

# **Adaptation dynamics**



# Workflow

Day One: Movement dynamics

Day Two: Adaptation dynamics

Day Three: Environment dynamics



# Day Two focus

Inferring:

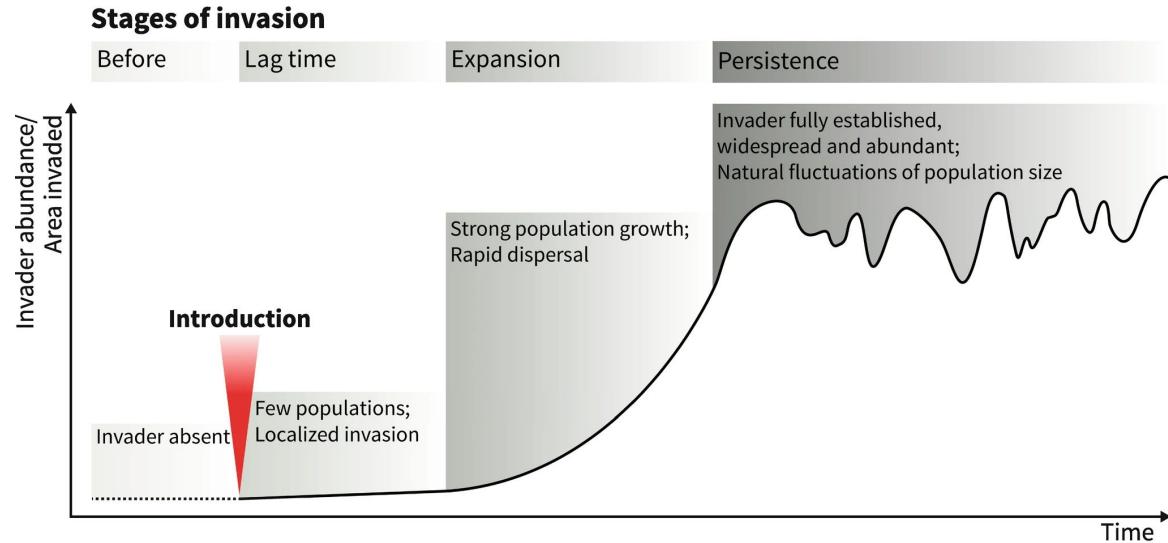
Selection/adaptive processes in the invaded range\*

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

\*The methods are location-independent!



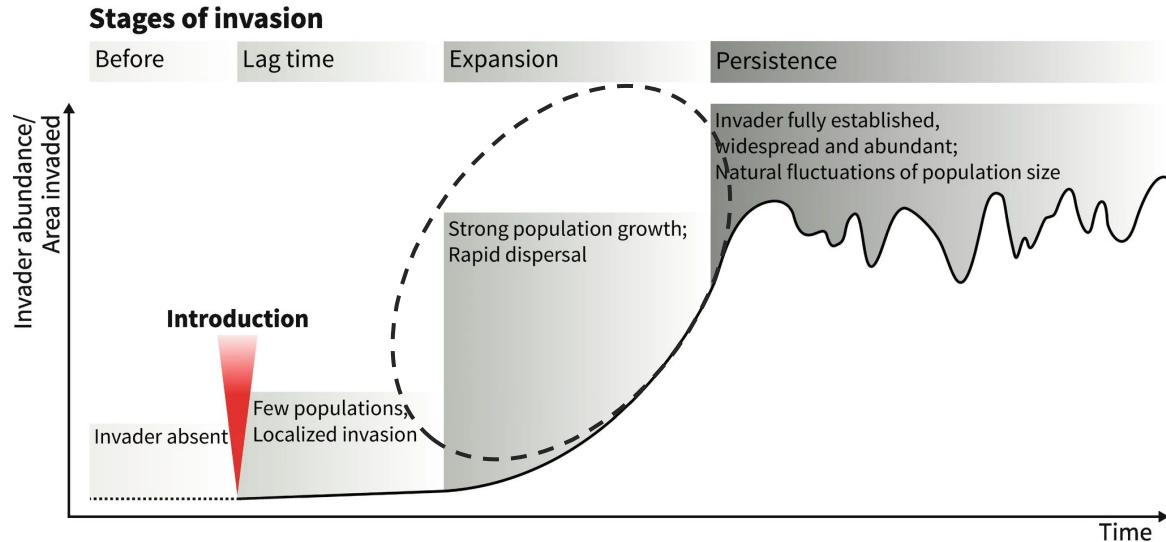
# Invasion curve



## 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

# Adaptation dynamics

Why should we care about adaptation in invasion biology?

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

# Genomic data

What kinds of questions can we ask?

What genomic regions are under selection?

Do different methods show the same outliers?

What genes are affected?

What is the function of those genes?

How might they be facilitating invasion processes?

# Common analyses

Genome outlier scans

Different flavours:

'Core'

Relate to phenotypes

Related to environmental/landscape variables

Target – identifying SNPs that correlate with those variables



# Selection definition

Adaptation occurs when a species evolves to become a better ‘fit’ to its environment in response to a selective pressure

Particularly common in changing environments (e.g., as experienced by new invaders!)

What we see in a microevolution sense is a *change in allele frequencies*

# Selective sweeps

Population genomic methods can be used to identify footprints of selection in the genome following adaptation processes → selective sweeps

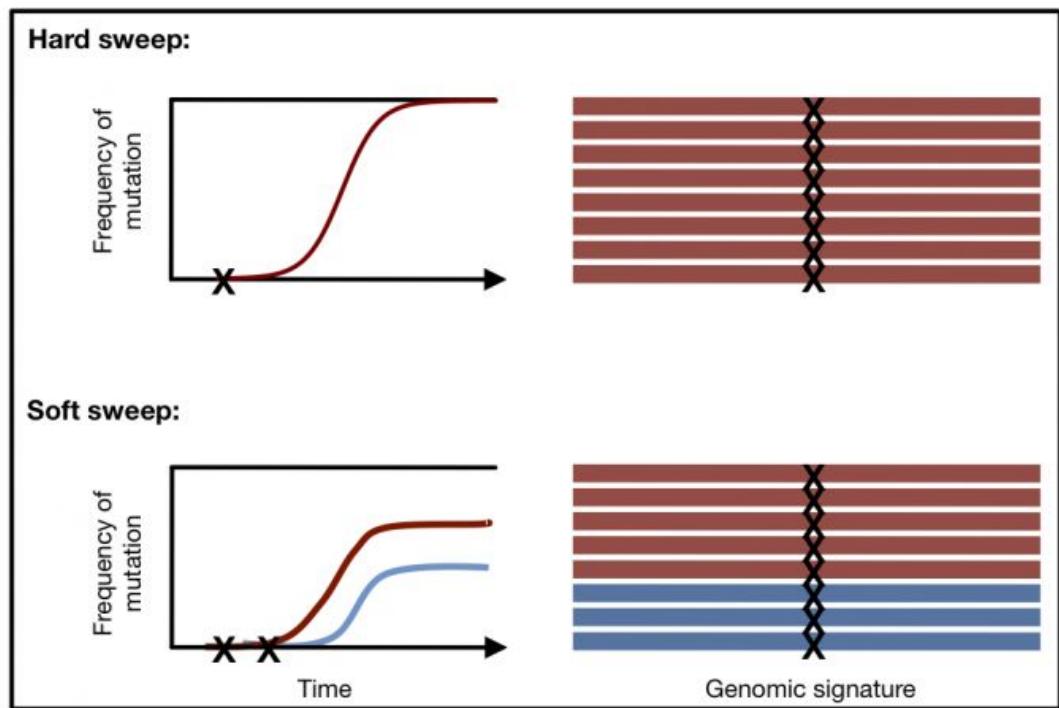
**Hard** selective sweeps occur when a single adaptive mutation rises in frequency

**Soft** selective sweeps occur when multiple adaptive mutations at the same locus sweep through the population simultaneously

# Selective sweeps

In both hard and soft selective sweeps, the alleles surrounding the adaptive mutation will also increase in frequency (i.e., as they're swept along)

