

# **Environment dynamics**



# Workflow

Day One: Movement dynamics

Day Two: Adaptation dynamics

Day Three: Environment dynamics



# Day Three focus

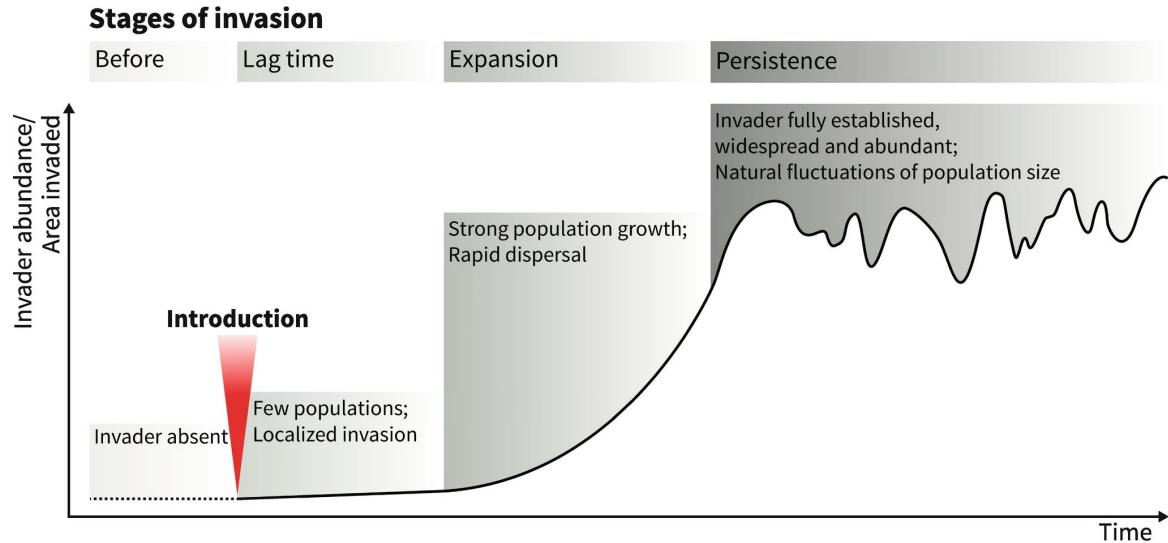
Inferring:

Adaptation to local conditions using  
genotype-by-environment associations (GEAs)

Introduction to environmental association analysis and its  
potential use for understanding invasive species adaptation



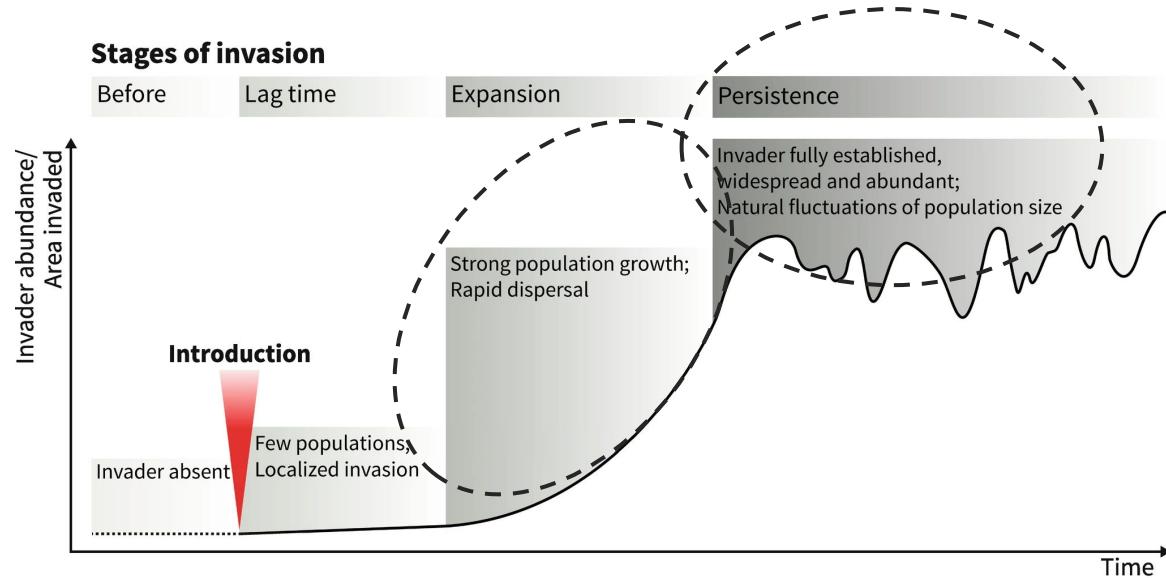
# Invasion curve



## Stages of management

Prevention	Eradication	Containment	Long-term management
Control of vectors and pathways	Measures for eradication might be successful	Prevention of further spread; Public awareness typically begins	Minimise impact of invader; Protection of native species and resources

# Invasion curve



## Stages of management

Prevention	Eradication	Containment	Long-term management
Control of vectors and pathways	Measures for eradication might be successful	Prevention of further spread; Public awareness typically begins	Minimise impact of invader; Protection of native species and resources

# **Environment** dynamics

Why should we care about GEA in invasion biology?

How can we use genomic data to tell us about EAs?

# Genomic data

What kinds of questions can we ask?

What genomic loci (SNPs) are significantly associated with environmental or landscape variables?

Do different methods identify the same associations?

What are the functions of candidate genes?

How might they be facilitating invasion processes?

# GEA premise

Environmental conditions vary across time and space

These differences can cause fitness variation in traits that can be acted on by natural selection

Thus the key focus is to test for associations among environmental, geo-spatial location, or phenotypic factors and genomic variation (*neutral and/or adaptive*)

Null hypothesis is that genomic variation is only due to limited dispersal and genetic drift, i.e., no correlation between allele frequencies and environmental variable(s)



# GEA premise

Selection tests from Day Two (as we ran them) do not identify the environmental factors (potential selective pressures) that drive local adaptation

Various methods – rely on the simplifying assumption that allele frequencies are linearly correlated with env variables, but can account for the confounding effects of underlying neutral genetic structure

Accounting is necessary to avoid high false positives → but can result in false negatives if env factors driving local adaptation are correlated with pop structure (e.g., from patterns of post-glacial recolonisation)



# Common analyses

Regression analyses

RDA, pRDA

GLMs

LFMM

Gradient forest

Target – identifying SNPs that correlate with environmental variables



# Climatic variables

## Getting the variables

Various options; we're using WorldClim;  
<https://www.worldclim.org/data/index.html>



This is WorldClim version 2.1 climate data for 1970-2000. This version was released in January 2020.

There are monthly climate data for minimum, mean, and maximum temperature, precipitation, solar radiation, wind speed, water vapor pressure, and for total precipitation. There are also 19 "bioclimatic" variables.

The data is available at the four spatial resolutions, between 30 seconds (-1 km<sup>2</sup>) to 10 minutes (-340 km<sup>2</sup>). Each download is a "zip" file containing 12 GeoTiff (.tif) files, one for each month of the year (January is 1; December is 12).

variable	10 minutes	5 minutes	2.5 minutes	30 seconds	
minimum temperature (°C)	tmin 10m	tmin 5m	tmin 2.5m	tmin 30s	
maximum temperature (°C)	tmax 10m	tmax 5m	tmax 2.5m	tmax 30s	
average temperature (°C)	tavg 10m	tavg 5m	tavg 2.5m	tavg 30s	
precipitation (mm)	prec 10m	prec 5m	prec 2.5m	prec 30s	
solar radiation (kJ m <sup>-2</sup> day <sup>-1</sup> )	srad 10m	srad 5m	srad 2.5m	srad 30s	
wind speed (m s <sup>-1</sup> )	wind 10m	wind 5m	wind 2.5m	wind 30s	
water vapor pressure (kPa)	vapr 10m	vapr 5m	vapr 2.5m	vapr 30s	

# Climatic variables

Checking for correlations as first step

They are coded as follows:

BIO1 = Annual Mean Temperature

BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))

BIO3 = Isothermality (BIO2/BIO7) (\*100)

BIO4 = Temperature Seasonality (standard deviation \*100)

BIO5 = Max Temperature of Warmest Month

BIO6 = Min Temperature of Coldest Month

BIO7 = Temperature Annual Range (BIO5-BIO6)

BIO8 = Mean Temperature of Wettest Quarter

BIO9 = Mean Temperature of Driest Quarter

BIO10 = Mean Temperature of Warmest Quarter

BIO11 = Mean Temperature of Coldest Quarter

BIO12 = Annual Precipitation

BIO13 = Precipitation of Wettest Month

BIO14 = Precipitation of Driest Month

BIO15 = Precipitation Seasonality (Coefficient of Variation)

