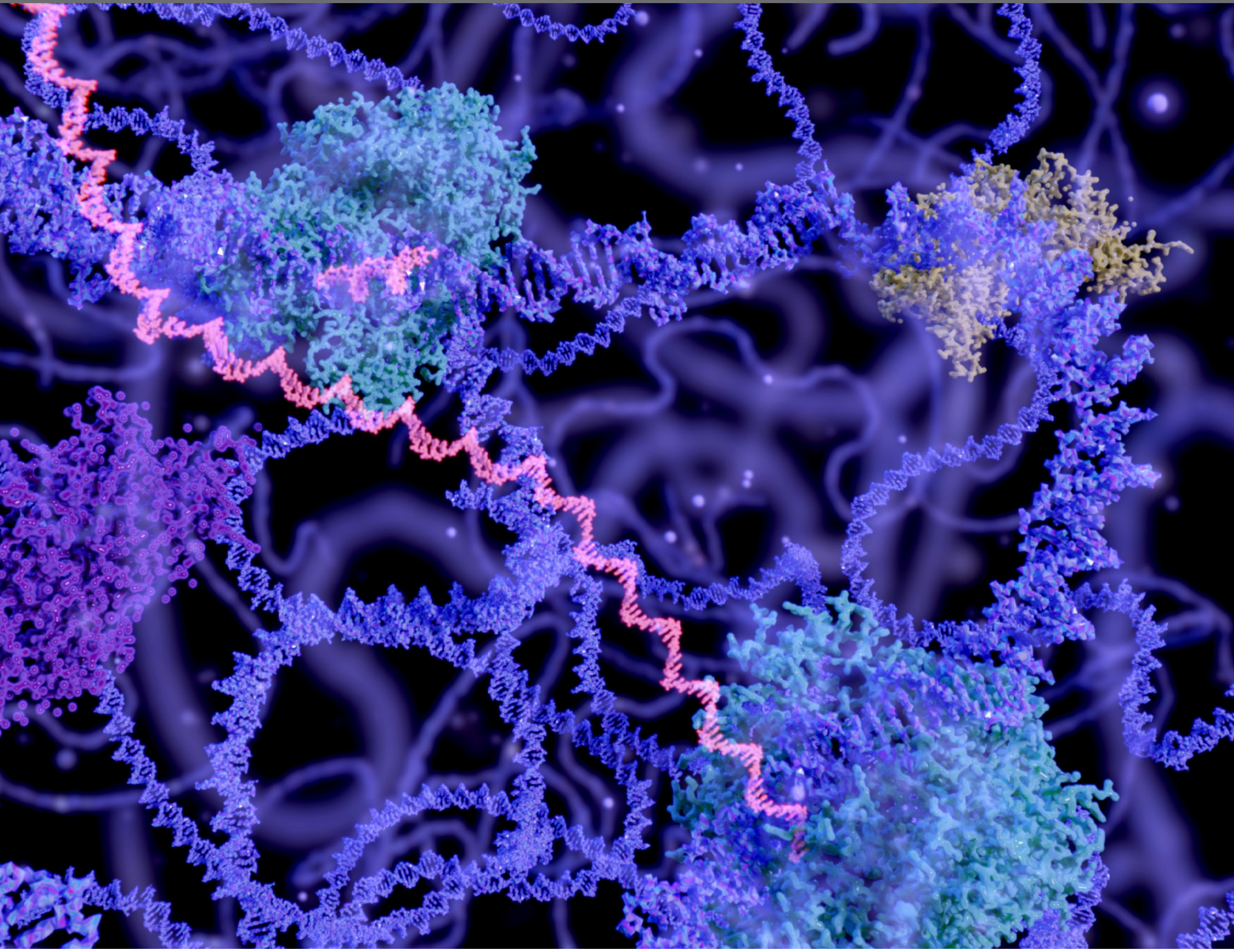
EVOLUTIONARY FEEDBACKS BETWEEN POPULATION BIOLOGY AND GENOME ARCHITECTURE

EDITED BY: Tariq Ezaz and Scott V. Edwards

PUBLISHED IN: Frontiers in Genetics and Frontiers in Plant Science



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-641-3
DOI 10.3389/978-2-88945-641-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

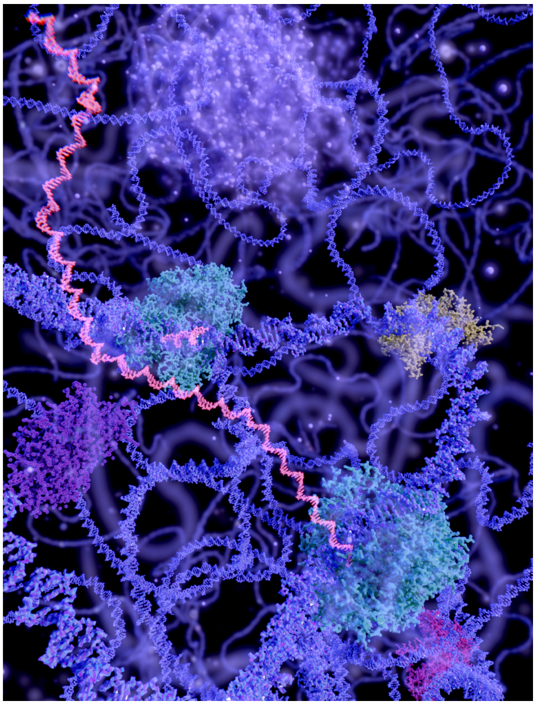
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

EVOLUTIONARY FEEDBACKS BETWEEN POPULATION BIOLOGY AND GENOME ARCHITECTURE

Topic Editors:
Tariq Ezaz, University of Canberra, Australia
Scott V. Edwards, Harvard University, United States



Cover image: Juan Gaertner/Shutterstock.com

This eBook presents all 10 articles published under the Frontiers Research Topic "Evolutionary Feedbacks Between Population Biology and Genome Architecture", edited by Scott V. Edwards and Tariq Ezaz. With the rise of rapid genome sequencing across the Tree of Life, challenges arise in understanding the major evolutionary forces influencing the structure of microbial and eukaryotic genomes, in particular the prevalence of natural selection versus genetic drift in shaping those genomes. Additional complexities in understanding genome architecture arise with the increasing incidence of interspecific hybridization as a force for shaping genotypes and phenotypes. A key paradigm shift facilitating a more nuanced interpretation of genomes came with the rise of the nearly neutral theory in the 1970s, followed by a greater appreciation for the contribution of nonadaptive forces such as genetic drift to genome structure in the 1990s and 2000s. The articles published in this eBook grapple with these issues and provide an update as to the ways in which modern population genetics and genome informatics deepen our understanding of the subtle interplay between these myriad forces. From intraspecific to macroevolutionary studies, population biology and population genetics are now major tools for understanding the broad landscape of how genomes evolve across the Tree of Life. This volume is a celebration across diverse taxa of the contributions of population genetics thinking to genome studies. We hope it spurs additional research and clarity in the ongoing search for rules governing the evolution of genomes.

Citation: Ezaz, T., Edwards, S. V., eds (2018). Evolutionary Feedbacks Between Population Biology and Genome Architecture. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-641-3

Table of Contents

I. GENOMES AND MACROEVOLUTION

II. CHROMOSOME STRUCTURE AND SPECIATION

III. INTROGRESSION AND EVOLUTION OF ORGANELLE GENOMES

IV. METHODS

frontiers in Genetics

EDITORIAL
published: 21 August 2018
doi: 10.3389/fgene.2018.00329

Editorial: Evolutionary Feedbacks Between Population Biology and Genome Architecture

Tariq Ezaz[1] *and Scott V. Edwards*[2]*

[1] *Institute for Applied Ecology, University of Canberra, Canberra, ACT, Australia,* [2] *Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States*

Keywords: microRNA, nearly neutral theory, vertebrate, CNVs, transposable element (TE), natural selection

Editorial on the Research Topic

Evolutionary Feedbacks Between Population Biology and Genome Architecture

Joe Felsenstein famously quipped in 1988 that "systematists and evolutionary geneticists don't often talk to each other," laying bare the schism between macroevolutionary thinking, such as building phylogenies, and population genetics (Felsenstein, 1988). Thirty years ago, in its early phase, studies of genome evolution were already beginning to incorporate population genetics when comparing the DNA of distantly related species (Dover and Flavell, 1982), but the field was far from mature. Today this schism between population genetics and genome evolution is much healed, being bridged by novel statistical methods for detecting natural selection (Kreitman and Akashi, 1995), as well as to the monumental book by Michael Lynch (2007), "The Origins of Genome Architecture," which in turn built upon the foundational nearly neutral theory of Tomoko Ohta and Motoo Kimura (Ohta, 1973, 1992). This Frontiers Research Topic celebrates this increasing infusion of population biology perspectives into studies of genome evolution, which has facilitated key advances in our understanding of how eukaryotic and microbial genomes evolve and the evolutionary forces influencing their structure.

This research topic includes 10 papers that encompass a wide range of model and non-model systems from bacteria, plants, crustaceans, and vertebrates. In addition to linking population biology and genome architecture, these chapters also study diverse components of the genome, including organelle and nuclear genomes, karyotypes, sex chromosomes, RNA transcripts, small non-coding RNAs, and repetitive elements, to understand genome evolution, speciation, and population divergence. For example, Nagai et al. combined mitochondrial and candidate sex determining gene Sox3 sequences to demonstrate speciation events as well as gene flow among 16 populations of frogs from Japan. In a similar approach that delved into interactions between the mitochondrial and nuclear genomes in two avian and one crustacean species, Sunnucks et al. demonstrated how genomic architecture might facilitate better understanding of co-adoption of mitonuclear interactions and enhance biochemical efficiencies of oxidative phosphorylation. Potter et al. present another approach by combining genome sequencing data with cytogenetics to understand chromosome rearrangements leading to speciation and population divergence. Using a unique Australian native marsupial, the rock wallaby, Potter and colleagues demonstrate the value of combined approaches of cytogenetics and genomics to understand evolution of genome

OPEN ACCESS

Edited and Reviewed by:
Samuel A. Cushman,
United States Forest Service (USDA),
United States

****Correspondence:***
Scott V. Edwards
sedwards@fas.harvard.edu

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *16 June 2018*
Accepted: *31 July 2018*
Published: *21 August 2018*

Citation:
Ezaz T and Edwards SV (2018)
Editorial: Evolutionary Feedbacks
Between Population Biology and
Genome Architecture.
Front. Genet. 9:329.
doi: 10.3389/fgene.2018.00329

architecture as driver of speciation. Addressing the variability of LINE element composition in vertebrates, Ruggiero et al. resequenced 13 genomes of green anole (*Anolis carolinensis*) from two populations and found high variation in the frequencies of polymorphic LINE elements, concluding that large effective population size and negative selection together curb the proliferation of LINEs in this species. In another inquiry into repeated sequences, Samelak-Czajka et al. expand on a clever and accurate approach for quantifying CNVs in *Arabidopsis* genomes, offering a promising approach to quantifying these widespread structures in plants. Each of these studies illustrates how molecular and population processes can interact in unexpected ways to shape the structure of eukaryotic genomes, an interdisciplinarity that requires consideration of the population context in which molecular variation is being studied.

Other studies in the Research Topic attempt to bridge population biology and genome evolution at macroevolutionary scales. Hua and Bromham tackle the perennial question of whether rates of lineage diversification are linked to rates of genome evolution and discuss a wide range of results and theoretical connections between the two. França et al. bring a genomic microscope to examining lineage diversification by reviewing the role of microRNAs in altering gene expression between species, thereby playing a potential role in phenotypic evolution and disease. An Miao An et al. turn attention to lineage fusion when studying phenotypic and molecular introgression in oaks in China. They find both morphological and molecular signals of introgression in this complex, in some cases leading to genetic swamping and likely mediated by habitat degradation. Bobay and Ochman update our knowledge of broad trends in bacterial genome architecture, showing how genome size is a product of the interaction of nearly neutral forces of drift and mutation bias toward deletions and how the interplay of selection and drift cause deviations from a strict correlation between genome size and gene number. Finally, Romiguier and Roux synthesize information on the interactions among of GC-biased gene conversion, recombination and natural selection in modulating GC-content in eukaryotes. They illustrate the interplay between micro- and macroevolution by showing how variation in GC-content among lineages can strongly bias our estimation of phylogenies, natural selection, and the extent of codon bias.

The articles in this Research Topic offer a useful snapshot of how research in genome evolution naturally incorporates insights and theories from population genetics, but also some of the challenges of doing so. To what extent should macroevolutionary models remain phenomenological, or instead incorporate the minutiae of evolutionary forces to predict observed patterns today (Hua and Bromham)? How can modern genomics effectively update the classical theories of chromosomal speciation (Potter et al.), or abundance of transposable elements (Ruggiero et al.), and apply them to natural populations? Today, coalescent methods of phylogenetic inference (Liu et al., 2015; Xu and Yang, 2016) have largely answered Felsenstein's quip. Similarly, the nearly neutral theory is widely applied on a genome-wide scale today (e.g., Yi, 2006; Akashi et al., 2012; Gossmann et al., 2012; Denisov et al., 2014), but even more noteworthy is the wide range of species and settings in which it is applied (e.g., Figuet et al., 2016; Chen et al., 2018). As illustrated by the articles in this Research Topic, population biology thinking has been shown to be useful in most areas of genome evolution where researchers care to apply it, from the evolution of sex chromosomes, microRNAs or transposable elements, to GC-content and genome size. This Research Topic shows the breadth of situations, both genomic and ecological, in which population-thinking is helping us to interpret genome evolution. And, in doing so, few can say today that genome evolutionists and population biologists rarely talk to each other.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We thank the Editorial team at Frontiers and the many authors and reviewers who contributed to the success of this Research Topic. SE acknowledges support from NSF grant DEB-1355343.

REFERENCES

Akashi, H., Osada, N., and Ohta, T. (2012). Weak selection and protein. *Evolution* 192, 15–31. doi: 10.1534/genetics.112.140178

Chen, J., Ni, P., Li, X., Han, J., Jakovlić, I., Zhang, C., et al. (2018). Population size may shape the accumulation of functional mutations following domestication. *BMC Evol. Biol.* 18:6. doi: 10.1186/s12862-018-1120-6

Denisov, S. V., Bazykin, G. A., Sutormin, R., Favorov, A. V., Mironov, A. A., Gelfand, M. S., et al. (2014). Weak negative and positive selection and the drift load at splice sites. *Genome Biol. Evol.* 6:1437–1447. doi: 10.1093/gbe/evu100

Dover, G. A., and Flavell, R. B. (1982). *Genome Evolution.* New York, NY: Academic Press.

Felsenstein, J. (1988). Phylogenies and quantitative characters. *Ann. Rev. Ecol. Syst.* 19, 445–471. doi: 10.1146/annurev.es.19.110188.002305

Figuet, E., Nabholz, B., Bonneau, M., Carrio, E. M., Nadachowska-Brzyska, K., Ellegren, H., et al. (2016). Life history traits, protein evolution, and the nearly neutral theory in amniotes. *Mol. Biol. Evol.* 33, 1517–1527. doi: 10.1093/molbev/msw033

Gossmann, T. I., Keightley, P. D., and Eyre-Walker, A. (2012). The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol. Evol.* 4, 658–667. doi: 10.1093/gbe/evs027

Kreitman, M., and Akashi, H. (1995). Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* 26, 403–422. doi: 10.1146/annurev.es.26.110195.002155

Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. (2015). Estimating phylogenetic trees from genome-scale data. *Ann. N. Y. Acad. Sci.* 1360:36–53. doi: 10.1111/nyas.12747

Lynch, M. (2007). *The Origins of Genome Architecture.* Sunderland, MA, Sinauer Associates, Inc.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* 246, 96–98. doi: 10.1038/246096a0

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23, 263–286. doi: 10.1146/annurev.es.23.110192.001403

Xu, B., and Yang, Z. (2016). Challenges in species tree estimation under the multispecies coalescent model. *Genetics* 204, 1353–1368. doi: 10.1534/genetics.116.190173

Yi, S. V. (2006). Non-adaptive evolution of genome complexity. *Bioessays* 28, 979–982. doi: 10.1002/bies.20478

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

HYPOTHESIS AND THEORY
published: 07 February 2017
doi: 10.3389/fgene.2017.00012

Darwinism for the Genomic Age: Connecting Mutation to Diversification

Xia Hua and Lindell Bromham*

Centre for Macroevolution and Macroecology, Research School of Biology, Australian National University, Canberra, ACT, Australia

A growing body of evidence suggests that rates of diversification of biological lineages are correlated with differences in genome-wide mutation rate. Given that most research into differential patterns of diversification rate have focused on species traits or ecological parameters, a connection to the biochemical processes of genome change is an unexpected observation. While the empirical evidence for a significant association between mutation rate and diversification rate is mounting, there has been less effort in explaining the factors that mediate this connection between genetic change and species richness. Here we draw together empirical studies and theoretical concepts that may help to build links in the explanatory chain that connects mutation to diversification. First we consider the way that mutation rates vary between species. We then explore how differences in mutation rates have flow-through effects to the rate at which populations acquire substitutions, which in turn influences the speed at which populations become reproductively isolated from each other due to the acquisition of genomic incompatibilities. Since diversification rate is commonly measured from phylogenetic analyses, we propose a conceptual approach for relating events of reproductive isolation to bifurcations on molecular phylogenies. As we examine each of these relationships, we consider theoretical models that might shine a light on the observed association between rate of molecular evolution and diversification rate, and critically evaluate the empirical evidence for these links, focusing on phylogenetic comparative studies. Finally, we ask whether we are getting closer to a real understanding of the way that the processes of molecular evolution connect to the observable patterns of diversification.

Keywords: molecular evolution, macroevolution, phylogeny, comparative studies, reproductive isolation

OPEN ACCESS

Edited by:
Scott V. Edwards,
Harvard University, USA

Reviewed by:
Brian C. Verrelli,
Virginia Commonwealth University, USA
Amy Litt,
University of California, Riverside, USA

****Correspondence:***
Xia Hua
xia.hua@anu.edu.au

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *09 August 2016*
Accepted: *19 January 2017*
Published: *07 February 2017*

Citation:
Hua X and Bromham L (2017) Darwinism for the Genomic Age: Connecting Mutation to Diversification. Front. Genet. 8:12. doi: 10.3389/fgene.2017.00012

INTRODUCTION

Darwinism unites genetics (heritable characteristics) with population biology (change in frequency of heritable traits through differential reproduction) and biodiversity (formation of new lineages through speciation). The neo-Darwinian synthesis codifies the belief that these processes are all parts of a single evolutionary process, such that the generation of genetic variation that makes individuals differ from each other feeds the change in the characteristics of populations by selection and drift which drives speciation and the diversification of lineages. Thus the neo-Darwinian synthesis "puts an equals sign between microevolution and macroevolution" (Dobzhanksy, 1937). Most biologists accept this unified view of evolutionary process.

Yet evolution is still predominantly studied in a remarkably partite fashion, with most researchers concentrating on one particular aspect of this process. While discipline specialization is a necessary part of modern science, there are some evolutionary phenomena that require an appreciation of many different levels of biological organization.

One such phenomenon is the observed correlation between differences in the mutation rate, estimated by comparing DNA sequences of different species, with net diversification rate, estimated from the number of extant species in different lineages (Barraclough and Savolainen, 2001; Pagel et al., 2006; Eo and DeWoody, 2010; Lanfear et al., 2010; Duchene and Bromham, 2013; Bromham et al., 2015). This relationship might initially seem surprising: how could the rate at which mistakes or damage are repaired in individual genomes lead to differences in species richness between lineages? The correlates of diversity are usually examined at the level of lineages, regions and biological communities. For example, diversification rate has been found to be associated with life history, dispersal rate, generalized feeding strategies, geographic range size, environmental energy and latitude (Cardillo, 1999; Cardillo et al., 2003; Davies et al., 2004; Phillimore et al., 2006). But evolutionary change must ultimately start with heritable changes to genetic information, so the supply of variation is a critical first step in the formation of new evolutionary lineages.

Our aim with this Hypothesis and Theory paper is to examine the possible links in the chain of causation that connects the average mutation rate, estimated by comparing nucleotide sequences, to differences in lineage net diversification rate, as measured from molecular phylogenies. In order to investigate how mutation rate could be linked to diversification rate, we need to consider how an increase in the supply of genetic variation could speed the rate of formation of new species. To do this we will first consider why mutation rate varies between species. Then we will examine the factors that shape the number and type of mutations that become fixed features of the genome within a population (substitutions). Then we consider how the accumulation of these substitutions can cause population to diverge from each other, until the genetic differences between their genomes prevent populations from exchanging genetic material. At this point, the populations can no longer interbreed, and we consider them separate species. Lastly, we propose a conceptual approach to relating the formation of new species to bifurcations on molecular phylogenies.

We explore both the theoretical foundations and empirical evidence for the links in the causal chain between mutation rates and macroevolution. These connections bring together topics typically considered in different biological disciplines, namely molecular evolution, population genetics and macroevolution, but we illustrate some of the approaches that allow us to effectively view all of these processes simultaneously by comparing DNA sequences between individuals, populations, species and lineages. Because the relationship has been noted by comparing gene sequences between lineages, we confine our discussion to mutations that change nucleotide sequences. However, there are many other kinds of genetic change that may also be important drivers of diversification, such as chromosomal rearrangement, regulatory changes or epigenetic modification.

MUTATION RATES EVOLVE

To explain why variation in mutation rate is associated with differences in diversification rate, the first topic we have to address is why species differ in their average mutation rate. Mutation rate is often in the background of molecular evolutionary studies. It is frequently assigned an arbitrary constant value, providing a steady drip of genetic variation into populations. But mutation rate is a dynamic and highly variable phenomenon, changing across the genome and over time, varying among individuals and lineages. We should not be surprised that mutation rate is highly responsive to evolutionary pressures: it is, after all, the fundamental process at the heart of all evolutionary change.

Because we wish to examine the links between diversification rate and the rate of molecular evolution as measured from DNA sequence comparisons, we will consider only point mutations that change single nucleotides in gene sequences. Point mutations arise when damage to DNA is imperfectly repaired or errors in DNA replication are not fully corrected, resulting in permanent change to the base sequence such that the changed sequence will be included in any subsequent copies. DNA is a large, complex molecule: it is inevitable that it will occasionally suffer damage that affects the genetic information it encodes. In addition, the genome must be replicated every time a cell divides, so even a phenomenally low replication error rate (typically only one error per billion bases copied) will result in frequent changes to the DNA sequence. To protect the integrity of the genetic information needed to make essential components of the organism, there is a diverse and sophisticated set of cellular machinery that is directed at detecting and correcting any damage to the DNA or mistakes made in copying the genome. The costs of DNA repair to the cell are not well known, but it seems fair to say that investment in repair apparatus must come at cost of other cellular functions. Therefore, while we expect organisms to invest resources in reducing the risk of harmful mutation, there might come a point where the payoffs of further reduction in mutation rate are outweighed by the increasing costs of DNA fidelity. All species must walk a tightrope, finding a balance between fidelity and error, but they may find different points of balance between repair and mutation (Denamur and Matic, 2006; Bromham, 2009; Lynch, 2010).

One aspect of this balancing act, initially proposed on the basis of information theory, is the need to maintain mutation rates below an “error threshold,” the copy error rate above which a replicating sequence effectively goes extinct by failing to produce sufficiently faithful copies of itself (Eigen, 1971). In the earliest life forms, this may have led to a trade-off between genome size and mutation rate: for a given per-base error rate, a small genome is more likely to be copied without error, but a small genome may also have limited capacity for evolving mechanisms that reduce the mutation rate (Maynard Smith and Szathmáry, 1995). The error threshold model appears to place real-world constraints on the evolution of RNA viruses (Holmes, 2003). The reliance of RNA viruses on the error-prone reverse transcriptase to complete their life cycle endows them with a mutation rate more than an order of magnitude greater than DNA viruses (Sanjuán et al., 2010). While the high mutation rate in RNA viruses has been considered to contribute to their rapid evolvability, it may also place restrictions on adaptation by limiting genome size. Small genomes may speed replication, but it could also place constraints on coding capacity. Therefore it has been suggested that the high mutation rate of RNA viruses may, by limiting genome size, constrain the capacity for evolving new genetic features (Holmes, 2003). Increase in mutation rate beyond the error threshold has been reported as causing extinction of yeast lines (Herr et al., 2011a).

It has been suggested that many species settle on a similar point of balance between the per-genome copy error rate and genome size that represents a chance of error roughly once in every three hundred genome copies (Drake et al., 1998), at least relative to the "non-frivolous fraction" of the genome in which mutations can influence fitness (Drake, 1991). A larger genome with more functional sequences provides more targets for deleterious mutations, so a lower per-base mutation rate is required to have the same per-genome-replication risk of fitness-damaging mutations. The balancing point for the per-genome mutation rates has been most frequently examined for single-celled organisms. It is less clear whether the same relationship between genome size and mutation rate applies to large, complex multicellular organisms. There is surprisingly sparse data on the association between mutation rate and genome size in large multicellular organisms (Bromham et al., 2015), but observed patterns have been considered consistent with an overarching relationship between the effective genome size and the per-generation mutation rate (Lynch, 2010).

The balancing point between the costs and benefits of mutation and repair can be altered by circumstance. Both theory and experiment have suggested that populations of microbes subject to rapidly changing conditions can show a transient increase in mutation rate because "mutator" alleles that increase the rate of mutation can hitchhike to fixation through their association with advantageous mutations (Chao and Cox, 1983; Haraguchi and Sasaki, 1996; Sniegowski et al., 1997; Taddei, 1997; Giraud, 2001). These theoretical models and laboratory experiments have found additional support in observations from medical settings. For example, bacteria infecting the lungs of patients with cystic fibrosis can only persist if they can adapt to ongoing antibiotic treatment and a changing environment as the patient's lung condition deteriorates, resulting in a higher frequency of bacteria with raised mutation rates (Oliver et al., 2000; Oliver and Mena, 2010).

Both models and experiments suggest that it is possible for mutation rates to be shaped by adaptation to circumstance. However, most of these experiments and simulations produce only transient increases in mutation rate, as the association between the mutator allele and beneficial mutations may be uncoupled by recombination, and the burden of deleterious mutations will result in selection against the mutator allele (McDonald et al., 2012). Can selection produce long-term differences in mutation rate between species through balancing the costs and benefits of mutation and repair? There are several observations that suggest it can. Firstly, in many populations there is naturally occurring heritable variation in mutation rates. For example, individuals can vary in the proofreading efficiency of their polymerase enzymes or their mismatch repair systems (Oliver et al., 2002; Sundin and Weigand, 2007; Reha-Krantz, 2010). Mutations that disable essential DNA repair genes can lead to disease phenotypes with dramatically increased mortality for affected individuals (Bradford et al., 2011), but slight variations in DNA repair genes that cause relatively mild reductions in DNA repair efficiency and fidelity can also influence fitness (e.g., Mohrenweiser et al., 2003; Gokkusu et al., 2013; Kabzinski et al., 2016). Although most studied "antimutator" alleles are changes that can compensate for deficiencies in mutator strains (Herr et al., 2011b), there is some evidence of naturally occurring alleles that can increase DNA repair efficiency or copy fidelity (Schaaper, 1998; Foury and Szczepanowska, 2011). The existence of spatial variation in DNA repair enzymes provides further evidence of the evolvability of mutation rate in natural populations (Miner et al., 2015; Svetec et al., 2016). So the raw material for selection to act on mutation rates does not seem to be lacking in natural populations.

Secondly, it seems fair to suppose that DNA repair is costly, such that organisms must find a level of investment that maximizes survival yet minimizes costs. Higher DNA repair efficiency may come at cost of other key cellular functions. For example, costs of repair may explain why the induction of resource-intensive stress response pathways in bacteria can lead to an increased mutation rate (Torres-Barceló et al., 2013). The potential trade-off between investment in DNA repair and other cellular functions is supported by the observation that some DNA repair systems are inducible under mutagenic conditions. For example, plants can have repair systems that are turned on or increased under high UV conditions, to prevent an upsurge in DNA damage (Ries et al., 2000). This suggests that maximum DNA repair efficiency is not maintained at all times, instead the repair effort is scaled to a level that allows the maintenance of cellular processes and reproductive potential. The metabolic cost of DNA repair is supported by observations that individuals in poor condition or subjected to mild stress can have elevated mutation rates, presumably because they are unable to invest as much in DNA repair (Agrawal and Wang, 2008; Goho and Bell, 2000). For example, it has been reported that male birds with lower levels of antioxidant carotenoids have higher rates of DNA damage and also reduced survivorship and lower mating success (Freeman-Gallant et al., 2011).

Thirdly, species can adapt to different levels of risk of mutation through altered investment in DNA repair. Species living in highly mutagenic environments might require greater investment in DNA repair in order to be able to maintain a persistent population against mutational meltdown. Microbes living in high-altitude lakes, exposed to high salinity, high levels of UV radiation and high concentrations of heavy metals, have been shown to have enhanced levels of DNA repair (Albarracin et al., 2012). Bacteria adapted to survive long periods of desiccation, which results in accumulation of DNA double-strand breaks, can have enhanced DNA repair that incidentally allows them to survive other mutagens such as ionizing radiation (Mattimore and Battista, 1996). *Escherichia coli* selected to survive ionizing radiation can evolve DNA repair proteins that are more efficient and less prone to inhibition by perturbations of normal metabolism, potentially making these repair proteins able to function under a broader range of environmental conditions (Piechura et al., 2015).

Conversely, species living in low mutagen environments may need to invest less in DNA repair. For example, species living in low UV environments may have lost some of their photolyase repair genes (Lucas-Lledó and Lynch, 2009). Populations can have consistently different levels of efficiency of UV-induced DNA repair, apparently reflecting different points of adaptation of DNA repair efficiency to match local environmental conditions. For example, water fleas from natural ponds with higher levels of UV exposure have more efficient DNA repair of light-induced mutation (Miner et al., 2015) and fruit fly populations from different latitudes show different levels of UV repair efficiency (Svetec et al., 2016). The adjustment of DNA repair to the level of mutation risk might also explain the puzzling lack of evidence for the prediction that basal metabolic rate should have a direct influence on the mutation rate, due to the production of free oxygen radicals that can damage DNA (Gillooly et al., 2007). Although this relationship has been modeled based on body size and temperature, comparative studies have found no significant variation of rate of molecular evolution with metabolic rate, above and beyond the association with other life history traits such as body size or generation time (Bromham et al., 1996; Lanfear et al., 2007; Galtier et al., 2009b). One possible explanation for the lack of a significant association between metabolic rate and mutation rate is that DNA repair efficiency may be adjusted to ameliorate any additional damage, so that species with high metabolic rates also evolve greater levels of protection against damage from free-oxygen radicals (Lanfear et al., 2007).

DNA repair proteins thus can be considered as a “highly adaptable scaffold readily tailored by evolution to the requirements for genome maintenance in each particular organism” (Piechura et al., 2015). Evolution shapes mutation rates just as it shapes other species traits, balancing costs and benefits to suit the species form and lifestyle. In the next section, we will consider how differences in mutation rates between species can vary predictably with species traits.

SPECIES TRAITS INFLUENCE MUTATION RATE EVOLUTION

Given that DNA repair is likely to impose a non-trivial cost on individuals, that higher rates of mutation can lead to lower survival and reproduction, and that individuals show genetic variation in DNA repair efficiency, we should expect selection to find a balance between the cost of repair and the cost of mutation. But the balance between these competing costs might vary with other species characteristics (Sniegowski et al., 2000).

In multicellular organisms, the costs of mutation are expected to increase with increasing body size (Bromham, 2011). The larger the body, the more cell generations it takes to build it. Every cell division requires the genome to be copied in its entirety, and every genome replication brings the risk of copy errors that might ruin important DNA sequences. The influence of number of genome replications on the mutation rate is supported by the observation of higher per-generation mutation rates in males than in females. Due to the large number of cell generations required for sperm production, the male germline is copied many times more per generation that the female germline, and so the majority of *de novo* mutations arise in males (Wilson Sayres and Makova, 2011). Furthermore, germline mutations accumulate with male age, suggesting that the increasing number of germline divisions per gamete results in more mutations (Kong et al., 2012; Venn et al., 2014). Species with stronger potential for sperm competition, and therefore selective pressure to increase sperm production, have been found in some cases to have higher rates of molecular evolution (Bartosch-Harlid et al., 2003), though this pattern is not supported in all studies (Wilson Sayres et al., 2011). The influence of number of DNA replications on mutation rate has been proposed as the explanation for the generation time effect on rates of molecular evolution, on the assumption that species with faster generation turnovers copy their germline DNA more often per unit time (Gaut et al., 1992; Bromham et al., 1996; Welch et al., 2008; Thomas et al., 2010; Lehtonen and Lanfear, 2014).

However, the copy error effect on its own does not seem sufficient to explain the observed life history patterns in mutation rate (Nabholz et al., 2008; Welch et al., 2008; Gaut et al., 2011; Hua et al., 2015). The difference in mutation rates between species is not commensurate with the difference in number of genome copies per unit time. For example, mice can go through 50 generations for every human generation, yet the rate of molecular evolution in mice is only a few times faster than that in humans. This suggests that the influence of copy number on mutation rate is modulated by other factors. Other life history characteristics have a significant relationship with mutation rate above and beyond their covariation with generation time, particularly fecundity and longevity (Nabholz et al., 2008; Welch et al., 2008). The significant association between generation time and molecular evolution rates may be partly due to it covarying with other causal traits, but being measured more reliably. Most studies of longevity are based on maximum recorded lifespan, which is strongly influenced by number of observations made (that is, it can only go up as more data points are included), so it is possible that for species with relatively little data on longevity, generation time is a better proxy for life history differences.

What effect will selection for longer life spans have on the evolution of a species’ mutation rate? Larger-bodied species tend to have longer lives, requiring the maintenance of more genome copies from incidental damage over a longer time period. Unrepaired damage to any cell’s genome can result in life-shortening damage, such as somatic mutations causing cancer. In particular, the accumulation of mutations in mitochondrial genomes has been proposed as a driver of aging (Fridovich, 2004; Kujoth et al., 2005; Loeb et al., 2005; Larsson, 2010; Yang et al., 2013). Reactive oxygen species produced in the mitochondria as a byproduct of metabolism can cause damage to DNA and other biomolecules. As damage accumulates over time, mitochondrial function may be impeded. So evolving a longer life span may require increased investment in cellular mechanisms that reduce the overall mutation rate in order to keep lifetime mutation risk to a tolerable level. For example, the mutation rate in long-lived species might be reduced by greater investment in mechanisms that prevent oxidative damage (Pamplona and Barja, 2011). This prediction is borne out by observations that long-lived species of mammals, birds and fish tend to have lower per-base mutation rates (as measured by the synonymous substitution rate), above and beyond the association between longevity and other aspects of life history (Nabholz et al., 2008; Welch et al., 2008; Galtier et al., 2009a,b; Hua et al., 2015).

Yet, molecular phylogenetic studies measure the germline mutation rate, not the somatic mutation rate. Although there must be evolutionary pressure on the mutation rate in both somatic and germline cells, the somatic mutation rate need not be the same as the mutation rate in the germline. Indeed, one of the possible evolutionary advantages of multicellularity is the ability to set aside a quiescent germ line in which DNA (particularly organelle DNA) is relatively protected from the harmful byproducts of metabolism, and for which expression levels are kept at a minimum (Bendich, 2010). Keeping germline copies in a quiescent state might be particularly valuable for mitochondrial genomes, hence the evolution of separate sexes: the mitochondrial genomes of motile and metabolically active male gametes can be discarded at fertilization, leaving only the quiescent, relatively unimpaired mitochondria from the immobile female gamete (Allen, 1995; de Paula et al., 2013).

While there is good reason to believe that selection acts on mutation rates, the power of natural selection to shape DNA repair might in itself by limited by species traits, specifically by traits that influence the mutation rate (affecting the supply of variation) and population size (affecting the power of selection). In small populations, fewer mutations that improve DNA repair will arise (Baer et al., 2007), and selection will be less effective at promoting slight improvements to DNA repair or removing mutations that increase the mutation rate, potentially limiting the effectiveness of selection in optimizing mutation rate (Knight et al., 2005; Lynch, 2007; Hodgkinson and Eyre-Walker, 2011).

Whatever the mechanistic or evolutionary forces that shape the differences in mutation rate between species – whether an incidental effect of copy frequency, an adaptive compromise between competing costs, or some other phenomena – the practical upshot is that even closely related species can differ in their average rate of mutation. So the rate of supply of new genetic variation to a species' gene pool is at least partly dependent on a variety of species traits that influence the mutation rate. What effect will differences in the supply of variation have on the evolutionary process? The role of mutation rate in governing the rate of evolution has been given relatively little attention for non-microbial taxa, perhaps due to a general conviction that adaptation in complex multicellular creatures is typically not mutation-limited. But we should not be quick to dismiss the rate of mutation as playing an important role in population diversification. Although there is debate about the relationship between mutation rate, standing variation, and rate of adaptation (Barrett and Schluter, 2008), empirical studies suggest that adaptation to rapid environmental change or strong selection pressures can come from both standing variation (existing alleles) and *de novo* mutations (Hartley et al., 2006; Durand et al., 2010; Jerome et al., 2011). Even if adaptation is not mutation limited, population divergence can also be driven by non-adaptive genome evolution that may be influenced by the mutation rate.

Species-specific differences in mutation rate will only influence the evolutionary rate of lineage divergence if the supply of variation influences the rate of fixation of genetic differences: that is, if the mutation rate at least partly determines the substitution rate. So now that we have explored the evolutionary factors that shape the supply of variation to populations at any point in time, we need to consider how the supply of variation through mutation contributes to rates of genome evolution.

Mutation Rate Influences Substitution Rate

Mutations occur in individual DNA molecules. For a mutation to be detected as a consistent difference in the DNA sequences sampled from different populations or species, it must clear several hurdles. First it must be copied from the genome it occurred in and passed on to that individual's offspring, in order to enter a new generation. If the mutation is included in one or more offspring in the next generation, then each of those individuals has a chance to reproduce and pass the mutation to their offspring. While the mutation is at low frequency in the population, there is a high chance of being lost by chance if those few individuals fail to reproduce. If it increases in representation in the population, the mutation becomes established as a polymorphism, carried by some but not all members of the population. Eventually it may rise in frequency until it replaces all other variants in the population, so now all new individuals born in that population will carry a copy of that mutation (barring a new mutation occurring at the same site). At this point, we call the mutation a substitution and say that it has been fixed in the population. In this section, we will consider the factors that influence the rate at which new mutations become substitutions, thereby contribute to population divergence.

For each neutral mutation that has no effect on fitness, such as most synonymous substitutions that change the DNA sequence of a gene but not the amino acid sequence it codes for, the probability of being passed to future generations is entirely due to chance, and so its frequency in the population is determined only by the mutation rate and the effective population size (N_e). However, the observed level of standing genetic variation within populations is often much lower than would be expected (Lande, 1976; Turelli, 1984; Lynch and Hill, 1986) if the amount of variation was shaped only by the processes of mutation and genetic drift (Frankham, 2012; Hodgins-Davis et al., 2015). This suggests that that there must be additional factors limiting the accumulation of genetic variation in populations (Eyre-Walker and Keightley, 2007).

For non-neutral mutations, such as most non-synonymous mutations that change the amino acid sequence of a protein, their frequency in the population will be determined not only by the mutation rate and the effective population size, but also the selection coefficient of the mutation (s). The influence of chance events on allele frequencies is potentially much greater in small populations. For example, the lucky survival and reproduction of an individual with a mutation that slightly lowers fitness will have a proportionally greater effect on allele frequencies in a small population than in a large population. In a large population, random fluctuations in allele frequencies are less likely to result in chance fixation, because they have proportionally less effect on average frequencies from one generation to the next. The larger the population, the less influence of chance events on allele frequencies. Therefore a beneficial mutation ($s > 0$) is more likely to increase in frequency in larger population than in a smaller population. Mutations with a substantial fitness costs ($s << 0$) will not become substitutions, as they will be removed from the population by purifying selection. But slightly deleterious changes ($s < 0$) will be nearly neutral and are more likely to go to fixation by chance in smaller population (Ohta, 1992).

We can describe the probability that a new mutation will becomes a substitution, found in all genomes in the population. We call this probability the fixation probability P_{fix}. Theoretical studies provide an analytical approximation of the fixation probability (Waxman, 2011):

$$P_{fix} = \frac{1 - E[e^{-4N_e sX(T)}]}{1 - e^{-4N_e s}}.$$

This approximation accounts for a general situation where the values of effective population size and selection coefficient can change over T amount of time. So $T = 0$ means the population has constant effective population size and selection coefficient. According to the equation, the fixation probability depends on the product of the effective population size and the selection coefficient of the mutation, and on the frequency of the mutation at time T [$X(T)$].

Over evolutionary time periods, the overall substitution rate in the population approximates the product of the mutation rate in the population and the fixation probability of each mutation, that is, $2N_e uP_{fix}$, where u is the mutation rate. When most mutations are neutral ($s = 0$) and when effective population size stays constant, P_{fix} depends only on the initial frequency of a mutation. For a new mutation occurring in a population of diploid individuals, its initial frequency is $1/2N$, where N is the number of individuals in the population. Under these conditions, substitution rate equals uN_e/N. To investigate the impact of changing effective population size over time on the substitution rate, we simulated changes in the frequency of a mutation when the effective population size increases, decreases, or stays constant. The resulting frequencies were used to calculate the expectation value of $e^{-4N_e sX(T)}$ in the equation for P_{fix}. Results suggest that changing the effective population size does not have a significant effect on the neutral substitution rate (**Figure 1**). Therefore, the rate of neutral substitutions always approximates uN_e/N, regardless of the demographic history of the population.

Theory predicts the rate of fixation of neutral mutations will be unaffected by population size. The substitution of beneficial mutations will be faster in large population and the effect is greater in an expanding population (**Figure 1**). The substitution of deleterious mutations will be faster in small populations and the effect is greater in a declining population (**Figure 1**). Given that advantageous mutations are relatively rare compared to deleterious mutations (Eyre-Walker and Keightley, 2007), we expect smaller populations to have the greater overall substitution rate for a given mutation rate.

We can test these predictions by comparing different classes of substitutions in DNA sequence comparisons. Since all types of substitution rates are related to mutation rate, taking the ratio between any two of them will cancel out the influence of mutation rate. Therefore, a common test for the effect of population size on substitution rate is to correlate population size to the ratio between non-synonymous to synonymous substitution rates (*dN/dS*). Similarly, radical substitutions (from one class of amino acid residue to another) are more likely to be deleterious than conservative substitutions (within the same functional category) because radical substitutions cause more changes in the physiochemical properties of the protien, so the ratio between them (*Kr/Kc*) is expected to show the same pattern with population size as *dN/dS* (**Figure 1**; Zhang, 2000; Smith, 2003).

The predicted negative relationship between *dN/dS* and population size has been supported by observed correlations between *dN/dS* and life history traits that scale with population size (Nikolaev et al., 2007; Popadin et al., 2007; Lartillot and Delsuc, 2012; Romiguier et al., 2013; Figuet et al., 2014). It has also been supported by comparing sister lineages that differ in effective population size, such as domesticated populations to wild populations (Björnerfeldt et al., 2006; Wang et al., 2011) and island lineages to their mainland close relatives (Johnson and Seger, 2001; Woolfit and Bromham, 2005), although island lineages do not necessarily have smaller effective population size than mainland lineages (Wright et al., 2009; James et al., 2016). A puzzling exception is that *dN/dS* in birds shows no correlation to body size (which is often related to population size), but the correlation between *Kr/Kc* and body size is as predicted (Nabholz et al., 2013; Weber et al., 2014; Figuet et al., 2016). While there are many possible causes of the lack of relationship in birds, such as a lack of correlation between life history traits and historical population size or unreliable estimates of *dN/dS* (Lartillot, 2013), there is thusfar a puzzling lack of evidence that they provide an explanation for the lack of correlation between dN/dS and body size in birds (Figuet et al., 2016).

Interestingly, the formula of fixation probability never suggests a linear relationship between *dN/dS* and population size. In fact, we show in **Figure 1** that in the range of effective population size where it has strong negative relationship with *Kr/Kc*, *dN/dS* may show weak or even no dependence on population size particularly when population declines. So, a simple correlation test between *dN/dS* and population size may

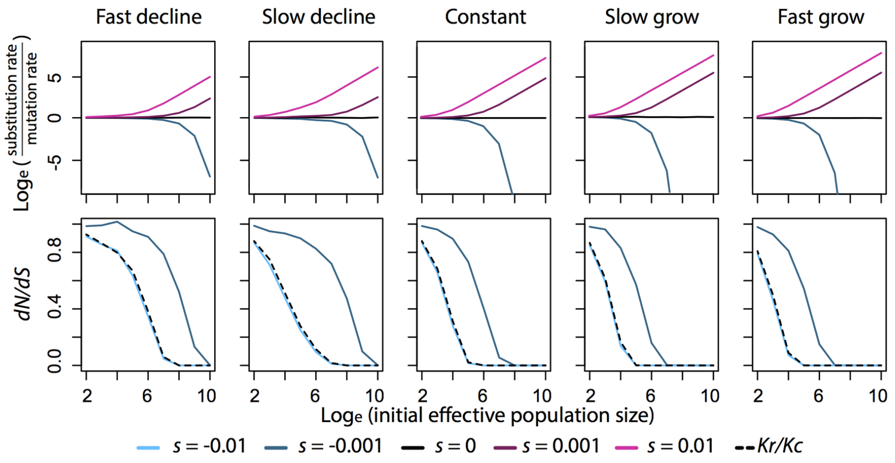


FIGURE 1 | Relationship between substitution rate, *dN/dS*, *Kr/Kc*, and effective population size (N_e) under different scenarios of population growth and selection coefficient (s). Fixation probability is calculated by the general formula described in the text. To approximate the expectation value in the formula, 500000 replicates of an allele are simulated under a Wright-Fisher model described in Waxman (2011). In each simulation, selection coefficient stays constant and population size changes deterministically over generations. When population grows, the final population size is 1.9 times the initial population size. When population declines, the final population size is 0.1 times the initial population size. Fast population change takes 20 generations to reach the final population size. Slow population change takes 200 generations. Substitution rate is then calculated from the fixation probability and is plotted on natural log scale with respect to mutation rate, so when substitution rate equals mutation rate, the value is 0 in the plots. *dN/dS* is calculated as the ratio between the fixation probability of a slightly deleterious mutation ($s = -0.001$) to that of a neutral mutation ($s = 0$). *Kr/Kc* is calculated as the ratio between the fixation probability of a deleterious mutation ($s = -0.01$) to that of a slightly deleterious mutation ($s = -0.001$).

not be a robust way to test our theoretical understanding of substitutions against empirical data. We suggest that the formula of fixation probability should be explicitly accounted in future analyses on the relationship between population size and substitution rates. In particular, the formula can be written into a form similar to logistic regression, where the selection coefficient is the regression coefficient and the expectation term becomes a parameter in the link function to estimate.

We have considered the way that increases in mutation rate should flow through to increases in substitution rate, in many classes of substitution (neutral, advantageous, slightly deleterious). Now we can look at the way the accumulation of substitutions makes populations genetically distinct from each other. In the next section, we will consider how isolation between populations is achieved either by spatial separation or through the acquisition of genetic traits that reduce the chances of interbreeding. Once populations are isolated, they will accumulate unique sets of substitutions.

ISOLATED POPULATIONS ACQUIRE DIFFERENT SUBSTITUTIONS

The simplest way for populations to become isolated is through the imposition of a physical barrier to the movement of alleles from one population to another. When a once-continuous population is divided by unsuitable habitat, there are three possible outcomes. One is that the barrier continues to prevent gene flow between the divided populations. The second is that the barrier is removed, allowing individuals or gametes to move between populations once more. The third is that one or both of the daughter populations acquires adaptations that allow them to overcome the barrier and re-establish contact with the other population. The first two processes are entirely determined by the frequency at which isolating mechanisms are created or removed. But the third possibility is a biological process and its frequency will depend not only on the nature of the isolating mechanism but also on the ability of either daughter population to adapt to the novel intervening habitat and overcome the barrier. Adaptation is found to occur on both new mutations and standing genetic variation (Olson-Manning et al., 2012). Standing genetic variation is also positively correlated with direct estimates of mutation rate (Vigouroux et al., 2002; Phillips et al., 2009). So, higher mutation rate could facilitate local adaptation, potentially allowing secondary contact between divided populations (Barton, 2001; Sexton et al., 2009).

It's possible for substitutions to cause genetic isolation between populations in the absence of a physical barrier to interbreeding. Divergent selection for different substitutions within the population may contribute to non-random mating (Servedio et al., 2011). For example, divergent selection on flowering time has been detected in two genetically distinct populations of *Mimulus guttatus*, causing temporal isolation between the two populations (Lowry et al., 2008), and divergent selection on color pattern mimicry has been found to cause assortative mating in sister species *Heliconius melpomene* (Jiggins et al., 2001). Such substitutions can happen rapidly under local adaptation, generating genetically isolating populations (Servedio et al., 2011), so the effectiveness of this process will depend on how common these kinds of disruptive traits are, and on the rate of generation of relevant alleles by mutation. Therefore, the major limit on the evolution of non-random mating is the initial frequency of alleles associated with these disruptive traits in the populations, which is related to the mutation rate and the rarity of the traits. In theory, any substitutions that make individuals more likely to mate with individuals of the same population will facilitate genetic isolation between populations. The most general case is when substitutions make immigrants much less viable and/or fertile than local individuals (Servedio et al., 2011).

If divergent selection is removed, then non-random mating may breakdown. For example, light gradient has been found to drive divergent evolution of female sensory bias in some cichlid fish, such that females at different water depths prefer different male colors (Seehausen et al., 2008). But the same species living in turbid water does not show the same preference, presumably because light of the certain wavelengths is absorbed so fast that most water depths have monochromatic light, under which females cannot discriminate color (Seehausen and van Alphen, 1998).

We have now considered how sub-populations can acquire different sets of substitutions, either because physical barriers prevent movement of substitutions between them, or because divergent selection fixes different traits in different parts of a species' range which directly or indirectly engender genetic isolation between populations. Now we will consider how the acquisition of these different sets of substitutions can contribute to genetic incompatibility between populations, paving the way for the formation of new species.

SUBSTITUTIONS LEAD TO INCOMPATIBILITY

Given some level of population isolation, genetic differences between populations will inevitably accumulate over time. If these genetic differences reduce the chance of successful reproduction between members of different populations, they contribute to the reproductive isolation (RI), preventing or reducing gene flow between populations. Broadly speaking, there are three ways that substitutions can contribute to reproductive isolation: by local adaptation of populations making hybrids unfit in either of the parent environment (e.g., Via et al., 2000), by influencing the chance of successful mating or fertilization (prezygotic isolation: e.g., Quinn et al., 2000), or by generating sets of alleles that reduce hybrid viability and/or fertility when mixed with alleles from other populations (post-zygotic isolation, or genomic incompatibility: e.g., Yamamoto et al., 2010).

However, there may be no clear line between the different ways that alleles contribute to RI. For example, local adaptation that involves changes to metabolic genes might involve coordinated changes to both mitochondrial and nuclear genes, potentially causing cytonuclear conflicts if mitochondrial alleles from one population are combined with nuclear alleles from another population (Johnson, 2010). Furthermore, if mutations conferring local adaptation are physically linked on the same chromosome to mutations causing genomic incompatibility, locally incompatible regions of the genomes ("speciation islands") can reduce interbreeding between populations and accelerate their genetic differentiation (Via, 2009, 2012). It is possible for RI to evolve within a population, for example through polyploidy or chromosomal rearrangement, as long as the mutant individuals can reproduce by selfing or mating with other mutants, but have greatly reduced chance of successfully combining chromosomes with wildtype members of the population (Ptacek et al., 1994; Guelbeogo et al., 2005; Wood et al., 2009; Twyford and Friedman, 2015).

Following Dobzhansky's pioneering experiments in the early 1900s that involved introgressing small regions of the genome from one species to another (Dobzhansky, 1936), "speciation genes" have been identified which confer low fitness in hybrid genetic backgrounds but not in their original genetic backgrounds. The genomic incompatibility caused by these genes stems from antagonistic interactions between parental genomes. Empirical studies on speciation genes have agreed on four general patterns (Coyne and Orr, 2004; Welch, 2004; Presgraves, 2010). First, the completion of RI can involve multiple incompatibilities. Second, the evolution of each incompatibility can be driven by multiple substitutions. Third, these substitutions are often found in uniparentally inherited genes. Fourth, these substitutions cause significant decrease in individual fitness only when they are in particular combinations.

Most of the models developed to account for these patterns are derivatives of the Dobzhansky–Muller-Incompatibility (DMI) model (Dobzhanksy, 1937; Muller, 1942). The basic argument of a DMI model is that combinations of alleles found in high frequencies in a given population must be compatible with each other; otherwise they would be removed from the population by purifying selection. Given that any individual unfortunate enough to inherit a deleterious combination of alleles has lower fitness, so will pass on fewer copies of those alleles to the next generation, we expect incompatible alleles to reduce in frequency within an interbreeding population. But alleles in one isolated population are not tested against alleles in other populations, so combinations of alleles that can only be formed by hybridization between populations may have low overall fitness. The DMI model is particularly good at explaining the four general patterns of "speciation genes," because it predicts that incompatibility is a feature of particular allele combinations, and it assumes that each incompatibility involves multiple substitutions as it requires different substitutions in different populations (Welch, 2004). Note that, although the DMI model was originally proposed to describe the evolution of genomic incompatibility, it can be generalized to model the other two ways that substitutions can contribute to reproductive isolation through prezygotic incompatibility or local adaptation. We can use the DMI model in any cases by where substitutions have positive or neutral effects on fitness in their own populations, but hybrids may have deleterious combinations of alleles from both parent populations and so contribute to RI.

When a population is divided, each subpopulation will acquire substitutions, and every substitution has some chance of creating incompatibility between the subpopulations. Orr (1995) formulated the DMI model by assuming that each derived allele in a population has equal probability of being incompatible with each of the derived alleles in the other population. As a result, the number of incompatibilities is the number of all possible combinations of derived alleles, which increases with the square of the number of substitutions fixed between the two populations. This prediction that the rate of acquisition of RI increases quadratically with genetic differences between populations is called the snowball effect (Orr, 1995). While empirical studies have given some support for the snowball effect (Matute et al., 2010; Moyle and Nakazato, 2010; Wang et al., 2015), it has been suggested that these results suffer from inaccurate estimates of genetic divergence between species (Stadler et al., 2012) or overestimates in the frequency of hybrid inviability (Barbash, 2011).

The assumption that every derived allele in a population has equal probability of being incompatible with each of the derived alleles in the other population does not seem realistic, as we might expect alleles associated with the same function to have a greater risk of being incompatible than substitutions affecting distinct functions that do not interact. For example, if different copies of a duplicated gene are silenced by mutations in different populations, then a hybrid may inherit both silenced copies and so lose the gene function (Masly et al., 2006; Bikard et al., 2009). Similarly, if one population fixes two alleles in succession, one of which compensates for the deleterious effect of the other allele, then a hybrid may suffer reduction in fitness if it inherits only one of the alleles without the compensating effect of the other allele. Some speciation genes are associated with the suppression of molecular drive by cytoplasmic genomes, pathogens or selfish genetic elements, which favor their own transmission at the expense of fitness of gametes not carrying them (Johnson, 2010; Presgraves, 2010). A hybrid between two separate populations might have incompatible sets of alleles, having the elements from one population but the suppression mechanisms of another. Situations such as these that involve pairs of compatible alleles could drive a linear relationship between the number of incompatibilities and genetic divergence between populations. So if paired substitutions affecting related functions are the primary cause of genomic incompatibility, then we might expect the rate of increase in RI to be less than quadratic.

But we should not expect all incompatible alleles to be paired up. Indeed, observation suggests that multiple substitutions are often needed to confer an incompatibility (Coyne and Orr, 2004; Welch, 2004). Given the complexity in the way derived alleles interact to cause incompatibilities, we might not expect either a strict linear or quadratic increase in RI with the number of substitutions fixed between two populations. If incompatibilities can be identified from genetic mapping data, then the number of incompatibilities can be regressed against the number of substitutions fixed between species (Matute et al., 2010; Moyle and Nakazato, 2010). Then we could use the degree of greater-than-linear increase as a continuous measure of how many alleles in one population can be incompatible with each allele in another population. We can also explicitly model the accumulation of incompatibilities along phylogenies under different models and compare the goodness-of-fit between models against the observed number of incompatibilities among species (Wang et al., 2013).

We have seen that species differ in the rate of supply of new mutations, and that this should influence the rate of acquisition of substitutions. Isolation between populations – whether by physical, behavioral or genetic barriers to interbreeding – will cause different sets of substitutions to accumulate in sister populations. These substitutions must be compatible with other alleles in the population, but may be incompatible with substitutions accumulated independently in the sister populations. Now we will consider how the accumulation of incompatible substitutions leads to the formation of new species.

INCOMPATIBILITY LEADS TO SPECIATION

Both spatial isolation and genetic isolation may drive speciation if the isolating conditions can be sustained for a sufficiently long time. The DMI model suggests that isolation makes speciation inevitable, as divided populations will eventually accumulate sufficient differences to prevent the formation of successful hybrids during secondary contact. Even if RI is not complete during secondary contact, as long as there is some form of selection against hybridization and the selection is strong enough, traits that facilitate premating isolation can be selected for to complete RI (Servedio and Noor, 2003; Otto et al., 2008; Bank et al., 2012). The more substitutions contribute to RI, the stronger is the selection against hybridization. So in order to link mutation, substitution and speciation, we need to ask how the rate of accumulation of substitutions is linked to the rate of formation of reproductive isolation.

To answer this question, Gavrilets and Gravner (1997) extended the DMI model to the holey landscape model, which considers an adaptive landscape in which the "holes" are unfit genotypes. The holey landscape model demonstrates the conditions under which incompatible substitutions can be fixed in different populations, and provides quantitative predictions concerning the number of substitutions and the amount of time required to reach RI. Because strongly deleterious mutations are unlikely to be fixed, we need to look for the conditions under which incompatible substitutions can be fixed in different populations without incurring large fitness costs. In other words, we want to find evolutionary paths that move along the adaptive landscape without falling into a hole. Given a DMI prediction on the relationship between incompatibilities and substitutions, the holey landscape model is able to predict the number of evolutionary paths that a genotype can move along (Gavrilets, 2004).

When all the incompatible substitutions are paired, the number of incompatibilities is a linear function of the number of substitutions εd, where ε is the probability that each substitution causes reduction in fitness and d is the number of substitutions that differ from a starting genotype. If the starting genotype has L mutational targets, there are L number of steps the genotype can take with one substitution and the probability that any step will not lead to a hole is the chance that the next substitution is compatible with any previous substitutions and so causes no reduction in fitness: 1–ε. Then the expected number of paths for the starting genotype is $L(1–ε)$. So the conditions where the starting genotype can move along the fitness landscape without falling into a hole is when $L(1–ε) > 1$, which is almost always true. Following similar arguments, the conditions for a snowball effect (quadratic increase in incompatibility) is $ε < \ln L/L$ (Gavrilets, 2004). In this case, having more mutational targets (L) should have lower values of ε to fulfill the condition of a snowball effect. In principle, ε can be estimated from a regression model between the number of incompatibilities and the number of substitutions fixed between species.

Assuming the simplest case where any one pair of incompatible substitutions causes complete RI, and all substitutions have the same chance of being incompatible, we can consider the probability that two populations differ in d substitutions can still interbreed. Two populations can interbreed if the path of d substitutions between them along the landscape does not fall in a hole, the probability of which is $(1-ε)^d$. The probability that the k-th substitution causes RI is $(1-ε)^{k-1}ε$. So the expectation of the number of substitutions required to complete RI is $\sum_{k=1}^{\infty} k(1-ε)^{k-1}ε = 1/ε$. The expected amount of time to complete RI is then the ratio between the number of substitutions to reach RI and the substitution rate (Gavrilets, 2004). This model allows us to make predictions concerning the relationship between mutation rates, substitution rates and time to speciation. If each step along the evolutionary paths of the populations is neutral or nearly neutral, the overall substitution rate in a population should be primarily determined by the mutation rate (see Mutation Rate Influences Substitution Rate). If the incompatible substitutions are the result of adaptation, then this will reduce the time to reach complete RI, proportional to the product between effective population size and selection coefficient (Gavrilets, 2004). Because this calculation is based solely on new mutations and neglects the role of standing genetic variation within populations, it may overestimate time to achieve RI (Gavrilets, 2004). To account for standing genetic variation, one can numerically model changes in genetic variation both within and between populations (Gavrilets, 1999).

Clearly, the waiting time to speciation will depend on many interacting factors, including the rate of supply of genetic variation, the level of gene flow between isolated populations, the nature of the genetic changes underlying reproductive isolation, population size and nature of selective pressures. The holey landscape model only provides general predictions on the relationship between mutation rate, substitution rate, and speciation rate. But general predictions are useful for comparative studies, in which the influences of confounding factors may be treated as random effects when sample size is large.

Now that we have seen how species differ in mutation rates, and how these will influence the rate at which populations acquire substitutions that will cause them to become genetically incompatible, we can explore the relationship between the formation of genetically isolated populations and the rate of diversification. To do this, we need to consider the possible evolutionary fates of newly isolated populations once they have formed.

SPECIATION DRIVES DIVERSIFICATION

Diversification is the net result of the processes that change the number of independent evolving lineages. The possible component processes of diversification are speciation adding lineages, extinction removing lineages, and merger of existing lineages through hybridization. So to understand how the processes that lead to genetically isolated populations contribute to diversification, we need to consider the factors that influence whether these new isolated populations will persist. If they persist, they may potentially divide again. Or, one or more of them may go extinct, resulting in the loss of any unique substitutions that had accumulated in that species. Or, if RI is not complete, isolated populations may reconnect and merge genetically with each other during secondary contact, losing their separate identity and becoming a single intermixed lineage.

While the process of speciation is typically studied by comparing closely related populations, diversification is usually studied by comparing the diversity of lineages over time and space. Because we are focusing only on evolutionary analysis of DNA sequences, we will not attempt to consider the ways that palaeontological, taxonomic and biogeographic data are used to shed light on the process of diversification. Instead we will consider only how diversification rate is measured from molecular phylogenies. Rather than considering the existence of RI between populations, phylogenetic studies of diversification rate typically rely on counting the number of recognized taxa within genetically distinct lineages. Speciation rate is typically estimated from the distribution of branching events in a phylogeny (e.g., Pagel et al., 2006), and differences in the net diversification rate estimated by comparing the number of extant species per lineage (e.g., Bromham et al., 2015).

These phylogenetic measures of diversification rate do not map exactly to the process being considered in speciation models. Taxonomic counts of species in a lineage are typically based on the number of physically or biogeographically distinct forms rather than direct measures of reproductive isolation. Phylogenetic identification of species recognizes populations that show deep genetic divergence (Simpson, 1962; de Queiroz, 1998), but not all the genetic differences between populations cause RI, and in some cases previously isolated populations can rejoin and fuse (Coyne and Orr, 2004), and physically or spatially distinct forms may be connected by interbreeding populations (Irwin et al., 2001). So phylogenetic estimates of diversification are not direct measures of time to reach RI as predicted by studies on speciation (Wiens, 2004).

To understand the relationship between phylogenetic patterns and speciation processes, we need to consider the interplay between population isolation, secondary contact, and reproductive isolation, and the relative amounts of evolutionary time between them. First we can define the expected waiting time between isolating events, which we will call T_I (time to

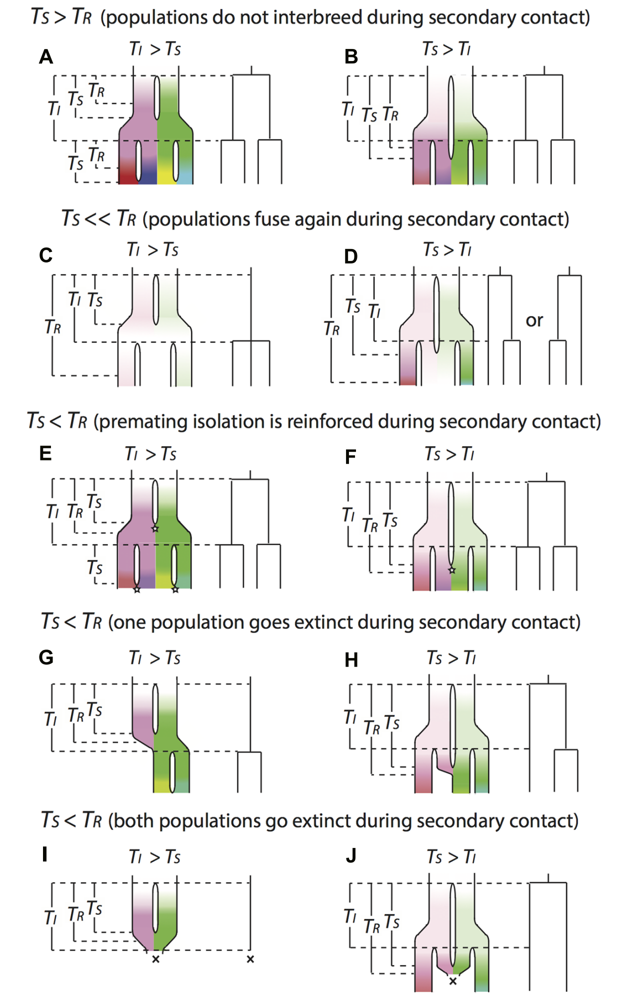

$T_S > T_R$ (populations do not interbreed during secondary contact)
A
$T_I > T_S$
T_I T_S T_R
T_S T_R
B
$T_S > T_I$
T_I T_S T_R
$T_S << T_R$ (populations fuse again during secondary contact)
C
$T_I > T_S$
T_R T_I T_S
D
$T_S > T_I$
T_R T_S T_I
or
$T_S < T_R$ (premating isolation is reinforced during secondary contact)
E
$T_I > T_S$
T_I T_R T_S
T_S
F
$T_S > T_I$
T_I T_R T_S
$T_S < T_R$ (one population goes extinct during secondary contact)
G
$T_I > T_S$
T_I T_R T_S
H
$T_S > T_I$
T_I T_R T_S
$T_S < T_R$ (both populations go extinct during secondary contact)
I
$T_I > T_S$
T_I T_R T_S
J
$T_S > T_I$
T_I T_R T_S

FIGURE 2 | Illustration for the conceptual approach to relating events of reproductive isolation to bifurcations on molecular phylogenies. Each scenario includes the true history of diversification on the left and the reconstructed molecular phylogeny on the right, assuming the phylogeny is accurately reconstructed. Gaps between lineages in the true history indicate population isolation. The color indicates the genetic distinctiveness of each lineage (in terms of how many different substitutions the populations have acquired). The brighter the color is, the more distinctive the lineage is to the others. Diversification starts with a white lineage that bifurcates into a pink and a green lineage. During the process, populations are expanding and encounter new isolation events. Then pink lineage bifurcates into a red and a purple lineage, and green lineage bifurcates into a yellow and a blue lineage. T_I is the expected waiting time between isolating events, which are kept the same under all the scenarios in order to make the internodal branch lengths of the reconstructed phylogenies under different scenarios comparable to each other. T_S is the amount of time that populations remain isolated. T_R is the amount of time that populations accumulate enough incompatibilities to reach complete RI. Stars represent reinforcement events and crosses represent extinction events. **(A–J)** describe different evolutionary scenarios: see text for details.

isolation). If isolating events are physical barriers preventing interbreeding, then T_I may depend on chance environmental factors that divide populations. If isolating events are caused by substitutions that cause non-random mating, then T_I depends not only on the occurrence of environmental factors that trigger divergent selection on the subpopulations, but also the amount of time the populations take to evolve non-random mating.

Similarly, we can consider the expected waiting time for two previously divided populations to come back into contact, reestablishing the potential for gene flow (T_S, time to secondary contact). In the case of physical barriers, T_S could depend, like T_I, on chance environmental factors that remove isolating barriers (e.g., a river changing course), or T_S may be due to ecological succession (e.g., burnt forest re-growing). Alternatively, T_S might be governed by genetic change within the population, if one or both of the divided populations is able to adapt to the intervening unsuitable habitat. So when populations are isolated by unsuitable habitat, T_S equals the length of time the intervening habitat remains unsuitable or the amount of time it takes for populations to adapt to the unsuitable habitat, whichever is shorter. When isolation is due to non-random mating, T_S may be rapid (e.g., if a sudden change in environment removes divergent selection) or more gradual (e.g., if the breakdown in divergent selection is due to selection for the same new adaptive trait in both populations).

Finally, we can consider the waiting time to complete RI, T_R, which describes the amount of time needed to accumulate sufficient incompatibilities to become irreversibly genetically incompatible, even if secondary contact between populations is restored. At T_R, members of the population can no longer interbreed even if brought back into contact.

We can consider the pattern of lineage evolution under different relative values of T_I, T_S and T_R (**Figure 2**). When the waiting time to secondary contact tends to be greater than the time needed to complete RI ($T_S > T_R$), previously isolated populations won't be able to interbreed if they come into secondary contact. In these circumstances, isolation leads to distinct species, and the internodal branch lengths of an accurately reconstructed phylogeny should equal T_I (**Figures 2A–B**). In this situation, the diversification rate estimated from branching events reflects the frequency of the isolating events.

When waiting time to secondary contact is much less than the time needed to develop complete reproductive isolation ($T_S \ll T_R$), previously isolated populations can interbreed if they come into secondary contact. If waiting time to secondary contact is shorter than the typical period between events that separate populations ($T_I > T_S$), then separated populations are able to fuse again before the next isolation event happens. If we were to reconstruct this history on a phylogeny, we would not be able to detect the period of isolation, instead we would detect only the most recent isolating event that causes a split (**Figure 2C**). If waiting time to secondary contact is longer than the period between isolating events ($T_I < T_S$), the next isolation event happens before populations have a chance to merge, keeping the most spatially distant populations isolated (**Figure 2D**). If we were to reconstruct this history on a phylogeny, we would observe the most spatially distant populations having deepest genetic divergence (even though they may not be the oldest isolating events), and the phylogenetic relationships between these populations and the population that is spatially distributed in the middle of them might not be fully resolved.

When waiting time to secondary contact is slightly less than the time needed to develop complete reproductive isolation ($T_S < T_R$), some degree of reproductive isolation has evolved between populations, so hybrids produced by secondary contact will have reduced fitness, in which case there are two possible outcomes. First, traits that facilitate premating isolation may be selected for to prevent production of low-fitness hybrids, and this selection will favor mutations that bring about complete RI. Under this scenario, we would observe similar phylogenetic patterns as the case of $T_S > T_R$ (**Figures 2E,F**). Second, the loss of reproductive output on unsuccessful hybrid mating could be severe enough to result in extinction of one or both populations (Todesco et al., 2016). If one population goes extinct, we would observe similar phylogenetic patterns as the case of $T_S \ll T_R$ (**Figures 2G,H**). If isolation events are less frequent than secondary contact events ($T_I > T_S$; **Figure 2I**), then both populations may go extinct on merger due to loss of reproductive output. However, an isolation event may save a population from extinction by hybridization, allowing persistence of the lineage, leading to a pattern of species with discrete spatial distributions. In this case each internode on the phylogeny would represent multiple isolating and merging events (**Figure 2J**).

In our illustration, we assumed that the values of T_I, T_S, and T_R were constant over the phylogeny, but it might be more realistic to model them as random variables. If the values can change over time and between lineages, then all the different scenarios illustrated in **Figure 2** may be found in a single reconstructed phylogeny reconstructed from the sequence data of a species group.

Using our conceptual approach to relating events of reproductive isolation to bifurcations on molecular phylogenies, we can make some predictions about the role of mutation rate in shaping average speciation rates. As outlined in the previous section, we expect mutation rate to have a role in determining the rate of formation of reproductive isolation between populations, so higher mutation rate should result in shorter T_R. When the rate of secondary contact is controlled primarily by environmental change, populations with shorter T_R are more likely to evolve reproductive isolation before secondary contact, so these populations will show the phylogenetic patterns under the scenario of $T_S > T_R$ (**Figures 2A,B**), where internodal branch lengths are shorter and speciation rate is higher than other scenarios. Higher mutation rate may also result in faster evolution of non-random mating under divergent selection and so shorter T_I. Populations with shorter T_I are more likely to show the phylogenetic patterns under scenarios of $T_1 < T_S$ (**Figures 2B,D,F,H,J**), where speciation rate is also higher than the corresponding scenarios of $T_1 > T_S$ (**Figures 2A,C,E,G,I**). As a result, greater mutation rates may lead to faster speciation due to more rapid accumulation of reproductive isolation and/or faster evolution of non-random mating. But when secondary contact is driven by one or both of the divided populations evolving to occupy the intervening habitat, we would expect populations with faster mutation rate to have both shorter T_S and T_R, because greater genetic variation in populations may speed both adaptation to novel environments and the evolution of reproductive isolation. Under this situation, whether faster mutation can accelerate speciation depends on the relative magnitude of T_S versus T_S, which seems unlikely to have a general pattern.

It is also possible that mutation rate could influence diversification rate through relative extinction rates. Several theoretical studies have suggested that populations with faster mutation rate should adapt to novel environments faster (Lynch, 1996; Barton and Partridge, 2000; Stockwell et al., 2003; Orr and Unckless, 2008). This may provide some reduction in extinction rates if extinction is largely driven by failing to adapt to environmental changes. However, some population genetic models have predicted an opposite effect, where the accumulation of deleterious mutations in extremely small populations causes extinction by “mutational meltdown” (Lynch et al., 1995; Lande, 1998). Due to difficulties in measuring the extinction rate for most lineages, it is difficult to test these hypotheses directly. In terms of the effect on diversification rates measured from phylogenies, extinction influences phylogenetic branch lengths because in most cases we expect to have direct evidence for only a relatively small proportion of all of the extinct species. Missing data from the extinct lineages increases internodal branch lengths in the phylogeny and decreases estimates of diversification rate.

LINKING MUTATION RATES TO DIVERSIFICATION RATES

We have discussed how the rate of supply of new genetic variation is determined by the mutation rate that varies between species. We have also seen how patterns and rates of fixation of mutations in populations are influenced by factors that are in themselves shaped by species traits. Therefore we expect that patterns and rates of accumulation of genetic differences will differ between lineages. The acquisition of substitutions causes populations to diverge from each other, until they become so different that they cannot freely interbreed. These links – more mutations, more substitutions, more incompatibilities, a higher rate of speciation, and a higher rate of diversification – connect biochemical events in single cells to the generation of biodiversity. This is not to say that the production of variation by itself explains the process of diversification. Clearly, individual speciation events will be driven by particular local circumstances and the biology of the organism in question. But given that a link between mutation and diversification has been detected in comparative studies, we conclude that the production of variation makes some degree of contribution to the rate of evolutionary change at both the microevolutionary and macroevolutionary levels.

Empirical support for this connection between mutation and diversification comes primarily from molecular phylogenetic studies that show a correlation between estimates of rate of molecular evolution (estimated from phylogenies either from the tips, all branches, or root-to-tip paths) and measures of net diversification rate (either species richness or number of nodes in a tree). Association between rates of molecular evolution and species richness has been noted for a wide range of plants and animals (Barraclough and Savolainen, 2001; Davies et al., 2004; Pagel et al., 2006; Eo and DeWoody, 2010; Lancaster, 2010; Duchene and Bromham, 2013; Ezard et al., 2013; Bromham et al., 2015; Dugo-Cota et al., 2015). While individual studies may be subject to measurement biases or analytical artifacts, the diversity of approaches taken and the wide variety of data analyzed supports the contention that these studies reveal a widespread phenomenon.

While comparative studies demonstrate statistically significant and consistent patterns across many taxa, they don't reveal the underlying cause of the relationship between rates of molecular evolution and diversification. We have hypothesized how mutation rate could promote diversification, either by providing more raw materials for adaptation, or by contributing to the evolution of reproductive isolation, or both. Conversely, it has been suggested that the process of speciation could cause acceleration in rates of molecular evolution, resulting in longer phylogenetic branch lengths in more species-rich lineages (Pagel et al., 2006). However, this hypothesis is difficult to reconcile with studies that have identified a correlation between the synonymous substitution rate and species richness (Barraclough and Savolainen, 2001; Lanfear et al., 2007; Duchene and Bromham, 2013; Bromham et al., 2015). Given that variation in synonymous substitution rate is considered to reflect differences in the mutation rate, it is difficult to see how speciation can directly influence mutation rate.

Alternatively, the association between molecular rates and diversification rate may be an incidental consequence of both being associated with some other factor, although it is difficult to think of a convincing indirect link. While body size and other associated life history characteristics correlate with rate of molecular evolution in a wide range of taxa, size is a surprisingly poor predictor of species richness in animals (Owens et al., 1999; Orme et al., 2002; Stuart-Fox and Owens, 2003; Isaac et al., 2005). Population size could offer an indirect link, if greater rates of speciation or extinction were associated with consistent reduction in population size, which could increase the fixation of slightly deleterious substitutions, however, there is currently little empirical evidence to support either a consistently lower population size or increased dN/dS in species-rich lineages (Bromham et al., 2015).

Various attempts have been made to provide a general theory of the tempo and mode of evolution by following the causal chain from biochemical processes to macroevolutionary change, for example by linking available kinetic energy to individual metabolic rate to both mutation rate and generation time, which may then influence the rate of evolutionary change (Gillooly and Allen, 2007). Such theories have been used to explain spatial patterns in biodiversity, on the assumption that higher temperatures drive greater rates of genetic change, either directly through an effect on rate of biochemical reactions, or indirectly through faster life histories reducing generation time (Rohde, 1992; Gillman and Wright, 2013). If rate of phenotypic evolution or niche change is increased by higher mutation rate or faster generation turnover, then lineages in warmer environments might diversify more rapidly (Smith and Beaulieu, 2009). However, theories linking energy availability to molecular change to diversity have been challenged on both empirical (noting exceptions to the rule) and theoretical grounds (questioning the validity of the underlying assumptions) (Duncan et al., 2007; Price et al., 2012). The proposed links in the causal chains might often be overwhelmed by other evolutionary forces operating on particular species (Dowle et al., 2013; Bromham et al., 2015; Glazier, 2015).

In particular, we might expect some of the evolutionary feedback loops discussed in this paper to have an impact on the knock-on effects of environmental temperature on rate of molecular evolution. Species are not passive in the face of environmental variation in mutagens such as UV or temperature, instead there is evidence that they adapt to their local conditions (Albarracin et al., 2012; Miner et al., 2015; Svetec et al., 2016), which may iron out some of the predicted environmental variation in mutation rates. Similarly, while variation in growth rates and generation turnover may influence the rate of accumulation of DNA replication errors, it seems that copy frequency effects are modulated by selection on copy fidelity mechanisms in order to produce acceptable levels of per generation error rates (Drake et al., 1998; Bromham, 2011; Sung et al., 2012). Life history and mutation rate must be matched to the environment and optimized to each other if a lineage is to persist through evolutionary time.

CONCLUSION

The search for simple unifying theories in macroevolution and macroecology seems unlikely to succeed given the vast number of factors that can influence a particular lineage's evolutionary trajectory, including rare events and the weight of history. Patterns in biodiversity are shaped by a great many factors, both intrinsic and extrinsic to organisms. Both evidence and theory suggests that one such factor is variation in the mutation rate between species. But the explanatory power of the observed relationship between molecular rates and biodiversity is relatively modest, so it does not provide anything like the predictive power that might be hoped for in a unifying theory. However, we feel that the evidence is growing that, in addition to the many and varied influences on the generation of diversity, the differential rate supply of variation through species-specific differences in mutation rate has some role to play in generating different rates of diversification.

Consideration of the forces shaping molecular evolution provides one piece of an intricate macroevolutionary puzzle. Molecular phylogenetic analysis has given us the ability to be able to consider both molecular processes and diversification rates simultaneously, giving us a new tool with which to explore the connections between the supply of variation and the production of biodiversity. We can't help but think that Darwin would be pleased with these new views of the evolutionary process that molecular analyses afford us, as it offers the potential to demonstrate the links in the Darwinian chain that connects variation between individuals to divergence between populations to the generation of biodiversity:

> *"It cannot be asserted that organic beings in a state of nature are subject to no variation; it cannot be proved that the amount of variation in the course of long ages is a limited quantity; no clear distinction has been, or can be, drawn between species and well-marked varieties. It cannot be maintained that species when intercrossed are invariably sterile, and varieties invariably fertile; or that sterility is a special endowment and sign of creation.... But the chief cause of our natural unwillingness to admit that one species has given birth to other and distinct species, is that we are always slow in admitting any great change of which we do not see the intermediate steps." (Darwin, 1859)*

AUTHOR CONTRIBUTIONS

XH and LB designed the work and wrote the article. XH developed models and conducted analyses.

FUNDING

We thank the Australian Research Council for funding this work.

ACKNOWLEDGMENTS

We thank David Waxman for allowing us use his simulation code.

REFERENCES

Agrawal, A. F., and Wang, A. D. (2008). Increased transmission of mutations by low-condition females: evidence for condition-dependent DNA repair. *PLoS Biol.* 6:e30. doi: 10.1371/journal.pbio.0060030

Albarracin, V., Pathak, G., Douki, T., Cadet, J., Borsarelli, C., Gurtner, W., et al. (2012). Extremophilic *Acinetobacter* strains from high-altitude lakes in Argentinean Puna: remarkable UV-B resistance and efficient DNA damage repair. *Orig. Life Evol. Biosph.* 42, 201–221. doi: 10.1007/s11084-012-9276-3

Allen, J. (1995). Separate sexes and the mitochondrial theory of ageing. *J. Theor. Biol.* 180, 135–140. doi: 10.1006/jtbi.1996.0089

Baer, C., Miyamoto, M. M., and Denver, D. R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.* 8, 619–631. doi: 10.1038/nrg2158

Bank, C., Hermisson, J., and Kirkpatrick, M. (2012). Can reinforcement complete speciation? *Evolution* 66, 229–239. doi: 10.1111/j.1558-5646.2011.01423.x

Barbash, D. A. (2011). Comment on "a test of the snowball theory for the rate of evolution of hybrid incompatibilities". *Science* 333:1576. doi: 10.1126/science.1202876

Barraclough, T. G., and Savolainen, V. (2001). Evolutionary rates and species diversity in flowering plants. *Evolution* 55, 677–683. doi: 10.1111/j.0014-3820.2001.tb00803.x

Barrett, R. D., and Schluter, D. (2008). Adaptation from standing genetic variation. *Trends Ecol. Evol.* 23, 38–44. doi: 10.1016/j.tree.2007.09.008

Barton, N. (2001). "Adaptation at the edge of a species' range," in *Integrating Ecology and Evolution in a Spatial Context*, eds J. Silvertown and J. Antonovics (London: Blackwell), 365–392.

Barton, N., and Partridge, L. (2000). Limits to natural selection. *Bioessays* 22, 1075–1084. doi: 10.1002/1521-1878(200012)22:12<1075::AID-BIES5<3.0.CO;2-M

Bartosch-Harlid, A., Berlin, S., Smith, N. G. C., Ller, A. P., and Ellegren, H. (2003). Life history and the male mutation bias. *Evolution* 57, 2398–2406. doi: 10.1111/j.0014-3820.2003.tb00251.x

Bendich, A. J. (2010). Hypothesis mitochondrial DNA, chloroplast DNA and the origins of development in eukaryotic organisms. *Biol. Direct* 5:42. doi: 10.1186/1745-6150-5-42

Bikard, D., Patel, D., Le Mette, C., Giorgi, V., Camilleri, C., Bennett, M. J., et al. (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* 323, 623–626. doi: 10.1126/science.1165917

Björnerfeldt, S., Webster, M. T., and Vilà, C. (2006). Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res.* 16, 990–994. doi: 10.1101/gr.5117706

Bradford, P. T., Goldstein, A. M., Tamura, D., Khan, S. G., Ueda, T., Boyle, J., et al. (2011). Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterizes the role of DNA repair. *J. Med. Genet.* 48, 168–176. doi: 10.1136/jmg.2010.083022

Bromham, L. (2009). Why do species vary in their rate of molecular evolution? *Biol. Lett.* 5, 401–404. doi: 10.1098/rsbl.2009.0136

Bromham, L. (2011). The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philos. Trans. Roy. Soc. B* 366, 2503–2513. doi: 10.1098/rstb.2011.0014

Bromham, L., Hua, X., Lanfear, R., and Cowman, P. F. (2015). Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. *Am. Nat.* 185, 507–524. doi: 10.1086/680052

Bromham, L., Rambaut, A., and Harvey, P. H. (1996). Determinants of rate variation in mammalian DNA sequence evolution. *J. Mol. Evol.* 43, 610–621. doi: 10.1007/BF02202109

Cardillo, M. (1999). Latitude and rates of diversification in birds and butterflies. *Proc. R. Soc. B* 266, 1221–1225. doi: 10.1098/rspb.1999.0766

Cardillo, M., Huxtable, J., and Bromham, L. (2003). Geographic range size, life history and rates of diversification in Australian mammals. *J. Evol. Biol.* 16, 282–288. doi: 10.1046/j.1420-9101.2003.00513.x

Chao, L., and Cox, E. C. (1983). Competition between high and low mutating strains of *Escherichia coli*. *Evolution* 37, 125–134. doi: 10.2307/2408181

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

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection: or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Davies, T. J., Savolainen, V., Chase, M. W., Moat, J., and Barraclough, T. G. (2004). Environmental energy and evolutionary rates in flowering plants. *Proc. R. Soc. B* 271, 2195–2200. doi: 10.1098/rspb.2004.2849

de Paula, W. B. M., Lucas, C. H., Agip, A. N., Vizcay-Barrena, G., and Allen, J. F. (2013). Energy, ageing, fidelity and sex: oocyte mitochondrial DNA as a protected genetic template. *Philos. Trans. R. Soc. B* 368, 20120263. doi: 10.1098/rstb.2012.0263

de Queiroz, K. (1998). "The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations," in *Endless Forms: Species and Speciation*, eds D. J. Howard and S. H. Berlocher (New York, NY: Oxford University Press), 57–75.

