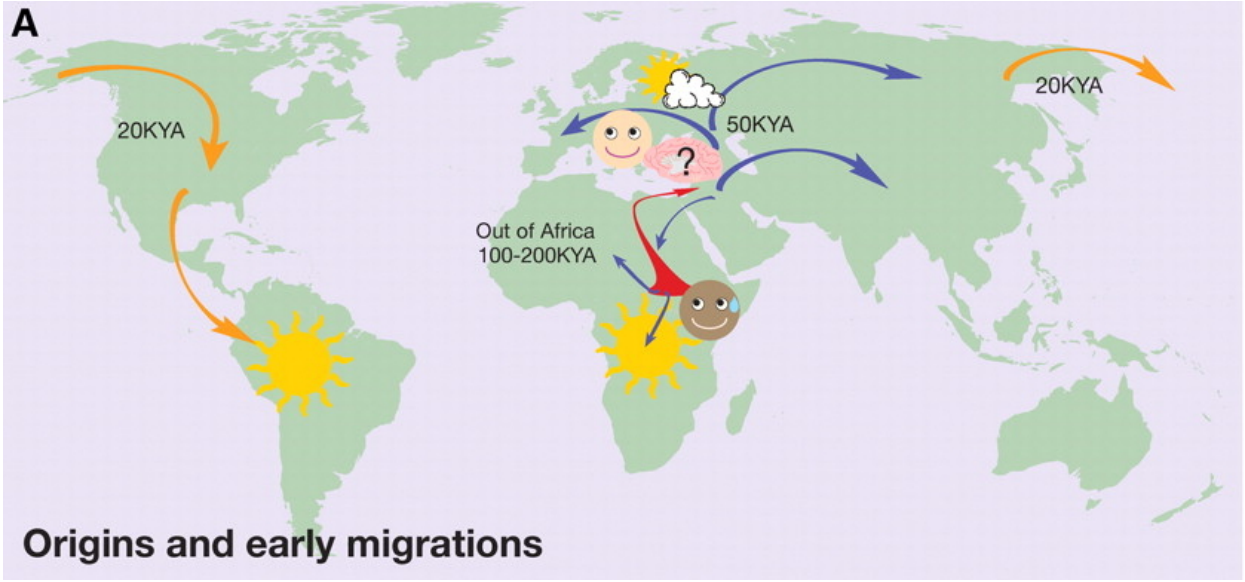
Evolutionary genomics
Data analysis module - Day 4

Detecting signatures of natural selection

April 16th 2015

Evolution and adaptation



- After the dispersal out-of-Africa, colonisation of a wide range of environments

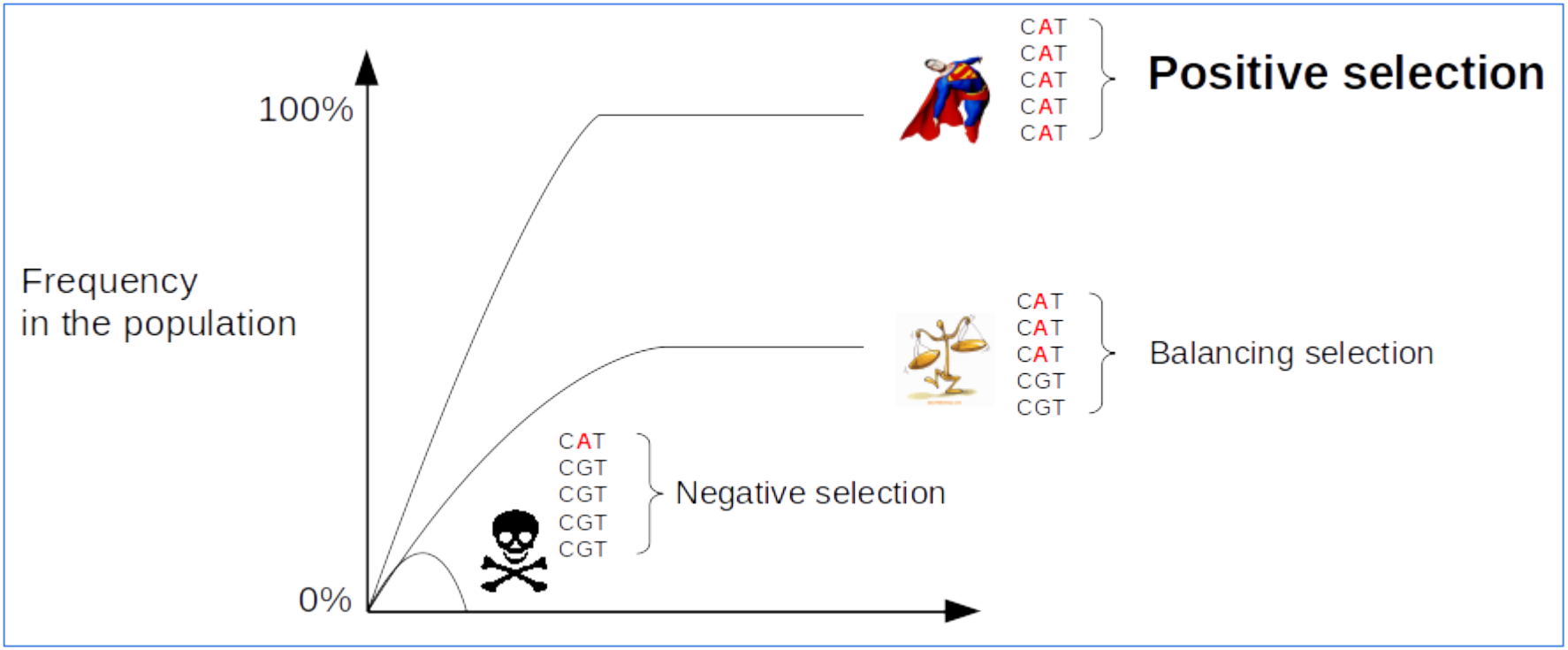
- Genetic adaptation to:
- Climate (e.g. UV light exposure)
- Diet (e.g. milk consumption)
- Pathogens

Balaresque et al. Hum Mol Genet. 2007

Natural selection

Heritable traits that increase the fitness of the become more common

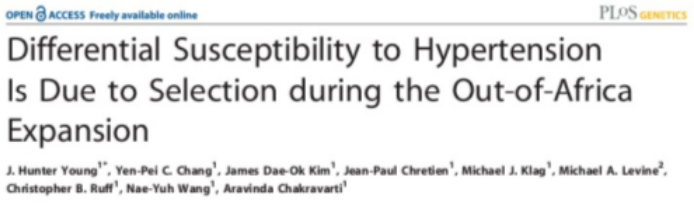
1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier



2) Sites targeted by natural selection are likely to harbour important functionality

Relevance to human health

- The presence of a **functional** variant is the prerequisite for selection.
- These variants might have played some adaptive role and have a role in human **disease**/condition.
- Recent studies have identified genes related to skin pigmentation, metabolic traits or immune defense targeted by natural selection in the human genome.


OPEN ACCESS Freely available online PLoS GENETICS

Differential Susceptibility to Hypertension Is Due to Selection during the Out-of-Africa Expansion

J. Hunter Young[1*], Yen-Pei C. Chang[1], James Dae-Ok Kim[1], Jean-Paul Chretien[1], Michael J. Klag[1], Michael A. Levine[2], Christopher B. Ruff[1], Nae-Yuh Wang[1], Aravinda Chakravarti[1]

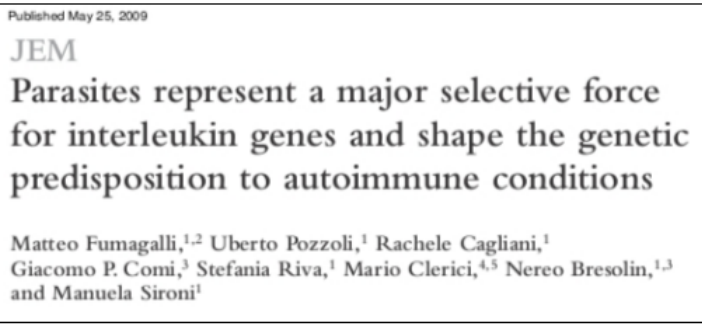
Hypertension: variants that allowed water and sodium retention and increased vascular reactivity now cause hypertension.

Obesity and insulin resistance: thrifty variants became unfavorable with a shift in diet.


OPEN ACCESS Freely available online PLoS GENETICS

Adaptations to Climate in Candidate Genes for Common Metabolic Disorders

Angela M. Hancock[1], David B. Witonsky[1], Adam S. Gordon[1], Gidon Eshel[2¤], Jonathan K. Pritchard[1], Graham Coop[1], Anna Di Rienzo[1*]


Published May 25, 2009

JEM

Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions

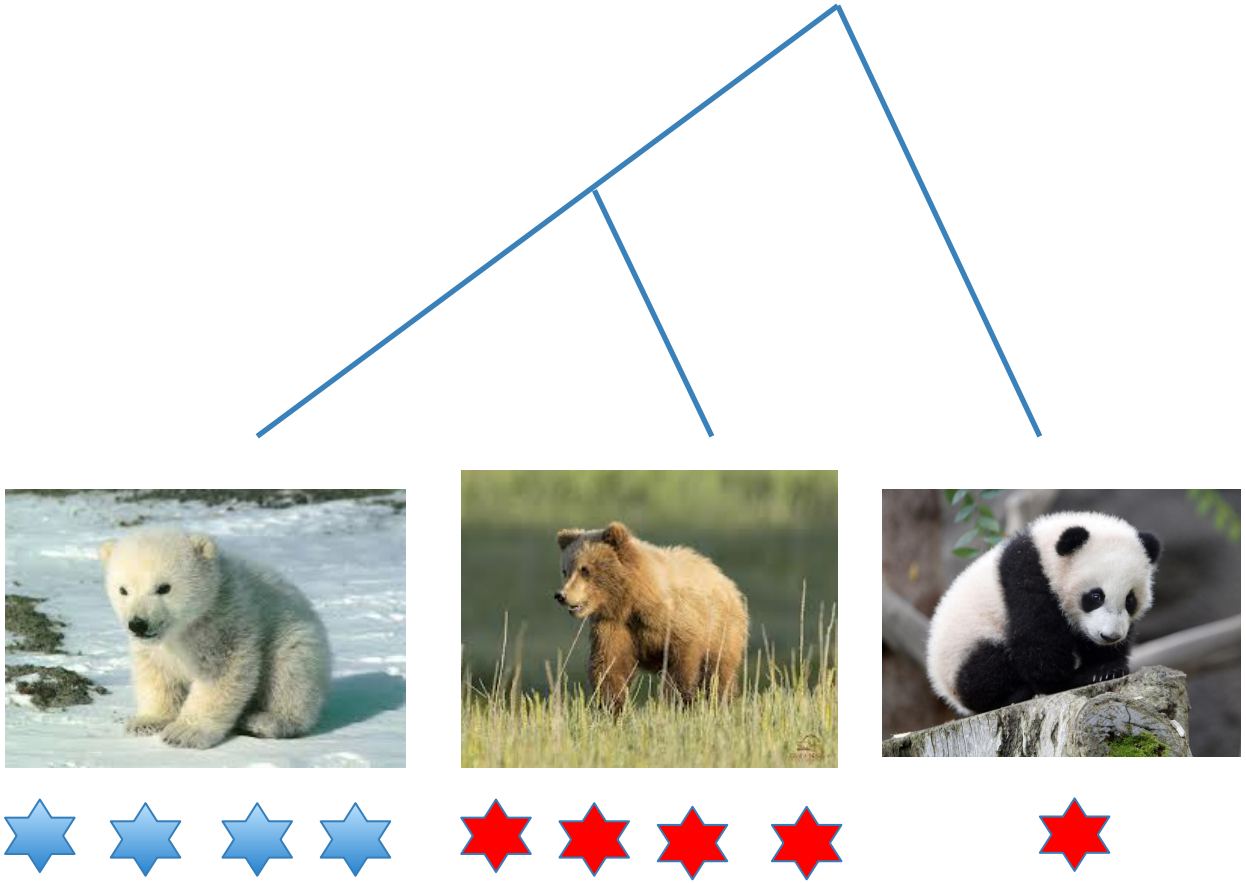
Matteo Fumagalli,[1,2] Uberto Pozzoli,[1] Rachele Cagliani,[1] Giacomo P. Comi,[3] Stefania Riva,[1] Mario Clerici,[4,5] Nereo Bresolin,[1,3] and Manuela Sironi[1]

Hygiene hypothesis: lack of exposure to pathogens in early life determines immune imbalances and predisposes to atopic and autoimmune diseases.

Methods to infer selection (I)

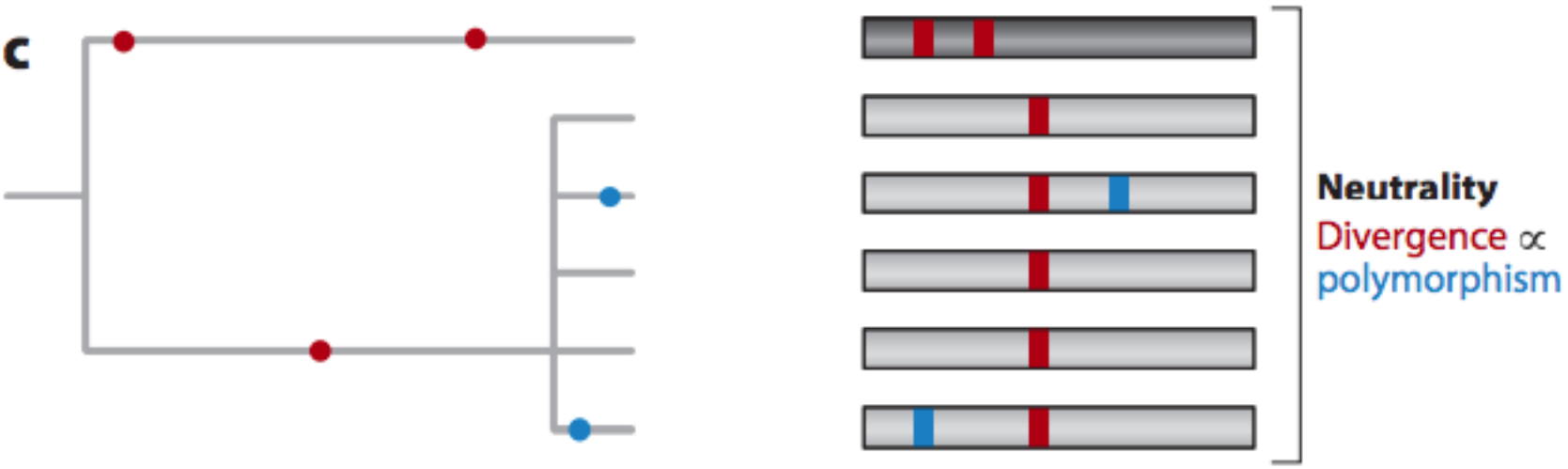
- between-species:

  events in the deep past, macro-evolutionary trends, selection between species



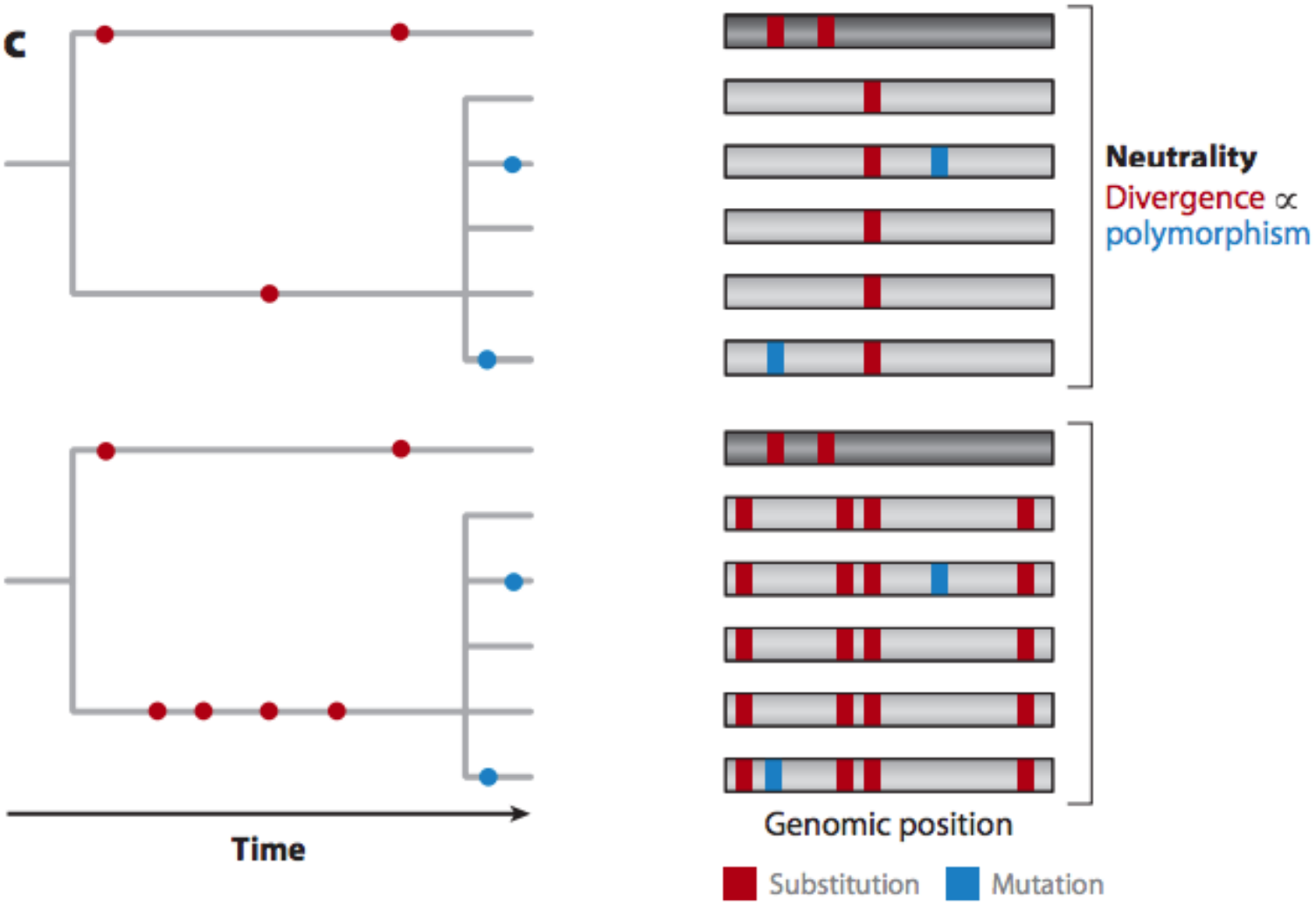
HKA test

Polymorphism levels depends on local mutation rates: under neutral evolution, the amount **of within- and between-species diversity** is expected to be similar across all loci in the genome.