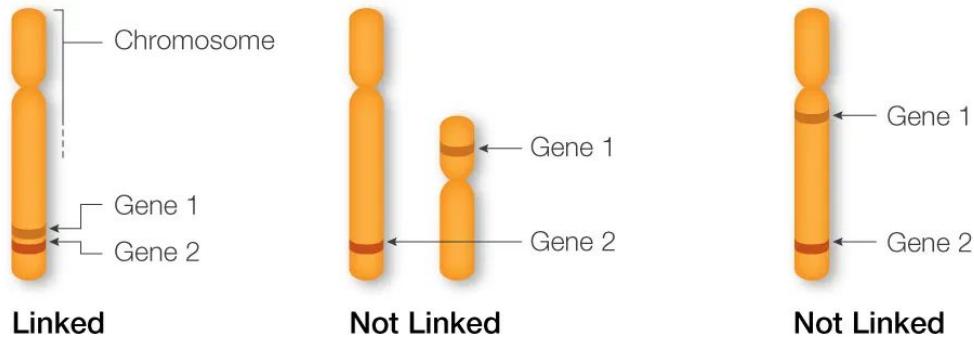
In a hard sweep, a single haplotype should be present at high frequency, whereas in a soft sweep many haplotypes should be present at high frequencies



# Selective sweeps

In practice, genetic diversity is reduced because the selected allele sweeps to fixation and surrounding linked loci go along with it ('hitchhike')

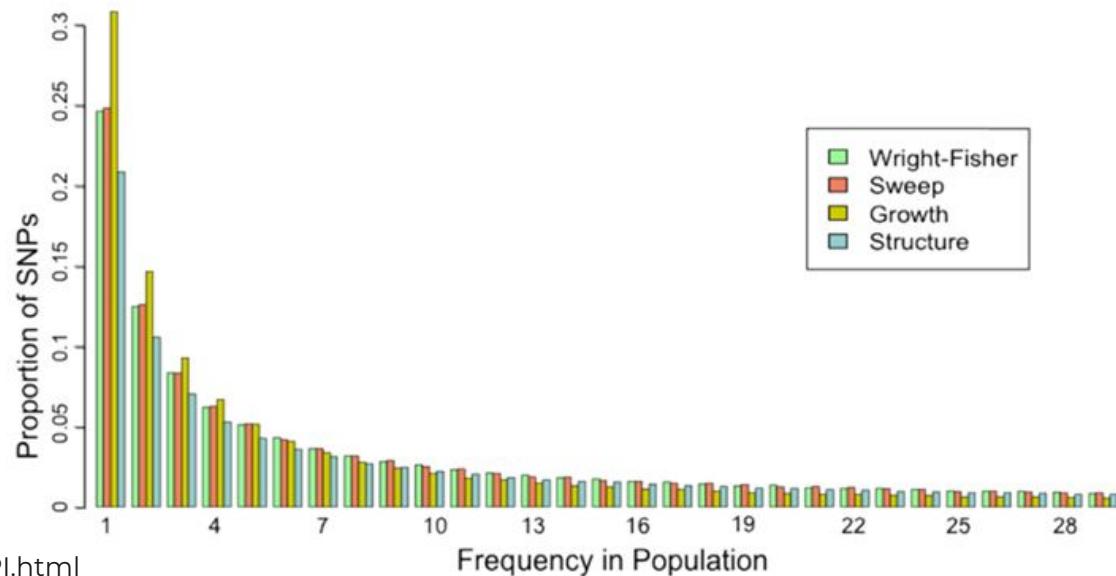
The size of the swept region depends on the strength of selection (thus how fast the sweep occurs) and recombination (LD) at either side of the selected site



# Selective sweeps

Identification of hard sweeps is based on patterns of very low genetic diversity, with new mutations showing an excess of low-frequency-derived alleles of the background level → i.e., scan for low diversity, excess of rare alleles (Tajima's D), shift in SFS toward high-frequency derived alleles

Soft sweeps leave a subtler signature: lower reduction in genetic diversity, no large shift in SFS as the selected allele is present on multiple haplotypes



# Selective sweeps

We can use statistical methods to identify both hard and soft sweeps

Scan the genome for areas that show different patterns (e.g., in the test statistic) than expected under an assumption of neutrality

Methods based on ‘extended haplotype homozygosity’, which look for derived alleles on very long haplotypes thought better for soft sweep detection (ML as well?)

Those based on diversity and SFS better for hard sweeps

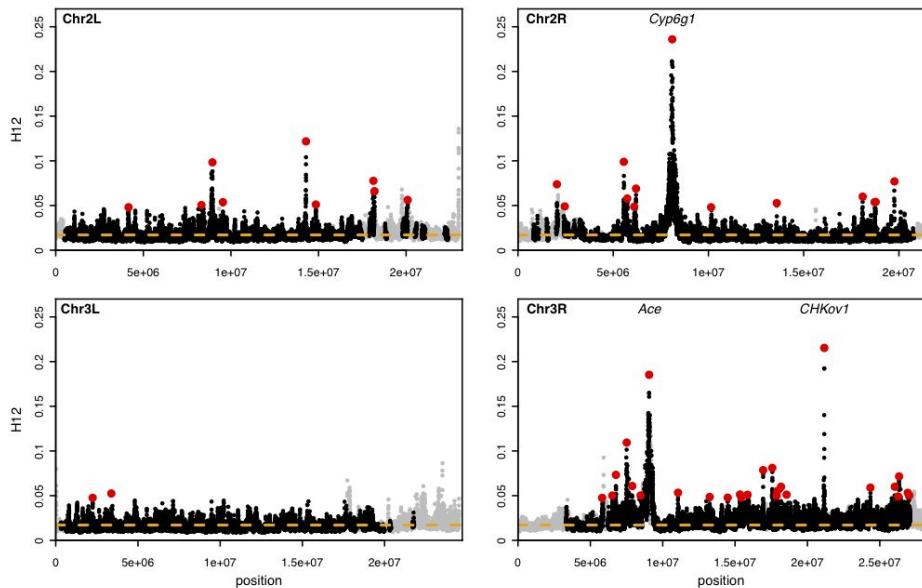
# Selective sweeps

Scan for hard and soft sweeps in *Drosophila melanogaster* data

Each point = H12 value (window size 400 SNPs); grey points = low LD regions (excluded as they can generate high H12 values)

Orange line = highest H12 under neutrality; red points = top 50 H12 peaks relative to the orange line

Positive controls identified



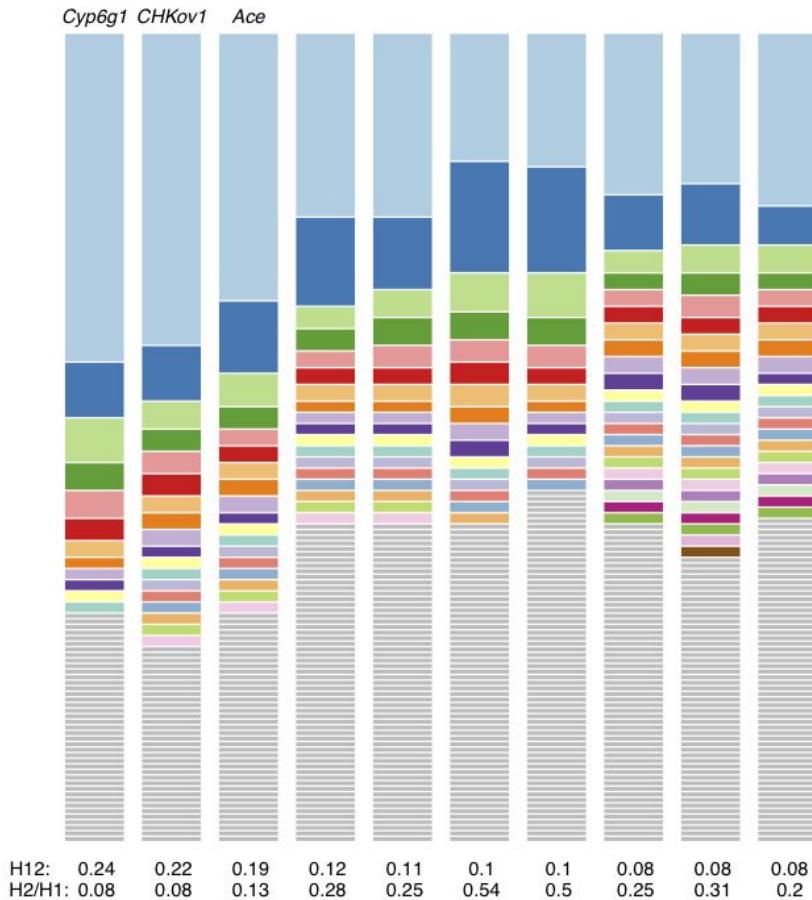
# Selective sweeps

Hard or soft?

