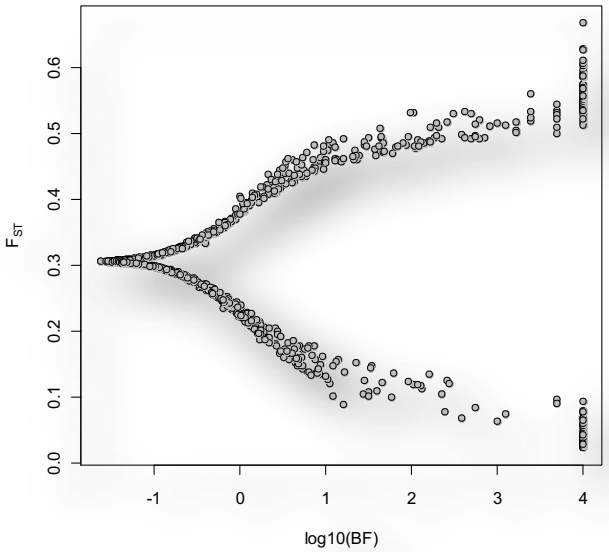
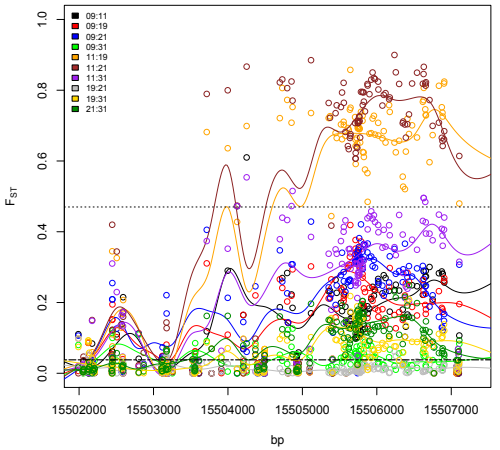
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



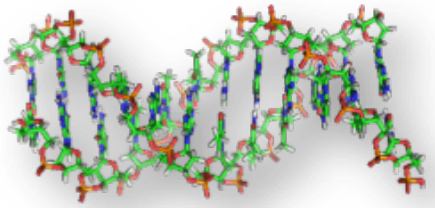


Detection of the genomic signature of selection

Martin C. Fischer
Institute of Integrative Biology
ETH Zürich

martin.fischer@env.ethz.ch

September 17th, 2018



Diversity and adaptation

Why is there so much variation among and within species?

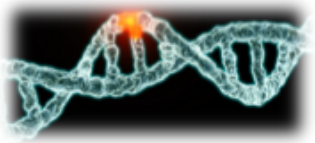
How do organisms adapted to their environment?

Genetic diversity

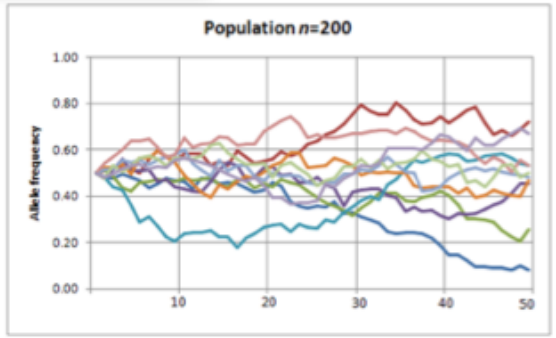
- Evolutionary changes at the molecular level are caused by…

Neutral processes

- Mutations



- Genetic drift



- Population history

Mutations

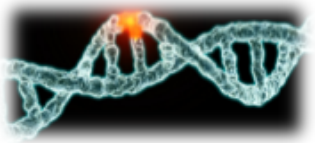
- Small scale:
  - Point mutations (SNPs)
  - Deletions
  - Insertions
- Large scale:
  - Copy number variation
  - Translocations
  - Inversions

Genetic diversity

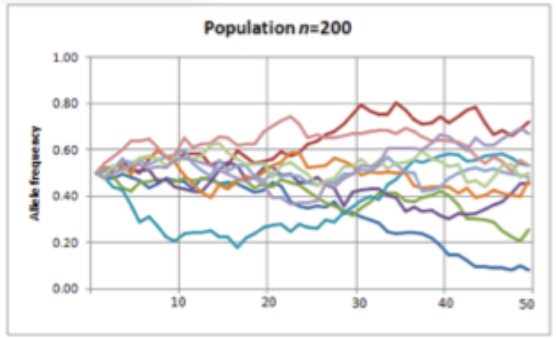
- Evolutionary changes at the molecular level are caused by…

Neutral processes

- Mutations



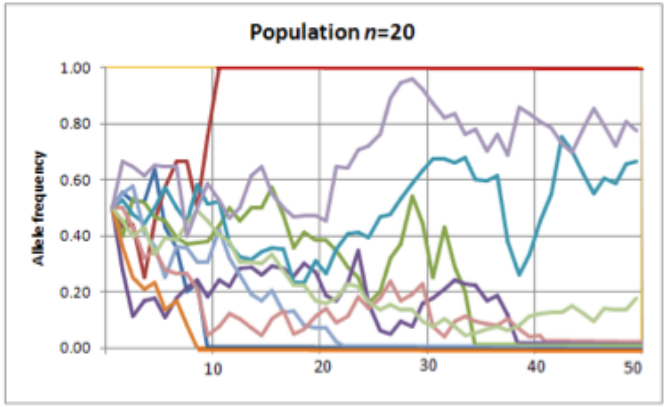
- Genetic drift



- Population history

Genetic drift

- Change of allele frequencies over generations in a population due to random sampling
- Population size:
  Drift is largest in small populations

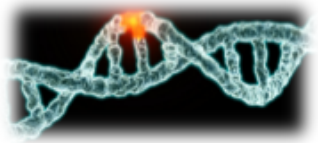


Genetic diversity

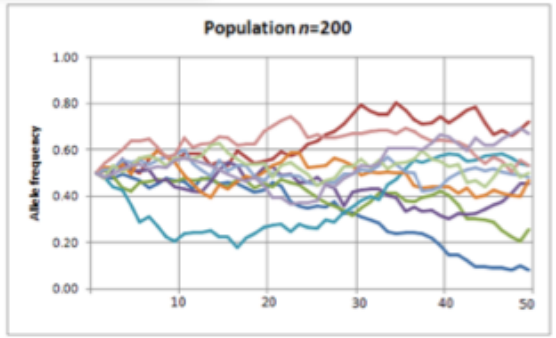
- Evolutionary changes at the molecular level are caused by…

Neutral processes

- Mutations

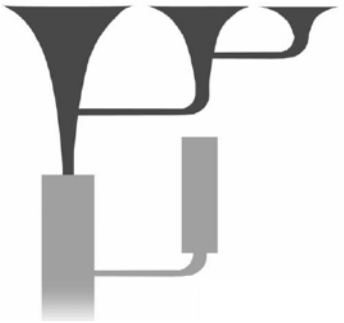


- Genetic drift

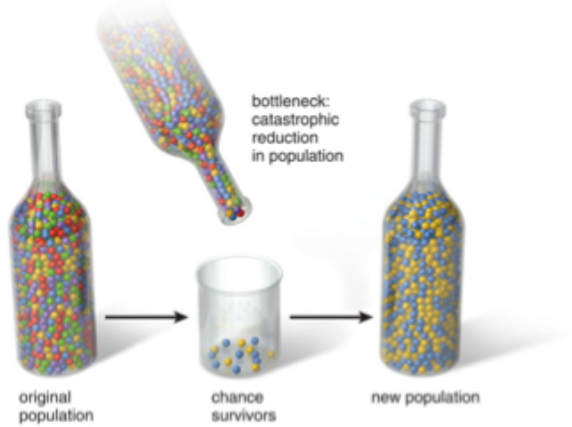


- Population history

Population history



- Demography
- Bottleneck



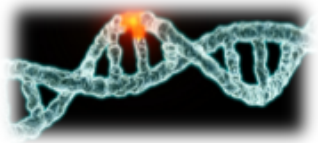
- Migration / gene-flow

Genetic diversity

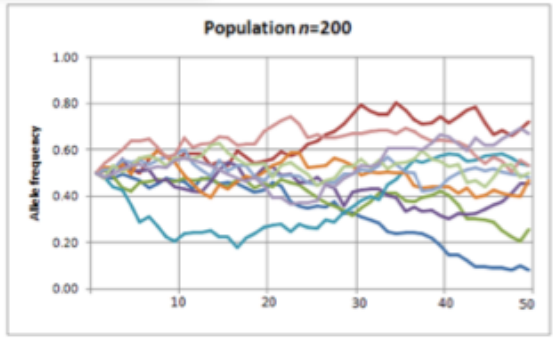
- Evolutionary changes at the molecular level are caused by…

Neutral processes

- Mutations

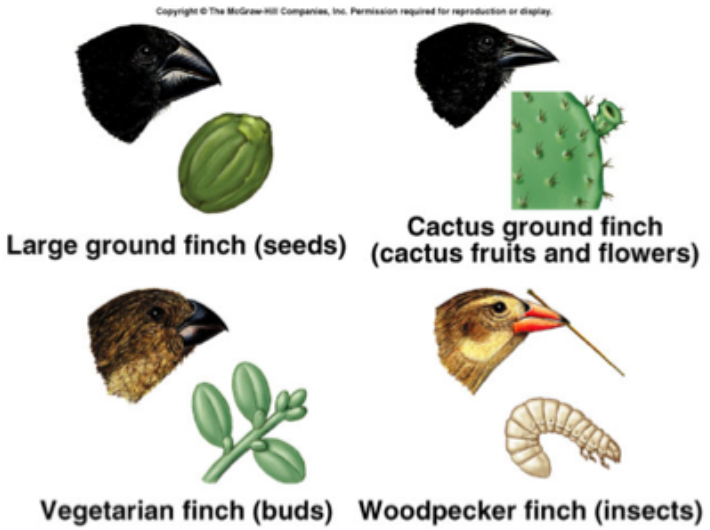


- Genetic drift

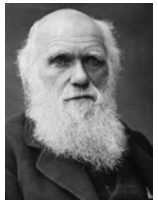


- Population history

Selection (natural, …)



Darwin‘s Finches



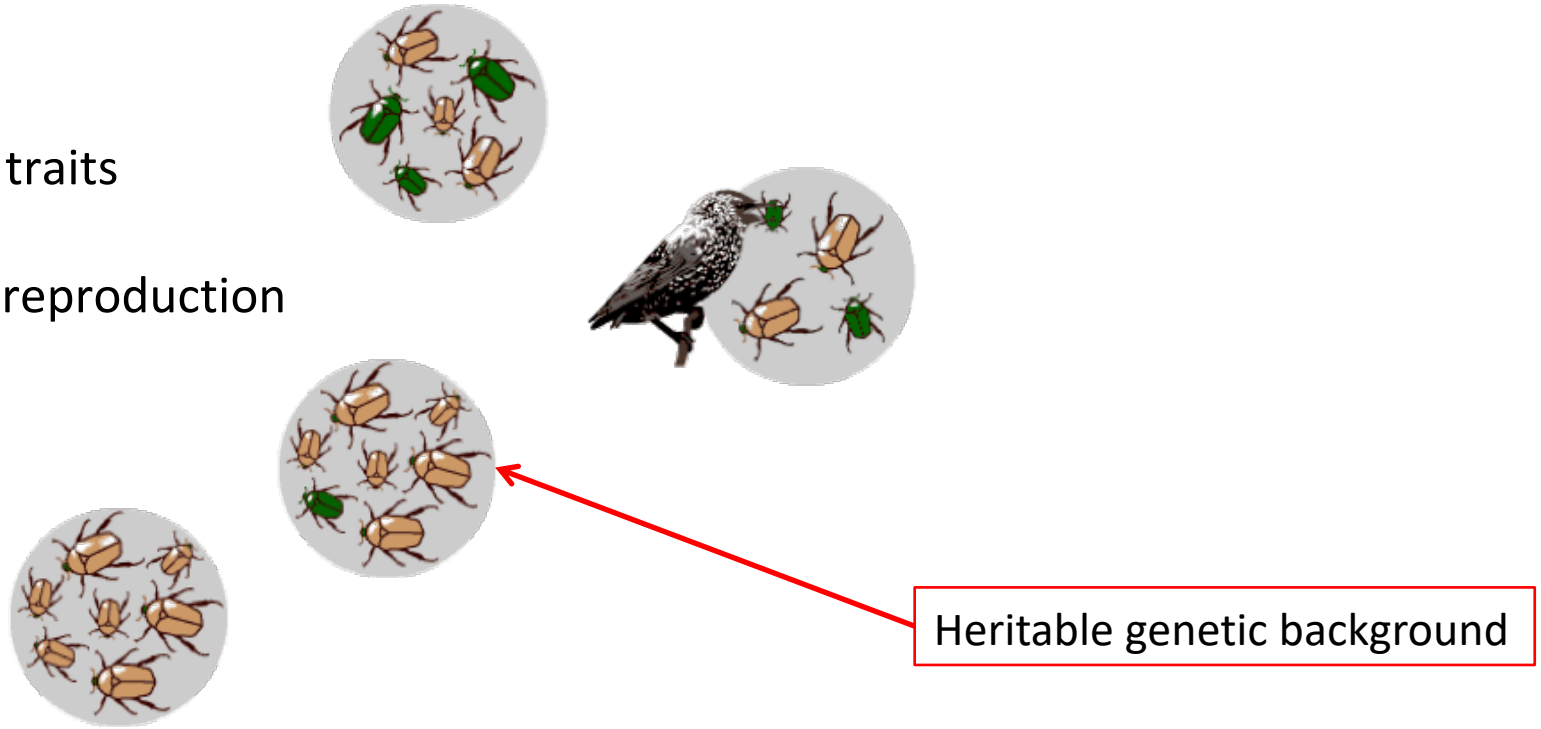
Lamichaney *et al.* 2015

Natural selection

Natural selection is one of the basic mechanisms of evolution

Population of beetles:

1. There is variation in traits
2. There is differential reproduction
3. There is heredity
4. End result



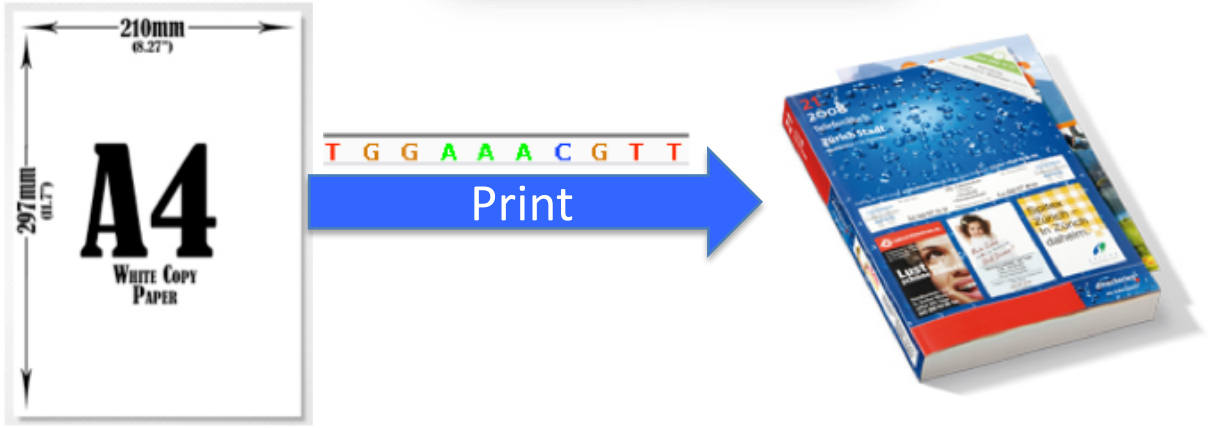
http://evolution.berkeley.edu/evolibrary/article/evo_25

The genomic signature of selection

Where do we fine the genomics signature of selection ?

