Selection & Adaptation

Leonie Moyle
lmoyle@indiana.edu

Workshop on Population and Speciation Genomics, Cesky Krumlov 2022

Selection and Adaptation: Today

A. Context: What is selection, what is adaptation?

B. Detecting selection: within populations

- sequence-based tests of selection
- association studies

C. Detecting selection: between populations

- outlier analyses
- environmental association analyses

D. Case study: Landscape genomics of adaptation to abiotic climate

E. Questions/chat

goals

(from the perspective of an end user, e.g., me)

broad overview of

general rationale underlying empirical tests of selection (and adaptation)

inferential structure of (some) tests of selection/adaptation, at varying scales

(some) factors that can mislead genomic inferences

(some) practical considerations for sampling and experimental decisions

Selection and Adaptation

the evolutionary force that maintains or increases the frequency of variants that contribute to fitness

(classically) a consistent difference in survival and/or reproduction among individuals that differ in one or more traits (alleles)

Flavours of natural selection

In a perfect world, depending upon the variant, selection:

‘directional’ selection

- removes deleterious (fitness reducing) mutations

  ‘negative’ or ‘purifying’ selection

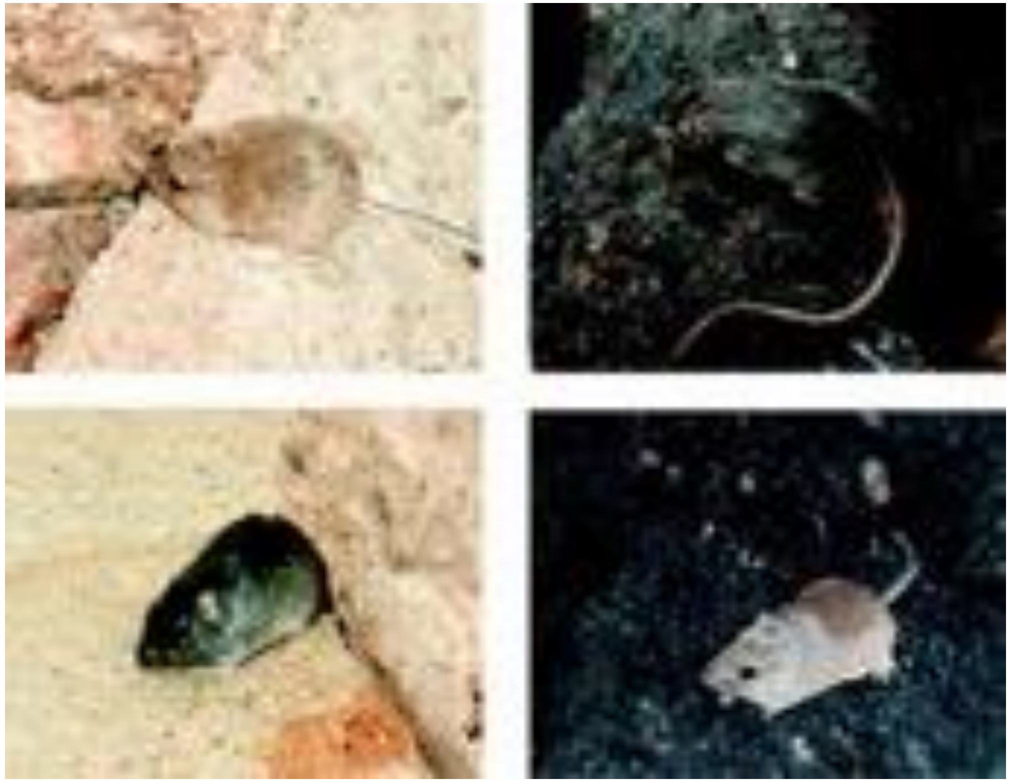
- promotes advantageous (fitness enhancing) mutations

  ‘positive’, or ‘divergent’, selection

- maintains advantageous (fitness enhancing) variation

  ‘balancing’ or ‘diversifying’ selection

Selection and Adaptation



the product of fitness-enhancing selection

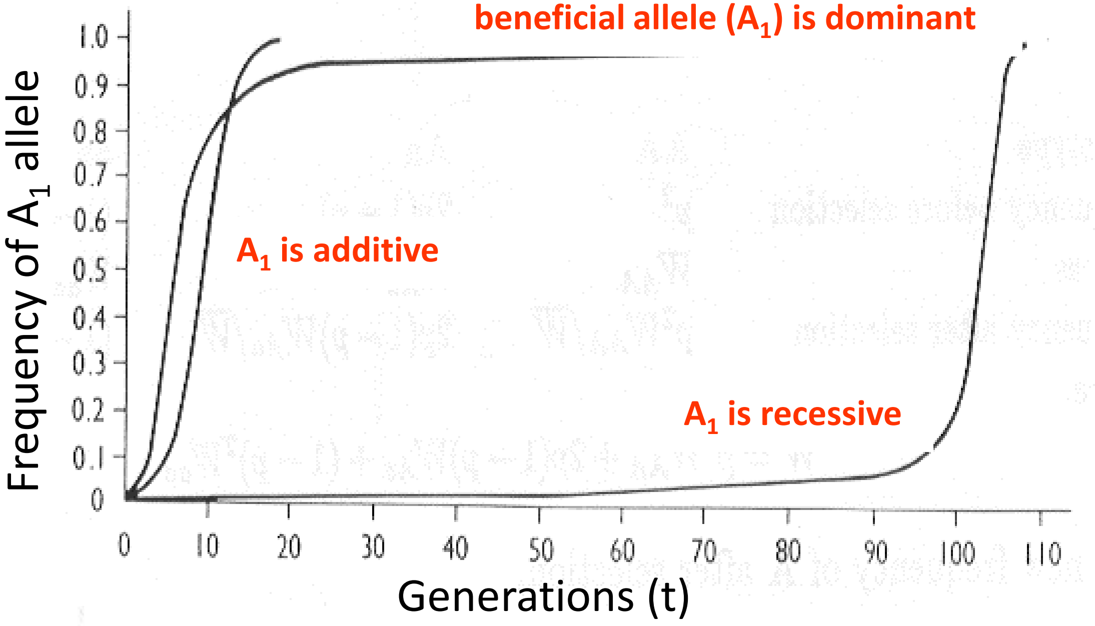
adaptation: a trait or characteristic that increases survival and/or reproduction in a given environment

the process of evolutionary change whereby a lineage of organisms increases in average fitness (within an environmental context)

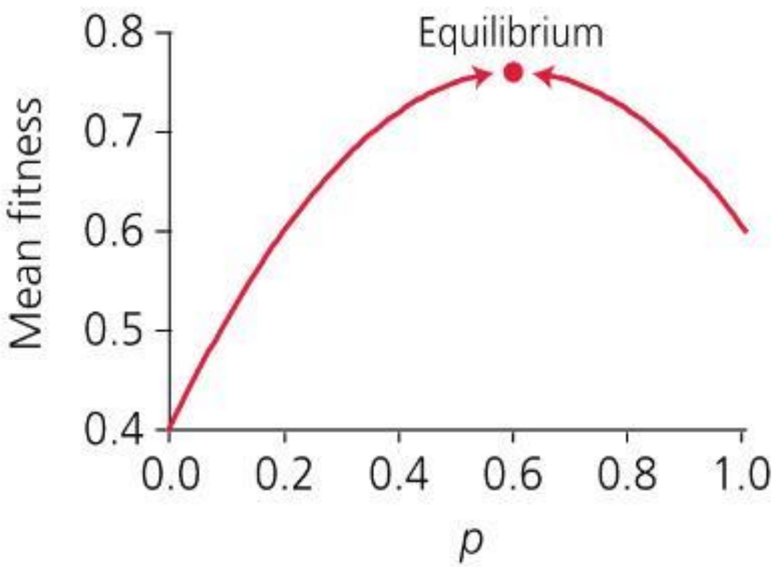
Single-locus models of selection

(e.g., selection on single SNPs)

$\hat{p}_A = 1 \text{ or } 0$



$\hat{p}_A = t / (s + t)$



at equilibrium

Selection on quantitative traits

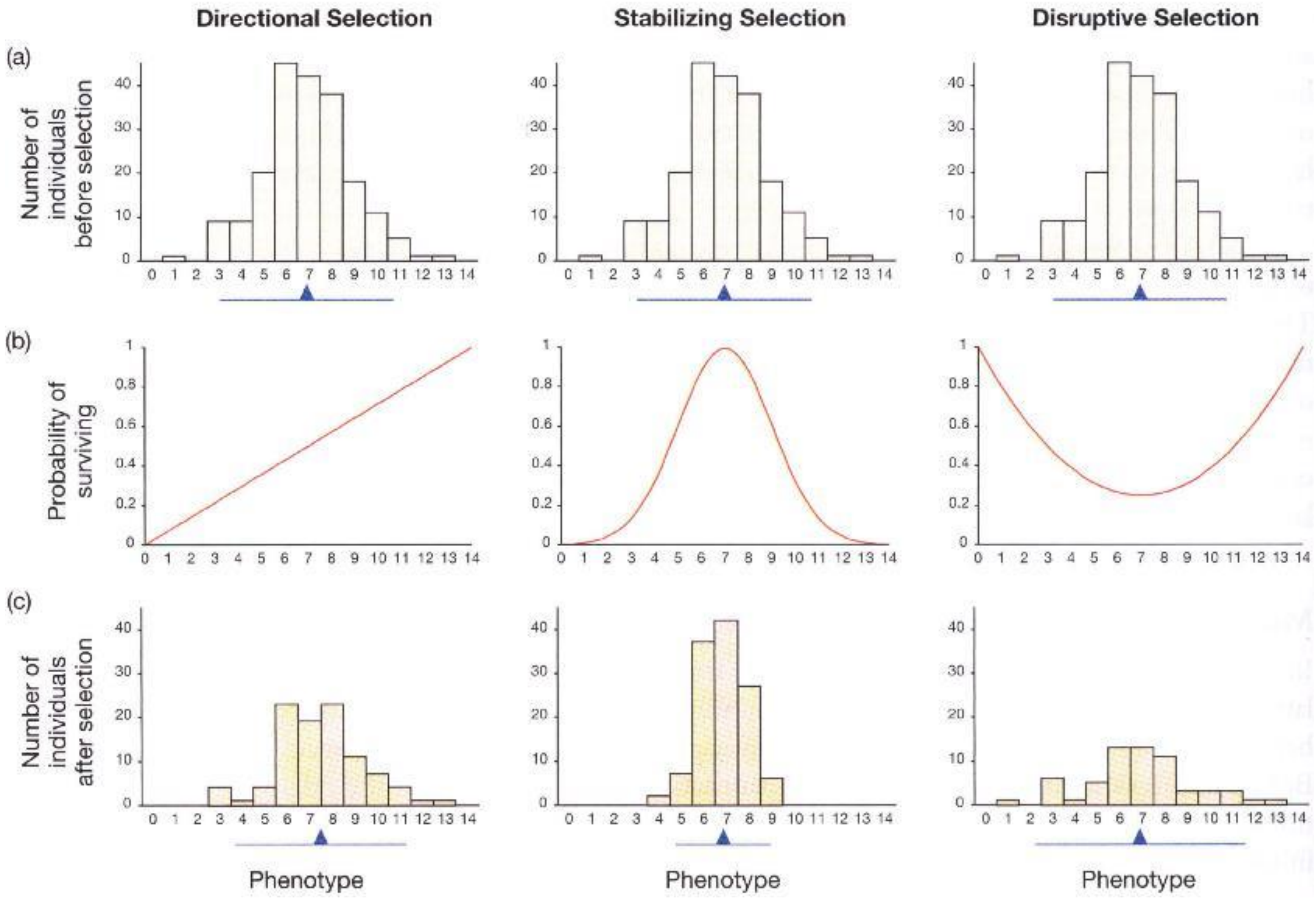
Starting phenotype

↓

Selection gradient

↓

Change in phenotype



why study selection and adaptation?

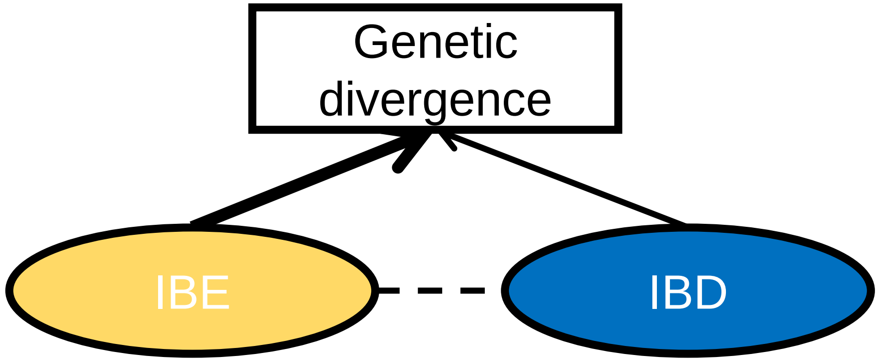
“Mechanisms”

- Genetic basis

- Key selective agents (ecological forces)

Interactions with other forces

- Relative importance compared to other evolutionary processes (geographic isolation, demographic history, relatedness, etc.)



Genetic divergence

IBE

IBD

why study selection and adaptation?

the genetic process of adaptation

- what distribution and order of phenotypic effects, rate over time?

- what is the genetic architecture underlying adaptations?

  - simple versus complex genetic basis
  - few versus many genes, allelic effects, epistasis, etc.
  - distribution and average size of genetic effects

- what is the genetic source of adaptation? new mutation versus standing genetic variation (versus introgression)

“Theoretical”: to understand how evolution works (in nature)

why study selection and adaptation?

<u>ecological and evolutionary context</u>

- are there common patterns of selection and adaptation (across populations or species) with respect to demography, traits, or history?

- how does gene flow interact with local selection to shape genetic/adaptive responses?

- does (local) adaptation act in parallel across species or environments?

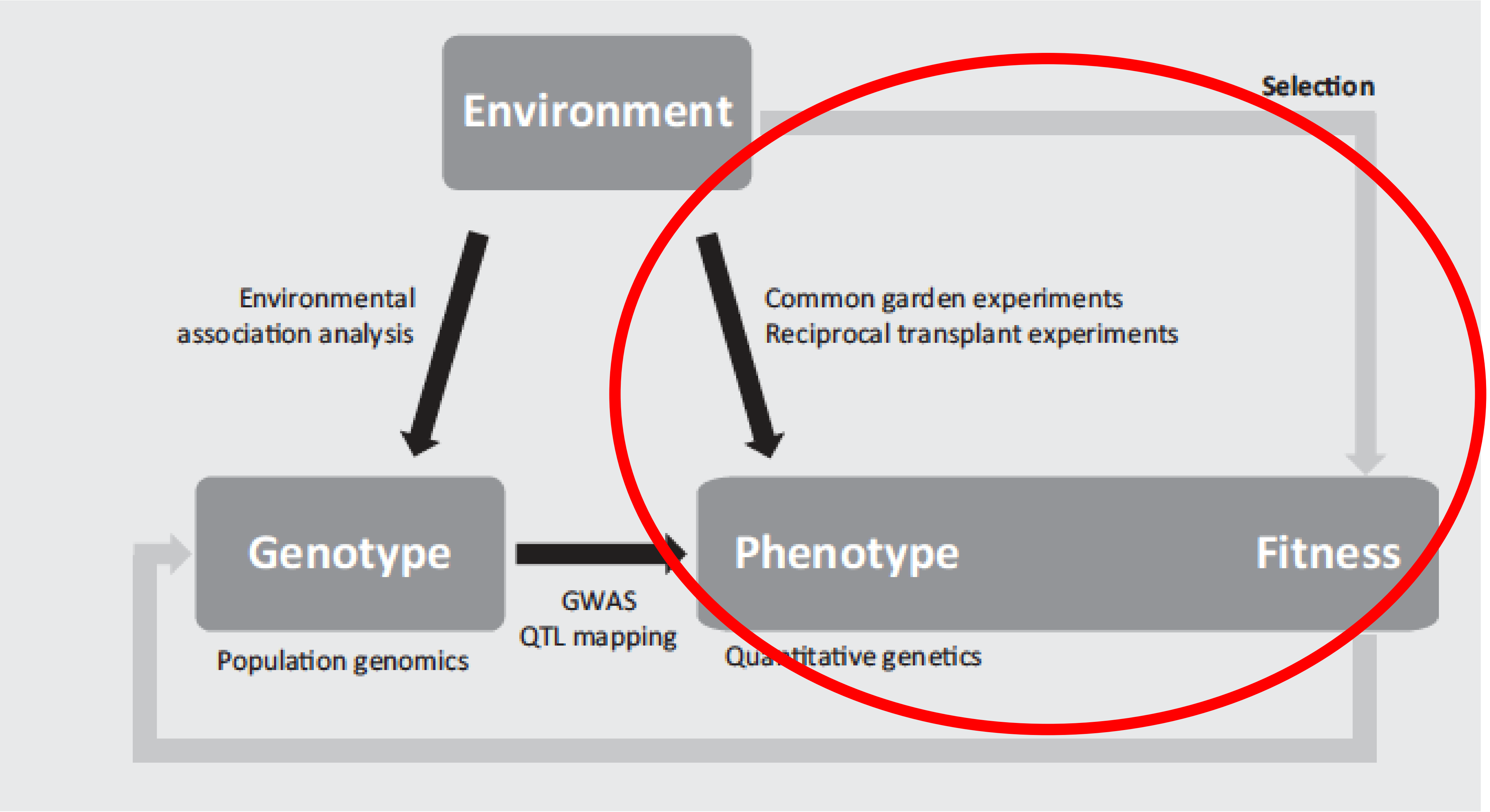
“Theoretical”: to understand how evolution works (in nature)

why study selection and adaptation?

- how is adaptive genetic variation distributed across a species range?
- what allows or constrains species range expansion/invasion?
- what genetic and ecological factors limit adaptation to future change (e.g. climate change)?
- what is the evolutionary potential of specific lineages or species?

Practical/applied: to understand, predict, manipulate populations

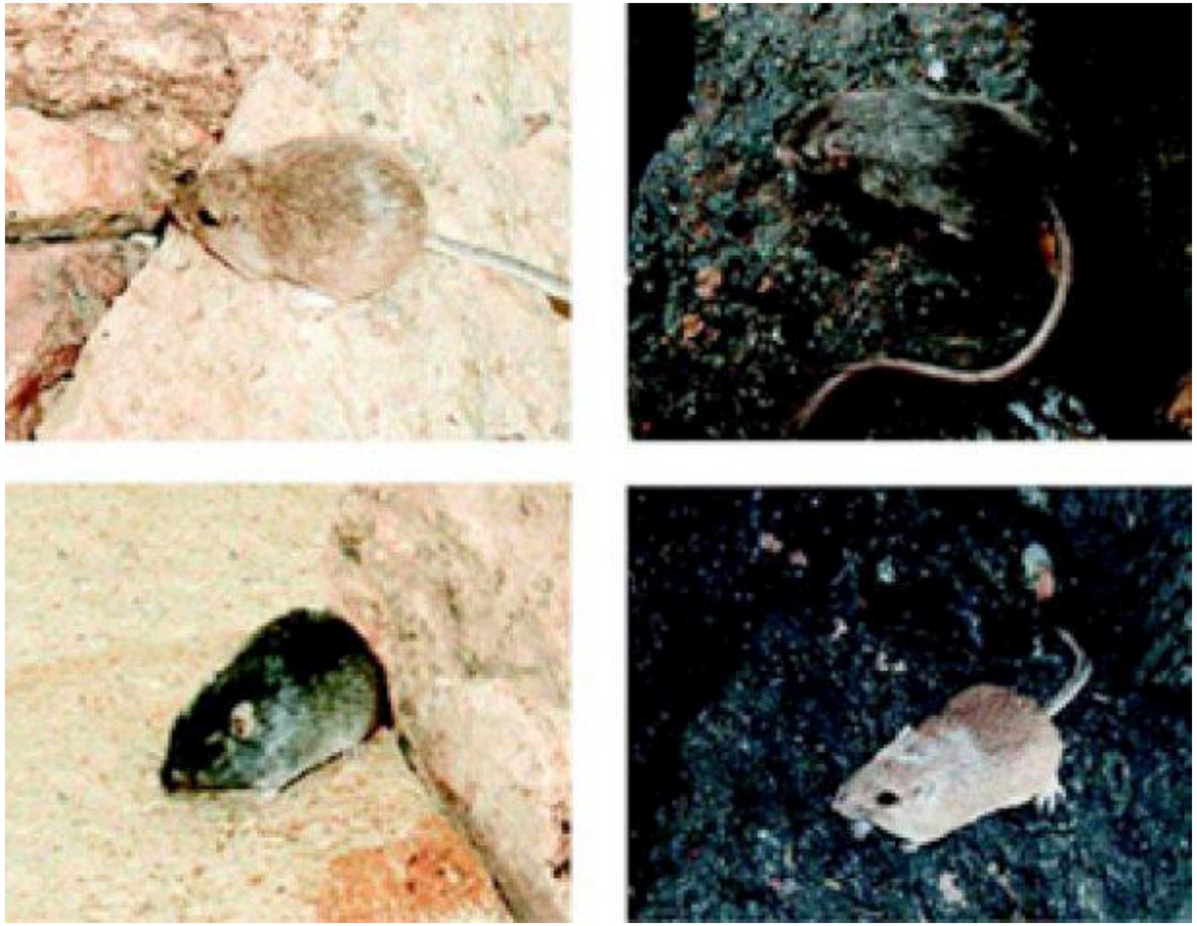
Detecting selection with pre-genomic data



(Rellstab et al. 2015)

Detecting selection with pre-genomic data

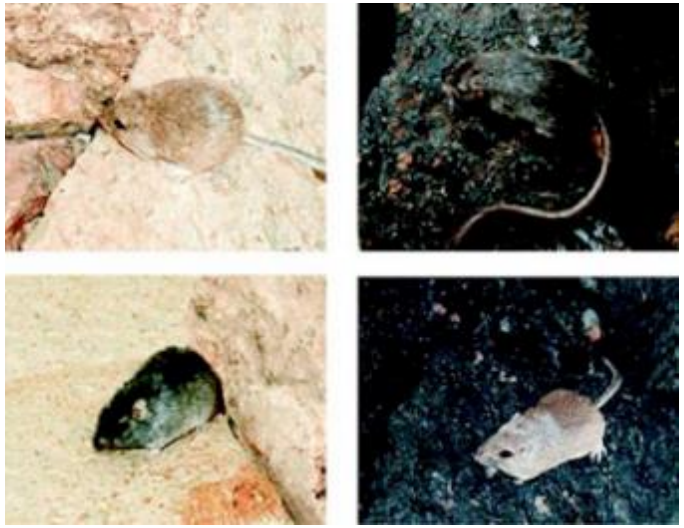
Trait-environment correlations /associations

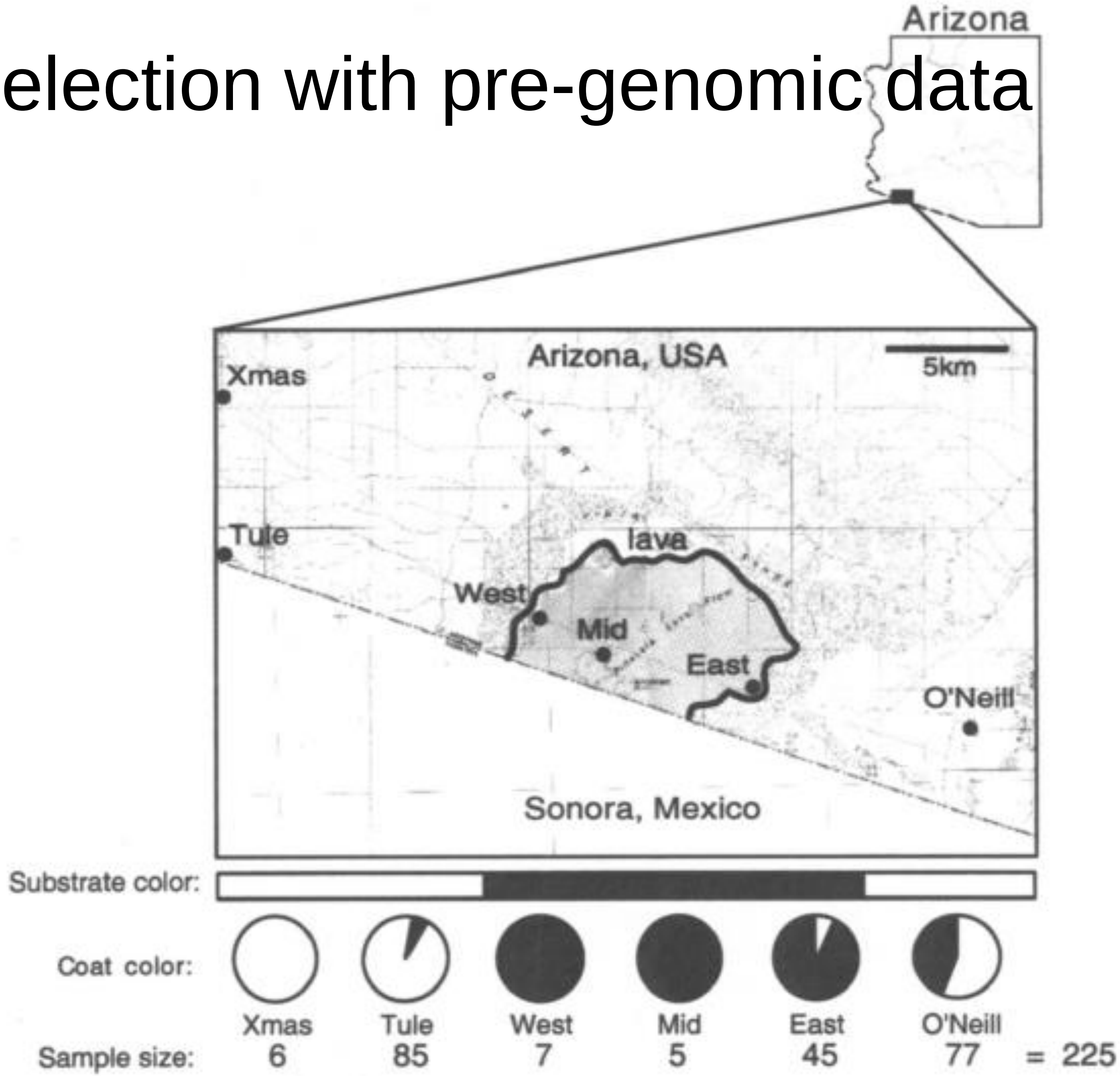


Classical evidence of adaptation

Detecting selection with pre-genomic data

Trait-environment correlations /associations





Hoekstra et al. 2004, Evolution

Detecting selection with pre-genomic data

Trait-environment correlations /associations

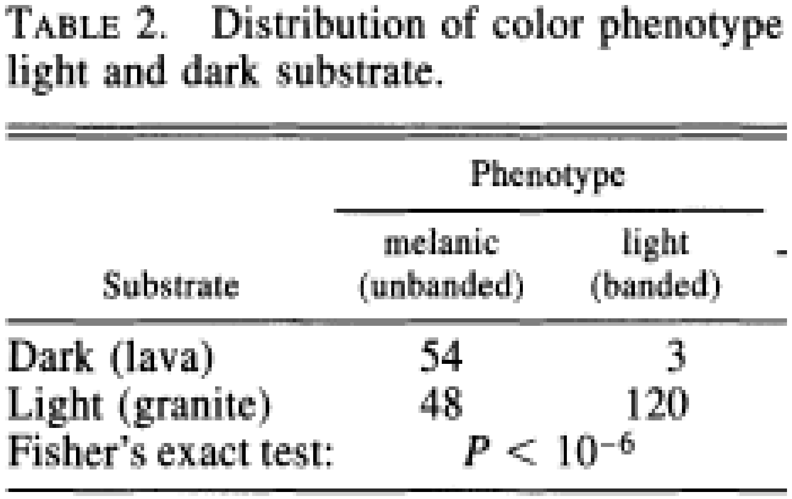
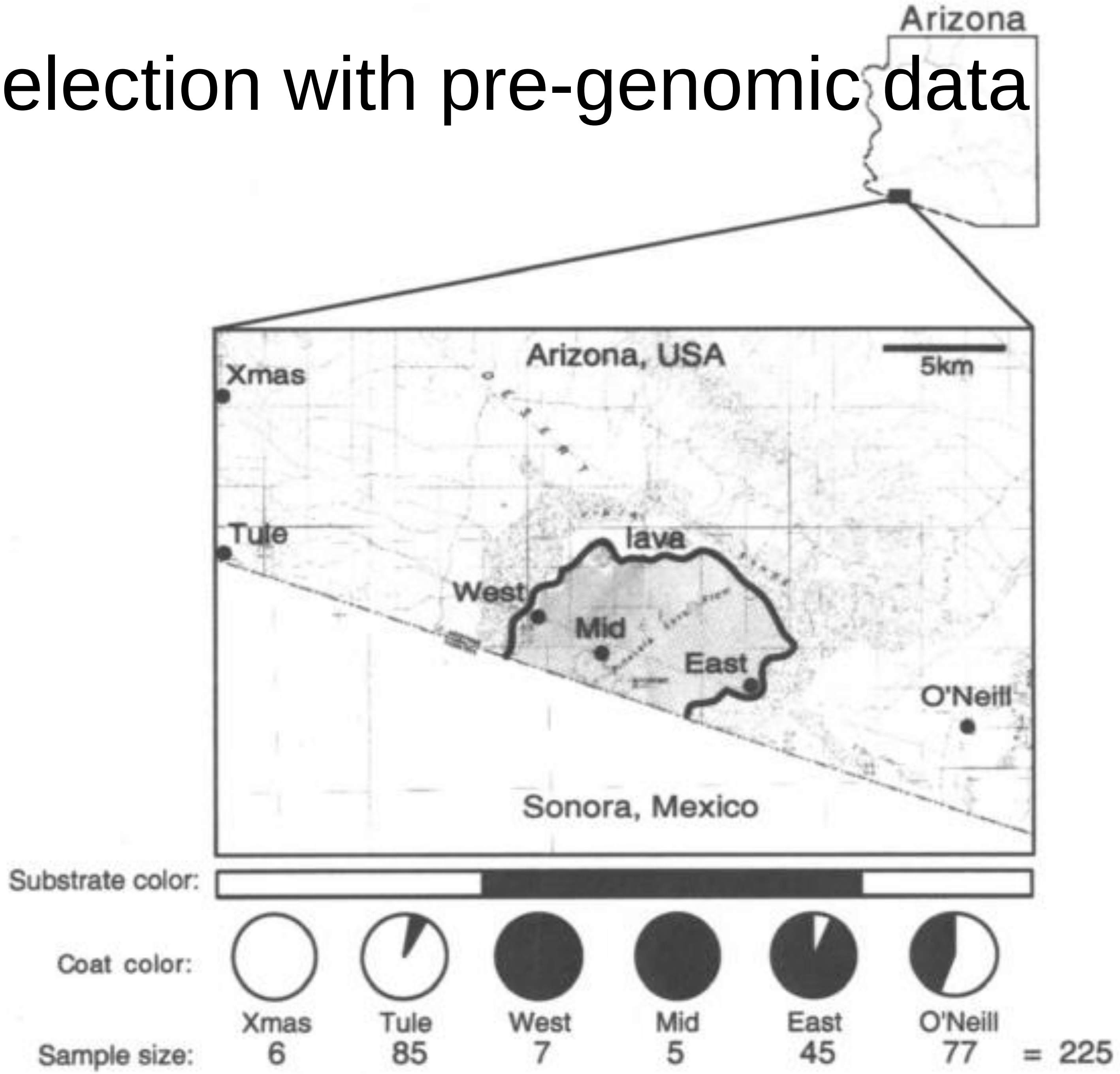


TABLE 2. Distribution of color phenotype light and dark substrate.

