Searching for Signatures of Selection with Selscan

Ryan Hernandez

ryan.hernandez@ucsf.edu

The Effect of Positive Selection

Adaptive

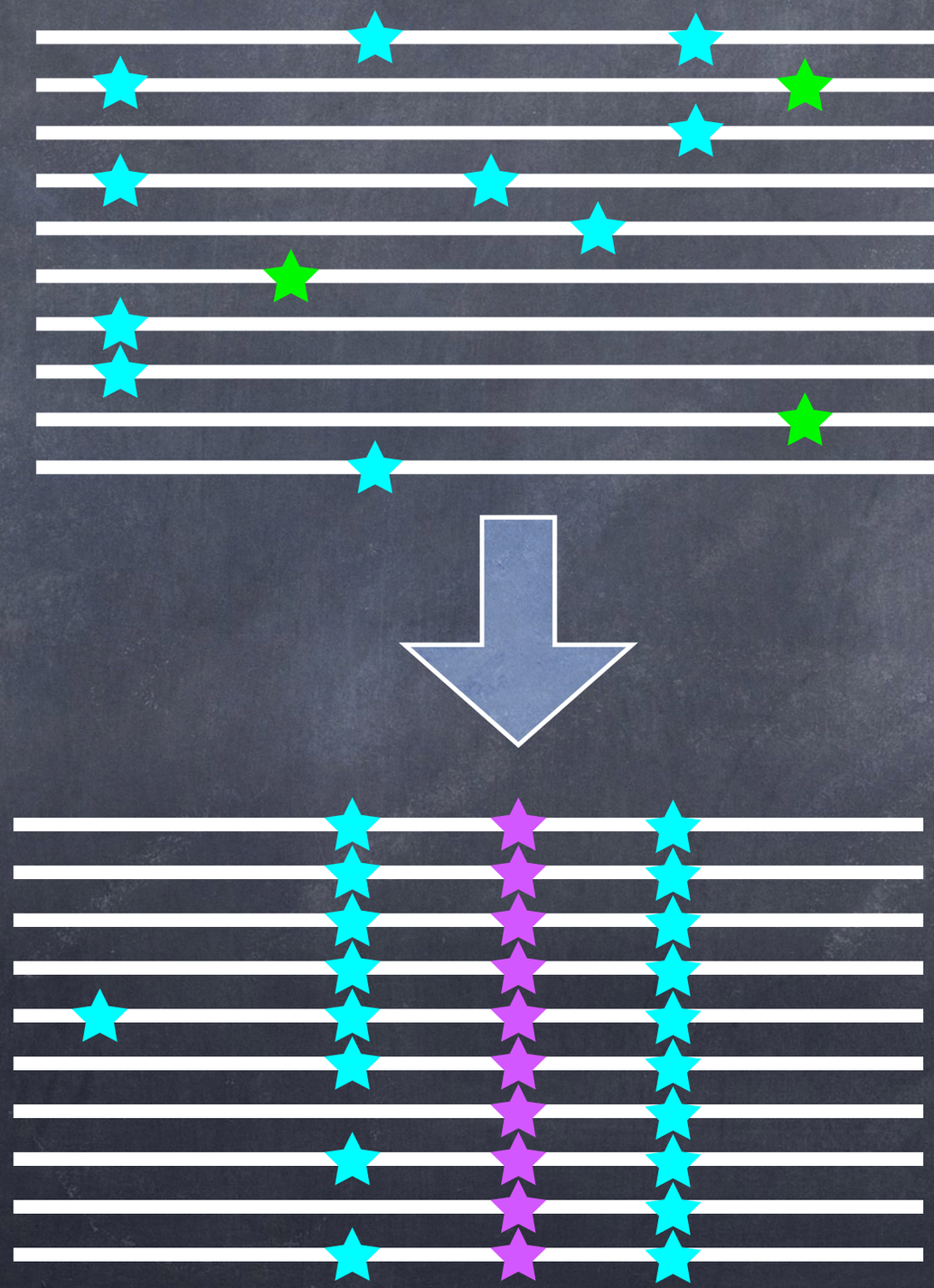
Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious



How do we capture this process in a statistic?

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

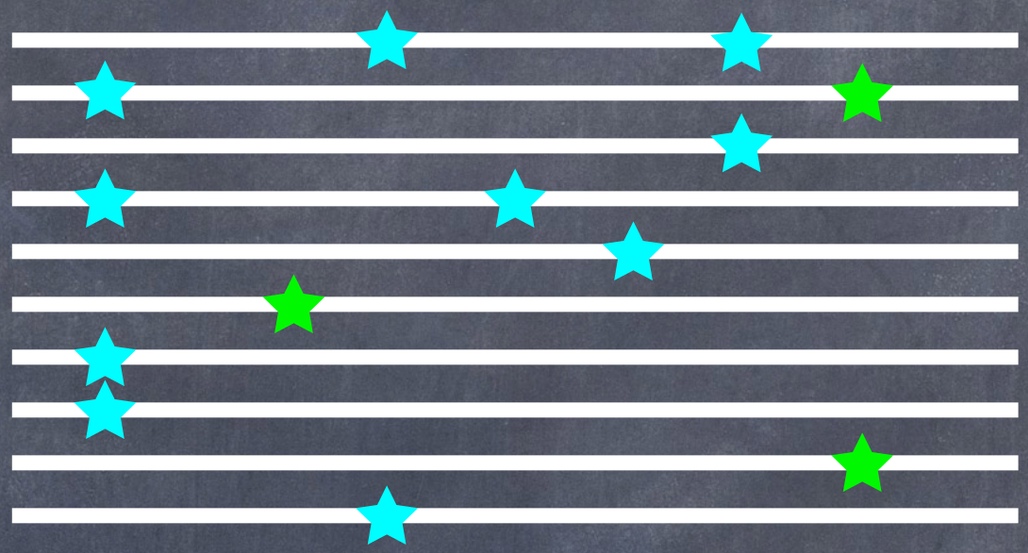
Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