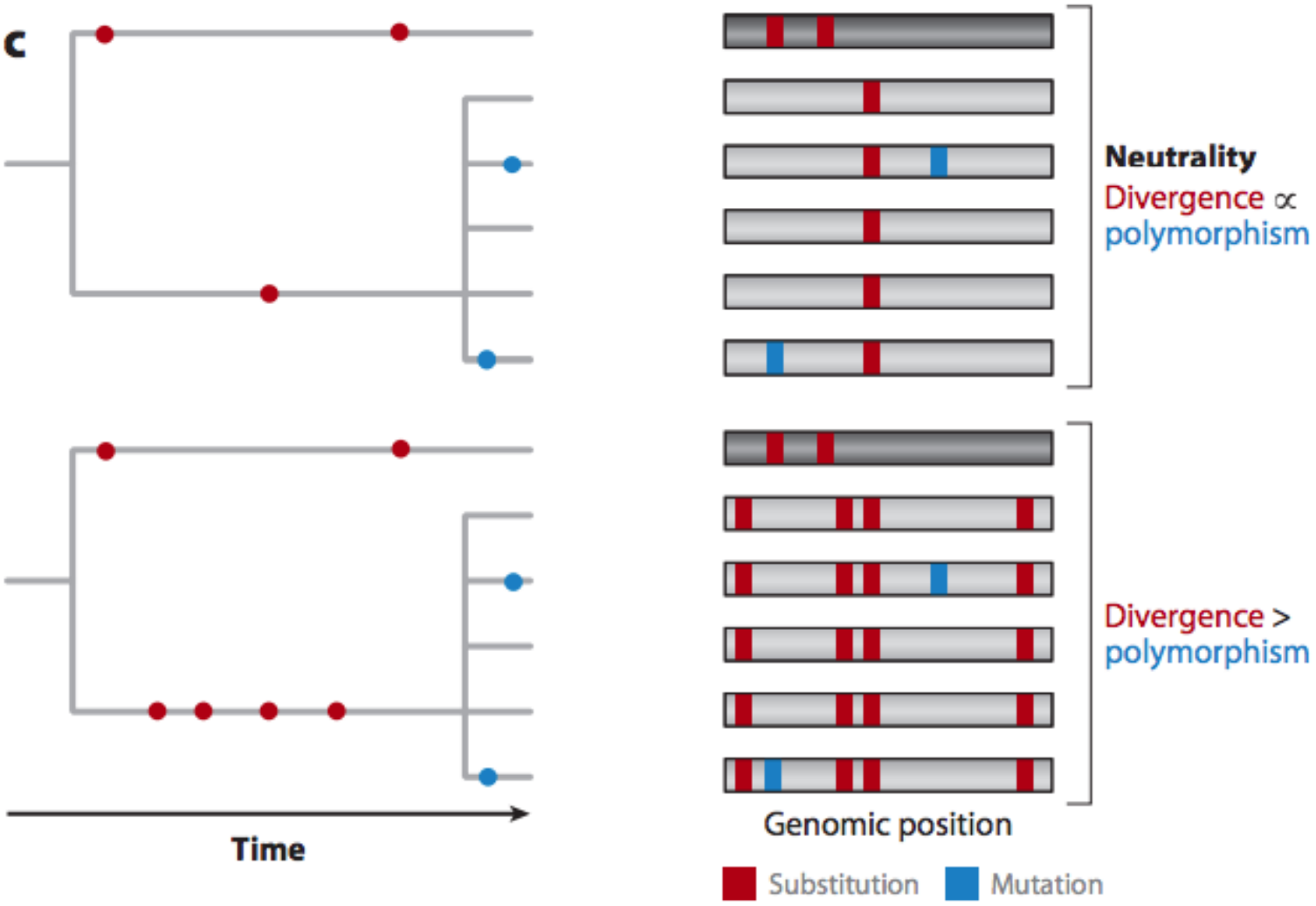
Vitti et al. Annu Rev Genet. 2013

HKA test



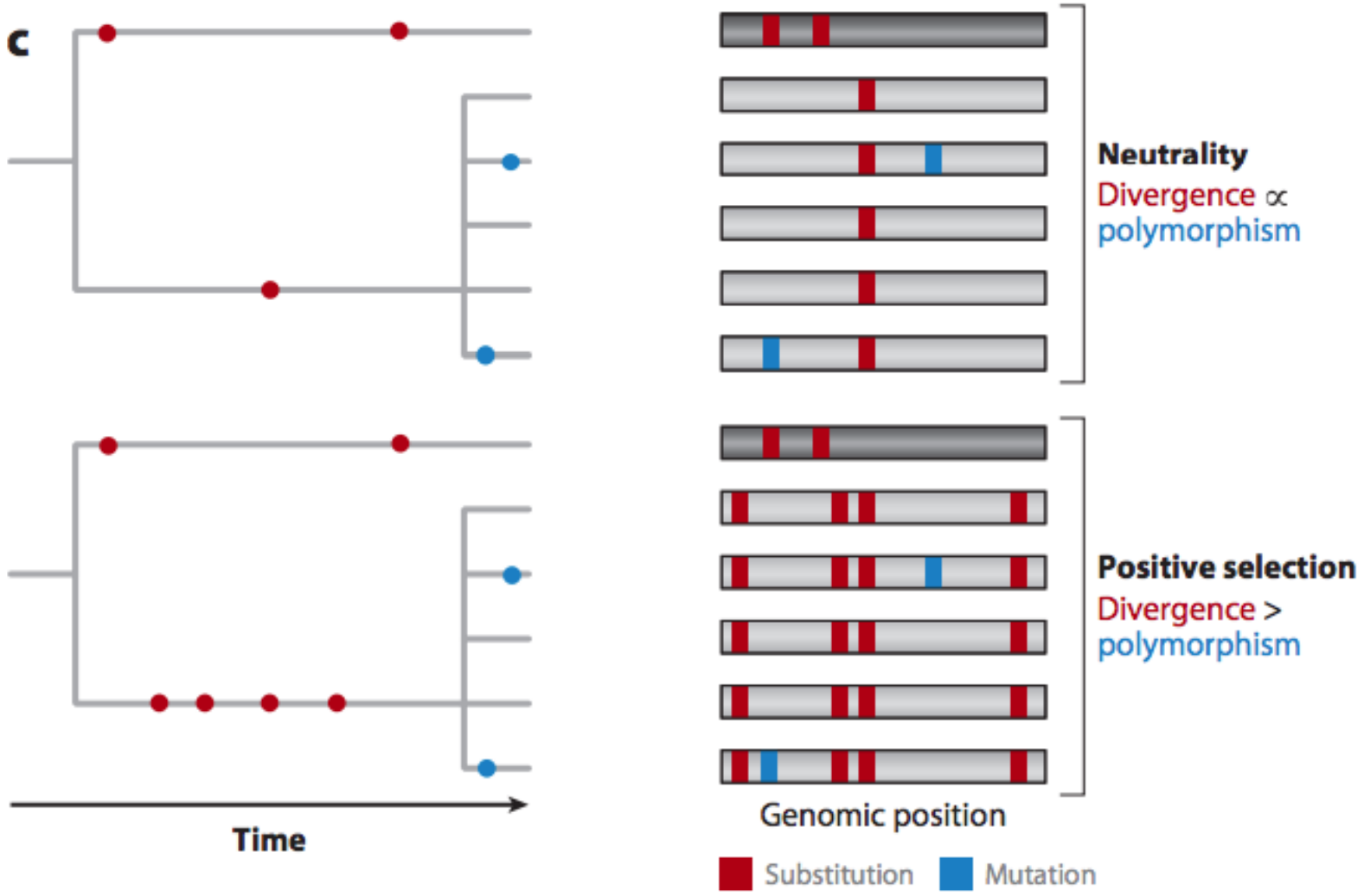
Vitti et al. Annu Rev Genet. 2013

HKA test



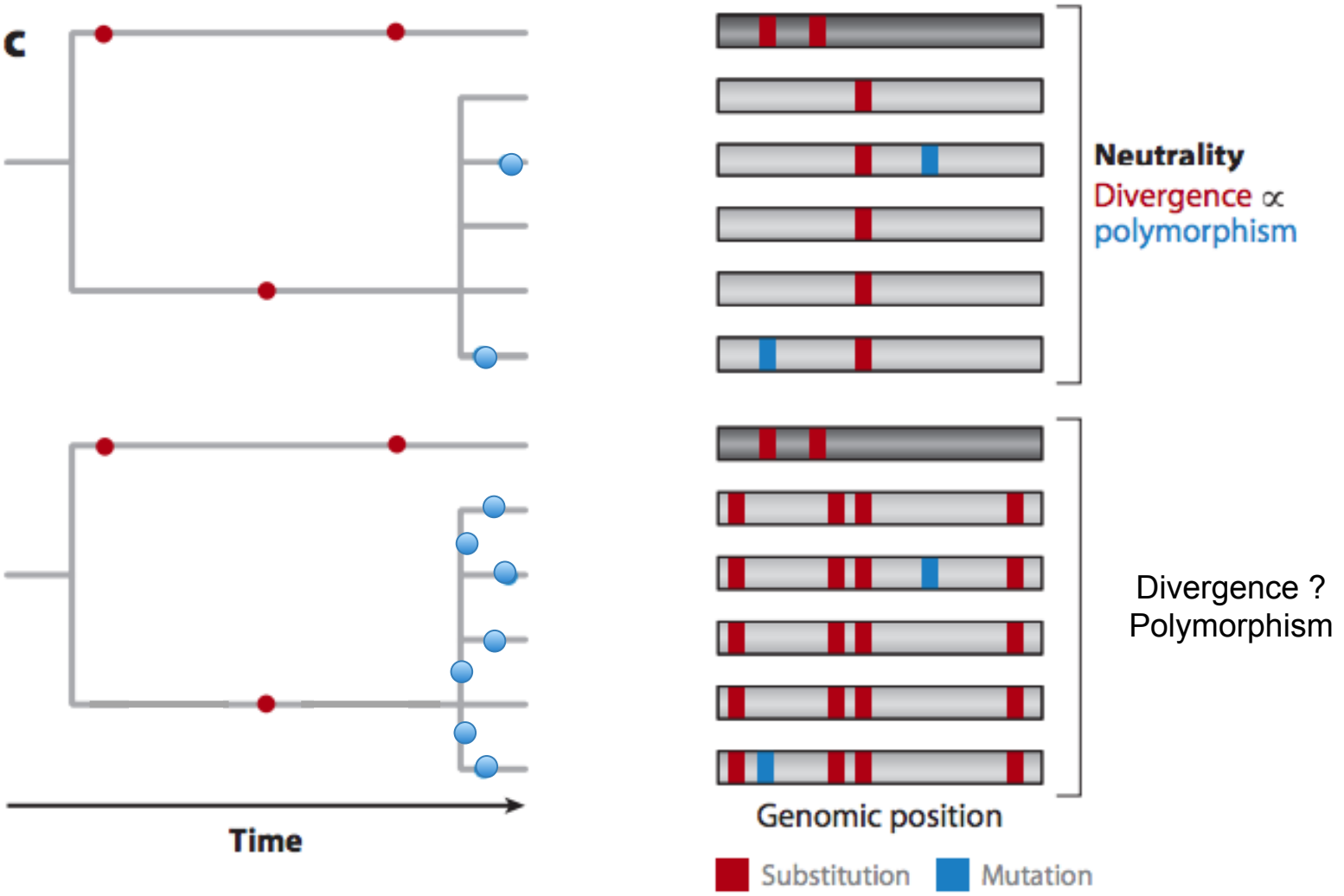
Vitti et al. Annu Rev Genet. 2013

HKA test – positive selection



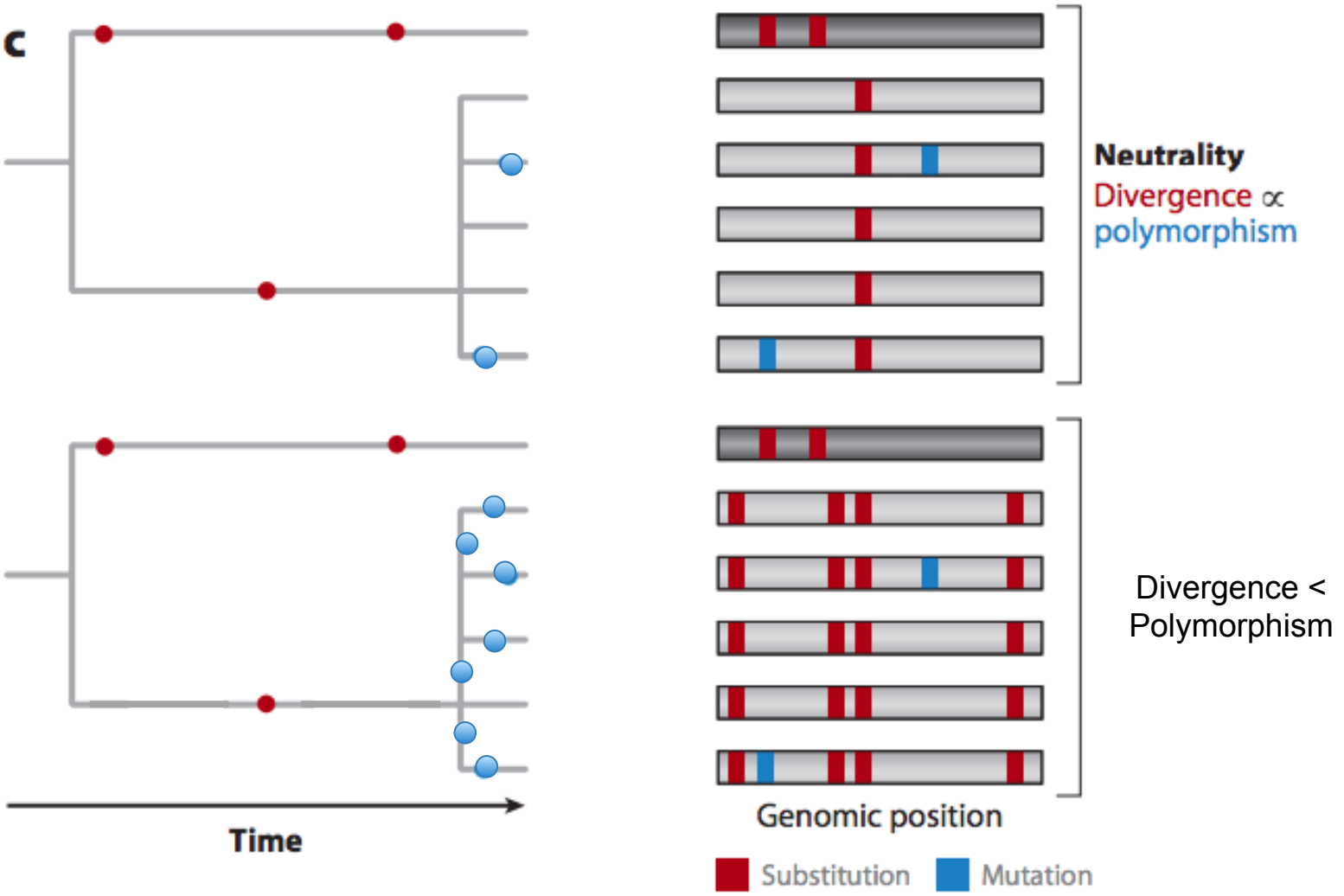
Vitti et al. Annu Rev Genet. 2013

HKA test



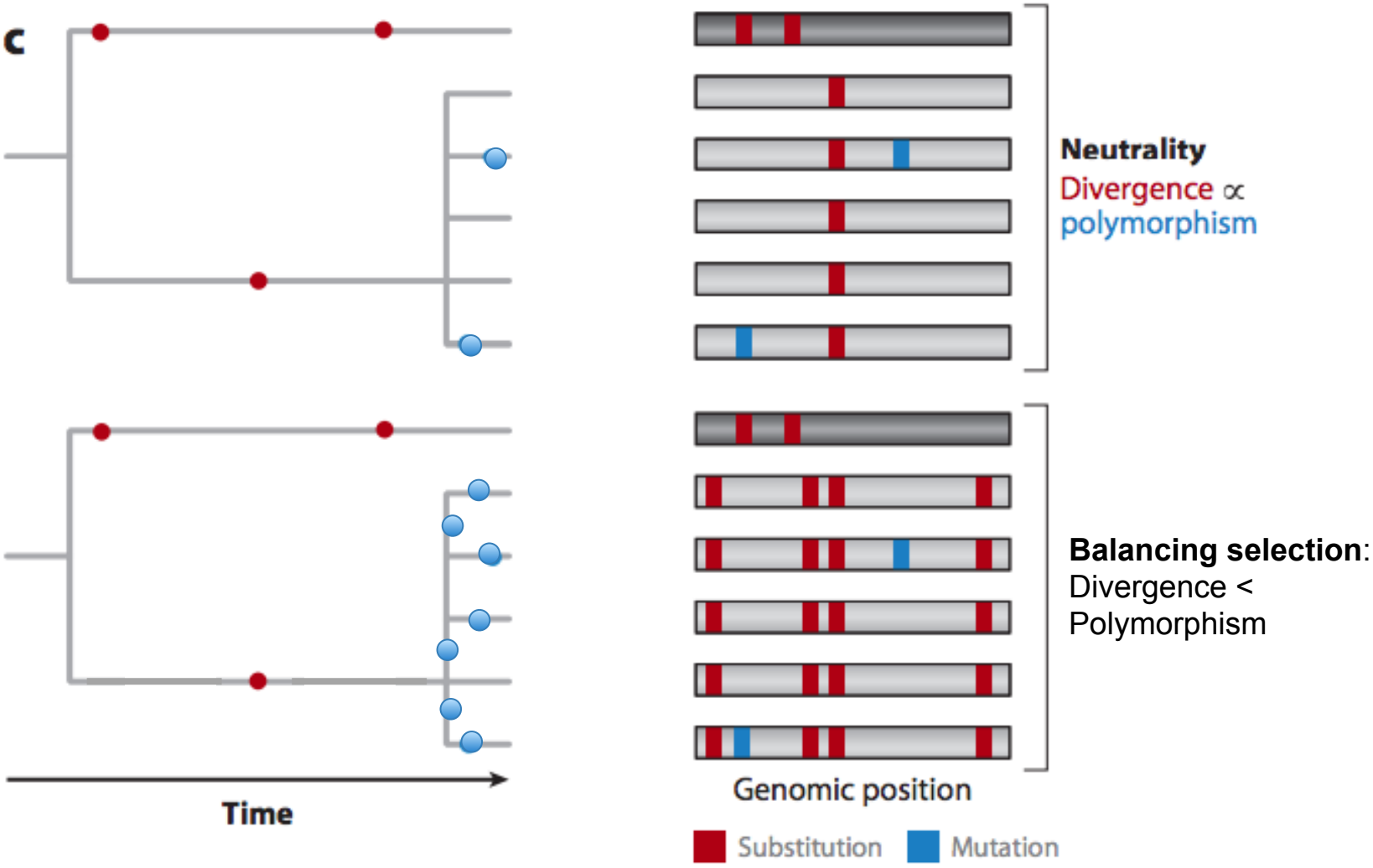
Vitti et al. Annu Rev Genet. 2013

HKA test



Vitti et al. Annu Rev Genet. 2013

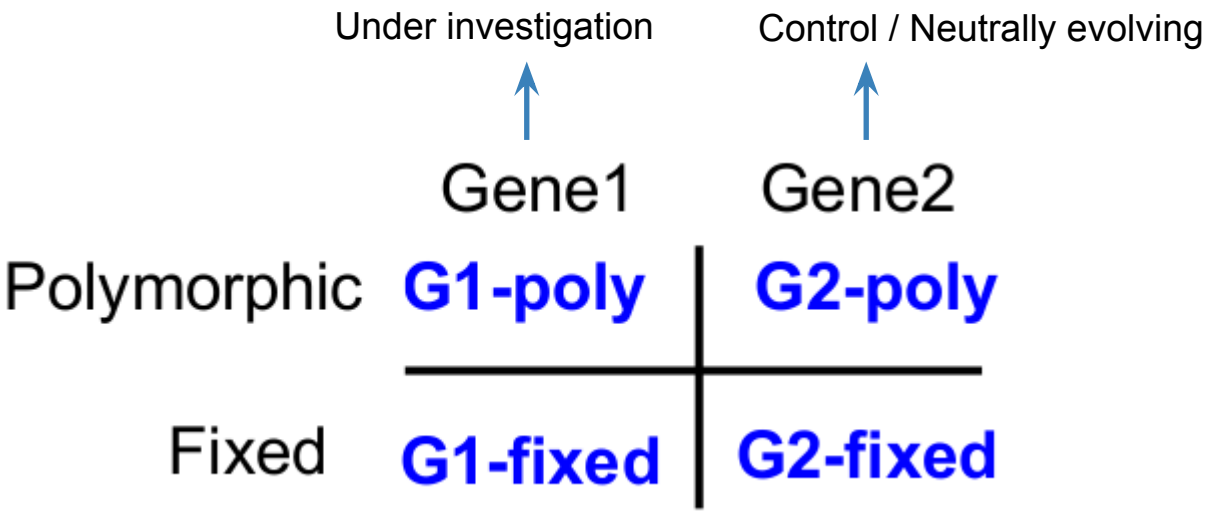
HKA test – balancing selection



Vitti et al. Annu Rev Genet. 2013

HKA test

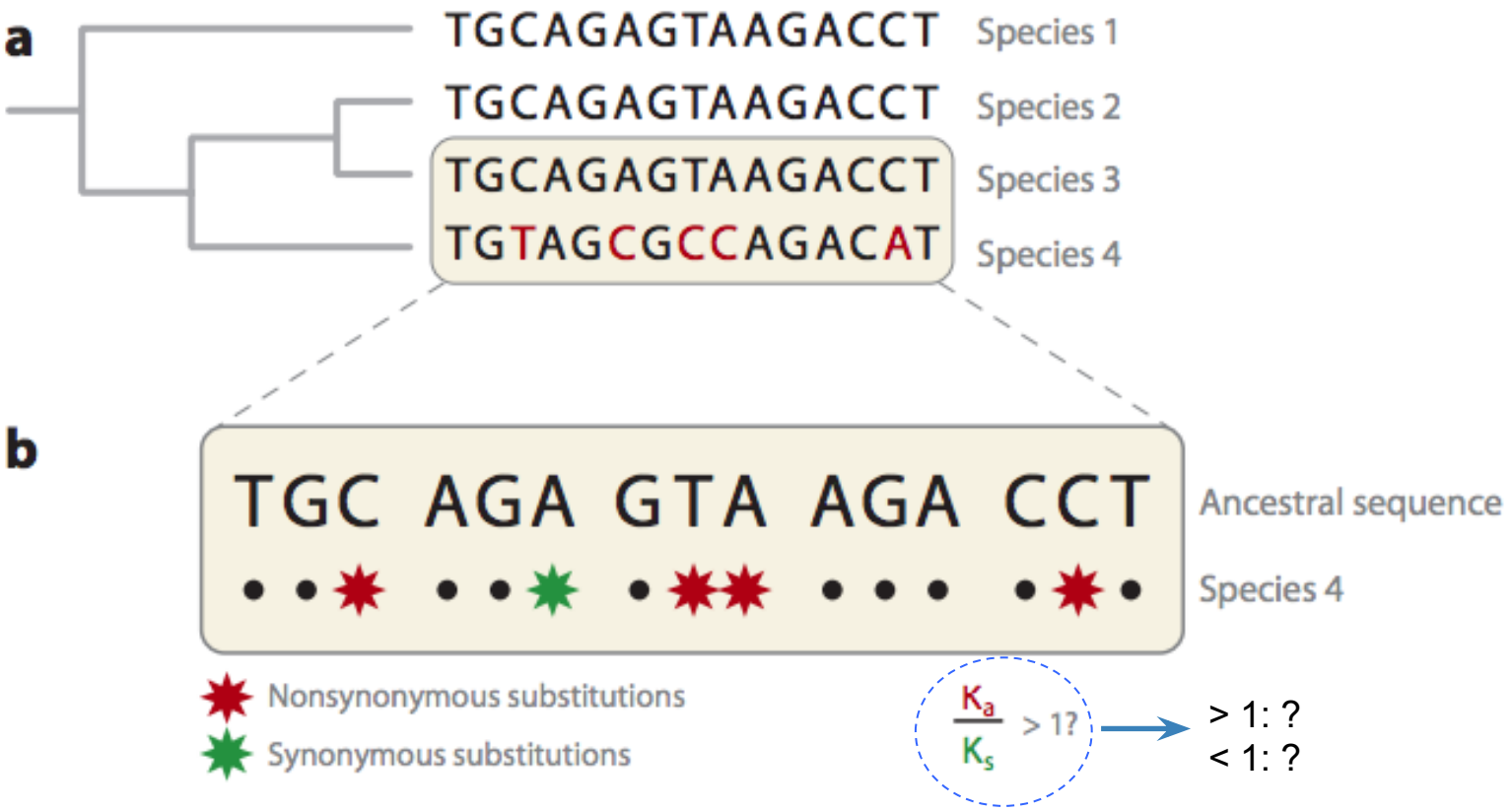
Hudson-Kreitman-Aguadè (HKA, Hudson et al. 1987) test



Contingency table 2x2: chi-square test

d_N/d_S

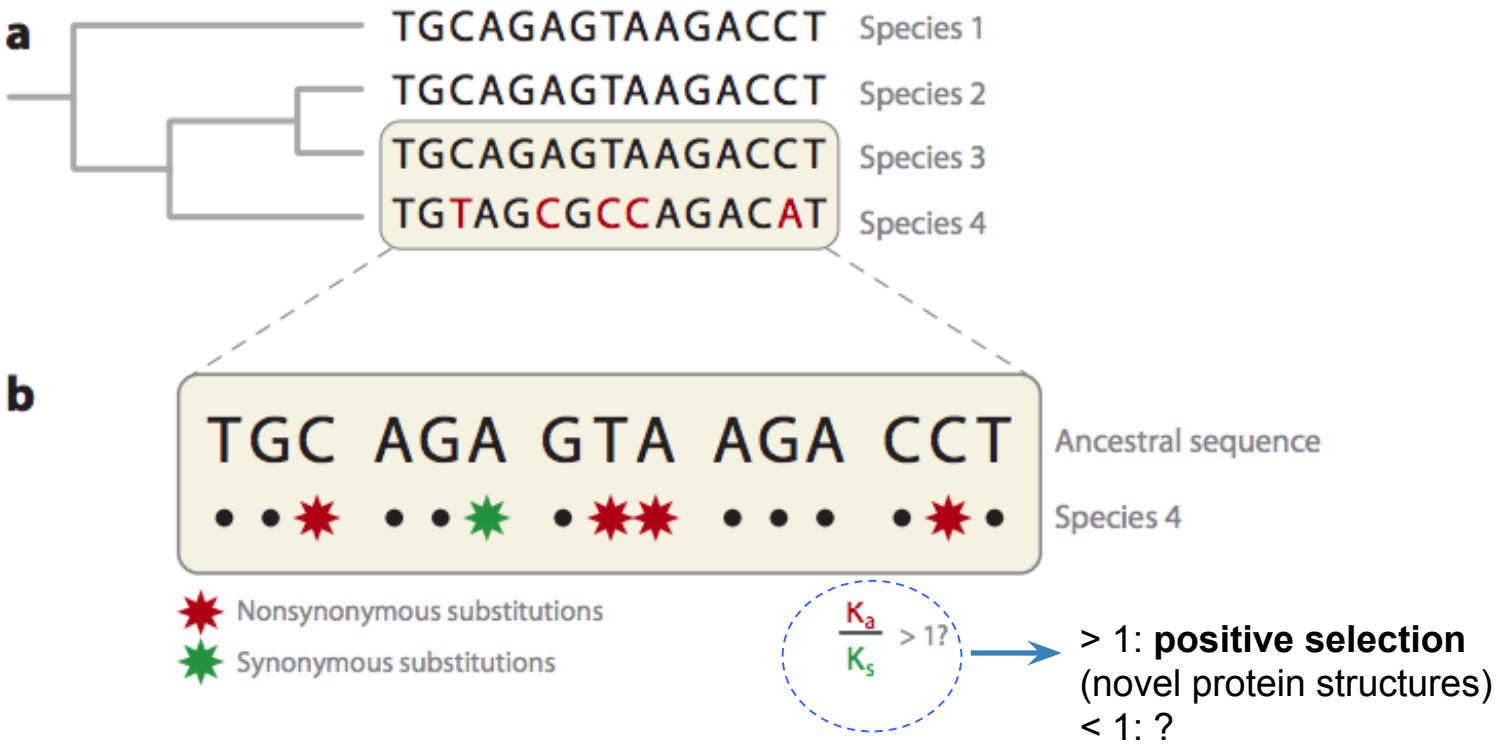
Comparison between rate of **nonsynonymous** substitutions and rate of **synonymous** substitutions (also called K_a/K_s or ω)



Vitti et al. Annu Rev Genet. 2013

d_N/d_S

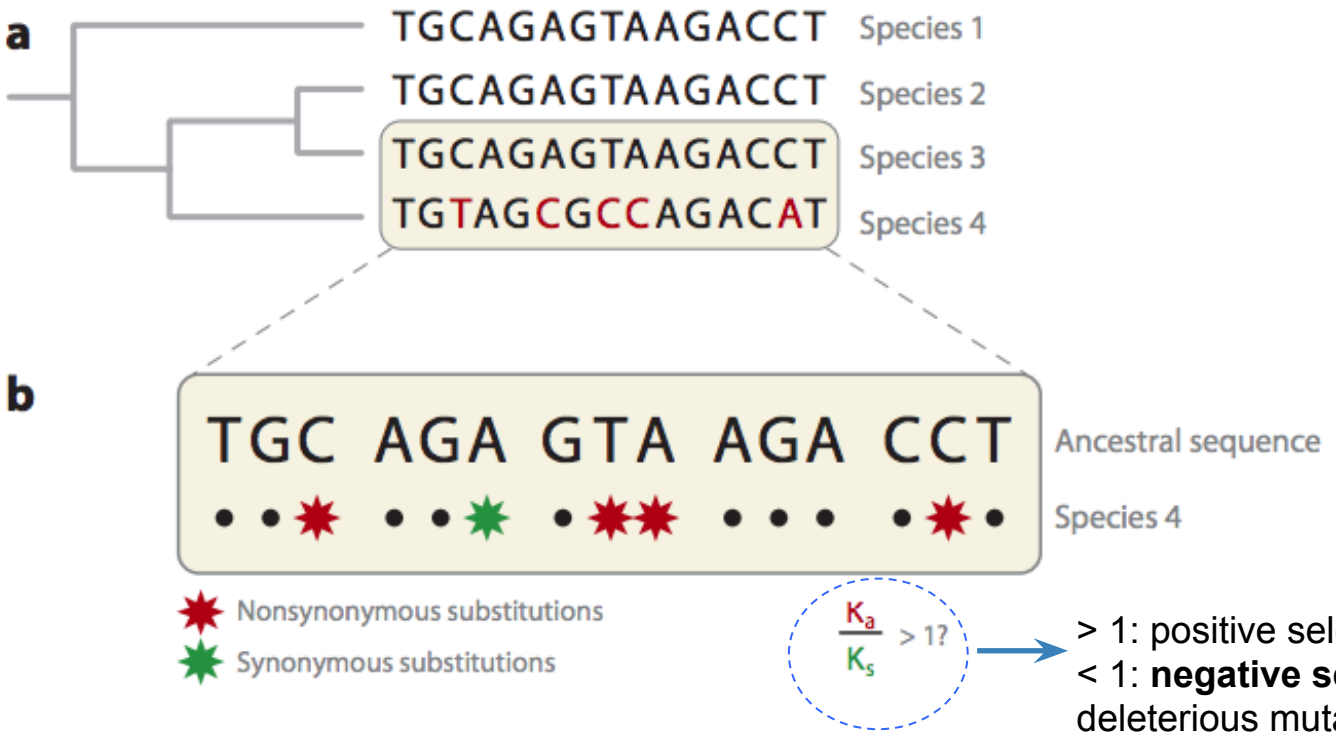
Comparison between rate of **nonsynonymous** substitutions and rate of **synonymous** substitutions (also called K_a/K_s or ω)



Vitti et al. Annu Rev Genet. 2013

d_N/d_S

Comparison between rate of **nonsynonymous** substitutions and rate of **synonymous** substitutions (also called K_a/K_s or ω)



Vitti et al. Annu Rev Genet. 2013

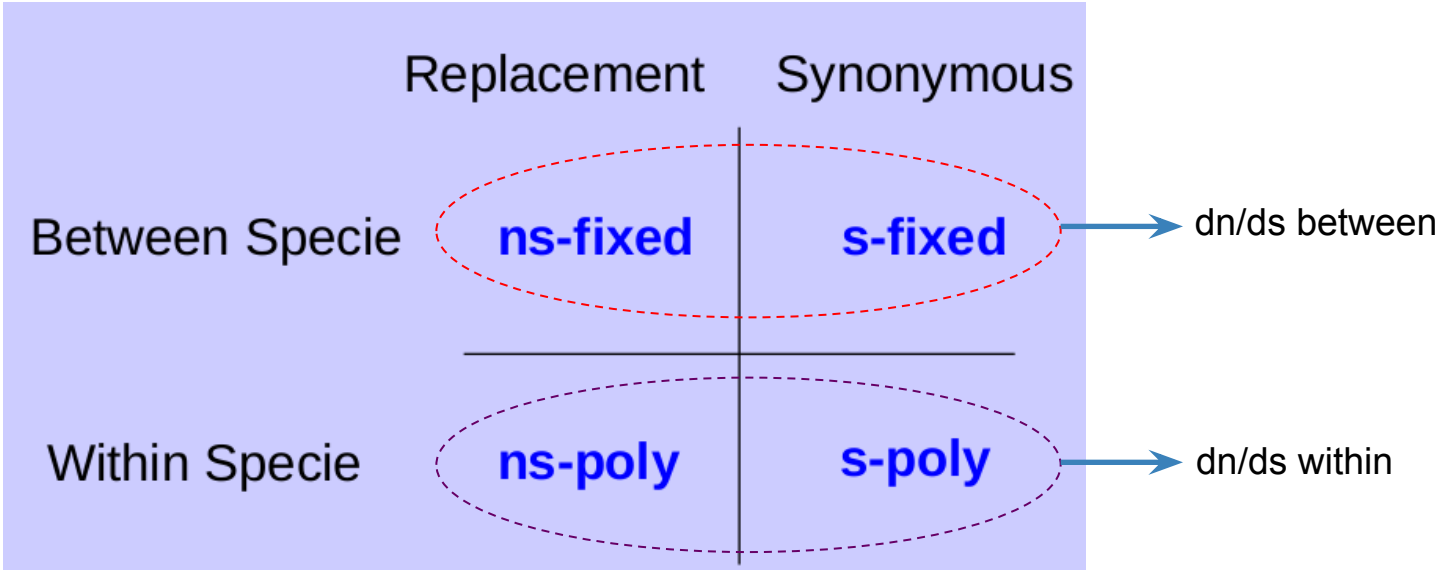
MK test

The MKT compares the amount of variation within a species to the divergence between species at **two types of sites**, one of which is putatively neutral and used as the reference to detect selection at the other types of sites.

