Searching for Signatures of Selection with Selscan

Ryan Hernandez

ryan.hernandez@ucsf.edu

The Effect of Positive Selection

Adaptive

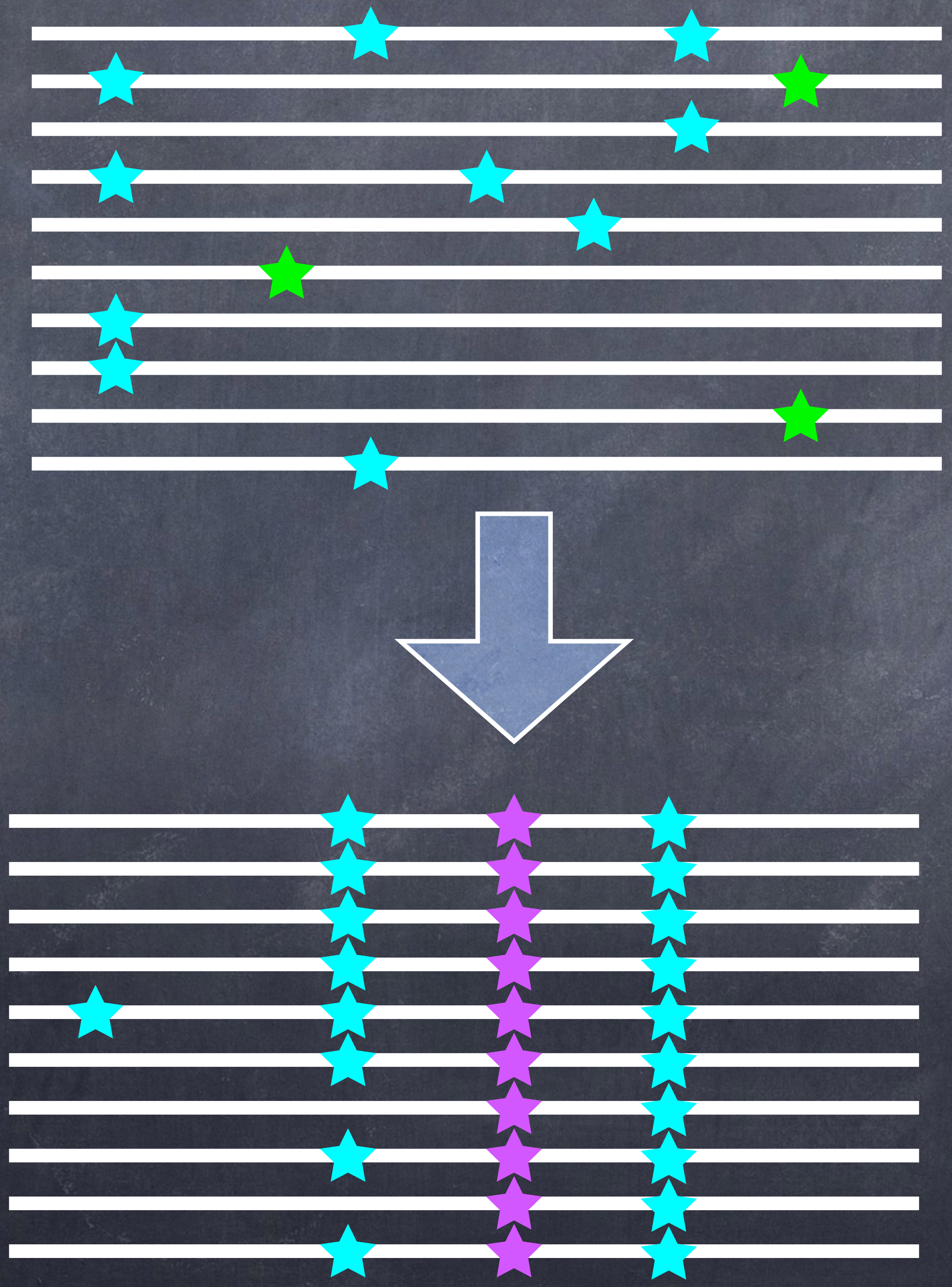
Neutral

Nearly Neutral

Mildly Deleterious

Fairly Deleterious

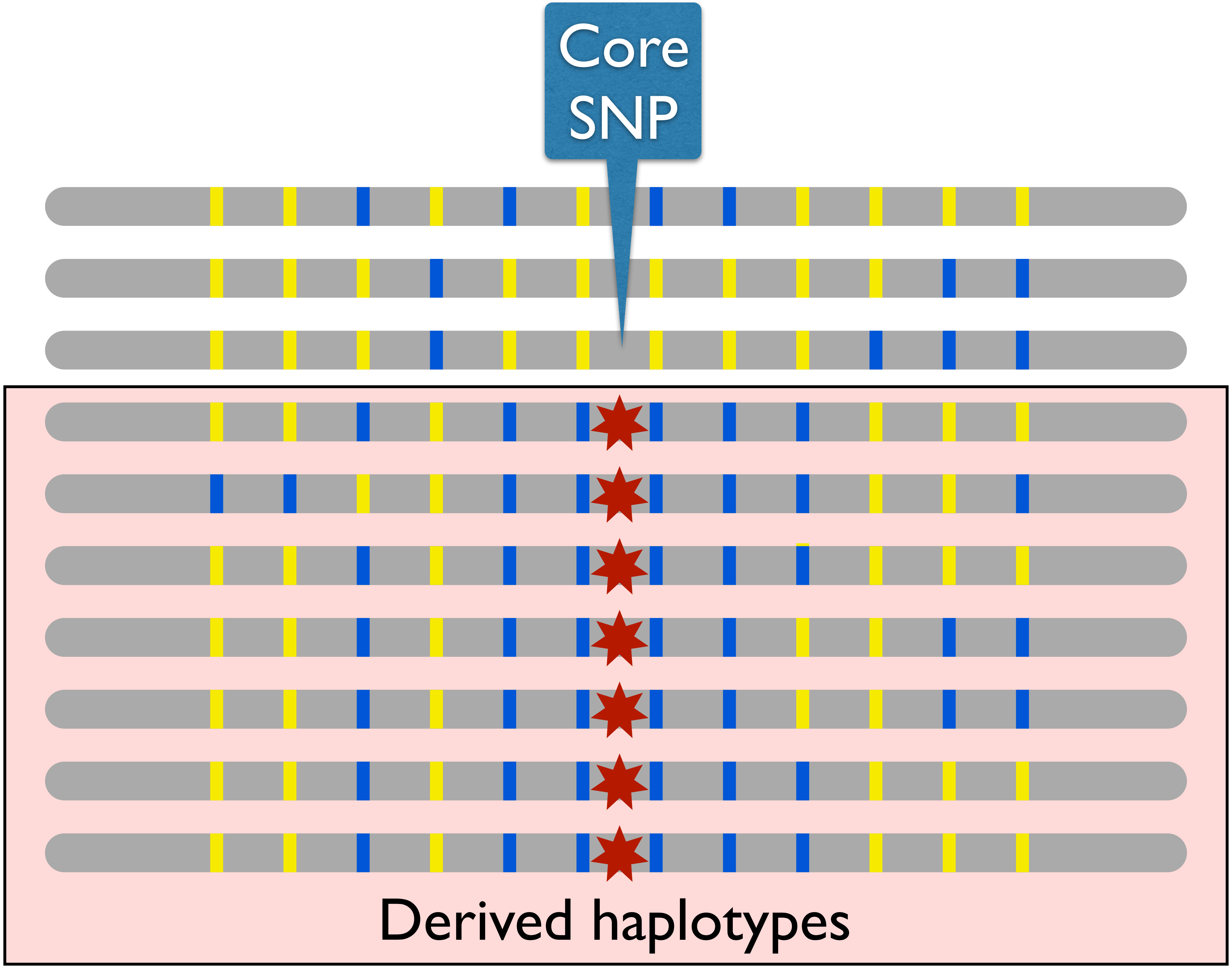
Strongly Deleterious



How do we capture this process in a statistic?

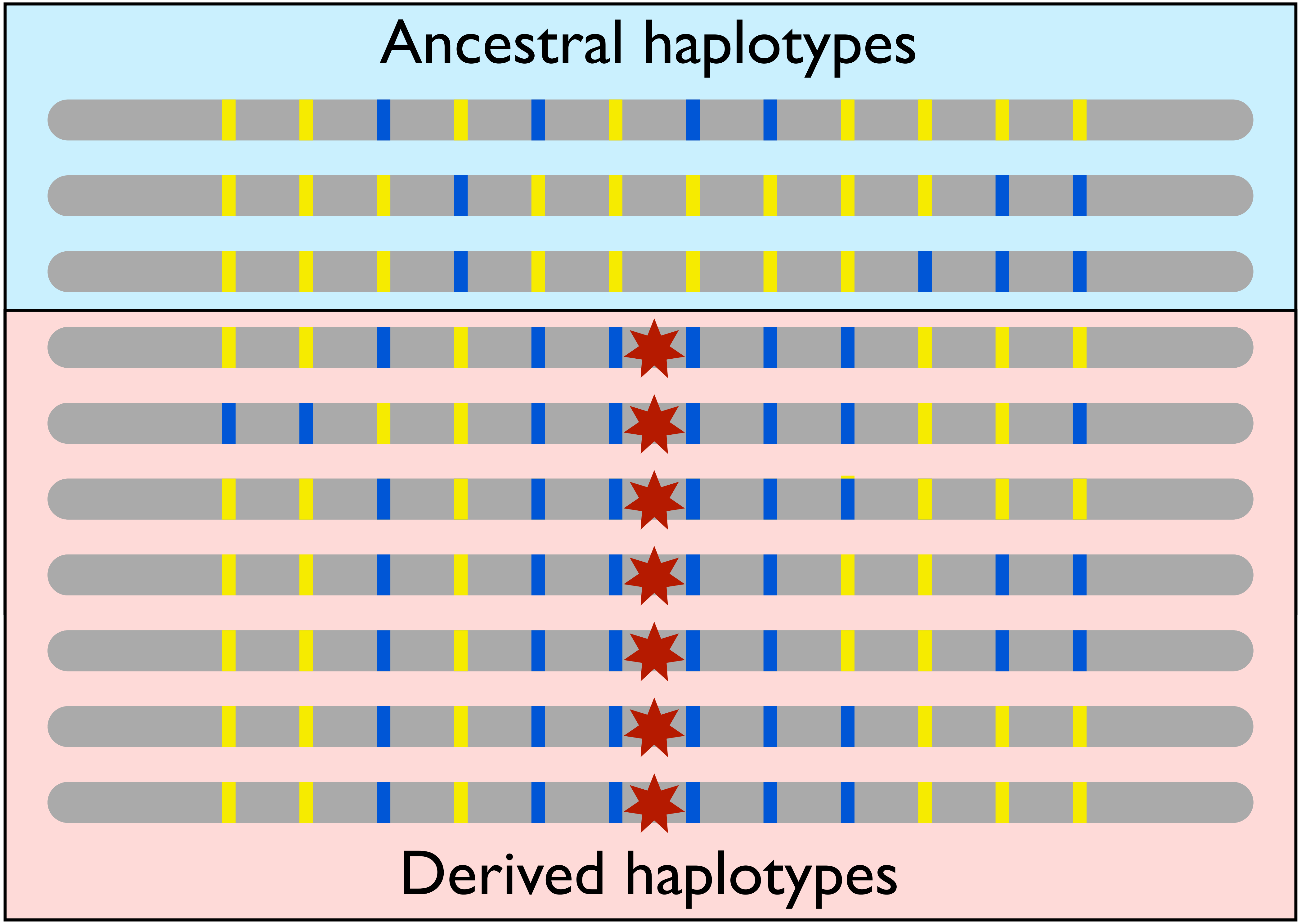
Extended Haplotype Homozygosity

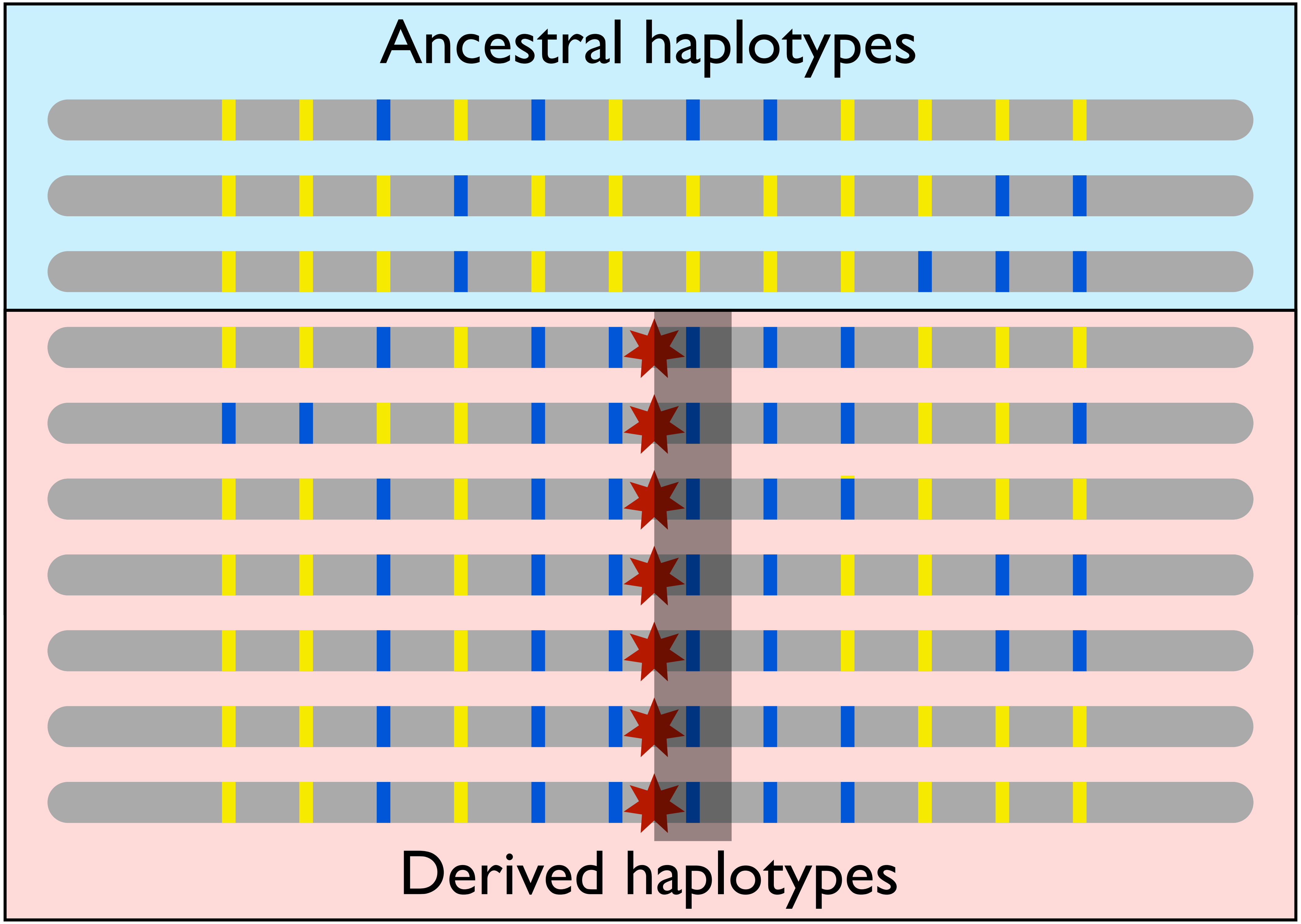
- Sabeti, et al. (*Nature*, 2002) proposed EHH
- Designed to track the decay of haplotype identity away from a locus of interest
    - If selection acts quickly enough
- Originally derives from ideas in Hudson, et al. (*Genetics*, 1994).