- *Arabidopsis thaliana*: 0.15 Gb
- *Homo sapiens*: 3.2 Gb

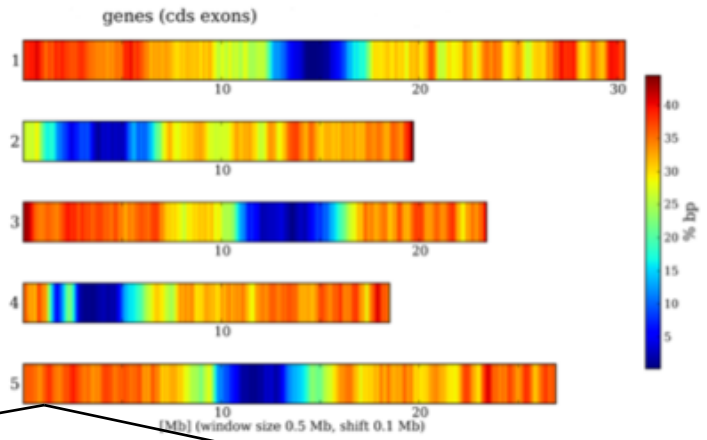
- A4 => Font 10



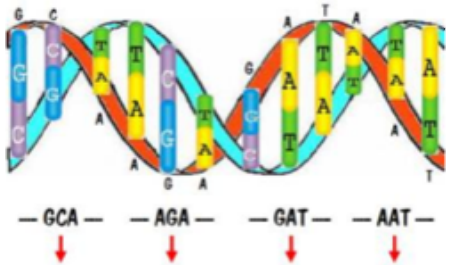
A. thaliana	~26,000 pages	~26 Zurich phone books
Homo sapiens	~532,000 pages	~532 Zurich phone books

The genome and where selection acts

	Human	*A. thaliana*
Genome	3.165 Gb	0.157 Gb
Chromosomes	22, XY	5
Chr length	50-250 Mb	20-34 Mb
Genes	~25,000	~28,000



Exon

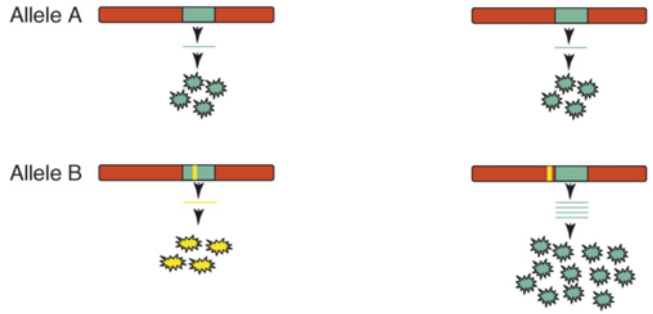


Intron

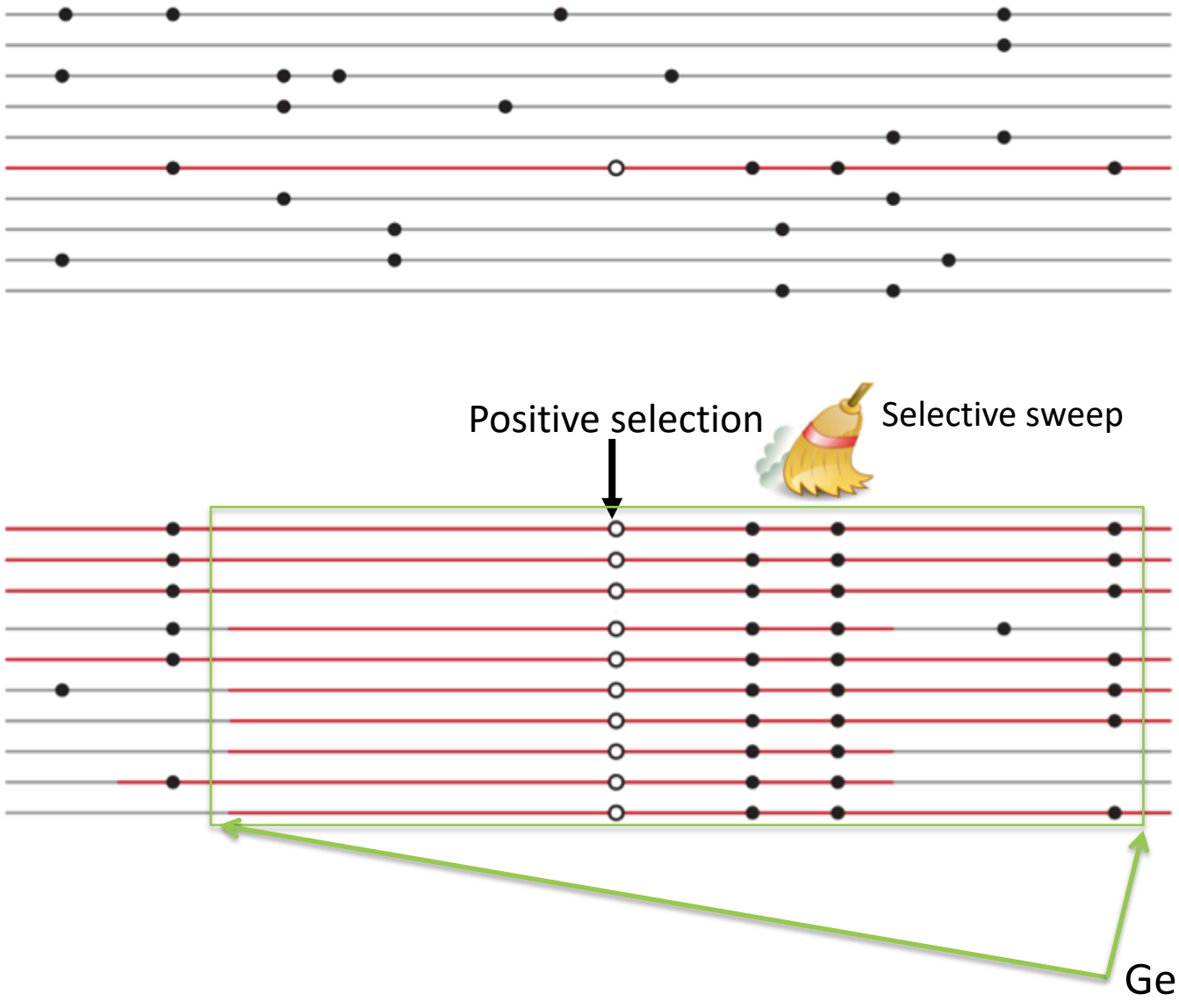
Intergenic regions

Functional mutation

Gene expression changes



Selective sweeps



Selective sweep:

- An allele that increases fitness arises and ‘sweeps’ to fixation in a population

- Recombination

- Linkage disequilibrium
- Long haplotypes

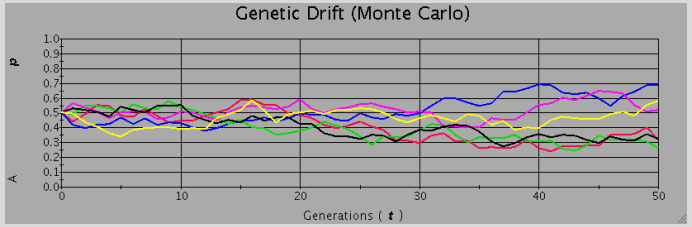
Figure 1 of Storz 2005, *Molecular Ecology*

Locus under selection behaves differently

Neutral processes:

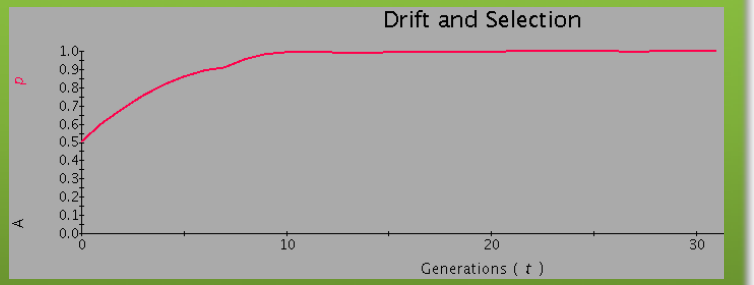
- Genetic drift
- Demographic history

➢ Affect all loci similarly



Selection:

➢ Affect only single locus



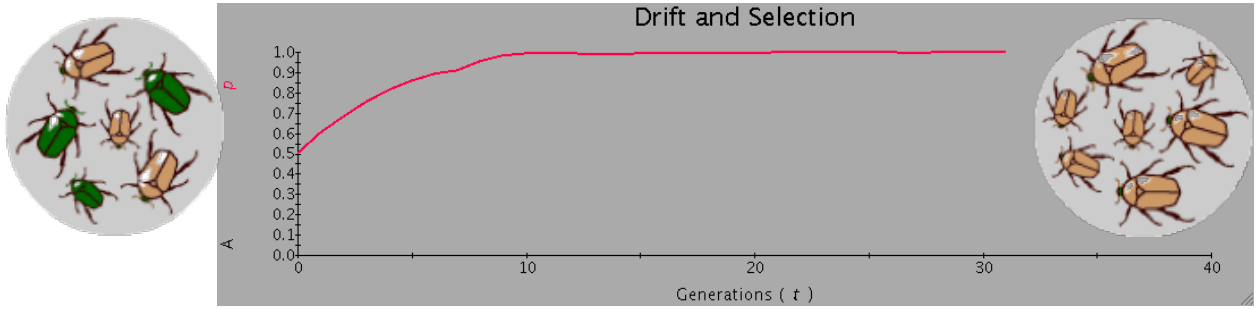
➢ Outlier detection

Outlier compared to the **rest of the genome**

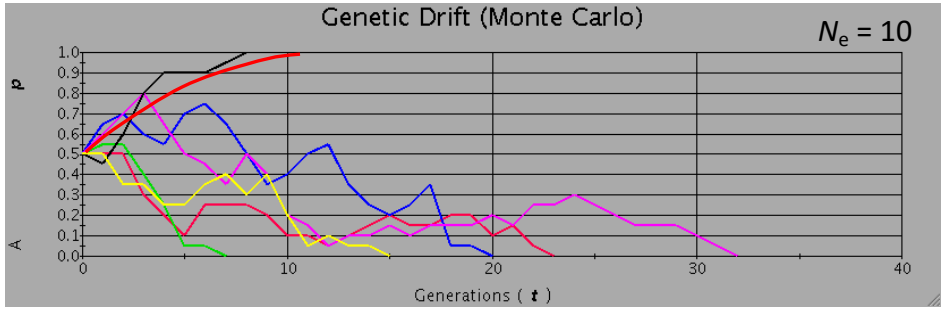
for F_{ST} see: Lewontin & Krakauer 1973; Beaumont & Nichols 1996

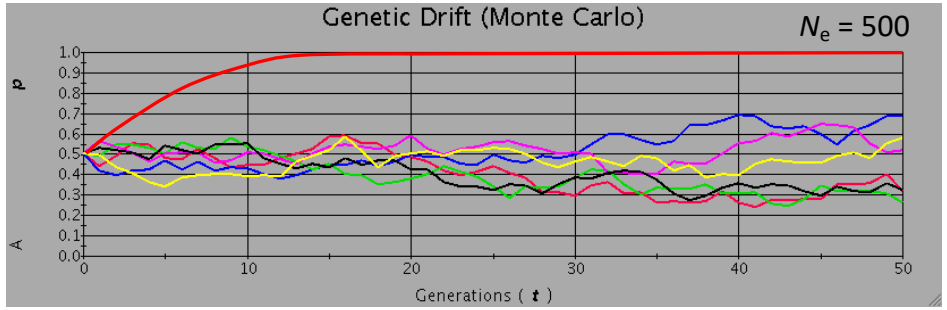
Population size matters….

- Selection



- Drift



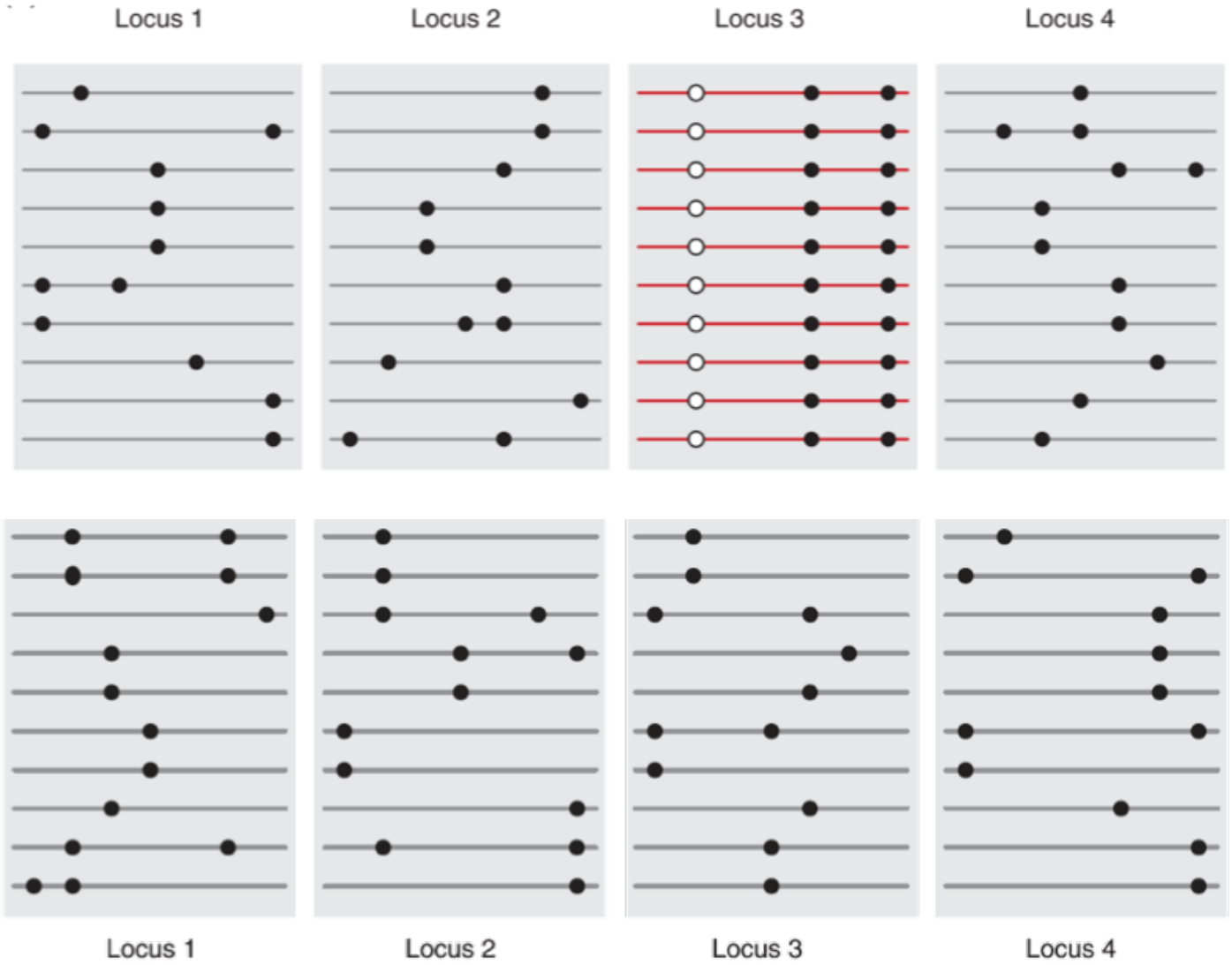


- Selection coefficient $s >> (1/2N_e)$

- Population under selection needs to have a minimum N_e to overcome drift, or s needs to be very strong

- Selection acts most efficient on large populations

Selective sweeps



Single locus is affected

- Recombination

Skewed allele frequency

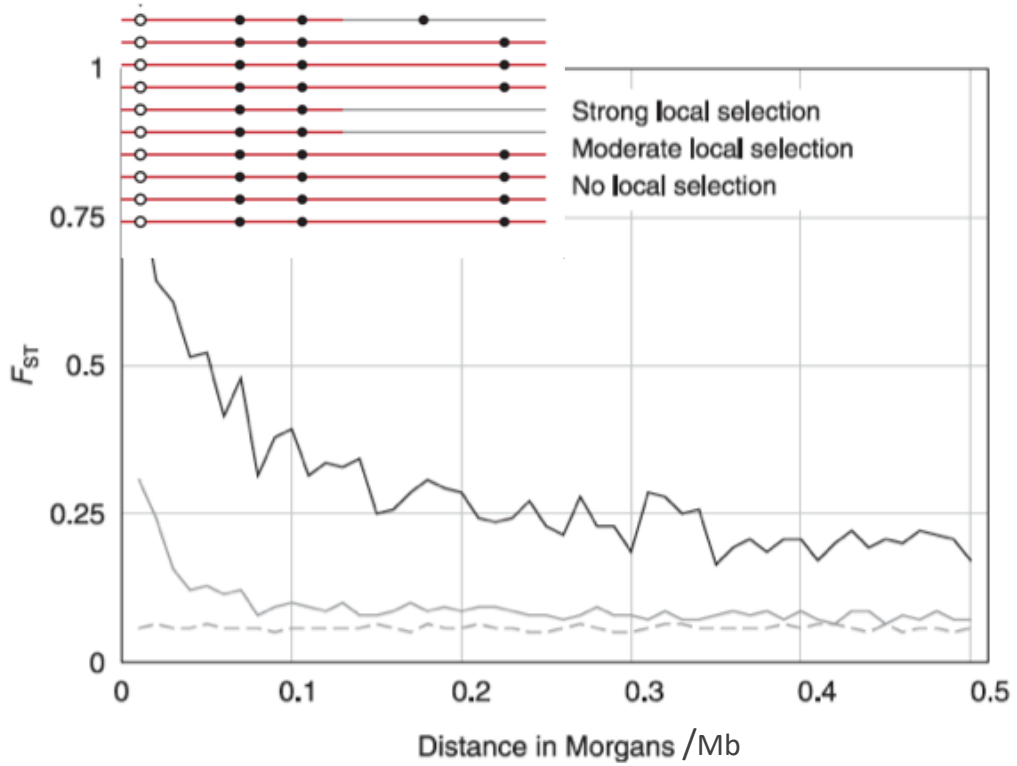
Reduced diversity (π, S)

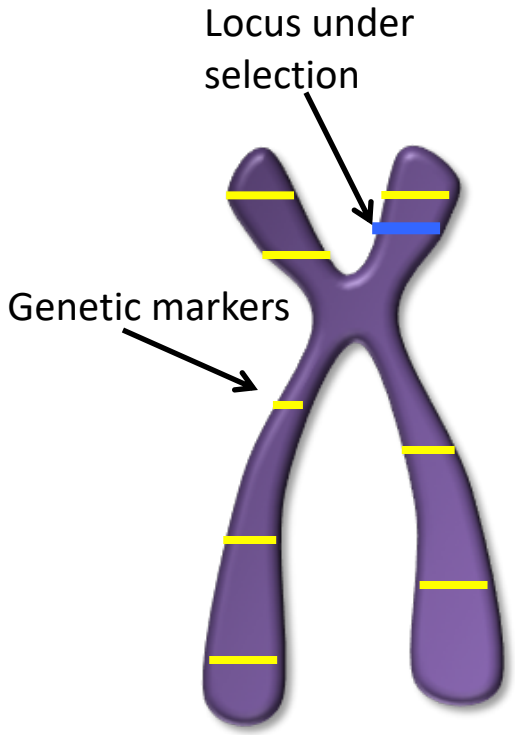
Strong locus specific differentiation

- F_{ST}-outlier
- $F_{ST} = \pi_{total} - \pi_{subpopulations}/\pi_{total}$

Figure 1 of Storz 2005, *Molecular Ecology*

Genetic hitchhiking





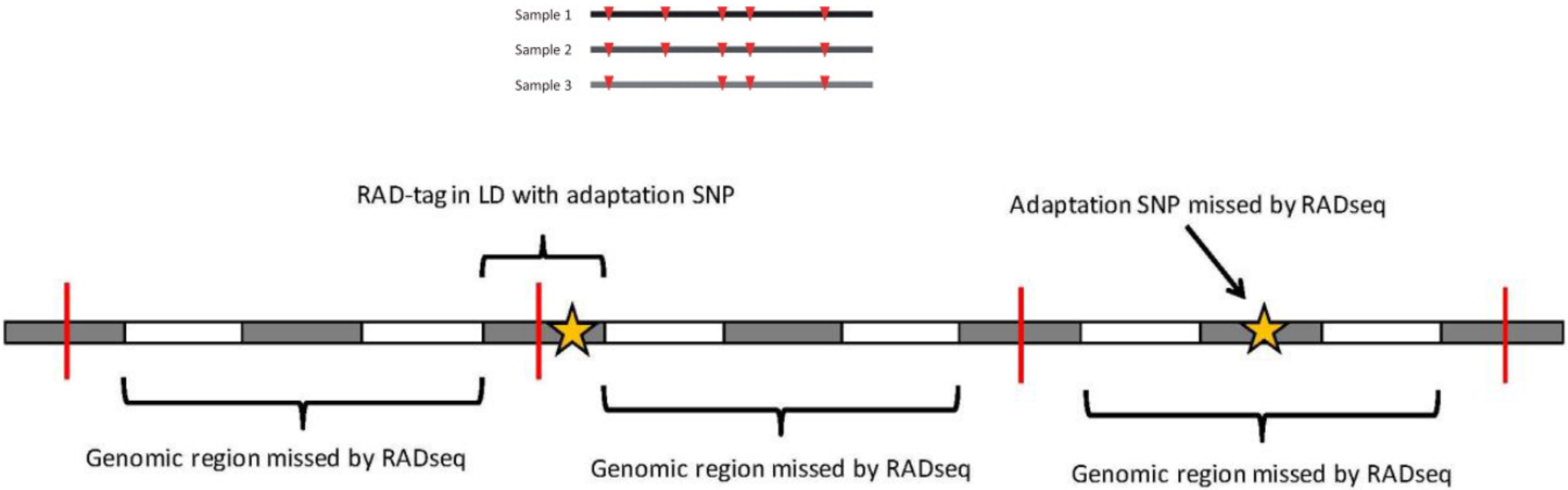
- The signature of selection decays with increasing distance from the locus under selection
  => **genetic hitchhiking**