	Replacement	Synonymous
Between Specie	**ns-fixed**	**s-fixed**
Within Specie	**ns-poly**	**s-poly**

MK test

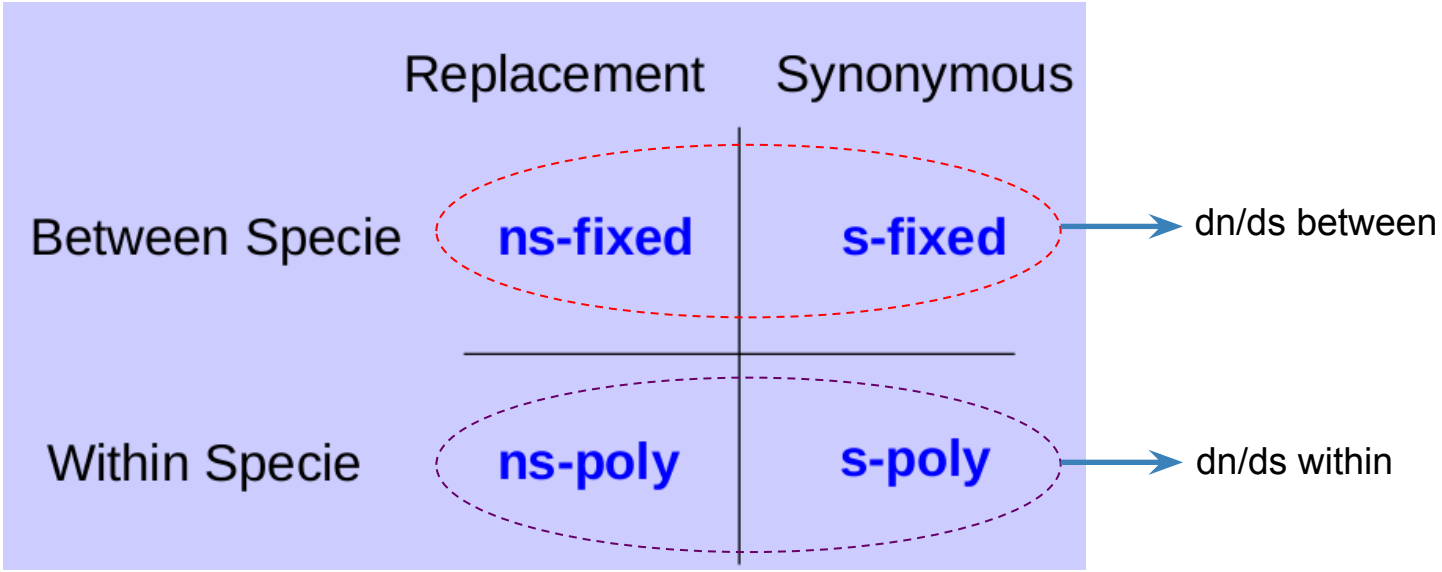
The MKT compares the amount of variation within a species to the divergence between species at **two types of sites**, one of which is putatively neutral and used as the reference to detect selection at the other types of sites.



MK test

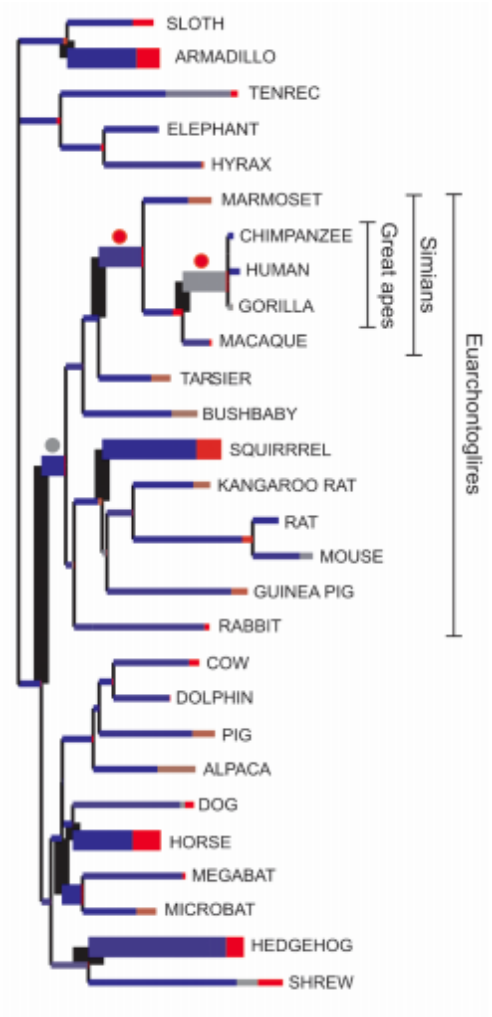
Under neutrality, these rates should be equal.

If the between-species ratio exceeds the within-species ratio indicates **positive selection** between species (a larger within-species value suggests balancing selection or excess or weakly deleterious variants).



Comparative genomics

Identify multiple-species conserved genomic elements which show an accelerated rate of substitution in a particular lineage.




```
                159                                 191
                 |                  *                  |
Sloth            LREAQLRVENDQECSIRFPLYTGQTQICVGDPR
Tenrec           LQDVQLRIQRDQECHNRFRQYRSLLQICVGSPE
Elephant         LQDMQLGVETAKECRIRFSFYTDRLQICVGNPR
Hyrax            LQDVQLRVERDEECSVRFSFYTNQLQICVGNPR
Marmoset         LREVQLRVQRDEQCFRTYRAYNSRTQLCVGDPR
Chimpanzee       LREVQLRVQRDSQCLRIFDSYDPRRQICVGDRR
Human            LREVQLRVQRDRQCLRIFGSYDPRRQICVGDRR
Gorilla          LREVQLRVQRDSQCLRIFGSYDPRRQICVGDRR
Macaque          LREVQLRVQRDRECFRLFSSYDGGSQICVGDRR
Bushbaby         LQEVQLRVQRDQQCQNRFRFYIGHTQICVGDPR
Kangaroo rat     LQEVMLRVQRDQKCSSRFSTYTSQTQLCVGNPR
Rat              LQEVKLRVQTDQTCANRFQFYNSQTQICVGNPR
Mouse            LQEVQLRVQMDQMCANRFQFYNSQTQICVGNPR
Guinea pig       LQEVQLRVQSDRKCIRRFSIYSSKTEICVGDPR
Rabbit           LHEVQLKVQRDRECSRRFSFYHGQTQIYVGNPR
Cow              LQCVQLRVQRDRVCRRRFMLYYGRTQICVGDPR
Dolphin          LQDVQLTVQRNGECSRRFMFYTGQTQICVGDPR
Dog              LHEVQLTVQRDSRCRDRFSFYTGQTQICVGNPS
Horse            LQDVRLRVQRDQECSDRFDFYTRQTQICVGDPR
Hedgehog         LHEVQLQIESNQECNQRFSFFNLQTQICVGDSR
Shrew            LQEVRLRVQRDEPCTRRFRFYNGLTQICVGDPT
```

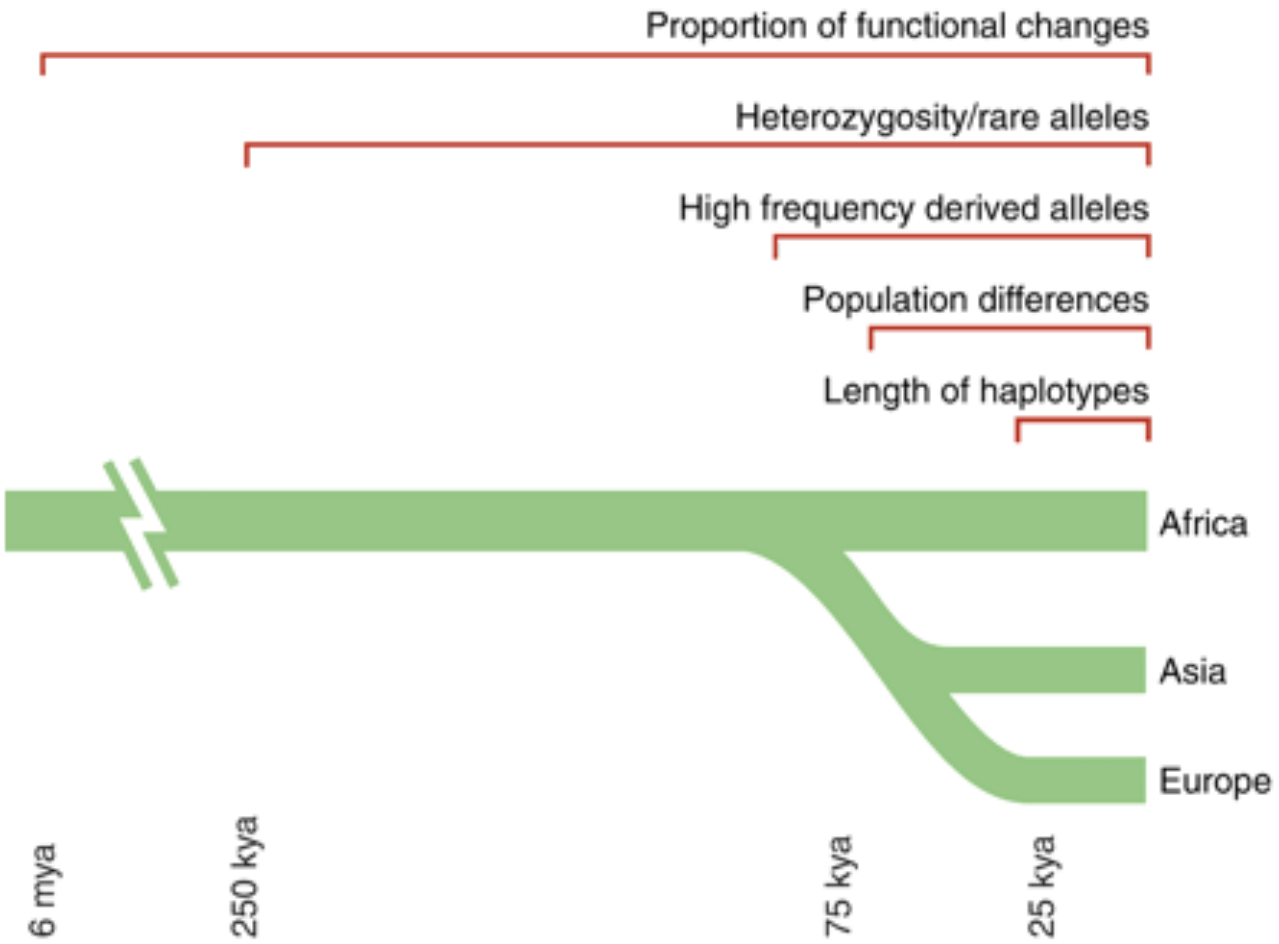
Forni et al. Plos Genetics 2014

Methods to infer selection (II)

- within-species:

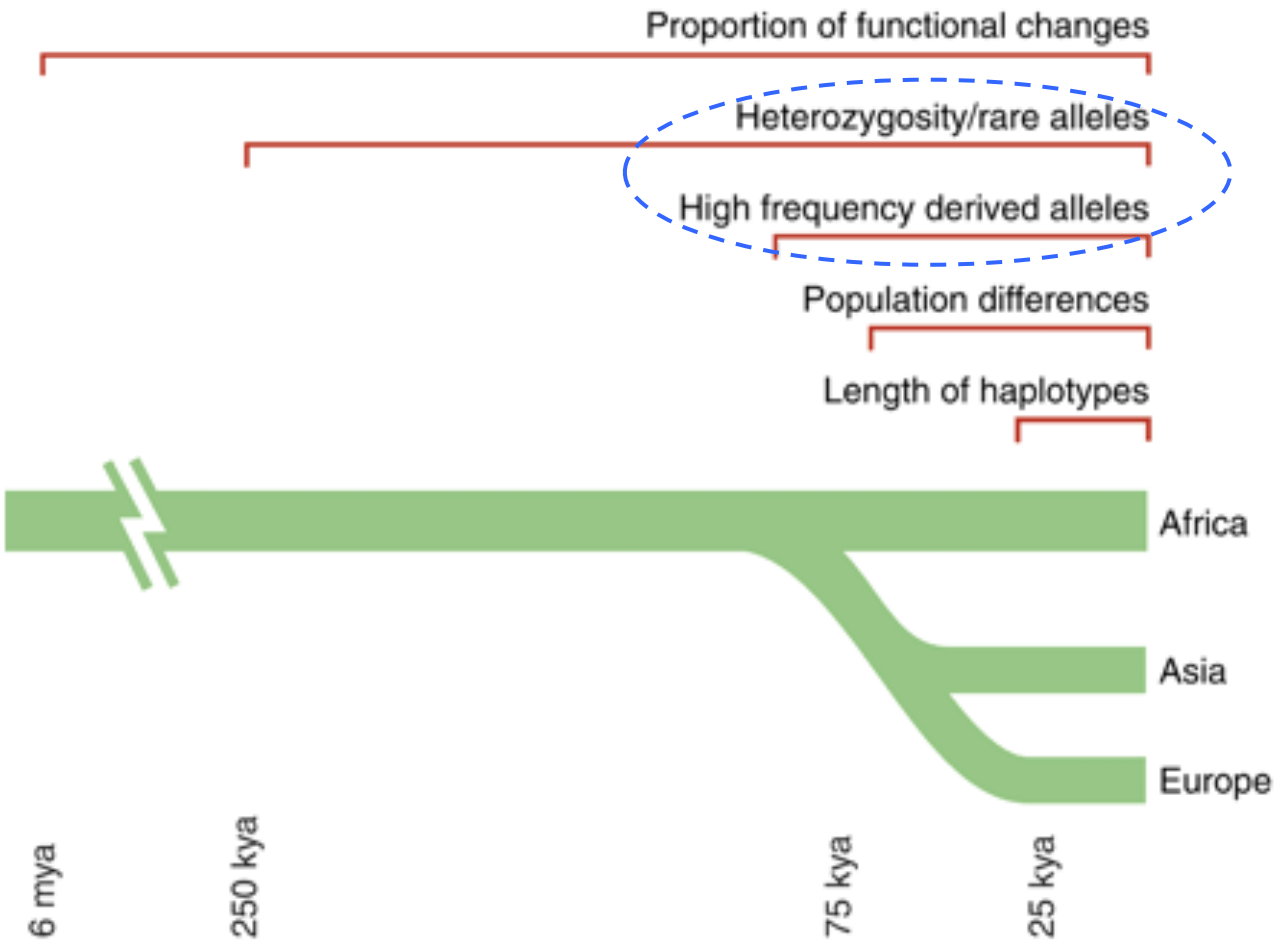
  Micro-evolutionary events between populations, local adaptation

Methods to infer recent selection



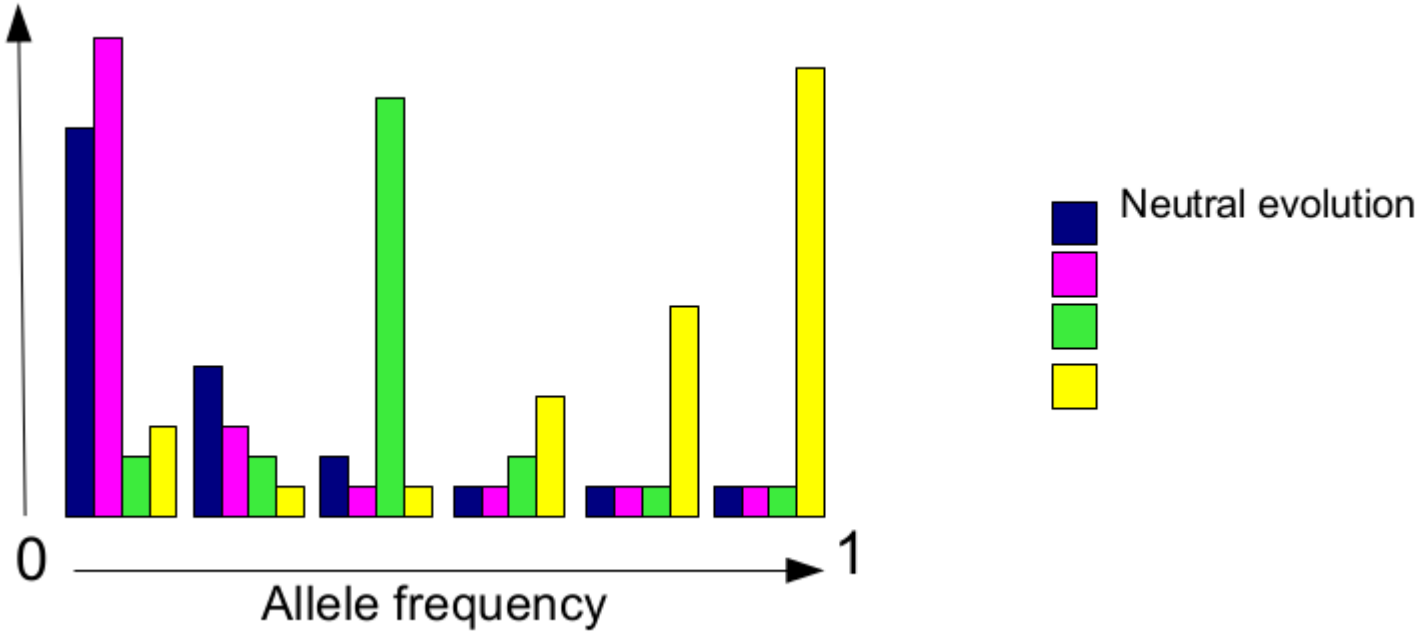
Sabeti et al. 2006 Science

Methods to infer recent selection

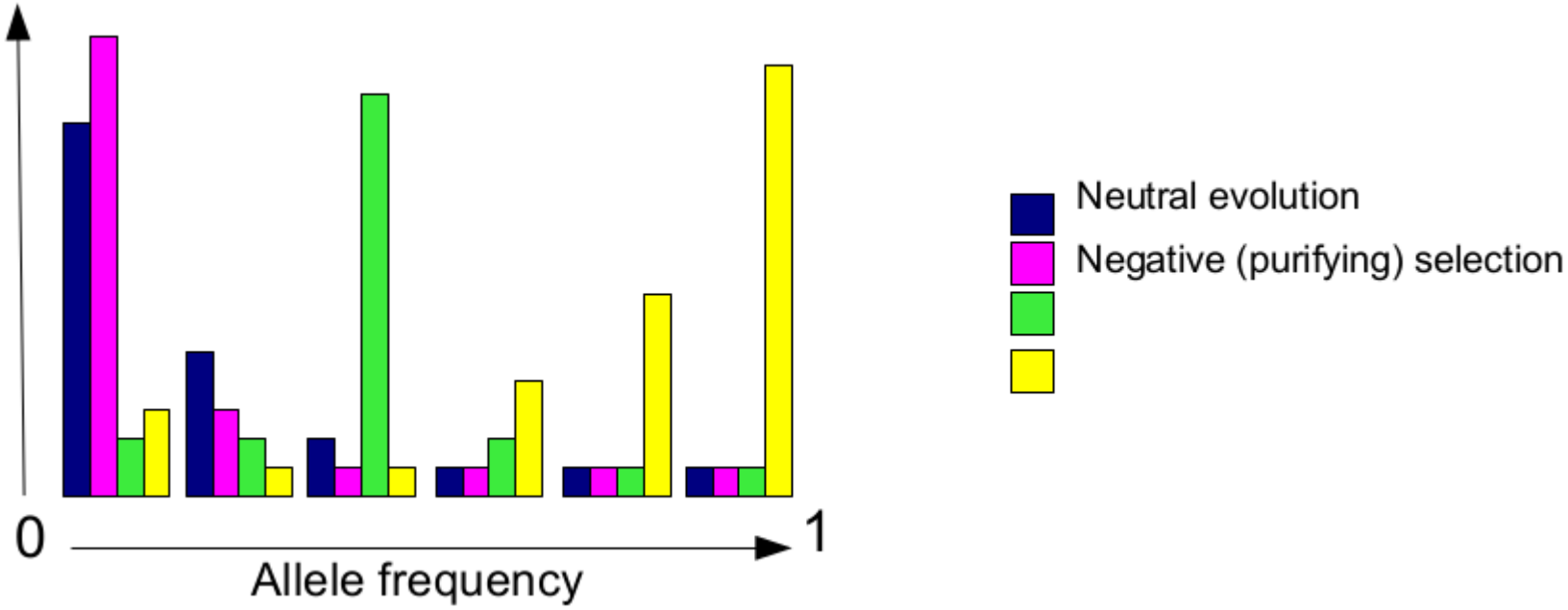


Sabeti et al. 2006 Science

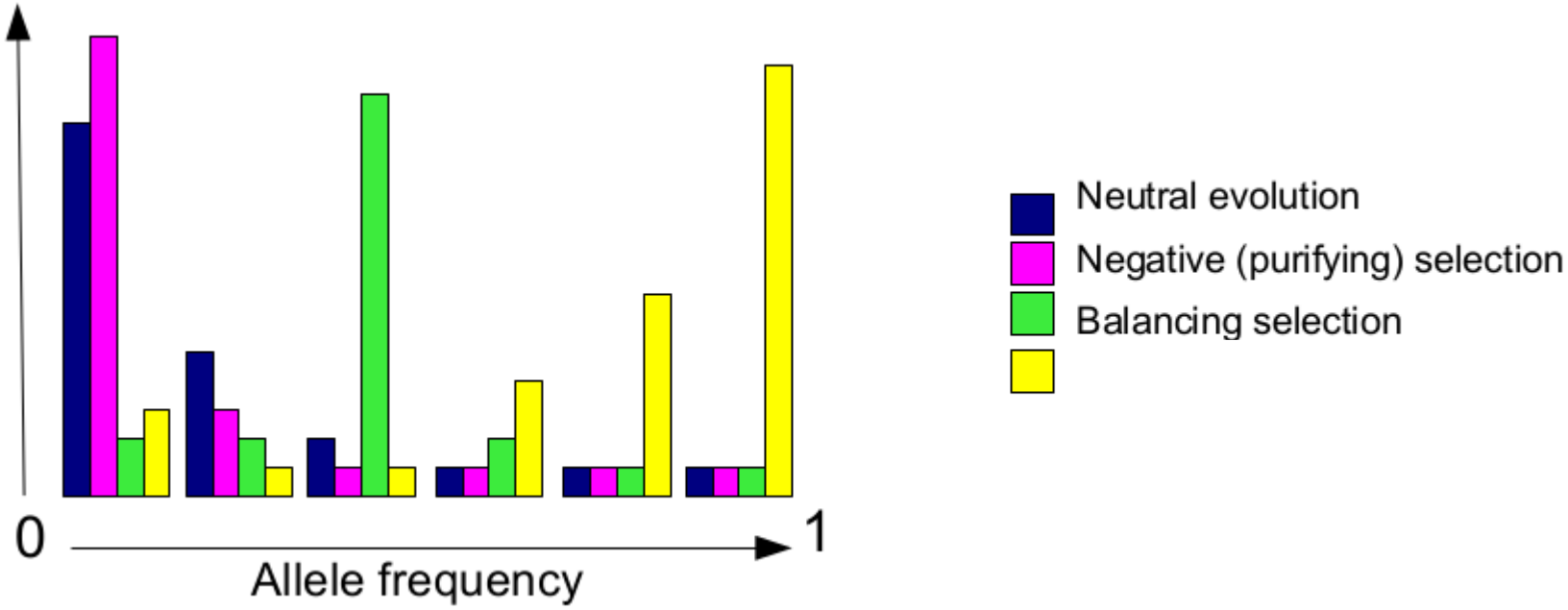
SFS-based statistics



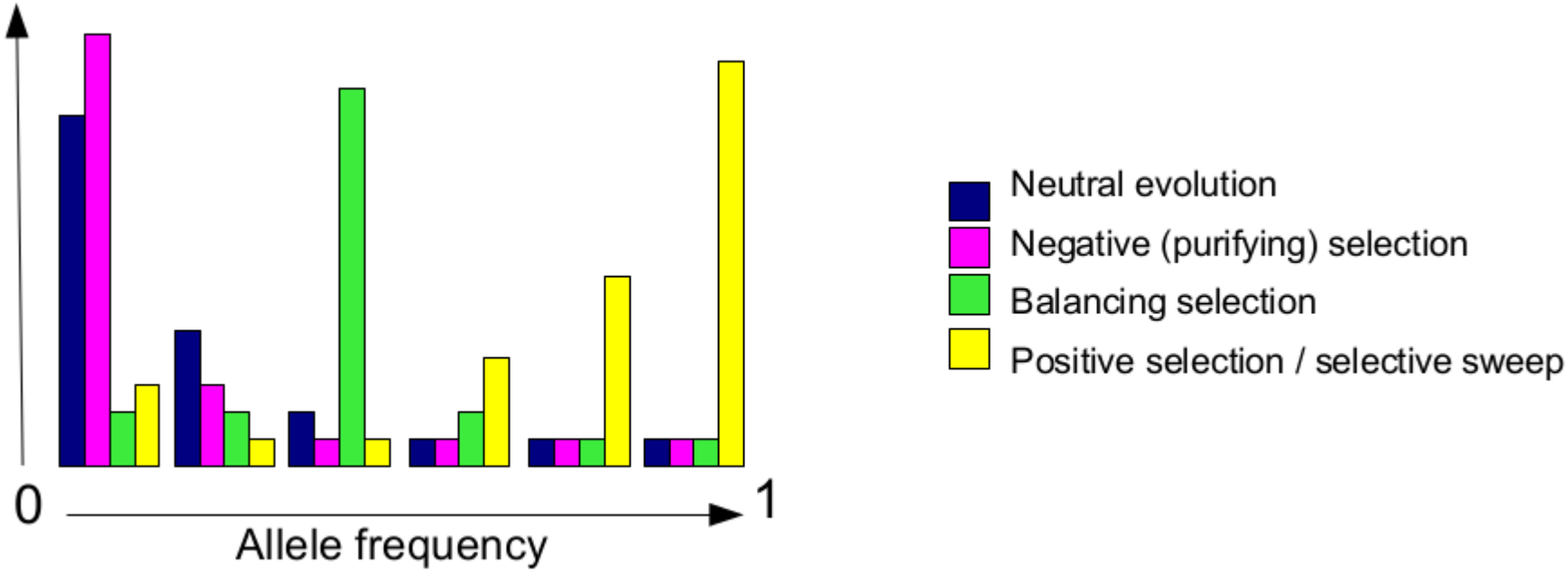
SFS-based statistics



SFS-based statistics



SFS-based statistics



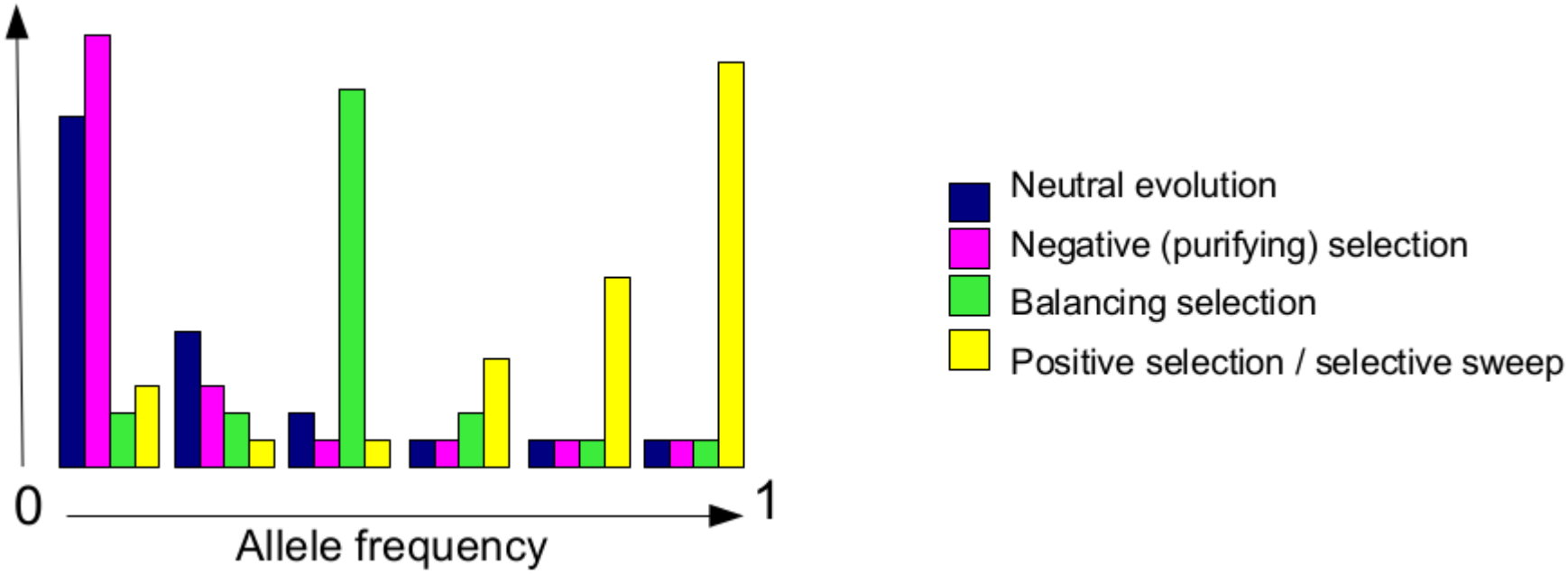
SFS-based statistics

Tajima's D (1989) is the standardized difference of two nucleotide diversity estimates π and θ_W.

