More Introduction to Positive Selection

Ryan Hernandez
Tim O'Connor

ryan.hernandez@ucsf.edu

Genome-wide scans

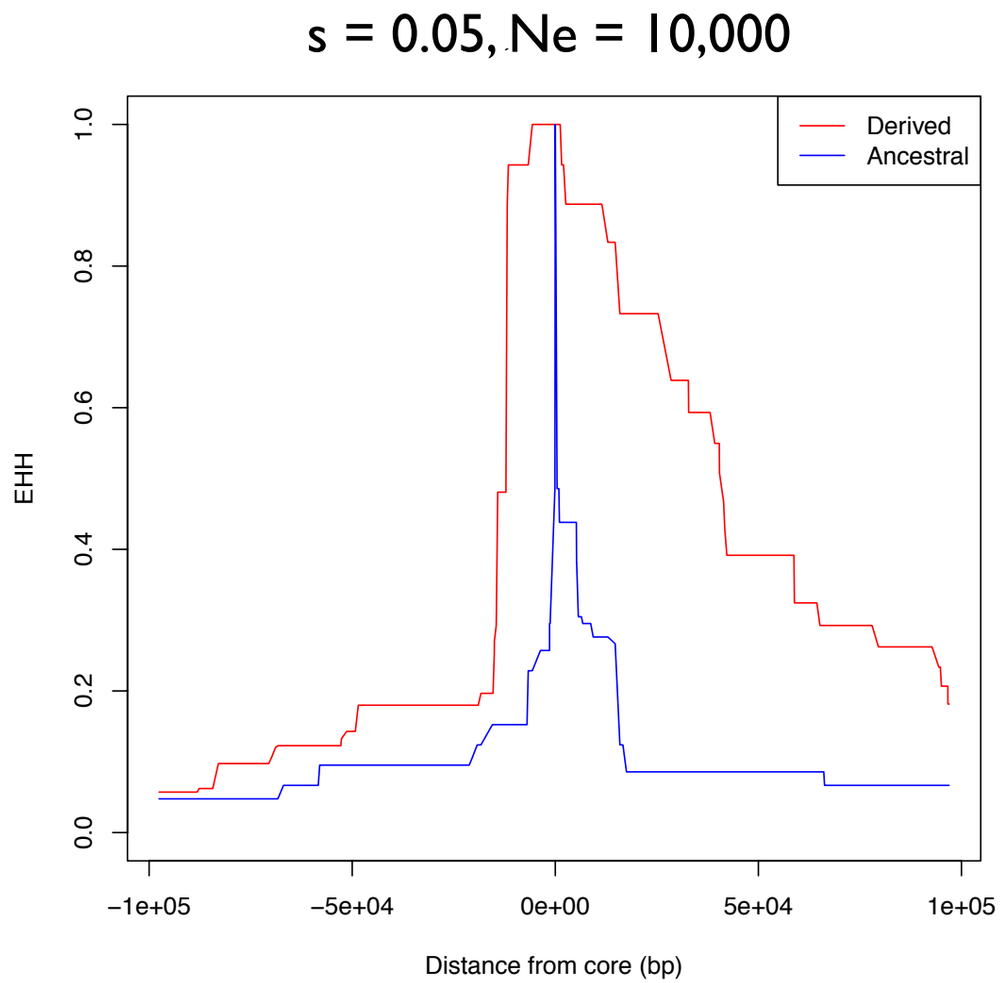
- The EHH approach does not lend itself to a genome-wide scan.

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

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

iHS

- If neutral, ancestral and derived EHH curves should have equal area.
- If a haplotype is positively selected, this curve should have larger area.



iHS

- Let the area under the ancestral haplotype EHH curve be iHH_A and the area under the derived haplotype EHH curve be iHH_D
- Then we define (unstandardized) iHS to be $\ln\left(\frac{iHH_A}{iHH_D}\right)$

iHS

$$\ln\left(\frac{iHH_A}{iHH_D}\right) < 0$$

iHS

$$\ln\left(\frac{iHH_A}{iHH_D}\right) < 0 \longrightarrow$$

Derived haplotype unusually long

iHS

$$\ln\left(\frac{iHH_A}{iHH_D}\right) < 0 \longrightarrow$$

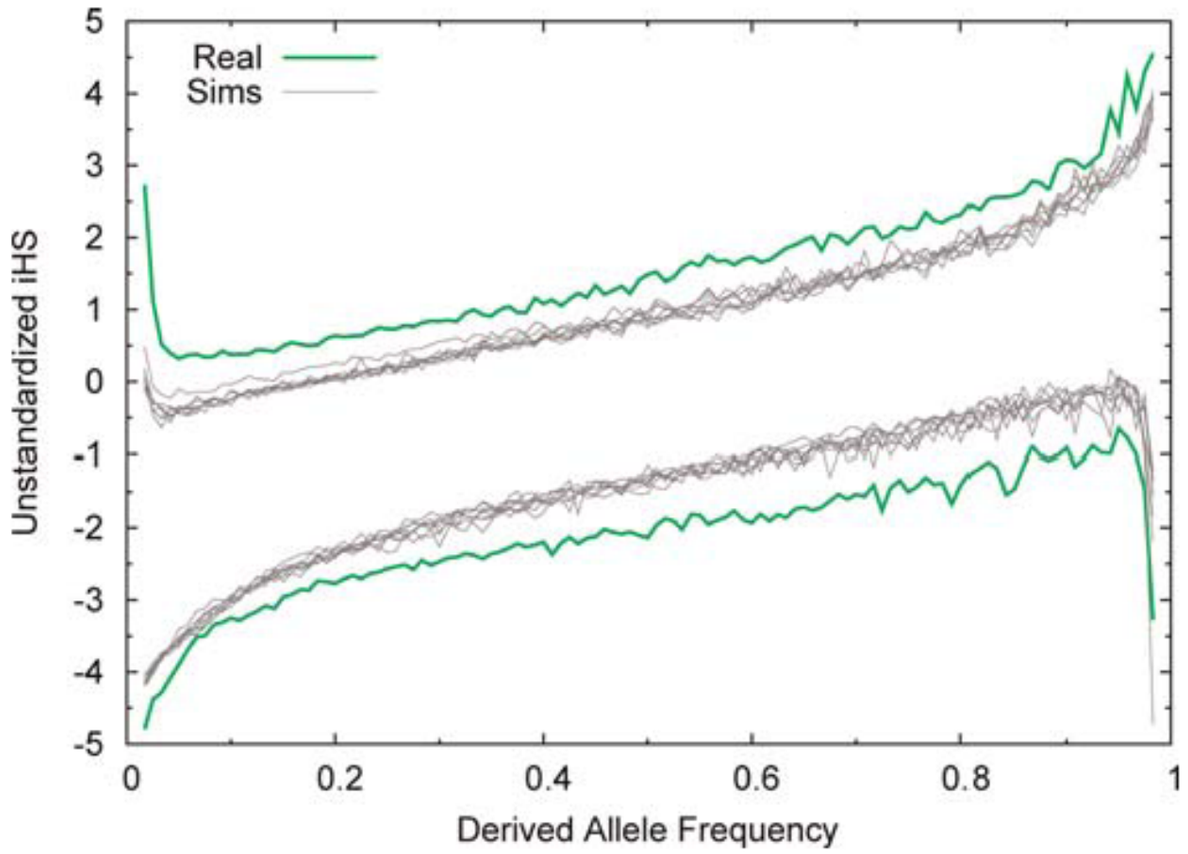
Derived haplotype unusually long

$$\ln\left(\frac{iHH_A}{iHH_D}\right) > 0 \longrightarrow$$

Ancestral haplotype unusually long

iHS

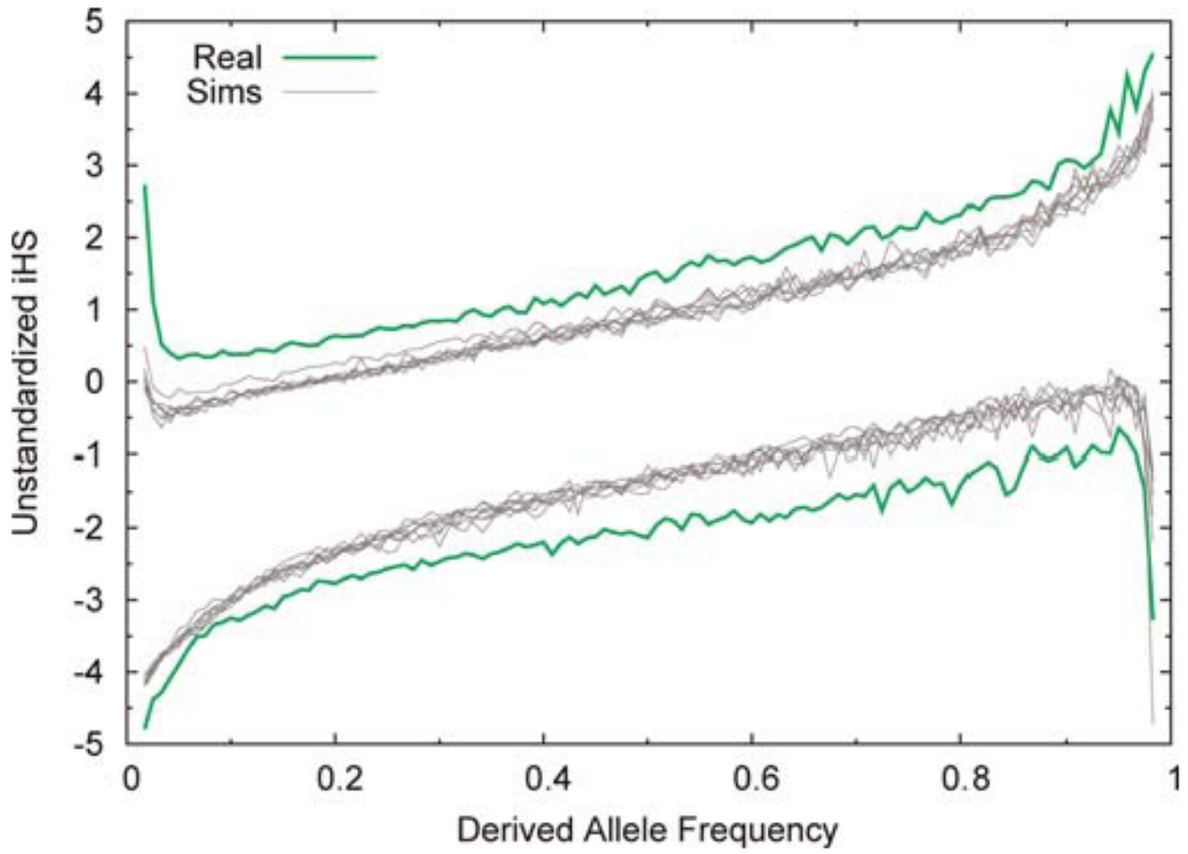
- Unstandardized iHS is correlated with allele frequency.
- Low frequency variants tend to be younger and therefore reside on longer haplotypes.



