ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

The genomic signature of selection

Martin C. Fischer
Institute of Integrative Biology
ETH Zürich
June 19th 2015

Diversity and adaptation

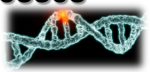
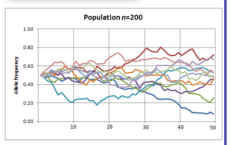
Why is there so much variation among and within species?

How are organisms adapted to their environment?

Genetic diversity

- Evolutionary changes at the molecular level are caused by…

Neutral processes

- Mutations 
- Genetic drift 

- Population history 

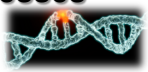
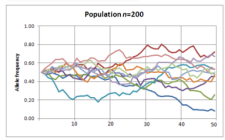
Mutations

- Small scale:
- Large scale:

Genetic diversity

- Evolutionary changes at the molecular level are caused by…

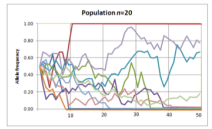
Neutral processes

- 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

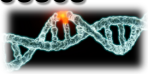


Genetic diversity

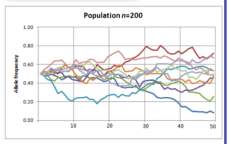
- Evolutionary changes at the molecular level are caused by…

Neutral processes

- Mutations



- Genetic drift

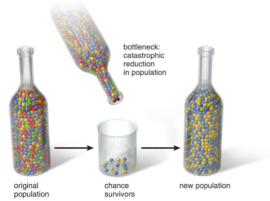


- Population history

Population history

- Demography

- Bottleneck

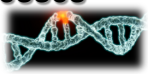


Genetic diversity

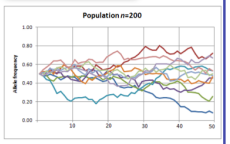
- Evolutionary changes at the molecular level are caused by…

Neutral processes

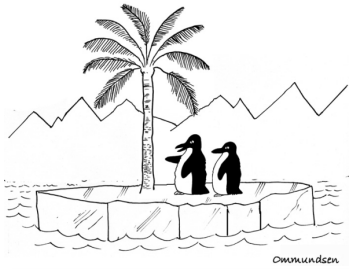
- Mutations



- Genetic drift



- Population history

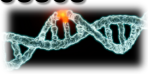


Genetic diversity

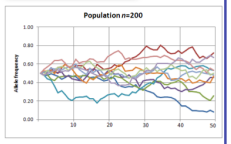
- Evolutionary changes at the molecular level are caused by…

Neutral processes

- Mutations

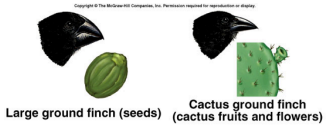


- Genetic drift

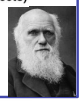


- Population history

Natural selection



Darwin's Finches



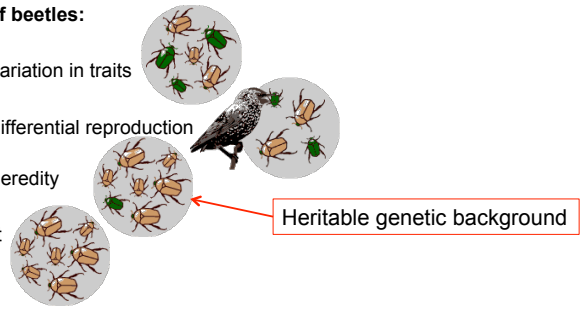
Natural selection

Natural selection

Natural selection is one of the basic mechanisms of evolution, along with *mutation*, *migration*, and *genetic drift*.

Population of beetles:

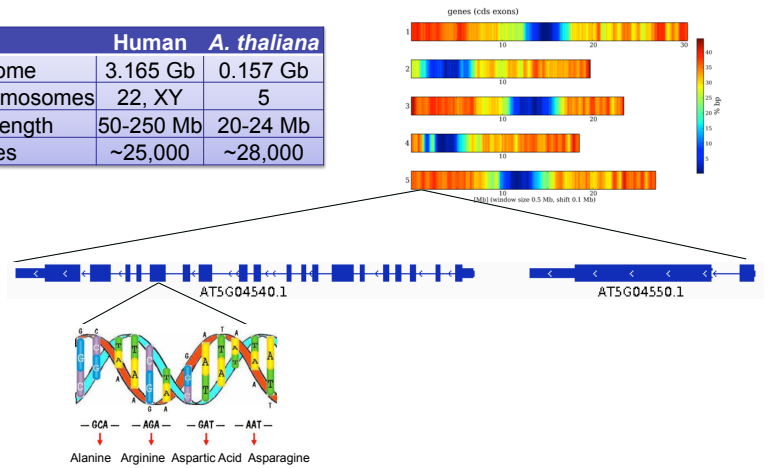
1. There is variation in traits
2. There is differential reproduction
3. There is heredity
4. End result