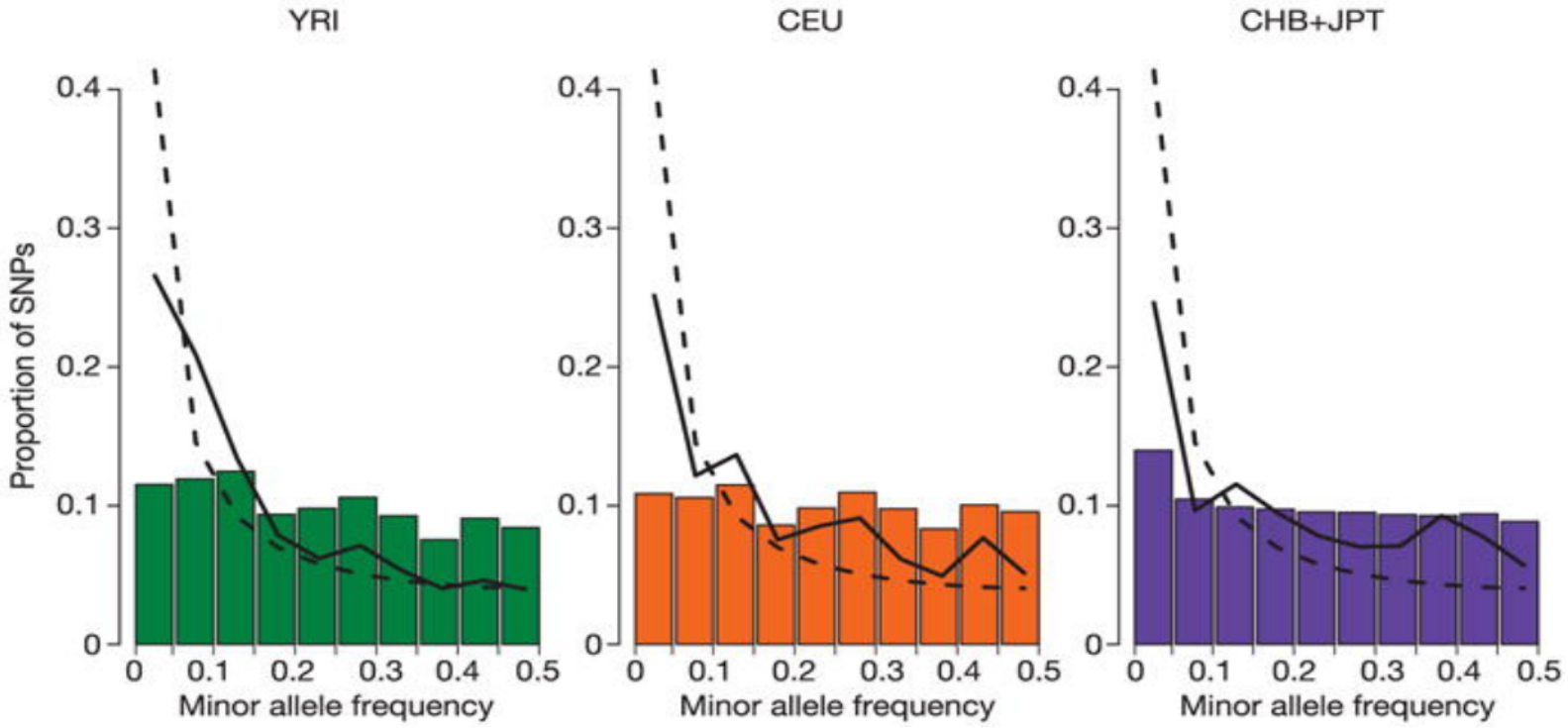
→ Positive values of D indicate an excess of intermediate frequency alleles;

→ Negative values of D indicate an excess of low frequency alleles.

Fu and Li (1993) extended Tajima's test by including information regarding the genealogy of the sample.

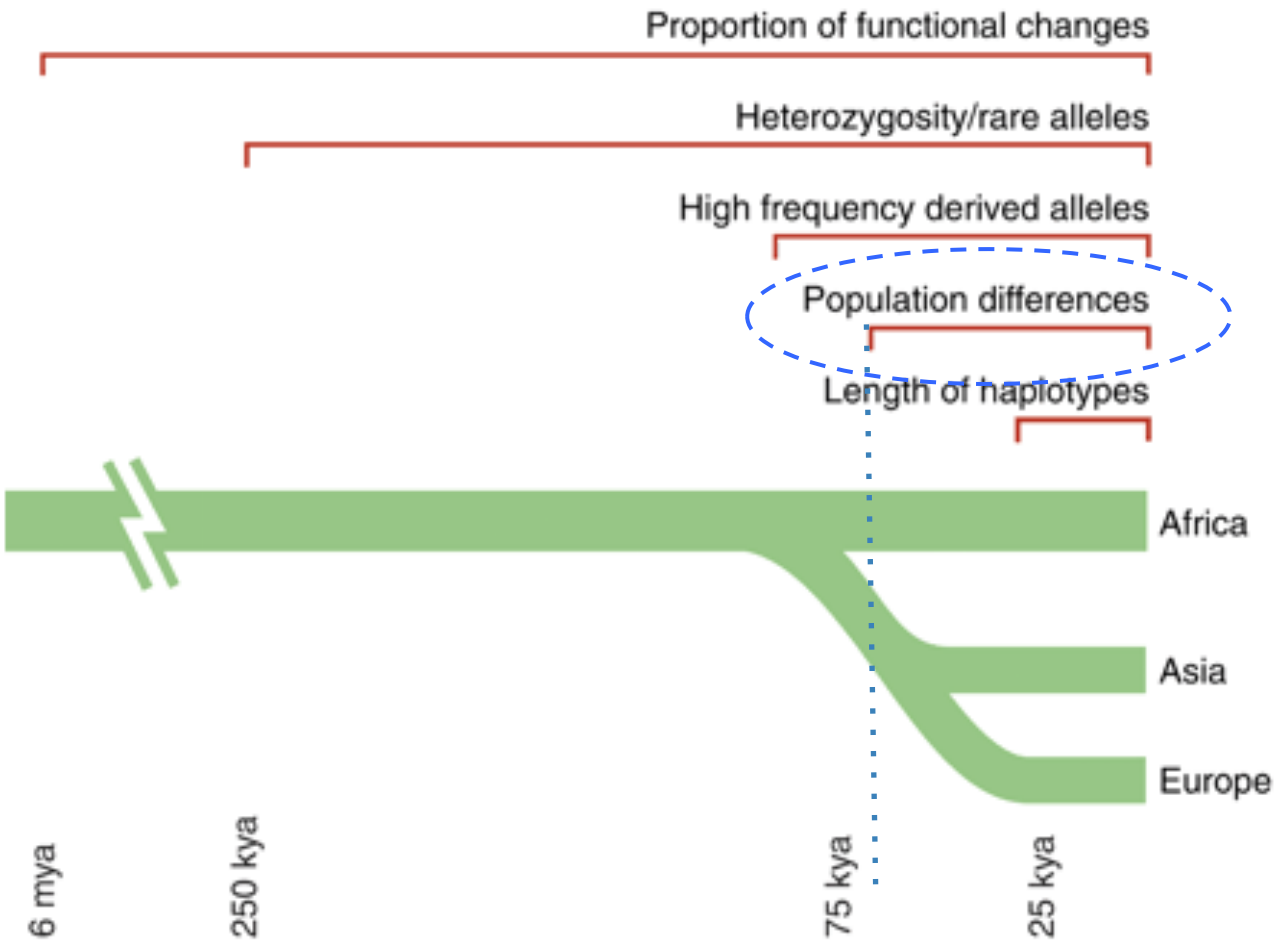


Ascertainment bias



HapMap Consortium. Nature 2005

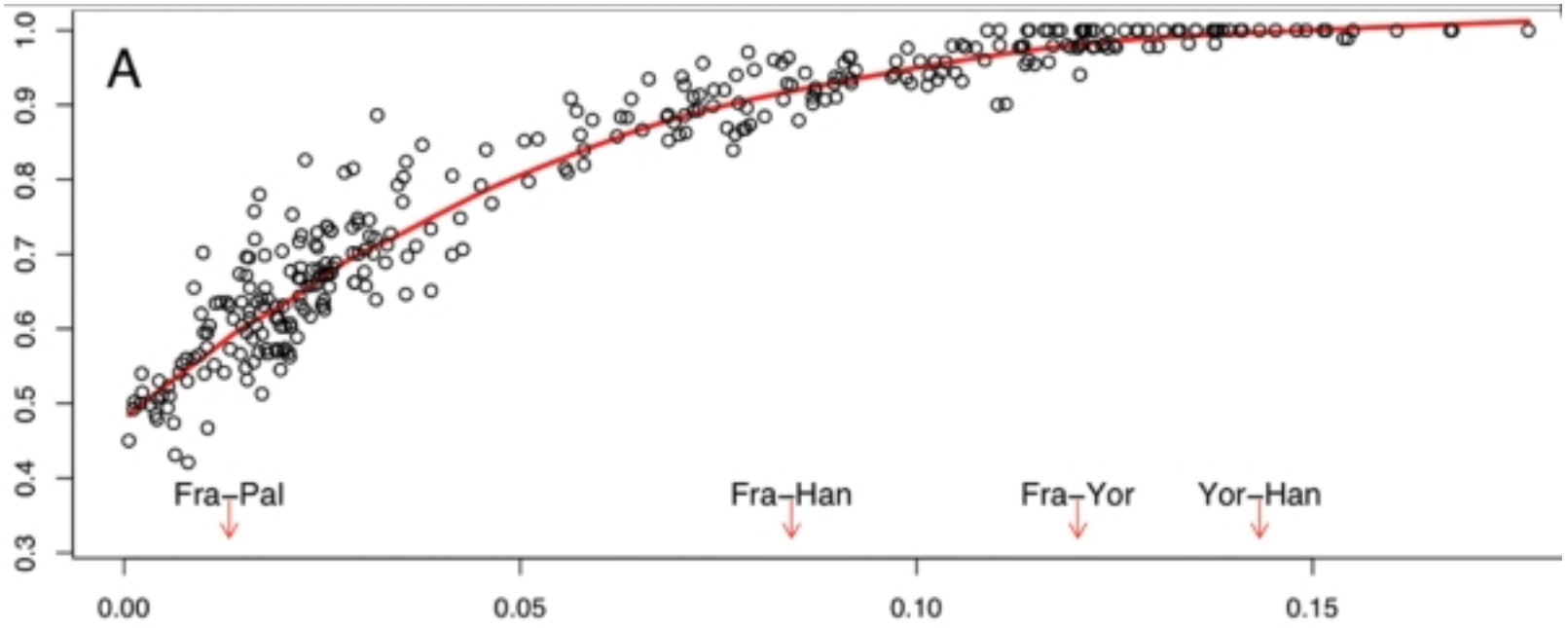
Methods to infer recent selection



Sabeti et al. 2006 Science

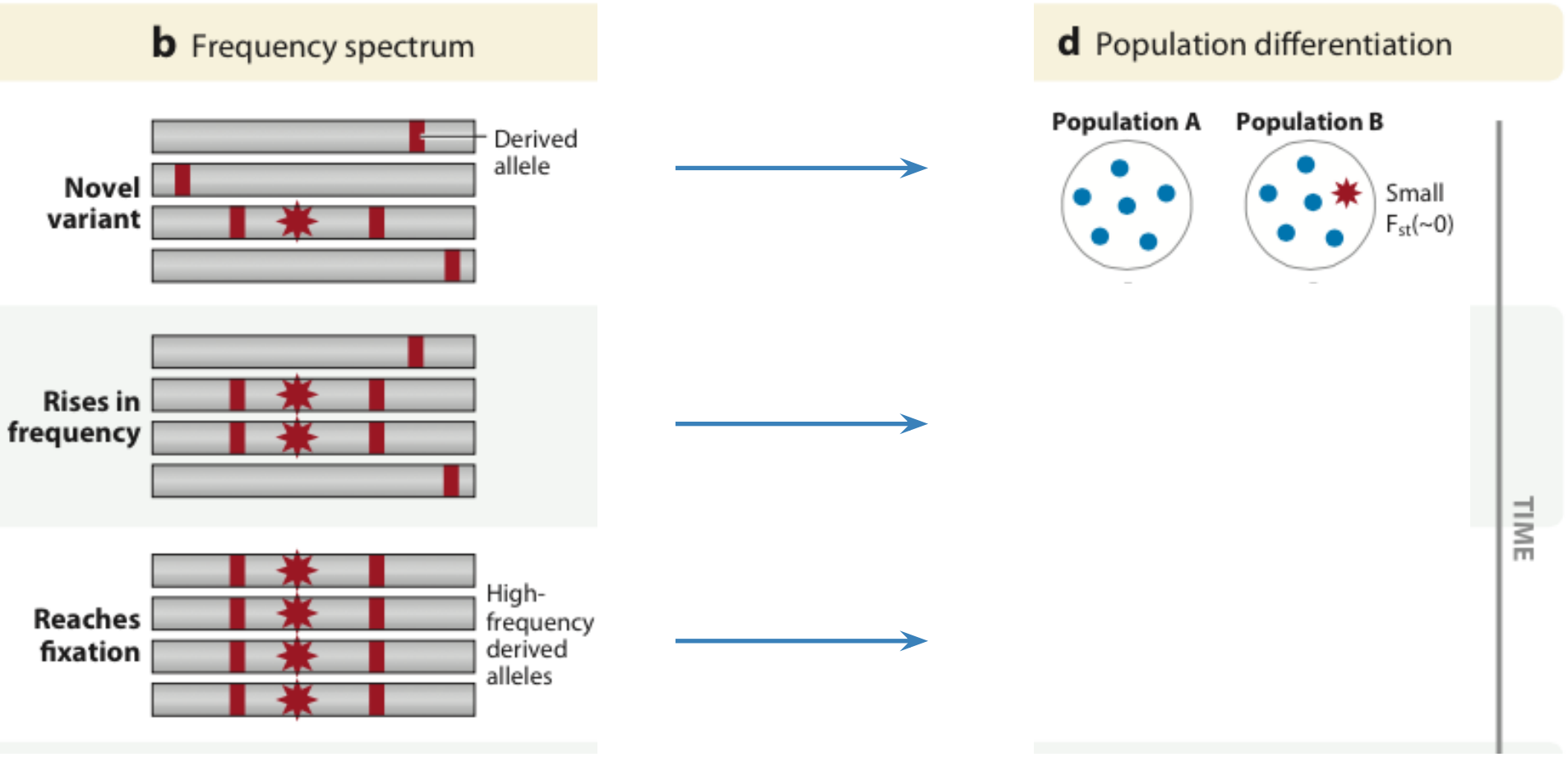
F_{ST}

Method-of-moments or maximum-likelihood estimators of F_{ST}.



