Population genomics and diversity: from molecular ecology to transposable elements

Génomique des populations et diversité : de l'écologie moléculaire aux éléments transposables

Yann Bourgeois

Membres du Jury : Dr. Emmanuelle Lerat
Dr. Sylvain Glémin
Dr. Benoit Nabholz
Pr. Cristina Vieira-Heddi
Pr. Vincent Castric

Institut de Recherche pour le Développement
Université de Montpellier

Manuscrit soumis en vue de l'obtention de
l'Habilitation à Diriger des Recherches

UMR DIADE

November 2024

Déclaration

Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche.

Yann Bourgeois
November 2024

Acknowledgements

Thank you to the reviewers of this work for the time spent assessing it.
Thank you to my colleagues at UMR DIADE for welcoming me back in France in 2023, in particular the ones sharing my office, Cécile Triay and Julie Orjuela-Bouniol.
Thank you to my long-time friends, Anne Roulin, Muriel Gros-Balthazard, Joris Bertrand, Anaïs Gibert, Julien Rimour, Pauline Berthier, Pierre-Jean Malé, Marion Chartier.
Thank you to my colleagues from the University of Portsmouth for their support despite being clearly overworked: Steven Dodsworth, Adele Julier, Natalia Przelomska, Frank Schubert, Lena Grinsted, Sam Robson.
Thank you to my previous post-doctoral and PhD supervisors for their support over the years: Dieter Ebert, Stéphane Boissinot, Christophe Thébaud, Borja Milá.

To Florentina

"*Din încreţirea lungii rochii*
Răsai ca marmura în loc –
S-atârnă sufletu-mi de ochii
Cei plini de lacrimi şi noroc."

In *Atât de fragedă*, M. Eminescu.

Table of contents

List of figures

List of tables

Chapter 1

Extended *Curriculum Vitae*

1.1 CV, research grants and publications

PERSONAL INFORMATION

Birth date	June 12th 1987 (37 years old)
Web profiles	Scholar and Orcid
Nationality	French
Languages	French (native), English (fluent), German and Romanian (intermediate, B1)
Research Interests	Population genomics, bioinformatics, evolutionary biology, selection, transposable elements, host-parasite interaction, colour polymorphisms

EDUCATION

University of Toulouse, France *Sept. 2009- Aug. 2013*
Ph.D. student. Investigating the proximate causes of a color polymorphism in a passerine bird, *Zosterops borbonicus*. Defense: 23th of July 2013. Supervisor: Christophe Thébaud.

École Normale Supérieure de Lyon *Sept. 2007-Aug. 2009*
Masters in Biosciences (M1, M2)

École Normale Supérieure de Lyon *Sept. 2006-Aug. 2007*
Admission by competitive examination to École Normale Supérieure de Lyon, France. Bachelor of Science, major in Molecular and Cell Biology.

EMPLOYMENT HISTORY

Permanent Research Scientist *Since Jan. 2023*
Researcher at IRD (*Institut de Recherche pour le Développement*), Montpellier, France. Visiting lecturer at the University of Portsmouth.

Lecturer in Bioinformatics *Jan. 2020 - Dec. 2022*
Permanent lectureship (equivalent to associate professor) at the University of Portsmouth, UK

Post-doctoral researcher *Sept. 2016 - Dec. 2019*
Research associate in Pr. Stéphane Boissinot's group at New York University Abu Dhabi, United Arab Emirates. Evolutionary genomics of vertebrates using whole-genome sequences.

Post-doctoral researcher *Sept. 2013 - Aug. 2016*
Post-doctoral researcher in Pr. Dieter Ebert's group in Basel, Switzerland. Study of the evolutionary dynamics of host-parasite interaction in *Daphnia magna*.

RESEARCH PROJECTS AND GRANTS (MAIN INVESTIGATOR) ***CA €650,000***

French National Research Agency (ANR) ***426,159 euros*** *Jan. 2024 - Dec. 2027*
Research grant on the project **DaTEPalm**: "Determining the impact of Transposable Elements in a keystone crop, the date Palm (*Phoenix dactylifera*)".

The Royal Society, UK ***£19,920*** *Jan. 2023 - Dec. 2023*
Research grant on the project: "Studying transposable elements during domestication of the date palm (*Phoenix dactylifera*)"

NERC Environmental Omics Facility (NEOF, UK) ***£10,468*** *Sept. 2022 - Dec. 2022*
Research grant on the project: The evolutionary dynamics of transposable elements during domestication of the date palm (*Phoenix dactylifera*)"

University of Paris-Saclay, France ***3500 euros*** *Sept. 2022*
Visiting professor grant to visit Dr. Amandine Cornille (CNRS)

Horizon Europe ***190,000 euros*** *Feb. 2020*
Marie Curie Fellowship (Returning Investigator). Project **BrachyAdapt**: adaptation to the urban environment in *Brachypodium distachyon* (score of 92.8%). Declined to take a permanent position in Portsmouth.

University of Toulouse ***1500 euros*** *Apr. 2010*
ATUPS program travel grant from Toulouse University. One month in Hopi Hoekstra's laboratory, Harvard, United States.

French Ministry of Higher Education ***ca 50,000 euros*** *Sept. 2010 - Aug. 2013*
PhD grant for three years

French Ministry of Higher Education ***ca 67,000 euros*** *Sept. 2006 - Aug. 2010*
Grant covering Masters studies and first year of PhD.

RESEARCH PROJECTS AND GRANTS (CO-INVESTIGATOR) ***CA €1,000,000***

PlantAlliance/IRD (France) ***120,000 euros*** *Jun. 2024 - Sept. 2023*
Partner in the project Savanache, involving multiple private partners, including Syngenta and Florimon-Depre. Project focusing on the development of an advanced pan-genome visualization tool, allowing selection of individuals and projection onto an interchangeable reference. Lead PI: Dr. François Sabot.

I+D+i, Comunitat Valenciana, Spain ***ca 20,000 euros*** *Sept. 2022 - Sept. 2023*
Regional grant on the project: Functional connectivity in vertebrate populations under global change: simulation of impacts and mitigation measures. Lead PI: Dr. M. Victoria Jiménez Franco.

BBSRC, UK ***ca £70,000/student*** *Sept. 2021 - Sept. 2025*
Grants for two funded Ph.D. obtained from the South Coast Doctoral Training Program. I) Selection on transposable elements during independent events of domestication (with Dr. Bousios, Pr. Eyre-Walker, Sussex Uni). II) Phenotypic constraints on crop improvement and the domestication of novel crops (with Pr. Chapman, Southampton, and Pr. Perez-Barrales, Granada, Spain).

French National Research Agency (ANR) ***ca 444,000 euros*** *Jan. 2022 - Dec. 2025*
Partner in the project PLEASURE, on the Population genomics of transposable elements in fruit trees. Lead PI: Dr. Amandine Cornille, University of Paris-Saclay.

Conseil Régional de La Réunion (France) ***103,000 euros*** *Sept. 2020 - Aug. 2023*
Co-investigator on a conservation project focusing on the endangered Reunion harrier (*Circus maillardi*). Lead PI: Steve Augiron, SEOR. Role: investigating the mutational load and past history of the species.

COVID-19 Genomics UK Consortium ***£288,800*** *Apr. 2020 - Sept. 2022*
Co-investigator on a project investigating SARS-CoV2 dynamics in Portsmouth region. Lead PI: Samuel Robson, Portsmouth University.

Post-doctoral supervision *Since Feb. 2024*

Co-supervision (50%) of Dr. Qindong Tang and Dr. Samuel Gornard with Dr. Ben Warren, French Museum of Natural History (MNHN). Analysis of modern and past diversity of Mascarene birds from museum and subfossil samples.

Co-supervision of Ernesto Testé (10% with M. Gros-Balthazard, IRD). Investigation of date palm cultivation in the Levantine region, based on (ancient) DNA and seed morphometric analyses.

PhD supervision in France *Since 2023*

Funder and co-supervisor of Valentin Grenet (50%, with R. Guyot, HDR) on the population dynamics of transposons in date palm.

Co-supervision of two PhD students, Maxime Criado (10%, with A. Cornille, Université Paris Saclay) and Margot Beisseiche (30%, with M. Gros-Balthazard and F. Sabot HDR, IRD).

Masters student supervision in France *Jan. 2024 - Aug. 2024*

Main supervision of Valentin Grenet (Dynamics of transposons in date palm, 90%) and Kilian Dolci (Dynamics of transposons in *Coffea canephora*, 50%, with V. Poncet HDR, IRD).

PhD supervision in the UK *From Jan. 2021*

Main supervision of one PhD student, Thomas Heller (academic supervisor with J. Viruel, Kew Gardens). Supervision reallocated to Dr. Steven Dodsworth following departure from Portsmouth University.

Co-supervision of Anastasia Kolesnikova (10%, with M. Chapman, University of Southampton).

Co-supervision of Snata Chakraborty, Thomas Roberts-McEwen (10%, with L. Grinsted, University of Portsmouth).

Undergraduate and Masters student supervision in the UK *Jan. 2020 - Dec. 2023*

Supervision of more than 20 Honours projects (third year bachelor students), six Msc students, two MRes student (Harry Simmonds, Daniel Bedford).

Lectures in the UK *Jan. 2020 - Dec. 2022*

Lectures and workshops in Bioinformatics (Master level, module leader), population genetics (module leader), Python and R (undergraduate), functional genomics (undergraduate), and Introduction to Biology (120 hours/year).

Pastoral care of students and supervision of third year Honours projects (150 hours/year).

Student supervision at NYUAD *Sept. 2017 - Aug. 2019*

Supervision of an undergraduate student (Imtiyaz Hariyani): population genomics of transposable elements in a complex of Ethiopian frogs (genus *Ptychadena*). Evaluation of undergraduate research theses in population genetics.

Supervision of undergraduate students on the population genetics and signatures of balancing selection at immune genes in endemic Ethiopian frogs.

Student supervision at the University of Basel *Sept. 2014 - June 2015*

Basics in Population Genetics for bachelor students. 20 hours.

Participation to BlockKurse at the University of Basel. Co-supervision of four bachelor students for a short research project (three months).

Student supervision at the University of Toulouse *Oct. 2009 - Jun. 2012*

Supervision of bachelor students. Co-supervision of an ornithology module.

Animal anatomy and Plant organization practical classes for bachelor students (18 hours).

SCIENTIFIC SERVICE

Mentorship of PhD students

Significant contribution to scientific discussions, analyses and writing on the model *Brachypodium distachyon*: Nikolaos Minadakis and Christoph Stritt (PhD students at Zurich University). Main supervisor: Dr. Anne Roulin.

Contribution to scientific writing and interpretation of analyses on colour variation in *Zosterops borbonicus*: Claire Mould (PhD student at Toulouse University). Main supervisor: Pr. Christophe Thébaud.

Contribution to scientific writing, analyses and their interpretation on host-parasite interactions in *Daphnia magna*: Camille Ameline (PhD student at Basel University). Main supervisor: Pr. Dieter Ebert.

Supervision over genomic analyses and related scientific writing in *Testudo graeca*: Andrea Mira Joves (Miguel Hernández University at Elche). Main supervisor: Pr. Eva Gracia Martinez.

Workshops and conference organization

Physalia course on population genomics with Thibault Leroy (November 2024).

Organizer of a Journal Club in Evolutionary Biology in the DIADE research team at IRD (Since 2023).

Co-organizer of Symposium S15 at the ESEB congress, Prague, 2022. Rapid evolution of colour patterns.

Group leader for the Ecology and Evolution research group at the University of Portsmouth.

Organization of regular meetings, social events and research talks (2021-2022).
Co-organizer of the post-doctoral day in Basel (Sept. 2015), with short talks from post-doctoral researchers in biology.

Memberships
The Royal Genetics Society, The Society for Molecular Biology and Evolution, The Société Française d'Écologie et d'Évolution.

Review and editorial activities
Associate Editor for the Botanical Journal of the Linnean Society (from 2023).
Reviewer for Annals of Botany, The New Phytologist, Nature Communications, Scientific Reports, eLife, Genome Biology, PCI Genomics, Genome Biology and Evolution, Molecular Biology and Evolution, Science Advances, Ibis, The Auk, PeerJ, Genetica, Molecular Ecology, Molecular Ecology Resources, BMC Evolutionary Biology, Global Change Biology, Biology Letters, Genetics and Molecular Biology.
Grant Reviewer for the National Science Foundation (USA), the Czech Science Foundation, the Hong Kong Research Grants Council and the US-Israel Binational Science Foundation.

Involvement in research consortium
Involved in COG-UK, a network aiming at analyzing COVID-19 genome sequences in the United Kingdom.

Involvement in committees
External jury member (2024) for a position on transposable elements in yam based at CIRAD (French agricultural research and cooperation organization).
Assessor for PhD. Committees: Laurie Bédouet (University of Franche Comté), Gwenaelle Vigo (University of Montpellier), Francesca Noyce (University of Portsmouth), Christina Scott (University of Portsmouth), Nikolaos Minadakis (Botanical Institute, University of Zurich), Matthieu Breil (University of Montpellier), James Robbins (University of Portsmouth).
Examiner for PhD viva (2020): Werner Struss (University of Portsmouth).
Representative of Ph.D. students in Toulouse (2010).

MISCELLANEOUS

Certifications
Animal manipulation certifications: The CITI Basic Course in Laboratory Animal Welfare

for Investigators, Staff and Students, Reducing Pain and Distress in Laboratory Mice and Rats, Working with Amphibians in Research Settings (2018).

Fieldwork

Five months of fieldwork on leatherback turtles in French Guiana (2008).
Two months of fieldwork on endemic birds in Reunion island (2009-2014).
Three weeks of fieldwork in Siberia (collection of *Daphnia magna* samples, 2015).
Three months of fieldwork in Ethiopia (bird and amphibian collection, 2016-2018).

PEER-REVIEWED PUBLICATIONS

---

* *Co-first author.*

$\vee$ *: Publication of interest.*

$\wedge$ *: These publications were carried out as part of the COG-UK consortium, dedicated to the study of the SARS-COV2 pandemic in the United Kingdom, and do not constitute the core of my research. However, they do demonstrate my involvement in the collection and analysis of phylodynamic data in Portsmouth during the SARS-COV-2 epidemic. The reader can refer to the section on past collaborations and research projects for more information.*

The names of PhD students whom I contributed to supervise during the writing of a given article are highlighted in orange

2024

(52) **Andrea Mira-Jover**, A., Graciá, E., Giménez, A., Fritz, U., Rodríguez-Caro, R.C., **Bourgeois, Y.** (2024). Taking advantage of reference-guided assembly in a slowly-evolving lineage: application to *Testudo graeca*. *PloS One*.

(51) **Minadakis, N.**, Kaderli, L., Horvath, R., **Bourgeois, Y.**, Xu, W., Thieme, M., Woods, D. P. & Roulin, A. C. Polygenic architecture of flowering time and its relationship with local environments in the grass *Brachypodium distachyon*. Accepted in *Genetics*.

(50) Horvath, R., **Minadakis, N.**, **Bourgeois, Y.** & Roulin, A. C. The evolution of transposable elements in *Brachypodium distachyon* is governed by purifying selection, while neutral and adaptive processes play a minor role. *eLife* 12. https://doi.org/10.7554/eLife.93284.3 (2024) (Feb. 2024). Version of record.

(49) **Bourgeois, Y.**, Warren, B., Augiron, S. (2024). The burden of anthropogenic changes and mutation load in a critically endangered harrier from the Reunion biodiversity hotspot, *Circus maillardi*. In press in *Molecular Ecology*.

(48) Reifová, R., Ament-Velásquez, S. L., **Bourgeois, Y.**, Coughlan, J., Kulmuni, J., Lipinska, A. P., Okude, G., Stevison, L., Yoshida, K., & Kitano, J. (2023, ahead of press). Mechanisms of Intrinsic Postzygotic Isolation: From Traditional Genic and Chromosomal Views to Genomic and Epigenetic Perspectives. *Cold Spring Harbor Perspectives in Biology*, *15*(10), a041607.

2023

(47) **Minadakis, N.**, Williams, H., Horvath, R., Caković, D., **Stritt, C.**, Thieme, M., **Bourgeois, Y.**, & Roulin, A. C. (2023). The demographic history of the wild crop relative *Brachypodium distachyon* is shaped by distinct past and present ecological niches. *Peer Community Journal*, *3*.

(46) **Mould, M. C.**, Huet, M., Senegas, L., Milá, B., Thébaud, C., **Bourgeois, Y.**, & Chaine, A. S. (2023). Beyond morphs: Inter-individual colour variation despite strong genetic determinism of colour morphs in a wild bird. *Journal of Evolutionary Biology*, *36*(1), 82–94.

(45) Gros-Balthazard, M., Battesti, V., Flowers, J. M., Ferrand, S., Breil, M., Ivorra, S., Terral, J.-F., Purugganan, M. D., Wing, R. A., & Mohammed, N., **Bourgeois, Y.** (2023). What lies behind a fruit crop variety name? A case study of the barnī date palm from al-ʻUlā oasis, Saudi Arabia. *Plants, People, Planet*, *5*(1), 82–97. ∨

(44) Cotton, S., McHugh, M. P., Dewar, R., Haas, J. G., Templeton, K., Robson, S. C., Connor, T. R., Loman, N. J., Golubchik, T., Nunez, R. T. M., **The COG-UK consortium** (2023). Investigation of hospital discharge cases and SARS-CoV-2 introduction into Lothian care homes. *Journal of Hospital Infection*, *135*, 28–36. ∧

2022

(43) Manthey, J. D., **Bourgeois, Y.**, Meheretu, Y., & Boissinot, S. (2022). Varied diversification patterns and distinct demographic trajectories in Ethiopian montane forest bird (Aves: Passeriformes) populations separated by the Great Rift Valley. *Molecular Ecology*, *31*(9), 2664–2678.

(42) Goutte, S., Hariyani, I., Utzinger, K. D., **Bourgeois, Y.**, & Boissinot, S. (2022). Genomic Analyses Reveal Association of *ASIP* with a Recurrently evolving Adaptive Color Pattern in Frogs. *Molecular Biology and Evolution*, *39*(11), msac235. ∨

(41) Thieme, M., Brêchet, A., **Bourgeois, Y.**, Keller, B., Bucher, E., & Roulin, A. C. (2022). Experimentally heat-induced transposition increases drought tolerance in *Arabidopsis thaliana*. *New Phytologist*, *236*(1), 182–194. ∨

(40) Twohig, K. A., Nyberg, T., Zaidi, A., Thelwall, S., Sinnathamby, M. A., Aliabadi, S., Seaman, S. R., Harris, R. J., Hope, R., Lopez-Bernal, J., **The COG-UK consortium** (2022). Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B. 1.617. 2) compared with alpha (B. 1.1. 7) variants of concern: A cohort study. *The Lancet Infectious Diseases*, *22*(1), 35–42. ∧

2021

(39) Almojil, D., **Bourgeois, Y.**, Falis, M., Hariyani, I., Wilcox, J., & Boissinot, S. (2021). The structural, functional and evolutionary impact of transposable elements in eukaryotes. *Genes*, *12*(6), 918.

(38) **Bourgeois, Y**, Fields, P. D., Bento, G., & Ebert, D. (2021). Balancing selection for pathogen resistance reveals an intercontinental signature of Red Queen coevolution. *Molecular Biology and Evolution*, *38*(11), 4918–4933. ∨

(37) **Ameline, C.**, **Bourgeois, Y.**, Vögtli, F., Savola, E., Andras, J., Engelstädter, J., & Ebert, D. (2021). A two-locus system with strong epistasis underlies rapid parasite-mediated evolution of host resistance. *Molecular Biology and Evolution*, *38*(4), 1512–1528.

(36) **Bourgeois, Y. X.**, & Warren, B. H. (2021). An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Molecular Ecology*, *30*(23), 6036–6071. ∨

(35) Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., Nascimento, F. F., **The COG-UK consortium** (2021). Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell*, *184*(1), 64-75. e11. ∧

(34) Graham, M. S., Sudre, C. H., May, A., Antonelli, M., Murray, B., Varsavsky, T., Kläser, K., Canas, L. S., Molteni, E., Modat, M., **The COG-UK consortium** (2021). Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B. 1.1. 7: An ecological study. *The Lancet Public Health*, *6*(5), e335–e345. ∧

(33) Elliott, P., Haw, D., Wang, H., Eales, O., Walters, C. E., Ainslie, K. E., Atchison, C., Fronterre, C., Diggle, P. J., Page, A. J., **The COG-UK consortium** (2021). Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science*, *374*(6574), eabl9551. ∧

(32) de Silva, T. I., Liu, G., Lindsey, B. B., Dong, D., Moore, S. C., Hsu, N. S., Shah, D., Wellington, D., Mentzer, A. J., Angyal, A., **The COG-UK consortium** (2021). The

impact of viral mutations on recognition by SARS-CoV-2 specific T cells. *Iscience*, *24*(11). ∧

(31) Meng, B., Kemp, S. A., Papa, G., Datir, R., Ferreira, I. A., Marelli, S., Harvey, W. T., Lytras, S., Mohamed, A., Gallo, G., **The COG-UK consortium** (2021). Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B. 1.1. 7. *Cell Reports*, *35*(13). ∧

2020

(30) **Bourgeois, Y.**, Ruggiero, R. P., Hariyani, I., & Boissinot, S. (2020). Disentangling the determinants of transposable elements dynamics in vertebrate genomes using empirical evidences and simulations. *PLoS Genetics*, *16*(10), e1009082. ∨

(29) **Bourgeois, Y. X.**, Bertrand, J. A., Delahaie, B., Holota, H., Thébaud, C., & Milá, B. (2020). Differential divergence in autosomes and sex chromosomes is associated with intra-island diversification at a very small spatial scale in a songbird lineage. *Molecular Ecology*, *29*(6), 1137–1153. ∨

(28) **The COG-UK consortium**. (2020). An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe*, *1*(3), e99. ∧

2019

(27) Boissinot, S., **Bourgeois, Y.**, Manthey, J. D., & Ruggiero, R. P. (2019). The mobilome of reptiles: Evolution, structure, and function. *Cytogenetic and Genome Research*, *157*(1–2), 21–33.

(26) **Bourgeois, Y.**, & Boissinot, S. (2019a). On the population dynamics of junk: A review on the population genomics of transposable elements. *Genes*, *10*(6), 419. ∨

(25) **Bourgeois, Y.**, & Boissinot, S. (2019b). Selection at behavioural, developmental and metabolic genes is associated with the northward expansion of a successful tropical colonizer. *Molecular Ecology*, *28*(15), 3523–3543.

(24) **Bourgeois, Y.**, Ruggiero, R. P., Manthey, J. D., & Boissinot, S. (2019). Recent secondary contacts, linked selection, and variable recombination rates shape genomic diversity in the model species *Anolis carolinensis*. *Genome Biology and Evolution*, *11*(7), 2009–2022. ∨

2018

(23) **Bourgeois, Y.**, **Stritt, C.**, Walser, J.-C., Gordon, S. P., Vogel, J. P., & Roulin, A. C. (2018b). Genome-wide scans of selection highlight the impact of biotic and abiotic

constraints in natural populations of the model grass Brachypodium distachyon. *The Plant Journal*, *96*(2), 438–451. ∨

(22) Reyes-Velasco, J., Manthey, J. D., **Bourgeois, Y.**, Freilich, X., & Boissinot, S. (2018). Revisiting the phylogeography, demography and taxonomy of the frog genus Ptychadena in the Ethiopian highlands with the use of genome-wide SNP data. *PloS One*, *13*(2), e0190440.

(21) Roman, I.*, **Bourgeois, Y.***, Reyes-Velasco, J., Jensen, O. P., Waldman, J., & Boissinot, S. (2018). Contrasted patterns of divergence and gene flow among five fish species in a Mongolian rift lake following glaciation. *Biological Journal of the Linnean Society*, *125*(1), 115–125.

(20) Toenshoff, E. R., Fields, P. D., **Bourgeois, Y. X.**, & Ebert, D. (2018). The end of a 60-year riddle: Identification and genomic characterization of an iridovirus, the causative agent of white fat cell disease in zooplankton. *G3: Genes, Genomes, Genetics*, *8*(4), 1259–1272.

2017

(19) Bento, G., Routtu, J., Fields, P. D., **Bourgeois, Y.**, Du Pasquier, L., & Ebert, D. (2017). The genetic basis of resistance and matching-allele interactions of a host-parasite system: The *Daphnia magna-Pasteuria ramosa* model. *PLoS Genetics*, *13*(2), e1006596. ∨

(18) **Bourgeois, Y. X.**, Delahaie, B., Gautier, M., Lhuillier, E., Malé, P.-J. G., Bertrand, J. A., Cornuault, J., Wakamatsu, K., Bouchez, O., & Mould, C. (2017). A novel locus on chromosome 1 underlies the evolution of a melanic plumage polymorphism in a wild songbird. *Royal Society Open Science*, *4*(2), 160805. ∨

(17) **Bourgeois, Y.***, Roulin, A. C.*, Müller, K., & Ebert, D. (2017). Parasitism drives host genome evolution: Insights from the *Pasteuria ramosa–Daphnia magna* system. *Evolution*, *71*(4), 1106–1113.

(16) Delahaie, B., Cornuault, J., Masson, C., Bertrand, J. A., **Bourgeois, Y. X.**, Milá, B., & Thébaud, C. (2017). Narrow hybrid zones in spite of very low population differentiation in neutral markers in an island bird species complex. *Journal of Evolutionary Biology*, *30*(12), 2132–2145.

(15) Ruggiero, R. P.*, **Bourgeois, Y.***, & Boissinot, S. (2017). LINE insertion polymorphisms are abundant but at low frequencies across populations of *Anolis carolinensis*. *Frontiers in Genetics*, *8*, 44. ∨

2016

(14) Roulin, A. C., **Bourgeois, Y.**, Stiefel, U., Walser, J.-C., & Ebert, D. (2016). A photoreceptor contributes to the natural variation of diapause induction in Daphnia magna. *Molecular Biology and Evolution*, *33*(12), 3194–3204.

(13) **Bourgeois, Y. X.**, Bertrand, J. A., Delahaie, B., Cornuault, J., Duval, T., Milá, B., & Thébaud, C. (2016a). Candidate gene analysis suggests untapped genetic complexity in melanin-based pigmentation in birds. *Journal of Heredity*, *107*(4), 327–335.

(12) Bertrand, J. A.*, **Bourgeois, Y. X.***, & Thébaud, C. (2016). Population density of the Réunion Grey White-eye Zosterops borbonicus within the summit ecosystems of Réunion, Mascarene Islands. *Ostrich*, *87*(1), 85–88.

(11) Besnard, G., Bertrand, J. A., Delahaie, B., **Bourgeois, Y. X.**, Lhuillier, E., & Thébaud, C. (2016). Valuing museum specimens: High-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura). *Biological Journal of the Linnean Society*, *117*(1), 71–82.

(10) Bertrand, J. A., Delahaie, B., **Bourgeois, Y. X.**, Duval, T., García-Jiménez, R., Cornuault, J., Pujol, B., Thébaud, C., & Milá, B. (2016). The role of selection and historical factors in driving population differentiation along an elevational gradient in an island bird. *Journal of Evolutionary Biology*, *29*(4), 824–836.

2015

(9) van de Crommenacker, J., **Bourgeois, Y. X. C**., Warren, B. H., Jackson, H., Fleischer-Dogley, F., Groombridge, J., & Bunbury, N. (2015). Using molecular tools to guide management of invasive alien species: Assessing the genetic impact of a recently introduced island bird population. *Diversity and Distributions*, *21*(12), 1414–1427.

(8) Casquet, J., **Bourgeois, Y. X.**, Cruaud, C., Gavory, F., Gillespie, R. G., & Thébaud, C. (2015). Community assembly on remote islands: A comparison of Hawaiian and Mascarene spiders. *Journal of Biogeography*, *42*(1), 39–50.

(7) Cornuault, J., Delahaie, B., Bertrand, J. A., **Bourgeois, Y. X.**, Milá, B., Heeb, P., & Thébaud, C. (2015). Morphological and plumage colour variation in the Réunion grey white-eye (Aves: Zosterops borbonicus): Assessing the role of selection. *Biological Journal of the Linnean Society*, *114*(2), 459–473.

2014

(6) Bertrand, J. A. M., **Bourgeois, Y. X. C.**, Delahaie, B., Duval, T., García-Jiménez, R., Cornuault, J., Heeb, P., Milá, B., Pujol, B., & Thébaud, C. (2014). Extremely reduced dispersal and gene flow in an island bird. *Heredity*, *112*(2), 190–196.

2013

(5) **Bourgeois, Y. X.**, Lhuillier, E., Cézard, T., Bertrand, J. A., Delahaie, B., Cornuault, J., Duval, T., Bouchez, O., Milá, B., & Thébaud, C. (2013). Mass production of SNP markers in a nonmodel passerine bird through RAD sequencing and contig mapping to the zebra finch genome. *Molecular Ecology Resources*, *13*(5), 899–907. ∨

(4) Cornuault, J., Khimoun, A., Harrigan, R. J., **Bourgeois, Y. X.**, Milá, B., Thébaud, C., & Heeb, P. (2013). The role of ecology in the geographical separation of blood parasites infecting an insular bird. *Journal of Biogeography*, *40*(7), 1313–1323.

2012

(3) Warren, B. H., Bermingham, E., **Bourgeois, Y.**, Estep, L. K., Prys-Jones, R. P., Strasberg, D., & Thébaud, C. (2012). Hybridization and barriers to gene flow in an island bird radiation. *Evolution*, *66*(5), 1490–1505.

(2) Bertrand, J. A., García-Jiménez, R., **Bourgeois, Y.**, Duval, T., Heeb, P., Thébaud, C., & Milá, B. (2012). Isolation and characterization of twelve polymorphic microsatellite loci for investigating an extreme case of microgeographical variation in an island bird (*Zosterops borbonicus*). *Conservation Genetics Resources*, *4*, 323–326.

(1) **Bourgeois, Y. X.**, Bertrand, J. A., Thebaud, C., & Mila, B. (2012). Investigating the role of the melanocortin-1 receptor gene in an extreme case of microgeographical variation in the pattern of melanin-based plumage pigmentation. *PLoS One*, *7*(12), e50906.

BIBLIOMETRICS

The official instructions from the GAIA doctoral school (link) require that the impact factor of journals be indicated. This is in clear violation of the San Francisco Declaration On Research Assessment (DORA) that is now followed by most modern universities. Metrics were extracted from *Google Scholar* using the get_journalrank() function in the *R* package "scholar". The function was applied on all journals extracted with the get_publications() function, using the profile ID "p61wpJUAAAAJ".

Table 1.1 Summary of journal metrics

Journal	Scimago Journal Rank	Scimago Best Quartile	Impact Factor
Biological Journal of the Linnean Society	0.771	Q1	2.17
Cell	25.716	Q1	45
Cell Reports	4.845	Q1	9.64
Cold Spring Harbor Perspectives in Biology	4.738	Q1	9.48
Conservation Genetics Resources	0.283	Q3	0.98
Cytogenetic and Genome Research	0.435	Q3	1.99
Diversity and Distributions	1.688	Q1	5.52
eLife	4.752	Q1	7.94
Evolution	1.56	Q1	3.28
Frontiers in Genetics	1.096	Q2	4.37
G3: Genes, Genomes, Genetics	1.09	Q1	3.08
Genes	1.032	Q2	3.96
Genome Biology and Evolution	1.441	Q1	3.69
Heredity	1.107	Q2	3.56
Iscience	1.592	Q1	5.74
Journal of Biogeography	1.607	Q1	4.81
Journal of Evolutionary Biology	0.934	Q1	2.27
Journal of Heredity	0.869	Q2	2.37
Journal of Hospital Infection	1.333	Q1	6.54
Molecular Biology and Evolution	5.341	Q1	6.49
Molecular Ecology	1.96	Q1	5.55
Molecular Ecology Resources	2.496	Q1	8.64
New Phytologist	3.009	Q1	9.23
Ostrich	0.435	Q3	1.08
Peer Community Journal: Evolutionary Biology	NA	NA	NA
Plants, People, Planet	1.009	Q1	4.6
PLoS Genetics	2.7	Q1	5.36
PLoS One	0.852	Q1	3.58
Royal Society Open science	0.758	Q1	3.41
Science	14.589	Q1	15.19
The Lancet Infectious Diseases	10.236	Q1	15.66
The Lancet Public Health	11.372	Q1	18.59
The Lancet Microbe	13.313	Q1	16.57
The Plant Journal	2.101	Q1	6.19

Bourgeois Y., **Grenet V.**, Gros-Balthazard M. (2024). Advances in the study of population dynamics of transposable elements and application to date palm genomics. XXth International Botanical Congress (IBC 2024, Madrid).

Bourgeois Y., Boissinot S. (2022). Population genomics perspective of transposable elements (TEs) dynamics: new methods and lines of research. Invited seminar at Sussex University (UK).

Bourgeois Y., Boissinot S. (2022). Population genomics perspective of transposable elements (TEs) dynamics: new methods and lines of research. Invited seminar at IDEEV, Paris Saclay (France).

Bourgeois Y., Boissinot S. (2021). A population genomics perspective on transposable elements (TEs) dynamics. Genetrop seminar, IRD Montpellier (France).

Bourgeois Y., Boissinot S. (2020). A population genomics perspective on transposable elements (TEs) dynamics. Invited seminar at the University of Durham (UK).

Bourgeois Y., Boissinot S. (2018). Population genomics of a successful colonizer: linking molecular approaches to ecology in a squamate. Speciation meeting, IST Austria, Vienna.

Bourgeois Y., Stephane Boissinot. (2018). Population genomics of the green anole reveals evolutionary forces shaping diversity in a reptile. NYU Abu Dhabi Research Conference, United Arab Emirates.

Bourgeois Y., Fields P., Bento G., Roulin A., McTaggart S., Little T., Obbard D., Ebert D. (2015). Increased diversity at a locus involved in resistance to parasitism in Daphnia magna. ESEB 2015, Lausanne.

Bento G., Routtu J., **Bourgeois Y.**, Ebert D. (2015). Genetics of natural variation of Daphnia magna resistance to a bacterial pathogen. ESEB 2015, Lausanne.

Ebert D., Routtu J., Bento G., **Bourgeois Y.** (2014). Mapping of a parasite resistant locus in the Daphnia magna genome. EMBO Conference on the Mighty Daphnia: Past, Present and Future. Birmingham.

Bourgeois Y.X.C., SNPs characterization by RAD-sequencing for studying color polymorphism in an island bird (2013). Journée d'échanges et de retours sur les développements technologiques de la plateforme génomique de Toulouse. Presentation at the Toulouse genomic platform.

Bourgeois Y.X.C., Boissinot S. (2019). Population genomics of transposable elements in the green anole. SMBE Manchester

Bourgeois Y.X.C., Boissinot S., Manthey J., Ruggiero R., Reyes-Velasco J. (2018). Population genomics of green anole (*Anolis carolinensis*) reveals evolutionary forces shaping diversity in a reptile. Evolution Joint Congress Montpellier

Bourgeois Y.X.C., Fields P., Bento G., Ebert D. (2017). Widespread balancing selection at a resistance locus in the water flea *Daphnia magna*. ESEB Groningen

Milá B., **Bourgeois Y.**, Bertrand J., Cornuault J., Delahaie B., Thébaud C. (2015). Divergent selection and reduced dispersal drive phenotypic diversification at a very small spatial scale in an island bird. ESEB 2015, Lausanne.

Bento G., Routtu J., **Bourgeois Y.X.C.**, Hall M., Kaberer N., Ebert D. (2014). Host - pathogen coevolution in the *Daphnia magna -Pasteuria ramosa* system. EMBO Conference on the Mighty Daphnia: Past, Present and Future. Birmingham.

Milá B, **Bourgeois Y.X.C.**, Bertrand J.A.M., Delahaie B. Thébaud C (2013). Inter- and intra-island speciation in a tropical passerine bird: inference from genetic, genomic and ecomorphological data. 4th Meeting of the Spanish Society for Evolutionary Biology (SESBE) - Barcelone, Spain.

Bourgeois Y.X.C., Milà B., Thébaud C. (2013). RADseq phylogenomics reveal the recent diversification history of a polymorphic songbird (*Zosterops borbonicus*) on the island of Reunion. II Iberian Congress of Biological Systematics, Barcelona, Spain.

Bourgeois Y.X.C., Bertrand J., Duval T., Warren B.H., Milà B., Thébaud C. (2012). Genetic mechanisms driving to melanic polymorphism in an island bird. Evolution Congress, Ottawa, Canada.

Bertrand J.A.M., **Bourgeois Y.X.C.**, Duval T., García-Jiménez R., Cornuault J., Milá B., Thébaud C. (2012). Selection-constrained dispersal drives fine-scale genetic differentiation in an island bird (*Zosterops borbonicus*). Evolution Congress, Ottawa, Canada.

1.2 Reflexive account of student and project supervision

1.2.1 Supervision and teaching philosophy

Over the course of my career, I have engaged with a broad diversity of undergraduate and postgraduate students, acting either as a project supervisor or as a teacher. My personal

experience is that mentoring can occur in a simultaneous state of excitement and frustration. Engaging with students in discussions about fundamental and technical concepts in biology is both stimulating and humbling, and has been one of the best experiences in my academic life. Established researchers have a (moral) duty towards their students or post-doctoral colleagues to bring them as close as possible to their own goals. However, in academia, the professional development of a researcher relies on their students. Supervisors often forget how long it took them to master a subject. Results are expected, schedules are to be followed, administrative tasks take time, productivity is paramount. This time pressure can generate frustration and divert the supervisor from their primary role, and sometimes contributes to the unfortunate perception of teaching as a distraction from more productive endeavours. I have learned that in managing research projects, it is crucial to maintain a broader, long-term perspective that **focuses on the individuals being supervised**.

Most of my teaching has been developed at the University of Portsmouth (UK), where I have worked as a lecturer for three years. Teaching load was heavy, and involved substantial pastoral care. Instead of considering research and teaching as two disconnected aspects, I tried to blend them as much as possible. I used teaching preparation time to review current literature, which gave me ideas for my own research as well as material to design up-to-date workshops and projects for students. For example, my current ANR project (*DaTEPalm*) has **benefited from the development of lectures** in functional genomics. These lectures introduced long-read sequencing technology and concepts revolving around multi-omics integration that substantially strengthened the project.

I planned and designed my bioinformatics and genomics lectures with a strong connection to research in mind. I found particularly rewarding how students appropriated themselves the core skills, and developed a critical perspective when interpreting their results instead of simply running scripts. I have kept this approach in the **design of projects** involving Masters and PhD students. In my supervision, I tend to focus on two aspects: i) provide a broader evolutionary context to the analyses undertaken, so students adopt a question-oriented approach; ii) introduce the most recent approaches available to extract relevant information from genomic data.

Teaching and pastoral care has made me acutely aware of the impact that supervisors may have on the future of their students. When planning a research project, I aim at assigning tasks so risks and benefits are balanced in a way that I deem appropriate given the career stage. For **Master students**, who typically have short-term projects, I consider that most if not all data have to be available before they start. The project itself must be on a topic in which I am deeply knowledgeable, both theoretically and practically. This greatly improves efficiency and helps to keep the students engaged. Although these requirements can be

relaxed for **PhD projects**, I consider important to have a **core set of data** that are already available, as well as **clear questions** that will ensure a good outcome for the student.

I have recently obtained funding to hire a **post-doctoral researcher** (see ANR project in Chapter 3). Considering that post-doctoral researchers are already independent and autonomous, I will rely on them to develop research questions and slightly riskier protocols that I would not easily master myself. I intend to involve them in the **supervision of students**, as well as supporting them in their **applications to grants** and **permanent positions**. I have retained the idea from my previous international experience that projects involving post-doctoral researchers must include a component that they can incorporate into their own research programme.

1.2.2 Open Science and Diffusion of methods

Current anthropic changes will have a dramatic impact on biodiversity. The growing availability of high-quality genomic datasets promises to yield critical insights into the adaptive potential of populations. It also fuels the much-needed move beyond the current set of model organisms. The deployment of genomic approaches will be essential to understand the past and future dynamics of new model species, such as traditional crops in the South, or endangered species. Unfortunately, this knowledge is still very poorly communicated to a broader audience. The pressing issue of anthropogenic change deserves enhanced efforts to disseminate methods that can leverage genomic data, ultimately improving our understanding of the response of biodiversity to environmental change. Many practitioners are receptive to using genetic tools, but do not always have access to the relevant expertise. Because of these limitations, I actively share my own experience in using population genomic tools. I have worked on species from major biodiversity hotspots, such as amphibians from Ethiopian highlands and endemic birds from Réunion island (*1*) in the Indian Ocean. I have used genome-wide markers to reconstruct the past history of populations and understand how genetic variation may be adaptive or deleterious. I share workshops and scripts based on this work on my Github account, sharing scripts on phylodynamic analyses (link), or landscape genomics (link). Based on this experience, I recently wrote a review addressed to practitioners to help choosing methods (*2*). I host and manage a website (www.methodspopgen.com) where I list the most recent tools developed in population genomics. I am committed to the FAIR (Findability, Accessibility, Interoperability, and Reuse) principles when sharing datasets and scripts.

1.2.3 Masters supervision

I started supervising post-graduate students in an official capacity in 2020, during my first year as a lecturer in Portsmouth. There are two main types of Masters degrees in the UK: Master of Sciences (*Msc*), and Master of Research (*MRes*). *Msc* are mostly taught courses, with a short research project effectively lasting two months. I supervised six *Msc* students between 2020 and 2022. Most of these students were foreigners, and could not reach the UK due to travel restrictions during the SARS-CoV-2 pandemic. I designed projects based on the analysis of publicly available data that could run on personal computers, and also arranged for the students to have access to the Sciama computing cluster at the University of Portsmouth. Most of these projects revolved around the phylodynamic analysis of SARS-CoV-2 epidemics in particular cities and countries, taking advantage of databases such as **GISAID** and Nexstrain. I provided as an example a preliminary analysis (see **Github**) that I carried out in the context of my participation to the COG-UK consortium (see also CV and Chapter 2).

In the context of this *HDR* report, *MRes* supervision is more directly relevant. Students usually spend a whole year on their research project. This duration echoes the six months that second year Master students (M2) spend as trainees in a host laboratory in France. In Portsmouth, students were evaluated on their ability to write a grant proposal describing their project, a literature report, as well as a research report written in the format of a scientific article. I supervised two *MRes* students, **Harry Simmonds** (2020-2021) and **Daniel Bedford** (2021-2022). In the remaining sections, boxes summarize students' projects on which I acted as main supervisor.

Harry Simmonds

Project: The *MRes* project consisted in investigating divergence and diversity at flowering genes across multiple populations of pale flax (*Linum bienne*). Mr Simmond's work suggests that recent positive selection occurred at a few well-known flowering genes, such as *TFL1*. The results he obtained will be included in a future publication describing the genetic diversity and past demographic history of pale flax, the closest relative to domesticated flax (*Linum usitatissimum*).
Professional outcome: Mr Simmonds is now a **PhD student** at the **University of Reading**, where he will investigate strategies to improve crop resistance to disease by taking into account diversity at plant resistance genes before applying fungicides.

Daniel Bedford

Project: The project consisted in investigating transposable elements (TE) diversity in a clade of "great speciators", the White-Eyes (Zosterops genus). Recent TE activity is mostly restricted to LTR retrotransposons in this clade. A phylogeny based on TEs was consistent with the ones obtained with SNP variants. Island species displayed a higher amount of TEs compared to continental species, although this difference was not significant due to a limited number of species with resequencing data available. A possible explanation could be the accumulation of TE load in island populations due to their lower effective population size.
Professional outcome: Mr Bedford is now a **Research analyst** at Juniper Research, a company offering consulting services in the telecom trade.

I recently obtained funding from the Agence Nationale de la Recherche to investigate the dynamics of transposable elements in date palms (*Phoenix dactylifera*). This project provides funding for a Masters student, **Valentin Grenet**, who started his project in February 2024. His rapid progress has led to him pursuing a PhD under my supervision.

Valentin Grenet

Project: Valentin Grenet was a M2 student at Polytech Nice Sophia. His project consisted in annotating LTR retrotransposons in the date palm reference genome. He has succeeded in obtaining consensus sequences for the most abundant lineages of LTR-RTs, has described their diversity through phylogenetic methods (see Figure 3.6 in Chapter 3), and has estimated the age of complete LTR-RTs and solo LTRs. The age and abundance of elements will be contrasted with genomic features (recombination, gene density, base content etc.) to investigate the interaction between TEs and their host. Nanopore data will be used to further investigate the abundance and frequency of polymorphic elements.
Professional outcome: Valentin has started his PhD under my supervision (HDR supervisor, R. Guyot).

I have also been involved in the supervision of **Kilian Dolci** , in collaboration with Valérie Poncet (*Institut de Recherche pour le Développement*).

Kilian Dolci

Project: Kilian Dolci was a M2 student at Toulouse University. His project focused on polymorphic TEs in divergent populations of *Coffea canephora*. He used MEGAnE (*3*) to call polymorphic TEs in a set of more than 70 individuals, identified candidate TEs for adaptation and quantified TE load in domesticated cultivars. He compared frequency spectra and TE diversity across genetic groups, and ran a genome-wide scan of association between TEs and environmental conditions using *LFMM2* (*4*)
Professional outcome: Applying to PhD positions.

1.2.4 PhD supervision

In 2021 and 2022, I have been the main supervisor of a part-time PhD student, **Thomas Heller**, based at **Kew Gardens**. The supervisors at Kew were Dr. Juan Viruel and Dr. Martin Hamilton. Mr Heller has worked at Kew gardens as a research engineer for several years, giving him independence in fieldwork and ecological assessments. My contribution was mostly focusing on report writing and population genetic analyses (using Angiosperm353 kits(*5*) and RAD-sequencing(*6*)).

Thomas Heller

Project: Quantifying the genetic diversity of a threatened endemic Caribbean tree, *Zanthoxylum thomasianum*.
The project focuses on threatened plants in the Caribbean, with active projects in the Virgin Islands (British and US) and Puerto Rico. Mr Heller is particularly interested in the evolution of the genus Zanthoxylum (Rutaceae) in the Caribbean through field study and molecular approaches (population genetics and phylogenomics). Molecular aspects of the project involve phylogenetic analyses to clarify the long-term evolutionary history of the genus, using well as population genetics/landscape genetics analyses in *Z. thomasianum*, to identify threats due to lack of connectivity and recent loss of genetic diversity
Professional outcome: In progress.

Due to my appointment at the *Institut de Recherche pour le Développement* in 2023, Mr Heller's main supervisor at the University of Portsmouth is now my colleague, Dr. Steven Dodsworth. I am nevertheless still involved in Mr Heller's supervision. At the time of my departure, the project was well advanced, with nearly 400 samples collected across Puerto

Rico and the British Virgin Islands, and clear prospects for further analyses. We expect sequencing data in 2024, and publication of results around 2025.

Since September 2022 I have also been involved in the supervision of **Anastasia Kolesnikova** at the **University of Southampton**, with Pr. Mark Chapman as main supervisor. Her project focuses on comparing genetic diversity between domesticated plants and their undomesticated relatives, to determine whether domesticated plants exhibit desirable properties in terms of mutation rate, mutation load, and pre-existing variation upon which artificial selection could act.

I am still involved in collaborations at the **University of Portsmouth**. I am acting as co-supervisor for Ms **Snata Chakraborty** and Mr **Thomas Roberts-McEwen**, with Dr. Lena Grinsted as main supervisor. Their projects revolve around the ecology and evolution of *Cyrtophora citricola*, a species of group-living spiders that display low intra-specific aggressiveness. Mr Roberts-McEwen's project will focus on using this species for pest biocontrol in South-East Asia. Ms Chakraborty will focus more on the evolution of the species, building on RAD-sequencing results collected by another undergraduate student I supervised, Mr Nathaniel Holmes.

I have benefited from **extensive supervisor training** in the UK, received from the University of Portsmouth (see resources here), as well as from the BBSRC/SoCoBio doctoral school. I have also acted as an examiner in several PhD committees (see CV).

Since January 2023 I have been **co-supervising** two PhD students in **France**. At the *Institut de Recherche pour le Développement*, I am supervising **Margot Beisseiche** (25%), with Dr. Muriel Gros-Balthazard and Dr. François Sabot (HDR). Ms Beisseche's project focuses on collecting ancient DNA from date palm remains to retrace the past history of domestication and diffusion of the date palm through the Middle-East and North Africa. Ms Beisseiche is now in her third year and has successfully managed to extract and amplify endogenous DNA from 1000 years old seeds from Libya. She has contributed to the creation of an archaeological database that will be freely accessible online. Since November 2024, I have been supervising Valentin Grenet, with the official supervision of Dr. R. Guyot (HDR). Mr Grenet will pursue his research on the population dynamics of transposable elements in the date palm, combining Nanopore sequencing with advanced population genomics to obtain estimates of TEs fitness effects and mutation load in this species. I am also involved in **Maxime Criado**'s supervision at Paris-Saclay University (10%), with Dr. Amandine Cornille and Dr. Elodie Marchadier. Mr Criado is currently starting his second year, and is currently developing genomic offset pipelines on fruit trees. I have provided advice on population genomic analyses and bioinformatic procedures.

1.2.5 Contribution to scientific articles involving PhD students

I have provided **significant support to five PhD students** during data analysis and writing of their manuscripts.

List of articles where I significantly contributed to the PhD student's project

(7) **Mira-Jover, A.**, Graciá, E., Fritz, U., Giménez, A., **Bourgeois, Y.** Taking advantage of reference-guided assembly in a slowly-evolving lineage: application to *Testudo graeca*. PloS One. See **Appendix B.**

(6) **Minadakis, N.**, Kaderli, L., Horvath, R., **Bourgeois Y.**, Xu, W., Thieme, M., Woods, D. P. Roulin, A. C. (2024). Polygenic architecture of flowering time and its relationship with local environments in the grass *Brachypodium distachyon.* Genetics.

(5) Horvath, R., **Minadakis, N.**, **Bourgeois Y**. Roulin, A. C. (2024). The evolution of transposable elements in *Brachypodium distachyon* is governed by purifying selection, while neutral and adaptive processes play a minor role. eLife 12. https://doi.org/10.7554/eLife.93284.3

(4) **Minadakis, N.**, Williams, H., Horvath, R., Caković, D., Stritt, C., Thieme, M., **Bourgeois, Y.**, Roulin, A. C. (2023). The demographic history of the wild crop relative *Brachypodium distachyon* is shaped by distinct past and present ecological niches. Peer Community Journal, 3.

(3) **Mould, M. C.**, Huet, M., Senegas, L., Milá, B., Thébaud, C., **Bourgeois, Y.** Chaine, A. S. (2023). Beyond morphs: Inter-individual colour variation despite strong genetic determinism of colour morphs in a wild bird. Journal of Evolutionary Biology, 36(1), 82–94.

(2) **Ameline, C.**, **Bourgeois, Y.**, Vögtli, F., Savola, E., Andras, J., Engelstädter, J., Ebert, D. (2021). A two-locus system with strong epistasis underlies rapid parasite-mediated evolution of host resistance. Molecular Biology and Evolution, 38(4), 1512–1528.

(1) **Bourgeois, Y.**, **Stritt, C.**, Walser, J.-C., Gordon, S. P., Vogel, J. P., Roulin, A. C. (2018). Genome-wide scans of selection highlight the impact of biotic and abiotic constraints in natural populations of the model grass *Brachypodium distachyon*. The Plant Journal, 96(2), 438–451.

I have been part of **Nikolaos Minadakis**'s PhD committee over the course of three years (2021-2024). Dr Minadakis's PhD was supervised by Dr. Anne Roulin (Botanical Institute, University of Zurich), and aimed at studying local adaptation and the evolution of flowering time in the model grass *Brachypodium distachyon*. I contributed to his supervision and

followed his progress. I assisted with demographic inference and genotype-environment association analyses (see also Chapter 3). I was involved in the work of another PhD student from the same laboratory, **Christoph Stritt**. I led an analysis of demography and selection in the same model, which served as a basis for the work undertaken by Dr Minadakis.

I have assisted **Claire Mould** in her analyses of the genetic bases of colour polymorphism in an endemic bird from Réunion (*Zosterops borbonicus*), that I studied during my PhD. I contributed to the development and analysis of diagnostic markers for plumage colouration. I also provided assistance in the interpretation of a few discrepancies between genotypes and phenotypes, explaining them as likely events of recombination between the causal colour locus and diagnostic markers.

I have significantly contributed to **Camille Ameline**'s PhD project under the supervision of Pr. Dieter Ebert at the University of Basel. I contributed to a Genome-Wide Association Study aimed at identifying loci underlying *Daphnia magna*'s resistance to multiple strains of *Pasteuria ramosa*. I contributed to the interpretation of results, revealing how epistatic interaction between two major loci resulted in the observed combinations of resistance phenotypes (see (*7*)).

All the aforementioned students have successfully defended their PhDs.

1.2.6 Evidence for successful PhD supervision as leading investigator

While my previous roles as postdoctoral researcher did not formally allow for independent PhD supervision for six years after my PhD, I have proactively developed my supervisory skills through substantial mentoring and project leadership. I have fully engaged as soon as possible in the supervision of several PhD students in the UK, France and Spain, including my significant involvement in **Andrea Mira Jover**'s PhD project at Miguel Hernández University in Elche (Spain). Ms Mira-Jover, who had no prior experience in population genomics, successfully developed expertise under my guidance. I structured her work program **from project conception to publication**, culminating in a publication where I served as **last and corresponding author**, and I also funded this open-access publication. In addition, I facilitated her access to critical resources, including a High Performance Computing Cluster, and organized workshops and meetings to explain key steps of reference-guided genome assembly. This work has resulted in a high-quality, contiguous reference genome for the vulnerable Greek Tortoise (*Testudo graeca*, NCBI BioProject PRJNA1086345). We are currently working on two new publications on the population genomics of tortoises from North Africa. I have supported her **application to mobility grants** in Spain, which she successfully obtained to visit me in Montpellier from October to December 2024. **These experiences demonstrate my ability to mentor students in developing technical expertise,**

navigating complex research workflows, funding research, and achieving impactful scientific outcomes. Moving forward, I am committed to fostering the academic growth, independence, and career development of PhD students through innovative projects and personalized mentorship.

1.2.7 Post-doctoral supervision

In 2024, I contributed to the recruitment of **Qindong Tang** and **Samuel Gornard** on an ANR post-doctoral project led by my collaborator Ben Warren (*Museum National d'Histoire Naturelle*). Dr Tang is comparing extant and past genetic diversity of endemic bird species from the Mascarenes. Dr Gornard will focus on the detection of recent adaptation and mutation load accumulation, comparing genome-wide allele frequencies between modern and museum samples. I am particularly involved in advising over the use of genomic tools that can handle the low sequencing depths obtained from subfossil and toepad samples (for example the ANGSD suite (*8*)), as well as methods dedicated to the inference of recent demographic events (e.g. HapNe (*9*)), correcting for differential drift in modern and ancient samples (e.g. Factor Analysis in (*10*)), or aimed at detecting positive selection (*11*) using time-series. I recently welcomed Dr Tang, Dr Gornard, and Dr Warren in Montpellier where we discussed over the population genomic pipelines that could be deployed.

1.3 Summary of ongoing collaborations

More information about some of my current and future research projects can be found in Chapter 3. I provide below a short description of my main active collaborations. They can be divided along three main axes: the genomics of transposable elements, biogeography studies, and conservation genomics.

Active collaborations on the population genomics of transposable elements

- Collaboration with **Dr Roulin** (Agroscope, Switzerland) on TEs in *Brachypodium distachyon*. **Planned work:** distribution of TEs fitness effects and signatures of local adaptation.
- Collaboration with **Dr Cornille** (NYUAD, Paris-Saclay, CNRS) on TEs in fruit trees. **Planned work:** support on population genomic analyses (demographic inference, selection, mutation load), and study of polymorphic TEs in fruit trees.
- Collaboration with **Dr Gros-Balthazard** (IRD) on date palm population genomics. **Planned work:** support on population genomic analyses (demographic inference, selection). Dr Gros-Balthazard provides access to samples and genomic data for my own study of polymorphic TEs in the date palm.
- Collaboration with **Pr Boissinot** (NYUAD). **Planned work:** adapation of anole lizards to the urban environment, population dynamics of TEs in house mice (with data from Pr Boursot, Montpellier University).
- Collaboration with **Dr Sabot** (IRD). **Planned work:** developing approaches to visualize and characterize TE polymorphisms in pangenome graphs.
- Collaboration with **Dr Poncet** and **Dr Guyot** (IRD). **Planned work:** adaptation and adaptability to biotic and abiotic changes in the *Coffea* genus, with a focus on TE variation.

Active collaborations in biogeography

- Collaboration with **Pr Perez-Barrales** (University of Granada) and **Pr Adrian Brennan** (Durham University). **Planned work:** Biogeography and genetic bases of flowering time variation in pale flax (*Linum bienne*), the wild relative of domesticated flax (*Linum usitatissimum*). Analyses based on low-depth whole genome resequencing data.
- Collaboration with **Dr Grinsted** (University of Portsmouth). **Planned work:** Biogeography of *Cyrtophora citricola*, a group-living species of spiders displaying low intra-specific aggressiveness. RAD-sequencing analyses led by Nathaniel Holmes (3rd year undergraduate student).

Active collaborations in conservation genomics

- Collaboration with **Dr Steve Augiron** (Société d'Etudes Ornithologiques de La Réunion). **Planned work:** Estimating dispersal, mutation load and recent history of the endangered Reunion harrier (*1*). Work is currently underway to increase sample size to refine landscape genetic analyses. Whole genome sequencing is also planned, to test for a link between inbreeding, mutation load and low fitness.

- Collaboration with **Dr Martinez and Dr Jimenez Franco** (Miguel Hernández University of Elche). **Planned work:** Investigating the landscape genomics and demographic history of Greek tortoises (*Testudo graeca*) in Southern Spain and Morocco. The project has so far generated a reference genome (see article in Annex) as well as RAD-sequencing and whole-genome resequencing data that will be analyzed in the coming year.

  All these projects are realized in close collaboration with local practitioners (Réunion National Park and Region Council, UK Overseas Territories, Doñana National Park), and will translate into conservation actions informed by genomics. For example, our study on the Réunion harrier revealed cryptic population structure associated with ecology, strong inbreeding, recent purging of the most deleterious mutations and accumulation of moderately deleterious ones (*1*). Strong philopatry and exposure of moderately deleterious mutations at the homozygous state call for careful population management to improve connectivity and more detailed studies on the genetic health of these populations.

Chapter 2

Summary of Past Research

My research has mostly revolved around molecular ecology and evolutionary biology. I am especially interested in using advanced genomic tools to understand the past history of species that are representative of biodiversity, as well as quantifying the link between fitness and genotypes. More recently, I have started developing projects on the population dynamics of transposable elements, which are involved in a strong coevolutionary interaction with their host. The study of TEs is fascinating as it contributes to break away from a linear view of the genome (see also Chapter 3).

2.1 PhD project: study of colour variation in *Zosterops borbonicus*

As populations and species diverge from each other, a progressive loss of shared polymorphism and accumulation of fixed alleles is expected. This is impacted by neutral processes (e.g. genetic drift), recombination, sexual and natural selection, that can have antagonistic effects on genetic diversity (*12*). In addition, the accumulation of incompatibilities between genomes from different populations can lead to genetic conflicts and reproductive isolation that is independent of the environment (intrinsic isolation, see (*12*)), even in the presence of gene flow (*13*). The interaction between these processes may vary along the genome, creating a mosaic of regions displaying different rates of divergence (*14*). My research has aimed at solving which mechanisms impact genetic variation in a range of species. One example highlighting this interest is found in my PhD project. I addressed a fundamental question that has puzzled evolutionary biologists for more than a century: why do organisms differ in color and how are color polymorphisms maintained in nature? I focused on a bird from the Mascarene archipelago, the Grey White-Eye (*Zosterops borbonicus*) (*15–17*).

Fig. 2.1 A Grey White-Eye from Réunion (grey morph)

This songbird is endemic to the volcanic island of Réunion (2,500km2) and displays a striking pattern of variation across parapatric plumage colour forms that vary in the extent of brown feathers on the back (Figure 2.2). Such diversity is not expected given the relatively high dispersal abilities of a bird compared to the modest size of the island (ca 50km radius). However, there is evidence that the Grey White Eye is extremely philopatric and a poor disperser(*18*, *19*), which may have an impact on the structure of its populations. Three forms occupy discrete geographical regions in the lowlands, with a completely brown form (lowland brown-headed brown form; hereafter LBHB), a grey-headed brown form (hereafter GHB) and a grey-headed brown form with a brown nape (brown-naped brown form; hereafter BNB). A fourth form is found at high elevation (up to 2,500m), where two colour morphs (highland brown-headed brown and grey; hereafter HBHB and GRY, respectively) coexist. In addition, patterns of coloration among forms and morphs are stable over time, with no apparent sex effect (see (*18*, *20*)). Brown feathers show a deposit of orange pheomelanin where grey feathers do not (*15*). I initially adopted a candidate-gene approach to identify the genetic bases of this color polymorphism, first targeting *MC1R* (*21*), a gene encoding a receptor often found involved in vertebrates pigmentation (*22–24*). The lack of association at this candidate gene led to another study on a set of seven other candidate genes (*25*), which did not reveal any association either.

A well known cognitive bias consists in looking for answers where one already expects to find something (the streetlight effect). This is not entirely unreasonable in the case of colour variation since only a handful of genes (such as *MC1R* or *Agouti/ASIP*) may display minimal pleiotropic effects (*26*). However, it contributes to storytelling and inflates the importance of evolutionary hotspots as major players of phenotypic variation. In the case of the Grey White Eye, a more agnostic approach was needed.

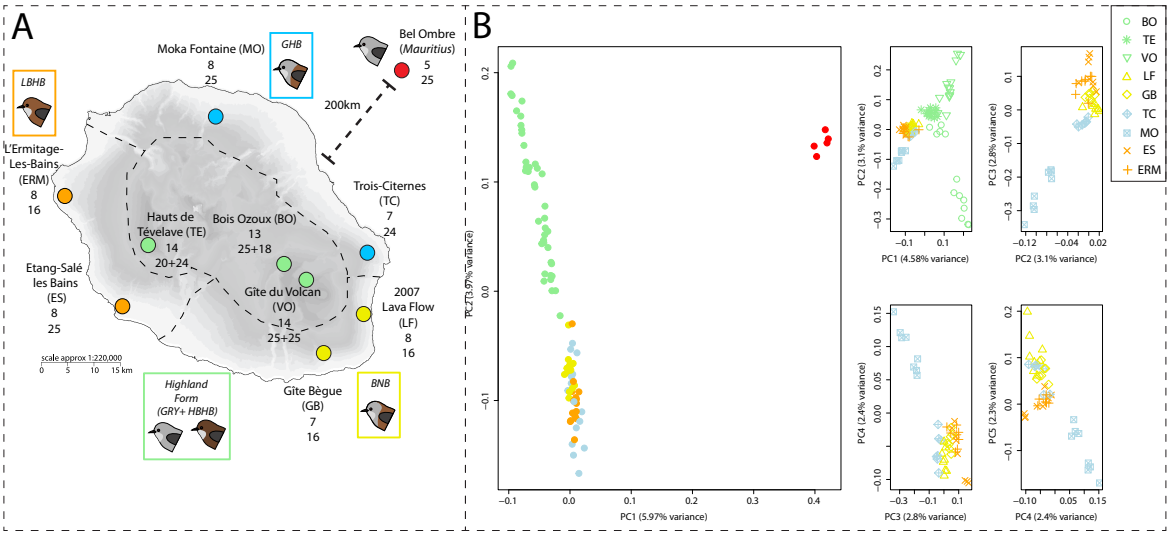


Fig. 2.2 Illustration of the geographic and genetic structure of colour morphs on Réunion island. A: The map shows localities sampled and a description of population structure using principal components analysis (PCA) on autosomal GBS data used to produce PCAs in panel B. For each locality, the sample size for the individual GBS data set is followed by the sample size for each pool. Populations are polymorphic at higher elevation. Distribution limits between the different geographical forms are indicated by dashed lines and elevation indicated by shades of grey. B. PCA results including the sister species Zosterops mauritianus (left panel) or not (right panels). Points corresponding to high-elevation individuals were removed on the last three panels for the sake of clarity

These early negative findings prompted me to develop a RAD-sequencing approach (*27*) at the Genotoul platform in Toulouse (*28*). Due to limited funding, I also used pooled-sequencing (Pool-Seq (*29*)), using individuals from several localities covering the different color forms (Figure 2.2). Individual data were obtained later, using a Genotyping-by-Sequencing protocol (GBS (*30*), Figure 2.4) that generated a much lower density of markers but proved convenient to confirm the validity of the Pool-Seq approach and for demographic reconstruction.

Using this protocol, I first investigated the genomics of color polymorphism in the form where HBHB and GRY individuals were found in sympatry, capitalizing on a collaboration that I initiated with Dr Mathieu Gautier at the CBGP in Montpellier. At the time, Pool-Seq was starting to be considered a viable way to obtain relatively cheap information on genome-wide allele frequencies (*31–33*), but there were still few methods able to properly use allele frequencies estimated from pooled data. Through collaboration with Dr. Gautier, I had the opportunity to utilize a beta version of the BAYPASS software (*34*), which has since become a staple tool in the molecular ecologist's toolkit.

I showed, through a comparative analysis of Pool-Seq, GBS and whole genome resequencing data, that strong signals of recent positive selection and association all pointed to the same genomic region on chromosome 1, in which no previously known color gene could be identified (*15*). I also showed the lack of genetic structure between color morphs apart for this single locus (*15*)), suggesting that some form of balancing selection may maintain color morphs across all populations at high elevation. The exact nature of the selective pressure remains to be identified.

Fig. 2.3 Slopes of Réunion island. The island features steep environmental gradients, primarily due to the towering volcanic peaks (2,500 to 3,000 meters high) located at its core.

Ultimately, I started investigating the genomic landscape of differentiation across all color forms, including parapatric ones found at lower elevation. Using a combination of population genomic tools, I showed that most of the differentiation between colour forms below 1,500 m was found on the Z sex chromosome (*35*). Sexual selection and possibly incompatibilities are associated with the maintenance of colour forms in parapatry despite extensive gene flow at multiple autosomal loci (Figure 2.4). On the other hand, divergence associated with elevation was mostly found at autosomal loci. I ran genome-wide scans of association to pinpoint loci involved in adaptation to elevation and reproductive isolation between the different color forms, highlighting an excess of loci involved in reproduction, immunity and stress-response. I identified *TYRP1*, a sex-linked gene, as a candidate for color patterns in lowland forms. Since this gene displays few pleiotropic effects, and given that the

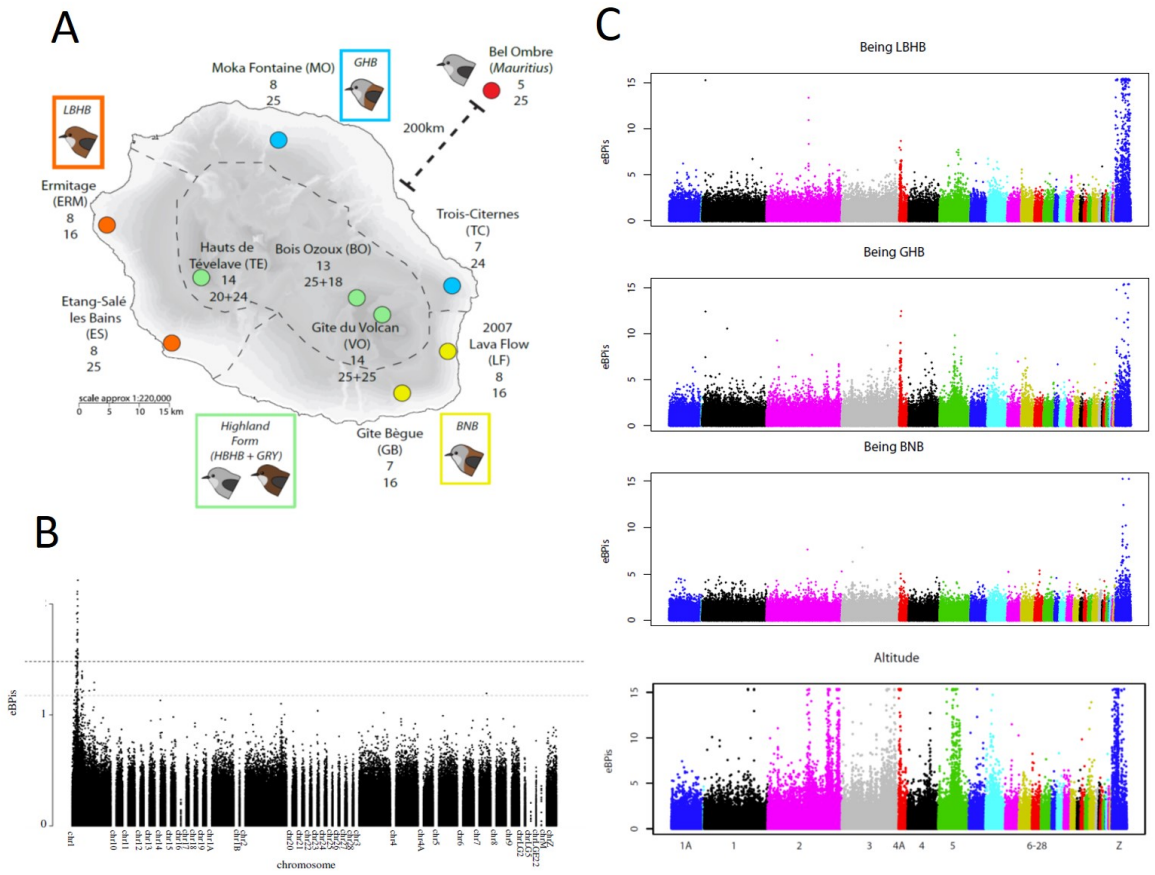


Fig. 2.4 Summary of genome scans of association with colour variation obtained using BAYPASS (34). A: Sampling scheme, as in Fig. 2.2. B: plot of genome-wide empirical Bayesian p-values obtained when comparing GRY and HBHB morphs. A clear peak on chromosome 1 is observed. C: plot of genome-wide empirical Bayesian p-values obtained when comparing each colour form to all the others, as well as a test of association with elevation

White-Eye is able to discriminate between color forms (*36*), it suggests that color itself may be a signal for assortative mating.

In conclusion, while isolation by ecology seems to play an important role in separating populations from low and high elevation, sexual selection and incompatibilities may underlie the maintenance of divergent forms even at a small geographical scale. This work highlights the importance of moving from model species and candidate gene approaches to provide a comprehensive picture of the genetic bases of phenotypic diversity of natural populations.

The methods that I used during this PhD project greatly helped identifying a candidate colour locus in an Ethiopian amphibian of the Ptychadena genus (*37*). This study showed, for the first time in an amphibian, that *ASIP* (*38*), a well-known genetic hotspot of the melanocortin pathway, was also involved in colour variation (Figure 2.5). These two projects

Fig. 2.5 Vertebral stripe colour polymorphism in Ptychadena robeensis. The three possible vertebral stripe morphs are shown. From left to right: wide striped, thin striped, unstriped. Two alleles at an ASIP homolog underlie this variation. Reproduced with permission from (37)

contributed to tackle two issues in the study of phenotypic variation: i) the lack of agnostic genomic scans of association leads to a possible overemphasis over the recurrent role of genetic hotspots and ii) the lack of taxonomic diversity in the models used may also reduce our ability to adequately quantify the rules of genetic convergence across the tree of life. Both projects also benefited from quantitative characterization of the phenotype (i.e. dosage of melanin contents for *Z. borbonicus*, transcriptomic and histological studies in *P. robeensis*), providing a more detailed, mechanistic picture.

2.2 Post-doctoral project: Population genomics of host-parasite interaction in *Daphnia magna - Pasteuria ramosa*

Host-parasite coevolution provides the opportunity to assess how selection that changes through space and time shapes polymorphism in resistance traits, a key aspect in epidemiology and agronomy (*39*, *40*). Hosts and parasites engage in highly dynamic, conflictual interactions, which have a strong impact on genetic diversity at hosts' resistance genes (*41*, *42*). Well-known examples include genes of the Major Histocompatibility Complex (MHC) in Vertebrates and R genes in plants (*43–45*). Factors that may have an impact on this diversity are directional selection, heterozygote advantage, negative-frequency-dependent selection or fluctuating selection (*46*, *47*). Evidence for a role of introgression in maintaining diversity is also growing (*48–50*). However, little is known yet about the relative importance of these factors and how endogenous constraints on the immune pathway affect the likelihood for a gene to be repeatedly the target of selection (*39*, *51*). Therefore, an important research topic lies in understanding which combination of selective and demographic factors lead to the maintenance of stable polymorphism (trench warfare) or repeated fixation of alleles

(arms race), and how the position of a gene in the immune pathway determines its likelihood to engage in a strong co-evolutionary dynamics (*52*).

During my first post-doctoral project in Basel in Dieter Ebert's team (2013-2016), I focused on the host-parasite coevolution between a bacterium (*Pasteuria ramosa*) and a freshwater crustacean (*Daphnia magna*) that display all theoretical conditions for negative-frequency dependent selection (a form of balancing selection) to occur. Resistance follows a matching allele model (MAM) (*56*) and displays a high genetic diversity within populations (*57*, *58*). Since no *D. magna* genotype nor *P. ramosa* strains are universally resistant or infectious, this system meets all the theoretical conditions for negative frequency-dependent selection to occur, a process where rare genotypes or strains have a fitness advantage.

I contributed to characterize the gene complex behind resistance to the pathogen by assisting in bioinformatic analyses and the identification of orthologous regions (*53*). Based on this knowledge, it was possible to directly test for balancing selection at this genomic region and address how host genetic variability is affected by parasite selection. I expanded the study of host-parasite interaction to natural populations in *Daphnia magna*.

I demonstrated that the resistance locus identified in laboratory conditions was involved in natural populations as well. This locus displayed signals of divergent selection between European populations, suggesting that spatial turnover of parasites was at least partly responsible for the maintenance of genetic polymorphism at this locus (*59*).

I have shown that resistance to the bacterial parasite is maintained at a broad geographical scale in *Daphnia magna*. Using whole-genome resequencing and extensive phenotypic data 125 individuals across the species range, I demonstrated that the interaction between *D. magna* and *P. ramosa* constituted a new and powerful model to quantify how host-parasite interactions shape the host's diversity and genome architecture (Figure 2.6). I could show that the resistance locus displayed clear signals of long-term balancing selection and local adaptation (*55*). Coalescence times are older at the vicinity of the resistance locus (Figure 2.6), with more allele sharing across metapopulations, a pattern consistent with the lack of strong geographical structure of resistotypes.

These global patterns could be explained by disruptive selection fixing distinct alleles in separate ponds, depending on the local pool of parasites. To determine whether diversity was also maintained within a single pond, I performed genome scans of selection using a known polymorphic population of *D. magna* in Switzerland, that has been followed by Pr Ebert's team for over 15 years. I could show that selection induced by parasites led to the active maintenance of polymorphism near the resistance locus within a panmictic population, with evidence for selection on ancient haplotypes already present in the metapopulation and maintained through negative frequency-dependent selection. Recombination between these

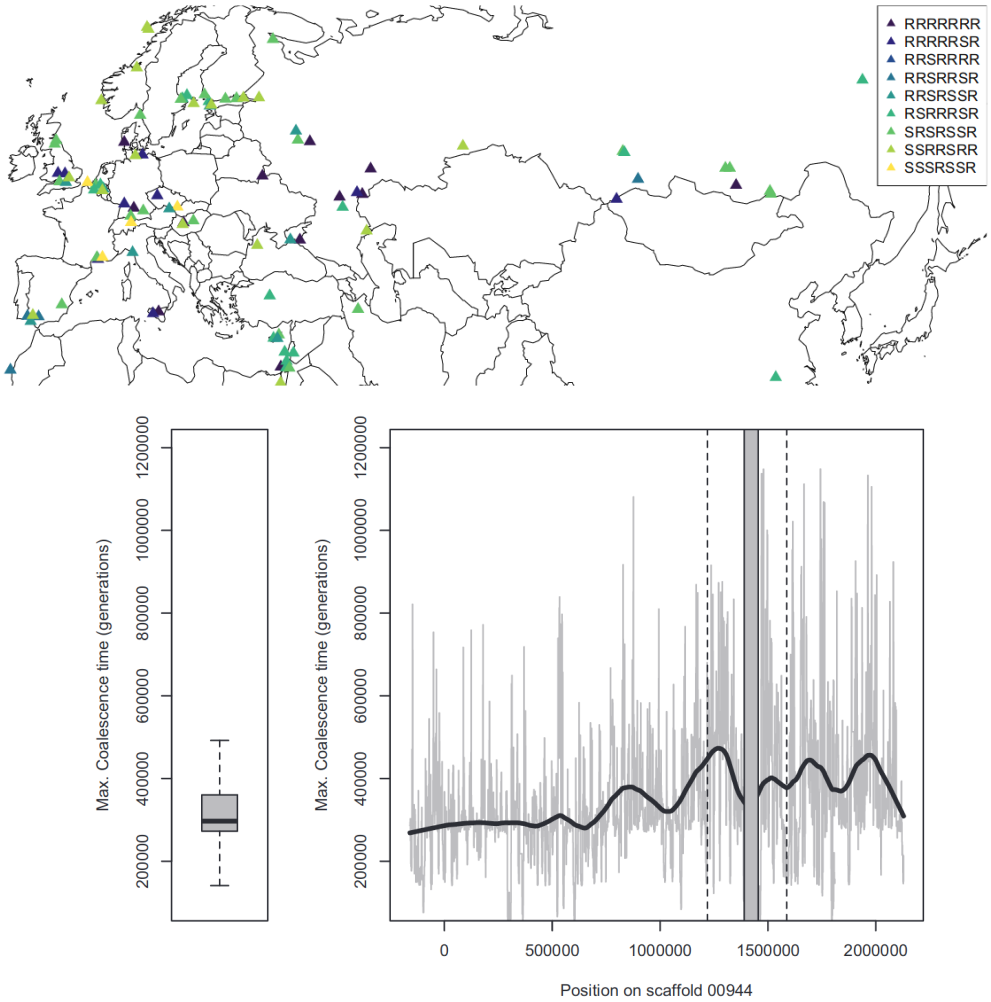


Fig. 2.6 Top: Map showing sampling of 125 D. magna individuals, along with seven resistance phenotype to five distinct strains of P. ramosa. No significant signal of spatial clustering could be detected. Bottom: The scaffold from D. magna's genome (v 2.4) shown here was initially identified through a QTL analysis (53). Time since the Most Recent Common Ancestor for 1-kb windows at the vicinity of the resistance QTL (indicated by flanking vertical dotted lines), estimated with ARGWeaver(54). Coalescence times are given in equivalent generations (sexual + asexual). Approximate times in years can be obtained by dividing by ten, assuming ten generations a year. Boxplots summarize the distribution of statistics from two comparably large scaffolds totaling more than 6 Mb. Taken from (55)

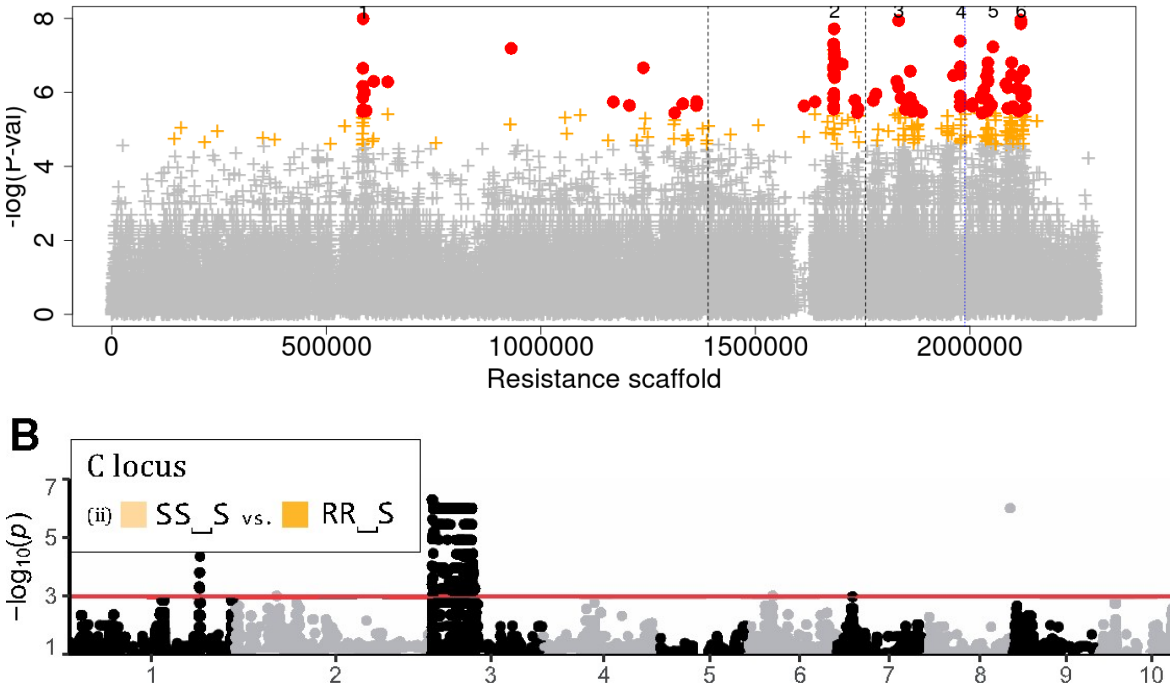


Fig. 2.7 The same locus displays signals of association at the species and population level. Top: P-value of association with resistance to four P. ramosa strains in 125 D. magna clones sampled across all Eurasia and the Middle-East. Phenotypes were coded as a succession of seven letters (S for susceptible to a given strain, R for resistant), and genome-wide scans of association with these discrete phenotypes was implemented using the multinomial logistic regression tool in Trinculo (60). The six main peaks of association are numbered, and the original QTL for resistance to P. ramosa (53) is indicated by vertical dotted lines. Unpublished result. Bottom: Association study for individuals collected from a single pond in Switzerland. Note the strong signal of association on Linkage Group 3, which nearly fully overlaps with the resistance scaffold shown on the top panel. The signal covers nearly half of LG3 and drops abruptly, which is highly suggestive of a large structural rearrangement. Figure taken from (61).

haplotypes was limited in this population due to a (likely) large-scale inversion covering 2 Mb (*61*). This suggests that the patterns observed at the global scale may result from a combination of local adaptation and negative-frequency dependent selection acting within ponds, as suggested by previous studies on sediment cores (*62*). This work was completed in collaboration with Camille Ameline over the course of her PhD project. I shared scripts and supervised over the generation of genome-wide association scans that revealed a second genomic region associated with resistance to parasites (*61*). There are clear epistatic interactions between the two loci, which calls for more work on their evolutionary dynamics. The discovery of what seems to be large structural variants under balancing selection during this post-doc project has made me aware of the importance of such variants in the evolutionary process. Future research should make full use of long-read sequencing technologies to properly assess the diversity of highly divergent super-alleles found at the core of the resistance loci (*53*), and better correct for alignment issues due to repetitive elements. Other possible research avenues could include an assessment of mutation load at the vicinity of resistance loci to test for the presence of 'sheltered load' maintained by overdominance (*51*).

2.3 Post-doctoral project: Molecular ecology and population genomics of transposable elements in *Anolis carolinensis*

The green anole (*Anolis carolinensis*) is a model to understand physiology, sexual behavior and adaptation in squamates. It is found in a diversity of environments, from the subtropical tip of Florida to the winter exposed flank of the Appalachians in Tennessee. Despite the availability of a reference genome, there has been so far no work using whole genomes to investigate at a microevolutionary scale the genomic diversity of squamates. This has limited our ability to link macro- and microevolutionary perspectives in the evolution of vertebrates.

During this project, I used whole-genome resequencing data to answer questions about intrinsic and extrinsic forces shaping the diversity of a squamate. The genetic structure of green anoles consists in five main genetic clusters, three of them having diversified in Florida in the last six million years, and two others having emerged through northward colonization of temperate environments in southern United-States in the last 300,000 years (*63–65*).

This recent expansion is interesting since anoles are otherwise exclusively tropical, and may have been associated with new adaptations (Figure 2.8). However, demography and intrinsic properties of genomes such as variation in recombination rates have often been overlooked in genomic scans for selection (*66*, *67*). This is problematic given their known

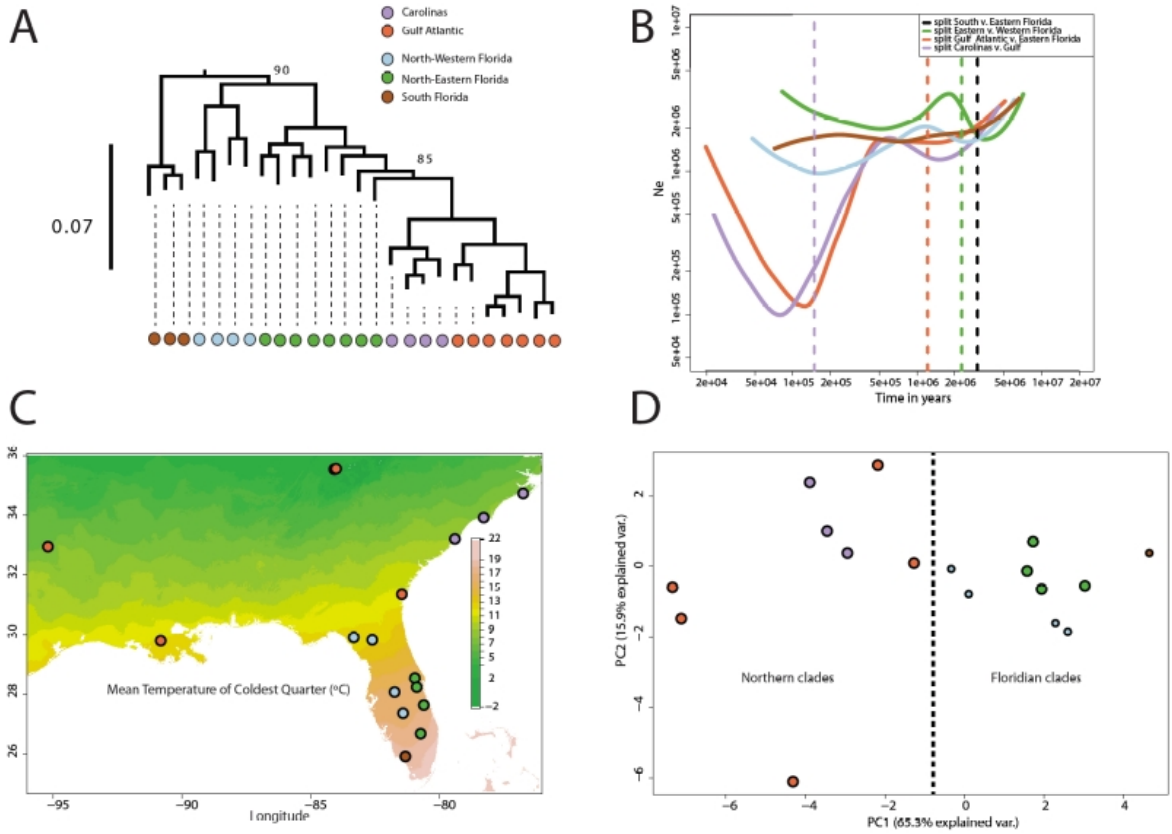


Fig. 2.8 A: Phylogeny of 27 green anoles included in a whole-genome resequencing study. B: Changes in effective population size (Ne) with time. Note the recent bottleneck around 200,000 years ago in northern populations (orange and purple lines). C: Map of average temperatures during winter in North America and locations of samples from the five distinct genetic clades. D: PCA over environmental variables for sampling sites. Larger dots highlight the northern clades and their sister Floridian clade.

impact on the rate of false positives in the set of candidates for selection. The first part of this work (*68*) therefore consisted in better understanding the non-adaptive forces shaping the genomic landscape of differentiation in green anoles. Using recent population genomics methods and model comparison, I showed that populations from Florida have remained stable in size and likely underwent repeated episodes of secondary contact due to past variation in sea level that turned Florida into an archipelago (*69*). I estimated effective recombination rates ($\rho = 4 \times N \times r$ with N the effective population size and r the recombination rate) along genomes, and demonstrated the role of low recombination and purifying selection in generating locally high levels of differentiation along the genome, but did not obtain strong evidence of loci resisting gene flow after secondary contact. Unlike previous claims that squamates genomes may lack biased gene conversion (BGC) (*70*, *71*), I found clear correlations between recombination rates, diversity and GC content in coding sequences, consistent with a bias from A/T to G/C during DNA repair through recombination (*72*). I also showed that the levels of diversity on the sex chromosome (X) were significantly lower than expected, even when taking into account their lower effective population size compared to autosomes, suggesting either strong selection reducing diversity at this chromosome, or male-biased dispersal reducing the effective number of X copies in populations that colonized temperate environments. Using this background, I then applied a set of recent methods to quantify selection in temperate populations, and showed that the colonization of this new environment may have been linked to behavioral shifts as well as physiological adaptation in the green anole (*73*). It is important to remind here that genome scans of selection are still heavily debated in the context of the long-standing (*74*) debate between adaptationists and near-neutralists (for recent discussions, see (*75–79*)). Their interpretation and asserted robustness hinge upon assumptions made by the researcher beforehand, such as the anticipated proportion of the genome subject to (strong) positive selection. It's important to emphasize that the ensuing results should not be viewed as conclusive evidence of adaptation but rather as pointers towards candidates warranting further investigation in subsequent research.

Briefly, selection impacts genomic features such as: i) the length of haplotypes around the selected locus, ii) the coalescence times and nucleotide diversity, iii) the allele frequency spectrum, iv) differentiation between localities in different environments and v) association with a phenotype known to be under selection. I applied some of the most recent methodological advances in GWSS (*54*, *80*) to scan the genomes for genes under divergent selection between temperate and tropical populations. Among these approaches, machine learning (diploS/HIC (*80*)) showed quite powerful. This method uses pseudo-observed datasets to generate expectations for several summary statistics under scenarios of neutrality, selection

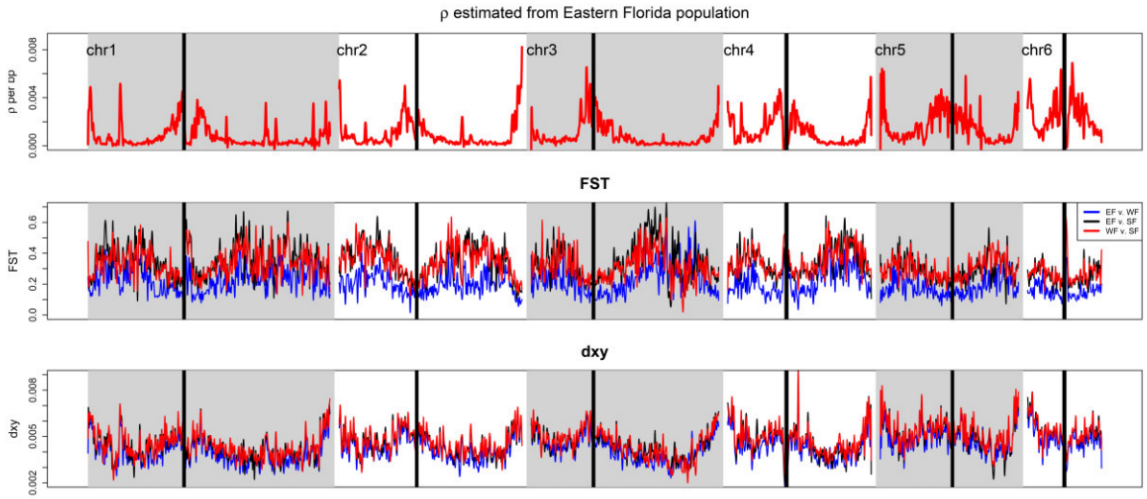


Fig. 2.9 Summary statistics for recombination and differentiation along chromosomes.($\rho = 4 \times N \times r$, with r the recombination rate per bp and per generation and N the effective population size. F_{ST} and d_{XY} are relative and absolute measures of differentiation that are correlated with the amount of shared heterozygosity and coalescence time across populations, respectively. The three lines show differentiation for the three genetic clusters having diverged for the longest time period. Statistics were averaged over nonoverlapping 5-kb windows and a smoothing line was fit to facilitate visual comparison. Repetitive centromeric regions that are masked from the green anole genome are highlighted by black rectangles.

on de novo mutations (hard sweep) or on standing variation (soft sweep). It has the desirable property of directly including the confounding effects of demography in its implementation, and gives an estimate of false positive and false negative rates (but see (*75*)). I naively scanned the genome for genes of interest and assessed the overlap between multiple tests. Strikingly, I identified many genes involved in limb development for which multiple tests clearly indicated positive selection. This is an interesting finding, given the importance of limb length variation in anole ecomorphs (*81*, *82*). Using a candidate genes strategy, I also determined that genes involved in response to cold or behavior displayed more frequently signals of selection, while controlling for local recombination rate, gene clustering and gene length. Genome scans are not anymore limited to recent positive selection but can also track signatures of balancing selection (see also previous section) Despite being often overlooked (*83*), this type of selection is commonly found at immune and resistance genes (*39*, *84*) and may therefore be relevant in a temperate context where biotic interactions can diverge drastically from tropical settings. I found signatures of balancing selection at immune genes in all investigated genetic groups, but also at genes involved in neuronal and anatomical development in Florida. This suggests that while pressures exerted by pathogens may maintain diversity across all populations despite bottlenecks, shifts in behavior and development

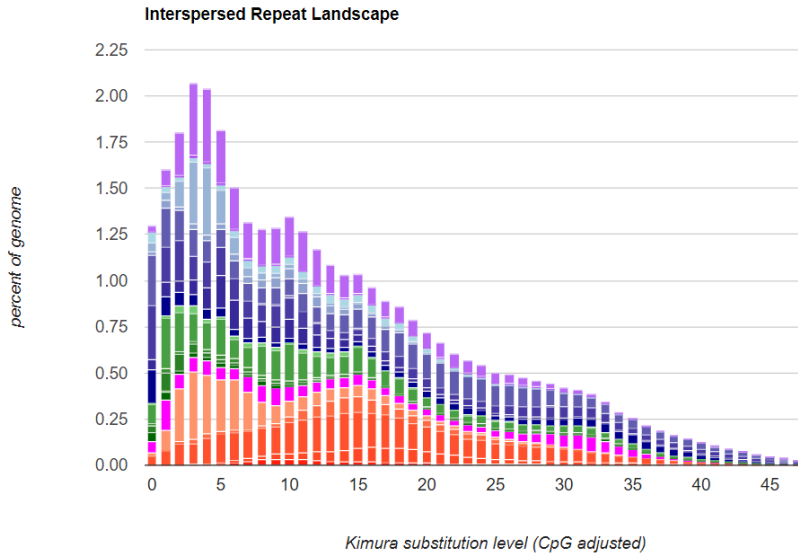


Fig. 2.10 Repeat landscape of the green anole. The plot shows TE divergence from their consensus sequence, and can be taken as a proxy for the time since they inserted. Recent and ongoing activity should correspond to the bar at 0% divergence. Red shades correspond to DNA transposons, green shades to LTR retrotransposons, blue shades to non-LTR retrotransposons (LINEs and SINEs).

during northward expansion may have been associated to changes in the selective mode at associated genes.

Studies on adaptation and adaptability have tremendously benefited from the study of Single Nucleotide Polymorphisms (SNPs). However, they have neglected transposable elements (*85*) that can have substantial effects on genome size and organization (*86*). TEs are DNA fragments that move and duplicate through genomes, sometimes inserting within or near genes. They can therefore have both deleterious or advantageous fitness effects. The green anole's genome contains an extraordinary diversity of TEs (Figure 2.10). I characterized TEs frequencies across several TEs families and used Single Nucleotide Polymorphisms (SNPs) as a near-neutral contrast (*87*). I showed that TE insertions displayed an allele frequency spectrum skewed towards singletons, which could not be explained by demographic stochasticity alone, but was instead consistent with purifying selection. I took advantage of the population genomics investigation described above to quantify how selection, demography and recombination have shaped the genomic distribution of elements (*88*). I showed that short TE insertions were nearly-neutral, while longer ones were more deleterious due to ectopic recombination and gene disruption. I did not find any evidence of TEs domestication through recent adaptation. I demonstrated through simulations that TEs frequency and abundance showed distinct correlations with recombination rate (estimated

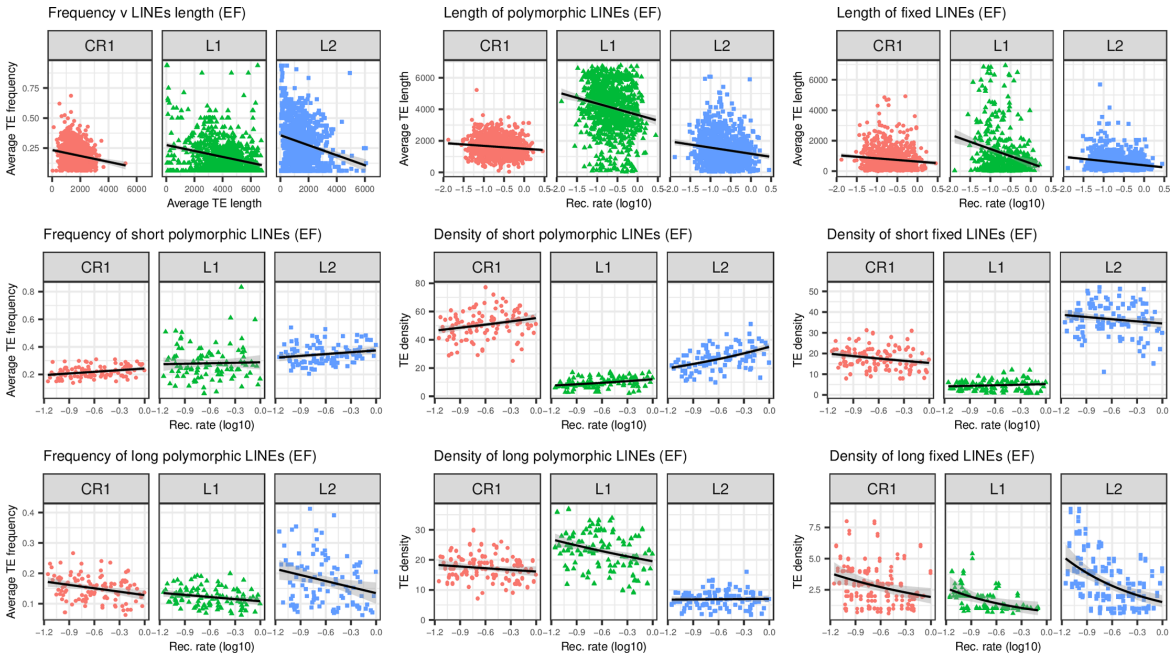


Fig. 2.11 Genome-wide diversity of non-LTR retrotransposons (LINEs). Top: Plots of TE length against recombination rate. Middle: Plots of average frequency, density of polymorphic insertions and density of fixed insertions for short LINEs, Bottom: same as middle row, for long LINEs. For middle and bottom plots, average frequencies and densities are computed for 10Mb windows. In (88)

from SNPs). This is because recombination rate is correlated with the probability of fixation of mutations and TEs insertions.

These distinct patterns were useful to determine how selection, preferential insertion of TEs and rates of transposition interacted to shape TEs abundance and frequency along the genome. For example, direct selection against deleterious TEs may be stronger in regions of high recombination, due to increased probability of ectopic recombination. This results in TEs that are less abundant and at lower frequency in regions of high recombination. On the other hand, the effects of Hill-Robertson interference and linked selection may lead to an accumulation of nearly-neutral TEs that are either fixed or at very low frequency in regions of low recombination. This is the same mechanism that leads to the widely observed negative correlation between F_{ST} and d_{XY} in many species (*89*): background selection eliminates polymorphism but increases the number of substitutions.

An illustration of how these features may be used can be derived from Figure 2.11. Longer TEs are expected to be more deleterious in regions of high recombination, possibly because they are more likely to engage in ectopic recombination. This pattern is clearly observed for the three main clades of LINEs investigated: TEs are shorter in regions of higher recombination, while the abundance and average frequency of long elements decreases with

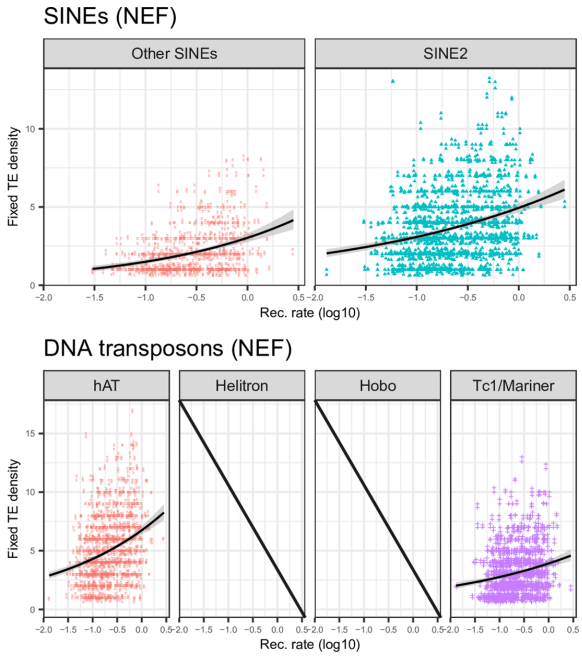


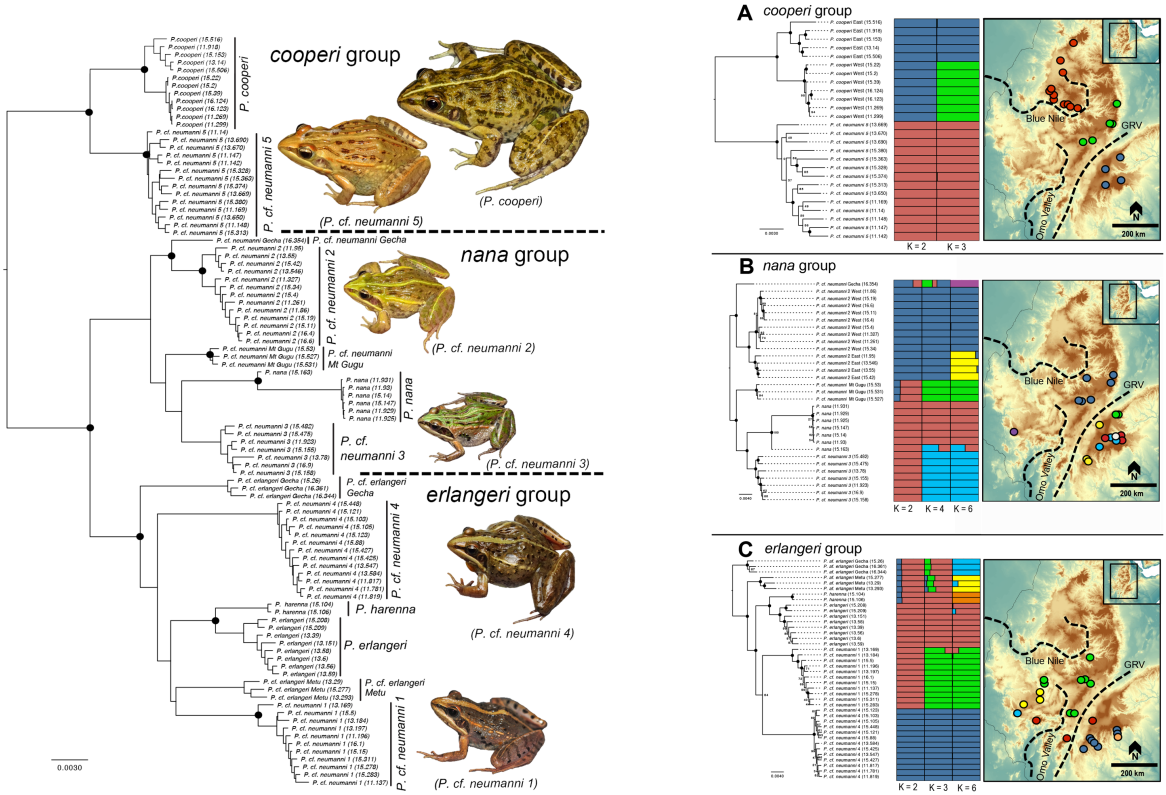
Fig. 2.12 Density of fixed SINEs and DNA transposon insertions against recombination rate. Helitron and Hobo insertions are too few to be shown. Figure taken from (88)

recombination. Shorter elements display an opposite pattern however, which can mostly be explained by the effects of background selection on near-neutral insertions. However, background selection alone should result in a decrease in the density of fixed insertions as recombination increases. This is the case for CR1 and L2, but a (non-significant) positive correlation was observed for L1. A very clear positive correlation is also observed for non-autonomous SINEs and DNA transposons (Figure 2.15). In my simulations, such a pattern could only be obtained if elements insert preferentially in regions of higher recombination. These results are promising, and show that considering TEs from a population genomics perspective may help in deciphering the mechanisms behind their diversity (e.g. transposition rates, selection, and intrinsic factors such as preference for specific motifs or open chromatin).

2.4 Other projects

2.4.1 Landscape genetics of Ethiopian birds and amphibians

The Ethiopian Highlands are a major biodiversity hotspot, with steep mountain ranges separated by two main barriers, the Rift Valley and the Blue Nile Valley. Barriers and steep environmental clines have a strong influence on speciation and local adaptation. During my post-doctoral position at New York University Abu Dhabi, I contributed to the study

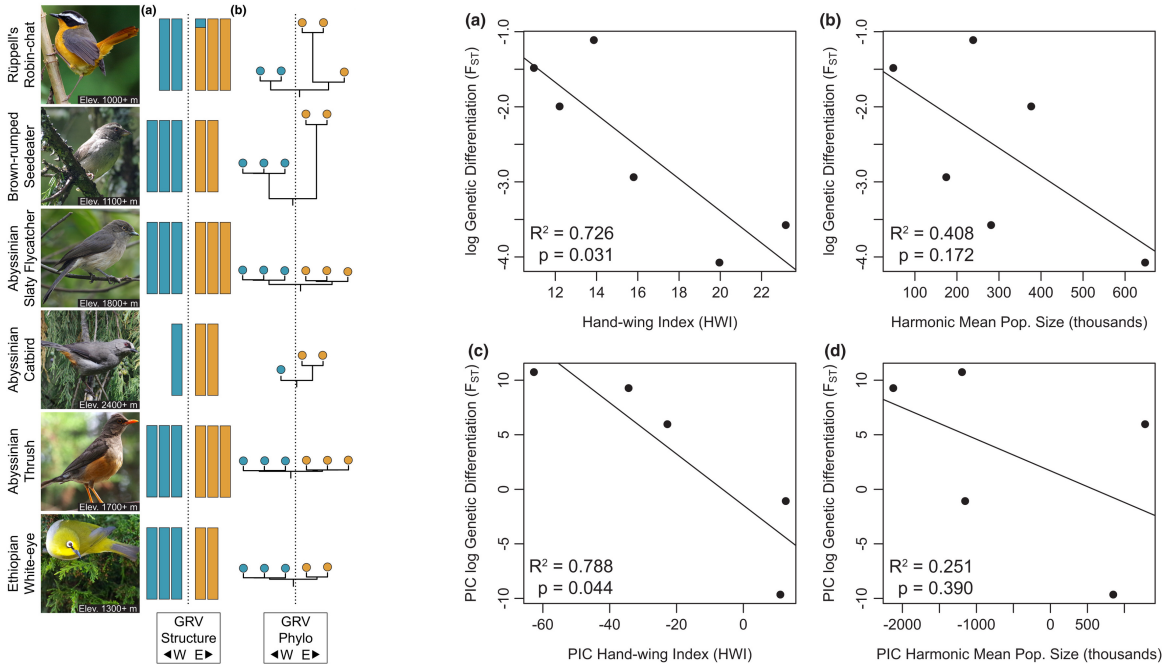


(a) *(b)*

Fig. 2.13 Ptychadena phylogeography. Three main groups can be identified in the phylogeny, the cooperi group, the nana group, and the erlangeri group. The nana group occurs at higher elevation than the erlangeri group, which favours forested habitats. The impact of the Rift Valley (GRV) and the Blue Nile on differentiation appears clearly in clustering analyses and phylogenies. Figure reproduced from (90)

(a) The Blue Nile *(b) Ethiopian Mountains*

Fig. 2.14 Biogeographical barriers in Ethiopia



(a) Structure in Ethiopian birds *(b) Correlations between dispersal and diversity/divergence*

Fig. 2.15 Based on whole genome resequencing data, there is a clear genetic dichotomy between endemic Ethiopian birds from the East and the West side of the Rift Valley. Birds with a better dispersal ability (measured as the Hand-wing index) also tend to display higher effective population sizes and lower differentiation across the Rift. PIC: phylogenetic independent contrast. Figure reproduced from (92)

of genetic variation in several species of endemic amphibians and birds. Using whole-genome data and RAD-sequencing, we showed that frogs of the genus Ptychadena can be clustered into three species groups, themselves subdivided into 13 lineages (Figure 2.13). These endemic lineages evolved in allopatry and are restricted to distinct elevation layers, demonstrating the important role of ecology and geography in shaping the exceptional biodiversity of Ethiopia (*91*). This work was also used to demonstrate the importance of combining morphological and genetic data when describing the taxonomic diversity of such clades. Avoiding species oversplitting and wrong assignation of specimen types to species is critical. A multi-disciplinary approach is needed to properly describe and ultimately protect biodiversity.

I also contributed to the first genetic assessment of six Ethiopian Highland forest bird species (*92*), in collaboration with Bale Mountain National Park and Mekelle University (Dr Yonas Meheretu). Using whole-genome sequencing, we showed how the Great Rift Valley (GRV) might act as a biogeographic barrier in species with relatively good dispersal ability. In particular, we tested whether this barrier could be overcome by species with better

dispersal abilities (Figure 2.15). We did observe significant, long-term (ca 350,000 years) genetic differentiation, which was negatively correlated with dispersal ability. Nevertheless, the distance across the GRV remains difficult to cross, even for highly dispersive species. This study highlights how accounting for differences in life history traits is informative when interpreting the state of genetic diversity and differentiation across multiple species. While the study already encompasses a significant fieldwork effort, additional sequencing would be required to conduct a comprehensive landscape genetic study and verify the Rift's role as a barrier.

2.4.2 Participation to COG-UK: phylodynamic analysis of the SARS-CoV-2 epidemic in Hampshire, UK

In 2020, at the start of the SARS-COV-2 pandemic, I contributed to a grant application led by Dr Sam Robson to integrate the COVID19 Genomics UK (COG-UK) Consortium (*93*). This application was successful and made Portsmouth an important node to collect sequencing data from local hospitals in Hampshire. I contributed to the phylodynamic analyses of data, and trained local students to use phylogenetic methods. Tracking epidemics requires heavy logistics, and may miss many cases due to insufficient testing capabilities or fast spread. The examination of viral genomes may help from this perspective. Genealogies of viral sequences contain information about the processes that generate them. Recent advances in the modelling of the coalescence process during epidemics give access to relevant epidemiological parameters such as the effective reproductive number or prevalence. From an epidemiological perspective, Portsmouth is an important city in the UK, being second only to London for its population density. It is therefore vulnerable to a fast spread of coronavirus. It is also a University city, where many students settle at the start of the term (end of September/beginning of October). Such large population movements might have an impact on the spread of the virus. I used the birth-death coalescent framework implemented in the BEAST2 (*94*) BDSKY package (*95*) to reconstruct the dynamic of the epidemic in student accommodations and hospitals in Portsmouth from August 2020 to January 2021. This approach estimated the reproductive number (R) of the virus over discrete time intervals that can be defined *a priori*. The R number can be seen as the number of new infections caused by a single infected individual. A number larger than 1 suggests a spread of the epidemics, while a number lower than 1 suggests a die out of the outbreak (Figure 2.16).

There was a strong increase of the reproductive number during the first week of October 2020 as well as in January 2021, indicating a significant rise in cases during these times. The increase in infection rates in January was attributed to the simultaneous rise of the B.1.1.7

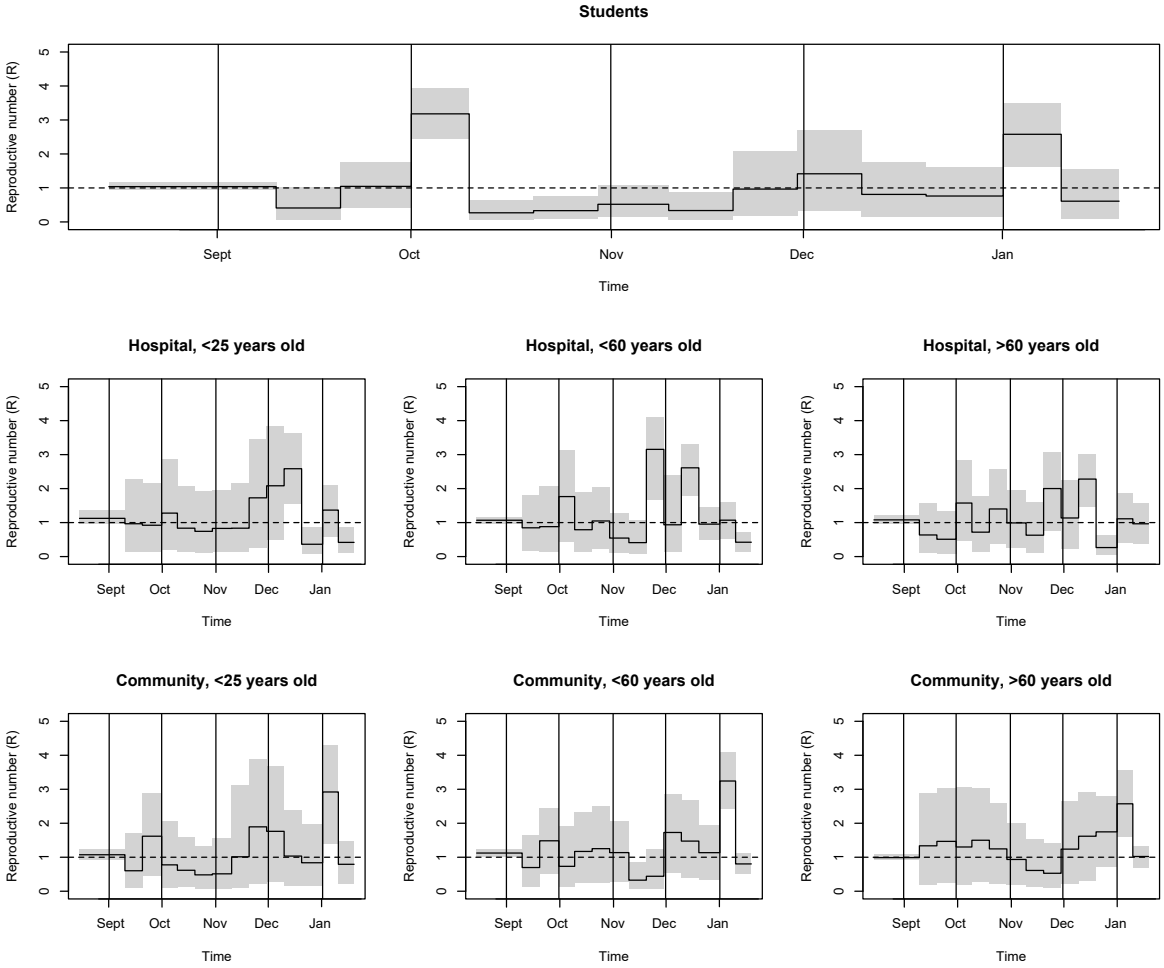


Fig. 2.16 Phylodynamics of SARS-CoV-2 in Portsmouth. The estimates of R (the reproductive number) from August 2020 to February 2021 obtained from resequencing the whole SARS-CoV-2 genome for hundreds of students, hospital patients, and members of the community are shown. I ran separate analyses for distinct age groups for both hospital samples and community samples (<25 years, <60 years and >60 years old).

lineage. The increase at the start of October may have been associated with the start of the new University term. A similar but earlier increase in R was seen in younger members of the community. Increasing infection rates at the beginning of October may therefore have been due to more infections amongst younger people in general rather than increased student influx into Hampshire. The R number dropped well below 1 soon afterwards, likely in response to the implementation of control procedures by the University. Overall, the impact of control policies and major social events was well reflected in this analysis, highlighting the interest of the massive sequencing programme put in place in the UK.

This project has fed my interest in deploying approaches based on coalescence to better characterize the dynamics of transposable elements (TEs) in their host genome. Phylodynamic approaches could be used on TEs to identify nucleotide variants along their sequence that may alter transposition or insertion preference (see for example (*96*)), in the same way phylodynamic analyses have been used to estimate the selective advantage of SARS-CoV-2 variants (*97*). Increasingly affordable long-read technologies that give access to the entire sequence of insertions will likely facilitate this.

Chapter 3

Current and Future Research Projects

3.1 Population genomics of transposable elements (TEs)

"*In the future, attention undoubtedly will be centered on the genome, with greater appreciation of its significance as a highly sensitive organ of the cell that monitors genomic activities and corrects common errors, senses unusual and unexpected events, and responds to them, often by restructuring the genome*" (*98*). **McClintock, 1984**

"*The rejection of one adaptive story usually leads to its replacement by another, rather than to a suspicion that a different kind of explanation might be required. Since the range of adaptive stories is as wide as our minds are fertile, new stories can always be postulated*" (*74*). **Lewontin and Gould, 1979**

"*Natural selection operating within genomes will inevitably result in the appearance of DNAs with no phenotypic expression whose only 'function' is survival within genomes. Prokaryotic transposable elements and eukaryotic middle-repetitive sequences can be seen as such DNAs, and thus no phenotypic or evolutionary function need be assigned to them*" (*99*). **Doolittle and Sapienza, 1980**

The interaction between Transposable Elements (TEs) and their hosts is the most intricate co-evolutionary process found in nature. TEs are DNA fragments that move and duplicate through genomes, sometimes inserting within or near genes. TEs are broadly classified into two classes: class I elements (or retrotransposons), which are mobilized by the reverse-transcription of an RNA intermediate, and class II elements (DNA transposons), which use a DNA intermediate. Retrotransposons are further divided into long terminal repeats (LTR) and non-LTR retrotransposons, based on the presence of long terminal repeats (LTR).

LTR retrotransposons, which include the *Ty1/Copia* and *Ty3/Gypsy* elements, are mobilized by a process similar to retroviruses. The RNA is reverse-transcribed in the cytoplasm into a double-strand cDNA, which is inserted back into the genome by an integrase. Non-LTR retrotransposons, which include the Long Interspersed Nuclear Elements (*LINEs*) and *Penelope* elements, are mobilized by a mechanism termed target-primed reverse transcription, where the RNA is reverse-transcribed at the site of insertion (*100*). The reverse transcriptase of non-LTR retrotransposons can also act on other transcripts and is responsible for the amplification of non-autonomous elements (also called Short INterspersed Elements, or SINEs), which can considerably outnumber their autonomous counterparts (*101*). Class II elements include elements that use a cut-and-paste transposition, such as the *hAT* and *mariner* elements, or elements that have a circular DNA intermediate (*Helitrons*). Class II elements can also mediate the transposition of non-autonomous copies, which, similar to *SINEs*, can amplify to extremely high copy numbers.