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

iHS

- Unstandardized iHS is correlated with allele frequency.
- Low frequency variants tend to be younger and therefore reside on longer haplotypes.



$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}$$

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

iHS

- In theory, we would want to search for strong negative iHS scores.
- In practice, ancestral alleles may be linked to the true beneficial allele, and therefore we often consider |iHS|.

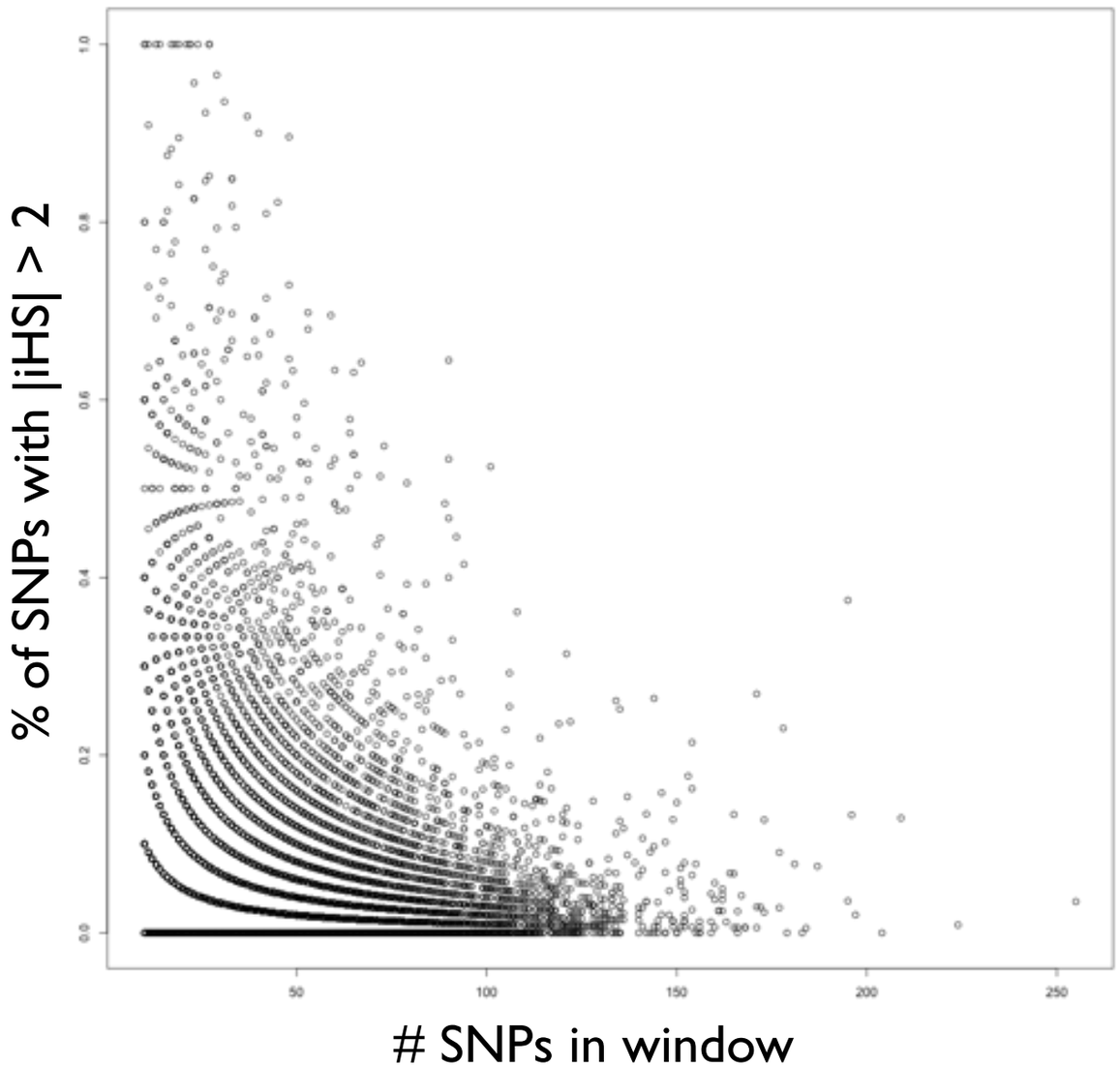
iHS

- Although large |iHS| values are possible even under neutrality, Voight, et al. found that these tend to occur uniformly across the genome.
- Under positive selection, large |iHS| values tended to cluster near the beneficial locus.

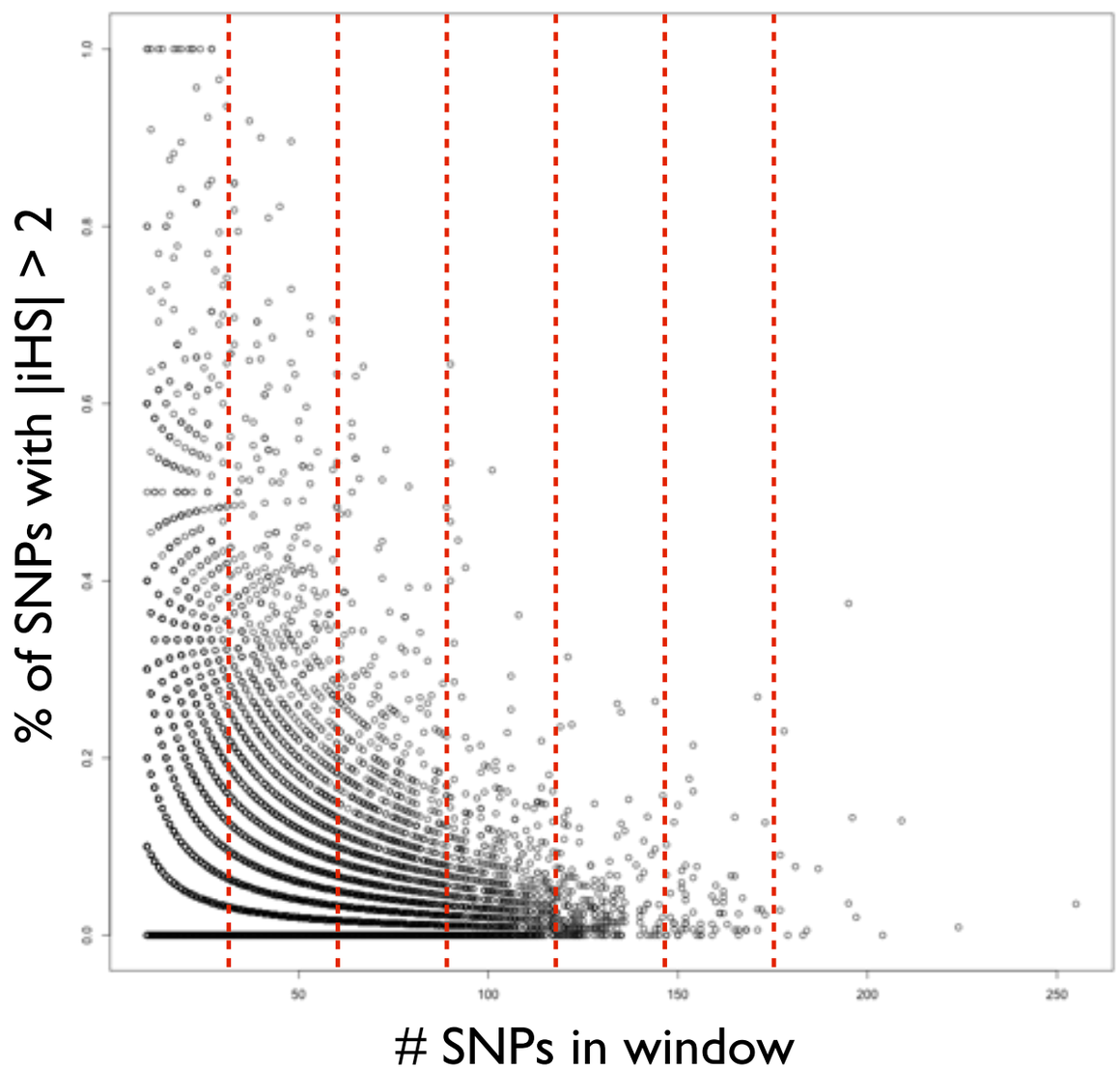
- Consider the fraction of SNPs with |iHS| > 2 in 51 SNP windows
  - Take the top 1% of windows
- Alternatively, consider fixed 100 kb windows across the genome