- Recombination

Figure 1 of Storz 2005, *Molecular Ecology*

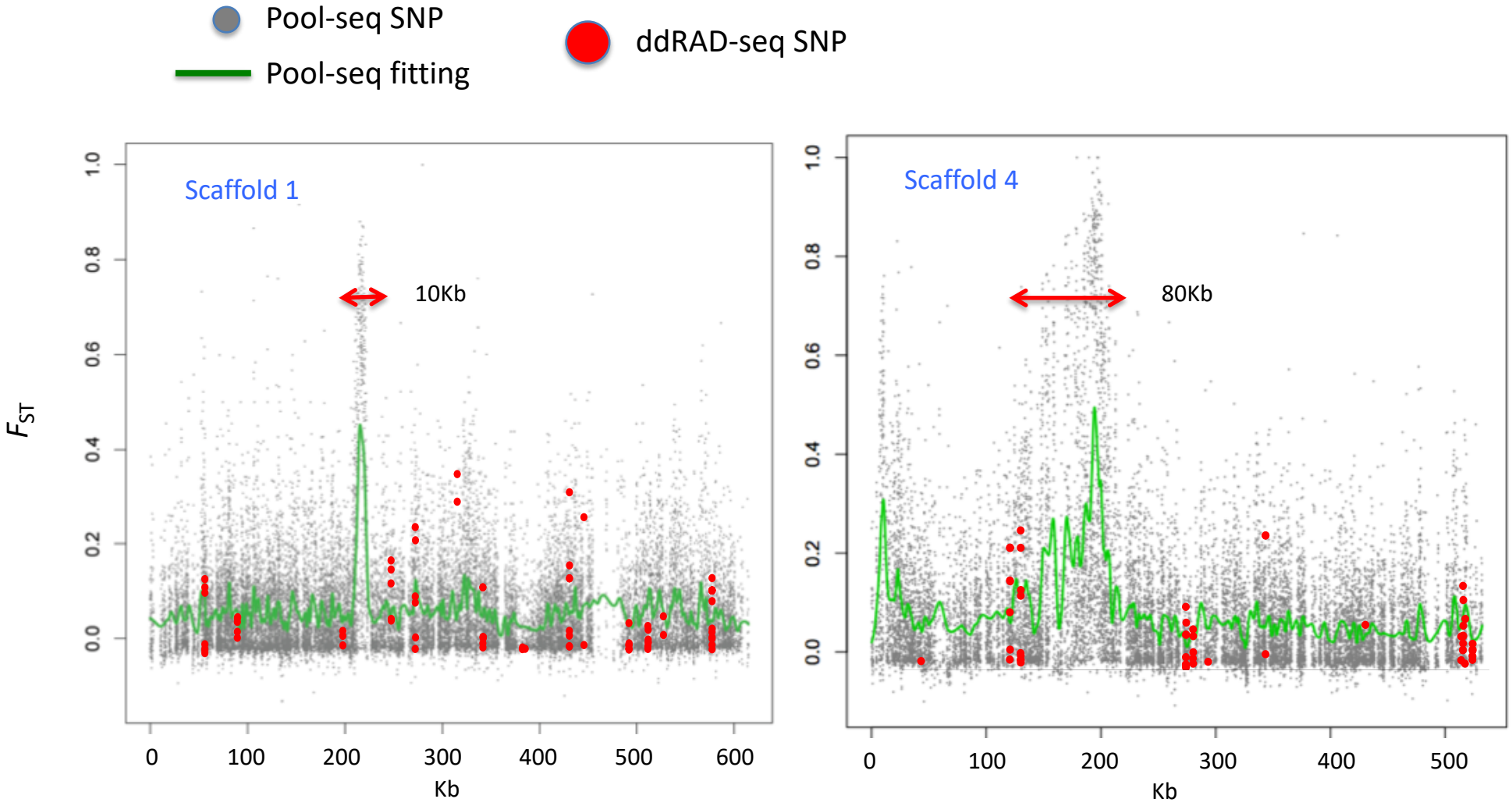
Marker density

- E.g. reduced representation sequencing (e.g. ddRAD-seq)

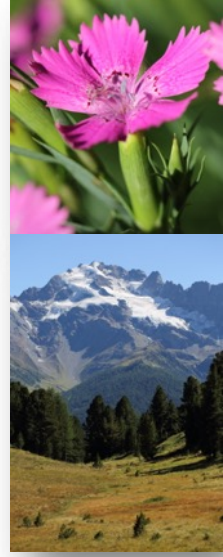


Lowery *et al.* 2016

Marker density is important to detect outliers!



RAD-seq data can't detect most of the 'islands of selection'



Provided by S. Fior

Four flavors of selection

Positive (directional)

- ’New’ (non-synonymous) mutations selected for
- Evolution of novel protein function

Diversifying

- Geographically restricted selection (e.g. due to spatial variation in climate)

Balancing

- Maintenance of multiple alleles within-population
- E.g. heterozygote advantage (sickle cell anemia), frequency dependent selection

Negative (purifying)

- New (non-synonymous) mutations selected against
- Retention of existing protein function

Case studies: Detect the genomic signature of selection

Different methods to detect selection:

- Reduced level of genetic variation (e.g. π)
- Linkage disequilibrium (*LD*)
- Skew of allele frequency spectra (Tajima's *D*)
- Locus specific population differentiation
  - F_{ST}-outlier approach
  - Population Branch Statistics (PBS)
- Environmental association analysis (EAA)
- Low coverage sequencing

Comparison:

- Within a population
- Among populations
- Among species

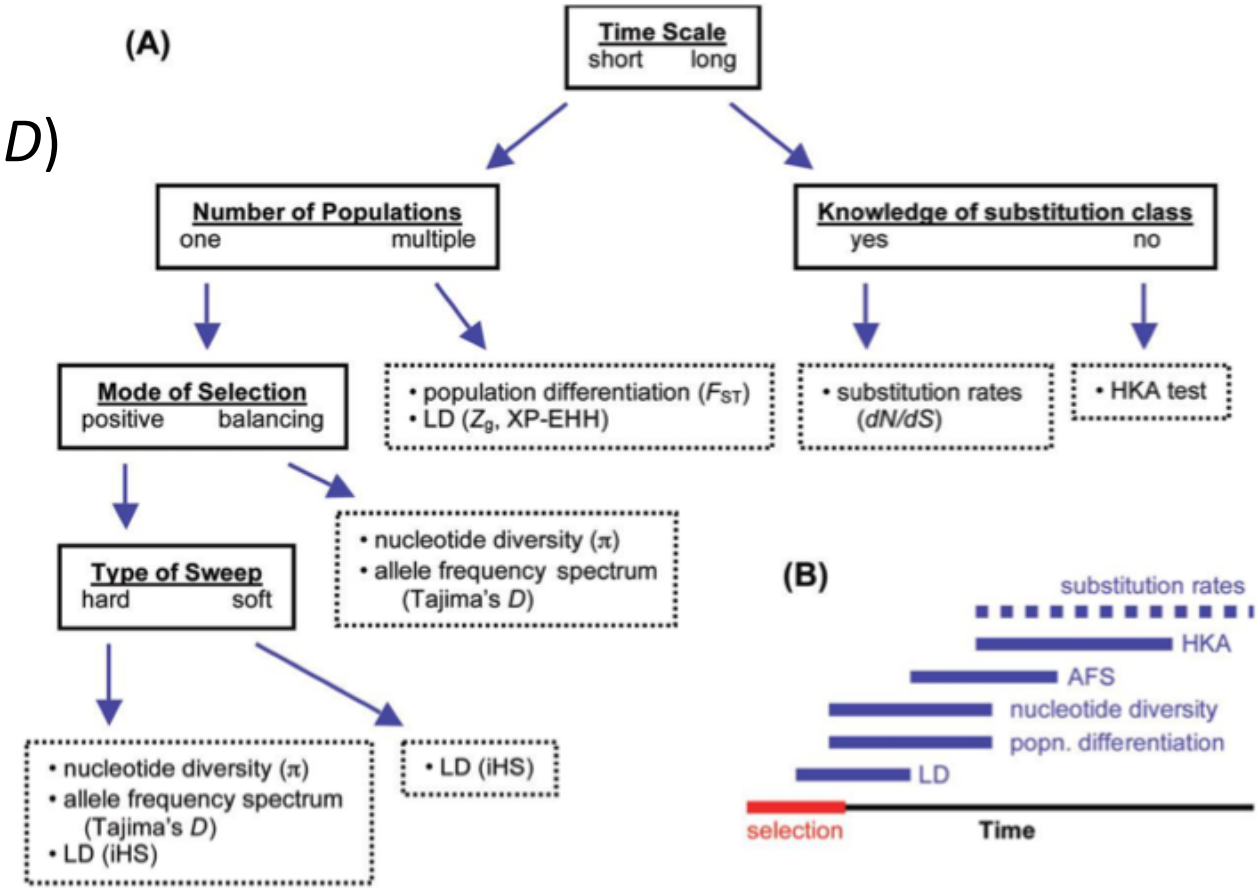


Figure 1 of Hohenlohe *et al*. (2010)

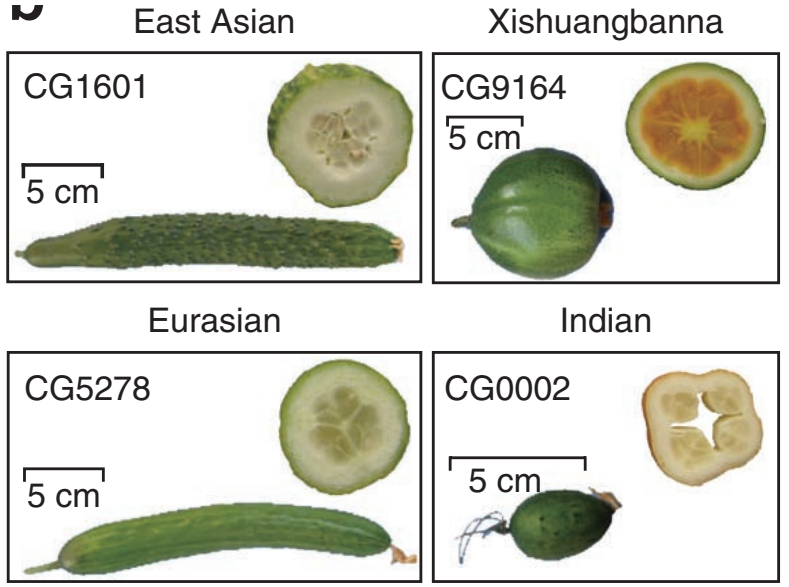
Reduced level of genetic variation (π)

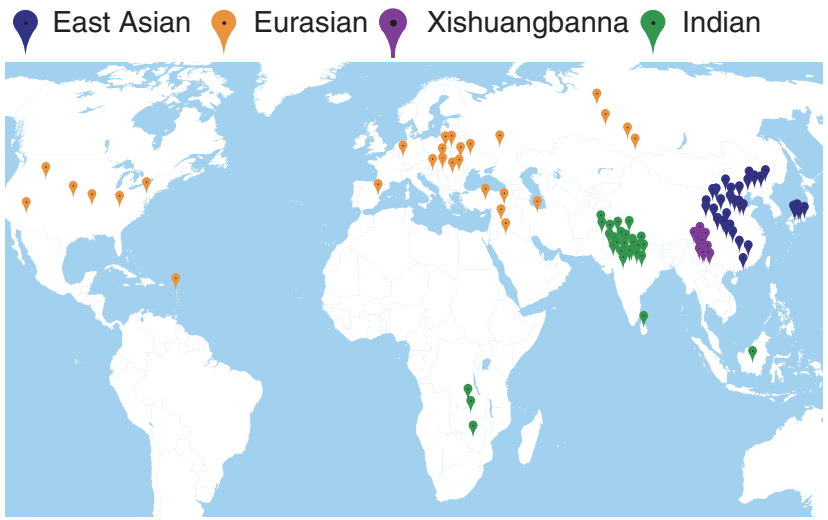
Cucumber

- 3 cultivated (C) and 1 wild (W) cucumber groups
- Morphologically different

- Genome re-sequencing (n=115)

- π: nucleotide diversity
  - mean number of nucleotide substitutions per site between any two randomly selected DNA sequences in a population





Qi *et al*. 2013

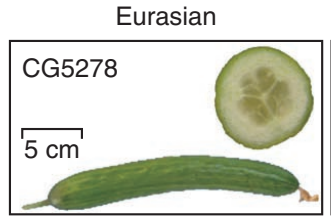
Reduced level of genetic variation (π)

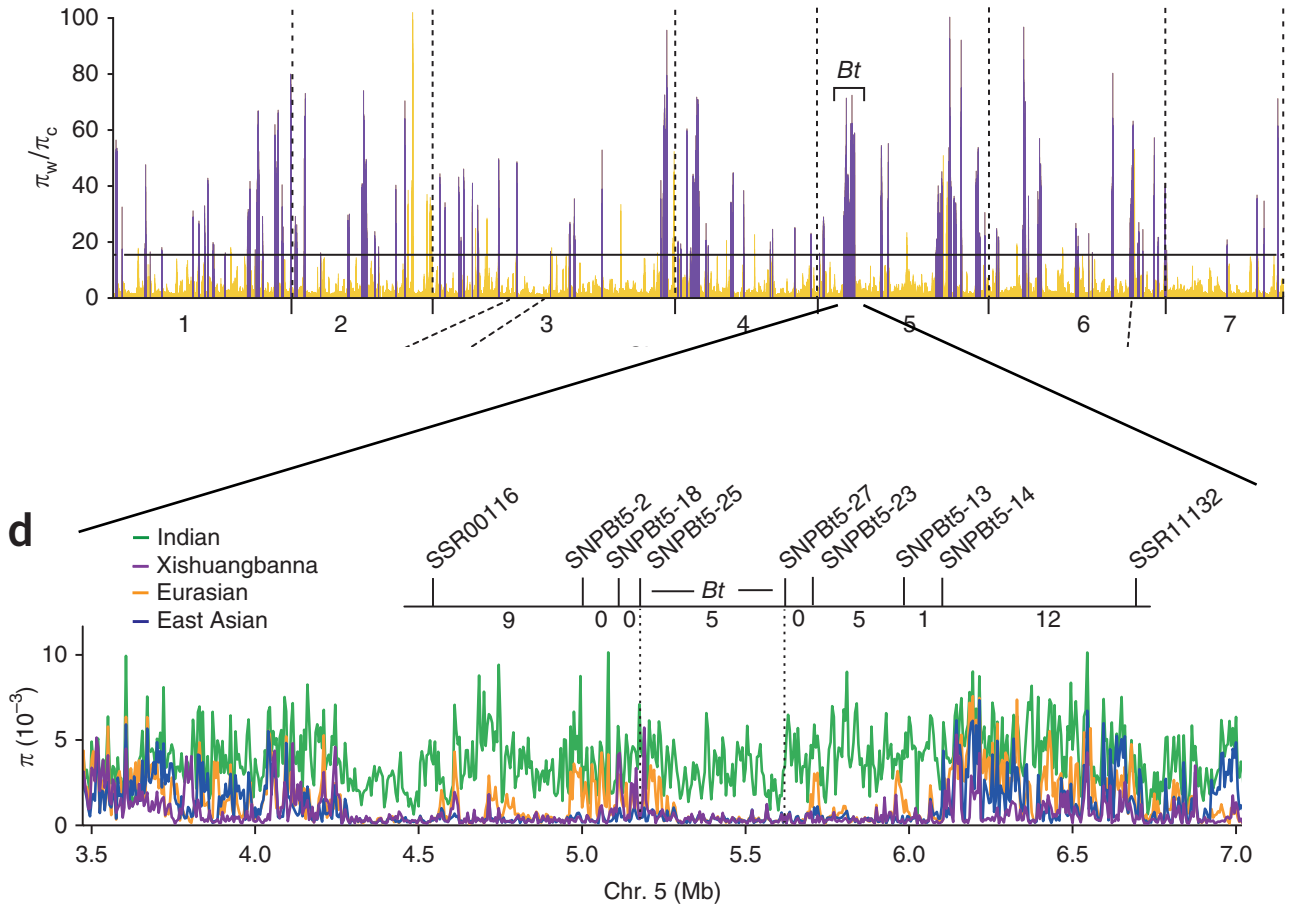
Domestication sweep

- Comparing π_W/π_C
- 112 regions detected

- ***Bt*** locus
  - Fruit bitterness

- Reduced π in cultivated cucumbers
- 2 Mb in length (Bt: 442 kb)



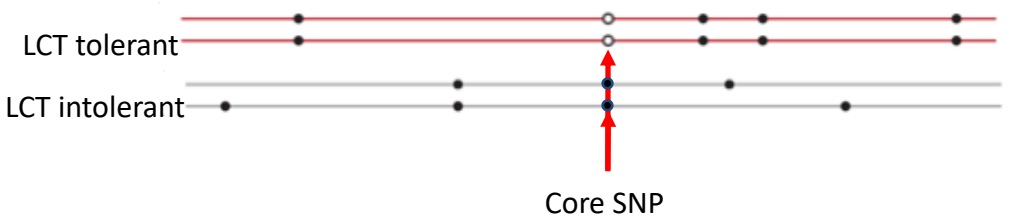


Qi *et al*. 2013

Linkage disequilibrium (*LD*)