Strongly Deleterious



The Effect of Positive Selection

Adaptive

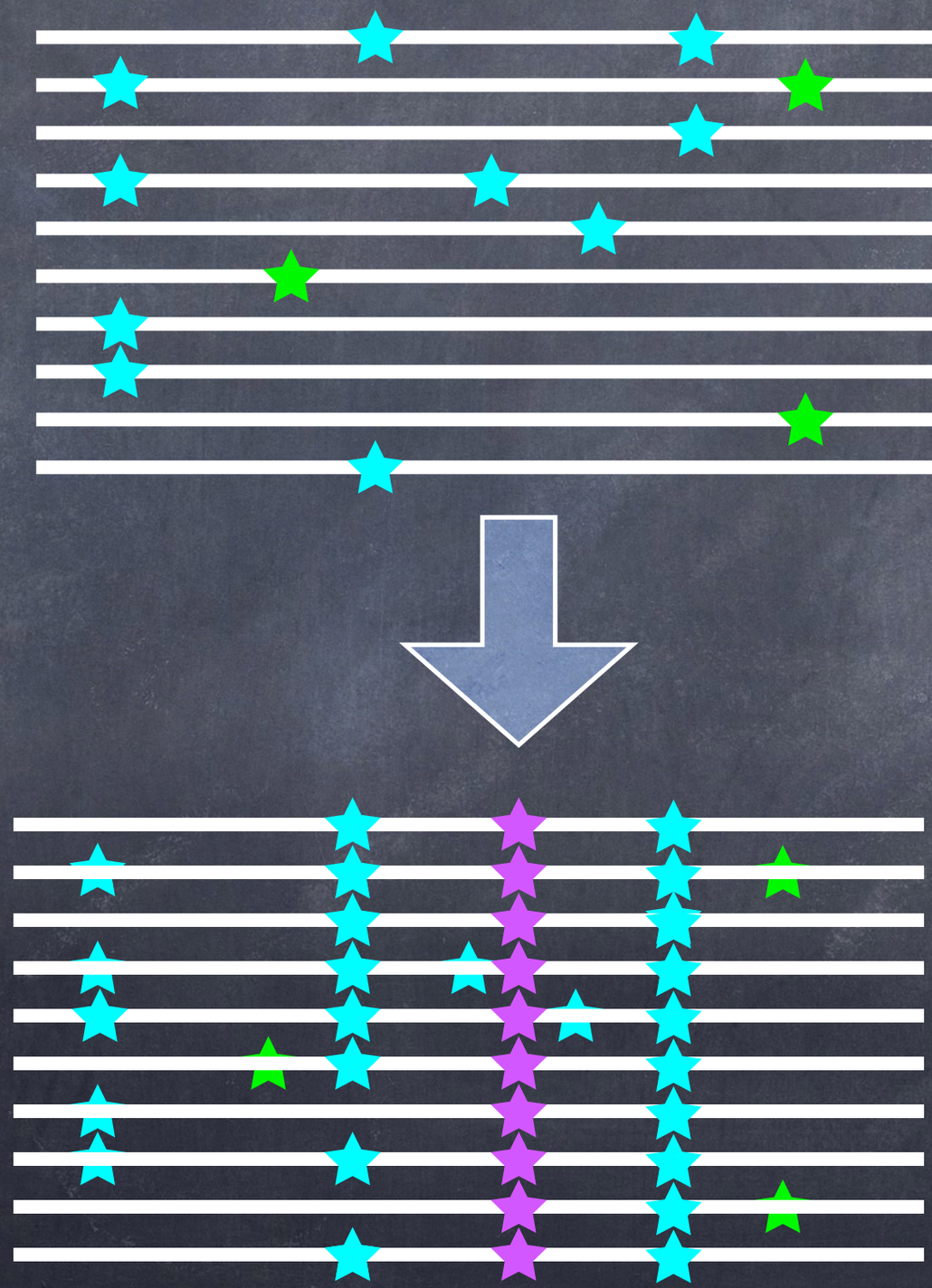
Neutral

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

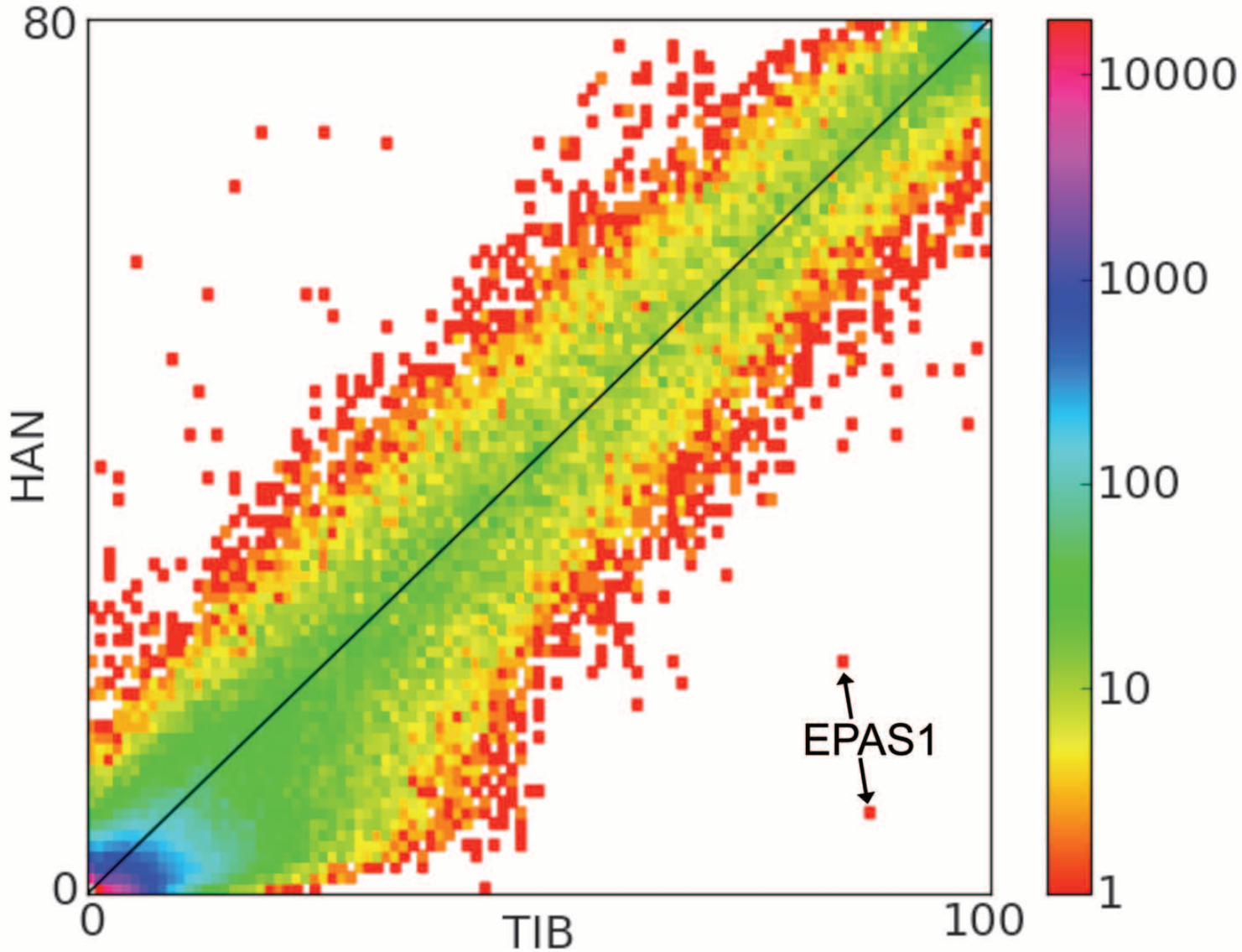
Testing for Population Divergence

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

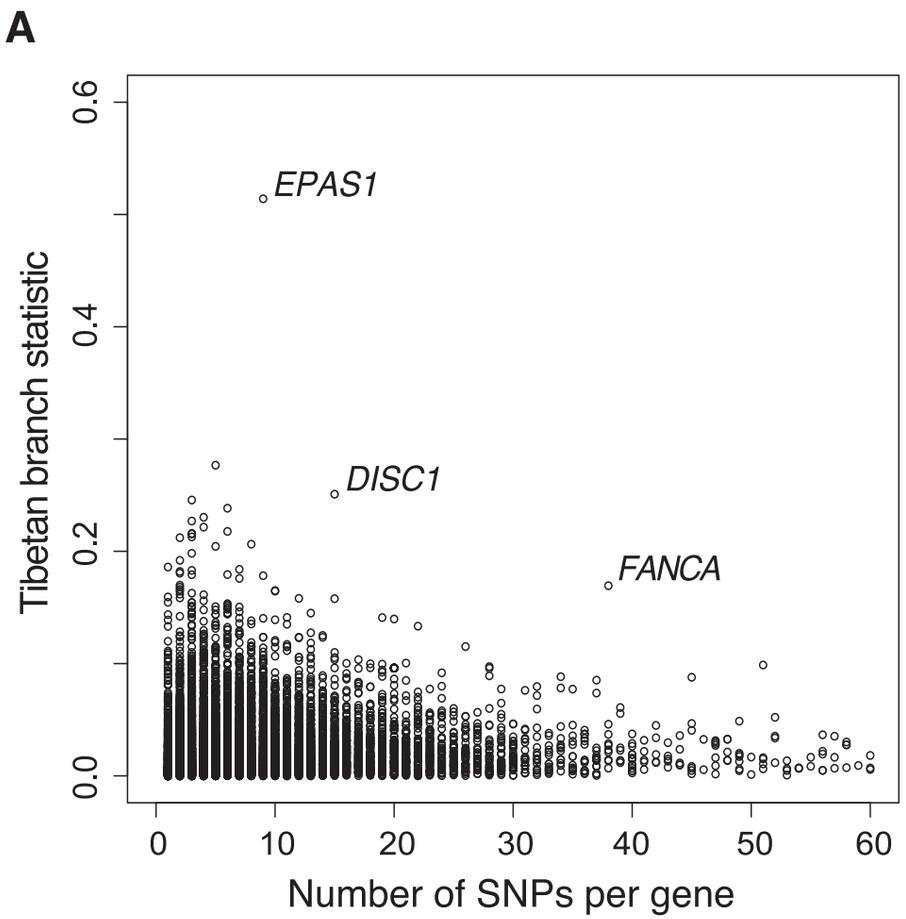


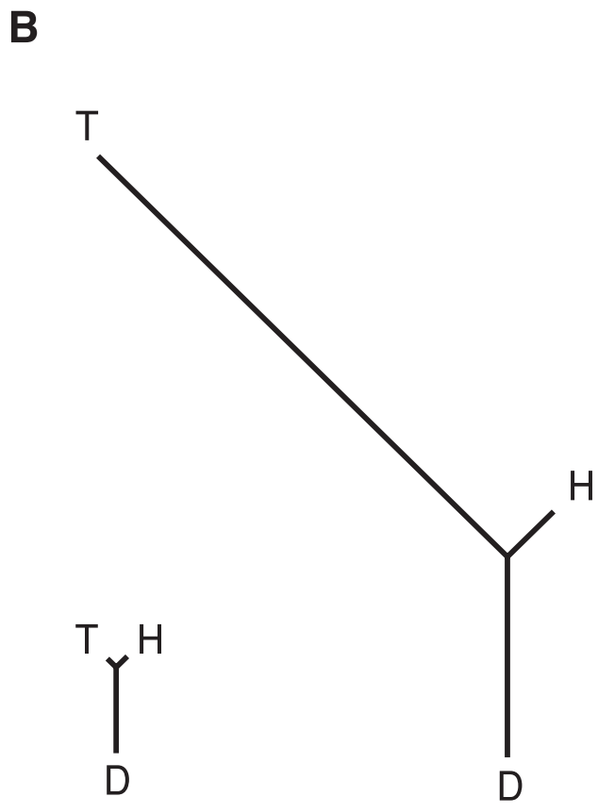
EPAS1: a transcription factor involved in response to hypoxia

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

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

Testing for Population Divergence



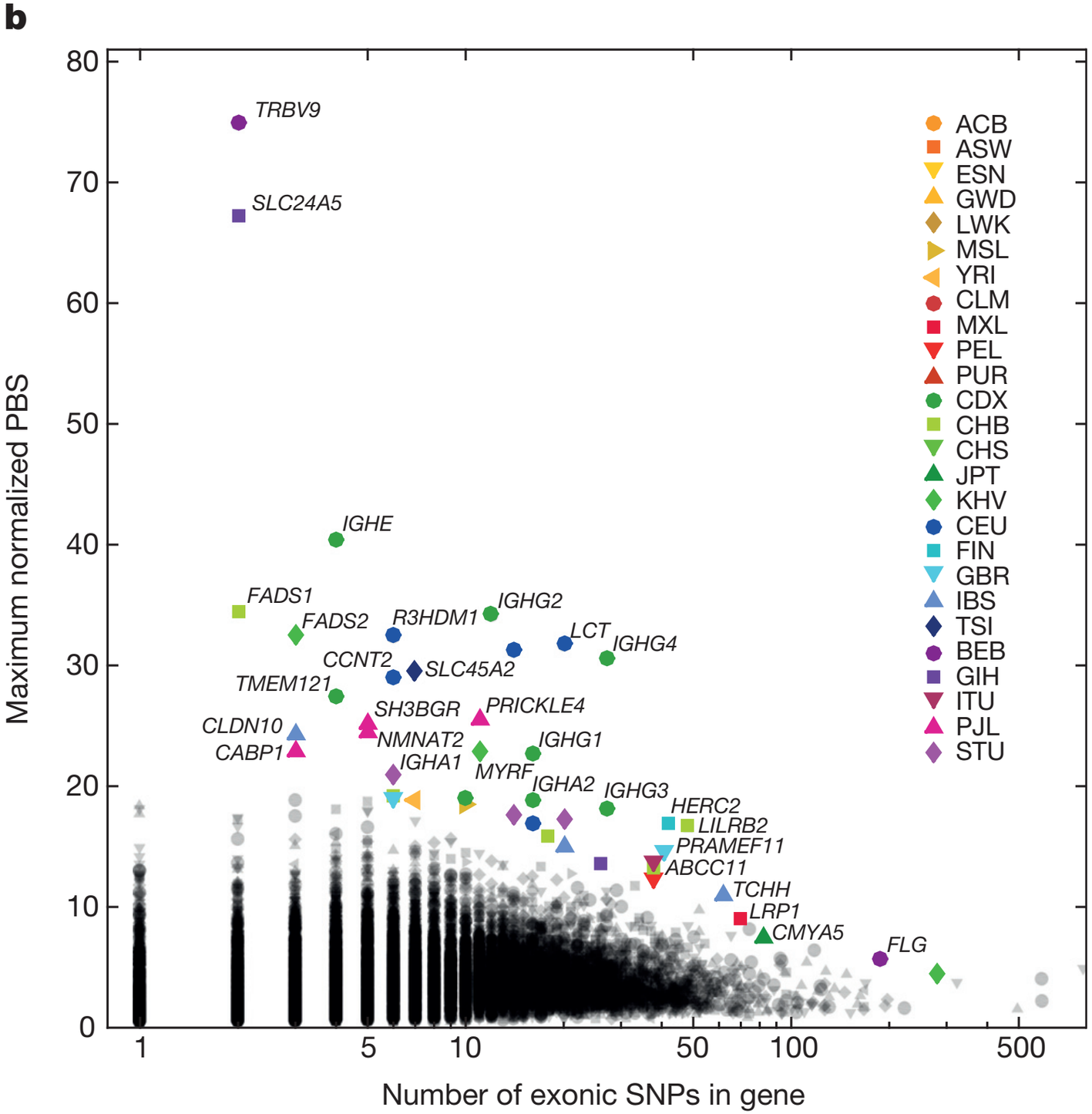


EPAS1: a transcription factor involved in response to hypoxia

- 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

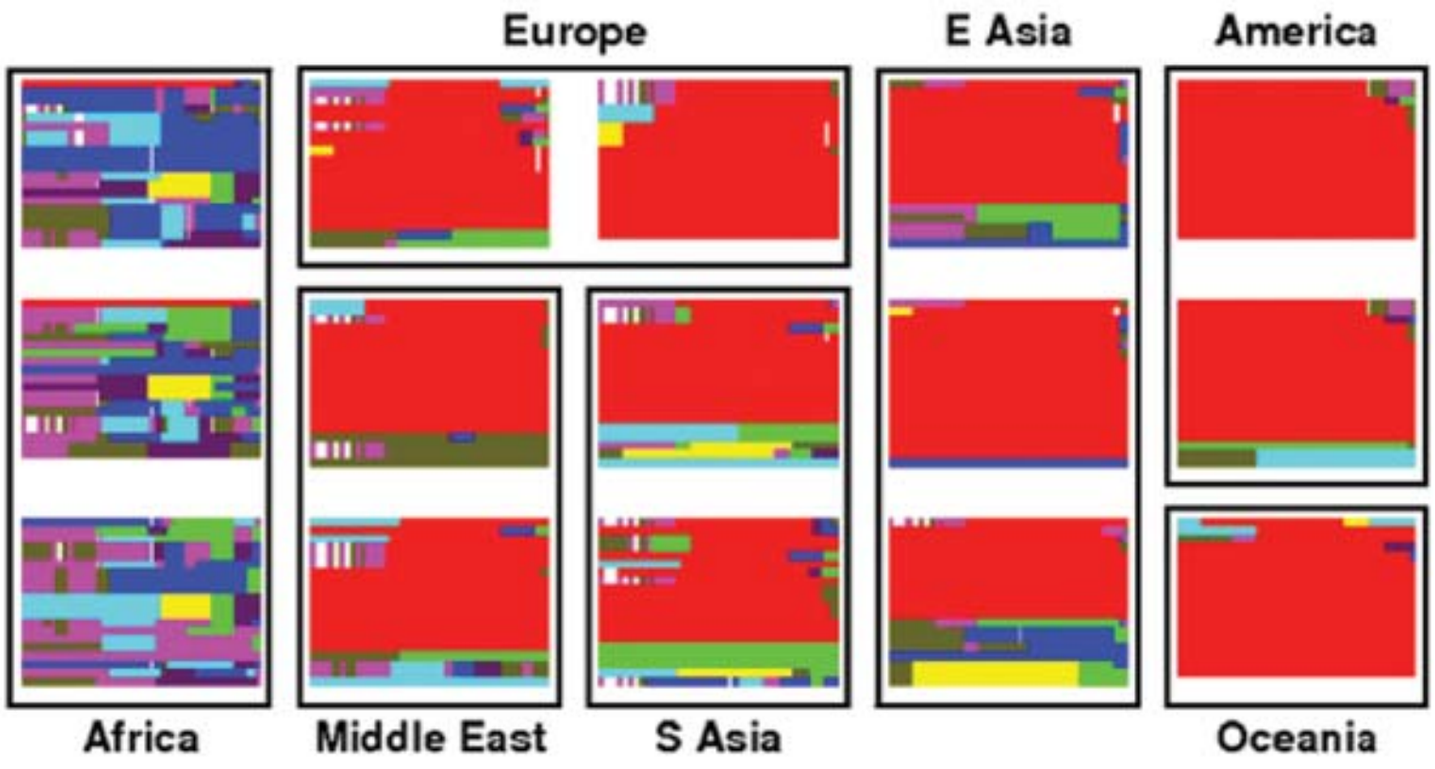
- Selection operates on a new mutation
  - Selection will act to increase the frequency of the allele
  - Results in a young allele at relatively high frequency
  - The test is simple:
    - Are there young alleles at unusually high frequency?