- Because of correlation, we split windows into bins based on # SNPs

- Take top 1% from within each bin



- Because of correlation, we split windows into bins based on # SNPs
- Take top 1% from within each bin



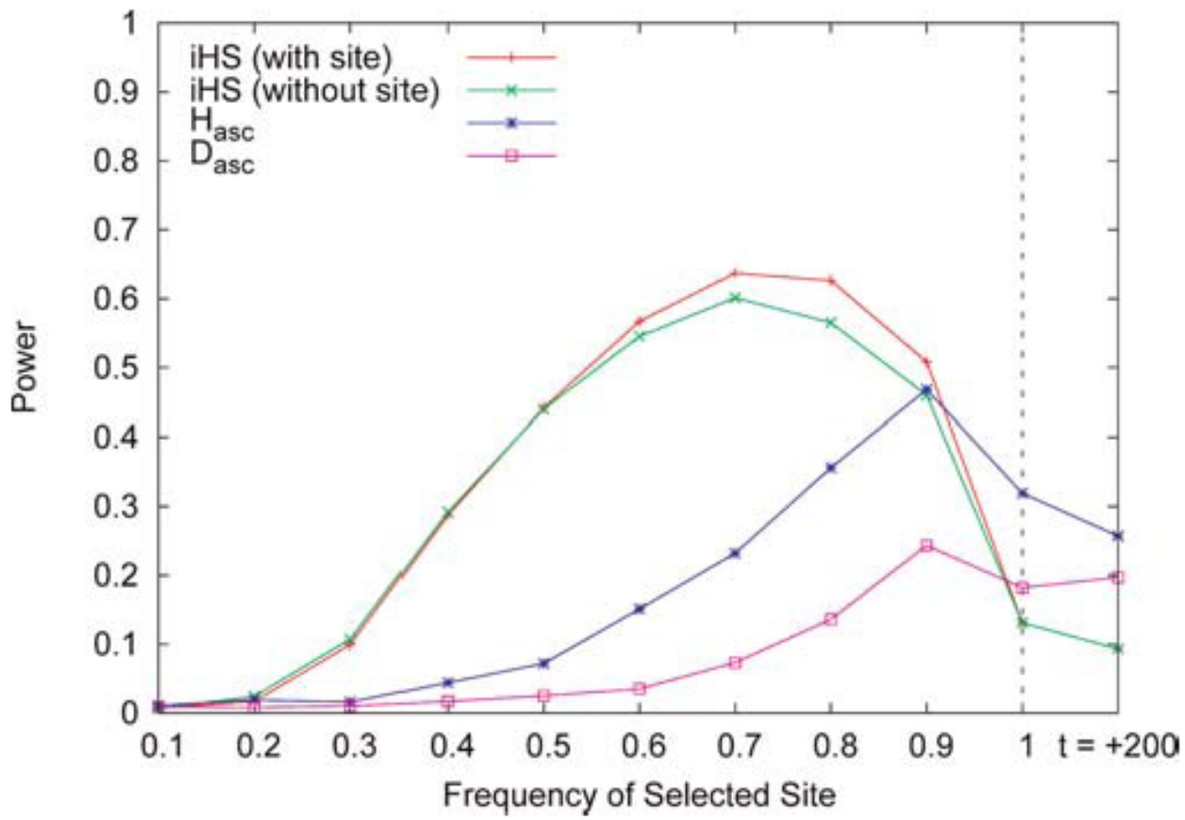
iHS

- In order to identify putative regions under positive selection, take the top 1% of windows.

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

iHS

- In order to identify putative regions under positive selection, take the top 1% of windows.



$$s = 0.0075, N_e = 10,000$$

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

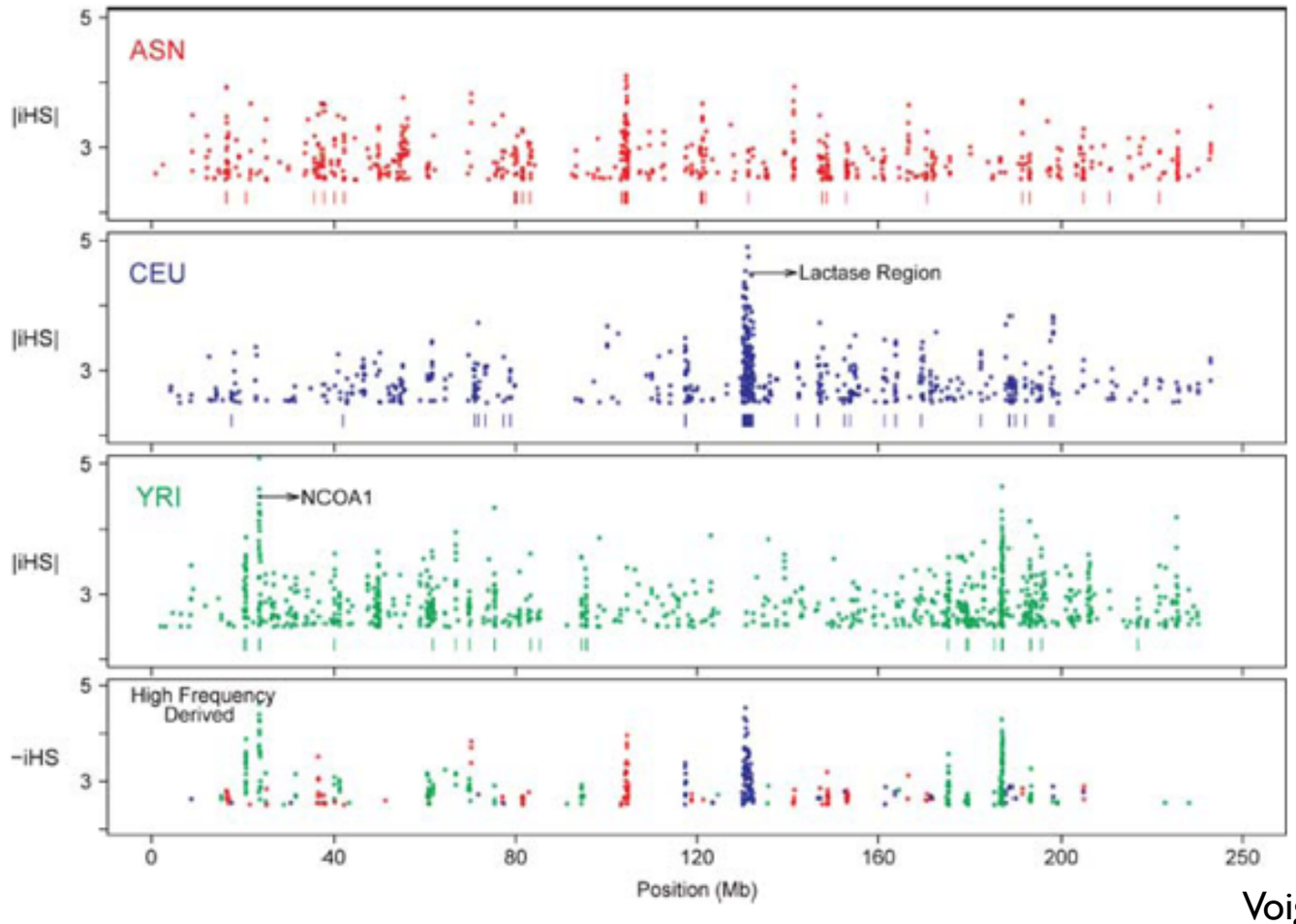
iHS

- Voight, et al. scan ~800k markers in three populations (ASN, CEU, YRI).

