Natural Selection

Ryan D. Hernandez

ryan.hernandez@me.com

qb3 ucb·ucsc·ucsf

UCSF University of California San Francisco

Department of Bioengineering and Therapeutic Sciences
a joint department of the UCSF Schools of Pharmacy and Medicine

McGill

The Effect of Positive Selection

Adaptive

Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious

The Effect of Positive Selection

Adaptive

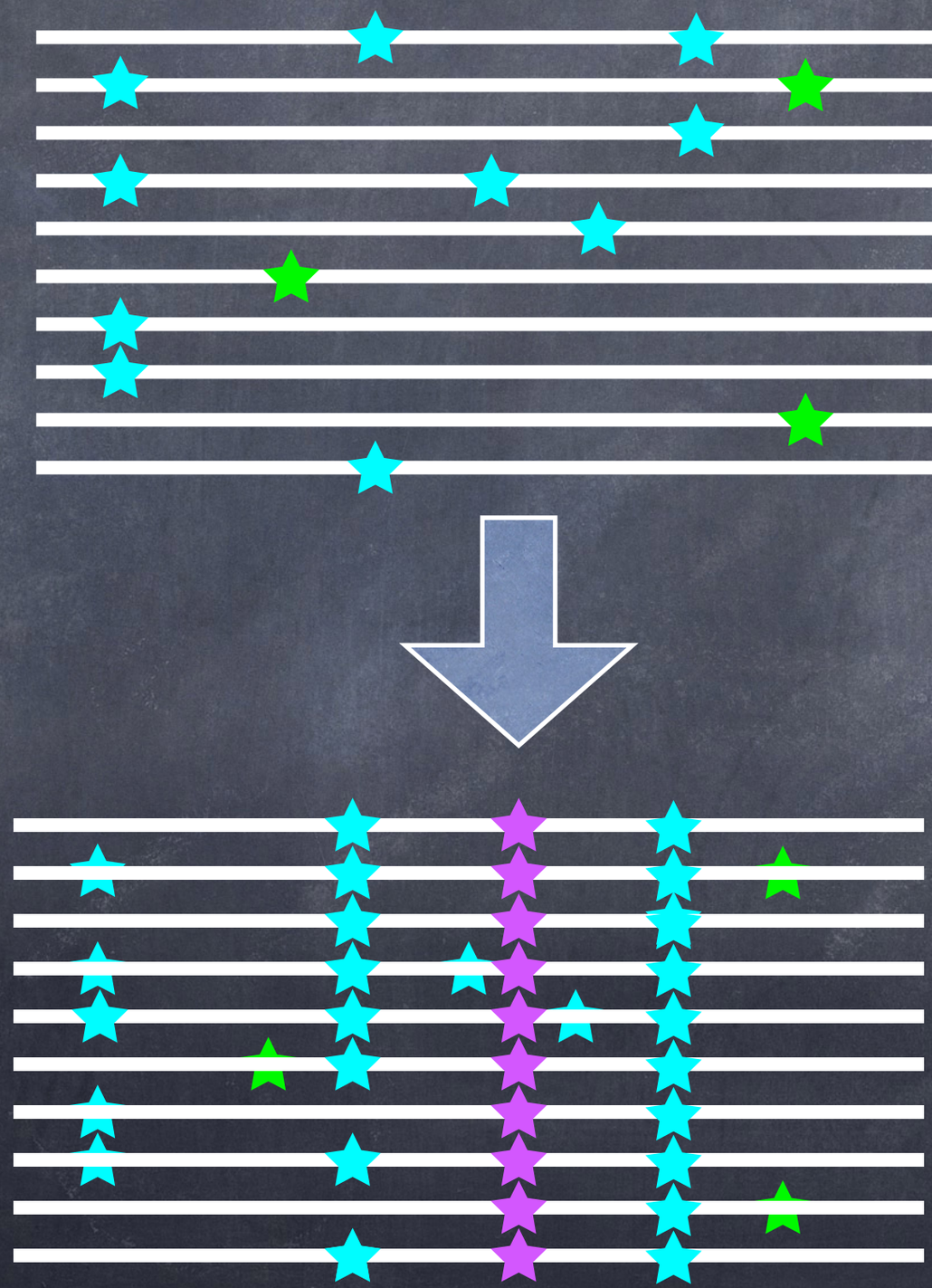
Neutral

Nearly Neutral

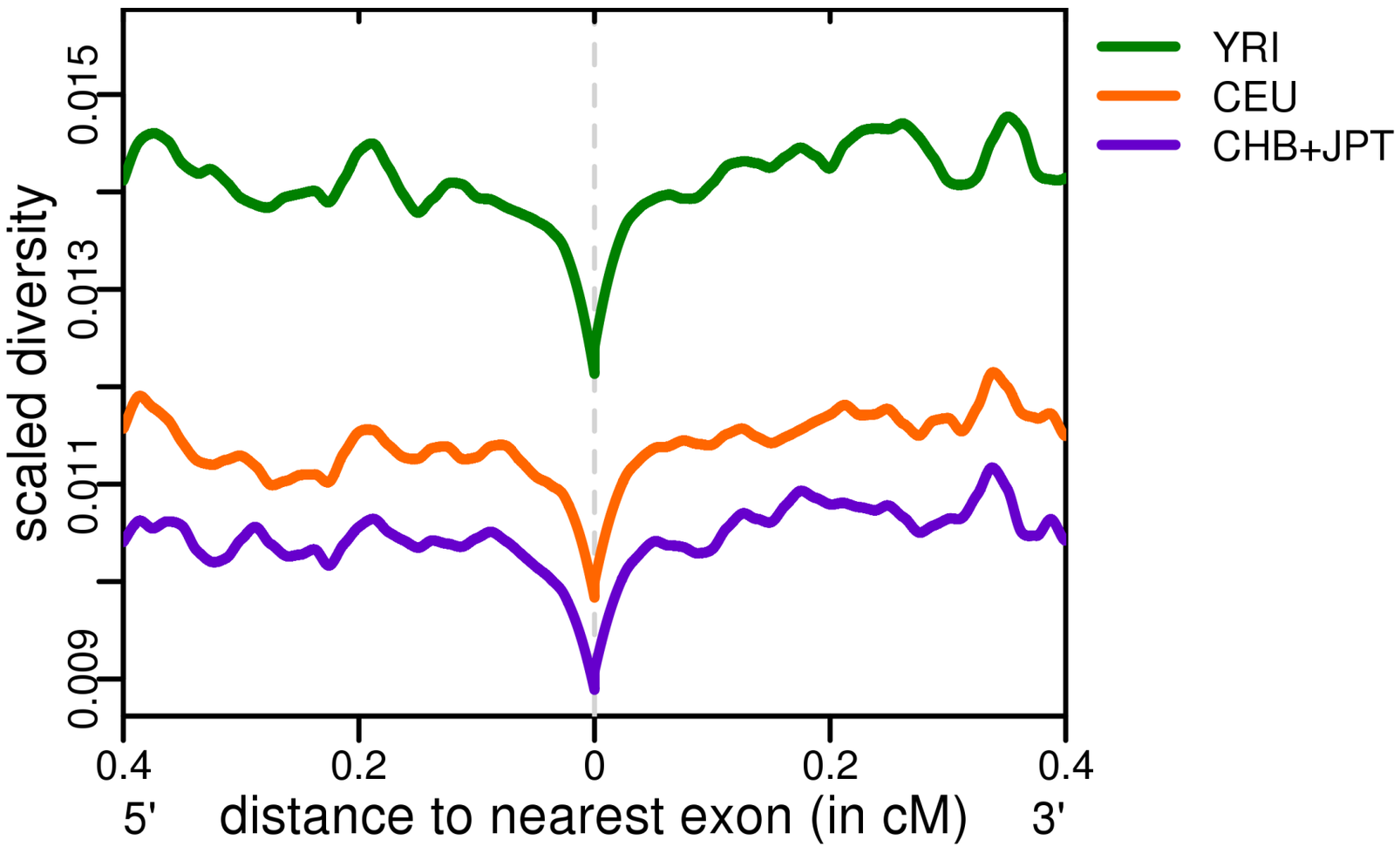
Mildly Deleterious

Fairly Deleterious

Strongly Deleterious



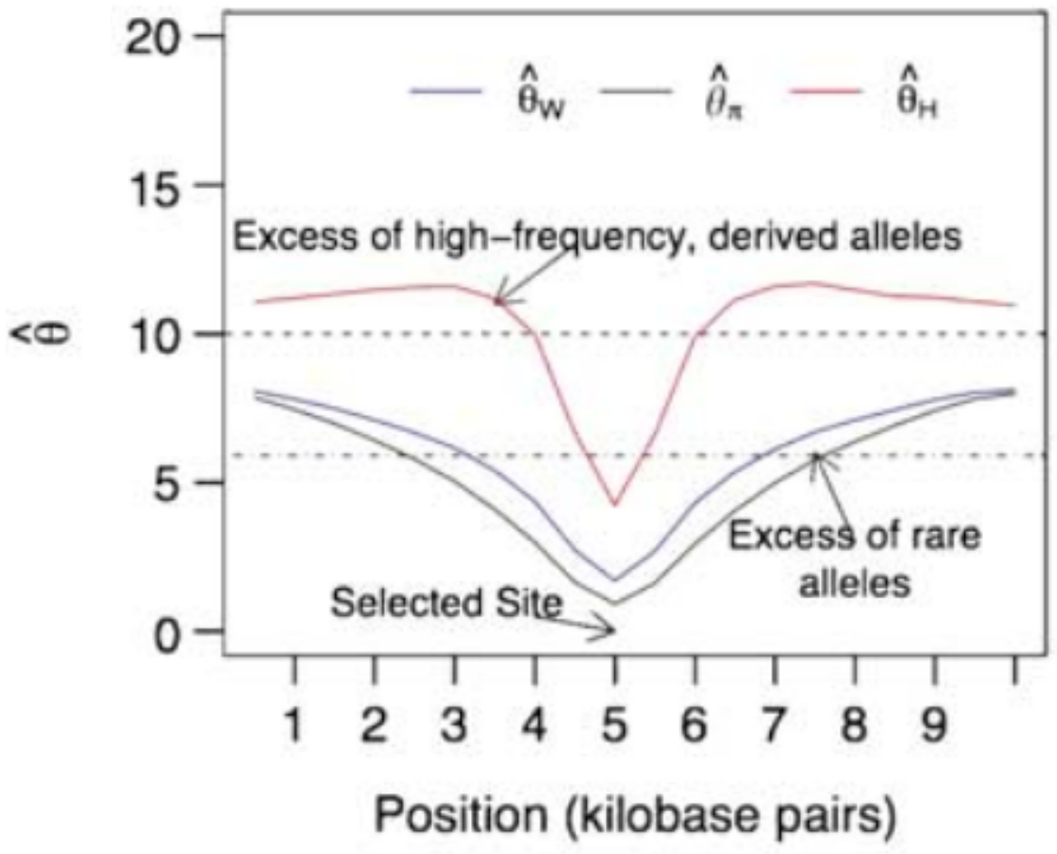
Coding regions tend to have the lowest levels of diversity in the genome



What are the predominant evolutionary forces driving human genomes?!

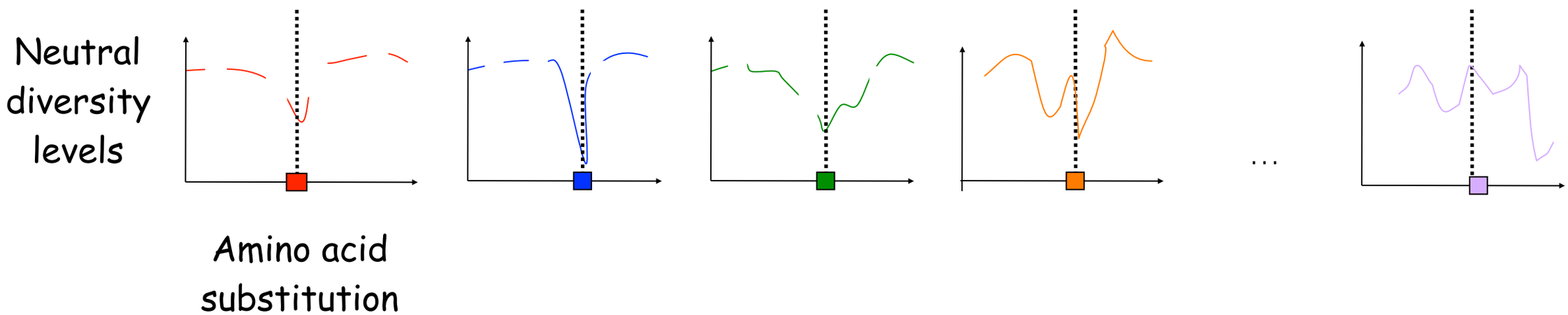
Eyre-Walker & Keightley (2009)	**~40%** of amino acid substitutions were **advantageous**
Boyko et al (2008)	**10-20%** of amino acid substitutions were **advantageous**
Williamson et al (2007)	**10%** of the genome affected by **selective sweeps**

Diversity levels around a selective sweep

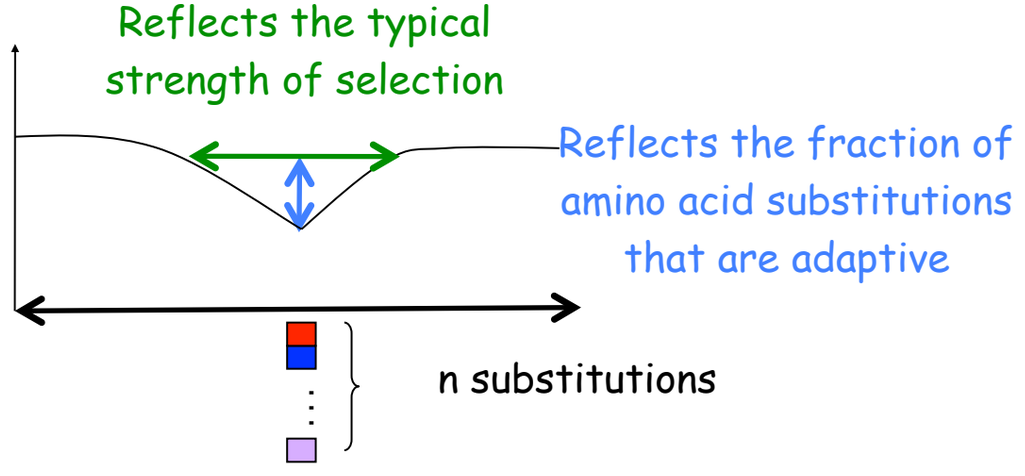


Thornton et al (2007): Simulation of patterns of **neutral** diversity around a **selective sweep**

The footprint of adaptive amino acid substitutions

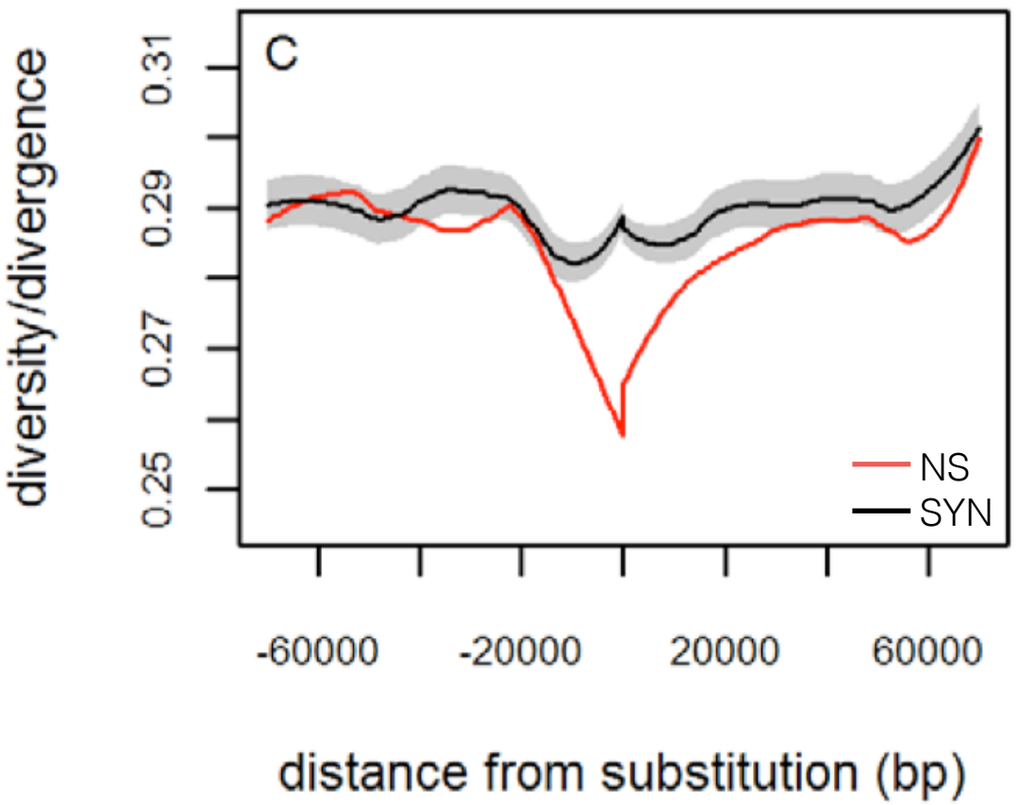


- Goal: compare the pattern around **amino acid substitutions** to the pattern around **synonymous substitutions**.



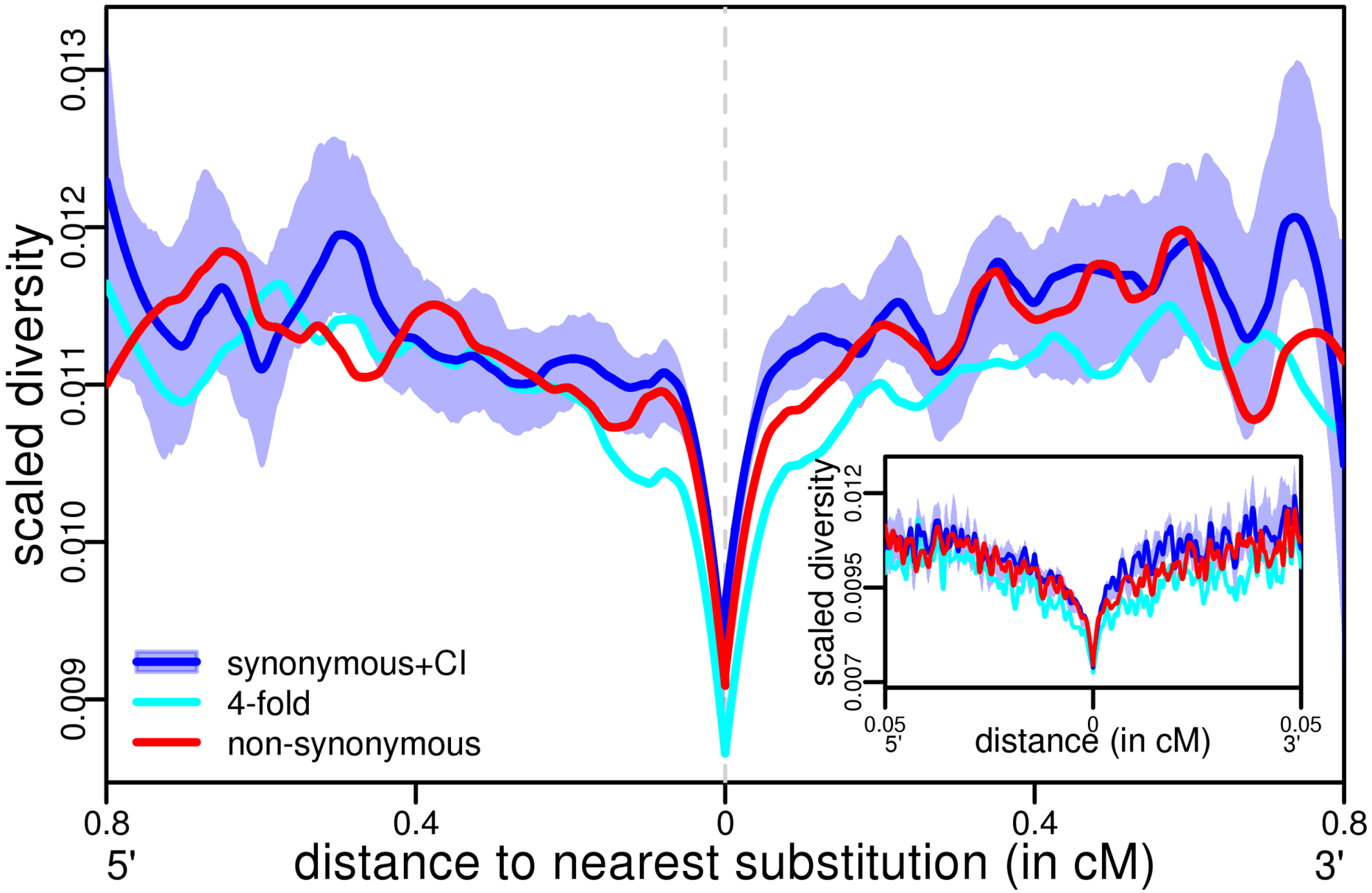
Other organisms…

Drosophila



Sattath et al (2011) estimate ~13% of amino acid substitutions were adaptive.