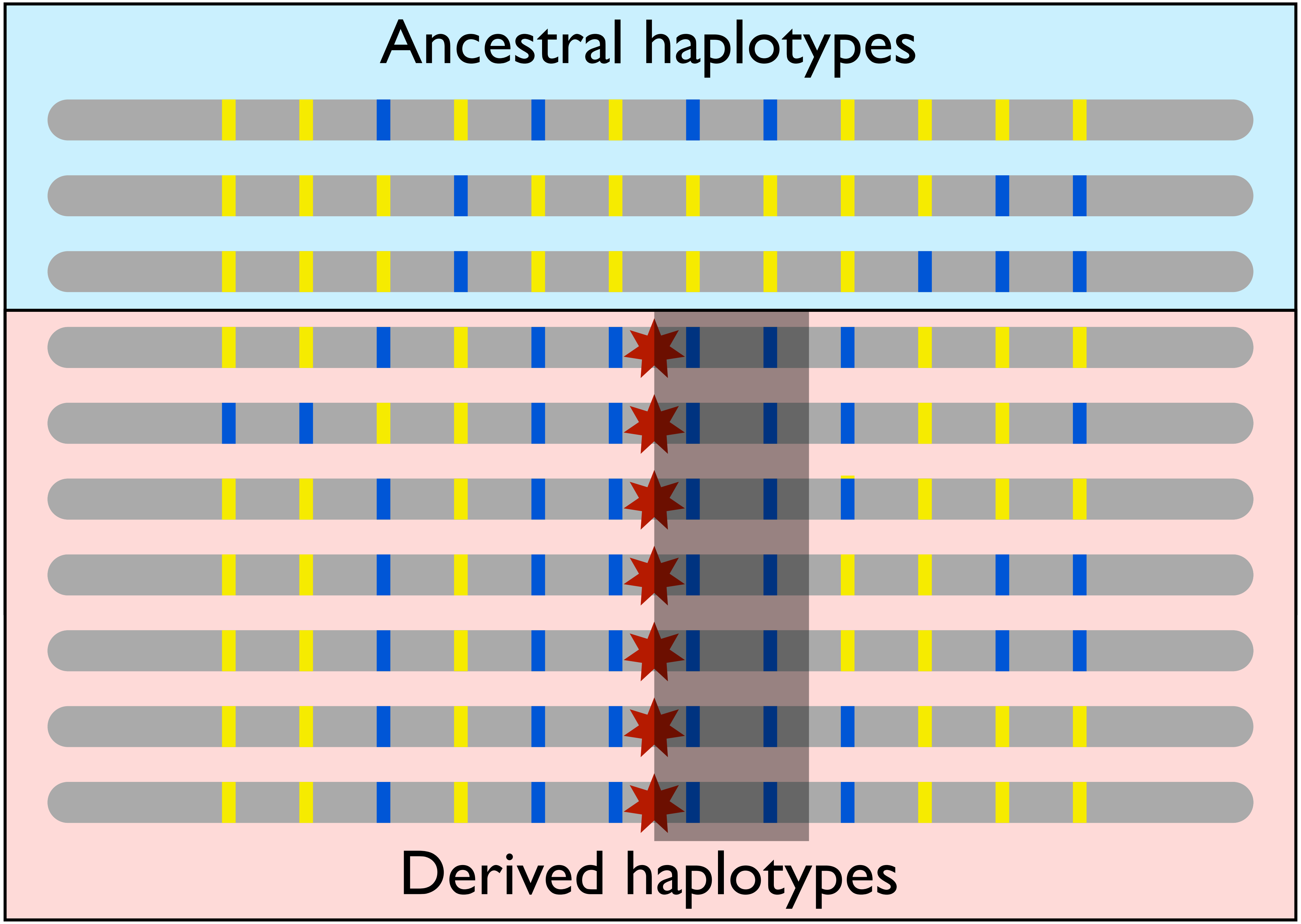


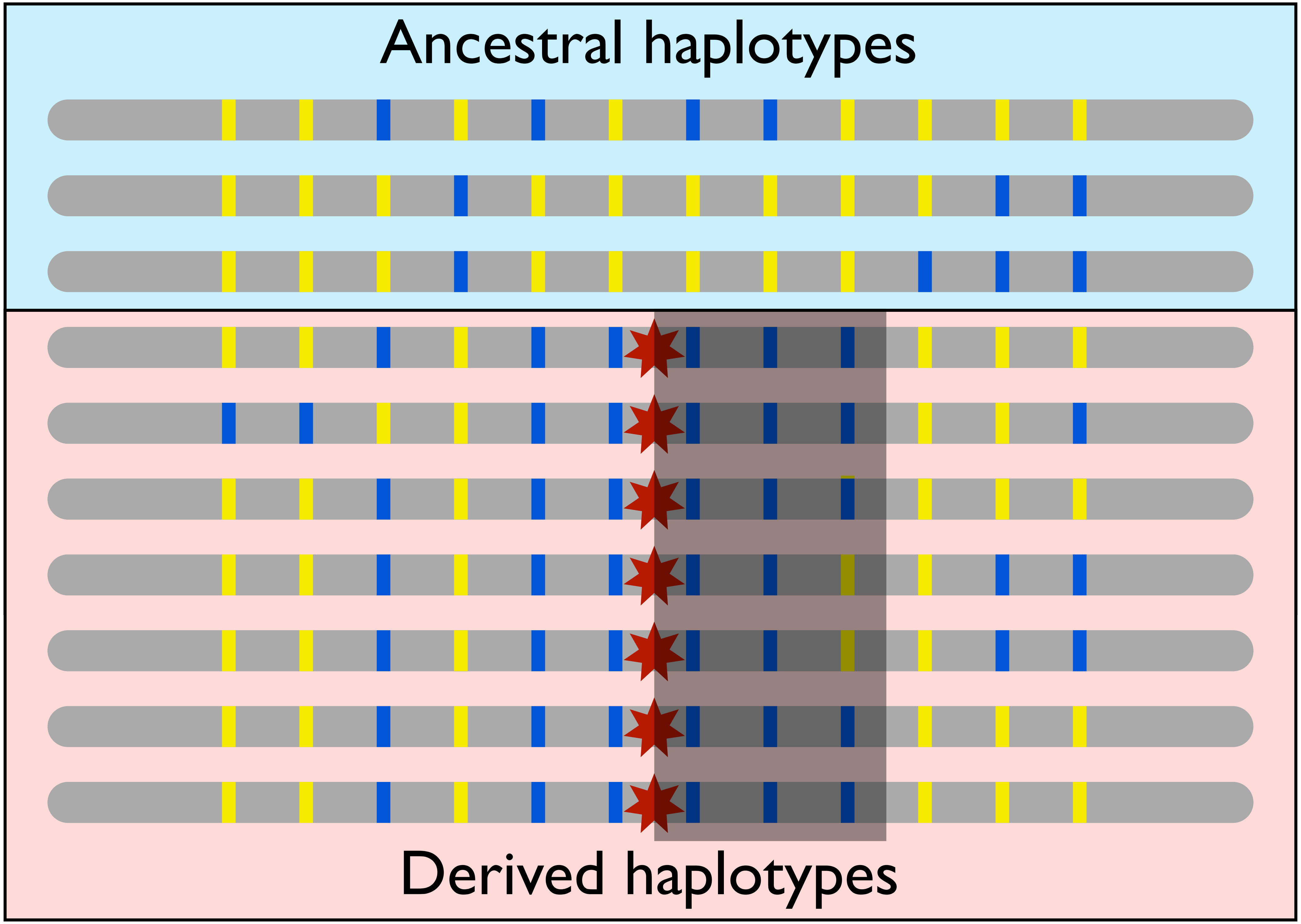
Core SNP

Derived haplotypes


Ancestral haplotypes
Derived haplotypes

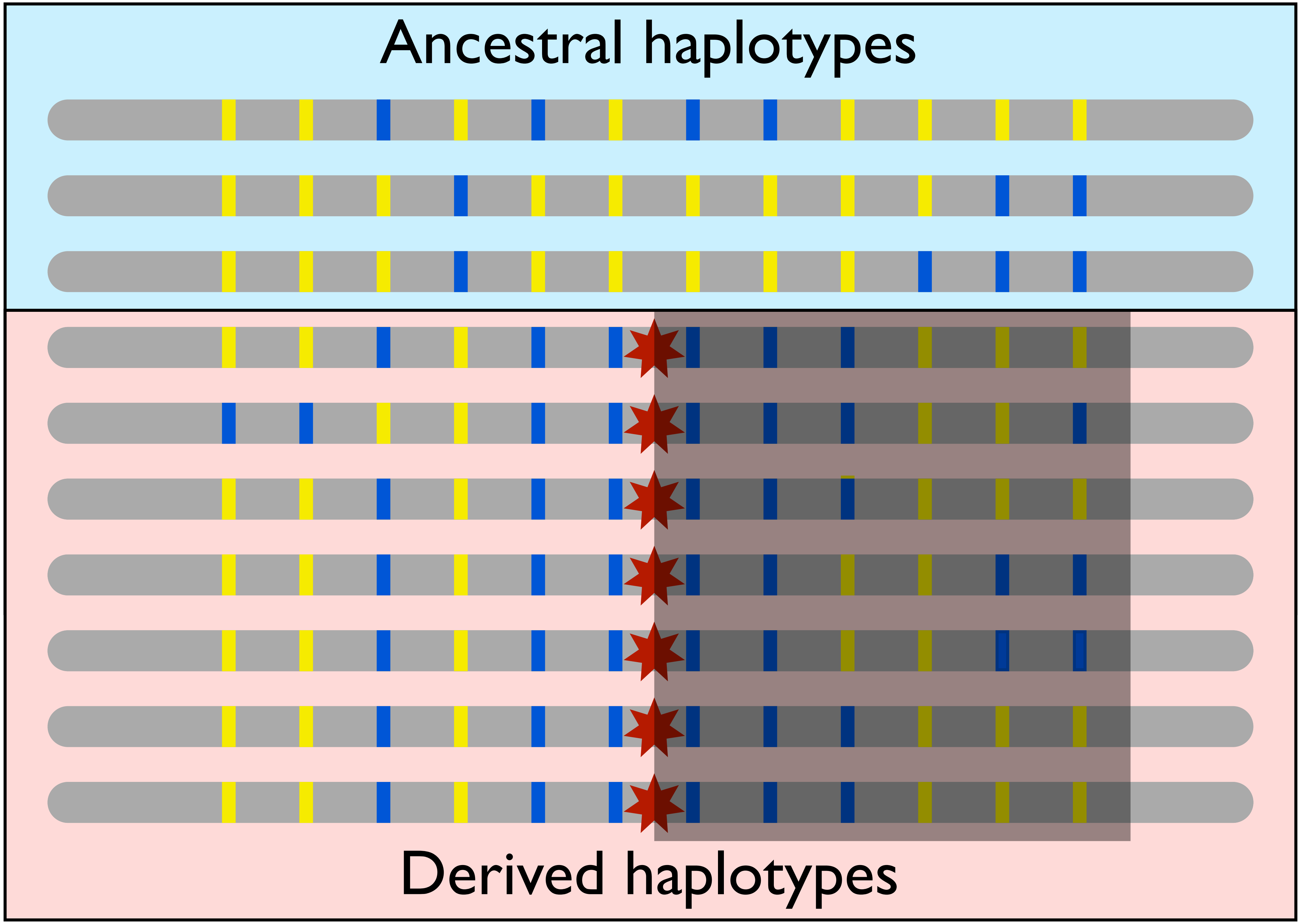

Ancestral haplotypes
Derived haplotypes


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 $\mathtt{c}$ out to marker $\mathtt{x}$, then

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

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

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

$$EHH_c(x) = \sum_{h \in \mathcal{H}(c,x)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}} \quad \longleftarrow \text{\# of ways to choose two } \mathtt{h} \text{ haplotypes}$$

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

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

Calculating EHH

- If $EHH_c(x)$ is the extended haplotype homozygosity of the core haplotype 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

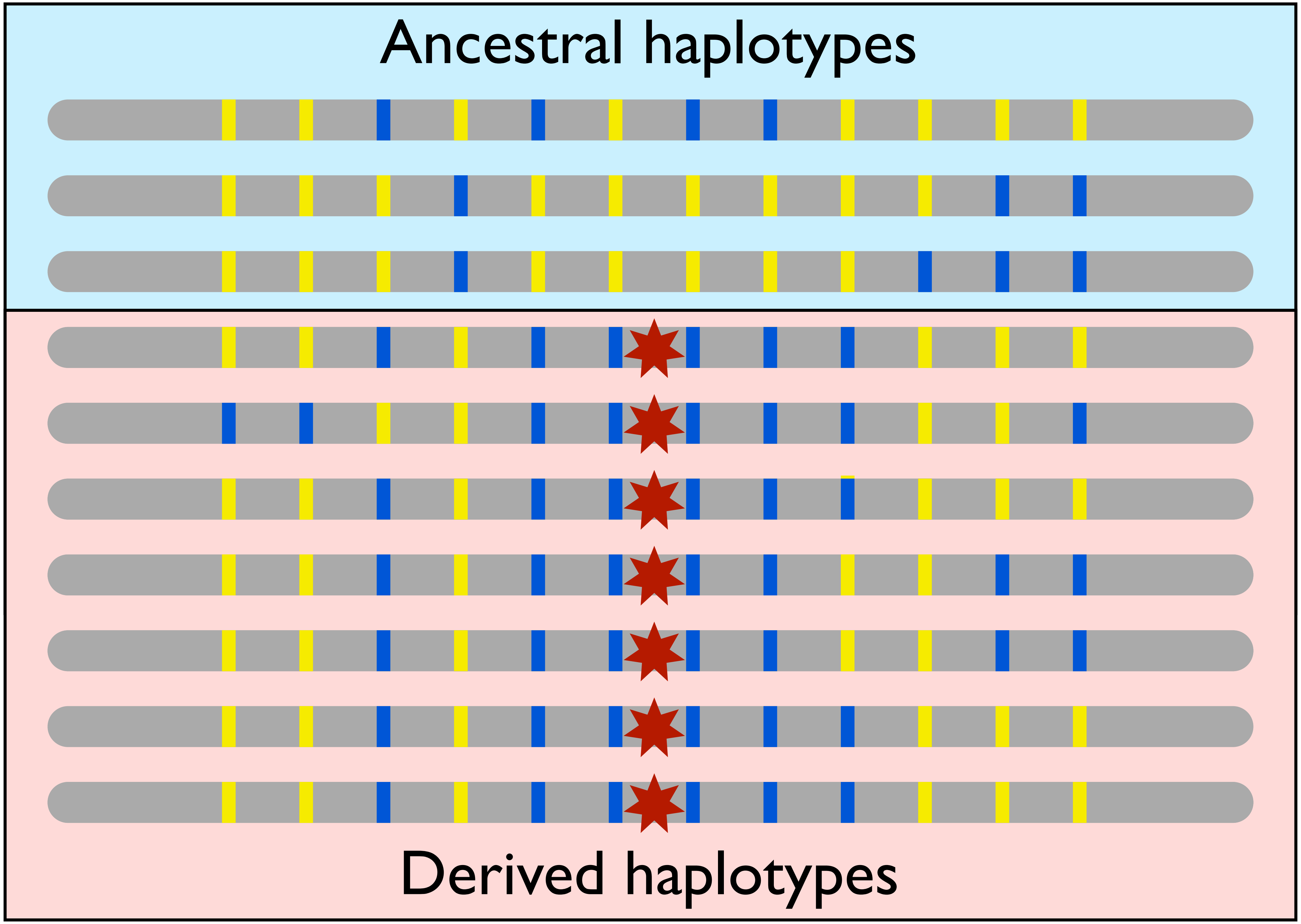
# of ways to choose two core haplotypes

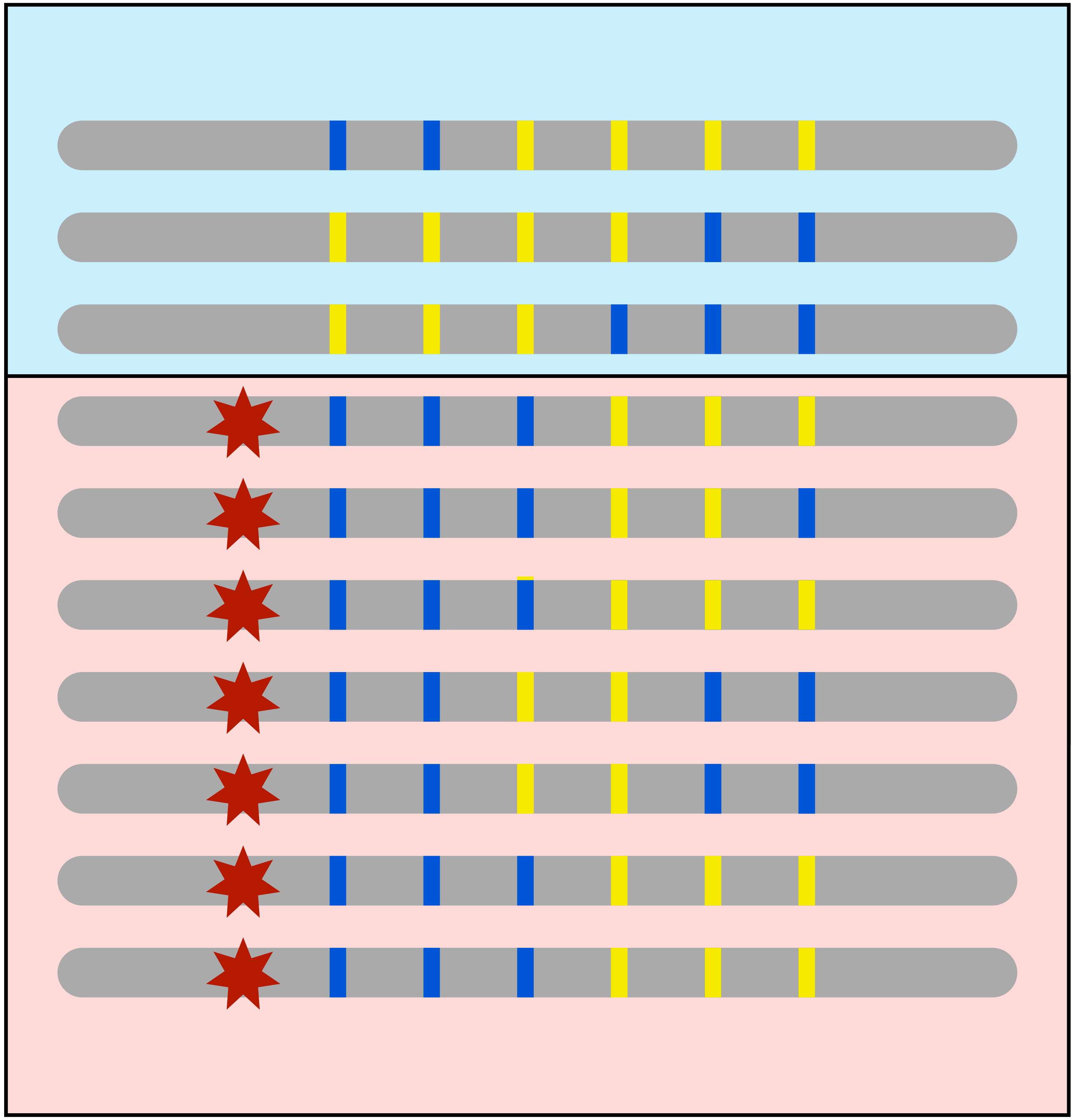
$$\binom{n}{2} := 0 \ \ \forall n < 2$$

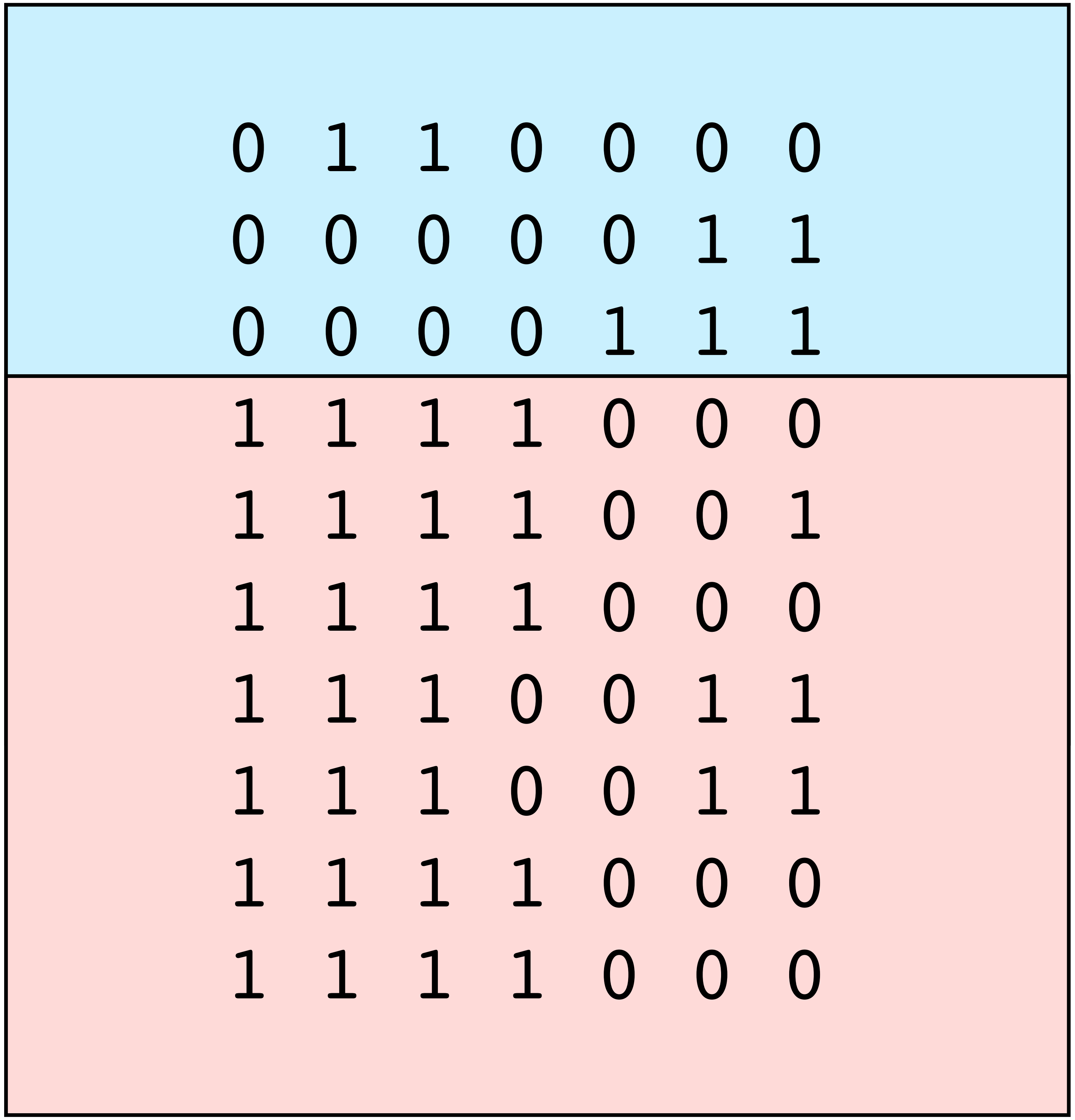
Szpiech and Hernandez (2014) *Molecular Biology and Evolution*