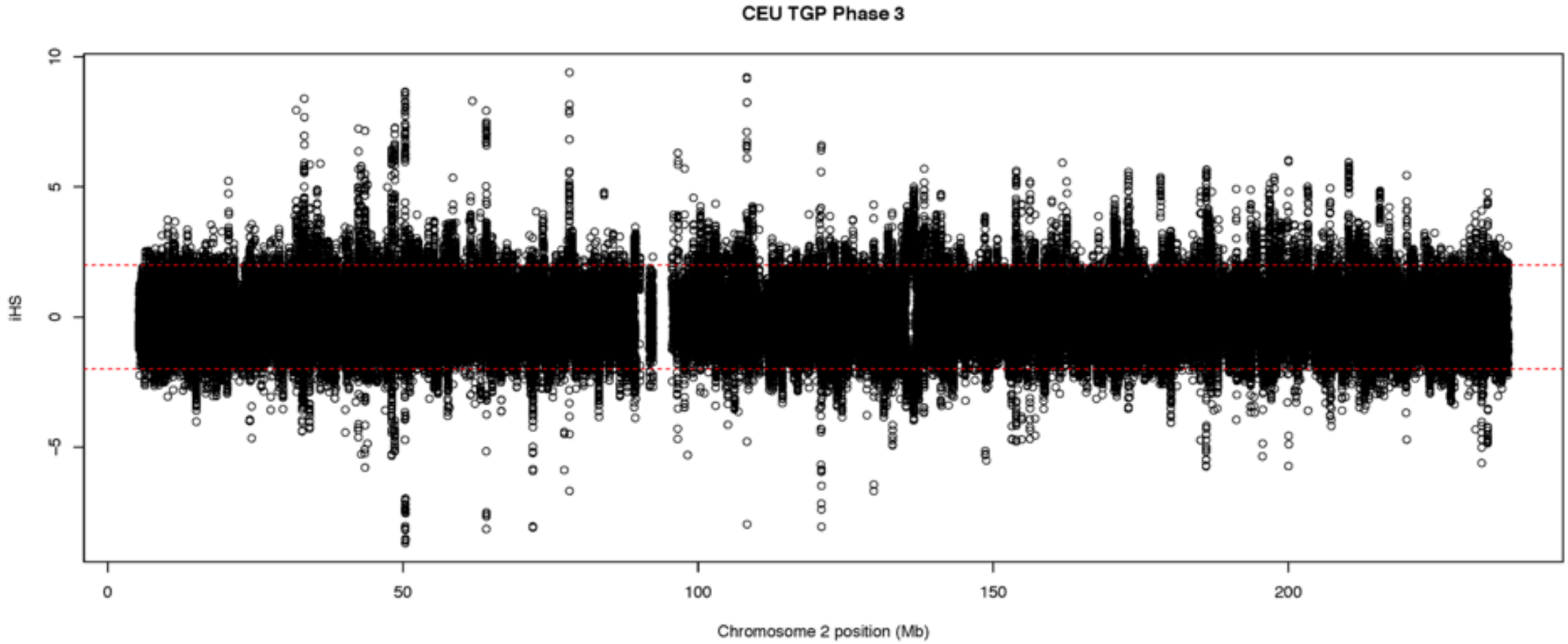
Voight, et al. (2006) *PLoS Biology*

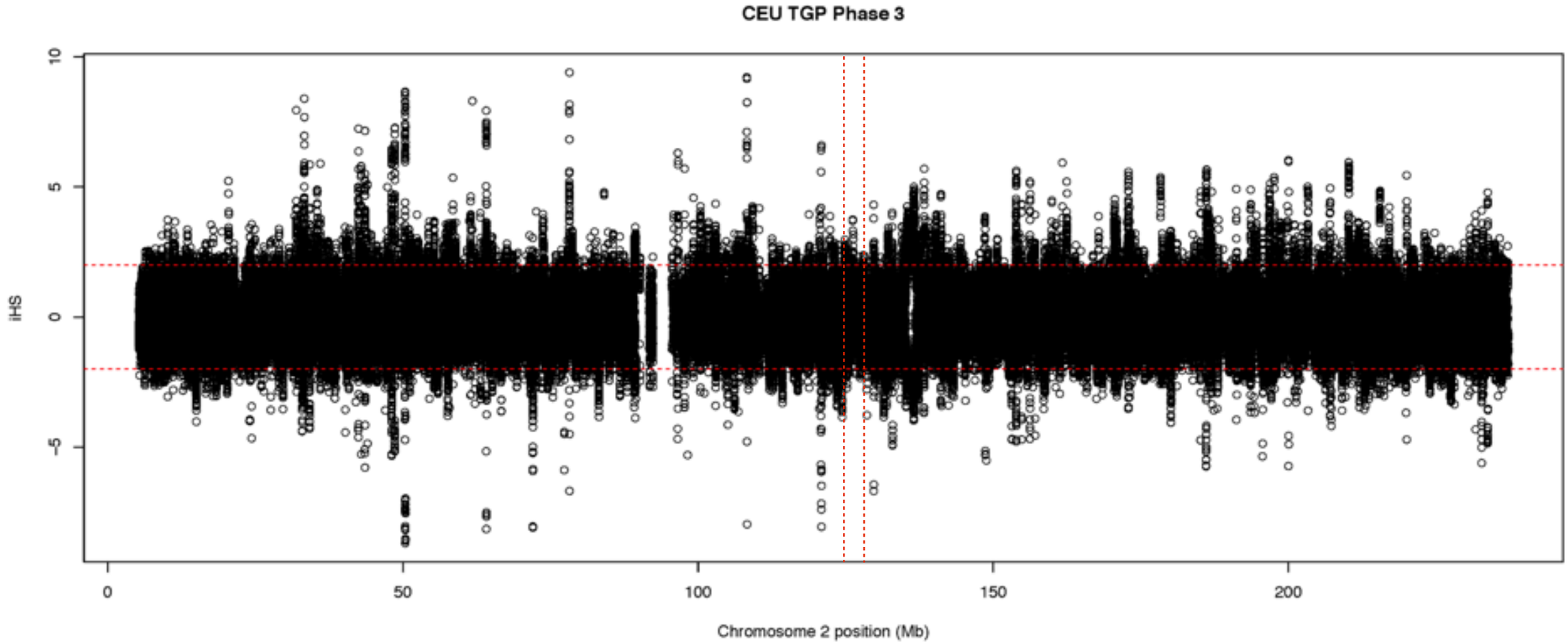
iHS

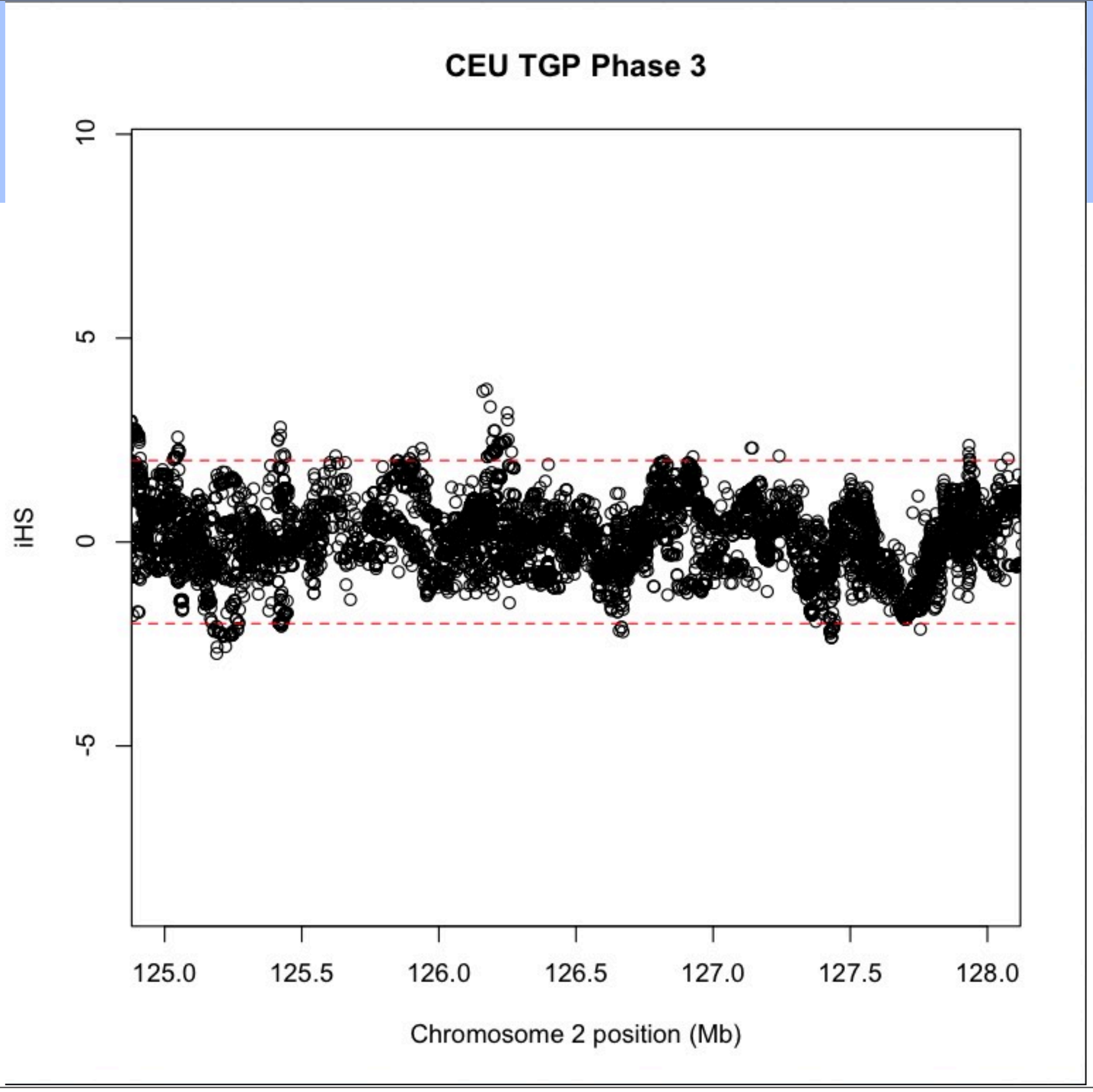
- Voight, et al. scan ~800k markers in three populations (ASN, CEU, YRI).