Haplotype frequency spectra for the top 10 H12 peaks in previous figure

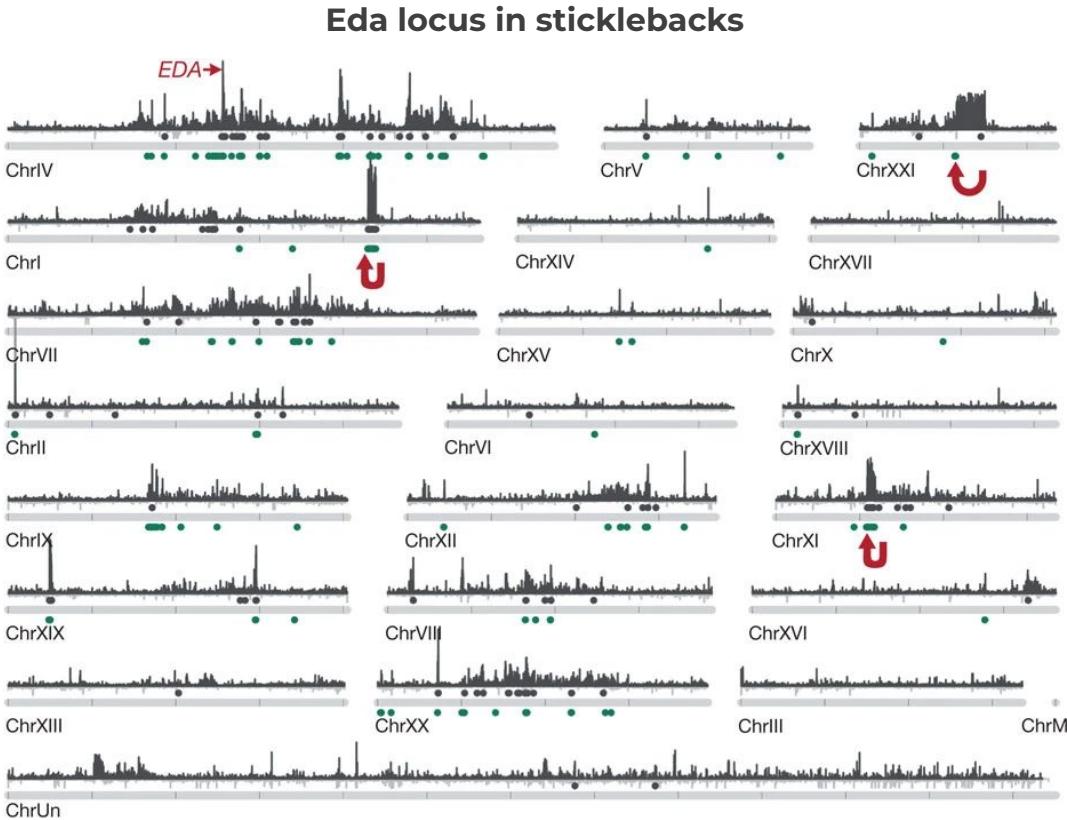
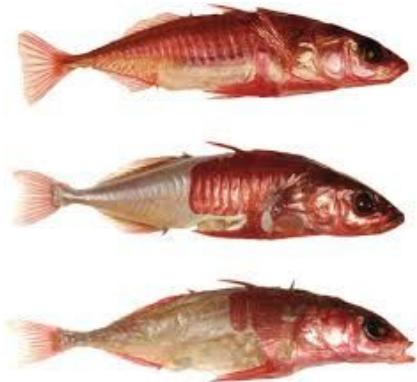
Height of the first bar (light blue) = frequency of the most prevalent haplotype, then 2nd, 3rd, etc (grey bars = singletons)

All peaks show multiple haplotypes at high frequency = soft sweeps



# Selective sweeps

Genomic data can be combined with phenotypic and environmental data to identify how these factors underpin the response to positive selection



# Selection detection methods

Genome scans can be based on identifying regions with

Low diversity, an excess of rare alleles (Tajima's D), a shift in SFS

Specific test statistics (e.g., H12, H2/H1, XtX, C2)

# Challenges

Different methods can result in false positives and false negatives

Demographic confounding: bottlenecks and drift can mimic selection

Population confounding: differentiation patterns (e.g., from  $F_{ST}$  scans) can be due to population isolation vs selection

Polygenic adaptation: small shifts across many loci are harder to detect

Temporal sampling: may need time-series data to confirm ongoing selection

# PCAdapt

See <https://bcm-uga.github.io/pcadapt/articles/pcadapt.html>

This method aims to detect adaptive genetic markers using outlier detection based on PCA → no population designation required (within-group analysis)

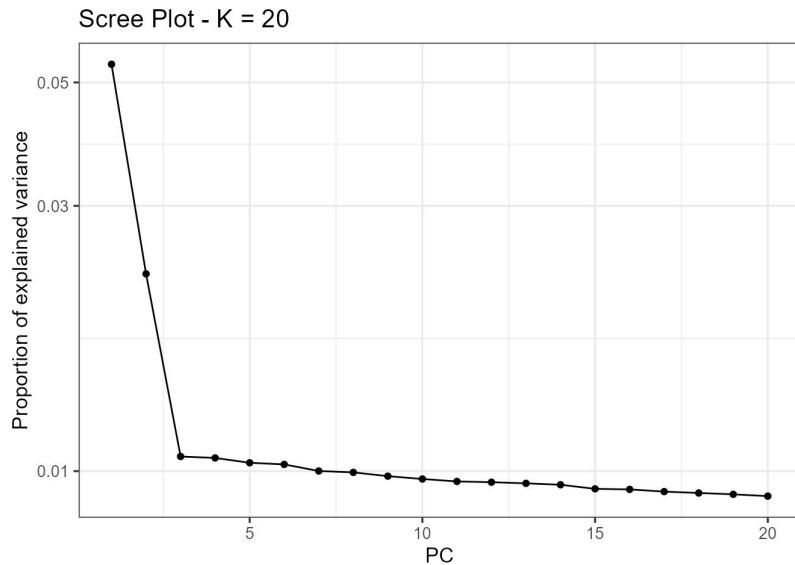
The method works by reading in genotype data, and performing PCA on the genotype matrix, thus finding outliers wrt background population structure

Next, test statistics and associated p-values are calculated based on correlations between the SNPs and the PCs

# PCAdapt

An important step is working out the number of PCs to use

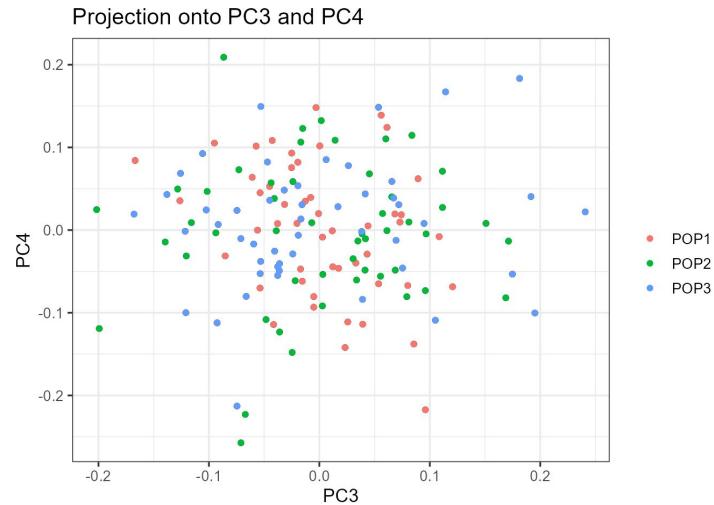
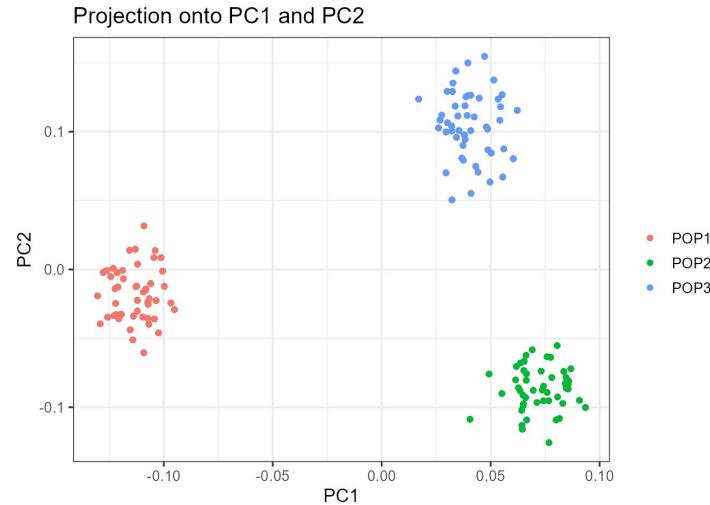
A scree plot shows the decreasing percentage of variance explained by each PC ( $K = 2$  in the example, as you should use the no. of PCs that correspond to the left of the point at which the line bends or levels off)