TEs constitute a significant portion of most eukaryotic genomes, explaining most differences in genome size across eukaryotes, and accounting for more than 85% of the genome in some species, such as bread wheat (*102*). TEs can disrupt gene function (through loss of function) or disturb pairing of chromatids during meiosis (ectopic recombination), which gives them the potential to be strongly deleterious (*103*). TE transposition can be triggered by stress, leading to bursts of insertions that then become the target of positive or negative selection (*104*, *105*). This sensitivity of TEs to stress and changing environments has fascinated their discoverer, Barbara McClintock, who saw them as agents through which the cell could react to environmental stresses by restructuring its genome (*98*). Since then, many studies on TEs have emphasized their importance for adaptation in a broad range of eukaryotic organisms (e.g. (*106–109*)).

The other traditional view is that TEs are genomic parasites, the "purpose" of which lies in propagation. Unlike viruses, which escape their host cell to infect others, TEs are (mostly) restrained to a single genome. As a consequence, insertions that cause too high a cost on fitness will be removed from the host population through purifying selection (*110*). In addition, TE insertions are directly impacted by the demography of their host population, and can be lost through drift (e.g. (*111*, *112*)).

All these views are true in the sense that there are clear examples of TEs being adaptive, deleterious, or under the influence of neutral processes. However, there is still a lack of quantitative assessment on the relative importance of these processes in shaping TE diversity. With the wealth of data that are about to be generated thanks to third-generation (long reads) sequencing technologies, combined with vast improvements in computing power and

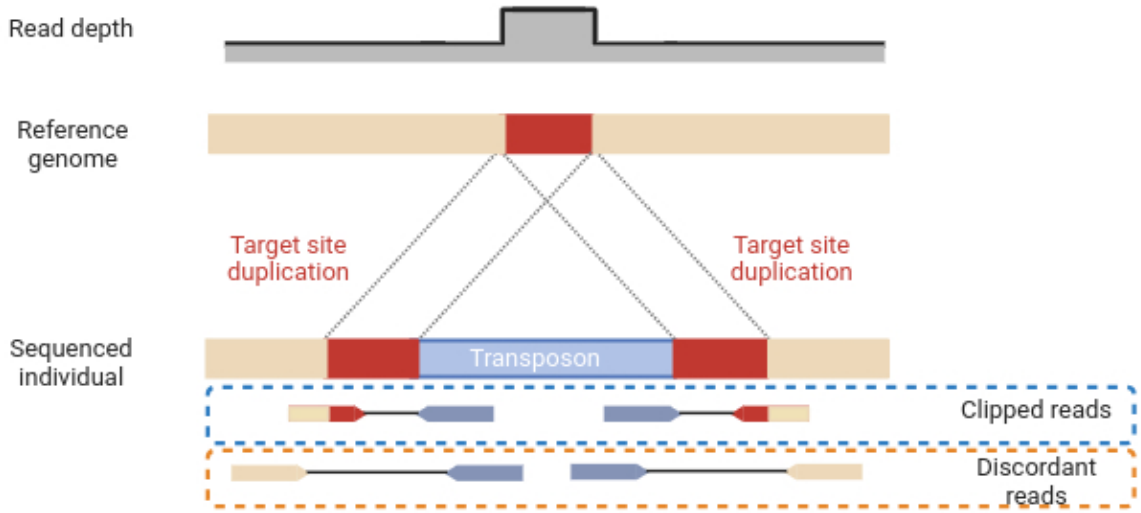


Fig. 3.1 TE detection using paired-end short reads. At their core, methods aiming at detecting TEs are similar to various Indel callers (such as Delly (113)). Reads that only partially align to the reference genome are extracted to identify putative insertion points. Clipped and discordant reads are then compared to a database of TE consensi, to confirm the presence of a TE insertion. The presence of a Target Site Duplication (TSD) is a hallmark of TEs, and is also considered during calling. Genotyping of the insertion accounts for sequencing depth at the TSD (1.5x genome-wide median depth for heterozygous TEs, 2x for homozygous TEs).

analytical pipelines, a path may well be opening for a quantitative assessment of TE impact on their host, by estimating the distribution of their fitness effects (DFE).

3.1.1 Some challenges of studying TE population dynamics

Explaining the processes behind TE spread and extinction is crucial to understand their dynamics and impact on the host. Most studies have focused on the molecular mechanisms underlying TE insertion, replication and control by the host. This view tends to see genomes as static and constant environments, where the effects of transposition are countered by purifying selection. Nevertheless, a more holistic approach connecting TEs and host evolutionary dynamics has been proposed, **building upon the concepts of population genetics**. The first population genetic models examining the specific issue of TE polymorphisms were developed in the 1980s, taking into account both intrinsic and extrinsic factors acting on TE dynamics, such as effective population size, transposition and excision rates, and purifying selection (*114*, *115*). For the last 40 years, while the mechanisms controlling the activity and copy number of TEs have been investigated at macroevolutionary scales, their microevolutionary dynamic has been understudied using the tools and datasets of the genomic era.

Despite their importance, transposable elements have been largely ignored in the search for the molecular bases of fitness variation using the tools and datasets of the genomic era (*116*). This is partly attributed to the absence of suitable methods capable of identifying polymorphic TE insertions within populations using short-read sequencing. Recent progress has been made however, with methods able to recover a significant fraction of polymorphic elements with reasonably low rates of false positives (e.g. MELT (*117*), MEGAnE (*3*)). These methods rely on discordant and clipped paired-end reads (usually > 100bp) to discover and genotype TE insertions in a resequenced individual (Figure 3.1). To classify insertions, most of these methods rely on an existing, curated database of elements to map discordant reads to a clearly annotated TE consensus (but see (*118*)). They also cannot access the internal sequence of the element, unless those are very short (< 500bp).

The development of third generation, long-read sequencing, is extremely well suited to the study of TEs. Both Oxford Nanopore and PacBio platforms sequence >10kb fragments that cover most TEs and their flanking sequence, replacing TEs along the genome with extremely high precision. These methods give access to the internal sequence of insertions. Nanopore basecalling can also include modified bases such as methylated cytosines, and may also be used soon to detect non-B DNA (*119*). This is interesting given the role of methylation in regulating TE activity, particularly in plants. Long reads can also be used to accurately detect and annotate transcribed TEs in RNA-sequencing data. Dedicated methods to genotype TEs using long reads are now starting to emerge. At the time of writing, TREMOLO (*120*) and GraffiTE (*121*) were among the most promising candidates. Both these methods rely on an existing database of TE consensi to identify TE-related indels and generate a graph of TEs identified in a given sample. GraffiTE provides the option to realign short reads on this graph, with performance on par with MELT or MEGAnE (*121*). A recent software, Pantera (*122*), uses insertions of similar lengths to identify polymorphic repeats on the fly, improving the annotation of polymorphic elements while reducing the amount of manual curation –a classical bottleneck in TE analysis for new model species.

Another challenge resides in the somewhat inconstant transposition of TEs over evolutionary times (*123*). This is an issue when comparing summary statistics for TEs and SNPs (Figure 3.2). I will elaborate on this issue in the next section.

3.1.2 How to obtain the distribution of TE fitness effects

As stated above, TEs can account for a large proportion of eukaryotic genomes, and there is a clear positive correlation between genome size and TE content. There are several hypotheses about the determinism of TE accumulation in genomes. One of the most famous (mutational hazard) was proposed by Lynch & Conery in 2003 (*124*), and relies on the near-neutral

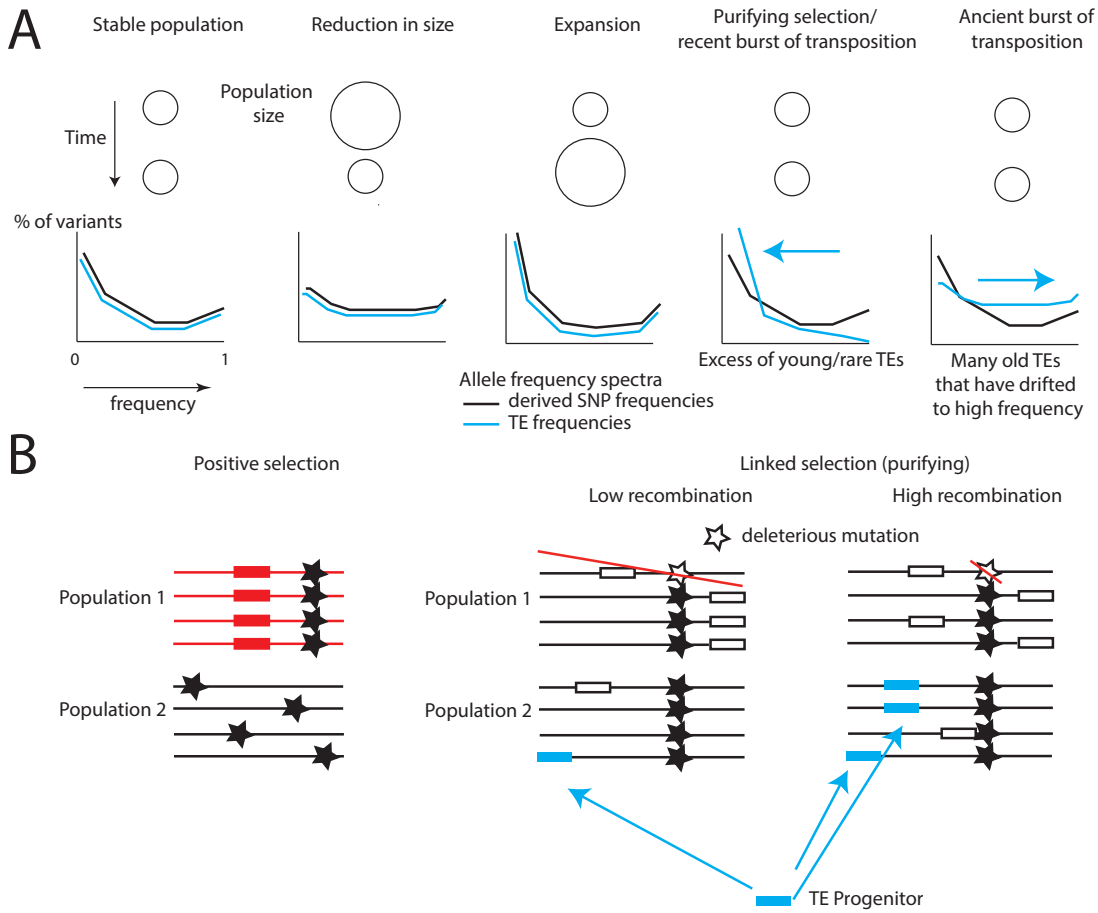


Fig. 3.2 Summary of the population genetics forces that impact TE diversity. A: Demography has an impact on the allele frequency spectrum of TEs. For neutral TEs, the shape of the spectrum is impacted in the same way as SNPs: for example, reduction in population sizes are associated with more alleles at intermediate frequencies. For TE families that are deleterious for the host, insertions are rapidly removed by selection, leading to an excess of rare TEs. B: TEs (rectangles) may also be recruited by positive selection, which should reduce diversity at associated sites in the population where they provide an advantage (red insertion in population 1). However, genome-wide diversity (including near-neutral SNPs and TEs) is strongly shaped by the effects of selection at linked sites (linked selection). If a site (black star) is under purifying selection, deleterious mutations at this site will be removed from the gene pool, along with linked SNPs and TEs. This effect is particularly pronounced in regions of low recombination. In regions of high recombination, adjacent loci are more independent, which prevents the erosion of diversity observed in regions of low recombination, and allows the maintenance of TEs at higher frequency. At last, preferential insertion of TEs shape their density along the genome (here, the blue progenitor generates copies that insert more in regions of high recombination).

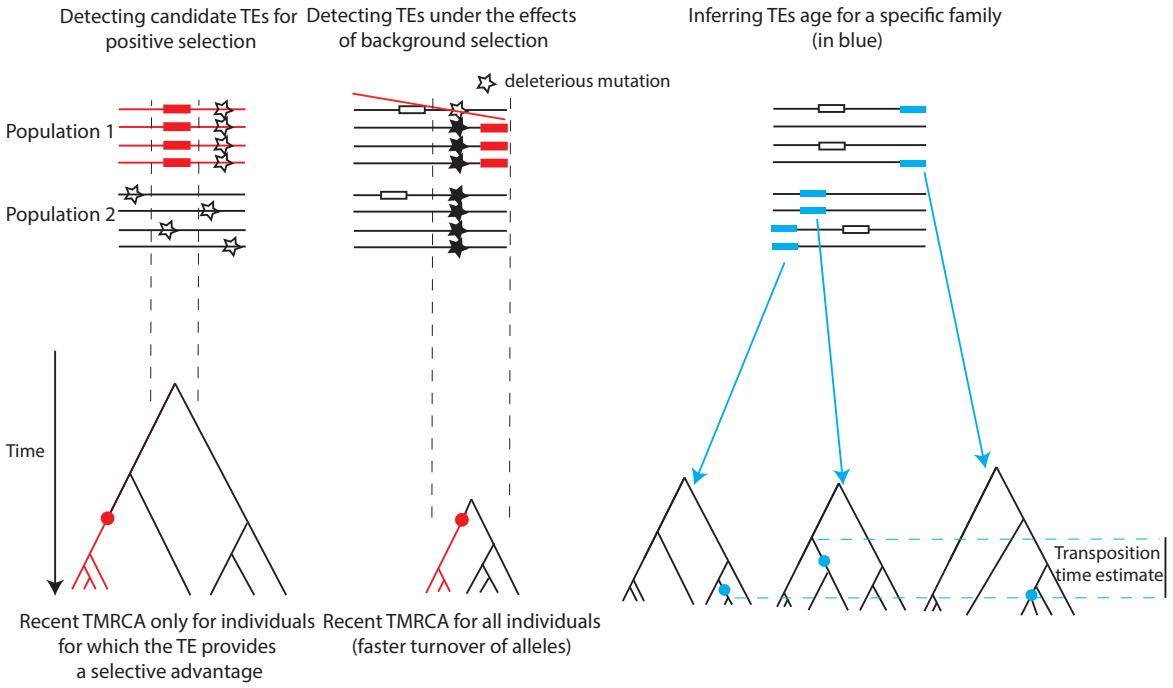


Fig. 3.3 How local genealogies can inform the processes behind TEs diversity. TEs that are under recent positive selection (left) should be associated with genealogies where all branches corresponding to haplotypes carrying the TE coalesce at a recent time. However, the Time since the Most Recent Common Ancestor (TMRCA) for sequences flanking the TE insertion should not drastically differ from the genomic background, since a few lineages escape the selective sweep, or because the allele does not provide an advantage in all populations. TEs that reach high frequencies in one population due to the effects of selection at linked sites, but are not under strong selection themselves, should be found associated with genealogies where the TMRCA is more recent compared to the genomic background. The TMRCA of haplotypes carrying TEs can also be used to estimate the minimum time at which a TE inserted, and can be collected for all insertions assigned to a given family.

theory of molecular evolution (*125*). In this framework, the fate of a variant depends on its scaled selection coefficient $S = N \times s$, with s the selection coefficient and N the haploid effective population size. If $-1 < S < 0$, the allele trajectory of deleterious variants is mostly influenced by genetic drift. In Lynch & Conery's theory, the larger genome sizes observed in many 'complex' organisms compared to unicellular prokaryotes and eukaryotes may result from the lower effective population size (*N)* of the former, preventing the removal of slightly deleterious DNA insertions (e.g. TEs) from the genome. Support for this hypothesis can be elusive when examining lower taxonomic levels (*126*, *127*).

The other term in the aforementioned equation is s, representing the selection coefficient. An alternative hypothesis to differences in TE content posits that TEs or larger genomes may differ in their deleterious effects, depending on their host. For instance, in palm trees, a

negative correlation exists between genome size and aridity (*128*). Similarly, flying tetrapods tend to have smaller genomes (*129*). Selective pressures favoring smaller genomes are linked to disparities in life history traits; for example, aridity or high metabolic demands may drive the evolution of smaller cells in palms and flying tetrapods, respectively. Tetrapod lineages with larger genomes also exhibit higher rates of increase (*129*), suggesting a 'run-away' effect: as the genome becomes increasingly composed of TEs, new insertions are less likely to occur within genes, thereby reducing their average deleterious effects.

Testing such hypotheses requires estimating a distribution of S for all insertions along a genome. Beyond its fundamental interest, accurately quantifying the **distribution of TE fitness effects** holds practical significance for real-world applications. This includes understanding the impact of quantitative genetic variation associated with TEs in human disease or assessing mutation load in endangered species.

One of the first approach (see also (*114*, *130*)) used to estimate S (but not its distribution) was established by Petrov et al., and relied on diffusion models, deriving the likelihood of S for TEs (*111*, *131–133*). These early models could also take into account ascertainment bias: first studies on TEs focused primarily on insertions found in the reference genome but absent in some of the resequenced individuals. With the emergence of second-generation sequencing, obtaining data for both SNPs and TE variants has become increasingly common. This enables the comparison of Allele Frequency Spectra (AFS) and other summary statistics between TEs and (neutral) SNPs (*87*, *111*, *134*). This is similar in spirit to methods estimating the distribution of fitness effects (DFE) for coding variants (*135*): differences between the AFS of non-synonymous and synonymous variants may be attributed to selection (Figure 3.2). However, such methods assume constant mutation (or transposition) rates, which may not always be the case for TEs (*123*). A possible way to overcome this issue consists in contrasting the age of a TE insertion with its frequency, and infer selection if a discrepancy with the near-neutral expectation is observed. For example, a TE insertion under positive selection will have at any given time a higher population frequency than expected under neutrality.

Changes in TE transposition rates and TE half-life may be estimated through their age distribution (*136*). For example, the age of LTR-RTs can be estimated by the divergence between their Long Terminal Repeats. For other TE families, divergence from other insertions can also be used to obtain an upper estimate of their age. This is the approach taken in a pioneering work by Justin Blumenstiel in *Drosophila melanogaster* (*137*). The model uses the age distribution of elements to infer their expected AFS under neutrality, and estimate S. This method requires that the TE's internal sequence be known. As highlighted in the

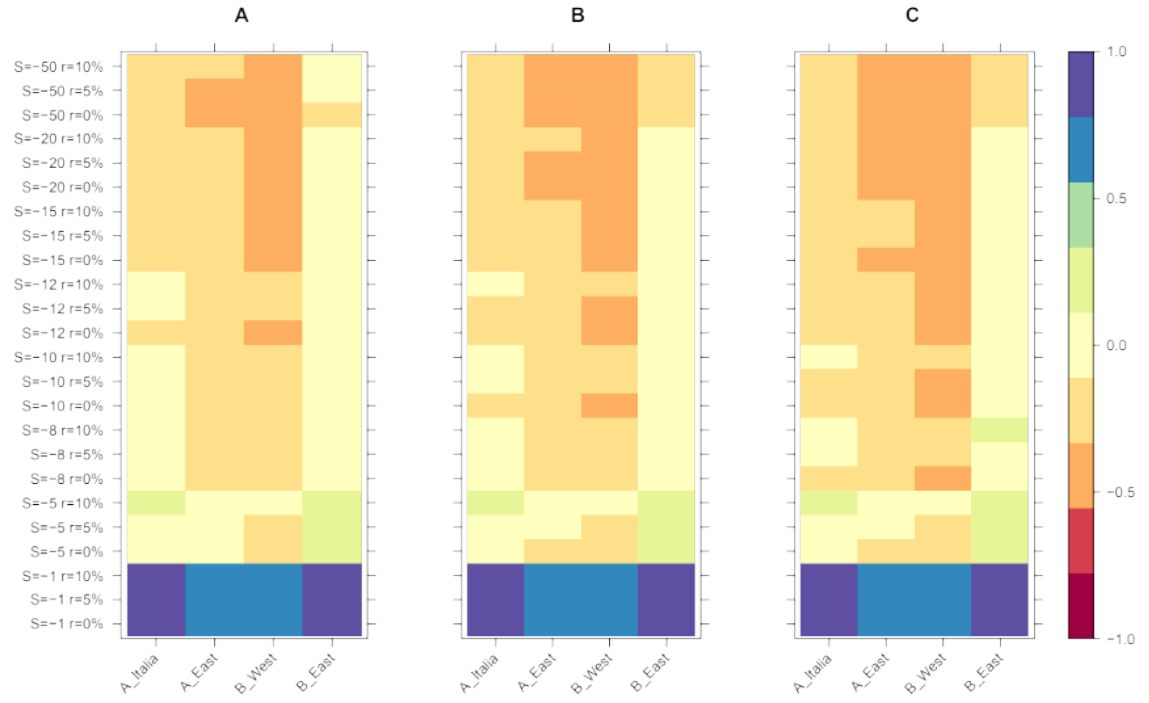


Fig. 3.4 Difference in the age of simulated and observed TEs in four populations of Brachypodium distachyon. S represents the scaled selection coefficient ($S = N \times s$ with s the selection coefficient against TEs and N the effective population size. r represents the fraction of neutral TEs in the simulations. The figure shows the relative age difference ((mutation age in simulations - observed mutation age)/maximum absolute age difference) between simulated and observed data for the oldest polymorphic TEs (>20,000 generations old). A: 25% quantile of simulated ages; B: 50% quantile; C: 75% quantile. The simulations closest to the data were obtained for a <10% fraction of neutral TEs and an average S of -8 or -5. More details can be found in (110).

previous section however, this limitation may be solved soon, with the emergence of long read sequencing and pangenome graphs.

Even if the entire TE sequence is known, mutation rates are usually too low to obtain precise estimates for the age of very recent (and typically rare) insertions. An alternative consists in estimating the age of TEs by inspecting the diversity of flanking genomic sequences (Figure 3.3). This type of approach has been applied by Robert Horvath, a post-doctoral researcher working with my colleague Anne Roulin at the Zurich Botanical Institute (Figure 3.4). The study to which I contributed (*110*) used *GEVA* (*138*) to estimate the age of TE insertions. The frequency of TEs was then compared to the frequency of derived SNPs in discrete age bins of the same size. A simple DFE (with two categories, neutral and under purifying selection) could be qualitatively fitted to the data (Figure 3.4). A difficulty in this work was to estimate the age of rare alleles, which can be informative about purifying

selection given their abundance, but are difficult to date accurately with *GEVA*. A possible alternative could be using *runtc* (*139*), a method estimating the age of first coalescence between alleles. Although not initially designed for TEs, another recent method estimating lengths of pairwise identity-by-state for low frequency variants could be used to estimate the DFE of (TE) variants at a particular frequency (*140*). While powerful, approaches based on tracks of identity by state usually require very large datasets to be applicable (ca 1000 - 10,000 individuals).

My future projects will continue exploring this research avenue. One potentially fruitful approach could involve integrating information obtained from the AFS with TE abundance and frequency along the genome. I am particularly interested in reinforcing the link between functional and population genomics, for example by including regulatory annotations, selective constraints, linked selection, recombination and DNA conformation in models of selection on TEs (see for example (*141–143*)). Possible ways to determine how S fluctuates along the genome from population and functional genomics data could lie in Approximal Bayesian Computation (ABC) or machine/deep learning (*144–147*), to efficiently compare observed data with the results of simulations (obtained with SLiM (*148*) for example). I am also a partner on an ANR project currently under evaluation and led by Thomas Aubier (CNRS, Toulouse), who intends to develop individual-based models that will efficiently model TEs in combination with other genomic features (e.g. large scale inversions).

3.1.3 TEs and plant domestication: an application to a perennial plant, the date palm

Agrobiodiversity is facing a major crisis due to monoculture, the loss of arable lands, reduced genetic diversity, and rapidly changing environments as a result of anthropic impact. Consequently, the potential for adaptation of crops may be impeded, which in turn threatens food security. This is particularly pronounced in the global South, due to the proportionally stronger impact of climate change in developing countries, the use of crops poorly adapted to local environments, and the limited technological transfer from the North. The scale of these challenges calls for enhanced efforts to describe and use genetic diversity in traditional crops used in the Global South. High genetic diversity provides plant populations with a broader set of hereditary phenotypic variation that provides resilience to environmental change, higher yield, or higher nutritive value (*150*). This resilience is critical to efficiently tackle the growing issues of climate change and monoculture. Genetic studies of traditional crops will directly contribute to achieving major Sustainable development objectives, such as

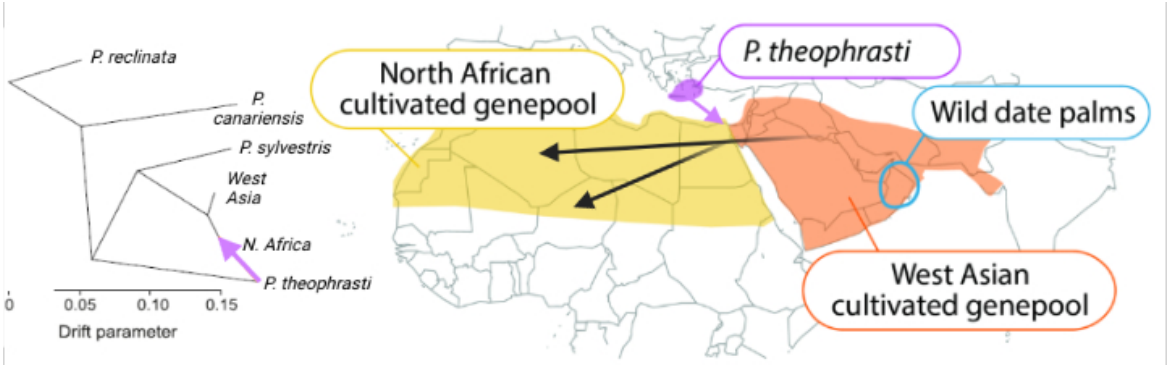


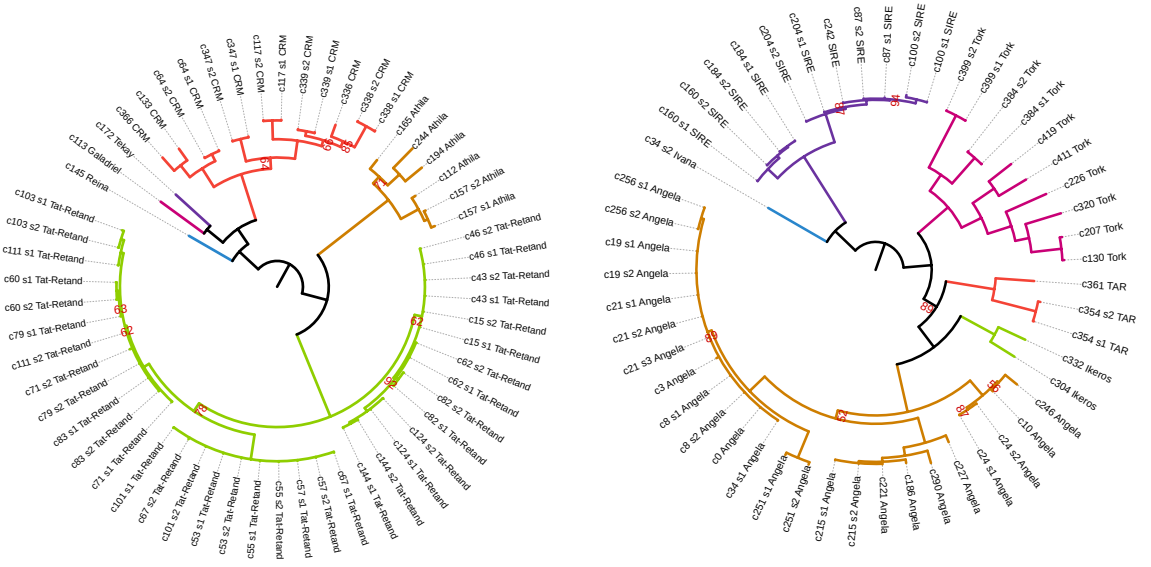
Fig. 3.5 Summary of the evolutionary history of date palm domestication (adapted from (149)). The tree on the left shows relatedness between populations, while the arrow indicates the direction of the most likely introgression event.

reducing inequalities (goal 10), Responsible production (goal 12), Life on land (goal 15), and Zero hunger (goal 2).

In the recent years, the quantification of the adaptive potential of plants has benefited from the development of new sequencing technologies, giving access to genome-wide variation in large samples. This in turn has fuelled the field of population genetics, which aims at understanding how the frequency of variants changes through time and space. Population genetics provides a way to reconstruct the past history of species and to identify genes under selection, also, it benefits from fast and efficient modelling tools that make it possible to predict the fate of alleles and populations. The application of population genetics methods to the study of traditional crops is essential to assess biodiversity, and can contribute to real world applications (*151*). This can include the development of diagnostic tools such as marker-assisted backcrossing, identification of the genetic bases of performant genotypes, or restoring genetic diversity in agricultural practices.

Domestication has a major impact on the DFE of polymorphic variants (*152*). Positive artificial selection may lead to the rapid rise in frequency of variants previously neutral in natural populations. On the other hand, domestication is often associated with strong bottlenecks that reduce the adaptive genetic potential, leading to an increase in the so-called genetic load in crops ("cost of domestication"). These mechanisms have come under increased scrutiny with the advances of second-generation sequencing. This has resulted in valuable insights regarding the role and impact of SNPs on adaptability (see (*153*) for a review in crops).

While most TE insertions are at least slightly deleterious, TEs are recruited by selection during crop domestication. They can either disrupt a gene with minimal pleiotropic effects by inserting in their coding sequence, or modify the expression of nearby genes through



(a) Ty3/Gypsy *(b) Ty1/Copia*

Fig. 3.6 Ongoing work by Valentin Grenet. Phylogeny based on the consensi of the reverse transcriptase for abundant (> 5 full-length copies) LTR-RTs in the date palm genome. Bayesian support values (MrBayes(155)) lower than 95% are indicated. The annotation of each lineage is also indicated. Branches corresponding to the same lineage are indicated with the same colour. Possible subfamilies identified during manual curation are indicated with a "s" suffix followed by a number. For example, c60_s1_Tat-Retand and c60_s2_Tat-Retand would cluster together using Wicker's 80-80-80 rule (156)

.

epigenetic silencing and regulatory changes (*107*). For example, retrotransposons contain promoters, which can lead to the over-expression of downstream genes. In several species, TEs inserting in promotors are found associated with resistance to drought (*154*).

Studies that did examine the role of TEs in domestication have focused on annual crops, biasing our understanding of the domestication process in plants (*157*). The involvement of TEs in perennial domestication is even more elusive than for annuals, but anecdotal evidence suggest they have been important in the domestication of perennial crops (*158*). Beyond these examples, little is known about the cumulative effect of these 'jumping genes' on their host. Which fraction of TEs is near-neutral, deleterious or advantageous has yet to be quantified (see previous section). Perennial plants are more often outcrossers than annuals, and are less likely to have gone through severe bottlenecks during domestication, broadening the genetic diversity upon which (positive or negative) selection can act effectively. Introgression is another process that introduces genetic diversity and is frequent in perennial crops. On the other hand, modern monovarietal practices are reducing effective population sizes, and

perennials display longer generation times and lower mutation rates than annual crops (*159*), which may reduce their resilience in front of fast environmental changes.

The date palm constitutes an ideal perennial crop to study the evolutionary impact of TEs during domestication. Studying TEs in date palms is of course of agronomic relevance. The date palm is a keystone species of the oasis agrosystems in the Middle-East and North Africa. While date consumption increases every year, its cultivation is threatened, especially by more frequent droughts and increased water salinity. Biotic threats are also an issue, with a recent collapse of palm groves due to pests such as red palm weevil (*Rynchophorus ferrugineus*). A precise characterization of its TE diversity is essential to understand the adaptability of this species to growing environmental pressures.

TEs account for almost half of the date palm genome (*160*). Most of these elements are Long-Terminal-Repeat retrotransposons (LTR-RTs) and belong to the *Ty1/Copia*(26% of the genome) and *Ty3/Gypsy* (12%) superfamilies. DNA transposons make up for less than 5% of the genome. The subfamilies and lineages of LTR-RTs found in the date palm genome are also extremely diverse, as highlighted by an ongoing study by Valentin Grenet (Figure 3.6). These elements have been recently active and are polymorphic, with copies showing low divergence from their consensi accounting for 3% of its 850Mb genome (Figure 3.7). There is evidence that TEs play a role in phenotypic variation: for example, an insertion is strongly associated with variation in fruit colour (*160*). Moreover, relative TE abundance is associated with aridity in palm trees (*128*), with *Ty1/Copia* being more abundant in species living in arid conditions. TEs are polymorphic and likely active. I recently genotyped polymorphic TEs in short read data for ca 600 individuals, revealing more than 130,000 polymorphic insertions (including both reference and non-reference TEs). We are currently improving TE annotation, particularly for DNA transposons, in collaboration with Pr. Josep Casacuberta's team in Barcelona (CRAG) and Dr. Clémentine Vitte (CNRS, Paris-Saclay University).

In March 2023, I started receiving pilot data for PacBio HiFi long-read sequencing (funded by a NEOF grant, ten samples, average coverage of 10X), and called TEs for a curated, abundant *Ty3* family using a graph-based approach (GraffiTE (*121*)). Short-reads and long-reads approaches show an overlap of 70% (Figure 3.7). Discrepancies between the two methods may be due to the relatively low depth of preliminary PacBio sequencing data (8-10X) and poor sensitivity in highly repetitive regions for MEGAnE. Transposition events may also have occurred independently in samples used for short and long reads sequencing, as those were obtained from distinct clones of the same variety. An important aspect from a population genetics perspective is the good match between genotypes obtained from short and long reads, with MEGAnE possibly underestimating the number of homozygote insertions. Overall, these results confirm the interest of combining short and long reads data

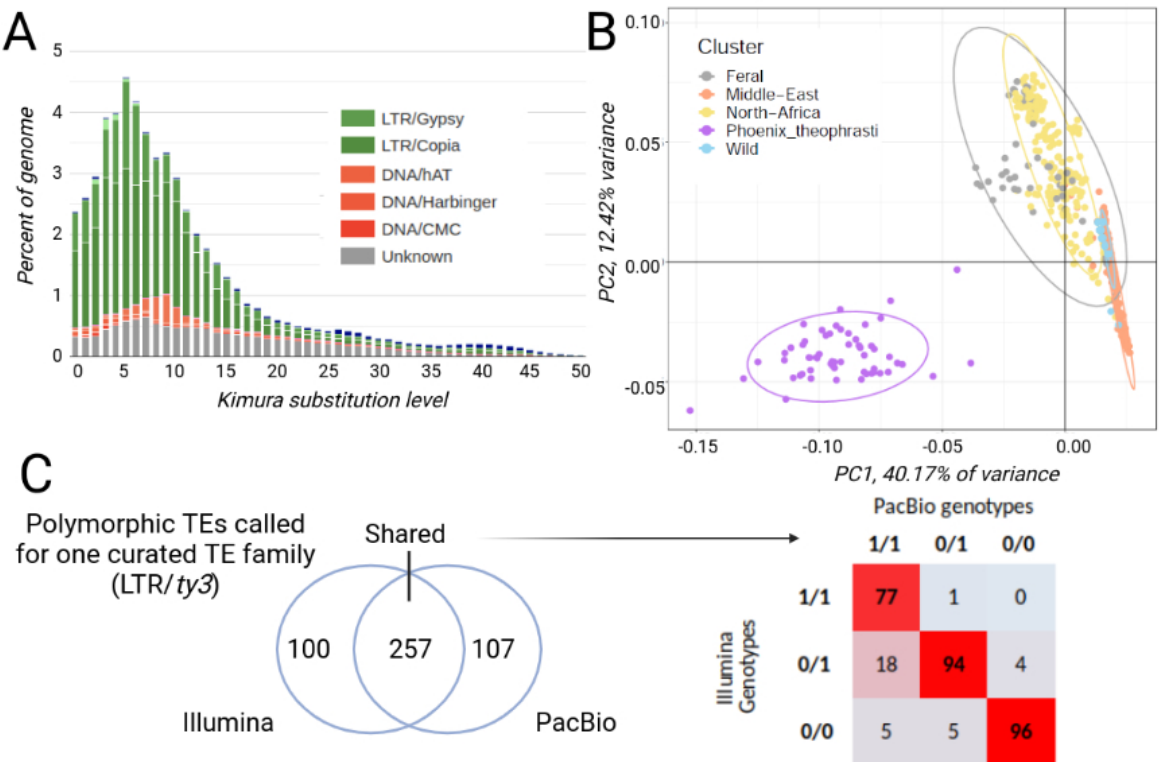


Fig. 3.7 A. Distribution of TE divergence from their consensus. Results are based on annotation using INPACTOR2 (161) to obtain full length LTR-RT elements along with their classification, as well as RepeatModeler/RepeatMasker (162), with redundant/overlapping TEs merged using the 80-80-80 rule (156)), giving a total of 877 TE families.B. Principal Component Analysis on TE genotypes for P. dactylifera and P. theophrasti (short read data). Structure is highly consistent with SNP-based results (149). C. Number of TE insertions identified with short and long reads in four clonal varieties. Right: Percentage of Illumina insertion genotypes matching with PacBio genotypes.

to benchmark genotype callers and possibly correct for biases. A more detailed analysis with Valentin Grenet is ongoing, and will use data from 22 palm trees sequenced with both Illumina and Oxford Nanopore at 40X (read N50 of 10kb for the latter, funded by the Royal Society). This dataset will likely solve the possible issue of relatively low depth.

I received funding in January 2024 from an ANR JCJC grant (DaTEPalm) to quantify the impact of TEs of their date palm host during domestication. I intend to implement a trans-disciplinary approach (Figure 3.8) which combines inference from population genomics analyses (Objective I) with detailed annotations of (epi-)genomes and gene-by-gene interactions (Objective II). The project comes with funding for one PhD student (Objective I) and one post-doctoral researcher (Objective II), and will be include collaborators from France, the UK, and the United States. The access given by Dr. Gros-Balthazard to hundreds of re-sequenced genomes from wild relatives, sister species, introgressed and feral populations

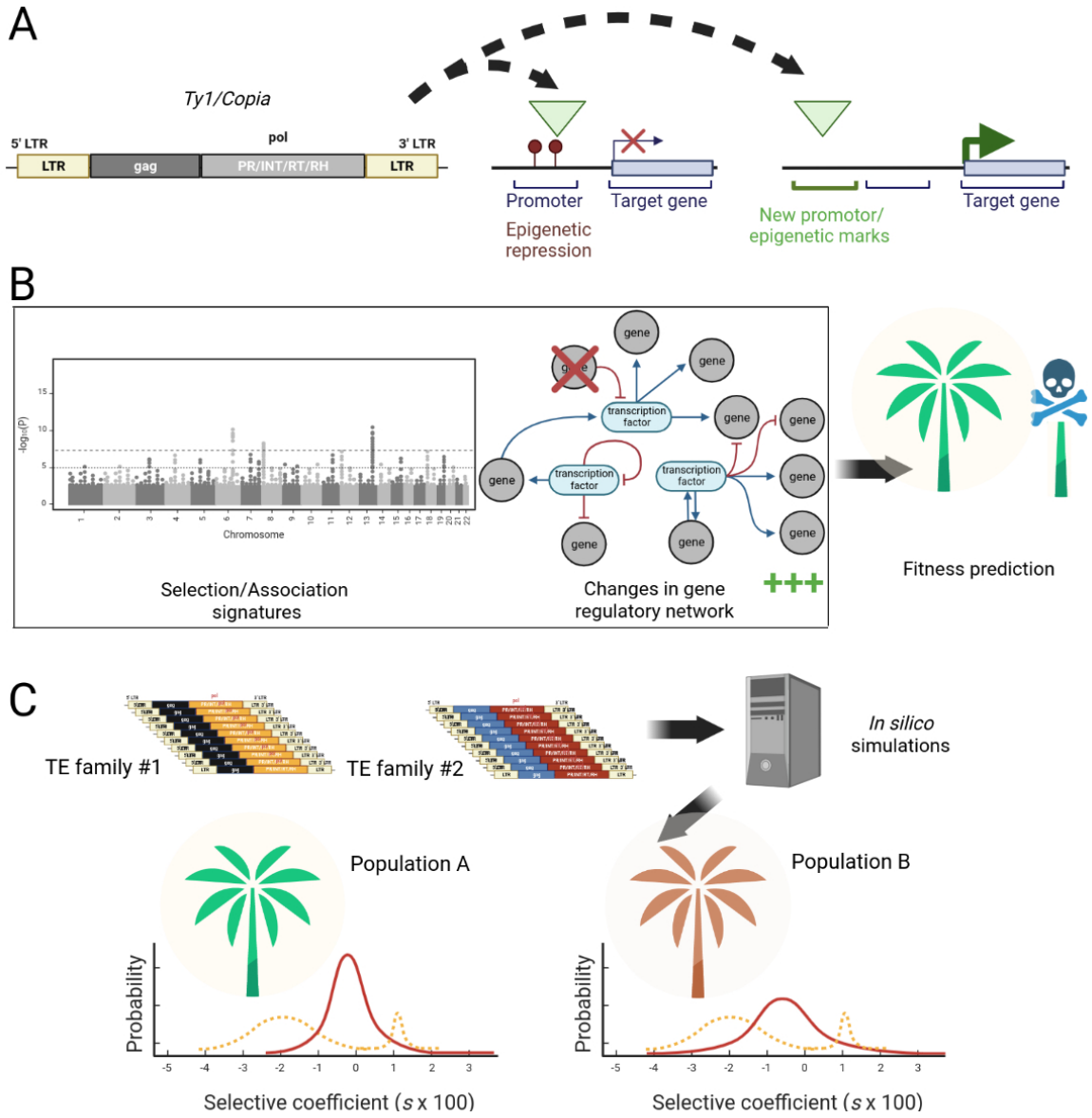


Fig. 3.8 A graphical summary of the project on TE diversity in date palms. A. A TE inserts, leading to changes in genome structure and regulation. B. These changes can be the target of negative (X) or positive (+++) selection, altering the TE frequency at the insertion site, and disrupting gene networks. These signatures of selection could be used to determine the fitness effect of a single element. C. The ultimate aim consists in using population and functional signatures of selection to estimate the average fitness effects for cohorts of polymorphic TEs, and even the distribution of these effects. Comparisons can then be made across TE families, genomic regions, host populations, species, while the effect of genomic context ("niches") can be addressed.

gives high power to test for the effects of selection in shaping the genomes during domestication. Moreover, collaborators at Michael Purugannan's lab (New York University) are generating a pangenome, which will greatly facilitate annotation.

Objective I: Using population genomics to quantify positive and negative selection on TEs.

The PhD student and I will identify TEs associated with genomic signatures of selection and retrace their historical changes in frequency to confirm their involvement in domestication. These scans for selection will carefully consider sources of false positives. We will run simulations and examine allele genealogies near TEs to disentangle whether they are the actual target of positive selection or because they hitch-hike with another advantageous variant. We will also test whether deleterious TEs have accumulated in domesticated palms compared to their undomesticated relatives. We will contrast TEs age, abundance and frequency to estimate their deleterious effects across populations, and estimate the timing of their activity. We will build population genetics models to estimate genome-wide fitness effects of TEs and their distribution.

Some TEs may be variants of large effects under positive selection. We will test whether TEs can be found in candidate regions for (recent) adaptation and adaptive introgression, contrasting low-diversity and elite varieties from the West Asia with more diverse North-African populations, including wild relatives. Many variants that are deleterious in the wild may become positively selected or effectively neutral. We will also test whether TEs have been recruited by selection from preexisting standing variation or originate from *P. theophrasti*, using simulations built on neutral demographic models to correct for false positives.

Given their deleterious potential, TEs may constitute a substantial fraction of crops mutation load and that the distribution of their fitness effects differs between wild, domesticated and feral populations (*163*). However, in perennials, this divergence may be mitigated by less intense selective constraints and bottlenecks. We will use population genetics approaches to estimate and compare fitness effects of the most abundant TE families between undomesticated and domesticated populations to test for an accumulation of deleterious TEs in the latter. We will compare these results with estimates of selective coefficients inferred from Single Nucleotide Polymorphisms (SNPs).

Objective II: Gene network topology and selective constraints on TE insertions.

Fitness effects of TE insertions may be predicted from the position in the regulatory network of TEs target genes. The post-doctoral researcher hired on the project and I will combine the population genetics approaches from Objective I with functional genomics to quantify the selective constraints on TEs. By combining these two layers of information, we will build integrated models estimating the fitness effects of TEs, and support cases of positive selection of TEs with functional evidence. Using transcriptomics and epigenomic approaches, we will annotate regulatory regions of the date palm genome sequence, with the help and expertise of Dr. Clémentine Vitte and Dr. Maud Fagny (CNRS and INRAE, Paris-Saclay University). Using a systems biology approach, we will reconstruct gene regulatory networks for two distinct tissues, and will document the genes and corresponding functions that are altered by TE insertions.

Studies of TEs in various palms have highlighted a positive correlation between the abundance of *Ty1/Copia* and aridity. Genome size also appeared more constrained in arid environments. One may therefore expect different levels of activity and control between the two most abundant LTR-RT found in the date palm genome (*Ty1* and *Ty3*). *Ty3* elements may be more uniformly repressed, whereas stress-inducible *Ty1* (*164*) may be more dynamic and have a broader DFE. This builds upon the work he is currently leading. We will test this by examining epigenetic marks on copies of these two TE clades, and contrast the positions of the genes they disrupt in the gene expression network. We also plan to investigate differences in TE activity using phylodynamic models in collaboration with Valentin Grenet. We will also examine the genomic distribution of TEs across clones of the same cultivar to obtain more information about ongoing TE activity and insertion preferences. Gene interactions can be summarized by a network, which can be divided itself into modules of genes involved in a common biological process.The position of a gene within a regulatory network may be related to the amount of selective constraints that act upon it (*165*). Genes that are at the core of the entire network are under strong selective constraints, and are unlikely to be the target of positive selection (*165*, *166*). TE insertions in these genes are therefore expected to be extremely rare, and TE polymorphisms should display typical patterns of purifying selection (i.e. low frequency and young age). On the other hand, the probability for genes to be under positive selection increases towards the edges of the network (optimal pleiotropy). We will particularly focus on TEs found near these genes as candidates for positive selection. We will test for consistency between population genomics estimates of fitness effects and the distribution of insertions across the regulatory network, clarifying the impact of TEs on their host.

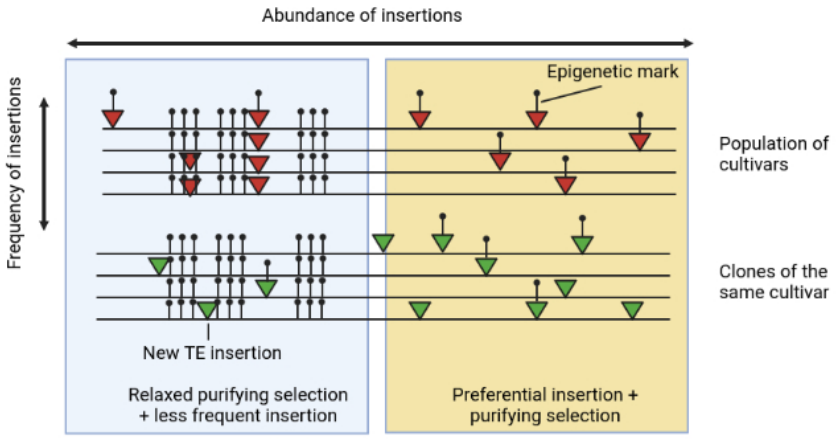


Fig. 3.9 Information about the functional impact of TEs (position along the genome and regulatory networks, epigenetic variation at TEs and flanking regions) will help build more comprehensive models of selection. Purifying selection should prevent TEs from reaching high frequencies, while differences in insertion rates may affect the abundance of polymorphic and fixed insertions. Clonal haplotypes cannot recombine, and new TEs insertions are heterozygous, shielding them from the effects of selection. Examining TEs segregating between clones may provide information about TEs dynamics before (strong) selection acts.

Development of partnerships

The project is strongly anchored in the Global South, particularly in the Mediterranean area, and focuses on a keystone species of oasis agrosystems. It will contribute to the transfer of methodologies and technologies, and benefit from the extensive network of partnerships in Djibouti, Tunisia and Morocco; which has already been developed in my host UMR, DIADE. I have extensive teaching experience, and have taught in front of a diverse international audience. I intend to bring this experience to the project. I will contribute to the development of local research expertise and infrastructure through regular stays in the South Mediterranean region, either through expatriation or *Mission Longue Durée*. Tunisia is a particularly interesting location for an expatriation project: its central location in North Africa will facilitate the collection of new samples for the sequencing of new varieties, and the extensive partnership with IRD will facilitate the local development of collaborations and workshops. I will also actively engage in the supervision of students from the Global South (Msc, PhD projects). The production and analysis of genetic data over the course of this project will follow the Nagoya Protocol, which ensures the fair and equitable sharing of benefits arising from the utilization of resources. The ultimate goal of this study consists in creating an openly accessible and well annotated genetic resource for practitioners. In collaboration with IRD, I will develop online resources for the date palm and any other species for which data are obtained.

3.2 Conclusions

I have directed my research profile towards molecular ecology, and the population genomics of TEs. Today and in the future, I aim to remain at this interface by developing -omics projects, as illustrated by the DaTEPalm project (see section above).

Throughout my career, I have used population genomics to link proximate and ultimate causes of phenotypic variation (*167*), and reconstruct the past history of species. I have witnessed and benefited from a dramatic technological burst over the last twelve years. Towards the end of my PhD, I employed Pool-Seq and RAD-sequencing techniques. Currently, I am engaged in whole-genome resequencing using portable platforms, which generate long reads with 98-99% accuracy, capturing both genetic and epigenetic variations. This endeavor has involved leveraging a diverse array of bioinformatic tools and establishing robust collaborations with ecologists, bioinformaticians, and conservationists.

Quite strikingly, despite – or, perhaps, because of – technological advancements, many debates in population genetics continue to revolve around the same themes, such as the definition of biological function (*168*), or the adaptationist-neutralist perspectives (*78*, *79*). Some of these controversies arise from advertising strategies, underscoring the significance of clear conceptual understanding and awareness of limitations, as well as exercising caution in the face of narrative storytelling.

In the near future, it appears improbable that data production will continue to pose a significant challenge for medium to large laboratories of the Global North, especially for species with genomes smaller than 2 or 3 Gb. However, recent years have witnessed a surge in the development of bioinformatic pipelines and analytical methods for population genomics. Keeping pace with methodological advancements has become progressively challenging. Nonetheless, there remains a gap between wet lab procedures and analytical pipelines, with the latter sometimes failing to fully exploit the wealth of information contained in the data.

To help alleviate these issues at my own scale, I have shared and organized my own experience and intend to keep working in this direction. An increasing part of my work is carried out through active supervision of students and assistance to colleagues who are not familiar with population genomics techniques. This allows me to multiply myself on various research themes simultaneously. In return, these partners evolve in a multidisciplinary and collaborative framework that, I hope, will offer them a wide range of opportunities to develop their own careers in the field of their choice. Ultimately, I would like the fundamental insights gathered over the course of my research to inform discussions on conservation and management strategies.

Chapter 4

Résumé en Français

4.1 Récapitulatif des activités de supervision et gestion de la recherche

4.1.1 Encadrement de Master

J'ai commencé à superviser des étudiants de troisième cycle à titre officiel en 2020, au cours de ma première année d'enseignement à Portsmouth. Les étudiants de *MRes* (Master of Research) étaient évalués sur leur capacité à rédiger une demande de subvention décrivant leur projet, un rapport bibliographique, ainsi qu'un rapport de recherche rédigé dans le format d'un article scientifique. J'ai supervisé deux étudiants en *MRes*, Harry Simmonds (2020-2021) et Daniel Bedford (2021-2022).

Harry Simmonds

Projet : Le projet de *MRes* consistait à étudier la divergence et la diversité des gènes de floraison dans plusieurs populations de lin pâle (Linum bienne). Les travaux de M. Simmonds suggèrent qu'une sélection positive récente s'est produite au niveau de quelques gènes de floraison bien connus, tels que *TFL1*. Les résultats obtenus seront inclus dans une future publication décrivant la diversité génétique et l'histoire démographique passée du lin pâle, le plus proche parent du lin domestique (*Linum usitatissimum*).

Devenir professionnel : M. Simmonds est maintenant **doctorant** à l'Université de Reading, où il développe des stratégies visant à améliorer la résistance des cultures aux pathogènes en tenant compte de la diversité des gènes de résistance des plantes.

Daniel Bedford

Projet : Le projet consistait à étudier la diversité des éléments transposables (ETs) dans un clade de "grands spéciateurs", les oiseaux du genre Zosterops. L'activité récente des ETs est principalement limitée aux rétrotransposons à LTR dans ce clade. Une phylogénie basée sur les ETs est cohérente avec celles obtenues avec les variantes SNP. Les espèces insulaires présentent un nombre plus élevé d'éléments transposables que les espèces continentales, bien que cette différence ne soit pas significative en raison du nombre limité d'espèces pour lesquelles des données de reséquençage sont disponibles. Une explication possible pourrait être l'accumulation de la charge des ETs dans les populations insulaires en raison de leur taille efficace plus réduite.
Devenir professionnel : M. Bedford est aujourd'hui analyste de recherche chez Juniper Research, une société qui offre des services de conseil dans le domaine des télécommunications.

De retour en France en 2023, j'ai récemment obtenu un financement de l'Agence Nationale de la Recherche pour étudier la dynamique des éléments transposables (ETs) des palmiers dattiers (*Phoenix dactylifera*). Ce projet finance notamment un étudiant de M2, Valentin Grenet, qui a commencé son projet en février 2024.

Valentin Grenet

Projet : Valentin Grenet est étudiant en M2 à Polytech Nice Sophia. Son projet consiste à annoter les rétrotransposons à *LTR* dans le génome de référence du palmier dattier. Il a déjà réussi à obtenir des séquences consensus pour les lignées les plus abondantes, a décrit leur diversité par des méthodes phylogénétiques (voir figure 3.6 au chapitre 3), et estime l'âge des rétroéléments à *LTR*. L'âge et l'abondance des éléments seront comparés aux caractéristiques génomiques (recombinaison, densité des gènes, contenu en bases, etc.) afin d'étudier l'interaction entre les TE et leur hôte. Si le temps le permet, des données Nanopore pour 22 palmiers seront utilisées pour étudier plus en détail l'abondance et la fréquence des éléments polymorphes.
Devenir professionnel : M. Grenet est désormais en thèse sous ma supervision ainsi que celle de R. Guyot (HDR).

Je participe également à la supervision de Kilian Dolci, en collaboration avec Valérie Poncet (Institut de Recherche pour le Développement).

Kilian Dolci

Projet : Kilian Dolci est étudiant en M2 à l'Université de Toulouse. Son projet porte sur les ETs polymorphiques dans les populations divergentes de *Coffea canephora*. Il a utilisé avec succès MEGAnE (*3*) pour appeler les ETs polymorphiques dans un ensemble de plus de 70 individus, et vise à identifier les ETs candidats à l'adaptation et à quantifier la charge mutationelle des cultivars. Il compare actuellement les spectres de fréquence allélique et la diversité des ETs entre groupes génétiques, et réalise une analyse d'association environnementale à l'échelle du génome à l'aide de *LFMM2* (*4*).
Devenir professionnel : En recherche de thèse.

4.1.2 Encadrement doctoral

En 2021 et 2022, j'ai été le principal superviseur d'un doctorant à temps partiel, Thomas Heller, basé à Kew Gardens. Les superviseurs à Kew étaient le Dr Juan Viruel et le Dr Martin Hamilton. Mr Heller a travaillé aux jardins de Kew en tant qu'ingénieur de recherche pendant plusieurs années, ce qui lui a permis d'acquérir une certaine indépendance dans le travail de terrain et les évaluations écologiques. Ma contribution s'est principalement concentrée sur la rédaction de rapports et les analyses génétiques des populations (à l'aide des kits Angiosperm353 (*5*) et de séquençage *RAD* (*169*)).

Thomas Heller

Projet : Quantification de la diversité génétique d'un arbre endémique menacé des Caraïbes, *Zanthoxylum thomasianum*. Le projet se concentre sur les plantes menacées dans les Caraïbes, avec des projets actifs dans les îles Vierges (britanniques et américaines) et à Porto Rico. M. Heller s'intéresse particulièrement à l'évolution du genre Zanthoxylum (Rutaceae) dans les Caraïbes par le biais d'études de terrain et d'approches moléculaires (génétique des populations et phylogénomique). Les aspects moléculaires du projet comprennent des analyses phylogénétiques visant à clarifier l'histoire de l'évolution à long terme du genre, ainsi que des analyses de génétique des populations et de génétique du paysage chez *Z. thomasianum*, afin d'identifier les menaces liées au manque de corridors écologiques et à la perte récente de diversité génétique.
Devenir professionnel : En cours

En raison de ma prise de poste à l'Institut de recherche pour le développement en 2023, le principal superviseur de M. Heller à l'université de Portsmouth est désormais mon collègue,

Dr Steven Dodsworth. Je reste néanmoins impliqué dans la supervision de M. Heller. Au moment de mon départ, le projet était en bonne voie, avec près de 400 échantillons collectés à travers Porto Rico et les îles Vierges britanniques. Les données de séquençage sont attendues pour 2024 et la publication des résultats pour 2025.

Depuis septembre 2022, je participe également à la supervision d'Anastasia Kolesnikova à l'université de Southampton, avec le Pr. Mark Chapman comme superviseur principal. Son projet se concentre sur la comparaison de la diversité génétique entre les plantes domestiquées et leurs parents non domestiqués, afin de déterminer si les plantes domestiquées présentent des propriétés désirables en termes de taux de mutation, de charge mutationnelle et de variation préexistante sur laquelle la sélection humaine aurait pu agir.

Je suis toujours impliqué dans des collaborations avec l'Université de Portsmouth. Je suis notamment le co-superviseur de Mme Snata Chakraborty et de M. Thomas Roberts-McEwen avec le Dr Lena Grinsted comme superviseur principal. Leurs projets portent sur l'écologie et l'évolution de *Cyrtophora citricola*, une espèce d'araignée vivant en groupe et faisant preuve d'une faible agressivité intraspécifique. Le projet de M. Roberts-McEwen se concentrera sur l'utilisation de cette espèce pour le biocontrôle des ravageurs en Asie du Sud-Est. Mme Chakraborty se concentrera davantage sur l'évolution de l'espèce, en s'appuyant sur les résultats de séquençage *RAD* recueillis par un autre étudiant de premier cycle que j'ai supervisé, M. Nathaniel Holmes. J'ai bénéficié d'une formation approfondie à la supervision au Royaume-Uni, dispensée par l'université de Portsmouth, ainsi que par l'école doctorale BBSRC/SoCoBio. J'ai également été membre de jury dans plusieurs comités de suivi de thèse (voir CV, Chapitre 1).

Depuis janvier 2023, je co-supervise deux doctorants en France. À l'Institut de recherche pour le développement, je supervise Margot Beisseiche (25 %), avec le Dr Muriel Gros-Balthazard et le Dr François Sabot (HDR). Le projet de Mme Beisseche repose sur la collecte d'ADN ancien à partir de restes de palmiers dattiers afin de retracer l'histoire de la domestication et de la diffusion du palmier dattier à travers le Moyen-Orient et l'Afrique du Nord. Mme Beisseiche est désormais en deuxième année et est parvenue à extraire l'ADN de graines libyennes vieilles de 1000 ans. Elle a contribué à la création d'une base de données archéologiques qui sera librement accessible en ligne. Depuis novembre 2024, j'encadre Valentin Grenet, sous la direction officielle du Dr R. Guyot (HDR). M. Grenet poursuivra ses recherches sur la dynamique des populations d'éléments transposables chez le palmier dattier, en combinant le séquençage Nanopore avec une génomique de population avancée pour obtenir des estimations des effets de fitness des ET et de la charge de mutation chez cette espèce. Je participe également à la supervision de Maxime Criado à l'Université Paris-Saclay (10%), avec le Dr. Amandine Cornille et le Dr. Elodie Marchadier. M. Criado

est actuellement en première année, et développe des pipelines de prédiction génomique aux conditions environnementales futures (*genomic offset*) sur les arbres fruitiers. J'ai prodigué des conseils sur les analyses de génomique des populations et les procédures bioinformatiques.

4.1.3 Contribution à des articles scientifiques impliquant des doctorants

J'ai apporté un soutien significatif à cinq doctorants lors de l'analyse des données et de la rédaction de leurs manuscrits.

Liste d'articles impliquant une contribution significative au projet du doctorant

(7) **Mira-Jover, A.**, Graciá, E., Fritz, U., Giménez, A., **Bourgeois, Y.** Taking advantage of reference-guided assembly in a slowly-evolving lineage: application to *Testudo graeca*. PloS One. Voir **Annexe B.**

(6) **Minadakis, N.**, Kaderli, L., Horvath, R., **Bourgeois Y.**, Xu, W., Thieme, M., Woods, D. P. Roulin, A. C. (2024). Polygenic architecture of flowering time and its relationship with local environments in the grass *Brachypodium distachyon.* Genetics.

(5) Horvath, R., **Minadakis, N.**, **Bourgeois Y**. Roulin, A. C. (2024). The evolution of transposable elements in *Brachypodium distachyon* is governed by purifying selection, while neutral and adaptive processes play a minor role. eLife 12. https://elifesciences.org/reviewed-preprints/93284v2

(4) **Minadakis, N.**, Williams, H., Horvath, R., Caković, D., Stritt, C., Thieme, M., **Bourgeois, Y.**, Roulin, A. C. (2023). The demographic history of the wild crop relative *Brachypodium distachyon* is shaped by distinct past and present ecological niches. Peer Community Journal, 3.

(3) **Mould, M. C.**, Huet, M., Senegas, L., Milá, B., Thébaud, C., **Bourgeois, Y.** Chaine, A. S. (2023). Beyond morphs: Inter-individual colour variation despite strong genetic determinism of colour morphs in a wild bird. Journal of Evolutionary Biology, 36(1), 82–94.

(2) **Ameline, C.**, **Bourgeois, Y.**, Vögtli, F., Savola, E., Andras, J., Engelstädter, J., Ebert, D. (2021). A two-locus system with strong epistasis underlies rapid parasite-mediated evolution of host resistance. Molecular Biology and Evolution, 38(4), 1512–1528.

(1) **Bourgeois, Y.**, **Stritt, C.**, Walser, J.-C., Gordon, S. P., Vogel, J. P., Roulin, A. C. (2018). Genome-wide scans of selection highlight the impact of biotic and abiotic constraints in natural populations of the model grass *Brachypodium distachyon*. The Plant Journal, 96(2), 438–451.

J'ai fait partie du comité de doctorat de Nikolaos Minadakis pendant trois ans (2021-2024). Le doctorat de M. Minadakis a été supervisé par Mme Anne Roulin (Institut de botanique, Université de Zurich) et visait à étudier l'adaptation locale et l'évolution du temps de floraison chez la graminée modèle *Brachypodium distachyon*. J'ai contribué à sa supervision et suivi ses progrès. J'ai contribué aux analyses démographique et d'association génotype-environnement (voir Chapitre 3). J'ai participé aux travaux d'un autre doctorant du même laboratoire, Christoph Stritt.

J'ai assisté Claire Mould dans ses analyses des bases génétiques du polymorphisme de couleur chez un oiseau endémique de la Réunion (*Zosterops borbonicus*), que j'ai étudié pendant mon doctorat. J'ai contribué au développement et à l'analyse de marqueurs diagnostiques pour la coloration du plumage. J'ai également contribué à l'interprétation de quelques divergences entre les génotypes et les phénotypes, en les expliquant comme des événements probables de recombinaison entre le locus de couleur causal et les marqueurs diagnostiques. J'ai contribué de manière significative au projet de doctorat de Camille Ameline sous la supervision du Pr. Dieter Ebert à l'Université de Bâle. J'ai contribué à une étude d'association à l'échelle du génome visant à identifier les loci qui sous-tendent la résistance de *Daphnia magna* à des souches distinctes du parasite *Pasteuria ramosa*. J'ai contribué à l'interprétation des résultats, révélant comment l'interaction épistatique entre deux loci majeurs produit les phénotypes de résistance (voir (*61*)).

Tous les étudiants susmentionnés ont défendu leur doctorat avec succès.

Je développe actuellement des projets dans le domaine de la génomique de la conservation et je participe activement au projet de doctorat d'Andrea Mira Joves à l'université Miguel Hernández d'Elche (Espagne). J'ai organisé le partage de scripts, l'accès à un cluster de calcul, et organisé ateliers et réunions détaillant les principales étapes de l'assemblage de génome. Ce travail a conduit à la production d'un génome de référence pour la tortue grecque vulnérable (*Testudo graeca*, NCBI BioProject PRJNA1086345). J'ai également contribué à la rédaction d'un manuscrit décrivant ce travail, qui sera soumis dans les mois à venir à *PloS One* (cf en annexe).

4.1.4 Supervision de chercheurs post-doctoraux

En 2024, j'ai contribué au recrutement de Qindong Tang et Samuel Gronard sur un projet de post-doctorat ANR mené par mon collaborateur Ben Warren (Muséum National d'Histoire Naturelle). Le Dr Tang compare la diversité génétique existante et passée d'espèces d'oiseaux endémiques des Mascareignes. Je suis particulièrement impliqué dans le conseil sur l'utilisation d'outils génomiques capables de gérer les faibles profondeurs de séquençage obtenues à partir d'échantillons subfossiles et toepad (par exemple la suite *ANGSD* (*8*)), ainsi que les méthodes dédiées à l'inférence d'événements démographiques récents (par exemple *HapNe* (*9*), la correction de la dérive différentielle dans les échantillons modernes et anciens (analyse factorielle dans (*10*)), ou la détection de sélection positive en cours dans les données temporelles (*11*). J'ai récemment accueilli les Dr Tang, Gronard, et Warren à Montpellier où nous avons discuté des pipelines de génomique des populations qui pourraient être déployés.

Je suis également impliqué depuis 2024 dans la co-supervision de Ernesto Testé, avec ma collègue Muriel Gros-Balthazard (IRD). M. Testé s'efforce de retracer l'histoire de la culture

du palmier dattier au Levant, et se fondera sur l'ADN ancien pour comparer la diversité passée et actuelle. Là encore, je procure des conseils sur l'utilisation des outils génomiques modernes.

4.2 Recherche passée

4.2.1 Projet de thèse (2009 - 2013)

Je suis un biologiste de l'évolution ayant travaillé sur de nombreux modèles animaux et végétaux. Mes travaux consistent en l'analyse de données de génomique obtenues à l'échelle de populations pour reconstruire l'histoire passée des espèces et quantifier le rôle de la sélection, positive comme négative. J'ai obtenu mon doctorat en 2013, après avoir identifié une région génomique liée au polymorphisme de coloration du plumage chez un passereau endémique de La Réunion (*Zosterops borbonicus*).

J'ai montré, grâce à une analyse comparative des données de Pool-Seq, de GBS et de reséquençage du génome entier, que des signaux forts de sélection positive récente et d'association pointaient tous vers la même région génomique sur le chromosome 1, dans laquelle aucun gène de couleur précédemment décrit n'a pu être identifié (*15*). J'ai également démontré l'absence de structure génétique entre les morphes de couleur en dehors de ce locus unique (*15*), ce qui suggère qu'une certaine forme de sélection balancée pourrait maintenir les morphes de couleur dans toutes les populations à d'altitude. La nature exacte de la pression sélective reste néanmoins à identifier.

J'ai également estimé la différenciation génomique entre toutes les formes de couleur, en incluant les formes parapatriques trouvées à plus basse altitude. En utilisant une combinaison d'outils de génomique des populations, j'ai montré que la majeure partie de la différenciation entre les formes de couleur à moins de 1 500 m se trouvait sur le chromosome sexuel Z. La sélection sexuelle et possiblement des incompatibilités sont associées au maintien des formes de couleur en parapatrie malgré un flux de gènes important aux loci autosomiques (Figure 2.4). Par ailleurs, la divergence associée à l'altitude a été principalement observée au niveau des loci autosomiques.

J'ai effectué des analyses d'association à l'échelle du génome pour identifier les loci impliqués dans l'adaptation à l'altitude et l'isolement reproducteur entre les différentes formes de couleur, en mettant en évidence un possible excès de loci impliqués dans la reproduction, l'immunité et la réponse au stress (*35*). J'ai identifié *TYRP1*, un gène du chromosome Z, comme un candidat sous-tendant les différences de couleur entre formes parapatriques. Comme ce gène présente peu d'effets pléiotropes et que *Z. borbonicus* est

capable de discriminer les formes de couleur, il est possible que la couleur elle-même constitue un signal pour l'accouplement assortatif.

En conclusion, si l'isolement par l'écologie semble jouer un rôle important dans la séparation des populations de basse et haute altitude, la sélection sexuelle et les incompatibilités pourraient sous-tendre le maintien de formes divergentes même à petite échelle géographique. Ce travail souligne l'importance de passer des espèces modèles et des approches de gènes candidats pour fournir une image complète des bases génétiques de la diversité phénotypique des populations naturelles.

4.2.2 Projet de post-doctorat (2013 - 2016)

Mon premier séjour post-doctoral (2013-2016) s'est effectué sous la supervision du Pr. Dieter Ebert à Bâle. J'ai analysé des données génomiques afin de tester l'existence d'un signal de sélection balancée à un QTL associé à la résistance au parasite *Pasteuria ramosa* chez le crustacé modèle *Daphnia magna*.

J'ai contribué à caractériser les super-allèles à l'origine de la résistance au pathogène en participant aux analyses bioinformatiques et à l'identification des régions orthologues (*53*). Sur la base de ces découvertes, il a été possible de tester directement la sélection balancée dans cette région génomique et de déterminer comment la variabilité génétique de l'hôte est affectée par la sélection parasitaire. J'ai étendu l'étude de l'interaction hôte-parasite aux populations naturelles de *Daphnia magna*.

J'ai démontré que le locus de résistance identifié en laboratoire était également présent dans les populations naturelles. Ce locus présentait des signaux cohérents avec une pression de sélection des parasites spatialement hétérogène, maintenant partiellement le polymorphisme génétique à ce locus (*59*).

J'ai montré que la résistance au parasite est maintenue à une large échelle géographique. En utilisant des données de reséquençage de génome complet et les phénotypes de résistance de 125 individus échantillonnés à travers l'aire de répartition de l'espèce, j'ai démontré que l'interaction entre *D. magna* et *P. ramosa* façonnait la diversité et l'architecture du génome de l'hôte (Figure 2.6). J'ai pu montrer que le locus de résistance présentait des signes clairs de sélection balancée ancienne et d'adaptation locale. Les temps de coalescence sont plus longs à proximité du locus de résistance (Figure 2.6), avec davantage d'allèles partagés entre métapopulations, un résultat cohérent avec l'absence de structure géographique forte des phénotypes de résistance.

4.2.3 Projet de post-doctorat (2016 - 2019)

Mon second post-doc (2016-2019) à l'Université de New York (campus d'Abou Dhabi, Emirats Arabes Unis) s'est quant à lui centré sur la biogéographie d'espèces d'amphibiens et d'oiseaux d'Ethiopie, ainsi qu'à la dynamique populationnelle des éléments transposables (ETs) chez l'anole vert (*Anolis carolinensis*). C'est ce dernier axe de recherche qui m'a conduit à proposer le projet que je mène actuellement à l'IRD.

Les études sur l'adaptation et l'adaptabilité ont énormément bénéficié de l'étude des polymorphismes ponctuels de l'ADN (*SNPs*). Cependant, elles ont négligé les éléments transposables (ETs) (*85*) qui peuvent présenter des effets substantiels sur la taille et l'organisation du génome (*86*). Les ETs sont des fragments d'ADN qui se déplacent et se dupliquent dans les génomes, s'insérant parfois dans ou à proximité des gènes. Ils peuvent avoir des effets neutres, délétères ou avantageux sur la *fitness*. La proportion d'éléments présente dans chaque catégorie demeure néanmoins peu claire. Le génome de l'anole vert contient une extraordinaire diversité d'éléments transposables. J'ai quantifié les fréquences de plusieurs familles d'ETs et utilisé les SNPs comme contraste neutre (*87*). J'ai montré que les insertions d'ETs présentaient un spectre de fréquence d'allèles biaisé en faveur des singletons, qui ne pouvait pas être expliqué par la seule stochasticité démographique, mais était au contraire cohérent avec la sélection purificatrice.

J'ai tiré parti d'une étude de génomique des populations menée en parallèle (*69*) pour quantifier la façon dont la sélection, la démographie et la recombinaison ont façonné la distribution génomique des éléments. J'ai montré que les insertions courtes d'ETs étaient presque neutres, tandis que les insertions plus longues s'avéraient plus délétères en raison de la recombinaison ectopique et de la perturbation des gènes. Je n'ai trouvé aucune preuve claire d'ETs sous sélection positive dans les clades ayant récemment colonisé les climats tempérés depuis la Floride tropicale. J'ai démontré par des simulations que la fréquence et l'abondance des éléments transgéniques présentaient des corrélations distinctes avec le taux de recombinaison. En effet, ce dernier est corrélé à la probabilité de fixation des mutations et des insertions d'éléments transgéniques.

Ces schémas distincts ont permis de déterminer comment la sélection, l'insertion préférentielle d'ETs et les taux de transposition influençaient l'abondance et la fréquence des ETs le long du génome. Par exemple, la sélection directe contre les ETs délétères peut être plus forte dans les régions de forte recombinaison, en raison de la probabilité accrue de recombinaison ectopique. Il en résulte que les éléments transgéniques sont moins abondants et moins fréquents dans les régions de forte recombinaison. D'autre part, les effets de l'interférence de Hill-Robertson et de la sélection liée peuvent conduire à une accumulation d'ETs presque neutres qui sont soit fixés, soit très peu fréquents dans les régions de faible recombinaison.

C'est le même mécanisme qui conduit à la corrélation négative souvent observée entre F_{ST} et d_{XY} chez de nombreuses espèces (*89*) : la sélection de fond élimine le polymorphisme mais augmente le nombre de substitutions.

4.2.4 Projets indépendants (2016 - 2019)

Mon premier poste permanent de *lecturer* (maître de conférences) à l'Université de Portsmouth au Royaume-Uni, obtenu en 2020, m'a permis de développer mon profil d'enseignement et de supervision. Je compte mettre à profit cette expérience afin de développer des ateliers et des cours au sein de mon UMR d'accueil à l'IRD. Durant mon séjour en Angleterre, j'ai initié et mené des recherches sur la biologie de la conservation d'espèces endémiques et/ou menacées au Sud (*Circus maillardi* à La Réunion, tortues grecques en Afrique du Nord). J'ai poursuivi ma recherche sur les éléments transposables en étudiant leur dynamique chez diverses espèces (*Mus musculus*, *Brachypodium distachyon*). J'ai également contribué à l'analyse phylogénétique de l'épidémie de SARS-COV-2 à Portsmouth et en Angleterre, en intégrant le consortium COG-UK. Enfin, j'ai entamé une collaboration avec le Dr. Muriel Gros-Balthazard (UMR DIADE, IRD) sur la génomique des populations de dattiers, ce qui m'a conduit à postuler au concours de CRCN de l'IRD en 2022.

4.3 Recherche présente, et futurs projets

4.3.1 Financement de la recherche

J'ai obtenu une ANR JCJC d'un montant de 426 000 euros, ayant débuté en Janvier 2024. Ceci me permettra de financer mes projets de recherche au cours des quatre prochaines années, ainsi que de proposer un contrat post-doctoral et un contrat de thèse, permettant de développer mon profil de superviseur.

4.3.2 Génomique des populations des éléments transposables

L'interaction entre les éléments transposables et leur hôte est l'un des processus co-évolutifs les plus fascinants trouvés dans la nature. La quantification de l'impact des ETs sur la valeur sélective de leur hôte (*fitness*) possède une valeur appliquée, en permettant d'identifier les bases moléculaires de phénotypes sélectionnés ou d'évaluer le fardeau mutationnel affectant les populations domestiquées ("coût de la domestication").

Le projet sur lequel je me focalise en priorité s'intéresse à l'impact des ETs sur la fitness de l'hôte, en usant du palmier dattier (*Phoenix dactylifera*) comme modèle. La culture du

dattier est d'une importance économique et sociale cruciale au Moyen-Orient et en Afrique du Nord, et sa consommation ne cesse d'augmenter chaque année. Cependant, le dattier est menacé par des sécheresses toujours plus fréquentes et une salinité croissante des sols du fait des changements climatiques en cours. Une caractérisation précise de sa diversité génétique est donc essentielle pour comprendre l'adaptabilité de cette espèce aux pressions environnementales anthropiques. Le projet utilise des données génomiques pour examiner le rôle des ETs polymorphiques dans la sélection et la différenciation des variétés de palmiers dattiers à travers l'aire de répartition de l'espèce.

Deux questions seront examinées :

- Quelle est la distribution des effets sur la fitness des ETs chez les palmiers dattiers ?
- Peut-on en partie prédire cette distribution à partir de données de génomique fonctionnelle telles que les réseaux de régulation de gènes ?

Je développe et applique de nouvelles méthodes en génomique des populations pour estimer la fraction d'ETs avantageux et délétères. Je comparerai ces résultats avec des annotations fonctionnelles de novo de l'organisation de la chromatine et des réseaux de gènes pour dresser une vue d'ensemble de l'impact des ETs sur une espèce pérenne d'intérêt agronomique majeur pour le développement.

Concrètement, dans l'année qui a suivi mon recrutement en janvier 2023, j'ai pu construire une base de données des éléments transposables de type LTR-RTs chez le palmier dattier, entamé l'annotation des éléments dans une ébauche de pan-génomes, en collaboration avec le Professeur Michael Purugganan à NYU et Josep Casacuberta à Barcelone. J'analyse des données Nanopore (40X, N50 de 10kb) sur une vingtaine de variétés de palmiers dattiers qui me permettront d'analyser plus précisément la dynamique populationnelle des ETs. Je contribue également de manière plus large à l'étude de l'histoire évolutive du palmier dattier.

References

(*1*) Bourgeois, Y., Warren, B. H., and Augiron, S. (2024). The burden of anthropogenic changes and mutation load in a critically endangered harrier from the Reunion biodiversity hotspot, *Circus maillardi*. *Molecular Ecology 33*, e17300.

(*2*) Bourgeois, Y. X., and Warren, B. H. (2021). An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Molecular Ecology 30*, ISBN: 0962-1083, 6036–6071.

(*3*) Kojima, S. et al. (2023). Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk. *Nature Genetics 55*, 939–951.

(*4*) Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution 30*, arXiv: 1205.3347 ISBN: 1537-1719 (Electronic)\r0737-4038 (Linking), 1687–1699.

(*5*) Johnson, M. G. et al. (2019). A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Systematic Biology 68*, ed. by Renner, S., 594–606.

(*6*) Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. a., and Johnson, E. a. (2011). Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PloS one 6*, e18561.

(*7*) Ameline, C., Bourgeois, Y., Vögtli, F., Savola, E., Andras, J., Engelstädter, J., and Ebert, D. (2021). A two-locus system with strong epistasis underlies rapid parasite-mediated evolution of host resistance. *Molecular biology and evolution 38*, Publisher: Oxford University Press, 1512–1528.

(*8*) Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics 15*, ISBN: 9783319072111, 356.

(*9*) Fournier, R., Tsangalidou, Z., Reich, D., and Palamara, P. F. (2023). Haplotype-based inference of recent effective population size in modern and ancient DNA samples. *Nature Communications 14*, 7945.

(*10*) François, O., and Jay, F. (2020). Factor analysis of ancient population genomic samples. *Nature Communications 11*, 4661.

(*11*) Schubert, M. et al. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences 111*, 201416991.

(*12*) Seehausen, O. et al. (2014). Genomics and the origin of species. *Nat Rev Genet 15*, Publisher: Nature Publishing Group ISBN: 1471-0064 (Electronic)\r1471-0056 (Linking), 176–192.

(*13*) Kulmuni, J., and Westram, A. M. (2017). Intrinsic incompatibilities evolving as a by-product of divergent ecological selection: Considering them in empirical studies on divergence with gene flow. *Molecular Ecology 26*, ISBN: 4955139574, 3093–3103.

(*14*) Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M. A. F., Mehlig, B., and Westram, A. M. (2017). Interpreting the genomic landscape of speciation: finding barriers to gene flow. *Journal of Evolutionary Biology 30*, 1450–1477.

(*15*) Bourgeois, Y., Roulin, A. C., Müller, K., and Ebert, D. (2017). Parasitism drives host genome evolution: Insights from the *Pasteuria ramosa - Daphnia magna* system. *Evolution*, 1–21.

(*16*) Bourgeois, Y. X. C., Bertrand, J. A. M., Delahaie, B., Holota, H., Thébaud, C., and Milá, B. (2020). Differential divergence in autosomes and sex chromosomes is associated with intra-island diversification at a very small spatial scale in a songbird lineage. *Molecular Ecology 29*, 1137–1153.

(*17*) Bourgeois, Y. X. C., Bertrand, J. A. M., Delahaie, B., Cornuault, J., Duval, T., Milá, B., and Thébaud, C. (2016). Candidate Gene Analysis Suggests Untapped Genetic Complexity in Melanin-Based Pigmentation in Birds. *Journal of Heredity 107*, 327–335.

(*18*) Gill, F. (1973). Intra-island variation in the Mascarene White-eye Zosterops borbonica. *Ornithological Monographs 12*, Publisher: JSTOR.

(*19*) Bertrand, J. A. M., Bourgeois, Y. X. C., Delahaie, B., Duval, T., García-Jiménez, R., Cornuault, J., Heeb, P., Milá, B., Pujol, B., and Thébaud, C. (2014). Extremely reduced dispersal and gene flow in an island bird. *Heredity 112*, ISBN: 1365-2540 Publisher: Nature Publishing Group, 190–196.

(*20*) Milá, B., Warren, B. H., Heeb, P., and Thébaud, C. (2010). The geographic scale of diversification on islands: genetic and morphological divergence at a very small spatial scale in the Mascarene grey white-eye (Aves: Zosterops borbonicus). *BMC evolutionary biology 10*, 158.

(*21*) Bourgeois, Y., Bertrand, J., Thébaud, C., and Milá, B. (2012). Investigating the Role of the Melanocortin-1 Receptor Gene in an Extreme Case of Microgeographical Variation in the Pattern of Melanin-Based Plumage Pigmentation. *PLoS ONE 7*, DOI: 10.1371/journal.pone.0050906.

(*22*) Cheviron, Z., Hackett, S. J., and Brumfield, R. T. (2006). Sequence variation in the coding region of the melanocortin-1 receptor gene (MC1R) is not associated with plumage variation in the blue-crowned manakin (Lepidothrix coronata). *Proceedings. Biological sciences / The Royal Society 273*, 1613–8.

(*23*) Steiner, C. C., Römpler, H., Boettger, L. M., Schöneberg, T., and Hoekstra, H. E. (2009). The genetic basis of phenotypic convergence in beach mice: similar pigment patterns but different genes. *Molecular biology and evolution 26*, 35–45.

(*24*) Uy, J. A. C. et al. (2016). Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. *Proc. R. Soc. B 283*, 2115–2118.

(*25*) Bourgeois, Y. X., Bertrand, J. A., Delahaie, B., Cornuault, J., Duval, T., Milá, B., and Thébaud, C. (2016). Candidate gene analysis suggests untapped genetic complexity in melanin-based pigmentation in birds. *Journal of Heredity 107*, ISBN: 0022-1503 Publisher: Oxford University Press US, 327–335.

(*26*) Ducrest, A.-L., Keller, L., and Roulin, A. (2008). Pleiotropy in the melanocortin system, coloration and behavioural syndromes. *Trends in ecology & evolution 23*, 502–10.

(*27*) Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one 3*, e3376.

(*28*) Bourgeois, Y. X. C., Lhuillier, E., Cézard, T., Bertrand, J. a. M., Delahaie, B., Cornuault, J., Duval, T., Bouchez, O., Milá, B., and Thébaud, C. (2013). Mass production of SNP markers in a nonmodel passerine bird through RAD sequencing and contig mapping to the zebra finch genome. *Molecular Ecology Resources 13*, ISBN: 1755-0998, 899–907.

(*29*) Futschik, A., and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics 186*, 207–18.

(*30*) Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE 6*, e19379.

(*31*) Cutler, D. J., and Jensen, J. D. (2010). To pool, or not to pool? *Genetics 186*, 41–3.

(*32*) Kofler, R., Pandey, R. V., and Schlötterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics 27*, 3435–6.

(*33*) Zhu, Y., Bergland, A. O., González, J., and Petrov, D. a. (2012). Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster. *PloS one 7*, e41901.

(*34*) Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics 201*, 1555–1579.

(*35*) Bourgeois, Y. X., Bertrand, J. A., Delahaie, B., Holota, H., Thébaud, C., and Milá, B. (2020). Differential divergence in autosomes and sex chromosomes is associated with intra-island diversification at a very small spatial scale in a songbird lineage. *Molecular Ecology 29*, ISBN: 0962-1083, 1137–1153.

(*36*) Cornuault, J., Delahaie, B., Bertrand, J. A. M., Bourgeois, Y. X. C., Mila, B., Heeb, P., and Thébaud, C. (2015). Morphological and plumage colour variation in the Réunion grey white-eye (Aves: Zosterops borbonicus): Assessing the role of selection. *Biological Journal of the Linnean Society 114*, 459–473.

(*37*) Goutte, S., Hariyani, I., Utzinger, K. D., Bourgeois, Y., and Boissinot, S. (2022). Genomic Analyses Reveal Association of ASIP with a Recurrently evolving Adaptive Color Pattern in Frogs. *Molecular Biology and Evolution 39*, ISBN: 0737-4038 Publisher: Oxford University Press US, msac235.

(*38*) Manceau, M., Domingues, V. S., Mallarino, R., and Hoekstra, H. E. (2011). The developmental role of Agouti in color pattern evolution. *Science (New York, N.Y.) 331*, 1062–5.

(*39*) Tellier, A., and Brown, J. K. M. (2011). Spatial heterogeneity, frequency-dependent selection and polymorphism in host-parasite interactions. *BMC Evolutionary Biology 11*, Publisher: BioMed Central Ltd, 319.

(*40*) Penczykowski, R. M., Laine, A. L., and Koskella, B. (2016). Understanding the ecology and evolution of host-parasite interactions across scales. *Evolutionary Applications 9*, ISBN: 1752-4571, 37–52.

(*41*) Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics 2*, e64.

(*42*) Tellier, A., Moreno-Gamez, S., and Stephan, W. (2014). Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics. *Evolution 68*, 2211–2224.

(*43*) Bergelson, J., Kreitman, M., Stahl, E. A., and Tian, D. C. (2001). Evolutionary dynamics of plant R-genes. *Science 292*, 2281–2285.

(*44*) Hedrick, P. W. (2006). Genetic Polymorphism in Heterogeneous Environments: The Age of Genomics. *Annual Review of Ecology, Evolution, and Systematics 37*, 67–93.

(*45*) Leffler, E. M., Gao, Z. Y., Pfeifer, S., Segurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J. D., Sella, G., Donnelly, P., McVean, G., and Przeworski, M. (2013). Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science 339*, ISBN: 0036-8075, 1578–1582.

(*46*) Takahata, N., Satta, Y., and Klein, J. (1992). Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics 130*, ISBN: 0016-6731, 925–938.

(*47*) Piertney, S., and Oliver, M. (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity 96*, ISBN: 0018-067X (Print) 0018-067X (Linking), 7–21.

(*48*) Grossen, C. et al. (2014). Introgression from Domestic Goat Generated Variation at the Major Histocompatibility Complex of Alpine Ibex. *PLoS Genetics 10*, DOI: 10.1371/journal.pgen.1004438.

(*49*) Dannemann, M., Andrés, A. M., and Kelso, J. (2016). Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *American Journal of Human Genetics 98*, 22–33.

(*50*) Bechsgaard, J., Jorgensen, T. H., and Schierup, M. H. (2017). Evidence for Adaptive Introgression of Disease Resistance Genes Among Closely Related Arabidopsis Species. *G3 (Bethesda, Md.) 7*, 2677–2683.

(*51*) van Oosterhout, C. (2009). A new theory of MHC evolution: beyond selection on the immune genes. *Proceedings. Biological sciences / The Royal Society 276*, ISBN: 0962-8452 (Print)\r0962-8452 (Linking), 657–665.

(*52*) Obbard, D. J., Welch, J. J., Kim, K.-W., and Jiggins, F. M. (2009). Quantifying adaptive evolution in the Drosophila immune system. *PLoS genetics 5*, e1000698.

(*53*) Bento, G., Routtu, J., Fields, P. D., Bourgeois, Y., Du Pasquier, L., and Ebert, D. (2017). The genetic basis of resistance and matching-allele interactions of a host-parasite system: The Daphnia magna-Pasteuria ramosa model. *PLoS Genetics 13*, ISBN: 1553-7390 Publisher: Public Library of Science San Francisco, CA USA, e1006596.

(*54*) Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics 10*, arXiv: q-bio.PE/1306.5110, DOI: 10.1371/journal.pgen.1004342.

(*55*) Bourgeois, Y., Fields, P. D., Bento, G., and Ebert, D. (2021). Balancing selection for pathogen resistance reveals an intercontinental signature of Red Queen coevolution. *Molecular Biology and Evolution 38*, ISBN: 0737-4038 Publisher: Oxford University Press, 4918–4933.

(*56*) Luijckx, P., Fienberg, H., Duneau, D., and Ebert, D. (2013). A matching-allele model explains host resistance to parasites. *Current Biology 23*, Publisher: Elsevier Ltd, 1085–8.

(*57*) Carius, H. J., Little, T. J., and Ebert, D. (2001). Genetic variation in a host-parasite association: potential for coevolution and frequency-dependent selection. *Evolution 55*, 1136–45.

(*58*) Andras, J. P., and Ebert, D. (2013). A novel approach to parasite population genetics: experimental infection reveals geographic differentiation, recombination and host-mediated population structure in Pasteuria ramosa, a bacterial parasite of Daphnia. *Molecular Ecology 22*, 972–86.

(*59*) Bourgeois, Y., Roulin, A. C., Müller, K., and Ebert, D. (2017). Parasitism drives host genome evolution: insights from the Pasteuria ramosa–Daphnia magna system. *Evolution 71*, ISBN: 1558-5646 Publisher: Blackwell Publishing Inc Malden, USA, 1106–1113.

(*60*) Jostins, L., and McVean, G. (2016). Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. *Bioinformatics 32*, ISBN: 0387954570, 1898–1900.

(*61*) Ameline, C., Bourgeois, Y., Vögtli, F., Savola, E., Andras, J., Engelstädter, J., and Ebert, D. (2020). A Two-Locus System with Strong Epistasis Underlies Rapid Parasite-Mediated Evolution of Host Resistance. *Molecular Biology and Evolution*, DOI: 10.1093/molbev/msaa311.

(*62*) Decaestecker, E., Gaba, S., Raeymaekers, J. a. M., Stoks, R., Van Kerckhoven, L., Ebert, D., and De Meester, L. (2007). Host-parasite 'Red Queen' dynamics archived in pond sediment. *Nature 450*, 870–3.

(*63*) Tollis, M., and Boissinot, S. In *Repetitive DNA*, MA, G.-R., Ed.; Karger: 2012, pp 68–91.

(*64*) Tollis, M., and Boissinot, S. (2014). Genetic Variation in the Green Anole Lizard (Anolis carolinensis) Reveals Island Refugia and a Fragmented Florida During the Quaternary. *Genetica 1*, arXiv: 15334406 ISBN: 0000000000000, 59–72.

(*65*) Manthey, J. D., Tollis, M., Lemmon, A. R., Moriarty Lemmon, E., and Boissinot, S. (2016). Diversification in wild populations of the model organism Anolis carolinensis : A genome-wide phylogeographic investigation. *Ecology and Evolution 6*, 8115–8125.

(*66*) Pavlidis, P., Jensen, J. D., Stephan, W., and Stamatakis, A. (2012). A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution 29*, ISBN: 0737-4038, 3237–3248.

(*67*) Haasl, R. J., and Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology 25*, arXiv: 15334406 ISBN: 1365-294X, 5–23.

(*68*) Bourgeois, Y., Ruggiero, R. P., Manthey, J. D., and Boissinot, S. (2019). Recent secondary contacts, linked selection, and variable recombination rates shape genomic diversity in the model species Anolis carolinensis. *Genome biology and evolution 11*, ISBN: 1759-6653 Publisher: Oxford University Press, 2009–2022.

(*69*) Bourgeois, Y., Ruggiero, R. P., Manthey, J. D., and Boissinot, S. (2019). Recent secondary contacts, linked selection, and variable recombination rates shape genomic diversity in the model species Anolis carolinensis. *Genome biology and evolution 11*, Publisher: Oxford University Press, 2009–2022.

(*70*) Alföldi, J. et al. (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature 477*, ISBN: 1476-4687 (Electronic)\r0028-0836 (Linking), 587–91.

(*71*) Fujita, M. K., Edwards, S. V., and Ponting, C. P. (2011). The *Anolis* lizard genome: An amniote genome without isochores. *Genome Biology and Evolution 3*, ISBN: 1759-6653, 974–984.

(*72*) Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in genetics : TIG 19*, 330–8.

(*73*) Bourgeois, Y., and Boissinot, S. (2019). Selection at behavioural, developmental and metabolic genes is associated with the northward expansion of a successful tropical colonizer. *Molecular ecology 28*, ISBN: 0962-1083, 3523–3543.

(*74*) (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences 205*, 581–598.

(*75*) Soni, V., Johri, P., and Jensen, J. D. (2023). Evaluating power to detect recurrent selective sweeps under increasingly realistic evolutionary null models. *Evolution 77*, ed. by Clark, N., and Chapman, T., 2113–2127.

(*76*) Harris, R. B., Sackman, A., and Jensen, J. D. (2018). On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *bioRxiv*, ISBN: 1111111111, 1–21.

(*77*) Schrider, D. R., Shanku, A. G., and Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics 204*, ISBN: 7065420965, 1207–1223.

(*78*) Kern, A. D., and Hahn, M. W. (2018). The neutral theory in light of natural selection. *Molecular Biology and Evolution 35*, 1366–1371.

(*79*) Jensen, J. D., Payseur, B. A., Stephan, W., Aquadro, C. F., Lynch, M., Charlesworth, D., and Charlesworth, B. (2019). The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution 73*, 111–114.

(*80*) Schrider, D. R., and Kern, A. D. (2016). S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLoS Genetics 12*, ISBN: 10.1371/journal.pgen.1005928, 1–31.

(*81*) Lapiedra, O., Schoener, T. W., Leal, M., Losos, J. B., and Kolbe, J. J. (2018). Predator-driven natural selection on risk-taking behavior in anole lizards. *Science (New York, N.Y.) 360*, 1017–1020.

(*82*) Losos, J. B., Schoener, T. W., and Spiller, D. A. (2004). Predator-induced behaviour shifts and natural selection in field-experimental lizard populations. *Nature 432*, ISBN: 1476-4687 (Electronic)\r0028-0836 (Linking), 505–508.

(*83*) Siewert, K. M., and Voight, B. F. (2017). Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Molecular Biology and Evolution 34*, ISBN: 0737-4038, 2996–3005.

(*84*) Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in zoology 2*, 16.

(*85*) Villanueva-Cañas, J. L., Rech, G. E., de Cara, M. A. R., and González, J. (2017). Beyond SNPs: how to detect selection on transposable element insertions. *Methods in Ecology and Evolution 8*, 728–737.

(*86*) Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., and Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome biology 19*, Publisher: Genome Biology, 199.

(*87*) Ruggiero, R. P., Bourgeois, Y., and Boissinot, S. (2017). LINE insertion polymorphisms are abundant but at low frequencies across populations of Anolis carolinensis. *Frontiers in genetics 8*, ISBN: 1664-8021 Publisher: Frontiers Media SA, 44.

(*88*) Bourgeois, Y., Ruggiero, R. P., Hariyani, I., and Boissinot, S. (2020). Disentangling the determinants of transposable elements dynamics in vertebrate genomes using empirical evidences and simulations. *PLoS genetics 16*, ISBN: 1553-7390 Publisher: Public Library of Science San Francisco, CA USA, e1009082.

(*89*) Cruickshank, T. E., and Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology 23*, ISBN: 1365-294X, 3133–3157.

(*90*) Reyes-Velasco, J., Manthey, J. D., Bourgeois, Y., Freilich, X., and Boissinot, S. (2018). Revisiting the phylogeography, demography and taxonomy of the frog genus Ptychadena in the Ethiopian highlands with the use of genome-wide SNP data. *PLoS One 13*, Publisher: Public Library of Science San Francisco, CA USA, e0190440.

(*91*) Reyes-Velasco, J., Manthey, J. D., Bourgeois, Y., Freilich, X., and Boissinot, S. (2018). Revisiting the phylogeography, demography and taxonomy of the frog genus Ptychadena in the Ethiopian highlands with the use of genome-wide SNP data. *PloS one 13*, ISBN: 1932-6203 Publisher: Public Library of Science San Francisco, CA USA, e0190440.

(*92*) Manthey, J. D., Bourgeois, Y., Meheretu, Y., and Boissinot, S. (2022). Varied diversification patterns and distinct demographic trajectories in Ethiopian montane forest bird (Aves: Passeriformes) populations separated by the Great Rift Valley. *Molecular Ecology 31*, 2664–2678.

(*93*) The COVID-19 Genomics UK (COG-UK) Consortium (2020). An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe 1*, Publisher: The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license, e99–e100.

(*94*) Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology 10*, ISBN: 1553-734X, 1–6.

(*95*) Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences of the United States of America 110*, 228–233.

(*96*) Boissinot, S., and Furano, A. V. (2001). Adaptive Evolution in LINE-1 Retrotransposons. *Molecular Biology and Evolution 18*, 2186–2194.

(*97*) Volz, E., Hill, V., McCrone, J. T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., and Nascimento, F. F. (2021). Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell 184*, ISBN: 0092-8674 Publisher: Elsevier, 64–75. e11.

(*98*) McClintock, B. (1984). The Significance of Responses of the Genome to Challenge. *Science 226*, 792–801.

(*99*) Doolittle, W. F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature 284*, 601–603.

(*100*) Luan, D. D., Korman, M. H., Jakubczak, J. L., and Eickbush, T. H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell 72*, 595–605.

(*101*) Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics 35*, 41–48.

(*102*) Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R. H., De Oliveira, R., Mayer, K. F., Paux, E., and Choulet, F. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology 19*, Publisher: Genome Biology, 1–18.

(*103*) Kent, T. V., Uzunović, J., and Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences 372*, DOI: 10.1098/rstb.2016.0458.

(*104*) Horváth, V., Merenciano, M., and González, J. (2017). Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. *Trends in Genetics 33*, Publisher: Elsevier, 832–841.

(*105*) Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., and Schulman, A. H. (2000). Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proceedings of the National Academy of Sciences 97*, ISBN: 0027-8424 (Print)\r0027-8424 (Linking), 6603–6607.

(*106*) González, J., and Petrov, D. A. (2009). The adaptive role of transposable elements in the Drosophila genome. *Gene 448*, Publisher: Elsevier B.V. ISBN: 0378-1119, 124–133.

(*107*) Rey, O., Danchin, E., Mirouze, M., Loot, C., and Blanchet, S. (2016). Adaptation to Global Change: A Transposable Element-Epigenetics Perspective. *Trends in Ecology and Evolution 31*, Publisher: Elsevier Ltd, 514–526.

(*108*) Jangam, D., Feschotte, C., and Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics 33*, Publisher: Elsevier Ltd ISBN: 0168-9525 (Print) 0168-9525 (Linking), 817–831.

(*109*) Feiner, N. (2016). Accumulation of transposable elements in Hox gene clusters during adaptive radiation of Anolis lizards. *Proceedings. Biological sciences 283*, DOI: 10.1098/rspb.2016.1555.

(*110*) Horvath, R., Minadakis, N., Bourgeois, Y., and Roulin, A. C. (2024). The evolution of transposable elements in Brachypodium distachyon is governed by purifying selection, while neutral and adaptive processes play a minor role. *eLife 12*, DOI: 10.7554/eLife.93284.2.

(*111*) Lockton, S., Ross-Ibarra, J., and Gaut, B. S. (2008). Demography and weak selection drive patterns of transposable element diversity in natural populations of Arabidopsis lyrata. *Proc Natl Acad Sci U S A 105*, ISBN: 1091-6490 (Electronic), 13965–13970.

(*112*) Mérel, V., Gibert, P., Buch, I., Rodriguez Rada, V., Estoup, A., Gautier, M., Fablet, M., Boulesteix, M., and Vieira, C. (2021). The Worldwide Invasion of Drosophila suzukii Is Accompanied by a Large Increase of Transposable Element Load and a Small Number of Putatively Adaptive Insertions. *Molecular Biology and Evolution 38*, 4252–4267.

(*113*) Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics 28*, ISBN: 1367-4811 (Electronic)\r1367-4803 (Linking), 333–339.

(*114*) Charlesworth, B., and Charlesworth, D. (1983). The Population Genetics of Transposable Elements. *Genetic Research 42*, Publisher: NYU School of Medicine, 1–27.

(*115*) Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature 371*, ISBN: 0028-0836 (Print), 215–220.

(*116*) Bourgeois, Y., and Boissinot, S. (2019). On the population dynamics of junk: a review on the population genomics of transposable elements. *Genes 10*, ISBN: 2073-4425 Publisher: MDPI, 419.

(*117*) Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., Mills, R. E., 1000 Genomes Project Consortium, 1. G. P., and Devine, S. E. (2017). The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome research*, gr.218032.116.

(*118*) Burke, D., Chuong, E., Taylor, W., and Layer, R. TEPEAK : A novel method for identifying and characterizing polymorphic transposable elements in non-model species populations, en, 2023.

(*119*) Hosseini, M., Palmer, A., Manka, W., Grady, P. G. S., Patchigolla, V., Bi, J., O'Neill, R. J., Chi, Z., and Aguiar, D. (2023). Deep statistical modelling of nanopore sequencing translocation times reveals latent non-B DNA structures. *Bioinformatics 39*, i242–i251.

(*120*) Mohamed, M., Sabot, F., Varoqui, M., Mugat, B., Audouin, K., Pélisson, A., Fiston-Lavier, A.-S., and Chambeyron, S. (2023). TrEMOLO: accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches. *Genome Biology 24*, 63.

(*121*) Groza, C., Chen, X., Wheeler, T. J., Bourque, G., and Goubert, C. GraffiTE: a Unified Framework to Analyze Transposable Element Insertion Polymorphisms using Genome-graphs, en, preprint, Bioinformatics, 2023.

(*122*) Sierra, P., and Durbin, R. Identification of transposable element families from pangenome polymorphisms, en, 2024.

(*123*) Arkhipova, I. R. (2018). Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. *Molecular biology and evolution 35*, ISBN: 1537-1719 (Electronic) 0737-4038 (Linking), 1332–1337.

(*124*) Lynch, M., and Conery, J. S. (2003). The Origins of Genome Complexity. *Science 302*, arXiv: 1011.1669v3 ISBN: 1095-9203 (Electronic) 0036-8075 (Linking), 1401–1404.

(*125*) Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics 23*, 263–286.

(*126*) Marino, A., Debaecker, G., Fiston-Lavier, A.-S., Haudry, A., and Nabholz, B. Effective population size does not explain long-term variation in genome size and transposable element content in animals, en, 2024.

(*127*) Daubin, V., and Moran, N. A. (2004). Comment on "The Origins of Genome Complexity". *Science 306*, 978–978.

(*128*) Schley, R. J., Pellicer, J., Ge, X. J., Barrett, C., Bellot, S., Guignard, M. S., Novák, P., Suda, J., Fraser, D., Baker, W. J., Dodsworth, S., Macas, J., Leitch, A. R., and Leitch, I. J. (2022). The ecology of palm genomes: repeat-associated genome size expansion is constrained by aridity. *New Phytologist 236*, 433–446.

(*129*) Organ, C. L., and Shedlock, A. M. (2009). Palaeogenomics of pterosaurs and the evolution of small genome size in flying vertebrates. *Biology Letters 5*, 47–50.

(*130*) Montgomery, E., Charlesworth, B., and Langley, C. (1987). A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster. *Genet Res. 49*, 31–41.

(*131*) Petrov, D. A., Aminetzach, Y. T., Davis, J. C., Bensasson, D., and Hirsh, A. E. (2003). Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. *Molecular Biology and Evolution 20*, ISBN: 0737-4038, 880–892.

(*132*) Boissinot, S., Davis, J., Entezam, A., Petrov, D., and Furano, A. V. (2006). Fitness cost of LINE-1 (L1) activity in humans. *Proceedings of the National Academy of Sciences 103*, ISBN: 0027-8424, 9590–9594.

(*133*) Hazzouri, K. M., Mohajer, A., Dejak, S. I., Otto, S. P., and Wright, S. I. (2008). Contrasting patterns of transposable-element insertion polymorphism and nucleotide diversity in autotetraploid and allotetraploid Arabidopsis species. *Genetics 179*, ISBN: 0016-6731, 581–592.

(*134*) Xue, A. T., Ruggiero, R. P., Hickerson, M. J., and Boissinot, S. (2018). Differential effect of selection against LINE retrotransposons among vertebrates inferred from whole-genome data and demographic modeling. *Genome Biology and Evolution 10*, 1265–1281.

(*135*) Eyre-Walker, A., Woolfit, M., and Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics 173*, 891–900.

(*136*) Dai, X., Wang, H., Zhou, H., Wang, L., Dvořák, J., Bennetzen, J. L., and Müller, H.-G. (2018). Birth and Death of LTR-Retrotransposons in *Aegilops tauschii*. *Genetics 210*, 1039–1051.

(*137*) Blumenstiel, J. P., Chen, X., He, M., and Bergman, C. M. (2014). An age-of-allele test of neutrality for transposable element insertions. *Genetics 196*, arXiv: 1209.3456 ISBN: 1943-2631 (Electronic)\n0016-6731 (Linking), 523–538.

(*138*) Albers, P. K., and McVean, G. (2020). Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biology 18*, ISBN: 1111111111, 1–26.

(*139*) Platt, A., Pivirotto, A., Knoblauch, J., and Hey, J. (2019). An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations. *PLoS Genetics 15*, ISBN: 1111111111, 1–25.

(*140*) Ortega-Del Vecchyo, D., Lohmueller, K. E., and Novembre, J. (2022). Haplotype-based inference of the distribution of fitness effects. *Genetics 220*, DOI: 10.1093/genetics/iyac002.

(*141*) Campos-Sánchez, R., Cremona, M. A., Pini, A., Chiaromonte, F., and Makova, K. D. (2016). Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. *PLoS Computational Biology 12*, 1–41.

(*142*) Makova, K. D., and Weissensteiner, M. H. (2023). Noncanonical DNA structures are drivers of genome evolution. *Trends in Genetics 39*, 109–124.

(*143*) Guiblet, W. M., Cremona, M. A., Cechova, M., Harris, R. S., Kejnovská, I., Kejnovsky, E., Eckert, K., Chiaromonte, F., and Makova, K. D. (2018). Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Research 28*, 1767–1778.

(*144*) Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in ecology & evolution 25*, 410–8.

(*145*) Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics 35*, arXiv: 1605.05537, 1720–1728.

(*146*) Sheehan, S., and Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLoS Computational Biology 12*, arXiv: 028175 ISBN: 1553-7358 (Electronic)\r1553-734X (Linking), 1–28.

(*147*) Schrider, D. R., and Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics 34*, 301–312.

(*148*) Haller, B. C., and Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Molecular Biology and Evolution 36*, 632–637.

(*149*) Flowers, J. M., Hazzouri, K. M., Gros-Balthazard, M., Mo, Z., Koutroumpa, K., Perrakis, A., Ferrand, S., Khierallah, H. S., Fuller, D. Q., Aberlenc, F., Fournaraki, C., and Purugganan, M. D. (2019). Cross-species hybridization and the origin of North African date palms. *Proceedings of the National Academy of Sciences of the United States of America 116*, 1651–1658.

(*150*) Swarup, S., Glenn, K. C., Cargill, E. J., Crosby, K., Flagel, L., and Kniskern, J. (2021). Genetic diversity is indispensable for plant breeding to improve crops. 839–852.

(*151*) Glaszmann, J. C., Kilian, B., Upadhyaya, H. D., and Varshney, R. K. (2010). Accessing genetic diversity for crop improvement. *Current Opinion in Plant Biology 13*, Publisher: Elsevier Ltd, 167–173.

(*152*) Gaut, B. S., Seymour, D. K., Liu, Q., and Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication. *Nature Plants 4*, Publisher: Springer US, 512–520.

(*153*) Turner-Hissong, S. D., Mabry, M. E., Beissinger, T. M., Ross-Ibarra, J., and Pires, J. C. (2020). Evolutionary insights into plant breeding. *Current Opinion in Plant Biology 54*, Publisher: Elsevier Ltd, 93–100.

(*154*) Niu, C. et al. (2022). Methylation of a MITE insertion in the MdRFNR1-1 promoter is positively associated with its allelic expression in apple in response to drought stress. *The Plant cell 34*, 3983–4006.

(*155*) Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics 17*, 754–755.

(*156*) Wicker, T., Sabot, F., Hua-van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., Sanmiguel, P., and Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics 8*, 973–982.

(*157*) Gaut, B. S., Díez, C. M., and Morrell, P. L. (2015). Genomics and the Contrasting Dynamics of Annual and Perennial Domestication. *Trends in Genetics 31*, Publisher: Elsevier Ltd, 709–719.

(*158*) Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., and Martin, C. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell 24*, 1242–1255.

(*159*) Gaut, B., Yang, L., Takuno, S., and Eguiarte, L. E. (2011). The Patterns and Causes of Variation in Plant Nucleotide Substitution Rates. *Annual Review of Ecology, Evolution, and Systematics 42*, 245–266.

(*160*) Hazzouri, K. M. et al. (2019). Genome-wide association mapping of date palm fruit traits. *Nature Communications 10*, Publisher: Springer US, 1–14.

(*161*) Orozco-Arias, S., Humberto Lopez-Murillo, L., Candamil-Cortés, M. S., Arias, M., Jaimes, P. A., Rossi Paschoal, A., Tabares-Soto, R., Isaza, G., and Guyot, R. (2023). Inpactor2: a software based on deep learning to identify and classify LTR-retrotransposons in plant genomes. *Briefings in Bioinformatics 24*, bbac511.

(*162*) Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., and Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America 117*, 9451–9457.

(*163*) Huang, X., Fortier, A. L., Coffman, A. J., Struck, T. J., Irby, M. N., James, J. E., León-Burguete, J. E., Ragsdale, A. P., and Gutenkunst, R. N. (2021). Inferring Genome-Wide Correlations of Mutation Fitness Effects between Populations. *Molecular Biology and Evolution 38*, 4588–4602.

(*164*) Thieme, M., Brêchet, A., Bourgeois, Y., Keller, B., Bucher, E., and Roulin, A. C. (2022). Experimentally heat-induced transposition increases drought tolerance in Arabidopsis thaliana. *New Phytologist 236*, ISBN: 0028-646X, 182–194.

(*165*) Wollenberg Valero, K. C. (2024). Brief Communication: The Predictable Network Topology of Evolutionary Genomic Constraint. *Molecular Biology and Evolution 41*, ed. by Teeling, E., msae033.

(*166*) Fagny, M., and Austerlitz, F. (2021). Polygenic Adaptation: Integrating Population Genetics and Gene Regulatory Networks. *Trends in Genetics 37*, Publisher: The Authors, 631–638.

(*167*) Mayr, E. (1961). Cause and effect in biology. *Science 134*, ISBN: 0036807500368075, 1501–1506.

(*168*) Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. a., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution 5*, 578–90.

(*169*) Hohenlohe, P. a., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. a., and Cresko, W. a. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS genetics 6*, e1000862.

Appendix A

Selected peer-reviewed publications

Received: 14 September 2019 | Revised: 12 February 2020 | Accepted: 20 February 2020

DOI: 10.1111/mec.15396

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Differential divergence in autosomes and sex chromosomes is associated with intra-island diversification at a very small spatial scale in a songbird lineage

Yann X. C. Bourgeois[1,2] | Joris A. M. Bertrand[2,3] | Boris Delahaie[2,4] | Hélène Holota[2] | Christophe Thébaud[2] | Borja Milá[5]

[1]School of Biological Sciences, University of Portsmouth, Portsmouth, UK

[2]Laboratoire Évolution et Diversité Biologique (EDB), UMR 5174 Centre National de la Recherche Scientifique (CNRS), Institut de Recherche pour le Développement (IRD), Université Paul Sabatier, Toulouse, France

[3]Laboratoire Génome & Développement des Plantes, UMR 5096, Université de Perpignan Via Domitia, Perpignan, France

[4]Department of Plant Sciences, University of Cambridge, Cambridge, UK

[5]National Museum of Natural Sciences, Spanish National Research Council (CSIC), Madrid, Spain

Correspondence
Yann X. C. Bourgeois, King Henri Building Room 6.14, School of Biological Sciences, University of Portsmouth, Portsmouth PO1 2DY, UK.
Email: yann.x.c.bourgeois@gmail.com

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-10-LABX-41 and ANR-2006-BDIV002; Fondation pour la Recherche sur la Biodiversité (FRB); Centre National de la Recherche Scientifique; National Geographic Society

Abstract

Recently diverged taxa showing marked phenotypic and ecological diversity provide optimal systems to understand the genetic processes underlying speciation. We used genome-wide markers to investigate the diversification of the Reunion grey white-eye (*Zosterops borbonicus*) on the small volcanic island of Reunion (Mascarene archipelago), where this species complex exhibits four geographical forms that are parapatrically distributed across the island and differ strikingly in plumage colour. One form restricted to the highlands is separated by a steep ecological gradient from three distinct lowland forms which meet at narrow hybrid zones that are not associated with environmental variables. Analyses of genomic variation based on single nucleotide polymorphism data from genotyping-by-sequencing and pooled RAD-seq approaches show that signatures of selection associated with elevation can be found at multiple regions across the genome, whereas most loci associated with the lowland forms are located on the Z sex chromosome. We identified *TYRP1*, a Z-linked colour gene, as a likely candidate locus underlying colour variation among lowland forms. Tests of demographic models revealed that highland and lowland forms diverged in the presence of gene flow, and divergence has progressed as gene flow was restricted by selection at loci across the genome. This system holds promise for investigating how adaptation and reproductive isolation shape the genomic landscape of divergence at multiple stages of the speciation process.

KEYWORDS

genomics, natural selection, plumage colour evolution, population differentiation, sex chromosome, speciation, *Zosterops*

1 | INTRODUCTION

As populations and lineages diverge from each other, a progressive loss of shared polymorphisms and accumulation of fixed alleles is expected. This is influenced by neutral processes (e.g., genetic drift), but also by natural and sexual selection, and the interaction between these processes may vary between different parts of the genome, creating a mosaic pattern of regions displaying different rates of divergence (Nosil, Harmon, & Seehausen, 2009; Wu, 2001). However, genomic regions directly involved in local adaptation and

Thébaud and Milá authors contributed equally to this work.

reproductive isolation may experience reduced effective gene flow compared to the genomic background (Ravinet et al., 2017). In addition, the effects of selection at linked sites can also locally increase divergence and magnify the effects of nonequilibrium demography over large genomic regions (Van Belleghem et al., 2018; Burri, 2017; Burri et al., 2015; Cruickshank & Hahn, 2014). Thus, establishing how different processes such as drift, selection and gene flow shape the rates of divergence at the genomic scale is critical to understand the links between speciation processes and their genetic and genomic consequences (Gavrilets, 2014).

Identifying the main drivers of genome-wide differentiation (i.e., isolation by environment versus reproductive isolation driven by nonadaptive factors) remains a complex question (Cruickshank & Hahn, 2014; Ravinet et al., 2017; Wolf & Ellegren, 2016). Recent studies have displayed varied results, and those focusing on the early stages of speciation have often emphasized ecological divergence over sexual selection and intrinsic incompatibilities (Bierne, Welch, Loire, Bonhomme, & David, 2011; Seehausen et al., 2014). In this context, studies of closely related taxa or populations that show phenotypic and ecological diversity, and are at different stages of divergence, hold promise to help clarify the chronology and relative importance of these underlying evolutionary mechanisms (Delmore et al., 2015; Mořkovský et al., 2018; Pryke, 2010; Sætre & Sæether, 2010; Safran, Scordato, Symes, Rodríguez, & Mendelson, 2013; Seehausen et al., 2014).

We used the Reunion grey white-eye (*Zosterops borbonicus*; taxonomy following Gill & Donsker, 2019), a songbird endemic to the small (2,512 km^2) volcanic ocean island of Reunion (Mascarene archipelago, southwestern Indian Ocean), to quantify genome-wide patterns of divergence across its range and better understand underlying evolutionary factors. This species is characterized by complex patterns of plumage colour and size variation, with five distinct variants recognized across the island (Gill, 1973). These variants can be grouped into four parapatrically distributed geographical forms with adjoining ranges that came into secondary contact after diverging in allopatry (Bertrand et al., 2016; Cornuault et al., 2015; Delahaie et al., 2017). Three lowland forms differ primarily in plumage colour (Cornuault et al., 2015; Gill, 1973) and show a unique distribution pattern, with each form being separated from the other two by narrow physical barriers such as rivers or lava fields (Gill, 1973). These forms differ strikingly in head coloration and include a light brown form (lowland brown-headed brown form; hereafter LBHB), a grey-headed brown form (GHB) with a brown back and a grey head, and a brown-naped brown form (BNB) with a brown back and nape, and a grey crown (Figure 1 and Figure S1; see Cornuault et al., 2015 for a detailed description). A fourth form, restricted to the highlands (between 1,400 and 3,000 m), is relatively larger than the lowland forms and comprises two very distinct colour morphs, with birds showing predominantly grey (GRY) or brown (highland brown-headed brown form, HBHB) plumage, respectively (Bertrand et al., 2016; Cornuault et al., 2015; Gill, 1973; Milá, Warren, Heeb, & Thébaud, 2010). Both of these latter morphs occur in sympatry and represent a clear case of plumage colour polymorphism (Bourgeois et al., 2017). This highland form is separated from all three lowland forms by relatively narrow contact zones located along the elevational gradient (Gill, 1973). One such contact zone was recently studied and was found to correspond to an ecotone between native habitat (>1,400 m above sea level [a.s.l.]) and anthropogenic landscapes (<1,400 m a.s.l.), suggesting a possible role of environmental differences in influencing the location of these zones (Bertrand et al., 2016). While plumage colour differences between the two apparently similar all-brown variants (LBHB and HBHB) may appear subtle, they are in fact significant when considering bird vision and using a visual model to project these colours in an avian-appropriate, tetrachromatic colour space (Cornuault et al., 2015). Patterns of coloration among forms and morphs are stable over time, with no apparent sex effect (see Gill, 1973; Milá et al., 2010).

Recent studies have revealed that dispersal and gene flow must be limited in the Reunion grey white-eye, with low levels of historical and/or contemporary gene flow among populations, unless they are very close geographically (<10 km) (Bertrand et al., 2014), and that more variation exists among the different geographical forms than would be expected under drift for both morphological and plumage colour traits (Cornuault et al., 2015). Thus, it is the combination of reduced dispersal and divergent selection that seems to explain why white-eyes were able to differentiate into multiple geographical forms within Reunion, as originally proposed by Gill (1973). The island has a dramatic topography (maximum elevation: 3,070 m), and a steep elevational gradient was found to be associated with strong divergent selection on phenotypes and marked genetic structure for autosomal microsatellites, a pattern that is consistent with isolation by ecology between lowland and highland forms (Bertrand et al., 2016). In contrast, lowland forms show no association with neutral genetic (microsatellite) structure or major changes in vegetation characteristics and associated climatic variables and are separated by very narrow hybrid zones centred on physical barriers to gene flow (Delahaie et al., 2017). Although the autosomal markers used could not provide information on sex-linked loci, these patterns of genetic differentiation suggest that while a sharp ecological transition between lowlands and highlands could drive differentiation at many autosomal loci through local adaptation, phenotypic divergence between lowland forms involves either fewer loci or loci concentrated in a narrower genomic region not covered by microsatellites (Delahaie et al., 2017).