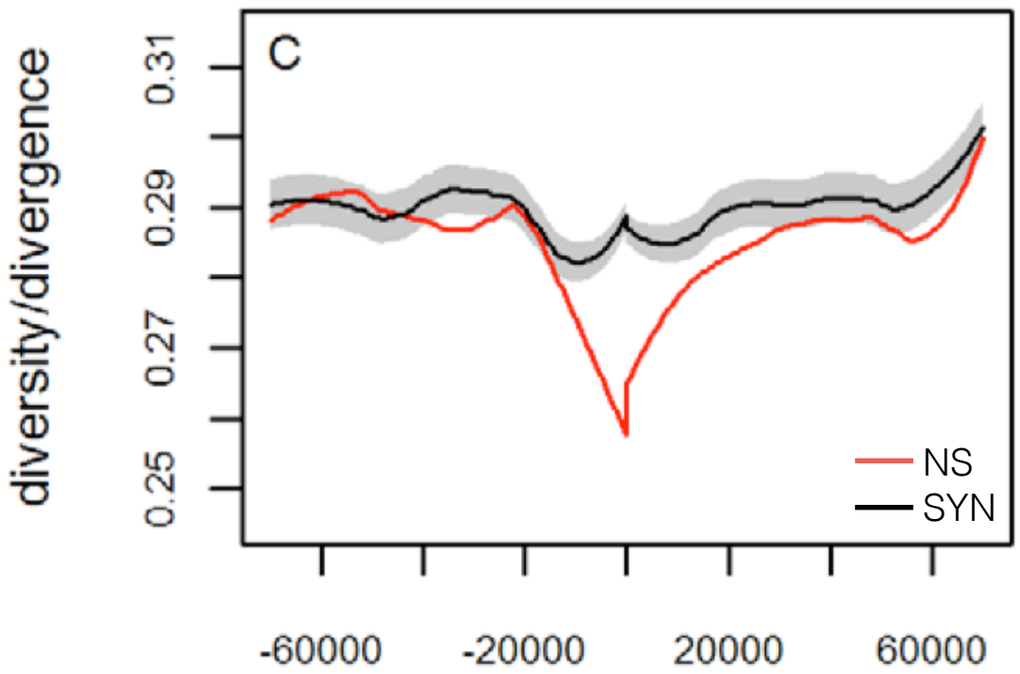
Observed Patterns of Diversity Around Human Substitutions



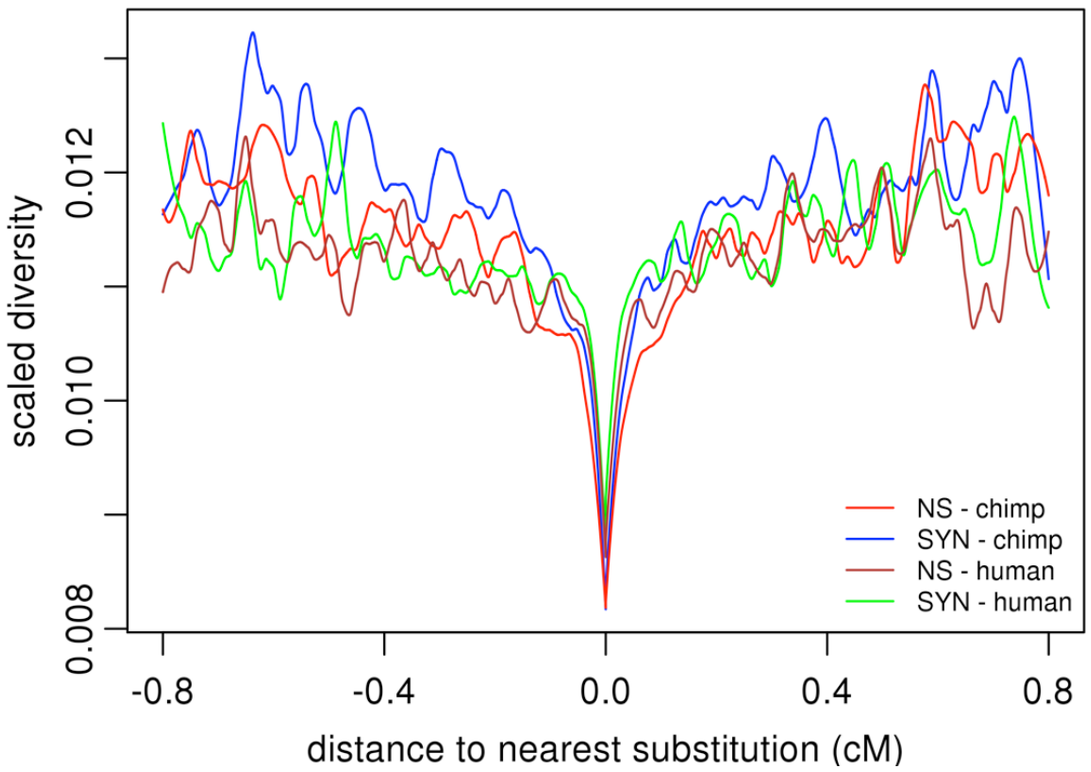
Hernandez, et al. Science (2011)

Other organisms...

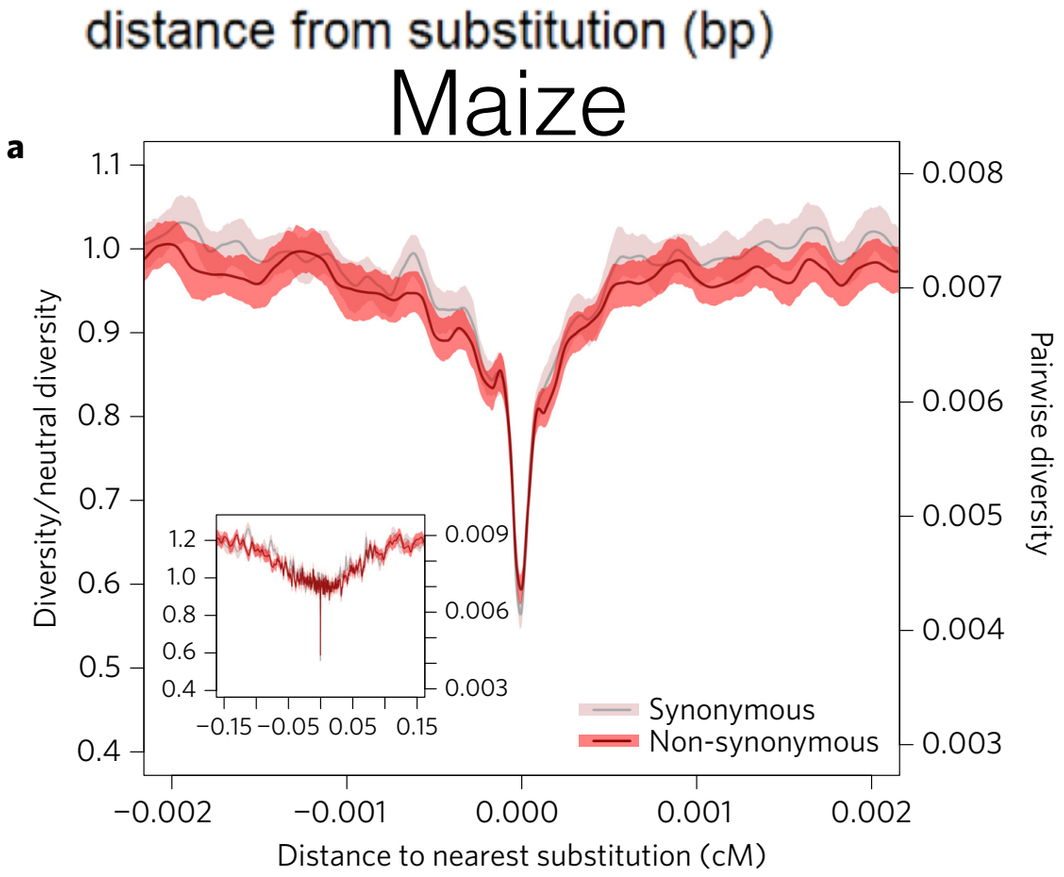
Drosophila



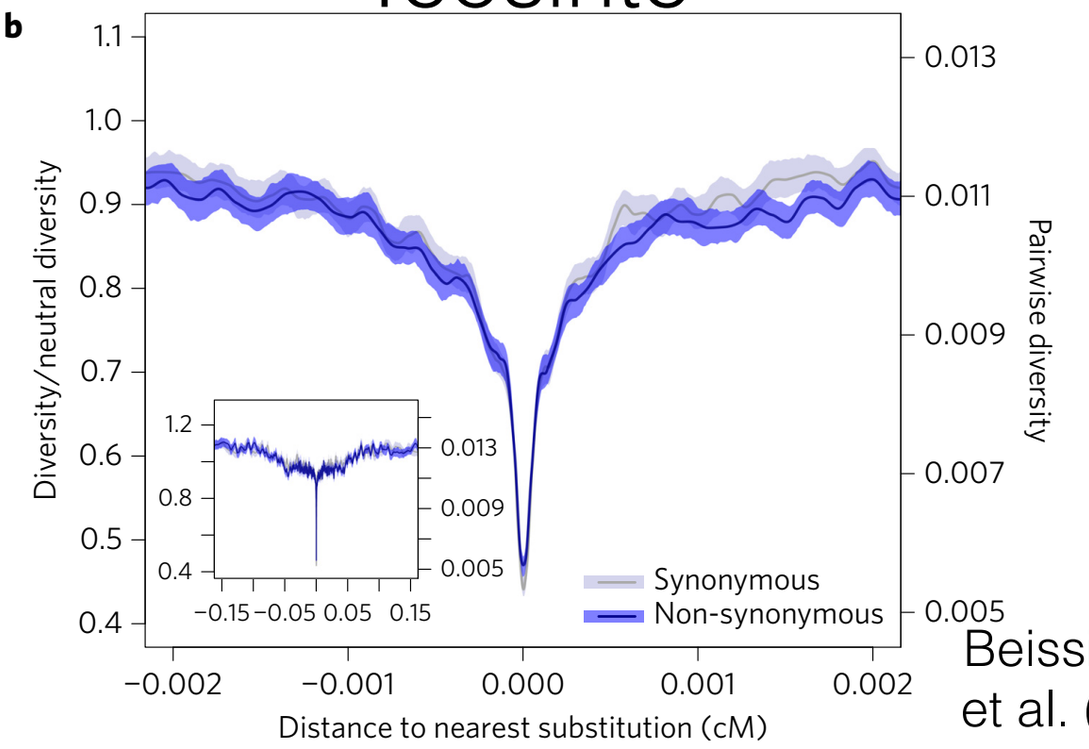
Chimpanzee



Maize



Teosinte



Beissinger , et al. (2016)

The Effect of Negative Selection

Adaptive

Neutral

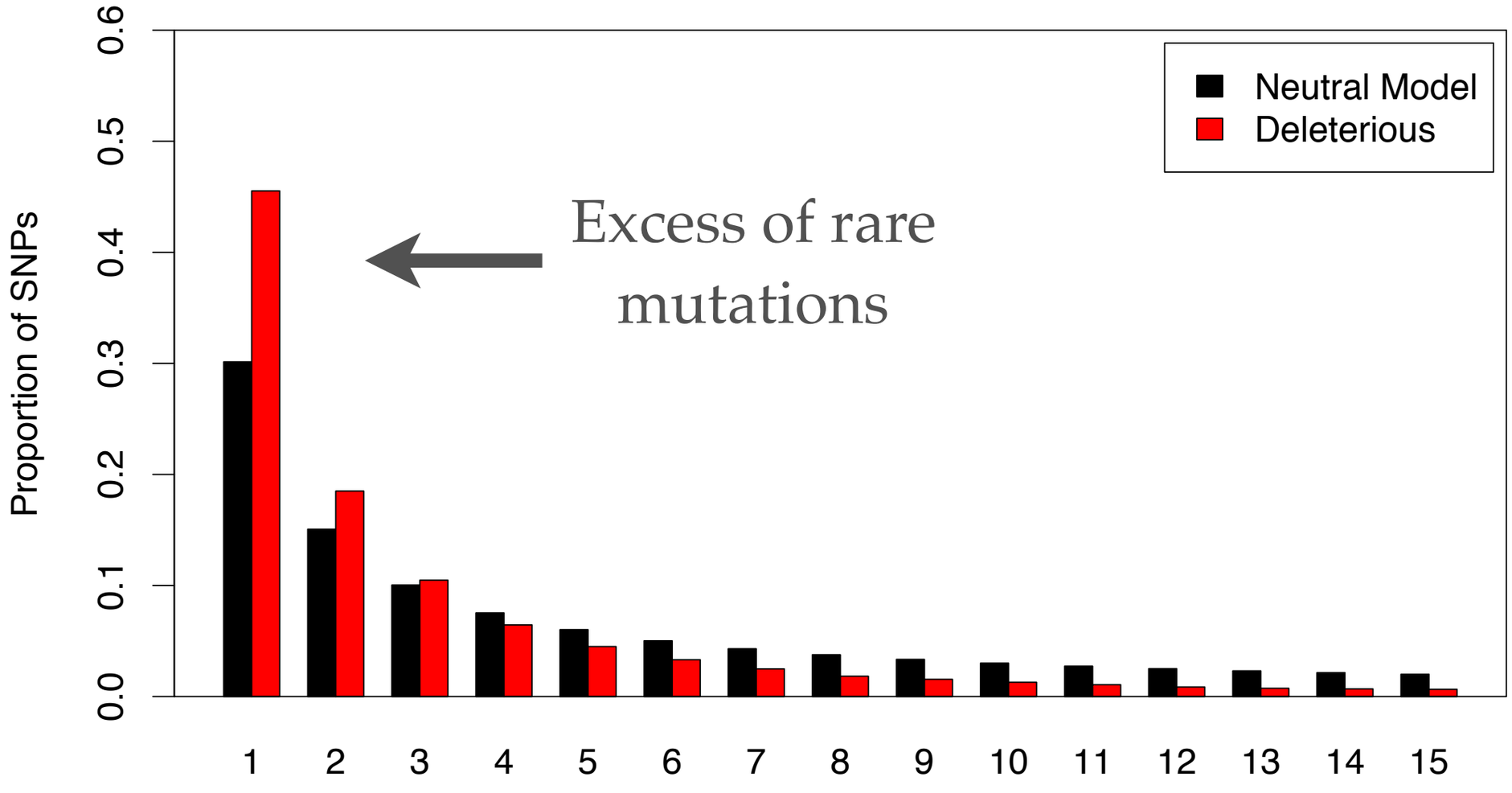
Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious

Site-Frequency Spectrum



The Effect of Negative Selection

Consequences:

- Some proportion of chromosomes eliminated each generation

Background selection

➡ Decreased effective population size ($f_0 N_e$)

➡ Decreased neutral variation ($f_0\boldsymbol{\pi}$)

- While neutral variation can be lost, some neutral mutations may increase in frequency

Background selection (BGS)

- Definition: The reduction of diversity at a **neutral** locus due to the effects of linked deleterious selection
- Can estimate the effect of BGS by comparing **observed** diversity at neutral sites compared to the level of diversity you would **expect** under neutrality!
- $\boldsymbol{\pi/\pi_0}$

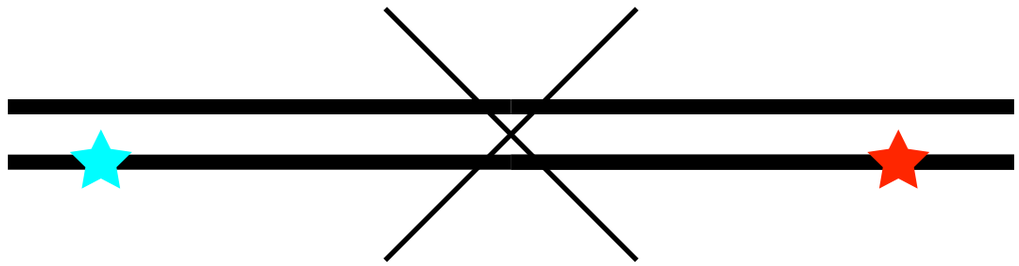
Earlier Theoretical Work

Hudson & Kaplan (1995)

$$f_0 = \exp\left(-\frac{U}{s+R}\right)$$

- U = deleterious mutation rate
- s = selection coefficient
- R = recombination rate

Effect of Recombination

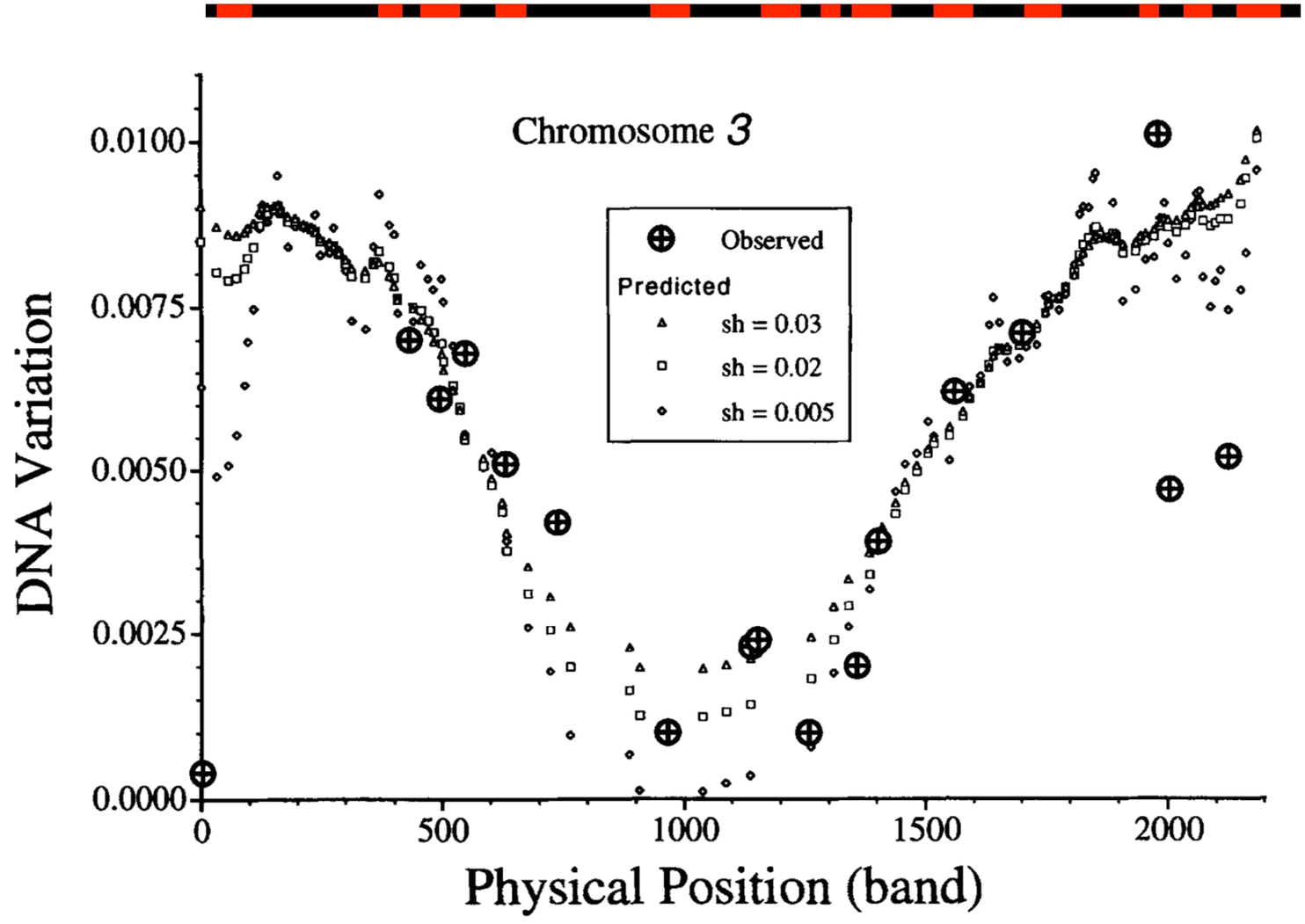


With recombination, neutral mutations can escape the grip of deleterious mutations.