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

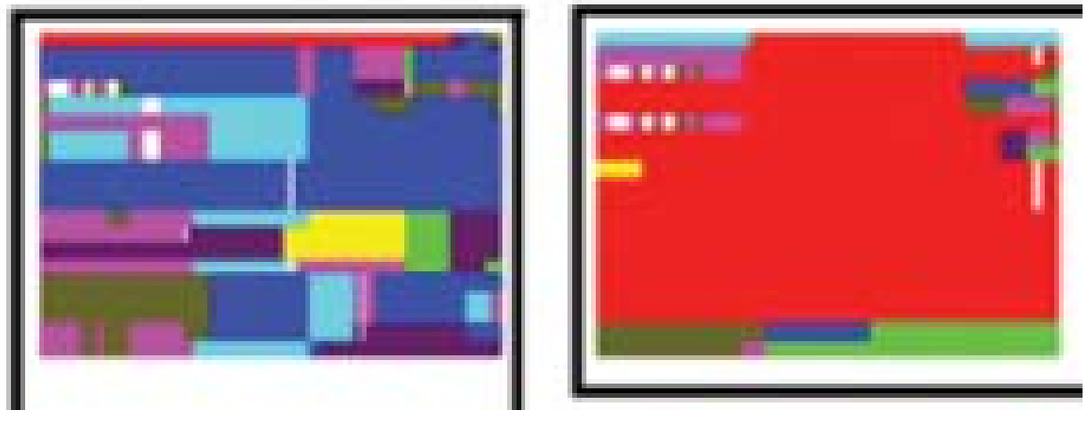
☑ Selection acts in one population but not another

☑ 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

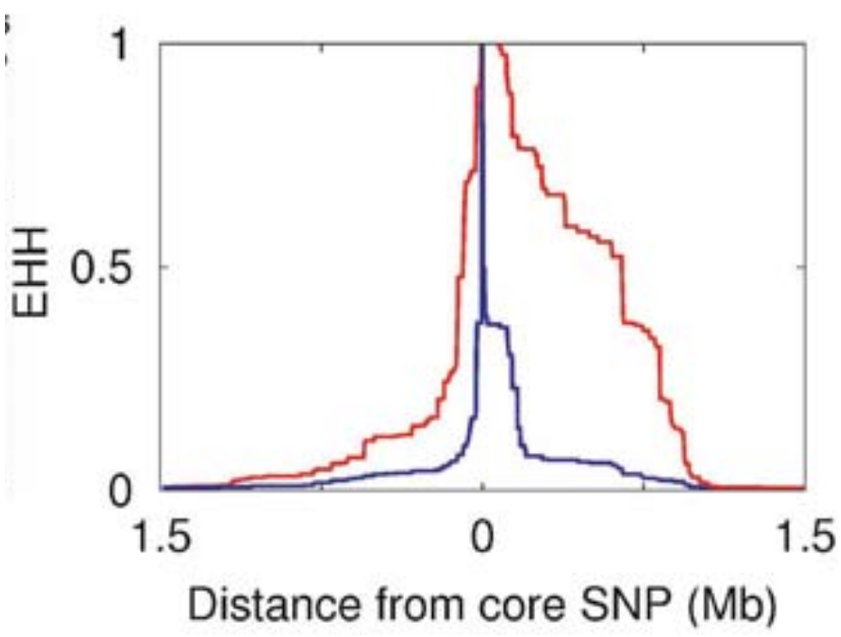
Divergence of a Young Allele

- Recall the haplotype patterns before for just two populations:



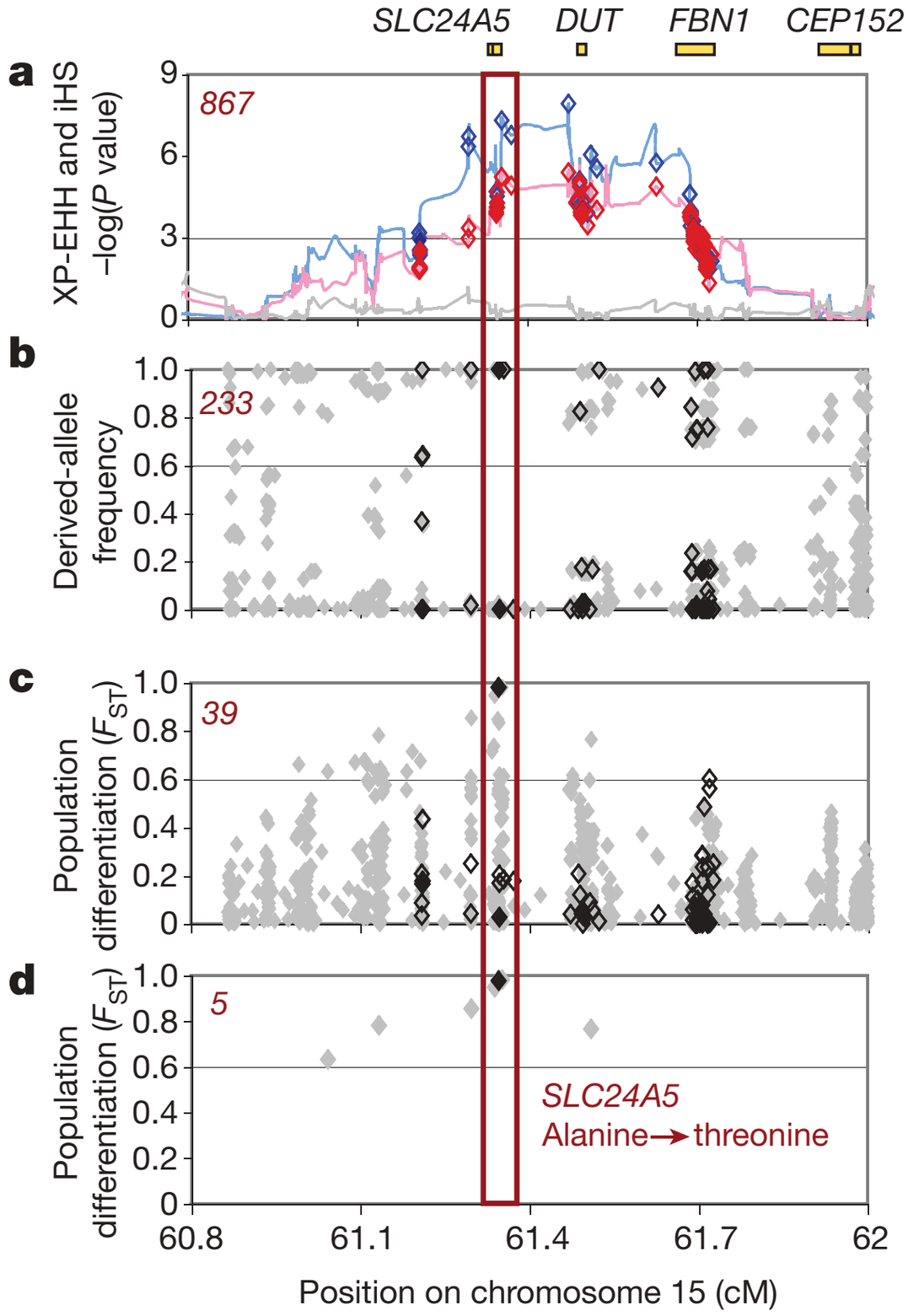
- These can be plotted as the probability that two randomly chosen individuals have an identical haplotype as a function of distance from the core SNP:

- 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

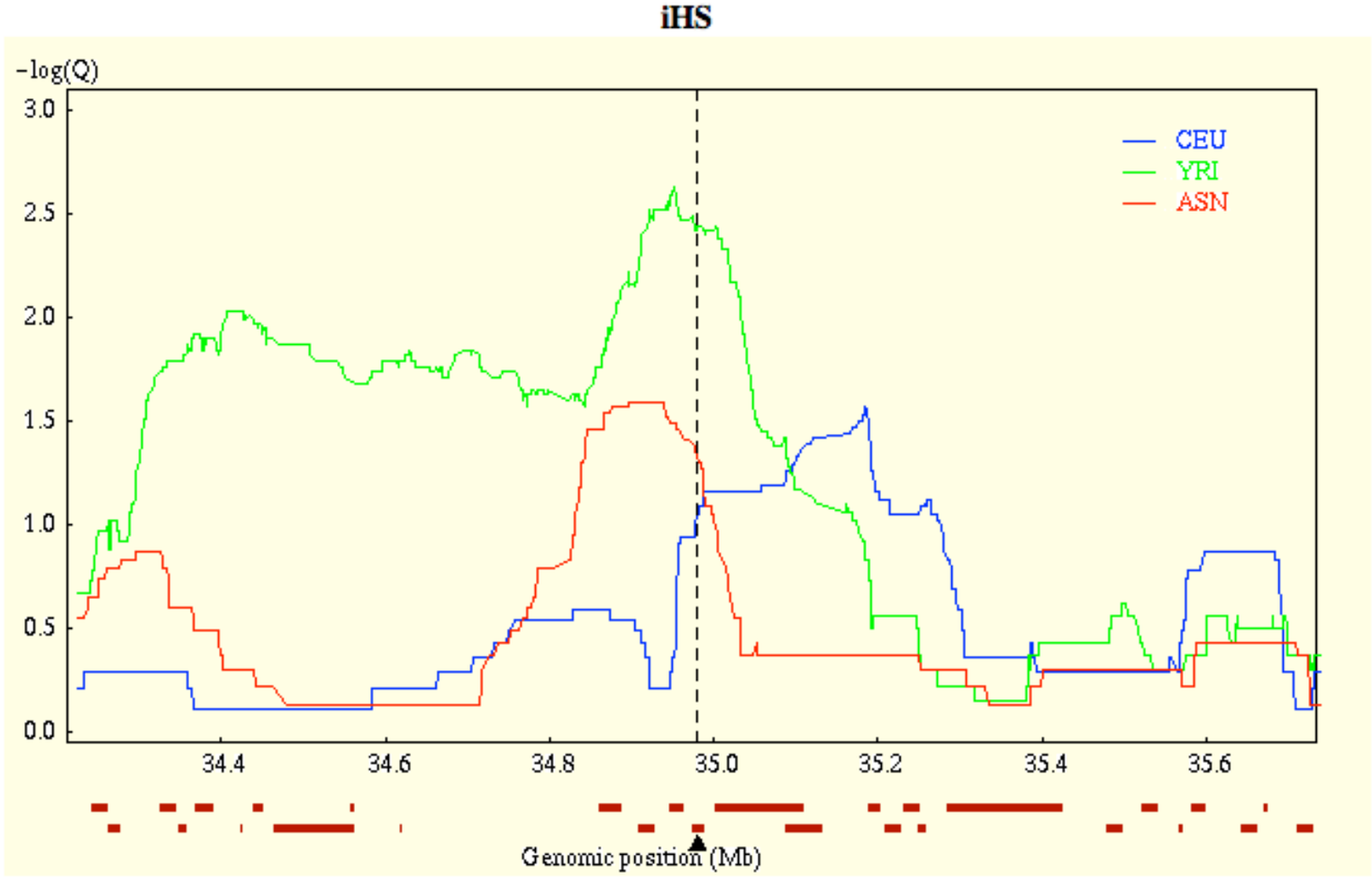
- Individuals of African descent have much higher incidence of kidney disease than individuals of European descent.
- GWAS had previously implicated the gene MYH9 with moderate effects ($p<10^{-8}$)
- But there was no clear biological story.

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):

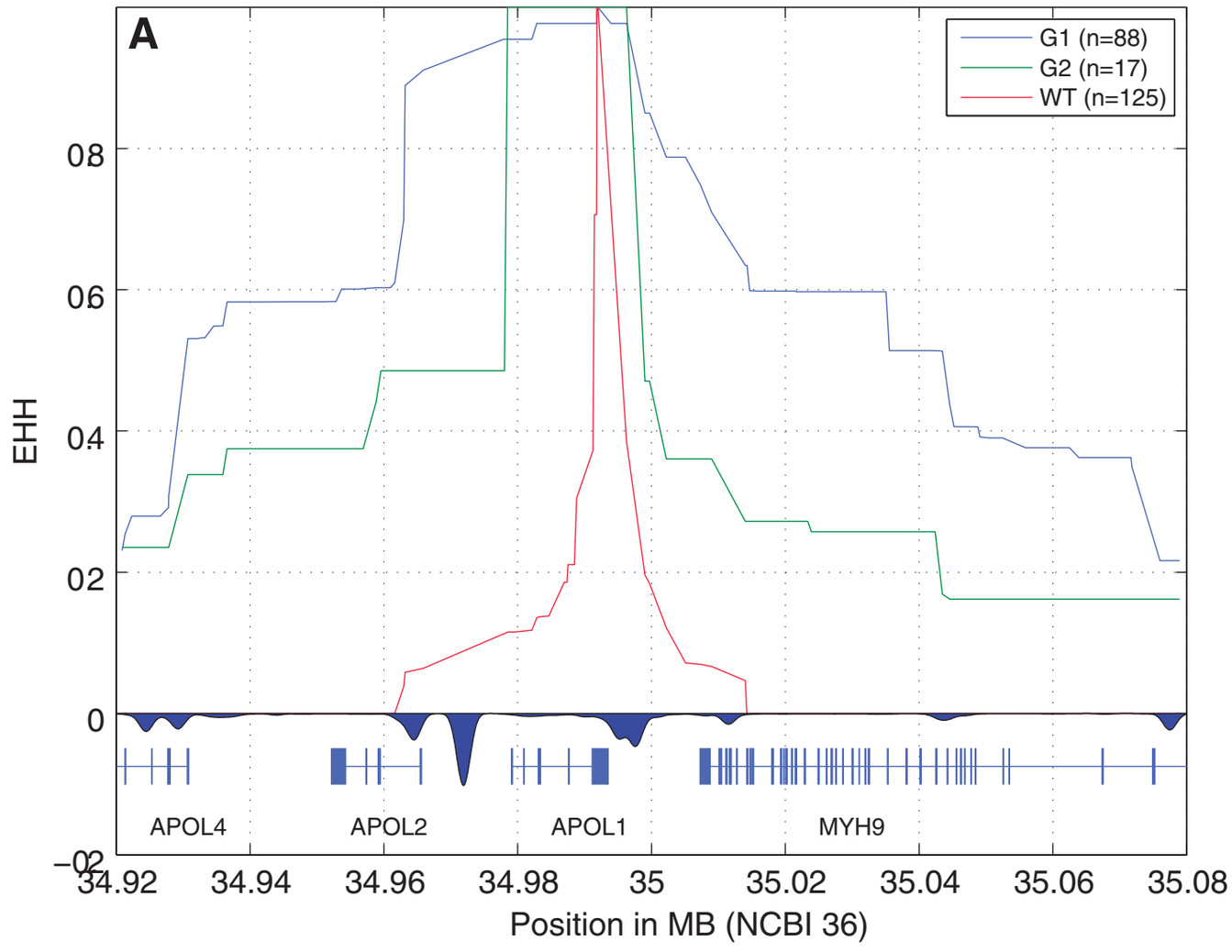
Gene	iHS p-value
APOL1	0.0033
MYH9	0.014