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

The genome

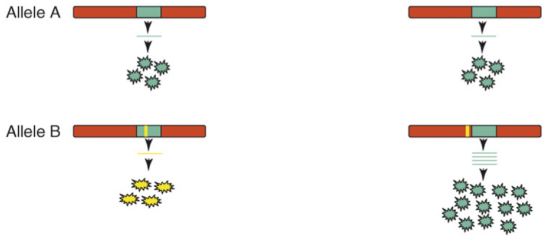
	Human	*A. thaliana*
Genome	3.165 Gb	0.157 Gb
Chromosomes	22, XY	5
Chr length	50-250 Mb	20-24 Mb
Genes	~25,000	~28,000



Causes of adaptive changes

Functional mutation

Gene expression changes



Morgante *et al.* 2007

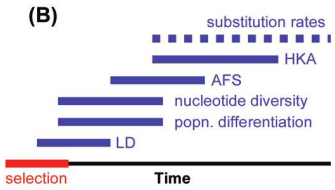
The genomic signature of selection

Different methods to detect selection:

- Reduced level of genetic variation (e.g. π)
- Linkage disequilibrium (*LD*)
- Skew of allele frequency spectra (Tajimas's *D*)
- Population differentiation (F_{ST} outlier approach)
- Substitution rates (e.g. d*N*/d*S*)

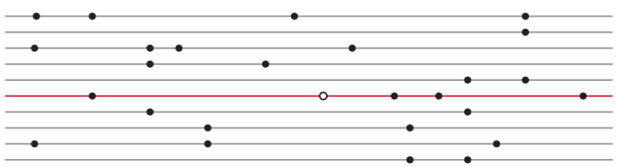
Comparison:

- Within population
- Among populations
- Among species



see Box 1 Hohenlohe *et al.* 2010

Selective sweeps

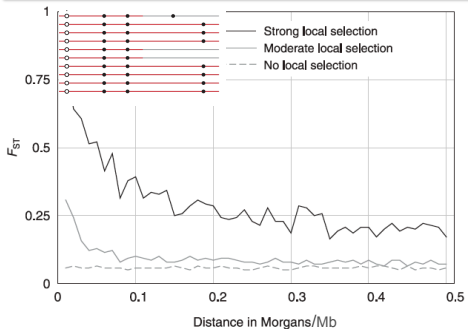


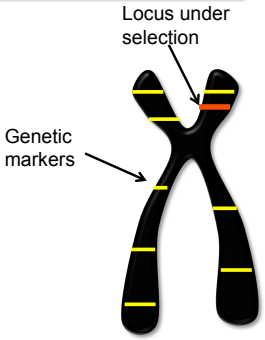
Selective sweep:
An allele that increases fitness arises and 'sweeps' to fixation in a population

- Recombination
- Linkage disequilibrium
- Hitchhiking

Figure 1 of Storz 2005, *Molecular Ecology*

Genetic hitchhiking

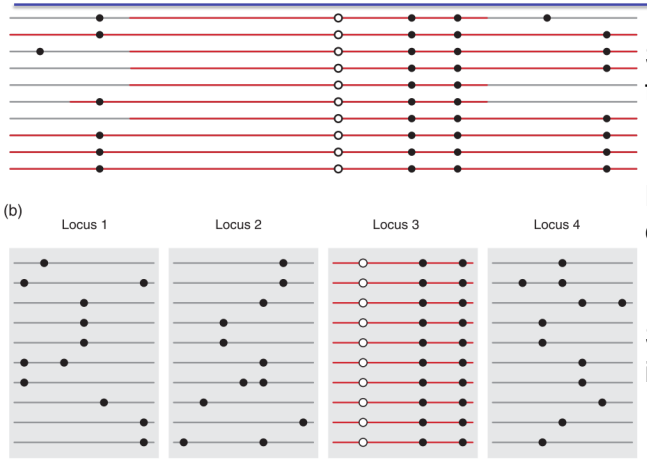




- Neutral loci close to a locus under selection gets "hijacked" and increase its allele frequency
- The signature of selection decays with increasing distance from the locus under selection => **genetic hitchhiking**