Denamur, E., and Matic, I. (2006). Evolution of mutation rates in bacteria. *Mol. Microbiol.* 60, 820–827. doi: 10.1111/j.1365-2958.2006.05150.x

Dobzhansky, T. (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics* 21, 113–135.

Dobzhanksy, T. (1937). *Genetics and the Origin of Species*. New York, NY: Columbia University Press.

Dowle, E., Morgan-Richards, M., and Trewick, S. (2013). Molecular evolution and the latitudinal biodiversity gradient. *Heredity* 110, 501–510. doi: 10.1038/hdy.2013

Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. U.S.A.* 88, 7160–7164. doi: 10.1073/pnas.88.16.7160

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics* 148, 1667–1686.

Duchene, D., and Bromham, L. (2013). Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species-richness in the Proteaceae. *BMC Evol. Biol.* 13:1. doi: 10.1186/1471-2148-13-65

Dugo-Cota, Á., Castroviejo-Fisher, S., Vilà, C., and Gonzalez-Voyer, A. (2015). A test of the integrated evolutionary speed hypothesis in a Neotropical amphibian radiation. *Global Ecol. Biogeogr.* 24, 804–813. doi: 10.1111/geb.12318

Duncan, R. P., Forsyth, D. M., and Hone, J. (2007). Testing the metabolic theory of ecology: allometric scaling exponents in mammals. *Ecology* 88, 324–333. doi: 10.1890/0012-9658200788[324:TTMTOE]2.0.CO;2

Durand, E., Tenaillon, M. I., Ridel, C., Coubriche, D., Jamin, P., Jouanne, S., et al. (2010). Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds. *BMC Evol. Biol.* 10:2. doi: 10.1186/1471-2148-10-2

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58, 465–523. doi: 10.1007/bf00623322

Eo, S. H., and DeWoody, J. A. (2010). Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles. *Proc. R. Soc. B* 277, 3587–3592. doi: 10.1098/rspb.2010.0965

Eyre-Walker, A., and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8, 610–618. doi: 10.1038/nrg2146

Ezard, T. H., Thomas, G. H., and Purvis, A. (2013). Inclusion of a near-complete fossil record reveals speciation-related molecular evolution. *Methods Ecol. Evol.* 4, 745–753. doi: 10.1111/2041-210X.12089

Figuet, E., Nabholz, B., Bonneau, M., Carrio, E. M., Nadachowska-Brzyska, K., Ellegren, H., et al. (2016). Life-history traits, protein evolution, and the nearly neutral theory in amniotes. *Mol. Biol. Evol.* 33, 1517–1527. doi: 10.1093/molbev/msw033

Figuet, E., Romiguier, J., Dutheil, J. Y., and Galtier, N. (2014). Mitochondrial DNA as a tool for reconstructing past life-history traits in mammals. *J. Evol. Biol.* 27, 899–910. doi: 10.1111/jeb.12361

Foury, F., and Szczepanowska, K. (2011). Antimutator alleles of yeast DNA polymerase gamma modulate the balance between DNA synthesis and excision. *PLoS ONE* 6:e27847. doi: 10.1371/journal.pone.0027847

Frankham, R. (2012). How closely does genetic diversity in finite populations conform to predictions of neutral theory? Large deficits in regions of low recombination. *Heredity* 108, 167–178. doi: 10.1038/hdy.2011.66

Freeman-Gallant, C. R., Amidon, J., Berdy, B., Wein, S., Taff, C. C., and Haussmann, M. F. (2011). Oxidative damage to DNA related to survivorship and carotenoid-based sexual ornamentation in the common yellowthroat. *Biol. Lett.* 7, 429–432. doi: 10.1098/rsbl.2010.1186

Fridovich, I. (2004). Mitochondria: are they the seat of senescence? *Aging Cell* 3, 13–16. doi: 10.1046/j.1474-9728.2003.00075.x

Galtier, N., Blier, P. U., and Nabholz, B. (2009a). Inverse relationship between longevity and evolutionary rate of mitochondrial proteins in mammals and birds. *Mitochondrion* 9, 51–57. doi: 10.1016/j.mito.2008.11.006

Galtier, N., Jobson, R. W., Nabholz, B., Glemin, S., and Blier, P. U. (2009b). Mitochondrial whims: metabolic rate, longevity and the rate of molecular evolution. *Biol. Lett.* 5, 413–416. doi: 10.1098/rsbl.2008.0662

Gaut, B., Yang, L., Takuno, S., and Eguiarte, L. E. (2011). The patterns and causes of variation in plant nucleotide substitution rates. *Annu. Rev. Ecol. Evol. Syst.* 42, 245–266. doi: 10.1146/annurev-ecolsys-102710-145119

Gaut, B. S., Muse, S. V., Clark, W. D., and Clegg, M. T. (1992). Relative rates of nucleotide subsitution at the rbcL locus of monocotyledonous plants. *J. Mol. Evol.* 35, 292–303. doi: 10.1007/BF00161167

Gavrilets, S. (1999). A dynamical theory of speciation on holey adaptive landscapes. *Am. Nat.* 154, 1–22. doi: 10.1016/S0169-5347(97)01098-7

Gavrilets, S. (2004). *Fitness Landscapes and the Origin of Species*. Princeton, NJ: Princeton University Press.

Gavrilets, S., and Gravner, J. (1997). Percolation on the fitness hypercube and the evolution of reproductive isolation. *J. Theor. Biol.* 184, 51–64. doi: 10.1006/jtbi.1996.0242

Gillman, L. N., and Wright, S. D. (2013). Patterns of evolutionary speed: in search of a causal mechanism. *Diversity* 5, 811–823. doi: 10.3390/d5040811

Gillooly, J. F., and Allen, A. P. (2007). Linking global patterns in biodiversity to evolutionary dynamics using metabolic theory. *Ecology* 88, 1890–1894. doi: 10.1890/06-1935.1

Gillooly, J. F., Mccoy, M. W., and Allen, A. P. (2007). Effects of metabolic rate on protein evolution. *Biol. Lett.* 3, 655–660. doi: 10.1098/rsbl.2007.0403

Giraud, A. (2001). Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* 291, 2606–2608. doi: 10.1126/science.1056421

Glazier, D. S. (2015). Is metabolic rate a universal 'pacemaker' for biological processes? *Biol. Rev.* 90, 377–407. doi: 10.1111/brv.12115

Goho, S., and Bell, G. (2000). Mild environmental stress elicits mutations affecting fitness in *Chlamydomonas*. *Proc. R. Soc. B* 267, 123–129. doi: 10.1098/rspb.2000.0976

Gokkusu, C., Cakmakoglu, B., Dasdemir, S., Tulubas, F., Elitok, A., Tamer, S., et al. (2013). Association between genetic variants of DNA repair genes and coronary artery disease. *Genet. Test. Mol. Biomarkers* 17, 307–313. doi: 10.1089/gtmb.2012.0383

Guelbeogo, W. M., Grushko, O., Boccolini, D., Ouedraogo, P. A., Besansky, N. J., Sagnon, N. F., et al. (2005). Chromosomal evidence of incipient speciation in the Afrotropical malaria mosquito Anopheles funestus. *Med. Vet. Entomol.* 19, 458–469. doi: 10.1111/j.1365-2915.2005.00595.x

Haraguchi, Y., and Sasaki, A. (1996). Host-parasite arms-race in mutation modifications-indefinate escalation despite a heavy load? *J. Theor. Biol.* 183, 121–137. doi: 10.1006/jtbi.1996.9999

Hartley, C., Newcomb, R., Russell, R., Yong, C., Stevens, J., Yeates, D., et al. (2006). Amplification of DNA from preserved specimens shows blowflies were preadapted for the rapid evolution of insecticide resistance. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8757–8762. doi: 10.1073/pnas.0509590103

Herr, A. J., Ogawa, M., Lawrence, N. A., Williams, L. N., Eggington, J. M., Singh, M., et al. (2011a). Mutator suppression and escape from replication error?induced extinction in yeast. *PLoS Genet.* 7:e1002282. doi: 10.1371/journal.pgen.1002282

Herr, A. J., Williams, L. N., and Preston, B. D. (2011b). Antimutator variants of DNA polymerases. *Crit. Rev. Biochem. Mol. Biol.* 46, 548–570. doi: 10.3109/10409238.2011.620941

Hodgins-Davis, A., Rice, D. P., and Townsend, J. P. (2015). Gene expression evolves under a house-of-cards model of stabilizing selection. *Mol. Biol. Evol.* 32, 2130–2140. doi: 10.1093/molbev/msv094

Hodgkinson, A., and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Rev. Genet.* 12, 756–766. doi: 10.1038/nrg3098

Holmes, E. C. (2003). Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* 11, 543–546. doi: 10.1016/j.tim.2003.10.006

Hua, X., Cowman, P., Warren, D., and Bromham, L. (2015). Longevity is linked to mitochondrial mutation rates in rockfish: a test using poisson regression. *Mol. Biol. Evol.* 32, 2633–2645. doi: 10.1093/molbev/msv137

Irwin, D. E., Irwin, J. H., and Price, T. D. (2001). Ring species as bridges between microevolution and speciation. *Genetica* 112, 223–243. doi: 10.1023/A:1013319217703

Isaac, N. J., Jones, K. E., Gittleman, J. L., and Purvis, A. (2005). Correlates of species richness in mammals: body size, life history, and ecology. *Am. Nat.* 165, 600–607. doi: 10.1086/429148

James, J. E., Lanfear, R., and Eyre-Walker, A. (2016). Molecular evolutionary consequences of island colonization. *Genome Biol. Evol.* 8, 1876–1888. doi: 10.1093/gbe/evw120

Jerome, J. P., Bell, J. A., Plovanich-Jones, A. E., Barrick, J. E., Brown, C. T., and Mansfield, L. S. (2011). Standing genetic variation in contingency loci drives the rapid adaptation of *Campylobacter jejuni* to a novel nost. *PLoS ONE* 6:e16399. doi: 10.1371/journal.pone.0016399

Jiggins, C. D., Naisbit, R. E., Coe, R. L., and Mallet, J. (2001). Reproductive isolation caused by colour pattern mimicry. *Nature* 411, 302–305. doi: 10.1038/35077075

Johnson, K. P., and Seger, J. (2001). Elevated rates of nonsynonymous substitution in island birds. *Mol. Biol. Evol.* 18, 874–881. doi: 10.1093/oxfordjournals.molbev.a003869

Johnson, N. A. (2010). Hybrid incompatibility genes: remnants of a genomic battlefield? *Trends Genet.* 26, 317–325. doi: 10.1016/j.tig.2010.04.005

Kabzinski, J., Mucha, B., Cuchra, M., Markiewicz, L., Przybylowska, I. K., Dziki, A., et al. (2016). Efficiency of base excision repair of oxidative DNA damage and its impact on the risk of colorectal cancer in the polish population. *Oxid. Med. Cell. Longev* 2016:9. doi: 10.1155/2016/3125989

Knight, C. A., Molinari, N. A., and Petrov, D. A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* 95, 177–190. doi: 10.1093/aob/mci011

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., et al. (2012). Rate of de novo mutations, father's age, and disease risk. *Nature* 488, 471–475. doi: 10.1038/nature11396

Kujoth, G. C., Hiona, A., Pugh, T. D., Someya, S., Panzer, K., Wohlgemuth, S. E., et al. (2005). Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science* 309, 481–484. doi: 10.1126/science.1112125

Lancaster, L. T. (2010). Molecular evolutionary rates predict both extinction and speciation in temperate angiosperm lineages. *BMC Evol. Biol.* 10:1. doi: 10.1186/1471-2148-10-162

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30, 314–334. doi: 10.2307/2407703

Lande, R. (1998). Risk of population extinction from fixation of deleterious and reverse mutations. *Genetica* 102, 21–27. doi: 10.1023/a:1017018405648

Lanfear, R., Ho, S. Y. W., Love, D., and Bromham, L. (2010). Mutation rate influences diversification rate in birds. *Proc. Natl Acad. Sci. U.S.A.* 107, 20423–20428. doi: 10.1073/pnas.1007888107

Lanfear, R., Thomas, J. A., Welch, J. J., and Bromham, L. (2007). Metabolic rate does not calibrate the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.* 104, 15388–15393. doi: 10.1073/pnas.0703359104

Larsson, N.-G. R. (2010). Somatic mitochondrial DNA mutations in mammalian aging. *Annu. Rev. Biochem.* 79, 683–706. doi: 10.1146/annurev-biochem-060408-093701

Lartillot, N. (2013). Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Mol. Biol. Evol.* 30, 356–368. doi: 10.1093/molbev/mss231

Lartillot, N., and Delsuc, F. (2012). Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution* 66, 1773–1787. doi: 10.1111/j.1558-5646.2011.01558.x

Lehtonen, J., and Lanfear, R. (2014). Generation time, life history and the substitution rate of neutral mutations. *Biol. Lett.* 10:20140801. doi: 10.1098/rsbl.2014.0801

Loeb, L. A., Wallace, D. C., and Martin, G. M. (2005). The mitochondrial theory of aging and its relationship to reactive oxygen species damage and somatic mtDNA mutations. *Proc. Natl. Acad. Sci.U.S.A.* 102, 18769–18770. doi: 10.1073/pnas.0509776102

Lowry, D. B., Modliszewski, J. L., Wright, K. M., Wu, C. A., and Willis, J. H. (2008). The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philos. Trans. R. Soc. B* 363, 3009–3021. doi: 10.1098/rstb.2008.0064

Lucas-Lledó, J. I., and Lynch, M. (2009). Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. *Mol. Biol. Evol.* 26, 1143–1153. doi: 10.1093/molbev/msp029

Lynch, M. (1996). "A quantitative-genetic perspective on conversation issues," in *Conservation Genetics: Case Studies From Nature*, eds J. C. Avise and J. L. Hamrick (London: Chapman & Hall), 471–501.

Lynch, M. (2007). *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Assoc.

Lynch, M. (2010). Evolution of the mutation rate. *Trends Genet.* 26, 345–352. doi: 10.1016/j.tig.2010.05.003

Lynch, M., Conery, J., and Burger, R. (1995). Mutation accumulation and the extinction of small populations. *Am. Nat.* 146, 489–518. doi: 10.1086/285812

Lynch, M., and Hill, W. G. (1986). Phenotypic evolution by neutral mutation. *Evolution* 40, 915–935. doi: 10.2307/2408753

Masly, J. P., Jones, C. D., Noor, M. A. F., Locke, J., and Orr, H. A. (2006). Gene transposition as a cause of hybrid sterility in *Drosophila*. *Science* 313, 1448–1450. doi: 10.1126/science.1128721

Mattimore, V., and Battista, J. R. (1996). Radioresistance of *Deinococcus radiodurans*: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. *J. Bacteriol.* 178, 633–637. doi: 10.1128/jb.178.3.633-637.1996

Matute, D. R., Butler, I. A., Turissini, D. A., and Coyne, J. A. (2010). A test of the snowball theory for the rate of evolution of hybrid incompatibilities. *Science* 329, 1518–1521. doi: 10.1126/science.1193440

Maynard Smith, J., and Szathmáry, E. (1995). *The Major Transitions in Evolution*. Oxford: W. H. Freeman.

McDonald, M. J., Hsieh, Y. Y., Yu, Y. H., Chang, S. L., and Leu, J. Y. (2012). The evolution of low mutation rates in experimental mutator populations of *Saccharomyces cerevisiae*. *Curr. Biol.* 22, 1235–1240. doi: 10.1016/j.cub.2012.04.056

Miner, B. E., Kulling, P. M., Beer, K. D., and Kerr, B. (2015). Divergence in DNA photorepair efficiency among genotypes from contrasting UV radiation environments in nature. *Mol. Ecol.* 24, 6177–6187. doi: 10.1111/mec.13460

Mohrenweiser, H. W., Wilson, D. M., and Jones, I. M. (2003). Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes. *Mutat. Res.* 526, 93–125. doi: 10.1016/S0027-5107(03)00049-6

Moyle, L. C., and Nakazato, T. (2010). Hybrid incompatibility "snowballs" between *Solanum* Species. *Science* 329, 1521–1523. doi: 10.1126/science.1193063

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

Nabholz, B., Glemin, S., and Galtier, N. (2008). Strong variations of mitochondrial mutation rate across mammals - the longevity hypothesis. *Mol. Biol. Evol.* 25, 120–130. doi: 10.1093/molbev/msm248

Nabholz, B., Uwimana, N., and Lartillot, N. (2013). Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. *Genome Biol. Evol.* 5, 1273–1290. doi: 10.1093/gbe/evt083

Nikolaev, S. I., Montoya-Burgos, J. I., Popadin, K., Parand, L., Margulies, E. H., Antonarakis, S. E., et al. (2007). Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20443–20448. doi: 10.1073/pnas.0705658104

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23, 263–286. doi: 10.1146/annurev.es.23.110192.001403

Oliver, A., Baquero, F., and Blázquez, J. (2002). The mismatch repair system (mutS, mutL and uvrD genes) in *Pseudomonas aeruginosa*: molecular characterization of naturally occurring mutants. *Mol. Microbiol.* 43, 1641–1650. doi: 10.1046/j.1365-2958.2002.02855.x

Oliver, A., Cantoìn, R., Campo, P., Baquero, F., and Blaìzquez, J. (2000). High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* 288, 1251–1253. doi: 10.1126/science.288.5469.1251

Oliver, A., and Mena, A. (2010). Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. *Clin. Microbiol. Infect.* 16, 798–808. doi: 10.1111/j.1469-0691.2010.03250.x

Olson-Manning, C. F., Wagner, M. R., and Mitchell-Olds, T. (2012). Adaptive evolution: evaluating empirical support for theoretical predictions. *Nat. Rev. Genet.* 13, 867–877. doi: 10.1038/nrg3322

Orme, C., Quicke, D., Cook, J., and Purvis, A. (2002). Body size does not predict species richness among the metazoan phyla. *J. Evol. Biol.* 15, 235–247. doi: 10.1046/j.1420-9101.2002.00379.x

Orr, H. A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* 139, 1805–1813.

Orr, H. A., and Unckless, R. L. (2008). Population extinction and the genetics of adaptation. *Am. Nat.* 172, 160–169. doi: 10.1086/589460

Otto, S. P., Servedio, M. R., and Nuismer, S. L. (2008). Frequency-dependent selection and the evolution of assortative mating. *Genetics* 179, 2091–2112. doi: 10.1534/genetics.107.084418

Owens, I. P., Bennett, P. M., and Harvey, P. H. (1999). Species richness among birds: body size, life history, sexual selection or ecology? *Proc. R. Soc. B* 266, 933–939. doi: 10.1098/rspb.1999.0726

Pagel, M., Venditti, C., and Meade, A. (2006). Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* 314, 119–121. doi: 10.1126/science.1129647

Pamplona, R., and Barja, G. (2011). An evolutionary comparative scan for longevity-related oxidative stress resistance mechanisms in homeotherms. *Biogerontology* 12, 409–435. doi: 10.1007/s10522-011-9348-1

Phillimore, A. B., Freckleton, R. P., Orme, C. D., and Owens, I. P. (2006). Ecology predicts large-scale patterns of phylogenetic diversification in birds. *Am. Nat.* 168, 220–229. doi: 10.1086/505763

Phillips, N., Salomon, M., Custer, A., Ostrow, D., and Baer, C. F. (2009). Spontaneous mutational and standing genetic (co)variation at dinucleotide microsatellites in *Caenorhabditis briggsae* and *Caenorhabditis elegans*. *Mol. Biol. Evol.* 26, 659–699. doi: 10.1093/molbev/msn287

Piechura, J. R., Tseng, T.-L., Hsu, H.-F., Byrne, R. T., Windgassen, T. A., Chitteni-Pattu, S., et al. (2015). Biochemical characterization of RecA variants that contribute to extreme resistance to ionizing radiation. *DNA Repair* 26, 30–43. doi: 10.1016/j.dnarep.2014.12.001

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., and Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13390–13395. doi: 10.1073/pnas.0701256104

Presgraves, D. C. (2010). The molecular evolutionary basis of species formation. *Nat. Rev. Genet.* 11, 175–180. doi: 10.1038/nrg2718

Price, C. A., Weitz, J. S., Savage, V. M., Stegen, J., Clarke, A., Coomes, D. A., et al. (2012). Testing the metabolic theory of ecology. *Ecol. Lett.* 15, 1465–1474. doi: 10.1111/j.1461-0248.2012.01860.x

Ptacek, M. B., Gerhardt, H. C., and Sage, R. D. (1994). Speciation by polyploidy in treefrogs: multiple origins of the tetraploid, *Hyla versicolor*. *Evolution* 48, 898–908. doi: 10.2307/2410495

Quinn, T. P., Unwin, M. J., and Kinnison, M. T. (2000). Evolution of temporal isolation in the wild: genetic divergence in timing of migration and breeding by introduced chinook salmon populations. *Evolution* 54, 1372–1385. doi: 10.1111/j.0014-3820.2000.tb00569.x

Reha-Krantz, L. J. (2010). DNA polymerase proofreading: multiple roles maintain genome stability. *Biochim. Biophys. Acta* 1804, 1049–1063. doi: 10.1016/j.bbapap.2009.06.012

Ries, G., Heller, W., Puchta, H., Sandermann, H., Seidlitz, H. K., and Hohn, B. (2000). Elevated UV-B radiation reduces genome stability in plants. *Nature* 406, 98–101. doi: 10.1038/35017595

Rohde, K. (1992). Latitudinal gradients in species diversity: the search for the primary cause. *Oikos* 65, 514–527. doi: 10.2307/3545569

Romiguier, J., Ranwez, V., Douzery, E. J. P., and Galtier, N. (2013). Genomic evidence for large, long-lived ancestors to placental mammals. *Mol. Biol. Evol.* 30, 5–13. doi: 10.1093/molbev/mss211

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. (2010). Viral mutation rates. *J. Virol.* 84, 9733–9748. doi: 10.1128/jvi.00694-10

Schaaper, R. M. (1998). Antimutator mutants in bacteriophage T4 and *Escherichia coli*. *Genetics* 148, 1579–1585.

Seehausen, O., Terai, Y., Magalhaes, I. S., Carleton, K. L., Mrosso, H. D. J., Miyagi, R., et al. (2008). Speciation through sensory drive in cichlid fish. *Nature* 455, 620–623. doi: 10.1038/nature07285

Seehausen, O., and van Alphen, J. J. M. (1998). The effect of male coloration on female mate choice in closely related Lake Victoria cichlids (*Haplochromis nyererei* complex). *Behav. Ecol. Sociobiol.* 42, 1–8. doi: 10.1007/s002650050405

Servedio, M. R., and Noor, M. A. F. (2003). The role of reinforcement in speciation: theory and data. *Annu. Rev. Ecol. Evol. Syst.* 34, 339–364. doi: 10.1146/annurev.ecolsys.34.011802.132412

Servedio, M. R., Van Doorn, G. S., Kopp, M., Frame, A. M., and Nosil, P. (2011). Magic traits in speciation: 'magic' but not rare? *Trends Ecol. Evol.* 26, 389–397. doi: 10.1016/j.tree.2011.04.005

Sexton, J. P., McIntyre, P. J., Angert, A. L., and Rice, K. J. (2009). Evolution and ccology of species range limits. *Annu. Rev. Ecol. Evol. Syst.* 40, 415–436. doi: 10.1146/annurev.ecolsys.110308.120317

Simpson, G. G. (1962). *Principles of Animal Taxonomy*. New York, NY: Columbia University Press.

Smith, N. G. C. (2003). Are radical and conservative substitution rates useful statistics in molecular evolution? *J. Mol. Evol.* 57, 467–478. doi: 10.1007/s00239-003-2500-z

Smith, S. A., and Beaulieu, J. M. (2009). Life history influences rates of climatic niche evolution in flowering plants. *Proc. R. Soc. B* 276, 4345–4352. doi: 10.1098/rspb.2009.1176

Sniegowski, P. D., Gerrish, P. J., Johnson, T., and Shaver, A. (2000). The evolution of mutation rates: separating causes from consequences. *Bioessays* 22, 1057–1066. doi: 10.1002/1521-1878(200012)22:12<1057::AID-BIES3<3.0.CO;2-W

Sniegowski, P. D., Gerrish, P. J., and Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387, 703–705. doi: 10.1038/42701

Stadler, T., Florez-Rueda, A. M., and Paris, M. (2012). Testing for "snowballing" hybrid incompatibilities in *Solanum*: impact of ancestral polymorphism and divergence estimates. *Mol. Biol. Evol.* 29, 31–34. doi: 10.1093/molbev/msr218

Stockwell, C. A., Hendry, A. P., and Kinnison, M. T. (2003). Contemporary evolution meets conservation biology. *Trends Ecol. Evol.* 18, 94–101. doi: 10.1016/S0169-5347(02)00044-7

Stuart-Fox, D., and Owens, I. P. F. (2003). Species richness in agamid lizards: chance, body size, sexual selection or ecology? *J. Evol. Biol.* 16, 659–669. doi: 10.1046/j.1420-9101.2003.00573.x

Sundin, G. W., and Weigand, M. R. (2007). The microbiology of mutability. *FEMS Microbiol. Lett.* 277, 11–20. doi: 10.1111/j.1574-6968.2007.00901.x

Sung, W., Tucker, A. E., Doak, T. G., Choi, E., Thomas, W. K., and Lynch, M. (2012). Extraordinary genome stability in the ciliate Paramecium tetraurelia. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19339–19344. doi: 10.1073/pnas.1210663109

Svetec, N., Cridland, J. M., Zhao, L., and Begun, D. J. (2016). The adaptive significance of natural genetic variation in the DNA damage response of *Drosophila melanogaster*. *PLoS Genet.* 12:e1005869. doi: 10.1371/journal.pgen.1005869

Taddei, F. (1997). Role of mutator alleles in adaptive evolution. *Nature* 387, 700–702. doi: 10.1038/42696

Thomas, J. A., Welch, J. J., Lanfear, R., and Bromham, L. (2010). A generation time effect on the rate of molecular evolution in invertebrates. *Mol. Biol. Evol.* 27, 1173–1180. doi: 10.1093/molbev/msq009

Todesco, M., Pascual, M. A., Owens, G. L., Ostevik, K. L., Moyers, B. T., Hübner, S., et al. (2016). Hybridization and extinction. *Evol. Appl.* 9, 892–908. doi: 10.1111/eva.12367

Torres-Barceló, C., Cabot, G., Oliver, A., Buckling, A., and MacLean, R. C. (2013). A trade-off between oxidative stress resistance and DNA repair plays a role in the evolution of elevated mutation rates in bacteria. *Proc. R. Soc. B* 280:20130007. doi: 10.1098/rspb.2013.0007

Turelli, M. (1984). Heritable genetic variation via mutation selection balance: Lerch's Zeta meets the abdominal bristle. *Theor. Popul. Biol.* 25, 138–193. doi: 10.1016/0040-5809(84)90017-0

Twyford, A. D., and Friedman, J. (2015). Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. *Evolution* 69, 1476–1486. doi: 10.1111/evo.12663

Venn, O., Turner, I., Mathieson, I., de Groot, N., Bontrop, R., and McVean, G. (2014). Strong male bias drives germline mutation in chimpanzees. *Science* 344, 1272–1275. doi: 10.1126/science.344.6189.1272

Via, S. (2009). Natural selection in action during speciation. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9939–9946. doi: 10.1073/pnas.0901397106

Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philos. Trans. R. Soc. B* 367, 451–460. doi: 10.1098/rstb.2011.0260

Via, S., Bouck, A. C., and Skillman, S. (2000). Reproductive isolation between divergent races of pea aphids on two hosts. II. Selection against migrants and hybrids in the parental environments. *Evolution* 54, 1626–1637. doi: 10.1111/j.0014-3820.2000.tb00707.x

Vigouroux, Y., Jaqueth, J. S., Matsuoka, Y., Smith, O. S., Beavis, W. F., Smith, J. S. C., et al. (2002). Rate and pattern of mutation at microsatellite loci in maize. *Mol. Biol. Evol.* 19, 1251–1260. doi: 10.1093/oxfordjournals.molbev.a004186

Wang, R. J., Ane, C., and Payseur, B. A. (2013). The evolution of hybrid incompatibilities along a phylogeny. *Evolution* 67, 2905–2922. doi: 10.1111/evo.12173

Wang, R. J., White, M. A., and Payseur, B. A. (2015). The pace of hybrid incompatibility evolution in house mice. *Genetics* 201, 229–242. doi: 10.1534/genetics.115.179499

Wang, Z., Yonezawa, T., Liu, B., Ma, T., Shen, X., Su, J., et al. (2011). Domestication relaxed selective constraints on the yak mitochondrial genome. *Mol. Biol. Evol.* 28, 1553–1556. doi: 10.1093/molbev/msq336

Waxman, D. (2011). A unified treatment of the probability of fixation when population size and the strength of selection change over time. *Genetics* 188, 907–913. doi: 10.1534/genetics.111.129288

Weber, C. C., Nabholz, B., Romiguier, J., and Ellegren, H. (2014). Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. *Genome Biol.* 15, 542. doi: 10.1186/s13059-014-0542-8

Welch, J. J. (2004). Accumulating Dobzhansky-Muller incompatibilities: reconciling theory and data. *Evolution* 58, 1145–1156. doi: 10.1554/03-502

Welch, J. J., Bininda-Emonds, O. R. P., and Bromham, L. (2008). Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol. Biol.* 8:53. doi: 10.1186/1471-2148-8-53

Wiens, J. J. (2004). What is speciation and how should we study it? *Am. Nat.* 163, 914–923. doi: 10.1086/386552

Wilson Sayres, M. A., and Makova, K. D. (2011). Genome analyses substantiate male mutation bias in many species. *Bioessays* 33, 938–945. doi: 10.1002/bies.201100091

Wilson Sayres, M. A., Venditti, C., Pagel, M., and Makova, K. D. (2011). Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution* 65, 2800–2815. doi: 10.1111/j.1558-5646.2011.01337.x

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13875–13879. doi: 10.1073/pnas.0811575106

Woolfit, M., and Bromham, L. (2005). Population size and molecular evolution on islands. *Proc. R. Soc. B* 272, 2277–2282. doi: 10.1098/rspb.2005.3217

Wright, S. D., Gillman, L. N., Ross, H. A., and Keeling, D. J. (2009). Slower tempo of microevolution in island birds: implications for conservation biology. *Evolution* 63, 2275–2287. doi: 10.1111/j.1558-5646.2009.00717.x

Yamamoto, E., Takashi, T., Morinaka, Y., Lin, S., Wu, J., Matsumoto, T., et al. (2010). Gain of deleterious function causes an autoimmune response and Bateson–Dobzhansky–Muller incompatibility in rice. *Mol. Genet. Genomics* 283, 305–315. doi: 10.1007/s00438-010-0514-y

Yang, J. N., Seluanov, A., and Gorbunova, V. (2013). Mitochondrial inverted repeats strongly correlate with lifespan: mtDNA inversions and aging. *PLoS ONE* 8:e73318. doi: 10.1371/journal.pone.0073318

Zhang, J. Z. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.* 50, 56–68. doi: 10.1007/s002399910007

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

PERSPECTIVE
published: 15 February 2017
doi: 10.3389/fgene.2017.00016

Analytical Biases Associated with GC-Content in Molecular Evolution

Jonathan Romiguier and Camille Roux*

Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

Molecular evolution is being revolutionized by high-throughput sequencing allowing an increased amount of genome-wide data available for multiple species. While base composition summarized by GC-content is one of the first metrics measured in genomes, its genomic distribution is a frequently neglected feature in downstream analyses based on DNA sequence comparisons. Here, we show how base composition heterogeneity among loci and taxa can bias common molecular evolution analyses such as phylogenetic tree reconstruction, detection of natural selection and estimation of codon usage. We then discuss the biological, technical and methodological causes of these GC-associated biases and suggest approaches to overcome them.

Keywords: GC-content, positive selection, biased gene conversion, codon usage bias, phylogeny, methodological biases

OPEN ACCESS

Edited by:
Scott V. Edwards,
Harvard University, USA

Reviewed by:
Jacob A. Tennessen,
Oregon State University, USA
Chris Organ,
Montana State University, USA

****Correspondence:***
Jonathan Romiguier
jonathan.romiguier@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *01 December 2016*
Accepted: *06 February 2017*
Published: *15 February 2017*

Citation:
Romiguier J and Roux C (2017) Analytical Biases Associated with GC-Content in Molecular Evolution. Front. Genet. 8:16. doi: 10.3389/fgene.2017.00016

INTRODUCTION

GC-content is shaped by a complex balance among mutation, selection, recombination, and genetic drift (Bulmer, 1991; Eyre-Walker and Hurst, 2001; Duret et al., 2002). As a consequence of variation in this subtle balance, it has been observed that GC-content varies considerably at two levels: (i) among genomes from different species and (ii) along chromosomes of a single species (Bernardi et al., 1985). Among species, the average genomic GC-content ranges from 13 to 75% (Pagani et al., 2011). Within the same genome, large chromosomal regions can also greatly differ in their nucleotide composition as first described in humans (Bernardi et al., 1985). For instance, GC-content is distributed across the human genome over successive long stretches of >100 kb that can be either GC-rich (with a GC-content ∼60%) or GC-poor (with a GC-content ∼35%; International Human Genome Sequencing Consortium, 2001).

After several years of debate among neutral or selective hypotheses [reviewed in Duret and Galtier (2009)], it is now widely accepted that one of the major drivers of base composition heterogeneity is GC-biased gene conversion (gBGC), a repair bias that favors GC over AT alleles during meiotic recombination (Eyre-Walker, 1993; Galtier et al., 2001; Montoya-Burgos et al., 2003; Duret and Arndt, 2008; Kent et al., 2012; Arbeithuber et al., 2015; Mugal et al., 2015). As a result of this link between GC and recombination, local GC-content increases faster in genomic hotspots of recombination (Spencer, 2006) while genome-wide GC-content increases faster in species with higher recombination rates per time unit (Romiguier et al., 2010, 2013b; Figuet et al., 2014; Weber et al., 2014). By conferring a higher transmission probability of GC alleles over AT in heterozygotes, gBGC mimics natural selection but is frequently overlooked in molecular evolution studies. Here, we revisit how much intra-genomic and inter-specific variations in base composition have a strong power to bias popular analyses in molecular evolution such as phylogenetic tree reconstruction, detection of natural selection and estimation of codon usage bias (**Figure 1**).

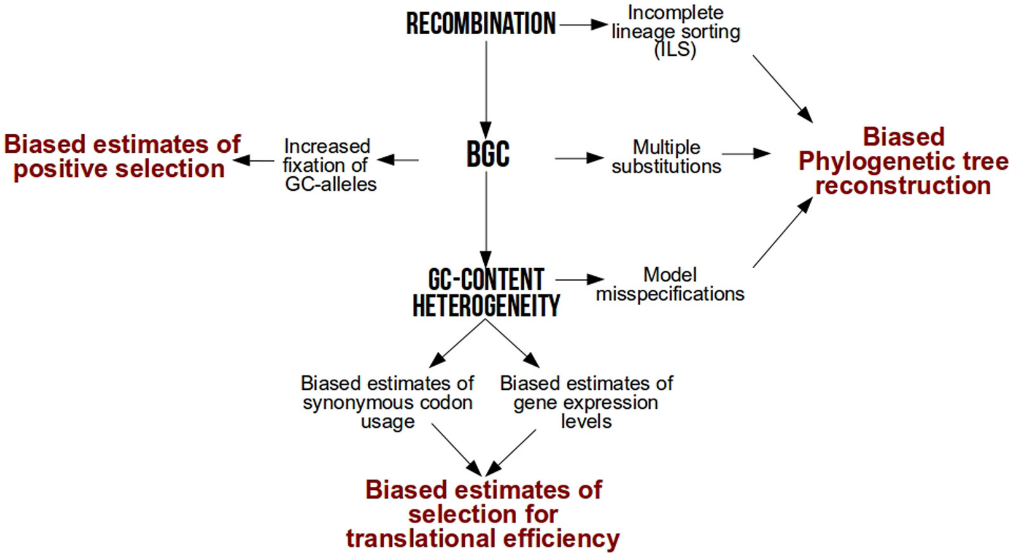

RECOMBINATION
Incomplete lineage sorting (ILS)
Biased estimates of positive selection
Increased fixation of GC-alleles
BGC
Multiple substitutions
Biased Phylogenetic tree reconstruction
GC-CONTENT HETEROGENEITY
Model misspecifications
Biased estimates of synonymous codon usage
Biased estimates of gene expression levels
Biased estimates of selection for translational efficiency

FIGURE 1 | Methodological biases associated to recombination, GC-biased gene conversion (BGC) and GC-content heterogeneity.

PHYLOGENETIC TREE RECONSTRUCTION

Reconstructions of phylogenetic trees from molecular datasets are central in evolutionary biology. While initially limited to a handful of loci with limited power to resolve difficult phylogenetic relationships, phylogenetic tree reconstruction is no longer restricted by the number of genetic markers. However, some phylogenetic relationships of the Tree of Life remain unresolved (Philippe et al., 2011). This difficulty stems from the mosaic nature of genomes gathering alternative and conflicting gene trees (Degnan and Rosenberg, 2009), where some but not all loci support the true species genealogy. Determining which loci are reliable phylogenetic markers is thus one of the biggest challenges in phylogenomics. While mixed historical signals along genomes are likely to have different natures, we will here focus to issues related to base composition.

A recent phylogenomic study reported that base composition is a relevant criterion to select markers carrying unambiguous phylogenetic signal: gene GC-content (average GC% of the sequences of an alignment) and GC-heterogeneity (variance of GC% among sequences of an alignment) were proved to bias species tree reconstructions (Romiguier et al., 2013a). As illustrated with mammalian genomes, phylogenetic trees of genes located in GC-rich regions produce five times more contradicting topologies than GC-poor genes, leading to important reconstruction biases and a poor resolution for both accepted and controversial nodes. This negative analytical effects of GC-content on tree reconstructions is widespread across the tree of life, as reported in basal eukaryote lineages (Rodríguez-Ezpeleta et al., 2007), yeasts (Collins et al., 2005), beetles (Sheffield et al., 2009), bees (Romiguier et al., 2016), hexapods (Delsuc, 2003), fishes (Li and Ortí, 2007; Betancur-R et al., 2013), birds (Nabholz et al., 2011), and bats (Teeling et al., 2000). Despite the accumulating empirical evidence demonstrating the pervasiveness of base composition issues in phylogeny, the reasons underlying such strong biases are unexplored. Here, we suggest three non-mutually exclusive hypotheses to explain this negative GC-effect in phylogenomics studies.

First, some aspects of the GC-bias are likely to be due to model misspecifications. Probabilistic methods for phylogenetic reconstruction (maximum likelihood or Bayesian inference) are indeed generally based on models of sequence evolution that assume a homogeneous base composition along the tree. However, this assumption is often violated (Phillips et al., 2004). Indeed, average GC-content of an alignment correlates strongly with GC-heterogeneity among sequences as a result of variation in the dynamic of gBGC among sampled species (Romiguier et al., 2013a). Such departures from the assumption of base composition homogeneity can lead to severe biases by incorrectly grouping distantly related taxa that converge in extreme nucleotide composition on a given locus (Phillips et al., 2004). This type of issues can be, however, easily solved by model-based solutions (see last paragraph of this section for more details).

The second hypothesis proposes that incomplete lineage sorting (ILS) is more important in GC-rich than GC-poor regions. ILS is known to produce conflicts among gene trees and the species tree because of the retention of ancestral polymorphisms (Degnan and Rosenberg, 2009). At the scale of the whole genome, the amount of incompletely sorted genes increases when the time of divergence is small relative to the average effective population sizes (Clark, 1997). Genomic variation in ILS was also empirically reported to be associated to GC-content in hominid genomes (Hobolth et al., 2011). This indirect (Charlesworth et al., 1993) relationship between GC-content and ILS can be explained by the dual effects of local recombination rates on base composition and linkage disequilibrium. High local recombination rates increase GC-content through gBGC, but also decrease the effect of genetic interferences, i.e., background selection (Charlesworth et al., 1993) and hitchhiking (Smith and Haigh, 1974). By being less affected by linked selective processes, GC-rich regions are thus expected to have relatively higher effective population sizes than GC-poor regions, leading to an extended retention time of ancestral polymorphism.

The third hypothesis is that gBGC is associated to saturation in multiple substitutions. Following the rapid birth and death of local recombination hotspots, gBGC is expected to occur in short, intense episodes (Duret and Galtier, 2009) where deleterious GC substitutions are likely to occur (Necşulea et al., 2011). Following a gBGC episode, natural selection is likely to revert such deleterious substitutions through AT replacement (Galtier et al., 2009). This toggling between GC deleterious and AT compensatory substitutions at the same nucleotide site is expected to lead to homoplasy, a direct consequence of multiple substitutions causing spurious similarity not due to common ancestry (Philippe et al., 2011). This type of AT/GC toggling is expected to be particularly fast and difficult to track because of the short-life of gBGC episodes that depends on the self-destructive nature of recombination hotspots (Coop and Myers, 2007). Even at very short evolutionary scales such as the Denisovan/Modern human divergence (0.4–0.8 Myrs), local recombination hotspots are not conserved (Lesecque et al., 2014), which could imply a complete loss of phylogenetic signals due to multiple turnovers between gBGC and natural selection at larger evolutionary scales. Although genomically small (1–2 kb), these short-lived recombination hotspots tend to arise and disappear in the same genomic regions of 1–2 Mb (Duret and Galtier, 2009) exhibit homoplasy issues. Common in fast-evolving sequences, homoplasy is also at the origin of the so-called and undesired "long branch attraction artifact" (Felsenstein, 1978). Reinforcing the idea that GC-rich genes might be affected by such biases, GC-rich and GC-heterogeneous genes have fast rates of evolution (Romiguier et al., 2013a, 2016). These abnormally fast-evolving genes are then likely to cause long-branch attraction artifacts, but also more general issues related to heterotachy-driven biases (Philippe et al., 2005). Even if long-branch attraction is generally considered as a minor problem in likelihood-based phylogenetics compared to parsimony, maximum likelihood methods using GC-rich genes can have a biased support toward topologies grouping long branches together (Romiguier et al., 2013a, 2016).

One solution to cope with base composition issues is the use of models of sequence evolution that takes into account heterogeneity in GC-content (Galtier and Gouy, 1998; Foster, 2004; Blanquart and Lartillot, 2006; Boussau and Gouy, 2006; Gowri-Shankar and Rattray, 2007; Dutheil and Boussau, 2008). However, these so-called non-homogeneous models are computationally costly. Albeit useful to alleviate GC-heterogeneity issues in phylogeny, empirical studies illustrate their limits to retrieve high bootstrap supports in the most GC-heterogeneous sequences (Betancur-R et al., 2013; Romiguier et al., 2016), shedding light on other GC-dependent biases in phylogeny such as ILS and gBGC-driven homoplasy. To date, the best practice recommended to discard noisy signals in sequences is the use of non-homogeneous models and/or the use of GC-poor phylogenetic markers. In this regard, it is noteworthy that coding sequences tend to be clustered in recombination hotspots and GC-rich regions (Duret and Galtier, 2009). Consequently, the use of the rare phylogenetic markers located in AT-rich regions is recommended. This is the case of ultra-conserved non-coding elements (UCE) that have the advantage to be AT-rich and evolve particularly slowly (McCormack et al., 2012). Compared to these non-coding AT-rich markers, clusters of AT-rich coding genes in low-recombining regions could undergo a higher rate of background selection, decreasing the effective population size and then, the amount of ILS. It is noteworthy that UCE and AT-rich genes both support the same topology for the controversial rooting of placental mammals (McCormack et al., 2012; Romiguier et al., 2013a), highlighting relevance of these markers to overcome GC-biases. Other strategies might involve to compare these markers with markers that cannot be affected by recombination and gBGC, such as mitochondrial genes. Further methodological improvements could come from coalescent-based supertree methods (Liu et al., 2009) that account for ILS. By weighting the confidence in each gene tree according to the GC-content of an alignment, they may allow the integration of most of the available information and alleviate the spurious signal inherent to GC-rich markers. To date, methods computing the exact likelihood of alternative topologies are restricted to relatively simple models neglecting direct and indirect effects of background selection, selective sweep, gBGC and ILS on phylogenetic reconstruction. But these processes are now implemented in recent simulators (Haller and Messer, 2017), allowing them to be treated as nuisance parameters during computational evolution of sequences. Although such highly complex models are currently intractable by maximum likelihood approaches, the possibility to simulate them within an approximate Bayesian computation (ABC) framework (Beaumont et al., 2002; Csilléry et al., 2010a,b; Pudlo et al., 2016) could bring new methodological perspectives in phylogenetic reconstruction. ABC has been proved to be a powerful framework to compare complex evolutionary scenarios for large datasets (Roux et al., 2016), illustrating the recent improvements made in flexible machine learning algorithms. Applied to phylogenetic reconstructions, efficient computational tools like SLiM 2 are already available to simulate models with gBGC episodes and multiple substitutions along a branch as well as statistical packages to compute the probabilities of alternative scenarios (Csilléry et al., 2012; Pudlo et al., 2016). Altogether, current available softwares already provide stimulating leads for future developments in phylogeny.

DETECTION OF POSITIVE SELECTION

Identifying candidate loci for natural selection is a central goal explored by two traditional approaches in adaptation-genomics: top-down (GWA and QTL) and bottom-up (genomic scan) approaches. With the advent of high-throughput sequencing, genomic scans became a popular approach to detect candidate target of selection. Such scans have the merit to identify candidates without the *a priori* expectation of a candidate gene approach (Ellegren, 2014). However, they have various limitations with false-positive issues (Mallick et al., 2009; Bierne et al., 2011), narrow signatures of balancing selection (Roux et al., 2012), and over-interpretation of outlier loci (Pavlidis et al., 2012). Here, we detail how GC-content can lead to important additional bias during genome scans for detecting natural selection.

Genome scans of positive selection often rely on methods that look for lineage-specific accelerations in the protein rate of evolution. Such accelerations are classically measured through *dN/dS*, which calculates the excess of amino-acid substitutions (*dN*: non-synonymous mutation rate per site) relative to *dS*, the substitution rate per site used as a proxy of the neutral clock. This *dN/dS* ratio is generally smaller than 1, reflecting the pervasiveness of purifying selection that eliminates non-synonymous mutations to preserve the protein structure. Conversely, a *dN/dS* ratio greater than 1 is considered as a signature of positive selection that favors the fixation of beneficial non-synonymous mutations. From a population genetics point of view, gBGC mimics positive selection by favoring the fixation of AT- > GC mutations, regardless of their beneficial or deleterious status (Nagylaki, 1983). Because GC alleles are actively selected by the repair systems of meiotic recombination, they are over-represented in the gamete pool and benefit of increased transmission to the next generation in a similar way than beneficial mutations subject to positive selection. Consequently, many accelerations of the substitution rate attributed to positive selection during genome scans are actually due to gBGC episodes (Galtier and Duret, 2007; Berglund et al., 2009; Galtier et al., 2009; Ratnakumar et al., 2010; Kostka et al., 2012). When a mutation toward GC is deleterious, gBGC can counteract positive selection and maintain or fix deleterious alleles. High fixation rates of non-synonymous mutations at a locus should thus not be systematically interpreted as being beneficial for the fitness of the individual, particularly when considering that gBGC has been proved to be able to maintain deleterious mutations associated to human diseases (Necşulea et al., 2011; Capra et al., 2013; Lachance and Tishkoff, 2014).

Confusion between positive selection and gBGC could be avoided through two different ways. The first is by filtering the results of classical tests of positive selection and consider with caution positive selection signatures in GC-rich regions. This is particularly true for selection tests that rely more on overall evolutionary rate rather than *dN/dS* (Pollard et al., 2006; Kostka et al., 2012). Even if gBGC can increase *dN/dS* in some conditions (Galtier et al., 2009; Bolívar et al., 2015), AT- > GC mutations are more likely to happen in synonymous sites, which limits the effect of gBGC on *dN/dS* compared to the evolutionary rate. Several criterions can be used in both cases to differentiate gBGC from positive selection, such as the number of mutations toward GC in the surrounding non-coding regions (Galtier and Duret, 2007). The second would be to develop methods that restrict *dN/dS* estimations to GC-conservative substitutions in the context of codon-models aimed to detect positive selection events (Yang and Nielsen, 2002; Lartillot, 2013).

CODON USAGE BIAS

Popular analytical methods in molecular evolution rely on a strong assumption: synonymous mutations are neutral. GC-content at synonymous positions is frequently claimed to be exposed only to the mutation/drift equilibrium. However, natural selection was proposed to be superimposed to these two evolutionary forces at synonymous codons (Urrutia, 2003; Comeron, 2004; Plotkin et al., 2004). Although initially challenged (Williamson et al., 2005), natural selection acting on standing synonymous variation was found to be associated to gene expression level, the most expressed genes using a set of preferred codons (Comeron, 2004). This association is explained by selection for increased translational efficiency. The analysis of >1,000 genes in *Drosophila* demonstrated that the most used synonymous codons corresponded to the most available tRNAs in the genome (Moriyama and Powell, 1997). Translational efficiency would then be optimized by increasing the usage of the preferred synonymous codons. Such a process can be tested in coding sequences by measuring the effective number of codons (ENc) in a given gene. ENc takes a value of 61 when all codons of the genetic code (minus the three stop codons) are used without bias, and decreases to 20 (the number of amino-acids) for the most biased genes. In agreement with the hypothesis of selection for translational efficiency, population genetics analyses in *Drosophila* described signatures of selection on synonymous mutations (Akashi, 1995; Akashi and Schaeffer, 1997).

A study of codon usage bias in *Caenorhabditis elegans*, *Drosophila melanogaster,* and *Arabidopsis thaliana* has shed light on the over-expression of genes featuring codon preference, with a large predominance of preferred codons ending with G or C (Duret and Mouchiroud, 1999). However, the GC-content at third coding positions (GC3) is also correlated to the GC-content of the surrounding non-coding regions (Kliman and Hey, 1994; Akashi et al., 1998), which suggests the action of gBGC shaping local base compositions. By locally increasing GC-content, gBGC mechanically restricts the number of used codons and reduces the measured ENc independently of selection for translational efficiency. The measured ENc is thus biased by gBGC and must be corrected with local background nucleotide compositions. In addition, variation in GC-content also impacts measures of gene expression. With the advent of high-throughput sequencing technologies, it is now a standard practice to approximate gene expression levels by counting the number of reads mapping a target in ChIP-seq or RNA-seq analysis. However, sequencing biases artificially over-represent genomic regions with intermediate levels of GC-content (50%), which in turn bias the estimates of gene expression levels (Chouvarine et al., 2016). Testing selection for translational efficiency by measuring the correlation between ENc and gene expression levels therefore requires the use of both GC-corrected ENc and GC-corrected expression levels. ENc estimates can be corrected by GC-content of neighborhood regions (Novembre, 2002), while GC-corrected expression levels can be obtained by applying local LOESS regression (Miller et al., 2011; Benjamini and Speed, 2012; Chandrananda et al., 2014) or quantile normalization-methods (Risso et al., 2011), i.e., by normalizing the raw number of mapped reads by the local GC-content.

The ongoing surge of transcriptomic data will permit measurement of GC-content heterogeneity, preferred codons usage and expression levels across a large number of loci and species. This type of large-scale analysis could open the door to a better understanding of the relationship linking effective population sizes (*Ne*) and codon usage. As theoretically predicted (Bulmer, 1991), selection on synonymous codons might be stronger in species with large *Ne*. While the *Ne*-hypothesis to explain variation in selection on codon usage remains untested by empirical studies, a descriptive study of the *Ne*-effect on variation in gBGC will be necessary to avoid entangling the two effects. Future projects aiming to test these hypotheses are expected to be strongly biased if GC-content biases are naively neglected regarding estimates of gene expression levels or codon usage.

CONCLUSION

GC-content is associated to multiple biases of different nature (**Figure 1**). Whether through technological reasons (sequencing technologies biases), biological reasons (GC-biased gene conversion) or methodological reasons (models of sequence evolution limitations), all these biases affect the results of downstream analyses. With the surge of genomic data from various non-model species, comparative genomics have the opportunity to solve many unresolved questions in evolution. However, one should be aware of the methodological challenges associated to the GC-content heterogeneity inherent to large scale studies, whether it be for a large number of species or loci.

AUTHOR CONTRIBUTIONS

JR had the idea of the project. JR and CR wrote the article.

FUNDING

This work was supported by a Federation of European Biochemical Societies (FEBS) long-term fellowship to JR.

REFERENCES

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* 139, 1067–1076.

Akashi, H., Kliman, R. M., and Eyre-Walker, A. (1998). Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. *Genetica* 10, 49–60. doi: 10.1023/A:1017078607465

Akashi, H., and Schaeffer, S. W. (1997). Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. *Genetics* 146, 295–307.

Arbeithuber, B., Betancourt, A. J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2109–2114. doi: 10.1073/pnas.1416622112

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162, 2025–2035.

Benjamini, Y., and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40:e72. doi: 10.1093/nar/gks001

Berglund, J., Pollard, K. S., and Webster, M. T. (2009). Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol.* 7:e26. doi: 10.1371/journal.pbio.1000026

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., et al. (1985). The mosaic genome of warm-blooded vertebrates. *Science* 228, 953–958. doi: 10.1126/science.4001930

Betancur-R, R., Li, C., Munroe, T. A., Ballesteros, J. A., and Ortí, G. (2013). Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). *Syst. Biol.* 62, 763–785. doi: 10.1093/sysbio/syt039

Bierne, N., Welch, J., Loire, E., Bonhomme, F., and David, P. (2011). The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol. Ecol.* 20, 2044–2072. doi: 10.1111/j.1365-294X.2011.05080.x

Blanquart, S., and Lartillot, N. (2006). A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* 23, 2058–2071. doi: 10.1093/molbev/msl091

Bolívar, P., Mugal, C. F., Nater, A., and Ellegren, H. (2015). Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill–Robertson interference, in an avian system. *Mol. Biol. Evol.* 33, 216–227. doi: 10.1093/molbev/msv214

Boussau, B., and Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* 55, 756–768. doi: 10.1080/10635150600975218

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129, 897–907.

Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S., and Siepel, A. (2013). A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet.* 9:e1003684. doi: 10.1371/journal.pgen.1003684

Chandrananda, D., Thorne, N. P., Ganesamoorthy, D., Bruno, D. L., Benjamini, Y., Speed, T. P., et al. (2014). Investigating and correcting plasma DNA sequencing coverage bias to enhance aneuploidy discovery. *PLoS ONE* 9:e86993. doi: 10.1371/journal.pone.0086993

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303.

Chouvarine, P., Wiehlmann, L., Losada, P. M., DeLuca, D. S., and Tümmler, B. (2016). Filtration and normalization of sequencing read data in whole-metagenome shotgun samples. *PLoS ONE* 11:e0165015. doi: 10.1371/journal.pone.0165015

Clark, A. G. (1997). Neutral behavior of shared polymorphism. *Proc. Natl. Acad. Sci. U.S.A.* 94, 7730–7734. doi: 10.1073/pnas.94.15.7730

Collins, T. M., Fedrigo, O., and Naylor, G. J. P. (2005). Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. *Syst. Biol.* 54, 493–500. doi: 10.1080/10635150590947339

Comeron, J. M. (2004). Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* 167, 1293–1304. doi: 10.1534/genetics.104.026351

Coop, G., and Myers, S. R. (2007). Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genet.* 3:e35. doi: 10.1371/journal.pgen.0030035

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010a). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.* 25, 410–418. doi: 10.1016/j.tree.2010.04.001

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010b). Invalid arguments against ABC: reply to A.R. Templeton. *Trends Ecol. Evol.* 25, 490–491. doi: 10.1016/j.tree.2010.06.011

Csilléry, K., François, O., and Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* 3, 475–479. doi: 10.1111/j.2041-210X.2011.00179.x

Degnan, J. H., and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340. doi: 10.1016/j.tree.2009.01.009

Delsuc, F. (2003). Comment on "Hexapod origins: monophyletic or paraphyletic?" *Science* 301, 1490–1491. doi: 10.1126/science.1086558

Duret, L., and Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4:e1000071. doi: 10.1371/journal.pgen.1000071

Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* 10, 285–311. doi: 10.1146/annurev-genom-082908-150001

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 96, 4482–4487. doi: 10.1073/pnas.96.8.4482

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. (2002). Vanishing GC-rich isochores in mammalian genomes. *Genetics* 162, 1837–1847.

Dutheil, J., and Boussau, B. (2008). Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol. Biol.* 8:255. doi: 10.1186/1471-2148-8-255

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 29, 51–63. doi: 10.1016/j.tree.2013.09.008

Eyre-Walker, A. (1993). Recombination and mammalian genome evolution. *Proc. R. Soc. B Biol. Sci.* 252, 237–243. doi: 10.1098/rspb.1993.0071

Eyre-Walker, A., and Hurst, L. D. (2001). The evolution of isochores. *Nat. Rev. Genet.* 2, 549–555. doi: 10.1038/35080577

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410. doi: 10.2307/2412923

Figuet, E., Ballenghien, M., Romiguier, J., and Galtier, N. (2014). Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biol. Evol.* 7, 240–250. doi: 10.1093/gbe/evu277

Foster, P. (2004). Modeling compositional heterogeneity. *Syst. Biol.* 53, 485–495. doi: 10.1080/10635150490445779

Galtier, N., and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23, 273–277. doi: 10.1016/j.tig.2007.03.011

Galtier, N., Duret, L., Glémin, S., and Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25, 1–5. doi: 10.1016/j.tig.2008.10.011

Galtier, N., and Gouy, M. (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15, 871–879. doi: 10.1093/oxfordjournals.molbev.a025991

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159, 907–911. doi: 10.1038/35091126

Gowri-Shankar, V., and Rattray, M. (2007). A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. *Mol. Biol. Evol.* 24, 1286–1299. doi: 10.1093/molbev/msm046

Haller, B. C., and Messer, P. W. (2017). SLiM 2: flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* 34, 230–240. doi: 10.1093/molbev/msw211

Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H., and Mailund, T. (2011). Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* 21, 349–356. doi: 10.1101/gr.114751.110

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 412, 860–921. doi: 10.1038/35057062

Kent, C. F., Minaei, S., Harpur, B. A., and Zayed, A. (2012). Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18012–18017. doi: 10.1073/pnas.1208094109

Kliman, R. M., and Hey, J. (1994). The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* 137, 1049–1056.

Kostka, D., Hubisz, M. J., Siepel, A., and Pollard, K. S. (2012). The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol. Biol. Evol.* 29, 1047–1057. doi: 10.1093/molbev/msr279

Lachance, J., and Tishkoff, S. A. (2014). Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am. J. Hum. Genet.* 95, 408–420. doi: 10.1016/j.ajhg.2014.09.008

Lartillot, N. (2013). Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Mol. Biol. Evol.* 30, 356–368. doi: 10.1093/molbev/mss231

Lesecque, Y., Glémin, S., Lartillot, N., Mouchiroud, D., and Duret, L. (2014). The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. *PLoS Genet.* 10:e1004790. doi: 10.1371/journal.pgen.1004790

Li, C., and Ortí, G. (2007). Molecular phylogeny of Clupeiformes (Actinopterygii) inferred from nuclear and mitochondrial DNA sequences. *Mol. Phylogenet. Evol.* 44, 386–398. doi: 10.1016/j.ympev.2006.10.030

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58, 468–477. doi: 10.1093/sysbio/syp031

Mallick, S., Gnerre, S., Muller, P., and Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* 19, 922–933. doi: 10.1101/gr.086512.108

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.* 22, 746–754. doi: 10.1101/gr.125864.111

Miller, C. A., Hampton, O., Coarfa, C., and Milosavljevic, A. (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS ONE* 6:e16327. doi: 10.1371/journal.pone.0016327

Montoya-Burgos, J. I., Boursot, P., and Galtier, N. (2003). Recombination explains isochores in mammalian genomes. *Trends Genet.* 19, 128–130. doi: 10.1016/S0168-9525(03){\break}00021-0

Moriyama, E. N., and Powell, J. R. (1997). Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* 45, 514–523. doi: 10.1007/PL00006256

Mugal, C. F., Weber, C. C., and Ellegren, H. (2015). GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species. *Bioessays* 37, 1317–1326. doi: 10.1002/bies.201500058

Nabholz, B., Künstner, A., Wang, R., Jarvis, E. D., and Ellegren, H. (2011). Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol. Biol. Evol.* 28, 2197–2210. doi: 10.1093/molbev/msr047

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. U.S.A.* 80, 6278–6281. doi: 10.1073/pnas.80.20.6278

Necşulea, A., Popa, A., Cooper, D. N., Stenson, P. D., Mouchiroud, D., Gautier, C., et al. (2011). Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum. Mutat.* 32, 198–206. doi: 10.1002/humu.21407

Novembre, J. A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.* 19, 1390–1394. doi: 10.1093/oxfordjournals.molbev.a004201

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. A., Smirnova, T., Nosrat, B., et al. (2011). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 40, D571–D579. doi: 10.1093/nar/gkr1100

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. *Mol. Biol. Evol.* 29, 3237–3248. doi: 10.1093/molbev/mss136

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., et al. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9:e1000602. doi: 10.1371/journal.pbio.1000602

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* 5:50. doi: 10.1186/1471-2148-5-50

Phillips, M. J., Delsuc, F., and Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21, 1455–1458. doi: 10.1093/molbev/msh137

Plotkin, J. B., Robins, H., and Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12588–12591. doi: 10.1073/pnas.0404957101

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443, 167–172. doi: 10.1038/nature05113

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics* 32, 859–866. doi: 10.1093/bioinformatics/btv684

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., et al. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 2571–2580. doi: 10.1098/rstb.2010.0007

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 12:480. doi: 10.1186/1471-2105-12-480

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* 56, 389–399. doi: 10.1080/10635150701397643

Romiguier, J., Cameron, S. A., Woodard, S. H., Fischman, B. J., Keller, L., and Praz, C. J. (2016). Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. *Mol. Biol. Evol.* 33, 670–678. doi: 10.1093/molbev/msv258

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., and Douzery, E. J. P. (2013a). Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol.* 30, 2134–2144. doi: 10.1093/molbev/mst116

Romiguier, J., Ranwez, V., Douzery, E. J. P., and Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20, 1001–1009. doi: 10.1101/gr.104372.109

Romiguier, J., Ranwez, V., Douzery, E. J. P., and Galtier, N. (2013b). Genomic evidence for large, long-lived ancestors to placental mammals. *Mol. Biol. Evol.* 30, 5–13. doi: 10.1093/molbev/mss211

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., and Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biol.* 14:e2000234. doi: 10.1371/journal.pbio.2000234

Roux, C., Pauwels, M., Ruggiero, M.-V., Charlesworth, D., Castric, V., and Vekemans, X. (2012). Recent and ancient Signature of balancing selection around the S-Locus in *Arabidopsis halleri* and *A. lyrata*. *Mol. Biol. Evol.* 30, 435–447. doi: 10.1093/molbev/mss246

Sheffield, N. C., Song, H., Cameron, S. L., and Whiting, M. F. (2009). Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. *Syst. Biol.* 58, 381–394. doi: 10.1093/sysbio/syp037

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35. doi: 10.1017/S0016672300014634

Spencer, C. C. A. (2006). Human polymorphism around recombination hotspots. *Biochem. Soc. Trans.* 34, 535–536. doi: 10.1042/BST0340535

Teeling, E. C., Scally, M., Kao, D. J., Romagnoli, M. L., Springer, M. S., and Stanhope, M. J. (2000). Molecular evidence regarding the origin of echolocation and flight in bats. *Nature* 403, 188–192. doi: 10.1038/35003188

Urrutia, A. O. (2003). The signature of selection mediated by expression on human genes. *Genome Res.* 13, 2260–2264. doi: 10.1101/gr.641103

Weber, C. C., Boussau, B., Romiguier, J., Jarvis, E. D., and Ellegren, H. (2014). Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* 15:549. doi: 10.1186/s13059-014-0549-1

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7882–7887. doi: 10.1073/pnas.0502300102

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917. doi: 10.1093/oxfordjournals.molbev.a004148

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

MINI REVIEW
published: 30 May 2017
doi: 10.3389/fgene.2017.00072

The Evolution of Bacterial Genome Architecture

Louis-Marie Bobay and Howard Ochman*

Department of Integrative Biology, University of Texas, Austin, TX, United States

The genome architecture of bacteria and eukaryotes evolves in opposite directions when subject to genetic drift, a difference that can be ascribed to the fact that bacteria exhibit a mutational bias that deletes superfluous sequences, whereas eukaryotes are biased toward large insertions. Expansion of eukaryotic genomes occurs through the addition of non-functional sequences, such as repetitive sequences and transposable elements, whereas variation in bacterial genome size is largely due to the acquisition and loss of functional accessory genes. These properties create the situation in which eukaryotes with very similar numbers of genes can have vastly different genome sizes, while in bacteria, gene number scales linearly with genome size. Some bacterial genomes, however, particularly those of species that undergo bottlenecks due to recent association with hosts, accumulate pseudogenes and mobile elements, conferring them a low gene content relative to their genome size. These non-functional sequences are gradually eroded and eliminated after long-term association with hosts, with the result that obligate symbionts have the smallest genomes of any cellular organism. The architecture of bacterial genomes is shaped by complex and diverse processes, but for most bacterial species, genome size is governed by a non-adaptive process, i.e., genetic drift coupled with a mutational bias toward deletions. Thus, bacteria with small effective population sizes typically have the smallest genomes. Some marine bacteria counter this near-universal trend: despite having immense population sizes, selection, not drift, acts to reduce genome size in response to metabolic constraints in their nutrient-limited environment.

Keywords: genetic drift, genome, bacterial, genome evolution, horizontal gene transfer, population dynamics

OPEN ACCESS

Edited by:
Scott V. Edwards,
Harvard University, United States

Reviewed by:
Isabel Gordo,
Instituto Gulbenkian de Ciência, Portugal
Matthew B. Hamilton,
Georgetown University, United States

****Correspondence:***
Louis-Marie Bobay
lbobay@utexas.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *03 January 2017*
Accepted: *12 May 2017*
Published: *30 May 2017*

Citation:
Bobay L-M and Ochman H (2017) The Evolution of Bacterial Genome Architecture. Front. Genet. 8:72. doi: 10.3389/fgene.2017.00072

INTRODUCTION

The overall structure and organization of bacterial genomes were well resolved before the golden era of genome sequencing. It was known that bacterial genomes varied in size by at least an order of magnitude and even that there could be considerable variation in genome size within a bacterial species (Herdman, 1985); that bacterial genomes typically comprised one circular chromosome but often harbored extrachromosomal elements in the form of plasmids or phages (Lederberg, 1998); that base composition was relatively uniform along the chromosome but highly variable across species, ranging from 13 to 75% G + C (Thomas et al., 2008; McCutcheon and Moran, 2010); that bacterial genomes consisted mostly of functional protein-coding regions, with little non-coding or intervening sequences (Mira et al., 2001); that genetic maps (and hence, gene order and gene content) remained fairly stable among related species (Rocha, 2008); that genome architecture could be altered by insertions, duplications, inversions, and translocations, fostered, in part, by mobile elements (Eisen et al., 2000; Tillier and Collins, 2000; Rocha, 2008); and that the bacterial chromosome is configured into domains that relate to its replication and packaging (Boccard et al., 2005).

Many of these features of bacterial genomes contrast those of eukaryotic genomes, which are often partitioned into multiple linear chromosomes and are generally much larger due both to increases in gene number and to the proliferation of non-coding and repetitive DNA. Although the properties of genomes and the variation in genome architecture across the tree of life were recognized by cytogeneticists and molecular biologists alike, it was not until large numbers of bacterial genome sequences became available that the processes underlying their evolution could be fully appreciated.

MICROEVOLUTIONARY PROCESSES DRIVE GENOME ARCHITECTURE

Like other biological features, the mechanisms forging the content and organization of bacterial genomes rely on selection and drift, whose relative contributions are dictated by the effective population size (N_e) and the selection coefficient (s) associated with a trait (Wright, 1931; Kimura, 1968). In a haploid organism, evolution is driven by stochastic processes (i.e., drift) when $|2 \times Ne \times s| << 1$, whereas selection dominates when $|2 \times Ne \times s| >> 1$ (Kimura, 1968). This central concept of population genetics was further expanded under the nearly neutral theory of evolution of Ohta (1973), who put forward the notion that, although selection does not change, the response to selection depends on the effective size of populations. Indeed, the selection coefficient s is a variable parameter that only depends on the impact of a given gene variant on the fitness of the individual relative to others in the population. In contrast, the long-term effective population size is a parameter that influences the impact of selection relative to drift, since smaller populations are more strongly affected by the random sampling of genotypes at each generation. Note that at the extremes, such that a trait is either essential (i.e., its disruption is lethal) or completely neutral (i.e., its disruption is inconsequential), effective population size does not affect its fixation; however, the fate of all other variants, and of the vast majority of sequences in a bacterial genome, depends on the interplay of selection and drift.

Considering these factors, it has been proposed that the architecture of genomes varies as a function of the effective population size (N_e) and the mutation rate (μ), under the so-called "mutational hazard hypothesis" (Lynch and Conery, 2003; Lynch et al., 2011). Those species with small effective population sizes, such as many animals and plants, will experience strong effects of drift-guided evolution and accumulate large amounts of moderately deleterious DNA, including mobile elements, pseudogenes, and introns (Lynch et al., 2011). In humans, whose effective population size is estimated to be lower than 10,000 (Takahata, 1993; Tenesa et al., 2007), sequences encoding functional proteins represent only <5% of genomic DNA, due to the genome-wide expansion of numerous genetic elements, such as introns, LINEs, and SINEs. Amassing these sequences is thought to represent a substantial mutational burden, since intron splice sites can represent potential targets for mutations and each new mobile element can potentially insert into and disrupt a functional region (Lynch, 2002; Lynch et al., 2011). In contrast, species with large effective population sizes evolve predominantly through selection, thereby preventing the accumulation of hazardous elements.

Relative to multicellular organisms, bacteria exhibit small, gene-rich genomes, typically under 10 Mb in length (Kuo et al., 2009). At first glance, these features seem to fit with the mutational hazard hypothesis, such that the large population sizes of bacteria increase the efficacy of selection, which fosters the removal of deleterious sequences and results in compact genomes consisting mostly of the functional genes (Lynch, 2006). However, the trend in bacteria actually runs *opposite* to the predictions of the mutational hazard hypothesis (Daubin and Moran, 2004; Kuo et al., 2009): bacterial species with the lowest effective population sizes, such as endosymbiotic bacteria whose effective population sizes approximate those of their animal hosts, typically have the smallest and most compact genomes, whereas those with the largest populations exhibit the expansive genomes (Kuo et al., 2009). This circumstance raises questions about why the genomic trends in bacteria differ from those of eukaryotes; and in this review, we resolve the population-level parameters as well as the mutational mechanisms that shape the structure, content, and evolution of bacterial genomes.

DEFINING BACTERIAL SPECIES AND POPULATIONS

Due to their unicellularity and uniformity in genome structure, bacteria are typically viewed as simple organisms. However, many of the most basic features of their populations remain obscure, often making it difficult to evaluate and quantify microevolutionary processes. The first issue surrounds the definition of a bacterial species (Shapiro et al., 2016). Sexual organisms are usually classified into species that represent units that are genetically and phenotypically cohesive, and the most widely applied species definition—the Biological Species Concept—allows for a simple and uniform classification of species across all sexual organisms (Mayr, 1942). The delineation of bacterial species is much more problematic, since no biologically relevant species concept is appropriate for asexual organisms that sporadically exchange or acquire genes by recombination or lateral gene transfer (Shapiro and Polz, 2014, 2015). Different conceptual frameworks, such as the ecotype definition, have been proposed (Cohan, 2001) but are difficult in practice to apply. In contrast, sequence-similarity thresholds are easy to apply but need not be biologically relevant (Konstantinidis and Tiedje, 2005; Hugenholtz et al., 2016; Bobay and Ochman, 2017). Estimation of several population genetic parameters relies on assessments of the allelic variation in conspecifics, so the arbitrary assignment of bacterial strains to species can (and has) lead to many contradictory conclusions about bacterial evolution.

Apart from delineation of species, the estimation of effective population sizes (N_e) is difficult in bacteria, both because they are difficult to observe and because they violate some of the assumptions of the Wright–Fisher model (Hartl and Clark, 2007). Aside from those few host-associated bacteria whose transmission dynamics are known, estimates of N_e for most bacterial species vary over several orders of magnitude depending on how and which populations are being assessed. Genomic-based strategies for estimating N_e are usually based on the extent of genomic diversity at neutral sites. N_e for haploid organisms is given by $\theta = 2 \times N_e \times \mu$ (Watterson, 1975), where θ is the number of segregating sites and μ is the mutation rate. The existence of truly neutral sites in bacteria has been called into question, since codon usage and nucleotide composition appear to be under weak selection in many species (Rocha and Feil, 2010). If this is the case, estimates based on such metrics should be considered prudently, especially in those species with large population sizes, since the effectiveness of selection at such sites would be enhanced as N_e becomes larger.

Estimating θ may be confounded by the fact that bacteria reproduce clonally, and the linkage of alleles makes them highly susceptible to Hill–Robertson effects (i.e., background selection, hitchhiking, and Muller's ratchet; Hill and Robertson, 1966; Felsenstein, 1974; Smith and Haigh, 1974; Charlesworth et al., 1993), such that selection on a beneficial or detrimental allele in a given genotype will lead to the loss of allelic diversity. Because deleterious mutations are expected to be frequent, it has been predicted that background selection leads to the loss of substantial genetic diversity in bacterial populations (Betancourt et al., 2009; Price and Arkin, 2015). It is important to note, however, that very few bacteria are truly clonal and that most engage in some homologous recombination (Vos and Didelot, 2009), which liberates alleles from genomic linkage and counteracts Hill–Robertson effects (Betancourt et al., 2009). Unlike recombination, whose rate is unpredictable for a given bacterial species, it is thought that μ is relatively constant across species. Mutation rates are fairly similar in most of the 10 or so bacterial species that have been assayed in the laboratory; however, they are still unknown for the vast majority of bacterial species and can vary up to 100-fold (Sung et al., 2016). Together, these factors make estimations of N_e based on the neutral expectations an imperfect metric.

A more convenient though indirect measure of N_e is based on assessment of K_a/K_s or d_N/d_S ratios, which represent the effectiveness of selection and scale negatively with N_e, since smaller populations promote the fixation of slightly deleterious mutations thereby increasing K_a (or d_N) (Daubin and Moran, 2004; Kryazhimskiy and Plotkin, 2008). Although d_N/d_S ratios are not constant over time when computed on genomes of the same species (Rocha et al., 2006; Kryazhimskiy and Plotkin, 2008) and can vary when genes are under different selective constraints (Batut et al., 2014), it provides a more robust metric for comparing N_e across species when adjusted for divergence times (e.g., by applying *dS* thresholds) and limited to comparisons of identical sets of genes in different species.

When analyzed across a diverse array of taxa, K_a/K_s ratios proved to be a fairly reliable proxy for N_e, since the values seemed to fit with what was known about the natural history of the specific bacterial groups. For example, endosymbiotic, parasitic, and other obligatory host-associated bacteria displayed high K_a/K_s ratios and are known to have effective population size that are small, approximating those of their animal hosts. In contrast, broadly distributed, environmental bacteria, presumed to have very large effective population sizes, displayed the lowest K_a/K_s ratios. It was also determined that K_a/K_s ratios scaled with genome size, such that bacteria with higher values (i.e., smaller N_e) have more highly reduced genomes, and this association holds across phylogenetically divergent bacteria (Kuo et al., 2009).

HOW LARGE ARE THE EFFECTIVE POPULATION SIZES OF BACTERIA?

Although the estimation of N_e is challenging, studies based on nucleotide diversity at neutral sites suggest that most bacterial species have an effective population size in the range of 10^6–10^9 (Sung et al., 2012). However, estimates based on d_N/d_S ratios—but including some additional species—yielded average estimates ranging from 10^6 to 10^{12} (Sela et al., 2016). It is surprising that the most abundant species on the planet, the marine bacterium *Prochlorococcus*, was estimated to have an N_e of only 1.5×10^9, since based on its census population, N_e could reach 10^{13} in this "species" (Kashtan et al., 2014). The N_e estimated from allelic diversity is likely an underestimation, as might occur if synonymous positions are not strictly neutral. But because the population dynamics of *Prochlorococcus* is largely unknown, it is possible that N_e is indeed much lower than the census population size due to frequent and drastic demographic variations, such as genotype sweeps and bottlenecks.

On the other end of the spectrum, endosymbionts experienced strong reductions in population sizes. Being confined within the cells of their hosts, and in the most extreme cases, transmitted by exclusively maternal lines, endosymbionts experience severe bottlenecks during propagation (Moran, 1996; Moran et al., 2009). In the aphid endosymbiont *Buchnera aphidicola*, N_e was estimated to be $\sim 10^6$ (Funk et al., 2001; Moran et al., 2009), but its mutation rate has not been directly estimated in the lab. The only small-genomed bacterium whose mutation rate has been accurately measured is the intracellular bacterium *Mesoplasma florum*, and its N_e was also estimated to be 10^6 (Sung et al., 2012), again among the lowest determined for bacteria.

THE MUTATIONAL HAZARD HYPOTHESIS AND BACTERIA

Because genome size in bacteria scales positively with N_e, bacteria defy the predictions of the mutational hazard hypothesis. Bacteria tend to have larger genomes when selection is more effective (Kuo et al., 2009; Sela et al., 2016), whereas eukaryotes have more streamlined genomes when selection is more effective (Lynch and

Conery, 2003; Lynch et al., 2011). This raises a paradox as to how and why the same force leads to opposite effects in bacteria and eukaryotes.

The answer resides in differences in the mutational processes: in bacteria, there is a strong mutational bias toward deleting superfluous sequences (Andersson and Andersson, 2001; Mira et al., 2001). It has long been known that gene number increases linearly with genome size in bacteria and that pseudogenes are rare or absent from bacterial genomes. This contrasts that situation in eukaryotic lineages in which there is little correlation between genome size and gene number—the "C-value paradox"—and there are pseudogenized copies of most genes (Lynch, 2007). In bacteria, deletional bias is apparent at all levels of genome organization: individual strains in culture incur large deletions encompassing up to 5% of their genome (Nilsson et al., 2005), comparisons of pseudogenes to their functional counterparts show that inactivated regions perpetually erode by small deletions (Mira et al., 2001; Kuo et al., 2009), and broad phylogenetic comparisons indicate that lineages of host-associated bacteria with small genomes derive from ancestors with large genomes over evolutionary timescales (Ochman, 2005).

The reason that bacterial species undergoing less effective selection (i.e., lower N_e) have smaller genomes is that they have accrued and tolerated more deleterious mutations due to drift. This is particularly evident in the genomes of pathogens and symbionts since their host-associated lifestyle both increases the fixation of slightly deleterious mutations and renders many previously useful genes redundant in the nutrient-rich host environment, thereby generating large numbers of non-essential regions that are subsequently removed by the pervasive mutational bias toward deletions. Note that the primary force countering gene erosion and elimination is natural selection, with the result that bacterial genomes, both large and small, maintain a high density of functional sequences (Ochman and Moran, 2001).

Genetic drift, coupled with deletional bias, are major determinants of bacterial genome size, such that species with the smallest N_e have the smallest genomes. But some—the marine bacteria—do not follow this trend and represent a curious exception. Marine bacteria have very large census population sizes but possess highly reduced genomes, on the order ∼1.5 Mb in length (Giovannoni et al., 2014; Kashtan et al., 2014). Moreover, these genomes harbor the smallest amount of intergenic DNA, with a median spacer length of only 3 bp between coding regions (Giovannoni et al., 2005). It has been hypothesized that genome reduction in marine species results from the efficacy of selection that can only occur in extremely large populations: these organisms live in nutrient-limited environments such that elimination of each non-essential nucleotide imparts an advantage by reducing the metabolic costs associated with DNA replication and processing (Giovannoni et al., 2014). In most populations, fitness differences this small would not be discriminated by selection; however, marine species provide a special case where selection, not genetic drift, governs genome size reduction.

EFFECTS OF POPULATION SIZE ON GENOME CONTENT AND COMPLEXITY

The linear relationship between genome size and gene number in bacteria implies that the proportion of non-coding and intergenic DNA is the same in all genomes. The effects of population size are also evident on bacterial genome complexity, i.e., the number and fraction of functional genes in a genome. Whereas intergenic regions typically constitute 10 ± 5% of a bacterial genome, species subject to drift sometimes can have much greater amounts of DNA that do not specify functional proteins. In particular, the genomes of bacteria that have sustained episodes of strong reductions in population size, such as pathogens and symbionts have recently become associated with hosts, contain large numbers of pseudogenes and/or mobile elements.

Most bacterial genomes maintain very low numbers of insertion sequence (IS) elements (<10; Touchon and Rocha, 2007) whereas several recent pathogens (e.g., *Shigella* spp. and *Rickettsia* spp.; Fuxelius et al., 2007; Touchon et al., 2009) and symbionts (e.g., *Sodalis glossinidius* and *Serratia symbiotica*; Toh et al., 2006; McCutcheon and Moran, 2012; Manzano-Marin and Latorre, 2014) possess hundreds of copies. Similarly, many host-associated bacteria, such as *Mycobacterium leprae* and *Endomicrobium* spp. (Cole et al., 2000; Zheng et al., 2016) harbor large numbers of pseudogenes when compared to their free-living relatives (Lerat and Ochman, 2005). The surge in the numbers of IS elements and pseudogenes in recent pathogens and symbionts conforms with the expectations of the mutational hazard hypothesis: severe reductions in population size result in less effective selection, which promotes the accumulation of non-functional and slightly deleterious sequences. Note that the proliferation of IS elements and pseudogenes is observed only during the initial stages of genome reduction since these sequences will eventually be purged from the genome by mutational processes (Moran and Plague, 2004).

In contrast to IS elements and pseudogenes, the proportion of bacterial genomes occupied by prophages increases with genome size (Touchon et al., 2016), a surprising relationship given that population sizes are larger, and selection more effective, in bacteria with larger genomes. While prophages may occasionally encode beneficial functions, most of their genes are of no consequence to their bacterial host (Ptashne, 1992; Casjens, 2003) and are expected to be eliminated. However, bacteria harboring prophages could be favored in a competitive environment, since these elements can potentially be used to eliminate competitors (Brown et al., 2006). When considering all bacteria, the majority of genome size variation is due to the gain and loss of accessory genes (Touchon et al., 2009) whose functions are thought to help bacteria cope with different niches or lifestyle. That bacteria with larger population sizes accommodate more accessory genes could reflect the fact that large populations likely span more diverse ecological conditions and require larger gene repertoires (Juhas et al., 2009) or that larger populations experience more competition, since many accessory genes are now known to be involved in bacterial warfare (Wexler et al., 2016). Hence, accessory genes, and perhaps prophages, represent a diverse arsenal that allows bacteria to adapt to their ever-changing and competitive environments. The ability of a bacterial species to capture and maintain a diverse repertoire of accessory genes likely constitutes a key feature to occupying a wide range of environments and maintaining large population sizes.

Because bacteria can undergo frequent bouts of horizontal gene acquisition (HGT; Ochman et al., 2000), the genome contents and architecture of closely related strains within a bacterial species can vary in ways that are not apparent in eukaryotes. Members of the same eukaryote species typically do not vary in their gene repertoires, and the acquisition of functional sequences in eukaryotes rarely results from HGT (Keeling, 2009). These key differences between bacteria and eukaryotes help drive, in addition to their respective biases toward insertions and deletions, the evolution of genome sizes toward opposite directions when exposed to drift. Thus, bacterial genomes increase in size by aggregating adaptive gene modules when exposed to new selective pressures, whereas eukaryotic genomes increase in size by accumulating large amounts of non-functional DNA when exposed to drift.

AUTHOR CONTRIBUTIONS

Both authors, HO and L-MB, contributed equally to the conception, contents, and writing of this manuscript.

FUNDING

The work was supported by National Institutes of Health grant number R35GM118038 awarded to HO.

REFERENCES

Andersson, J. O., and Andersson, S. G. (2001). Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes. *Mol. Biol. Evol.* 18, 829–839. doi: 10.1093/oxfordjournals.molbev.a003864

Batut, B., Knibbe, C., Marais, G., and Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat. Rev. Microbiol.* 12, 841–850. doi: 10.1038/nrmicro3331

Betancourt, A. J., Welch, J. J., and Charlesworth, B. (2009). Reduced effectiveness of selection caused by a lack of recombination. *Curr. Biol.* 19, 655–660. doi: 10.1016/j.cub.2009.02.039

Bobay, L. M., and Ochman, H. (2017). Biological species are universal across Life's domains. *Genome Biol. Evol.* doi: 10.1093/gbe/evx026 [Epub ahead of print].