# PCAdapt

A score plot can also be used to assess this – it shows the projections of individuals onto the specified PCs

In the example, you can see there's no structure on PC3/4, confirming K = 2

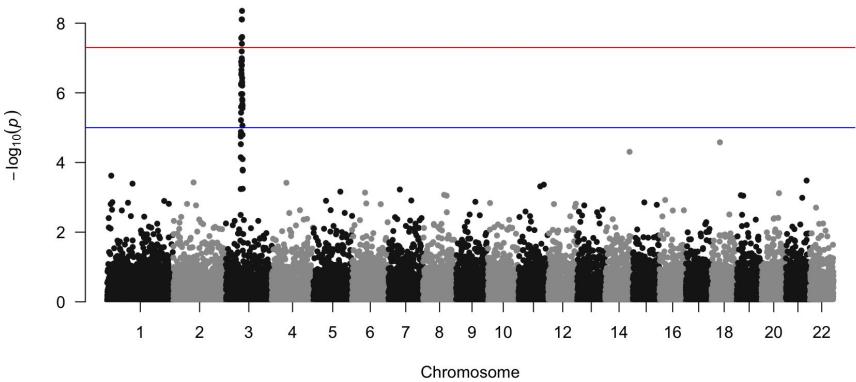


# PCAdapt

The next step is to compute the test statistic based on the PCA

Here, the test statistic is based on the z-scores that are obtained when the SNPs are regressed against the K PCs – essentially, the method identifies outlier SNPs as those that are more distant from the mean based on the z-score distribution

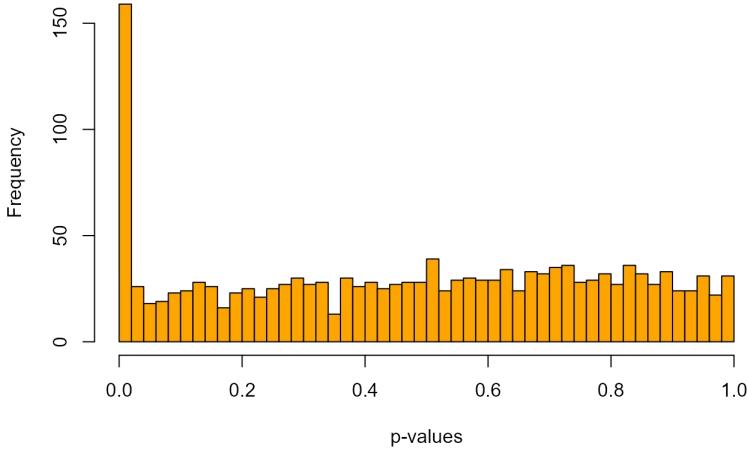
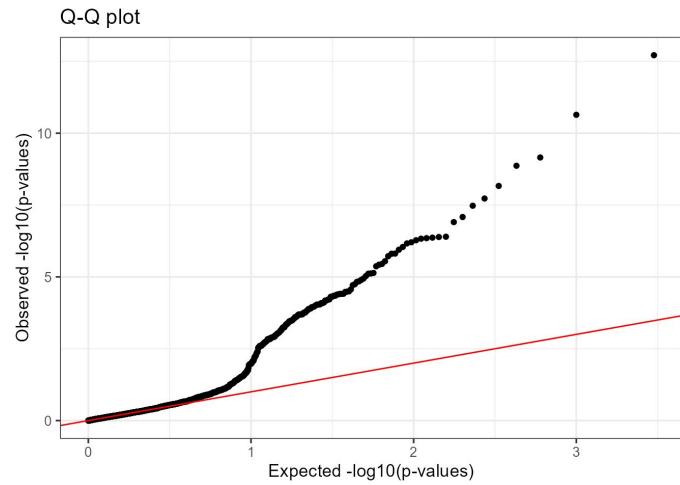
Results are plotted using Manhattan plots  
(showing p-values at each SNP locus along the genome)



# PCAdapt

QQ-plots are used to show that most of the p-values follow a uniform distribution, but the smallest ones are smaller than expected (i.e., are outliers)

This can also be visualised in a histogram



# PCAdapt

Working out a threshold for p-value significance involves:

Transforming p-values into q-values and applying a multiple-testing correction, e.g.:

FDR, Benjamini-Hochberg, Bonferroni correction

# Case study 1

JOURNAL OF  
**Evolutionary Biology**

Issues More Content ▾ Submit ▾ Purchase Alerts About ▾ Journal of Evolution ▾



Volume 37, Issue 8  
August 2024

JOURNAL ARTICLE EDITOR'S CHOICE

## Historical museum samples reveal signals of selection and drift in response to changing insecticide use in an agricultural pest moth

Elahe Parvizi, Andy Bachler, Andreas Zwick, Tom K Walsh, Craig Moritz, Angela McGaughran 

*Journal of Evolutionary Biology*, Volume 37, Issue 8, August 2024, Pages 967–977, <https://doi.org/10.1093/jeb/voae068>

# Case study 1

*Helicoverpa armigera* – a significant pest of cotton and other crops

Highly polyphagous, strong flier/disperser, rapid spread of insecticide resistance

254 museum samples spanning > 100 years, pre1950s (before insecticides)  
compared to post-1950s decades, when insecticides were being used in Australia

Exon capture with focus on insecticide resistance gene families

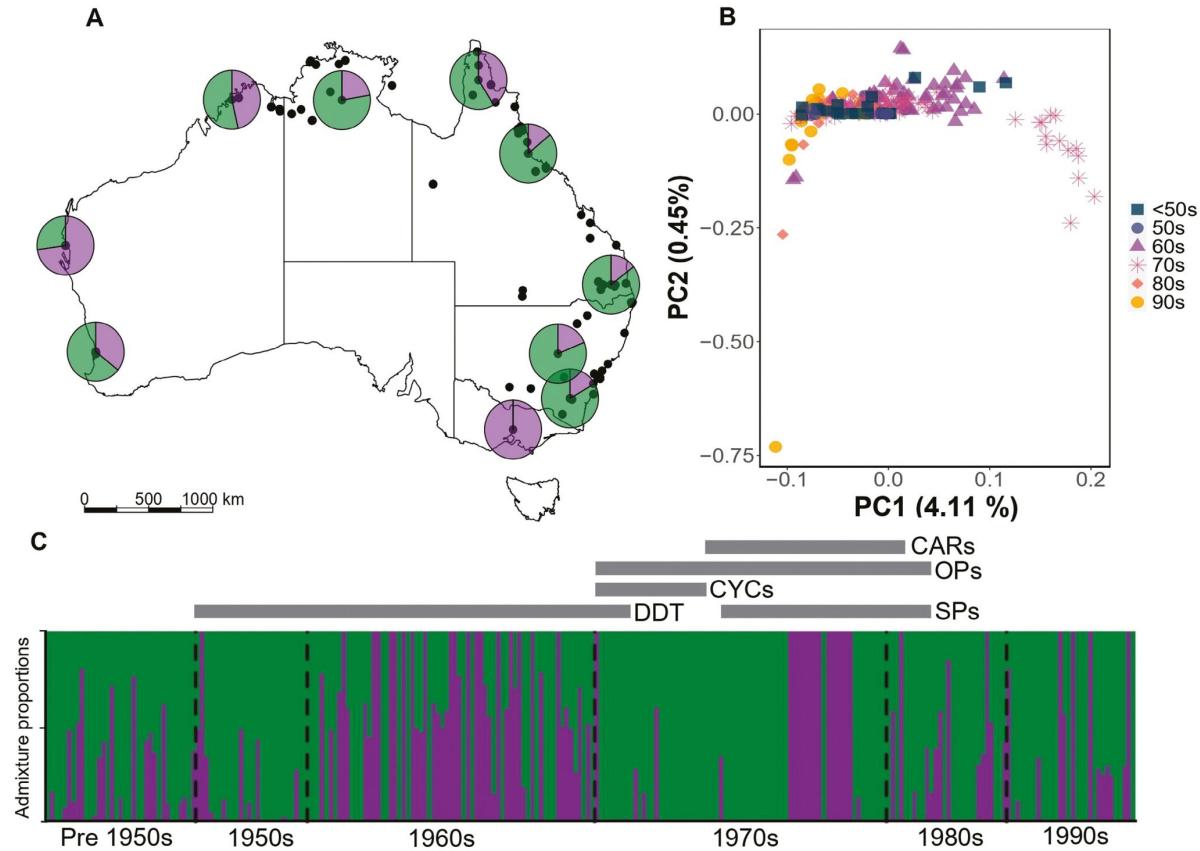


# Case study 1

No real structure except for 1970s/WA

Some signs of reduced gene flow following use of pesticides?