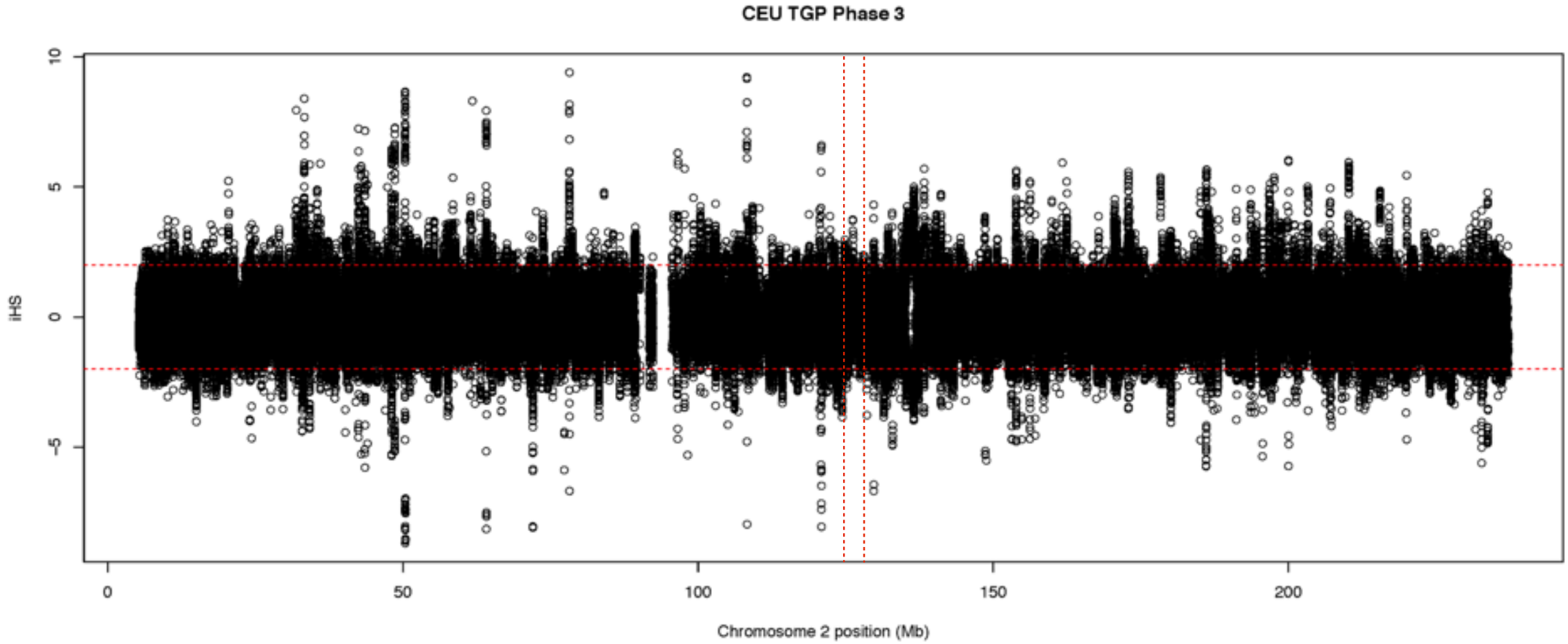


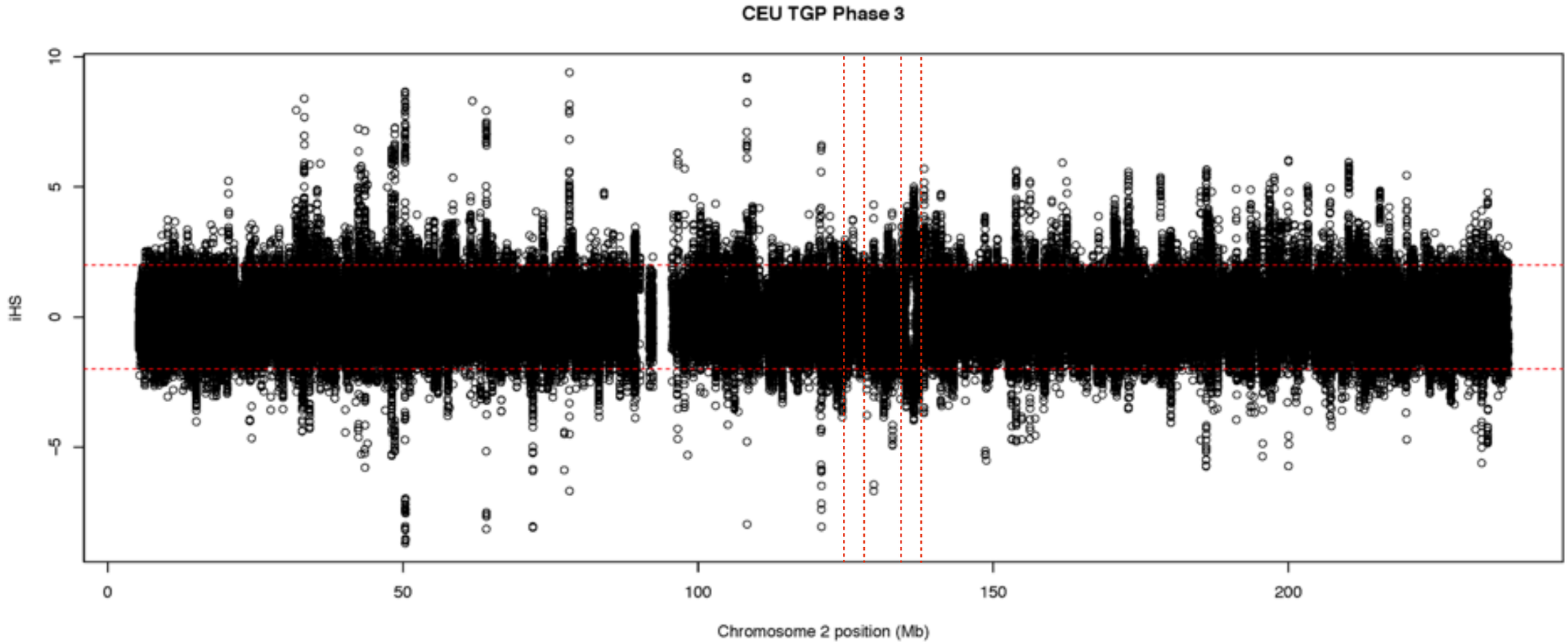
Voight, et al. (2006) *PLoS Biology*

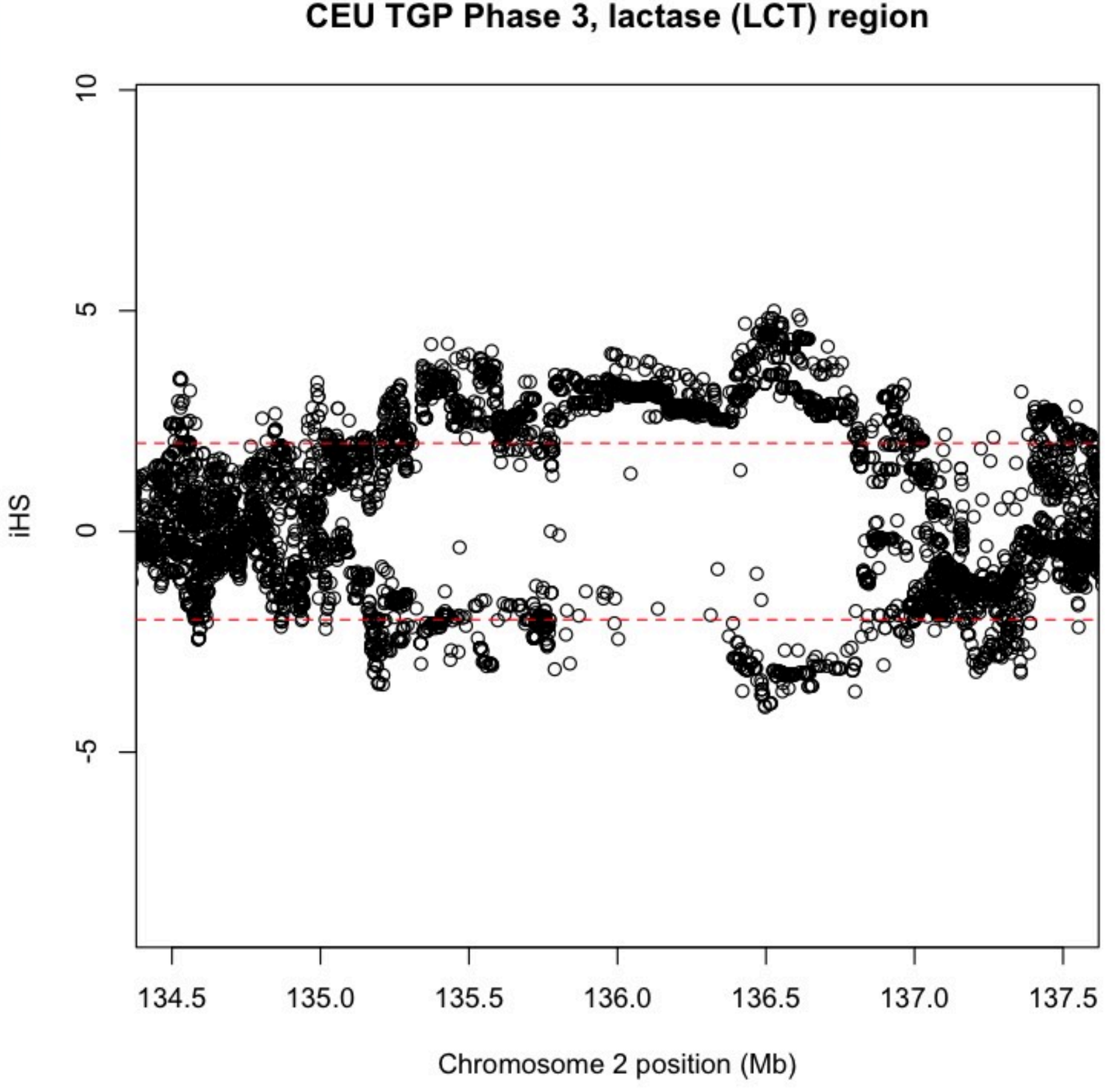

CEU TGP Phase 3
iHS
10
5
0
−5
0
50
100
150
200
Chromosome 2 position (Mb)


CEU TGP Phase 3
iHS
10
5
0
−5
0
50
100
150
200
Chromosome 2 position (Mb)


CEU TGP Phase 3
iHS
10
5
0
-5
125.0
125.5
126.0
126.5
127.0
127.5
128.0
Chromosome 2 position (Mb)


CEU TGP Phase 3
iHS
10
5
0
−5
0
50
100
150
200
Chromosome 2 position (Mb)


CEU TGP Phase 3
iHS
10
5
0
−5
0
50
100
150
200
Chromosome 2 position (Mb)


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)

XP-EHH

- Sabetti, et al. (2007) develop XP-EHH as a modification to iHS.
- XP-EHH compares EHH decay between populations.
- It seeks to discover variants near/at fixation on long haplotypes in one population but remains polymorphic in others.