Boccard, F., Esnault, E., and Valens, M. (2005). Spatial arrangement and macrodomain organization of bacterial chromosomes. *Mol. Microbiol.* 57, 9–16. doi: 10.1111/j.1365-2958.2005.04651.x

Brown, S. P., Le Chat, L., De Paepe, M., and Taddei, F. (2006). Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. *Curr. Biol.* 16, 2048–2052. doi: 10.1016/j.cub.2006.08.089

Casjens, S. (2003). Prophages and bacterial genomics: What have we learned so far? *Mol. Microbiol.* 49, 277–300. doi: 10.1046/j.1365-2958.2003.03580.x

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303.

Cohan, F. M. (2001). Bacterial species and speciation. *Syst. Biol.* 50, 513–524. doi: 10.1080/10635150118398

Cole, S. T., Honore, N., and Eiglmeier, K. (2000). Preliminary analysis of the genome sequence of *Mycobacterium leprae*. *Lepr. Rev.* 71(Suppl.), S162–S164; discussion S164–S167. doi: 10.5935/0305-7518.20000088

Daubin, V., and Moran, N. A. (2004). Comment on "the origins of genome complexity". *Science* 306:978; author reply 978. doi: 10.1126/science.1098469

Eisen, J. A., Heidelberg, J. F., White, O., and Salzberg, S. L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* 1:RESEARCH0011. doi: 10.1186/gb-2000-1-6-research0011

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics* 78, 737–756.

Funk, D. J., Wernegreen, J. J., and Moran, N. A. (2001). Intraspecific variation in symbiont genomes: bottlenecks and the aphid-buchnera association. *Genetics* 157, 477–489.

Fuxelius, H. H., Darby, A., Min, C. K., Cho, N. H., and Andersson, S. G. (2007). The genomic and metabolic diversity of Rickettsia. *Res. Microbiol.* 158, 745–753. doi: 10.1016/j.resmic.2007.09.008

Giovannoni, S. J., Cameron Thrash, J., and Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME J.* 8, 1553–1565. doi: 10.1038/ismej.2014.60

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309, 1242–1245. doi: 10.1126/science.1114057

Hartl, D. L., and Clark, A. G. (2007). *Principles of Population Genetics*, 4th Edn. Sunderland, MA: Sinauer Associates.

Herdman, M. (1985). "The evolution of bacterial genomes," in *The Evolution of Genome Size*, ed. T. Cavalier-Smith (New York, NY: John Wiley and Sons), 37–68.

Hill, W. G., and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet. Res.* 8, 269–294. doi: 10.1017/S0016672300010156

Hugenholtz, P., Skarshewski, A., and Parks, D. H. (2016). Genome-based microbial taxonomy coming of age. *Cold Spring Harb. Perspect. Biol.* 8:a018085. doi: 10.1101/cshperspect.a018085

Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., and Crook, D. W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.* 33, 376–393. doi: 10.1111/j.1574-6976.2008.00136.x

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* 344, 416–420. doi: 10.1126/science.1248575

Keeling, P. J. (2009). Functional and ecological impacts of horizontal gene transfer in eukaryotes. *Curr. Opin. Genet. Dev.* 19, 613–619. doi: 10.1016/j.gde.2009.10.001

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624–626. doi: 10.1038/217624a0

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2567–2572. doi: 10.1073/pnas.0409727102

Kryazhimskiy, S., and Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genet.* 4:e1000304. doi: 10.1371/journal.pgen.1000304

Kuo, C. H., Moran, N. A., and Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res.* 19, 1450–1454. doi: 10.1101/gr.091785.109

Lederberg, J. (1998). Plasmid (1952–1997). *Plasmid* 39, 1–9. doi: 10.1006/plas.1997.1320

Lerat, E., and Ochman, H. (2005). Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res.* 33, 3125–3132. doi: 10.1093/nar/gki631

Lynch, M. (2002). Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6118–6123. doi: 10.1073/pnas.092595699

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.* 60, 327–349. doi: 10.1146/annurev.micro.60.080805.142300

Lynch, M. (2007). *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Associates.

Lynch, M., Bobay, L. M., Catania, F., Gout, J. F., and Rho, M. (2011). The repatterning of eukaryotic genomes by random genetic drift. *Annu. Rev. Genomics Hum. Genet.* 12, 347–366. doi: 10.1146/annurev-genom-082410-101412

Lynch, M., and Conery, J. S. (2003). The origins of genome complexity. *Science* 302, 1401–1404. doi: 10.1126/science.1089370

Manzano-Marin, A., and Latorre, A. (2014). Settling down: the genome of *Serratia symbiotica* from the aphid *Cinara tujafilina* zooms in on the process of accommodation to a cooperative intracellular life. *Genome Biol. Evol.* 6, 1683–1698. doi: 10.1093/gbe/evu133

Mayr, E. (1942). *Systematics and the Origin of Species*. New York, NY: Columbia University Press.

McCutcheon, J. P., and Moran, N. A. (2010). Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol. Evol.* 2, 708–718. doi: 10.1093/gbe/evq055

McCutcheon, J. P., and Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10, 13–26.

Mira, A., Ochman, H., and Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17, 589–596. doi: 10.1016/S0168-9525(01)02447-7

Moran, N. A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 93, 2873–2878. doi: 10.1073/pnas.93.7.2873

Moran, N. A., McLaughlin, H. J., and Sorek, R. (2009). The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323, 379–382. doi: 10.1126/science.1167140

Moran, N. A., and Plague, G. R. (2004). Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* 14, 627–633. doi: 10.1016/j.gde.2004.09.003

Nilsson, A. I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J. C., and Andersson, D. I. (2005). Bacterial genome size reduction by experimental evolution. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12112–12116. doi: 10.1073/pnas.0503654102

Ochman, H. (2005). Genomes on the shrink. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11959–11960. doi: 10.1073/pnas.0505863102

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304. doi: 10.1038/35012500

Ochman, H., and Moran, N. A. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292, 1096–1099. doi: 10.1126/science.1058543

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* 246, 96–98. doi: 10.1038/246096a0

Price, M. N., and Arkin, A. P. (2015). Weakly deleterious mutations and low rates of recombination limit the impact of natural selection on bacterial genomes. *mBio* 6:e01302. doi: 10.1128/mBio.01302-15

Ptashne, M. (1992). *Genetic Switch: Phage Lambda and Higher Organisms*. Cambridge, MA: Blackwell.

Rocha, E. P. (2008). The organization of the bacterial genome. *Annu. Rev. Genet.* 42, 211–233. doi: 10.1146/annurev.genet.42.110807.091653

Rocha, E. P., and Feil, E. J. (2010). Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS Genet.* 6:e1001104. doi: 10.1371/journal.pgen.1001104

Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., et al. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* 239, 226–235. doi: 10.1016/j.jtbi.2005.08.037

Sela, I., Wolf, Y. I., and Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* 113, 11399–11407. doi: 10.1073/pnas.1614083113

Shapiro, B. J., Leducq, J. B., and Mallet, J. (2016). What is speciation? *PLoS Genet.* 12:e1005860. doi: 10.1371/journal.pgen.1005860

Shapiro, B. J., and Polz, M. F. (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol.* 22, 235–247. doi: 10.1016/j.tim.2014.02.006

Shapiro, B. J., and Polz, M. F. (2015). Microbial Speciation. *Cold Spring Harb. Perspect. Biol.* 7:a018143. doi: 10.1101/cshperspect.a018143

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35. doi: 10.1017/S0016672300014634

Sung, W., Ackerman, M. S., Dillon, M. M., Platt, T. G., Fuqua, C., Cooper, V. S., et al. (2016). Evolution of the insertion-deletion mutation rate across the tree of life. *G3* 6, 2583–2591. doi: 10.1534/g3.116.030890

Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18488–18492. doi: 10.1073/pnas.1216223109

Takahata, N. (1993). Allelic genealogy and human evolution. *Mol. Biol. Evol.* 10, 2–22.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., et al. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17, 520–526. doi: 10.1101/gr.6023607

Thomas, S. H., Wagner, R. D., Arakaki, A. K., Skolnick, J., Kirby, J. R., Shimkets, L. J., et al. (2008). The mosaic genome of *Anaeromyxobacter dehalogenans* strain 2CP-C suggests an aerobic common ancestor to the delta-*proteobacteria*. *PLoS ONE* 3:e2103. doi: 10.1371/journal.pone.0002103

Tillier, E. R., and Collins, R. A. (2000). Genome rearrangement by replication-directed translocation. *Nat. Genet.* 26, 195–197. doi: 10.1038/79918

Toh, H., Weiss, B. L., Perkin, S. A., Yamashita, A., Oshima, K., Hattori, M., et al. (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res.* 16, 149–156. doi: 10.1101/gr.4106106

Touchon, M., Bernheim, A., and Rocha, E. P. (2016). Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.* 10, 2744–2754. doi: 10.1038/ismej.2016.47

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., et al. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5:e1000344. doi: 10.1371/journal.pgen.1000344

Touchon, M., and Rocha, E. P. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* 24, 969–981. doi: 10.1093/molbev/msm014

Vos, M., and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 3, 199–208. doi: 10.1038/ismej.2008.93

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276. doi: 10.1016/0040-5809(75)90020-9

Wexler, A. G., Bao, Y., Whitney, J. C., Bobay, L. M., Xavier, J. B., Schofield, W. B., et al. (2016). Human symbionts inject and neutralize antibacterial toxins to persist in the gut. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3639–3644. doi: 10.1073/pnas.1525637113

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* 16, 97–159.

Zheng, H., Dietrich, C., Hongoh, Y., and Brune, A. (2016). Restriction-modification systems as mobile genetic elements in the evolution of an intracellular symbiont. *Mol. Biol. Evol.* 33, 721–725. doi: 10.1093/molbev/msv264

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Genetics

MINI REVIEW
published: 21 March 2017
doi: 10.3389/fgene.2017.00034

Unveiling the Impact of the Genomic Architecture on the Evolution of Vertebrate microRNAs

Gustavo S. França[1†], Ludwig C. Hinske[2], Pedro A. F. Galante[3] and Maria D. Vibranovski[1*]*

[1] Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo, Brazil, [2] Department of Anesthesiology, Clinic of the University of Munich, Ludwig Maximilian University of Munich, Munich, Germany, [3] Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo, Brazil

OPEN ACCESS

Edited by:
*Scott V. Edwards,
Harvard University, USA*

Reviewed by:
*Antonio Marco,
University of Essex, UK
Bastian Fromm,
Oslo University Hospital, Norway
Danillo Pinhal,
São Paulo State University (UNESP),
Brazil*

***Correspondence:**
*Gustavo S. França
gustavo.starvaggifranca@nyumc.org
Maria D. Vibranovski
mdv@ib.usp.br*

†Present address:
*Gustavo S. França,
Institute for Computational Medicine,
NYU School of Medicine,430 East
29th Street, New York, NY, USA*

Specialty section:
*This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics*

Received: *09 December 2016*
Accepted: *09 March 2017*
Published: *21 March 2017*

Citation:
França GS, Hinske LC, Galante PAF and Vibranovski MD (2017) Unveiling the Impact of the Genomic Architecture on the Evolution of Vertebrate microRNAs. Front. Genet. 8:34. doi: 10.3389/fgene.2017.00034

Eukaryotic genomes frequently exhibit interdependency between transcriptional units, as evidenced by regions of high gene density. It is well recognized that vertebrate microRNAs (miRNAs) are usually embedded in those regions. Recent work has shown that the genomic context is of utmost importance to determine miRNA expression in time and space, thus affecting their evolutionary fates over long and short terms. Consequently, understanding the inter- and intraspecific changes on miRNA genomic architecture may bring novel insights on the basic cellular processes regulated by miRNAs, as well as phenotypic evolution and disease-related mechanisms.

Keywords: intragenic, intergenic region, new and old miRNAs, host gene, target interactions, expression breadth

INTRODUCTION

Recent genome-wide projects have revealed an outstanding transcriptome diversity, especially of non-coding RNAs (ncRNAs), as well as a wealth of regulatory mechanisms and gene product interactions that compound the molecular basis of phenotypes (Carninci et al., 2005; Mele et al., 2015). A notable feature that soon became clear is the interleaved nature of eukaryotic genomes, despite their typical large sizes. This means that a particular genomic region can be suited for different purposes, with an extensive overlap of transcriptional units either in sense or antisense DNA strands (Kapranov et al., 2007).

The interleaved model opens up numerous possibilities for regulatory mechanisms. For instance, products of antisense transcription, which is believed to occur in more than 30% of gene loci in humans (Galante et al., 2007), can regulate gene activity through many different ways (reviewed in Pelechano and Steinmetz, 2013). In the interleaved genome, transcription units may show high interdependency, whereby neighboring or overlapping genes can be co-regulated by shared regulatory elements; yet, structural changes in the chromatin environment can also influence their expression coordinately (Mellor et al., 2016). Complex transcriptional networks thus emerge from a modular architecture that can either be shaped by evolutionary advantages and constraints (Mercer and Mattick, 2013), but also as a result of neutral processes (Graur et al., 2015). Such interleaved architecture is particularly striking in regard to microRNAs (miRNAs). Ever since the first large-scale studies on their genomic organization (Rodriguez et al., 2004), it is commonly observed that these small non-coding RNAs overlap to protein-coding genes, with vertebrate miRNAs mapping to intronic regions more than expected by chance (Baskerville and Bartel, 2005; Hinske et al., 2010, 2014; Campo-Paysaa et al., 2011; Meunier et al., 2013). As they comprise an essential class of gene expression regulators in basic biological processes and diseases, genomic context analyses are pivotal to uncover unique aspects of miRNA biology. Here, we discuss recent advances in this topic focusing on the importance of the genomic context to miRNA expression and their target interactions. In this framework, we highlight the evolutionary consequences for the fixation of newly emerged miRNAs and functional properties arising from miRNA–genomic context relationships over long-and short-evolutionary terms.

THE IMPACT OF THE GENOMIC CONTEXT ON miRNA EXPRESSION AND FUNCTION

As any other gene, the evolutionary processes that gives rise to new miRNAs – mainly by duplication or *de novo* origin (Berezikov, 2011; Meunier et al., 2013) – takes place on certain regions of the genome that may overlap or not to preexisting gene loci. In a recent study, Meunier et al. (2013) showed that all vertebrate species analyzed (Chicken, Platypus, Opossum, Mouse, Macaque, and Human) have a significant excess of intragenic miRNAs, with on average 54% of them overlapping to introns. Curiously, the proportions of intronic miRNAs are even higher for those of recent origin, suggesting that introns are hotspots for new miRNA origination. Moreover, the transcriptional orientation of intragenic miRNAs is highly biased (∼80%) toward the same strand orientation of their host genes (Rodriguez et al., 2004; Campo-Paysaa et al., 2011; Meunier et al., 2013; Hinske et al., 2014).

Given the large size of vertebrate genomes, why do miRNAs apparently have such preference to emerge in intragenic regions? Which evidences support the role of natural selection shaping this pattern, and what advantages miRNAs might take from such genomic organization? To address these questions, França et al. (2016) investigated the patterns of emergence and expression of human miRNAs along the vertebrate evolution considering the evolutionary origin of their host genes, i.e., whether miRNAs are intergenic, mapped to old protein-coding genes (originated before fish and tetrapods divergence), or to young protein-coding genes (originated after the divergence). Similar to previous studies (Iwama et al., 2013), it was shown that most human miRNAs (∼70%) have a relatively recent origin, emerging in the primate order. Though an interesting pattern was revealed, the majority of those young miRNAs are intragenic and preferentially embedded within old host genes, even when controlled by host gene length (including intronic region) and expression level. Expression breadth analyses showed that young miRNAs hosted by old genes were more broadly expressed (expression in more tissues) than their intergenic counterparts. On the other hand, miRNAs hosted by young genes showed a bias to tissue-specific expression when compared to the intergenic ones or those within old genes. The same conclusions held when a very stringent miRNA annotation provided by Fromm et al. (2015) was considered, since several miRBase entries do not represent bonafide miRNAs (Chiang et al., 2010; Taylor et al., 2014; Fromm et al., 2015). It is well established that expression breadth is negatively correlated with evolutionary rates (Wolf et al., 2009; Park and Choi, 2010), meaning that overall conserved genes are highly and broadly expressed, whereas less conserved genes tend to have low and narrow expression. What turns out is that the expression of intragenic miRNAs is tightly coupled to their genomic environment, especially in regard to the evolutionary ages of their host genes. In a mechanistic way, this is clearly connected with the co-expression of miRNA–host gene pairs by shared regulatory elements, a very well-documented event (Baskerville and Bartel, 2005; Ozsolak et al., 2008; Marsico et al., 2013). Hence, the maintenance of miRNAs embedded in genic regions may be indicative of some evolutionary constraint, since young and older intragenic miRNAs are biased toward host gene sense orientation, as well as preferential emergence within old host genes. In addition, same age miRNAs show differential expression breadth depending on their genomic context, a pattern that is maintained not only during recent (e.g., primates) but also over longer periods. Such pattern is observed for miRNAs originated in amniotes (e.g., chicken) or in placental mammals (e.g., mouse) presenting higher or lower expression breadth depending on the age of their host genes (França et al., 2016).

In particular for young intragenic miRNAs, being hosted by old genes could be beneficial at least during an initial adaptive phase, because of the expression broadness achieved through a presumably favorable transcriptional environment. Instead of readily relying on the settlement of their own regulatory apparatus, young miRNAs would initially been benefited by their hosts' regulatory elements, albeit they may acquire independent regulation afterward (França et al., 2016). Supporting this notion, it has been suggested that young and middle-aged intragenic miRNAs are more likely to be regulated by shared promoters, whereas old miRNAs are frequently regulated by their independent intronic promoters (Marsico et al., 2013). In addition, as old host genes provide higher expression breadth for those young miRNAs, it would, in principle, increase the opportunities for new target interactions in different tissues. From such perspective, the host transcriptional environment could facilitate the initial expression of young miRNAs and thereafter contribute to the process of miRNA functionalization.

The location of a gene in the genome is clearly related to its expression, as revealed by transgene insertion experiments (Mlynárová et al., 2002) and global expression analyses of gene neighborhoods (Caron et al., 2001; Purmann et al., 2007; Michalak, 2008). Nevertheless, some of the observed expression changes in gene vicinity may not be subjected to selection, but rather it would be a consequence of expression changes in a close gene under strong selection. Recently, Ghanbarian and Hurst (2015) demonstrated that expression changes in humans, relative to the human–chimp common ancestor, coordinately drive changes in expression of the neighbors of a focal gene, and that this effect is stronger as the distance between genes are shorter (<100kbp). Therefore, the genomic context still may yield important effects on the expression, and perhaps the fixation of novel miRNAs that are not under direct selection.

EVOLUTIONARY CONSERVATION AND NOVELTIES FROM miRNAs' GENOMIC CONTEXT

The phylogenetic distribution of miRNAs in vertebrates is distinguished by the presence of deeply conserved and abundant clade or species-specific repertoires (Berezikov et al., 2006; Wheeler et al., 2009; Meunier et al., 2013; Fromm et al., 2015). Although the evolution of miRNA sequences have been investigated (Lyu et al., 2014; Ninova et al., 2014), the conserved patterns and evolutionary innovations that arose due to interspecific differences in the genomic context are largely underexplored. One of the few studies to address this issue compared the genomic location and expression of ∼100 miRNAs during developmental stages of medaka fish, zebrafish, chicken, and mouse (Ason et al., 2006). It was demonstrated that spatial expression differences can be related to changes either in the miRNA location and copy number variation rather than to sequence divergence (Ason et al., 2006). Actually, the miRNA genomic location is thought to influence their expression divergence, as old- and middle-aged intragenic miRNAs tend to be more similarly expressed among species than intergenic ones (França et al., 2016).

Such kind of expression constraint linked to a conserved genomic context is clearly observed for *miR-490* and its host gene *CHRM2* (França et al., 2016). Homologous sequences of *miR-490* are found across amniotes, with identical mature sequences from human to chicken. Gene order and location of *miR-490* in the second intron of *CHRM2* are also preserved (**Figure 1A**). Although *miR-490* is annotated as intergenic in chicken, predicted transcripts with an intron overlapping *miR-490* are annotated. Expression analyses reveal a strongly conserved pattern among human, rhesus macaque, mouse, and chicken; indicating concomitant expression of *miR-490* and *CHRM2* (Shen et al., 2015) with highest abundance in heart (**Figure 1A**). The host gene is a muscarinic cholinergic receptor involved in acetylcholine-mediated cardiac chronotropic (heart rate) and inotropic (strength of muscle contraction) effects (Brodde and Michel, 1999), and it has been associated with cardiomyopathy (Zhang et al., 2008). Notably, dysregulation of *miR-490* is also reported in cardiac disease (Cooley et al., 2012) and is involved with proliferation of human coronary artery smooth cells (Sun et al., 2013), suggesting an important functional connection between *miR-490* and *CHRM2*.

As mentioned earlier, the transcriptional environment of host genes may act as a key factor to promote the expression of newly emerged miRNAs. This phenomenon is well illustrated by the primate-specific *miR-625* encoded within *FUT8* (**Figure 1B**). This host gene is a fucosyltransferase well-conserved throughout animals (Costache et al., 1997; Juliant et al., 2014) that catalyzes fucosylation of glycoproteins, which is essential for activating growth factor receptors (Liu et al., 2011), while its deletion has lethal effects in mice (Wang et al., 2005). *FUT8* is ubiquitously expressed in human tissues (Mele et al., 2015) and *miR-625* seems to follow its host expression pattern (**Figure 1B**). Considering the young evolutionary age of *miR-625*, its expression levels and breadth are unusually high, thus being frequently altered in different types of cancer (Zhou et al., 2014; Zheng et al., 2015). It is interesting that *miR-625* has emerged as a promising predictive biomarker in colorectal cancer (Verma et al., 2015; Rasmussen et al., 2016), exhibiting strong association with oxaliplatin (a chemotherapeutic agent used in the treatment of metastatic colorectal cancer) resistance (Rasmussen et al., 2016).

Another singular feature of miRNAs is their frequent occurrence in clusters, originated through tandem or non-local duplications or by *de novo* mutations either in introns or intergenic regions (Berezikov, 2011). Such genomic organization is prone to greatly affect the evolution of newly emerged miRNAs. According to Wang et al. (2016), members of the same cluster tend to exhibit coordinated expression and to target overlapping sets of genes. The authors proposed that clustering arrangement and by developing functions related to the pre-existing miRNAs in the same cluster would help the initial survival of these young miRNAs, until the cluster is settled up by purifying selection. Otherwise, the most usual fate of *de novo* newly emerged miRNAs would to undergo rapid degeneration. In further support of this "functional co-adaptation" model, clustered young miRNAs indeed present significant signs of adaptive changes that probably drive them to functional constraints associated with the older members of the cluster (Wang et al., 2016).

miRNA–TARGET INTERACTIONS: FUNCTIONAL AND EVOLUTIONARY IMPLICATIONS

If a recently emerged miRNA is expressed and integrated into regulatory networks through consistent and biologically relevant target interactions, it will have more chances to become functional and be retained afterward over long periods (Chen and Rajewsky, 2007; Lyu et al., 2014). Therefore, young miRNAs originated in a genomic context able to boost their expression in multiple tissues would favor target recognition. This idea is consistent with the previous observation that young miRNAs emerged within old host genes are expressed in more tissues and tend to have more predicted targets compared to young intergenic ones (França et al., 2016). We, therefore, suggested a miRNA evolution model that takes into account not only the miRNAs themselves, but also their genomic context (França et al., 2016) (**Figure 2**). Hence, young miRNAs (or "proto" miRNAs) hosted by old genes would gain higher expression breadth benefited by their host's transcriptional activity, thus enabling many target interactions that, at first glance, are mostly neutral (Chen and Rajewsky, 2007; Nozawa et al., 2016), but could be stabilized by natural selection over time. On the other hand, as young intergenic miRNAs tend to have narrower expression, and apparently less targets to interact with, they could undergo faster degeneration (**Figure 2**). This degeneration scenario is also most likely to happen with miRNAs emerged within young hosts, because of their general tissue-specific expression signature (França et al., 2016).

Evolutionary sequence conservation has been successfully introduced to reduce the number of false-positive and to increase the signal-to-noise ratio in target predictions. Instead

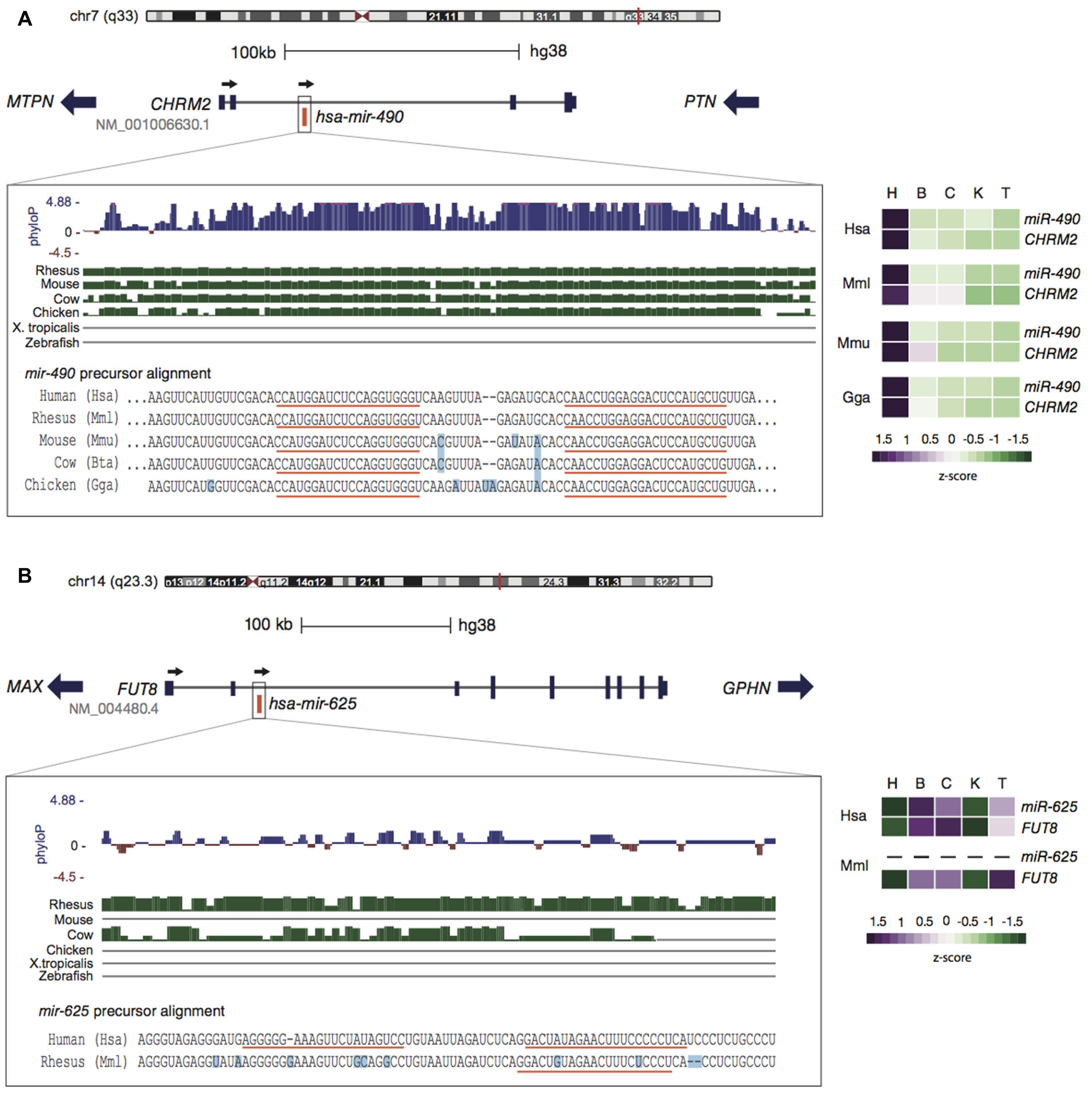


FIGURE 1 | Genomic context conservation of intragenic miRNAs. (A) The human *miR-490* embedded within *CHRM2* reveals a highly conserved pattern in terms of sequence (left panel) and expression (right panel). Alignments from the UCSC genome browser indicate the preservation of *miR-490* throughout amniotes (green bars) with few differing bases (light blue squares) and identical mature sequences (orange lines). High-phyloP base scores indicate strong purifying selection on this region. *MiR-490* and *CHRM2* are co-expressed with highest levels in heart, a pattern conserved in other species. **(B)** The human *miR-625*, encoded within *FUT8*, has homologous sequences only in primates. The expression of *miR-625* follows its host pattern, with higher levels in brain and cerebellum, possibly reflecting rapid evolution. Expression of *miR-625* in rhesus was not detected. Expression data were obtained from Brawand et al. (2011) and Meunier et al. (2013) and processed in França et al. (2016). Tissues are: heart (H), brain (B), cerebellum (C), kidney (K), and testis (T).

of helping identifying conserved pathways and relationships among miRNAs and their targets (Hausser and Zavolan, 2014), this requirement comes with a drawback, since it can only be applied to miRNAs and target genes that have conservation data available and which are not species-specific. Indeed, a recent study demonstrated that target sites identified by cross-linking immunoprecipitation data are rarely conserved between distantly related species, but extensive conservation is observed between closely related ones (Xu et al., 2013). Even when considering species-specific sites, there is evidence of selective constraints compared to non-target sites across the 3′UTR region, suggesting that most of non-conserved targets might be functional at least for a short evolutionary period. A striking example of this condition is the human-specific target site for *miR-183* in the 3vUTR of the transcription factor *FOXO1*, whose regulation altered FOXO1-dependent phenotypes, such as proliferation and migration, in a species-specific manner (McLoughlin et al., 2014). Despite of the recent advances on the characterization of operating mechanisms that guide miRNA–target interactions, we are only on the verge of understanding how newly emerged

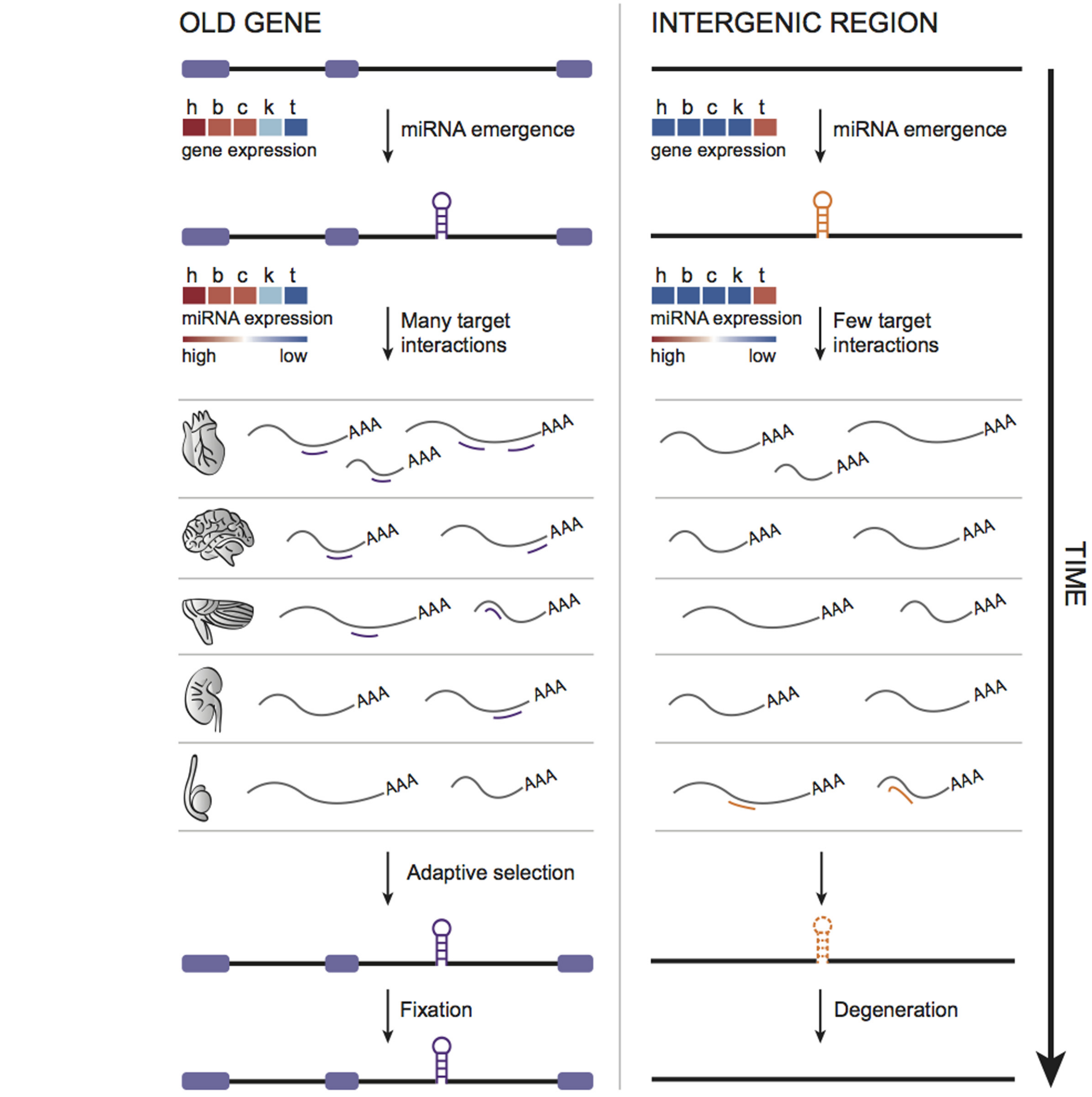


FIGURE 2 | Model of miRNA evolution. Young miRNAs emerged within old genes are expressed in more tissues and, therefore, could interact with diverse set of targets, possibly enhancing the chances of functionalization and fixation through time. In contrast, as young intergenic miRNAs tend to be tissue-specific (likely expressed in testis), very limited target interactions could contribute to their faster degeneration.

miRNAs in different genomic contexts are integrated into regulatory networks, as well as how their novel target interactions contribute to phenotypic plasticity.

POPULATION BIOLOGY PERSPECTIVE FOR THE GENOME ARCHITECTURE OF miRNAS

Population biology studies at the genome level have been proved to be promising tools, enhancing our understanding on how genetic elements are interconnected spatially and temporally (Barrón et al., 2014; Sudmant et al., 2015). Most of miRNA population studies have focused on the impact of single nucleotide variants localized inside the seed and the mature regions to analyze conservation patterns, target diversification, and differential disease susceptibility (e.g., Barbash et al., 2014; Rawlings-Goss et al., 2014; Gallego et al., 2016). Except for few studies of miRNA expression quantitative trait loci (e.g., Huan et al., 2015), the evolution of miRNA genomic architecture has not been deeply investigated using a population biology framework.

It is still unknown if variation in miRNAs sequence, expression, and target sites across populations are more relevant for uncovering the mechanisms of phenotypic evolution and disease than other genetic variation. On one hand, due to its folded structure and small size, miRNAs are more likely to emerge *de novo* than novel protein coding genes (Berezikov, 2011). Diversification of miRNA target repertoire may be more prone to appear as result of simple sequence modifications such as direct mutation, seed or hairpin shifting, and arm switching (Berezikov, 2011). Therefore, variation on miRNA-binding sites indeed can lead to phenotypic innovation, as exemplified by the lineage diversification of cichlid fishes (Loh et al., 2011; Franchini et al., 2016). On the other hand, as target mRNAs can be regulated subtly by several miRNAs, detecting phenotypical effects by population variation seems to be harder than for genetic variation in regulatory or coding regions. Indeed, most of single nucleotide polymorphisms (SNPs) involved in the creation of novel miRNA target sites does not correlate with phenotypic differences among humans (Saunders et al., 2007).

Nonetheless, it is possible that genomic comparisons of different individuals can give insights on the origination process of miRNAs, as previously done for other genetic elements (Hatcher, 2000; Schlötterer, 2015). For instance, the basis of retrogene origination in metazoans has been recently deciphered through *Drosophila* population data. Flanking regions signatures of polymorphic retrocopies revealed that long terminal repeat (LTR) retrotransposons have mediate their formation (Tan et al., 2016). miRNAs are mostly originated *de novo* or by duplication (Meunier et al., 2013), but mechanistic details on how those processes occur are still unknown. Population genomics might help uncover those components through the identification of mutational signatures attached to polymorphic miRNAs that are usually erased by time and throughout their fixation.

In addition, comparing fixed patterns present in different species to polymorphic states observed in a group of individuals are useful tools for contrasting genomic features driven by natural selection to patterns produced by mutation bias (Long et al., 2013). Notable, this type of comparison helped to support the hypothesis in which natural selection drives retrogene duplication from the X chromosome to the autosomes in *Drosophila* and humans (Schrider et al., 2011, 2013; Navarro and Galante, 2015). Therefore, the analyses of different human populations can give further support to the adapted pattern of miRNAs organized inside old protein coding host genes.

Furthermore, as miRNAs expression and targeting has been shown to be implicated in a wide of human diseases (Mendell and Olson, 2012), seed, and mature region variants found among ethnic populations become clinically important (Rawlings-Goss et al., 2014). More specifically, there are distinct miRNA profiles in diseases between African and European descendants (e.g., Huang et al., 2011; Heegaard et al., 2012) which could be responsible for differences among those populations in susceptibility to diseases, drug sensitiveness, and biomarker diagnostics (Rawlings-Goss et al., 2014). Therefore, should worth investigating if ethnic group variation on miRNA genomic context have also significant role in human health.

From the discussion above, it turns out that the genomic context, as an outcome of natural selection, imposes evolutionary constraints to maintain the structural and functional integrity of its genetic elements. Moreover, it can also propel the evolutionary fate of new elements that arise in a suitable environment, eventually accelerating the process of functionalization. Therefore, evolutionary models tackling the 3D chromatin organization will be of extreme value to pursue the general principles that afford those processes take place throughout genomes.

AUTHOR CONTRIBUTIONS

GSF and MDV conceived the study. GSF, LCH, PAFG, and MDV wrote the manuscript. All authors revised and approved it for publication.

ACKNOWLEDGMENTS

This work was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) (381128/2016-4, to G.S.F) and FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) (15/20844-4, to M.D.V).

REFERENCES

Ason, B., Darnell, D. K., Wittbrodt, B., Berezikov, E., Kloosterman, W. P., Wittbrodt, J., et al. (2006). Differences in vertebrate microRNA expression. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14385–14389. doi: 10.1073/pnas.0603529103

Barbash, S., Shifman, S., and Soreq, H. (2014). Global coevolution of human micrornas and their target genes. *Mol. Biol. Evol.* 31, 1237–1247. doi: 10.1093/molbev/msu090

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., and González, J. (2014). Population genomics of transposable elements in *Drosophila*. *Annu. Rev. Genet.* 48, 561–581. doi: 10.1146/annurev-genet-120213-092359

Baskerville, S., and Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* 11, 241–247. doi: 10.1261/rna.7240905

Berezikov, E. (2011). Evolution of microRNA diversity and regulation in animals. *Nat. Rev. Genet.* 12, 846–860. doi: 10.1038/nrg3079

Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., et al. (2006). Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* 38, 1375–1377. doi: 10.1038/ng1914

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature* 478, 343–348. doi: 10.1038/nature10532

Brodde, O.-E., and Michel, M. C. (1999). Adrenergic and muscarinic receptors in the human heart. *Pharmacol. Rev.* 51, 651–690.

Campo-Paysaa, F., Sémon, M., Cameron, R. A., Peterson, K. J., and Schubert, M. (2011). microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evol. Dev.* 13, 15–27. doi: 10.1111/j.1525-142X.2010.00452.x

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., et al. (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563. doi: 10.1126/science.1112014

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., et al. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291, 1289–1292. doi: 10.1126/science.1056794

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* 8, 93–103. doi: 10.1038/nrg1990

Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeng, V. C., Spies, N., Baek, D., et al. (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.* 24, 992–1009. doi: 10.1101/gad.1884710

Cooley, N., Cowley, M. J., Lin, R. C. Y., Marasco, S., Wong, C., Kaye, D. M., et al. (2012). Influence of atrial fibrillation on microRNA expression profiles in left and right atria from patients with valvular heart disease. *Physiol. Genomics* 44, 211–219. doi: 10.1152/physiolgenomics.00111.2011

Costache, M., Apoil, P. A., Cailleau, A., Elmgren, A., Larson, G., Henry, S., et al. (1997). Evolution of fucosyltransferase genes in vertebrates. *J. Biol. Chem.* 272, 29721–29728. doi: 10.1074/jbc.272.47.29721

França, G. S., Vibranovski, M. D., and Galante, P. A. F. (2016). Host gene constraints and genomic context impact the expression and evolution of human microRNAs. *Nat. Commun.* 7:11438. doi: 10.1038/ncomms11438

Franchini, P., Xiong, P., Fruciano, C., and Meyer, A. (2016). The role of microRNAs in the repeated parallel diversification of lineages of midas cichlid fish from nicaragua. *Genome Biol. Evol.* 8, 1543–1555. doi: 10.1093/gbe/evw097

Fromm, B., Billipp, T., Peck, L. E., Johansen, M., Tarver, J. E., King, B. L., et al. (2015). A Uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.* 49, 213–242. doi: 10.1146/annurev-genet-120213-092023

Galante, P. A. F., Vidal, D. O., de Souza, J. E., Camargo, A. A., and de Souza, S. J. (2007). Sense-antisense pairs in mammals: functional and evolutionary considerantions. *Genome Biol.* 8:R40. doi: 10.1186/gb-2007-8-3-r40

Gallego, A., Melé, M., Balcells, I., García-Ramallo, E., Torruella-Loran, I., Fernández-Bellon, H., et al. (2016). Functional implications of human-specific changes in great ape microRNAs. *PLoS ONE* 11:e0154194. doi: 10.1371/journal.pone.0154194

Ghanbarian, A. T., and Hurst, L. D. (2015). Neighboring genes show correlated evolution in gene expression. *Mol. Biol. Evol.* 32, 1748–1766. doi: 10.1093/molbev/msv053

Graur, D., Zheng, Y., and Azevedo, R. B. (2015). An evolutionary classification of genomic function. *Genome Biol. Evol.* 7, 642–645. doi: 10.1093/gbe/evv021

Hatcher, M. J. (2000). Persistence of selfish genetic elements: population structure and conflict. *Trends Ecol. Evol.* 15, 271–277. doi: 10.1016/S0169-5347(00)01875-9

Hausser, J., and Zavolan, M. (2014). Identification and consequences of miRNA-target interactions–beyond repression of gene expression. *Nat. Rev. Genet.* 15, 599-612. doi: 10.1038/nrg3765

Heegaard, N. H., Schetter, A. J., Welsh, J. A., Yoneda, M., Bowman, E. D., and Harris, C. C. (2012). Circulating micro-RNA expression profiles in early stage nonsmall cell lung cancer. *Int. J. Cancer* 130, 1378–1386. doi: 10.1002/ijc.26153

Hinske, L. C., França, G. S., Torres, H. A. M., Ohara, D. T., Lopes-Ramos, C. M., Heyn, J., et al. (2014). miRIAD-integrating microRNA inter- and intragenic data. *Database (Oxford).* 2014, 1–9. doi: 10.1093/database/bau099

Hinske, L. C. G., Galante, P. A. F., Kuo, W. P., and Ohno-Machado, L. (2010). A potential role for intragenic miRNAs on their hosts' interactome. *BMC Genomics* 11:533. doi: 10.1186/1471-2164-11-533

Huan, T., Rong, J., Liu, C., Zhang, X., Tanriverdi, K., Joehanes, R., et al. (2015). Genome-wide identification of microRNA expression quantitative trait loci. *Nat. Commun.* 6:6601. doi: 10.1038/ncomms7601

Huang, R. S., Gamazon, E. R., Wen, Y., Im, H. K., Zhang, W., Wing, C., et al. (2011). Population differences in microRNA expression and biological implications. *RNA Biol.* 8, 692–701. doi: 10.4161/rna.8.4.16029

Iwama, H., Kato, K., Imachi, H., Murao, K., and Masaki, T. (2013). Human microRNAs originated from two periods at accelerated rates in mammalian evolution. *Mol. Biol. Evol.* 30, 613–626. doi: 10.1093/molbev/mss262

Juliant, S., Harduin-Lepers, A., Monjaret, F., Catieau, B., Violet, M. L., Cérutti, P., et al. (2014). The α1, 6-fucosyltransferase gene (FUT8) from the Sf9 lepidopteran insect cell line: insights into FUT8 evolution. *PLoS ONE* 9:e110422. doi: 10.1371/journal.pone.0110422

Kapranov, P., Willingham, A. T., and Gingeras, T. R. (2007). Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* 8, 413–423. doi: 10.1038/nrg2083

Liu, Y.-C., Yen, H.-Y., Chen, C.-Y., Chen, C.-H., Cheng, P.-F., Juan, Y.-H., et al. (2011). Sialylation and fucosylation of epidermal growth factor receptor suppress its dimerization and activation in lung cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11332–11337. doi: 10.1073/pnas.1107385108

Loh, Y. E., Yi, S. V., and Streelman, J. T. (2011). Evolution of microRNAs and the diversification of species. *Genome Biol. Evol.* 3, 55–65. doi: 10.1093/gbe/evq085

Long, M., VanKuren, N. W., Chen, S., and Vibranovski, M. D. (2013). New gene evolution: little did we know. *Annu. Rev. Genet.* 47, 307–333. doi: 10.1146/annurev-genet-111212-133301

Lyu, Y., Shen, Y., Li, H., Chen, Y., Guo, L., Zhao, Y., et al. (2014). New microRNAs in *Drosophila*–birth, death and cycles of adaptive evolution. *PLoS Genet.* 10:e1004096. doi: 10.1371/journal.pgen.1004096

Marsico, A., Huska, M. R., Lasserre, J., Hu, H., Vucicevic, D., Musahl, A., et al. (2013). PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol.* 14:R84. doi: 10.1186/gb-2013-14-8-r84

McLoughlin, H. S., Wan, J., Spengler, R. M., Xing, Y., and Davidson, B. L. (2014). Human-specific microRNA regulation of FOXO1: implications for microRNA recognition element evolution. *Hum. Mol. Genet.* 23, 2593–2603. doi: 10.1093/hmg/ddt655

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). The human transcriptome across tissues and individuals. *Science* 348, 660–665. doi: 10.1126/science.aaa0355

Mellor, J., Woloszczuk, R., and Howe, F. S. (2016). The interleaved genome. *Trends Genet.* 32, 57–71. doi: 10.1016/j.tig.2015.10.006

Mendell, J. T., and Olson, E. N. (2012). MicroRNAs in stress signaling and human disease. *Cell* 148, 1172–1187. doi: 10.1016/j.cell.2012.02.005

Mercer, T. R., and Mattick, J. S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome Res.* 23, 1081–1088. doi: 10.1101/gr.156612.113

Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., et al. (2013). Birth and expression evolution of mammalian microRNA genes. *Genome Res.* 23, 34–45. doi: 10.1101/gr.140269.112

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* 91, 243–248. doi: 10.1016/j.ygeno.2007.11.002

Mlynárová, L., Loonen, A., Mietkiewska, E., Jansen, R. C., and Nap, J. P. (2002). Assembly of two transgenes in an artificial chromatin domain gives highly coordinated expression in tobacco. *Genetics* 160, 727–740.

Navarro, F. C. P., and Galante, P. A. F. (2015). A genome-wide landscape of retrocopies in primate genomes. *Genome Biol. Evol.* 7, 2265–2275. doi: 10.1093/gbe/evv142

Ninova, M., Ronshaugen, M., and Griffiths-Jones, S. (2014). Fast-evolving microRNAs are highly expressed in the early embryo of *Drosophila virilis*. *RNA* 20, 360-372. doi: 10.1261/rna.041657.113

Nozawa, M., Fujimi, M., Iwamoto, C., Onizuka, K., Fukuda, N., Ikeo, K., et al. (2016). Evolutionary transitions of MicroRNA-target pairs. *Genome Biol. Evol.* 8, 1621–1633. doi: 10.1093/gbe/evw092

Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G., et al. (2008). Chromatin structure analyses identify miRNA promoters Chromatin structure analyses identify miRNA promoters. *Genes Dev*. 22, 3172–3183. doi: 10.1101/gad.1706508

Park, S. G., and Choi, S. S. (2010). Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol. Biol.* 10:241. doi: 10.1186/1471-2148-10-241

Pelechano, V., and Steinmetz, L. M. (2013). Gene regulation by antisense transcription. *Nat. Rev. Genet.* 14, 880–893. doi: 10.1038/nrg3594

Purmann, A., Toedling, J., Schueler, M., Carninci, P., Lehrach, H., Hayashizaki, Y., et al. (2007). Genomic organization of transcriptomes in mammals: coregulation and cofunctionality. *Genomics* 89, 580–587. doi: 10.1016/j.ygeno.2007.01.010

Rasmussen, M. H., Lyskjær, I., Jersie-Christensen, R. R., Tarpgaard, L. S., Primdal-Bengtson, B., Nielsen, M. M., et al. (2016). miR-625-3p regulates oxaliplatin resistance by targeting MAP2K6-p38 signalling in human colorectal adenocarcinoma cells. *Nat. Commun.* 7:12436. doi: 10.1038/ncomms12436

Rawlings-Goss, R. A., Campbell, M. C., and Tishkoff, S. A. (2014). Global population-specific variation in miRNA associated with cancer risk and clinical biomarkers. *BMC Med. Genomics* 7:53. doi: 10.1186/1755-8794-7-53

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., and Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Res.* 14, 1902–1910. doi: 10.1101/gr.2722704

Saunders, M. A., Liang, H., and Li, W. H. (2007). Human polymorphism at microRNAs and microRNA target sites. *Proc. Natl. Acad. Sci. U.S.A.* 104, 3300–3305. doi: 10.1073/pnas.0611347104

Schlötterer, C. (2015). Genes from scratch–the evolutionary fate of de novo genes. *Trends Genet.* 31, 215–219. doi: 10.1016/j.tig.2015.02.007

Schrider, D. R., Navarro, F. C. P., Galante, P. A. F., Parmigiani, R. B., Camargo, A. A., Hahn, M. W., et al. (2013). Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.* 9:e1003242. doi: 10.1371/journal.pgen.1003242

Schrider, D. R., Stevens, K., Cardeño, C. M., Langley, C. H., and Hahn, M. W. (2011). Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*. *Genome Res.* 21, 2087–2095. doi: 10.1101/gr.116434.110

Shen, J., Xiao, Z., Wu, W. K. K., Wang, M. H., To, K. F., Chen, Y., et al. (2015). Epigenetic silencing of miR-490-3p reactivates the chromatin remodeler SMARCD1 to promote *Helicobacter pylori*-induced gastric carcinogenesis. *Cancer Res.* 75, 754–765. doi: 10.1158/0008-5472.CAN-14-1301

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81. doi: 10.1038/nature15394

Sun, Y., Chen, D., Cao, L., Zhang, R., Zhou, J., Chen, H., et al. (2013). MiR-490-3p modulates the proliferation of vascular smooth muscle cells induced by ox-LDL through targeting PAPP-A. *Cardiovasc. Res.* 100, 272–279. doi: 10.1093/cvr/cvt172

Tan, S., Cardoso-Moreira, M., Shi, W., Zhang, D., Huang, J., Mao, Y., et al. (2016). LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Res.* 26, 1–13. doi: 10.1101/gr.204925.116

Taylor, R. S., Tarver, J. E., Hiscock, S. J., and Donoghue, P. C. J. (2014). Evolutionary history of plant microRNAs. *Trends Plant Sci.* 19, 175–182. doi: 10.1016/j.tplants.2013.11.008

Verma, A. M., Patel, M., Aslam, M. I., Jameson, J., Pringle, J. H., Wurm, P., et al. (2015). Circulating plasma microRNAs as a screening method for detection of colorectal adenomas. *Lancet* 385, S100. doi: 10.1016/S0140-6736(15)60415-9

Wang, X., Inoue, S., Gu, J., Miyoshi, E., Noda, K., Li, W., et al. (2005). Dysregulation of TGF-beta1 receptor activation leads to abnormal lung development and emphysema-like phenotype in core fucose-deficient mice. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15791–15796. doi: 10.1073/pnas.0507375102

Wang, Y., Luo, J., Zhang, H., and Lu, J. (2016). microRNAs in the same clusters evolve to coordinately regulate functionally related genes. *Mol. Biol. Evol.* 33, 2232–2247. doi: 10.1093/molbev/msw089

Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S., et al. (2009). The deep evolution of metazoan microRNAs. *Evol. Dev.* 11, 50–68. doi: 10.1111/j.1525-142X.2008.00302.x

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., and Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7273–7280. doi: 10.1073/pnas.0901808106

Xu, J., Zhang, R., Shen, Y., Liu, G., Lu, X., and Wu, C. -I. (2013). The evolution of evolvability in microRNA target sites in vertebrates. *Genome Res.* 23, 1810–1816. doi: 10.1101/gr.148916.112

Zhang, L., Hu, A., Yuan, H., Cui, L., Miao, G., Yang, X., et al. (2008). A missense mutation in the CHRM2 gene is associated with familial dilated cardiomyopathy. *Circ. Res.* 102, 1426–1432. doi: 10.1161/CIRCRESAHA.107.167783

Zheng, H., Ma, R., Wang, Q., Zhang, P., Li, D., Wang, Q., et al. (2015). MiR-625-3p promotes cell migration and invasion via inhibition of SCAI in colorectal carcinoma cells. *Oncotarget* 6, 27805–27815. doi: 10.18632/oncotarget.4738

Zhou, X., Zhang, C. Z., Lu, S. -X., Chen, G. G., Li, L. -Z., Liu, L. -L., et al. (2014). miR-625 suppresses tumour migration and invasion by targeting IGF2BP1 in hepatocellular carcinoma. *Oncogene* 34, 965–977. doi: 10.1038/onc.2014.35

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 February 2017
doi: 10.3389/fgene.2017.00010

Chromosomal Speciation in the Genomics Era: Disentangling Phylogenetic Evolution of Rock-wallabies

Sally Potter[1,2], Jason G. Bragg[3], Mozes P. K. Blom[4], Janine E. Deakin[5], Mark Kirkpatrick[6], Mark D. B. Eldridge[2] and Craig Moritz[1]*

[1] Research School of Biology, Australian National University, Acton, ACT, Australia, [2] Australian Museum Research Institute, Australian Museum, Sydney, NSW, Australia, [3] National Herbarium of New South Wales, The Royal Botanic Gardens and Domain Trust, Sydney, NSW, Australia, [4] Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden, [5] Institute for Applied Ecology, University of Canberra, Bruce, ACT, Australia, [6] Department of Integrative Biology, University of Texas, Austin, TX, USA

OPEN ACCESS

Edited by:
Tariq Ezaz,
University of Canberra, Australia

Reviewed by:
Rachel O'Neill,
University of Connecticut, USA
Lukas Kratochvil,
Charles University, Czechia

****Correspondence:***
Sally Potter
sally.potter@anu.edu.au

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *08 December 2016*
Accepted: *18 January 2017*
Published: *10 February 2017*

Citation:
Potter S, Bragg JG, Blom MPK, Deakin JE, Kirkpatrick M, Eldridge MDB and Moritz C (2017) Chromosomal Speciation in the Genomics Era: Disentangling Phylogenetic Evolution of Rock-wallabies. Front. Genet. 8:10. doi: 10.3389/fgene.2017.00010

The association of chromosome rearrangements (CRs) with speciation is well established, and there is a long history of theory and evidence relating to "chromosomal speciation." Genomic sequencing has the potential to provide new insights into how reorganization of genome structure promotes divergence, and in model systems has demonstrated reduced gene flow in rearranged segments. However, there are limits to what we can understand from a small number of model systems, which each only tell us about one episode of chromosomal speciation. Progressing from patterns of association between chromosome (and genic) change, to understanding processes of speciation requires both comparative studies across diverse systems and integration of genome-scale sequence comparisons with other lines of evidence. Here, we showcase a promising example of chromosomal speciation in a non-model organism, the endemic Australian marsupial genus *Petrogale*. We present initial phylogenetic results from exon-capture that resolve a history of divergence associated with extensive and repeated CRs. Yet it remains challenging to disentangle gene tree heterogeneity caused by recent divergence and gene flow in this and other such recent radiations. We outline a way forward for better integration of comparative genomic sequence data with evidence from molecular cytogenetics, and analyses of shifts in the recombination landscape and potential disruption of meiotic segregation and epigenetic programming. In all likelihood, CRs impact multiple cellular processes and these effects need to be considered together, along with effects of genic divergence. Understanding the effects of CRs together with genic divergence will require development of more integrative theory and inference methods. Together, new data and analysis tools will combine to shed light on long standing questions of how chromosome and genic divergence promote speciation.

Keywords: chromosome rearrangement, speciation, rock-wallaby, divergence, genomics

INTRODUCTION

Differences in how the genome is packaged – chromosome variation – have long been known to influence how genetic variation is transmitted and redistributed within and among populations (Darlington, 1958; White, 1973). Today, with increasing availability of high quality genome assemblies, the capacity for genome-scale resequencing and the tools of molecular cytogenetics and population- and phylo-genomic analysis, we are returning to a whole-genome perspective on evolution. At the same time, evidence from genome comparisons is revealing that reticulate evolution, including introgression across distantly related species, is far more common in animals than previously thought (Mallet et al., 2016) with implications for gene-tree – species-tree discordance (Edwards et al., 2016). This directs attention to the potential for differing extents of introgression within and outside rearranged regions of the genome (Noor and Bennett, 2009; Crawford et al., 2015).

Here, we revisit the early history of thinking about how chromosomal rearrangements (CRs) affect population and speciation processes. We then highlight a case study that emphasizes how a combined knowledge of genome architecture and genomic sequence divergence is important for understanding the history of CRs in association with speciation and being able to assess whether gene flow is reduced in rearranged regions. Finally, we return to broader themes, considering what combinations of evidence and theory are necessary to gain a holistic understanding of how chromosome change can promote incipient divergence and ultimately translate into species diversification.

Chromosome Change, Population Processes, and Speciation – A Potted History

Observations on differences in chromosome number and form were some of the earliest data available on genetic differences among species. Inevitably, this led to consideration of whether and how such large-scale restructuring of the genome could cause reproductive isolation – speciation – as well as the role of CRs in adaptive evolution within species (Sturtevant, 1938; Dobzhansky, 1950; Stebbins, 1950; Grant, 1964; White, 1973).

The initial focus on adaptive evolution of CRs was largely for paracentric inversions, and their role in recombination suppression and thus accumulation of linked adaptive genes (Dobzhansky, 1950). More broadly, consideration of how multiple CRs (e.g., reciprocal translocations) could lead to long chains of chromosomes with no recombination lead to concepts of "genetic systems" and their role in maintaining heterozygosity (Darlington, 1958; James, 1982). This thread connecting chromosome organization with adaptive evolution continues today with the proposal that the recombination suppression associated with CRs can promote local adaptation and the accumulation of genetic incompatibilities between species (Navarro and Barton, 2003; Kirkpatrick and Barton, 2006; reviewed in Faria and Navarro, 2010; Ortiz-Barrientos et al., 2016). In one powerful example, Shaw et al. (1986) found that shifts in recombination positions in chromosome heterozygotes of *Caledia* grasshoppers was associated with hybrid breakdown, and more so than genetic distance *per se*. Association between range size and rate of inversions in birds also support these models, albeit indirectly (Hooper and Price, 2015).

As evidence of marked differences in chromosome organization among species continued to accumulate, various concepts of chromosomal speciation developed (reviewed by White, 1978; King, 1993). For the most part, these focused on types of CRs that potentially reduce fertility of heterozygotes – "sterility models" – because of disruptions of segregation, or meiotic silencing of unsynapsed chromosomes (MSUC) during meiosis (Garagna et al., 2014). The obvious challenge is to explain how a new mutation that reduces the fitness of its heterozygous carrier can survive selection against it, to establish within a local population. Stimulated by the observation that such changes are often seen in taxa that form small isolated populations (e.g., Bush et al., 1977), various models based on strong genetic drift or founder events followed, some analogous to Wright's Shifting Balance Theory of alternating drift and adaptive evolution in metapopulations (Wright, 1982). Such models were immediately controversial, especially when they invoked variants of sympatric speciation (Key, 1968; Futuyma and Mayer, 1980) or rapid fixation in founder populations (Templeton, 1981). This led to strong skepticism of the view that individual chromosome changes, though reduced fertility, could be a primary and common driver of speciation (Walsh, 1982; Coyne et al., 2000; Coyne and Orr, 2004). Nonetheless, in chromosomally diverse butterflies and *Drosophila*, differences in chromosome number accumulate more rapidly between sympatric than allopatric species and are linked to reinforcing selection for pre-mating isolation (Noor et al., 2001; Lukhtanov et al., 2005; Kandul et al., 2007). An association between speciation and chromosomal evolution was identified in mammals (Bush et al., 1977), and more recently in a diverse genus of lizards, *Sceloporus*, where a phylogenomic analysis revealed higher speciation rates in clades with extensive Robertsonian fusions (Leaché et al., 2016).

Mechanisms that could promote fixation of chromosome changes despite reduced hybrid fertility include: (i) meiotic drive, (ii) establishment of recombination suppression which facilitates adaptive evolution, and simply, (iii) beneficial effects of CRs on gene expression. (i) *Meiotic drive* – (segregation distortion) is a powerful evolutionary force that can drive mutations that otherwise reduce fitness to fixation by biased transmission of chromosomes (reviewed in Lindholm et al., 2016; see also Pardo-Manuel de Villena and Sapienza, 2001). Meiotic drive has been observed to favor Robertsonian fusions (metacentric) over unfused (acrocentric) chromosomes in shrews (Wyttenbach et al., 1998; Fedyk and Chętnicki, 2007) but evidence for this in *Mus* is mixed (Nachman and Searle, 1995; Chmátal et al., 2014). Meiotic drive might also underpin large-scale patterns of chromosome diversity in fish (Yoshida and Kitano, 2012; Molina et al., 2014). Sex chromosomes have been shown to be frequently involved in fusions in fish and amniotes (see Pokorná et al., 2014; Pennell et al., 2015). (ii) *Recombination suppression and adaptation* – selection to reduce negative effects of chromosomal heterozygosity, including shifts in recombination (chiasma) positions, non-homologous pairing and synaptic adjustment. For synapsis to occur during meiosis, chromosomes need to pair to allow crossing over and this process uses the synaptonemal complex. Evidence from a variety of organisms – mice (Johannisson and Winking, 1994; Borodin et al., 2005; Manterola et al., 2009), humans (Guichaoua et al., 1986), chickens (Kaelbling and Fechheimer, 1985) and *Caenorhabditis elegans* (Henzel et al., 2011), highlight that homology of chromosomes is not required to complete this process and synaptic adjustment (reviewed in Zickler and Kleckner, 1999) can overcome issues of non-homology. There is also evidence that this occurs broadly in eutherian mammals between the sex chromosomes (pairing of X and Y), where only a short domain is homologous (pseudo-autosomal region) allowing for non-homologous synapsis (Bergero and Charlesworth, 2009). However, the ability to overcome non-homology depends on a number of factors including the size of the rearrangement, the gene content, the location with respect to centromeres and telomeres and the genetic background (see Torgasheva and Borodin, 2010). This can favor production of balanced gametes for a variety of rearrangements including deletions, insertions, inversions, Robertsonian fusions (Kingswood et al., 1994; Vozdova et al., 2014) and duplications (reviewed in Torgasheva and Borodin, 2010). In addition, recombination suppression may drive adaptive evolution by bringing together advantageous gene combinations (Hoffmann and Rieseberg, 2008; see also Navarro and Barton, 2003). Theory on effects of recombination suppression focuses primarily on inversions (e.g., Kirkpatrick and Barton, 2006) but also considers fusions (Guerrero and Kirkpatrick, 2014) and centric shifts, which may occur via pericentric inversion, three break rearrangements or establishment of neocentromeres, and in the vicinity of centromeres involved in fusion/fissions events (Rieseberg, 2001; Navarro and Barton, 2003). Finally, the simplest possibility is (iii) *a beneficial mutation* – a rearrangement could generate a beneficial effect of relocating genes into a different regulatory environment, long referred to as position effects (Muller, 1930). As with most mutations, such changes will most often be deleterious (as in humans – Harewood and Fraser, 2014).

The well-known Bateson Dobzhansky Muller (BDM) model (based on work of Bateson, 1909; Dobzhansky, 1936; Muller, 1942) can operate for CRs as it does for genic mutations, avoiding the hybrid-sterility conundrum. Independent chromosome changes arise within isolates, and proceed to fixation by drift or adaptive evolution, followed, on secondary contact, by reduced fertility of heterozygotes for multiple rearrangements (see Coyne and Orr, 2004). Comparative and experimental data on *Mus* (reviewed in Garagna et al., 2014), *Sorex* shrews (Polyakov et al., 2011; Horn et al., 2012) and *Rhogeessa* bats (Baird et al., 2009), appear to be exemplify the BDM process, where the focus is on systems with multiple chromosomal fusions with one or more common arms in different fusion arrangements, i.e., monobrachial homology (Baker and Bickham, 1986).

Putting aside contention over whether chromosomal speciation is common, empirical systems where closely related species differ by multiple, complex CRs are frequently observed (White, 1973; King, 1993; Coyne and Orr, 2004; Dobigny et al., 2017). However, our current understanding of CRs is largely based on changes that are visible by classical cytology and chromosome banding. With the tools of molecular cytogenetics and high resolution genome sequencing, yet more, often substantial, CRs are being discovered between species thought to have few changes (e.g., human vs. chimpanzee; Prado-Martinez et al., 2013; Farré et al., 2015).

So, how do we revisit these old questions and debates with new theory and empirical evidence? Despite recent advances in chromosomal speciation theory (Kirkpatrick and Barton, 2006; Faria and Navarro, 2010; Kirkpatrick, 2010, 2017; Guerrero et al., 2012b; Guerrero and Kirkpatrick, 2014), more needs to be done to develop inference methods that can exploit genomic comparisons (see Prospectus section). From the empirical perspective, one fruitful approach is to apply genome-scale analyses to systems that exemplify chromosome change among closely related taxa. Sites and Moritz (1987) proposed that models of chromosomal speciation that require strong genetic drift could be tested using simple predictions for reduced genetic polymorphism and elevated divergence, but both the empirical and inference tools available at the time were limiting. This has now changed substantially, with the ability to sequence thousands of loci across populations of any organism and to use coalescent and network methods to infer divergence history (Edwards et al., 2016). The key challenge for recently diverged taxa is to disentangle the effects of retained ancestral polymorphism (incomplete lineage sorting – ILS) from subsequent gene flow. While this remains challenging, the emergence of isolation-with-migration models (Pinho and Hey, 2010) and phylogenetic network methods (Nakhleh, 2013), when combined with genome-scale data, offer some hope. Recent research into the *Anopheles* system has highlighted the value of genomic data in disentangling ILS from introgression (Fontaine et al., 2015; Wen et al., 2016), as has sliding window analysis of genomes in *Xiphophorus* fishes (Cui et al., 2013). Whole genomes allow for a suite of new analyses to identify introgression (e.g., using the ABBA-BABA discordance test; Green et al., 2010; Durand et al., 2011; Martin et al., 2015; Nater et al., 2015), but currently there are still limitations based on genome sequencing and alignments, where phasing errors can lead to over-estimation of recombination or mutations (e.g., Qi et al., 2014; Rasmussen et al., 2014). Further, it may be that comparative genome screening alone will not be sufficient to resolve different effects of CRs on divergence (e.g., Suh, 2016; but see Prospectus).

Inferring Divergence Histories of Candidates for Chromosomal Speciation

To resolve whether CRs initiate divergence or follow genic speciation, we need to focus on recently diverged taxa (Coyne and Orr, 2004). We need to identify organisms that can help address questions in chromosomal speciation and apply integrative tools to them. In particular, cytogenetic and molecular data can be combined to infer the sequence and timing of CRs in systems with complex chromosome change (Faria and Navarro, 2010). This is especially important to interpret signatures of genetic divergence associated with these CRs (Noor and Bennett, 2009). It should be reiterated that it remains a formidable challenge to resolve relationships and reticulations among recently separated species (e.g., Leaché et al., 2016).

Several recent comparative studies of species with high quality reference genomes have used extensive resequencing to resolve divergence histories and contrast levels of introgression among recently separated taxa that differ by chromosomal inversions (e.g., Primates – Carbone et al., 2014, *Drosophila* – Kulathinal et al., 2009; McGaugh and Noor, 2012; Lohse et al., 2015; *Anopheles* – Wen et al., 2016). There has, however, been mixed support for recombination suppression models (see Faria and Navarro, 2010). By contrast to chromosomal inversions, there have been few genome-scale analyses of closely related taxa with complex Robertsonian fusions. In the Robertsonian fusion races of *Mus*, increased genetic divergence has been observed at microsatellite loci near the centromeres of fused chromosomes (Franchini et al., 2010; Förster et al., 2016) and simulations of recombination suppression versus hybrid breakdown reveal that hybrid breakdown alone could explain the patterns in *Mus* from Italy (Giménez et al., 2013). Like *Mus*, reduced gene flow (higher divergence) is evident within CRs in *Sorex* shrews (Basset et al., 2006; Yannic et al., 2009).

While analyses of model systems, such as the above, have provided important insights into causes and consequences of CRs, it remains important to extend analyses of effects of chromosome change to systems with distinct genomic features and population structures (Payseur and Rieseberg, 2016). In the following, we present one such example and then conclude with a prospectus for how to advance this and other non-model systems. With this and other such systems, we hope to obtain a greater insight into the processes driving variation in genomic architecture, that lead to divergence and speciation.

CASE STUDY: *Petrogale* ROCK-WALLABIES

The rock-wallaby (*Petrogale*) system has been considered a classical model for chromosomal speciation due to extensive chromosome repatterning, combined with their habitat specialization (rocky environments) which causes populations to be isolated and small (King, 1993). Rock-wallabies are medium sized marsupials (1–12 kg) that inhabit complex rocky areas distributed across continental Australia and some offshore islands, (Eldridge, 2008). A strong propensity for isolation among disjunct rocky habitats (e.g., Pope et al., 1996; Hazlitt et al., 2006) is thought to increase their rate of speciation and contribute to the fixation of novel CRs (Eldridge and Close, 1993). *Petrogale* includes 17 recognized species corresponding to 23 chromosomal taxa (Supplementary Table 1; **Figure 1**). This is the most chromosomally diverse genus of marsupials, which in general have a conserved karyotype across all five Australasian and American super-families ($2n = 14$; Rofe and Hayman, 1985; Hayman, 1990; see O'Neill et al., 1999; Graves and Renfree, 2013). Macropodids (kangaroos and wallabies) show variable karyotypes (see O'Neill et al., 1999), but the ancestral macropodid $2n = 22$ karyotype is only found in *Petrogale* (*P. lateralis*, *P. persephone*, *P. rothschildi*, and *P. xanthopus*; Eldridge et al., 1992a) and *Thylogale* (Pademelons). The $2n = 22$ macropodid ancestral karyotype is itself derived from the widespread $2n = 14$ marsupial karyotype by a series of fissions (Rofe, 1979; Hayman, 1990). CRs are extensive across *Petrogale*, and range from simple to complex. A majority of the rearrangements are Robertsonian fusions, but there is also cytogenetic evidence for inversions and centric shifts (centromeric transpositions) (Eldridge et al., 1989, 1990, 1991, 1992b; Eldridge and Close, 1992, 1993). In addition, the X chromosome is frequently variable in morphology among taxa and also sometimes within taxa (Eldridge and Close, 1997). The highest chromosomal diversity occurs in groups that are sympatric or parapatric (*brachyotis* and *penicillata* groups; Eldridge et al., 1992b; Eldridge and Close, 1993; **Figure 1**), and these are also the most speciose. This pattern matches the predictions of chromosomal speciation models (see Faria and Navarro, 2010) and the recombination suppression model. This could reflect yet another scenario where fixation rate of chromosomal rearrangements correlates with parapatry and sympatry – suggesting adaptation and divergent selection could be a dominant process driving fixation (e.g., Hooper and Price, 2015).

The *brachyotis* group of rock-wallabies includes four species and five sub-species distributed across northwestern Australia which have the most complex rearrangements found in *Petrogale* (Supplementary Table 2) and large amounts of centromeric constitutive heterochromatin not present in other *Petrogale* (Maynes, 1989; Sharman et al., 1990; Eldridge et al., 1992b; Eldridge and Close, 1993). Within the *lateralis* group, there are two chromosomal races and three sub-species (Eldridge et al., 1991; **Figure 1**). These races/sub-species are recently diverged and are distinguished by single autosomal rearrangements or fusions (Eldridge and Close, 1993, 1997; **Figure 1**). However, the most interesting group are the recently diverged (∼0.5–2.7 mya; Potter et al., 2012a) Queensland *penicillata* group taxa. Six parapatric species display extensive variation in karyotypes ranging from simple to complex – including fusions, inversions and centric shifts (see Eldridge and Close, 1993). Early research was driven by cytogenetic analyses (reviewed in Eldridge and Close, 1997) and captive breeding experiments that showed evidence of reproductive isolation including infertile male hybrids and reduced fertility of female hybrids (Eldridge and Close, 1992). This resulted in the description of three new species (Eldridge and Close, 1992) and a focus on the role of chromosomal variation in speciation. Meiotic irregularities, including problems with more extensive rearrangements and X-autosome associations have been reported (Close et al., 1996) – patterns also seen in model systems (e.g., *Mus*). Recent genetic analysis of the *penicillata* group using microsatellites and mitochondrial DNA (mtDNA) found extensive sharing of alleles between some of the most chromosomally divergent species (Potter et al., 2015). This could be a consequence of introgression or ILS. Further analysis of nuclear markers across the genome is required to assess the genomic divergence between these species and assess if speciation with gene flow is occurring between these taxa, or if more complex interactions between genomic architecture and genic divergence is at play. The

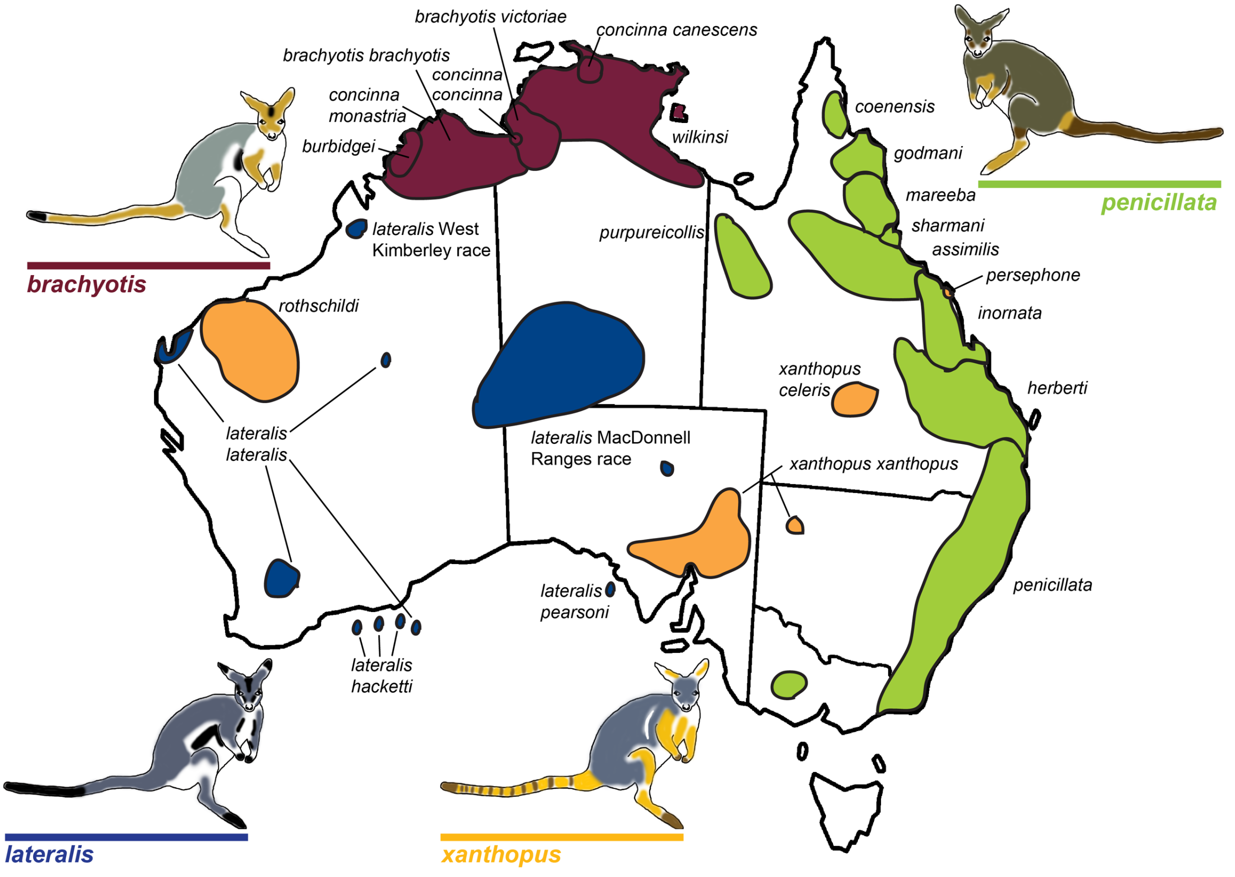


FIGURE 1 | Map of rock-wallaby (*Petrogale*) taxa distributions across Australia. Map modified from Eldridge and Close (1993). Taxa are colored in accordance with their chromosomal groupings: the *brachyotis* group = red; the *xanthopus* group = yellow; the *lateralis* group = blue; and the *penicillata* group = green.

characteristics of this genus, specifically their rapid radiation and extensive chromosome variation, make them a valuable model for understanding chromosome evolution and speciation.

Phylogenetic analyses of the rock-wallabies have not previously included representatives of all 23 chromosomal taxa, nor have they been able to resolve phylogenetic relationships, particularly among the more recently evolved species within the *penicillata* group (Campeau-Péloquin et al., 2001; Potter et al., 2012a). This, in addition to evident homoplasy of rearrangements (see Eldridge and Close, 1993), has precluded tracing the evolution of chromosomal changes. The phylogenetic relationship of *P. xanthopus* and *P. purpureicollis* has also been difficult to resolve (see Eldridge et al., 1991; Eldridge and Close, 1993; Potter et al., 2012a), which has hindered interpreting chromosome evolution as these taxa retain the ancestral chromosome number.

Here, we report results from targeted capture for ~2000 exons from two individuals per taxon to resolve the relationships across the genus (all supplementary material and methods are outlined in Supplementary Datasheet 1). These data allow us to understand the evolution of chromosomes in this group. In particular, they provide insight into phylogenetic and sequence divergence signals of discordance across the X, rearranged and non-rearranged chromosome arms that could reflect effects of CRs on gene flow (see Supplementary Table 3). We focus on sets of concatenated loci, rather than individual gene trees as individual exons have low phylogenetic resolution at this scale. While it is desirable to use multispecies coalescent approaches (e.g., *BEAST and ASTRAL), such programs are confounded by introgression across non-sister data and are therefore unsuitable for this system (see Solis-Lemus et al., 2016). Hence, we explore multispecies coalescent network approaches that allow for introgression (see below; reviewed in Nakhleh, 2013; Edwards et al., 2016). We expect to find discordant phylogenies and divergence levels between these categories of loci, particularly for the recent radiation of Queensland taxa.

The phylogenetic relationships amongst taxa using the entire dataset of 1961 exons and ~1 million bp firmly resolves, for the first time, relationships within *Petrogale* (**Figure 2**). The *brachyotis* group with the most extensive chromosomal rearrangements is also phylogenetically basal, which is consistent with previous genetic data (Campeau-Péloquin et al., 2001; Potter et al., 2012a). The *xanthopus* chromosomal group is paraphyletic. *P. rothschildi* forms the sister taxon to the

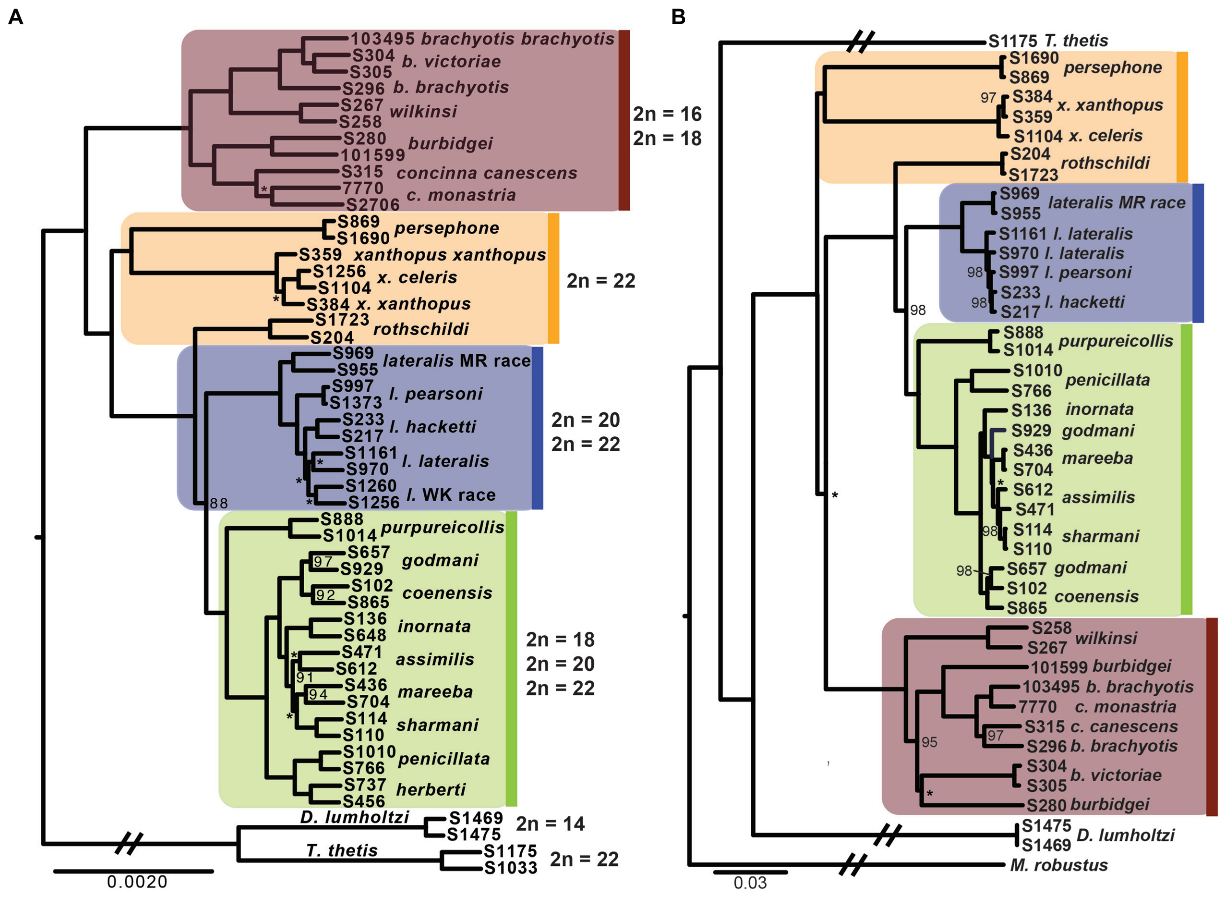


FIGURE 2 | (A) Phylogenetic relationships of rock-wallabies (*Petrogale*) based on a maximum likelihood analysis of concatenated nuclear data (1961 loci). Bootstrap support < 100% is outlined on the nodes, * = < 50%; nodes with neither have support of 100%. Four chromosomal groups are highlighted on the phylogeny: the *brachyotis* group = red; the *xanthopus* group = yellow; the *lateralis* group = blue; and the *penicillata* group = green. Karyotype variation (2*n*) for each of the four chromosomal groups is highlighted. *Dendrolagus lumholtzi* (tree kangaroo) and *Thylogale thetis* (pademelon) are used as outgroups. **(B)** A maximum likelihood mitochondrial phylogeny of *Petrogale* based on all mitochondrial coding genes (12 loci). Chromosomal groups are highlighted to match **(A)**, as is bootstrap support.

lateralis and *penicillata* groups, where as *P. persephone* and *P. xanthopus* form a well supported monophyletic group. There is deep divergence between all three taxa that retain the ancestral karyoptype. The cytogenetically conservative *lateralis* group has similar branch lengths among taxa to those among species within the more chromosomally diverse *penicillata* group. Despite recent speciation, each taxon within the *lateralis* and *penicillata* groups is monophyletic, albeit with lower support for *P. lateralis lateralis*, *P. l.* West Kimberley race and *P. assimilis*. Relationships amongst some of the most closely related taxa are also not strongly resolved, but we note that monophyly of *P. mareeba* and *P. sharmani* is consistent with their chromosomal evolutionary history, since both share a derived fusion between chromosomes 5 and 10. Although the phylogenetic position of *P. purpureicollis* has been previously unresolved (Sharman et al., 1990; Eldridge et al., 1991; Campeau-Péloquin et al., 2001; Potter et al., 2012a), these new data strongly resolve it as sister to the *penicillata* group. The sub-species of both *P. brachyotis* and *P. xanthopus* (with no known chromosomal differences) did not form monophyletic lineages, suggesting recent divergence or some nuclear gene flow.