What would you expect to see?



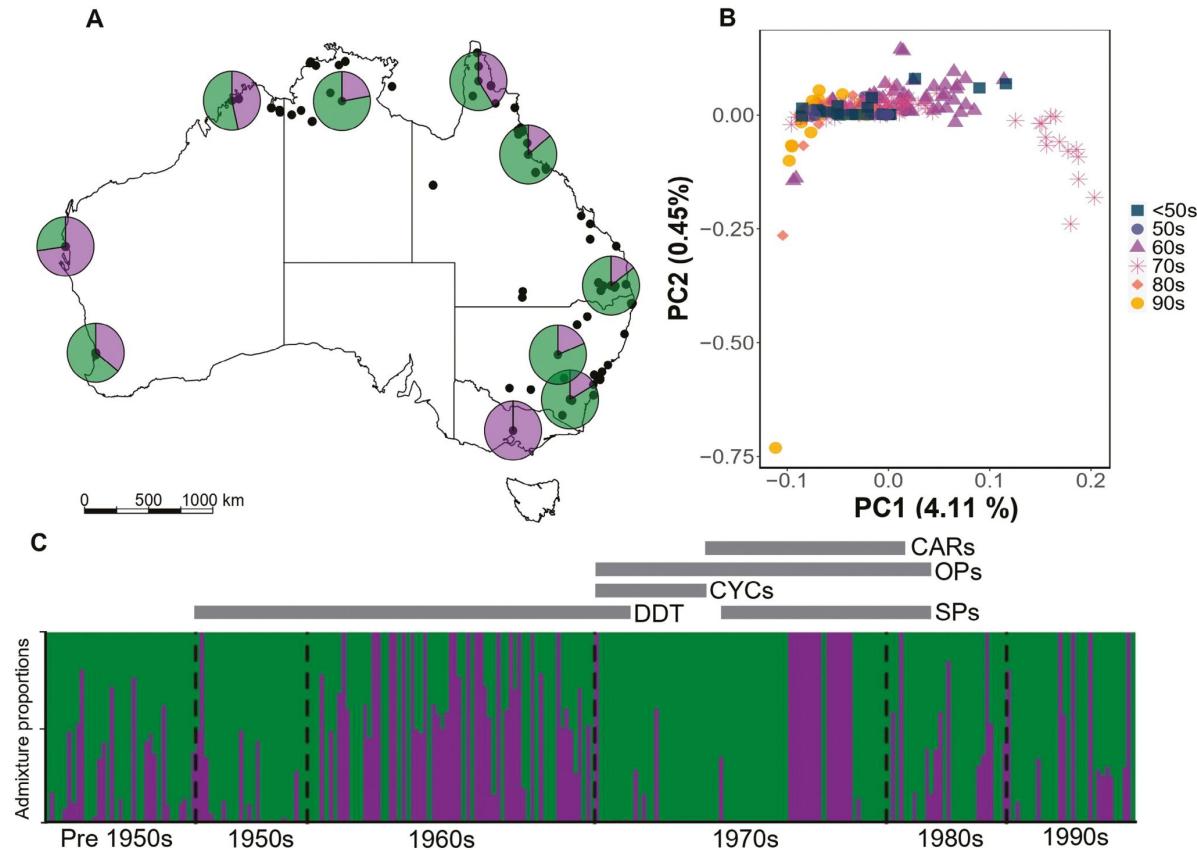
# Case study 1

No real structure except for 1970s/WA

Some signs of reduced gene flow following use of pesticides?

What would you expect to see?

Effects of demography and selection



# Case study 1

Investigating signatures of recent positive selection

PCAdapt, two PCs, MAF of 0.05, Bonferroni correction to the *p*-values, setting a false discovery rate of 5%; ran overall (no population inference) and in a population-specific manner to identify outliers in each decade

Also used an  $F_{ST}$  outlier approach, defining outliers as those with  $F_{ST}$  higher than three standard deviations above the mean for the given population comparison

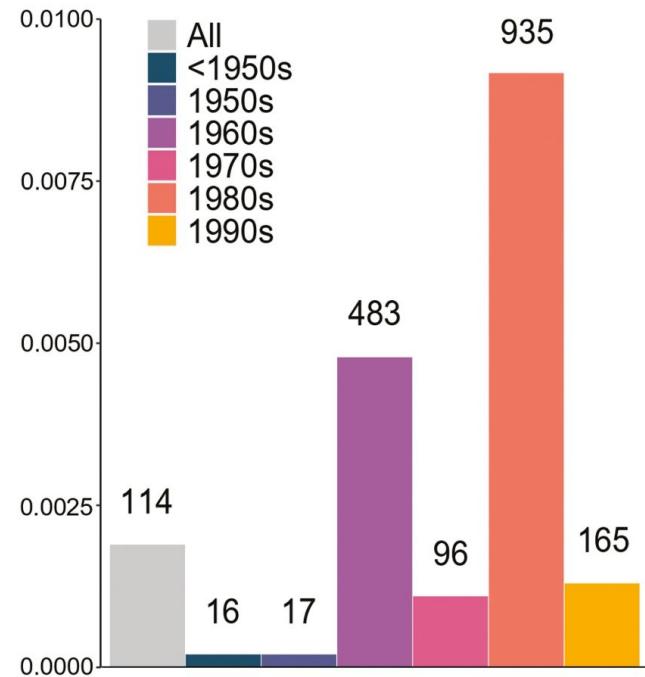
For both methods, counted number of times outlier SNPs were present in each decade, and in each functional gene group; also annotated SNP outliers that were common to both methods to generate candidate gene lists per decade

# Case study 1

Significant variation in number of outliers per decade identified with PCAdapt

Highest in the 1960s and 1980s

Lower in pre1950s samples and 1950s

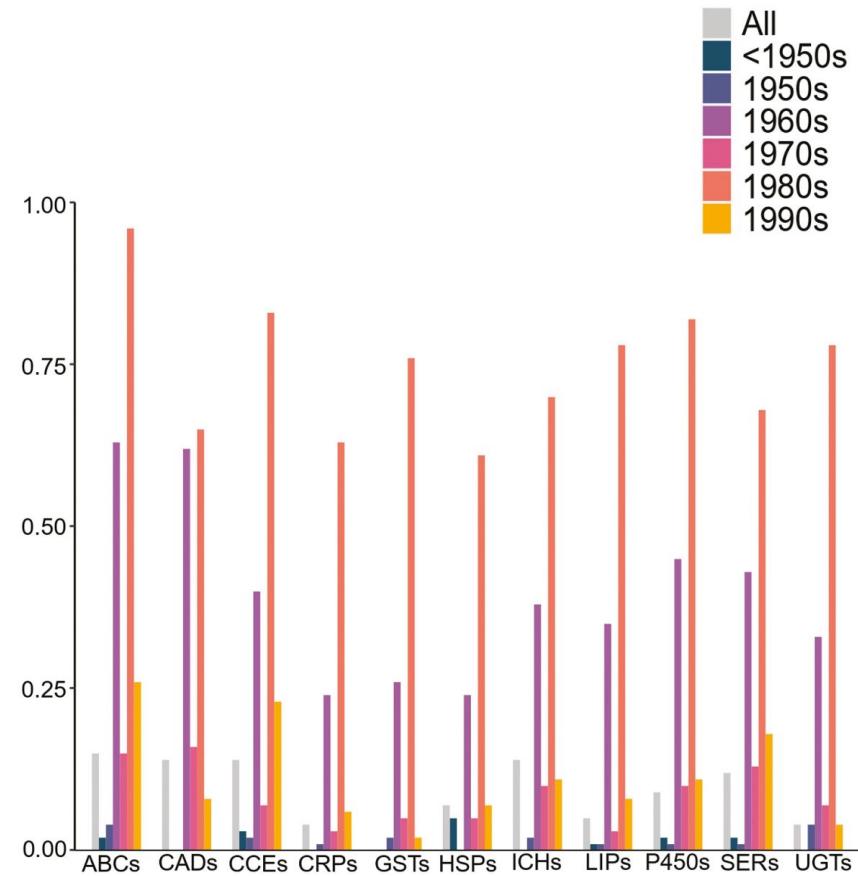


# Case study 1

Same pattern for PCAdapt outliers in each functional gene category across time

Low numbers of outliers in samples from the pre1950s and 1950s

Much higher in 1960s and 1980s



# Case study 1

Overlap with  $F_{ST}$  and annotation:

Almost all of the PCAdapt outliers were represented in  $F_{ST}$  outlier analysis

Annotation of the outliers common between the two methods identified highly relevant gene functions e.g., gustatory and odourant receptors, cytochrome P450, heat shock proteins ...