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

Case Study: Kidney Disease in African Americans

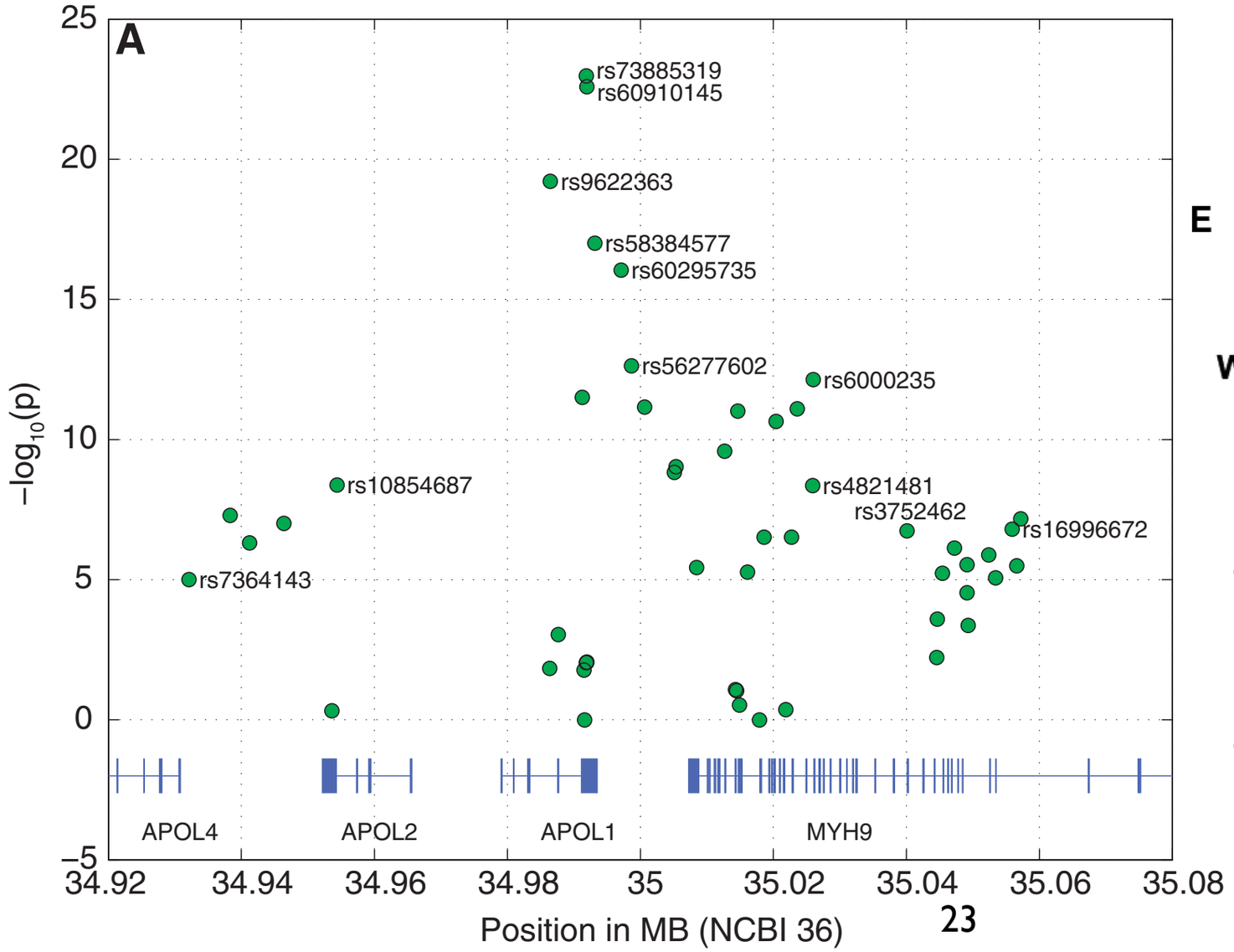
- Tag SNPs chosen across a broader region, and calculated EHH based on higher resolution data

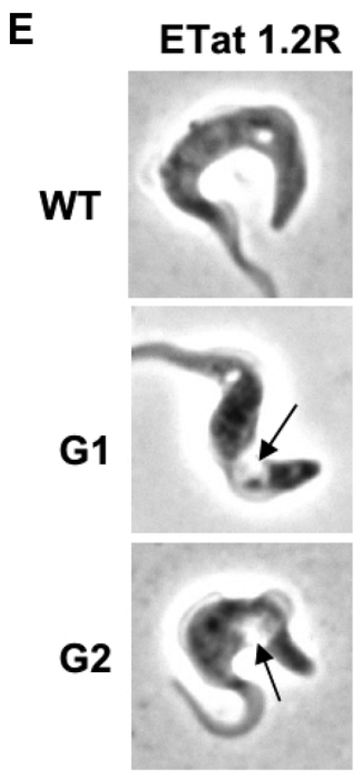


Genovese, G. et al. *Science* **329**, 841-845 (2010).

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).

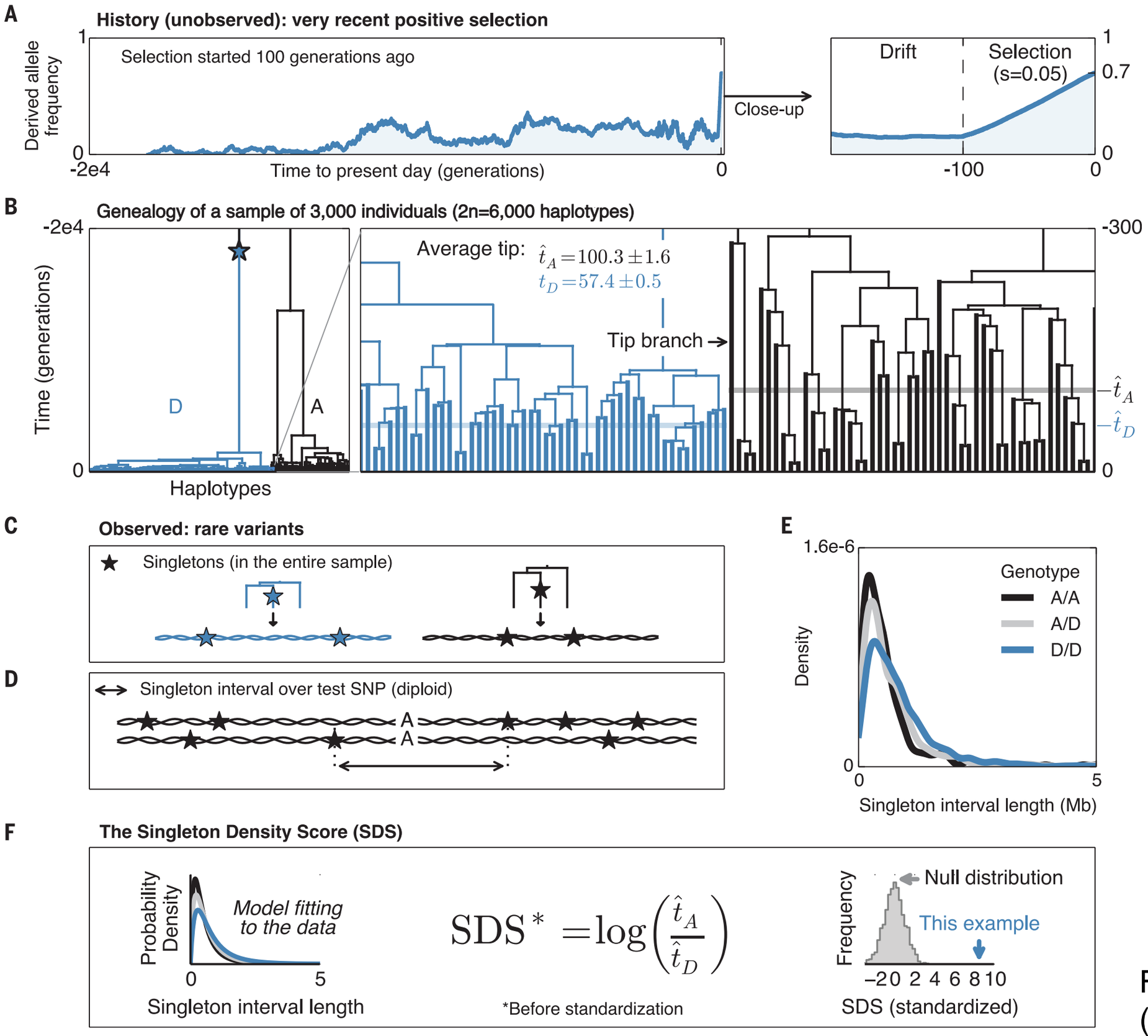
WGS

- 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

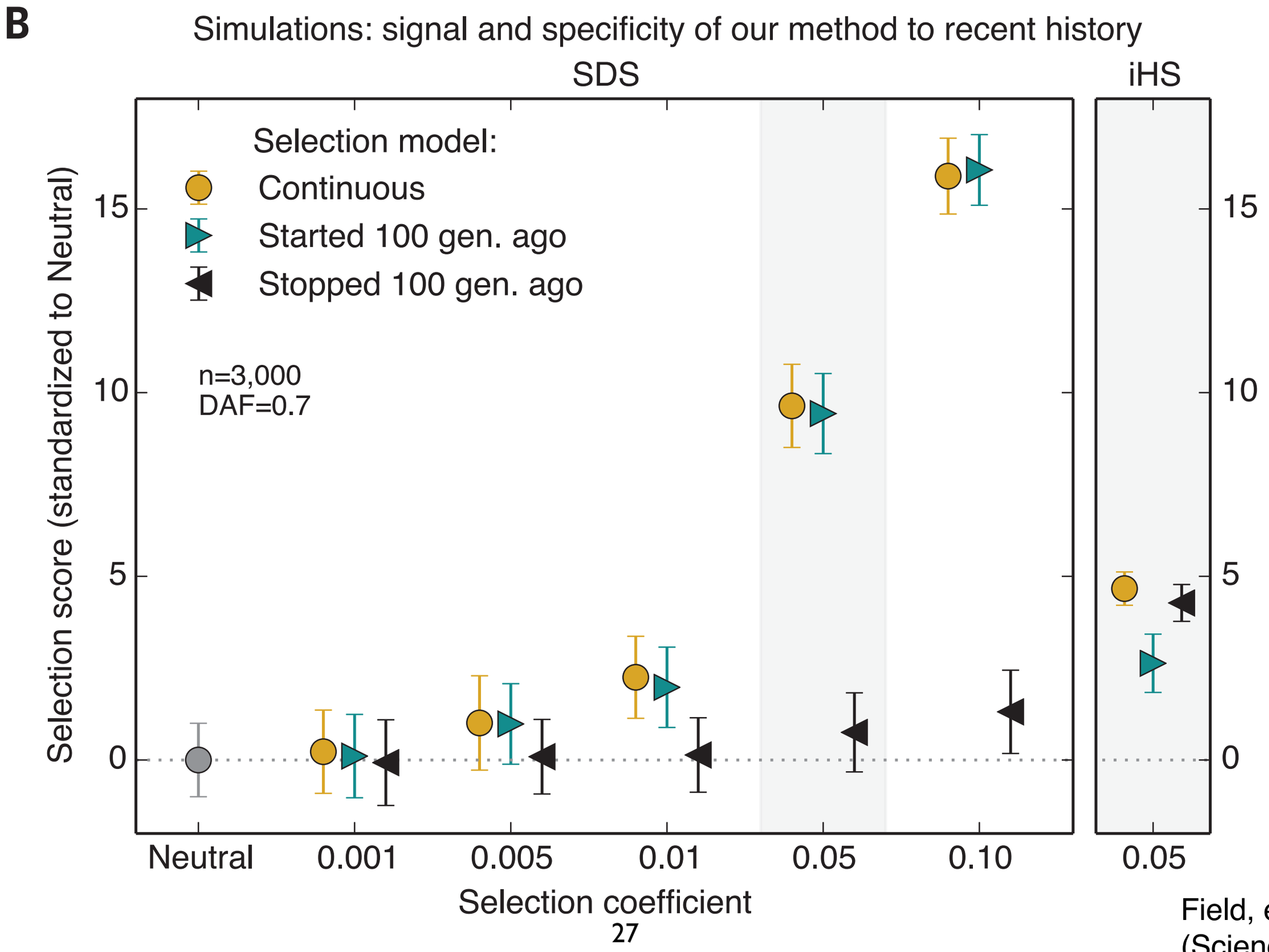
- Field, et al. (*Science*, 2016) introduced the Singleton Density Score (SDS) to capitalize on WGS data with very large samples.
- In the presence of a sweep, the distribution of distances (across individuals) to the nearest singleton will be skewed towards longer distances.