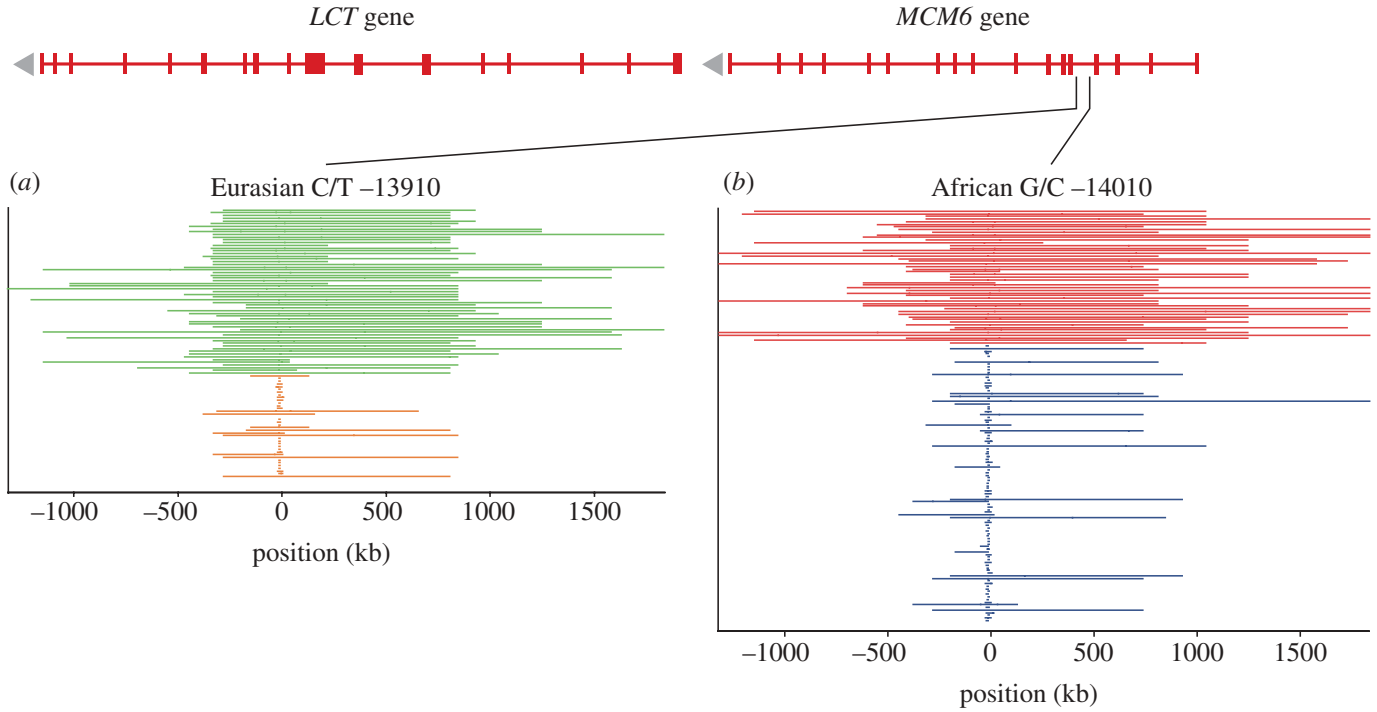
EHH: extended haplotype homozygosity

- Measures the decay of homozygosity from a ‘core’ SNP



LCT: lactose persistent gene in humans

- 2 Mb haplotypes
- Independent evolution in Africa and Europe



Tishkoff *et al.* (2007); Oleksyk *et al.* (2010)

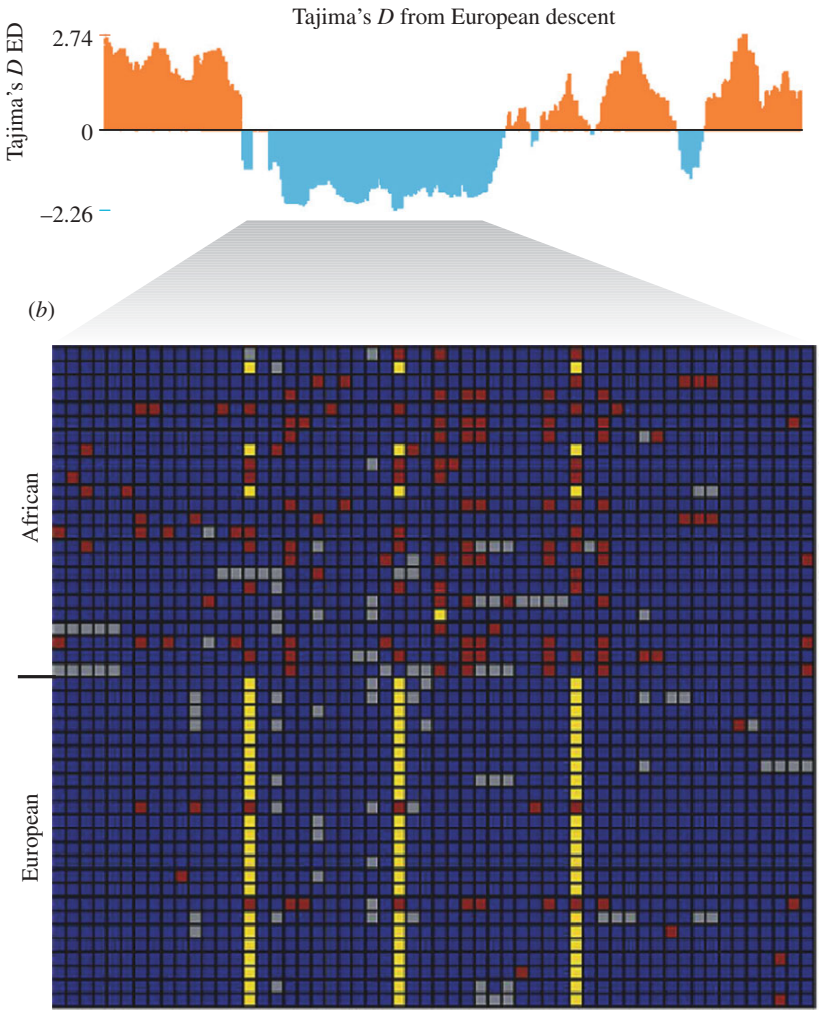
Skew of allele frequency spectra (Tajima's *D*)

Tajima's *D*

- Normalized difference between π and segregating sites (S, Θ_W)
- $d = \pi - \Theta_W$

$$D = \frac{d}{\sqrt{V(d)}}$$

- **Balancing selection**
  excess intermediate-freq. SNPs:
  $\pi > \Theta_W$, **+*D***
- **Positive selection**
  excess of low-frequency SNPs:
  $\pi < \Theta_W$, **-*D***
- Correct for demographic effects
  - Compare values against genome-wide estimates



Carlson *et al.* 2005; Oleksyk *et al.* 2010

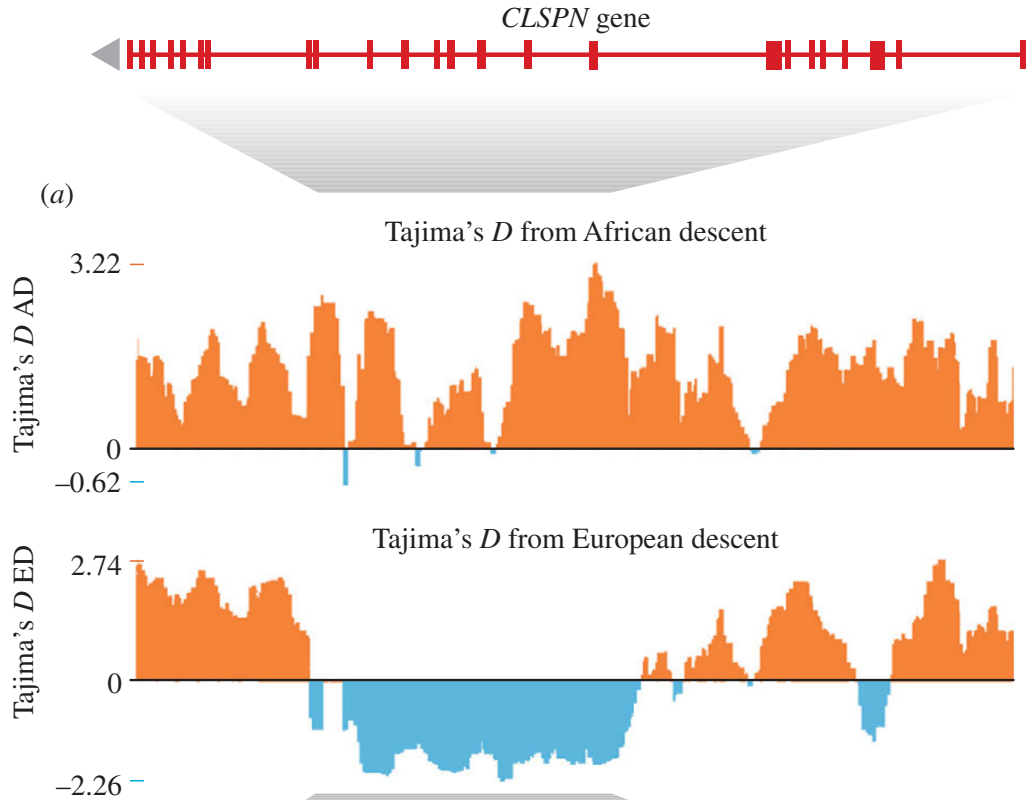
Skew of allele frequency spectra

Human *CLSPN* gene

- Inferred from dense SNP data
- Tajima's *D* plot

- Positive selection in European (-*D*)
- 1.5 Mb

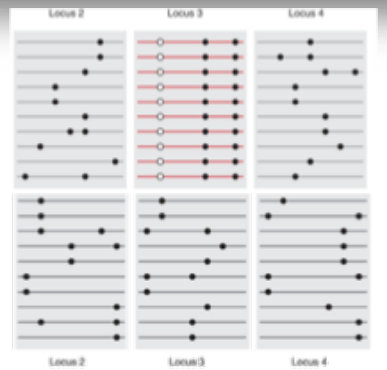
- Unknown function!

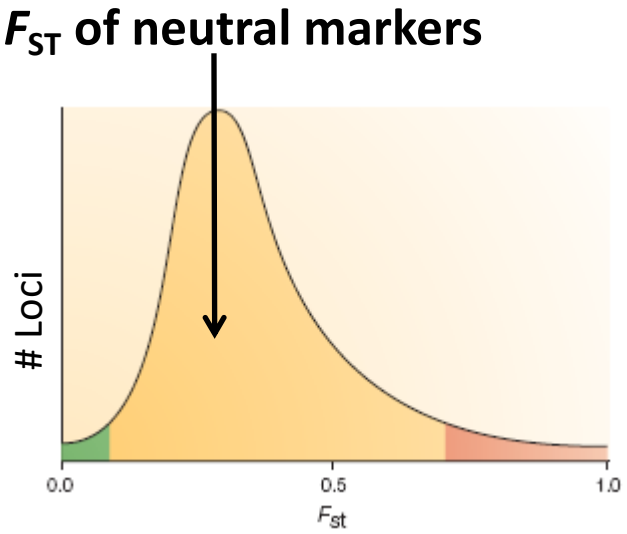


Carlson *et al.* 2005; Oleksyk *et al.* 2010

Model free F_{ST}-outlier approach

- >3 populations
- Screen many loci (>10,000; up to whole genome)
- **Outliers:** e.g. 95% quantile of F_{ST} distribution





Balancing selection

➢ relatively uniform frequencies across populations

⇒ low F_{ST} values

Fischer *et al.* 2014 PLoS One

Positive selection

➢ increased level of differentiation among populations

⇒ high F_{ST} values

Fischer *et al.* 2011/2013 MolEcol

Luikart *et al.* 2003

Alpine environments

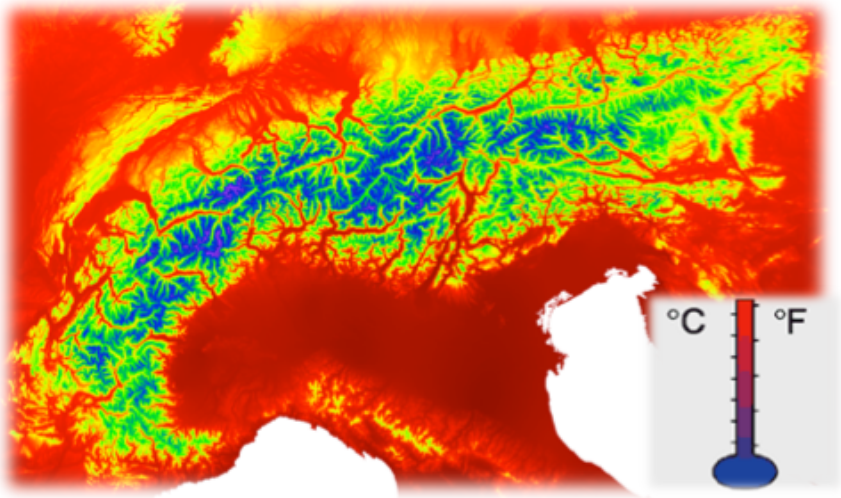
- Highly heterogeneous
- Strong environmental gradients

- Genetic basis of adaptation

MOLECULAR ECOLOGY
Molecular Ecology (2013) 22, 5594–5607 doi: 10.1111/mec.12521

Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps

MARTIN C. FISCHER,* CHRISTIAN RELLSTAB,† ANDREW TEDDER,‡ STEFAN ZOLLER,§ FELIX GUGERLI,† KENTARO K. SHIMIZU,‡ ROLF HOLDEREGGER*† and ALEX WIDMER*



Fischer *et al.* 2013

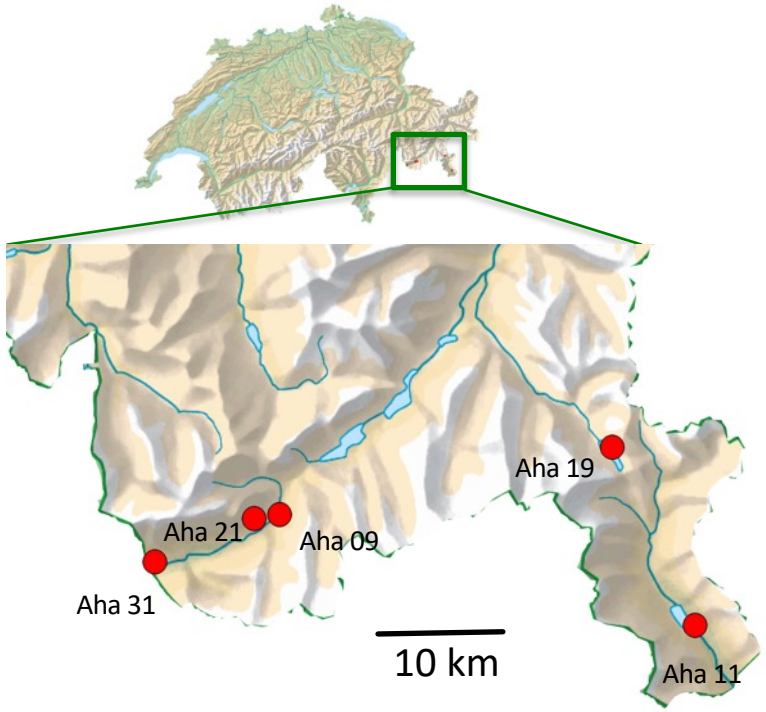
Study organism

Arabidopsis halleri

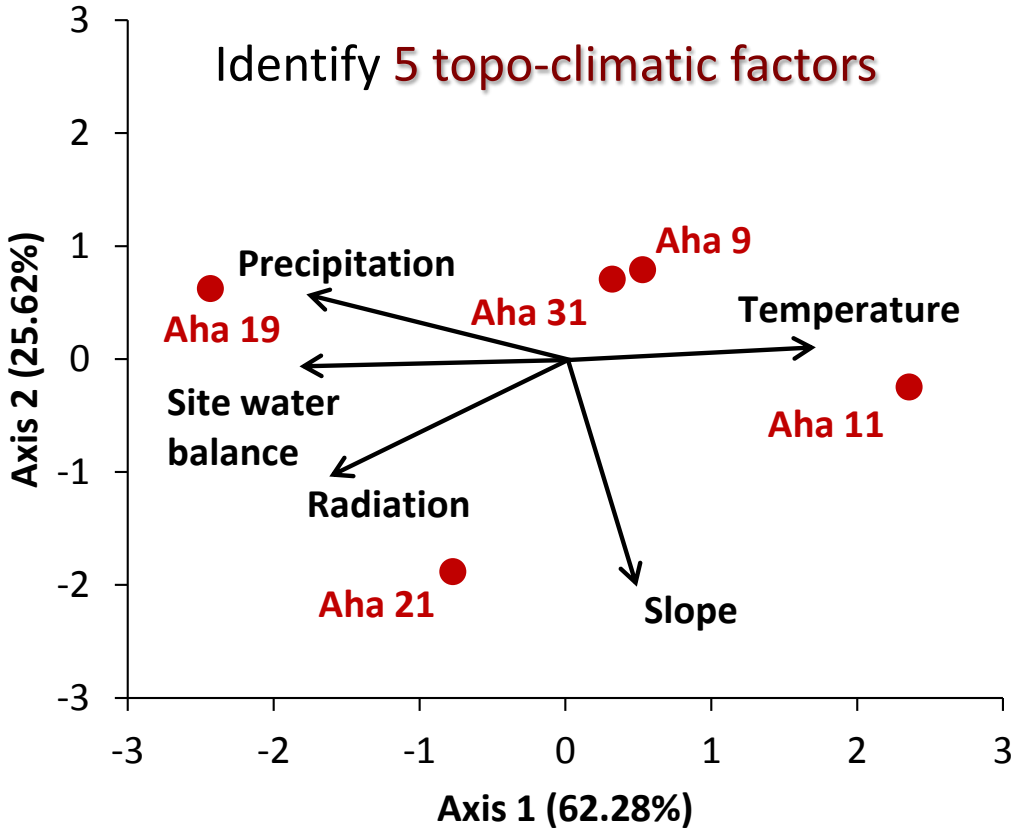
- Close relative of the model organism *A. thaliana*
- Genome size 255 Mbp
- Strictly outcrossing
- 300 – 2400 m a.s.l.

Heterogeneous Alpine environments

- 5 populations in close vicinity (2 – 45 km)
- 20 individuals each
- Cover wide range of abiotic environmental conditior
  - E.g. 790 – 2308 m a.s.l.