# Case study 1

Implications:

Temporal genome scans showed extensive evidence of selection across decades

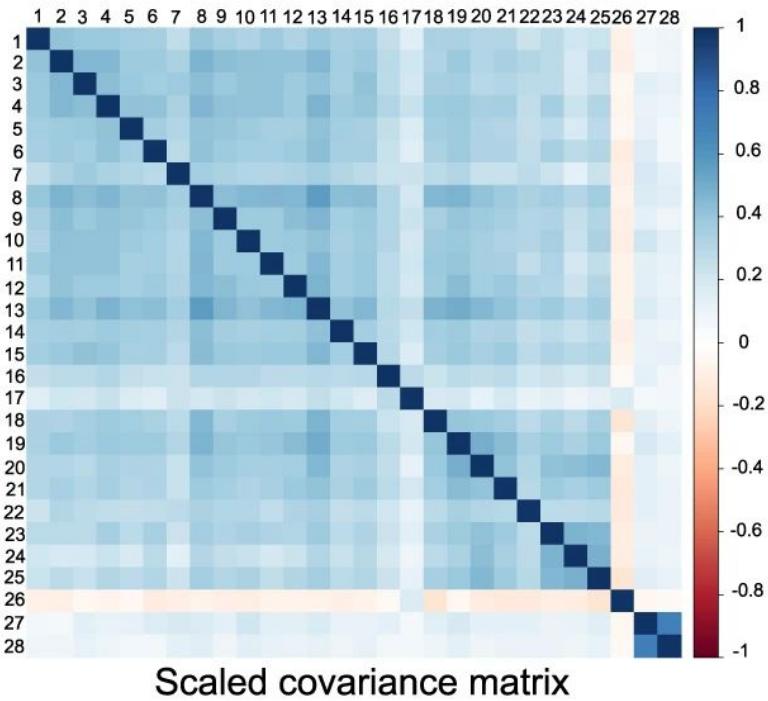
Annotation of outliers to functionally important genes suggests these processes may have facilitated rapid adaptive responses among moth populations to intense insecticide-driven pressures over time

# BayPass

See [https://forge.inrae.fr/mathieu.gautier/bypass\\_public](https://forge.inrae.fr/mathieu.gautier/bypass_public)

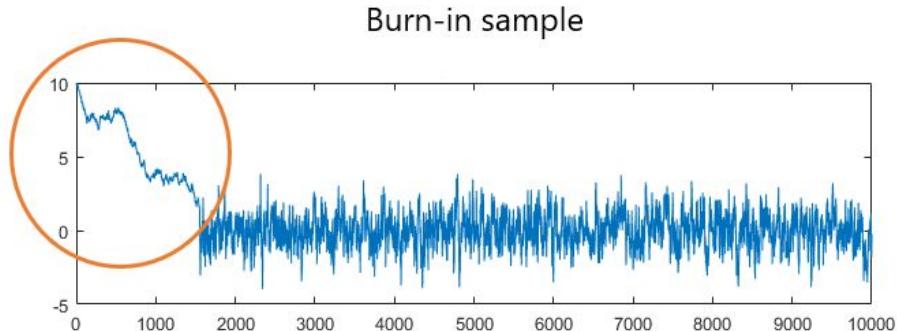
Identifies genetic markers subject to selection (and/or associated with population-specific covariates like environmental or phenotypic data)

The model explicitly accounts for (and estimates) the covariance structure among population allele frequencies due to shared demography and history of the populations



# BayPass

Bayesian method that uses a Monte Carlo Markov Chain (MCMC) algorithm



Should be run at least three times independently, with different runs converging on same values i.e., the estimated model hyper-parameters should be highly consistent across multiple runs

Support for association of each SNP with the corresponding covariate evaluated using the median Bayes Factor (BF) for  $XtX$  statistic

$BF\ 20 = 100:1$  odds,  $30 = 1,000:1$ ,  $40 = 10,000:1$ , etc.

# BayPass

In the tutorial, we are using the C2 statistic of BayPass to estimate contrasts between pairwise populations → a very nice use for invasive/native population comparisons; the software for C2 analysis simply requires

The allele frequency data in BayPass format

The contrasts file to set up the two populations that should be contrasted

The output file from BayPass then specifies the C2 statistic and log10 p-value for each SNP; these can be nicely plotted with Manhattan plots and different contrasts can be tested (e.g., to explore parallel selection patterns for different population pairs)

# BayPass

Thus, our use of BayPass (C2 statistic) compares the allele frequencies (corrected for population structure) between the two groups of populations specified by the binary covariate of interest (e.g., ‘invasive’, ‘native’)

But note that BayPass can also be used in other ways (XtX statistic), e.g.:

A core model to assess population structure and identify outlier SNPs

A covariate model relating outliers to environmental data

# Case study 2

## Current Biology



Volume 28, Issue 20, 22 October 2018, Pages 3296-3302.e7

Report

### The Genomic Basis of Color Pattern Polymorphism in the Harlequin Ladybird

Mathieu Gautier<sup>1 15</sup>, Junichi Yamaguchi<sup>2 15</sup>, Julien Foucaud<sup>1</sup>, Anne Loiseau<sup>1</sup>,  
Aurélien Ausset<sup>1</sup>, Benoit Facon<sup>1 10</sup>, Bernhard Gschloessl<sup>1</sup>, Jacques Lagnel<sup>1 11</sup>,  
Etienne Loire<sup>1 12 13</sup>, Hugues Parrinello<sup>3</sup>, Dany Severac<sup>3</sup>, Celine Lopez-Roques<sup>4</sup>,  
Cecile Donnadieu<sup>4</sup>, Maxime Manno<sup>4</sup>, Helene Berges<sup>5</sup>, Karim Gharbi<sup>6 14</sup>, Lori  
Lawson-Handley<sup>7</sup>, Lian-Sheng Zang<sup>8</sup>, Heiko Vogel<sup>9</sup>, Arnaud Estoup<sup>1 16</sup>   
Benjamin Prud'homme<sup>2 16 17</sup>

## Case study 2

*Harmonia axyridis* lady beetle commonly known as harlequin, Asian, or multicoloured Asian lady beetle

One of the most variable lady beetle species, with exceptionally wide range of colour forms (>200 described) via allelic variation at a single, unknown locus)

Native to eastern Asia; introduced to other areas as a biocontrol agent for aphids and scale insects, now widely established globally



## Case study 2



Four dominant colour forms in natural populations at high frequencies: three melanic forms with patterns Black-2Spots, Black-4Spots, and Black-nSpots, and a non-melanistic form, called Red-nSpots

Collected individuals from eight geographic locations and carried out genome scans studies → association with the proportion of individuals of a given colour pattern form in pooled samples