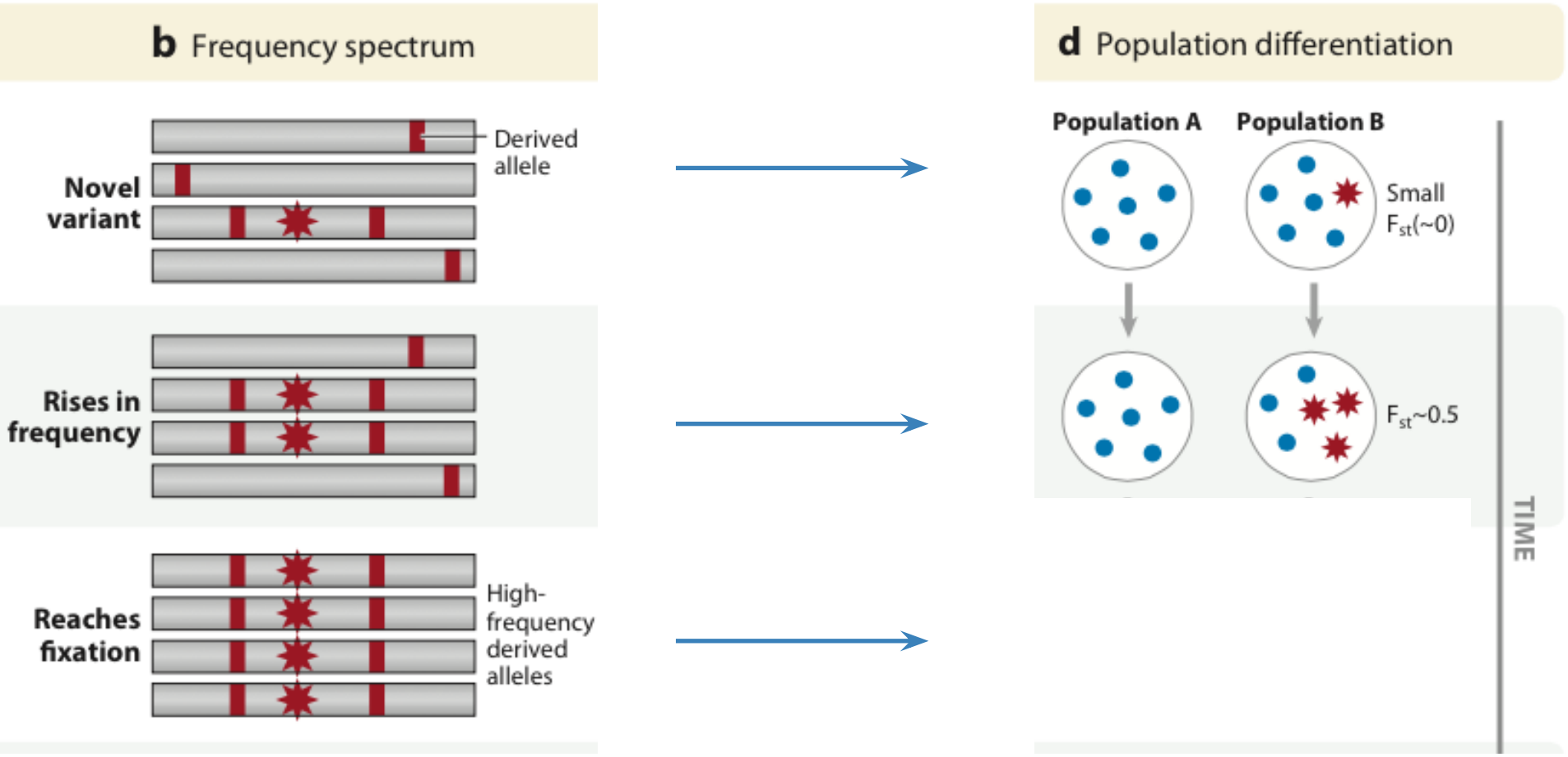
Coop et al. Plos genetics 2009

F_{ST}



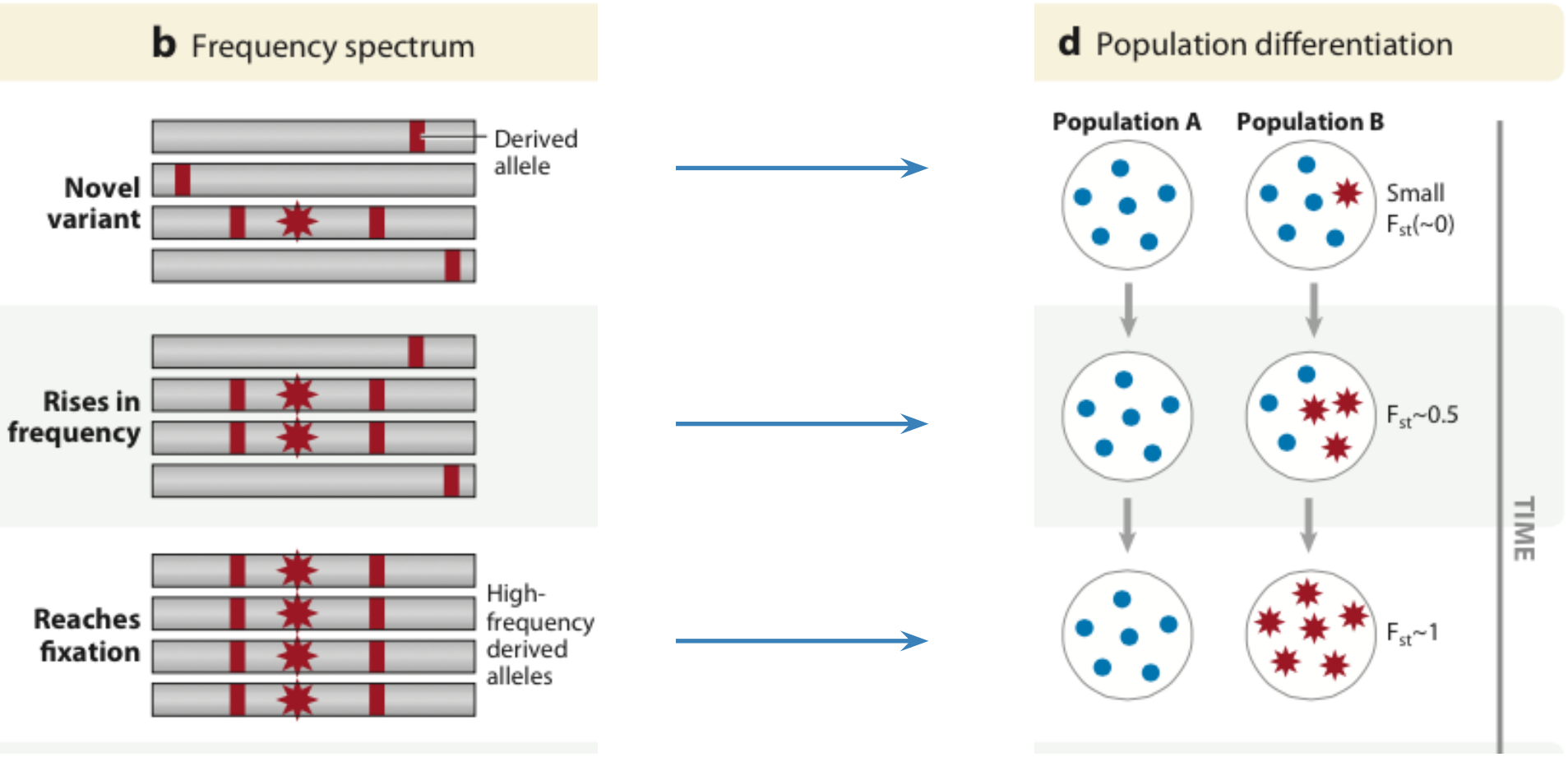
Vitti et al. Annu Rev Genet. 2013

F_{ST}



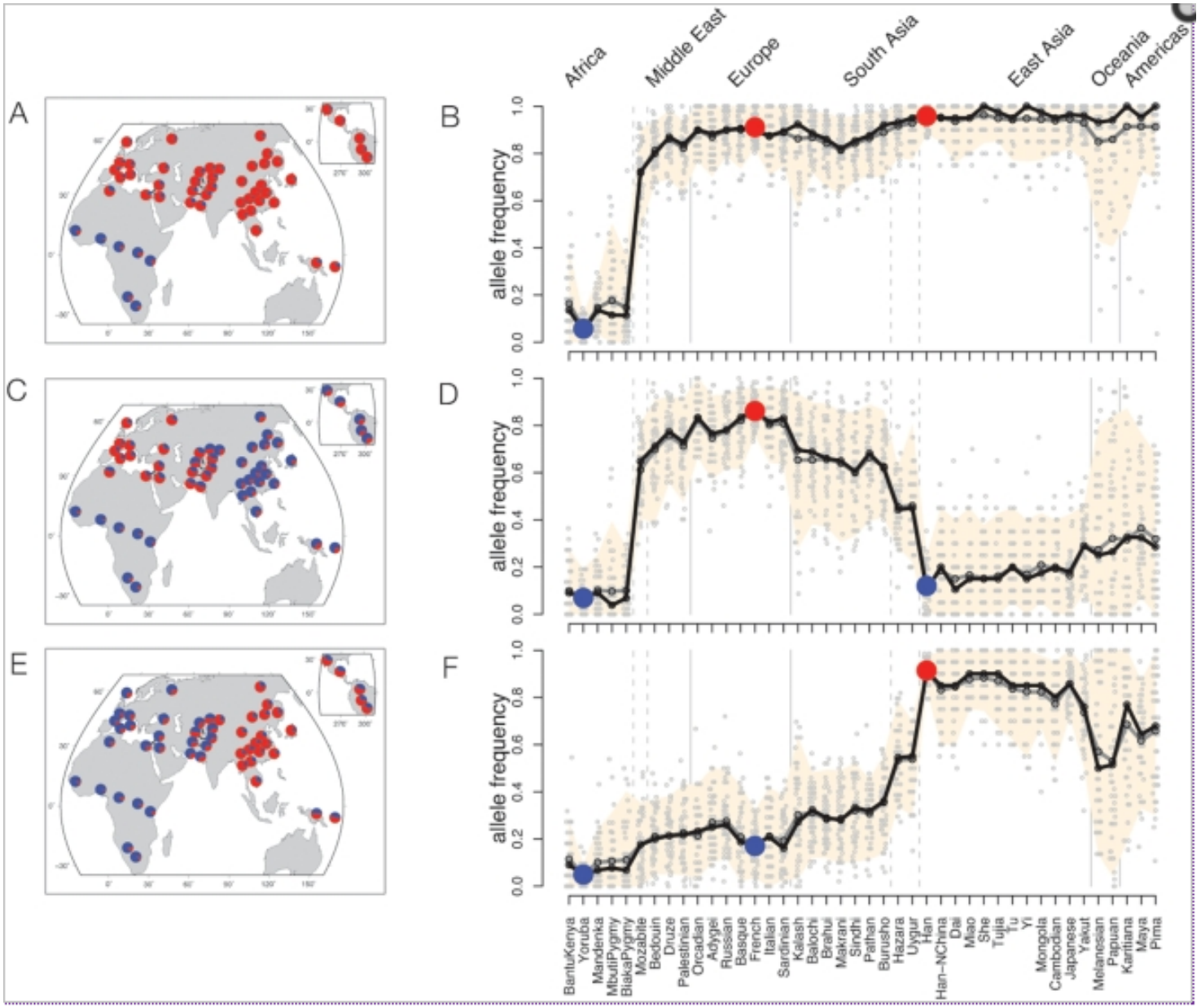
Vitti et al. Annu Rev Genet. 2013

F_{ST}



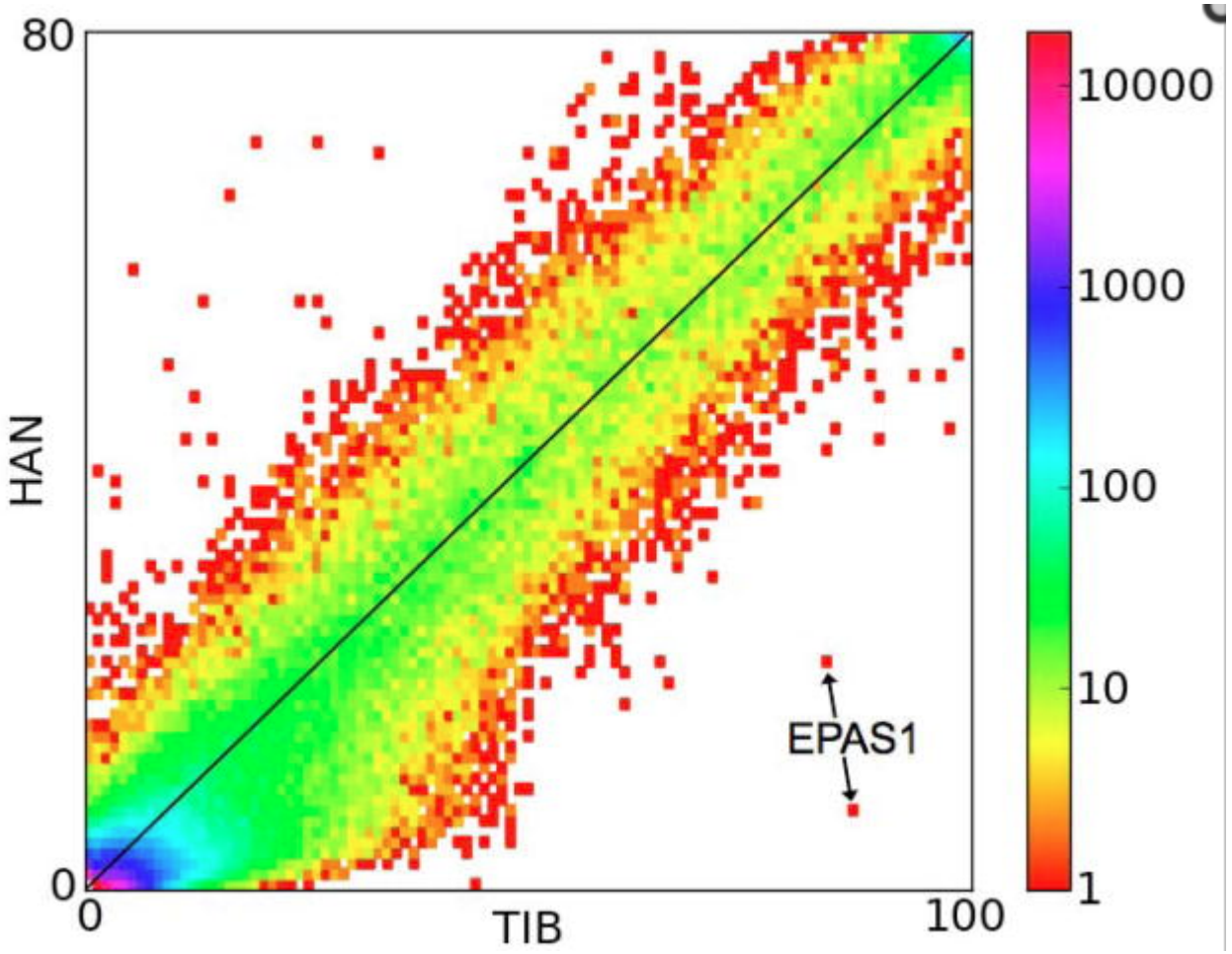
Vitti et al. Annu Rev Genet. 2013

Genetic differentiation



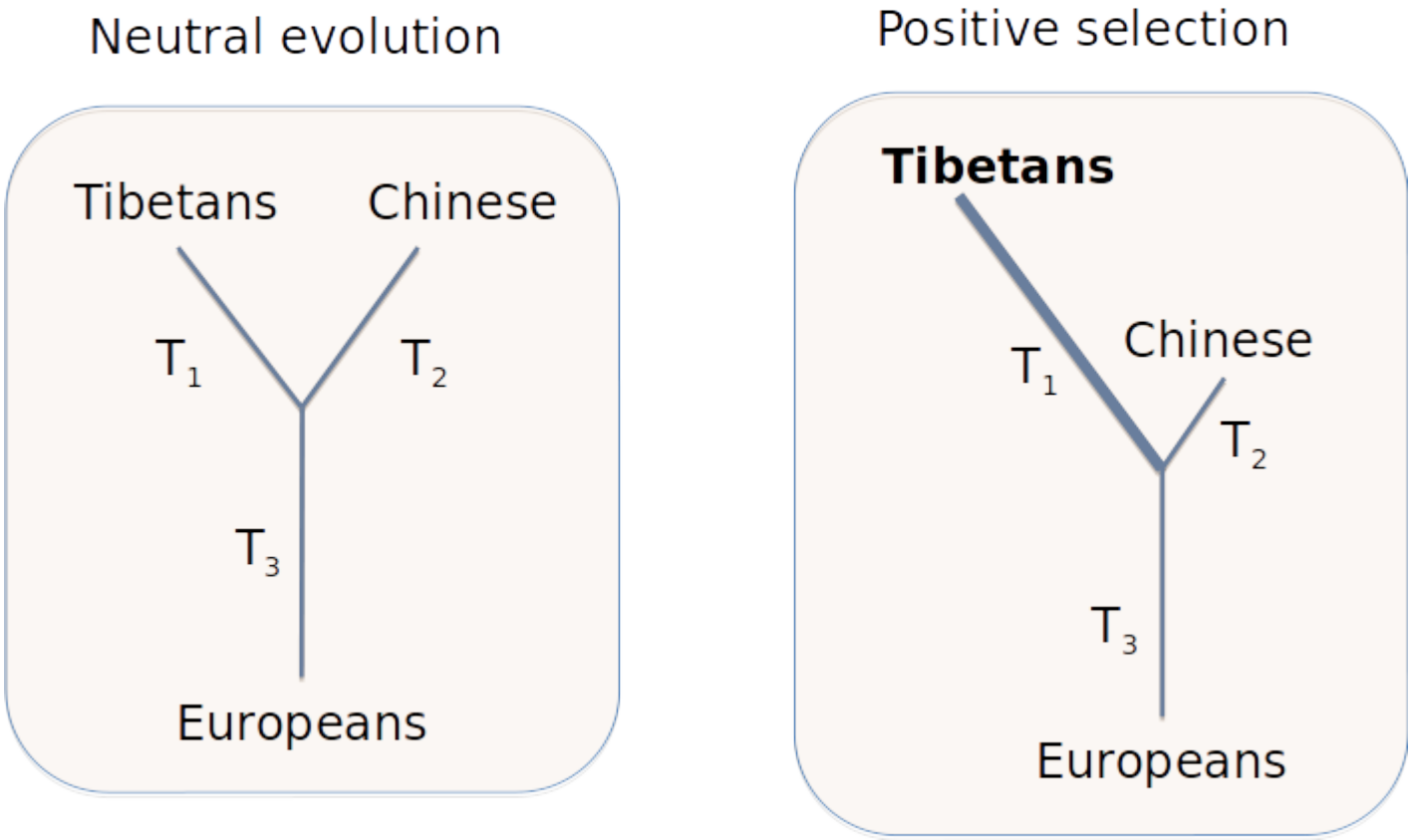
Coop et al. Plos genetics 2009

2D-SFS



Yi et al. Science 2010

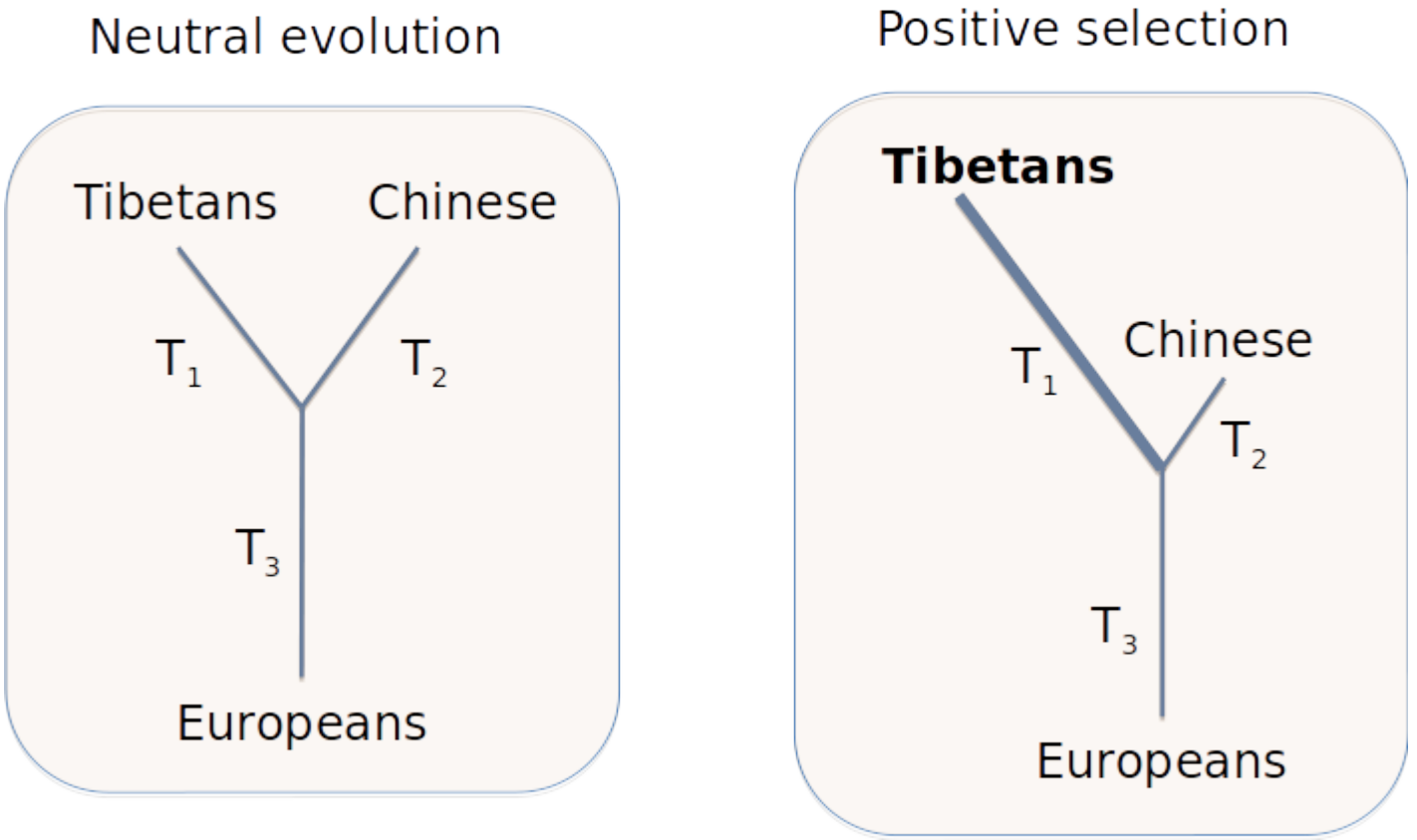
Population Branch Statistic



Population Branch Statistic (PBS) ≈ ?

Yi et al. Science 2010

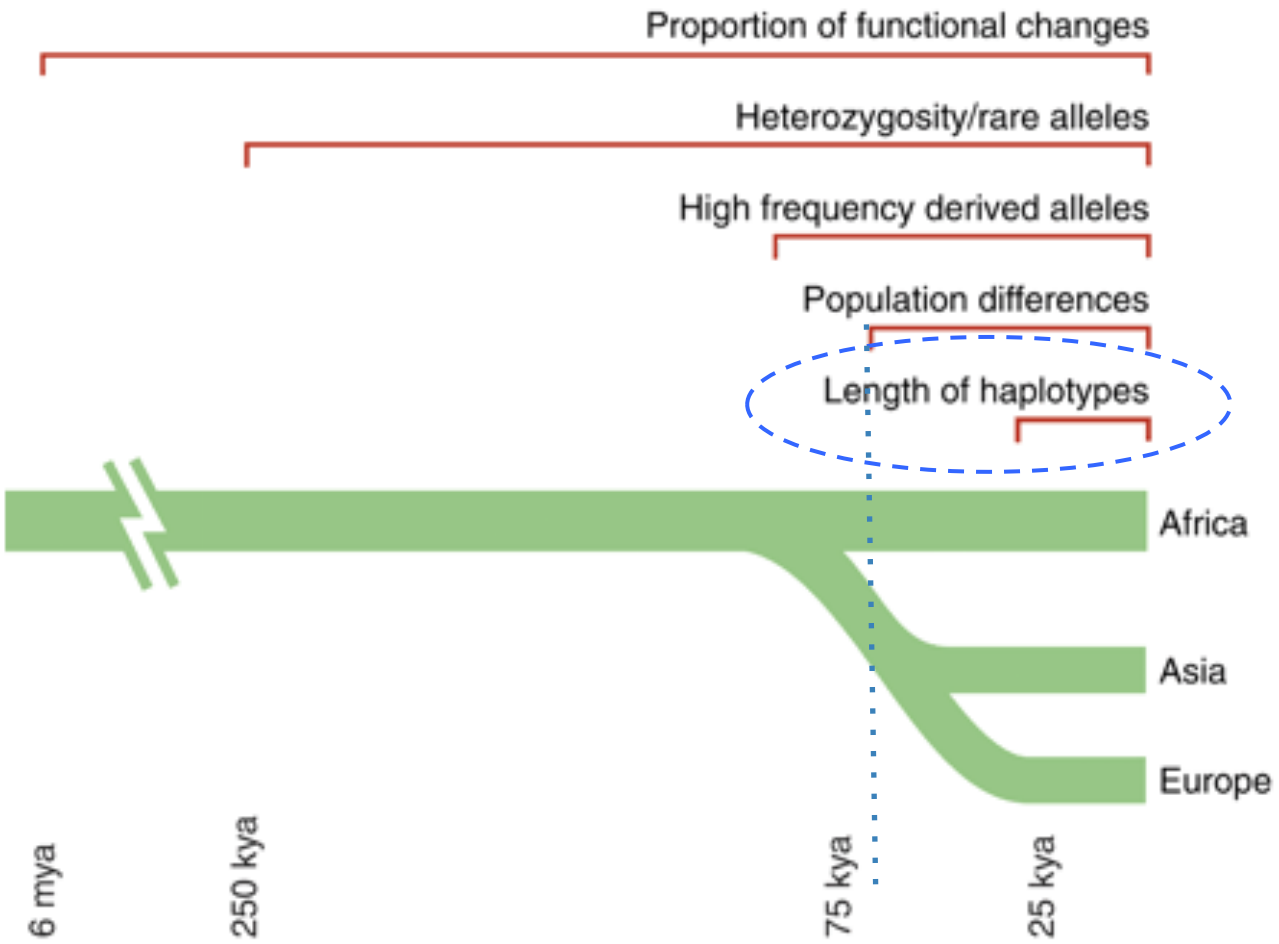
Population Branch Statistic



Population Branch Statistic (PBS) ≈ $T_{12} + T_{13} - T_{23}$

Yi et al. Science 2010

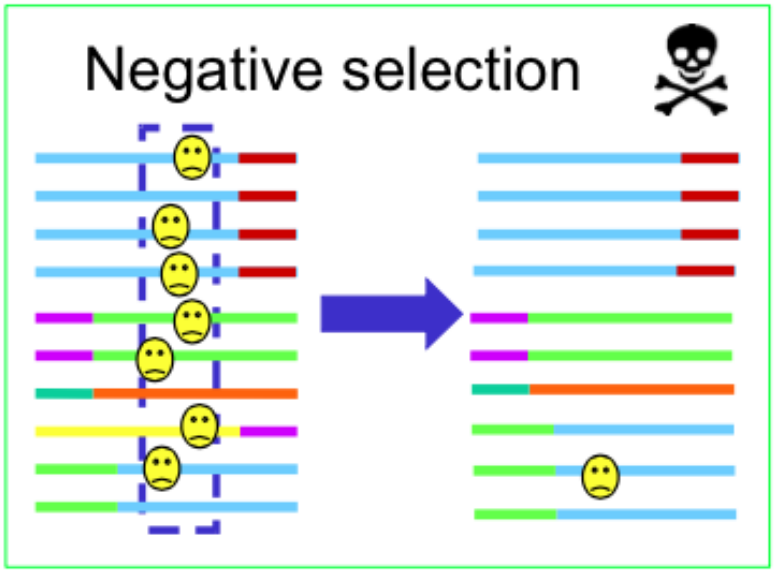
Methods to infer recent selection

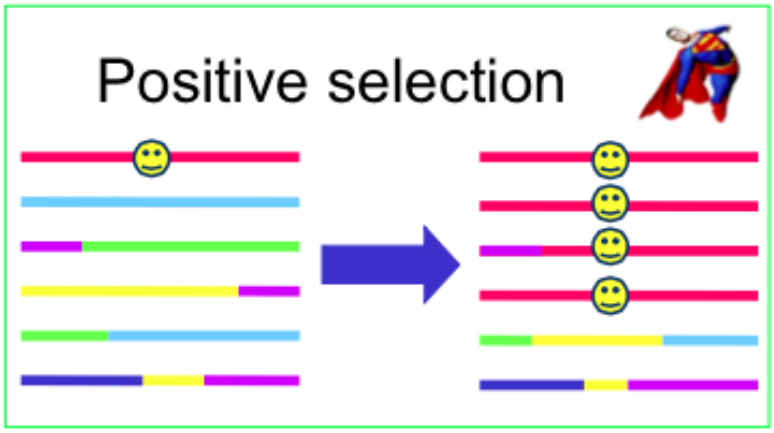


Sabeti et al. 2006 Science

Selective sweep

Effect of genetic hitchhiking

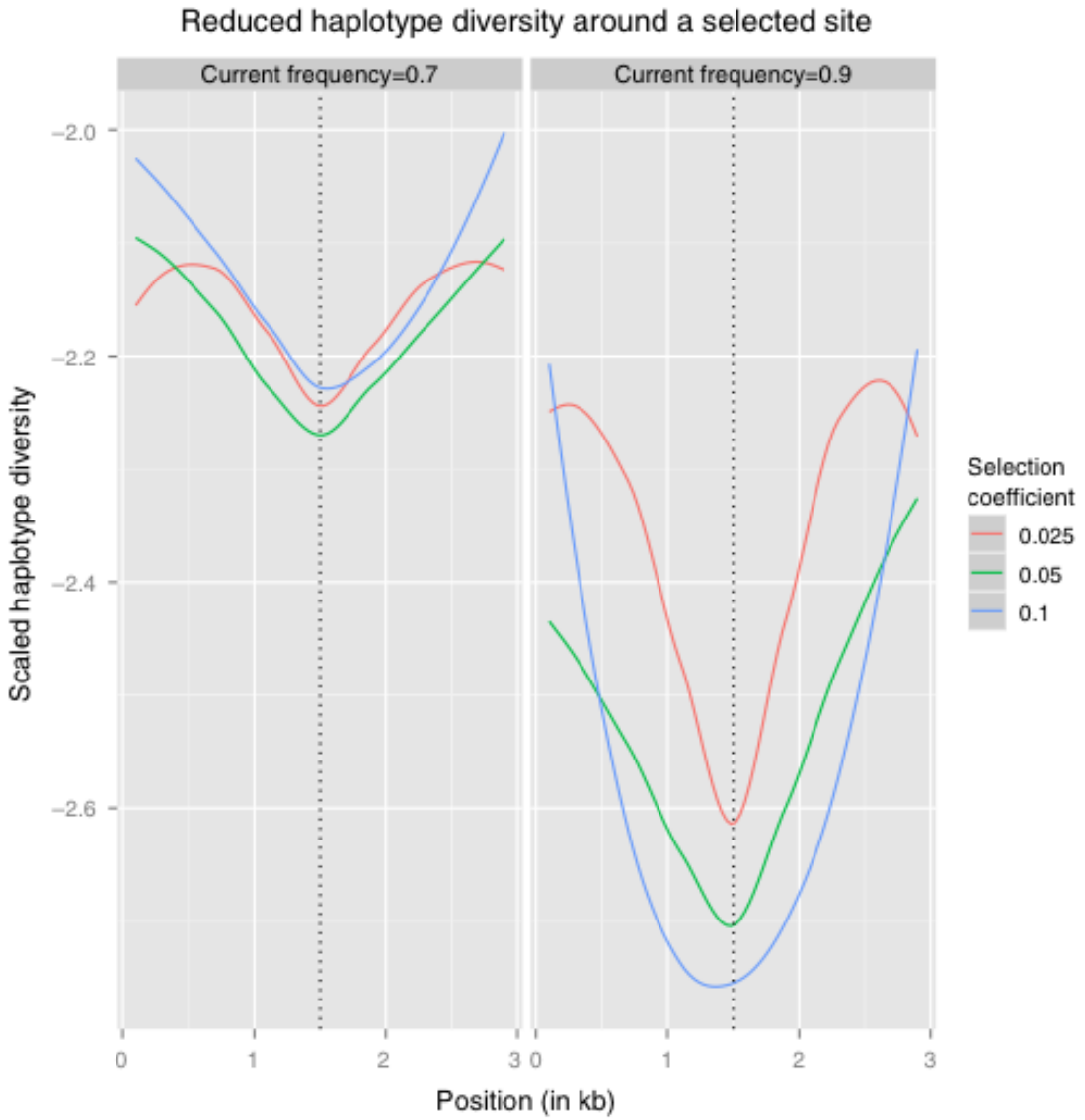




- Selective sweep: the beneficial allele is driven to fixation
- Partial sweep: when the beneficial allele hasn't reached fixation yet.