BIO16 = Precipitation of Wettest Quarter

BIO17 = Precipitation of Driest Quarter

BIO18 = Precipitation of Warmest Quarter

BIO19 = Precipitation of Coldest Quarter

# RDA

Redundancy analysis (RDA) is an ordination method that extends linear regression to multivariate response data (i.e., allele frequencies of multiple SNPs)

First, linear regressions are computed between allele frequencies at each population and explanatory environmental variables at each SNP

Then fitted values of those regressions are analysed in a PCA to produce ordination axes that are linear combinations of the original explanatory variables

In contrast to LFMM (next), RDA can be run simultaneously on multiple env variables

# RDA

Tests for the effect of environmental variables on allele frequencies

Checks for multicollinearity among predictors (e.g., VIF score)

Typically includes latitude and longitude of the sampling sites as explanatory variables to control for the effect of potential neutral spatial structure

Outlier SNPs identified on each of the selected ordination axes (e.g., those SNPs that are  $\pm 3$  SD from the mean score for that axis)

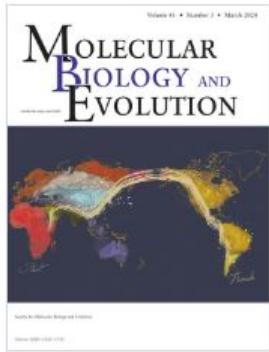
Then calculate correlation between the allele frequency at each outlier and environmental variable(s) → significant at a p-value  $< 0.05$  = candidates for associations with predictor variable(s)

## A note on pRDA

Disentangle the effects of the various predictors (climate, geographic location, demographic/population history) on the genomic variation

pRDA-based variance partitioning compares the variance explained by a full model to that explained by one set of variables (e.g., climate) when the other variables are included as conditional (e.g., genetic)

# Case study 1



Volume 41, Issue 3  
March 2024

JOURNAL ARTICLE

## Rapid and Repeated Climate Adaptation Involving Chromosome Inversions following Invasion of an Insect

Li-Jun Ma, Li-Jun Cao, Jin-Cui Chen, Meng-Qing Tang, Wei Song, Fang-Yuan Yang, Xiu-Jing Shen, Ya-Jing Ren, Qiong Yang, Hu Li ... Show more

*Molecular Biology and Evolution*, Volume 41, Issue 3, March 2024, msae044,  
<https://doi.org/10.1093/molbev/msae044>

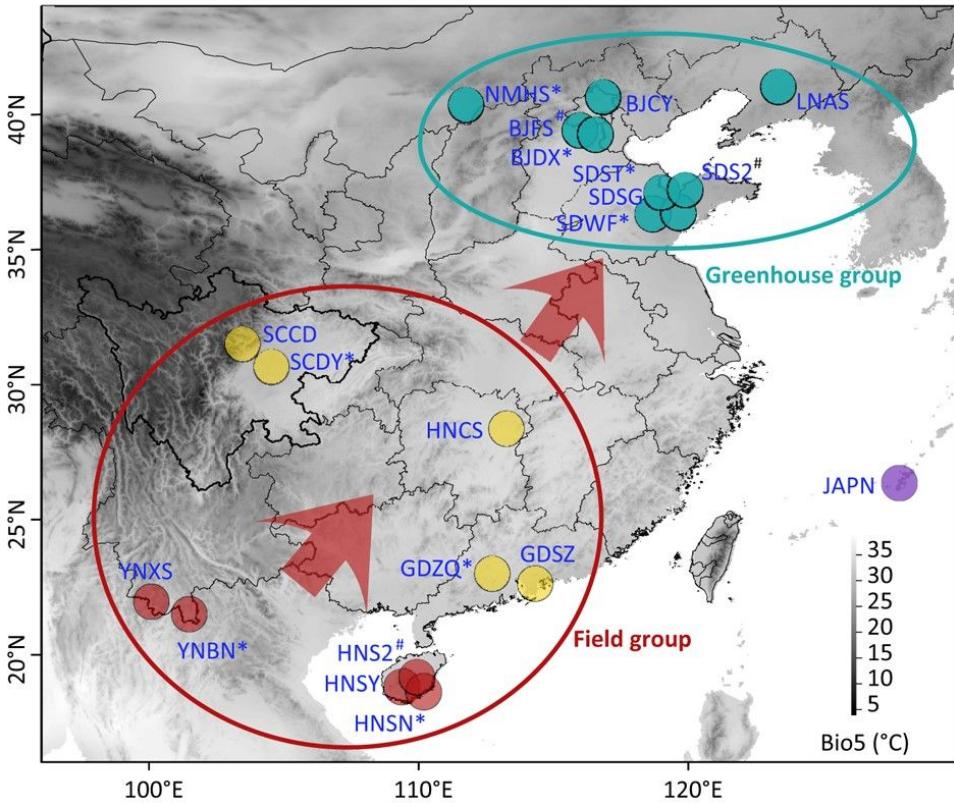
# Case study 1

Insect pest, *Thrips palmi*



Contemporary range expansion across a sharp climate gradient from the subtropics to temperate (warmer) areas

Have populations become adapted to local climates across the range post-invasion?