	Phenotype	
Substrate	melanic (unbanded)	light (banded)
Dark (lava)	54	3
Light (granite)	48	120
Fisher's exact test:	$P < 10^{-6}$	

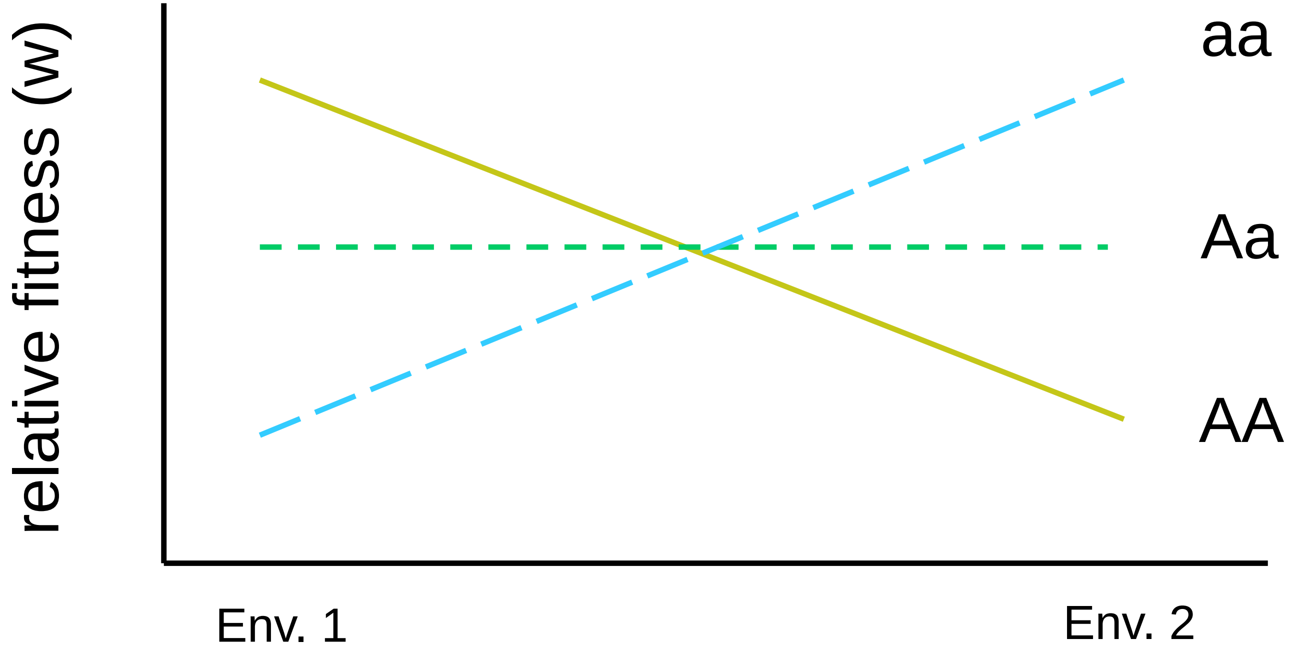


Hoekstra et al. 2004, Evolution

Detecting selection with pre-genomic data

Genotype x Environment Interaction

Change in relative **fitness** of genotypes across environments

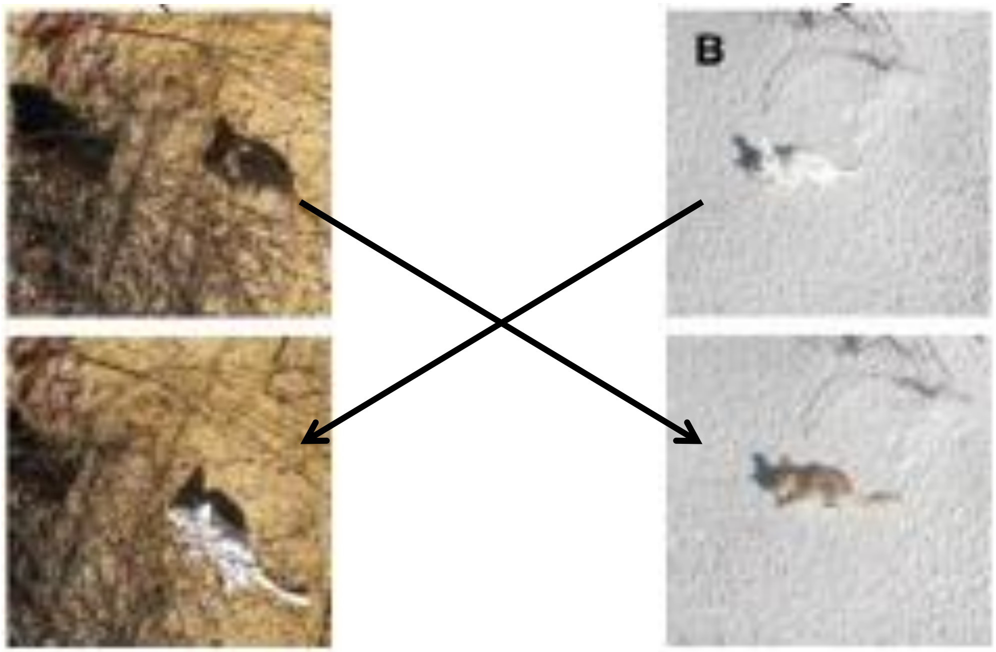


a crossing reaction norm for **fitness** == local adaptation

Detecting selection with pre-genomic data

Genotype x Environment Interaction

Change in relative **fitness** of genotypes across environments

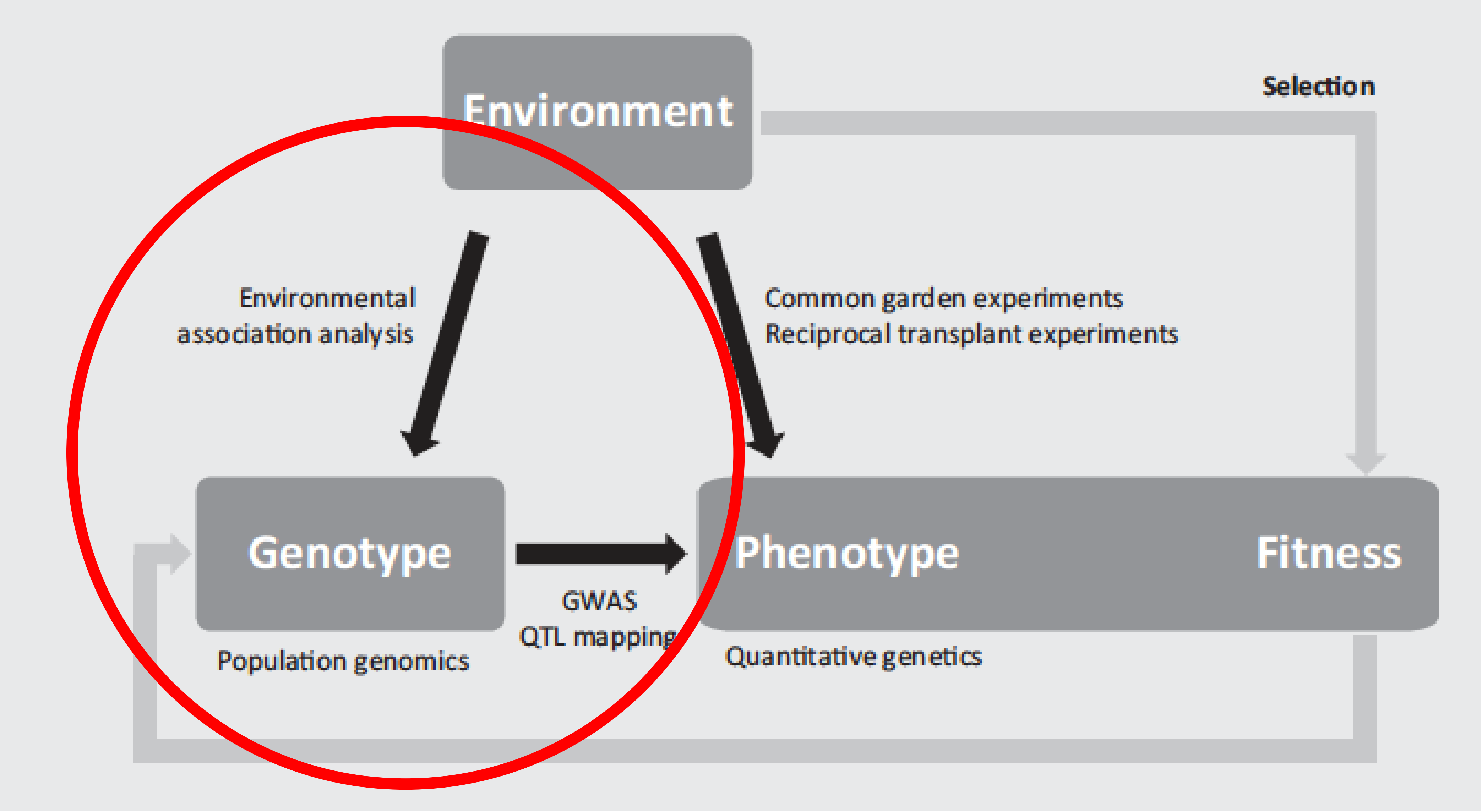


$w_{RR} > w_{RC} > w_{CC}$

$w_{RR} < w_{RC} < w_{CC}$

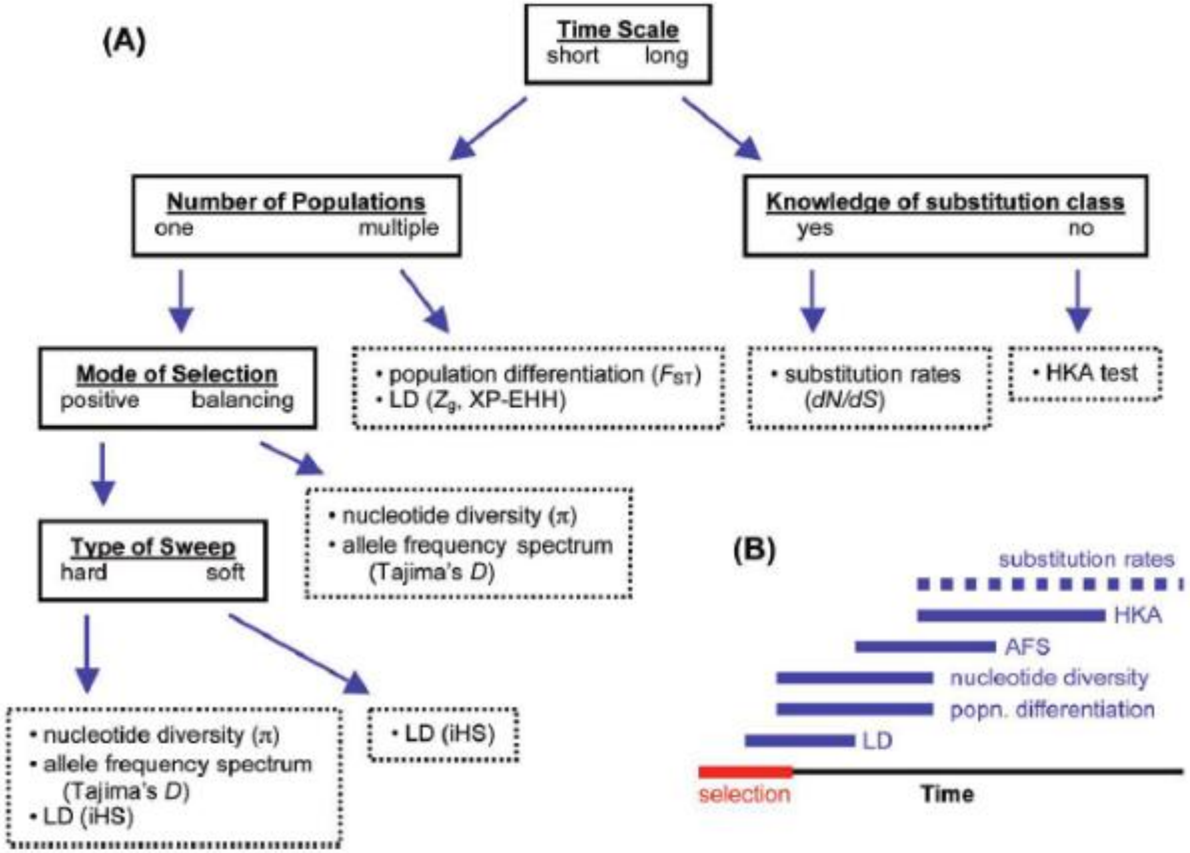
a crossing reaction norm for **fitness** == local adaptation

Detecting selection with genomic data

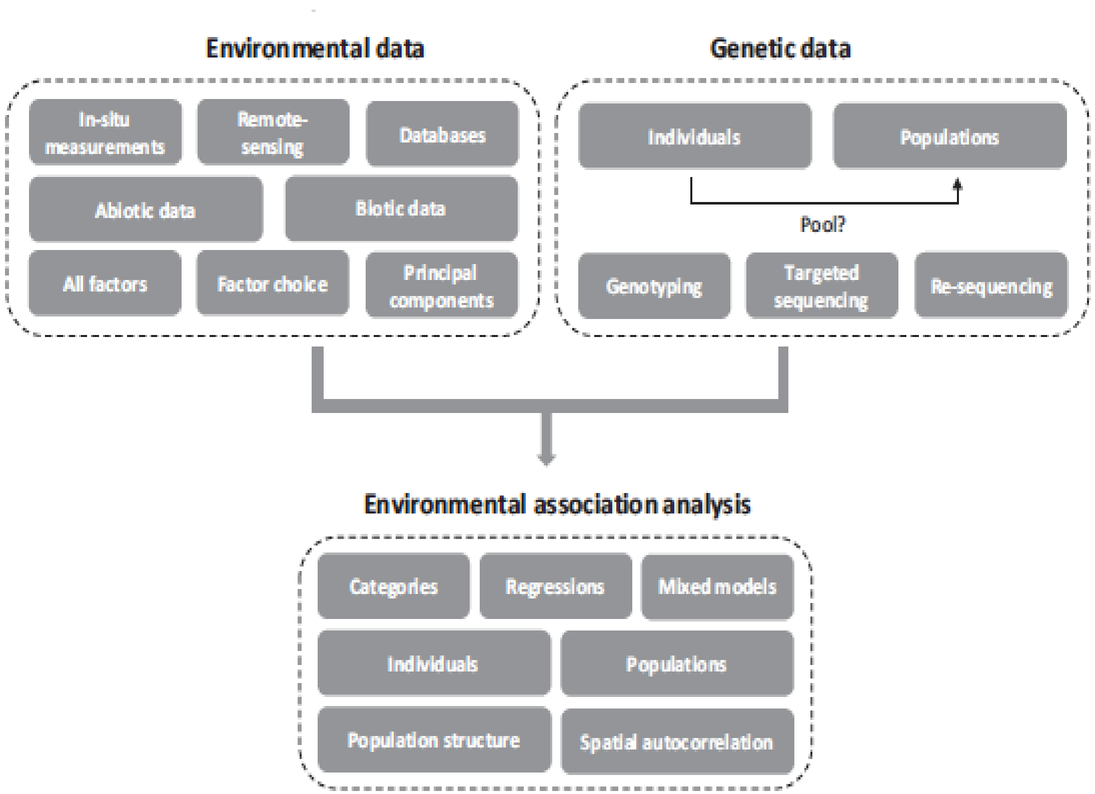


(Rellstab et al. 2015)

Detecting selection with genomic data



(Hohenlohe et al. 2010)



(Rellstab et al. 2015)

using only (or primarily) variant data

using variant and other (phenotypic, environmental, fitness) data

Detecting selection with genomic data

selection is locus-specific,
whereas historical and/or demographic effects act genome-wide

therefore must be able to:

1. describe the background genomic context (demography/history)
2. differentiate it from the target signature (selection)

Detecting selection with genomic data

genomic heterogeneity in summary statistics, incl. those used to infer selection

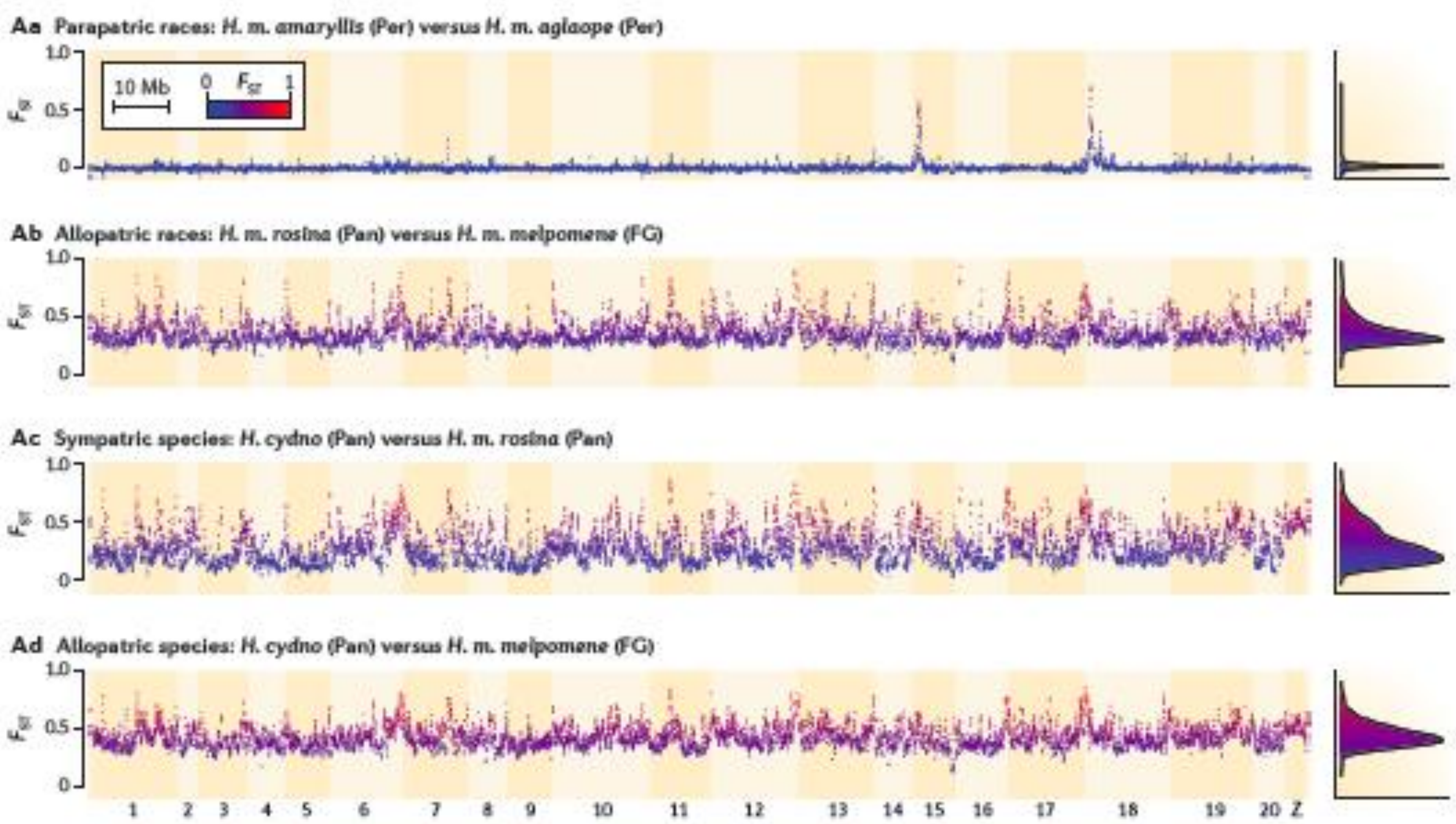
genetic structure or historical relatedness among individuals

CORE CHALLENGE =

accounting for/incorporating background variation

genomic heterogeneity in summary statistics

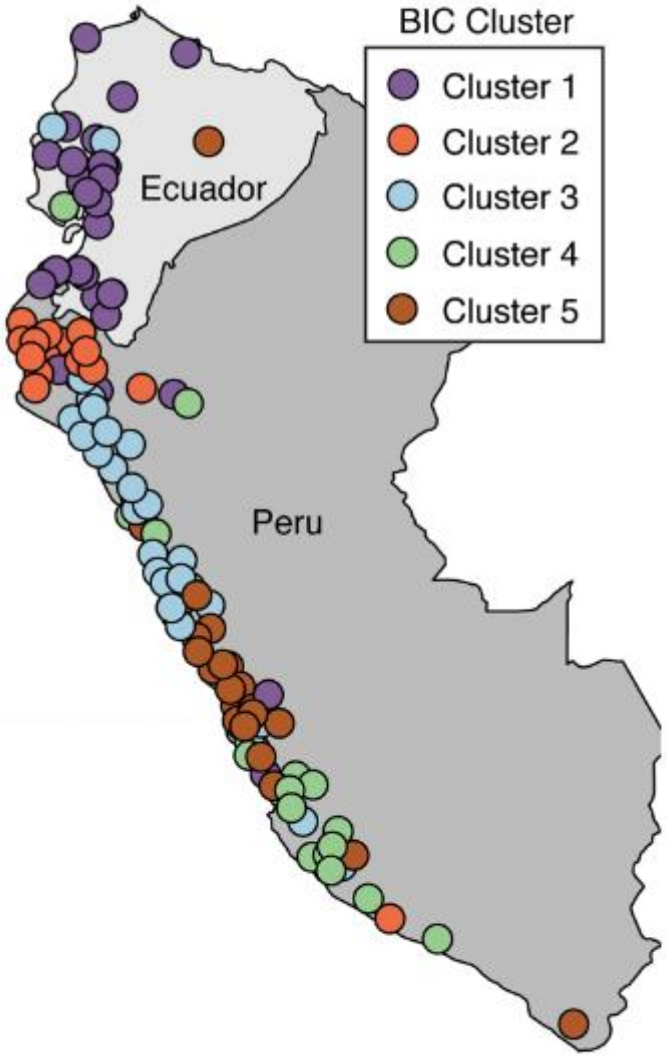
(often spatially correlated across the genome)



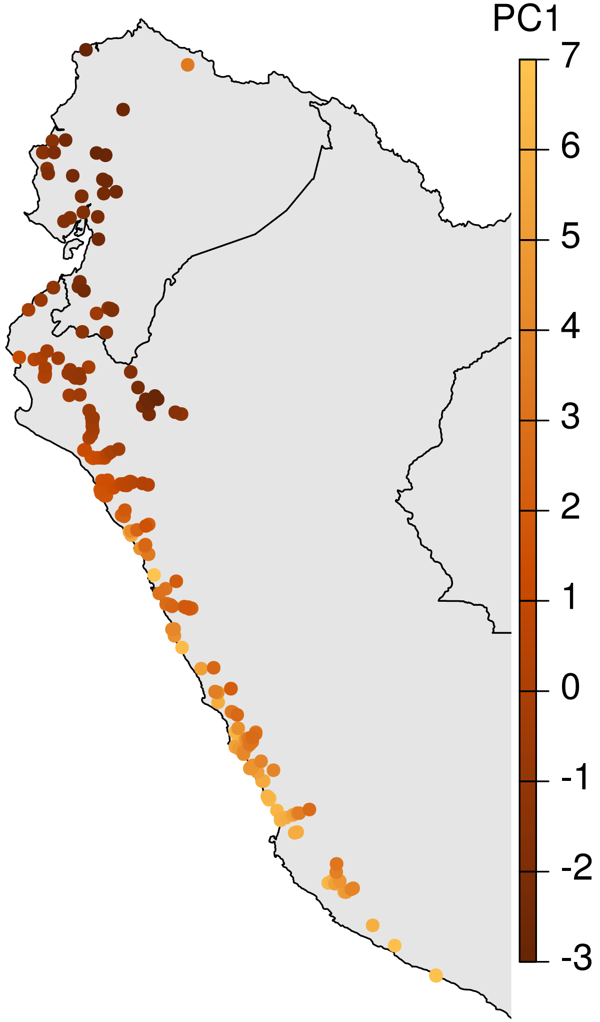
Examples…. different species pairs of Heliconius butterflies

genetic structure or historical relatedness among individuals

(often spatially correlated across geography)



Genetic structure



Environmental structure

Example….

spatial structure in
wild tomato *S. pimpinellifolium*

Detecting selection with genomic data

lots (most?) of population genomics aims to characterize these genome-wide/ 'background' features

but this often
isn't easy…

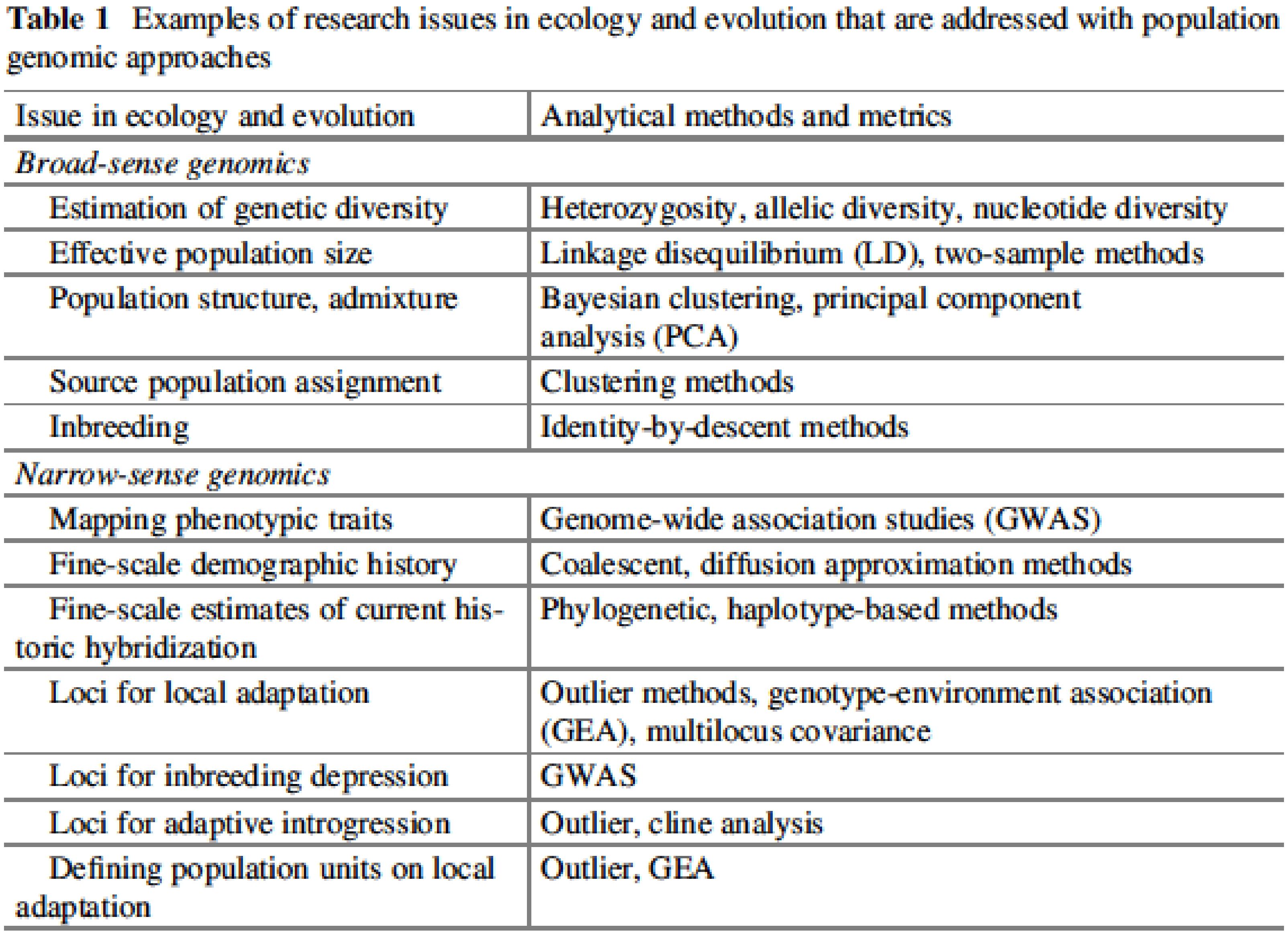


Table 1 Examples of research issues in ecology and evolution that are addressed with population genomic approaches

Issue in ecology and evolution	Analytical methods and metrics
Broad-sense genomics	
Estimation of genetic diversity	Heterozygosity, allelic diversity, nucleotide diversity
Effective population size	Linkage disequilibrium (LD), two-sample methods
Population structure, admixture	Bayesian clustering, principal component analysis (PCA)
Source population assignment	Clustering methods
Inbreeding	Identity-by-descent methods
Narrow-sense genomics	
Mapping phenotypic traits	Genome-wide association studies (GWAS)
Fine-scale demographic history	Coalescent, diffusion approximation methods
Fine-scale estimates of current historic hybridization	Phylogenetic, haplotype-based methods
Loci for local adaptation	Outlier methods, genotype-environment association (GEA), multilocus covariance
Loci for inbreeding depression	GWAS
Loci for adaptive introgression	Outlier, cline analysis
Defining population units on local adaptation	Outlier, GEA

(Hohenlohe et al. 2018)

Detecting selection with genomic data

contemporary
recent
older

time and/or spatial scale

within populations
between populations
between species

your approach to detecting selection will depend
upon your sample design and study goal

selection within populations

goal: identify loci	undergoing recent selection (with or w/out phenotype)	underlying important functional variation
signature: variants/regions	that depart from neutral or null expectations	associated with segregating functional variation
approaches:	sequence-based tests of selection	association studies

sequence-based tests of selection

Goal:

Identify markers/variants/SNPs that deviate from generic, null, or genome-wide patterns, due to the action of recent selection

Rationale:

- selection generates predictable changes in the kind, amount, and distribution of genetic variation

- targets of (recent) selection should be detectable based on characteristic *patterns of population genetic statistics at local genomic locations*, that **differ from background regions**

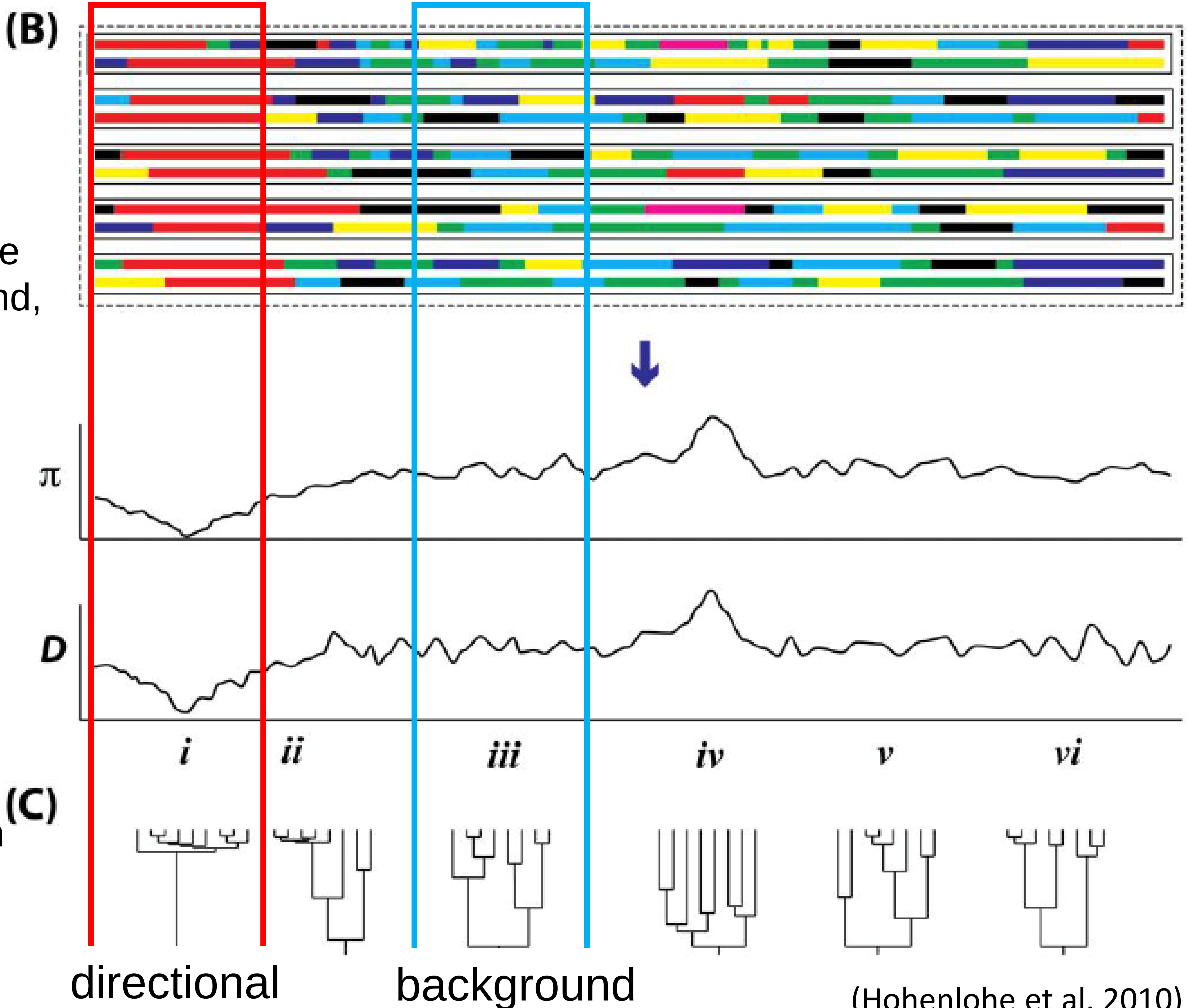
sequence-based tests of selection

in comparison to the genomic background, selection changes:

amount of sequence diversity

allele frequency spectrum

topology and depth of the coalescent



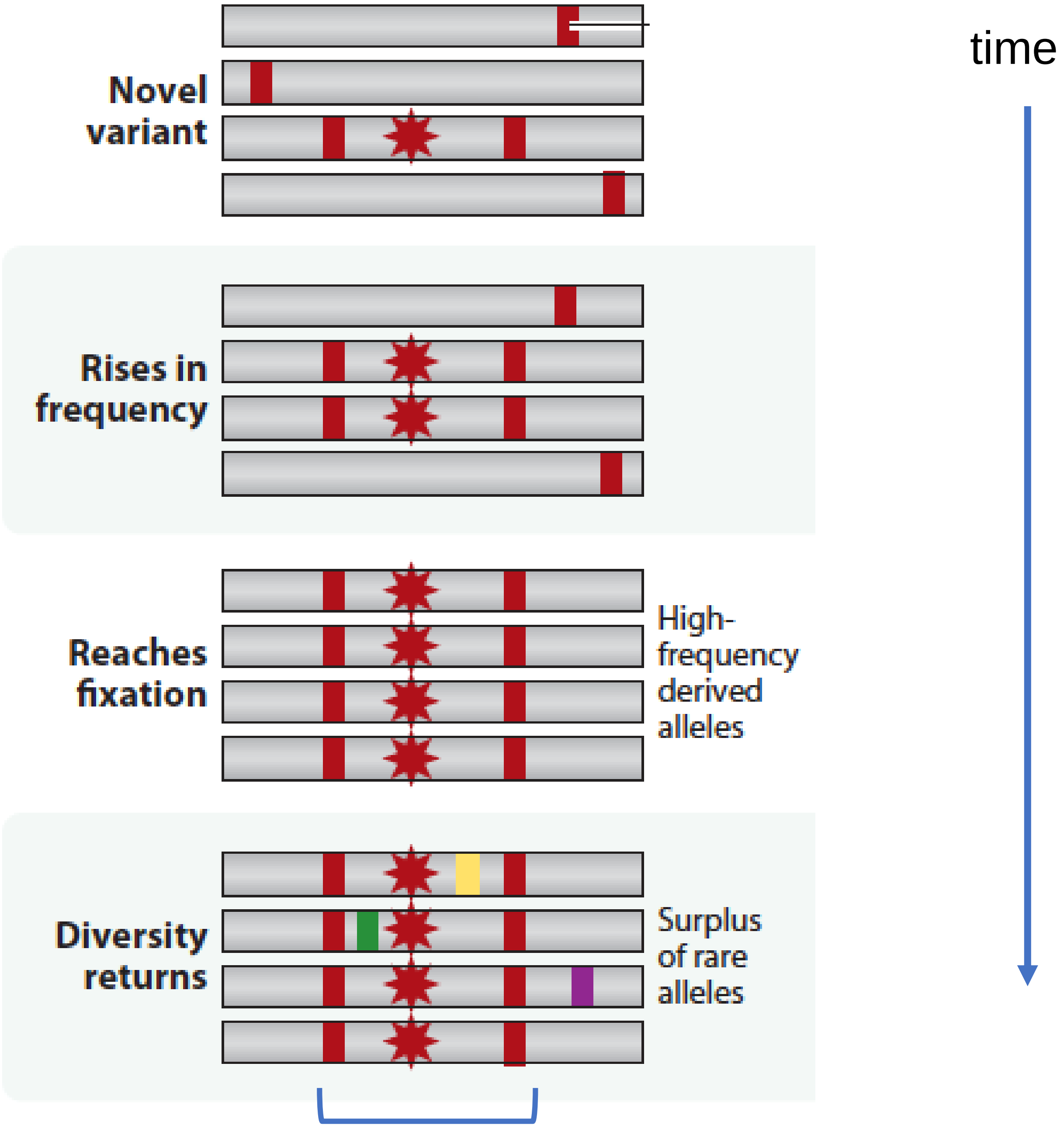
(Hohenlohe et al. 2010)

selective sweeps

in comparison to the genomic background, at the site of a selective sweep:

reduced sequence diversity

shifted allele frequency spectrum (excess rare variants)

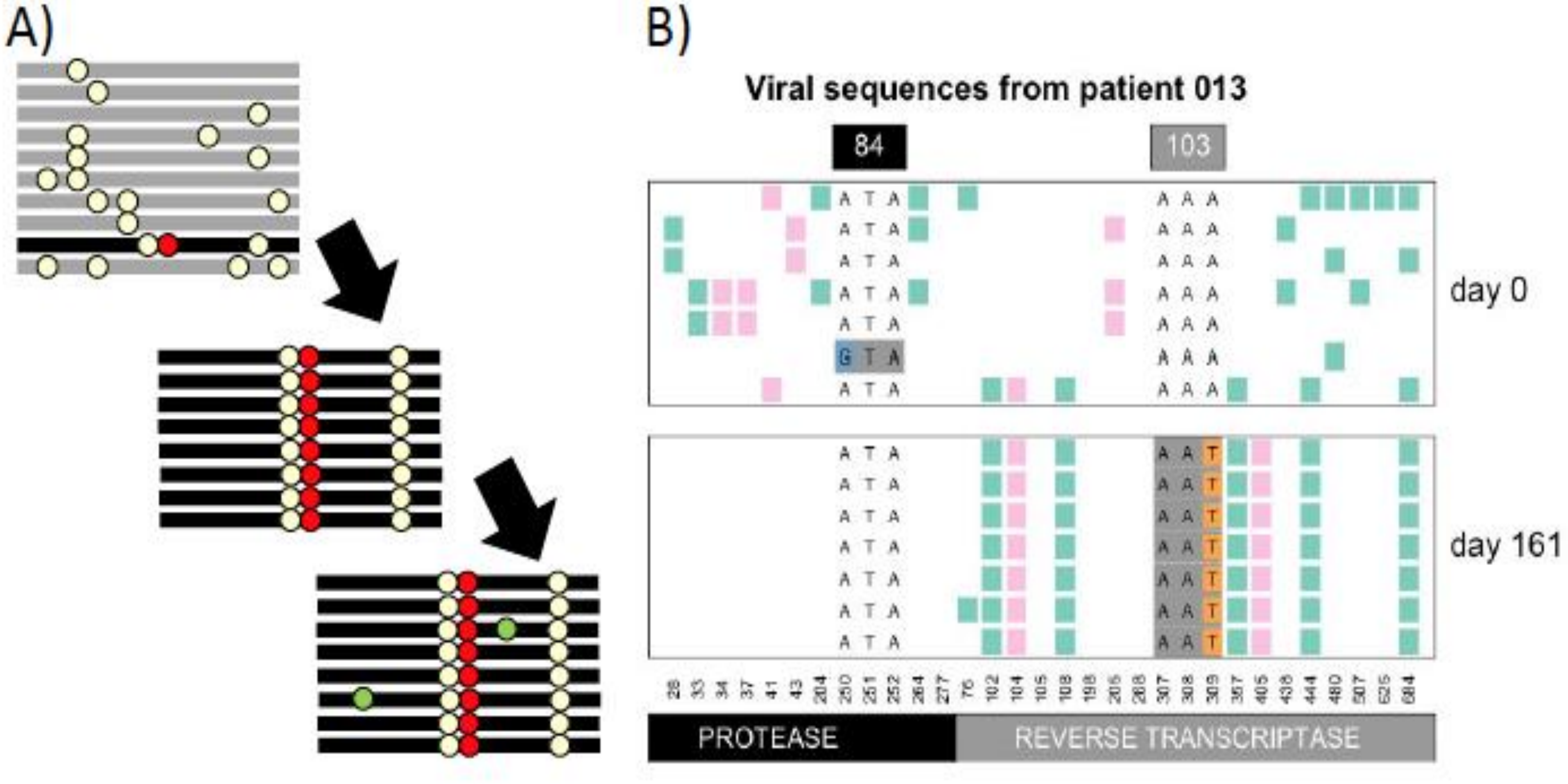


selective sweeps

in comparison to the genomic background, at the site of a selective sweep:

reduced sequence diversity

shifted allele frequency spectrum (excess rare variants)



(Coop 2022; originally Williams and Pennings 2019)

selective sweeps

in comparison to the genomic background, at the site of a selective sweep:

reduced sequence diversity

shifted allele frequency spectrum (excess rare variants)