In this work, we aim to: (i) identify the genomic variation associated with phenotypic differentiation between forms in relation to ecological variation (low versus high elevation) and divergence in signalling traits (conspicuous variation in plumage colour between lowland forms in the absence of abrupt ecological transitions); (ii) determine whether divergence peaks are found on autosomes or sex chromosomes, respectively; and (iii) identify potential candidate genes associated with divergent genomic regions. We used individual genotyping by sequencing (GBS) (Elshire et al., 2011) to characterize the amount of divergence between forms. We further used

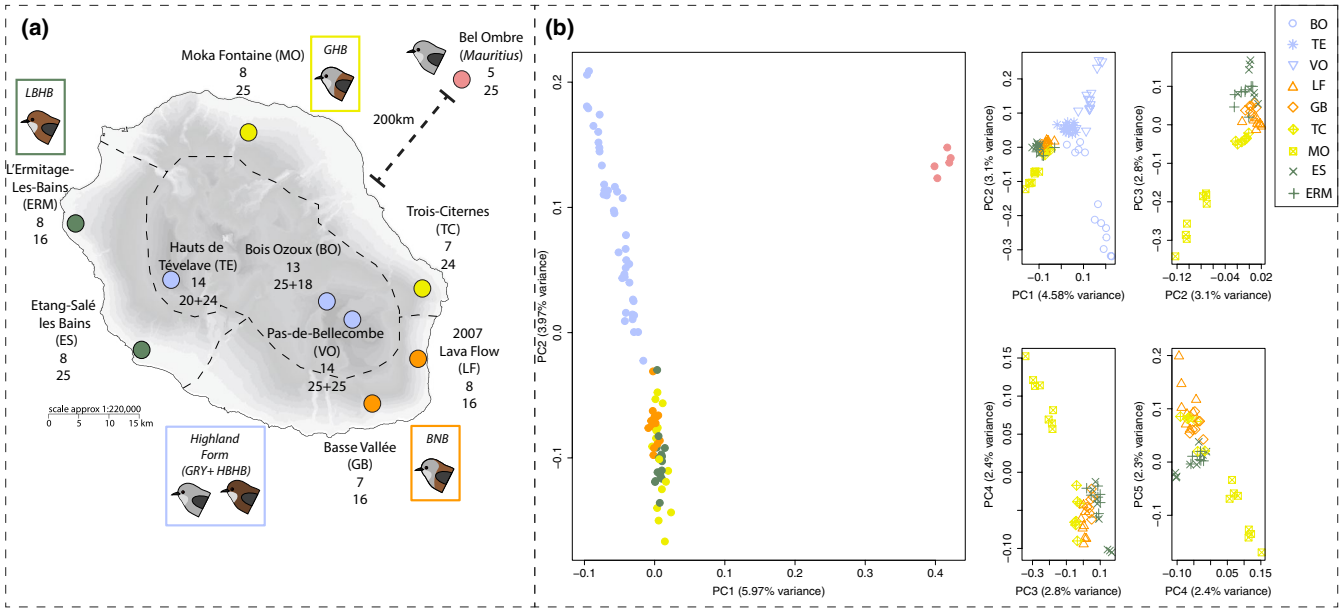


FIGURE 1 Map of Reunion showing *Zosterops* localities sampled in this study and a description of population structure using principal components analysis (PCA) on autosomal GBS data. (a) For each locality, the sample size for the individual GBS data set is followed by the sample size for each pool. In the three localities found at high elevation, where populations are polymorphic, the size of pools is given for grey + brown individuals. Distribution limits between the different geographical forms are indicated by dashed lines. (b) PCA results including *Zosterops mauritianus* (left panel) or not (right panels). Points corresponding to high-elevation individuals were removed on the last three panels for the sake of clarity [Colour figure can be viewed at wileyonlinelibrary.com]

a pooled RAD-sequencing (RAD-seq) (Baird et al., 2008) approach that produced a high density of markers to characterize with greater precision the genomic landscape of divergence and assess the extent of differentiation between the different colour forms. Finally, to test whether incomplete lineage sorting or gene flow explained shared genetic variation among forms, we use coalescent models to test alternative demographic scenarios of divergence, including models with different temporal patterns of gene flow and effective population size changes over time.

2 | MATERIAL AND METHODS

2.1 | Field sampling

We sampled a total of 259 Reunion grey white-eyes between 2007 and 2012 from nine locations that were chosen to extensively cover the species' range and the different geographical forms. We also sampled 25 Mauritius grey white-eyes (*Zosterops mauritianus*) from a single location on Mauritius to be used as an outgroup in some of our analyses as this species and the Reunion grey white-eye are sister taxa (Warren, Bermingham, Prys-Jones, & Thébaud, 2006). Birds were captured using mist-nets, marked with a uniquely numbered aluminium ring, and ~50 µl of blood was collected from each individual and preserved in Queen's lysis buffer (Seutin, White, & Boag, 1991). All manipulations were conducted under a research permit issued by the Centre de Recherches sur la Biologie des Populations d'Oiseaux (CRBPO) – Muséum National d'Histoire Naturelle (Paris). Individuals were sexed using PCR (polymerase chain reaction) (Griffiths, Double, Orr, & Dawson, 1998) to infer the number of distinct Z chromosomes included in each genetic pool. We included 152 females and 132 males in this study, among which 47 females and 48 males were included in the GBS experiment (see below).

2.2 | GBS using individual DNA samples

We performed GBS (Elshire et al., 2011) on 95 individuals, including 90 Reunion and five Mauritius grey white-eyes (Figure 1). We included 7–14 individuals from each Reunion location and two locations per geographical form; such a sampling scheme should be sufficient to retrieve patterns of differentiation and diversity at the scale of forms, as highlighted by both theoretical (Willing, Dreyer, & Oosterhout, 2012) and empirical studies (Jeffries et al., 2016; Nazareno, Bemmels, Dick, & Lohmann, 2017). GBS is similar to RAD-seq, but involves fewer preparation steps (Elshire et al., 2011) and it samples loci at a lower resolution across the genome. Approximately one microgram of DNA was extracted with a Qiagen DNeasy Blood & Tissue kit following the manufacturer's instructions and sent to the BRC Genomic Diversity Facility at Cornell University (see Elshire et al., 2011) for single-end sequencing on a single lane of an Illumina HiSeq2000 device after digestion with *Pst*I. Read length was 100 bp. Three individuals had to be removed from subsequent analyses due to the extremely low number of reads obtained (Table S1). Raw reads were trimmed with TRIMMOMATIC (version 0.33; Bolger, Lohse, & Usadel, 2014) with a minimum

base quality of 20. We used the recently assembled *Zosterops lateralis* genome (Cornetti et al., 2015) to map reads back onto this reference with BWA MEM (version 0.7.12; Li & Durbin, 2009) and SAMTOOLS (version 1.3.1; Li et al., 2009), instead of creating consensuses directly from the data as in Bourgeois et al. (2013). Reads with a mapping score below 20 were excluded (samtools view -q 20). We then aligned contigs and scaffolds from a congeneric white-eye species, *Z. lateralis*, on the zebra finch (*Taeniopygia guttata*) passerine reference genome (version July 2008, assembly WUGSC version 3.2.4) using LASTZ (version 1.03.54; Harris, 2007; Schwartz, Kent, & Smit, 2003). We used the following options and thresholds: --masking = 254 --hspthresh = 4,500 --gappedthresh = 3,000. The first option means that any locus found mapping more than 254 times is automatically masked and does not appear in the final pairwise alignment. The --hspthresh parameter is an option that excludes any alignment with a score lower than 4,500 during the gap-free extension stage. The --gappedthresh option controls the maximum size of the gaps allowed to join best local alignments; the higher the score, the fewer gaps are allowed. We used the same set of options previously used in comparisons between other related bird genomes, such as chicken and grouse (e.g., Kozma, Melsted, Magnússon, & Höglund, 2016). Scaffolds were then assigned to chromosomal regions based on their alignment scores. We note that synteny is well conserved in birds (Derjusheva, Kurganova, Habermann, & Gaginskaya, 2004), and that misalignment is therefore unlikely to constitute a major source of errors.

SNPs were called using FREEBAYES (version 0.9.15–1; Garrison & Marth, 2012) and filtered with VCFTOOLS (version 0.1.12b) using the following criteria for autosomal markers: (a) a sequencing depth between 8 × and 100 × for each individual genotype; (b) a minimal genotype quality of 20; and (c) no more than nine missing genotypes. Missing data per individual before filtering and after removing individuals with low read count was at most 50% (average 30%, *SD* = 5%, see Table S1). Average sequencing depth was 9.75× (*SD* = 3.05). The range of sequencing depth for filtering was chosen based on visual examination of histograms produced by SAMTOOLS (option depth), to remove loci with a clear excess of mapping reads that may indicate repetitive sequences and those loci with very low depth for which genotypes may not be called confidently, while retaining enough information for inference. For Z-linked markers, we first listed scaffolds mapping on the Zebra finch's Z chromosome based on LASTZ alignments. We then used VCFTOOLS to extract genotypes found on Z scaffolds (providing a list of these scaffolds with the option --bed). SNPs were filtered in male individuals only, using the same criteria as for autosomes, except that no more than five missing genotypes were allowed. We then extracted these sites in females only, using VCFTOOLS (option --positions), and removed markers displaying more than three heterozygous females, allowing for some tolerance because FREEBAYES attempts to balance the count of heterozygotes in a diploid population. This led to the removal of 171 sites out of 1,136. We then recalled SNPs with FREEBAYES in females only, assuming haploidy (option --ploidy 1). The constraint on sequencing depth and genotype quality was removed in females to consider the fact that a single Z copy is found in these individuals, therefore reducing depth of coverage at Z-linked markers. Because we excluded reads with a mapping quality below 20 when creating BAM files, we considered that a single read was enough to call a site in females for the Z scaffold. This decision was taken to maximize the number of markers available for this chromosome. The final data set consisted of 34,951 autosomal markers and 965 Z-linked markers. Recent studies have suggested the existence of a neo sex chromosome in Sylvioidea, consisting of a fusion between ancestral Z and W chromosomes with the first 10 Mb of the zebra finch's chromosome 4A (Pala et al., 2012). We therefore excluded this region from our analyses and studied it separately, focusing on males only. In males, 917 and 730 SNPs called by FREEBAYES were found polymorphic on the Z and the 4A sex-linked fragment, respectively.

2.3 | Pooled RAD-seq

To identify loci and genomic regions associated with ecological variation (low versus. high elevation) and divergence plumage colour between lowland forms, we used a paired-end RAD-seq protocol, using a data set partially described elsewhere (Bourgeois et al., 2013) in which six pools of 20–25 individuals from the same three locations as those sampled for the high-elevation form in the GBS experiment were sequenced. We added seven more pools of 16–25 individuals from the lowland forms to cover the same localities as the GBS data set (Figure 1; Table S2). This protocol was used because it produced a higher density of markers along the genome relative to the GBS approach described above, thus increasing the ability to detect outlier genomic regions. This approach resulted in ~ 600,000 contigs with an average size of 400 bp, covering about 20% of the genome (Bourgeois et al., 2013). The larger number of individuals included in each pool should also increase the ability to detect shifts in allele frequencies between populations. We modified the bioinformatics protocol used in Bourgeois et al. (2013) by mapping the reads on the *Z. lateralis* genome using BWA MEM instead of creating contigs from the RAD-seq reads. PCR duplicates were removed using SAMTOOLS (Li et al., 2009). SNPs were called using POPOOLATION2 (version 1.201; Kofler, Pandey, & Schlötterer, 2011), using a minimal allele count of two across all pools, and a depth between 10 × and 300 × for each pool to remove loci that were clear outliers for sequencing depth while keeping a high density of markers along the genome. We used BEDTOOLS (version 2.25.0; Quinlan & Hall, 2010) to estimate the proportion of sites covered at a depth between 10 × and 300 × in each pool (option genomecov). Overall, more than 1,104,000 SNPs for autosomes and 42,607 SNPs for the Z chromosome were obtained, covering between 12% and 18% of the genome (Table S2). We accounted for the unequal number of alleles between autosomal markers and Z-linked markers in all subsequent analyses.

2.4 | Genetic structure

To assess population genetic structure within and between geographical forms, we first performed a principal components analysis

(PCA; Patterson, Price, & Reich, 2006) on all GBS autosomal markers, using the Bioconductor package SEQVARTOOLS (version 1.24.0; Huber et al., 2015), excluding markers with a minimal allele frequency below 0.05. We then evaluated population structure for both autosomal and Z-linked markers using the software ADMIXTURE (version 1.3.0; Alexander & Novembre, 2009). This software is a fast and efficient tool for estimating individual ancestry coefficients. It does not require any a-priori grouping of individuals by locality but requires defining the expected number (*K*) of clusters to which individuals can be assigned. Importantly, ADMIXTURE allows us to specify which scaffolds belong to sex chromosomes, and corrects for heterogamy between males and females. "Best" values for *K* were assessed using a cross-validation (CV) procedure using 10 CV replicates. In this context, CV consists in masking alternately one-fifth of the data set, then using the remaining data set to predict the masked genotypes. Predictions are then compared with actual observations to infer prediction errors. This procedure is therefore sensitive to heterogeneity in structure across markers induced by, for example, selection. Therefore, we present results for all values of *K* as they may reveal subtle structure supported by only a subset of markers under selection. Based on patterns of linkage disequilibrium (LD)-decay, we thinned the data set to limit the effects of linkage, with a minimal distance between two adjacent markers of 1,000 bp (Figure S2). Pairwise LD (measured as r^2, which does not require phasing) between all pairs of markers was computed in VCFTOOLS (version 0.1.12b).

To further explore whether changes in SNP caller and the number of markers could affect the ADMIXTURE analysis and observed differences between autosomes and Z-linked markers, we called Z-linked SNPs using ANGSD (version 0.923; Korneliussen, Albrechtsen, & Nielsen, 2014), following an approach similar to that used with FREEBAYES. We first called SNPs in all individuals assuming diploidy, using a uniform prior based on allele frequencies but not assuming Hardy–Weinberg equilibrium in samples (option -doPost 2). SNP likelihoods were computed following the model implemented in SAMTOOLS (-GL = 1). We then called SNPs in females only, assuming haploid markers and calling the consensus base (option -doHaploCall 2). We filtered reads so they mapped to a single site in the genome (-uniqueOnly 1 -remove_bads 1), had a mapping quality of at least 20 (-minMapQ 20) and a minimum read quality of 20 (-minQ 20), and were covered in at least two-thirds of individuals with a minimum individual depth of 6 × in males (-geno_minDepth 6). We corrected for excessive mismatch with the reference and excess of SNPs with indels (-C 50 -baq 1). This resulted in 2,282 Z-linked SNPs.

To assess the relative proportion of genetic variance contributing to the differentiation of geographical forms (estimated by F_{CT}) while taking into account population substructure (F_{SC} and F_{ST}), we conducted an analysis of molecular variance (AMOVA) in ARLEQUIN version 3.5 (Excoffier & Lischer, 2010) using as groups either islands (Reunion versus Mauritius), lowlands and highlands, or the forms themselves. *F*-statistics for the whole data set are weighted averages. Significance was assessed with 1,000 permutations.

We assessed relationships between populations from pooled data using POPTREE2 (Takezaki, Nei, & Tamura, 2010) to compute F_{ST} matrices across populations using allele frequencies. A neighbor-joining tree was then estimated from these matrices. We included 20,000 random SNPs with a minimum minor allele count of 2. Branch support was estimated through 1,000 bootstraps. As a supplementary control, we also report the correlation matrix estimated from the variance–covariance matrix obtained by the software BAYPASS (version 2.1) (Gautier, 2015) for both GBS and pooled data. The variance–covariance matrix reflects covariation of allele frequencies within and between populations. The correlation matrix describing pairwise relatedness between populations was then derived using the R function cov2cor() provided with BAYPASS. The function hierclust() was used to perform hierarchical clustering based on matrix coefficients. For POPTREE2 analysis and AMOVAs, we accounting for the different number of alleles between males and females for Z-linked markers, including one allele for females and two for males.

2.5 | Demography

To help distinguish between the presence of extensive gene flow between forms and incomplete lineage sorting as an explanation for the generally low differentiation, we performed model comparison under the likelihood framework developed in FASTSIMCOAL2.6 (Excoffier, Dupanloup, Huerta-Sanchez, Sousa, & Foll, 2013) using frequency spectra inferred by ANGSD from autosomal and Z-linked GBS data. We focused on the split between populations from high and low elevations. This split was the clearest across all analyses, with very little genetic differentiation between lowland forms for these markers (see Results). Given the weak genetic substructure for autosomal markers, and because a large number of individuals within groups is required to infer very recent demographic events (Robinson, Coffman, Hickerson, & Gutenkunst, 2014), we pooled individuals into three groups: highlands, lowlands and Mauritius. We acknowledge that models incorporating substructure could be built, but this would come at the cost of adding more parameters. We preferred to limit this study to simple models that can serve as a basis for future, more detailed work using more markers and individuals (Otto & Day, 2007).

We used GBS autosomal and Z-linked markers as they could be filtered with higher stringency than pooled markers and were more likely to follow neutral expectations, with only 1.6% of the genetic variance explained by GBS loci harbouring significant differentiation between highland and lowland forms (see AMOVAs in Results). SNPs mapping on scaffolds corresponding to the neo sex chromosome region on 4A were discarded from the analysis (see Results). We extracted the joint derived site frequency spectrum (SFS) using ANGSD, which takes into account genotypic uncertainties to directly output the most likely SFS. We used the reference genome as an outgroup to assign alleles to ancestral or derived states. Using ANGSD should correct for biases that may occur when calling genotypes from low- and medium-depth sequencing. We filtered markers using the same criteria used for the ADMIXTURE analysis on Z-linked loci (-uniqueOnly 1 -remove_bads 1 -minMapQ 20 -minQ 20 -doPost 2 -geno_minDepth 6 -C 50 -baq 1),

but excluded two individuals with very low average sequencing depth (993 and 11_0990), and removed sites that were not covered in all remaining individuals. For Z-linked markers, we extracted the SFS from male individuals only, given that ANGSD cannot simultaneously extract the spectrum from samples with mixed ploidy. Entries in the joint SFS were examined to exclude potential paralogues displaying strong heterozygosity. We did not filter on minimal allele frequency because singletons are important to properly estimate parameters and likelihoods. We compared four distinct demographic models (Figure 3), one in which all forms were essentially treated as a single population (by forcing a very recent split 10 generations ago) that went through a change in effective population sizes after the split from Mauritius, one with no gene flow between highland and lowland forms, one allowing constant and asymmetric gene flow between lowland and highland forms, and a last model in which gene flow could vary at some time in the past between the present and the split between lowland and highland forms. Population sizes could vary at each splitting time and each group was assigned a specific effective population size. Parameters were estimated from the joint SFS using the likelihood approach implemented in FASTSIMCOAL2.6 (Excoffier et al., 2013). Parameters with the highest likelihood were obtained after 20 cycles of the algorithm, starting with 50,000 coalescent simulations per cycle, and ending with 100,000 simulations. This procedure was replicated 50 times and the set of parameters with the highest final likelihood was retained as the best point estimate. The likelihood estimated by FASTSIMCOAL2.6 is a composite likelihood, which can be biased by covariance between close markers (Excoffier et al., 2013). To properly compare likelihoods and keep the effects of linkage to a minimum, we first used a thinned data set with SNPs separated by at least 10,000 bp (see LD decay, Figure S2). We then used the complete data set for parameter estimation. We used a fixed divergence time of 430,000 years between Reunion and Mauritius grey white-eyes (Warren et al., 2006) and assumed a generation time of 1 year to calibrate parameters and obtain an estimate in demographic units for the timing of diversification in the Reunion grey white-eye. We used python scripts implemented in ∂A∂I (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009) to visualize and compare predicted and observed SFS. To assess deviation from neutrality and stable demography, we estimated Tajima's *D* (Tajima, 1989) from the spectra with ∂A∂I.

We estimated 95% confidence intervals (CIs) using a nonparametric bootstrap procedure, bootstrapping the observed SFS 100 times using ANGSD and repeating the parameter estimation procedure on these data sets, using 10 replicates per bootstrap run to reduce computation time. To visualize whether the model fitted the observed data, we compared the observed SFS with 100 SFS simulated using the set of parameters obtained from the model with the highest likelihood. Coalescent simulations were carried out in FASTSIMCOAL2.6, simulating DNA fragments of the size of GBS loci (91 bp) and using a mutation rate of 3.6×10^{-9} mutations per generation (Axelsson, Smith, Sundström, Berlin, & Ellegren, 2004) until the number of SNPs matched the number of segregating sites in the observed data set. Parameters for each simulation were uniformly drawn from the CIs of the best model. We summarized SFS by PCA using the gfitpca() function of the ABC package (Csilléry, François, & Blum, 2012), including only entries with at least 0.1% of the total number of segregating sites to reduce variance.

2.6 | Selection and environmental association

To detect loci displaying a significant association with geographical forms and environment, we performed an association analysis on the pooled RAD-seq data. We used the approach implemented in BAYPASS (version 2.1) to detect SNPs displaying high differentiation (Gautier, 2015). This approach is designed to robustly handle uncertainties in allele frequencies due to pooling and uneven depth of coverage, by directly using read count data. It performs well in estimating differentiation and population structure (Hivert, Leblois, Petit, Gautier, & Vitalis, 2018). Briefly, BAYPASS estimates a variance–covariance matrix reflecting correlations between allele frequencies across populations. Divergence at each locus is quantified through a Bayesian framework using the X^TX statistic, which can be seen as an SNP-specific F_{ST} corrected by this matrix. BAYPASS also offers the option to estimate an empirical Bayesian *p*-value (*eBPis*) which can be seen as the support for a nonrandom association between alleles and any population-specific covariable. We also computed *eBPis* to determine the level of association of each SNP with geographical form and elevation. Specifically, we tested associations between allele frequencies and a binary covariable stating whether pools belonged or did not belong to Mauritius grey white-eye, GHB, BNB and LBHB forms. We also tested for an association with elevation, coded as a continuous variable. BAYPASS was run using default parameters under the core model with 25 pilot runs and a final run with 100,000 iterations thinned every 100 iterations. Because estimates of the variance–covariance matrix are robust to minor allele count thresholds (Gautier, 2015), we only included SNPs with a minor allele count of 10 over the entire data set to reduce computation time and ran separate analyses on the Z chromosome and autosomes to account for their distinct patterns of differentiation and allele counts.

2.7 | GO enrichment analysis

To gain insight into the putative selective pressures acting on Reunion grey white-eye populations, we performed a Gene Ontology (GO) enrichment analysis, selecting for each association test SNPs in the top 1% for both *eBPis* and X^TX, considering separately the Z chromosome and autosomes. Gene annotations within 100-kb windows flanking selected SNPs were extracted using the zebra finch reference and the intersect function in BEDTOOLS version 2.25.0 (Quinlan & Hall, 2010). We chose 100 kb based on the average territory size of protein-coding genes in the zebra finch (87.8 kb, see table 1 in Warren et al., 2010). We adjusted the gene universe by removing zebra finch genes not mapping onto the *Z. lateralis* genome. GO enrichment analysis was performed using the package topGO in R (Alexa and Rahnenfuhrer, 2016), testing

for significant enrichment using a Fisher's test for over-representation. We present raw *p*-values instead of *p*-values corrected for multiple testing, following recommendations from the topGO manual, and because we wanted to detect any interesting trend in the data set that could then be further explored. We present the top 50 GO terms associated with biological processes, ranked by their raw *p*-values.

3 | RESULTS

3.1 | Genetic structure and relationships among geographical forms

We first assessed whether geographical forms could be distinguished based on the genomic data available for individuals (GBS). A PCA on autosomal allele frequencies revealed a clear distinction between Mauritius and Reunion grey white-eyes on the first axis, as well as a distinction between localities from lowlands and highlands on the second axis (Figure 1b). When excluding Mauritius white-eyes, the main distinction remained between localities from high and low elevation, reflecting differentiation between highland and lowland forms. Further principal components did not reveal any strong clustering of the different forms from the lowlands (Figure 1b). This pattern was further confirmed by the ADMIXTURE analysis, where low and high clustered separately for both autosomal and Z-linked markers (Figure 2). The CV procedure gave $K = 2$ and $K = 7$ as best models for autosomal and Z-linked markers respectively (Figure S3). However, CV scores were low for all K values ranging from 5 to 7 for Z-linked markers, making it difficult to clearly identify an optimal value. Given the subtle genetic structure, we present results for values of K ranging from 2 to 7.

Clustering was consistent with the PCA, with a distinction between localities from high and low elevations at $K = 2$ and 3, for Z-linked and autosomal markers respectively (Figure 2a). For autosomal markers, higher values of K highlighted a structure consistent with sampling sites, but structure according to geographical forms was more elusive. For Z-markers, clustering tended to be more consistent with forms at low elevation. At $K = 7$, one cluster corresponded to localities at high elevation, three others grouped lowland localities by forms, while a fifth cluster included Mauritian individuals (Figure 2a). Two localities (LF and TC) displayed stronger signals of mixed ancestry, probably due to gene flow between those two localities which are close to a zone of contact between the GHB and BNB forms. The same pattern was observed using the set of Z-linked SNPs called with ANGSD, the distinction between GHB and BNB forms being even clearer, probably due to the larger number of markers (Figure S4). We note that using this set of markers improved the discrimination between colour forms, probably because of the higher number of SNPs remaining after filtering with ANGSD (2,282 loci) instead of FREEBAYES (965 loci).

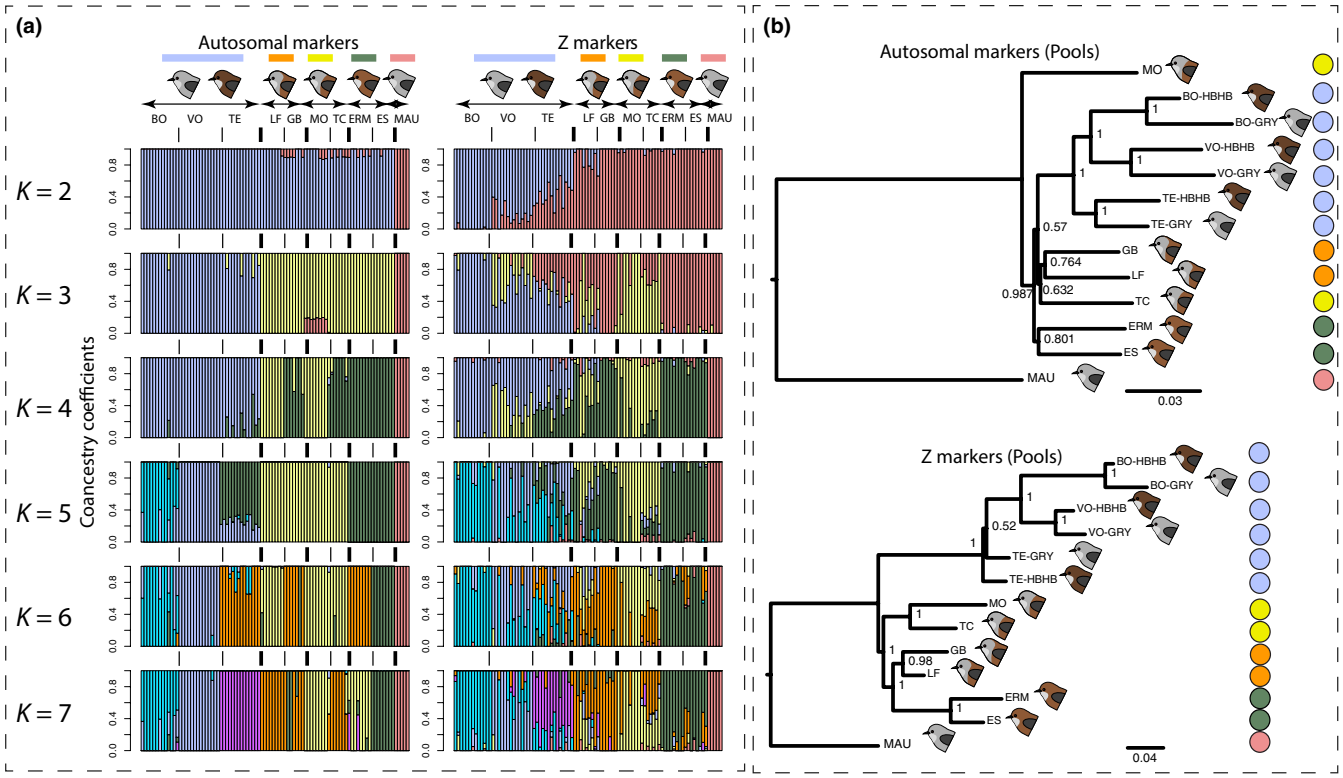


FIGURE 2 (a) Co-ancestry coefficients obtained from ADMIXTURE for K values ranging from 2 to 7. Population codes as in Figure 1. Separate analyses were run on autosomal and Z markers. Bold and thin vertical lines indicate limits between forms and populations, respectively. (b) POPTREE2 analysis obtained from 20,000 markers randomly sampled from pooled data. Branch support was obtained from 1,000 bootstraps [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Pairwise AMOVAs comparing geographical forms within Reunion grey white-eye and between Reunion and Mauritius grey white-eyes. All values are significant (1,000 permutations). To properly compare patterns observed on sex-linked and autosomal markers, results for autosomal markers using only males are also shown

	LBHB versus BNB		GHB versus LBHB		GHB versus BNB		High versus Low		Reunion versus Mauritius	
Comparison	***F*-statistics**	**% of variance explained**	***F*-statistics**	**% of variance explained**	***F*-statistics**	**% of variance explained**	***F*-statistics**	**% of variance explained**	***F*-statistics**	**% of variance explained**
Autosomes (without 4A sex-linked markers)										
Among groups (F_{CT})	0.005	0.48	0.010	1.03	0.006	0.59	0.016	1.57	0.173	17.31
Among populations within groups (F_{SC})	0.033	3.31	0.042	4.11	0.035	3.49	0.040	3.97	0.049	4.09
Within populations (F_{ST})	0.038	96.21	0.051	94.86	0.041	95.92	0.052	94.46	0.214	78.61
Z-linked markers										
Among groups (F_{CT})	0.068	6.82	0.108	10.84	0.034	3.45	0.126	12.59	0.206	20.61
Among populations within groups (F_{SC})	0.038	3.56	0.093	8.28	0.095	9.13	0.123	10.78	0.191	15.17
Within populations (F_{ST})	0.104	89.62	0.191	80.88	0.095	87.42	0.234	76.63	0.358	64.21
Autosomes (without 4A sex-linked markers, males only)										
Among groups (F_{CT})	0.007	0.69	0.009	0.93	0.005	0.53	0.019	1.86	0.166	16.6
Among populations within groups (F_{SC})	0.013	1.32	0.029	2.87	0.027	2.72	0.040	3.92	0.051	4.25
Within populations (F_{ST})	0.02	97.98	0.038	96.2	0.032	96.75	0.058	94.22	0.209	79.15
4A–1–10 Mb (males only)										
Among groups (F_{CT})	0.044	4.44	0.069	6.92	0.048	4.81	0.080	7.95	0.370	28.58
Among populations within groups (F_{SC})	0.050	4.80	0.059	5.47	0.039	3.68	0.075	6.89	0.118	8.44
Within populations (F_{ST})	0.044	90.76	0.124	87.61	0.085	91.51	0.148	85.16	0.286	62.98

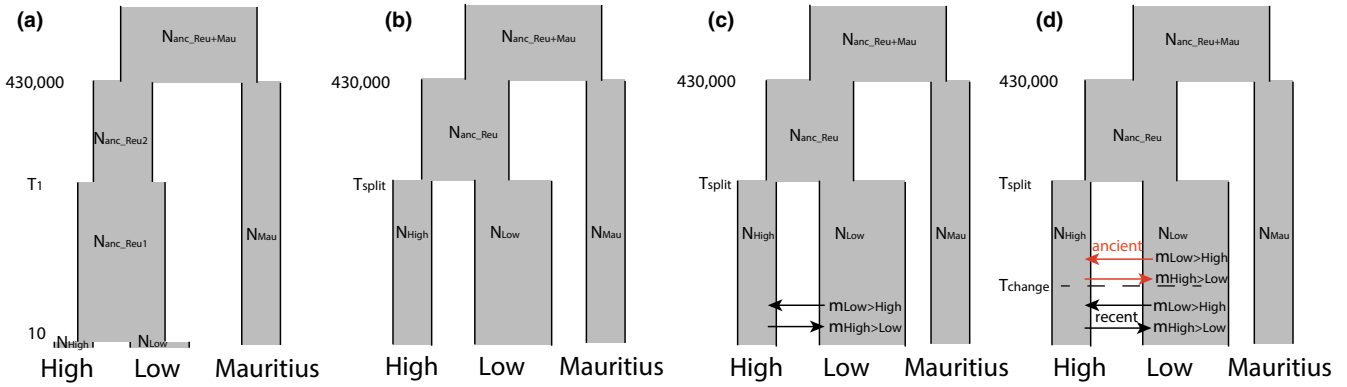


FIGURE 3 Summary of the four demographic models and their parameters evaluated using GBS data [Colour figure can be viewed at wileyonlinelibrary.com]

Recent studies revealed the existence of neo-sex chromosomes in Sylvioidea, to which Zosteropidae belong (Pala et al., 2012), and their existence was recently confirmed in the Reunion grey white-eye (Leroy et al., 2019). These neo-sex chromosomes emerged from the fusion of Z and W chromosomes with the first 10 Mb of chromosome 4A. We extracted GBS markers mapping to this region and confirmed that they were sex-linked, as shown by a PCA on allele frequencies (Figure S5). Visual examination of genotypes revealed that females (ZW) displayed a strong excess of heterozygous markers, in contrast to males (ZZ), due to divergence between neo-Z and neo-W chromosomes. We further investigated population structure in males only, using a set of combined Z and 4A markers, and an ADMIXTURE analysis tended to better discriminate geographical forms at $K = 6$ with this set of markers (Figure S6).

The same pattern was observed with the POPTREE2 analysis on the pooled data set. The autosomal topology supported a grouping of localities from high elevation together and supported a grouping of localities belonging to BNB and LBHB forms, yet there was no support for grouping together localities from the GHB form. In contrast, topology based on Z markers provided good support for a grouping of localities by geographical form (Figure 2b).

We further investigated population structure by estimating the variance–covariance matrix obtained from allele frequencies for both pooled and GBS data using BAYPASS (Gautier, 2015). Localities from high elevation were systematically found clustering together, a pattern consistent with previous analyses (Figure S7). Again, both analyses on Z-linked markers for GBS and pooled data revealed a closer relationship between populations from the same geographical form when compared to autosomal markers. LBHB and BNB forms clustered together, with the GHB form branching off first within the lowland group.

To estimate quantitatively the proportion of the genome discriminating among geographical forms while accounting for population structure within forms, we performed an AMOVA on GBS data. This analysis confirmed the previous patterns, with a proportion of variance explained by forms or elevation not higher than 1.6% for autosomal markers (Table 1). The strongest differentiation was observed between localities from low and high elevations and between Mauritius and Reunion populations. For Z-linked markers, however, the proportion of variance explained by forms and elevation was more substantial, ranging from 3.4% to 12.6% (Table 1). This higher differentiation could not be explained by differences in sample size between autosomal and Z-linked data. For autosomal markers, AMOVAs performed only on males did not show substantial deviations from the results obtained with the full data set (Table 1).

3.2 | Demographic history

We compared four nested models (Figure 3), allowing for no or constant gene flow after the split between populations. The highest

TABLE 2 Likelihoods of the four demographic models compared with FASTSIMCOAL2.6, and their AIC scores. Note that we report likelihoods in natural logs, while likelihoods produced by FASTSIMCOAL are expressed as $\log_{10}$

Marker	Model	Log(Likelihood)	Number of parameters	AIC	ΔAIC
Autosomes	D	−36,333.59	11	72,689.19	0.00
	C	−36,372.41	8	72,760.81	71.62
	B	−36,453.23	6	72,918.45	229.27
	A	−37,509.72	7	75,033.44	2,344.25
Z-linked	D	−1611.58	11	3,245.17	0.00
	C	−1628.72	8	3,273.44	28.28
	B	−1662.54	6	3,337.09	91.92
	A	−1957.25	7	3,928.51	683.34

TABLE 3 Parameter estimates for a model with migration rates using SFS from ANGSD. Population sizes are haploid sizes (with *N* representing the number of diploid individuals)

Marker	Parameter	$2N_{Mauritius}$	$2N_{Low}$	$2N_{High}$	$2N_{anc\_Reu}$	$2N_{anc\_Reu+Mau}$	T_{split}	T_{change}	$m_{Low\to High}$ (recent)	$m_{High\to Low}$ (recent)	$m_{Low\to High}$ (ancient)	$m_{High\to Low}$ (ancient)
Autosomal	Best estimate	1,081,493	2,957,776	606,586	1,266,228	1,022,327	400,228	79,188	1.63E−05	2.03E−06	2.00E−07	3.11E−06
Autosomal	2.5% lower bound	906,204	2,303,253	480,897	596,984	916,712	374,624	62,305	1.54E−05	2.76E−07	1.40E−08	2.36E−06
Autosomal	97.5% upper bound	1,158,440	2,957,776	686,191	1,332,774	1,140,005	423,302	204,547	2.17E−05	3.60E−06	3.82E−06	5.64E−06
Z-linked	Best estimate	640,280	1,595,330	228,792	996,150	905,370	404,812	133,746	6.66E−06	3.10E−07	1.92E−08	2.28E−09
Z-linked	2.5% lower bound	490,028	1,195,174	140,680	596,824	565,930	348,043	66,246	4.34E−06	3.76E−08	1.47E−09	6.91E−10
Z-linked	97.5% upper bound	878,312	1,807,538	338,704	1,175,232	1,194,501	419,568	166,238	1.05E−05	1.14E−06	2.36E−06	2.73E−07

likelihoods were found for the model allowing change in gene flow at some time (T_{change}) after the initial split at T_{split} between lowlands and highlands (model D, Figure 3), while the other models were clearly rejected (Table 2). Such a pattern of higher gene flow in recent times is consistent with a scenario of introgression through secondary contact.

Assuming a conservative divergence time between *Z. borbonicus* and *Z. mauritianus* of 430,000 years, parameter estimates suggested a split between high- and low-elevation populations 400,000 years ago and an increase in gene flow 80,000 years ago (Table 3). Overall, point estimates of effective migration rates ($2Nm$, with N the diploid population size) were high ($2N_{High}$ $m_{Low\to High}$ = 0.1 gene copies per generation, $2N_{Low}$ $m_{High\to Low}$ = 9.2 before T_{change}, then reaching $2N_{High}$ $m_{Low\to High}$ = 9.9 and $2N_{Low}$ $m_{High\to Low}$ = 6.0 80,000 years ago), consistent with homogenization of genomes through migration. This model was able to explain our observed data set, as indicated by visual comparisons of the observed and predicted SFS (Figure S8a). Coalescent simulations using parameters drawn from CIs of the best model produced SFS similar to the observed one (Figure S8b).

Genes that are involved in reproductive isolation between populations are expected to resist gene flow due to counterselection of maladapted or incompatible alleles. This should result in an increased divergence at these loci when compared to the genomic background (Cruickshank & Hahn, 2014). We tested whether reduced gene flow explained increased divergence in the Z chromosome by estimating parameters from the four models but using this time all Z-linked markers, because they explained most of the differentiation between forms. We used the same filtering criteria that we used for autosomes but focused our analysis on diploid males only. The highest likelihood was also found for the model allowing gene flow to change after T_{split} (Table 2), again providing support for a scenario of introgression through secondary contact. As expected, given hemizygosity, effective population size estimates were lower for this set of markers than for autosomes (Table 3). Estimates of gene flow were also historically lower than for autosomes, while time since the split between highland and lowland groups was similar. This combination of lower population sizes and lower gene flow is expected to lead to increased divergence at Z-linked loci, in accordance with the stronger differentiation observed at these markers in AMOVAs and other tests.

To further explore whether differences in demographic inferences may be due to stronger effects of linked selection on the Z chromosome, we computed Tajima's *D* for each group (Tajima, 1989). This statistic should be negative (<2) in the case of recent population expansion, or if positive/purifying selection is acting. It should be positive (>2) in the case of a recent bottleneck or balancing selection. Tajima's *D* values for the autosomal spectra obtained by ANGSD were − 0.29, −0.73 and − 1.18 for Mauritius, highlands and lowlands respectively. Tajima's *D* values were higher for Z-linked markers, at 0.12, 0.09 and − 0.76, but followed the same trend, being lower in lowlands than in highlands

and Mauritius. This suggests that widespread effects of linked selection on the Z chromosome are unlikely to explain its higher differentiation.

3.3 | Genome scan for association and selection analysis

We used BAYPASS (Gautier, 2015) on pooled data to retrieve markers displaying high levels of differentiation (X^TX) and association (empirical Bayesian *p*-values; *eBPis*) with five different features (elevation, being GHB, being BNB, being LBHB, being Mauritian). Overall, this pooled data set included a total of 284 individuals sequenced at an average depth of ~250× over all pools. We examined Z and autosomal markers separately due to their distinct demographic histories. The results revealed a strong association of Z-linked markers with geographical forms and their associated colour phenotypes and elevation (Figure 4). SNPs discriminating Mauritius from other populations were distributed uniformly along the genome. Most of the peaks displaying a large X^TX were also found associated with elevation. Strikingly, the clearest peaks on chromosomes 2, 3 and 5 covered large genomic regions, spanning several hundreds of kilobases. The sex-linked region on chromosome 4A was the clearest outlier. Because the neo-W chromosome is highly divergent and found in all females, an excess of variants with frequencies correlated to the proportion of females in the pool is expected. This may lead to high differentiation between pools with different sex ratios. Despite this, the strong association with elevation and geographical form on chromosome 4A is genuine because the expected proportion of divergent W-linked alleles in each pool was not correlated with those variables in our experiment (Figure S9), making confounding effects unlikely.

To assess the possible role of these genomic regions in adaptation, we performed a GO enrichment analysis using the zebra finch (*Taeniopygia guttata*) annotation (Tables S3-S6). Genes found in regions associated with elevation displayed enrichment for GO terms linked to development, body growth, gene expression, RNA metabolism, DNA organization, immune-system development, haematopoiesis and haemoglobin synthesis (GO:0005833: haemoglobin complex, three genes found over four in total, $p = .0031$). For genes associated with each one of the parapatric lowland forms, we mostly found associations with immune response, metabolic process, haematopoiesis, reproduction, morphogenesis and development (Tables S4-S6).

To identify candidates for plumage colour variation between forms, we screened the genes found in outlier regions for GO terms linked to melanin synthesis and metabolism (GO:0042438, GO:0046150, GO:0006582). *TYRP1* (*tyrosinase-related protein 1*), located on the Z chromosome, was the only gene that was systematically found associated with BNB, GHB and LBHB forms. Another gene, *WNT5A*, was found associated only with the GHB form.

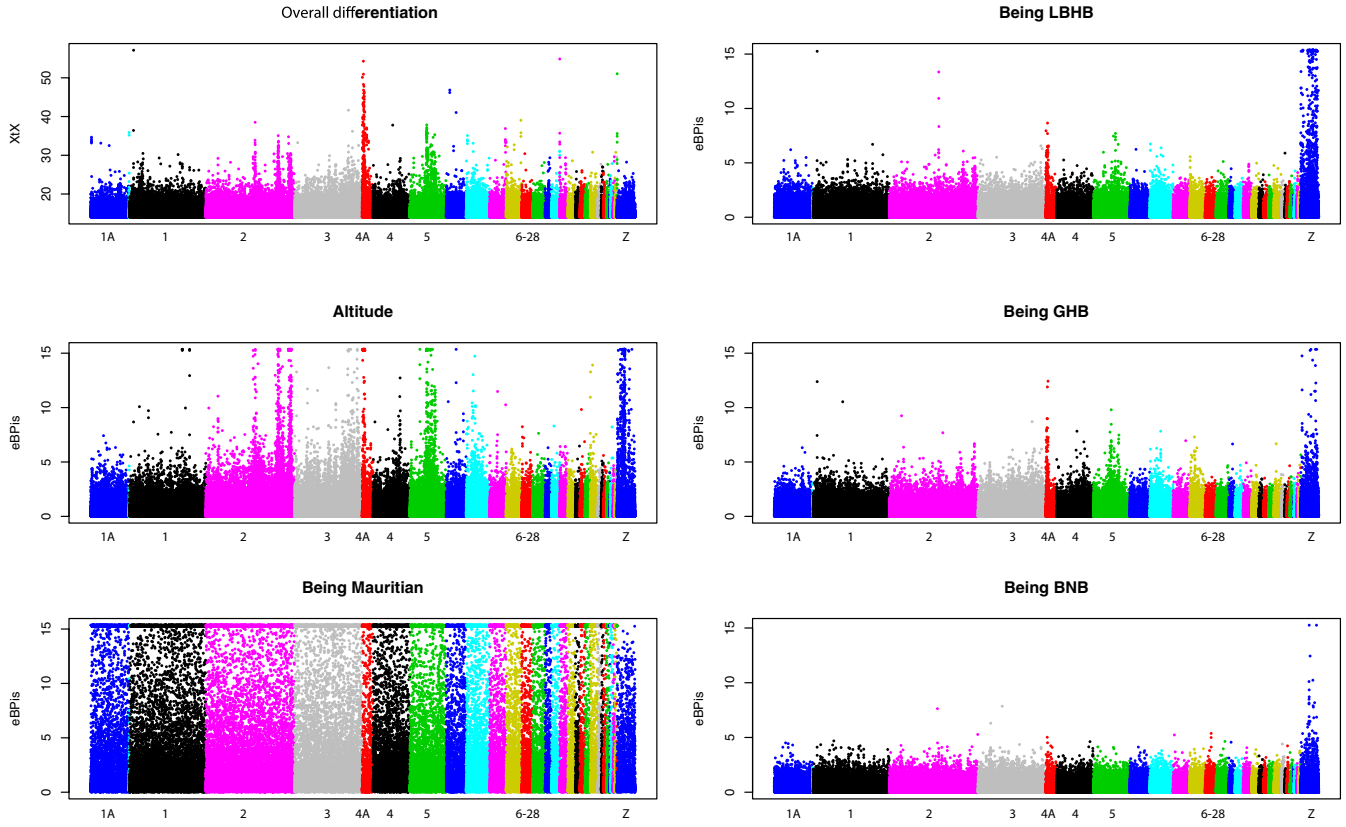


FIGURE 4 Manhattan plots of X^TX and Bayesian empirical *p*-values (*eBPis*) for an association with elevation and geographical form. Note that X^TX values are not directly comparable for autosomes and Z chromosomes as they were analysed separately and display different demographic histories (see Figure 3) [Colour figure can be viewed at wileyonlinelibrary.com]

4 | DISCUSSION

4.1 | Genetic structure and genomic islands of differentiation

Our results quantify the relative importance of autosomal and sex-linked genetic variation underlying a potential case of incipient speciation in the Reunion grey white-eye. We confirm previous findings based on microsatellite data regarding the existence of a fine-grained population structure, with low but significant F_{ST} between localities (Bertrand et al., 2014), and a clear divergence between populations from low and high elevations (Bertrand et al., 2016; Cornuault et al., 2015). These events of divergence have probably occurred in the last 430,000 years (Warren et al., 2006), and perhaps even more recently as this divergence time estimate was based on mitochondrial data that tend to yield overestimates of population splitting times (Moore, 1995). A striking result is the clear contrast between patterns of variation in autosomes and the Z chromosome, the latter discriminating more clearly the different geographical forms and systematically displaying markers associated with plumage colour phenotypes and elevational ranges. Similar results were found independently for both pooled and individual data sets, and with two different SNP callers. These observations combined with demographic analyses suggest that while extensive gene flow has occurred between populations from high and low elevations, the Z chromosome acts as a barrier to it.

4.2 | Reproductive isolation between forms may explain divergence at the Z chromosome

Reproductive isolation between nascent species can arise as incompatibilities between interacting loci accumulate in the genome, as described by the Bateson–Dobzhansky–Muller model (Dobzhansky 1934; Muller 1940, 1942; Orr, 1996). Sex chromosomes are particularly likely to accumulate such incompatibilities, because the genes they harbour are not transmitted by the same rules between males and females (Seehausen et al., 2014) and are therefore prone to genetic conflicts. Indeed, sex chromosomes often harbour many genes causing disruption of fertility or lower viability in hybrids (Good, Dean, & Nachman, 2008; Masly & Presgraves, 2007; Storchová et al., 2004; Storchová, Reif, & Nachman, 2010), which may lead to increased divergence (Carling & Brumfield, 2008; Ellegren et al., 2012; Macholán et al., 2011), and faster emergence of postzygotic isolation (Lima, 2014). Hemizygosity of sex chromosomes may also lead to the exposure of recessive incompatibilities in heterogametic individuals (Qvarnström & Bailey, 2009). This large effect of sex chromosomes has been a common explanation of Haldane's rule (Haldane, 1922; Orr 1997; Coyne & Orr, 2004), which states that in hybrids, the heterogametic sex often displays a stronger reduction in fitness, and may explain why an excess of highly differentiated regions on sex chromosomes is often observed during the early stages of speciation (e.g., Backström et al., 2010). An excess of highly differentiated regions on sex chromosomes at the early stages of speciation therefore suggests a role for intrinsic barriers to gene flow in promoting divergence (Backström et al., 2010).

In addition, premating and prezygotic isolation affect sexually dimorphic traits that are under the control of sex-linked genes (Pryke, 2010). While this type of isolation would lead to divergence across the entire genome due to the isolation of gene pools, divergence is expected to be accelerated at loci controlling traits under sexual selection, especially if hybrids and backcrosses have lower mating success (Svedin, Wiley, Veen, Gustafsson, & Qvarnstrom, 2008). Finally, effective recombination rates are lower in sex chromosomes because they recombine in only one sex (males in birds), which facilitates linkage of genes involved in pre- and post-zygotic barriers. This linkage would ultimately promote reinforcement of isolation between species or populations (Pryke, 2010). Because of these processes, gene flow at isolating sex-linked loci is expected to be impeded between diverging populations, leaving a stronger signature of differentiation than in the rest of the genome (e.g., Mořkovský et al., 2018).

Our findings are in line with these theoretical expectations, and show an excess of highly differentiated markers on the Z chromosome that display evidence for resisting gene flow, a pattern that is also consistent with Haldane's rule (e.g., Carling & Brumfield, 2008). Importantly, highly differentiated autosomal markers were mostly associated with differences in elevation ranges but Z chromosome variation was found to be associated with both elevation and plumage colour differences between lowland forms. This suggests that the divergence between lowland and highland forms is due to both polygenic adaptation to different elevational ranges and behavioural differentiation (e.g., mating preferences) (Sæther et al., 2007).

We acknowledge that we have no direct evidence yet of assortative mating, character displacement or lower hybrid fitness in contact zones between the different geographical forms of our study species, but: (a) a previous study showed that Reunion grey white-eyes can perceive colour differences between forms, and that both within- and between-form differences in plumage colour can be discriminated (Cornuault et al., 2015); (b) hybrid zones are narrow among lowland forms (Delahaie et al., 2017) and quite likely also between lowland and highland forms (Bertrand et al., 2016), suggesting that hybrid phenotypes must have a lower fitness; and (c) high gene flow within forms can erase signatures of character displacement if alleles responsible for premating isolation are nearly neutral in populations that are distant from contact zones (Servedio & Noor, 2003).

4.3 | Role of drift and recent selective sweeps in Z chromosome divergence

The higher differentiation observed on the Z chromosome may also be due to its reduced effective population size, which leads to faster drift and lineage sorting between forms. Such a mechanism may explain the generally higher rate of divergence observed on Z

chromosomes (fast-Z effect). This effect can become even stronger in the presence of a recent bottleneck and population size change (Van Belleghem et al., 2018; Pool & Nielsen, 2007). However, we did not find any evidence for strong bottlenecks or abrupt changes in population sizes, making faster accumulation of divergent alleles in recently established populations unlikely. Moreover, another recent study on *Z. borbonicus* has also shown that there is weak support for a fast-Z effect in the clade to which this species belongs (Leroy et al., 2019). Another mechanism that might lead to stronger differentiation at Z-linked loci is female-biased dispersal, which is frequent in passerine birds (Greenwood, 1980). However, dispersal in *Z. borbonicus* is extremely reduced for both sexes (Bertrand et al., 2014), which should attenuate the contrast between autosomal and Z-linked markers. In addition, although it could lead to stronger differentiation between localities, such a mechanism alone is unlikely to explain why such a high proportion of variance is associated with forms on the Z chromosome. Finally, we do not observe higher differentiation on Z markers or autosomes in males as compared to all individuals (Table 1).

Lower effective recombination rates on the sex chromosome may enhance the effects of selection at sites linked to loci involved in ecological adaptation, further reducing diversity on the Z chromosome. In addition, a possible explanation of a fast-Z effect in birds may result from positive selection on recessive beneficial alleles in heterogametic females (Dean, Harrison, Wright, Zimmer, & Mank, 2015). Faster drift on the Z chromosome may lead to a faster accumulation of incompatibilities (Janoušek et al., 2019), which will further increase divergence. Ultimately, drift and selection are interconnected processes that are difficult to disentangle. Unfortunately, we cannot provide with our data set alone a detailed understanding of how processes such as linked selection, chromosomal rearrangements and barriers to gene flow can interact (Bierne et al., 2011; Ravinet et al., 2017). Future studies using whole-genome resequencing data should provide a clearer picture, by providing information on genealogies and age of both autosomal and sex-linked haplotypes.

However, we note that the allele frequency spectra of lowland and highland forms do not display the expected signature of pervasive linked selection, with Tajima's *D* actually higher for Z markers than for autosomes. Our demographic analyses also show that the effective population sizes estimated from Z markers are reduced by a factor of 0.38–0.89 when compared to autosomes, the neutral expectation being 0.75. We acknowledge that polygenic adaptation to divergent environmental pressures is probably partly responsible for the observed genomic landscape of differentiation, possibly leading to localized divergence at autosomal and Z loci (see below). However, reproductive isolation seems to be at play in this system, either through premating isolation based on plumage characters, and/or post-zygotic isolation. More research about the ecology of the Reunion grey white-eye would also be useful to quantify the extent of sex-biased dispersal, assortative mating, parental imprinting, intrinsic incompatibilities, and how these factors may interact in this system (Pryke, 2010; Seehausen et al., 2014).

4.4 | Autosomal divergence is mostly associated with elevation

Adaptation to local environmental conditions or natural selection against hybrids is more likely to be driven by genes scattered across the genome, assuming polygenic selection (Qvarnström & Bailey, 2009; Seehausen et al., 2014). In addition to Z-linked loci, we found many differentiated loci on autosomes, mostly in association with elevation. Genomic regions associated with elevation displayed an enrichment of genes involved in development and body growth, and included the cluster of haemoglobin subunits A, B and Z on chromosome 14. The function of these genes is consistent with biological expectations, given the wide elevational range (from 0 to 3,000 m) occupied by white-eyes on Reunion.

Large chromosomal rearrangements are powerful drivers of differentiation, as they prevent recombination between several consecutive genes, facilitating the maintenance of divergent allele combinations between populations. A famous example of adaptive inversions facilitating the maintenance of colour (geographical) forms and species has been reported in *Heliconius* butterflies (Joron et al., 2011) or the white-throated sparrow (Tuttle et al., 2016), and these rearrangements have been predicted to take place in isolation followed by secondary contact (Feder, Gejji, Powell, & Nosil, 2011). In this study, autosomal regions associated with geographical forms and elevation sometimes spanned more than 1 Mb. This suggests a possible role for large-scale rearrangements and linked selection in regions of low recombination as a substrate for divergence in Reunion grey white-eyes. Our results remind at a much smaller spatial scale what has been previously observed in *Ficedula* flycatchers, with high differentiation on the Z chromosome, linked selection and chromosomal rearrangements (Burri et al., 2015; Ellegren et al., 2012).

The low number of Z-linked markers found in this study and their higher level of population differentiation may limit the interpretation of results when comparing patterns of divergence with autosomes. Whole-genome resequencing data and refined demographic models building on those used in this study will be critical to precisely quantify the evolutionary dynamics of the Z chromosome compared to autosomes, and identify at a higher resolution the loci displaying strong divergence. Future studies should also focus on the variation in allele frequencies along hybrid zones and test whether loci that are more likely to be involved in local adaptation (such as immune genes or haemoglobin subunits) display changes in frequencies that are as sharp as Z-linked loci, as the latter are more probably involved in both pre- and post-zygotic isolation. Overall, our results suggest an extreme case of divergence with gene flow that can bring valuable insights into the relative order at which pre- and post-zygotic isolation mechanisms occur during speciation.

4.5 | Genetics of colour

TYRP1, a well-characterized colour gene in both model species and natural populations (Abolins-Abols et al., 2018; Backström

et al., 2010; Delmore, Toews, Germain, Owens, & Irwin, 2016; Nadeau, Mundy, Gourichon, & Minvielle, 2007), was the only known colour gene found in the regions associated with each of the three lowland forms. This gene had been previously studied using a candidate gene approach in *Z. borbonicus*, but no clear association with plumage colour phenotypes was found, probably owing to the low levels of polymorphism displayed among lowland forms (Bourgeois et al., 2016). The *WNT5A* gene was found associated with the GHB form, with a brown back and a grey head. This gene is known to regulate *TYRP1* expression (Zhang et al., 2013) and is differentially expressed between black carrion and grey-coated hooded crows (Poelstra, Vijay, Hoeppner, & Wolf, 2015). However, this gene is not only involved in melanogenesis but also in cell migration and differentiation, making it a less straightforward candidate in this system.

Our pooled RAD-seq approach covered only about 10%–20% of the genome, and may therefore have missed some colour loci. The likelihood is high, however, that recent selective sweeps would have been detected if such genes had been targeted by selection. For example, the locus underlying colour polymorphism in the high-elevation form shows clear signs of a selective sweep reducing diversity over 500 kb (see Figure 3 in Bourgeois et al., 2017), a region which is large enough to be covered by tens of RAD-seq loci. However, cases of long-term balancing selection may be harder to detect because of the short haplotypes typically found in this case due to extensive recombination. Together with previous findings (Bourgeois, Bertrand, Thébaud, & Milá, 2012; Bourgeois et al., 2017), this suggests that a large part of plumage colour variation between the geographical forms of the Reunion grey white-eye may be controlled by a set of a few loci of major effect. More detailed studies of hybrid zones between the different lowland forms may help to characterize the exact association of alleles that produce a given plumage colour phenotype.

ACKNOWLEDGMENTS

We thank Joseph Manthey, Stéphane Boissinot, Maëva Gabrielli and Thibault Leroy for insightful comments that improved the manuscript. We also thank the Reunion National Park for granting us permission to conduct fieldwork and to collect blood samples. Thomas Duval, Guillaume Gélinaud, Josselin Cornuault, Philipp Heeb, Dominique Strasberg, Ben Warren and Juli Broggi assisted with fieldwork. Emeline Lhuillier and Olivier Bouchez assisted with the development of pooled RAD-seq. This research was carried out on the High-Performance Computing resources at New York University Abu Dhabi and the Genotoul HPC cluster. This work was supported by Fondation pour la Recherche sur la Biodiversité (FRB), Agence Française pour le Développement (AFD), Agence Nationale de la Recherche (ANR-2006-BDIV002), Centre National de la Recherche Scientifique (CNRS) through a PEPS grant, The National Geographic Society, and the "Laboratoire d'Excellence" TULIP (ANR-10-LABX-41). The first author was supported by an MESR (Ministère de l'Enseignement Supérieur et de la Recherche) PhD scholarship during this study.

AUTHOR CONTRIBUTIONS

B.M. and C.T. initiated, coordinated and supervised the project; Y.B., B.M. and C.T. conceived the study and designed the experiments; B.M., C.T., Y.B., J.A.B. and B.D. conducted the fieldwork; molecular data were generated by Y.B. and H.H.; Y.B. analysed the data; and Y.B., B.M. and C.T. wrote the paper with input from the other authors. All authors gave final approval for publication.

DATA AVAILABILITY STATEMENT

All data associated with this manuscript are published on DRYAD (VCF files, allele counts for pooled data and position of *Z. lateralis* scaffolds on zebra finch chromosomes; https://doi.org/10.5061/dryad.z34tmpg8z) and European Nucleotide Archive (BAM files for pools and fastq files for individual GBS data; accession number PRJEB36701, https://www.ebi.ac.uk/ena/browser/view/PRJEB36701).

ORCID

Yann X. C. Bourgeois https://orcid.org/0000-0002-1809-387X
Joris A. M. Bertrand https://orcid.org/0000-0002-3379-1019
Borja Milá https://orcid.org/0000-0002-6446-0079

REFERENCES

Abolins-Abols, M., Kornobis, E., Ribeca, P., Wakamatsu, K., Peterson, M. P., Ketterson, E. D., … Milá, B. (2018). Differential gene regulation underlies variation in melanic plumage coloration in the dark-eyed junco (*Junco hyemalis*). *Molecular Ecology*, *27*, 4501–4515.

Alexa, A., & Rahnenfuhrer, J. (2016). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.26.0.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664.

Axelsson, E., Smith, N. G. C., Sundström, H., Berlin, S., & Ellegren, H. (2004). Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and Turkey. *Molecular Biology and Evolution*, *21*, 1538–1547.

Backström, N., Lindell, J., Zhang, Y. U., Palkopoulou, E., Qvarnström, A., Saetre, G.-P., & Ellegren, H. (2010). A high-density scan of the Z chromosome in *Ficedula* flycatchers reveals candidate loci for diversifying selection. *Evolution*, *64*, 3461–3475.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*, e3376.

Bertrand, J. A. M., Bourgeois, Y. X. C., Delahaie, B., Duval, T., García-Jiménez, R., Cornuault, J., … Thébaud, C. (2014). Extremely reduced dispersal and gene flow in an island bird. *Heredity*, *112*, 190–196.

Bertrand, J. A. M., Delahaie, B., Bourgeois, Y. X. C., Duval, T., García-Jiménez, R., Cornuault, J., … Milá, B. (2016). The role of selection and historical factors in driving population differentiation along an elevational gradient in an island bird. *Journal of Evolutionary Biology*, *29*, 824–836.

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis, why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*, 2044–2072.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic, A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120.

Bourgeois, Y. X. C., Bertrand, J. A. M., Delahaie, B., Cornuault, J., Duval, T., Milá, B., & Thébaud, C. (2016). Candidate gene analysis suggests untapped genetic complexity in melanin-based pigmentation in birds. *Journal of Heredity*, *107*, 327–335.

Bourgeois, Y. X. C., Bertrand, J. A. M., Thébaud, C., & Milá, B. (2012). Investigating the role of the melanocortin-1 receptor gene in an extreme case of microgeographical variation in the pattern of melanin-based plumage pigmentation. *PLoS ONE*, *7*, e50906.

Bourgeois, Y. X. C., Delahaie, B., Gautier, M., Lhuillier, E., Malé, P.-J., Bertrand, J. A. M., … Thébaud, C. (2017). A novel locus on chromosome 1 underlies the evolution of a melanic plumage polymorphism in a wild songbird. *Royal Society Open Science*, *4*(2), 160805.

Bourgeois, Y. X. C., Lhuillier, E., Cézard, T., Bertrand, J. A. M., Delahaie, B., Cornuault, J., … Thébaud, C. (2013). Mass production of SNP markers in a nonmodel passerine bird through RAD sequencing and contig mapping to the zebra finch genome. *Molecular Ecology Resources*, *13*, 899–907.

Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, *1*, 118–131.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L. et al (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Research*, *25*, 1656–1665.

Carling, M. D., & Brumfield, R. T. (2008). Haldane's rule in an avian system, using cline theory and divergence population genetics to test for differential introgression of mitochondrial, autosomal, and sex-linked loci across the *Passerina* bunting hybrid zone. *Evolution*, *62*, 2600–2615.

Cornetti, L., Valente, L. M., Dunning, L. T., Quan, X., Black, R. A., Hébert, O., & Savolainen, V. (2015). The genome of the 'great speciator' provides insights into bird diversification. *Genome Biology and Evolution*, *7*, 2680–2691.

Cornuault, J., Delahaie, B., Bertrand, J. A. M., Bourgeois, Y. X. C., Mila, B., Heeb, P., Thébaud, C. (2015). Morphological and plumage colour variation in the Réunion grey white-eye (Aves, *Zosterops borbonicus*), assessing the role of selection. *Biological Journal of the Linnean Society*, *114*, 459–473.

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sunderland, MA: Sinauer Associates.

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*, 3133–3157.

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc, an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*, 475–479.

Dobzhansky, T. (1934). Studies on Hybrid Sterility. I. Spermatogenesis in pure and hybrid Drosophila pseudoobscura. *Zeitschrift für Zellforschung und mikroskopische Anatomie*. *21*(2), 169–221

Dean, R., Harrison, P. W., Wright, A. E., Zimmer, F., & Mank, J. E. (2015). Positive selection underlies faster-Z evolution of gene expression in birds. *Molecular Biology and Evolution*, *32*, 2646–2656.

Delahaie, B., Cornuault, J., Masson, C., Bertrand, J. A. M., Bourgeois, Y. X. C., Milá, B., & Thébaud, C. (2017). Narrow hybrid zones in spite of very low population differentiation in neutral markers in an island bird species complex. *Journal of Evolutionary Biology*, *30*, 2132–2145.

Delmore, K. E., Hübner, S., Kane, N. C., Schuster, R., Andrew, R. L., Câmara, F., … Irwin, D. E. (2015). Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. *Molecular Ecology*, *24*, 1873–1888.

Delmore, K. E., Toews, D. P. L., Germain, R. R., Owens, G. L., & Irwin, D. E. (2016). The genetics of seasonal migration and plumage color. *Current Biology*, *26*, 2167–2173.

Derjusheva, S., Kurganova, A., Habermann, F., & Gaginskaya, E. (2004). High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds. *Chromosome Research*, *12*, 715–723.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., … Wolf, J. B. W. (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, *491*, 756–760.

Elshire, R. J., Glaubitz, J. C., Sun, Q. I., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*, e19379.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS*, *Genet*, *9*.

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5, a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*, 564–567.

Feder, J. L., Gejji, R., Powell, T. H. Q., & Nosil, P. (2011). Adaptive chromosomal divergence driven by mixed geographic mode of evolution. *Evolution*, *65*, 2157–2170.

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. arXiv Prepr arXiv12073907, *9*.

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, *201*, 1555–1579.

Gavrilets, S. (2014). Models of speciation, Where are we now? *Journal of Heredity*, *105*, 743–755.

Gill, F. B. (1973). Intra-island variation in the Mascarene White-eye *Zosterops borbonica*. *Ornithol. Monogr*, *12*. pp. iii-vi, 1–66.

Gill, F., & Donsker, D. (2019). *IOC World Bird List v9.1*. Int Ornithol. Union Comm Nomencl.

Good, J. M., Dean, M. D., & Nachman, M. W. (2008). A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics*, *179*, 2213–2228.

Greenwood, P. J. (1980). Mating systems, philopatry and dispersal in birds and mammals. *Animal Behavior*, *28*, 1140–1162.

Griffiths, R., Double, M., Orr, K., & Dawson, R. J. G. (1998). A DNA test to sex most birds. *Molecular Ecology*, *7*, 1071–1075.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS*, *Genet*, 5.

Haldane, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. *Journal of genetics*, *12*(2), 101–109.

Harris, R. S. (2007) *Improved pairwise alignment of genomic DNA*. PhD thesis. The Pennsylvania State University.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics*, *210*, 315–330.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*, 115–121.

Janoušek, V., Fischerová, J., Mořkovský, L., Reif, J., Antczak, M., Albrecht, T., & Reifová, R. (2019). Postcopulatory sexual selection reduces Z-linked genetic variation and might contribute to the large Z effect in passerine birds. *Heredity*, *122*, 622–635.

Jeffries, D. L., Copp, G. H., Handley, L. L., Håkan Olsén, K., Sayer, C. D., & Hänfling, B. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, *25*, 2997–3018.

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., … ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*, 203–206.

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2, identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*, 3435–3436.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD, analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356.

Kozma, R., Melsted, P., Magnússon, K. P., & Höglund, J. (2016). Looking into the past - The reaction of three grouse species to climate change over the last million years using whole genome sequences. *Molecular Ecology*, *25*, 570–580.

Leroy, T., Anselmetti, Y., Tilak, M.-K., Berard, S., Csukonyi, L., Gabrielli, M., … Nabholz, B. (2019). A Bird's white-eye view on avian sex chromosome evolution. *bioRxiv*, *505610*, 1–50.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079.

Lima, T. G. (2014). Higher levels of sex chromosome heteromorphism are associated with markedly stronger reproductive isolation. *Nature Communications*, *5*, 4743.

Macholán, M., Baird, S. J. E., Dufková, P., Munclinger, P., Bímová, B. V., & Piálek, J. (2011). Assessing multilocus introgression patterns: A case study on the mouse X chromosome in central europe. *Evolution*, *65*, 1428–1446.

Masly, J. P., & Presgraves, D. C. (2007). High-resolution genome-wide dissection of the two rules of speciation in Drosophila. *PLoS Biology*, *5*, 1890–1898.

Milá, B., Warren, B. H., Heeb, P., & Thébaud, C. (2010). The geographic scale of diversification on islands, genetic and morphological divergence at a very small spatial scale in the Mascarene grey white-eye (Aves, *Zosterops borbonicus*). *BMC Evolutionary Biology*, *10*, 158.

Moore, W. S. (1995). Inferring phylogenies from mtDNA variation, mitochondrial-gene trees versus nuclear-gene trees. *Evolution*, *49*, 718–726.

Mořkovský, L., Janoušek, V., Reif, J., Rídl, J., Pačes, J., Choleva, L., … Reifová, R. (2018). Genomic islands of differentiation in two songbird species reveal candidate genes for hybrid female sterility. *Molecular Ecology*, *27*, 949–958.

Muller, H. J. (1940). Bearing of the Drosophila work on systematics. In J. Huxley (Ed.), *The new systematics* (pp. 185–268). London, UK: Oxford University Press.

Muller, H. J. (1942). Isolating mechanisms, evolution and temperature. *Biology Symposium*, *6*, 71–125.

Nadeau, N. J., Mundy, N. I., Gourichon, D., & Minvielle, F. (2007). Association of a single-nucleotide substitution in TYRP1 with roux in Japanese quail (*Coturnix japonica*). *Animal Genetics*, *38*, 609–613.

Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics, an empirical study from an Amazonian plant species. *Molecular Ecology Resources*, *17*, 1136–1147.

Nosil, P., Harmon, L. J., & Seehausen, O. (2009). Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution*, *24*, 145–156.

Orr, H. A. (1996). Dobzhansky, Bateson, and the genetics of speciation. *Genetics.*, *144*(4), 1331–1335.

Orr, H. A. (1997). Haldane's rule. *Annual Review of Ecology and Systematics*, *28*(1), 195–218.

Otto, S. P., & Day, T. (2007). *A biologist's guide to mathematical modeling in ecology and evolution*, Vol. *13*. Princeton, NJ: Princeton University Press.

Pala, I., Naurin, S., Stervander, M., Hasselquist, D., Bensch, S., & Hansson, B. (2012). Evidence of a neo-sex chromosome in birds. *Heredity*, *108*, 264–272.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*, 2074–2093.

Poelstra, J. W., Vijay, N., Hoeppner, M. P., & Wolf, J. B. W. (2015). Transcriptomics of colour patterning and coloration shifts in crows. *Molecular Ecology*, *24*, 4617–4628.

Pool, J. E., & Nielsen, R. (2007). Population size changes reshape genomic patterns of diversity. *Evolution*, *61*, 3001–3006.

Pryke, S. R. (2010). Sex chromosome linkage of mate preference and color signal maintains assortative mating between interbreeding finch morphs. *Evolution*, *64*, 1301–1310.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools, a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842.

Qvarnström, A., & Bailey, R. I. (2009). Speciation through evolution of sex-linked genes. *Heredity*, *102*, 4–15.

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M. … Westram, A. M. (2017). Interpreting the genomic landscape of speciation, finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*, 1450–1477.

Robinson, J. D., Coffman, A. J., Hickerson, M. J., & Gutenkunst, R. N. (2014). Sampling strategies for frequency spectrum-based population genomic inference. *BMC Evolutionary Biology*, *14*, 1–16.

Saether, S. A., Saetre, G.-P., Borge, T., Wiley, C., Svedin, N., Andersson, G., … Qvarnstrom, A. (2007). Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. *Science*, *318*, 95–97.

Sætre, G. P., & Sæther, S. A. (2010). Ecology and genetics of speciation in Ficedula flycatchers. *Molecular Ecology*, *19*, 1091–1106.

Safran, R. J., Scordato, E. S. C., Symes, L. B., Rodríguez, R. L., & Mendelson, T. C. (2013). Contributions of natural and sexual selection to the evolution of premating reproductive isolation, A research agenda. *Trends in Ecology & Evolution*, *28*, 643–650.

Schwartz, S., Kent, W., & Smit, A. (2003). Human–mouse alignments with BLASTZ. *Genome Research*, *13*, 103–107.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., … Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, *15*, 176–192.

Servedio, M. R., & Noor, M. A. F. (2003). The role of reinforcement in speciation, theory and data. *Annual Review of Ecology Evolution and Systematics*, *34*, 339–364.

Seutin, G., White, B. N., & Boag, P. T. (1991). Preservation of avian blood and tissue samples for DNA analyses. *Canadian Journal of Zoology*, *69*, 82–90.

Storchová, R., Gregorová, S., Buckiová, D., Kyselová, V., Divina, P., & Forejt, J. (2004). Genetic analysis of X-linked hybrid sterility in the house mouse. *Mammalian Genome*, *15*, 515–524.

Storchová, R., Reif, J., & Nachman, M. W. (2010). Female heterogamety and speciation, Reduced introgression of the z chromosome between two species of nightingales. *Evolution*, *64*, 456–471.

Svedin, N., Wiley, C., Veen, T., Gustafsson, L., & Qvarnstrom, A. (2008). Natural and sexual selection against hybrid flycatchers. *Proceedings of the Royal Society B-Biological Sciences*, *275*, 735–744.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*, 585–595.

Takezaki, N., Nei, M., & Tamura, K. (2010). POPTREE2, Software for constructing population trees from allele frequency data and computing other population statistics with windows interface. *Molecular Biology and Evolution*, *27*, 747–752.

Tuttle, E. M., Bergland, A. O., Korody, M. L., Brewer, M. S., Newhouse, D. J., Minx, P., … Balakrishnan, C. N. (2016). Divergence and functional degradation of a sex chromosome-like supergene. *Current Biology*, *26*, 344–350.

Van Belleghem, S. M., Baquero, M., Papa, R., Salazar, C., McMillan, W. O., Counterman, B. A., … Martin, S. H. (2018). Patterns of Z chromosome divergence among Heliconius species highlight the importance of historical demography. *Molecular Ecology*, *27*, 3852–3872.

Warren, B. H., Bermingham, E., Prys-Jones, R. P., & Thébaud, C. (2006). Immigration, species radiation and extinction in a highly diverse songbird lineage, white-eyes on Indian Ocean islands. *Molecular Ecology*, *15*, 3769–3786.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. D. W., Künstner, A., … Wilson, R. K. (2010). The genome of a songbird. *Nature*, *464*, 757–762.

Willing, E. M., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers. *PLoS ONE*, *7*, 1–7.

Wolf, J. B. W., & Ellegren, H. (2016). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, *18*, 87–100.

Wu, C. I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*, 851–865.

Zhang, J., Li, Y., Wu, Y., Yang, T., Yang, K. E., Wang, R., … Guo, H. (2013). Wnt5a inhibits the proliferation and melanogenesis of melanocytes. *Internation Journal of Medical Sciences*, *10*, 699–706.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Bourgeois YXC, Bertrand JAM, Delahaie B, Holota H, Thébaud C, Milá B. Differential divergence in autosomes and sex chromosomes is associated with intra-island diversification at a very small spatial scale in a songbird lineage. *Mol Ecol*. 2020;29:1137–1153. https://doi.org/10.1111/mec.15396

Balancing Selection for Pathogen Resistance Reveals an Intercontinental Signature of Red Queen Coevolution

Yann Bourgeois,*,† Peter D. Fields, Gilberto Bento, and Dieter Ebert*

Zoology, Department of Environmental Sciences, University of Basel, Basel, Switzerland

†Present address: School of Biological Sciences, University of Portsmouth, Portsmouth, United Kingdom

***Corresponding authors:** E-mails: yann.x.c.bourgeois@gmail.com; dieter.ebert@unibas.ch.

Associate editor: Kelley Harris

Abstract

The link between long-term host–parasite coevolution and genetic diversity is key to understanding genetic epidemiology and the evolution of resistance. The model of Red Queen host–parasite coevolution posits that high genetic diversity is maintained when rare host resistance variants have a selective advantage, which is believed to be the mechanistic basis for the extraordinarily high levels of diversity at disease-related genes such as the major histocompatibility complex in jawed vertebrates and R-genes in plants. The parasites that drive long-term coevolution are, however, often elusive. Here we present evidence for long-term balancing selection at the phenotypic (variation in resistance) and genomic (resistance locus) level in a particular host–parasite system: the planktonic crustacean *Daphnia magna* and the bacterium *Pasteuria ramosa*. The host shows widespread polymorphisms for pathogen resistance regardless of geographic distance, even though there is a clear genome-wide pattern of isolation by distance at other sites. In the genomic region of a previously identified resistance supergene, we observed consistent molecular signals of balancing selection, including higher genetic diversity, older coalescence times, and lower differentiation between populations, which set this region apart from the rest of the genome. We propose that specific long-term coevolution by negative-frequency-dependent selection drives this elevated diversity at the host's resistance loci on an intercontinental scale and provide an example of a direct link between the host's resistance to a virulent pathogen and the large-scale diversity of its underlying genes.

***Key words:* *Daphnia magna*, coevolution, *Pasteuria ramosa*, negative frequency-dependent selection, Red Queen, population genomics.**

Introduction

Hosts and parasites engage in specific interactions that are believed to select for and maintain genetic diversity at host resistance genes (Sackton et al. 2007; Ebert and Fields 2020; Radwan et al. 2020). If pathogens evolve to overcome the resistance of common host alleles, rare resistance alleles have a selective advantage until they also become common. This form of time-lagged negative-frequency-dependent selection (NFDS), often referred to as Red Queen coevolution, is believed to increase genetic polymorphism at loci that interact with the antagonist (Charlesworth 2006; Thrall et al. 2015; Rabajante et al. 2016). Indeed, the Red Queen hypothesis has gained so much popular support that regions in host genomes that show elevated genetic diversity are taken as potential indicators of antagonistic coevolution, even when the coevolving antagonists are unknown. The Red Queen model was originally conceived to be a process that acts within populations, but host–parasite interactions undergoing NFDS also shape genetic diversity among populations (reviewed in Ebert and Fields [2020]). Because resistance alleles that migrate into host populations are rare, they may be favored by selection, resulting in a higher effective migration rate than other alleles in the genome (Charlesworth et al. 1997; Thrall et al. 2012; Jousimo et al. 2014; Bolnick and Stutz 2017). Nevertheless, the random loss of genotypes in small populations and strong selection from local parasites can also quickly lead to genetic divergences between neighboring populations (Lively and Dybdahl 2000; Bourgeois et al. 2017). Given this combination of regional and local dynamics, even nearby populations can display high divergence at resistance loci, whereas distant populations may show low divergence (Charlesworth et al. 1997). On large geographic scales, thus, one would expect genomic regions with resistance loci involved in coevolution to display signatures of higher genetic diversity than the rest of the genome, balancing selection, and reduced spatial structure. Evidence for these predictions has been found in the vertebrate MHC loci (Eizaguirre et al. 2012; Kaufman 2018) and in R-genes in plants (Bergelson et al. 2001; Tellier and Brown 2011), although, for both these groups of genes, the functional link between the resistance genes and the long-term coevolving parasites is missing. In other systems, although coevolutionary dynamics between hosts and specific parasites have been demonstrated, the underlying genetics are not known (Thrall et al. 2012; Gibson et al. 2018).

Here we test the hypothesis that host–parasite coevolution causes balancing selection at a host resistance gene cluster in the water flea *Daphnia magna*, coevolving with the

Article

obligate bacterial endoparasite *Pasteuria ramosa*. In this system, both the host and the parasite have a wide natural distribution covering nearly the entire Holarctic. Infections bear extreme fitness costs for the host (Luijckx et al. 2012). Resistance follows a matching allele model, preventing individual hosts and parasite genotypes from reaching fixation (Luijckx et al. 2013), and displays high diversity within populations (Andras and Ebert 2013). Coevolution has been indicated in this system based on a study of sediment cores showing the temporal dynamics of *D. magna*–*P. ramosa* interactions over about three decades (Decaestecker et al. 2007). To test for predicted patterns of genetic diversity within and between populations, we used a panel of *D. magna* genotypes consisting of single clonal lines collected from 125 populations in Eurasia and North Africa (fig. 1*A*), each with information about geographic origin and genome sequences. Notably, for each host genotype we also possessed resistance phenotype data for five parasite genotypes. To test for signatures of balancing selection, we analyzed patterns of diversity at both the phenotypic and genetic level. The latter focused especially on a genomic region in *D. magna* that explains the most variance in its resistance to *Pasteuria* (Bento et al. 2017). This region (positions 1,368,860 to 1,506,215 on scaffold00944 of the *D. magna* reference genome, version 2.4, here called "resistance QTL") contains a supergene that has been found to harbor extremely diverged haplotypes (Bento et al. 2017). Evidence from several sources suggests that this region plays a role in resistance to *P. ramosa* in natural populations. Genome scans for selection and association show a significant signal for this cluster and its flanking genomic regions across European *D. magna* populations (Bourgeois et al. 2017). The same association signal is found within a single panmictic population in Switzerland (Ameline et al. 2021). It has historically been difficult to establish a functional link between resistance, genetic diversity, and the consequences for coevolution, as the underlying genes of either the coevolving parasites or the host were unknown. Nevertheless, the architecture of resistance to *P. ramosa* has now been characterized for *D. magna* (Routtu and Ebert 2015; Bento et al. 2017; Bourgeois et al. 2017; Ameline et al. 2021), making it possible to test directly for a signature of balancing selection at this region and understand how coevolution with a virulent and widespread parasite affects host genetic variability.

Results and Discussion

No Geographical Structure for Resistance Phenotypes

Parasite-driven NFDS is expected to result in a geographic mosaic of resistance phenotypes and genotypes across the host's range with no or weak geographic structure as compared with the genetic background (Kaltz and Shykoff 1998; Tellier and Brown 2011; Ebert and Fields 2020). To test this, we investigated whether polymorphism for resistance phenotypes displayed a signal of spatial structure. We isolated five parasite strains (*P. ramosa* C1, C19, P15, P20, and P21) from natural populations across Europe (Luijckx et al. 2011) and phenotyped *D. magna* clones for resistance to these by assessing whether labeled spores attached to the host's foregut (all five parasites) or hindgut (two parasites: P15 and P21) (Duneau et al. 2011) (supplementary table S1, Supplementary Material online). This resulted in seven different resistance phenotypes, which we summarized with 7-letter codes (R for resistance and S for susceptibility for each of the five foregut and two hindgut phenotypes; fig. 1*A*). Note that this phenotypic assessment covers only a fraction of the total phenotypic variation and should be seen as a sample for the actual diversity in parasites and host resistotypes. Resistance phenotypes were found to be uniformly distributed across the entire study region without a pattern of isolation by distance (IBD; fig. 1*A* and supplementary table S2, Supplementary Material online). This was further confirmed by a global Distance-based Moran's eigenvector maps (dbMEM) analysis, which did not detect any significant positive spatial correlation in the spatial repartition of the seven resistotypes ($_{adj}R^2 = 0.006$, $P = 0.068$). The same observation held when considering each resistotype independently ($_{adj}R^2$ between -0.12 and 0.002, all $P > 0.1$). To understand the biogeographic context for this absence of a spatial pattern, we compared this analysis to a similar analysis using single-nucleotide polymorphism (SNP) data derived from the genomic sequences of the 125 *D. magna* clones. We found a strong pattern of IBD for genomic data, where average relatedness between individual host clones decreased with geographic distance ($N = 125$, Mantel $R = -0.56$, 1,000,000 permutations, $P < 10^{-6}$). This pattern is consistent with a previous study of the same *Daphnia* species (Fields et al. 2015). Moreover, in *D. magna*, other phenotypic traits show a clear geographic structure (Yampolsky et al. 2014; Seefeldt and Ebert 2019), underscoring that the lack of geographic structure for *Pasteuria* resistance is unique.

We further examined whether resistance polymorphism (only to *Pasteuria* C1, C19, P15, and P20). The P21 isolate was isolated only later also held true on the single-population scale. To do so, we obtained resistance phenotypes for *D. magna* individuals hatched from resting eggs from 23 populations from the Western Palaearctic for which we could successfully phenotype at least five host genotypes—20 of them polymorphic (R or S) for at least one *Pasteuria* strain. There was no correlation between variation in resistotype frequencies and pairwise distance (fig. 1*C*) based on Mantel tests (supplementary table S3, Supplementary Material online). In contrast, SNP data across the genomes revealed strong IBD for the same 23 populations (fig. 1*B* and supplementary table S3, Supplementary Material online). This lack of positive spatial correlation was confirmed by a dbMEM analysis ($_{adj}R^2 < 0$). Thus, phenotypic diversity for resistance against *Pasteuria* infections, which did not show a spatial pattern, contrasted strongly with the genomic background, which was shaped by IBD. This uniform diversity in resistance on a very large geographic scale (entire Palaearctic) coincides with the theory of host–parasite coevolution by balancing selection, which projects that phenotypic diversity is maintained at loci of functional importance for the interaction of the antagonists. To our knowledge, this finding has not been shown before for phenotypic traits under coevolution.

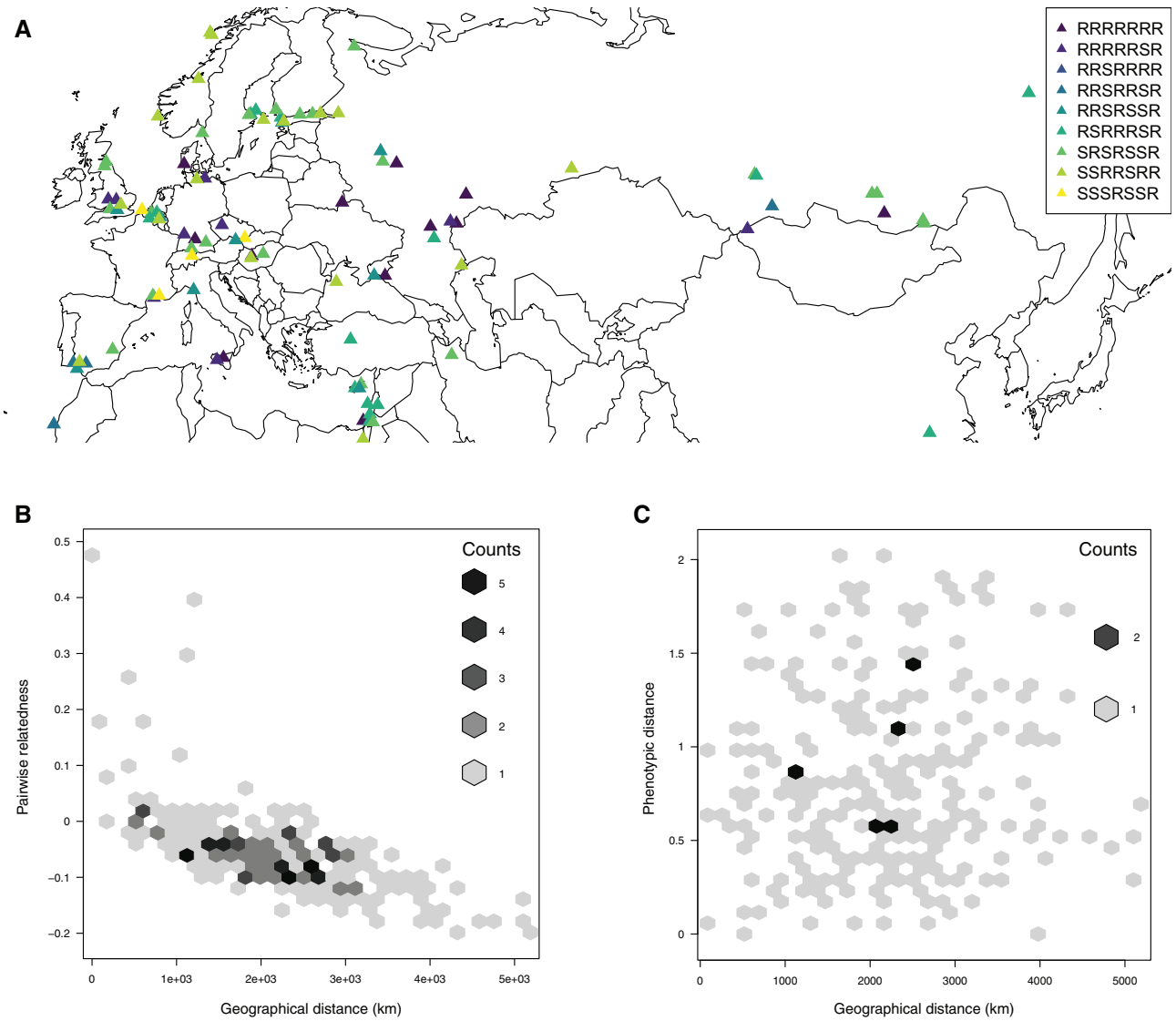


Fig. 1. (*A*) Resistotypes designations for the 125 *Daphnia magna* clones from across Eurasia and North Africa used in this study. Seven-letter codes indicate R (resistant to spore attachment) or S (susceptible) for the following parasite clones (in order): C1, C19, P15 (hindgut attachment), P15 (foregut attachment), P20, P21 (hindgut attachment), and P21 (foregut attachment). To improve readability, only resistotypes found at least four times are shown. (*B*) Plot of relatedness using genomic SNP data for 23 clones sampled from the same populations as in B against their pairwise geographic distance. Counts indicate overlaying data points. (*C*) Plot of pairwise geographic distance and pairwise distance of resistance phenotypes for 23 *D. magna* populations. Phenotypic distance is measured as the pairwise Euclidean distance incorporating population differences in the frequencies of resistotypes.

Genomic Data Reveal a Signature of Population Structure and Postglacial Expansion

To investigate patterns of genetic diversity and divergence, population structure and history should be taken into account. Using a discriminant analysis on principal components (DAPC; Jombart et al. 2010) based on genotypes (supplementary fig. S1*A*, Supplementary Material online) for all 125 *D. magna* clones, we identified three geographic clusters. As the model with three clusters had the lowest Bayesian Information criterion (BIC), we assigned individuals to three putative geographical clusters, which we called, for simplicity, Europe+, Middle East, and East-Asian ($N = 100$, 11, and 14, respectively). Our results supported previous studies in revealing substantial divergence between East-Asian samples and the other groups (Fields et al. 2015, 2018); indeed the East-Asian cluster was clearly separated from other clones by the first discriminant function with the highest eigenvalue (fig. 2*A*). In addition, estimates of differentiation measured by F_{ST} over 1-kb windows along the genome (see Materials and Methods) were also substantially higher for the East-Asian cluster (average $F_{ST} = 0.32$, 0.37, and 0.124 for Europe+ vs. East-Asia, Middle-East vs. East-Asia, and Europe+ vs. Middle-East, respectively, Wilcoxon signed rank tests, all $P < 2.2 \times 10^{-16}$).

Past reductions in effective population size can produce genome-wide signatures that are similar to balancing selection (Charlesworth 2006). Indeed, demographic analyses reveal a clear signature of expansion and population splits

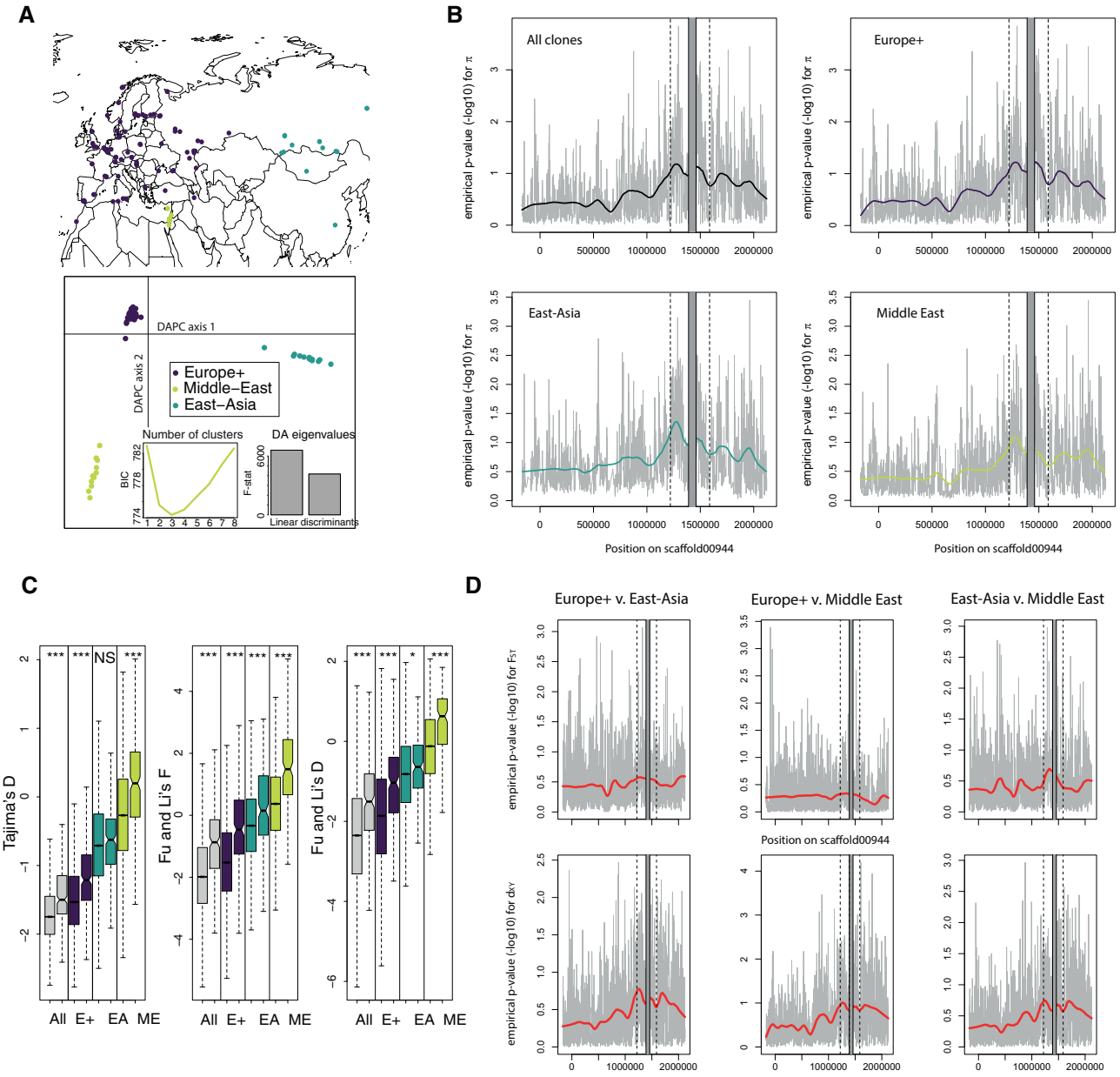


FIG. 2. Genetic diversity and population genetic parameters in the genomic region flanking the *D. magna*'s resistance QTL. (*A*) Sites of origin and DAPC on 8,978 genome-wide SNPs with no missing data sampled every kb for 125 *D. magna* genotypes. The DAPC analysis identified three major groups: Europe+ (E+), East-Asia (EA), and Middle-East (ME). (*B*) Empirical *P* values for nucleotide diversity in 1-kb windows for all 125 *D. magna* clones and the three geographic groups. Diversity statistics are ranked in decreasing order to obtain *P* values, so low *P* values correspond to high diversity. The resistance supergene region (QTL locus ± 100 kb) is located between the two dotted lines. The supergene itself is masked in gray due to very poor mapping of short reads to this region (positions 1,435,000 to 1,490,000 on scaffold00944). Coordinates correspond to *D. magna* 2.4 genome. Negative coordinates correspond to a region in the PacBio scaffold that mapped outside the original scaffold00944 (see supplementary fig. S1, Supplementary Material online). (*C*) Neutrality statistics (over 1-kb windows) in the region around the resistance supergene compared with genome-wide values (excluding scaffolds shorter than 10 kb in genome version 2.4). In all pairwise comparisons, the boxplots on the left and right correspond to the genomic background and the region around the resistance supergene, respectively. For Fu and Li's *F* and Fu and Li's *D*, *Daphnia similis* was used as an outgroup; higher values are associated with frequency spectra skewed toward ancestral variants and alleles at intermediate frequencies, supporting balancing selection. *P* values were obtained from Wilcoxon rank-sum tests (NS: nonsignificant; *: $P < 0.05$; ***: $P < 0.001$). Color codes as in figure 2. (*D*) Empirical *P* values for divergence statistics. The upper panels show the F_{ST}, which is expected to be reduced if balancing selection is present, for all three pairwise comparisons among the geographic regions Europe+, East-Asia, and Middle-East. In that case, F_{ST} values are ranked in increasing order to obtain the empirical *P* value. The lower panel shows the absolute divergence, d_{xy}, for the same pairs, which is expected to increase if there are ancient polymorphic alleles.

following the last glacial maximum (supplementary fig. S1 and table S4, Supplementary Material online for exact point estimates and confidence intervals), which coincides with previous studies based on mitochondrial data (Fields et al. 2018). Such demographic events are thought to skew the genome-wide allele frequency spectrum (AFS) toward more

rare alleles, whereas balancing selection would maintain alleles at higher frequency. The absence of a strong recent bottleneck suggests that false positive evidence for balancing selection due to demography most likely do not explain the patterns we observed near the resistance QTL.

High Nucleotide Diversity and Skewed Allele Frequency Spectra near the Resistance Locus

To first assess whether the region around the resistance QTL displayed elevated nucleotide diversity, as would be expected under balancing selection (Charlesworth 2006), we improved the quality of scaffold00944 by using a PacBio contig from the same individual used to build the reference genome (supplementary fig. S2, Supplementary Material online), as described in a previous study (Bento et al. 2017). Nevertheless, divergence between variants within the supergene (roughly located between positions 1,435,000 and 1,490,000 on scaffold00944) were so high that alignment of short Illumina reads on the reference was not possible. Because the supergene haplotypes are not homologous and are difficult to assemble due to their repeat richness, typical population genetic approaches that rely on the alignment of diverged haplotypes are impossible for this region (Bento et al. 2017). In all subsequent analyses, thus, we excluded the supergene region to avoid the unreliable mapping of reads and instead focused on its flanking regions. As a result, our divergence measures and other population genetic summary statistics based on SNP variation in these flanking regions strongly underestimate the actual polymorphism in the region of highest diversity. All geographic clusters displayed a considerable increase in nucleotide diversity for 1-kb windows between positions encompassing the resistance QTL and the following 1 Mb (fig. 2*B*), with most windows in the top 5–10% genome-wide, and peaks in the top 0.1%. For all clones as well as for the three geographic clusters, nucleotide diversity was higher at the resistance region (hereafter defined as the resistance QTL ±100 kb), than the rest of the genome (Wilcoxon rank-sum tests, all $P < 2.2 \times 10^{-16}$).

We then tested the hypothesis that alleles with intermediate frequencies should be more common in genomic regions under balancing selection than in the genomic background (Charlesworth 2006). Indeed, the resistance region showed an abundance of alleles at intermediate frequencies with a significantly elevated Tajima's *D* (fig. 2). Using the closely related species *Daphnia similis* (Cornetti et al. 2019) as an outgroup to define ancestral alleles, we further found elevated Fu and Li's *F and D* (Fu and Li 1993), indicating an excess of ancestral polymorphisms at intermediate frequencies (fig. 2*C*). This supports the hypothesis that balancing selection acts at the resistance region. We also note that genome-wide values for Tajima's *D* were generally negative for the Europe+ and East-Asia clusters, and closer to 0 for the Middle-East cluster, consistent with the recent expansion inferred by our demographic analyses and evidence on mitochondrial genomes (Fields et al. 2018). The glacial refugium for the European *D. magna* was suggested to be in South-Eastern Europe/Middle East (Fields et al. 2018).

Low Relative but High Absolute Measures of Spatial Differentiation

Another hallmark of balancing selection is reduced differentiation at selected loci among populations at large spatial scales. Migrating alleles at loci under balancing selection are likely to be rare upon arrival, giving them an advantage and, as a consequence, increases their effective migration rate. Neutral alleles, on the other hand, would only increase in the recipient population if they hitchhike with alleles under selection (Laine et al. 2011; Thrall et al. 2012; Phillips et al. 2018; Ebert and Fields 2020). In addition, alleles under balancing selection are less likely to go extinct in a given population because they are advantageous when rare. Balancing selection, thus, can be expected to reduce the turnover rate of alleles and to facilitate long-term persistence of polymorphism within populations (Charlesworth 2006; Leffler et al. 2013). Distinct populations would share polymorphisms and show reduced estimates of population differentiation at the genes under selection (Charlesworth 2006; Rico et al. 2015). To test this theory, we estimated genome-wide variation in relative (F_{ST}) and absolute (d_{XY}) differentiation using the three geographic clusters determined by the DAPC (fig. 2*A*). Introgression and selection are both expected to impact population differentiation and estimates of F_{ST} *and* d_{XY} in different ways: balancing selection should decrease F_{ST} and increase d_{XY}, whereas recent introgression would reduce both statistics, as it is an absolute measure of divergence that captures the number of sequence differences since the most recent common ancestor (TMRCA) (Cruickshank and Hahn 2014). Our data clearly support balancing selection, showing reduced relative population divergence (significantly lower F_{ST} for Europe+ vs. East-Asia and East-Asia vs. Middle-East comparisons; Wilcoxon rank-sum test, all $P < 2.2 \times 10^{-16}$, fig. 2*D*), and increased absolute population divergence in the resistance region (higher d_{XY} compared with genome-wide mean [all $P < 2.2 \times 10^{-16}$, fig. 2*D*]). This suggests that alleles in the resistance region have a higher chance of spreading across populations and being maintained within populations, consistent with an advantage for being rare.

Observed Patterns of Diversity Are Consistent with Simulations under Negative-Frequency-Dependent Selection but Not with Neutrality

To test whether differentiation and diversity statistics deviated significantly from neutral expectations, we generated 10 million coalescent simulations under our demographic model. We then performed a principal component analysis (PCA) on summary statistics (fig. 3*A*). The first PC axis (PC1) explains 42% of the total variance, and is clearly correlated with diversity statistics and Tajima's *D*. The second axis (PC2) explains 18% of the variance and is mostly correlated with F_{ST}. Predicted values for the resistance region are consistent with our previous observations, with high PC1 scores (high diversity and Tajima's *D*), and low PC2 scores (low F_{ST}). There is a high density of windows deviating from neutrality in the

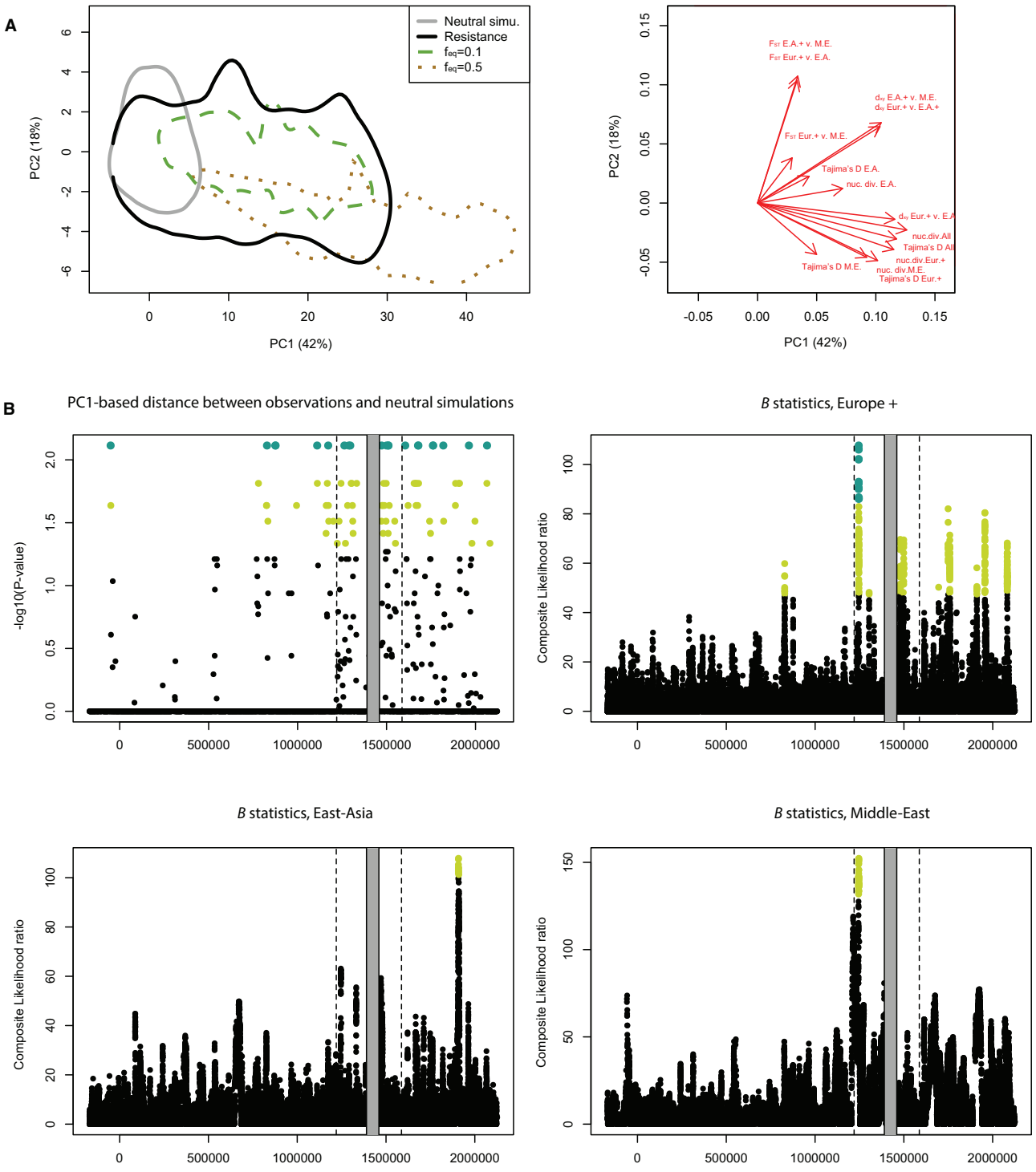


Fig. 3. Comparisons of diversity between the resistance region, simulations, and the rest of the genome. (*A*) Principal components analysis (PCA) of 10 million neutral coalescent simulations. The statistics used include nucleotide diversity, pairwise divergence statistics, and Tajima's *D* (correlation circle displayed in the right panel). Predicted values for the resistance region and two sets of SLiM3 NFDS simulations are also shown. The SLiM3 simulations were obtained with a fraction of new mutations recruited by selection of 0.1% and equilibrium frequencies (f_{eq}) of 10% and 50%. The envelopes cover 95% of points from each category. (*B*) The upper-left panel shows Bonferroni-corrected *P* values obtained from comparing observations and neutral coalescent simulations for each 1-kb window. Light green points indicate *P* values < 0.05 and large dark green dots indicate *P* values < 0.01. The three other panels show the $B_{0,MAF}$ statistics. Composite likelihood ratio for each of the three geographic groups. The statistics compares local allele frequency spectra to the genome-wide spectrum and compares the likelihood of a model with balancing selection against a neutral model. Light green points indicate the highest 1% of scores genome wide, whereas large dark green dots indicate those among the top 0.5%.

resistance region (fig. 3*B*). This is further confirmed by a scan for balancing selection using the $B_{0,MAF}$ statistics (Cheng and Degiorgio 2020). The test contrasts allele frequency spectra around a focal SNP to the genome-wide frequency spectrum to estimate the likelihoods of models with and without balancing selection, under a broad range of equilibrium frequencies. We observe clear signals in Europe+ and Middle-East clusters (fig. 3*B*), with many regions above the top 1% genome-wide threshold.

To test the conditions under which NFDS could produce the observed patterns of diversity, we ran simulations using the forward-in-time simulator SliM3 on the demographic history estimated from whole-genome data. We simulated 1-kb windows with mutation and recombination rates consistent with current knowledge about *D. magna*. We varied the fraction of new mutations recruited by selection and the equilibrium frequency at which balanced polymorphisms are maintained. As expected in NFDS simulations, nucleotide diversity, Tajima's *D*, and d_{XY} were higher than in neutral simulations, whereas F_{ST} was lower (fig. 3*A* and supplementary fig. S3, Supplementary Material online). In scenarios where 0.01% of new mutations were recruited by selection, diversity was generally lower than our observations in the resistance region. However, the skew in the frequency spectrum (Tajima's *D*) was consistent with observations. A closer match between simulations and observed data occurred in scenarios where 0.1% of new mutations were recruited by selection with an equilibrium frequency of 0.1 (fig. 3*A*). For higher equilibrium frequencies, Tajima's *D* values were much higher than our observations (supplementary fig. S3, Supplementary Material online), consistent with a stronger skew of the AFS toward higher frequencies. This suggests that balanced polymorphisms in the resistance region may not necessarily reach very high frequencies at the geographical scale considered here, which is consistent with a fast tracking of host genotypes by quickly evolving pathogens that rapidly reduce the selective advantage of the most common resistotypes (Decaestecker et al. 2007).

It is important to note that we scaled down effective population sizes and times by 100 in our simulations to ensure fast running times, which limits the maximum strength of selection that we could simulate. We assumed a selective coefficient of 0.005 for newly established mutations under NFDS before scaling parameters. In actual populations, the strength of selection is likely higher, as *P. ramosa* castrates the host, reducing the residual fitness of the infected female by about 90% (Ebert et al. 2016). Such strong selection should counteract the effects of recombination over large genomic intervals, especially given the large effective population sizes considered here. In that case, an even lower fraction of mutations recruited by selection would be enough to generate the patterns of diversity observed in the resistance region. Although it seems clear that neutrality can be rejected, further simulations at multiple spatial scales would be needed to properly compare various NFDS scenarios. Nevertheless, our simulations suggest that even moderate selection and low equilibrium frequencies can lead to a marked increase in diversity in our system.

The Resistance Region Displays an Excess of Ancient Alleles and Older Coalescence Times

Reduced allele turnover during balancing selection implies that alleles have a longer lifespan. Thus, regions under long-term balancing selection should display older coalescence times and higher local effective population sizes (Charlesworth 2006). Our findings of higher absolute divergence (d_{XY}, fig. 2*C*) in the resistance region suggest that the alleles in this region are already older than alleles in other parts of the genome. We confirmed this by computing ancestral recombination graphs (ARG) (Rasmussen et al. 2014) on a set of three large scaffolds, including the scaffold with the resistance region, retrieving local genealogies at all nonrecombinant blocks. We restricted the analysis to 48 host clones that had fewer than 5% missing genotypes to achieve a high quality of the estimates and to reduce computational burden (we could not handle with the entire genomes of all genotypes). We also included priors on past changes in effective population sizes based on our demographic analyses. The region displaying a local reduction in F_{ST} also displayed longer times to the TMRCA (coalescence time; fig. 4), further supporting long-term balancing selection at this region. This long coalescence time was not driven by a few older haplotypes, since the half coalescence time (HCT; the minimum time at which half of lineages coalesce) was also substantially older than the background (fig. 4).

Higher diversity near the resistance region may be caused by the reduced influence of linked selection if recombination rates are particularly high (Charlesworth 2013; Cruickshank and Hahn 2014; Burri 2017). While in the upper range when compared with other scaffolds, recombination rates estimated by ARGWeaver in the resistance region are not extreme (fig. 4 and supplementary fig. S4, Supplementary Material online), suggesting that higher recombination rates and weaker linked selection do not explain the locally high concentration of windows of high diversity.

Genome Scans for Ancient Balancing Selection Pinpoint Genes Involved in Glycolipid and Glycoprotein Synthesis

It is predicted that the flanking regions surrounding a genomic region under balancing selection will show little signature of balancing selection because older polymorphisms give recombination enough time to erode a signal on the flanking regions (Charlesworth 2006). However, our study found clear evidence of signals of balancing selection in the flanking region of the resistance supergene, hinting that further polymorphisms in these flanking regions might be under balancing selection. Furthermore, an earlier resistance mapping study also suggested that genes outside the resistance supergene are linked to phenotypic variation in resistance (Bourgeois et al. 2017). We therefore conducted genome scans of balancing selection using recently developed β statistics (Siewert and Voight 2017) to investigate at a higher resolution which regions near the resistance supergene displayed the strongest signals of selection. β should be sensitive to alleles at an equilibrium allele frequency between 0.2 and

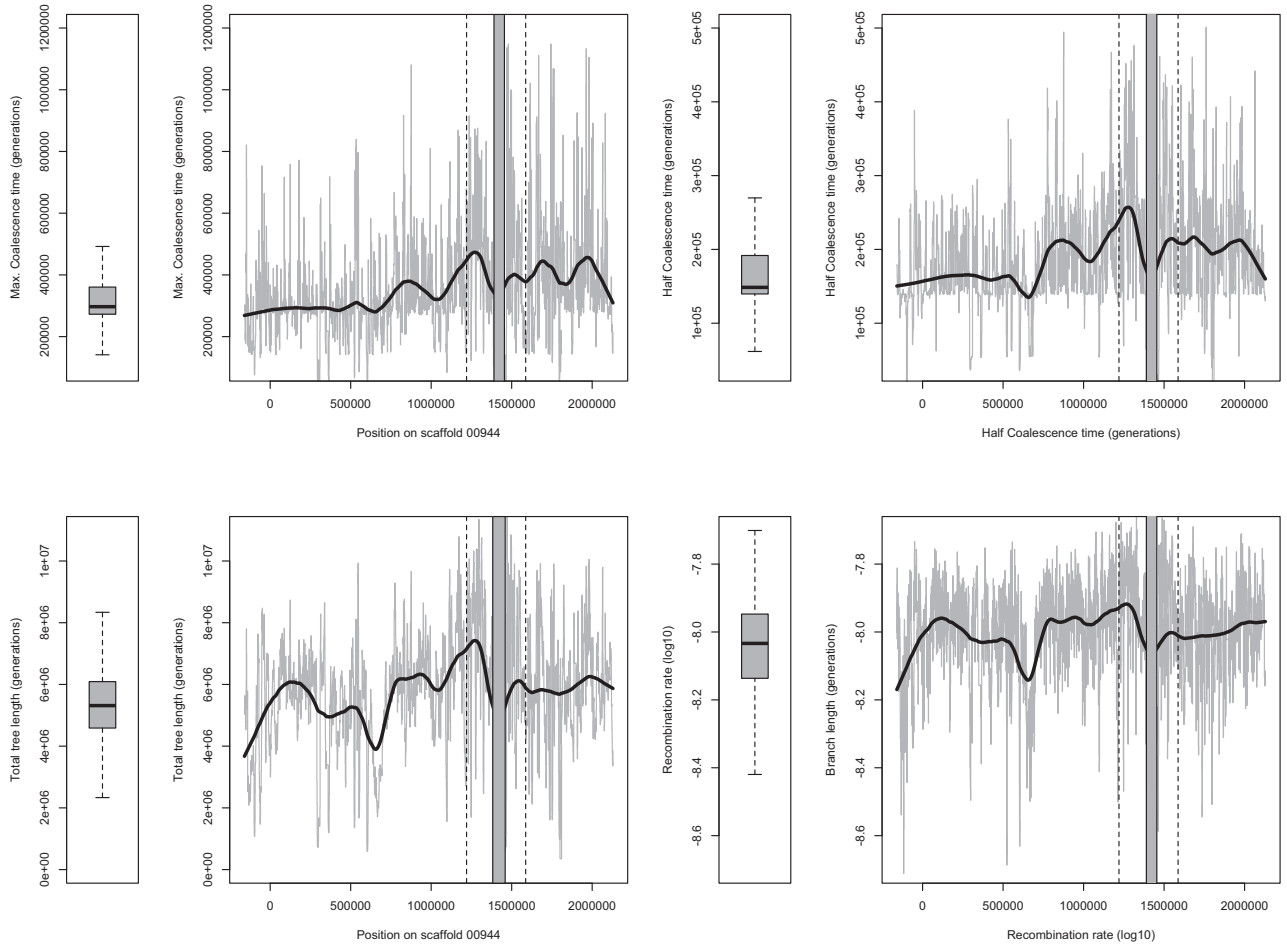


FIG. 4. Coalescence analysis for 1-kb windows across the resistance region (indicated by flanking vertical dotted lines). Coalescence times are given in equivalent generations (sexual + asexual). Approximate times in years can be obtained by dividing by ten, assuming ten generations a year. Boxplots summarize the distribution of statistics from two other large scaffolds (00024 and 00512) totaling more than 6 Mb. Half-coalescence time is defined as the minimum time at which half of the lineages coalesce (see main text). Total tree length corresponds to the average sum of all branches in genealogies of nonrecombining blocks. Recombination rates are estimated by ARGWeaver and log10 transformed.

0.5, covering the range of equilibrium frequencies explored in our simulations. It should also be robust to the lower sample sizes in the Middle-East and East-Asia clusters (Siewert and Voight 2017). We also expect β to be more powerful than other statistics such as *NCD* (Bitarello et al. 2018), which mostly target alleles with equilibrium frequencies in the 0.3–0.5 range. We found several signals of balancing selection in the region around the resistance supergene, particularly in the Europe+ lineage where our sample size was largest (fig. 5*A*). We then extracted genealogies from the ARGWeaver output for nonrecombining blocks that overlapped the three main peaks identified in the targeted region (fig. 5*B*). These genealogies displayed very long branches with very old coalescence times (~2 million generations). For two (#1, #2) of the three peaks of beta score close to the resistance QTL, clones from all three geographic regions were represented in both clades. For the third peak (#3), relatedness between non-European clones coincided with the clusters identified by our DAPC analysis, with individuals from the same genetic group clustering together in the phylogeny. This lack of strong geographic structure was common along the resistance region. Correlations between geographic and phylogenetic distance between clones were weaker for hundreds of ARGWeaver trees randomly sampled in the resistance region than for trees sampled in other scaffolds (fig. 5*C*, Wilcoxon rank-sum test, $P < 4.3 \times 10^{-14}$), consistent with the maintenance of ancestral polymorphism in the resistance region.