Given a well-resolved phylogeny, we can now investigate the history of chromosome change in the genus. Using parsimony mapping of CRs, as identified by G-banding (Eldridge and Close, 1993) we were able to resolve ancestral nodes where rearrangements occurred, in particular a single origin of CR 7a (a = acrocentric). However, for some chromosomes, we could not distinguish between different hypotheses (see **Figure 3**; Supplementary Table 2). Apparent multiple independent origins of the 3a, 4a, 4sm (sm = submetacentric) and 5i (i = inversion) rearrangements suggest there could be regions of the genome susceptible to rearrangement processes ("hotspots"), which have also been implicated in chromosome change in other macropodids (Bulazel et al., 2007). This highlights the potential for convergent evolution of rearrangements, including chromosomal fusions (e.g., 6 and 10 fusion), inversions and centromere shifts. The alternate hypotheses are multiple reversals to an ancestral chromosome morphology, or the introgression of chromosomes between taxa, but further analysis (e.g.,

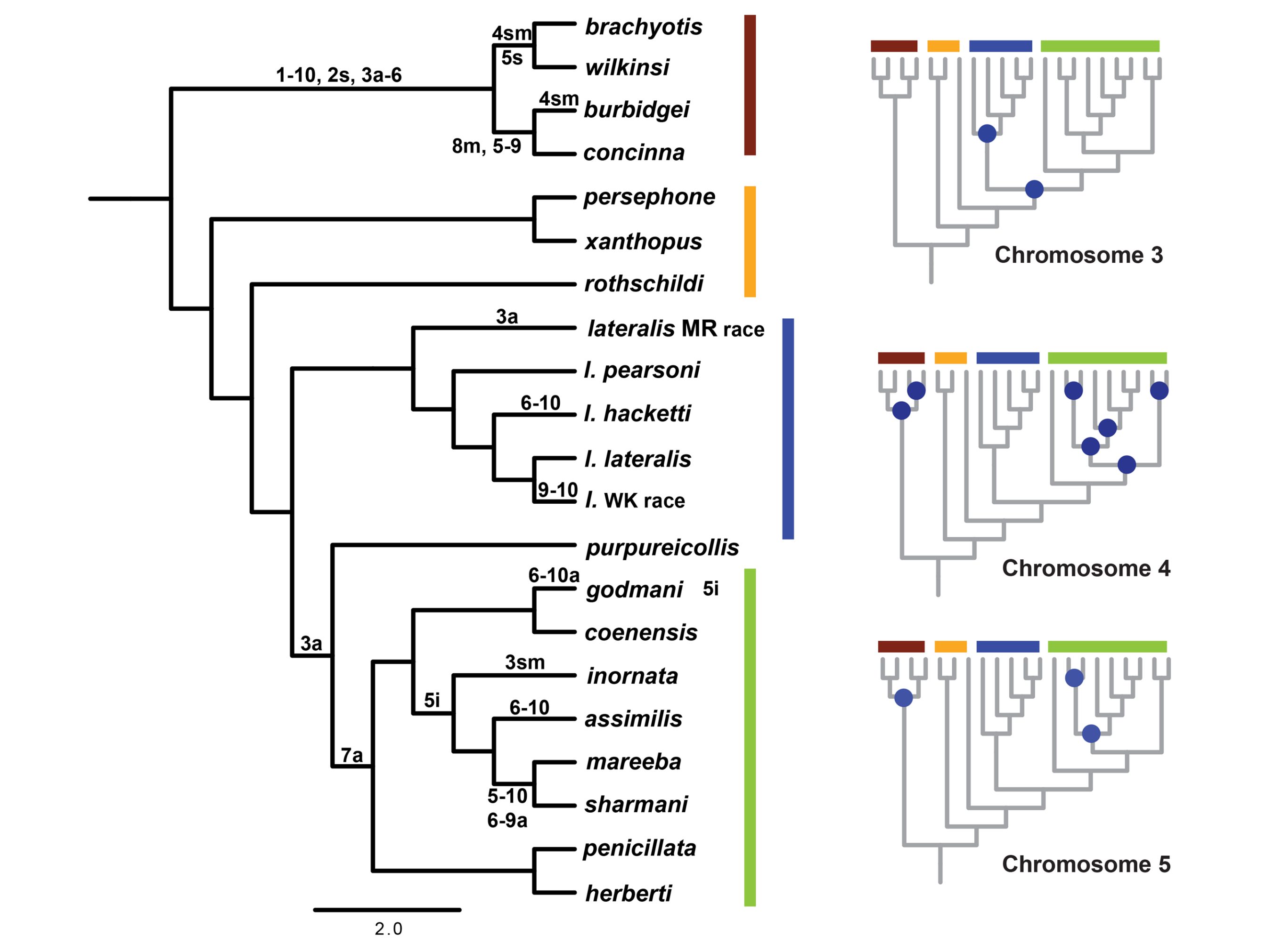


FIGURE 3 | Reconstruction of chromosomal rearrangements based on parsimony analysis for *Petrogale* using only known chromosomal karyotypes (e.g., no sub-species for *P. brachyotis*, *P. concinna* or *P. xanthopus*). Reconstructions of ancestral karyotypes are highlighted on the main phylogeny and those that could not be resolved for nodes in the phylogeny are indicated in blue for chromosomes 3, 4, and 5. See Supplementary Table 2 for character state matrix. Chromosomal groups are again outlined in color: *brachyotis* = red; *xanthopus* = yellow; *lateralis* = blue; and *penicillata* = green. Chromosomal rearrangements include: centric shifts, a = acrocentric, m = metacentric, sm = submetacentric; inversions (i); and fusions between two chromosomes (–).

sequencing of breakpoints) is required to distinguish between them. In addition to parallel evolution, the same chromosomes are involved in fusion events in different taxa (e.g., 5, 6, 9, and 10). Cell culture experiments using mitomycin C to induce centric fusions showed that chromosome 10 fused most frequently (Eldridge and Johnston, 1993). Despite all chromosomes being involved in fusions in this experiment, the higher frequency of chromosome 10 fusions *in vitro* matches the larger proportion of chromosome 10 being involved in fusions in the wild in *Petrogale* (five out of eight fusions). This together with higher frequencies of breakpoints from gamma radiation in chromosomes 5, 6 and 10 (Eldridge and Johnston, 1993) further support the notion of a nonrandom process of CR.

Next, we partitioned the sequenced exons into autosomal-non-rearranged ($N = 140$ exons; 75,296 bp), autosomal-rearranged ($N = 160$ exons; 36,168 bp), and the X chromosome ($N = 21$ exons; 8,951 bp). This approach was motivated by the expectation that autosomes will have less phylogenetic signal than the X because of their higher gene flow rates and larger N_e (see Supplementary Table 3 for mapped loci). Mean divergence between the four chromosomal groups varies across the X chromosome, rearranged and non-rearranged autosomes (**Figure 4**). When accounting for differences in sequence length, we find that the X has reduced diversity compared with the autosomes, although only slightly compared to the rearranged autosomes. In other mammalian systems (e.g., apes – Nam et al., 2014) several selective sweeps have created large regions of low diversity on the X chromosome. Assessment of more loci along the X is necessary to explore if the lower diversity on the X within *Petrogale* is associated with selection. The effects of small N_e on the X will need to be assessed to distinguish between selective sweeps and neutral models of evolution. However, given that sex chromosomes contribute disproportionately to post-zygotic isolation in many taxa (e.g., *Drosophila*, Presgraves, 2008;

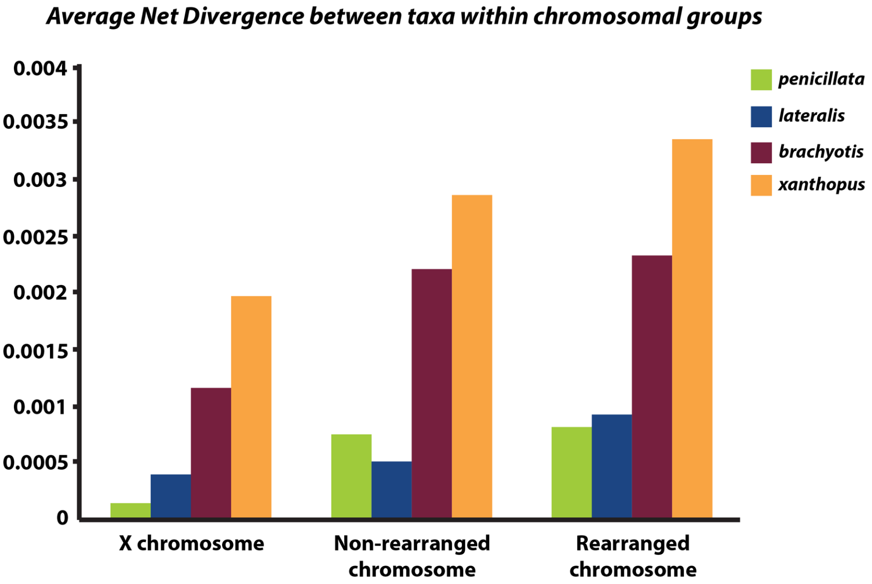


FIGURE 4 | Graph of average net divergence between taxa within each chromosomal group: *brachyotis*, *lateralis*, *penicillata*, and *xanthopus*. Divergences are estimated for loci on the X chromosome, non-rearranged chromosomes (2,4,7,8) and rearranged chromosomes (5,6,9,10).

Flycatchers, Saether et al., 2007), and evidence of the *Petrogale* system conforming to Haldane's rule, we would expect more loci on the X to conform to the true species phylogeny than autosomal loci (as argued by Fontaine et al., 2015).

Rearranged and non-rearranged chromosomes show similar levels of divergence in the *brachyotis* and *penicillata* groups, which are among the most chromosomally diverse *Petrogale* (**Figure 5**). Conversely, rearranged chromosomes showed the greatest increase in divergence relative to non-rearranged chromosomes in the *lateralis* group, which has relatively few rearrangements. These observations do not tend to support the hypothesis that CRs are a primary cause of speciation. Instead, the CRs may have fixed when species were already isolated. Alternatively, the existing data may be inadequate to distinguish between these hypotheses. Further work is needed.

We expect clearer phylogenetic resolution in rearranged than in non-rearranged regions of the genome, particularly within the *penicillata* group (**Figure 5**). Neither autosomal phylogenetic reconstructions are able to resolve the relationships of the *lateralis* and *penicillata* groups. The phylogeny based on rearranged regions does however, separate *P. herberti* and *P. penicillata* from the remaining taxa, as well as *P. coenensis* and *P. godmani*, compared to the non-rearranged chromosomes. Both autosomal phylogenies resolve the *brachyotis* and *xanthopus* groups, but the rearranged phylogeny places the *xanthopus* group as basal instead of the *brachyotis* group. By contrast, the X loci resolve nodes deeper in the tree but appears to lack enough information to resolve all of the internal relationships of the *lateralis* and *penicillata* groups. The X however, does generally group individuals of a taxon together, unlike the autosomes (**Figure 2**).

We then asked how often the individuals sampled from each species formed a monophyletic group for the different subsets of exons. If rearrangements result in reduced introgression, we expect to see higher concordance in concatenated loci from rearranged than non-rearranged chromosome arms. Further, as the X chromosome is frequently found to be resistant to gene flow, we also expected higher congruence across the X-linked loci. On average, the rearranged chromosomes have greater monophyly of taxa than the non-rearranged for the *penicillata* group, supporting our hypothesis of higher concordance. We do, however, find the opposite pattern for the *lateralis* group (Supplementary Table 4). This may be because the *lateralis* group has fewer rearranged loci. Overall it had lower concordance of monophyletic individuals compared to the *penicillata* group. The X chromosome had the greatest average monophyly. This may result from a smaller N_e of the X, faster divergence of the X (e.g., Charlesworth et al., 1987), or greater divergent selection on the X. The evolution of the sex chromosomes needs further investigation.

Analysis of all mitochondrial coding genes (12 genes; 11,373 bp), albeit still a single linkage group, reveals some strong conflicts between mitochondrial and nuclear evolutionary history (**Figure 2**). Previously, it has been highlighted that introgression, retained ancestral polymorphism (or ILS) has resulted in paraphyletic species complexes within both the *brachyotis* group (see Potter et al., 2012a,b) and the *penicillata* group (Briscoe et al., 1982; Bee and Close, 1993; Potter et al., 2015). This is the first analysis using all coding genes across the mitochondrial genome, providing the most phylogenetic information and highlight discrepancies to the nuclear phylogeny, in particular – the placement of *P. persephone* and *P. xanthopus* as basal branches; the paraphyly of *brachyotis* group taxa; and lack of monophyly for *P. assimilis*, *P. coenensis* and *P. godmani* within the *penicillata* group. These inconsistencies further highlight areas of potential

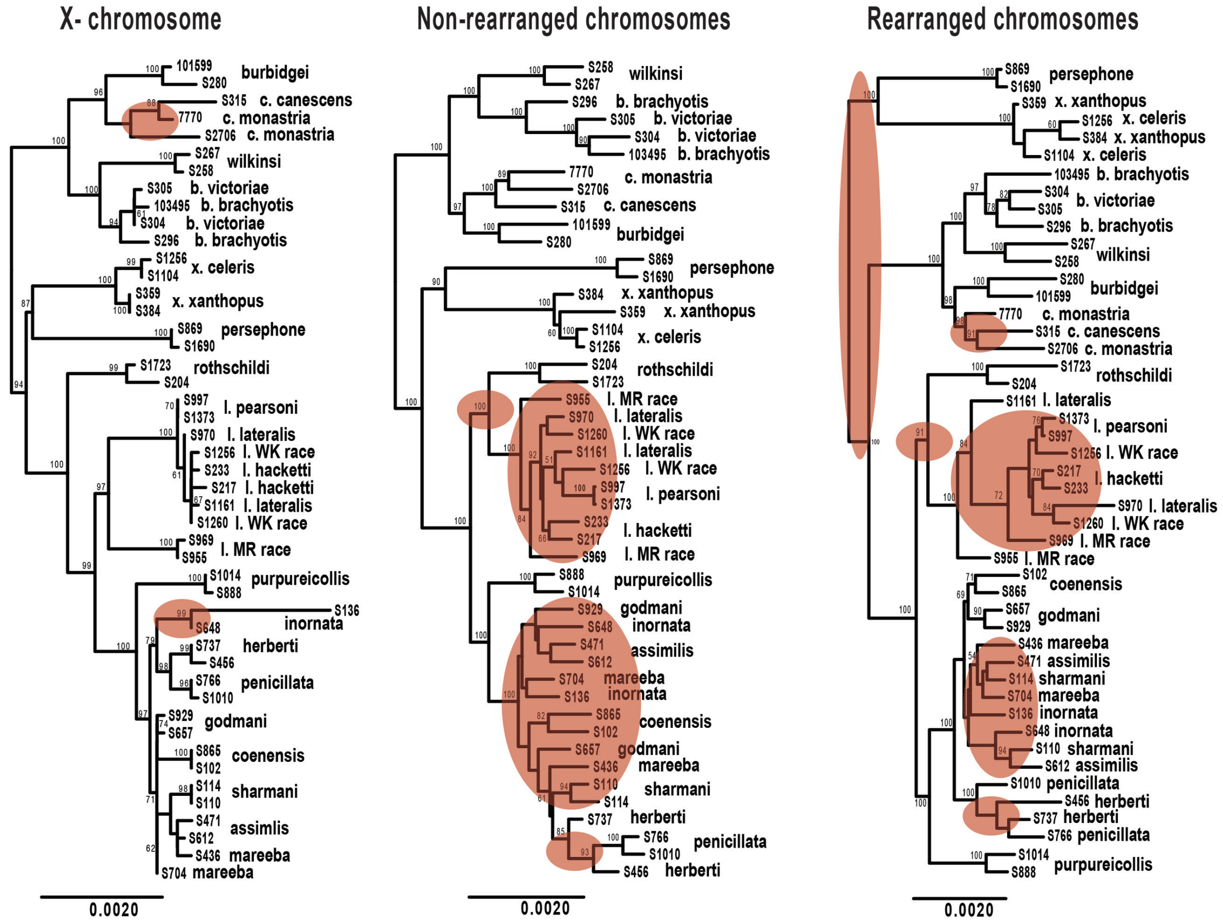


FIGURE 5 | Maximum likelihood phylogenies of *Petrogale* taxa based on loci from the X chromosome, rearranged autosomal chromosomes (5,6,9,10) and non-rearranged autosomal chromosomes (2,4,7,8). Bootstrap support is highlighted on the nodes of the tree. Incongruence in the phylogenetic reconstruction to the concatenated nuclear phylogeny (**Figure 2A**) is highlighted in red on each tree.

reticulation in the history of *Petrogale*; in particular, the potential for repeated episodes of introgression in the history of the genus.

We then explored patterns of reticulation among chromosomal groups. We based these analyses on concatenated nuclear loci, effectively ignoring coalescent variance in gene trees. We first used an exploratory statistical approach (Neighbor-Net in Splits tree – Bryant and Moulton, 2002; Huson and Bryant, 2006, based on average distances). The results suggested reticulation amongst *P. coenensis* and *P. godmani*, between *P. assimilis*, *P. mareeba* and *P. sharmani,* as well as between *P. lateralis lateralis* and *P. lateralis* West Kimberley race (**Figure 6**). The *brachyotis* group does not indicate any strong evidence of reticulation between species. The *penicillata* group results match previous results of microsatellites and mtDNA, which inferred gene flow between these taxa, even those with complex CRs (Potter et al., 2015).

We next used an approach based on the multispecies network coalescent (PhyloNet – Than et al., 2008; Yu et al., 2013, 2014). We find evidence of reticulation for all three chromosomal groups (**Figure 6**; Supplementary Table 5). Based on our analysis of up to three reticulation events, the results support 2–3 reticulation nodes in the *brachyotis* group, which suggests historical introgression may explain the discordance between mtDNA and nuclear loci. This included reticulation at nodes of ancestral branches in the *brachyotis* group. For both the *lateralis* and *penicillata* comparisons there was support for a single reticulation node. Within the *lateralis* group, the reticulation node involved *P. lateralis* West Kimberley race. Reticulation could reflect ILS between *P. l. lateralis* and *P. l. hacketti* or introgression with *P. l. lateralis*. In the *penicillata* group, the analyses were less concordant, and the reticulation node included *P. assimilis* and *P. sharmani* for the independent analysis. Both network analyses reflect greater reticulation across all three species and include introgression with *P. mareeba* and *P. assimilis*, as well as ILS amongst the three species. In all three chromosomal group comparisons, reticulation is evident between species with CRs (fusions and centric shifts). Further work is necessary to disentangle the effects of ILS from introgression, which will require further sampling in rearranged vs. non-rearranged regions of chromosomes.

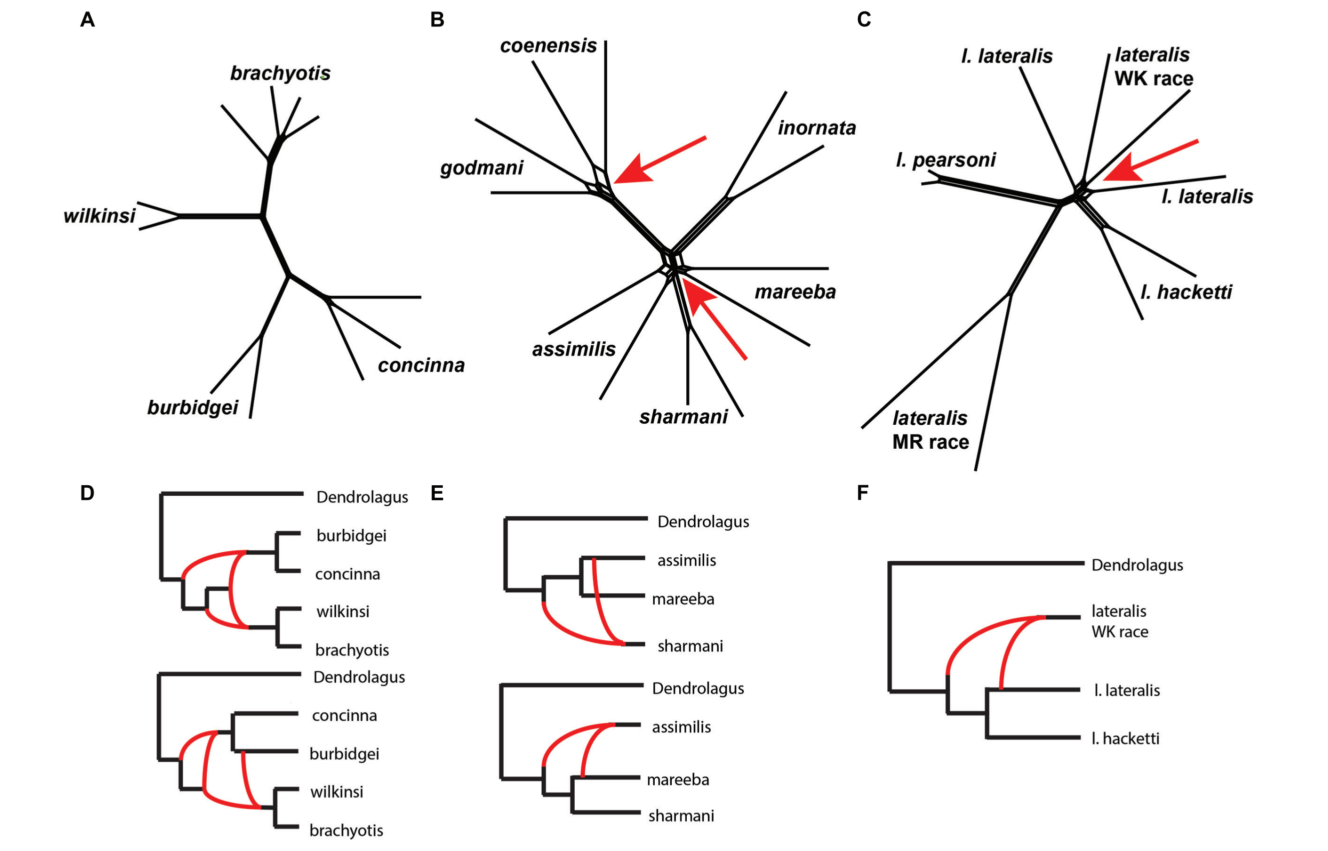


FIGURE 6 | (A–C) A phylogenetic network analysis of the chromosomal groups **(A)** *brachyotis*, **(B)** *lateralis*, and **(C)** *penicillata* estimated using a distance based approach in SplitsTree (Neighbor-Net). The red arrows highlight regions on the network where reticulation is inferred. **(D–F)** Model based analysis of reticulate evolutionary history based on analysis of 0–3 reticulations. The lowest log likelihood results are shown for each **(D)** *brachyotis*, **(E)** *lateralis*, and **(F)** *penicillata* chromosomal groups using PhyloNet. Analysis was performed on a single individual for each taxon and two replicate analyses, each including one of the two independent samples per taxon. **(D)** The *brachyotis* group supported 2–3 reticulations and highlight reticulation involving ancestors in the *brachyotis* group. **(E)** The *penicillata* group analysis include a three species complex (*P. assimilis*, *P. mareeba*, and *P. sharmani*) and support a single reticulation model but alternate topologies and individuals involved in reticulation based on the individuals used in the analysis. **(F)** The *lateralis* group analyses included three taxa (*P. l. lateralis*, *P. l. hacketti* and *P. l.* West Kimberley race). The results were congruent in identifying a single reticulation model, which involved the *P. lateralis* West Kimberley race.

Understanding the genome, the physical position of loci and how they interact is crucial in interpreting the evolutionary history of organisms. Our results highlight that if certain loci taken alone without any context of chromosome structure can yield completely different results and a misunderstanding of the mechanisms involved in reproductive isolation. We are still in the early stages of understanding the physical location of loci in this non-model system and as further work allows mapping of loci to chromosomes and regions of rearrangements, we will be better able to test for recombination suppression and reduced gene flow in these regions compared to the non-rearranged chromosomes. With the addition of X chromosome loci, we may be able to further understand the genetic variation on the X which itself has a high rate of intrachromosomal rearrangement in this system. The combined effects of faster X divergence and reduced recombination may affect speciation in this genus. Below, we outline the necessary steps to progress this non-model system – a framework we believe useful for any non-model speciation research.

PROSPECTUS

The case study above, for a fascinating yet "non-model" system, indicates both the promise and the challenges of arriving at a holistic understanding of how CRs affect divergence and adaptation. Hybrid infertility and recombination suppression (adaptive) models have largely been treated as exclusive, but in systems with complex rearrangements, both could well be in play (Faria and Navarro, 2010; Garagna et al., 2014). Further, such hypotheses should not be treated as exclusive of accumulation of genic incompatibilities, such as large X-effects (Presgraves, 2008), effects of recessive X-linked incompatibility alleles on sterility of the heterogametic sex (Haldane's rule; Haldane, 1922; Turelli, 1998; Presgraves, 2008, 2010) or cytonuclear incompatibilities (Turelli and Moyle, 2007). In addition to new evidence across diverse systems, we need further development of theory that incorporates these co-occurring processes (Feder and Nosil, 2009; Faria and Navarro, 2010). We should also consider extended models with meiotic drive, genomic conflict

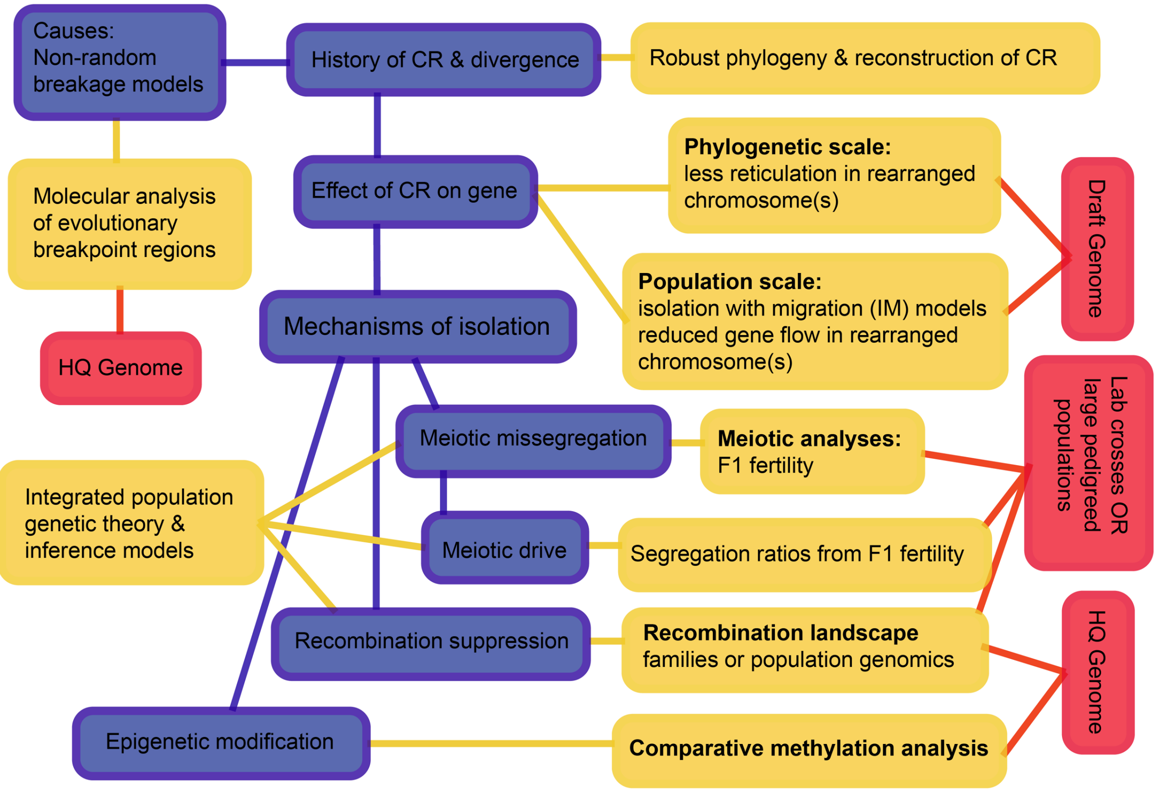

Causes:
Non-random breakage models
History of CR & divergence
Robust phylogeny & reconstruction of CR
Molecular analysis of evolutionary breakpoint regions
Effect of CR on gene
Phylogenetic scale: less reticulation in rearranged chromosome(s)
Draft Genome
Population scale: isolation with migration (IM) models reduced gene flow in rearranged chromosome(s)
HQ Genome
Mechanisms of isolation
Meiotic missegregation
Meiotic analyses: F1 fertility
Lab crosses OR large pedigreed populations
Integrated population genetic theory & inference models
Meiotic drive
Segregation ratios from F1 fertility
Recombination suppression
Recombination landscape families or population genomics
HQ Genome
Epigenetic modification
Comparative methylation analysis

FIGURE 7 | A schematic overview of an integrative approach to increase understanding how chromosomal rearrangements create divergence and speciation. The hierarchy outlined progresses from simple to complex and may not be feasible for all non-model systems, but outlines the ultimate goals. Blue boxes highlight the processes, yellow boxes highlight the analyses and red boxes highlight the data needed to achieve a holistic understanding of chromosomal speciation. CR, chromosome rearrangement; HQ, high quality.

and disruption of epigenetic programming (Brown and O'Neill, 2010).

Considering all these interacting processes, we suggest a hierarchy of questions and associated requirements that progress from simple to challenging (**Figure 7**). The first question is the history of speciation and CRs, as we have now begun to resolve for the rock-wallabies. We move on to testing the effects of CRs on the suppression of gene flow, and then to testing mechanisms that may have suppressed gene flow. The scale and diversity of data needed expands as we move from pattern to process. Of particular note is the need to develop new theory and inference methods that can distinguish the different processes by integrating genomic data, recombination landscapes, meiotic irregularities, segregation patterns, and epigenetics. Such methods involve complicated parameter estimation and may need to rely on approaches such as approximate Bayesian computation (ABC) to compare competing models (see Beaumont, 2010). Those models, in turn, will likely employ multispecies coalescence that includes speciation and demographic history, as well as the evolution of CRs (e.g., coalescent models for CRs – Guerrero et al., 2012a,b; Ayala et al., 2013; Peischl et al., 2013). The quality of genome assemblies needed increases as we move from pattern to process. This is crucial to remove errors that can influence inference of rearrangements as well as recombination rate changes.

In the following, we explore three key issues that have emerged: (i) what do we know of differential origins of new (or recurrent) CRs; (ii) how do CRs interact with the epigenetic programming of genes and chromosome segments; and (iii) what new theory and inference methods do we need to exploit comparative genomic data in the context of individual and combined effects of CRs.

Origins of CRs

Recent reviews highlight that the breakpoints of CRs are associated with repetitive elements, and are influenced by functional constrains and meiotic recombination (Farré et al., 2015). These features have been discovered from whole genome comparisons of distantly and closely related species. Repetitive sequences, such as segmental duplications and transposable elements, appear to provide the substrates for non-allelic homologous recombination (Bailey et al., 2004; Cordaux and Batzer, 2009), resulting in CRs such as inversions. These breakpoint regions are also common in gene-dense regions of the genome, with the breaks occurring predominantly in intergenic regions where they are less likely to silence a gene (Lemaitre et al., 2009). Breakpoint regions are enriched with genes involved in adaptive processes, such as immune response genes, and CRs causing changes to expression of these genes or otherwise rendering a gene non-functional could provide a selective advantage to result in the fixation of this rearrangement (Larkin et al., 2009; Ullastres et al., 2014). Furthermore, gene-rich regions have been shown to be more actively transcribed, possessing epigenetic characteristics of open, and therefore accessible, chromatin (Lemaitre et al., 2009; Capilla et al., 2016). This makes sense in light of the observation that meiotic double strand breaks required for CRs typically occur in transcriptionally active, open chromatin regions of the genome (Smagulova et al., 2011).

Rearrangements involving centromeres are common in many systems, including *Petrogale*. Similar to the breakpoint regions described above, centromeres contain highly repetitive sequences. Robertsonian fusions may be formed from illegitimate recombination occurring among these highly repetitive sequences (Slijepcevic, 1998; Garagna et al., 2001; Ruiz-Herrera et al., 2006). Nuclear architecture also appears to play a role in the formation of Robertsonian fusions. In mice, the pericentric regions of telocentric chromosomes converge during the leptotene stage of prophase I in meiosis, placing pericentromeric regions of different chromosomes in close proximity (Garagna et al., 2014). Prophase I is also the time when DNA is damaged and the cell is repairing the double-stranded breaks by homologous recombination for synapsis of homologous chromosomes (Neale and Keeney, 2006). The combination of these two factors lends this phase of meiosis to being a time of potential CR (Garagna et al., 2014). More generally, the location of chromosomes within the nucleus should be considered when studying CRs as it could help to understand why some chromosomes or chromosome arms more commonly involved in CRs than other chromosomes (Wesche and Robinson, 2012).

In some cases, the CR involving the centromere is a centric shift. This could occur via a pericentric inversion. Alternatively, neocentromeres may be established from epigenetic changes to repetitive sequences located elsewhere on the chromosome and old centromeres decommissioned (O'Neill et al., 2004). Also, a three-break rearrangement that allows the centromere to be excised and then reinserted in a different position further along the chromosome could occur (Eldridge et al., 1988). O'Neill et al. (2004) suggested that heterochromatin stabilizes centromere position and its absence opens the way for neocentromerization. Although cause and consequence remain difficult to distinguish, and further exploration of this hypothesis in *Petrogale* would be of value.

The characterization of breakpoint features and the role of centromeres in CRs has made it evident that sequence content, an open chromatin conformation and chromosome territories within the nucleus in the germline are all important for understanding the origin of CRs and should be considered together (i.e., the "Integrative Breakage Model," Farré et al., 2015). Detailed studies of epigenomic features and chromatin conformation are possible for model species like the mouse (Capilla et al., 2016), testing the Integrative Breakage Model may be more challenging at present for non-model species as a high-quality reference genome is essential for interpreting epigenomic data. However, advances in sequencing technology and the corresponding genome assembly pipelines (e.g., Putnam et al., 2016), make this achievable for non-model species.

Interaction of CRs with Epigenetic Programming

Epigenetic variation has recently been recognized to cause genetic incompatibilities that can lead to reproductive isolation (Brown and O'Neill, 2010; Durand et al., 2012). The only known "speciation gene" in mammals is *Prdm9* in mice (Mihola et al., 2009). It encodes for the meiotic specific-protein responsible for marking the location of recombination hotspots (Grey et al., 2011). Recombination is essential recognition of chromosome homologues during prophase I of meiosis and its disruption results in sterile male hybrids. An interaction between the autosomal *Prdm9* and the X-linked *Meir1* gene contributes to reproductive isolation of *Mus musculus musculus* and *Mus m. domesticus* subspecies (Balcova et al., 2016). Robertsonian fusions alter the epigenetic marks of H3K9me3 and γH2AX (Capilla et al., 2014). The accumulation of γH2AX is associated with the MSUC mechanism, which is similar to meiotic sex chromosome inactivation (MSCI) for silencing the unsynapsed X and Y in males. These mechanisms involve a suite of epigenetic modifications to achieve transcriptional silencing of the sex chromosomes (MSCI) or unsynapsed region (MSUC) (reviewed in Turner, 2007). When CRs reduce recombination, MSUC may therefore silence genes critical to meiosis (Shiu et al., 2001) and cause infertility (Garagna et al., 2014).

The frequent involvement of centromeres in CRs makes it interesting to consider the role of these important structures in speciation. Sequences at centromeres rapidly evolve, differing markedly between closely related species. Chromatin-binding proteins, such as various histone proteins, are important for the normal function of a centromere during meiosis and mitosis. Reproductive isolation could result from incompatibilities in these proteins, with chromatin-binding proteins from one species failing to recognize the repeat sequence of the other (Sawamura et al., 2012). Centromere incompatibilities could then lead to centromere-drive where there is an imbalance in centromere strength during meiosis, contributing to post-zygotic isolation (Henikoff et al., 2001; Malik and Henikoff, 2003).

Population Genetics – Theory and Inference

Substantial progress has been made toward understanding speciation using sequence data. Faria and Navarro (2010) develop a framework based on the models of Navarro and Barton (2003) and Kirkpatrick and Barton (2006) to study how CRs contribute to speciation and how they first become fixed in different populations. This links population genetics theory to speciation and encompasses a model of how CRs have become fixed outside of previous models based on drift (Walsh, 1982; Lande, 1985; Spirito, 1998), selective advantages (Nei et al., 1967), meiotic drive (Nei et al., 1967), or epistatic interactions (Charlesworth, 1974).

There is the need for development of models that incorporate synergistic effects of genic and chromosomal variation, as well as the effects of drift and selection. Such models might provide opportunities to evaluate when a single process can explain observed patterns of variation in a dataset, and when multiple interacting processes need to be invoked. However, the reality is that genic and chromosomal divergences occur in parallel, CRs potentially have multiple effects, and that such models will only be capable of distinguishing among hypothesized processes in the simplest of systems. Here comparative genomic datasets, in particular lineages with a history of recent CRs, might offer unique opportunities. With such data, it is possible to test predictions of molecular evolution and biogeography, but empirical data can produce similar signatures under different scenarios (Noor and Bennett, 2009). One possible way forward is to exploit the different types of evidence (**Figure 7**) to set up competing/complementary models of single and joint effects that could then be evaluated using comparative sequence data.

In systems with multiple CRs, like the *Petrogale* system described here, we can view each CR as a semi-independent evolutionary replicate, providing opportunities to examine how different independent CRs, fixing in populations at different times, relate to genic divergence. This in turn can inform our understanding of processes. For instance, evidence of synchrony in fixation of different rearrangements may provide evidence of meiotic drive (e.g., Pardo-Manuel de Villena and Sapienza, 2001). Under meiotic drive hypotheses, we expect rearrangements will be of similar ages. If, however, CRs generate beneficial fitness effects (spread by positive selection), we do not expect fixation to occur at similar times. With combined cytogenetic understanding, this allows us to fit models to different regions along each chromosome to capture their unique evolutionary histories. If rearrangements are important to divergence, we expect the times at which they are established to coincide with speciation events. Using coalescent models to date speciation events and then mapping rearrangements on these phylogenies will help elucidate which CRs could have been involved in speciation (also previously suggested by Faria and Navarro, 2010). The key here will be finding systems where CRs are relatively recent, to distinguish between genic divergence post speciation vs. mutations causing reproductive isolation in relation to genomic architecture.

CONCLUSION

Integration of cytogenetic, genomic, and epigenetic data using a holistic approach will be crucial to improving our understanding of how genomic architecture influences and potentially drives reproductive isolation amongst organisms. The non-model system highlighted here has value for many reasons aside from its recent origin and chromosomal diversity. These include: the effect of reversed recombination rate between the sexes in marsupials relative to most other eutherians (males > females), X and Y chromosomes do not pair during meiosis (they lack the pseudoautosomal region), and epigenetic mechanisms vary significantly from eutherians. This one non-model system illustrates how biological variation can provide valuable contrasts to model systems. The growth of genomic and computational technology is opening new vistas on fundamental questions about how genomic architecture influences evolution.

DATA ACCESSIBILITY

Raw sequencing reads associated with this study are available at NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra; BioProject PRJNA360868; BioSamples SAMN06233288-SAMN06233338). A Dryad Repository (http://dx.doi.org/10.5061/dryad.mm856) contains code that was used in data analysis and Sequence Read Archive accession details.

ETHICS STATEMENT

This study did not use live animals and therefore does not require animal ethics approval. Samples used were tissues from the Australian Museum collection. Animal research at the Australian Museum and in Australia is governed by the "Australian code for the care and use of animals for scientific purposes." The Code only requires a research project to be approved by an ethics committee if the research involves live animals.

AUTHOR CONTRIBUTIONS

SP had substantial contribution to the conception, acquisition, design, analysis and interpretation of the work, drafting the work, final approval of the version to be published and agreement to be accountable for all aspects of the work. JB had substantial contribution to the conception, design, analysis and interpretation of the work, drafting the work, final approval of the version to be published and agreement to be accountable for all aspects of the work. MB had substantial contribution to the analysis of the work, drafting the work, final approval of the version to be published and agreement to be accountable for all aspects of the work. JD had substantial contribution to the conception, acquisition and design of the work, drafting the work, final approval of the version to be published and agreement to be accountable for all aspects of the work. MK had substantial contribution to the conception of the work, drafting the work, final approval of the version to be published and agreement to be accountable for all aspects of the work. ME and CM had substantial contribution to the conception, acquisition, design and interpretation of the work, drafting the work, final approval of the version to be published and agreement to be accountable for all aspects of the work.

FUNDING

This research was supported by an Australian Research Council Discovery Grant (DP160100187) as well as and ARC Laureate Fellowship awarded to CM (ARC FL110100104).

ACKNOWLEDGMENTS

We would like to thank the Australian Museum and South Australian Museum for samples. In addition, we would like to thank all the individuals that contributed tissues that were used in this study. We would like to thank Rachel O'Neill for access to the tammar wallaby genome. We thank Maxine Piggott for assistance with laboratory work. We would also like to thank members of the Moritz and Kirkpatrick labs for ongoing discussions relating to this study, in particular Josh Penalba and Ana Silva for advice on analyses. We thank Ke Bi and Sonal Singhal for providing scripts and for discussions. This work used the ACRF Biomolecular Resource Facility at ANU.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00010/full#supplementary-material

REFERENCES

Ayala, D., Guerrero, R. F., and Kirkpatrick, M. (2013). Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution* 67, 946–958. doi: 10.1111/j.1558-5646.2012.01836.x

Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D., and Eichler, E. E. (2004). Hotspots of mammalian chromosomal evolution. *Genome Biol.* 5:R23. doi: 10.1186/gb-2004-5-4-r23

Baird, A. B., Hillis, D. M., Patton, J. C., and Bickham, J. W. (2009). Speciation by monobrachial centric fusions: a test of the model using nuclear DNA sequences from the bat genus *Rhogeessa*. *Mol. Phylogenet. Evol.* 50, 256–267. doi: 10.1016/j.ympev.2008.11.001

Baker, R. J., and Bickham, J. W. (1986). Speciation by monobrachial centric fusions. *Proc. Natl. Acad. Sci. U.S.A.* 83, 8245–8248. doi: 10.1073/pnas.83.21.8245

Balcova, M., Faltusova, B., Gergelits, V., Bhattacharyya, T., Mihola, O., Trachtulec, Z., et al. (2016). Hybrid sterility locus on chromosome X controls meiotic recombination rate in mouse. *PLoS Genet.* 12:e1005906. doi: 10.1371/journal.pgen.1005906

Basset, P., Yannic, G., Brünner, H., and Hausser, J. (2006). Restricted gene flow at specific parts of the shrew genome in chromosomal hybrid zones. *Evolution* 60, 1718–1730. doi: 10.1554/06-181.1

Bateson, W. (1909). "Heredity and variation in modern lights," in *Darwin and Modern Science*, ed. A. C. Seward (Cambridge: Cambridge University Press), 85–101.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* 41, 379–406. doi: 10.1146/annurev-ecolsys-102209-144621

Bee, C. A., and Close, R. L. (1993). Mitochondrial DNA analysis of introgression between adjacent taxa of rock-wallabies, *Petrogale* species (Marsupialia: Macropodidae). *Genet. Res.* 61, 21–37. doi: 10.1017/S0016672300031074

Bergero, R., and Charlesworth, D. (2009). The evolution of restricted recombination in sex chromosomes. *Trends Ecol. Evol.* 24, 94–102. doi: 10.1016/j.tree.2008.09.010

Borodin, P. M., Ladygina, T. Y., Rodionova, M. I., Zhelezova, A. I., Zykovich, A. S., and Axenovich, T. I. (2005). Genetic control of chromosome synapsis in mice heterozygous for a paracentric inversion. *Russ. J. Genet.* 41, 602–607. doi: 10.1007/s11177-005-0133-6

Briscoe, D. A., Calaby, J. M., Close, R. L., Maynes, G. M., Murtagh, C. E., and Sharman, G. B. (1982). "Isolation, introgression and genetic variation in rock wallabies," in *Species at Risk: Research in Australia*, eds R. H. Groves and W. D. L. Ryde (Canberra, ACT: Australian Academy of Science), 73–87.

Brown, J. D., and O'Neill, R. J. (2010). Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annu. Rev. Genomics Hum. Genet.* 11, 291–316. doi: 10.1146/annurev-genom-082509-141554

Bryant, D., and Moulton, V. (2002). "NeighborNet: an agglomerative method for the construction of planar phylogenetic networks," in *International Workshop on Algorithms in Bioinformatics*, R. Guigó and D. Gusfield (Heidelberg: Springer), 375–391.

Bulazel, K. V., Ferreri, G. C., Eldridge, M. D. B., and O'Neill, R. J. (2007). Species-specific shifts in centromere sequence composition are coincident with breakpoint reuse in karyotypically divergent lineages. *Genome Biol.* 8:R170. doi: 10.1186/gb-2007-8-8-r170

Bush, G. L., Case, S. M., Wilson, A. C., and Patton, J. L. (1977). Rapid speciation and chromosome evolution in mammals. *Proc. Natl. Acad. Sci. U.S.A.* 74, 3942–3946. doi: 10.1073/pnas.74.9.3942

Campeau-Péloquin, A., Kirsch, J. A., Eldridge, M. D., and Lapointe, F. J. (2001). Phylogeny of the rock-wallabies, *Petrogale* (Marsupialia: Macropodidae) based on DNA/DNA hybridisation. *Aust. J. Zool.* 49, 463–486. doi: 10.1071/ZO01034

Capilla, L., Medarde, N., Alemany-Schmidt, A., Oliver-Bonet, M., Ventura, J., and Ruiz-Herrera, A. (2014). Genetic recombination variation in wild Robertsonian mice: on the role of chromosomal fusions and Prdm9 allelic background. *Proc. R. Soc. Lond.* 281:20140297. doi: 10.1098/rspb.2014.0297

Capilla, L., Sánchez-Guillén, R. A., Farré, M., Paytuví-Gallart, A., Malinverni, R., Ventura, J., et al. (2016). Mammalian comparative genomics reveals genetic and epigenetic features associated with genome reshuffling in Rodentia. *Genome Biol. Evol.* doi: 10.1093/gbe/evw276

Carbone, L., Harris, R. A., Gnerre, S., Veeramah, K. R., Lorente-Galdos, B., Huddleston, J., et al. (2014). Gibbon genome and the fast karyotype evolution of small apes. *Nature* 513, 195–201. doi: 10.1038/nature13679

Charlesworth, B. (1974). Inversion polymorphism in a two-locus genetic system. *Genetical Res.* 23, 259–280. doi: 10.1017/S0016672300014919

Charlesworth, B., Coyne, J. A., and Barton, N. H. (1987). The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* 130, 113–146. doi: 10.1086/284701

Chmátal, L., Gabriel, S. I., Mitsainas, G. P., Martínez-Vargas, J., Ventura, J., Searle, J. B., et al. (2014). Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Curr. Biol.* 24, 2295–2300. doi: 10.1016/j.cub.2014.08.017

Close, R. L., Bell, J. N., Dollin, A. E., and Harding, H. R. (1996). Spermatogenesis and synaptonemal complexes of hybrid *Petrogale* (Marsupialia). *J. Hered.* 87, 96–107. doi: 10.1093/oxfordjournals.jhered.a0229829

Cordaux, R., and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691–703. doi: 10.1038/nrg2640

Coyne, J. A., Barton, N. H., and Turelli, M. (2000). Is Wright's shifting balance process important in evolution? *Evolution* 54, 306–317. doi: 10.1111/j.0014-3820.2000.tb00033.x

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

Crawford, J. E., Riehle, M. M., Guelbeogo, W. M., Gneme, A., Sagnon, N. F., Vernick, K. D., et al. (2015). Reticulate speciation and barriers to introgression in the *Anopheles gambiae* species complex. *Genome Biol. Evol.* 7, 3116–3131. doi: 10.1093/gbe/evv203

Cui, R., Schumer, M., Kruesi, K., Walter, R., Andolfatto, P., and Rosenthal, G. G. (2013). Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. *Evolution* 67, 2166–2179. doi: 10.1111/evo.12099

Darlington, C. D. (1958). *The Evolution of Genetic Systems*, 2nd Edn. New York, NY: Basic Books Inc.

Dobigny, G., Britton-Davidian, J., and Robinson, T. J. (2017). Chromosomal polymorphism in mammals: an evolutionary perspective. *Biol. Rev. Camb. Philos. Soc.* 92, 1–21. doi: 10.1111/brv.12213

Dobzhansky, T. (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics* 21, 113–135.

Dobzhansky, T. (1950). Genetics of natural populations XIX Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. *Genetics* 35, 288–302.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* 28, 2239–2252. doi: 10.1093/molbev/msr048

Durand, S., Bouché, N., Strand, E. P., Loudet, O., and Camilleri, C. (2012). Rapid establishment of genetic incompatibility through natural epigenetic variation. *Curr. Biol.* 22, 326–331. doi: 10.1016/j.cub.2011.12.054

Edwards, S. V., Potter, S., Schmitt, C. J., Bragg, J. G., and Moritz, C. (2016). Reticulation, divergence, and the phylogeography–phylogenetics continuum. *Proc. Natl. Acad. Sci. U.S.A.* 113, 8025–8032. doi: 10.1073/pnas.1601066113

Eldridge, M. D. B. (2008). "Rock-wallabies: *Petrogale*," in *The Mammals of Australia*, 3rd Edn, ed. S. Van Dyck and R. Strahan (Chatswood, NSW: New Holland Publishers Pty Ltd), 361.

Eldridge, M. D. B., and Close, R. L. (1992). Taxonomy of rock wallabies, *Petrogale* (Marsupialia, Macropodidae). I. A revision of the eastern *Petrogale* with the description of 3 new species. *Aust. J. Zool.* 40, 605–625. doi: 10.1071/ZO9920605

Eldridge, M. D. B., and Close, R. L. (1993). Radiation of chromosome shuffles. *Curr. Opin. Genet. Dev.* 3, 915–922. doi: 10.1016/0959-437X(93)90014-G8

Eldridge, M. D. B., and Close, R. L. (1997). Chromosomes and evolution in rock-wallabies, *Petrogale* (Marsupialia: Macropodidae). *Aust. Mammal.* 19, 123–136.

Eldridge, M. D. B., Close, R. L., and Johnston, P. G. (1990). Chromosomal rearrangements in rock wallabies, *Petrogale* (Marsupialia: Macropodidae). III. G-banding analysis of the *Petrogale inornata* and *P. penicillata*. *Genome* 33, 798–802. doi: 10.1139/g90-120

Eldridge, M. D. B., Close, R. L., and Johnston, P. G. (1991). Chromosomal rearrangements in rock wallabies, *Petrogale* (Marsupialia: Macropodidae). IV. G-banding analysis of the *Petrogale lateralis* complex. *Aust. J. Zool.* 39, 621–627. doi: 10.1071/ZO9910629

Eldridge, M. D. B., Dollin, A. E., Johnston, P. G., Close, R. L., and Murray, J. D. (1988). Chromosomal rearrangements in rock wallabies, *Petrogale* (Marsupialia, Macropodidae). I. The P. assimilis species complex. G-banding and synaptonemal complex analysis. *Cytogenet. Cell Genet.* 48, 228–232. doi: 10.1159/000132634

Eldridge, M. D. B., and Johnston, P. G. (1993). Chromosomal rearrangements in rock wallabies, *Petrogale* (Marsupialia: Macropodidae). VIII. An investigation of the non-random nature of karyotypic change. *Genome* 36, 524–534. doi: 10.1139/g93-072

Eldridge, M. D. B., Johnston, P. G., and Close, R. L. (1992a). Chromosomal rearrangements in rock wallabies, *Petrogale* (Marsupialia: Macropodidae). VI. Determination of the plesiomorphic karyotype: G-Banding comparison of *Thylogale* with *Petrogale persephone*, *P. xanthopus*, and *P. l. lateralis*. *Cytogenet. Cell Genet.* 61, 29–33.

Eldridge, M. D. B., Johnston, P. G., Close, R. L., and Lowry, P. S. (1989). Chromosomal rearrangements in rock wallabies, *Petrogale* (Marsupialia: Macropodidae). II. G-banding analysis of *Petrogale godmani*. *Genome* 32, 935–940. doi: 10.1139/g89-5348

Eldridge, M. D. B., Johnston, P. G., and Lowry, P. S. (1992b). Chromosomal rearrangements in rock wallabies, *Petrogale* (Marsupialia: Macropodidae). VII. G-banding analysis of *P. brachyotis* and *P. concinna*: species with dramatically altered karyotypes. *Cytogenet. Cell Genet.* 61, 34–39. doi: 10.1159/0001333649

Faria, R., and Navarro, A. (2010). Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends Ecol. Evol.* 25, 660–669. doi: 10.1016/j.tree.2010.07.008

Farré, M., Robinson, T. J., and Ruiz-Herrera, A. (2015). An integrative breakage model of genome architecture, reshuffling and evolution. *Bioessays* 37, 479–488. doi: 10.1002/bies.201400174

Feder, J. L., and Nosil, P. (2009). Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions? *Evolution* 63, 3061–3075. doi: 10.1111/j.1558-5646.2009.00786.x

Fedyk, S., and Chętnicki, W. (2007). Preferential segregation of metacentric chromosomes in simple Robertsonian heterozygotes of *Sorex araneus*. *Heredity* 99, 545–552. doi: 10.1038/sj.hdy.6801036

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., et al. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347:1258524. doi: 10.1126/science.1258524

Förster, D. W., Jones, E. P., Jóhannesdóttir, F., Gabriel, S. I., Giménez, M. D., Panithanarak, T., et al. (2016). Genetic differentiation within and away from the chromosomal rearrangements characterising hybridising chromosomal races of the western house mouse (*Mus musculus domesticus*). *Chromosome Res.* 24, 271–280. doi: 10.1007/s10577-016-9520-1

Franchini, P., Colangelo, P., Solano, E., Capanna, E., Verheyen, E., and Castiglia, R. (2010). Reduced gene flow at pericentromeric loci in a hybrid zone involving chromosomal races of the house mouse *Mus musculus domesticus*. *Evolution* 64, 2020–2032. doi: 10.1111/j.1558-5646.2010.00964.x

Futuyma, D. J., and Mayer, G. C. (1980). Non-allopatric speciation in animals. *Syst. Zool.* 29, 254–271. doi: 10.2307/2412661

Garagna, S., Marziliano, N., Zuccotti, M., Searle, J. B., Capanna, E., and Redi, C. A. (2001). Pericentromeric organization at the fusion point of mouse Robertsonian translocation chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* 98, 171–175. doi: 10.1073/pnas.98.1.171

Garagna, S., Page, J., Fernandez-Donoso, R., Zuccotti, M., and Searle, J. B. (2014). The Robertsonian phenomenon in the house mouse: mutation, meiosis and speciation. *Chromosoma* 123, 529–544. doi: 10.1007/s00412-014-0477-6

Giménez, M. D., White, T. A., Haufe, H. C., Panithanarak, C., and Searle, J. B. (2013). Understanding the basis of diminished gene flow between hybridizing chromosome races of the house mouse. *Evolution* 67, 1446–1462. doi: 10.1111/evo.12054

Grant, V. (1964). The architecture of the germ plasm. *Soil Sci.* 98:212. doi: 10.1097/00010694-196409000-00032

Graves, J. A. M., and Renfree, M. B. (2013). Marsupials in the age of genomics. *Annu. Rev. Genomics Hum. Genet.* 14, 393–420. doi: 10.1146/annurev-genom-091212-153452

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science* 328, 710–722. doi: 10.1126/science.1188021

Grey, C., Barthès, P., Chauveau-Le Friec, G., Langa, F., Baudat, F., and De Massy, B. (2011). Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biol.* 9:e1001176. doi: 10.1371/journal.pbio.1001176

Guerrero, R. F., and Kirkpatrick, M. (2014). Local adaptation and the evolution of chromosome fusions. *Evolution* 68, 2747–2756. doi: 10.1111/evo.12481

Guerrero, R. F., Kirkpatrick, M., and Perrin, N. (2012a). Cryptic recombination in the ever-young sex chromosomes of Hylid frogs. *J. Evol. Biol.* 25, 1947–1954. doi: 10.1111/j.1420-9101.2012.02591.x

Guerrero, R. F., Rousset, F., and Kirkpatrick, M. (2012b). Coalescent patterns for chromosomal inversions in divergent populations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 430–438. doi: 10.1098/rstb.2011.0246

Guichaoua, M. R., Delafontaine, D., Taurelle, R., Taillemite, J. L., Morazzani, M. R., and Luciani, J. M. (1986). Loop formation and synaptic adjustment in a human male heterozygous for two pericentric inversions. *Chromosoma* 93, 313–320. doi: 10.1007/BF003275895

Haldane, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* 12, 101–109. doi: 10.1007/BF02983075

Harewood, L., and Fraser, P. (2014). The impact of chromosomal rearrangements on regulation of gene expression. *Hum. Mol. Gen.* 23, R76–R82. doi: 10.1093/hmg/ddu278

Hayman, D. L. (1990). Marsupial cytogenetics. *Aust. J. Zool.* 37, 331–349. doi: 10.1071/ZO9890331

Hazlitt, S. L., Goldizen, A. W., and Eldridge, M. D. B. (2006). Significant patterns of population genetic structure and limited gene flow in a threatened macropodid marsupial despite continuous habitat in southeast Queensland, Australia. *Conserv. Genet.* 7, 675–689. doi: 10.1007/s10592-005-9101-x

Henikoff, S., Ahmad, K., and Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293, 1098–1102. doi: 10.1126/science.1062939

Henzel, J. V., Nabeshima, K., Schvarzstein, M., Turner, B. E., Villeneuve, A. M., and Hillers, K. J. (2011). An asymmetric chromosome pair undergoes synaptic adjustment and crossover redistribution during *Caenorhabditis elegans* meiosis: implications for sex chromosome evolution. *Genetics* 187, 685–699. doi: 10.1534/genetics.110.124958

Hoffmann, A. A., and Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* 39, 21–42. doi: 10.1146/annurev.ecolsys.39.110707.173532

Hooper, D. M., and Price, T. (2015). Rates of karyotypic evolution in Estrildid finches differ between island and continental clades. *Evolution* 69, 890–903. doi: 10.1111/evo.12633

Horn, A., Basset, P., Yannic, G., Banaszek, A., Borodin, P. M., Bulatova, N. S., et al. (2012). Chromosomal incompatibilities do not seem to affect the gene flow in hybrid zones between karyotypic races of the common shrew (*Sorex araneus*). *Evolution* 66, 882–889. doi: 10.1111/j.1558-5646.2011.01478.x

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

James, S. H. (1982). "Coadaptation of the genetic system and the evolution of isolation among populations of Western Australian native plants," in *Mechanisms of Speciation*, ed. C. Barigozzi (New York, NY: Alan R. Liss Inc.), 461–470.

Johannisson, R., and Winking, H. (1994). Synaptonemal complexes of chains and rings in mice heterozygous for multiple Robertsonian translocations. *Chromosome Res.* 2, 137–145. doi: 10.1007/BF015534922

Kaelbling, M., and Fechheimer, N. S. (1985). Synaptonemal complex analysis of a pericentric inversion in chromosome 2 of domestic fowl, *Gallus domesticus*. *Cytogenet. Genome Res.* 39, 82–86. doi: 10.1159/0001321122

Kandul, N. P., Lukhtanov, V. A., and Pierce, N. E. (2007). Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution* 61, 546–559. doi: 10.1111/j.1558-5646.2007.00046.x

Key, K. H. L. (1968). The concept of stasipatric speciation. *Syst. Biol.* 17, 14–22. doi: 10.3390/insects2010049

King, M. (1993). *Species Evolution: The Role of Chromosome Change*. Cambridge: Cambridge University Press.

Kingswood, S. C., Kumamoto, A. T., Sudman, P. D., Fletcher, K. C., and Greenbaum, I. F. (1994). Meiosis in chromosomally heteromorphic goitered gazelle, *Gazella subgutturosa* (Artiodactyla, Bovidae). *Chromosome Res.* 2, 37–46. doi: 10.1007/BF015394529

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biol.* 8:e1000501. doi: 10.1371/journal.pbio.1000501

Kirkpatrick, M. (2017). The evolution of genome structure by natural and sexual selection. *J. Hered.* 108, 3–11. doi: 10.1093/jhered/esw041

Kirkpatrick, M., and Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics* 173, 419–434. doi: 10.1534/genetics.105.047985

Kulathinal, R. J., Stevison, L. S., and Noor, M. A. F. (2009). The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genet.* 5:e1000550. doi: 10.1371/journal.pgen.1000550

Lande, R. (1985). The fixation of chromosomal rearrangements in a subdivided population with local extinction and recolonisation. *Heredity* 54, 323–332. doi: 10.1038/hdy.1985.430

Larkin, D. M., Pape, G., Donthu, R., Auvil, L., Welge, M., and Lewin, H. A. (2009). Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res.* 19, 770–777. doi: 10.1101/gr.086546.108

Leaché, A. D., Banbury, B. L., Linkem, C. W., and de Oca, A. N. M. (2016). Phylogenomics of a rapid radiation: is chromosomal evolution linked to increased diversification in north american spiny lizards (Genus *Sceloporus*)? *BMC Evol. Biol.* 16:63. doi: 10.1186/s12862-016-0628-x

Lemaitre, C., Zaghloul, L., Sagot, M. F., Gautier, C., Arneodo, A., Tannier, E., et al. (2009). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC Genomics* 10:335. doi: 10.1186/1471-2164-10-335

Lindholm, A. K., Dyer, K. A., Firman, R. C., Fishman, L., Forstmeier, W., Holman, L., et al. (2016). The ecology and evolutionary dynamics of meiotic drive. *Trends Ecol. Evol.* 31, 315–326. doi: 10.1016/j.tree.2016.02.001

Lohse, K., Clarke, M., Ritchie, M. G., and Etges, W. J. (2015). Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. *Evolution* 69, 1178–1190. doi: 10.1111/evo.12650

Lukhtanov, V. A., Kandul, N. P., Plotkin, J. B., Dantchenko, A. V., Haig, D., and Pierce, N. E. (2005). Reinforcement of pre-zygotic isolation and karyotype evolution in *Agrodiaetus* butterflies. *Nature* 436, 385–389. doi: 10.1038/nature03704

Malik, H. S., and Henikoff, S. (2003). Phylogenomics of the nucleosome. *Nat. Struct. Mol. Biol.* 10, 882–891. doi: 10.1038/nsb996

Mallet, J., Besansky, N., and Hahn, M. W. (2016). How reticulated are species? *Bioessays* 38, 140–149. doi: 10.1002/bies.201500149

Manterola, M., Page, J., Vasco, C., Berrios, S., Parra, M. T., Viera, A., et al. (2009). A high incidence of meiotic silencing of unsynapsed chromatin is not associated with substantial pachytene loss in heterozygous male mice carrying multiple simple Robertsonian translocations. *PLoS Genet.* 5:e1000625. doi: 10.1371/journal.pgen.1000625

Martin, S. H., Davey, J. W., and Jiggins, C. D. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* 32, 244–257. doi: 10.1093/molbev/msu269

Maynes, G. M. (1989). "Zoogeography of the Macropodoidea," in *Kangaroos, Wallabies and Rat-Kangaroos*, 1st Edn, ed. G. Grigg, P. Jarman, and I. Hume (Chipping Norton, NSW: Surrey Beatty & Sons Pty, Ltd.), 47–66.

McGaugh, S. E., and Noor, M. A. F. (2012). Genomic impacts of chromosomal inversions in parapatric *Drosophila* species. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 422–429. doi: 10.1098/rstb.2011.0250

Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C., and Forejt, J. (2009). A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* 323, 373–375. doi: 10.1126/science.1163601

Molina, W. F., Martinez, P. A., Bertollo, L. A. C., and Bidau, C. J. (2014). Evidence for meiotic drive as an explanation for karyotype changes in fishes. *Mar. Genomics* 15, 29–34. doi: 10.1016/j.margen.2014.05.001

Muller, H. (1942). Isolating mechanisms, evolution and temperature. *Biol. Symp.* 6, 71–125.

Muller, H. J. (1930). Types of visible variations induced by X-rays in *Drosophila*. *J. Genet.* 22, 299–334. doi: 10.1007/BF02984195

Nachman, M. W., and Searle, J. B. (1995). Why is the house mouse karyotype so variable? *Trends Ecol. Evol.* 10, 397–402. doi: 10.1016/S0169-5347(00)89155-7

Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol. Evol.* 28, 719–728. doi: 10.1016/j.tree.2013.09.004

Nam, K., Munch, K., Hobolth, A., Dutheil, J. Y., Veeramah, K., Woerner, A., et al. (2014). *Strong Selective Sweeps Associated with Ampliconic Regions in Great Ape X Chromosomes*. Available at: https://arxiv.org/abs/1402.5790

Nater, A., Burri, R., Kawakami, T., Smeds, L., and Ellegren, H. (2015). Resolving evolutionary relationships in closely related species with whole-genome sequencing data. *Syst. Biol.* 64, 1000–1017. doi: 10.1093/sysbio/syv045

Navarro, A., and Barton, N. H. (2003). Accumulating postzygotic isolation in parapatry: a new twist on chromosomal speciation. *Evolution* 57, 447–459. doi: 10.1111/j.0014-3820.2003.tb01537.x

Neale, M. J., and Keeney, S. (2006). Clarifying the mechanics of DNA strand exchange in meiotic recombination. *Nature* 442, 153–158. doi: 10.1038/nature04885

Nei, M., Kojima, K. I., and Schaffer, H. E. (1967). Frequency changes of new inversions in populations under mutation-selection equilibria. *Genetics* 57, 741–750.

Noor, M. A., and Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* 103, 439–444. doi: 10.1038/hdy.2009.151

Noor, M. A., Grams, K. L., Bertucci, L. A., and Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proc. Natl. Acad. Sci. U.S.A.* 98, 12084–12088. doi: 10.1073/pnas.221274498

O'Neill, R. W., Eldridge, M. D. B., Toder, R., Ferguson-Smith, M. A., O'Brien, P. C., and Graves, J. A. M. (1999). Chromosome evolution in kangaroos (Marsupialia: Macropodidae): cross species chromosome painting between the tammar wallaby and rock wallaby spp. *with the* 2 *n*=22 ancestral macropodid karyotype. *Genome* 42, 525–530. doi: 10.1139/g98-159

O'Neill, R. J., Eldridge, M. D. B., and Metcalfe, C. J. (2004). Centromere dynamics and chromosome evolution in marsupials. *J. Hered.* 95, 375–381. doi: 10.1093/jhered/esh063

Ortiz-Barrientos, D., Engelstädter, J., and Rieseberg, L. H. (2016). Recombination rate evolution and the origin of species. *Trends Ecol. Evol.* 31, 226–236. doi: 10.1016/j.tree.2015.12.016

Pardo-Manuel de Villena, F., and Sapienza, C. (2001). Female meiosis drives karyotypic evolution in mammals. *Genetics* 159, 1179–1189.

Payseur, B. A., and Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Mol. Ecol.* 25, 2337–2360. doi: 10.1111/mec.13557

Peischl, S., Koch, E., Guerrero, R. F., and Kirkpatrick, M. (2013). A sequential coalescent algorithm for chromosomal inversions. *Heredity* 111, 200–209. doi: 10.1038/hdy.2013.38

Pennell, M. W., Kirkpatrick, M., Otto, S. P., Vamosi, J. C., Peichel, C. L., Valenzuela, N., et al. (2015). Y fuse? Sex chromosome fusions in fishes and reptiles. *PLoS Genet.* 11:e1005237. doi: 10.1371/journal.pgen.1005237

Pinho, C., and Hey, J. (2010). Divergence with gene flow: models and data. *Annu. Rev. Ecol. Evol. Syst.* 41, 215–230. doi: 10.1146/annurev-ecolsys-102209-144644

Pokorná, M., Altmanová, M., and Kratochvíl, L. (2014). Multiple sex chromosomes in the light of female meiotic drive in amniote vertebrates. *Chromosome Res.* 22, 35–44. doi: 10.1007/s10577-014-9403-2

Polyakov, A. V., White, T. A., Jones, R. M., Borodin, P. M., and Searle, J. B. (2011). Natural hybridization between extremely divergent chromosomal races of the common shrew (*Sorex araneus*, Soricidae, Soricomorpha): hybrid zone in Siberia. *J. Evol. Biol.* 24, 1393–1402. doi: 10.1111/j.1420-9101.2011.02266.x

Pope, L. C., Sharp, A., and Moritz, C. (1996). Population structure of the yellow-footed rock-wallaby *Petrogale* xanthopus (Gray, 1854) inferred from mtDNA sequences and microsatellite loci. *Mol. Ecol.* 5, 629–640. doi: 10.1111/j.1365-294X.1996.tb00358.x6

Potter, S., Cooper, S. J., Metcalfe, C. J., Taggart, D. A., and Eldridge, M. D. B. (2012a). Phylogenetic relationships of rock-wallabies *Petrogale* (Marsupialia: Macropodidae) and their biogeographic history within Australia. *Mol. Phylogenet. Evol.* 62, 640–652. doi: 10.1016/j.ympev.2011.11.005

Potter, S., Eldridge, M. D. B., Taggart, D. A., and Cooper, S. J. B. (2012b). Multiple biogeographical barriers identified across the monsoon tropics of northern Australia: phylogeographic analysis of the brachyotis group of rock-wallabies. *Mol. Ecol.* 21, 2254–2269. doi: 10.1111/j.1365-294X.2012.05523.x

Potter, S., Moritz, C., and Eldridge, M. D. B. (2015). Gene flow despite complex Robertsonian fusions among rock-wallaby (*Petrogale*) species. *Biol. Lett.* 11:20150731. doi: 10.1098/rsbl.2015.0731

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., et al. (2013). Great ape genetic diversity and population history. *Nature* 499, 471–475. doi: 10.1038/nature12228

Presgraves, D. C. (2008). Sex chromosomes and speciation in *Drosophila*. *Trends Genet.* 24, 336–343. doi: 10.1016/j.tig.2008.04.007

Presgraves, D. C. (2010). Darwin and the origin of interspecific genetic incompatibilities. *Am. Nat.* 176, S45–S60. doi: 10.1086/657058

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., et al. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 26, 342–350. doi: 10.1101/gr.193474.115

Qi, J., Chen, Y., Copenhaver, G P., and Ma, H. (2014). Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10007–10012. doi: 10.1073/pnas.1321897111

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* 10:e1004342. doi: 10.1371/journal.pgen.1004342

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16, 351–358. doi: 10.1016/S0169-5347(01)02187-5

Rofe, R., and Hayman, D. (1985). G-banding evidence for a conserved complement in Marsupialia. *Cytogenet. Cell Genet.* 39, 40–50. doi: 10.1159/0001321018

Rofe, R. H. (1979). *G-Banding and Chromosome Evolution in Marsupials*. Ph.D. thesis, University of Adelaide, Adelaide SA.

Ruiz-Herrera, A., Castresana, J., and Robinson, T. J. (2006). Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol.* 7:R115. doi: 10.1186/gb-2006-7-12-r115

Saether, S. A., Saether, G. P., Borge, T., Wiley, C., Svedin, N., Andersson, G., et al. (2007). Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. *Science* 318, 95–97. doi: 10.1126/science.1141506

Sawamura, K., Ting, C.-T., Kopp, A., and Moyle, L. C. (2012). Mechanisms of Speciation. *Int. J. Evol. Biol.* 2012:820358. doi: 10.1155/2012/820358

Sharman, G. B., Close, R. L., and Maynes, G. M. (1990). Chromosomal evolution, phylogeny and speciation of rock wallabies (*Petrogale*: Macropodidae). *Aust. J. Zool.* 37, 351–363.

Shaw, D. D., Coates, D. J., and Wilkinson, P. (1986). Estimating the genic and chromosomal components of reproductive isolation within and between subspecies of the grasshopper *Caledia captiva*. *Can. J. Genet. Cytol.* 28, 686–695. doi: 10.1139/g86-098

Shiu, P. K. T., Raju, N. B., Zickler, D., and Metzenberg, R. L. (2001). Meiotic silencing by unpaired DNA. *Cell* 107, 905–916. doi: 10.1016/S0092-8674(01)00609-2

Sites, J. W., and Moritz, C. (1987). Chromosomal evolution and speciation revisited. *Syst. Zool.* 36, 153–174. doi: 10.2307/2413266

Slijepcevic, P. (1998). Telomeres and mechanisms of Robertsonian fusion. *Chromosoma* 107, 136–140. doi: 10.1007/s0041200502892

Smagulova, F., Gregoretti, I. V., Brick, K., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* 472, 375–378. doi: 10.1038/nature09869

Solis-Lemus, C., Yang, M., and Ané, C. (2016). Inconsistency of species-tree methods under gene flow. *Syst. Biol.* 65, 843–851. doi: 10.1093/sysbio/syw030

Spirito, F. (1998). "The role of chromosomal rearrangements in speciation," *Endless Forms*, ed. D. J. Howard and S. H. Berlocher (Oxford: Oxford University Press), 320–329.

Stebbins, G. L. (1950). *Variation and Evolution in Plants*. New York, NY: Columbia University Press.

Sturtevant, A. H. (1938). Essays on evolution. III. On the origin of interspecific sterility. *Q. Rev. Biol.* 13, 333–335. doi: 10.1086/394565

Suh, A. (2016). The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zool. Scripta* 45, 50–62. doi: 10.1111/zsc.12213

Templeton, A. R. (1981). Mechanisms of speciation—a population genetics approach. *Annu. Rev. Ecol. Syst.* 12, 23–48. doi: 10.1146/annurev.es.12.110181.000323

Than, C., Ruths, D., and Nakhleh, L. (2008). PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary histories. *BMC Bioinformatics* 9:322. doi: 10.1186/1471-2105-9-322

Torgasheva, A. A., and Borodin, P. M. (2010). Synapsis and recombination in inversion heterozygotes. *Biochem. Soc. Trans.* 38, 1676–1680. doi: 10.1042/BST0381676

Turelli, M. (1998). The causes of Haldane's rule. *Science* 282, 889–891. doi: 10.1126/science.282.5390.889

Turelli, M., and Moyle, L. C. (2007). Asymmetric postmating isolation: Darwin's corollary to Haldane's rule. *Genetics* 176, 1059–1088. doi: 10.1534/genetics.106.065979

Turner, J. M. (2007). Meiotic sex chromosome inactivation. *Development* 134, 1823–1831. doi: 10.1242/dev.000018

Ullastres, A., Farré, M., Capilla, L., and Ruiz-Herrera, A. (2014). Unraveling the effect of genomic structural changes in the rhesus macaque-implications for the adaptive role of inversions. *BMC Genomics* 15:530. doi: 10.1186/1471-2164-15-530

Vozdova, M., Sebestova, H., Kubickova, S., Cernohorska, H., Awadova, T., Vahala, J., et al. (2014). Impact of Robertsonian translocation on meiosis and reproduction: an impala (*Aepyceros melampus*) model. *J. Appl. Genet.* 55, 249–258. doi: 10.1007/s13353-014-0193-1

Walsh, J. B. (1982). Rate of accumulation of reproductive isolation by chromosome rearrangements. *Am. Nat.* 120, 510–532. doi: 10.1086/284008

Wen, D., Yu, Y., Hahn, M. W., and Nakhleh, L. (2016). Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Mol. Ecol.* 25, 2361–2372. doi: 10.1111/mec.13544

Wesche, P. L., and Robinson, T. J. (2012). Different patterns of Robertsonian fusion pairing in Bovidae and the house mouse: the relationship between chromosome size and nuclear territories. *Genet. Res.* 94, 97–111. doi: 10.1017/S0016672312000262

White, M. J. D. (1973). *Animal Cytology and Evolution*. Cambridge: Cambridge University Press.

White, M. J. D. (1978). *Modes of Speciation*. San Francisco, CA: W. H. Freeman.

Wright, S. (1982). The shifting balance theory and macroevolution. *Annu. Rev. Genet.* 16, 1–20. doi: 10.1146/annurev.ge.16.120182.000245

Wyttenbach, A., Borodin, P., and Hausser, J. (1998). Meiotic drive favors Robertsonian metacentric chromosomes in the common shrew (*Sorex araneus*, Insectivora, Mammalia). *Cytogenet. Genome Res.* 83, 199–206.

Yannic, G., Basset, P., and Hausser, J. (2009). Chromosomal rearrangements and gene flow over time in an inter-specific hybrid zone of the *Sorex araneus* group. *Heredity* 102, 616–625. doi: 10.1038/hdy.2009.19

Yoshida, K., and Kitano, J. (2012). The contribution of female meiotic drive to the evolution of neo-sex chromosomes. *Evolution* 66, 3198–3208. doi: 10.1111/j.1558-5646.2012.01681.x

Yu, Y., Dong, J., Liu, K., and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. U.S.A.* 111, 16448–16453. doi: 10.1073/pnas.1407950111

Yu, Y., Ristic, N., and Nakhleh, L. (2013). Fast algorithms and heuristics for phylogenomics under hybridization and incomplete lineage sorting. *BMC Bioinformatics* 14:S6. doi: 10.1186/1471-2105-14-S15-S6

Zickler, D., and Kleckner, N. (1999). Meiotic chromosomes: integrating structure and function. *Annu. Rev. Genet.* 33, 603–754. doi: 10.1146/annurev.genet.33.1.603

Conflict of Interest Statement: The handling Editor declared a shared affiliation, though no other collaboration, with one of the authors JD and states that the process nevertheless met the standards of a fair and objective review.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 13 April 2017
doi: 10.3389/fgene.2017.00044

LINE Insertion Polymorphisms are Abundant but at Low Frequencies across Populations of *Anolis carolinensis*

Robert P. Ruggiero[†], *Yann Bourgeois*[†] *and Stéphane Boissinot**

New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

OPEN ACCESS

Edited by:
Frederic J.J. Chain,
University of Massachusetts Lowell,
USA

Reviewed by:
Juan Caballero,
Autonomous University of Queretaro,
Mexico
Josefa Gonzalez,
Institut of Evolutionary Biology –
Consejo Superior de Investigaciones
Científicas-Universitat Pompeu Fabra,
Spain

****Correspondence:***
Stéphane Boissinot
sb5272@nyu.edu

[†]These authors have contributed equally to this work.

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *23 December 2016*
Accepted: *29 March 2017*
Published: *13 April 2017*

Citation:
Ruggiero RP, Bourgeois Y and Boissinot S (2017) LINE Insertion Polymorphisms are Abundant but at Low Frequencies across Populations of Anolis carolinensis. Front. Genet. 8:44. doi: 10.3389/fgene.2017.00044

Vertebrate genomes differ considerably in size and structure. Among the features that show the most variation is the abundance of Long Interspersed Nuclear Elements (LINEs). Mammalian genomes contain 100,000s LINEs that belong to a single clade, L1, and in most species a single family is usually active at a time. In contrast, non-mammalian vertebrates (fish, amphibians and reptiles) contain multiple active families, belonging to several clades, but each of them is represented by a small number of recently inserted copies. It is unclear why vertebrate genomes harbor such drastic differences in LINE composition. To address this issue, we conducted whole genome resequencing to investigate the population genomics of LINEs across 13 genomes of the lizard *Anolis carolinensis* sampled from two geographically and genetically distinct populations in the Eastern Florida and the Gulf Atlantic regions of the United States. We used the Mobile Element Locator Tool to identify and genotype polymorphic insertions from five major clades of LINEs (CR1, L1, L2, RTE and R4) and the 41 subfamilies that constitute them. Across these groups we found large variation in the frequency of polymorphic insertions and the observed length distributions of these insertions, suggesting these groups vary in their activity and how frequently they successfully generate full-length, potentially active copies. Though we found an abundance of polymorphic insertions (over 45,000) most of these were observed at low frequencies and typically appeared as singletons. Site frequency spectra for most LINEs showed a significant shift toward low frequency alleles compared to the spectra observed for total genomic single nucleotide polymorphisms. Using Tajima's D, F_{ST} and the mean number of pairwise differences in LINE insertion polymorphisms, we found evidence that negative selection is acting on LINE families in a length-dependent manner, its effects being stronger in the larger Eastern Florida population. Our results suggest that a large effective population size and negative selection limit the expansion of polymorphic LINE insertions across these populations and that the probability of LINE polymorphisms reaching fixation is extremely low.

Keywords: retrotransposon, LINE, *Anolis carolinensis*, genome resequencing, transposable element, selection

INTRODUCTION

The complete sequencing of dozens of vertebrate genomes representing most extant lineages has been an extraordinary source of information, thereby revolutionizing the field of genetics, development and evolutionary biology. However, those genomes vary considerably in size and structure and understanding the cause(s) of these differences is fundamental for meaningfully interpreting genomic annotations (Elliott and Gregory, 2015). Among the features that show the most variation across vertebrate taxa is the abundance and diversity of non-LTR retrotransposons [also called LINEs for Long Interspersed Nuclear Elements; reviewed in Tollis and Boissinot (2012)]. LINEs are autonomously replicating retroelements, meaning they encode the molecular machinery necessary for their own replication. LINEs are ubiquitous components of eukaryotic genomes and the origin of the main LINE lineages is very ancient, possibly predating the origin of eukaryotes (Malik et al., 1999). LINEs are classified into a number of clades based on the presence of conserved features (Malik et al., 1999; Kapitonov et al., 2009). The most basal clades of LINEs (e.g., R2, R4, RTE) contain a single open-reading frame (ORF) encoding a reverse transcriptase domain, while the most derived lineages contains two ORFs (e.g., L1, L2, CR1). The mechanism of transposition was characterized for the R2 and L1 elements and it is assumed that other LINEs transpose using a similar mechanism (Luan et al., 1993; Cost et al., 2002). Following transcription and export of LINE mRNA to the cytoplasm, LINE-encoded proteins are translated and form an RNA-protein complex that is reimported in the nucleus. In the nucleus, reverse transcription takes place at the site of insertion, through a process called target-primed reverse transcription. Although there is a strong *cis* preference (Wei et al., 2001), the replicative machinery of LINEs can act on other transcripts and is responsible for the amplification of the non-autonomous SINEs and of retrotransposed pseudogenes (Ohshima et al., 1996; Dewannieux et al., 2003; Dewannieux and Heidmann, 2005; Piskurek et al., 2009).

Long Interspersed Nuclear Elements are ubiquitous in vertebrates and constitute the dominant category of autonomously replicating retroelements in most vertebrate genomes (Tollis and Boissinot, 2012). They have considerably affected the size and structure of these genomes and it is believed that LINE abundance is one of the major determinants of haploid genome size differences among vertebrates. At one extreme, mammalian genomes contain extremely large numbers of LINEs that can account for as much as 30% of their size (Lander et al., 2001; Waterston et al., 2002). LINEs in placental mammals are represented by a single clade, L1. The vast majority of L1 elements are the product of past amplification and in most species only the most recently evolved family of elements is active at a time (Furano, 2000). Fish, amphibians and non-avian reptile genomes contain a much larger diversity of active LINE families, generally representing multiple clades (Volff et al., 2003; Duvernell et al., 2004; Furano et al., 2004; Novick et al., 2009; Blass et al., 2012; Chalopin et al., 2015). These families are usually represented by small numbers of very similar copies, suggesting that the majority of insertions are recent (Furano et al., 2004; Novick et al., 2009; Blass et al., 2012).

In mammals, the evolutionary dynamics of LINEs is relatively well understood. Population genetics and genomics studies in humans have shown that the majority of L1 elements behave as neutral alleles and accumulate readily in the genome of their host (Boissinot et al., 2006). This does not mean that L1 activity is fully neutral. In humans, a fitness cost related to the length of L1 elements has been demonstrated (Boissinot et al., 2001, 2006). This suggests that the deleterious effect of L1 result from the ability of long elements to mediate ectopic recombination events (Myers et al., 2005; Song and Boissinot, 2007). However, this cost is insufficient to prevent the fixation of most elements, hence the extremely large number of L1 copies in mammals. By comparison the dynamics of LINEs in non-mammalian genomes is not as well understood. The young age and relatively small number of LINEs in fish and reptile genomes could be interpreted as evidence for a lower rate of fixation of novel insertions in non-mammalian genomes. Studies in stickleback and in lizard suggest that, indeed, LINE insertions tend to be negatively selected, yet a number of insertions do reach fixation (Blass et al., 2012; Tollis and Boissinot, 2013). In addition, population genetics data in the pufferfish show that the frequency spectrum of recent insertions is consistent with neutrality (Neafsey et al., 2004). Thus we have been unable to exclude the possibility that LINEs are neutral or weakly deleterious in non-mammalian vertebrates and that their copy number is controlled by other means, possibly by a faster decay due to a higher rate of DNA loss (Novick et al., 2009; Blass et al., 2012).

At this point, our understanding of LINE population dynamics is heavily biased toward their dynamics in humans. However, the extreme abundance and low diversity of LINEs in mammals constitute a derived state relative to other vertebrates. Thus, inferences drawn from studies in mammals are unlikely to apply to other vertebrates. In addition, results obtained from previous studies in non-mammalian vertebrates provide only a partial picture since those studies relied on a relatively small number of polymorphisms, principally collected from the published reference genomes (Neafsey et al., 2004; Blass et al., 2012; Shen et al., 2013; Tollis and Boissinot, 2013). Thus, we decided to investigate the population dynamics of LINEs in a non-mammalian vertebrate, the green anole *Anolis carolinensis*, using a complete genome re-sequencing approach. The anole genome is a particularly good model because it is among the most diverse vertebrate genomes in terms of LINE diversity (Novick et al., 2009; Chalopin et al., 2015). Five LINE clades are simultaneously active in anole: L1, L2, CR1, R4 and RTE. These elements differ considerably in structure, copy number, and diversity (**Table 1**). For example, the L1 and the L2 clades contain 20 and 17 highly divergent families, respectively, whereas the CR1 clade is represented by only 4 closely related families. Since these clades and families coexist within the same genome, they are equally affected by the demography of their host. It is thus possible to assess their relative impact on fitness and to infer the evolutionary processes determining their diversification and replicative success.