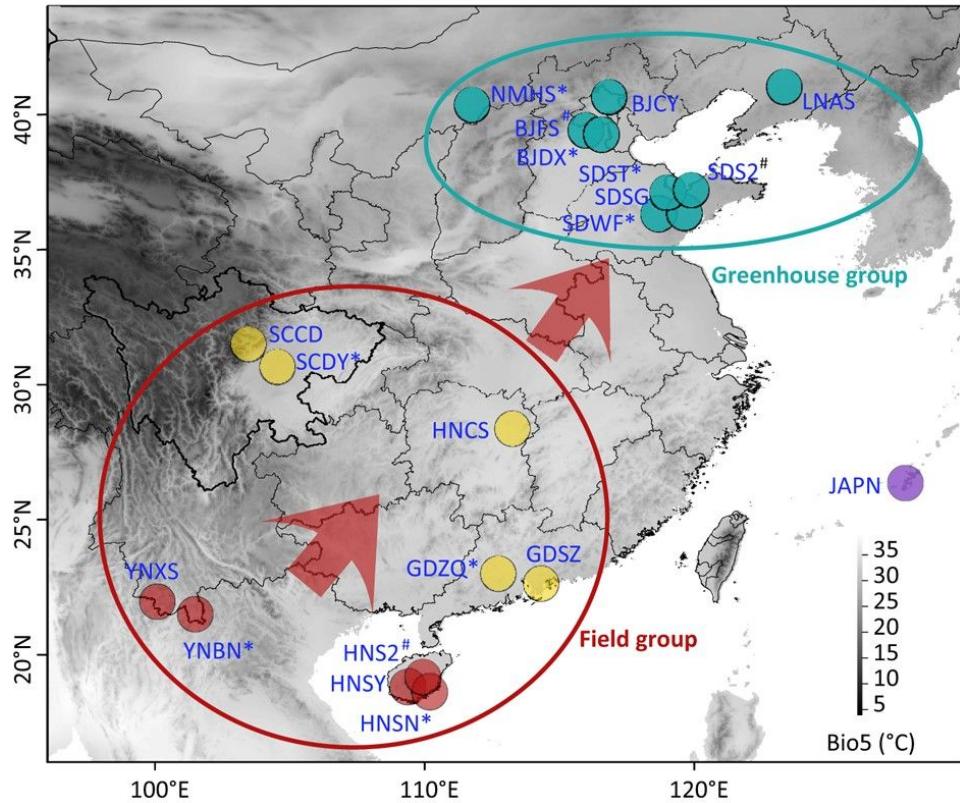


# Case study 1

Conducted whole-genome resequencing

249 individuals, 17 geographical populations across distribution

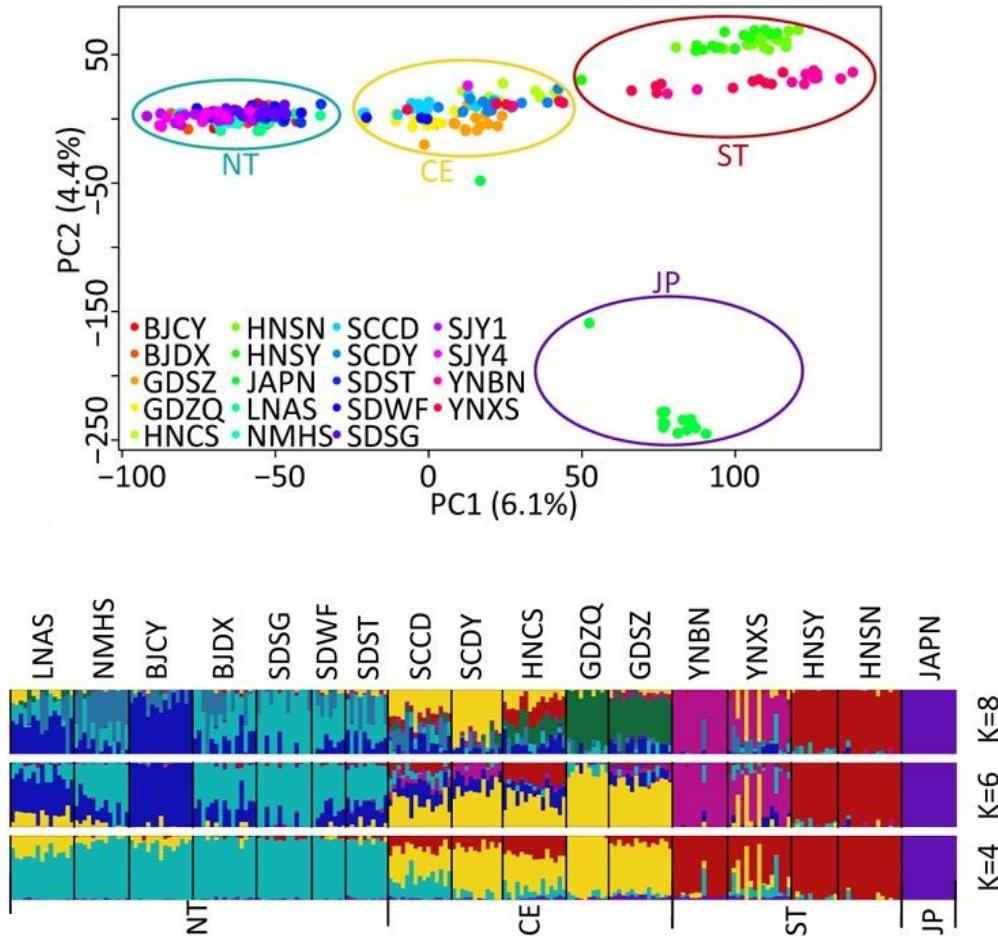
Heat stress assays in common garden experiments



# Case study 1

Eight genetic clusters: unique Japanese population; other 7 clusters mainly distributed in southern, central, and northern groups

Southern and central groups in areas where *T. palmi* can overwinter; Northern group areas where *T. palmi* survive winter climates in greenhouses



# Case study 1

Significant correlation between genetic and geographical distances among all (Mantel test,  $r = 0.594$ ,  $P < 0.0001$ )

Stronger effects of climate than geography on genomic variation

RDA analysis: ~35% (all populations), ~40% (field group), and ~35% (greenhouse group) of genetic variation explained by combined effects of neutral genetic structure, climatic, geography, and their interactions

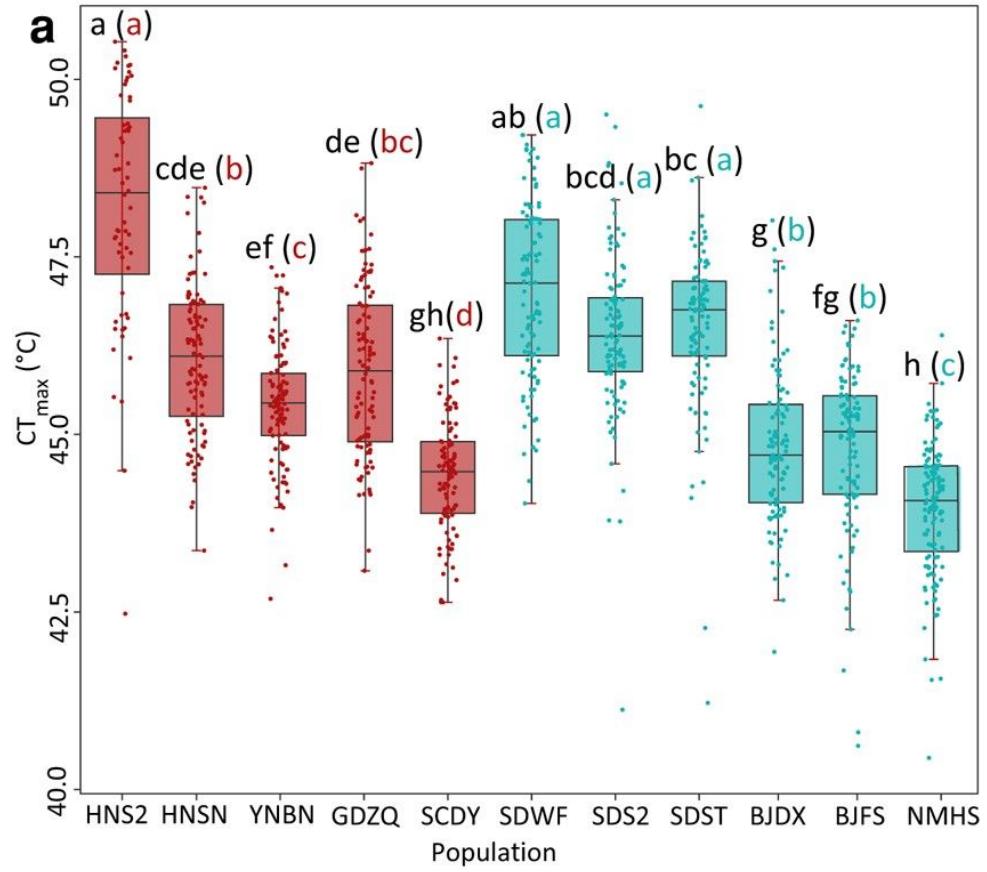
After controlling for neutral genetic variation, climate (all: ~19%; field: ~20%; greenhouse: ~18%) explained more of genetic variance than geography (all: ~6%; field: ~3%; greenhouse: 0%)

Population	RDA model	Variance	$R^2$	P
All populations	Full model	222,800	0.2566	0.001
	Only climate	41,579	0.0479	0.001
	Only structure	48,749	0.0561	0.001
	Only geography	12,754	0.0147	0.001
	Confounded climate/structure/geography	119,718		
	Total unexplained	422,609		
	Total variance	645,409		

# Case study 1

Common garden experiments,  
field and greenhouse groups  
showed clinal patterns in  $CT_{\max}$