Haplotype diversity under selection

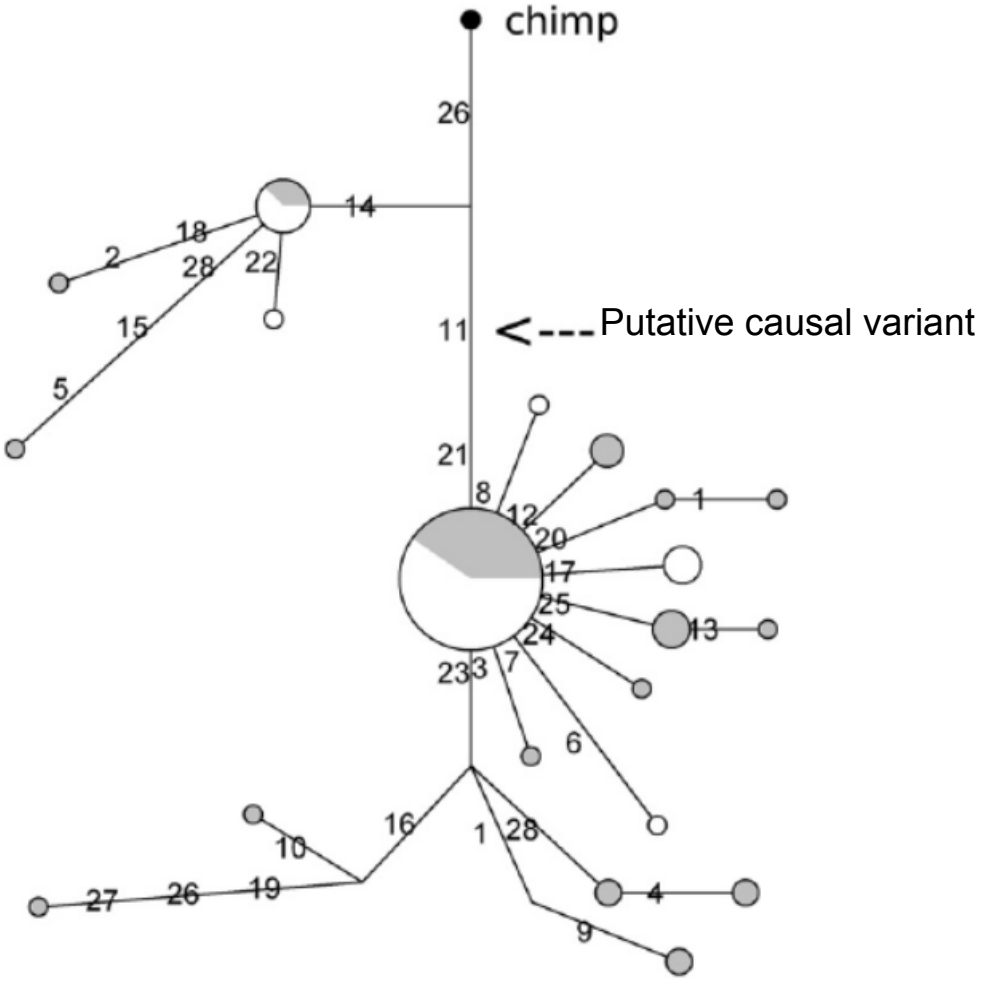
- Reduction of polymorphisms
- Decay of haplotype diversity is proportional to selective strength
- Recombination breaks the signal
- The effect is stronger in proximity of the selected site



Fumagalli and Sironi. 2014 Current Opinion in Immunology

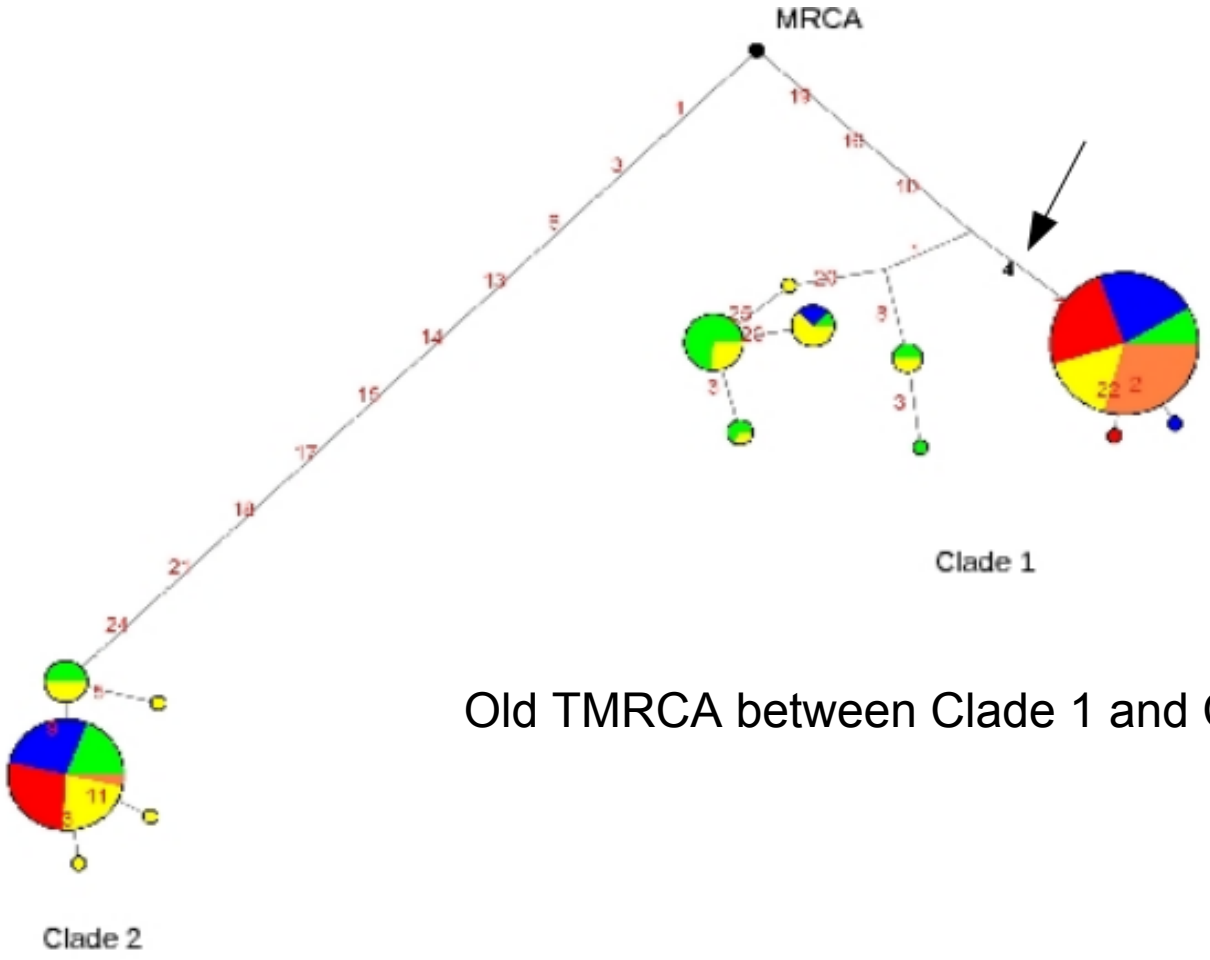
Haplotype genealogy

Median-joining network



Cagliani et al. AJHG. 2009

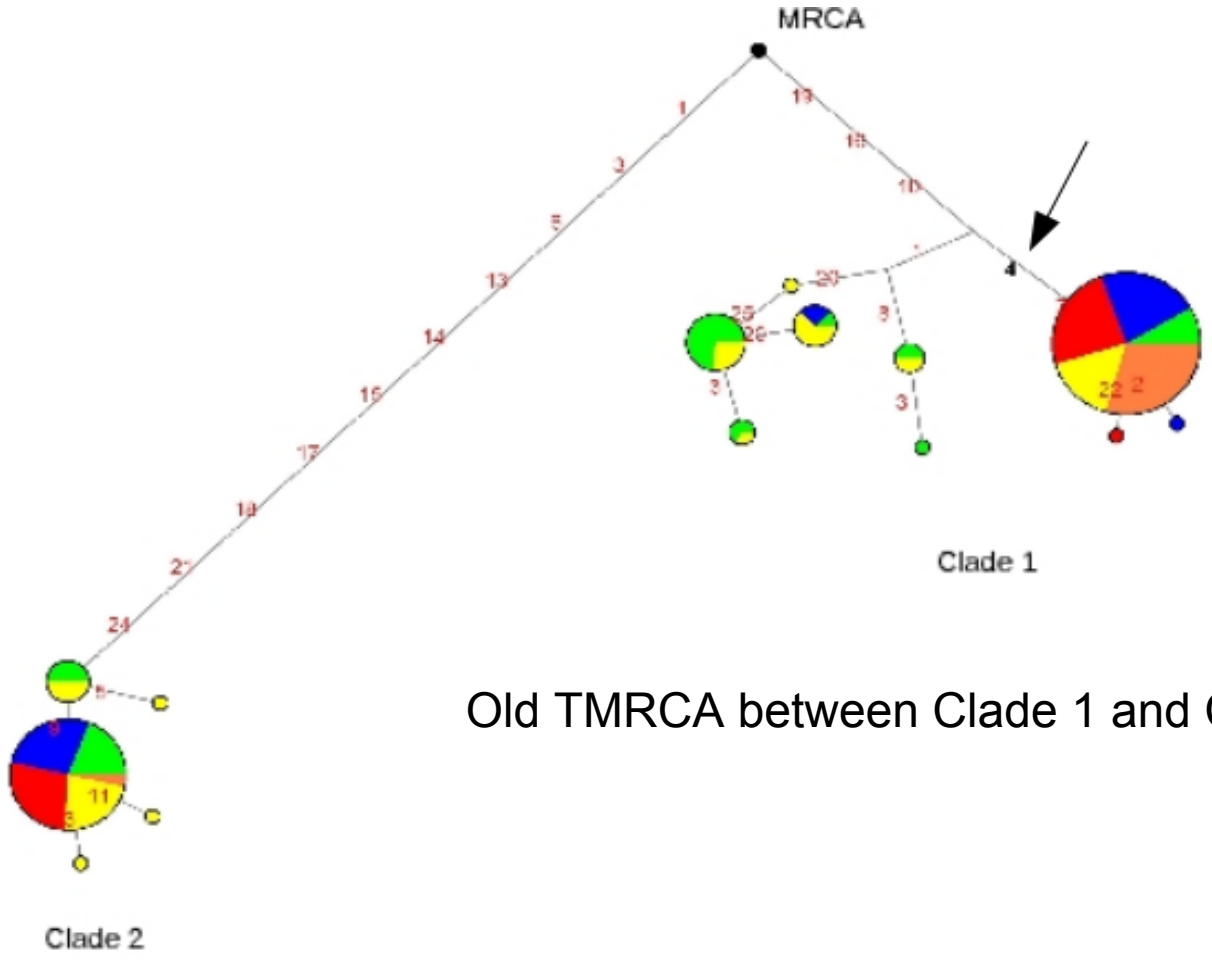
Haplotype genealogy (under ? selection)



Old TMRCA between Clade 1 and Clade 2

Cagliani et al. BMC Evol Biol. 2011

Haplotype genealogy (under balancing selection)

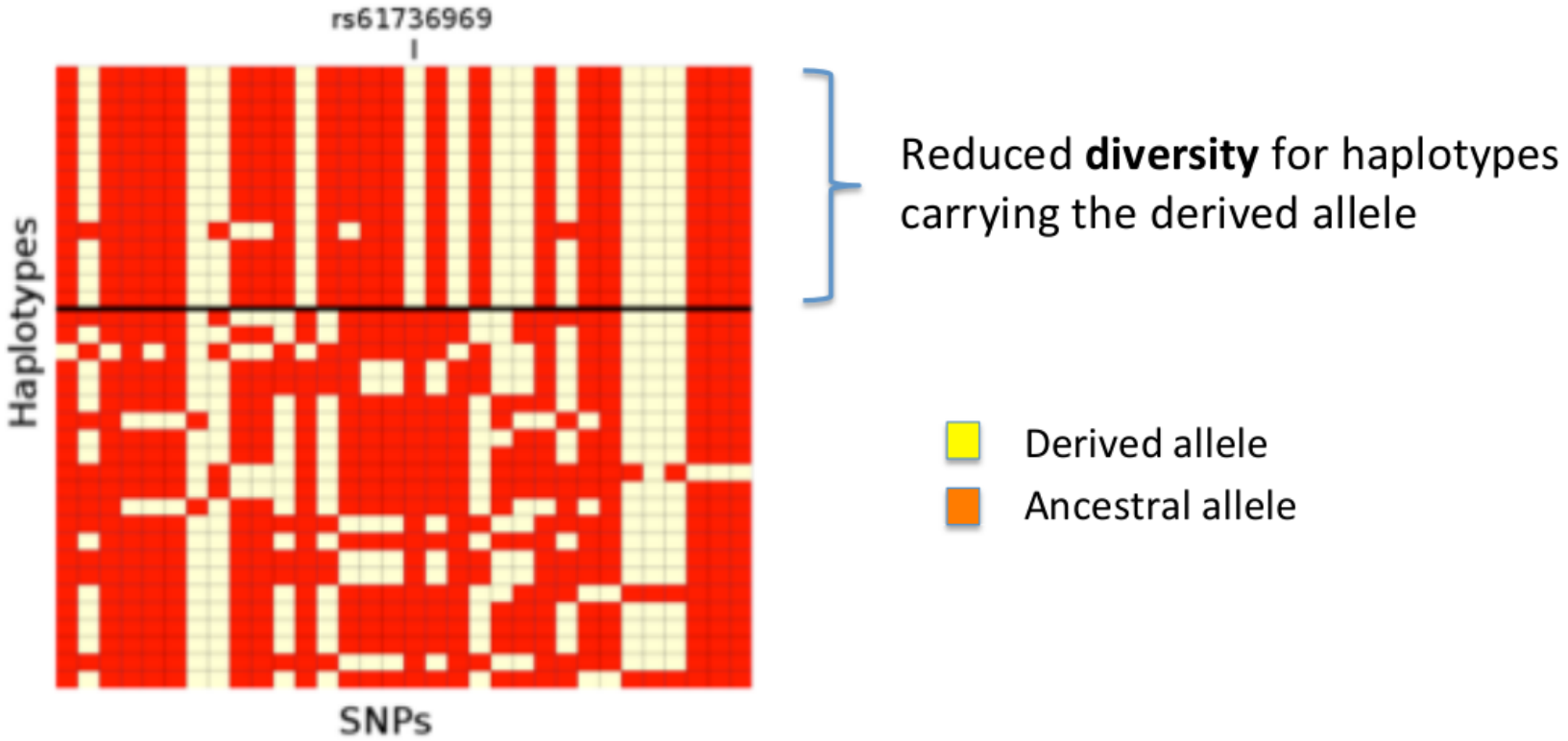


Old TMRCA between Clade 1 and Clade 2

Cagliani et al. BMC Evol Biol. 2011

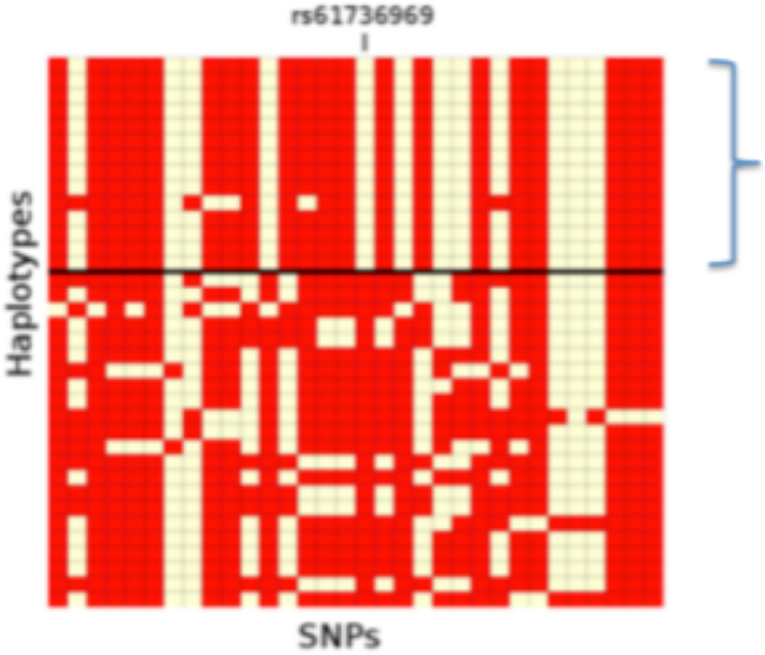
Haplotype diversity

Quantification of the haplotype diversity



Moltke et al. Nature 2014

Haplotype diversity



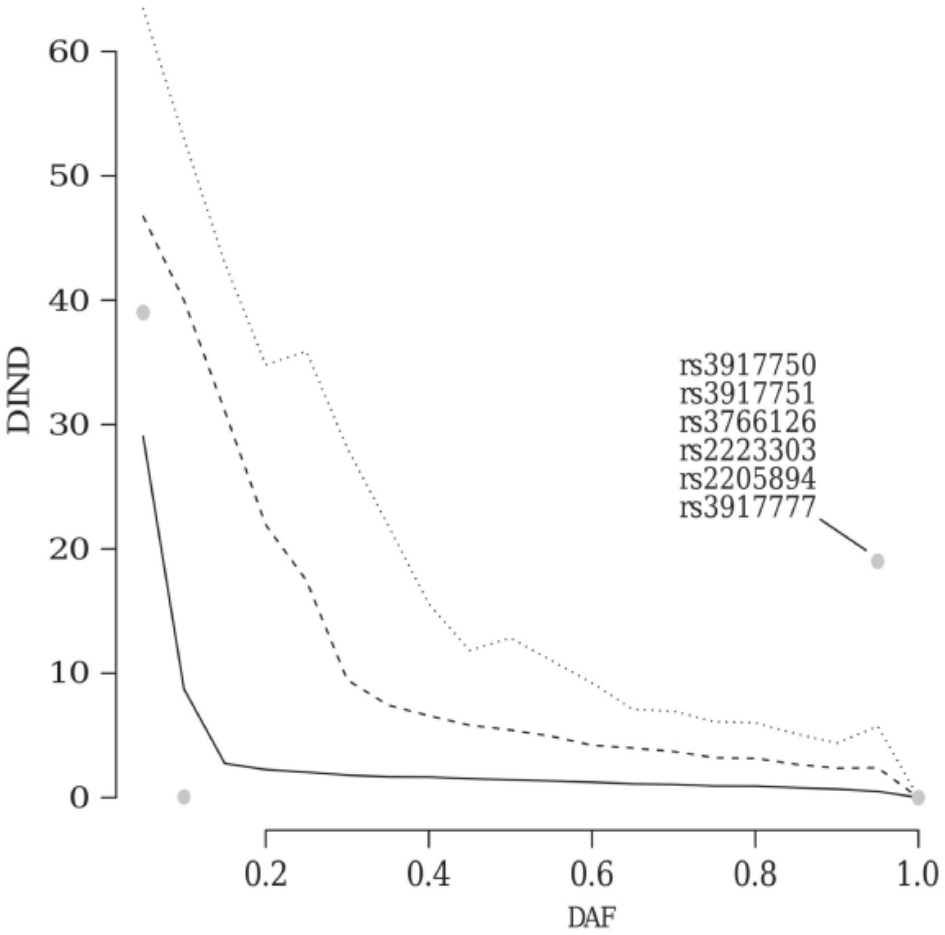
DIND (Derived Intra-allelic Nucleotide Diversity):

$$i\pi_{D,x} = \frac{\sum_{k=1}^{n_D-1} \sum_{l=k+1}^{n_D} d_{kl}}{{}_{n_D}C_2}$$

DIND ~ Diversity of haplotypes carrying the derived allele / Diversity of haplotypes carrying the ancestral allele

Barreiro et al. Plos genetics 2009

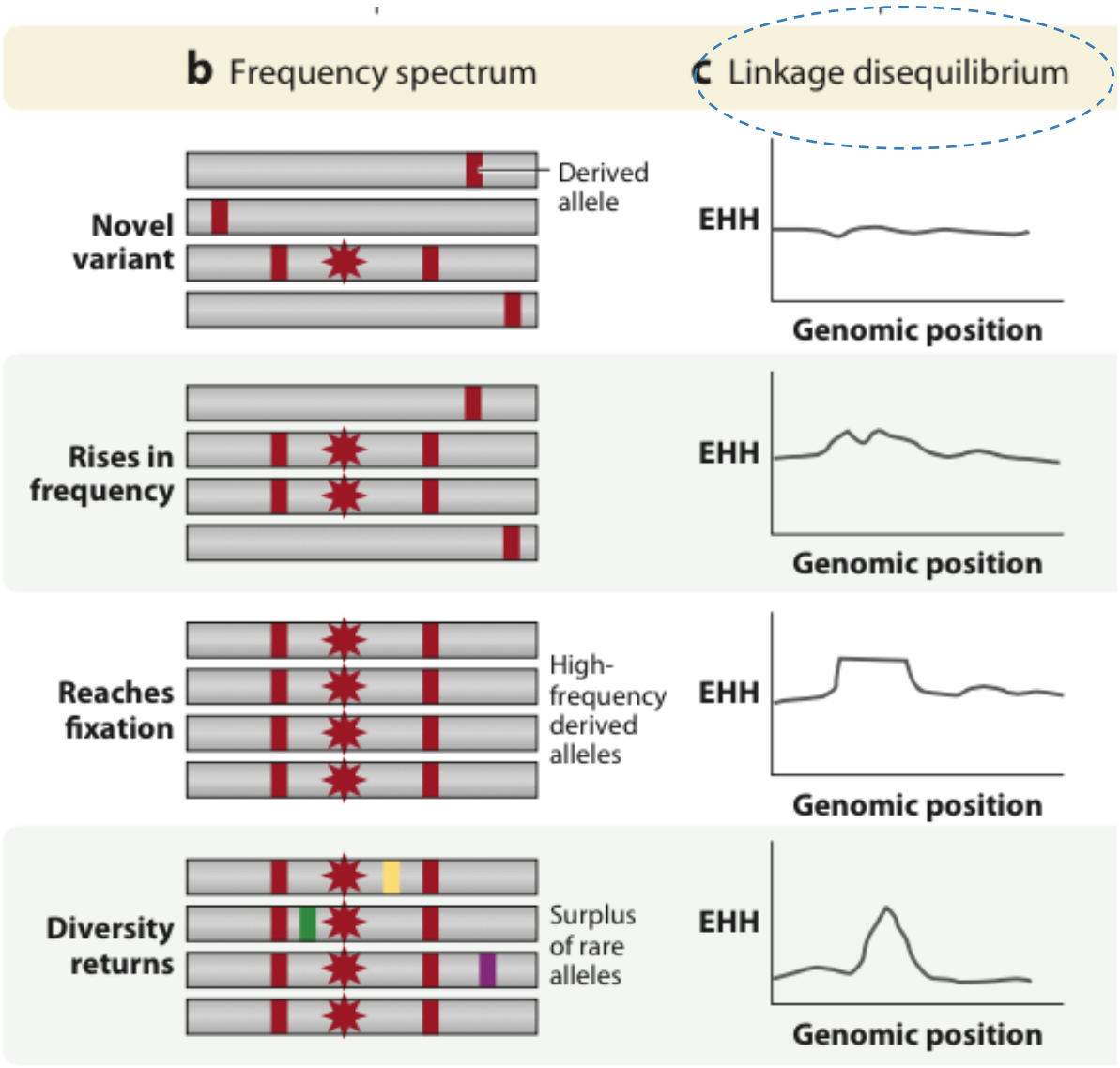
DIND



DIND is useful to identify the causal variant by looking at outliers in the DIND/DAF distribution

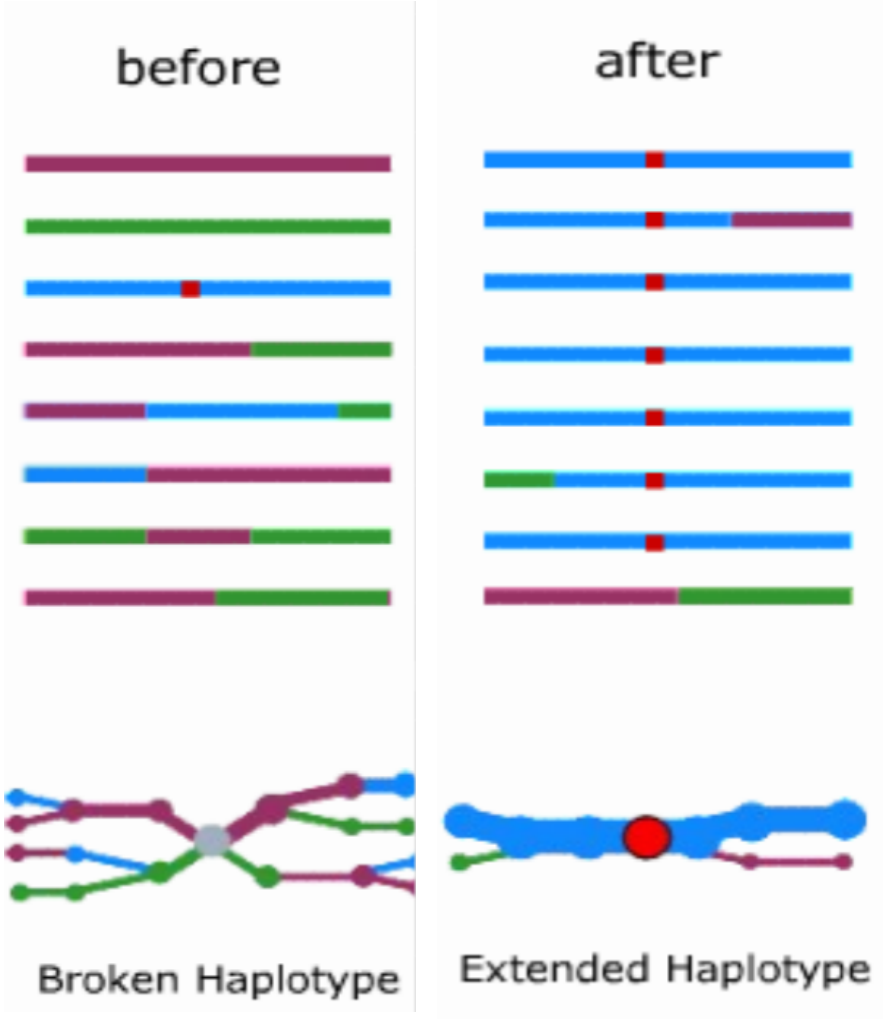
Fumagalli et al Heredity 2012

Extended Linkage Disequilibrium



Vitti et al. Annu Rev Genet. 2013

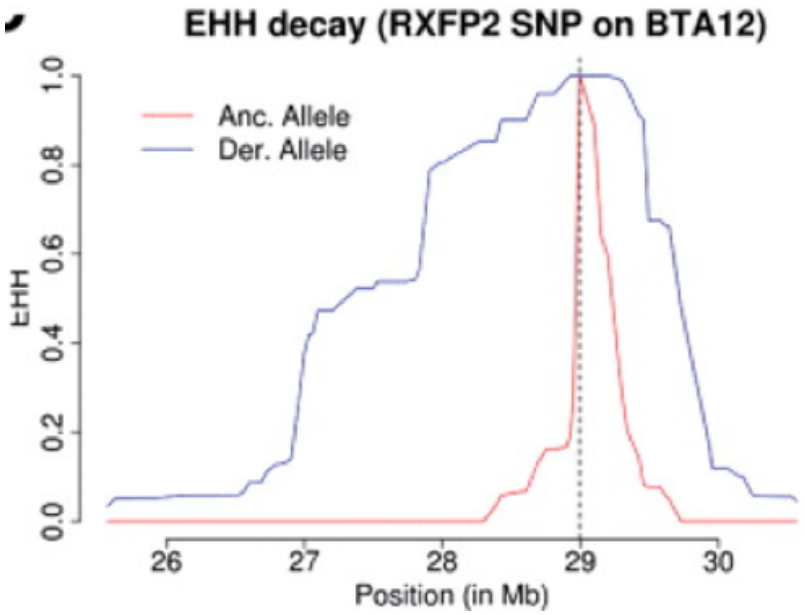
Extended Haplotype Homozygosity (EHH)



LD-based approaches:

- ❏ identifying variants under partial or incomplete selective sweep
- ❏ EHH from a core allele to a specified distance
- ❏ EHH decreases with distance by the action of recombination

EHH-derived statistics



- **iHS** compares the area under the curve defined by EHH for the derived and ancestral variants as one travels further in genetic distance from the core region
- cross-population extended haplotype homozygosity (**XP-EHH**) compares haplotype lengths between populations.