Identify 5 topo-climatic factors

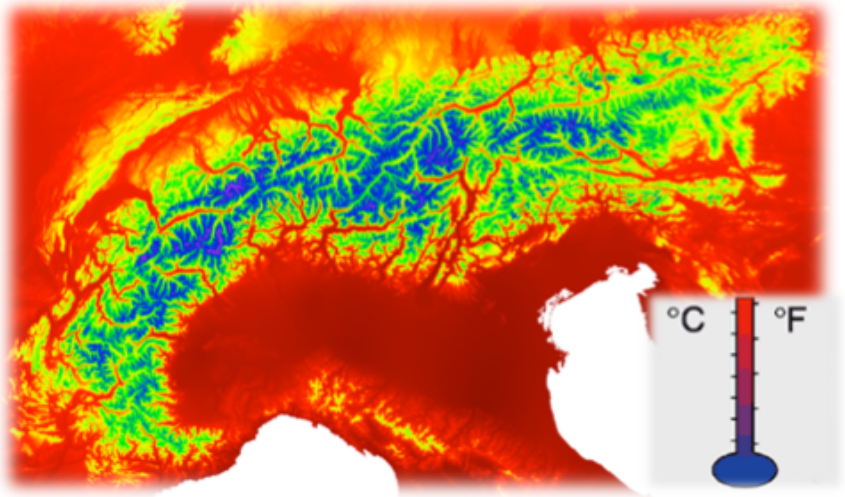


Zimmermann & Kienast 1999; Fischer *et al.* 2013

Adaptive genomics

Many abiotic factors are changing on short distance

- Expected signature of selection?
  - No signature of selection
  - Some genes with major effects
  - Many different genes under selection



- Which genes are involved in adaptation?
  - Gene functions?
  - Abiotic factors?

- Population genomic approach

Whole genome re-sequencing

- No bias from insufficient marker density or distribution

✧ **Pool-Seq**; pooled population approach

- ➢ Cost effective => 5 libraries
- ➢ Reduces amount of DNA required

✧ 20 diploid genomes pooled per population

Futschik & Schlötterer 2010; Turner et al. 2010; Rellstab et al. 2013; Schlötterer et al. 2014

Accuracy of Pool-Seq approach

OPEN ACCESS Freely available online

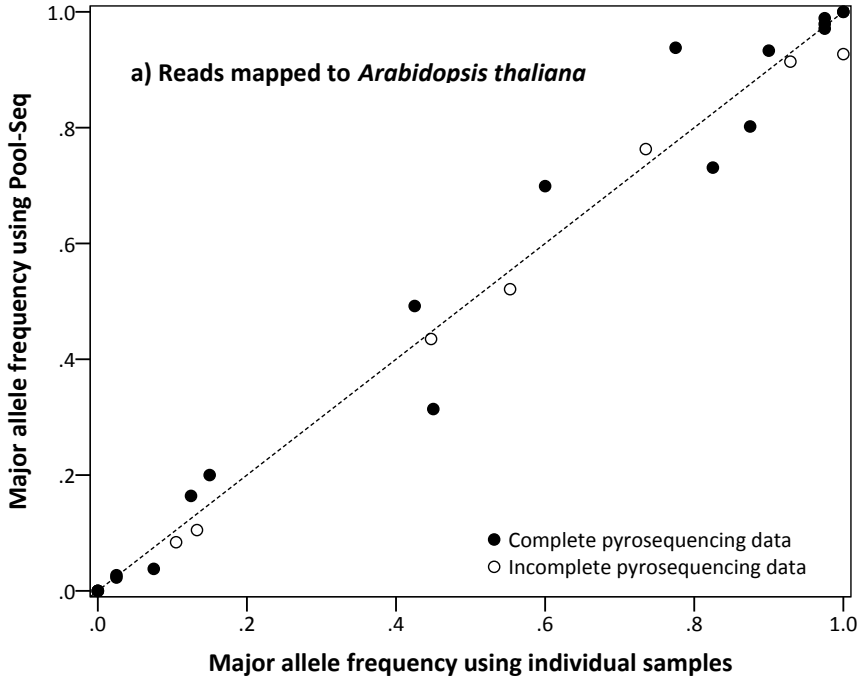
PLOS ONE

Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species

Christian Rellstab[1], Stefan Zoller[2], Andrew Tedder[3], Felix Gugerli[1], Martin C. Fischer[4*]

1 Biodiversity and Conservation Biology, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland, **2** Genetic Diversity Centre, ETH Zürich, Zürich, Switzerland, **3** Institute of Evolutionary Biology and Environmental Studies and Institute of Plant Biology, University of Zürich, Zürich, Switzerland, **4** Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland

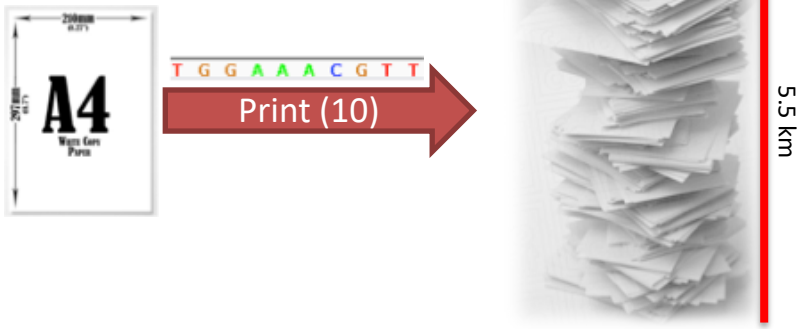
- Individual SNP genotyping
- 3 populations
- 9 SNPs validated
- PyroMark
- $R^2 = 0.98$



Rellstab *et al.* 2013

Population genomics

- Mapping of reads onto *A. thaliana* ref. genome
  - BWA (Li & Durbin 2009)
  - ~60x coverage
  - >120 billion bases

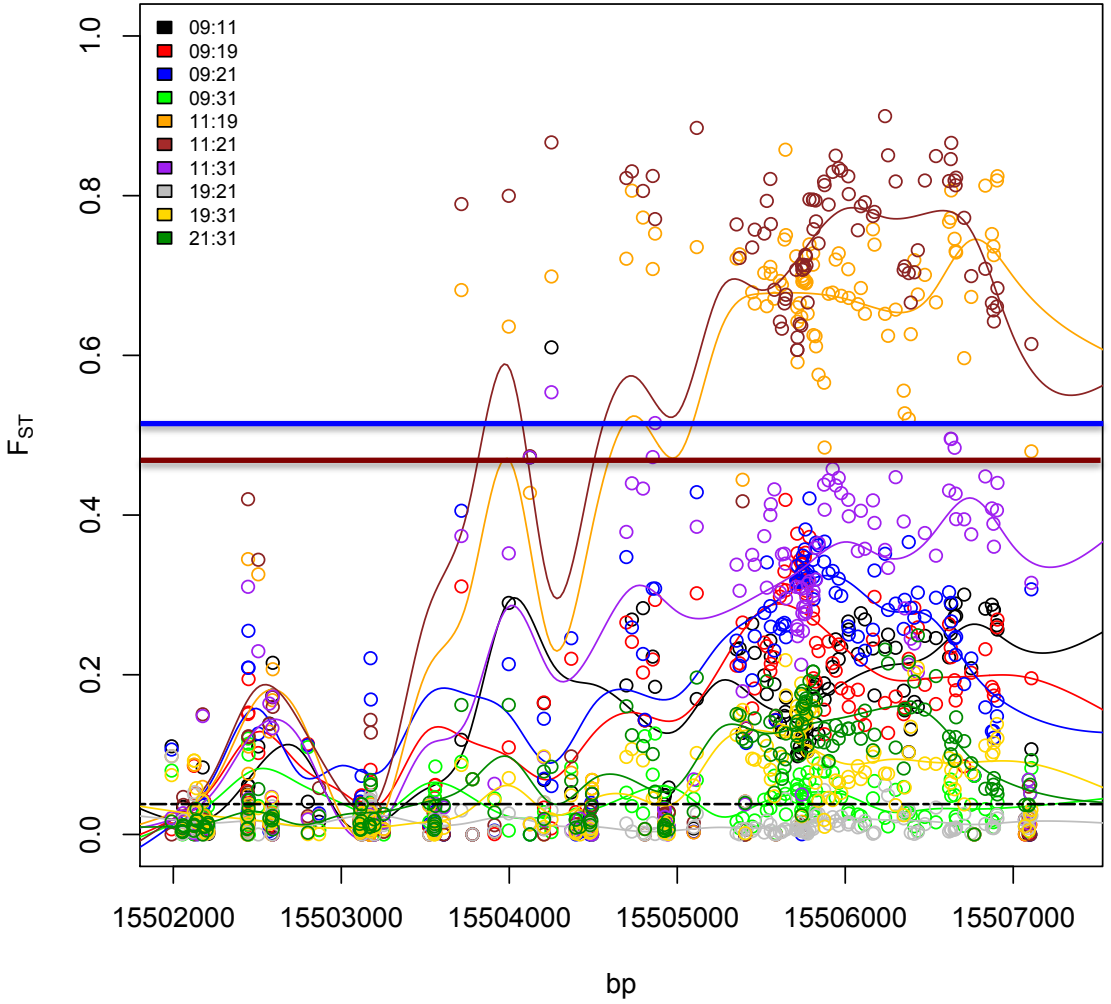


- PoPoolation2 (Kofler *et al*. 2011)
  - Identified **2,091,957 SNPs**
  - 25,764 genes covered
  - Genome-wide pairwise F_{ST}
    - Mean: 0.038

F_{ST}-outlier detection

3 step model free approach:

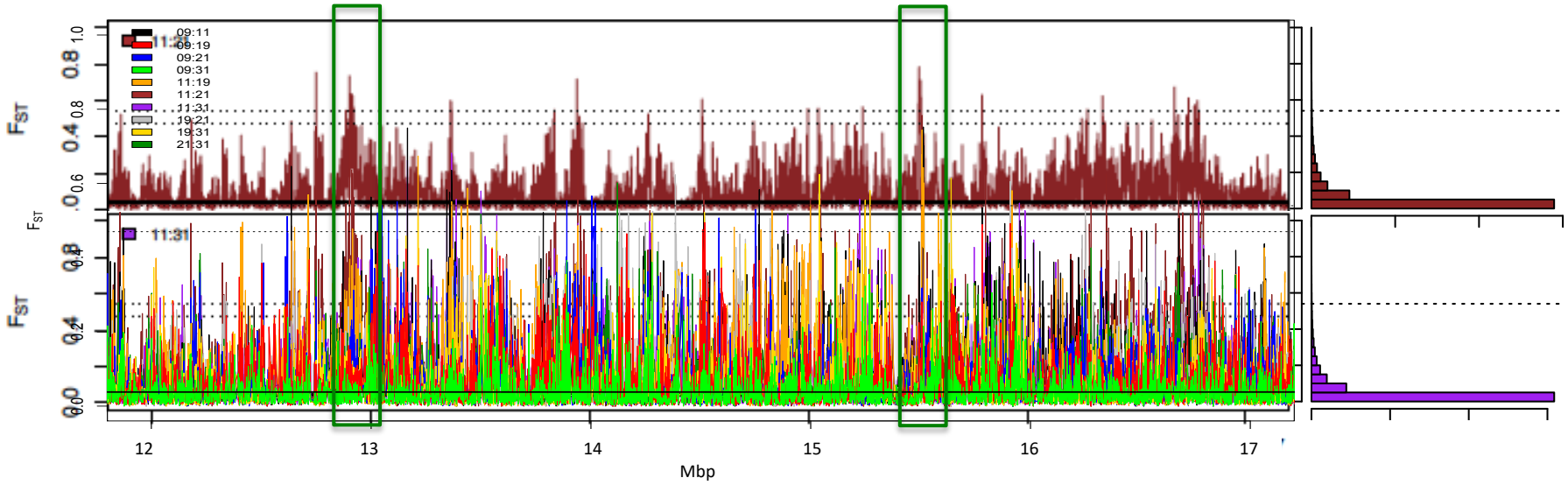
- 0.1% of highest ***F_{ST}* regions**: $F_{ST} > 0.47$
  - F_{ST} **sliding windows** approach of 500 bp
- 0.1% of highest ***F_{ST}* SNPs**: $F_{ST} > 0.54$
  - Corrected for population structure
- Corrects for coverage
  - *Fisher's exact test*
  - $p < 2.39*10^{-9}$
  - Strong allele frequency differences



Kofler *et al.* 2011; Fischer *et al.* 2013

Genomic Signature of Selection

- 4282 strong outlier SNPs
  - 0.2% of all SNPs
- 571 outlier genes

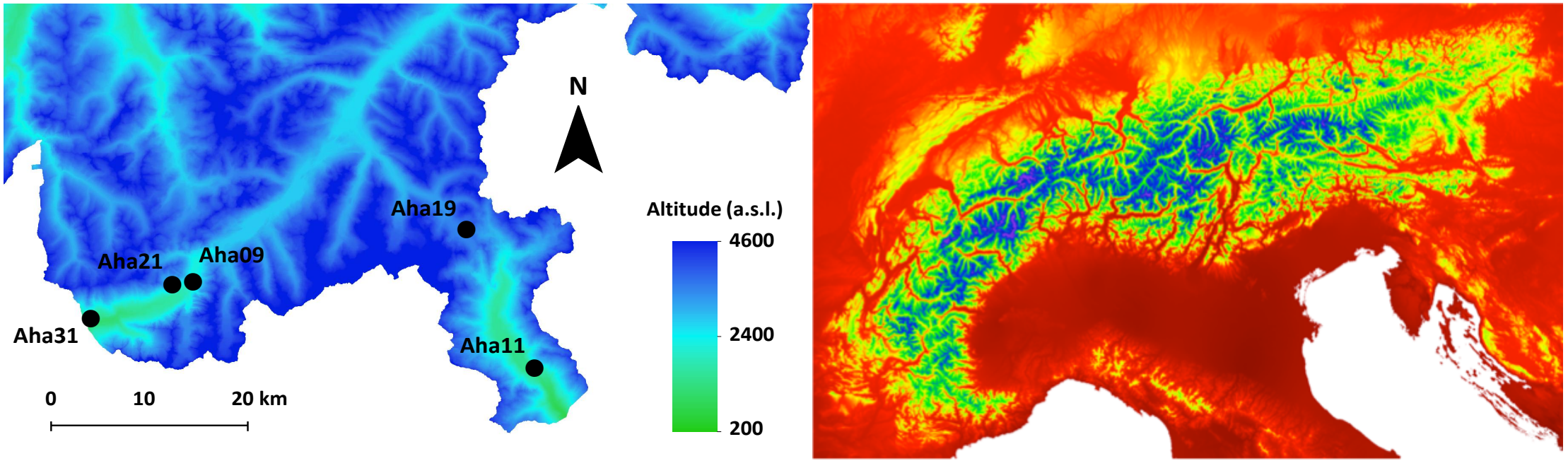


Fischer *et al.* 2013 MolEcol

Gene functions

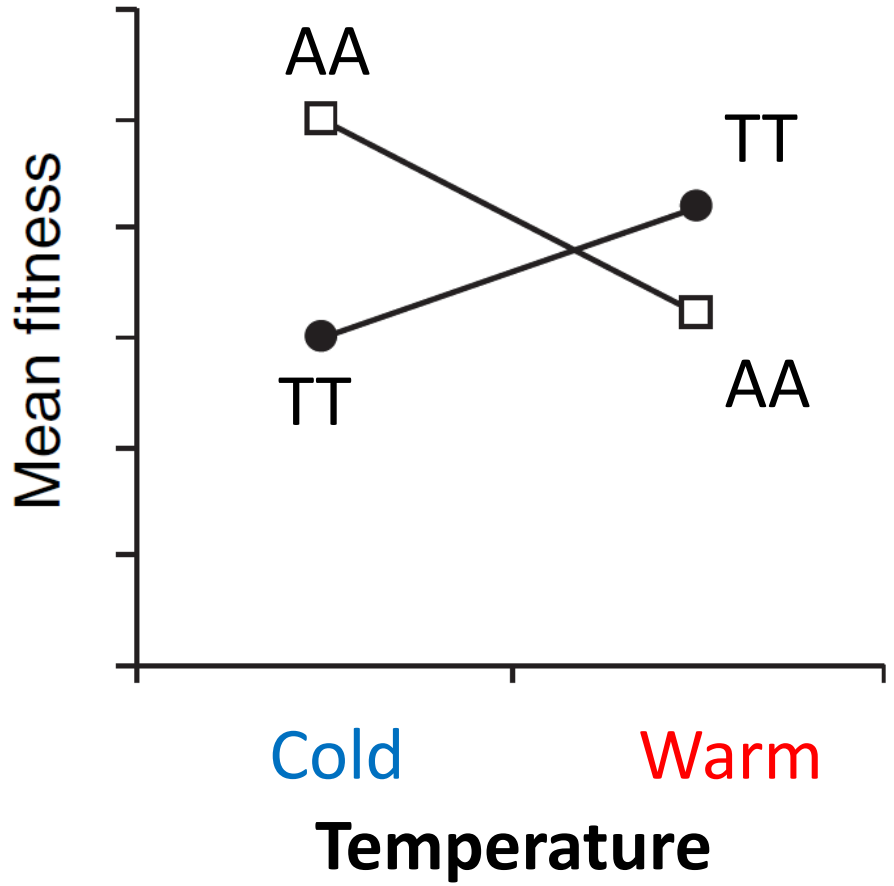
- 571 outlier genes
  - 139 genes (24%) unknown function