Multiple Targets of Deleterious Mutations

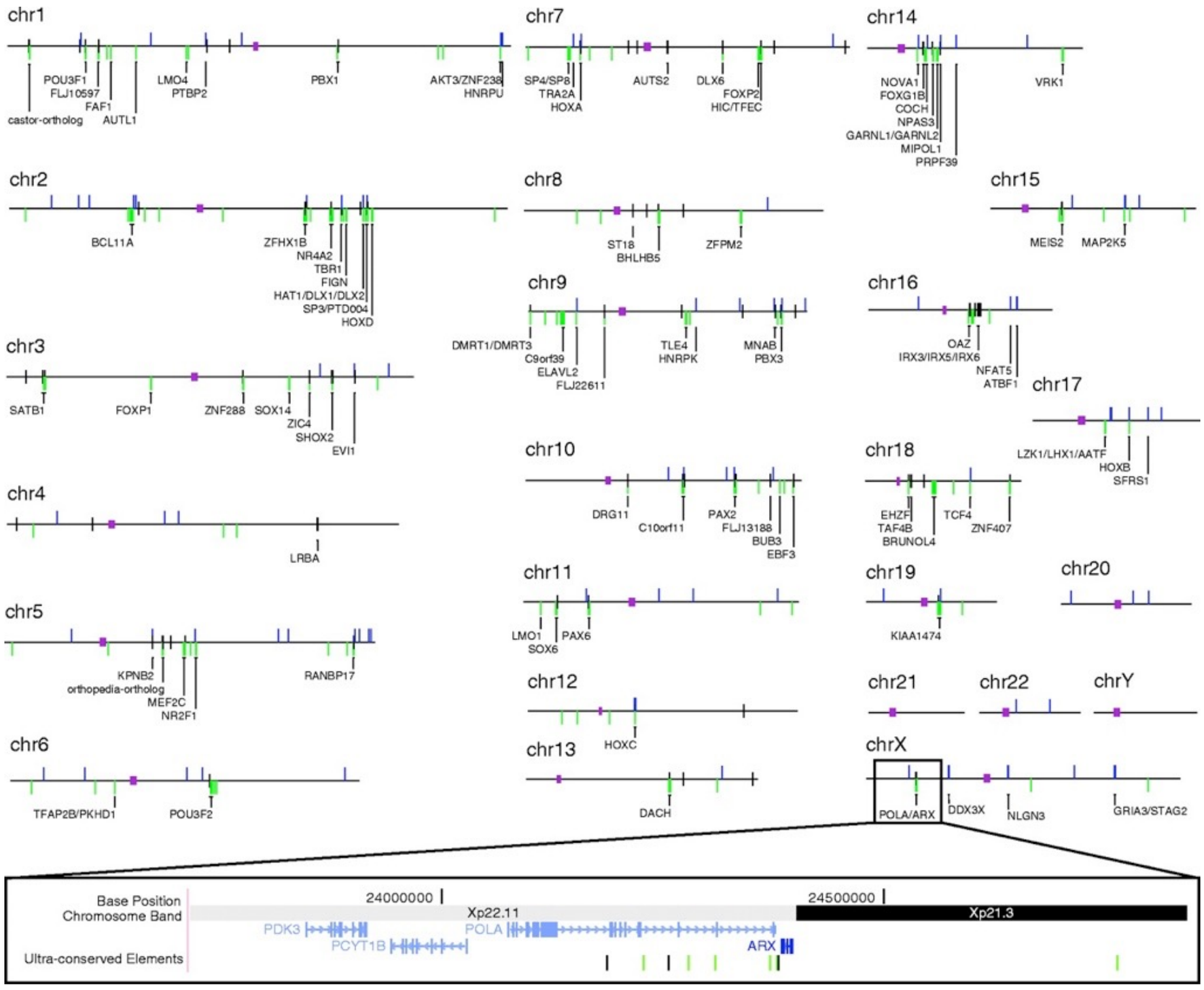
Consider a chromosome composed of neutral loci and deleterious loci

Drosophila



Hudson & Kaplan (1995)

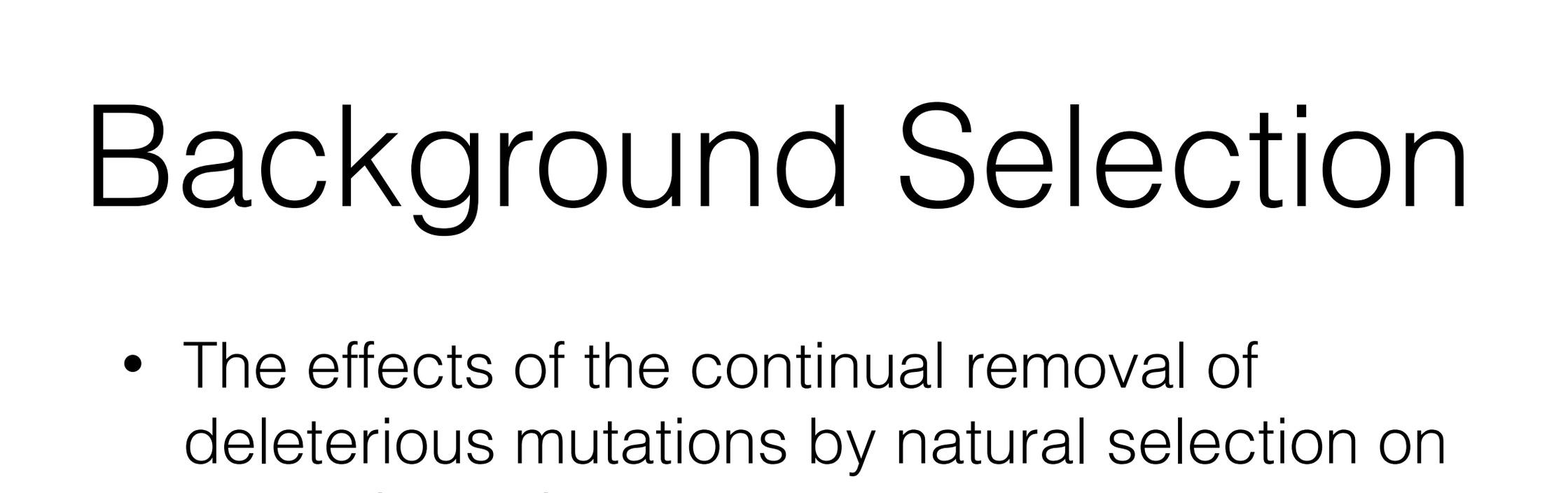
Distribution of Ultraconserved Elements in the Human Genome



Bejerano et al. (2004)

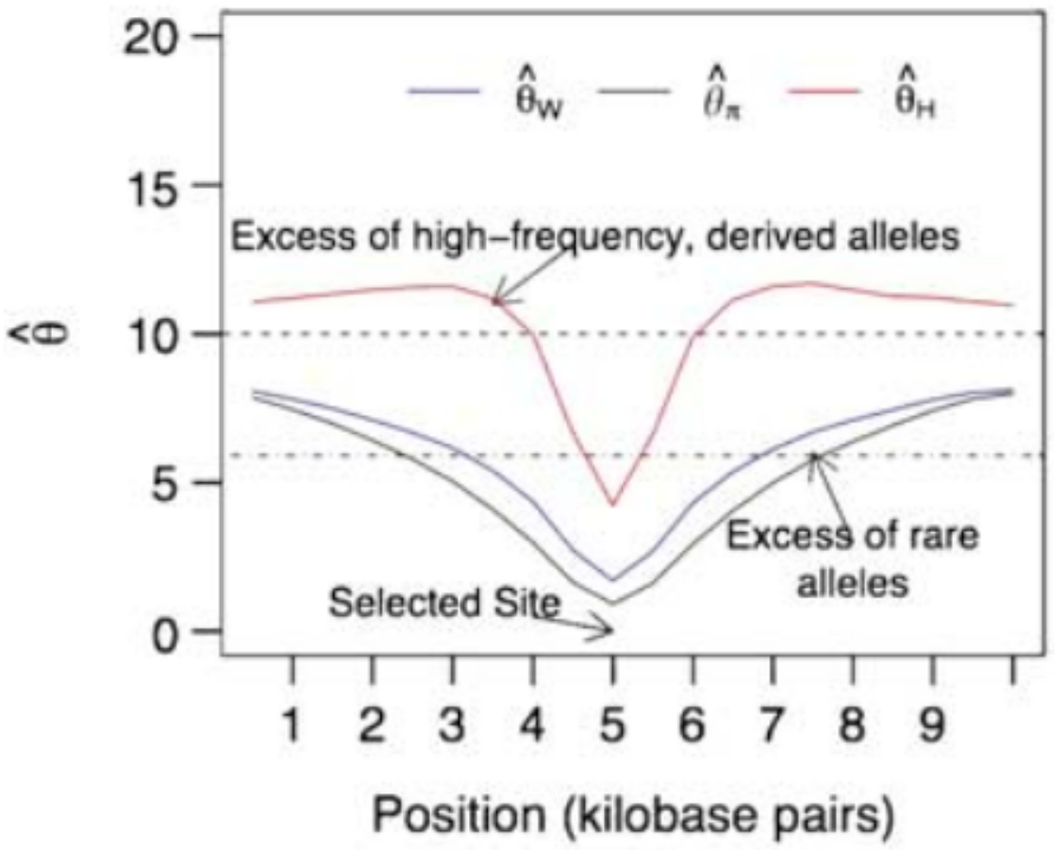
Background Selection

- The effects of the continual removal of deleterious mutations by natural selection on variability at linked sites.

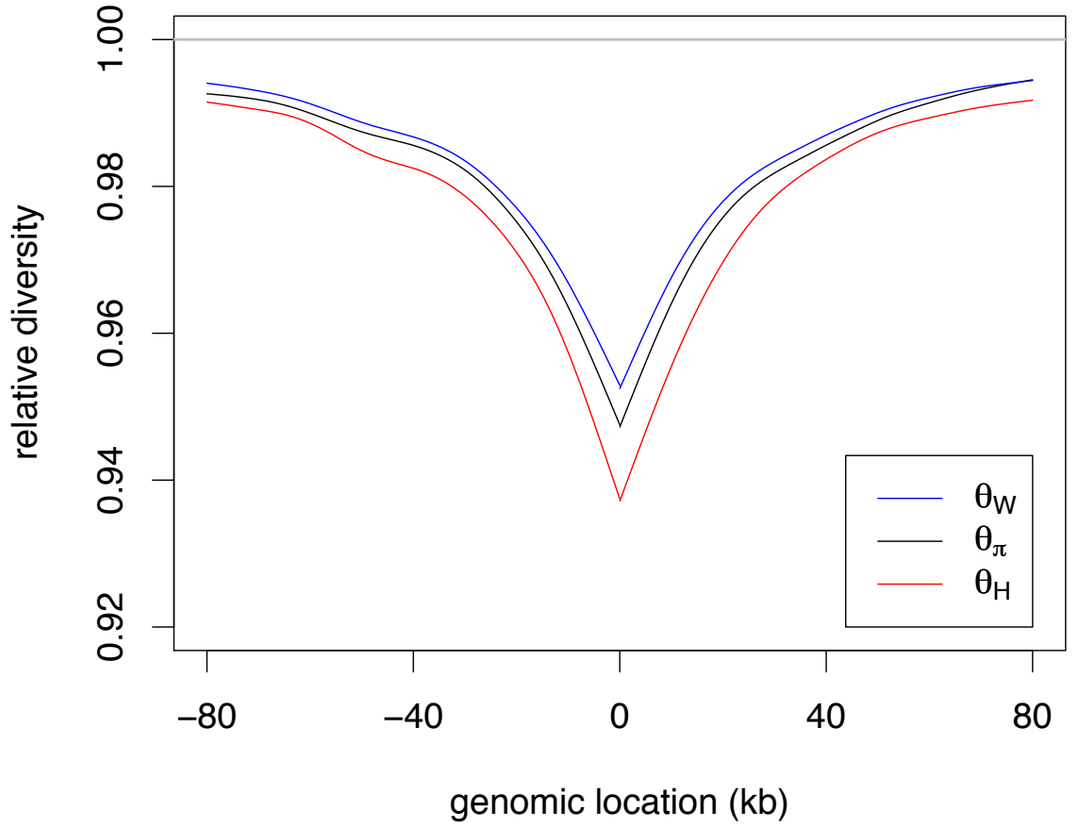


McVicker, G. et al. *PLoS Genet* (2009).

Diversity levels around sites subject to natural selection

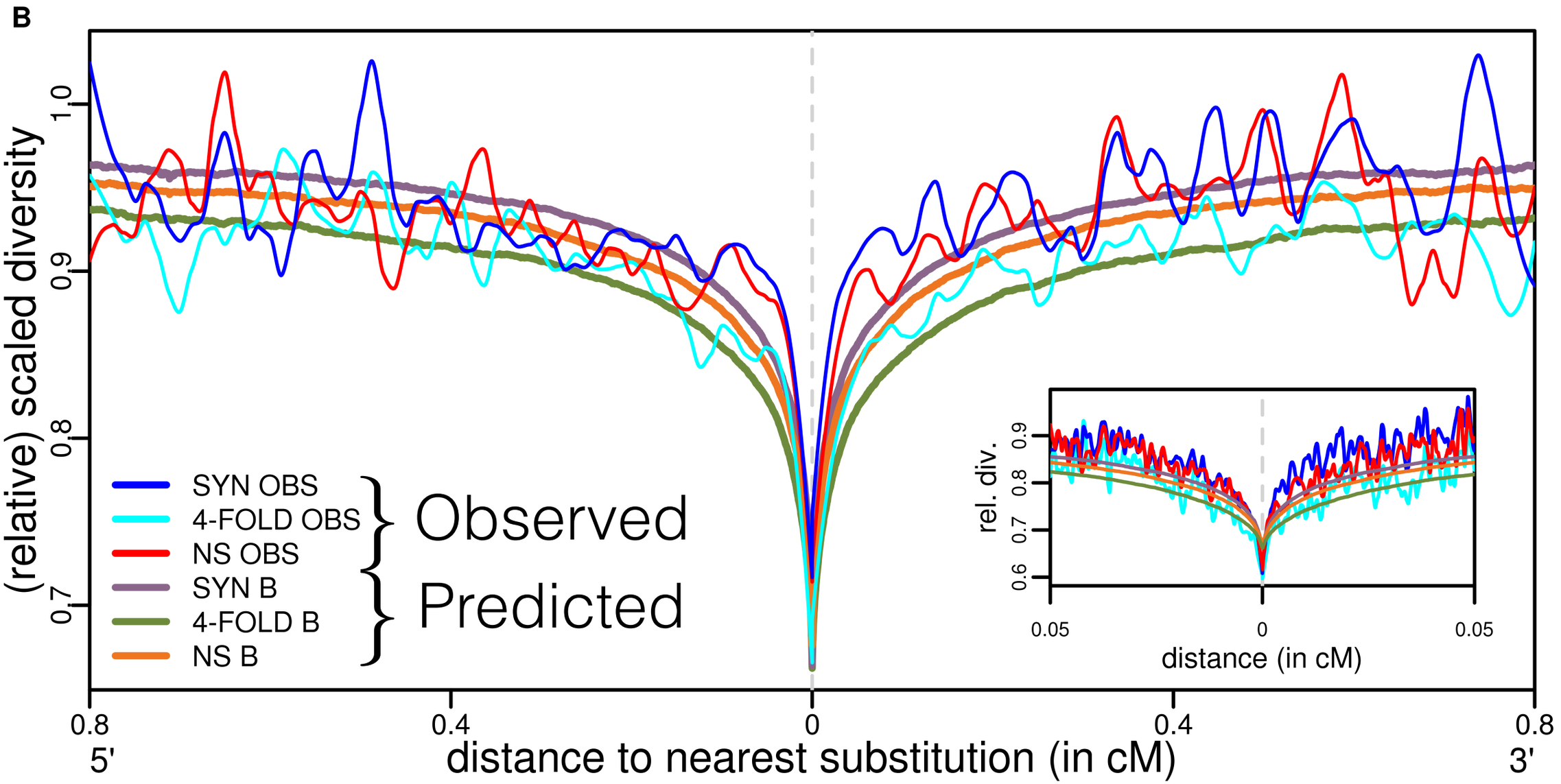


Thornton et al (2007): Simulation of patterns of **neutral** diversity around a **selective sweep**

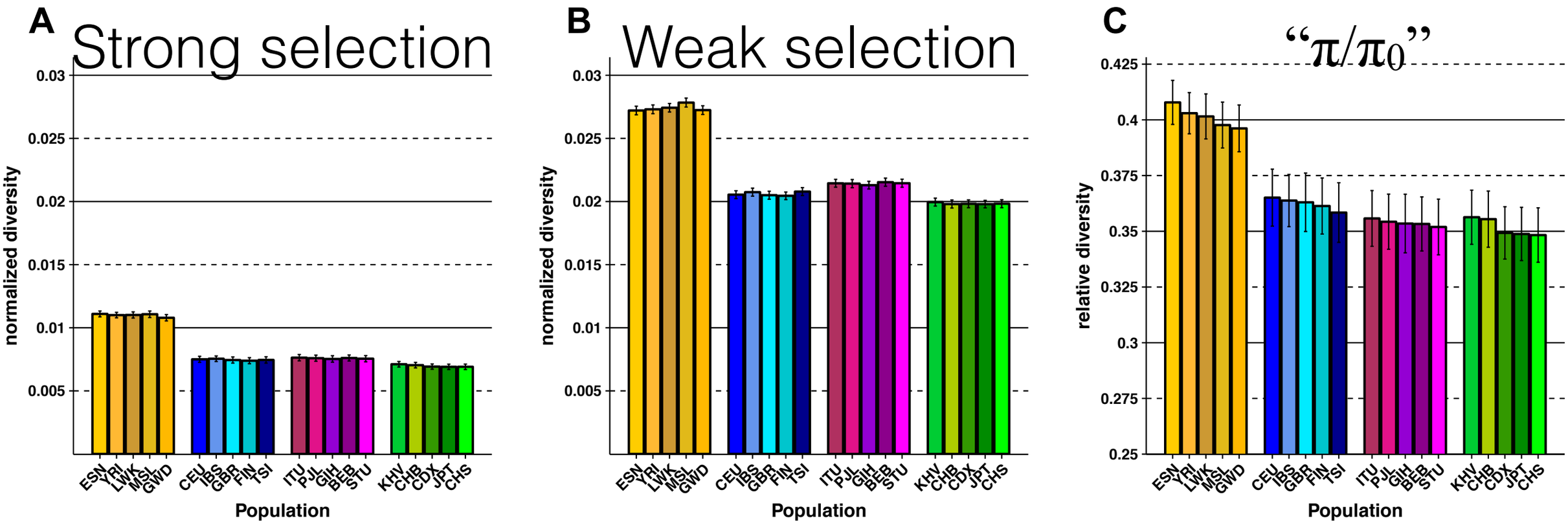


Simulation of patterns of **neutral** diversity around a 700bp **deleterious locus** with γ=-5.