Sabetti, et al. (2007) *Nature*

XP-EHH

- iHS compares ancestral vs. derived EHH decay in the same population.
- XP-EHH compares EHH decay at the same locus between two populations.
- Note that EHH in a population does *not* necessarily start at 1.
  - Only if the starting site is fixed in the sample of that population

XP-EHH

iHH_A

XP-EHH

iHH_A ⟵ Integrated EHH in population A

XP-EHH

iHH_A ⟵ Integrated EHH in population A

iHH_B ⟵ Integrated EHH in population B

XP-EHH

iHH_A ⟵ Integrated EHH in population A

iHH_B ⟵ Integrated EHH in population B

$$\ln\left(\frac{iHH_A}{iHH_B}\right) < 0$$ ⟵ Unusually long haplotypes in population B

XP-EHH

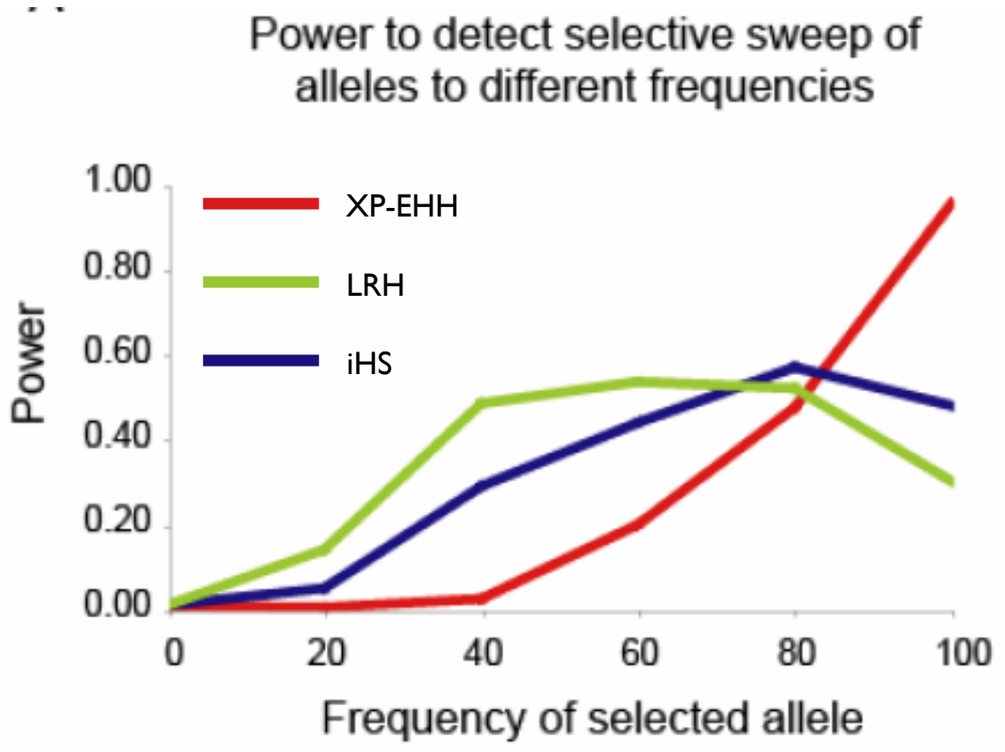
iHH_A ⟵ Integrated EHH in population A

iHH_B ⟵ Integrated EHH in population B

$\ln\left(\frac{iHH_A}{iHH_B}\right) < 0$ ⟵ Unusually long haplotypes in population B

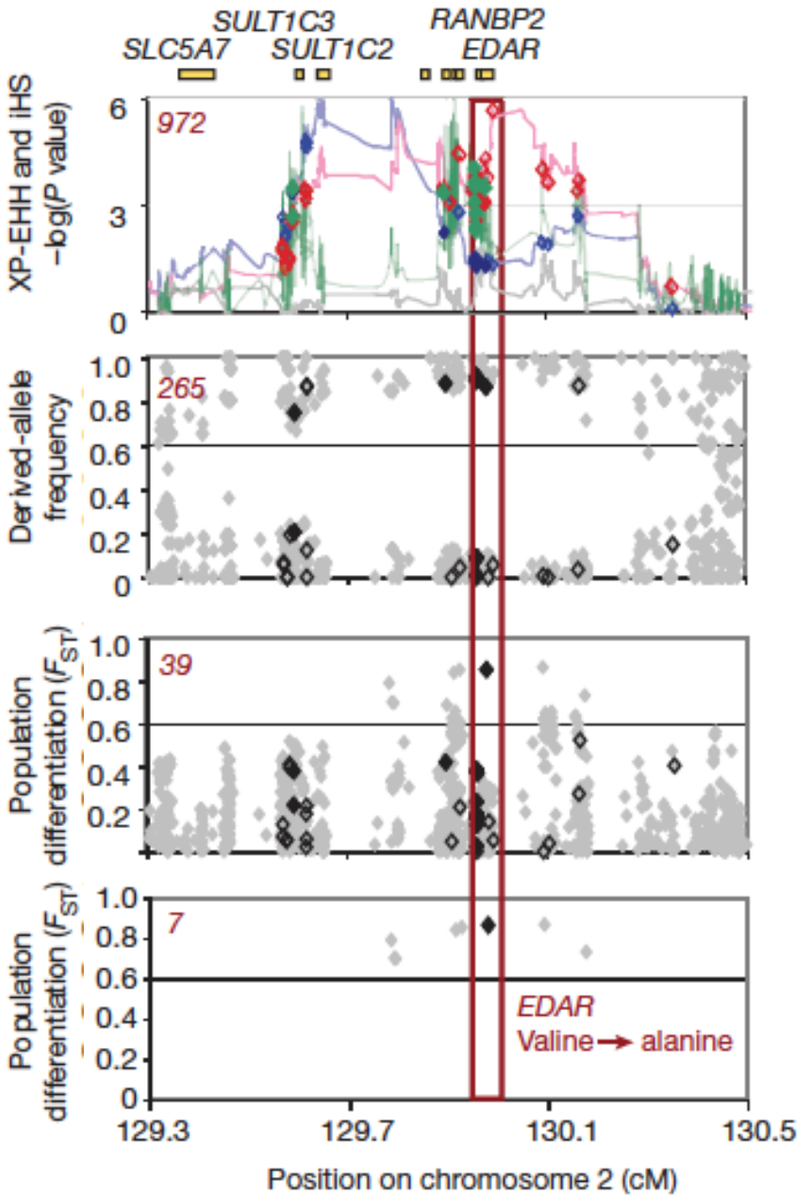
$\ln\left(\frac{iHH_A}{iHH_B}\right) > 0$ ⟵ Unusually long haplotypes in population A

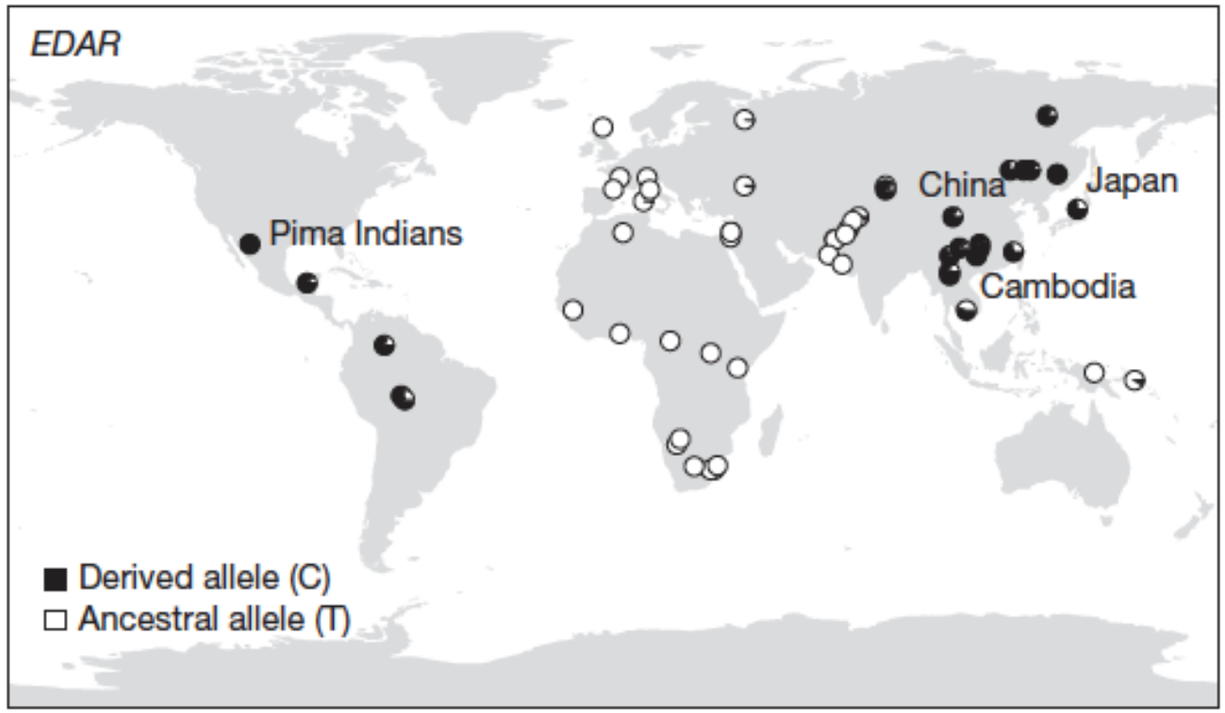
XP-EHH



Sabetti, et al. (2007) *Nature*

XP-EHH





Sabetti, et al. (2007) *Nature*

EDAR

- They follow up in a mouse model, knock-in EDAR V370A
- Increased hair thickness
- Higher number of active eccrine glands
- Temperature and humidity as selective forces?