SDS



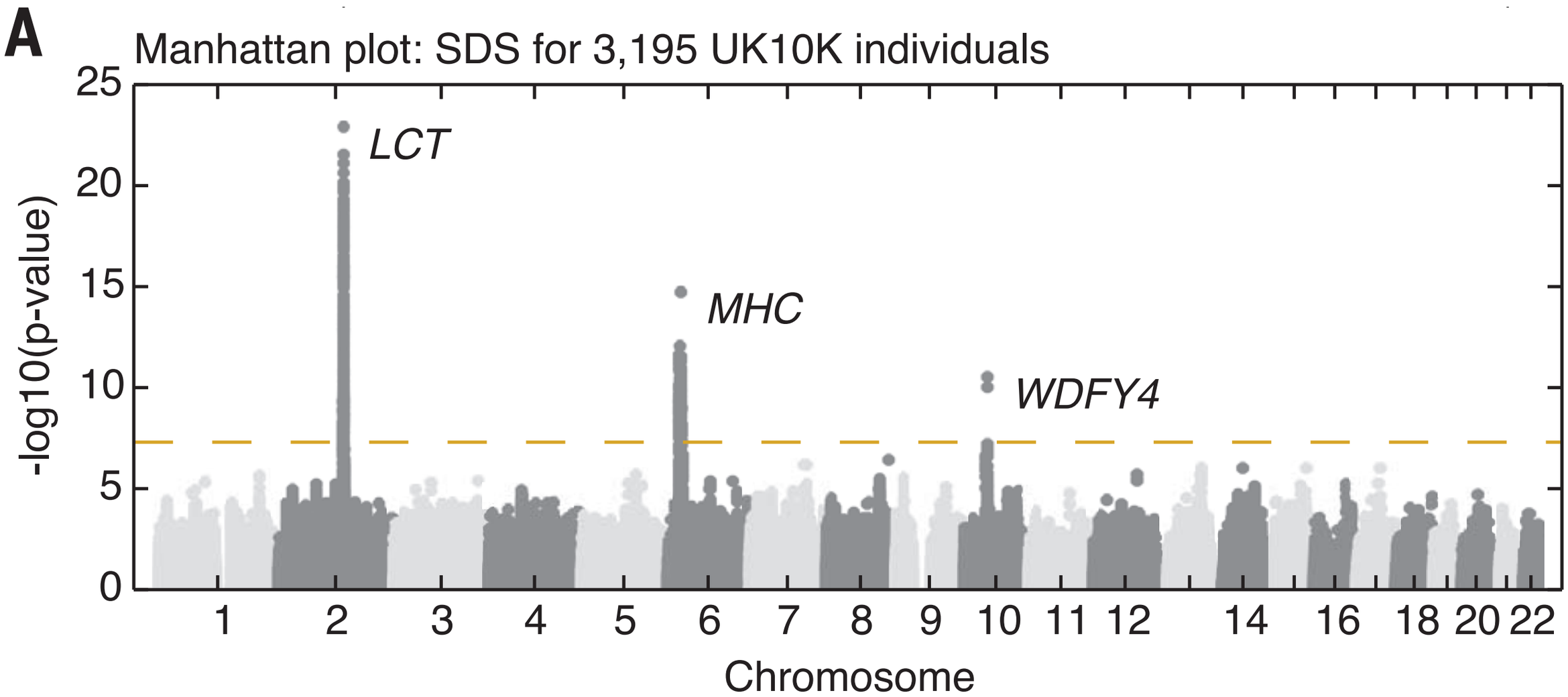
Field, et al. (Science, 2016)

SDS



Field, et al.
(Science, 2016)

SDS



Field, et al.
(Science, 2016)

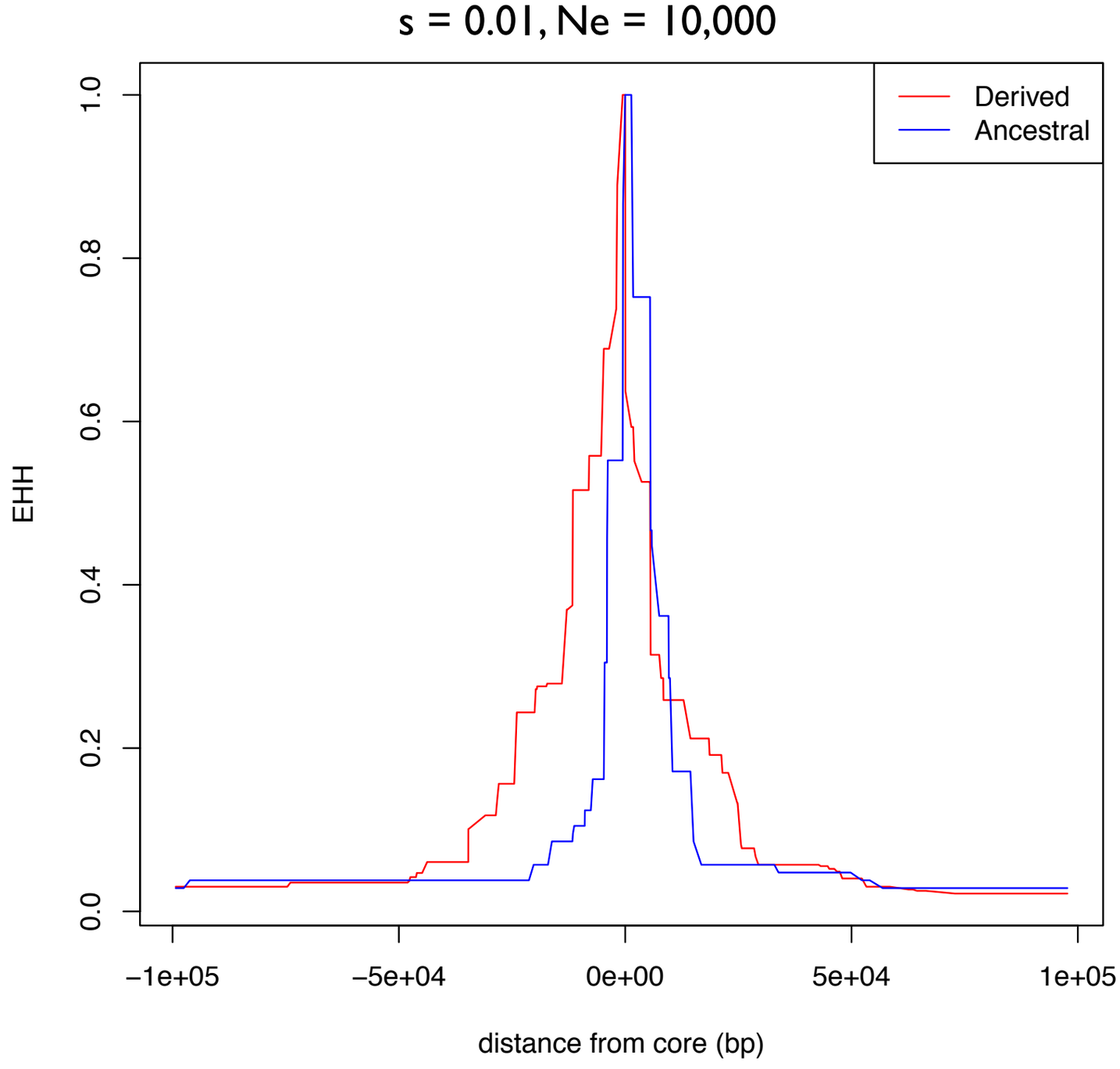
Conclusions

- 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.

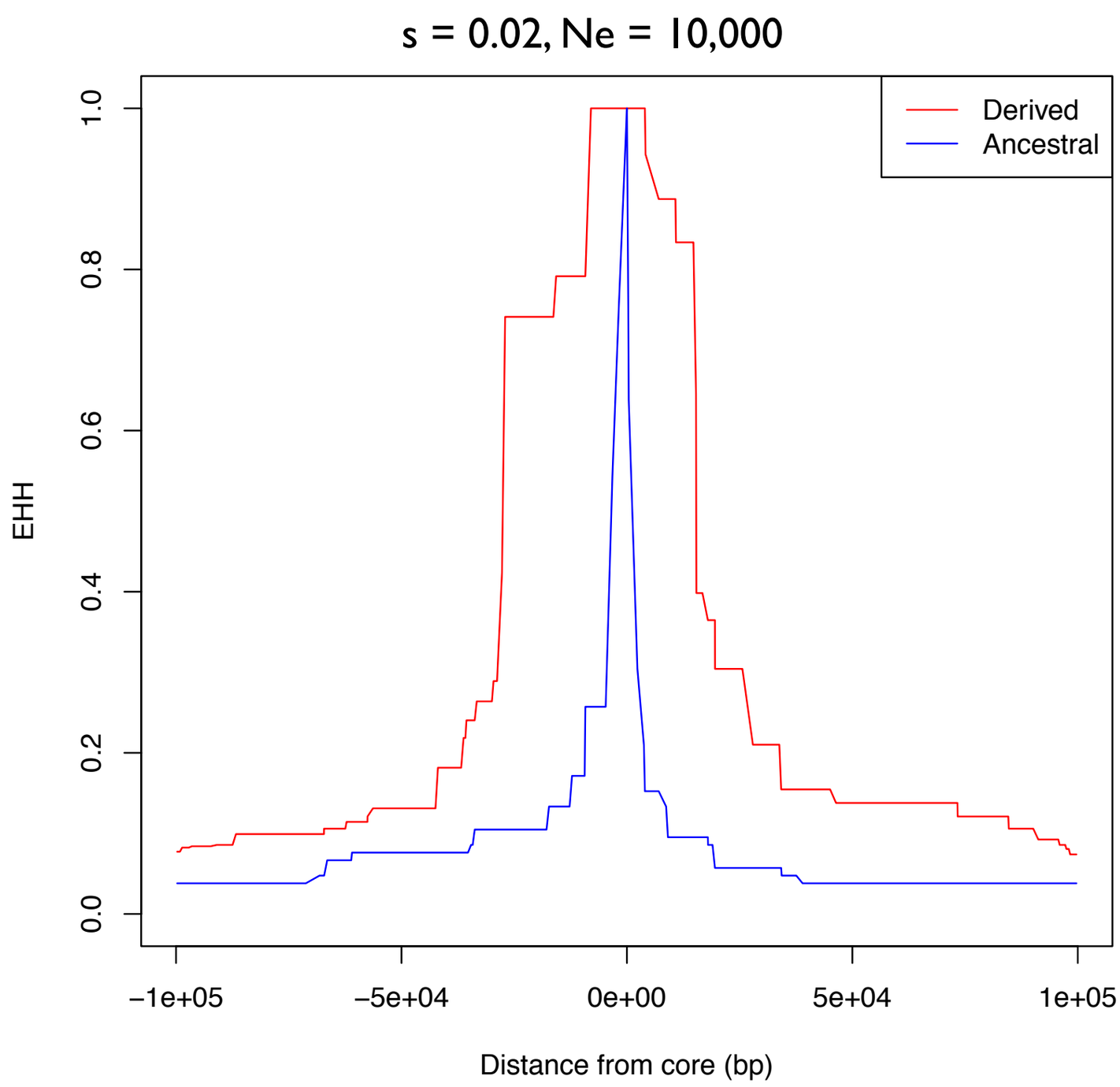
Extended Haplotype Homozygosity

- 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).