Modeling the data



BGS Features



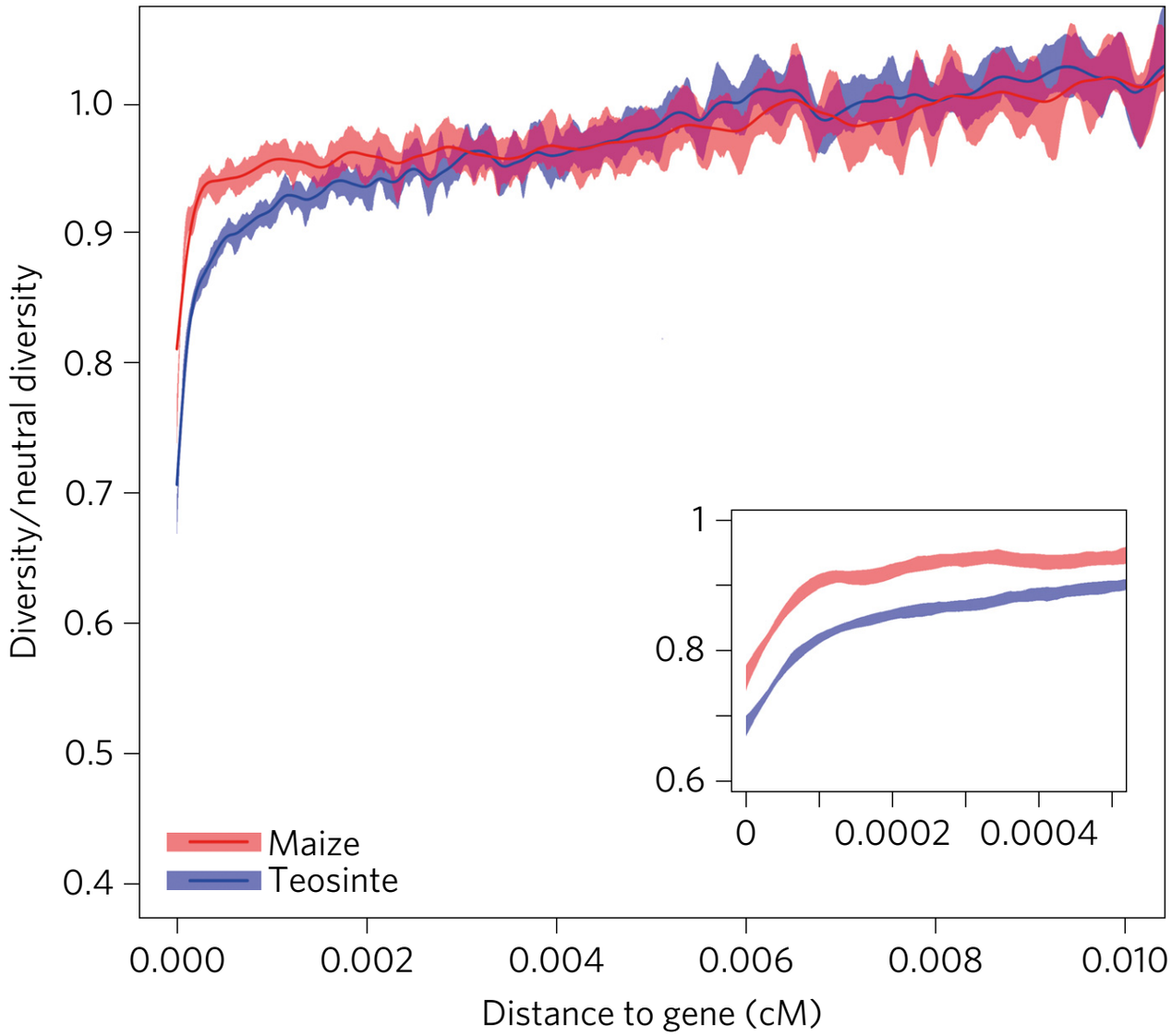
- Neutral sites in 1000 Genomes Project data: 20 non-admixed populations

- The strength of background selection varies across populations!

    - Stronger effects in bottlenecked Out-Of-Africa populations

Torres, et al (2018)

BGS Features



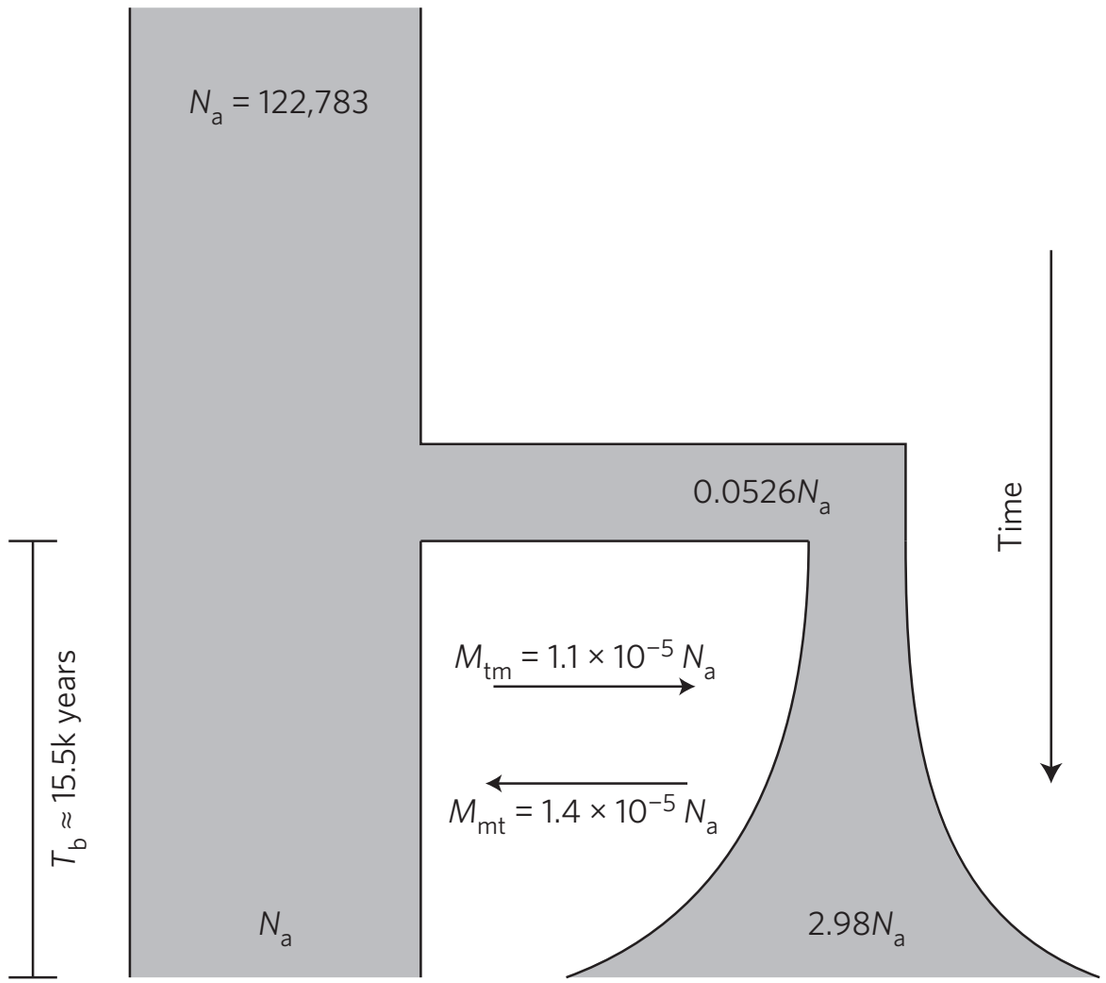


Figure 2 | Estimated demographic history of maize and teosinte. Parameter estimates for a basic bottleneck model of maize domestication. See Methods for details.

- Strength of BGS varies between Maize and Teosinte

Genetic Load

- Genetic load is the reduction in population mean fitness due to deleterious mutations compared to a (hypothetical) mutation-free population.

- Load is the outcome of the evolutionary process of a population.

- But, unlike other features of genetic variation, it cannot be directly observed.

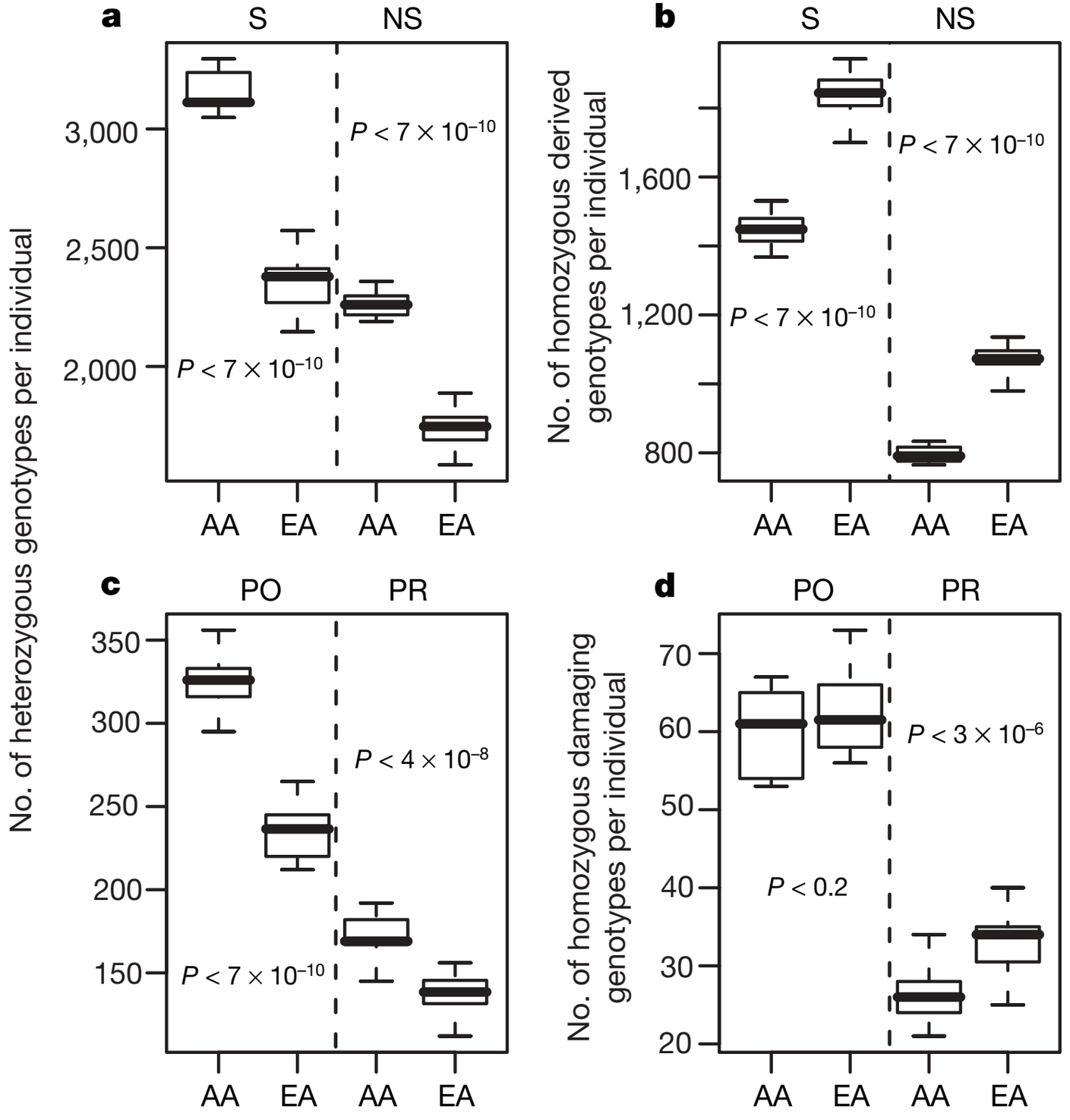
- Must be indirectly inferred.

Inferring Genetic Load

- Empirical counting approaches:
  - –Under an additive model, the number of derived deleterious alleles will be proportional to genetic load
  - –Under a recessive model, the number of homozygous derived genotypes will be proportional to load

Inferring Genetic Load

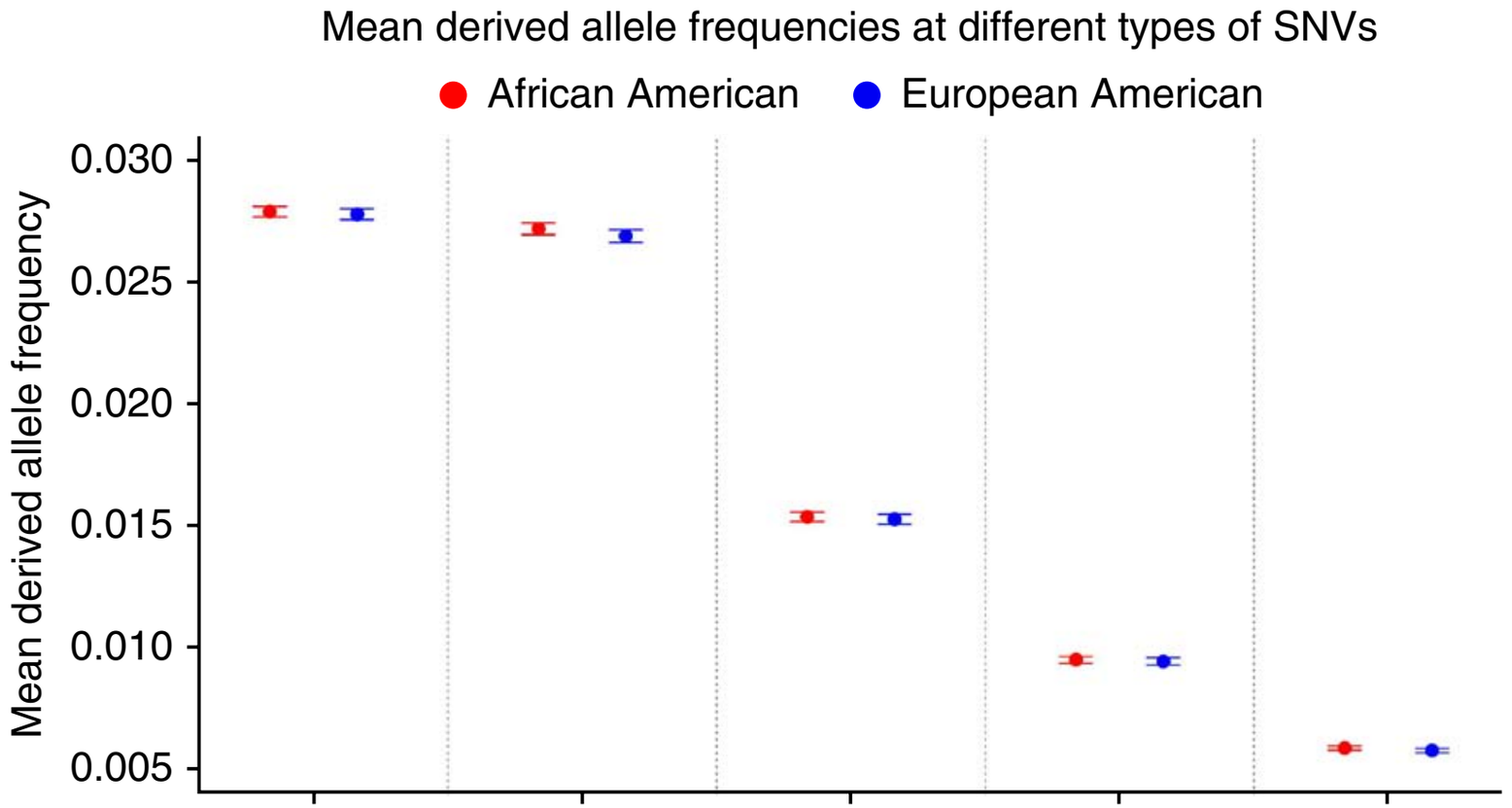
- It is widely appreciated that African ancestry individuals have more variation overall than individuals with European ancestry.
- However, European individuals have more homozygous variation.
- Increased load?