- Identify genes involved in environmental adaptation



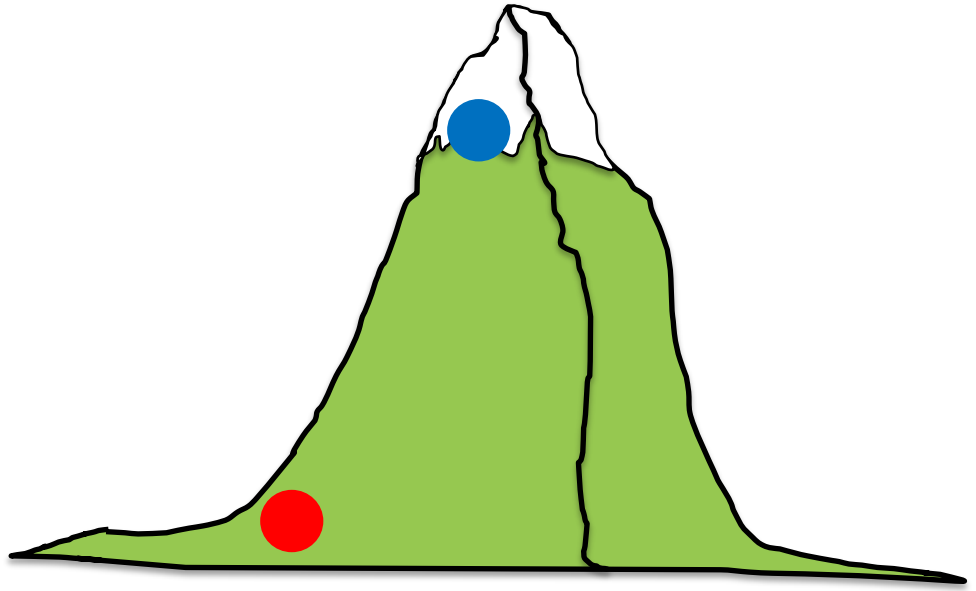
Environmental association analyses

□ Population from cold habitat
● Population from warm habitat



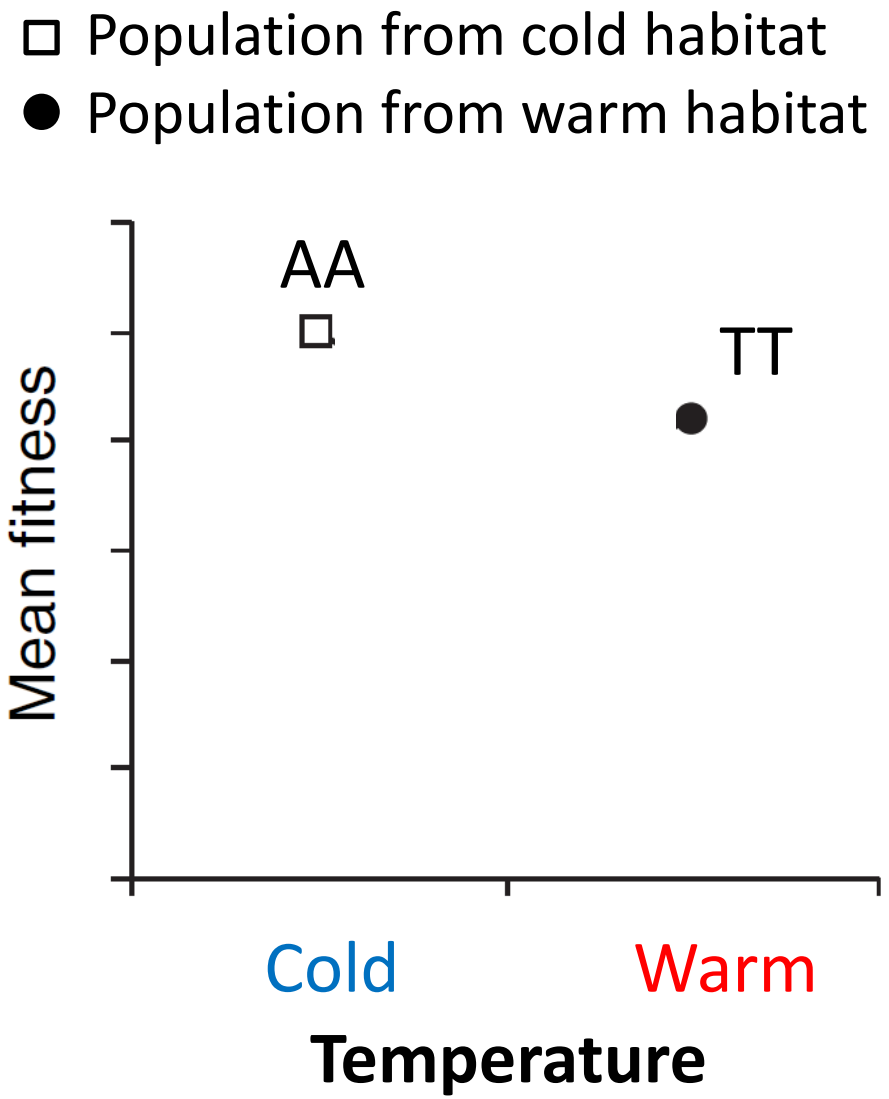
ACGTGTAGCAGTAGC**A**TGATGCTGATCGATT
ACGTTTAGCAGTAGC**T**TGATGCTGATCGTTT

Fitness difference is genetically controlled



Kawecki & Ebert 2004

Environmental association analyses

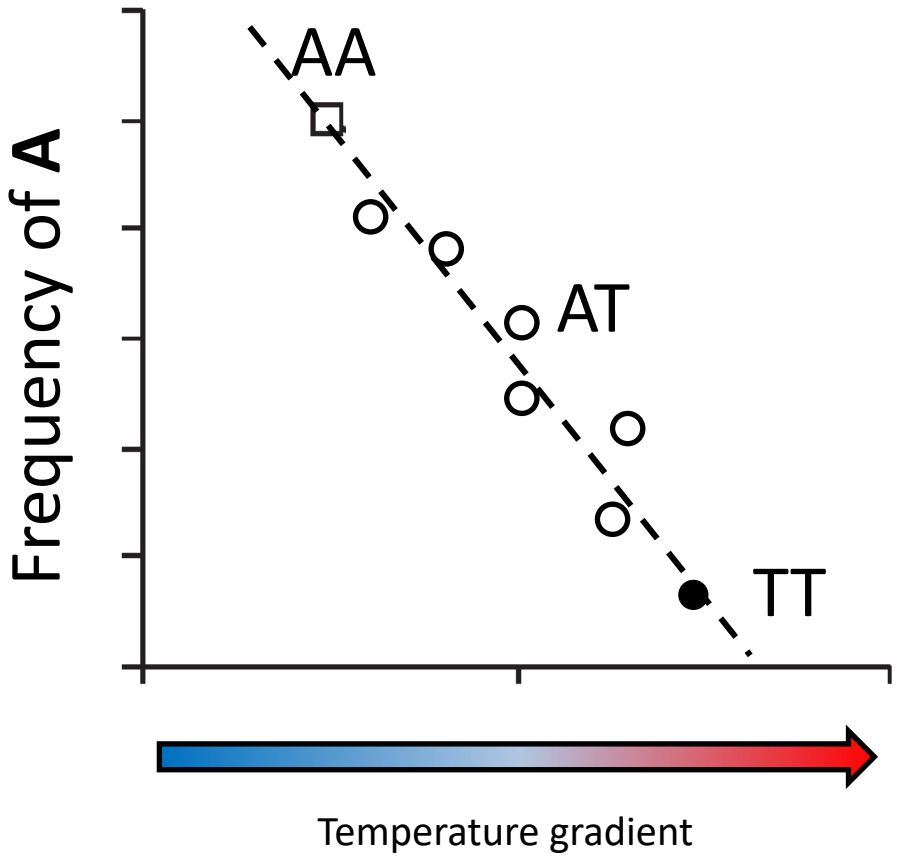


All populations are adapted to their local environmental conditions

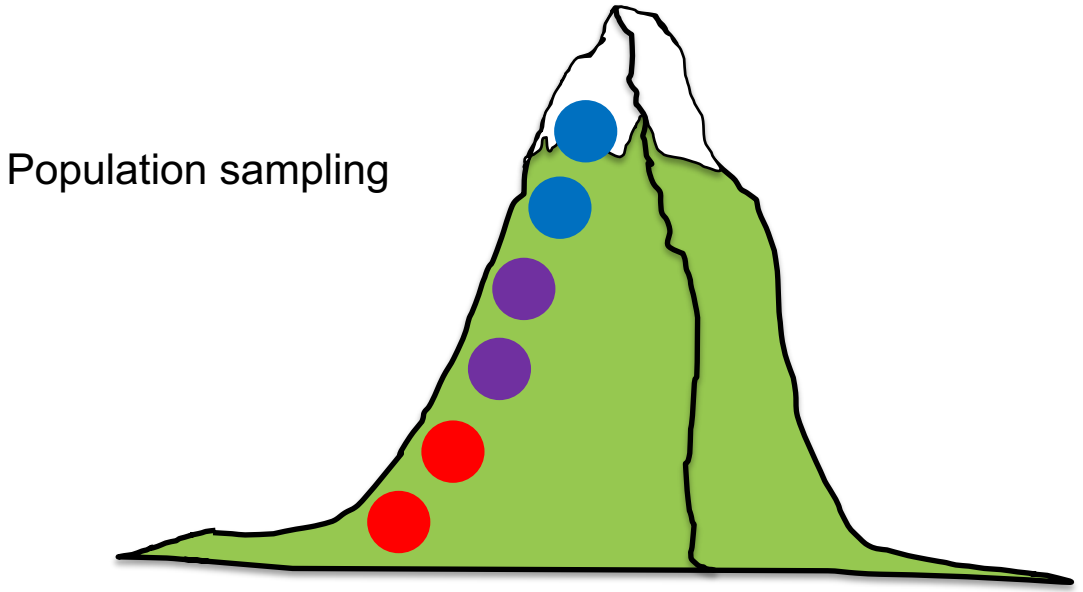
Kawecki & Ebert 2004

Environmental association analyses

- Population from cold habitat
- Population from warm habitat

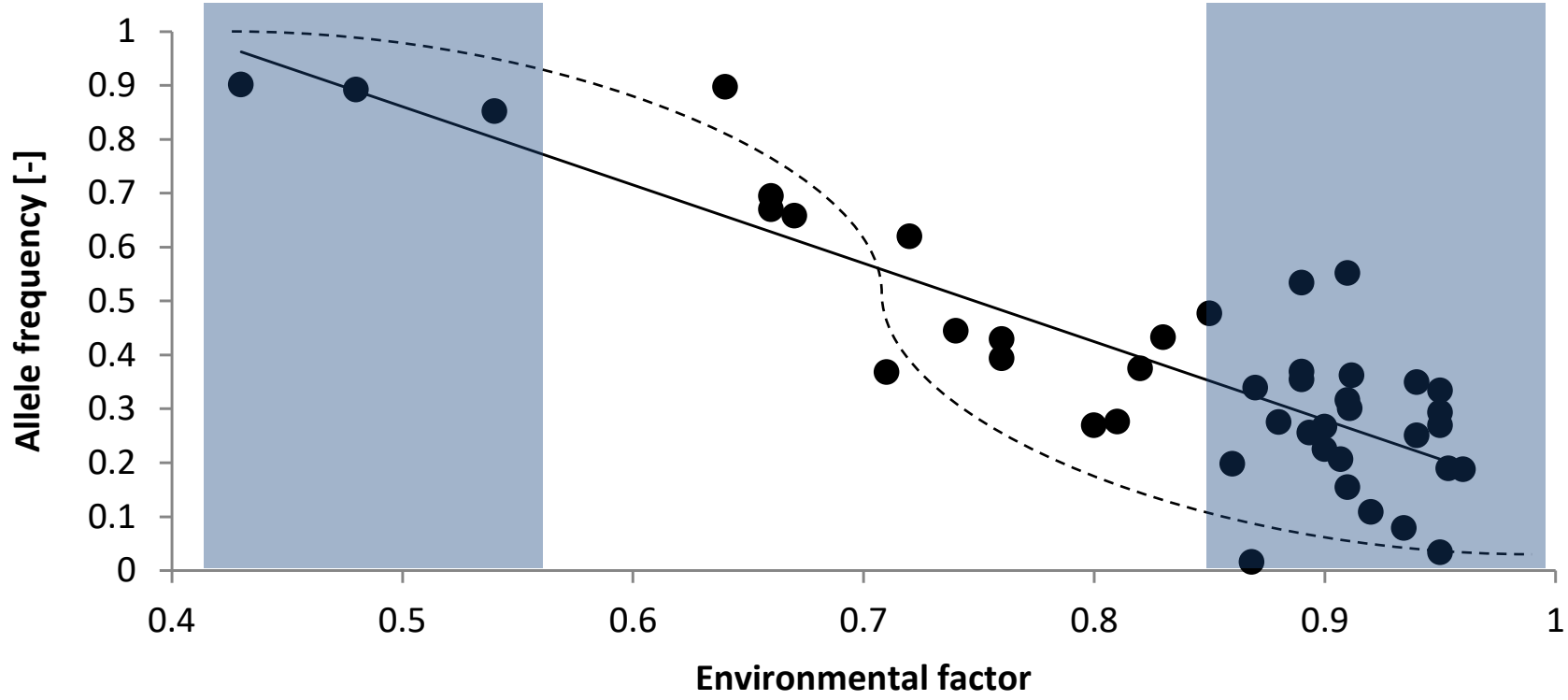


Additive genetic effects



Kawecki & Ebert 2004

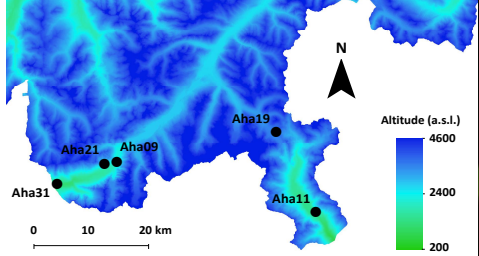
Environmental association analyses



- **Correlation/association of genotype with environmental factors**
- Identify alleles, SNPs, or genes associated with particular environmental factors (temperature, rain, soil, UV, …)
- Identify environmental factors that drive local adaptation

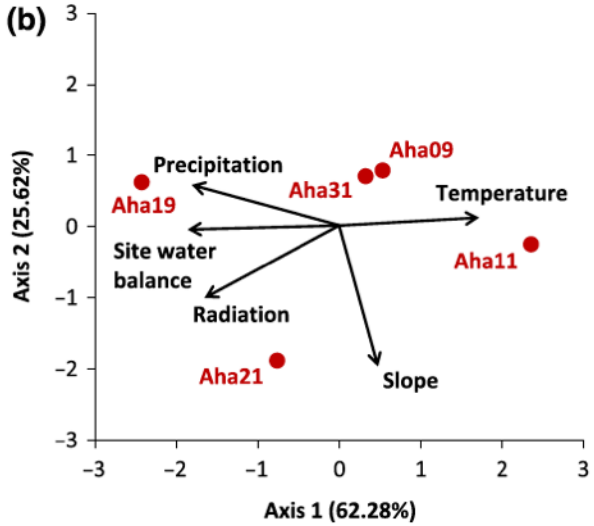
Kawecki & Ebert 2004

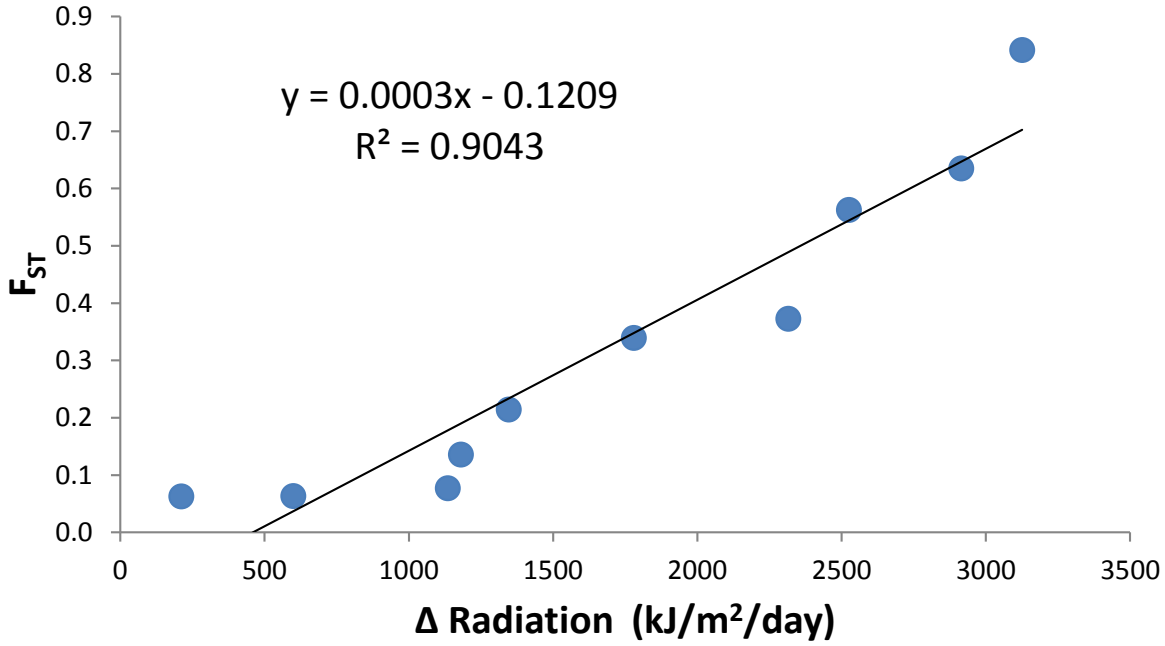
Environmental Associations



Partial Mantel tests

- Correlates two distance matrices:
  - Pairwise genetic distance of outlier SNPs
  - Pairwise climatic distance of environmental factors
- Controlling for population structure