We then identified which genes near the resistance region were associated with β scores higher than the top 1% genome-wide (table 1). These genes displayed homology with glucosyltransferases, chitinases, and transcription factors, echoing the function of genes previously found in the two haplotypes of the resistance supergene (Bento et al. 2017). Other peaks on the scaffold outside the resistance supergene itself were mostly concentrated between positions 1,820,000 and 2,000,000 on scaffold00944 and overlapped genes with similar annotations (table 1), suggesting that the resistance QTL may belong to a larger region recruited by NFDS. We also found genes annotated as digestive enzymes such as trypsins and serine-proteases, an interesting observation given that *P. ramosa* starts its infection process in the host's oesophagus and hindgut (Duneau et al. 2011; Bento et al. 2020) and

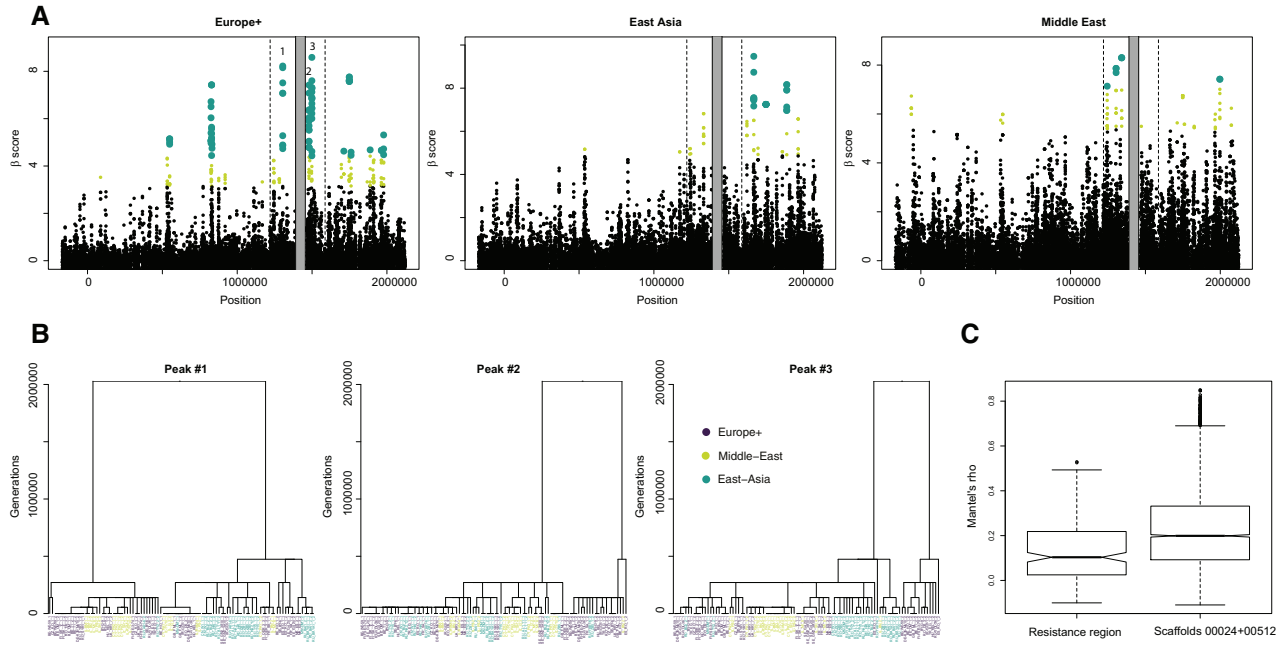


FIG. 5. Scan for balancing selection in the resistance region and flanking sites. (*A*) Results from the Beta scan analysis. Light green points indicate the highest 1% of scores genome wide, whereas large dark green dots indicate those among the top 0.5%. (*B*) Local topologies obtained from ARGWeaver for nonrecombining blocks overlapping with SNPs at the three peaks are highlighted in (*A*). (*C*) Mantel's correlation coefficients obtained by comparing the matrix of geographical distance between clones with 5,000 matrices of phylogenetic distance inferred from 5,000 trees randomly sampled across scaffolds of the genome.

attaches to host cells through collagen-like proteins (Mouton et al. 2009; Andras et al. 2020).

Conclusion

The *D. magna–P. ramosa* system has become a model for the study of antagonistic coevolution in natural population (Mitchell et al. 2004; Decaestecker et al. 2007; Goren and Ben-Ami 2013; Auld et al. 2016). The discovery of a major resistance supergene in the host genome (Routtu and Ebert 2015; Bento et al. 2017) has allowed us to further explore the evolution of this region and to test whether it is under balancing selection, as the model of Red Queen coevolution would predict. Our study demonstrates that the region around this supergene diversity is indeed maintained through balancing selection. Furthermore, variation in this region is ancestral and older than variation in other regions of the genome. For the first time, our analysis links large-scale phenotypic diversity for parasite resistance with the underlying genomic region for a host–parasite system. As in other systems with balancing selection, such as mating types and incompatibility alleles (Joly and Schoen 2011; Roux et al. 2013), we found high phenotypic and genetic diversity combined with the absence of a large-scale geographic pattern. Textbook examples of balancing selection in genes related to immune function typically lack functional evidence of the interaction between host resistance genes and a specific coevolving parasite. Mechanistically, balancing selection at the *Pasteuria* resistance locus may be maintained by a matching allele resistance matrix that links host and parasite genotypes on a functional level (Luijckx et al. 2013; Metzger et al. 2016). The host locus underlaying this matching allele matrix is part of the supergene (Bento et al 2017), which is the center of the current study.

Earlier studies that provide evidence for balancing selection at disease loci in hosts typically only speculate about the coevolving parasite or assume a community of different parasites that vary in space and time (hence diffuse coevolution, see Ebert and Fields 2020). This study tested for balancing selection in a region of the host genome known to interact specifically with the widespread, virulent bacterial parasite *Pasteuria ramosa*, but not with other parasites (Routtu and Ebert 2015; Krebs et al. 2017; Keller et al. 2019). Specific coevolution—which we believe explains the polymorphisms at the resistance supergene region examined here—is the heart of the Red Queen hypothesis of antagonistic coevolution, as it was originally proposed by Clarke (although not under the name Red Queen) (Clarke 1976; Hamilton 1980) and taken up by others (Hamilton 1980; Frank 1991; Tellier and Brown 2007). The simplicity of this model, combined with the fascinating complexity produced by the intricate interactions between antagonists, have made it frequently cited for antagonistic coevolution and further, for phenomena linked to coevolution, such as the evolution of genetic recombination (Hamilton et al. 1990; Lively 2010). The *Daphnia–Pasteuria* system is among the few systems where Red Queen model's assumption of specific coevolution is demonstrated and shown with strong evidence (Decaestecker et al. 2007; Duneau et al. 2011; Luijckx et al. 2011, 2013). Our finding of long-term balancing selection on the *Pasteuria* resistance locus here further reinforces this key prediction of the Red Queen model for specific coevolution, leaving a strong impact on the genome of the host.

Table 1. List of candidate genes with a signature of balancing selection on scaffold00944, highlighting the geographical clusters in which they were identified.

Start	End	Gene Name	Populations with Outlier β Score	Region	ME Max LR	EA Max LR	E+ Max LR
166339	169222	Noncoding RNA	E+	scaffold00944	16.97	44.83	31.92
587440	597064	Putative Beta-1,3-glucosyltransferase	ME; EA	scaffold00944	47.55	32.82	26.54
596562	601020	Noncoding RNA	EA	scaffold00944	47.55	25.30	24.87
597069	598912	Chymotrypsin-2-like	EA	scaffold00944	47.39	25.30	24.87
606651	608619	Noncoding RNA	E+	scaffold00944	33.21	14.06	21.01
866318	872452	Uncharacterized, similar to integumentary mucin C.1 protein (94% coverage, 99% identity, *D. magna*)	E+	scaffold00944	47.30	37.07	59.83
868009	868917	Uncharacterized	E+	scaffold00944	47.30	37.07	59.83
872613	877571	Uncharacterized	E+	scaffold00944	47.30	37.07	59.83
913115	920541	Noncoding RNA	E+	scaffold00944	33.87	19.77	45.27
915177	937472	Putative neuropeptide receptor	E+	scaffold00944	37.66	19.77	45.27
962327	980655	Rap1 GTPase-activating protein	E+	scaffold00944	37.66	17.91	11.74
1198279	1199273	Uncharacterized, similar to protein FAM98B-like (100% coverage, 96% identity, *D. magna*)	E+	scaffold00944	34.00	0.79	33.58
1199954	1206543	Disintegrin and metalloproteinase domain-containing protein 28	EA	scaffold00944	11.90	33.38	8.88
1274156	1284959	Anion exchange protein/Sodium bicarbonate transporter-like protein 11	ME; EA; E+	Resistance region	151.92	63.13	107.69
1308110	1311274	Uncharacterized	E+	Resistance region	32.30	23.83	48.18
1311330	1312503	Uncharacterized	E+	Resistance region	32.30	23.83	48.18
1331506	1334662	Hypothetical, homology with matrix metalloproteinase 1 (70% coverage, 59% identity, *Daphnia pulex*) and Galactose-3-O-sulfotransferase 2 (70% coverage, 43% identity *D. magna*)	ME; E+	Resistance region	32.30	23.83	48.18
1359580	1364008	Uncharacterized, possible homology with matrix metalloproteinase 1 (68% coverage, 58% identity, *Daphnia pulex*)	East	Resistance region	51.48	55.54	34.60
1370390	1372988	Putative metal-responsive transcription factor 1 protein	ME	Resistance QTL	43.42	18.51	10.02
1431210	1433374	Phytanoyl-CoA dioxygenase	ME; EA; E+	Resistance QTL	27.57	28.52	15.70
1494522	1497956	Beta-1,3-*N*-acetylglucosaminyltransferase	ME	Resistance QTL	23.81	59.30	69.64
1501990	1503649	Uncharacterized, similar to *N*-acetylneuraminate 9-O-acetyltransferase-like (79% coverage, 98.6% identity, *D. magna*)	E+	Resistance QTL	23.81	59.30	69.64
1503794	1504979	Alpha1,3 fucosyltransferase	E+	Resistance QTL	23.81	59.30	69.64
1505080	1510969	Putative WSC domain-containing protein 1 (sulfotransferase activity)	E+	Resistance QTL	23.81	59.30	69.64
1518381	1524265	Putative vascular endothelial growth factor receptor 3/brain chitinase and chia	E+	Resistance region	33.53	17.63	69.44
1639127	1641901	Uncharacterized	EA	scaffold00944	60.12	37.24	39.95
1639490	1640158	Noncoding RNA	EA	scaffold00944	60.12	37.24	39.95
1652260	1690450	Uncharacterized	EA	scaffold00944	73.78	43.65	50.26
1678117	1683180	Uncharacterized, similar to trypsin-like isoform X1 (*D. magna*), 100% coverage, 88.7% identity	EA	scaffold00944	73.78	43.65	50.26
1823839	1829476	Popeye domain-containing protein 3	ME	scaffold00944	32.02	10.68	19.79
1854814	1864203	Multidrug resistance-associated protein 7-like	EA	scaffold00944	13.29	27.78	12.63
1867949	1870932	Histone deacetylase 8	E+	scaffold00944	6.99	23.46	22.54

(continued)

Table 1. Continued

Start	End	Gene Name	Populations with Outlier β Score	Region	ME Max LR	EA Max LR	E+ Max LR
1885017	1888137	Clip-domain serine protease, similar to trypsin Blo t 3-like (100% coverage, 96.8% identity, *D. magna*)	E+	scaffold00944	15.86	30.42	36.06
1888238	1902202	High choriolytic enzyme/putative Metalloendopeptidase	E+	scaffold00944	43.75	38.63	18.50
1902029	1906857	Clip-domain serine protease/putative Trypsin-7	E+	scaffold00944	62.49	89.36	33.86
1907645	1910718	Clip-domain serine protease/putative Trypsin-7	E+	scaffold00944	60.72	107.84	58.13
1910898	1913791	High choriolytic enzyme/putative metalloendopeptidase	E+	scaffold00944	53.68	86.37	31.78
1953036	1963323	Lactosylceramide/alpha-1,4-*N*-acetylglucosaminyltransferase	E+	scaffold00944	19.62	39.60	80.49
1963325	1965982	Lactosylceramide. Similar to *N*-acetylneuraminate 9-O-acetyltransferase (99% coverage, 72.7% identity, *D. magna*)	E+	scaffold00944	27.28	48.74	42.63
1966223	1972322	Putative vascular endothelial growth factor, brain chitinase, and chia	ME; E+	scaffold00944	29.91	14.44	18.13
1971615	1981027	Brain chitinase and chia, similar to vascular endothelial growth factor (63% coverage 93.1% identity, *D. magna*)	E+	scaffold00944	30.21	23.25	17.48
1996043	1999375	Putative GMP synthase	ME	scaffold00944	42.88	10.16	13.79
2069556	2075588	Putative eukaryotic translation initiation factor 4B	ME	scaffold00944	60.45	10.05	7.76

Note.—For some uncharacterized proteins, a protein–protein BLAST search was performed at https://blast.ncbi.nlm.nih.gov/Blast.cgi to identify possible homologs. In those cases, we report the percentage of coverage, identity, and the species in which the homolog was found. For each gene, we highlight whether it was found in the original resistance QTL (excluding the supergene), in the region around the resistance supergene (QTL $\pm$ 100 kb), or elsewhere on scaffold00944. For each candidate, we also indicate the maximum value for the $B_{0,MAF}$ statistics composite likelihood ratio in each of the three geographic groups (see also fig. 3).

Materials and Methods

Spatial Variability in Resistance Phenotypes

For each of the 125 *D. magna* clones, seven resistance phenotypes (resistotypes) were obtained using the attachment test (Duneau et al. 2011). Resistance (failure of the parasite to attach to the host cuticle) was coded as "R," and susceptibility (attachment), as "S" (Andras and Ebert 2013; Luijckx et al. 2013, Bento et al. 2020). Resistotypes are defined by the R–S sequence for the seven tests each host clone underwent (at least three replicates per host–parasite combination). Pairwise phenotypic distance between individuals was coded as 0 when resistotypes were the same and 1 when they differed. Pairwise genetic differences between individuals were estimated from genomic data (see below) using the relatedness function in VCFTOOLS v0.1.12b (Danecek et al. 2011), which computes the Ajk statistics (Yang et al. 2010). This statistic should vary between 0 (for pairs of unrelated individuals) and 1 (for an individual with itself). Resistotype frequencies were calculated for 23 population samples (supplementary table S5, Supplementary Material online). Distance measures, including all phenotypes, were estimated as a Euclidean distance, using each resistotype as a distinct dimension. IBD was assessed by Mantel tests with the ecodist package in R (v3.6.3) (Dray and Dufour 2007). dbMEM were also used to assess the geographical variables influencing resistotypes composition at the different spatial scales in our study (Legendre et al. 2015). Resistotypes presence/absence or abundance data were first Hellinger transformed to avoid overweighting rare resistotypes and significant linear trends were removed, using the adespatial package in R.

Whole-Genome Resequencing

Genomic DNA was extracted from the 125 *D. magna* and one *D. similis* clone (three times selfed) as in Fields et al. (2015) (see supplementary table S6, Supplementary Material online for details). Individuals were treated with antibiotics to evacuate their guts and reduce DNA from microbiota and food following the protocol of Dukić et al. (2016). DNA was extracted using an isopropanol precipitation protocol. Paired-end 125 cycle sequencing was performed by the Quantitative Genomics Facility service platform at the Department of Biosystem Science and Engineering (D-BSSE, ETH) in Basel, Switzerland, on an Illumina HiSeq 2000. Read quality was assessed with FastQCv0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc, last accessed August 9, 2021), and Trimmomatic (v0.32) (Bolger et al. 2014) was subsequently used to remove low quality bases,

sequencing adapter contamination and systematic base calling errors. Sequences were aligned using BWA MEM (v0.7.15) on the *D. magna* genome assembly (v. 2.4) (Li et al. 2009). We improved the quality of this reference for the resistance region by replacing a part of the original scaffold944 (length = 2,137,955 bp) with a PacBio contig from the same Xinb3 reference (supplementary fig. S1, Supplementary Material online) as described in a previous study (Bento et al. 2017); however, only minor differences between the analyses using the old and the updated reference were observed. We converted coordinates for the PacBio contig into coordinates on scaffold00944 by carrying a BLAST (v2.6.0) search analysis between the two sequences (Altschul et al. 1997). BAM alignment files were filtered for quality, and PCR duplicates were removed using PICARD tools (v 2.0.1, http://broadinstitute.github.io/picard/, last accessed August 9, 2021). SNP calling was performed simultaneously on all samples using freebayes (v. 0.9.15-1). Freebayes is a haplotype caller, which automatically performs indels realignment and base quality recalibration. VCF files were then filtered using VCFTOOLS v0.1.12b (Danecek et al. 2011) to include SNPs with a minimum quality of twenty, a minimum genotype quality of 30, a minimum depth of coverage of 8X/genotype, and a mean maximum sequencing depth of 70×. We removed polymorphic indels and homopolymers before using the vcfallelicprimitive script in vcflib (https://github.com/vcflib, last accessed August 9, 2021) to convert the haploid calls into pointwise SNPs. Only SNPs that passed filters in at least 90% of samples were included in subsequent analyses (no more than 10% missing data). For simplicity, contigs and scaffolds shorter than 10 kb were excluded from analyses of the genomic background (see supplementary table S6, Supplementary Material online for quality statistics after filtering). Poor mapping could lead to an excess of false-positive polymorphisms and an excess of heterozygotes in the resistance region. We further examined six quality statistics for the whole genome and the resistance region: the mapping quality of the reference and alternate alleles, the proportion of reference and alternate alleles supported by properly paired-ends reads, the ratio between depth of coverage at heterozygous sites normalized by individual depth of coverage, and sequencing depth. No substantial differences were observed between the resistance region and the rest of the genome (supplementary fig. S5, Supplementary Material online).

Structure and Descriptive Statistics

To characterize population structure, we used DAPC in the R package adegenet (v2.1.2) to perform a clustering analysis (Jombart et al. 2010) on a set of 8978 SNPs with no missing data; these were thinned every 1,000 bp to limit the effects of linkage and of variation in SNP density (supplementary fig. S4, Supplementary Material online). DAPC first decomposes the variance in the data set into principal components (PC), then performs a discriminant analysis on these PC to identify the most likely genetic clusters. We selected the clustering model with the highest support using BIC and retained 14 PC that explained about 21.4% of the total variance and two of the linear discriminants. These numbers were chosen through a cross-validation procedure that suggested perfect assignation to clusters with a 0% mean-square error (Jombart and Collins 2015). Tajima's *D*, F_{ST} *and* d_{xy} were calculated for nonoverlapping 1-kb windows using the R package PopGenome (v2.2.5) (Pfeifer et al. 2014), and *D. similis* was used as an outgroup for computing Fu and Li's *F*, and *D*. This windows length was chosen to ensure independence between windows, based on the rapid decay in linkage disequilibrium over 1,000 bp in *Daphnia* genome (supplementary fig. S4, Supplementary Material online). Windows with less than 5 segregating sites were excluded.

Demographic Model

We fitted a demographic model on SNP data using the likelihood algorithm implemented in fastsimcoal2.6 (Excoffier and Foll 2011). The model consisted of one ancestral population that split into three with gene flow, and allowed one population size change after each split to reflect the recent postglacial expansion in *D. magna* (Fields et al. 2018). The three populations corresponded to the three geographical clusters identified by the DAPC. To obtain accurate spectra and limit the impact of missing data, we used a subset of 48 clones with less than 5% missing data (supplementary table S7, Supplementary Material online), covering the whole species range as well as common resistotypes. Migration rates, along with current effective population sizes and time since divergence between populations, were estimated from the joint folded AFS with 30 independent runs, and included 2,458,902 SNPs with no missing data. We estimated the total number of callable sites with the coverage tool in BEDTOOLS v2.25.0 to exclude genomic intervals covered at less than 10× depth in each single individual (Quinlan and Hall 2010). Each run used 40 cycles of likelihood optimization, with 100,000 coalescent simulations per cycle. We present the results from the run with the highest likelihood. Time in years and effective population size were obtained by assuming a mutation rate of 8.96×10^{-9} substitution/generation (Ho et al. 2020) and ten generations (asexual and sexual) per year (Haag et al. 2009). The same procedure was applied to 100 bootstrapped frequency spectra to obtain confidence intervals for all parameters.

Forward-in-Time Simulations

To understand how diversity statistics in 1-kb windows may be affected by demography, variable proportion of mutations under balancing selection, and equilibrium frequency, we performed simulations using the forward-in-time simulator SLiM 3 (Haller and Messer 2019) for 125 diploid individuals drawn from three populations following the demographic model inferred by fastsimcoal2.6. For consistency with previous analyses of genetic diversity in *Daphnia magna*, we assumed a mutation rate of 8.96×10^{-9}/generation (sexual + asexual combined), and a recombination rate of 6.78×10^{-8}/sexual generation (Dukić et al. 2016), equivalent to 6.78×10^{-9}/generation (sexual + asexual combined). For scenarios with balancing selection, 0.01% or 0.1% of new mutations were under NFDS. At equilibrium frequency, the selective coefficient *s* was equal to 0 and varied, so that $s = 0.005$ *

$(f_{eq} - f_{obs})/f_{eq}$, where f_{eq} is the equilibrium frequency, and f_{obs} the frequency of the allele at a given generation in a given population. This results in a dynamic where s approaches 0.005 as f_{obs} approaches 0, and −0.005 as f_{obs} approaches 2 * f_{eq}. To shorten run times, we scaled all parameters inferred by fastsimcoal2.6 by a factor of 100: migration, mutation and recombination rates were multiplied by 100, whereas effective population sizes, times in generation, and selection coefficients were divided by the same factor. This scaling maintains constant parameters that control mutation-selection-drift balance, such as $N\mu$, Nm, Nr, and Ns, with μ the mutation rate, m the migration rate, r the recombination rate, s the selection coefficient, and N is the effective population size. Simulations were run without any demographic event for 10,000 generations (after scaling) to ensure that mutation-selection-drift balance was achieved. We ran 1,000 simulations for each combination of parameters, producing a VCF file for each; summary statistics were computed with PopGenome.

Test for Neutrality and Composite Likelihood Ratio Test for Balancing Selection

To test for deviation from neutrality, we generated 10 million coalescent simulations with fastsimcoal without scaling parameters and converted the outputs into VCF files to obtain statistics with PopGenome. Divergence and diversity statistics were summarized through a PCA using the prcomp function in R. We then predicted PCA scores for windows in the resistance region and for SLiM3 simulations. Envelopes containing 95% of points for each category were obtained using the locfit package in R (https://cran.r-project.org/web/packages/locfit/index.html). Deviation from neutrality was estimated for each 1-kb window in the genome by counting the proportion of simulations with a higher score on the first PC axis. The resulting P values were Bonferroni-corrected. We also calculated the $B_{0,MAF}$ statistics (Cheng and Degiorgio 2020) for each of the three geographical groups identified by DAPC. The statistics does not require specifying a window's size. Allele count data were extracted from the VCF file using VCFTOOLS.

Ancestral Recombination Graphs and Alleles Age

We conducted coalescent analyses using ARGweaver (Rasmussen et al. 2014; Hubisz et al. 2020) on the same 48 clones with less than 5% missing data that we used in the demographic analyses (supplementary table S7, Supplementary Material online). ARGWeaver estimates local recombination rates and time since coalescence along the genome by reconstructing genealogies along the genome as well as changes in their branching due to recombination events ARG. To limit computation time, we focused on three scaffolds larger than 2 Mb (PacBio contig + end of scaffold00944, scaffold00512, and scaffold00024) that belonged to distinct linkage groups. We used the VCF file as an input, which makes ARGWeaver estimate the phase for each diploid genome. We used the same mutation and recombination rates as those used in demographic inference and simulations (see above). We also allowed changes in effective population size over time, using results from our fastsimcoal2.6 inference as a prior. We set the number of time points at which coalescence events could happen at 10 and set the maximum coalescence time at 5 million generations. Because we are mostly interested in ancient balancing selection, we also set the −delta parameter at 0.00001 so that coalescence events were less biased toward recent times than with the default value. The algorithm was run over 6,000 iterations and the MCMC chain sampled every 30 iterations. Observation of the likelihood values showed that convergence was achieved after 2,000 MCMC iterations, which were discarded as burn-in. We then extracted time since TMRCA for each nonrecombining block, the minimal time since coalescence for half of the samples, the recombination rate, and the total length of genealogies. To obtain statistics over 1-kb windows, we averaged estimates across nonrecombining blocks using the package regioneR (v1.20.0) (Gel et al. 2015).

Refined Scans for Balancing Selection

For each SNP we computed β score (Siewert and Voight 2017), a statistic that identifies SNPs of allele clusters that segregate at similar frequencies, a pattern associated with long-term balancing selection. The length of the windows we examined around each given SNP was chosen using the formula provided in Siewert and Voight (2017). The distribution of haplotypes sizes is exponential with rate parameter $T * \rho$, with T being the time since balancing selection and ρ the recombination rate. Assuming $T = 3 \times 10^6$ generations (which is about ten times older than the average coalescence time retrieved by ARGWeaver) and $\rho = 6.78 \times 10^{-9}$/generation, 95% of haplotypes flanking a selected site should be shorter than 147 bp. We used a window size of 125 bp on each side of each focal SNP (for a total size of 250 bp, option -w 250), which, assuming 10 generations/year, should guarantee the detection of events that occurred in the last 300,000 years. Because alleles with equilibrium frequencies below 0.1 are more likely to be erased by drift, the statistic was not reported for SNPs at frequencies beneath this threshold (option -m 0.1). We performed analyses within each of the three clusters identified by DAPC to minimize confounding effects of population structure and regional adaptation and included all 125 clones. For some candidate regions, we extracted genealogies from the ARGWeaver output that overlapped with the SNP with the highest β score, sampling a random genealogy from the post burn-in MCMC iterations. We also sampled 5,000 random genealogies across the scaffolds, estimated pairwise phylogenetic distances between all pairs of samples using the R package ape, and performed Mantel test between these distance matrices and the matrix of geographical distances between samples. These analyses were conducted using the R packages ape (Paradis et al. 2004) and ecodist.

Identification of Candidate Genes

We identified a set of strong candidates for balancing selection by first selecting SNPs and genomic regions in the top 1% for β scores (threshold estimated using all scaffolds larger than 10 kb). A mappability score was estimated using

GenMap (Pockrandt et al. 2020) (v1.0.2), with a score of 1 indicating no repetitive sequence at a given position. We replaced regions from scaffold00944 covered by the improved PacBio scaffold. We filtered out regions that had overlapping, repetitive content, that is, sequences of at least 125 bp (125-mers, length of a single Illumina read) and scores <1. We allowed for up to four mismatches between repeated 125-mers. To further eliminate possible issues with copy-number variants that could artificially inflate diversity, we performed a one-tailed test for an excess of heterozygotes in all 125 individuals and removed regions where SNPs harbored P values lower than 1×10^{-4}. Windows satisfying these conditions, and genes overlapping them, were extracted using BEDTOOLS (Quinlan and Hall 2010). We also extracted the highest $B_{0,MAF}$ value at each candidate gene using BEDTOOLS.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Jürgen Hottinger and Urs Stiefel for laboratory support. A Natural Environment Research Council (NERC) grant NE/J010790/1 to Darren Obbard and Tom Little (Edinburgh University) contributed to the sequencing of some of the *Daphnia* clones used here. We thank the members of the Ebert group for helpful discussions and comments on the manuscript. Suzanne Zweizig improved the language of the manuscript. We also thank two anonymous reviewers who made constructive comments on an earlier version of the manuscript. This work was supported by the Swiss National Science Foundation. The research was carried out on the High-Performance Computing resources at New York University Abu Dhabi, and the Sciama High Performance Compute (HPC) cluster supported by the ICG, SEPNet and the University of Portsmouth.

Data Availability

All sequencing data (.BAM files) have been made available through the NCBI BioProject PRJNA745967. Supplementary Material is available for this article. All code and software are freely available. Correspondence and requests for materials should be addressed to Yann Bourgeois and Dieter Ebert.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389–253402.

Ameline C, Bourgeois Y, Vögtli F, Savola E, Andras J, Engelstädter J, Ebert D. 2021. A two-locus system with strong epistasis underlies rapid parasite-mediated evolution of host resistance. *Mol Biol Evol.* 38(4):1512–1528.

Andras JP, Ebert D. 2013. A novel approach to parasite population genetics: experimental infection reveals geographic differentiation, recombination and host-mediated population structure in *Pasteuria ramosa*, a bacterial parasite of *Daphnia*. *Mol Ecol.* 22(4):972–986.

Andras JP, Fields PD, Pasquier LD, Fredericksen M, Ebert D. 2020. Genome-wide association analysis identifies a genetic basis of infectivity in a model bacterial pathogen. *Mol Biol Evol.* 37(12):3439–3452.

Auld SKJR, Tinkler SK, Tinsley MC. 2016. Sex as a strategy against rapidly evolving parasites. *Proc R Soc B.* 283(1845):20162226.

ΘBento G, Fields PD, Duneau D, Ebert D. 2020. An alternative route of bacterial infection associated with a novel resistance locus in the *Daphnia–Pasteuria* host–parasite system. *Heredity* 125(4):173–183.

Bento G, Routtu J, Fields P, Bourgeois Y, Du Pasquier L, Ebert D. 2017. The genetic basis of resistance and matching-allele interactions of a host-parasite system: the *Daphnia magna-Pasteuria ramosa* model. *PLoS Genet.* 13(2):e1006596.

Bergelson J, Kreitman M, Stahl EA, Tian D. 2001. Evolutionary dynamics of plant R-genes. *Science* 292(5525):2281–2285.

Bitarello BD, De Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D, Andres AM. 2018. Signatures of long-term balancing selection in human genomes. *Genome Biol Evol.* 10(3):939–955.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

Bolnick DI, Stutz WE. 2017. Frequency dependence limits divergent evolution by favouring rare immigrants over residents. *Nature* 546(7657):285–288.

Bourgeois Y, Roulin AC, Müller K, Ebert D. 2017. Parasitism drives host genome evolution: insights from the *Pasteuria ramosa–Daphnia magna* system. *Evolution* 71(4):1106–1113.

Burri R. 2017. Interpreting differentiation landscapes in the light of long-term linked selection. *Evol Lett.* 1(3):118–131.

Charlesworth B. 2013. Background selection 20 years on. *J Hered.* 104(2):161–171.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res.* 70(2):155–174.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2(4):e64.

Cheng X, Degiorgio M. 2020. Flexible mixture model approaches that accommodate footprint size variability for robust detection of balancing selection. *Mol Biol Evol.* 37(11):3267–3291.

Clarke BC. 1976. Genetic aspects of host-parasite relationships. In: Taylor AER, Muller RM, editors. The ecological relationship of host-parasite relationships. Oxford: Blackwell. p. 87–104.

Cornetti L, Fields PD, Van Damme K, Ebert D. 2019. A fossil-calibrated phylogenomic analysis of *Daphnia* and the Daphniidae. *Mol Phylogenet Evol.* 137:250–262.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol.* 23(13):3133–3157.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.

Decaestecker E, Gaba S, Raeymaekers J. A M, Stoks R, Van Kerckhoven L, Ebert D, De Meester L. 2007. Host-parasite "Red Queen" dynamics archived in pond sediment. *Nature* 450(7171):870–873.

Dray S, Dufour AB. 2007. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw.* 22:1–20.

Dukić M, Berner D, Roesti M, Haag CR, Ebert D. 2016. A high-density genetic map reveals variation in recombination rate across the genome of *Daphnia magna*. *BMC Genet.* 17(1):137

Duneau D, Luijckx P, Ben-Ami F, Laforsch C, Ebert D. 2011. Resolving the infection process reveals striking differences in the contribution of environment, genetics and phylogeny to host-parasite interactions. *BMC Biol.* 9:11.

Ebert D, Duneau D, Hall MD, Luijckx P, Andras JP, Du Pasquier L, Ben-Ami F. 2016. A Population Biology Perspective on the Stepwise Infection Process of the Bacterial Pathogen Pasteuria ramosa in Daphnia. *Adv Parasitol.* 91:265–310.

Ebert D, Fields PD. 2020. Host–parasite co-evolution and its genomic signature. *Nat Rev Genet.* 21(12):754–768.

Eizaguirre C, Lenz TL, Kalbe M, Milinski M. 2012. Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nat Commun*. 3:621.
Excoffier L, Foll M. 2011. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27(9):1332–1334.
Fields PD, Obbard DJ, McTaggart SJ, Galimov Y, Little TJ, Ebert D. 2018. Mitogenome phylogeographic analysis of a planktonic crustacean. *Mol Phylogenet Evol*. 129:138–148.
Fields PD, Reisser C, Dukic M, Haag CR, Ebert D. 2015. Genes mirror geography in *Daphnia magna*. *Mol Ecol*. 24(17):4521–4536.
Frank SA. 1991. Ecological and genetic models of host-pathogen coevolution. *Heredity* 67(Pt 1):73–83.
Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133(3):693–709.
Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. 2015. RegioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* 32:289–291.
Gibson AK, Delph LF, Vergara D, Lively CM. 2018. Periodic, parasite-mediated selection for and against sex. *Am Nat*. 192(5):537–551.
Goren L, Ben-Ami F. 2013. Ecological correlates between cladocerans and their endoparasites from permanent and rain pools: patterns in community composition and diversity. *Hydrobiologia* 701(1):13–23.
Haag CR, McTaggart SJ, Didier A, Little TJ, Charlesworth D. 2009. Nucleotide polymorphism and within-gene recombination in *Daphnia magna* and *D. pulex*, two cyclical parthenogens. *Genetics* 182(1):313–323.
Haller BC, Messer PW. 2019. SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol*. 36(3):632–637.
Hamilton WD. 1980. Sex versus non-sex versus parasite. *Oikos* 35:282–290.
Hamilton WD, Axelrod R, Tanese R. 1990. Sexual reproduction as an adaptation to resist parasites (A review). *Proc Natl Acad Sci U S A*. 87(9):3566–3573.
Ho EKH, Macrae F, Latta LC, McIlroy P, Ebert D, Fields PD, Benner MJ, Schaack S. 2020. High and highly variable spontaneous mutation rates in *Daphnia*. *Mol Biol Evol*. 37(11):3258–3266.
Hubisz MJ, Williams AL, Siepel A. 2020. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genet*. 16(8):e1008895.
Joly S, Schoen DJ. 2011. Migration rates, frequency-dependent selection and the self-incompatibility locus in Leavenworthia (Brassicaceae). *Evolution* 65(8):2357–2369.
Jombart T, Collins C. 2015. A tutorial for discriminant analysis of principal components (DAPC) using adegenet. R Vignette. Available from: https://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf. Accessed August 9, 2021.
Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet*. 11:94.
Jousimo J, Tack AJM, Ovaskainen O, Mononen T, Susi H, Tollenaere C, Laine AL. 2014. Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science* 344(6189):1289–1293.
Kaltz O, Shykoff JA. 1998. Local adaptation in host–parasite systems. *Heredity* 81(4):361–370.
Kaufman J. 2018. Unfinished business: evolution of the MHC and the adaptive immune system of jawed vertebrates. *Annu Rev Immunol*. 36(36):383–409.
Keller D, Kirk D, Luijckx P. 2019. Four QTL underlie resistance to a microsporidian parasite that may drive genome evolution in its *Daphnia* host. bioRxiv.
Krebs M, Routtu J, Ebert D. 2017. QTL mapping of a natural genetic polymorphism for long-term parasite persistence in *Daphnia* populations. *Parasitology* 144(13):1686–1694.
Laine AL, Burdon JJ, Dodds PN, Thrall PH. 2011. Spatial variation in disease resistance: from molecules to metapopulations. *J Ecol*. 99(1):96–112.
Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339(6127):1578–1582.
Legendre P, Fortin MJ, Borcard D. 2015. Should the mantel test be used in spatial analysis? *Methods Ecol Evol*. 6(11):1239–1247.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
Lively CM. 2010. A review of Red Queen models for the persistence of obligate sexual reproduction. *J Hered*. 101(Suppl):S13–S20.
Lively CM, Dybdahl MF. 2000. Parasite adaptation to locally common host genotypes. *Nature* 405(6787):679–681.
Luijckx P, Ben-Ami F, Mouton L, Du Pasquier L, Ebert D. 2011. Cloning of the unculturable parasite *Pasteuria ramosa* and its *Daphnia* host reveals extreme genotype-genotype interactions. *Ecol Lett*. 14(2):125–131.
Luijckx P, Fienberg H, Duneau D, Ebert D. 2012. Resistance to a bacterial parasite in the crustacean *Daphnia magna* shows Mendelian segregation with dominance. *Heredity*108(5):547–551.
Luijckx P, Fienberg H, Duneau D, Ebert D. 2013. A matching-allele model explains host resistance to parasites. *Curr Biol*. 23(12):1085–1088.
Metzger CMJA, Luijckx P, Bento G, Mariadassou M, Ebert D. 2016. The Red Queen lives: epistasis between linked resistance loci. *Evolution* 70(2):480–487.
Mitchell SE, Read AF, Little TJ. 2004. The effect of a pathogen epidemic on the genetic structure and reproductive strategy of the crustacean *Daphnia magna*. *Ecol. Lett*. 7(9):848–858.
Mouton L, Traunecker E, McElroy K, Du Pasquier L, Ebert D. 2009. Identification of a polymorphic collagen-like protein in the crustacean bacteria *Pasteuria ramosa*. *Res Microbiol*. 160(10):792–799.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.
Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient swiss army knife for population genomic analyses in R. *Mol Biol Evol*. 31(7):1929–1936.
Phillips KP, Cable J, Mohammed RS, Herdegen-Radwan M, Raubic J, Przesmycka KJ, van Oosterhout C, Radwan J. 2018. Immunogenetic novelty confers a selective advantage in host-pathogen coevolution. *Proc Natl Acad Sci U S A*. 115(7):1552–1557.
Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. 220. GenMap: ultra-fast computation of genome mappability. Bioinformatics. 36(12):3687–3692.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
Rabajante JF, Tubay JM, Ito H, Uehara T, Kakishima S, Morita S, Yoshimura J, Ebert D. 2016. Host-parasite Red Queen dynamics with phase-locked rare genotypes. *Sci Adv*. 2(3):e1501548.
Radwan J, Babik W, Kaufman J, Lenz TL, Winternitz J. 2020. Advances in the evolutionary understanding of MHC polymorphism. *Trends Genet*. 36(4):298–311.
Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. *PLoS Genet*. 10(5):e1004342.
Rico Y, Morris-Pocock J, Zigouris J, Nocera JJ, Kyle CJ. 2015. Lack of spatial immunogenetic structure among wolverine (*Gulo gulo*) populations suggestive of broad scale balancing selection. *PLoS One*. 10(10):e0140170.
Routtu J, Ebert D. 2015. Genetic architecture of resistance in *Daphnia* hosts against two species of host-specific parasites. *Heredity* 114(2):241–248.
Roux C, Pauwels M, Ruggiero M-V, Charlesworth D, Castric V, Vekemans X. 2013. Recent and ancient signature of balancing selection around the S-locus in *Arabidopsis halleri* and *A. lyrata*. *Mol Biol Evol*. 30(2):435–447.
Sackton TB, Lazzaro BP, Schlenke T. A, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in Drosophila. *Nat Genet*. 39(12):1461–1468.

Seefeldt L, Ebert D. 2019. Temperature- versus precipitation-limitation shape local temperature tolerance in a Holarctic freshwater crustacean. *Proc Biol Sci*. 286(1907):20190929.

Siewert KM, Voight BF. 2017. Detecting long-term balancing selection using allele frequency correlation. *Mol Biol Evol*. 34(11):2996–3005.

Tellier A, Brown JKM. 2007. Polymorphism in multilocus host-parasite coevolutionary interactions. *Genetics* 177(3):1777–1790.

Tellier A, Brown JKM. 2011. Spatial heterogeneity, frequency-dependent selection and polymorphism in host-parasite interactions. *BMC Evol Biol*. 11(319):319.

Thrall PH, Barrett LG, Dodds PN, Burdon JJ. 2015. Epidemiological and evolutionary outcomes in gene-for-gene and matching allele models. *Front Plant Sci*. 6:1084.

Thrall PH, Laine AL, Ravensdale M, Nemri A, Dodds PN, Barrett LG, Burdon JJ. 2012. Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation. *Ecol Lett*. 15(5):425–435.

Yampolsky LY, Zeng E, Lopez J, Williams PJ, Dick KB, Colbourne JK, Pfrender ME. 2014. Functional genomics of acclimation and adaptation in response to thermal stress in *Daphnia*. *BMC Genomics*. 15:859.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 42(7):565–569.

A Two-Locus System with Strong Epistasis Underlies Rapid Parasite-Mediated Evolution of Host Resistance

Camille Ameline [1,*,1], Yann Bourgeois [1,2], Felix Vögtli,[1] Eevi Savola [1,3], Jason Andras [1,4], Jan Engelstädter [5] and Dieter Ebert [1]

[1]Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland
[2]School of Biological Sciences, University of Portsmouth, Portsmouth, United Kingdom
[3]Institute of Evolutionary Biology, Ashworth Laboratories, University of Edinburgh, Edinburgh, United Kingdom
[4]Department of Biological Sciences, Clapp Laboratory, Mount Holyoke College, South Hadley, MA, USA
[5]School of Biological Sciences, The University of Queensland, Brisbane, QLD, Australia

*Corresponding author: E-mail: cameline8@gmail.com.

Associate editor: Deepa Agashe

Abstract

Parasites are a major evolutionary force, driving adaptive responses in host populations. Although the link between phenotypic response to parasite-mediated natural selection and the underlying genetic architecture often remains obscure, this link is crucial for understanding the evolution of resistance and predicting associated allele frequency changes in the population. To close this gap, we monitored the response to selection during epidemics of a virulent bacterial pathogen, *Pasteuria ramosa*, in a natural host population of *Daphnia magna*. Across two epidemics, we observed a strong increase in the proportion of resistant phenotypes as the epidemics progressed. Field and laboratory experiments confirmed that this increase in resistance was caused by selection from the local parasite. Using a genome-wide association study, we built a genetic model in which two genomic regions with dominance and epistasis control resistance polymorphism in the host. We verified this model by selfing host genotypes with different resistance phenotypes and scoring their F1 for segregation of resistance and associated genetic markers. Such epistatic effects with strong fitness consequences in host–parasite coevolution are believed to be crucial in the Red Queen model for the evolution of genetic recombination.

Key words: **parasite-mediated selection, zooplankton, resistance, genetic architecture, epistasis, dominance, multi-locus genetics, *Daphnia magna*, *Pasteuria ramosa*.**

Article

Introduction

Darwinian evolution is a process in which the phenotypes that are best adapted to the current environment produce more offspring for the next generation. Genetic variants that code for these phenotypes are thus expected to increase in frequency in the population. Although this concept is fundamental in evolutionary biology, it remains difficult to connect the phenotype under selection with the underlying changes in the gene pool of natural populations (Ellegren and Sheldon 2008; Whitlock and Lotterhos 2015; Hoban et al. 2016). Although single-gene effects have been shown to explain the phenotype–genotype interplay in some naturally evolving populations (Daborn 2002; Cao et al. 2016; van't Hof et al. 2016), the genetic architecture underlying a phenotype is often complex. In addition, the way the environment influences the expression of a trait, and genotype × environment interactions may further obscure the link between phenotype and genotype. It is, thus, often impossible to predict genetic changes in a population that result from selection on specific phenotypes. Among the most potent drivers of evolutionary change in host populations are parasites; parasite-mediated selection can raise the frequency of resistant phenotypes rapidly (Schmid-Hempel 2011; Kurtz et al. 2016; Morgan and Koskella 2017; Koskella 2018) and is thought to contribute to many biological phenomena, such as biodiversity (Laine 2009), speciation (Schlesinger et al. 2014), and the maintenance of sexual recombination in the host (Lively 2010; Gibson et al. 2018).

To link patterns produced by parasite-mediated selection with evolutionary theory, we need to know the genetic architecture that underlies resistance; this includes the number of loci, their relative contribution to the phenotype, and the interaction between loci (epistasis) and alleles (dominance). In this way, we may be able to predict the outcome of selection, test theoretical models, and understand epidemiological dynamics (Hamilton 1980; Galvani 2003; Schmid-Hempel 2011). In a few cases, resistance to parasites has been found to be determined by single loci with strong effects, for example, in plants (Gómez-Gómez et al. 1999; Li and Cowling 2003; Li et al. 2017), invertebrates (Juneja et al. 2015; Xiao et al.

2017), and vertebrates (Samson et al. 1996). However, the genetic architecture is often obscured by intrinsic complexity and confounding factors that may influence the phenotype. Resistance might be determined by multiple loci with qualitative or quantitative effects, present distinct dominance patterns, and display interactions with other genes or the environment. Indeed, multilocus genetic architecture of resistance can create more diversity, and is thus thought to be more common than single loci (Sasaki 2000; Tellier and Brown 2007; Wilfert and Schmid-Hempel 2008). Multilocus architecture was described in *Drosophila melanogaster*, for example, where resistance was found to be determined mostly by a few large-effect loci (Bangham et al. 2008; Magwire et al. 2012) and some additional small-effect loci (Cogni et al. 2016; Magalhães and Sucena 2016). Quantitative resistance has also been found in crops where it may be used as a pathogen control strategy (Pilet-Nayel et al. 2017). In the water flea *Daphnia magna*, resistance has been found to be quantitative to a microsporidian parasite, but qualitative to a bacterial pathogen (Routtu and Ebert 2015). Although resistance tends to be dominant (Hooker and Saxena 1971; Carton et al. 2005), resistant alleles have been found to be both dominant and recessive in plants (Gómez-Gómez et al. 1999; Li and Cowling 2003; Li et al. 2017) and invertebrates (Luijckx et al. 2012; Juneja et al. 2015; Xiao et al. 2017). Epistasis between resistance loci has also been found in diverse plants and animals (Kover and Caicedo 2001; Wilfert and Schmid-Hempel 2008; Jones et al. 2014; González et al. 2015; Metzger et al. 2016), emphasizing its crucial role in the evolution of resistance. The link between genetic architecture and natural selection for resistance remains weak, however, mainly limited to the theoretical extrapolation of results from laboratory experiments.

Dominance and epistasis describe the nonadditive interaction among alleles of the same or different loci, respectively, making them crucial for the evolutionary response to selection. Epistasis among resistance genes could contribute to the maintenance of genetic diversity by reducing fixation rates at individual loci, and is thus thought to be pervasive (Tellier and Brown 2007). In the Red Queen model for the evolution of sex, thus, epistasis among resistance loci helps maintain genetic diversity and recombination in the host (Hamilton 1980; Hamilton et al. 1990; Howard and Lively 1998; Salathé et al. 2008; Engelstädter and Bonhoeffer 2009; Kouyos et al. 2009). Important with regard to the role of epistasis for the evolution of host–parasite interactions is furthermore, that the interacting loci must be polymorphic within the same natural populations. However, the importance of genetic architecture for understanding the evolution of resistance stands in stark contrast to the limited amount of available data on natural populations (Alves et al. 2019). In this study, we investigate the evolution of resistance in a natural population of the planktonic crustacean *D. magna* as it experiences epidemics of the virulent bacterial pathogen *Pasteuria ramosa*. We link parasite-mediated selection to its associated allele frequency change by resolving the underlying genetic architecture of host resistance.

In recent years, water fleas of the genus *Daphnia* (Crustacea, Cladocera) and their microparasites have become one of the best understood systems for studying the evolution and ecology of host–parasite interactions (Ebert 2005; Vale et al. 2011; Izhar and Ben-Ami 2015; González-Tortuero et al. 2016; Strauss et al. 2017; Shocket et al. 2018; Turko et al. 2018; Rogalski and Duffy 2020). Parasite selection in natural *Daphnia* populations has been shown to alter the phenotypic distribution of resistance (Little and Ebert 1999; Decaestecker et al. 2007; Duffy and Sivars-Becker 2007; Duncan and Little 2007), and genetic mapping studies identified loci involved in host resistance (Luijckx et al. 2012, 2013; Routtu and Ebert 2015; Metzger et al. 2016; Bento et al. 2017, 2020) and parasite infectivity (Andras et al. 2020); however, because studies on host resistance largely involved crosses among populations, the results may not reflect genetic variation within populations. Genetic changes in natural host populations have been observed but so far it was not possible to link this change to parasite resistance loci (Mitchell et al. 2004; Duncan and Little 2007). Understanding the link between parasite-mediated selection on host resistance and the underlying genetic architecture would enable us to determine and predict the tempo and mode of evolution in natural populations and to link observed phenotypic changes to frequency changes of alleles under selection. This study provides such a phenotype–genotype link. We quantified the change in frequency of resistance phenotypes over time in a natural *D. magna* population and, through experiments, showed that the locally dominant, virulent parasite genotype of *P. ramosa* played a major role in the observed phenotypic changes. A genome-wide association study (GWAS) and genetic crosses revealed the underlying genetic architecture of resistance in our study population and provided a genetic model for inheritance of resistance. This genetic model comprises two resistance loci presenting distinct dominance patterns and strongly linked with epistasis. These results strongly support the Red Queen model of host–parasite coevolution and the maintenance of genetic recombination.

Results

Parasite-Mediated Selection Explains Phenotypic Dynamics

Monitoring

We monitored the Aegelsee *D. magna* population from fall 2010 to fall 2015, whereas the present study focuses on the 2014 and 2015 planktonic seasons. In this population, *D. magna* diapauses during winter as resting eggs, whereas the active season spans from early April to early October. Each summer, we observed a *P. ramosa* epidemic that typically started in early May, about a month after *Daphnia* emerged from diapause, and lasted throughout the summer (fig. 1A) with peak prevalence of 70–95%. *Pasteuria ramosa* infection in the host is characterized by gigantism, a reddish-brownish opaque coloration, and castration, that is, an empty brood pouch. *Pasteuria ramosa* is a virulent parasite, stripping the host of 80–90% of its residual reproductive success and

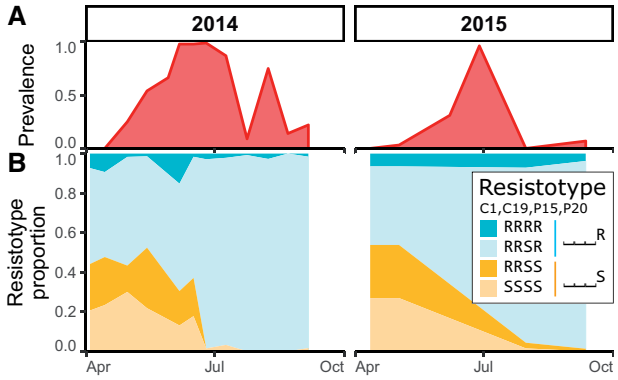


FIG. 1. Prevalence and resistotype dynamics observed in the Aegelsee *Daphnia magna* population. (*A*) *Pasteuria ramosa* prevalence across two summer epidemics. (*B*) Resistotype (resistance phenotype) frequencies across time ($n = 60–100$ *D. magna* clones from each sampling date in 2- to 3-week intervals). Resistotypes = resistance to *P. ramosa* C1, C19, P15, and P20, consecutively.

killing it after 6–10 weeks, at which point it releases millions of long-lasting spores into the environment (Ebert et al. 1996, 2016).

Animals sampled from the field were cloned, and their resistance phenotypes (resistotypes) were scored. *Daphnia magna* produces asexual clonal eggs which are used in the laboratory to produce clonal lines, a.k.a. genotypes. Individuals castrated by the parasite received an antibiotic treatment to allow clonal reproduction. Resistance to the bacteria is indicated when parasite spores are unable to attach to the gut wall of the host (Duneau et al. 2011; Luijckx et al. 2011). We thus defined host clone resistotypes according to the ability of parasite spores of given isolates to attach to the host gut wall or not. The host's overall resistotype is its combined resistotypes for the four *P. ramosa* isolates in the following order: C1, C19, P15, and P20, for example, a clone susceptible to all four isolates will have the SSSS resistotype. P20 had been isolated from our study population in May 2011; isolates C1, C19, and P15 had previously been established in the laboratory from other *D. magna* European populations. Overall, we found three predominant resistotypes: RRSR, RRSS, and SSSS, which together accounted for 95.1 ± 1.0% of all tested animals over the active season in 2014 ($n = 995$) and 2015 ($n = 260$). RRRR represented a much smaller proportion of the resistotypes (4.9 ± 1.0%) (fig. 1B). Excluding the resistotype data for *P. ramosa* isolate P15, for which over 95% of the hosts were susceptible, the study population was mainly composed of the three resistotypes: RR␣R, RR␣S, and SS␣S. When one isolate was not considered, we used the placeholder "␣" for that resistotype: for example, "RR␣R resistotype." A few other resistotypes that were absent in the 2014 and 2015 samples were observed in other samples. Notably, the SR␣S resistotype was found in 0.3% of hatched animals from *D. magna* resting eggs sampled during the winter 2014 diapause. The SR␣R resistotype has never been found in the field samples but was found in the selfed offspring of the rare resistotype SR␣S. Resistotypes RS␣␣ and SS␣R were not observed in this population.

In 2011, we sampled a subset of infected animals ($n = 113$) to characterize *P. ramosa* diversity among infected hosts throughout the active season and found that the P20 genotype represented about 50% of the parasite diversity among infected hosts when the epidemics began. This proportion decreased to zero during the epidemic, as other *P. ramosa* genotypes took over (supplementary fig. S1, Supplementary Material online).

The temporal dynamics revealed an increase in animals resistant to P20 (RRSR and RRRR, in short: RR␣R, or ␣␣␣R) soon after the onset of the epidemics, whereas animals susceptible to P20 (RRSS and SSSS, or ␣␣␣S) declined accordingly (fig. 1B) in both study years. Resistance to C1, C19, and P15 did not seem to play a strong role in the selection process during the epidemics. In the result described next, we tested the hypothesis that selection by *P. ramosa* isolate P20 is the main driver of natural resistotype dynamics in our study population during the early planktonic season. As a reminder, P20 has been isolated from a spring sample of the here-studied population.

Experimental and Field Infections

First, we tested the impact of the parasite on the different resistotypes to associate disease phenotype with resistotype. To do this, we obtained a sample of the spring cohort of the *D. magna* population by hatching resting eggs collected in February 2014. These animals represented, in total, 70 clones of the four most common resistotypes (RRSR, RRSS, SSSS, RRRR), with each clone replicated five times ($n = 350$). We exposed these clonal offspring to a mixture of *P. ramosa* spores that represented the diversity of the parasite population during the early phase of the epidemic. We then monitored the hosts for infection (looking for visible signs) and fecundity (counting the number of produced clutches). Sixteen animals died before we could test their infection status, resulting in a total sample size of $n = 334$. Individuals with resistotypes RRSS and SSSS (susceptible to P20) were infected far more frequently than RRSR and RRRR (resistant to P20) individuals (fig. 2A; null deviance = 461.3 on 333 df, residual deviance = 358.4 on 329 df, $P < 0.001$). The analysis also compared P20-susceptible and P20-resistant resistotypes, confirming the high susceptibility of the P20-susceptible animals (fig. 2A; null deviance = 461.3 on 333 df, residual deviance = 360.7 on 331 df, $P < 0.001$). Infected P20-susceptible individuals produced on an average about one less clutch before parasitic castration ($n = 136$, 1.83 ± 0.07 clutches) than did infected P20-resistant individuals ($n = 19$, 2.53 ± 0.3 clutches) (fig. 2B; null deviance = 76.9 on 154 df, residual deviance = 74.0 on 152 df, $P = 0.023$). Accordingly, the average time period until visible infection was shorter in P20-susceptible clones (15.7 ± 0.2 days) than in P20-resistant clones (19.4 ± 1 days) (fig. 2C; null deviance = 94.4 on 154 df, residual deviance = 85.1 on 152 df, $P = 0.0018$). These results clearly support the hypothesis that early season *P. ramosa* strains from the field select on the P20 resistotype.

In the following year, we looked at the relationship of disease phenotype and P20 resistotype only in the field by

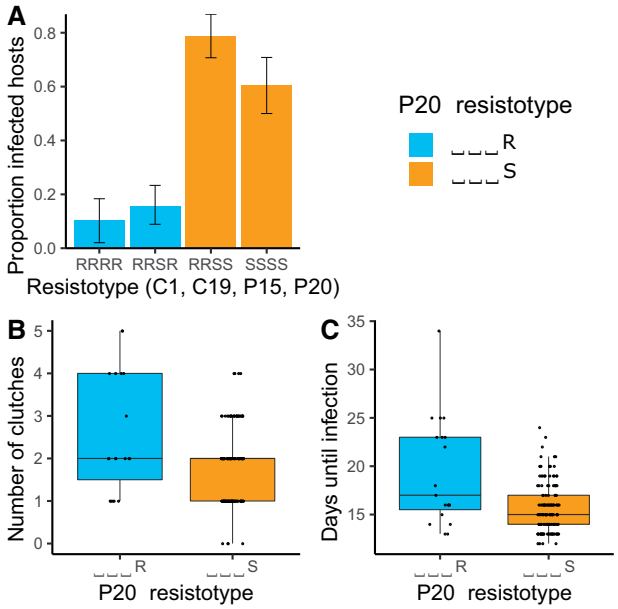


FIG. 2. Experimental infections of *Daphnia magna* with different resistotypes (resistance phenotype). Resistotypes RRSR, RRSS, SSSS ($n = 20$ clones for each), and RRRR ($n = 10$ clones) were infected with parasite spores from the early phase of the epidemic. Five repeats were performed for each clone (total $n = 334$). Controls ($n = 210$) remained uninfected and are not shown here. (*A*) Proportion of infected *D. magna* among the four resistotypes. (*B*) Number of clutches produced before parasitic castration in the infected P20-resistant (⎵⎵⎵R) and susceptible (⎵⎵⎵S) animals ($n = 115$). (*C*) Time before visible infection in P20-resistant and P20-susceptible individuals ($n = 115$).

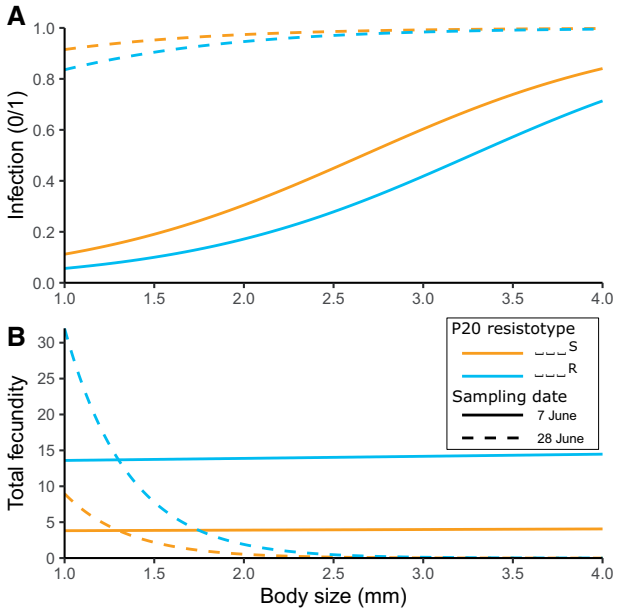


FIG. 3. Fitted models of infection phenotypes in field-collected *Daphnia magna* relative to their body size at capture (*x* axis) and their resistance to P20 for two sampling dates in June 2015. (*A*) P20-susceptible (orange) animals have a higher likelihood to be infected than P20-resistant (blue) ones for any body size. (*B*) Infected P20-susceptible animals have a lower total fecundity than P20-resistant ones for any body size. Differences between the data are partially due to the difference in parasite prevalence on the two sampling dates (31% on June 7 and 96% on June 28).

measuring the parasite's impact on P20-resistant and P20-susceptible hosts. We collected animals in the field during the early half of the *P. ramosa* epidemic and raised them individually in the laboratory, recording their disease symptoms. We then cured infected animals with antibiotics, allowed them to produce clonal offspring, and determined their P20 resistotype. Our analysis revealed higher infection rates (size corrected) for P20-susceptible than for P20-resistant individuals in these natural conditions (fig. 3*A*; Fitted model: glm [Infected (1/0) ~ P20 resistotype + Body_size + Sampling_date], family = quasibinomial(), $n = 331$; null deviance = 415.1 on 330 df, residual deviance = 209.1 on 327 df, $P = 0.025$). Field-caught infected P20-susceptible individuals also produced, on an average, fewer offspring before parasitic castration than infected P20-resistant ones (fig. 3*B*; Fitted model: glm.nb [Fecundity ~ P20 resistotype + Body_size × Sampling_date], $n = 224$; null deviance = 127.9 on 223 df, residual deviance = 92.9 on 219 df, $P = 0.014$). In both models, the sampling date also had a significant effect. Parasite prevalence on the two sampling dates differed strongly (31% on June 7 and 96% on June 28, 2015). We observed on the first sampling date that larger individuals were more infected and consequently produced less offspring. This size difference is not visible anymore on the second sampling date, where almost all individuals were infected. The overall pattern in relation to the P20 resistotype remained the same, even though the difference in infection and fecundity between field-collected P20 resistotypes was less pronounced than in the controlled infection experiment (compare figs. 2 and 3). In summary, the results of the two experiments clearly support the hypothesis that early season *P. ramosa* from the field select on the P20 resistotype.

Linking Resistance Phenotypes to Genotypes

Excluding the P15 resistotype, which has very low variability because most animals are P15-susceptible, the study population was composed mainly of three resistotypes: RR⎵R, RR⎵S, and SS⎵S. A supergene for resistance to C1 and C19 has been described in *D. magna* using QTL mapping (Routtu and Ebert 2015; Bento et al. 2017), and the genetic architecture of resistance at this so-called ABC-cluster, or *P. ramosa* resistance (PR) locus, has been further resolved using genetic crosses among host genotypes (Metzger et al. 2016). According to this genetic model, an SS⎵⎵ resistotype (susceptible to C1 and C19) has an "aabbcc" genotype (lower case letters indicate recessive alleles), whereas RS⎵⎵ individuals (resistant to C1 and susceptible to C19) are "A---cc" (upper case letters indicate dominant alleles and a dash "-" indicates alleles that do not influence the phenotype); SR⎵⎵ individuals are "aaB-cc," and RR⎵⎵ individuals are "----C-." In other words, allele A epistatically nullifies variation at the B locus, and allele C nullifies variation at the A and B loci (Metzger et al. 2016; Bento et al. 2017). See also supplementary figure S2, Supplementary Material online. Considering this genetic model, we assume that the recessive allele at the A locus is fixed in our study population ("aa" genotype)

and that the dominant allele at the B locus is very rare, as we never observed RS_ _ individuals and only found SR_ _ in very low proportions. In our study population, the SS_ _/RR_ _ polymorphism can therefore be best described by the C-locus polymorphism, that is, genotypes "aabbcc" and "aabbC-," respectively, with C being the dominant allele for resistance. Given this, we assume, in the following sections, that variation at the C locus underlies the resistance polymorphism for C1 and C19.

Genomic Regions of Resistance to the Parasite

We sequenced the genomes of 16, 10, and 11 clones with resistotypes RR_R, RR_S, and SS_S, respectively and conducted a GWAS comparing five pairs of these resistotypes to identify candidates for resistance to C1, C19, and P20: (i) SS_ _ versus RR_ _, (ii) SS_S versus RR_S, (iii) _ _ _S versus _ _ _R, (iv) RR_S versus RR_R and (v) SS_S versus RR_R. Comparisons (i) and (ii) (variation at C1 and C19 resistotypes) revealed a strong signal on linkage group (LG) 3 (fig. 4*A* and *B*). This region encompasses the super gene described earlier by Routtu and Ebert (2015) and Bento et al. (2017), the so-called ABC-cluster, or PR locus. Comparisons (iii) and (iv) (variation at P20 resistotype) revealed a strong signal on LG 5 (fig. 4*C* and *D*), hereafter called the E-locus region. In the present host–parasite system, the D locus determines resistance to P15 and is not considered here (Bento et al. 2020). The E-locus region has not yet been associated with resistance, and no PR gene has been described on the same linkage group in *D. magna*. Finally, comparison (v) (variation at C1, C19, and P20 resistotypes) indicated a strong signal at both the ABC cluster and the E-locus region (fig. 4*E*). The genomic regions associated with resistotypes in our GWAS were not sharp peaks, but rather table-like blocks of associated SNPs (fig. 4). This structure was expected for the C locus, which is a known supergene—a large block of genome space with apparently little or no recombination that contains many genes (Bento et al. 2017). Figure 4 indicates that the same may be the case for the E-locus region, where the block of associated SNPs makes up nearly half of the linkage group. A few single SNPs also showed significant association in all the comparisons (fig. 4), but because of the strength of the observed pattern and because we expected a large region to be associated with resistance, we do not consider these single SNPs further.

The E-locus region encompassed 22 scaffolds and one contig on version 2.4 of the *D. magna* reference genome, with a cumulative length of more than 3 Mb (3,101,076 bp) (supplementary table S1, Supplementary Material online). We found 485 genes on all associated scaffolds. The strongest signals of association were found on scaffolds 2,167 and 2,560, which harbored 82 genes. Some of these genes were similar to genes identified in a previous study of the ABC cluster on LG 3 (Bento et al. 2017), with a glucosyltransferase found on scaffold 2,167. Three other sugar transferases (galactosyltransferases) were identified, two of them on scaffold 2,560 (supplementary table S2, Supplementary Material online).

Genetic Model of Resistance Inheritance

Mean allele frequencies at associated SNPs showed that SS_ _ individuals (susceptible to C1 and C19) display a single allele at the C locus, whereas RR_ _ individuals display two distinct alleles at the C locus. This suggests that SS_ _ individuals are homozygous at the C locus, whereas RR_ _ individuals comprise homo- and heterozygotes. At the E locus, _ _ _R individuals (resistant to P20) are presumably homozygous, whereas _ _ _S individuals (susceptible to P20) comprise homo- and heterozygous individuals (fig. 4, right panel). These results indicate that resistance to C1 and C19 is governed by a dominant allele ("C-" genotype), as shown before (Metzger et al. 2016). In contrast, resistance to P20 is determined by a recessive allele ("ee" genotype), as was shown before for a different resistance locus (D locus, Bento et al. 2020). Screening individual genomes revealed that some SS_S individuals (susceptible to C1 and C19, and P20) present the "ee" genotype at the E locus (underlying P20 resistotype), although this genotype should confer resistance to P20. This was not observed in RR_S individuals (resistant to C1 and C19, but susceptible to P20) (supplementary table S3, Supplementary Material online), which we hypothesize to be explained by an epistatic relationship linking the C and the E loci. This epistasis confers P20 susceptibility to individuals susceptible to C1 and C19, that is, presenting the "cc" genotype, regardless of their genotype at the E locus. This genetic model is presented in figure 5 (without variation at the B locus, see below). In the present study, we mostly considered variation at the C and E loci, as they seem to play a major role in the diversity of resistotypes in our study population.

To test the genetic model derived from the GWAS, we investigated segregation of resistotypes among selfed offspring of *D. magna* genotypes with diverse resistotypes. *Daphnia magna* reproduces by cyclical parthenogenesis, in which asexual eggs produce clonal lines and sexual reproduction allows to perform genetic crosses. Our genetic model allowed us to predict the segregation of genotypes and phenotypes, which can then be compared with the observed segregation patterns among selfed offspring. From 24 host genotypes (F0 parent clones), we produced 24 groups of selfed F1 offspring. Twenty-two F0 clones included animals with all possible combinations of alleles at the C and E loci, whereas two F0s showed the rare variation at the B locus and variation at the E locus. Expected and observed resistotype frequencies are presented in tables 1 and 2 and detailed for each F1 group in supplementary tables S4–S15, Supplementary Material online. In the 22 F1 groups showing variation at the C and E loci, segregation of offspring followed the predictions of our genetic model (fig. 5), that is, we observed all expected resistotypes and saw no significant deviations from the expected frequencies based on the model. These data clearly support the genetic model for resistance at the C and E loci.

As described above, earlier research (Metzger et al. 2016; Bento et al. 2017) has shown that the dominant allele at the C locus interacts epistatically with the A and B loci (all are part of the ABC cluster), such that variation at the A and B loci

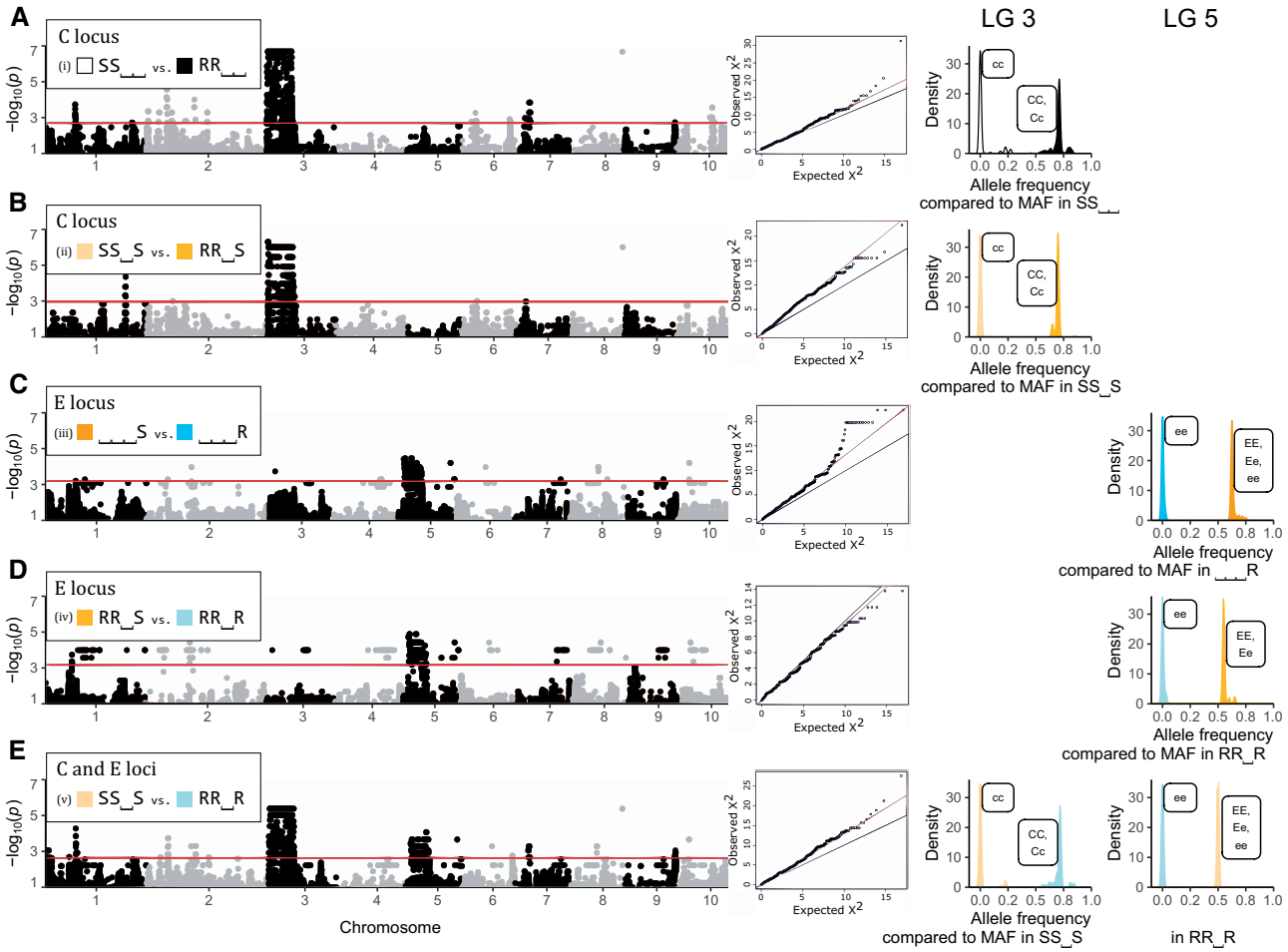


Fig. 4. GWAS analysis comparing the most common resistance phenotypes (resistotypes) in the Aegelsee *Daphnia magna* population. The resistotype depicts resistance (R) or susceptibility (S) to *Pasteuria ramosa* isolates C1, C19, P15, and P20. (i) SS_ _ versus RR_ _; (ii) SS_S versus RR_S; (iii) _ _ _S versus _ _ _R; (iv) RR_S versus RR_R, and (v) SS_S versus RR_R. Comparisons (i) and (ii) (variation at C1 and C19 resistotypes) revealed a strong signal on linkage group (LG) 3 corresponding to the C locus. Comparisons (iii) and (iv) (variation at P20 resistotype) revealed a strong signal on LG 5 corresponding to the E locus. Comparison (v) (variation at C1 and C19, and P20 resistotypes) revealed a strong signal on both regions. Left panel: Manhattan plots of relationships between different resistotype groups (showing only SNPs with $P_{corrected} < 0.01$). The *x* axis corresponds to SNP data mapped on the 2.4 *D. magna* reference genome (Routtu et al. 2014), representing only SNPs, not physical distance on the genome. Middle panel: Quantile-quantile plots of noncorrected *P* values excluding SNPs from linkage groups 3 and 5, since these scaffolds displayed an excess of strongly associated markers. Right panel: Comparison of allele frequencies between resistotype groups at the C and the E loci. Significant SNPs on LG 3 or LG 5 were used (SNPs with $P < P_{lim}/100$, with P_{lim} as defined in the Materials and Methods section, eq. 1). For each SNP, the allele with the minor allele frequency (MAF) within resistotype groups that presented only one allele at the C or the E locus (all homozygous individuals) was used for comparisons. Hence, the *x* axis represents allele frequency of the dominant allele within resistotype groups (considering total allele number, or chromosome number: 2n). Labels attached to peaks describe the inferred possible genotypes at the C or the E locus within resistotype groups. In comparisons at the C locus on LG 3, resistotype groups susceptible to C1 and C19 presented only one allele, that is, they contained only homozygous recessive individuals at the C locus (dominant allele frequency of zero). Resistotype groups resistant to C1 and C19 did contain the dominant allele (frequency between 0.5 and 1), showing that resistance is dominant at the C locus, as the resistant group contains heterozygous individuals. Similarly, in comparisons at the E locus on LG 5, resistotype groups resistant to P20 do not present the dominant allele (frequency of zero), whereas resistotype groups susceptible to P20 do, that is, contain heterozygous individuals (dominant allele frequency between 0.5 and 1). This shows that, in contrast with the C locus, susceptibility is dominant at the E locus. Screening individual genomes revealed that some SS_S individuals (susceptible to C1 and C19, and P20) presented the "ee" genotype at the E locus (resistance to P20), although susceptibility is dominant at the E locus. This was not observed in RR_S individuals (resistant to C1 and C19 but susceptible to P20) (supplementary table S3, Supplementary Material online). This observation can be explained by an epistatic relationship linking the C and the E loci. This epistasis confers susceptibility to P20 to individuals susceptible to C1 and C19, that is, presenting the "cc" genotype regardless of the genotype at the E locus. In contrast, with groups containing SS_S individuals, that is, comparisons (iii) and (iv), some SS_S individuals present the "ee" genotype at the E locus. In these groups, the frequency of the dominant allele can be lower than 0.5.

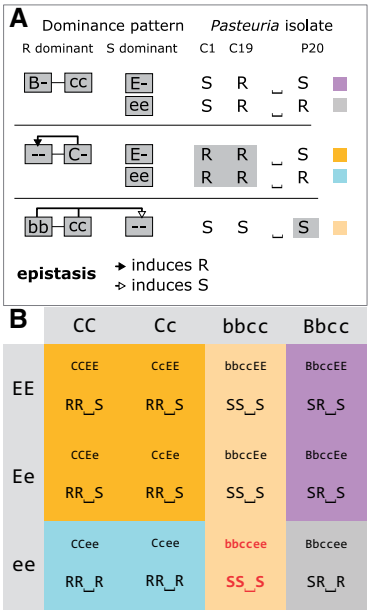


Fig. 5. Model for the genetic architecture of resistance to C1, C19, and P20 *Pasteuria ramosa* isolates in the Aegelsee *Daphnia magna* population as inferred from the GWAS analysis (fig. 4) and the genetic crosses (tables 1 and 2). (*A*) Schematic representation of the genetic model. Resistance to C1 and C19 is determined by the ABC cluster as described in Metzger et al. (2016), and the model is extended to include the newly discovered E locus. The dominant allele at the B locus induces resistance (R) to C19 and susceptibility (S) to C1. The dominant allele at the C locus confers resistance to both C1 and C19, regardless of the genotype at the B locus (epistasis). The newly discovered E locus contributes to determining resistance to P20. Resistance is dominant at the C locus (resistance to C1 and C19) but recessive at the E locus (resistance to P20). Homozygosity for the recessive allele at the B and C loci induces susceptibility to P20, regardless of the genotype at the E locus (epistasis). Hence epistasis can only be observed phenotypically in the "bbccee" genotype, which has the resistotype SS␣S. Without epistasis, the "bbccee" genotype is expected to have the phenotype SS␣R, a phenotype we never observed in the population or in our genetic crosses. (*B*) Multilocus genotypes and resistotypes at the B, C, and E loci. Resistotypes are grouped by background color. As the C allele epistatically nullifies the effect of the B locus, only combinations of the B and E loci are shown where the C locus is homozygous for the c allele. This model does not consider variation at the A locus, as the recessive allele at this locus is believed to be fixed in the Aegelsee *D. magna* population.

becomes neutral when a C allele is present. We assume that the a allele is fixed in the Aegelsee *D. magna* population, so that only variation at the B locus influences the C1 and C19 resistotypes in individuals with the "cc" genotype (see above). As variation at the B locus is very rare in our *D. magna* study population and could not be included in the GWAS analysis, we selfed two *D. magna* genotypes that presented the very rare SR␣S resistotype, whose underlying genotype at the ABC cluster we expect to be "B-cc" (probably "Bbcc," considering the B allele is rare in the population). In the F1 offspring of the two F0 parents with the SR␣S resistotype and the "Bbcc--" genotype, we observed SR␣R individuals. We speculate that SR␣R animals have the genotype "B-ccee," indicating that the epistatic relationship previously described between the C and the E loci ("cc" acts epistatically on the E locus) should also include the B locus. If this is the case, "bbcc" acts epistatically on the E locus (fig. 5). The two groups of selfed F1 offspring involving "Bb" heterozygotes showed a good fit between this expectation in the expanded model and the observed phenotypic segregation. We observed one SR␣R offspring produced from a SR␣S parent with the inferred "aaB-ccEE" genotype, which is not expected in our model (table 2B, lower panel), but typing mistakes cannot be fully ruled out.

Linking the Genomic Regions and the Genetic Model of Resistance

To test whether the segregation of the genomic regions, we discovered in the GWAS and the segregation of resistotypes in our crosses agreed with each other, we designed size-polymorphic markers in the genomic regions of the C and the E loci (two for each locus). We tested whether these markers cosegregated with the resistotypes as predicted by our genetic model. Of the four markers, DMPR1 (C locus) and DMPR3 (E locus) showed better linkage with their respective resistance loci (99.6% and 94.8% match, respectively) compared with DMPR2 (C locus) and DMPR4 (E locus) (91.4% and 69.4% match, respectively) (supplementary tables S16–S18, Supplementary Material online). These numbers reflect that our markers are close to the actual resistance loci, but that recombination between them is possible, leading to non-perfect association. We further based our scoring of resistance genotypes on the more predictive marker genotypes of DMPR1 and DMPR3. In 20 of 22 F1 groups representing all possible combinations of alleles at the C and E loci (table 1), the segregation of marker genotypes in the F1 offspring followed our genetic model predictions, that is, all expected genotypes were observed, with no statistically significant deviations from the expected frequencies. In two F1 groups from "CCEe" and "CcEe" F0 parents, the E-locus markers appeared not to be linked to the E locus (supplementary tables S5 and S8, Supplementary Material online). Based on the genotype markers results, we had assigned the "EE" genotype to the F0 parent and F1 offspring, but phenotypic segregation in the F1 offspring indicated the parent would have the "Ee" genotype. We speculate that recombination had uncoupled the genetic marker and the resistance loci in these two-parent genotypes. Together, these results show that the genomic regions found in the GWAS are indeed associated with the segregation of resistotype in the F1 selfed offspring, supporting our genetic model for the segregation of resistance (fig. 5).

Discussion

This present study aims to assess how annual epidemics by a parasitic bacterium, *P. ramosa*, influence resistance and the frequencies of the underlying genes in a natural host population of the crustacean *D. magna*. Over the course of epidemics in two consecutive years, we observed drastic changes in resistance phenotype (resistotype). Using experimental infections and fitness measurements on wild-caught

Table 1. Genetic Crosses of Resistance Phenotypes (resistotypes) from the Aegelsee *Daphnia magna* Population, Where Only the C and the E Loci Are Considered.

A

RR_S CcEe	CE	Ce	cE	ce
CE	RR_S	RR_S	RR_S	RR_S
Ce	RR_S	RR_R	RR_S	RR_R
cE	RR_S	RR_S	SS_S	SS_S
ce	RR_S	RR_R	SS_S	SS_S

B

F0 parent		Selfed F1 offspring							
				Proportion					
					Observed				
					Repeat (F1 groups from different parent clones)				
Inferred genotype	Resistotype	Inferred genotype	Resistotype	Expected	a	b	c	d	C-M-H test on counts
					n = 43	*n = 37*			
CCEE	RR_S	CCEE	RR_S	1	1	1			NA
					n = 89	*n = 31*	*n = 79*		X^2=0.85,
CCEe	RR_S	CCE-	RR_S	0.75	0.79	0.74	0.84		*df*=1,
		CCee	RR_R	0.25	0.21	0.26	0.16		*p*=0.36
					n = 39	*n = 42*	*n = 70*	*n = 79*	
CCee	RR_R	CCee	RR_R	1	1	1	1	1	NA
					n = 19				*Fisher test on counts*
CcEE	RR_S	C-EE	RR_S	0.75	0.74				*p*=1
		ccEE	SS_S	0.25	0.26				
					n = 48	*n = 34*	*n = 64*	*n = 65*	
CcEe	RR_S	C-E-	RR_S	0.56	0.47	0.56	0.65	0.51	M^2=4.61,
		cc--	SS_S	0.25	0.30	0.06	0.13	0.21	*df*=2,
		C-ee	RR_R	0.19	0.23	0.38	0.22	0.28	*p*=0.10
					n = 36	*n = 49*	*n= 22*		X^2=0.0062,
Ccee	RR_R	C-ee	RR_R	0.75	0.75	0.80	0.62		*df*=1,
		ccee	SS_S	0.25	0.25	0.20	0.38		*p*=0.94
					n = 87				
ccEE	SS_S	ccEE	SS_S	1	1				NA
					n = 84	*n = 65*			
ccEe	SS_S	ccE-	SS_S	1	1	1			NA
					n = 74	*n = 35*			
ccee	SS_S	ccee	SS_S	1	1	1			NA

NOTE.—A, Punnett square for all possible gamete combinations according to our genetic model of resistance inheritance. The table shows the resistotypes (grouped by background color) from the 16 combinations of gametes from a double heterozygote for the C and the E loci. The bottom right cell (red font, italics) represents offspring individuals where the epistatic interaction between the C and the E loci is revealed (fig. 5); B, Results from selfing of *D. magna* clones. Resistotypes of F0 mothers and F1 offspring groups were obtained using the attachment test, and resistance genotypes of F0 parents at the C and E loci were inferred from their resistotypes and the segregation patterns of resistotypes in their F1 offspring. Expected resistotype proportions within F1 groups were calculated using the genetic model presented in the Punnett square and the R package "peas" (fig. 5 and supplementary doc. S1, Supplementary Material online). Detailed results and statistical analyses for each cross are presented in supplementary tables S4–S12 and S15, Supplementary Material online. Segregation of offspring is presented as proportions, although statistical tests were run on counts. One to four crosses using distinct mother clonal lines (repeats a to d) were conducted for each F0 mother resistance genotype at the C and E loci. No variation at the B locus was observed (all F0 mothers are inferred to have the "bb" genotype according to F1 resistotype segregation).

individuals, we showed that these changes in resistotype frequency were caused by a local parasite type common during the early phases of the epidemics. A GWAS and laboratory crosses enabled us to locate the resistance genes that responded to this selection and to uncover their mode of inheritance. We pinpointed the genetic architecture of resistance to two genomic regions with dominance and epistasis, thus bridging the gap between natural selection on phenotypes and the underlying genetics.

Parasite-Mediated Selection

Over the two consecutive years of this study, resistotype frequencies in the host population changed drastically during the parasite epidemics, but remained stable outside of the epidemics (fig. 1)—a pattern consistent with the host population being under strong selection for resistance to *P. ramosa*. The P20 *P. ramosa* isolate, collected during the early epidemic, turned out to be representative of the parasite population during the early part of the two epidemics studied here: Host genotypes characterized by their susceptibility to *P. ramosa* P20 drastically decreased in proportion during the epidemics and were much more susceptible to the local parasite than P20-resistant individuals in experimental infections (fig. 2*A*). Infected P20-susceptible genotypes also became infected earlier and produced fewer offspring than P20-resistant individuals (fig. 2*B* and *C*), revealing a stronger fitness impact of

Table 2. Genetic Crosses of Resistance Phenotypes (resistotypes) from the Aegelsee *Daphnia magna* Population Considering the B and the E Loci, with the C Locus Fixed for Genotype "cc."

A

SR_S BbccEe	BcE	Bce	bcE	bce
BcE	SR_S	SR_S	SR_S	SR_S
Bce	SR_S	SR_R	SR_S	SR_R
bcE	SR_S	SR_S	SS_S	SS_S
bce	SR_S	SR_R	SS_S	SS_S

B

F0 parent		Selfed F1 offspring				
				Proportion		
Inferred genotype	Resistotype	Inferred genotype	Resistotype	Expected	Observed	Fisher test on counts
					n = 37	
BbccEE	SR_S	B-ccEE	SR_S	0.75	0.74	
		bbccEE	SS_S	0.25	0.23	*p*=1
		B-ccee	SR_R	0.00	0.03	
					n = 38	
BbccEe	SR_S	B-ccE-	SR_S	0.56	0.56	
		bbcc--	SS_S	0.25	0.33	*p*=0.58
		B-ccee	SR_R	0.19	0.11	

NOTE.—A, Punnett square for all possible gamete combinations according to our genetic model of resistance inheritance. The table shows the resistotypes (grouped by background color) resulting from the 16 combinations of gametes from a double heterozygote for the B and the E loci. The bottom right cell (red font, italics) represents offspring individuals where the epistatic interaction between the B, the C, and the E loci is revealed (fig. 5); B, Results from selfing of *D. magna* clones. Resistotypes of F0 parents and F1 offspring were obtained using the attachment test, and resistance genotypes of F0 parents at the B, C, and E loci were inferred from their resistotypes and the segregation patterns of resistotypes in their F1 offspring. Expected resistotype proportions of F1 were calculated following the genetic model outlined in the Punnett square and using the R package "peas" (fig. 5 and supplementary doc. S2, Supplementary Material online). The detailed results and statistical analyses for each cross are presented in supplementary tables S13–S15, Supplementary Material online. Segregation of offspring is presented as proportions, although the statistical tests were run on counts.

infection by the local parasite. Field data confirmed this result, as wild-caught P20-susceptible individuals were infected more frequently and produced fewer offspring than infected P20-resistant individuals (fig. 3), again showing the higher virulence of the parasite in these P20-susceptible individuals. This effect of the parasite seemed less strong in field-collected animals than for those infected in the laboratory. Multiple factors may contribute to this, including differences among field and laboratory, and differences in the host and parasite populations from the two study years (2014 and 2015). Our findings reveal nevertheless a strong and rapid response to parasite-mediated selection on host resistotypes, that are characterized by their interaction with the P20 *P. ramosa* isolate, in the natural Aegelsee *D. magna* population.

In field samples, smaller individuals were found to be less infected than larger ones. This is not surprising, as older—hence bigger—animals have longer exposure to the parasite than younger animals. For a chronic disease like *P. ramosa* infections, it is expected that, with increasing size and age, prevalence will increase. These results are thus not in conflict with reports showing that younger—hence smaller *Daphnia*—were more susceptible to parasitic infections (Garbutt et al. 2014; Izhar and Ben-Ami 2015; Ben-Ami 2019). Differences in age-related susceptibility might, however, influence the shape of the body size—prevalence relationship observed in the field.

Although the parasite P20 was isolated during the early phase of the yearly epidemics, previous research also shows other parasite genotypes in the Aegelsee population (Andras and Ebert 2013) that, as we observed in an earlier year, become more common in infected hosts later in the epidemics (supplementary fig. S1, Supplementary Material online). We speculate that these later-season isolates may represent different parasite infectotypes (infection phenotypes). Consistent with this, we observed that animals resistant to P20 did, in fact, become infected, both in the field and in the laboratory (figs. 1–3), which cannot be explained with P20-infectotype parasites alone. The present study focuses on natural selection during the early part of the epidemics, which, as our data and data from other years shows, has a fairly consistent selection pattern (Ameline C, Vögtli F, Andras J, Engelstädter J, Ebert D, unpublished data), being mainly defined by a drastic increase in P20-resistant individuals from around 50% to almost 100% within a period of 2–3 months (fig. 1).

The composition of the resistotypes at the beginning of the two seasons (2014 and 2015) in which we monitored this system was strikingly similar, which is surprising given that selection increased resistance over the course of the summer 2014. Although answering this question is not part of the current study, there are a few tentative explanations for this observation. First, part of the yearly resting eggs yield, which form the basis of the new population in the following

spring, are produced as early as mid-June before selection has diminished some of the resistotypes. Second, epistasis and dominance can protect alleles from natural selection, thus slowing down the response to selection (Feldman et al. 1975; Otto 2009). Our study, as well as earlier studies on this system (Luijckx et al. 2012, 2013; Metzger et al. 2016), all indicate strong epistasis and dominance for resistance loci. Further studies are needed to understand how much resting egg production and the genetic architecture of resistance explain the slow response to selection observed across seasons in the Aegelsee *D. magna* population.

Genetic Architecture of Resistance

To understand the genetic architecture of resistance loci under selection in our study population, we combined a GWAS using *D. magna* genotypes with different resistotypes together with a series of genetic crosses. We found that the most diversity in host resistance to the bacteria is determined by variation at the C locus, situated in a previously described supergene, the PR locus containing the ABC cluster (Bento et al. 2017), and at a newly discovered locus on a different chromosome, the E locus (fig. 4). Taken alone and in the right genetic background, that is, when there is no epistatic relationship, each of these two loci show Mendelian segregation with resistance being dominant (C locus) or recessive (E locus) (fig. 4, right panel). The two loci interact epistatically with each other, resulting in a complex pattern of inheritance (fig. 5). Balancing selection is hypothesized to maintain diversity at resistance genes (Llaurens et al. 2017; Wittmann et al. 2017; Connallon and Chenoweth 2019), and these genes are often found to have different dominance patterns and epistatic interactions (Saavedra-Rodriguez et al. 2008; González et al. 2015; Conlon et al. 2018).

The E locus is situated on linkage group (LG) 5 (genome version 2.4: Routtu et al. 2014) and appears as a large region of 3.1 Mb (fig. 4). In this regard, the E locus is similar to the ABC cluster, a well-characterized, nonrecombining, and extremely divergent region on LG 3 (Bento et al. 2017). Nonrecombining genomic structures, that is, supergenes, are suggested to facilitate adaptation via association of advantageous alleles in host–parasite coevolution (Joron et al. 2011; Llaurens et al. 2017). Such large, diverse genomic regions are difficult to study because the absence of recombination hampers fine mapping (Bento et al. 2017). Therefore, we do not know where the actual resistance loci lie within the ABC- and E-loci regions. This may also explain why our genetic markers are not perfectly linked to the resistance loci (supplementary tables S17 and S18, Supplementary Material online). Supergenes may also harbor several resistance loci, thus variation at the C or the E locus could actually represent variation at several loci physically very close to each other. Within the E-locus region, we find four sugar transferases. Glycosylation genes are candidates to explain variation of resistance in this system (Bento et al. 2017; Bourgeois et al. 2017).

In the *D. magna*–*P. ramosa* system, the ABC cluster has been shown to play a major role in host resistance and the evolutionary dynamics of resistance (Routtu and Ebert 2015; Bento et al. 2017; Bourgeois et al. 2017). Our results confirm the role of this cluster in a natural population and describe a new resistance region in the *D. magna* genome that is polymorphic in the Aegelsee population. Multilocus polymorphisms have been shown to underlie parasite resistance in host–parasite coevolution (Sasaki 2000; Tellier and Brown 2007; Cerqueira et al. 2017). In the Aegelsee *D. magna* population, there seems to be no variation at the A locus and little variation at the B locus. The observed variation at the B and C loci is consistent with the genetic model of resistance at the ABC cluster described in Metzger et al. (2016). In addition, resistance to *P. ramosa* isolate P15 (influenced by the D locus, Bento et al. 2020) remains fairly consistent, with the vast majority of animals being susceptible to P15 (fig. 1). Resistance to *P. ramosa* P21, also isolated from our study population, varies only toward the end of the summer epidemic (Ameline C, Vögtli F, Andras J, Engelstädter J, Ebert D, unpublished data). In summary, the ability to resist P20 plays a major role in the early epidemics and most resistotype diversity we measured in the Aegelsee *D. magna* population is well explained by genotypic variation at these loci. As we use more parasite isolates in further research, we might find other resistance regions in the *D. magna* genome. This is likely to be of importance in the later phase of the epidemics in the Aegelsee population.

Resistance segregation in *D. magna* is currently best explained by a genetic model where each locus contains just two alleles. This model was compiled by studies that used either mapping panels created from a few *D. magna* genotypes or, as here, host genotypes from one focal population. Additional resistance alleles may be revealed instead of new resistance regions if we test the genetic model on a larger panel of host and parasite genotypes.

We created 22 F1 offspring groups from the three common resistotypes in our study population. Segregation of resistance phenotypes and genotypes among the selfed F1 strongly supported the genetic model of resistance, consisting of the C and E loci, each with two alleles, and their epistatic interaction, produced by the GWAS (tables 1 and 2; supplementary tables S4–S15, Supplementary Material online). Two F1 offspring groups showed rare variation at the B locus, suggesting yet an additional epistatic interaction in this model besides the previously described role of the B locus for the *P. ramosa* C1 and C19 resistotypes. This consisted of the "bbcc" genotype that causes susceptibility to P20, irrespective of the genotype at the E locus (fig. 5). However, this modified model needs to be further investigated and verified with more genetic crosses.

Conclusion

In this study, we demonstrate rapid parasite-mediated selection in a natural plankton population. We find the genomic regions associated with resistance under selection and describe their mode of inheritance. This knowledge will allow us to conduct direct measurements of resistance allele frequency changes over time and to test theories on the dynamics of host and parasite coevolution, for example by

tracing genetic changes in the resting stages of *D. magna* derived from the layered sediments in ponds and lakes (Decaestecker et al. 2007). Pinpointing resistance loci can also be used to infer mechanisms of selection in the host with the molecular evolution tool box (Charlesworth 2006; Fijarczyk and Babik 2015; Hahn 2018). Our model of resistance consists of a few loci linked with epistasis and different dominance patterns, characteristics that have been shown to be relevant in coevolution, in particular when balancing selection maintains diversity at resistance genes (Tellier and Brown 2007; Engelstädter 2015; Conlon et al. 2018). The genomic regions we pinpoint can now be further studied, for example, by testing for genomic signatures for balancing selection (Charlesworth 2006; Ebert and Fields 2020). Hence, a precise knowledge of the genetic architecture of resistance opens the door to addressing wider evolutionary questions. For example, the Red Queen theory states that host–parasite interactions may explain the ubiquity of sex and recombination (Salathé et al. 2008).

Materials and Methods

Study Site

Our study site is the Aegelsee, a pond near Frauenfeld, Switzerland (code: CH-H for Hohliberg; coordinates: 47.557769 N, 8.862783 E, about 30,000 m^2 surface area) where *D. magna* is estimated to have a census population size over ten million individuals and an overwintering resting egg bank of about the same size. Every year from early October, the pond is used as a waste repository by a sugar factory: they progressively lower the water level from May to September and from October, warm ammoniacal condensation water is released into the pond, warming the water temporarily to 40–60 °C (Seefeldt and Ebert 2019) and killing all zooplankton, but not the resting eggs. In winter the pond usually freezes over, and in April, *Daphnia* and other invertebrates hatch from resting eggs. We sampled the pond in February 2014 and March 2015 and did not find *D. magna*, suggesting little or no overwintering. Besides *D. magna*, the plankton community includes *D. pulex*, *D. curvirostis* and a diverse array of other invertebrates, among them copepods, ostracods, rotifers, and corixids. The waste-water treatment prevents fish from invading the pond. The *D. magna* population experiences strong yearly epidemics of *P. ramosa*, reaching prevalence of 70–95%. Infections by other parasites were only rarely observed. The other *Daphnia* species in the pond were never observed to be infected by *P. ramosa*.

Temporal Monitoring

In 2014 and 2015, we sampled the host population every 2–3 weeks from early April to early October to study the impact of the pathogen epidemics. For each sampling date, we aimed to obtain about 100 cloned host lines (produced as iso-female lines). To achieve this, we randomly took about 200–300 female *D. magna* from the sample, placed them in 80-ml jars filled with ADaM (Artificial *Daphnia* Medium, Klüttgen et al. 1994, as modified by Ebert 1998) and let them reproduce asexually. Oversampling was necessary during the hot summer months, as many animals would die for unknown reasons within 48 h under laboratory conditions. This mortality was, to the best of our knowledge, not disease related. Over the following 3 weeks, we screened animals for *P. ramosa* infections by checking for the typical signs of disease: gigantism, reddish-brownish opaque body coloration, and empty brood pouch. Infected animals that had not yet reproduced asexually were treated with tetracycline (50 mg l^{-1}) (an antibiotic which kills Gram-positive bacteria) until an asexual clutch was observed, usually after about 2 weeks. They were fed 25 million cells of the unicellular green algae *Scenedesmus* sp. three times a week, and the medium was renewed every 2 weeks. Feeding and fresh medium protocols were adapted according to the size and number of animals in a jar when necessary.

Resistotype Assessment: The Attachment Test

We assessed resistance phenotype (resistotype) for all cloned hosts using four *P. ramosa* isolates (C1, C19, P15, and P20). We isolated the parasite, P20, from our study population at the start of the epidemic on May 13, 2011 and subsequently passaged it three times through a susceptible *D. magna* host clone from the same population. The three other *P. ramosa* clones or isolates had been previously established in the laboratory: C1 (clone), originated from a *D. magna* population in Russia (Moscow), C19 (clone) from Germany (Gaarzerfeld) and P15 (isolate) from Belgium (Heverlee) (Luijckx et al. 2011; Bento et al. 2020). We used these three *P. ramosa* allopatric isolates in the present study to implement our working genetic model for resistance (Luijckx et al. 2012, 2013; Metzger et al. 2016). Parasite transmission stage (= spore) production in the laboratory followed the protocol by Luijckx et al. (2011).

The resistotypes of *D. magna* clones were assessed using a spore attachment test (Duneau et al. 2011). Bacterial spores attach to the foregut or the hindgut of susceptible host clones. Attachment is a prerequisite for subsequent infection. We call these genotypes susceptible, otherwise they are considered resistant. A genotype allowing attachment and penetration of the parasite into the host, may sometimes still resist infection, based on subsequent immune defense (Hall et al. 2019). To test for attachment, we exposed each individual *Daphnia* to 8,000 (C1, C19) or 10,000 (P15, P20) fluorescently labeled spores following the protocol of Duneau et al. (2011). We used higher spore doses for P15 and P20 because previous observations had shown that fewer of these isolate spores attach to the host esophagus, resulting in a weaker fluorescent signal. Three repeats were used for C1, C19, and P15, whereas six to nine repeats were used for P20. A clone was considered susceptible to the bacterial isolate when more than half of its replicates showed clear attachment. Its overall resistotype is the combination of its resistotypes to the four individual *P. ramosa* isolates in the following order: C1, C19, P15, and P20, for example, a clone susceptible to all four isolates would have resistotype SSSS. Since resistance to P15 had low variability in our study population, this isolate was only considered in the first experiment presented here and was otherwise represented with the placeholder "␣", for example, "RR␣R resistotype."

Experimental Infections of Resistotypes

As an initial assessment of the parasite's fitness impact on the host population, we conducted experimental infections on a representative sample of the spring 2014 host population. We collected surface sediment from five different points in the pond in February 2014, before onset of the natural hatching season and placed 100 *D. magna* ephippia from each replicate in 80-l containers with 30 l ADaM. The five containers were placed outdoors under direct sunlight and checked for hatchlings every 2 days. We recorded hatching dates and cloned hatchlings in the laboratory where we then scored their resistotypes. For the infection experiment, we used parasites collected from the ongoing epidemic in the pond. We collected three pools of 20 randomly chosen infected individuals during the first phase of the epidemic in early June 2014. These field-infected animals were kept in the laboratory under ad libitum feeding conditions. Shortly before their expected death, we pooled all 60 animals, homogenized them to produce a spore suspension, and froze it at −20 °C. A placebo suspension was produced from 60 homogenized uninfected *D. magna*. The parasite spore mixture was not passaged before we used it, so, in contrast to the isolates used for the attachment test, it represents a population sample of the parasite.

Among the four predominant resistotypes, we observed in the cloned cohort of spring hatchlings (SSSS, RRSS, RRSR, and RRRR), we used 20 clones each from the more common resistotypes SSSS, RRSS, and RRSR and ten of the less common resistotype RRRR for an infection experiment, due to limited availability. From each of these 70 clones, we produced five replicate lines, and these 350 lines were maintained individually in 80-ml jars. To reduce maternal effects before the experiment, we kept all lines for three generations in the same experimental conditions: 20 °C, 16:8 light:dark cycle, ADaM medium, and daily ad libitum feeding of 8 million *Scenedesmus* sp. cells per jar. The three generations were produced as follows: as soon as a female produced a clutch, she was discarded and the offspring were kept. When these offspring were mature, a single female was kept in the jar until she in turn produced a clutch. The medium was changed every 4 days or when the females released offspring. We exposed 2- to 3-day-old juveniles from all replicates to the parasite spore suspension by placing individual *D. magna* in 10 ml of medium with 10,000 spores. Additionally, three controls from the third-generation offspring were randomly taken from among the five replicates for each clone ($n = 210$) and were exposed to the equivalent volume of placebo suspension. Three days after exposure, the jars were filled to 80 ml. Medium was changed after ten days, and then every 4 days until the end of the experiment. Jars were monitored daily for 35 days. We recorded infection occurrence, clutch number, and time when visible signs of infection were observed. Controls did not get infected and produced offspring at regular intervals.

We tested both the effects of the full resistotype and of the P20 resistotype only on the three dependent variables: infection (binary: 1/0), clutch number (integer), and time of infection (continuous). Replicates were nested within clones, which were nested within resistotypes. We fitted general linear models using binomial data family type for infection and quasi-Poisson for clutch number and time to infection. For clutch size and time to infection, only data on infected individuals were used.

Infection Phenotypes of Field-Collected Hosts

As a second assessment of the impact of the local parasite on the host population, we measured fitness traits of animals caught during the epidemics. Because the infection experiment described above (carried out in the previous year) indicated that P20 played a strong role, we focused on this parasite isolate. On June 7 and 28, 2015, we collected large *D. magna* samples from our study site and measured body length, from the top of the head through the eye to the base of the tail spine. We kept all females ($n = 331$) individually under ad libitum feeding conditions, each in about 80 ml medium. We recorded clutches (time and size) and the onset time of disease symptoms over the following 3 weeks. After parasitic castration was evident, we cured animals with tetracycline. These data have also been reported in a paper describing the disease phenotype under natural conditions (Savola and Ebert 2019). The current data set is however smaller than the published data, as we report here only those animals for which we were able to score the resistotypes.

Using generalized linear models, we tested the effect of the P20 resistotype on infection and fecundity, taking body size into account. Sampling date was included as a fixed effect since there are only two sampling dates. Interaction terms were excluded from the model when not significant ($P > 0.1$). We fitted a general linear model using quasibinomial data family type for infection, and a negative binomial generalized linear model for total fecundity (R packages used: MASS: Venables and Ripley 2002, lme4: Bates et al. 2015).

Genome-Wide Association Study

Because our experiments revealed that resistance to P20 plays a major role in the disease dynamics in both laboratory experiments and the field, we used a genome-wide association approach to investigate the genetic architecture of resistance with 37 clones that presented the three most common resistotypes in our study population, excluding P15 resistotype ($n = 16$ RR_R, 10 RR_S, and 11 SS_S). All 37 clones were derived directly from our study population (supplementary table S3, Supplementary Material online).

Whole-Genome DNA Extraction, Sequencing, and Bioinformatics

To remove microbial DNA, individuals were treated for 72 h with three antibiotics (streptomycin, tetracycline, ampicillin at a concentration of 50 mg l^{-1} each in filtered water) and fed twice daily with 200 µl of a dextran bead solution (Sephadex G-25 Superfine by Sigma Aldrich: 20–50 µm diameter at a concentration of 5 g l^{-1}) to remove algae from the gut. DNA was extracted from 15 to 20 adult animals using an isopropanol precipitation protocol (QIAGEN DNeasy Blood

& Tissue Kit). Paired-end 125-cycle sequencing was performed on an Illumina HiSeq 2000.

$$P_{\text{corrected}} = P_{\chi 2}\left(\frac{\chi 2}{\lambda(\chi 2_{\text{LG}\neq 3\&5})}\right) \quad (1)$$

Raw reads were aligned using BWA MEM (Li and Durbin 2009) on the *D. magna* draft genome (v.2.4) and a genetic map (Routtu et al. 2014). BAM alignment files were then filtered for quality, and PCR duplicates were removed using PICARD tools (http://broadinstitute.github.io/picard/, last accessed December 2020). Variant calling was performed using freebayes (v. 0.9.15-1). VCF files were then filtered using VCFTOOLS v. 0.1.12b (Danecek et al. 2011) to include SNPs with a minimum quality of 20, a minimum genotype quality of 30, and a mean sequencing depth between 10× and 50×. Only SNPs that passed filters in every clone sample were included in subsequent analyses, resulting in a data set of 510,087 SNPs. Association analyses were performed using the command "-assoc" in PLINK (Purcell et al. 2007), which compares allele counts between cases and controls and outputs a *P* value from a χ^2 test with one degree of freedom. Five pairwise comparisons were performed to identify possible candidates for resistance to C1, C19, and P20: (i) SS␣␣ versus RR␣␣, (ii) SS␣S versus RR␣S, (iii) ␣␣␣S versus ␣␣␣R, (iv) RR␣S versus RR␣R, and (v) SS␣S versus RR␣R. We corrected for the genomic inflation of *P* values (λ) that may have resulted from relatedness between samples using the R package GenABEL (Aulchenko et al. 2007). Lambda was calculated excluding SNPs from linkage groups 3 and 5, since these scaffolds displayed an excess of strongly associated markers. We divided raw χ^2 scores by λ to obtain corrected *P* values using R commands "pchisq" and "estlambda." For each SNP:

Histograms of corrected *P* values were examined to confirm their uniform distribution. We estimated the minimum false discovery rate incurred when a given *P* value was identified as significant (so-called *q*-value) from the set of corrected *P* values using the R package "qvalue" (Storey et al. 2015).

$$Q = q\text{value}(P_{\text{corrected}}) \quad (2)$$

The minimum significant threshold for a given association was then calculated as the maximum corrected *P* value with a *q*-value <5%.

$$P_{\text{lim}} = \max(P_{Q<0.05}) \quad (3)$$

The "gg.manhattan" function in R was used to display manhattan plots of the comparisons between different resistotypes (https://github.com/timknut/gg.manhattan/, last accessed December 2020). We used BEDTOOLS (v 2.25.0) to extract genes found in the associated candidate regions, using the 2011 annotation of the genome (available at: wfleabase.org, last accessed December 2020).

Assessment of Resistotype Segregation

The genetic model that resulted from the GWAS analysis allowed us to make predictions about the segregation of resistotypes in sexually reproducing *D. magna* lines. To test these predictions, we selfed *D. magna* clones with different resistotypes. Selfing is possible with *D. magna* because the same clonal line can produce sons (asexual production) as well as eggs by sexual production. The latter need fertilization by males. The resulting sexual eggs must undergo an obligatory resting phase before they can hatch (Ślusarczyk et al. 2019). The resistotypes of the selfed offspring (F1) were examined to assess whether their segregation matched expectations from the genetic model derived from the GWAS.

All clones used for the genetic crosses derived from the study population. We selfed five to ten *D. magna* clones of the three common resistotypes (RR␣R, RR␣S, and SS␣S) and two clones of a rare resistotype (SR␣S), following the protocol from Luijckx et al. (2012). Hatching of selfed offspring is not always successful, resulting in uneven sample sizes. We obtained between 19 and 89 selfed offspring from each of 22 parent clones (supplementary table S19, Supplementary Material online). Their resistotypes were assessed with the attachment test. Samples from each clonal line were stored at −20 °C for future DNA extraction and genotyping.

Predictions of Segregation Patterns

We compared the resistotype segregation patterns in the selfed offspring with predictions in our genetic model. To calculate proportions of expected phenotypes, we developed an R package called "peas" (https://github.com/JanEngelstaedter/peas, last accessed December 2020) that enables the user to predict distributions of offspring genotypes and phenotypes in complex genetic models with Mendelian inheritance (supplementary docs. S1 and S2, Supplementary Material online). We compared these predictions to the segregation patterns from our selfed offspring using the Cochran–Mantel–Haenszel (C–M–H) test for repeated tests of independence. The C–M–H test is applied either to 2 × 2 tables and outputs a chi-square statistic (χ^2) or to larger tables (generalized C–M–H test), where it outputs a M^2 statistic. When there was only one repeat per parent genotype, we used the Fisher's test. When there was only one category of expected and observed phenotype (i.e., no segregation), no test was possible, and expectation and observation showed a perfect match. Following each C–M–H test, assumption of homogeneity of the odds ratio across repeats was confirmed using a Breslow-Day test (R package DescTools: Signorell et al. 2018). However, this test can only be used with 2 × 2 tables. We ran a Fisher's test of independence on each comparison (expected vs. observed for each repeat, Bonferroni corrected) to detect differences in opposite directions across repeats, which would have resulted in a nonsignificant C–M–H test, but no such differences in direction were detected (see supplementary table S15, Supplementary Material online for detailed results of statistical analyses). Tests were run on counts, but for better illustration, we present here segregation of offspring as proportions.

Linking the Phenotype to the Genotype

We designed PCR-based diagnostic markers physically linked to the resistance loci that the GWAS identified (DMPR1 to 4 for "*Daphnia magna*–*Pasteuria ramosa*" markers, supplementary table S20, Supplementary Material online) and tested if these markers (and their corresponding resistance regions) are indeed associated with the resistotypes, by comparing expected and observed association between marker genotypes and resistotypes (supplementary tables S16–S18, Supplementary Material online). We then used these markers to confirm genotyping of the selfed parents.

DNA Extraction and PCR-Based Markers Analysis

DNA of parents and selfed offspring was extracted on 96-well PCR plates using a 10% Chelex bead solution (Bio-Rad) adapted from Walsh et al. (1991). First, individuals were crushed in the wells with 20 µl of deionized water using a customized rack of metallic pestles. We added 150 µl of 10% Chelex solution and 10 µl of proteinase K and incubated samples for 2 h at 55 °C followed by 10 min at 99 °C. Fragment amplification, genotyping, and allele scoring were done following the protocol described in Cabalzar et al. (2019) (see supplementary table S21, Supplementary Material online for PCR reaction details).

Statistical Software

Unless otherwise stated, all statistical analyses and graphics were performed using R software version 3.6.1 (R Core Team 2019). Graphics were edited in Inkscape v. 1.0.1 (https://inkscape.org/, last accessed December 2020). Mean values are presented with standard error: mean ± SE (Package RVAideMemoire v. 09-45-2, Hervé 2015). Packages used in R for package installation, data manipulation, and graphics are the following: package development, documentation, and installation: devtools v. 2.2.1 (Wickham, Hester, et al. 2019) and roxygen2 v. 6.1.1 (Wickham et al. 2018), data manipulation: dplyr v. 0.8.3 (Wickham, Francois, et al. 2019), tidyr v. 1.0.0 (Wickham and Henry 2019), tidyquant v. 0.5.8 (Dancho and Vaughan 2019), tidyverse v. 1.2.1 (Wickham 2017), xlsx v. 0.6.1 (Dragulescu and Arendt 2018), graphics: ggplot2 v. 3.3.0 (Wickham 2016), extrafont v. 0.17 (Chang 2014), scales v. 1.0.0 (Wickham 2018), cowplot v. 1.0.0 (Wilke 2019), gridExtra v. 2.3 (Auguie 2017), ggpubr v. 0.2.3 (Kassambara 2019), ggplotify v. 0.0.4 (Yu 2019), magick v. 2.2 (Ooms 2019), egg v. 0.4.5 (Auguie 2019), ggsci v. 2.9 (Xiao 2018), and png v. 0.1-7 (Urbanek 2013).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Jürgen Hottinger, Urs Stiefel, Kristina Müller, Michelle Krebs, Samuel Pichon, Dita Vizoso, and Jelena Rakov for help in the field and laboratory. Sequencing for the GWAS analysis was performed at the Genomics Facility at the Department of Biosystem Science and Engineering (D-BSSE, ETH) in Basel. The AB3130xl sequencer used for the markers analysis was operated by Nicolas Boileau. Members of the Ebert group, Jonathon Stillman and Luís Teixeira provided valuable feedback on the study and the manuscript. Suzanne Zweizig improved the language of the manuscript. Two anonymous reviewers provided a thoughtful and thorough review of the manuscript. This work was supported by the Swiss National Science Foundation (SNSF) (Grant No. 310030B_166677 to D.E.); the Freiwillige Akademische Gesellschaft Basel (FAG) to C.A.; the University of Basel to D.E. and C.A.; and the Australian Research Council through a Future Fellowship (FT140100907 to J.E.).

Author Contributions

D.E. and J.A. designed the overall study. F.V. and D.E. designed the infection experiment. F.V. conducted the infection experiment and analyzed the data. E.S. and D.E. designed the fitness measurements. E.S. conducted the fitness measurements and analyzed the data. Y.B. and D.E. designed the GWAS analysis. Y.B. conducted the GWAS analysis. C.A. and D.E. designed the crossings. C.A. conducted the crossings and analyzed the data. J.E. developed the "peas" R package. C.A. analyzed the data, wrote the manuscript, and designed the figures. All authors reviewed the manuscript.

Data Availability

The data underlying this article and analysis scripts are available in the Figshare Repository https://doi.org/10.6084/m9.figshare.13259828.v1. Raw sequence data from the GWAS analysis are deposited in the NCBI Bioproject PRJNA680821.

References

Alves JM, Carneiro M, Cheng JY, Lemos de Matos A, Rahman MM, Loog L, Campos PF, Wales N, Eriksson A, Manica A, et al. 2019. Parallel adaptation of rabbit populations to myxoma virus. *Science* 363(6433):1319–1326.

Andras JP, Ebert D. 2013. A novel approach to parasite population genetics: experimental infection reveals geographic differentiation, recombination and host-mediated population structure in *Pasteuria ramosa*, a bacterial parasite of *Daphnia*. *Mol Ecol*. 22(4):972–986.

Andras JP, Fields PD, Du Pasquier L, Fredericksen M, Ebert D. 2020. Genome-wide association analysis identifies a genetic basis of infectivity in a model bacterial pathogen. *Mol Biol Evol*.

Auguie B. 2017. gridExtra: miscellaneous functions for "Grid" graphics. Available from: https://CRAN.R-project.org/package=gridExtra. Last accessed on December 2020.

Auguie B. 2019. egg: extensions for "ggplot2": custom geom, custom themes, plot alignment, labelled panels, symmetric scales, and fixed panel size. Available from: https://CRAN.R-project.org/package=egg. Last accessed on December 2020.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23(10):1294–1296.

Bangham J, Knott SA, Kim K-W, Young RS, Jiggins FM. 2008. Genetic variation affecting host-parasite interactions: major-effect quantitative trait loci affect the transmission of sigma virus in *Drosophila melanogaster*. *Mol Ecol*. 17(17):3800–3807.

Bates D, Maechler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 67:1–48.

Ben-Ami F. 2019. Host age effects in invertebrates: epidemiological, ecological, and evolutionary implications. *Trends Parasitol*. 35(6):466–480.