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

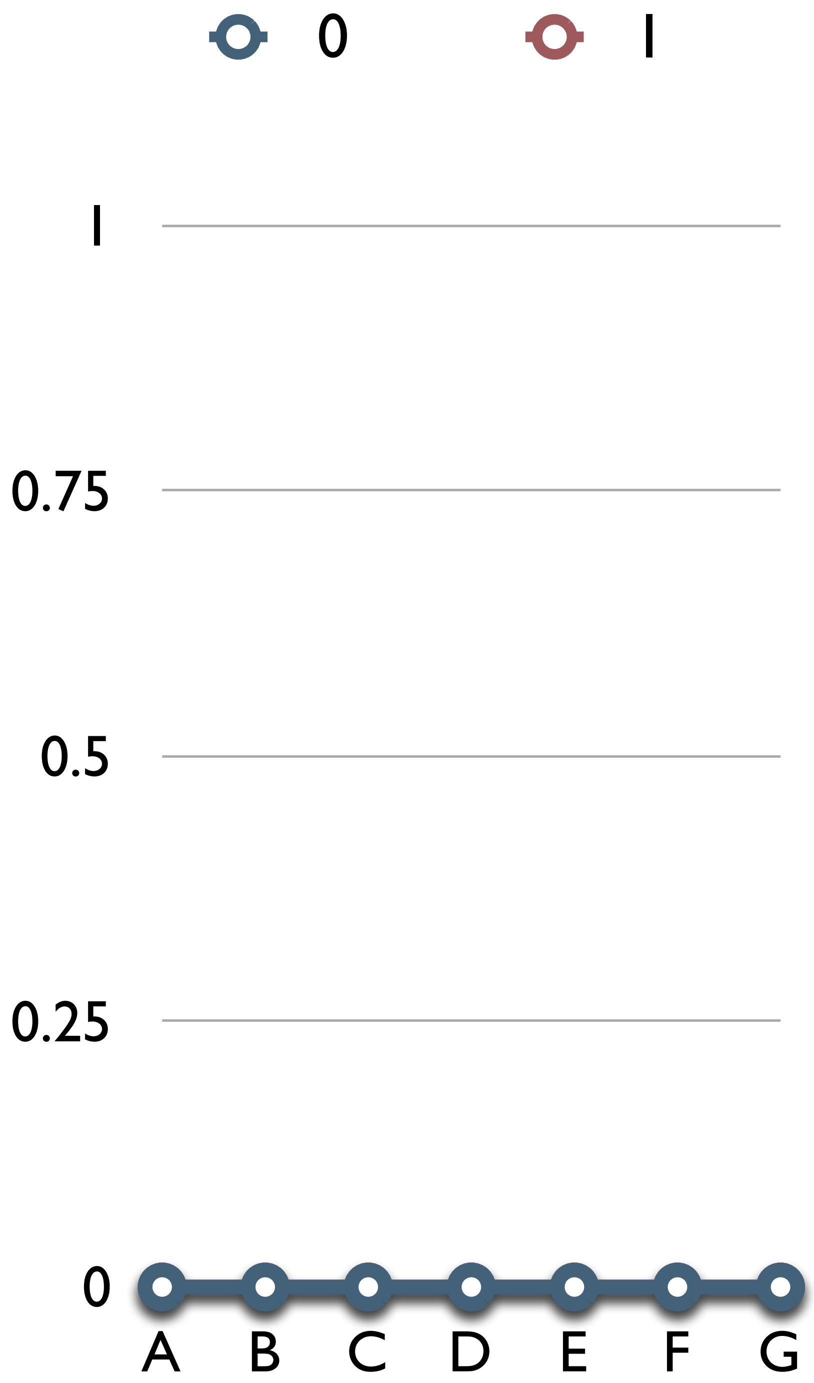




0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

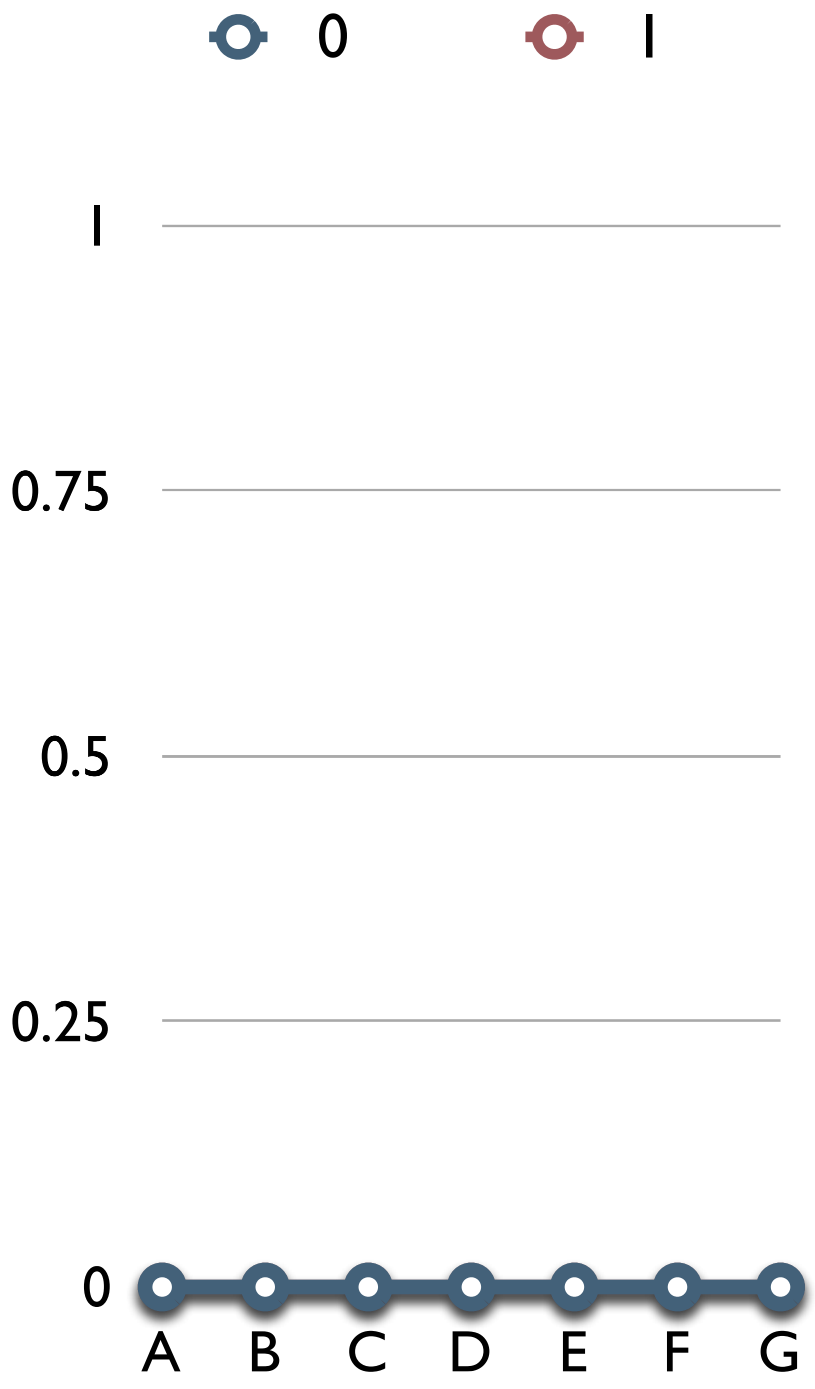
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

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

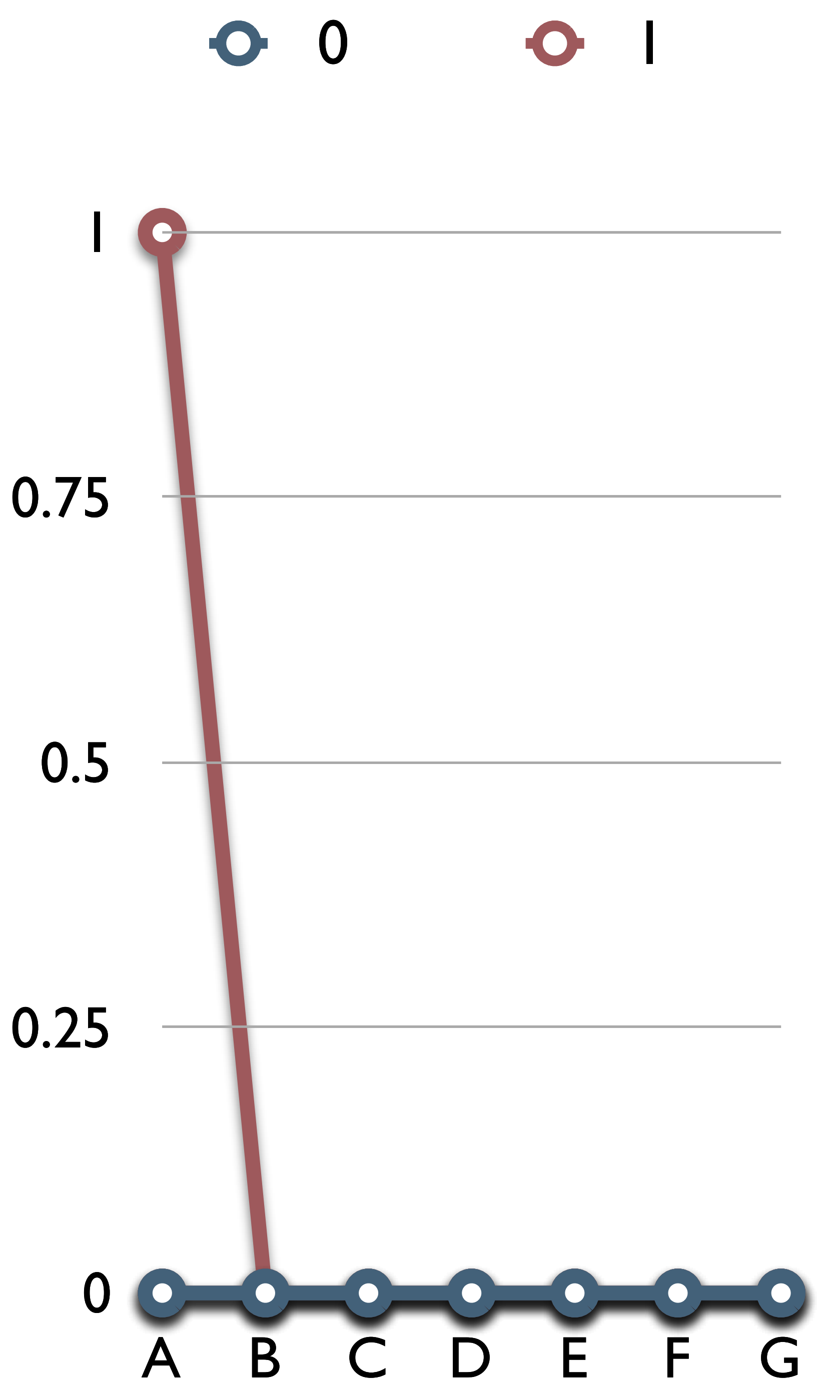
$$n_c = n_1 = 7$$



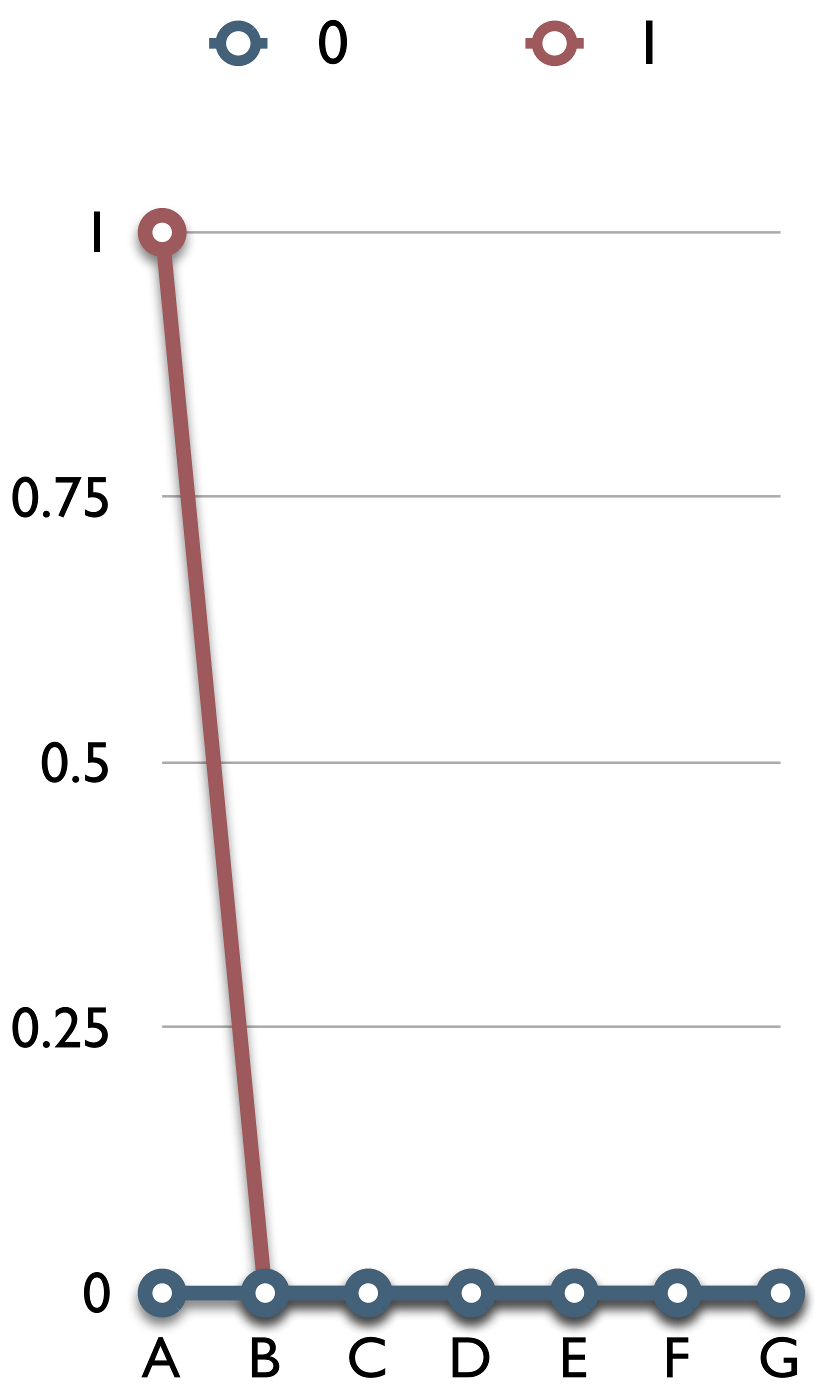
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_c = n_1 = 7$$