Figure 3 of Storz 2005

Selective sweeps



Skewed allele frequency

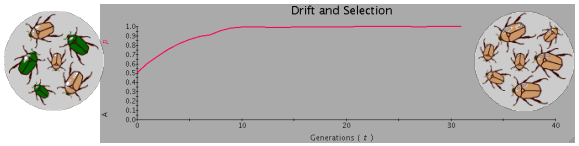
Reduced diversity (π)

Single locus is affected

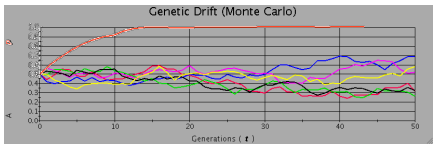
Figure 1 of Storz 2005, *Molecular Ecology*

Population size matters….

- Selection



- Drift



- Selection coefficient ***s* >> (1/4N_e)**

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

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: 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 (Tajimas's *D*)
- Population differentiation (F_{ST} outlier approach)
- Substitution rates (e.g. d*N*/d*S*)

Comparison:

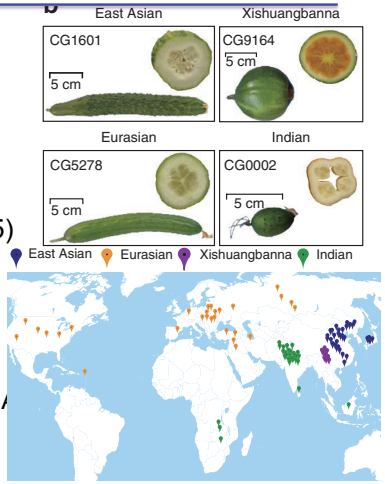
- Within population
- Among populations
- Among species

see Box 1 Hohenlohe *et al.* 2010

Reduced level of genetic variation (π)

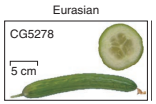
Cucumber

- 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



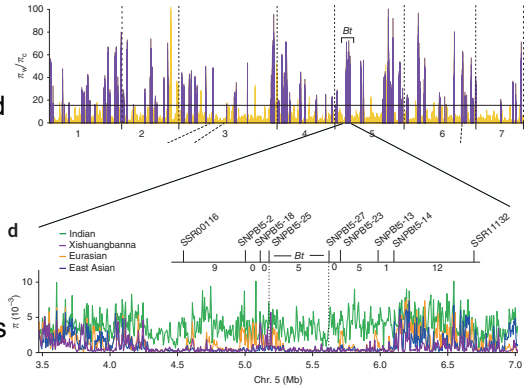
Qi *et al.* 2013

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
- 442-kb in length



Qi *et al.* 2013

Linkage disequilibrium (*LD*)

- Correlation coefficient (*r*) between pairs of loci
- Extended homozygosity

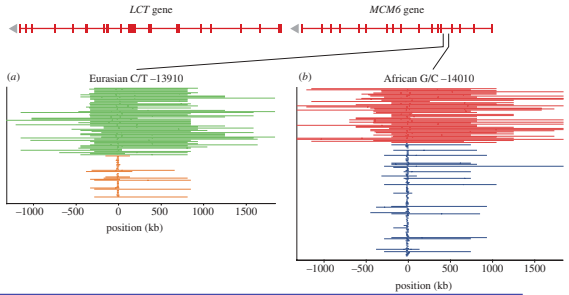
EHH: extended haplotype homozygosity

- Measures the decay of homozygosity from a 'core' SNP

LCT: lactose persistent gene in humans

2 Mb haplotypes

Independent evolution in Africa and Europe

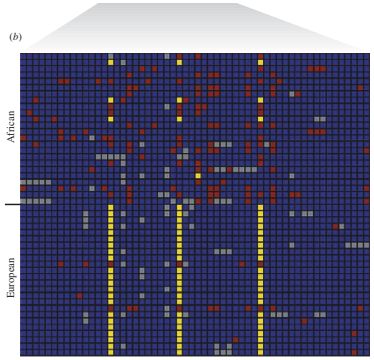


Tishkoff *et al.* (2007); Oleksyk *et al.* (2010)