In this article we present the first population genetic analysis of LINEs using re-sequencing data in a non-mammalian vertebrate. We sequenced thirteen individuals, from two populations with different demographic histories, at a depth of coverage ranging from 8 to 16×. For each resequenced genome we then characterized the single nucleotide polymorphisms (SNPs) and polymorphic sites containing LINE insertions not found in the reference genome. We determined that the number of insertion polymorphisms generated by LINEs in this species is large, exceeding 45,000 insertions, with substantial differences in replicative success among clades. We also determined that the vast majority of these insertions exist at very low frequency in natural populations as a result of the very large effective population size of *A. carolinensis* and of purifying selection against those insertions.

MATERIALS AND METHODS

Sampling

There are five geographically and genetically distinct anole populations in North America (Campbell-Staton et al., 2012; Tollis et al., 2012; Tollis and Boissinot, 2014; Manthey et al., 2016). We decided to focus our re-sequencing effort on two of those populations, the Eastern Florida population and the Gulf-Atlantic population (**Table 2**). The Eastern Florida population is restricted to a ∼50 Km band along the eastern coast, extending from Jacksonville in the north to West Palm Beach in the south. Demographically, this population has remained relatively stable during the Pleistocene, with a slight signature of expansion (Manthey et al., 2016). The Gulf-Atlantic population is about 10 times smaller, although it is widely distributed from the Atlantic coast of Georgia and North Carolina to Texas in the west. It has experienced a bottleneck followed by demographic expansion (Manthey et al., 2016). This study was carried out in accordance with the recommendations of the American Veterinary Medical Association for the euthanasia of ectotherms. The protocol was approved by the Queens College Institutional Animal Care and Use Committee (Animal welfare assurance number: A32721-01; protocol number: 135).

DNA Extraction and Whole Genome Sequencing

DNA samples were retrieved from ethanol-preserved tissue and isolated with Ampure beads using the manufacturer's protocol. For each sample 200 ng of DNA was used to prepare Illumina TRU-Seq paired end libraries and sequenced on an Illumina HiSeq 2500, at the NYUAD Center for Genomics And Systems

TABLE 1 | Long Interspersed Nuclear Element clades found in the *A. carolinensis* genome.

Clades	Number of families	Number of RT hits[1]	Total number of copies in published genome[1]	Number of full-length copies in published genome[1]	Length of full length elements[1]	Number of polymorphic insertions[2]	Number of full-length polymorphic insertions[2]
R4	2	7,682	3,000	994	3.8 Kb	1,729	712
RTE	2	18,554	3,516	217	3.2–3.9 Kb	3,367	1782
CR1	4	86,802	1,594	117	4.6–5.8 Kb	27,802	2,578
L2	17	38,607	3,800	380	4.8–6.3 Kb	11,210	769
L1	20	7,441	806	170	5.2–6.8 Kb	2,508	1,089

[1] *Data from Novick et al. (2009);* [2] *This study.*

TABLE 2 | Origin of the samples sequenced, sequencing depth, and number of polymorphic insertions per individual.

Sample	Clade	Locality	Latitude	Longitude	Depth	Number of polymorphic insertions present	Number of polymorphic full-length insertions present
AC_36_1	Gulf-Atlantic	Blount, Tennessee	35.53855	−84.07625	15×	7,557	839
AC_38_4	Gulf-Atlantic	Blount, Tennessee	35.5558	−84.00245	10×	6,367	699
AC_8_13	Gulf-Atlantic	Thibodaux, Louisiana	29.797883	−90.8129	9×	6,402	629
AC_8_8	Gulf-Atlantic	Thibodaux, Louisiana	29.797883	−90.8129	16×	7,849	861
AC_27_3	Gulf-Atlantic	Darien, Georgia	31.35295	−81.447467	10×	5,626	565
AC_27_4	Gulf-Atlantic	Darien, Georgia	31.35295	−81.447467	10×	5,135	500
CC3	East Florida	Cocoa, Florida	28.243611	−80.870556	16×	9,969	863
CC8	East Florida	Cocoa, Florida	28.243611	−80.870556	16×	11,965	1,130
SB3	East Florida	South Bay, Florida	26.683333	−80.716884	12×	11,839	1,069
SB4	East Florida	South Bay, Florida	26.683333	−80.716884	8×	8,371	621
TV8	East Florida	Titusville, Florida	28.5437777	−80.9421666	8×	8,557	740
VB6	East Florida	Vero Beach, Florida	27.640278	−80.59475	10×	10,393	890
VB7	East Florida	Vero Beach, Florida	27.640278	−80.59475	9×	10,451	924

Biology Sequencing Core[1]. Sequencing was conducted twice, once to generate higher depth of coverage (two individuals per lane) and once to generate a broader sampling (four individuals per lane) at lower depth of coverage. Quality assessment was conducted using FastQCv0.11.5[2] followed by quality trimming. We used Trimmomatic (Bolger et al., 2014) to trim off low quality bases, sequencing adapter contamination and systematic base calling errors. The specific parameters we used were "trimmomatic_adapter.fa:2:30:10 TRAILING:3 LEADING:3 SLIDINGWINDOW:4:15 MINLEN:36." For the higher depth of coverage runs an average of 1,519,339,234 read pairs were generated: after quality trimming read pairs, we retained 93.3% as paired reads and 6.3% as single reads. For the lower depth of coverage runs an average of 99,464,570 read pairs were generated: after quality trimming read pairs, we retained 89.8% as paired reads and 9.9% as single reads (Supplementary Table S1). Sequencing data from this study have been submitted to the Sequencing Read Archive[3] under the BioProject designation PRJNA376071.

Sequence Alignment and SNP Calling

Surviving reads were aligned to the May 2010 assembly of the *A. carolinensis* reference genome (Broad AnoCar2.0/anoCar2; GCA_000090745.1; Alfoldi et al., 2011) and processed for SNP detection with the assistance of the NYUAD Bioinformatics Core, using NYUAD variant calling pipeline. For each sample, quality-trimmed reads were aligned to the reference genome using Bowtie2 (Langmead and Salzberg, 2012). The resulting SAM file for each individual was sorted, converted into BAM format and indexed using SAMtools (Li et al., 2009). These files were then checked for insertions, deletions and duplications using Picard tools[4] and GATK was applied for indel realignment, SNP and indel discovery and genotyping according to GATK Best Practices (DePristo et al., 2011; Van der Auwera et al., 2013). To maximize the sensitivity and confidence of variant calls, joint genotyping was conducted using GATK. To do this we first generated genomic VCF (g.VCF) files for each individual, then applied the GenotypeGVCFs command, using the previously generated g.VCF as input, to generate a group VCF file containing SNPs for the 13 genomes from the two *Anolis* populations considered here. To confirm the efficacy of this approach we selectively compared high quality genotype calls from the GATK to results from SAMtools *mpileup* (Li et al., 2009).

SNP Filtering

Our goal was to compare the frequency of polymorphic LINE insertions to the frequency of SNPs across the genome (excluding LINEs), requiring a high confidence collection of SNPs. SNPs were filtered using VCFTOOLS (Danecek et al., 2011), by applying the following criteria: a minimum Phred-score of 20, a minimum sequencing depth of 6× for each genotype, a minimum genotype quality of 20. Indels were removed and only SNPs genotyped in all individuals after quality trimming were kept for further analysis. SNPs were sampled every 1,000 SNPs to limit the effect of linkage disequilibrium while retaining enough markers for precise parameters estimation (332,839 SNPs). Options in VCFTOOLS were thus as follows: –minDP 6 –minGQ 20 –minQ 20 –max-missing 1 –min-alleles 2 –max-alleles 2 –remove-indels. Filtering might lead to biases when estimating the neutral allele frequency spectrum (Kim et al., 2011). However, our filtering criteria did not result in any strong bias in summary statistics when compared to the unfiltered VCF file, suggesting that bias in allele frequency estimates due to filtering remained limited.

Mobile Element Polymorphism Detection

To characterize LINE insertion polymorphisms, we used the Mobile Element Locator Tool (MELT[5]; Sudmant et al., 2015). MELT identifies, characterizes and genotypes polymorphic transposable element insertions and has been used successfully for extensive analyses of LINE and SINE polymorphisms in the human genome (1000 Genomes Project Consortium et al., 2015; Sudmant et al., 2015). MELT exhibits high precision and recall of LINE insertions in low depth of coverage genomes (Rishishwar et al., 2016). MELT identifies the presence and absence of insertions based on the appearance of target mobile element sequence in split or discordant reads. For our analyses we selected target sequences from previously described, potentially active LINE families from the CR1, L1, L2, R4 and RTE clades (Novick et al., 2009). These sequences were identified based on the presence of a characteristic reverse transcriptase domain using Genome Parsing Suite software (McClure et al., 2005), exist as full length copies in the *Anolis* reference genome and exhibit low divergence (typically less than 2% divergence between copies and consensus sequence), indicative of recent activity by members of these groups (Novick et al., 2009). Previously published consensus sequences for these elements were collected from Repbase (Bao et al., 2015) to be used as target sequences, and cleared of ambiguities, when they occurred, by direct comparison to full-length genomic copies. Based on the low divergence exhibited by these groups (Novick et al., 2009), and our intention to generate a conservative estimate, we selected an acceptable error rate of 2%.

Mobile Element Locator Tool operates on BWA-aligned re-sequenced genomes, so for each *Anolis* sample, quality-trimmed FastQ reads were aligned to the AnoCar2.0 genome using the BWA-mem short read alignment approach (Li and Durbin, 2009). Each BWA-aligned sample genome was sorted and converted to BAM format using Samtools (Li et al., 2009). The MELT Preprocess software was then run on each sample genome BAM file to prepare it for analysis. For our analyses we used the MELT-SPLIT pathway, which consists of four runtime stages: individual analysis (IndivAnalysis), group analyses (GroupAnalysis), genotyping (Genotype) and VCF file construction (makeVCF). Individual analyses identify evidence of target element insertions in BAM files. Results from individual analyses are merged during group analysis, and the

[1] http://nyuad.nyu.edu/en/research/infrastructure-and-support/core-technology-platforms.html

[2] http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[3] https://www.ncbi.nlm.nih.gov/sra

[4] http://broadinstitute.github.io/picard/

[5] http://melt.igs.umaryland.edu/

pooled data is used to produce improved calls regarding each insertion, including breakpoints, insertion length, strand, and target site duplication. Genotyping is conducted on each genome individually to determine its genotype for each polymorphic locus. Finally, the data from individual genotyping are merged to form a VCF file for the population. For each of the 41 specific LINE subgroup consensus sequences, every BWA-aligned and preprocessed genome was analyzed and used to produce VCF files for individuals from the East Florida and Gulf Atlantic *Anolis* populations. These files were then combined and filtered to remove any polymorphic loci that failed to exhibit coverage in all samples or exhibited low quality calls. Where duplicate calls occurred (i.e., when multiple LINE insertions of different families occurred within 50 bp from each other) only the longest was kept in the VCF file. This study focused exclusively on the presence and predicted length of polymorphic LINE insertions and at no point do we analyze or discuss mutations occurring within these insertions since it is nearly impossible to match a SNP within a LINE with its specific genomic location.

Descriptive Statistics

We used several statistics to describe the allele frequency spectra and allele sharing between populations, of both SNPs and LINE insertion polymorphisms. Tajima's D (Tajima, 1989) is a statistic that is commonly used to detect selection. It reflects the difference between θw and π, which are two different estimators of the effective population size scaled by mutation rate ($4Ne\mu$) that should be positively correlated under neutrality. At mutation-drift equilibrium, the expected value of Tajima's D is zero, while positive values indicate population reduction or balancing selection, and negative values indicate population expansion or purifying and positive selection. We computed the mean number of pairwise differences for the whole dataset and each population, as well as the number of private and fixed polymorphisms. We also computed the mean F_{ST} between populations for each category of markers. These statistics were calculated using VCFTOOLS (Danecek et al., 2011) and the R package PopGenome (Pfeifer et al., 2014). An element was considered as complete if its size was at least 90% of the maximum size for its family. The vcflib script vcffilter[6] was used to split VCFs between complete and truncated elements for each family.

Demographic Parameters Estimation from SNPs

To assess whether LINE variation deviated from a neutral model, we estimated the demographic history of the two populations using the SNP dataset. We fitted a model of isolation with migration, allowing for one population size change in each derived population. Time since divergence between the two species was fixed at 1.34 Mya (Tollis et al., 2012). Parameters were estimated from the joint allele frequency spectrum (SFS) using the likelihood approach implemented in fastsimcoal2.5 (Excoffier et al., 2013). Parameters with the highest likelihood were obtained after 40 cycles of the algorithm, starting with 50,000 coalescent simulations per cycle, and ending with 250,000 simulations. This procedure was replicated 100 times and the set of parameters with the highest final likelihood was retained.

We estimated 95% confidence intervals (CI) by simulating coalescence under the best model for the same number of SNPs as in the original dataset. We performed parameter estimation for 150 of these pseudo-observed datasets to infer CI. Coalescence simulations were performed using fastsimcoal2.5 (Excoffier and Foll, 2011). We further checked whether our model fit the observed data by sampling parameters from the 95% CI range for 10,000 simulations and comparing observed and simulated datasets. We summarized allele frequency spectra using Principal Components Analysis [gfitpca function in the R package abc (Csillery et al., 2012)].

Simulations and Deviation from Neutral Expectations

To estimate if the LINE SFS deviated significantly from neutral expectations, we simulated for each family the derived allele frequency spectrum in fastsimcoal2.5. Parameters were sampled from the CI obtained for SNPs. We performed 5,000 simulations for each dataset, assuming unlinked LINE insertion sites, and obtained *p*-values from the comparison between the observed Tajima's D or F_{ST} value to the distribution obtained under a neutral model. We also performed a non-parametric bootstrap on the actual SNP dataset and extracted random sets of 100–500 SNPs along each chromosome, computing Tajima's D and comparing the resulting distribution to the values observed for LINEs.

RESULTS

LINE Insertion Polymorphisms are Numerous and Their Abundance Varies by Clade

We sequenced six *A. carolinensis* genomes from the Gulf-Atlantic population and seven from the East Florida population with a sequencing depth of coverage ranging from 8 to 16× after alignment to the reference genome (**Table 2**). We detected extensive LINE insertion polymorphism in both populations (summarized in **Tables 3**, **4**) with a total of 46,616 polymorphic insertions across the 13 individuals. The East Florida population appears to maintain a greater total number of LINE polymorphisms, with a mean of 10,022 polymorphic LINE insertions per individual (from 8,371 to 11,965 insertions). In the Gulf Atlantic population the mean number of insertions per individual was substantially lower at 6,489 (from 5,135 to 7,849 insertions). Across all genomes roughly 10% of all polymorphic insertions approximated their full length, though individual populations varied: for individuals from the Gulf-Atlantic, 10.5% of polymorphic LINE insertions were full length (4,093 out of 38,936), while in the East Florida population only 8.7% (6,237 out of 71,545) were full length.

The five clades of LINEs investigated (R4, RTE, CR1, L1, and L2) exhibited notable variation in their success at generating new insertions (**Tables 1**, **3**). The most successful group was the

[6] https://github.com/vcflib/vcflib

TABLE 3 | Summary statistics for all LINE clades, families and subgroups considered in this study.

		Mean number of differences in polymorphic insertions			Tajima's D		Number of polymorphic loci	% of private insertions		% of fixed differences	% of shared differences	Mean F_{ST}
Dataset		**All**	**Florida**	**Gulf-Atl**	**Florida**	**Gulf-Atl**		**Florida**	**Gulf-Atl**			
SNPs		0.21	0.22	0.36	−0.62	0.47	314575	60.25	15.85	0.19	23.72	0.12
L1	All	0.15	0.22	0.31	−1.39***	−0.48**	2508	65.67	19.46	0	14.87	0.04**
L1_AC1 to 16	FL	0.15	0.21	0.32	−1.46***	−0.13	454	71.81	18.5	0	9.69	0.04*
	TR	0.18	0.25	0.30	−0.95	−0.5**	1062	59.13	15.35	0	25.52	0.04***
L1_AC17 to 20	FL	0.11	0.17	0.28	−2.06***	−0.78***	635	68.82	27.09	0	4.09	0.03*
	TR	0.14	0.20	0.31	−1.6***	−0.24*	357	71.71	19.33	0	8.96	0.04
L2	All	0.15	0.23	0.28	−1.27***	−0.74***	11210	61.06	23.76	0	15.18	0.05**
	FL	0.13	0.20	0.28	−1.65***	−0.75***	769	67.1	25.1	0	7.80	0.04***
	TR	0.15	0.23	0.28	−1.24***	−0.74***	10440	60.61	23.66	0	15.73	0.05**
CR1	All	0.15	0.22	0.31	−1.31***	−0.29*	27802	70.35	18.02	0.02	11.62	0.05
	FL	0.14	0.21	0.30	−1.51***	−0.49**	2578	68	23.27	0	8.73	0.05
	TR	0.16	0.22	0.31	−1.29***	−0.27*	25224	70.59	17.48	0.02	11.91	0.05
R4	All	0.17	0.24	0.25	−1.04*	−1.1***	1729	49.1	20.76	0	30.13	0.03***
	FL	0.16	0.23	0.25	−1.16**	−1.21***	1017	47.79	20.94	0	31.27	0.02***
	TR	0.18	0.25	0.27	−0.87	−0.93***	712	50.98	20.51	0	28.51	0.04***
RTE-1	All	0.11	0.18	0.23	−1.91***	−1.42***	2853	62.57	33.16	0	4.28	0.02***
	FL	0.11	0.18	0.22	−2.00***	−1.52***	1774	61.72	35.17	0	3.10	0.02***
	TR	0.12	0.19	0.24	−1.77***	−1.23***	1079	63.95	29.84	0	6.21	0.02***
RTEBovB	All	0.25	0.31	0.34	−0.08^{+}	0.06	514	37.74	12.84	0	49.42	0.05***
	FL	0.27	0.38	0.33	0.76	−0.06	8	25	25	0	50.00	0.14
	TR	0.25	0.31	0.34	−0.1	0.06	506	37.94	12.65	0	49.41	0.05

*Mean pairwise divergence was computed only for loci polymorphic within a population and represents the average number of differences in polymorphic insertions or SNPs computed for all pairs of individuals (equivalent to nucleotide diversity). The number of private, shared, and fixed polymorphisms are provided as proportions of sites polymorphic in the whole sample. For each group for which coalescence simulations were performed, we provide a one-tailed p-value based on where the observed value for the statistics fell in the simulated distribution. We did not perform separate simulations for truncated and complete elements in RTEBovB due to the low number of polymorphic complete insertions. All, All elements; FL, full-length; TR, truncated; +, Tajima's D fell in the 0.1% upper tail of the distribution. *p-value < 0.05; **p-value < 0.01; ***p-value < 0.001.*

CR1 clade, for which we found 27,802 polymorphic insertions. The L2 clade also has a large number of insertions: 11,210. Far fewer polymorphisms were found for the remaining families: the RTE clade had 3,367 polymorphisms, the L1 clade 2,508, and the R4 clade 1,729. Within each clade we also found substantial differences in the success of active families (**Table 4**). The L1 clade consists of 20 highly divergent families (Novick et al., 2009; Boissinot and Sookdeo, 2016). We used the consensus sequence for each of these families to search for polymorphisms and found a highly uneven fraction of polymorphic insertions across these families. No polymorphisms were found for three L1 families (L1AC03, L1AC10, and L1AC18), indicating these families are inactive in the populations we studied (**Table 4**). Most families had polymorphic insertions numbering less than 100, however, two families appeared at much higher numbers: L1AC07, which had 532 polymorphic insertions, and L1AC17, which had 763 polymorphic insertions. Together, these two families account for the majority (52%) of all L1 polymorphic insertions we identified. The L2 Clade has 17 known families in the *Anolis* genome but their differences in replicative success were not as large as those in the L1 clade. All L2 families exhibit polymorphisms and the most frequent group, L2AC09, only constitute 14% of L2 insertions. The RTE clade has only two representatives, RTE-1 and the ancient RTEBovB family. There are nearly six times more RTE-1 polymorphisms than RTEBovB (2853 versus 514, respectively), which is consistent with the idea that RTEBovB may be extinct in *Anolis*. The two R4 and the four CR1 families previously described (Novick et al., 2009) are nearly identical in sequence over most of their length and it was not possible to distinguish them using this dataset.

Our prior expectations for the complement of LINE insertions have been shaped in part by published analyses conducted on the *Anolis* genome assembly using GPS-RT (McClure et al., 2005) and by BLAST searches using the 3′ termini of consensus sequences (Novick et al., 2009). Those two earlier analyses were conducted on a single sequence assembly, representing an individual. We compared our results to the results of these earlier analyses to assess how much LINE-generated polymorphisms there are in natural populations relative to the reference genome. The number of polymorphic CR1 insertions we identified is more than 17-fold the total number of insertions from the BLAST search of the reference genome (**Table 1**). This discrepancy is best explained by the large number of severely truncated insertions (<50 bp) that could have been missed by the BLAST search (which used the entire 3′UTR). The number of polymorphic insertions from the L2 clade is slightly less than threefold the number of insertions in the published genome. This is similar to L1, which has just over threefold more polymorphic insertions than insertions in the reference genome, though L1 has far fewer total insertions than L2 (2,500 L1 versus 11,000, respectively). Roughly the same number (∼3,300) of polymorphic RTE insertions were found as were previously detected by BLAST, and the R4 clade was found to have less than half as many polymorphic insertions as insertions identified by BLAST. These differences among clades possibly reflect differences in the fractions of fixed insertions relative to polymorphic ones among clades, which could be due to differential chance of fixation or to different timing of amplification of the LINE clades during the evolution of *A. carolinensis*. The number of RT hits detected by GPS are 3–5 times higher than the number of polymorphisms

TABLE 4 | Copy numbers of L1 and L2 families.

L1 Clade		L2 clade		RTE clade	
Families	**Copy number**	**Families**	**Copy number**	**Families**	**Copy number**
L1AC01	68	L2AC01	507	RTE-1	2853
L1AC02	18	L2AC02	336	RTEBovB	514
L1AC03	0	L2AC03	301		
L1AC04	43	L2AC04	504		
L1AC05	27	L2AC05	276		
L1AC06	87	L2AC06	569		
L1AC07	532	L2AC07	543		
L1AC08	95	L2AC08	1424		
L1AC09	82	L2AC09	1661		
L1AC10	0	L2AC10	131		
L1AC11	90	L2AC11	720		
L1AC12	52	L2AC12	206		
L1AC13	103	L2AC13	948		
L1AC14	85	L2AC14	256		
L1AC15	181	L2AC15	1177		
L1AC16	53	L2AC16	388		
L1AC17	763	L2AC17	1263		
L1AC18	0				
L1AC19	23				
L1AC20	206				

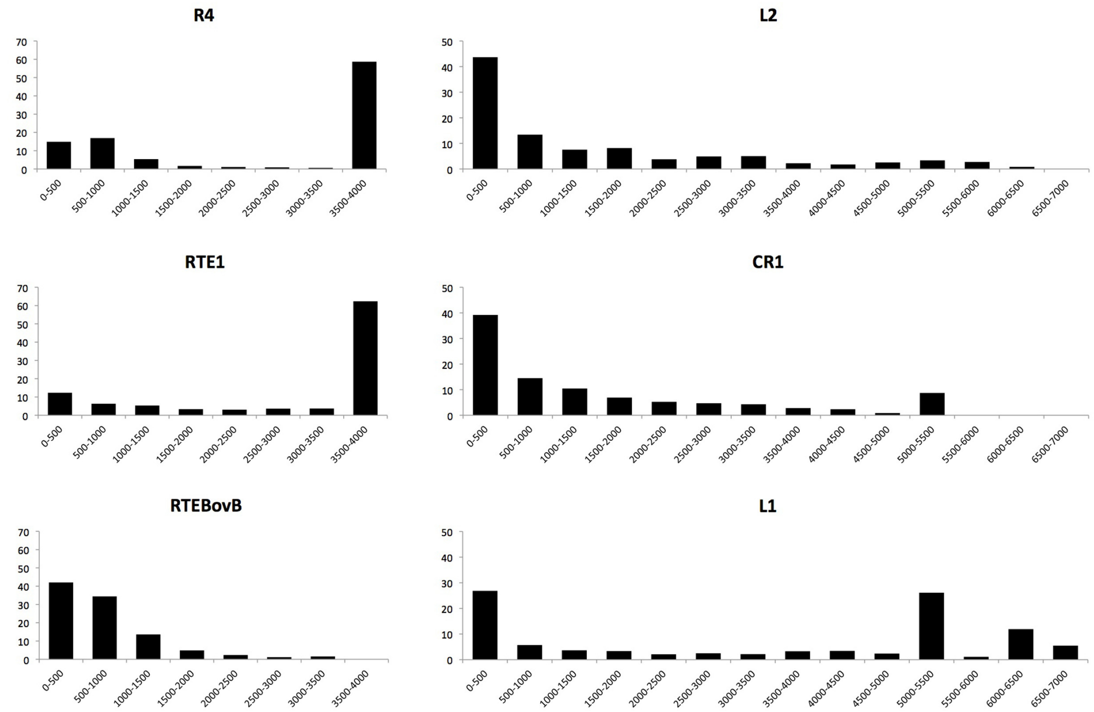


FIGURE 1 | Length distribution of polymorphic LINE insertions. The majority of R4 and RT1 insertions approach full-length, while all other LINEs were primarily found as truncated copies. Note that the L1 clade has multiple subfamilies whose full-length copies vary in length and this results in the multiple peaks at the right of the length distribution.

but the numbers are roughly proportional in the sense that the clades with the largest number of polymorphisms (CR1 and L2) are also the clades with the most RT hits. This difference in the total number of counts probably reflects the ability of GPS to identify the entire complement of RT-containing elements, including ancient elements that have long been fixed in the *Anolis* genome.

LINE Clades Show Distinct Patterns of Insertion Length and Success

The total number of polymorphic insertions found for each clade is not directly related to the number of full-length insertions. In most clades (with the notable exception of RTE) more truncated than full-length elements were found. All the truncated elements had their 3′ extremity and were truncated in 5′. This pattern is typical of LINEs and is caused by premature termination of the reverse-transcription reaction at the site of insertion (Ostertag and Kazazian, 2001; Martin et al., 2005). The CR1 clade has the largest number of insertions but the fraction of full-length CR1 insertions is less than 10%. For the L2 clade, which is also abundant, less than 7% of these insertions were full-length (769). In contrast, the majority (53%) of the RTE insertions are full-length and ∼40% of the L1 and R4 insertions are complete. It is unlikely that the differences we observe result from differences in the length of LINEs. L1 consensus sequences are the longest (5.2–6.8 kb) whereas the R4 consensus is substantially shorter (3.8 kb), yet the same fraction of insertions is full-length in these two clades. The consensus sequences of L1 and L2 are of similar length but the fraction of full-length insertions is six times larger for L1 than for L2. These differences are likely due to variations in the mode of truncation of the elements at the time of insertions. **Figure 1** depicts the length distribution of the different clades. It shows that truncation in R4 and RTE1 can occur anywhere along the length of the element but a large fraction of the elements are transposed all the way to their 5′ end. The probability of truncation in CR1 and L2 decreases proportionally to the distance to the 3′ end and a minority of the elements insert as full-length. L1 elements either truncate early on during transposition (and don't reach 1 Kb), or if they do, they tend to be complete, hence the large fraction of elements longer than 5 Kb. It should be noted that complete elements fall into two length categories: elements between 5 and 5.5 Kb and elements longer than 6 Kb. These two types correspond to two sub-clades of L1, the families with short (∼230 bp long) 5′UTR (families L1AC16 to 20) and the families with long (800–1,500 bp) 5′UTRs (families L1AC1 to 15) (Boissinot and Sookdeo, 2016). Finally, the RTEBovB family contains a very small number of full-length elements, which is probably related to the fact that this family is on its way to extinction.

Most Polymorphic LINE Insertions Exist at Low Frequencies

Strikingly few insertions occurred at high allelic frequencies or are fixed in either population (**Figure 2A**). We found 16

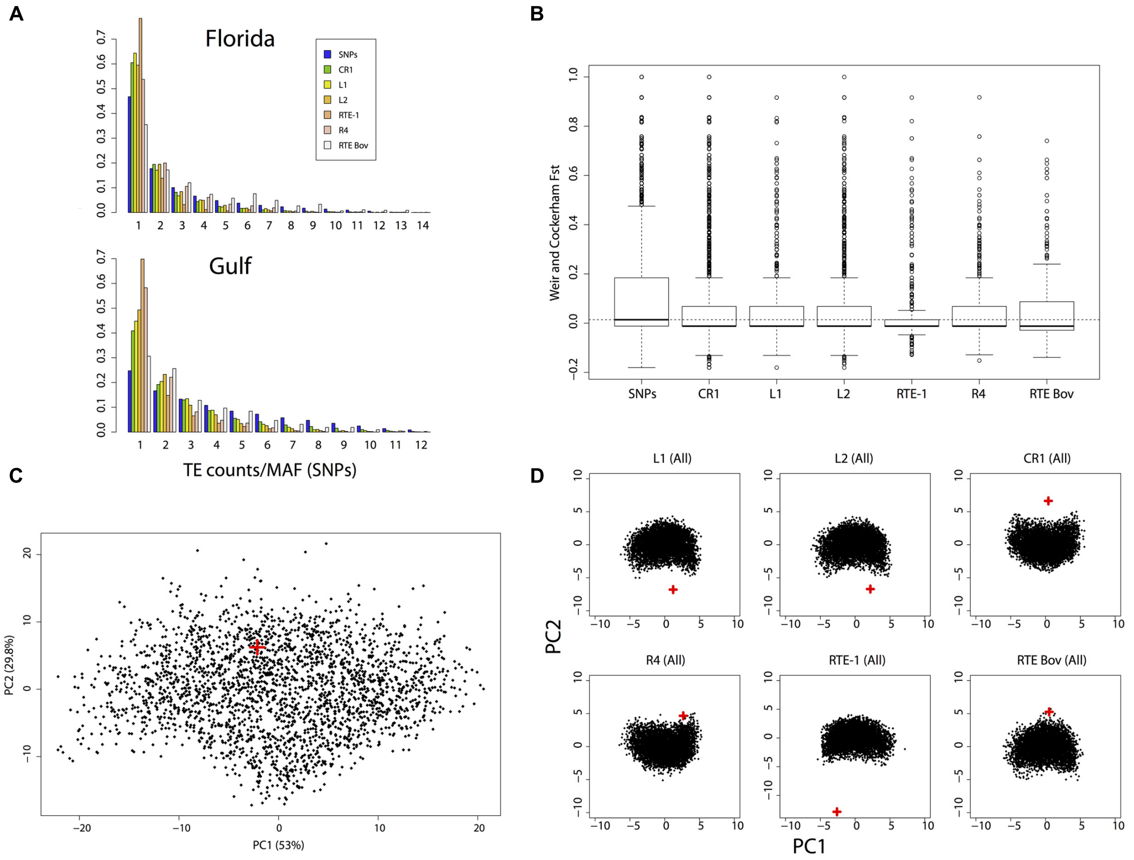


FIGURE 2 | Summary of allele frequency spectra and simulations. (A) Allele frequency spectra for SNPs and LINE insertions in the East Florida and Gulf-Atlantic populations. For SNPs, the frequency of the minor allele in each population was considered. **(B)** F_{ST} distribution for SNPs and transposons clades. The dotted line represents the median for SNPs. **(C)** Principal Component Analysis (PCA) summarizing the joint allele frequency spectrum for SNP simulations. **(D)** PCA obtained after simulating insertion polymorphism in the six main clades. For all PCAs, the red crosses indicate the predicted position of the observed dataset.

LINE insertions that were fixed in the East Florida population but absent in the reference genome (12 CR1, three L2 and one R4), and 28 LINE insertions that were fixed in the Gulf population but absent in the reference genome (27 CR1, four L2, two R4 and one L1). Only two insertions were found to be fixed across all the genomes sequenced here but absent in the reference sequence and both were from the CR1 clade. The site frequency spectrum (SFS) of insertions is consistently skewed toward low frequencies when compared to SNPs' minor allele frequencies (**Figure 2A**), which we used as a proxy for the "neutral" demographic history of the two populations. The only exception to this pattern is RTEBovB, where insertions at intermediate and high frequencies were more common in both populations. The skew in SFS was captured by Tajima's D, which takes negative values for all categories of LINEs and for both populations, and average pairwise differences over the two populations, which were almost always lower for LINE insertions than for SNPs (**Table 3**). These two statistics are consistent with there being an excess of singletons and rare variants. This pattern was especially strong for RTE-1 and R4 clades in the Gulf population (**Figure 2A**), with a significant reduction in the mean number of pairwise differences even compared with other LINE clades (pairwise comparisons, Wilcoxon rank sum tests, all $P < 1.7 \times 10^{-6}$). This reduced polymorphism was also reflected by the lower F_{ST} values observed for insertions when compared to SNPs (**Figure 2B**). The proportion of alleles found exclusively in Florida (private alleles) was higher than the proportion of private alleles in the Gulf-Atlantic population (Wilcoxon signed rank test on all subgroups in **Table 3**, $V = 91$, $p = 2.4 \times 10^{-4}$), suggesting a reduced genetic diversity in the Gulf population. Similarly, Tajima's D was consistently higher in the Gulf population ($V = 88$, $P = 3.3 \times 10^{-3}$). This pattern was, however, not observed for RTEBovB, which displayed a higher proportion of shared alleles between populations than the other LINEs analyzed here.

Polymorphic LINE Insertions are Negatively Selected

Estimates of current effective population sizes assessed using the SNP dataset confirmed a large Florida population (diploid population size), and a smaller (but still large) population in Gulf-Atlantic (see **Table 5** for more details). This pattern is consistent with the higher number of polymorphic sites

TABLE 5 | Summary of parameters (in demographic units) estimated with fastsimcoal2.5.

Parameter	2.50%	Maximum Likelihood estimate	97.50%
Ancestral size (Gulf)	379795	1422722	8838592
Ancestral size (Florida)	366002	751115	1756393
Ancestral size (All)	564492	1167977	1488644
Current size (Florida)	1959085	3316203	4603720
Current size (Gulf)	101238	235789	351645
Time since size change (Gulf)	57331	274157	559121
Time since size change (Florida)	275163	802462	1110215
Migration rate (Gulf from Florida)	2.96E-07	3.94E-07	5.51E-07
Migration rate (Florida from Gulf)	2.19E-07	3.38E-07	9.00E-07

Parameters for modeling insertions were sampled from a uniform distribution bounded by the 95% CI.

observed in Florida for all markers. Simulated joint SFS based on the demographic model inferred from SNPs were consistent with the observed SFS (**Figure 2C**), suggesting a good fit of the model. Summary statistics obtained from simulations displayed more negative values for Tajima's D than the ones obtained from random sampling of 100–500 SNPs across the genome. This suggests that our model is conservative for detecting signatures of purifying selection under insertion/drift equilibrium. Nonetheless, observed SFS for LINE insertions never matched the simulations (**Figure 2D**) and the simulated summary statistics such as F_{ST} or Tajima's D were generally larger than the observed ones (**Table 3**). Again, the only exception to this pattern was RTEBovB, which even displayed a higher Tajima's D than expected in Florida.

Since previous studies in other organisms have determined that complete elements are found at lower frequencies than truncated ones, we compared the frequency of these two types of elements. We assessed whether there was any difference between these two categories by comparing Tajima's D, F_{ST} and the mean number of pairwise differences between truncated and complete elements (**Figure 3**). In Florida, Tajima's D was significantly skewed toward more negative values for complete elements than for truncated ones (26 polymorphic families, $V = 69$, P-value $= 5.6 \times 10^{-3}$). The average pairwise differences were consistent with this pattern, being always significantly lower for complete elements than for truncated elements in Florida (**Table 6**). In the Gulf-Atlantic population, the values for Tajima's D tend to be lower for full-length CR1, R4 and RTE1 than for truncated ones, but those differences are not significant. However, the average pairwise differences were significantly different between full-length and truncated elements RTE-1, R4 and CR1 but not for L1 and L2 (**Table 6**).

DISCUSSION

Using whole genome resequencing data, we investigated the population dynamics of polymorphic LINEs in the lizard *A. carolinensis*. We found that LINEs generate a considerable amount of structural polymorphism in this species, in excess of 45,000 insertions, including close to 7,000 full-length elements. This is considerably more than the 998 polymorphic L1 insertions identified by the 1,000 genomes project in the global human population (Stewart et al., 2011) but similar to the number of LINE polymorphisms (∼40,000) found across 17 classical and wild derived mouse strains, which evolution roughly covers a similar time span (∼2 my) (Nellaker et al., 2012). The number of polymorphisms detected here is about four times larger than the number of copies detected in the published genome, which is consistent with the idea that most insertions do not reach fixation (Novick et al., 2009). Additionally, it is important to note that the estimates of LINE polymorphism presented here is likely a conservative one. There are several reasons for this: we used stringent criteria (a maximum of 2% divergence) when identifying LINE insertions, greater depth of coverage could potentially improve the sensitivity of our analyses, and our approach assumes that all insertions in the reference genome are fixed. Together, this will bias our analyses against the identification of rare or degenerate LINE insertions, however, a reduction in this bias would only further support the observations and conclusions described here.

We report substantial differences in the replicative success of LINEs in anoles (**Table 1**). CR1 accounts for more than half of these insertions, followed in abundance by L2, RTE, L1 and R4. Interestingly the total number of insertion generated by a specific clade is not related to the number of potential progenitors. For instance 62% of the ∼2,850 RTE1 insertions, 41% of the ∼1,730 R4 and 43% of the ∼2,500 L1 are complete whereas only 7% of the ∼11,210 L2 and 9% of the ∼27,800 CR1 are complete. This pattern is clearly related to the probability of truncation of LINEs (**Figure 1**). These different patterns of truncation are indicative of variations in the processivity of the reverse-transcription reaction among clades that will need to be further explored experimentally. The inverse relation between copy number and fraction of complete elements suggests that clades are using different strategies to ensure their long-term success. Elements that have a low probability of generating full-length copies (CR1 and L2) tend to generate a much larger number of insertions, increasing the odds that some of these insertions are full-length and potential progenitors. By analogy with the field of ecology, this strategy would be similar to a species with an *r* reproductive strategy, i.e., a strategy where many offspring are produced thus compensating for the low survival to adulthood. In contrast, there is no pressure for L1, RTE1 and R4 to produce a large number of copies since many of the new insertions will be full-length and capable of further transpositions. This is similar to the K strategy where the number of offspring is limited but their chance to propagate the species is high.

In all clades and families examined (with the notable exception of RTEBovB which is discussed below), polymorphic LINE insertions were found at very low frequency and the vast majority were observed from only a single chromosome in our sample. We also showed that the frequency distribution of LINE polymorphisms is significantly skewed toward lower values than the SNP distribution, which presumably reflects the effect of purifying selection acting on LINEs. In addition, we found this

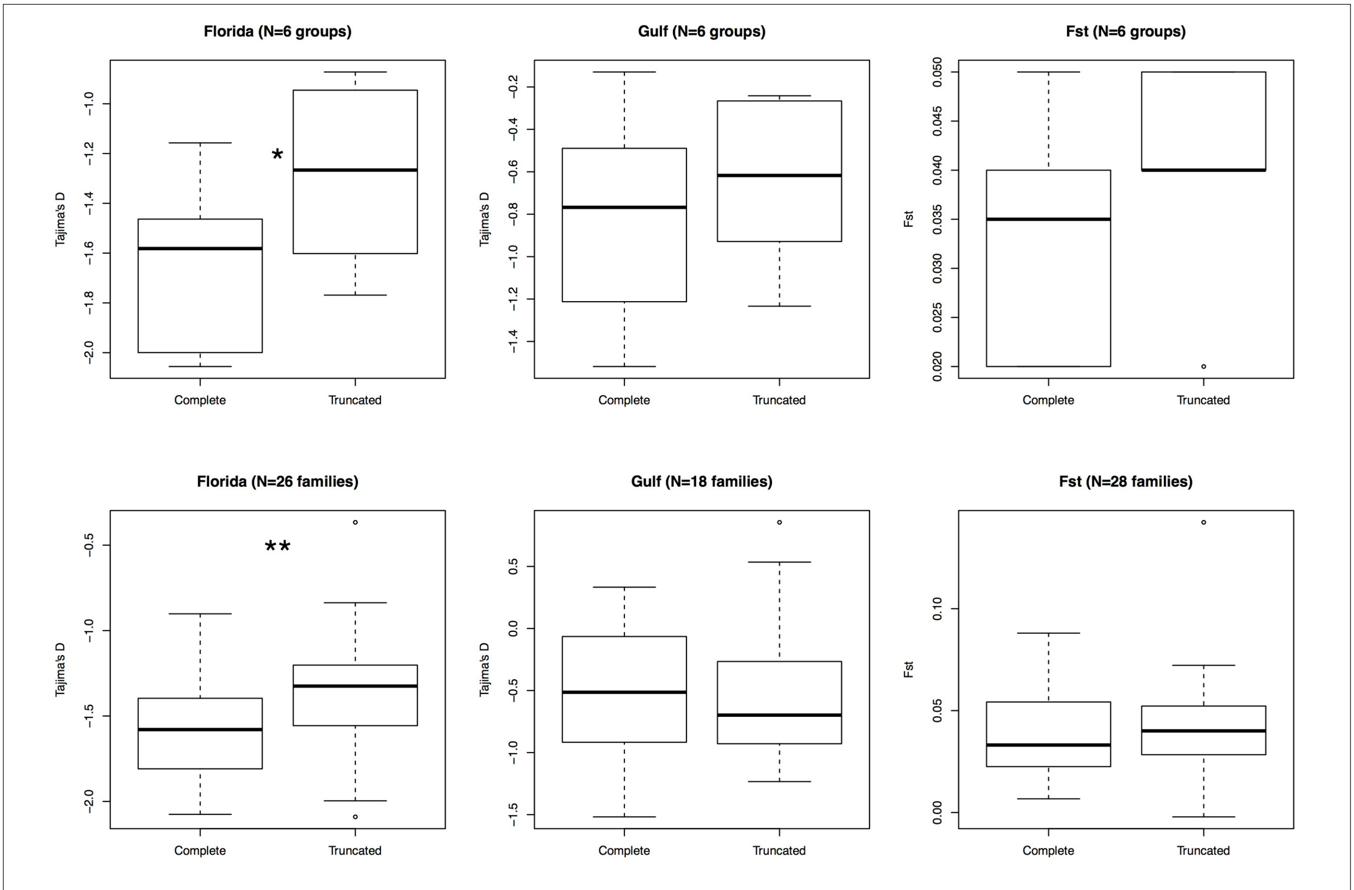


FIGURE 3 | Comparison of variation between truncated and complete elements. Six groups were considered: L1 (families AC1 to 16), L1 (families 17 to 20), L2, R4, RTE-1, and CR1 (top row). A more detailed analysis comparing all families within clades is also displayed (bottom). Only polymorphic families with at least 10 polymorphic sites in a population were retained. Wilcoxon signed rank test; **P*-value < 0.05, ***P*-value < 0.01.

TABLE 6 | Comparison of the mean number of pairwise divergence for complete and truncated elements in the two populations.

Clade	Florida, complete	Florida, truncated	W summary statistics	*P*-value	Gulf, complete	Gulf, truncated	W summary statistics	*P*-value
CR1	**0.209**	**0.225**	**19360000**	**6.41E-07**	**0.297**	**0.313**	**2865600**	**0.001878**
L1 (AC 1 to 16)	**0.213**	**0.249**	**142220**	**6.79E-06**	0.322	0.296	30660	0.06041
L1 (AC 17 to 20)	**0.172**	**0.203**	**57887**	**2.30E-05**	0.276	0.314	8754	0.05928
L2	**0.200**	**0.229**	**2043000**	**5.99E-07**	0.278	0.280	517370	0.8827
R4	**0.234**	**0.254**	**210700**	**0.01054**	**0.246**	**0.266**	**85305**	**0.02438**
RTE01	**0.176**	**0.192**	**403250**	**0.0001458**	**0.225**	**0.245**	**122830**	**0.01861**

Significance for each comparison between truncated and complete elements was assessed using a Wilcoxon rank sum test. Significant comparisons are highlighted in bold.

skew to be more pronounced for the Floridian population than for the Gulf-Atlantic population and for long elements than for the truncated ones. Purifying selection efficiently prevents the fixation of LINE insertions in anoles because the effective population size of extant and ancestral anole populations is large, ranging from ~236,000 individuals for the extant Gulf-Atlantic population to ~3,332,000 for Florida (**Table 5**). Under such demographic conditions, the chance of fixation of a novel insertion, deleterious or neutral, is very low (Gonzalez and Petrov, 2012). In fact, the observation that more private alleles are detected in Florida than in the Gulf population (as well as a higher proportion of polymorphic sites, and a SFS skewed toward low frequencies and singletons) is consistent with Florida's larger population size compared to the Gulf population (Tollis et al., 2012; Tollis and Boissinot, 2014; Manthey et al., 2016) and is suggestive of a stronger effect of drift on the Gulf-Atlantic population, as previously noted (Tollis and Boissinot, 2013). Thus, the low frequency distribution of LINEs in *A. carolinensis* results both from the effect of selection and a large effective population size. However, previous studies have shown that a number of insertions present in the published genome sequence are fixed (Tollis and Boissinot, 2013). Under the current demographic conditions, it is unlikely that the fixation of the elements occurred recently. Instead it is plausible that these insertions reached fixation when the effective population size of *A. carolinensis* was smaller, possibly at the time of the colonization of North America from Cuba (Glor et al., 2005). Comparison of LINE polymorphisms with genomic sequence from the Cuban species *A. porcatus* and *A. allisoni* will be necessary to answer this question.

The case of RTEBovB is unique among the LINEs analyzed here because it exemplifies the dynamics of a family going extinct. This family is mostly constituted of truncated elements, and is likely ancient (Novick et al., 2009). It displays the highest proportion of shared alleles (49.42%), suggesting that many insertions rose to relatively high frequencies even before the split between populations. It is also the only family for which we observed a higher Tajima's D than expected, possibly due to ancient demographic variation that is not even captured by the SNPs. The observed pattern is thus consistent with the age of the family and suggests that these elements were not eliminated by selection.

The excess of singletons and the general lower frequency of LINE polymorphisms than SNPs suggest that LINEs are negatively selected and constitute a genetic load for their host. This pattern is consistent with the very low divergence calculated between elements from the same family (Novick et al., 2009; Tollis and Boissinot, 2013) and supports a turnover model in which insertions rarely reach fixation and in which novel insertions are eliminated from the population as new insertions are generated. We also determined that the intensity of selection is stronger against complete elements. This is in line with previous studies in human, fruit fly, and stickleback populations, which showed that selection against TEs is length dependent (Petrov et al., 2003; Boissinot et al., 2006; Blass et al., 2012). However, truncated elements are also found at lower frequency in the populations than expected under neutrality (**Table 3**) suggesting that they are negatively selected. This result contrasts with studies in humans where truncated insertions were shown to behave like neutral alleles (Boissinot et al., 2006). Thus, the negative effect of LINEs does not seem to be limited to long elements in *Anolis*, although those seem to be more deleterious. It was proposed that the deleterious effect of LINEs in vertebrates result mostly from their ability to mediate ectopic recombination leading to chromosomal rearrangements (Furano et al., 2004; Boissinot et al., 2006; Song and Boissinot, 2007; Tollis and Boissinot, 2013), and our observation that complete elements are under stronger purifying selection than truncated ones supports this model. However, the lower frequency of truncated insertions compared with SNPs raises the possibility that ectopic recombination in anoles could also involve short elements, thus providing support to the hypothesis that ectopic recombination may not be as tightly regulated in non-mammalian vertebrates as it is in mammals (Furano et al., 2004; Novick et al., 2009; Tollis and Boissinot, 2013), and that LINEs may impose a stronger genetic load on reptile genomes than they do in mammals.

An alternative explanation for the observed excess of singletons is a departure from transposition-selection equilibrium. Our coalescence simulations implicitly assume a constant mutation/transposition rate. However, it has been shown that transposable elements can go through bursts of transposition, leading to an excess of insertions having the same age. Thus, a recent burst of transposition can also lead to an excess of recent insertions compared to the expectation under equilibrium, even if LINEs are not under purifying selection (Bergman and Bensasson, 2007; Blumenstiel et al., 2014). However, we observed an excess of singletons across all clades (except RTEBovB), which should not be the case unless all families went through a recent, coordinated burst in both populations. In addition, most clades display elements that are shared between the two populations, and were therefore present in the ancestral population, suggesting that the low frequency of these polymorphisms is not caused by a very recent burst. However, differences in the rate of transposition cannot be fully excluded and could contribute to some of the differences we observe. For example, the RTE1 family, which shows the most negative values of Tajima's D and the most skewed frequency distribution, is also the one with the smallest fraction of shared polymorphism, suggesting that a recent increase in the rate of transposition could contribute to the excess of singletons in this family. From this perspective, the inclusion of other *A. carolinensis* populations should help characterize the extent of shared polymorphism at the species scale, allowing us to better evaluate the likelihood of recent bursts of activity in distinct populations.

Even if non-equilibrium explanations for the excess of rare insertions are considered unlikely (Petrov et al., 2011; Barron et al., 2014), neutral models would benefit from new ways to model the transposition process and provide even more conservative assessments of either negative or positive selection (Bergman and Bensasson, 2007). Future studies should focus in more detail on the relationship between TE frequencies and genomic features such as recombination hotspots, coding and intergenic regions. Combining information about TE position and SNP variation in regions flanking insertion sites is also a powerful way to detect TEs under selection, and should provide fundamental insights into the dynamics of transposable elements in *Anolis* and vertebrates in general.

AUTHOR CONTRIBUTIONS

RR and SB designed the project. RR and YB analyzed the data. SB and YB prepared the original artwork. RR, YB, and SB wrote the manuscript. All authors have made substantial intellectual contributions to the research project and approved the final manuscript.

FUNDING

This research was supported by New York University Abu Dhabi Research funds AD180 (to SB). NYUAD Sequencing and Bioinformatics Cores are supported by NYUAD Research Institute grant G1205-1205A to the Center for Genomics and Systems Biology at NYUAD.

ACKNOWLEDGMENTS

We thank the NYUAD High-Performance Computing, the NYUAD Core Technology Platforms, the NYUAD high throughput Sequencing Core, and the NYUAD Bioinformatics Core for support. We thank Marc Arnoux and Mehar Sultana for helping with the sequencing. We thank Nizar Drou and Jillian D. Rowe for helping with the bioinformatics analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00044/full#supplementary-material

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Alfoldi, J., Di Palma, F., Grabherr, M., Williams, C., Kong, L., Mauceli, E., et al. (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* 477, 587–591. doi: 10.1038/nature10390

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* 6, 11. doi: 10.1186/s13100-015-0041-9

Barron, M. G., Fiston-Lavier, A. S., Petrov, D. A., and Gonzalez, J. (2014). Population genomics of transposable elements in *Drosophila*. *Annu. Rev. Genet.* 48, 561–581. doi: 10.1146/annurev-genet-120213-092359

Bergman, C. M., and Bensasson, D. (2007). Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11340–11345. doi: 10.1073/pnas.0702552104

Blass, E., Bell, M., and Boissinot, S. (2012). Accumulation and rapid decay of non-LTR retrotransposons in the genome of the three-spine stickleback. *Genome Biol. Evol.* 4, 687–702. doi: 10.1093/gbe/evs044

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, 523–538. doi: 10.1534/genetics.113.158147

Boissinot, S., Davis, J., Entezam, A., Petrov, D., and Furano, A. V. (2006). Fitness cost of LINE-1 (L1) activity in humans. *Proc. Natl. Acad. Sci. U.S.A.* 103, 9590–9594. doi: 10.1073/pnas.0603334103

Boissinot, S., Entezam, A., and Furano, A. V. (2001). Selection against deleterious LINE-1-containing loci in the human lineage. *Mol. Biol. Evol.* 18, 926–935. doi: 10.1093/oxfordjournals.molbev.a003893

Boissinot, S., and Sookdeo, A. (2016). The evolution of Line-1 in vertebrates. *Genome Biol. Evol.* 8, 3485–3507. doi: 10.1093/gbe/evw247

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Campbell-Staton, S. C., Goodman, R. M., Backstrom, N., Edwards, S. V., Losos, J. B., and Kolbe, J. J. (2012). Out of Florida: mtDNA reveals patterns of migration and Pleistocene range expansion of the Green Anole lizard (*Anolis carolinensis*). *Ecol. Evol.* 2, 2274–2284. doi: 10.1002/ece3.324

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J. N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol. Evol.* 7, 567–580. doi: 10.1093/gbe/evv005

Cost, G. J., Feng, Q., Jacquier, A., and Boeke, J. D. (2002). Human L1 element target-primed reverse transcription in vitro. *EMBO J.* 21, 5899–5910. doi: 10.1093/emboj/cdf592

Csillery, K., Francois, O., and Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* 3, 475–479. doi: 10.1111/j.2041-210X.2011.00179.x

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* 35, 41–48. doi: 10.1038/ng1223

Dewannieux, M., and Heidmann, T. (2005). L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. *J. Mol. Biol.* 349, 241–247. doi: 10.1016/j.jmb.2005.03.068

Duvernell, D. D., Pryor, S. R., and Adams, S. M. (2004). Teleost fish genomes contain a diverse array of L1 retrotransposon lineages that exhibit a low copy number and high rate of turnover. *J. Mol. Evol.* 59, 298–308. doi: 10.1007/s00239-004-2625-8

Elliott, T. A., and Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370, 20140331. doi: 10.1098/rstb.2014.0331

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9:e1003905. doi: 10.1371/journal.pgen.1003905

Excoffier, L., and Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27, 1332–1334. doi: 10.1093/bioinformatics/btr124

Furano, A. V. (2000). The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Prog. Nucleic Acid Res. Mol. Biol.* 64, 255–294. doi: 10.1016/S0079-6603(00)64007-2

Furano, A. V., Duvernell, D., and Boissinot, S. (2004). L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet.* 20, 9–14. doi: 10.1016/j.tig.2003.11.006

Glor, R. E., Losos, J. B., and Larson, A. (2005). Out of Cuba: overwater dispersal and speciation among lizards in the *Anolis carolinensis* subgroup. *Mol. Ecol.* 14, 2419–2432. doi: 10.1111/j.1365-294X.2005.02550.x

Gonzalez, J., and Petrov, D. A. (2012). Evolution of genome content: population dynamics of transposable elements in flies and humans. *Methods Mol. Biol.* 855, 361–383. doi: 10.1007/978-1-61779-582-4_13

Kapitonov, V. V., Tempel, S., and Jurka, J. (2009). Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* 448, 207–213. doi: 10.1016/j.gene.2009.07.019

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., et al. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* 12:231. doi: 10.1186/1471-2105-12-231

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

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. doi: 10.1016/0092-8674(93)90078-5

Malik, H. S., Burke, W. D., and Eickbush, T. H. (1999). The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* 16, 793–805. doi: 10.1093/oxfordjournals.molbev.a026164

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. *Ecol. Evol.* 6, 8115–8125. doi: 10.1002/ece3.2547

Martin, S. L., Li, W.-H. P., Furano, A. V., and Boissinot, S. (2005). The structures of mouse and human L1 elements reflect their insertion mechanism. *Cytogenet. Genome. Res.* 110, 223–228. doi: 10.1159/000084956

McClure, M. A., Richardson, H. S., Clinton, R. A., Hepp, C. M., Crowther, B. A., and Donaldson, E. F. (2005). Automated characterization of potentially active retroid agents in the human genome. *Genomics* 85, 512–523. doi: 10.1016/j.ygeno.2004.12.006

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310, 321–324. doi: 10.1126/science.1117196

Neafsey, D. E., Blumenstiel, J. P., and Hartl, D. L. (2004). Different regulatory mechanisms underlie similar transposable element profiles in pufferfish and fruitflies. *Mol. Biol. Evol.* 21, 2310–2318. doi: 10.1093/molbev/msh243

Nellaker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., et al. (2012). The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol.* 13:R45. doi: 10.1186/gb-2012-13-6-r45

Novick, P. A., Basta, H., Floumanhaft, M., McClure, M. A., and Boissinot, S. (2009). The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard *Anolis carolinensis* shows more similarity to fish than mammals. *Mol. Biol. Evol.* 26, 1811–1822. doi: 10.1093/molbev/msp090

Ohshima, K., Hamada, M., Terai, Y., and Okada, N. (1996). The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. *Mol. Cell. Biol.* 16, 3756–3764. doi: 10.1128/MCB.16.7.3756

Ostertag, E. M., and Kazazian, H. H. Jr. (2001). Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* 11, 2059–2065. doi: 10.1101/gr.205701

Petrov, D., Aminetzach, Y. T., Davis, J. C., Bensasson, D., and Hirsh, A. E. (2003). Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol. Biol. Evol.* 20, 880–892. doi: 10.1093/molbev/msg102

Petrov, D. A., Fiston-Lavier, A. S., Lipatov, M., Lenkov, K., and Gonzalez, J. (2011). Population genomics of transposable elements in *Drosophila melanogaster*. *Mol. Biol. Evol.* 28, 1633–1644. doi: 10.1093/molbev/msq337

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E., and Lercher, M. J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* 31, 1929–1936. doi: 10.1093/molbev/msu136

Piskurek, O., Nishihara, H., and Okada, N. (2009). The evolution of two partner LINE/SINE families and a full-length chromodomain-containing Ty3/Gypsy LTR element in the first reptilian genome of *Anolis carolinensis*. *Gene* 441, 111–118. doi: 10.1016/j.gene.2008.11.030

Rishishwar, L., Marino-Ramirez, L., and Jordan, I. K. (2016). Benchmarking computational tools for polymorphic transposable element detection. *Brief. Bioinform.* doi: 10.1093/bib/bbw072 [Epub ahead of print].

Shen, J. J., Dushoff, J., Bewick, A. J., Chain, F. J., and Evans, B. J. (2013). Genomic dynamics of transposable elements in the western clawed frog (Silurana tropicalis). *Genome Biol. Evol.* 5, 998–1009. doi: 10.1093/gbe/evt065

Song, M., and Boissinot, S. (2007). Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* 390, 206–213. doi: 10.1016/j.gene.2006.09.033

Stewart, C., Kural, D., Stromberg, M. P., Walker, J. A., Konkel, M. K., Stutz, A. M., et al. (2011). A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* 7:e1002236. doi: 10.1371/journal.pgen.1002236

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81. doi: 10.1038/nature15394

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

Tollis, M., Ausubel, G., Ghimire, D., and Boissinot, S. (2012). Multi-locus phylogeographic and population genetic analysis of *Anolis carolinensis*: historical demography of a genomic model species. *PLoS ONE* 7:e38474. doi: 10.1371/journal.pone.0038474

Tollis, M., and Boissinot, S. (2012). The evolutionary dynamics of transposable elements in eukaryote genomes. *Genome Dyn* 7, 68–91. doi: 10.1159/000337126

Tollis, M., and 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, 1754–1768. doi: 10.1093/gbe/evt133

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* 142, 59–72. doi: 10.1007/s10709-013-9754-1

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43, 11–33. doi: 10.1002/0471250953.bi1110s43

Volff, J. N., Bouneau, L., Ozouf-Costaz, C., and Fischer, C. (2003). Diversity of retrotransposable elements in compact pufferfish genomes. *Trends Genet.* 19, 674–678. doi: 10.1016/j.tig.2003.10.006

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562. doi: 10.1038/nature01262

Wei, W., Gilbert, N., Ooi, S. L., Lawler, J. F., Ostertag, E. M., Kazazian, H. H., et al. (2001). Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* 21, 1429–1439. doi: 10.1128/MCB.21.4.1429-1439.2001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 February 2017
doi: 10.3389/fpls.2017.00229

Introgression Threatens the Genetic Diversity of *Quercus austrocochinchinensis* (Fagaceae), an Endangered Oak: A Case Inferred by Molecular Markers

Miao An [1,2], *Min Deng* [1,2]*, *Si-Si Zheng* [1,2], *Xiao-Long Jiang* [1,2] *and Yi-Gang Song* [1,2]

[1] Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, Shanghai, China, [2] Shanghai Chenshan Plant Science Research Center, Chinese Academy of Sciences, Shanghai, China

Natural introgression can cause negative effects where rare species experience genetic assimilation and invade by their abundant congeners. *Quercus austrocochinchinensis* and *Q. kerrii* (subgenus *Cyclobalanopsis*) are a pair of closely related species in the Indo-China area. Morphological intermediates of the two species have been reported in this region. In this study, we used AFLP, SSR and two key leaf morphological diagnostic traits to study the two *Q. austrocochinchinensis* populations, two pure *Q. kerrii* and two putative hybrid populations in China. Rates of individual admixture were examined using the Bayesian clustering programs STRUCTURE and NewHybrids, with no a priori species assignment. In total, we obtained 151 SSR alleles and 781 polymorphic loci of AFLP markers. Population differentiation inferred by SSR and AFLP was incoherent with recognized species boundaries. Bayesian admixture analyses and principal coordinate analysis identified more hybrids and backcrossed individuals than morphological intermediates in the populations. SSR inferred a wide genetic assimilation in *Q. austrocochinchinensis*, except for subpopulation D2 in the core area of Xi-Shuang-Ban-Na Nature Reserve (XSBN). However, AFLP recognized more *Q. austrocochinchinensis* purebreds than SSR. Analysis using NewHybrids on AFLP data indicated that these hybridized individuals were few F_2 and predominantly backcrosses with both parental species. All these evidences indicate the formation of a hybrid swarm at XSBN where the two species co-exist. Both AFLP and SSR recognized that the core protected area of XSBN (D2) has a high percentage of *Q. austrocochinchinensis* purebreds and a unique germplasm. The Hainan population and the other subpopulations of XSBN of the species might have lost their genetic integrity. Our results revealed a clear genetic differentiation in the populations and subpopulations of *Q. austrocochinchinensis* and ongoing introgression between *Q. austrocochinchinensis* and *Q. kerrii* at the disturbed contact areas. Combining the results from genetic and morphological analyses, the conservation of subpopulation D2 should

OPEN ACCESS

Edited by:
*Scott V. Edwards,
Harvard University, USA*

Reviewed by:
*Octavio Salgueiro Paulo,
Universidade de Lisboa, Portugal
Yongpeng Ma,
Kunming Institute of Botany (CAS),
China*

***Correspondence:**
*Min Deng
dengmin@sibs.ac.cn*

Specialty section:
*This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Plant Science*

Received: 03 October 2016
Accepted: 06 February 2017
Published: 21 February 2017

Citation:
An M, Deng M, Zheng S-S, Jiang X-L and Song Y-G (2017) Introgression Threatens the Genetic Diversity of Quercus austrocochinchinensis (Fagaceae), an Endangered Oak: A Case Inferred by Molecular Markers. Front. Plant Sci. 8:229. doi: 10.3389/fpls.2017.00229

be prioritized. Conservation and restoration of the integrity of tropical ravine rainforest is an important long-term goal for the successful conservation of *Q. austrocochinchinensis*. The fine-scale landscape might play an essential role in shaping the spatial patterns of hybridization. Further studies are needed to evaluate these patterns and dynamics.

Keywords: *Quercus* subgenus *Cyclobalanopsis*, AFLP, SSR, introgression, endangered species

INTRODUCTION

Natural hybridization is a frequent phenomenon in plants, occurring in 25% of extant species (Mallet, 2005; Whitney et al., 2010). The F_1 hybrids without reproductive barriers can bridge the gene flow between the two parental species by facilitating further backcrossing to the parental species; this process leads to introgression, which is an important evolutionary process (Arnold, 1992; Barton, 2001). By interspecific genetic exchange, introgression increases the genetic diversity of one or both parental species and can lead to novel adaptations and speciation events (Grant, 1981; Rieseberg, 1995, 1997; Mallet, 2007). However, hybridization and introgression events can also have harmful effects on the progenitor's fate. An endemic or rare species may go extinct when it undergoes introgression with common congeners or a more reproductively successful prevalent species (Rieseberg, 1995; Levin et al., 1996; Rhymer and Simberloff, 1996; Lepais et al., 2009). By repeated backcrossing, ancestral alleles of rare species become diluted after a certain number of generations (Briggs and Walters, 1997).

The genus *Quercus* s.l. contains ~400 to 600 species (Govaerts and Frodin, 1998) and can be divided into the subgenera *Quercus* and *Cyclobalanopsis*, based on whether the cupule is imbricate-scaled or lamellate. The subgenus *Cyclobalanopsis* is one of the dominant tree taxa in evergreen broad-leaf forests (EBLFs) of eastern and southeastern Asia, with ~90 to 122 species (Govaerts and Frodin, 1998; Deng, 2007). Natural introgression is common in oaks (Valbuena-Carabaña et al., 2005; Curtu et al., 2007; Burgarella et al., 2009; Salvini et al., 2009; Ortego and Bonal, 2010; Moran et al., 2012). Renowned as "worst case scenario for the biological species concepts" (Coyne and Orr, 2004) due to apparent local interspecific gene flow (Burger, 1975; Whittemore and Schaal, 1991; Lexer et al., 2006), widespread oak species of the subgenus *Quercus,* nonetheless, exhibit genetic coherence across a broad geographic range (Muir et al., 2000; Hipp and Weber, 2008; Cavender-Bares and Pahlich, 2009). However, most of these studies were conducted on species of the subgenus *Quercus*.

Although trees of the subgenus *Cyclobalanopsis* are the keystone elements in EBLFs of mainland Asia, studies on hybridization and introgression among the species of this subgenus are rather rare. Only two sympatric species (*Q. sessilifolia* and *Q. acuta*) distributed in Korea and Japan were investigated and revealed the introgression between the two species (Tamaki and Okada, 2014). Indo-China is the diversification center for the subgenus *Cyclobalanopsis*, with about 70 species occurring in EBLFs in this region (Lou and Zhou, 2001). Of these, one-third are endemic and rare species, and a large number of them have sympatric distributions, but maintain their prominent ecological niche and morphological variation. So far, no studies have applied genotyping methods to test the existence of gene flow among the different species of the subgenus *Cyclobalanopsis* in Indo-China. *Quercus kerrii* and *Q. austrocochinchinensis* is a pair of species closely genetically related to the subgenus Cyclobalanopsis (Deng et al., 2013). *Q. kerrii* is widespread common species in open slopes of EBLFs in Indo-China, while the distribution of *Q. austrocochinchinensis* is rather restricted, with only four known sites, of which two are in China, and other two are located in Northern Vietnam and Northern Thailand respectively (Huang et al., 1999; Phengklai, 2006).

We have previously described the morphological intermediates *Q. austrocochinchinensis* and *Q. kerrii* using leaf morphological traits, indicating that the two species can form hybrids (Song et al., 2015). Intermediate morphology has been widely used to reveal the status of hybrids in former studies of plant hybridization (Kleinschmit et al., 1995; Craft et al., 2002; Kremer et al., 2002). However, morphological diagnostic traits have limited power to accurately identify the hybrids and pure parental species (López-Caamal and Tovar-Sánchez, 2014). Compared to morphologic methods, DNA markers are more reliable and powerful tools compared (Harrison, 1993) and can also precisely predict the ancestral states in later generation hybrids (Pritchard et al., 2000; Falush et al., 2003; Evanno et al., 2005). Preserving the genetic distinction of endangered species is critical for their conservation and by using molecular approaches, it is possible to select out non- or less-hybridized subpopulations from a hybrid zone to use in *ex-situ* conservation.

In this follow-up study, we aim to (1) investigate whether and to what extent introgression exists between *Q. austrocochinchinensis* and *Q. kerrii*; (2) discuss the genetic extinction risk of the rare oak species *Q. austrocochinchinensis* and its possible conservation management; (3) compare the results of different approaches (morphological traits, AFLP, and SSR), and discuss the diagnostic power in distinguishing hybrids.

MATERIALS AND METHODS

Ethics Statement

Sampling of endangered oak species *Quercus austrocochinchinensis* and *Q. kerrii* was granted and supported by National Forestry Bureau of China and Local National Nature Reserves.

Population Sampling and Species Identification

In total, 57 and 36 individuals with typical traits of *Q. austrocochinchinensis* and *Q. kerrii*, respectively, were used and

15 morphological intermediates were included in this study. The 108 individuals were sampled in five populations, A–E (**Figure 1** and **Table 1**). According to a previous investigation (Song et al., 2015), *Q. austrocochinchinensis* trees can only be found in populations D and E. Of these, E was considered as a pure *Q. austrocochinchinensis* population (Song et al., 2015) and D is located in the contact zone which contains both *Q. kerrii* and *Q. austrocochinchinensis* trees. Six sub-populations, D1–D6, were sampled within D population regions. Two putative *Q. kerrii* purebred populations were sampled from populations A and B. Population C is a putative hybrid population with morphological intermediates, but trees with typical traits of *Q. austrocochinchinensis* cannot be found in or close to region C.

Prior to the experiment, all individuals and their voucher specimens were carefully inspected and identified based on key morphological diagnostic traits, e.g., shape of cupule and trichomes on leaf abaxial surface. The results were used as species information for the studied samples in later analyses. Leaf tissues used for DNA extractions were collected from each individual and dried instantly using silica gel. All vouchers specimens of each tree were stored at the herbarium of the Shanghai Chenshan Botanical Garden (CSH).

Measurement of Leaf Morphological Traits

In a previous study, we found that the two macro-morphological features leaf apex shape and leaf length-to-width ratio are the key diagnostic features that can be applied to identify *Q. kerrii* and *Q. austrocochinchinensis* and even to assign their hybrids (Song et al., 2015). To link the genetic pattern to the morphological features, we used ratios instead of absolute lengths to measure leaf morphological traits for the two parental species and the intermediates. The two ratios were leaf length-to-width ratio and leaf apex length-to-width ratio. The measuring method of the two ratios is shown in **Figure 2**. At least five leaves were measured from each voucher specimen to compute an average for the drawing of the scatter plot.

SSR and AFLP Analysis

Total genomic DNA for each sampled individual was extracted from the silica gel-dried leaf tissue using the modified CTAB method (Huang et al., 2000). The DNA quality was checked by loading DNA on a 1.0% agarose gel, and the DNA concentration of each sample was measured using a TBS-380 Fluorometer (Turner BioSystems Inc., Sunnyvale, CA). The samples were screened for 11 SSR markers. Of these loci, Qk15874, Qk17139, Qk17611, and Qk20944 have been developed for *Q. kerrii* from transcriptome data (An et al., 2016), QmC00693, QpZAG9, QpZAG16, QpZAG36, QpZAG110, CG371 were from the non-coding region (Steinkellner et al., 1997; Ueno et al., 2008; Tong et al., 2012), and CR627959 was predicted in the coding region of Cys-3-His zinc finger protein in nuclear genome (Ueno and Tsumura, 2008). All of these loci show a relatively high degree of transferability within the genus *Quercus* and an adequate degree of polymorphism of the studied taxa. For the 11 SSR primer pairs, 5′ end of forward primers were labeled with fluorescent dye tags (6FAM or HEX or ROX) (Sangon, Shanghai, China). The PCR reactions were performed in 20 μl reaction volume containing 10 μl TIANGEN PCR Master Mix (TIANGEN, Beijing, China), 0.3 μl/L of each primer (10 mM), and 20 ng genomic DNA; PCR reactions were performed as follows: 5 min initial denaturation at 94°C, 35 cycles of 40 s at 94°C, 30 s at 55°C, 1 min elongation at 72°C, and 7 min extension at 72°C. Finally, Gene-Scan-500 LIZ size standard (Applied Biosystem^{TM}) was added to all the samples before loading on an automated sequencer ABI 3730 (Applied Biosystem^{TM}). The final step was performed by a private professional commercial lab (Shanghai Majorbio Bio-pharm Technology Co., Ltd, Shanghai, China).

The AFLP method was performed following the protocol by Vos et al. (1995), with minor modifications. Briefly, 500 ng of genomic DNA were double-digested using 5 U of *Eco*RI and 2 U of *Mse*I (New England BioLabs). The digestion mixtures were incubated at 37°C for 3 h and the digested mixture was then incubated at 70°C for 10 min to denature the enzymes. Subsequently, 4 μL of digested DNA were added to 16 μL of ligation mix containing 2 U T4 DNA ligase (New England BioLabs), 5 pmol *Eco*RI, and 50 pmol *Mse*I adaptor. The mixture was incubated at 10°C for 14 h and then denatured at 70°C for 10 min. The ligated DNA samples were diluted 5-fold with double sterile water. Pre-selective amplification reactions were carried out using *Eco*RI-A (5′-GACTGCGTACCAATTCA-3′) and *Mse*I-C (5′-GATGAGTCCTGAGTAAC-3′) in a 50 μL volume containing 1.5 mmol/L $MgCl_2$, 200 μmol/L of each dNTP, 1.25 μmol/L of each primer, and 0.6 U *r*Taq DNA polymerase (Takara Biotechnology, Dalian, China) under the following cycle: 3 min at 72°C, 30 cycles of 30 s denaturing at 94°C, 30 s annealing at 56°C, 1 min extension at 72°C, and a final extension for 5 min at 72°C. After a 1:20 dilution of pre-selective PCR products, seven primer combinations were performed in selective amplification (*Eco*RI-ACG/*Mse*I-CAC, *Eco*RI-AGG/*Mse*I-CAA, *Eco*RI-AGG/*Mse*I-CAC, *Eco*RI-AGG/*Mse*I-CAG, *Eco*RI-AGG/*Mse*I-CTA, *Eco*RI-AGG/*Mse*I-CTC, *Eco*RI-AGG/*Mse*I-CTT). The *Eco*RI primers were fluorescently labeled with 6-FAM. These primer pairs were chosen because they generated clear and fewer bands (thus decreasing the risk of fragment non-homology) with sufficient variability in preliminary tests. Selective PCRs were carried out in a 20 μL volume containing 2.5 mmol/L $MgCl_2$, 200 μmol/L of each dNTP, 1.25 μmol/L of each primer, and 0.2 U of rTaq DNA polymerase (Takara Biotechnology, Dalian, China) and under the following cycle: 3 min at 94°C, 9 cycles of 40 s at 94°C, 30 s at 65–57°C touchdown (reducing the temperature at 1°C per cycle), 15 min at 72°C, 20 cycles of 40 s at 94°C, 30 s at 56°C, 1.5 min at 72°C, and a final extension for 7 min at 60°C. The PCR products were 10-fold diluted and mixed with Gene-Scan-500 LIZ size standard (Applied Biosystem^{TM}); products from each primer combination were loaded separately on an automated sequencer ABI 3730 (Applied Biosystem^{TM}) by the same commercial service provider mentioned above.

Raw data of SSR and AFLP samples were collected and analyzed using GeneMarker®v2.2.0. The samples with low quality of size calibrations or peaks were excluded from allele calling. The allele sizes of SSR were read manually and MicroChecker v 2.2.3 (Van Oosterhout et al., 2004) was used to check for potential errors. For AFLP, the variable fragments in

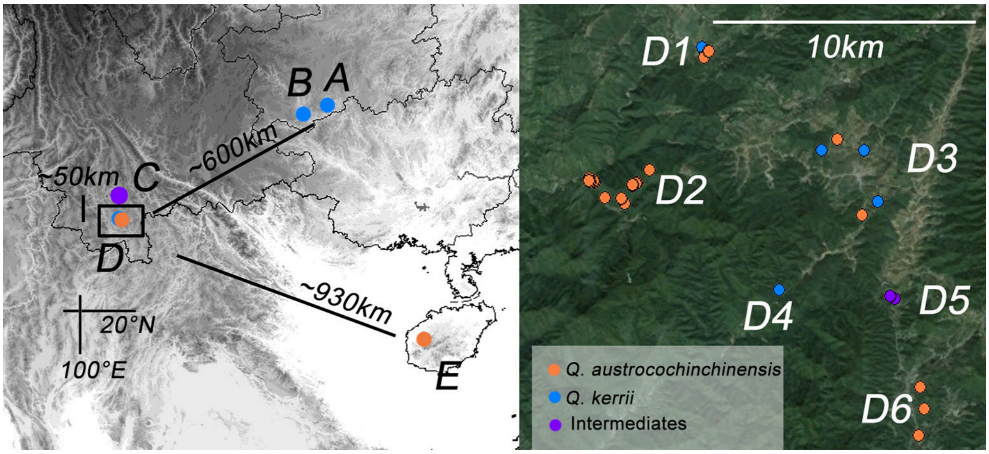


FIGURE 1 | Geographic distribution of the study populations of *Quercus kerrii* and *Q. austrocochinchinensis*. Population A and B are pure *Q. kerrii* populations. C is a putative hybrid population without co-occurrence of any *Q. austrocochinchinensis* trees. D1–D9 are located in the contact zone. E is a population of *Q. austrocochinchinensis* that has been reported previously in a morphological study as a purebred population (Song et al., 2015). Individuals were colored according to parental species or intermediates (orange, blue, or purple) pre-identified by morphological features.

TABLE 1 | Sampling information of the study populations.

Site ID	Population type	Location	N(a)	N(k)	N(i)	Long (E)	Lat (N)	Elev (m)
A	Pure site	Luo-dian, Guizhou	0	11	0	106°40′	25°15′	526
B	Pure site	Heng-xian, Guizhou	0	17	0	105°53′	24°59′	459
C*	Intermediate	Si-mao, Yunnan	0	0	10	100°50′	22°50′	1077
D1	Mixed site	XSBN, Yunnan	2	2	0	100°48′	22°05′	919
D3	Mixed site	XSBN, Yunnan	14	4	0	100°52′	22°19′	1030
D2	Pure site	XSBN, Yunnan	25	0	0	100°46′	22°19′	867
D6	Pure site	XSBN, Yunnan	7	0	0	100°53′	22°14′	867
D4	Pure site	XSBN, Yunnan	0	2	0	100°50′	22°17′	1036
D5	Intermediate	XSBN, Yunnan	0	0	5	100°53′	22°17′	936
E	Pure site	BWL, Hainan	9	0	0	109°05′	19°07′	247

XSBN, Xi-Shuang-Ban-Na National Nature Reserve; BWL, Ba-Wang-Ling Nature Reserve; N (a/k/i): sampling number of Q. austrocochinchinensis (a), Q. kerrii (k) and morphological intermediates (i). Long, Longitude; Lat, Latitude; Elev, Elevation.

the size range 50–500 base pairs (bp) were manually scored as present (1) or absent (0). We only considered fragments with similar fluorescence profile and intensities across the samples to maximize the probability of homology.

Genetic Diversity Analysis

A set of statistical tests on SSRs, including allelic richness (Na), allele frequency distribution, Shannon's Information Index (I), Observed and Expected Heterozygosity (H_O and H_E), were performed by GenAlEx version 6.5 (Peakall and Smouse, 2012), using all the individuals of *Q. kerrii* and *Q. austrocochinchinensis*. In addition, Unbiased Expected Heterozygosity (uH_E) was estimated as $(2N/(2N-1)) * H_E$. Fixation Index (F_{IS}), or inbreeding coefficient was estimated as $(H_E–H_O)/H_E$ (Nei and Li, 1979). We used GenePop v 4.2 (Rousset, 2008) to test the departure from Hardy-Weinberg equilibrium (HWE) (heterozygote deficiency or excess) for each locus of the 11 SSR loci, and to test for homogeneity of alleles distributions between species. We also counted the number of private alleles for each species.

For AFLP, percentage of polymorphic loci, unbiased estimates of genetic diversity (H_j, analogous to H_E), and differentiation statistics were calculated using the AFLP-SURV v. 1.0 software (Vekemans et al., 2002). With this software, allelic frequencies at AFLP loci were calculated from the observed frequencies of fragments using the Bayesian approach proposed by Zhivotovsky (1999) for diploid species. A non-uniform prior distribution of allelic frequencies was assumed with its parameters derived from the observed distribution of fragment frequencies among loci. These allelic frequencies were used as the input for the analysis of genetic diversity within and between samples following the method described in Lynch and Milligan (1994).

Genetic and Phenotypic Differentiation

The comparison of F_{ST} and Q_{ST} provides a basis to distinguish neutral from adaptive divergence (Leinonen et al., 2013). To

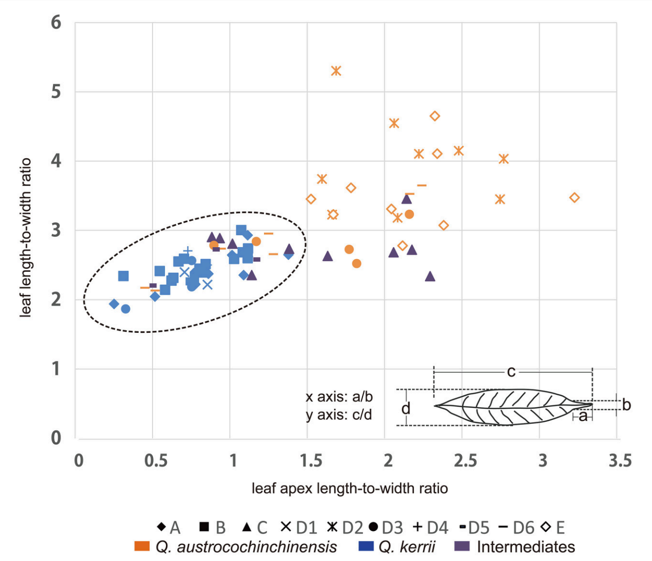


FIGURE 2 | Distributions of two leaf morphological measures. Graph displaying the relationship between the two morphological features that were measured in this study, where the leaf apex length-to-width ratio is on the x-axis and the leaf length-to-width ratio is on the y-axis. "◆" represents Population A; "■" represents Population B; "▲" represents Population C; "◇" represents Population E; the remaining symbols belong to Population D, of which "×" represents sub-population D1; "Ж" represents sub-population D2; "•" represents sub-population D3; "+" represents subpopulation D4; "■" represents subpopulation D5; "▬" represents sub-population D6.

investigate genetic differentiation between the two species and among populations, F_{ST} values were measured on AFLP and SSR data using AFLP-SURV v. 1.0 (Vekemans et al., 2002) and GenePop v. 4.2 (Rousset, 2008), respectively. The parameter Q_{ST} estimates the among-population proportion of the total additive genetic variance of a genetic quantitative trait. If $Q_{ST} = F_{ST}$, trait divergence among populations could have been driven only by genetic drift. If $Q_{ST} > F_{ST}$, the populations are likely to have been caused by directional selection, and if $Q_{ST} < F_{ST}$, there is evidence for uniform selection or stabilizing selection across the populations. Because this is not possible for the breeding designs for this study, we used P_{ST}, a proxy for Q_{ST}, to compare with F_{ST} (Leinonen et al., 2013). In this study, P_{ST} was calculated between all the *Q. austrocochinchinensis* and *Q. kerrii* individuals, using the equation of $\sigma^2_{GB}/(\sigma^2_{GB}+2\sigma^2_{GW})$, where σ^2_{GW} and σ^2_{GB} are within- and among-population components of variance. Putative hybrids were excluded from F_{ST} and P_{ST} estimation.

In addition, F_{ST} for each locus was also estimated. For AFLP, allele frequencies of *Q. austrocochinchinensis* and *Q. kerrii* were calculated separately. We calculated F_{ST} values between 57 for *Q. austrocochinchinensis* and 36 for *Q. kerrii* individuals for all polymorphic loci via the formula $F_{ST} = 1\text{-}H_S/H_T$ (Nei, 1973). The variable H_S represents average within-population heterozygosity and H_T represents expected heterozygosity for the total population; H_S and H_T were calculated using the following formula: $H_S = 1/2(2p_1q_1+2p_2q_2)$ and $H_T = 1/2(p_1+p_2)*(q_1+q_2)$, with $q = 1\text{-}p$. For SSR markers, F_{ST} for each loci was calculated using GenePop v. 4.2 (Rousset, 2008). Finally, the frequency distribution of F_{ST} values for both markers were plotted in a histogram.

To detect the outlier loci under selection of an AFLP dataset (781 loci), program BayeScan v. 2.1 (Foll and Gaggiotti, 2008; Fischer et al., 2011) was used to identify candidate loci under natural selection across all 10 populations. A threshold value for determining loci under selection was evaluated in accordance with Jeffreys (1961) interpretation, which is a logarithmic scale for model choice as follows: $\log_{10}PO > 0.5$ (substantial), $\log_{10} PO > 1.0$ (strong), $\log_{10} PO > 1.5$ (very strong), and $\log_{10}PO > 2.0$ (decisive support for accepting a model) (Fischer et al., 2011). We employed a threshold of $\log_{10} PO > 2.0$ for the rejection of the null hypothesis in each of the conducted tests. BayeScan analysis was conducted with a burn-in of 50,000 iterations, a thinning interval of 50, and a sample size of 5,000. The number of pilot runs was kept at 20, with a length of 50,000 each. The SSR loci were not used to calculate outliers because of their limited number.