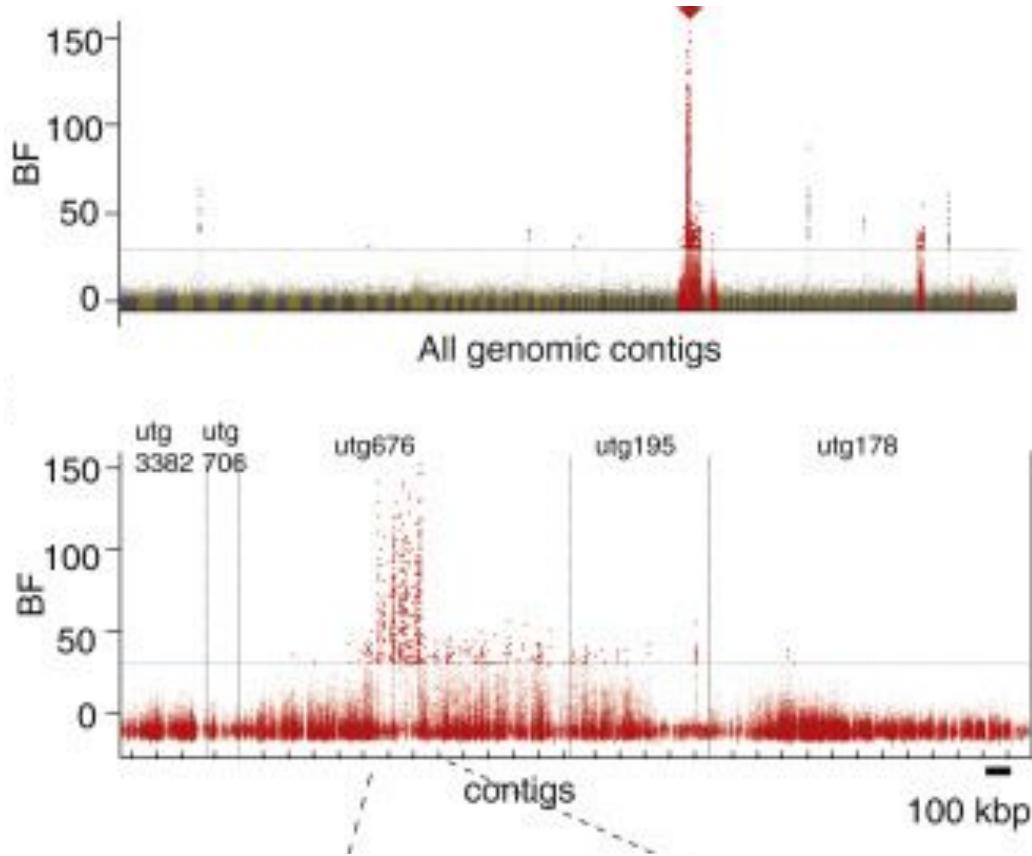
Took one SNP every 200 SNPs along the genome

Analysed with BayPass core model with outliers (associated with colour prevalence covariate) identified via median Bayes Factors (BF) over three independent runs

## Case study 2

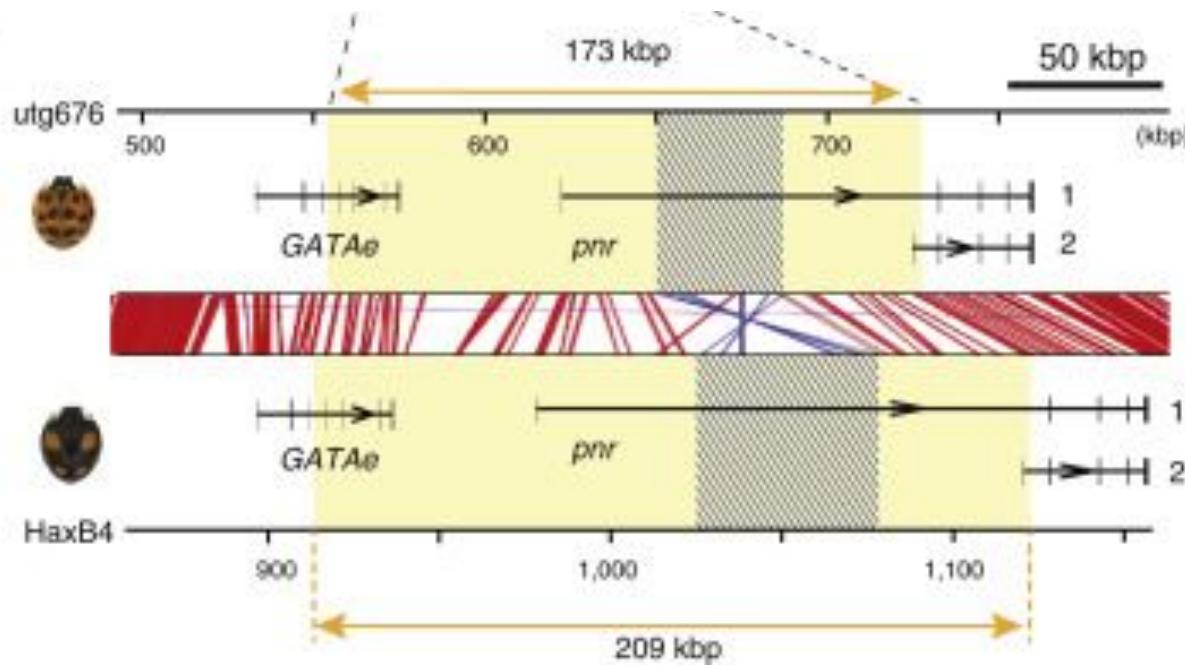
Found 710 (out of 18,425,210) SNPs strongly associated with proportion of Red-nSpots form, 86% of which located within a single 1.3 Mb contig

Same region identified for Black-4Spots, Black-2Spots, or Black-nSpots individuals



## Case study 2

*HaxR* gene strongest candidate for the colour pattern locus (orthologue of the *Drosophila* gene *pannier*)

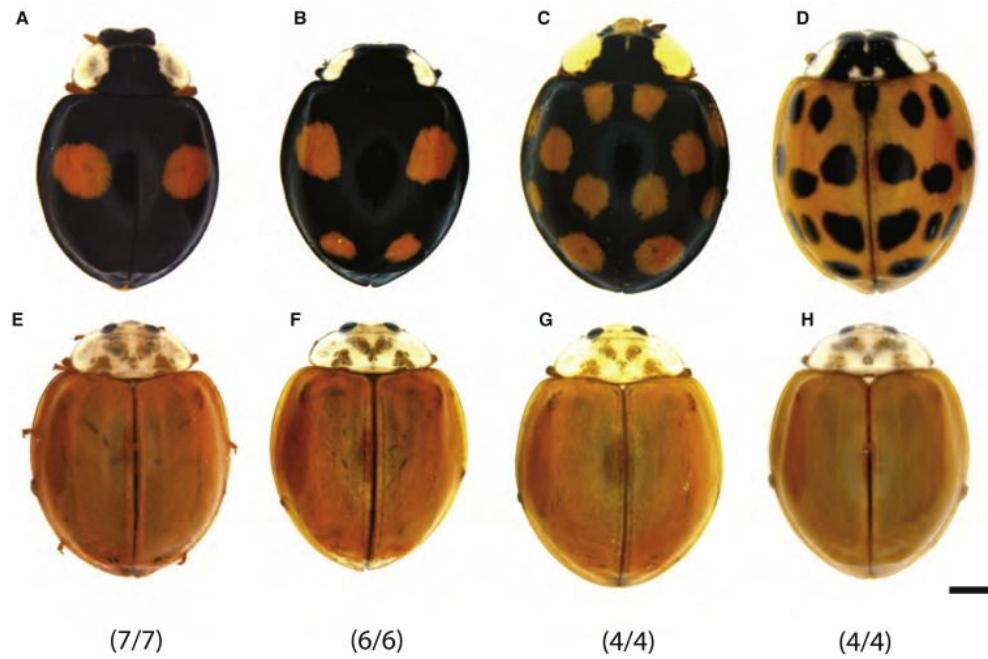


# Case study 2

Functional follow-up

Knockdown of *pannier* dramatically reduced the formation of black pigment in all four forms

Adults almost homogeneous red across most of the body



# Case study 3



► Mol Biol Evol. 2020 Apr 17;37(8):2369–2385. doi: [10.1093/molbev/msaa098](https://doi.org/10.1093/molbev/msaa098) ↗

## A Whole-Genome Scan for Association with Invasion Success in the Fruit Fly *Drosophila suzukii* Using Contrasts of Allele Frequencies Corrected for Population Structure

Laure Olazcuaga<sup>m1</sup>, Anne Loiseau<sup>m1</sup>, Hugues Parrinello<sup>m2</sup>, Mathilde Paris<sup>m3</sup>, Antoine Fraimout<sup>m1</sup>, Christelle Guedot<sup>m4</sup>, Lauren M Diepenbrock<sup>m5</sup>, Marc Kenis<sup>m6</sup>, Jinping Zhang<sup>m7</sup>, Xiao Chen<sup>m8</sup>, Nicolas Borowiec<sup>m9</sup>, Benoit Facon<sup>m10</sup>, Heidrun Vogt<sup>m11</sup>, Donald K Price<sup>m12</sup>, Heiko Vogel<sup>m13</sup>, Benjamin Prud'homme<sup>m3</sup>, Arnaud Estoup<sup>m1,✉,#</sup>, Mathieu Gautier<sup>m1,✉,#</sup>