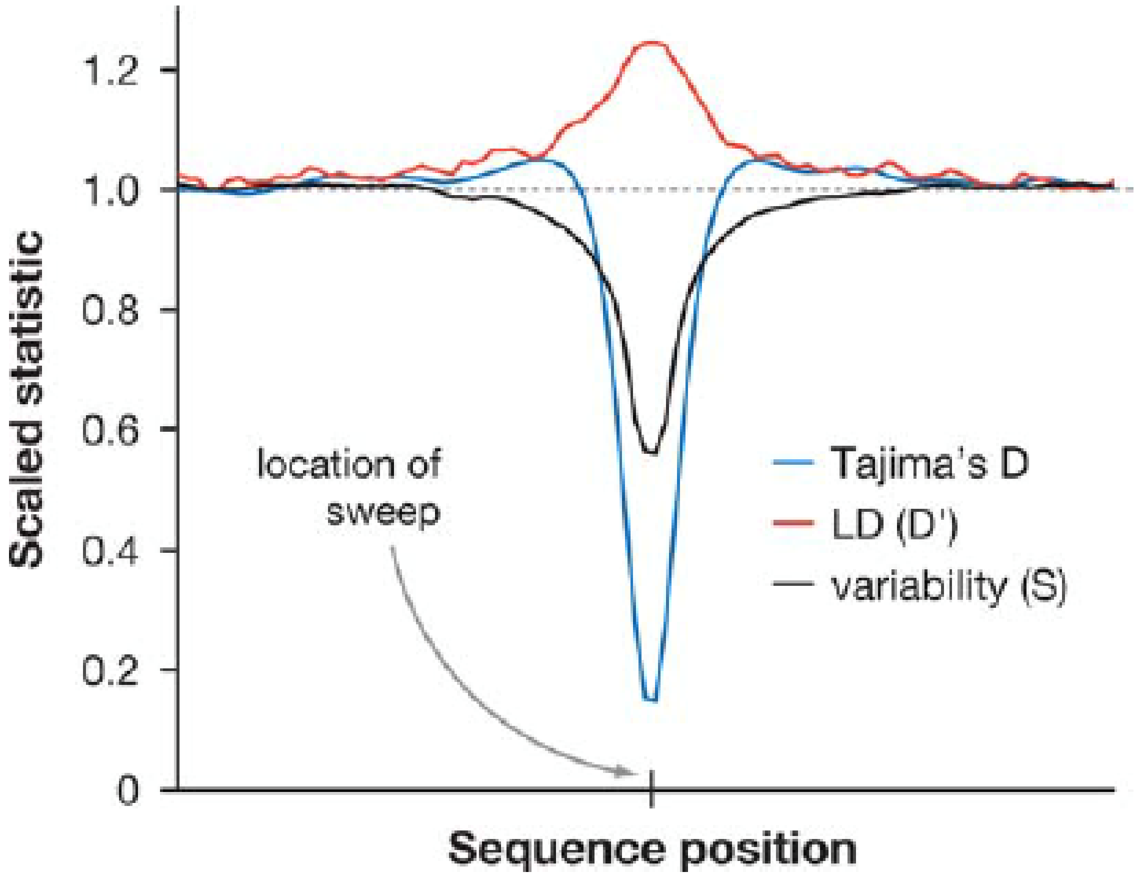


Figure 1

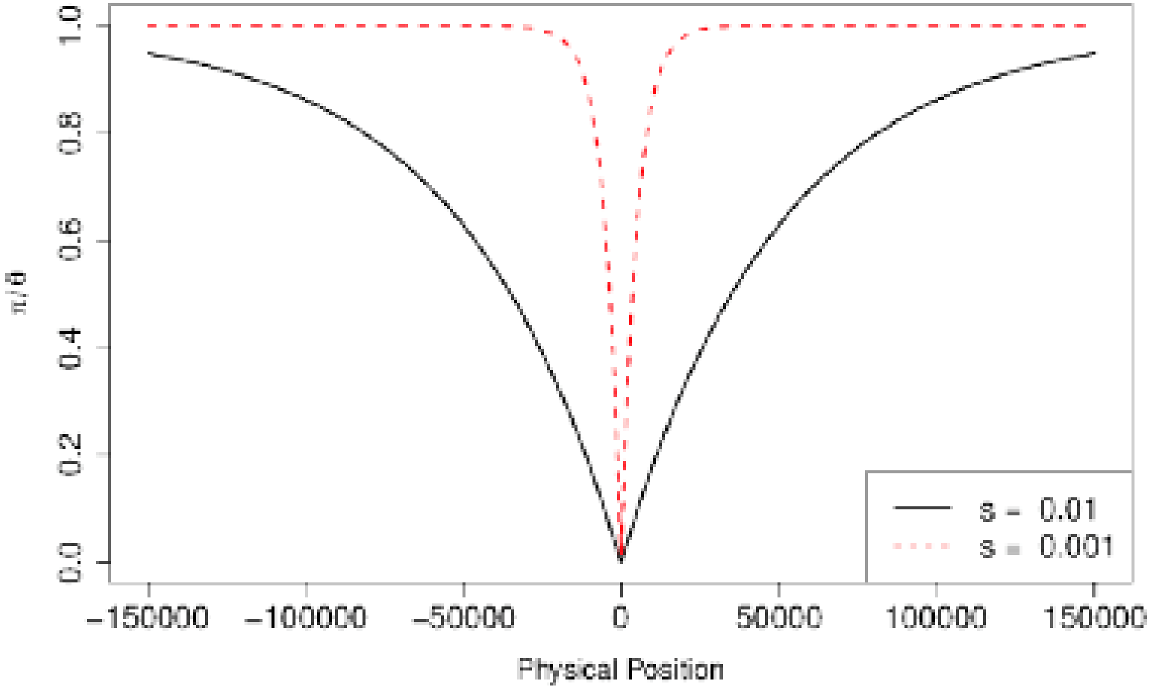
The effect of a selective sweep on genetic variation. The figure is based on averaging over 100 simulations of a strong selective sweep. It illustrates how the number of variable sites (variability) is reduced, LD is increased, and the frequency spectrum, as measured by Tajima's D, is skewed, in the region around the selective sweep. All statistics are calculated in a sliding window along the sequence right after the advantageous allele has reached frequency 1 in the population. All statistics are also scaled so that the expected value under neutrality equals one.

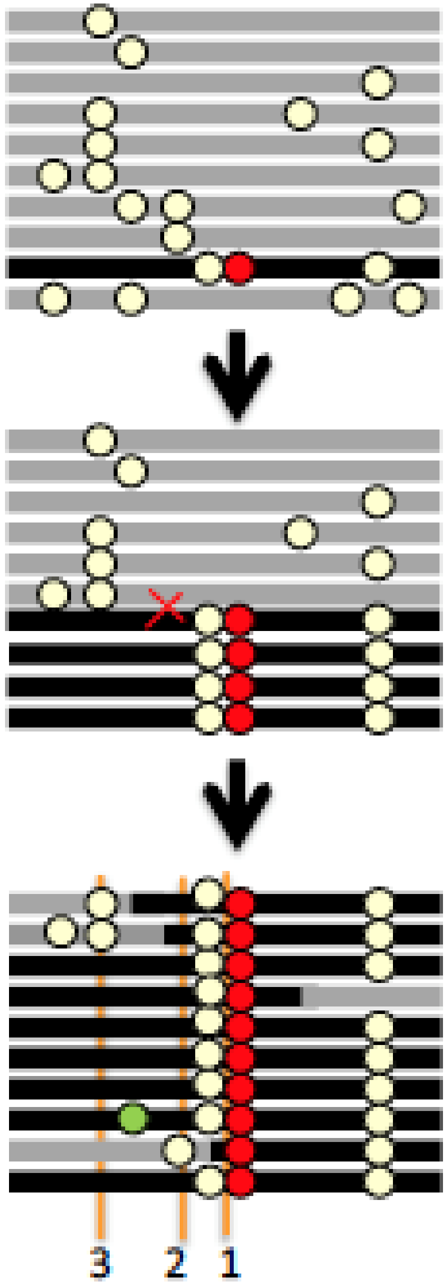
(Nielson 2005)

selective sweeps

extent of *hitchhiking region* depends on (among other things):

selection coefficient
(how advantageous the new variant is)



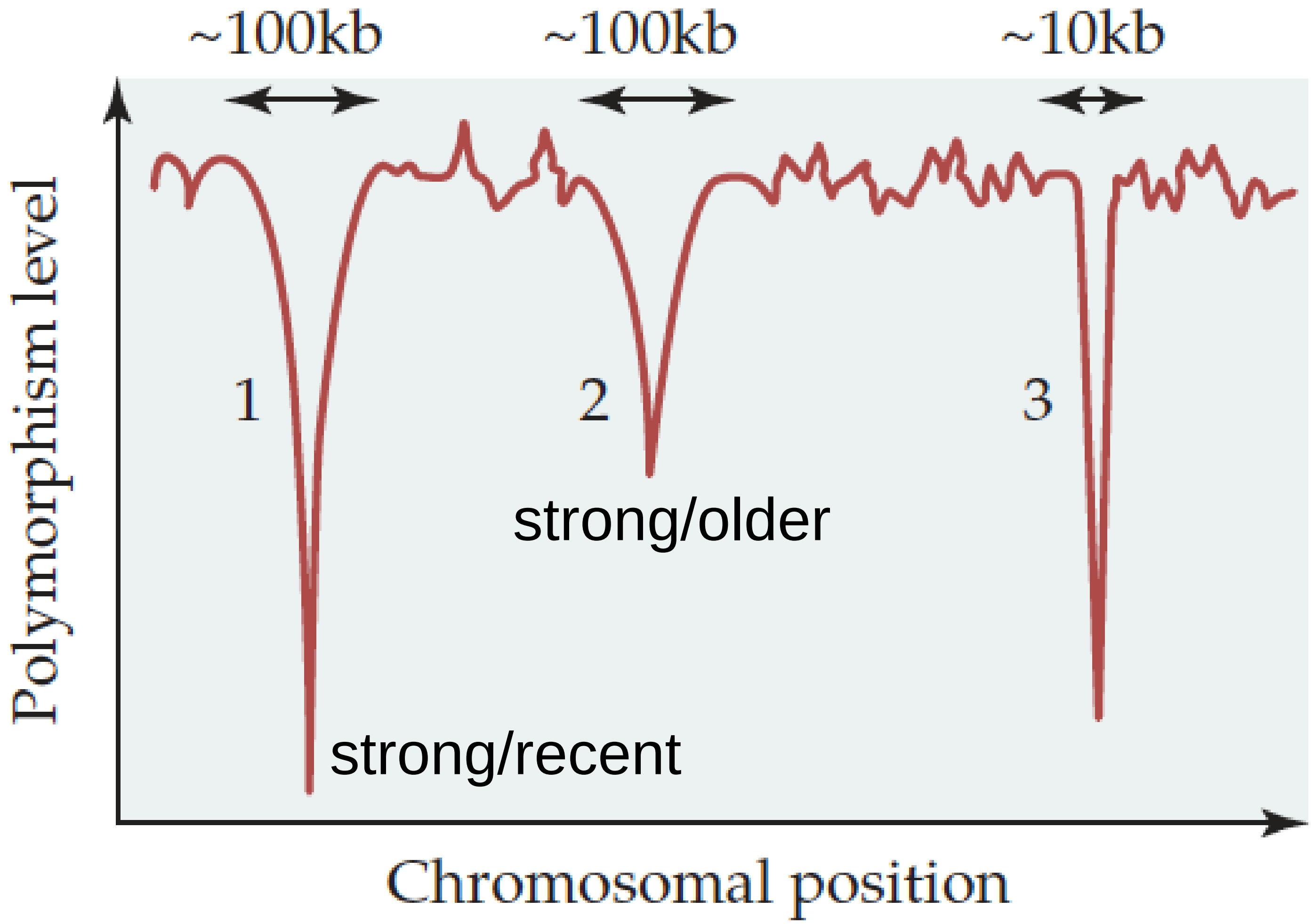


features of the sweep can tell you something about the strength of selection (and/or interaction with recombination (time))

recombination around the selected variant

(Coop 2022, Chapter 13)

selective sweeps



features of the sweep can tell you something about the strength of selection (and/or interaction with recombination (time))

(Figure 8.2, from Hahn 2018)

balancing selection

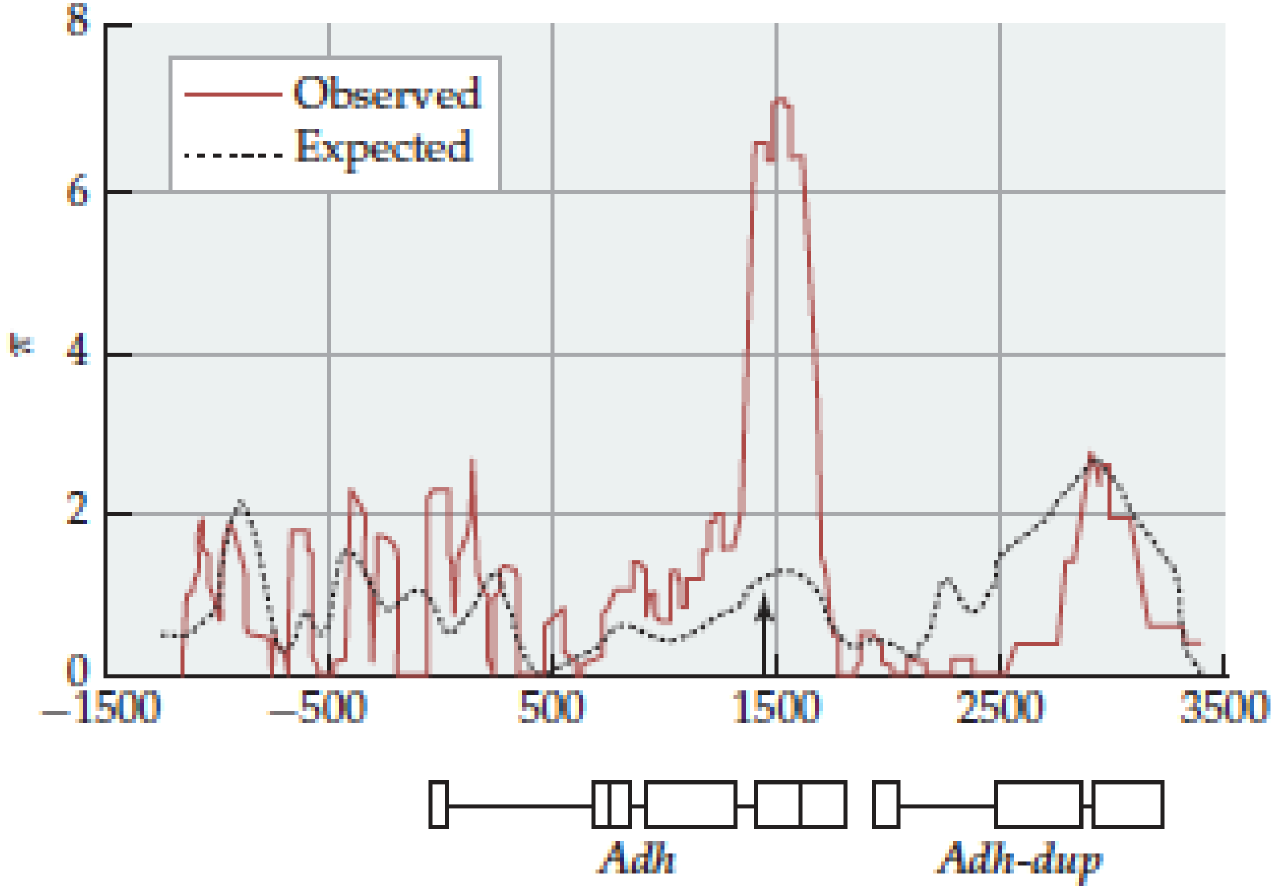
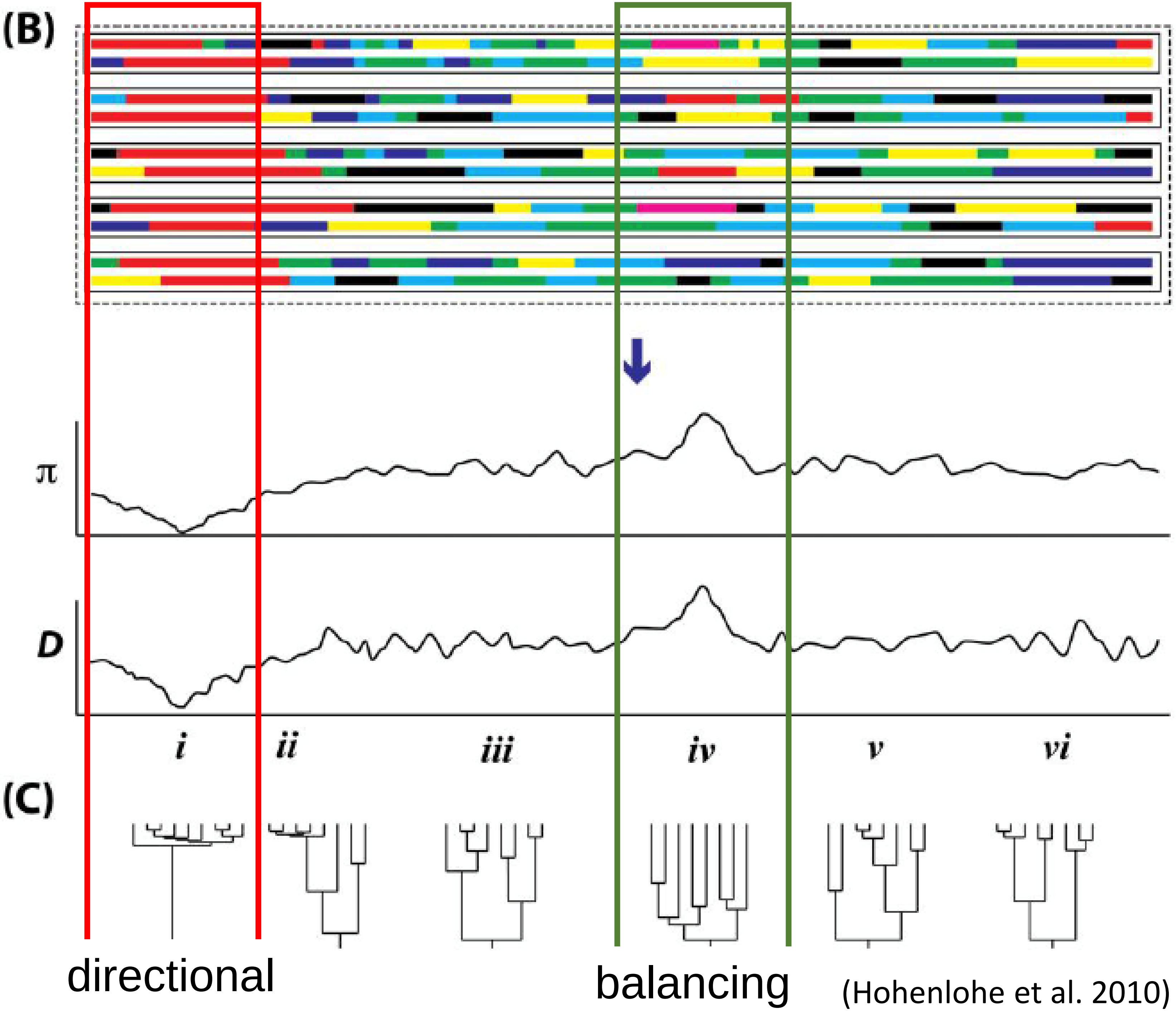


FIGURE 8.4 Balancing selection at *Adh* in *D. melanogaster*. The solid line shows a 100-bp average of π across the region, while the dotted line shows the value expected under a neutral model. The site of the suspected balanced polymorphism is marked with an arrow. (From Kreitman and Hudson 1991.)

(from Hahn 2018)

sequence-based tests of selection

potentially very powerful



(Hohenlohe et al. 2010)

sequence-based tests of selection

limitations

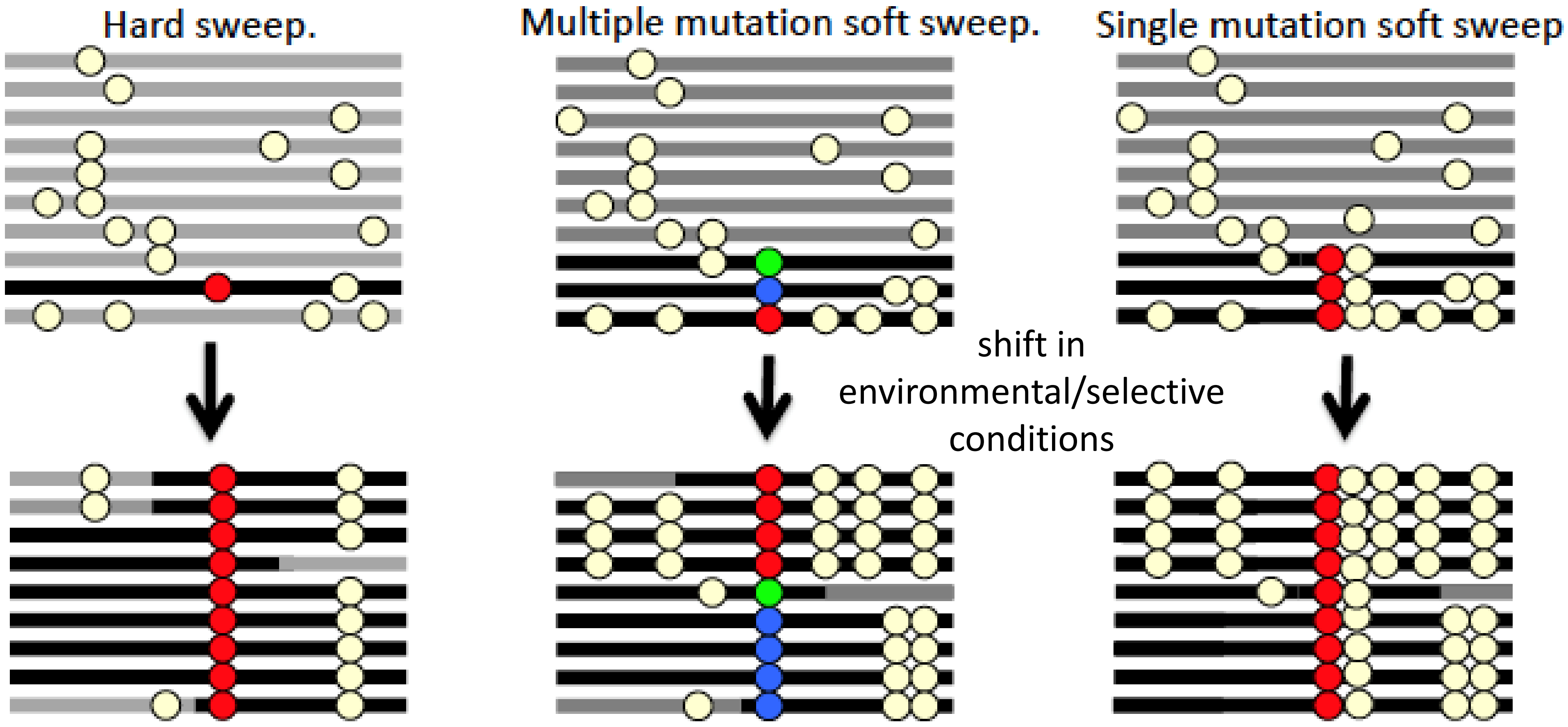
- need tonnes of data (med to high coverage), for good inferences
- need genomic position information

- soft selective sweeps
- polygenic basis to traits
- epistatic interactions among genes

selection based on incremental &/or collective changes at 2+ loci

selective sweeps

new variation **standing variation** **standing variation**



shift in environmental/selective conditions

monogenic traits? polygenic traits

(Coop 2022, Chapter 13)

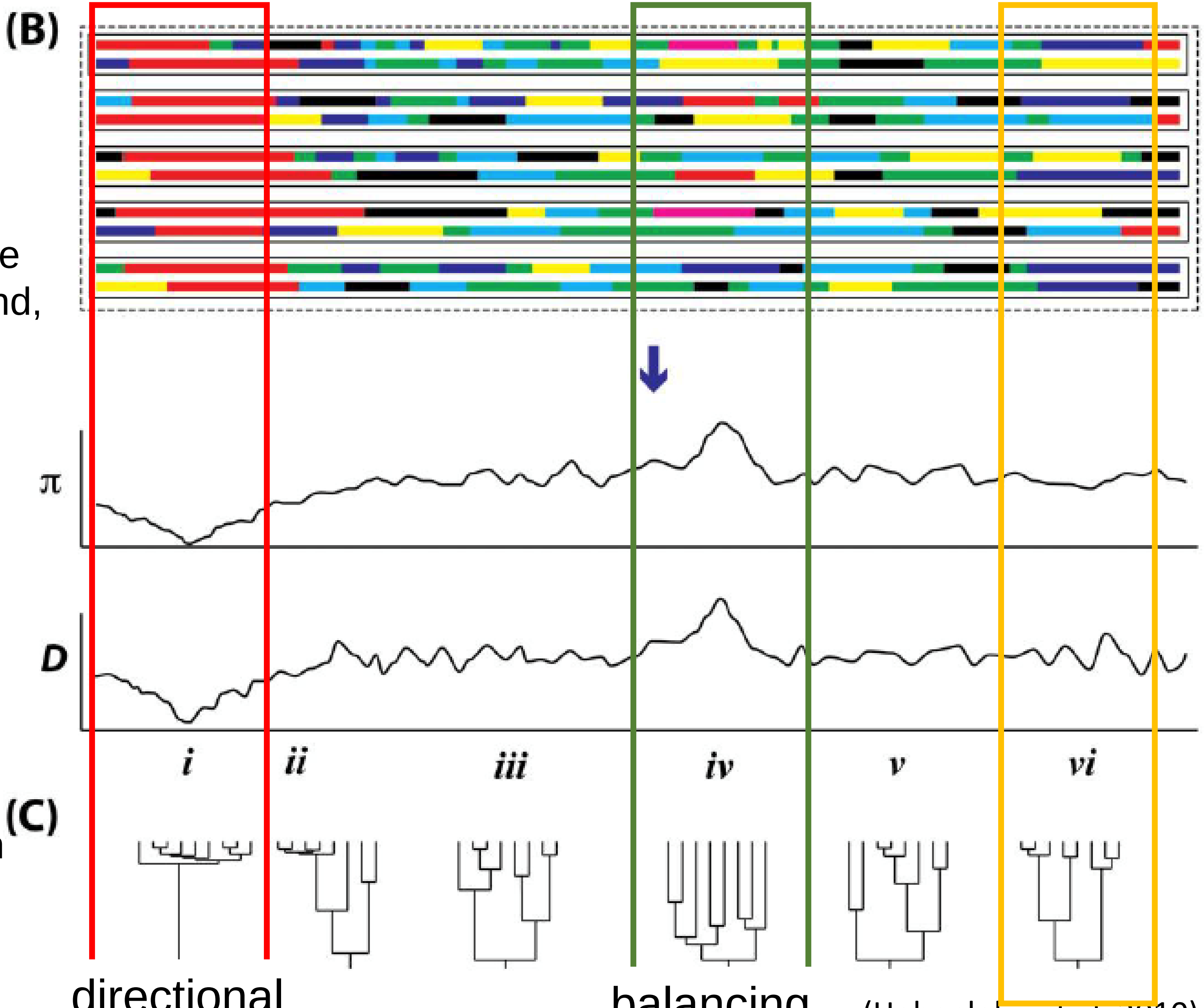
sequence-based tests of selection

in comparison to the genomic background, selection changes:

amount of sequence diversity

allele frequency spectrum

topology and depth of the coalescent



(Hohenlohe et al. 2010)

sequence-based
tests of selection

limitations

- need tonnes of data (med to high coverage), for good inferences
- need genomic position information

- soft selective sweeps
- polygenic basis to traits
- epistatic interactions among genes

selection based on incremental &/or collective changes at 2+ loci

- too little, or too much, time since selection

FALSE NEGATIVES

sequence-based
tests of selection

limitations

no phenotypes, no fitness,
so no direct information on:

- selective conditions/agents
- locus identity (depending on system…)
- functional importance (“adaptation”)

anonymous tests are a double-edged sword

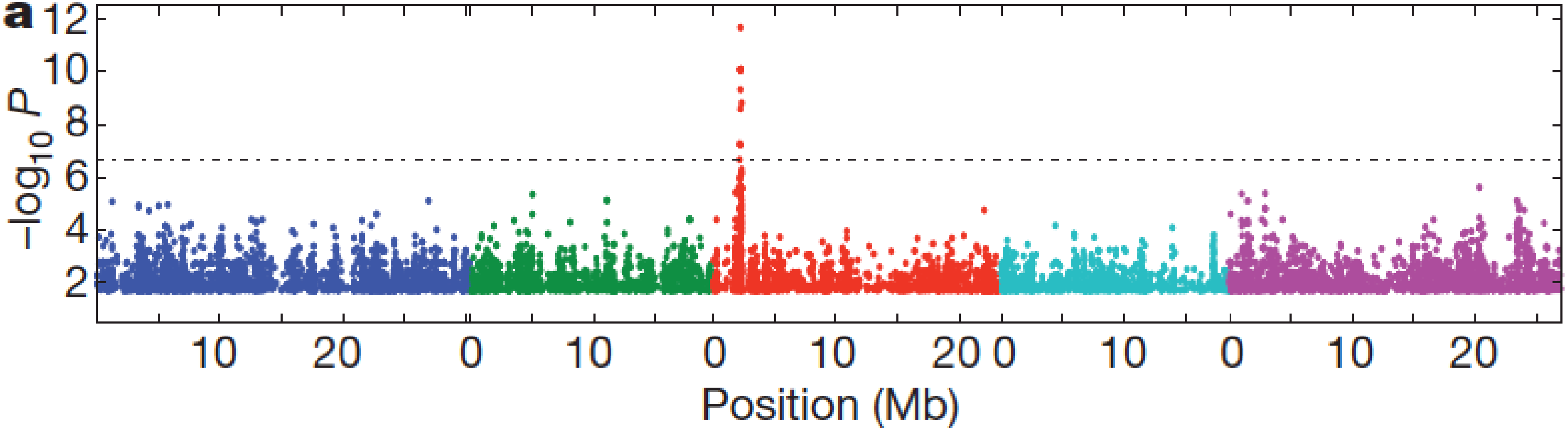
selection within populations

goal: identify loci	undergoing recent selection (with or w/out phenotype)	underlying important functional variation
signature: variants/regions	that depart from neutral or null expectations	associated with segregating functional variation
approaches:	sequence-based tests of selection	association studies

association studies

(genome-wide) analysis of statistical associations between traits and markers in large population samples.

GWAS (genome-wide association study)



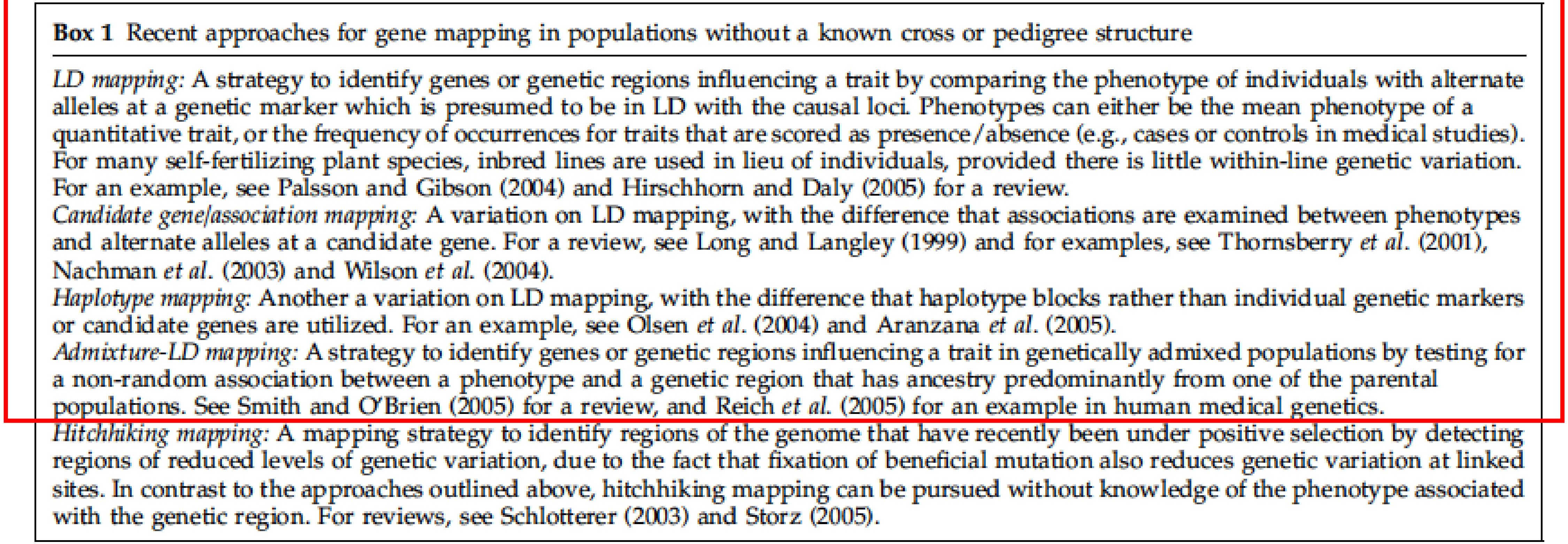
association score

chromosome position

association studies

(genome-wide) analysis of statistical associations between traits and markers in large population samples.

GWAS (genome-wide association study)



Box 1 Recent approaches for gene mapping in populations without a known cross or pedigree structure

LD mapping: A strategy to identify genes or genetic regions influencing a trait by comparing the phenotype of individuals with alternate alleles at a genetic marker which is presumed to be in LD with the causal loci. Phenotypes can either be the mean phenotype of a quantitative trait, or the frequency of occurrences for traits that are scored as presence/absence (e.g., cases or controls in medical studies). For many self-fertilizing plant species, inbred lines are used in lieu of individuals, provided there is little within-line genetic variation. For an example, see Palsson and Gibson (2004) and Hirschhorn and Daly (2005) for a review.

Candidate gene/association mapping: A variation on LD mapping, with the difference that associations are examined between phenotypes and alternate alleles at a candidate gene. For a review, see Long and Langley (1999) and for examples, see Thornsberry *et al*. (2001), Nachman *et al*. (2003) and Wilson *et al*. (2004).

Haplotype mapping: Another a variation on LD mapping, with the difference that haplotype blocks rather than individual genetic markers or candidate genes are utilized. For an example, see Olsen *et al*. (2004) and Aranzana *et al*. (2005).

Admixture-LD mapping: A strategy to identify genes or genetic regions influencing a trait in genetically admixed populations by testing for a non-random association between a phenotype and a genetic region that has ancestry predominantly from one of the parental populations. See Smith and O'Brien (2005) for a review, and Reich *et al*. (2005) for an example in human medical genetics.

Hitchhiking mapping: A mapping strategy to identify regions of the genome that have recently been under positive selection by detecting regions of reduced levels of genetic variation, due to the fact that fixation of beneficial mutation also reduces genetic variation at linked sites. In contrast to the approaches outlined above, hitchhiking mapping can be pursued without knowledge of the phenotype associated with the genetic region. For reviews, see Schlotterer (2003) and Storz (2005).

(Stinchcombe & Hoekstra 2008)

association studies

(genome-wide) analysis of statistical associations between traits and markers in large population samples.

Goal:

Identify markers/variants/SNPs statistically associated with variation in traits of interest, due to LD with causal loci

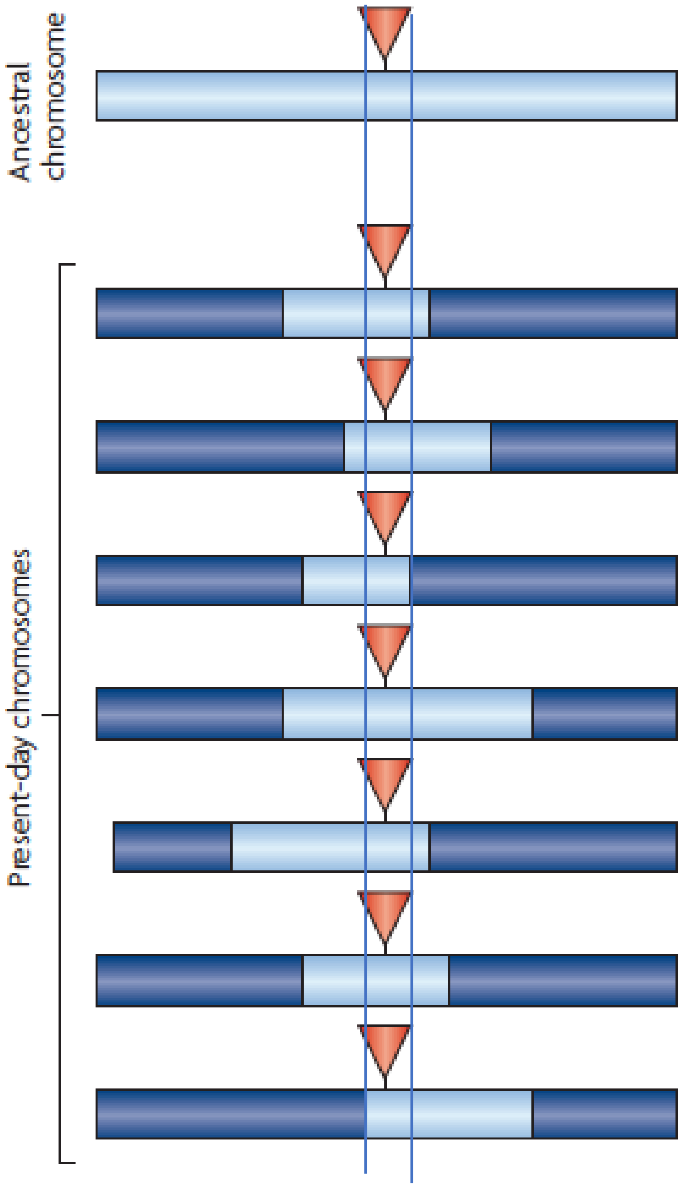
Reminder: what is LD (linkage disequilibrium)?