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

Tajima's *D*

- Normalized difference between π and segregating sites (S, θ_W)
- $d = \boldsymbol{\pi - \theta_W}$
  $D = d/V(d)^{1/2}$
- **Positive selection**
  excess of low-frequency SNPs:
  $\pi < \theta_W$, ***-D***
- **Balancing selection**
  excess intermediate-freq. SNPs:
  $\pi > \theta_W$, ***+D***

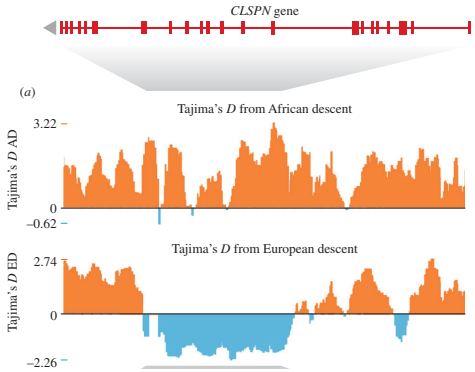


Carlson *et al.* 2005; Oleksyk *et al.* 2010

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!



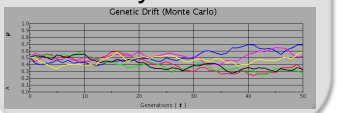
Carlson *et al.* 2005; Oleksyk *et al.* 2010

Population differentiation (F_{ST} outlier approach)

Neutral processes:

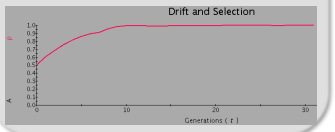
- Genetic drift
- Demographic history

➢ Affect **all loci** similarly



Selection:

➢ Affect only **single locus**



Outlier detection

- Population differentiation

Lewontin & Krakauer 1973; Beaumont & Nichols 1996

F_{ST} outlier detection

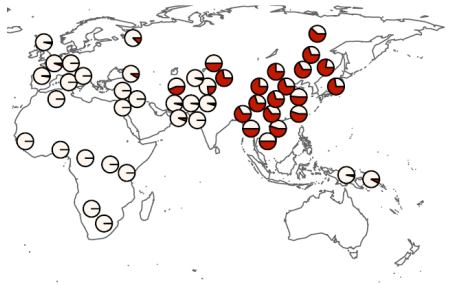
Local **positive selection**

- Increased level of differentiation among populations

➢ High F_{ST} values

- *mc1r*

➢ Adaptation to different environment



Gerstenblith *et al.* 2007; Nachman *et al.* 2003

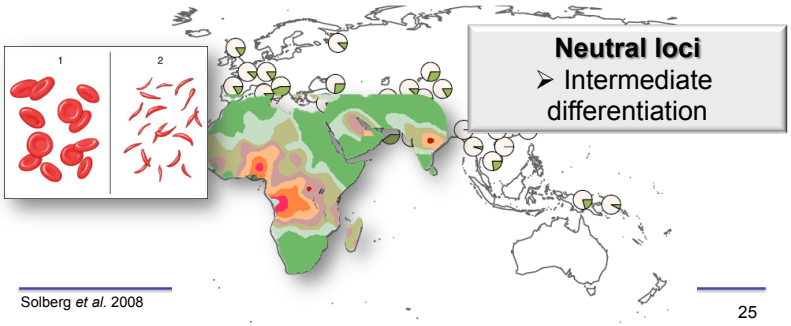
F_{ST} outlier detection

Balancing selection
- Relatively uniform frequencies across populations

➢ Low F_{ST} values

- *HLA-C* gene (*MHC*)
- Sickle cell anaemia

➢ Immune system
(Andres *et al.* 2009)

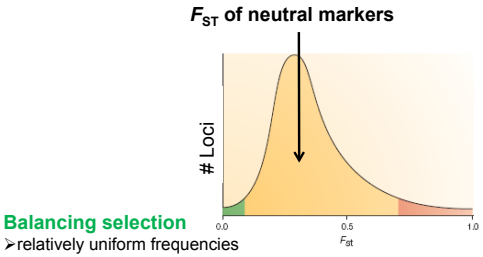


Neutral loci
➢ Intermediate differentiation

Solberg *et al.* 2008

Model free F_{ST} outlier approach

- >2 populations
- Screen many loci (>100; up to whole genome)
- Outliers: e.g. 95% confidence interval of F_{ST} distribution