Lohmueller, et al. (Nature, 2008)

Inferring Genetic Load

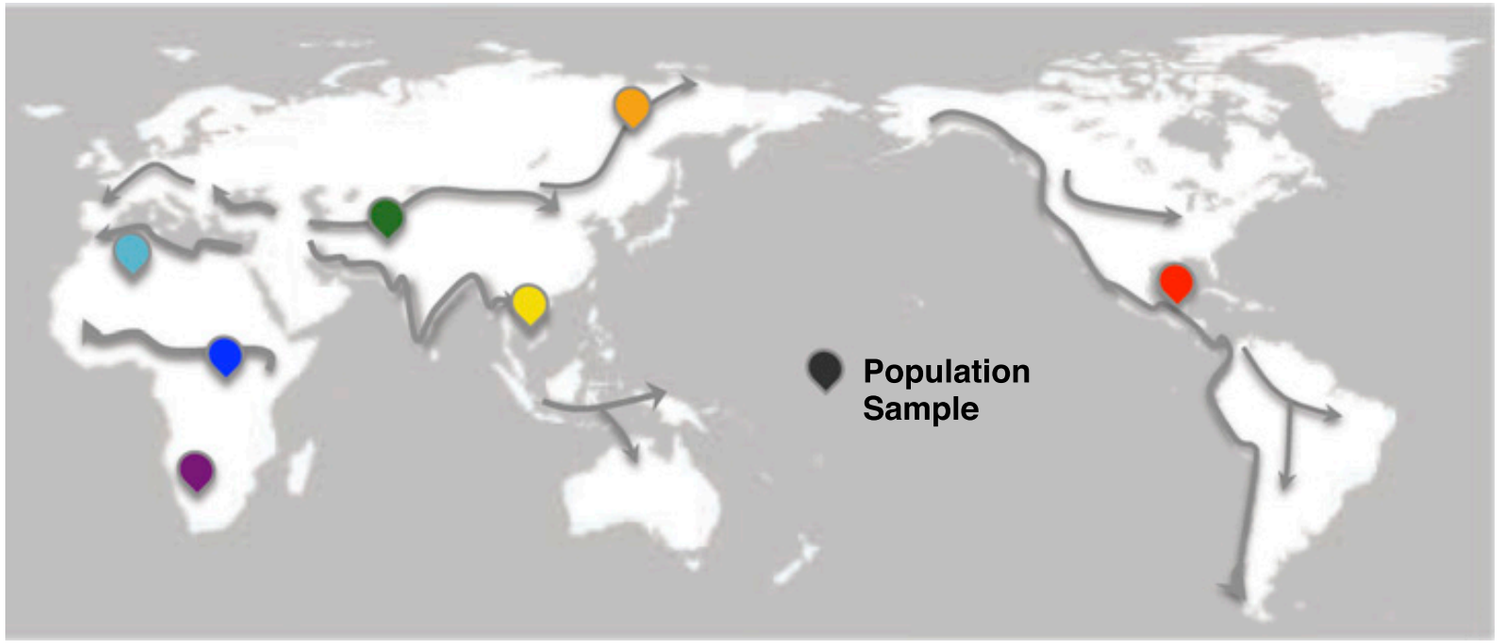
- However, het. and hom. derived alleles appear to balance between African and European Americans.
- All individuals have same number of derived alleles!

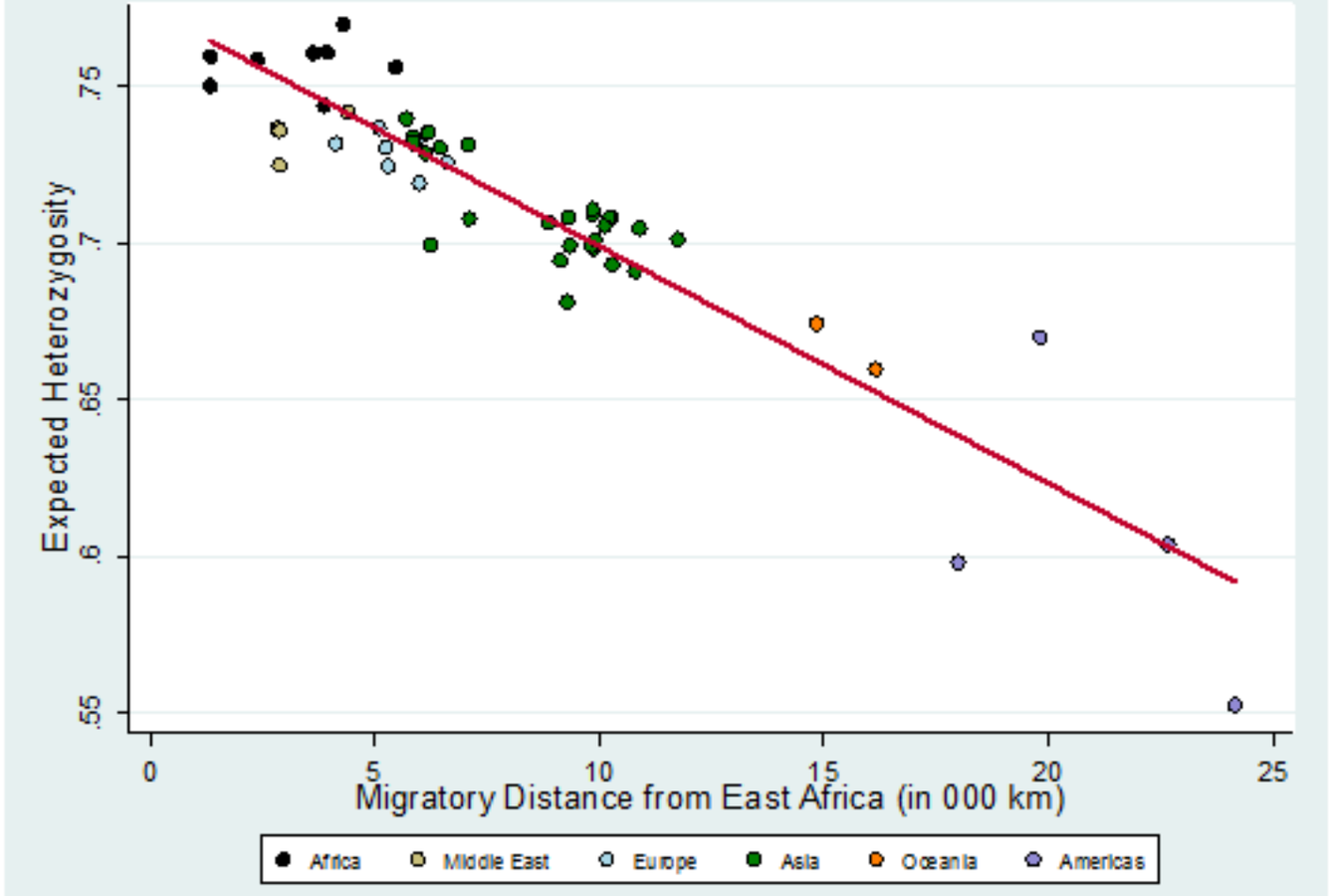


	Noncoding	Synonymous	Benign nonsynonymous	Possibly damaging	Probably damaging
Number per individual, AA:	21,421	15,401	5,373	1,695	2,002
Number per individual, EA:	21,345	15,231	5,338	1,682	1,969

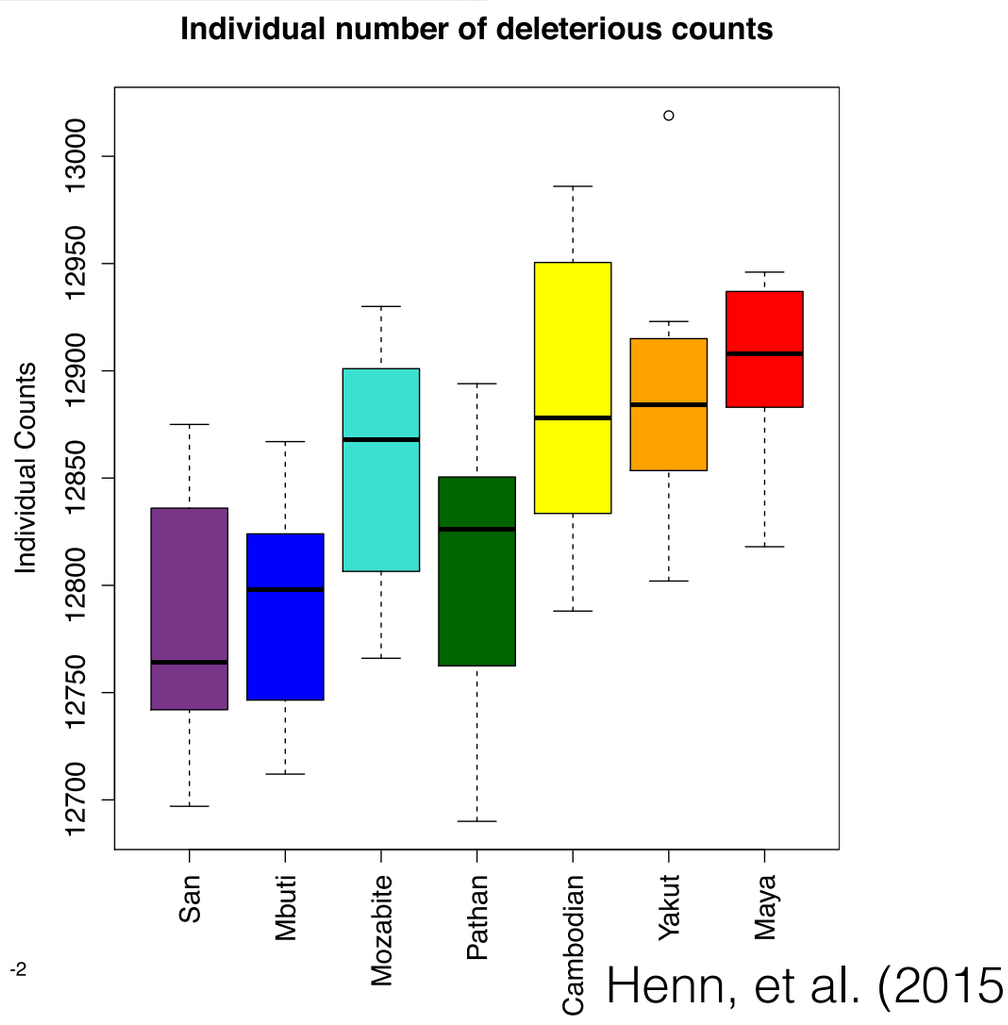
Simons, et al. (2014)

Serial Founder Effects on Genetic Load





Ramachandran, et al. (2005)



Henn, et al. (2015)

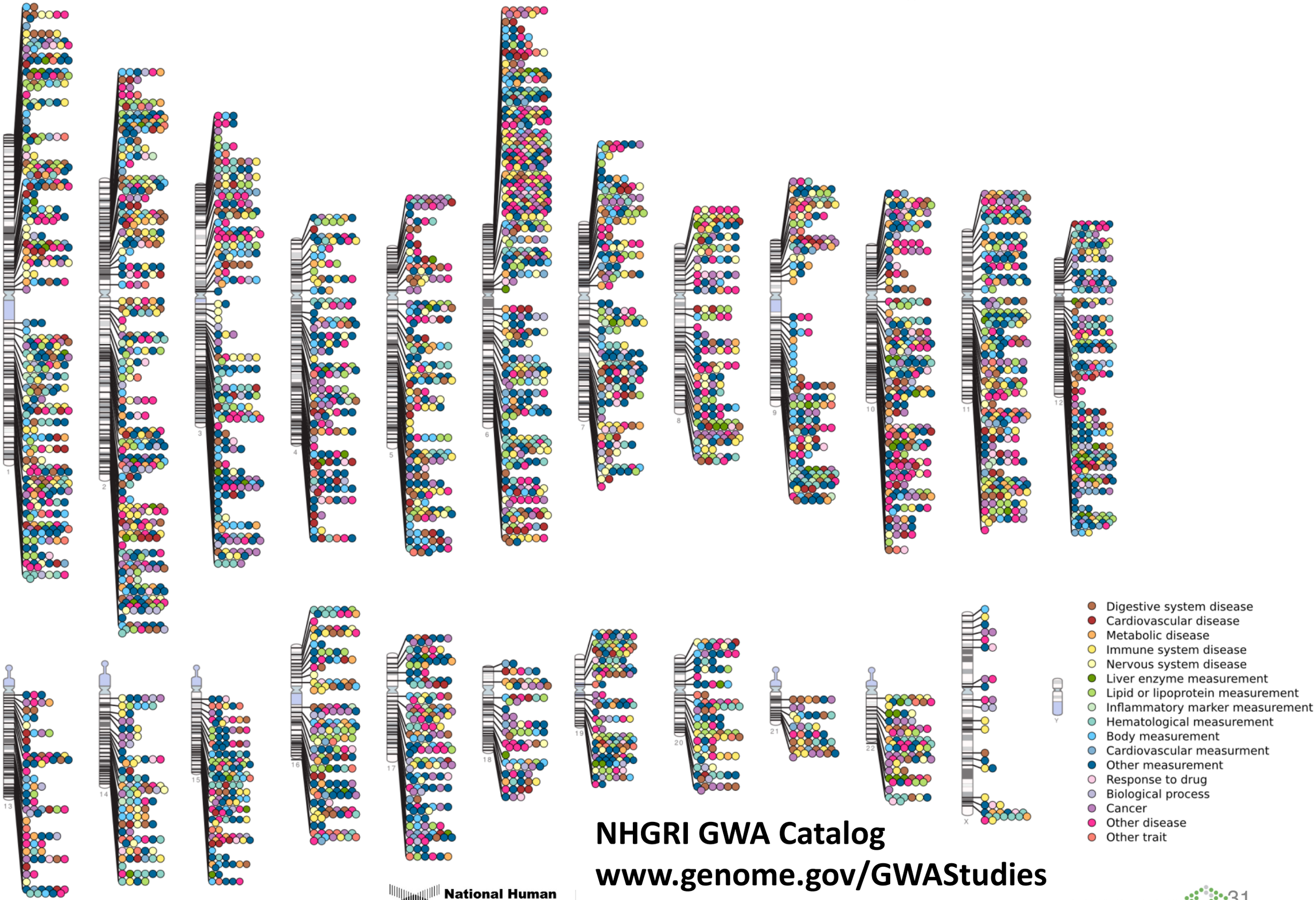
Background Selection & Disease?

Background selection drives patterns of genetic variation.

- But does it matter?
- Does it have implications for studying complex traits?

To find out, we looked at the NHGRI GWAS database: www.genome.gov/gwastudies/

Published Genome-Wide Associations through 12/2013
Published GWA at p≤5X10^{-8} for 17 trait categories

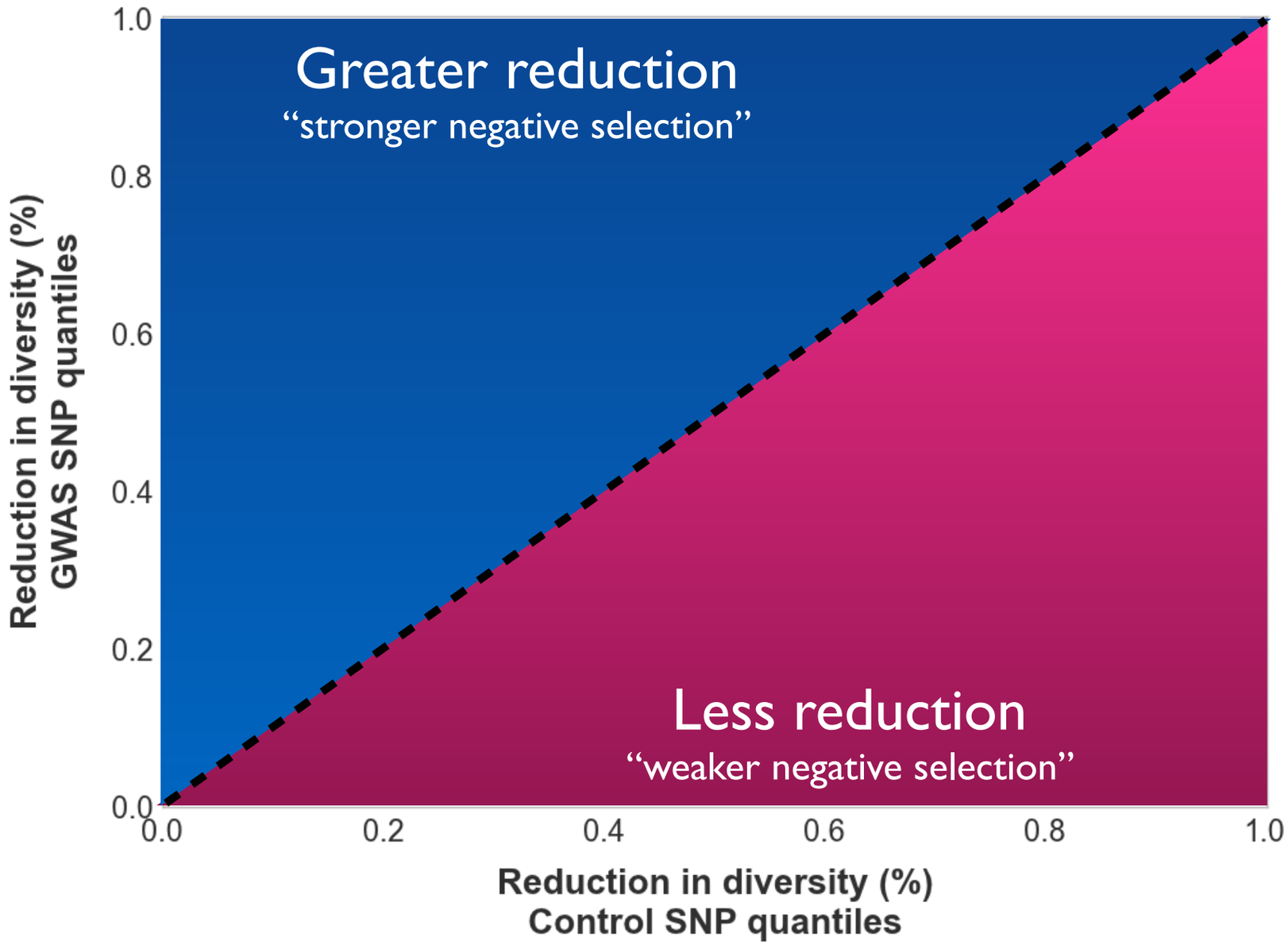


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

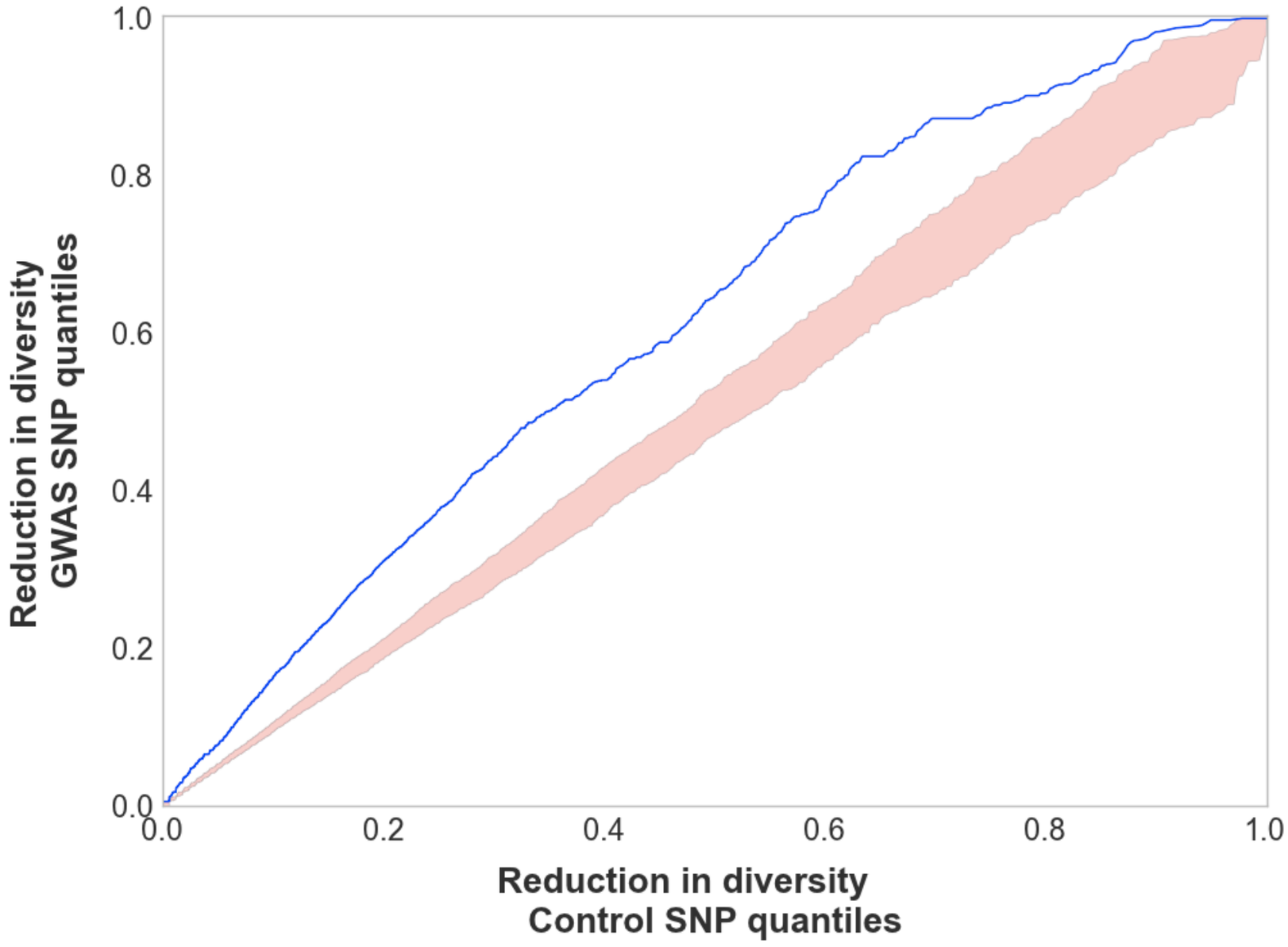
EMBL-EBI

Effects of Linked Selection



- QQ-plot of the reduction in diversity around GWAS hits compared to background.