A **statistical association** between markers or loci, such that: alternative alleles at 2 (or more) loci are found **together more often than expected by chance** (e.g. mendelian ratios)

L.D. can be due to (for example):

- chromosomal association (physical linkage) between loci
- historical/geneological associations (population structure) between alleles at different loci
- selection for/against particular allelic associations

association studies



origin of causal mutation

subsequent recombination

Physically adjacent markers will remain associated with target locus (SNP) through many recombination events

when LD is short, need high marker density so at least a few remain in LD with target locus

markers in perfect LD with target locus

(Kruglyak 2008)

association studies

(genome-wide) analysis of statistical associations between traits and markers in large population samples.

Goal:

Identify markers/variants/SNPs statistically associated with variation in traits of interest, due to LD with causal loci

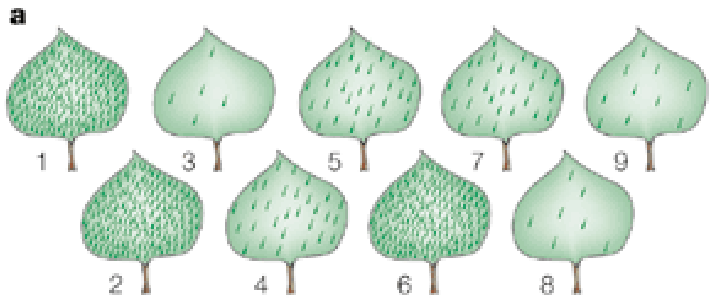
Rationale:

- markers physically linked with (adjacent to) causal locus should be statistically associated with phenotypic effect of that locus

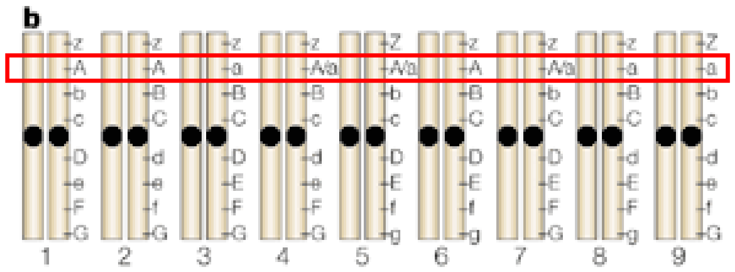
- natural populations (‘wild’ samples) have accumulated many recombination events (therefore resolution is very fine-scaled)

contrast with: QTL mapping

Phenotypes of mapping population

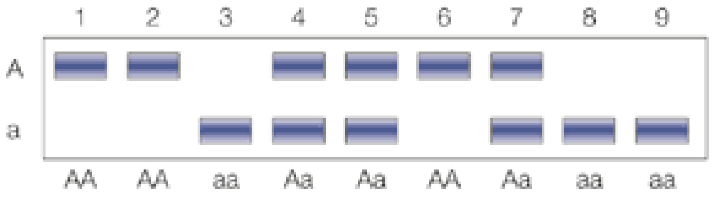


Genotypes of mapping population (1 csome)



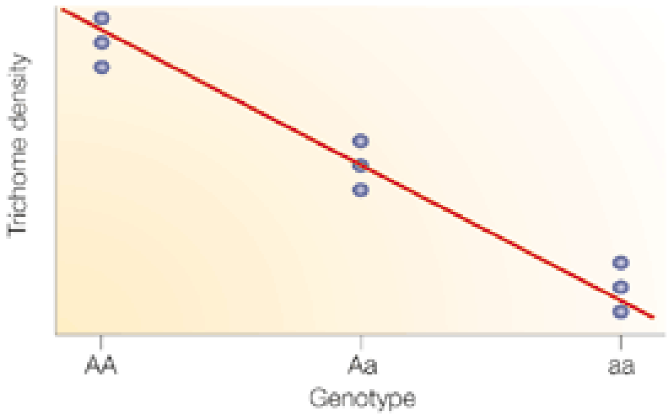
Genotypes at locus A

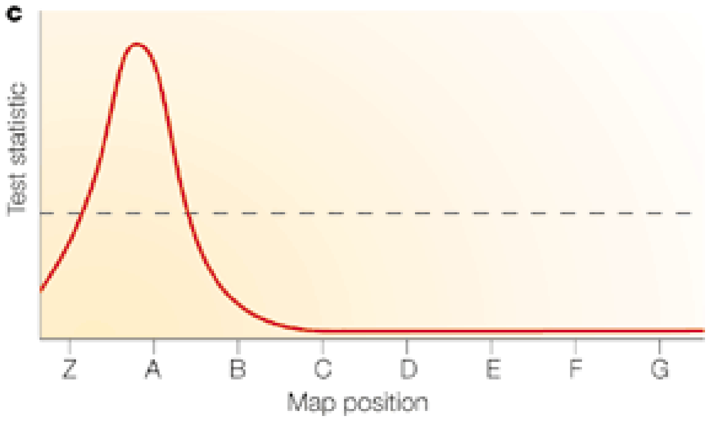
Genotype alleles at locus A



Analyses estimate the degree of covariation/association between:

- each marker (allele a vs A), and
- phenotype (trait measurements)

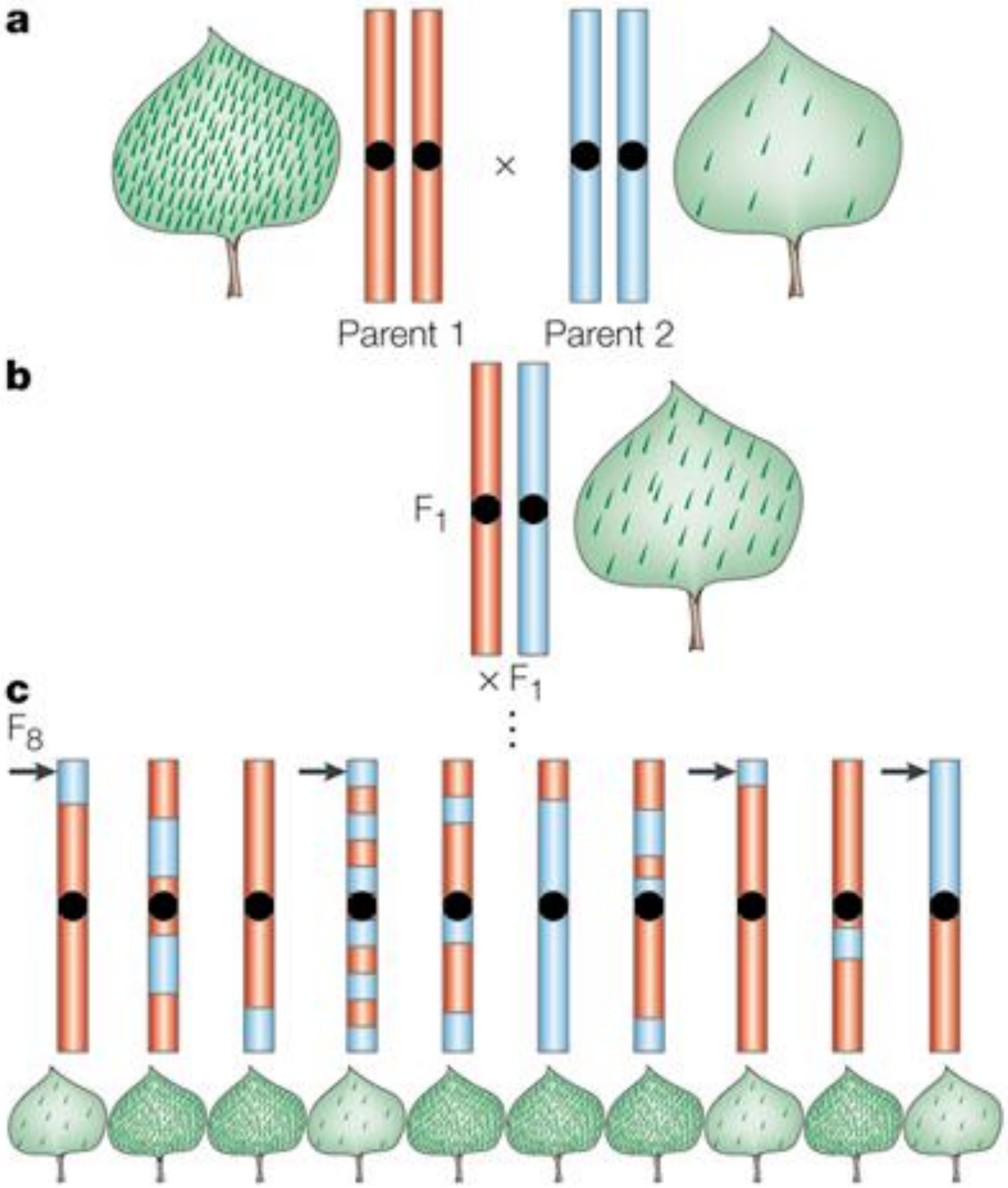




contrast with:
QTL
mapping

artificially segregating populations

finite (small) populations means limited # recombination events, therefore limited resolution



association studies

(genome-wide) analysis of statistical associations between traits and markers in large population samples.

Goal:

Identify markers/variants/SNPs statistically associated with variation in traits of interest, due to LD with causal loci

Rationale:

- markers physically linked with (adjacent to) causal locus should be statistically associated with phenotypic effect of that locus

- natural populations ('wild' samples) have accumulated many recombination events (therefore resolution is very fine-scaled)

association studies

(genome-wide) analysis of statistical associations between traits and markers in large population samples.

Requires:

- markers/variant sites (1000’s to WG) that differ between individuals
- linkage map or genome sequence
- quantitative phenotypes/trait variation
- methods to associate trait/genotype (and exclude confounding factors, correct for multiple testing)

In general:

- test marker by marker associations (or sometimes haplotypes)
- assess and control/account for population processes (**especially historical demography and/or population structure/relatedness**)

Why do we care about population structure?

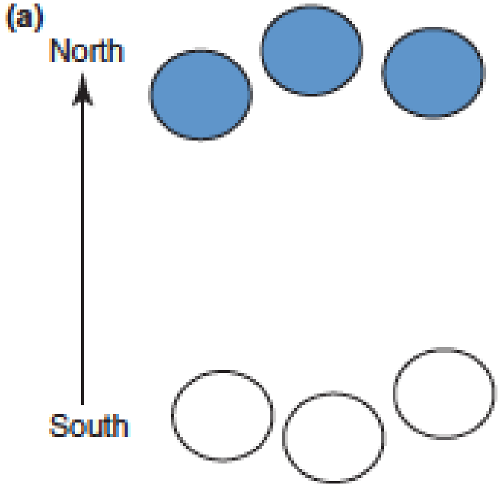
population structure—heterogeneous genetic relationships among individuals—**creates patterns of LD in a dataset**

that have NOTHING to do with adaptive trait variation.

association studies

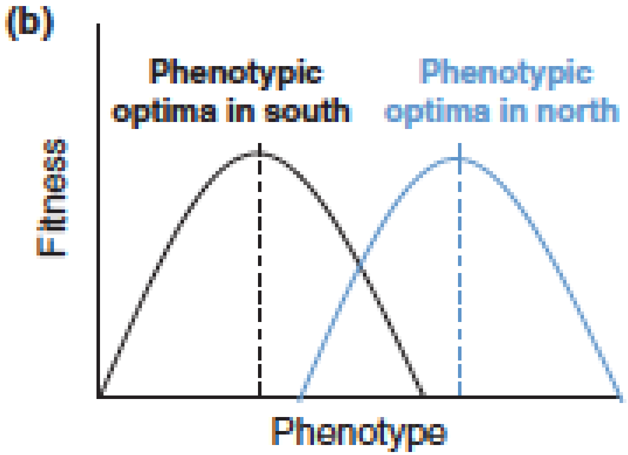
Population structure produces LD among unlinked loci

individuals are more closely related to each other, share SNPs



Historical structure due to selection and drift

individuals are more closely related to each other, share SNPs

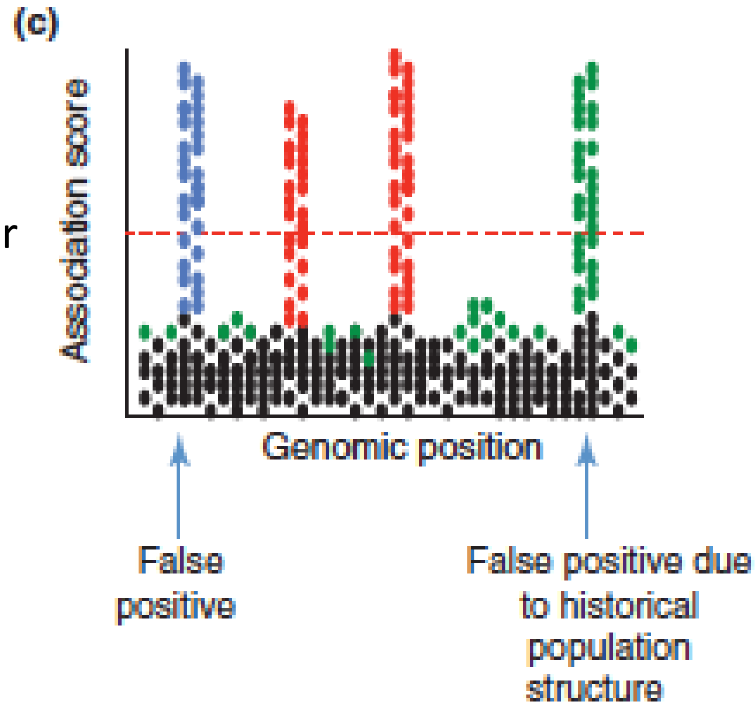


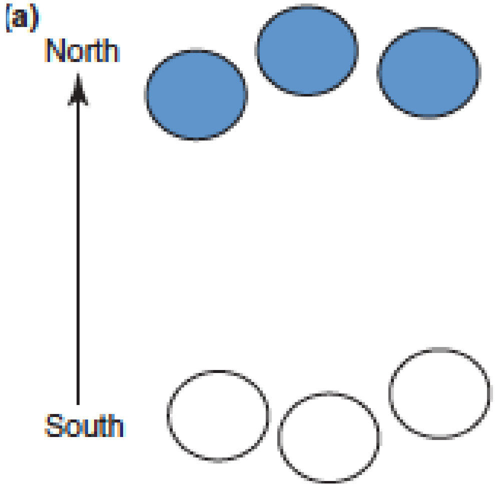
(Anderson, Willis, Mitchell-Olds 2011)

association studies

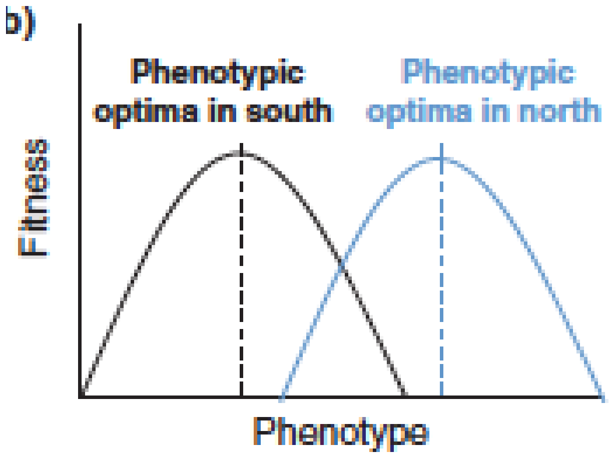
Population structure produces LD among unlinked loci

Associations without correcting for population structure





Historical structure due to selection and drift



(Anderson, Willis, Mitchell-Olds 2011)

Why do we care about population structure?

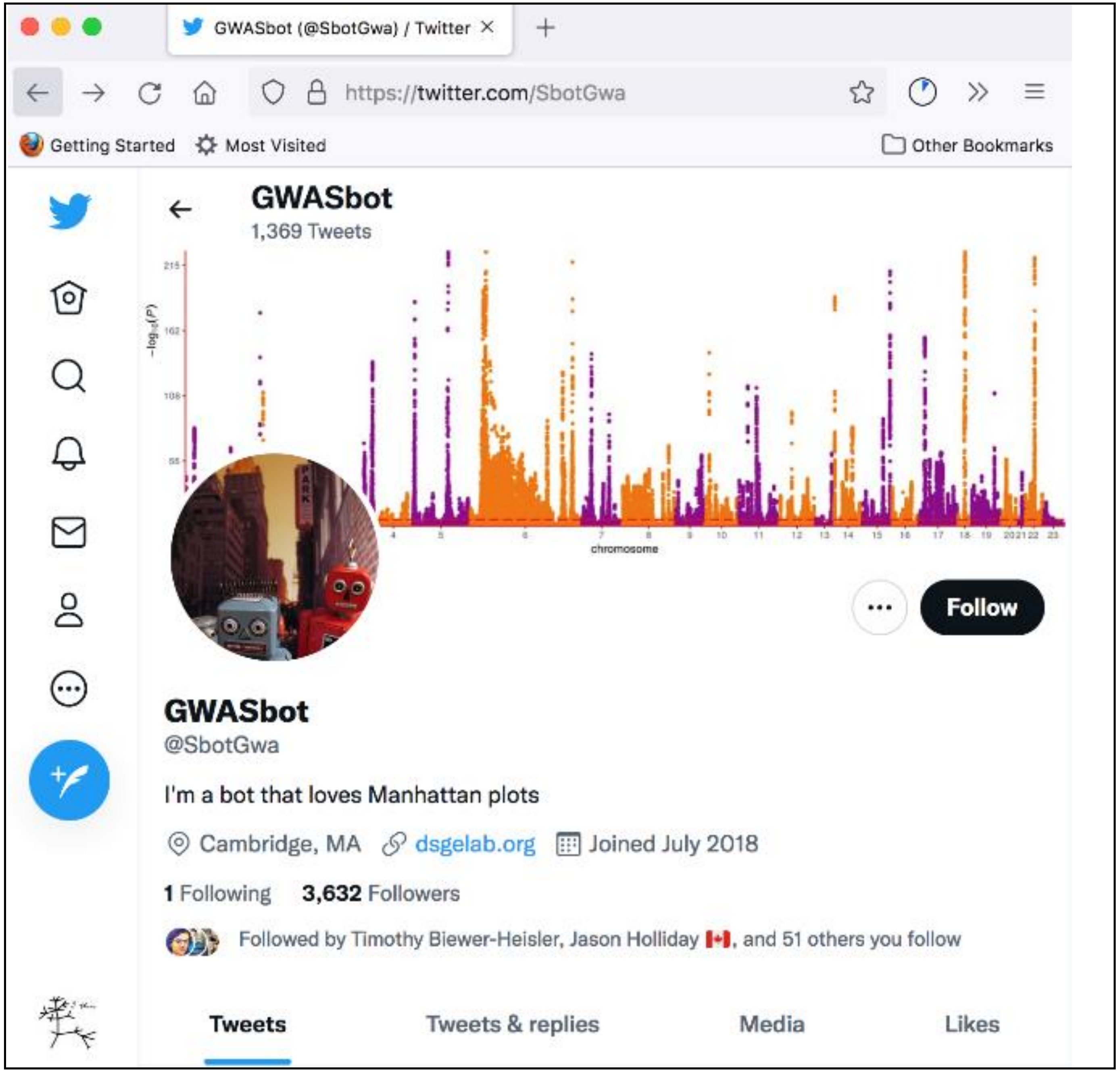
population structure—heterogeneous genetic relationships among individuals—**creates patterns of LD in a dataset**

that have NOTHING to do with adaptive trait variation.

!!when population history is correlated with distribution of trait variation, false positives!!

(!!similarly, when population history is correlated with the environment, false positives!!)

association studies



GWASbot

1,369 Tweets

GWASbot

@SbotGwa

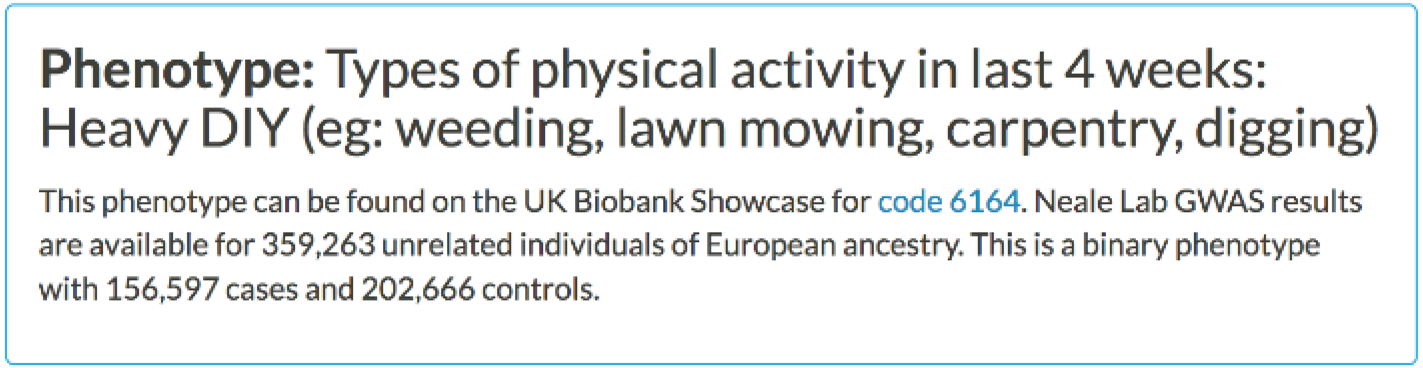
I'm a bot that loves Manhattan plots

Cambridge, MA dsgelab.org Joined July 2018

1 Following 3,632 Followers

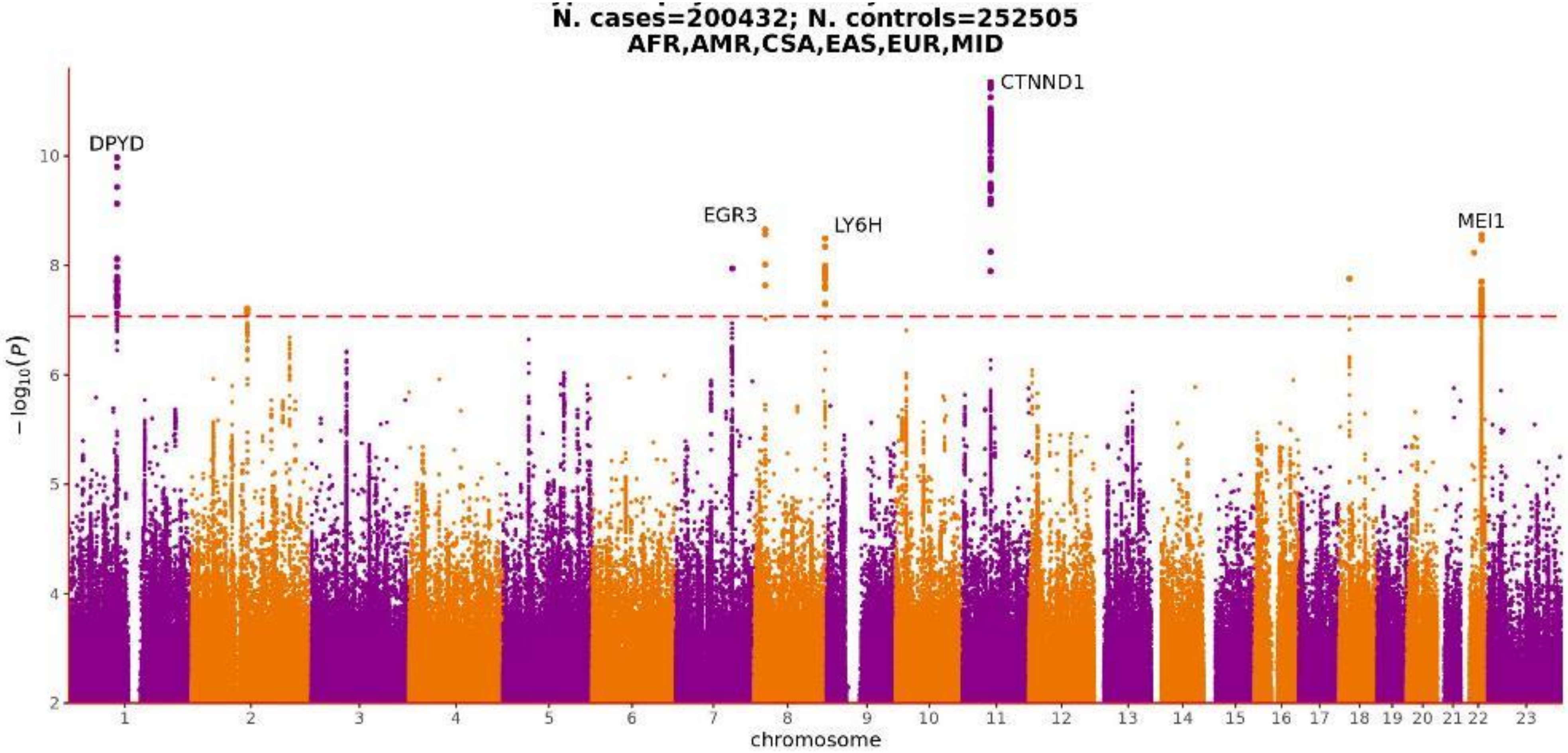
Followed by Timothy Biewer-Heisler, Jason Holliday 🇨🇦, and 51 others you follow

association studies



Phenotype: Types of physical activity in last 4 weeks: Heavy DIY (eg: weeding, lawn mowing, carpentry, digging)

This phenotype can be found on the UK Biobank Showcase for code 6164. Neale Lab GWAS results are available for 359,263 unrelated individuals of European ancestry. This is a binary phenotype with 156,597 cases and 202,666 controls.



association studies

limitations

one of the biggest ‘problems’ with GWAS (etc.) is that trait variation is often confounded with historical/spatial population structure

producing spurious (non-causal) associations
between markers and traits

FALSE POSITIVES

- correcting for population structure can overcompensate

FALSE NEGATIVES

- need tonnes more data: collecting (high quality) trait data is hard…
- still several steps away from direct causal inference

selection within populations

goal: identify loci	undergoing recent selection (with or w/out phenotype)	underlying important functional variation
signature: variants/regions	that depart from neutral or null expectations	associated with segregating functional variation
approaches:	sequence-based tests of selection	association studies
…and others…..	select and re-sequence (change over one or few generations)	

Detecting selection with genomic data

contemporary
recent
older

time and/or spatial scale

within populations
between populations
between species

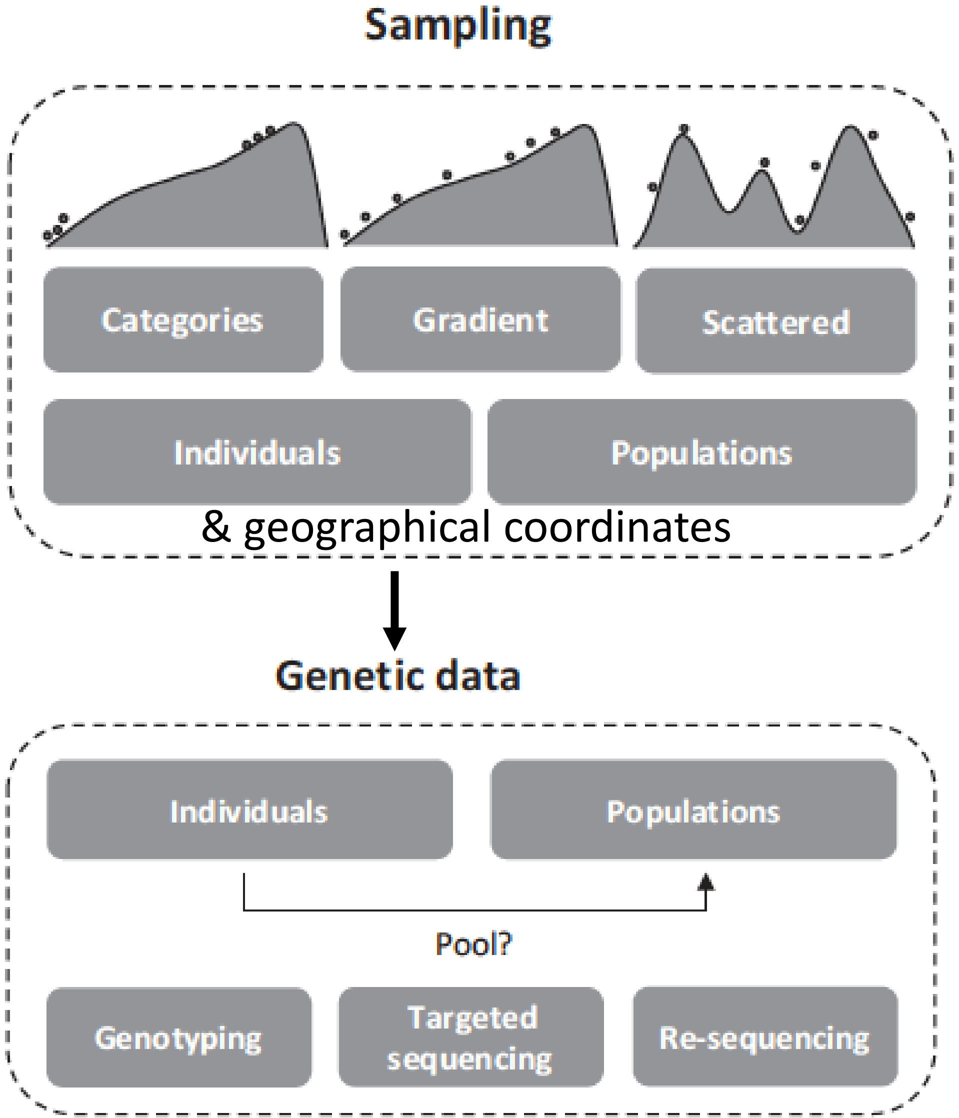
your approach to detecting selection will depend upon your sample design and study goal

selection between populations

goal: identify loci	undergoing recent selection (with or w/out phenotype) *divergent* across space	underlying important functional variation across space
signature: variants/regions	that depart from neutral or null expectations	associated with segregating functional variation
approaches:	sequence-based tests of selection differentiation-based tests environmental association analyses	association studies

differentiation-based tests

population genomics + space



sequence-based tests (in 2+ pops)

differentiation-based analyses
e.g. Fst outliers

differentiation-based tests

Goal:

Identify markers/variants/SNPs that show interesting (elevated) patterns of differentiation among 2+ populations

Rationale:

• populations in different (spatial) locations experience different selective conditions

• markers physically linked with (adjacent to) locally-adapted loci should show *elevated/exaggerated patterns of differentiation*, above background levels of population differentiation

differentiation-based tests

Requires:

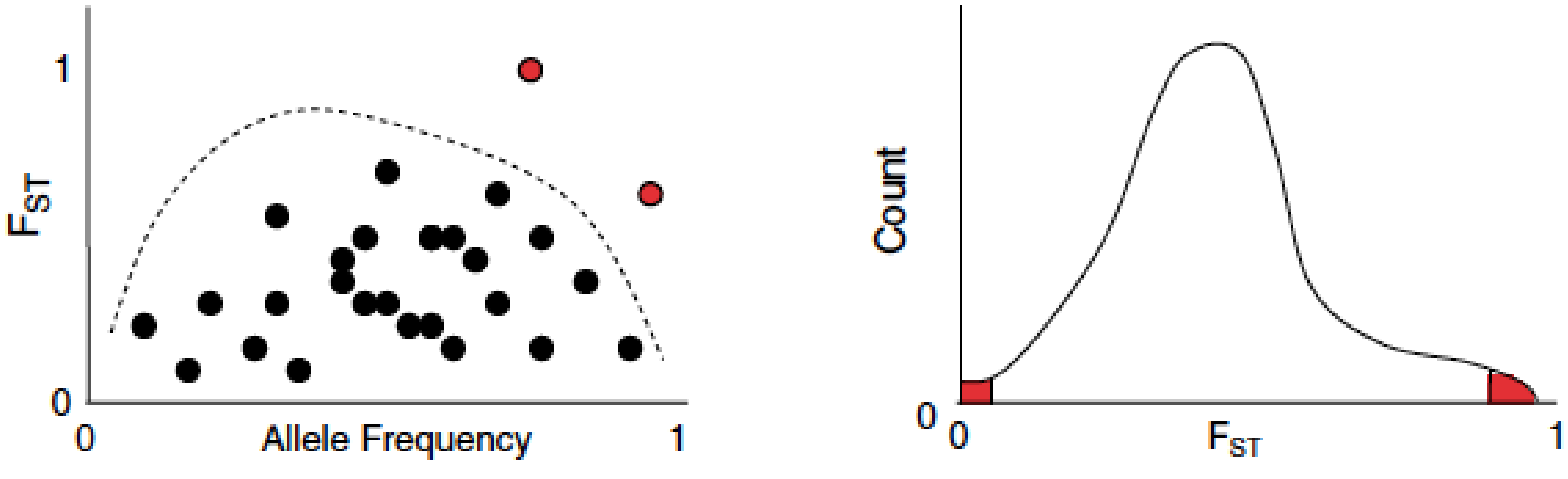
- markers/variant sites (1000’s to WG) that differ between individuals in 2+ populations
- null pop gen. or demographically informed models of expected variation among populations

!!super easy!!

In general:

- assess differentiation at every marker/locus across whole dataset
- identify markers loci that are *more differentiated* than expected (given historical demography and/or population structure)

differentiation-based tests

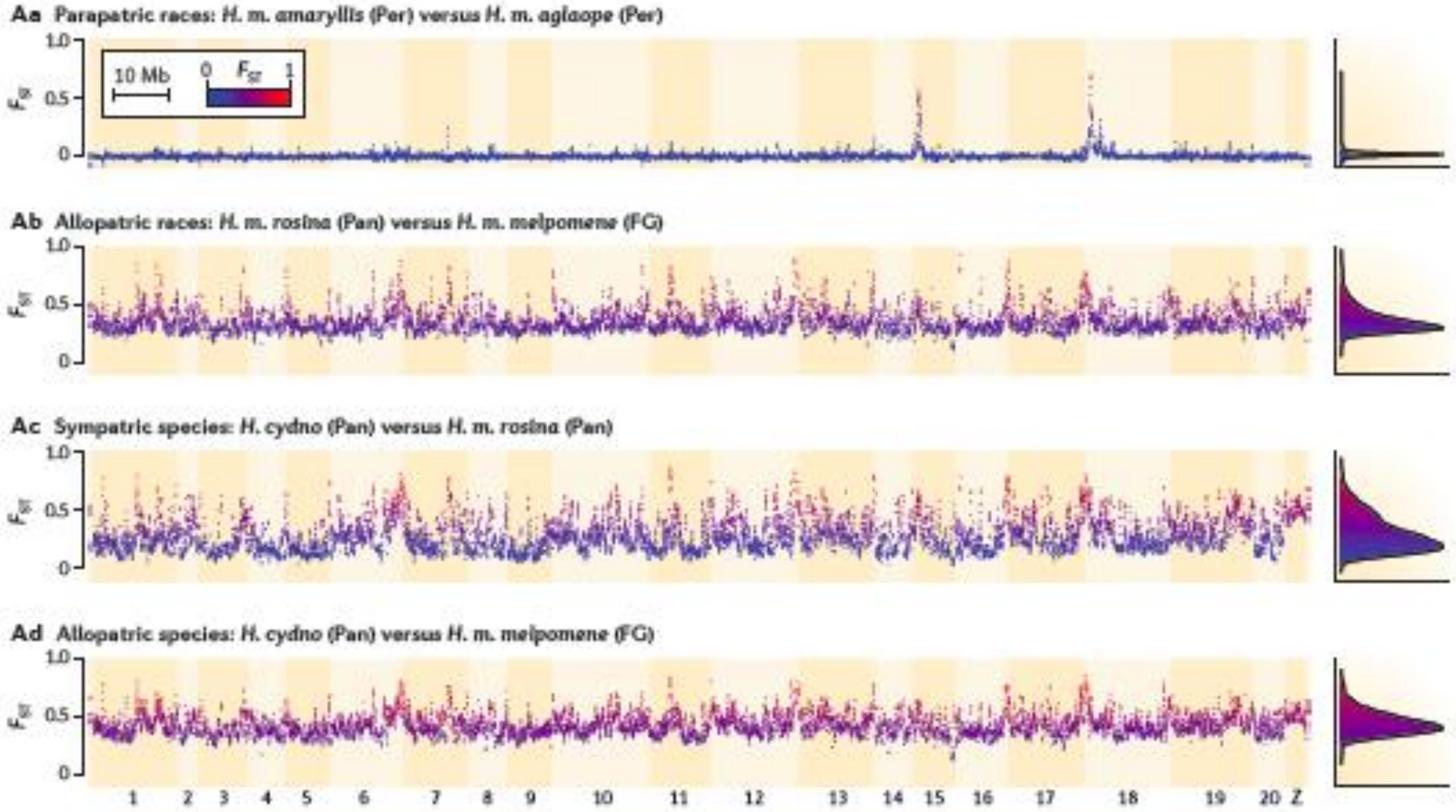


Outlier loci
identified statistically
(AFLP, microsat, SNP)

Are there any potential problems with this?