Balancing selection
➢relatively uniform frequencies across populations
⇒ low F_{ST} values

Positive selection
➢increased level of differentiation among populations
⇒ high F_{ST} values

Luikart *et al.* 2003

NGS case study

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

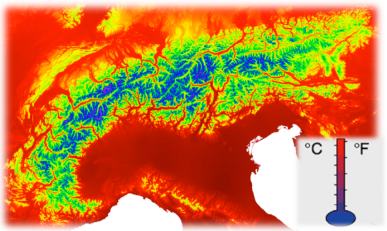
Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps

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

Alpine environments

- Highly heterogeneous
- Strong environmental gradients

➢ Genetic basis of adaptation



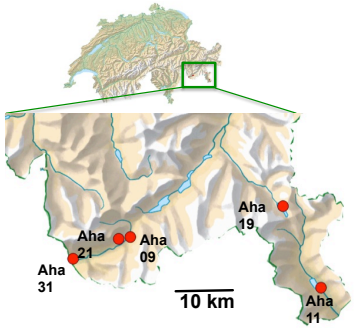
Study organism

Arabidopsis halleri

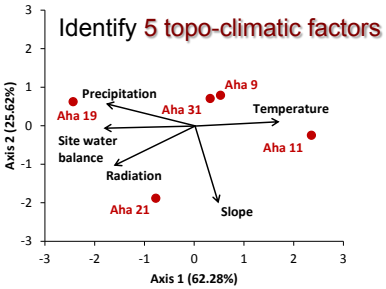
- Close relative of the model organism *A. thaliana*
- Genome size 255 Mbp
- Strictly outcrossing
- 300 – 2400 m a.s.l.

Heterogeneous Alpine environments

- 5 populations in close vicinity (2 – 45 km)
- 20 ind. each
- Cover wide range of abiotic environmental conditions
  - 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

Accuracy of Pool-Seq approach

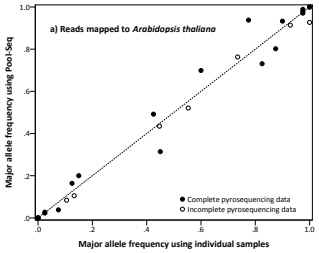
OPEN ACCESS Freely available online PLOS ONE

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]*

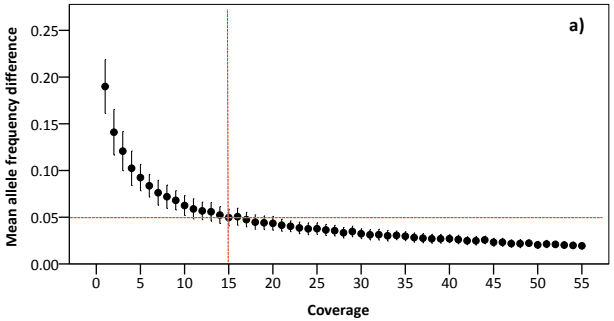
1 Biodiversity and Conservation Biology, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland, **2** Genetic Diversity Centre, ETH Zürich, Zürich, Switzerland, **3** Institute of Evolutionary Biology and Environmental Studies and Institute of Plant Biology, University of Zürich, Zürich, Switzerland,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

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



Accuracy of Pool-Seq approach

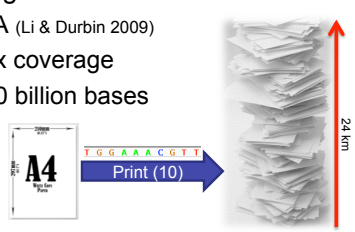
- Influence of the coverage for Pool-Seq



Rellstab *et al. 2013*

Population genomics

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

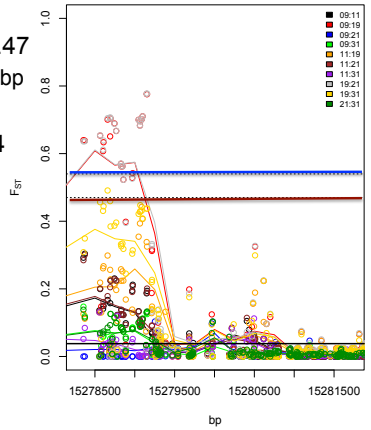


- PoPoolation2 (Kofler *et al.* 2011)
  - Identified **2,091,957 SNPs**
  - 25,764 genes covered
  - Genome wide pairwise F_{ST}
    - Mean: 0.038