Repeated phenotypic adaptation  
to climate and latitude



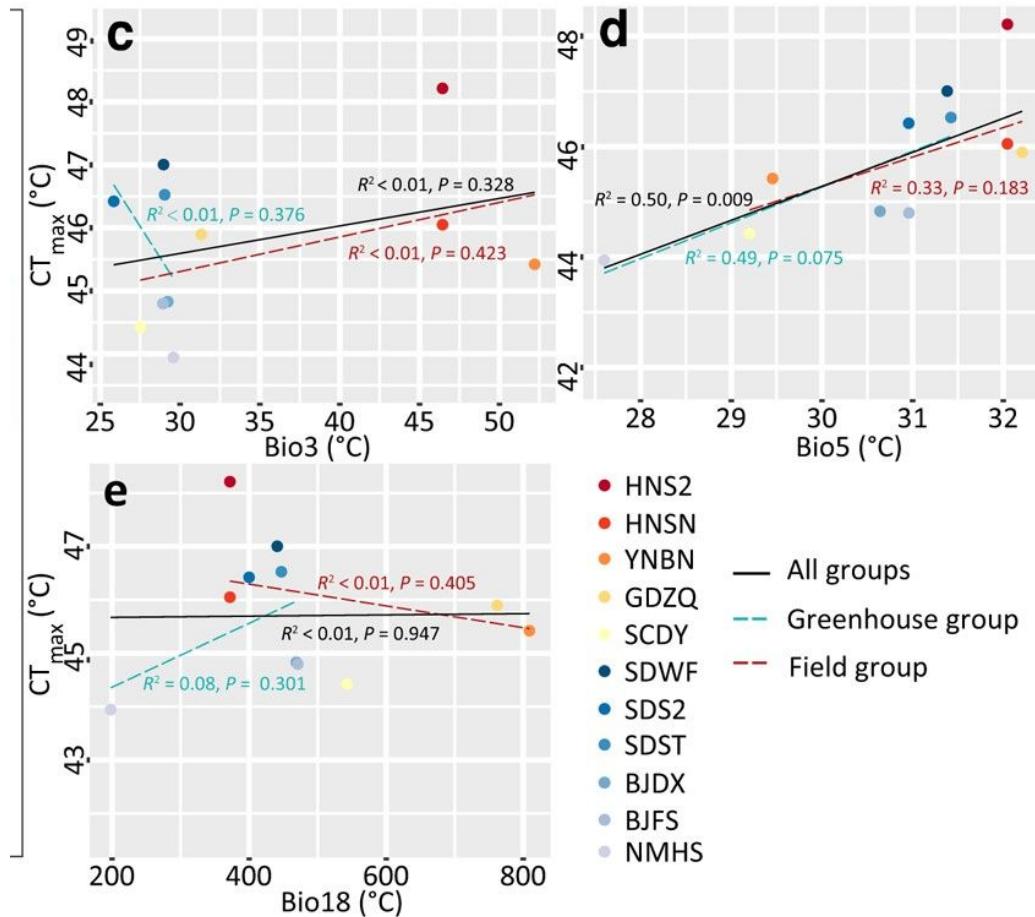
# Case study 1

Three uncorrelated bioclimatic variables

$CT_{max}$  linked with latitude and temperature variables

Temperature-related bio5 ( $R^2 = 0.50, 0.33, 0.49, P = 0.009, 0.183, 0.075$  for all, field, greenhouse)

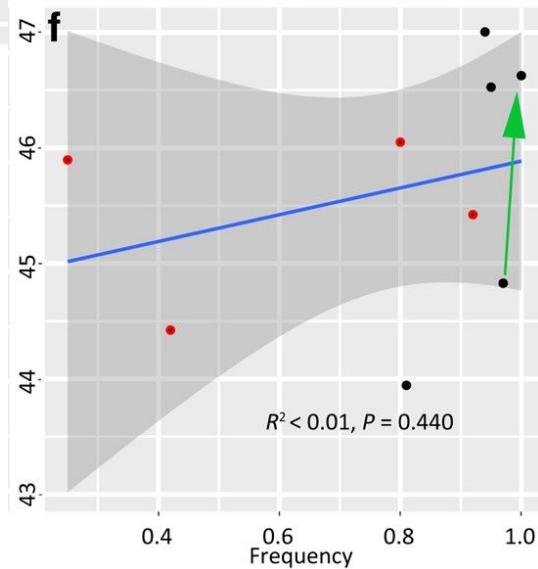
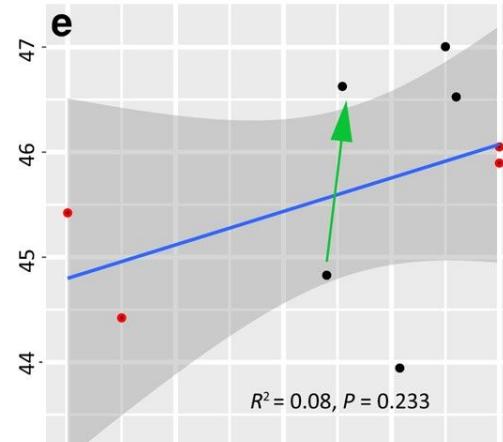
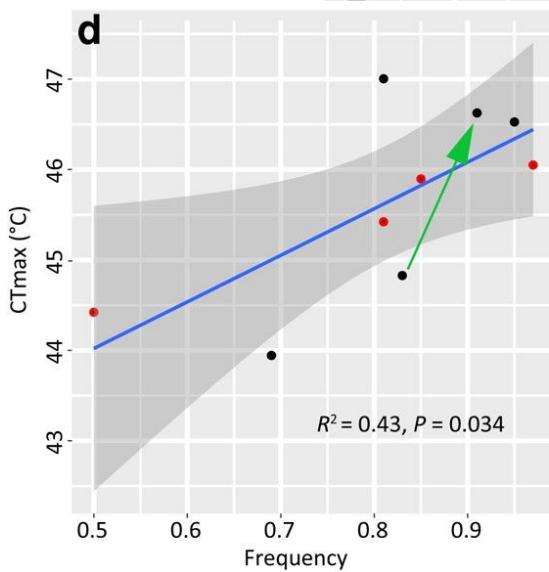
Weaker for bio3 (isothermality), and precipitation-related bio18



# Case study 1

Related patterns back to genetic basis of heat stress adaptation by looking at SVs

Three chromosome inversions identified, correlated with variance in  $CT_{\max}$  among all, greenhouse, field populations



# Case study 1

Implications:

Rapid adaptation to thermal temperature

Importance of inversions as large-effect alleles in climate adaptation and invasion success

Paper also showcases power of different experimental techniques to work together to create a comprehensive picture of adaptive processes

# LFMM

Latent factor mixed models (LFMM) are hierarchical Bayesian mixed linear models that test for correlations/associations between specific SNPs and climate-based principal components (as predictor variables)

Use the covariance structure of all loci being tested to adjust for neutral population history through latent factors (which represent background residual variation that could result from population structure or unmeasured environmental pressures)

Must correct for collinearity in the climatic variables

Method shown to have low false positives rates, performs well under complex population structure and polygenic adaptation, robust to sampling designs and population demography

# LFMM

Parameters estimated in a Bayesian context – replicated models run with MCMC chains and burn-in removed

Median z-scores used to calculate genomic inflation factor, used to calculate adjusted p-values (which are then used to determine significant associations)

Number of latent factors (unobserved variables) specified by the user; corresponds to selection of a K-value to correct for population structure (software like LEA or STRUCTURE used to pick the number of latent factors i.e., K)

Can be good to test more than one value of K but depends on your dataset

Outlier SNPs identified as candidates for selection via association

## Case study 2

# Evolutionary Applications

Open Access



ORIGINAL ARTICLE

Open Access



## Assessing the Impacts of Adaptation to Native-Range Habitats and Contemporary Founder Effects on Genetic Diversity in an Invasive Fish

Thaïs A. Bernos , Zdenek Lajbner, Petr Kotlík, Jacklyn M. Hill, Silvia Marková, Jonah Yick, Nicholas E. Mandrak, Ken M. Jeffries

First published: 04 October 2024 | <https://doi.org/10.1111/eva.70006> | Citations: 1

## Case study 2

Invasiveness can be facilitated by rapid adaptation following species introduction

But, adaptive divergence within the native range can also enhance invasion success

Examine this question in *Tinca tinca*, a fish native to Eurasia and introduced to all continents except Antarctica for aquaculture and recreational fishing

Negative impacts include changes in water quality, disruption of native communities, introduction of pathogens

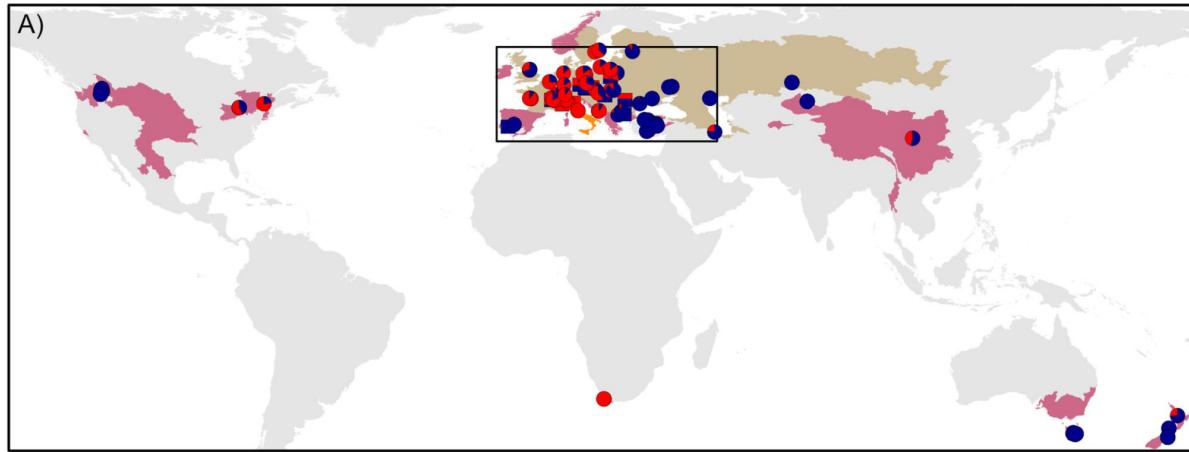


## Case study 2

GBS dataset of 7,058 SNPs

36 native, 40 introduced, 19 farmed populations (768 individuals)