(Stinchcombe &Hoekstra 2008)

genomic heterogeneity in summary statistics
(often spatially correlated across the genome)



Examples…. different species pairs of Heliconius butterflies

differentiation-based tests

Requires:

- markers/variant sites (1000’s to WG) that differ between individuals in 2+ populations
- null pop gen. or demographically informed models of expected variation among populations

!!super easy!! …& potentially super misleading

In general:

- assess differentiation at every marker/locus across whole dataset
- identify markers loci that are *more differentiated* than expected (given historical demography and/or population structure)

differentiation-based tests

Requires:

- markers/variant sites (1000’s to WG) that differ between individuals in 2+ populations
- null pop gen. or demographically informed models of expected variation among populations
- additional data on genomic location and/or
- data on ecological or evolutionarily relevant variation….

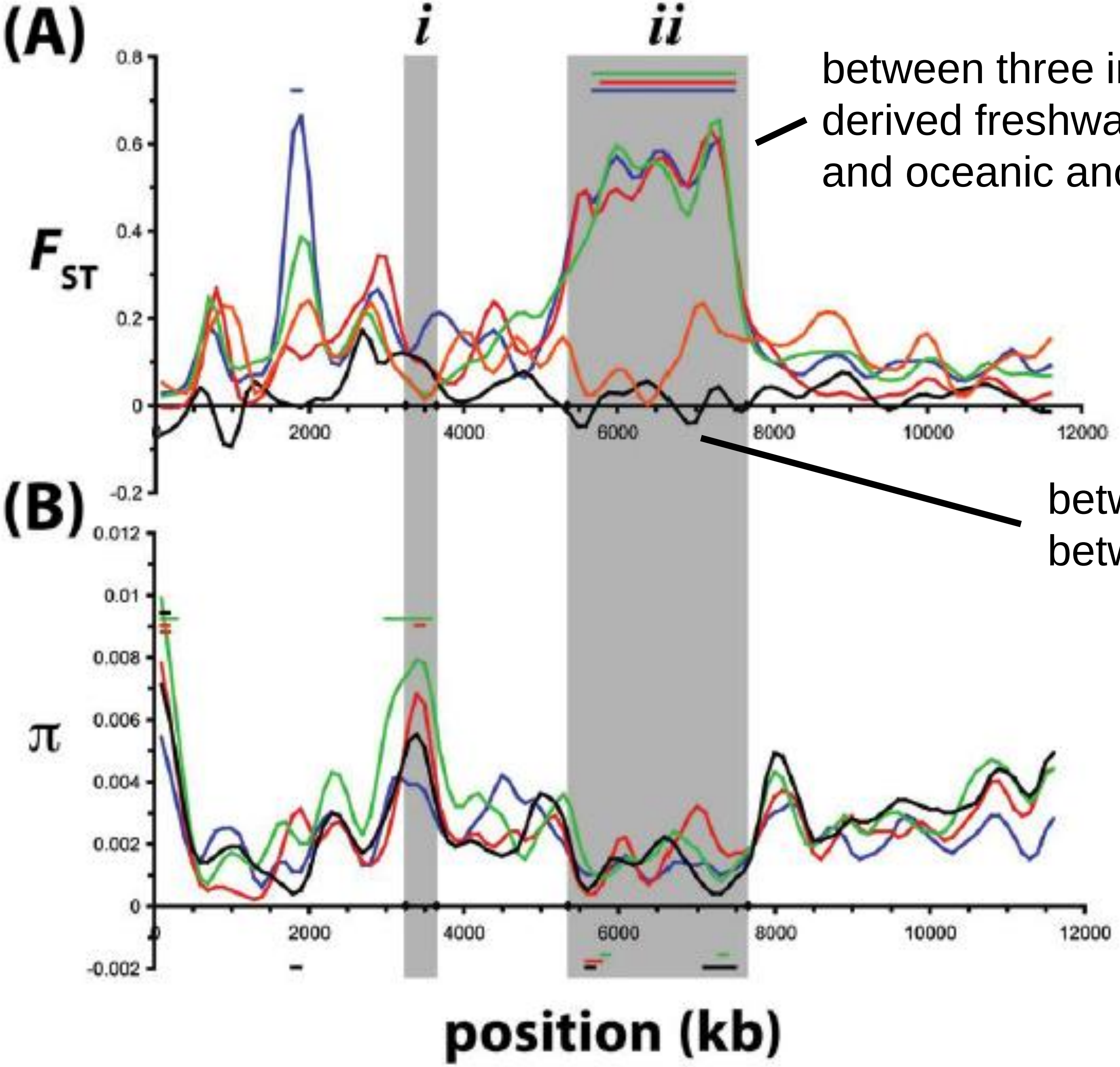
In general:

- assess differentiation at every marker/locus across whole dataset
- identify markers loci that are *more differentiated* than expected (given historical demography and/or population structure)

differentiation-based tests

Oceanic versus freshwater threespine stickleback

(Hohenlohe et al. 2010b)



between three independently derived freshwater populations and oceanic ancestral pops

between oceanic or between freshwater

(Figure 4 in Hohenlohe et al. 2010))

differentiation-based tests

limitations

- genome-wide heterogeneity in differentiation or diversity statistics can **produce spurious (non-causal) signals of elevated differentiation**

FALSE POSITIVES

- need additional data on ecological or evolutionary context to interpret patterns of pairwise differentiation

- still several steps away from direct causal inference….

differentiation-based tests

limitations

no phenotypes, no fitness,
so no direct information on:

- selective conditions/agents
- locus identity (depending on system…)
- functional importance (“adaptation”)

selection between populations

goal: identify loci	undergoing recent selection (with or w/out phenotype) *divergent* across space	underlying important functional variation across space
signature: variants/regions	that depart from neutral or null expectations	associated with segregating functional variation
approaches:	sequence-based tests of selection differentiation-based tests environmental association analyses	association studies

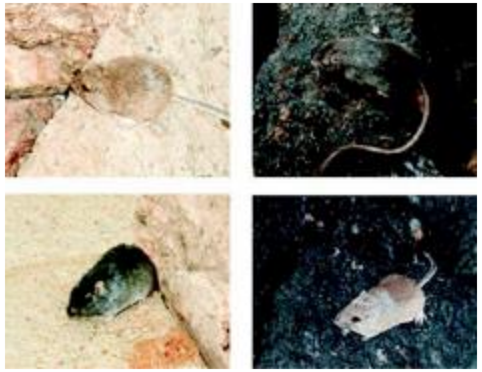
divergent selection between populations

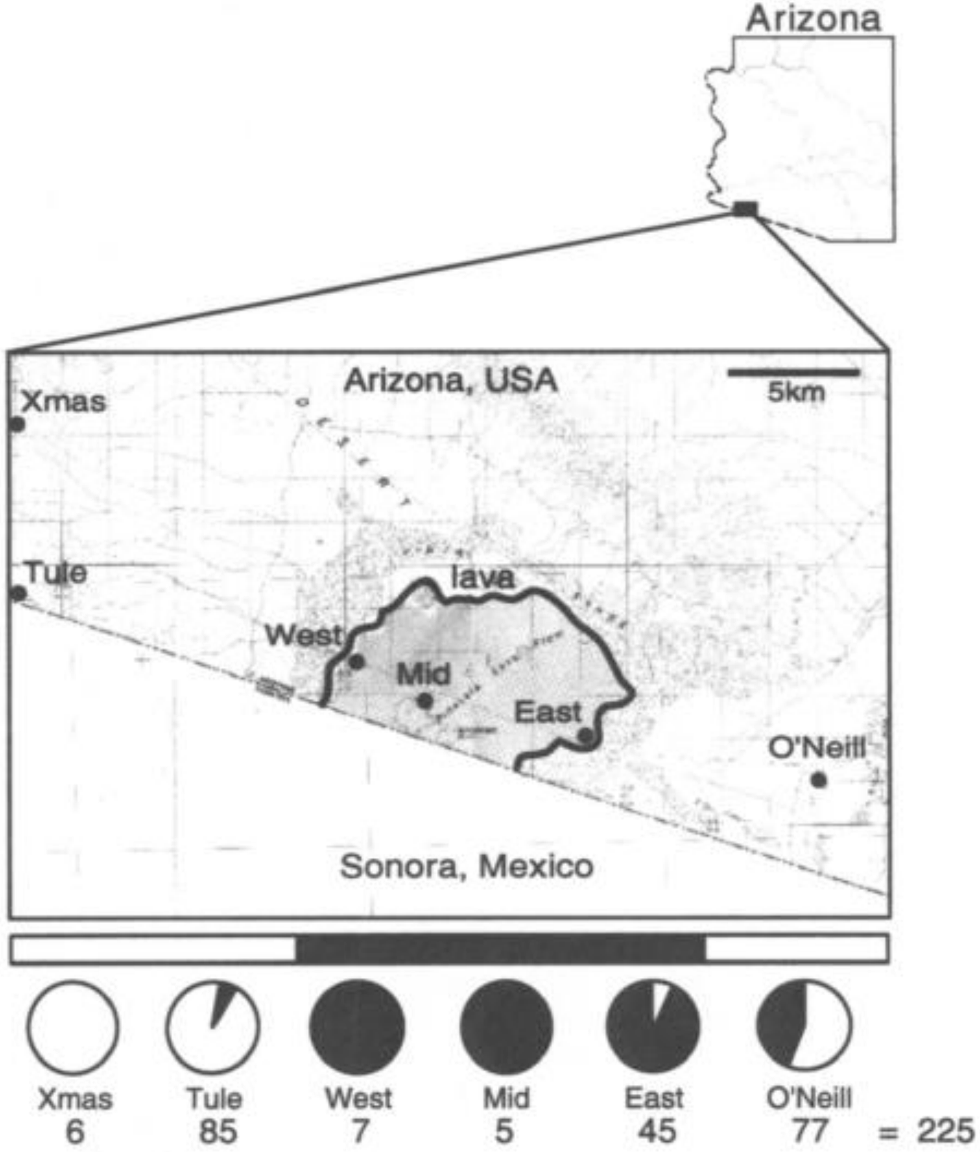
population genomics + space + environmental variation

environmental association analyses (EAA)
genotype x environment analyses (GEA)
(within “landscape genomics”)

environmental association analyses (EAA)

the conceptual origins of EAA are from classical clinal analyses





SNP

trait-environment associations

environmental association analyses (EAA)

EAAs are essentially association studies but association with environments not traits

surprise!

Goal:

Identify markers/variants/SNPs statistically associated with variation in environmental factors of interest, due to LD with causal loci

Rationale:

- populations in different (spatial) locations experience different selective conditions

- markers physically linked with locally-adapted loci should show *statistical associations with the causal selective agent*, above background levels of SNP-environment associations

environmental association analyses (EAA)

use SNP-environmental associations to infer things like:

- specific genomic targets of environmental selection (loci)
- specific environmental components that impose selection (agents)
- contribution of spatially-varying (abiotic) selection
  to genome-wide genomic variation
- parallel versus unique responses to
  repeated environmental gradients

environmental association analyses (EAA)

Requires:

- markers/variant sites (1000’s to WG) that differ between individuals
- linkage map or genome sequence (ideally)
- quantitative environmental data (univariate or multivariate)
- methods to associate environment/genotype (and exclude confounding factors, correct for multiple testing)

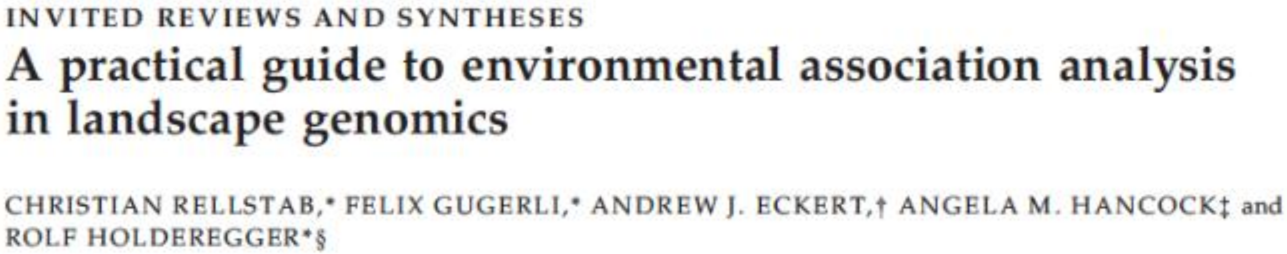
In general:

- test each marker OR composite genotypes associations with single environmental factors OR multivariate environmental variation

- assess and control/account for population processes (**especially historical demography and/or population structure/relatedness**) EITHER sequentially or simultaneously.

EAA is really a heterogeneous set of tools and approaches

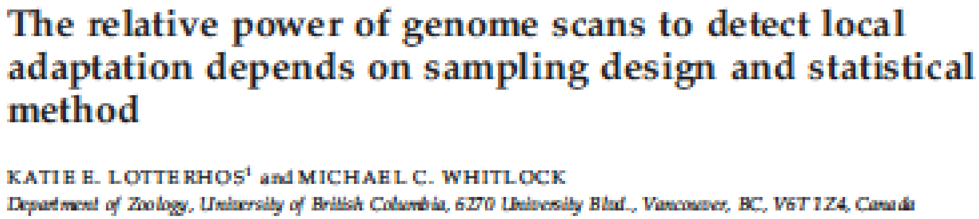
environmental association analyses (EAA)



INVITED REVIEWS AND SYNTHESES

A practical guide to environmental association analysis in landscape genomics

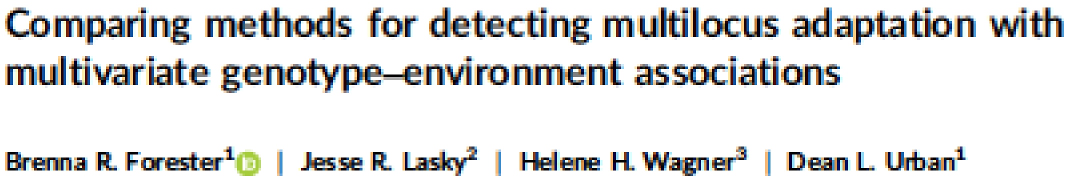
CHRISTIAN RELLSTAB,* FELIX GUGERLI,* ANDREW J. ECKERT,† ANGELA M. HANCOCK‡ and ROLF HOLDEREGGER*§



The relative power of genome scans to detect local adaptation depends on sampling design and statistical method

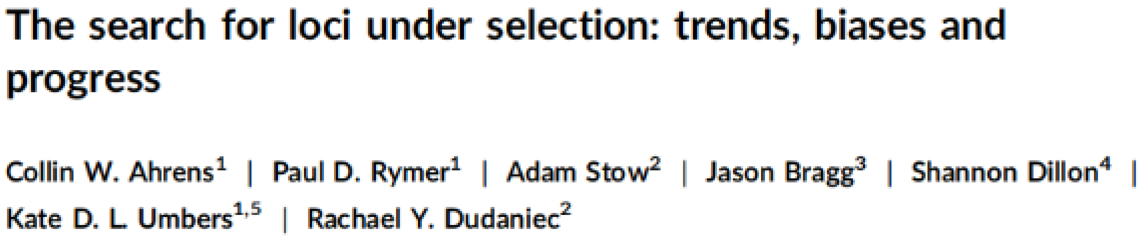
KATIE E. LOTTERHOS[1] and MICHAEL C. WHITLOCK

Department of Zoology, University of British Columbia, 6270 University Blvd., Vancouver, BC, V6T 1Z4, Canada



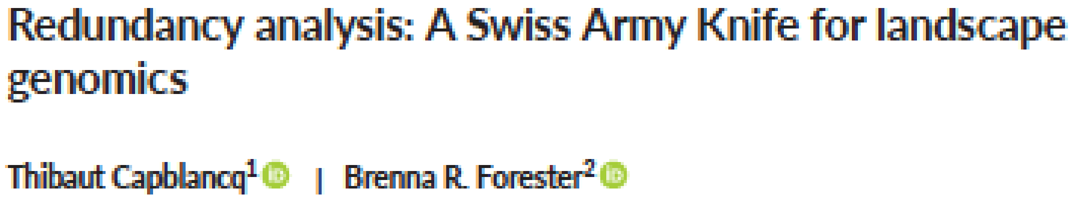
Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations

Brenna R. Forester[1] | Jesse R. Lasky[2] | Helene H. Wagner[3] | Dean L. Urban[1]



The search for loci under selection: trends, biases and progress

Collin W. Ahrens[1] | Paul D. Rymer[1] | Adam Stow[2] | Jason Bragg[3] | Shannon Dillon[4] | Kate D. L. Umbers[1,5] | Rachael Y. Dudaniec[2]



Redundancy analysis: A Swiss Army Knife for landscape genomics

Thibaut Capblancq[1] | Brenna R. Forester[2]

and more…

EAA is really a heterogeneous set of tools and approaches

environmental association analyses (EAA)

(half of)
Table 1 from
Rellstab et al. 2015

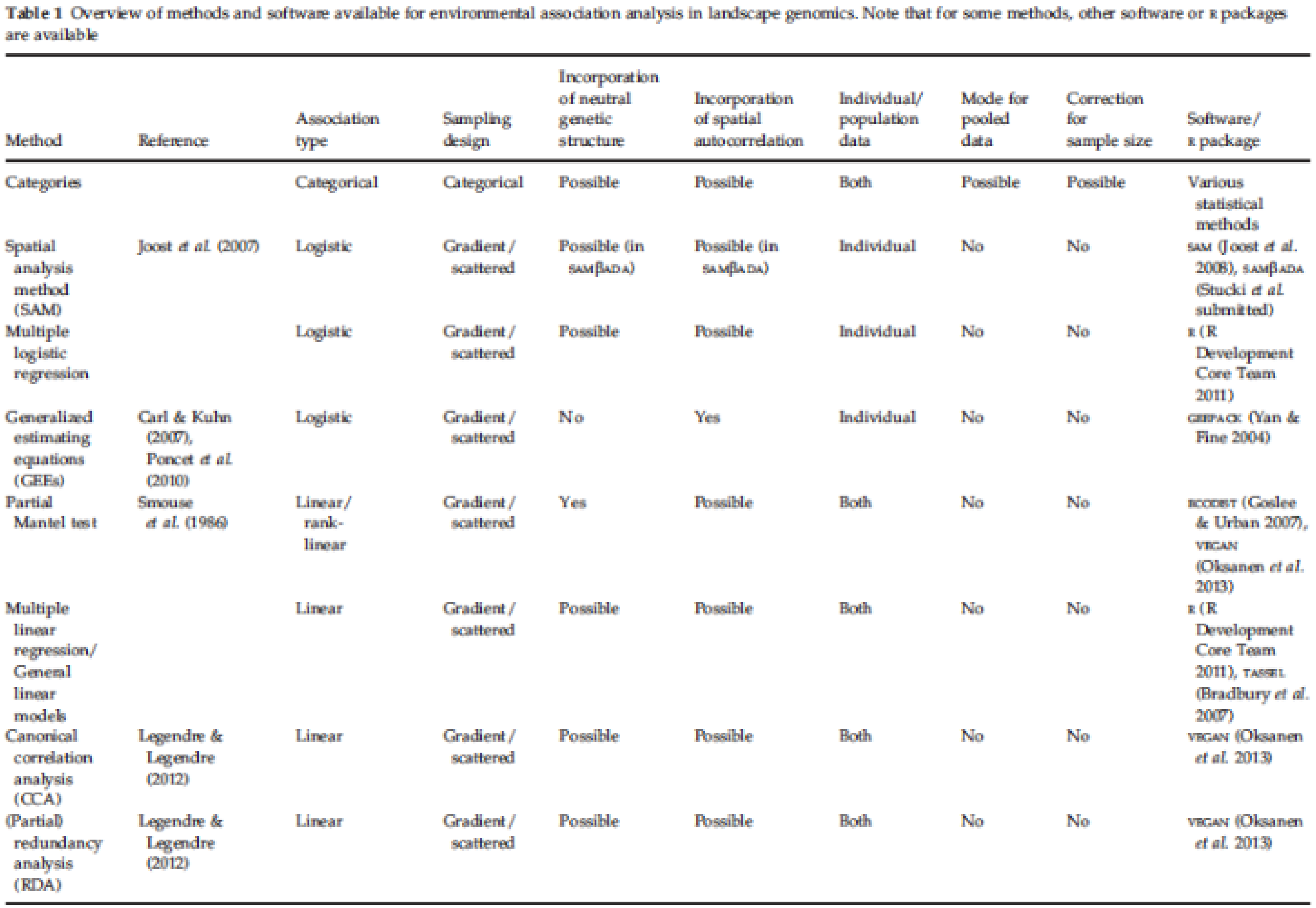


Table 1 Overview of methods and software available for environmental association analysis in landscape genomics. Note that for some methods, other software or R packages are available

Method	Reference	Association type	Sampling design	Incorporation of neutral genetic structure	Incorporation of spatial autocorrelation	Individual/ population data	Mode for pooled data	Correction for sample size	Software/ R package
Categories		Categorical	Categorical	Possible	Possible	Both	Possible	Possible	Various statistical methods
Spatial analysis method (SAM)	Joost *et al.* (2007)	Logistic	Gradient/ scattered	Possible (in SAMβADA)	Possible (in SAMβADA)	Individual	No	No	SAM (Joost *et al.* 2008), SAMβADA (Stucki *et al.* submitted)
Multiple logistic regression		Logistic	Gradient/ scattered	Possible	Possible	Individual	No	No	R (R Development Core Team 2011)
Generalized estimating equations (GEEs)	Carl & Kuhn (2007), Poncet *et al.* (2010)	Logistic	Gradient/ scattered	No	Yes	Individual	No	No	GEEPACK (Yan & Fine 2004)
Partial Mantel test	Smouse *et al.* (1986)	Linear/ rank-linear	Gradient/ scattered	Yes	Possible	Both	No	No	ECODIST (Goslee & Urban 2007), VEGAN (Oksanen *et al.* 2013)
Multiple linear regression/ General linear models		Linear	Gradient/ scattered	Possible	Possible	Both	No	No	R (R Development Core Team 2011), TASSEL (Bradbury *et al.* 2007)
Canonical correlation analysis (CCA)	Legendre & Legendre (2012)	Linear	Gradient/ scattered	Possible	Possible	Both	No	No	VEGAN (Oksanen *et al.* 2013)
(Partial) redundancy analysis (RDA)	Legendre & Legendre (2012)	Linear	Gradient/ scattered	Possible	Possible	Both	No	No	VEGAN (Oksanen *et al.* 2013)

EAA is really a heterogeneous set of tools and approaches

environmental association analyses (EAA)

tools vary depending upon the question(s), and:

➡ • distribution of samples across space and/or environment

• type of model (e.g. logistic regression, matrix correlation, mixed-effects models)

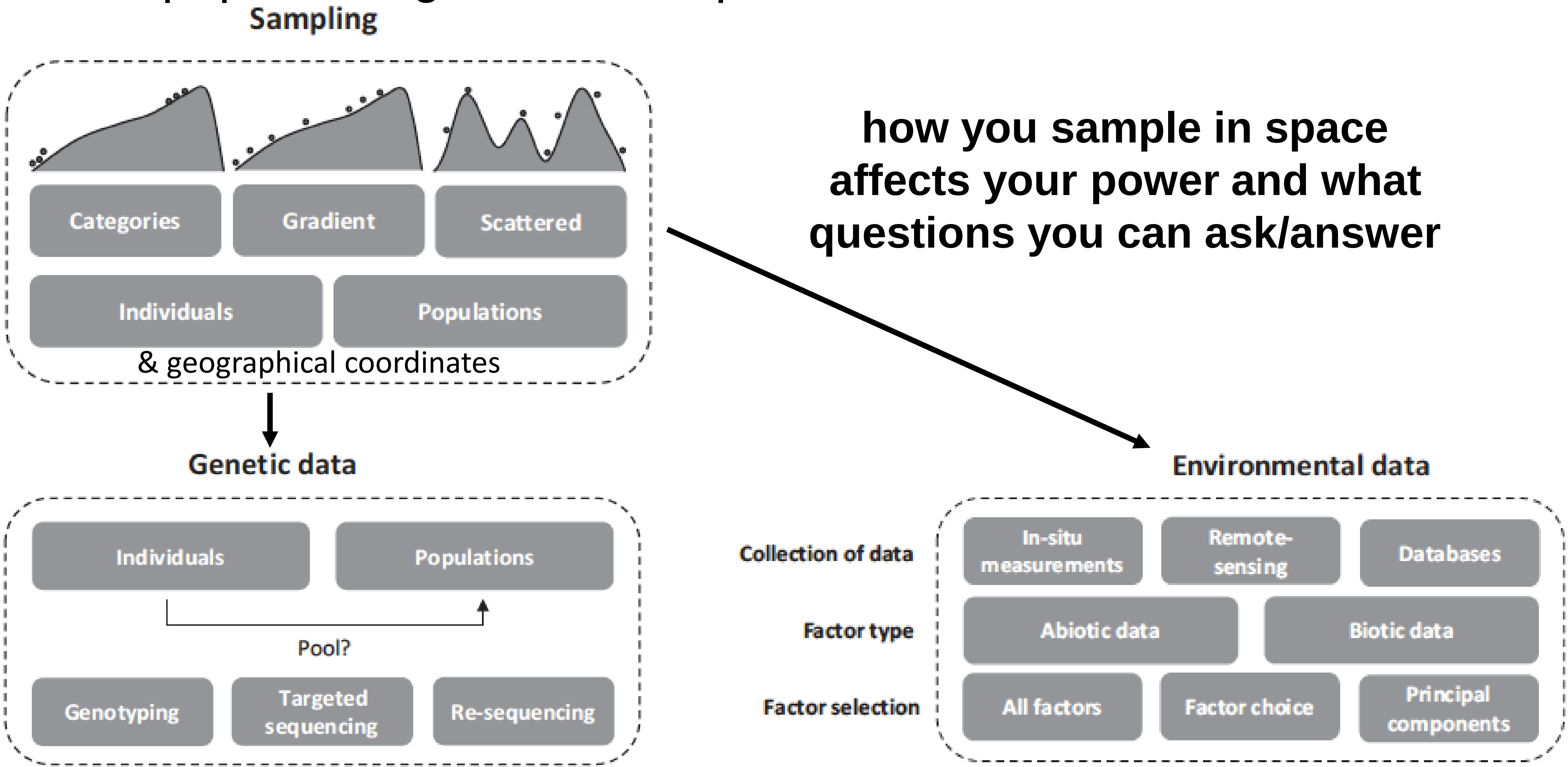
• statistical procedure used (e.g. FDR, p-values)

➡ • method of handling/accounting for population structure

EAA is really a heterogeneous set of tools and approaches

divergent selection between populations

population genomics + space + environmental variation



how you sample in space affects your power and what questions you can ask/answer

Sampling

Categories

Gradient

Scattered

Individuals

Populations

& geographical coordinates

Genetic data

Individuals

Populations

Pool?

Genotyping

Targeted sequencing

Re-sequencing

Environmental data

Collection of data

In-situ measurements

Remote-sensing

Databases

Factor type

Abiotic data

Biotic data

Factor selection

All factors

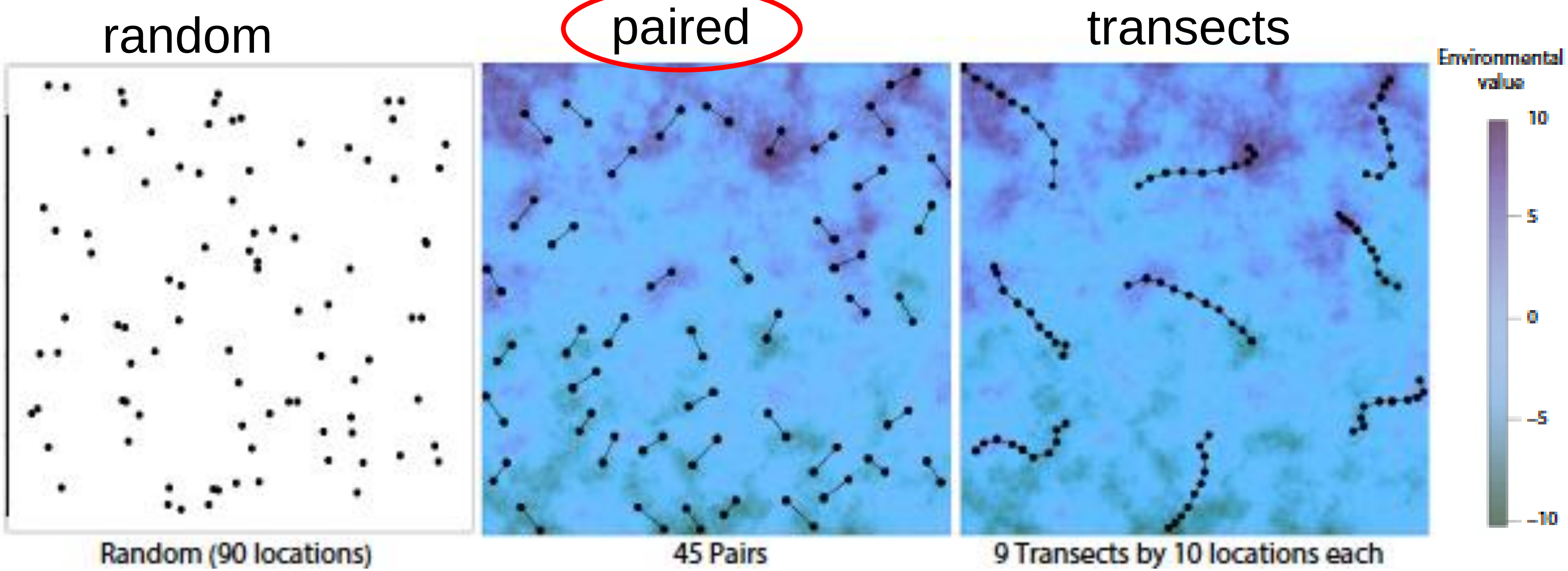
Factor choice

Principal components

sampling for EAA

e.g. power

can detect weaker selection (but also depends on many other factors…)



Lotterhos & Whitlock 2015

sampling for EAA

e.g. power

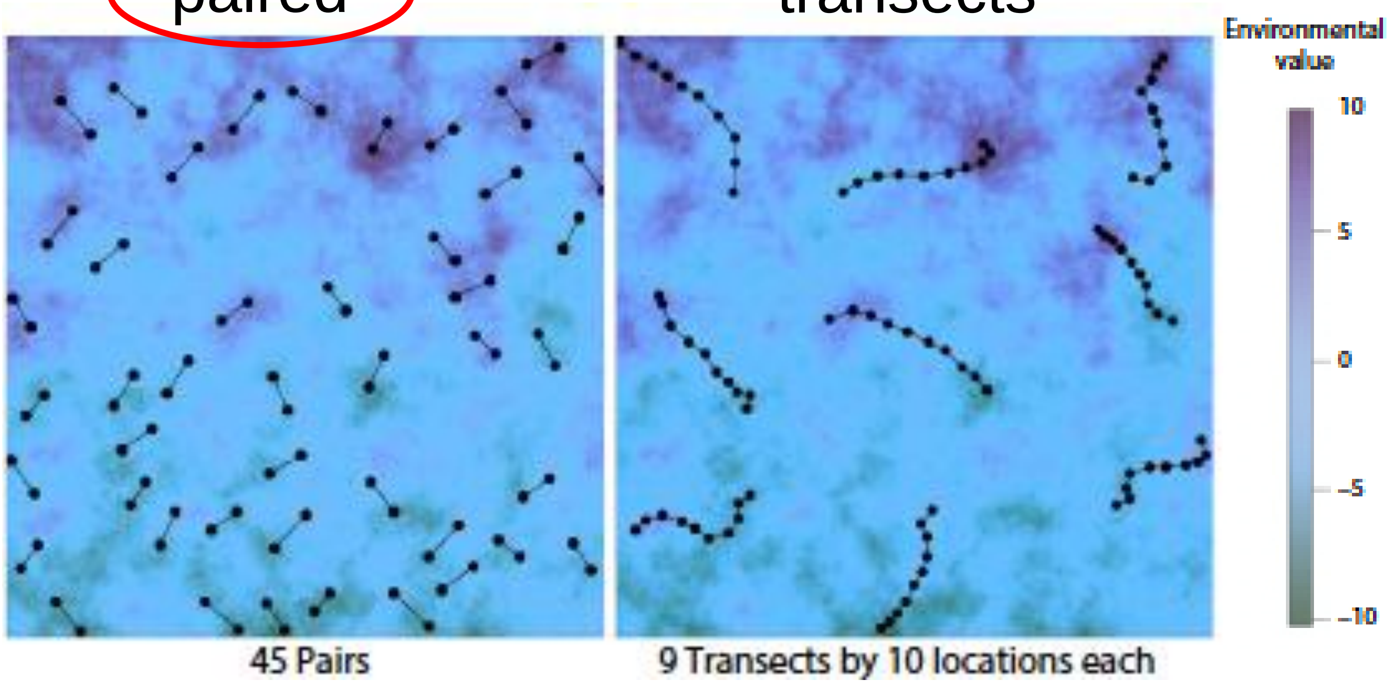
when pairs span repeated categorical contrast: "quasi-experimental"
(Rellstab et al. 2015)

hot-cold
high-low
on-off

can detect weaker selection (but also depends on many other factors…)

paired

transects



Lotterhos & Whitlock 2015

sampling for EAA

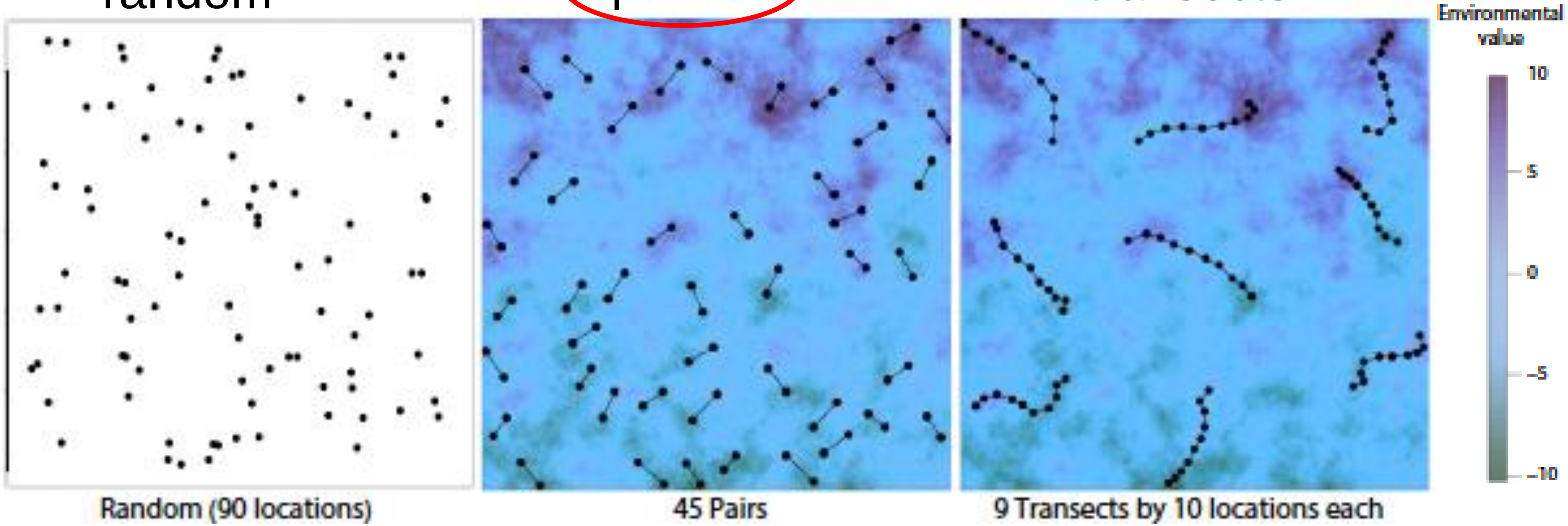
e.g. power

can detect weaker selection (but also depends on many other factors…)

random

paired

transects



Random (90 locations)

45 Pairs

9 Transects by 10 locations each

no *a priori* knowledge required

binary selective agent?

clinal selective agent?

e.g. distribution of environmental factors

parallel responses to selection

Lotterhos & Whitlock 2015

sampling for EAA

e.g. distribution of environmental factors

how you sample in space affects your power and what questions you can ask/answer

e.g. population genomic factors

individual-based analyses better when:

- many coordinates
- enviro data has high variation across sampling area
- local Ne is low

population-based analyses better when:

- samples are clustered at local sites
- enviro data changes at scales >> than local samples
- local Ne is higher

sampling for EAA

e.g. distribution of environmental factors

how you sample in space affects your power and what questions you can ask/answer

e.g. population genomic factors

individual- versus population-based analyses

Both also affect how to incorporate demographic/historical/neutral genetic structure into an EAA

Why do we care about population structure?

population structure—heterogeneous genetic relationships among individuals—**creates patterns of LD in a dataset**

that have NOTHING to do with adaptive variation.