F_{ST} outlier detection

3 step model free approach:

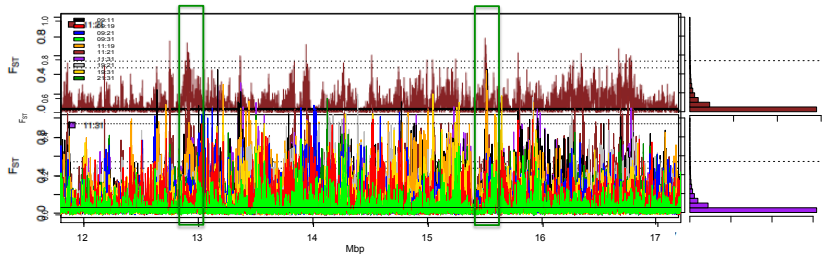
- 0.1% of highest ***F_{ST}* regions**: $F_{ST} > 0.47$
  - F_{ST} sliding windows approach of 500 bp
- 0.1% of highest ***F_{ST}* SNPs**: $F_{ST} > 0.54$
  - Corrected for population structure
- Fisher's exact test
  - $p < 2.39*10^{-9}$
  - Strong allele frequency differences
  - Corrects for coverage



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

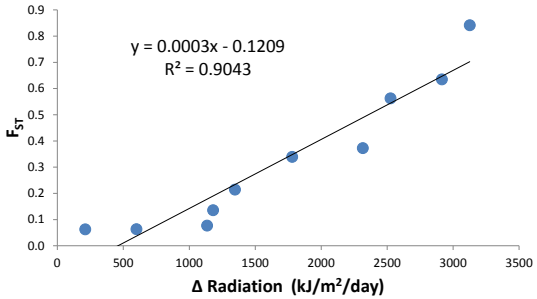
Gene Ontology Terms (Biological process)
Defense response to bacterium
Aging
Cell surface receptor signaling pathway
Seed coat development
Cofactor catabolic process
Chromatin modification
Recognition of pollen
Gibberellic acid mediated signaling pathway
DNA methylation
Chromatin silencing
Regulation of root development
Protein N-linked glycosylation

Genes with a signature of selection are involved in diversity of biological processes

Environmental Associations

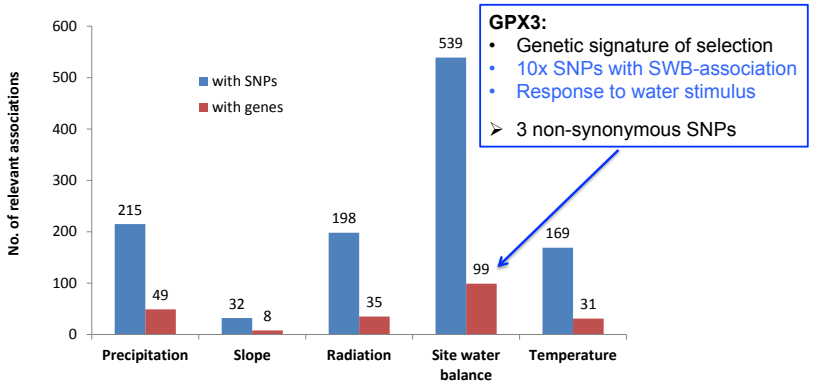
Partial Mantel tests

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



Abiotic environmental associations

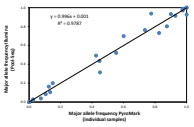
- 175 genes with environmental associations
  - 41 genes (23%) of unknown gene function



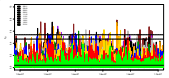
Fischer *et al.* 2013 MolEcol

Conclusions

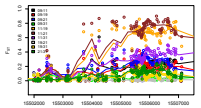
- Population genomic analyses of non-model species using a Pool-Seq approach are:
  - Feasible and provide accurate SNP estimates
  - Reveal evidence for selection across the genome



- The molecular signature of adaptation in Alpine populations of *A. halleri* is highly complex



- Many genes showing interesting patterns of variation have unknown functions – need for ecological gene annotation