Infer GEAs within native populations and assess whether adaptation to native conditions may aid success of introduced populations

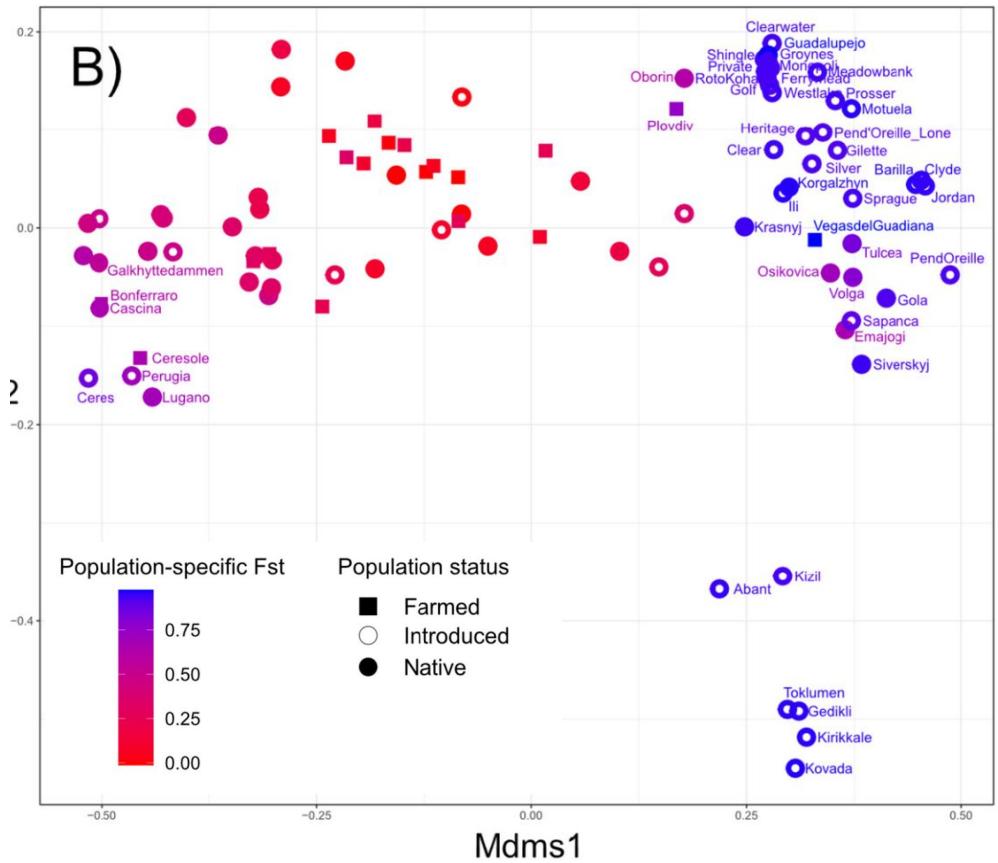


# Case study 2

Population structure

Quite well-mixed

At broadest spatial scale, optimal number of clusters unclear (no clear K-value)



## Case study 2

Used uncorrelated (< 0.7) bioclimatic variables (min and max temp of the coldest and the warmest months, mean annual temp, temp seasonality (→ related to thermal tolerance and/or influence fish physiology and metabolism), and annual precipitation (affects water chemistry and productivity)

Focused on the 36 native populations

pRDA and the LFMM to examine GEAs (at multilocus and single-loci levels, respectively) then ran additional RDA on the putative candidates from LFMM/pRDA

## Case study 2

Also extrapolated a 'genotype-environment index' of genetic composition correlated with climates based on the native range

Ultimately explores the extent of genetic change needed for the genetic composition of source populations to correlate with environmental conditions in the invaded range (i.e., the genetic offset)

Calculate difference between the GE index of the environmental pixel where a tench population was introduced and that of the most closely related native population (based on pairwise  $F_{ST}$ )

Small values = genotype occurs in invaded range under similar env conditions as native range; large = mismatch between the env under which a genotype occurs in native vs invasive range i.e., genomic offset/maladaptation

# Case study 2

Climate, geographic location, and demographic history explained 78% of the total genetic variance

Individually, climate explained 6% of total genetic variation, geography 4%, demographic history 31%

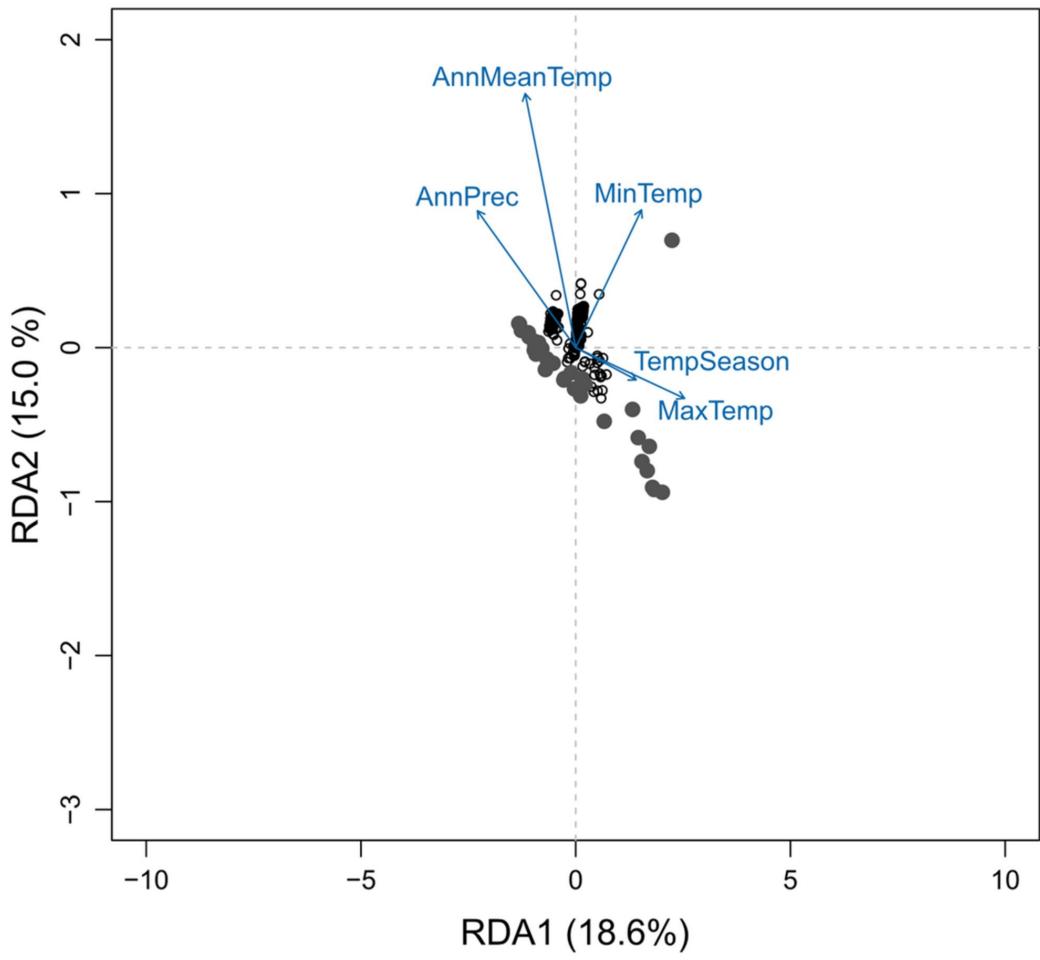
pRDA models	Inertia	<i>R</i> <sup>2</sup>	<i>p</i>	Explainable variance (%)	Total variance (%)
Full model (clim. + geog. + stru.)	863.6	0.785	0.001	1	0.78
Climate (clim. + [geog., demo.])	61.2	0.056	0.036	0.07	0.06
Geography (geog. + [clim., demo])	43.4	0.039	0.001	0.05	0.04
Demography (demo. + [geog., clim.])	339.3	0.308	0.001	0.40	0.31
Confounded	396.4			0.52	0.41
Total unexplained	236.7				0.22
Total inertia	1100.3				1

## Case study 2

The LFMM and the pRDA identified 357 and 408 candidate loci, respectively (122 shared)

pRDA, many candidates strongly associated with annual mean temp, then min temp of coldest month, max temp of warmest month