!!when population history is correlated with distribution of trait variation, false positives!!

(!!similarly, **when population history is correlated with the environment**, false positives!!)

FALSE POSITIVES

different methods incorporate population structure in different ways

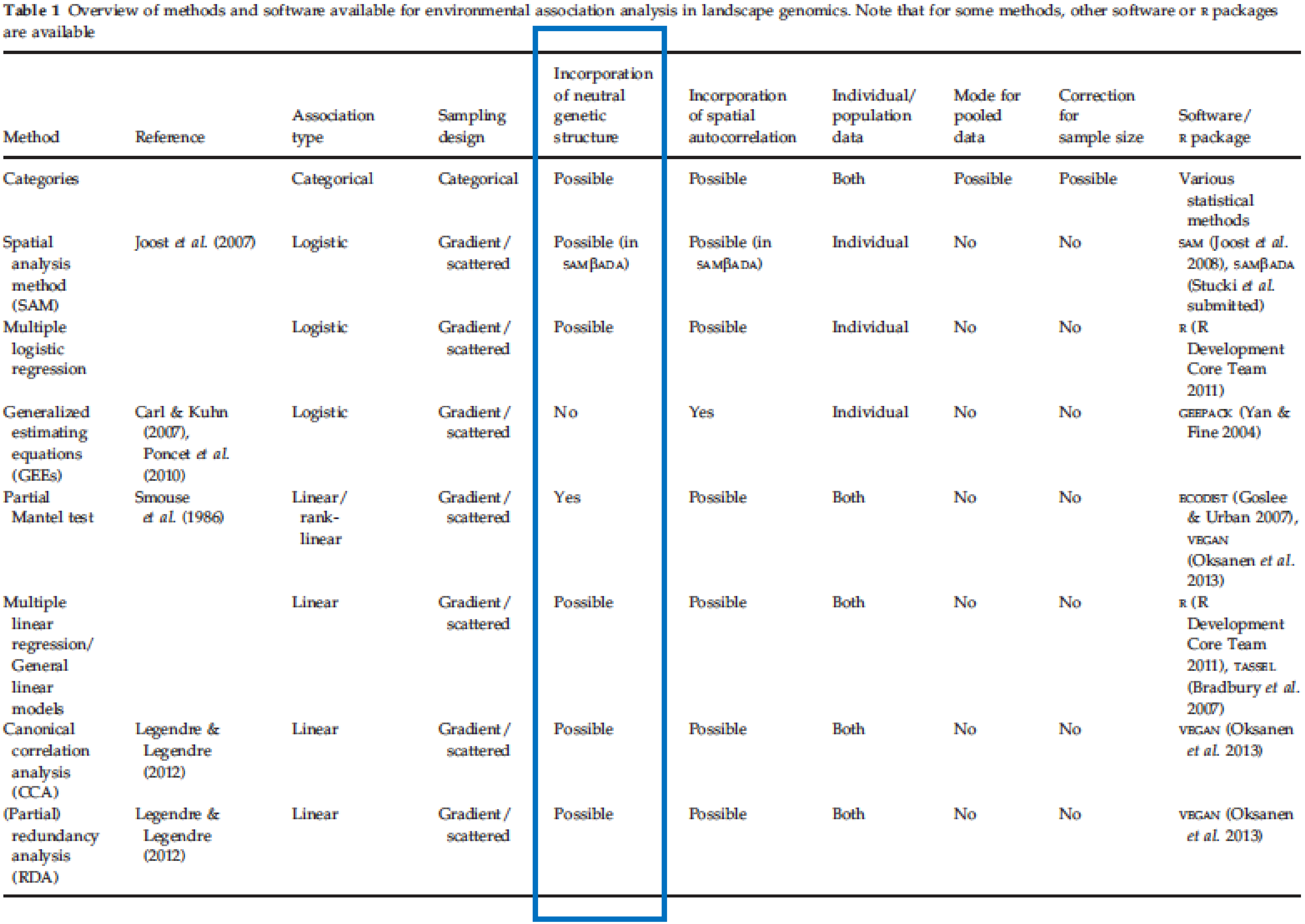


Table 1 Overview of methods and software available for environmental association analysis in landscape genomics. Note that for some methods, other software or R packages are available

Method	Reference	Association type	Sampling design	Incorporation of neutral genetic structure	Incorporation of spatial autocorrelation	Individual/ population data	Mode for pooled data	Correction for sample size	Software/ R package
Categories		Categorical	Categorical	Possible	Possible	Both	Possible	Possible	Various statistical methods
Spatial analysis method (SAM)	Joost *et al.* (2007)	Logistic	Gradient/ scattered	Possible (in SAMβADA)	Possible (in SAMβADA)	Individual	No	No	SAM (Joost *et al.* 2008), SAMβADA (Stucki *et al.* submitted)
Multiple logistic regression		Logistic	Gradient/ scattered	Possible	Possible	Individual	No	No	R (R Development Core Team 2011)
Generalized estimating equations (GEEs)	Carl & Kuhn (2007), Poncet *et al.* (2010)	Logistic	Gradient/ scattered	No	Yes	Individual	No	No	GEEPACK (Yan & Fine 2004)
Partial Mantel test	Smouse *et al.* (1986)	Linear/ rank-linear	Gradient/ scattered	Yes	Possible	Both	No	No	ECODIST (Goslee & Urban 2007), VEGAN (Oksanen *et al.* 2013)
Multiple linear regression/ General linear models		Linear	Gradient/ scattered	Possible	Possible	Both	No	No	R (R Development Core Team 2011), TASSEL (Bradbury *et al.* 2007)
Canonical correlation analysis (CCA)	Legendre & Legendre (2012)	Linear	Gradient/ scattered	Possible	Possible	Both	No	No	VEGAN (Oksanen *et al.* 2013)
(Partial) redundancy analysis (RDA)	Legendre & Legendre (2012)	Linear	Gradient/ scattered	Possible	Possible	Both	No	No	VEGAN (Oksanen *et al.* 2013)

(half of)
Table 1 from
Rellstab et al. 2015

some common approaches for EAA

e.g. (LFMM) Latent Factor MM

LMM that uses environment (specific climate variables) as a fixed effect

incorporates population structure
by using K (e.g. STRUCTURE) as latent factors (representing random effects)

environmental effect and population structure are assessed simultaneously

some common approaches for EAA

e.g. BAYENV

LMM method to assess evidence for
correlation (of SNPs) with environment
(specific climate variables)

<u>incorporates population structure</u>
by generating a kinship matrix from allelic data, to
estimate a null model of demographic structure

compares models (in a Bayesian framework)
that do (alternative) and do not (null)
include environment

some common approaches for EAA

e.g. Redundancy Analysis (RDA)

Multiple linear regression method for testing associations between SNPs and multivariate environment

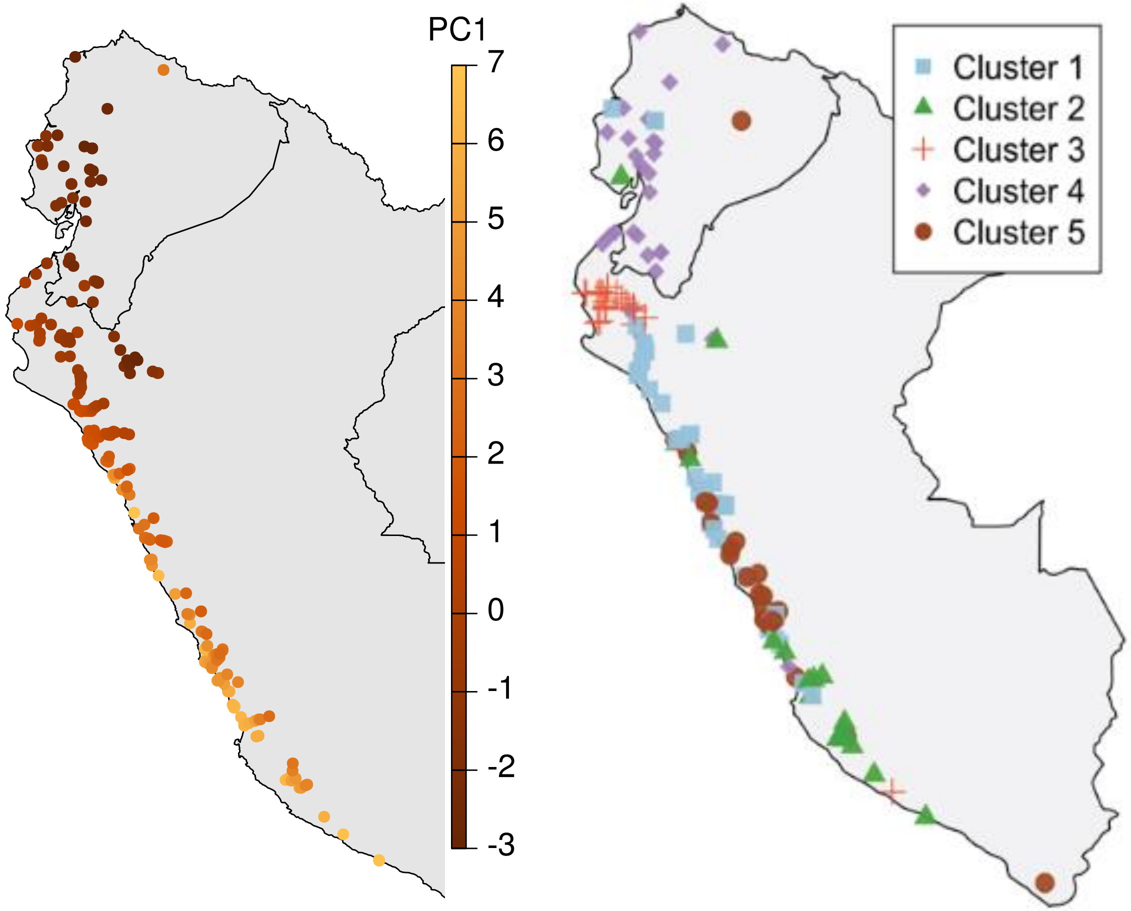
incorporates population structure
via constrained ordination matrix of spatial relationships

multivariate environmental effects and spatial (population) structure are assessed simultaneously

environmental association analyses (EAA)

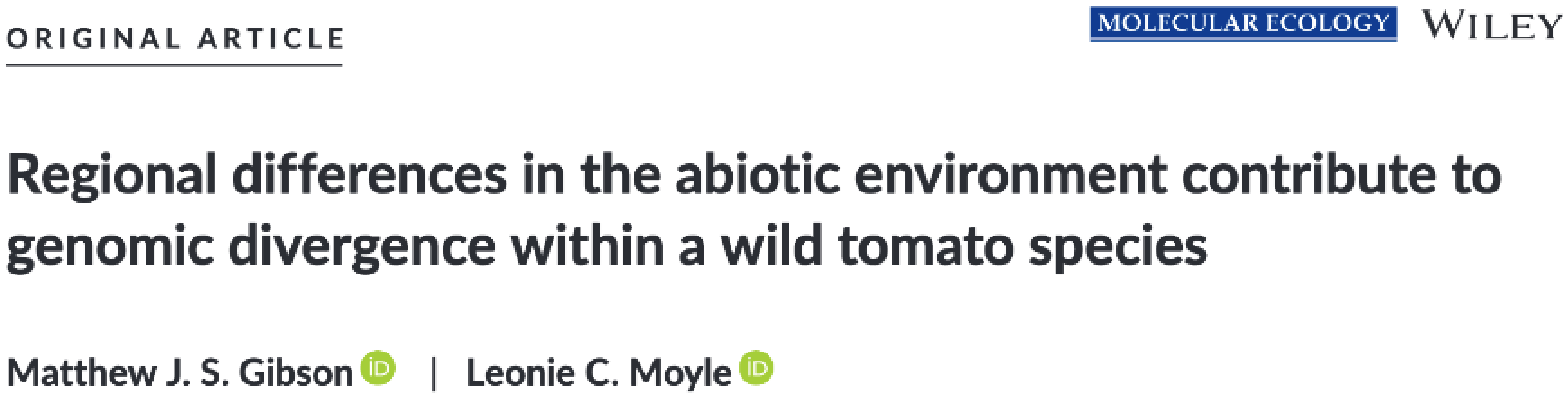
CASE STUDY

Case study: Landscape genomics of adaptation to abiotic climates



Gibson & Moyle, 2020 *Molecular Ecology*

Case study: Landscape genomics of adaptation to abiotic climates



ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

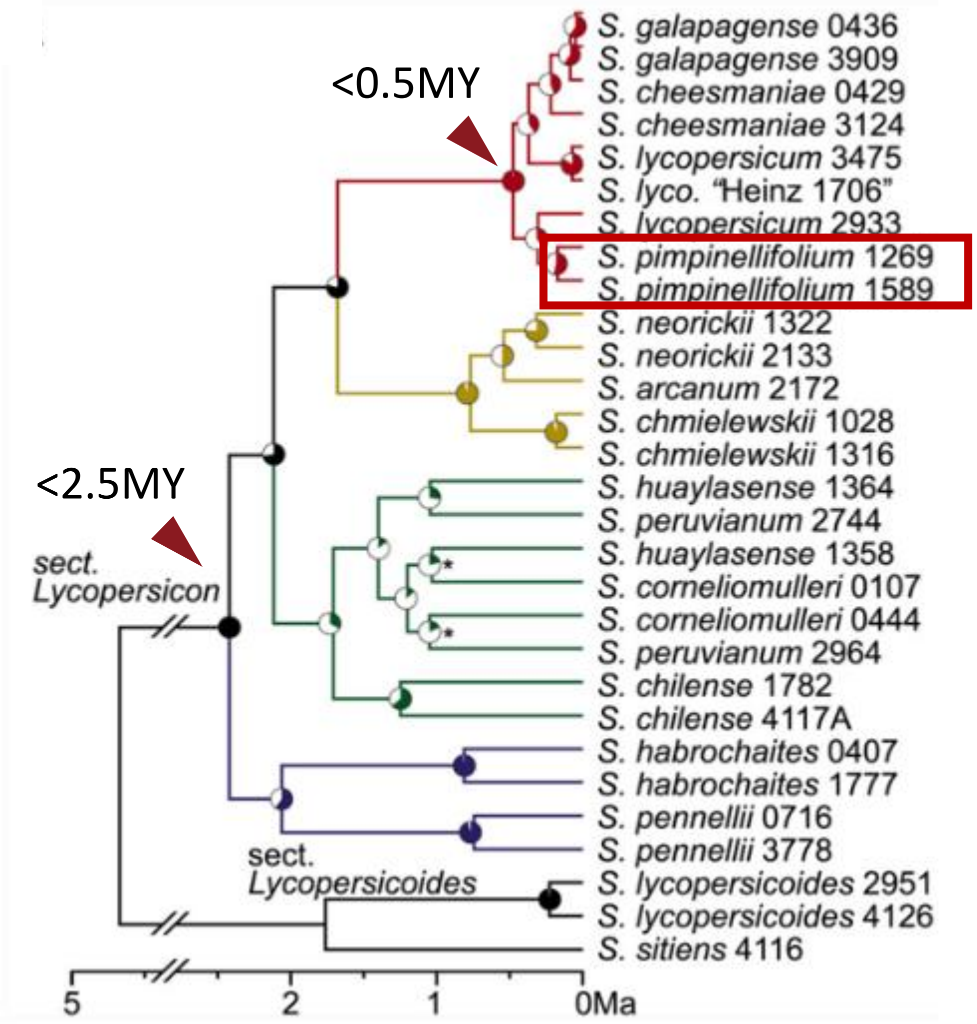
Regional differences in the abiotic environment contribute to genomic divergence within a wild tomato species

Matthew J. S. Gibson | Leonie C. Moyle

Matthew Gibson

Wild tomatoes

S. pimpinellifolium



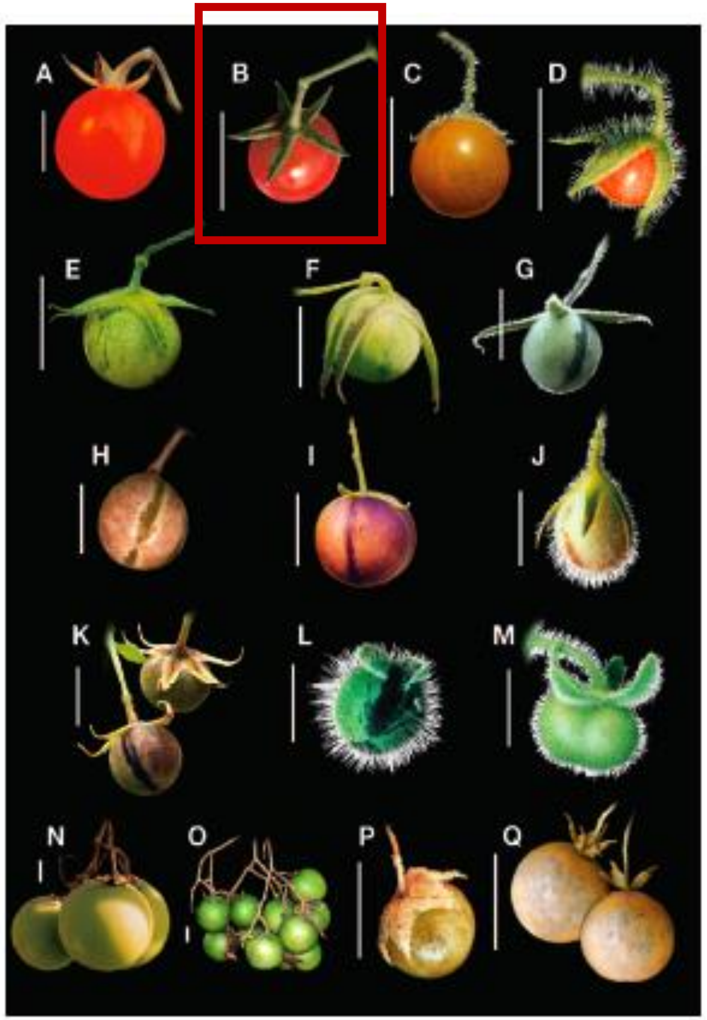
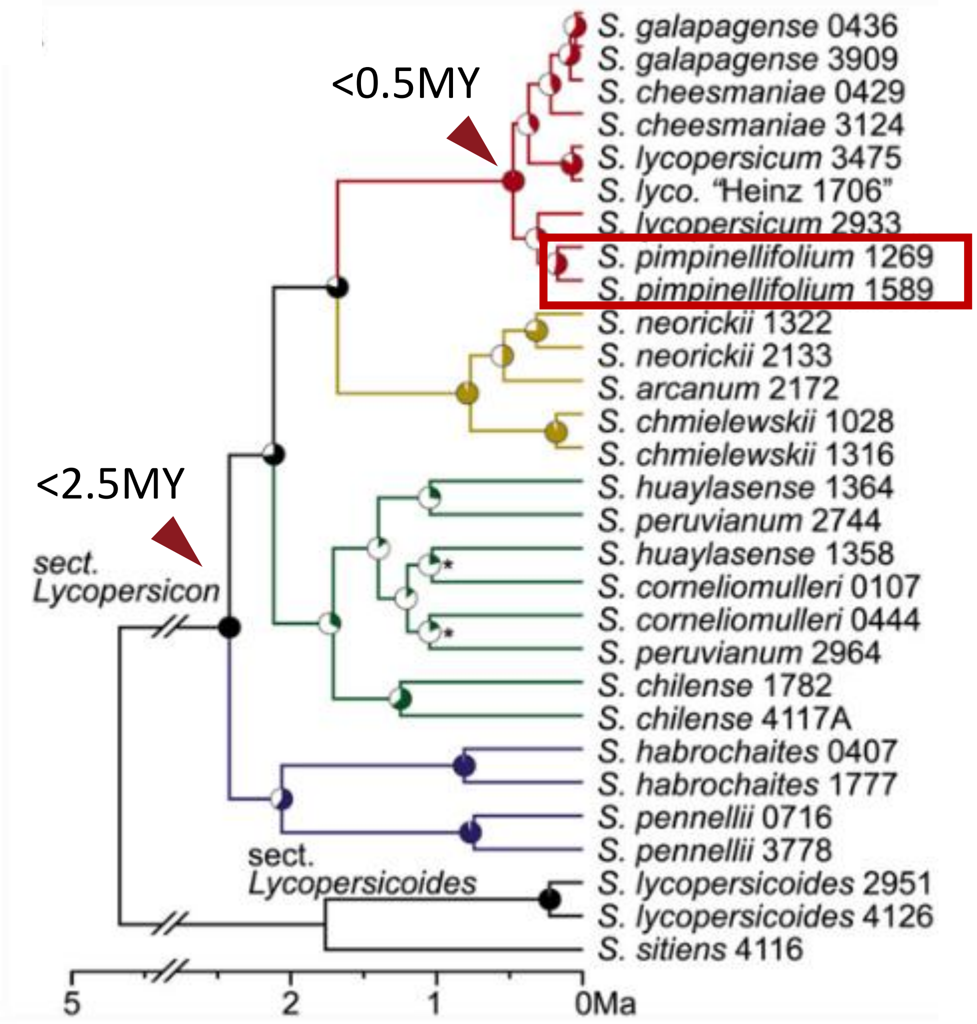


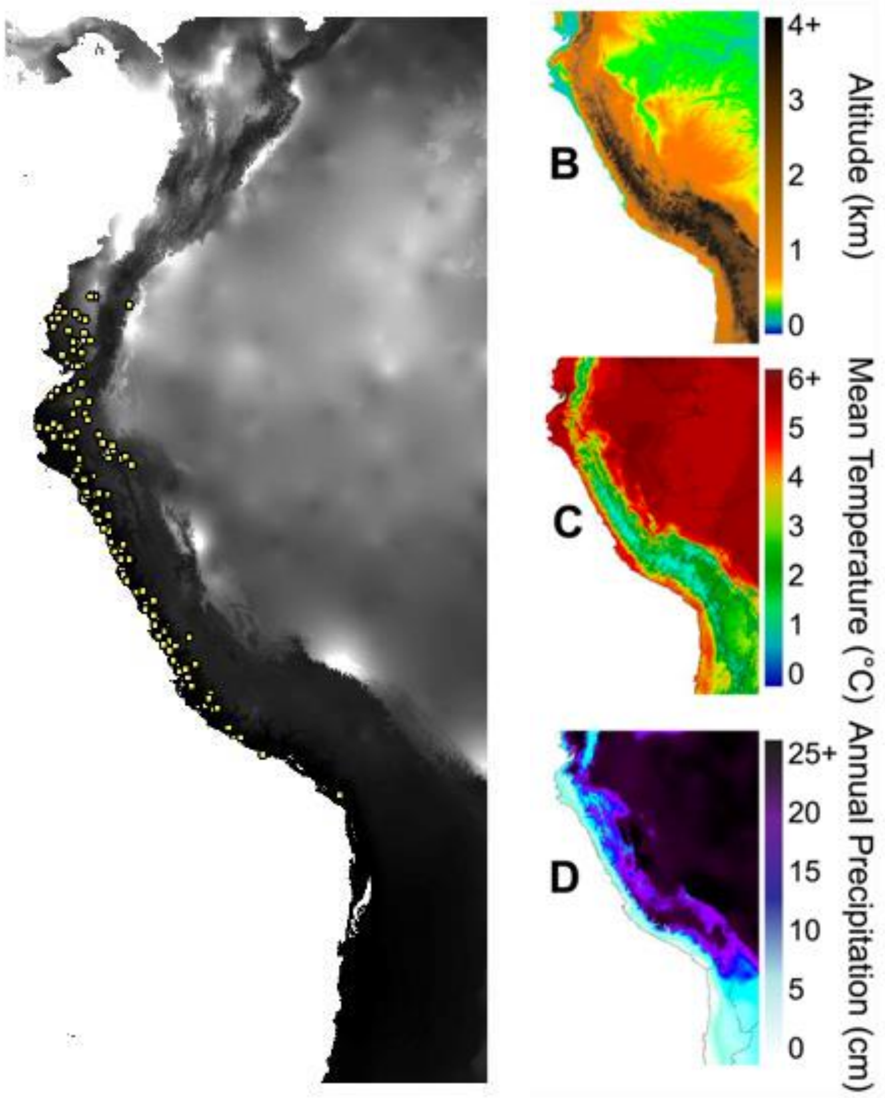
Image: Peralta et al. 2008

(Pease et al., 2016 *PLoS Biology)*

Wild tomatoes

S. pimpinellifolium



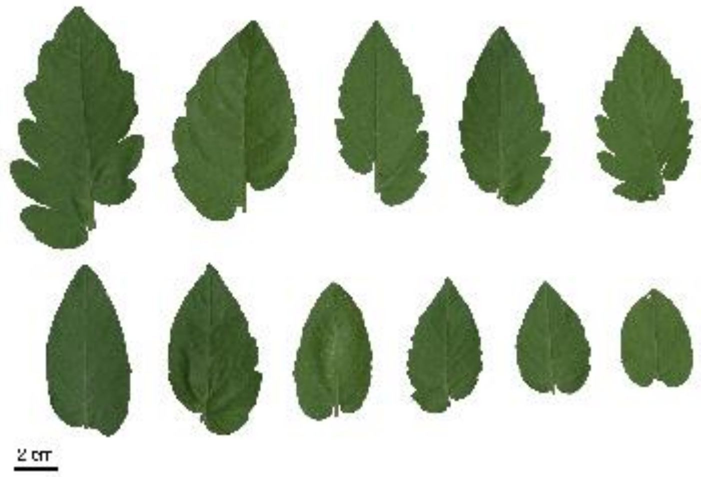


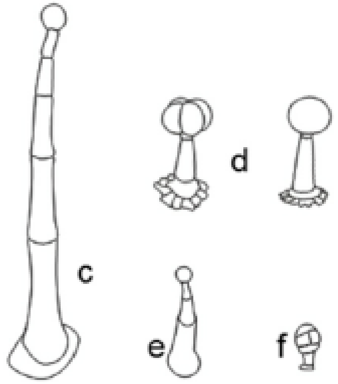
(Pease et al., 2016 *PLoS Biology)*

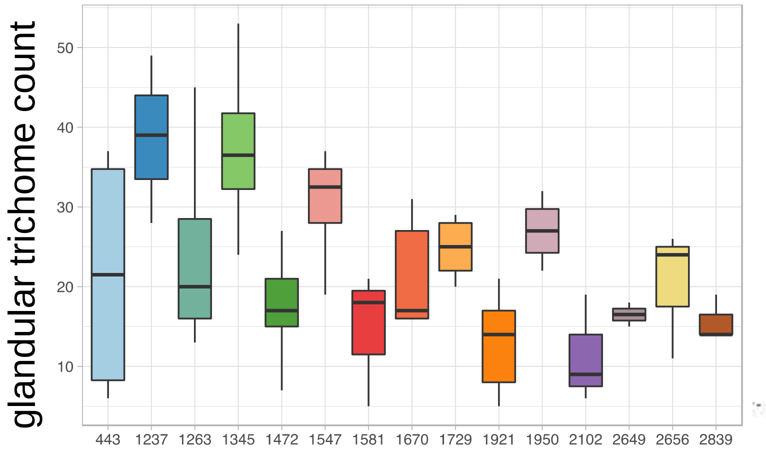
variable abiotic habitats

Wild tomatoes

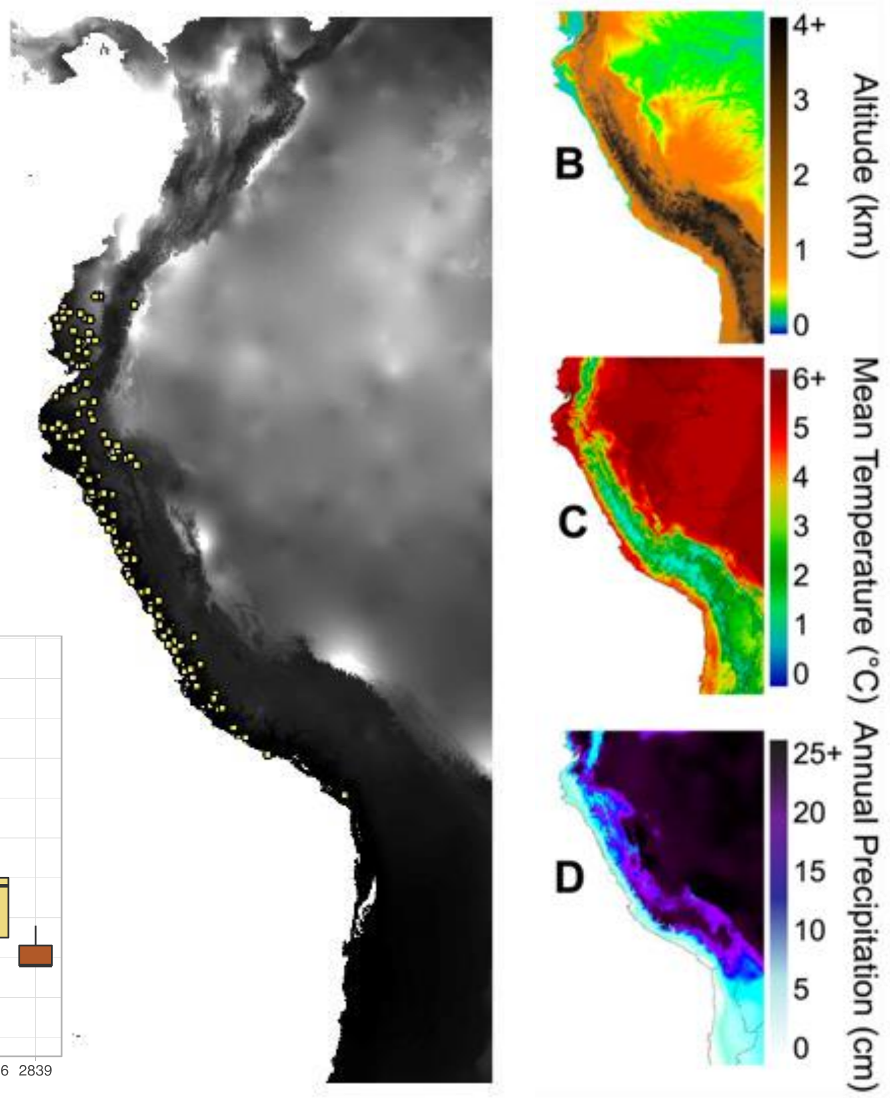
S. pimpinellifolium







quantitative trait diversity



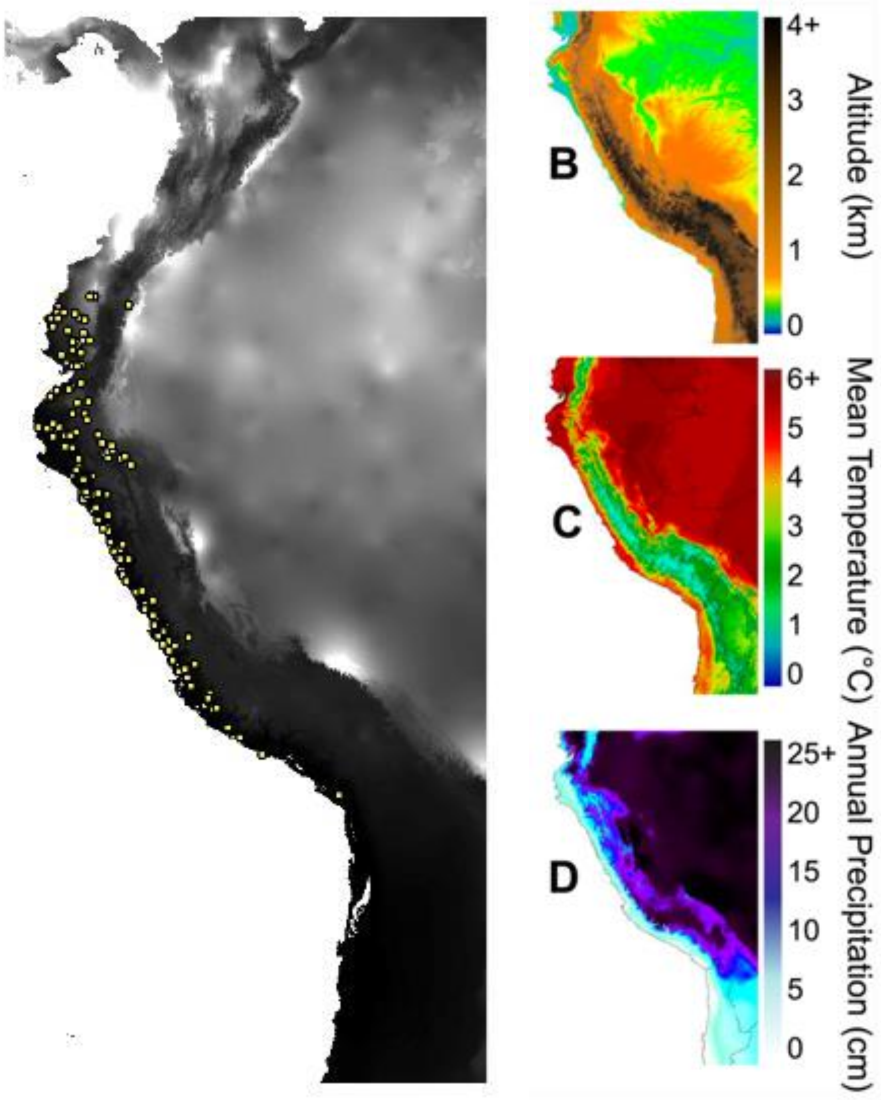
variable abiotic habitats

Wild tomatoes

S. pimpinellifolium

Abiotic conditions are proposed to shape numerous traits

- Days to wilting (Nakazato et al., 2008, 2010)
- Leaf shape (Chitwood et al, 2012)
- Shade response (Chitwood et al, 2012)
- Rooting depth (Nakazato et al., 2008)



quantitative trait diversity

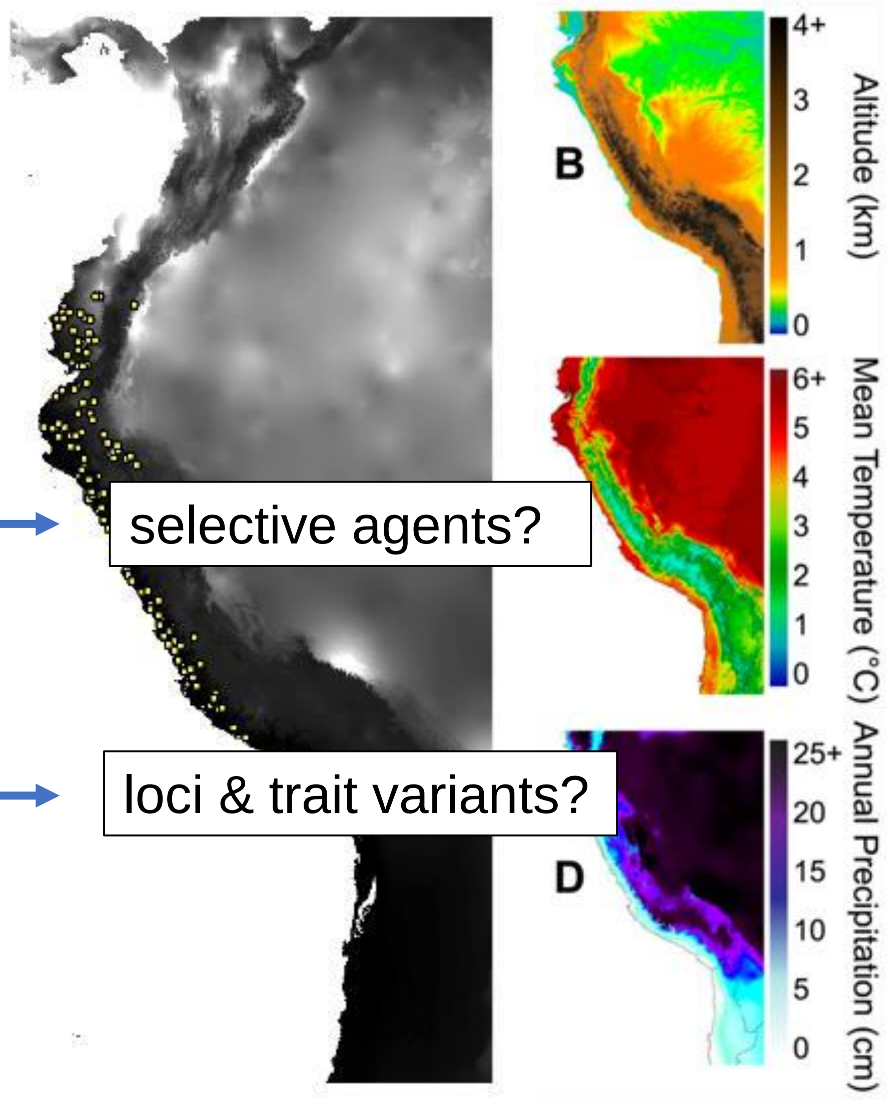
variable abiotic habitats

Goals

1. Estimate the independent contributions of climate and space to explaining genome-wide diversity

2. Infer abiotic climate variables most predictive of gene-environment associations

3. Identify genetic variants most strongly associated with major axes of multivariate climate



selective agents?

loci & trait variants?