Population Cluster Analysis

To cluster individuals into genetically distinct groups, a Bayesian clustering approach was employed using STRUCTURE v. 2.3.4 (Pritchard et al., 2000; Falush et al., 2003), without consideration of sampling information. We adopted the admixture model with correlated allele frequencies (Lepais et al., 2009; Zalapa et al., 2009). No prior knowledge of the species was included in the analyzed data sets. To determine the optimal number of groups (K), we ran STRUCTURE, with K varying from 1 to 10 and with 10 runs for each K value. The ΔK was calculated using the mean log-likelihood for each K according to Evanno et al. (2005). For SSR and AFLP data sets, each run was performed for 100,000 Markov Chain Monte Carlo (MCMC) repetitions with a burn-in period of 50,000.

The admixture coefficient (*q*-value) generated from STRUCTURE was used to classify individuals into purebred and hybrids, using a threshold *q*-value of = 0.1, where samples with *q*-values < 0.1 or > 0.9 were classified as purebred and those with *q*-values between 0.1 and 0.9 as hybrids, including F_1 and backcrosses (Vähä and Primmer, 2006; Lepais et al., 2009). The F_1 hybrids result in *q*-values = 0.5, but the coefficient of backcrosses would be biased toward one of the parental species and produce *q*-values between 0.5 and 0.9 (Lepais et al., 2009). Taking errors into consideration, individuals with 0.6 < *q*-values < 0.9 were recognized as backcrosses.

As the SSR loci were tested to violate the HWE assumption (see result), we also used the program InStruct (Gao et al., 2007), an alternative software of STRUCTURE, to ensure that the results obtained from STRUCTURE were reliable. The HWE within loci for co-dominant markers is not compulsory for the model in InStruct. The optimal K (number of clusters) value was selected according to Deviance Information Criteria (DIC) and is presented in the result section. Mode 2 was selected to run 50,000 MCMC repetitions with 10,000 burn-in periods.

The NewHybrids software v. 1.1 beta, employing a Bayesian analysis (Anderson and Thompson, 2002; Anderson, 2008), was used as a second method to confirm the presence of hybrids in our dataset. It calculates the posterior probability that sampled individuals fall into one set of hybrid categories (Parent K, Parent A, F_1, F_2, BC to Parent K; BC to Parent A, thus covering parents and two generations of offspring). All the 11 SSR loci were analyzed using this software. For AFLP data, NewHybrids will assign the individuals more accurately, using the loci with high differentiation at interspecies level. Therefore, we filtered the AFLP loci according to F_{ST}. We applied three approaches to compare the results obtained by NewHybrids, using 100 loci with highest F_{ST} value at interspecies level, 279 loci ($F_{ST} \geq 0.1$), and 450 AFLP loci ($F_{\mathrm{ST}} \geq 0.03$), respectively, for analysis. A burn-in period of 7,500 MCMC repetitions was defined and 10,000 iteration were run thereafter.

Principal coordinate analysis (PCoA) aims to visualize similarities or dissimilarities of individual data based on a distance matrix. This method is also an alternative algorithm without any assumptions about the population genetic model. For AFLP and SSR data, the pairwise Euclidian distance matrix was constructed, and the first two principal co-ordinates were visualized by GenAlEx v. 6.5 (Peakall and Smouse, 2012).

RESULTS

Morphological Analysis

Morphological measurements showed that the two species could be distinguished using a combination of two ratio values, which were leaf length-to-width ratio (LR) and leaf apex length-to-width ratio (LAR) (**Figure 2**). We found that all *Q. kerrii* individuals were grouped together, with an average of LR = 2.45 (1.88–3.02) and LAR = 0.78 (0.23–1.36). Of the sampled *Q. austrocochinchinensis* individuals, only the samples from D2 and E were totally separated from *Q. kerrii* samples. They had larger LR and LAR values compared to other populations, with an average of LR = 3.71 (2.79–5.32) and LAR = 2.06 (1.50–3.20). Morphological variation was higher in *Q. austrocochinchinensis* individuals than in *Q. kerrii* specimens, with standard deviations (σ) for *Q. austrocochinchinensis* and *Q. kerrii* of 0.72 and 0.25 for LR and 0.64 and 0.25 for LAR, respectively.

The putative hybrids of D5 within the hybrid zone had morphological values similar to *Q. kerrii*, with an average of LR = 2.52 (2.22–2.74) and LAR = 0.82 (0.47–1.13). However, the putative hybrids of population C had a sharper leaf apex with LAR = 1.54 (0.86–2.27), although the leaves were as broad as those of *Q. kerrii* with LR = 2.77 (2.35–3.47). In addition, there were seven *Q. austrocochinchinensis* individuals from D3 and D6 with leaf morphological traits similar to those of *Q. kerrii*.

Population Diversity and Differentiation

In total, 151 alleles were obtained from the 11 SSR loci used in this study. Some SSRs were highly variable, e.g., CG371, QpZAG16, and QpZAG110, containing more than 20 alleles. These SSRs also had higher Observed Heterozygosity (H_{O}) and Expected Heterozygosity (H_{E}) (**Table 2**) than other loci. Although most frequent alleles were shared by *Q. austrocochinchinensis* and *Q. kerrii*, some loci showed greater variation in allele frequency (**Figure S1**). Mean Expected Heterozygosity (H_{E}) across all loci in *Q. austrochichinensis* was 0.706 and higher than that in *Q. kerrii* with 0.595. Species-specific alleles were found at several loci, especially rare alleles restricted to *Q. austrocochinchinensis* (**Table 3**). Frequency of the private alleles was 34.4 and 12.0% in *Q. austrocochinchinensis* and *Q. kerrii*, respectively. All SSR loci except Qk17611 significantly deviated from HWE ($p < 0.01$). Heterozygote deficiency was found at all loci except Qk17611 and QpZAG36 ($p < 0.01$). The estimation of genetic differentiation between *Q. austrocochinchinensis* and *Q. kerrii* was low across all SSR loci ($F_{\mathrm{ST}} = 0.138$). The differentiation among population within species was also low for both species, with F_{ST} values of 0.130 and 0.042 for *Q. austrocochinchinensis* and *Q. kerrii*, respectively.

Application of the seven AFLP primer pairs to 108 individuals resulted in 859 loci, of which 781 were polymorphic. The levels of diversity within each species, either at the population (H_{W}) or the whole sample level (H_{t}), were very similar (**Table 4**). Genetic differentiation between *Q. austrocochinchinensis* and *Q. kerrii* across the 781 markers was low ($F_{\mathrm{ST}} = 0.095$, $p < 0.01$) (**Table 4**). Similarly, the differentiation among populations within species was also low, with $F_{\mathrm{ST}} = 0.0246$ ($p < 0.01$) for *Q. austrocochinchinensis* and $F_{\mathrm{ST}} = 0.0667$ ($p < 0.01$) for *Q. kerrii*. The F_{ST} distribution of 781 AFLP alleles followed an L-shaped curve, which suggested that most of the alleles (471 of 781, 60.3%) had lower F_{ST} values ($F_{\mathrm{ST}} < 0.1$). Only a few alleles exhibited higher F_{ST} values (**Figure S2**). The P_{ST} values for all individuals of the two species, except for the putative hybrids, were larger (0.338) than F_{ST} (0.095 for AFLP, 0.082 for SSR). By using BayeScan, 3 out of 781 AFLP loci (0.38%) were identified as outlier loci under directional selection at a threshold of $\log_{10}$ PO > 2 (posterior probabilities higher than 0.99) (**Figure S3**).

Structure Analysis

The most likely number of clusters (K) of STRUCTURE analysis can be inferred by the peak of ΔK. We found that $K = 2$ was the optimal K value for both AFLP and SSR (**Figure S4**), representing two species, respectively. Following ten independent STRUCTURE runs with $K = 2$, individuals morphologically identified as *Q. kerrii* were assigned to one cluster with high probability, whereas those morphologically identified as *Q. austrocochinchinensis* were assigned to the other cluster with similarly high probability. Therefore, these two clusters were determined to represent *Q. kerrii* and *Q. austrocochinchinensis*, respectively. The STRUCTURE analysis results of SSR and AFLP are illustrated in **Figures 3A,C**, respectively, but the results are very different. Among individuals morphologically identified as *Q. kerrii*, both AFLP and SSR recognized populations A and B are *Q. kerrii* purebred population, but the results of the remaining populations are largely contradictory. For example, SSR results assigned most of the individuals in populations/subpopulations C, D1, D3, D4, D5, D6, E, and two individual of D2 to the *Q. kerrii* cluster, and the remaining individuals in D2 to another cluster, regardless of morphological identification. Based on the AFLP results, among the individuals identified as *Q. kerrii* in populations C, D1, D3, and D4, the mean proportion of *Q. kerrii* was 0.575 (0.412–0.941). The two morphological intermediate putative hybrid populations C and D5 contained a high proportion of *Q. austrocochinchinensis* germplasm (with a mean value of 0.901 [0.677–0.992]). Among the individuals

TABLE 2 | Genetic diversity of 11 SSR loci for all individuals of *Quercus asutrocochinchinensis* and *Q. kerrii*.

Locus	N	Na	Ne	I	H_O	H_E	uH_E	F_{IS}
CG371	96	24	9.352	2.593	0.781	0.893	0.898	0.125
QK20944	101	3	2.610	1.028	0.386	0.617	0.620	0.374
QK17611	69	5	1.615	0.796	0.420	0.381	0.383	−0.104
QK17139	93	11	5.269	1.822	0.688	0.810	0.815	0.151
QK15874	94	17	6.073	2.286	0.617	0.835	0.840	0.261
QpZAG36	93	5	2.884	1.177	0.484	0.653	0.657	0.259
QpZAG16	57	23	15.112	2.894	0.719	0.934	0.942	0.230
QpZAG110	99	28	9.343	2.728	0.677	0.893	0.898	0.242
QpZAG9	96	19	6.958	2.344	0.604	0.856	0.861	0.294
CR627959	93	8	1.612	0.869	0.269	0.380	0.382	0.292
QmC00963	61	8	2.717	1.373	0.459	0.632	0.637	0.274

*N, number or alleles**;** Na, number of different alleles; Ne, number of effective alleles; I, Shannon's Information Index; H_O, Observed Heterozygosity; H_E, Expected Heterozygosity; uH_E, Unbiased Expected Heterozygosity; F_{IS}**,** Fixation Index.*

TABLE 3 | Comparison of genetic diversity and differentiation between *Quercus austrocochinchinensis* and *Q. kerrii* based on 11 SSR loci.

	N		An		Ap		H_O		H_E		F_{ST}		
Locus	**QA**	**QK**	**QA**	**QK**	**QA**	**QK**	**QA**	**QK**	**QA**	**QK**	**Between species**	**QA**	**QK**
CG371	54	30	18	17	6	5	0.778	0.800	0.841	0.906	0.041	0.076	−0.001
QK20944	53	33	3	3	0	0	0.302	0.606	0.579	0.501	0.224	0.034	−0.011
QK17611	38	31	5	4	1	0	0.526	0.290	0.467	0.259	0.025	0.027	0.040
QK17139	53	30	9	7	2	0	0.623	0.767	0.770	0.760	0.087	0.185	0.023
QK15874	53	27	15	5	11	1	0.698	0.519	0.892	0.502	0.201	0.009	0.079
QpZAG36	52	29	3	3	0	0	0.385	0.586	0.358	0.439	0.557	−0.002	0.007
QpZAG16	26	22	16	11	10	5	0.615	0.909	0.898	0.875	0.056	−0.123	0.056
QpZAG110	54	31	22	17	8	3	0.611	0.774	0.862	0.887	0.027	0.030	0.070
QpZAG9	50	31	15	15	4	4	0.600	0.677	0.851	0.877	0.004	0.249	0.113
CR627959	51	29	7	3	4	0	0.353	0.172	0.526	0.161	0.076	0.347	0.103
QmC00963	36	11	8	2	6	0	0.500	0.273	0.718	0.351	0.072	0.309	−0.144
All			121	87	52	18	0.545	0.579	0.706	0.593	0.138	0.130	0.042

*QA, two populations of Q. austrocochinchinensis; QK, three populations of Q. kerrii; N**,** The number of individuals having the allele in each species; An, number of alleles over all the populations for each species; Ap, number of private alleles; H_O, observed heterozygosity; H_E, expected heterozygosity; F_{ST}, differentiation between populations (species).*

TABLE 4 | Comparison of genetic diversity and differentiation between *Quercus austrocochinchinensi*s and *Q. kerrii* based on 781 AFLP loci.

Populations	N	H_j	H_t	SE (H_t)	H_W	SE (H_W)	F_{ST}	Lower 99% F_{ST}	Upper 99% F_{ST}
Between species	2	0.2808	0.3084	<0.001	0.0294	0.0025	0.0953	−0.005	0.0096
Q. austrocochinchinensis	2	0.2765	0.2847	<0.001	0.2778	0.0015	0.0246	−0.0135	0.0203
Q. kerrii	3	0.2815	0.3009	<0.001	0.2808	0.0021	0.0667	−0.0154	0.0156

N, number of populations; H_t, total diversity; H_W, average diversity within population; F_{ST}, differentiation between populations.

identified as *Q. austrocochinchinensis,* in populations D1, D2, and E, the mean proportion of *Q. austrocochinchinensis* germplasm was 89.1% (with the lowest value, 21.7%, found in individual Ca44), but mean proportion was considerably lower in D3 with 34.5% (1.1–71.3%). Given that we considered individuals with *q*-values between 0.6 and 0.9 as backcrosses, we did not identify any backcrossed individuals based on SSRs. However, for AFLP, 25 out of 108 (23.1%) samples were recognized as backcrosses.

The software InStruct (Gao et al., 2007) was used to verify the reliability of SSR results generated by STRUCTURE. The result was highly consistent to the results obtained from STRUCTURE. The optimal *K* value selected by DIC was also 2, similar to the best *K* inferred by STRUCTURE. The D2 subpopulation with pure *Q. austrocochinchinensis* was still easily distinguished from the analysis (**Figure S5**).

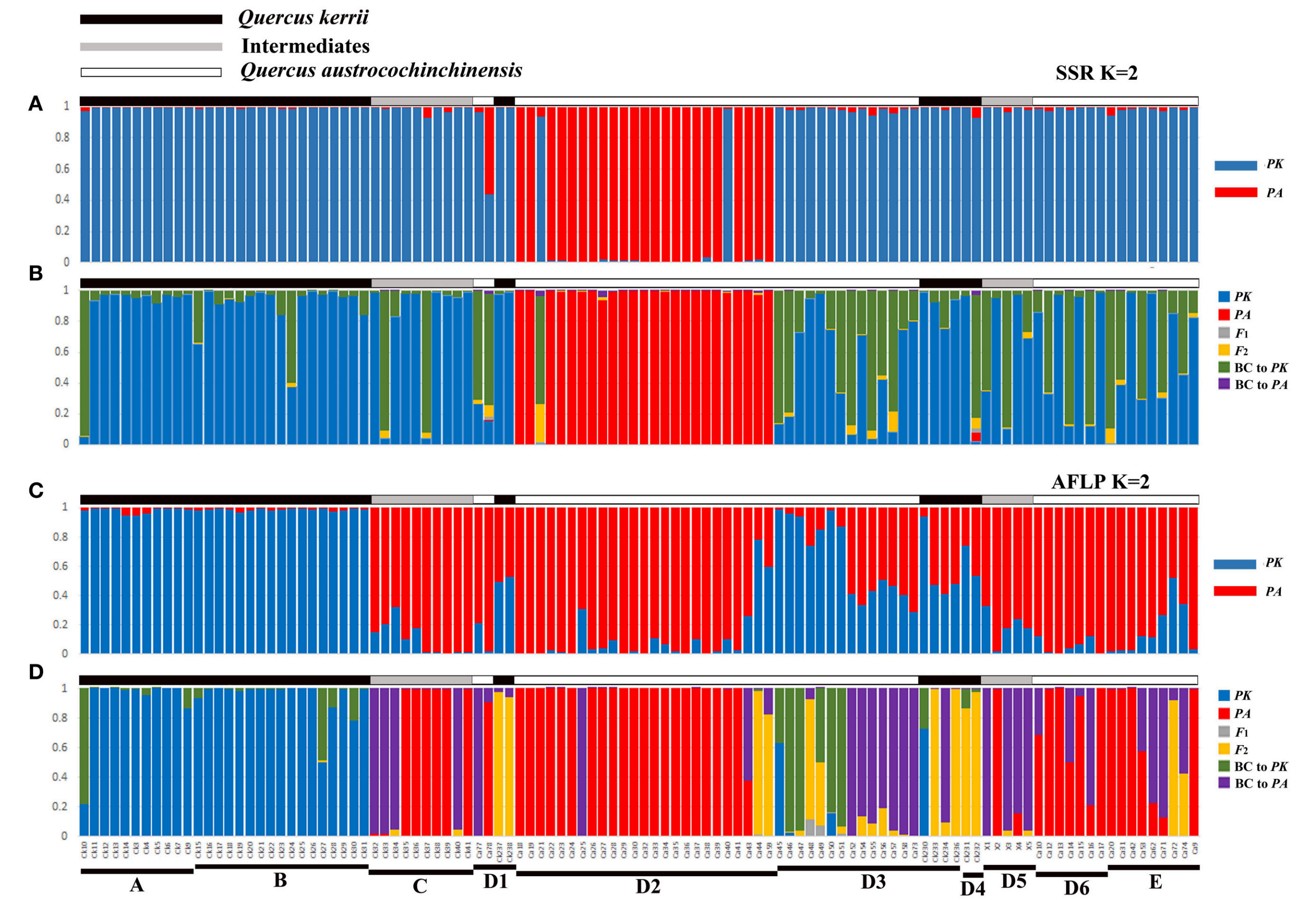


FIGURE 3 | Genotype class assignment of all 108 individuals of *Q. austrocochinchinensis*, *Q. kerrii*, and putative hybrids based on the programs STRUCTURE and NewHybrids using SSR (A,B) and AFLP **(C,D)** data. Bayesian assignment proportion for K = 2 cluster was determined in STRUCTURE for 11 SSR loci **(A)** and 781 AFLP loci **(C)**. Genotype class assignment using NewHybrids was based on 11 SSR loci **(B)** and 100 AFLP loci with highest F_{ST} **(D)**. There is a one-to-one correspondence among individuals between the four histograms.

NewHybrids Analysis

Analysis of SSR data using NewHybrids did not obtain results in agreement with the morphological identity of *Q. kerii* and *Q. austrocochinchinensis* individuals examined. Most individuals previously identified as *Q. austrocochinchinensis*, based on morphological characters from subpopulation D2, were still assigned to *Q. austrocochinchinensis*, with high posterior probabilities (>0.974), except for one individual (Ca21). The remaining individuals identified as *Q. austrocochinchinensis* in D1, D3, D6, and E were assigned to either *Q. kerrii* or backcrosses to *Q. kerrii*, as well as the morphological intermediates individuals in Population C and D5. Among the 36 individuals identified as *Q. kerrii*, most were assigned to *Q. kerrii*, with mediate to high probabilities, two (CK10, CK232) were assigned to backcrosses to *Q. kerrii*, with probabilities of 0.935 and 0.802, respectively (**Figure 3B**).

The assignment of individuals to certain genotypes using AFLP data better fits the morphological identification, but is still very messy. The results obtained from 100, 279, and 450 loci of AFLP were similar in terms of assigning the individuals of *Q. austrocochinchinensis* and *Q. kerrii* from the D region, but the results were different in terms of assigning the individuals in *Q. kerrii* purebred populations A and B, as more backcrosses to *Q. kerrii* were found when adding more loci with low divergence of F_{ST} to the analysis (**Figure 3**, **Figures S5B,C**). To accurately estimate hybrid classes, alleles with large divergence of F_{ST} are required, or misclassification of backcrosses and purebred parental individuals are very likely to happen (Vähä and Primmer, 2006). Therefore, the NewHybrids results obtained from 100 AFLP loci with highest F_{ST} value were more reliable and sensitive on assigning the genealogical classes.

Based on morphological features, 57 individuals were identified as *Q. austrocochinchinensis* in populations D and E, of which 30 were assigned to pure *Q. austrocochinchinensis*, one was assigned to F_2 hybrids, six to backcrosses to *Q. austrocochinchinensis*, and four to backcrosses to *Q. kerrii* with high probabilities. The remaining genotypes were mixtures of certain amounts of *Q. kerrii*, backcrosses to both parental species, and F_2 (**Figure 3D**). Among the 15 morphological intermediate individuals of populations C and D5, seven were assigned to

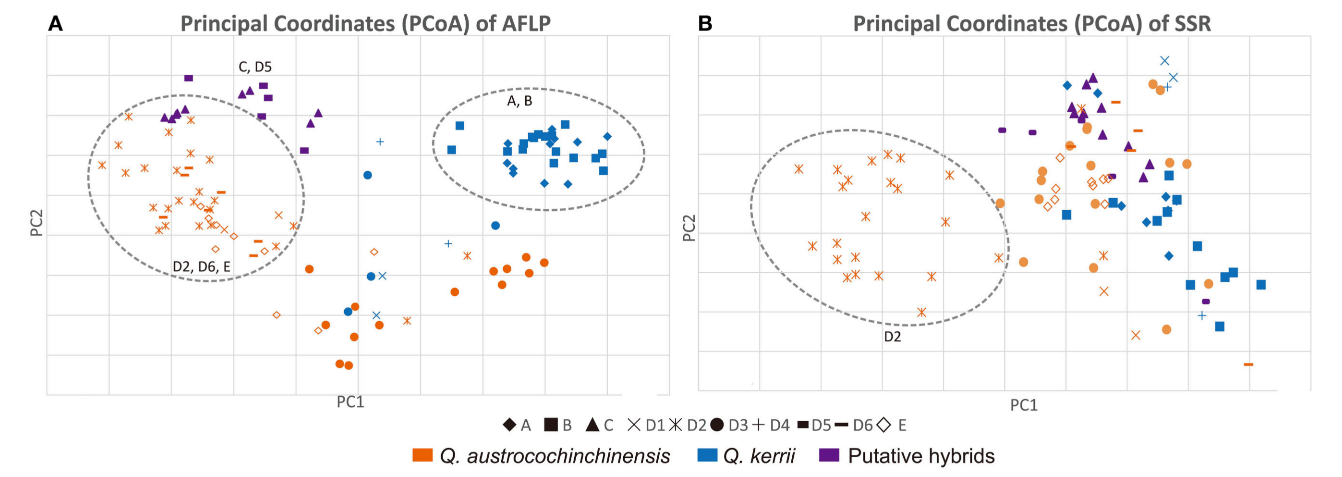


FIGURE 4 | PCoA plots of AFLP (A) and SSR **(B)** data. Shapes of the data points on the plot indicate a particular population, as “◆”represents Population A; “■” represents Population B; “▲” represents Population C;“”represents Population E; The rest symbols belongs to Population D, of which “×” represents sub-population D1; “Ж” represents sub-population D2; “●” represents sub-population D3; “+” represents subpopulation D4; “■” represents subpopulation D5; “▬” represents sub-population D6; *Q. austrocochinchinensis*, *Q. kerrii* and morphological intermediates are represented by orange, blue, and purple, respectively. Clearly clustered individuals are shown within a dashed line circle.

Q. austrocochinchinensis purebreds, seven to backcrosses to *Q. austrocochinchinensis* with high probabilities, and one was an admixture of *Q. austrocochinchinensis* and its backcrosses. The morphologically identified individuals of *Q. kerrii* in A and B populations were predominantly assigned to pure *Q. kerrii.* The remaining genotypes were mixtures, with few admixtures of certain amounts of *Q. kerrii* and backcrosses to *Q. kerrii.* In D3 and D4, no purebred parental individuals existed, three individuals were assigned to F_2 with high probability. The genotypes the remaining individuals were either backcrosses to both parental species or a wide spectrum of admixtures of the backcrosses to both parental species, F_1, F_2, and *Q. kerrii* purebreds. Of the six individuals of *Q. kerrii* in subpopulations D3 and D4, three were assigned to backcrosses to *Q. austrocochinchinensis*, one to backcross to *Q. kerrii*, and two were identified as F_2 (**Figure 3D**).

Principal Coordinate Analysis

The PCoA results of the AFLP and the SSR data are shown in **Figures 4A,B**, respectively. The PCoA results were in agreement with the results of the STRUCTURE analysis. Individuals of the two species were mostly separated for both AFLP and SSR data, which indicated that interspecific differentiation was stronger than intraspecific differentiation.

For the AFLP data, most *Q. austrocochinchensis* individuals in D2, D6, and E showed little mixing with *Q. kerrii* and were grouped together (left dashed line circle in **Figure 4A**). The *Q. kerrii* purebred populations A and B were grouped unambiguously (right dashed line circle in **Figure 4A**). Populations D1 and D3, which contained both species, were grouped between the two dashed line circles that represented each species. The morphological intermediates in C and D5 were grouped together in the middle position and biased toward *Q. austrocochinchensis.* For SSR analysis, most *Q. austrocochinchinensis* individuals in D2 were grouped together with smaller PC1 values (dashed line circle in **Figure 4B**). Other samples did not separate into distinct clusters. Still, the result of the PCoA seems to be more reliable, as it revealed more groups than that inferred by STRUCTURE, which only recognized two groups. Except for D2, the putative pure parental species and hybrids were mostly grouped together, but the resolution was not as high as the AFLP data.

DISCUSSION

Hybridization and Introgression between *Q. kerrii* and *Q. austrocochinchinensis*

Quercus austrocochinchinensis is a rare species with only two known distribution sites in China: Xi-Shuang-Ban-Na (XSBN) Nature Reserve in Yunnan province and Ba-Wang-Ling (BWL) Nature Reserve in Hainan Province (Huang et al., 1999; Song et al., 2015). Trees with intermediate morphological form between the two parental species have been found previously in XSBN, suggesting ongoing natural hybridization (Song et al., 2015). However, morphological intermediacy is not invariably associated with hybrids, as it can be a result of hybridization or phenotypic plasticity of the species in these areas (Rieseberg et al., 1993). Therefore, one major objective of this study is to confirm the possible hybridization and introgression in the two species.

Overall, our study revealed that there were fewer *Q. austrocochinchinensis* purebreds than previously expected based on morphological diagnostic traits. Although AFLP and SSR have different distinguishing power on the genotypes of examined samples, both markers revealed the presence of hybridization and introgression between the two species. The SSR data indicated that subpopulation D2 of *Q. austrocochinchinensis* has a unique germplasm composition and might be the only existing purebred. Further STRUCTURE analysis identified that only one individual (Ca78), located at contact zone subpopulation D1, is F_1. NewHybrids analysis is more sensitive to assign the genotype to different genetic categories and detected a general presence of backcrosses to *Q. kerrii*, both in morphologically identified *Q. kerrii* and *Q. austrocochinchinensis* and the intermediate populations, but no F_1, F_2, or backcrosses to *Q. austrocochinchinensis*. The SSR results indicate unidirectional introgression from *Q. kerrii* to *Q. austrocochinchinensis*.

For the AFLP data set, both Bayesian clustering approaches used (implemented in STRUCTURE and NewHybrids) detected an unexpected high number of backcrosses and hybrid genotypes. The threshold values and loci used on assigning individuals to different genetic categories are different in STRUCTURE and NewHybrids. Therefore, the two methods provided different percentages on F_1 and F_2 hybrids and backcrosses. STRUCTURE inferred the existence of F_1 hybrids and backcrosses, whereas NewHybrids inferred absences of F_1, but predominant backcrosses. STURECTURE is more efficient to evaluate the presence of hybrids in wild populations, whereas NewHybrids algorithm explicitly searches for hybrid and parental classes with assumption of two parental classes, which generally showed higher assignments accuracy than STRUCTURE (Marie et al., 2011). The NewHybrids result suggested that the formation of first-generation hybrids is less likely to occur than the interbreeding of hybrids with purebreds or with other hybrids. The occurrence of F_2 hybrids and the predominance of first-generation backcrosses to both parental species also reflect recent hybridization between the two species. Meanwhile, the genotypes of some individuals fell between F_2 hybrids, the first-generation backcrosses, pure *Q. kerrii*, and pure *Q. austrocochinchinensis*, with NewHybrids having no category available to assign them. Those individuals might in reality be second- or later-generation backcrosses or hybrids. This phenomenon is quite prominent in subpopulations D3 and D4, as no purebred individuals exist and the individuals are all hybrids, with varying percentages of backcrossing and parental types. Such evidence indicates that the two locations (D3 and D4) contain historical and ongoing gene flow between the two species.

The intermediate individuals in populations C and D5 are morphologically similar to *Q. kerrii*, mainly in terms of the shallow cupule, persistent trichomes on the leaf abaxial surface, and a relatively thick bark. However, their sharp leaf apex and leaf margin teeth more resemble *Q. austrochochinensis* (Song et al., 2015). The STRUCTURE results of the AFLP and SSR data on individuals of C and D5 were contradictory. The AFLP result indicated that individuals of C and D5 were *Q. austrocochinchinensis* purebred and backcrosses to *Q. austrocochinchinensis*, but the SSR suggested that all individuals in C and D5 were *Q. kerrii* purebreds. In the PCoA analysis, morphologically intermediate individuals from C and D5 were grouped together and located between two parental purebreds (**Figure 4**). Interestingly, C and D5 are geographically distant. Population D5 is located between regions with *Q. kerrii* and *Q. austrocochinchinensis*, but population C does not have any individuals with the typical features of *Q. austrocochinchinensis*; in addition, all the individuals are young trees, as the area is almost entirely occupied by farming land. The AFLP data revealed a high percentage germplasm of *Q. austrocochinchinensis* in C and D5. A similar situation was also found in population E and subpopulations D1 and D6, indicating that they are genetically "swamped" and that this "swamping" occurred recently, as a high percentage of germplasms of *Q. austrocochinchinensis* can still be detected.

A hybrid swarm is often characterized by a wide spectrum of phenotypic variation, the existence of backcrosses, and high genetic variation (Cockayne and Allan, 1926; Keim et al., 1989). Gathering all the evidences from molecular markers and morphology, there is incidence that a hybrid swarm had been established at the contact region of *Q. kerrii* and *Q. austrocochinchinensis* in XSBN Nature Reserve, especially at populations D3 and D4. The trees of *Q. austrocochinchinensis* in the adjacent regions (locations C, D1, D3–D6, and parts of D2) might have been already genetically swamped. The same situation might also have occurred in population E in Hainan, as SSR indicated that no *Q. austrocochinchinensis* purebreds exist, and the AFLP inferred co-existence of *Q. austrocochinchinensis* purebreds and its backcrosses.

It is worth noting that NewHybrids detected that the genotypes of few individuals contain different admixture levels of backcrosses to *Q. kerrii* and *Q. kerrii* purebreds in the two *Q. kerrii* purebred populations A and B (e.g., CK10, CK27, and CK30), which are distant from all the known populations of *Q. austrocochinchinensis*. Interspecific gene flow is a widespread and ongoing process among oaks, especially in species with close genetic relationship (Coart et al., 2002; Burgarella et al., 2009; Lepais et al., 2009; Moran et al., 2012). There is strong evidence that Neogene climatic changes had little impact on plant distribution in tropical and subtropical Asia (An, 2000; Su et al., 2013; Jacques et al., 2014), although evergreen oaks and other Fagaceae still experienced range shifts in this region (Xu et al., 2015; Jiang et al., 2016; Sun et al., 2016). A recent biogeographical study has indicated that the tropical zone could have extended further north in the geological past than it does today, e.g., the line 20°30′N was the northern biogeographical boundary of the tropical zone in south and southeastern China during the mid-Holocene (Zhu, 2013). The germplasm of *Q. austrocochinchinensis* recovered in *Q. kerrii* purebred populations probably reflects the historical gene flow between the two species and a once wider distribution of *Q. austrocochinchinensis* at this geological time in Indo-China. Future work will need to include more populations of both species, using both maternal markers and high throughput SNP markers to explore the genetic structure and couple the niche modeling to estimate the historical population size; such an approach could provide a better understanding on the genetic patterns of the two oak species.

Ecological Preference

Previous leaf anatomical work has suggested that population E is a *Q. austrocochinchinensis* purebred population with very distinct features compared to *Q. kerrii,* such as narrow leaves, sharp leaf apex, and margin tips (Song et al., 2015). However, the AFLP data showed evidences of hybridization in population

E. Across all surveyed populations, only D2 was identified as a pure *Q. austrocochinchinensis* subpopulation by both molecular markers. It is situated in the core region of the Xi-Shan-Ban-Na Nature Reserve with thick woods and geographically isolated from the *Q. kerrii* population, which likely reduces the chances of gene flow and hybridization. Population D2 has a much higher forest canopy density than that of other populations, including population E.

According to our field observation, *Q. austrocochinchinensis* and *Q. kerrii* have different habitat preferences, as the former is likely to grow in closed and moist forests, but the latter tends to grow on open slopes or by roadsides (**Figure S6**). The fine-scale closed and shady habitats with high humidity might favor the growth of *Q. austrocochinchinensis*, but are also crucial to maintain its genetic distinction. Subpopulations D3, D4, and D5 are located in the buffer area of XSBN Nature Reserve, where human activities are intensified. These regions were inferred as the location of the hybrid swarm of the two species based on our study. Opening habitat as a result of deforestation will likely favor colonization by *Q. kerrii* and facilitate pollen invasion. Our study also showed that introgression of the two oaks occurred at disturbed habitats, and pure *Q. austrocochinchinensis* was found at core protection areas with minimum disturbance. Habitat selection is intimately tied to niche differentiation and coexistence in plant communities (Bazzaz, 1991). Although there are no studies demonstrating the habitat preference of these two species, our data suggests that fine-scale heterogeneous habitats may play an important role in shaping the genetic structure of the two species in areas where they co-exist.

Discrepancies between AFLP and SSR Data

The results of STRUCTURE and NewHybrids showed discrepancies between AFLP and SSR markers (**Figure 3**). For many samples, AFLP and SSR provided different classifications of species from the same set of samples. Such similar results have also been detected in other plant groups, e.g., *Abies ziyuanensis* (Tang et al., 2008) and the arctic-alpine genus *Draba* (Skrede et al., 2009). This is probably caused by the limited genetic differentiation between the two species, especially in a fine sampling scale. If the two species are not fully differentiated, they will have a number of incompletely differentiated alleles, which may reduce the diagnostic power to distinguish one species from another. In the cases of species with limited genetic differentiation, sampling large numbers of loci across the genome is required when using a molecular approach. The AFLP method has advantages over the SSR method in sampling loci numbers; consequently, AFLP generally reveals higher polymorphism than SSR does (Varshney et al., 2007; Sun et al., 2008; Skrede et al., 2009); it also has a higher assignment success and solution compared to SSR loci (Zeng et al., 2010). Our results also demonstrated that the AFLP data was more consistent with morphology in STRUCTURE and NewHybrids analysis, and the two species separated much better based on using AFLP in PCoA than by using SSR markers. Therefore, we speculate that the results generated from AFLP are more reliable than those obtained from SSR. Our results also demonstrated that when studying hybridization between two genetically closely related oaks, high throughput markers e.g., AFLP or RAD, are more appropriate.

In another aspect, the discrepancies between the two molecular markers might be due to different selective pressures. Similar to the results found by Scotti-Saintagne et al. (2004), the F_{ST} values of 781 AFLP alleles fit an L-shaped curve. Most alleles (60.3%) have low F_{ST} values ($F_{ST} < 0.1$). The results of BayeScan also suggested that only three loci (0.38%) are potentially under selection among the 10 populations. Therefore, most of the AFLP markers are selectively neutral. However, SSR markers may have been subjected to more selective pressure. Morgante et al. (2002) mentioned that SSR repeats tend to occur at transcribed regions of the genome. Selective pressures acting on coding regions are higher than in non-coding regions. In turn, such pressure may lead to genetic differentiation and as a result, SSR loci may have higher genetic differentiation levels than AFLP. We particularly used SSR loci that had been developed from coding sequences, e.g., Qk15874, Qk17139, Qk17611, Qk20944, and CR627959.

Although SSR provided less information than AFLP, it is still a valuable method to compare different marker types. For a pair of species undergoing hybridization, the impermeable genomic regions that are under high selective pressure often serve as a way to maintain species integrity, and thus, these regions accumulate genetic divergence, whereas the regions with low selective pressure are permeable to introgression and show decreased differentiation (Wu, 2001). Because of the higher selective pressures, SSR is more representative of regions with high species integrity, while AFLP is more likely to reflect the effects of gene flow. It is important to note that SSR identified more *Q. kerrii* than AFLP in the STRUCTURE analysis. Moreover, in the PCoA analysis for AFLP, the morphological intermediate populations C and D5 were clustered with *Q. austrocochinchinensis* purebreds. However, for SSR, they were closer to *Q. kerrii* purebreds instead. As discussed above, we suggest that although the trees in the swarm were influenced by the gene flow of *Q. austrocochinchinensis*, the alleles of *Q. kerrii* were preferred in these hybrids. If this is the case, it indicates the high genetic extinction risk of *Q. austrocochinchinensis*.

Discrepancies between Morphological and Molecular Approaches

Our study revealed the discrepancies between the results inferred by morphological traits and molecular markers. For example, although morphological intermediates only make up a small proportion of samples in the hybrid zone (15 of 61), the molecular method detected more hybrids than expected. Moreover, individuals with one parental morphological characteristic may be identified as another parental purebred, e.g., a part of trees from population D3. There may be two explanations for these discrepancies. First, trees with clear parental characteristics could in fact be F_1 hybrids or backcrosses, and the morphological phenotypes may be a result of some of the loci having dominant effects on controlling a certain morphological trait. Moreover, hybrids that are in the second or later generation may return to homozygosis at specific loci and thus resemble one parent only (Stebbins, 1950; Rieseberg et al., 1993). However, this offers only a partial explanation, because of the low chance that all loci controlling morphological diagnostic characteristics return to homozygosis at the same time.

Another possibility of the inconsistency is that the phenotype with one of the parental phenotype may be under positive selection in natural habitats, as described previously (Hipp and Weber, 2008). As discussed above, the two species appear to have different habitat preferences. The two populations with typical *Q. austrocochinchinensis* in appearance were all located in the national nature reserve, where canopy density is usually high. The stellate and fused fascicular trichomes and dense stomata, as well as the thick leaf cuticles are important features in seasonal dry areas with strong sun light (Mediterranean-type) climate (Tattini et al., 2007). The glabrous leaves with sharp, spiny teeth were generally an adaptation trait to humid and shady environments (Sun et al., 2003), which may explain why the trees are distributed in dense canopy areas of the Xi-Shuang-Ban-Na Nature Reserve of Yunnan Province and Ba-Wang-Ling Nature Reserves likely show the phenotypes of typical *Q. austrocochinchinensis*, although introgression is common, as this phenotype might have a selective advantage over the hairy, obtuse toothed phenotype in ravine habitats where the environment is usually humid and shaded. Meanwhile, we also could not rule out the possibility that the morphological variation in *Q. kerrii* might encompass the typical morphology of *Q. austrocochinchinensis*. In this case, the morphological features used to distinguish *Q. austrocochinchinensis* from *Q. kerrii* need to be further clarified. Common garden experiment coupling the high throughput genotyping on the different populations of the two taxa and their hybrids are need to reveal how and why traits evolved in response to nature selection and local adaptation in the future.

Potential Threats to *Q. austrocochinchinensis* and Conservation Strategies

Our data indicates ongoing and historical introgression in the two *Q. austrocochinchinensis* populations. Although the introgression directions inferred by SSR and AFLP are not the same (as SSR indicated unidirectional introgression from *Q. kerrii* to *Q. austrocochinchinensis*, but bidirectional introgression was inferred by AFLP). BayeScan analysis indicated most AFLP loci were selection neutral, but almost all the SSR loci violated HWE assumption. Therefore, AFLP result is more reliable to infer the gene flow between the two species. Recently simulation model study demonstrated when outercrossing rate between the two parental species is different, hybridization can facilitate invasions of the species with high outcrossing rate, even without enhancing local adaption (Mesgaran et al., 2016). Considering the fast shrinking and degradation of habits favoring *Q. austrocochinchinensis* at middle to low elevation of tropical Asia, the ongoing hybrid swarm and the predominance of *Q. kerrii* at co-occur region, *Q. austrocochinchinensis* faces critical extinction risks both in terms of lost habitat and genetic assimilation. Furthermore, our results indicate that only subpopulation D2 maintains its genetic integrity, as the most pure individuals of *Q. austrocochinchinensis* are still purebreds. The other subpopulations in XSBN Nature Reserve and adjacent areas formed a hybrid swarm. Once introgression is detected in an endangered species, protective intervention measures should be adopted immediately to prevent species integrity loss. Otherwise, all subpopulations in the reserve could become hybrid swarms.

Firstly, *in situ* conservation of species genetic resource can be realized by establishing nature reserves. According to our results, the most pure *Q. austrocochinchinensis* population D2 has the highest forest canopy density and lowest human disturbance. Although it is difficult to prove the correlation between ecological integrity and hybrid degree in this case study, we hypothesize that it is important for *in situ* conservation to maintain the existing ecological balance of the habitat of population D2. To avoid further asymmetrical introgressions, forest landscape restoration is also essential for the hybrid *Q. austrocochinchinensis* populations. Secondly, considering the abundance of *Q. kerrii* and the efficient pollen dispersal abilities of oaks, *ex situ* conservation of pure *Q. austrocochinchinensis* should also be considered. Although we cannot claim that *Q. austrocochinchinensis* homozygous individuals still exist, molecular markers and leaf morphological features all identified D2 as the most pure *Q. austrocochinchinensis* population. Therefore, we suggest that the D2 population area should be designated as the priority conservation zone.

CONCLUSION

This study suggests that *Q. austrocochinchinensis* in China is experiencing introgression from *Q. kerrii*. The incoherent genetic structure inferred by AFLP and SSR, as well as morphological traits, might be due to the different selective pressures. The extinction risk of *Q. austrocochinchinensis* is higher than previously expected, as much less *Q. austrocochinchinensis* purebreds were detected based on molecular markers; in addition, the species faces ongoing hybrid swarm with *Q. kerrii* and habitat loss in tropical Asia. The subpopulation D2 of *Q. austrocochinchinensis* in the core area of the XSBN Nature Reserve, with unique germplasm and vulnerable to disturbance, should be prioritized for protection. The habitat with high forest canopy density and humidity may act on shaping the genetic structure of the two species at the contact zone. Further studies using high throughput molecular markers and coupling the environmental parameters at fine scale to study the genetic diversity patterns at the co-occurring area of the two species and scan more *Q. kerrii* purebred populations can provide a better understanding on the dynamics of the hybrid zone and determine the underlying genes involved in local adaptation.

AUTHOR CONTRIBUTIONS

MA and MD designed the study, wrote, and revised the manuscript; MD was responsible for the manuscript submission.

SZ and MA performed the experiments; XJ assisted to analyze the data; MD and YS collected the plant materials.

ACKNOWLEDGMENTS

This work was supported by funding from the National Natural Science Foundation (NSFC) (31270267), the Shanghai Municipal Administration of Forestation and City Appearances (G162404, G162405, G152428), the Science and Technology Commission of Shanghai Municipality (14DZ2260400). We are grateful to Mr. Quan-Jian LI for helping to collect the plant samples and DNA extraction.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00229/full#supplementary-material

Figure S1 | Allele distributions in *Q. austrocochinchinensis* and *Q. kerrii* at 11 target loci. (A–K) Allele frequency of 11 SSR loci. Alleles are arranged on the x-axis and allele frequencies on the y-axis. Blue bar indicates *Q. austrocochinchinensis*; Orange bar indicates *Q. kerrii*. **(A)** Locus GB371; **(B)** locus QK20944; **(C)** locus QK17611; **(D)** locus QK17139; **(E)** locus QK15874; **(F)** QpZAG36; **(G)** locus QpZAG16; **(H)** locus QpZAG110; **(I)** locus QpZAG9; **(J)** locus CR627959; **(K)** locus QmC00963. **(L)** Mean allelic patterns between *Q. austrocochinchinensis* and *Q. kerrii*. **Na**, Number of different alleles; **Na Freq. $\geq$ 5%**, number of different alleles with a frequency $\geq$ 5%; **Ne**, number of effective alleles; **I,** Shannon's Information Index; No. Private Alleles, number of alleles unique to *Q. austrocochinchinensis* and *Q. kerrii*; **No. LComm Alleles** ($\leq$ **25%**), Number of locally common alleles (freq. $\geq$ 5%) found in 25% or fewer Populations; **No. LComm Alleles** ($\leq$ **50%**), number of locally common alleles (freq. $\geq$ 5%) found in 50% or fewer populations; $\boldsymbol{H_E}$, Expected Heterozygosity; **u**$\boldsymbol{H_E}$, unbiased expected heterozygosity.

Figure S2 | $\boldsymbol{F_{ST}}$ value distribution of AFLP and SSR loci between *Q. austrocochinchinensis* and *Q. kerrii*. The x-axis represents the F_{ST} values and the y-axis represents the number of loci.

Figure S3 | BayeScan plots of 781 AFLP loci in 10 sampled populations of *Q. austrocochinchinensis* and *Q. kerrii*. The vertical read line is the threshold ($\text{Log}_{10}(\text{PO}) = 2$) used for identifying outlier loci. Dots that fall to the right of the threshold line are identified as outlier loci.

Figure S4 | Changes of $\boldsymbol{\Delta K}$ from each $\boldsymbol{K}$ cluster in program STRUCTURE. ΔK is used to identify the most likely number of clusters. For the AFLP and SSR data, $K = 2$ was the most likely K value.

Figure S5 | Genotype class assignment of all 108 individuals of *Q. austrocochinchinensis*, *Q. kerrii*, and putative hybrids based on the programs InStruct and NewHybrids using SSR (A) and AFLP **(B,C)** data. $K = 2$ cluster was determined in InStruct for 11 SSR loci **(A)**. The 249 **(B)** and 450 **(C)** AFLP loci with highest F_{ST} were analyzed, respectively, using NewHybrids.

Figure S6 | Habitat preference of *Q. austrocochinchinensis* and *Q. kerrii.* *Q. austrocochinchinensis* tends to grow in closed and moist habitat **(A)**, while *Q. kerrii* prefers open and dry habitat **(B)**.

REFERENCES

An, M., Deng, M., Zheng, S. S., and Song, Y. G. (2016). *De novo* transcriptome assembly and development of SSR markers of oaks *Quercus austrocochinchinensis* and *Q. kerrii* (Fagaceae). *Tree Genet. Genomes* 12, 103. doi: 10.1007/s11295-016-1060-5

An, Z. S. (2000). The history and variability of the East Asian paleomonsoon climate. *Quat. Sci. Rev*. 19, 171–187. doi: 10.1016/S0277-3791(99)00060-8

Anderson, E. C. (2008). Bayesian inference of species hybrids using multilocus dominant genetic markers. *Philos. Trans. R. Soc. B Biol. Sci*. 363, 2841–2850. doi: 10.1098/rstb.2008.0043

Anderson, E. C., and Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* 160, 1217–1229.

Arnold, M. L. (1992). Natural hybridization as an evolutionary process. *Annu. Rev. Ecol. Syst*. 23, 237–261. doi: 10.1146/annurev.es.23.110192.001321

Barton, N. (2001). The role of hybridization in evolution. *Mol. Ecol.* 10, 551–568. doi: 10.1046/j.1365-294x.2001.01216.x

Bazzaz, F. A. (1991). Habitat selection in plants. *Am. Nat.* 137, s116–s130. doi: 10.1086/285142

Briggs, D., and Walters, S. M. (1997). *Plant Variation and Evolution*. Cambridge: Cambridge University Press.

Burgarella, C., Lorenzo, Z., Jabbour-Zahab, R., Lumaret, R., Guichoux, E., Petit, R. J., et al. (2009). Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). *Heredity* 102, 442–452. doi: 10.1038/hdy.2009.8

Burger, W. (1975). The species concept in Quercus. *Taxon* 24, 45–50. doi: 10.2307/1218998

Cavender-Bares, J., and Pahlich, A. (2009). Moluecular, morphological and ecological niche differentiation of sympatric sister oak species, *Quercus virginiana* and *Q. geminata* (Fagaceae). *Am. J. Bot.* 96, 1690–1702. doi: 10.3732/ajb.0800315

Coart, E., Lamote, V., De Loose, M., Van Bockstaele, E., Lootens, P., and Roldán-Ruiz, I. (2002). AFLP markers demonstrate local genetic differentiation between two indigenous oak species [*Quercus robur* L. and *Quercus petraea* (Matt.) Liebl.] in Flemish populations. *Theor. Appl. Genet.* 105, 431–439. doi: 10.1007/s00122-002-0920-6

Cockayne, L., and Allan, H. H. (1926). The naming of wild hybrid swarms. *Nature* 118, 623–624. doi: 10.1038/118623a0

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

Craft, K. J., Ashley, M. V., and Koenig, W. D. (2002). Limited hybridization between *Quercus lobata* and *Quercus douglasii* (Fagaceae) in a mixed stand in central coastal California. *Am. J. Bot.* 89, 1792–1798. doi: 10.3732/ajb.89.11.1792

Curtu, A. L., Gailing, O., and Finkeldey, R. (2007). Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community. *BMC. Evol. Bio.* 7:e218. doi: 10.1186/1471-2148-7-218

Deng, M. (2007). *Anatomy, Taxonomy, Distribution & Phylogeny of Quercus subg. Cyclobalanopsis (Oersted) Schneid. (Fagaceae)*. dissertation/PhD's thesis. Kunming Institute of Botany, Chinese Academy of Sciences, Kunming.

Deng, M., Zhou, Z. K., and Li, Q. S. (2013). Taxonomy and systematics of Quercus subgenus Cyclobalanopsis. *International Oaks* 24, 48–60.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587.

Fischer, M. C., Foll, M., Excoffier, L., and Heckel, G. (2011). Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). *Mol. Ecol.* 20, 1450–1462. doi: 10.1111/j.1365-294X.2011.05015.x

Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics* 180, 977–993. doi: 10.1534/genetics.108.092221

Gao, H., Williamson, S., and Bustamante, C. D. (2007). A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* 176, 1635–1651. doi: 10.1534/genetics.107.072371

Govaerts, R., and Frodin, D. G. (1998). *World Checklist and Bibliography of Fagales (Betulaceae, Corylaceae, Fagaceae and Ticodendraceae)*. London: Kew Publishing.

Grant, V. (1981). *Plant Speciation*. New York, NY: Columbia University Press.

Harrison, R. G. (1993). *Hybrid Zones and the Evolutionary Process*. Oxford: Oxford University Press.

Hipp, A. L., and Weber, J. A. (2008). Taxonomy of Hill's oak (*Quercus ellipsoidalis*: Fagaceae): evidence from AFLP data. *Syst. Bot.* 33, 148–158. doi: 10.1600/036364408783887320

Huang, C. C., Chang, Y. T., and Bartholomew, B. (1999). "Fagaceae," in *Flora of China*, *Vol. 4*, eds C. Y. Wu and P. H. Raven (Beijing; St. Louis, MI: *Science Press and Missouri Botanical Garden Press)*, 380–400.

Huang, J., Ge, X., and Sun, M. (2000). Modified CTAB protocol using a silica matrix for isolation of plant genomic DNA. *BioTechniques* 28, 432–434.

Jacques, F. M. B., Su, T., Spicer, R. A., Xing, Y. W., Huang, Y. J., and Zhou, Z. K. (2014). Late Miocene southwestern Chinese floristic diversity shaped by the southeastern uplift of the Tibetan Plateau. *Paleogeogr. Paleoclimatol. Paleoecol.* 411, 208–215. doi: 10.1016/j.palaeo.2014.05.041

Jeffreys, H. (1961). *Theory of Probability, 3rd Edn*. London: *Oxford University Press*, 432.

Jiang, X. L., Deng, M., and Li, Y. (2016). Evolutionary history of subtropical evergreen broad-leaved forest in Yunnan Plateau and adjacent areas: an insight from *Quercus schottkyana* (Fagaceae). *Tree Genet. Genomes* 12, 104. doi: 10.1007/s11295-016-1063-2

Keim, P., Paige, K. N., Whitham, T. G., and Lark, K. G. (1989). Genetic analysis of an interspecific hybrid swarm of Populus: occurrence of unidirectional introgression. *Genetics* 123, 557–565.

Kleinschmit, J. R., Bacilieri, R., Kremer, A., and Roloff, A. (1995). Comparison of morphological and genetic traits of pedunculate oak (*Quercus robur* L.) and sessile oak (*Q. petrea* (Matt.) Liebl.). *Silvae. Genet.* 44, 256–269.

Kremer, A., Dupouey, J. L., Deans, J. D., Cottrell, J., Csaikl, U., Finkeldey, R., et al. (2002). Leaf morphological differentiation between *Quercus robur* and *Quercus petraea* is stable across western European mixed oak stands. *Ann. For. Sci.* 59, 777–787. doi: 10.1051/forest:2002065

Leinonen, T., McCairns, R. J. S., O'Hara, R. B., and Merila, J. (2013). QST-FST comparisons: evolutionary and ecological insights from genomic heterogeneity. *Nat. Rev. Genet.* 14, 179–190. doi: 10.1038/nrg3395

Lepais, O., Petit, R., Guichoux, E., Lavabre, J., Alberto, F., Kremer, A., et al. (2009). Species relative abundance and direction of introgression in oaks. *Mol. Ecol.* 18, 2228–2242. doi: 10.1111/j.1365-294X.2009.04137.x

Levin, D. A., Francisco-Ortega, J., and Jansen, R. K. (1996). Hybridization and the extinction of rare plant species. *Conserv. Biol.* 10, 10–16. doi: 10.1046/j.1523-1739.1996.10010010.x

Lexer, C., Kremer, A., and Petit, R. J. (2006). Shared alleles in sympatric oaks: recurrent gene flow is a more parsimonious explanation than ancestral polymorphism. *Mol. Ecol.* 15, 2007–2012. doi: 10.1111/j.1365-294X.2006.02896.x

López-Caamal, A., and Tovar-Sánchez, E. (2014). Genetic, morphological, and chemical patterns of plant hybridization. *Rev. Chil. Hist. Nat.* 87, 1–14. doi: 10.1186/s40693-014-0016-0

Lou, Y., and Zhou, Z. K. (2001). Phytogeography of Quercus subg. Cyclobalanopsis. *Acta Bot. Yunnan* 23, 1–16.

Lynch, M., and Milligan, B. G. (1994). Analysis of population genetic structure with RAPD markers. *Mol. Ecol.* 3, 91–99. doi: 10.1111/j.1365-294X.1994.tb00109.x

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20, 229–237. doi: 10.1016/j.tree.2005.02.010

Mallet, J. (2007). Hybrid speciation. *Nature* 446, 279–283. doi: 10.1038/nature05706

Marie, A. D., Bernatchez, L., and Garant, D. (2011). Empirical assessment of software efficiency and accuracy to detect introgression under variable stocking scenarios in brook charr (*Salvelinus fontinalis*). *Conserv. Genet.* 12, 1215–1227. doi: 10.1007/s10592-011-0224-y

Mesgaran, M. B., Lewis, M. A., Ades, P. K., Donohue, K., Ohadi, S., Li, C. J., et al. (2016). Hybridization can facilitate species invasions, even without enhancing local adaptation. *Proc. Natl. Acad. Sci. U.S.A.* 113, 10210–10214. doi: 10.1007/s10592-011-0224-y

Moran, E. V., Willis, J., and Clark, J. S. (2012). Genetic evidence for hybridization in red oaks (Quercus sect. Lobatae, Fagaceae). *Am. J. Bot.* 99, 92–100. doi: 10.3732/ajb.1100023

Morgante, M., Hanafey, M., and Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30, 194–200. doi: 10.1038/ng822

Muir, G., Fleming, C. C., and Schlotterer, C. (2000). Taxonomy - Species status of hybridizing oaks. *Nature* 405, 1016–1016. doi: 10.1038/35016640

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321

Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U.S.A.* 76, 5269–5273. doi: 10.1073/pnas.76.10.5269

Ortego, J., and Bonal, R. (2010). Natural hybridisation between kermes (*Quercus coccifera* L.) and holm oaks (*Q. ilex L.*) revealed by microsatellite markers. *Plant Biol.* 12, 234–238. doi: 10.1111/j.1438-8677.2009.00244.x

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Phengklai, C. (2006). A synoptic account of the Fagaceae of Thailand. *Thai For. Bull.* 34, 53–175.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Rhymer, J. M., and Simberloff, D. (1996). Extinction by hybridization and introgression. *Annu. Rev. Ecol. Syst.* 27, 83–109. doi: 10.1146/annurev.ecolsys.27.1.83

Rieseberg, L. H. (1995). The role of hybridization in evolution: old wine in new skins. *Am. J. Bot.* 82, 944–953. doi: 10.2307/2445981

Rieseberg, L. H. (1997). Hybrid origins of plant species. *Annu. Rev. Ecol. Syst.* 28, 359–389. doi: 10.1146/annurev.ecolsys.28.1.359

Rieseberg, L. H., Ellstrand, N., and Arnold, M. (1993). What can molecular and morphological markers tell us about plant hybridization? *Crit. Rev. Plant Sci.* 12, 213–241. doi: 10.1080/07352689309701902

Rousset, F. (2008). Genepop'007: a complete re-implementation of the Genepop software for Windows and Linux. *Mol. Ecol. Res.* 8, 103–106. doi: 10.1111/j.1471-8286.2007.01931.x

Salvini, D., Bruschi, P., Fineschi, S., Grossoni, P., Kjaer, E. D., and Vendramin, G. G. (2009). Natural hybridisation between *Quercus petraea* (Matt.) Liebl. and *Quercus pubescens* Willd. within an Italian stand as revealed by microsatellite fingerprinting. *Plant. Biol.* 11, 758–765. doi: 10.1111/j.1438-8677.2008.00158.x

Scotti-Saintagne, C., Mariette, S., Porth, I., Goicoechea, P. G., Barreneche, T., Bodénès, C., et al. (2004). Genome scanning for interspecific differentiation between two closely related oak species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]. *Genetics* 168, 1615–1626. doi: 10.1534/genetics.104.026849

Skrede, I., Borgen, L., and Brochmann, C. (2009). Genetic structuring in three closely related circumpolar plant species: AFLP versus microsatellite markers and high-arctic versus arctic-alpine distributions. *Heredity* 102, 293–302. doi: 10.1038/hdy.2008.120

Song, Y., Deng, M., Hipp, A., and Li, Q. J. (2015). Leaf morphological evidence of natural hybridization between two oak species (*Quercus austrocochinchinensis* and *Q. kerrii)* and its implications for conservation management. *Eur. J. For. Res.* 134, 139–151. doi: 10.1007/s10342-014-0839-x

Stebbins Jr. C. (1950). *Variation and Evolution in Plants*. New York, NY: Columbia University Press.

Steinkellner, H., Lexer, C., Turetschek, E., and Glössl, J. (1997). Conservation of (GA) n microsatellite loci between Quercus species. *Mol. Ecol.* 6, 1189–1194. doi: 10.1046/j.1365-294X.1997.00288.x

Su, T., Jacques, F. M. B., Spicer, R. A., Liu, Y. S., Huang, Y. J., Xing, Y. W., et al. (2013). Post-Pliocene establishment of the present monsoonal climate in SW China: evidence from the late Pliocene Longmen megaflora. *Clim. Past* 9, 1911–1920. doi: 10.5194/cp-9-1911-2013

Sun, B., N., Cong, P., Yan, D., and Xie, S. (2003). Cuticular structure of two angiosperm fossils in neogene from tengchong, yunnan province and its palaeoenvironmental significance. *Acta Palaeont. Sin.* 42, 216–222.

Sun, Q. B., Li, L. F., Li, Y., Wu, G. J., and Ge, X. J. (2008). SSR and AFLP markers reveal low genetic diversity in the biofuel plant *Jatropha curcas* in China. *Crop Sci.* 48, 1865–1871. doi: 10.2135/cropsci2008.02.0074

Sun, Y., Surget-Groba, Y., and Gao, S. (2016). Divergence maintained by climatic selection despite recurrent gene flow: a case study of *Castanopsis carlesii* (Fagaceae). *Mol. Ecol.* 25, 4580–4592. doi: 10.1111/mec.13764

Tamaki, I., and Okada, M. (2014). Genetic admixing of two evergreen oaks, *Quercus acuta* and *Q. sessilifolia* (subgenus Cyclobalanopsis), is the result of interspecific introgressive hybridization. *Tree Genet. Genomes* 10, 989–999. doi: 10.1007/s11295-014-0737-x

Tang, S., Dai, W., Li, M., Zhang, Y., Geng, Y., Wang, L., et al. (2008). Genetic diversity of relictual and endangered plant *Abies ziyuanensis* (Pinaceae) revealed by AFLP and SSR markers. *Genetica* 133, 21–30. doi: 10.1007/s10709-007-9178-x

Tattini, M., Matteini, P., Saracini, E., Traversi, M. L., Giordano, C., and Agati, G. (2007). Morphology and biochemistry of non-glandular trichomes in *Cistus salvifolius* L. leaves growing in extreme habitats of the Mediterranean basin. *Plant Biol.* 9, 411–419. doi: 10.1055/s-2006-924662

Tong, X., Xu, N. N., Li, L., and Chen, X. Y. (2012). Development and characterization of polymorphic microsatellite markers in *Cyclobalanopsis glauca* (Fagaceae). *Am. J. Bot.* 99, e120–e122. doi: 10.3732/ajb.1100448

Ueno, S., and Tsumura, Y. (2008). Development of ten microsatellite markers for *Quercus mongolica* var. crispula by database mining. *Conserv. Genet.* 9, 1083–1085. doi: 10.1007/s10592-007-9462-4

Ueno, S., Taguchi, Y., and Tsumura, Y. (2008). Microsatellite markers derived from *Quercus mongolica* var. crispula (Fagaceae) inner bark expressed sequence tags. *Genes Genet. Syst.* 83, 179–187. doi: 10.1266/ggs.83.179

Vähä, J. P., and Primmer, C. R. (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Mol. Ecol.* 15, 63–72. doi: 10.1111/j.1365-294X.2005.02773.x

Valbuena-Carabaña, M., González-Martínez, S. C., Sork, V. L., Collada, C., Soto, A., Goicoechea, P. G., et al. (2005). Gene flow and hybridisation in a mixed oak forest (*Quercus pyrenaica* Willd. and *Quercus petraea* (Matts.) Liebl.) in central Spain. *Heredity* 95, 457–465. doi: 10.1038/sj.hdy.6800752

Van Oosterhout, C., Hutchinson, W. F., Wills, D. P. M., and Shipley, P. (2004). Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* 4, 535–538. doi: 10.1111/j.1471-8286.2004.00684.x

Varshney, R. K., Chabane, K., Hendre, P. S., Aggarwal, R. K., and Graner, A. (2007). Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. *Plant Sci.* 173, 638–649. doi: 10.1016/j.plantsci.2007.08.010

Vekemans, X., Beauwens, T., Lemaire, M., and Roldán-Ruiz, I. (2002). Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. *Mol. Ecol.* 11, 139–151. doi: 10.1046/j.0962-1083.2001.01415.x

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23, 4407–4414. doi: 10.1093/nar/23.21.4407

Whitney, K. D., Ahern, J. R., Campbell, L. G., Albert, L. P., and King, M. S. (2010). Patterns of hybridization in plants. *Perspect. Plant Ecol. Evol. Syst.* 12, 175–182. doi: 10.1016/j.ppees.2010.02.002

Whittemore, A. T., and Schaal, B. A. (1991). Interspecific gene flow in sympatric oaks. *Proc. Natl. Acad. Sci. U.S.A* 88, 2540–2544. doi: 10.1073/pnas.88.6.2540

Wu, C. I. (2001). The genic view of the process of speciation. *J. Evol. Biol.* 14, 851–865. doi: 10.1046/j.1420-9101.2001.00335.x

Xu, J., Deng, M., Jiang, X. L., Westwood, M., Song, Y. G., and Turkington, R. (2015). Phylogeography of *Quercus glauca* (Fagaceae), a dominant tree of East Asian subtropical evergreen forests, based on three chloroplast DNA interspace sequences. *Tree Genet. Genomes* 11, 805. doi: 10.1007/s11295-014-0805-2

Zalapa, J. E., Brunet, J., and Guries, R. P. (2009). Patterns of hybridization and introgression between invasive *Ulmus pumila* (Ulmaceae) and native *U. rubra. Am. J. Bot.* 96, 1116–1128. doi: 10.3732/ajb.0800334

Zeng, Y. F., Liao, W. J., Petit, R. J., and Zhang, D. Y. (2010). Exploring species limits in two closely related Chinese oaks. *PLoS ONE* 5:e15529. doi: 10.1371/journal.pone.0015529.g001

Zhivotovsky, L. A. (1999). Estimating population structure in diploids with multilocus dominant DNA markers. *Mol. Ecol.* 8, 907–913. doi: 10.1046/j.1365-294x.1999.00620.x

Zhu, H. (2013). Geographical elements of seed plants suggest the boundary of the tropical zone in China. *Paleogeogr. Paleoclimatol. Paleoecol.* 386, 16–22. doi: 10.1016/j.palaeo.2013.04.007

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

PERSPECTIVE
published: 03 March 2017
doi: 10.3389/fgene.2017.00025

Integrative Approaches for Studying Mitochondrial and Nuclear Genome Co-evolution in Oxidative Phosphorylation

Paul Sunnucks[1]*, Hernán E. Morales[1,2], Annika M. Lamb[1], Alexandra Pavlova[1] and Chris Greening[1]

[1] School of Biological Sciences, Monash University, Clayton, VIC, Australia, [2] Department of Marine Sciences, University of Gothenburg, Gothenburg, Sweden

In animals, interactions among gene products of mitochondrial and nuclear genomes (mitonuclear interactions) are of profound fitness, evolutionary, and ecological significance. Most fundamentally, the oxidative phosphorylation (OXPHOS) complexes responsible for cellular bioenergetics are formed by the direct interactions of 13 mitochondrial-encoded and ~80 nuclear-encoded protein subunits in most animals. It is expected that organisms will develop genomic architecture that facilitates co-adaptation of these mitonuclear interactions and enhances biochemical efficiency of OXPHOS complexes. In this perspective, we present principles and approaches to understanding the co-evolution of these interactions, with a novel focus on how genomic architecture might facilitate it. We advocate that recent interdisciplinary advances assist in the consolidation of links between genotype and phenotype. For example, advances in genomics allow us to unravel signatures of selection in mitochondrial and nuclear OXPHOS genes at population-relevant scales, while newly published complete atomic-resolution structures of the OXPHOS machinery enable more robust predictions of how these genes interact epistatically and co-evolutionarily. We use three case studies to show how integrative approaches have improved the understanding of mitonuclear interactions in OXPHOS, namely those driving high-altitude adaptation in bar-headed geese, allopatric population divergence in *Tigriopus californicus* copepods, and the genome architecture of nuclear genes coding for mitochondrial functions in the eastern yellow robin.

Keywords: mitochondrial, nuclear, mitonuclear, oxidative phosphorylation, OXPHOS, co-evolution, genome architecture

OPEN ACCESS

Edited by:
Tariq Ezaz,
University of Canberra, Australia

Reviewed by:
Dan Mishmar,
Ben-Gurion University of the Negev, Israel
Alan Brelsford,
University of California, Riverside, USA

****Correspondence:***
Paul Sunnucks
paul.sunnucks@monash.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 January 2017*
Accepted: *17 February 2017*
Published: *03 March 2017*

Citation:
Sunnucks P, Morales HE, Lamb AM, Pavlova A and Greening C (2017) Integrative Approaches for Studying Mitochondrial and Nuclear Genome Co-evolution in Oxidative Phosphorylation. Front. Genet. 8:25. doi: 10.3389/fgene.2017.00025

INTRODUCTION

Rapid improvements in genomics hold much promise in advancing one of the most important but demanding tasks in evolutionary biology: establishing genotype-to-phenotype links for features of organisms that are important in adaptation and speciation (Savolainen et al., 2013; Seehausen et al., 2014). The main challenge is that fitness-conferring characteristics in complex organisms are typically quantitative traits, controlled by many loci with small individual effect sizes

(Mackay et al., 2009). This is compounded by the astronomical numbers of both meaningful gene interactions and spurious correlations that arise from population structure and history. Accordingly, adaptation can be implicated in species evolution only when disentangled from population history (Hoban et al., 2016). Identifying genotype-to-phenotype links of complex traits can be made more tractable by focussing on genomic variation expected to bestow major fitness differences based on prior knowledge. If such predictions are consistent with population genomic analyses, this will increase confidence that the candidate genes and mechanisms are true positives worthy of the demanding empirical investigations in wild populations needed to test them (Cheviron et al., 2014; Gompert et al., 2014; Egan et al., 2015).

An excellent opportunity to study the interplay between population biology and genome architecture is presented by interactions between mitochondrial proteins encoded by genes of the mitochondrial and nuclear genomes. Such mitonuclear interactions are required for fundamental physiological processes such as cellular respiration (Bar-Yaacov et al., 2012) and thus influence processes at multiple levels of biological organization: cellular function, organismal fitness, and ecosystem processes (Dowling et al., 2008; Wolff et al., 2014; Latorre-Pellicer et al., 2016). Moreover, these interactions are so central to evolutionary and ecological processes, including adaptation and speciation, that the term 'mitonuclear ecology' was recently proposed for their study (Hill, 2015, 2016).

While we have an incomplete understanding of most mitonuclear interactions (Pagliarini et al., 2008), we have a rich knowledge of a fundamental subset of them: those that form the core protein complexes responsible for oxidative phosphorylation (OXPHOS) (Rand et al., 2004; Gershoni et al., 2009). This essential system is responsible for the availability of nearly all cellular energy in eukaryotes, and thus through metabolic, trophic and thermal biology, at some level underpins virtually all eukaryotic ecological and evolutionary phenomena (Rand et al., 2004). These interactions are amenable to experimental investigation through interdisciplinary approaches. Essential for fitness, tractably complex, and relatively well-understood, these interactions thus represent strong study systems for understanding the evolution of adaptive traits. In this article, we present principles and case studies of investigations of the mitonuclear co-evolution of OXPHOS complexes in wildlife. We suggest an integrated experimental approach to this key issue in evolutionary biology, including a novel perspective on the role of genomic architecture in optimizing mitonuclear interactions.

OXIDATIVE PHOSPHORYLATION AS AN EVOLUTIONARY STUDY SYSTEM

OXPHOS Depends on Intimate Mitonuclear Interactions

Oxidative phosphorylation depends on the interaction of protein complexes in the inner mitochondrial membrane. In most animals, the four core complexes mediating OXPHOS are encoded by the 13 protein-encoding mitochondrial genes and an estimated 80 nuclear-encoded genes (Nicholls and Ferguson, 2013). The respirasome (comprising complexes I, III, and IV) uses the energy liberated during electron transfer from NADH to O_2 to drive proton-translocation to the intermembrane space and thus establish a proton gradient across the inner mitochondrial membrane (**Figure 1**) (Gu et al., 2016; Letts et al., 2016; Wu et al., 2016). The ATP synthase (complex V) uses the proton-motive force thus generated to chemiosmotically drive ATP synthesis (Allegretti et al., 2015; Hahn et al., 2016). Nuclear-encoded proteins also serve as electron carriers (e.g., cytochrome *c*), alternative electron inputs (e.g., complex II), and assembly factors (e.g., NDUFC1, SURF1) throughout the OXPHOS system (Mashkevich et al., 1997; Nicholls and Ferguson, 2013; Stroud et al., 2016).

Intimate associations between mitochondrial- and nuclear-encoded subunits are required for the electron transport chain of the respirasome and ATP synthase activity for efficient mitochondrial 'coupling,' i.e., the ratio of ATP synthesis per unit substrate and O_2 consumed (Lowell and Spiegelman, 2000). The intimacy of such interactions is particularly reflected by OXPHOS complex I. This enzyme couples electron transfer through its nuclear-encoded hydrophilic arm to proton translocation through its four mitochondrially encoded proton pumps. This depends on long-range conformational changes mediated through protein–protein interactions of the mitochondrially encoded subunits with core and supernumerary nuclear-encoded subunits (Fiedorczuk et al., 2016; Zhu et al., 2016). Genetic studies have demonstrated that assembly of this complex depends strictly on 39 of its 45 subunits (Stroud et al., 2016), and even single amino acid substitutions can alter coupling efficiency (Mimaki et al., 2012; Gershoni et al., 2014). In addition to serving as the primary electron input in the respiratory chain (Nicholls and Ferguson, 2013), OXPHOS complex I is also the main site of cellular reactive oxygen species (ROS) production (Murphy, 2009).

The central physiological importance of OXPHOS means that mitonuclear compatibilities are required for optimal fitness. Even minor biochemical inefficiencies can have major fitness consequences for an organism by modulating their energetic efficiency and oxidative stress levels. There are therefore strong selective pressures to maintain optimal mitonuclear interactions in the OXPHOS system (Rand et al., 2004; Dowling et al., 2008; Burton et al., 2013).

Mitonuclear Interactions Are Linked to Thermal and Redox Adaptation

There are multiple lines of evidence that OXPHOS is important for local adaptation. Experimental approaches with model organisms have allowed researchers to test an impressive array of mitonuclear combinations and assess their functional effects under a wide set of conditions (Dowling et al., 2007; Arnqvist et al., 2010; Paliwal et al., 2014; Ma et al., 2016; Mossman et al., 2016). These efforts have revealed that mismatched mitonuclear interactions (gene–gene interactions) can have profound consequences, such as reduced metabolic performance, fecundity, and lifespan. When mitonuclear combinations are

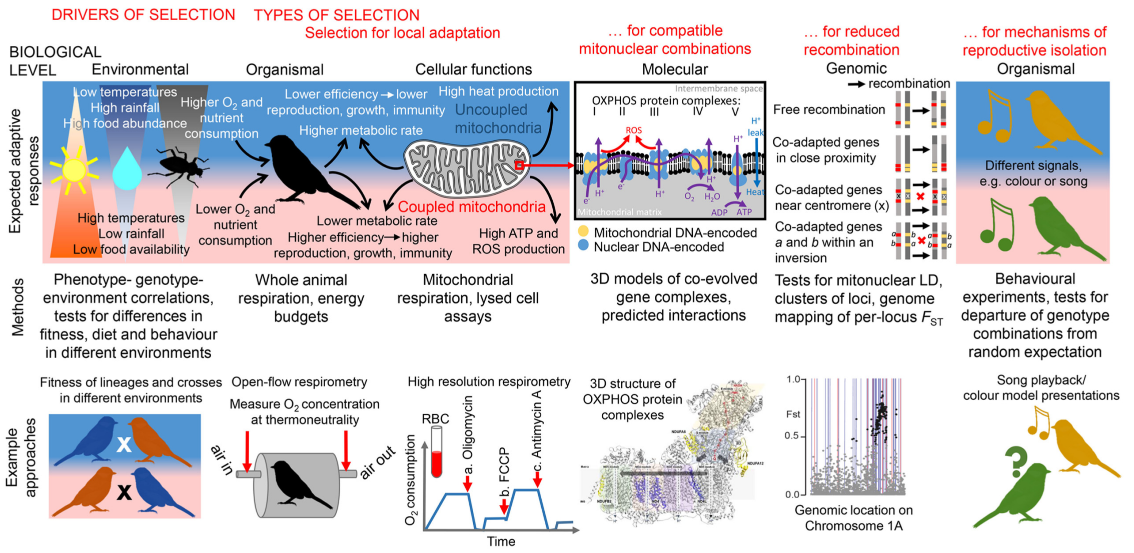

DRIVERS OF SELECTION
TYPES OF SELECTION
Selection for local adaptation
… for compatible mitonuclear combinations
… for reduced recombination
… for mechanisms of reproductive isolation
BIOLOGICAL LEVEL
Environmental
Organismal
Cellular functions
Molecular
Genomic
Organismal
Expected adaptive responses
Low temperatures
High rainfall
High food abundance
Higher O_2 and nutrient consumption
Lower efficiency → lower reproduction, growth, immunity
Higher metabolic rate
High heat production
Uncoupled mitochondria
High temperatures
Low rainfall
Low food availability
Lower O_2 and nutrient consumption
Lower metabolic rate
Higher efficiency → higher reproduction, growth, immunity
Coupled mitochondria
High ATP and ROS production
Intermembrane space
OXPHOS protein complexes:
I
II
III
IV
V
ROS
H^+ leak
H^+
e^-
O_2
H_2O
Heat
ADP
ATP
Mitochondrial matrix
● Mitochondrial DNA-encoded
● Nuclear DNA-encoded
→ recombination
Free recombination
Co-adapted genes in close proximity
Co-adapted genes near centromere (x)
Co-adapted genes *a* and *b* within an inversion
a
b
x
Different signals, e.g. colour or song
Methods
Phenotype- genotype-environment correlations, tests for differences in fitness, diet and behaviour in different environments
Whole animal respiration, energy budgets
Mitochondrial respiration, lysed cell assays
3D models of co-evolved gene complexes, predicted interactions
Tests for mitonuclear LD, clusters of loci, genome mapping of per-locus F_{ST}
Behavioural experiments, tests for departure of genotype combinations from random expectation
Example approaches
Fitness of lineages and crosses in different environments
x
x
Open-flow respirometry
Measure O_2 concentration at thermoneutrality
air in
air out
High resolution respirometry
RBC
a. Oligomycin
b. FCCP
c. Antimycin A
O_2 consumption
Time
3D structure of OXPHOS protein complexes
1.0
Fst
0.5
0.0
Genomic location on Chromosome 1A
Song playback/ colour model presentations
?

FIGURE 1 | Overview of expected responses to selective pressures related to thermal metabolism at various levels of biological organization and integrative approaches to studying the mitonuclear interactions they modulate. Environmental level: Significant differences in temperature and precipitation can drive differences in food abundance and selection for local adaptation. Significant correlations between phenotype, genotype, and environment, after controlling for confounding factors (e.g., genetic drift), can suggest the presence of local adaptation. Fitness/metabolic performance of organisms with diverged mitolineages measured in different environments can indicate the presence of local metabolic adaptation, whereas fitness/metabolic performance of several generations of crosses can show whether mitonuclear incompatibilities have evolved between lineages. **Organismal level:** Heat produced from less-coupled mitochondria may be adaptive in colder environments for endothermic organisms (Pörtner et al., 1998); individuals with less-coupled metabolism are expected to produce fewer ATP molecules (leading to lower amount of energy available for growth, immune function, or reproduction) and fewer ROS (leading to lower oxidative stress and greater longevity) per unit of O_2/nutrients consumed (Stier et al., 2014a,b). Higher O_2/nutrient consumption could be expected to compensate for metabolic inefficiency. O_2 consumption at thermoneutrality can be measured with an open-flow respirometry system (Lighton, 2008) after some acclimation time in a metabolic chamber; this can be used to calculate an organism's resting metabolism, expected to be lower in organisms adapted to warmer environments (White et al., 2007). **Cellular function level:** The level of mitochondrial coupling between substrate oxidation and ATP production determines the amount of ATP and heat (through proton leak) produced per unit of O_2/substrate consumed. A low level of coupling resulting in high heat production might be adaptive in cold climates. Mitochondrial respiration in birds, fish, amphibians and reptiles can be measured non-destructively from red blood cells (RBC) instead of liver or muscle tissues; O_2 consumption (blue line) can be measured at a baseline for comparison with responses to the additions of (a) the ATP synthase inhibitor oligomycin (measures residual O_2 consumption during proton leakage), (b) the mitochondrial uncoupler FCCP (measures maximal uncoupled O_2 consumption), and (c) the inhibitor of mitochondrial respiration antimycin A (measures non-mitochondrial oxygen consumption) (Stier et al., 2016). **Molecular level**: Complex I (NADH dehydrogenase), complex II (succinate dehydrogenase), complex III (*bc*$_1$ complex) and complex IV (cytochrome *c* oxidase) transport electrons (e^-) from NADH, succinate and FAD-linked substrates (not shown) to create a proton-motive force (H^+ gradient). Complex V (ATP synthase) uses energy released by backflow of protons to create ATP from ADP. Proteins of complexes I, III, IV, and V are encoded by both nuclear and mitochondrial genomes, leading to strong selection for compatible functional allele combinations. Mapping genes with amino acid candidates for positive selection onto 3D models of OXPHOS complexes enables better understanding of mitonuclear interactions. Here, the 3D structure of OXPHOS complex I (Fiedorczuk et al., 2016) is shown, with mitochondrially encoded subunits ND4 and ND4L, found to contain positively selected amino acids in two eastern yellow robin lineages (Morales et al., 2016a), highlighted in purple. **Genomic level:** Selection for co-transmission of co-adapted nuclear-encoded mitochondrial allele combinations with mitochondrial DNA lineages can drive the evolution of genomic architecture that suppresses recombination between co-adapted genes; examples of such mechanisms include close proximity of co-adapted genes and location of co-adapted genes near a centromere or within an inversion. Mapping single-nuclear-locus F_{ST} between populations fixed for alternative mitochondrial lineages to a reference genome can help detect clusters of loci co-inherited with mtDNA. Here, F_{ST}s between two eastern yellow robin lineages mapped to chromosome 1A (dots; black dots- top 1% outliers) show the presence of the mtDNA-linked cluster of loci (Morales et al., 2016a); background lines show the location of genes with predicted mitochondrial functions (red lines- OXPHOS genes), for which this genomic region was enriched. **Organismal level (reproductive isolation and incompatibilities):** Selection against incompatible mitonuclear combinations [postzygotic reproductive isolation can drive evolution of prezygotic reproductive isolation and result in speciation (Sloan et al., 2016)]. For example, organisms can advertize their mitonuclear genotypes through differences in color or vocalization (Hill and Johnson, 2013; Hill, 2016). Behavioral experiments involving model presentations can elucidate whether individuals mate assortatively according to their mitonuclear genotype, implying late stages of speciation.

assessed in multiple environments (e.g., diet, temperature or hypoxia), interaction effects are commonly context dependent (gene–gene-environment interactions; Koevoets et al., 2012; Hoekstra et al., 2013).

In the wild, OXPHOS traits have been correlated with a wide range of environmental pressures, including heat stress (Morales et al., 2016a,b), cold stress (Cheviron et al., 2014; Stier et al., 2014a,b), nutrient limitation (da Fonseca et al., 2008), and hypoxia (da Fonseca et al., 2008; Scott et al., 2011). Consistently, there is evidence for positive selection and climate-linked differences in the sequences and expression of OXPHOS genes in a range of animal species with wide biogeographic ranges (Mishmar et al., 2003; Ruiz-Pesini et al., 2004; Toews and Brelsford, 2012; Garvin et al., 2015; Morales et al., 2016a). Accordingly, genes encoding the OXPHOS machinery are frequently candidates for positive selection. This likely reflects the high levels of mitochondrial DNA variation within and among populations combined with the selection pressures for optimally adapted phenotypes (Gershoni et al., 2009; Bar-Yaacov et al., 2012).

As mechanistic understanding of OXPHOS activity continues to improve, it should be increasingly possible to make more specific predictions of what kinds of protein-level changes should be adaptive. For example, it is hypothesized that the coupling efficiencies of OXPHOS complexes are closely linked to adaptive thermal biology. OXPHOS generates chemical energy and heat in proportions that depend on the coupling efficiency of the respirasome and ATP synthase (Lowell and Spiegelman, 2000). It is proposed that the heat produced from less-coupled mitochondria may be particularly beneficial for adaptation of endothermic organisms to colder environments. In contrast, heat production may be deleterious in warm environments, necessitate higher nutritional intake, and is associated with high oxidative stress due to increased ROS production (**Figure 1**) (Pörtner et al., 1998; Brand, 2000; Somero, 2002; Fangue et al., 2009; Stier et al., 2014a,b). It is important to account for variation in such predictions among organisms and environments. For example, contrary to the expectations for endotherms, cold adaptation in fishes is linked to higher mitochondrial densities in muscle (White et al., 2011), and so the associated high energy demands for synthesis and maintenance of mitochondria may favor genotypes with high coupling efficiency (Pavlova et al., 2017).

Nuclear Genome Architecture May Facilitate Co-evolution of Mitochondrial-Encoded and Nuclear-Encoded Mitochondrial Genes

Mitonuclear co-evolution should be enforced under strong selection given the complex interactions and essential functions mediated by OXPHOS (Burton et al., 2013). Challenges to positive co-evolution include the fast mutation rate of the mitochondrial genome (due to its proximity to ROS production, high rate of replication and lack of efficient repair mechanisms), typically maternal inheritance, and lack of recombination, which generate a mutation load that the nuclear genome must counter by compensatory mutation (Rand et al., 2004; Lynch et al., 2006; Osada and Akashi, 2012; Havird et al., 2015; Havird and Sloan, 2016). In addition, mitonuclear co-evolution can be disrupted by mechanisms that generate genetic variation or promote gene flow. In particular, substantial gene flow and recombination in nuclear DNA will tend to break up optimally functioning allele combinations of co-adapted genes in each sexual generation (Rand et al., 2004; Burton and Barreto, 2012; Burton et al., 2013). Accordingly, we propose that nuclear genomic architecture should tend to evolve to suppress recombination and prevent the segregation of genome regions that mediate epistatic functions of nuclear-encoded mitochondrial genes.