2nd RDA on adaptive loci explained ~37% of total variance



## Case study 2

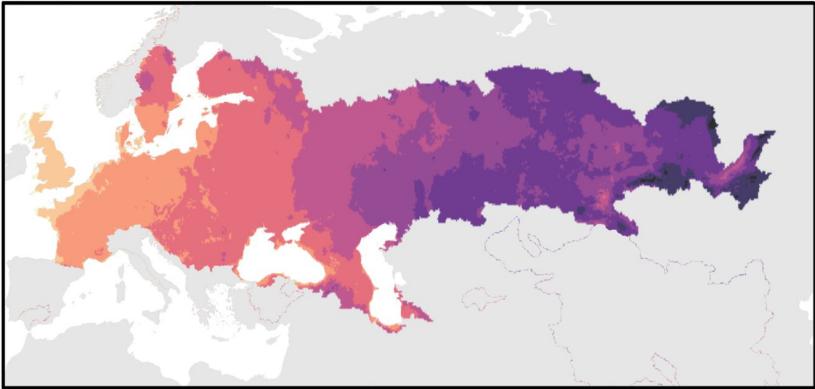
In the native range, GE index gradient contrasted Western (low value) and Eastern (high value) regions

40 invasive populations showed a range of higher and lower values

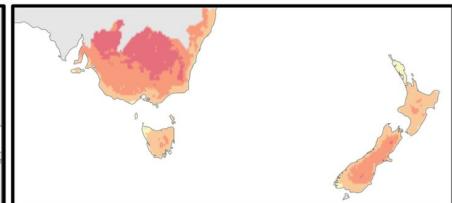
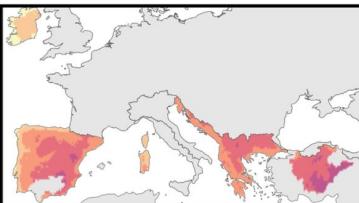
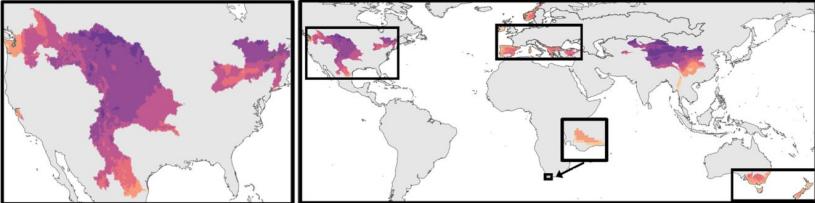
27 had small ( $< 2.83$ ) genomic offset (i.e., difference in GE index compared to their most genetically similar native population), while four had high genomic offset (but some of these still widespread)



A) Projected across the native range



B) Projected across the introduced range



## Case study 2

Implications:

This study showed adaptation in the native range

But, not necessarily important for invasion success, i.e., some widespread invaders with high GO

Lack of prior adaptation in the native range may not always be a problem for invasives, especially those that have expanded distributions in their native range that encompass broad environmental breadths

# GF

Gradient forest (GF) is an extension of the random forest machine learning approach

Models non-linear change of allele frequency at each locus along an environmental gradient

Assesses allele-specific ‘turnover functions’ that identify major ‘tipping point’ environmental conditions

# GF

Putative selective loci obtained via some method (e.g., PCAdapt) used as predictors in the GF analysis

Account for spatial structure using ‘Principal Coordinates of Neighbor Matrices’ (PCNMs)

Cumulative distributions of the standardised importance for each SNP then calculated separately and collectively, and scaled by  $R^2$ , for every examined bioclimatic variable

Then, ‘predict’ function in the “raster” package used to simulate a value for each tested env variable for every pixel in the geographic distribution → plotted visually to assess how allele frequencies change across the env gradient

# Case study 3

## scientific reports

[Explore content](#) | [About the journal](#) | [Publish with us](#)

[nature](#) > [scientific reports](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 26 October 2024

### Global invasion history with climate-related allele frequency shifts in the invasive Mediterranean fruit fly (Diptera, Tephritidae: *Ceratitis capitata*)

[Pablo Deschepper](#)✉, [Sam Vanbergen](#), [Massimiliano Virgilio](#), [Andrea Sciarretta](#), [Marco Colacci](#), [Vasilis G. Rodovitis](#), [Josep A. Jaques](#), [Mario Bjeliš](#), [Kostas Bourtzis](#), [Nikos T. Papadopoulos](#) & [Marc De Meyer](#)

[Scientific Reports](#) **14**, Article number: 25549 (2024) | [Cite this article](#)

**2424** Accesses | **4** Citations | **1** Altmetric | [Metrics](#)

## Case study 3

Mediterranean fruit fly (*Ceratitis capitata*)

Significant pest of fruit crops



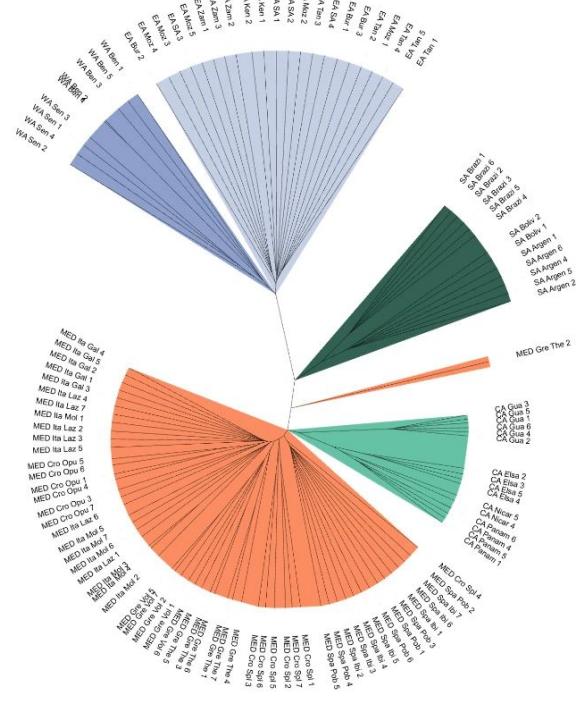
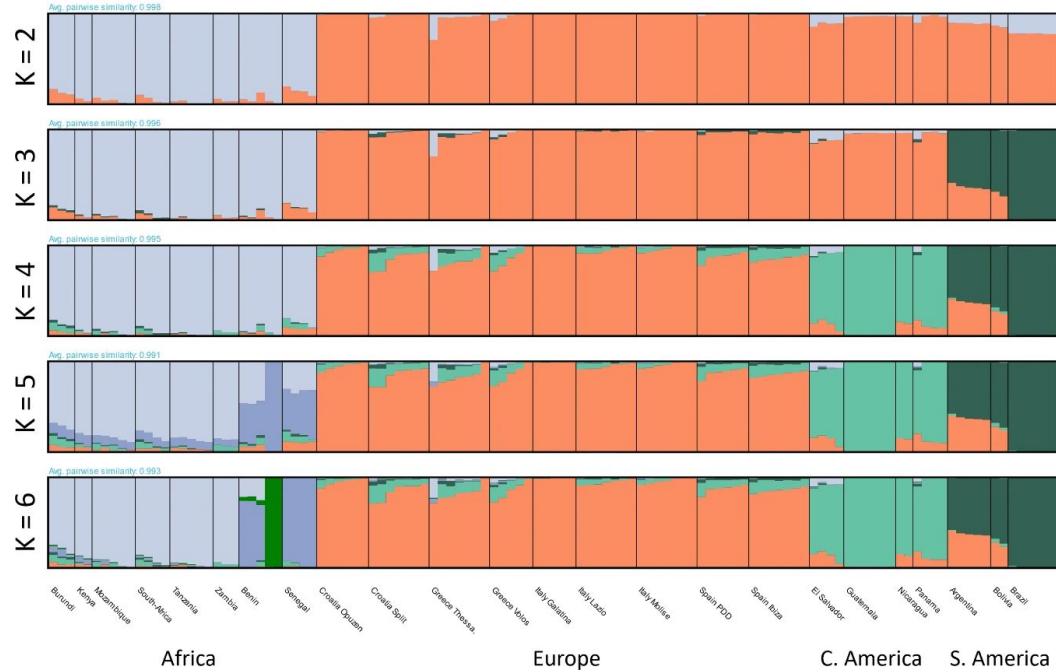
Native to tropical sub-Saharan Africa, invasive in northern Africa, all Mediterranean countries, Central and South America, Western Australia, and the Hawaiian islands; also commonly detected in parts of US and central Europe