Model based F_{ST} outlier approach

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

- **BayeScan**

➢ Bayesian model comparison

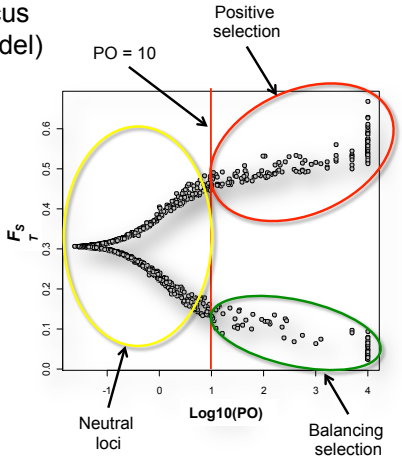
$$PO = \frac{\Pr(M_1|Data)}{\Pr(M_2|Data)}$$

Model with selection

Neutral model

Posterior odds

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



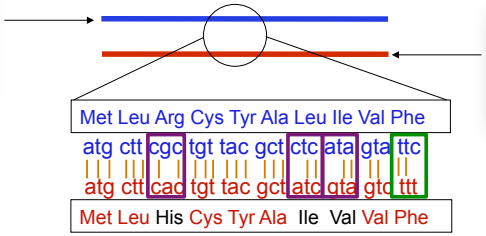
Foll & Gaggiotti 2008; Fischer *et al*. 2011

Substitution rates (e.g. d*N*/d*S*)

13,731 orthologous genes compared

Human

Chimpanzee



Non-synonymous substitution
Fitness effects

Synonymous substitution
No fitness effects

d*N* = number of non-synonymous substitutions / number of non-synonymous sites

d*S* = number of synonymous substitutions / number of synonymous sites

Nielsen et al. 2005

Substitution rates (e.g. d*N*/d*S*)

- Universal genetic code

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Substitution rates (e.g. d*N*/d*S*)

13,731 orthologous genes compared

Human

Chimpanzee

- d*N*/d*S* >1 => Positive selection => 733 genes
- d*N*/d*S* <1 => Purifying selection

- Sensory perception or **immune defenses genes** detected.

- Selection over long time scale
- Requires multiple amino acid substitutions

Nielsen et al. 2005

The genomic signature of selection

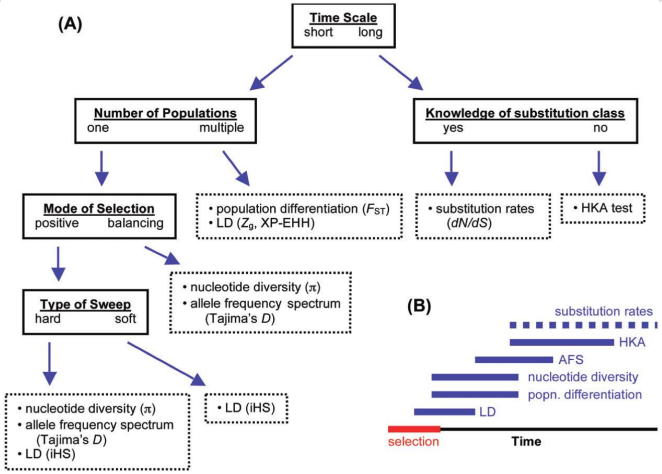
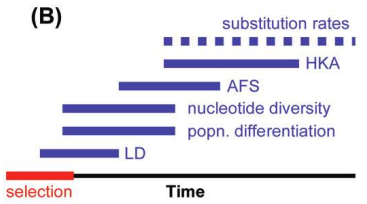


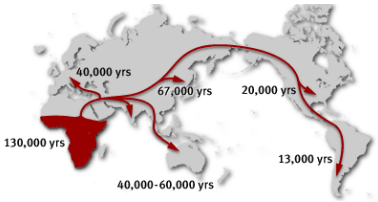
Figure 1 of Hohenlohe *et al.* 2010

Different time scale of selection tests

Example in humans

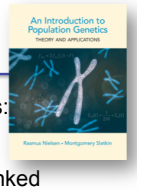
• Substitution rates (d*N*/d*S*)	>1'000'000 years
• Skew of allele frequency spectra	<200'000 years
• Reduced level of genetic variation (π)	<200'000 years
• Population differentiation (F_{ST} outlier)	<80'000 years
• Linkage disequilibrium (*LD*)	<30'000 years





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.

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.

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.


© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
search ID: dyen460
DNA ANALYSIS

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).