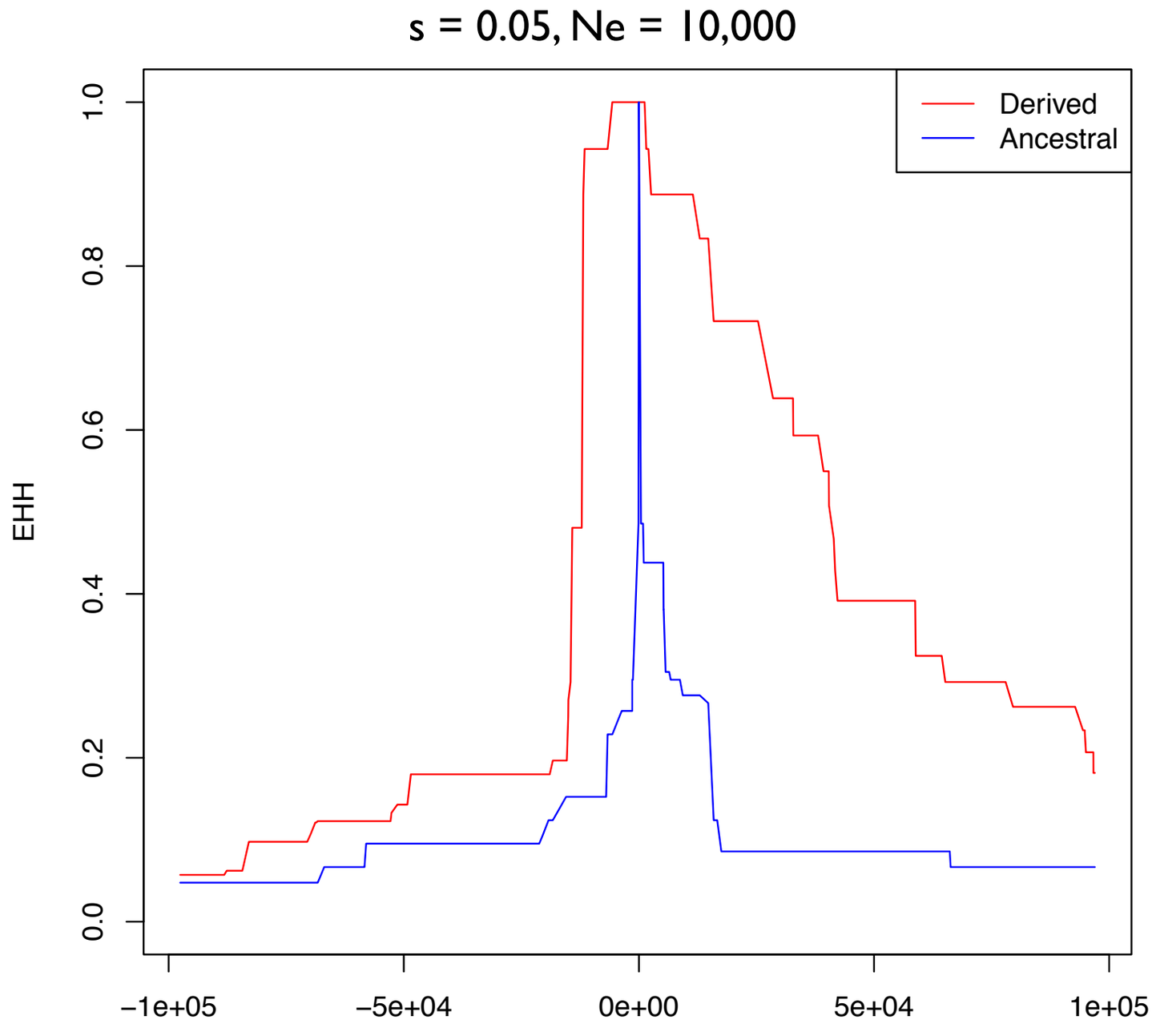
EHH


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

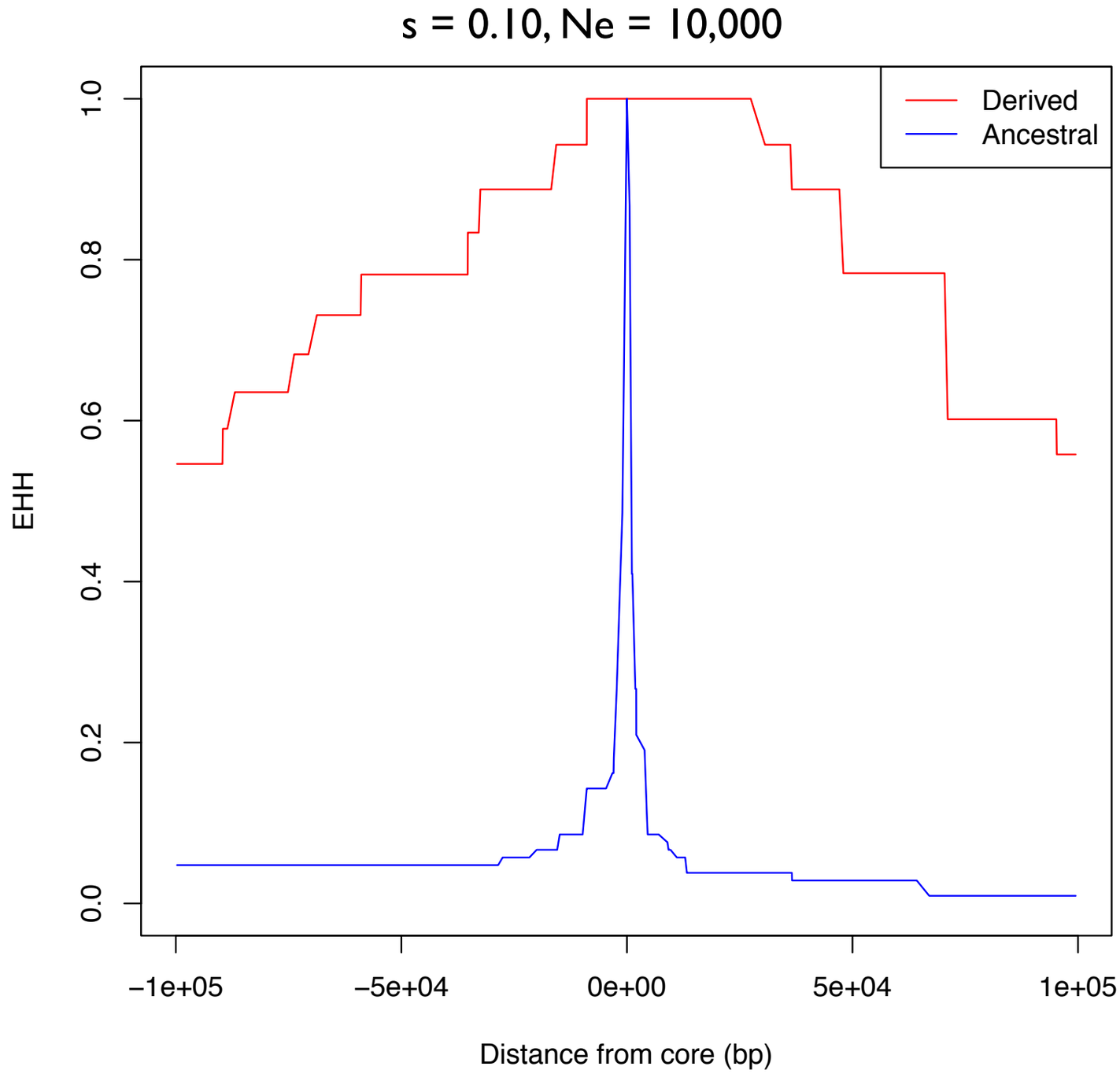
EHH


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)

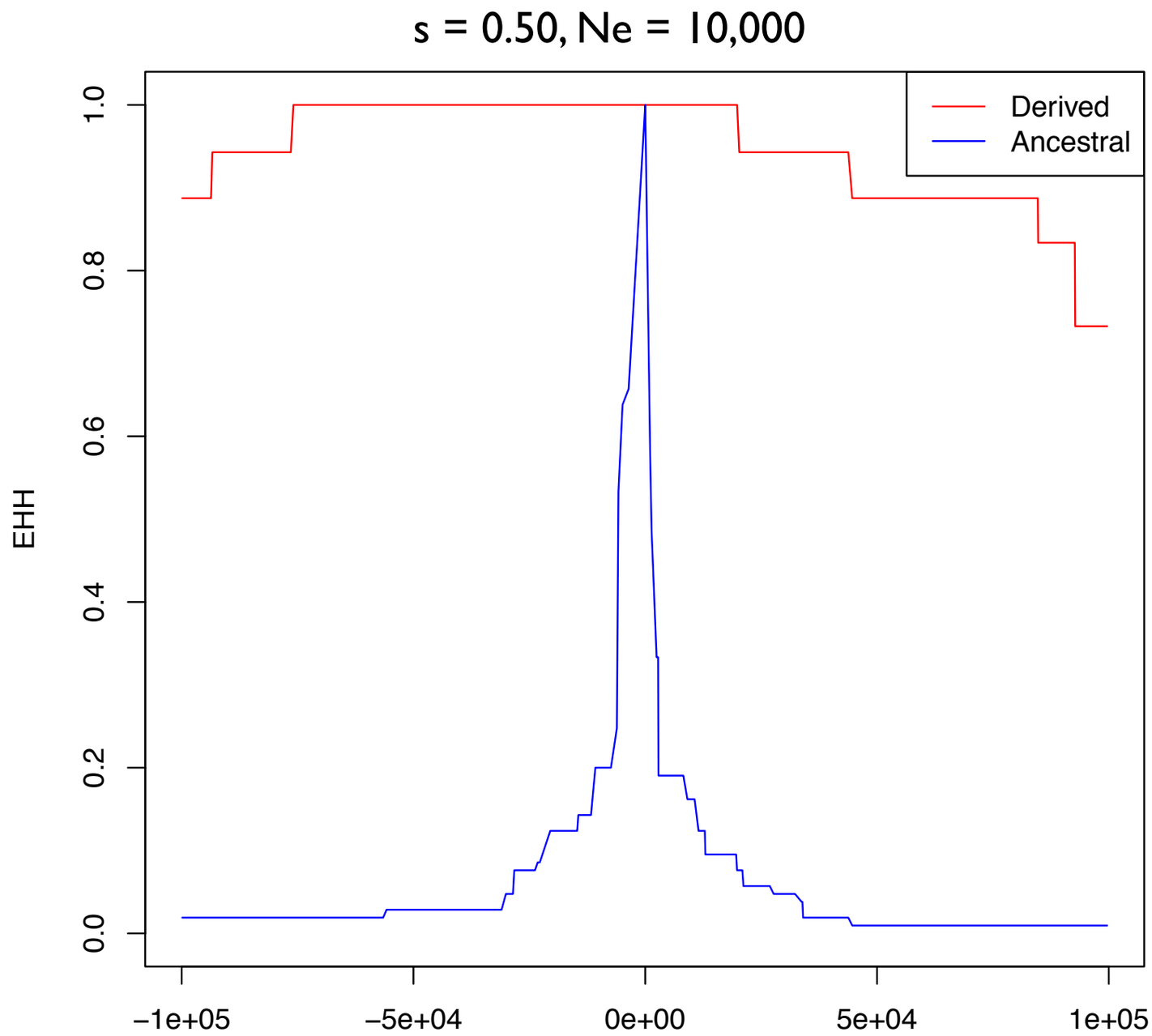
EHH


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
1.0
0.8
0.6
0.4
0.2
0.0
−1e+05
−5e+04
0e+00
5e+04
1e+05
Distance from core (bp)

EHH


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

EHH

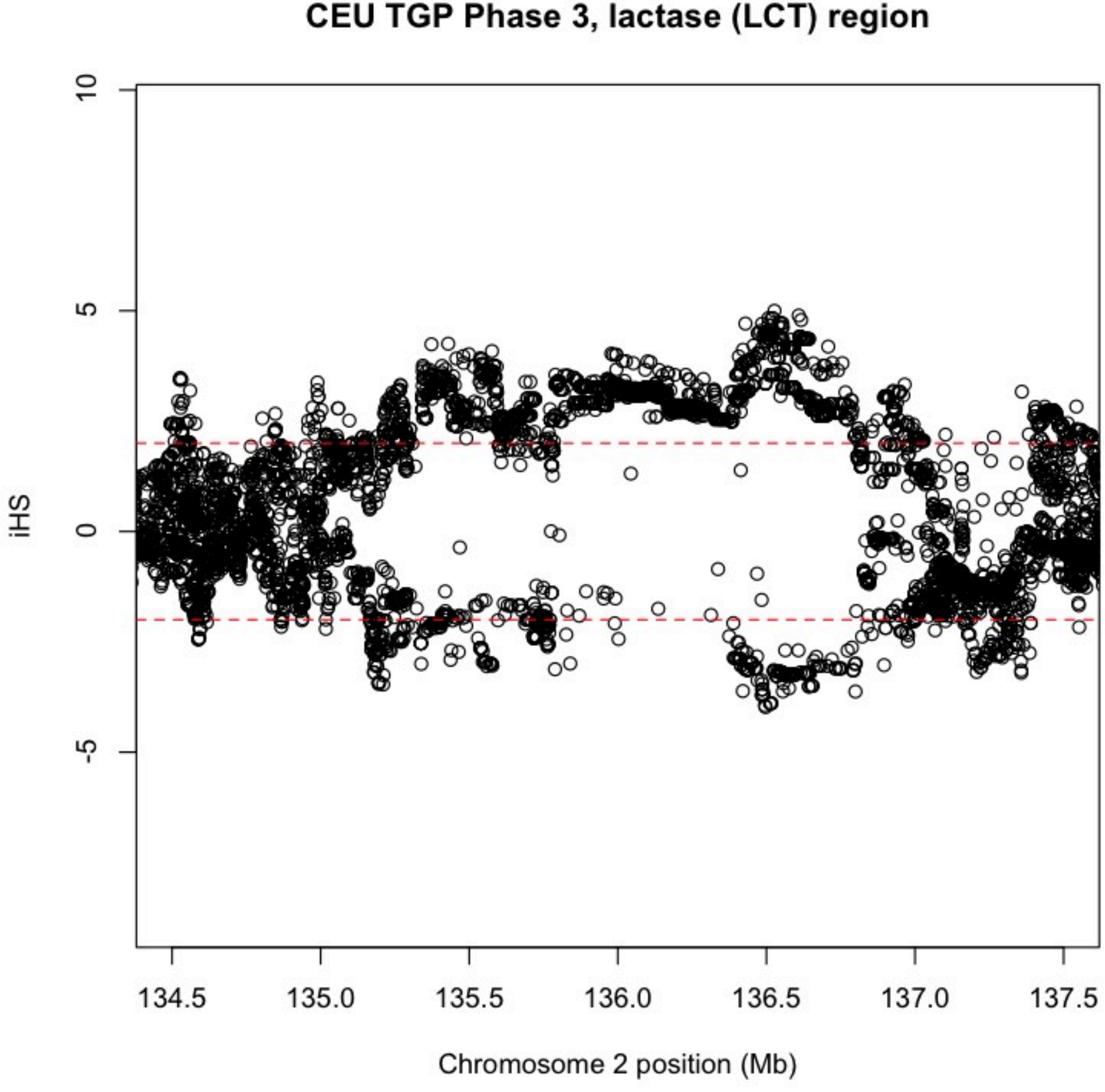
- When querying a specific region of the genome, for each core haplotype, calculate EHH for successively longer surrounding haplotypes.
- Statistical significance is determined by comparing EHH scores to neutral simulations and random control regions of the genome.