Kamberov, et al. (2013) *Cell*

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!
- 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:**
    - `cp ..\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.hap
--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:
```

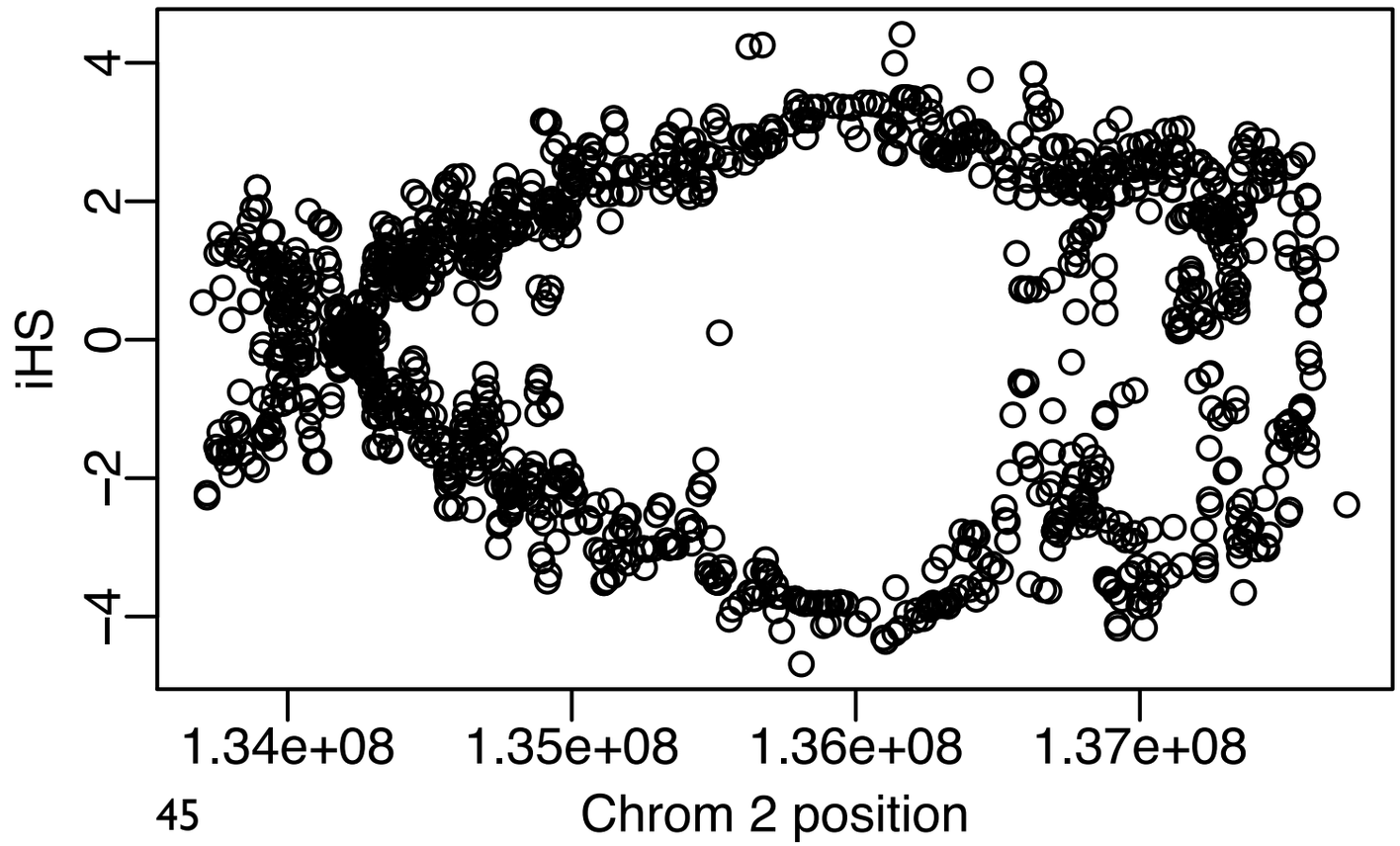
iHS

- 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

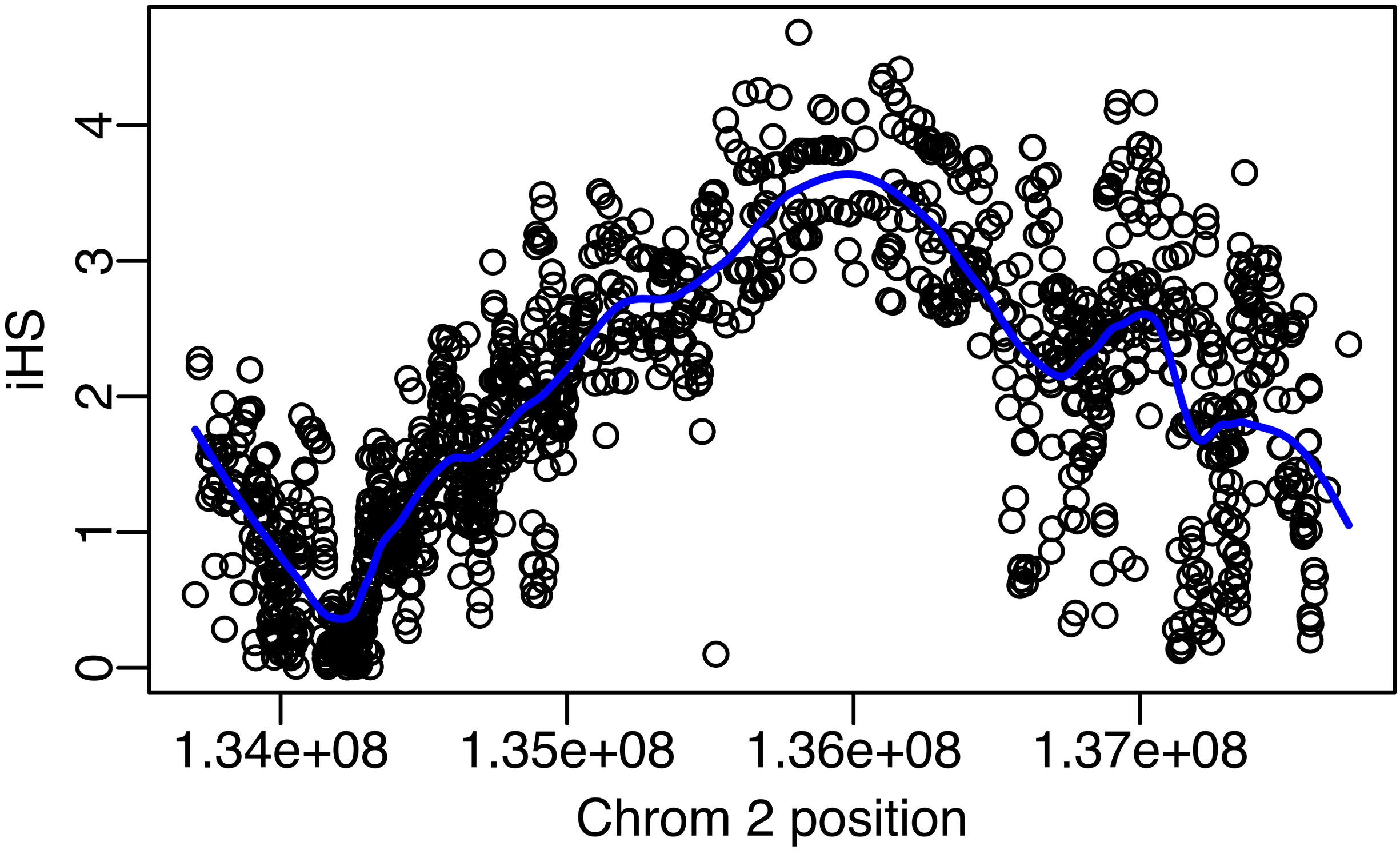
CEU.loess=loess(y~CEU.x,span=SP,data.frame(x=CEU.x,y=y)); #step 1

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


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

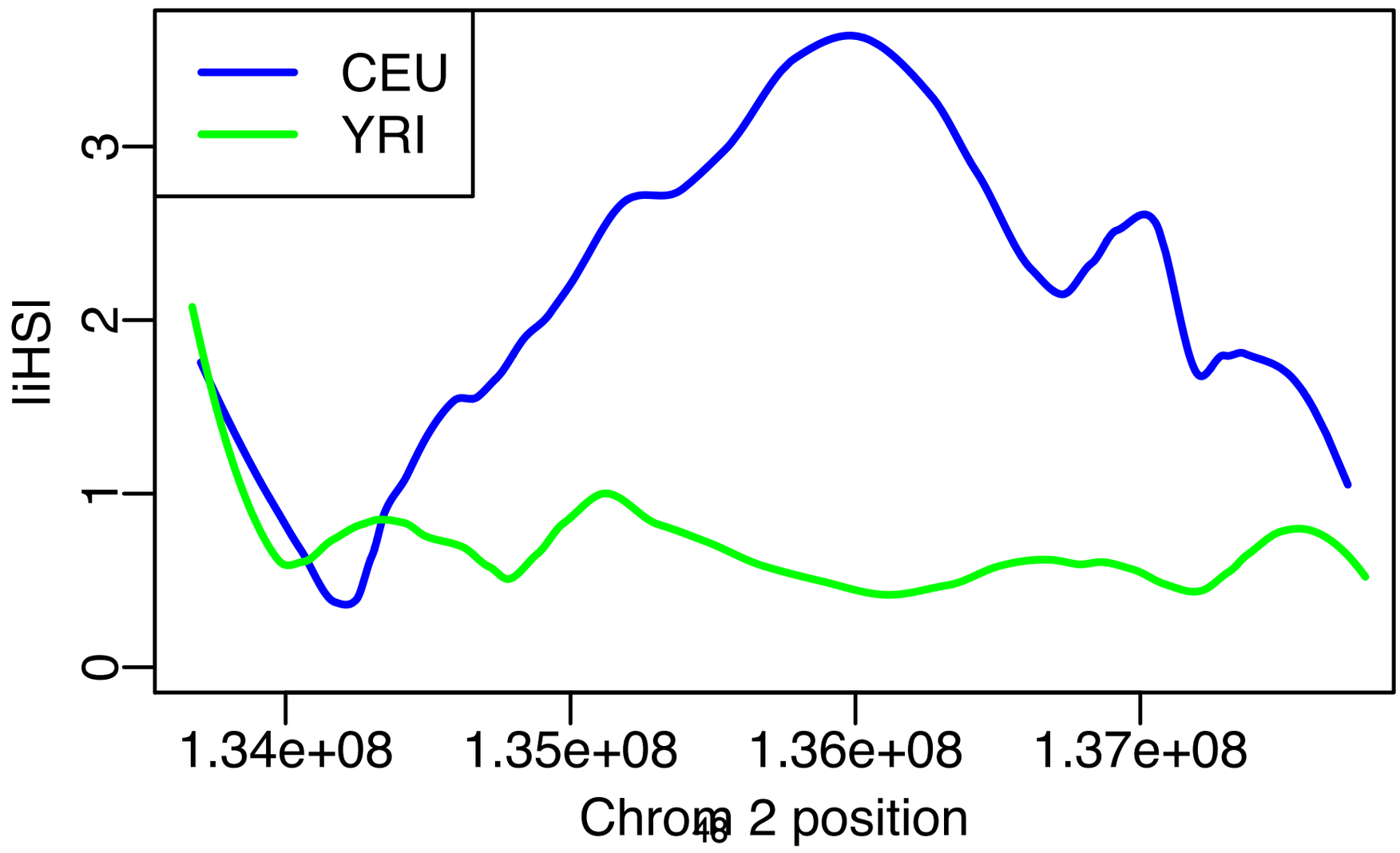
lines(CEU.x, CEU.predict, lwd=2, col='blue')
```