With improvements in techniques to explore genome structure, examples are building of how genomic architecture can drive evolutionary adaptation, for example the 'supergene' of 125 genes associated with differences in male mating strategies in birds (Küpper et al., 2016). Natural selection can locate co-adapted loci in genome areas of low recombination (Schwander et al., 2014; Thompson and Jiggins, 2014) or promote genomic clustering of synergistically adaptive alleles (Yeaman, 2013) so that they can be co-inherited and/or co-regulated. The three-dimensional organization of the genome can dictate how and which loci should be subject to genome changes that will favor their co-location (Lanctôt et al., 2007; Wijchers and de Laat, 2011). Reduced recombination among co-adapted genes (increasing their co-inheritance) can occur through the evolution of recombination modifiers or chromosomal re-arrangements, such as transposition of a gene to a location close to a co-adapted gene, movement of co-adapted genes toward a centromere or into a region within an inversion between diverged lineages (**Figure 1**) (Rieseberg, 2001; Butlin, 2005; Kirkpatrick and Barton, 2006; Yeaman and Whitlock, 2011; Yeaman, 2013; Ortiz-Barrientos et al., 2016).

To date, mitonuclear genomic architecture (encompassing mitochondrial-nuclear and nuclear-nuclear) remains relatively underexplored, except in the context of biased co-transmission of mitochondrial and nuclear-encoded mitochondrial genes on sex chromosomes. Because mtDNA, due to its maternal inheritance, accumulates mutations that are deleterious in males (mother's curse), selection to restore fitness in males that drives compensatory evolution of nuclear-encoded mitochondrial genes could be expected to prevent concentrations of such genes on female-linked chromosomes (Havird and Sloan, 2016). Results for different taxa variously support overrepresentation, underrepresentation, or unbiased distribution of nuclear-encoded mitochondrial genes on X and Z chromosomes with respect to autosomes, supporting multiple theories of the distribution of nuclear-encoded mitochondrial genes: co-adaptation, sexual conflict and sexual selection (Drown et al., 2012; Hill and Johnson, 2013; Dean et al., 2014, 2015; Hill, 2014; Rogell et al., 2014). More recently, we uncovered an autosomal genomic island of divergence associated with mitonuclear interactions in a passerine (**Figure 1**) (Morales et al., 2016a). This genomic island of divergence is implicated in maintaining deep mitochondrial divergence between two parapatric lineages connected by nuclear gene flow. Observations of the fluidity of positioning of nuclear-encoded mitochondrial genes in some systems (but see Dean et al., 2015) raise questions about the role of genome organization in rates of co-evolution between mitochondria and nuclear genes (Hill, 2014). We contend that exploration of genomic architecture may be crucial for understanding mitonuclear co-evolution, and also vice versa given the crucial roles and considerable number of genes concerned with mitochondrial function (Pagliarini et al., 2008).

INTEGRATIVE APPROACHES FOR STUDYING MITONUCLEAR CO-EVOLUTION

While signals of selection have frequently been identified in OXPHOS-encoded genes, few studies have examined changes in function due to observed substitutions; thus empirical demonstration of local adaptation is limited (Burton et al., 2013; Levin et al., 2014). However, it is possible to develop relatively strong genotype-to-phenotype links by bridging population genetic studies with biochemical and physiological approaches developed for studying OXPHOS. In the five decades since its discovery by Mitchell (1961), a wealth of physiological, biochemical, and structural studies have developed a rich understanding of oxidative phosphorylation (Nicholls and Ferguson, 2013), and much of the knowledge and methodology can be translated to wild populations.

Here we suggest a flexible framework that draws on recent technical advances in multiple fields for testing the significance of mitonuclear interactions. First, candidate interacting loci can be identified by improved methods for inferring loci under selection. Second, these candidates can then be examined through structural mapping and modeling to develop hypotheses about biochemical interactions relevant to the species biology in question. Third, these hypotheses can then be tested by measuring phenotypic responses at different scales, notably whole cell, whole animal, and fitness in the wild. Finally, experimental approaches could be used to test for reproductive isolation between differently adapted lineages (**Figure 1**). It is particularly desirable to compare multiple species for repeated signals of selection in the same genomic regions: common signals of selection between lineages in the context of their geographic arrangement relative to selection pressures provide strong evidence of adaptation (Garvin et al., 2014). Adopting this proposed framework should increase comparability among studies.

Genetic Approaches for Detecting Natural Selection

Detection of natural selection is one of the most contentious and active fields in evolutionary biology. Here we highlight, in the context of mitonuclear co-evolution, the more general issues that are explored in depth elsewhere (Nielsen, 2005; Haasl and Payseur, 2016; Manel et al., 2016; Stephan, 2016).

The power to detect candidate loci that evolve under natural selection rests on the molecular tools available for a given system: reduced representation genomic scans (SNPs), sequence-based genomic scans (candidate genes, exome-sequencing, or RNA-sequencing), whole genome re-sequencing, and/or physical linkage maps (Manel et al., 2016; Stephan, 2016). The key limitation to detecting natural selection in the wild is that several ecological and evolutionary processes can leave a similar signature to selection and lead to a high rate of false positives. Confounding factors include demographic processes (e.g., population size change and structure), background selection, and heterogeneous mutation and recombination rates. Given that mitochondrial and nuclear genomes can have largely independent evolutionary histories (e.g., different introgression patterns and mutation load), knowledge of the demographic history of the study system is especially useful to interpret patterns of mitonuclear co-evolution (Bar-Yaacov et al., 2015; Morales et al., 2016a,b; Pereira et al., 2016; Sloan et al., 2016). Given that mitonuclear co-evolution is likely to respond to environmental variation (Burton et al., 2013), approaches to detecting selection that rely on gene-environment associations could be particularly useful to identify candidate loci under selection (Rellstab et al., 2015; Forester et al., 2016).

A common starting point in the search for signatures of selection in mitonuclear co-evolution is sequencing full mitochondrial genomes. A family of methods proven to be especially useful in the context of mitogenome evolution are codon-based approaches, which rely on the estimation of the non-synonymous to synonymous ratio ($\omega = d_N/d_S$) [HyPhy and Datamonkey (Pond and Frost, 2005); PAML (Yang, 2007)]. There are multiple examples in the literature of how complementary codon-based approaches have been combined to discriminate positive and relaxed purifying selection in mitogenome-encoded OXPHOS components (da Fonseca et al., 2008; Garvin et al., 2011; Morales et al., 2015; Wertheim et al., 2015; Pavlova et al., 2017). However, these types of methods have important limitations: they require data across multiple species, or sequences that are reasonably diverged, and only allow selection inference within coding regions (Gonçalves da Silva, 2017). It is important to consider this last limitation since mitonuclear incompatibilities have been mapped to non-coding regulatory genes, non-coding sequences such as transfer RNAs and the mitochondrial control region (Montooth et al., 2009; Meiklejohn et al., 2013; Rollins et al., 2016; Jhuang et al., 2017).

A natural follow-up is to look for evidence of natural selection in nuclear-encoded mitochondrial genes and signals of mitonuclear co-evolutionary adaptation (Mishmar et al., 2006; Gagnaire et al., 2012; Bar-Yaacov et al., 2015; Pereira et al., 2016). For example, such approaches have identified supernumerary and assembly factors of OXPHOS complex I implicated in local adaptation (Garvin et al., 2016; Morales et al., 2016a).

Protein Mapping and Modeling Enable Development of Mechanistic Hypotheses

Recent advances in understanding structure-function relationships in oxidative phosphorylation enable better prediction of how genetic substitutions affect mitochondrial function. Largely as a result of recent advances in cryo-electron microscopy, complete atomic-resolution structures of all components in the mammalian electron transport chain are now available, including the mitochondrially co-encoded complex I (**Figure 1**) (Fiedorczuk et al., 2016; Zhu et al., 2016), complex III (Iwata et al., 1998), complex IV (Tsukihara et al., 1996), and the respirasome supercomplex (Gu et al., 2016; Letts et al., 2016; Wu et al., 2016). In addition, near-complete structures of yeast ATP synthase are also available (Allegretti et al., 2015; Hahn et al., 2016).

With these newly available protein structures, it is now possible to map the locations of subunits and amino acids predicted to be under selection using protein visualization software (Pettersen et al., 2004) and to develop homology models using public servers (Källberg et al., 2012; Kelley et al., 2015). Such approaches have been used to predict the mechanistic effects of amino acid substitutions observed in

OXPHOS subunits across diverse species (Scott et al., 2011; Finch et al., 2014; Caballero et al., 2015; Zhang and Broughton, 2015; Campana et al., 2016; Morales et al., 2016a). This approach was first highlighted by Garvin et al. (2011) who detected an amino acid under positive selection in the Pacific salmon in an unusual region of OXPHOS complex I: the piston-like horizontal helix (helix HL) of ND5. Meta-analysis has since suggested that this helix is the most common region of positive selection in the mitogenomes of diverse animal species (Garvin et al., 2014). While the function of the helix HL remains unresolved, it is hypothesized to influence coupling by propagating conformational changes from proximal to distal proton pumps; hence fine-tuning its properties may have adaptive consequences for heat and energy production (Torres-Bacete et al., 2011; Sazanov, 2014).

Bridging Gaps through Mitochondrial, Cellular, and Organismal Physiology

In animal systems, it is challenging to validate experimentally that certain amino acid substitutions affect mitochondrial function. Due to their membrane localisation, multi-subunit cofactor-bound composition, and complex assembly pathways, OXPHOS complexes are incompatible with recombinant protein expression and can rarely be purified natively (Nicholls and Ferguson, 2013). However, a suite of physiological techniques enable us to measure the activities, kinetics, and efficiencies of OXPHOS complexes. For example, classical respirometry techniques enable measurement of the rates of substrate oxidation or oxygen consumption by whole cells or purified mitochondria; it is possible to calculate mitochondrial coupling efficiencies and to probe the activities of specific protein complexes by systematically comparing basal respiration rates with those in the presence of specific agonists, inhibitors, and uncouplers (**Figure 1**) (Nicholls and Ferguson, 2013; Stier et al., 2013, 2016; Toews et al., 2014). Well-established assays using lysed cells also enable measurement of potentially relevant parameters such as the expression levels, protein content, and kinetic parameters of individual OXPHOS complexes (Scott et al., 2011).

Another development that enhances the ability to measure biologically relevant mitochondrial function is the recent discovery that non-mammalian animals harbor functional mitochondria in their erythrocytes. This presents options for non-destructive sampling of wild populations (Stier et al., 2013, 2015, 2016). These cellular measurements of mitochondrial respiration can be complemented with whole-organism measurements of basal and maximal metabolic rates; while rarely adopted in mitonuclear ecology, such approaches may have value for understanding relationships between nutritional intake and energy expenditure (White et al., 2007, 2011; Halsey and White, 2010).

Taking an approach amenable to experimental manipulations, laboratory-based crossings have also been used to assess the effects of intraspecific and interspecific mitonuclear compatibilities using individuals sampled from wild populations from different environments. As elaborated in a case study below, there are several examples of how crosses have been combined with measurements of enzymatic, cellular, or organismal performance to consolidate genotype-to-phenotype links (Edmands and Burton, 1999; Willett and Burton, 2001; Arnqvist et al., 2010; Chang et al., 2015; Dordevic et al., 2016). In the few study systems where cell lines can be established, experimental cellular approaches can provide valuable functional insights in study systems (Blier et al., 2006), for example through the construction of mitonuclear hybrid cell systems (cybrids) that allow testing of the effects of mitogenome variation on fitness in a constant nuclear background (e.g., Barrientos et al., 1998; Moreno-Loshuertos et al., 2006; Dingley et al., 2014). Within these kinds of manipulative approaches, as well as more broadly, rapid advances in molecular genomics, including the increasing tractability of RNA sequencing, are facilitating investigations of the roles, mechanisms and evolutionary genomics of genes of interest in adaptation and divergence of wild species (Harrisson et al., 2014; Havird and Sloan, 2016; Latorre-Pellicer et al., 2016).

Gaining Insights into the Genomic Architecture of Mitonuclear Co-adaptation

The level of resolution that can be reached in examining genomic architecture depends on the genomic resources available. Full resolution requires assembled genomes and physical linkage maps, rarely available for wild organisms. However, powerful approximations can be made by mapping genetic variants of interest (e.g., candidate genes or loci under selection) on to a reference annotated genome of the same species or a close relative with known conserved synteny. Mapped variation provides the presumed order, position and identity of loci of interest (e.g., nuclear-encoded mitochondrial genes).

Linkage (gametic) disequilibrium (LD) is the non-random association of alleles at different loci within individuals. These correlations can arise through genes being near each other on a chromosome, via population subdivision, and driven by epistatic selection. Accordingly, LD can arise between markers on different nuclear chromosomes, and between the mitochondrial and nuclear genomes (Sloan et al., 2015). Linkage disequilibrium has proven a powerful tool for studying the genomic architecture of population divergence, local adaptation, and reproductive isolation (Nosil et al., 2009; Servedio, 2009; Smadja and Butlin, 2011). Natural selection can favor the evolution of high LD when multiple loci that influence a trait experience the same divergent selection (Nielsen, 2005). Strong LD between nuclear-encoded mitochondrial alleles could signal co-adapted genes responding to the same selective drivers, which may or may not be maintained by genomic architecture favoring reduced levels of recombination. Significant clustering of nuclear genes encoding mitochondrial or chloroplast proteins in *Arabidopsis* has been demonstrated (Alexeyenko et al., 2006). Similar analysis for animal taxa has rarely been performed, not least because fully assembled genomes are unavailable for many of the organisms for which mitonuclear co-evolution might be relevant, but some significant mitochondrial-nuclear LD is present in humans (Sloan et al., 2015). As genomic resources for non-model organisms expand, we expect to see more studies of mitonuclear genomic architecture (e.g., genomic re-arrangements of nuclear-encoded mitochondrial genes), as is increasingly the case for other traits linked to reproductive isolation (Lowry and Willis, 2010; Jones et al., 2012; Egan et al., 2015).

Even without a reference genome, LD can be estimated by analyzing LD clustering with LDna (Kemppainen et al., 2015). This tool finds clusters of loci with similarly high levels of LD independently of their position in the genome. This is a valuable first step to studying the genomic architecture of organisms without genomic resources. It is important to note that high LD can also emerge through processes unrelated to selection, such as population history and structure, which should be accounted for (Mangin et al., 2012; Goicoechea et al., 2015). Demographic factors, however, should impact many loci across the genome, so significant excesses of high LD among nuclear genes with mitochondrial functions is indicative of non-neutral processes (Morales et al., 2016a). Comparative genomic approaches are also recommended to investigate whether re-arranged or ancestral genomic architectures in closely related taxa are more or less prone to evolution of mitonuclear interactions, and whether nuclear-encoded mitochondrial gene re-arrangements can be favored to reduce recombination between locally co-adapted alleles and can promote adaptive genetic divergence (Faria et al., 2011; Yeaman, 2013).

CASE STUDIES HIGHLIGHTING INTEGRATIVE APPROACHES

The strength of oxidative phosphorylation as an evolutionary study system has been highlighted by several in-depth studies on specific organisms. Each of these studies integrated different combinations of the techniques described above to address different questions about the role of OXPHOS in evolutionary and ecological processes.

Mitochondrial Selection in High-Altitude Adaptation in Bar-Headed Geese

One of the richest examples of the role of mitochondrial evolution in local adaptation comes from studies of the bar-headed goose *Anser indicus*. During its well-documented migrations over the Himalayas, this bird sustains high metabolic rates as a result of a multitude of physiological adaptations (Bishop et al., 2015). Comparative physiological studies show that adaptations at multiple levels of organization, from protein activity to organ morphology, enable this species to enhance O_2 supply and modulate O_2 usage compared to low-altitude geese species (Scott et al., 2009, 2011).

Among the potentially adaptive differences observed is a difference in the substrate kinetics of OXPHOS complex IV in bar-headed geese compared to other species. Cardiac muscle measurements show that this enzyme has a comparatively high affinity but low activity for its nuclear-encoded substrate cytochrome *c*. Scott et al. (2011) proposed that this may be adaptively relevant by enabling the mitochondrion to maintain redox balance in response to limitations and fluctuations in their O_2 supply during their extreme flights. The authors inferred the genotypic basis of these changes by comparing the sequences of the mtDNA and nuclear-encoded mitochondrial genes from complex IV subunits between low- and high-altitude species. This revealed several non-synonymous substitutions in bar-headed geese, including a unique W116R substitution in COX3, as well as subtle differences in the expression levels of the complex (Scott et al., 2011).

Despite these strong phenotype-to-genotype links, it remains to be determined how these substitutions affect the biochemistry of the complex. Homology modeling suggests that the W116R mutation disrupts inter-subunit interactions in complex IV, but it is unclear if and how this affects the binding of cytochrome *c* (Scott et al., 2011). These gaps reflect the major challenges associated with purifying this enzyme complex for kinetic or structural characterisation. A striking contrast is provided by studies on why the O_2-affinity of hemoglobin is so high in bar-headed geese; the relative ease of purifying this protein from erythrocytes has facilitated structural studies showing that a single amino acid substitution markedly shifts the cooperative behavior of the hemoglobin tetramer and in turn modulates O_2 affinity (Zhang et al., 1996; Liang et al., 2001).

These studies of the bar-headed goose offer a strong example of integrated research leading to an understanding of the mechanistic links between genes affecting mitochondrial function and ecophysiological phenotype. While there are multiple classes of gene that might be expected to contribute to the adaptive phenotype (Scott et al., 2011), there is no particular expectation that strong genome architecture is required to promote the co-evolution of these: the species does not interbreed with another, and there are no apparent differently adapted lineages within the bar-headed goose, so gene flow should not disrupt co-adapted combinations.

Mitonuclear Co-evolution in the Marine Crustacean *Tigriopus californicus*

Through extensive studies, Burton and colleagues have demonstrated that mitonuclear discordance has contributed to allopatric population divergence of *Tigriopus californicus* copepods. There is extraordinary genetic differentiation between populations of this intertidal copepod across geographic barriers in the Californian coast, with mtDNA divergence exceeding 18% between Santa Cruz and San Diego populations (Burton and Lee, 1994; Burton, 1998). Elegantly designed aquarium experiments revealed that F_1 hybrids from the *T. californicus* populations are viable, but subsequent generations exhibit a range of fitness defects and reduced ATP production rates (Burton and Lee, 1994; Ellison and Burton, 2006, 2008). Consistent with a mitochondrial origin, maternal but not paternal backcrossing can restore fitness of progeny (Ellison and Burton, 2008).

Targeted sequencing revealed that there are high levels of divergence in genes encoding key determinants of the mitochondrial electron transport chain. While most of these substitutions appear to be neutral, ω-based approaches provided strong evidence for positive selection for substitutions in mitochondrially encoded complex IV and its nuclear-encoded substrate cytochrome *c* (Rawson et al., 2000; Willett and Burton, 2004). Interpopulation crossing experiments substantiated this by showing that mitonuclear compatibility was required for optimal complex IV activity (Edmands and Burton, 1999) and that the mitotype modulated segregation ratios of cytochrome *c* (Willett and Burton, 2001). The authors went further by validating these predictions using biochemical approaches. Intrapopulation pairs of complex IV and cytochrome *c* consistently showed up to fourfold higher activity than did interpopulation pairs (Rawson and Burton, 2002). Moreover, recombinantly produced cytochrome *c* variants (reflecting different nuclear backgrounds) interacted differentially with complex IV in tissue homogenates (reflecting different mitochondrial backgrounds) (Harrison and Burton, 2006). This work proved that single substitutions are sufficient to cause mitonuclear incompatibilities in wild populations.

Despite these accomplishments, the evolutionary processes and pressures that result in allopatric population divergence remain under investigation. Pereira et al. (2016) recently approximated the contribution of genetic drift and natural selection in *T. californicus* divergence by comparing whole-transcriptome sequences of allopatric populations at different stages of divergence (Pereira et al., 2016). They found that the pattern of shared polymorphism could be partially explained by genetic drift, as lower effective population sizes led to less shared polymorphism between populations, and higher mutation load. However, natural selection possibly drives accelerated evolution of some genes, including nuclear-encoded mitochondrial ones. The authors predict that genomic architecture should regulate the efficiency of selection and the impact of drift, for example by modulating recombination rates, but this prediction was not tested.

Genomic Architecture of Mitonuclear Interactions in the Eastern Yellow Robin

Our studies on the population structure of eastern yellow robin *Eopsaltria australis* have emphasized the importance of studying genomic architecture (Morales et al., 2016a). This songbird is one of multiple animals that maintains functional mitonuclear interactions despite discordance between its mitochondrial and nuclear genomes (Toews and Brelsford, 2012). Population genetic data have shown that the major axis of nuclear DNA differentiation runs north-south through the species range in Eastern Australia, whereas mitochondrial DNA has diverged into two mitolineages in the perpendicular coastal-inland direction (Pavlova et al., 2013; Morales et al., 2016a). Coalescent analyses suggest that the two genomes initially differentiated together in a north–south direction during the early Pleistocene, but their evolutionary history became separated due to two mitochondrial introgression events in the mid-to-late Pleistocene (Pavlova et al., 2013; Morales et al., 2016b). The two mitolineages show sharp climate-correlated differences in their distributions. This suggests that the mitochondrial introgression and resulting divergence were driven by natural selection (Morales et al., 2016b).

There is evidence of positive selection for non-conservative amino acid differences in the proton pumps ND4 and ND4L of OXPHOS complex I between the mitolineages of the eastern yellow robin. These polymorphisms are predicted to cause differences in electrostatic subunit-subunit interactions and in turn influence coupling efficiencies of the complex, though this remains to be validated experimentally (Morales et al., 2015). Comparison of fixation indexes in the nuclear genomes between eastern yellow robin populations across their biogeographic range revealed the existence of two genomic islands of divergence against a background of low differentiation (Morales et al., 2016a). One of these islands, located on autosome 1A (**Figure 1**), is statistically overrepresented with nuclear-encoded genes with predicted mitochondrial functions; among them are three complex I supernumerary subunits proposed to be functionally linked to the mitochondrially encoded ND4 and ND4L genes. Moreover, markers within the genomic island of divergence exhibit very strongly elevated LD, suggesting genome architecture that promotes reduced recombination between nuclear-encoded genes with mitochondrial functions. Further research will disentangle whether mitonuclear co-evolution promoted the evolution of this particular genomic architecture or pre-existing genomic architecture enabled this tight mitonuclear co-evolution.

CONCLUSION

In this perspective, we have summarized some of the wealth of information of the adaptive consequences of mitochondrial-nuclear interactions, with particular focus on OXPHOS functions. We make the case that the powerful fitness consequences of mitonuclear gene interactions are likely to select for optimized genome architecture that will hold together effective combinations of nuclear-encoded mitochondrial gene in the face of gene flow. Demonstrating the phenotypic consequences of genome variation in wildlife species is challenging: we suggest a workflow that utilizes advances in detection of candidates of selection, biochemical understanding of OXPHOS, phenotyping and studying genome organization. The number of studies demonstrating major evolutionary impacts of mitonuclear interactions in wildlife is currently limited, but we anticipate they will be found to be common under the application of the strong emerging methods of investigation such as we present here. Comparative genomic approaches will be important for deriving the deepest insights into mitonuclear evolution and genome architecture, and accordingly, we encourage the application of consistent methodologies.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

Our work on eastern yellow robin has not yet been funded by a large grant. Partial funding was provided by the

Australian Research Council Linkage Grant (LP0776322). HM was supported by the Holsworth Wildlife Research Endowment and Stuart Leslie Bird Research Award from BirdLife Australia, a Monash Graduate Scholarship (MGS), a Monash Faculty of Science Dean's International Postgraduate Research Scholarship and a Monash Postgraduate Publication Award. CG was supported by Australian Research Council DECRA Fellowship DE170100310.

ACKNOWLEDGMENTS

We thank Scott Edwards and Tariq Ezaz for the invitation to write this perspective. We are grateful to the many researchers with whom we have discussions on topics covered here, including Antoine Stier, Ricardo Pereira, Geoff Hill, Damian Dowling, Jonci Wolff, and our research group members including Stephanie Falk and Lynna Kvistad. The comments of Dan Mishmar and Alan Brelsford helped improve the manuscript.

REFERENCES

Alexeyenko, A., Millar, A. H., Whelan, J., and Sonnhammer, E. L. L. (2006). Chromosomal clustering of nuclear genes encoding mitochondrial and chloroplast proteins in *Arabidopsis*. *Trends Genet.* 22, 589–593. doi: 10.1016/j.tig.2006.09.002

Allegretti, M., Klusch, N., Mills, D. J., Vonck, J., Kühlbrandt, W., and Davies, K. M. (2015). Horizontal membrane-intrinsic alpha-helices in the stator a-subunit of an F-type ATP synthase. *Nature* 521, 237–240. doi: 10.1038/nature14185

Arnqvist, G., Dowling, D. K., Eady, P., Gay, L., Tregenza, T., Tuda, M., et al. (2010). Genetic architecture of metabolic rate: environment specific epistasis between mitochondrial and nuclear genes in an insect. *Evolution* 64, 3354–3363. doi: 10.1111/j.1558-5646.2010.01135.x

Barrientos, A., Kenyon, L., and Moraes, C. T. (1998). Human xenomitochondrial cybrids cellular models of mitochondrial complex I deficiency. *J. Biol. Chem.* 273, 14210–14217. doi: 10.1074/jbc.273.23.14210

Bar-Yaacov, D., Blumberg, A., and Mishmar, D. (2012). Mitochondrial-nuclear co-evolution and its effects on OXPHOS activity and regulation. *Biochim. Biophys. Acta* 1819, 1107–1111. doi: 10.1016/j.bbagrm.2011.10.008

Bar-Yaacov, D., Hadjivasiliou, Z., Levin, L., Barshad, G., Zarivach, R., Bouskila, A., et al. (2015). Mitochondrial involvement in vertebrate speciation? The case of mito-nuclear genetic divergence in chameleons. *Genome Biol. Evol.* 7, 3322–3336. doi: 10.1093/gbe/evv226

Bishop, C. M., Spivey, R. J., Hawkes, L. A., Batbayar, N., Chua, B., Frappell, P. B., et al. (2015). The roller coaster flight strategy of bar-headed geese conserves energy during Himalayan migrations. *Science* 347, 250–254. doi: 10.1126/science.1258732

Blier, P. U., Breton, S., Desrosiers, V., and Lemieux, H. (2006). Functional conservatism in mitochondrial evolution: insight from hybridization of arctic and brook charrs. *J. Exp. Zool. Part B Mol. Dev. Evol.* 306, 425–432. doi: 10.1002/jez.b.21089

Brand, M. D. (2000). Uncoupling to survive? The role of mitochondrial inefficiency in ageing. *Exp. Gerontol.* 35, 811–820. doi: 10.1016/S0531-5565(00)00135-2

Burton, R. S. (1998). Intraspecific phylogeography across the Point Conception biogeographic boundary. *Evolution* 52, 734–745. doi: 10.2307/2411268

Burton, R. S., and Barreto, F. S. (2012). A disproportionate role for mtDNA in Dobzhansky–Muller incompatibilities? *Mol. Ecol.* 21, 4942–4957. doi: 10.1111/mec.12006

Burton, R. S., and Lee, B. N. (1994). Nuclear and mitochondrial gene genealogies and allozyme polymorphism across a major phylogeographic break in the copepod *Tigriopus californicus*. *Proc. Natl. Acad. Sci. U.S.A.* 91, 5197–5201. doi: 10.1073/pnas.91.11.5197

Burton, R. S., Pereira, R. J., and Barreto, F. S. (2013). Cytonuclear genomic interactions and hybrid breakdown. *Annu. Rev. Ecol. Evol. Syst.* 44, 281–302. doi: 10.1146/annurev-ecolsys-110512-135758

Butlin, R. K. (2005). Recombination and speciation. *Mol. Ecol.* 14, 2621–2635. doi: 10.1111/j.1365-294X.2005.02617.x

Caballero, S., Duchêne, S., Garavito, M. F., Slikas, B., and Baker, C. S. (2015). Initial evidence for adaptive selection on the NADH Subunit two of freshwater dolphins by analyses of mitochondrial genomes. *PLoS ONE* 10:e0123543. doi: 10.1371/journal.pone.0123543

Campana, M. G., Parker, L. D., Hawkins, M. T. R., Young, H. S., Helgen, K. M., Gunther, M. S., et al. (2016). Genome sequence, population history, and pelage genetics of the endangered African wild dog (*Lycaon pictus*). *BMC Genomics* 17, 1013. doi: 10.1186/s12864-016-3368-9

Chang, C.-C., Rodriguez, J., and Ross, J. (2015). Mitochondrial-nuclear epistasis impacts fitness and mitochondrial physiology of interpopulation *Caenorhabditis briggsae* hybrids. *G3* (Bethesda). 6, 209–219. doi: 10.1534/g3.115.022970

Cheviron, Z. A., Connaty, A. D., McClelland, G. B., and Storz, J. F. (2014). Functional genomics of adaptation to hypoxic cold-stress in high-altitude deer mice: transcriptomic plasticity and thermogenic performance. *Evolution* 68, 48–62. doi: 10.1111/evo.12257

da Fonseca, R. R., Johnson, W. E., O'Brien, S. J., Ramos, M. J., and Antunes, A. (2008). The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics* 9, 1. doi: 10.1186/1471-2164-9-119

Dean, R., Zimmer, F., and Mank, J. E. (2014). The potential role of sexual conflict and sexual selection in shaping the genomic distribution of mito-nuclear genes. *Genome Biol. Evol.* 6, 1096–1104. doi: 10.1093/gbe/evu063

Dean, R., Zimmer, F., and Mank, J. E. (2015). Deficit of mitonuclear genes on the human X chromosome predates sex chromosome formation. *Genome Biol. Evol.* 7, 636–641. doi: 10.1093/gbe/evv017

Dingley, S. D., Polyak, E., Ostrovsky, J., Srinivasan, S., Lee, I., Rosenfeld, A. B., et al. (2014). Mitochondrial DNA variant in COX1 subunit significantly alters energy metabolism of geographically divergent wild isolates in *Caenorhabditis elegans*. *J. Mol. Biol.* 426, 2199–2216. doi: 10.1016/j.jmb.2014.02.009

Dordevic, M., Stojkovic, B., Savkovic, U., Immonen, E., Tucic, N., Lazarevic, J., et al. (2016). Sex-specific mitonuclear epistasis and the evolution of mitochondrial bioenergetics, ageing, and life history in seed beetles. *Evolution* 71, 274–288. doi: 10.1111/evo.13109

Dowling, D. K., Friberg, U., Hailer, F., and Arnqvist, G. (2007). Intergenomic epistasis for fitness: within-population interactions between cytoplasmic and nuclear genes in *Drosophila melanogaster*. *Genetics* 175, 235–244. doi: 10.1534/genetics.105.052050

Dowling, D. K., Friberg, U., and Lindell, J. (2008). Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends Ecol. Evol.* 23, 546–554. doi: 10.1016/j.tree.2008.05.011

Drown, D. M., Preuss, K. M., and Wade, M. J. (2012). Evidence of a paucity of genes that interact with the mitochondrion on the X in mammals. *Genome Biol. Evol.* 4, 875–880. doi: 10.1093/gbe/evs064

Edmands, S., and Burton, R. S. (1999). Cytochrome c oxidase activity in interpopulation hybrids of a marine copepod: a test for nuclear-nuclear or nuclear-cytoplasmic coadaptation. *Evolution* 53, 1972–1978. doi: 10.2307/2640456

Egan, S. P., Ragland, G. J., Assour, L., Powell, T. H. Q., Hood, G. R., Emrich, S., et al. (2015). Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. *Ecol. Lett.* 18, 817–825. doi: 10.1111/ele.12460

Ellison, C. K., and Burton, R. S. (2006). Disruption of mitochondrial function in interpopulation hybrids of *Tigriopus californicus*. *Evolution (N. Y).* 60, 1382–1391.

Ellison, C. K., and Burton, R. S. (2008). Interpopulation hybrid breakdown maps to the mitochondrial genome. *Evolution (N. Y).* 62, 631–638. doi: 10.1111/j.1558-5646.2007.00305.x

Fangue, N. A., Richards, J. G., and Schulte, P. M. (2009). Do mitochondrial properties explain intraspecific variation in thermal tolerance? *J. Exp. Biol.* 212, 514–522. doi: 10.1242/jeb.024034

Faria, R., Neto, S., Noor, M. A. F., and Navarro, A. (2011). "Role of natural selection in chromosomal speciation," in *Encyclopedia of Life Sciences (ELS)* (Chichester: John Wiley & Sons Ltd.). doi: 10.1002/9780470015902.a0022850

Fiedorczuk, K., Letts, J. A., Degliesposti, G., Kaszuba, K., Skehel, M., and Sazanov, L. A. (2016). Atomic structure of the entire mammalian mitochondrial complex I. *Nature* 538, 406–410. doi: 10.1038/nature19794

Finch, T. M., Zhao, N., Korkin, D., Frederick, K. H., and Eggert, L. S. (2014). Evidence of positive selection in mitochondrial complexes I and V of the African elephant. *PLoS ONE* 9:e92587. doi: 10.1371/journal.pone.0092587

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., and Lasky, J. R. (2016). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol. Ecol.* 25, 104–120. doi: 10.1111/mec.13476

Gagnaire, P.-A., Normandeau, E., and Bernatchez, L. (2012). Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American eels. *Mol. Biol. Evol.* 29, 2909–2919. doi: 10.1093/molbev/mss076

Garvin, M. R., Bielawski, J. P., and Gharrett, A. J. (2011). Positive Darwinian selection in the piston that powers proton pumps in complex I of the mitochondria of Pacific salmon. *PLoS ONE* 6:e24127. doi: 10.1371/journal.pone.0024127

Garvin, M. R., Bielawski, J. P., Sazanov, L. A., and Gharrett, A. J. (2014). Review and meta-analysis of natural selection in mitochondrial complex I in metazoans. *J. Zool. Syst. Evol. Res.* 53, 1–17. doi: 10.1111/jzs.12079

Garvin, M. R., Templin, W. D., Gharrett, A. J., DeCovich, N., Kondzela, C. M., Guyon, J. R., et al. (2016). Potentially adaptive mitochondrial haplotypes as a tool to identify divergent nuclear loci. *Methods Ecol. Evol.* doi: 10.1111/2041-210X.12698

Garvin, M. R., Thorgaard, G. H., and Narum, S. R. (2015). Differential expression of genes that control respiration contribute to thermal adaptation in redband trout (*Oncorhynchus mykiss gairdneri*). *Genome Biol. Evol.* 7, 1404–1414. doi: 10.1093/gbe/evv078

Gershoni, M., Levin, L., Ovadia, O., Toiw, Y., Shani, N., Dadon, S., et al. (2014). Disrupting mitochondrial-nuclear coevolution affects OXPHOS Complex I integrity and impacts human health. *Genome Biol. Evol.* 6, 2665–2680. doi: 10.1093/gbe/evu208

Gershoni, M., Templeton, A. R., and Mishmar, D. (2009). Mitochondrial bioenergetics as a major motive force of speciation. *Bioessays* 31, 642–650. doi: 10.1002/bies.200800139

Goicoechea, P. G., Herrán, A., Durand, J., Bodénès, C., Plomion, C., and Kremer, A. (2015). A linkage disequilibrium perspective on the genetic mosaic of speciation in two hybridizing Mediterranean white oaks. *Heredity (Edinb).* 114, 373–386. doi: 10.1038/hdy.2014.113

Gompert, Z., Comeault, A. A., Farkas, T. E., Feder, J. L., Parchman, T. L., Buerkle, C. A., et al. (2014). Experimental evidence for ecological selection on genome variation in the wild. *Ecol. Lett.* 17, 369–379. doi: 10.1111/ele.12238

Gonçalves da Silva, A. G. (2017). "Measuring natural selection," in *Bioinformatics: Data, Sequence Analysis, and Evolution*, ed. J. M. Keith (New York, NY: Springer), 315–347.

Gu, J., Wu, M., Guo, R., Yan, K., Lei, J., Gao, N., et al. (2016). The architecture of the mammalian respirasome. *Nature* 537, 639–643. doi: 10.1038/nature19359

Haasl, R. J., and Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.* 25, 5–23. doi: 10.1111/mec.13339

Hahn, A., Parey, K., Bublitz, M., Mills, D. J., Zickermann, V., Vonck, J., et al. (2016). Structure of a complete ATP synthase dimer reveals the molecular basis of inner mitochondrial membrane morphology. *Mol. Cell* 63, 445–456. doi: 10.1016/j.molcel.2016.05.037

Halsey, L. G., and White, C. R. (2010). Measuring energetics and behaviour using accelerometry in cane toads Bufo marinus. *PLoS ONE* 5:e10170. doi: 10.1371/journal.pone.0010170

Harrison, J. S., and Burton, R. S. (2006). Tracing hybrid incompatibilities to single amino acid substitutions. *Mol. Biol. Evol.* 23, 559–564. doi: 10.1093/molbev/msj058

Harrisson, K. A., Pavlova, A., Telonis-Scott, M., and Sunnucks, P. (2014). Using genomics to characterize evolutionary potential for conservation of wild populations. *Evol. Appl.* 7, 1008–1025. doi: 10.1111/eva.12149

Havird, J. C., Hall, M. D., and Dowling, D. K. (2015). The evolution of sex: a new hypothesis based on mitochondrial mutational erosion. *Bioessays* 37, 951–958. doi: 10.1002/bies.201500057

Havird, J. C., and Sloan, D. B. (2016). The roles of mutation, selection, and expression in determining relative rates of evolution in mitochondrial vs. nuclear genomes. *Mol. Biol. Evol.* 33, 3042–3053. doi: 10.1093/molbev/msw185

Hill, G. E. (2014). Sex linkage of nuclear-encoded mitochondrial genes. *Heredity (Edinb).* 112, 469–470. doi: 10.1038/hdy.2013.125

Hill, G. E. (2015). Mitonuclear ecology. *Mol. Biol. Evol.* 32, 1917–1927. doi: 10.1093/molbev/msv104

Hill, G. E. (2016). Mitonuclear coevolution as the genesis of speciation and the mitochondrial DNA barcode gap. *Ecol. Evol.* 6, 5831–5842. doi: 10.1002/ece3.2338

Hill, G. E., and Johnson, J. D. (2013). The mitonuclear compatibility hypothesis of sexual selection. *Proc. Biol. Sci.* 280:20131314. doi: 10.1098/rspb.2013.1314

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., et al. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.* 188, 379–397. doi: 10.1086/688018

Hoekstra, L. A., Siddiq, M. A., and Montooth, K. L. (2013). Pleiotropic effects of a mitochondrial–nuclear incompatibility depend upon the accelerating effect of temperature in *Drosophila*. *Genetics* 195, 1129–1139. doi: 10.1534/genetics.113.154914

Iwata, S., Lee, J. W., Okada, K., Lee, J. K., Iwata, M., Rasmussen, B., et al. (1998). Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science* 281, 64–71. doi: 10.1126/science.281.5373.64

Jhuang, H., Lee, H., and Leu, J. (2017). Mitochondrial–nuclear co-evolution leads to hybrid incompatibility through pentatricopeptide repeat proteins. *EMBO Rep.* 18, 87–101. doi: 10.15252/embr.201643311

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., et al. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55–61. doi: 10.1038/nature10944

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., et al. (2012). Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* 7, 1511–1522. doi: 10.1038/nprot.2012.085

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi: 10.1038/nprot.2015.053

Kemppainen, P., Knight, C. G., Sarma, D. K., Hlaing, T., Prakash, A., Maung, M., et al. (2015). Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Mol. Ecol. Resour.* 15, 1031–1045. doi: 10.1111/1755-0998.12369

Kirkpatrick, M., and Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics* 173, 419–434. doi: 10.1534/genetics.105.047985

Koevoets, T., Van de Zande, L., and Beukeboom, L. W. (2012). Temperature stress increases hybrid incompatibilities in the parasitic wasp genus *Nasonia*. *J. Evol. Biol.* 25, 304–316. doi: 10.1111/j.1420-9101.2011.02424.x

Küpper, C., Stocks, M., Risse, J. E., dos Remedios, N., Farrell, L. L., Mcrae, S. B., et al. (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet*. 48, 79–83. doi: 10.1038/ng.3443

Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G., and Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genet.* 8, 104–115. doi: 10.1038/nrg2041

Latorre-Pellicer, A., Moreno-Loshuertos, R., Lechuga-Vieco, A. V., Sánchez-Cabo, F., Torroja, C., Acín-Pérez, R., et al. (2016). Mitochondrial and nuclear DNA matching shapes metabolism and healthy ageing. *Nature* 535, 561–565. doi: 10.1038/nature18618

Letts, J. A., Fiedorczuk, K., and Sazanov, L. A. (2016). The architecture of respiratory supercomplexes. *Nature* 537, 644–648. doi: 10.1038/nature19774

Levin, L., Blumberg, A., Barshad, G., and Mishmar, D. (2014). Mito-nuclear co-evolution: the positive and negative sides of functional ancient mutations. *Front. Genet.* 5:448. doi: 10.3389/fgene.2014.00448

Liang, Y., Hua, Z., Liang, X., Xu, Q., and Lu, G. (2001). The crystal structure of bar-headed goose hemoglobin in deoxy form: the allosteric mechanism of a hemoglobin species with high oxygen affinity. *J. Mol. Biol.* 313, 123–137. doi: 10.1006/jmbi.2001.5028

Lighton, J. R. B. (2008). *Measuring Metabolic Rates: A Manual for Scientists*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195310610.001.0001

Lowell, B. B., and Spiegelman, B. M. (2000). Towards a molecular understanding of adaptive thermogenesis. *Nature* 404, 652–660.

Lowry, D. B., and Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* 8:e1000500. doi: 10.1371/journal.pbio.1000500

Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. *Science* 311, 1727–1730. doi: 10.1126/science.1118884

Ma, H., Gutierrez, N. M., Morey, R., Van Dyken, C., Kang, E., Hayama, T., et al. (2016). Incompatibility between nuclear and mitochondrial genomes contributes to an interspecies reproductive barrier. *Cell Metab.* 24, 283–294. doi: 10.1016/j.cmet.2016.06.012

Mackay, T. F. C., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* 10, 565–577. doi: 10.1038/nrg2612

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., et al. (2016). Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Mol. Ecol.* 25, 170–184. doi: 10.1111/mec.13468

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb).* 108, 285–291. doi: 10.1038/hdy.2011.73

Mashkevich, G., Repetto, B., Glerum, D. M., Jin, C., and Tzagoloff, A. (1997). SHY1, the yeast homolog of the mammalian SURF-1 gene, encodes a mitochondrial protein required for respiration. *J. Biol. Chem.* 272, 14356–14364. doi: 10.1074/jbc.272.22.14356

Meiklejohn, C. D., Holmbeck, M. A., Siddiq, M. A., Abt, D. N., Rand, D. M., and Montooth, K. L. (2013). An incompatibility between a mitochondrial tRNA and its nuclear-encoded tRNA synthetase compromises development and fitness in *Drosophila*. *PLoS Genet.* 9:e1003238. doi: 10.1371/journal.pgen.1003238

Mimaki, M., Wang, X., McKenzie, M., Thorburn, D. R., and Ryan, M. T. (2012). Understanding mitochondrial complex I assembly in health and disease. *Biochim. Biophys. Acta* 1817, 851–862. doi: 10.1016/j.bbabio.2011.08.010

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., et al. (2003). Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. U.S.A.* 100, 171–176. doi: 10.1073/pnas.0136972100

Mishmar, D., Ruiz-Pesini, E., Mondragon-Palomino, M., Procaccio, V., Gaut, B., and Wallace, D. C. (2006). Adaptive selection of mitochondrial complex I subunits during primate radiation. *Gene* 378, 11–18. doi: 10.1016/j.gene.2006.03.015

Mitchell, P. (1961). Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature* 191, 144–148. doi: 10.1038/191144a0

Montooth, K. L., Abt, D. N., Hofmann, J. W., and Rand, D. M. (2009). Comparative genomics of *Drosophila* mtDNA: novel features of conservation and change across functional domains and lineages. *J. Mol. Evol.* 69, 94–114. doi: 10.1007/s00239-009-9255-0

Morales, H. E., Pavlova, A., Amos, N., Major, R., Bragg, J., Kilian, A., et al. (2016a). Mitochondrial-nuclear interactions maintain a deep mitochondrial split in the face of nuclear gene flow. *bioRxiv* 095596. doi: 10.1101/095596

Morales, H. E., Pavlova, A., Joseph, L., and Sunnucks, P. (2015). Positive and purifying selection in mitochondrial genomes of a bird with mitonuclear discordance. *Mol. Ecol.* 24, 2820–2837. doi: 10.1111/mec.13203

Morales, H. E., Sunnucks, P., Joseph, L., and Pavlova, A. (2016b). Perpendicular axes of incipient speciation generated by mitochondrial introgression. *bioRxiv* 072942. doi: 10.1101/072942

Moreno-Loshuertos, R., Acín-Pérez, R., Fernández-Silva, P., Movilla, N., Pérez-Martos, A., de Cordoba, S. R., et al. (2006). Differences in reactive oxygen species production explain the phenotypes associated with common mouse mitochondrial DNA variants. *Nat. Genet.* 38, 1261–1268. doi: 10.1038/ng1897

Mossman, J. A., Biancani, L. M., Zhu, C.-T., and Rand, D. M. (2016). Mitonuclear epistasis for development time and its modification by diet in *Drosophila*. *Genetics* 203, 463–484. doi: 10.1534/genetics.116.187286

Murphy, M. P. (2009). How mitochondria produce reactive oxygen species. *Biochem. J.* 417, 1–13. doi: 10.1042/BJ20081386

Nicholls, D. G., and Ferguson, S. (2013). *Bioenergetics*. Cambridge, MA: Academic Press.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet.* 39, 197–218. doi: 10.1146/annurev.genet.39.073003.112420

Nosil, P., Funk, D. J., and Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* 18, 375–402. doi: 10.1111/j.1365-294X.2008.03946.x

Ortiz-Barrientos, D., Engelstädter, J., and Rieseberg, L. H. (2016). Recombination rate evolution and the origin of species. *Trends Ecol. Evol.* 31, 226–236. doi: 10.1016/j.tree.2015.12.016

Osada, N., and Akashi, H. (2012). Mitochondrial–nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome c oxidase complex. *Mol. Biol. Evol.* 29, 337–346. doi: 10.1093/molbev/msr211

Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S.-E., et al. (2008). A mitochondrial protein compendium elucidates complex I disease biology. *Cell* 134, 112–123. doi: 10.1016/j.cell.2008.06.016

Paliwal, S., Fiumera, A. C., and Fiumera, H. L. (2014). Mitochondrial-nuclear epistasis contributes to phenotypic variation and coadaptation in natural isolates of *Saccharomyces cerevisiae*. *Genetics* 198, 1251–1265. doi: 10.1534/genetics.114.168575

Pavlova, A., Amos, J. N., Joseph, L., Loynes, K., Austin, J. J., Keogh, J. S., et al. (2013). Perched at the mito-nuclear crossroads: divergent mitochondrial lineages correlate with environment in the face of ongoing nuclear gene flow in an australian bird. *Evolution (N. Y).* 67, 3412–3428. doi: 10.1111/evo.12107

Pavlova, A., Gan, H. M., Lee, Y. P., Austin, C. M., Gilligan, D. M., Lintermans, M., et al. (2017). Purifying selection and genetic drift shaped Pleistocene evolution of the mitochondrial genome in an endangered Australian freshwater fish. *Heredity (Edinb)*. doi: 10.1038/hdy.2016.120 [Epub ahead of print].

Pereira, R. J., Barreto, F. S., Pierce, N. T., Carneiro, M., and Burton, R. S. (2016). Transcriptome-wide patterns of divergence during allopatric evolution. *Mol. Ecol.* 25, 1478–1493. doi: 10.1111/mec.13579

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Pond, S. L. K., and Frost, S. D. W. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21, 2531–2533. doi: 10.1093/bioinformatics/bti320

Pörtner, H.-O., Hardewig, I., Sartoris, F. J., and Van Dijk, P. L. M. (1998). Energetic aspects of cold adaptation: critical temperatures in metabolic, ionic and acid-base regulation. *Cold Ocean Physiol.* 66, 88–120. doi: 10.1017/CBO9780511661723.005

Rand, D. M., Haney, R. A., and Fry, A. J. (2004). Cytonuclear coevolution: the genomics of cooperation. *Trends Ecol. Evol.* 19, 645–653. doi: 10.1016/j.tree.2004.10.003

Rawson, P. D., Brazeau, D. A., and Burton, R. S. (2000). Isolation and characterization of cytochrome c from the marine copepod *Tigriopus californicus*. *Gene* 248, 15–22. doi: 10.1016/S0378-1119(00)00145-1

Rawson, P. D., and Burton, R. S. (2002). Functional coadaptation between cytochrome c and cytochrome c oxidase within allopatric populations of a marine copepod. *Proc. Natl. Acad. Sci. U.S.A.* 99, 12955–12958. doi: 10.1073/pnas.202335899

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370. doi: 10.1111/mec.13322

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16, 351–358. doi: 10.1016/S0169-5347(01)02187-5

Rogell, B., Dean, R., Lemos, B., and Dowling, D. K. (2014). Mito-nuclear interactions as drivers of gene movement on and off the X-chromosome. *BMC Genomics* 15, 1. doi: 10.1186/1471-2164-15-330

Rollins, L. A., Woolnough, A. P., Fanson, B. G., Cummins, M. L., Crowley, T. M., Wilton, A. N., et al. (2016). Selection on mitochondrial variants occurs between and within individuals in an expanding invasion. *Mol. Biol. Evol.* 33, 995–1007. doi: 10.1093/molbev/msv343

Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V., and Wallace, D. C. (2004). Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* 303, 223–226. doi: 10.1126/science.1088434

Savolainen, O., Lascoux, M., and Merilä, J. (2013). Ecological genomics of local adaptation. *Nat. Rev. Genet.* 14, 807–820. doi: 10.1038/nrg3522

Sazanov, L. A. (2014). The mechanism of coupling between electron transfer and proton translocation in respiratory complex I. *J. Bioenerg. Biomembr.* 46, 247–253. doi: 10.1007/s10863-014-9554-z

Schwander, T., Libbrecht, R., and Keller, L. (2014). Supergenes and complex phenotypes. *Curr. Biol.* 24, R288–R294. doi: 10.1016/j.cub.2014.01.056

Scott, G. R., Richards, J. G., and Milsom, W. K. (2009). Control of respiration in flight muscle from the high-altitude bar-headed goose and low-altitude birds. *Am. J. Physiol. Integr. Comp. Physiol.* 297, R1066–R1074. doi: 10.1152/ajpregu.00241.2009

Scott, G. R., Schulte, P. M., Egginton, S., Scott, A. L. M., Richards, J. G., and Milsom, W. K. (2011). Molecular evolution of cytochrome c oxidase underlies high-altitude adaptation in the bar-headed goose. *Mol. Biol. Evol.* 28, 351–363. doi: 10.1093/molbev/msq205

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., et al. (2014). Genomics and the origin of species. *Nat. Rev. Genet.* 15, 176–192. doi: 10.1038/nrg3644

Servedio, M. R. (2009). The role of linkage disequilibrium in the evolution of premating isolation. *Heredity (Edinb).* 102, 51–56. doi: 10.1038/hdy.2008.98

Sloan, D. B., Fields, P. D., and Havird, J. C. (2015). Mitonuclear linkage disequilibrium in human populations. *Proc. Biol. Sci.* 282:20151704. doi: 10.1098/rspb.2015.1704

Sloan, D. B., Havird, J. C., and Sharbrough, J. (2016). The on-again, off-again relationship between mitochondrial genomes and species boundaries. *Mol. Ecol* doi: 10.1111/mec.13959 [Epub ahead of print].

Smadja, C. M., and Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. *Mol. Ecol.* 20, 5123–5140. doi: 10.1111/j.1365-294X.2011.05350.x

Somero, G. N. (2002). Thermal physiology and vertical zonation of intertidal animals: optima, limits, and costs of living. *Integr. Comp. Biol.* 42, 780–789. doi: 10.1093/icb/42.4.780

Stephan, W. (2016). Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.* 25, 79–88. doi: 10.1111/mec.13288

Stier, A., Bize, P., Roussel, D., Schull, Q., Massemin, S., and Criscuolo, F. (2014a). Mitochondrial uncoupling as a regulator of life-history trajectories in birds: an experimental study in the zebra finch. *J. Exp. Biol.* 217, 3579–3589. doi: 10.1242/jeb.103945

Stier, A., Bize, P., Schull, Q., Zoll, J., Singh, F., Geny, B., et al. (2013). Avian erythrocytes have functional mitochondria, opening novel perspectives for birds as animal models in the study of ageing. *Front. Zool.* 10:33. doi: 10.1186/1742-9994-10-33

Stier, A., Massemin, S., and Criscuolo, F. (2014b). Chronic mitochondrial uncoupling treatment prevents acute cold-induced oxidative stress in birds. *J. Comp. Physiol. B* 184, 1021–1029. doi: 10.1007/s00360-014-0856-6

Stier, A., Reichert, S., Criscuolo, F., and Bize, P. (2015). Red blood cells open promising avenues for longitudinal studies of ageing in laboratory, non-model and wild animals. *Exp. Gerontol.* 71, 118–134. doi: 10.1016/j.exger.2015.09.001

Stier, A., Romestaing, C., Schull, Q., Lefol, E., Robin, J. P., Roussel, D., et al. (2016). How to measure mitochondrial function in birds using red blood cells: a case study in the king penguin and perspectives in ecology and evolution. *Methods Ecol. Evol.* doi: 10.1111/2041-210X.12724

Stroud, D. A., Surgenor, E. E., Formosa, L. E., Reljic, B., Frazier, A. E., Dibley, M. G., et al. (2016). Accessory subunits are integral for assembly and function of human mitochondrial complex I. *Nature* 538, 123–126. doi: 10.1038/nature19754

Thompson, M. J., and Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity (Edinb).* 113, 1–8. doi: 10.1038/hdy.2014.20

Toews, D. P. L., and Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Mol. Ecol.* 21, 3907–3930. doi: 10.1111/j.1365-294X.2012.05664.x

Toews, D. P. L., Mandic, M., Richards, J. G., and Irwin, D. E. (2014). Migration, mitochondria, and the yellow-rumped warbler. *Evolution* 68, 241–255. doi: 10.1111/evo.12260

Torres-Bacete, J., Sinha, P. K., Matsuno-Yagi, A., and Yagi, T. (2011). Structural contribution of C-terminal segments of NuoL (ND5) and NuoM (ND4) subunits of complex I from *Escherichia coli*. *J. Biol. Chem.* 286, 34007–34014. doi: 10.1074/jbc.M111.260968

Tsukihara, T., Aoyama, H., Yamashita, E., and Tomizaki, T. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 angstrom. *Science* 272, 1136. doi: 10.1126/science.272.5265.1136

Wertheim, J. O., Murrell, B., Smith, M. D., Pond, S. L. K., and Scheffler, K. (2015). RELAX: detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* 32, 820–832. doi: 10.1093/molbev/msu400

White, C. R., Alton, L. A., and Frappell, P. B. (2011). Metabolic cold adaptation in fishes occurs at the level of whole animal, mitochondria and enzyme. *Proc. Biol. Sci.* 279, 1740–1747. doi: 10.1098/rspb.2011.2060

White, C. R., Blackburn, T. M., Martin, G. R., and Butler, P. J. (2007). Basal metabolic rate of birds is associated with habitat temperature and precipitation, not primary productivity. *Proc. Biol. Sci.* 274, 287–293. doi: 10.1098/rspb.2006.3727

Wijchers, P. J., and de Laat, W. (2011). Genome organization influences partner selection for chromosomal rearrangements. *Trends Genet.* 27, 63–71. doi: 10.1016/j.tig.2010.11.001

Willett, C. S., and Burton, R. S. (2001). Viability of cytochrome c genotypes depends on cytoplasmic backgrounds in *Tigriopus californicus*. *Evolution (N. Y).* 55, 1592–1599. doi: 10.1111/j.0014-3820.2001.tb00678.x

Willett, C. S., and Burton, R. S. (2004). Evolution of interacting proteins in the mitochondrial electron transport system in a marine copepod. *Mol. Biol. Evol.* 21, 443–453. doi: 10.1093/molbev/msh031

Wolff, J. N., Ladoukakis, E. D., Enríquez, J. A., and Dowling, D. K. (2014). Mitonuclear interactions: evolutionary consequences over multiple biological scales. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130443. doi: 10.1098/rstb.2013.0443

Wu, M., Gu, J., Guo, R., Huang, Y., and Yang, M. (2016). Structure of mammalian respiratory supercomplex I1III2IV1. *Cell* 167, 1598–1609. doi: 10.1016/j.cell.2016.11.012

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yeaman, S. (2013). Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1743–E1751. doi: 10.1073/pnas.1219381110

Yeaman, S., and Whitlock, M. C. (2011). The genetic architecture of adaptation under migration–selection balance. *Evolution (N. Y).* 65, 1897–1911. doi: 10.1111/j.1558-5646.2011.01269.x

Zhang, F., and Broughton, R. E. (2015). Heterogeneous natural selection on oxidative phosphorylation genes among fishes with extreme high and low aerobic performance. *BMC Evol. Biol.* 15, 173. doi: 10.1186/s12862-015-0453-7

Zhang, J., Hua, Z., Tame, J. R. H., Lu, G., Zhang, R., and Gu, X. (1996). The crystal structure of a high oxygen affinity species of haemoglobin (bar-headed goose haemoglobin in the oxy form). *J. Mol. Biol.* 255, 484–493. doi: 10.1006/jmbi.1996.0040

Zhu, J., Vinothkumar, K. R., and Hirst, J. (2016). Structure of mammalian respiratory complex I. *Nature* 536, 354–358. doi: 10.1038/nature19095

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 05 April 2018
doi: 10.3389/fgene.2018.00079

The Distributions and Boundary of Two Distinct, Local Forms of Japanese Pond Frog, *Pelophylax porosus brevipodus*, Inferred From Sequences of Mitochondrial DNA

*Yukari Nagai[1], Toshio Doi[2], Kunio Ito[3], Yoshiaki Yuasa[4], Takeshi Fujitani[5], Jun-ichi Naito[6], Mitsuaki Ogata[7] and Ikuo Miura[8]**

[1] Department of Biology, Graduate School of Science, Hiroshima University, Higashihiroshima, Japan, [2] Environmental Assessment and Symbiosis Promotion Division, Kobe Municipal Office, Kobe, Japan, [3] Kawasaki Senior High School Attached to Kawasaki Medical School, Kurashiki, Japan, [4] Himeji City Aquarium, Himeji, Japan, [5] Higashiyama Zoo and Botanical Gardens Information, Nagoya, Japan, [6] Society for the Study of Natural History of Nishi-Chugoku Mountains, Hiroshima, Japan, [7] Preservation and Research Center, The City of Yokohama, Yokohama, Japan, [8] Amphibian Research Center, Hiroshima University, Higashihiroshima, Japan

OPEN ACCESS

Edited by:
Tariq Ezaz,
University of Canberra, Australia

Reviewed by:
Tony Gamble,
Marquette University, United States
Tim Page,
Griffith University, Australia

****Correspondence:***
Ikuo Miura
imiura@hiroshima-u.ac.jp

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *21 February 2017*
Accepted: *26 February 2018*
Published: *05 April 2018*

Citation:
Nagai Y, Doi T, Ito K, Yuasa Y, Fujitani T, Naito J-i, Ogata M and Miura I (2018) The Distributions and Boundary of Two Distinct, Local Forms of Japanese Pond Frog, Pelophylax porosus brevipodus, Inferred From Sequences of Mitochondrial DNA. Front. Genet. 9:79. doi: 10.3389/fgene.2018.00079

The Nagoya Daruma pond frog *Pelophylax porosus brevipodus* is distributed in western Japan and is traditionally divided into two local forms: the Okayama form in the west and the Nagoya form in the east. These two forms are genetically differentiated, but have never been defined taxonomically because their distributions are unclear to date. To complete the distributions and identify the boundary of the two forms, we genetically investigated 16 populations including eight populations located within the unexamined area. We found that the distributional boundary is located within a small area of Hyogo Prefecture where haplotypes of mitochondrial *cytochrome b* (*cytb*) and D-loop region corresponding to the two forms co-existed. On the other hand, the polymorphic site of the nuclear gene *SOX3* revealed introgression over the boundary into Okayama *cytb* clade. These results suggest that the two forms were geographically isolated from each other in the past, and secondarily contacted and then accepted one-way introgression. As a next step of the research, taxonomic approach is expected to define the two forms.

Keywords: Japanese pond frog, *cytochrome b*, D-loop, *SOX3*, two major forms

INTRODUCTION

Two pond frog species live in the Japanese islands, *Pelophylax nigromaculatus* and *Pelophylax porosus*. The latter species is endemic to Japan and is called the Daruma pond frog. It is similar to a traditional Japanese Daruma doll with its round shape. This species is comprised of two subspecies: *P. p. porosus* (Tokyo Daruma pond frog), which is distributed in eastern Japan, and *P. p. brevipodus* (Nagoya Daruma pond frog), which is distributed in western Japan. *P. p. brevipodus* is traditionally divided into two distinct, local forms called the Okayama form in the west and the Nagoya form in the east (Ito, 1941; Moriya, 1954; Kawamura, 1962; Matsui and Hikida, 1985). They are genetically differentiated from each other by their external morphologies (**Figure 1**), mating calls, sex chromosomes, allozymes and mitochondrial genes (Moriya, 1951, 1954; Nishioka et al., 1992; Nishioka and Sumida, 1994; Ueda, 1994; Sumida et al., 1998, 2000a,b; Komaki et al., 2015). However, the two forms have never been defined taxonomically because their distributions are unclear. Since the genetic researches on the two local forms to date were always restricted to several representative populations, the area covering around 150 km between the two forms remains unstudied. It is still unknown whether the two forms are geographically separated or distributed sympatrically with mutual genetic introgression. Such information is definitely necessary for judging taxonomic positions of the two forms. Recently, the geographic populations of the Okayama form have been declining and are concerned about their possible extinction (Okochi et al., 1997). The degradation is especially severe in the western edge of the distribution, Hiroshima Prefecture, where only a few tiny populations have survived (Naito et al., 2014). Conservation of the population and environment is an urgent issue and taxonomic definition of the form is expected to assist the conservation activities.

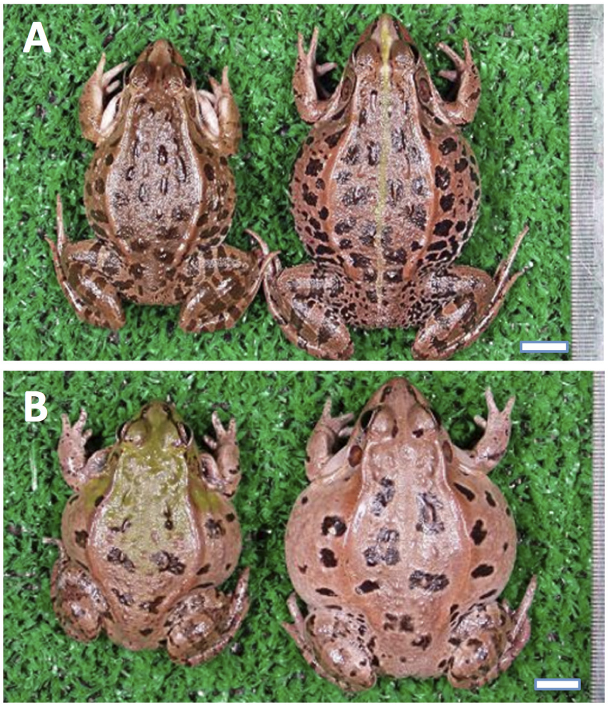


FIGURE 1 | External appearance of *Pelophylax porosus brevipodus* belonging to the two local forms. **(A)** The Nagoya form and **(B)** the Okayama form. Males and females are placed on the left and right, respectively. Central line on the back is seen in the female of Nagoya form. The line is absent in all frogs of Okayama form ever examined. The black spots on the back are larger and lower in number in the Okayama form than Nagoya form.

In this study, we collected samples of the two major forms in western Japan and investigated sequences of mitochondrial and nuclear genes in order to assess whether the two forms are separated geographically or are distributed sympatrically with mutual genetic introgression. In particular, the eight populations in Okayama and Hyogo Prefectures are located between the known distributions of the two forms and were genetically examined for the first time in 63 years since the primary morphological study of Moriya (1954) (**Figure 2**).

MATERIALS AND METHODS

Frogs

The number of frogs of *Pelophylax porosus brevipodus*, *P. p. porosus* and *P. nigromaculatus* used for sequence analyses are listed in **Table 1** and their collecting locations are shown in **Figure 2**. We collected three frogs of *P. p. brevipodus* each from Aichi and Gifu Prefectures, and reared them at our laboratory, while all other tissue samples were taken from the toe-clips in the fields, and stored in 100% ethanol until use. The frogs were thereafter released to the fields. Animal care and experimental procedures were conducted under approval of the Committee for Ethics in Animal Experimentation at Hiroshima University (Permit Number: G13-3).

DNA Extraction and PCR Amplification

Genomic DNA was extracted from the tissue samples using DNeasy blood and tissue kit (QIAGEN) according to the manufacture's instruction. Mitochondrial *cytochrome b* and nuclear *SOX3* fragments were amplified in 50 μl solution including 1.0 μl of DNA solution, 0.2 μl GXL Taq polymerase (TaKaRa), 5 μl of 10× Buffer, 4 μl of 2.5 mM dNTP, and 1 μl of 12.5 mM primers at 98°C for 5 s followed by 30 cycles of 98°C

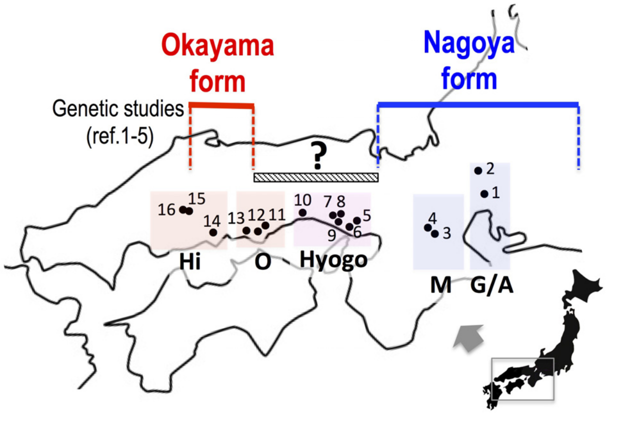


FIGURE 2 | The collecting localities of *Pelophylax porosus brevipodus*. Populations 5 to 12, which are placed under a hatched bar (150 km) with a question mark, have never been examined genetically since the primary morphological study of Moriya (1954). References 1–5: Nishioka et al. (1992), Sumida et al. (1998, 2000a,b), Komaki et al. (2015). See the localities (shown in number) in **Table 1**. Hi, Hiroshima prefecture; O, Okayama prefecture; Hyogo, Hyogo prefecture; M, Mie prefecture; G/A, Gifu and Aichi prefectures.

TABLE 1 | Populations, haplotype of mitochondrial *cytochrome b,* D-loop region, and genotype of nuclear *SOX3*.

Locality No.	Species	Population	Prefecture	City	Town or area	No. of frogs (♂, ♀, juvenile)	*cytb* haplotype	Repeats in D-loop region	*SOX3* (233rd)
1	*Pelophylax*	K-Nagoya	Aichi	Kita-Nagoya	Shikatsu	3 (2,1,0)	A1, A3	AB	GG
2	*porosus*	Gifu	Gifu	Gifu		3 (1,2,0)	A2, A3, A4		GG
3	*brevipodus*	Iga A	Mie	Iga	(A)	4 (1,3,0)	B1		GG
4		Iga B		Iga	(B)	2 (0,2,0)	B3	ABAB	GG
5		Kobe-O	Hyogo	Kobe	Oshibedani	9 (7,2,0)	B1	AB	GG
6		Kobe-H		Kobe	Hirano-machi	10 (3,2,5)	B1		GG
7		Kakogawa-YA		Kakogawa	Yahata (A)	30 (3,6,21)	B2, B1,C3		GT, TT
8		Kakogawa-YB		Kakogawa	Yahata (B)	6 (1,1,4)	B1, C3		GG, GT, TT
9		Kagogawa-I		Kakogawa	Inami	10 (2,5,3)	B1, C3	ABABAB, ABA	GT, TT
10		Ako		Ako	Fukuura	12 (4,8,0)	C2, C3		GT, TT
11		Okayama-S	Okayama	Okayama	Seto	11 (5,5,1)	C1, C3	ABA	TT
12		Okayama-N		Okayama	Nodono	20 (5,11,4)	C1		GG, GT, TT
13		Kurashiki		Kurashiki	Mabi	8 (1,2,5)	C1		TT
14		Fukuyama	Hiroshima	Fukuyama	Kannabe	8 (3,5,0)	C1	ABA	TT
15		Miyoshi-Y		Miyoshi	kisa, Yasuda	10 (0,0,10)	C1	ABA	TT
16		Miyoshi-K		Miyoshi	Kisa, Kaitahara	10 (0,0,10)	C1	ABA	TT
	P. p. porosus	Itako	Ibaraki	Itako		11 (5,6,0)	P1,P2,P3,P4		GG
	P. nigromaculatus	Outgroup	Hiroshima	Miyoshi	Kisa, Kaitahara	1 (0,0,1)			GG
	P. fukienensis	Outgroup	Taiwan				AB029941.1		

10 s, 64°C for 40 s, and 72°C for 60 s. The amplified product was purified using GFX PCR DNA and Gel band purification kit (GE Healthcare), and was used for nucleotide sequence determination with 3130XL sequencing machine (ABI).

Mitochondrial fragments including D-loop region (300∼500 bp) were amplified and purified by the above methods, and were cloned into pUC118 vector using Mighty cloning kit (TaKaRa) with competent cell DH5α (Ecos, Nippon gene) according to the manufactures' instructions. One to three colonies were picked up and the nucleotide sequences were determined by the method described above. Gene trees were constructed based on the nucleotide sequence of cytochrome-b gene by the methods of maximum likelihood (ML), neighbor joining (NJ) and maximum parsimony (MP) methods using Mega 7 software (Kumar et al., 2016). *p*-distance was also calculated using the above software. Primers used are forward 5′-CCA TGC ACT ACA CAG CCG ACA-3′ and reverse 5′-AGG TTT TTG CGA TAG GGC GGA-A3′ for *cytochrome b* (designed in this study using software Genetix ver. 7.3, Genetix corp.), S1 5′-GTG CGC TCC TCC TGC TTC TTT-3′ and A1 5′-TCC TCA AGT TTT CTG CAT TCT GAT-3′ for *SOX3* (Miura et al., 2016), and F23 5′-ATG AAT GCT ATA ATG ACA TAA TGT-3′ and R21 5′-TGC TGG CTC CTA AGG CCA GTG GAG GGC TGT-3′ for D-loop region (Sumida et al., 2000a). The sequences of ten haplotypes (A1–4, B1–3, and C1–3) of *cytochrome b* have been deposited with the DDBJ Data Libraries under the accession numbers LC217488-LC217457, and the sequences of *SOX3* (Kurashiki and Iga populations), under the accession numbers LC316654 and LC316655, respectively.

RESULTS

Mitochondrial *Cytochrome b*

We collected samples consisting of 156 specimens (38 males, 55 females, and 63 juveniles) from 16 populations covering their present habitat in western Japan (**Figure 2** and **Table 1**). We determined the nucleotide sequences of 566 base pairs of the mitochondrial cytochrome-b gene. Ten haplotypes were identified and the gene tree was constructed using the maximum likelihood (ML) method (**Figure 3** and **Table 1**). The haplotypes formed two distinct clades, which are designated PB-N and PB-O because they correspond to the Nagoya form and Okayama form, respectively. The genetic (*p*) distance between the two clades was 0.055 and that between the two subspecies was 0.054, suggesting that the genetic relationships among *P. p. porosus* and the two local forms of *P. p. brevipodus* are within almost equal range of each other. Notably, two haplotypes of PB-O and PB-N were detected in Kakogawa-YA, -YB and -I populations of Hyogo Prefecture (7, 8, and 9 in **Figures 2–4**), which were located immediately east over the Kakogawa River. Three of 30, one of six, and three of ten specimens examined in the populations had PB-O haplotypes, while the others possessed

PB-N haplotypes. This indicates that the boundary between the two clades is restricted to the small area of Hyogo Prefecture.

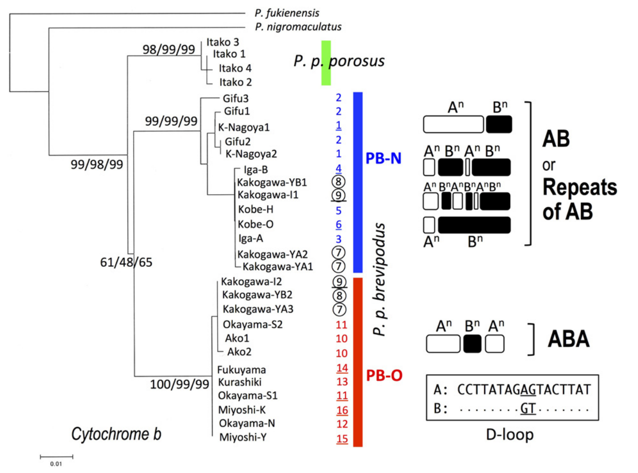


FIGURE 3 | Maximum-likelihood (ML) tree of *Pelophylax porosus brevipodus* collected from 16 populations based on mitochondrial cytochrome-b gene and the repeating pattern of 17-bp sequences in D-loop region. The numbers at each node on the tree are ML/NJ/MP bootstrap values. The numbers put alongside vertical bars are population numbers shown in **Table 1** and **Figure 2**: the Kakogawa populations 7, 8, and 9 are circled and the populations, of which D-loop regions were examined, are underlined. Repeats of type A and B in D-loop region are indicated by open and closed round squares, respectively, and the sequences of type A and B are shown in the box. Dots in type B indicate the same nucleotides as type A.

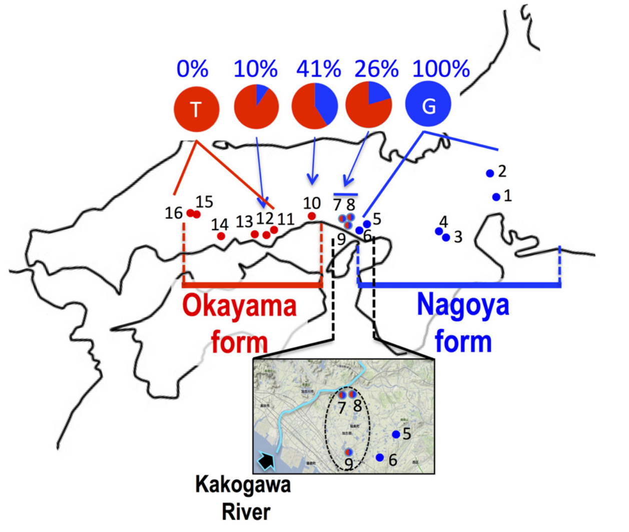


FIGURE 4 | Haplogroup of *cytochrome b* and frequency of haplotype at 233rd position of *SOX3* in 16 populations of *P. p. brevipodus*. The red and blue circles indicate haplogroups of clade PB-O and PB-N, respectively. Red-blue (half-half) circles indicate sympatric distribution of the two haplogroups within a single population. The area including populations 5–9 are magnified with an arrow showing Kakogawa River. The frequencies of thymine and guanine haplotypes at 233rd positions of *SOX3* are indicated in red and blue, respectively. See the localities (shown in number) in **Table 1**.

Repeated Sequence in D-loop Region

The D-loop region of the mitochondrial genome includes a highly repeated sequence. We cloned this region and determined the nucleotide sequences of specimens from eight populations of *P. p. brevipodus* (**Table 1**). The repeated region comprised of repeats of two kinds of 17-bp units designated types A and B of which nucleotides at the 9th and 10th positions were different: AG and GT, respectively (**Figure 3** and Supplementary Table 1). The repeated pattern was different among populations (**Figure 3** and Supplementary Table 1). Pattern AB was specific to the Kita-Nagoya population (population No. 1 in **Table 1** and **Figure 3**), while pattern ABA was observed in the Kakogawa I2, Okayama-S1 and three Hiroshima populations (9, 11, and 14–16). In the Iga, Kobe-O and Kakogawa I1 populations (4, 6, and 9), the observed patterns were ABAB, AB, and ABABAB, respectively. All the repeat patterns were thus classified into two types: AB (or repeats of AB) and ABA. The two types corresponded with the two major clades of *cytochrome b*: PB-N with AB or repeats of AB and PB-O with ABA, respectively. In the Kakogawa-I population (9), the specimen with the PB-N *cyt-b* haplotype had the ABABAB pattern while that with the PB-O haplotype possessed the ABA pattern.

SOX3

The sequence of 860 base pairs of the nuclear *SOX3* gene was determined for 140 specimens from 16 populations. The nucleotide at position 233 varied by population (**Figure 4** and **Table 1**). In the six eastern populations (1–6: Kita-Nagoya, Gifu, Iga-A, Iga-B, Kobe-O, and Kobe-H), all specimens were homozygous for guanine. On the other hand, in the five western populations (11, 13–16: Okayama, Kurashiki, and three of Hiroshima Prefecture), all were homozygous for thymine. In the other five populations (7–10, 12) located at the intermediate regions, the specimens were heterozygous or homozygous for guanine or thymine (**Figure 4**). The frequency of guanine in these populations varied from 10 to 41%.

DISCUSSION

Based on the mitochondrial *cytochrome b*, the two major *P. p. brevipodus* forms of Okayama and Nagoya were identified as distinct clades, and two major types of the D-loop region supported the *cytb* clades. The genetic distance (*p*-distance, 0.055) between the two forms was very similar to that (0.054) between the two subspecies. These genetic relationships are well supported by another study that used mitochondrial and nuclear genes (Komaki et al., 2015). The distribution boundary between the two forms was for the first time found in this study. It is located at a very small area that included the Kakogawa populations (7–9 in **Figures 2**, **4**) of Hyogo Prefecture and was where two haplotypes of the Okayama and Nagoya forms co-existed. This shows that the two forms were geographically isolated from each other in the past and have secondarily contacted at the small area after they were genetically differentiated. The molecular clock based on *cytochrome b* and seven nuclear genes estimates that the two forms were separated from each other around 1.3

MYA (Komaki et al., 2015). Currently, no remarkable barrier of geographic structure could be identified around the boundary that separates the two forms of *Pelophylax porosus brevipodus*, or no geographic event that actually occurred 1.3 MYA is known. However, some geographic barrier must have existed in the past and prevented crossings across the boundary area, because many other animals, such as grasshopper, harvestman, frog, landing snail, and monkey, are likewise genetically differentiated between the west and the east of the boundary region (Tsurusaki et al., 1991; Kawakami, 1999; Nishi and Sota, 2005; Kawamoto et al., 2007; Nishizawa et al., 2011). Conversely, it was found that nuclear gene *SOX3* showed introgression over the boundary from eastern Nagoya form into the western Okayama form. The genetic affinity between the two forms is also confirmed by the results of artificial crossings in the study of Moriya (1960a,b), showing fertile hybrids between the two forms. However, it was quite difficult in this study to recognize the genetic introgression in external morphology: for example, a central line on the back, which was normally observed in 44% frogs of Nagoya form, was not found in any populations of the Okayama form (except one specimen in Kakogawa I population, No. 9 on the map). A deeper analysis on nuclear genomes of the two forms focusing on the populations around the boundary is required to verify the on-going introgression of the genomes.

CONCLUSION

It is evident that the two local forms had been once isolated from each other and accumulated their genetic differences, and thereafter they have secondarily contacted immediately east over the Kakogawa River (**Figure 4**) and possibly the Okayama form is now accepting introgression from the Nagoya form. We speculate that the ancestral lineage of the Okayama form remains around the eastern edge of the range (Hiroshima Prefecture and the western region of Okayama Prefecture in **Figure 2**) where population declining and extinction are concerned. At a next step of the research, taxonomic definition of the two forms are expected (for example, name of Okayama Daruma pond frog is given to the Okayama form), because they are precisely identified and the geographic boundary between the two forms is very clear based on the mitochondrial DNA. Unfortunately, the previous study on morphology (Moriya, 1954) used no statistical analyses and examined only one population of the Okayama form, and the previous mating call analysis (Ueda, 1994) was restricted to just one or two populations of each form, which are located at the extremes in distribution. Hence, a future taxonomic approach needs to consider the distribution range for choosing populations and complete investigation on the morphology and mating calls of the two forms.

AUTHOR CONTRIBUTIONS

IM and MO designed the study and wrote the manuscript. YN, MO, and IM performed the experiments and analyses. TD, KI, YY, TF, and J-iN collected the specimens and discussed about the results. All authors read and approved the final manuscripts.

ACKNOWLEDGMENTS

We thank T. Morita and the late T. Utsunomiya for providing us the tissue samples of the specimens from Fukuyama city in Hiroshima Prefecture.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00079/full#supplementary-material

REFERENCES

Ito, R. (1941). On the two types of *Rana nigromaculata*. *Rep. Nagoya Biol. Sci.* 8, 77–88.

Kawakami, Y. (1999). Geographic variation of the brachypterous grasshopper *Parapodisma setouchiensis* group in western Honshu, with its taxonomic revision. *Species Divers.* 4, 43–61.

Kawamoto, Y., Shotake, T., Nozawa, K., Kawamoto, S., Tomari, K., Kawai, S., et al. (2007). Postglacial population expansion of Japanese macaques (*Macaca fuscata*) inferred from mitochondrial DNA phylogeography. *Primates* 48, 27–40. doi: 10.1007/s10329-006-0013-2

Kawamura, T. (1962). On the names of some Japanese frogs. *J. Sci. Hiroshima Univ. Ser. B Div. 1* 20, 181–193.

Komaki, S., Igawa, T., Lin, S., Tojo, K., Min, M., and Sumida, M. (2015). Robust molecular phylogeny and palaeodistribution modeling resolve a complex evolutionary history: glacial cycling drove recurrent mtDNA introgression among *Pelophylax* frogs in East Asia. *J. Biogeogr.* 42, 2159–2171. doi: 10.1111/jbi.12584

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Matsui, M., and Hikida, T. (1985). *Tompoptera porosa* Cope, 1868, a senior synonym of *Rana brevipoda* Ito, 1941 (Ranidae). *J. Herpetol.* 19, 423–425. doi: 10.2307/1564274

Miura, I., Ohtani, H., Ogata, M., and Ezaz, T. (2016). Evolutionary changes in sensitivity to hormonally induced gonadal sex reversal in a frog species. *Sex. Dev.* 10, 79–90. doi: 10.1159/000445848

Moriya, K. (1951). On isolating mechanisms between two subspecies of pond frog, *Rana nigromaculata* HALLOWELL. I. Differences in the morphological characters. *J. Sci. Hiroshima Univ. Ser. B Div. 1* 12, 47–56.

Moriya, K. (1954). Studies on the five races of the Japanese pond frog, *Rana nigromaculata* HALLOWELL. I. Differences in the morphological characters. *J. Sci. Hiroshima Univ. Ser. B Div. 1* 15, 1–21.

Moriya, K. (1960a). Studies on the five races of the Japanese pond frog, *Rana nigromaculata* HALLOWELL. II. Differences in character of development. *J. Sci. Hiroshima Univ. Ser. B Div. 1* 18, 109–124.

Moriya, K. (1960b). Studies on the five races of the Japanese pond frog, *Rana nigromaculata* HALLOWELL. III. Sterility in interracial hybrids. *J. Sci. Hiroshima Univ. Ser. B Div. 1* 18, 125–156.

Naito, J., Sakamura, A., Nakayama, T., and Matsubara, C. (2014). The conservation on the Daruma pond frog (*Rana porosa brevipoda*) in biotope area of Haizuka Dam. *Hibakagaku* 250, 1–27.

Nishi, H., and Sota, T. (2005). Phylogenetic study of the land snail *Euhadra* in Chugoku district based on analysis of mitochondrial DNA sequences. *Bull. Hoshizaki Green Found.* 8, 185–196.

Nishioka, M., and Sumida, M. (1994). The position of sex-determination in the *Rana nigromaculata* and *Rana brevipoda*. *Sci. Rep. Lab. Amphibian Biol. Hiroshima Univ.* 13, 51–97.

Nishioka, M., Sumida, M., and Ohtani, H. (1992). Different ion of 70 populations in the *Rana nigromaculata* group by method of electrophoretic analysis. *Sci. Rep. Lab. Amphibian Biol. Hiroshima Univ.* 12, 1–70.

Nishizawa, T., Kurabayashi, A., Kunihara, T., Sano, N., Fujii, T., and Sumida, M. (2011). Mitochondrial DNA diversification, molecular phylogeny, and biogeography of the primitive rhacophorid genus *Buergeria* in East Asia. *Mol. Phylogenet. Evol.* 59, 139–147. doi: 10.1016/j.ympev.2011.01.015

Okochi, I., Utsunomiya, T., Utsunomiya, Y., and Numasawa, M. (1997). Captive breeding andre-inforcement to an endangered population of *Rana porosa brevipoda* Ito (Ranidae: Amphibia). *Jpn. J. Conserv. Ecol.* 2, 135–146.

Sumida, M., Kaneda, H., Kato, Y., Kanamori, Y., Yonezawa, H., and Nishioka, M. (2000a). Sequence variation and structural conservation in the D-loop region and flanking genes of mitochondrial DNA from Japanese pond frogs. *Genes Genet. Syst.* 75, 79–92.