Genome-wide scans

- The EHH approach does not lend itself to a genome-wide scan.

- Voight, et al. (2006) create a genome-wide scan statistic based on EHH called integrated Haplotype Score (iHS).

Voight, et al. (2006) *PLoS Biology*


CEU TGP Phase 3, lactase (LCT) region
iHS
10
5
0
-5
134.5
135.0
135.5
136.0
136.5
137.0
137.5
Chromosome 2 position (Mb)

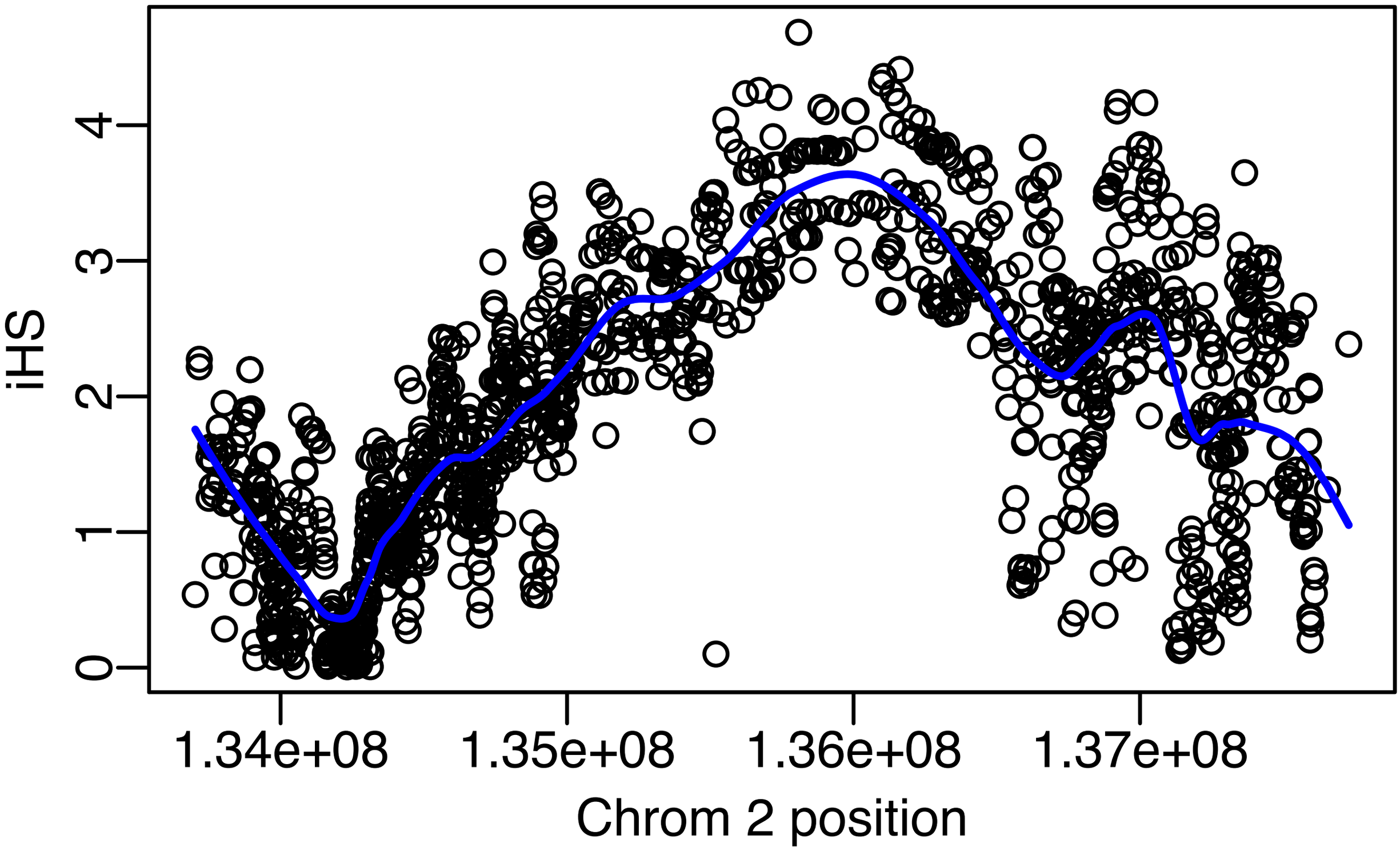
Caveats

- Power may be overstated.
  - If a large proportion of the genome is non-neutral, we lose power to detect the weakest selected variants because of genome-wide normalization.
- iHS no formal test to decide significance.
  - Take top 1% of signals
- XP-EHH more sensitive to demographics
  - i.e. comparing populations with serial bottlenecks separating them
- Important to combine *multiple lines* of evidence!

Breakout Groups!!
Running iHS with `selscan`

- Open up your command prompt (i.e., rev your engines)
- Let's give iHS a go!
- Let's consider the LCT gene.
- Follow commands in Day3_AM2_CMD.txt
    - You will need `selscan.zip`
    - In terminal run:
        - `selscan` …
        - `norm` …
    - Plot it in R!

iHS



Other populations??

- Now run selscan on the YRI population
- YRI is a sample of individuals from Yoruba, Nigeria, where they do not have a long tradition of domesticating cows.
- Update the selscan commands by replacing "CEU" with "YRI"
- Breakout Groups!!

Do you think there is selection in this region in the YRI population?

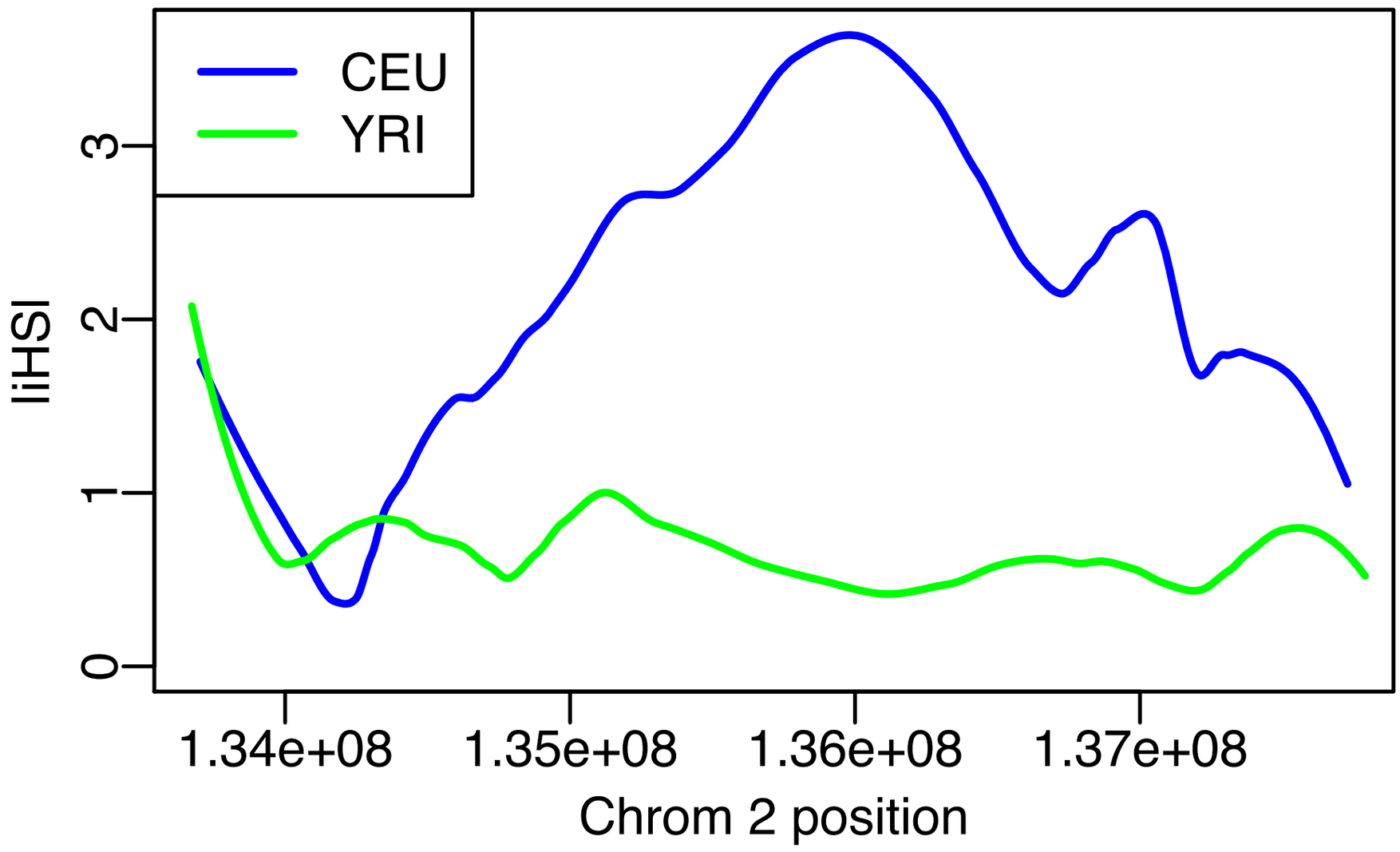
Yes!

No!

Unclear

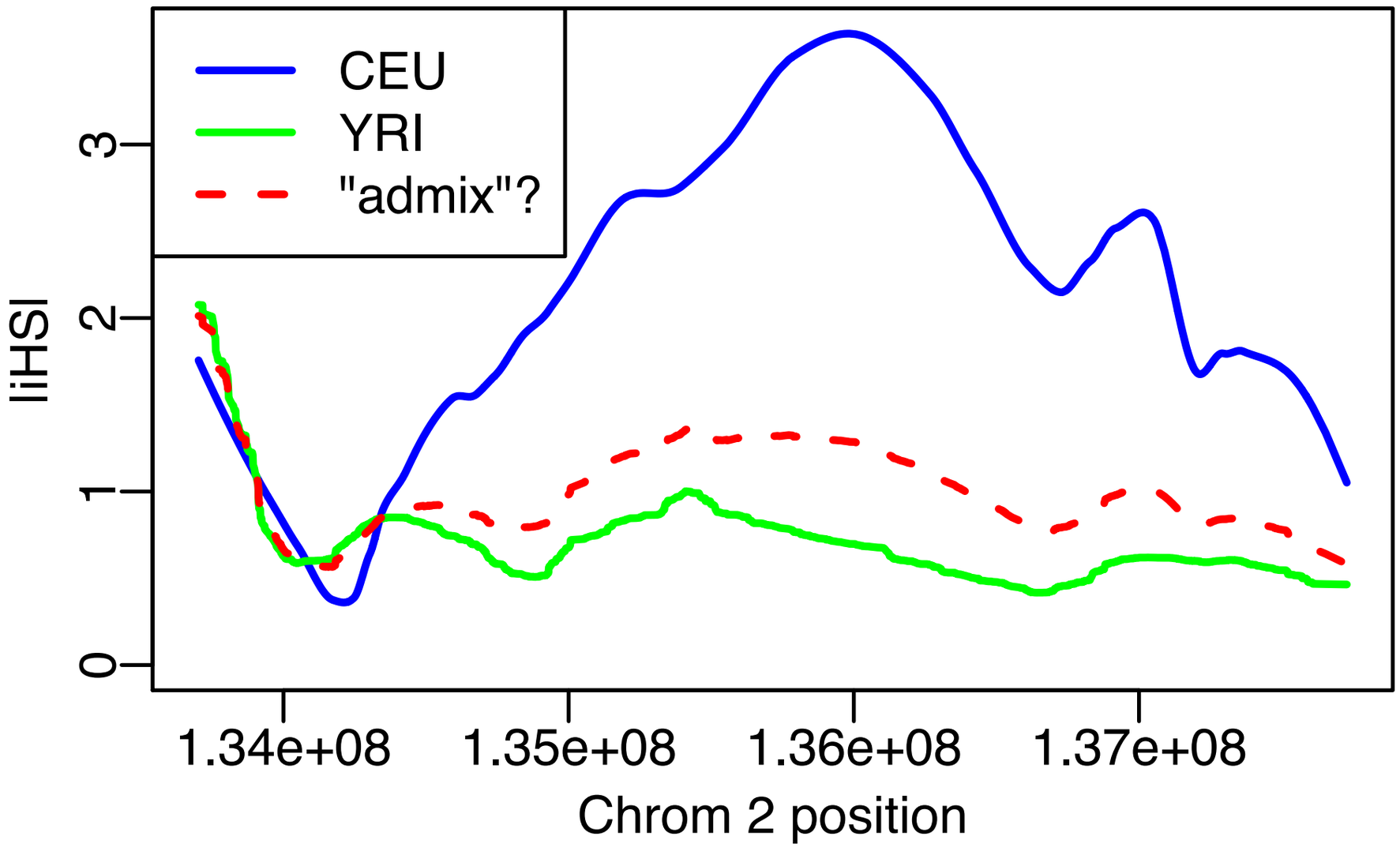
Other populations??