Datasets

140 georeferenced accessions of
S. pimpinellifolium (TGRC; Davis, CA)

Geographic/spatial data

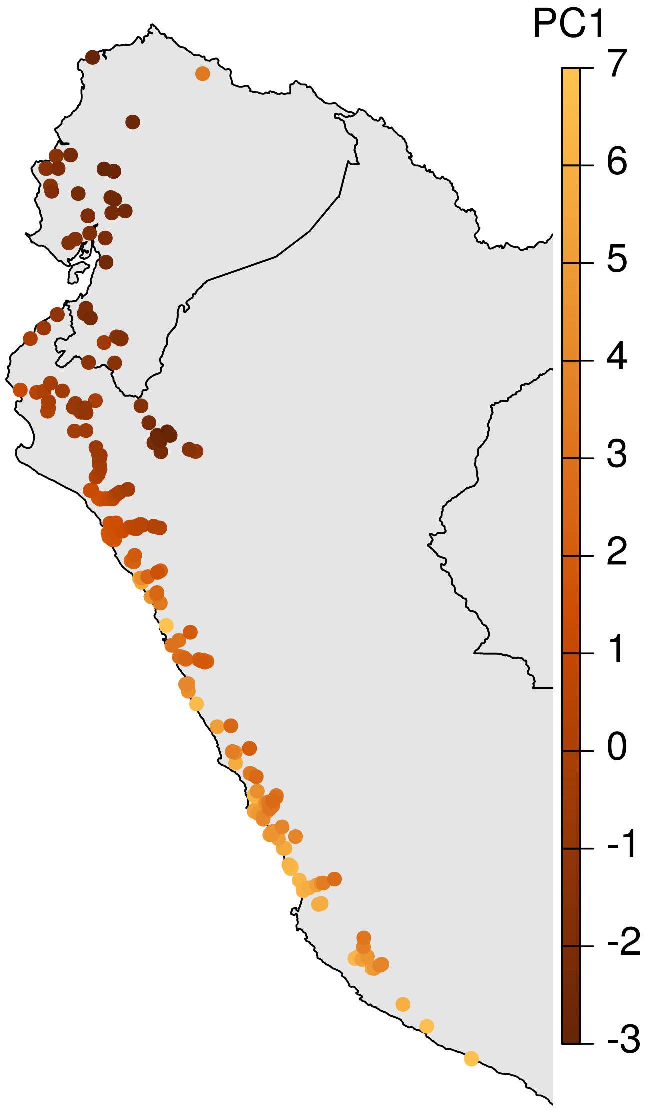
lat/long of collection locations

Environment/climate data

29 (of 54) non-redundant abiotic variables
at each location
(*WorldClim*, *CGIAR*, *ClimateSA*, and *SoilGrids*)

PCA on centered, scaled data
(multivariate climate variation)

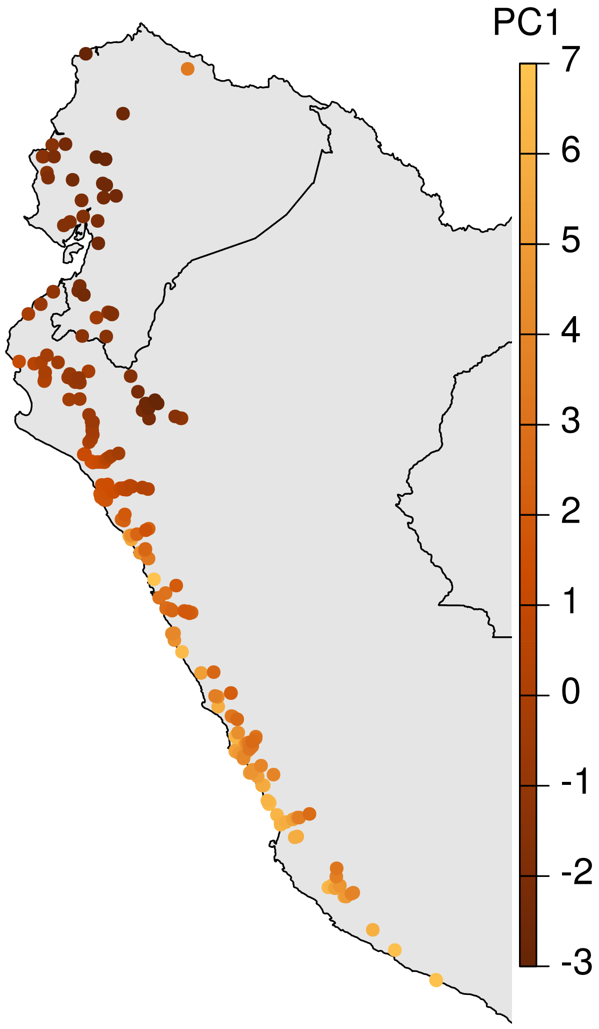
Genetic data



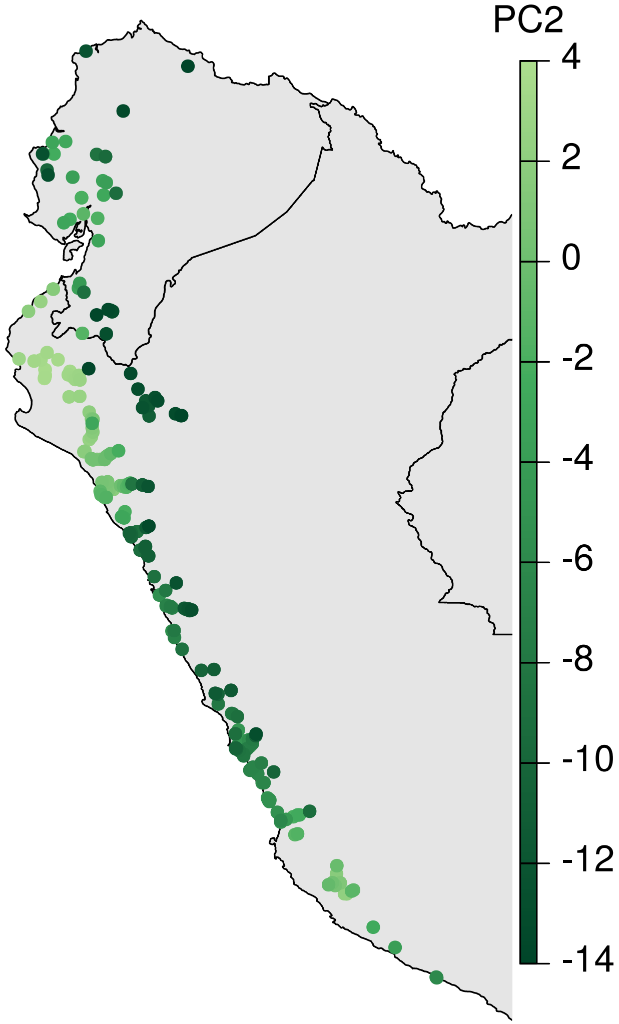
environmental variation follows spatial clines

PCA: 29 bioclimatic variables for accession locations

first 2 axes ~70% variance



(EnvPC1: 48.0%)



(EnvPC2: 22.7%)

Datasets

Geographic/spatial data

lat/long of collection locations

Environment/climate data

29 (of 54) non-redundant abiotic variables
at each location
(*WorldClim*, *CGIAR*, *ClimateSA*, and *SoilGrids*)

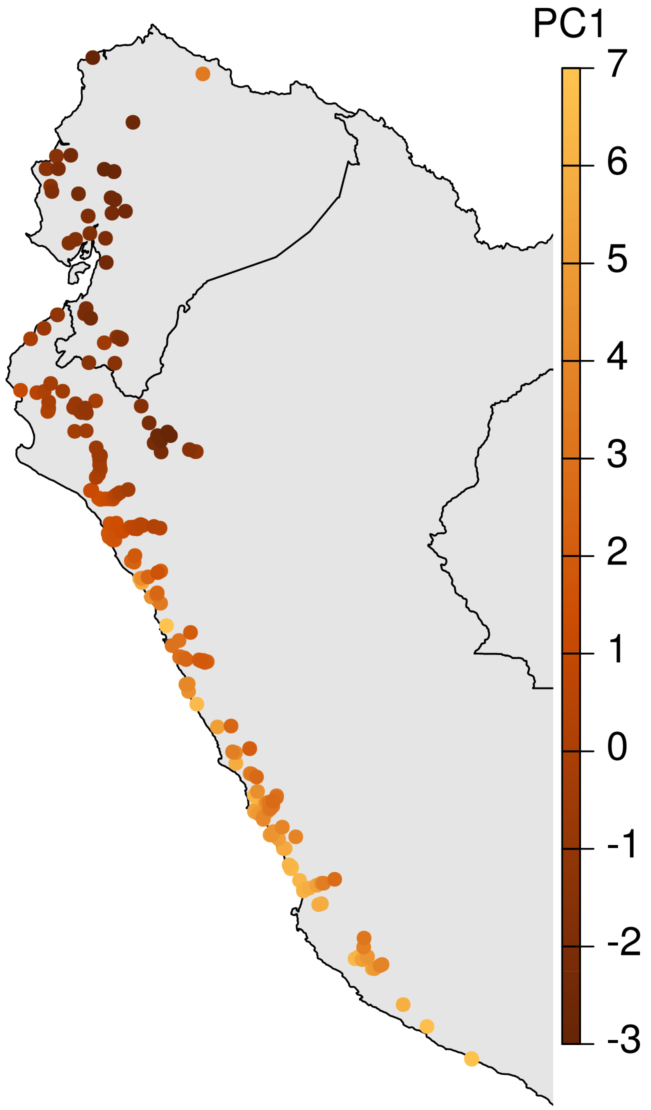
PCA on centered, scaled data
(multivariate climate variation)

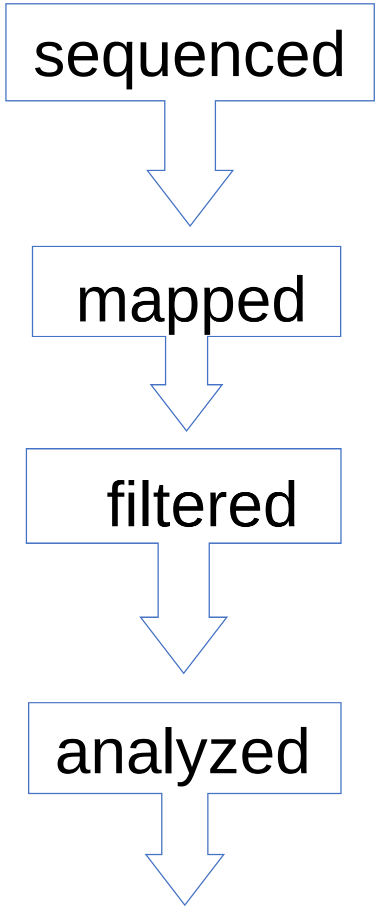
Genetic data

ddRAD (*PstI* & *EcoRI)* and Stacks
ref_map genotyping pipeline

model-based (*fastStructure*) and
non-model based (*PCA*) methods

140 georeferenced accessions of
S. pimpinellifolium (TGRC; Davis, CA)





sequenced

- tagged >450,000 loci
- average coverage: 66x (s.d. 36.7x)

mapped

- reference based (tomato genome, ITAG 3.2)
- ~360,000 SNPs (single nucleotide polymorphisms)

filtered

- low missing, high depth SNPs: 44,064
- LD-filtered SNPs: 17,358

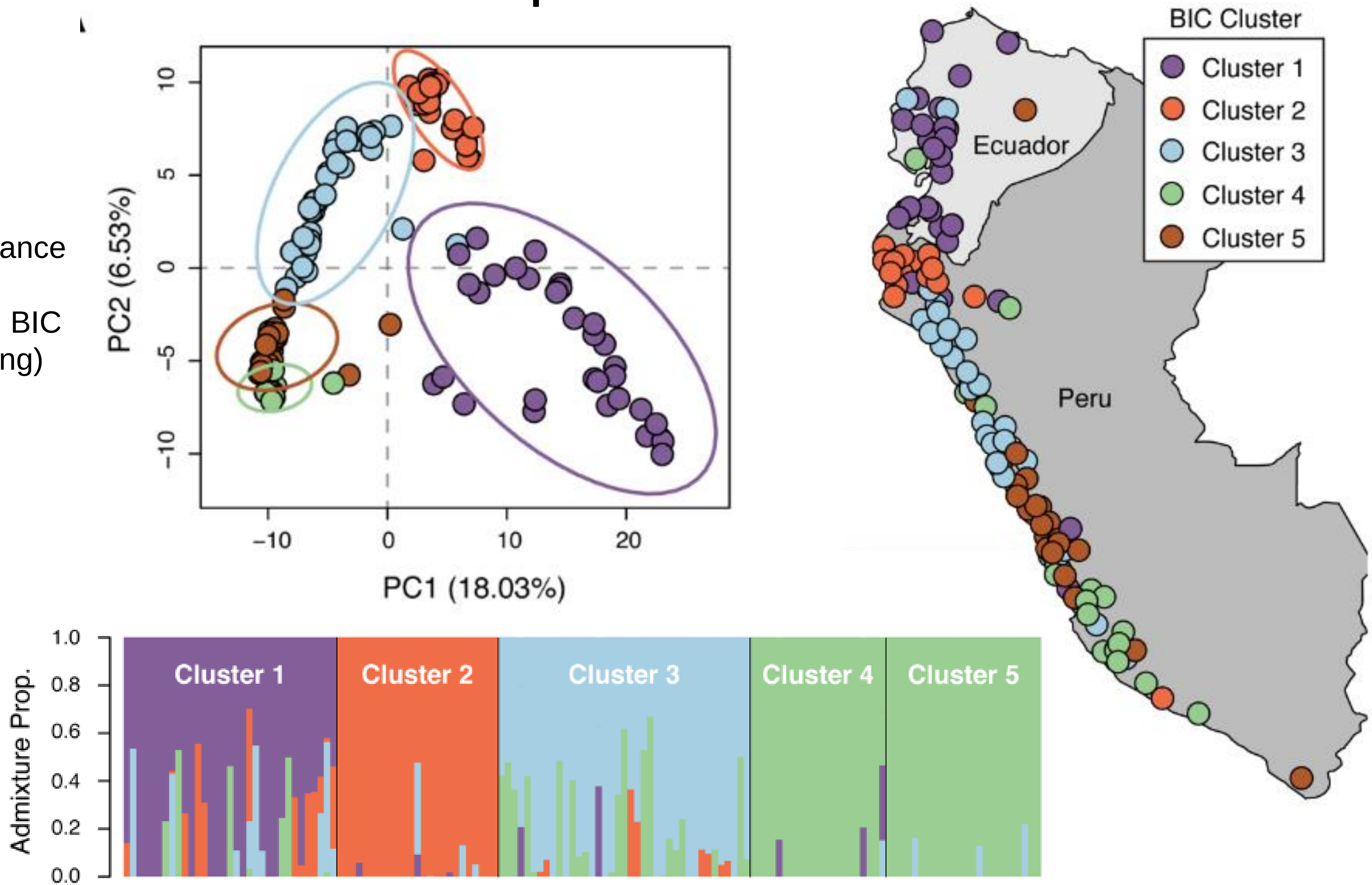
analyzed

- genomic distribution
- predicted variant categories

Genetic structure also follows spatial clines

Multilocus PCA:
first 2 axes ~22% variance

5 clusters (minimizing BIC with K-means clustering)

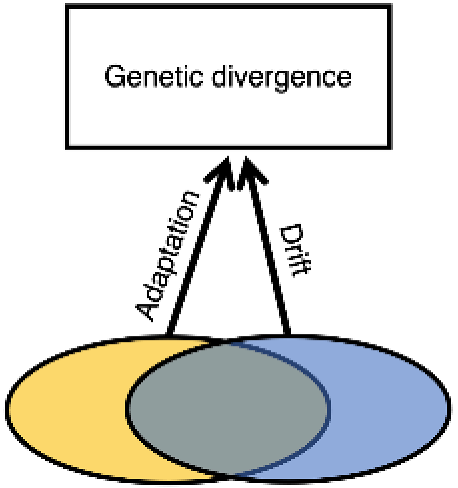


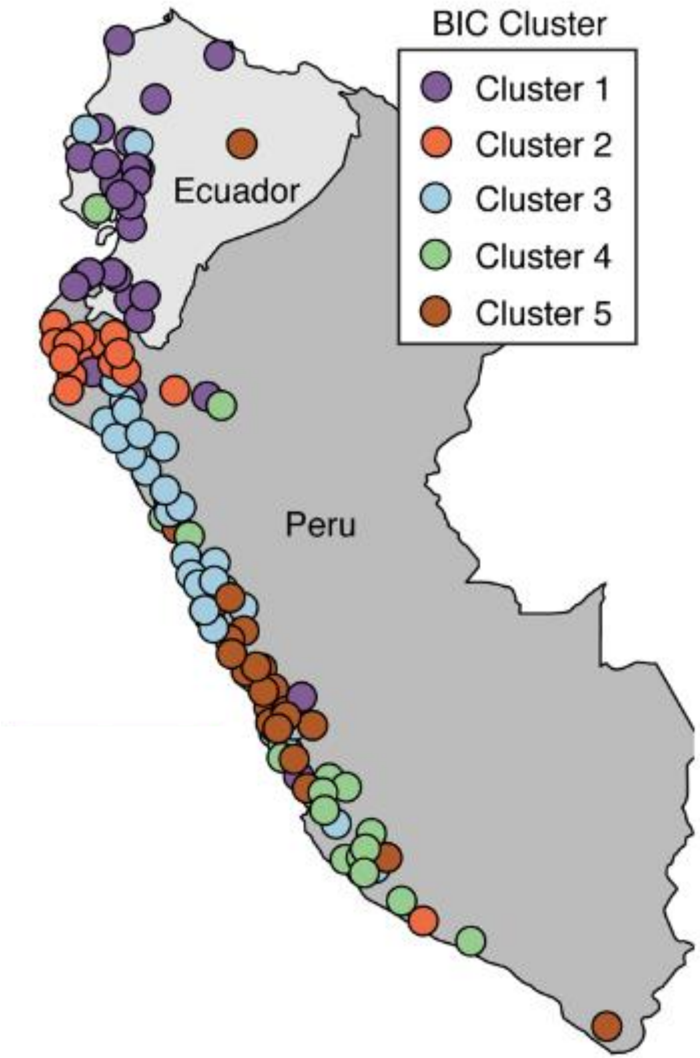
fastStructure:
4 populations
(maximizing marginal likelihood over K)

(*fastStructure*, Raj et al., 2014)

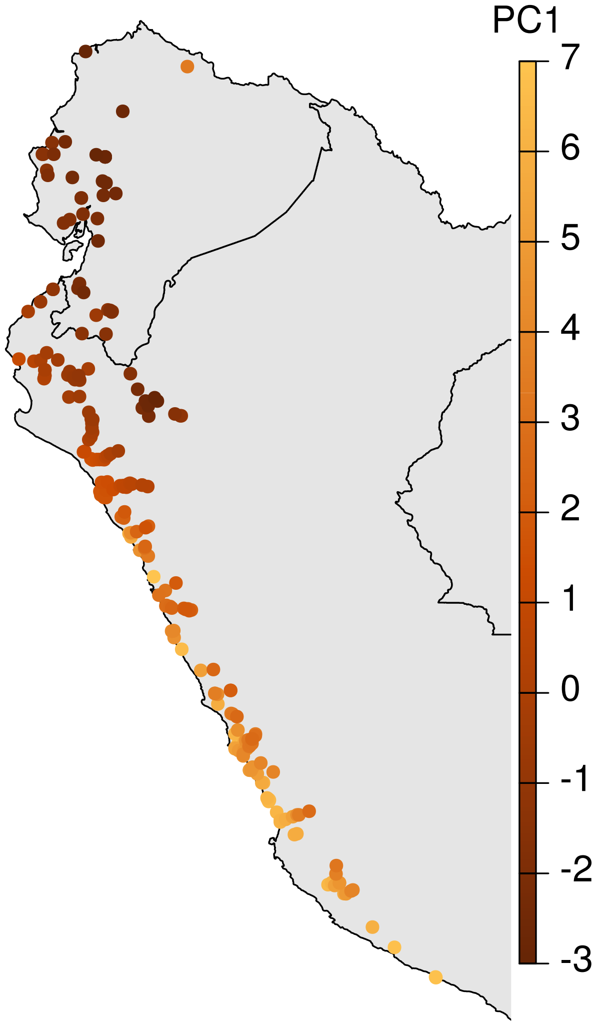
!!Collinearity!!

latitude is a very strong driver in this species





Genetic structure



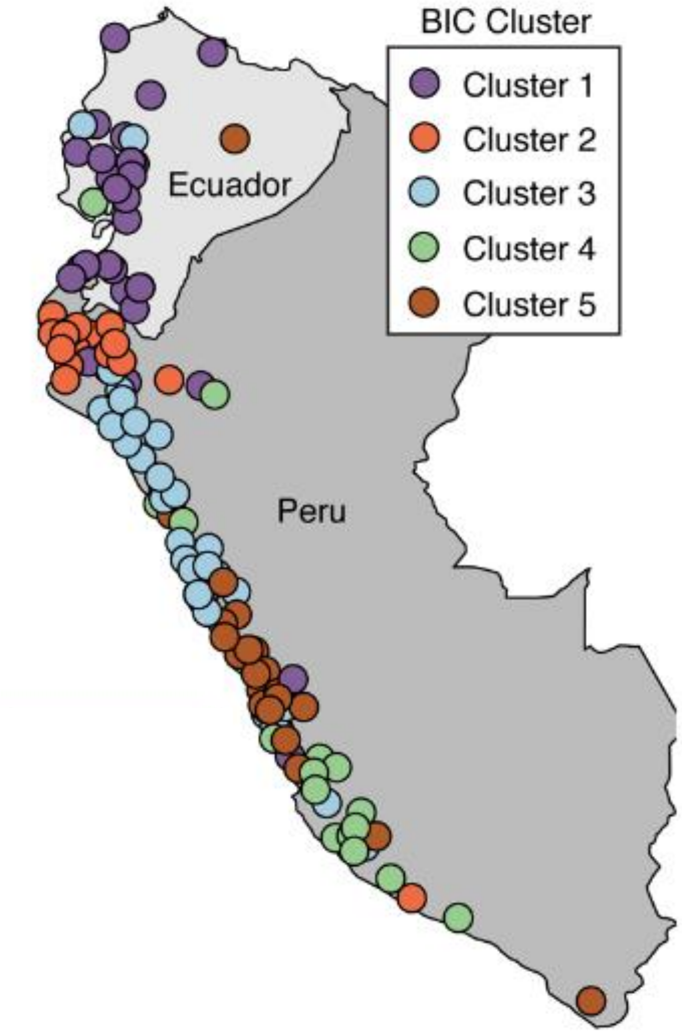
Environmental structure

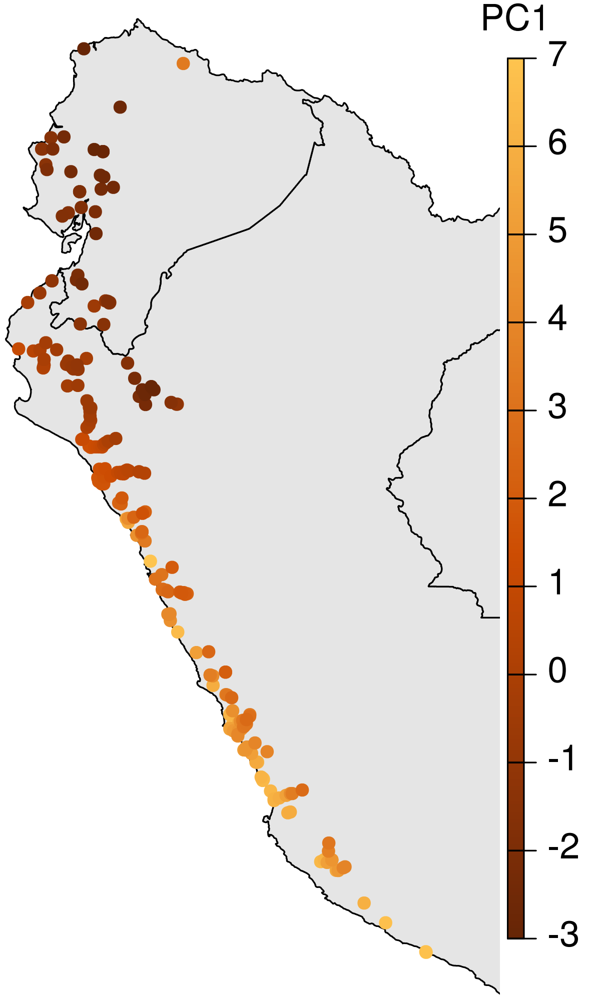
independent contributions of climate vs space (historical structure) to genetic variation

Variance partitioning by Redundancy Analysis (RDA) (*vegan*; Oksansen, 2018)

Structural equation modeling (SEM) (*lavaan*; Rosseel, 2012)

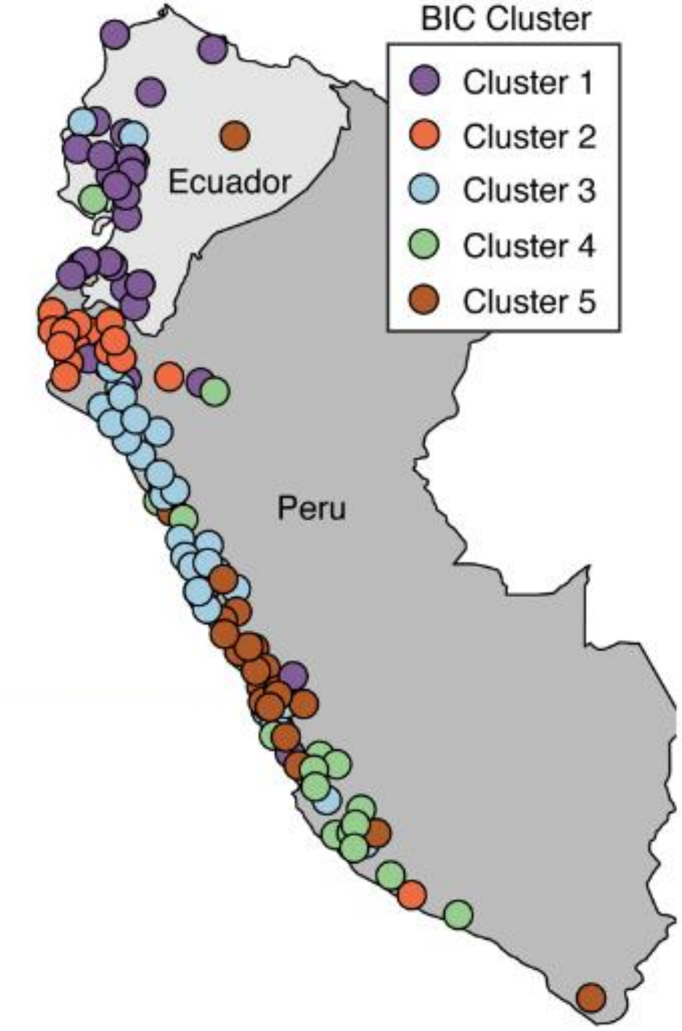
Generalized dissimilarity modeling (GDM) (*lgdm*; Manion, 2018)

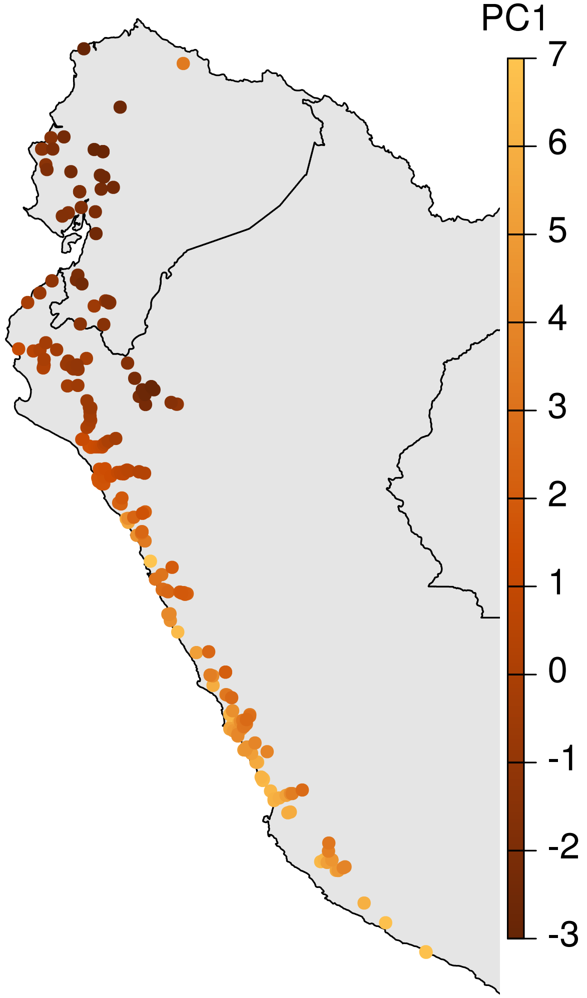




independent contributions of climate vs space (historical structure) to genetic variation

Variance partitioning by Redundancy Analysis (RDA)
(*vegan*; Oksansen, 2018)





Multiple linear regression: multiple response variables on multiple explanatory variables

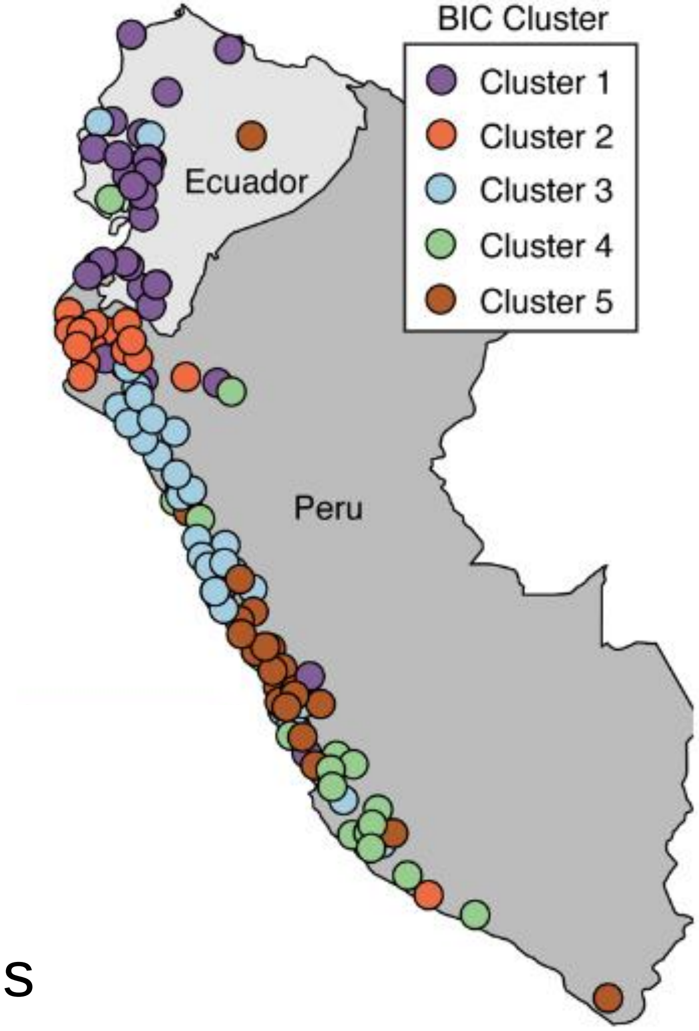
independent contributions of climate vs space (historical structure) to genetic variation

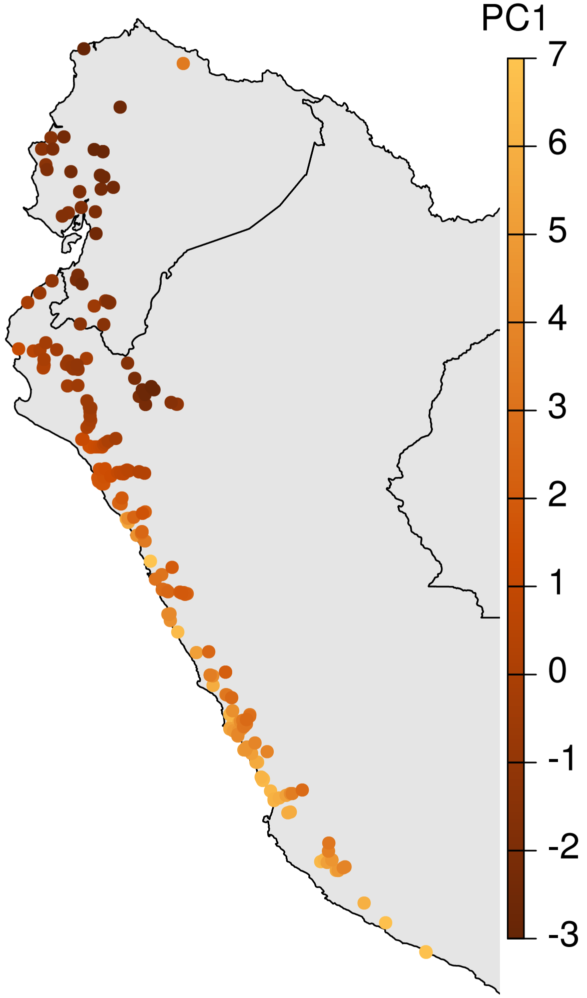
Variance partitioning by Redundancy Analysis (RDA)
(*vegan*; Oksansen, 2018)

SPACE:
truncated ordination matrix
(transformed euclidean distances)

ENVIRONMENT:
matrix of multivariate
environmental differences

GENETICS
matrix of multivariate SNP genotypes

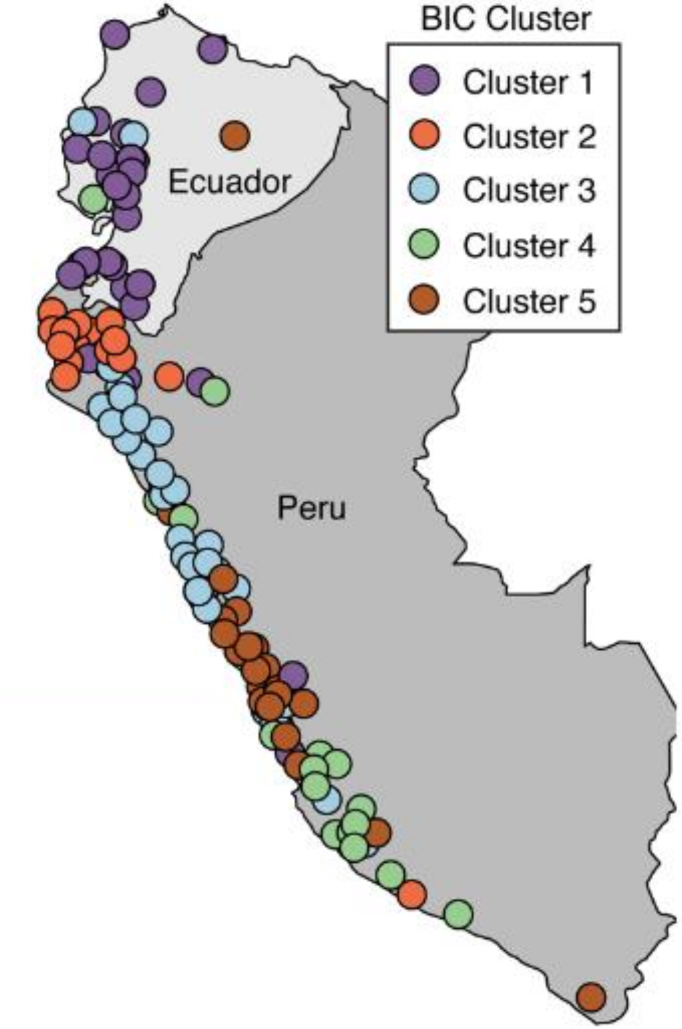


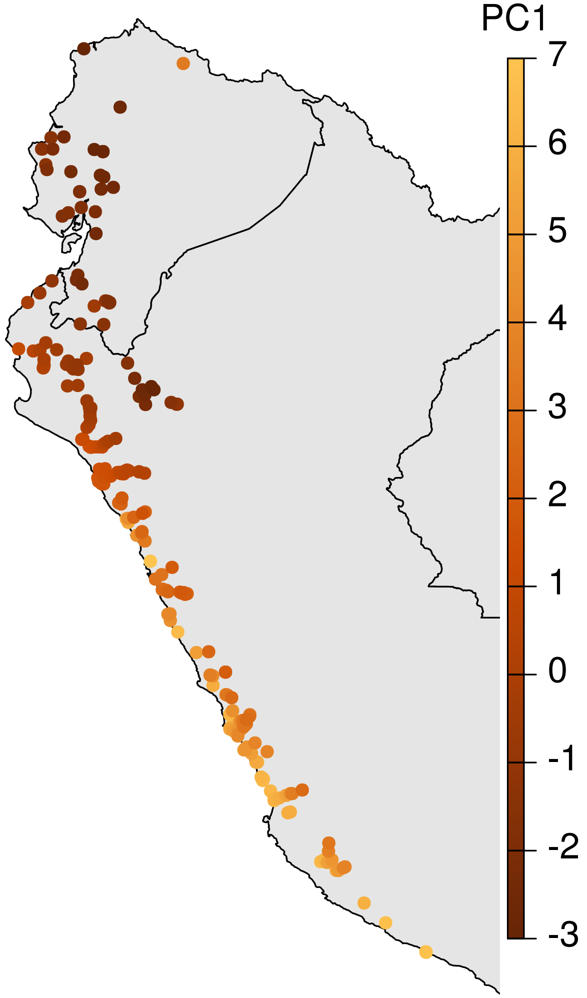


independent contributions of climate vs space (historical structure) to genetic variation