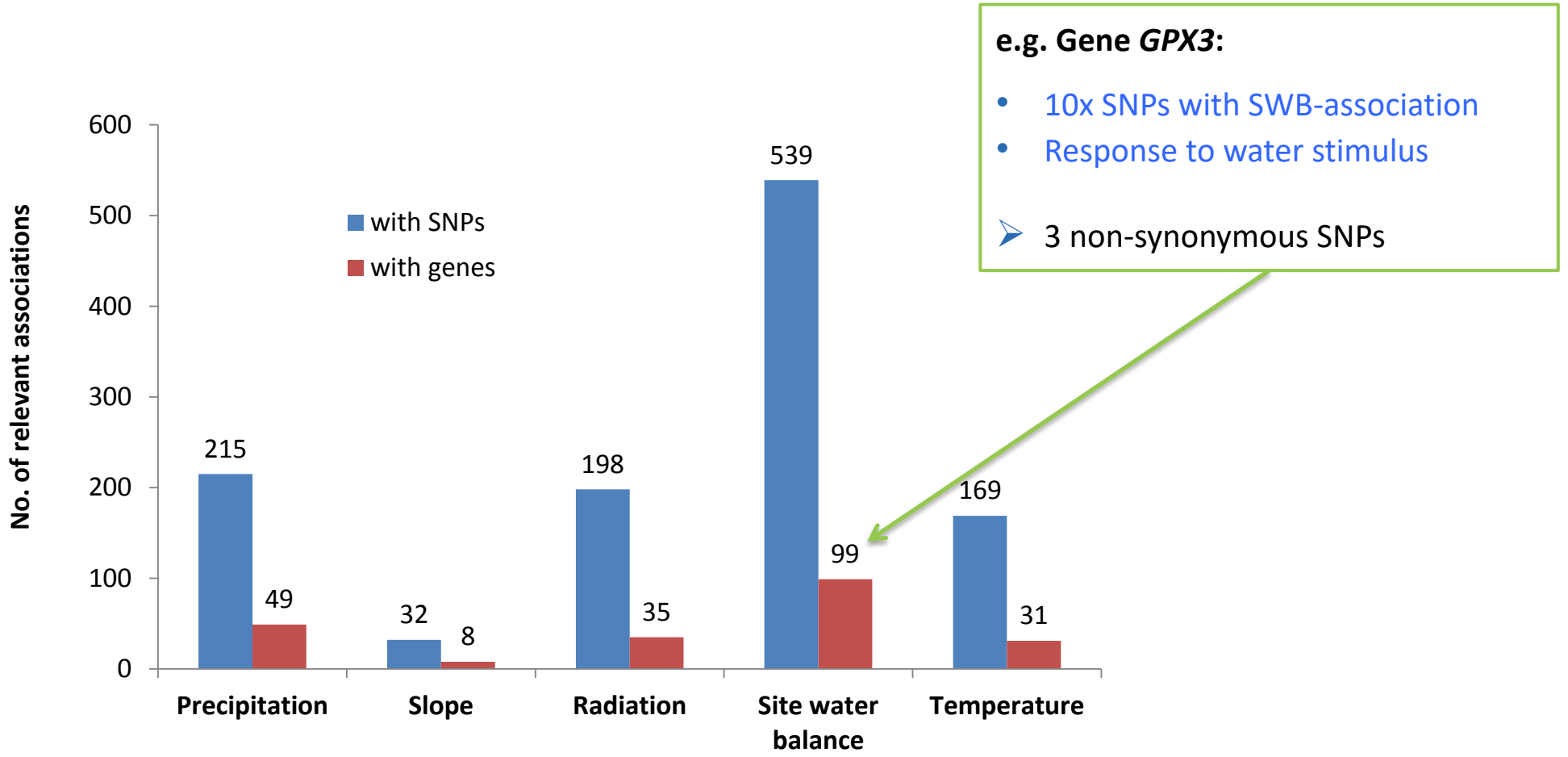




Fischer *et al.* 2013

Abiotic environmental associations

✧ 175 genes with environmental associations (**30.6%**; out of 571 genes)



e.g. Gene *GPX3*:

- 10x SNPs with SWB-association
- Response to water stimulus

➢ 3 non-synonymous SNPs

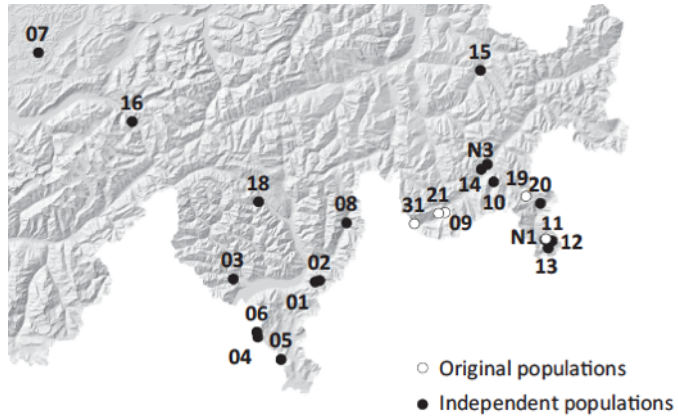
Fischer *et al.* 2013 MolEcol

Local remains local… … in a highly heterogeneous environment

ORIGINAL ARTICLE

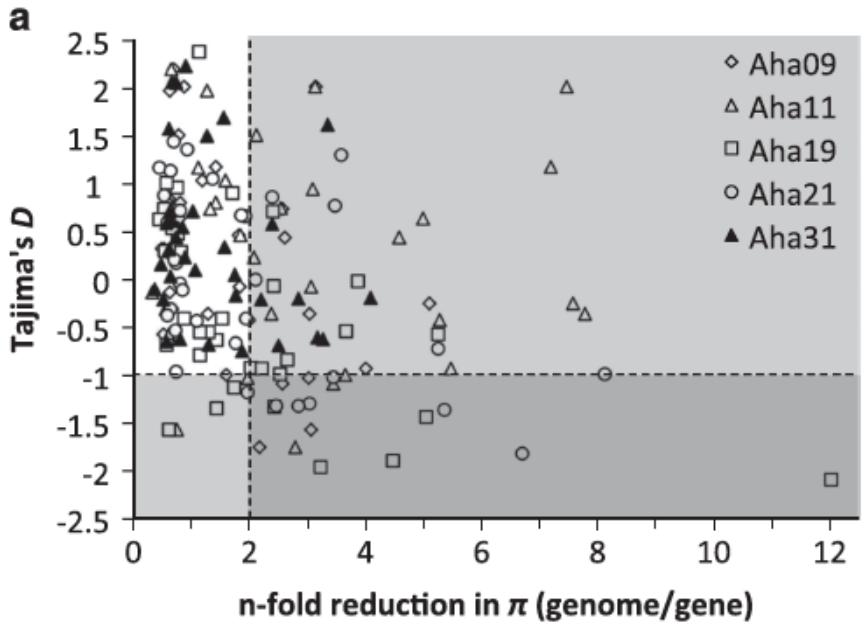
Local adaptation (mostly) remains local: reassessing environmental associations of climate-related candidate SNPs in *Arabidopsis halleri*

C Rellstab[1], MC Fischer[2], S Zoller[3], R Graf[1], A Tedder[4], KK Shimizu[4], A Widmer[2], R Holderegger[1,2] and F Gugerli[1]

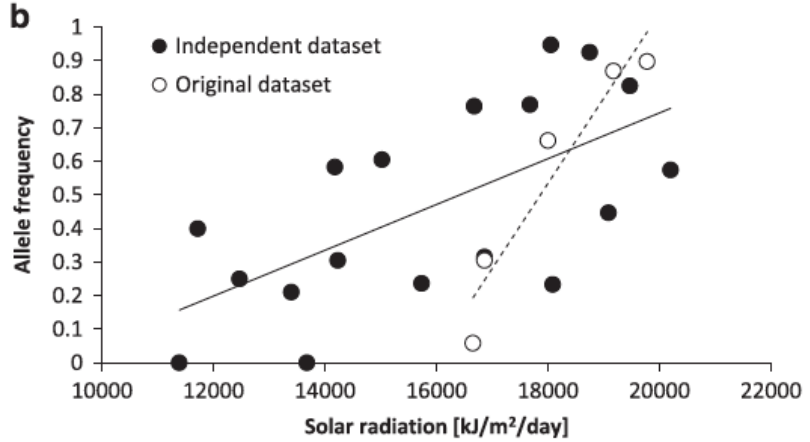


- 74 candidate SNPs tested for EAA on a large geographic scale
- 444 individuals in **23 populations**

- **Signature of selection** was mainly observed in singe populations



- LFMM and partial Mantel Test



- **31%** of the genes revealed the same **association**

Rellstab *et al.* 2017; Heredity

Last glacial maximum (18,000 years)

- The molecular signature of adaptation in Alpine populations of *A. halleri* is highly complex and local
- F_{ST}-outlier approached detect **recent selection**
- Selective signals mostly found in low elevation populations
- Alpine plants are already adapted to harsh environments
- Adaptation to **warm** and **dry** environments

Model based F_{ST}-outlier approach

Estimates probability of each locus to be under selection (Island model)

- **BayeScan** (Foll & Gaggiotti 2008; Fischer e*t al.* 2011)

- Bayesian model comparison

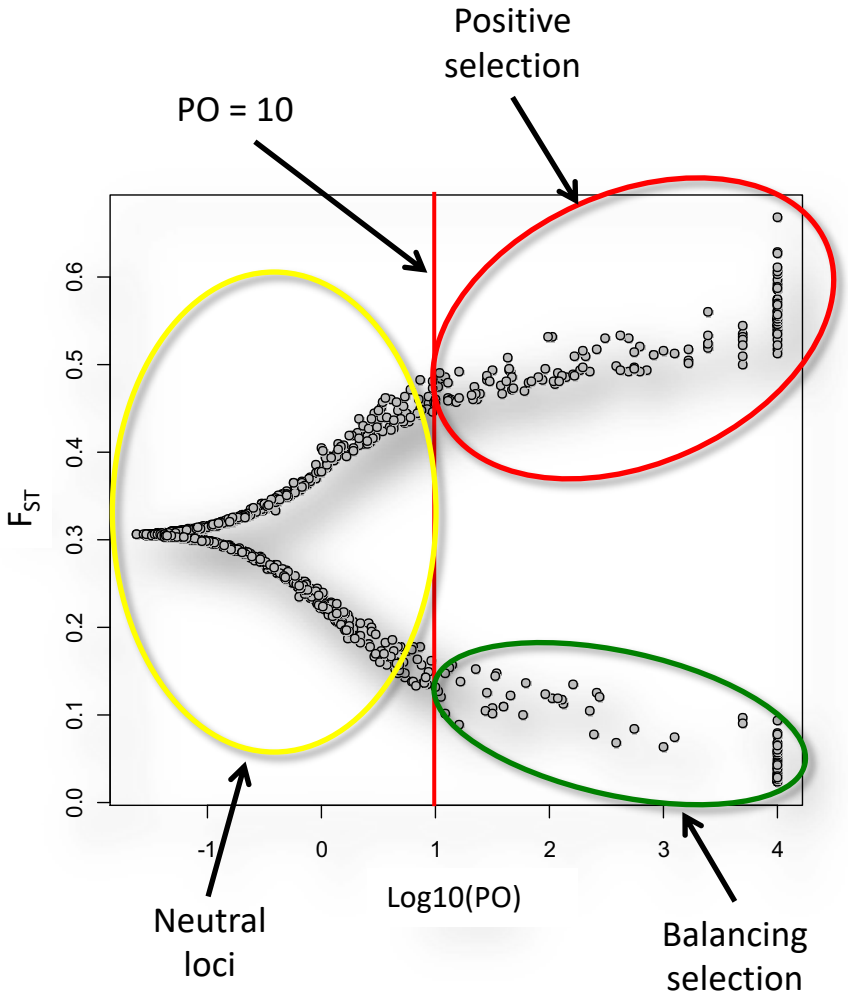
$$PO = \frac{\Pr(M_1|Data)}{\Pr(M_2|Data)}$$

Model with selection

Posterior odds

Neutral model

- PO >10 strong evidence for accepting a model (Jeffreys 1961)



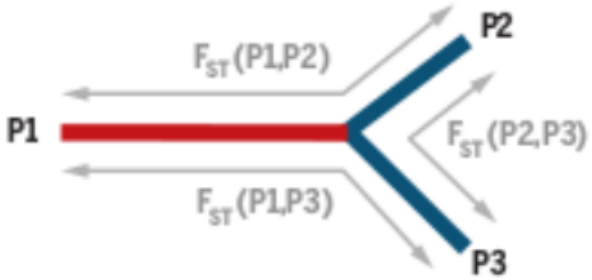
Foll & Gaggiotti 2008; Fischer e*t al.* 2011

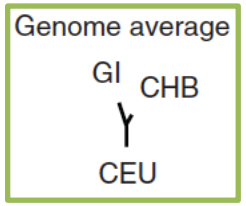
Population Branch Statistics (PBS)

Adaptation of Greenland Intuits to high altitude

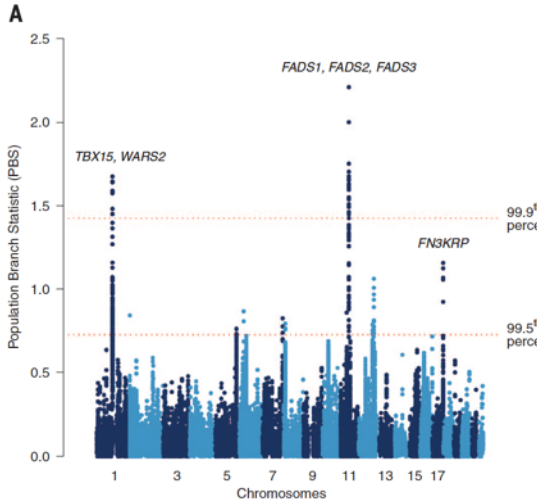
- 191 **Greenland Intuits** (**GI**; P1)
- 44 Han Chines (**CHB**; P2)
- 40 European(**CEU**; P3)

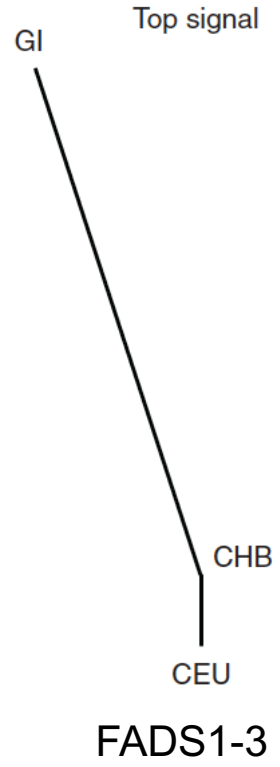
- Locus specific branch length
  - $PBS = (T^{12} + T^{13} - T^{23})/2$
  - $T^{12} = -\log(1-F^{12}_{ST})$





- FADS1-3
  - Fatty acid desaturase
  - Mutation compensate for high-fat diet





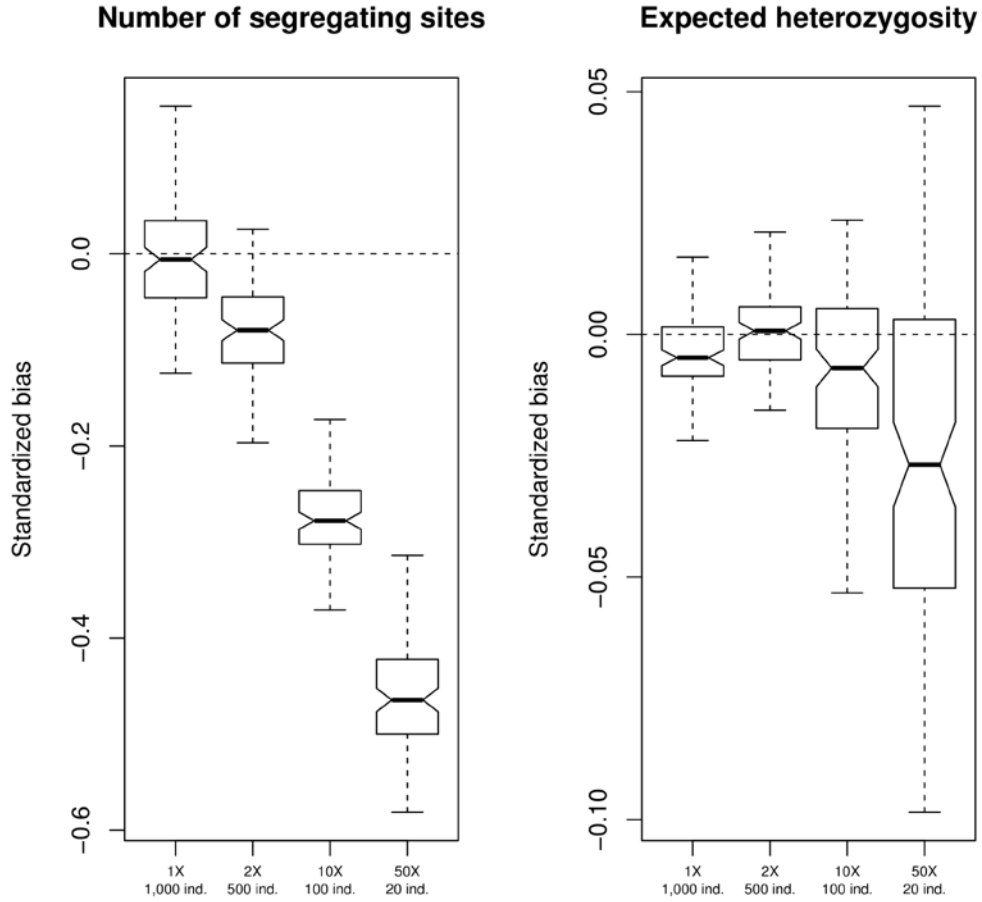
Fumagalli *et al.* 2015; Yi *et al.* 2010

Genotype likelihoods (ANGSD)

- Medium/**low coverage** individual sequencing: **1x**-10x
- Ultra low coverage individual sequencing: <1x

- Sequencing many individuals!

- Cheap library preparation, ~5.- CHF (Therkildsen & Palumbi 2017)



Fumagalli *et al.* 2010

Genotype likelihoods

- Using **genotype likelihoods** (ANGSD) instead of genotypes for downstream analysis

Summarise the data in 10 genotype likelihoods

bases (b):
TCCTTTTTTTT
quality scores (Q):
GHSSBBTTTTG

$\mapsto$

	A	C	G	T
A	1	2	3	4
C		5	6	7
G			8	9
T				10

The likelihood

$P(Data|G = \{A_1, A_2\}) \propto P(X|G = \{A_1, A_2\}) = P(X|G)$
where $A \in \{A, C, G, T\}$

10 genotype likelihoods

	A	C	G	T
A	0.0	0.001	0.0	0.01
C		0.02	0.001	0.12
G			0.0	0.003
T				0.001

TCCTTTTTTTT
CCCTTTTTTTT

Base (b)

- Nr. of reads
- Base quality (Error)
- (Mapping quality)

- Pop. allele freq. as prior => Gen**otype posterior** used for down stream analyis

Korneliussen *et al.* 2014; BMC Bioinformatics

ANGSD (Genotype likelihoods)

- ANGSD

Table 1 Overview of analyses implemented in ANGSD

Analysis	Basis	Reference
Contamination estimates based on the X-chromosomes	BC	[19][b]
Type specific error estimation estimated by simultaneously estimating allele frequencies and genotype likelihoods	GL	[10]
Type specific error estimation based on an outgroup and a high quality genome	BC	[20][ab]
Genotype likelihoods (GL) (diploids)	BC/Seq	[6,8,10,15]
Allele frequencies for a site	BC/GL/GP	[21][b] [10]
SNP discovery (LRT) used for rejecting that the allele frequency is different from zero	GL	[10]
Genotype posteriors (GP) can be used for calling genotypes by specifying a cutoff	GL/SAF	[9,10]
Sample allele frequencies (SAF) the probability of all read data given the sample allele frequency	GL/GP	[9][b]
Population differentiation statistics F_{st}	SAF	[14][ac]
Population structure via principle components analysis **(PCA)**	GP	[14][ac]
Admixture analysis (NGSadmix) NGS data	GL	[22][ab]
Detection of ancient admixture **ABBA-BABA/d-statistics**	BC	[20][b]
Estimation of **SFS (1D)**	SAF	[9][ab]
Estimation of **SFS (2D)**	SAF	
Selection scans, Neutrality tests (e.g θ's and Tajima's D)	SAF	[12][ab]
Estimation of individual and site-wise **Inbreeding** coefficients. Also MAF and GP estimation for inbreed individuals	GL	[13][abc]
Allele frequency based association for case/control data)	GL	[10]
Association score test in a generalized linear model framework for both quantitative and case/control data while allowing for additional covariates	GL-GP	[11][b]

Table of the supported analyses in ANGSD. [a]indicates methods that require a secondary program in ANGSD package. [b]indicates methods for which ANGSD is the *de facto* implementation and [c]are user supplied extensions for ANGSD. The basis for each analysis is either the sequencing data (Seq), base counts (BC), genotype likelihood (GL), sample allele frequencies (SAF) or genotype probabilities (GP).