Effects of Linked Selection



- Greater reduction in diversity around GWAS hits indicates a strong, local burden of negative selection.

Maher, et al. *Human Heredity* (2012).

Key Feature of Natural Selection

- Alleles change frequency unusually fast
    - Positive selection tends to increase frequency
    - Negative selection tends to decrease frequency
- All tests for natural selection seek to identify this feature using different aspects of the data.
- While negative selection shapes majority of patterns of variation in many species, positive selection may drive patterns of local variation.

The Effect of Positive Selection

Adaptive

Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious

The Effect of Positive Selection

Adaptive

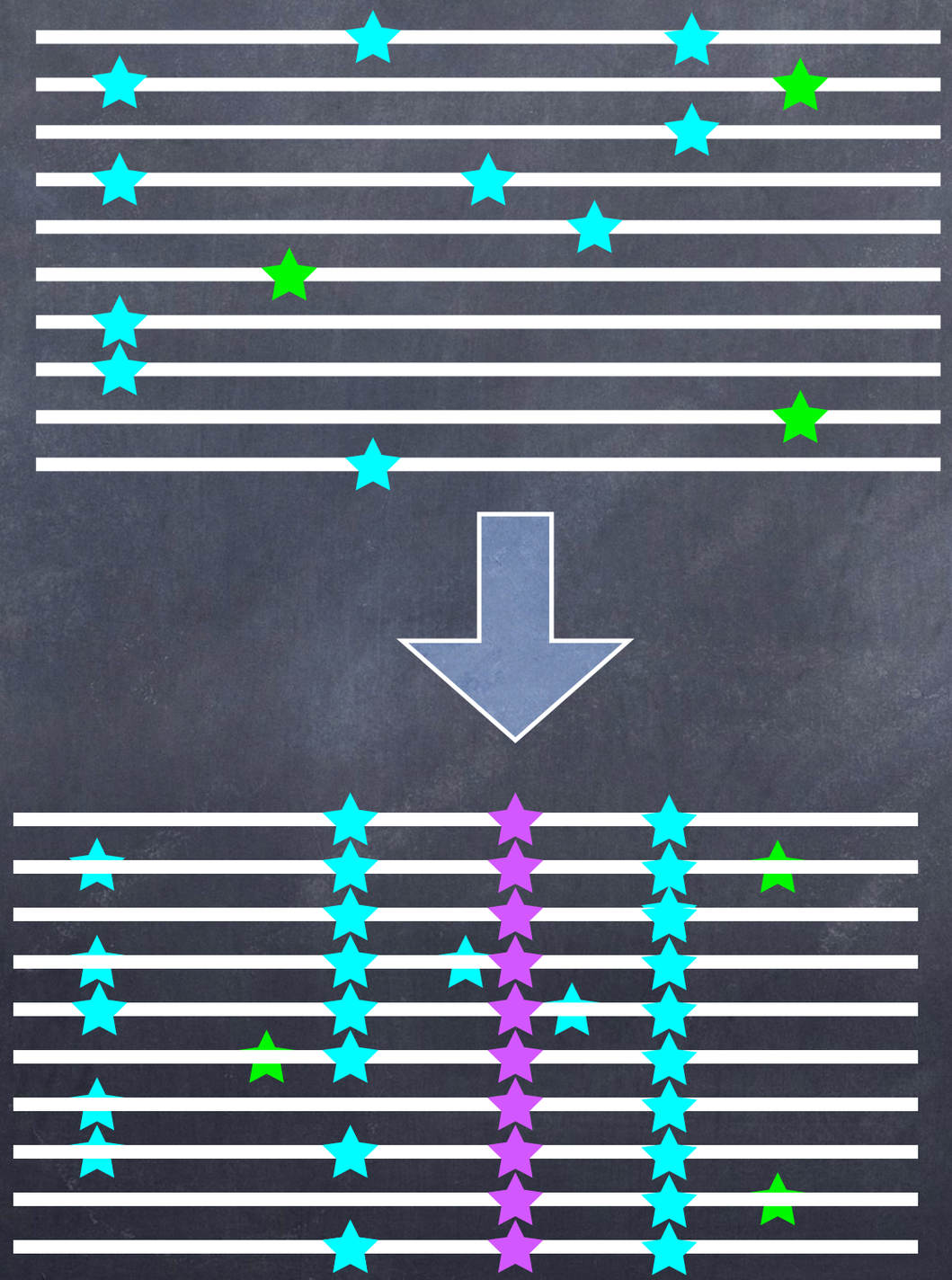
Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious



Types of Positive Selection

- Selection acts in one population but not another
    - Frequencies of the selected alleles in one population will go up relatively quickly compared to the frequencies of those same alleles in the other population.
- The test is simple:
    - Are there alleles that have unusually large allele frequency differences between two populations?

Testing for Population Divergence

- Imagine two populations diverged several thousand years ago.
- One population stayed where it was, but the other migrated up a mountain to the Tibetan Plateau.
  - Many environmental changes...
  - Not obvious where in the genome to look for adaptations
  - Try exome sequencing

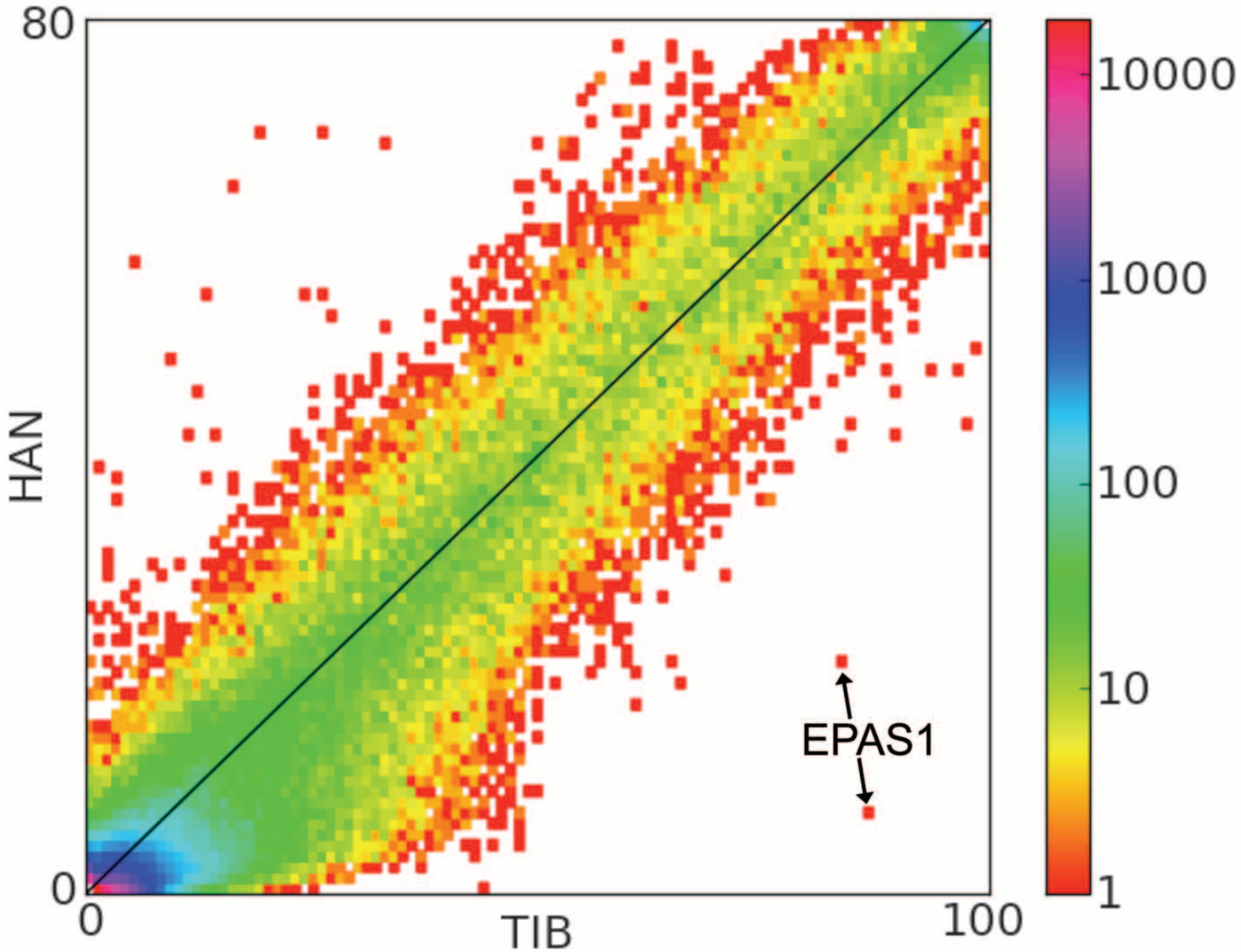
Testing for Population Divergence

Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude

Xin Yi,[1,2]* Yu Liang,[1,2]* Emilia Huerta-Sanchez,[3]* Xin Jin,[1,4]* Zha Xi Ping Cuo,[2,5]* John E. Pool,[3,6]* Xun Xu,[1] Hui Jiang,[1] Nicolas Vinckenbosch,[3] Thorfinn Sand Korneliussen,[7] Hancheng Zheng,[1,4] Tao Liu,[1] Weiming He,[1,8] Kui Li,[2,5] Ruibang Luo,[1,4] Xifang Nie,[1] Honglong Wu,[1,9] Meiru Zhao,[1] Hongzhi Cao,[1,9] Jing Zou,[1] Ying Shan,[1,4] Shuzheng Li,[1] Qi Yang,[1] Asan,[1,2] Peixiang Ni,[1] Geng Tian,[1,2] Junming Xu,[1] Xiao Liu,[1] Tao Jiang,[1,9] Renhua Wu,[1] Guangyu Zhou,[1] Meifang Tang,[1] Junjie Qin,[1] Tong Wang,[1] Shuijian Feng,[1] Guohong Li,[1] Huasang,[1] Jiangbai Luosang,[1] Wei Wang,[1] Fang Chen,[1] Yading Wang,[1] Xiaoguang Zheng,[1,2] Zhuo Li,[1] Zhuoma Bianba,[10] Ge Yang,[10] Xinping Wang,[11] Shuhui Tang,[11] Guoyi Gao,[12] Yong Chen,[5] Zhen Luo,[5] Lamu Gusang,[5] Zheng Cao,[1] Qinghui Zhang,[1] Weihan Ouyang,[1] Xiaoli Ren,[1] Huiqing Liang,[1] Huisong Zheng,[1] Yebo Huang,[1] Jingxiang Li,[1] Lars Bolund,[1] Karsten Kristiansen,[1,7] Yingrui Li,[1] Yong Zhang,[1] Xiuqing Zhang,[1] Ruiqiang Li,[1,7] Songgang Li,[1] Huanming Yang,[1] Rasmus Nielsen,[1,3,7]† Jun Wang,[1,7]† Jian Wang[1]†

Yi, X. et al. *Science* **329**, 75-78 (2010).

Testing for Population Divergence

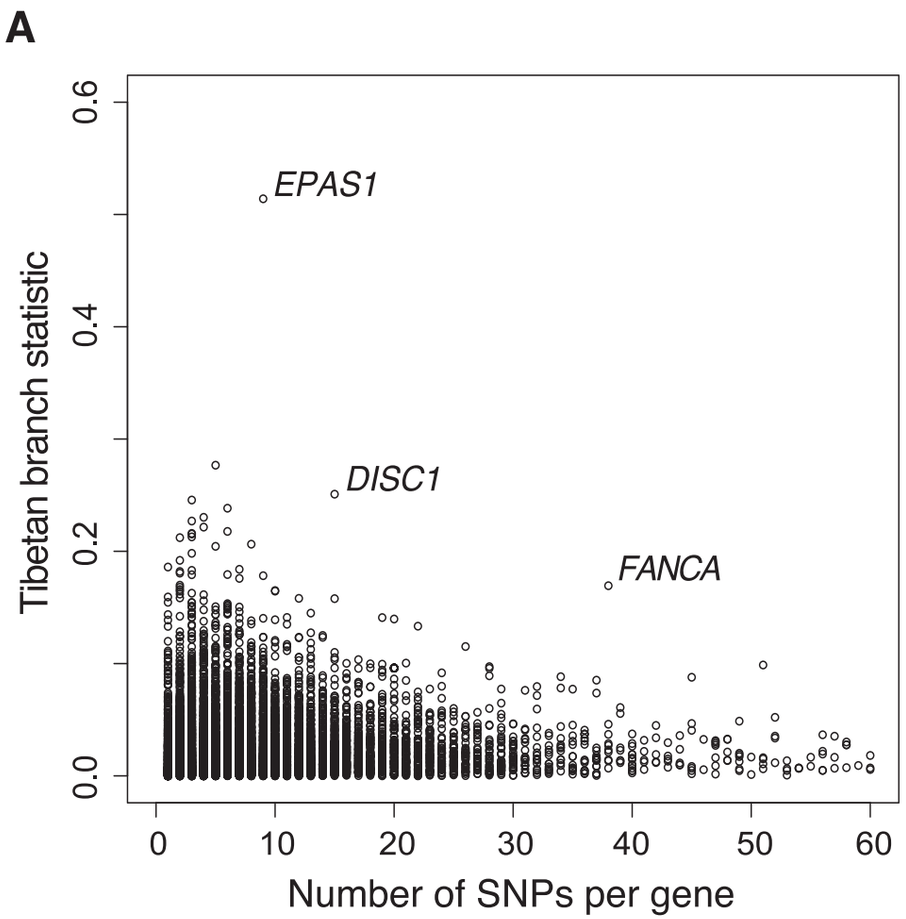


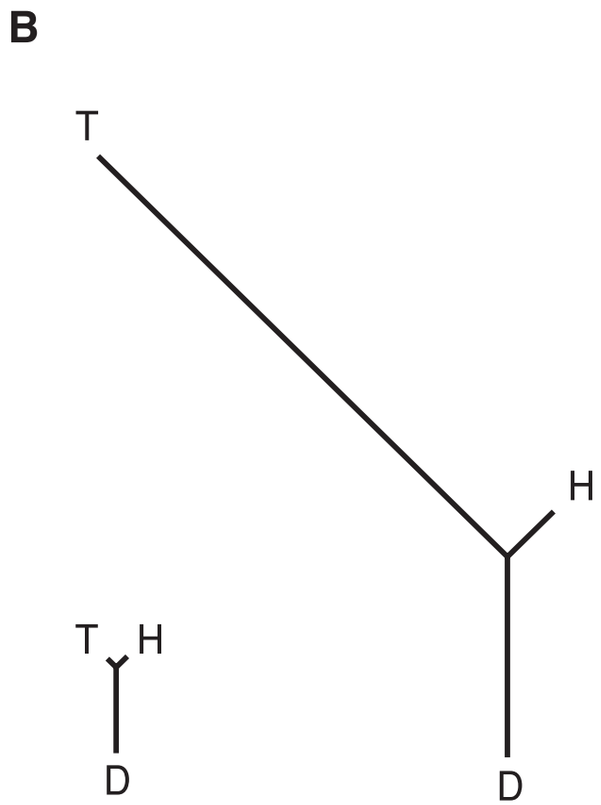
EPAS1: a transcription factor involved in response to hypoxia

- To find these types of signatures:
  - Compare allele frequencies using Fst

Yi, X. et al. *Science* **329**, 75-78 (2010).

Testing for Population Divergence



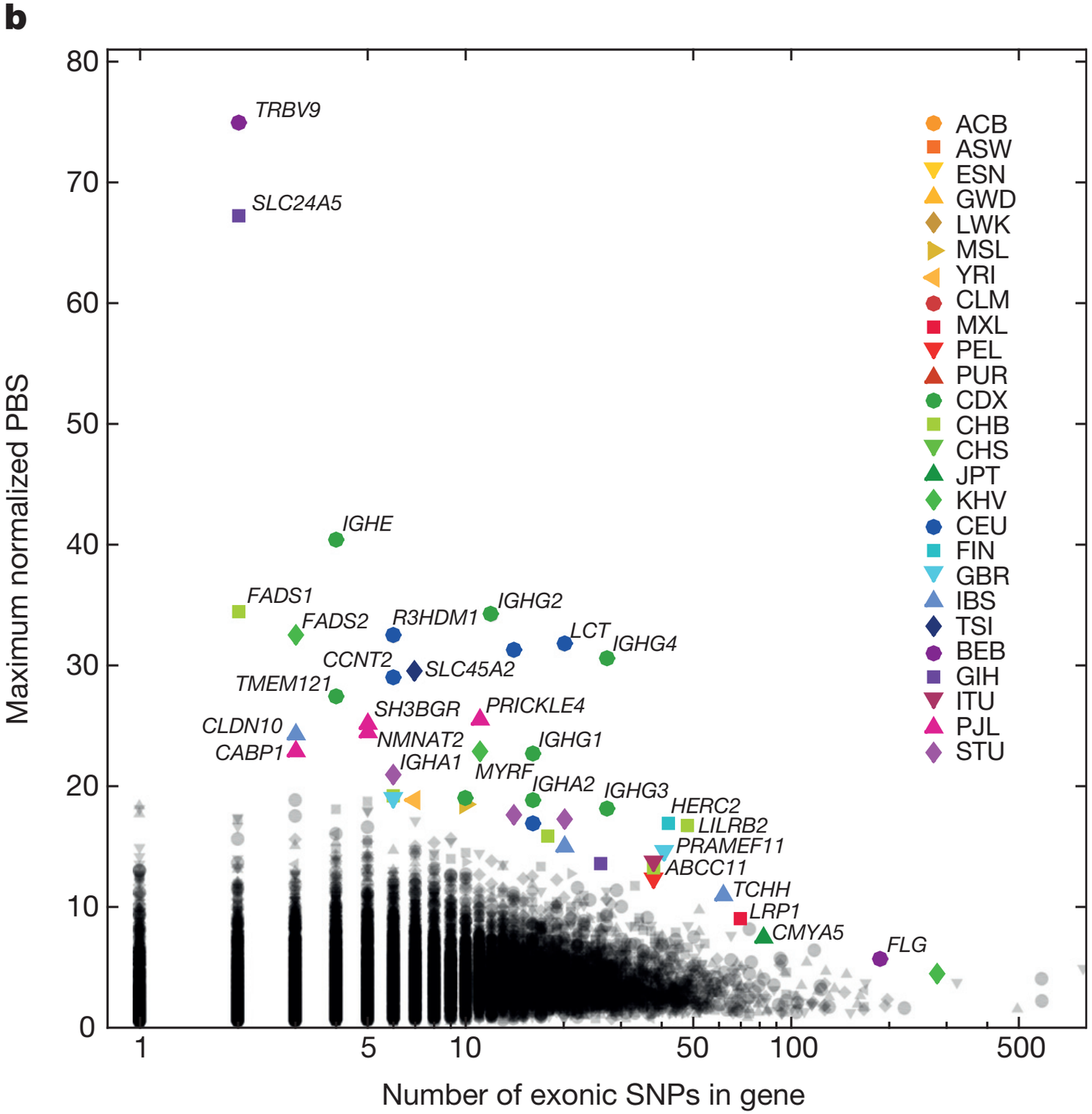


EPAS1: a transcription factor involved in response to hypoxia

- To find these types of signatures:
  - Compare allele frequencies using Fst

Yi, X. et al. *Science* **329**, 75-78 (2010).

Testing for Population Divergence



- Applying this statistic to 26 human populations
- Several known genes
- Several novel ones

1000 Genomes Project Consortium. ***Nature***, 526, 68–74 (2015).

Types of Positive Selection

☑ Selection acts in one population but not another

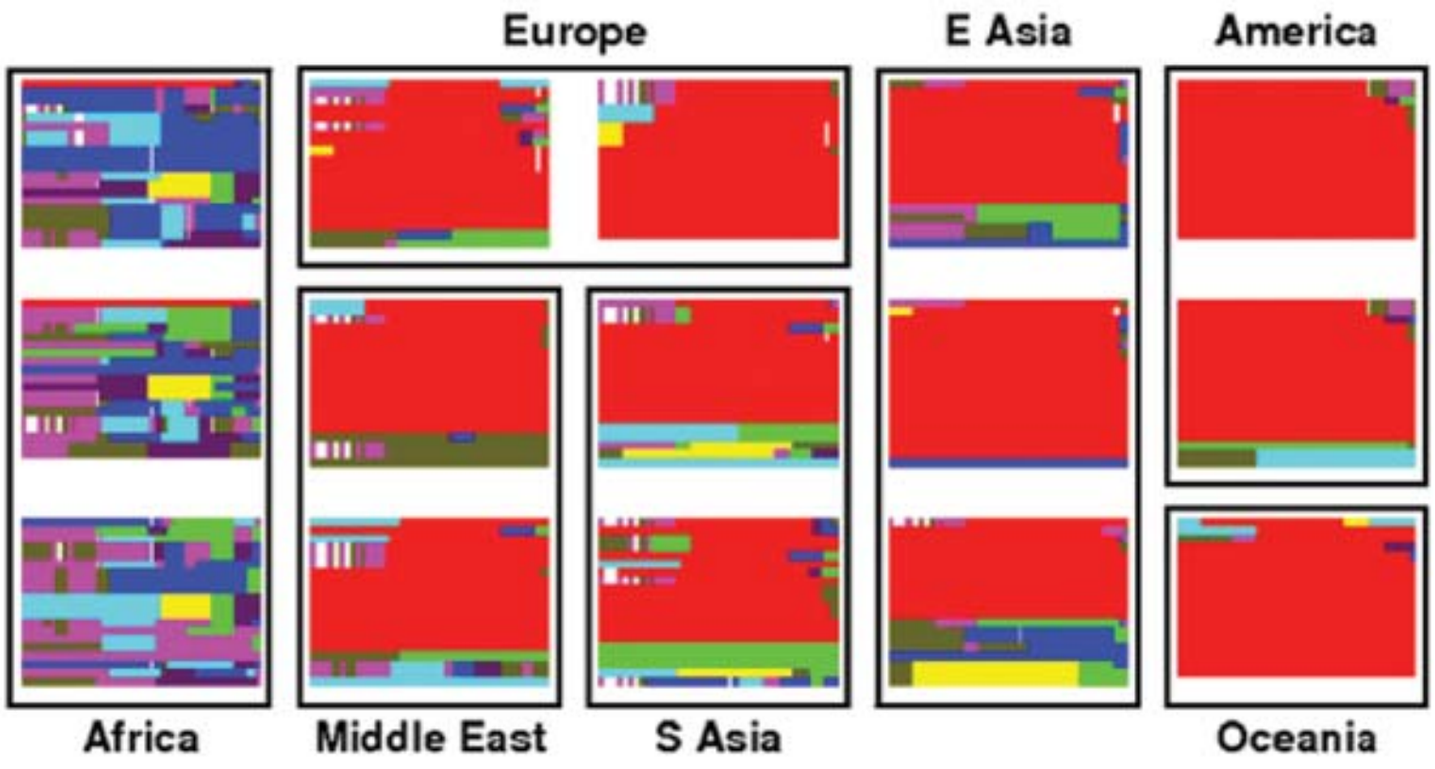
- Selection operates on a new mutation
  - Selection will act to increase the frequency of the allele
  - Results in a young allele at relatively high frequency
  - The test is simple:
    - Are there young alleles at unusually high frequency?

Testing for High Freq. Young Alleles

- The age of an allele can be assessed by measuring the amount of genetic variation around the allele.
    - As time passes:
        - Mutations occur nearby
        - Recombination breaks down the correlation between the allele and others nearby

Testing for High Freq. Young Alleles

- Example: Skin pigmentation
  - KITLG is a gene known to contribute to lighter skin in non-African populations.



- Each plot is a population.
- Each row is an individual's haplotype.
- Identical haplotypes have the same color.
- Large red blocks indicate long haplotypes with very little variation (i.e., young).

Testing for High Freq. Young Alleles

- Detecting these types of signatures:
    - Long Range Haplotype (**LRH**) or Extended Haplotype Homozygosity (**EHH**) {Sabeti, P. C. et al. Nature 419, 832-837 (2002)}.
    - integrated Haplotype Score (**iHS**) {Voight, B. F. et al. PLoS Biol 4, e72 (2006)}.
    - Composite Likelihood Ratio (**CLR**) {Williamson, S. H. et al. PLoS Genet 3, e90 (2007)}.

Types of Positive Selection

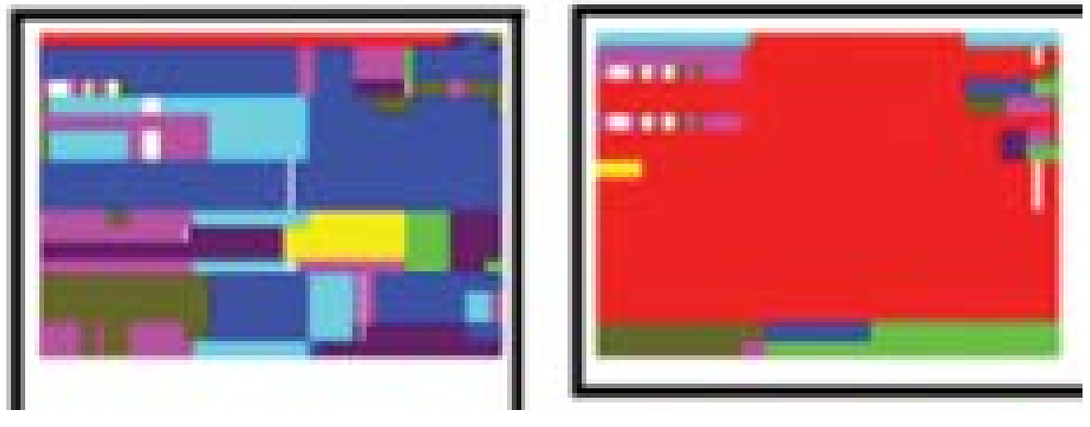
☑ Selection acts in one population but not another

☑ Selection acts on a new mutation

- Selection acts on <u>new mutations</u> primarily in one population
  - In this case, we expect high divergence and long haplotypes in one population

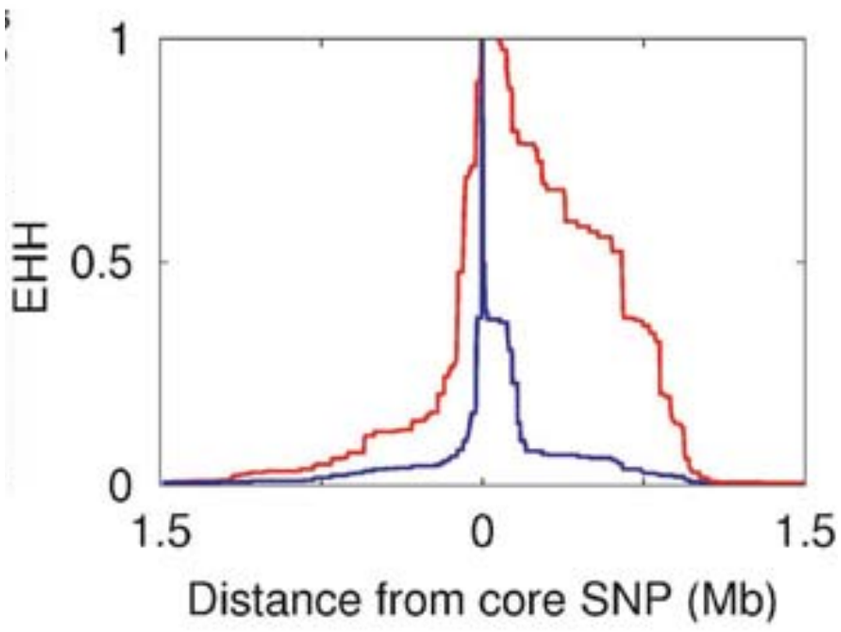
Divergence of a Young Allele

- Recall the haplotype patterns before for just two populations:



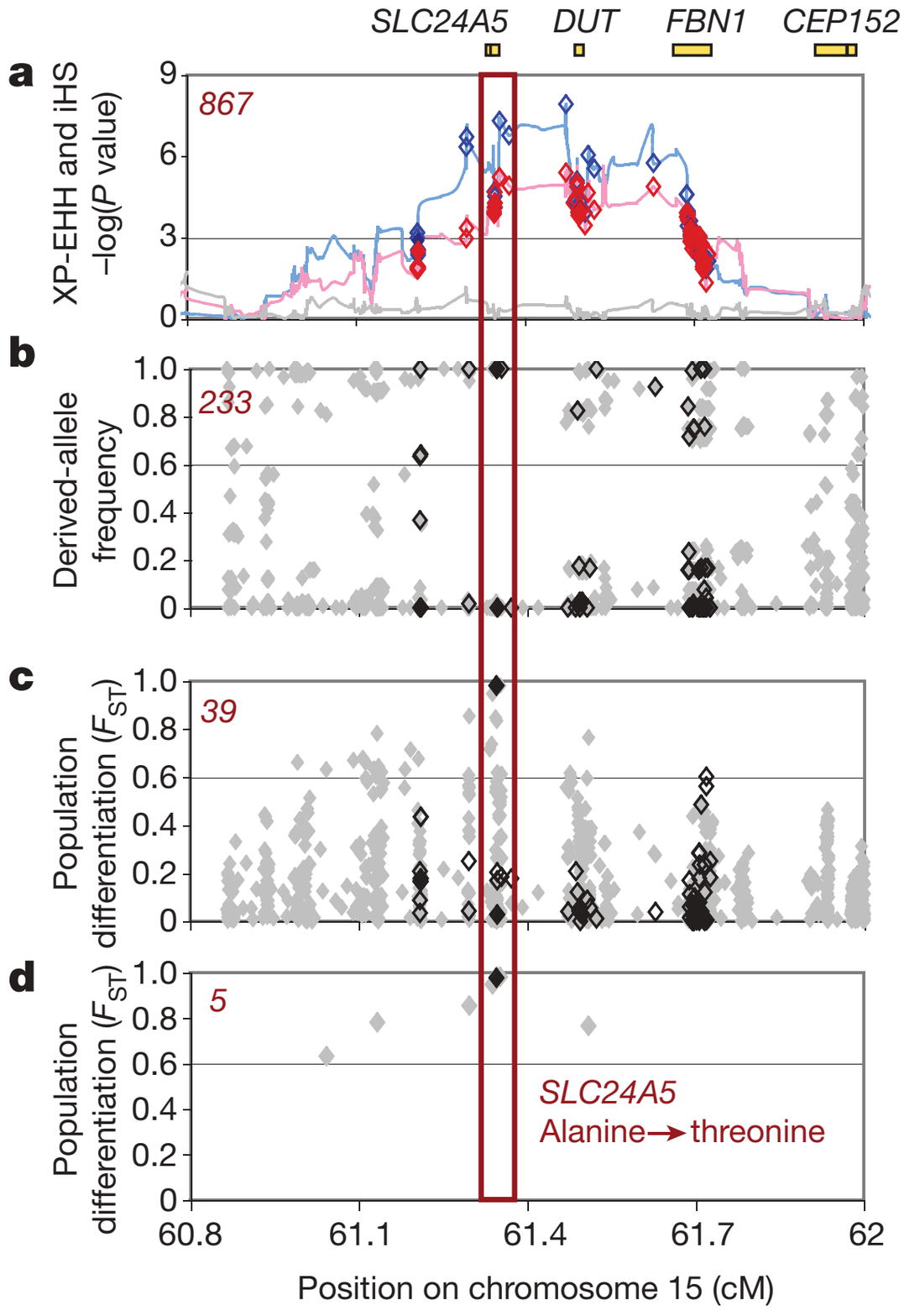
- These can be plotted as the probability that two randomly chosen individuals have an identical haplotype as a function of distance from the core SNP:

- Comparing the area under these two curves is the basis for XP-EHH



Divergence of a Young Allele

- XP-EHH rediscovers a nonsynonymous variant in SLC24A5 contributing to lighter skin outside Africa.



Sabeti, P. C. et al. *Nature* **449**, 913-918 (2007).

Motivation

- Why should we care about finding signatures of natural selection?
  - It's cool... It's what often drives speciation
  - Understanding disease/complex traits

Case Study: Kidney Disease in African Americans

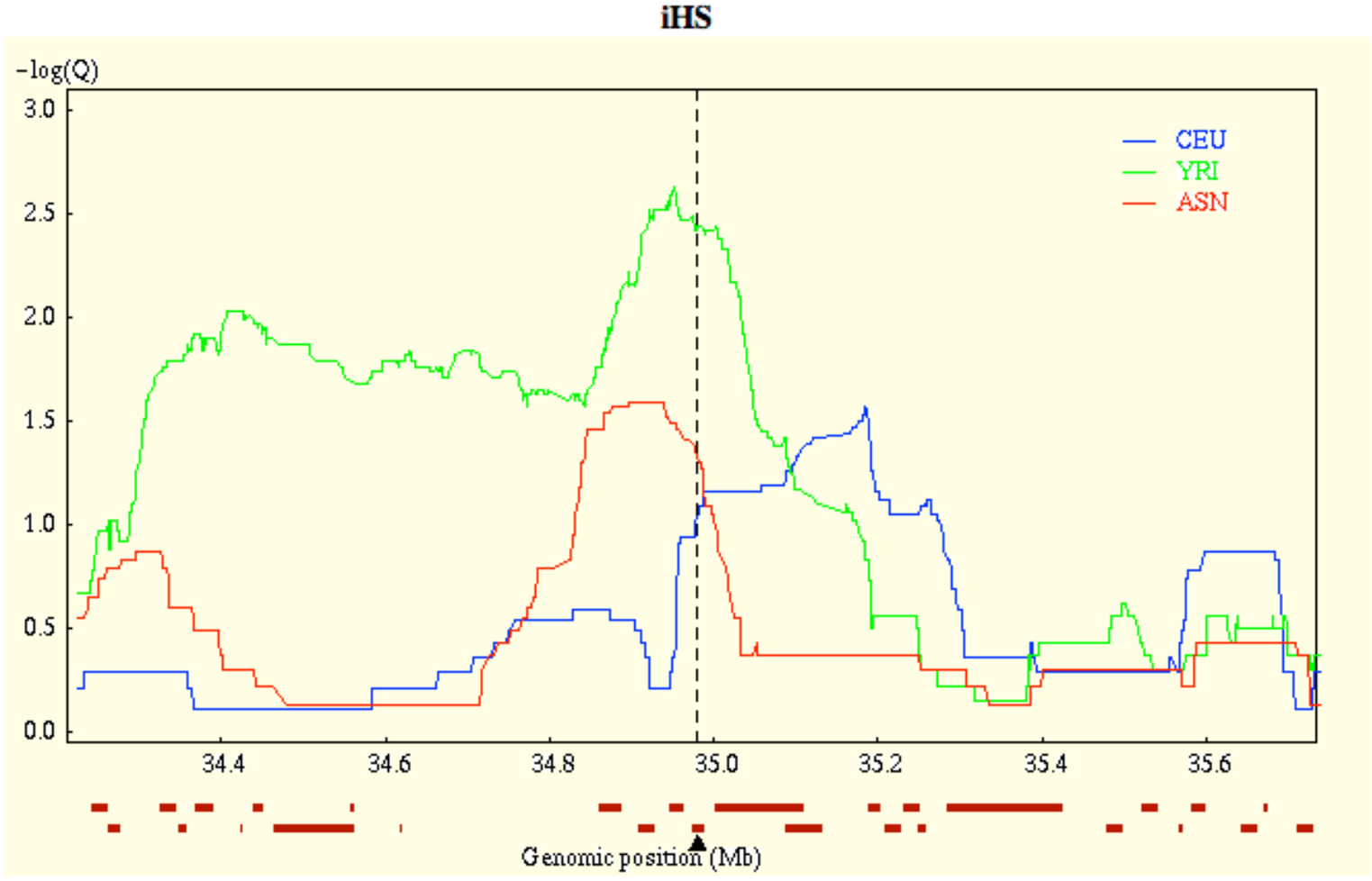
- Individuals of African descent have much higher incidence of kidney disease than individuals of European descent.
- GWAS had previously implicated the gene MYH9 with moderate effects ($p<10^{-8}$)
- But there was no clear biological story.

Genovese, G. et al. *Science* **329**, 841-845 (2010).