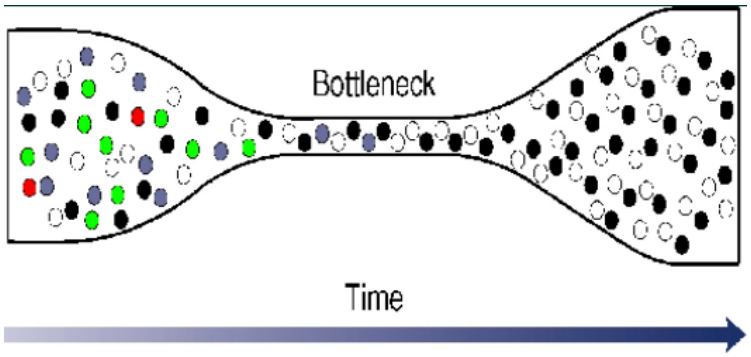
Gautier and Vitalis. Bioinformatics 2012

Software available

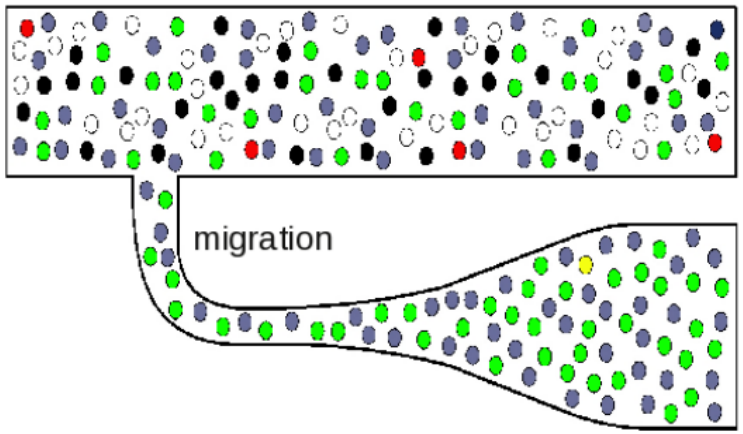
- DnaSP (http://www.ub.edu/dnasp/)
- Arlequin (http://cmpg.unibe.ch/software/arlequin35/)
- BayeScan (http://cmpg.unibe.ch/software/BayeScan/)
- libsequence (http://www.molpopgen.org/software.html)
- sweep (http://www.broadinstitute.org/mpg/sweep/)
- iHS (http://coruscant.itmat.upenn.edu/software.html)
- nSL (http://cteg.berkeley.edu/~nielsen/resources/software/)
- Pre-computed values (USCS genome browser tables)
- Homemade scripts
- …

Neutral confounding factors

Various demographic scenarios may result in the rejection of the null hypothesis of neutrality.



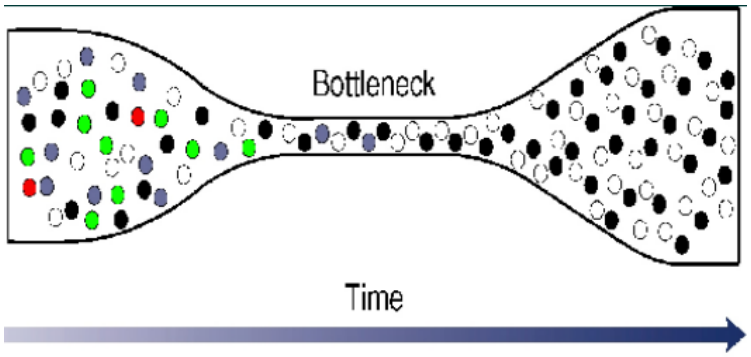
?



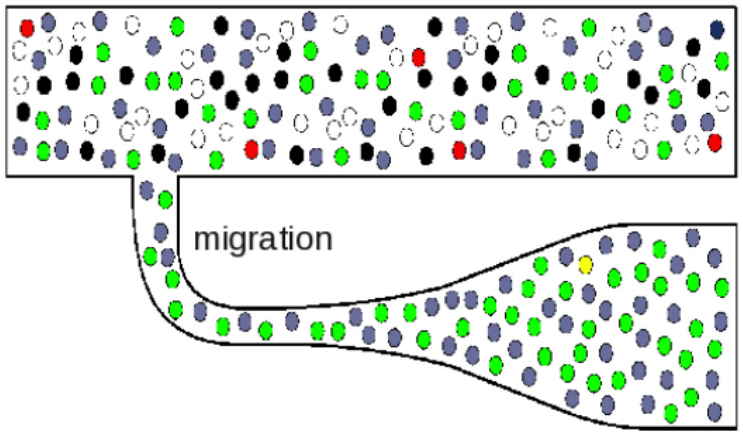
?

Neutral confounding factors

Various demographic scenarios may result in the rejection of the null hypothesis of neutrality.



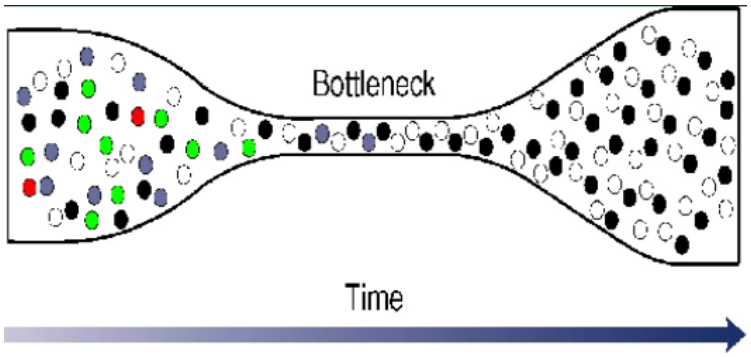
- Bottleneck: excess of intermediate frequency alleles



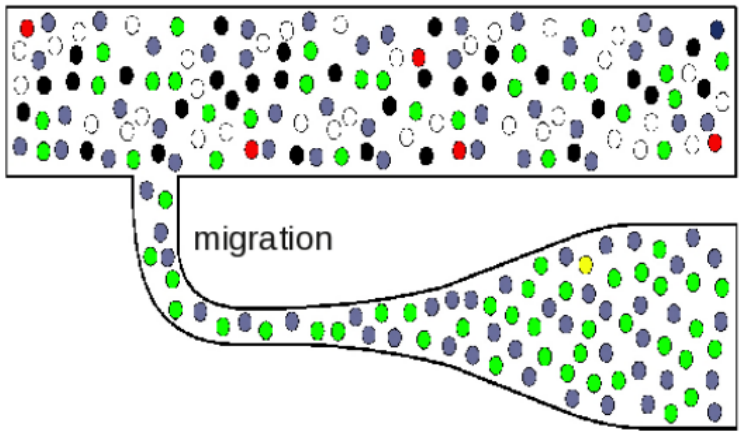
- Founder effect: decreased diversity
- Expansion: ?

Neutral confounding factors

Various demographic scenarios may result in the rejection of the null hypothesis of neutrality.



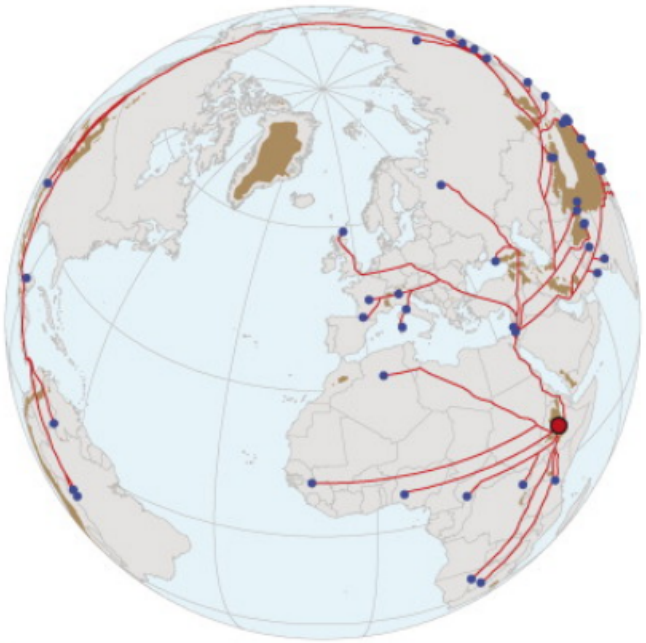
- Bottleneck: excess of intermediate frequency alleles

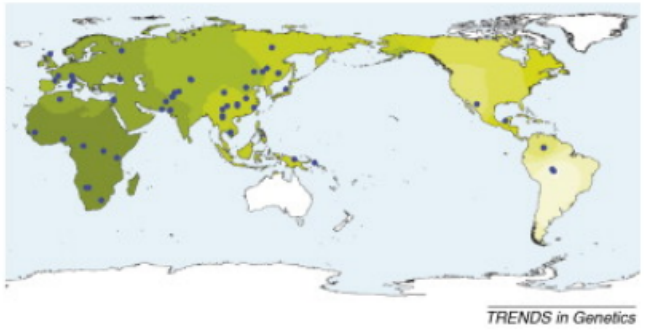


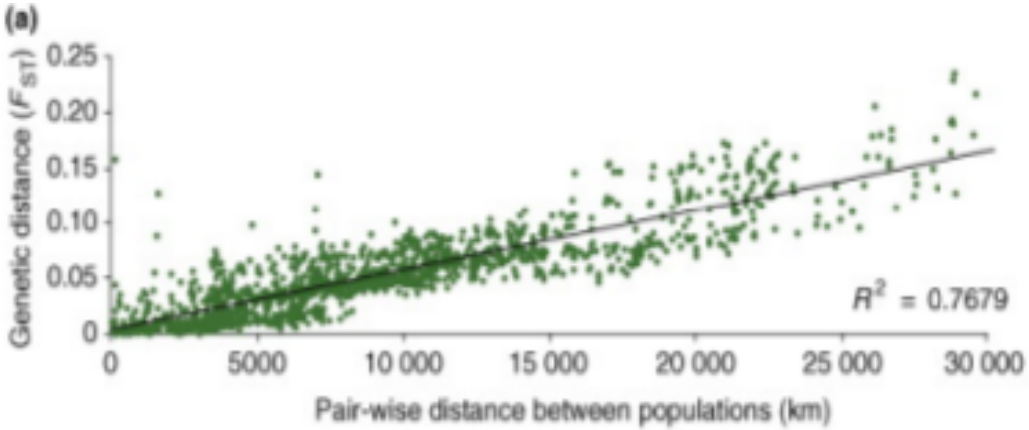
- Founder effect: decreased diversity
- Expansion: excess of low frequency alleles

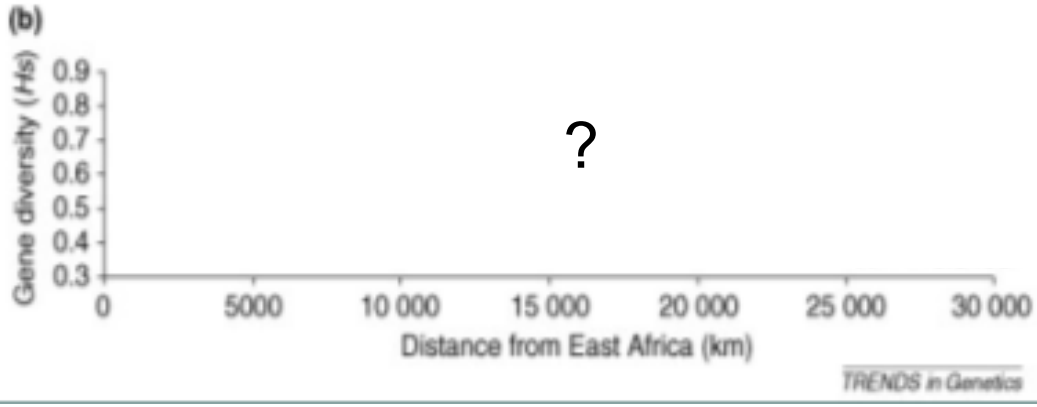
Isolation by distance

- Human populations are not independent.
- Global human genetic variation is greatly affected by geography.







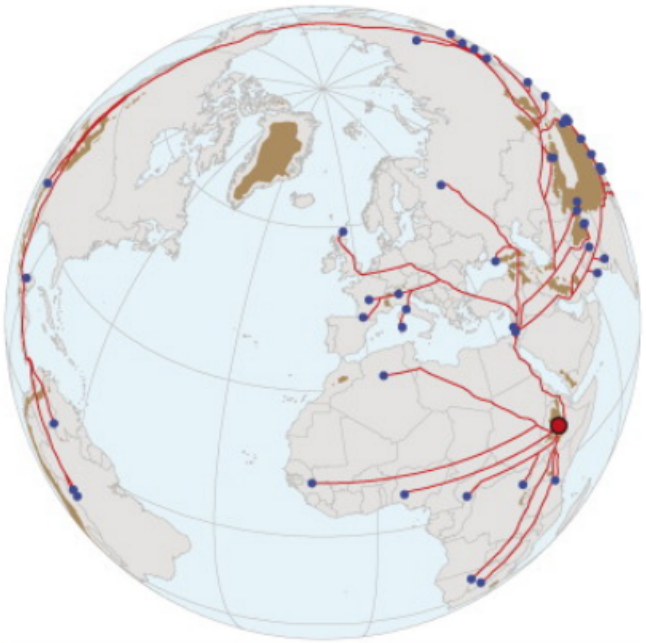


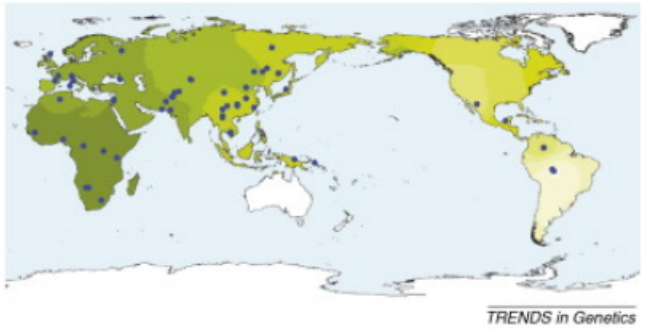
?

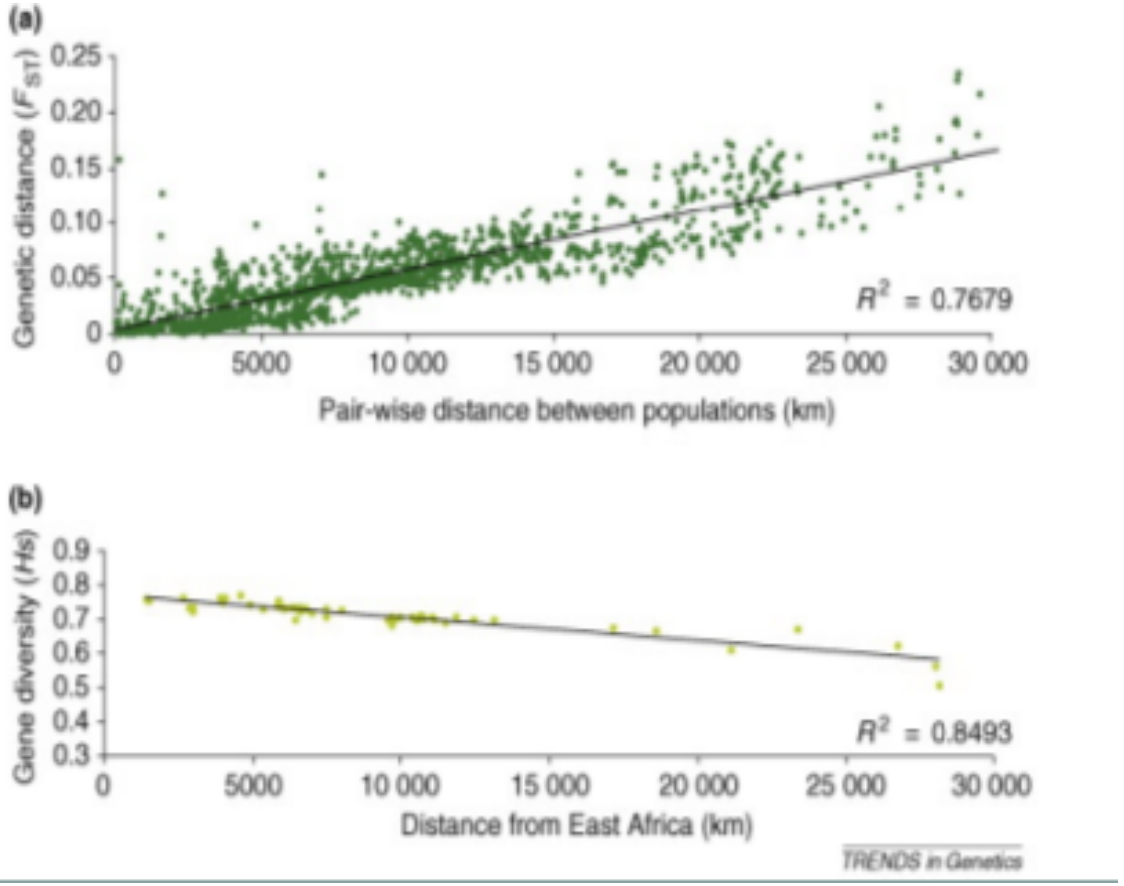
Prugnolle et al. Trends in Genetics 2007

Isolation by distance

- Human populations are not independent.
- Global human genetic variation is greatly affected by geography.







Prugnolle et al. Trends in Genetics 2007

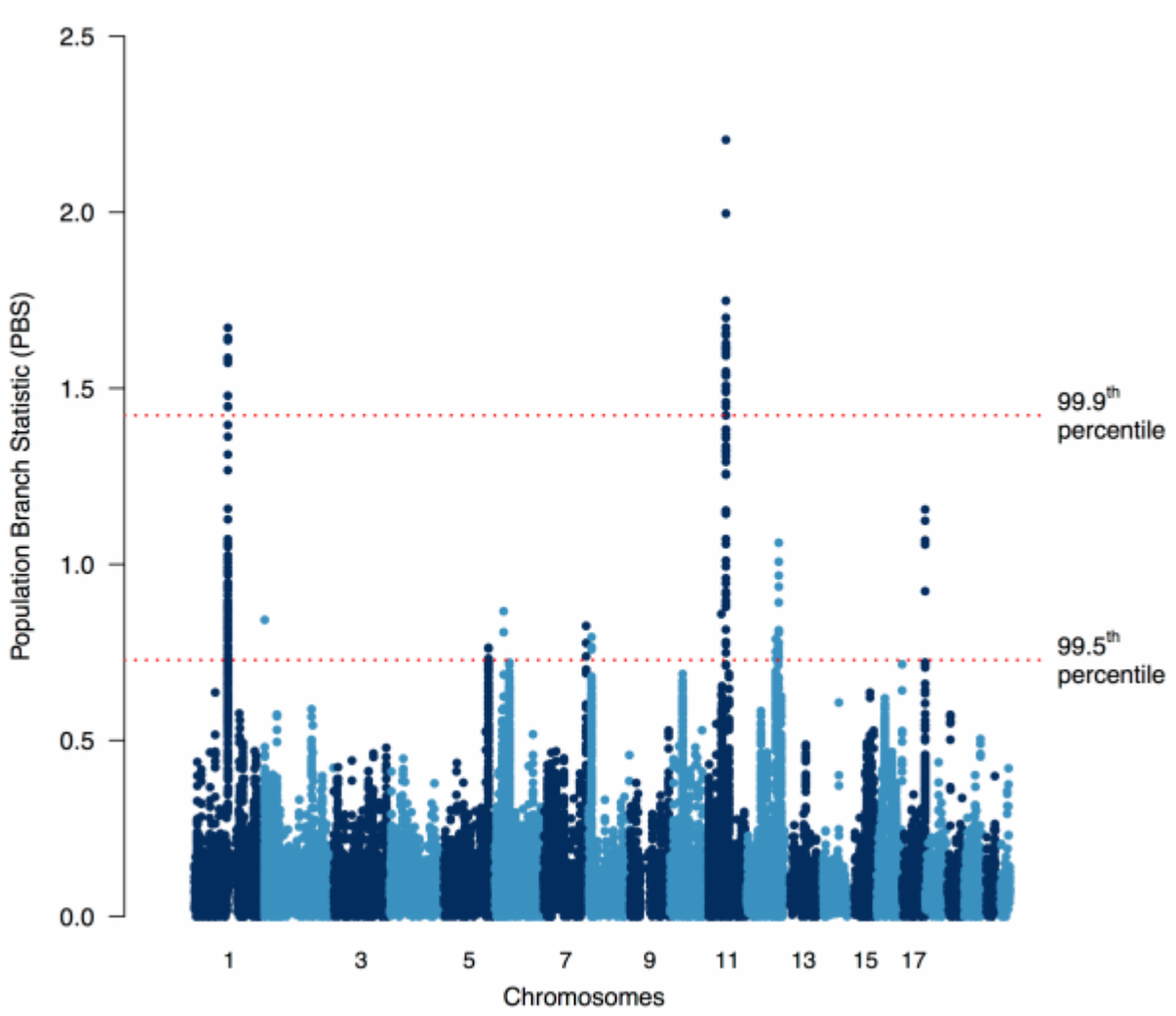
Assessing statistical significance

1) **Comparison** with other loci in the genome. While selection acts on a locus, demography equally affects the whole genome. We can calculate the percentile rank of the statistic for the gene/region of interest.

2)

Empirical distribution

Outlier approach



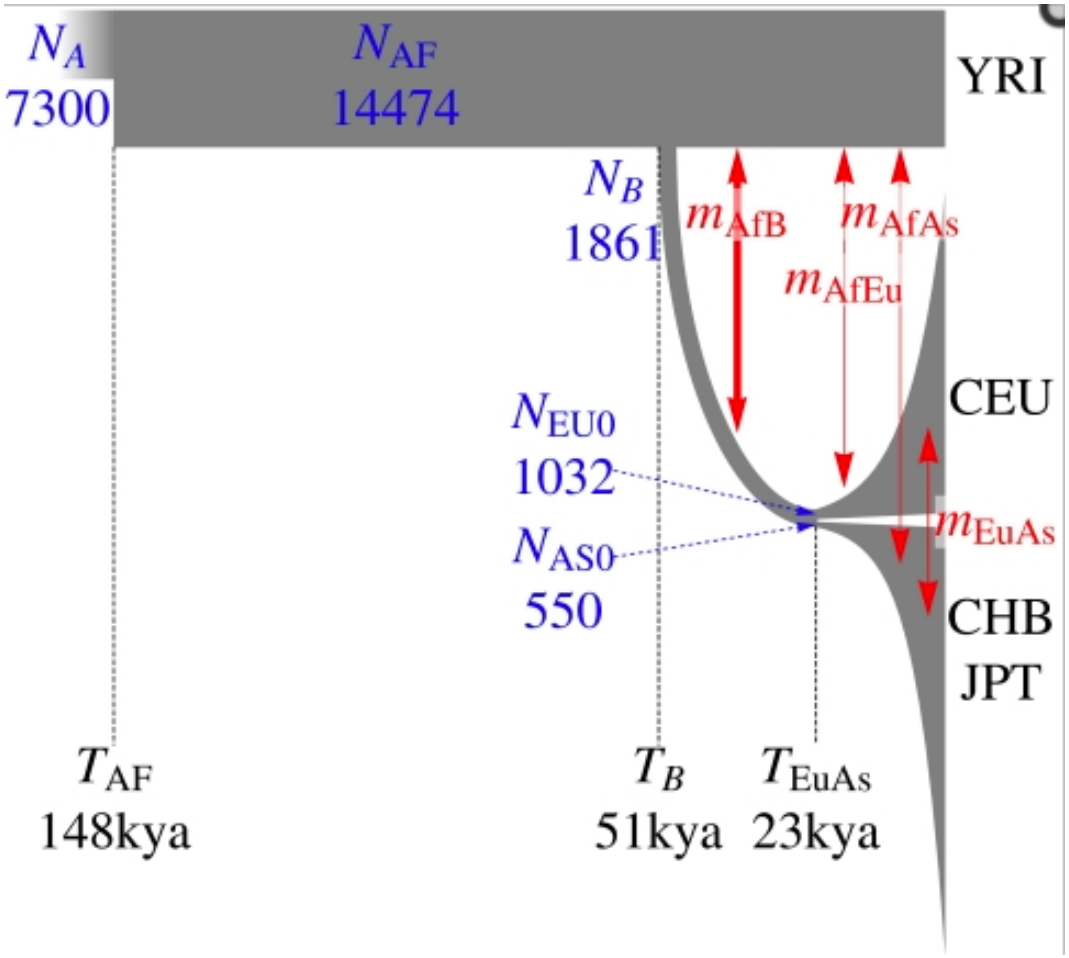
Assessing statistical significance

1) **Comparison** with other loci in the genome. While selection acts on a locus, demography equally affects the whole genome. We can calculate the percentile rank of the statistic for the gene/region of interest.

2) **Simulations**. Coalescence is the process in which, looking backwards in time, the genealogies of alleles merge at a common ancestor. We can incorporate a demographic model for species' evolution.

Demographic model

Simulations under a neutral demographic model



Gravel et al. PNAS 2011

Adaptation to environments

Detect putative loci targeted by natural selection showing extreme or unusual values of statistics based on site frequencies spectrum, haplotypes structure, population genetic differentiation.

But:

1) **power** to detect a selective event only if the process is recent or strong enough or the beneficial allele is nearly fixed to leave a clear footprint;

2) unable to underline the **environmental** factor (pathogens, diet or climate) that acted as selective **pressure**.