iHS



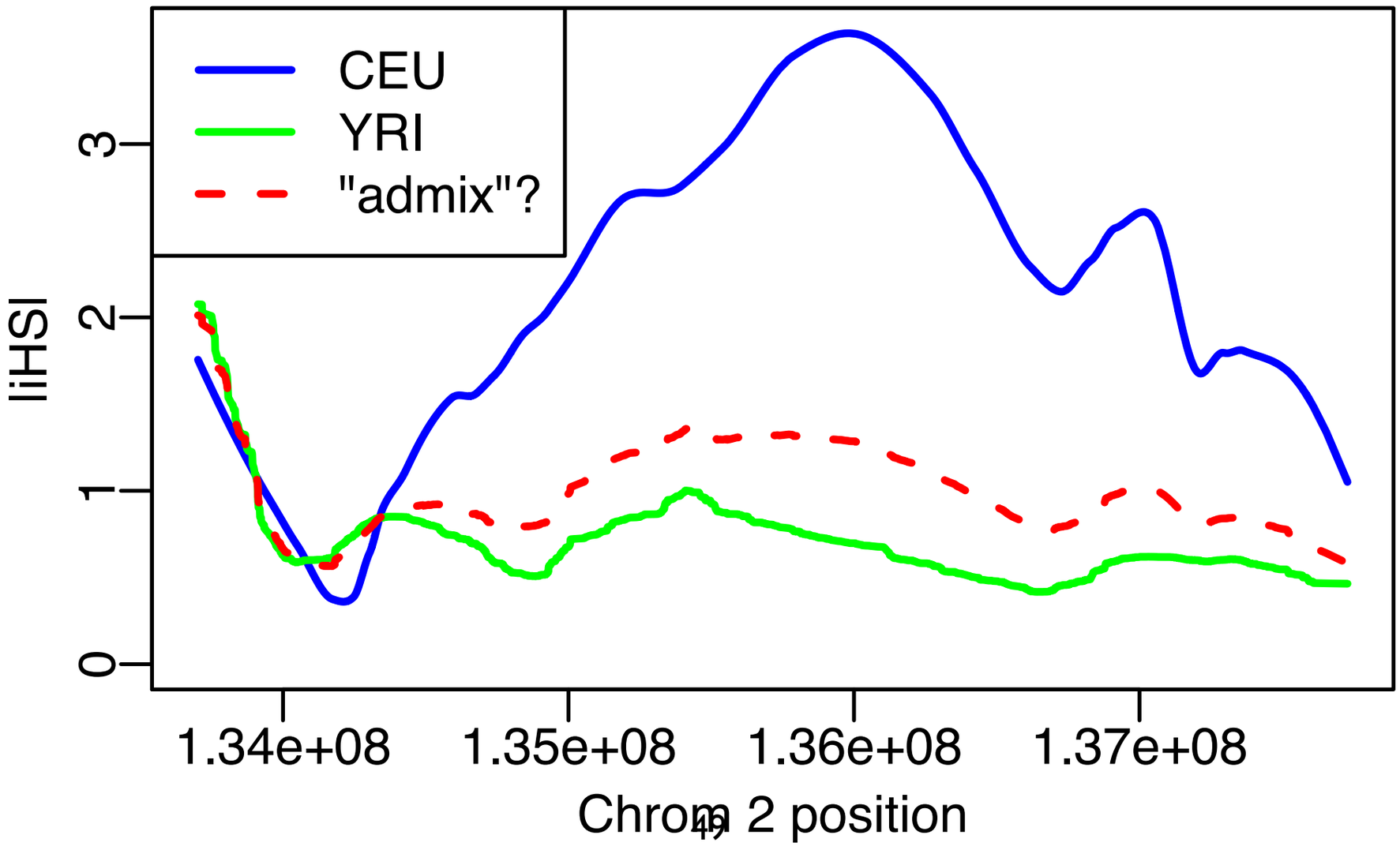
Other populations??

- 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"



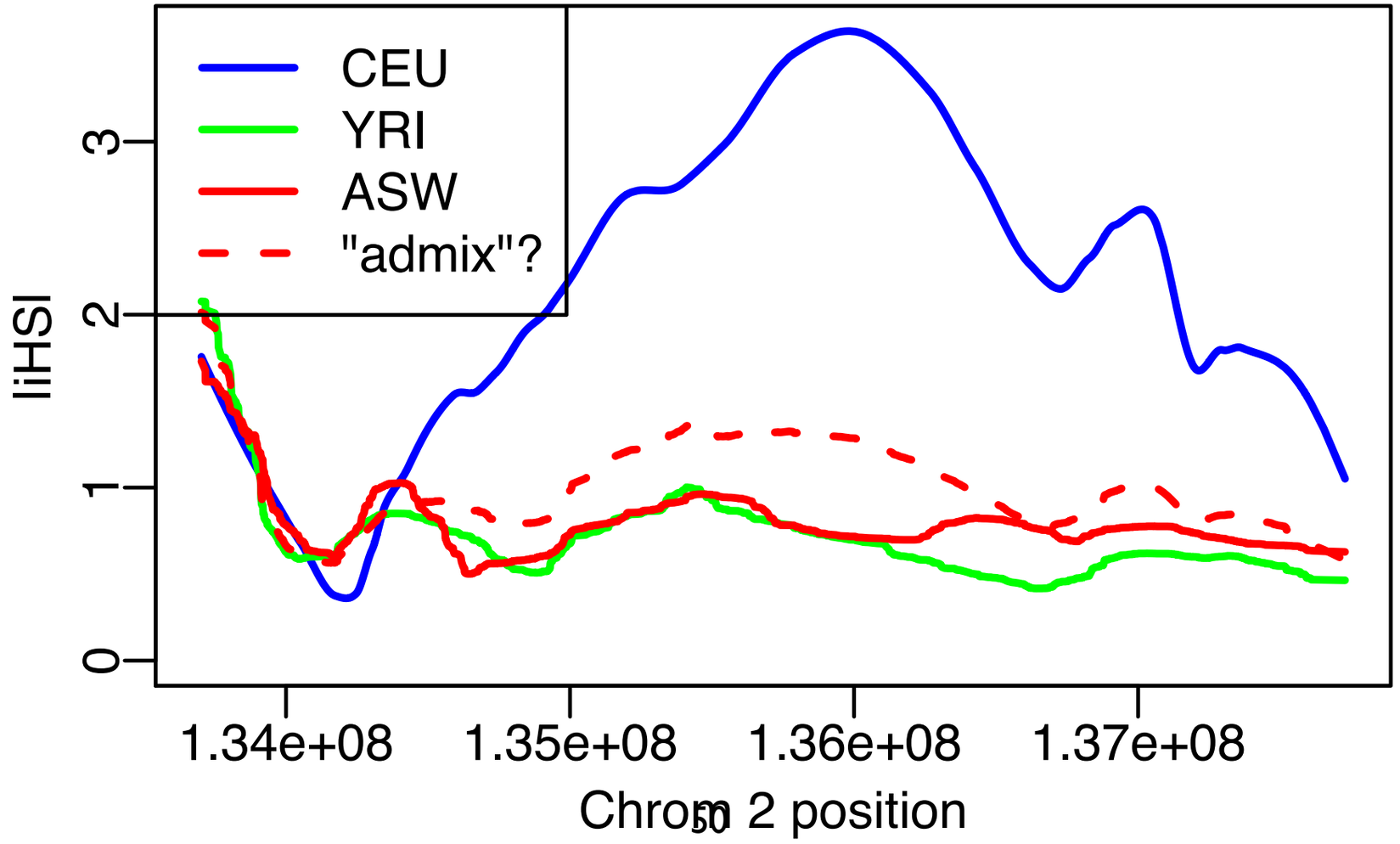
What about admixture?

- 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.