Wide invasive range, exposed to diverse climatic regimes

Coping with diverse climatic conditions likely a factor in invasion success

# Case study 3

117 samples, 25 locations in 20 countries; K = 5



Distinct native African cluster; American samples more distinct from K=3

# Case study 3

Five uncorrelated climatic variables, then LFMM (using K=5), RDA, and PCAdapt

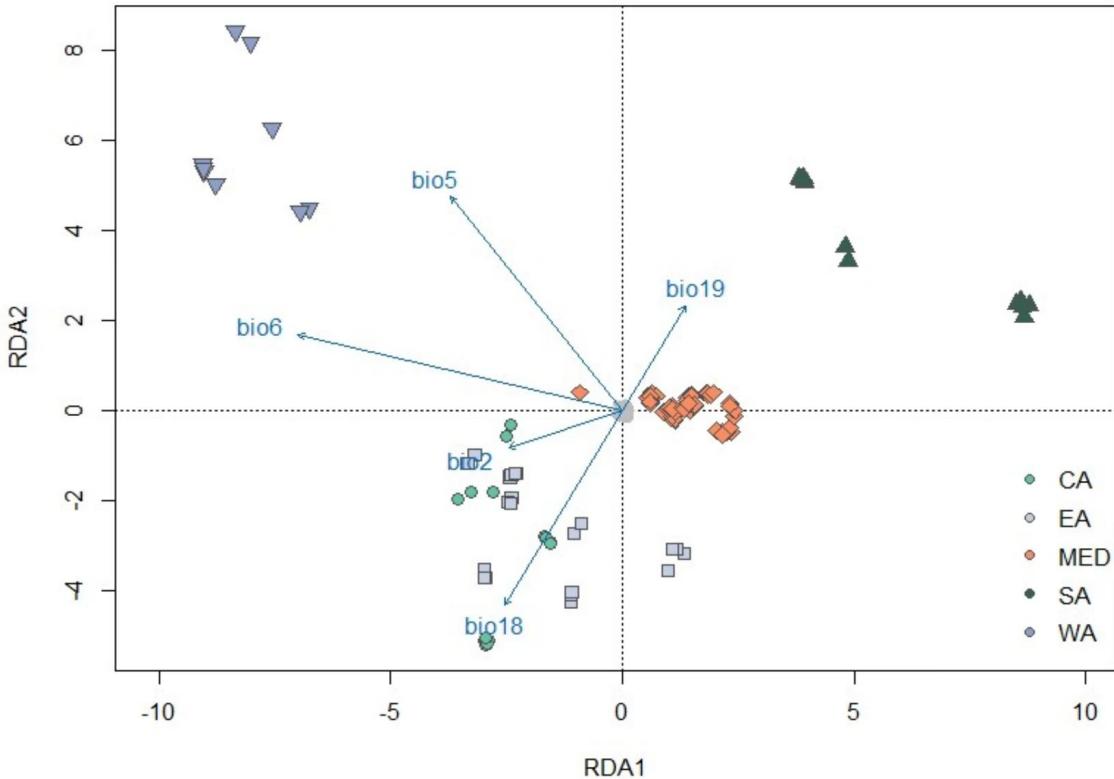
GF applied to map the allelic turnover (genomic composition) of putative selective SNPs from PCAdapt

# Case study 3

LFMM = 72 SNPs associated to the uncorrelated climatic variables

RDA = 945 putatively selective SNPs; significant RDA outliers most associated with BIO5 (447 SNPs), BIO18 ( $n=173$ ), BIO2 ( $n=159$ ), BIO6 ( $n=152$ ) and BIO19 ( $n=13$ )

PCAdapt = 7,102 outlier loci



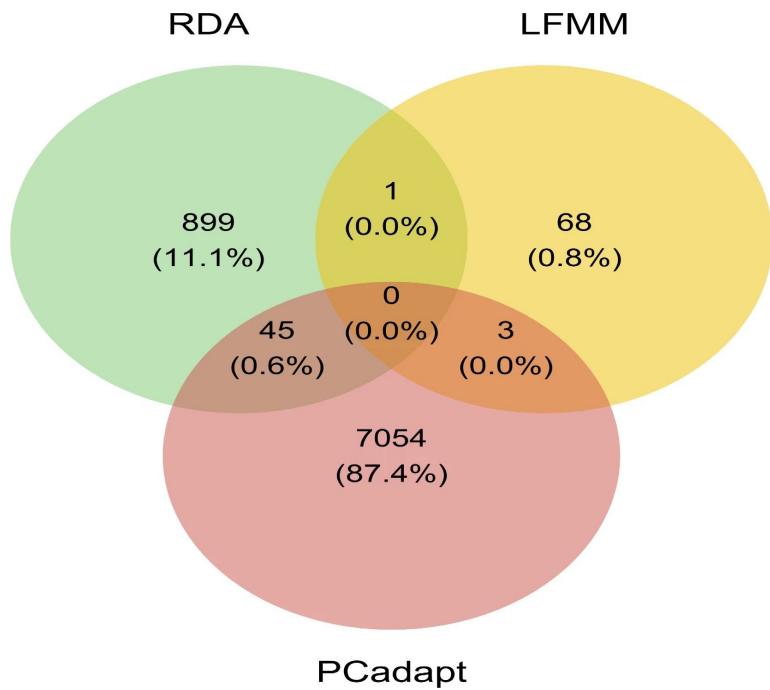
# Case study 3

One SNP shared between LFMM and RDA

PCadapt shared 3 and 45 SNPs with LFMM and RDA approach, respectively

No SNPs shared across all three methods

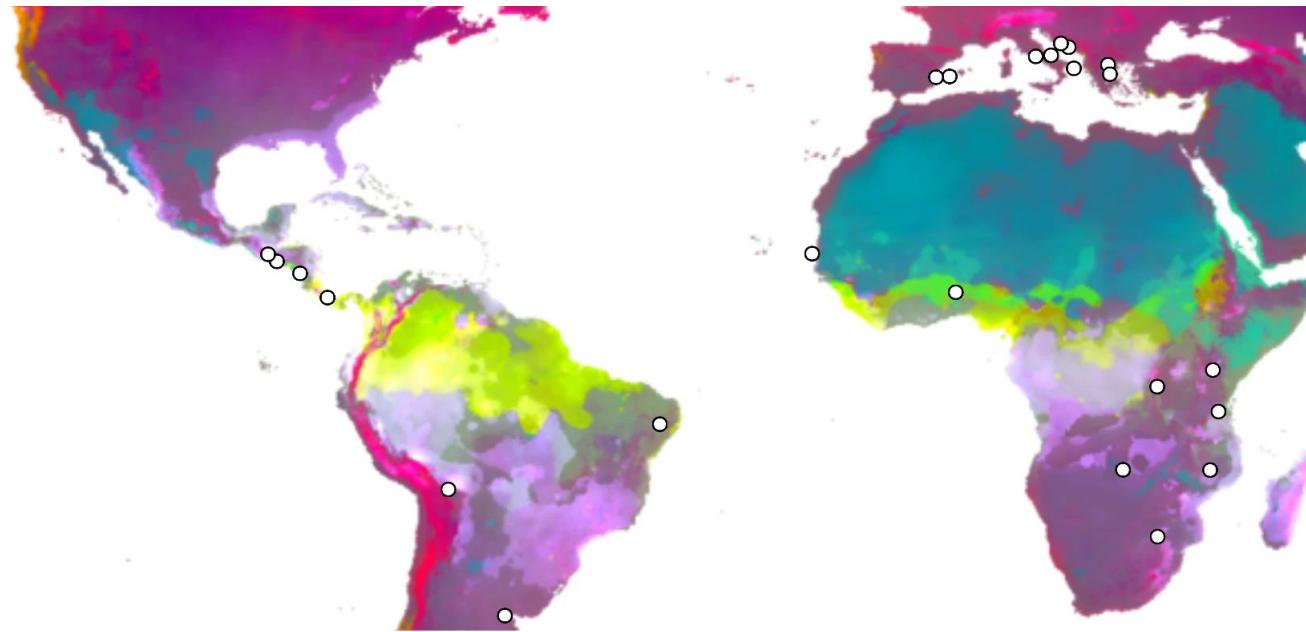
Authors proceeded with 49 SNPs (shared by at least 2 methods) for GF



# Case study 3

GF analysis found that BIO18 was most important (explained ~7% of variation), then BIO5 (6.5%), BIO19 (6.4%), BIO6 (5.9%)