Adaptation to environments

But:

1) **power** to detect a selective event only if the process is recent or strong enough or the beneficial allele is nearly fixed to leave a clear footprint;

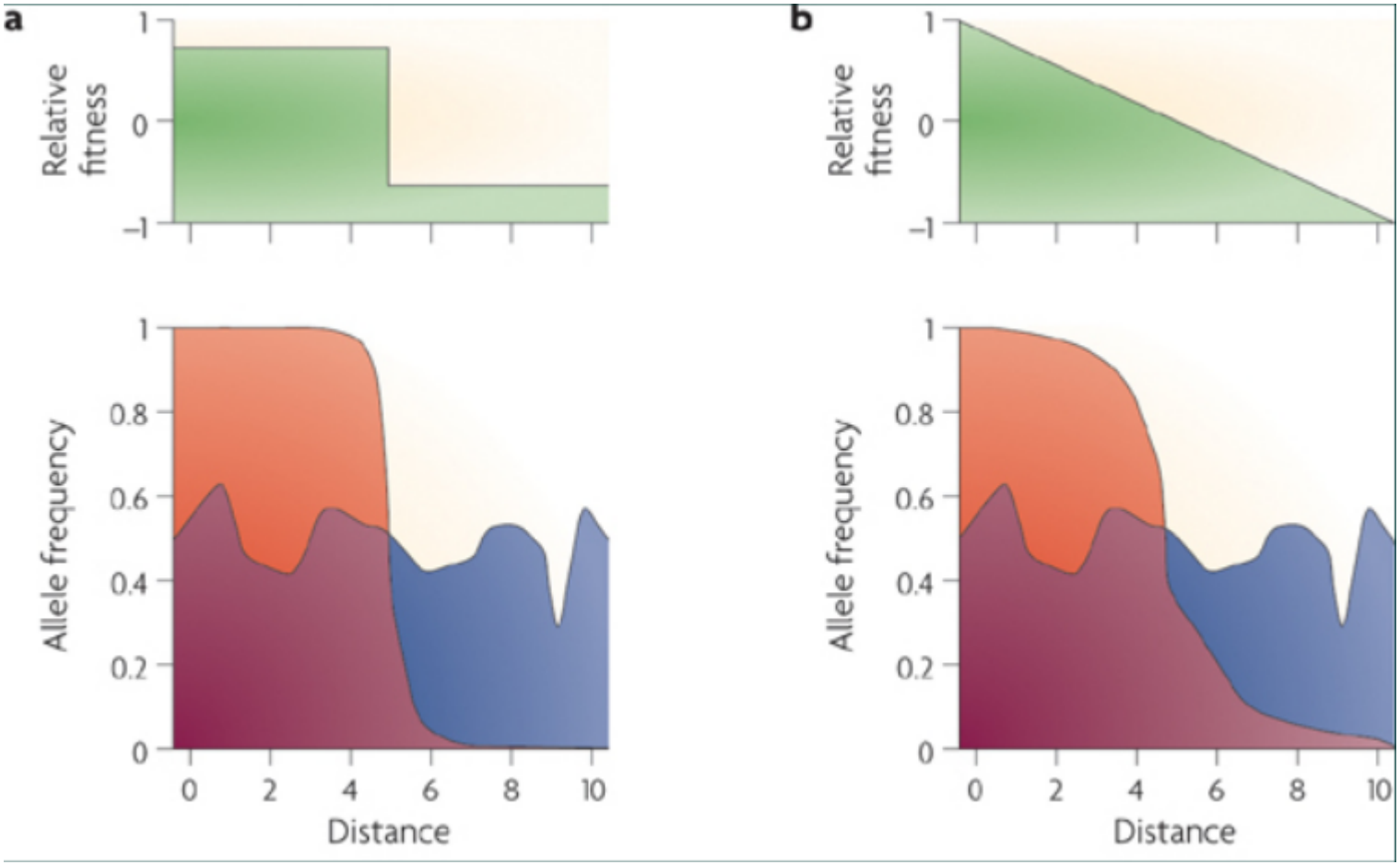
2) unable to underline the **environmental** factor (pathogens, diet or climate) that acted as selective **pressure**.

Solutions:

1) Alternative models might better explain the complex signatures of adaptation (selective sweeps from standing variation or polygenic adaptation)

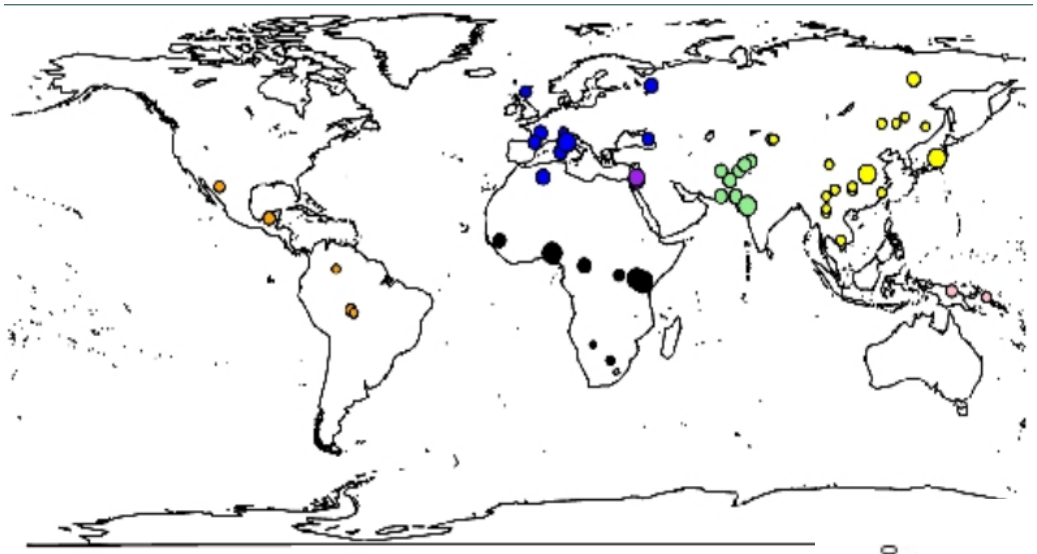
2) Methods detecting polymorphisms whose frequencies are greatly correlated with environmental variables.

Spatial variation



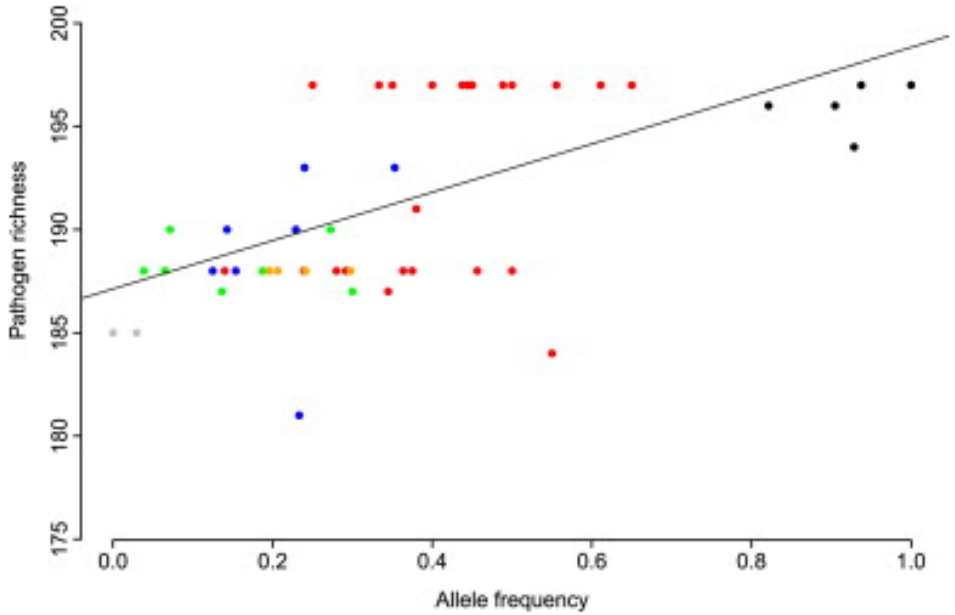
Novembre and Di Rienzo. Nat Rev Genet. 2009

Correlations



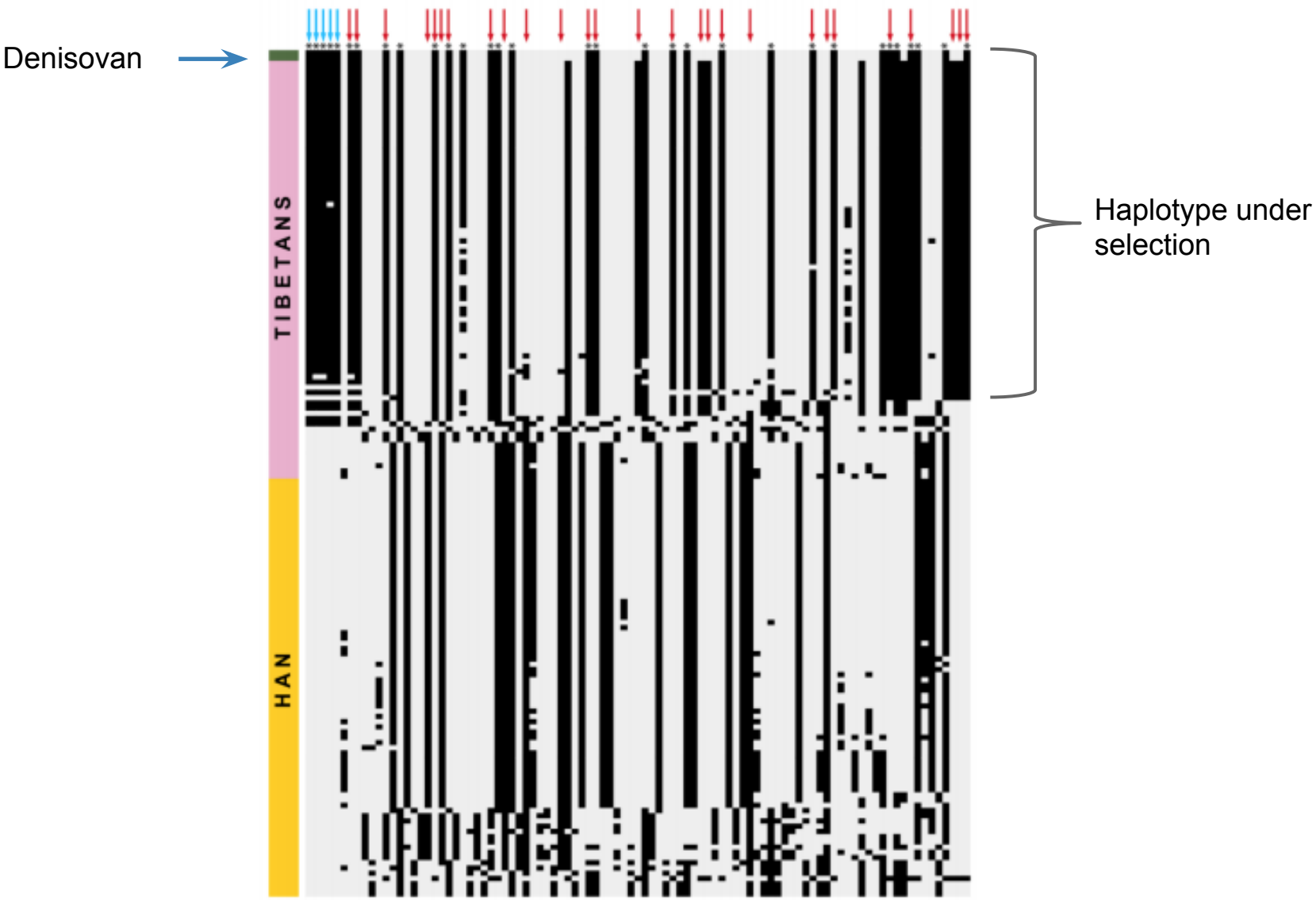
- Worldwide samples
- Ecological variables

- Mantel correlations
- Bayesian methods (BayEnv)
- Mixed models (LFMM)



Sironi and Clerici. Microbes Infect. 2010

Adaptive introgression



Huerta-Sanchez et al. Nature 2014

Adaptive admixture

Figure 2: Tibetans as a mixture of the HA-proxy and Han Chinese-related ancestral populations in the scaffold tree.

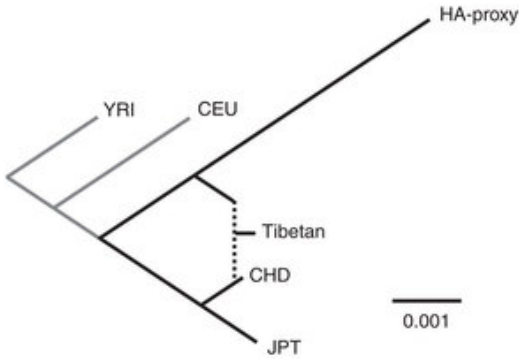
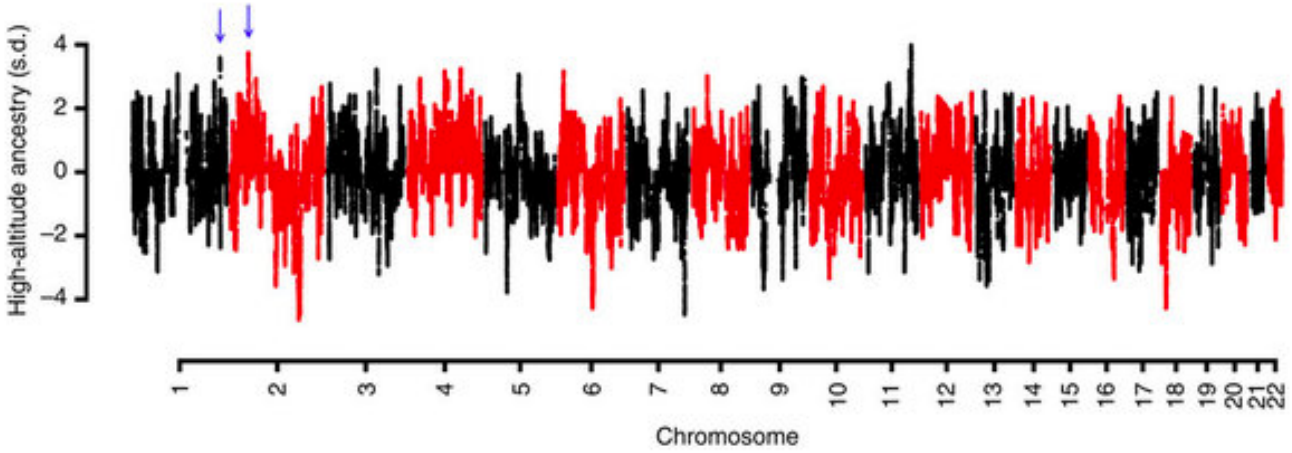


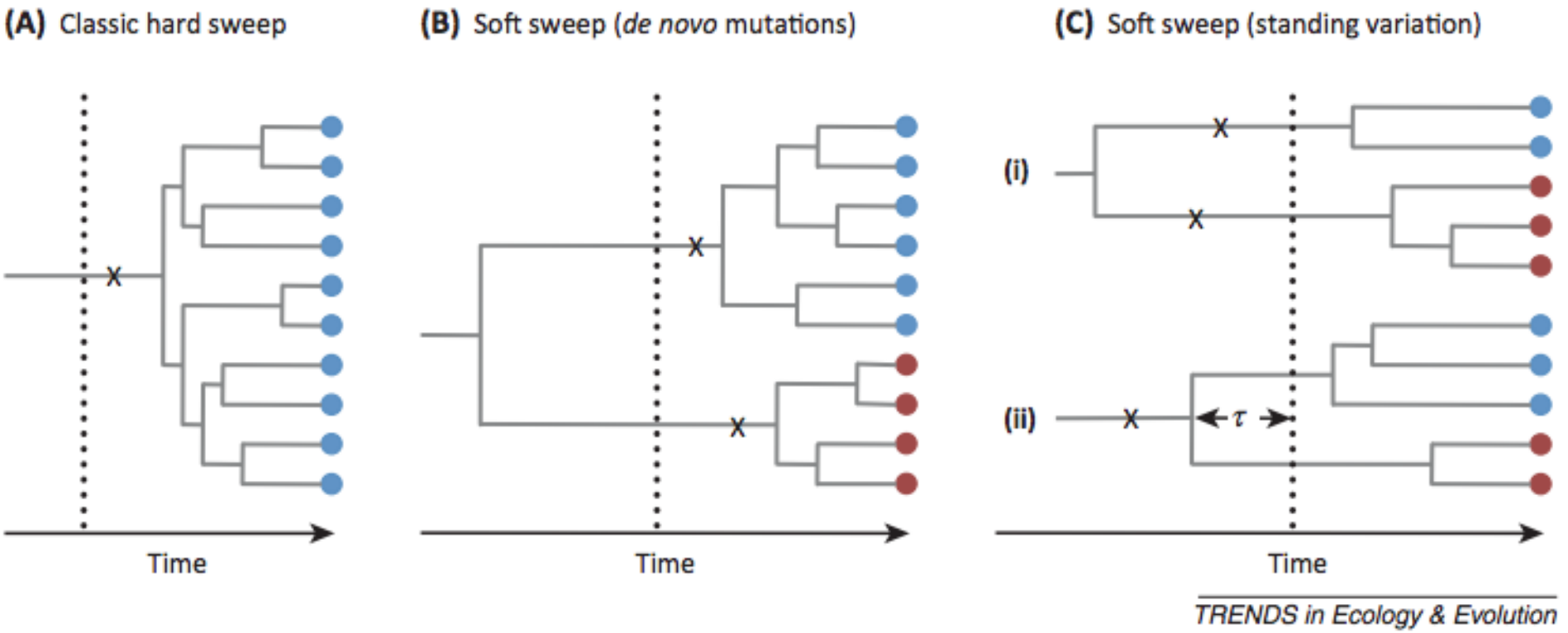
Figure 4: The distribution of high-altitude ancestry proportions across the Tibetan genome.



Blue arrows mark the positions of *EGLN1* (in chromosome 1) and *EPAS1* (in chromosome 2).

Jeong et al. Nature Comm. 2014

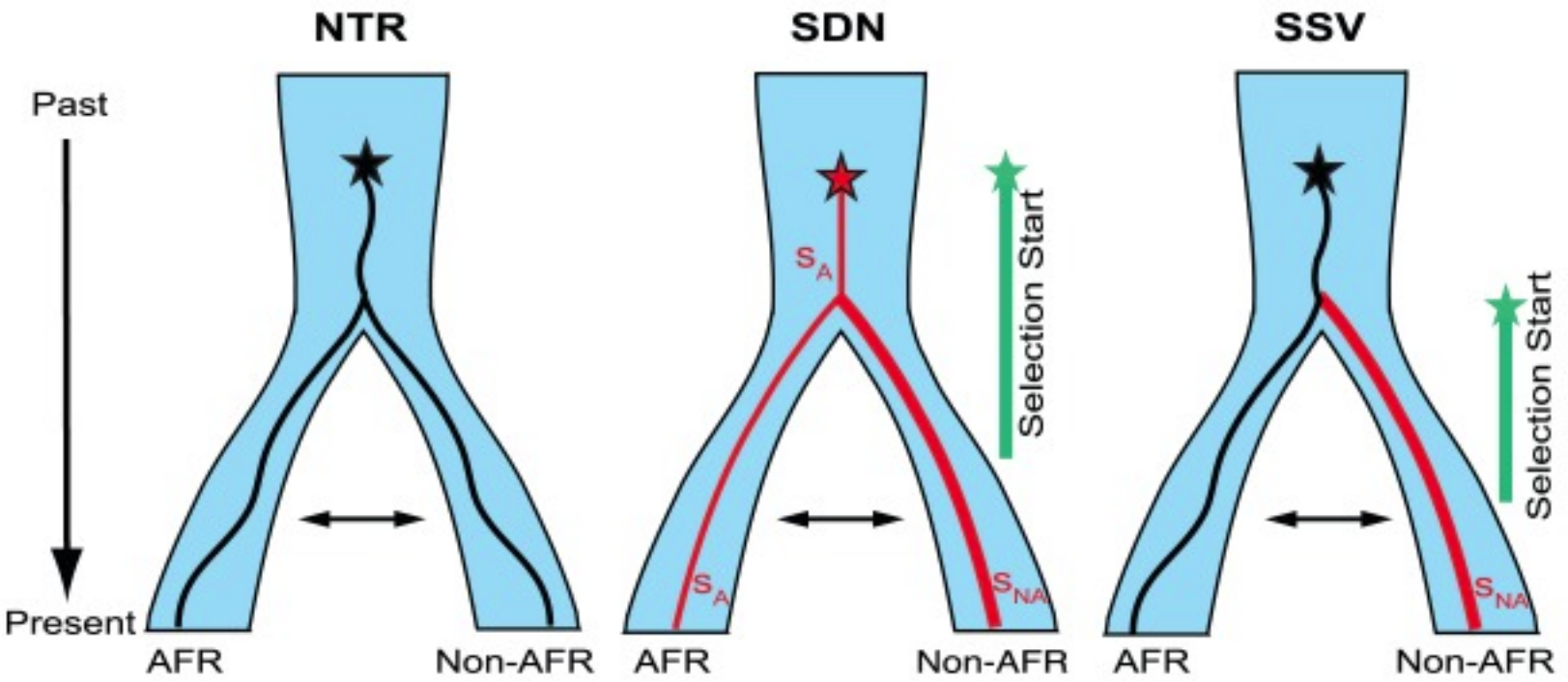
Soft sweep



Messer and Petrov. 2013

Simulation-based approaches

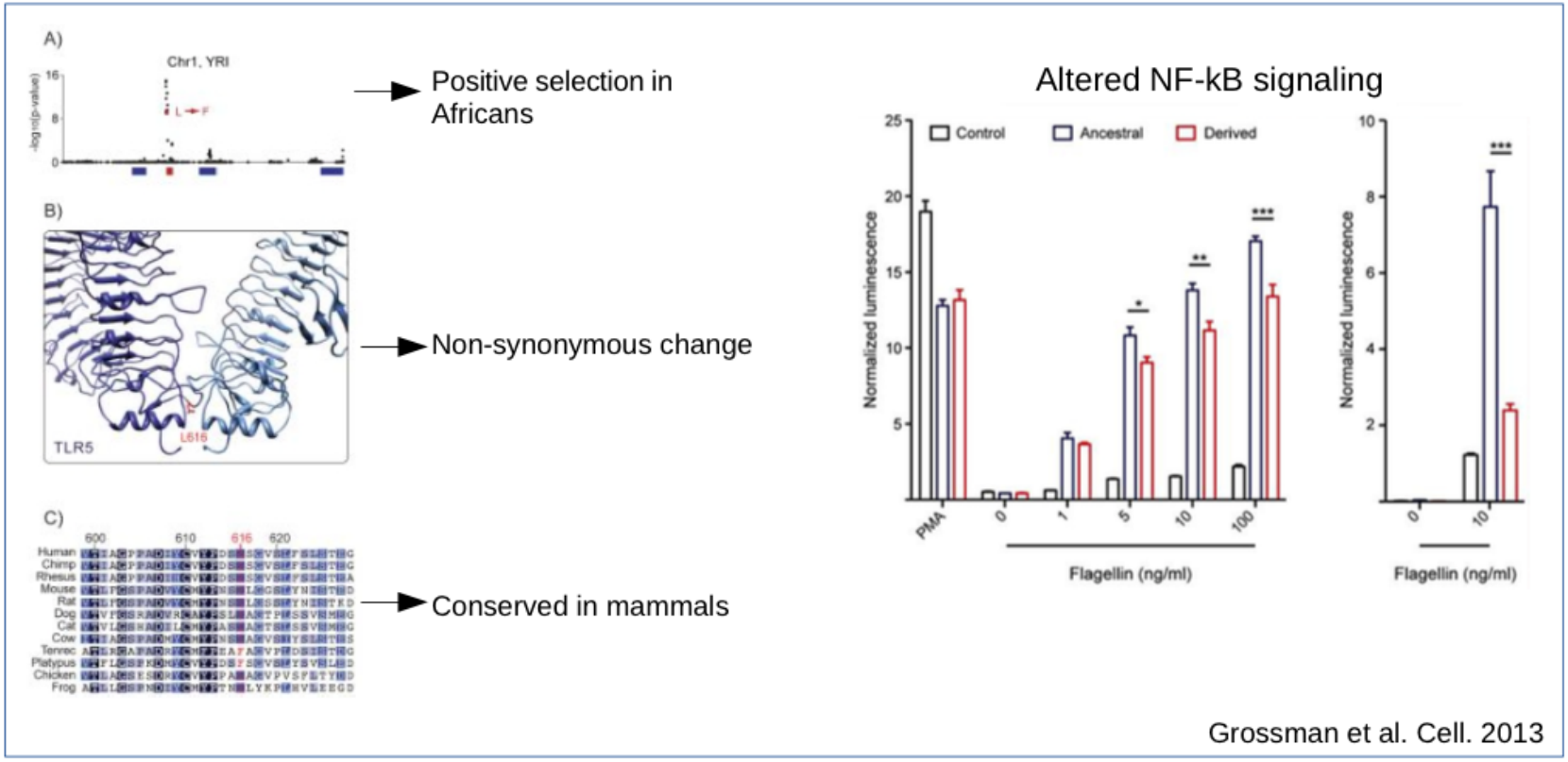
Approximate Bayesian Computation (ABC)



Key et al. PLoS Genetics 2014

Future perspectives

Integrative approach with functional validation of candidate adaptive variants



Paper discussion

Cell

Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears

Shiping Liu,[1,2,20] Eline D. Lorenzen,[3,4,20] Matteo Fumagalli,[3,20] Bo Li,[1,20] Kelley Harris,[5] Zijun Xiong,[1] Long Zhou,[1] Thorfinn Sand Korneliussen,[4] Mehmet Somel,[3,21] Courtney Babbitt,[6,7,22] Greg Wray,[6,7] Jianwen Li,[1] Weiming He,[1,2] Zhuo Wang,[1] Wenjing Fu,[1] Xueyan Xiang,[1,8] Claire C. Morgan,[9] Aoife Doherty,[10] Mary J. O'Connell,[9] James O. McInerney,[10] Erik W. Born,[11] Love Dalén,[12] Rune Dietz,[13] Ludovic Orlando,[4] Christian Sonne,[13] Guojie Zhang,[1,14] Rasmus Nielsen,[1,3,15,16,*] Eske Willerslev,[4,*] and Jun Wang[1,16,17,18,19,*]