Bento G, Fields PD, Duneau D, Ebert D. 2020. An alternative route of bacterial infection associated with a novel resistance locus in the *Daphnia–Pasteuria* host–parasite system. *Heredity* 125(4):173–183.
Bento G, Routtu J, Fields PD, Bourgeois Y, Du Pasquier L, Ebert D. 2017. The genetic basis of resistance and matching-allele interactions of a host-parasite system: the *Daphnia magna-Pasteuria ramosa* model. *PLoS Genet*. 13(2):e1006596.
Bourgeois Y, Roulin AC, Müller K, Ebert D. 2017. Parasitism drives host genome evolution: insights from the *Pasteuria ramosa–Daphnia magna* system. *Evolution* 71(4):1106–1113.
Cabalzar AP, Fields PD, Kato Y, Watanabe H, Ebert D. 2019. Parasite-mediated selection in a natural metapopulation of *Daphnia magna*. *Mol Ecol*. 28(21):4770–4785.
Cao C, Magwire MM, Bayer F, Jiggins FM. 2016. A polymorphism in the processing body component Ge-1 controls resistance to a naturally occurring Rhabdovirus in Drosophila. *PLoS Pathog*. 12:e1005387.
Carton Y, Nappi AJ, Poirie M. 2005. Genetics of anti-parasite resistance in invertebrates. Dev. Comp. Immunol. 29(1):9–32.
Cerqueira GC, Cheeseman IH, Schaffner SF, Nair S, McDew-White M, Phyo AP, Ashley EA, Melnikov A, Rogov P, Birren BW, et al. 2017. Longitudinal genomic surveillance of *Plasmodium falciparum* malaria parasites reveals complex genomic architecture of emerging artemisinin resistance. *Genome Biol*. 18(1):13.
Chang W. 2014. extrafont: tools for using fonts. Available from: https://CRAN.R-project.org/package=extrafont. Last accessed on December 2020.
Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet*. 2(4):e64.
Cogni R, Cao C, Day JP, Bridson C, Jiggins FM. 2016. The genetic architecture of resistance to virus infection in *Drosophila*. *Mol Ecol*. 25(20):5228–5241.
Conlon BH, Frey E, Rosenkranz P, Locke B, Moritz RFA, Routtu J. 2018. The role of epistatic interactions underpinning resistance to parasitic *Varroa* mites in haploid honey bee (*Apis mellifera*) drones. *J Evol Biol*. 31(6):801–809.
Connallon T, Chenoweth SF. 2019. Dominance reversals and the maintenance of genetic variation for fitness. *PLoS Biol*. 17(1):e3000118.
Daborn PJ. 2002. A single P450 allele associated with insecticide resistance in *Drosophila*. *Science* 297(5590):2253–2256.
Dancho M, Vaughan D. 2019. tidyquant: tidy quantitative financial analysis. Available from: https://CRAN.R-project.org/package=tidyquant. Last accessed on December 2020.
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158.
Decaestecker E, Gaba S, Raeymaekers JAM, Stoks R, Van Kerckhoven L, Ebert D, De Meester L. 2007. Host–parasite 'Red Queen' dynamics archived in pond sediment. *Nature* 450(7171):870–873.
Dragulescu AA, Arendt C. 2018. xlsx: read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. Available from: https://CRAN.R-project.org/package=xlsx. Last accessed on December 2020.
Duffy MA, Sivars-Becker L. 2007. Rapid evolution and ecological host-parasite dynamics. *Ecol Lett*. 10(1):44–53.
Duncan AB, Little TJ. 2007. Parasite-driven genetic change in a natural population of *Daphnia*. *Evolution* 61(4):796–803.
Duneau D, Luijckx P, Ben-Ami F, Laforsch C, Ebert D. 2011. Resolving the infection process reveals striking differences in the contribution of environment, genetics and phylogeny to host-parasite interactions. *BMC Biol*. 9(1):1–11.
Ebert D. 1998. Experimental evolution of parasites. *Science* 282(5393):1432–1436.
Ebert D. 2005. Ecology, epidemiology, and evolution of parasitism in *Daphnia*. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.
Ebert D, Duneau D, Hall MD, Luijckx P, Andras JP, Du Pasquier L, Ben-Ami F. 2016. A population biology perspective on the stepwise infection process of the bacterial pathogen *Pasteuria ramosa* in *Daphnia*. *Adv Parasitol*. 91:265–310.
Ebert D, Fields PD. 2020. Host–parasite co-evolution and its genomic signature. *Nat Rev Genet*. 21:754–768.
Ebert D, Rainey P, Embley TM, Scholz D. 1996. Development, life cycle, ultrastructure and phylogenetic position of *Pasteuria ramosa* Metchnikoff 1888: rediscovery of an obligate endoparasite of *Daphnia magna* Straus. *Philos Trans R Soc B Biol Sci*. 351:1689–1701.
Ellegren H, Sheldon BC. 2008. Genetic basis of fitness differences in natural populations. *Nature* 452(7184):169–175.
Engelstädter J. 2015. Host-Parasite coevolutionary dynamics with generalized success/failure infection genetics. *Am Nat*. 185(5):E117–E129.
Engelstädter J, Bonhoeffer S. 2009. Red Queen dynamics with non-standard fitness interactions. *PLoS Comput Biol*. 5:e1000469.
Feldman MW, Lewontin RC, Frankin IR, Christiansen RB. 1975. Selection in complex genetic systems III. An effect of allele multiplicity with two loci. *Genetics* 79(2):333–347.
Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. *Mol Ecol*. 24(14):3529–3545.
Galvani AP. 2003. Epidemiology meets evolutionary ecology. *Trends Ecol Evol*. 18(3):132–139.
Garbutt JS, O'Donoghue AJP, McTaggart SJ, Wilson PJ, Little TJ. 2014. The development of pathogen resistance in *Daphnia magna*: implications for disease spread in age-structured populations. *J Exp Biol*. 217(21):3929–3934.
Gibson AK, Delph LF, Vergara D, Lively CM. 2018. Periodic, parasite-mediated selection for and against sex. *Am Nat*. 192(5):537–551.
Gómez-Gómez L, Felix G, Boller T. 1999. A single locus determines sensitivity to bacterial flagellin in *Arabidopsis thaliana*. *Plant J*. 18(3):277–284.
González AM, Yuste-Lisbona FJ, Rodiño AP, De Ron AM, Capel C, García-Alcázar M, Lozano R, Santalla M. 2015. Uncovering the genetic architecture of *Colletotrichum lindemuthianum* resistance through QTL mapping and epistatic interaction analysis in common bean. *Front Plant Sci*. 6:1–13.
González-Tortuero E, Rusek J, Turko P, Petrusek A, Maayan I, Piálek L, Tellenbach C, Gieler S, Spaak P, Wolinska J. 2016. *Daphnia* parasite dynamics across multiple *Caullerya* epidemics indicate selection against common parasite genotypes. *Zoology* 119(4):314–321.
Hahn MW. 2018. Molecular population genetics. New York: Oxford University Press.
Hall MD, Routtu J, Ebert D. 2019. Dissecting the genetic architecture of a stepwise infection process. *Mol Ecol*. 28:1–16.
Hamilton WD. 1980. Sex versus non-sex versus parasite. *Oikos* 35(2):282.
Hamilton WD, Axelrod R, Tanese R. 1990. Sexual reproduction as an adaptation to resist parasites (a review). *Proc Natl Acad Sci U S A*. 87(9):3566–3573.
Hervé M. 2015. RVAideMemoire: diverse basic statistical and graphical functions. Available from: http://CRAN.R-project.org/package=RVAideMemoire. Last accessed on December 2020.
Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed LK, Storfer A, Whitlock MC. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am Nat*. 188(4):379–397.
Hooker AL, Saxena KMS. 1971. Genetics of disease resistance in plants. *Annu Rev Genet*. 5(1):407–424.
Howard RS, Lively CM. 1998. The maintenance of sex by parasitism and mutation accumulation under epistatic fitness functions. *Evolution* 52(2):604–610.
Izhar R, Ben-Ami F. 2015. Host age modulates parasite infectivity, virulence and reproduction. *J Anim Ecol*. 84(4):1018–1028.
Jones AG, Bürger R, Arnold SJ. 2014. Epistasis and natural selection shape the mutational architecture of complex traits. *Nat Commun*. 5:1–10.
Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* 477(7363):203–206.

Juneja P, Ariani CV, Ho YS, Akorli J, Palmer WJ, Pain A, Jiggins FM. 2015. Exome and transcriptome sequencing of *Aedes aegypti* identifies a locus that confers resistance to *Brugia malayi* and alters the immune response. *PLoS Pathog*. 11(3):e1004765.
Kassambara A. 2019. ggpubr: "ggplot2" based publication ready plots. Available from: https://CRAN.R-project.org/package=ggpubr. Last accessed on December 2020.
Klüttgen B, Dülmer U, Engels M, Ratte HT. 1994. ADaM, an artificial freshwater for the culture of zooplankton. *Water Res*. 28(3):743–746.
Koskella B. 2018. Resistance gained, resistance lost: an explanation for host–parasite coexistence. *PLoS Biol*. 16(9):e3000013.
Kouyos RD, Salathé M, Otto SP, Bonhoeffer S. 2009. The role of epistasis on the evolution of recombination in host–parasite coevolution. *Theor Popul Biol*. 75(1):1–13.
Kover PX, Caicedo AL. 2001. The genetic architecture of disease resistance in plants and the maintenance of recombination by parasites. *Mol Ecol*. 10(1):1–16.
Kurtz J, Schulenburg H, Reusch TBH. 2016. Host-parasite coevolution-rapid reciprocal adaptation and its genetic basis. *Zoology* 119(4):241–243.
Laine A-L. 2009. Role of coevolution in generating biological diversity: spatially divergent selection trajectories. *J Exp Bot*. 60(11):2957–2970.
Li C, Cowling W. 2003. Identification of a single dominant allele for resistance to blackleg in *Brassica napus* "Surpass 400". *Plant Breed*. 122(6):485–488.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
Li W, Zhu Z, Chern M, Yin J, Yang C, Ran L, Cheng M, He M, Wang K, Wang J, et al. 2017. A natural allele of a transcription factor in rice confers broad-spectrum blast resistance. *Cell* 170(1):114–126.e15.
Little TJ, Ebert D. 1999. Associations between parasitism and host genotype in natural populations of *Daphnia* (Crustacea: Cladocera). *J Anim Ecol*. 68(1):134–149.
Lively CM. 2010. A review of Red Queen models for the persistence of obligate sexual reproduction. J. Hered. 101:S13–S20.
Llaurens V, Whibley A, Joron M. 2017. Genetic architecture and balancing selection: the life and death of differentiated variants. *Mol Ecol*. 26(9):2430–2448.
Luijckx P, Ben-Ami F, Mouton L, Du Pasquier L, Ebert D. 2011. Cloning of the unculturable parasite *Pasteuria ramosa* and its *Daphnia* host reveals extreme genotype-genotype interactions. *Ecol Lett*. 14(2):125–131.
Luijckx P, Fienberg H, Duneau D, Ebert D. 2012. Resistance to a bacterial parasite in the crustacean *Daphnia magna* shows Mendelian segregation with dominance. *Heredity* 108(5):547–551.
Luijckx P, Fienberg H, Duneau D, Ebert D. 2013. A matching-allele model explains host resistance to parasites. *Curr Biol*. 23(12):1085–1088.
Magalhães S, Sucena É. 2016. Genetics of host–parasite interactions: towards a comprehensive dissection of *Drosophila* resistance to viral infection. *Mol Ecol*. 25(20):4981–4983.
Magwire MM, Fabian DK, Schweyen H, Cao C, Longdon B, Bayer F, Jiggins FM. 2012. Genome-wide association studies reveal a simple genetic basis of resistance to naturally coevolving viruses in Drosophila melanogaster. *PLoS Genet*. 8:e1003057.
Metzger CMJA, Luijckx P, Bento G, Mariadassou M, Ebert D. 2016. The Red Queen lives: epistasis between linked resistance loci. *Evolution* 70(2):480–487.
Mitchell SE, Read AF, Little TJ. 2004. The effect of a pathogen epidemic on the genetic structure and reproductive strategy of the crustacean *Daphnia magna*. *Ecol Lett*. 7(9):848–858.
Morgan AD, Koskella B. 2017. Coevolution of host and pathogen (Second Edition). In: Tibayrenc M, editor. Genetics and evolution of infectious diseases. Elsevier. Sara Tenney. p. 115–140. Available from: http://linkinghub.elsevier.com/retrieve/pii/B9780127999425000068. Last accessed on December 2020.
Ooms J. 2019. magick: advanced graphics and image-processing in R. Available from: https://CRAN.R-project.org/package=magick. Last accessed on December 2020.
Otto SP. 2009. The evolutionary enigma of sex. *Am Nat*. 174(S1):S1–S14.
Pilet-Nayel M-L, Moury B, Caffier V, Montarry J, Kerlan M-C, Fournet S, Durel C-E, Delourme R. 2017. Quantitative resistance to plant pathogens in pyramiding strategies for durable crop protection. *Front Plant Sci*. 8:1838.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81(3):559–575.
R Core Team. 2019. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org. Last accessed on December 2020.
Rogalski MA, Duffy MA. 2020. Local adaptation of a parasite to solar radiation impacts disease transmission potential, spore yield, and host fecundity. Evolution 74:1856–1864.
Routtu J, Ebert D. 2015. Genetic architecture of resistance in *Daphnia* hosts against two species of host-specific parasites. *Heredity* 114(2):241–248.
Routtu J, Hall MD, Albere B, Beisel C, Bergeron RD, Chaturvedi A, Choi J-H, Colbourne J, De Meester L, Stephens MT, et al. 2014. An SNP-based second-generation genetic map of *Daphnia magna* and its application to QTL analysis of phenotypic traits. *BMC Genomics* 15(1):1033–1015.
Saavedra-Rodriguez K, Strode C, Flores Suarez A, Fernandez Salas I, Ranson H, Hemingway J, Black WC. 2008. Quantitative trait loci mapping of genome regions controlling permethrin resistance in the mosquito *Aedes aegypti*. *Genetics* 180(2):1137–1152.
Salathé M, Kouyos R, Bonhoeffer S. 2008. The state of affairs in the kingdom of the Red Queen. *Trends Ecol Evol*. 23(8):439–445.
Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber C-M, Saragosti S, Lapouméroulie C, Cognaux J, Forceille C, et al. 1996. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 382(6593):722–725.
Sasaki A. 2000. Host-parasite coevolution in a multilocus gene-for-gene system. *Proc R Soc Lond B*. 267(1458):2183–2188.
Savola E, Ebert D. 2019. Assessment of parasite virulence in a natural population of a planktonic crustacean. *BMC Ecol*. 19(1):14.
Schlesinger KJ, Stromberg SP, Carlson JM. 2014. Coevolutionary immune system dynamics driving pathogen speciation. *PLoS One* 9:e102821.
Schmid-Hempel P. 2011. Evolutionary parasitology. The integrated study of infections, immunology, ecology, and genetics. New York: Oxford University Press.
Seefeldt L, Ebert D. 2019. Temperature- versus precipitation-limitation shape local temperature tolerance in a Holarctic freshwater crustacean. *Proc R Soc B*. 286(1907):20190929.
Shocket MS, Vergara D, Sickbert AJ, Walsman JM, Strauss AT, Hite JL, Duffy MA, Cáceres CE, Hall SR. 2018. Parasite rearing and infection temperatures jointly influence disease transmission and shape seasonality of epidemics. *Ecology* 99(9):1975–1987.
Signorell A, et al. 2018. DescTools: tools for descriptive statistics. R package version 0.99.26.
Ślusarczyk M, Chlebicki W, Pijanowska J, Radzikowski J. 2019. The role of the refractory period in diapause length determination in a freshwater crustacean. *Sci Rep*. 9:11905.
Storey J, Bass A, Dabney A, Robinson D. 2015. qvalue: Q-value estimation for false discovery rate control. *Available from: http://github.com/jdstorey/qvalue. Last accessed on December 2020.*
Strauss AT, Hite JL, Shocket MS, Cáceres CE, Duffy MA, Hall SR. 2017. Rapid evolution rescues hosts from competition and disease but—despite a dilution effect—increases the density of infected hosts. *Proc R Soc B*. 284(1868):20171970.
Tellier A, Brown JKM. 2007. Polymorphism in multilocus host-parasite coevolutionary interactions. *Genetics* 177(3):1777–1790.
Turko P, Tellenbach C, Keller E, Tardent N, Keller B, Spaak P, Wolinska J. 2018. Parasites driving host diversity: incidence of

disease correlated with *Daphnia* clonal turnover. *Evolution* 72(3):619–629.
Urbanek S. 2013. png: read and write PNG images. Available from: https://CRAN.R-project.org/package=png. Last accessed on December 2020.
Vale PF, Wilson AJ, Best A, Boots M, Little TJ. 2011. Epidemiological, evolutionary, and coevolutionary implications of context-dependent parasitism. *Am Nat*. 177(4):510–521.
van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* 534(7605):102–105.
Venables WN, Ripley BD. 2002. Modern applied statistics with S. 4th ed. New York: Springer.
Walsh PS, Metzger DA, Higuchi R. 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques* 10(4):506–513.
Whitlock MC, Lotterhos KE. 2015. Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. *Am Nat*. 186(13):S24–S36.
Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.
Wickham H. 2017. tidyverse: easily install and load the "Tidyverse." Available from: https://CRAN.R-project.org/package=tidyverse. Last accessed on December 2020.
Wickham H. 2018. scales: scale functions for visualization. Available from: https://CRAN.R-project.org/package=scales. Last accessed on December 2020.
Wickham H, Danenberg P, Eugster D. 2018. roxygen2: in-line documentation for R. Available from: https://CRAN.R-project.org/package=roxygen2. Last accessed on December 2020.
Wickham H, Francois R, Henry L, Müller K. 2019. dplyr: a grammar of data manipulation. Available from: https://CRAN.R-project.org/package=dplyr. Last accessed on December 2020.
Wickham H, Henry L. 2019. tidyr: tidy messy data. Available from: https://CRAN.R-project.org/package=tidyr. Last accessed on December 2020.
Wickham H, Hester J, Chang W. 2019. devtools: tools to make developing R packages easier. Available from: https://CRAN.R-project.org/package=devtools. Last accessed on December 2020.
Wilfert L, Schmid-Hempel P. 2008. The genetic architecture of susceptibility to parasites. *BMC Evol Biol*. 8(1):187.
Wilke CO. 2019. cowplot: streamlined plot theme and plot annotations for "ggplot2." Available from: https://CRAN.R-project.org/package=cowplot. Last accessed on December 2020.
Wittmann MJ, Bergland AO, Feldman MW, Schmidt PS, Petrov DA. 2017. Seasonally fluctuating selection can maintain polymorphism at many loci via segregation lift. *Proc Natl Acad Sci U S A*. 114(46):E9932–E9941.
Xiao N. 2018. ggsci: scientific journal and sci-fi themed color palettes for "ggplot2." Available from: https://CRAN.R-project.org/package=ggsci. Last accessed on December 2020.
Xiao Y, Dai Q, Hu R, Pacheco S, Yang Y, Liang G, Soberón M, Bravo A, Liu K, Wu K. 2017. A single point mutation resulting in cadherin mislocalization underpins resistance against *Bacillus thuringiensis* toxin in cotton bollworm. *J Biol Chem*. 292(7):2933–2943.
Yu G. 2019. ggplotify: convert plot to "grob" or "ggplot" object. Available from: https://CRAN.R-project.org/package=ggplotify.

Received: 30 December 2020 | Revised: 26 April 2021 | Accepted: 11 May 2021

DOI: 10.1111/mec.15989

SPECIAL ISSUE

MOLECULAR ECOLOGY WILEY

An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes

Yann X. C. Bourgeois[1] | Ben H. Warren[2]

[1]School of Biological Sciences, University of Portsmouth, Portsmouth, UK

[2]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, UA, CP 51, Paris, France

Correspondence
Yann X. C. Bourgeois, School of Biological Sciences, University of Portsmouth, King Henry Building, Portsmouth, UK.
Email: yann.bourgeois@port.ac.uk

Abstract

Characterizing the population history of a species and identifying loci underlying local adaptation is crucial in functional ecology, evolutionary biology, conservation and agronomy. The constant improvement of high-throughput sequencing techniques has facilitated the production of whole genome data in a wide range of species. Population genomics now provides tools to better integrate selection into a historical framework, and take into account selection when reconstructing demographic history. However, this improvement has come with a profusion of analytical tools that can confuse and discourage users. Such confusion limits the amount of information effectively retrieved from complex genomic data sets, and impairs the diffusion of the most recent analytical tools into fields such as conservation biology. It may also lead to redundancy among methods. To address these isssues, we propose an overview of more than 100 state-of-the-art methods that can deal with whole genome data. We summarize the strategies they use to infer demographic history and selection, and discuss some of their limitations. A website listing these methods is available at www.methodspopgen.com.

KEYWORDS
bioinformatics, demography, population genomics, selection, whole-genome sequencing

1 | INTRODUCTION

Comprehensive analyses of species history and selection contribute to our understanding of causation in biology, an effort that has included genetics, developmental science and ecology (Laland et al., 2011). The number of population genomic studies aimed at elucidating the history of natural populations has increased enormously in the last 10 years. A few examples include an improved understanding of the history of human migrations, admixture and adaptation (e.g., Abi-Rached et al., 2011; Li & Durbin, 2011; Sabeti et al., 2002), the origin of domesticated species (e.g., Axelsson et al., 2013; Cubry et al., 2018; Schubert et al., 2014) and the genetic basis of local adaptation (e.g., Kolaczkowski et al., 2011; Kubota et al., 2015; Legrand et al., 2009; Roux et al., 2013). Developments in whole-genome resequencing have continually improved the throughput of genetic data, while reducing the time and cost of their production. Increased data production has been accompanied by a drive to develop efficient computational methods to interpret patterns of genetic variation at the genomic scale. These interconnected developments have allowed species histories to be inferred even when little preliminary knowledge is available. Investigating variation across multiple genomes sampled across populations or closely related species is now a common task for teams studying evolutionary processes, who can rely on a diverse array of methods to infer demography and selection. Such progress has confirmed the value of population genomics in understanding biological diversity, beyond the initial

handful of model species upon which most of the field was built (Abzhanov et al., 2008; Ellegren et al., 2012; Jenner & Wills, 2007; Mandoli & Olmstead, 2000; Poelstra et al., 2014; Weber et al., 2013; White et al., 2010). Such advances are needed to broaden our view of evolutionary processes and improve sampling of distant clades. Ultimately, this should provide a more balanced picture than the one brought by the study of a few model species (Abzhanov et al., 2008). From an applied perspective, genomic approaches also have the potential to improve conservation genetic inference by scaling up the amount of data available (Shafer et al., 2015), understanding the past history of species (Leitwein et al., 2020), and identifying loci and alleles important for local adaptation, which can then be used to define relevant conservation units (Fraser & Bernatchez, 2001).

Much effort has recently been made in facilitating the dissemination of sometimes complex, state-of-the-art methods. Nevertheless, the last comprehensive review of methods for population genetics was performed more than 10 years ago (Excoffier & Heckel, 2006). Recent methodological advances have brought increased analytical complexity to the field, and an inflation in the number of methods covering any one topic. The widespread use of sophisticated analytical tools is made difficult by the lack of communication between fields (Shafer et al., 2015), little user-friendliness of software, inflation of data formats (Lischer & Excoffier, 2012) and the ever-increasing number of methods made available. As a consequence, it has become increasingly difficult for all potential users (and also developers) to follow developments and be sure of selecting the most appropriate method for the question and data at hand. Combining approaches is one of the current grand challenges in evolutionary biology (Cushman, 2014). While large-scale collaborations and sharing of skills between researchers allow for detailed analyses, a global summary of methods that can handle whole-genome resequencing data would be valuable for smaller research teams, so they can quickly start new projects and evaluate their experimental design. It would also facilitate communication between different subfields of evolutionary biology, by providing a common resource that can be used to identify methodological convergence and possible synergies. It may also avoid situations where similar methods are developed in parallel. Furthermore, the issue of anthropogenic environmental change and decline in biodiversity is pressing, and merits enhanced efforts to disseminate methods that can leverage genomic data, ultimately improving our understanding of the response of biodiversity to environmental change. Many conservation practitioners are receptive to using genetic tools, but do not always have access to the relevant expertise (Taylor et al., 2017). A freely accessible methodological summary may be useful in this context.

In this review, we assume that the reader is already familiar with the main concepts and current issues in population genomics, but needs an overview of the different methods associated with these concepts. We promote the idea that multiple approaches must be used and compared in any population genomics project. This has several benefits: it gives the investigator a better idea of the robustness of results and may reveal issues in raw data processing. In addition, different methods aim to detect slightly different signals, and their combination may provide a more comprehensive overview of the processes acting. We aim at providing a resource that, if not fully comprehensive, can act as an efficient starting point for researchers investigating whole-genome variation in the next few years. This article can be used in combination with other recent methodological reviews on selection (Haasl & Payseur, 2016; Koropoulis et al., 2020), demographic inference or simulation-based approaches (Schrider & Kern, 2018; Smith & Flaxman, 2020).

For the sake of simplicity, we divide our review into two sections (Figure 1): (i) methods devoted to the study of population structure and history (Tables 1 and 2), and (ii) detecting signatures of evolutionary processes along the genome (Tables 3 and 4). We end this review by outlining how different analyses can be combined, and present future directions that may be taken by the field of population genomics. We particularly insist on the interest—but also the challenges—of model-based approaches to test specific hypotheses, benchmark different methods and incorporate intrinsic properties of genomes (Table 4). The tables and a summary of the methods discussed in this paper will be kept updated to follow improvements, and are available at www.methodspopgen.com. Contributions are of course welcome, and can be sent to the following email address: methodspopgen@gmail.com.

2 | POPULATION STRUCTURE AND HISTORY

Genetic diversity and its genome-wide variance are directly impacted by variation in many factors including effective population sizes, population structure, inbreeding and migration. Moreover, the effects of selection on diversity at linked sites depends directly on local variation in the recombination rate. All these factors are important to characterize in any study of genome-wide variation. In this section, we describe methods aiming to quantify these aspects (see also Tables 1 and 2).

2.1 | Estimating familial relationships and reconstructing pedigrees

Understanding relatedness and structure both within and between populations is an important starting point for any study making inferences of selection or demographic history. Methods for estimating the relatedness of individuals are suited to studies relying on pedigree information (for example in quantitative genetics studies), or if there are reasons to suspect that familial relationships and inbreeding can play a major role in shaping the genetic structure of the population(s) considered. The most powerful methods in this category are likelihood-based and make use of heterozygous sites in each individual (e.g., COLONY in Wang, 2019, see also the detailed review in Huisman, 2017). Each pedigree configuration can be assigned a likelihood at each locus which depends on the probability of observing a given genotype conditional on the genotypes of assigned parents. Assuming independent loci, a composite likelihood can then be derived for a set of unlinked single nucleotide

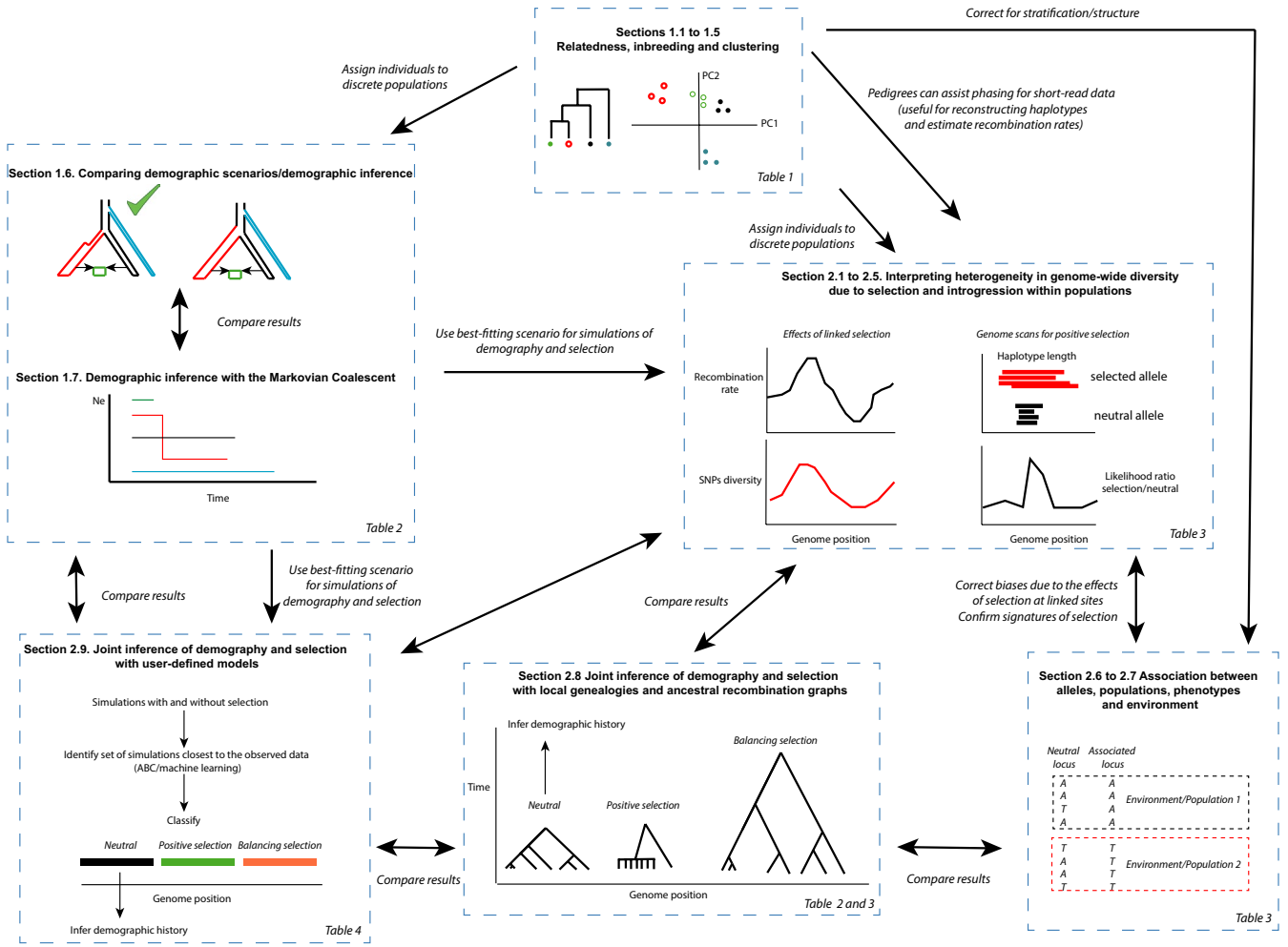


FIGURE 1 Graphical summary of this review. This work is divided into two main sections: the first section covers methods that generally assume neutrality and are generally used for demographic inference. The second section covers methods that aim at identifying loci under the direct and indirect effects of selection. The table listing the relevant methods is indicated at the bottom right of each box. The linear structure of this review does not necessarily reflect the network of possible comparisons between the results of different methods. These possible comparisons are indicated by double arrows. Results obtained from methods listed in different sections can be used to inform the next steps of an analysis (single arrows). We acknowledge that there is no one-size-fits-all pipeline, and elements of this general framework may be entirely omitted from an analysis depending on the research question

polymorphisms (SNPs). Information about pedigrees is important in order to filter out related individuals before carrying other population genetics analyses. Furthermore, mendelian constraints provide important information about haplotypes that can be used by phasing programs. Including related individuals can be useful when attempting to phase genotypes and generate a reference panel for further phasing in unrelated samples. The popular phasing algorithm Shapeit (Delaneau et al., 2019; Williams et al., 2012) can include familial information when reconstructing phased haplotypes.

2.2 | Using nonsupervised models to estimate relatedness and population structure

An elegant and efficient class of methods relies on using multivariate approaches such as principal component analysis (PCA) to infer relatedness between individuals and populations without a priori knowledge. These methods apply a dimension reduction procedure to matrices of individual genotypes, projecting genotypic variability along several axes of variation (Jombart et al., 2009). These approaches have been especially useful to study the consistency between geographical and genetic structure in human populations of Europe (Novembre et al., 2008). Procrustes rotation (Novembre et al., 2008) can be used to match geographical coordinates with PCA axes, showing how isolation by distance has shaped genetic structure. Since these methods do not have underlying assumptions based on (diploid) population genetics, they are suitable for analysing species displaying polyploidy or mixed-ploidy (Dufresne et al., 2014). They go beyond a mere description of data, since projections of individuals on PCA axes can be used to infer admixture proportions, and contain information about demographic processes shaping genetic diversity (McVean, 2009). PCA can be used as a summary of genetic

TABLE 1 Summary of methods dedicated to data description and assessing population structure. VCF: variant call format (see Danecek et al., 2011)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SNMF	Clustering and characterizing admixture	Grouping individuals in clusters maximizing Hardy–Weinberg (HW) equilibrium and LD between loci	Fast (30× than ADMIXTURE)	Still slow computation time for very large data sets	http://membres-timc.imag.fr/Olivier.Francois/snmf/index.htm	Frichot et al. (2014)
STRUCTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	User-friendly interface. Bayesian inference	Not suited for large whole genomes. Requires specific input format. Might be used on a small set of high-quality markers for small genomes	http://pritchardlab.stanford.edu/structure.html	Pritchard et al. (2000)
FASTSTRUCTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	~100× faster than STRUCTURE	Approximate inference of the original STRUCTURE model	http://rajanil.github.io/fastStructure/	Raj et al. (2014)
ADMIXTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	Maximum likelihood, faster than STRUCTURE. Can handle sex-linked markers	Often slower than its counterparts	https://www.genetics.ucla.edu/software/admixture/index.html	Alexander and Novembre (2009)
FINESTRUCTURE/GLOBETROTTER	Clustering and characterizing admixture	Chromosome painting, admixture and clustering	Estimates time since admixture, fast, set of scripts to facilitate analysis	Relies on STRUCTURE and FASTSTRUCTURE assumptions. Requires phased data	http://paintmychromosomes.com/	Hellenthal et al. (2014)
PCADMIX	Clustering and characterizing admixture	Chromosome painting	Fast, uses HMM to smooth out windows and limit noise due to low-confidence ancestry	Requires a priori definition of ancestral populations and phased haplotypes	https://sites.google.com/site/pcadmix/	Brisbin et al. (2012)
MOSAIC	Clustering and characterizing admixture	Chromosome painting, estimating admixture time and proportions	Can handle several source populations. These populations do not have to be good surrogates of populations that actually mixed	Requires phased data, but performs phasing error correction	https://maths.ucd.ie/~mst/MOSAIC/	Salter-Townshend and Myers (2019)
TFA	Clustering and characterizing admixture	Summarizing variance across loci and visualizing interindividual genetic distance	Uses latent factors to correct for drift and to position ancient samples in a PCA-like framework.	NA	https://bcm-uga.github.io/tfa/	François and Jay (2020)
CONSTRUCT	Clustering and characterizing admixture	Perform clustering while taking into account isolation by distance	Aims to extend STRUCTURE while avoiding the over-clustering that is produced by isolation by distance	Slow for large data sets	http://www.genescape.org/construct.html	Bradburd et al. (2017)

(Continues)

TABLE 1 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
DYSTRUCT	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	This method explicitly takes into account the age of samples. Useful when analysing mixtures of modern and ancient samples	Requires a genotype matrix in the eigenstrat format. Primarily tested on human data	https://github.com/tyjo/dystruct	Joseph and Pe'er (2019)Joseph and Pe'er (2019)
BEDASSLE	Differentiation and MCMC model testing	Identifies contribution of environment and geographical distance to population differentiation	Less biased than Mantel tests, provides tools for model testing	Uses population-level data.	https://cran.r-project.org/web/packages/BEDASSLE/index.html	Bradburd et al. (2013)
LOSTRUCT	Differentiation/ diversity	Chromosome painting	Performs local PCA along the genome. Identifies regions showing discrepancies with genome-wide structure, as often happens due to inversions	NA	https://github.com/petrelharp/local_pca	
NPSTAT	Differentiation/ diversity	Extracting summary statistics from pooled data	Explicitly corrects for sampling bias in pooled data. Allows computing tests using an outgroup (MK test, HKA test, Fay and Wu's *H*) and characterizing coding mutations	Mostly limited to summary statistics, but more complete than POPOOLATION	https://github.com/lucaferretti/npstat	Ferretti et al. (2013)
POPOOLATION/ POPOOLATION2/ POPOOLATION TE	Differentiation/ diversity/ recombination	Extracting summary statistics from pooled data	Explicitly corrects for sampling bias in pooled data. Can be used to detect TE polymorphisms.	Mostly limited to a few summary statistics. A pipeline dedicated to TE detection is also available	https://sourceforge.net/p/popoolation/wiki/Main/	Kofler, Orozco-terWengel, et al. (2011), Kofler, Pandey, et al. (2011)
POPGENOME	Differentiation/ diversity/ recombination	Computing summary statistics based on AFS and LD along genomes	Accepts VCF and GFF/GFT files, efficient and fast. Tests for admixture available (ABBA–BABA test). Includes basic coalescence simulations (MS and MSMS)	Mostly limited to summary statistics (but coalescent simulations are possible). No built-in SNP calling module	http://catchenlab.life.illinois.edu/stacks/	Pfeifer et al. (2014)
ANGSD	Differentiation/ diversity/ recombination	Computing summary statistics based on AFS and LD along genomes	Able to process BAM files, built-in procedures for data filtering, admixture analysis. Suited for low-depth data. Includes a suite of methods to estimate relatedness (NGSRELATE).	Mostly limited to summary statistics. Tutorials not always up-to-date.	https://github.com/ANGSD/angsd https://github.com/ANGSD/NgsRelate	Korneliussen et al. (2014), Hanghøj et al. (2019)
VCFTOOLS	Differentiation/ diversity/ recombination	Computing summary statistics based on AFS and LD along genomes	Fast. VCFTOOLS can also be used for SNP filtering	Less summary statistics than POPGENOME	https://vcftools.github.io/man_latest.html	Danecek et al. (2011)

(Continues)

TABLE 1 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ATLAS	Differentiation/ diversity/ recombination	Low-depth sequencing/ancient samples analysis	Particularly suited for analysing ancient samples. Includes sets of tools to call variants, estimate post-mortem damage, inbreeding, genetic diversity. Produces the input file for PSMC (demography from a single diploid genome)	Better used in combination with GATK pipelines. Still in development	https://bitbucket.org/wegmannlab/atlas/wiki/Home	Link et al. (2017)
POPTREE2	Genetic differentiation	Visualizing a matrix of pairwise differentiation statistics as a tree	Can be used for pooled data sets, several statistics can be used	Differentiation measures alone do not necessarily retrieve the actual history of populations	http://www.med.kagawa-u.ac.jp/~genomelb/takezaki/poptree2/index.html	Takezaki et al. (2010)
EEMS	Landscape genomics	Estimating barriers to gene flow in a spatial context	Estimates pairwise relatedness between all samples, and compares it to isolation-by-distance expectations to identify barriers to gene flow and corridors of higher connectivity. Can handle both haploid and diploid data	Requires to convert VCF file into PLINK binary format. Estimates effective migration rates (does not disantangle migration rates and effective population sizes). Setting parameters for the MCMC chain requires some trial-and-error	https://github.com/dipetkov/eems	Petkova et al. (2015)
MAPS	Landscape genomics	Estimating barriers to gene flow in a spatiotemporal context	Expands on EEMS, but takes into account the phase to reconstruct past changes in connectivity. Can disentantle migration rates and effective population sizes (unlike EEMS)	Relies on identity-by-descent tracks, requiring phasing (e.g., using BEAGLE). A pipeline to obtain IBD tracks is available, with a few details here: https://github.com/halasadi/ibd_data_pipeline/issues/1	https://github.com/halasadi/MAPS	Al-Asadi et al. (2019)
TESS3R	Landscape genomics	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	Incorporates geographical information of samples. Can run genome scans of selection based on contrasting ancestral and modern allele frequencies.	Importing data requires using conversion tools found in the LEA suite	https://bcm-uga.github.io/TESS3_encho_sen/	Caye et al. (2016)

(Continues)

TABLE 1 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SPACEMIX	Landscape genomics	Create maps based on genetic distance, and identify admixture	Creates a "geogenetic map" by embedding genetic distances in a map; anomalously high similarity in this map can be indicative of admixture	Can be slow for large data sets	http://www.genescape.org/spacemix.html	Bradburd et al. (2016)
SNPRELATE	Multivariate analysis	Summarizing variance across loci and visualizing interindividual genetic distance	Fast. Can use VCF files as an input	Requires careful interpretation (Jombard et al. 2009)	https://bioconductor.org/packages/release/bioc/html/SNPRelate.html	Zheng et al. (2012)
EIGENSTRAT/ SMARTPCA	Multivariate analysis	Summarizing variance across loci and visualizing interindividual genetic distance	Fast. Can use VCF files as an input	Requires careful interpretation (Jombard et al. 2009)	https://github.com/DReichLab/EIG/tree/master/EIGENSTRAT	Price et al. (2006)
DAPC (ADEGENET)	Multivariate analysis/ clustering	Maximizes divergence between groups identified by PCA	Fast. Less sensitive to HW eqilibrium assumptions. Claims to be more efficient than STRUCTURE	Requires careful interpretation (Jombard et al. 2009)	http://adegenet.r-forge.r-project.org/	Jombart et al. (2010)
SPCA (ADEGENET)	Multivariate analysis/ clustering	Spatially explicit model to assess population structure	Spatially explicit and able to detect cryptic structure. Fast	Does not take into account HW equilibrium or LD	http://adegenet.r-forge.r-project.org/	Jombart et al. (2008)
LAMP	Pedigree, identity by descent/ state	Chromosome painting, relatedness	LAMP also allows for association and pedigree analyses	Identifies local ancestry in windows (source of noise), requires phased data	http://lamp.icsi.berkeley.edu/lamp/	Baran et al. (2012)
PLINK	Pedigree, Identity by descent/ state	Estimating inbreeding and relatedness	Allows studying identity by descent and by state. PLINK is a multipurpose tool, facilitating data analysis within the same software	NA	http://pngu.mgh.harvard.edu/~purcell/plink/	Purcell et al. (2007)
VCFTOOLS	Pedigree, identity by descent/ state	Estimating inbreeding and relatedness	Computes unadjusted Ajk and kinship coefficient	NA	https://vcftools.github.io/man_latest.html	Danecek et al. (2011)
KING	Pedigree, identity by descent/ state	Estimating inbreeding and relatedness, multivariate analysis	Mendelian error checking, testing family structure, highly accurate kinship coefficient, association analysis, population structure inference	Kinship coefficient also computed in VCFTOOLS	http://people.virginia.edu/~wc9c/KING/Download.htm	Manichaikul et al. (2010)

(Continues)

TABLE 1 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
COLONY	Pedigrees	Pedigree inference from SNPs	Robust even with high error rates (e.g., low-depth sequencing). Can handle haplo-diploid systems (e.g., ants). Multithreaded	Can only simulate genotypes with the Windows version	https://www.zsl.org/science/software/colony	Wang (2019)
SEQUOIA	Pedigrees	Pedigree inference from SNPs	Can be applied to large pedigrees (>1,000 individuals). Accomodates unknown birth times	Handles hundreds of SNPs. For whole-genome data, preliminary filtering and LD-pruning may be recommended. Efficient with ~100 SNPs	https://cran.r-project.org/web/packages/sequoia/index.html	Huisman (2017)
LDHAT	Recombination	Estimating variation in recombination rates along a genome	Handles unphased and missing data, underlying model can be used for organisms such as viruses or bacteria	Limited to 300 sequences, specific format (not VCF), model for recombination hotspots based on human data	http://ldhat.sourceforge.net/	McVean et al. (2002)
LDHOT	Recombination	Identifying recombination hotspots	Specifically designed for detecting recombination hotspots	Requires data to be phased, working with LDHat	https://github.com/auton1/LDhot	Myers et al. (2005)
ISMC	Recombination	Recombination from a single diploid genome	No phasing needed. Accepts VCF files as input	Introgression and demographic misspecification may bias results. No detailed tutorial	https://github.com/gvbarroso/iSMC	Barroso et al. (2019)
LDHELMET	Recombination	Estimating variation in recombination rates along a genome	Higher accuracy than LDHat	Requires phased data. Does not handle VCF, only fasta and fastq formats. Requires dividing the genome into short segments to be analysed in parallel	https://sourceforge.net/projects/ldhelmet/	Chan et al. (2012)

TABLE 2 Summary of methods for demographic inference, detecting introgression and comparing demographic scenarios.

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
DSUITE	ABBA–BABA	Identifying past events of admixture between populations	Fast, handles VCF format. Suited for low-depth sequencing (handles uncertainties on genotypes). Provides a set of summary statistics that are useful to investigate complex admixture events	Requires an outgroup sequence. The methods cannot estimate the direction of gene flow.	https://github.com/millanek/Dsuite	Malinsky et al. (2021)
RENT+	Ancestral recombination graphs/ coalescence	Retracing the whole process of recombination and coalescence along a genome	Faster than first version of ARGWEAVER	Requires phased haplotypes. Specific input format. No built-in functions to extract information from genealogies	https://github.com/SajadMirzaei/RentPlus	Mirzaei and Wu (2017)
TREEMIX	Clustering and characterizing admixture	Admixture graph, infers most likely admixture events in a tree	Based on allele frequencies and can be used for pooled data	Requires multiple runs to properly assess the likelihood of each model	https://bitbucket.org/nygcresearch/treemix/src	Pickrell and Pritchard (2012)
G-PHOCS	Coalescence/ Bayesian	Estimating population divergence and migration parameters using a coalescent framework	Bayesian + MCMC, handles ancient samples	Parameters scaled by mutation rate, no admixture	http://compgen.cshl.edu/GPhoCS/	Gronau et al. (2011)
ABLE	Coalescence/ composite likelihood	Model comparison and parameter estimation	Uses both allele frequency spectrum and linkage disequilibrium within blocks of a prespecified size	Relies on MS syntax. Determining the most informative size for blocks requires performing pilot runs	https://github.com/champost/ABLE	Beeravolu et al. (2018)
STAIRWAY2	Coalescence/ composite likelihood	Inferring change in N_e with time	User-friendly. Fast. Suitable for pools or low-depth sequencing	Cannot handle migration or population splits	https://github.com/xiaoming-liu/stairway-plot-v2	Liu and Fu (2020)
FASTSIMCOAL2	Coalescence/ likelihood	Model comparison and parameter estimation	Performs coalescent simulations, parameter estimation and model testing using a fast likelihood method. Can handle arbitrarily complex scenarios for any type of marker	The maximum-likelihood method only uses the allele frequency spectrum. Several runs (20–100) are needed to explore the likelihood space	http://cmpg.unibe.ch/software/fastsimcoal2/	Excoffier et al. (2013)

(Continues)

TABLE 2 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
∂A∂I	Diffusion approximation of the AFS	Model comparison and parameter estimation	Run time does not depend on the number of SNPs included, does not require coalescent simulations, handles arbitrarily complex scenarios. Fast estimation of confidence intervals around parameter estimates (Godambe method). Suitable for pools/low-depth sequencing	Requires some knowledge of Python. Limited to three populations. Several runs (20–100) are needed to explore the likelihood space.	https://bitbucket.org/gutenkunstlab/dadi	Gutenkunst et al. (2009)
MOMENTS	Diffusion approximation of the AFS	Model comparison and parameter estimation	Based on Python, syntax similar to ∂A∂I. Can handle selection. Can use VCF files as input	Requires some knowledge of Python. Limited to five populations. Several runs (20–100) are needed to explore the likelihood space	https://bitbucket.org/simongravel/moments/src/master/	Jouganous et al. (2017)
MOMI2	Diffusion approximation of the AFS	Model comparison and parameter estimation	Can scale to 10 populations. Can simulate and read data in the VCF format. Detailed tutorials available	Does not handle continuous gene flow	https://github.com/popgenmethods/momi2	Kamm et al. (2020)
KIMTREE	Diffusion approximation/ Bayesian	Estimating divergence time between populations and testing for topologies. Estimate divergence times and past effective sex-ratio along branches of a populations tree	Fast and user-friendly. R scripts to obtain plots are available. Suitable for pools/low-depth sequencing. The method is conditional on a prior topology provided by the user. It computes DIC for a given topology, allowing to test for the best one	Strong selection on the sex chromosome can produce male-biased sex-ratios. Times are given in diffusion timescale, and can be converted in demographic times using independent estimates of N_e	http://www1.montpellier.inra.fr/CBGP/software/kimtree/download.html	Clemente et al. (2018)
GADMA	Genetic algorithm	Model comparison and parameter estimation	Based on moments and ∂A∂I. Automates the search for the best set of models explaining a given frequency spectrum	Limited to three populations at the moment	https://github.com/ctlab/GADMA	Noskova et al. (2020)
DORIS	Identity by descent (IBD) tract	Testing various demographic scenario	Uses variation in IBD tracts length to test for various demographic models	IBD must be inferred first with (e.g., BEAGLE). Handles a limited set of demographic scenarios. Modification in the code is required for more complex scenarios	https://github.com/pierpal/DoRIS	Palamara and Pe'er (2013)
UNNAMED.	Identity by state (IBS) tract	Predict observed patterns of IBS along a genome by fitting an appropriate, arbitrary complex demographic model	Allows bootstrapping and estimating confidence over parameter estimates with MS	Specific input format (similar to MSMC or ARGWEAVER)	https://github.com/kelleyharris/Inferring-demography-from-IBS	Harris and Nielsen (2013)

(Continues)

TABLE 2 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ASTRAL-2	Phylogeny	Builds species trees using short nonrecombining sequences	Coalescence-based. Suitable for short loci (e.g., RAD-seq and GBS)	More reliable under high incomplete lineage sorting than SVDQUARTETS and NJST (Chou et al. 2015)	https://github.com/smirarab/ASTRAL	Mirarab and Warnow (2015)
BEAST2	Phylogeny	Network reconstruction and phylogenetic relationships	User-friendly. Can be used to track changes in effective population sizes (Bayesian Skyline Plots). Possible to estimate divergence times	Slow for large data sets. Requires sequence data that can be produced by, for example, STACKS for RAD-seq data	http://beast2.org/	Drummond and Rambaut (2007), Bouckaert et al. (2014)
IQ-TREE 2	Phylogeny	Divergence time estimation and phylogenetic relationships	User-friendly, can be run locally or on a web server, very detailed tutorials. Fast and accurate	Still no tutorial for analysing big data (last checked December 2020)	http://www.iqtree.org/	Minh et al. (2020)
MCMCTREE AND MCMCTREER	Phylogeny	Divergence time estimation and phylogenetic relationships	Included in PAML. An R program is designed to help choose relevant priors and interpret results https://github.com/PuttickMacroevolution/MCMCtreeR	Bayesian, sensitive to priors. Requires a resolved phylogeny and an alignment. Slow for large data sets. Not suited for recent divergence and high gene flow	http://abacus.gene.ucl.ac.uk/software/paml.html	Yang (2007), Puttick (2019)
NJST	Phylogeny	Builds species trees using short nonrecombining sequences	Available in the R package PHYBASE. Estimates populations/species tree from gene trees	Requires splitting part of the genome into nonrecombining "loci"	https://github.com/bomeara/phybase/	Liu and Yu (2010), Liu and Yu (2011)
PHRAPL	Phylogeny	Admixture graph, reticulated evolution	Uses trees in NEWICK format as an input to infer topology, migration rates, divergence times. Similar to ABC in spirit, using tree topology as a summary statistics	Cannot handle more than 16 taxa at a time, and requires subsetting larger data sets	http://www.phrapl.org/	Jackson et al. (2017)
PHYML	Phylogeny	Phylogenetic relationships	Maximum likelihood inference of phylogenetic relationships. An online version is available	Should be used on complex of species or divergent populations with little migration. Can be run on genomic windows to detect introgression (with, e.g., TWISST, DSUITE)	http://www.atgc-montpellier.fr/phyml/binaries.php	Guindon et al. (2010)
RAXML	Phylogeny	Network reconstruction and phylogenetic relationships	Maximum likelihood inference of phylogenetic relationships	Should be used on complex of species or divergent populations with little migration	http://sco.h-its.org/exelixis/web/software/raxml/index.html	Stamatakis (2014)

(Continues)

TABLE 2 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SNAPP	Phylogeny	Phylogenetic relationships	Handles SNP data	Remains slow for medium to large data sets (>1,000 SNPs)	http://beast2.org/snapp/	Bryant et al. (2012)
SNPHYLO	Phylogeny	Network reconstruction and phylogenetic relationships	Complete pipeline from SNP filtering to tree reconstruction	Should be used on complex of species or divergent populations with little migration	http://chibba.pgml.uga.edu/snphylo/	Lee et al. (2014)
SVDQUARTETS	Phylogeny	Phylogenetic relationships	Estimates populations/species tree from gene trees	Remains slow for large data sets. Requires PAUP*	https://www.asc.ohio-state.edu/kubatko.2/software/SVDquartets/	Chifman and Kubatko (2014)
SVDQUEST	Phylogeny	Phylogenetic relationships	Estimates populations/species tree from gene trees	Faster than SVDQUARTETS	https://github.com/pranjalv123/SVDquest	Vachaspati and Warnow (2018)
*BEAST	Phylogeny and species tree inference	Divergence time estimation and phylogenetic relationships	Outputs a species tree instead of concatenated gene tree. Allows for testing consistency between phylogenetic signals at different loci	Slow for large data sets. Requires sequence data. Not suited for situations where gene flow/admixture is important	http://beast2.org/	Heled and Drummond (2010)
SPLITSTREE	Phylogeny/network	Network reconstruction and phylogenetic relationships	User-friendly interface, proposes a variety of methods for network reconstruction	Mostly descriptive	http://www.splitstree.org/	Huson and Bryant (2006)
DICAL2	Sequentially Markovian coalescent	Testing any arbitrary demographic scenario	Works with smaller, more fragmented data sets than PSMC. Handles more complex demographic models than MSMC (including admixture)	Requires phased whole genome data and a model to be defined	https://sourceforge.net/projects/dical2/	Sheehan et al. (2013)
MSMC AND MSMC-IM	Sequentially Markovian coalescent	Inferring change in N_e and migration rates with time between two populations	Allows tracking of population size changes in time without a priori. Allows estimating variation in cross-coalescence rate between two populations	Limited to the study of eight diploid individuals from two populations at once. Requires whole genome phased data and masking regions with insufficient sequencing depth	https://github.com/stschiff/msmc and https://github.com/wangke16/MSMC-IM	Schiffels and Durbin (2014)
SMC++	Sequentially Markovian coalescent	Inferring change in N_e with time and splitting time between two populations	Can analyse hundreds of individuals at a time and does not require phasing	Masking regions as in MSMC. The ancestral allele is assumed to be the reference allele by default. Assumes a clean split for population divergence. Future versions should allow gene flow inference	https://github.com/popgenmethods/smcpp	Terhorst et al., (2016)

(Continues)

TABLE 2 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
TWISST	Topology weighting	Chromosome painting, clustering and branching between populations	Retrieves the most likely coalescence pattern between several taxa along the genome. Can be seen as an extension of the ABBA/BABA test	Needs a priori grouping of individuals into taxa. Requires at least four taxa. Impractical for more than six taxa. Windows size must include enough SNPs to retrieve the correct topology but at the risk that regions with different histories are included	https://github.com/simonhmartin/twisst	Martin and Van Belleghem (2017)
BAYPASS/BAYENV	Variance/covariance matrix	Building a population covariance matrix across population allele frequencies, similar to TREEMIX	Can handle pooled data	Matrices are mostly designed to provide a neutral model for assessing selection, but can be used to infer population structure	http://www1.montpellier.inra.fr/CBGP/software/baypass/; https://bitbucket.org/tguenther/bayenv2_public/src	Günther and Coop (2013), Gautier (2015)

variation in a discriminant analysis, allowing clusters of individuals with highest genetic differentiation to be identified (e.g., using discriminant analysis of principal components [DAPC]; Jombart et al., 2010), and with potential to incorporate temporal sampling (e.g., using DYSTRUCT or TFA; Joseph & Pe'er, 2019; François & Jay, 2020), which is relevant for museum and ancient DNA studies.

2.3 | Model-based inference of population structure

Unlike the previous set of "algorithmic" approaches (see taxonomy proposed in Alexander & Novembre, 2009), model-based approaches model the probability of observing a set of genotypes given a predefined number of clusters (*K*). Some of these methods use a Bayesian (e.g., STRUCTURE; Pritchard et al., 2000, FASTSTRUCTURE; Raj et al., 2014) or a maximum-likelihood framework (e.g., ADMIXTURE; Alexander & Novembre, 2009) and are usually run for a range of *K* values. The optimal number of clusters can then be determined based on likelihood, although examining population structure for a range of *K* can allow substructure to be better identified. The main interest of these approaches is that they provide an estimate of coancestry coefficients, which are the proportions of an individual genome originating from multiple ancestral gene pools. Such information is more difficult to retrieve with approaches such as PCA (though not impossible, see McVean, 2009). There have been criticisms, however, regarding whether ambiguous assignment should actually be interpreted as a signal of admixture, and detailed inference requires thorough model testing and estimating the goodness of fit of a model with admixture (see Lawson et al., 2018).

2.4 | Heterogeneous structure in space: Landscape genomics

Some methods can explicitly use spatial information to inform clustering, allowing improved consideration of the effect of landscape heterogeneity on selection against migrants and drift (e.g., SPACEMIX, TESS3, Table 1). This spatial perspective can be useful to visualize the location and shape of hybrid zones (Guedj & Guillot, 2011). Simple Mantel tests have been popular to routinely investigate relationships between ecological variables and genetic differentiation while accounting for geographical distances. However, these tests are biased by spatial autocorrelation, assume linear dependence between variables, and do not allow testing the relative contribution of each variable (Guillot & Rousset, 2013; Legendre & Fortin, 2010). More recent methods such as EEMS (Petkova et al., 2015) divide the landscape with a dense geographical grid, and identify edges between samples where the effective migration rates are higher or lower than expectations based on isolation-by-distance. The method, however, cannot differentiate between scenarios that lead to the same amount of divergence between samples (e.g., divergence in isolation followed by secondary contact or a geographical barrier with constant, low gene flow). However, a recent expansion of the model, MAPS (Al-Asadi et al., 2019), makes the most of the information provided

TABLE 3 Summary of common methods for identifying loci under positive and balancing selection. The table also lists methods targeting loci associated with environmental features and phenotypes of interest. Note that GENABEL is no longer maintained

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ARGWEAVER/ ARGWEAVER-D	Ancestral recombination graphs/ coalescence	Retracing the whole process of recombination and coalescence along a genome	Provides quantitative estimates for time to the most recent common ancestor (TMRCA) and topologies at each locus. ARGWEAVER-D can estimate introgression. Estimates effective population size. Provides tools to extract summary statistics for the topologies retrieved. Does not require phasing (but slower)	High computing cost. Slower on unphased or low depth data. ARGWEAVER-D is not part of the Anaconda (Python) distribution (http://compgen.cshl.edu/ARGweaver/doc/argweaver-d-manual.html)	Can be installed via conda: conda install -c genomedk argweaver and https://github.com/mjhubisz/argweaver and http://compgen.cshl.edu/ARGweaver/doc/argweaver-d-manual.html	Rasmussen et al. (2014); Hubisz et al. (2020)
GAPIT3	Association	Detecting association with environmental/ phenotypic features	Includes most methods for GWAS studies, including procedures for fast computation, mixed linear models, efficient mixed model association, Bayesian methods such as BLINK, diagnostics such as QQ plots and genotype filtering	May be slow for very large data sets	https://github.com/jiabowang/GAPIT3	Wang and Zhang (2020)
GEMMA	Association	Detecting association with environmental/ phenotypic features	Computationnally efficient for large-scale data sets	Imports data from PLINK format	http://www.xzlab.org/software.html	Zhou and Stephens (2012)
PLINK	Association	Detecting association with environmental/ phenotypic features	Handles a variety of tests for population structure and relatedness	Population structure/kinship need to be assessed in prior association analysis	http://pngu.mgh.harvard.edu/~purcell/plink/	Purcell et al. (2007)
TRINCULO	Association	Detecting association with environmental/ phenotypic features	Specifically designed to handle categorical variables with more than two categories. Performs multinomial logistic regression and provides frequentist and Bayesian frameworks	Requires lapack library in Unix. Allows fine-mapping by testing for corrrelations between adjacent markers	https://sourceforge.net/projects/trinculo/	Jostins and McVean (2016)
SAMBADA	Association/ environmental association	Detecting association with environmental/ phenotypic features	Designed to be fast, underlying models have been kept simple. Allows conversion from PLINK format. Takes into account spatial autocorrelation of individual genotypes. Allows correction for population structure	Does not work with pooled data. Possibly high levels of false positives. Relatedness between samples should be assessed independently. Should be used in combination with LFMM or BAYPASS	http://lasig.epfl.ch/sambada	Stucki et al. (2017)

(Continues)

TABLE 3 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
RELATE	Coalescence with recombination	Reconstruct genome-wide genealogies for hundreds of samples	Provides quantitative estimates for TMRCA and topologies at each locus. Infers past demography (similar to PSMC methods). Infers changes in mutation rates. Performs scans for positive selection over discrete time periods	Requires an outgroup to polarize alleles as ancestral/derived. Requires a recombination map. Does not reconstruct ARG *sensu stricto*, and does not estimate uncertainty of the local genealogies	https://myersgroup.github.io/relate/index.html	Speidel et al. (2019)
DICAL-IBD	Coalescent with recombination/IBD	Predicting IBD tracts from demographic models	High IBD sharing suggests recent positive selection.	Uses DICAL output to obtain expectations based on demographic scenarios	https://sourceforge.net/projects/dical-ibd/	Tataru et al. (2014)
VOLCANOFINDER	Composite likelihood test	Adaptive introgression	Detects a specific signature of increase then drop in diversity near a selected locus brought in a population through introgression	Private input format. Computationally intensive, needs to be run in parallel.	http://degiorgiogroup.fau.edu/vf.html	Setter et al. (2020)
SCCT	Conditional coalescent tree	Detecting positive selection	Designed for detecting recent positive selection. Claims to be more precise at identifying selected sites	The ancestral state of alleles must be obtained through an outgroup	https://github.com/wavefancy/scct	Wang et al. (2014)
LFMM	Environmental association	Detecting adaptation to environmental features	Corrects for population structure using latent factors, faster than BAYENV for large data sets	Only performs association with environment	http://membres-timc.imag.fr/Olivier.Francois/lfmm/software.htm	Frichot et al. (2013)
CLUES	Genealogies at selected loci	Estimate the time at which a beneficial allele rises in frequency	Previous version used ARGWEAVER output, current version uses RELATE. Provides scripts to plot the trajectory of selected alleles	Assumes a panmictic population, neglects the effects of selection at linked sites	https://github.com/35ajstern/clues	Stern et al. (2019)
PALM	Genealogies at selected loci	Estimate the strength and timing of selection on polygenic traits	Uses genealogies estimated from RELATE and results from GWAS to estimate timing and strength of selection for polygenic traits. Should be robust to pleiotropy and residual structure in GWAS	May overestimate selection for older events. Only tested in humans	https://github.com/35ajstern/palm	Stern et al. (2021)
STARTMRCA	Genealogies at selected loci	Estimate the time at which a beneficial allele rises in frequency	Compares genealogies between carriers and noncarriers of an advantageous mutation, assuming a star-genealogy at selected loci. Can handle VCF files	Requires a reference panel of noncarrier haplotypes. Sensitive to local diversity before the sweep, and to migration events during a sweep. More indicated for recent sweeps	https://github.com/jhavsmith/startmrca	Smith, Coop, Stephens, and Novembre (2018)
ANCESTRY _ HMM-S	Identity-by-state tracts	Adaptive introgression	Estimates the selective coefficient of the introgressed loci through a hidden-Markov chain approach	Requires the time and extent of introgression to be defined by the user	https://github.com/jesvedberg/Ancestry_HMM-S/	Svedberg et al. (2021)

(Continues)

TABLE 3 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
H12 TEST	LD	Detecting selection using signatures of high LD	Does not require phased data. Designed for detecting soft sweeps	Coalescent simulations are recommended to evaluate the likelihood of selection	https://github.com/ngarud/SelectionHapStats/	Garud et al. (2015)
LDNA	LD	Detecting selection using signatures of high LD	Can be used to address population structure or detect large inversions or indel polymorphism through LD	The user needs to play with parameters to ensure robustness of SNPs significantly linked	https://github.com/petrikemppainen/LDna	Kemppainen et al. (2015)
REHH	LD	Detecting selection using signatures of high LD	Can compute both XP-EHH and *Rsb*. Handles several input formats	Requires phased data and high density of markers	https://cran.r-project.org/web/packages/rehh/index.html	Gautier and Vitalis (2012)
SCAN FOR EPISTATIC INTERACTION (BASED ON LD)	LD	Polygenic selection/epistatic interactions	Uses genome-wide LD between a candidate locus and the rest of the genomes to identify epistatic interactions. Can test SNP–SNP interaction, or between genomic windows (summarizes genotypes through PCA)	Lack of a detailed tutorial	https://github.com/leaboyrie/LD_corpc1	Boyrie et al. (2020)
SELSCAN	LD	Detecting selection using signatures of high LD	Includes the *nSL* statistics dedicated to soft sweep detection	Does not include utilities to specify the ancestral state of alleles. Requires phased data and high density of markers	https://github.com/szpiech/selscan	Szpiech and Hernandez (2014)
BALLET	Likelihood test for balancing selection	Detecting balancing selection	Designed for detecting ancient balancing selection. Does not require phasing	Requires whole-genome data and recombination map. The ancestral state of alleles must be obtained through an outgroup	http://www.personal.psu.edu/mxd60/ballet.html	DeGiorgio et al. (2014)
BETASCAN2	Local associations of allele frequencies	Detecting balancing selection	Uses correlations in frequencies between genomically proximate SNPs to compute a score. Can incorporate information about ancestral/derived alleles, fixed derived variants and normalizes the statistics depending on the number of sites in a given genomic window. Very detailed tutorial and utilities	Requires estimating the length distribution of ancestral fragments on each side of the selected site. The 95% percentile can be estimated with the formula L = −log(0.05)/(T*rho), where T is the time since selection in generations and rho is the effective recombination rate/generation	https://github.com/ksiewert/BetaScan	Siewert and Voight (2017), Siewert and Voight (2020)
NCD STATISTICS	Local associations of allele frequencies	Detecting balancing selection	Examines the observed and expected frequency spectra of polymorphisms in genomic windows to test for selection. Can incorporate fixed differences with an outgroup (NCD2), but not mandatory (NCD1)	Private input format, requires simulations to calibrate the statistics. Requires to define the expected equilibrium frequency of alleles (usually between 0.3 and 0.5). Low sensitivity below these frequencies	https://github.com/bbitarello/NCD-Statistics	Bitarello et al. (2018)
PCADAPT	Population differentiation	Detecting positive selection and local adaptation	Does not require to define populations. Handles admixed populations and pooled data sets	False positive rate can be high	http://membres-timc.imag.fr/Michael.Blum/PCAdapt.html	Duforet-Frebourg et al. (2016)

(Continues)

TABLE 3 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SELESTIM	Population differentiation	Detecting positive selection and local adaptation	Can estimate the coefficients of selection. Calibration using a simulated data set (can be used in combination with the R function simulate.baypass() in BAYPASS)	Assumes a Wrigth–Fisher island model.	http://www1.montpellier.inra.fr/CBGP/software/selestim/	Vitalis et al. (2014)
BAYENV, BAYPASS	Population differentiation/association	Detecting positive selection and adaptation to environmental features	Less sensitive to population demographic history than previous methods. Handles pooled data sets	Significance thresholds need to be determined from simulated data sets. Calibration with neutral SNPs is recommended. BAYPASS better estimates the kinship matrix	http://www1.montpellier.inra.fr/CBGP/software/baypass/; https://bitbucket.org/tguenther/bayenv2_public/src	Günther and Coop (2013), Gautier (2015)
FLK	Population differentiation/association	Detecting positive selection and local adaptation	Less sensitive to population demographic history than previous methods	Requires an outgroup population	https://qgsp.jouy.inra.fr/index.php?option=com_content&view=article&id=50&Itemid=55	Bonhomme et al. (2010)
LSD	Population differentiation/population-branch test	Detecting positive selection and local adaptation	Compares the level of exclusively shared differences between internal and external branches of a population tree. Allows testing selection occurring on the ancestral branch leading to two populations	Requires several populations to perform the test. May be less sensitive to selection on standing variation	https://bitbucket.org/plibrado/LSD	Librado and Orlando (2018)
POPBAM	Summary statistics	Detecting selection using AFS, differentiation	Extracts summary statistics directly from BAM files	Does not allow for sophisticated filtering and SNP calling	http://popbam.sourceforge.net/	Garrigan (2013)
VCFTOOLS	Summary statistics	Detecting selection using AFS, differentiation	Extracts summary statistics from VCF files. Also allows VCF filtering and conversion	Set of summary statistics not as extensive as POPGENOME	http://vcftools.sourceforge.net/	Danecek et al. (2011)
RAISD	Summary statistics/allele frequency spectrum +LD	Detecting positive selection and local adaptation	Scans the genome for composite signals of selective sweeps summarized by the μ statistics. Corrects for the effects of background selection by estimating a threshold value for the statistics based on simulations with background selection	Uses a single population of interest	https://github.com/alachins/raisd	Alachiotis and Pavlidis (2018)
TASSEL	Summary statistics/association	Detecting association with phenotype	User friendly (Java interface), corrects for relatedness, allows computing summary statistics (LD, diversity)	Requires relatedness to be assessed externally (with, e.g., STRUCTURE)	http://www.maizegenetics.net/tassel	Bradbury et al. (2007)

(Continues)

TABLE 3 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ANGSD	Summary statistics/ association/ population branch test	Detecting selection using AFS, differentiation, association with functional traits	Allows for association using generalized linear models	Descriptive statistics. *p*-values need to be evaluated through coalescent simulations	http://www.popgen.dk/angsd/index.php/ANGSD	Korneliussen et al. (2014)
SWEED	Summary statistics/ composite likelihood test	Designed for whole genome data (or large continuous regions)	Supports Fasta and VCF formats. Estimates selection coefficients	NA	http://pop-gen.eu/wordpress/software/sweed	Degiorgio et al. (2016)
SELECTIONTOOLS	Summary statistics/ LD	Detecting selection using AFS, differentiation and LD statistics	Allows combining several tools in a single pipeline. Includes phasing tools	Set of available summary statistics remains limited (same as VCFTOOLS +Fay and Wu's *H*)	https://github.com/MerrimanLab/selectionTools	Cadzow et al. (2014)
GRROSS	Summary statistics/ allele frequency spectrum	Detecting selection in populations with complex admixture history	Computes the S_B statistics, which detects loci/regions deviating from neutral expectations for each branch leading to current populations. Supports VCF format (converter tools available). Runs with R	Requires that the history of admixture is known, and described with an admixture graph	https://github.com/FerRacimo/GRoSS	Refoyo-Martínez et al. (2019)
PAML/CODEML	Summary statistics/ phylogeny	Distribution of fitness effects/ selection on coding variation	Estimates selection along branches in a phylogeny for genes of interest, contrasting patterns of synonymous and nonsynonymous substitutions. A detailed tutorial is available here: https://link.springer.com/protocol/10.1007%2F978-1-4939-1438-8_4#Sec29	Slow for large data sets. Needs to be parallelized	http://abacus.gene.ucl.ac.uk/software/paml.html	Yang (2007)
POLYDFE2.0	Summary statistics/ phylogeny	Distribution of fitness effects/ selection on coding variation	Can test for invariance of DFEs across data sets (genomic regions within species, or different species). No need for divergence estimates (does not assume that the same DFE is shared between species and outgroup). Very detailed tutorial available here: https://link.springer.com/protocol/10.1007/978-1-0716-0199-0_6	Comparisons require a large number of SNPs for each data set for comparisons to be meaningful	https://github.com/paula-tataru/polyDFE	Tataru and Bataillon (2019)
POPGENOME	Summary statistics/ population branch test	Detecting selection using AFS, differentiation	Fast, embedded in R, allows using annotation files (GFF/GTF format)	Does not perform association, but can be used in combination with GENABEL within R	https://cran.r-project.org/web/packages/PopGenome/index.html	Pfeifer et al. (2014)

TABLE 4 Summary of common methods for simulating genome-wide data and performing simulation-based parameter inference and model comparison (supervised machine learning and Approximate Bayesian Computation)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ABC/ABCRF	ABC	Performs all steps for model-checking and parameter estimation for ABC analyses. ABCRF includes random forest methods (a type of supervised machine-learning)	Informative vignette, allows graphical representation, complete and robust	Does not perform coalescent simulations (but can be used in combination with coala)	https://cran.r-project.org/web/packages/abc/index.html https://cran.r-project.org/web/packages/abcrf/index.html	Csilléry et al. (2012), Raynal et al. (2019)
ABCTOOLBOX	ABC	Complete ABC analysis, from simulations to model checking and parameter estimation	Modular, facilitates the computation of summary statistics	NA	https://bitbucket.org/wegmannlab/abctoolbox/wiki/Home	Wegmann et al. (2010)
DIYABC	ABC	Complete ABC analysis, from simulations to model checking and parameters estimation	User-friendly. Many ways to check goodness-of-fit. Good introduction to ABC models	Does not model continuous gene flow	http://www1.montpellier.inra.fr/CBGP/diyabc/	Cornuet et al. (2008)
POPSIZEABC	ABC	Inferring change in N_e using whole-genome data	Supposed to better assess recent events. Uses a set of summary statistics for the AFS and LD between markers. Handles multiple individuals	Approximate Bayesian approaches do not retrieve the whole information	https://forge-dga.jouy.inra.fr/projects/popsizeabc/	Boistard et al. (2016)
COALA	ABC/coalescent simulations	Combining coalescent simulators within a single framework	Facilitates the building of scenarios and computes summary statistics for simulations. Can be easily combined with the ABC or ABC-RF packages in R	Includes so far MS, MSMS and SCRM	https://cran.r-project.org/web/packages/coala/index.html	Staab and Metzler (2016)
FACSEXCOALESCENT	Coalescent simulations	Simulate demographic scenarios for asexual/facultatively sexual species	Can handle varying levels of sexual reproduction, inbreeding, selfing and cloning	Does not handle population size changes nor selection yet	https://github.com/MattHartfield/FacSexCoalescent https://github.com/MattHartfield/FacSexCoalescentR	Hartfield et al. (2016)
FASTSIMCOAL2	Coalescent simulations	Building any arbitrary scenario using a coalescent framework	Any arbitrary scenario can be implemented. Handles SNP, microsatellites and sequence data	Does not handle selection. Slower than MS with no recombination, much faster with recombination (see manual)	http://cmpg.unibe.ch/software/fastsimcoal2/	Excoffier and Foll (2011)
MS, MSMS, MSABC	Coalescent simulations	Building any arbitrary scenario using a coalescent framework	Any arbitrary scenario can be implemented. Handles SNP, microsatellites and sequence data. MSMS can include selection in the model	Syntax can be difficult to handle for new users compared to, e.g., FASTSIMCOAL2 (but see COALA)	http://www.bio.lmu.de/~pavlidis/home/?Software:msABC	Hudson (2002), Ewing and Hermisson (2010), Pavlidis et al. (2010)

(Continues)

TABLE 4 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
MSMS	Coalescent simulations	Simulate demographic scenarios including selection	Flexible, syntax similar to MS, handles arbitratily complex models. Can be used in an ABC framework to include selection as a parameter to be estimated	Syntax can be difficult to handle for the naive user (but see coala)	http://www.mabs.at/ewing/msms/index.shtml	Ewing and Hermisson (2010)
MSPRIME	Coalescent simulations	Building any arbitrary scenario using a coalescent framework	Faster than MS, Python interface. Syntax is more explicit than MS	Requires some knowledge of Python	https://github.com/tskit-dev/msprime	Kelleher et al. (2016)
SCRM	Coalescent simulations	Fast simulation of chromosome-scale sequences	Syntax similar to MS, handles any arbitrary scenario	Does not handle gene conversion and fixed number of segregating sites (unlike MS)	https://scrm.github.io/	Staab et al. (2015)
SPLATCHE3	Coalescent simulations	Simulating demographic scenarios in their spatial context	Coalescent simulator for genetic data, forward-in-time for demography in space. Spatially explicit.	Simulations can be slow (>1 hr) for large data sets (>100,000 SNPs) over more than 1,000 generations. Does not incorporate selection	http://www.splatche.com/splatche3	Currat et al. (2019)
COALESCENCE	Discoal	Simulate selective sweeps under arbitrary demographic scenarios	Relatively fast for short genomic fragments. Designed to simulate "hard" and "soft" sweeps	Mostly used with DIPLOS/HIC. Other simulators such as MSMS may be more suited for some scenarios	https://github.com/kr-colab/discoal	Kern and Schrider (2016)
QUANTINEMO2	Forward-in-time simulations	Simulating demographic and selection scenarios in their spatial context	Comprehensive simulator. Designed for the study of selection in a spatially explicit context. Simulates quantitative traits, fitness landscapes and underlying genetic variation with migration. Includes both population and individual-based simulations	Scan be slow for large/complex models	https://www2.unil.ch/popgen/softwares/quantinemo/	Currat et al. (2019), Neuenschwander et al. (2019)
SLIM3	Forward-in-time simulations	Simulating genomic sequences with intrinsic and extrinsic factors	One of the most comprehensive simulators. Can simulate genetic data in their spatiotemporal context, the effects of selection at linked sites, coding and noncoding variation, inbreeding and selfing. Supports tree-sequence recording for faster simulations. Large community	Slow for large genomic regions/large populations	https://messerlab.org/slim/	(Haller & Messer, 2019)

(Continues)

TABLE 4 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
DIPLOS/HIC	Supervised machine learning	Detecting selective sweeps	Classifies genomic windows as neutral, selected, or impacted by selection at linked sites. Also distinguishes between selection on standing and *de novo* variation. Uses a set of summary statistics describing frequency spectrum and LD, does not require phasing. Good tutorial explaining the pipeline	Good performance depends on the parameters used to simulate sweeps (window size, selective coefficient, demography). Requires some trial and error for new model species. Interpretation of "soft" and "hard" sweeps remains discussed	https://github.com/kr-colab/diploSHIC	Schrider and Kern (2016), Kern and Schrider (2018)
EVONET	Supervised machine learning	Detecting selective sweeps, balancing selection, and estimate demographic history	Uses deep-learning algorithms to classify genomic regions as selected or neutral, and estimate effective population sizes. Flexible (any number of summary statistics can be provided by the investigator)	Requires summary statistics as an input. Difficult for a naïve user	https://sourceforge.net/projects/evonet/?source=typ_redirect	Sheehan and Song (2016)
FASTEPRR	Supervised machine learning	Estimating effective recombination rates	Uses regression to estimate effective recombination rates from SNP alignments. Can use the VCF format. No clear bias due to phasing errors observed. Can incorporate demographic history (using MS command line)	Requires phased data	https://www.picb.ac.cn/evolgen/softwares/index.html	Gao et al. (2016)
FILET	Supervised machine learning	Detecting introgression	Uses Extra Trees classifiers and dedicated summary statistics to classify genomic windows as being introgressed or not. Identifies the direction of introgression	Targets pulse of introgression rather than continuous gene flow, but can detect the latter. Requires phased data in a fasta format	https://github.com/kr-colab/FILET	Gao et al. (2016)
GENOMATNN	Supervised machine learning	Desstecting adaptive introgression	Uses convolutional neural networks to identify adaptive introgression. Trained using the tree-sequence records obtained from SLIM3. Can handle VCF files and unphased data	Strong computational bottleneck with SLIM simulations	https://github.com/grahamgower/genomatnn	Gower et al. (2020)

(Continues)

TABLE 4 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
IMAGENE	Supervised machine learning	Detecting selective sweeps	Uses convolutional neural networks to classify genomic windows in bins of distinct selection coefficients. Directly uses the image of the alignment, avoiding compression (i.e., using summary statistics)	Can be slow for large data sets	https://github.com/mfumagalli/ImaGene	Torada et al. (2019)
RELERNN	Supervised machine learning	Estimating recombination rates	Uses recurrent neural networks to estimate recombination rates from SNP alignments. Handles unphased and pooled data. Uses MSPRIME (Python implementation of MS) to generate simulations upon which the algorithm is trained. Can incorporate known demographic history provided by the user	Can be computationnally intensive for large effective population sizes. Accuracy on pooled data is modest for low depth of coverage. Absolute estimates of recombination rates depend on the accuracy of the mutation rate used for simulations	https://github.com/kr-colab/ReLERNN	Adrion, Galloway, et al. (2020)
SWIFR	Supervised machine learning	Detecting selective sweeps	Uses averaged one-dependence estimator to classify genomic regions as selected or neutral. Flexible in terms of which summary statistics are used. Can incorporate demographic history	Requires summary statistics as an input. Only distinguishes between selective sweeps and neutral regions	https://github.com/ramachandran-lab/SWIFr/blob/master/README.md	Sugden et al. (2018)

by whole genomes. By using phased haplotypes instead of genotypes, it can jointly estimate migration rates and local effective population sizes across the landscape. Because the length of shared haplotypes is associated with the time since they coalesced, analyses can focus on specific classes of haplotype lengths to reconstruct the past migration landscape at different time periods.

There exist methods that complement the approaches described above, by identifying which combination of geographical and ecological distance limits dispersal. A good example is BEDASSLE, which uses the deviation of allele frequencies at unlinked sites in local populations from the global average, and estimates genetic covariance between all pairs of populations. It then uses a spatial model to estimate the strength of association of covariance with environmental features, assuming a negative relationship between genetic and environmental distances. However, disentangling these effects has proved to be complex. A deeper analysis of genes more strongly impacted by either geography or ecology may be more informative when it comes to the proximate causes of reduced dispersion and differentiation, such as biased dispersal (Bolnick & Otto, 2013; Edelaar & Bolnick, 2012) or selection against migrants (Hendry, 2004). Landscape genomics now extends its focus to adaptive genetic variation, and benefits from new methods targeting signatures of selection (see below).

2.5 | Inferring phylogenetic relationships

Recent advances in molecular phylogenetic methods, and the employment of different types of next-generation sequencing (NGS) data is well beyond the scope of this review (see, e.g., Moriarty Lemmon & Lemmon 2013; Cruaud et al. 2014; Wen et al. 2015). In this respect, both maximum likelihood and Bayesian approaches have become popular to investigate evolutionary relationships between individuals from different populations, even when divergence is very recent (e.g., Wagner et al., 2013). These methods are implemented in software such as RAXML (Stamatakis, 2014) and BEAST2 (Drummond & Rambaut, 2007). Ultimately, all molecular phylogenies reconstruct the genealogy of the genes with which they have been constructed. A problem when applying phylogenetic methods, especially in the context of recent divergence, is the assumption that gene trees are representative of lineage history. This assumption is likely to be violated at the population level, since the influences of gene flow and incomplete lineage sorting are strong at this scale, and may cause gene trees to deviate from population history. Fortunately, many recent coalescent-based methods in phylogenomics explicitly model gene trees to fit inside a speciation framework.

When using genome-wide data at the population level, methods specifically dedicated to reconstructing multiple species coalescent (MSCs) models such as *BEAST (STAR-BEAST) may be preferred over concatenation (Edwards et al., 2016), since they accommodate fluctuations in genealogical history across the genome, allowing discordance between species trees and individual gene trees to be identified. However, in the presence of strong gene flow, MSC models can underestimate divergence times and overestimate effective population sizes (Leaché et al., 2014), because they attempt to explained the observed diversity with a strict isolation model. This issue is partially tackled by methods such as PHRAPL (Jackson et al., 2017), which estimates the likelihood of complex histories by examining genealogies at multiple genes and comparing them with coalescent simulations. Such integration is particularly needed for species and populations that are in the "grey zone of speciation" (Roux et al., 2016).

While useful to infer topologies, caution is advised when using branches lengths obtained from SNP-only data sets, foe example to calculate divergence times between different groups or species (Leaché et al., 2015). For this purpose, it might be more straightforward and reliable to extract from the data both variant and invariant sites at several genes (e.g., coding or conserved sequences), and analyse the whole sequences in software like BEAST2. Such analyses can also be performed in two steps: first, estimate the phylogenetic relationships between samples, then apply a molecular clock model to obtain times since divergence. This style of approach is implemented in the Bayesian method MCMCTree in the PAML package (Yang, 2007).

2.6 | Inferring demographic history with likelihood methods based on the allele frequency spectrum

The allele frequency spectrum (AFS) is the distribution of allele frequencies at polymorphic loci in one or several populations (called in that case joint or multipopulation spectrum). Different patterns of gene flow and demographic events all shape the AFS in specific ways (e.g., alleles are likely to occur at more similar frequencies if divergence is recent or if populations are highly connected). Several methods use the AFS to infer the demographic events explaining current genomic diversity. Two of the most popular methods (∂A∂I and FASTSIMCOAL2; Gutenkunst et al., 2009; Excoffier et al., 2013) fit population genetics model specified by the user to the observed spectrum using a maximum-likelihood approach. The AFS expected under a given scenario is obtained through simulation, either using a diffusion approach (∂A∂I), or coalescent simulations (FASTSIMCOAL2). These approaches quickly estimate parameters using composite likelihoods, but do not explicitly take into account correlations induced by linkage disequilibrium (LD) between physically linked markers (but see ABLE; Beeravolu et al., 2018). This might limit power to detect recent demographic events (e.g., migration, Jenkins et al., 2012). Including SNPs that are physically close together should not strongly bias parameter estimation. However, such an approach prevents direct comparisons of likelihoods from different models. Therefore, physically independent SNPs should be used to consider composite likelihoods as quasi likelihoods for model comparison (Excoffier et al., 2013). Using allele frequencies estimated from pooled data sets is also feasible, as illustrated by a recent study on hybridization in *Populus* species where AFS was estimated from pooled whole genome resequencing data (Christe et al., 2016). The same applies for low-depth-sequencing data, with software such as ANGSD or ATLAS that are able to extract the most likely AFS and other relevant summary statistics. Such approaches are particularly promising to analyse whole-genome data from species with large genomes, ancient

DNA samples, or when sequencing costs would otherwise be too prohibitive (Box 1).

2.7 | Inferring past demography with hidden-Markov model and sequentially Markovian coalescent methods

Methods have been developed to infer variation in population sizes with time using the whole genome of one or several diploid individuals. Briefly, these methods model successive genealogies along the genome sequence as a Markov process: the genealogy at one locus only depends on the genealogy at the previous locus. Changes in the topology

BOX 1 Analyzing pooled sequences, ancient DNA samples and low-depth data

Despite decreasing costs, whole-genome sequencing remains quite expensive, especially for species with large genomes. Classical experimental designs usually target a sequencing depth of about 20–40×. However, several options exist in situations in which this depth is not achievable. Pooled sequencing (Futschik & Schlötterer, 2010), in which individuals from the same sampling site/population are sequenced as a single library, can be an option to reduce costs. Summary statistics along the genome and allele frequency spectra can then be extracted for each population (e.g., using methods such as POPOOLATION; Kofler, OrozcoterWengel, et al., 2011; Kofler et al., 2011). Since individual information is not available, variation in LD across individuals cannot be fully exploited, but methods such as ∂A∂I can still be used to test complex demographic scenarios (Gutenkunst et al., 2009). Shallow shotgun sequencing (1–5× per individual) is another approach that gives access to individual information for a similar cost (Buerkle & Gompert, 2013), but might prevent using methods requiring accurate phasing and unbiased individual genotypes. Nevertheless, recent methods such as those implemented in the packages ANGSD (Korneliussen et al., 2014) or ATLAS (Link et al., 2017) are promising. For example, ATLAS includes an approach to reconstruct past demographic histories by applying the pairwise sequentially Markovian coalescent (PSMC) to low-depth ancient DNA samples. ANGSD comes with several methods that estimate relatedness in low-depth samples (NGSRELATEV2; Hanghøj et al., 2019), and can estimate allele frequency spectra that can be used for demographic and selection inference. Among the most powerful methods available, recent versions of ARG-WEAVER are promising since they can take into account genotype quality when reconstructing genealogies along the genome, and can therefore be applied to "low-quality" samples. One of the main drawbacks is that such analyses take time, making ARG-WEAVER more suited to investigating genealogies in a limited set of genomic regions of interest.

BOX 2 Efficiently simulate whole-genome data

Simulations of whole-genome data are poised to become a standard tool for researchers, and recent initiatives such as STDPOPSIM, an open library of population genetics simulation models for multiple species, might help design reproducible simulations (Adrion, Cole, et al., 2020). More than 145 genetic simulators are currently available, but not all can handle genome-sized data (see https://surveillance.cancer.gov/genetic-simulation-resources/). Simulated data can be used to define significance thresholds for summary statistics when trying to scan the genome for regions under selection. Simulations are also at the core of simulation-based algorithms such as ABC or supervised machine-learning. By comparing simulations with observed data, these methods can identify the processes that underlie diversity in any given genomic region.

There are two main categories of simulators, those based on coalescent ("backward in time" simulators), and forward-in-time simulators. The MS software, with its extensions (such as MSMS, Table 4), is one of the most versatile available. Coalescent simulators are generally fast, and can simulate large genomic regions of hundreds of kilobases efficiently. An important limitation of these simulators is that most only simulate SNP data, and were not intended to simulate other categories such as transposable elements. Moreover, despite the abundance of species that practice self-fertilization and asexual reproduction, only FACSEXCOALESCENT is able to model coalescence in facultatively sexual species.

Forward-in-time simulators such as SLIM3 (Haller & Messer, 2019) bypass the aforementioned limitations. They can accomodate an impressive diversity of scenarios and model genomic data in their spatiotemporal context, incorporate purifying and positive selection, and even go beyond Wright–Fisher approximations, for example by allowing overlapping generations. This comes at the cost of speed: long genome sequences can take days or weeks to be simulated. Simulation time can be reduced by scaling mutation rates, selective coefficients, times of demographic events and population sizes, but can still remain relatively long, requiring massive parallelization. However, SLIM3 now supports tree-sequence recording, which greatly reduces simulation time. Instead of explicitly simulating neutral mutations, the method outputs genealogies upon which mutations can be added at a later stage using the coalescent simulator msprime, implemented in Python (Kelleher et al., 2016).

are due to recombination events reconnecting branches in the tree. The whole genealogy is usually not estimated, however, which results in drastic gains in speed. Such methods have the advantage of requiring only a small number of individuals (1–10), no a priori knowledge of population history, and permitting time-varying gene flow to be incorporated (see MSMC-IM). One general drawback, however, is that they are limited to rather simple scenarios, and do not handle more than two populations as yet (but see DICAL2, Table 2). While powerful, they are sensitive to confounding factors such as population structure (OrozcoterWengel, 2016) that lead to false signatures of expansion or bottlenecks. These methods also do not allow extremely recent demographic events to be investigated, since the coalescence of two alleles from a single individual in the recent past (a few tens to hundreds of generations) is infrequent. Moreover, most of these methods require the data to be phased (but see SMC++; Terhorst et al., 2016), for example with FASTPHASE (Scheet & Stephens, 2006) or BEAGLE (Browning & Browning, 2011). In addition, phasing errors can lead to strong biases in parameter estimates for recent times (Terhorst et al., 2016). An extension of these methods takes into account population structure and aims to identify the number of islands contributing to a single genome, assuming it is sampled from a Wright *n*-island metapopulation (Mazet et al., 2015; Rodríguez et al., 2018). Such developments should improve the amount of information retrieved from only a few genomes.

Methods based on tracts of identity-by-descent (IBD, Palamara & Pe'er, 2013) constitute an interesting alternative for more complex model testing when whole genomes are available in large number. Such methods allow recent demographic events to be inferred with relative precision. They are used to predict the length of haplotypes shared by two individuals that are inherited from a common ancestor without recombination. However, IBD detection requires large cohorts and accurate phasing, and therefore application of these methods has been largely restricted to human populations so far (Browning & Browning, 2011; Palamara & Pe'er, 2013). Another approach has used tracts of identity-by-state (IBS) to perform demographic inference over a range of timescales (Harris & Nielsen, 2013). IBS tracts are directly observable since they are simply the intervals between pairwise differences in an alignment of sequences and do not require any assumption about coancestry to be defined. The method predicts the length distribution of IBS tracts for pairs of haplotypes under a range of demographic parameters. These predicted spectra are then compared to empirical data under a likelihood framework, as with methods based on the AFS.

3 | DETECTING LOCAL SIGNATURES OF EVOLUTIONARY PROCESSES ALONG THE GENOME

3.1 | Selection, introgression and their impact on sequence variation

While demographic forces such as drift and migration will affect the whole genome, selection in the presence of recombination is expected to be specific to particular portions of the genome, and therefore yield discrepancies with genome-wide polymorphism (Lewontin & Krakauer, 1973; but see section 3.9). Both positive and negative selection have long-distance effects on sites that are adjacent to those under selection, an effect often put under the umbrella term of "linked selection" (Cruickshank & Hahn, 2014; Ravinet et al., 2017). These effects are stronger in regions of low recombination, and may explain the correlations observed between nucleotide diversity, divergence metrics and recombination rates that are observed across many clades (Charlesworth et al., 1997; Cruickshank & Hahn, 2014). Using whole-genome resequencing data, it is possible to estimate the effective recombination rate along the genome (see the Recombination class of methods in Table 1). Such estimates are particularly useful in the absence of pre-existing genetic maps to assess how recombination and linked selection may bias estimates of diversity statistics or scans for selection. It also provides a way to determine a suitable window size to compute "independent" statistics along genomic windows.

In the sections that follow, we describe different methods aiming at identifying regions under selection by contrasting local patterns of diversity and divergence with genome-wide patterns. We begin with approaches focusing on single populations, and then summarize those focused on multiple populations.

3.2 | Quantifying positive and purifying selection on coding regions

The ratio between the number of nonsynonymous and synonymous mutations (also called dn/ds, K_A/K_S or ω) is often used to detect whether a specific gene is undergoing negative ($\omega < 1$) or positive ($\omega > 1$) selection. It is also useful to estimate the effects of demography on mutational load and ultimately extinction risk. An excess of nonsynonymous mutations can signal positive or balancing selection, or a relaxation of selective constraints on a given gene. More sophisticated tests, such as the MK test (McDonald & Kreitman, 1991), can use population data and compare the proportion of nonsynonymous and synonymous variation segregating within and between species. However, these approaches require an annotated genome and an outgroup to detect synonymous and nonsynonymous variants. Annotation of mutations can be performed with a dedicated software (e.g., SNPDAT; Doran & Creevey, 2013). The main issue with estimating ω from a single pair of species is that its value rarely exceeds 1, even in the case of positive selection, due to long-term effects of purifying selection. A more powerful approach lies in the comparison of nonsynonymous and synonymous mutations between orthologues from different species, and can be performed in packages such as PAML and CODEML (Yang, 2007). These methods are model-based and estimate the likelihood of different models of sequence evolution that can include selection at a specific codon, gene or branches along the phylogeny while accommodating variation in substitution rates, base composition or transition/transversion ratios.

The comparison of the AFS of synonymous (assumed neutral) and nonsynonymous polymorphisms is also useful to infer the distribution

of fitness effects (DFE), an informative measure in quantitative genetics regarding the adaptive potential of populations (Eyre-Walker & Keightley, 2007). This allows estimation of a fundamental parameter for coding sequences, α, the proportion of variants fixed by adaptive evolution. Several probability distributions have been proposed to fit the DFE (usually deriving from the Γ distribution, see Eyre-Walker & Keightley, 2007). Methods aiming at estimating the DFE derive the expected AFS for synonymous and nonsynonymous mutations under different probability distributions, and treat the effects of unknown demography and polarization errors as nuisance parameters shared by both categories of polymorphisms. The DFE is then obtained through comparison of the maximum-likelihood of different models. A well-developed set of models and distributions can be compared and tested in POLYDFE (Tataru & Bataillon, 2019). Note that a very detailed tutorial with scripts is available in Tataru and Bataillon (2020) for the latter method.

3.3 | Detecting selective sweeps (recent positive selection)

Selective sweeps reduce diversity in genomic regions flanking the selected site(s). This leads to local deviations in the shape of the AFS that can be captured by several summary statistics computed over genomic windows, such as π, the nucleotide diversity (Nei & Li, 1979), Tajima's *D* (Tajima, 1989), and Fay and Wu's *H* (Fay & Wu, 2000). Using a combination of these statistics allows targets of selection to be identified with greater precision, and minimizes the confounding effects of demography. However, defining a threshold beyond which the values of a set of statistics supports selection is nontrivial. Recent developments in machine learning and Approximate Bayesian Computation (ABC) may assist in this regard (see section 3.9 below). Directly contrasting genome-wide with local AFS is another option that does not require combining results from multiple summary statistics. This approach has been used to develop composite tests, such as the composite likelihood ratio (CLR) test (Degiorgio et al., 2016; Stamatakis et al., 2013) that aims to detect recent selective sweeps by maximizing the likelihood of a model with selection in a genomic window, and comparing it to a model built on SNPs sampled from the genomic background.

In regions near to a selected allele, it is expected that LD is increased and diversity is decreased, especially after recent positive selection. A class of methods are aimed at targeting those regions that display an excess of long homozygous haplotypes, such as the extended haplotype homozygosity (EHH) test (Sabeti et al., 2002). It is also possible to compare haplotype extension across populations, with the Cross Population Extended Haplotype Homozygosity test (*XP-EHH*; Sabeti et al., 2007) or *Rsb* (the standardized ratio of EHH at a given SNP site; Tang et al., 2007). These methods require data to be phased in order to reconstruct haplotypes, which can make them susceptible to switch-errors. Nevertheless, methods based on LD may be more sensitive to selection on standing variation or on multiple alleles that leave a more subtle signature (so called soft sweeps). Statistics dedicated to the detection of soft sweeps include the *nSL* statistics (Ferrer-Admetlla et al., 2014) in SELSCAN or the *H2*/*H1* statistics (Garud et al., 2015). These statistics usually examine the distribution of the length of homozygous haplotypes (in number of SNPs), comparing ancestral and derived haplotypes (*nSL*), or the second most frequent derived haplotype with the most frequent one (H2/H1). Further studies are still needed to understand to what extent hard and soft sweeps can actually be distinguished (Schrider et al., 2015), as well as their relative importance (Jensen, 2014; Messer & Petrov, 2013). Even hard selective sweeps can be challenging to detect with LD-based statistics especially under unstable demography and weak selection (Jensen, 2014). It is advisable to combine several approaches to improve confidence when pinpointing candidate genes for selection. Methods based on LD alone can sometimes miss the actual variants under selection due to the impact of recombination on local polymorphism that can mimic soft or ongoing hard sweeps (Schrider et al., 2015).

3.4 | Detecting long-term balancing selection

Unlike directional selection, balancing selection can lead to the maintenance of polymorphism at selected loci over long periods of time. This type of selection is extremely relevant for evolutionary biologists (Sellis et al., 2011), since it is at the core of strong coevolutionary dynamics such as host–parasite interactions (Ebert & Fields, 2020). Despite its importance, balancing selection has often been overlooked. This is mostly due to its narrow effects, particularly in the case of long-term balancing selection where recombination erodes association between loci under selection and neutral neighbouring loci. Nevertheless, the emergence of whole-genome resequencing data has facilitated the investigation of these narrow signals. Several recent methods and summary statistics (see Table 3, "Detecting balancing selection") have been specifically developed to detect this type of selection (Bitarello et al., 2018; DeGiorgio et al., 2014; Rasmussen et al., 2014; Siewert & Voight, 2017). These methods are all based on the AFS to some extent. Some methods examine the strength of correlations between allele frequencies at adjacent SNPs (Siewert & Voight, 2020), while others use a CLR approach, contrasting the likelihood of a model with selection in candidate windows with the likelihood computed for all sites in the genome (DeGiorgio et al., 2014). On the other hand, recent balancing selection may look similar to an incomplete selective sweep (Charlesworth, 2006), and be detected by methods aimed at detecting long haplotypes and low diversity.

3.5 | Detecting introgressed genomic regions

Understanding the origin of genomic regions under selection highlights the evolutionary history of adaptive alleles (e.g., Abi-Rached et al., 2011) and contributes to our understanding of the origin and maintenance of reproductive isolation. Studies focusing on hybrid

zones and introgression have provided inspiring examples of adaptive introgression (Hedrick, 2013), as demonstrated by recent work on localized introgression and inversions at a colour locus in *Heliconius* butterflies (The Heliconius Genome Consortium et al., 2012) or adaptive introgression of anticoagulant resistance alleles in mice (Song et al., 2011).

Summary statistics can be useful to obtain a first set of candidates for introgression and selection. One may, for example, plot the distribution of a differentiation measure such as F_{ST} (Weir & Cockerham, 1984) between populations, estimates of effective recombination rates and nucleotide diversity along the genome. Such an approach has been used in Darwin's finches, which uncovered genomic islands of divergence with low recombination rates resisting gene flow (Han et al., 2017). Other approaches, such as chromosome painting (Table 1), extend PCA and ADMIXTURE-like methods by incorporating information about the relative order of markers in the genome, allowing identification of regions for which ancestry differs from the rest of the genome. Recent developments also provide a fast and efficient way to test complex patterns of heterogeneous introgression along the genome (see, for example, Dsuite in Malinsky et al., 2021). These methods build upon the well-known ABBA–BABA statistics (Durand et al., 2011) and provide a variety of estimators that can be estimated for the whole genome or along genomic windows. They require that phylogenetic relationships between populations and species are known (see Section 2.5 above). Other methods allow the user to test the relative contribution of different topologies expected with and without gene flow (e.g., the topology weighting method implemented in TWISST; Martin & Van Belleghem, 2017).

3.6 | Identifying highly differentiated loci and associations between allele frequencies and environmental features

When an allele is under positive selection in a population, its frequency tends to rise to fixation, unless gene flow from other populations or strong drift prevents this from happening (Charlesworth et al., 1997). It is therefore possible to contrast patterns of differentiation between populations adapted to their local environment to detect loci under divergent selection (e.g., displaying a high F_{ST}). However, it is essential to control for population structure, as it may strongly affect the distribution of differentiation measures and produce high rates of false positives. Modern methods based on this principle (Table 3) correct for relatedness across populations, and can test association between allele frequencies and environmental features (see the extensive review by François et al., 2015). Methods such as BAYPASS (Gautier, 2015) are convenient in both describing population structure and providing preliminary insights into the proportion of loci that do not follow neutral expectations. When this proportion is not too high, outliers can be removed to avoid bias (Schrider et al., 2016) and the remaining loci can be used for demographic inference and model-testing. These estimated parameters can then be used to simulate sequences or independent SNPs and generate a neutral expectation. Loci that are more likely to be neutral can be used to further calibrate tests for selection (Lotterhos & Whitlock, 2014).

Detecting an association between environment and allele frequencies does not necessarily imply a role for local adaptation. For example, in the case of secondary contact, intrinsic genetic incompatibilities can lead to the emergence of tension zones that may shift until they reach an environmental barrier where they can be trapped (Bierne et al., 2011). In addition, the effects of selection at linked sites might generate false positives. The sampling strategy must take into account the particular historical and demographic features of the species investigated to gain power (Nielsen et al., 2007). The sequencing strategy must also be carefully considered to control for spatial autocorrelation of genotypes due to IBD and shared demographic history. For example, localized range expansion may produce a spurious association between environmental features and allele frequencies due to repeated founder effects and allele surfing (Excoffier & Ray, 2008). Including samples from populations not affected by such an expansion may avoid reaching biased conclusions by examining signatures of association at a broader scale.

3.7 | Identifying significant genotype–phenotype associations and and epistatic interactions between variants

The methods described above focus on allele frequencies at the population scale, but do not test association with traits that vary between individuals within populations (e.g., resistance to a pathogen, symbiotic association, individual size or flowering time). These traits can be under directional selection, but also under stabilizing selection across multiple populations (e.g., height). For this task, methods performing genome-wide association analysis (GWAS) are better suited. Detailed reviews on these methods and their biases are available (Liu & Yan, 2019; Tam et al., 2019; Wang et al., 2019). Initiatives such as GAPIT3 (Wang & Zhang, 2020) provide most of the currently available tools for GWAS in a single framework. The recent development of multivariate methods also allows loci putatively under selection to be identified in admixed or continuous populations without requiring information about individual phenotype (Duforet-Frebourg et al., 2016).

Uncovering the genetic basis of complex, polygenic traits remains challenging, even in model species (Pritchard & Di Rienzo, 2010; Rockman, 2012). It may be unavoidable as a first step to focus only on traits that are under relatively simple genetic determinism. This can, however, lead to the overrepresentation of loci of major phenotypic effect, a fact that should be acknowledged when discussing the impact of selection on genome variation. The fact that loci of major effect are the easiest to target does not imply that they are necessarily the main substrate of selection (Rockman, 2012). Association methods may help to target variants undergoing soft sweeps, weak selection or those involved in polygenic control of traits (Pritchard et al., 2010).

In such cases, signatures of selection may be subtle and sometimes difficult to retrieve from allele frequency data. Nevertheless, recent tools may have a higher sensitivity to polygenic selection (see section below on Ancestral recombination graphs), and a recent method uses genome-wide patterns of LD between a candidate gene (the "bait") and loci along the genome to detect candidate genes that may be involved in epistatic interactions (Boyrie et al., 2020). Such developments hold great promise in addressing the issue of nonadditive genetic effects.

3.8 | Inferring differences in history along the genome with ancestral recombination graphs

Ancestral recombination graphs (ARGs) are a generalization of the coalescent and describe the sequence of genealogies along a sample of recombining sequence. Genealogies are estimated for each nonrecombining block, and recombination between adjacent blocks is described by breaking the branch leading to the recombining haplotype and allowing it to recoalesce to the rest of the tree. This succession of local trees joined by recombination events provides a full description of the genealogical history of the data and is therefore a promising approach to characterize all modes of selection, introgression and demography while taking into account variation in recombination and mutation rate. Methods that are able to estimate or approximate these genealogies have long been computationally intensive and unapplicable to whole genomes. Fortunately, recent improvements make their application to whole genomes feasible (Table 4). A good example is ARGWEAVER (Rasmussen et al., 2014), which has allowed candidate genes for long-term balancing selection to be recovered from human data, and has recently been used in combination with machine learning methods to study speciation in capuchino seedeater birds (Hejase et al., 2020) and introgression in humans (Kuhlwilm et al., 2016). However, ARGWEAVER remains slow when analysing more that 50 diploid genomes, and is even slower for low-depth or unphased data. Another promising method is implemented in the RELATE software (Speidel et al., 2019). RELATE is able to estimate genome-wide genealogies, and uses this information to reconstruct past demographic trajectories, changes in mutation rate over time, identify loci under positive selection, and estimate when selection acted on candidate mutations. RELATE can handle thousands of genomes in a manageable amount of time, but requires an outgroup sequence to polarize derived alleles, and a recombination map. RELATE comes with two add-on methods, CLUES and PALM, which can use the RELATE output to estimate the strength of selection for single loci and a set of candidate loci identified by GWAS respectively. The latter method therefore provides a way to quantify polygenic selection and adaptive introgression at multiple loci, and constitutes a major advance in the field of population genomics.

3.9 | Jointly inferring demographic history and selection using ABC and supervised machine learning

It has recently become clear that the interactions between mutation and recombination rate, introgression, demography, selective sweeps and background selection have to be integrated into analyses of genetic variation (Andrew et al., 2013; Li et al., 2012; Ravinet et al., 2017). Simulation-based approaches hold great promise for incorporating this complexity. Two nonexclusive methodologies are promising: ABC and supervised machine learning approaches. Both rely on simulated data simulated under a range of parameters which are used to identify the combination of parameter values that are most likely to have generated the observed data (see Box 2 for a discussion on simulators). Both methodologies are flexible and powerful, and have become increasingly popular in population genomics (Csilléry et al., 2010; Schrider & Kern, 2018). These methods can accomodate any type of marker and arbitrarily complex models. By measuring the distance between carefully chosen summary statistics describing each simulation with those from the observed data set, it is possible to infer which combination of selective and demographic parameters best explains the data. These methods enable the rate of false positives to be estimated, for example by estimating how many times simulations of neutral sequences are classified as selected.

However, using summary statistics leads to the loss of potentially useful information (Robert et al., 2011). Machine learning presents three advantages with respect to this problem. First, whereas ABC is prone to show lower performance as the number of statistics summarizing simulations increases, machine learning tends to display better performance. Second, ABC usually excludes many simulations by retaining only those closest to the observed data, whereas machine learning makes use of all simulations to form a model. Third, machine learning algorithms can identify the set of summary statistics that is the most useful for inference, and some neural networks algorithms may also be trained directly on images of aligned sequences (e.g., IMAGENE, Table 4). An example of application of ABC and deep learning is provided in a study of African populations of *Drosophila melanogaster* (Sheehan & Song, 2016), in which the authors use these approaches to identify genomic regions under balancing and positive selection, and infer past demography.

The flexibility of ABC and supervised machine learning means that researchers can adapt the pipeline to their need, for example by using the simulator that generate simulations that are as close to the specifics of their study system as possible. However, this flexibility also means a steeper training curve for researchers learning these methods, due to the lack of a clearly unified analytical pipeline. Moreover, recent discussions on the power of machine learning pipelines to detect selective sweeps have highlighted that a careful consideration of demographic null models and methodological limitations is imperative (Harris et al., 2018).

4 | CONCLUSION

As illustrated by sections 3.8 and 3.9, the field of population genomics is now moving towards both better integrating the demographic framework in inferences of selection, and, conversely, taking into account selection when reconstructing demographic history. The joint

inference of loci under selection and quantification of demographic dynamics is of crucial importance in fields such as landscape genomics or the study of ongoing speciation. It might provide insights into the role of selection, recombination and gene flow in promoting or impairing local adaptation to new habitats. The growing availability of genome-wide data for nonmodel species is therefore promising, but requires caution and high stringency in our interpretation of observed patterns. With the decreasing cost of sequencing, it has been suggested that NGS will rapidly broaden our perspective on complex evolutionary processes, from biogeography (Lexer et al., 2013) to the genetic basis of traits (Hohenlohe, 2014) and the maintenance of polymorphisms (Hedrick, 2006). However, the study of DNA sequence variation is already challenging in its own right, and prone to storytelling (Pavlidis et al., 2012). In order to be informative about processes such as selection and demography, population genomics should ultimately be combined with other disciplines such as ecology and functional analyses (Habel et al., 2015). This can be achieved, for example, by assessing the function of selected genes, the consistency of demographic history with information retrieved from the fossil record or geological history, and the broader integration of population genomics with other fields and methods whenever possible, such as niche modelling, common garden experiments or the study of macro-evolutionary patterns of selection and diversification.

ACKNOWLEDGEMENTS

The University of Basel, New York University Abu Dhabi and the Univeristy of Portsmouth have supported Y.B.'s research in this area. We thank Stephane Boissinot, Joris Bertrand, Muriel Gros-Balthazard, Khaled Hazzouri, Anne Roulin and three anonymous reviewers for their insightful comments on previous versions of the manuscript. We also thank Gabriel Renaud and Peter Ralph for suggesting additional methods.

AUTHOR CONTRIBUTION

YB wrote the first draft of the manuscript and maintained the website. YB and BW agreed on the structure and revised the initial draft.

DATA AVAILABILITY STATEMENT

No data to provide. The tables shown in this article are available at www.methodspopgen.com

ORCID

Yann X. C. Bourgeois https://orcid.org/0000-0002-1809-387X
Ben H. Warren https://orcid.org/0000-0002-0758-7612

REFERENCES

Abi-Rached, L., Jobin, M. J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F. A., Kimani, J., Carrington, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S. G. E., Maiers, M., Guethlein, L. A., Tavoularis, S., … Parham, P. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, *334*, 89–95. https://doi.org/10.1126/science.1209202

Abzhanov, A., Extavour, C. G., Groover, A., Hodges, S. A., Hoekstra, H. E., Kramer, E. M., & Monteiro, A. (2008). Are we there yet? Tracking the development of new model systems. *Trends in Genetics*, *24*(7), 353–360. https://doi.org/10.1016/j.tig.2008.04.002

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., & Kern, A. D. (2020). A community-maintained standard library of population genetic models. *eLife*, *9*, e54967. https://doi.org/10.7554/eLife.54967

Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the landscape of recombination using deep learning. *Molecular Biology and Evolution*, *37*(6), 1790–1808. https://doi.org/10.1093/molbev/msaa038

Alachiotis, N., & Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, *1*(79), https://doi.org/10.1038/s42003-018-0085-8

Al-Asadi, H., Petkova, D., Stephens, M., & Novembre, J. (2019). Estimating recent migration and population-size surfaces. *PLoS Genetics*, *15*(1), 1–21. https://doi.org/10.1371/journal.pgen.1007908

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.vidual

Andrew, R. L., Bernatchez, L., Bonin, A., Buerkle, C. A., Carstens, B. C., Emerson, B. C., Garant, D., Giraud, T., Kane, N. C., Rogers, S. M., Slate, J., Smith, H., Sork, V. L., Stone, G. N., Vines, T. H., Waits, L., Widmer, A., & Rieseberg, L. H. (2013). A road map for molecular ecology. *Molecular Ecology*, *22*(10), 2605–2626. https://doi.org/10.1111/mec.12319

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., & Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, *495*(7441), 360–364. https://doi.org/10.1038/nature11837

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., & Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, *28*(10), 1359–1367. https://doi.org/10.1093/bioinformatics/bts144

Barroso, G. V., Puzović, N., & Dutheil, J. Y. (2019). Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, *15*(11), e1008449. https://doi.org/10.1371/journal.pgen.1008449

Beeravolu, C. R., Hickerson, M. J., Frantz, L. A. F., & Lohse, K. (2018). ABLE: Blockwise site frequency spectra for inferring complex population histories and recombination. *Genome Biology*, *19*, 145. https://doi.org/10.1186/s13059-018-1517-y

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Bitarello, B. D., De Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andres, A. M. (2018). Signatures of long-term balancing selection in human genomes. *Genome Biology and Evolution*, *10*(3), 939–955. https://doi.org/10.1093/gbe/evy054

Boistard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data – An approximate Bayesian computation approach. *PLoS Genetics*, 858–865, https://doi.org/10.1371/journal.pgen.1005877

Bolnick, D. I., & Otto, S. P. (2013). The magnitude of local adaptation under genotype-dependent dispersal. *Ecology and Evolution*, *3*(14), 4722–4735. https://doi.org/10.1002/ece3.850

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J. M., Blott, S., & San Cristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, *186*, 241–262. https://doi.org/10.1534/genetics.110.117275

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, *10*(4), 1–6. https://doi.org/10.1371/journal.pcbi.1003537

Boyrie, L., Moreau, C., Frugier, F., Jacquet, C., & Bonhomme, M. (2020). A linkage disequilibrium-based statistical test for Genome-Wide Epistatic Selection Scans in structured populations. *Heredity*, *126*(1), 77–91. https://doi.org/10.1038/s41437-020-0349-1

Bradburd, G. S., Coop, G. M., & Ralph, P. L. (2017). Inferring continuous and discrete population genetic structure across space. *Genetics*, *210*(September), 33–52. https://doi.org/10.1101/189688

Bradburd, G. S., Ralph, P. L., & Coop, G. M. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, *67*(11), 3258–3273. https://doi.org/10.1111/evo.12193

Bradburd, G. S., Ralph, P. L., & Coop, G. M. (2016). A spatial framework for understanding population structure and admixture. *PLoS Genetics*, *12*(1), 1–38. https://doi.org/10.1371/journal.pgen.1005703

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics (Oxford, England)*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G., & Bustamante, C. D. (2012). PCAdmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology*, *84*(4), 343–364. https://doi.org/10.3378/027.084.0401

Browning, B. L., & Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, *88*(2), 173–182. https://doi.org/10.1016/j.ajhg.2011.01.010

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & Roychoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932. https://doi.org/10.1093/molbev/mss086

Buerkle, C. A., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: How low should we go? *Molecular Ecology*, *22*(11), 3028–3035. https://doi.org/10.1111/mec.12105

Cadzow, M., Boocock, J., Nguyen, H. T., Wilcox, P., Merriman, T. R., & Black, M. A. (2014). A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics*, *5*(AUG), 1–8. https://doi.org/10.3389/fgene.2014.00293

Caye, K., Deist, T. M., Martins, H., Michel, O., & François, O. (2016). TESS3: Fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, *16*(2), 540–548. https://doi.org/10.1111/1755-0998.12471

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics*, *8*(12), e1003090. https://doi.org/10.1371/journal.pgen.1003090

Charlesworth, B., Nordborg, M., & Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research*, *70*(02), 155–174. https://doi.org/10.1017/S0016672397002954

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, *2*(4), e64. https://doi.org/10.1371/journal.pgen.0020064

Chifman, J., & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, *30*(23), 3317–3324. https://doi.org/10.1093/bioinformatics/btu530

Chou, J., Gupta, A., Yaduvanshi, S., Davidson, R., Nute, M., Mirarab, S., & Warnow, T. (2015). A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, *16*, S2.

Christe, C., Stolting, K. N., Paris, M., Fraisse, C., Bierne, N., & Lexer, C. (2016). Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Molecular Ecology*, *26*(1), 59–76. https://doi.org/10.1111/mec.13765

Clemente, F., Gautier, M., & Vitalis, R. (2018). Inferring sex-specific demographic history from SNP data. *PLoS Genetics*, *14*(1), 1–32. https://doi.org/10.1371/journal.pgen.1007191

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T., & Estoup, A. (2008). Inferring population history with DIY ABC: A user-friendly approach to approximate Bayesian computation. *Bioinformatics*, *24*(23), 2713–2719. https://doi.org/10.1093/bioinformatics/btn514

Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Sauné, L., Genson, G., Dubois, E., Nidelet, S., Deuve, T., & Rasplus, J.-Y. (2014). Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution*, *31*(5), 1272–1274. https://doi.org/10.1093/molbev/msu063

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–3157. https://doi.org/10.1111/mec.12796

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Cubry, P., Tranchant-Dubreuil, C., Thuillet, A.-C., Monat, C., Ndjiondjop, M.-N., Labadie, K., Cruaud, C., Engelen, S., Scarcelli, N., Rhoné, B., Burgarella, C., Dupuy, C., Larmande, P., Wincker, P., François, O., Sabot, F., & Vigouroux, Y. (2018). The rise and fall of african rice cultivation revealed by analysis of 246 new genomes. *Current Biology*, *28*(14), 2274–2282.e6. https://doi.org/10.1016/j.cub.2018.05.066

Currat, M., Arenas, M., Quilodràn, C. S., Excoffier, L., & Ray, N. (2019). SPLATCHE3: Simulation of serial genetic data under spatially explicit evolutionary scenarios including long-distance dispersal. *Bioinformatics*, *35*(21), 4480–4483. https://doi.org/10.1093/bioinformatics/btz311

Cushman, S. A. (2014). Grand challenges in evolutionary and population genetics: The importance of integrating epigenetics, genomics, modeling, and experimentation. *Frontiers in Genetics*, *5*(JUL), 1–5. https://doi.org/10.3389/fgene.2014.00197

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Degiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016). SWEEPFINDER 2: Increased sensitivity, robustness, and flexibility. *Bioinformatics*, 10.111/mec.13351.RR.*32*(12), 1895–1897. https://doi.org/10.1093/bioinformatics/btw051

DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genetics*, *10*(8), e1004561. https://doi.org/10.1371/journal.pgen.1004561

Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10*(1), 24–29. https://doi.org/10.1038/s41467-019-13225-y

Doran, A. G., & Creevey, C. J. (2013). Snpdat: Easy and rapid annotation of results from de novo snp discovery projects for model and

non-model organisms. *BMC Bioinformatics*, *14*(1), 45. https://doi.org/10.1186/1471-2105-14-45

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*, 214. https://doi.org/10.1186/1471-2148-7-214

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., & Blum, M. G. B. (2016). Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Molecular Biology and Evolution*, *33*(4), 1082–1093. https://doi.org/10.1093/molbev/msv334

Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, *23*(1), 40–69. https://doi.org/10.1111/mec.12581

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Ebert, D., & Fields, P. D. (2020). Host–parasite co-evolution and its genomic signature. *Nature Reviews Genetics*, *21*(12), 754–768. https://doi.org/10.1038/s41576-020-0269-1

Edelaar, P., & Bolnick, D. I. (2012). Non-random gene flow: An underappreciated force in evolution and ecology. *Trends in Ecology & Evolution*, *27*(12), 659–665. https://doi.org/10.1016/j.tree.2012.07.009

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., & Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, *94*, 447–462.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., & Wolf, J. B. W. (2012). The genomic landscape of species divergence in *Ficedula flycatchers*. *Nature*, *491*(7426), 756–760. https://doi.org/10.1038/nature11584

Ewing, G., & Hermisson, J. (2010). MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, *26*(16), 2064–2065. https://doi.org/10.1093/bioinformatics/btq322

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., & Foll, M. (2011). Fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*(9), 1332–1334. https://doi.org/10.1093/bioinformatics/btr124

Excoffier, L., & Heckel, G. (2006). Computer programs for population genetics data analysis: A survival guide. *Nature Reviews. Genetics*, *7*(10), 745–758. https://doi.org/10.1038/nrg1904

Excoffier, L., & Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology and Evolution*, *23*(7), 347–351. https://doi.org/10.1016/j.tree.2008.04.004

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405–1413. https://doi.org/10.1093/genetics/155.3.1405

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), 1275–1291. https://doi.org/10.1093/molbev/msu077

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576. https://doi.org/10.1111/mec.12522

François, O., & Jay, F. (2020). Factor analysis of ancient population genomic samples. *Nature Communications*, *11*(4661). https://doi.org/10.1038/s41467-020-18335-6

François, O., Martins, H., Caye, K., & Schoville, S. (2015). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, *25*, 454–469. https://doi.org/10.1111/mec.13513

Fraser, D. J., & Bernatchez, L. (2001). Adaptive evolutionary conservation: Towards a unified concept for defining conservation units. *Molecular Ecology*, *10*(12), 2741–2752. https://doi.org/10.1046/j.1365-294X.2001.t01-1-01411.x

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, *196*(4), 973–983. https://doi.org/10.1534/genetics.113.160572

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, *30*(7), 1687–1699. https://doi.org/10.1093/molbev/mst063

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, *186*(1), 207–218. https://doi.org/10.1534/genetics.110.114397

Gao, F., Ming, C., Hu, W., & Li, H. (2016). New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3 Genes|genomes|genetics*, *6*(6), 1563–1571. https://doi.org/10.1534/g3.116.028233

Garrigan, D. (2013). POPBAM: Tools for evolutionary analysis of short read sequence alignments. *Evolutionary Bioinformatics*, *2013*(9), 343–353. https://doi.org/10.4137/EBO.S12751

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*, *11*(2), 1–32. https://doi.org/10.1371/journal.pgen.1005004

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, *201*(4), 1555–1579. https://doi.org/10.1534/genetics.115.181453

Gautier, M., & Vitalis, R. (2012). Rehh an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, *28*(8), 1176–1177. https://doi.org/10.1093/bioinformatics/bts115

Gower, G., Picazo, P. I., Fumagalli, M., & Racimo, F. (2020). Detecting adaptive introgression in human evolution using convolutional neural networks. *BioRxiv*. https://doi.org/10.1101/2020.09.18.301069

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, *43*(10), 1031–1034. https://doi.org/10.1038/ng.937

Guedj, B., & Guillot, G. (2011). Estimating the location and shape of hybrid zones. *Molecular Ecology Resources*, *11*(6), 1119–1123. https://doi.org/10.1111/j.1755-0998.2011.03045.x

Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, *4*(4), 336–344. https://doi.org/10.1111/2041-210x.12018

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*(1), 205–220. https://doi.org/10.1534/genetics.113.152462

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*(1), 5–23. https://doi.org/10.1111/mec.13339

Habel, J., Zachos, F., Dapporto, L., Rödder, D., Radespiel, U., Tellier, A., & Schmitt, T. (2015). Population genetics revisited – Towards a multidisciplinary research field. *Biological Journal of the Linnean Society*, *115*, 1–12. https://doi.org/10.1111/bij.12481

Haller, B. C., & Messer, P. W. (2019). Slim 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Han, F., Lamichhaney, S., Rosemary Grant, B., Grant, P. R., Andersson, L., & Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research*, *27*(6), 1004–1015. https://doi.org/10.1101/gr.212522.116

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, *8*(5), 1–9. https://doi.org/10.1093/gigascience/giz034

Harris, K., & Nielsen, R. (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, *9*(6), https://doi.org/10.1371/journal.pgen.1003521

Harris, R. B., Sackman, A., & Jensen, J. D. (2018). On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *PloS Genetics*, *14*(12), e1007859–, https://doi.org/10.1371/journal.pgen.1007859

Hartfield, M., Wright, S. I., & Agrawal, A. F. (2016). Coalescent times and patterns of genetic diversity in species with facultative sex: Effects of gene conversion, population structure, and heterogeneity. *Genetics*, *202*(1), 297–312. https://doi.org/10.1534/genetics.115.178004

Hedrick, P. W. (2006). Genetic polymorphism in heterogeneous environments: The age of genomics. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 67–93. https://doi.org/10.1146/annurev.ecolsys.37.091305.110132

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *22*(18), 4606–4618. https://doi.org/10.1111/mec.12415

Hejase, H. A., Salman-Minkov, A., Campagna, L., Hubisz, M. J., Lovette, I. J., Gronau, I., & Siepel, A. (2020). Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30554–30565. https://doi.org/10.1073/pnas.2015987117

Heled, J., & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, *27*(3), 570–580. https://doi.org/10.1093/molbev/msp274

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(February), 747–751. https://doi.org/10.1126/science.1243518

Hendry, A. P. (2004). Selection against migrants contributes to the rapid evolution of ecologically dependent reproductive isolation. *Evolutionary Ecology Research*, *6*(8), 1219–1236.