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

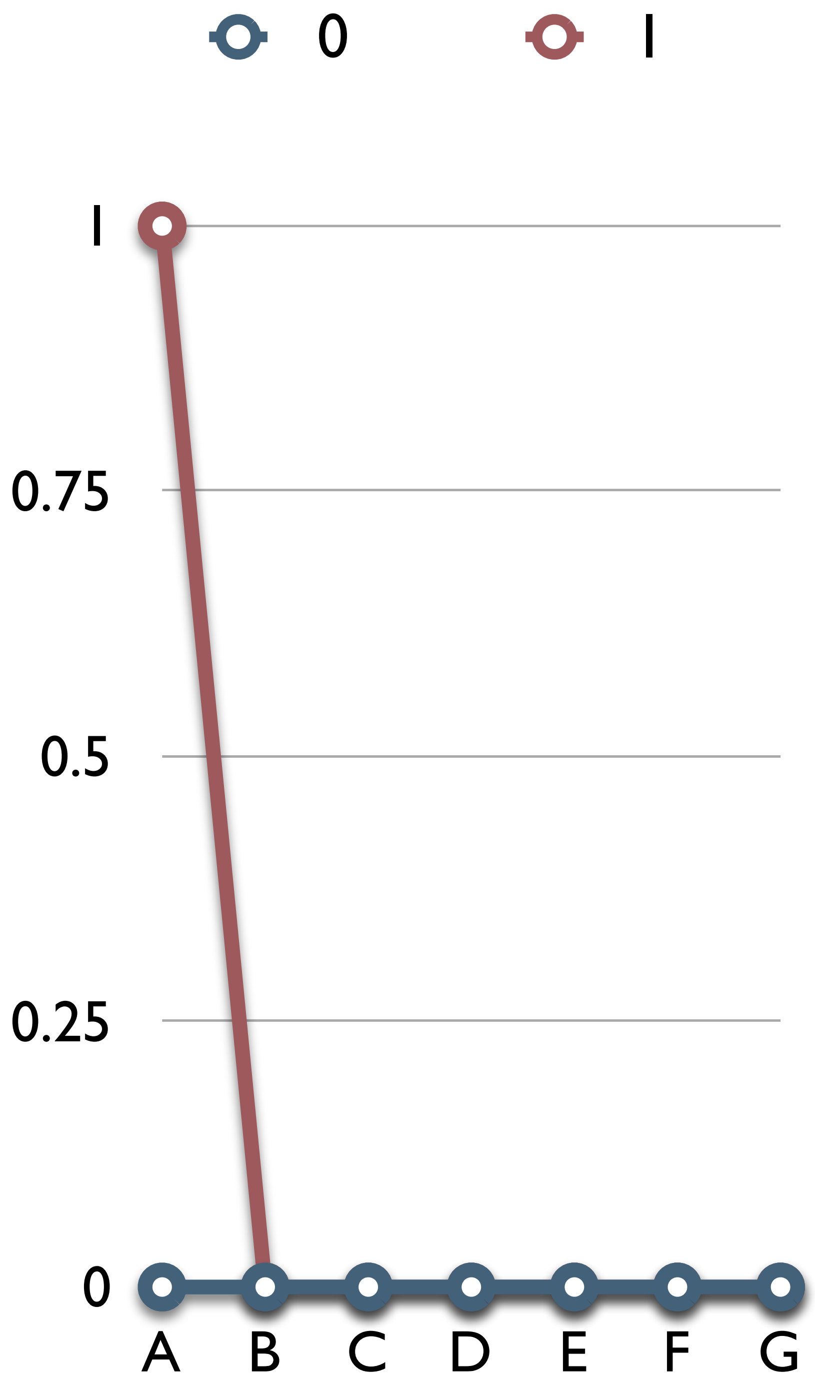


A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0



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

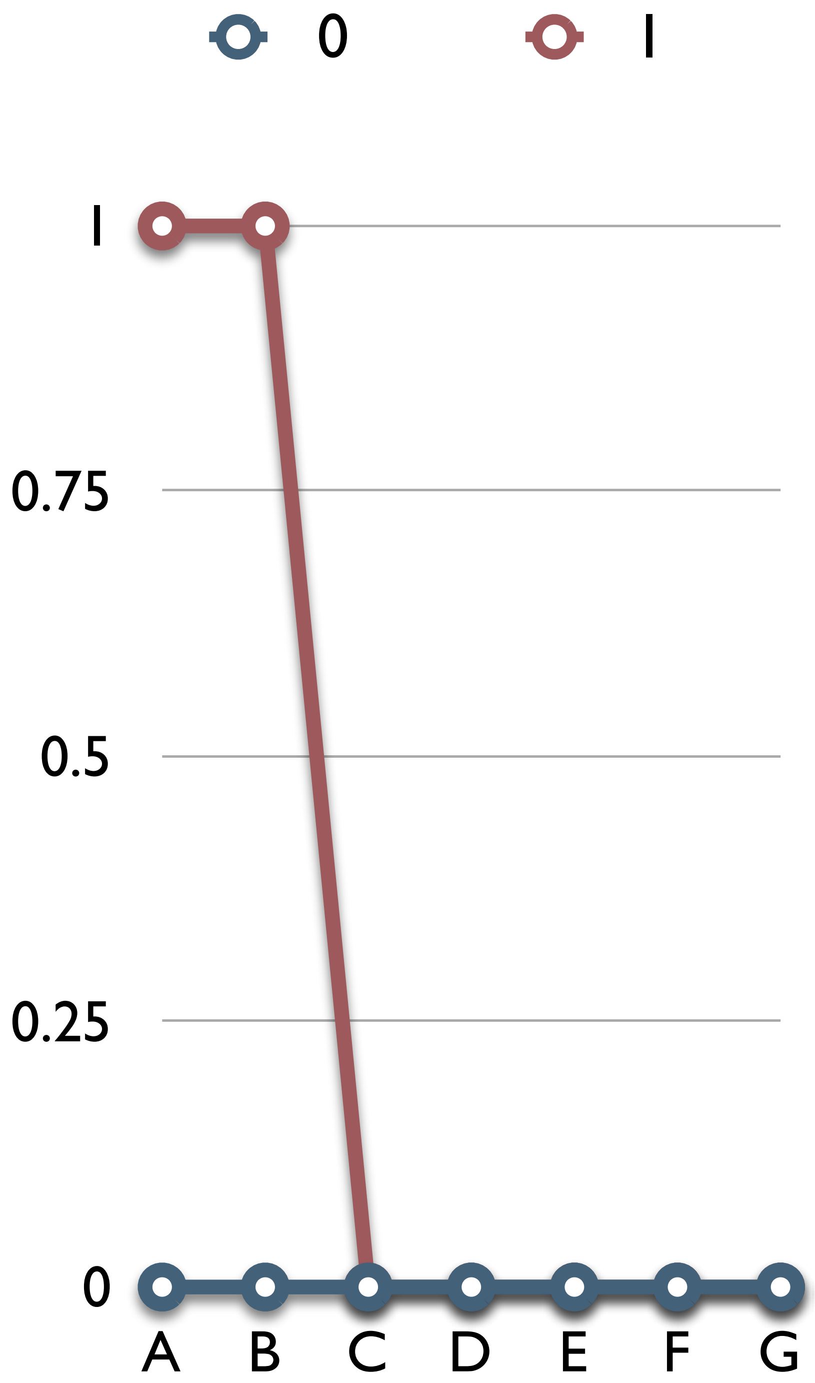
$$n_{11} = 7$$



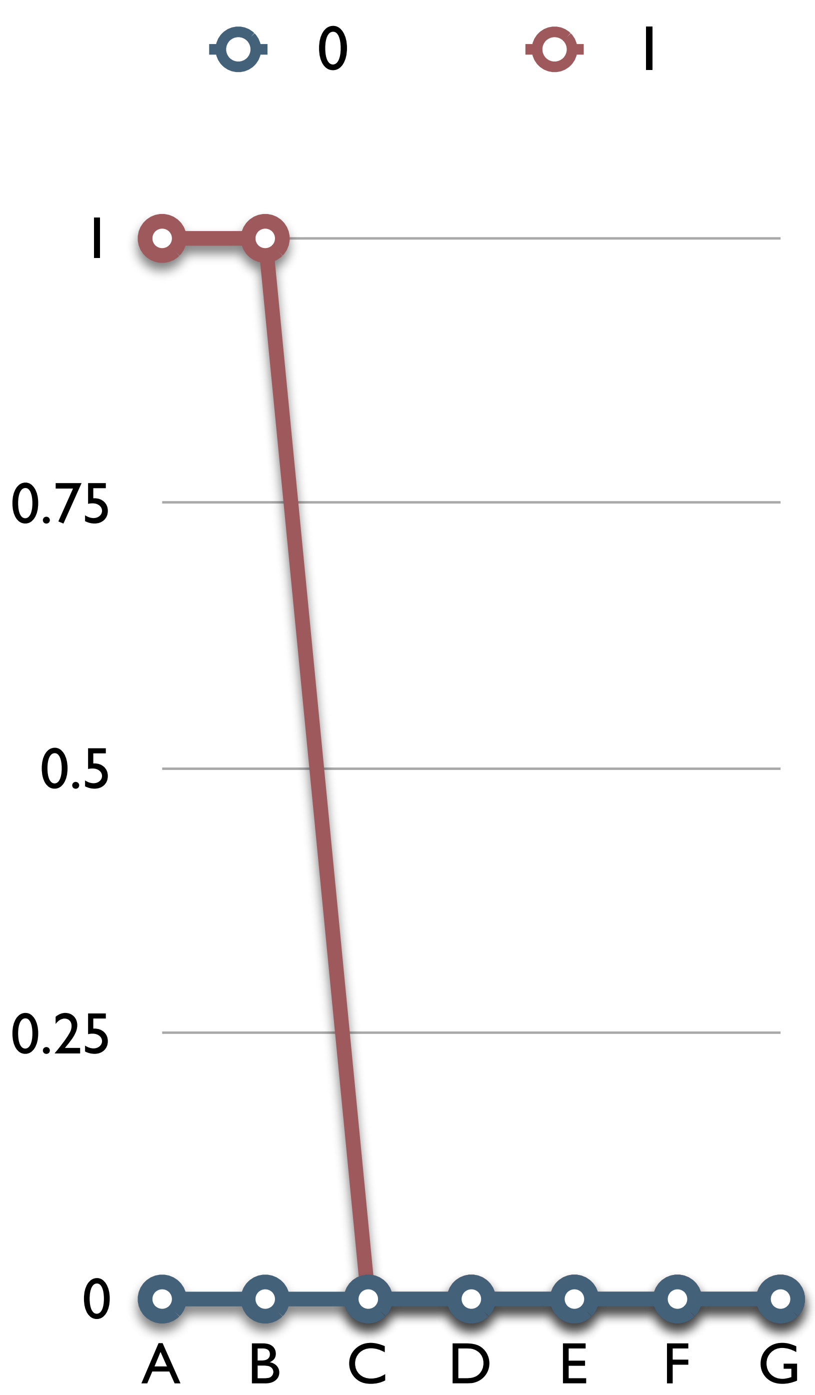
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_{11} = 7$$

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

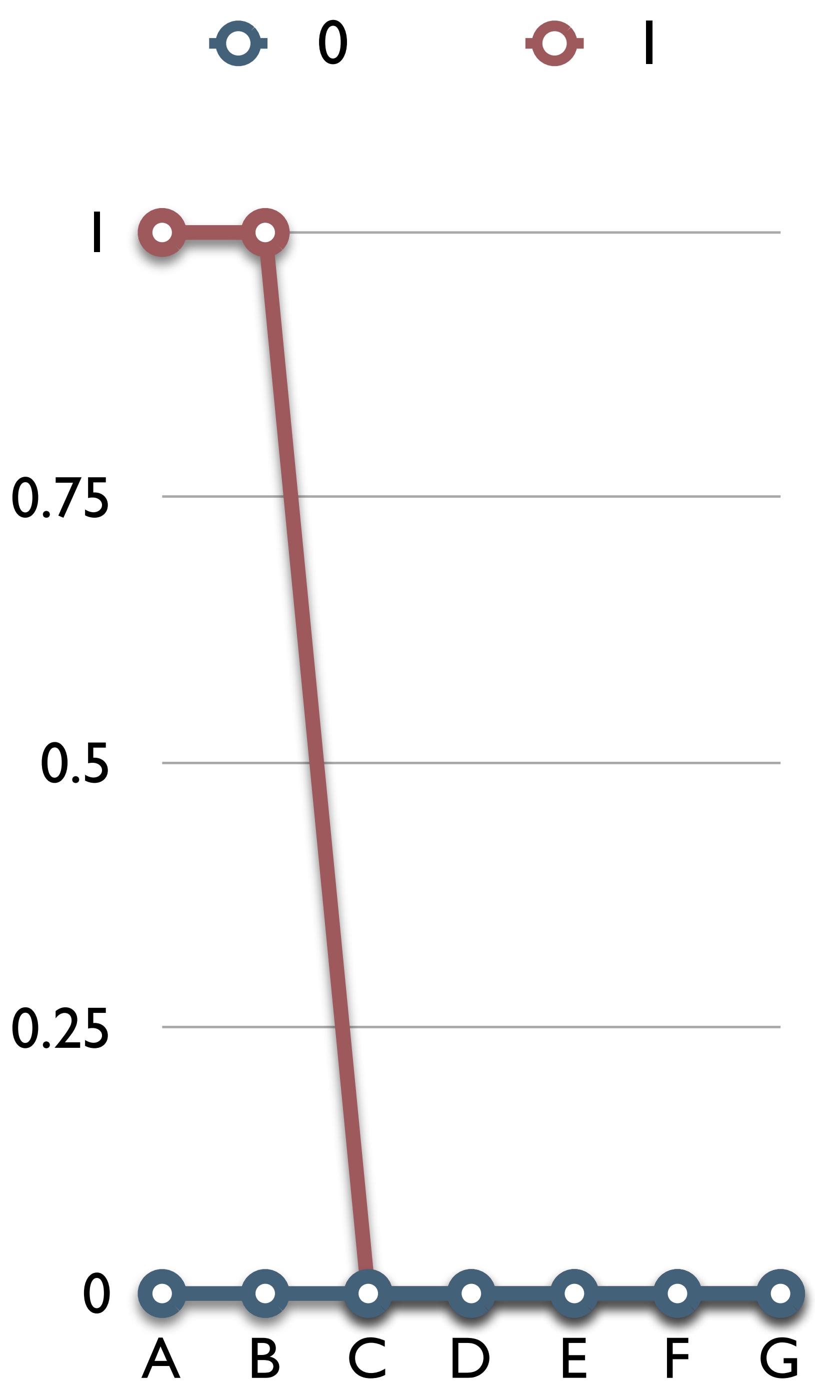


A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0



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

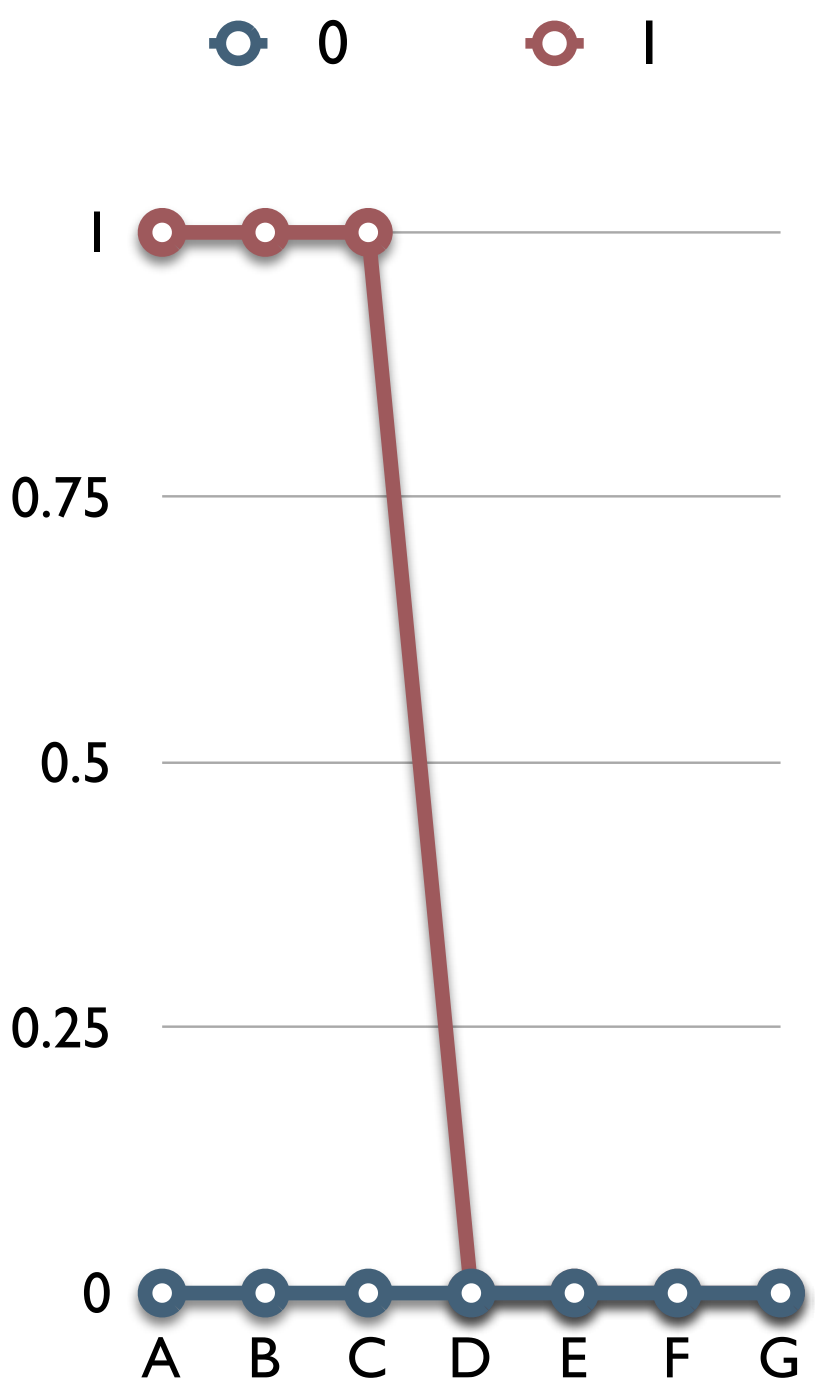
$$n_{111} = 7$$



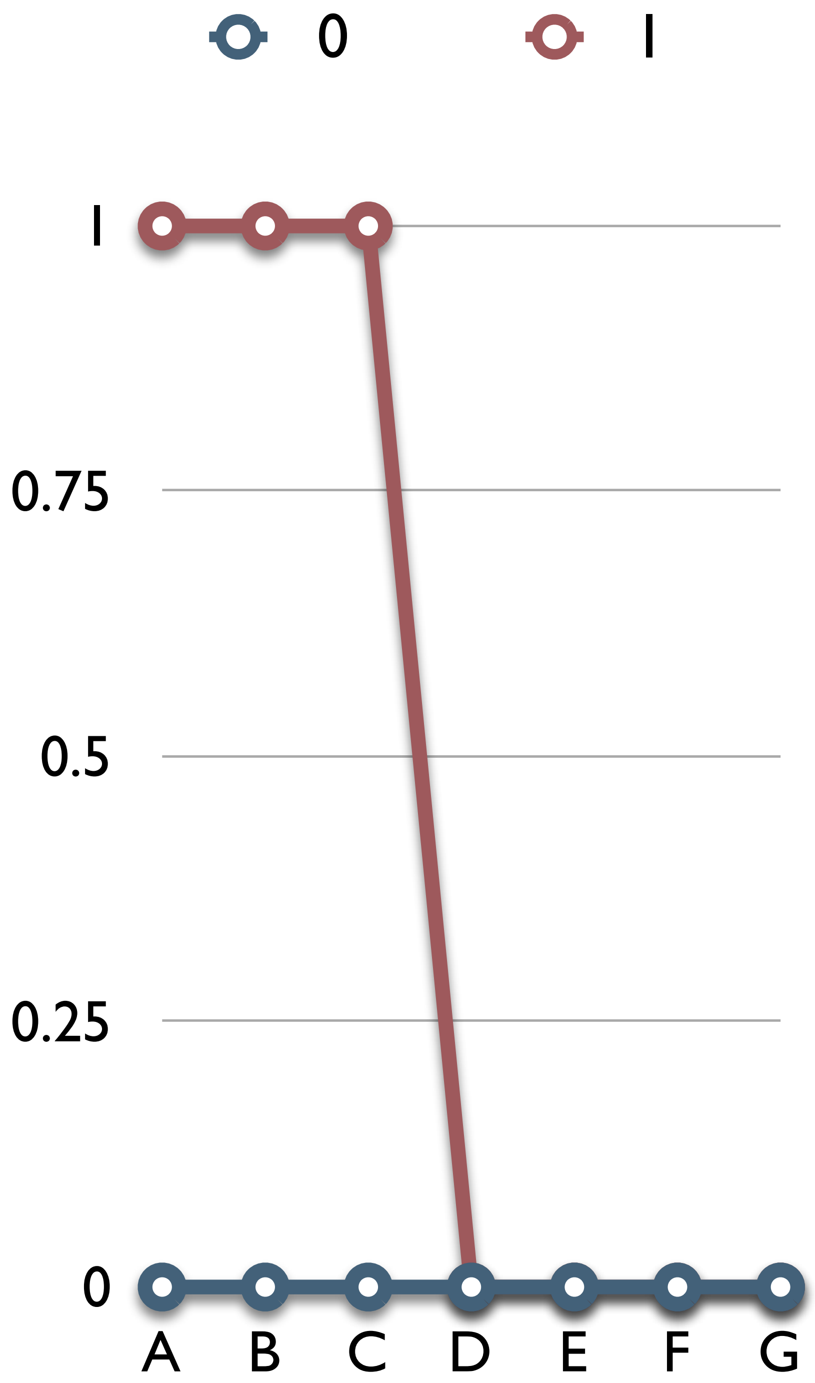
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_{111} = 7$$

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

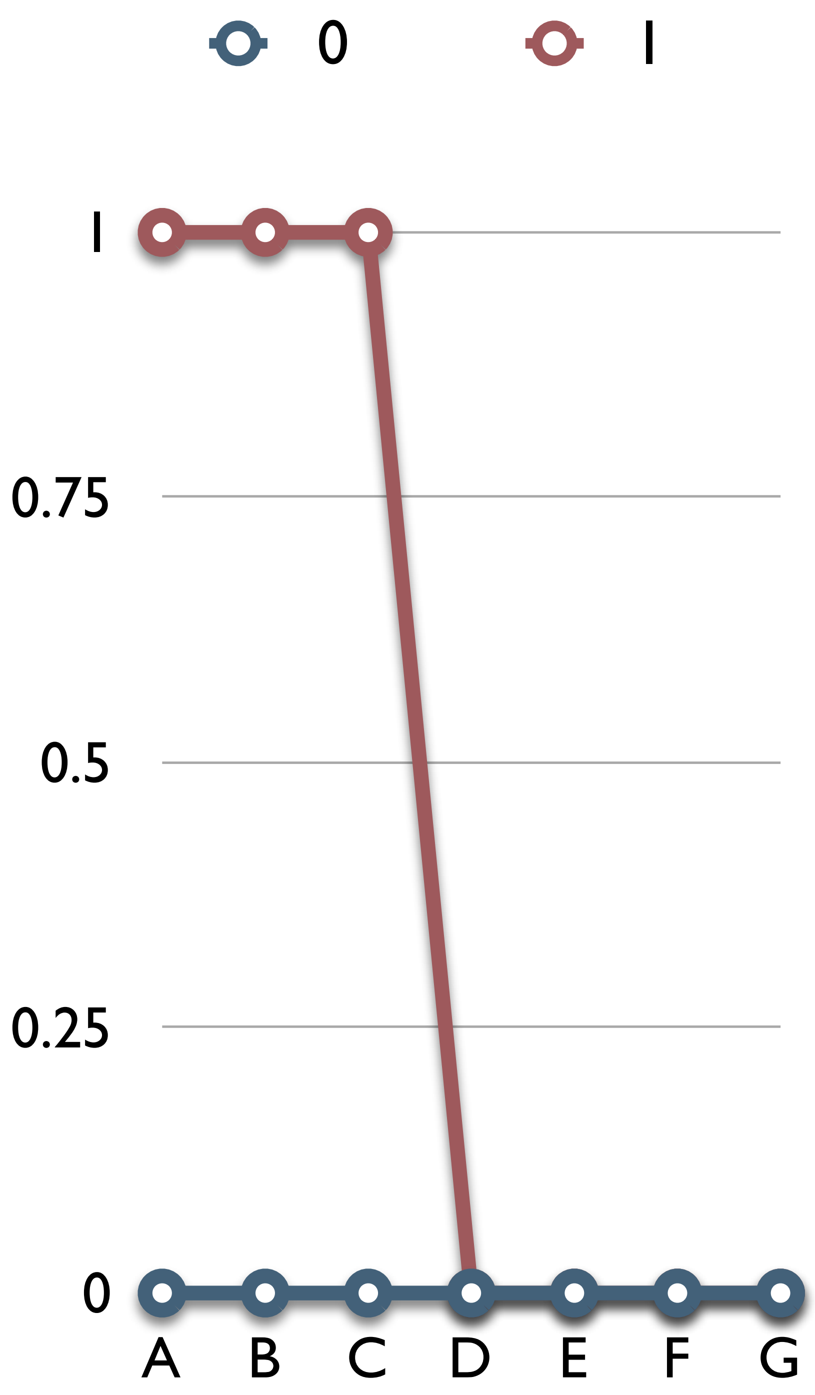


A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0



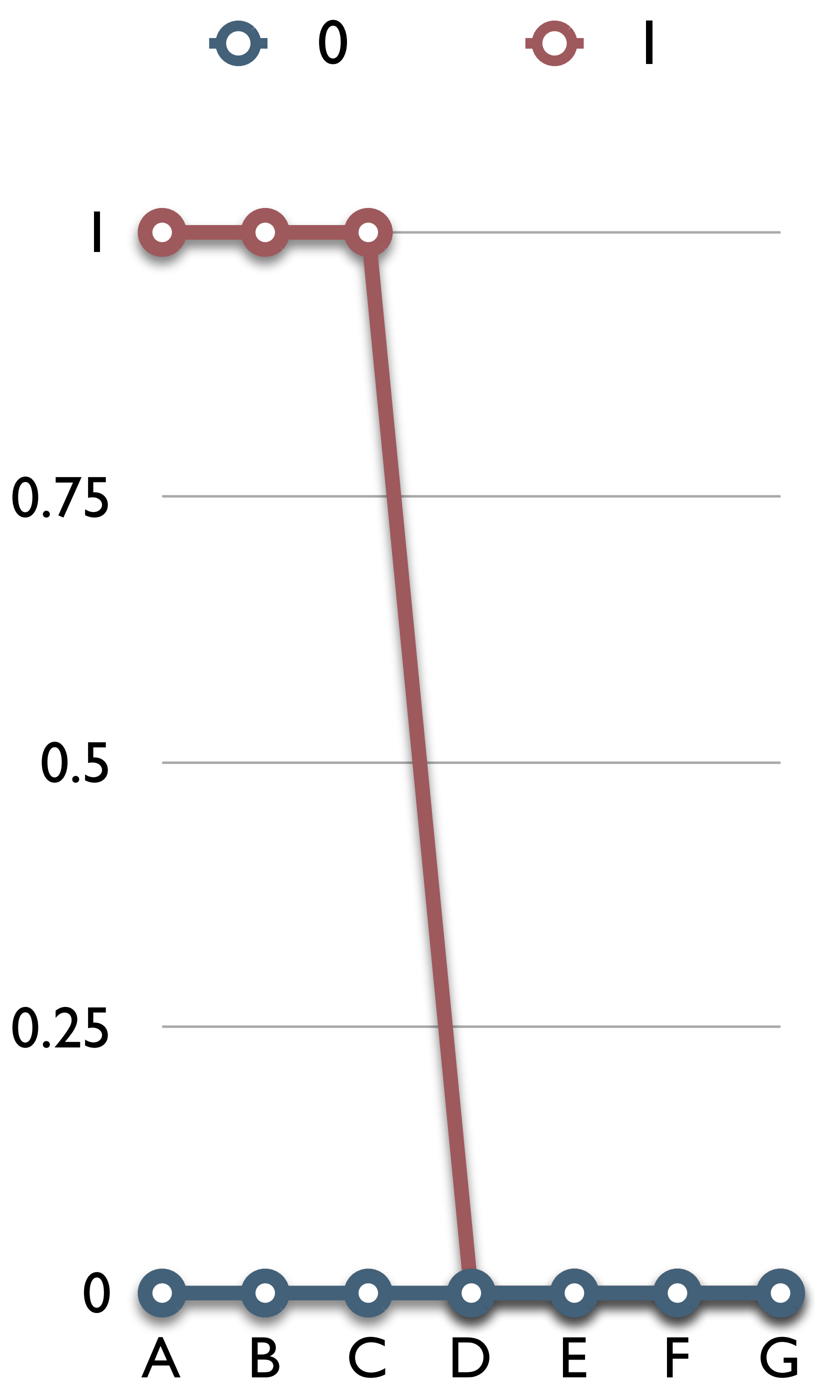
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