Variance partitioning by Redundancy Analysis (RDA)
(*vegan*; Oksansen, 2018)

what is the explanatory power of multivariate predictors (enviro & spatial variables) for multivariate responses (SNP genotypes)?





independent contributions of climate vs space (historical structure) to genetic variation

Variance partitioning by Redundancy Analysis (RDA)
(*vegan*; Oksansen, 2018)

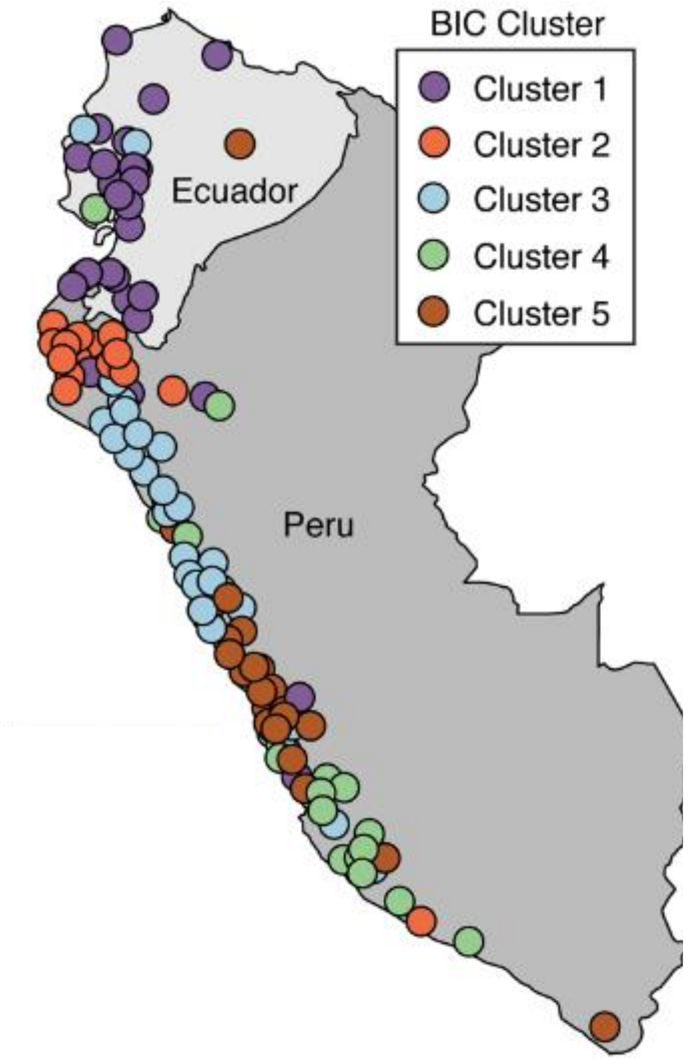
% SNP variance explained	
22.0	Total
17.0	Climate+Space
2.0	Space only
3.0	Climate only

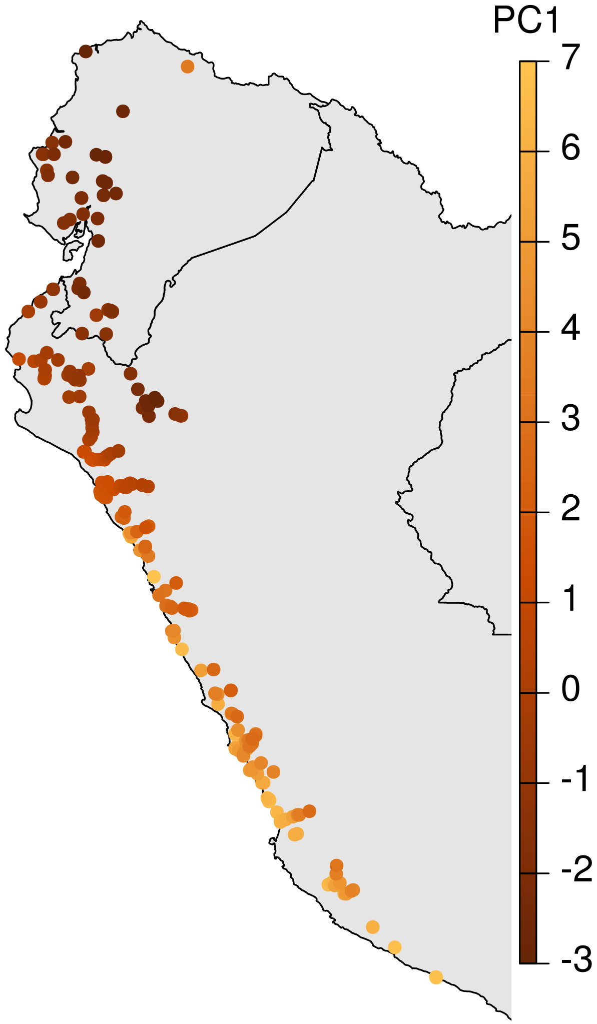
P < 0.001 for all proportions

colinear with both spatial and environmental variation

genetic variation explained by environment alone

what is the explanatory power of multivariate predictors (enviro & spatial variables) for multivariate responses (SNP genotypes)?





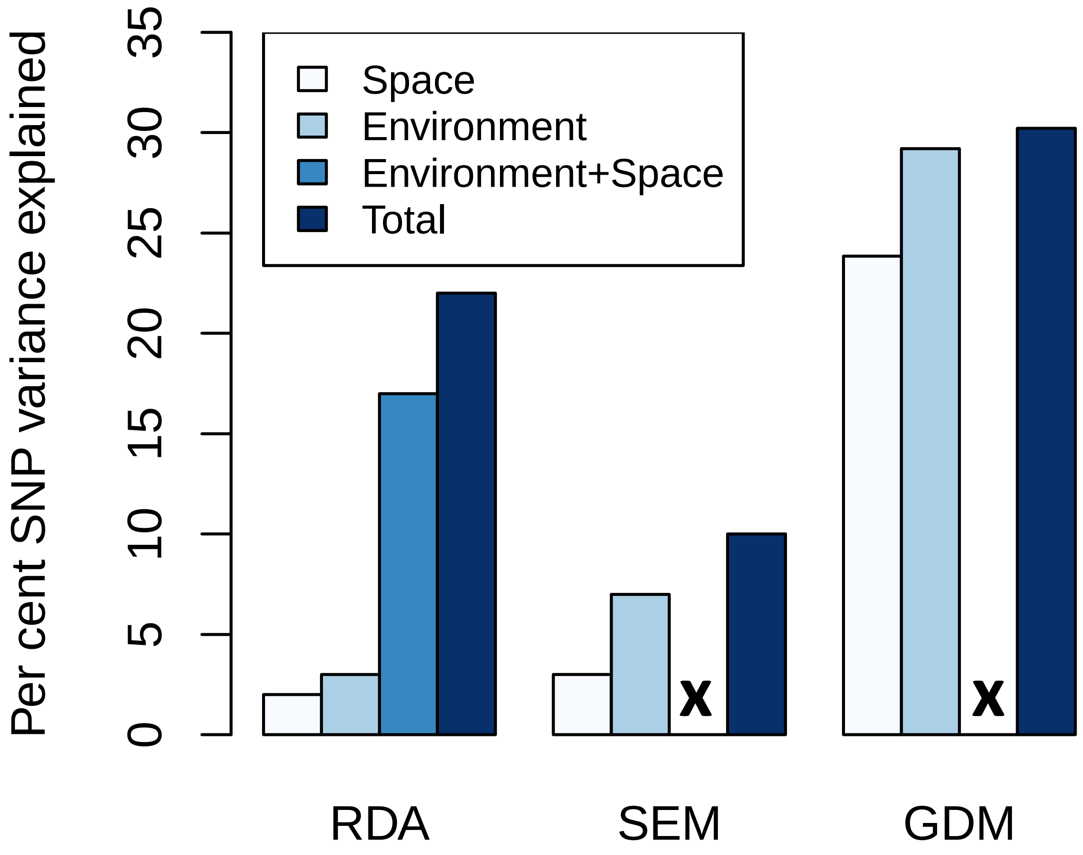
both are correlated with latitude!

independent contributions of climate vs space (historical structure) to genetic variation

Variance partitioning by
Redundancy Analysis (RDA)
(*vegan*; Oksansen, 2018)

Structural equation modeling (SEM)
(*lavaan*; Rosseel, 2012)

Generalized dissimilarity modeling (GDM)
(*lgdm*; Manion, 2018)



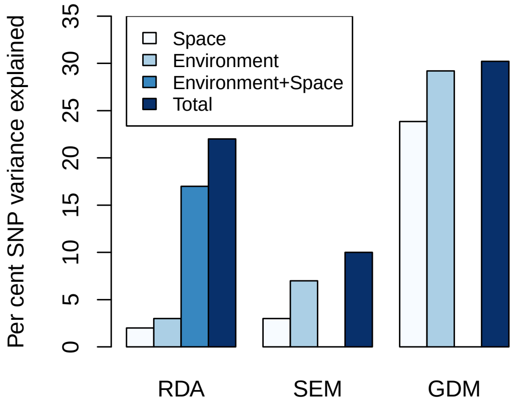
P < 0.001 for all proportions

Goals

1. Estimate the independent contributions of climate and space to explaining genome-wide diversity

2. Infer abiotic climate variables most predictive of gene-environment associations

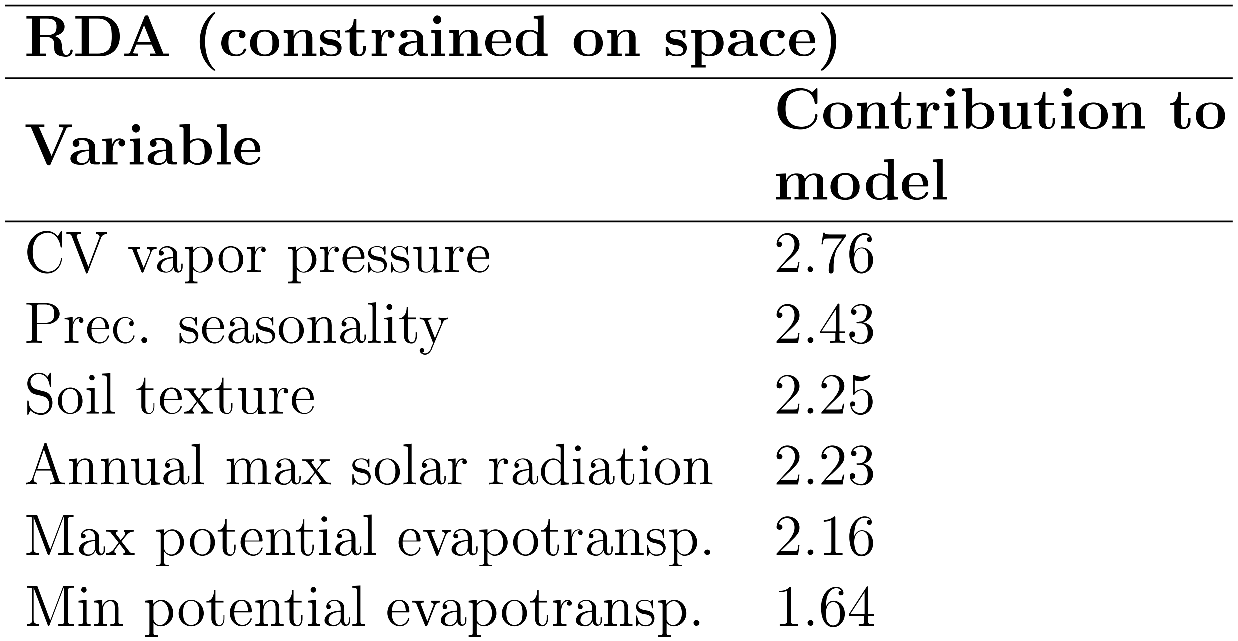
3. Identify genetic variants most strongly associated with major axes of multivariate climate



The abiotic environment explains more SNP variation than spatial structure

environmental variables most predictive of SNP variation

Variance partitioning by Redundancy Analysis (RDA)
(*vegan*; Oksansen, 2018)



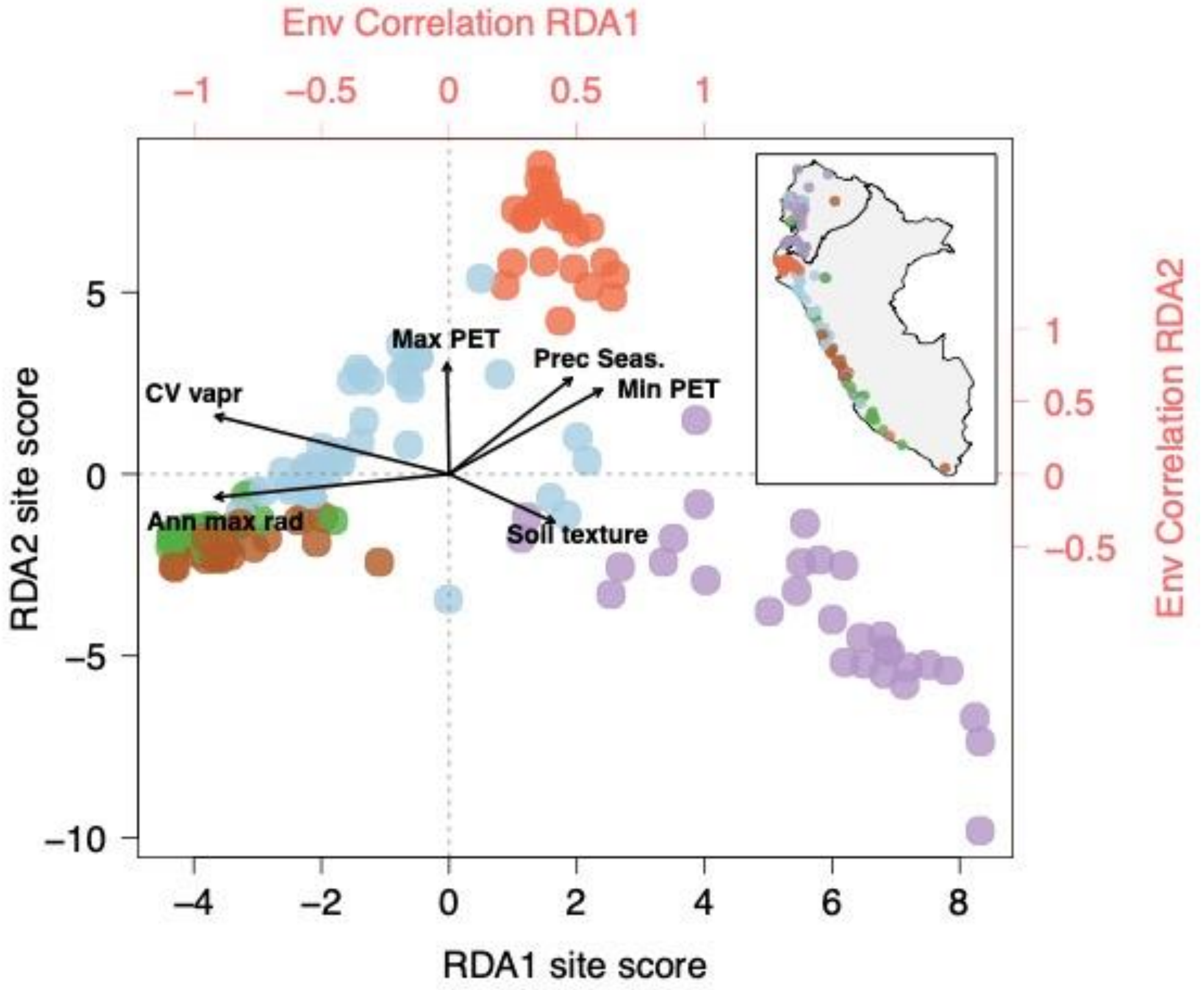
RDA (constrained on space)	
Variable	**Contribution to model**
CV vapor pressure	2.76
Prec. seasonality	2.43
Soil texture	2.25
Annual max solar radiation	2.23
Max potential evapotransp.	2.16
Min potential evapotransp.	1.64

evapotranspiration and seasonality variables are the strongest contributors

conditioned on spatial structure, what is the contribution of each environmental predictor to the RDA model?

environmental variables most predictive of SNP variation

especially **<u>variation</u> in vapor pressure** and **precipitation**



warning: conservative!!

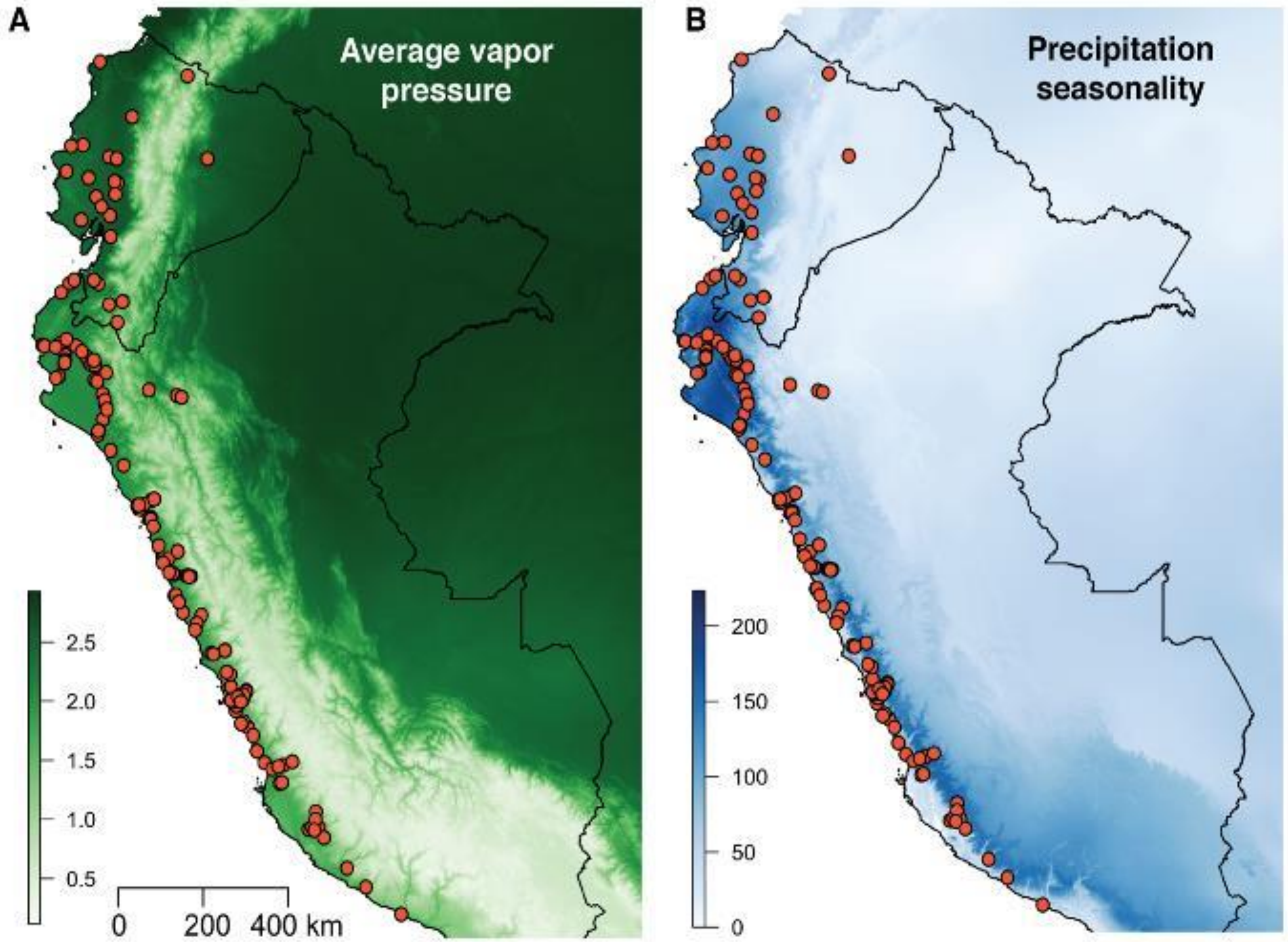
evapotranspiration and seasonality variables are the strongest contributors

environmental variables most predictive of SNP variation

RDA (constrained on space)	
Variable	**Contribution to model**
CV vapor pressure	2.76
Prec. seasonality	2.43
Soil texture	2.25
Annual max solar radiation	2.23
Max potential evapotransp.	2.16
Min potential evapotransp.	1.64

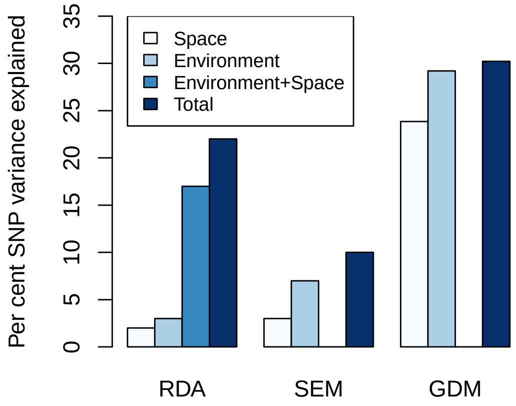
warning: conservative!!

evapotranspiration and seasonality variables are the strongest contributors



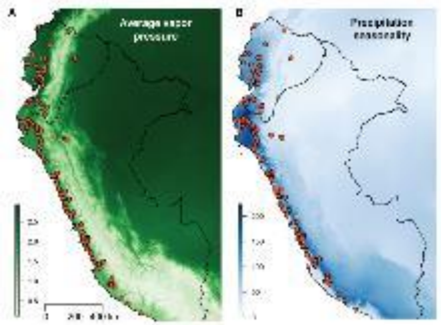
Goals

1. Estimate the independent contributions of climate and space to explaining genome-wide diversity



The abiotic environment explains more SNP variation than spatial structure

2. Infer abiotic climate variables most predictive of gene-environment associations

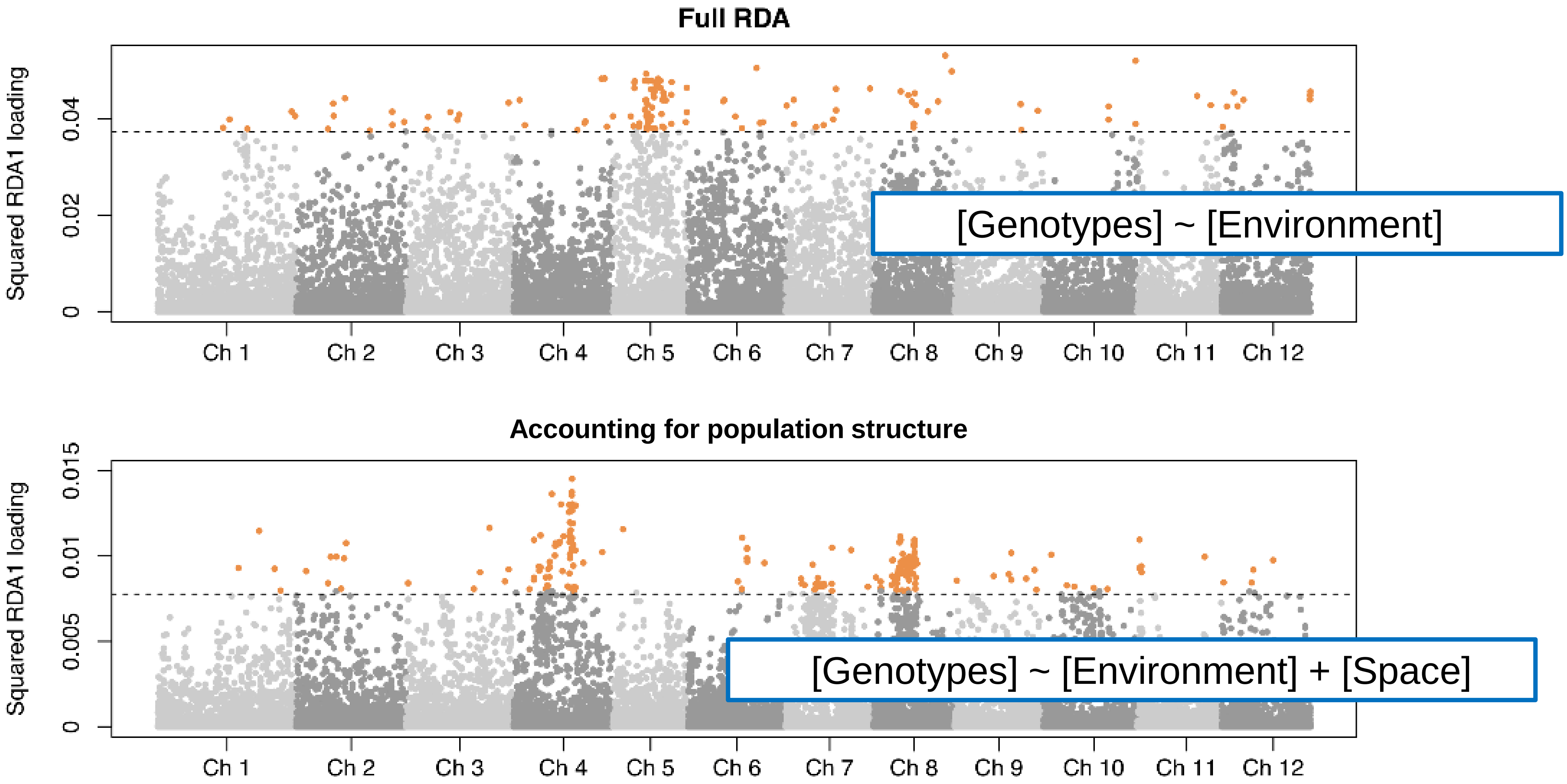


***evapotranspiration** and **seasonality** variables are the strongest contributors**

3. Identify genetic variants most strongly associated with major axes of multivariate climate

SNPs with strongest environmental associations

estimated by strength of loading on first RDA axis

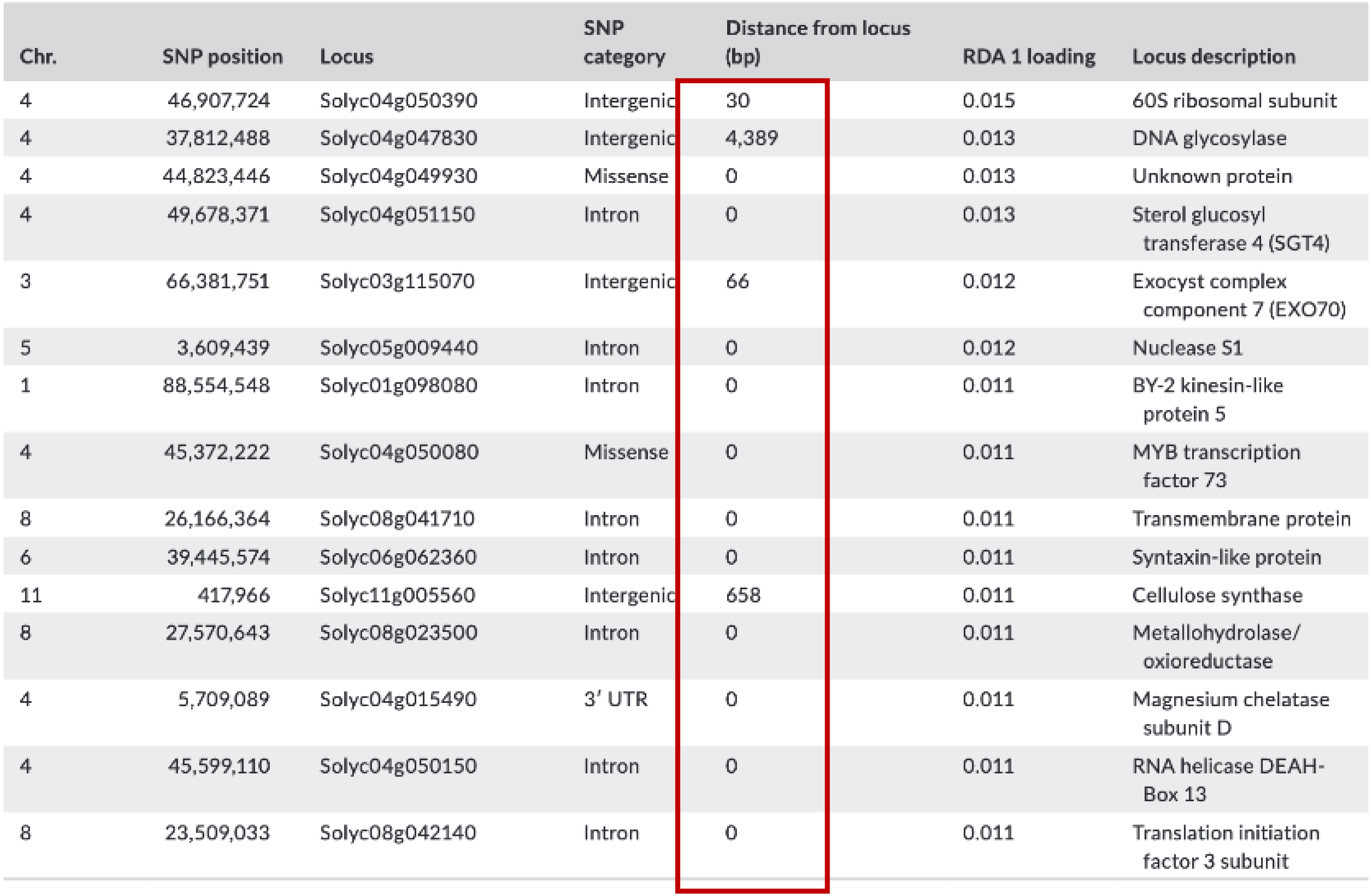


[Genotypes] ~ [Environment]

[Genotypes] ~ [Environment] + [Space]

SNPs with strongest environmental associations

Top 15 associations with RDA axis 1



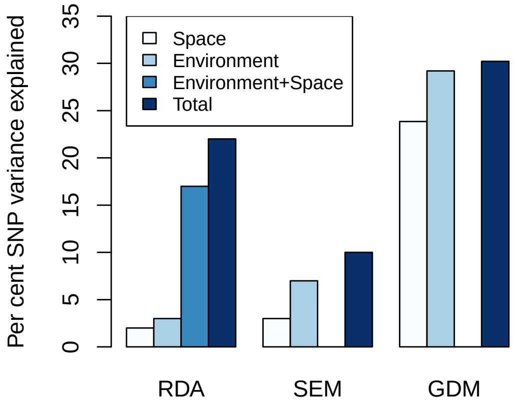
Chr.	SNP position	Locus	SNP category	Distance from locus (bp)	RDA 1 loading	Locus description	
4	46,907,724	Solyc04g050390	Intergenic	30	0.015	60S ribosomal subunit	
4	37,812,488	Solyc04g047830	Intergenic	4,389	0.013	DNA glycosylase	
4	44,823,446	Solyc04g049930	Missense	0	0.013	Unknown protein	
4	49,678,371	Solyc04g051150	Intron	0	0.013	Sterol glucosyl transferase 4 (SGT4)	**
3	66,381,751	Solyc03g115070	Intergenic	66	0.012	Exocyst complex component 7 (EXO70)	
5	3,609,439	Solyc05g009440	Intron	0	0.012	Nuclease S1	
1	88,554,548	Solyc01g098080	Intron	0	0.011	BY-2 kinesin-like protein 5	
4	45,372,222	Solyc04g050080	Missense	0	0.011	MYB transcription factor 73	**
8	26,166,364	Solyc08g041710	Intron	0	0.011	Transmembrane protein	
6	39,445,574	Solyc06g062360	Intron	0	0.011	Syntaxin-like protein	
11	417,966	Solyc11g005560	Intergenic	658	0.011	Cellulose synthase	
8	27,570,643	Solyc08g023500	Intron	0	0.011	Metallohydrolase/ oxioreductase	
4	5,709,089	Solyc04g015490	3′ UTR	0	0.011	Magnesium chelatase subunit D	
4	45,599,110	Solyc04g050150	Intron	0	0.011	RNA helicase DEAH-Box 13	
8	23,509,033	Solyc08g042140	Intron	0	0.011	Translation initiation factor 3 subunit	

in or near known genes

environmental response functions

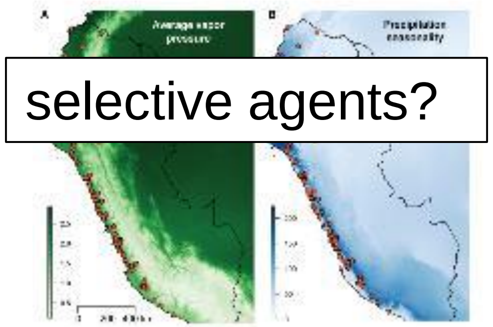
Goals

1. Estimate the independent contributions of climate and space to explaining genome-wide diversity



The abiotic environment explains more SNP variation than spatial structure

2. Infer abiotic climate variables most predictive of gene-environment associations

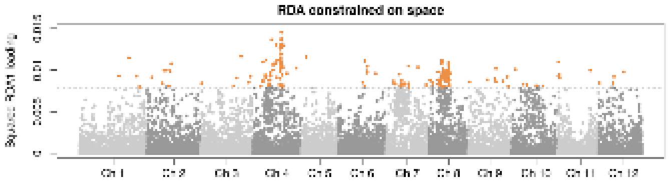


selective agents?

evapotranspiration *and* ***seasonality*** *variables are the strongest contributors*

3. Identify genetic variants most strongly associated with major axes of multivariate climate

loci & trait variants?



extreme SNPs are associated with genes relevant to climate adaptation

environmental association analyses (EAA)

use SNP-environmental associations to infer things like:

- specific genomic targets of environmental selection (loci)

- specific environmental components that impose selection (agents)

- contribution of spatially-varying (abiotic) selection to genome-wide genomic variation

- parallel versus unique responses to repeated environmental gradients

environmental assoc. analyses

limitations

- environmental variation can be confounded with historical/spatial population structure **producing spurious (non-causal) associations**

FALSE POSITIVES

- correcting for population structure can overcompensate

FALSE NEGATIVES

- collecting (high quality, relevant) environmental data can be challenging

- still several steps away from direct causal inference…

selection within and between populations

goal: identify loci	undergoing recent selection (with or w/out phenotype) incl.*divergent* across space	underlying important functional variation incl. across space
signature: variants/regions	that depart from neutral or null expectations	associated with segregating functional variation
approaches:	sequence-based tests of selection differentiation-based tests environmental association analyses	association studies

take-homes

all approaches have limitations
(being aware of these is imp!!)

most are still challenging
except in ‘developed’ systems

all are (at least) several steps
from direct causal inferences
about adaptation

References (in order of appearance)

Rellstab et al. 2015. Molecular Ecology. 24: 4348–4370

Hoekstra et al. 2004. Evolution. 58: 1329-1341

Hohenlohe et al. 2010. Int. J. Plant Sci. 171(9): 1059–1071.

Hohenlohe et al. 2018 pp 483-510, in Om P. Rajora (ed.), Population Genomics: Concepts, Approaches and Applications, Population Genomics. Springer International, 2018.

Coop 2022, Population and Quantitative Genetics (https://github.com/cooplab/popgen-notes/releases)

Nielson. 2005. Annual Reviews Genetics 39:197–218

Hahn 2018. Molecular Population Genetics. Sinauer.

Stinchcombe & Hoekstra. 2008. Heredity: 100, 158–170

Kruglyak. 2008. Nature Reviews Genetics. 9: 314–318.

Anderson, Willis, Mitchell-Olds. 2011. Trends in Genetics. 27(7): 258-266

Hohenlohe et al. 2010b. PLoS Genetics 6: e1000862.

Lotterhos & Whitlock 2015. Molecular Ecology. 24: 1031–1046

Gibson & Moyle. 2020. Molecular Ecology. 29:2204–2217.

Pease et al., 2016 PLoS Biology 14(2): e1002379. doi:10.1371/journal.pbio.1002379