Hohenlohe, P. A. (2014). Ecological genomics in full colour. *Molecular Ecology*, *23*(21), 5129–5131. https://doi.org/10.1111/mec.12945

Hubisz, M. J., Williams, A. L., & Siepel, A. (2020). Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genetics*, *16*(8), 1–24. https://doi.org/10.1371/JOURNAL.PGEN.1008895

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Huisman, J. (2017). Pedigree reconstruction from SNP data: Parentage assignment, sibship clustering and beyond. *Molecular Ecology Resources*, *17*(5), 1009–1024. https://doi.org/10.1111/1755-0998.12665

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Jackson, N. D., Morales, A. E., Carstens, B. C., & O'Meara, B. C. (2017). PHRAPL: Phylogeographic inference using approximate likelihoods. *Systematic Biology*, *66*(6), 1045–1053. https://doi.org/10.1093/sysbio/syx001

Jenkins, P. A., Song, Y. S., & Brem, R. B. (2012). Genealogy-based methods for inference of historical recombination and gene flow and their application in *Saccharomyces cerevisiae*. *PLoS One*, *7*(11), e46947. https://doi.org/10.1371/journal.pone.0046947

Jenner, R. A., & Wills, M. A. (2007). The choice of model organisms in evo-devo. *Nature Reviews. Genetics*, *8*(4), 311–319. https://doi.org/10.1038/nrg2062

Jensen, J. D. (2014). On the unfounded enthusiasm for soft selective sweeps. *Nature Communications*, *5*, 5281. https://doi.org/10.1038/ncomms6281

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Jombart, T., Devillard, S., Dufour, A.-B., & Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, *101*, 92–103. https://doi.org/10.1038/hdy.2008.34

Jombart, T., Pontier, D., & Dufour, A.-B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity*, *102*, 330–341. https://doi.org/10.1038/hdy.2008.130

Joseph, T. A., & Pe'er, I. (2019). Inference of population structure from time-series genotype data. *American Journal of Human Genetics*, *105*(2), 317–333. https://doi.org/10.1016/j.ajhg.2019.06.002

Jostins, L., & McVean, G. (2016). Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. *Bioinformatics*, *32*(12), 1898–1900. https://doi.org/10.1093/bioinformatics/btw075

Jouganous, J., Long, W., Ragsdale, A. P., & Gravel, S. (2017). Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, *206*(3), 1549–1567. https://doi.org/10.1534/genetics.117.200493

Kamm, J., Terhorst, J., Durbin, R., & Song, Y. S. (2020). Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, *115*(531), 1472–1487. https://doi.org/10.1080/01621459.2019.1635482

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, *12*(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Kemppainen, P., Knight, C. G., Sarma, D. K., Hlaing, T., Prakash, A., Maung Maung, Y. N., Somboon, P., Mahanta, J., & Walton, C. (2015). Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Molecular Ecology Resources*, *15*(5), 1031–1045. https://doi.org/10.1111/1755-0998.12369

Kern, A. D., & Schrider, D. R. (2016). Discoal: Flexible coalescent simulations with selection. *Bioinformatics*, *32*(24), 3839–3841. https://doi.org/10.1093/bioinformatics/btw556

Kern, A. D., & Schrider, D. R. (2018). diploS/HIC: An updated approach to classifying selective sweeps. *G3 Genes|genomes|genetics*, *8*(6), 1959–1970. https://doi.org/10.1534/g3.118.200262

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., & Schlötterer, C. (2011). PoPoolation: A toolbox for population genetic analysis of next generation sequencing

data from pooled individuals. *PLoS One*, *6*(1), e15925. https://doi.org/10.1371/journal.pone.0015925

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Kolaczkowski, B., Kern, A. D., Holloway, A. K., & Begun, D. J. (2011). Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. *Genetics*, *187*(1), 245–260. https://doi.org/10.1534/genetics.110.123059

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Koropoulis, A., Alachiotis, N., & Pavlidis, P. (2020). Detecting positive selection in populations using genetic data. In J. Y. Dutheil (Ed.). *Statistical population genomics* (pp. 87–123). Springer. https://doi.org/10.1007/978-1-0716-0199-0_5

Kubota, S., Iwasaki, T., Hanada, K., Nagano, A. J., Fujiyama, A., Toyoda, A., Sugano, S., Suzuki, Y., Hikosaka, K., Ito, M., & Morinaga, S.-I. (2015). A genome scan for genes underlying microgeographic-scale local adaptation in a wild arabidopsis species. *PLoS Genetics*, *11*(7), 1–26. https://doi.org/10.1371/journal.pgen.1005361

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., & Castellano, S. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, *530*(7591), 429–433. https://doi.org/10.1038/nature16544

Laland, K. N., Sterelny, K., Odling-Smee, J., Hoppitt, W., & Uller, T. (2011). Cause and effect in biology revisited: Is Mayr's proximate-ultimate dichotomy still useful? *Science (New York, N.Y.)*, *334*(6062), 1512–1516. https://doi.org/10.1126/science.1210879

Lawson, D. J., van van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun*, 066431, *9*, 3258https://doi.org/10.1038/s41467-018-05257-7

Leaché, A. D., Banbury, B. L., Felsenstein, J., De Oca, A. N. M., & Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*, *64*, 1032–1047.

Leaché, A. D., Harris, R. B., Rannala, B., & Yang, Z. (2014). The influence of gene flow on species tree estimation: A simulation study. *Systematic Biology*, *63*, 17–30.

Lee, T.-H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. (2014). SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, *15*(1), 162. https://doi.org/10.1186/1471-2164-15-162

Legendre, P., & Fortin, M. J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, *10*(5), 831–844. https://doi.org/10.1111/j.1755-0998.2010.02866.x

Legrand, D., Tenaillon, M. I., Matyot, P., Gerlach, J., Lachaise, D., & Cariou, M.-L. (2009). Species-wide genetic variation and demographic history of *Drosophila sechellia*, a species lacking population structure. *Genetics*, *182*(4), 1197–1206. https://doi.org/10.1534/genetics.108.092080

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P. A., & Bernatchez, L. (2020). Using haplotype information for conservation genomics. *Trends in Ecology and Evolution*, *35*(3), 245–258. https://doi.org/10.1016/j.tree.2019.10.012

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 99–121.

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175–195. https://doi.org/10.1093/genetics/74.1.175

Lexer, C., Mangili, S., Bossolini, E., Forest, F., Stölting, K. N., Pearman, P. B., Zimmermann, N. E., & Salamin, N. (2013). 'Next generation' biogeography: Towards understanding the drivers of species diversification and persistence. *Journal of Biogeography*, *40*(6), 1013–1022. https://doi.org/10.1111/jbi.12076

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, J., Li, H., Jakobsson, M., Li, S., SjÖdin, P., & Lascoux, M. (2012). Joint analysis of demography and selection in population genetics: Where do we stand and where could we go? *Molecular Ecology*, *21*(1), 28–44. https://doi.org/10.1111/j.1365-294X.2011.05308.x

Librado, P., & Orlando, L. (2018). Detecting signatures of positive selection along defined branches of a population tree using LSD. *Molecular Biology and Evolution*, *35*, 1520–1535. https://doi.org/10.1093/molbev/msy053

Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., & Wegmann, D. (2017). ATLAS: Analysis tools for low-depth and ancient samples. *BioRxiv*. https://doi.org/10.1101/105346

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

Liu, H. J., & Yan, J. (2019). Crop genome-wide association study: A harvest of biological relevance. *Plant Journal*, *97*(1), 8–18. https://doi.org/10.1111/tpj.14139

Liu, L., & Yu, L. (2010). Phybase: An R package for species tree analysis. *Bioinformatics*, *26*(7), 962–963. https://doi.org/10.1093/bioinformatics/btq062

Liu, L., & Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic Biology*, *60*(5), 661–667. https://doi.org/10.1093/sysbio/syr027

Liu, X., & Fu, Y. X. (2020). Stairway Plot 2: Demographic history inference with folded SNP frequency spectra. *Genome Biology*, *21*(1), 1–9. https://doi.org/10.1186/s13059-020-02196-9

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, *23*(9), 2178–2192. https://doi.org/10.1111/mec.12725

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite – Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*, 584–595. https://doi.org/10.1111/1755-0998.13265

Mandoli, D. F., & Olmstead, R. (2000). The importance of emerging model systems in plant biology. *Journal of Plant Growth Regulation*, *19*(3), 249–252. https://doi.org/10.1007/s003440000038

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Martin, S. H., & Van Belleghem, S. M. (2017). Exploring evolutionary relationships across the genome using topology weighting. *Genetics*, *206*(1), 429–438. https://doi.org/10.1534/genetics.116.194720

Mazet, O., Rodriguez, W., & Chikhi, L. (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology*, *104*, 46–58. https://doi.org/10.1016/j.tpb.2015.06.003

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, *351*(6328), 652–654. https://doi.org/10.1038/351652a0

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*(10), e1000686. https://doi.org/10.1371/journal.pgen.1000686

McVean, G., Awadalla, P., & Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene

sequences. *Genetics*, *160*(3), 1231–1241. https://doi.org/10.1093/genetics/160.3.1231

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*, *28*(11), 659–669. https://doi.org/10.1016/j.tree.2013.08.003

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Teeling, E. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*(12), i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Mirzaei, S., & Wu, Y. (2017). RENT+: An improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, *33*(7), 1021–1030. https://doi.org/10.1093/bioinformatics/btw735

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, *310*(5746), 321–324. https://doi.org/10.1126/science.1117196

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Neuenschwander, S., Michaud, F., & Goudet, J. (2019). QuantiNemo 2: A Swiss knife to simulate complex demographic and genetic scenarios, forward and backward in time. *Bioinformatics*, *35*(5), 886–888. https://doi.org/10.1093/bioinformatics/bty737

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, *8*(11), 857–868. https://doi.org/10.1038/nrg2187

Noskova, E., Ulyantsev, V., Koepfli, K.-P., O'Brien, S. J., & Dobrynin, P. (2020). GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *GigaScience*, *9*(3), 1–18. https://doi.org/10.1093/gigascience/giaa005

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

Orozco-terWengel, P. (2016). The devil is in the details: The effect of population structure on demographic inference. *Heredity*, *116*(4), 349–350. https://doi.org/10.1038/hdy.2016.9

Palamara, P. F., & Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics*, *29*(13), 180–188. https://doi.org/10.1093/bioinformatics/btt239

Pavlidis, P., Jensen, J. D., Stephan, W., & Stamatakis, A. (2012). A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, *29*(10), 3237–3248. https://doi.org/10.1093/molbev/mss136

Pavlidis, P., Laurent, S., & Stephan, W. (2010). MsABC: A modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, *10*(4), 723–727. https://doi.org/10.1111/j.1755-0998.2010.02832.x

Petkova, D., Novembre, J., & Stephens, M. (2015). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, *48*(1), 94–100. https://doi.org/10.1038/ng.3464

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Baglione, V., & Wolf, J. B. W. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, *344*(6190), 1410–1414.

Price, A., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Pritchard, J. K., & Di Rienzo, A. (2010). Adaptation – Not by sweeps alone. *Nature Reviews Genetics*, *11*(10), 665–667. https://doi.org/10.1038/nrg2880

Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, *20*(4), R208–R215. https://doi.org/10.1016/j.cub.2009.11.055

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Puttick, M. N. (2019). MCMCtreeR: Functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics*, *35*(24), 5321–5322. https://doi.org/10.1093/bioinformatics/btz554

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, *10*(5), https://doi.org/10.1371/journal.pgen.1004342

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: Finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*, 1450–1477. https://doi.org/10.1111/jeb.13047

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, *35*(10), 1720–1728. https://doi.org/10.1093/bioinformatics/bty867

Refoyo-Martínez, A., Da Fonseca, R. R., Halldórsdóttir, K., Árnason, E., Mailund, T., & Racimo, F. (2019). Identifying loci under positive selection in complex population histories. *Genome Research*, *29*(9), 1506–1520. https://doi.org/10.1101/gr.246777.118

Robert, C. P., Cornuet, J.-M., Marin, J.-M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15112–15117. https://doi.org/10.1073/pnas.1102900108

Rockman, M. V. (2012). The QTN program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*, *66*(1), 1–17. https://doi.org/10.1111/j.1558-5646.2011.01486.x

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., & Chikhi, L. (2018). The IICR and the non-stationary structured coalescent: Towards demographic inference with arbitrary changes in population structure. *Heredity*, *121*(6), 663–678. https://doi.org/10.1038/s41437-018-0148-0

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, *14*(12), e2000234. https://doi.org/10.1371/JOURNAL.PBIO.2000234

Roux, C., Pauwels, M., Ruggiero, M.-V., Charlesworth, D., Castric, V., & Vekemans, X. (2013). Recent and ancient signature of balancing selection around the S-locus in *Arabidopsis halleri* and *A.*

lyrata. *Molecular Biology and Evolution*, *30*(2), 435–447. https://doi.org/10.1093/molbev/mss246

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837. https://doi.org/10.1038/nature01027.1

Sabeti, P. C., Varilly, P., Fry, B., McCarroll, S. A., Frazer, K. A., Ballinger, D. G., & Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

Salter-Townshend, M., & Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics*, *212*(July), 869–889. https://doi.org/10.1534/genetics.119.302139

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*(4), 629–644. https://doi.org/10.1086/502802

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

Schrider, D. R., & Kern, A. D. (2016). S/HIC: Robust identification of soft and hard sweeps using machine learning. *PLoS Genetics*, *12*(3), 1–31. https://doi.org/10.1371/journal.pgen.1005928

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Schrider, D. R., Mendes, F. K., Hahn, M. W., & Kern, A. D. (2015). Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics*, *200*(1), 267–284. https://doi.org/10.1534/genetics.115.174912

Schrider, D. R., Shanku, A. G., & Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, *204*(3), 1207–1223. https://doi.org/10.1534/genetics.116.190223

Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., & Orlando, L. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(52), 201416991. https://doi.org/10.1073/pnas.1416991111

Sellis, D., Callahan, B. J., Petrov, D. A., & Messer, P. W. (2011). Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), 20666–20671. https://doi.org/10.1073/pnas.1114573108

Setter, D., Mousset, S., Cheng, X., Nielsen, R., DeGiorgio, M., & Hermisson, J. (2020). VolcanoFinder: Genomic scans for adaptive introgression. *PLoS Genetics*, *16*(6), 1–44. https://doi.org/10.1371/journal.pgen.1008867

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., Colling, G., Dalén, L., De Meester, L., Ekblom, R., Fawcett, K. D., Fior, S., Hajibabaei, M., Hill, J. A., Hoezel, A. R., Höglund, J., Jensen, E. L., Krause, J., Kristensen, T. N., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, *30*(2), 78–87. https://doi.org/10.1016/j.tree.2014.11.009

Sheehan, S., Harris, K., & Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, *194*, 647–662. https://doi.org/10.1534/genetics.112.149096

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS Computational Biology*, *12*(3), 1–28. https://doi.org/10.1371/journal.pcbi.1004845

Siewert, K. M., & Voight, B. F. (2017). Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, *34*(11), 2996–3005. https://doi.org/10.1093/molbev/msx209

Siewert, K. M., & Voight, B. F. (2020). BetaScan2: Standardized statistics to detect balancing selection utilizing substitution data. *Genome Biology and Evolution*, *12*(2), 3873–3877. https://doi.org/10.1093/gbe/evaa013

Smith, C. C. R., & Flaxman, S. M. (2020). Leveraging whole genome sequencing data for demographic inference with approximate Bayesian computation. *Molecular Ecology Resources*, *20*(1), 125–139. https://doi.org/10.1111/1755-0998.13092

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., Nachman, M. W., & Kohn, M. H. (2011). Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, *21*(15), 1296–1301. https://doi.org/10.1016/j.cub.2011.06.043

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321–1329. https://doi.org/10.1038/s41588-019-0484-x

Staab, P. R., & Metzler, D. (2016). Coala: An R framework for coalescent simulation. *Bioinformatics*, *32*(12), 1903–1904. https://doi.org/10.1093/bioinformatics/btw098

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). Scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, *31*(10), 1680–1682. https://doi.org/10.1093/bioinformatics/btu861

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stamatakis, A., Alachiotis, N., Pavlidis, P., & Daniel, Z. (2013). SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, *30*(9), 2224–2234. https://doi.org/10.1093/molbev/mst112

Stern, A. J., Speidel, L., Zaitlen, N. A., & Nielsen, R. (2021). Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *The American Journal of Human Genetics*, *108*(2), 219–239, https://doi.org/10.1016/j.ajhg.2020.12.005

Stern, A. J., Wilton, P. R., & Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genetics*, *15*(9), e1008384. https://doi.org/10.1371/journal.pgen.1008384

Stucki, S., Orozco-terWengel, P. Forester, B. R. Duruz, S. Colli, L. Masembe, C., Negrini, R., Landguth, E., Jones, M. R., The NEXTGEN Consortium, Bruford, M. W., Taberlet, P., & Joost, S. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, *17*(5), 1072–1089. https://doi.org/10.1111/1755-0998.12629

Sugden, L. A., Atkinson, E. G., Fischer, A. P., Rong, S., Henn, B. M., & Ramachandran, S. (2018). Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nature Communications*, *9*(703), https://doi.org/10.1038/s41467-018-03100-7

Svedberg, J., Shchur, V., Reinman, S., Nielsen, R., & Corbett-Detig, R. (2021). Inferring Adaptive Introgression Using Hidden Markov Models. *Molecular Biology and Evolution*, *38*(5), 2152–2165. https://doi.org/10.1093/molbev/msab014

Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, *31*(10), 2824–2827. https://doi.org/10.1093/molbev/msu211

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Takezaki, N., Nei, M., & Tamura, K. (2010). POPTREE2: Software for constructing population trees from allele frequency data and

computing other population statistics with windows interface. *Molecular Biology and Evolution*, *27*(4), 747–752. https://doi.org/10.1093/molbev/msp312

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tang, K., Thornton, K. R., & Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, *5*(7), 1587–1602. https://doi.org/10.1371/journal.pbio.0050171

Tataru, P., & Bataillon, T. (2019). PolyDFEv2.0: Testing for invariance of the distribution of fitness effects within and across species. *Bioinformatics*, *35*(16), 2868–2869. https://doi.org/10.1093/bioinformatics/bty1060

Tataru, P., & Bataillon, T. (2020). polyDFE: Inferring the distribution of fitness effects and properties of beneficial mutations from polymorphism data. In J. Y. Dutheil (Ed.). *Statistical population genomics* (pp. 125–146). Springer. https://doi.org/10.1007/978-1-0716-0199-0_6

Tataru, P., Nirody, J. A., & Song, Y. S. (2014). DiCal-IBD: Demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics*, *30*(23), 3430–3431. https://doi.org/10.1093/bioinformatics/btu563

Taylor, H. R., Dussex, N., & van Heezik, Y. (2017). Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation*, *10*, 231–242. https://doi.org/10.1016/j.gecco.2017.04.001

Terhorst, J., Kamm, J. A., & Song, Y. S. (2016). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309. https://doi.org/10.1038/ng.3748

The Heliconius Genome Consortium, Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., & Jiggins, C. D. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, *487*(7405), 94–98. https://doi.org/10.1038/nature11041

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., & Fumagalli, M. (2019). ImaGene: A convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, *20*(Suppl 9), 1–12. https://doi.org/10.1186/s12859-019-2927-x

Vachaspati, P., & Warnow, T. (2018). SVDquest: Improving SVDquartets species tree estimation using exact optimization within a constrained search space. *Molecular Phylogenetics and Evolution*, *124*, 122–136. https://doi.org/10.1016/j.ympev.2018.03.006

Vitalis, R., Gautier, M., Dawson, K. J., & Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics*, *196*(3), 799–817. https://doi.org/10.1534/genetics.113.152991

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., & Seehausen, O. (2013). Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, *22*(3), 787–798. https://doi.org/10.1111/mec.12023

Wang, J. (2019). Pedigree reconstruction from poor quality genotype data. *Heredity*, *122*(6), 719–728. https://doi.org/10.1038/s41437-018-0178-7

Wang, J., & Zhang, Z. (2020). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *BioRxiv*, https://doi.org/10.1101/2020.11.29.403170

Wang, M. H., Cordell, H. J., & Van Steen, K. (2019). Statistical methods for genome-wide association studies. *Seminars in Cancer Biology*, *55*, 53–60. https://doi.org/10.1016/j.semcancer.2018.04.008

Wang, M., Huang, X., Li, R., Xu, H., Jin, L., & He, Y. (2014). Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree. *Molecular Biology and Evolution*, *31*(11), 3068–3080. https://doi.org/10.1093/molbev/msu244

Weber, J. N., Peterson, B. K., & Hoekstra, H. E. (2013). Discrete genetic modules are responsible for complex burrow evolution in Peromyscus mice. *Nature*, *493*(7432), 402–405. https://doi.org/10.1038/nature11816

Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: A versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, *11*, 116. https://doi.org/10.1186/1471-2105-11-116

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370.

Wen, J., Liu, J., Ge, S., Xiang, Q.-Y., & Zimmer, E. A. (2015). Phylogenomic approaches to deciphering the tree of life. *Journal of Systematics Evolution*, *53*, 369–370. https://doi.org/10.1111/jse.12175

White, B. J., Cheng, C., Simard, F., Costantini, C., & Besansky, N. J. (2010). Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular Ecology*, *19*(5), 925–939. https://doi.org/10.1111/j.1365-294X.2010.04531.x

Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H., & Reich, D. (2012). Phasing of many thousands of genotyped samples. *American Journal of Human Genetics*, *91*(2), 238–251. https://doi.org/10.1016/j.ajhg.2012.06.013

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310

Jombart, T., Pontier, D., & Dufour, A. -B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity (Edinb)*, *102*, 330–341.

How to cite this article: Bourgeois YX, Warren BH. An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Mol Ecol*. 2021;00:1–36. https://doi.org/10.1111/mec.15989

Recent Secondary Contacts, Linked Selection, and Variable Recombination Rates Shape Genomic Diversity in the Model Species *Anolis carolinensis*

Yann Bourgeois[1,*], Robert P. Ruggiero[1], Joseph D. Manthey[1,2], and Stéphane Boissinot[1,*]

[1]New York University Abu Dhabi, United Arab Emirates

[2]Department of Biological Sciences, Texas Tech University

*Corresponding authors: E-mails: yann.x.c.bourgeois@gmail.com; stephane.boissinot@nyu.edu.

Accepted: May 23, 2019

Data deposition: This project has been deposited to the Sequencing Read Archive under the Bioproject designation PRJNA533001.

Abstract

Gaining a better understanding on how selection and neutral processes affect genomic diversity is essential to gain better insights into the mechanisms driving adaptation and speciation. However, the evolutionary processes affecting variation at a genomic scale have not been investigated in most vertebrate lineages. Here, we present the first population genomics survey using whole genome resequencing in the green anole (*Anolis carolinensis*). Anoles have been intensively studied to understand mechanisms underlying adaptation and speciation. The green anole in particular is an important model to study genome evolution. We quantified how demography, recombination, and selection have led to the current genetic diversity of the green anole by using whole-genome resequencing of five genetic clusters covering the entire species range. The differentiation of green anole's populations is consistent with a northward expansion from South Florida followed by genetic isolation and subsequent gene flow among adjacent genetic clusters. Dispersal out-of-Florida was accompanied by a drastic population bottleneck followed by a rapid population expansion. This event was accompanied by male-biased dispersal and/or selective sweeps on the X chromosome. We show that the interaction between linked selection and recombination is the main contributor to the genomic landscape of differentiation in the anole genome.

Key words: *Anolis carolinensis*, recombination, divergence, selection.

Introduction

Nucleotide variation along a DNA sequence results from the interactions between multiple processes that either generate new alleles (e.g., recombination, mutation) or affect the fate of these alleles in populations (e.g., selection, demography, and speciation). The variable outcome of these interactions along the genome can result in heterogeneous patterns of diversity and divergence at both intra- and interspecific scales (Begun and Aquadro 1992; Nachman and Payseur 2012; Cruickshank and Hahn 2014; Roux et al. 2014; Seehausen et al. 2014; Wolf and Ellegren 2017). Given their importance in divergence and speciation, quantifying these processes has been at the core of evolutionary genomics for the last decade. With the advent of next-generation sequencing and the continuous development of novel analytical tools, it has become possible to properly quantify the impact of recombination (Booker et al. 2017; Kawakami et al. 2017), selection (Begun and Aquadro 1992; Barrett et al. 2008; Mullen and Hoekstra 2008; Cai et al. 2009), and demographic history (Gutenkunst et al. 2009; Excoffier et al. 2013; Roux et al. 2016) on diversity patterns in several vertebrates. Ultimately, such investigations have the power to answer outstanding biological questions such as the role of sex chromosomes, the nature of reproductive barriers, or the timing of gene flow and selection during differentiation (Wolf and Ellegren 2017). Assessing the effects of these mechanisms in a range of organisms is crucial to inform current debates about their relative importance (Bierne et al. 2011; Kern and Hahn 2018; Pouyet et al. 2018; Jensen et al. 2019) with a broader set of empirical data.

Thorough analyses of the factors affecting genome diversity at the peri-specific level have been performed in a small number of vertebrate species (see Ellegren et al. 2012; Sousa et al. 2013; Poelstra et al. 2014; Booker et al. 2017; Han et al. 2017; Kawakami et al. 2017) but several major clades of vertebrates have not been investigated at all. Among those are the nonavian reptiles, a speciose group of vertebrates that harbor a wide diversity of morphology and adaptation. Anoles in particular have been abundantly studied to understand the mechanisms underlying adaptation. This neotropical group of squamates diversified during the Cenozoic, and constitute a model system for understanding speciation and adaptation in ectotherms (Losos et al. 2004; Glor et al. 2005; Losos 2009; Kolbe et al. 2017; Lapiedra et al. 2018). The analysis of anoles genomes has also provided considerable insights on genome evolution in vertebrates (Alföldi et al. 2011; Fujita et al. 2011; Tollis and Boissinot 2011, 2013; Figuet et al. 2015; Costantini et al. 2016; Ruggiero et al. 2017). At last, they display a dramatic physiological, morphological, and behavioral diversity (Lailvaux et al. 2004; Glor et al. 2005; Wade 2012; Campbell-Staton et al. 2018; Lapiedra et al. 2018). Given this importance as a model species, we decided to perform a study on genome-wide variation in the green anole (*Anolis carolinensis*) to better understand its microevolutionary dynamics, expanding previous genetic, and genomic work (Tollis et al. 2012; Tollis and Boissinot 2014; Manthey et al. 2016).

The green anole is the first nonavian reptile for which the whole genome was sequenced (Alföldi et al. 2011) and its population structure is relatively well known (Tollis et al. 2012; Tollis and Boissinot 2014; Campbell-Staton et al. 2016; Manthey et al. 2016; Ruggiero et al. 2017). Whole-genome resequencing of green anoles populations would be an opportunity to better understand the drivers and constraints that act on their radiation at a resolution that was not allowed by the previous genetic data sets.

The green anole colonized Florida from Cuba (Glor et al. 2005; Campbell-Staton et al. 2012; Tollis et al. 2012; Tollis and Boissinot 2014; Manthey et al. 2016) between 6 and 12 Ma and diversified into five genetic groups: South Florida (SF), Eastern Florida (EF), Western Florida (WF), Gulf Atlantic (GA), and Carolinas (CA). Populations in Florida likely diverged in allopatry on island refugia before coming back into contact due to sea-level oscillations during the Pleistocene. Colonization of the rest of North America seems to be more recent, with two clades having probably expanded in the last 500,000 years (Manthey et al. 2016). It is the only species in the *Anolis* genus to have colonized temperate climates without human intervention.

Here, we present results obtained from whole-genome resequencing of five genetic clusters of the green anole. We provide a detailed assessment of the multiple factors that are likely to impact the green anole's genetic diversity at a genome-wide scale. We demonstrate that the combined effects of regional variation in recombination rate, linked selection, and migration are responsible for the heterogeneous genomic landscape of diversity and divergence in the green anole.

Materials and Methods

DNA Extraction and Whole Genome Sequencing

Whole genome sequencing libraries were generated from *A. carolinensis* liver tissue samples collected between 2009 and 2011 (Tollis et al. 2012), and *Anolis porcatus* and *Anolis allisoni* tissue samples generously provided by Breda Zimkus from the Museum of Comparative Zoology at Harvard University. For each of the 29 samples, DNA was isolated from ethanol preserved tissue using Ampure beads per the manufacturers' protocol. Illumina TRU-Seq paired end libraries were generated using 200 ng of DNA per sample and sequenced at the NYUAD Center for Genomics And Systems Biology Sequencing Core (http://nyuad.nyu.edu/en/research/infrastructure-and-support/core-technology-platforms.html; Last accessed on June 7, 2019) with an Illumina HiSeq 2500. Read quality was assessed with FastQCv0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc; Last accessed on June 7, 2019) and Trimmomatic (Bolger et al. 2014) was subsequently used to remove low quality bases, sequencing adapter contamination and systematic base calling errors. Specifically, the parameters "trimmomatic_adapter.fa: 2:30:10 TRAILING:3 LEADING:3 SLIDINGWINDOW:4:15 MINLEN:36" were used. Samples had an average of 1,519,339,234 read pairs, and after quality trimming 93.3% were retained as paired reads and 6.3% were retained as single reads. Sequencing data from this study have been submitted to the Sequencing Read Archive (https://www.ncbi.nlm.nih.gov/sra; last accessed: June 7, 2019) under the BioProject designation PRJNA533001.

Sequence Alignment and SNP Calling

Quality trimmed reads were aligned to the May 2010 assembly of the *A. carolinensis* reference genome (Broad AnoCar2.0/anoCar2; GCA_000090745.1; Alföldi et al. 2011) and processed for SNP detection with the assistance of the NYUAD Bioinformatics Core, using NYUAD variant calling pipeline (last accessed in June 2017). Briefly, the quality-trimmed FastQ reads of each sample were aligned to the AnoCar2.0 genome using the BWA-mem short read alignment approach (Li and Durbin 2011) and resulting SAM files were converted into BAM format, sorted, and indexed using SAMtools (Li et al. 2009). Picard was then used to identify insertions, deletions, and duplications in the sorted BAM files (http://broadinstitute.github.io/picard/; Last accessed on June 7, 2019) and evaluated using SAMtools (stats and depth). Alignments contained an average of 204,459,544 reads that passed QC, 97.75% mapping and 91.93% properly

paired (supplementary table S1, Supplementary Material online). Each individual resequenced genome was then processed with GATK for indel realignment, SNP and indel discovery, and genotyping, following GATK Best Practices (Depristo et al. 2011; Van Der Auwera et al. 2014). GATK joint genotyping was conducted with HaplotypeCaller for increased sensitivity and confidence, and results were selectively compared with results generated from SAMtools mpileup (Li et al. 2009). Filtering was performed in VCFtools (Danecek et al. 2011), with the following criteria: a 6× minimum depth of coverage per individual, a 15× maximum average depth of coverage, no more than 40% missing data across all 29 samples, a minimum quality score of 20 per site, and a minimum genotype quality score of 20.

Population Structure

To assess genetic structure, we conducted a clustering analysis using discriminant analysis of principal components (DAPC) on a subset of ~6,500 SNPs with <20% missing data and randomly thinned every 10 kb to minimize linkage disequilibrium (LD) between markers while retaining enough variants for inference. DAPC (Jombart et al. 2010) first estimates principal components (PC) describing variance in SNP data sets, then performs a discriminant analysis on these PC axes to identify genetic groupings. We selected the clustering model with the highest support using the Bayesian Information Criterion (BIC). We retained two principal components that explained ~40% of the total variance, and two of the linear discriminants. The probability for each individual to be assigned to a specific cluster was summarized by a barplot with the function compoplot() provided with the DAPC R package. We also described relationships between individuals with the same data set using the network algorithm implemented in Splitstree v4 (Huson and Bryant 2006). Lastly, we filtered the entire SNP data set to include one million randomly sampled SNPs present in a minimum of 80% of the individuals for use as input in RAxML v8 (Stamatakis 2014). We used RAxML to create a maximum-likelihood phylogeny, using the GTRGAMMA model of sequence evolution, and 100 rapid bootstraps to assess support for the phylogeny with the highest likelihood.

We further examined patterns of diversity and the shape of the allele frequency spectrum in each cluster by computing two summary statistics, the average number of pairwise differences θ_π (or nucleotide diversity) per bp, and Tajima's *D*, for nonoverlapping 5-kb windows using the software POPGENOME (Pfeifer et al. 2014). We removed windows overlapping ambiguities in the green anole genome using BEDTOOLS v2.25.0 (Quinlan and Hall 2010).

Demographic Estimates without Gene Flow

We used the multiepoch model implemented in SMC++ (Terhorst et al. 2017) to reconstruct population size trajectories and time since population split for each of the five genetic clusters of green anoles. This software is an extension of the Pairwise Sequentially Markov Coalescent (Li and Durbin 2011) that uses the spatial arrangement of polymorphisms along genome sequences to naively infer variation in effective population sizes and splitting time between populations. It has the advantage of using both information related to the site frequency spectrum and patterns of LD to make demographic inferences. Another benefit of this algorithm is that it is phase-insensitive, limiting the propagation of phasing errors that can bias effective population size estimates for recent times (Terhorst et al. 2017).

Within each of the five genetic clusters, we created one data set per individual for each of the six autosomes and combined those individual data sets to reconstruct past variation in effective population sizes. A mutation rate of 2.1×10^{-10} per site per generation and a generation time of 1 year (Tollis and Boissinot 2014) were assumed to translate coalescence times into years. We set a polarization error of 0.5 since the ancestral allele could not be determined for many loci. We also estimated splitting times between genetic clusters. However, these estimates should be taken with caution as the method assumes that no gene flow occurs after the split.

Effective Sex-Ratio

Sex-biased contribution to the gene pool is a critical aspect of demographic dynamics and is often impacted by variation in social structure between populations. We used the algorithm implemented in KIMTREE (Gautier and Vitalis 2013; Clemente et al. 2018) to estimate branch lengths from our SNP data set and infer the effective sex-ratios (ESR) for each of the five genetic clusters. This method is robust to LD, small sample sizes, and demographic events such as bottlenecks and expansions. To increase the number of usable markers, and since the authors recommend working with recently diverged populations, we focused on the recent northward colonization, and included individuals from the East Florida, Gulf Atlantic, and Carolinas genetic clusters.

Briefly, the method builds a hierarchical Bayesian model to estimate the evolution of SNP frequencies along branches of a population tree provided by the user. Genetic drift along branches is estimated by a time-dependent diffusion approximation. In this framework, branch length τ is proportional to the time since divergence in generations (t) scaled by the effective population size (N_e), such that $\tau \equiv t/2N_e$. The method can jointly contrast allele frequencies between autosomal and sex-linked markers to estimate the relative contribution of males and females to each generation (ESR). The ESR can then be seen as a comparison of the effective population sizes estimates obtained from autosomes and the X chromosome.

We sexed individuals by taking advantage of the expected relationship between depths of coverage at autosomal and

sex-linked loci in males and females. Since females are XX and males XY, the latter are expected to display a two-times lower coverage at X-linked sites compared with autosomal loci (supplementary fig. 1, Supplementary Material online). We then adjusted allele frequencies for all X-linked scaffolds, including Linkage Group b (Alföldi et al. 2011) and several scaffolds (GL343282, GL343364, GL343550, GL343423, GL343913, GL343947, GL343338, GL343417) recently identified as belonging to the green anole's sex chromosome (Rupp et al. 2017). We counted one haplotype per male and two per female. To obtain confidence intervals over ESR estimates, we generated 50 pseudoreplicated data sets by randomly sampling 5,000 autosomal and 5,000 sex-linked SNPs with no missing data. The algorithm was started with 25 pilot runs of 1,000 iterations each to adjust the parameters of the Monte Carlo Markov Chain (MCMC). The MCMC itself was run for 100,000 generations and sampled every 25 iterations after a burn-in of 50,000 iterations. Convergence for all parameters was assessed by visually inspecting posterior sampling in R (R Core team 2016). For each replicate *i*, we estimated the support for biased sex-ratio (S_i) such as:

$$S_i = 1 - 2\,|p_i - 0.5\,|.$$

with $S_i < 0.05$ being interpreted as a strong support for biased sex-ratio and where p_i is the proportion of posterior MCMC samples with an ESR > 0.5.

Model Comparison of Demographic Scenarios

None of the previous population genetics studies of green anoles has ever precisely quantified the strength nor the timing of gene flow between genetic clusters. We addressed this issue by comparing different demographic scenarios for two pairs of sister clades (EF and GA; EF and WF) that included the most individuals (at least 11).We used the diffusion approximation-based likelihood approach implemented in the ∂a∂i software (Gutenkunst et al. 2009). We compared a set of scenarios of strict isolation (SI), isolation with migration (IM), ancient migration (AM) with one or two (PAM) periods of gene flow and secondary contact (SC) with one or two (PSC) periods of gene flow (see Christe et al. 2017 for a detailed summary). We added complexity to this set of basic scenarios by allowing for a combination of population expansion (prefix "ex"), heterogeneous asymmetric migration rates (suffix "2M2P") and heterogeneous effective population size (suffix "2N") among loci. These additions were made to incorporate the genome-wide effects of selection on linked neutral sites (so-called "linked selection") and model genomic islands resisting gene flow (Cruickshank and Hahn 2014). We also tested scenarios with both asymmetric migration rates and heterogeneous population sizes but were unable to reach convergence. Overall, we compared 34 scenarios combining these features, using a set of scripts available on dryad (Christe et al. 2017) and a modified version of ∂a∂i (v1.7.0) kindly provided by Christelle Fraïsse (available at http://methodspopgen.com/wp-content/uploads/2017/12/dadi-1.7.0_modif.zip; Last accessed on June 7, 2019). We extracted for each pairwise comparison a set of ~12,000 SNPs with no missing data and thinned every 100,000 bp to meet the requirement of independence among loci that is needed to properly compare the composite likelihoods estimated by ∂a∂i. We extracted the unfolded joint sites frequency spectra (SFS) by polarizing alleles using *A. porcatus* and *A. allisoni* as references. We considered ancestral the allele found at a minimal frequency of 75% in those two individuals or found fixed in one of them if the other individual was missing. We note that the ∂a∂i models include a parameter (O) estimating the proportion of correctly polarized sites. We evaluated each model 30 times and retained the replicate with the highest likelihood for model comparison. Models were compared using the Akaike information criterion (AIC). For the best model, we calculated uncertainties over the estimated parameters using a nonparametric bootstrap procedure, creating 100 pseudo-observed data sets by resampling with replacement from the SFS. We used the procedure implemented in the dadi.Godambe.GIM_uncert() script to obtain a maximum-likelihood estimate of 95% confidence intervals (Coffman et al. 2016). ∂a∂i parameters are scaled by the ancestral population size N_{ref}. For the sake of comparison with SMC++ estimates, parameters were converted into demographic units by estimating the ancestral effective population size as the harmonic mean of the SMC++ estimates before splitting time for all pairs of populations.

Estimating Recombination Rates

We used the LDHat software (McVean et al. 2002) to estimate effective recombination rates ($\rho = 4 N_e r$ with r the recombination rate per generation and N the effective population size) along the green anole genome. This method has been successfully used to obtain recombination maps for data sets similar to ours in terms of sequencing depth and sample sizes (Auton et al. 2012). Unphased genotypes were converted into LDHat format using VCFtools (option –ldhat). Since LDHat assumes that samples are drawn from a panmictic population, we focused on the Eastern Florida clade for which sampling effort was the highest ($n = 8$ diploid individuals). We used precomputed likelihood lookup tables with an effective population mutation rate (θ) of 0.001, which was the closest from the θ value estimated from our data set ($\theta \sim 0.004$) and used the lkgen module to generate a table fitting the number of observed samples (16 chromosomes). Recombination rates were estimated over 500-kb windows with 100-kb overlaps using the Bayesian reversible MCMC scheme implemented in the interval module. The chain was run for 1,000,000 iterations and sampled every 5,000 iterations with a large block

penalty of 20 to avoid overfitting and minimize random noise. The first 100,000 generations were discarded as burn-in. Convergence under these parameters was confirmed by visually inspecting MCMC traces for a subset of windows. We averaged ρ estimates over nonoverlapping 100-kb windows, or over coding sequences for subsequent analyses.

Summary Statistics for Differentiation and LD

To assess whether the joint effects of selection and low recombination on diversity and differentiation, we computed two measures of divergence (F_{ST} and d_{XY}) over nonoverlapping 100-kb windows for the three divergent Floridian lineages. These lineages were chosen because of their relative demographic stability (see Results). We picked 100,000 bp to reduce spatial autocorrelation between statistics of adjacent windows, since no further substantial LD decay could be observed over this distance for the two populations with the largest sample sizes (supplementary fig. 2, Supplementary Material online), pairwise LD (measured as r^2) was computed using VCFtools (Danecek et al. 2011). Comparison between those two statistics for a given genomic region has been proposed as a way to disentangle the effects of gene flow and selection (Cruickshank and Hahn 2014). For the sake of simplicity, correlations between differentiation statistics and recombination were estimated using a Spearman's correlation test in R, although we note that measurements cannot be fully considered independent. As a sanity check, we computed the *ZZ* statistics (Rozas et al. 2001) to assess whether LDHat estimates of ρ were consistent with the genomic distribution of LD. This statistic contrasts LD between adjacent pairs of SNPs to LD calculated over all pairwise comparisons in a given window. High values are suggestive of increased intragenic recombination. We also computed the average frequency of polymorphic derived alleles (average DAF) in the EF cluster since it included the most individuals, using *A. porcatus* and *A. allisoni* to polarize alleles, and excluding sites with at least six individuals genotyped over eight. DAF has been recently used to estimate the effects of linked selection in humans, an excess of ancestral alleles being expected in regions under the influence of background selection (Pouyet et al. 2018).

Because ρ is the effective recombination rate and depends on the effective population size, it is directly correlated to local reduction of diversity due to linked selection. Thus, a low value of ρ may be observed in regions linked to selection even if r itself is not significantly different from the rest of the genome. To reduce this correlation, we calculated ρ/θ_π, with θ_π the nucleotide diversity computed in POPGENOME. Since θ_π is an estimator of $4N\mu$, with μ the mutation rate, this statistic represents the ratio between r and the mutation rate μ (see Wang et al. 2016 for an example).

We examined the average DAF across five quantiles of ρ/θ_π to assess whether lower rates of recombination were associated with changes in the frequency of derived alleles that may be due to linked selection. We also compared the average DAF between the genomic background and regions of high relative and low absolute divergence which may be candidates for stronger linked selection. These regions were defined as regions belonging both to the top 20% quantile for F_{ST} and the lowest 20% quantile for d_{XY}.

Results

Statistics for Whole Genome Resequencing

Twenty-seven green anoles (*A. carolinensis*) sampled across the species' range and covering the five genetic clusters identified in previous analyses (Tollis and Boissinot 2014; Manthey et al. 2016) were chosen for whole-genome resequencing. We also included two samples from the closely related species *A. porcatus* and *A. allisoni* as outgroups. Sequencing depth was between 7.22× and 16.74×, with an average depth of 11.45× (supplementary table S1, Supplementary Material online). About 74,920,333 variants with <40% missing data were retained after the first round of filtering (Materials and Methods).

Population Structure and Nucleotide Variation Reveal a Reduced Diversity in Northern Populations

We first assessed geographic structure across the distribution of *A. carolinensis*. To this end, we used >6,500 SNPs thinned every 10 kb and with <20% missing data and identified k = 5 as the most likely number of genetic clusters with DAPC (fig. 1*A* and supplementary fig. 3, Supplementary Material online). Three groups were identified within Florida, while individuals from the rest of North-America were assigned to two clusters. These groups were consistent with the clusters identified in previous genetic studies (Tollis et al. 2012; Tollis and Boissinot 2014; Manthey et al. 2016). Possible introgression from Carolinas was observed for two Gulf Atlantic individuals (fig. 1*A*). A maximum-likelihood phylogeny estimated in RAxML based on one million random SNPs and a network analysis of relatedness in Splitstree further supported this clustering (fig. 1*B* and *C*). Results closely matched previous findings, with South Florida (SF) being the sister clade of all other groups. The two northernmost clusters, Gulf Atlantic (GA) and Carolinas (CA), clustered together in the RaxML phylogeny. Eastern Florida (EF) constituted a paraphyletic group in the phylogeny in which GA and CA were nested. This is likely due to incomplete lineage sorting induced by the high and constant effective population sizes of populations from Florida (see below), or to ongoing or recent gene flow resulting in the inclusion of loci with different coalescence times. At last, the Western Florida (WF) cluster was basal to all other groups except South Florida (SF).

Nucleotide diversity was the lowest in GA and CA (table 1) despite the large geographic area covered by these two

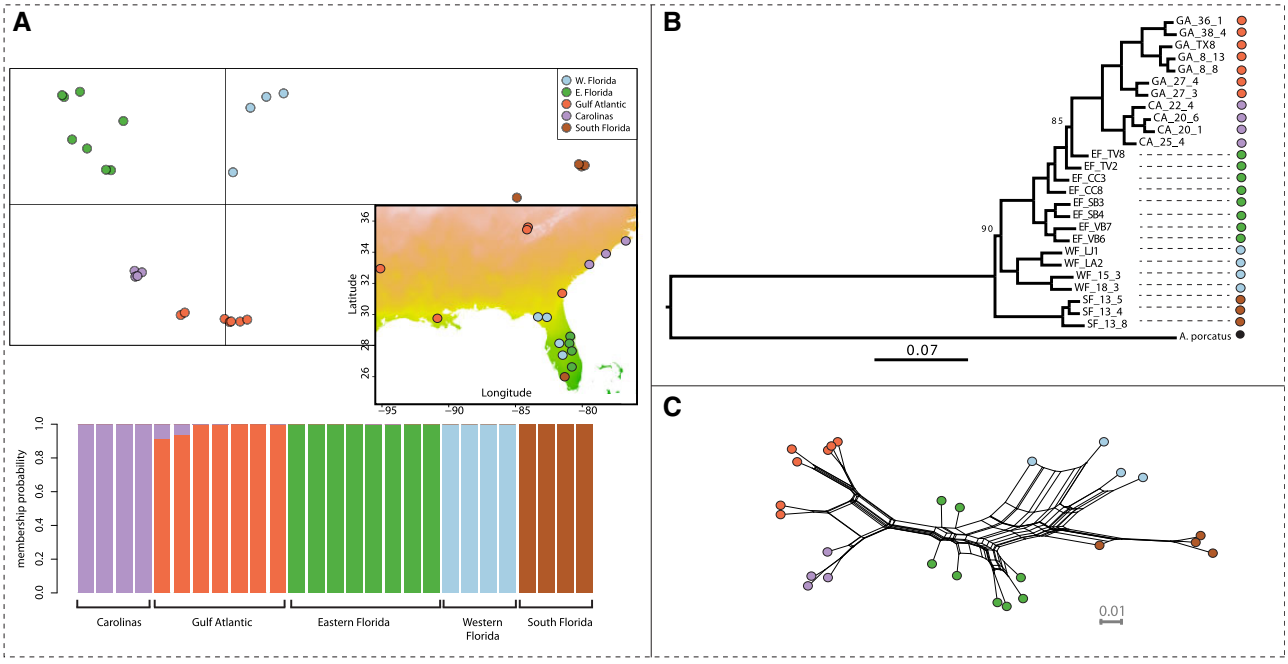


Fig. 1.—Genetic structure in *Anolis carolinensis* from whole-genome SNP data. (*A*) Results from the DAPC analysis highlighting the five clusters inferred from the analysis of ~6,500 SNPs thinned every 10 kb and with <20% missing data. The map reports the coordinates of the localities used in this study and the genetic clusters they belong to. (*B*) RAxML phylogeny based on one million SNPs randomly sampled across the genome. All 100 bootstrap replicates supported the reported topology, except for two nodes with support of 90 and 85. One individual from South Florida was removed due to a high rate of missing data. (*C*) Network representation of the relatedness between samples as inferred by Splitstree v4. Color codes match those in parts (*A*) and (*B*).

Table 1

Diversity and Tajima's *D* (±SD) for Each of the Five Genetic Clusters, Averaged over Nonoverlapping 5-kb Windows across the Genome

Statistics	CA	GA	EF	WF	SF
Nucleotide diversity	0.00155±0.00154	0.00177±0.00153	0.00330±0.0021	0.00341±0.0022	0.00279±0.002
Tajima's *D*	−0.17±1.49	0.14±0.0015	−0.73±0.002	−0.80±0.0022	−0.66±0.002

genetic clusters. Average Tajima's *D* values ranged between −0.8 (WF) and 0.14 (GA). Positive Tajima's *D* values suggest recent population contraction, while negative Tajima's *D* are expected in the case of recent population expansion (Tajima 1989), although both population substructure and linked selection may impact it. Northern clusters (CA and GA) displayed the highest average Tajima's *D*, possibly due to reductions in population sizes and relaxation of linked selection (see below).

Recent Population Expansion and Male-Biased Sex-Ratios in Northern Populations

We then reconstructed the demographic history of each genetic cluster. To this end, we used the whole set of filtered SNPs with <40% missing data to infer past changes in effective population sizes (N_e) without any a priori demographic model with SMC++ (fig. 2*A*). All populations from Florida showed rather stable demographic trajectories, with some evidence for population expansion in EF and WF. Assuming a mutation rate of 2.1×10^{-10}/bp per year (Tollis and Boissinot 2014), present population sizes were in the range of 500,000 to 5,000,000 individuals for each population, in accordance with previous analyses based on target capture markers (Manthey et al. 2016). Northern populations (CA and GA) showed a clear signature of expansion starting between 200,000 and 100,000 years ago, following a bottleneck that started between 500,000 and 1,000,000 years in the past. We also estimated the splitting times between the different groups but since this model assumes no gene flow after the split, the estimates are likely to be biased toward the present. The split between GA and CA occurred shortly before these populations expanded, in accordance with the previously proposed hypothesis of double colonization following the Gulf and Atlantic coasts (Tollis and Boissinot 2014). In Florida, divergence events took place between 3 and 2 Ma.

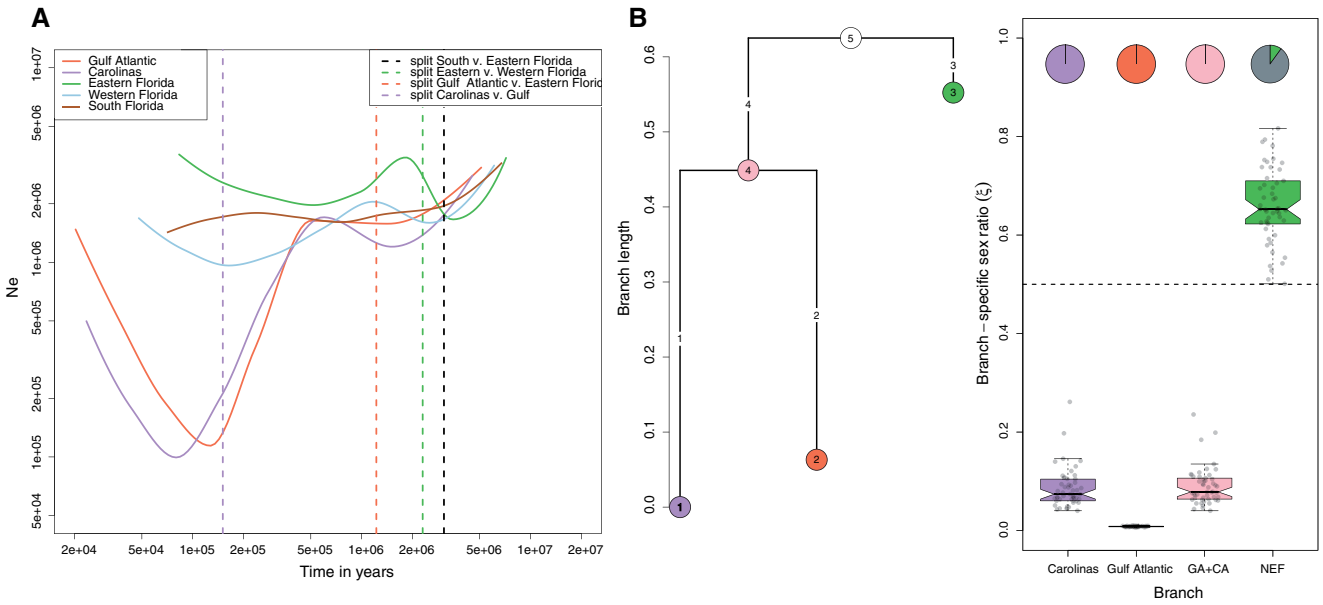


Fig. 2.—Variation in effective population sizes with time and comparison of drift between autosomes and sex-linked scaffolds. (*A*) Reconstruction of past variations in effective population sizes (N_e) inferred by SMC++. Dashed vertical lines correspond to the estimated splitting times between the five genetic clusters previously inferred. We assume a mutation rate of 2.1×10^{-10}/bp per generation and a generation time of 1 year. (*B*) Average branch lengths obtained from autosomal data and ESRs (ξ) inferred from KIMTREE. A set of 5,000 autosomal and 5,000 sex-linked markers were randomly sampled to create 50 pseudoreplicated data sets on which the analysis was run. The analysis was run on the three most closely related populations. Pie charts indicate the proportion of replicates for which we observed significant support (S_i<0.01) in favor of a biased sex-ratio.

The relative order of splitting events was consistent with the topology obtained from our phylogeny and previous studies.

Since anoles exhibit sex differences in dispersal (Johansson et al. 2008), we tested whether the recent colonization of new and possibly suboptimal habitats could lead to unequal contribution of males and females to the gene pool or selection at sex-linked loci (fig. 2*B*). We built a population tree and quantified genetic signatures of biased sex-ratio with the algorithm implemented in KIMTREE. We focused on the three populations that diverged most recently, GA, CA, and EF. Note that the length of branch i (τ_i) represents time in generations (t_i) scaled by the effective population size for this branch such as $\tau_i = t_i/2N_{e,i}$ (Clemente et al. 2018). Branch lengths were particularly high for the CA and GA lineages compared with EF, as expected in the case of stronger drift (fig. 2*B*). This is in line with their smaller effective population sizes and the bottleneck inferred by SMC++. We found evidence for a strongly male-biased ESR in CA and GA, but not in EF which was slightly female-biased. Indeed, nucleotide diversity was substantially more reduced at sex-linked scaffolds in GA than in EF when compared with autosomal diversity (supplementary fig. 4, Supplementary Material online). Note that sex-ratios are the proportion of females effectively contributing to the gene pool along each branch of the tree and should not be interpreted directly in terms of census size. The GA cluster displayed the strongest bias, with an estimated ratio of less than one female for 100 males, suggesting strong sex-bias in the founding population or strong male-biased dispersal during population expansion. The CA cluster and the inner branch leading to CA and GA showed a ratio of approximately 10 females for 100 males. All 50 replicates displayed a high support for a male-biased sex-ratio in CA and GA, while only five replicates supported a female-biased sex-ratio in EF (i.e., the Markov chain almost systematically explored sex-ratios >0.5 in only five replicates).

Gene Flow at SC Has Homogenized Green Anole Populations

We tested whether gene flow and its interruption may have played a role in shaping the genomic landscape of differentiation in green anoles (fig. 3). We focused on two pairs of genetic clusters. The first comparison was between the EF and WF clusters, which are two populations with high and stable population sizes (according to SMC++), that both live in subtropical Florida. This comparison should be suited to detect the long-term effects of interrupted gene flow since there is already evidence that these clades may have been isolated by sea rising during interglacial periods (Tollis and Boissinot 2014; Manthey et al. 2016). In addition, since populations from Florida have remained relatively stable over the last 2 Myr, our power to detect the expected correlations between recombination and diversity should be enhanced in the case of linked selection (Burri 2017; Torres et al.

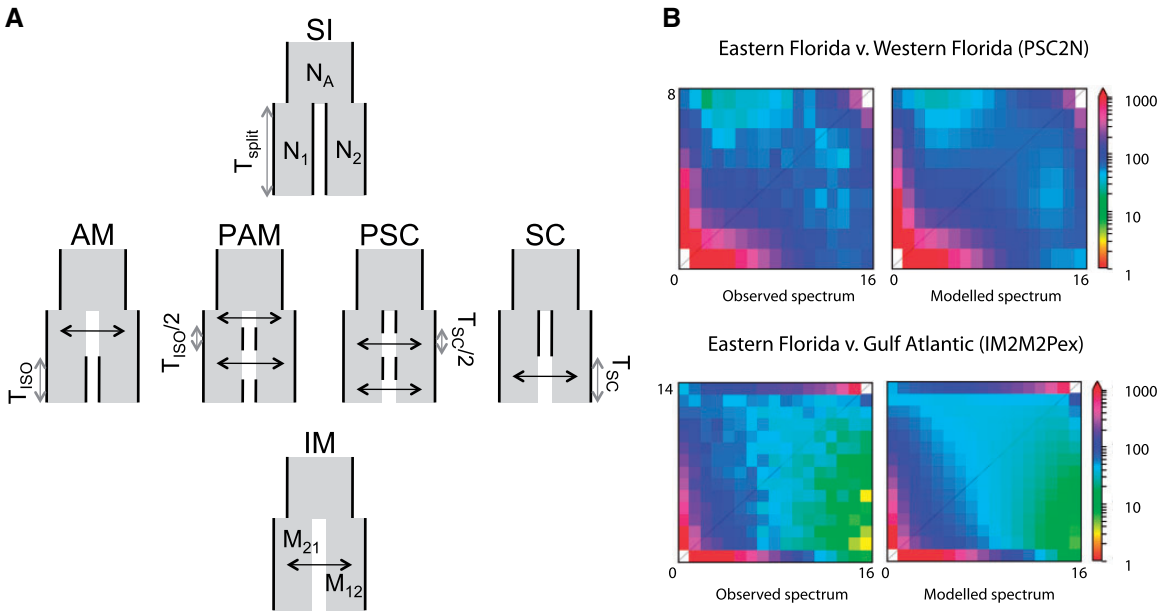


Fig. 3.—(*A*) Graphic description of the six categories of ∂a∂i models tested over pairs of green anole genetic clusters. Each model describes a scenario where two populations diverge from an ancestral one, with varying timing and strength of gene flow after their split. SI, strict isolation; AM, ancestral migration, where populations first exchange gene flow then stops T_{iso} generations ago; PAM, ancestral migration with two periods of contact lasting $T_{iso}/2$ generations; SC, secondary contact where populations still exchange gene flow at present time; PSC, secondary contact with two periods of contact lasting $T_{sc}/2$ generations; IM, isolation with constant migration and no interruption of gene flow. Models were constrained so that T_{iso} and T_{sc} lasted at least ~50,000 years. Reproduced with the authorization of Christelle Fraïsse. (*B*) Fitting of the best models for the EF ($N = 16$) versus GA ($N = 14$) and EF versus WF ($N = 8$) comparisons. Both models fit the observed data sets as indicated by the similar spectra between observation and simulation. The "2 *N*" suffix means that background selection was added to the base model by modeling heterogeneous effective population sizes across loci. The "2M2P" suffix means that heterogeneity in gene flow was incorporated into the model. The "ex" suffix means that exponential population size change was introduced in the base model.

2019). The second comparison was between the EF and GA clusters, the latter corresponding to a recent expansion northward and adaptation to temperate environments.

We ran ∂a∂i models to assess the effects of differential gene flow and linked selection. We compared a set of 34 divergence scenarios, allowing gene flow and effective population sizes to vary with time and across loci. Briefly, heterogeneity in gene flow (suffix 2M2P) was implemented by dividing the site frequency spectrum into three sets of loci with proportions $1\text{-}P_1\text{-}P_2$, P_1, and P_2. Assuming that the first population is WF and the second EF, the first set of loci ($1\text{-}P_{WF}\text{-}P_{EF}$) is modeled with all parameters from the base model. The two other sets are modeled with no gene flow toward WF (P_{WF}) or EF (P_{EF}) and represent genomic islands resisting gene flow in WF and EF, respectively. To simulate the reduction in diversity expected under purifying selection at linked, nonrecombinant (*nr*) sites, two sets of loci were modeled at frequencies $1-nr$ and nr (suffix $2N$). The first set was modeled with all parameters from the base model, the other with the same parameters but with effective population sizes reduced by a background selection factor (*bf*).

Strict-isolation models (SI) consistently displayed the lowest likelihood and highest AIC, clearly supporting a role for gene flow in homogenizing green anoles genomes. For the comparison between EF and WF, models including heterogeneous population sizes performed better than models with heterogeneous gene flow. Among scenarios with gene flow, SC with one and two periods of gene flow (SC and PSC) often received the highest support (fig. 3*B* and supplementary fig. 5, Supplementary Material online). Parameters estimated from the best models are shown in table 2. There was no substantial gain in likelihood when adding expansion to scenario of two SCs with background selection (PSC2N), and models with heterogeneous migration displayed lower likelihood. The PSC2N model supported a scenario where about $nr = 65\%$ of the genome was affected by selection at linked sites, suggesting a rather large effect of low recombination and purifying selection on diversity. These Eastern and Western Floridian genetic clusters experienced long periods of isolation lasting ~2 Myr, followed by periods of SC lasting ~125,000 years in total.

For the comparison between GA and EF, we confirmed the decrease in effective population size detected by SMC++ in GA compared with EF, with a present effective population

Table 2
Summary of Best-Supported Demographic Models

Comparison	Model	Na12	Na1	Na2	N1	N2	m2->1	m1->2	Tiso	Tsc	Tscg	P_1	P_2	*nr*	*bf*	O	logLikelihood	AIC
GA vs. EF	IM2M2Pex	2,156,641	3,538,759	2,438,866	371,048	7,132,328	2.42E-07	2.57E-07	NA	**521,532***	1,615,603	0.83	0.95	NA	NA	0.97	−1,045.99	2,113.97
GA vs. EF	IMex	2,156,641	5,971,380	2,137,291	364,329	6,755,470	1.93E-07	2.36E-07	NA	**61,217****	1,962,652	NA	NA	NA	NA	0.97	−1,048.14	2,114.29
GA vs. EF	SC2M2Pex	2,156,641	4,214,885	2,296,060	367,662	7,049,892	2.12E-07	2.58E-07	**139,258****	**229,611***	1,723,615	0.91	0.95	NA	NA	0.97	−1,046.01	2,116.01
GA vs. EF	PSCex	2,156,641	4,126,463	2,402,541	367,459	6,926,709	1.93E-07	2.35E-07	**58,466****	110,560	1,713,849	NA	NA	NA	NA	0.97	−1,048.03	2,116.06
GA vs. EF	SCex	2,156,641	4,150,102	2,359,593	369,470	6,934,978	1.91E-07	2.36E-07	**97,752****	**249,043***	1,719,867	NA	NA	NA	NA	0.97	−1,048.05	2,116.11
GA vs. EF	PSC2M2Pex	2,156,641	3,873,718	2,394,203	370,665	7,005,838	2.04E-07	2.58E-07	**79,128****	**124,565***	1,672,998	0.93	0.95	NA	NA	0.97	−1,046.12	2,116.24
GA vs. EF	IM2Nex	**2,156,641***	**3,106,584***	2,766,204	360,422	6,961,498	1.99E-07	2.34E-07	NA	615,678	1,461,807	NA	NA	**0.60****	**0.81***	0.97	−1,048.09	2,118.17
EF vs. WF	PSC2N	2,091,300	NA	NA	5,181,876	5,389,218	1.94E-06	6.93E-07	1,064,768	759,07	NA	NA	NA	0.60	0.25	0.98	−669.91	1,357.81
EF vs. WF	SC2N	2,091,305	NA	NA	**5,252,609***	**5,330,317***	1.18E-06	4.25E-07	1,952,690	**144,821***	NA	NA	NA	0.57	0.26	0.98	−676.66	1,371.33

NOTE.—PSC2N, secondary contact with two periods in isolation and heterogeneous effective population sizes across the genome; SCex, secondary contact with an episode of population expansion following secondary contact; SC2M2P and IM2M2P, models of secondary contact and constant gene flow with heterogeneous migration rates along the genome. The ancestral size (N_{a12}) before the split was calculated from the SMC++ output to facilitate comparisons. For the -ex models, following the initial split, populations have a population size of N_{1a} and N_{2a} followed by an exponential growth that leads to their current sizes N_1 and N_2. In basic models, populations have constant sizes N_1 and N_2 since the split. *nr*, proportion of the genome displaying an effective population size of *bf* times the population size displayed by the remaining 1−*nr* fraction not affected by linked selection; *O*, proportion of sites for which the ancestral state was correctly inferred; P_1 and P_2, the proportion of sites resisting gene flow in populations 1 and 2; T_{iso}, total time spent in isolation. For the PSC model, populations are isolated twice in their history for $T_{isc}/2$ generations and are connected twice for $T_{sc}/2$ generations (see fig. 3A). T_{iso}, time during which stable populations stay connected; T_{scg}, time since population size change (with gene flow). The total time during which populations were connected is $T_{scg}+T_{sc}$. For each model, the set of best estimates is shown. Uncertainties over parameters were measured by SDs obtained from 100 bootstrap replicates. No star, uncertainties are below ±20% of point estimates; *, uncertainties between ±20% and ±50%; **, uncertainties between ±50% and ±150%. Cells with "NA" values correspond to parameters that were not part of a given model.

size 20 times lower in GA than in EF. We note that unlike SMC++, ∂a∂i was not able to detect the recent rebound in size following the bottleneck, even after allowing for more past changes in effective population sizes (data not shown). The model with the smallest AIC was the IM2M2Pex model, followed by models of SC (PSCex, SCex, and SC2M2Pex). We therefore present results obtained for several representative models (table 2). All models supported a scenario with extensive gene flow, with high uncertainties for the time spent in isolation for SC models. Models with the highest likelihood and lowest AIC incorporated genomic barriers to gene flow in GA, with ∼10–20% of loci resisting introgression from Florida and ∼5% resisting gene flow from GA.

Recombination and Linked Selection Shape Genome Differentiation and Diversity

It has been suggested that SCs can lead to the emergence of genomic islands resisting gene flow, that display higher differentiation than regions that have been homogenized (Payseur et al. 2004; Teeter et al. 2008; Larson et al. 2014; Malinsky et al. 2015; McGee et al. 2016). The diversity of such islands may also be higher, as they diverged and accumulated mutations before gene flow resumed. On the other hand, selection at linked sites can also generate genomic islands, as it reduces diversity and lead to an increase of relative measures of differentiation (Noor and Bennett 2009; Cruickshank and Hahn 2014; Burri et al. 2015). Some of the best supported models in ∂a∂i suggested a widespread impact of selection in Florida, reducing diversity at linked sites over ∼60% of the genome. We therefore tested the role of low recombination in shaping the genomic landscape of diversity and differentiation in green anoles in a context of SC.

Recombination rates estimated by LDHat in the EF cluster were highly heterogeneous along chromosomes, with stronger recombination rates at the tips and toward centromeres, though they dropped at the immediate vicinity of the latter (fig. 4). This pattern was supported by the Rozas's *ZZ* statistic, suggesting stronger LD in the middle of chromosomes arms (supplementary fig. 6, Supplementary Material online). We observed higher relative differentiation (measured by F_{ST}) in regions of low recombination (Spearman's rank correlation test, all *P* values $<2.2\times10^{-16}$; fig. 4 and supplementary fig. 7, Supplementary Material online). The correlation was however opposite for measures of absolute differentiation (d_{XY}), a statistics directly related to diversity and to the average age of alleles across populations (Cruickshank and Hahn 2014). These correlations are consistent with selection reducing heterozygosity in regions of low recombination, and further support the ∂a∂i models of heterogeneous effective population sizes along the genome. These measures of differentiation were strongly correlated when examining all three pairwise comparisons within Florida (fig. 4 and supplementary fig. 8, Supplementary Material online). There was also a positive

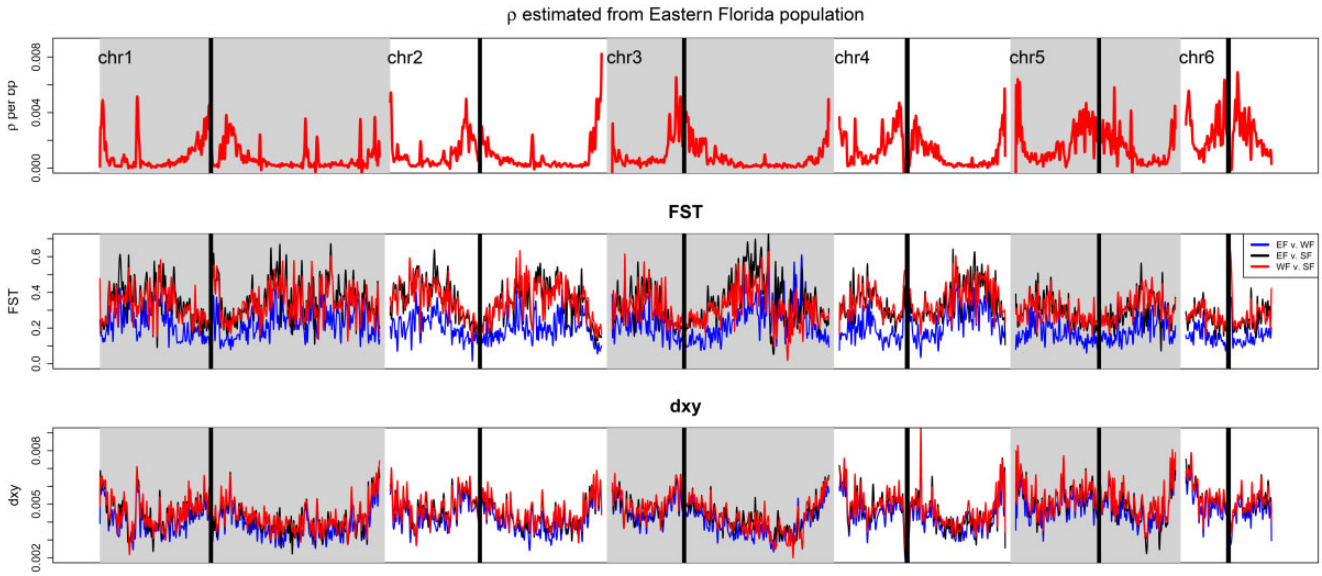


Fig. 4.—Summary statistics for recombination and differentiation along chromosomes. $\rho = 4 \times N_e \times r$, with r the recombination rate per bp and per generation and N_e the effective population size for the EF cluster. F_{ST} and d_{XY} are relative and absolute measures of differentiation that are correlated with the amount of shared heterozygosity and coalescence time across populations, respectively. We present differentiation for the three genetic clusters having diverged for the longest time period. Statistics were averaged over nonoverlapping 5-kb windows and a smoothing line was fit to facilitate visual comparison. Repetitive centromeric regions that are masked from the green anole genome are highlighted by black rectangles.

correlation between recombination rate and the average frequency of derived alleles (DAF) in the EF cluster (Spearman's rank correlation test, $\rho = 0.11$, P value < 0.001), although we did observe an increase in the average DAF for very low recombination rates that may be due to linked positive selection (supplementary fig. 9, Supplementary Material online). We did not observe any significant shift toward an excess of derived variants in regions of high divergence and low diversity (supplementary fig. 10, Supplementary Material online).

Given the possible effects of linked selection on diversity, we performed a SMC++ analysis masking 100-kb regions in the top 33% or low 33% ρ/θ_π. We did not notice any strong deviation in the relative timing of divergence. The estimated effective population sizes remained in the same range of 1,000,000–5,000,000 individuals for each population in the last 3 Myr when compared with the unmasked analysis. Events more recent than ~500 kya could not be reliably inferred due to masking. The strongest changes were observed for splitting times within Florida, for which estimates were ~500,000 years younger with masking, with divergence between South Florida and other clades starting ~2.3 Ma instead of 3.0 Ma.

Discussion

In this study, we investigated the processes responsible for genomic differentiation and variation in the green anole. We showed that a complex history of recent expansion and SCs associated with linked selection shaped the genomic landscape of differentiation and diversity. Our results provide an important assessment of the forces acting on the green anole genome.

A Dynamic Demographic History Has Shaped the Genomic Landscape of Differentiation

Green anole populations are strongly structured and it was hypothesized that successive splits and SC occurred in Florida during the Pleistocene (Tollis and Boissinot 2014; Manthey et al. 2016). Fluctuations in sea level may have generated temporary islands on which isolated populations could have diverged. At last, reconnection of Florida to the mainland would have provided the opportunity for expansion northward (Soltis et al. 2006). Our results support this claim in three ways. First, splitting times estimated by SMC++ and ∂a∂i suggest a series of splits in Florida between 3 and 2 Ma, a time range during which successions of glacial and interglacial periods may have led to several vicariance events (Lane 1994; Petuch 2004). Second, the models receiving the highest support in ∂a∂i were the ones allowing for several events of isolation followed by SC in Florida. Third, we found clear signatures of population expansion in GA and CA at the beginning of the Late Pleistocene, a time when lowering sea levels would have facilitated colonization (Lane 1994; Petuch 2004). We acknowledge that the exact timing of these events depends on the mutation rate used, which was previously established based on rates of divergence for three intronic nuclear markers compared with a mitochondrial one (Tollis and Boissinot 2014). The mutation rate used here is in the lower range of what is expected for nuclear markers

compared with the mitochondrial rate of 1.3×10^{-8} per year commonly used in lizards (Macey et al. 1999), being ~60 times lower while the average for squamates is ~26 (Allio et al. 2017). Despite an old history of divergence, our ∂a∂i analysis found clear evidence for gene flow between taxa having diverged in the last 2 Myr. We note for example that ∂a∂i models with the highest likelihoods for the Gulf Atlantic-Eastern Florida comparison included heterogeneous migration rates along the genome, and suggested barriers to gene flow limiting introgression from Florida. This could reflect local adaptation through reduced effective migration rates at loci under selection in northern latitudes (but see Bierne et al. 2011).

While linked selection seems to play a major role in populations from Florida (see below), our results do not preclude the existence of heterogeneous gene flow along the genome, since we could not properly test the likelihood of models incorporating both of these aspects at once. Instead, they highlight the important role of linked selection in producing heterogeneous landscapes of differentiation (Cruickshank and Hahn 2014), even in a context of SC where genomic islands resisting gene flow may be expected. Recent years have seen a growing interest for the so-called "genomic islands of speciation," regions that harbor higher differentiation than the genomic background (Feder and Nosil 2010; Ellegren et al. 2012; Nadeau et al. 2012; Wolf and Ellegren 2017). Several studies have since successfully highlighted the important role of heterogeneous migration and selection in shaping diversity in several organisms, such as mussels (Roux et al. 2014), sea bass (Tine et al. 2014), or poplars (Wang et al. 2016; Christe et al. 2017). This area of research has however been neglected so far in squamates, preventing any comparison of their genome dynamics at microevolutionary scales with other vertebrates. Our results call for more studies, for example using transects encompassing contact zones (Barton and Hewitt 1985), to assess how alleles diffuse across genetic clusters and better assess how heterogeneous gene flow may interact with recombination and linked selection to shape differentiation landscapes, which remains a challenging question (Nachman and Payseur 2012).

Unequal Diversity between the X and Autosomal Chromosomes

We detected a significant deviation from a balanced ESR in the two populations that recently expanded and colonized North America, with strongly reduced nucleotide diversity on the X chromosome in the Gulf Atlantic population when compared with autosomal diversity. The method we used takes into account the expected difference in effective population sizes between autosomes and sex chromosomes. This suggests that the number of females that contributed to the present diversity on the X chromosome may have been extremely reduced compared with the number of males. Since this signature was found only in expanding populations, a possible explanation would be that the colonization of suboptimal habitats (compared with the center of origin in Florida) favored male-biased dispersal. The limited number of available females in the newly colonized regions would have therefore led to a biased sex-ratio in the founding populations and smaller effective population sizes on the X chromosome compared with unbiased expectations.

In *Anolis roquet*, male-biased dispersal is associated with competition, since males disperse more when density increases and competition for females is stronger (Johansson et al. 2008). In *Anolis sagrei*, smaller males tend to disperse more while females are more likely to stay in high quality territories, independently of female density (Calsbeek 2009). The green anole is a polygynous species, with sexual dimorphism and high levels of competition between males (Jenssen et al. 2000). It is therefore likely that competition within sexes may lead to unequal contribution of males and females to the gene pool.

Another nonexclusive possibility lies in the action of positive selection on the X chromosome in northern populations. The X chromosome is extremely small compared with autosomes in green anoles, probably not exceeding 20 Mb (Rupp et al. 2017). This means that even a few recent selective sweeps would have widespread effects on the entire chromosome, reducing diversity and the effective population size. Since the method implemented in KIMTREE compares estimates of effective population sizes between autosomes and X chromosome, this would result in an artificially biased sex-ratio. Sexual or natural selection may be responsible for this pattern, and our finding calls for further comparisons of sex-biased dispersal and behavior between populations of the green anole.

Selection and Recombination Shape Nucleotide Composition and Diversity at Linked Sites

We observed strong heterogeneity in recombination rates along the green anole genome. Our results show that this heterogenous recombination landscape plays an important role in shaping genetic diversity in anoles. Both purifying and positive selections are expected to reduce diversity and increase genetic differentiation at linked sites (Cruickshank and Hahn 2014). We found evidence for an effect of linked selection in shaping differentiation between genetic clusters in Florida. These clusters have had relatively stable effective population sizes over the last 2 Myr, which should limit the stochasticity induced by drift in explaining heterogeneous patterns of differentiation along the genome. In that case, signatures of linked selection such as high differentiation and low diversity should be easier to detect in regions of low recombination. Indeed, regions of high diversity that are characterized by high d_{XY} displayed higher recombination rates and lower F_{ST} in the green anole, while regions with high

F_{ST} were found in regions of low recombination and diversity. More specifically, a prominent role for background selection in shaping the differentiation landscape is supported by the correlations we observed between diversity, differentiation, and recombination. We acknowledge that positive selection may lead to similar signatures of high differentiation and low diversity in regions of low recombination due to the effects of selection at linked sites (Josephs and Wright 2016), and we did observe an increase in the average frequency of derived alleles in regions of very low recombination (supplementary fig. 9, Supplementary Material online). However, regions of both high divergence and low diversity did not seem to harbor any excess of derived alleles (supplementary fig. 10, Supplementary Material online) which suggests that positive selection is not the main driver of the differentiation landscape within Florida. Moreover, the agreement between landscapes of differentiation for all pairwise comparisons (fig. 4) suggests a reduction in diversity in the ancestral population rather than population-specific events of selection, consistent with background selection, or positive selection in the lineage ancestral to all green anoles.

Since green anoles constitute an important model species to understand the mechanisms of adaptation, adopting a cautious (i.e., nonadaptationist) interpretation of divergence landscapes is primordial (Burri 2017). Disentangling recent positive selection from the confounding effects of demography and background selection is especially challenging, even in species for which extensive genomic and functional studies have been performed, as in humans. Genome scans for positive selection have often failed to identify common outliers (Pavlidis et al. 2012), and require to carefully consider the demographic history of populations (Li et al. 2012; Pavlidis et al. 2012; Elyashiv et al. 2016; Schrider and Kern 2016; Sheehan and Song 2016; Bourgeois et al. 2018). The green anole is an important system to understand local adaptation in reptiles (Campbell-Staton et al. 2017) and the incorporation of our findings in future studies will be useful to properly test for signals of local adaptation. This will be done by taking into account the possible biases induced by demography and the impact of selection at linked sites. Further studies of positive selection will require more detailed analyses, building on the results we show in the present study.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank three anonymous reviewers for their constructive comments that improved an earlier version of this article. We are grateful to Breda Zimkus from the Museum of Comparative Zoology Cryogenic Collection in Harvard and J. Rosado from the Herpetology Collection for providing the samples of *A. porcatus* and *A. allisoni*. We also thank Christelle Fraïsse for providing tutorials and the modified version of ∂a∂i that was needed to compare demographic models. We thank Justin Wilcox for his comments on the article. We thank Marc Arnoux from the Genome Core Facility at NYUAD for assistance with genome sequencing. This research was carried out on the High-Performance Computing resources at New York University Abu Dhabi. This work was supported by New York University Abu Dhabi (NYUAD) research funds AD180 (to S.B.). The NYUAD Sequencing Core is supported by NYUAD Research Institute grant G1205-1205A to the NYUAD Center for Genomics and Systems Biology.

Literature Cited

Alföldi J, et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477(7366):587–591.

Allio R, Donega S, Galtier N, Nabholz B. 2017. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. Mol Biol Evol. 34(11):2762–2772.

Auton A, et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. Science 336(6078):193–198.

Barrett RDH, Rogers SM, Schluter D. 2008. Natural selection on a major armor gene in threespine stickleback. Science 322(5899):255–257.

Barton NH, Hewitt GM. 1985. Analysis of hybrid zones. Annu Rev Ecol Syst. 16(1):113–148.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356(6369):519–520.

Bierne N, Welch J, Loire E, Bonhomme F, David P. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. Mol Ecol. 20(10):2044–2072.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Booker TR, Ness RW, Keightley PD. 2017. The recombination landscape in wild house mice inferred using population genomic data. Genetics 207(1):297–309.

Bourgeois Y, et al. 2018. Genome-wide scans of selection highlight the impact of biotic and abiotic constraints in natural populations of the model grass *Brachypodium distachyon*. Plant J. 96(2):438–451.

Burri R. 2017. Interpreting differentiation landscapes in the light of long-term linked selection. Evol Lett. 1(3):118–131.

Burri R, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. Genome Res. 25(11):1656–1665.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. PloS Genet 5(1):e1000336.

Calsbeek R. 2009. Sex-specific adult dispersal and its selective consequences in the brown anole, *Anolis sagrei*. J Anim Ecol. 78(3):617–624.

Campbell-Staton SC, Bare A, Losos JB, Edwards SV, Cheviron ZA. 2018. Physiological and regulatory underpinnings of geographic variation in reptilian cold tolerance across a latitudinal cline. Mol Ecol. 27(9):2243–2255.

Campbell-Staton SC, Edwards SV, Losos JB. 2016. Climate-mediated adaptation after mainland colonization of an ancestrally subtropical island lizard, *Anolis carolinensis*. J Evol Biol. 29(11):2168–2180.

Campbell-Staton SC, et al. 2012. Out of Florida: mtDNA reveals patterns of migration and pleistocene range expansion of the green anole lizard (*Anolis carolinensis*). Ecol Evol. 2(9):2274–2284.
Campbell-Staton SC, et al. 2017. Winter storms drive rapid phenotypic, regulatory, and genomic shifts in the green anole lizard. Science 357(6350):495–498.
Christe C, et al. 2017. Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. Mol Ecol. 26(1):59–76.
Clemente F, Gautier M, Vitalis R. 2018. Inferring sex-specific demographic history from SNP data. PLoS Genet. 14:1–32.
Coffman AJ, Hsieh PH, Gravel S, Gutenkunst RN. 2016. Computationally efficient composite likelihood statistics for demographic inference. Mol Biol Evol. 33(2):591–593.
Costantini M, Greif G, Alvarez-Valin F, Bernardi G. 2016. The *Anolis* lizard genome: an amniote genome without isochores? Genome Biol Evol. 8(4):1048–1055.
Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 23(13):3133–3157.
Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.
Depristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491–501.
Ellegren H, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491(7426):756–760.
Elyashiv E, et al. 2016. A genomic map of the effects of linked selection in Drosophila. PLoS Genet. 12:1–24.
Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PloS Genet 9(10): e1003905.
Feder JL, Nosil P. 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution 64(6):1729–1747.
Figuet E, Ballenghien M, Romiguier J, Galtier N. 2015. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biol Evol. 7(1):240–250.
Fujita MK, Edwards SV, Ponting CP. 2011. The *Anolis* lizard genome: an amniote genome without isochores. Genome Biol Evol. 3:974–984.
Gautier M, Vitalis R. 2013. Inferring population histories using genome-wide allele frequency data. Mol Biol Evol. 30(3):654–668.
Glor RE, Losos JB, Larson A. 2005. Out of Cuba: overwater dispersal and speciation among lizards in the *Anolis carolinensis* subgroup. Mol Ecol. 14(8):2419–2432.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PloS Genet 5(10): e1000695.
Han F, et al. 2017. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Res. 27(6):1004–1015.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23(2):254–267.
Jensen JD, et al. 2019. The importance of the Neutral Theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution 73(1):111–114.
Jenssen T, Orrell KS, Lovern MB, Ross S. 2000. Sexual dimorphisms in aggressive signal structure and use by a polygynous lizard, *Anolis carolinensis*. Copeia 2000(1):140–149.
Johansson H, Surget-Groba Y, Thorpe RS. 2008. Microsatellite data show evidence for male-biased dispersal in the Caribbean lizard *Anolis roquet*. Mol Ecol. 17(20):4425–4432.
Jombart T, et al. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11(1):94.
Josephs EB, Wright SI. 2016. On the trail of linked selection. PLoS Genet. 12:1–5.
Kawakami T, et al. 2017. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Mol Ecol. 26(16):4158–4172.
Kern AD, Hahn MW. 2018. The neutral theory in light of natural selection. Mol Biol Evol. 35(6):1366–1371.
Kolbe JJ, et al. 2017. An incipient invasion of brown anole lizards (*Anolis sagrei*) into their own native range in the Cayman Islands: a case of cryptic back-introduction. Biol Invasions. 19(7):1989–1998.
Lailvaux SP, Herrel A, Vanhooydonck B, Meyers JJ, Irschick DJ. 2004. Performance capacity, fighting tactics and the evolution of life-stage male morphs in the green anole lizard (*Anolis carolinensis*). Proc Biol Sci. 271(1556):2501–2508.
Lane, E. 1994. Florida's Geological History and Geological Resources. Special publication (Florida Geological Survey (1989)). Vol. 35. Published for the Florida Geological Survey; Tallahassee, FL.
Lapiedra O, Schoener TW, Leal M, Losos JB, Kolbe JJ. 2018. Predator-driven natural selection on risk-taking behavior in anole lizards. Science 360(6392):1017–1020.
Larson EL, White TA, Ross CL, Harrison RG. 2014. Gene flow and the maintenance of species boundaries. Mol Ecol. 23(7):1668–1678.
Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475(7357):493–496.
Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079.
Li J, et al. 2012. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol Ecol. 21(1):28–44.
Losos JB. 2009. Lizards in an evolutionary tree. Berkeley: University of California Press [Database]
Losos JB, Schoener TW, Spiller DA. 2004. Predator-induced behaviour shifts and natural selection in field-experimental lizard populations. Nature 432(7016):505–508.
Macey JR, et al. 1999. Molecular phylogenetics, tRNA evolution, and historical biogeography in Anguid lizards and related taxonomic families. Mol Phylogenet Evol. 12(3):250–272.
Malinsky M, et al. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science 350(6267):1493–1498.
Manthey JD, Tollis M, Lemmon AR, Moriarty Lemmon E, Boissinot S. 2016. Diversification in wild populations of the model organism *Anolis carolinensis*: a genome-wide phylogeographic investigation. Ecol Evol. 6(22):8115–8125.
McGee MD, Neches RY, Seehausen O. 2016. Evaluating genomic divergence and parallelism in replicate ecomorphs from young and old cichlid adaptive radiations. Mol Ecol. 25(1):260–268.
McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160(3):1231–1241.
Mullen LM, Hoekstra HE. 2008. Natural selection along an environmental gradient: a classic cline in mouse pigmentation. Evolution 62(7):1555–1570.
Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos Trans R Soc Lond B Biol Sci. 367(1587):409–421.
Nadeau NJ, et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos Trans R Soc Lond B Biol Sci. 367(1587):343–353.

Noor MAF, Bennett SM. 2009. Islands of speciation or mirages in the desert Examining the role of restricted recombination in maintaining species. Heredity 103(6):439–444.
Pavlidis P, Jensen JD, Stephan W, Stamatakis A. 2012. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. Mol Biol Evol. 29(10):3237–3248.
Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution 58(9):2064–2078.
Petuch, EJ. 2004. Cenozoic Seas : the View from Eastern North America. Boca Raton: CRC Press.
Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol. 31(7):1929–1936.
Poelstra JW, et al. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science 344(6190):1410–1414.
Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. 2018. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife 7:1–21.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.
Roux C, et al. 2014. Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a Mytilus hybrid zone. J Evol Biol. 27(8):1662–1675.
Roux C, et al. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol 14(12):e2000234.
Rozas J, Gullaud M, Blandin G, Aguade M. 2001. DNA variation at the *rp49* gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics 158(3):1147–1155.
Ruggiero RP, Bourgeois Y, Boissinot S. 2017. LINE insertion polymorphisms are abundant but at low frequencies across populations of *Anolis carolinensis*. Front Genet. 8:1–14.
Rupp SM, et al. 2017. Evolution of dosage compensation in *Anolis carolinensis*, a reptile with XX/XY chromosomal sex determination. Genome Biol Evol. 9(1):231–240.
Schrider DR, Kern AD. 2016. S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. 12:1–31.
Seehausen O, et al. 2014. Genomics and the origin of species. Nat Rev Genet. 15(3):176–192.
Sheehan S, Song YS. 2016. Deep learning for population genetic inference. PLoS Comput Biol. 12:1–28.
Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS. 2006. Comparative phylogeography of unglaciated eastern North America. Mol Ecol. 15(14):4261–4293.
Sousa VC, Carneiro M, Ferrand N, Hey J. 2013. Identifying loci under selection against gene flow in isolation-with-migration models. Genetics 194(1):211–233.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595.
Teeter KC, et al. 2008. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. 18(1):67–76.,
Terhorst J, Kamm JA, Song YS. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 49(2):303–309.
Tine M, et al. 2014. European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Commun. 5:5770.
Tollis M, Ausubel G, Ghimire D, Boissinot S. 2012. Multi-locus phylogeographic and population genetic analysis of *Anolis carolinensis*: historical demography of a genomic model species. PLoS One 7:1–14.
Tollis M, Boissinot S. 2011. The transposable element profile of the *Anolis* genome: how a lizard can provide insights into the evolution of vertebrate genome size and structure. Mob Genet Elements. 1(2):107–111.
Tollis M, Boissinot S. 2013. Lizards and LINEs: selection and demography affect the fate of L1 retrotransposons in the genome of the green anole (*Anolis carolinensis*). Genome Biol Evol. 5(9):1754–1768.
Tollis M, Boissinot S. 2014. Genetic variation in the green anole lizard (*Anolis carolinensis*) reveals island refugia and a fragmented florida during the quaternary. Genetica 1:59–72.
Torres R, Stetter MG, Hernandez RD, Ross-Ibarra J. 2019. The temporal dynamics of background selection in non-equilibrium populations. bioarXiv. doi: https://doi.org/10.1101/618389
Van Der Auwera GA, et al. 2014. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 11:11.10.1–11.10.33.
Wade J. 2012. Sculpting reproductive circuits: relationships among hormones, morphology and behavior in anole lizards. Gen Comp Endocrinol. 176(3):456–460.
Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016. Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. Mol Biol Evol. 33(7):1754–1767.
Wolf JBW, Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. 18(2):87–100.

Associate editor: Takashi Gojobori

PLOS GENETICS

RESEARCH ARTICLE

Disentangling the determinants of transposable elements dynamics in vertebrate genomes using empirical evidences and simulations

Yann Bourgeois[1,2]*, **Robert P. Ruggiero**[2,3], **Imtiyaz Hariyani**[2], **Stéphane Boissinot**[2]*

1 School of Biological Sciences, University of Portsmouth, Portsmouth, United Kingdom, **2** New York University Abu Dhabi, Saadiyat Island Campus, Abu Dhabi, United Arab Emirates, **3** Department of Biology, Southeast Missouri State University, Cape Girardeau, MO, United States of America

* yann.bourgeois@port.ac.uk (YB); stephane.boissinot@nyu.edu (SB)

OPEN ACCESS

Citation: Bourgeois Y, Ruggiero RP, Hariyani I, Boissinot S (2020) Disentangling the determinants of transposable elements dynamics in vertebrate genomes using empirical evidences and simulations. PLoS Genet 16(10): e1009082. https://doi.org/10.1371/journal.pgen.1009082

Editor: Cédric Feschotte, Cornell University, UNITED STATES

Received: April 23, 2020

Accepted: August 25, 2020

Published: October 5, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009082

Data Availability Statement: The scripts used to perform simulations using SLiM3 are available on Github (https://github.com/YannBourgeois/SLIM_simulations_TEs). All sequencing data are available

Abstract

The interactions between transposable elements (TEs) and their hosts constitute one of the most profound co-evolutionary processes found in nature. The population dynamics of TEs depends on factors specific to each TE families, such as the rate of transposition and insertional preference, the demographic history of the host and the genomic landscape. How these factors interact has yet to be investigated holistically. Here we are addressing this question in the green anole (*Anolis carolinensis*) whose genome contains an extraordinary diversity of TEs (including non-LTR retrotransposons, SINEs, LTR-retrotransposons and DNA transposons). We observed a positive correlation between recombination rate and frequency of TEs and densities for LINEs, SINEs and DNA transposons. For these elements, there was a clear impact of demography on TE frequency and abundance, with a loss of polymorphic elements and skewed frequency spectra in recently expanded populations. On the other hand, some LTR-retrotransposons displayed patterns consistent with a very recent phase of intense amplification. To determine how demography, genomic features and intrinsic properties of TEs interact we ran simulations using SLiM3. We determined that i) short TE insertions are not strongly counter-selected, but long ones are, ii) neutral demographic processes, linked selection and preferential insertion may explain positive correlations between average TE frequency and recombination, iii) TE insertions are unlikely to have been massively recruited in recent adaptation. We demonstrate that deterministic and stochastic processes have different effects on categories of TEs and that a combination of empirical analyses and simulations can disentangle these mechanisms.

Author summary

Transposable elements (TEs) are mobile DNA sequences that can replicate and insert in genomes. By doing so, they can disrupt gene function and meiotic process, but also generate evolutionary novelties. It is however unclear how different processes such as varying

on the European Nucleotide Archive (https://www.ncbi.nlm.nih.gov/sra) under the BioProject designation PRJNA376071 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA376071). All TEs counts and genotype files are available on Dryad (doi:10.5061/dryad.wpzgmsbjw).

Funding: This work was supported by New York University Abu Dhabi (NYUAD) research funds AD180 (to SB). The NYUAD Sequencing Core is supported by NYUAD Research Institute grant G1205-1205A to the NYUAD Center for Genomics and Systems Biology. The funding bodies had no role in designing the study, nor in data collection and interpretation.

Competing interests: The authors have declared that no competing interests exist.

rates of transposition, selection on TEs, linked selection and genome properties interact with each other. Here, we use the green anole (*Anolis carolinensis*) as a model, since it harbors one of the highest diversities of TEs found in a vertebrate (including non-LTR retrotransposons, LTR-Retrotransposons, DNA transposons and SINEs). By studying the population genomics of these different categories of TEs *within the same species*, we are able to disentangle processes that are specific to TE clades from general processes related to drift and selection. To do so, we use simulations of TEs in their genomic context to provide an interpretation of associations between recombination rate and statistics summarizing TE diversity and abundance. Our results highlight clear differences in TE dynamics across clades, with a clear dichotomy between SINEs/DNA-transposons and LTR-Retrotransposons/long LINEs. These differences can be mostly explained by changes in the relative impact of selection against TEs, linked selection, and insertional preferences.

Introduction

Transposable elements (TEs) are among the genomic features that display the most variation across the living world. The nature of the interactions between these genomic 'parasites' and their hosts has likely played a considerable role in determining the size, structure and function of eukaryotic genomes [1–3]. From the perspective of TEs, genomes can be seen as an ecosystem with distinct niches. Borrowing from community ecology concepts [4,5], variation in TE composition and diversity along the genome may be due to competition for resources between clades or constraints linked to changes in environmental conditions (niche-partitioning). An alternative model would posit that TE diversity be driven by stochastic events of population size changes in the host and drift that are independent of intrinsic TE properties such as selection or transposition (neutral theory) [6]. Within a given host species, these processes can be studied through the prism of population genetics, a field that conceptually inspired the study of ecological communities. Processes linked to niche-partitioning such as varying selection against new insertions [7], variability in the use of cellular machinery and access to chromatin by different TE clades [8,9], or domestication of elements [10], may shape TEs diversity in predictable ways. On the other hand, stochastic processes at the level of individual elements, but also demography at the scale of the host [11–13], may be sufficient to explain variation in the TE landscape [4]. In addition, stochastic processes may not be constant along the genome. For example, recent investigations have highlighted the importance of recombination rates in shaping genomic diversity, due to the effects of selection at linked sites. Because of Hill-Robertson interference, regions near a selected site see their genetic diversity drop, an effect that increases in regions of low recombination [14]. This drop may not only affect nucleotide diversity, but also TEs and other structural variants.

In this work, we investigate three main factors that may impact TE distribution and diversity in the genome: direct selection on TEs, Hill-Robertson interference, and differences in their properties (e.g. preferential insertion). Many of these mechanisms make predictions about the correlation between recombination rate and diversity. For example, it is often assumed that higher recombination rates may result in higher rates of ectopic recombination, making repetitive elements more deleterious in regions of high recombination (e.g. [7,15]). This should result in negative correlations between TEs abundance/frequency and recombination. Hill-Robertson interference leads to shorter coalescence times in regions of low recombination. This may result in a faster fixation of neutral and slightly deleterious mutations, but also in lower polymorphism than in regions of high recombination [16,17]. At last, because

recombination rate is often correlated with other genomic features such as exon density, DNA repair machinery, or open chromatin, variation in TE insertion mechanisms may be reflected in correlations between their density and recombination.

In vertebrates, most of the knowledge on the micro-evolutionary dynamics of TEs is provided by studies on humans [7]. It seems clear that mechanisms such as drift, selection and migration may play an important role in shaping TEs abundance and frequencies (e.g. [11]). In addition, TEs can insert within regulatory sequences and coding regions, and have a strong potential to reduce fitness. It is therefore likely that they are under purifying selection, which should leave specific signatures such as allele frequency spectra skewed towards rare variants in TEs compared to near-neutral markers such as SNPs [18]. In human, purifying selection acting against long TEs has been demonstrated and this pattern was explained by the greater ability of long elements to mediate deleterious ectopic recombination [19]. While the human model has provided deep insights about the dynamic of LINEs in mammals, it provides only a partial picture of the dynamics of TEs as a whole, given the absence of recent activity of other categories of TEs, such as DNA transposons, in the human genome. In fact, mammalian genomes are unique among vertebrates. They are typically dominated by a single category of autonomous element, *L1*, and related non-autonomous elements (e.g. *Alu* in primates).

Non-mammalian vertebrates display a much larger TE diversity, and often include both class I elements (i.e. elements that use an RNA intermediate in their life cycle) and class II elements (i.e. elements that don't use an RNA intermediate). Class I includes LTR-retrotransposons, non-LTR retrotransposons (*i.e*. LINEs and *Penelope*) and their non-autonomous counterparts (SINEs). Class II includes a wide diversity of elements including the widespread DNA transposons. Since TEs vary in their mode of transposition, length, regulatory content and structure, it is likely that the effect they have on host fitness and how they are in turn affected by host-specific response will differ. A potentially fruitful approach to this question would be to apply the conceptual and practical tools of population genetics in a model harboring a wide diversity of active TEs. This would facilitate direct comparisons between TE categories while removing the confounding effects of host demography since all elements within the same genome share the same demographic history. The growing availability of whole-genome resequencing data, as well as the development of new computational tools, has revived the interest of the evolutionary genomics community for the analysis of TE polymorphisms within species [20,21].

Whether TEs constitute a substrate for adaptation is another area of interest. Since TEs can lead to substantial regulatory and structural variation, they may constitute targets for fast adaptation and be domesticated by the host's genome [22]. Several possible cases have now been identified at short evolutionary scales, such as the involvement of a TE insertion in industrial melanism trait in peppered moth [23], or the association between some TEs and adaptation to temperate environment or pesticides [10,24] in *Drosophila*. Identifying candidate TEs (and more generally genomic regions) for positive selection is still challenging, and requires stringent filters to keep the number of false positives at a minimum. Combining genome scans obtained from SNP data with a screening of TEs displaying strong difference in frequencies across populations should, fulfill this goal [20,25].

In this study we investigate TE variation in the green anole (*Anolis carolinensis*), which is a particularly relevant model since it is extremely diverse in terms of TE content. Its genome contains four main TE categories, each represented by multiple clades of elements: non-LTR autonomous retrotransposons (nLTR-RT; including the *L1*, *CR1*, *L2* and *Penelope* clades), SINEs, LTR-retrotransposons (LTR-RT; including the *BEL*, *Copia*, *Gypsy* and *Dirs* clade), and DNA transposons (including *hAT*, *hobo*, *Tc1/Mariner* and *helitrons* clades). There is preliminary evidence that TEs may have been involved in adaptation in anoles, for example by

inserting in the *Hox* genes cluster [26]. Previous studies have investigated patterns of genetic structure and past history: the ancestor of the green anole originally colonized Florida from Cuba between 6 and 12 million years ago [27]. A first step of divergence occurred in Florida between 3 and 2 mya (S1 Fig) [28], producing three distinct genetic clusters in Florida, the North-Eastern Florida population (NEF), the North-Western Florida population (NWF) and the South Florida population (SF), the latter being the basal one. The ancestral population of lizards now living in temperate territories diverged from the NEF cluster approximately 1 Mya. This divergence was followed by expansion northwards from Florida to the remaining South-Eastern USA, across the Gulf Coastal Plain over the last 100,000–300,000 years [29,30]. This led to the emergence of the two current northern populations, Gulf-Atlantic (GA) and Carolinas (CA). A key aspect of these studies is that they revealed large effective population sizes in all clusters, which should increase the efficiency of selection on TEs and render it easier to detect. In addition, the broad set of environmental conditions encountered by the green anole should provide opportunities for recruitment of TEs by positive selection. At last, genetic diversity is highly variable along the green anole genome, reflecting the joined effects of heterogeneous recombination rates and linked selection [30].

We take advantage of previous studies that investigated the recombination and diversity landscape along the genome to assess i) how does diversity and genomic repartition vary across different TE clades; ii) if direct selection against TE insertions is detectable; iii) how the interaction between demography, counter-selection and linked selection may impact TE frequencies and local abundance; iv) whether there is any clear evidence for positive selection acting on TEs.

Results

Description of polymorphic insertions

A total of 339,149 polymorphic TE insertions with no missing genotype were recovered from resequencing data obtained from 28 anoles, including the five genetic clusters identified in previous studies [29,30]. This included both reference and non-reference insertion. Two of these genetic clusters (GA and CA, referenced as Northern populations) went through a bottleneck 100,000 years ago. Note that the individual used to build the reference genome was sampled in South Carolina, which places it in the Northern populations [31]. The most abundant category of polymorphic TE found in our dataset consisted in DNA transposons (N = 132,370), followed by nLTR-RTs (N = 97,586), LTR-RTs (N = 78,472), and SINEs (N = 30,721). At a finer taxonomic scale, we mostly identified elements belonging to the *CR1*, *L2*, *L1* and *Penelope* clades for nLTR-RTs, *Gypsy* and *DIRS* for LTR-RT, and *Hobo*, *Tc1/Mariner*, *hAT* and *Helitron* for DNA transposons (Table 1). Elements such as *R4*, *RTEX*, *RTE-BovB*, *Vingi* or *Neptune* were rare and mostly fixed (Table 1), probably due to their older age. The same was observed for ancient repeats, classified as *Eulor*, *MER*, *UCON* or *REP* for DNA transposons.

Diversity within individuals and genetic clusters

We first examined the possible impact of demography on TEs diversity and abundance. In each individual, we assessed whether heterozygous insertions were found in other green anoles or outgroups. We focused on shared heterozygosity at the individual level to better visualize intra- and inter-individual diversity (Fig 1). Singletons are more likely to be of recent origin, while heterozygous TEs shared between multiple individuals should be older, which may give information about the past and current dynamic of polymorphic elements. Given the low homoplasy of TE insertions [32,33], elements shared with the two outgroups were almost certainly found in the common ancestor, and may highlight how past demography impacted

Table 1. Summary of TE polymorphisms in the five genetic clusters identified in the green anole, and its two Cuban counterparts. For each cluster/outgroup, the number of polymorphic or fixed elements is given. Note that GA and CA (Northern populations) went through a bottleneck approximately 100,000 years ago.

			A. allisoni		*A. porcatus*		SF		NWF		NEF		GA (Northern pop)		CA (Northern pop)	
Category	**Clade**	**N**	**Fixed**	**Heterozygous**	**Fixed**	**Heterozygous**	**Fixed**	**Polymorphic**	**Fixed**	**Polymorphic**	**Fixed**	**Polymorphic**	**Fixed**	**Polymorphic**	**Fixed**	**Polymorphic**
nLTR-RTs	*CR1*	32804	3892	2613	3343	2795	3557	6712	3488	6783	3328	13601	4059	5175	4357	3476
	L2	26392	4261	4396	4469	5051	6577	5196	6604	4962	6259	6250	7004	2929	7266	2064
	Penelope	16208	978	1375	1313	1470	1074	2935	1056	3611	915	6426	1081	2670	1180	1876
	L1	14181	1057	1474	1088	1594	1023	2842	1012	3231	914	5414	1128	2137	1243	1491
	RTE1	3709	418	232	370	372	348	309	352	943	332	1152	362	606	375	221
	R4	1516	166	345	253	427	265	308	274	310	286	243	303	134	301	98
	RTE_BovB	920	240	209	302	235	358	167	377	151	347	187	380	83	392	66
	Vingi	860	360	174	306	151	777	75	783	77	758	102	823	37	826	34
	RTEX	496	128	134	206	146	416	73	425	68	380	116	447	49	465	31
	Neptune	376	14	4	4	12	6	37	4	57	3	215	3	56	4	28
	Other	124	14	18	18	24	25	22	26	18	24	45	25	28	28	15
DNA transposons	*Hobo*	45421	1380	6693	986	6900	244	11869	31	13042	2	19344	6	9900	39	6509
	Tc1/Mariner	37718	8380	6692	8072	8133	4533	6681	4600	6386	4190	11070	4759	4661	4935	3112
	hAT	25165	3520	5115	6705	5043	8679	5348	8778	5115	7841	7515	9058	4252	9461	2797
	Helitron	19266	1730	2008	1093	2491	147	3899	21	5007	2	7779	4	2424	24	1499
	Other	3517	569	1015	1783	878	2847	636	2958	554	2666	850	3098	419	3203	314
	Chapaev	1229	154	329	491	284	783	286	833	204	728	360	837	198	896	123
	MER	17	6	3	8	4	13	4	14	3	12	5	16	1	15	2
	Eulor	16	2	5	10	3	13	3	13	3	13	3	15	1	15	1
	UCON	11	0	1	4	5	7	4	9	2	9	2	10	1	10	1
	Chompy	5	1	2	3	1	4	1	5	0	5	0	4	1	5	0
	Harbinger	3	2	1	0	1	1	2	2	1	3	0	2	1	3	0
	REP	2	0	1	1	1	1	1	2	0	1	1	1	1	2	0
LTR-RTs	*Gypsy*	45625	1037	1223	985	1299	1338	7408	1029	9186	940	15157	1106	12939	1218	8020
	Other LTRs	13946	219	349	241	322	753	3719	441	4237	366	6215	419	2958	509	2153
	BEL	9391	183	189	125	234	166	1329	157	1154	148	4380	175	1929	194	1369
	Dirs	6873	391	577	297	607	362	1226	320	1618	315	1796	363	1332	393	930
	Copia	1962	324	268	315	550	257	279	239	392	233	465	256	233	276	132
	ERV	674	79	62	80	63	92	112	89	135	87	178	94	130	102	75
	Ultra-conserved	1	0	0	0	1	1	0	1	0	0	1	1	0	1	0
SINEs	*SINE2*	20716	4121	5185	4568	5757	5600	3275	5661	2954	5331	3846	5837	1769	5960	1294
	Non-assigned (ACASINE)	4802	997	1577	1427	1564	1908	818	1941	836	1839	1024	1984	554	2024	418
	SINE1	4115	271	440	195	569	57	1031	27	1030	23	1423	25	473	29	367
	SINE3	1083	87	365	297	380	500	191	518	173	463	243	536	115	549	84
	MIR-like	5	0	2	4	1	5	0	5	0	2	3	5	0	5	0

https://doi.org/10.1371/journal.pgen.1009082.t001

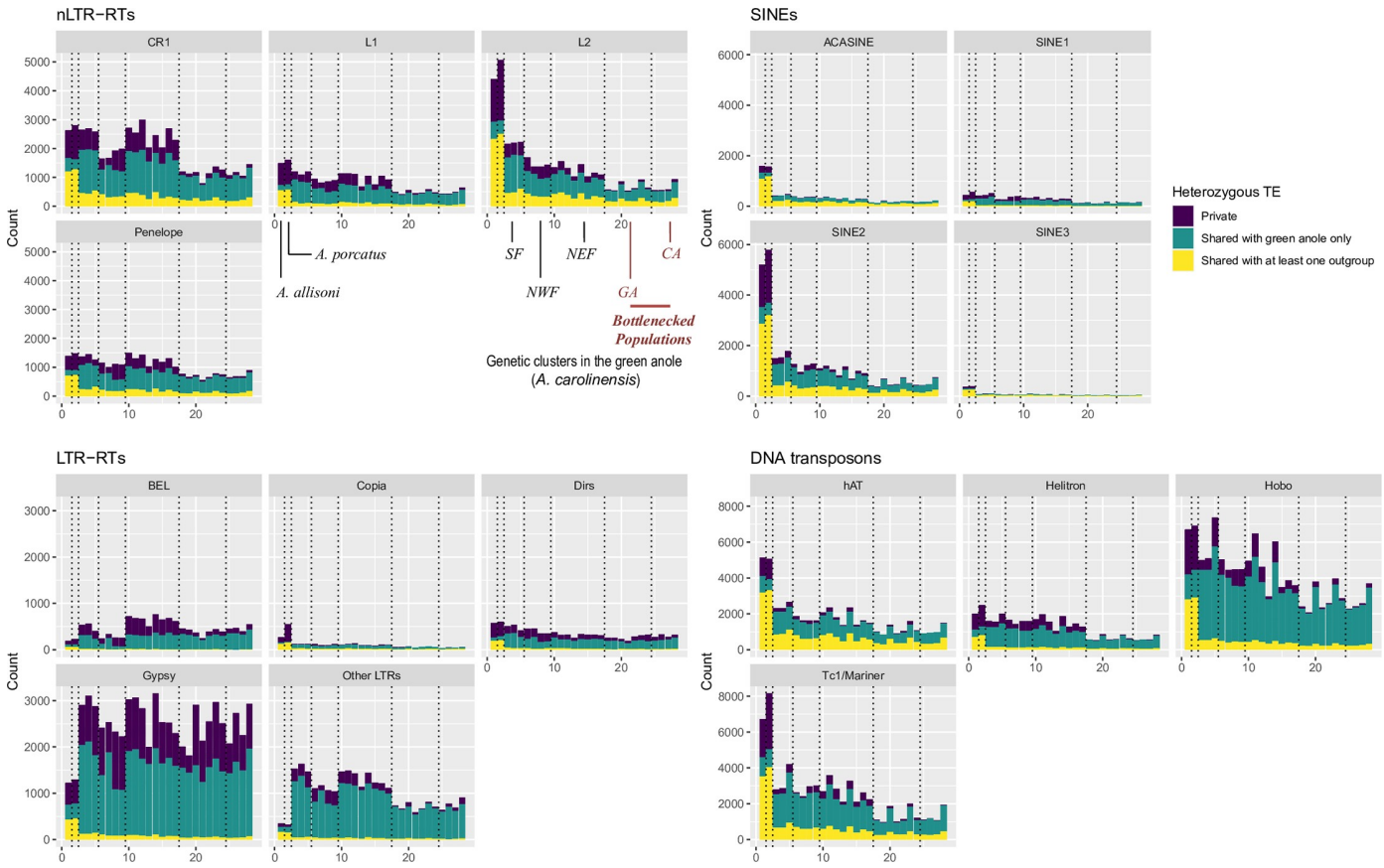


Fig 1. Count of heterozygous sites across all 28 individuals included in this study. Vertical dotted lined delimit the five main genetic clusters and the two outgroups in this order: *A. allisoni* and *A. porcatus*, SF, NWF, NEF, GA and CA. See S1 Fig for more details about these clusters.

https://doi.org/10.1371/journal.pgen.1009082.g001

individual TE landscapes. An examination of the repartition of polymorphic insertions across individuals showed a similar pattern across nLTR-RTs, SINEs and DNA transposons. On average, more heterozygous TEs were observed in individuals from the Floridian populations, which became established about two million years and remained stable and large (effective population size, N_e ~ 1 million) since colonization from Cuba. For these three categories, heterozygous TEs (private or shared) are more abundant in the outgroups (which correspond to the 2 Cuban anole species) and in the Floridian populations but become rarer in populations that expanded out of Florida, which is consistent with the loss of genetic variation experienced in those more recently established populations. In addition, for the most abundant clades, there were always more fixed insertions in GA and CA than in Floridian populations with similar sample sizes (Table 1). These patterns are consistent with drift leading to faster fixation or elimination of polymorphic TEs. For nLTR-RTs and SINEs, *L2* and *SINE2* elements displayed a large number of heterozygous TEs found only in the two outgroups, but also displayed a large proportion of heterozygous sites shared between *A. carolinensis* and either *A. porcatus* or *A. allisoni*. The same was observed for the DNA transposons *Tc1/Mariner* and *hAT*. This suggests that a substantial proportion of elements inserted before the split between these species, and that drift may have led to gradual loss of shared elements. *Hobo*, *Helitron*, *SINE1*, *L1*, *CR1* and *Penelope* maintained a relatively high proportion of private insertions in individuals from Florida, less shared heterozygous sites and similar number of heterozygous insertions when

compared to the outgroups. This is consistent with elements at lower frequencies in the common ancestor, either because of stronger purifying selection or more recent transposition activity, leading to less shared variation between present genetic groups and species.

On the other hand, for LTR-RTs, elements from the *Gypsy* and *BEL* clades displayed a large number of private insertions in the green anole, with many insertions found only in a single individual, and no clear pattern of reduced abundance in bottlenecked populations from the Northern cluster. This can be interpreted as a signature of recent and active transposition in the green anole lineage. This was especially clear for *Gypsy* elements, suggesting a burst of transposition following colonization from Cuba.