Sumida, M., Ogata, M., Kaneda, H., and Yonekawa, H. (1998). Evolutionary relationships among Japanese pond frogs inferred from mitochondrial DNA sequences of cytochrome b and 12S ribosomal RNA genes. *Genes Genet. Syst.* 73, 121–133. doi: 10.1266/ggs.73.121

Sumida, M., Ogata, M., and Nishioka, M. (2000b). Molecular phylogenetic relationships of pond frogs distributed in the Palearctic region inferred from DNA sequences of mitochondrial 12S ribosomal RNA and cytochrome b genes. *Mol. Phylogenet. Evol.* 16, 278–285.

Tsurusaki, N., Murakami, M., and Shimokawa, K. (1991). Geographic variation of chromosomes in the Japanese harvestman, *Gagrellopsis nodulifera*, with special reference to a hybrid zone in western Honshu. *Zool. Sci.* 8, 265–275.

Ueda, H. (1994). Mating calls of the pond frog species distributed in the Far East and their artificial hybrids. *Sci. Rep. Lab. Amphibian Biol. Hiroshima Univ.* 13, 197–232.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a past co-authorship with the authors MO and IM.

PROTOCOLS
published: 21 February 2017
doi: 10.3389/fpls.2017.00222

MLPA-Based Analysis of Copy Number Variation in Plant Populations

Anna Samelak-Czajka[1], *Malgorzata Marszalek-Zenczak*[2], *Malgorzata Marcinkowska-Swojak*[3], *Piotr Kozlowski*[3], *Marek Figlerowicz*[1,2] and *Agnieszka Zmienko*[1,2]*

[1] *Institute of Computing Science, Faculty of Computing, Poznan University of Technology, Poznan, Poland,* [2] *Department of Molecular and Systems Biology, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland,* [3] *Department of Molecular Genetics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*

Copy number variants (CNVs) are intraspecies duplications/deletions of large DNA segments (>1 kb). A growing number of reports highlight the functional and evolutionary impact of CNV in plants, increasing the need for appropriate tools that enable locus-specific CNV genotyping on a population scale. Multiplex ligation-dependent probe amplification (MLPA) is considered a gold standard in genotyping CNV in humans. Consequently, numerous commercial MLPA assays for CNV-related human diseases have been created. We routinely genotype complex multiallelic CNVs in human and plant genomes using the modified MLPA procedure based on fully synthesized oligonucleotide probes (90–200 nt), which greatly simplifies the design process and allows for the development of custom assays. Here, we present a step-by-step protocol for gene-specific MLPA probe design, multiplexed assay setup and data analysis in a copy number genotyping experiment in plants. As a case study, we present the results of a custom assay designed to genotype the copy number status of 12 protein coding genes in a population of 80 *Arabidopsis* accessions. The genes were pre-selected based on whole genome sequencing data and are localized in the genomic regions that display different levels of population-scale variation (non-variable, biallelic, or multiallelic, as well as CNVs overlapping whole genes or their fragments). The presented approach is suitable for population-scale validation of the CNV regions inferred from whole genome sequencing data analysis and for focused analysis of selected genes of interest. It can also be very easily adopted for any plant species, following optimization of the template amount and design of the appropriate control probes, according to the general guidelines presented in this paper.

Keywords: structural variation, MLPA, 1001 Arabidopsis Genomes project, CNV genotyping, multiplexing

OPEN ACCESS

Edited by:
Scott V. Edwards,
Harvard University, USA

Reviewed by:
Nathan Lewis Clark,
University of Pittsburgh, USA
John Malone,
University of Connecticut, USA
Pierre Baduel,
John Innes Centre (BBSRC), USA

****Correspondence:***
Agnieszka Zmienko
akisiel@ibch.poznan.pl

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *05 October 2016*
Accepted: *06 February 2017*
Published: *21 February 2017*

Citation:
Samelak-Czajka A, Marszalek-Zenczak M, Marcinkowska-Swojak M, Kozlowski P, Figlerowicz M and Zmienko A (2017) MLPA-Based Analysis of Copy Number Variation in Plant Populations. Front. Plant Sci. 8:222. doi: 10.3389/fpls.2017.00222

INTRODUCTION

The rise of high-throughput genomics techniques – DNA arrays and, more recently, whole-genome sequencing (WGS) – has revealed the structural complexity and dynamics of eukaryotic genomes. In particular, the ability to re-sequence and compare hundreds or even thousands of genomes of individuals within one species has paved the way for the investigation of the extent to which individual genomes differ from each other. One type of structural variation that is ubiquitous in the genomes of humans, animals and plants is copy number variation (CNV). This term refers to intraspecies duplications and deletions of large DNA segments, usually >1 kb [although variants >50 bp have been recently included in this spectrum (Alkan et al., 2011)]. The human genome is the most intensively studied eukaryotic genome in terms of the distribution and functional significance of CNVs and the mechanisms leading to the formation of copy number rearrangements (Zarrei et al., 2015). However, the number of species for which CNV regions have been inferred on the genome-wide scale is growing rapidly. For plants, this list includes maize, rice, sorghum, *Arabidopsis* (*Arabidopsis thaliana)*, soybean, wheat, and barley (Springer et al., 2009; Beló et al., 2010; Swanson-Wagner et al., 2010; Cao et al., 2011; Saintenac et al., 2011; Zheng et al., 2011; McHale et al., 2012; Muñoz-Amatriaín et al., 2013; Duitama et al., 2015; Bai et al., 2016). As in humans, CNV regions in plants are not uniformly distributed across the chromosomes. Although they are more common in the intergenic regions, they also co-localize with hundreds of protein-coding genes (Swanson-Wagner et al., 2010; Beló et al., 2010; McHale et al., 2012; Muñoz-Amatriaín et al., 2013). The ability to alter the gene structure and copy number makes CNV an important factor that influences gene expression (Żmieńko et al., 2014). By the gene dosage effect, CNVs can also affect the interaction of the genes' products within protein and metabolic networks (Hanada et al., 2011; Conant et al., 2014). Quite often, such variation accounts for adaptive traits or - as shown for humans - can underlie disease (Stankiewicz and Lupski, 2010; Zarrei et al., 2015). In plants, a growing number of studies highlight the shaping role of CNVs in genome evolution, phenotypic variation and – sometimes rapid - adaptation to environmental challenges (Gaines et al., 2010; Cook et al., 2012; Maron et al., 2013; Chang et al., 2015; Wang et al., 2015). Therefore, it is anticipated that the number of genetic studies focused on individual CNVs of interest will grow and that new CNV-associated traits will be revealed.

In-depth analysis of individual CNVs in plants has rarely been conducted (Gaines et al., 2010; Cook et al., 2012; Maron et al., 2013). Likewise, in plants for which the CNV regions were inferred from WGS data, the subsequent validation was not conducted or was limited to the PCR-based detection of CNV deletions (Swanson-Wagner et al., 2010; Cao et al., 2011; Tan et al., 2012; Bai et al., 2016). Therefore, there is an urgent need to widen the range of experimental studies of CNV in plants to contribute to the creation of high-confidence CNV maps and enhance association studies linking CNVs with phenotypic traits in plant species. In this context, the lack of validated experimental approaches for the analysis of individual CNVs in plants is apparent, as opposed to the well-established methods and standardized protocols available for the human genome.

The range of popular molecular methods used for DNA copy number genotyping in humans is wide (Ceulemans et al., 2012; Cantsilieris et al., 2013; Bharuthram et al., 2014). Among them, multiplex ligation-dependent probe amplification (MLPA), first introduced in 2002 (Schouten et al., 2002) and later developed by the MRC Holland company, is considered a gold standard in the diagnosis of numerous DNA copy number-related human diseases (Hömig-Hölzel and Savola, 2012). MLPA is a simple and robust method of relative quantification of DNA sequences on a population scale. The standard multiplex assay utilizes up to 50 probes targeting specific DNA regions (e.g., exons in a gene of interest). Each probe is composed of two half-probes (physically separate DNA fragments, one fully synthetic and one clone-derived) that match the target sequence in directly adjacent positions with their target-specific sequences (TSSs). Successful hybridization of both half-probes to the genomic DNA enables their ligation and linear amplification. The amplification products are then analyzed by capillary electrophoresis. Relative quantification of the signal peaks from fragments of unique size, generated by individual probes in the assay, provides information about the template DNA copy number. MLPA requires little genomic DNA input (Schouten et al., 2002). Additionally, the genomic sequence targeted by the probes is quite short (50–70 nt), which enables use of MLPA for the analysis of regions too small to be detected by the FISH method. MLPA has been shown to be superior to qPCR for gene copy number quantification (Perne et al., 2009; Cantsilieris et al., 2014). Additionally, it presents similar performance to droplet digital PCR in accurate quantification of up to eight gene copies, making it suitable for the analysis of multiallelic CNVs, i.e., those that exist in more than two genotypes in a population (Zmienko et al., 2016).

According to PubMed, the seminal MLPA work (Schouten et al., 2002) has been cited almost 450 times (∼220 times within 5 last years). Additionally, ∼2,000 articles in PubMed matched the search keyword "Multiplex Ligation-Dependent Probe Amplification". Among these papers, only 16 also matched the search keyword "plant". Those that actually described plant applications of MLPA involved alternative applications of this method: the detection of genetically modified organisms (GMO-MLPA) (Rudi et al., 2003), single nucleotide polymorphism (SNP) genotyping (Thumma et al., 2009), or gene expression analysis (RT-MLPA) (Li et al., 2009, 2011, 2013). However, none of these papers presented a primary MLPA application of copy number analysis. Several reasons might account for the fact that the MLPA approach has not been adopted by the plant community. One is much later recognition of the intraspecies variation and CNV prevalence in the plant genomes than in humans. Additionally, the commercial MLPA assays are focused on biomedical studies and cover only humans. Therefore, to assess plant genome variation with MLPA, it is necessary to self-design synthetic probes. It should be noted that, over the years, numerous modifications of the MLPA strategy have been introduced that simplify the probe design procedure (Marcinkowska et al., 2010; Ling et al., 2015, and references therein). In the current work, we present the optimized protocol for MLPA-based CNV analysis and provide guidelines for designing and performing MLPA assays in plants. The protocol is based on the MLPA adaptation developed previously by one of us (PK) that involves fully synthetic oligonucleotide probes, 90 to 200 nt in length, and allows for simultaneous genotyping of >30 different positions in the genomic DNA (Kozlowski et al., 2007). The protocol combines MLPA probe design, synthesis, experimental procedures, data preprocessing and analysis stages into one comprehensive procedure. The lack of MLPA-based genotyping studies in plants highlights the need for such an integrated resource. We also provided the probe design template, developed specifically for the presented MLPA variant. It allows for semi-automatic probe sequence setup, clarifies the idea of probe set composition and shortens the design process by days.

High and low copy level duplications may have different effects on the gene dosage and the phenotype, e.g., by triggering differences in gene expression level or inducing the silencing mechanisms in plants. Therefore, an important aspect of plant CNV genotyping studies is to estimate the actual gene copy numbers in the analyzed lines in order to analyze their influence on the trait of interest (Cook et al., 2014). To illustrate the performance of the MLPA method for precise DNA copy number genotyping in plant populations, we present exemplar assays for 12 genes with different levels of copy number diversity in a population of 80 *Arabidopsis* ecotypes, including multiallelic CNVs. We also describe the set of experimentally verified normalization control probes and the results of genomic DNA template amount optimization performed for this model species.

An advantage of the presented approach is that the assay - after it has been standardized for the particular organism – is always performed in the same conditions, regardless of the probe set composition. It may be utilized for the detailed analysis of a genomic region of interest using a set of MLPA probes scattered along this region or for large-scale validation/genotyping studies of WGS-based predicted CNVs, with 1-2 MLPA probes per inferred CNV.

MATERIALS AND EQUIPMENT

Materials

(1) High-quality genomic DNA for each analyzed sample, evaluated using a NanoDrop 2000 spectrophotometer (Thermo Scientific) and with standard gel electrophoresis; the working concentration is typically 0.4 to 50 ng/μl, depending on the species (see the following sections).

For *Arabidopsis*: We successfully genotyped CNVs using genomic DNA from 3-week-old rosette leaves extracted with a DNeasy Plant Mini Kit (Qiagen).

(2) Self-designed synthetic oligonucleotides (MLPA half-probes; see the following section for the probe design instructions) purchased from Integrated DNA Technologies (or similar provider) as 100 nmol oligo, purified by HPLC (for oligonucleotides up to 100 nt in length) or PAGE (for oligonucleotides over 100 nt in length); the right half-probes should be additionally modified by 5′ phosphorylation.

(3) Nuclease-free water (not DEPC-treated) (Ambion, cat. no. AM9938)

(4) SALSA MLPA EK-1 reagent kit (MRC-Holland, cat. no. EK1-FAM), which includes the following components:

SALSA MLPA Buffer
SALSA Ligase-65
Ligase Buffer A
Ligase Buffer B
SALSA PCR Primer MIX
SALSA Polymerase

(5) Consumables for capillary electrophoresis, depending on the instrument type; here, for the ABI Prism 3130XL Genetic Analyzer:

HiDi formamide (Thermo Fisher Scientific, cat. no. 4440753)
GeneScan 600 LIZ Size Standard (Thermo Fisher Scientific, cat. no 4366589)
POP7 Polymer (Thermo Fisher Scientific, cat. no 4352759).

Equipment

(1) 0.2 ml PCR strips and suitable caps, e.g., 8-Strip PCR tubes (Starlab, cat. no. I1402-3500) and 8-Strip caps (Starlab, cat. no. I1400-0800).
(2) Standard and multichannel pipettes.
(3) Thermocycler with heated lid (e.g., Bio-Rad T100 Thermal Cycler or equivalent).
(4) Vortex mixer (e.g., ELMI V-3 Sky Line or equivalent).
(5) Mini laboratory centrifuge with Eppendorf tube adapter and PCR strip adapter (e.g., Labnet Spectrafuge or equivalent).
(6) Capillary electrophoresis instrument (AppliedBiosystems ABI Prism 3130XL Genetic Analyzer or equivalent) or access to a capillary electrophoresis service provider.
(7) Software tool for the extraction of the intensity data after size-separation of MLPA reaction products (e.g., GeneMarker by SoftGenetics).

STEPWISE PROCEDURES

The general concept of the MLPA strategy is presented in **Figure 1**. The entire procedure involves three main stages: (A) designing the MLPA probes; (B) performing MLPA assay, which involves half-probes hybridization to DNA template, subsequent ligation and amplification; and (C) data collection and analysis, including the estimation of the copy number genotypes.

Stage A: Design the MLPA Probes (Time: Approximately 1 Week + Oligonucleotide Synthesis and Transportation by an External Provider)

The presented MLPA procedure based on fully synthetic oligonucleotide probes allows for simultaneous copy number analysis of ∼30 individual regions in the genomic DNA. Of these, at least 3 to 5 MLPA probes should target the confirmed non-variable control regions, distant from the studied genomic positions. These probes serve as normalization controls in the subsequent analysis of the MLPA data to account for the possible variation of the input DNA template amount and technical issues. The typical targets of the MLPA assays are protein-coding genes, as the changes in their copy number potentially affect the protein level and may contribute to the phenotype. The number of probes designed for each gene and their density in the covered genomic region depend on the user’s requirements.

The procedure for individual MLPA probe design has been graphically presented in Supplementary Figure S1 and is

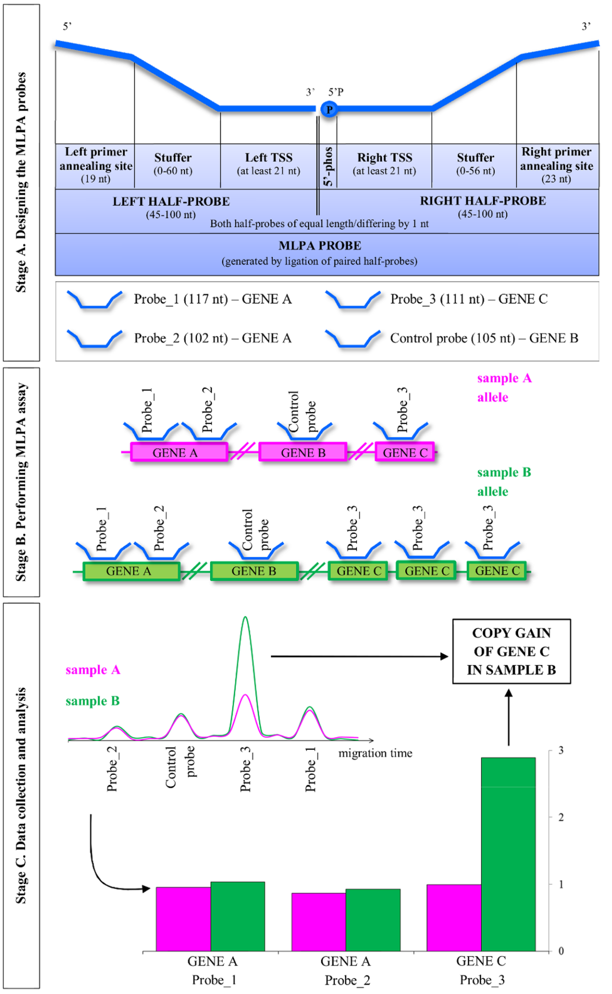

Stage A. Designing the MLPA probes
5'
3'
3'
5'P
P
Left primer annealing site (19 nt)
Stuffer (0-60 nt)
Left TSS (at least 21 nt)
5'-phos
Right TSS (at least 21 nt)
Stuffer (0-56 nt)
Right primer annealing site (23 nt)
LEFT HALF-PROBE (45-100 nt)
RIGHT HALF-PROBE (45-100 nt)
Both half-probes of equal length/differing by 1 nt
MLPA PROBE (generated by ligation of paired half-probes)
Probe_1 (117 nt) – GENE A
Probe_3 (111 nt) – GENE C
Probe_2 (102 nt) – GENE A
Control probe (105 nt) – GENE B
Stage B. Performing MLPA assay
sample A allele
Probe_1
Probe_2
Control probe
Probe_3
GENE A
GENE B
GENE C
sample B allele
Probe_1
Probe_2
Control probe
Probe_3
Probe_3
Probe_3
GENE A
GENE B
GENE C
GENE C
GENE C
Stage C. Data collection and analysis
sample A
sample B
COPY GAIN OF GENE C IN SAMPLE B
Probe_2
Control probe
Probe_3
Probe_1
migration time
3
2
1
0
GENE A Probe_1
GENE A Probe_2
GENE C Probe_3

FIGURE 1 | Overview of the multiplex ligation-dependent probe amplification (MLPA) method. MLPA is comprised of three main stages: designing the probes, performing the multiplex MLPA assay and data collection and analysis. The three stages are described in the detail in the main text. TSS, target-specific sequence; 5′-phos: phosphorylation at the 5′ end of the oligonucleotide. Note that *Arabidopsis*, as a self-pollinating plant, typically carries pairs of identical alleles. For simplicity, single alleles are depicted.

described in detail in the following sections. We used *Arabidopsis* gene *AT1G01040* encoding Dicer-like 1 protein as an example.

Select TSSs for the MLPA Probes

Step 1. Retrieve the genomic sequence of the gene of interest from the appropriate database, including the exon-intron positions. We recommend localizing the MLPA probes within the exon sequences because they display lower variation than the non-coding regions of genes.

For *Arabidopsis:* Use the gene locus identifier (e.g., *AT1G01040*) to localize that gene in the TAIR10 genomic sequence, available through the *Arabidopsis* genome browser[1], and display its splice variants, when applicable (Protein Coding Gene Models track). In *Arabidopsis*, protein coding genes have five exons on average, each with mean length of ~240 bp (Koralewski and Krutovsky, 2011). This length is sufficient for selecting two adjacent TSSs (one for each half-probe). Use the GBrowse navigation tools to zoom in to the selected exon and export its DNA sequence as a FASTA file.

Step 2. Ensure your sequence does not include any repetitive elements.

For *Arabidopsis,* rice, maize, wheat, and some other crops: Submit the extracted sequence to the CENSOR software tool (Kohany et al., 2006) that masks the repetitive elements in the query sequence using the collection of repeats for selected animal and plant species. Select a fragment of at least 100 nt that is not interrupted by any masked regions.

Step 3. If possible, check the selected sequence for the presence of SNPs and small indels.

For *Arabidopsis*: Use the 1001 Genomes Project VCF Subset tool[2] to download the subset of VCF files that contain full-genome VCF data for 1135 accessions (as of September 2016) (1001 Genomes Consortium, 2016). Download SNP information for the region and accessions of interest. Evaluate whether the selected sequence is free of common polymorphisms.

Step 4. From the selected region, choose two directly adjacent fragments of at least 21 nt (left and right TSS) and adjust their length and position so that the melting temperature (Tm) of each fragment will be as close as possible to 71°C (calculated with the free RaW program available from MRC Holland[3] with the following settings: method Go-Oli-Go, salt concentration 0.1 M, oligo concentration 1 μm). Avoid long homopolymer tracts and GC tracts of $\geq$4 bases.

Step 5. Join the adjacent left and right TSSs and use the resulting sequence in a homology search against the genomic sequence of the analyzed species to check for its specificity.

For *Arabidopsis*: Perform a BLAST search against *A. thaliana* NCBI reference genome with the following parameters: blastn algorithm, word size 7, match/mismatch scores 2;-3, gap costs 5;2, no sequence masking and filtering, *E*-value threshold 0.001.

Step 6. Repeat steps 3 to 5 until the pair of adjacent TSSs that satisfies all design criteria is found for a given gene.

Design the Half-Probes

Step 7. Add the respective PCR primer annealing sequence to each TSS and – optionally – the stuffer sequence, in the following order (see **Figure 1**):

for the left half-probe:

5′-left primer annealing sequence – stuffer – left TSS -3′, where the left primer annealing sequence is GGGTT CCCTAAGGGTTGGA;

for the right half-probe:

5′-right TSS – stuffer – right primer annealing sequence – 3′, where the right primer annealing sequence is TCTAGA TTGGATCTTGCTGGCGC.

For the stuffer, use the fragment of enterobacteria phage M13 sequence (NCBI/GenBank ID V00604, range: 3-119). This

[1] https://gbrowse.arabidopsis.org/cgi-bin/gb2/gbrowse/arabidopsis/

[2] http://tools.1001genomes.org

[3] http://www.mrc-holland.com/

fragment has no significant blastn matches to any eukaryotic genomic sequence deposited in the NCBI/RefSeq Representative Genome Database (accessed July 4th, 2016). It has been successfully applied as a stuffer in our previous MLPA assays performed for *Arabidopsis* and human DNA (Marcinkowska-Swojak et al., 2014; Klonowska et al., 2015; Zmienko et al., 2016).

Note: The addition of the optional stuffer sequence allows the user to adjust the length of the half-probes so that the resulting PCR amplification fragments would be of unique size and differ by 3 nt for probes in the 90-120 nt range and by 4 nt for probes >120 nt long. The length of the two half-probes in the pair should be the same or differ by 1 nt. For example, to obtain the MLPA probe of length 120, the left and right half-probe sequences should each be 60 nt long (and at least 21 nt of each half-probe should constitute TSS).

To facilitate the process of MLPA probe design and combining multiple MLPA probes in one experimental assay, we provided a Microsoft Excel template (Supplementary Table S1). This template includes the formulas that automatically adjust the length of the stuffer sequence and add the required adapter sequences to both the left and right half-probes. As a result, the final sequence of the MLPA probe of the desired length is returned. The user can choose the MLPA probe length. Typically, when fewer than the maximal number of MLPA probes are included in the assay, we recommend designing shorter probes to minimize the oligonucleotide synthesis costs. Often, the MLPA assays contain two or more probes targeting adjacent genomic regions. We recommend randomization of these probe MLPA lengths to minimize the influence of the possible biases or artifacts. Likewise, we recommend distributing the control probe lengths to cover the entire range of the MLPA probes in the assay.

For *Arabidopsis*: We provide pre-designed sequences for five control MLPA probes (ctrl1–ctrl5) that target genes located on chromosomes 1, 2, 4, and 5. The first gene is *DCL1*, coding for a RNA helicase involved in microRNA processing. The second gene encodes an oxidoreductase belonging to a zinc-binding dehydrogenase family protein. The third non-variable gene is *APG10*, coding for a BBMII isomerase involved in histidine biosynthesis. The fourth gene is *PDF5*, coding for a prefoldin, involved in unfolded protein binding. The fifth gene is *PS2*, coding for a pyrophosphate-specific phosphatase. The lengths of the probes cover the entire range of the MLPA assay (Supplementary Table S1). The regions were selected as not copy-number variable in *Arabidopsis* based on WGS data and were experimentally validated in 189 natural accessions (Zmienko et al., 2016).

Order the Oligonucleotide Synthesis

The synthesis of the designed MLPA probes is typically performed by an external service provider, such as Integrated DNA Technologies (IDT).

Step 8. Order the synthesis of left and right half-probes, each as separate oligonucleotides, at a 100-nmol scale. All right half-probes must be additionally modified at their 5′ ends (5′ phosphorylation).

Caution: 5′ phosphorylation of the right half-probes is essential for a successful ligation step (described below). The oligonucleotides designed for MLPA assays should be of high purity; therefore, we recommend selecting a PAGE or HPLC purification option, depending on the oligonucleotide length and according to the oligonucleotide manufacturer's recommendations.

Step 9. Re-dissolve the lyophilized oligonucleotides upon arrival in deionized water to a concentration of 20 μM. Alternatively, the oligonucleotides can be re-dissolved in 10 mM Tris-HCl, pH 8.2.

Step 10. Store the half-probe stocks at –20°C.

Stage B. Perform MLPA Assay (Time: 2 Days)

Note: When performing the MLPA assay, keep all reagents, stock solutions and working solutions on ice. Set up the reactions in PCR tubes or strips (recommended) at room temperature, unless indicated otherwise. Depending on the user's experience, we recommend running assays for 8–32 samples at once in 1–4 PCR strips.

Note: Whenever applicable, prepare the reagent master mixes for all assayed samples with 10% volume surplus to minimize sample-to-sample variation and save pipetting time. Distribute the master mix to eight tubes of a new PCR strip and then transfer the required amount to all PCR strips containing your samples with a multichannel pipette.

Note: Perform all incubation steps in a thermocycler, programmed as specified in **Table 1**.

Caution: Do not vortex the tubes containing Ligase-65 or Salsa Polymerase enzymes. Likewise, do not vortex the master mixes after adding any of these enzymes.

Prepare the MLPA Probe Set Mix

The correctly composed assay should include both half-probes (left and right) for each region of interest. Each pair of half-probes should generate a ligation product of unique length in the assay. The concentration of the MLPA probes in the final reaction mixture is very low (see below); therefore, it is convenient to perform a two-step oligonucleotide dilution during the probe set mix preparation as follows.

Step 1. Melt all half-probe stocks constituting one assay.

Step 2. Dilute each 20 μM stock with water to a 0.2 μM working solution (200 μl).

Step 3. Mix 2 μl of each half-probe working solution and fill to 400 μl with water.

The resulting 1 nM MLPA Probe Set Mix will contain all the desired pairs of half-probes in equal concentrations and is directly applicable in the reaction setup.

Note: MLPA Probe Set Mix can be stored at –20°C until later use.

Hybridize Half-Probes

For each genomic DNA sample, perform the MLPA assay in a separate tube. We recommend running MLPA assays in multiples of 8 in PCR strips with caps.

TABLE 1 | Programmed thermocycler conditions for multiplex ligation-dependent probe amplification (MLPA) assay.

Program		Action
Denaturation (Step 5)		
98°C, 5 min;		Denature samples.
25°C, ∞;		Cool down samples before removing.
Pause		Proceed to Step 6.
Hybridization (Steps 9-10)		
95°C, 1 min;		Hybridize half-probes to their genomic targets.
60°C, 16–20 h;		
54°C, ∞;		Adjust the temperature for the next step.
Pause		Proceed to Step 11.
Ligation (Step 14)		
54°C, 15 min;		Ligate adjacently hybridized half-probes.
98°C, 5 min;		Inactivate the enzyme.
20°C, ∞;		Cool down samples before removing.
Pause		Proceed to Step 15.
Amplification (Step 18)		
35 cycles of:	95°C, 30 s;	Amplify the correctly ligated MLPA probes.
	60°C, 30 s;	
	72°C, 1 min;	
72°C, 20 min;		Perform final extension of PCR products.
4°C, ∞;		Cool down samples before removing.
End		Proceed to Step 19.

Caution: Replace the strip caps with new ones at each opening during the entire procedure to prevent cross-contamination.

Step 4. Aliquot 5 µl of genomic DNA (0.4 to 50 ng/µl) to individual strip tubes to obtain a final template amount of 2–250 ng per assay, depending on the species.

Note: We recommend performing template optimization assays for each species.

For *Arabidopsis:* We successfully performed MLPA assays using 2, 5, 10, 15, 30, 60, and 100 ng genomic DNA per assay (see the next section).

Step 5. Insert the samples into the thermocycler. Heat for 5 mins at 98°C then let the samples cool to 25°C.

Step 6. Remove the samples from the thermocycler and centrifuge.

Step 7. Prepare master mix I. Briefly vortex and centrifuge the SALSA MLPA buffer and MLPA Probe Set Mix. Prepare the adequate amount of the master mix I by mixing 1.5 µl of SALSA MLPA buffer and 1.5 µl of 1 nM MLPA Probe Set Mix per sample, with 10% volume surplus. Vortex and centrifuge the tube.

Step 8. Add 3 µl of the master mix I to each denatured DNA sample and mix briefly by pipetting. Close the strips with the new caps and centrifuge. The reaction volume in each tube should be 8 µl.

Step 9. Put the samples back into the thermocycler and incubate for 1 min at 95°C, then for 16 to 18 h at 60°C.

Step 10. Adjust the thermoblock temperature to 54°C before proceeding to the next step.

Caution: Do NOT remove the samples from the thermocycler!

Ligate the Hybridized Half-Probes

Step 11. Prepare master mix II without enzyme. Briefly vortex and centrifuge Ligase Buffer A and Ligase Buffer B. Mix 3 µl of Ligase Buffer A, 3 µl of Ligase Buffer B, and 25 µl of nuclease-free water per sample, with 10% volume surplus. Vortex and centrifuge the tube.

Step 12. Centrifuge the tube containing SALSA Ligase-65 enzyme. Add 1 µl of the enzyme per sample with 10% volume surplus to the master mix II. Mix briefly by pipetting. Centrifuge the tube and store on ice until use. Proceed to the next step without delay.

Step 13. Without removing the strips from the thermocycler, add 32 µl of master mix II to each sample. Mix by pipetting and close the strips with new caps. The reaction volume in each tube should be 40 µl.

Step 14. Incubate the samples for 15 min at 54°C, followed by heat inactivation of the ligase enzyme (5 min at 98°C). Cool the thermoblock to 20°C and remove the samples.

Amplify the Ligated MLPA Probes

Step 15. Prepare master mix III. Briefly vortex and centrifuge the SALSA PCR primer mix. Mix 2 µl of SALSA PCR primer mix and 7.5 µl of nuclease-free water per sample, with 10% volume surplus. Vortex and centrifuge the tube.

Step 16. Centrifuge the tube containing SALSA Polymerase enzyme. Heat the tube in hands for approximately 10 s, then add 0.5 µl of the enzyme per sample with 10% volume surplus to master mix III. Mix briefly by pipetting. Centrifuge the tube and store on ice until use.

Step 17. Add 10 µl of master mix III to each sample and mix by pipetting. Close the strips with new caps and replace in the thermocycler. The final reaction volume in each tube should be 50 µl.

Step 18. Perform the PCR comprising 35 cycles of: 95°C for 30 s; 60°C for 30 s and 72°C for 1 min, followed by a 20 min final elongation at 72°C. Cool the thermoblock to 4°C.

Step 19. Store the samples at 4°C, protected from light, until the product size-separation (1–3 days).

Stage C. Collect and Analyze the Data (Time: 1 Day for the Data Collection, Variable for the Analysis)

Size-Separate the PCR Products by Capillary Electrophoresis

The product separation should be performed under denaturing conditions on any standard capillary DNA analyzer. The specific run parameters must be adjusted according to the recommendations of the instrument manufacturer.

We typically use the services of the local Molecular Biology Techniques facility (at the Department of Biology of Adam Mickiewicz University, Poznan, Poland) and separate the samples in ABI Prism 3130XL Genetic Analyzer (Applied Biosystems), using the following procedure.

Step 1. Each MLPA reaction sample is diluted 20× with nuclease-free water, mixed with 9 µl of HiDi formamide (Thermo

Fisher Scientific) containing GeneScan 600 LIZ Size Standard (Thermo Fisher Scientific) and denatured.

Step 2. Samples are injected at 1.2 kV voltage and separated on ABI Prism 3130XL Genetic Analyzer (Applied Biosystems) at 15 kV, in POP7 separation matrix (Thermo Fisher Scientific).

Analyze the Electropherograms

Evaluate the data quality and extract the signal intensity from the electropherograms. Numerous software tools are appropriate for this purpose. Below, we describe the step-by-step analysis performed with GeneMarker (SoftGenetics) (Supplementary Figure S2).

Note: The GeneMarker functions used here are accessible in the limited demo version of the software, freely downloadable from the manufacturer's web site. The details regarding use of these functions are described in the software manual, also available for download.

Step 3. Load the electropherogram data to GeneMarker.

Step 4. Analyze the raw data files with the MLPA analysis type option and appropriate DNA standard selected (depending on the capillary electrophoresis conditions). Select the size call method and data normalization approach (Supplementary Figure S2A).

Note: GeneMarker software provides two normalization options (intra-sample "Internal Control Probe Normalization" and inter-sample "Population Normalization") that aim to correct for the variation in signal intensity caused by the differences in the lengths of the probes in the multiplex assay. We typically use the intra-sample normalization against our control probes, although at this step it is not critical, because the range of the probe lengths in our assay (96–200 nt) is much smaller than in the case of commercial MLPA assays (130–490 nt).

Caution: Use the same parameter settings for all samples. When applying internal control probe normalization, use the same set of control probes for analysis of all samples in the MLPA assay.

Note: At the first analysis of a new MLPA assay, run the analysis for a selection of samples using the "NONE" panel selection. This will allow you to manually create the custom MLPA panel later by indicating the peak positions in your pre-processed samples (see Step 5). If the MLPA panel has already been created, select that panel for the final analysis of all your samples.

Step 5. Perform this step for the new MLPA assay only. Manually create the probe panel with the Panel Editor (Supplementary Figure S2B). Use the pre-processed set of representative MLPA electropherograms (see Step 4) to locate and insert the alleles at the expected positions. Label the alleles with the MLPA probe names. If you want to use the "Internal Control Probe Normalization" option during the analysis, mark the control probes as 1. Repeat Step 4 to re-run all samples using the newly created panel.

Note: In our assays, all peak sizes consistently appeared ∼3 bp shorter than the theoretical length of their attributed MLPA probes. This is not an unexpected result because the migration times of the peak maxima depend on many factors, including the amount of the sample injected, the temperature and the dye used. The capillary electrophoresis systems estimate the relative allele size (using internal standard) and do not necessarily report the true fragment size (McCord, 2003). Therefore, the observed shift is specific to the system and MLPA assay conditions. As long as the peaks are consistently observed at the same positions in all samples under comparison, it does not influence the peak discrimination and subsequent analysis of the MLPA data.

Step 6. Evaluate the quality of individual electropherograms in accordance with the peak pattern of the size standard, the electrophoresis baseline, signal sloping and overall signal intensity. Samples that show abnormalities should be excluded from the analysis.

Step 7. Configure the report layout and copy the results to MS Excel or similar program for further analysis (Supplementary Figure S2C).

Note: The processed data can be reported as the fluorescence intensity (peak height) or the peak area values for each allele. The choice of the output typically does not affect the downstream data analysis and we obtained comparable results with both options. We preferably use the fluorescence intensity data.

Estimate the DNA Copy Number

Step 8. Use the normalization controls to perform within-sample normalization of all your sample data before comparison.

For Arabidopsis: Use at least 3 of the provided control probes (ctrl1–ctrl5) for normalization. Divide each intensity value by the average intensity of the control probes, separately for each sample.

Step 9. For each region analyzed, compare the normalized intensity between the samples. Cluster the samples with the similar intensities and infer the copy numbers from analysis of histograms or two-dimensional plots (see next section). Whenever possible, use the (set of) positive and negative control samples with known copy number status to determine the duplication/deletion intensity thresholds (see the next section for exemplar results).

ANTICIPATED RESULTS

Exemplar MLPA Assay

Based on the available WGS data from 1001 Arabidopsis Genomes Project (1001 Genomes Consortium, 2016) and our own analysis of a subset of this data including 80 accessions, originally described in (Cao et al., 2011), we selected 12 genes that overlapped CNVs with various levels of structural complexity. Genes *AT1G47670* and *AT1G80830* do not present copy number changes. Genes *AT1G32300* and *AT4G19520* are biallelic; more specifically, they display presence-absence variation. The remaining eight genes are multiallelic and present duplications (*AT4G27080*, *AT5G09590*, and *AT5G61700*) or duplications and deletions (*AT1G27570*, *AT1G52950*, *AT3G21960*, *AT4G27080*, and *AT5G54710*). Additionally, gene *AT5G09590* overlaps CNV only partially, whereas *AT1G52950*, *AT5G54710*, and *AT1G27570* are members of multigene families and are localized in the regions of high structural diversity (manifested e.g., by the presence of adjacent or overlapping CNVs, presence of nearby transposable element genes or the presence of clusters of highly similar paralogs). To present the performance of the MLPA approach we set up a multiplex assay Ath.test for these genes (**Table 2**). We evaluated the genes' copy number status in 80 *Arabidopsis* accessions, characterized in the first stage of 1001 *Arabidopsis* Genomes Project (Cao et al., 2011). All seeds were obtained from The European *Arabidopsis* Stock Centre[4] and grown as described previously (Zmienko et al., 2016).

Optimization of the Template Amount

The multiplex MLPA-based strategy presented in this paper was originally developed for CNV genotyping of human DNA (Kozlowski et al., 2007; Marcinkowska et al., 2010). To adjust it for use with the *Arabidopsis* genome, we aimed to optimize the amount of DNA template. For humans, the typical MLPA assays include 50-250 ng genomic DNA per reaction. In our previous study, we successfully performed MLPA-based copy number analysis using 100 ng *Arabidopsis* genomic DNA (Zmienko et al., 2016). However, because the *Arabidopsis* genome is ∼20 times smaller than the human genome, we expected that the template amount could be substantially reduced without affecting the reaction performance. To evaluate the acceptable range of DNA amount for this species, we used the Col-0 accession, performed serial dilutions of the DNA template and performed MLPA assays for each of the following DNA amounts: 100, 60, 30, 15, 10, 5, and 2 ng, in three replicates. We observed that the intensity data showed little variance across all DNA concentrations tested and the peaks showed very good resolution and similar distribution, regardless of the template amount (**Figures 2A–C**; Supplementary Data Sheet S1). The normalized signal intensity data for various template amounts were highly correlated, with the results calculated for 2 ng DNA input showing only slightly lowered correlation than the other amounts (**Figure 2D**). From this comparison, we concluded that the whole range of tested DNA amounts generates valid data. Below, we used the smallest tested amount of DNA (2 ng) to perform the exemplar Ath.test MLPA assay.

Gene Copy Number Analysis

We generated MLPA data, processed it in GeneMarker and exported it to a Microsoft Excel worksheet (Supplementary Data Sheet S1). Three samples were excluded at this stage due to poor data quality. To enable sample-to-sample comparison, we normalized the data within each sample using the mean signal intensity of the control probes ctrl1–ctrl5. The data were then compared and the copy numbers were estimated relative to the Col-0 accession that has the basic copy number of each gene analyzed in this assay ($2n = 2$) and therefore served as the reference sample. To reveal groups of accessions with distinct gene copy numbers, the population data were displayed as dot plots, histograms of the signal intensities or (for genes targeted by two MLPA probes) as 2D plots. We set the duplication/deletion thresholds at <0.7 and >1.3 of the relative intensity, respectively, for all genes in the assay. Subsequently, for each gene, the samples passing the threshold values were clustered and the clusters were manually assigned the copy numbers, as demonstrated previously (Marcinkowska-Swojak et al., 2014; Zmienko et al., 2016).

Non-variable Regions

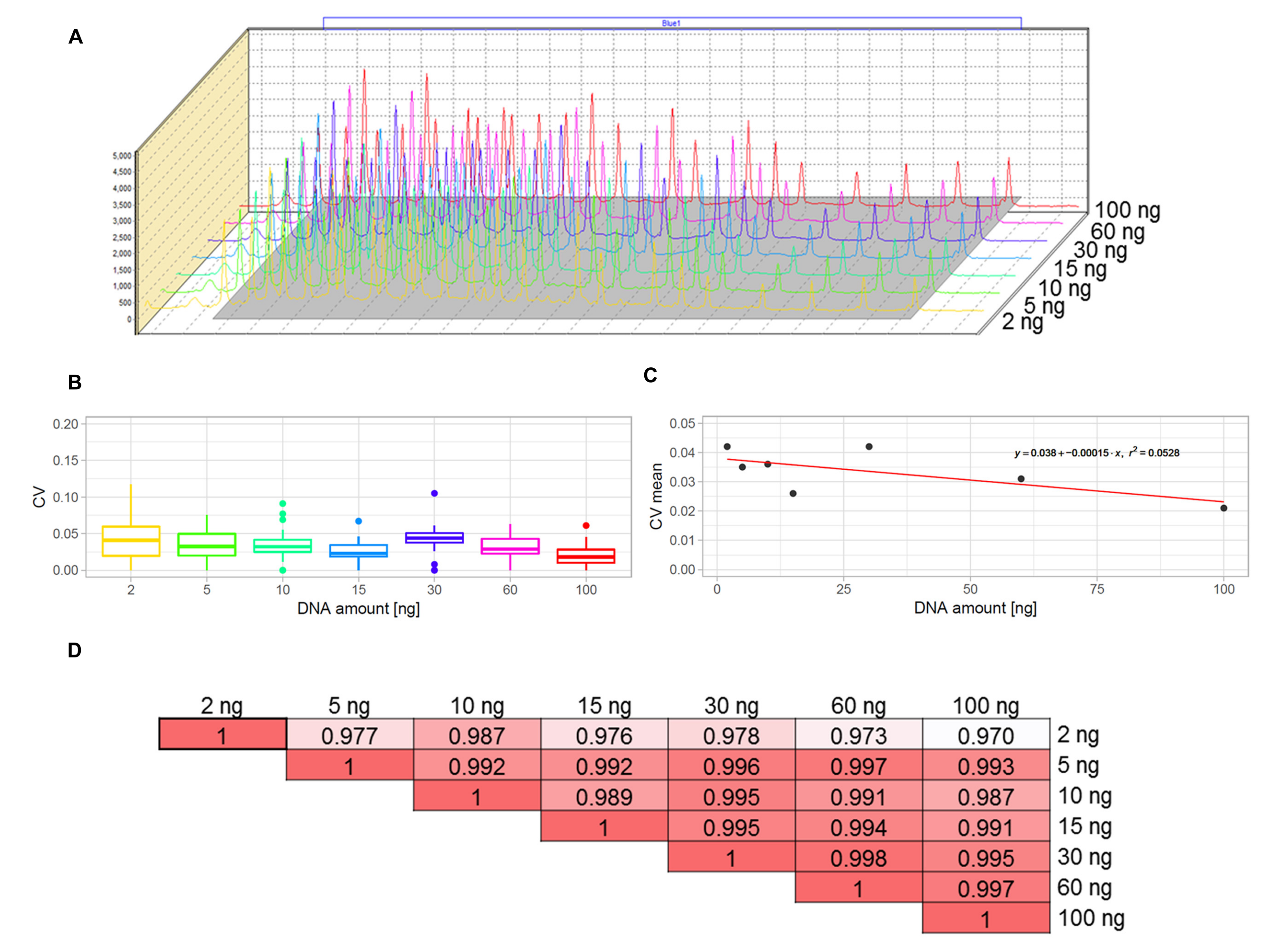
The probes mlpaA, mlpaB1, and mlpaB2 targeted two genes predicted to have the same copy number in all accessions: *AT1G47670*, coding for lysine histidine transporter-like 8 (mlpaA), and *AT1G80830*, coding for NRAMP1 transporter (mlpaB1 and mlpaB2). For all accessions, the relative signals

TABLE 2 | The probe composition and gene targets of Ath.test assay.

Probe name	Probe length	Target genomic site	Locus ID	Predicted CNV status	Source*
ctrl1	96 nt	Chr1:25593..25645	AT1G01040	Non-variable; normalization control	a
ctrl2	111 nt	Chr4:11476533..11476582	AT4G21580	Non-variable; normalization control	a
ctrl3	124 nt	Chr2:15194440..15194490	AT2G36230	Non-variable; normalization control	a
ctrl4	144 nt	Chr5:7847361..7847414	AT5G23290	Non-variable; normalization control	a
ctrl5	172 nt	Chr1:27465468..27465522	AT1G73010	Non-variable; normalization control	a
mlpaA	160 nt	Chr1:17539289..17539343	AT1G47670	Non-variable	b; c
mlpaB1; mlpaB2	90 nt 148 nt	Chr1:30374276..30374321 Chr1:30373647..30373699	AT1G80830	Non-variable	b; c
mlpaC	93 nt	Chr1:11651708..11651754	AT1G32300	Biallelic	b
mlpaD1; mlpaD2	105 nt 114 nt	Chr1:9575624..9575678 Chr1:9577003..9577055	AT1G27570	Multiallelic	b; c
mlpaE1; mlpaE2	136 nt 196 nt	Chr1:19726669..19726721 Chr1:19727385..19727439	AT1G52950	Multiallelic	b; c
mlpaF1; mlpaF2	99 nt 120 nt	Chr3:7737420..7737467 Chr3:7737872..7737929	AT3G21960	Multiallelic	b; c
mlpaG1; mlpaG2	128 nt 164 nt	Chr4:10641616..10641668 Chr4:10644628..10644679	AT4G19520	Biallelic	c
mlpaH	180 nt	Chr4:13592606..13592658	AT4G27080	Multiallelic	b; c
mlpaI	117 nt	Chr4:17705274..17705327	AT4G37685	Multiallelic	b
mlpaJ1; mlpaJ2	108 nt 156 nt	Chr5:2976409..2976464 Chr5:2978013..2978065	AT5G09590	Multiallelic; part of the gene	c
mlpaK1; mlpaK2	188 nt 102 nt	Chr5:22228424..22228479 Chr5:22229438..22229488	AT5G54710	Multiallelic	b; c
mlpaL	132 nt	Chr5:24796111..24796161	AT5G61700	Multiallelic	c

**The initial information about the gene CNV status comes from the following resources: a, Zmienko et al. (2016); b, Arabidopsis 1001 Genomes Project; c, our unpublished analysis of the WGS data originally presented in Cao et al. (2011).*

[4] http://arabidopsis.info/



	2 ng	5 ng	10 ng	15 ng	30 ng	60 ng	100 ng
2 ng	1	0.977	0.987	0.976	0.978	0.973	0.970
5 ng		1	0.992	0.992	0.996	0.997	0.993
10 ng			1	0.989	0.995	0.991	0.987
15 ng				1	0.995	0.994	0.991
30 ng					1	0.998	0.995
60 ng						1	0.997
100 ng							1

FIGURE 2 | Optimization of the DNA template amount for copy number variant (CNV) genotyping in *Arabidopsis*. The data were obtained from three replicates of MLPA assay per each DNA amount tested: 2, 5, 10, 15, 30, 60, and 100 ng. **(A)** comparison of the peak heights and localization on the electropherograms. **(B,C)** Variance of probe intensity measurements across tested DNA concentrations. Boxplots **(B)** present the distribution of coefficients of variation (CV) calculated separately for each probe and each DNA amount; The linear plot **(C)** presents regression analysis (CV vs input amount). For visibility, mean CV values (all probes, each in three replicates) per input amount are displayed on the plot; **(D)** pairwise correlation of results obtained for all template amounts tested, presented as coefficients of determination (R^2) of the MLPA probe intensity data.

from these three probes were at the same level as those in Col-0 (mean intensity 1.01, 1.03, and 0.93, respectively, see **Figure 3A**) and showed very little variance (CV 0.060, 0.089, and 0.064, respectively). Additional evaluation of the mlpaB1 and mlpaB2 probes on a 2D plot revealed that all samples were grouped in one cluster (**Figure 3B**).

Biallelic CNVs

We analyzed two genes with presence-absence variation revealed by the WGS data analysis: *AT1G32300* (coding for D-arabinono-1,4-lactone oxidase family protein) and *AT4G19520* (coding for TIR-NBS-LRR class disease resistance protein). We designed one probe (mlpaC) for *AT1G32300* exon 1 and two probes, mlpaG1 and mlpaG2, for *AT4G19520* exons 3 and 5, respectively. For *AT1G32300*, we observed a dominant population of samples with mean signal intensity 1.08, indicative of two gene copies per diploid genome. The remaining samples formed a distinct group with mean signal intensity 0.09, indicative of the absence of the analyzed gene in the respective accessions (**Figure 4A**). In the case of *AT4G19520*, the combined data for the mlpaG1 and mlpaG2 probes revealed the presence of two compact clusters (**Figure 4B**). One cluster included 29 accessions with no difference in copy number relative to Col-0 (mlpaG1 mean intensity 1.03; mlpaG2 mean intensity 1.01). The other cluster included 47 accessions with substantially reduced intensity (mlpaG1 mean intensity 0.14; mlpaG2 mean intensity 0.12), indicative of the deletion.

Multiallelic CNVs: One MLPA Probe Per Gene

For three genes that overlap multiallelic CNVs we designed 1 MLPA probe per gene in Ath.test assay (**Figure 5A**). Gene *AT4G37685* codes for a hypothetical protein and is targeted by the mlpaI probe. Majority of accessions (39) harbor two copies of this gene. Gene deletion was detected in eight accessions and

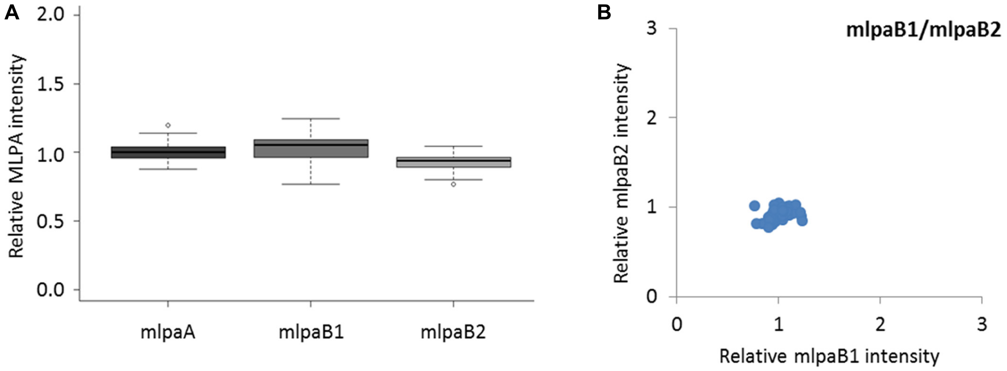


FIGURE 3 | Multiplex ligation-dependent probe amplification assay results for non-variable regions. (A) relative intensity distribution of the mlpaA probe, targeting the *AT1G47670* gene as well as mlpaB1 and mlpaB2 probes, both targeting the *AT1G80830* gene, among the accessions; **(B)** 2D plot of intensity from probes mlpaB1 and mlpaB2.

duplication in 30 accessions. Of the latter, 22 accessions had four copies, seven accessions had six copies, and one harbored a very high-level duplication, most likely ≥12 copies.

Gene *AT5G61700* codes for ATH16, a member of ABC transporter subfamily A and is targeted by probe mlpaL. In most analyzed accessions, the gene exists in two copies per diploid genome. In eight accessions, however, duplications were detected: four copies in three accessions, six copies in two accessions, and ≥10 copies in three accessions. It is worth noting that, in MLPA assays, the signal intensity is non-linearly related to the DNA copy number (Zmienko et al., 2016). This is manifested by reducing the distance between the clusters with different duplication levels for high copy numbers. Consequently, a large number of samples harboring high-level duplications is needed to precisely distinguish the clusters of 8 and more copies from each other.

Gene *AT4G27080* codes for a protein disulfide isomerase that is involved in cell redox homeostasis and is targeted by the mlpaH probe. From the WGS data, we predicted that majority of accessions harbor partial or full duplications of this gene. Likewise, MLPA analysis revealed that only nine accessions harbor two copies of *AT4G27080* gene, while duplications were detected in 68 accessions. Among them, we clearly identified a group of 44 accessions with four copies, but the remaining accessions were less distinctive and formed two heterogeneous groups which we named “medium-level duplications” (10 accessions) and “high-level duplications” (14 accessions). For 12 of these “high-level duplication” accessions, the mlpaH peak intensity counts reached the upper detection limits (see **Notes** section below for additional comments). We concluded that designing two or more MLPA probes targeting this genomic region and repeating the assay with adjusted capillary electrophoresis parameters would be helpful in more accurate distinction of the CNV genotypes or resolution of the structural complexity of the investigated gene.

Multiallelic CNVs: Two MLPA Probes Per Gene

For 2 other genes that overlap multiallelic CNVs we designed two MLPA probes per gene (**Figure 5B**). The *AT5G54710* gene codes for an ankyrin repeat family protein and is positioned between two other ankyrin repeat family protein coding genes, in the region that is highly copy number variable. We used two specific probes (mlpaK1 and mlpaK2), located in the fourth and third exons of *AT5G54710*, respectively, and confirmed that this gene is multiallelic. The high linear correlation of the mlpaK1 and mlpaK2 probe intensities allowed us distinguish several clusters of accessions with distinct copy numbers: 0 copies (2 accessions), 2 copies (54 accessions), 4 copies (8 accessions), 6 copies (6 accessions), and 8 copies (1 accession). We did not assign the integer copy numbers for 6 accessions which displayed uneven duplication level based on the mlpaK1 and mlpaK2 probe signal.

The *AT5G09590* gene, encoding mitochondrial heat shock protein MTHSC70-2, is localized in the breakpoint of a large CNV that encompasses loci *AT5G09590* – *AT5G09630*. Consequently, *AT5G09590* is only partially duplicated in several accessions. We designed two probes, localized outside of and within the CNV region (mlpaJ1, targeting fourth exon and mlpaJ2, targeting sixth exon, respectively). The results of the MLPA assay clearly revealed that only the 3′ part of *AT5G09590* (targeted by probe mlpaJ2) is duplicated: 43 accessions harbored four copies, two accessions harbored six copies, and one accession harbored at least 10 copies. The region targeted by probe mlpaJ1 invariantly had two copies in all accessions.

Complex Multiallelic CNVs

Some genomic regions, e.g., these that harbor clustered multigene families, may display high structural diversity in the populations. A gene may be fully duplicated/deleted in some accessions while in the other ones only part of this gene may display copy number alteration. Additionally, the duplicated DNA copies within one sample may differ from each other in length and sequence, which may affect the affinity of the MLPA probe to some (but not all) copies. Consequently, the copy number pattern revealed by the MLPA analysis may be complex. Below we present some examples of MLPA analysis in multiallelic CNVs with a complex structure (**Figure 5C**).

The *AT3G21960* gene is localized in the central part of a ~50 kb CNV, that encompasses 21 genes, mainly members of

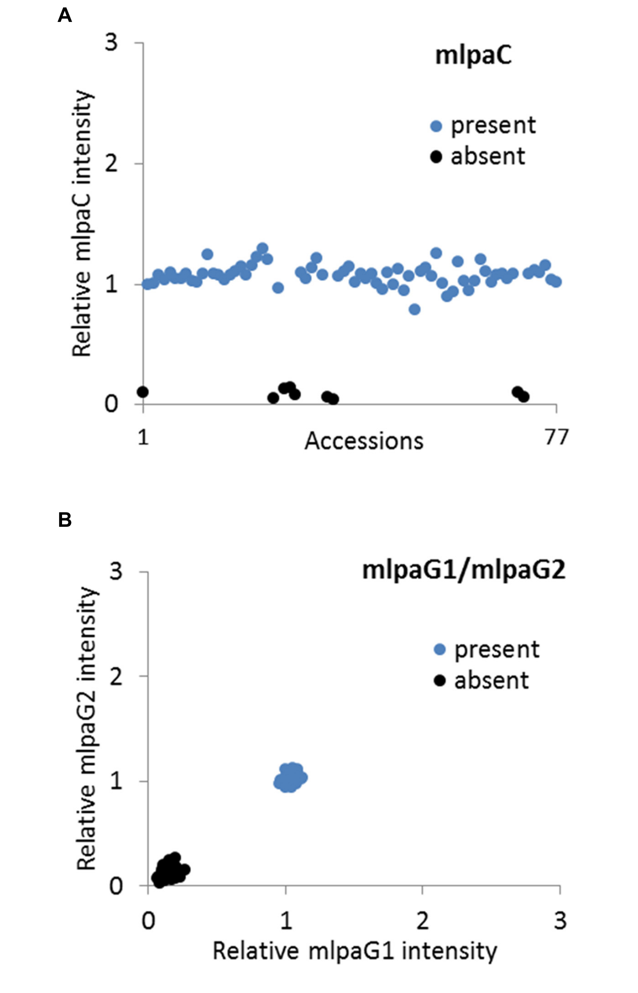


FIGURE 4 | Multiplex ligation-dependent probe amplification assay results for presence-absence CNVs. (A) relative intensity from probe mlpaC targeting the *AT1G32300* gene, in individual accessions; **(B)** 2D plot of relative signal from mlpaG1 and mlpaG2, both targeting the *AT4G19520* gene. Clusters are colored according to the deduced CNV status.

the receptor-like protein kinase-related family and genes coding for proteins with unknown domain DUF26. We assayed the *AT3G21960* gene with specific probes targeting exons 1 and 2 (probes mlpaF1 and mlpaF2, respectively). In 30 samples the signals from these probes were highly correlated and formed 4 distinct groups of: 0 copies (1 accession), 2 copies (26 accessions), 4 copies (1 accession) and 6 copies (2 accessions). In 6 accessions, however, only the mlpaF2 probe intensity was elevated (1.83–6.54), while mlpaF1 intensity was about 1. On the contrary, the remaining 41 accessions formed a compact cluster, with the mlpaF1 intensity below 0.7 (the value that has been set as the deletion threshold), and the mlpaF2 intensity about 1. A brief evaluation of the *AT3G21960* genomic sequence inferred from WGS data[5] (obtained with Pseudogenomes Download Tool) provided evidence that this complex pattern is true, as 519 out of 1135 accessions with available genomic data had 80–100% uncalled sites (Ns) in the exon 1 sequence, while only 3 accessions had 80–100% uncalled sites in exon 2 sequence.

Complex multiallelic CNVs are often related to the activity of mobile genetic elements, which may trigger partial or full deletion/duplication of the nearby genes. Gene *AT1G52950* codes for a nucleic acid-binding OB fold-like protein and is localized within one CNV region with a nearby transposable element gene *AT1G52960* (the two loci are separated by only 3.6 kb distance). We assayed the copy number status of *AT1G52950* using two probes, mlpaE1 to target exon 6 and mlpaE2 target exon 9. For 69 accessions, we detected compact clusters with distinct copy numbers (0 to 6 copies) and a high correlation between the two measurements ($R^2 = 0.9881$). Interestingly, in two cases, the intensity data suggested the existence of one copy and three copies of the *AT1G52950* gene per diploid genome in the surveyed individuals. *Arabidopsis* is a highly self-pollinating species for which most genomic loci are expected to exist in a homozygous state, therefore assaying additional individuals would be necessary to establish the representative gene copy number for these two accessions in a population study. For seven accessions, of which six originated from Southern Tyrol region and 1 was a Spanish relict accession (1001 Genomes Consortium, 2016), the copy number status indicated by probe mlpaE1 was always higher than the copy number status indicated by probe mlpaE2. This effect may have many reasons, e.g., partial duplication or deletion of a gene of interest, sequence divergence in some duplicated copies that affect the hybridization of one MLPA probe, etc. Unambiguous interpretation of these data would require additional region characterization by sequencing. Nevertheless, the signals from both probes were also well correlated ($R^2 = 0.9856$). Finally, one accession displayed an extremely high level of duplications at the mlpaE1 target site while no copy number changes were observed at the mlpaE2 site.

Effect of Non-specific Hybridization on MLPA Signal

To present the effect of compromised probe specificity on the MLPA results, we assayed a gene *AT1G27570*, which encodes the phosphatidylinositol 3- and 4-kinase family protein and is localized within the large multiallelic CNV (over 20 kb). We designed two probes, mlpaD1 and mlpaD2, targeting this gene, of which only mlpaD2 was specific to *AT1G27570*. Probe mlpaD1 had an alternative target site (with only two mismatches in the left TSS and one mismatch in the right TSS, distant from the ligation site) in the nearby gene *AT1G27590*, not copy number variable. As a result, the signal from the mlpaD1 probe was elevated by the background signal from the alternative target site. This background signal was stable (due to unchanged copy number of *AT1G27590* gene in all accessions) therefore the high correlation between the data for mlpaD1 and mlpaD2 probes was preserved (**Figure 6A**). As a rule, we suggest re-designing of the MLPA probes that produce non-specific signal. However, if a set of the control samples that carry confirmed deletion of the gene of interest can be defined, these samples may be used for the data correction. In the present example, we calculated the mean non-specific signal of probe mlpaD1 in the cluster of 15 samples with gene deletions (marked in black color in

[5] http://1001genomes.org

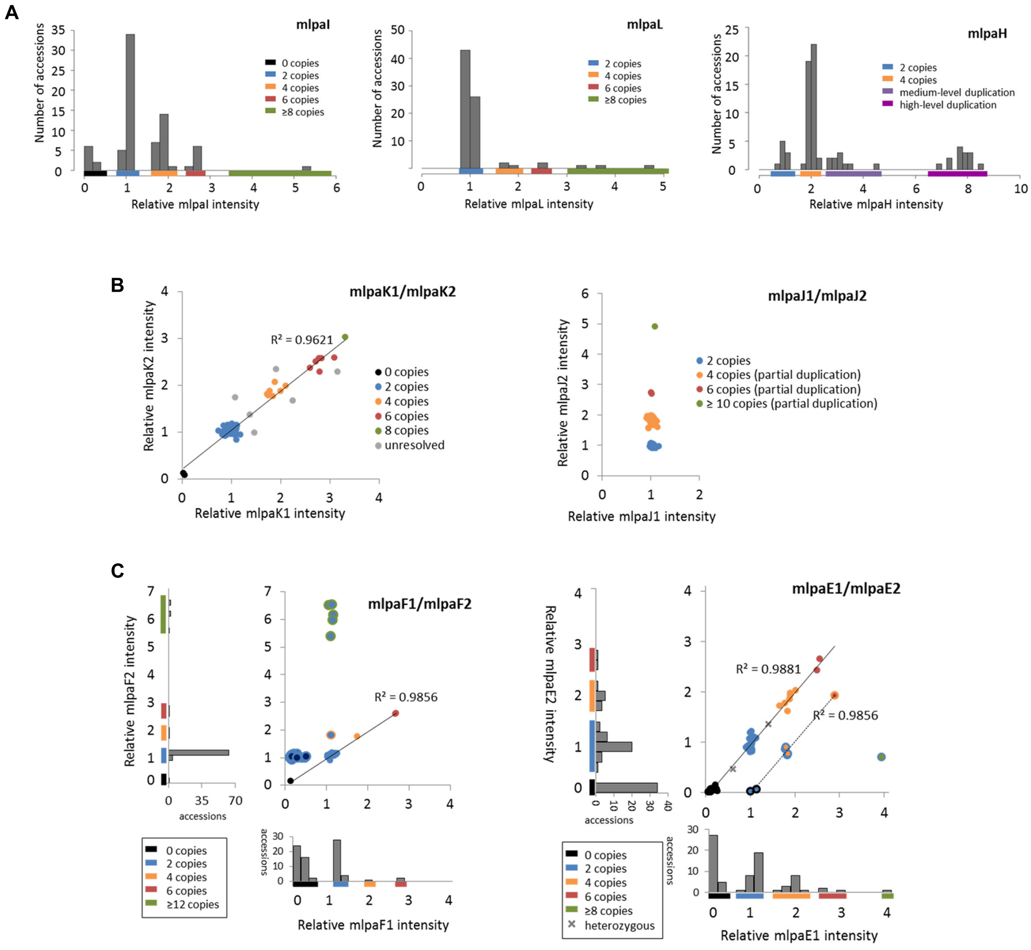


FIGURE 5 | Multiplex ligation-dependent probe amplification results for multiallelic CNVs. (A) CNV genotyping with one MLPA probe per gene. Histograms present the relative signal distribution from probe mlpaI (targeting the *AT4G37685* gene), probe mlpaL (targeting the *AT5G61700* gene), and probe mlpaH (targeting the *AT4G27080* gene). The histogram bin size is 0.2 in all plots; **(B)** CNV genotyping with two MLPA probes per gene. 2D plots present the relative signal from probes mlpaK1and mlpaK2 (both targeting the *AT5G54710* gene) and from probes mlpaJ1 and mlpaJ2 (both targeting the *AT5G09590* gene). Clusters are colored according to deduced CNV status. The coefficient of determination (R^2) is calculated for accessions with assigned copy numbers. **(C)** Genotyping complex multiallelic CNVs. 2D intensity plots present relative signal from probes mlpaF1and mlpaF2 (targeting exon 1 and exon 2 of the *AT3G21960* gene, respectively) and from probes mlpaE1and mlpaE2 (targeting exon 6 and exon 9 of the *AT1G52950* gene, respectively). Clusters are colored according to deduced CNV status. The coefficient of determination (R^2) is calculated for subsets of accessions, as detailed in the main text.

Figure 6A). This value was then subtracted from the probe mlpaD1 signal in each sample, before estimating the intensity ratio relative to Col-0 accession. The correction improved the relative intensity ratio observed for probe mlpaD1 (**Figure 6B**). We note here, that the process of data correction had no effect on the overall correlation between the signals from probes mlpaD1 and mlpaD2. This correlation was high ($R^2 = 0.9386$), therefore allowing to distinguish the copy number clusters on 2D plots pretty easily both before and after data correction.

NOTES

Below we included some notes on the limitations of the procedure, common mistakes and possible artifacts related to the presented application.

Probe Design

Oligonucleotide MLPA probes described in this procedure target specific sequences in the genome, typically 45–75 bp. Regions located outside of the probe's recognition sequence may have different copy number status. If partial gene duplication/deletion or insertion of duplicated sequence is suspected, additional probes, e.g., covering different exons of the gene should be included in the assay.

Compromised ability of MLPA probe to recognize the target sequence may be the source of false positive results. Sequence changes (SNPs, indels, point mutations) in the target sequence detected by a probe can negatively affect or completely prevent probe binding. The critical positions in the TSS sequence are these constituting the ligation site; the presence of a SNP at or near the ligation site will disrupt the ligation step and result in

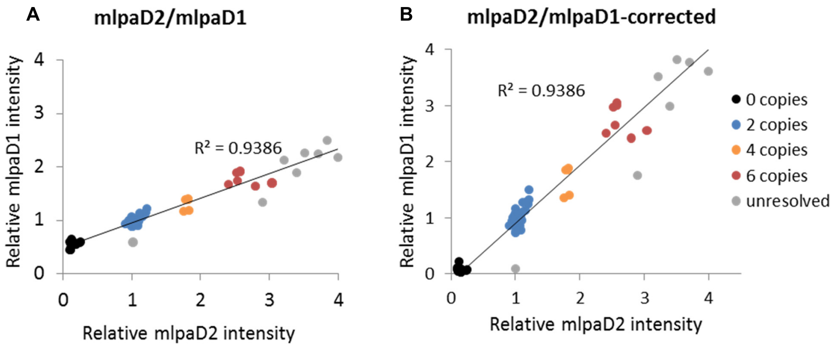


FIGURE 6 | Effect of non-specific probe hybridization on the MLPA results. Probes mlpaD1 and mlpaD2 target the *AT1G27570* gene. Additionally, probe mlpaD1 targets also the *AT1G27590* gene. **(A)** 2D intensity plot of relative signal from probes mlpaD1 and mlpaD2; **(B)** 2D intensity plot of relative signal from probe mlpaD1 corrected for the presence of non-specific hybridization signal (see main text for details) and probe mlpaD2. Clusters are colored according to the deduced CNV status and are identical for each sample on both plots. The coefficient of determination (R^2) is calculated for all accessions. The deletion of *AT1G27570* gene has been confirmed by PCR in accessions from the "0 copies" cluster (not shown).

no signal from the MLPA probe, falsely indicative of deletion of the region in the affected sample (Kim et al., 2016). Note that the MLPA technique can be also used for detecting small mutations (Marcinkowska-Swojak et al., 2016), but these applications are not covered in the present protocol.

The accuracy of the results is also strictly dependent on the MLPA probe specificity. If alternative target site exists in the genome (e.g., in a paralogue or a pseudogene), it will generate non-specific signal (see Effect of Non-specific Probe Hybridization on the MLPA Results Section). To this end, for plants with incomplete genome information we strongly advise designing ≥2 MLPA probes per gene, to minimize this risk.

In the case of newly designed MLPA probes we recommend verifying their performance on a (set of) well characterized reference samples. If no product is observed, make sure that the common mistakes interfering with the experimental steps are avoided (see below). If needed, re-design the MLPA probe.

Assay Design and Performing

Multiplex ligation-dependent probe amplification results may be compromised by multiple factors that will affect the enzymatic reactions and result in reduced peak signals. These factors include but are not limited to: DNA integrity and contamination, presence of PCR inhibitors in the samples, incomplete DNA denaturation, sample evaporation, suboptimal amount of the sample DNA used. In the Section "Stepwise Procedures" we included useful tips regarding the sample preparation and assay setup. Additional comments are given below.

If the DNA sample contamination is a suspected problem, perform new DNA extraction. From our experience, we advise using column-based methods, e.g., DNeasy Plant Mini Kit (Qiagen) for DNA extraction (or purification of DNA extracted with other methods) because they produce samples of high purity and comparable amounts.

Use multichannel pipettes to reduce the pipetting time and avoid sample evaporation.

Reduce sample-to-sample variability by simultaneous performing multiple assays, using strips (preferable) or multiwell PCR plates. Use the same MLPA Probe Set Mix preparation for all samples under comparison.

Replacing the strip caps on each opening minimizes the risk of sample cross-contamination.

Follow the capillary electrophoresis protocols (size standard, sample preparation, injection time and voltage) suitable for the instrument used. Decrease injection time if the peaks are out of range. We recommend prior optimization of the DNA template amount in the assay and capillary electrophoresis conditions on a validated reference sample.

Abnormal pictures after capillary electrophoresis may indicate capillary electrophoresis problems but they also may result from the PCR step troubles. See the MLPA troubleshooting wizard by MRC Holland[6] for common peak pattern problems and possible solutions.

Data Analysis and Copy Number Estimation

It is advisable to manually check the peaks identified by GeneMarker before further data processing. In our assay, we repeatedly observed that the software did not detect the peaks for probe mlpaH in 12 samples and reported "0" intensity for this probe (Supplementary Figure S3). In fact, high intensity peaks from probe mlpaH with their tops flattened (cut) were present in these samples, which indicated that the signal exceeded the capillary electrophoresis system detection limits. We manually corrected the peak localization and used the maximum reported values for copy number calculation, but this likely resulted in underestimation of the gene copy number in these samples in our study (see Section "Multiallelic CNVs: One MLPA Probe Per Gene"). To accurately quantify the probe signal, repeating the electrophoresis with lower injection time would be necessary. The results from high and low injection time electropherograms may be then merged after internal control probe normalization step, to preserve good resolution of the low intensity peaks.

[6] http://www.mlpa.com/elearning/tswizard/

Multiplex ligation-dependent probe amplification is a relative technique, therefore selecting well validated reference samples with basic copy number of the region of interest (usually two copies) is essential for accurate quantification. However, in case of population scale CNV genotyping of numerous independent genomic regions in a multiplex assay (similar to example provided in this paper) such a reference sample may not exist or remains unknown. Providing that sufficiently large number of samples in the population are genotyped, the presented protocol still allows for inferring the cluster copy numbers without a reference sample, under the assumption that the neighboring clusters of accessions/lines differ by two copies and that the distances between these clusters are ∼equal in the range of 0–4 copies (see Zmienko et al., 2016 for further discussion on the distances between the clusters in MLPA assays).

Validation of the Results

Regardless of the number of probes and samples used, we recommend to verify the positive MLPA results with an independent technique. We advise performing droplet digital PCR (ddPCR) on selected samples, as this approach allows for estimating gene copy numbers at the same or even higher range, as the MLPA procedure described in this protocol (Zmienko et al., 2016). Additionally, ddPCR generates amplicons of ∼60–200 bp, therefore allows for genome assaying at similar resolution as MLPA.

CONCLUSION

In this work, we described the protocol for the simple MLPA-based CNV genotyping in plants, with particular emphasis on the model plant *Arabidopsis*. We provided a description of the probe design process, experimental setup, and data analysis. We also discussed the results of the exemplar multiplex assay and showed that the MLPA method is very robust and is a rich source of information regarding the CNV in the analyzed samples. The abundant genomic data obtained for a growing number of species as a part of large-scale sequencing projects, highlight CNV as the major contributor to natural diversity at a genotype level (Zarrei et al., 2015; 1001 Genomes Consortium, 2016; Bai et al., 2016). Gene duplication has been considered the major factor driving long-term evolution and gene birth by sub- and neofunctionalization of the duplicated copies (Conant et al., 2014). Some regions in the genome may be more prone to CNV than the others, due to their specific structural features, that will locally induce the mechanisms leading to CNV formation, e.g., non-allelic recombination (Zmienko et al., 2016). The duplication / deletion events may have also consequences on organism's fitness and contribute to the adaptation to environmental challenges, as well as to coevolutionary interactions between host and pathogen or a symbiont (reviewed in: Kondrashov, 2012, Żmieńko et al., 2014). Remarkably, the protein coding genes displaying CNVs are often related to environmental stress response and pathogen resistance (Cook et al., 2012; Maron et al., 2013). The creation of high-confidence CNV maps and assessing the gene copy number in large populations will enhance the studies on the evolution of genomes in the context of CNV origin, fixation and the impact on the phenotype. These data can be later combined with the results of the transcriptomic, proteomic, metabolomics, protein interaction, phenotyping, and other studies). We recently used the MLPA method to genotype *MSH2*, *AT3G18530*, and *AT3G18535* copy number in a set of 189 natural accessions. Based on these results, we were subsequently able to reveal the recurrent nature of *AT3G18530* and *AT3G18535* duplications/deletions and to dissect the structural features that promoted non-allelic homologous recombination, leading to a widespread occurrence of the *AT3G18530* and *AT3G18535* genes deletion in nature (Zmienko et al., 2016).

This protocol will enable potential users to introduce the MLPA technique in plant genetic and population biology studies. The technique is multiplexable and very well suited for verification of WGS-based analyses or for rapid characterization of copy number status across a region of interest in large populations. Notably, once designed, the individual MLPA probes may be used in various combinations according to one's needs, providing that the lengths of the probes in one assay are unique. We believe that the MLPA protocol presented in the current work will contribute to accelerating the discovery of new associations between CNV and important traits in plants.

AUTHOR CONTRIBUTIONS

AS-C prepared DNA samples, performed MLPA assays, analyzed data, helped prepare figures, and draft the manuscript. MM-Z performed template optimization experiments. MM-S helped design the MLPA probes and set up the assay. PK analyzed the data and helped draft the manuscript. MF contributed to the conception of the work and revised the manuscript. AZ conceived of and designed the study, analyzed data, oversaw the research, prepared figures, and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Centre of Science grants (2011/01/B/NZ2/04816 and 2014/13/B/NZ2/03837). Publication of the results was supported by Polish Ministry of Science and Higher Education, under the KNOW program.

ACKNOWLEDGMENT

We thank Pawel Wojciechowski and Kaja Milanowska for help with WGS data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.00222/full#supplementary-material

REFERENCES

1001 Genomes Consortium (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 1–11. doi: 10.1016/j.cell.2016.05.063

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376. doi: 10.1038/nrg2958

Bai, Z., Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., et al. (2016). The impact and origin of copy number variations in the *Oryza* species. *BMC Genomics* 17:261. doi: 10.1186/s12864-016-2589-2

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B., and Rafalski, A. (2010). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor. Appl. Genet.* 120, 355–367. doi: 10.1007/s00122-009-1128-9

Bharuthram, A., Paximadis, M., Picton, A. C. P., and Tiemessen, C. T. (2014). Comparison of a quantitative Real-Time PCR assay and droplet digital PCR for copy number analysis of the CCL4L genes. *Infect. Genet. Evol.* 25, 28–35. doi: 10.1016/j.meegid.2014.03.028

Cantsilieris, S., Baird, P. N., and White, S. J. (2013). Molecular methods for genotyping complex copy number polymorphisms. *Genomics* 101, 86–93. doi: 10.1016/j.ygeno.2012.10.004

Cantsilieris, S., Western, P. S., Baird, P. N., and White, S. J. (2014). Technical considerations for genotyping multi-allelic copy number variation (CNV), in regions of segmental duplication. *BMC Genomics* 15:329. doi: 10.1186/1471-2164-15-329

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* 43, 956–963. doi: 10.1038/ng.911

Ceulemans, S., van der Ven, K., and Del-Favero, J. (2012). "Targeted screening and validation of copy number variations," in *Genomic Structural Variants: Methods and Protocols, Methods in Molecular Biology*, ed. L. Feuk (Berlin: Springer Science+Business Media), 369–384. doi: 10.1007/978-1-61779-507-7_18

Chang, C., Lu, J., Zhang, H.-P., Ma, C.-X., and Sun, G. (2015). Copy number variation of cytokinin oxidase gene Tackx4 associated with grain weight and chlorophyll content of flag leaf in common wheat. *PLoS ONE* 10:e0145970. doi: 10.1371/journal.pone.0145970

Conant, G. C., Birchler, J. A., and Pires, J. C. (2014). Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* 19, 91–98. doi: 10.1016/j.pbi.2014.05.008

Cook, D. E., Bayless, A. M., Wang, K., Guo, X., Song, Q., Jiang, J., et al. (2014). Distinct copy number, coding sequence, and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. *Plant Physiol.* 165, 630–647. doi: 10.1104/pp.114.235952

Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., et al. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science* 338, 1206–1209. doi: 10.1126/science.1228746

Duitama, J., Silva, A., Sanabria, Y., Cruz, D. F., Quintero, C., Ballen, C., et al. (2015). Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. *PLoS ONE* 10:e0124617. doi: 10.1371/journal.pone.0124617

Gaines, T. A., Zhang, W., Wang, D., Bukun, B., Chisholm, S. T., Shaner, D. L., et al. (2010). Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1029–1034. doi: 10.1073/pnas.0906649107

Hanada, K., Sawada, Y., Kuromori, T., Klausnitzer, R., Saito, K., Toyoda, T., et al. (2011). Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. *Mol. Biol. Evol.* 28, 377–382. doi: 10.1093/molbev/msq204

Hömig-Hölzel, C., and Savola, S. (2012). Multiplex ligation-dependent probe amplification (MLPA) in tumor diagnostics and prognostics. *Diagn. Mol. Pathol.* 21, 189–206. doi: 10.1097/PDM.0b013e3182595516

Kim, M. J., Cho, S. I., Chae, J. H., Lim, B. C., Lee, J. S., Lee, S. J., et al. (2016). Pitfalls of multiple ligation-dependent probe amplifications in detecting DMD exon deletions or duplications. *J. Mol. Diagn.* 18, 253–259. doi: 10.1016/j.jmoldx.2015.11.002

Klonowska, K., Ratajska, M., Czubak, K., Kuzniacka, A., Brozek, I., Koczkowska, M., et al. (2015). Analysis of large mutations in BARD1 in patients with breast and/or ovarian cancer: the Polish population as an example. *Sci. Rep.* 5:10424. doi: 10.1038/srep10424

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and censor. *BMC Bioinformatics* 7:474. doi: 10.1186/1471-2105-7-474

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* 279, 5048–5057. doi: 10.1098/rspb.2012.1108

Koralewski, T. E., and Krutovsky, K. V. (2011). Evolution of exon-intron structure and alternative splicing. *PLoS ONE* 6:e18055. doi: 10.1371/journal.pone.0018055

Kozlowski, P., Roberts, P., Dabora, S., Franz, D., Bissler, J., Northrup, H., et al. (2007). Identification of 54 large deletions/duplications in TSC1 and TSC2 using MLPA, and genotype-phenotype correlations. *Hum. Genet.* 121, 389–400. doi: 10.1007/s00439-006-0308-9

Li, X., Wu, H. X., Dillon, S. K., and Southerton, S. G. (2009). Generation and analysis of expressed sequence tags from six developing xylem libraries in *Pinus radiata* D. Don. *BMC Genomics* 10:41. doi: 10.1186/1471-2164-10-41

Li, X., Wu, H. X., and Southerton, S. G. (2011). Transcriptome profiling of *Pinus radiata* juvenile wood with contrasting stiffness identifies putative candidate genes involved in microfibril orientation and cell wall mechanics. *BMC Genomics* 12:480. doi: 10.1186/1471-2164-12-480

Li, X., Yang, X., and Wu, H. X. (2013). Transcriptome profiling of radiata pine branches reveals new insights into reaction wood formation with implications in plant gravitropism. *BMC Genomics* 14:768. doi: 10.1186/1471-2164-14-768

Ling, X.-Y., Zhang, G., Pan, G., Long, H., Cheng, Y., Xiang, C., et al. (2015). Preparing long probes by an asymmetric polymerase chain reaction-based approach for multiplex ligation-dependent probe amplification. *Anal. Biochem.* 487, 8–16. doi: 10.1016/j.ab.2015.03.031

Marcinkowska, M., Wong, K.-K., Kwiatkowski, D. J., and Kozlowski, P. (2010). Design and generation of MLPA probe sets for combined copy number and small-mutation analysis of human genes: EGFR as an example. *ScientificWorldJournal.* 10, 2003–2018. doi: 10.1100/tsw.2010.195

Marcinkowska-Swojak, M., Handschuh, L., Wojciechowski, P., Goralski, M., Tomaszewski, K., Kazmierczak, M., et al. (2016). Simultaneous detection of mutations and copy number variation of NPM1 in the acute myeloid leukemia using multiplex ligation-dependent probe amplification. *Mutat. Res.* 786, 14–26. doi: 10.1016/j.mrfmmm.2016.02.001

Marcinkowska-Swojak, M., Klonowska, K., Figlerowicz, M., and Kozlowski, P. (2014). An MLPA-based approach for high-resolution genotyping of disease-related multi-allelic CNVs. *Gene* 546, 257–262. doi: 10.1016/j.gene.2014.05.072

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., et al. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. U.S.A.* 110, 5241–5246. doi: 10.1073/pnas.1220766110

McCord, B. (2003). *Troubleshooting Capillary Electrophoresis Systems*. Available at: https://pl.promega.com/resources/profiles-in-dna/2003/troubleshooting-capillary-electrophoresis-systems/

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., et al. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* 159, 1295–1308. doi: 10.1104/pp.112.194605

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., et al. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol.* 14:R58. doi: 10.1186/gb-2013-14-6-r58

Perne, A., Zhang, X., Lehmann, L., Groth, M., Stuber, F., and Book, M. (2009). Comparison of multiplex ligation-dependent probe amplification and real-time PCR accuracy for gene copy number quantification using the beta-defensin locus. *Biotechniques* 47, 1023–1028. doi: 10.2144/000113300

Rudi, K., Rud, I., and Holck, A. (2003). A novel multiplex quantitative DNA array based PCR (MQDA-PCR) for quantification of transgenic maize in food and feed. *Nucleic Acids Res.* 31:e62. doi: 10.1007/s00217-009-1155-4

Saintenac, C., Jiang, D., and Akhunov, E. D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* 12:R88. doi: 10.1186/gb-2011-12-9-r88

Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F., and Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.* 30:e57. doi: 10.1093/nar/gnf056

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* 5:e1000734. doi: 10.1371/journal.pgen.1000734

Stankiewicz, P., and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* 61, 437–455. doi: 10.1146/annurev-med-100708-204735

Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., et al. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* 20, 1689–1699. doi: 10.1101/gr.109165.110

Tan, S., Zhong, Y., Hou, H., Yang, S., and Tian, D. (2012). Variation of presence/absence genes among *Arabidopsis* populations. *BMC Evol. Biol.* 12:86. doi: 10.1186/1471-2148-12-86

Thumma, B. R., Matheson, B. A., Zhang, D., Meeske, C., Meder, R., Downes, G. M., et al. (2009). Identification of a cis-acting regulatory polymorphism in a eucalypt COBRA-like gene affecting cellulose content. *Genetics* 183, 1153–1164. doi: 10.1534/genetics.109.106591

Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H., Xu, J., et al. (2015). Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet.* 47, 944–948. doi: 10.1038/ng.3346

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nat. Rev. Genet.* 16, 172–183. doi: 10.1038/nrg3871

Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., et al. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol.* 12:R114. doi: 10.1186/gb-2011-12-11-r114

Żmieńko, A., Samelak, A., Kozłowski, P., and Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* 127, 1–18. doi: 10.1007/s00122-013-2177-7

Zmienko, A., Samelak-Czajka, A., Kozlowski, P., Szymanska, M., and Figlerowicz, M. (2016). *Arabidopsis thaliana* population analysis reveals high plasticity of the genomic region spanning MSH2, AT3G18530 and AT3G18535 genes and provides evidence for NAHR-driven recurrent CNV events occurring in this location. *BMC Genomics* 17:893. doi: 10.1186/s12864-016-3221-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership