Deeper Introduction to Positive Selection

Ryan Hernandez
Tim O'Connor

ryan.hernandez@ucsf.edu

Goals

- Understand characteristic genomic signatures left by natural selection
- Learn about a few tools that can be used to search for natural selection (and web resources)
- Become familiar with searching for recent natural selection using iHS and XP-EHH.
- There are many other ways of detecting selection!

Overview

- Genetic variation comes in many forms:
  - tag SNPs in candidate regions (10-1000)
  - Genome wide SNP chip data (100,000-5,000,000)
    - candidate gene sequencing
    - exome sequencing
    - genome sequencing
- The signature of selection you look for depends on the type of data you have.

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.

The Effect of Positive Selection

Adaptive

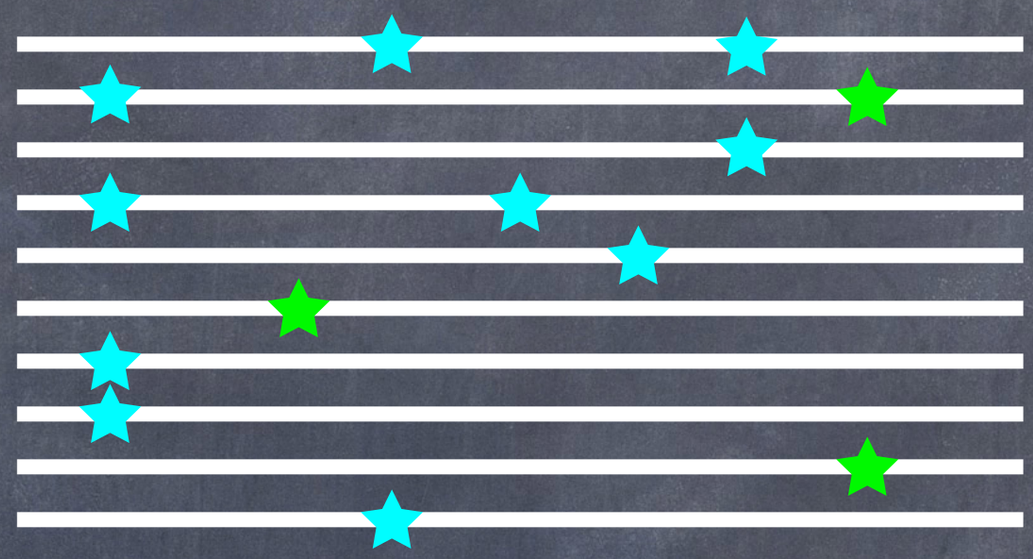
Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious



The Effect of Positive Selection

Adaptive

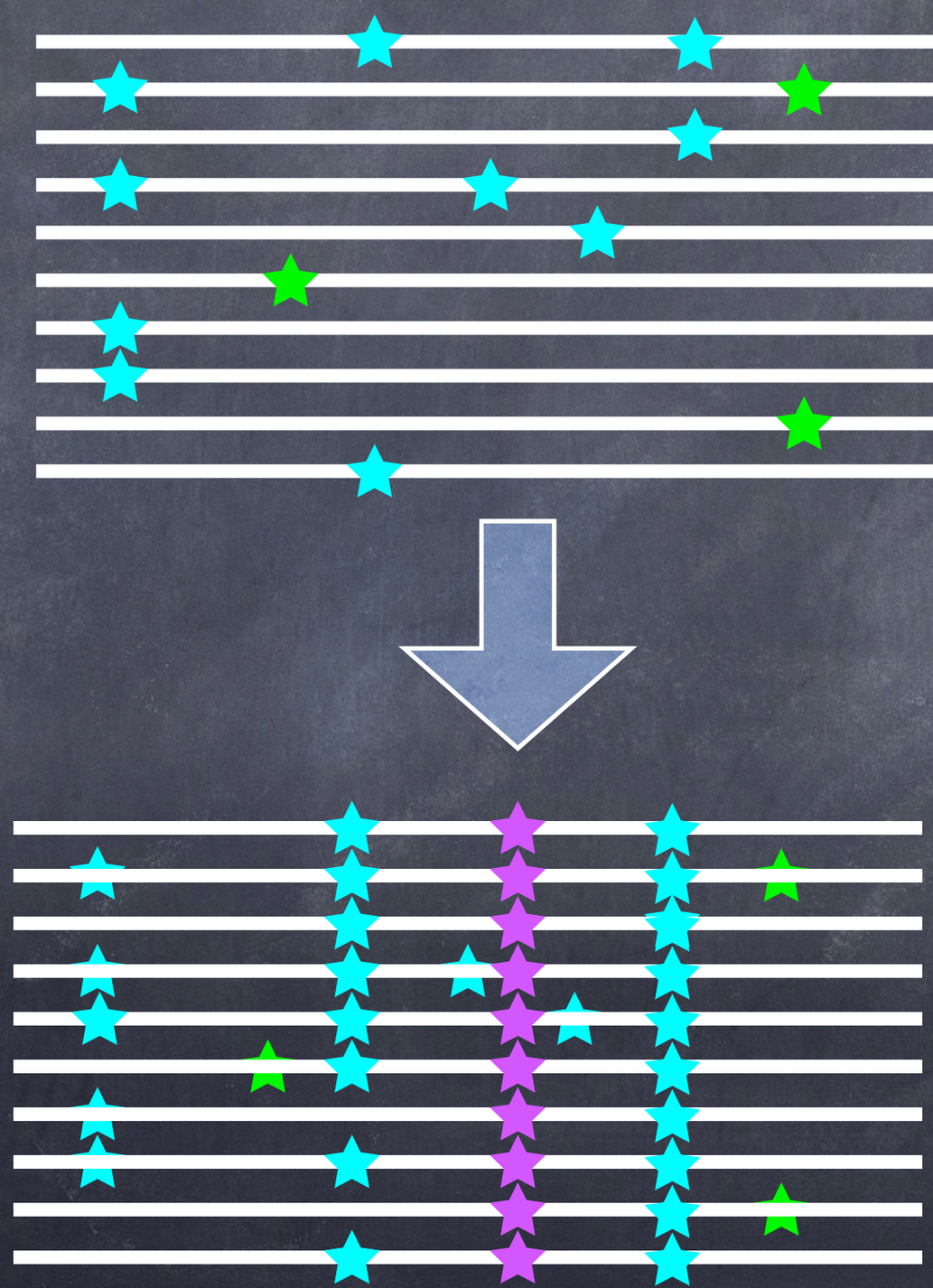
Neutral

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

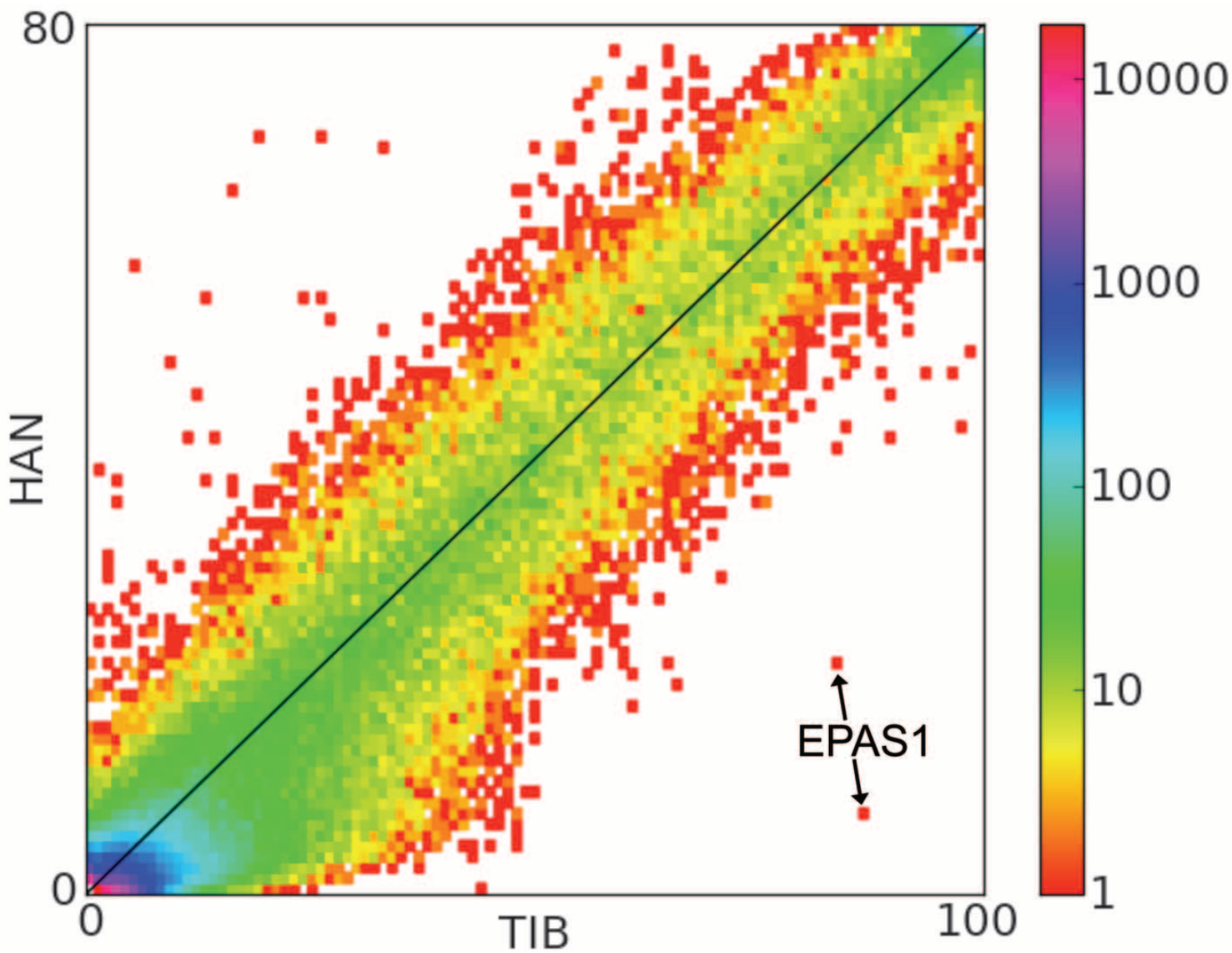
Testing for Population Divergence

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

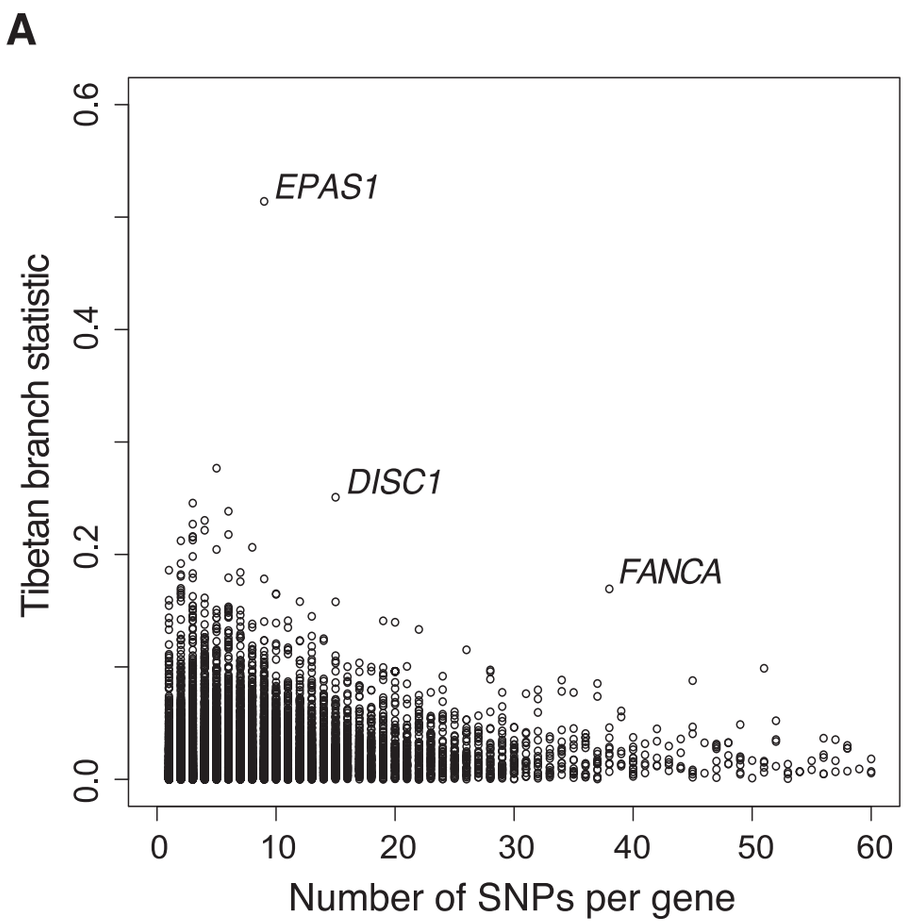


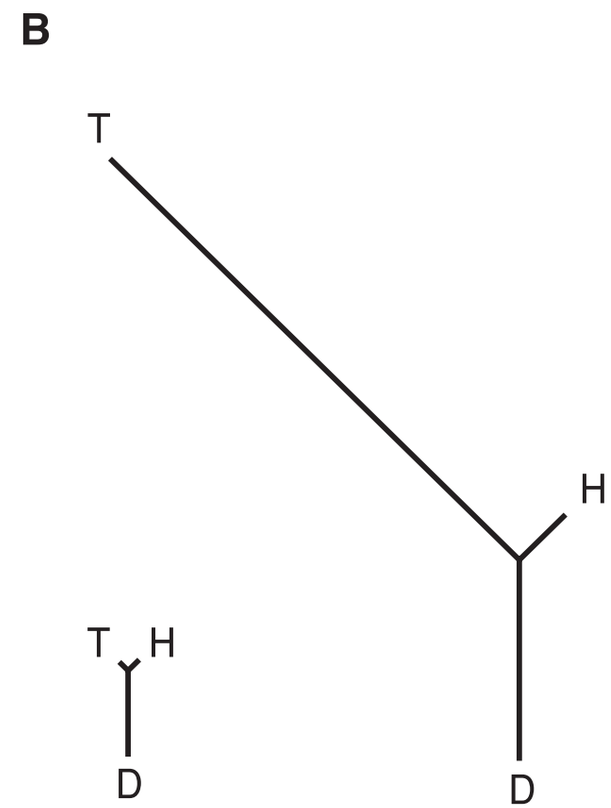
EPAS1: a transcription factor involved in response to hypoxia

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

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

Testing for Population Divergence



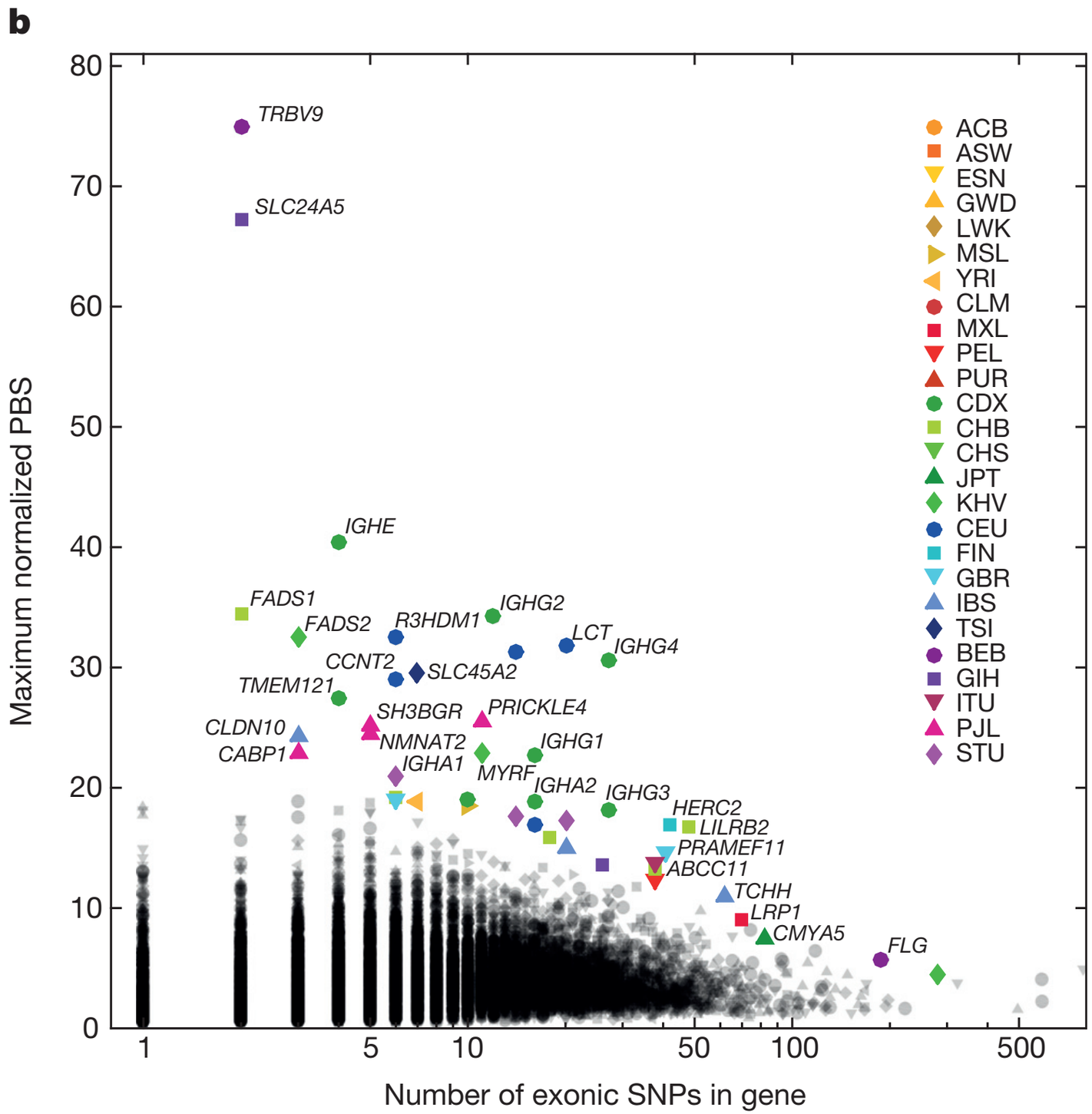


EPAS1: a transcription factor involved in response to hypoxia

- 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 populations from 1000 Genomes
- 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

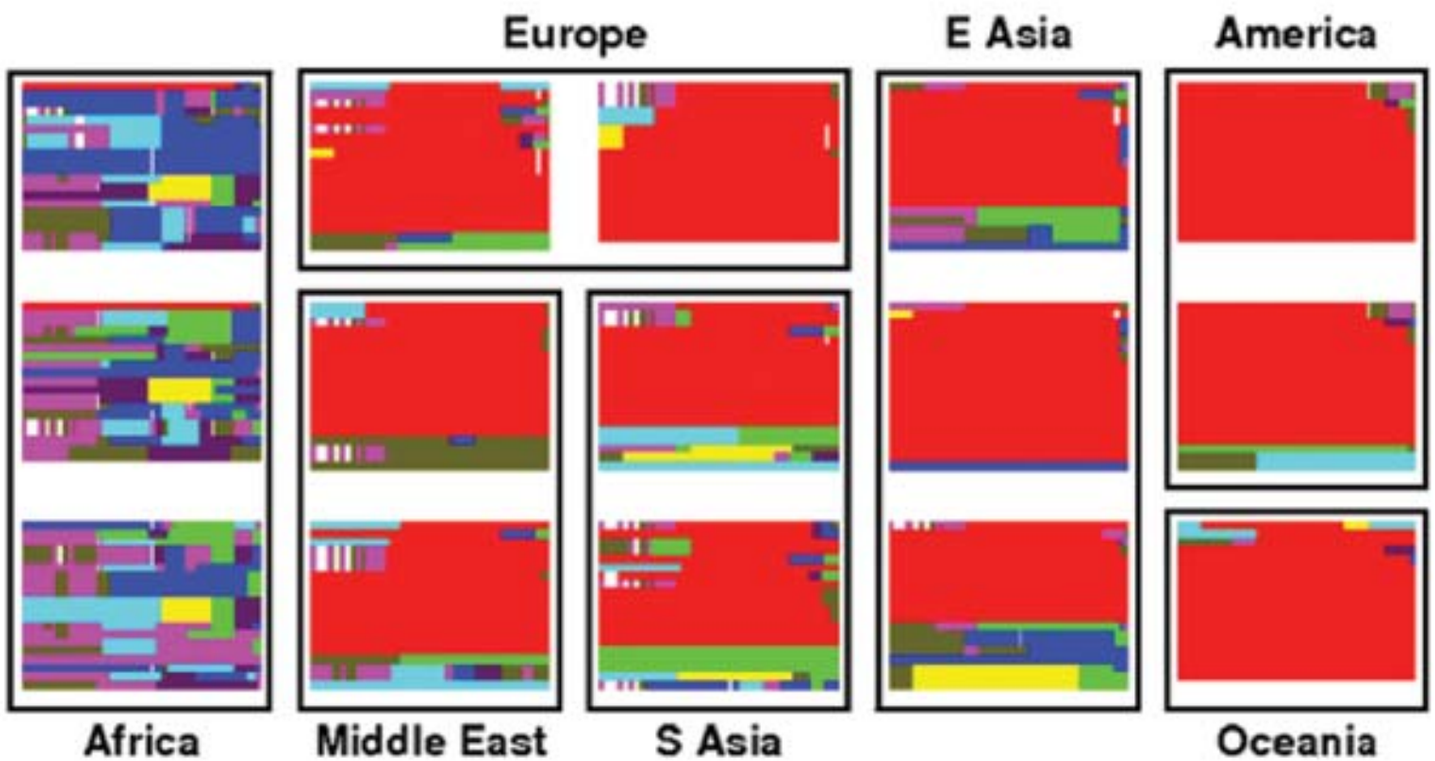
- Selection operates on a new mutation
  - Selection will act to increase the frequency of the allele
  - Results in a young allele at relatively high frequency
  - The test is simple:
    - Are there young alleles at unusually high frequency?

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

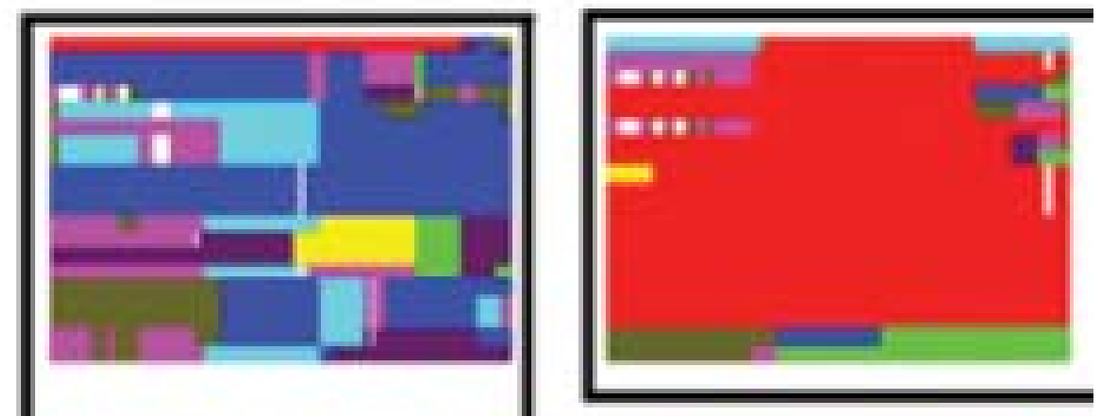
- Detecting these types of signatures:
  - Long Range Haplotype (**LRH**) or Extended Haplotype Homozygosity (**EHH**) {Sabeti, P. C. et al. Nature 419, 832-837 (2002)}.
  - integrated Haplotype Score (**iHS**) {Voight, B. F. et al. PLoS Biol 4, e72 (2006)}.
  - Composite Likelihood Ratio (**CLR**) {Williamson,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

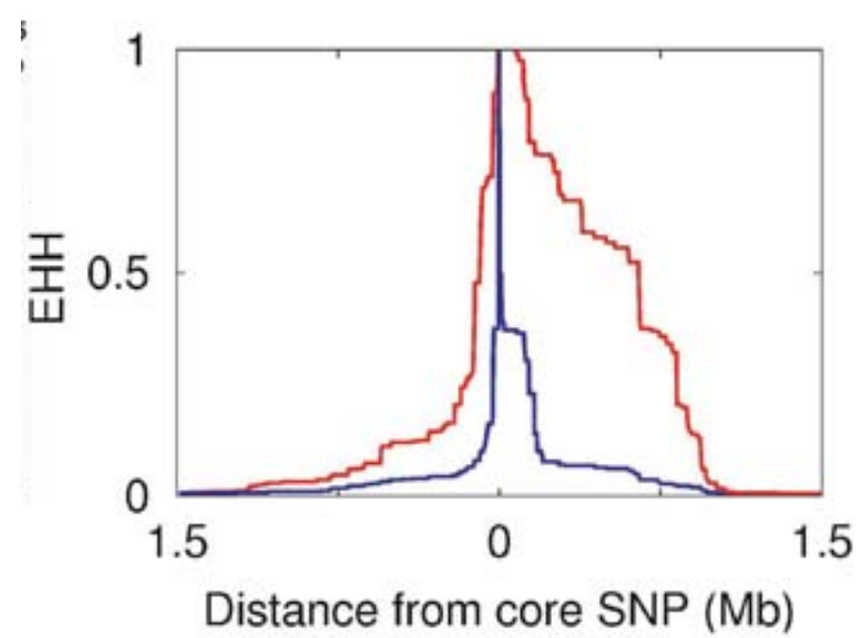
Divergence of a Young Allele

- Recall the haplotype patterns before for just two populations:



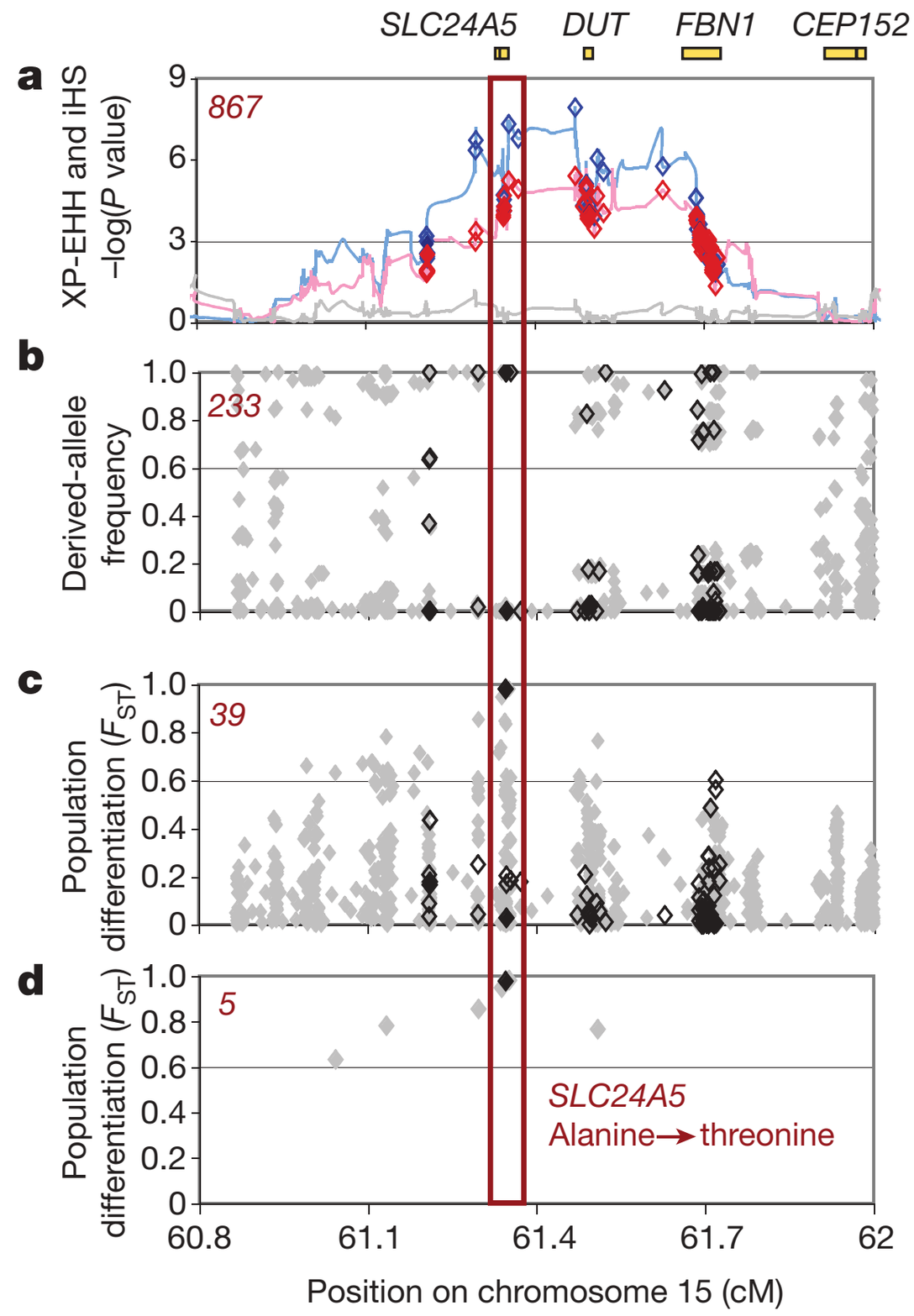
- These can be plotted as the probability that two randomly chosen individuals have an identical haplotype as a function of distance from the core SNP:

- 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 makes us human
  - Understanding disease

Case Study: Kidney Disease in African Americans

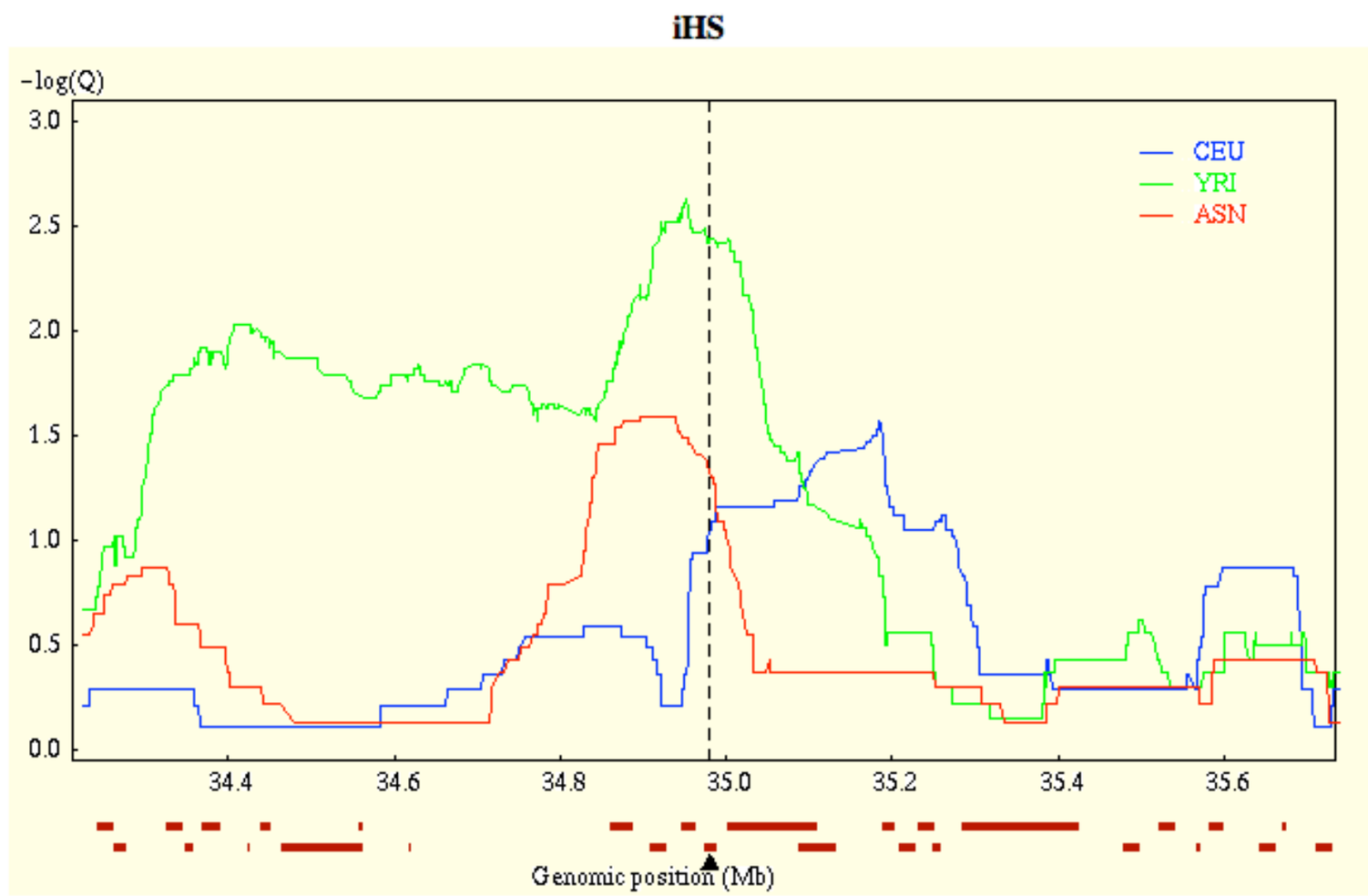
- Individuals of African descent have much higher incidence of kidney disease than individuals of European descent.
- GWAS had previously implicated the gene MYH9 with moderate effects ($p<10^{-8}$)
- But there was no clear biological story.

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):

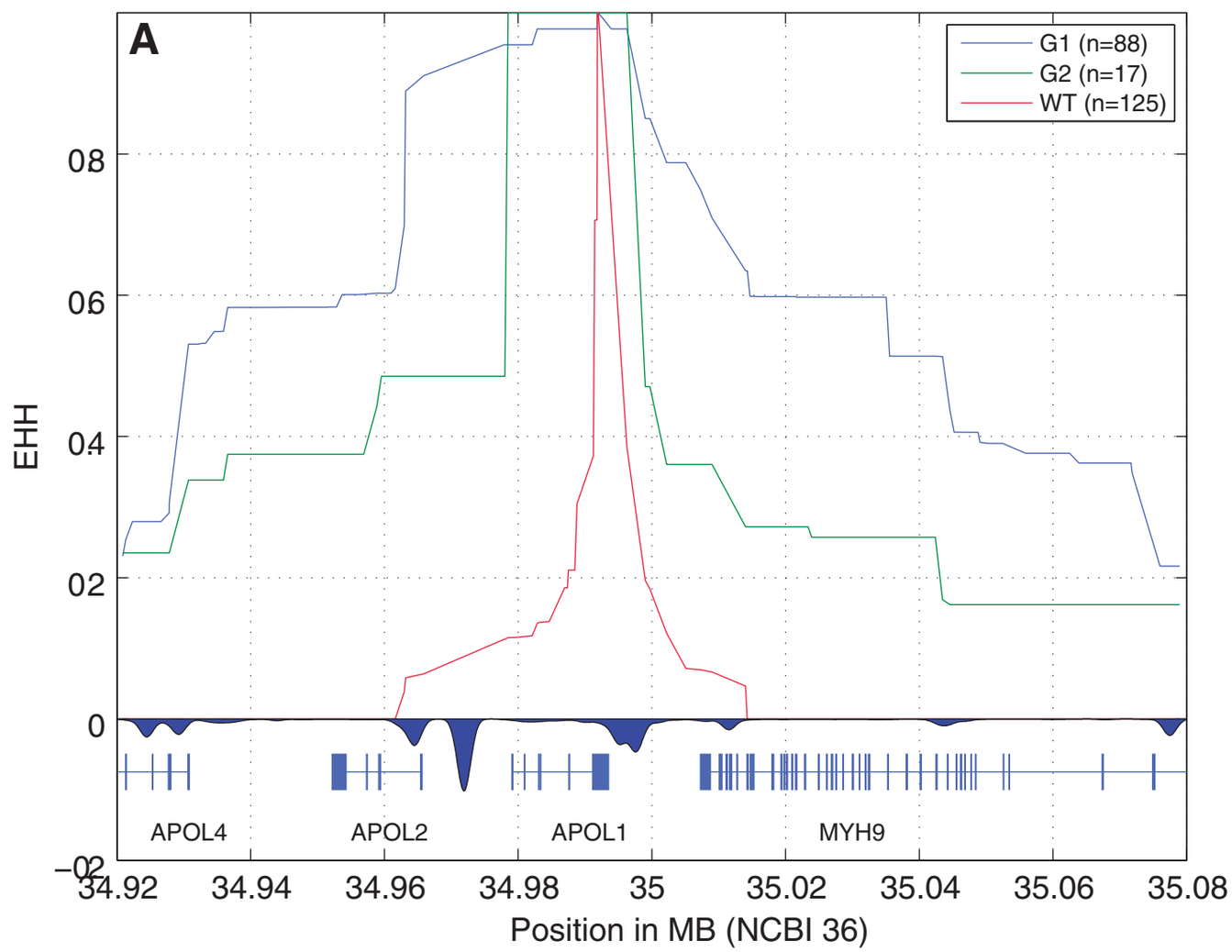
Gene	iHS p-value
APOL1	0.0033
MYH9	0.014



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

Case Study: Kidney Disease in African Americans

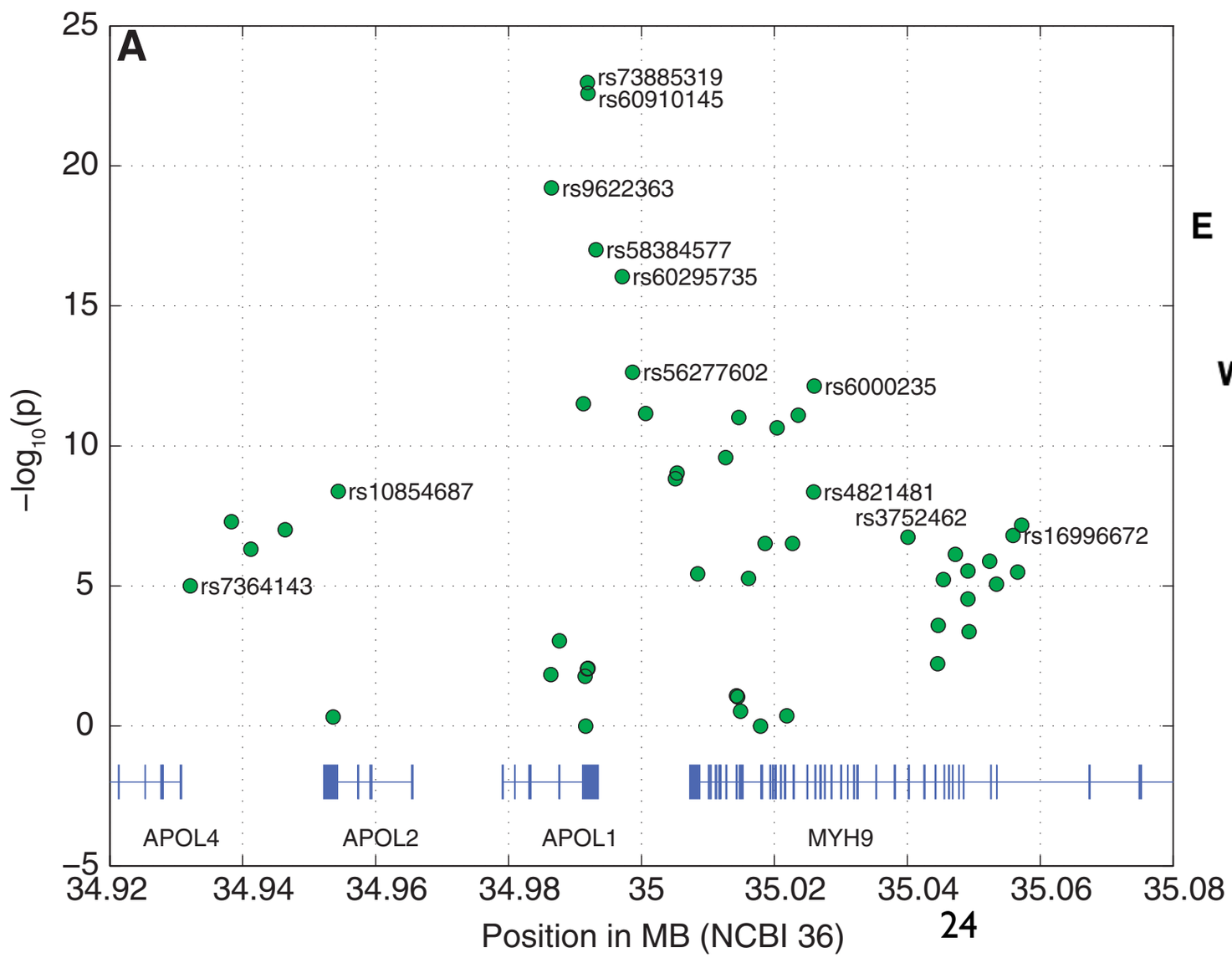
- Tag SNPs chosen across a broader region, and calculated EHH based on higher resolution data

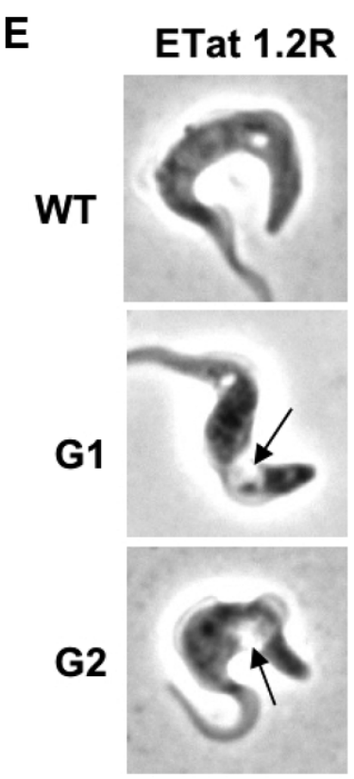


Genovese, G. et al. *Science* **329**, 841-845 (2010).

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).

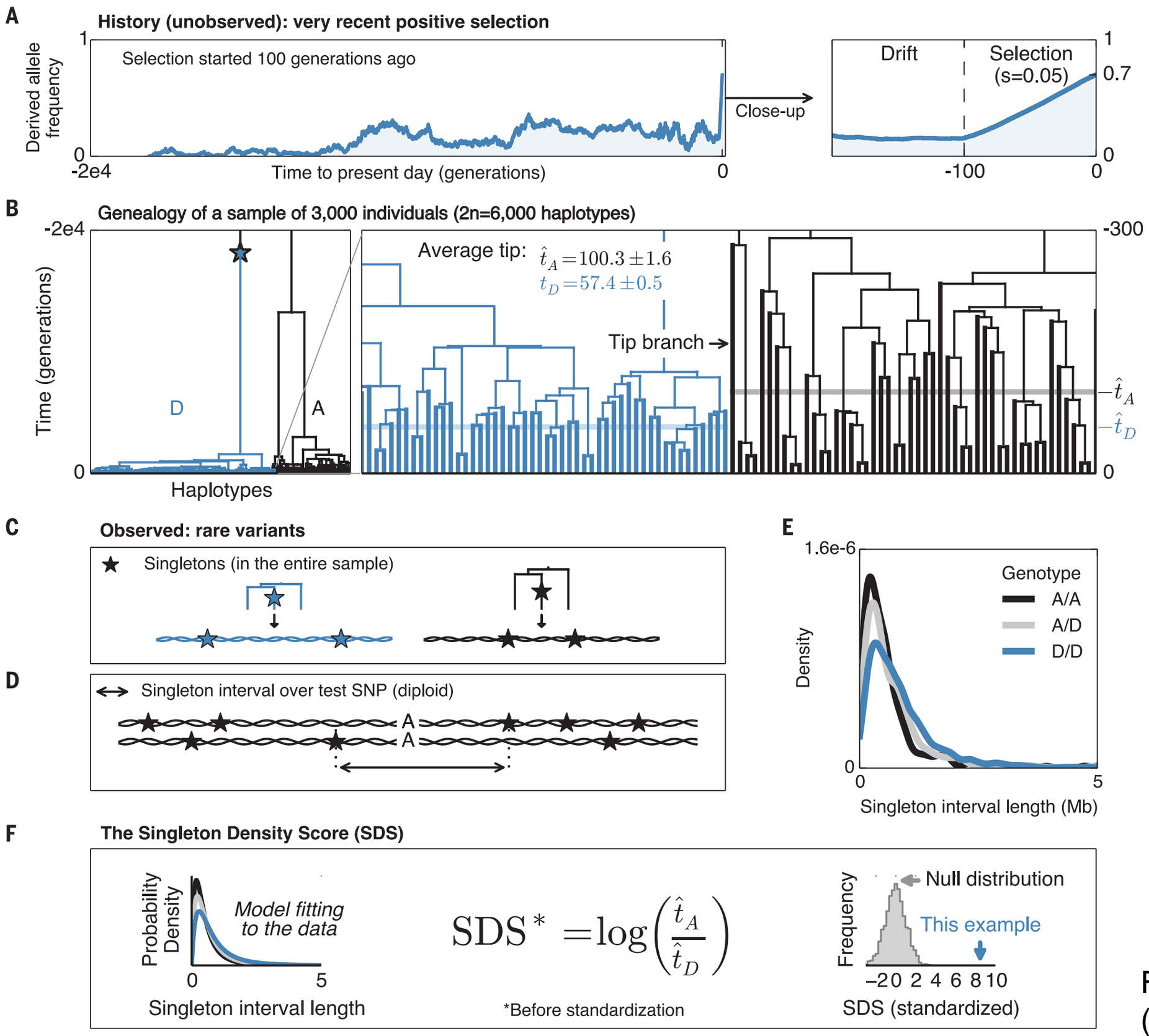
WGS

- 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

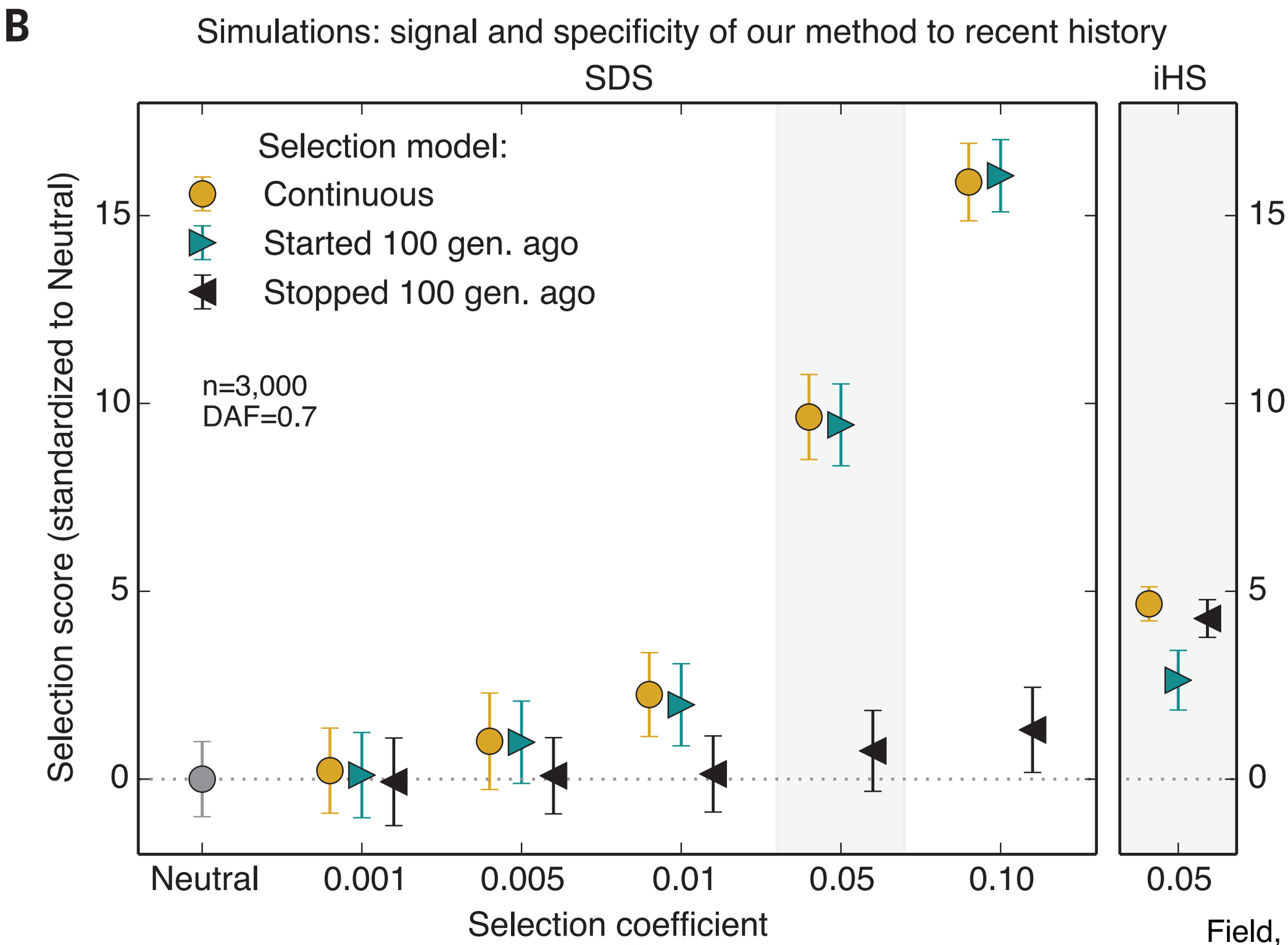
- Field, et al. (*Science*, 2016) introduced the Singleton Density Score (SDS) to capitalize on WGS data with very large samples.
- In the presence of a sweep, the distribution of distances (across individuals) to the nearest singleton will be skewed towards longer distances.

SDS



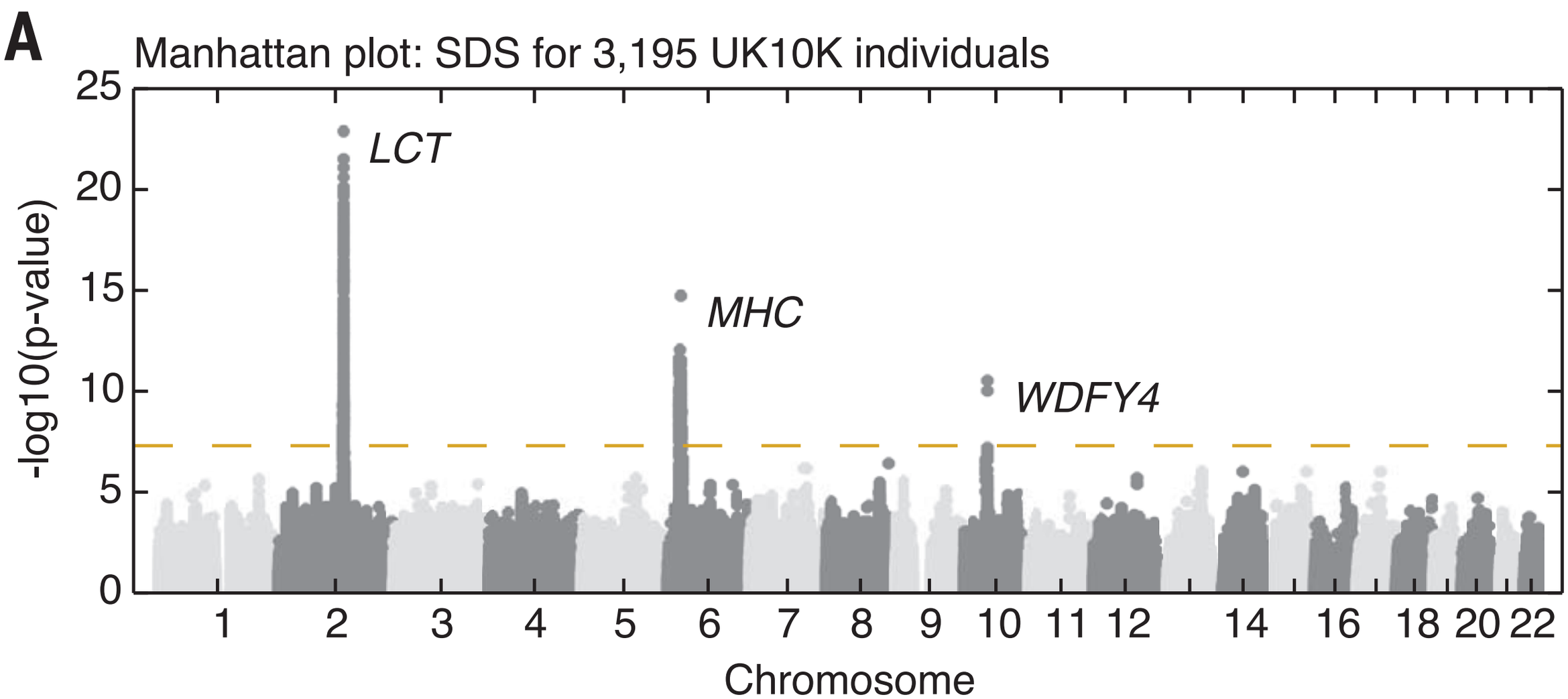
Field, et al. (Science, 2016)

SDS



Field, et al.
(Science, 2016)

SDS



Field, et al.
(Science, 2016)

Conclusions

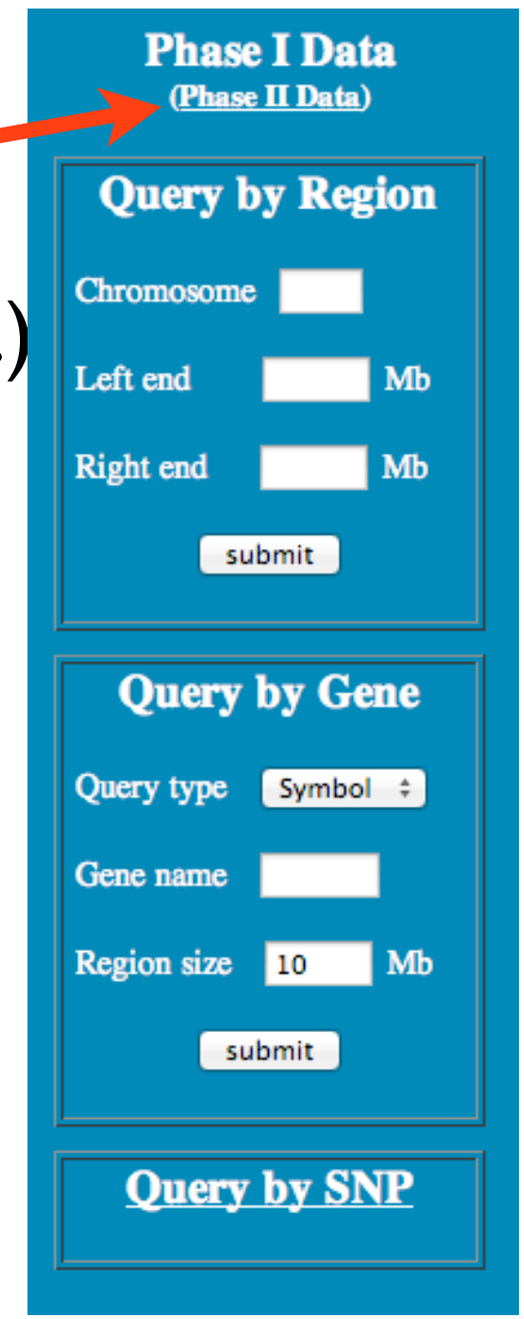
- Natural selection leaves distinctive footprints within patterns of genetic variation.
- This occurs because alleles driven by natural selection tend to be younger than neutral alleles at the same frequency.
- Characterizing signatures of natural selection around disease associated loci can sometimes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

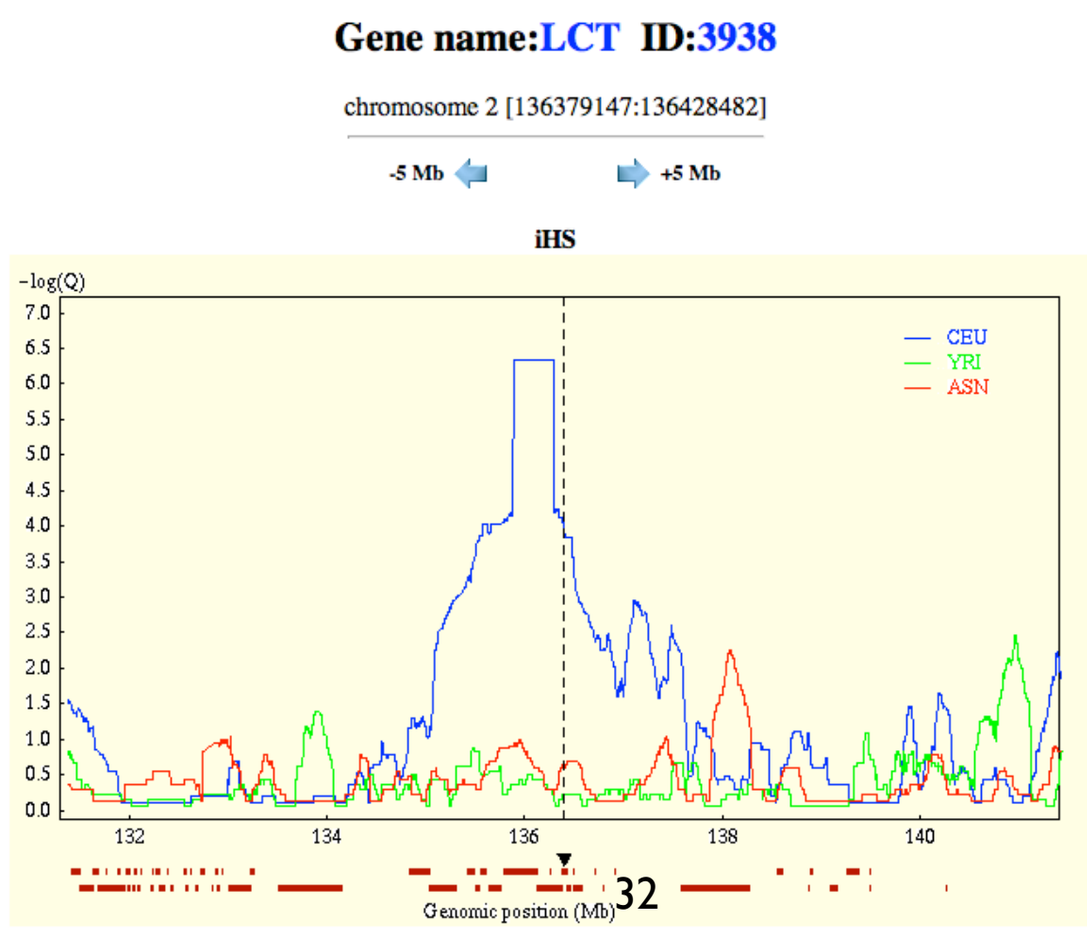
Web Resources

- Two easy web servers for signatures of natural selection:
  - http://haplotter.uchicago.edu/
    - Based on HapMap data
    - displays iHS, and two summary statistics of allele frequencies (H and D)
  - http://hgdp.uchicago.edu/
    - Based on Human Genome Diversity Panel (HGDP)
    - Calculates heterozygosity, iHS, Fst, and XP-EHH

Haplotter

- Send your browser to http://haplotter.uchicago.edu
- Click (Phase II Data) in upper left corner.
- Now enter your favorite gene into the Gene name box below (e.g., LCT, ApoL1, etc.)

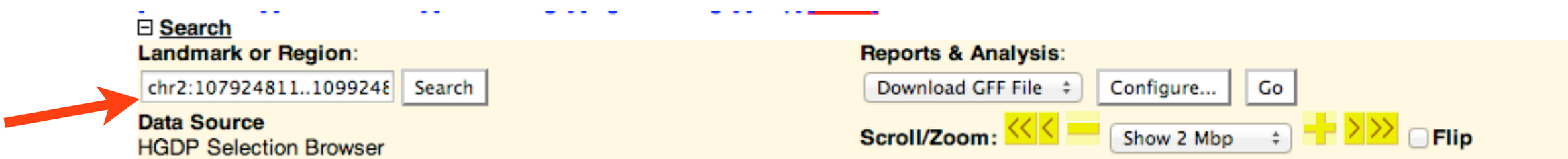




HGDP Selection Browser

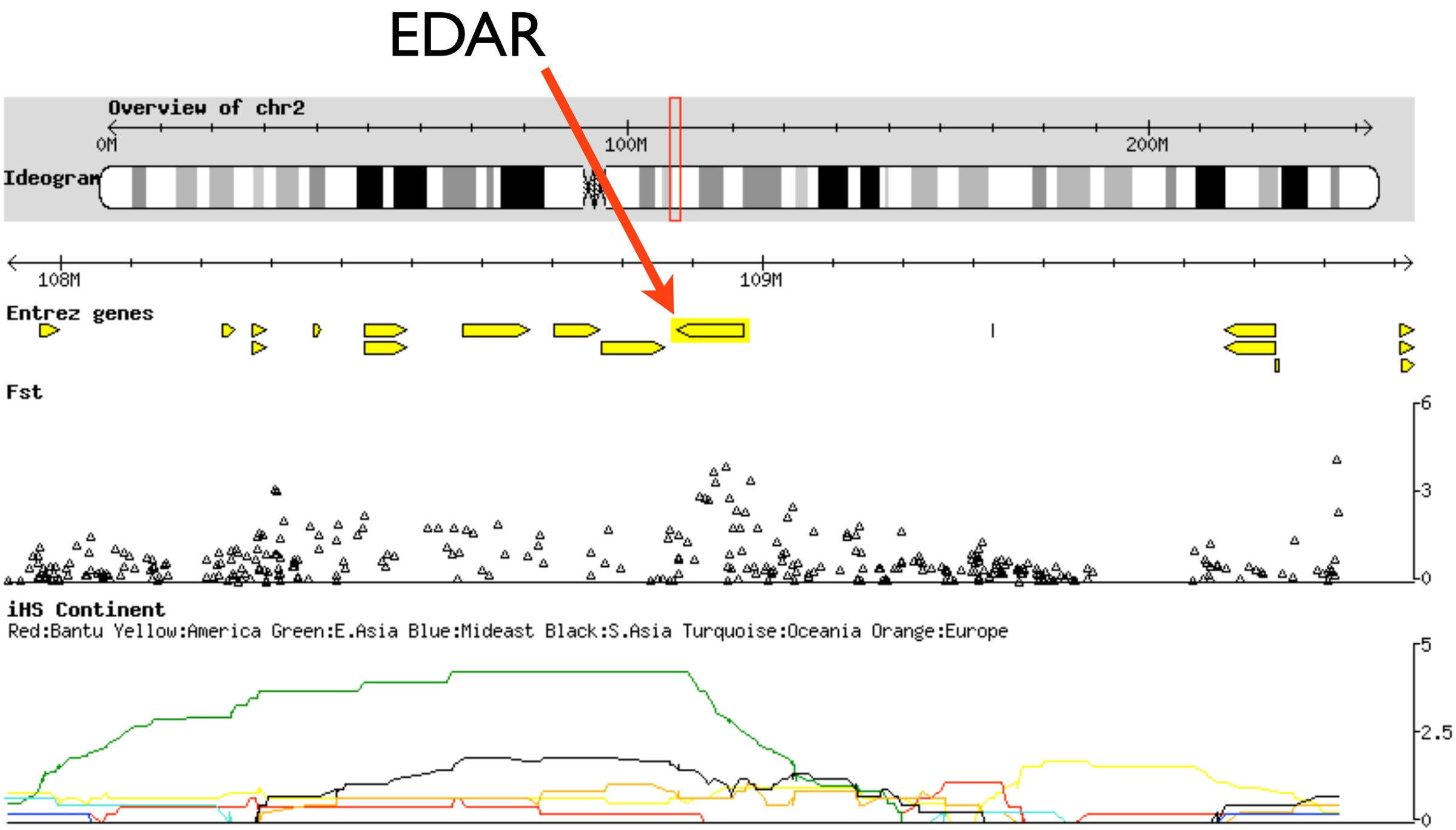
- Send your browser to http://hgdp.uchicago.edu and click blue banner:

- Enter gene into the Landmark or Region box (e.g., EDAR)



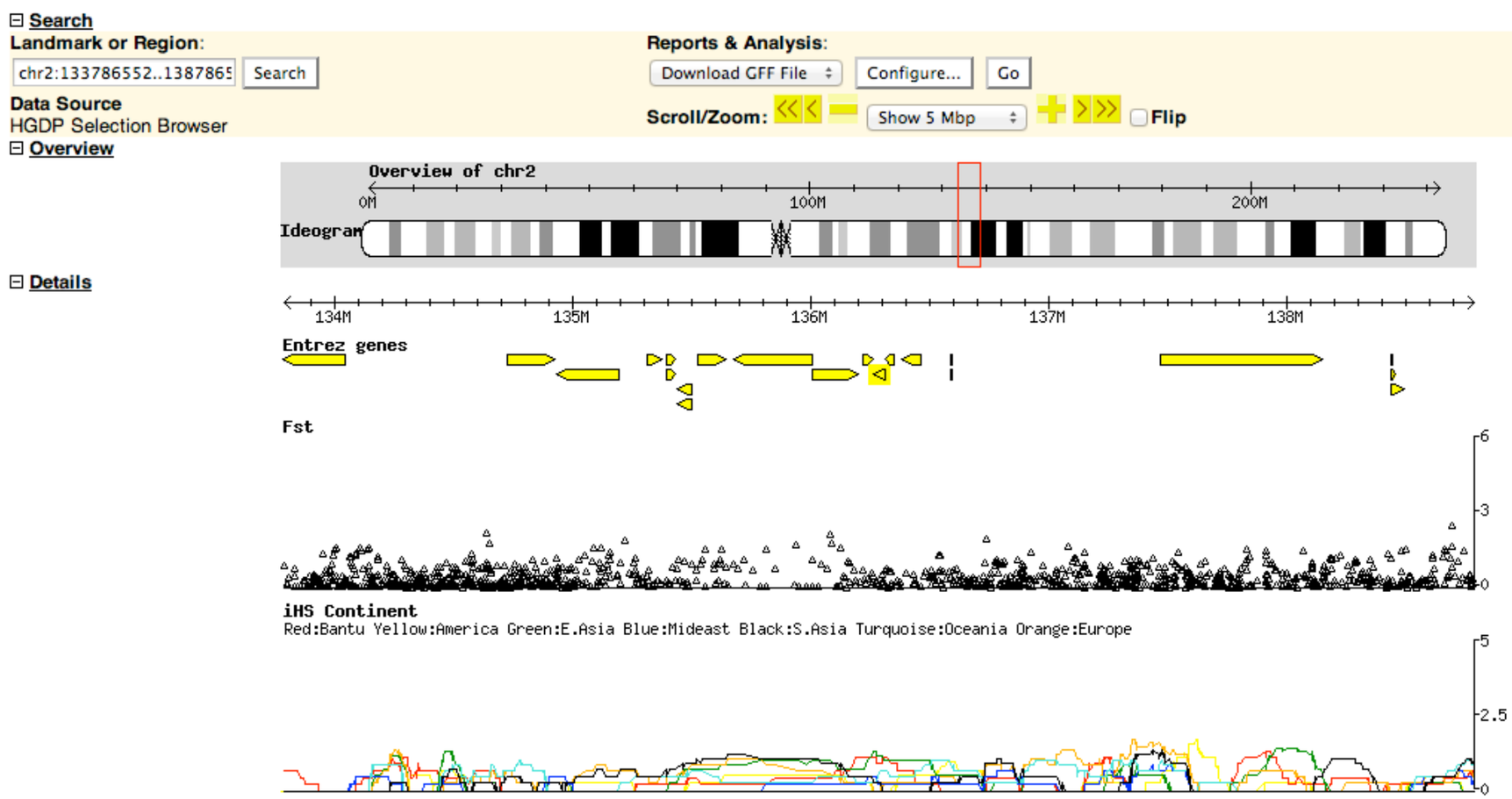
- Then adjust the Zoom (e.g., "Show 2 MB")

HGDP Selection Browser



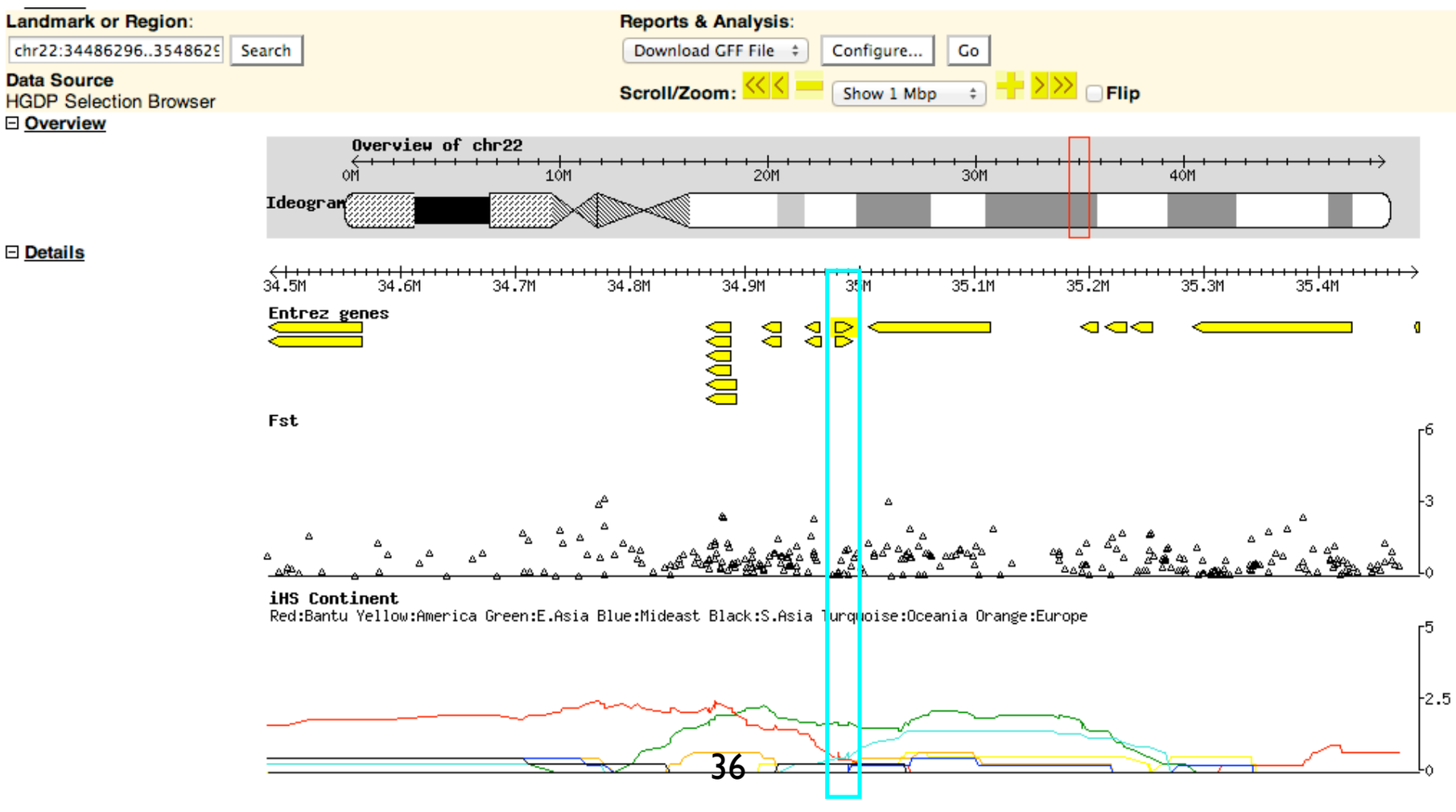
Warnings

- HapMap and HGDP are very different populations.
- Even though the same methods are applied, signatures of natural selection are sometimes not recapitulated.
- Here is LCT in hgdp (with 5MB window)



Warnings

- HapMap and HGDP are very different populations.
- Even though the same methods are applied, signatures of natural selection are sometimes not recapitulated.
- Here is APOL1 (with 1MB window)



Conclusions

- Signatures of natural selection are very much dependent on the population you are studying.
- If you want reliable results for your population of interest, you should calculate these statistics on your own data!
- PLEASE NOTE: seeing significant peaks in iHS or any other method does not necessarily mean there is selection.
- EVERY DISTRIBUTION HAS A TAIL. Unfortunately not everything in the tails are interesting.

Calculating statistics

The Effect of Positive Selection

Adaptive

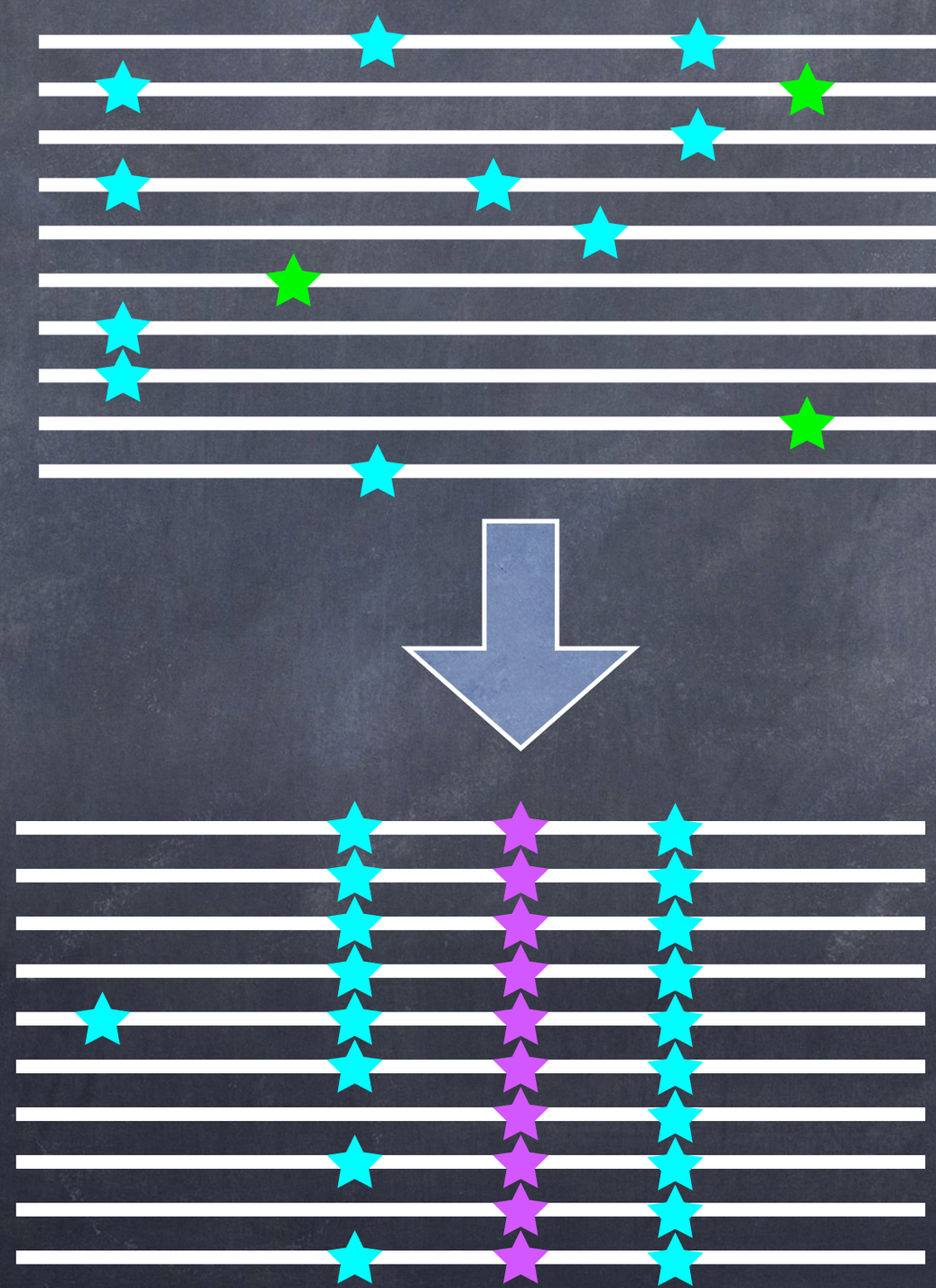
Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

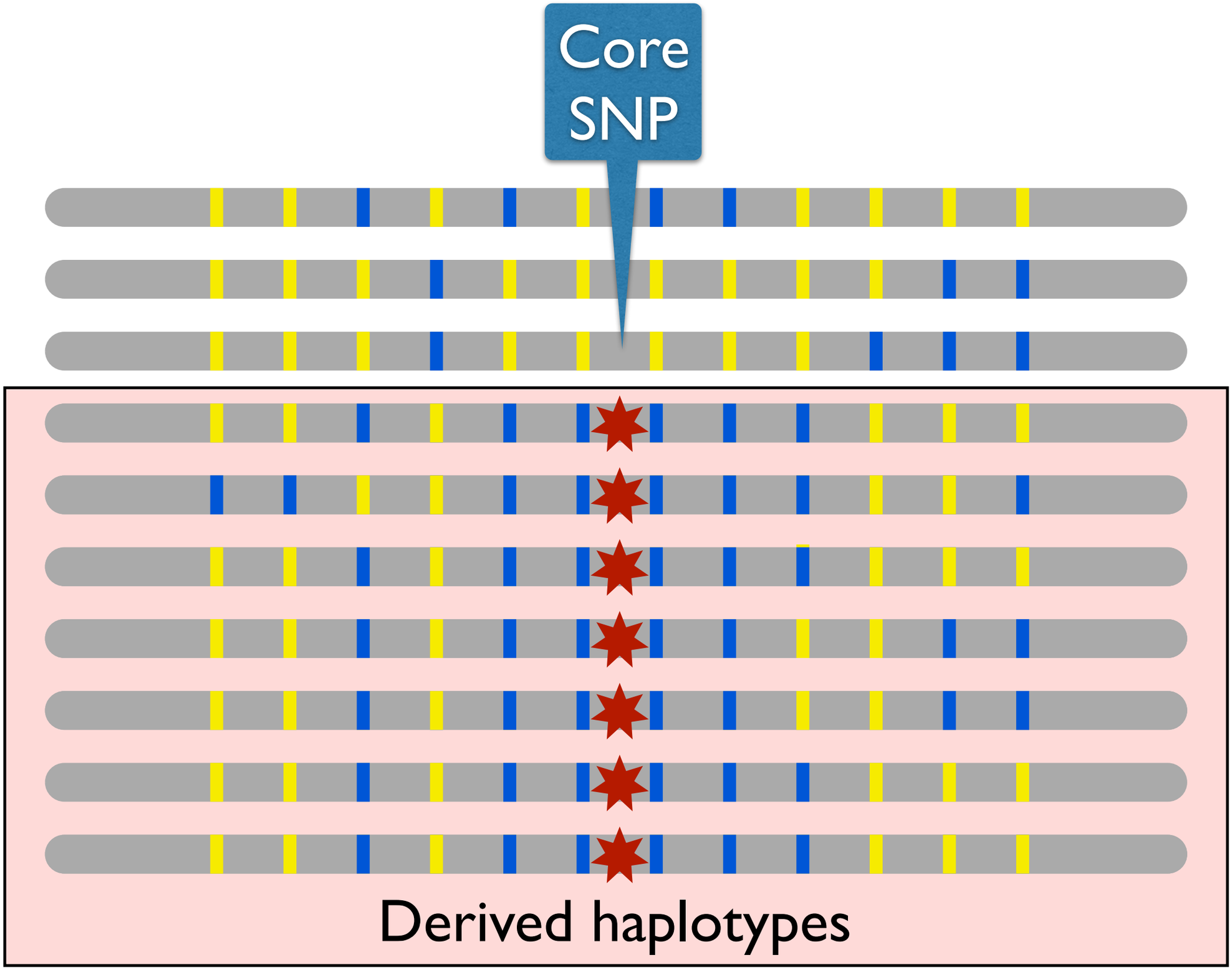
Strongly Deleterious

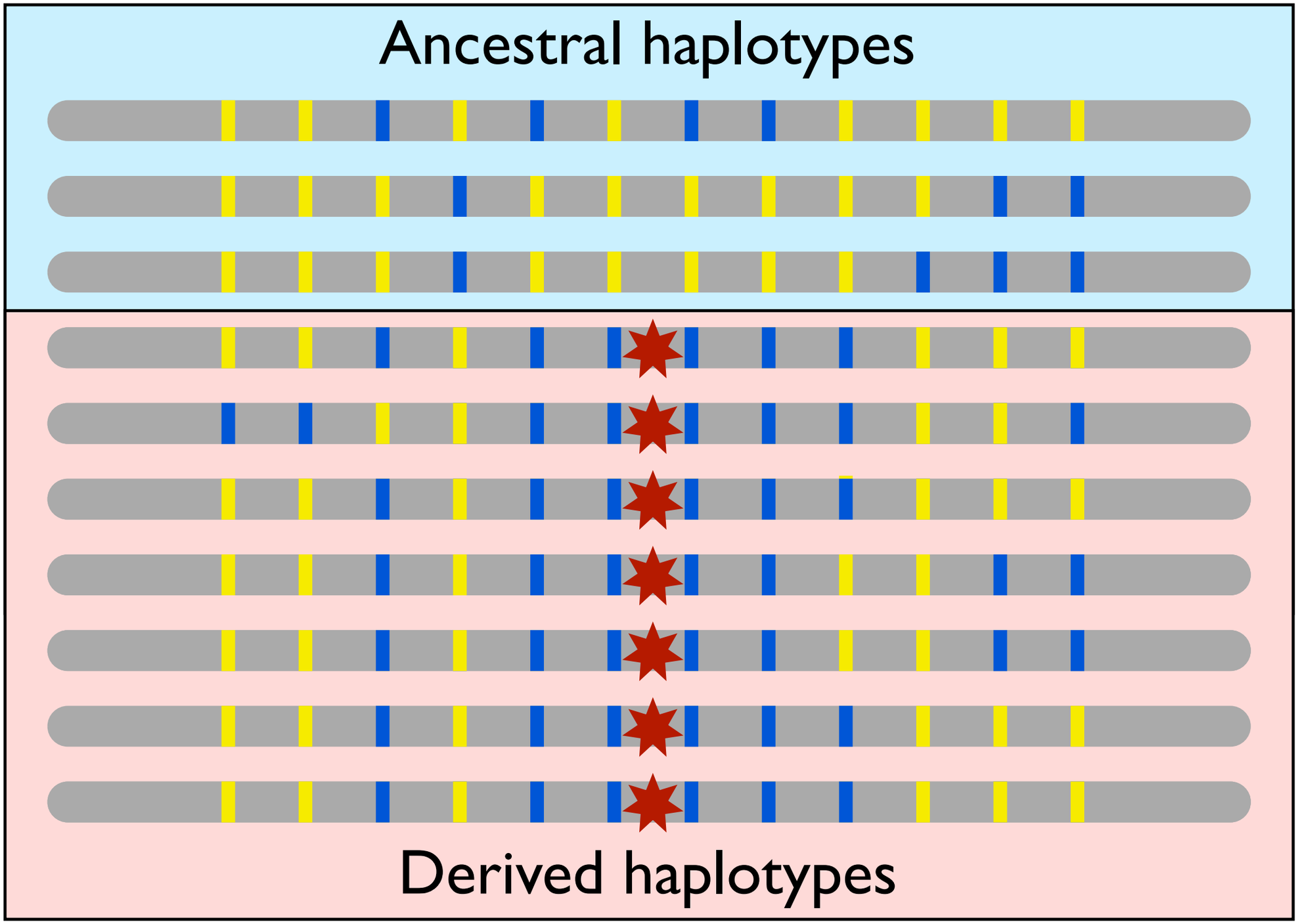


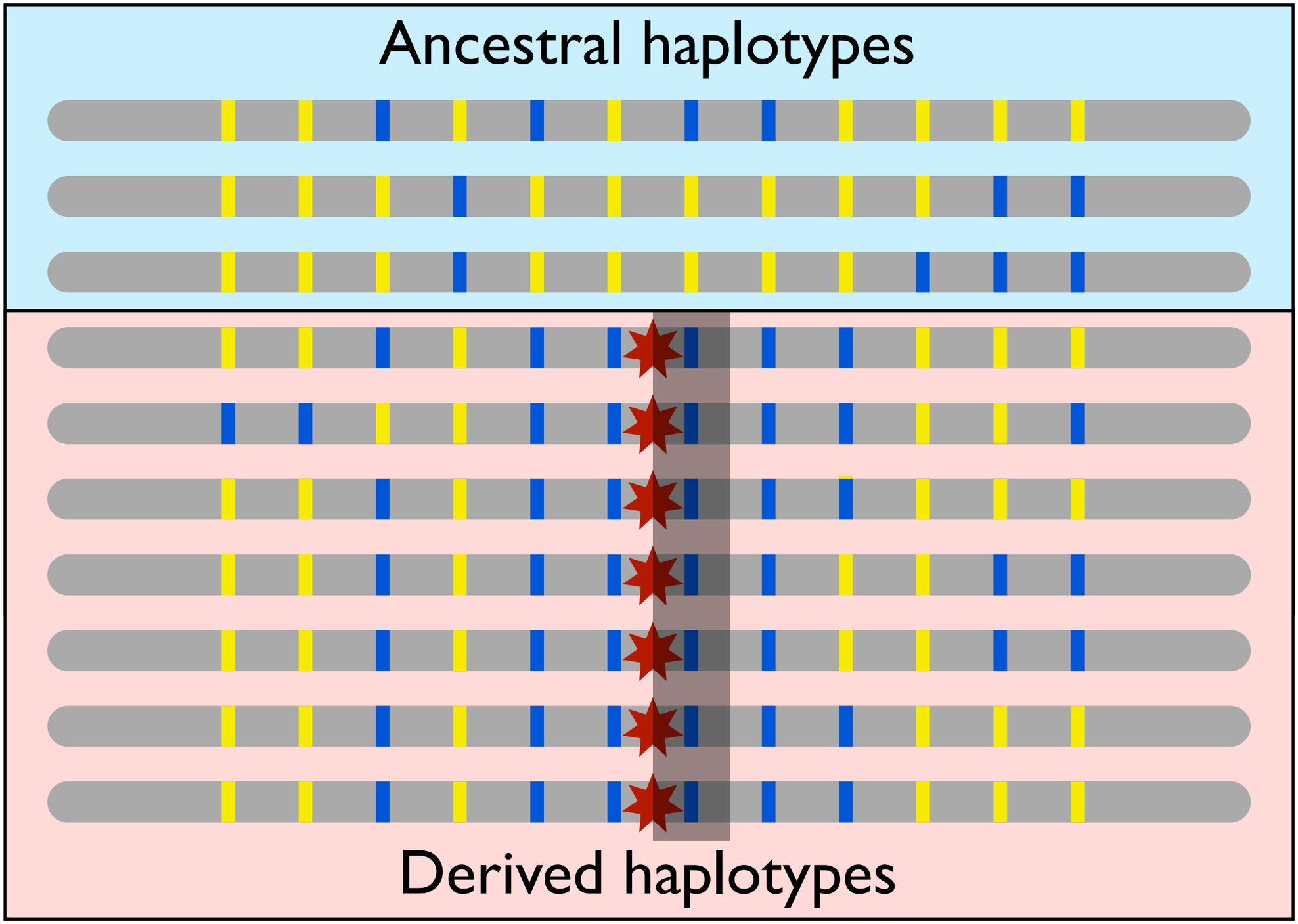
How do we capture this process in a statistic?

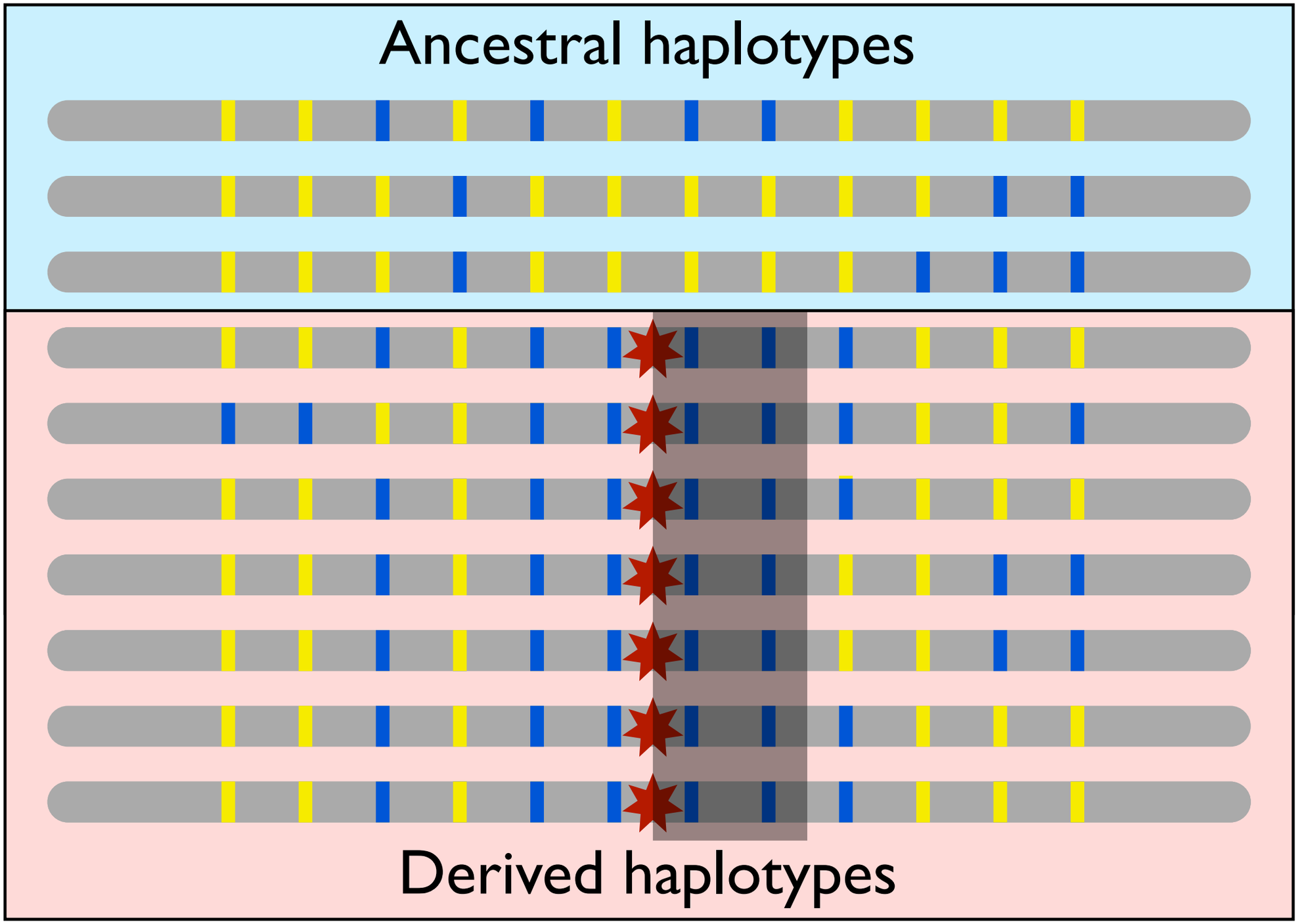
Extended Haplotype Homozygosity

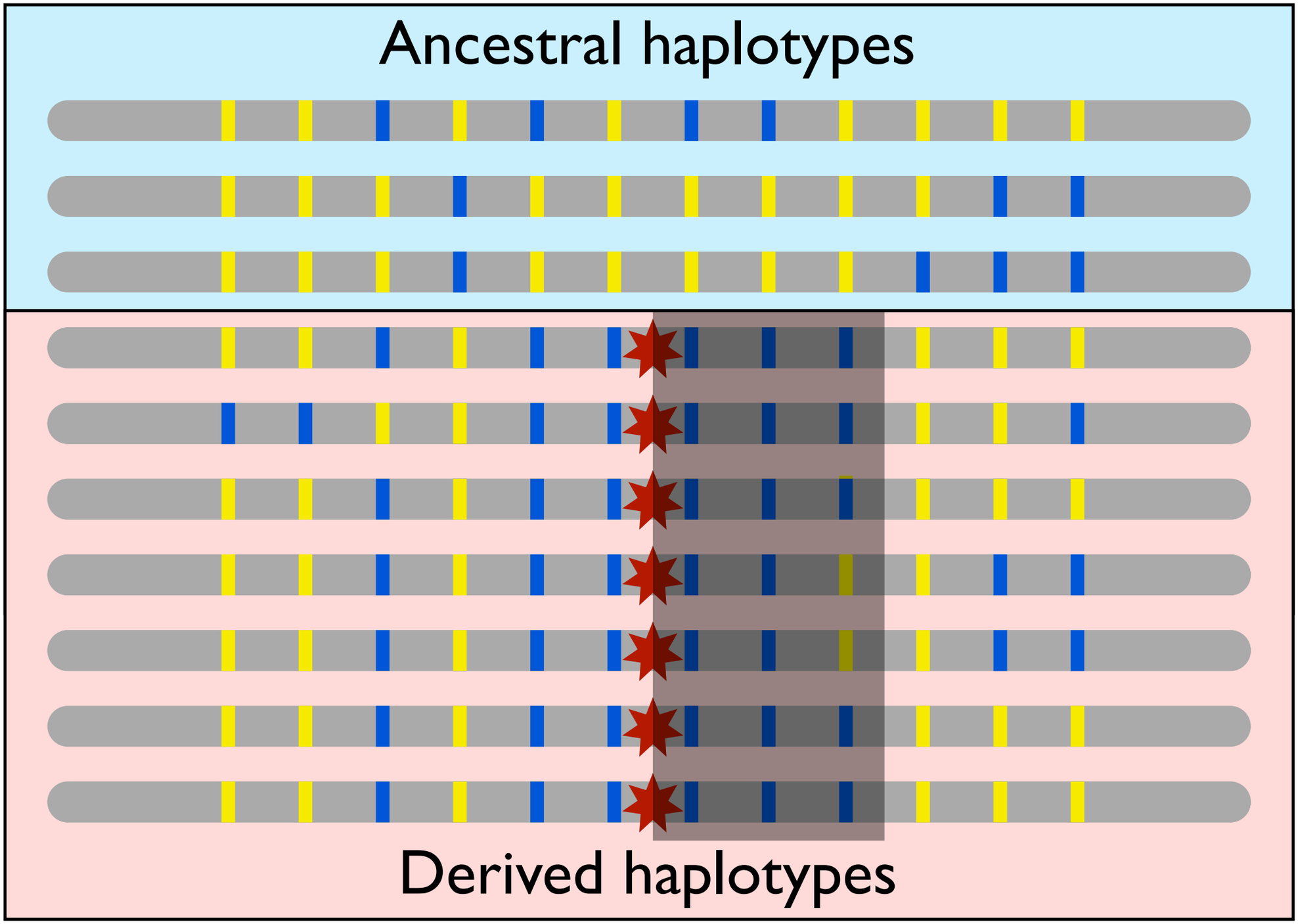
- Sabeti, et al. (*Nature*, 2002) proposed EHH
- Designed to track the decay of haplotype identity away from a locus of interest
  - If selection acts quickly enough
- Originally derives from ideas in Hudson, et al. (*Genetics*, 1994).

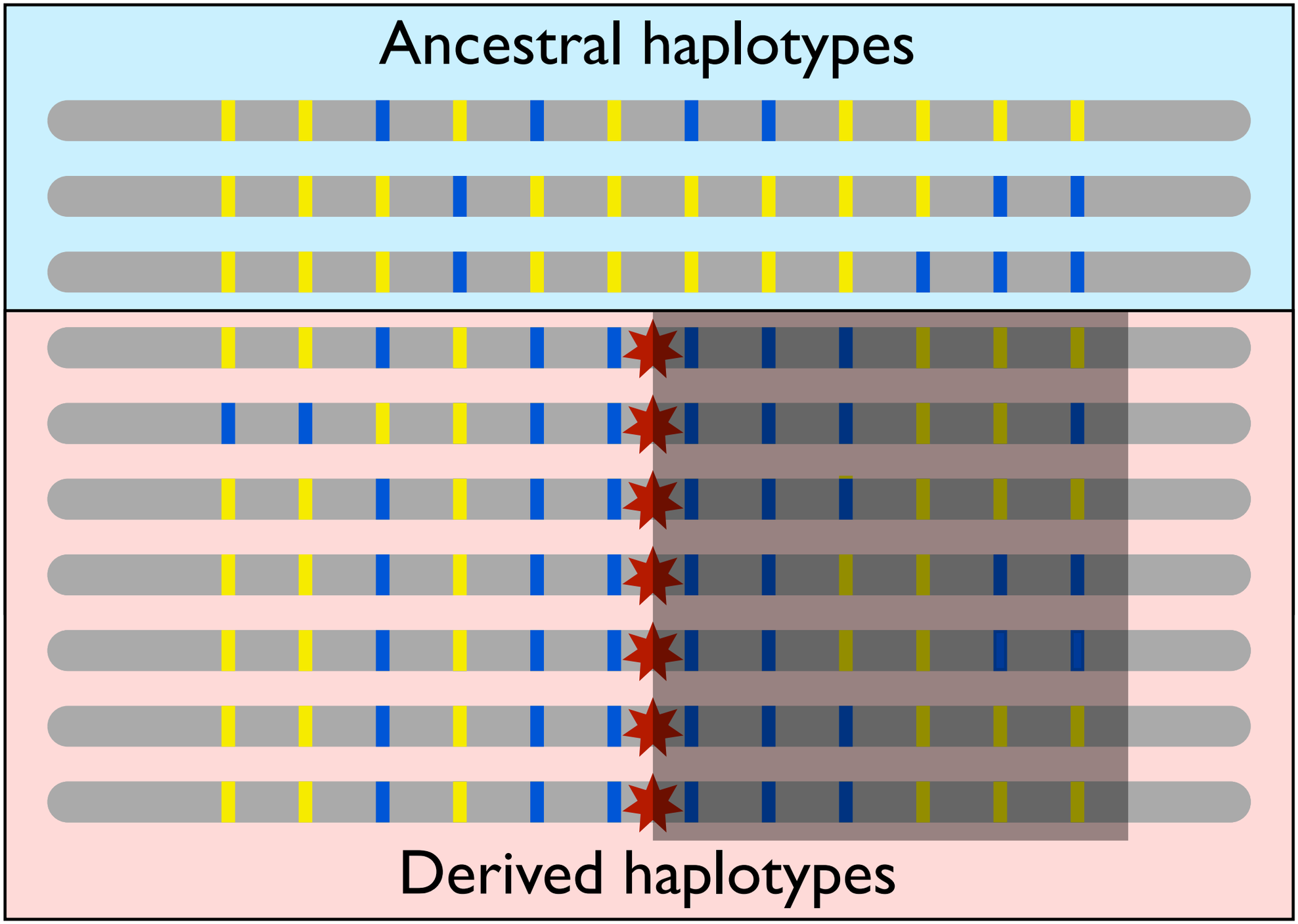

Core SNP
Derived haplotypes


Ancestral haplotypes
Derived haplotypes


Ancestral haplotypes
Derived haplotypes


Ancestral haplotypes
Derived haplotypes


Ancestral haplotypes
Derived haplotypes


Ancestral haplotypes
Derived haplotypes

Calculating EHH

- Given a locus of interest, $\mathcal{C}$ is the set of all distinct haplotypes at that locus.
- Select a "core" haplotype, $c \in \mathcal{C}$.
- $\mathcal{H}(c, x)$ is the set of all distinct haplotypes that extend from the locus of interest to marker `x` and contain the core haplotype `c`.
- For $h \in \mathcal{H}(c, x)$, n_h is the number of haplotypes of type `h`
- n_c is the number of the core haplotypes

Szpiech and Hernandez (2014) *Molecular Biology and Evolution*

Calculating EHH

- If $EHH_c(x)$ is the extended haplotype homozygosity of the core haplotype `c` out to marker `x`, then

$$EHH_c(x) = \sum_{h \in \mathcal{H}(c,x)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$\binom{n}{2} := 0 \;\; \forall n < 2$$

Szpiech and Hernandez (2014) *Molecular Biology and Evolution*

Calculating EHH

- If $EHH_c(x)$ is the extended haplotype homozygosity of the core haplotype `c` out to marker `x`, then

$$EHH_c(x) = \sum_{h \in \mathcal{H}(c,x)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

# of ways to choose two `h` haplotypes

$$\binom{n}{2} := 0 \;\; \forall n < 2$$

Szpiech and Hernandez (2014) *Molecular Biology and Evolution*

Calculating EHH

- If $EHH_c(x)$ is the extended haplotype homozygosity of the core haplotype $\mathtt{c}$ out to marker $\mathtt{x}$, then

$$EHH_c(x) = \sum_{h \in \mathcal{H}(c,x)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

# of ways to choose two $\mathtt{h}$ haplotypes

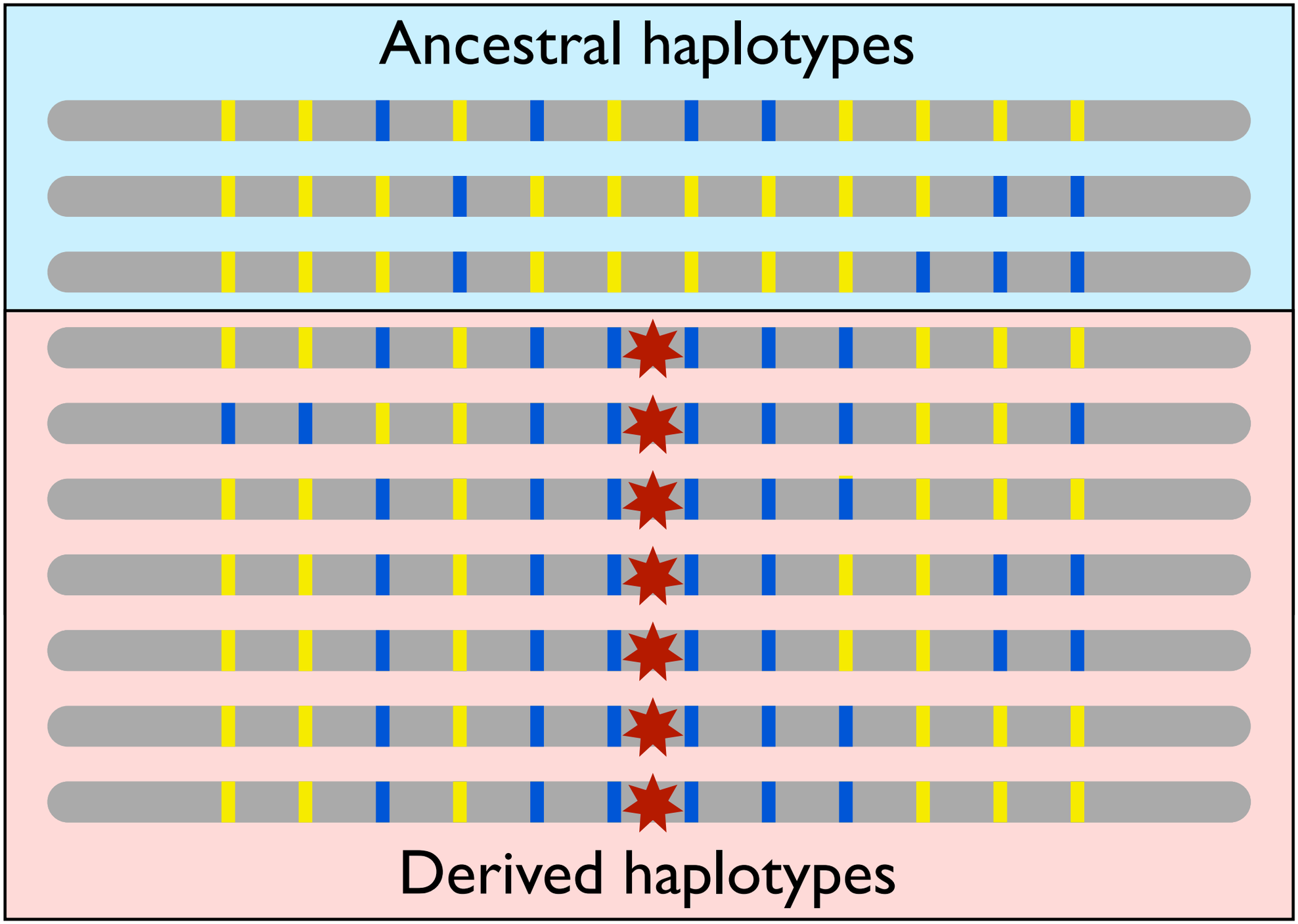
# of ways to choose two core haplotypes

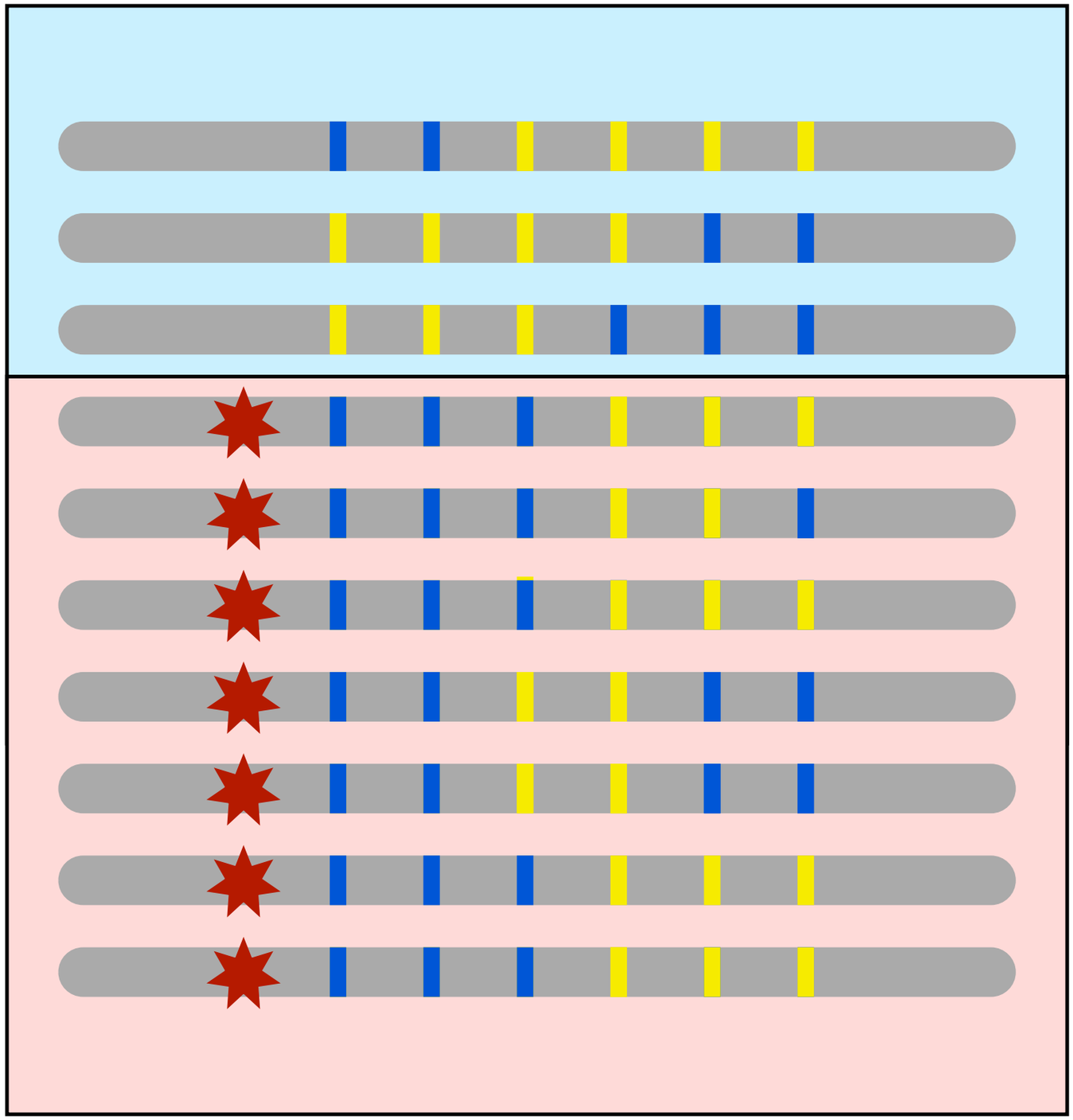
$$\binom{n}{2} := 0 \;\; \forall n < 2$$

Szpiech and Hernandez (2014) *Molecular Biology and Evolution*

Calculating EHH

- Notice that EHH at the core haplotype is necessarily 1 and that it tends to 0 as the number of distinct haplotypes tends to infinity.

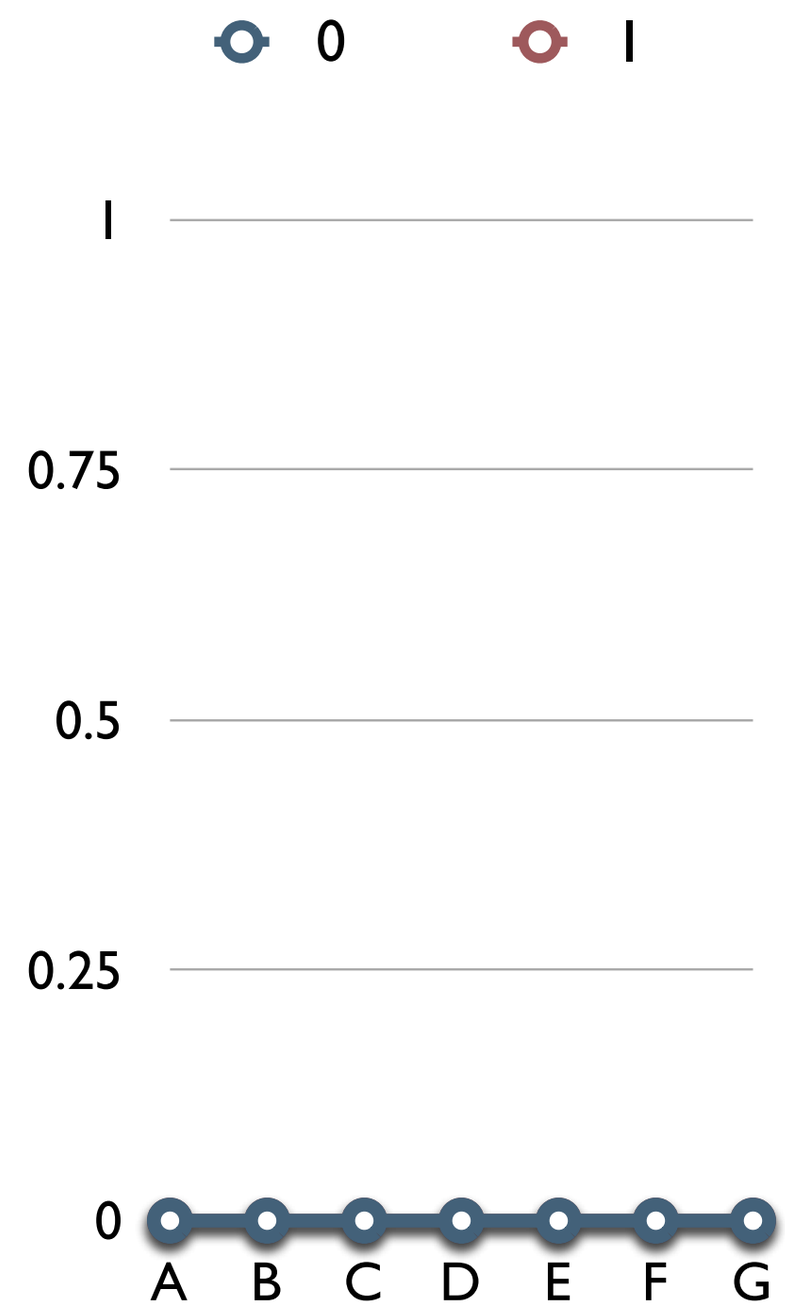

Ancestral haplotypes
Derived haplotypes



0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

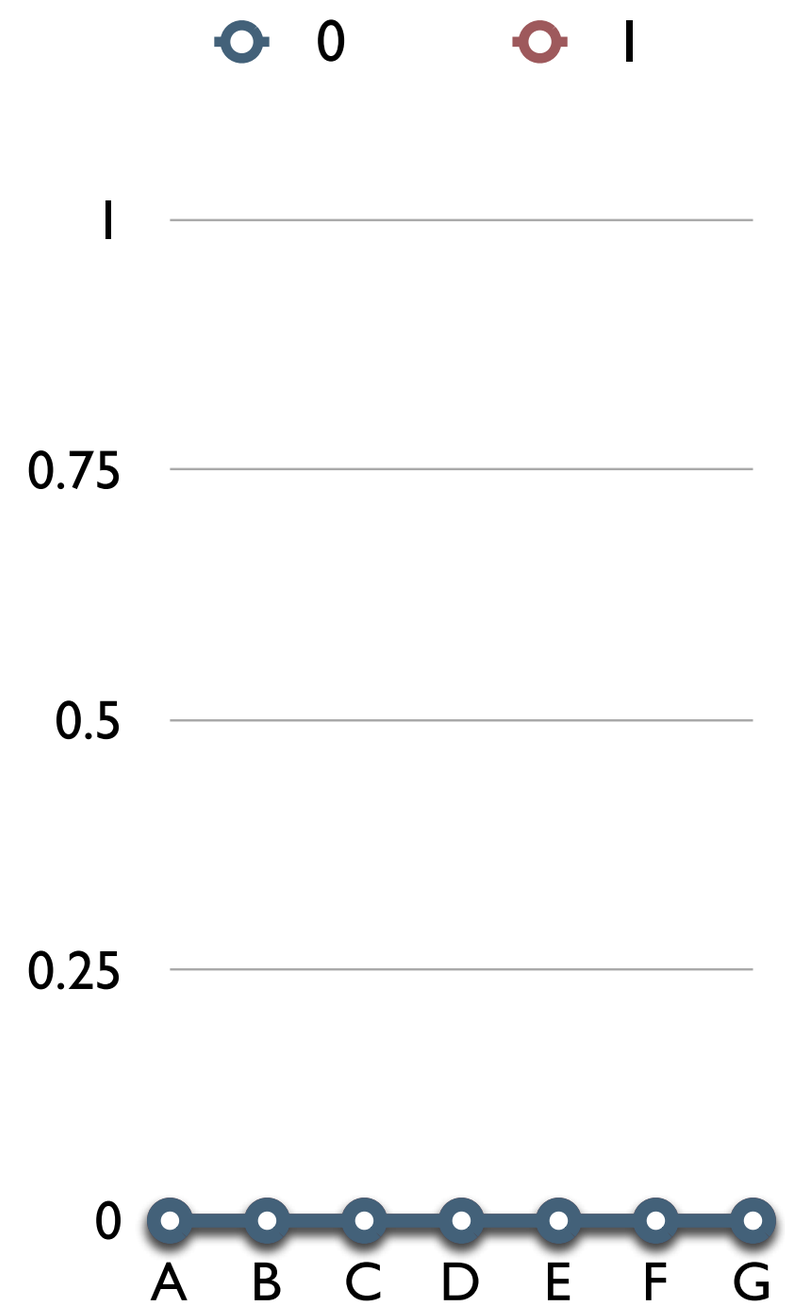
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

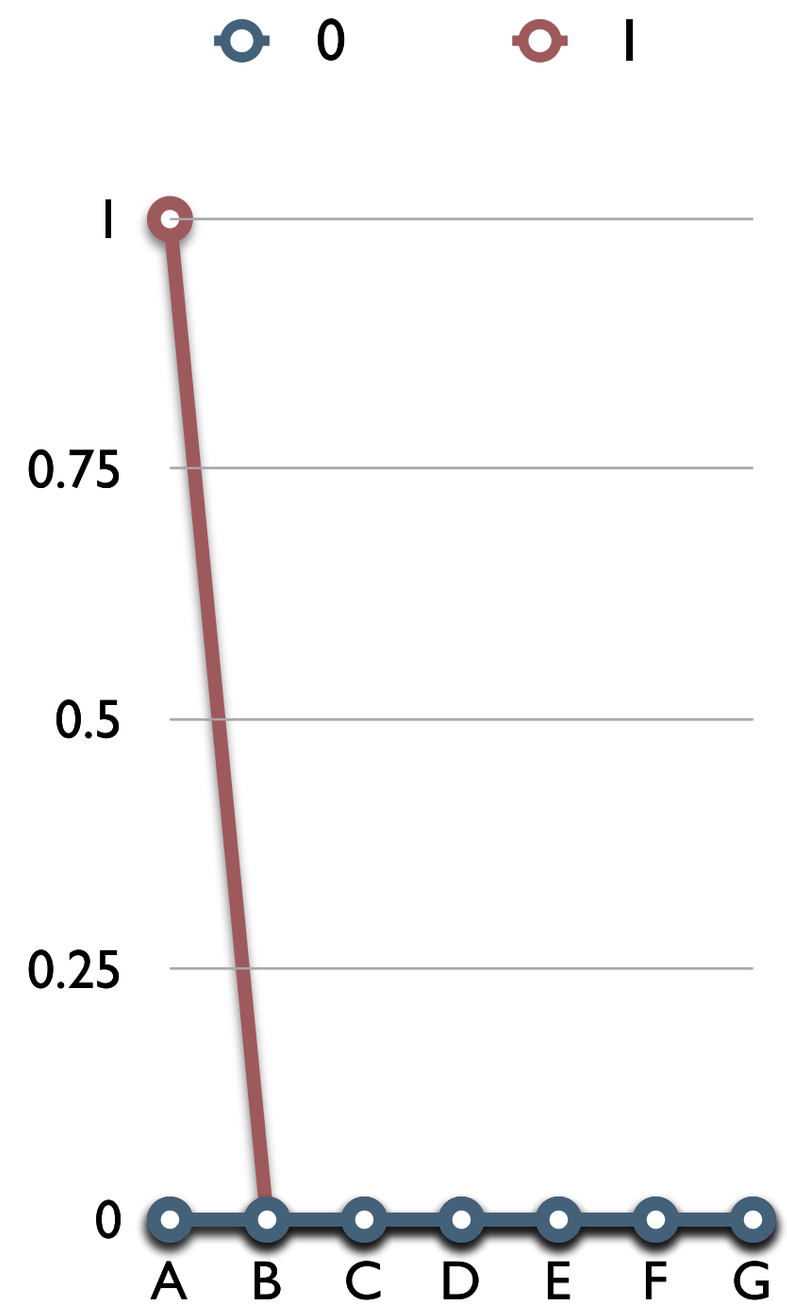
$$n_c = n_1 = 7$$



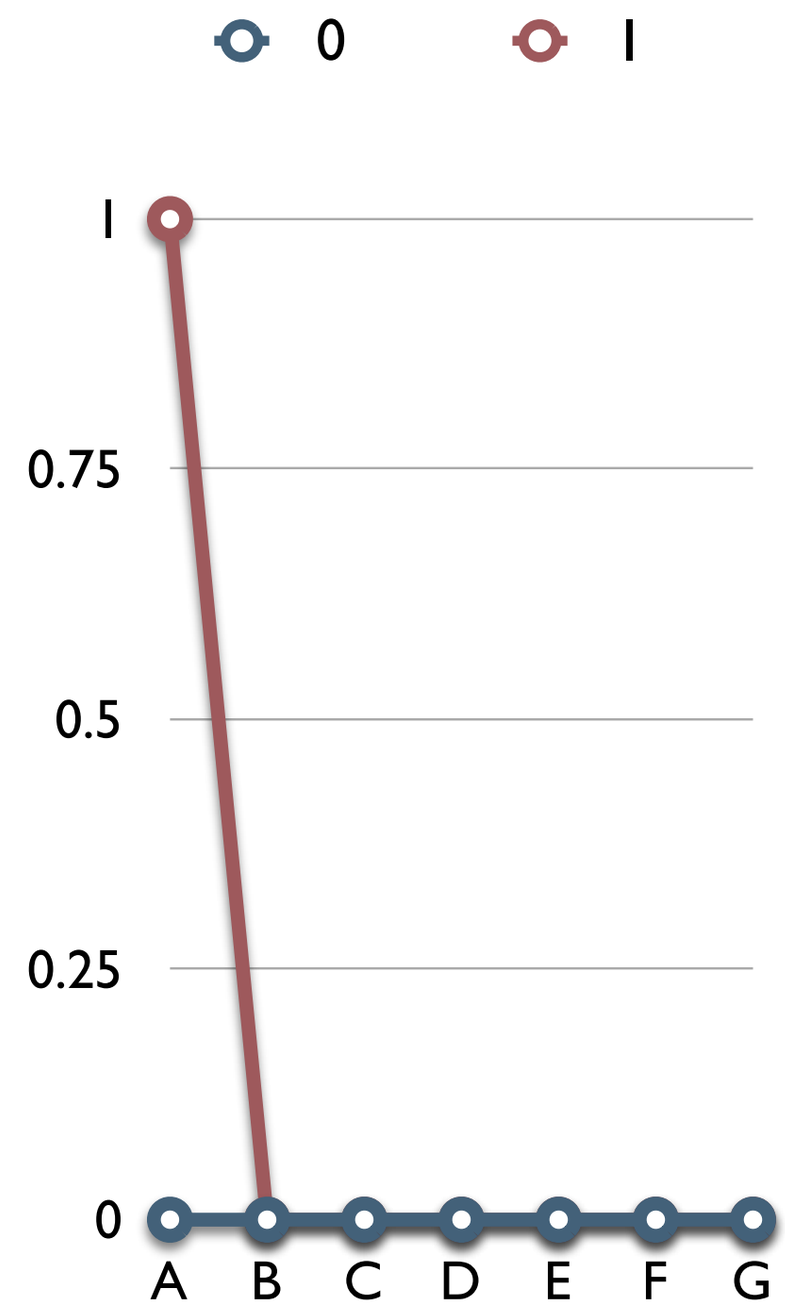
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_c = n_1 = 7$$

$$EHH_1(A) = \frac{\binom{7}{2}}{\binom{7}{2}} = 1$$

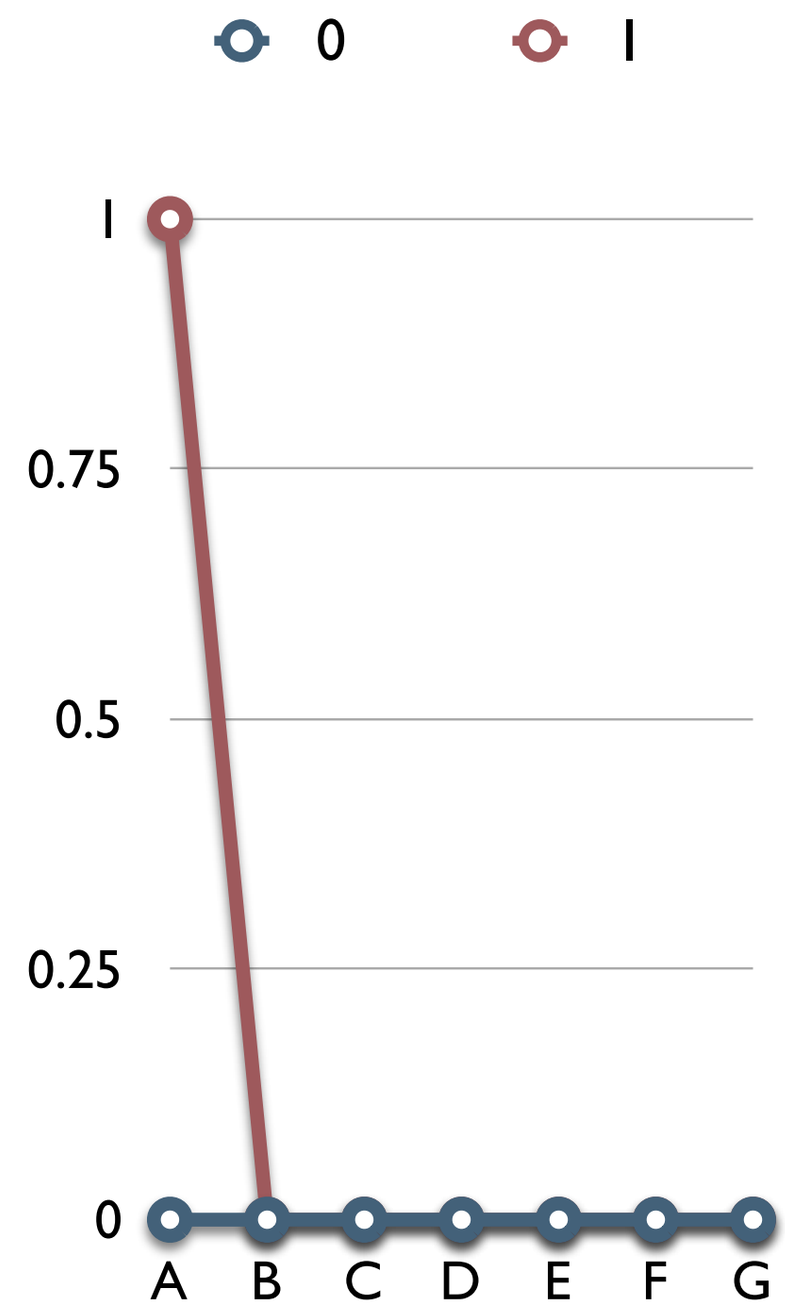


A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

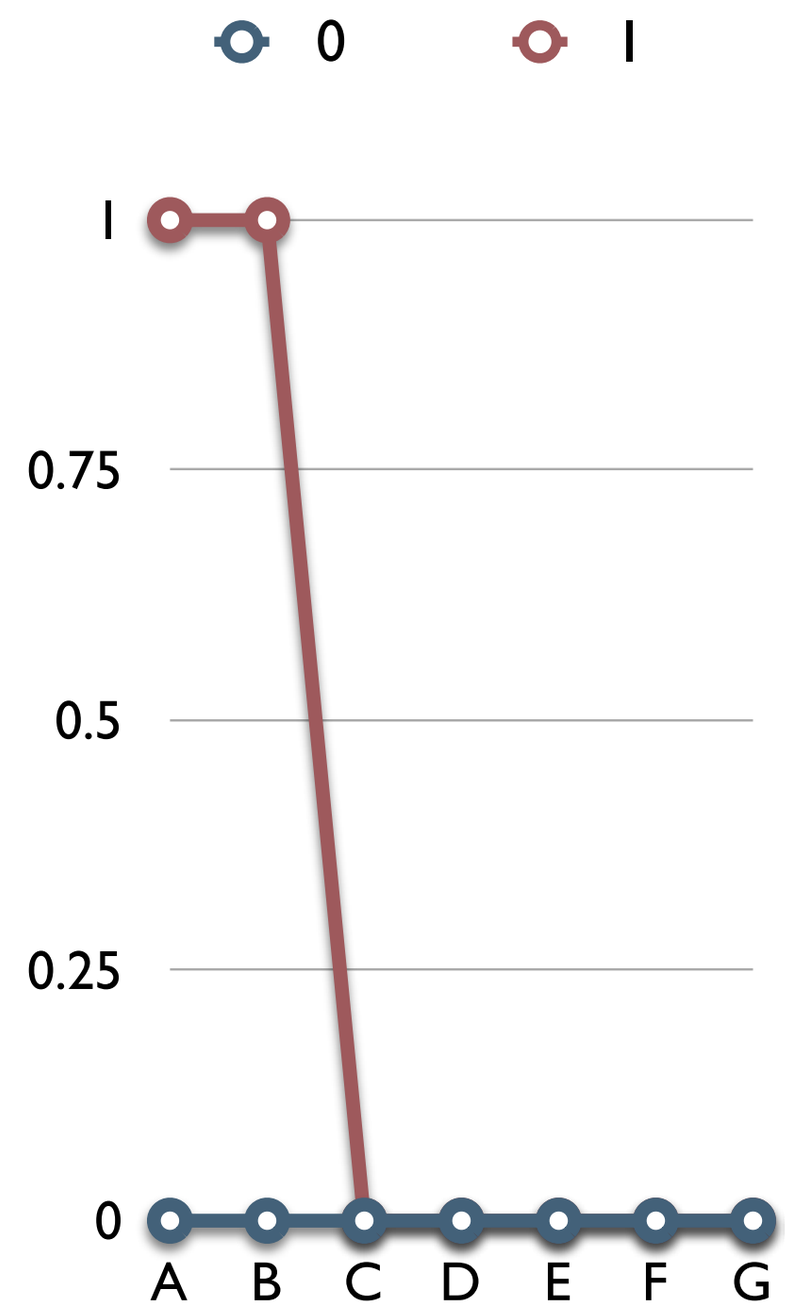
$$n_{11} = 7$$



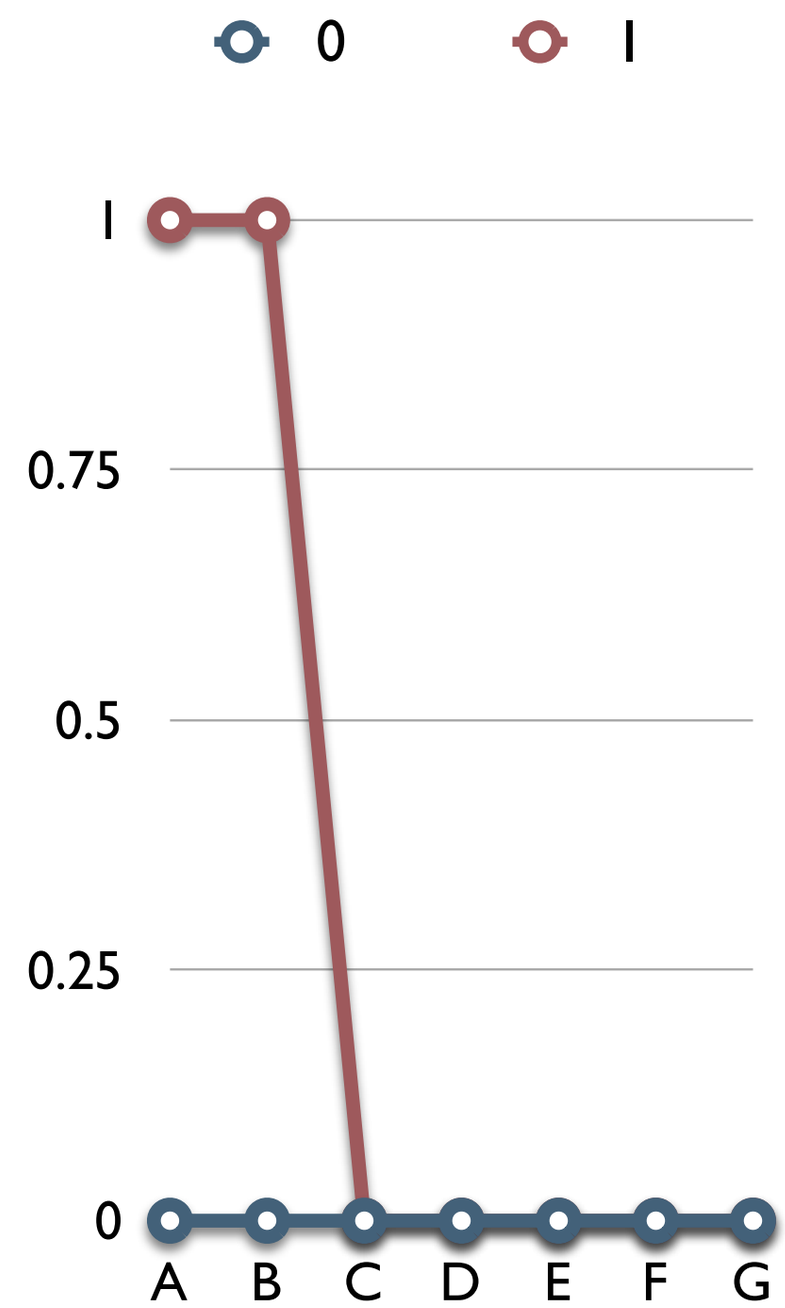
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_{11} = 7$$

$$EHH_1(B) = \frac{\binom{7}{2}}{\binom{7}{2}} = 1$$

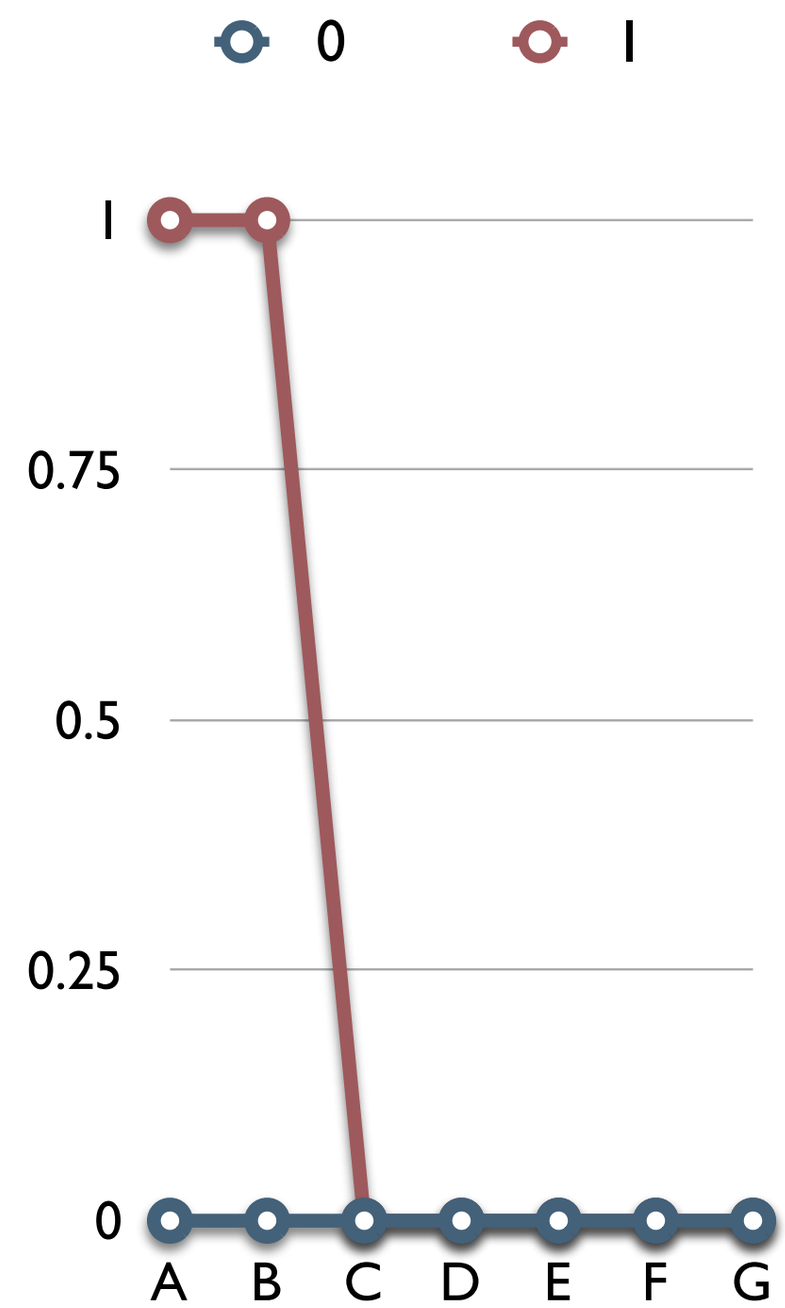


A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

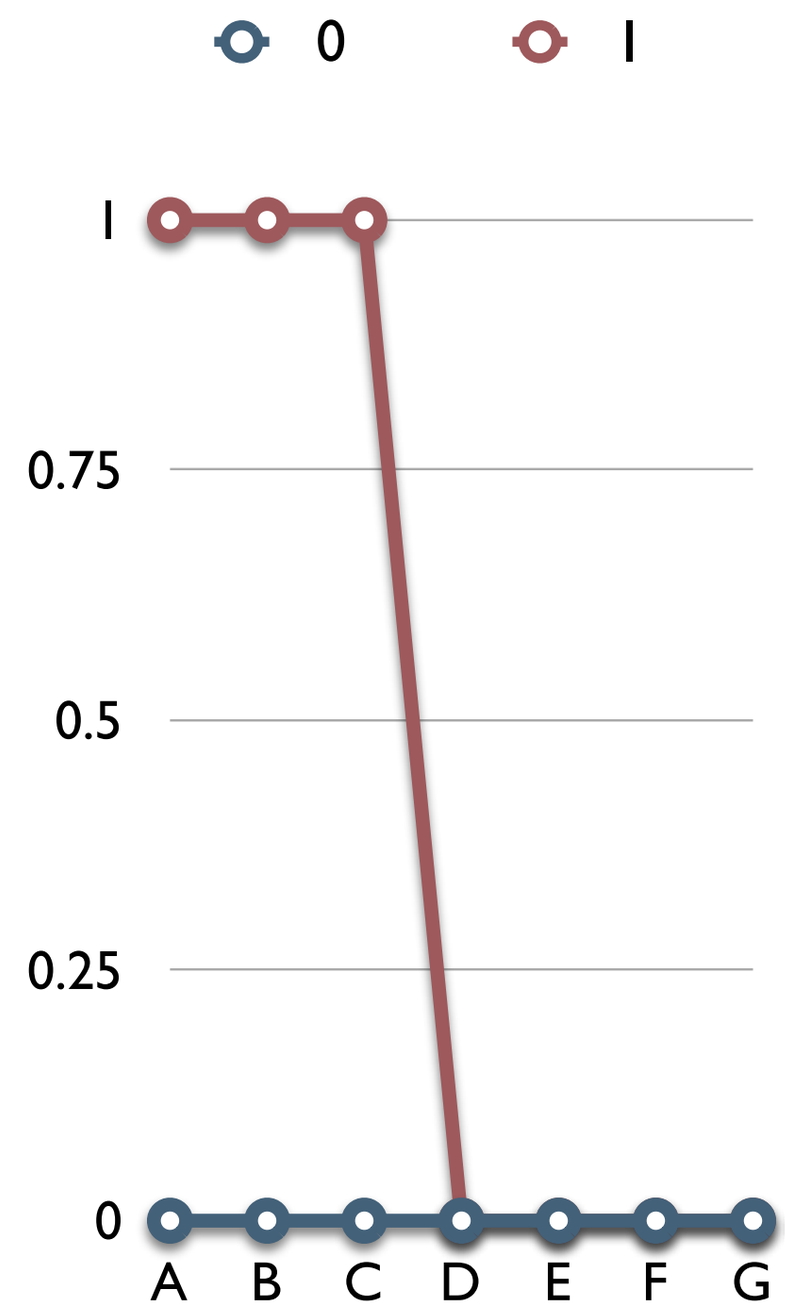
$$n_{111} = 7$$



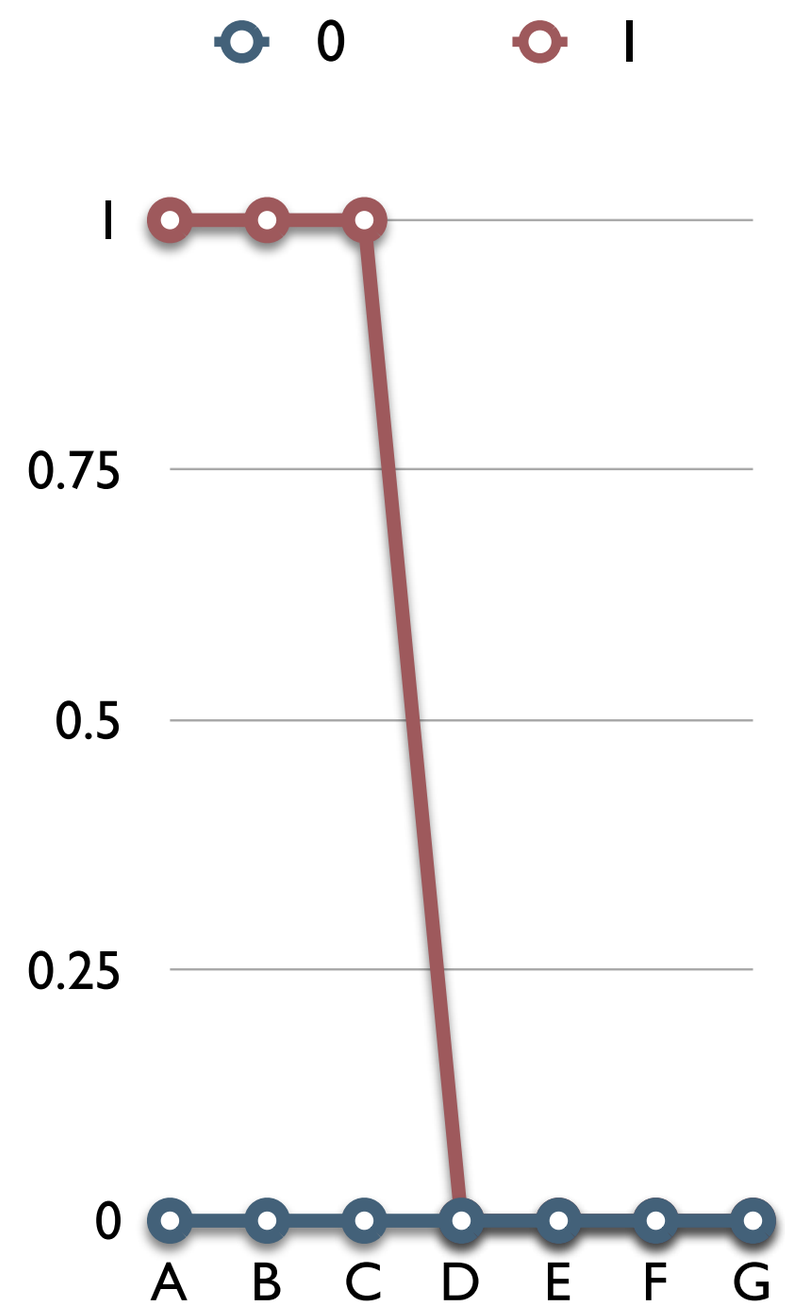
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_{111} = 7$$

$$EHH_1(C) = \frac{\binom{7}{2}}{\binom{7}{2}} = 1$$

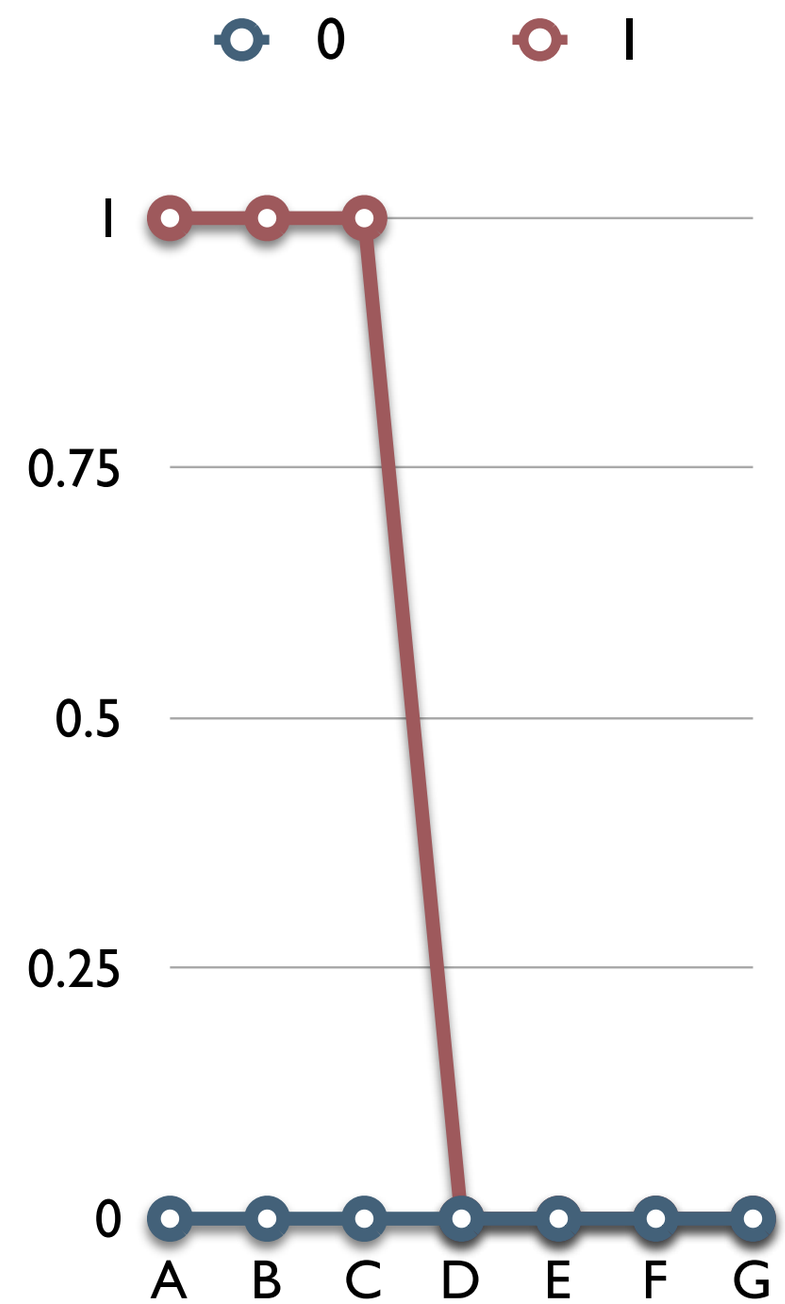


A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0



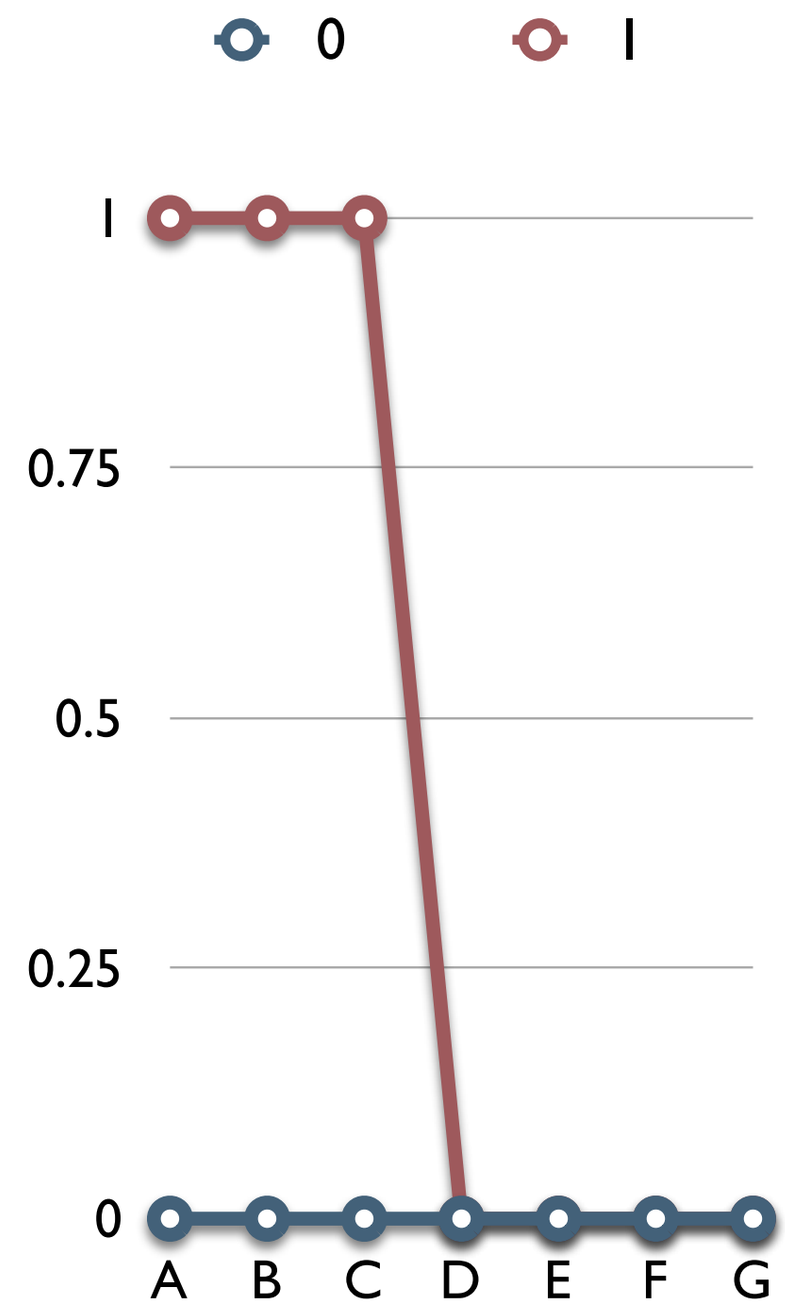
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_{1111} = 5$$



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

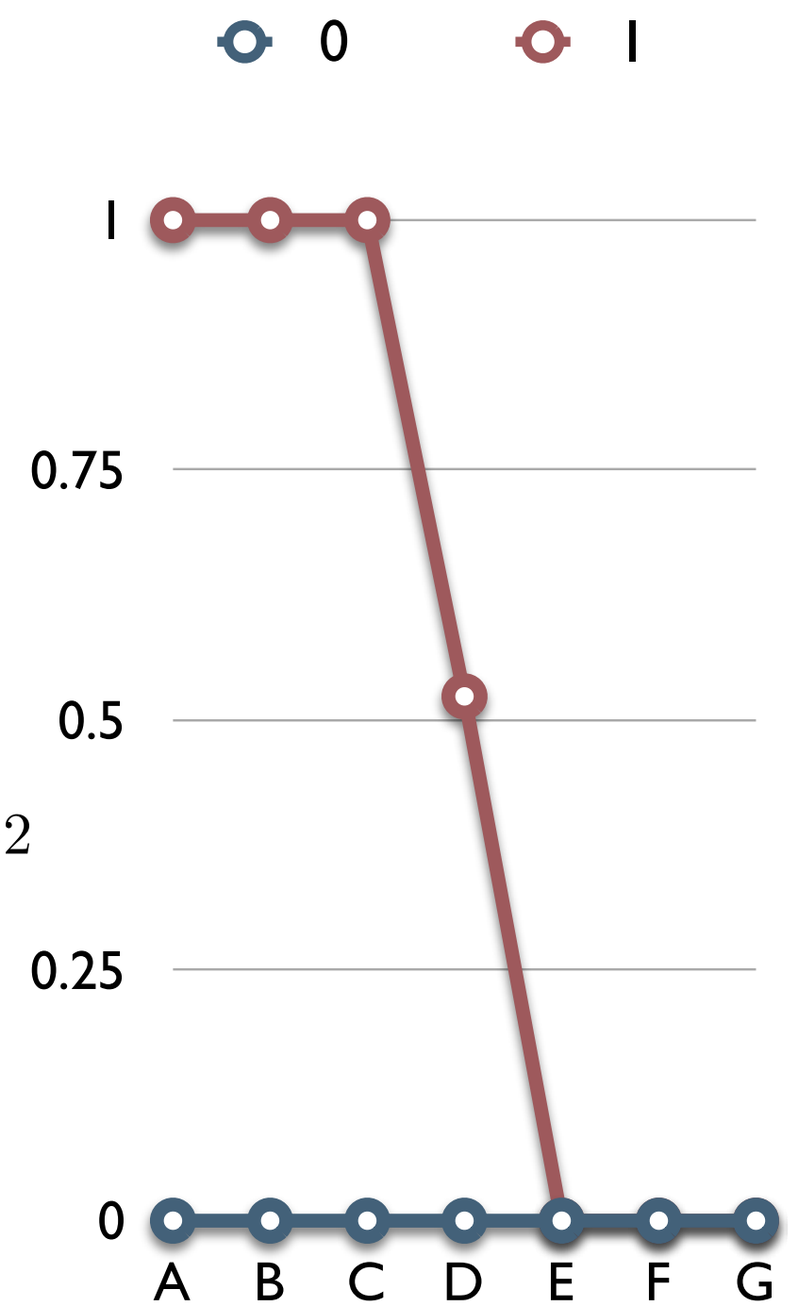
$$n_{1111} = 5$$
$$n_{1110} = 2$$



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

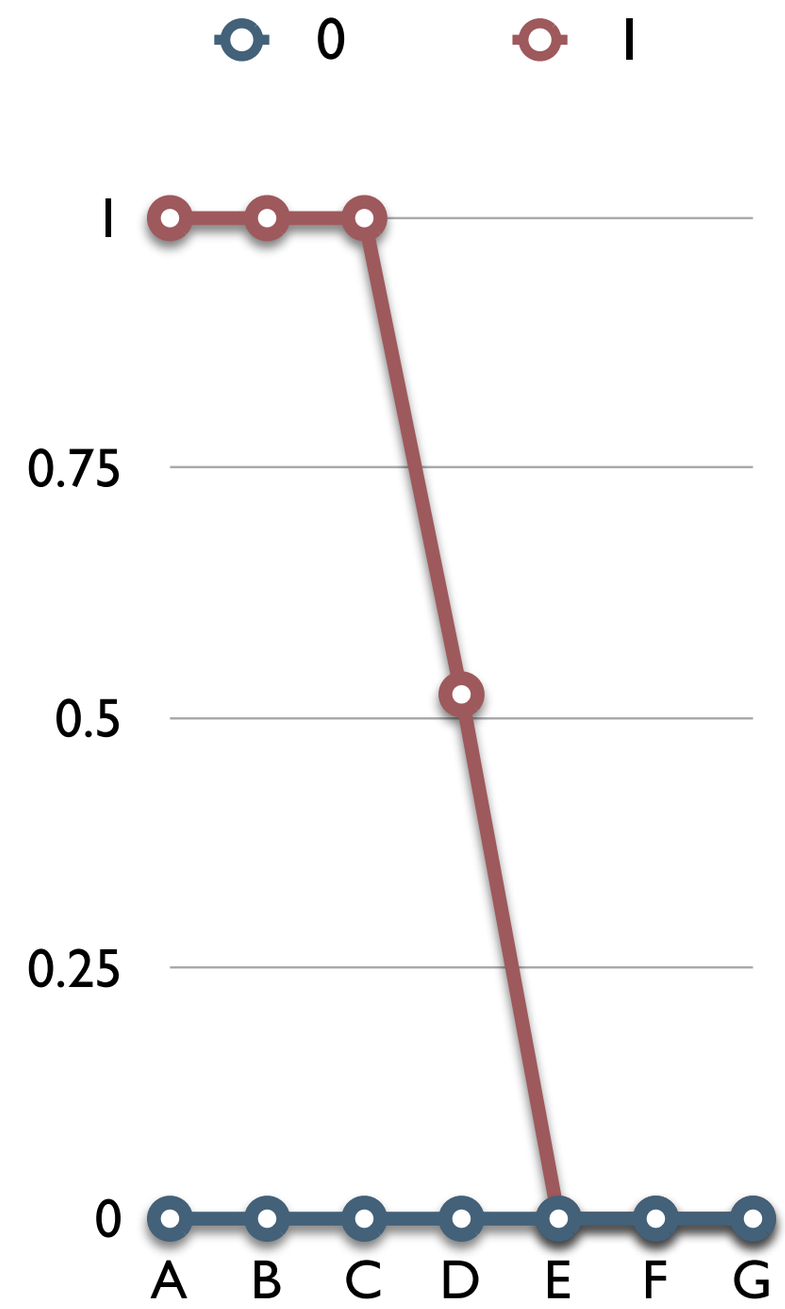
$$n_{1111} = 5$$
$$n_{1110} = 2$$

$$EHH_1(D) = \frac{\binom{5}{2}}{\binom{7}{2}} + \frac{\binom{2}{2}}{\binom{7}{2}} \approx 0.52$$



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

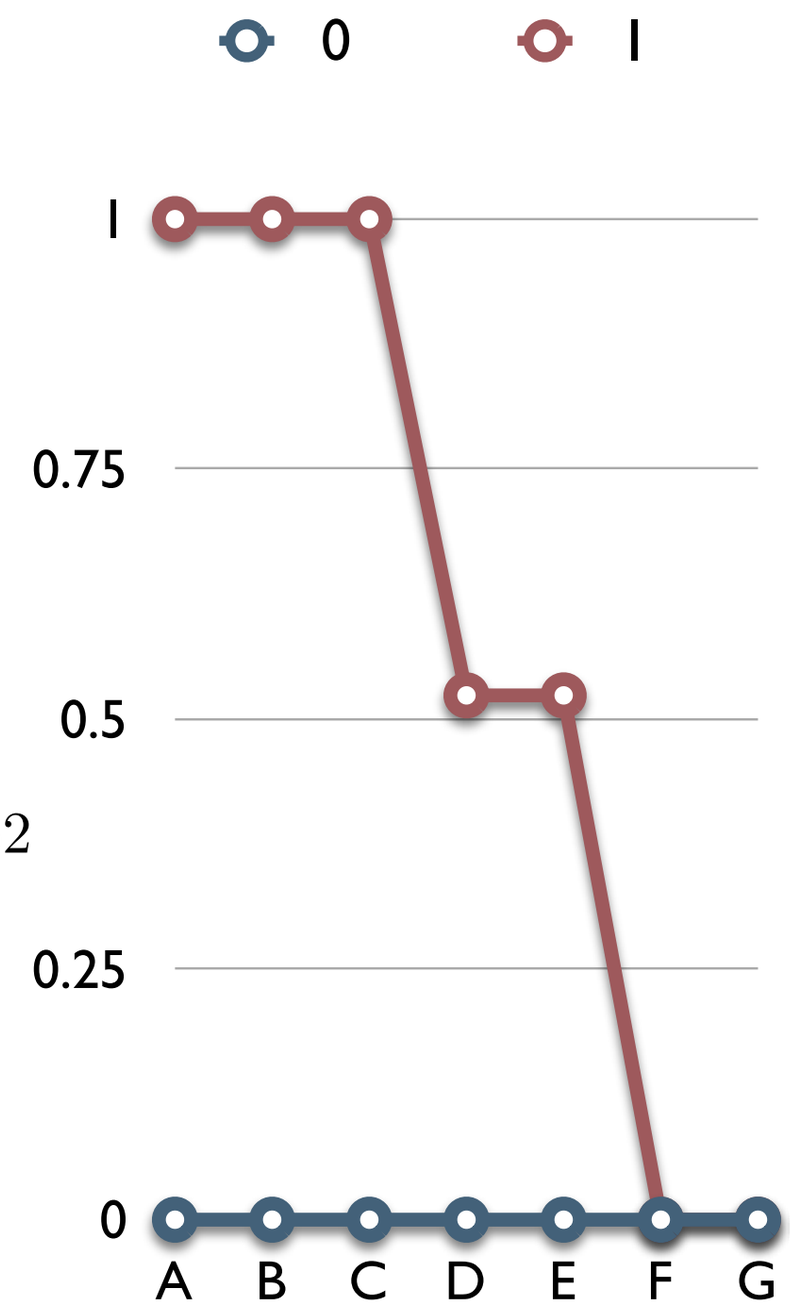
$$n_{11110} = 5$$
$$n_{11100} = 2$$



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

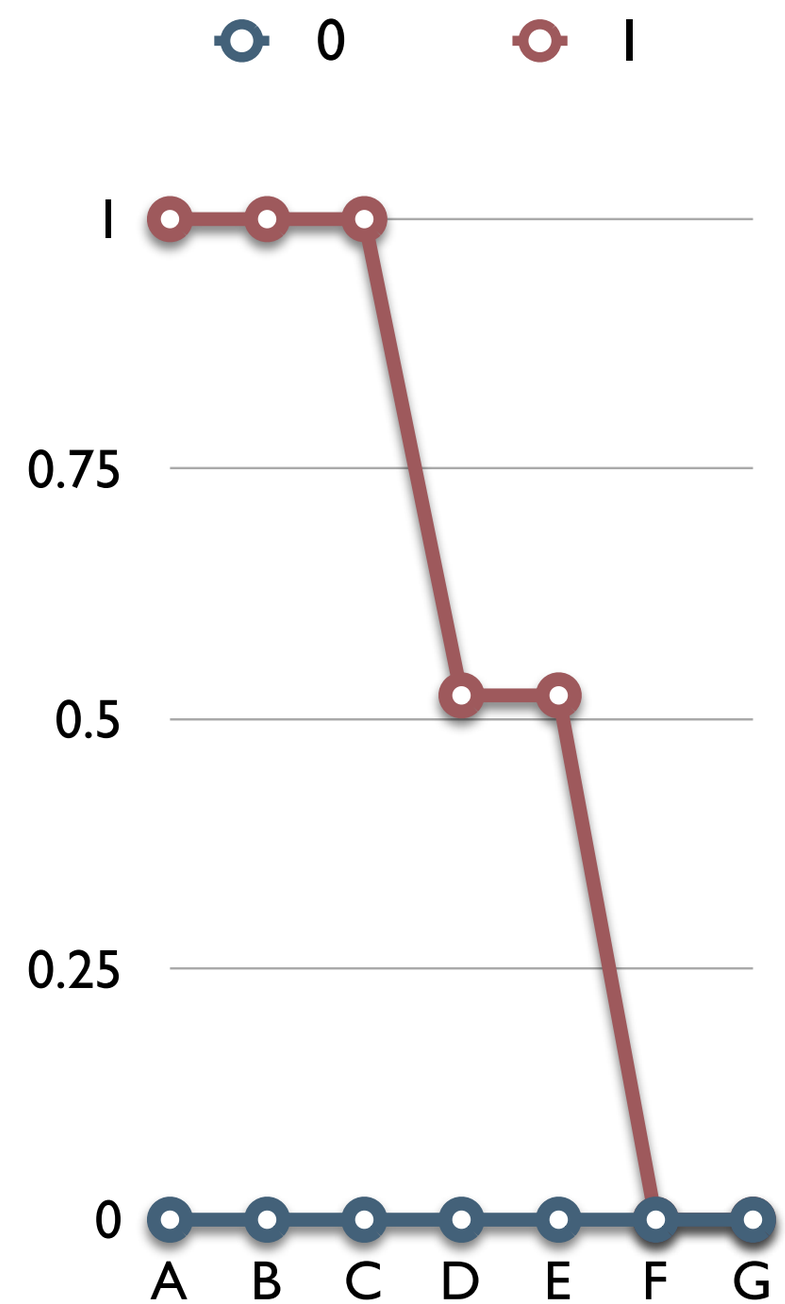
$$n_{11110} = 5$$
$$n_{11100} = 2$$

$$EHH_1(E) = \frac{\binom{5}{2}}{\binom{7}{2}} + \frac{\binom{2}{2}}{\binom{7}{2}} \approx 0.52$$



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

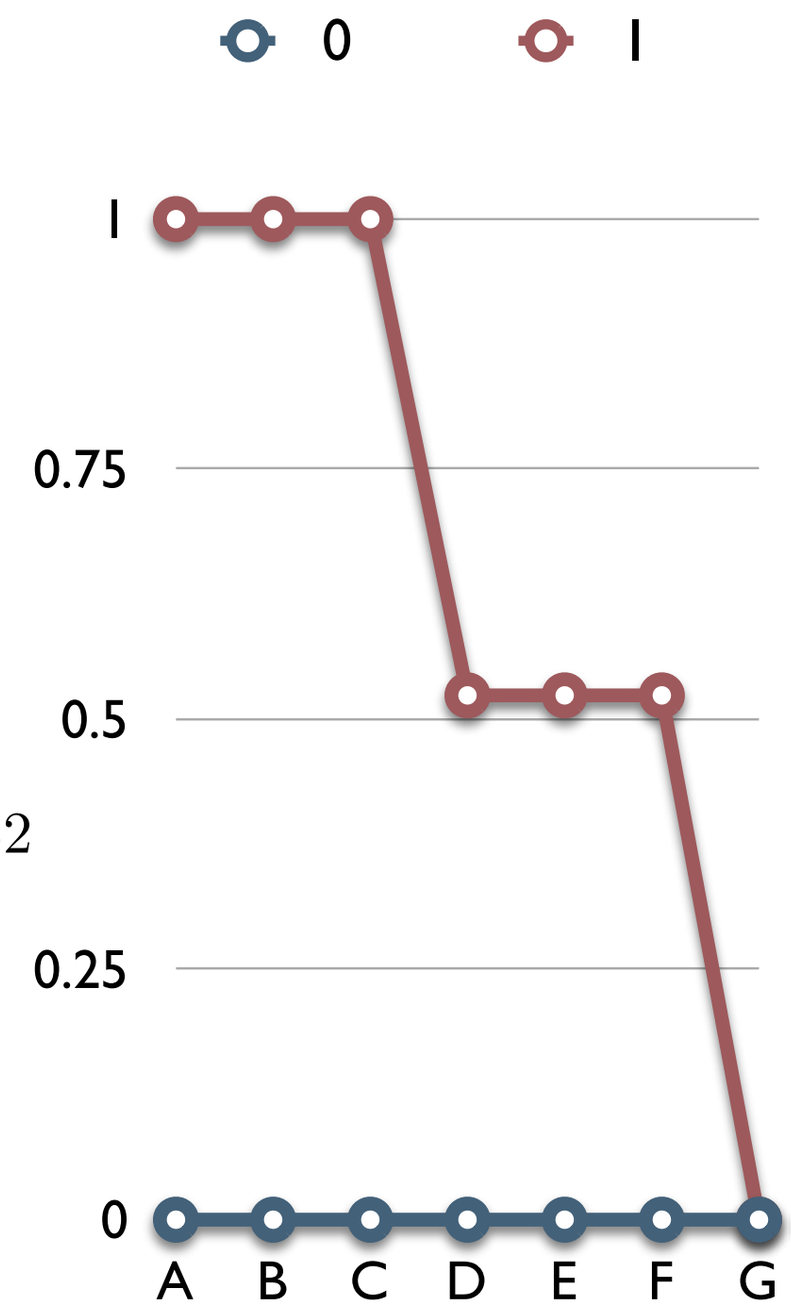
$$n_{111100} = 5$$
$$n_{111001} = 2$$



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

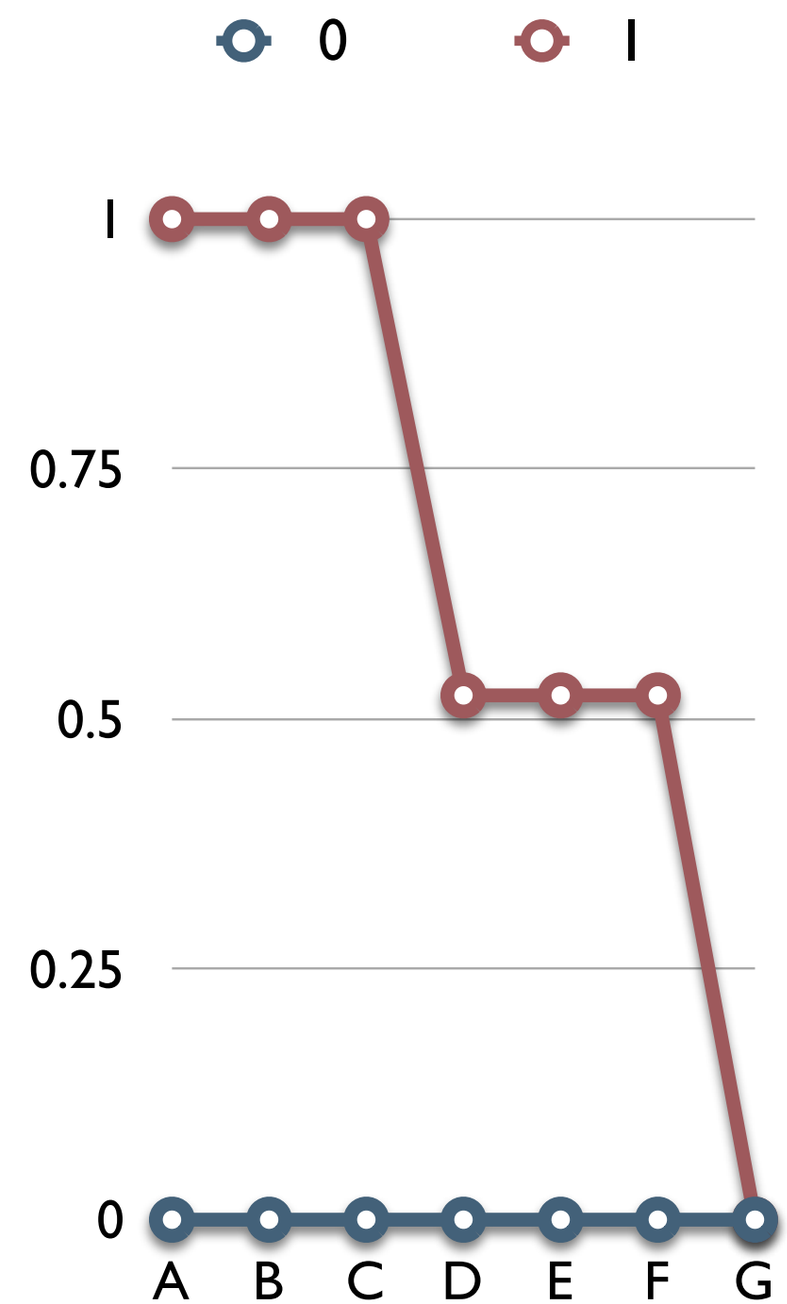
$$n_{111100} = 5$$
$$n_{111001} = 2$$

$$EHH_1(F) = \frac{\binom{5}{2}}{\binom{7}{2}} + \frac{\binom{2}{2}}{\binom{7}{2}} \approx 0.52$$



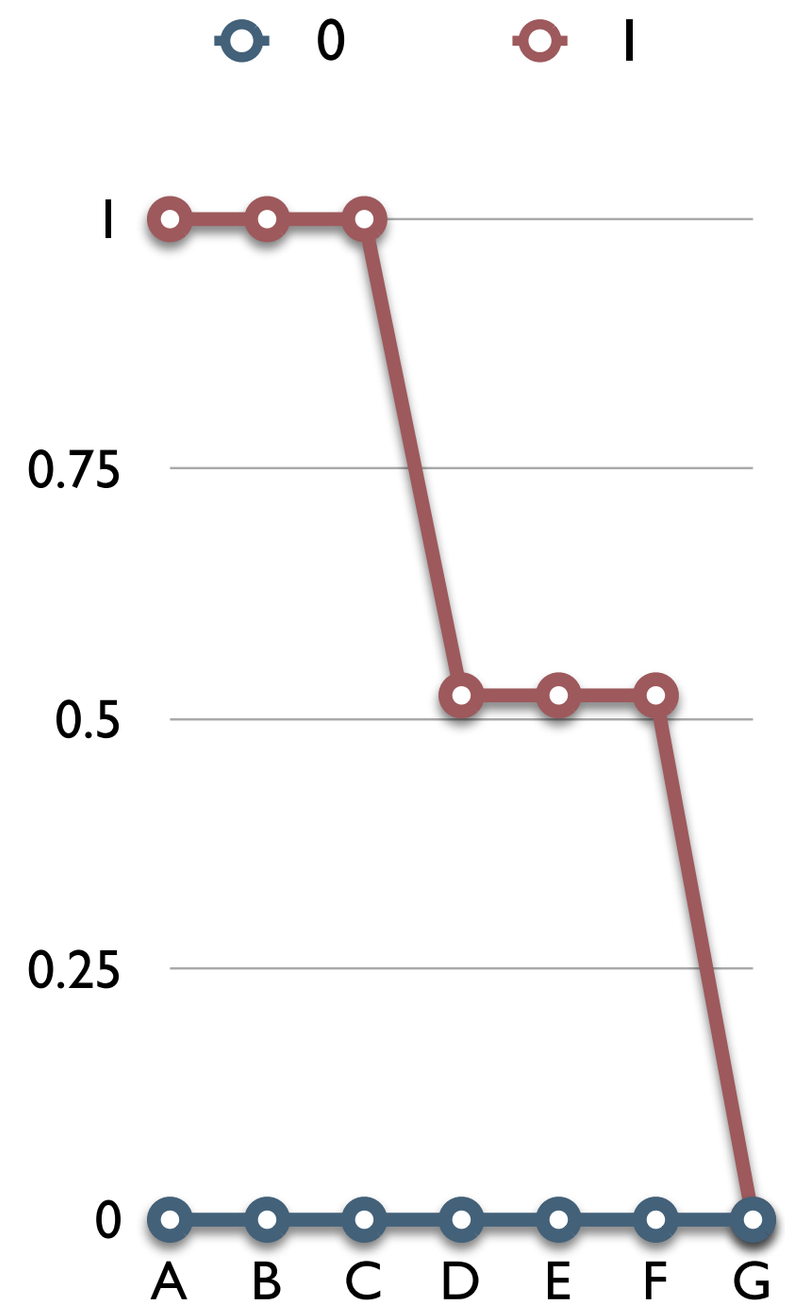
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_{1111000} = 4$$
$$n_{1110011} = 2$$



A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

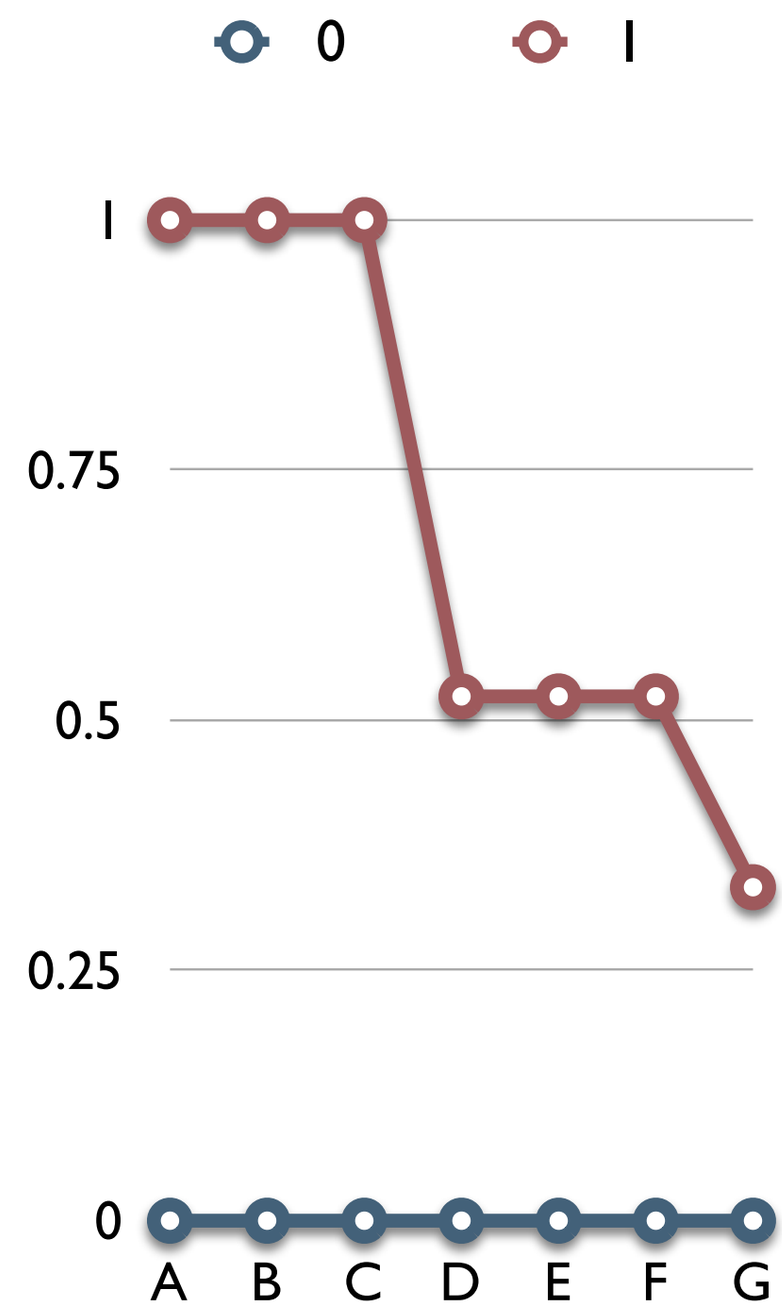
$$n_{1111000} = 4$$
$$n_{1110011} = 2$$
$$n_{1111001} = 1$$



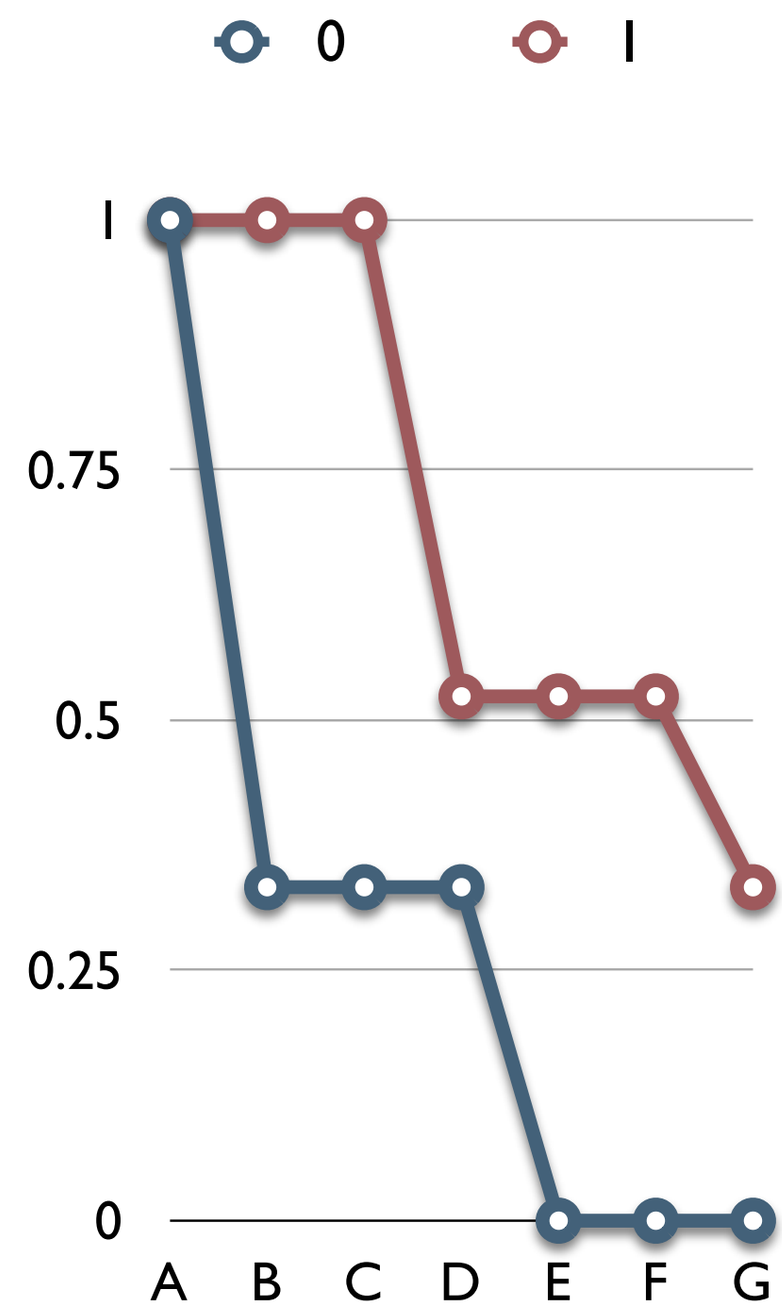
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_{1111000} = 4$$
$$n_{1110011} = 2$$
$$n_{1111001} = 1$$

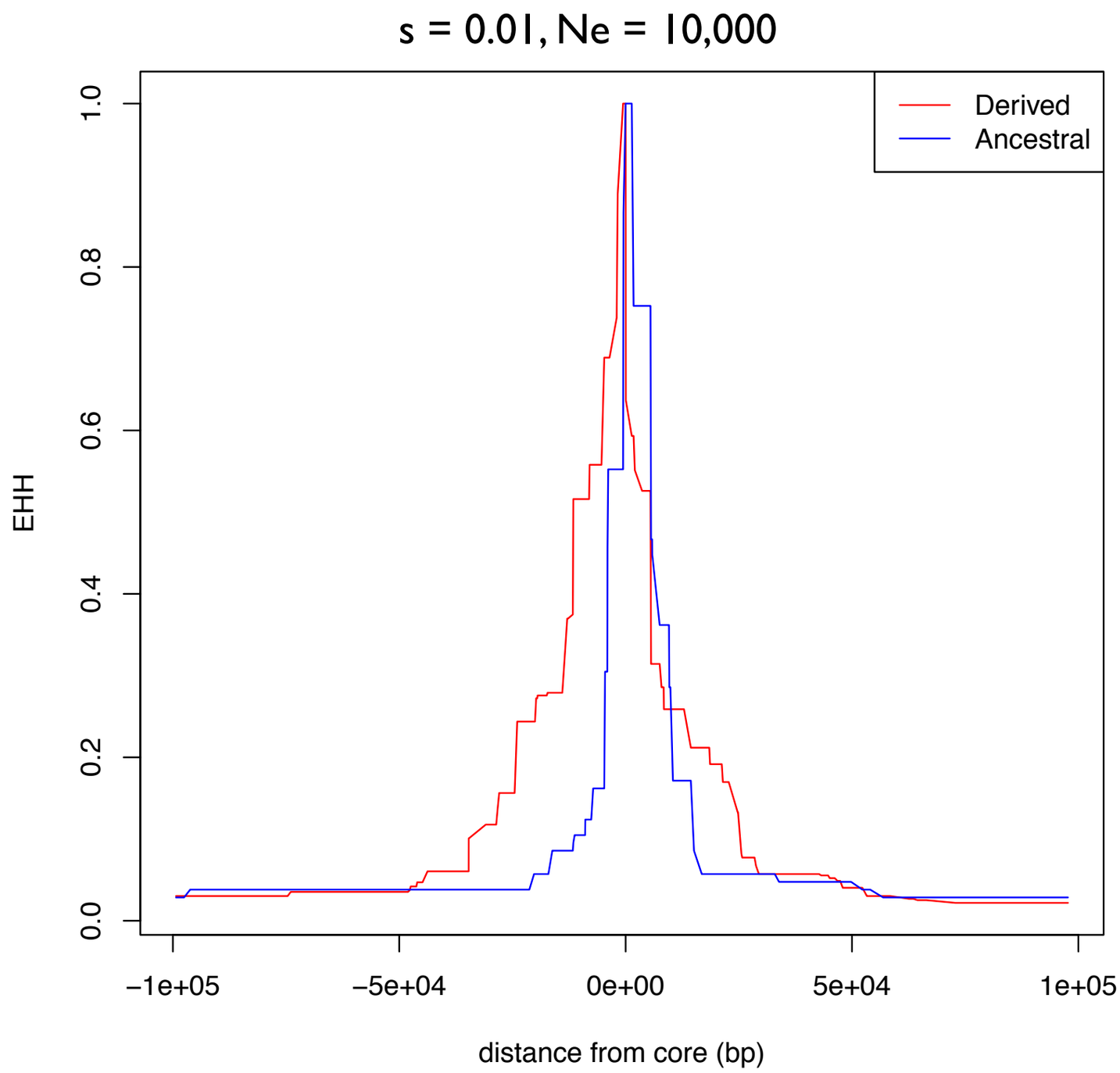
$$EHH_1(G) = \frac{\binom{4}{2}}{\binom{7}{2}} + \frac{\binom{2}{2}}{\binom{7}{2}} + \frac{\binom{1}{2}}{\binom{7}{2}} = \frac{1}{3}$$



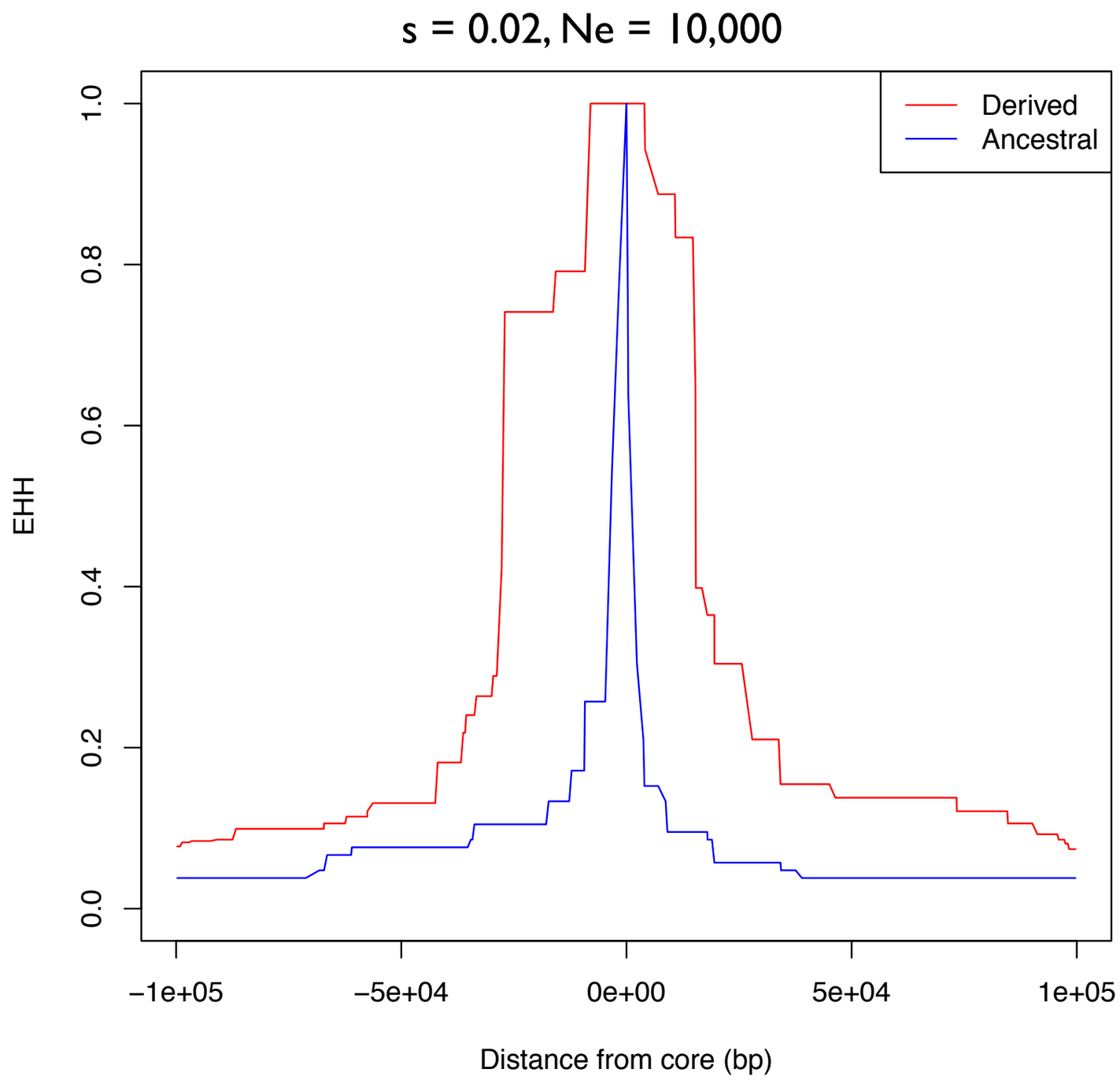
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0



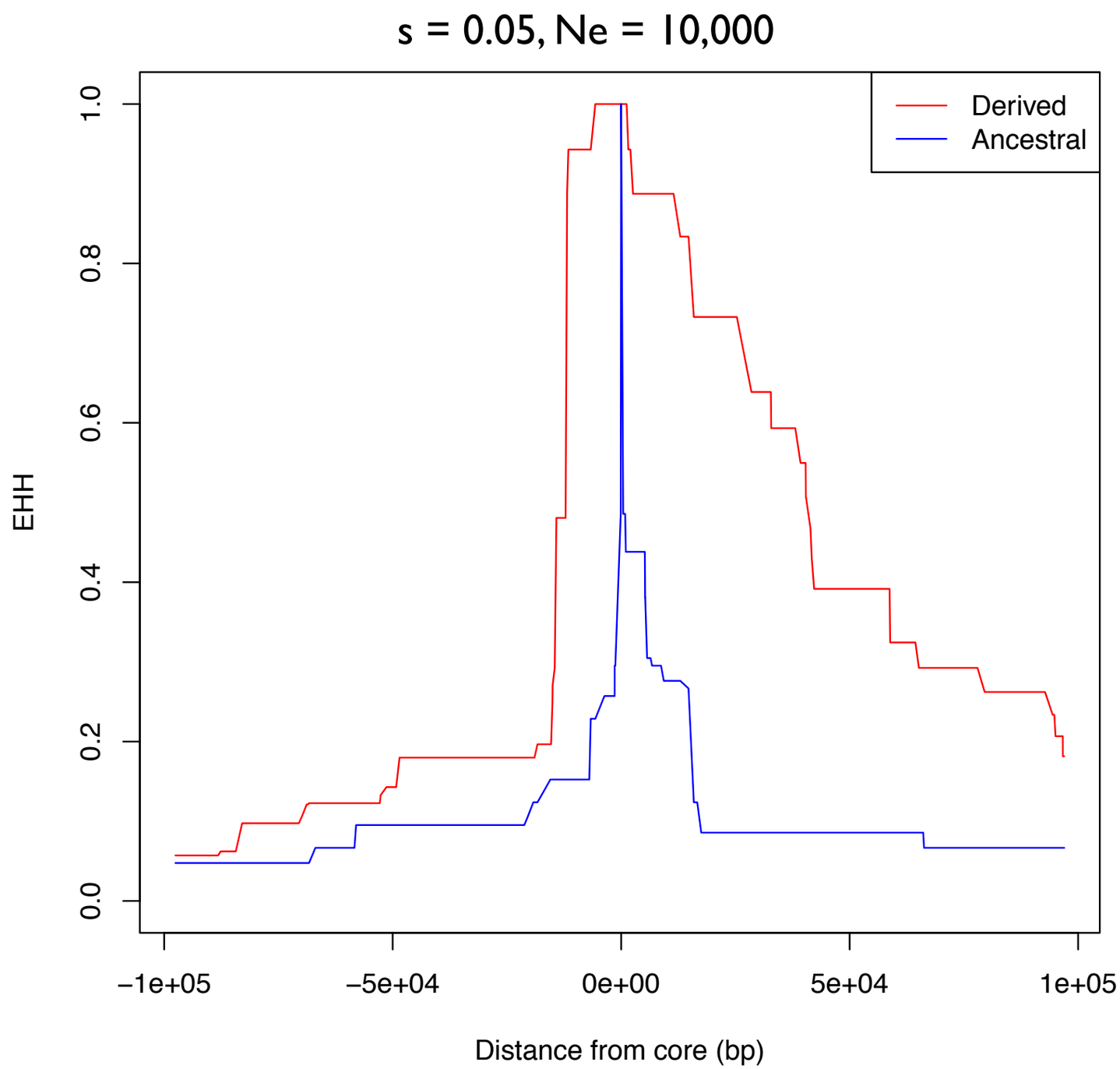
EHH


s = 0.01, Ne = 10,000
Derived
Ancestral
EHH
0.0
0.2
0.4
0.6
0.8
1.0
−1e+05
−5e+04
0e+00
5e+04
1e+05
distance from core (bp)

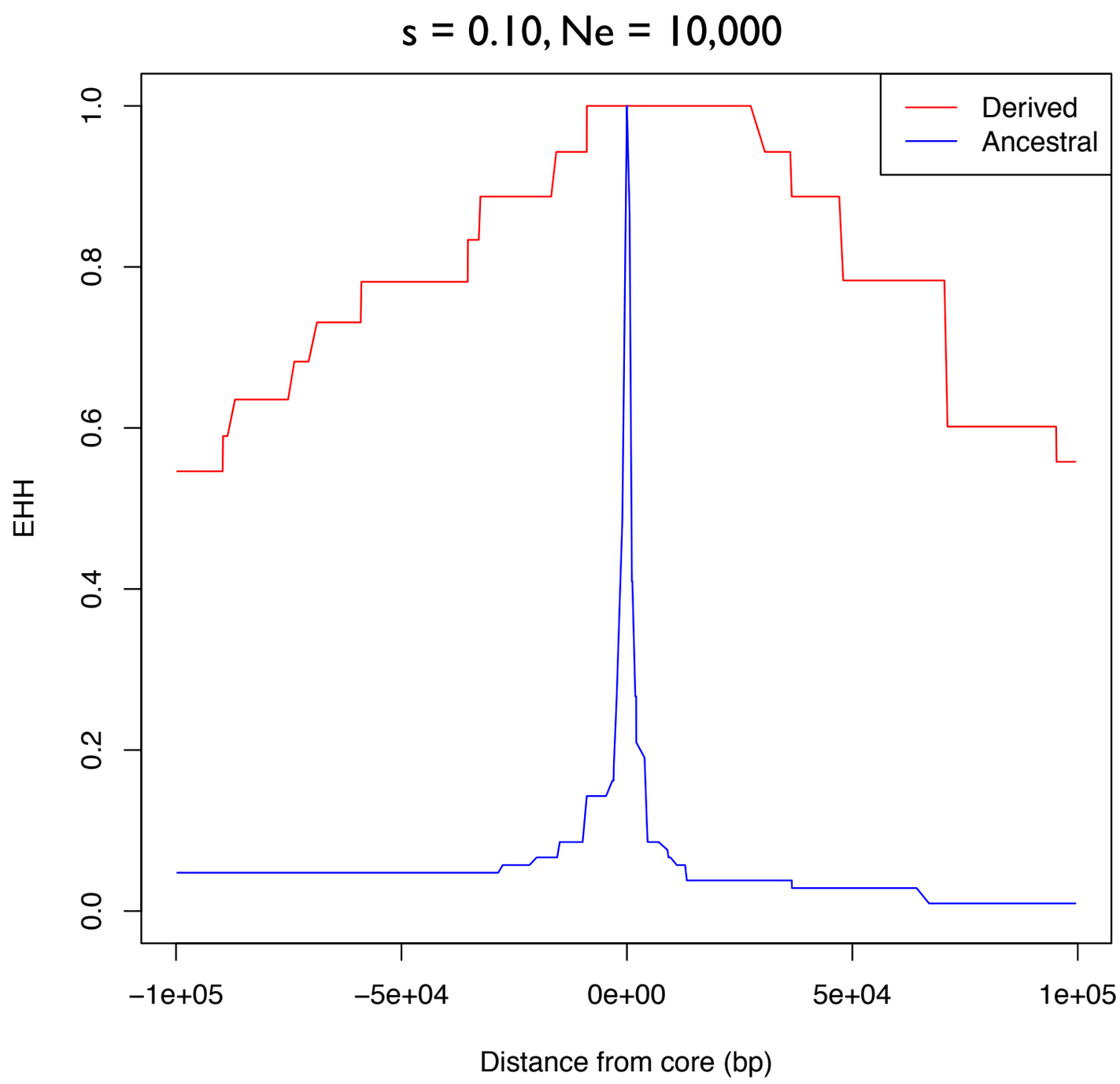
EHH


s = 0.02, Ne = 10,000
Derived
Ancestral
EHH
0.0
0.2
0.4
0.6
0.8
1.0
−1e+05
−5e+04
0e+00
5e+04
1e+05
Distance from core (bp)

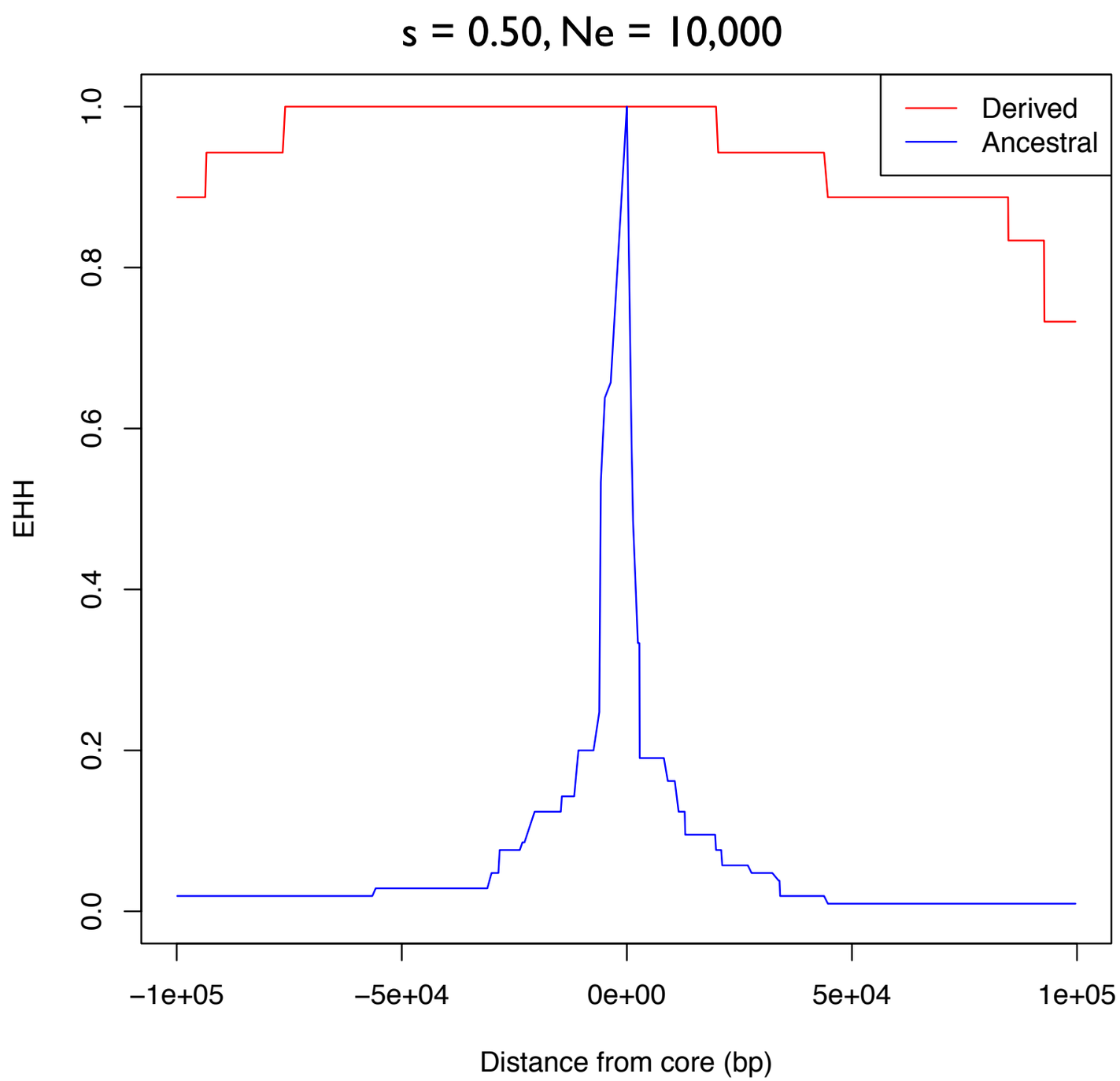
EHH


s = 0.05, Ne = 10,000
Derived
Ancestral
EHH
0.0
0.2
0.4
0.6
0.8
1.0
−1e+05
−5e+04
0e+00
5e+04
1e+05
Distance from core (bp)

EHH


s = 0.10, Ne = 10,000
Derived
Ancestral
EHH
0.0
0.2
0.4
0.6
0.8
1.0
−1e+05
−5e+04
0e+00
5e+04
1e+05
Distance from core (bp)

EHH


s = 0.50, Ne = 10,000
Derived
Ancestral
EHH
0.0
0.2
0.4
0.6
0.8
1.0
−1e+05
−5e+04
0e+00
5e+04
1e+05
Distance from core (bp)

EHH

- When querying a specific region of the genome, for each core haplotype, calculate EHH for successively longer surrounding haplotypes.
- Statistical significance is determined by comparing EHH scores to neutral simulations and random control regions of the genome.