Korneliussen *et al.* 2014; BMC Bioinformatics

The genomic signature of selection

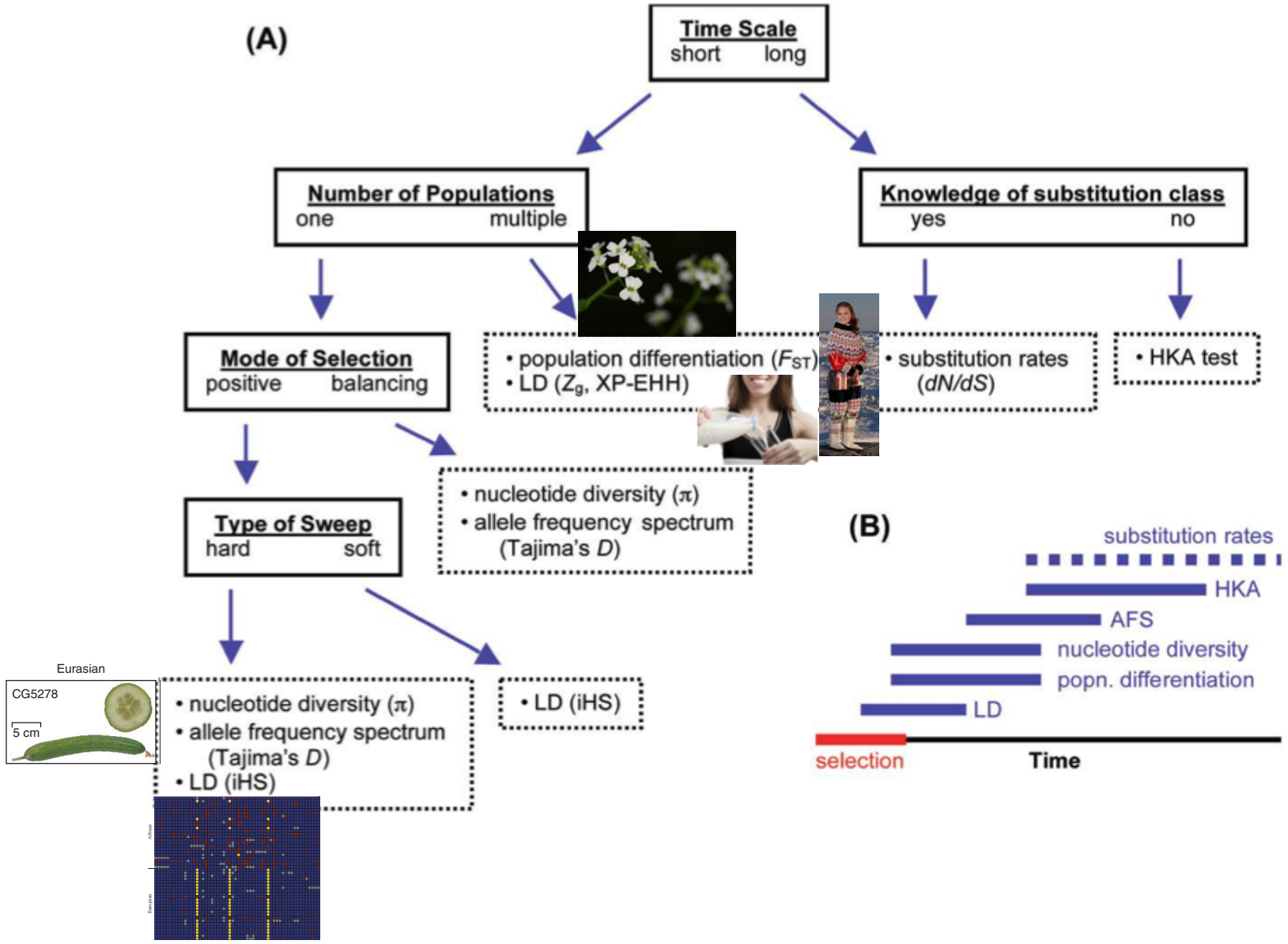
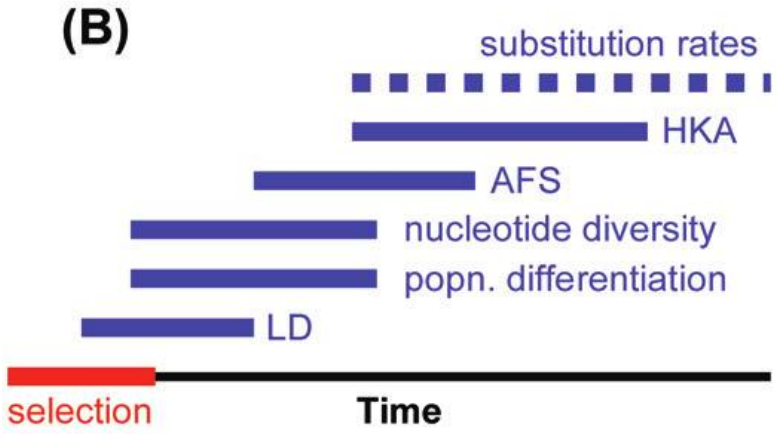


Figure 1 of Hohenlohe et al. 2010

Different time scale of selection tests

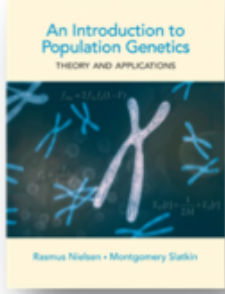
Example in humans

✧ Substitution rates (d*N*/d*S*)	>1,000,000 years	~ >50,000 generations
✧ Skew of allele frequency spectra	<200,000 years	~ <10,000 generations
✧ Reduced level of genetic variation (π)	<200,000 years	~ <10,000 generations
✧ Population differentiation (F_{ST}-outlier)	<80,000 years	~ <4,000 generations
✧ Linkage disequilibrium (*LD*)	<30,000 years	~ <1,500 generations



Oleksyk et al. 2010; Hoehenlohe et al. 2010

Further Reading



Nielsen R & Slatkin M (2013) An Introduction to Population Genetics: Theory and applications. *Sinauer*

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics* **14**, 262-274.

Fischer MC, Rellstab C, Leuzinger M*, et al.* (2017) Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics* **18**, 69.

Fischer MC, Foll M, Heckel G, Excoffier L (2014) Continental-scale footprint of balancing and positive selection in a small rodent (*Microtus arvalis*). *PLoS One* **9**, e112332.

Fischer MC, Rellstab C, Tedder A, et al. (2013) Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Molecular Ecology* **22**, 5594-5607.

Fischer MC, Foll M, Excoffier L, Heckel G (2011) Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). *Molecular Ecology* **20**, 1450-1462.

Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One* **8**, e79667.

Fumagalli M, Moltke I, Grarup N, et al. (2015) Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343-1347.

Hohenlohe PA, Phillips PC, Cresko WA (2010) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences* **171**, 1059-1071.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 1-13.

Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 185-205.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**, 671-688.

Rellstab C, Fischer MC, Zoller S*, et al.* (2017) Local adaptation (mostly) remains local: reassessing environmental associations of climate-related candidate SNPs in Arabidopsis halleri. *Heredity* **118**, 193-201.

Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC (2013) Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS One* **8**, e80422.

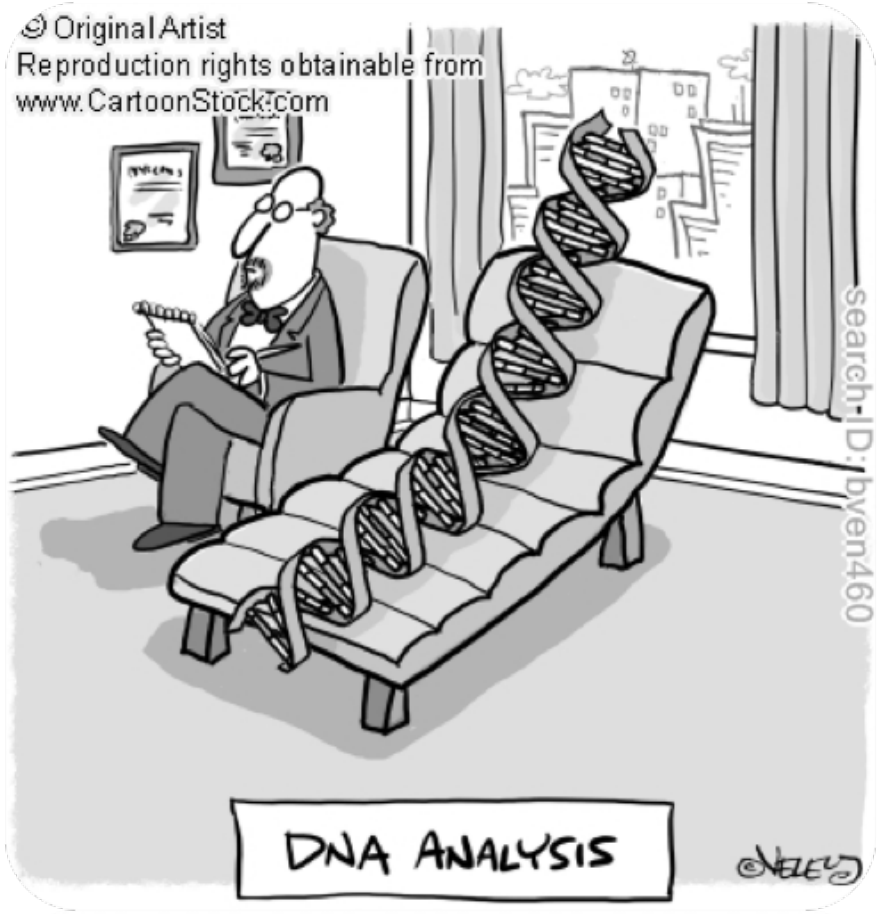
Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology* **24**, 4348-4370.

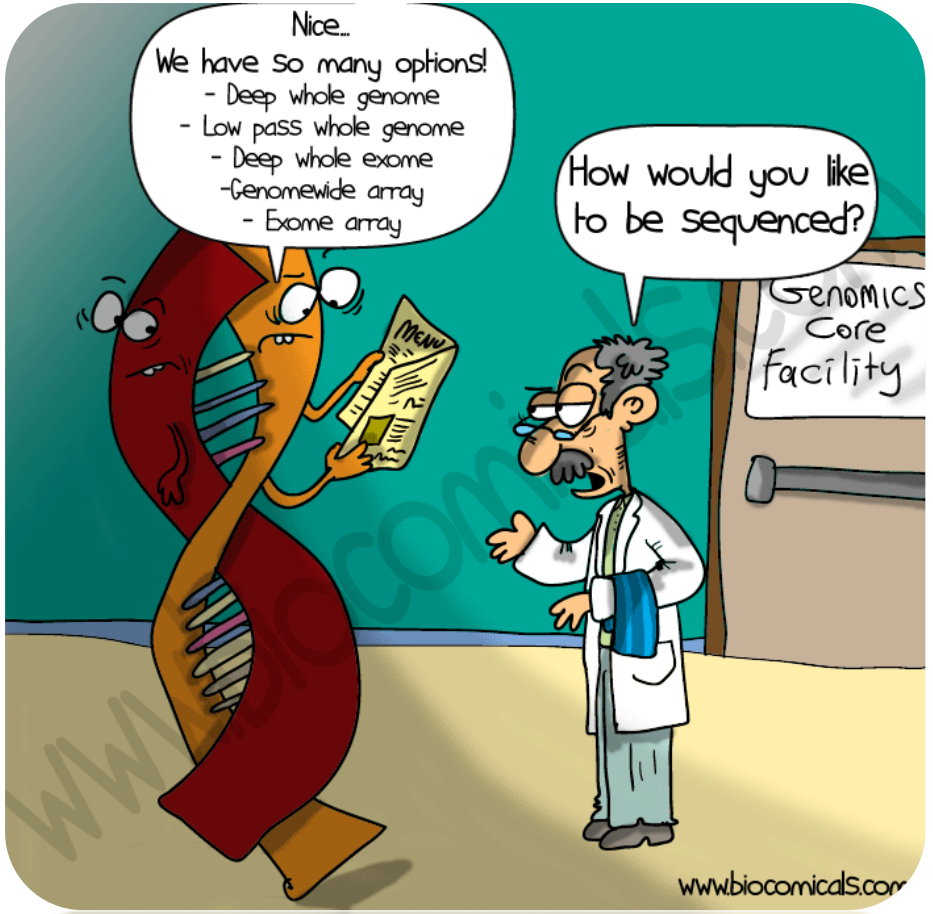
Yi X, Liang Y, Huerta-Sanchez E, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78.

Winter School: **Bioinformatics for Adaption Genomics (B@G)**

Bioinformatics resources: http://www.adaptation.ethz.ch/education/bag-winter-school-2018/teaching-resources.html

Thanks!


Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
search ID: byen460
DNA ANALYSIS


Nice...
We have so many options!
- Deep whole genome
- Low pass whole genome
- Deep whole exome
-Genomewide array
- Exome array
How would you like to be sequenced?
Menu
Genomics Core Facility
www.biocomicals.com

Box 1

Critical Population Genetic Concepts and Statistical Measures Used to Detect Selection in Population Genomics

Allele frequency spectrum (AFS): The distribution of frequencies across alleles in a sample. Tests based on AFS using DNA sequence data rely on a few related statistics, all of which are comparisons between estimates of the population genetic parameter $\theta = 4N\mu$. The statistics are calculated as the difference of two such estimates, normalized by the expected variance of the difference under a neutral model, so that values below −2 or greater than 2 roughly exceed the 95% confidence limits about the neutral expectation of 0. However, the actual mean may frequently deviate from 0 (Thornton 2005; Wares 2010). Simonsen et al. (1995) compared three measures and found Tajima's D to have the most statistical power:

Tajima's D: Normalized difference between π and S, the number of segregating sites (Tajima 1989).

Fu and Li's D^*: Normalized difference between S and the number of singletons η (alleles observed only once in a sample; Fu and Li 1993).

F^*: Normalized difference between π and η (Fu and Li 1993).

Background selection: Ongoing selection against deleterious mutations that can result in the loss of linked neutral variation (Charlesworth et al. 1993).

Balancing selection: Here we define balancing selection broadly as the class of selective forces that maintain polymorphism over time. This can include, for example, frequency-dependent selection or heterozygote advantage (Charlesworth 2006).

Coalescent theory: A theoretical framework for understanding genetic variation based on the retrospective pattern of shared ancestry among alleles in a sample (Wakeley 2009).

Divergent selection: Positive selection acting differentially between separate populations.

dN*/*dS: Ratio of nonsynonymous (amino acid–changing) to synonymous substitutions in a nucleotide sequence. Testing for selection based on this ratio typically uses aligned sequence data among populations or taxa and can detect selection over long timescales, although it requires multiple amino acid substitutions (i.e., recurrent selective sweeps).

F_{ST}: A statistic describing the partitioning of allelic variance within versus among populations; F_{ST} ranges from 0 (no population differentiation) to 1 (complete population differentiation). There are multiple ways of calculating F_{ST} that can occasionally have substantial effects on its value but rarely its relative magnitude among loci (Charlesworth 1998; Holsinger and Weir 2009). Commonly used population genomic tests for selection based on identifying outliers in F_{ST} are as follows:

LOSITAN (Antao et al. 2008) computer software implements the method of Beaumont and Nichols (1996) to identify F_{ST} outliers based on heterozygosity, which affects the predicted neutral distribution of F_{ST}.

ARLEQUIN (Excoffier et al. 2009) software performs the same analysis, accounting for hierarchical population structure.

BAYESFST (Beaumont and Balding 2004) assesses the significance of a locus-specific parameter that indicates selection in a model of F_{ST}.

BAYESCAN (Foll and Gaggiotti 2008) modifies the approach of Beaumont and Balding (2004) to estimate the posterior probability of a locus being subject to selection.

DETSEL (Vitalis et al. 2003) uses coalescent simulations in a simple two-population model to identify F_{ST} outliers.

Genetic draft: The loss of genetic diversity and changes in AFS at loci linked to a selected locus during a selective sweep (Gillespie 2000).

HKA test: A test of the neutral prediction for the relationship between within-population diversity and among-population divergence (Hudson et al. 1987).

Linkage disequilibrium (LD): The correlation between alleles across loci. Traditionally, LD has been calculated as a function of a pair of loci, regardless of their physical position (Slatkin 2008). This aspect of LD can be partitioned among populations in the statistic Z_g as a test of selection (Storz and Kelly 2008). Genome scans for selection also apply several of the following statistics that describe the decay of LD as a function of physical distance, also known as haplotype structure:

Extended haplotype homozygosity (EHH) measures the probability that any two randomly chosen haplotypes are identical over a given distance from a focal site (Sabeti et al. 2002).

Integrated haplotype score (iHS) integrates the area under the EHH curve (Voight et al. 2006). Huff et al. (2010) found this measure to have greater statistical power and to be more robust to complex demographics than two related alternatives.

Cross-population extended haplotype homozygosity (XP-EHH) compares EHH between two populations to test for interpopulation differences in the extent of LD (Sabeti et al. 2007).

π: A measure of nucleotide diversity, calculated as the proportion of pairwise differences in a sample; π can be estimated either within or between populations and is directly used in some calculations of F_{ST} (Charlesworth 1998).

Positive (directional) selection: Selection in which one or a class of alleles is favored.

Selective sweeps: The increase in frequency of one or a class of alleles favored by selection. **Hard sweeps** result from selection on a single allele, typically a new mutation that is favored immediately on its appearance in a population. **Soft sweeps** are selection on standing genetic variation or on variants supplied by recurrent mutation or migration during the selective phase, so that a number of different alleles are collectively favored and increase in frequency. These alleles are typically considered to be neutral or even deleterious before a shift in selective regime (Hermisson and Pennings 2005).