Spatial variables (PCNMs) and BIO2 <2.7% of the variation

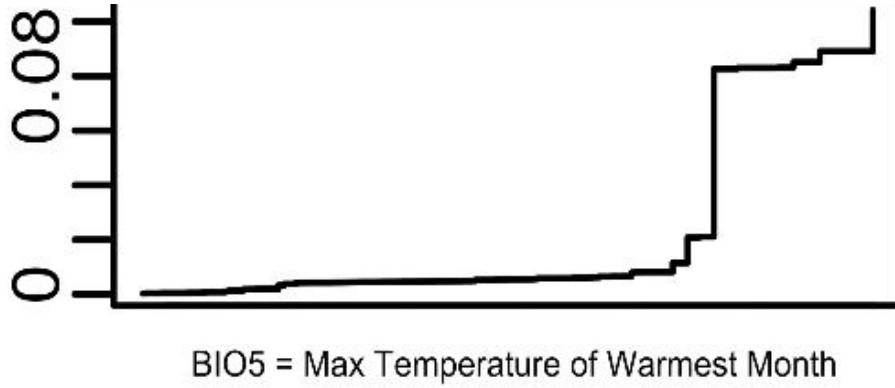


Plot shows predicted genetic composition by GF for all 49 putatively selective loci; colours = allelic-turnover across the map, including unsampled regions (sampling locations = white circles)

## Case study 3

Large allelic turnover observed for BIO5

The SNP that accounted for the highest variation (31.3%) also seen for BIO5

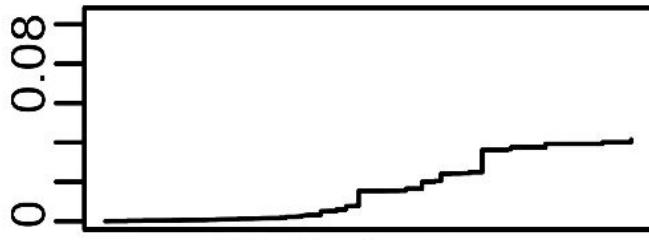


## Case study 3

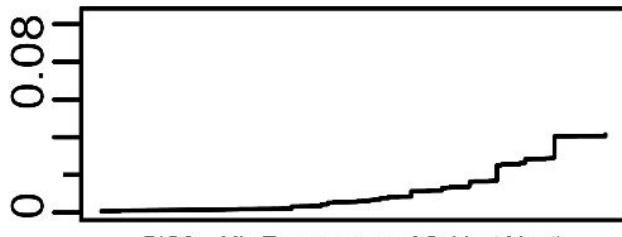
BIO2, BIO6, and BIO18 show more gradual changes in allelic composition

A key finding was a significant allele frequency shift related to adaptation to cold stress (BIO6; min temp coldest month)

Suggests potential to rapidly adapt to seasonal variations in colder climates



BIO2 = Mean Diurnal Range



BIO6 = Min Temperature of Coldest Month



BIO18 = Precipitation of Warmest Quarter

# Case study 3

Implications:

Evidence for rapid adaptation over steep climatic barriers (causing changes in allelic composition) at 49 outlier loci

Climatic barriers may impede gene flow due to maladaptation (e.g., may explain limited admixture between Central/South American populations)

Adaptation to winter cold stress in particular → a broader phenomenon to explore in other species?

→ Potential for invasive fruit flies to rapidly expand range by shifting allele frequencies in response to the climate during invasion

# Common approaches

As for selection analysis, it is common to run multiple GEA analyses with different assumptions/algorithms → we will run RDA/LFMM/GF

Look for union → outliers that show up in > 1 method more robust / higher confidence; but what about outliers unique to certain methods?

Typical to follow up with gene annotation (if you can!) → does the annotation make logical sense? Associations to known phenotypes and/or environmental covariates a powerful approach

Ultimately, functional validation required to confirm → what happens when we knock it out?

# Genomic insights

Understanding rapid adaptation at a mechanistic level by identifying the selective pressures in operation and how they are being responded to

Predicting spread and impact

Informing management strategies

# Future directions

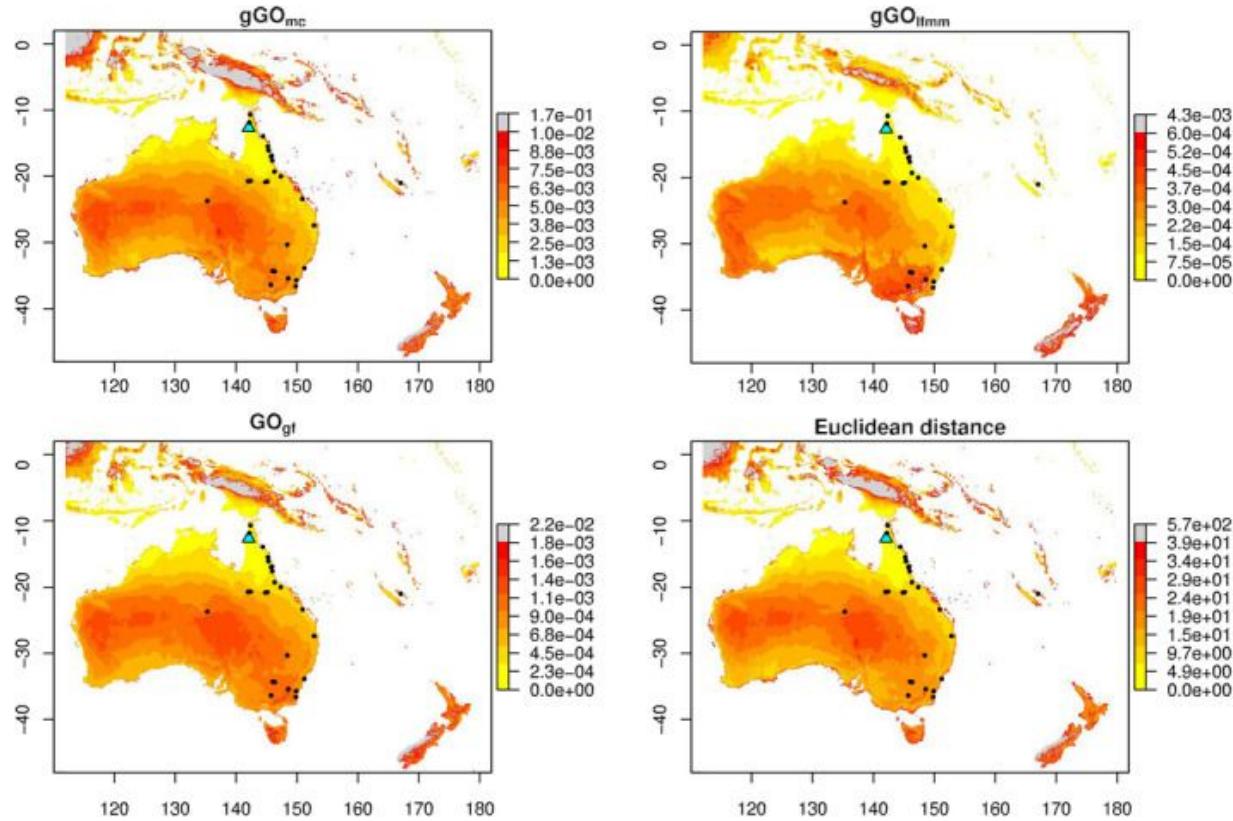
Ultimate aim is to link genotype, phenotype, environment

Role of epigenetics and gene regulation in rapid adaptation?

Combining genomic data with ecological niche models?

Genomic offset

# Future directions



Source = blue triangle

Yellow = lower GO  
(common in native range)

Red = higher GO  
(future invasion?)

# Food for thought

As for Day Two: can we use genomic data to predict which species will adapt successfully in new environments?

Particularly by determining which environments they will be well and maladapted to



## Conclusions

The environment is a key driver of species distributions via selective processes

Genomic data can identify selective pressures via their association with genomic variation; the genomic variants can potentially be annotated to identify candidate invasion-relevant genes

# Final notes from me

Things we didn't cover:

Demography, inbreeding depression, genomic load, etc

Not because they're not important!

# Final notes from me

Things we didn't cover:

Demography, inbreeding depression, genomic load, etc

Not because they're not important!

My hope that you continue on your invasion genomics journey and that this workshop has been helpful...!

# Final notes from me

Things we didn't cover:

Demography, inbreeding depression, genomic load, etc

Not because they're not important!

My hope that you continue on your invasion genomics journey and that this workshop has been helpful...!

A big thank you to Ellie!



# Final notes from me

Things we didn't cover:

Demography, inbreeding depression, genomic load, etc

Not because they're not important!

My hope that you continue on your invasion genomics journey and that this workshop has been helpful...!

A big thank you to Ellie!

And a big thank you to all of you :)



# Next steps

Please move to the Day Three tutorial

Remember to ask questions via slack!

**CREDIT:** SlideEgg PowerPoint template

[www.slideegg.com](http://www.slideegg.com)