$$n_{1111} = 5$$



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

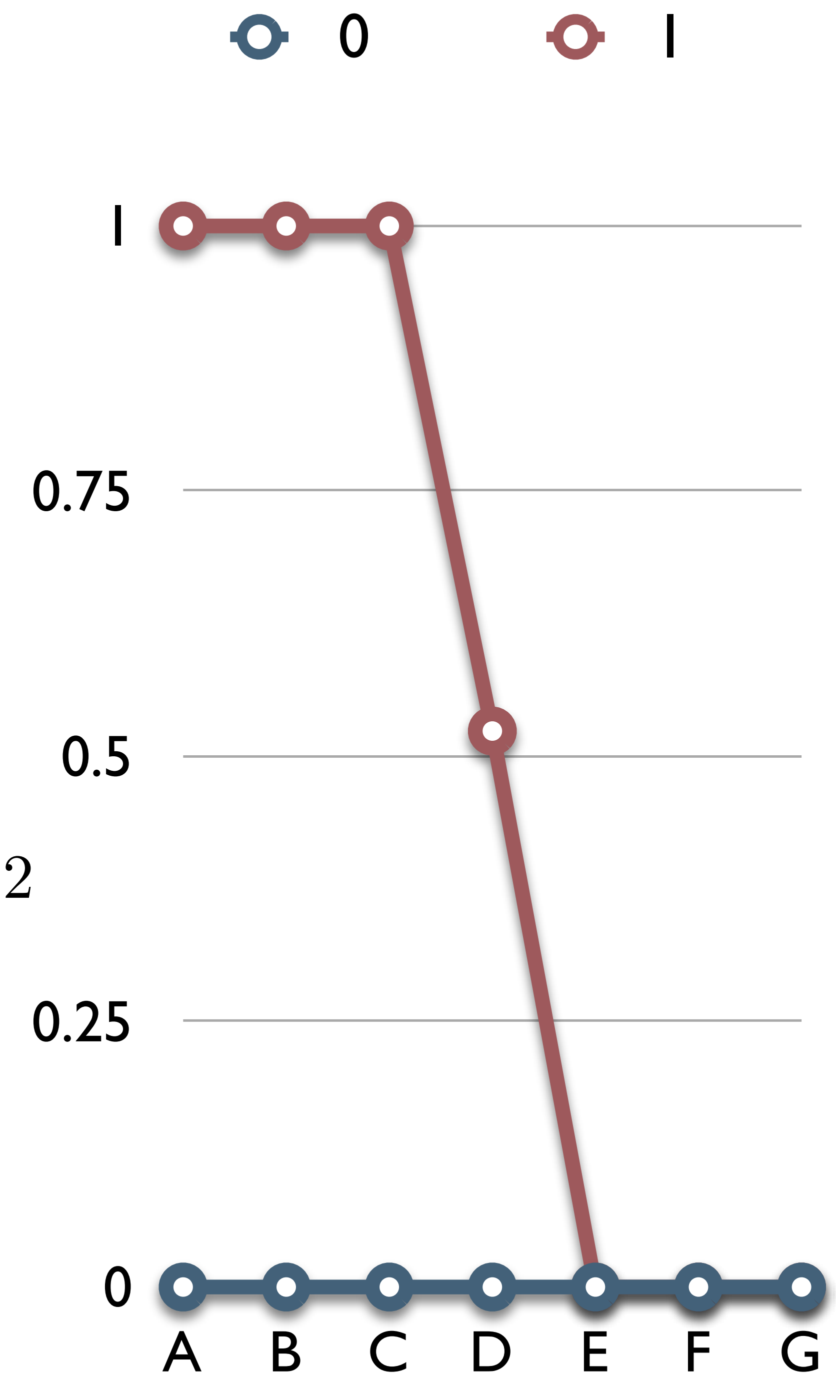
$$n_{1111} = 5$$
$$n_{1110} = 2$$



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

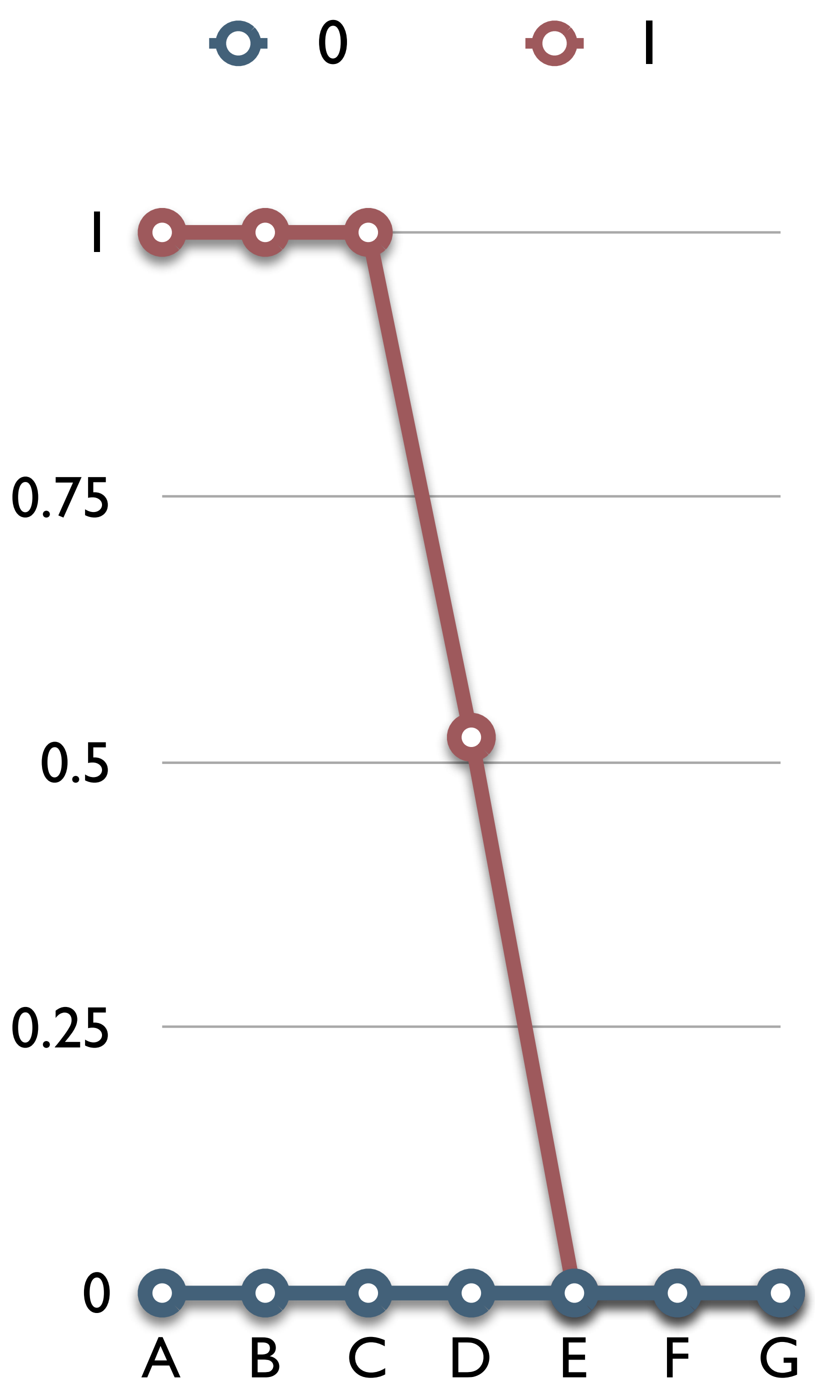
$$n_{1111} = 5$$
$$n_{1110} = 2$$

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



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

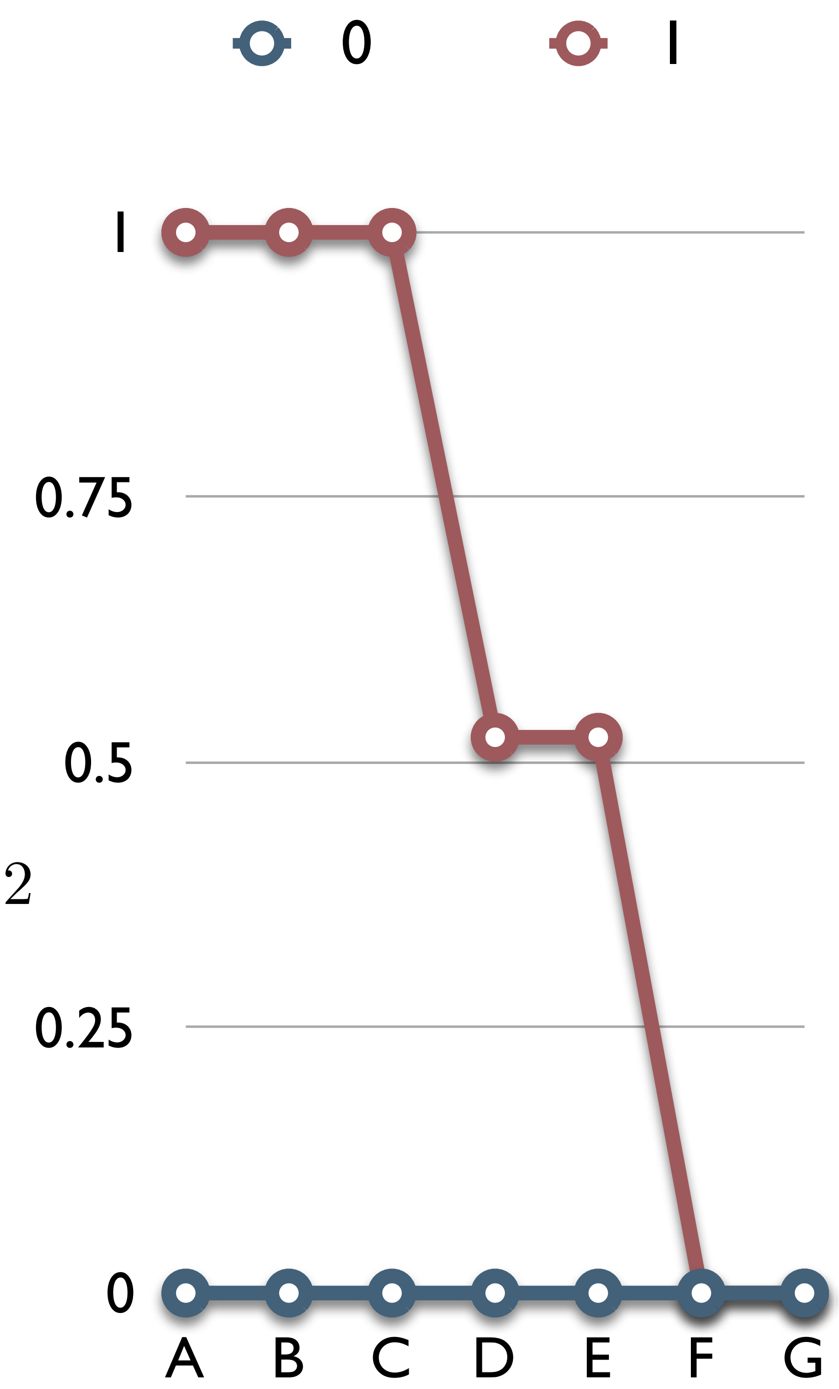
$$n_{11110} = 5$$
$$n_{11100} = 2$$



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

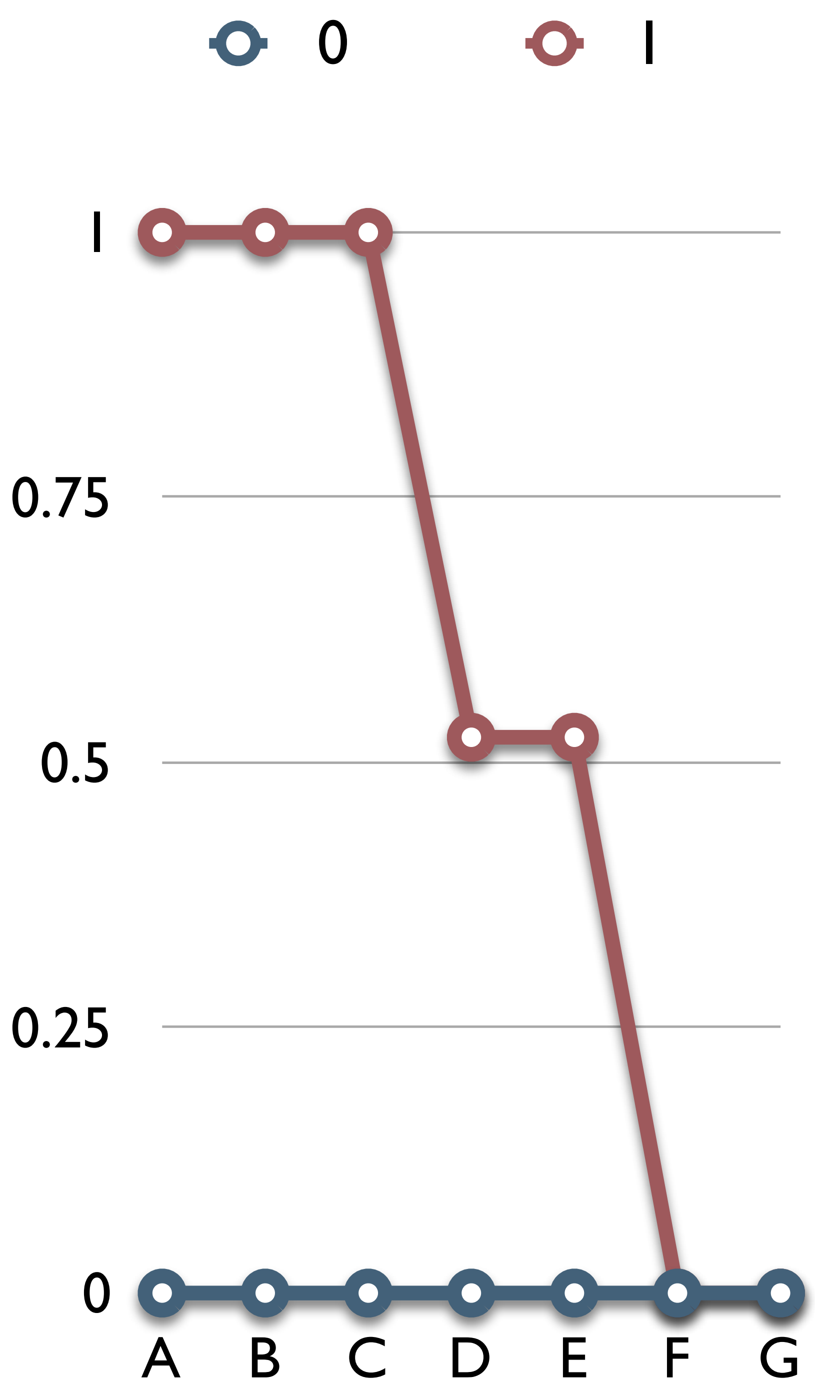
$$n_{11110} = 5$$
$$n_{11100} = 2$$

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



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

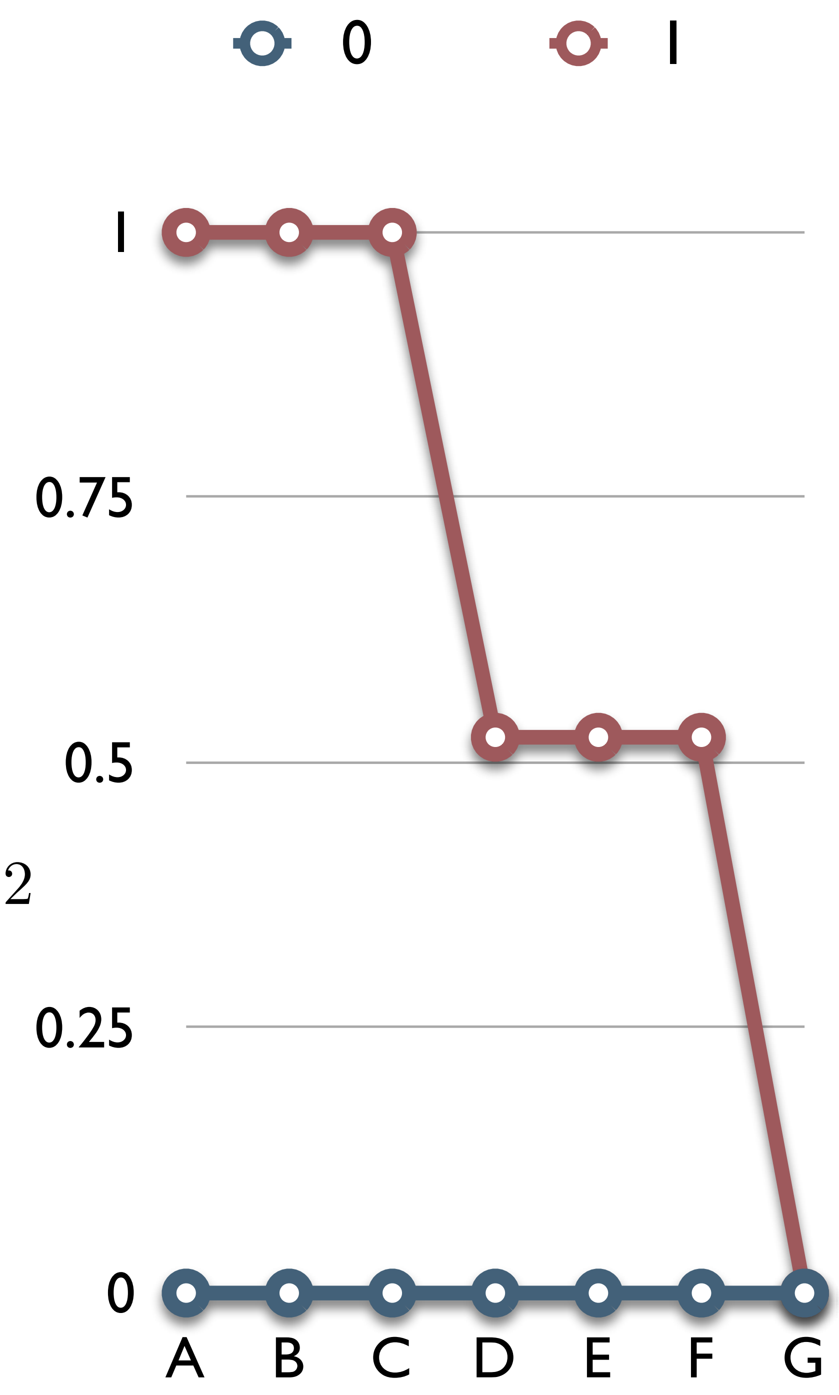
$$n_{111100} = 5$$
$$n_{111001} = 2$$



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