A visual inspection of allele frequency spectra (AFS) confirmed the effect of demography on TEs (Fig 2, S2–S5 Figs): for DNA transposons, nLTR-RTs and SINEs, spectra were skewed toward singletons in genetic clusters with large population sizes (SF, NEF, NWF), while this trend was less pronounced in clusters having been through a recent bottleneck (GA and CA). This was reflected by systematically higher average allele frequencies in GA than in NEF (Wilcoxon tests, $P < 5.7.10^{-12}$ except for *SINE3*; $P = 0.03$), the only exception being ACASINE for

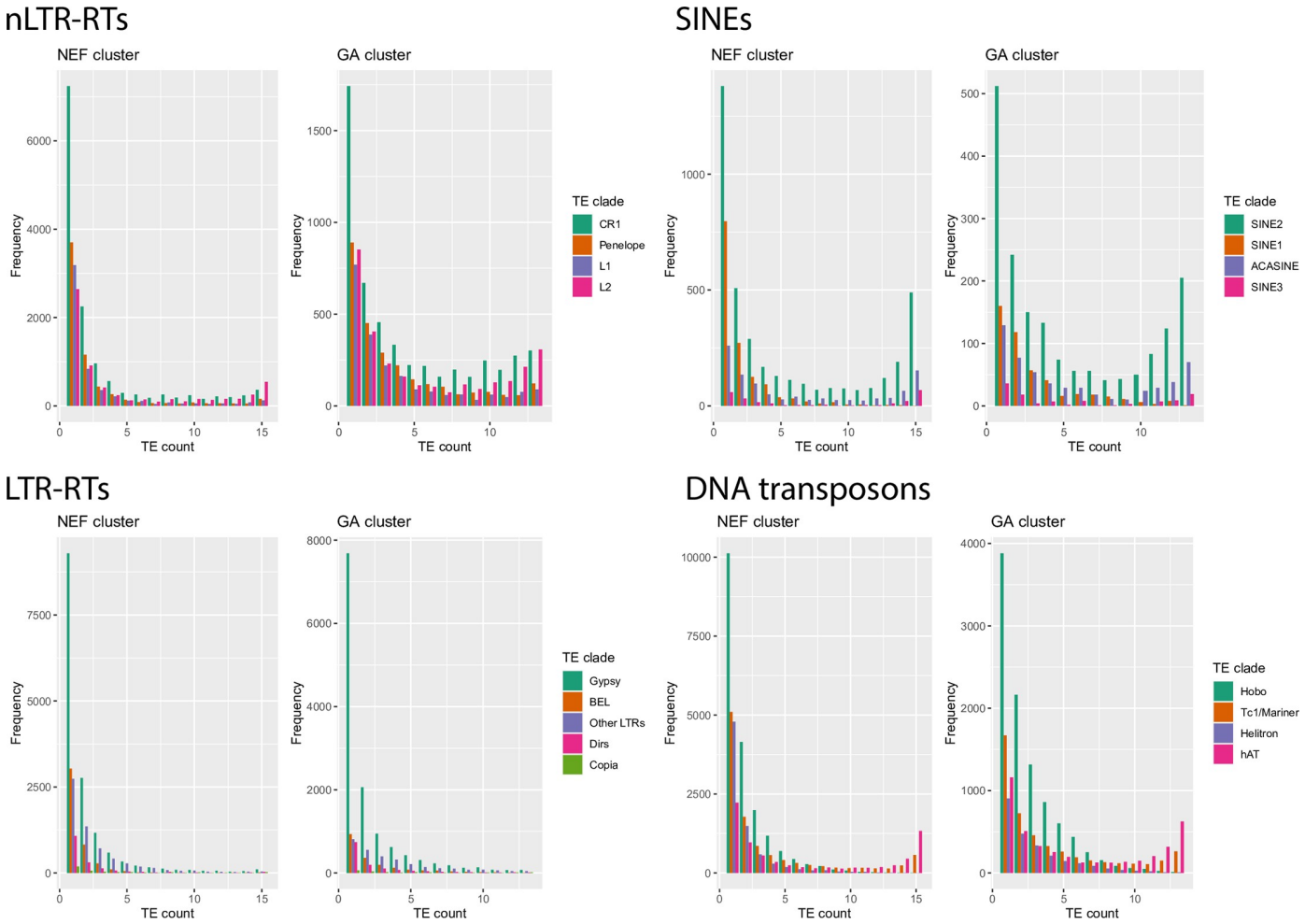


Fig 2. Allele frequency spectra for TEs belonging to two genetic clusters identified in the green anole. NEF (N = 8 diploid individuals) corresponds to a large, stable population from Florida, and GA (N = 7 diploid individuals) corresponds to a more recently established population having colonized northern environments in the last 100,000 years.

https://doi.org/10.1371/journal.pgen.1009082.g002

which no significant difference was observed. This is consistent with the excess of frequent alleles expected in the case of population contraction. These differences were however less clear for LTR-RTs, with spectra strongly skewed towards singletons in all populations. While AFS were clearly U-shaped for the other three types of elements, almost no LTR-RT insertion was found at very high frequencies. Such a pattern is consistent with recent activity and purifying selection preventing insertions to reach high frequencies. There were also differences within different types of elements. For non-LTR retrotransposons, elements such as *Poseidon* or *RTEBovB* were mostly found at high frequencies (Table 1). Elements such as *RTE1*, *L1*, *CR1* and *Penelope* displayed a stronger skew towards singletons than *L2*. In SINEs, *SINE1* had more singletons, while other elements were more frequent. For DNA transposons, the skew towards singletons was strongly pronounced for *Hobo* and *Helitron*, and very few fixed insertion were found (Table 1), suggesting either stronger purifying selection or a recent increase in transposition rate.

Correlation of TE density with recombination and differentiation reveals discordant patterns

Studies focusing on SNPs have revealed that regions of low recombination display lower diversity and stronger differentiation between populations due to the effects of linked selection [34,35]. First, we tested whether typical signals of linked selection could be observed along the genome by examining correlations between recombination rates, derived allele frequencies, and absolute (d_{XY}) and relative (F_{ST}) measures of differentiation computed over SNP data in non-overlapping 1Mb windows (Fig 3). We focused on the six main autosomes of the green anole. If linked selection shapes genomic diversity along the genome, there should be 1) positive correlations between diversity indices (average derived allele frequency, d_{XY}) and

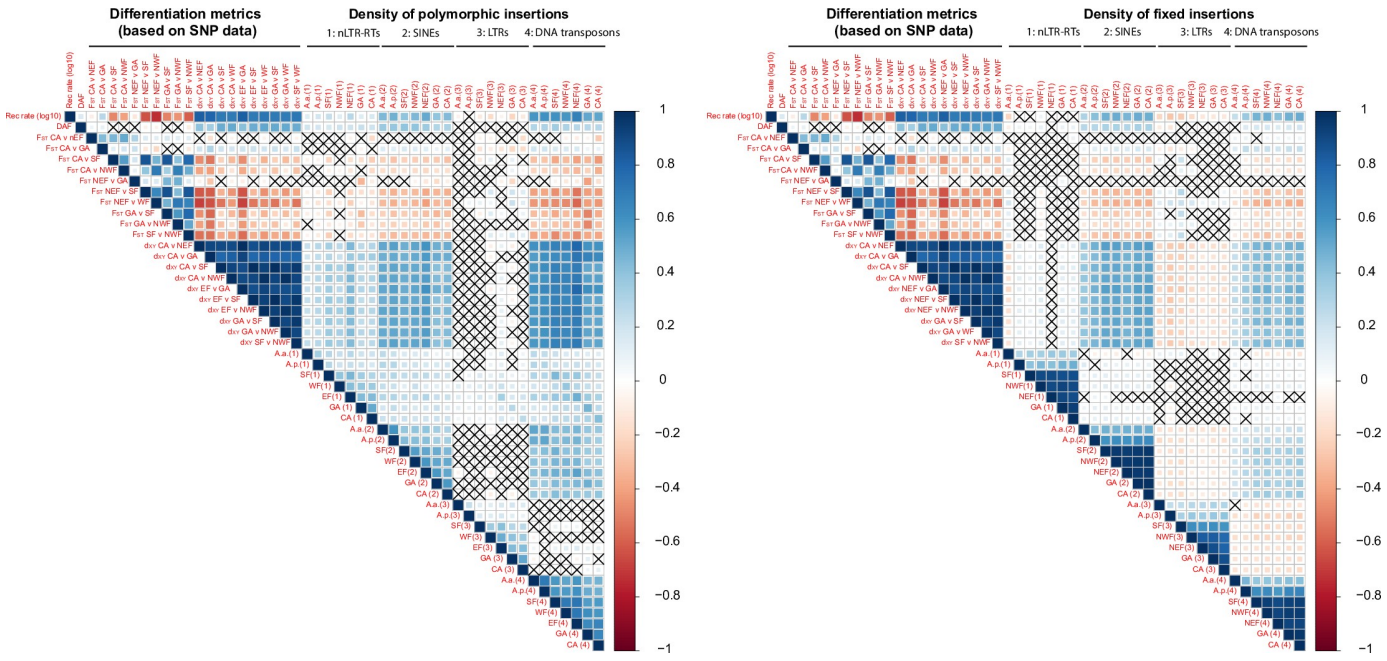


Fig 3. Correlograms illustrating Spearman's rank correlation coefficients between TE densities in 1 Mb windows and SNP-based statistics such as recombination rate (measured as r/μ, see Methods), pairwise relative (F_{ST}) and absolute (d_{XY}) measures of differentiation, and derived SNP frequency in the NEF cluster (DAF). Correlations with $P>0.05$ are indicated with a cross.

https://doi.org/10.1371/journal.pgen.1009082.g003

recombination, 2) negative correlations between differentiation measures (F_{ST}) and recombination, 3) consistency in genomic regions displaying high or low values for F_{ST} or d_{XY} across all pairwise comparisons. This is in line with our observations, with mostly positive correlations between recombination rate, diversity and absolute divergence for all pairwise comparisons between the five genetic clusters (Fig 3). Pairwise relative measures of differentiation (F_{ST}) were negatively correlated with recombination rate, d_{XY}, and derived allele frequencies, which is consistent with a role of linked selection reducing diversity in regions of low recombination across all genetic clusters. Indices of differentiation comparing CA or GA with other populations were less correlated with indices of differentiation estimated between pairs of clusters from Florida, suggesting a role for recent expansion in blurring the expected correlations.

Then, we examined densities of polymorphic and fixed TEs across four main categories of TEs (Fig 3). Assuming they are nearly neutral, linked selection should have a similar effect on TEs as on SNPs. The stronger Hill-Robertson interference observed in regions of low recombination should lead to a lower number of polymorphic TEs there. On the other hand, it is generally assumed that rates of ectopic recombination increase with crossover rates. In that case, elements involved in ectopic recombination should be under strong purifying selection, slowing the accumulation of TEs in regions of high recombination compared to regions of low recombination. This should result in decreasing densities of both polymorphic and fixed elements as recombination increases. TE densities were positively correlated with recombination rate, diversity and relative measures of differentiation for SINEs and DNA transposons. Correlations were weaker for nLTR-RTs, and almost absent for LTR-RTs. The density of fixed LTR-RTs even followed an opposite pattern, with more fixed insertions in regions of low recombination and high F_{ST}. For fixed nLTR-RTs, correlations were weak or absent. This suggests that purifying selection against LTR-RT and to some extent nLTR-RTs may explain the variation in their local abundance and diversity.

The lower abundance of some TE categories in regions of low recombination was not explained by a higher density of functional elements that could increase their deleterious effects (S6 Fig). Exon density was positively correlated with recombination rate (Spearman's $rho = 0.15$; $P = 9.1.10^{-7}$), which suggests that regions of high recombination may also be more frequently transcribed, and are therefore more often in an open chromatin state.

TE densities were positively correlated with each other across hosts' populations for all TEs, with correlations strengthening as comparisons involved more closely related pairs of populations. This effect is expected due to a longer shared history for related genetic clusters.

Comparison of TE diversity across TE clades in a demographically stable genetic cluster

We assessed whether purifying selection had a direct impact against TEs by examining average TE frequencies in 1Mb windows and comparing it to the frequencies of derived SNPs. To obtain a more accurate estimate of frequency, we focused on the population with the largest sample size and with a historically stable effective population size, NEF [30]. We also examined diversity at the clade level to highlight specific dynamics. We excluded TE clades with less than 5000 elements (Table 1), and merged SINEs that were not *SINE2* together to provide a comparison within the category. We examined these statistics for SNPs and the main clades within the four main TE categories (Fig 4). Average TE frequencies were lower for LTR-RTsthan for SNPs and the differences were statistically significant (frequencies of 0.10, 0.15, 0.13, 0.17, and 0.26 for *BEL*, *Dirs*, *Gypsy*, unclassified LTRs and derived SNPs respectively; paired-samples Wilcoxon tests, all $P<2.2.10^{-16}$) across all clades. This is consistent with either purifying selection against these elements, and/or their younger age. The same was observed for *CR1*, *L1* and

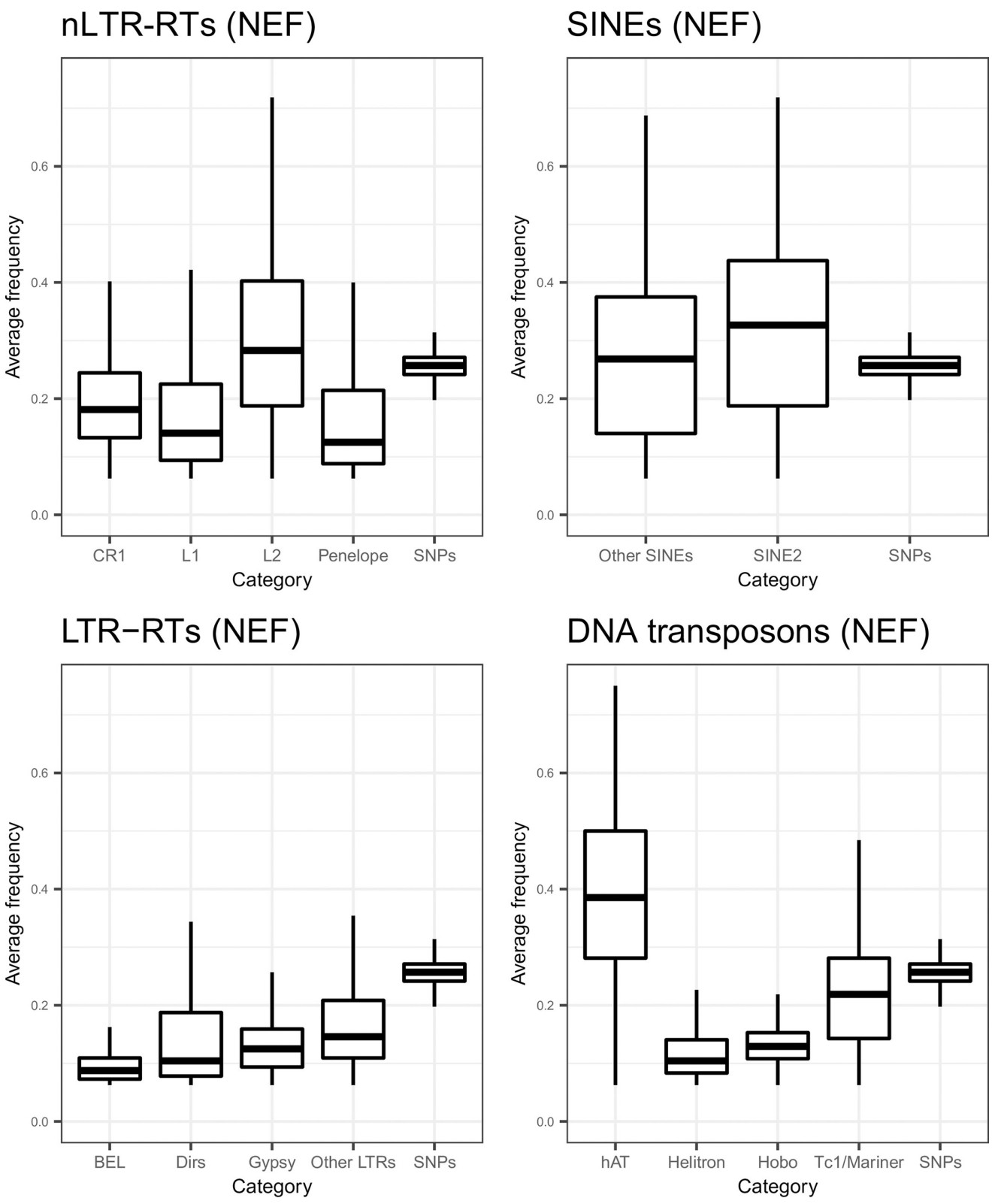


Fig 4. Boxplots of average TE frequency for each main TE category in the NEF population. For SNPs, the derived allele frequency was obtained by assigning variants to ancestral and derived states using *A. allisoni* and *A. porcatus*.

https://doi.org/10.1371/journal.pgen.1009082.g004

Penelope (frequencies of 0.19, 0.17 and 0.16), but not *L2* (frequency of 0.30), for which the average frequencies were significantly higher than derived SNPs (all $P<1.10^{-11}$). The average frequency of SINEs other than *SINE2* was 0.28, not substantially different from SNPs ($P = 0.88$), and was even higher for SINE2 (0.33, $P = 5.5.10^{-12}$). For DNA transposons, *Hobo*, *Helitron*, and to a lesser extent, *Tc1/Mariner* displayed lower frequencies than SNPs (0.13, 0.12 and 0.22 respectively, all $P<2.2.10^{-16}$). On the other hand, *hAT* displayed an average frequency of 0.39, substantially higher than SNPs ($P<2.2.10^{-16}$). Elements at a higher frequency than derived SNPs are likely ancient, and their high frequency is best explained by a non-equilibrium dynamic, with a lack of recent transposition resulting in a depletion in the lower frequencies of the allele frequency spectrum. Because DNA transposons replicate through a cut-and-paste mechanism, it may happen that some insertions be removed from a given insertion site. Nevertheless, the large effective population sizes considered here would make any substantial impact of occasional cut-and-paste extremely limited in terms of allele frequency.

TEs involved in ectopic recombination should be subject to purifying selection, becoming stronger in regions of high recombination. In addition, the higher exon density in these regions (S6 Fig) may increase the odds that these TEs alter gene expression. This should result in reduced frequency of polymorphic insertions and abundance of elements in regions of high recombination and high gene density. To test whether TEs from different clades followed this predicted pattern, we assessed whether their average frequency, density of polymorphic insertions, and density of fixed insertions, varied with the recombination rate (Figs 5, 6 and 7, Table 2). For all LTR-RTs, we observed negative correlations between recombination rate and average frequency (Fig 5). Weak, negative correlations were also observed when replacing frequency by the density of fixed insertions (Fig 7), the strongest trend being observed with

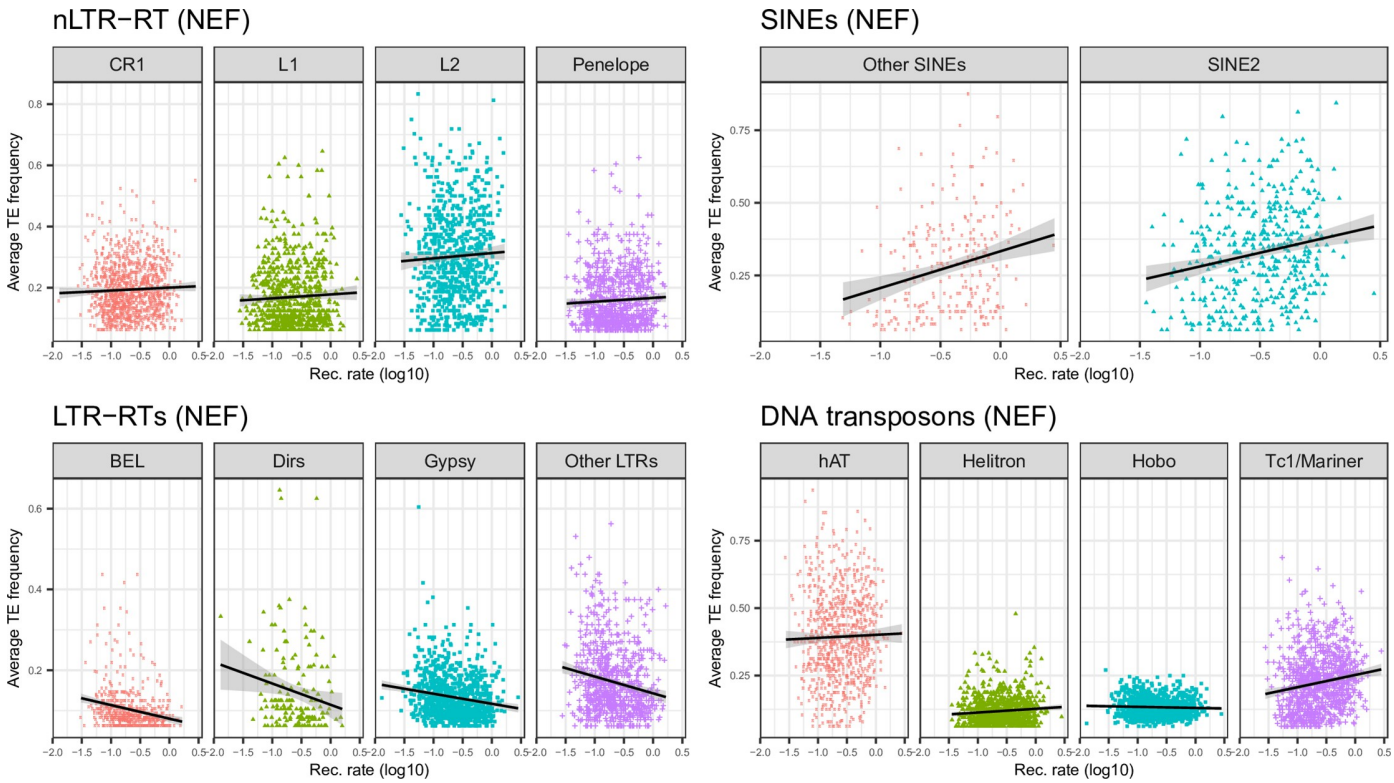


Fig 5. Plots of average TE frequency against recombination rate computed over 1Mb windows for each main TE clade in the NEF population.

https://doi.org/10.1371/journal.pgen.1009082.g005

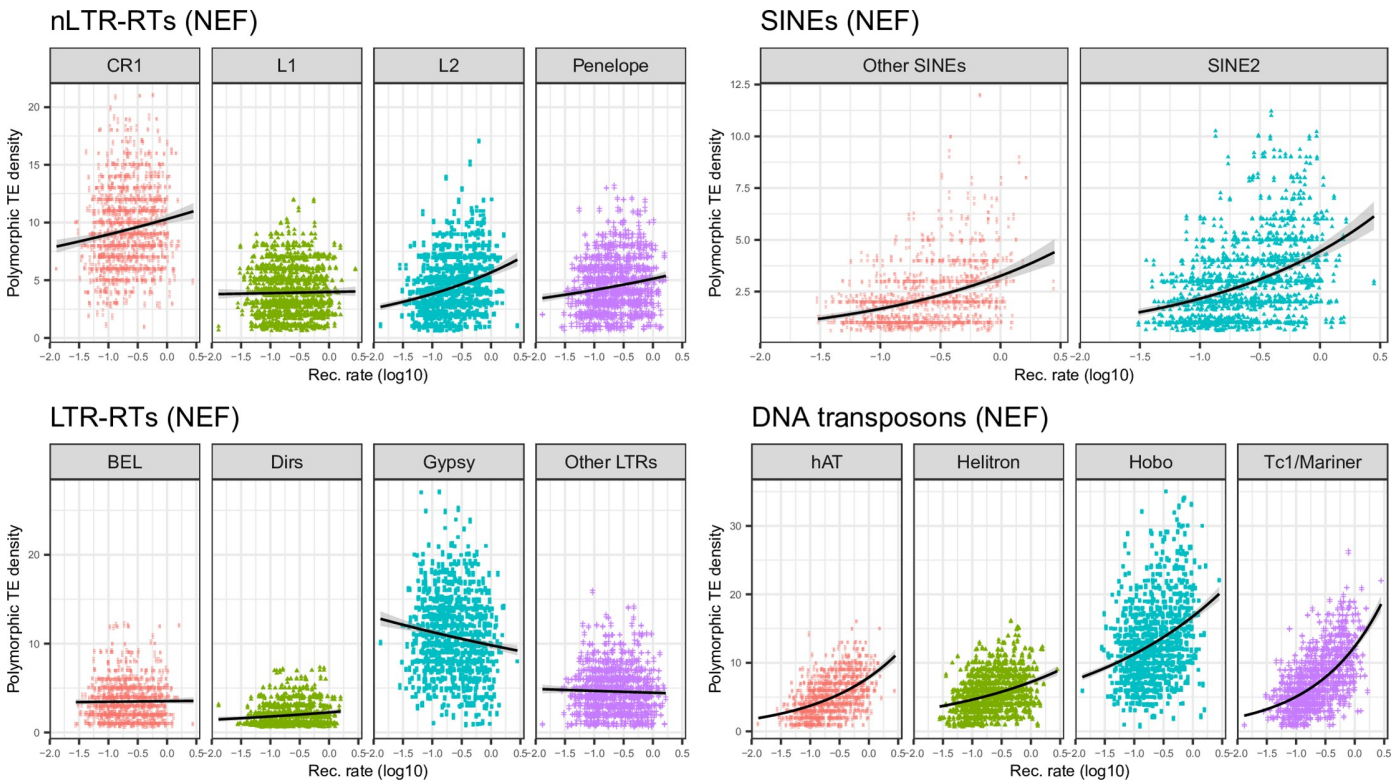


Fig 6. Plots of polymorphic TE density against recombination rate computed over 1Mb windows for each main TE clade in the NEF population.

https://doi.org/10.1371/journal.pgen.1009082.g006

Gypsy. For the latter, negative correlation between the density of polymorphic sites and recombination was observed (Fig 6). This pattern is clearly consistent with a stronger deleterious effect of these elements in regions of high recombination and gene density. Correlations were however weak (*BEL*, unclassified LTR-RT) for other LTR-RTs. They were significantly positive for *Dirs*. SINEs and DNA transposons (except *Hobo*) showed positive correlations between all three summary statistics and recombination rate, which may be partly explained by linked selection and a lack of strong purifying selection. For *Hobo*, the only significant correlation was found between recombination rate and the density of polymorphic sites, probably because of the rather low number of fixed insertions, obscuring correlations.

For nLTR-RTs, we did not observe significant correlations between recombination and TE frequency or the density of fixed insertions, except for *CR1* (Figs 5 and 7; Table 2). Positive correlations were however observed for *Penelope*, *CR1* and *L2* when examining the density of polymorphic sites. We however suspect that this lack of clear correlation may be due to variation in the strength of purifying selection among nLTR-RTs. Previous studies in vertebrates and *Drosophila*, [7,13,36,37] have shown that the effects of TE insertions on fitness may be correlated with their length. This may be due to the fact that the odds of homologous recombination rise with the length of homologous fragments [38], or because longer elements contain promoter sequences that may have more deleterious effects on nearby genes. Truncation in LINEs occurs at the 5' end of elements, which makes MELT estimates of their length accurate since it detects TEs based on reads mapping the ends of the insertion. To assess whether purifying selection acted more strongly on longer elements, we examined the correlation between recombination rates and the average length of fixed and polymorphic LINEs (which make

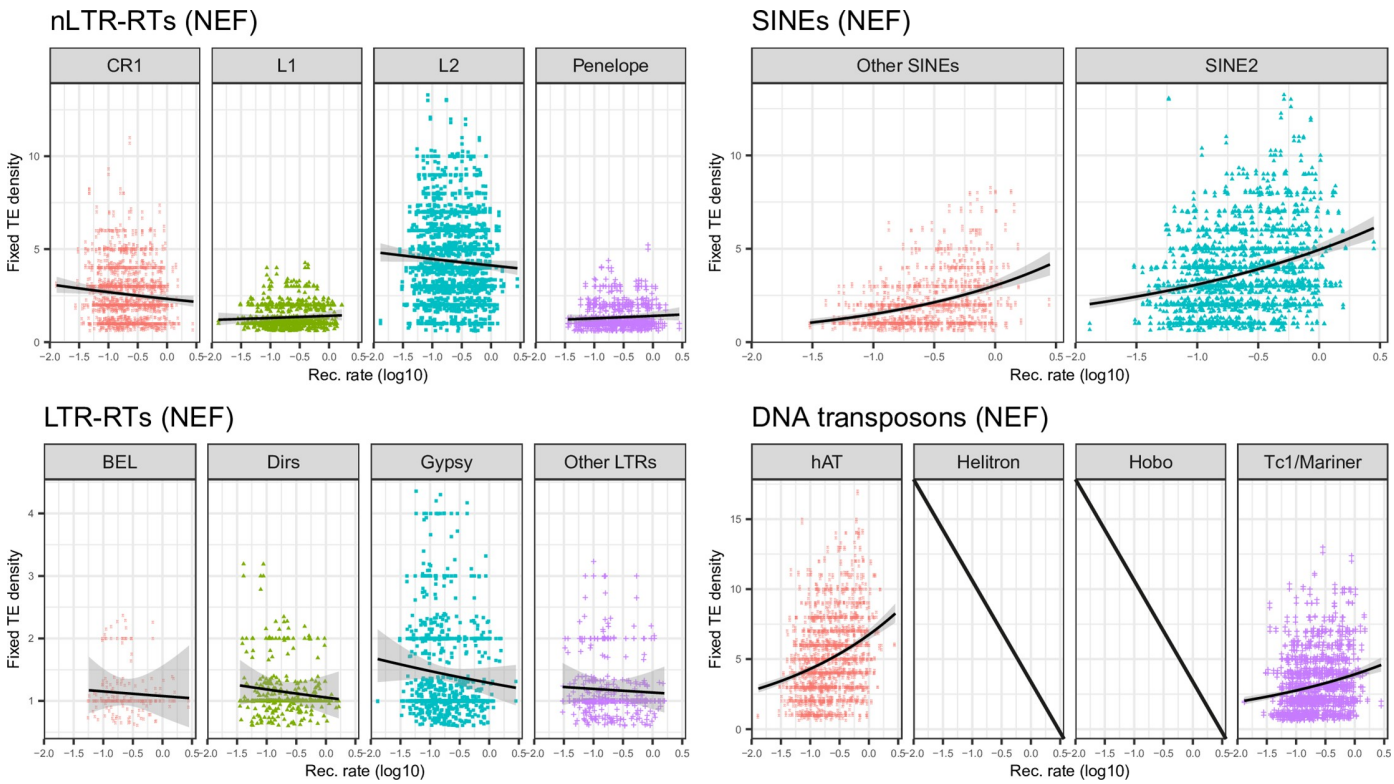


Fig 7. Plots of fixed TE density against recombination rate computed over 1Mb windows for each main clade in the NEF population. For *Helitron* and *Hobo*, there are not enough fixed insertions.

https://doi.org/10.1371/journal.pgen.1009082.g007

most of nLTR-RTs but exclude *Penelope*) in 1Mb windows (Fig 8), and observed a clear negative correlation between these two statistics (Spearman's *rho* = -0.16, -0.26, -0.21 for *CR1*, *L1* and *L2* respectively, $P<5.10–7$). LINEs that were fixed in the NEF population were also shorter than the polymorphic ones. We then focused on short LINEs (<20% of the maximum length of their respective clade) to assess whether they were also erased from regions of high recombination. We used 10Mb windows to increase the number of insertions and avoid losing too much information. We then reexamined the correlations between recombination rate and our three summary statistics (Fig 8). We found a positive correlation between frequency and recombination rate for short *CR1* and short *L2*. All short elements showed positive correlations between recombination and the density of polymorphic elements, while no clear correlation was observed for the density of fixed elements (Table 2, Fig 8). For long LINEs (>30% of the maximum length of their specific clade), we observed strong negative correlations between TE frequency, the density of fixed insertions, and recombination. The same was observed with the density of polymorphic insertions, except for *L2* (Table 2). These results suggest that weak correlations observed at the scale of the whole clade are explained by non-uniform, length-dependent selection against the elements. Short LINEs are therefore more likely under the influence of linked selection, while long LINEs display patterns that are closer to observations in LTR-RT, suggesting a stronger influence of purifying selection.

These results could also be explained by a higher rate of deletion in TE insertions located in regions of high recombination [39], or older elements containing more deletions. However, an examination of the start and end coordinates of insertions on their consensus did not reveal any substantial truncation at the 3' end (S7 Fig), which would be expected if deletion occurred

Table 2. Summary of correlations observed between average recombination rate, the average frequency of TEs, the density of polymorphic TEs and the density of fixed elements. For short and long nLTR-RTs, due to the low number of fixed insertions in 1Mb windows, we present results for 10MB windows instead. The last column provides an interpretation of the correlations obtained in simulations and observed in empirical data. For simulated TEs, we distinguish between outcomes where TEs are at high frequency (higher than SNPs) and low frequency (lower than SNPs). Pur. Selec. Ect. Rec.: Purifying Selection against ectopic recombination; Linked Sel.: Linked Selection; Pref. Ins.: Preferential Insertion in regions of high recombination/open chromatin; Anc. Burst: Ancient Burst of Transposition; (): the process may occur but does not impact the direction of correlations (for simulations), or is possible but no conclusive evidence is provided by the three summary statistics (for empirical observations). NA: for *Helitron* and *Hobo*, the lack of fixed insertions prevents the computation of these statistics. *: *P*-value<0.05; **: *P*-value<0.01; ***: *P*-value<0.001.

		Correlation between recombination rate and:			
Category	**Superfamily/simulation**	**Average frequency**	**Polymorphic density**	**Fixed density**	**Dominant process**
simulations	simulated TE	+	+	+	Linked Sel. + Pref. Ins.
	simulated TE	+	+	-	Linked Sel.
	simulated TE (high frequency)	-	+	+	Linked Sel. + Anc. Burst + Strong Pref. Ins.
	simulated TE (high frequency)	-	+	-	Linked Sel. + Anc. Burst
	simulated TE (low frequency)	-	+	-	Pur. Selec. Ect. Rec. + (Anc. Burst) + Pref. Ins.
	simulated TE	-	-	-	Pur. Selec. Ect. Rec. + (Anc. Burst)
nLTR-RTs	*CR1*	0.05	0.15***	-0.08 *	Mixture
	CR1 (short)	0.30**	0.25**	-0.24*	Linked Sel.
	CR1 (long)	-0.28 **	-0.14	-0.32 **	Pur. Selec. Ect. Rec.
	L1	0.02	0.01	0.08	Mixture
	L1 (short)	-0.006	0.35 ***	0.08	Linked Sel. + Pref. Ins.?
	L1 (long)	-0.25 *	-0.29 **	-0.57 ***	Pur. Selec. Ect. Rec.
	L2	0.05	0.29 ***	-0.06	Mixture
	L2 (short)	0.21 *	0.50 ***	-0.10	Linked Sel. + (Pref. Ins.)
	L2 (long)	-0.24 *	0.02	-0.32 **	Pur. Selec. Ect. Rec. + (Pref. Ins.)
	Penelope	0.04	0.15***	0.08	Linked Sel. + Pref. Ins.?
SINEs	*SINE2*	0.19 ***	0.42 ***	0.28 ***	Linked Sel. + Pref. Ins.
	Other SINEs	0.24 ***	0.37 ***	0.35 ***	Linked Sel. + Pref. Ins.
LTR-RTs	*Dirs*	-0.21 ***	0.14***	-0.09	Pur. Selec. Ect. Rec. + Pref. Ins.
	BEL	-0.28 ***	0.01	-0.05	Pur. Selec. Ect. Rec. + (Pref. Ins.)
	Gypsy	-0.18 ***	-0.13 ***	-0.11 *	Pur. Selec. Ect. Rec.
	Other LTRs	-0.18 ***	-0.04	-0.04	Pur. Selec. Ect. Rec. + (Pref. Ins.)
DNA transposons	*hAT*	0.04	0.51 ***	0.29 ***	Linked Sel. + Pref. Ins. + Anc. Burst
	Helitron	0.10 **	0.32 ***	NA	Linked Sel. + (Pref. Ins.)
	Hobo	-0.03	0.32 ***	NA	Linked Sel. + (Pref. Ins.)
	Tc1/Mariner	0.20 ***	0.59 ***	0.20 ***	Linked Sel. + Pref. Ins.

https://doi.org/10.1371/journal.pgen.1009082.t002

once the element is already inserted. We only observed truncation at the 5' end, which is consistent with truncation during the insertion process.

Simulations clarify the relative impact of purifying selection, linked selection and bursts of transposition on autonomous retrotransposons diversity

Our results reveal many combinations of correlations between TE diversity and recombination rate. To clarify and illustrate the conditions under which these combinations arise, we built a simple model of retrotransposon evolution in the forward-in-time simulator SLiM3 [40]. We simulated a 4Mb fragment with two recombination rates and negative selection on 10% of the non-coding SNPs. Recombination was high on the first and last Mb, and low for the 2Mb in the middle of the fragment. To reflect varying density of functional sites between regions of low and high recombination (S6 Fig), the density of coding sequences was 10,000 bp/Mb for the 2Mb in the middle of the fragment and 20,000 bp/Mb for the first and last Mb. Coding

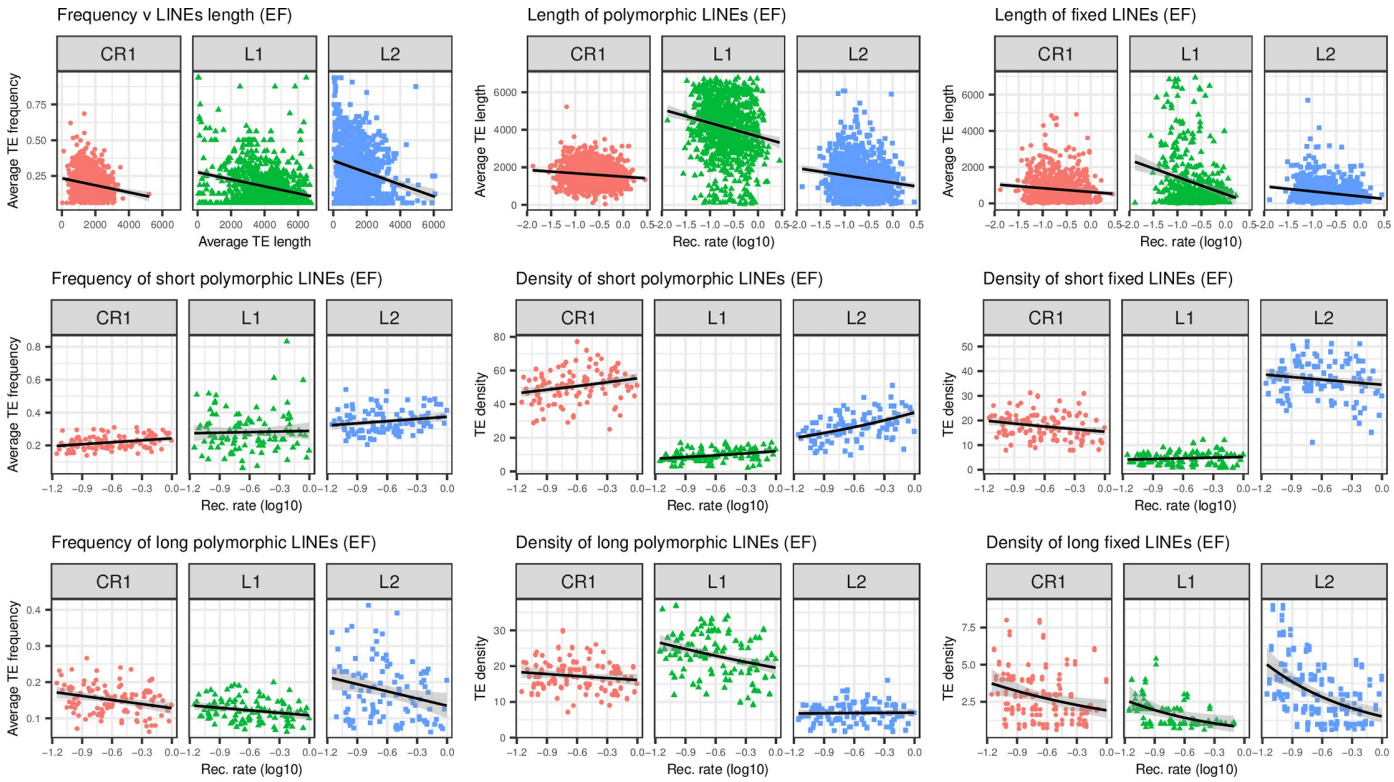


Fig 8. Top: plots of LINEs (i.e. nLTR-RTs excluding *Penelope*) length against recombination rate. Middle: Plots of average frequency, density of polymorphic insertions and density of fixed insertions for short LINEs, Bottom: same as middle row, for long LINEs. For middle and bottom plots, average frequencies and densities are computed for 10Mb windows.

https://doi.org/10.1371/journal.pgen.1009082.g008

sequences included negative selection on 70% of new mutations. Two categories of TEs were simulated, "short" TEs that were weakly deleterious (Fig 9, blue boxplots), and "long" TEs (red boxplots) that were more deleterious in regions of high recombination. Both long and short TEs falling in coding regions were strongly deleterious, with a selection coefficient s = -1. We then examined the same three summary statistics than earlier: the average frequency of polymorphic insertions, the density of polymorphic insertions, and the density of fixed insertions (Fig 9). Short TEs showed higher average frequencies in regions of high recombination when transposition was kept constant, a pattern consistent with expectations if linked selection increases lineage sorting in regions of low recombination (Fig 9, panels A). This trend was however reversed if transposition occurred as a single ancient burst (panels B). In that case, average TE frequencies were also higher, due to the older age of insertions. Moreover, because linked selection leads to faster lineage sorting in regions of low recombination, polymorphic insertions that survive after the burst reach higher frequencies, explaining the observed correlation. On the other hand, long TEs displayed lower average frequencies in regions of high recombination, due to their stronger deleterious effects, whether transposition was kept constant or not. Models including preference for TEs to insert in regions of high recombination (panels C and D) produced very similar results for this summary statistic.

The density of polymorphic insertions was higher in regions of high recombination for short TEs across all simulations, but the difference was even more pronounced when preference for regions of high recombination was added to the model (panels C and D). The trend was reversed for long TEs (panel A), but including preference for high recombination again

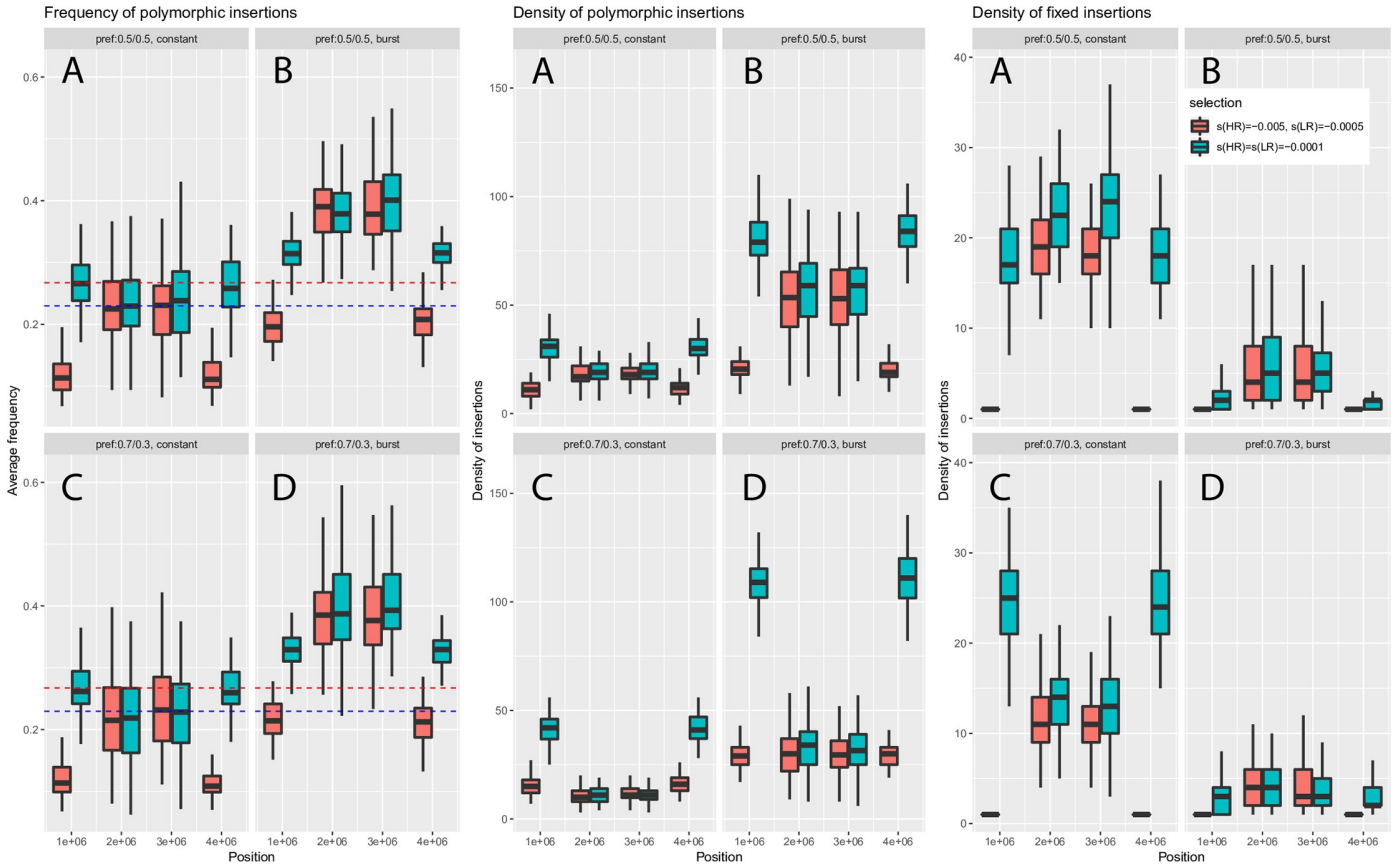


Fig 9. Summary of simulations of TEs using SLiM3, using parameters realistic for the NEF cluster. Eight diploid individuals were sampled to mimic our sampling scheme. Boxplots correspond to the results obtained over 100 simulations of a 4Mb fragment, divided into three regions of 1, 2 and 1 Mb. The first and last Mb correspond to regions of high recombination and high density of functional coding sites (respectively 10 times and 2 times higher than in the 2Mb central region). Coefficients of selection and other parameters are scaled using an effective population size of 1000 instead of 1,000,000 to reduce computation time (see Methods). $2N_es$ = -10 for 10% of non-coding sites and $2N_es$ = -100 for 70% of coding sites. Blue and red dotted lines correspond to average derived SNP frequencies in regions of low and high recombination respectively. A: model with constant transposition and no preferential insertion; B: model with a burst of transposition. C and D are the same than A and B respectively, but include preferential insertion in regions of higher recombination such that 70% of new elements go into these regions. TEs that fall in coding regions are strongly deleterious (selection coefficient 2Ns = -2000).

https://doi.org/10.1371/journal.pgen.1009082.g009

led to a positive correlation between recombination rate and the summary statistic (panel C), since more insertions could replace the ones erased by selection. Models where a burst of transposition occurred gave the same trends (panels B and D), although preference for high recombination did not fully reverse the correlation (panel D).

The density of fixed insertions was lower in regions of high recombination than in regions of low recombination in models with no preference (panels A and B). This result was observed for both short and long TEs, although the effect was enhanced for long TEs due to their stronger deleterious effect in regions of high recombination. In models where preferential insertion in regions of high recombination was added however, a positive correlation with recombination rate was observed under a constant transposition rate, and differences were less marked in the case of a transposition burst (panels C and D).

Our observations remained valid under scenarios with varying proportions of point mutations under purifying and positive selection, and selective coefficients (S8 to S12 Figs). Stronger purifying selection (S8 Fig, $2N_es$ = -400 in coding sequences, $2N_es$ = -40 in non-coding sequences) led to results similar to the ones shown in Fig 9, but the density of polymorphic

TEs tended to be even lower in regions of low recombination for "long" elements, reducing the contrast with regions of high recombination (panels A and B). In the case of weaker purifying selection (S9 and S10 Figs), we observed little difference in TE frequencies and densities across windows for "short" elements, consistent with their near-neutral behavior. At last, we note that adding modest amounts of positively selected sites to this latter model with weak purifying selection restored partially the correlations observed with strong purifying selection alone (S11 and S12 Figs).

We compared these trends with our actual observations (summarized in Table 2), which are consistent with either strong purifying selection against new insertions through ectopic recombination or predominant effects of linked selection. For short nLTR-RTs, and particularly *CR1*, we observed correlations consistent with linked selection, similar to the simulations for short elements highlighted in panels A of Fig 9. A possible effect of preferential insertion may explain the weak correlations observed between the density of fixed elements and recombination for *L1* and *L2* (panels C). For long nLTR-RTs, correlations for the three statistics were consistent with simulations obtained for long elements with no preferential insertion (Fig 9, panels A and B). The same was observed for *Gypsy* elements. For long *L2*, the lack of strong correlation between the density of polymorphic elements and recombination may reflect a situation closer from the simulations presented in panels C and D, with some effect of preferential insertion and past burst of transposition. The same reasoning may be applied to LTR-RT elements such as *BEL*. For *Dirs*, observations matched expectations for long elements in simulations shown in panel C, suggesting both selection against ectopic recombination and preferential insertion in regions of high recombination. For SINEs and *Tc1/Mariner*, the observed correlations clearly matched simulations for short elements including linked selection and preferential insertion (panel C). This scenario is also likely for *Hobo* and *Helitron*, although their weak frequencies obscures correlations between average allele frequencies, density of fixed sites, and recombination. The same issue makes any interpretation of patterns observed for *Penelope* difficult.

Given the high frequencies observed for *SINE2* and particularly *hAT*, it is possible that lower transposition rates in more recent times have led to a situation intermediate between our constant transposition and ancient burst scenarios for short elements (panels C and D respectively), weakening correlations between average frequencies and recombination.

Are TEs targeted by strong and recent positive selection in northern populations?

Because TEs can cause major regulatory changes, they may be recruited during local adaptation, especially in species encompassing a broad range of environmental conditions. If TE insertions were recruited during the recent colonization of northern environments, they should display a strong change in frequencies between the Floridian source and northern populations, and fall in regions displaying signatures of positive selection that can be detected through the use of SNP data. We first scanned all polymorphic insertions to identify a set of candidate TEs displaying high frequencies in Northern clusters and low frequency in Florida. We used two statistics to identify TEs that were potentially under positive selection, X^TX and *eBPis* [41]. X^TX is a measure of global differentiation that should be higher for markers displaying variation in allele frequencies that are not consistent with demographic expectations drawn from SNPs. *eBPis* is a complementary statistic that specifically contrasts frequencies between Floridian or Northern clusters. We identified a set of 34 insertions that were in the top 1% for both *eBPis* and X^TX statistics and showed a shift of at least 0.5 in their frequency compared to all samples in Southern Florida. We then filtered out insertions that did not fall

in a set of candidate windows displaying consistent signals of selection across three different approaches (diploS/HIC [42], BAYPASS [41] and LSD [43]; see Methods). Seventeen and 15 TE insertions overlapped with windows in the top 10% for the LSD score and BAYPASS score respectively (Table 3). Eight TE insertions fell in a window classified as a sweep by diploS/HIC. A total of six insertions were found in candidate windows for selection in all three tests (Table 3), four of them found within three distinct genes, a neurexin, *PTBP3*and *TCF-20-201*.

Discussion

Using empirical data in a model species harboring a large diversity of active TEs as well as simulations, we investigated the relative impact of selective and non-selective factors on the population dynamics of all the main TE categories active in vertebrates. We tested how the combination of linked selection in the host, direct selection against TEs and changes in transposition rate may explain heterogeneous TE frequency and abundance along the genome. By comparing the diversity of several of the most common TE categories found in vertebrates within the same organism, we clearly demonstrate that the interaction between these processes lead to sometimes drastically different outputs, even under a shared demographic history. It may be possible to disentangle these different processes using information about elements length, genomic location and frequency.

Demography shapes TE diversity across populations

We observed a clear effect of genetic drift on TE diversity across the genetic clusters examined in this study. Past work on green anole demography clearly showed that the GA and CA clusters expanded recently after a bottleneck when populations contracted to reach about 10% of their ancestral sizes [30]. This is associated with a reduction in the total number of polymorphic insertions found in these populations (Fig 1, Table 1), but also in an increase in the number of fixed elements compared to Floridian samples. Across families and clades, there were between 5 and 20% more fixed insertions in northern samples than in Florida (Table 1). This is a classical expectation: under a bottleneck, rare mutations frequently go extinct while frequent ones tend to reach fixation, leaving an excess of mutations at intermediate frequencies [44]. Fixation may also be facilitated by relatively less efficient selection due to lower effective population size, reducing $N_e s$. The strong impact of demography on TE abundance and frequencies has also been observed in a broad range of species and TE families, such as SINEs, nLTR-RTs, *Ac*-like elements and *Gypsy* in several species of *Arabidopsis* [11,45]. In *Drosophila subobscura*, recent bottlenecks may also explain the unusually high frequencies of *Gypsy* and *bilbo* elements [46].

Linked selection affects TE frequency, but does not fully explain TE density

We obtained intriguing results for SINEs, DNA transposons such as *Tc1/Mariner*, and short nLTR-RTs. Under the ectopic recombination hypothesis [36,47], which is usually invoked to explain genome-wide patterns of TE diversity, TEs tend to be removed from regions of high recombination through purifying selection. Such correlations have been commonly observed for several TE families in fruit flies and other vertebrates [7,36,48,49]. This should lead to negative correlations between recombination and TE diversity or abundance, assuming constant transposition. Instead, we observe a positive correlation between recombination and average frequency and density of polymorphic elements. Such positive correlation between allelic diversity and recombination is however a well-known feature of so-called "linked selection" [14,17]. Haplotypes harboring deleterious mutations tend to be longer in regions of low recombination, and competition between them reduces the efficacy of selection [17]. Similarly,

Table 3. Summary of the 34 TE insertions candidate for positive selection. None but two of these insertions were found in *A. allisoni* and *A. porcatus*. For each insertion, the putative length estimated by MELT is provided. LSD scores, median *eBPis* for SNPs (obtained with BAYPASS), and diploS/HIC classifications for windows containing the focal TE are given. Windows that are above the 90% percentile or classified as sweeps are highlighted with an asterisk.

Chromosome	Position	Clade	TE putative length	Nearest gene	Distance to nearest gene	Frequency in outgroups (/4)	Frequency in Florida (/30)	Frequency in North (/22)	diploS/HIC classification	median eBPis (SNPs)	LSD score	Number of tests above 90% percentile/ classified as sweeps
2	128671912	*ERV*	9190	*rapgef6*	17445	0	0	16	**Hard***	**3.86***	**0.61***	3
5	27319544	*CR1*	1751	*TCF20-201*	0	0	1	21	**Hard***	**6.38***	**0.88***	3
1	260526442	*L1*	1680	*Neurexin*	0	0	1	21	**Soft***	**5.06***	**0.87***	3
1	260803841	*LTR-RT*	859	*Neurexin*	0	0	1	21	**Soft***	**6.40***	**0.93***	3
2	57599175	*CR1*	3374	*PTBP3*	0	0	0	18	**Hard***	**4.25***	**0.72***	3
4	149706919	*CR1*	1434	*5S_rRNA*	137822	1	2	20	**Hard***	**5.06***	**0.83***	3
4	153568233	*Hobo*	1957	*COL20A1*	0	2	3	22	neutral	**3.10***	**0.74***	2
6	80481846	*Gypsy*	6269	*PRRG2*	0	0	0	19	neutral	**3.86***	**0.77***	2
2	57632946	*Helitron*	539	*PTBP3*	20057	0	0	16	neutral	**3.88***	**0.70***	2
4	110324469	*Dirs*	5787	*TESK2*	7154	0	4	22	neutral	**4.71***	**0.58***	2
1	260754973	*CR1*	1807	*ENSACAG00000005710*	0	0	2	22	neutral	**4.78***	**0.90***	2
1	89643089	*BEL*	718	*TNS1*	170148	0	0	13	neutral	**4.88***	**0.66***	2
5	26218266	*Gypsy*	1150	*L3MBTL2*	0	0	1	19	neutral	**5.36***	**0.73***	2
5	133673963	*BEL*	718	*5S_rRNA*	61322	0	0	14	**Hard***	1.87	0.55	1
6	64325965	*CR1*	921	*EFTUD2*	0	0	1	18	NA	0.65	**0.58***	1
GL343263	1240245	*Penelope*	2665	*PPL*	0	0	0	21	NA	0.91	**0.82***	1
4	10015690	*CR1*	4292	*ENSACAG00000030455*	72218	0	6	16	NA	**4.16***	0.38	1
1	182379221	*LTR-RT*	859	*SERINC1*	20124	0	0	12	neutral	1.29	**0.56***	1
4	10794679	*Gypsy*	6279	*NUDCD1*	0	0	2	22	neutral	1.83	**0.62***	1
5	45604395	*L2*	880	*PKD2*	0	0	4	20	neutral	**3.89***	0.49	1
3	104260103	*Gypsy*	969	*NAXD*	159686	0	0	12	**Soft***	1.41	0.32	1
GL343312	1468585	*Gypsy*	6277	*ENSACAG00000029426*	3958	0	0	18	NA	1.22	0.35	0
GL343285	1809184	*Gypsy*	301	*ENSACAG00000027710*	2326	0	0	12	NA	NA	NA	0
3	24577145	*Dirs*	32	*DOCK10*	0	0	0	12	neutral	0.29	-0.09	0
5	133150214	*Penelope*	2665	*ENSACAG00000029401*	21423	0	0	12	neutral	0.71	0.29	0
1	197986056	*Gypsy*	888	*ENSACAG00000026030*	8212	0	0	15	neutral	0.89	0.38	0
3	63538365	*Gypsy*	1174	*ADAM12*	9569	0	1	18	neutral	1.29	0.05	0
4	67205243	*CR1*	1944	*GABBR2*	0	0	3	21	neutral	1.42	0.46	0
5	67232158	*LTR-RT*	859	*HYAL4*	35781	0	1	21	neutral	1.49	0.08	0
5	106252550	*Penelope*	2665	*SGCZ*	11505	0	0	15	neutral	2.08	0.04	0
4	83252786	*Dirs*	5672	*VAV3*	0	0	0	13	neutral	2.10	0.29	0
3	24593897	*BEL*	718	*DOCK10*	0	0	0	12	neutral	2.11	0.30	0
GL343198	969347	*Dirs*	108	*RAB11FIP4*	0	0	0	14	neutral	2.24	0.17	0
2	53454989	*Hobo*	1914	*ENSACAG00000032857*	35068	0	0	13	neutral	2.67	0.13	0

https://doi.org/10.1371/journal.pgen.1009082.t003

the local reduction in diversity that comes with selective sweeps extends over longer genomic distances in regions of low recombination. Altogether, this leads to an effect similar to a local reduction of effective population sizes in regions of low recombination, reducing diversity and increasing the odds that deleterious alleles reach fixation. We note that these elements tend to be quite short in length, which may make them nearly neutral, and therefore more likely to be impacted by linked selection.

While some work has been done in examining whether Hill-Robertson interference between elements may increase the number of fixed insertions in regions of low recombination [16], there is not any study (to our knowledge) that examined the allele frequency spectrum of TE insertions under linked selection. In addition, the latter study considered only TE insertions and did not incorporate background selection or sweeps on SNPs. Our simulations suggest that linked selection may lead to positive correlations between polymorphic TE frequency and abundance: polymorphic TEs would stochastically reach frequencies of 0 or 1 at a faster rate in regions of lower recombination. This would therefore lead to a rise in the number of polymorphic TEs and average TE frequencies as recombination increases, but also to a reduction in the number of fixed TEs (as expected in the case of Hill-Robertson interference).

Unlike the ectopic recombination and linked-selection hypotheses, preferential insertion in regions of high recombination and open chromatin does predict a positive correlation between recombination rates and TEs density. This mechanism has been proposed to explain why LINEs and LTR-RT may be more abundant in regions of high recombination in *Ficedula* flycatchers and the zebra finch [8]. In the case of LTR-RTs, it may also be that higher recombination rates increase the frequency of solo-LTR formation, limiting their deleterious effects. It is commonly observed for several retrotransposons in a variety of species [9,50,51]. However, in humans, *L1* may actually not display strong preference for open chromatin and is more constrained by local replication timing [52,53]. In the green anole, nLTR-RTs and LTR-RTs do not display strong evidence of preferential insertion in regions of high recombination, which tend to harbor less fixed elements. We note that these families may be older in birds than in the green anole, having accumulated between 55 and 33 Mya [54], while a substantial proportion of these elements display less than 1% divergence from their consensus in green anoles (see repeat landscape at http://www.repeatmasker.org/species/anoCar.html, last accessed 25/03/2020). It is therefore possible that purifying selection has had more time to remove the most deleterious insertions in birds, increasing the signal of preferential insertion that may be masked in the green anole. It may also be that LTR-RT elements produce more frequently solo-LTR in birds than in lizards, which would make them less deleterious and more subject to drift and linked selection. Further studies at finer genomic scales will be helpful to precisely quantify how local genomic features impact TE abundance.

Our simulations suggest preferential insertion would probably not produce higher average TE frequencies in regions of high recombination. We interpret this as the fact that preferential insertion is analog to locally higher mutation rates for nucleotides: while this may affect local SNP density along the genome, it should have little effect on the shape of the allele frequency spectrum under mutation-drift equilibrium (under the assumption of infinite sites which should hold for low mutation or transposition rates [55]).

We therefore propose that SINEs, *Tc1/Mariner* and most short elements are under the influence of linked selection and preferentially insert into regions of high recombination, possibly because these are more likely to be associated with an open chromatin state. Indeed, combining these mechanisms in our simulations produced correlations matching our observations for SINEs and *Tc1/Mariner*. The average frequency of these elements was quite close from average derived SNP frequency. It is therefore unlikely that strong purifying selection acts against these elements (Figs 4, 5, 6 and 9).

For short nLTR-RTs, the negative correlation between recombination rate and the density of fixed elements may reflect a residual effect of stronger purifying selection in regions of high recombination, and/or weaker preference for regions of open chromatin. In the case of short *L1*, we observe a positive correlation between the density of polymorphic elements and recombination rate, but this correlation is weak when examining the density of fixed elements or average frequency. We note however that *L1s* are substantially longer than other LINEs (Fig 8; average length of 3700 bp, 1194 bp and 1448 bp for *L1*, *L2* and *CR1* respectively for all insertions in the dataset, Wilcoxon tests, all $P<2.2.10^{-16}$), which limits our power to study short elements.

Combination of bursts of transposition and linked selection leaves a specific signature

Sudden bursts of transposition are common in TEs, and have been documented in a variety of species [56–60]. This idiosyncrasy limits direct comparisons between TEs and SNPs, since mutation rates are usually considered constant for the latter. A general prediction is that the average frequency of elements should increase with their age, which is observed in *Drosophila* [48]. Our simulations also suggest that the positive correlation between average TE frequency and recombination rate observed for weakly deleterious TEs could be weakened and even reversed in the case of a sufficiently old transposition burst. This is due to the fact that the rarest elements have already been eliminated through drift, and the effects of linked selection lead to a faster accumulation of elements at high frequency in regions of low recombination.

We found that elements such as *hAT* and *L2* had substantially higher average frequencies, even higher than derived SNPs. For these two elements, correlations between their average frequencies and recombination rate were quite weak, even when considering only short *L2* that should be the least deleterious. This could reflect an intermediate situation compared to the extreme scenarios illustrated in Fig 9, such as multiple waves of transposition, or a younger burst than the one modeled, that may obscure correlations by flattening average allele frequencies. Examining the spectrum from more individuals may have the potential to reveal irregular transposition since local peaks in the spectrum should correspond to the age of each burst.

On the other hand, DNA transposons such as *Helitron* and *Hobo* are at very low frequencies, with almost no fixed insertion, but are more abundant in regions of high recombination. This pattern could be explained by a recent burst of transposition associated with weak purifying selection. Whether these elements share the preference of other DNA transposons for regions of high recombination remains difficult to assess due to the lack of fixed insertions.

Strong purifying selection against Dirs, LTR-RTs and long nLTR-RTs

There is evidence that strong purifying selection acts on *Dirs*, LTR-RTs and long nLTR-RTs: their average frequency is generally lower than the one of derived SNPs. Recent bursts of transposition alone may also be responsible for an excess of young, therefore rare, alleles [61]. While this seems clearly the case for *Gypsy* elements, which display many singletons and seem to be less impacted by recent demography, we also found evidence for lower average TE frequency and density of fixed TEs in regions of high recombination for long nLTR-RTs and LTR-RTs. According to our simulations, such a correlation can only be obtained through stronger purifying selection in regions of high recombination, consistent with the ectopic recombination model. For all LTR-RTs (except *Gypsy*) and long *L2*, we observed weak and even positive correlations between recombination rate and the density of polymorphic elements. This may reflect some preference for regions of high recombination compensating the loss of polymorphic elements through selection.

These results suggest that LTRs and long nLTR-RTs may be more harmful in regions of high recombination, which are also richer in functional elements. Assuming that our simulations are reasonably close from the actual processes taking place in the green anole, $N_e s$ against these elements is likely high, and possibly higher for elements with very low frequencies such as *Dirs*, *Gypsy* or *BEL*. Full-length LTR-RTs are very long elements (~5,000bp) in the green anole, that may be strongly deleterious under the ectopic recombination model. In addition, they harbor regulatory motifs that may increase their deleterious effects near coding regions. The length of an element seems to be strongly correlated with its impact on fitness, since short LINEs display a weakening and even a reversal of correlations with recombination rate. These results are consistent with the ectopic recombination hypothesis, since longer elements are more likely to mediate ectopic recombination events [7].

Strong recent positive selection on TEs is rare

Recent colonization of northern climates by the green anole may have been an opportunity for domestication of TEs, either through adaptation to the new selective pressures encountered or selection on dispersal promoting colonization of the new environment [62]. We did not find strong evidence that TEs be involved in adaptation in the northern populations. Only a few TEs displayed substantial differences in frequencies between northern and Floridian clusters. We found in total four elements that are serious candidate for positive selection, falling in introns of a neurexin gene, *PTBP3* and *TCF20-201*. *PTBP3* is involved in cell growth and erythropoiesis [63]. *Neurexins* are involved in the neurotransmitter release [64], while *TCF20-201* is a transcription factor associated with behavioral abnormalities [65,66]. While this suggests a potential impact on the nervous system and behavior, and echoes our findings from a previous study on positive selection in green anoles [62], further investigations are needed to formally validate the causal role of these elements and discard the possibility that they are only linked to a causal variant under selection. The fact that none of these elements was full-length (Table 3) makes substantial regulatory changes unlikely. Our results contrast with observations in *Drosophila*, where many TEs display steep clines in frequency that match environmental gradients and adaptive phenotypes [24,67,68]. Further investigations are needed to assess whether higher effective population sizes and more compact genome structure in *Drosophila* may explain higher rates of domestication.

There is a growing body of evidence that intrinsic properties of genomes (e.g. overdominance, Hill-Robertson effects, non-equilibrium demography) may lead to spurious signals of selection. We note that we are extremely stringent in our approach, requiring that at least three distinct tests of positive selection give a consistent signal, one of these tests explicitly incorporating demographic history in its implementation. While this could potentially limit our power to detect more subtle signals of positive selection (e.g. soft or partial sweeps), we caution against over interpreting results obtained from a single test, especially when demographic histories are complex. This is not to say that TEs are not more frequently involved over longer timescales: for example, TEs may be involved in speciation and morphological adaptation by shaping the *Hox* genes cluster in anoles [26]. Future studies on larger sample sizes may provide a more refined picture of the role of TEs in local adaptation.

Perspectives on modelling TE dynamics

We created a simple model of TE evolution that incorporated variable purifying selection against TEs, bursts of transposition, preferential insertion of TEs in regions of high recombination, and linked selection. While this model was designed as a way to illustrate how different combinations of parameters may impact correlations for the three main statistics examined in

this work, this constitutes a template for future, more detailed studies of TE evolution. For example, SLiM3 allows the incorporation of detailed maps of genomic features, complex demographic histories, multiple modes of selection, or asexual reproduction. This should facilitate the interpretation of TE diversity in species for which a reference genome is available, and improve our understanding in model species for which extensive genomic information exists. Simulated data could be used in an ABC-like approach [13], or to train machine learning algorithms [69]. Such approaches may have the power to directly quantify for each TE clade the strength of purifying selection and how other processes such as linked selection and transposition process may interact.

Materials and methods

Sampling and SNPs calling

Liver tissue samples from 27 *Anolis carolinensis* individuals were collected between 2009 and 2011 (Tollis et al. 2012), and *A*. *porcatus* and *A*. *allisoni* were generously provided by Breda Zimkus at Harvard University. Whole genome sequencing libraries were generated from these samples following the laboratory and bioinformatics procedures already presented in [30] and detailed in S1 Text. Sequencing depth was comprised between 7.22X and 16.74X, with an average depth of 11.45X. SNP data included 74,920,333 variants with less than 40% missing data. Sequencing data from this study have been submitted to the Sequencing Read Archive (https://www.ncbi.nlm.nih.gov/sra) under the BioProject designation PRJNA376071. We excluded one individual with low depth of coverage from subsequent analyses due to its large amount of missing data.

Calling TEs

We used the Mobile Element Locator Tool (MELT) to identify polymorphic insertions in the green anole genome [70]. This software performs well in identifying and genotyping polymorphic TEs in resequencing data of low and moderate coverage (5-20X), using TE consensus sequences (available at https://github.com/YannBourgeois/SLIM_simulations_TEs) to identify reads mapping to both the reference genome and the consensus. We followed the same pipeline used in previous studies [12,13], but included several clades of transposable elements covering SINEs, nLTR-RT LTR-RT and DNA transposons, using all available consensus sequences available on Repbase [71] to call TEs. Note that MELT can estimate the most likely breakpoints, insertion length, and strand for each insertion. We followed the MELT-SPLIT pathway, which consists of four main steps. First, TEs are called for each individual separately (IndivAnalysis). Then, calling is performed at the scale of the whole dataset to improve sensitivity and precision when estimating breakpoints and insertion length (GroupAnalysis). This information is then used to genotype each individual (Genotype), after what a VCF file is produced that lists all polymorphisms (makeVCF). To draw an accurate estimate of TE frequency spectra, we also used MELT-DELETION to identify polymorphic insertions found in the reference but not in all sequenced individuals. We called polymorphic TEs for each clade within the four main categories, using a threshold of 5% with the consensus sequence to attribute an element to a specific clade. The resulting VCF files were then merged for each of the four main categories considered. In case of a possible duplicate call (*i.e*. when two insertions were found at less than 2000bp from each other), only the insertion with the lowest divergence was kept. In case of equal divergence, the element with the highest calling rate was kept. We focused on TEs insertions with no missing data. While we acknowledge that these filters may be quite stringent, they should not have any impact on correlations with intrinsic genomic features and demography.

Correlations with genomic features and SNPs statistics

From the MELT output, we extracted information about the frequency of each insertion in each of the five genetic clusters found in the green anole, using the option—counts in VCFTOOLS (v0.1.14) [72]. We also estimated the number of heterozygous sites for each individual using the—012 option in VCFTOOLS. For nLTR-RTs, we extracted the length of each insertion using shell scripts. We counted the number of insertions, and the proportion of private and shared alleles for each clade using R scripts [73].

We also investigated how TE diversity correlated with intrinsic features of the genome such as the recombination rate, and statistics related to demographic processes. We focused on three commonly used statistics to describe TE diversity in each genetic cluster: the density of polymorphic TEs along the genome, the density of fixed TEs along the genome, and the average frequency of polymorphic TEs in the host's population. Note that we do not include TEs that are fixed in all 29 samples since our interest in on the most recent population dynamics. We averaged TEs frequencies and densities over 1Mb windows, a length chosen to recover enough TEs even at the clade level, while limiting the effects of linkage disequilibrium and autocorrelation between adjacent windows. Windows with no TEs or found on scaffolds not assigned to any of the six main autosomes were excluded. To estimate average TE frequencies, only windows with at least three polymorphic insertions were used. We also extracted the average effective recombination rate $\rho = 4N_e r$ in the NEF clade estimated by LDHat (v2.2) [74] in a previous study, with N_e the effective population size and r the recombination rate between two adjacent sites (see S1 Text and [30] for details). This population was chosen since it has the largest sample size and has a large, stable effective population size. This rate was divided by another estimator of the effective population size, the average number of pairwise differences ($\theta_\pi = 4N_e\mu$, μ being the mutation rate per base pair), to obtain an estimate r/μ less sensitive to local reductions in effective population sizes due to linked selection. Relative and absolute measures of differentiation such as d_{XY} and F_{ST} were also computed over 1 Mb windows, as well as the average frequency of derived SNPs in green anoles, using the two outgroups *A*. *porcatus* and *A*. *allisoni* to determine the derived alleles. These last statistics were obtained using the package POPGENOME (v2.7.5) [75]. Correlograms summarizing correlations between these summary statistics, TE frequencies, and TE densities for the four main orders were obtained using the R package corrplot. Significance and strength of correlations were assessed using Spearman's rank correlation tests. For plots of correlation, regression lines and their confidence intervals were added to improve visibility with the function geom_smooth in ggplot2 (v3.2.1) [76], using a Gaussian model for TE frequencies and a Poisson model for TE densities (which are counts per window).

SLiM3 simulations

In order to clarify how factors such as linked selection, bursts of transposition and preferential insertion of TEs may impact the three statistics examined in this study, we performed simulations using the forward-in-time simulator SLiM (v3.3.3) [40]. We modified a preexisting recipe (14.12) provided by Benjamin Haller and Philipp Messer. We simulated a 4Mb genomic fragment with parameters such as effective population size, exon density, mutation and recombination rates that were realistic for green anoles (S13 Fig). We simulated 8 diploid individuals drawn from a stable population with a N_e of one million diploid individuals, similar to the NEF clade (S1 Fig). The mutation rate for nucleotides was set at $2.1.10^{-10}$ mutation/generation/site [28]. To simulate the effects of linked selection, we set the recombination rate at 2.10^{-10} /generation on the first and last Mb of the fragment, and at 2.10^{-11} /generation in the 2Mb between. These rates encompassed those estimated with LDHat in previous studies

[30,62]. Because regions of higher recombination tend to display more exons (S6 Fig), we assigned to regions of low and high recombination 10,000 bp and 20,000 bp of coding sequences per Mb respectively. We simulated 160 bp exons (close to the average length of exons in the green anole) that were randomly placed until the desired density was reached. To explore the effects of linked selection due to deleterious and positively selected sites, we varied the proportion of nucleotide mutations under selection. In exons, we kept the proportion of new deleterious point mutation at 70% in all simulations, which seems reasonable given that dN/dS in anoles are about 0.15 [77], suggesting that most substitutions at non-silent sites are deleterious (see box 2 in [78]). To obtain the results shown in Fig 9, we assumed that deleterious mutations in exons would display a strong effects on fitness, with $2N_es$ = -100 ($s = 5.10^{-5}$). Of all new point mutations in non-coding regions, 10% were deleterious with $2N_es$ = -10 ($s = 5.10^{-6}$). There is not much information about the fitness effects of new mutations in non-coding sequences in vertebrates in general, and the green anole in particular. However, our estimate seems conservative given that in mice and humans, about 20–40% of mutations in conserved regions may have an $s > 3.10^{-4}$ [79]. To explore further how varying selective coefficients may impact our results, we examined results from simulations with $2N_es$ = -40 or $2N_es$ = -1 in 10% of non-coding sites, and $2N_es$ = -400 or $2N_es$ = -10 in coding regions (S8 and S9 Figs). We also examined a case with almost no purifying selection on nucleotides, with only 1% of non-coding sites being under purifying selection (S10 Fig). We acknowledge that positive selection may also play a major role in reducing diversity, and also explored how adding positive selection to the latter model with little purifying selection on nucleotides may restore the correlations we observed with strong purifying selection. We added positively selected substitutions, with 1% and 5% of new substitutions in coding regions with $2N_es$ = +10 (0.1% and 0.5% in non-coding regions, S11 and S12 Figs respectively)We assumed that there are 10 TE progenitors for a given TE clade in the whole genome that can jump and insert at $P = 1.10^{-3}$ elements/generation/genome at a constant rate, a value chosen to reflect known transposition rates in vertebrates and which produced a number of TEs close from our empirical observations for individual TE clades. This gave a probability of insertion in the 4Mb region of P x 4 / 1780, since the green anole genome is 1.78 Gb long. We also modelled bursts of transposition where P was set 100 times higher, but with transposition occurring only during a lapse of 100,000 years, starting 1,000,000 years ago. Note that in that latter case, TE insertions do not reach transposition-selection-drift equilibrium. Half of the newly generated elements were considered "short" and under weak purifying selection, with $2N_es$ = -0.1. The other half were considered "long", and had a stronger impact on fitness when falling in regions of high recombination ($2N_es$ = -10) than in regions of low recombination ($2N_es$ = -1). The justification for this is that long elements have a higher probability of mediating deleterious ectopic recombination events and those events are more likely to occur in regions of high recombination. Both long and short TEs falling in coding sequences were considered strongly deleterious ($2N_es$ = -2000).To improve the speed of simulations, we modelled a population of size N_e = 1000 diploid individuals, and rescaled all parameters accordingly: mutation, recombination and rates of insertion were multiplied by a factor 1000, and times in generation and selection coefficients divided by the same factor. Simulations were run over 20,000 generations to ensure that mutation-selection-drift balance was achieved for nucleotide mutations.

To account for the potential preference of elements to insert in regions of high recombination, which tend to be gene rich and are often associated with open chromatin [8,80], we also added a preference bias Q which could take the values 0.5 (TEs were as likely to insert in regions of low recombination than in regions of high recombination) or 0.7 (in that case, 70% of TEs jumping into the 4Mb region inserted in regions of high recombination and 30% in regions of low recombination). Note that values for selection coefficients and preferential

insertion were chosen to better visualize the trends that we observed across a range of other combinations, and because they produced results close from our empirical observations. The scripts used to simulate these data are available on Github (https://github.com/YannBourgeois/SLIM_simulations_TEs), and can be reused to explore in more details other combinations of parameters.

Overlap with scans for positive selection

We used the approach implemented in BAYPASS (v2.1) [41] to detect TEs displaying high differentiation in northern populations. Overall divergence at each locus was first characterized using the X^TX statistics, which is a measure of adaptive differentiation corrected for population structure and demography. Briefly, BAYPASS estimates a variance-covariance matrix reflecting correlations between allele frequencies across populations, a description that can incorporate admixture events and gene flow. This matrix is then used to correct differentiation statistics. BAYPASS offers the option to estimate an empirical Bayesian *p*-value (*eBPis*) and a correlation coefficient, which can be seen as the support for a non-random association between alleles and specific populations. BAYPASS was run using default parameters under the core model and using the matrix inferred from SNP data in [62]. We considered a TE as a candidate for selection in northern populations when belonging to the top 1% X^TX and 1% *eBPis*, and if the difference in frequency with Florida was at least 0.5.

We compared our set of candidate TEs with the results obtained from a previous study on positive selection in the same northern populations [62]. Briefly, three different methods were applied and their results compared. We first used diploS/HIC [42], which is a machine-learning approach that uses coalescent simulations with and without selection to estimate which genomic regions are more likely to be under selection. This method has the advantage of incorporating past fluctuations in population sizes, which may reduce the number of false positives due to demography. We also used LSD [43], an approach that compares genealogies along genomic windows and detects those harboring short branches in the focal population compared to its sister clades, a signal of disruptive selection. At last, we also used BAYPASS on SNP data. Further details can be found in S1 Text and [62]. The set of candidate TEs for selection was compared with the set of candidate windows for positive selection and the intersection was extracted using BEDTOOLS (v2.25.0) [81].

Supporting information

S1 Fig. Summary of population structure and environmental variation in green anoles (see [30] for further details). A: RAxML phylogeny on one million random SNPs. B: Demographic evolution of the five genetic clusters of green anoles reconstructed by SMC++ [82]. C: Sampling locations used in this study. Units for temperature are in tenth of Celsius degrees. D: PCA over environmental variables (BIOCLIM data) for the locations used in this study. Larger dots highlight the northern clades (GA and CA) and their sister Floridian clade (NEF).
(PDF)

S2 Fig. Allele frequency spectra for nLTR-RTs belonging to all five genetic clusters identified in the green anole.
(PDF)

S3 Fig. Allele frequency spectra for SINEs belonging to all five genetic clusters identified in the green anole.
(PDF)

S4 Fig. Allele frequency spectra for LTR-RTs belonging to all five genetic clusters identified in the green anole.
(PDF)

S5 Fig. Allele frequency spectra for DNA-transposon s belonging to all five genetic clusters identified in the green anole.
(PDF)

S6 Fig. Plot of the correlation between exon density and scaled recombination rate.
(PDF)

S7 Fig. Truncation of LINE elements that are assigned unambiguously to their consensus. Left: position of the start of an element relative to its consensus, reflecting 5' truncation. Right: position of the end of an element relative to its consensus.
(PDF)

S8 Fig. Summary of simulations of TEs using SLiM3. Legend is the same as Fig 9. Parameters: $2N_es$ = -40 for 10% of non-coding sites and $2N_es$ = -400 for 70% of coding sites.
(PDF)

S9 Fig. Summary of simulations of TEs using SLiM3. Legend is the same as Fig 9. Parameters: $2N_es$ = -1 for 10% of non-coding sites and $2N_es$ = -10 for 70% of coding sites.
(PDF)

S10 Fig. Summary of simulations of TEs using SLiM3. Legend is the same as Fig 9. Parameters: $2N_es$ = -1 for 1% of non-coding sites and $2N_es$ = -10 for 70% of coding sites.
(PDF)

S11 Fig. Summary of simulations of TEs using SLiM3. Legend is the same as Fig 9. Parameters: Same as S10, but includes positive selection with $2N_es$ = 10 for 0.1% of non-coding sites and 1% of coding sites.
(PDF)

S12 Fig. Summary of simulations of TEs using SLiM3. Legend is the same as Fig 9. Parameters: Same as S10, but includes positive selection with $2N_es$ = 10 for 0.5% of non-coding sites and 5% of coding sites.
(PDF)

S13 Fig. Graphical summary of SLiM3 simulation parameters. We simulate a 4Mb fragment, assuming the following unscaled parameters (see Methods for details about scaling): a stable effective population size of 1 million individuals, a mutation rate of $2.1.10^{-10}$/year, high recombination in the first and last Mb ($r = 2.10^{-10}$ /year), low recombination in the 2 Mb in the middle ($r = 2.10^{-11}$ /generation). Linked selection is modelled by introducing 10% of deleterious mutations with $2N_es$ = -10 in non-coding regions and 70% of deleterious mutations with $2N_es$ = -100 in coding regions. We assume that there are 10 TE progenitors in the whole genome that can jump P generations/genome (constant rate). We also model bursts of transposition where the probability of jumping is 100X higher, but transposition occurs during a lapse of 100,000 years, deviating from transposition-drift balance. We also add an insertion bias Q to model preferential insertion in regions of high recombination.
(PDF)

S1 Text. Supplementary Methods detailing the procedures used in previous studies to call SNPs, infer recombination rates and detect candidate regions for positive selection.
(DOCX)

Acknowledgments

We are grateful to Breda Zimkus from the Museum of Comparative Zoology Cryogenic Collection in Harvard and J. Rosado from the Herpetology Collection for providing the samples of *Anolis porcatus* and *Anolis allisoni*. We thank Marc Arnoux from the Genome Core Facility at NYUAD for assistance with genome sequencing. This research was carried out on the High-Performance Computing resources at New York University Abu Dhabi.

Author Contributions

Conceptualization: Yann Bourgeois, Imtiyaz Hariyani, Stéphane Boissinot.

Data curation: Yann Bourgeois, Robert P. Ruggiero, Stéphane Boissinot.

Formal analysis: Yann Bourgeois, Imtiyaz Hariyani.

Funding acquisition: Stéphane Boissinot.

Investigation: Yann Bourgeois, Stéphane Boissinot.

Methodology: Yann Bourgeois, Robert P. Ruggiero.

Project administration: Yann Bourgeois, Stéphane Boissinot.

Resources: Robert P. Ruggiero, Stéphane Boissinot.

Software: Yann Bourgeois, Robert P. Ruggiero, Imtiyaz Hariyani.

Supervision: Stéphane Boissinot.

Validation: Yann Bourgeois.

Visualization: Yann Bourgeois.

Writing – original draft: Yann Bourgeois, Stéphane Boissinot.

Writing – review & editing: Yann Bourgeois, Robert P. Ruggiero, Imtiyaz Hariyani, Stéphane Boissinot.

References

1. Sotero-Caio CG, Platt RN, Suh A, Ray DA. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol Evol. 2017; 9: 161–177. https://doi.org/10.1093/gbe/evw264 PMID: 28158585
2. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: From conflicts to benefits. Nat Rev Genet. 2017; 18: 71–86. https://doi.org/10.1038/nrg.2016.139 PMID: 27867194
3. Song MJ, Schaack S. Evolutionary Conflict between Mobile DNA and Host Genomes. Am Nat. 2018; 192: 263–273. https://doi.org/10.1086/698482 PMID: 30016164
4. Venner S, Feschotte C, Biémont C. Dynamics of transposable elements: towards a community ecology of the genome. Trends Genet. 2009; 25: 317–323. https://doi.org/10.1016/j.tig.2009.05.003 PMID: 19540613
5. Brookfield JFY. The ecology of the genome—Mobile DNA elements and their hosts. Nat Rev Genet. 2005; 6: 128–136. https://doi.org/10.1038/nrg1524 PMID: 15640810
6. Arkhipova IR. Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. Mol Biol Evol. 2018; 35: 1332–1337. https://doi.org/10.1093/molbev/msy083 PMID: 29688526
7. Boissinot S, Davis J, Entezam A, Petrov D, Furano A V. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci. 2006; 103: 9590–9594. https://doi.org/10.1073/pnas.0603334103 PMID: 16766655
8. Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, et al. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Mol Ecol. 2017; 26: 4158–4172. https://doi.org/10.1111/mec.14197 PMID: 28597534

9. Liu S, Yeh CT, Ji T, Ying K, Wu H, Tang HM, et al. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. 2009; 5. https://doi.org/10.1371/journal.pgen.1000733 PMID: 19936291

10. González J, Karasov TL, Messer PW, Petrov DA. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. PLoS Genet. 2010; 6: 33–35. https://doi.org/10.1371/journal.pgen.1000905 PMID: 20386746

11. Lockton S, Ross-Ibarra J, Gaut BS. Demography and weak selection drive patterns of transposable element diversity in natural populations of Arabidopsis lyrata. Proc Natl Acad Sci U S A. 2008; 105: 13965–13970. https://doi.org/10.1073/pnas.0804671105 PMID: 18772373

12. Ruggiero RP, Bourgeois Y, Boissinot S. LINE Insertion Polymorphisms Are Abundant but at Low Frequencies across Populations of Anolis carolinensis. Front Genet. 2017; 8: 1–14. https://doi.org/10.3389/fgene.2017.00001 PMID: 28179914

13. Xue AT, Ruggiero RP, Hickerson MJ, Boissinot S. Differential effect of selection against LINE retrotransposons among vertebrates inferred from whole-genome data and demographic modeling. Genome Biol Evol. 2018; 10: 1265–1281. https://doi.org/10.1093/gbe/evy083 PMID: 29688421

14. Burri R. Interpreting differentiation landscapes in the light of long-term linked selection. Evol Lett. 2017; 1: 118–131. https://doi.org/10.1002/evl3.14

15. Barron MG, Fiston-Lavier A-S, Petrov DA, Gonzalez J. Population Genomics of Transposable Elements in Drosophila. Annu Rev Genet. 2014; 48: 561–81. https://doi.org/10.1146/annurev-genet-120213-092359 PMID: 25292358

16. Dolgin ES, Charlesworth B. The effects of recombination rate on the distribution and abundance of transposable elements. Genetics. 2008; 178: 2169–2177. https://doi.org/10.1534/genetics.107.082743 PMID: 18430942

17. Hill WG, Robertson A. Local effects of limited recombination. Genet Res. 1966; 8: 269–294. PMID: 5980116

18. Charlesworth B, Charlesworth D. Elements of evolutionary genetics. Roberts and Company Publishers. 2010. https://doi.org/10.1525/bio.2011.61.5.12

19. Boissinot S, Entezam A, Furano A V. Selection Against Deleterious LINE-1-Containing Loci in the Human Lineage. Mol Biol. 2001; 18: 926–935.

20. Villanueva-Cañas JL, Rech GE, de Cara MAR, González J. Beyond SNPs: how to detect selection on transposable element insertions. Methods Ecol Evol. 2017; 8: 728–737. https://doi.org/10.1111/2041-210X.12781

21. Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, et al. Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. Am Nat. 2016; 188: 379–397. https://doi.org/10.1086/688018 PMID: 27622873

22. Jangam D, Feschotte C, Betrán E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends Genet. 2017; 33: 817–831. https://doi.org/10.1016/j.tig.2017.07.011 PMID: 28844698

23. van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 2016; 534: 102–105. https://doi.org/10.1038/nature17951 PMID: 27251284

24. González J, Petrov DA. The adaptive role of transposable elements in the Drosophila genome. Gene. 2009; 448: 124–133. https://doi.org/10.1016/j.gene.2009.06.008 PMID: 19555747

25. Bourgeois Y, Boissinot S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. Genes (Basel). 2019; 10: 419. https://doi.org/10.3390/genes10060419 PMID: 31151307

26. Feiner N. Accumulation of transposable elements in Hox gene clusters during adaptive radiation of Anolis lizards. Proceedings Biol Sci. 2016;283. https://doi.org/10.1098/rspb.2016.1555 PMID: 27733546

27. Glor RE, Losos JB, Larson A. Out of Cuba: Overwater dispersal and speciation among lizards in the Anolis carolinensis subgroup. Mol Ecol. 2005; 14: 2419–2432. https://doi.org/10.1111/j.1365-294X.2005.02550.x PMID: 15969724

28. Tollis M, Boissinot S. Genetic Variation in the Green Anole Lizard (Anolis carolinensis) Reveals Island Refugia and a Fragmented Florida During the Quaternary. Genetica. 2014; 1: 59–72. https://doi.org/10.1038/nbt.3121.ChIP-nexus

29. Manthey JD, Tollis M, Lemmon AR, Moriarty Lemmon E, Boissinot S. Diversification in wild populations of the model organism Anolis carolinensis: A genome-wide phylogeographic investigation. Ecol Evol. 2016; 6: 8115–8125. https://doi.org/10.1002/ece3.2547 PMID: 27891220

30. Bourgeois Y, Ruggiero RP, Manthey JD, Boissinot S. Recent Secondary Contacts, Linked Selection, and Variable Recombination Rates Shape Genomic Diversity in the Model Species Anolis carolinensis. Genome Biol Evol. 2019; 11: 2009–2022. https://doi.org/10.1093/gbe/evz110 PMID: 31134281

31. Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature. 2011; 477: 587–91. https://doi.org/10.1038/nature10390 PMID: 21881562

32. Ray DA, Xing J, Salem AH, Batzer MA. SINEs of a nearly perfect character. Syst Biol. 2006; 55: 928–935. https://doi.org/10.1080/10635150600865419 PMID: 17345674

33. Han KL, Braun EL, Kimball RT, Reddy S, Bowie RCK, Braun MJ, et al. Are transposable element insertions homoplasy free?: An examination using the avian tree of life. Syst Biol. 2011; 60: 375–386. https://doi.org/10.1093/sysbio/syq100 PMID: 21303823

34. Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. 2015; 25: 1656–1665. https://doi.org/10.1101/gr.196485.115 PMID: 26355005

35. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014; 23: 3133–3157. https://doi.org/10.1111/mec.12796 PMID: 24845075

36. Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol. 2003; 20: 880–892. https://doi.org/10.1093/molbev/msg102 PMID: 12716993

37. Boissinot S, Sookdeo A. The Evolution of Line-1 in Vertebrates. Genome Biol Evol. 2016; evw247. https://doi.org/10.1093/gbe/evw247 PMID: 28175298

38. Cooper DM, Schimenti KJ, Schimenti JC. Factors affecting ectopic gene conversion in mice. Mamm Genome. 1998; 9: 355–360. https://doi.org/10.1007/s003359900769 PMID: 9545491

39. Nam K, Ellegren H. Recombination drives vertebrate genome contraction. PLoS Genet. 2012; 8. https://doi.org/10.1371/journal.pgen.1002680 PMID: 22570634

40. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol Biol Evol. 2019; 36: 632–637. https://doi.org/10.1093/molbev/msy228 PMID: 30517680

41. Gautier M. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics. 2015; 201: 1555–1579. https://doi.org/10.1534/genetics.115.181453 PMID: 26482796

42. Kern AD, Schrider DR. diploS/HIC: An Updated Approach to Classifying Selective Sweeps. G3; Genes|Genomes|Genetics. 2018; g3.200262.2018. https://doi.org/10.1534/g3.118.200262 PMID: 29626082

43. Librado P, Orlando L. Detecting signatures of positive selection along defined branches of a population tree using LSD. Mol Biol Evol. 2018; 35: 1520–1535. https://doi.org/10.1093/molbev/msy053 PMID: 29617830

44. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123: 585–95. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1203831&tool=pmcentrez&rendertype=abstract PMID: 2513255

45. Hazzouri KM, Mohajer A, Dejak SI, Otto SP, Wright SI. Contrasting patterns of transposable-element insertion polymorphism and nucleotide diversity in autotetraploid and allotetraploid Arabidopsis species. Genetics. 2008; 179: 581–592. https://doi.org/10.1534/genetics.107.085761 PMID: 18493073

46. García Guerreiro MP, Chávez-Sandoval BE, Balanyà J, Serra L, Fontdevila A. Distribution of the transposable elements bilbo and gypsy in original and colonizing populations of Drosophila subobscura. BMC Evol Biol. 2008; 8: https://doi.org/10.1186/1471-2148-8-234 PMID: 18702820

47. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, Gonzalez J. Population Genomics of Transposable Elements in Drosophila melanogaster. Mol Biol Evol. 2011; 28: 1633–1644. https://doi.org/10.1093/molbev/msq337 PMID: 21172826

48. Kofler R, Betancourt AJ, Schlötterer C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. PLoS Genet. 2012; 8. https://doi.org/10.1371/journal.pgen.1002487 PMID: 22291611

49. Kapun M, Barrón M, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, et al. Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. Mol Biol Evol. 2020. https://doi.org/10.1093/molbev/msaa120 PMID: 32413142

50. Baller JA, Gao J, Voytas DF. Access to DNA establishes a secondary target site bias for the yeast retrotransposon Ty5. Proc Natl Acad Sci U S A. 2011; 108: 20351–20356. https://doi.org/10.1073/pnas.1103665108 PMID: 21788500

51. Yoshida J, Akagi K, Misawa R, Kokubu C, Takeda J, Horie K. Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. Sci Rep. 2017; 7: 1–18. https://doi.org/10.1038/s41598-016-0028-x PMID: 28127051

52. Flasch DA, Macia Á, Sánchez L, Ljungman M, Heras SR, García-Pérez JL, et al. Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. Cell. 2019; 177: 837–851.e28. https://doi.org/10.1016/j.cell.2019.02.050 PMID: 30955886

53. Sultana T, van Essen D, Siol O, Bailly-Bechet M, Philippe C, Zine El Aabidine A, et al. The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. Mol Cell. 2019; 74: 555–570.e7. https://doi.org/10.1016/j.molcel.2019.02.036 PMID: 30956044

54. Suh A, Smeds L, Ellegren H. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. Mol Ecol. 2018; 27: 99–111. https://doi.org/10.1111/mec.14439 PMID: 29171119

55. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol. 1983; 23: 183–201. https://doi.org/10.1016/0040-5809(83)90013-8 PMID: 6612631

56. de Boer JG, Yazawa R, Davidson WS, Koop BF. Bursts and horizontal evolution of DNA transposons in the speciation of pseudotetraploid salmonids. BMC Genomics. 2007; 8: 1–10. https://doi.org/10.1186/1471-2164-8-1 PMID: 17199895

57. Hellen EHB, Brookfield JFY. Transposable element invasions. Mob Genet Elements. 2013; 3: e23920. https://doi.org/10.4161/mge.23920 PMID: 23734297

58. Vieira C, Lepetit D, Dumont S, Biémont C. Wake up of transposable elements following Drosophila simulans worldwide colonization. Mol Biol Evol. 1999; 16: 1251–1255. https://doi.org/10.1093/oxfordjournals.molbev.a026215 PMID: 10486980

59. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. 2006; 16: 1262–1269. https://doi.org/10.1101/gr.5290206 PMID: 16963705

60. Manthey JD, Moyle RG, Boissinot S. Multiple and independent phases of transposable element amplification in the genomes of piciformes (woodpeckers and allies). Genome Biol Evol. 2018; 10: 1445–1456. https://doi.org/10.1093/gbe/evy105 PMID: 29850797

61. Blumenstiel JP, Chen X, He M, Bergman CM. An age-of-allele test of neutrality for transposable element insertions. Genetics. 2014; 196: 523–538. https://doi.org/10.1534/genetics.113.158147 PMID: 24336751

62. Bourgeois Y, Boissinot S. Selection at behavioural, developmental and metabolic genes is associated with the northward expansion of a successful tropical colonizer. Mol Ecol. 2019; 28: 3523–3543. https://doi.org/10.1111/mec.15162 PMID: 31233650

63. Sadvakassova G, Dobocan MC, Difalco MR, Congote LF. Regulator of Differentiation 1 (ROD1) Binds to the Amphipathic C-terminal Peptide of Thrombospondin-4 and Is Involved in Its Mitogenic Activity. J Cell Physiol. 2009; 1: 672–679. https://doi.org/10.1002/jcp.21817 PMID: 19441079

64. Missler M, Zhang W, Rohlmann A, Kattenstroth G, Hammer RE, Gottmann K, et al. α-neurexins couple Ca 2+ channels to synaptic vesicle exocytosis. Nature. 2003; 423: 939–948. https://doi.org/10.1038/nature01755 PMID: 12827191

65. Vetrini F, McKee S, Rosenfeld JA, Suri M, Lewis AM, Nugent KM, et al. De novo and inherited TCF20 pathogenic variants are associated with intellectual disability, dysmorphic features, hypotonia, and neurological impairments with similarities to Smith-Magenis syndrome. Genome Med. 2019; 11: 1–17. https://doi.org/10.1186/s13073-018-0611-9 PMID: 30609936

66. Schäfgen J, Cremer K, Becker J, Wieland T, Zink AM, Kim S, et al. De novo nonsense and frameshift variants of TCF20 in individuals with intellectual disability and postnatal overgrowth. Eur J Hum Genet. 2016; 24: 1739–1745. https://doi.org/10.1038/ejhg.2016.90 PMID: 27436265

67. González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA. High rate of recent transposable element-induced adaptation in Drosophila melanogaster. PLoS Biol. 2008; 6: 2109–2129. https://doi.org/10.1371/journal.pbio.0060251 PMID: 18942889

68. Rech GE, Bogaerts-Marquez M, Barron MG, Merenciano M, Villanueva-Canas JL, Horvath V, et al. Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. PloS Genet. 2018; 15: e1007900. https://doi.org/10.1101/380618

69. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet. 2018; 34: 301–312. https://doi.org/10.1016/j.tig.2017.12.005 PMID: 29331490

70. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. Genome Res. 2017; gr.218032.116. https://doi.org/10.1101/gr.218032.116 PMID: 28855259

71. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015; 6–11. https://doi.org/10.1186/s13100-015-0041-9 PMID: 26045719

72. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

73. R Development Core Team R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2011. https://doi.org/10.1007/978-3-540-74686-7

74. McVean G, Awadalla P, Fearnhead P. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics. 2002; 160: 1231–1241. PMID: 11901136

75. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. PopGenome: An efficient swiss army knife for population genomic analyses in R. Mol Biol Evol. 2014; 31: 1929–1936. https://doi.org/10.1093/molbev/msu136 PMID: 24739305

76. Ginestet C. ggplot2: Elegant Graphics for Data Analysis. J R Stat Soc Ser A (Statistics Soc. 2011. https://doi.org/10.1111/j.1467-985x.2010.00676_9.x

77. Tollis M, Hutchins ED, Stapley J, Rupp SM, Eckalbar WL, Maayan I, et al. Comparative Genomics Reveals Accelerated Evolution in Conserved Pathways during the Diversification of Anole Lizards. Genome Biol Evol. 2018; 10: 489–506. https://doi.org/10.1093/gbe/evy013 PMID: 29360978

78. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007; 8: 610–618. https://doi.org/10.1038/nrg2146 PMID: 17637733

79. Kryukov G V., Schmidt S, Sunyaev S. Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet. 2005; 14: 2221–2229. https://doi.org/10.1093/hmg/ddi226 PMID: 15994173

80. Myers S, Freeman C, Auton A, Donnelly P, McVean G. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet. 2008; 40: 1124–1129. https://doi.org/10.1038/ng.213 PMID: 19165926

81. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

82. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2016; 49: 303–309. https://doi.org/10.1038/ng.3748 PMID: 28024154

Received: 31 March 2022 | Revised: 13 July 2022 | Accepted: 19 July 2022

DOI: 10.1002/ppp3.10326

RESEARCH ARTICLE

Plants People Planet PPP Open Access

What lies behind a fruit crop variety name? A case study of the barnī date palm from al-ʻUlā oasis, Saudi Arabia

Muriel Gros-Balthazard[1,2] | Vincent Battesti[3] | Jonathan M. Flowers[2] | Sylvie Ferrand[2] | Matthieu Breil[3,4] | Sarah Ivorra[4] | Jean-Frédéric Terral[4] | Michael D. Purugganan[2,5,6] | Rod A. Wing[7] | Nahed Mohammed[7] | Yann Bourgeois[8]

[1]DIADE, University of Montpellier, CIRAD, IRD, Montpellier, France

[2]Center for Genomics and Systems Biology (CGSB), New York University Abu Dhabi, Abu Dhabi, UAE

[3]Eco-anthropologie, CNRS, Muséum national d'histoire naturelle, Université Paris Cité, at Musée de l'Homme, Paris, France

[4]ISEM, Institut des Sciences de l'Evolution-Montpellier, Université de Montpellier/CNRS/IRD/EPHE, Montpellier, France

[5]Center for Genomics and Systems Biology (CGSB), New York University, New York, USA

[6]Institute for the Study of the Ancient World, New York University, New York, USA

[7]Center for Desert Agriculture, Biological and Environmental Science and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

[8]School of Biological Sciences, University of Portsmouth, Portsmouth, UK

Correspondence
Muriel Gros-Balthazard, DIADE, University of Montpellier, CIRAD, IRD, 911 Avenue Agropolis, 34394/34830 Montpellier Cedex 5, France.
Email: muriel.gros-balthazard@ird.fr

Vincent Battesti, Eco-anthropologie, CNRS, Muséum national d'histoire naturelle, Université Paris Cité, at Musée de l'Homme, 17 place du Trocadéro, 75016 Paris, France.
Email: vincent.battesti@cnrs.fr

Funding information
French Agency for AlUla Development (AFALULA)

Societal Impact Statement

The oasis of al-ʻUlā is subject to a vast development operation by the central government of the Saudi monarchy. Agriculture is not strictly speaking the first objective of this initiative, the emphasis being on tourism and thus on the vast historical heritage and landscape qualities of the region. Nevertheless, agriculture and, in particular, phoeniculture remain the main resource for the inhabitants. Characterizing the local date palm agrobiodiversity is key to the sustainable development of oases. In al-ʻUlā, documenting indigenous knowledge about the locally predominant barnī variety and characterizing its genetic integrity and mode of propagation represents the essential leverage needed by farm development project planners to develop local production.

Summary

- Understanding how farmers name and categorize their crops in relation to the way they are propagated is critical for a proper assessment of agrobiodiversity. Yet, indigenous knowledge is often overlooked in genetic studies, which may result in an underestimation of crop diversity, thereby preventing its conservation and mobilization for developing sustainable agroecosystems.

Muriel Gros-Balthazard and Vincent Battesti contributed equally to this work.

Transliteration system
DIN-31635 (except API x for خ [ḫ]).
All qāf ق are pronounced gāf (IPA: g) in al-ʻUlā region.

- Here, we focus on the barnī date palm variety, a local elite variety of al-ʻUlā oasis, Saudi Arabia. We conducted an ethnobotanical survey on local phoeniculture practices and generated whole-genome data to determine whether or not barnī palms are exclusively clonally (vegetatively) propagated. Further, we contrasted the genomes of barnī and two other palms from al-ʻUlā with 112 *Phoenix* spp. to provide an initial insight into date palm diversity in this oasis.
- The survey reveals that the dates of the barnī palm bear distinct names, depending on their quality. Results show that barnī is a true-to-type cultivar, indicating clonal propagation by offshoots with name maintenance, even between distinct cultivating situations in al-ʻUlā and a nearby oasis. Nonetheless, it is distinct from the prominent barnī cultivated in Oman. Its ancestry is comparable to other West Asian date palms, but another palm from this oasis shows influence from North Africa.
- What lies behind the cultivar name barnī in al-ʻUlā and further afield in the Arabian Peninsula has been deciphered through the key disciplinary combination of social anthropology and genetics. Future studies will provide additional insights into the original genetic make-up of this millennia-old oasis.

KEYWORDS
agrobiodiversity, al-ʻUlā oasis (Saudi Arabia), date palm (*Phoenix dactylifera* L.), ethnobotany, indigenous knowledge, intra-varietal genetic variation, local categorization, population genetics

1 | INTRODUCTION

Agrobiodiversity is a proven lever for the resilience and adaptation of our food production systems (Jarvis et al., 2016). It is the spectacular product of farmers' work over millennia, and as such, it encompasses not only the genetic diversity of crops and their wild relatives but also the local knowledge and practices associated to their cultivation (Bahuchet, 2017). Yet, the biological and social dimensions of this diversity are rarely examined jointly, even though it would promote a better understanding and evaluation of it, in turn fostering its conservation and mobilization to mitigate the adverse effects of global change (Caillon & Degeorges, 2007; Gros-Balthazard et al., 2020; Leclerc & Coppens d'Eeckenbrugge, 2011).

In predominantly vegetatively (clonally) propagated crops, farmers promote the maintenance of valuable phenotypes and adaptive potential of a lineage through the use of mixed sexual/clonal reproduction systems (McKey et al., 2010). On one hand, farmers propagate interesting genotypes vegetatively, in which case a name (of a variety or cultivar) is assumed to refer to a single genotype. On the other hand, sexual reproduction creates new combinations of genes, and the resulting seedlings may be incorporated by farmers to the cultivated pool. Farmers may voluntarily select a new seedling with attractive characteristics to become a new variety, under a new name (e.g., in cassava, Elias et al., 2000 or in yam, Scarcelli et al., 2006). By contrast, farmers may incorporate seedlings under an existing variety name, if they consider that the seedling has the same phenotypic traits, thus prompting intra-varietal diversity (e.g., in oca, Bonnave et al., 2014, or in date palm, Gros-Balthazard et al., 2020). This latter practice may lead to underestimation of agrobiodiversity because variety names may not correspond to single genotypes.

So far, very few studies have explored what lies behind a crop variety name and its connection to cultivation practices. In date palm, our aforementioned study, focusing on the oasis of Siwa in Egypt, remains unique in its integration of social and biological sciences (Battesti, 2013; Battesti et al., 2018; Gros-Balthazard et al., 2020). It appears crucial to develop this pluri-disciplinary approach in other oases to establish a comprehensive evaluation of date palm agrobiodiversity, in terms of both variety names and genetic diversity.

The date palm (*Phoenix dactylifera* L.) is the main crop of desert oasis agrosystems of North Africa and West Asia, first cultivated for its fruits: the dates. All parts of the plant, however, can be used—for food, architecture, handicrafts, and so forth—and the plant itself is the oasis system engineer (Battesti, 2005). Its importance can be seen in its denominations: In Arabic in particular, its generic name is always the same (*naxla*, نخلة), but the names of the varieties—each of which has its own use, conservation, taste, harvest period, pedoclimatic needs, and so forth— are prolific. Some are well known and refer to elite commercial varieties, such as medjool (or mejhoul [*majhūl*], see Zaid & Oihabi, 2022) or khalas [*xalāṣ*], while many names are rather found locally. Inventories of varieties have been carried out locally (e.g., in Tunisia; Rhouma, 1994, 2005), but their overall number remains difficult to evaluate and could exceed 3000 worldwide (Zaid & Arias-Jiménez, 1999).

It is usually assumed that date palm varieties are vegetatively propagated by farmers (as in Krueger, 2011, and discussed in Battesti, 2013). Indeed, although this dioecious species can reproduce sexually, its multiplication is mainly carried out vegetatively by farmers, who cut and replant the offshoots growing at its base. This technique of vegetative propagation maintains the selected features, particularly that of the fruit. By contrast, palms grown from seeds have the drawback for farmers of being half males, which do not produce dates, and if females, they produce dates solely after 6–7 years, and those are typically different from that of the mother plant. Therefore, seedlings found in palm groves "merely grew by accident" (Popenoe & Bennett, 1913) and are almost always regarded by farmers as of lower quality than those of the mother plant, thus resulting in their uprooting (but see Johnson et al., 2013; Newton et al., 2013).

According to both main local and scientific narratives, a name matches a variety that matches a genotype (a true-to-type cultivar). But numerous studies have highlighted the existence of genetic variability within a name (e.g., Al-Khalifah & Askari, 2003; Al-Ruqaishi et al., 2008; Elhoumaizi et al., 2006; Sabir et al., 2014). Several hypotheses were raised to explain this variability, such as the presence of somatic mutations (Elmeer et al., 2020; Gros-Balthazard et al., 2020) or the existence of homonyms, that is, when two different lines of clones are called the same. The latter is particularly relevant given that names may refer to general features (e.g., sukarī for sugary) or a common anthroponym (e.g., nabtat saīf, Saīf's plant). Local practices of naming and categorizing palms may also promote diversity under a name. First, seedlings sharing some characteristics or qualities according to local standards, for example, fruit color or usage, may be called by a single name. The most obvious examples of such a so-called 'local category' (Battesti, 2013) are the categories of "males" or "seedlings" (the latter usually called khalt [*xalṭ*], Johnson et al., 2013). Second, a small number of clonal lines, all sharing, from farmers' local point of view, the same phenotypic characteristics (in particular the fruit), and vegetatively propagated by farmers, may be called deliberately by a single name; we coined the result of this practice as ethnovariety (Battesti, 2013; Battesti et al., 2018; Gros-Balthazard et al., 2020). Consequently, the number of cultivated genotypes in the field of this clonally propagated crop can be very different from the number of named types, which can lead to an underestimation of local agrobiodiversity if the system of naming and classification of palms remains undocumented (Gros-Balthazard et al., 2020).

In this study, we explore which kind of identity lies behind the named type "barnī" (برني), the local elite date palm of the oasis of al-'Ulā, province of Medina, in northwestern Saudi Arabia. "Barnī" is used today to designate a variety of date palm. In Saudi Arabia, it is grown in the provinces of Asir, Medina, Najran, Riyad, and Tabuk (Al-Khayri et al., 2015), but whether they belong to a single clonal line or are cases of homonymy remains to be elucidated. According to Zhang et al. (2015), for instance, the "Barni Al Madinah" date is a medium to long date, cylindrical in shape, and of brown color. This description matches the barnī cultivated in al-'Ulā, but it is unknown however if it is genetically the very same barnī. In addition, a "barni" variety is grown in Oman where it is one of the top 10 producing varieties (Al-Yahyai & Al-Khanjari, 2019), but Popenoe and Bennett noted that the "Oman variety apparently has no relation to the classical Birní of Arabia and North Africa" (1913, p. 227). This possible homonymy may be based on the antiquity of the term barnī coupled with its reputation or connotation. Indeed, barnī is one of the few (e.g., 'ajwa) date palm varieties mentioned in the collections of reports of the Sunnah of the Prophet Muhammad. In Sahih al-Bukhari, one of the collection of hadith, it is mentioned that Bilāl ibn Rabāḥ brought barnī to the Prophet (Muḥammad ibn Ismā'īl al-Bukhārī, 2312, Book 40, Hadith 12). In this saying, it is quite clear that barnī stands for a date of superior quality. The etymology of the term "barnī" is confusing. Popenoe and Bennett (1913, pp. 226–227) tentatively ventured this for the barnī variety grown in Oman: What is called Burní or Berni might derive from the name of a city named "Burn" or from the Persian "*bir*, fruit/drop" and "*nik*, good/heavy." Before them, the lexicographer E. W. Lane (1863, p. 196) compiled these same etymologies (an Arabicized Persian word) from classical Arabic authors, adding a meaning of "clay vessel" also mentioned by these authors. We can cautiously hypothesize that a variety of date may have taken the name of the pottery that preserved it (not long ago in al-'Ulā, for example, certain varieties of date were preserved and exported in paste form, see below, in goatskin called *šanna* or basketry called *mijlaḍ*).

In this study, we explore the identity of the date palm barnī in al-'Ulā and in the Arabian Peninsula. We assess whether the local identity of barnī, as given by the farmers, corresponds to a unique genetic identity (and therefore, strictly vegetatively propagated from offshoots locally) or if it refers to a multiplicity of genetic forms with an ethnobotanical survey and genetic analysis. The two approaches are essential and complementary. Indeed, an extensive ethnographical field survey will shed light on the local categorization processes and assist with designing an effective sampling strategy of the local diversity. The genetic data will permit an exploration of what lies behind the named type barnī in term of genetic variability and provide insight into how it relates to other date palms from the oasis and beyond.

2 | MATERIALS AND METHODS

2.1 | Ethnobotanic survey and sampling of the date palm named type barnī in al-'Ulā

We carried out an ethnobotanical survey in the oasis of al-'Ulā between April 2019 and November 2021, totaling more than 9 months divided into four fieldwork stays. It involved observations, non-directive and semi-structured interviews and theme-based group discussions with farmers. The purpose of the survey was to understand the local cognitive and practical relationships with plants and specifically the date palm of the named type barnī.

We sampled in situ, in collaboration with local farmers, 23 barnī palms while conducting this ethnobotanical study (Table 1; Figure 1; Figure S1). Twenty-two were identified as such by and with their owner or otherwise by their permanent manager. The last palm,

TABLE 1 Sampling of barnī and other palm accessions for the intra-varietal genetic study. We sampled 23 barnī from al-ʻUlā oasis and surroundings along with three date palms (two barnī from Khaybar, Saudi Arabia and from al-Kamil, Oman, and one mabrouma from the United Arab Emirates)

ID	Name type	Place of sampling and type of palm grove where barnī palms were sampled in al-ʻUlā
Samples of barnī from al-ʻUlā region		
barnī_00010	barnī	al-ʻUlā, Saudi Arabia (old palm grove)
barnī_00011	barnī	al-ʻUlā, Saudi Arabia (old palm grove)
barnī_00012	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00013	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00017	barnī	al-ʻUlā, Saudi Arabia (Bedouin palm grove)
barnī_00018	barnī	al-ʻUlā, Saudi Arabia (Bedouin palm grove)
barnī_00019	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00021	barnī	al-ʻUlā, Saudi Arabia (new 21st c. palm grove.)
barnī_00022	barnī	al-ʻUlā, Saudi Arabia (new 21st c. palm grove)
barnī_00023	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00024	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00025	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00026	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00029	barnī	al-ʻUlā, Saudi Arabia (new 21st c. palm grove.)
barnī_00030	barnī	al-ʻUlā, Saudi Arabia (Bedouin palm grove)
barnī_00031	barnī	al-ʻUlā, Saudi Arabia (Bedouin palm grove)
barnī_00032	barnī	al-ʻUlā, Saudi Arabia (new 21st c. palm grove)
barnī_00033	barnī	al-ʻUlā, Saudi Arabia (new 21st c. palm grove)
barnī_00034	barnī	al-ʻUlā, Saudi Arabia (new 21st c. palm grove)
barnī_00035	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00036	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00037	barnī	al-ʻUlā, Saudi Arabia (modern 20th c. palm grove)
barnī_00067	barnī	al-ʻUlā, Saudi Arabia (old palm grove)
Samples of barnī and mabrouma collected outside of al-ʻUlā		
barnī_00268	barnī	Khaybar, Saudi Arabia
barnī_Oman	barnī	al-Kamil, Oman
mabrouma	mabrouma	al-Shiwayb, United Arab Emirates
Clones used to calibrate the analyses		
barnī_00036A	barnī	al-ʻUlā, Saudi Arabia
barnī_00036r1	barnī	al-ʻUlā, Saudi Arabia
barnī_00036r2	barnī	al-ʻUlā, Saudi Arabia
ʻasēla_00169A	ʻasēla	al-ʻUlā, Saudi Arabia
ʻasēla_00169r1	ʻasēla	al-ʻUlā, Saudi Arabia
ʻasēla_00169r2	ʻasēla	al-ʻUlā, Saudi Arabia
ḏakar_00254	ḏakar	al-ʻUlā, Saudi Arabia
ḏakar_00254r1	ḏakar	al-ʻUlā, Saudi Arabia
ḏakar_00254r2	ḏakar	al-ʻUlā, Saudi Arabia

Note: Ten additional accessions (three sets of clones) were used to calibrate the analyses.

so-called "potential barnī" (barnī_00018), was sampled in a Bedouin palm grove where our Bedouin informant did not specifically point out the collected palm but stated that most of the palms there were supposed to be barnī; it was nonetheless identified as barnī by another informant who did not belong to the tribe owning the palm grove.

We voluntarily maximized both the diversity of the social criteria (sedentary and Bedouin, tribes, social groups, large and small landowners, and so forth) and the different types of palm groves (which include different farming conditions) (see Results; Figures 1 and 2) in order to verify whether some of these criteria could be explanatory.

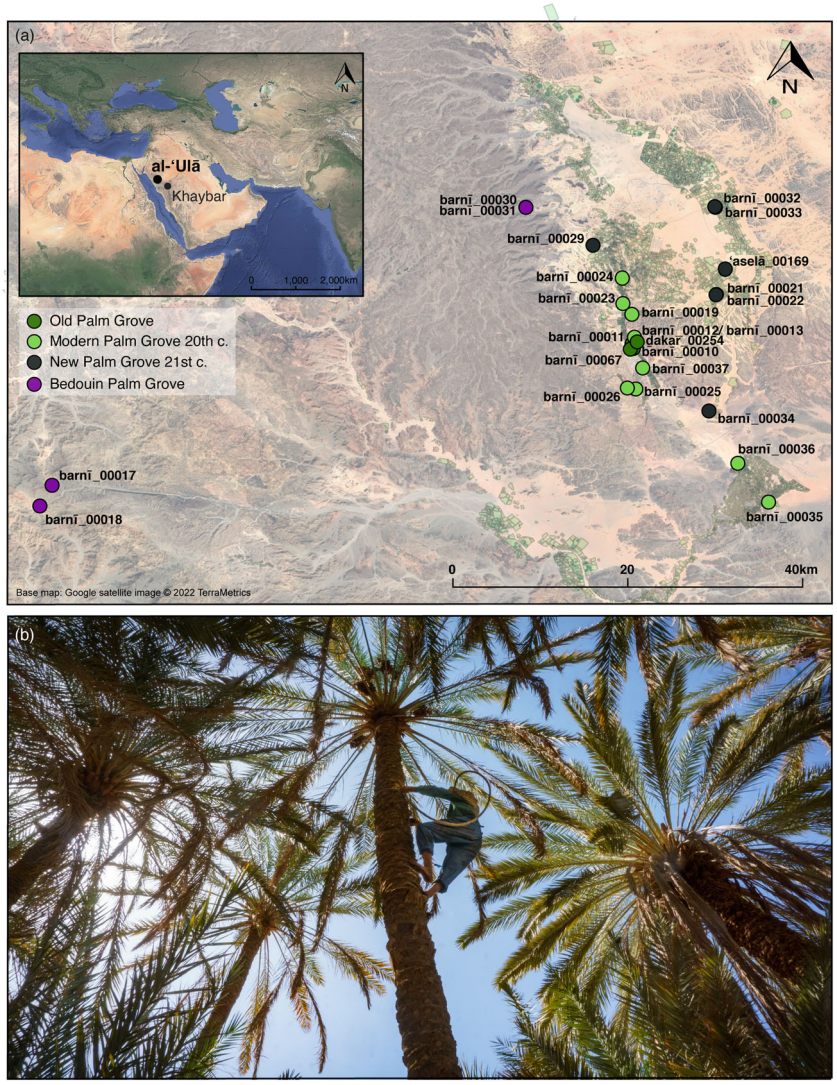


FIGURE 1 Date palm sampling. (a) Location of al-'Ulā and Khaybar oases, in Saudi Arabia, are presented in the top-left map, while the main map shows the sampling in al-'Ulā, (b) a farmer, workforce from Qena (Egypt), is climbing a barnī date palm for harvest, in al-Lutāt, an old palm grove of al-'Ulā. The same technique was used by us or the farmers to sample leaflets for analysis. October 25, 2021. Picture: Vincent Battesti

2.2 | Sampling of additional date palms and *Phoenix* spp.