Case Study: Kidney Disease in African Americans

- Looking at signatures of selection adds valuable insight.
- Consider iHS from haplotter.uchicago.edu (more on this later):

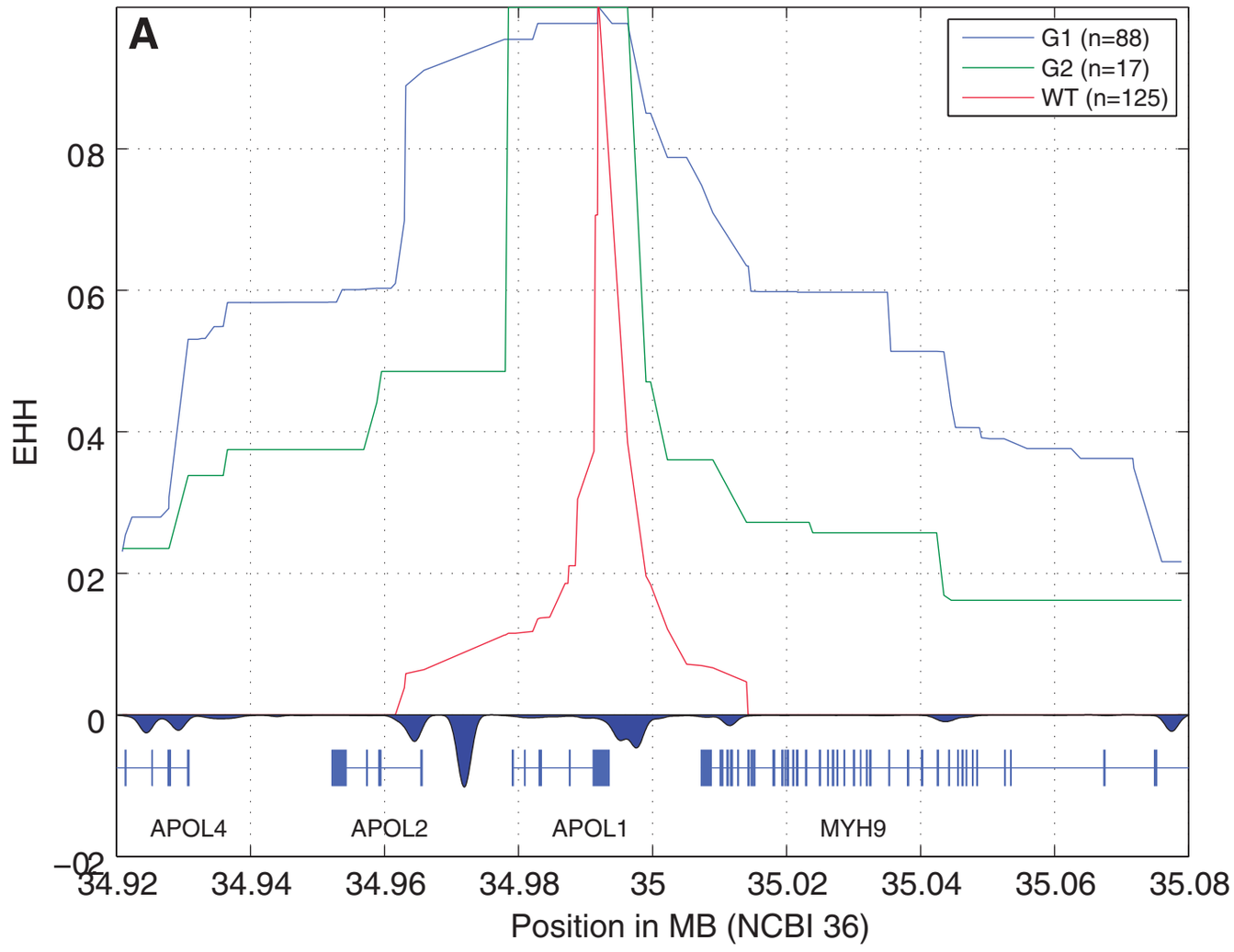
Gene	iHS p-value
APOL1	0.0033
MYH9	0.014



Genovese, G. et al. *Science* **329**, 841-845 (2010).

Case Study: Kidney Disease in African Americans

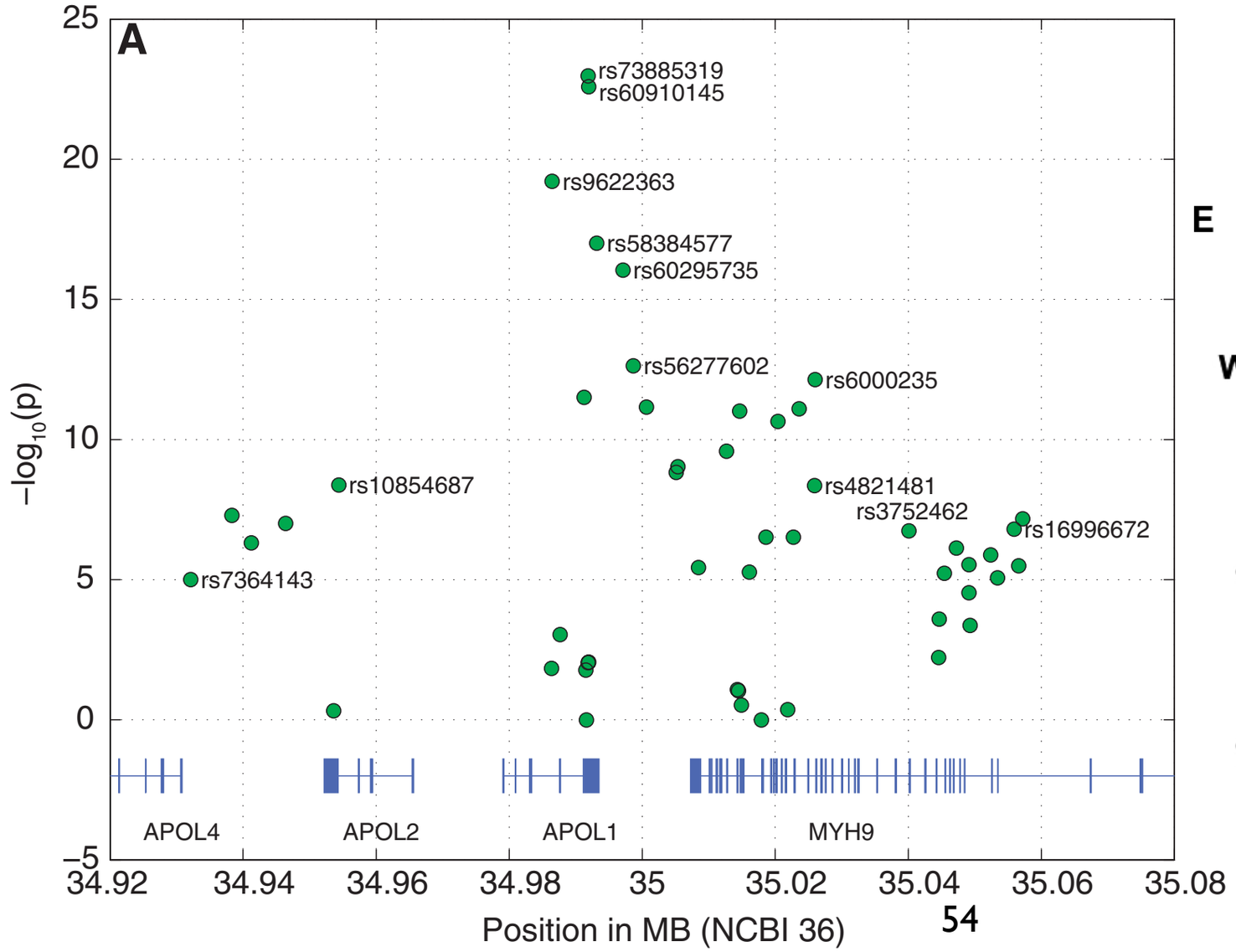
- Tag SNPs chosen across a broader region, and calculated EHH based on higher resolution data



Genovese, G. et al. *Science* **329**, 841-845 (2010).

Case Study: Kidney Disease in African Americans

- Subset of SNPs chosen based on signatures of selection genotyped on a larger panel strongly implicates APOL1!



Risk alleles confer resistance to trypanosomes (swelling of the lysosome).

Genovese, G. et al. *Science* **329**, 841-845 (2010).

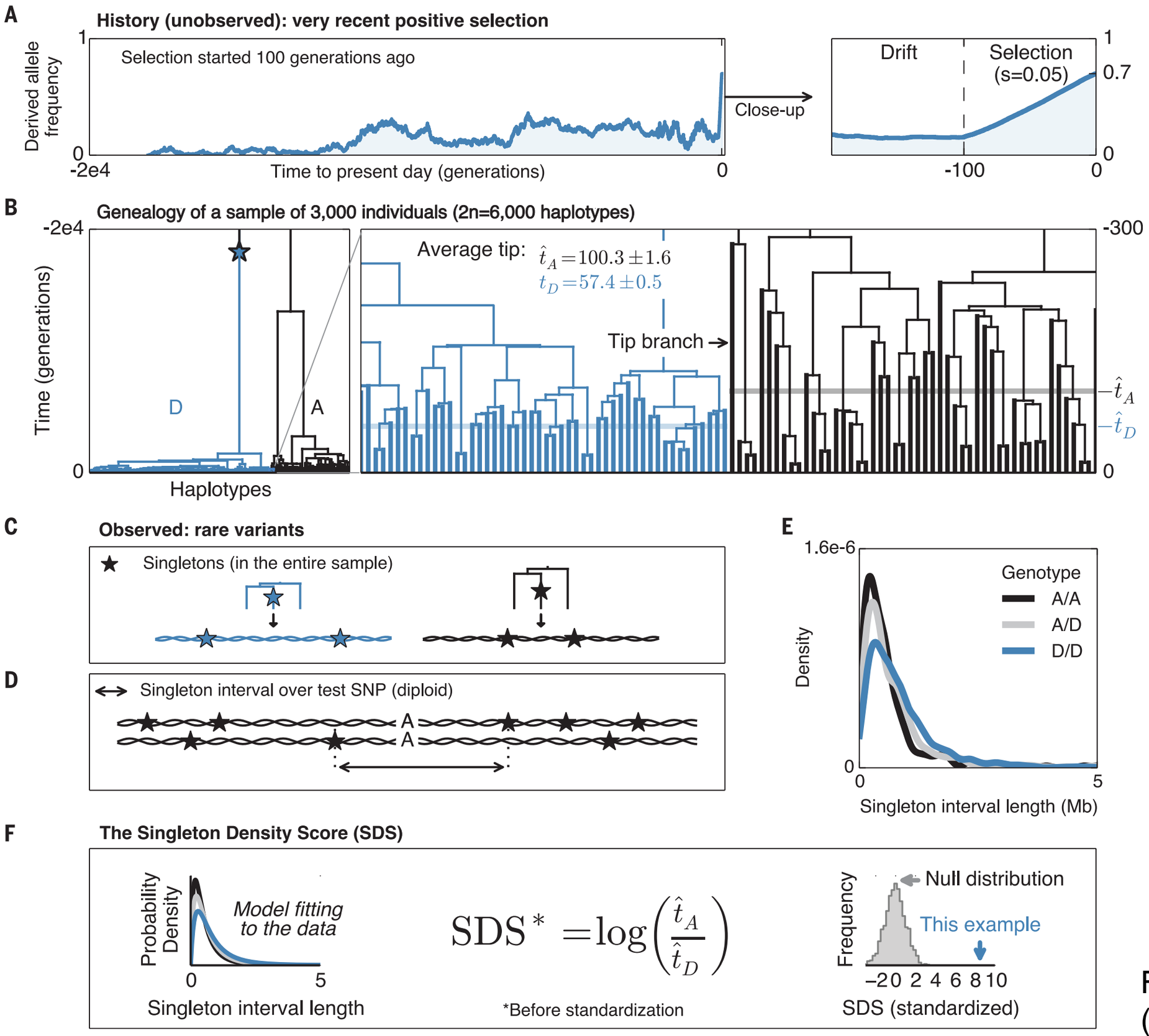
WGS

- The statistics described do not really handle whole genome sequencing data (WGS).
- Further, the timescale for when selection acted is not very well specified.
- With an abundance of rare variants, WGS should be informative about recent selection.
- Enter the Singleton Density Score (SDS).

SDS

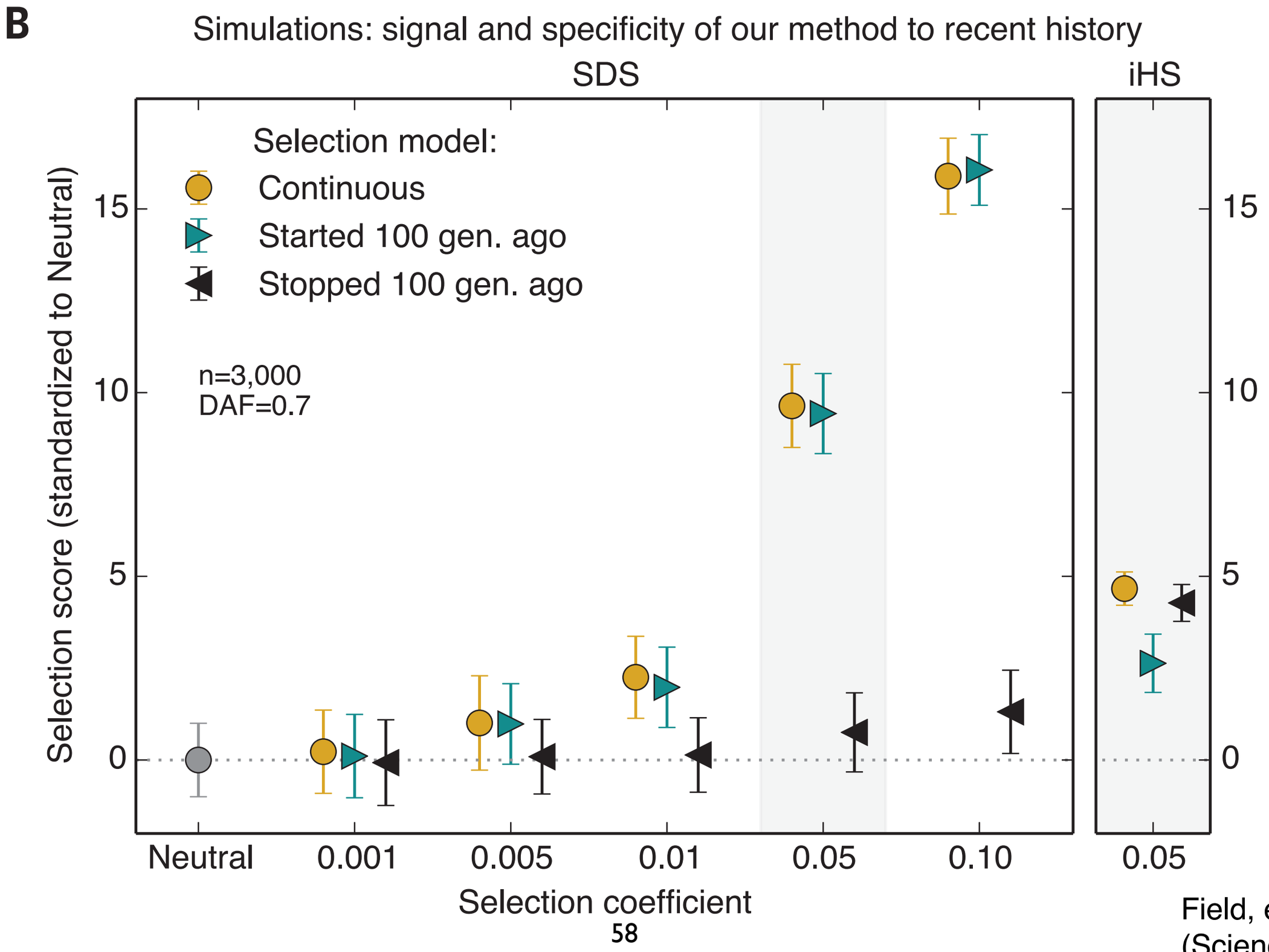
- Field, et al. (*Science*, 2016) introduced the Singleton Density Score (SDS) to capitalize on WGS data with very large samples.
- In the presence of a sweep, the distribution of distances (across individuals) to the nearest singleton will be skewed towards longer distances.

SDS



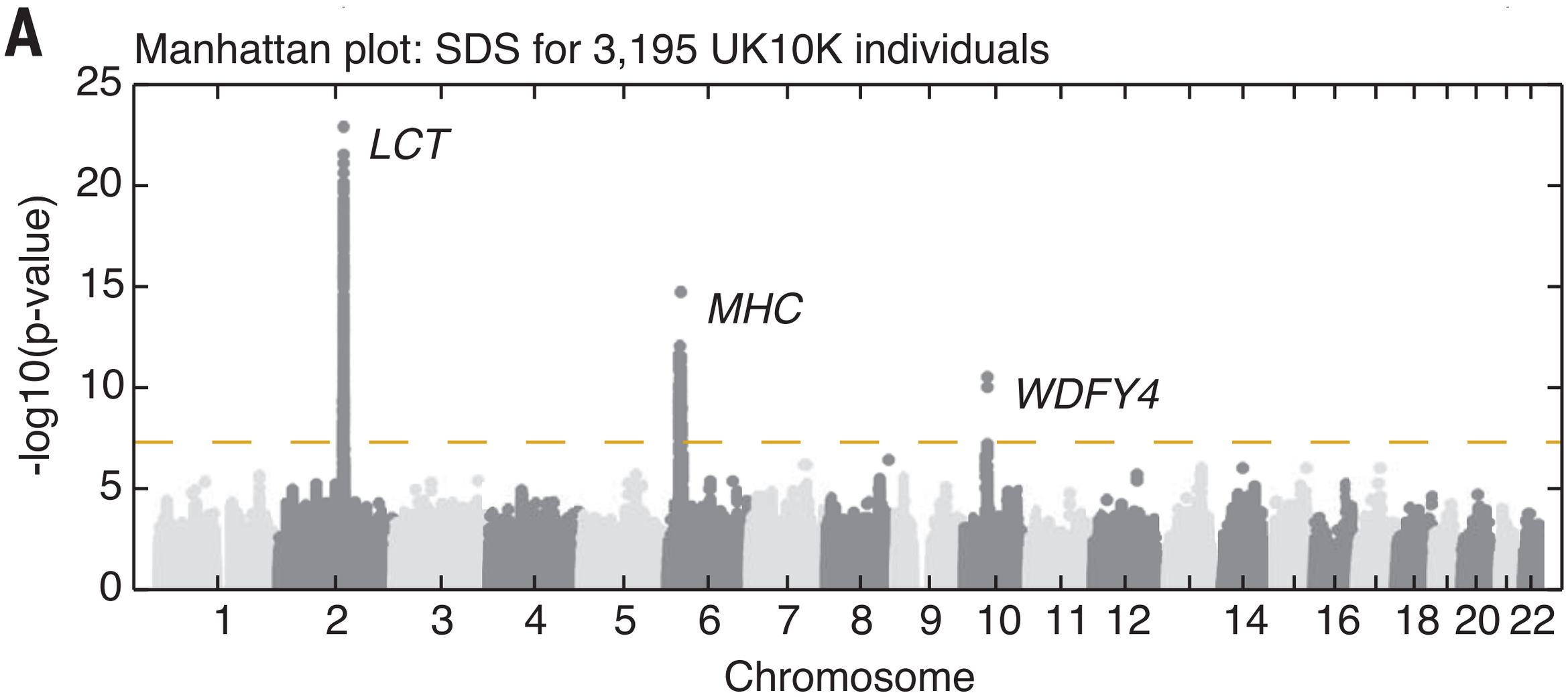
Field, et al. (Science, 2016)

SDS



Field, et al.
(Science, 2016)

SDS



Field, et al.
(Science, 2016)

Conclusions

- Natural selection leaves distinctive footprints within patterns of genetic variation.
- This occurs because alleles driven by natural selection tend to be younger than neutral alleles at the same frequency.
- Characterizing signatures of natural selection around disease associated loci can sometimes illuminate mechanistic relationships.