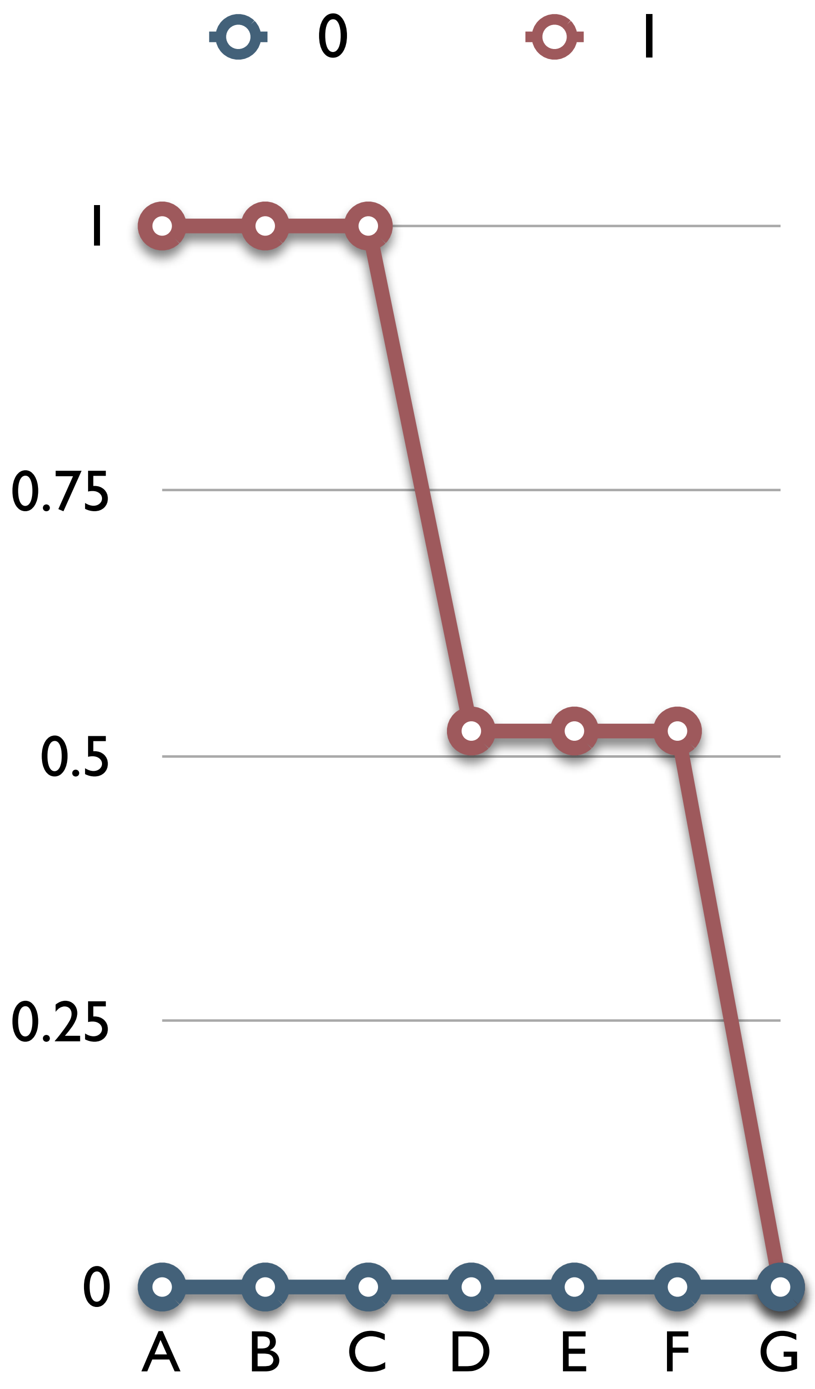
$$n_{111100} = 5$$
$$n_{111001} = 2$$

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



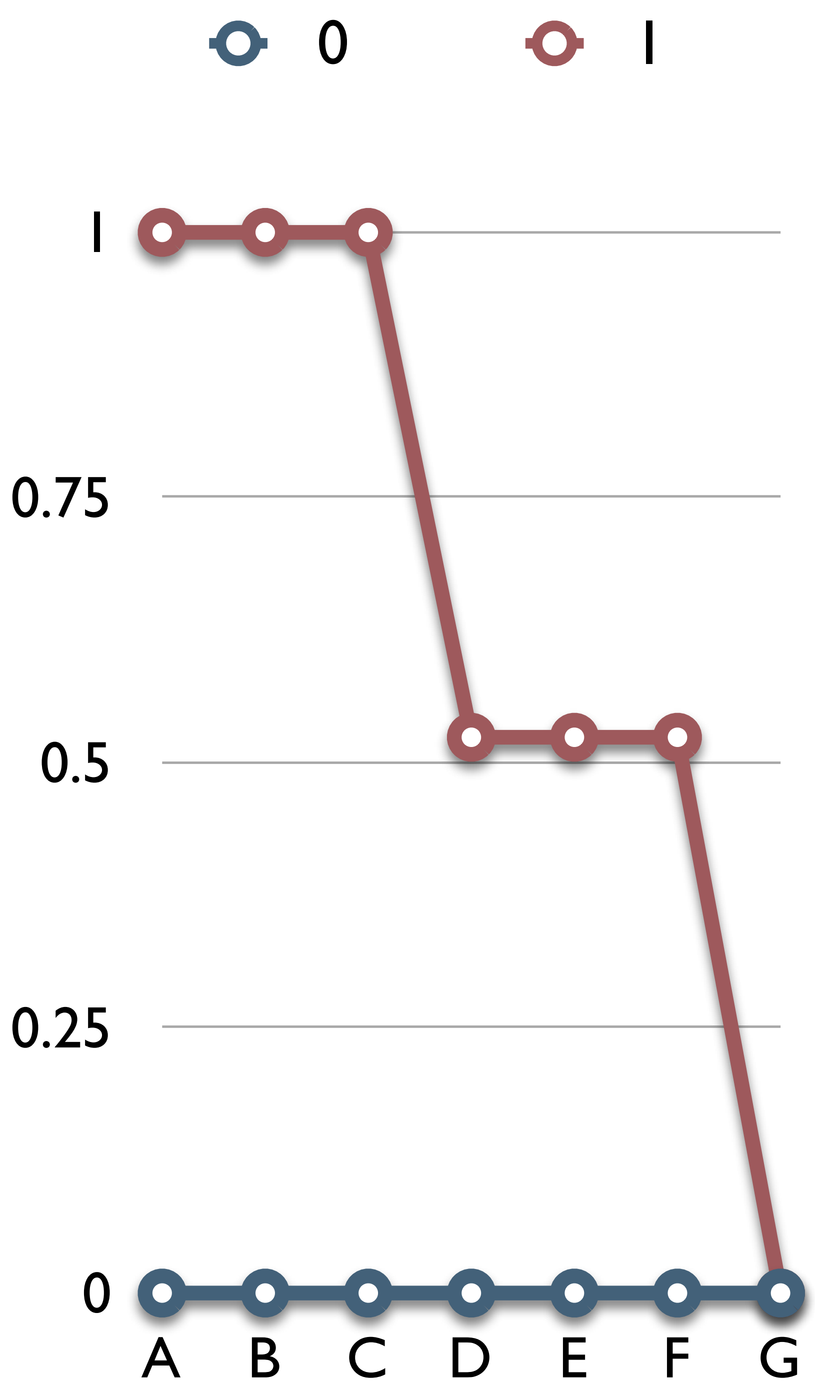
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

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



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

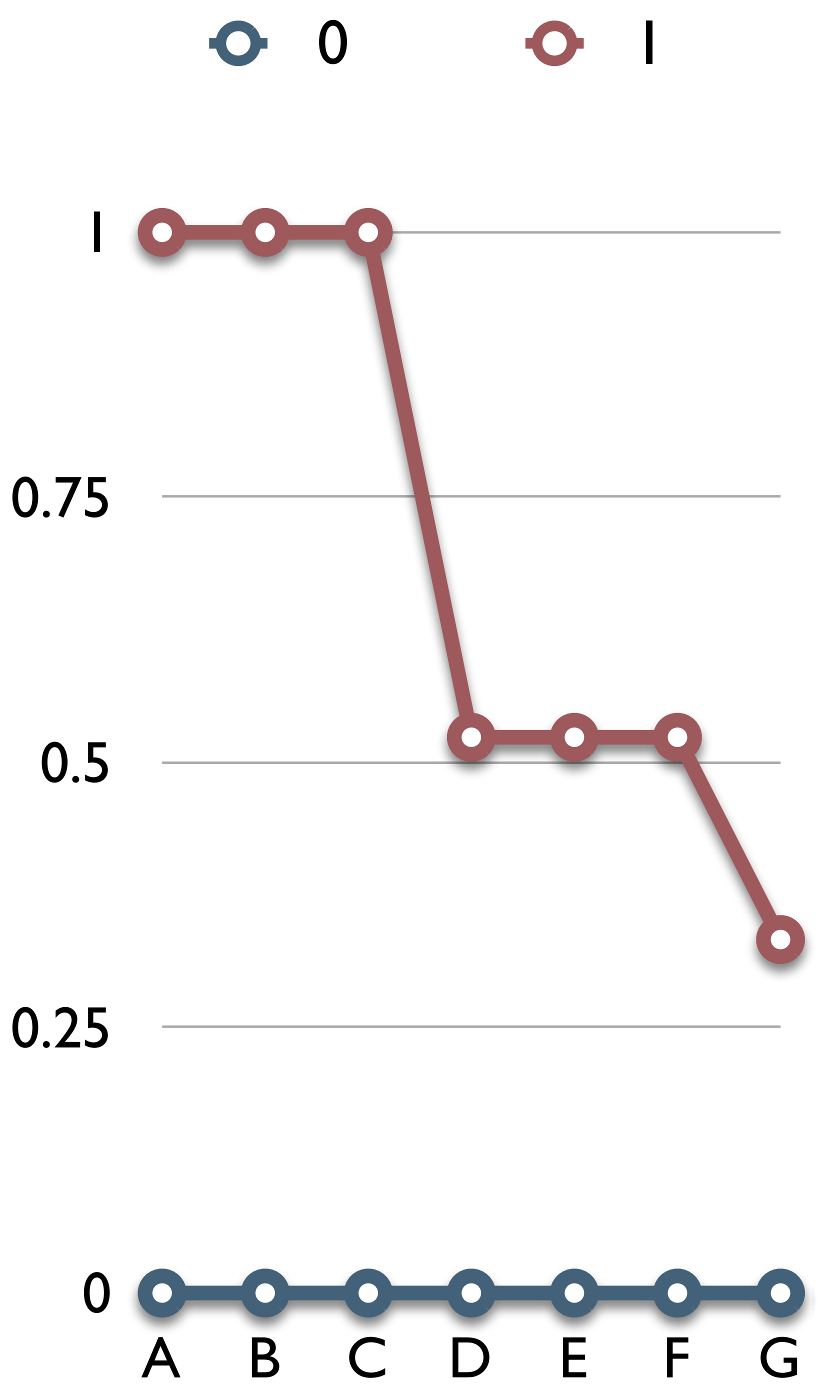
$$n_{1111000} = 4$$
$$n_{1110011} = 2$$
$$n_{1111001} = 1$$



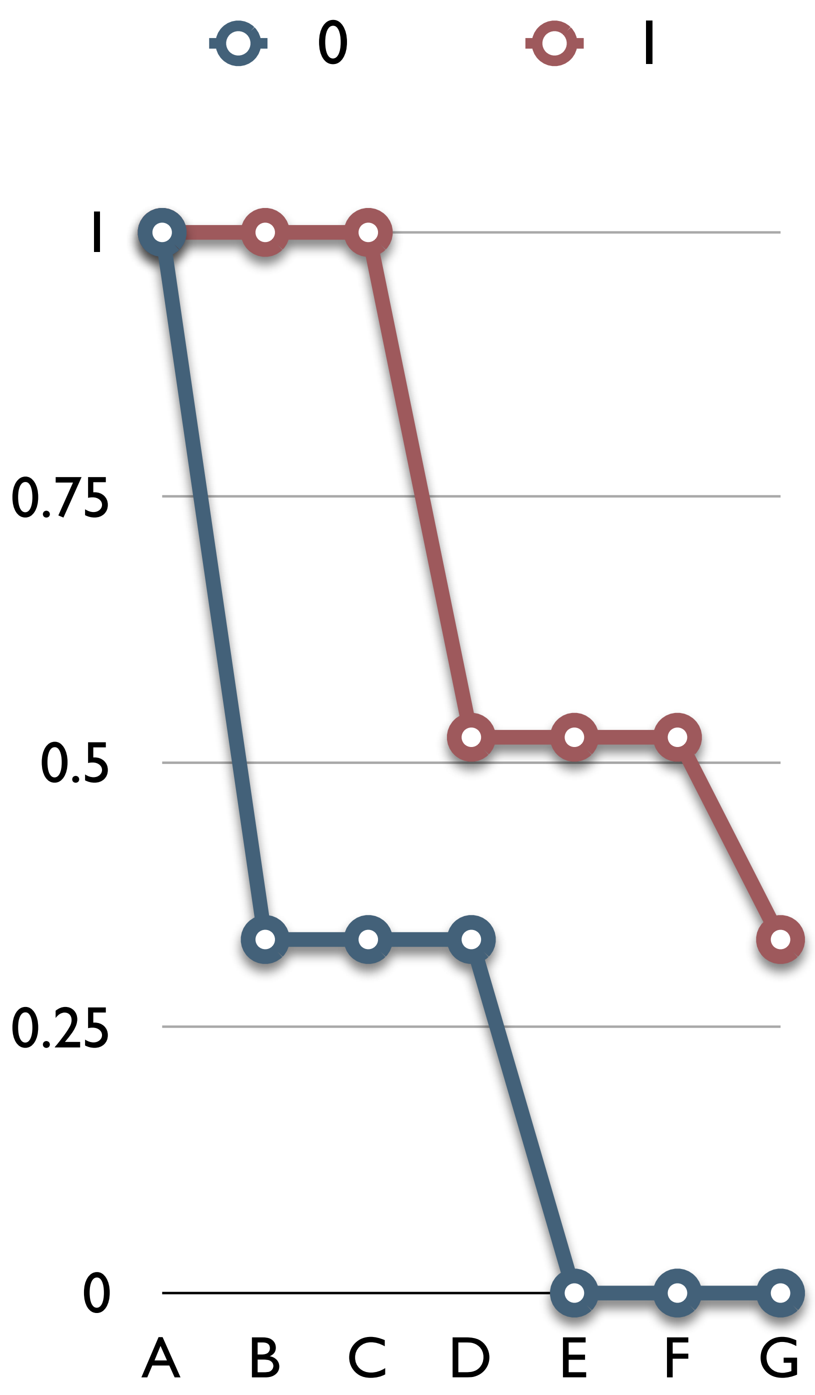
A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

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

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

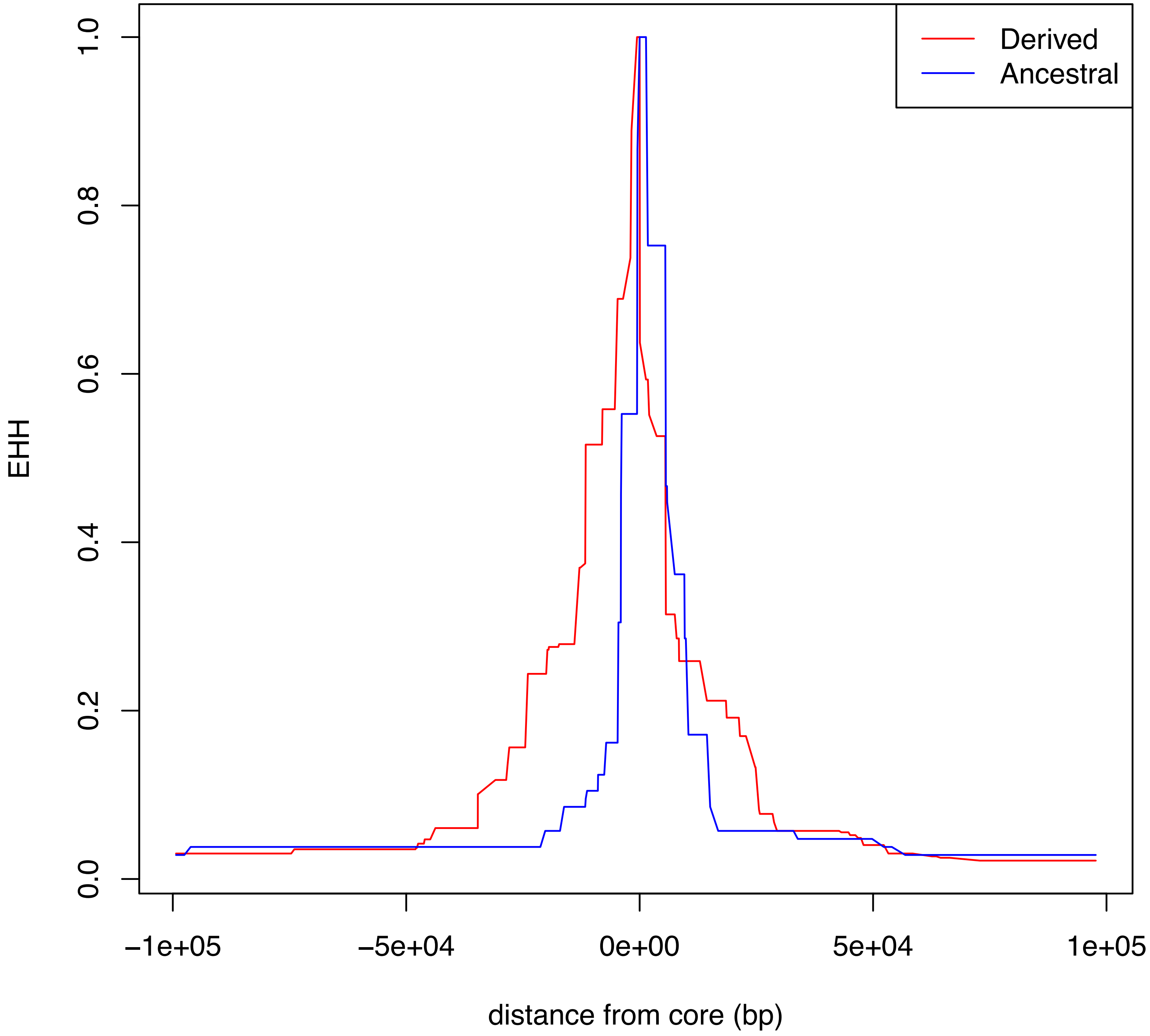


A	B	C	D	E	F	G
0	1	1	0	0	0	0
0	0	0	0	0	1	1
0	0	0	0	1	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	1
1	1	1	1	0	0	0
1	1	1	0	0	1	1
1	1	1	0	0	1	1
1	1	1	1	0	0	0
1	1	1	1	0	0	0