## Case study 3

*Drosophila suzukii*

Native to Southeast Asia, initially invaded N. America and Europe in 2008, then La Réunion Island (Indian Ocean) and South America in 2013

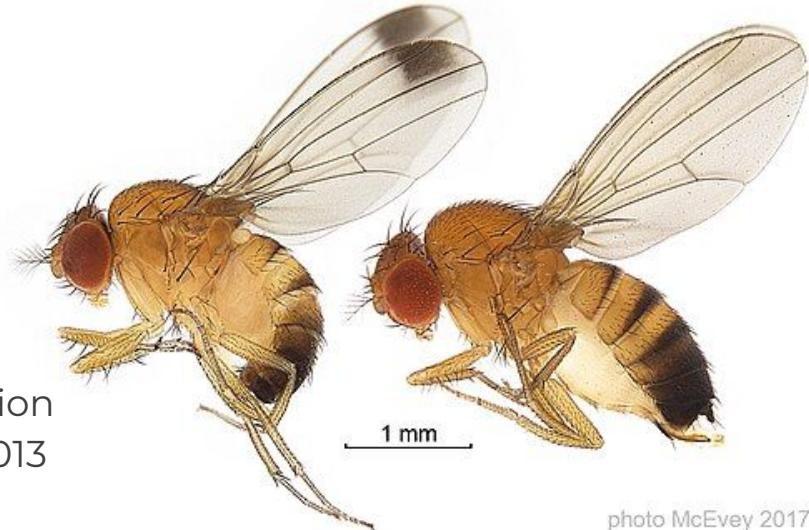


photo McEvey 2017

Key advantage compare to most Drosophilids → can lay eggs in unripe fruits

Dramatic losses in fruit production in agricultural areas, annual costs exceeding one billion euros globally

Rapid spread suggests strong adaptive responses

# Case study 3

Genome-wide association study to identify genes involved in adaptation during invasion

22 populations

C2 statistic, contrasting  
allele frequencies between  
invasive versus native

N. American and European  
populations separate invasions,  
also analysed as two independent  
invasion replicates

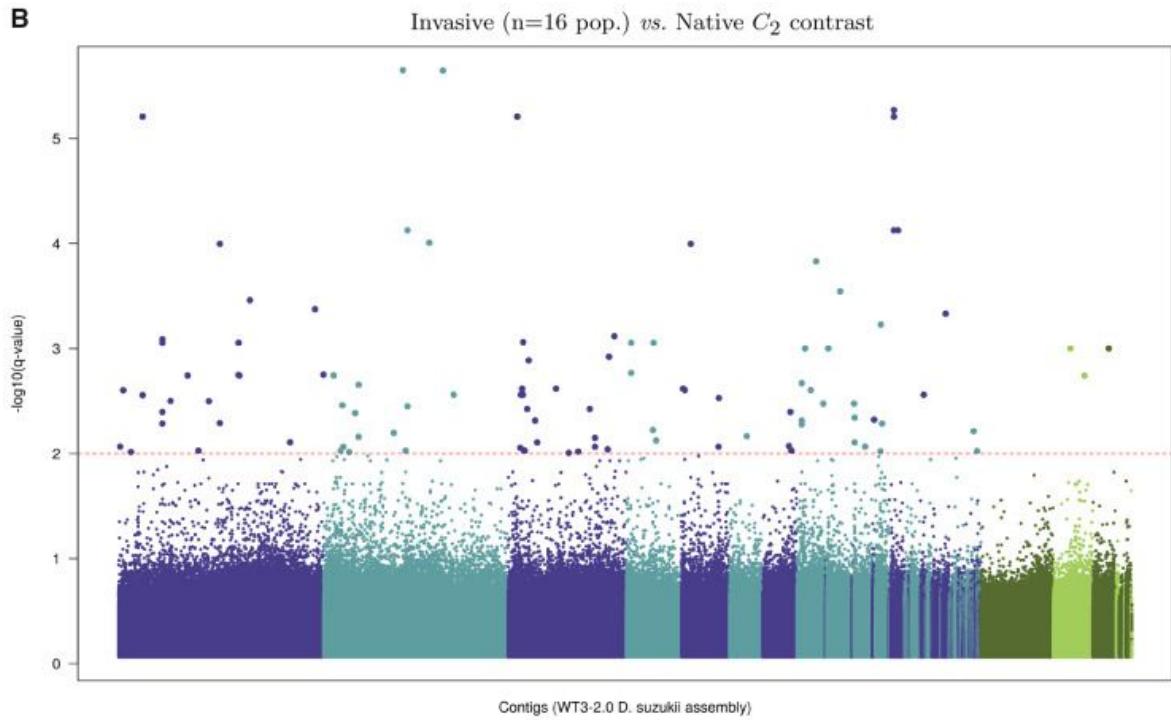


# Case study 3

204 outlier SNPs

Lack of overall clustering  
of outlier SNPs → low LD,  
consistent with large  
effective populations size?

Not unexpected since  
invasion success not the  
only selective constraint  
acting on these global  
populations



## Case study 3

C<sub>2</sub> statistic for the two separate invasion routes (C<sub>2</sub>EU, C<sub>2</sub>AM → vs rest of the pops = C<sub>2</sub>WW)

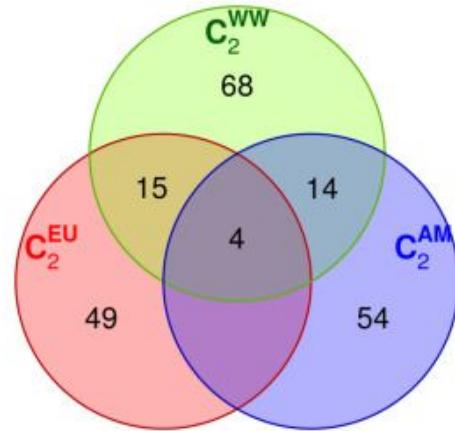
204 SNPs significant in at least one of the three contrasts

68 SNPs significant for the C<sub>2</sub>EU, 15 also significant for C<sub>2</sub>WW

72 SNPs significant for the C<sub>2</sub>AM, 14 also significant for C<sub>2</sub>WW

Majority of significant SNPs identified with either C<sub>2</sub>WW or C<sub>2</sub>AM specific to those invasion routes

Four SNPs common to all contrasts → strong candidates for association with global invasion success?



## Case study 3

Mapped 169 SNPs (out of the 204) onto 130 *D. melanogaster* genes

A subset of those genes associated with physiological functions and traits in *D. melanogaster*, but most lacked functional and phenotypic context

Two genes contained outlier SNPs in all three contrasts: *RhoGEF64C* (unknown function) and *cpo* (maybe related to diapause)

Median allele frequencies for the reference allele of *RhoGEF64C* SNP was 0.09 in the native populations and 0.93 and 0.87 in the two invasive populations; for *cpo*, median reference allele frequency 0.20 (native) and 0.99 (invasive)

Clearly important candidates!

# Case study 3

Implications:

A relatively small number of SNPs associated with invasive status (selection is complex!)

Independent contrast analyses of the two main invasion routes identified different subsets of outlier SNPs:

Source populations and/or invasion process differ in the two invaded areas

Polygenic nature of the traits underlying invasion success?

# Common approaches

Common to run multiple outlier scans with different assumptions/algorithms → we will run PCAdapt and BayPass

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 (Day Three)

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

# Genomic insights

Understanding rapid evolution at a mechanistic level by identifying the genes responding to selection

Predicting spread and impact

Informing management strategies

# Future directions

Generating adaptive datasets increasingly achievable, even for temporal contrasts

Broader, more comprehensive sampling

Adaptive databases? Machine learning to detect complex selection patterns?

Real-time and/or rapid monitoring of adaptive evolution using portable sequencers

Management outcomes with gene editing?

Genome-informed strategies

# Food for thought

What are the implications of rapid adaptation for biosecurity and control?

How can we confidently distinguish between selection and drift in invasive populations?

Can we use genomic data to predict which species will adapt successfully in new environments?



## Conclusions

Adaptation is a key feature of invasion success

Genomic data can identify areas of the genome under selection; with a relevant reference, these outliers can be annotated to identity candidate invasion genes

# Next steps

Please move to the Day Two tutorial

Remember to ask questions via slack!

**CREDIT:** SlideEgg PowerPoint template

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