In the first part of this study, we explored the intra-named type genetic variability of barnī through genome sequencing of 23 barnī palms sampled in al-'Ulā region, analyzed along with the genomes of 12 other date palms (Table 1; Figure 1; Dataset S1). Specifically, three accessions from other regions, having either the same name or a name that may also be used to refer to barnī, were included. The first one is a "barnī," identified as such by the owner, sampled in the oasis of Khaybar located 200 km from al-'Ulā (Figure 1). The second one is also named "barnī" and is grown in al-Kamil, in Oman. The third is a GenBank accession (from Hazzouri et al., 2019) of a date palm cultivated in the UAE (al-Shiwayb) and called "mabrouma" by its owner, easily associated with the name given in al-'Ulā to a category of barnī dates, *mabrūm* (see Results).

Further, three sets of clonal accessions consisting of one quartet and two triplets were sampled to calibrate the analyses aiming at identifying whether those palms represent a single clonal line (Table 1; Figure 1; Dataset S1; Figure S2). A quartet of barnī clones was constituted by sampling one accession a first time in 2019 (barnī_00036) and re-sampling the same palm 2 years later (barnī_00036A), along with two of its offshoots (barnī_00036r1 and barnī_00036r2). The two triplets were constituted by a male palm and a female variety (ḍakar_00254 and 'asēla_00169A, respectively) sampled along with two of their offshoots (ḍakar_00254r1/ḍakar_00254r2 and 'asēla_00169r1/'asēla_00169r2, respectively).

In the second part of the study, we assessed the diversity of date palms in al-'Ulā by analyzing three unique genomes from this region (barnī_000268, ḍakar_00254 and 'asēla_00169A) along with 112 *Phoenix* spp. genomes from previous studies (Dataset S1) (Flowers et al., 2019; Gros-Balthazard et al., 2017, 2021; Hazzouri et al., 2015). This includes 88 date palms from 13 countries of North Africa and West Asia, 8 *Phoenix sylvestris*, 18 *Phoenix. theophrasti*, and 1 *Phoenix reclinata*.

FIGURE 2 Photographs of four different types of palm groves: (a) old palm grove, here a *bustān*, a garden nearby the old city of al-ʿUlā oasis, April 15, 2019; (b) Bedouin palm grove in the bed of wādī Werd, about 100 km west of al-ʿUlā oasis, November 8, 2019; (c) modern 20th century palm grove, here a farm with only lined up barnī date palms, in al-Khurayba sector, north of the old palm grove of al-ʿUlā, September 23, 2021; (d) new 21st century palm grove, with a view of the agricultural farm *mazraʿa* planted in al-ʿOḏeyb district, north of al-ʿUlā, November 1, 2021. Pictures: Vincent Battesti

Overall, a total of 146 *Phoenix* spp. genomes were analyzed, of which 34 were new to this study and 112 were retrieved from GenBank SRA (Table 1; Figure 1; Dataset S1).

2.3 | DNA extraction, whole-genome sequencing, and bioinformatic processing

Genomic DNA was extracted from silica-dried leaf tissue using plant DNeasy mini kit (Qiagen, Venlo, Netherlands). Libraries (2 × 100 or 150 bp paired end) were constructed with either NEBNext Ultra II FS, Nextera DNA Flex, or Illumina truseq nano DNA library preparation kits, and sequenced on an Illumina NextSeq 550 or a NovaSeq 6000 system according to the manufacturer's protocols.

The detailed protocol for read processing, genome alignment, variant calling and filtering may be found in Methods S1. Briefly, we filtered reads based on quality and length before aligning them to the Barhee BC4 reference genome assembly (Hazzouri et al., 2019). We carried out low depth sequencing and performed population genetic analyses to identify the intra-varietal genetic variability by computing the genotype likelihoods from short read alignments. For the second part of our study, we called variants from higher coverage data using GATK v4.2.0.0 (McKenna et al., 2010) and filtered sites and genotypes based on several criteria detailed in Methods S1.

2.4 | Data analysis of intra-named type variability in the date palm barnī

Relatedness of the samples was quantified with the King-robust kinship estimator, given its robustness to SNP ascertainment bias and applicability to low-depth sequencing data (Waples et al., 2019), and calculated using NGSrelate v2 (Hanghøj et al., 2019). A principal component analysis was performed using PCAngsd v1.01 (Meisner & Albrechtsen, 2018). In both cases, genotype likelihoods were computed with ANGSD v0.933 (Korneliussen et al., 2014) using the GATK method (option -GL 2). Of note: only repeat masked annotated regions from the 18 linkage groups (Hazzouri et al., 2019) were used. Additionally, reads that did not map uniquely were discarded, and only those reads where the mate could be mapped were kept. We filtered out sites based on the following criteria: non-biallelic sites, minimum mapping quality and minimum base quality of 20, minimum number of individuals 12 (34%), minimum global depth 250 and max depth 415, minimum individual depth 5×, and SNPs with a p value < 1.10^{-6}.

Further, genetic distances among those samples were computed using ngsDist v1.0.10 (Vieira et al., 2016). ANGSD was used to compute genotype posterior probabilities with the same filtering options as above and downsampling the sites to obtain ~10,000 sites. Bootstrap replicates ($n = 100$) using blocks of 20 sites were generated, and fastME v2.1.5 (Guindon & Gascuel, 2003) was employed to compute the trees with default parameters. The *consensus* function from *ape R* package (Paradis & Schliep, 2019) was run to obtain a consensus tree where nodes found in 90% of the 100 bootstrap trees were represented. Finally, *TreeDyn* v198.3 (Chevenet et al., 2006) was used for plotting with mid-point rooting via the *phylogeny.fr* web interface (Dereeper et al., 2008).

2.5 | Data analysis of the genetic make-up of al-ʻUlā date palms

Three unique genotypes from al-ʻUlā (barnī, ʻaselā, and ḏakar) were compared with 112 *Phoenix* spp. genomes in order to have a first glimpse into the genetic diversity present within al-ʻUlā oasis. First, the structure of the genetic diversity was analyzed by estimating individual ancestries using ADMIXTURE (Alexander et al., 2009) v1.3.0 with a cross-validation of 100, and a principal component analysis (PCA) was run with the pcadapt (Luu et al., 2017) *R* package v4.3.3 filtering out SNPs with minor allele frequencies below 5%. A maximum likelihood tree was generated using RAXML-NG (Kozlov et al., 2019) v0.9.0. To do so, 20 maximum likelihood tree searches were performed using 10 random and 10 parsimony-based starting trees. The best scoring topology was picked and checked for robustness by performing 100 bootstrap replicates.

The fraction of heterozygote sites for each date palm accession was calculated using pixy (Korunes & Samuk, 2021). Finally, admixture tests were performed using the admixr *R* package v0.9.1 (Petr et al., 2019) (Methods S2).

To gain insight into the maternal origins of the date palms from al-ʻUlā, a bootstrapped chloroplast DNA tree was constructed using the Neighbor-joining method with the phangorn v2.8.1 package in *R* (Methods S1). The tree was rooted with *Phoenix reclinata* (PREC1).

Statistical analyses and plotting were conducted with the *R* Statistical Programing Language (R Core Team, 2022).

3 | RESULTS AND DISCUSSION

Our study focused on the date palm locally named barnī, the local elite date palm of al-ʻUlā oasis, Saudi Arabia. We first performed an ethnobotanic survey to both better understand folk categorization in conjunction with local date palm agrobiodiversity and set up an in situ sampling methodology with the essential cooperation of the local farmers. We then performed whole-genome sequencing of barnī date palms from al-ʻUlā region in order to assess whether this name refers to a unique clone, an ethnovariety or a local category with multiple genetic identities. By adding two barnī date palms from outside of al-ʻUlā (i.e., Khaybar, Saudi Arabia and al-Kamil, Oman) and one "mabrouma" (*mabrūm* is the name given in al-ʻUlā to a quality of barnī dates; see below) from the UAE, we further explored the genetic variability of barnī at the scale of the Arabian Peninsula. Finally, we studied three date palms from al-ʻUla, including a barnī, along with 112 *Phoenix* spp. to obtain a first glimpse into the genetic makeup of this millennia-old oasis.

3.1 | The cultivation of date palm in the oasis of al-ʻUlā

Our anthropological survey (a 10-month field survey in 2019–2022) of al-ʻUlā oasis and the region highlighted different types of palm groves, each with its own social and spatial organization, space, and cultural practices, depending on their location, their history and the social group that exploit and own them (Notes S1; Figure 2). To summarize, two of them reflected historical growing situations, namely, the date palms grown in the subsistence-type gardens (*basātīn*) of the palm grove near the old city of the oasis and those found in the Bedouin palm groves scattered in desert wadis outside the oasis (Figure 2a,b). The two other grove types were more recent and commercial: the modern palm groves established during the 20th century outside of the historical core area, but in its immediate vicinity and the more recent gardens established during the 21st century beyond the perimeter of the old and modern palm groves (Figure 2c,d).

Our fieldwork was performed among all local social categories with several hundred farmers being interviewed in Arabic following the ethnographic methodology. This ethnobotanical survey suggests a very rich agrobiodiversity for date palm alone, drawing up a complex picture of variety names ($n = 99$, at this stage of our survey). Farmers classify the two million palms in the region into categories and assign names to them by consensus based on local criteria and shared features. We found that they typically propagate their palms by offshoots, as is usual in palm groves in the Sahara and Arabia (Munier, 1973), but we lack assurance that the clonal propagation has been applied consistently and systematically throughout al-ʻUlā region for all varieties. As a matter of fact, we witnessed reproduction by seed and analyzed emic discourses of palm biology that enable such practices, but we do not yet know the extent to which this technique is used (in practice and over time): it may have emerged in recent decades due to a less extensive knowledge of date palms by newcomers to phoeniculture (sedentarized Bedouins in particular, as declared by both the social group of oasis sedentaries and the Bedouins themselves).

3.2 | The barnī, a socialized date palm in al-ʻUlā

3.2.1 | A local elite variety, but a recently increased supremacy over the local date agrobiodiversity

The ethnobotanical survey conducted on date agrobiodiversity in al-ʻUlā clearly revealed the special and shared status, today, of the named type barnī. Indeed, in al-ʻUlā, there is "the barnī" and "the rest," *al-bāqī* (الباقي), or "the [other] varieties," *al-aṣnāf* (الأصناف). It is by far the most cultivated variety and is found in all four agricultural contexts described above (Figure 2). It also is the most exported date variety of this oasis. It is considered by all farmers hardy and local and, as such, enjoys an elevated status among the inhabitants and is perceived to be a superior fruit and crop and to grow better. The alternate local elite variety, presently second in rank for all farmers, is the variety named ḥelwa (and when it is necessary to specify, the ḥelwa ḥamrā' to distinguish it from another local variety, the ḥelwa beyḍā'). Together, barnī and ḥelwa constitute the main varieties in the oasis and are usually the only two reported to be grown in Bedouin palm groves in the region.

The orientalists Jaussen and Savignac (1914, p. 40) noted in the early 20th century the elite status of the barnī and ḥelwa: "Dans

l'oasis d'el-'Ela on cultive principalement le palmier qui est la grande ressource du pays. On distingue ici deux principales espèces de dattes, les dattes douces, *ḥelweh* (حلوة) et les dattes suaves, *barnyeh* (برنية)." Similarly, Nasif (1988, p. 174) stated that "the best-known and best-liked of the various types of date produced is the ḥulwah or sweet date. (…) The barnī follows the ḥulwah in popularity." This barnī variety seems to have been the reference variety for payment by farmers (in kind, with grain as well) for the work of the *mu'allim*, the person responsible for sharing irrigation water (Nasif, 1988, p. 249). The barnī hence played a central role in agricultural life.

Actually, a distinction must be made between two issues: a numerical supremacy (the case of barnī) and a shared preference (the case of ḥelwa). Local accounts collected during our ethnobotanical survey corroborate the population's shared preference for the taste of ḥelwa over the barnī, especially among elders. It was previously noted by Doughty (1921: p. 153) that the ḥelwa was the most valued, tasted like honey, and widely exported with the pilgrimage. He added that the barnī constituted the "cheaper household food" of al-'Ulā. We should probably understand that dates in general, and among them especially the barnī, were the staple food of the oasis inhabitants. The above-mentioned export of ḥelwa is no longer what it once was. In the past, according to local narratives, dates were exported in the form of paste, which suited the soft ḥelwa well. Today, individual fruits are preferred over paste in export markets thus making the semi-soft barnī a higher value commodity. Consumer preferences outside of al-'Ulā have changed and the export market has been more supportive of barnī than ḥelwa.

This explains the very broad numerical supremacy that barnī date palms have acquired today in al-'Ulā palm groves. Its election is undoubtedly not recent, but might have been largely amplified by the administration and the market from the second part of the 20th century. Government incentives—notably through the conditions for agricultural loans for "modern palm groves" which target export (particularly, it seems, towards Turkey's market)—have pragmatically favored a barnī monoculture. Meanwhile, the inhabitants have diversified their consumption and diet (more rice than dates). Our survey reveals in parallel that local people substituted (sometimes clearly uprooting and replacing) rare local varieties with the elite barnī. Local farmers of the sedentary group recounted their memories of ḥelwa date palms (for instance) that were purposefully uprooted to plant barnī offshoots (we even witnessed it), and clearly identified that as a consequence of a market appeal. The supremacy of barnī has been further increased with the considerable extension of the area cultivated in palm groves in the region over the last decades, and particularly in varietal monoculture of barnī.

3.2.2 | Naming a palm and naming its dates according to their quality

The elite status of the barnī is also reflected by the fact that the palm, *the* barnī, produces dates that are not named/sold under this name. According to our study, in al-'Ulā, three names are used to sort its dates depending on date quality (*jawda*) for marketing (Figure 3): #1 *mabrūm*: the best quality; #2 *mašrūk*: almost good, but of lower quality (smaller, and has a *qešra* "skin," i.e., a more wrinkled epidermis); #3 *'ādī*: good only for livestock, damaged with too much "skin" (a cracked epidermis; the term *'ādī* here refers to the notion of "ordinary," "worthless"). The second quality or the merged second and third qualities are sometimes also referred to as the name *abū qešra* (because of that whitish skin, *qešra*, epidermis). In addition to the shape, the tastes differ with the qualities, according to most farmers. The *mabrūm* are said to taste better, be more presentable, and sell for much more (Notes S2). These different commercial

FIGURE 3 Different quality of the dates harvested from a barnī date palm. (a) A date bunch where various qualities of dates can be spotted. Ṣadar palm grove, August 24, 2021. (b) The three different qualities in the hand of a farmer. On the right, the *mabrūm* (the best quality), in the middle, the *mašrūk* (almost good, but of lower quality), and the *'ādī* (good only for livestock, damaged with too much "skin"). Jabāna/Muġeyra palm grove, August 17, 2021. (c) Team of workers busy sorting the harvest of barnī date palms, in al-Xaṭīb palm grove in al-'Ulā. In the foreground are boxes of the best dates, the so-called *mabrūm*. November 2, 2021. Pictures: Vincent Battesti

qualities of dates are present in the same bunch of the same date palm. It is said that the older the barnī palm, the greater the proportion of *mabrūm* in its yield. Besides, farmers say that this proportion varies across years, and since this variety now dominates the date economy (Notes S2), it is a sign of a bad or good phœnicultural harvest.

That barnī fruits go by different names seems to have already been noticed by Doughty a century ago (1921, p. 153), who stated that "there are many kinds." This is best explained later by Nasif (1988, p. 174), who highlighted that "this date is sometimes divided into two kinds, when the better quality of this type is selected because it is without a skin; this class of the barnī is known as mabrūm."

The case of naming a palm and its dates differently is unusual but not unique: One example is in Siwa (Egypt, a Berber-speaking oasis) of a variety, also elite, the tasutet palm whose dates are named ṣaɛidi (Battesti, 2013). Further, naming dates of the same variety according to their quality is also uncommon, although the case is not entirely unique either. For instance, in Siwa, the second elite cultivar (after the above-mentioned ṣaɛidi) is the true-to-type alkak date palm whose local name depends on fruit quality which varies according to growth conditions and age. The higher alkak date palms are called "alkak n amles," meaning bearing smooth or wrinkle-free alkak dates, while the lower alkak date palms bearing smaller dates, and three times cheaper, are called "alkak nifuġen" (Gros-Balthazard et al., 2020). The case is not entirely analogous to al-ʻUlā though: in Siwa, the palms bear either of these two names, and so does their production, while for the barnī, it is the dates from the same palm that are given different names.

3.3 | What lies behind the name barnī?

To study the genetic variability within the barnī, we sequenced the genomes of 35 date palms, including 23 barnī from al-ʻUlā collected in the four agricultural contexts (Figures 1 and 2), and obtained between 18 and 122 million of paired reads of size >76 bp (Table 1; Dataset S2). Sequencing reads from these runs and those retrieved from GenBank SRA were mapped to the Barhee BC4 genome assembly (Hazzouri et al., 2019), resulting mostly in low coverage alignments, ranging from 2.9× to 20.5× (9.3× on average; Dataset S2).

We assessed whether 26 barnī and mabrouma date palms from Saudi Arabia, Oman and the UAE are genetically identical using the King-robust kinship estimator (Waples et al., 2019) (Figure 4a; Dataset S3), a principal component analysis (Figure 4b), and a tree based on genetic distances (Figure 4c; Figure S3). In all three cases, 305,610,249 sites across the 18 linkage groups of the date palm genome were analyzed by ANGSD. After filtering, 1,508,939 sites

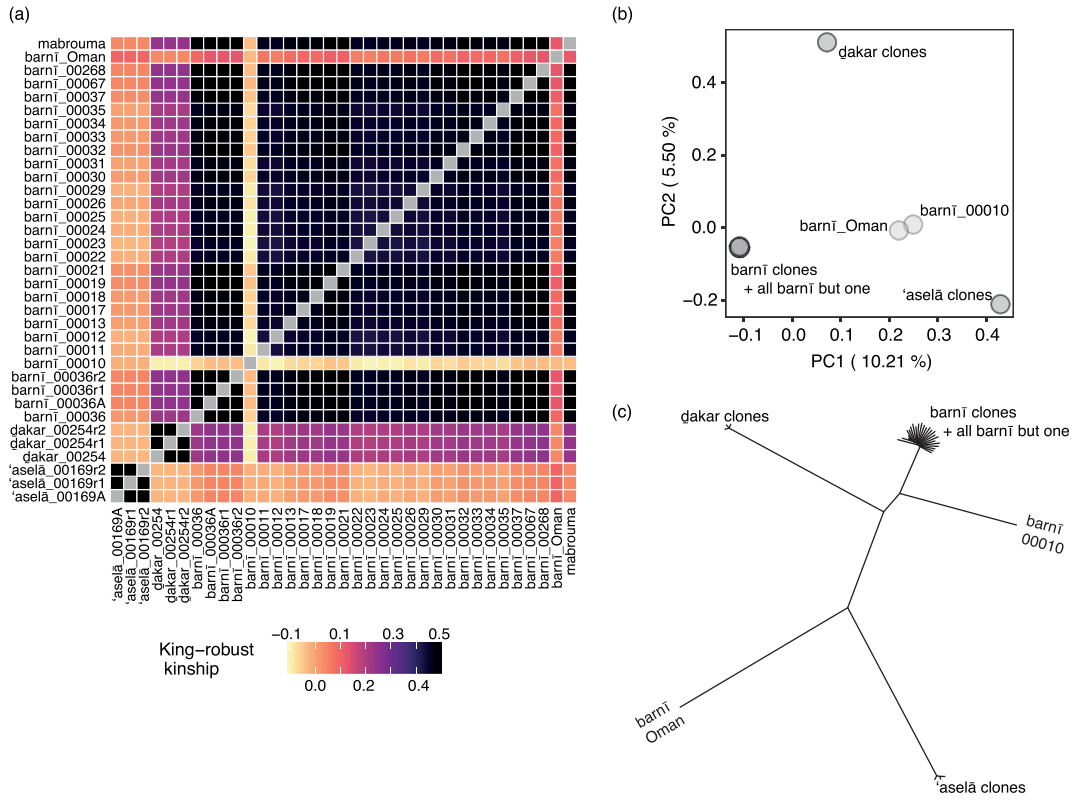


FIGURE 4 Intra-named type variability of the barnī date palms. (a) Heatmap of the king-robust kinship estimator calculated across 1,508,939 sites in 35 date palms; (b) principal component analysis of 35 date palms (1,508,939 sites). Variance explained by each principal component (PC) is provided within parentheses; (c) tree of genetic distance calculated across 10,742 sites in 35 date palms. The consensus tree obtained from 100 bootstrap replicates may be found in Figure S3.

were retained for both the relatedness analysis and PCA analysis, and we further downsampled these sites to get 10,742 sites for computing genetic distances while limiting the effects of linkage.

Low coverage sequencing may yield kinship estimates that differ from the expectation of 0.5 for members of a single clone particularly for heterozygous samples such as date palms. We therefore sequenced three clones multiple times to assess the deviation from 0.5 that can be expected using our approach to kinship estimation (Figure S2). The barnī we sampled and sequenced twice revealed a King-robust kinship of 0.482, while kinship estimates for the mother palm and two offshoots ranged between 0.481 and 0.499 (Figure 4a; Dataset S3). Sequences of two clone triplets (ḏakar and ʻaselā) that each consisted of a palm tree and two of its offshoots each yielded a kinship estimate of 0.499. Since these samples are known to be clonal, we attribute any differences in kinship estimates from the expectation of 0.5 to be attributable to our low coverage approach. In the PCA, the accessions from each triplet/quartet of clones overlap (Figure 4b), while they group together in separate clades in the genetic distance tree (Figure 4c; Figure S3), as expected given their genetic identity.

3.3.1 | Barnī is a true-to-type cultivar in al-ʻUlā region

We found that the 23 barnī from al-ʻUlā are genetically identical, except barnī_00010. Indeed, those 22 accessions cluster together in both the PCA (Figure 4b) and the tree (Figure 4c; Figure S3). Further, the King-robust kinship among them ranges from 0.424 to 0.491, nearing the theorical 0.5 expectation and the 0.481–0.499 range observed among the known clones (Figure 4a; Dataset S3). On a technical note, we hypothesized that lower kinship estimates among the 22 barnī, compared with the known clones, may be due to a lower coverage in the former (Dataset S3). We tested the relationship between the fraction of sites with information for two individuals (used to calculate pairwise kinship) and the King-robust kinship estimate, and indeed found that they are highly positively correlated (Notes S3; Figure S4).

To understand why barnī_00010 is genetically different, we returned to the field and found that it had been misidentified by the farm manager (a foreign worker) at the time of collection. The error is attributable to his lack of knowledge of the planting history of the sample as well as the youthfulness of the palm which made it difficult to identify using botanical characteristics. This demonstrates the need for meticulous sampling pre-informed by ethnobotanical and anthropological study.

The samples of the named type barnī were voluntarily collected from the different categories of palm groves of al-ʻUlā and these palm groves represent a great diversity of management, farming practices and social origins (Figure 2; Notes S1). Nevertheless, these differences did not lead to the selection of an ethnovariety. Indeed, all barnī whether from old, modern (20th c.) or new (21st c.) palm groves are identical. More intriguing, the four barnī from Bedouin palm grove (including the so-called "potential barnī" barnī_00018) also are identical to those found in the oasis palm groves. Those Bedouin palm groves found in the Balawī tribal territory about 100 km west of al-ʻUlā are of the picking palm grove type, i.e. characterized by very little labor investment, a monoculture of a few barely pollinated palms, in the bottom of the wādī, thus without necessary irrigation and without permanent habitation. This suggests a circulation, difficult over great distances, of palm offshoots between the apparently antagonistic Bedouin and sedentary worlds. The local saying that refers the Bedouin to "*ibel w ġanem*" (dromedary and sheep) and the sedentary to *naxla* (the date palm) does suggest very different ways of life and mode of production, and consequently distinct domains of knowledge: Bedouins are less knowledgeable about date palms than oasian people, but apparently enough to maintain a clonal lineage in their picking palm groves (Notes S1).

This result—barnī is a true-to-type cultivar—was, in a sense, expected, since the mode of reproduction is supposedly vegetative in al-ʻUlā (according to our field survey, in spite of some discordant evidence of cases of seed propagation). Besides, barnī is a local elite variety and we have previously hypothesized that local elite date palms are likely to be true-to-type cultivars, despite their prevalence and therefore the mechanical probability of becoming an ethnovariety (Gros-Balthazard et al., 2020). The system of ethnovariety and local category would allow farmers to organize the diversity of palms with lower commercial value (but potentially high local value) while not multiplying the denominations for the same characteristics (i.e., not creating named types for palms that are locally considered the same); it offers a fairly flexible management of agrobiodiversity (Gros-Balthazard et al., 2020).

On a practical note, processing and distributing (including identification and traceability) a heterogeneous or a homogeneous commercial product obviously differs. Our clarification of the status of this local elite variety in al-ʻUlā, covering thousands of hectares, is thus of great importance regarding the agronomic and economic development of the phoenicultural sector in this region. The barnī of al-ʻUlā is not only precisely identified by the farmers and multiplied strictly by offshoots, even by Bedouins, but can now also be easily identified through genetic fingerprints.

On a separate note, we found that the male accession used to calibrate the genetic identity analyses (ḏakar_00254) is closely related to barnī, with a King-robust kinship estimate ranging from 0.194 to 0.284 (Figure 4; Dataset S3). Since barnī is by far the most cultivated named type in al-ʻUlā and that males mostly arise from accidental seedlings, this male is probably an offspring of barnī.

3.3.2 | The date palm barnī across the Arabian Peninsula

We compared the barnī of al-ʻUlā with two barnī date palms from outside this oasis (Khaybar, Saudi Arabia and Nizwa, Oman) and with a palm identified as mabrouma (name given in al-ʻUlā to the best date category of barnī, see above; Figure 3) in a private collection from the UAE.

We found that the barnī from Khaybar, a sedentary oasis palm grove 200 km away, is genetically identical to the barnī from al-ʻUlā, indicating that the named type barnī refers to the same entity not only at the scale of al-ʻUlā oasis, but possibly at the scale of the region. The mabrouma from a palm grove in the UAE is also identical to the barnī of al-ʻUlā. It originates from a collector's farm (the exotic character is often valued by collectors). Perhaps originating from al-ʻUlā, this palm was supplied to the owner under the name "mabruma," possibly in reference to the *mabrūm* quality of its dates.

On the other hand, the barnī from Oman turns out to be different from that of al-ʻUlā (Figure 4). This alternate barnī appears to be present in Oman, mainly in the Northern and Southern al-Sharqia regions and is the tenth most cultivated variety in the country (Al-Yahyai & Al-Khanjari, 2019; Al-Yahyai & Khan, 2015). Although Popenoe and Bennett (1913, pp. 226–227) recognized the probable difference between these barnī, our genetic study confirmed this example of homonymy.

3.4 | A first glimpse into the date palm agrobiodiversity in al-ʻUlā

In Northwestern Saudi Arabia, date palms have dominated the oasis agricultural system since at least the fourth century BCE (Bouchaud, 2013), and al-ʻUlā region is known to be home to the two-millennia-old Nabataean site of Hegra (Madāʾīn Ṣāleḥ). In this area, dates have been consumed since at least the end of the 2nd millennia BCE (Rohmer et al., in press) and they have a particular symbolic importance in the Nabatean period as attested by the date necklace excavated in a tomb of this period (Bouchaud et al., 2015). The region is positioned in a strategic location, on a critical trade route, namely, the incense road, connecting the South of the Arabian Peninsula with the Levantine region roughly during the 7th century BCE and the 2nd century CE. It is also at the crossroads of date palm diversity, between the distinct North African and West Asian genepools (Flowers et al., 2019; Hazzouri et al., 2015).

So far, no studies have focused on the origin and extent of the diversity of Northwestern Saudi Arabia date palms, including al-ʻUlā oasis. A few studies focused on the genetic diversity of date palms varieties in Saudi Arabia (Aleid et al., 2015; Al-Khalifah & Askari, 2003; Al-Qurainy et al., 2011) but how those varieties, some potentially cultivated in al-ʻUlā, relate to cultivars from other regions remain to be elucidated.

Here, we analyzed the genome of the barnī of al-ʻUlā and of two other accessions from this oasis, one of the ʻaselā variety (ʻaselā_000169A) and a male (ḏakar_00254), along with that of 112 *Phoenix* spp. (Dataset S1). Given the genetic identity of all barnī palms from al-ʻUlā, we picked the accession with the highest coverage, namely barnī_000268, for downstream analyses. Aligning reads to the Barhee BC4 genome led to an average coverage of 17.3x across accessions (Dataset S2) and we identified 1,007,281 SNPs after quality filtering, which we used in subsequent analyses (Methods S1).

3.4.1 | Genetic relationships between date palms from al-ʻUlā and from North Africa and West Asia

The relationships and the genetic structure of the 3 date palms of al-ʻUlā and 88 date palms from North Africa and West Asia were determined by performing model-based genetic clustering (Figure 5a; Figures S5 and S6), reconstructing a phylogenetic tree (Figure 5b) and applying a PCA (Figure 5c; Dataset S4) using the SNP data. All three analyses (Figure 5) corroborated previous results on date palms, that is, that they can be split in two main clusters (North Africa and West Asia) with Egyptian accessions being a mix between the two (reviewed by Gros-Balthazard & Flowers, 2021).

Regarding the diversity in al-ʻUlā, our results revealed that it may represent a unique mixing among West Asian date palms (Figure 5): The three al-ʻUlā palms indeed cluster with West Asian date palms, but the male ḏakar shows influence from North African diversity. It is found on the edge of the West Asian group in the PCA, close to North African date palms (Figure 5c) and shows mixed ancestry from the West Asian and the North African clusters in the clustering analysis (Figure 5a).

3.4.2 | Genetic diversity of al-ʻUlā date palms

To determine the genetic diversity of al-ʻUlā date palms, the proportion of heterozygous sites in each date palm accession was calculated using pixy (Korunes & Samuk, 2021), which confirmed results from previous reports (Flowers et al., 2019; Gros-Balthazard et al., 2017, 2021; Hazzouri et al., 2015): i.e., North African date palms display a higher diversity (mean heterozygosity 0.094 ± 0.016%) than cultivated West Asian date palms (mean heterozygosity 0.062 ± 0.0069%; one-sided Wilcoxon rank sum test, $W = 1243$, $P = 1.11 \times 10^{-11}$). Regarding the three al-ʻUlā date palms analyzed in this study, their diversity (mean heterozygosity 0.065 ± 0.0064%) was on average comparable to that of cultivated West Asian date palms (Wilcoxon rank sum test, $W = 99$, $P = 0.41$), and lower than that found in African date palms (one-sided Wilcoxon rank sum test, $W = 3$, $P = 0.0021$).

3.4.3 | Evidence of gene flow between al-ʻUlā date palms and the wild relative *P. theophrasti*

Interspecific introgression has shaped the diversity of North African and Levantine date palms since at least 2000 years (Flowers et al., 2019; Gros-Balthazard et al., 2021; Pérez-Escobar et al., 2021). The date palm was presumably domesticated in the Persian Gulf region during the fifth millennium BCE and its cultivation then spread across Arabia and further across North Africa. Modern North African date palms and ancient Levantine date palms show higher genetic diversity than that found in West Asia, which may at least partially be explained by gene flow from a wild relative species, that is, *Phoenix theophrasti*, whose present-day distribution includes Crete and the coast of

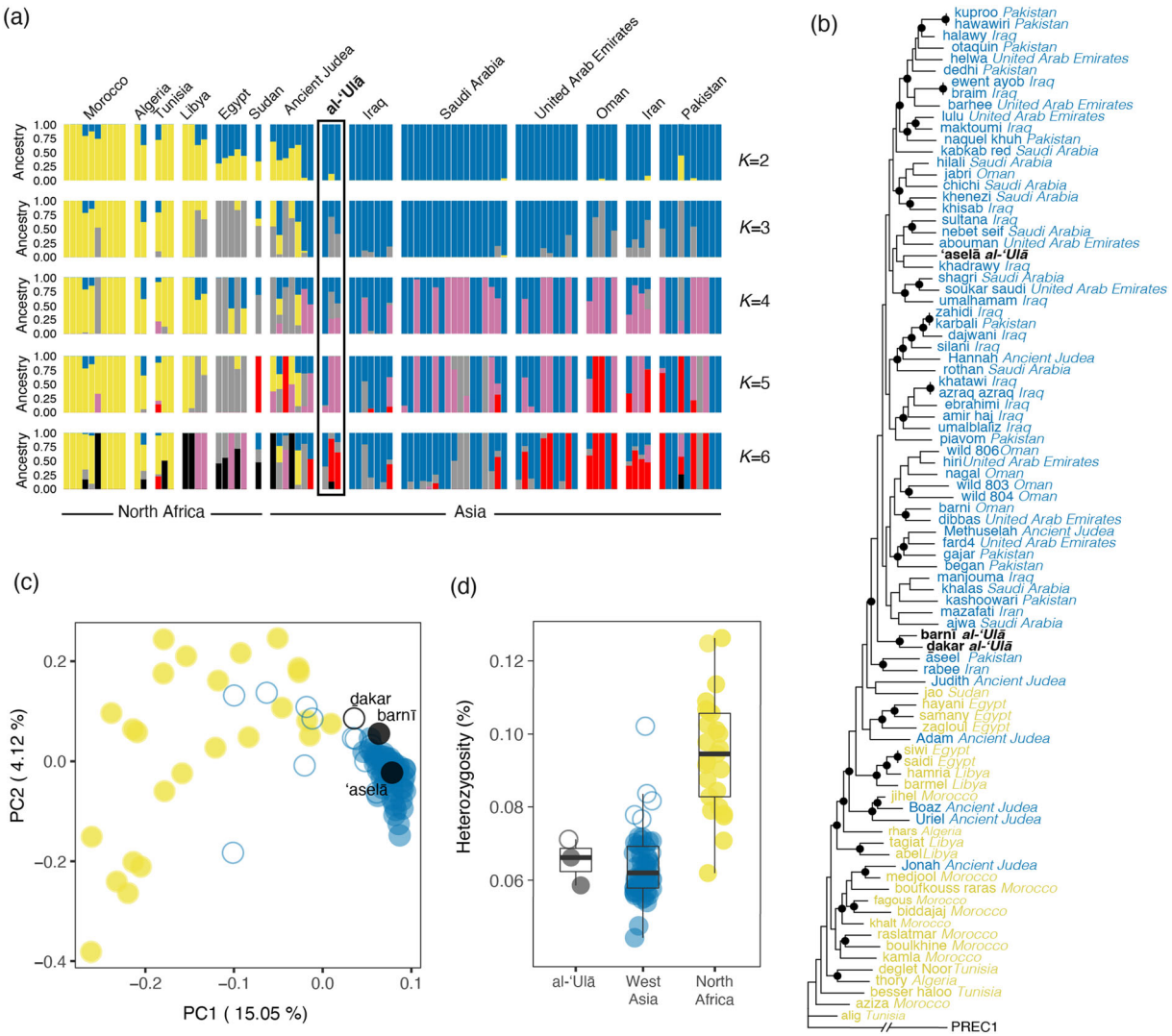


FIGURE 5 Structure and diversity of 88 date palms from North Africa and West Asia, including al-'Ulā. (a) Ancestry coefficients from the West of the distribution to the East inferred from 28,406 single nucleotide polymorphisms (SNPs). The date palms from al-'Ulā are ordered as follow (from left to right): 'aselā, ḍakar, and barnī. Labels and additional *K* values may be found in Figure S4; (b) phylogenetic relationships inferred from 50,171 SNPs. The tree was rooted with *Phoenix reclinata*. Black circles indicate nodes with >80% bootstrap support. Labels in yellow correspond to North African date palms, in blue to West Asian date palms, and in black to those collected in al-'Ulā; (c) principal component analysis inferred from 14,137 SNPs. Black dots correspond to the date palms from al-'Ulā. Yellow dots correspond to North African date palms. Full blue dots correspond to West Asian date palms, the empty ones are those having a fraction of their ancestry from the African cluster in (a). The variance explained by each principal component (PC) is given in parentheses. Coordinates on PC1 and 2, and additional PCs may be found in Dataset S4; (d) fraction of heterozygote sites across 88 date palm genomes calculated from 105,236,083 sites including 1,007,281 SNPs. Same legend as panel (c)

Turkey, or a *theophrasti*-like species (Flowers et al., 2019; Gros-Balthazard & Flowers, 2021). The oasis of al-'Ulā has long been connected to the Levantine region and i1ts ports on the Mediterranean Sea. Whether the genetic make-up of its date palms also displays ancestry from *P. theophrasti*, reflecting this possible bond, remains to be determined.

To evaluate whether the date palms from al-'Ulā showed evidence of introgression, both *D*- and f_4-statistics were computed (Figure 6; Methods S2; Datasets S5-S6). This revealed that both barnī and 'aselā do not show any evidence of introgression (Figure 6). On the other hand, North African date palms, and the ancient Judean date palms, that were previously shown to be admixed, displayed significant positive *D*- and f_4-statistics (Figure 6; Flowers et al., 2019; Gros-Balthazard et al., 2021). Interestingly, this is also the case for the male ḍakar_00254 from al-'Ulā (Figure 6).

We further estimated the fraction of *theophrasti* ancestry in the date palms showing evidence of admixture according to *D*- and f_4-statistics (Figure 6; Dataset S7), including the male ḍakar_00254. We found that the latter displays about 1.1% of its genome from *Phoenix theophrasti*. This is in the lower range of what is observed in North African and Judean date palms (average 11.23%, ranging from 0.59 to 21.58%; Dataset S7).

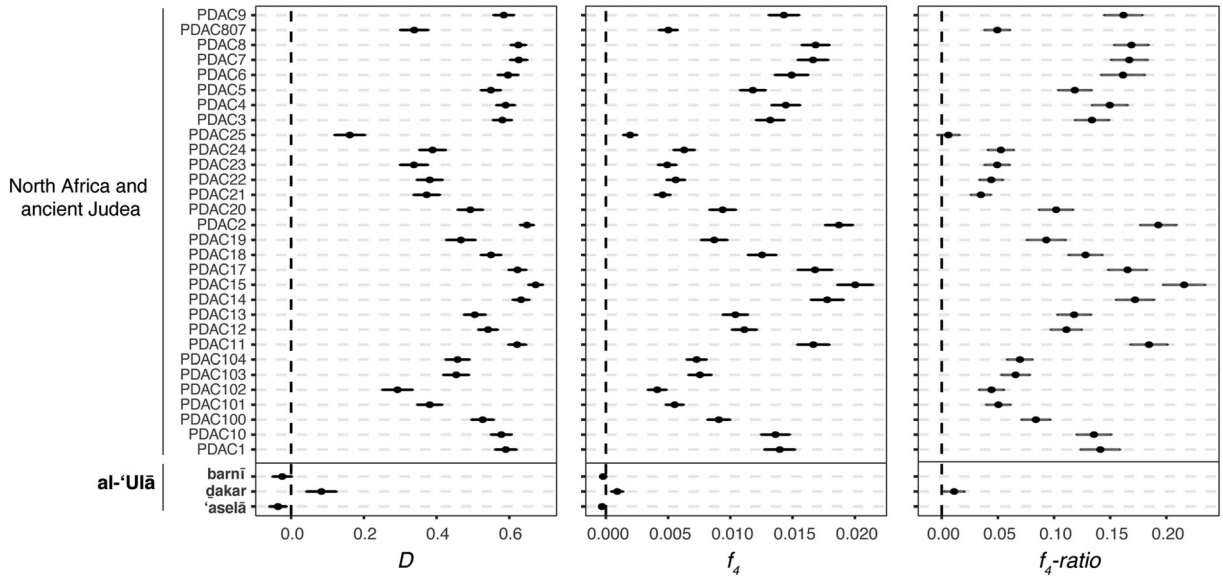


FIGURE 6 Admixture between date palms from North Africa, ancient Judea, al-'Ulā and *P. theophrasti*. *D*-statistic and f_4-statistic, both testing whether modern North African, ancient Judean, and al-'Ulā date palms show an excess of shared alleles with *P. theophrasti*, in which case they are significantly positive. We estimated the *D*-statistics and f_4-statistics using the following tree: (((test sample, West Asian date palms), *P. theophrasti*), *P. reclinata*). The f_4-ratio statistic, indicates the fraction of *P. theophrasti* genomes found in the test sample and calculated as f4(A,O; X,C)/f4(A,O; B,C), where X is the test sample; A is a sister species, namely *P. sylvestris*; B and C are the mixing populations, namely, West Asian date palms and *P. theophrasti*, respectively; and O is the outgroup, that is, *P. reclinata*. It was calculated only for samples showing evidence of an excess of shared alleles with *P. theophrasti* as evidenced by *D*- and f_4-statistics, that is all samples except barnī and 'aselā. A negative *D* and f_4 in this context imply a greater degree of allele sharing between West Asian samples and *theophrasti*, whereas positive values would imply greater sharing between the test sample and *theophrasti*. More details on the methods may be found in Methods S2 while the data are in Datasets S5, S6, and S7.

The evidence of gene flow from *P. theophrasti* found in date palms of the Levant has been hypothesized to be related to the growing control of the Roman Empire in the region 2000 years ago, favoring exchange of goods, including dates, with North Africa (Gros-Balthazard et al., 2021). Indeed, a changeover from absence or low *P. theophrasti* ancestry to ~10% of *P. theophrasti* ancestry coincided with a shift in imperial control of the region in favor of the Romans. The region of al-'Ulā, with the famous site of Hegra, was part of the Nabatean kingdom, and a stop on the trading routes connecting southern and eastern Arabia to Petra and the Mediterranean Sea. The site is also located, since 106 CE, on the border of the newly created Roman province of Arabia on the ruins of this kingdom. To observe evidence of ancestry that is rather characteristic of North Africa and the Levant is therefore not unlikely. Whether this reflects ancient (i.e., from the Nabateo-Roman period) or other gene flow remains to be elucidated.

3.4.4 | Maternal origin of al-'Ulā date palms

In date palm, two deeply diverged chlorotypes (so-called western and eastern) have been reported (Pintaud et al., 2013). The eastern (or oriental) chlorotype is found in most West Asian date palms, while in North Africa, the so-called occidental is prominent. Previous analysis of date palm chlorotypes showed a gradient in frequency of the oriental chlorotype from low in Northwestern Africa to ~50% in Egypt, which suggest strong gene flow from West Asia (Gros-Balthazard et al., 2017; Zehdi-Azouzi et al., 2015). Our analysis of chloroplast sequences from three al-'Ulā date palms revealed that all bear the oriental chlorotypes indicating a maternal origin from West Asia (Figure S7).

4 | CONCLUSION AND PROSPECTS

What lies behind variety names in clonally propagated crops has been insufficiently explored, although it is a key element for assessing agrobiodiversity (e.g., in oca, Bonnave et al., 2014 or in date palm, Gros-Balthazard et al., 2020). Many studies have highlighted variation within variety names; in date palms (Al-Khalifah & Askari, 2003; Al-Ruqaishi et al., 2008; Elhoumaizi et al., 2006; Sabir et al., 2014), and in other clonally propagated fruit crops as well (e.g., in grapevine, Meneghetti et al., 2012 or in olive, Lazović et al., 2018), although most do not reference deliberate cultivation practices as a source of this variation (but see Battesti et al., 2018; Gros-Balthazard et al., 2020).

By engaging with local farmers, we have established that barnī cultivated in al-'Ula and its surroundings, is a true-to-type cultivar, that is, its local identity given by the farmers corresponds to a unique genetic identity. It implies that, locally, barnī date palms have always been

strictly vegetatively propagated, even by Bedouins in remote desert areas and in different palm grove farming systems. But geographic scale matters: barnī from al-'Ulā oasis is distinct from that of Oman, and we therefore confirmed the homonymy foreseen by Popenoe and Bennett (1913).

In al-'Ulā alone, to date, we estimate that about 99 varieties are cultivated, some of which have been given names close to barnī: for example, barniyat al-'aīṣ, barniyat banāt sa'ad, and barniyat bader. The two latter are probably old local rare named types, while barniyat al-'aīṣ (the "barnī from al-'Aīṣ") clearly designates a variety said from the oasis of al-'Aīṣ (KSA), which is referred to in al-'Aīṣ as barnī. This stresses the importance of understanding the categorization and naming systems used by farmers in relation to the way they propagate palms for a proper understanding and assessment of crop biodiversity.

In addition, our joint analyses of three date palms from al-'Ulā and other *Phoenix* specimens, have revealed an intriguing diversity patterns, where, although clustering with West Asian date palms, one of them shows influence from North Africa. Further analyses, comprising all date palm diversity from al-'Ulā area, will provide further insights into the diversity and history of the keystone species of this antique oasis.

ACKNOWLEDGEMENTS

We are grateful to the farmers, who opened their gardens and farms to us, shared their knowledge and know-how, and offered us unfailing hospitality.

We thank Claire Newton for providing a sample; Marc Arnoux, Nizar Drou, Michael Dhar, and the New York University Abu Dhabi Bioinformatics Core for assistance with DNA sequencing and bioinformatic analyses; Luis Rivera Serna from the Center for Desert Agriculture at KAUST for assistance with bioinformatic analyses. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. We are particularly grateful to Shenglong Wang for his unwavering support while using NYU cluster.

This work was funded by the French Agency for AlUla Development (AFALULA) with his Saudi partner the Royal Commission for AlUla (RCU) through a grant awarded to Vincent Battesti and Muriel Gros-Balthazard (project al-'Ulā DPA: Ethnographic, genetic, and morphometric analyses of the date palm agrobiodiversity in al-'Ulā oasis). We wish to thank them for their financial and logistic support, and in particular Elisabeth Dodinet and Stéphane Forman. Legally, the Royal Commission for AlUla is the owner and provider of the material.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

MGB and VB conceived and designed the analysis. VB led the ethnobotanical survey accompanied by MGB. MGB, VB, SF, JMF, MDP, RW and NM contributed to data. MGB, VB, MB, SI, JFT, JMF, MDP, YB contributed to the analyses. MGB and VB wrote the paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supporting information of this article. Sources for all downloaded genomic data are stated in Dataset S1. Sequencing reads for the newly sequenced accessions can be found in the public sequence database GenBank under the BioProject ID PRJNA817028 and BioSample IDs can be found in Dataset S1.

ORCID

Muriel Gros-Balthazard https://orcid.org/0000-0002-2587-3946
Vincent Battesti https://orcid.org/0000-0002-5793-1098
Jonathan M. Flowers https://orcid.org/0000-0002-8752-205X
Sylvie Ferrand https://orcid.org/0000-0003-1088-5941
Matthieu Breil https://orcid.org/0000-0002-7981-7317
Sarah Ivorra https://orcid.org/0000-0003-0314-8054
Jean-Frédéric Terral https://orcid.org/0000-0003-1921-2161
Michael D. Purugganan https://orcid.org/0000-0002-9197-4112
Rod A. Wing https://orcid.org/0000-0001-6633-6226
Nahed Mohammed https://orcid.org/0000-0002-8857-3246
Yann Bourgeois https://orcid.org/0000-0002-1809-387X

REFERENCES

Aleid, S. M., Al-Khayri, J. M., & Al-Bahrany, A. M. (2015). Date palm status and perspective in Saudi Arabia. In J. M. Al-Khayri, S. M. Jain, & D. V. Johnson (Eds.), *Date palm genetic resources and utilization* (pp. 49–95). Springer Netherlands. https://doi.org/10.1007/978-94-017-9707-8_3

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Al-Khalifah, N. S., & Askari, E. (2003). Molecular phylogeny of date palm (*Phoenix dactylifera* L.) cultivars from Saudi Arabia by DNA fingerprinting. *Theoretical and Applied Genetics*, *107*(7), 1266–1270. https://doi.org/10.1007/s00122-003-1369-y

Al-Khayri, J. M., Jain, S. M., & Johnson, D. V. (Eds.) (2015). *Date palm genetic resources and utilization*. Springer Netherlands. https://doi.org/10.1007/978-94-017-9707-8

Al-Qurainy, F., Khan, S., Al-Hemaid, F. M., Ali, M. A., Tarroum, M., & Ashraf, M. (2011). Assessing molecular signature for some potential date (*Phoenix dactylifera* L.) cultivars from Saudi Arabia, based on chloroplast DNA sequences rpoB and psbA-trnH. *International Journal of Molecular Sciences*, *12*(10), 6871–6880. https://doi.org/10.3390/ijms12106871

Al-Ruqaishi, I. A., Davey, M., Alderson, P., & Mayes, S. (2008). Genetic relationships and genotype tracing in date palms (*Phoenix dactylifera* L.) in Oman, based on microsatellite markers. *Plant Genetic Resources*, *6*(1), 70–72. https://doi.org/10.1017/S1479262108923820

Al-Yahyai, R., & Al-Khanjari, S. (2019). Biodiversity of date palm in the Sultanate of Oman. *African Journal of Crop Science*, *7*(10), 1–7.

Al-Yahyai, R., & Khan, M. M. (2015). Date palm status and perspective in Oman. In J. M. Al-Khayri, S. M. Jain, & D. V. Johnson (Eds.), *Date palm genetic resources and utilization* (pp. 207–240). Springer Netherlands. https://doi.org/10.1007/978-94-017-9707-8_6

Bahuchet, S. (2017). *Les jardiniers de la nature*. Odile Jacob.

Battesti, V. (2005). *Jardins au désert, Évolution des pratiques et savoirs oasiens: Jérid tunisien*. Éditions IRD. https://doi.org/10.4000/books.irdeditions.10160

Battesti, V. (2013). L'agrobiodiversité du dattier (*Phoenix dactylifera* L.) dans l'oasis de Siwa (Égypte): Entre ce qui se dit, s'écrit et s'oublie. *Revue d'ethnoécologie*, *4*. https://doi.org/10.4000/ethnoecologie.1538

Battesti, V., Gros-Balthazard, M., Ogéron, C., Ivorra, S., Terral, J.-F., & Newton, C. (2018). Date palm agrobiodiversity (*Phoenix dactylifera* L.) in Siwa Oasis, Egypt: Combining ethnography, morphometry, and genetics. *Human Ecology*, *46*(4), 529–546. https://doi.org/10.1007/s10745-018-0006-y

Bonnave, M., Bleeckx, G., Rojas Beltrán, J., Maughan, P., Flamand, M.-C., Terrazas, F., & Bertin, P. (2014). Farmers' unconscious incorporation of sexually-produced genotypes into the germplasm of a vegetatively-propagated crop (*Oxalis tuberosa* Mol.). *Genetic Resources and Crop Evolution*, *61*(4), 721–740. https://doi.org/10.1007/s10722-013-0068-z

Bouchaud, C. (2013). Exploitation végétale des oasis d'Arabie: Production, commerce et utilisation des plantes. L'exemple de Madâ'in Sâlih (Arabie Saoudite) entre le IV^{e} siècle av. J.-C. et le VII^{e} siècle apr. J.-C. *Revue d'ethnoécologie*, *4*. https://doi.org/10.4000/ethnoecologie.1217

Bouchaud, C., Sachet, I., Dal Prà, P., Delhopital, N., Douaud, R., & Leguilloux, M. (2015). New discoveries in a Nabataean tomb. Burial practices and 'plant jewellery' in ancient Hegra (Madâ'in Sâlih, Saudi Arabia). *Arabian Archaeology and Epigraphy*, *26*(1), 28–42. https://doi.org/10.1111/aae.12047

Caillon, S., & Degeorges, P. (2007). Biodiversity: Negotiating the border between nature and culture. *Biodiversity and Conservation*, *16*(10), 2919–2931. https://doi.org/10.1007/s10531-007-9149-7

Chevenet, F., Brun, C., Bañuls, A.-L., Jacq, B., & Christen, R. (2006). TreeDyn: Towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics*, *7*, 439. https://doi.org/10.1186/1471-2105-7-439

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.-M., & Gascuel, O. (2008). Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, *36*(Web Server), W465–W469. https://doi.org/10.1093/nar/gkn180

Doughty, C. M. (1921). *Travels in Arabia Deserta* (Edition (Vol. 1). P.L. Warner and J. Cape. http://archive.org/details/travelsinarabiad01doug

Elhoumaizi, M. A., Devanand, P. S., Fang, J., & Chao, C.-C. T. (2006). Confirmation of 'Medjool' date as a landrace variety through genetic analysis of 'Medjool' accessions in Morocco. *Journal of the American Society for Horticultural Science*, *131*(3), 403–407. https://doi.org/10.21273/JASHS.131.3.403

Elias, M., Rival, L. M., & Mckey, D. (2000). Perception and management of cassava (*Manihot esculenta* Crantz) diversity among Makushi Amerindians of Guyana (South America). *Journal of Ethnobiology*, *20*, 239–265.

Elmeer, K., Mattat, I., Al Malki, A., Al-Mamari, A.-G., Al-Jabri, A., Buhendi, A., Alkhabaz, S., Abu-Idrees, A., Abdulkaream, A., Udupa, S. M., & Baum, M. (2020). Intra-cultivar variability at microsatellite loci in date palm cultivars across the GCC countries. *QScience Connect*, *2020*(1), 3. https://doi.org/10.5339/connect.2020.3

Flowers, J. M., Hazzouri, K. M., Gros-Balthazard, M., Mo, Z., Koutroumpa, K., Perrakis, A., Ferrand, S., Khierallah, H. S. M., Fuller, D. Q., Aberlenc, F., Fournaraki, C., & Purugganan, M. D. (2019). Cross-species hybridization and the origin of North African date palms. *Proceedings of the National Academy of Sciences*, *116*(5), 1651–1658. https://doi.org/10.1073/pnas.1817453116

Gros-Balthazard, M., Battesti, V., Ivorra, S., Paradis, L., Aberlenc, F., Zango, O., Zehdi-Azouzi, S., Moussouni, S., Naqvi, S. A., Newton, C., & Terral, J. (2020). On the necessity of combining ethnobotany and genetics to assess agrobiodiversity and its evolution in crops: A case study on date palms (*Phoenix dactylifera* L.) in Siwa Oasis, Egypt. *Evolutionary Applications*, *13*(8), 1818–1840. https://doi.org/10.1111/eva.12930

Gros-Balthazard, M., & Flowers, J. M. (2021). A brief history of the origin of domesticated date palms. In J. M. Al-Khayri, S. M. Jain, & D. V. Johnson (Eds.), *The date palm genome* (Vol. 1). Phylogeny, biodiversity and mapping. (pp. 55–74). Springer International Publishing. https://doi.org/10.1007/978-3-030-73746-7_3

Gros-Balthazard, M., Flowers, J. M., Hazzouri, K. M., Ferrand, S., Aberlenc, F., Sallon, S., & Purugganan, M. D. (2021). The genomes of ancient date palms germinated from 2,000 y old seeds. *Proceedings of the National Academy of Sciences*, *118*(19), e2025337118. https://doi.org/10.1073/pnas.2025337118

Gros-Balthazard, M., Galimberti, M., Kousathanas, A., Newton, C., Ivorra, S., Paradis, L., Vigouroux, Y., Carter, R., Tengberg, M., Battesti, V., Santoni, S., Falquet, L., Pintaud, J.-C., Terral, J.-F., & Wegmann, D. (2017). The discovery of wild date palms in Oman reveals a complex domestication history involving centers in the Middle East and Africa. *Current Biology*, *27*(14), 2211–2218.e8. https://doi.org/10.1016/j.cub.2017.06.045

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, *52*(5), 696–704. https://doi.org/10.1080/10635150390235520

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, *8*(5), giz034. https://doi.org/10.1093/gigascience/giz034

Hazzouri, K. M., Flowers, J. M., Visser, H. J., Khierallah, H. S. M., Rosas, U., Pham, G. M., Meyer, R. S., Johansen, C. K., Fresquez, Z. A., Masmoudi, K., Haider, N., El Kadri, N., Idaghdour, Y., Malek, J. A., Thirkhill, D., Markhand, G. S., Krueger, R. R., Zaid, A., & Purugganan, M. D. (2015). Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. *Nature Communications*, *6*(1), 8824. https://doi.org/10.1038/ncomms9824

Hazzouri, K. M., Gros-Balthazard, M., Flowers, J. M., Copetti, D., Lemansour, A., Lebrun, M., Masmoudi, K., Ferrand, S., Dhar, M. I., Fresquez, Z. A., Rosas, U., Zhang, J., Talag, J., Lee, S., Kudrna, D., Powell, R. F., Leitch, I. J., Krueger, R. R., Wing, R. A., … Purugganan, M. D. (2019). Genome-wide association mapping of date palm fruit traits. *Nature Communications*, *10*(1), 4680. https://doi.org/10.1038/s41467-019-12604-9

Jarvis, D. I., Hodgkin, T., Brown, A. H. D., Tuxill, J. D., López Noriega, I., Smale, M., Sthapit, B. R., & Samper, C. (2016). *Crop genetic diversity in the field and on the farm: Principles and applications in research practices*. YALE University Press.

Jaussen, A., & Savignac, R. (1914). *Mission archéologique en Arabie* (mars-mai 1907) 2. El-'Ela, d'Hégra à Teima, Harrah de Tebouk *(Ernest Leroux, Éditeur)*. Publications de la Société des fouilles archéologiques. https://archive.org/details/missionarcheolog21jaus

Johnson, D. V., Jain, S. M., & Al-Khayri, J. M. (2013). Seedling date palms (*Phoenix dactylifera* L.) as genetic resources. *Emirates Journal of Food and Agriculture*, *25*(11), 809. https://doi.org/10.9755/ejfa.v25i11.16497

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *13*, 356. https://doi.org/10.1186/s12859-014-0356-4

Korunes, K. L., & Samuk, K. (2021). PIXY: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/10.1111/1755-0998.13326

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Krueger, R. R. (2011). Date palm germplasm. In S. M. Jain, J. M. Al-Khayri, & D. V. Johnson (Eds.), *Date palm biotechnology* (pp. 313–336). Springer Netherlands. https://doi.org/10.1007/978-94-007-1318-5_16

Lane, E. W. (1863). *An Arabic-English lexicon* (Vol. 1-8). Williams and Norgate.

Lazović, B., Klepo, T., Adakalić, M., Šatović, Z., Arbeiter, A. B., Hladnik, M., Strikić, F., Liber, Z., & Bandelj, D. (2018). Intra-varietal variability and genetic relationships among the homonymic East Adriatic olive (*Olea europaea* L.) varieties. *Scientia Horticulturae*, *236*, 175–185. https://doi.org/10.1016/j.scienta.2018.02.053

Leclerc, C., & Coppens d'Eeckenbrugge, G. (2011). Social organization of crop genetic diversity. The G × E × S interaction model. *Diversity*, *4*(1), 1–32. https://doi.org/10.3390/d4010001

Luu, K., Bazin, E., & Blum, M. G. B. (2017). *pcadapt*: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKey, D., Elias, M., Pujol, B., & Duputié, A. (2010). The evolutionary ecology of clonally propagated domesticated plants. *New Phytologist*, *186*(2), 318–332. https://doi.org/10.1111/j.1469-8137.2010.03210.x

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731. https://doi.org/10.1534/genetics.118.301336

Meneghetti, S., Poljuha, D., Frare, E., Costacurta, A., Morreale, G., Bavaresco, L., & Calò, A. (2012). Inter- and intra-varietal genetic variability in Malvasia cultivars. *Molecular Biotechnology*, *50*(3), 189–199. https://doi.org/10.1007/s12033-011-9423-5

Munier, P. (1973). *Le palmier-dattier*. G.-P. Maisonneuve&Larose.

Nasif, A. A. (1988). *Al-'Ulā: An historical and archaeological survey with special reference to its irrigation system*. Riyadh: King Saud University. (Ph.D. thesis in Victoria University of Manchester, 1981).

Newton, C., Gros-Balthazard, M., Ivorra, S., Paradis, L., Pintaud, J.-C., & Terral, J.-F. (2013). *Phoenix dactylifera* and *P. sylvestris* in northwestern India: A glimpse of their complex relationships. *Palms*, *57*(1), 37–50.

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pérez-Escobar, O. A., Bellot, S., Przelomska, N. A. S., Flowers, J. M., Nesbitt, M., Ryan, P., Gutaker, R. M., Gros-Balthazard, M., Wells, T., Kuhnhäuser, B. G., Schley, R., Bogarín, D., Dodsworth, S., Diaz, R., Lehmann, M., Petoe, P., Eiserhardt, W. L., Preick, M., Hofreiter, M., … Baker, W. J. (2021). Molecular clocks and archeogenomics of a late period Egyptian date palm leaf reveal introgression from wild relatives and add timestamps on the domestication. *Molecular Biology and Evolution*, *38*(10), 4475–4492. https://doi.org/10.1093/molbev/msab188

Petr, M., Vernot, B., & Kelso, J. (2019). admixr—R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics*, *35*(17), 3194–3195. https://doi.org/10.1093/bioinformatics/btz030

Pintaud, J.-C., Ludeña, B., Aberlenc-Bertossi, F., Zehdi, S., Gros-Balthazard, M., Ivorra, S., Terral, J.-F., Newton, C., Tengberg, M., Abdoulkader, S., Daher, A., Nabil, M., Saro Hernández, I., González-Pérez, M. A., Sosa, P., Santoni, S., Moussouni, S., Si-Dehbi, F., & Bouguedoura, N. (2013). Biogeography of the date palm (*Phoenix dactylifera* L., arecaceae): Insights on the origin and on the structure of modern diversity. *Acta Horticulturae*, *994*, 19–38. https://doi.org/10.17660/ActaHortic.2013.994.1

Popenoe, P. B., & Bennett, C. L. (1913). *Date growing in the old world and the new/by Paul B. Popenoe (with a chapter on the food value of the date)*. West India Gardens. https://doi.org/10.5962/bhl.title.32190

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rhouma, A. (1994). *Le palmier dattier en Tunisie, I. Le patrimoine génétique* (Vol. 1). Arabesques, INRA Tunisie, GRIDAO France, PNUD/FAO.

Rhouma, A. (2005). *Le palmier dattier en Tunisie, I. Le patrimoine génétique* (Vol. 2). IPGRI, UNDP, GEF/FEM, INRAT.

Rohmer, J., Lesguer, F., Bouchaud, C., Purdue, L., Alsuhaibani, A., Tourtet, F., Monchot, H., Dabrowski, V., Decaix, A., & Desormeau, X. (in press). New clues to the development of the oasis of Dadan. Results from a test excavation at Tall al-Sālimīyyah (al-'Ulā, Saudi Arabia). *Proceedings of the Seminar for Arabian Studies*.

Sabir, J. S. M., Arasappan, D., Bahieldin, A., Abo-Aba, S., Bafeel, S., Zari, T. A., Edris, S., Shokry, A. M., Gadalla, N. O., Ramadan, A. M., Atef, A., Al-Kordy, M. A., El-Domyati, F. M., & Jansen, R. K. (2014). Whole mitochondrial and plastid genome SNP analysis of nine date palm cultivars reveals plastid heteroplasmy and close phylogenetic relationships among cultivars. *PLoS ONE*, *9*(4), e94158. https://doi.org/10.1371/journal.pone.0094158

Scarcelli, N., Tostain, S., Vigouroux, Y., Agbangla, C., Daïnou, O., & Pham, J.-L. (2006). Farmers' use of wild relative and sexual reproduction in a vegetatively propagated crop. The case of yam in Benin. *Molecular Ecology*, *15*(9), 2421–2431. https://doi.org/10.1111/j.1365-294X.2006.02958.x

Vieira, F. G., Lassalle, F., Korneliussen, T. S., & Fumagalli, M. (2016). Improving the estimation of genetic distances from next-generation sequencing data: Genetic distances from NGS data. *Biological Journal of the Linnean Society*, *117*(1), 139–149. https://doi.org/10.1111/bij.12511

Waples, R. K., Albrechtsen, A., & Moltke, I. (2019). Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Molecular Ecology*, *28*(1), 35–48. https://doi.org/10.1111/mec.14954

Zaid, A., & Arias-Jiménez, E. J. (Eds.). (1999). *Date palm cultivation*. FAO, Plant Production and Protection Paper.

Zaid, A., & Oihabi, A. (Eds.). (2022). Mejhoul Variety, The Jewel of Dates. Origin, distribution and international market. Khalifa International Award for Date Palm and Agricultural Innovation.

Zehdi-Azouzi, S., Cherif, E., Moussouni, S., Gros-Balthazard, M., Abbas Naqvi, S., Ludeña, B., Castillo, K., Chabrillange, N., Bouguedoura, N., Bennaceur, M., Si-Dehbi, F., Abdoulkader, S., Daher, A., Terral, J.-F., Santoni, S., Ballardini, M., Mercuri, A., Ben Salah, M., Kadri, K., … Aberlenc-Bertossi, F. (2015). Genetic structure of the date palm (*Phoenix dactylifera*) in the Old World reveals a strong differentiation between eastern and western populations. *Annals of Botany*, *116*(1), 101–112. https://doi.org/10.1093/aob/mcv068

Zhang, C.-R., Aldosari, S. A., & Nair, M. G. (2015). Determination of the variability of sugars in date fruit varieties. *Journal of Plantation Crops*, *43*(1), 53–61.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gros-Balthazard, M., Battesti, V., Flowers, J. M., Ferrand, S., Breil, M., Ivorra, S., Terral, J.-F., Purugganan, M. D., Wing, R. A., Mohammed, N., & Bourgeois, Y. (2023). What lies behind a fruit crop variety name? A case study of the barnī date palm from al-'Ulā oasis, Saudi Arabia. *Plants, People, Planet*, *5*(1), 82–97. https://doi.org/10.1002/ppp3.10326

Appendix B

Last author publication with a PhD student as first author: reference genome of *Testudo graeca*

PLOS ONE

RESEARCH ARTICLE

Taking advantage of reference-guided assembly in a slowly-evolving lineage: Application to *Testudo graeca*

Andrea Mira-Jover[1], Eva Graciá[1]*, Andrés Giménez[1], Uwe Fritz[2], Roberto Carlos Rodríguez-Caro[3], Yann Bourgeois[4]*

1 Ecology Area, University Institute for Agro-food and Agro-environmental Research and Innovation (CIAGRO), Miguel Hernández University, Elche, Carretera de Beniel, Orihuela (Alicante), Spain, **2** Museum of Zoology, Senckenberg Dresden, Dresden, Germany, **3** Departamento de Ecología, Universidad de Alicante, San Vicent del Raspeig, Spain, **4** DIADE, University of Montpellier, Montpellier, France

* egracia@umh.es (EG); yann.bourgeois@ird.fr (YB)

Citation: Mira-Jover A, Graciá E, Giménez A, Fritz U, Rodríguez-Caro RC, Bourgeois Y (2024) Taking advantage of reference-guided assembly in a slowly-evolving lineage: Application to *Testudo graeca*. PLoS ONE 19(8): e0303408. https://doi.org/10.1371/journal.pone.0303408

Editor: Murtada D. Naser, University of Basrah, IRAQ

Received: April 24, 2024

Accepted: July 22, 2024

Published: August 9, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0303408

Data Availability Statement: ***PA AT ACCEPT: Please check if data has been inputted in the NHI-NCB as links are currently empty on the database at RTC***"Yes - all data are fully available without

Abstract

Background

Obtaining *de novo* chromosome-level genome assemblies greatly enhances conservation and evolutionary biology studies. For many research teams, long-read sequencing technologies (that produce highly contiguous assemblies) remain unaffordable or unpractical. For the groups that display high synteny conservation, these limitations can be overcome by a reference-guided assembly using a close relative genome. Among chelonians, tortoises (Testudinidae) are considered one of the most endangered taxa, which calls for more genomic resources. Here we make the most of high synteny conservation in chelonians to produce the first chromosome-level genome assembly of the genus *Testudo* with one of the most iconic tortoise species in the Mediterranean basin: *Testudo graeca*.

Results

We used high-quality, paired-end Illumina sequences to build a reference-guided assembly with the chromosome-level reference of *Gopherus evgoodei*. We reconstructed a 2.29 Gb haploid genome with a scaffold N50 of 107.598 Mb and 5.37% gaps. We sequenced 25,998 protein-coding genes, and identified 41.2% of the assembly as repeats. Demographic history reconstruction based on the genome revealed two events (population decline and recovery) that were consistent with previously suggested phylogeographic patterns for the species. This outlines the value of such reference-guided assemblies for phylogeographic studies.

Conclusions

Our results highlight the value of using close relatives to produce *de novo* draft assemblies in species where such resources are unavailable. Our annotated genome of *T. graeca* paves the way to delve deeper into the species' evolutionary history and provides

restriction" "All genomic and sequence files are available from the NHI-NCBI BioProject database (accession number PRJNA1086345). All scripts used for this study are freely available at https://github.com/YannBourgeois/Scripts_Genome_assembly_Tgraeca

Funding: This work was supported by Project PID2019-105682RA-I00 and TED2021-130381B-I00, funded by the Spanish Ministry of Innovation, Science and Universities (MICIU/AEI/10.13039/501100011033), the last also with the support of the European Union "NextGenerationEU"/PRTR". Roberto Carlos Rodríguez Caro was supported by European Union-Next Generation EU in the Maria Zambrano Programme (ZAMBRANO 21-26).

Competing interests: The authors have declared that no competing interests exist.

a valuable resource to enhance direct conservation efforts on their threatened populations.

Introduction

Whole genome sequencing (WGS) has become a powerful tool in evolutionary and conservation biology due to the progressive reduction of practical and economical efforts to generate genomic libraries [1]. New high-throughput sequencing methods can be used to produce highly contiguous reference genomes for non-model or "obscure" organisms [1, 2]. Long-read DNA sequencing (e.g., Oxford Nanopore Technologies or PacBio) is a promising technique to generate high-quality reference genomes and is established as the future of *de novo* assemblies [3–6]. However, long read technologies remain expensive for species with large genomes, and require large amounts of high-molecular-weight DNA to be efficient [6]. The involved extraction protocols require fresh or flash-frozen tissues that cannot always be obtained for many laboratories and study systems [3, 5, 6]. Meanwhile short-read techniques remain cheaper and easier to use than long-read methods, and allow for using degraded samples [4–6]. Mapping-based and reference-guided assemblies (alignment of contigs/scaffolds to a close relative reference genome) provide a powerful tool to generate contiguous genomes using short reads [4]. The long scaffolds obtained from the close relative can be used to anchor the typically short contigs obtained with short reads [4, 7]. This method is particularly interesting for highly conserved syntenic genomes or even for extinct species, where mapping back the reference genome against extant relatives is the only option [4, 7–10]. Reference-guided assemblies provide an interesting feedback loop: the more high-quality reference genomes are obtained, the more likely it is that a close relative of the species of interest will be available [7, 9].

Chelonians, the vertebrate group that includes tortoises and turtles, is remarkable for its highly conserved synteny [11, 12]. The very wide diversity of ecological niches of chelonians is not reflected in their functional diversity or genome organization, which remains highly conserved across taxa while nucleotide divergence is particularly reduced and considered a slowly-evolving group (except for mitogenomes) [11–15]. High synteny is often observed in reptiles, including birds, while mammals tend to display more dynamic genomes (for a review see [16]). High synteny may be facilitated by long generation times typically found in chelonians, reducing the odds of rearrangements during meiosis. The lack of recently active transposable elements (TEs) may also contribute to prevent abnormal recombination between distant copies [17].

Genetic and genomic resources for chelonians have increased over the past few years. For example, a phylogeny obtained from 98 mitogenomes has contributed to address how ancestral extinctions, niche diversity and biogeography have impacted extant diversity [18]. However, microevolutionary processes like gene flow, genomic recombination, introgression, or hybridization, cannot be extensively addressed using only mitogenome trees [18]. Nuclear reference genomes are also becoming increasingly available. There are currently 38 reference genomes from 14 different families in the NCBI database (Table 1). These genomic resources have shed light on speciation events, ancient and recent demographic changes, and are also promising for addressing future studies, such as genomic determinants of aging, immunology, aridity tolerance, or gigantism in chelonians [12–14, 19–23]. However, the majority of these reference genomes represent aquatic species, particularly freshwater turtles (Table 1). Land tortoises or simply tortoises (Testudinidae) are the most threatened family of all chelonians [24], but only five annotated testudinid genomes have been published, with three from the

Table 1. State of the art of reference genome availability for chelonians. Species and subspecies are classified and divided into the main ecotypes (freshwater, marine, or terrestrial species). Estimated genome size is represented as Gb for all species. Assembly level represents the highest level for any object in the assembly (i.e., the sequence organization or connection among them). Sequencing technology is defined for every species jointly with N50, which indicates if the genome was sequenced using long-read (expressed on Mb) or short-read methods (expressed on kb).

Type	Species/subspecies	Genome size (Gb)	Assembly level	Sequencing technology	N50	NCBI Accession (GenBank)
Freshwater	*Actinemys marmorata*	2.30	Scaffold	PacBio Sequel II; Illumina NovaSeq; Dovetail OmniC	75.1 Mb	GCA_022086475.1
	Actinemys pallida	2.33	Scaffold	PacBio Sequel II and Sequel IIe; Dovetails OmniC; Illumina NovaSeq	94 Mb	GCA_023634205.1
	Apalone spinifera	1.90	Chromosome	Illumina HiSeq	14.7 kb	GCA_030068395.1
	Carettochelys insculpta	2.18	Chromosome	PacBio Sequel	126.5 Mb	GCA_033958435.1
	Chelydra serpentina	2.26	Scaffold	PacBio Revio	47.4 Mb	GCA_018859375.1
	Chrysemys picta bellii	2.48	Chromosome	454 Life Sciences	21.3 kb	GCA_000241765.5
	Cuora amboinensis	2.21	Scaffold	Illumina	47.2 kb	GCA_004028625.2
	Cuora mccordi	2.39	Scaffold	10X Genomics	74.3 kb	GCA_003846335.1
	Dermatemys mawii	1.87	Scaffold	10X Genomics Chromium	180 kb	GCA_007922305.1
	Emydoidea blandingii	2.3	Scaffold	PacBio Sequel; Illumina NovaSeq	43.1 Mb	GCA_036785055.1
	Emydura macquarii macquarii	1.92	Contig	Oxford Nanopore PromethION; Illumina HiSeq	17.1 Mb	GCA_026122565.1
	Emydura subglobosa	1.99	Scaffold	10X Genomics Chromium	351 kb	GCA_007922225.1
	Emys orbicularis	2.31	Chromosome	PacBio Sequel II HiFi; Bionano DLS; Arima Hi-C v2	91.3 Mb	GCA_028017835.1
	Glyptemys insculpta	2.32	Scaffold	PacBio Sequel; Illumina NovaSeq	95.7 Mb	GCA_032172135.1
	Graptemys geographica	2.3	Contig	PacBio Revio	107 Mb	GCA_037349215.1
	Macrochelys suwanniensis	2.13	Chromosome	PacBio Sequel II HiFi; Arima Hi-C v2	43.9 Mb	GCA_033296515.1
	Malaclemys terrapin pileata	2.21	Chromosome	PacBio Sequel II HiFi; Bionano Genomics DLS; Arima Hi-C v2	75.6 Mb	GCA_027887155.1
	Mauremys mutica	2.48	Chromosome	PacBio	15 Mb	GCA_020497125.1
	Mauremys reevesii	2.37	Chromosome	Oxford Nanopore; Illumina	33.4 Mb	GCA_016161935.1
	Mesoclemmys tuberculata	2.03	Scaffold	10X Genomics Chromium	146.4 kb	GCA_007922155.1
	Pelochelys cantorii	2.16	Chromosome	PacBio Sequel	41.4 Mb	GCA_032595735.1
	Pelodiscus sinensis	2.20	Scaffold	Illumina HiSeq 2000	22 kb	GCA_000230535.1
	Pelusios castaneus	2.04	Scaffold	10X Genomics Chromium	74.9 kb	GCA_007922175.1
	Platysternon megacephalum	2.32	Scaffold	Illumina	213.6 kb	GCA_003942145.1
	Podocnemis expansa	2.45	Scaffold	10X Genomics Chromium	134.8 kb	GCA_007922195.1
	Rafetus swinhoei	2.24	Chromosome	Oxford Nanopore PromethION	31 Mb	GCA_019425775.1
	Sternotherus odoratus	1.76	Scaffold	PacBio Sequel; Illumina NovaSeq	17 Mb	GCA_032164245.1
	Terrapene carolina triunguis	2.57	Scaffold	10X Genomics Chromium	76.6 kb	GCA_002925995.2
	Trachemys scripta elegans	2.13	Chromosome	Illumina NovaSeq; PacBio	140 Mb	GCA_013100865.1

(*Continued*)

Table 1. (Continued)

Type	Species/subspecies	Genome size (Gb)	Assembly level	Sequencing technology	N50	NCBI Accession (GenBank)
Marine	*Caretta caretta*	2.13	Chromosome	Illumina NovaSeq; Oxford Nanopore PromethION	18.2 Mb	GCA_023653815.1
	Chelonia mydas	2.13	Chromosome	PacBio Sequel I CLR; Illumina NovaSeq; Arima Genomics Hi-C; Bionano Genomics DLS	39.4 Mb	GCA_015237465.2
	Dermochelys coriacea	2.16	Chromosome	PacBio Sequel I CLR; llumina NovaSeq; Arima Genomics Hi-C; Bionano Genomics DLS	7 Mb	GCA_009764565.3
	Eretmochelys imbricata	2.30	Chromosome	PacBio Sequel	82 Mb	GCA_030012505.1
Terrestrial	*Aldabrachelys gigantea*	2.37	Chromosome	PacBio Sequel	58.7 Mb	GCA_026122505.1
	Chelonoidis niger abingdonii	2.30	Scaffold	Illumina HiSeq; PacBio	73.2 kb	GCA_003597395.1
	Gopherus agassizii	2.18	Scaffold	Illumina HiSeq	43.7 kb	GCA_002896415.1
	Gopherus evgoodei	2.30	Chromosome	PacBio Sequel I; 10X Genomics linked reads; Arima Genomics Hi-C; Bionano Genomics DLS	13 Mb	GCA_007399415.1
	Gopherus flavomarginatus	2.46	Chromosome	PacBio Sequel I CLR; Bionano Genomics DLS; Arima Genomics Hi-C; 10X Genomics linked reads	6.9 Mb	GCA_025201925.1

https://doi.org/10.1371/journal.pone.0303408.t001

same genus (North American desert tortoises), *Gopherus flavomarginatus*, *G. evgoodei*, and *G. agassizii*. The other two represent giant tortoises from Galapagos (*Chelonoidis niger abingdonii*) and Aldabra (*Aldabrachelys gigantea*) [13, 14, 22] (Table 1). Of them, only three assemblies are annotated at the chromosome level: *G. flavomarginatus*, *G. evgoodei*, and *A. gigantea* (Table 1), all of them using long-read techniques.

In the Testudinidae family, the genus *Testudo* comprises five species of Mediterranean tortoises [24–27], and three of them are listed as threatened by the IUCN: *Testudo graeca* and *T. horsfieldii* are considered vulnerable [VU], and *T. kleinmanni* is listed as Critically Endangered [CE]) [24]. The spur-thighed tortoise (*T. graeca Linnaeus*, 1758) is the most widespread *Testudo* species in the Western Palearctic and shows an intricate phylogeographic history. Eleven mitochondrial lineages are described for *T. graeca*, and are divided into two different groups. The first, the eastern group, spans through the Near and Middle East and southeastern Europe, and consists of *T. g. ibera*, *T. g. terrestris*, *T. g. buxtoni*, *T. g. zarudnyi*, and *T. g. armeniaca* [27]. The second, the western group, primarily inhabits North Africa, but also includes some isolated populations in southwestern Europe. It is represented by *T. g. graeca*, *T. g. whitei*, *T. g. marokkensis*, *T. g. nabeulensis*, *T. g. cyrenaica*, and an additional lineage awaiting its formal description [27]. Fossil-calibrated molecular clock analyses based on mitochondrial data suggest that the western group diverged from its sister taxon, *T. g. armeniaca*, during the Pliocene (7.95–3.48 Mya). Two independent diversification bursts took place during the Mio-Pliocene (8–2 Mya) for the eastern lineages, and during the Pleistocene (1–0.1 Mya) for the subspecies distributed in North Africa [27]. Southwestern European populations have their origin in North Africa [26] being historically introduced on Mallorca, Sardinia, and the Doñana National Park [27, 28]. An exception is a *T. g. whitei* population in southeastern Spain, with molecular markers indicating a range expansion from North Africa during the Late Pleistocene (20 kya) and subsequent natural expansion in southeastern Spain [26].

In the face of the conservation status of the *Testudo* species (and all other turtles and tortoises, with more than 50% of its species considered as Threatened [24]) and the singularity of the phylogeographic history of *T. graeca* throughout the Mediterranean, a reference genome for *Testudo* will greatly contribute to all future studies aimed at the conservation and better understanding of tortoises.

To address the lack of genomic resources for this genus, we present the first high-quality genome for *T. graeca*. We sequenced it using short-read technology on an Illumina platform to generate a draft assembly, and used an available reference genome of a close relative (*G. evgoodei*, diverged approximately 50 Mya) [18, 29] to scaffold and annotate the genome. Our work demonstrates the efficiency of the reference-guided assembly to create accurate *de novo* reference genomes that can serve for future studies.

Material and methods

Sample collection and sequencing

To sequence the whole genome of *T. graeca*, we sampled a fresh road-killed male tortoise from Murcia (southeastern Spain) (S1 Map of sample location). Both field sampling, and the collection and treatment of biological samples, were supported by the government of Murcia Region (AUF20140057) and Project Evaluation Agency of the Research Vice-Rectorate of Miguel Hernandez University (Elx, Spain) (UMH.DBA.EGM.03.19). The sample was stored at -18˚C. Tissues were extracted under sterile conditions and kept frozen until processing. DNA was extracted from muscle using the E.Z.N.A Tissue DNA kit (Omega Biotek) and eluting it in 100 μL. DNA quantification was performed by a Qubit High Sensitivity dsDNA Assay (Thermo Fisher Scientific) at a final concentration of 37.6 ng/μL. Genomic DNA libraries were constructed using the TruSeq Nano DNA kit and quality-checked in the TapeStation D1000 ScreenTape System (Agilent Technologies). Genomic libraries were sequenced on an Illumina NovaSeq with PE150 (paired-end) to obtain a total output of 220 Gb (*c.* 100X depth of coverage). Raw FASTQ files were quality-checked using FastQC v0.11.5 [30] (S1 Appendix). All the procedures were carried out by AllGenetics & Biology S.L. following its company protocols.

Genome assembly

Before starting the assembly, we adapter-trimmed all sequences and quality-filtered them using Trimmomatic 0.39 [31]. We discarded reads with a Phred quality score lower than 28 and trimmed reads when quality dropped below 5. We removed the Illumina adapters (TruSeq3-PE) and discarded reads shorter than 40 bp. Overlapping reads were merged employing Pear v0.9.11 (default overlap of 10 bp) [32]. Sequencing errors were corrected using SOAPec v. 2.0.3 by specifying k-mer size as 27, and the cut-off size as 3, for removing low-frequency k-mers. Assembly was performed using SOAPdenovo2 (version 2.04 release 242) [33] with a range of increasing k-mer values (27, 37, 47, 57, 67, 77, 87, 97, 107). We also tested using k-mer sizes (121 and 127 bp), predicted as being optimal by KmerGenie 1.7051 [34]. KmerGenie was also deployed to predict genome size.

The assemblies that employ short reads are generally fragmented and consist in thousands of short contigs. To improve our draft assemblies, we used ntJoin [4] to scaffold our draft assemblies with *G. evgoodei* as a reference given its high-quality chromosome-level assembly with a few unplaced scaffolds. We ran ntJoin with a range of word sizes (100 bp, 250 bp, 500 bp and 1000 bp) and a set of k-mer values (16, 24, 32, 40, 48, 56, 64). Any gaps between contigs were then closed using GapCloser v1.12 [33]. To confirm the efficiency of this approach, we examined the completeness of our assembly with BUSCO (BUSCO score v 5.3.0) [35]. We tested the continuity and the presence of 5310 shared genes of tetrapods (tetrapoda_0db10, from OrthoDB database) before and after applying ntJoin, and always after gap removal with GapCloser. We also compared the quality of the different assemblies by examining N50, L50 and other statistics using the stats.sh script from the bbmap suite (BBTOOLS 38.18). Quality criteria were assessed according to the percentage of gaps, the number of the longest scaffolds

covering half the assembly (L50), and the shortest length of those scaffolds (N50). Therefore, we retained the assembly with the lowest gap percentage, the highest N50 and the lowest L50.

Repeat analysis

For identifying repeated elements, we used RepeatModeler v2.0.2 [36] to create *de novo* predictions of repetitive sequences and to construct a library of repetitive elements for *T*. *graeca*. To mask the genome, we combined this *de novo* annotation with an existing consensus of repetitive sequences for tetrapods using the freely available resources (Dfam) provided with RepeatMasker v4.1.2 [36]. We ran the latter program with RMBLAST v2.11.0 to classify and annotate all the repeat families. Then we built a Repeat Landscape to compare *T*. *graeca* repeat content to other species. We explored the age distribution of TEs by examining the divergence among the different TE families with the calcDivergenceFromAlign.pl script from the RepeatMasker package.

Gene annotation

Gene finding was performed using BRAKER2 v2.1.6 [37], which incorporates a combination of tools to predict gene coordinates and generates gene structure annotations [37–39]. As we do not have access to the RNAseq data for our species, we applied the BRAKER pipeline using the "C" option to incorporate "proteins of any evolutionary distance" into our target species. Because these methods work better with proteins from related species, we combined the protein annotations available for *G*. *evgoodei*, *A*. *gigantea*, and *Gallus gallus* with a set of vertebrate protein data obtained from OrthoDB (tetrapoda_odb10) using DIAMOND [40] to remove any redundant genes between both sources. We ran gene predictions on our masked genome to avoid wrongly annotating TEs as genes. Briefly, the pipeline involves running ProtHint [41] to generate hints of protein prediction by identifying alignments with sequences from close or distant relatives for *T*. *graeca* in the provided protein database. Annotation is further improved by training AUGUSTUS [38, 40, 41] on the set of hints to obtain the coordinates and predictions of introns, exons, and start/stop codons. We obtained Gene Ontology (GO) terms and gene names for the predicted genes with EggNOG-mapper v2 [42], and by a high-precision search among orthologous groups.

We also transferred the *G*. *evgoodei* annotation to the *T*. *graeca* draft genome using Liftoff with default parameters [43].

Mitogenome reconstruction

We used MitoZ [44] with default parameters on a subset of 10 million pairs of reads to reconstruct the mitogenome of *T*. *graeca*. Several k-mer values were tested (59, 79, 99, 119, 141). The final assembly was obtained with a k-mer value 141. We aligned our mitogenome reference to other *Testudo* mitogenomes using MAFFT online with default parameters (https://mafft.cbrc.jp/). To further confirm the quality of our sequence, we checked its position in the phylogeny of complete *T*. *graeca* mitogenomes with a mitogenome from *Testudo marginata* as an outgroup (NCBI accession DQ080047.1). We also employed MAFFT online to run a Neighbor-Joining phylogenetic analysis on the alignment using 100 bootstrap replicates to calculate node support.

Demographic history inference

Historical changes in the effective population size were inferred with the MSMC2 v2.1.4 software [45]. MSMC2 uses a Hidden Markov Model to estimate the most recent time since coalescence among the haplotypes under recombination. The method can be applied to a single

diploid genome, but requires heterozygous sites to be identified to obtain coalescence times between the two haplotypes. To do so, we realigned the reads on the reference genome using BWA-MEM-2 [46]. We ran freebayes v1.3.2 [47] to call variants from the generated alignment file (in BAM format). We restricted the analysis to the nine longest scaffolds. To identify poorly mappable regions, we employed GenMap v1.3.0 [48] on the genome assembly. We used BEDTOOLS v2.29.2 [49] to obtain the depth of coverage along the genome (average of 100x) from the BAM file. We masked the regions with a mappability lower than 1 and a depth of coverage below 10X. With VCFtools v0.1.16 [50], we filtered the SNP variants from each chromosome and excluded the sites with a genotype quality lower than 30 and depth less than 10X, or more than 200X. Using the *generate_multihetsep.py* script (provided by msmc-tools, a repository containing utilities for MSMC2, https://github.com/stschiff/msmc-tools), we merged VCF outputs and mask files together to generate the input files for MSMC2. The software was run with default parameters by defining time segmentation as *'-p 1*2+25*1+1*2 +1*3'* and grouping the first and last two-time intervals to force the coalescent rate to remain constant.

The coalescence rates estimated by MSMC2 were converted into generations at a mutation rate of 6 x 10^{-10} bp/year based on the *c*. 6% divergence between the *G. evgoodei* and *T. graeca* genomes, which diverged *c*. 50 Mya (substitution rate of 3% per lineage over 50 My, or 6 x 10^{-10} substitutions per year) [29]. Generation time was estimated at 17.72 years, as in Graciá et al. [26].

Results

Genome sequencing and assembly

For whole genome sequencing, we generated a total of 2 x 913,404,107 high-quality paired-end short Illumina reads with an average sequence length of 151 bp and a GC content of 45%. After adapter and low-quality trimming, we conserved 872,311,739 sequenced reads. *KmerGenie* predicted an optimal k-mer value for the genome *de novo* assembly of 125 bp for an estimated genome size of 2,172,882,866 bp. This estimate is consistent with the sizes obtained for other chelonian genomes assembled at the chromosome level (ranging from 2.13 Gb for *Chelonia mydas* and 2.48 Gb for *Chrysemys picta bellii*). Assembling with SOAPdenovo and a k-mer size of 87, produced the assembly with the highest scaffold and contig L50 (5.67 kb and 4.01 kb, respectively; see also S1 Fig). This assembly was used for further scaffolding employing ntJoin. A word size of 100 and a k-mer size of 24 resulted in the reference-guided assembly with the highest contig N50 value (3.6 kb) and the smallest gap proportion (13.24%). The proportion of gaps dropped to 5.37% after running GapCloser (Table 2), but the contig N50 rose to 132,837 bp. This scenario suggests that a large proportion of contigs and scaffolds obtained by SOAPdenovo were correctly positioned in relation to one another to ensure efficient gaps filling in the reference-guided assembly (Table 2). The BUSCO complete score rose from 30.3% for the SOAPdenovo assembly to 96.7% after scaffolding with ntJoin and gaps filling, while the proportion of the fragmented and missing genes dropped from 28.8% to 1.1% and from 40.9% to 2.2%, respectively (Table 2 and Fig 1A).

Repeat content

The fraction of repetitive elements identified by RepeatModeler/RepeatMasker was 41.02% for a total of 956,680,633 bp. This falls in line with other related chelonian genomes (Table 3). Long interspersed nuclear elements (LINEs) were the most abundant class of repetitive elements (11.11%), followed by DNA transposons (7.19%), short interspersed nuclear elements (SINEs) (2.06%) and long terminal repeats retrotransposons (LTR-RTs) (3.11%). Unclassified elements accounted for 17.04% of the draft genome. Divergence of repeats from their

Table 2. Comparison of the assembly statics and BUSCO analysis before and after correcting the draft assembly by the ntJoin method.

	SOAPdenovo2 assembly (k = 87)	Reference-guided assembly after running GapCloser
Assembly statics		
Main genome scaffold (L/N)	3,895,059/2,588.160 MB	84,170/2,332.51 MB
Main genome contig (L/N)	4,236,106/2,563.699 MB	112,106/2,207.2 MB
GAPs	0.945%	5.37%
Main genome scaffold L/N50	118,065/5,676 bp	6/130.16 MB
Main genome contig L/N50	157,154/4,014 bp	4,605/132.84 KB
Main genome scaffold L/N90	817,946/240 bp	31/3.92 MB
Main genome contig L/N90	1,228,070/203 bp	21,136/13.46 KB
Max scaffold length	220,437 bp	348.492 MB
Max contig length	220,066 bp	1.21 MB
Scaffolds > 50 kb	157	144
% Main genome in scaffolds >50 kb	0.39%	92.15%
BUSCO v5.3.0 search		
Complete and single-copy	1561/29.4%	5082/95.7%
Complete and duplicated	49/0.9%	55/1%
Fragmented	1530/28.8%	56/1.1%
Missing	2170/40.9%	117/2.2%
Total groups searched	5310	5310

https://doi.org/10.1371/journal.pone.0303408.t002

consensus sequences showed a mode at 7% (S2 Fig), which suggests limited activity of transposable elements (TEs) over recent (< 10 Mya) evolutionary times.

Gene annotation

The *de novo* annotation of the genome obtained with BRAKER2 using a bank of vertebrate protein sequences recovered 24,397 genes with an average length of 6808 bp (Fig 1B). Although the total number of genes recovered is in line with the number estimated for other Testudinidae, their average length is one order of magnitude shorter than those of *G. evgoodei* or *A. gigantea* [14]. As the *G. evgoodei* annotation benefited from transcriptomic data, we transferred it to our own reference. Of the 19,808 coding genes, 462 could not be transferred to the *T. graeca* genome (Table 4). Most (70%) of the genes that were reconstructed *de novo* overlapped with a gene from the *G. evgoodei* annotation, which confirmed that our *de novo* annotation recovered the majority of coding genes, but likely not their full-length sequence.

Mitogenome reconstruction and phylogenetic placement

We reconstructed the mitogenome using MitoZ, and obtained a 16,928-bp-long circularized sequence with 37 annotated mitochondrial genes (Fig 1C). Depth of coverage was homogeneous along the sequence (average depth +/- s.d.: 172X +/- 19 when excluding the outlier control region), except for the control region where it peaked at more than 2000X. The comparison to the other two complete *T. graeca* mitogenomes (NCBI accession numbers: DQ080049.1, DQ080050.1) confirmed the completeness of our assembly outside the control region, and the whole sequence was fully aligned to the other references with 96.2% and 98.7% identity. Our mitogenome was shorter than the other references (lengths: 17,674 and 19,278 bp) due to a shorter control region. This region could not be fully assembled due to its

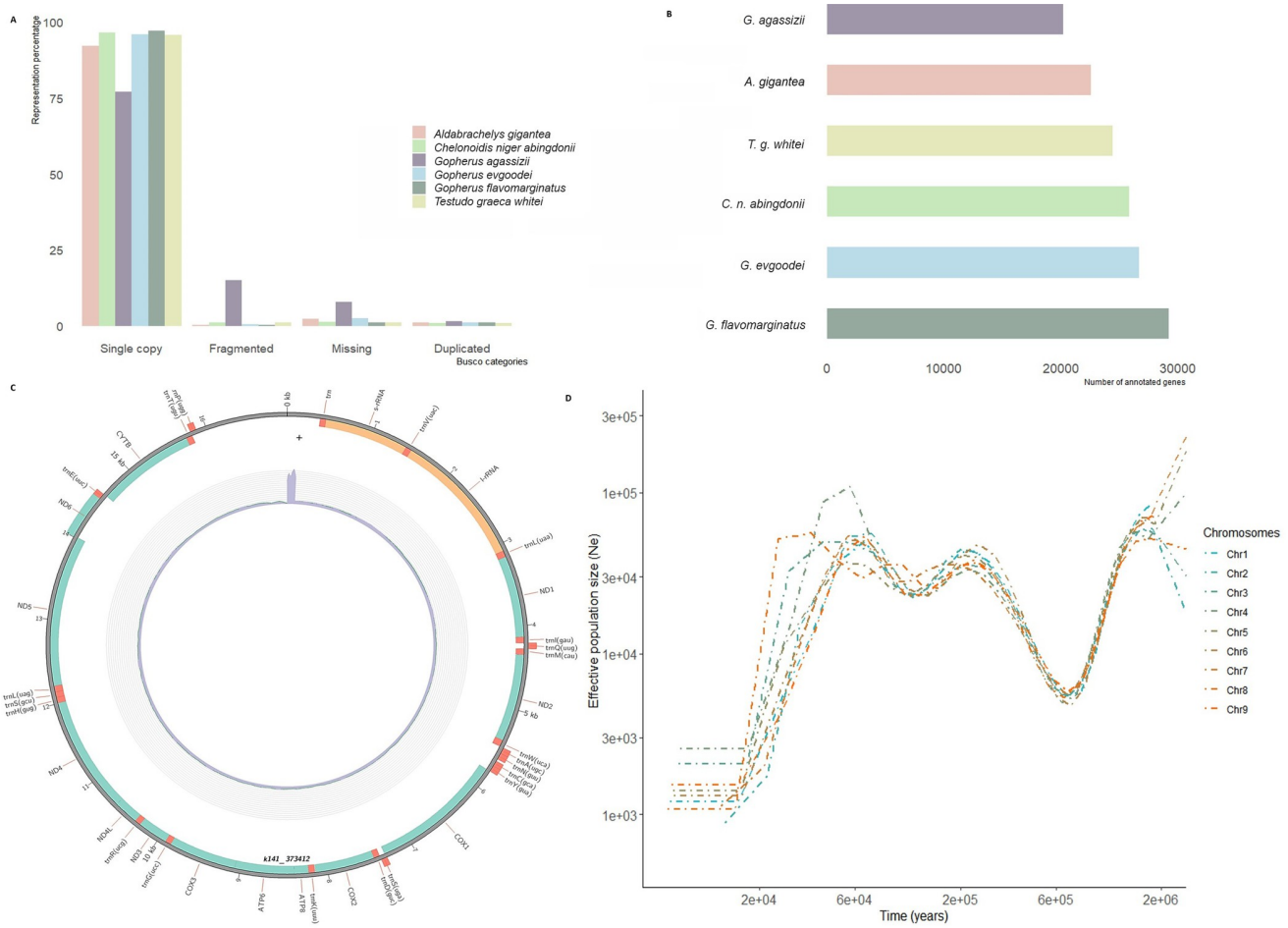


Fig 1. Representation of the main genome assembly results. (A) BUSCO completeness comparison with the five assembled Testudinidae genomes. (B) Number of genes annotated for the five Testudinidae genomes and for *T. graeca*. (C) Circos plot of the mitogenome, with the position of the annotated genes and depth of coverage (inner circle). (D) Demographic reconstruction of the large annotated chromosomes for *T. graeca*.

https://doi.org/10.1371/journal.pone.0303408.g001

repetitiveness and the relatively short length of our reads and inserts. A Neighbor-Joining phylogenetic analysis places our reference close to the Tunisian sample with high support, but more distant from the Turkish sample. All this is consistent with expectations given the species' biogeography (S3 Fig).

Demographic reconstruction

To evaluate the applicability of *T. graeca*'s reference genome for demographic analyses, we reconstructed its past demographic history using MSMC2 (Fig 1D). We estimated an effective population size (N_e) to have revolved around 30,000 individuals over the last 3 My, with two population decline events: the first one around 1 Mya, when N_e declined to *c.* 6000 individuals; the second decline more recently occurred at 40–20 kya, with N_e declining to *c.* 1000 individuals. Of these declines, we observed a significant recovery in N_e, approximately by one order of magnitude, during the period between 600–200 kya.

Table 3. Representation of the different families of repetitive elements found in the *Testudo graeca* assembly and other closely related available chelonians (*Gopherus agassizii* and *Aldabrachelys gigantea*). *Testudo graeca* and *A. gigantea* show similar rates of retroelements as DNA transposons over the genome, while the *G. agassizii* masked region results in a greater presence of retroelements, but fewer DNA transposons. We note that the *G. agassizii* study only ran RepeatModeler on unmasked regions, while the *T. graeca* and *A. gigantea* studies cover the whole genome. Unreported statistics are indicated as 'NA'.

	Number of elements/Length occupied (bp)		
	T. graeca	***G. agassizii***	***A. gigantea***
Interspersed repeats	944,683,527	NA	1,087,548,019
Retroelements	1,296,773 / 379,529,682	NA / 505,075,885	1,177,209 / 482,092,777
SINEs	331,628 / 47,948,990	NA / 44,092,705	51,461,867 / 326,746
Penelope	139,606 / 30,016,241	NA	NA
LINEs	838,188 / 259,068,640	NA / 276,159,275	293,395,900 / 695,701
LTR	126,957 / 72,512,052	NA / 184,823,905	137,235,010 / 154,762
DNA transposons	661,863 / 167,700,840	NA / 297,537,719	198,183,931 / 642,321
Unclassified	2,294,377 / 397,453,005	NA / 203,226,010	407,271,311 / 1,902,917
Others			
Small RNA	48,814 / 9,288,904	NA / 9,451,148	NA
Satellites	1,775 / 691,439	NA / 1,402,596	NA
Simple repeats	216,453 / 8,510,043	NA	NA
Low complexity	33,453	NA	NA

https://doi.org/10.1371/journal.pone.0303408.t003

Discussion

Using Illumina NovaSeq PE150 sequencing, we generated the first high-quality draft genome assembly for *T. graeca*, including its mitogenome. The reference-guided assembly notably increased sequence contiguity and facilitated annotation. This illustrates the efficiency of the reference-guide assembly for chromosome-level scaffolding and gene annotation by providing a resource for comparing genome organization and diversity within and across clades [4, 51]. However, there are always potential inherent biases towards the reference and *de novo* assembled genome (mainly due to divergent regions between the chosen and target assemblies or errors in reference sequence annotation) [7]. In our case, taking *G. evgoodei* as a reference drastically reduced the gap contents and the presence of any "fragmented" and "missing genes". BUSCO completeness scored favorably with other chelonians (Fig 1A), but it should be noted that Çilingir et al. [14] conducted a BUSCO analysis using OrthoDB v10 datasets from phylum (vertebrata_odb10) and class (sauropsida_odb10) instead of all the tetrapods.

BRAKER2 gene prediction estimated a similar number of genes to other Testudinidae (Fig 1A and 1B). However, lack of RNA-seq data prevented us from obtaining full-length transcripts and genes. By making the most of the high contiguity of our reference, and combined with conserved synteny and high identity with *G. evgoodei*, we were able to transfer the annotation of the latter to the *T. graeca*'s genome (Table 4).

As *T. graeca* shows accurate differences between population and lineages, the genome herein produced is valuable for further population genomics studies. Using reference genomes

Table 4. Comparison of the *de novo* annotation with BRAKER and Liftoff of the already existing annotation from the related *Gopherus evgoodei*.

Annotation	Number of exons	Number of coding exons	Number of genes	Number of private genes	Average gene length (+/- s.d.)	Average coding exon length (+/- s.d.)
BRAKER	12,9168	12,9168	24,397	7293	6808 +/- 10,306	225 +/- 352
Liftoff from G. *evgoodei*	32,4881	31,8841	19,346	2242	38,670 +/- 67,033	195 +/- 307

https://doi.org/10.1371/journal.pone.0303408.t004

from distantly related species can negatively impact SNP calling by underestimating the number of variants or biasing heterozygote calling [52]. This effect is significant in turtles and tortoises [12], and obtaining a reference from the same species ensures accurate future genotyping, thereby avoiding bias in the analyses. Highly contiguous genomes are also essential for proper gene annotation. This is clearly reflected by the drastic improvement in our BUSCO scores, with the proportion of complete single copies recovered rising from 29.4% to 95.7%. The average gene length of *c*. 38,000 bp in *Gopherus* is nearly one order of magnitude higher than the scaffold N50 of 5676 bp before reference-guided assembly. At last, contiguous reference genomes are critical for accurate population genetic inference. In humans, approaches such as MSMC2 and related methods lose in accuracy for scaffold lengths under 100 kb– 1 Mb [53]. Before reference-guided scaffolding, the longest scaffold of our assembly was 220,437 kb long, while only four scaffolds were longer than 100 kb. Even assuming human-like mutation and recombination rates, which are likely higher than in tortoises, the *T. graeca* reference would therefore have lacked of long enough scaffolds for MSMC-like approaches, impairing demographic reconstructions. Long scaffolds are also important for genome scans of selection, which rely on the lengths of haplotypes to detect possible selective sweeps [54].

Demographic history reconstructions can address biological questions and retrace the evolutionary dynamics underlying the current distribution and the population genetic status of species [45, 55]. The population size dynamics herein inferred aligns well with the past population changes proposed in previous studies that used mtDNA or microsatellites [26, 27]. The older population decline is compatible with the rapid radiation suggested for the North African lineages during the Pleistocene [27]. Today these subspecies show a clear niche differentiation in North Africa, particularly in relation to climate variables like rainfall [56]. The subsequent population recovery aligns with the diversification of *T. g. whitei*, and has been estimated to have occurred between 850 and 170 kya [27]. For this lineage, it has been suggested that it was confined to several refuge areas during the Last Glacial Maximum, from which it subsequently expanded [56], and a similar pattern can be anticipated during other glacial maxima. Repeated contractions and expansions may have greatly contributed to lineage diversification during the Pleistocene. Finally, the more recent decline is consistent with the bottleneck linked with the species' arrival in southeastern Spain, estimated to have occurred some 20 kya [26, 27].

As inferred for *T. g. whitei* in our demographic reconstruction, the currently available reference genome will increase our knowledge of the past population dynamics and other lineages' demographic history.

Our repetitive element analysis showed that *c*. 40% of the genome is made of TEs, which is a similar proportion to other chelonians, such as *Chelonia mydas*, *Chrysemys picta bellii*, or *Gopherus* spp., but lower than the estimates for *A. gigantea* (46.7%) or *Trachemys scripta elegans* (45%) (Table 2). The Interspersed Repeat Landscape suggests very low recent transposition given the observed age distribution of TEs. Recent TEs activity appears unlikely, and is possibly biased due to the difficulty of assembling highly repeated regions, and using reference-guided ones. However, this is unlikely given the Kmergenie estimates, which are consistent with the reconstructed genome length and consistent estimates with a c-value.

Conclusion

In this study, we report the first reference genome for the genus *Testudo* and add *T. graeca*'s genome to the "toolkit" of genomic resources for tortoises. Given the shared synteny of Testudinidae, we made the best of the high-quality assemblies for another tortoise species, *G.*

evgoodei, to scaffold and annotate a chromosome-level assembly genome from short-read sequences. Thanks to this approach, we avoided the higher costs and sample quality challenges of long-read techniques, and make the most of the low error rate and the cost effectiveness of short read sequencing.

This newly generated reference genome will be useful for answering questions about the evolutionary history and conservation of the *T*. *graeca* complex, and possibly of other *Testudo* species.

Supporting information

S1 Map. Sample location coordinates.
(PDF)

S1 Appendix. FastQC report.
(PDF)

S1 Fig. Summary Statics of ntJoin scaffolding. Upper row: summary statistics (scaffold, right, and contig L50, left) for SOAPdenovo2 assemblies using a range of k-mer size before scaffolding with ntJoin. The red dot indicates the assembly retained for the next step (scaffolding with ntJoin). Lower row: Percentage of gaps remaining after scaffolding the k = 87 SOAPdenovo2 assembly with ntJoin for k-mer size (right) and word size (left). The red dot indicates the assembly retained for the next step (GapCloser).
(PDF)

S2 Fig. Repeated Landscape of *Testudo graeca*. Kimura divergence of each repetitive element copy from its consensus is displayed as a barplot.
(PDF)

S3 Fig. Mitochondrial phylogeny. Phylogeny of full-length *Testudo graeca* mitochondrial sequences using *T*. *marginata* as an outgroup. Bootstrap support is indicated at nodes.
(PDF)

Acknowledgments

We thank AllGenetics SL sequencing services, Portsmouth University for computational resources and to all the Ecology Area of Miguel Hernández University (specially to Paco Botella), Serbal, Ecologistas en Acción de Murcia, and Andalucía and Murcia Region governments for their field support. Finally, we acknowledge all projects who provided new and high-quality reference genomes of chelonians. The genome assembly was done on the Sciama High Performance Compute (HPC) cluster which is supported by the ICG, SEPNet and the University of Portsmouth.

Author Contributions

Conceptualization: Eva Graciá, Yann Bourgeois.

Formal analysis: Andrea Mira-Jover, Yann Bourgeois.

Funding acquisition: Eva Graciá, Andrés Giménez.

Resources: Eva Graciá, Andrés Giménez, Roberto Carlos Rodríguez-Caro, Yann Bourgeois.

Supervision: Eva Graciá, Andrés Giménez, Yann Bourgeois.

Validation: Uwe Fritz, Roberto Carlos Rodríguez-Caro.

Visualization: Andrea Mira-Jover, Eva Graciá, Andrés Giménez, Uwe Fritz, Roberto Carlos Rodríguez-Caro, Yann Bourgeois.

Writing – original draft: Andrea Mira-Jover, Yann Bourgeois.

Writing – review & editing: Andrea Mira-Jover, Eva Graciá, Andrés Giménez, Uwe Fritz, Roberto Carlos Rodríguez-Caro, Yann Bourgeois.

References

1. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol. 2014; 29:51–63. https://doi.org/10.1016/j.tree.2013.09.008 PMID: 24139972
2. Matz MV. Fantastic beasts and how to sequence them: ecological genomics for obscure model organisms. Trends Genet. 2018; 34:121–32. https://doi.org/10.1016/j.tig.2017.11.002 PMID: 29198378
3. Blom MPK. Opportunities and challenges for high-quality biodiversity tissue archives in the age of long-read sequencing. Mol Ecol. 2021; 30(23):5935–48. https://doi.org/10.1111/mec.15909 PMID: 33786900
4. Coombe L, Nikolić V, Chu J, Birol I, Warren RL. NtJoin: Fast and lightweight assembly-guided scaffolding using minimizer graphs. Bioinformatics. 2020; 36(12):3885–7. https://doi.org/10.1093/bioinformatics/btaa253 PMID: 32311025
5. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021; 82(11):801–11. https://doi.org/10.1016/j.humimm.2021.02.012 PMID: 33745759
6. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021; 39:1348–65. https://doi.org/10.1038/s41587-021-01108-x PMID: 34750572
7. Lischer HEL, Shimizu KK. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics. 2017; 18(1): 474. https://doi.org/10.1186/s12859-017-1911-6 PMID: 29126390
8. Palkopoulou E, Lipson M, Mallick S, Nielsen S, Rohland N, Baleka S, et al. A comprehensive genomic history of extinct and living elephants. Proc Natl Acad Sci USA. 2018; 115(11): E2566–E2574. https://doi.org/10.1073/pnas.1720554115 PMID: 29483247
9. Liu K, Xie N, Wang Y, Liu X. The utilization of reference-guided assembly and in silico libraries improves the draft genome of Clarias batrachus and Culter alburnus. Mar Biotechnol. 2023; 25(6):907–17. https://doi.org/10.1007/s10126-023-10248-x PMID: 37661218
10. Barnett R, Westbury MV, Sandoval-Velasco M, Vieira FG, Jeon S, Zazula G, et al. Genomic adaptations and evolutionary history of the extinct scimitar-toothed cat, Homotherium latidens. Curr Biol. 2020; 30(24):5018–5025.e5. https://doi.org/10.1016/j.cub.2020.09.051 PMID: 33065008
11. Lee LS, Navarro-Domínguez BM, Wu Z, Montiel EE, Badenhorst D, Bista B, et al. Karyotypic evolution of sauropsid vertebrates illuminated by optical and physical mapping of the painted turtle and slider turtle genomes. Genes (Basel). 2020; 11(8):1–20. https://doi.org/10.3390/genes11080928 PMID: 32806747
12. Vilaça ST, Piccinno R, Rota-Stabelli O, Gabrielli M, Benazzo A, Matschiner M, et al. Divergence and hybridization in sea turtles: Inferences from genome data show evidence of ancient gene flow between species. Mol Ecol. 2021; 30(23):6178–92. https://doi.org/10.1111/mec.16113 PMID: 34390061
13. Tollis M, DeNardo DF, Cornelius JA, Dolby GA, Edwards T, Henen BT, et al. The Agassiz's desert tortoise genome provides a resource for the conservation of a threatened species. PLoS One. 2017; 12 (5): e0177708. https://doi.org/10.1371/journal.pone.0177708 PMID: 28562605
14. Çilingir FG, A'Bear L, Hansen D, Davis LR, Bunbury N, Ozgul A, et al. Chromosome-level genome assembly for the Aldabra giant tortoise enables insights into the genetic health of a threatened population. Gigascience. 2022; 11(1): giac090. https://doi.org/10.1093/gigascience/giac090 PMID: 36251273
15. Rodríguez-Caro RC, Graciá E, Blomberg SP, Cayuela H, Grace M, Carmona CP, et al. Anthropogenic impacts on threatened species erode functional diversity in chelonians and crocodilians. Nat Commun. 2023; 14(1):1542. https://doi.org/10.1038/s41467-023-37089-5 PMID: 36977697
16. Damas J, Corbo M, Lewin HA. Vertebrate Chromosome Evolution. Annu Rev Anim Biosci. 2021; 9:1–26. https://doi.org/10.1146/annurev-animal-020518-114924 PMID: 33186504
17. Klein SJ, O'Neill RJ. Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Res. 2018; 26:5–23. https://doi.org/10.1007/s10577-017-9569-5 PMID: 29332159
18. Kehlmaier C, Graciá E, Ali JR, Campbell PD, Chapman SD, Deepak V, et al. Ancient DNA elucidates the lost world of western Indian Ocean giant tortoises and reveals a new extinct species from Madagascar. Sci Adv. 2023; 9(2): eabq2574. https://doi.org/10.1126/sciadv.abq2574 PMID: 36630487

19. Vilaça ST, Hahn AT, Naro-Maciel E, Abreu-Grobois FA, Bowen BW, Castilhos JC, et al. Global phylogeography of ridley sea turtles (Lepidochelys spp.): evolution, demography, connectivity, and conservation. Conserv Genet. 2022; 23(6):995–1010.

20. Vilaça ST, Maroso F, Lara P, de Thoisy B, Chevallier D, Arantes LS, et al. Evidence of backcross inviability and mitochondrial DNA paternal leakage in sea turtle hybrids. Mol Ecol. 2023; 32(3):628–43. https://doi.org/10.1111/mec.16773 PMID: 36336814

21. Miller JM, Quinzin MC, Edwards DL, Eaton DAR, Jensen EL, Russello MA, et al. Genome-wide assessment of diversity and divergence among extant Galapagos giant tortoise species. J Hered. 2018; 109 (6):611–9. https://doi.org/10.1093/jhered/esy031 PMID: 29986032

22. Quesada V, Freitas-Rodríguez S, Miller J, Pérez-Silva JG, Jiang ZF, Tapia W, et al. Giant tortoise genomes provide insights into longevity and age-related disease. Nat Ecol Evol. 2019; 3(1):87–95. https://doi.org/10.1038/s41559-018-0733-x PMID: 30510174

23. Simison WB, Parham JF, Papenfuss TJ, Lam AW, Henderson JB. An annotated chromosome-level reference genome of the red-eared slider turtle (Trachemys scripta elegans). Genome Biol Evol. 2020; 12 (4):456–62. https://doi.org/10.1093/gbe/evaa063 PMID: 32227195

24. Turtle Taxonomy Working Group. Turtles of the world. Annotated checklist and atlas of taxonomy, synonymy, distribution, and conservation status (9th Ed.). Chelon Res Monogr. 2021;8:1–472.

25. Escoriza D, Hassine JB. Niche diversification of Mediterranean and southwestern Asian tortoises. PeerJ. 2022; 10: e13702. https://doi.org/10.7717/peerj.13702 PMID: 35846890

26. Graciá E, Giménez A, Anadón JD, Harris DJ, Fritz U, Botella F. The uncertainty of Late Pleistocene range expansions in the western Mediterranean: A case study of the colonization of south-eastern Spain by the spur-thighed tortoise, Testudo graeca. J Biogeogr. 2013; 40(2):323–34.

27. Graciá E, Vargas-Ramírez M, Delfino M, Anadón JD, Giménez A, Fahd S, et al. Expansion after expansion: Dissecting the phylogeography of the widely distributed spur-thighed tortoise, Testudo graeca (Testudines: Testudinidae). Biol J Linn Soc. 2017; 121(3):641–54.

28. Graciá E, Rodríguez-Caro RC, Andreu AC, Fritz U, Giménez A, Botella F. Human-mediated secondary contact of two tortoise lineages results in sex-biased introgression. Sci Rep. 2017; 7(1):1–12.

29. Lourenço JM, Glémin S, Chiari Y, Galtier N. The determinants of the molecular substitution process in turtles. J Evol Biol. 2013; 26(1):38–50. https://doi.org/10.1111/jeb.12031 PMID: 23176666

30. Wingett SW, Andrews S. Fastq screen: A tool for multi-genome mapping and quality control. F1000Res. 2018; 7:1338. https://doi.org/10.12688/f1000research.15931.2 PMID: 30254741

31. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

32. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014; 30(5):614–20. https://doi.org/10.1093/bioinformatics/btt593 PMID: 24142950

33. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012; 1(1):18. https://doi.org/10.1186/2047-217X-1-18 PMID: 23587118

34. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014; 30(1):31–7. https://doi.org/10.1093/bioinformatics/btt310 PMID: 23732276

35. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021; 38(10):4647–54. https://doi.org/10.1093/molbev/msab199 PMID: 34320186

36. Flynn J, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020; 117(17):9451–7. https://doi.org/10.1073/pnas.1921046117 PMID: 32300014

37. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021; 3(1):1–11. https://doi.org/10.1093/nargab/lqaa108 PMID: 33575650

38. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016; 32(5):767–9. https://doi.org/10.1093/bioinformatics/btv661 PMID: 26559507

39. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. In: Methods in Molecular Biology. Humana Press Inc.; 2019. p. 65–95.

40. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015; 12(1):59–60. https://doi.org/10.1038/nmeth.3176 PMID: 25402007

41. Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform. 2020; 2(2): lqaa026. https://doi.org/10.1093/nargab/lqaa026 PMID: 32440658

42. Cantalapiedra CP, Herņandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38(12):5825–9.

43. Shumate A, Salzberg SL. Liftoff: Accurate mapping of gene annotations. Bioinformatics. 2021; 37 (12):1639–43. https://doi.org/10.1093/bioinformatics/btaa1016 PMID: 33320174

44. Meng G, Li Y, Yang C, Liu S. MitoZ: A toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Res. 2019; 47(11): e63. https://doi.org/10.1093/nar/gkz173 PMID: 30864657

45. Schiffels S, Wang K. MSMC and MSMC2: The multiple sequentially Markovian coalescent. In: Methods in Molecular Biology. Humana Press Inc.; 2020. p. 147–66. https://doi.org/10.1007/978-1-0716-0199-0_7 PMID: 31975167

46. Md V, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: Proceedings– 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019. Institute of Electrical and Electronics Engineers Inc.; 2019. p. 314–24.

47. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012 Jul 17; Available from: http://arxiv.org/abs/1207.3907

48. Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. GenMap: Ultra-fast computation of genome mappability. Bioinformatics. 2020; 36(12):3687–92. https://doi.org/10.1093/bioinformatics/btaa222 PMID: 32246826

49. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

50. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

51. Chetruengchai W, Singchat W, Srichomthong C, Assawapitaksakul A, Srikulnath K, Ahmad SF, et al. Genome of Varanus salvator macromaculatus (Asian Water Monitor) reveals adaptations in the blood coagulation and innate immune system. Front Ecol Evol. 2022; 10:850817.

52. Duchen P, Salamin N. A Cautionary note on the use of genotype callers in phylogenomics. Syst Biol. 2021; 70(4):844–54. https://doi.org/10.1093/sysbio/syaa081 PMID: 33084875

53. Gower G, Tuke J, Rohrlach AB, Soubrier J, Llamas B, Bean N, et al. Population size history from short genomic scaffolds: how short is too short? bioRxiv. 2019; https://doi.org/10.1101/382036

54. Bourgeois YXC, Warren BH. An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. Mol Ecol. 2021; 30(23):6036–71. https://doi.org/10.1111/mec.15989 PMID: 34009688

55. Mather N, Traves SM, Ho SYW. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. Ecol Evol. 2020; 10:579–89. https://doi.org/10.1002/ece3.5888 PMID: 31988743

56. Anadón JD, Graciá E, Botella F, Giménez A, Fahd S, Fritz U. Individualistic response to past climate changes: Niche differentiation promotes diverging Quaternary range dynamics in the subspecies of Testudo graeca. Ecography. 2015; 38(9):956–66