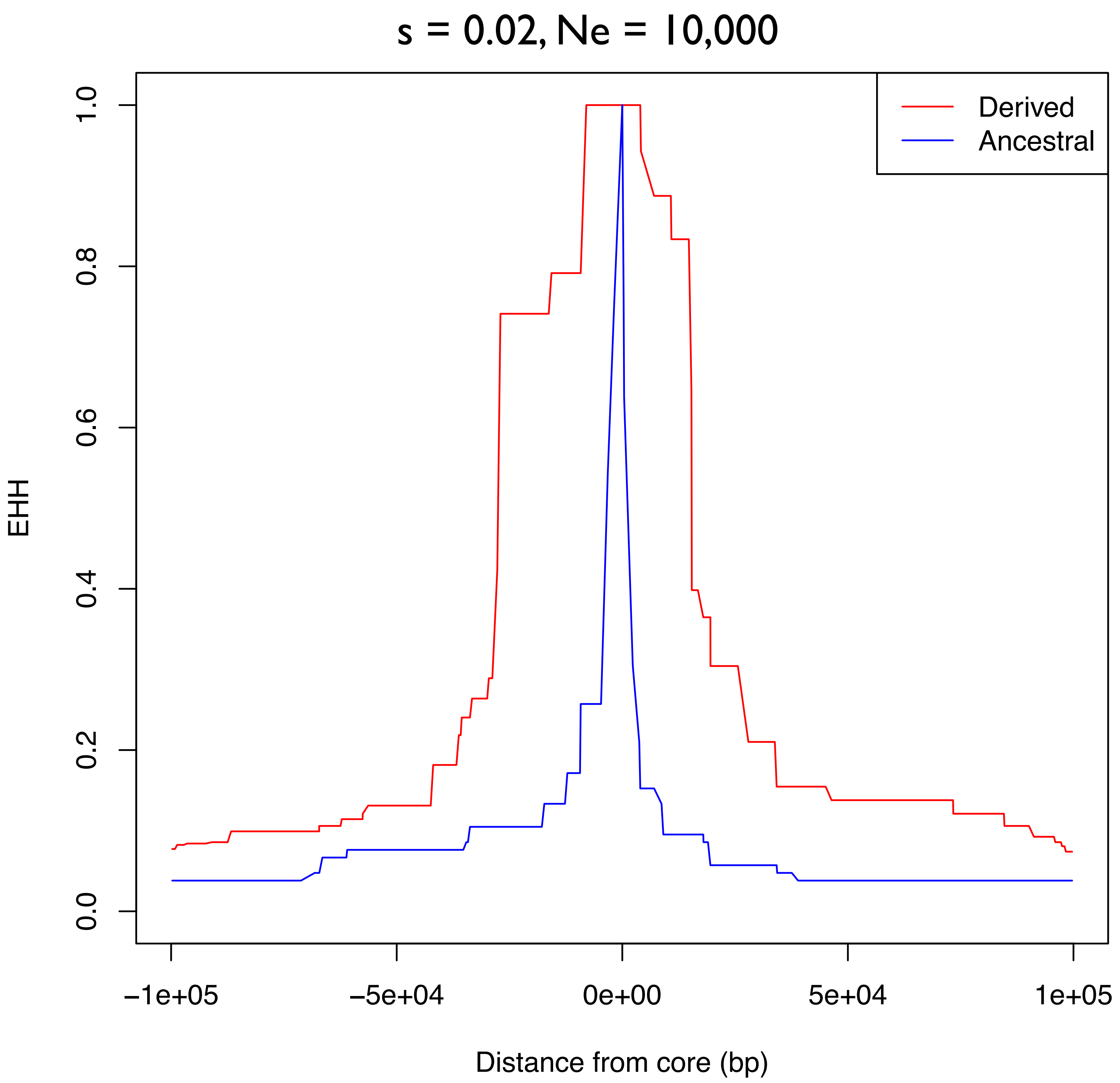


EHH

s = 0.01, Ne = 10,000

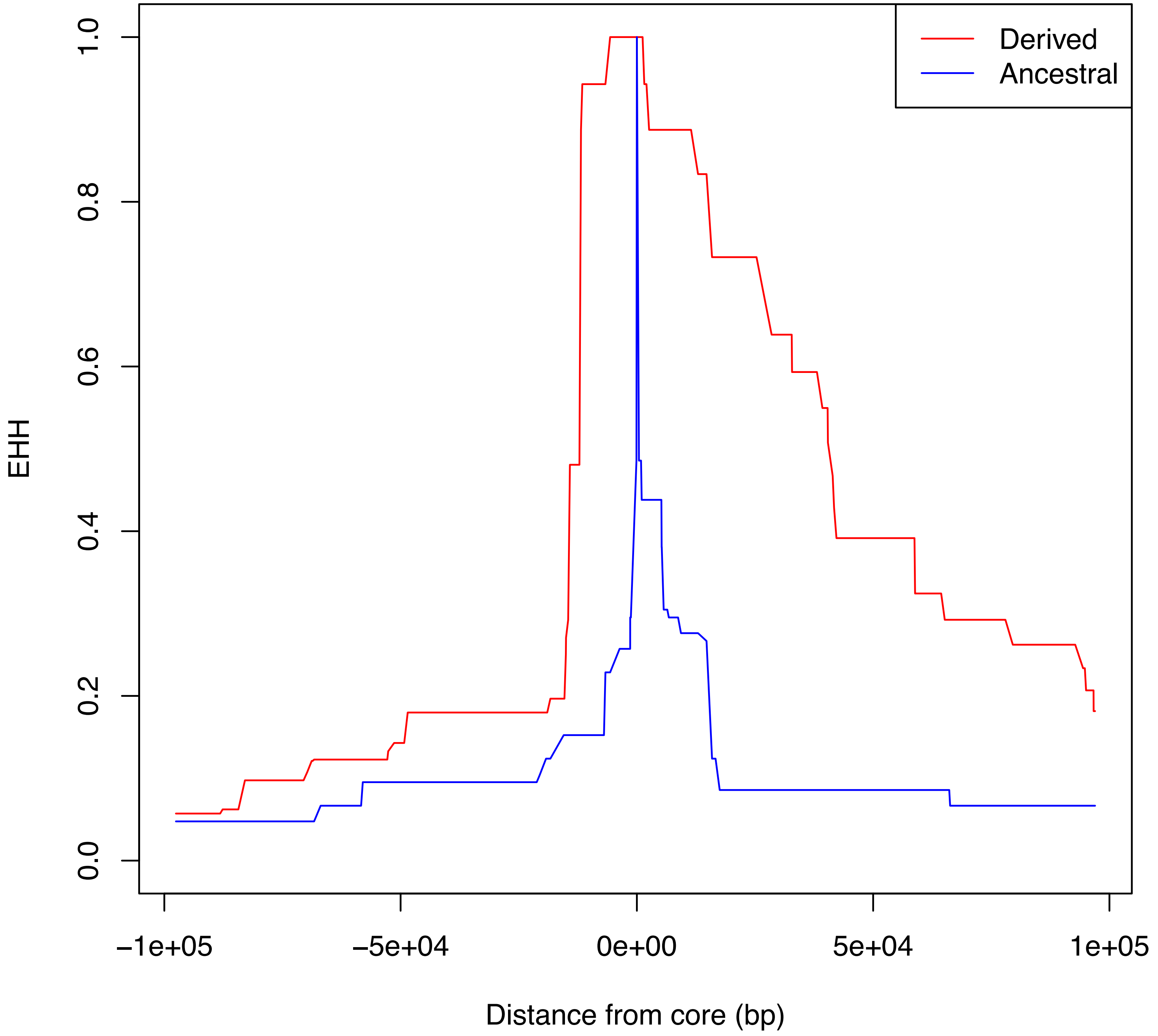


EHH


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)

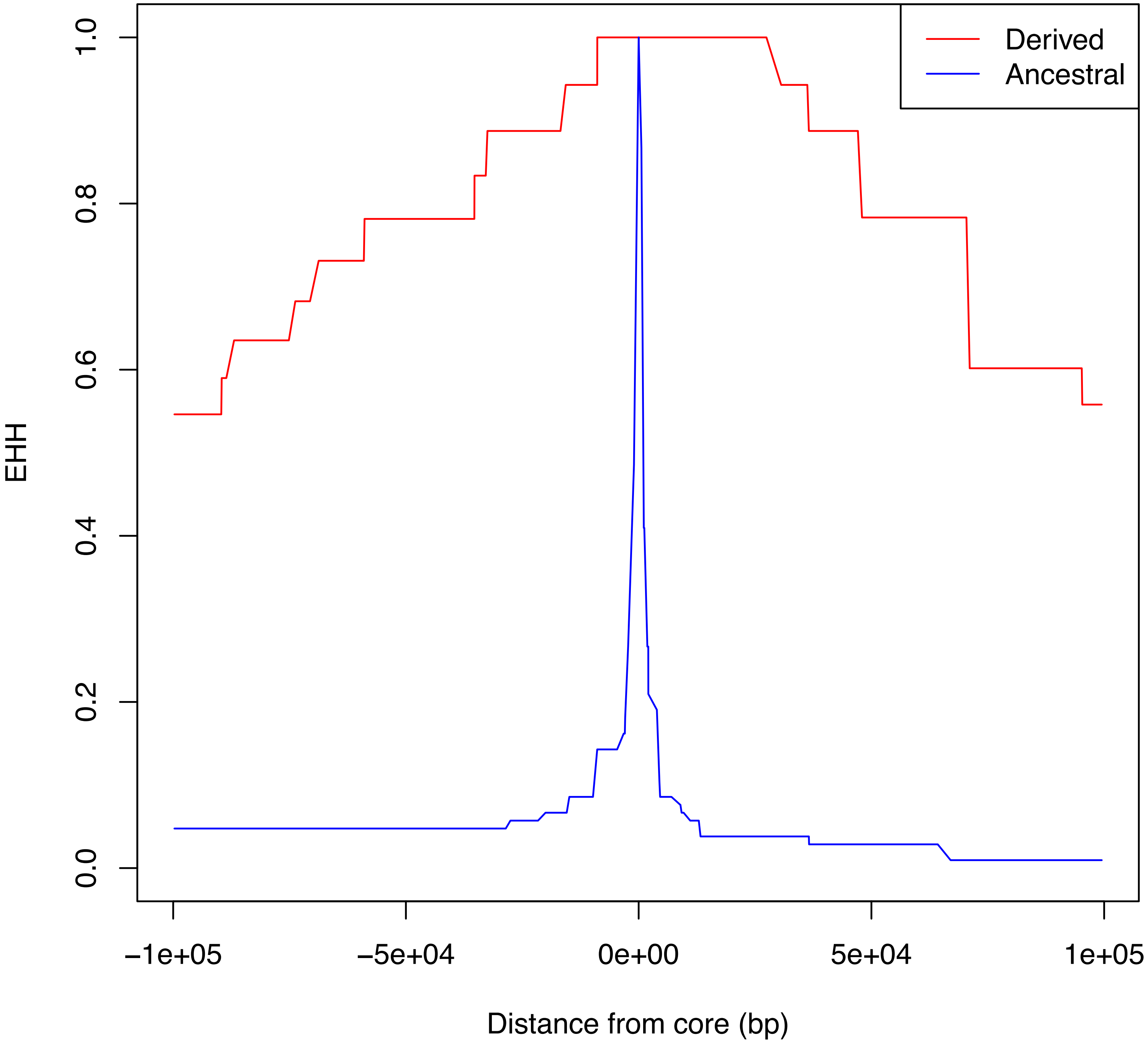
EHH

s = 0.05, Ne = 10,000



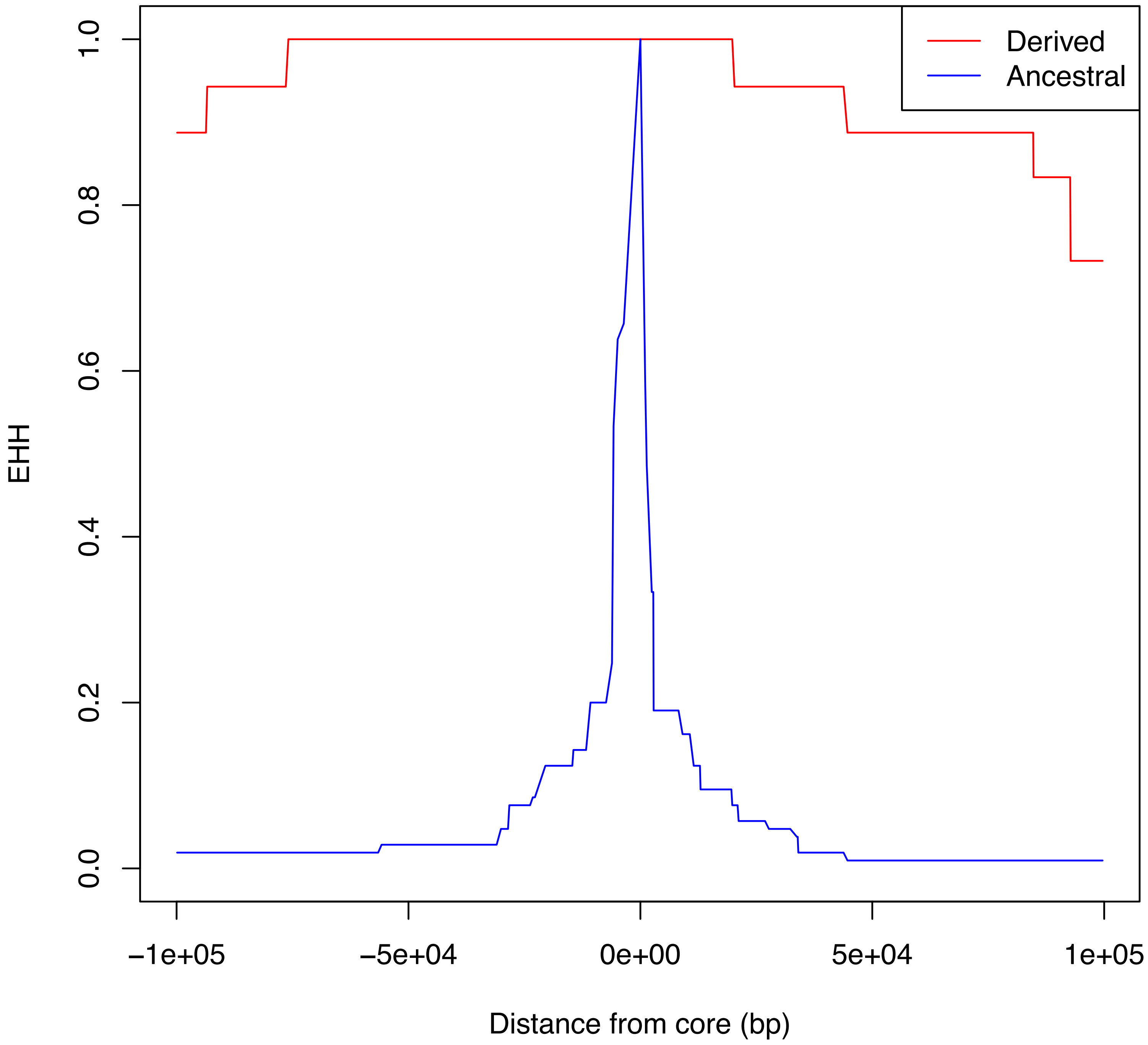
EHH

s = 0.10, Ne = 10,000



EHH

s = 0.50, Ne = 10,000



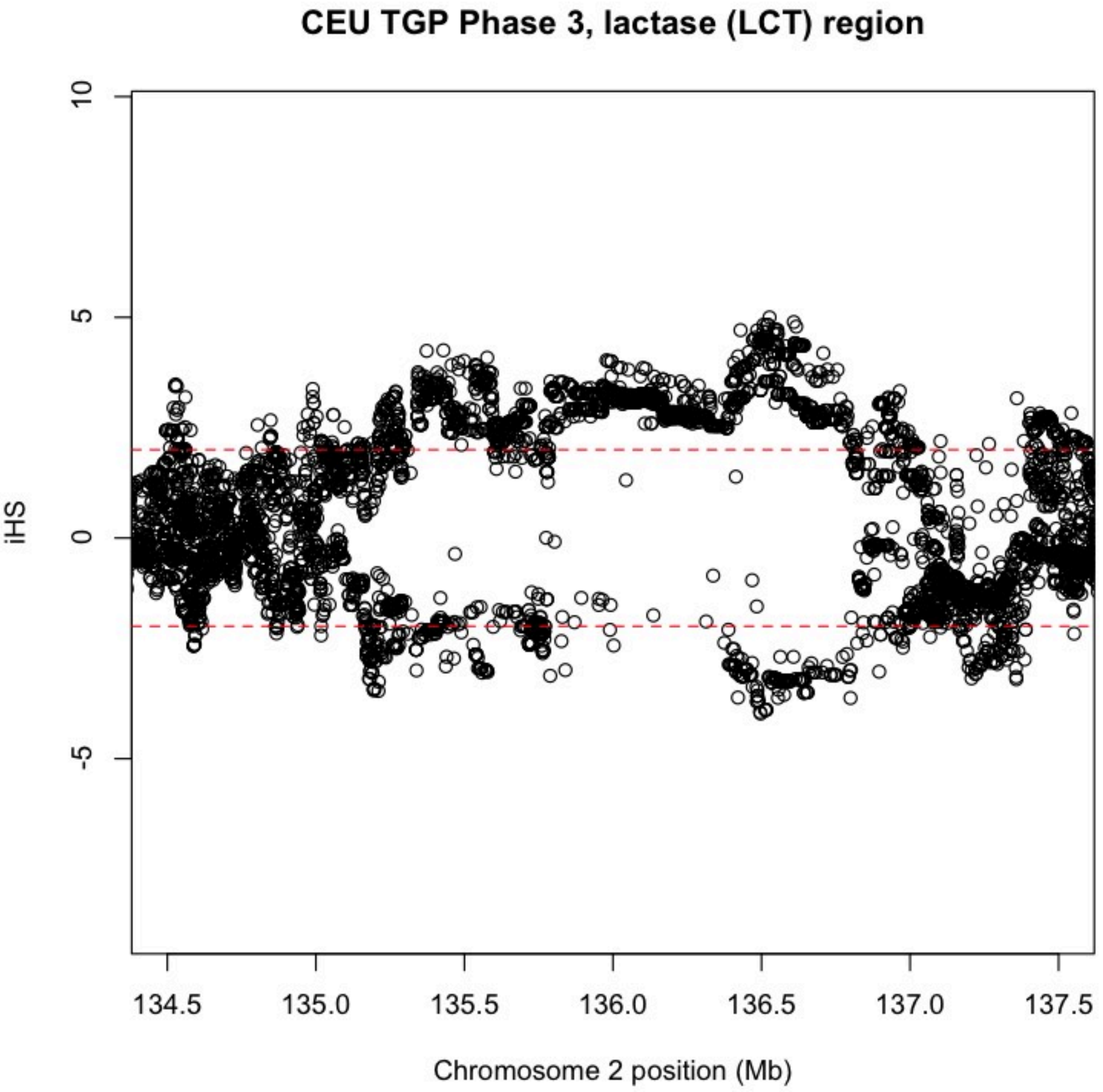
EHH

- 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)

Computational Tips

- Associative arrays for haplotype comparison and counting
    - O(log N)
- Instead of computing EHH until the end of the data stop after a certain distance away from the core
    - Either EHH < 0.05 or distance from core > 1Mb
- Multithreading
    - Adjacent SNPs don’t rely on each other to complete calculation
    - Compute adjacent scores on separate threads

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

Computational Tips

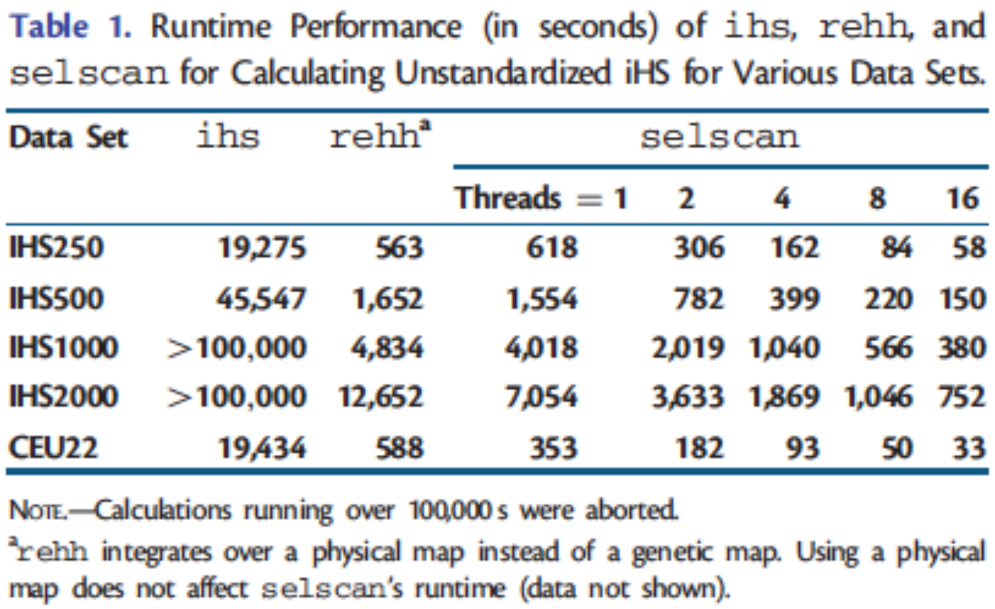


Table 1. Runtime Performance (in seconds) of ihs, rehh, and selscan for Calculating Unstandardized iHS for Various Data Sets.

Data Set	ihs	rehh[a]	selscan				
			Threads = 1	2	4	8	16
IHS250	19,275	563	618	306	162	84	58
IHS500	45,547	1,652	1,554	782	399	220	150
IHS1000	>100,000	4,834	4,018	2,019	1,040	566	380
IHS2000	>100,000	12,652	7,054	3,633	1,869	1,046	752
CEU22	19,434	588	353	182	93	50	33

Note.—Calculations running over 100,000 s were aborted.

[a] rehh integrates over a physical map instead of a genetic map. Using a physical map does not affect selscan's runtime (data not shown).

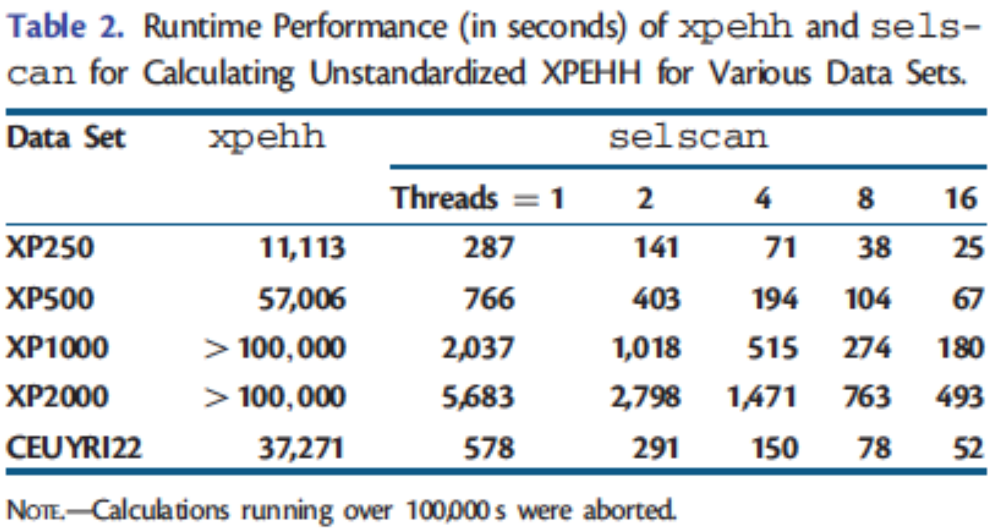


Table 2. Runtime Performance (in seconds) of xpehh and selscan for Calculating Unstandardized XPEHH for Various Data Sets.