- "CEU" vs "YRI"



What about admixture?

- African American genomes contain admixture with African ancestry (~80%) and European ancestry (~20%).
- ASW is one sample of African Americans (from the Southwest)
- One guess might be that it should be intermediate

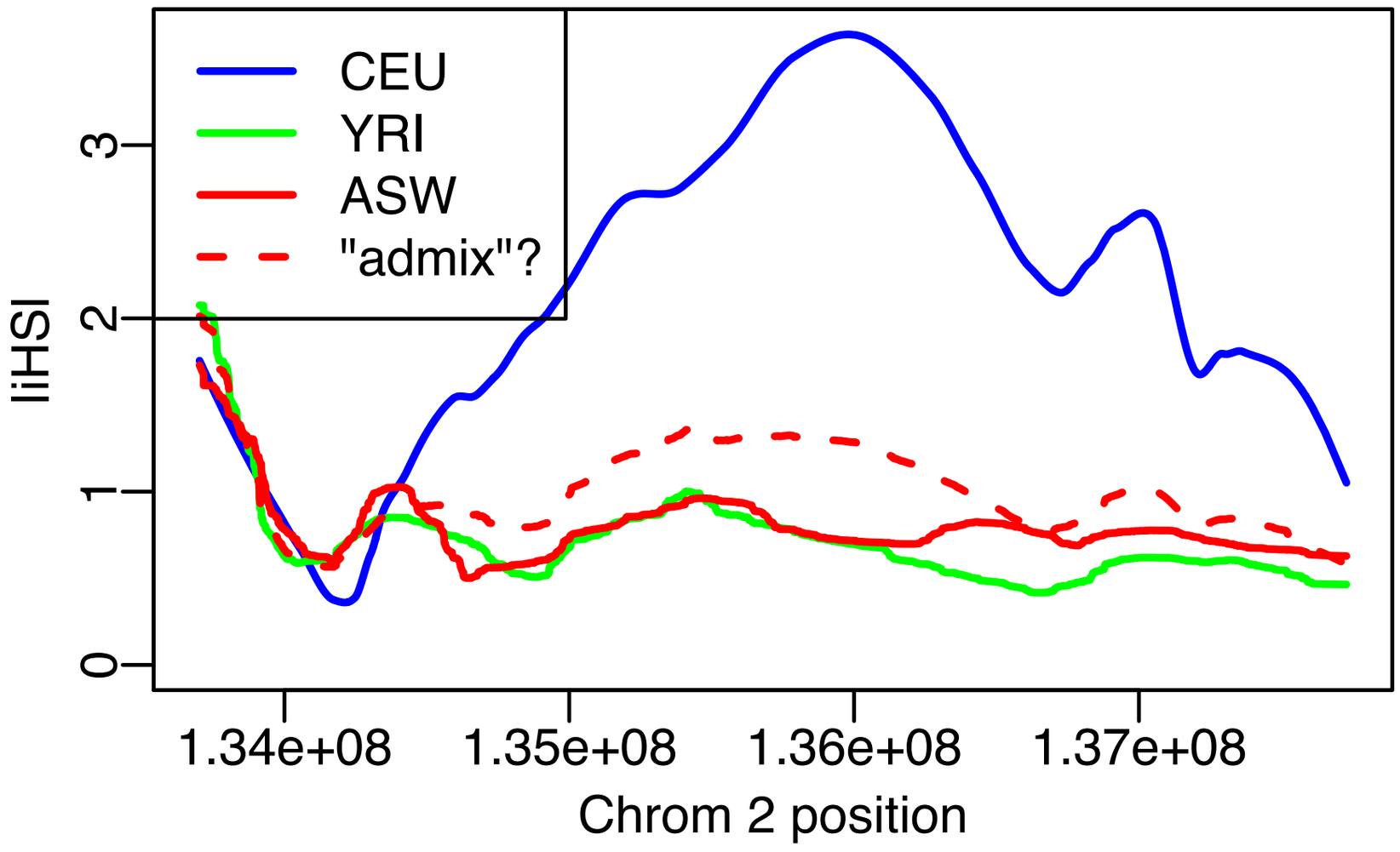


Other populations??

- Now run selscan on the ASW population
- Update the selscan command by replacing "CEU" with "ASW"
- Breakout groups!!

Other populations??

- Now run selscan on the ASW population
- Update the selscan command by replacing "CEU" with "ASW"
- In these data, ASW is much more similar to YRI than "expected".



Summary

- iHS is one example of a statistic geared toward detecting a "classic sweep".
- It is based on the idea that a new mutation has been selected, and quickly spread through the population.
- selscan is one piece of software that can run many different selection statistics in an efficient manner.

selscan

- Open your terminal/command prompt!
- Change to the new `selscan` directory
- For example:
    - `cd ~/Desktop/selscan/`
- There should 4 subdirectories:
    - `rhernandez$ ls`
      `data linux osx win`
- Change Directory to where the data are:
    - `cd data`

selscan

- All the commands we are running can be found in the selscan_CMD.txt file.
- Copy the appropriate executable to the data directory:
- **osx:**
    - `cp ../osx/selscan .`
- **linux:**
    - `cp ../linux/selscan .`
- **Windows:**
    - `copy ..\win\selscan.exe .`

selscan

- Test that it works:
    - **osx/linux:** `./selscan` (**Win:** `selscan.exe`)

```
selscan v1.1.0b
ERROR: Must specify one and only one of
EHH (—ehh)
iHS (--ihs)
XP-EHH (--xpehh)
PI (--pi)
nSL (--nsl)
```

selscan

- iHS requires 2 files, a **map** file and a **hap** file.
  - `--map <string>: A mapfile with one row per variant site.`
    - `Formatted with 4 columns:`
      - `<chr#> <locusID> <genetic pos> <physical pos>`
  - `--hap <string>: A hapfile with one row per haplotype, and one column per variant. Variants should be coded 0/1.`

selscan

- Now run it!
  - All in one line type:
    - `./selscan` (Win: `selscan.exe`)

```
--ihs
--map CEU.chr2.map
--hap CEU.chr2.ihshap
--out CEU.chr2
```

```
selscan v1.1.0b
Opening ../data/CEU.chr2.hap...
Loading 224 haplotypes and 1971 loci...
Opening ../data/CEU.chr2.map...
Loading map data for 1971 loci
--skip-low-freq set. Removing all variants < 0.05.
Removed 359 low frequency variants.
Starting iHS calculations with alt flag not set.
|=================================>                |
```

Normalize

- All in one line type:
  - `./norm`

```
--ihs
--files CEU.chr2.ihs.out bg.ihs.out
```

```
norm v1.1.0aYou have provided 2 output files for joint
normalization.
Opened ../data/CEU.chr2.ihs.out
Opened ../data/bg.ihs.out

Total loci: 666285
Reading all frequency and iHS data.
Calculating mean and variance per frequency bin:
```

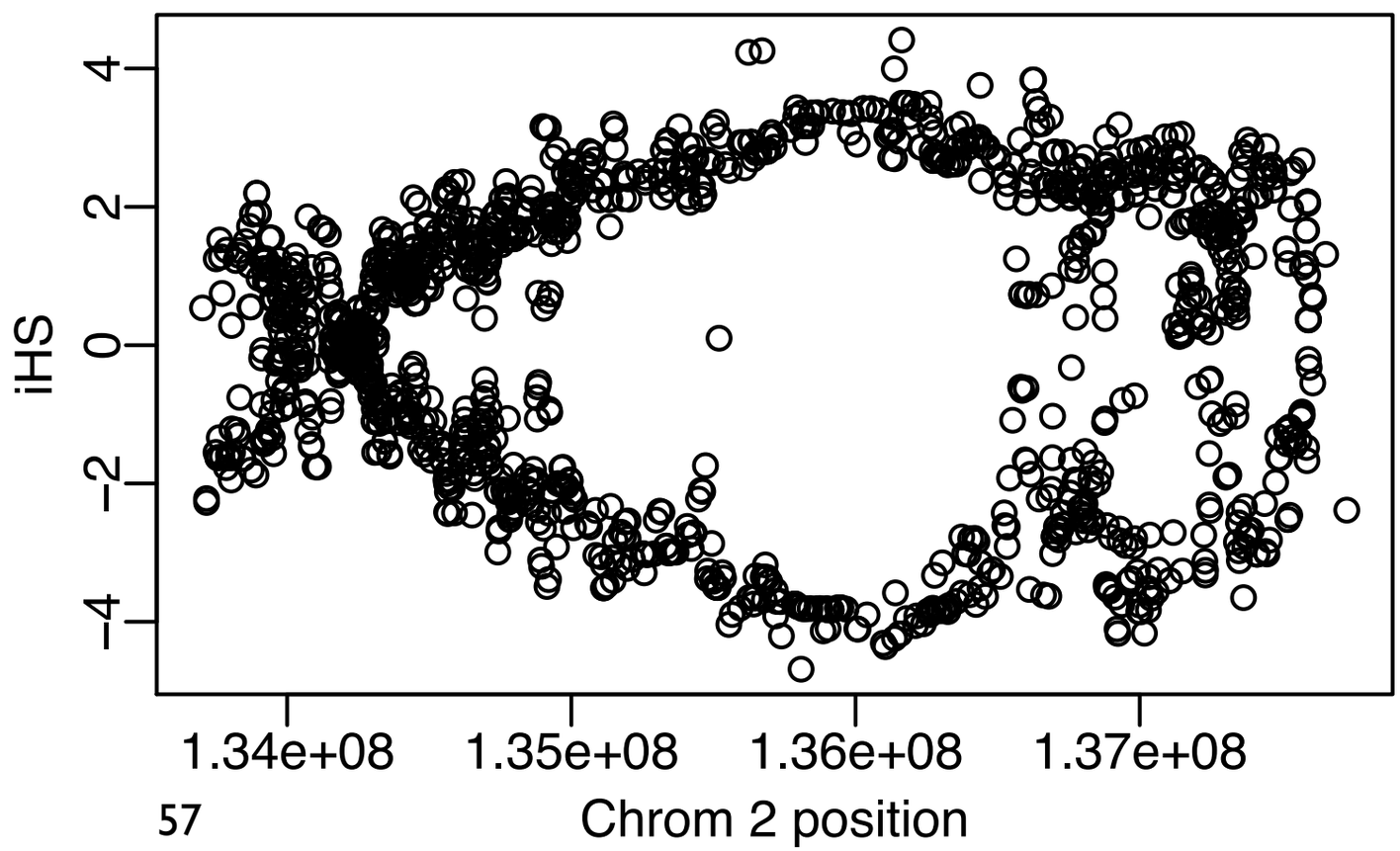
iHS

- Now let's plot it!
- Open R.
- Read in data for CEU:

```
setwd("cd ~/Desktop/selscan/data“)

CEU=read.table(“CEU.chr2.ihs.out.100bins.norm”)

plot(CEU[,2], CEU[,7])
```



iHS

- Often analyze absolute value, and smooth it out.
- My preferred method for smoothing is using loess

```
SP=0.2 #this is the span, a parameter you can change (higher = more
smoothing)

CEU.x=CEU[,2]; #the x-coordinates in Mb

y=abs(CEU[,7]) #iHS is actually the absolute value

CEU.loess=loess(y~CEU.x,span=SP,data.frame(x=CEU.x,y=y)); #step 1

CEU.predict=predict(CEU.loess,data.frame(x=CEU.x)); #step 2


plot(CEU[,2], abs(CEU[,7]))

lines(CEU.x, CEU.predict, lwd=2, col='blue')
```