Data Set	xpehh	selscan				
		Threads = 1	2	4	8	16
XP250	11,113	287	141	71	38	25
XP500	57,006	766	403	194	104	67
XP1000	> 100,000	2,037	1,018	515	274	180
XP2000	> 100,000	5,683	2,798	1,471	763	493
CEUYRI22	37,271	578	291	150	78	52

Note.—Calculations running over 100,000 s were aborted.

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

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!

Running `selscan`: iHS

- Open up your command prompt (i.e., rev your engines)
- Let’s give iHS a go!
- Let’s consider the LCT gene.
- First transfer data to your computer…
    - You will need `selscan.zip`
- Easy if you put it on your Desktop and unzip it:
    - `~/Desktop/selscan/`
- `selscan` also available: https://github.com/szpiech/selscan.

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

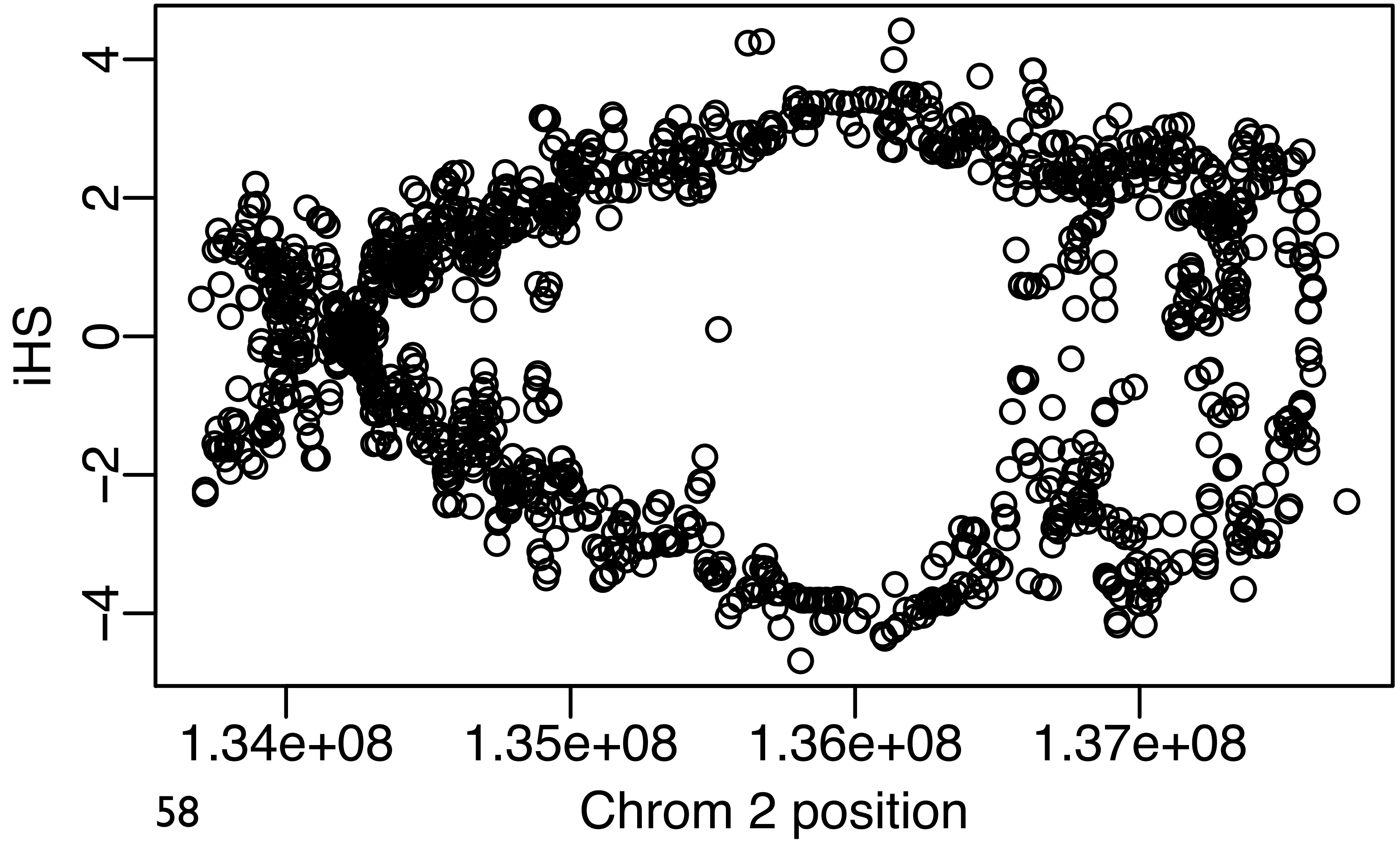
```
        --ihs
        --files CEU.chr2.ihs.out bg.ihs.out
```

```
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])
```

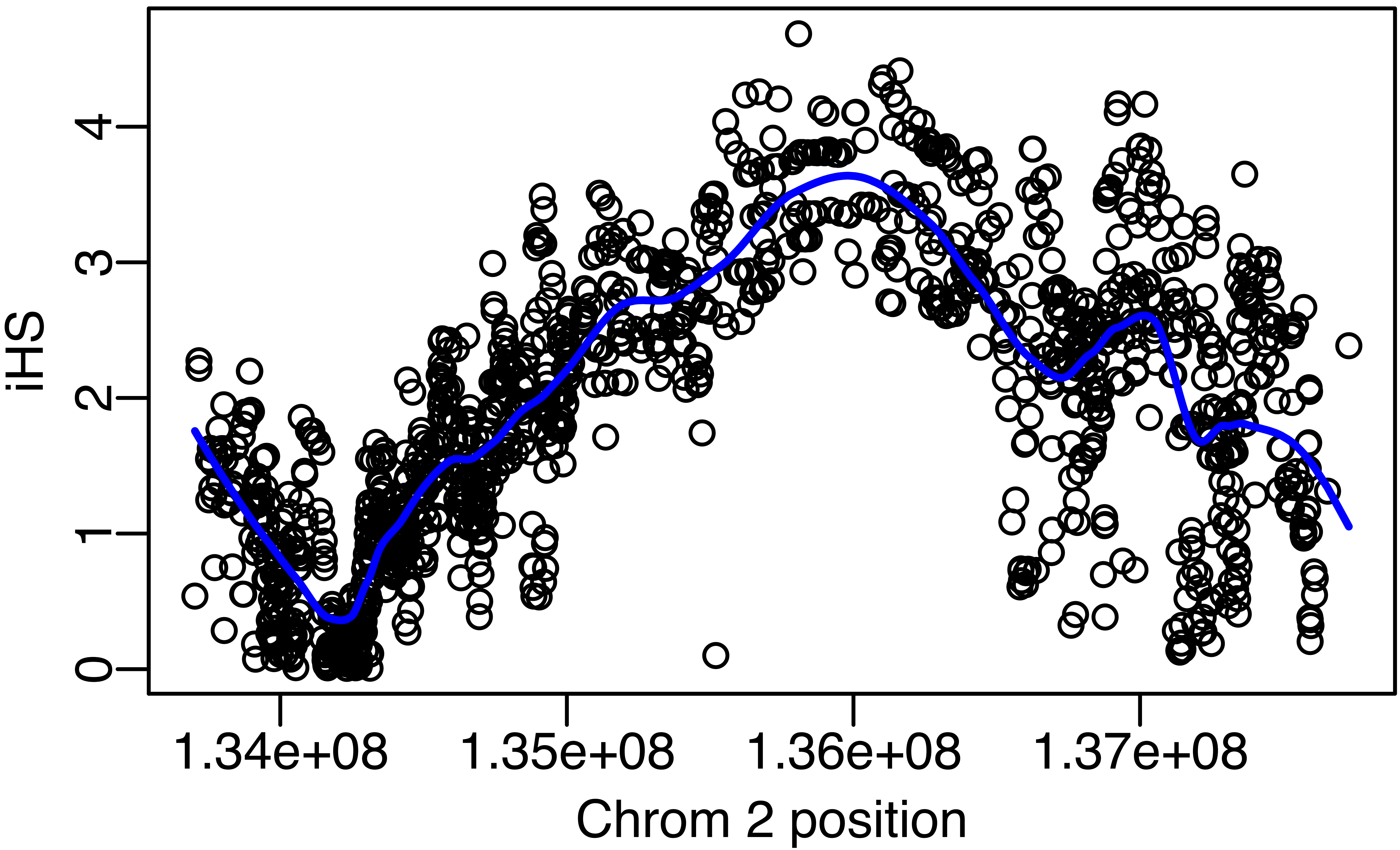


iHS

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

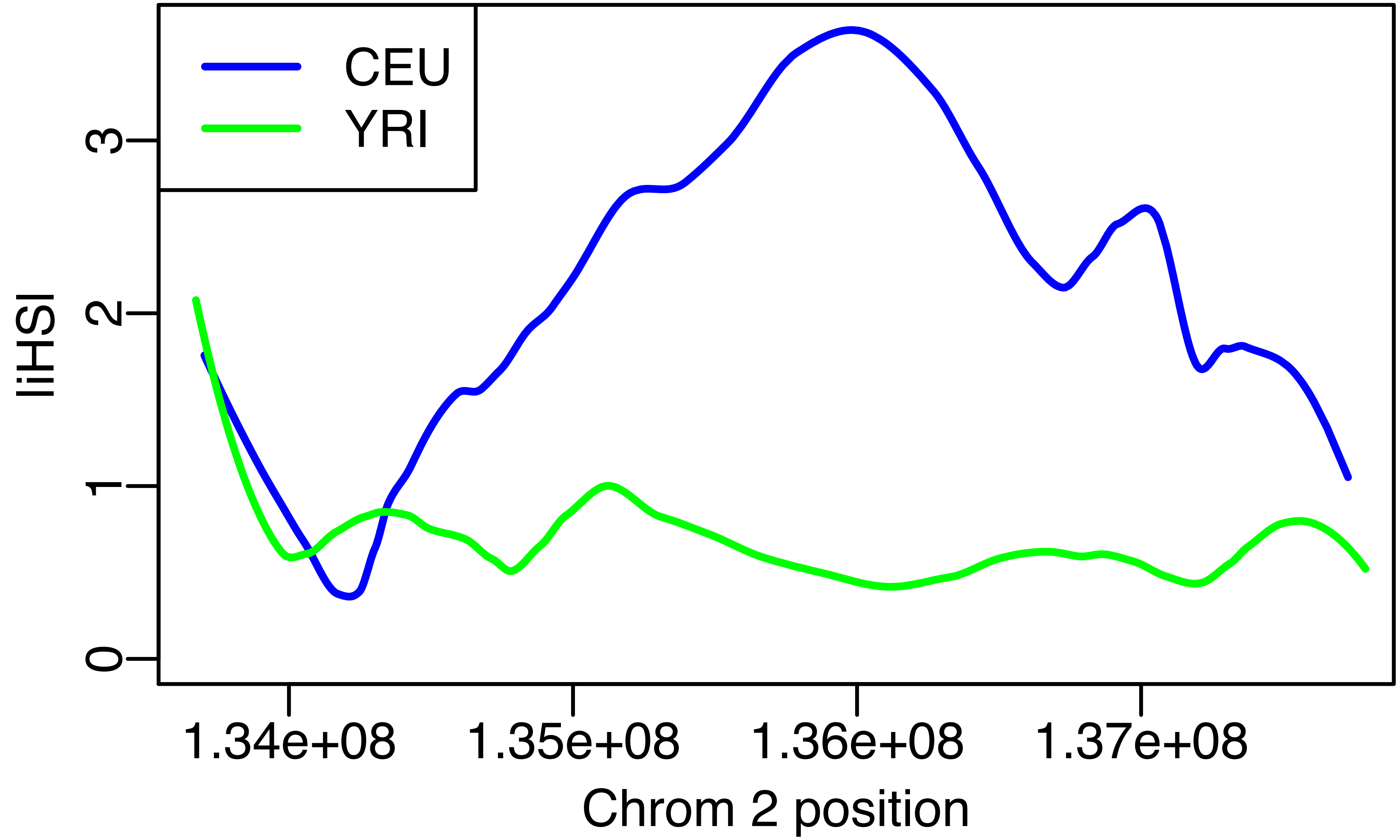
```
SP=0.2 #this is the span, a parameter you can change (higher = more
smoothing)

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')
```

iHS



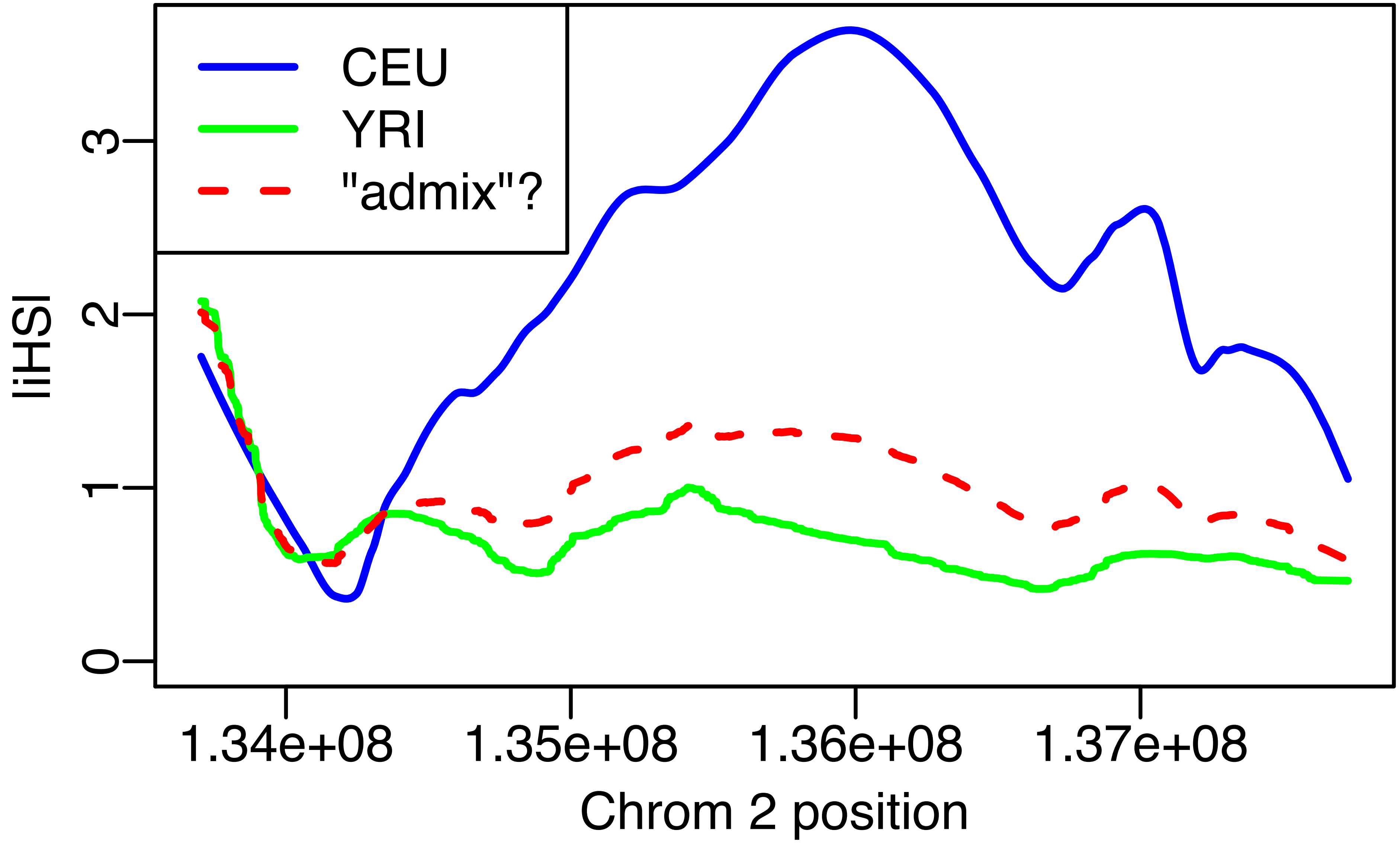
Other populations??

- 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"



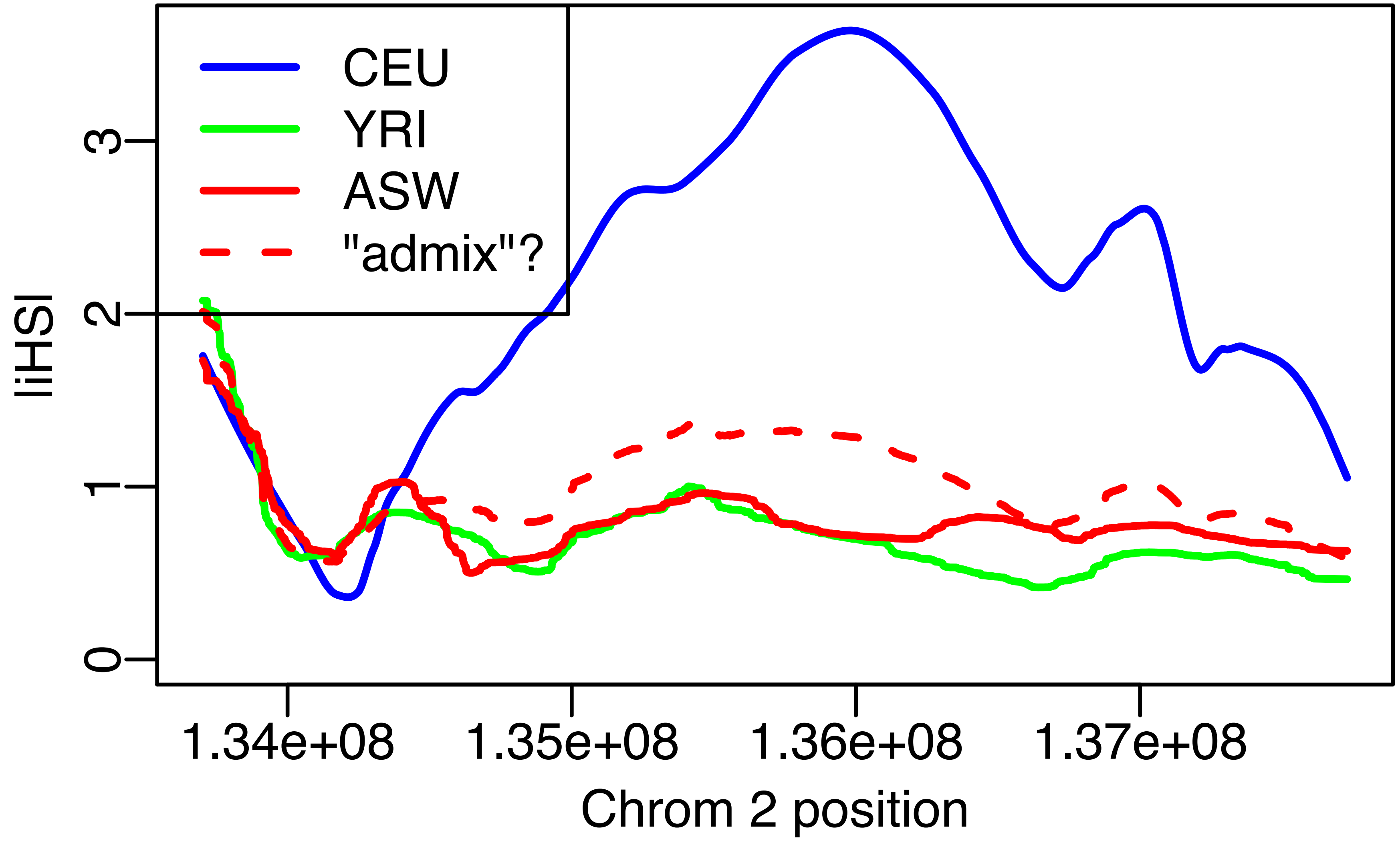
What about admixture?

- 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"
- 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.