Detecting Selective Sweeps

Alan R. Rogers

March 31, 2021

Outline

- Questions
  - Have humans evolved rapidly or slowly during the past 40 kyr?
  - What functional categories of gene have evolved most?
- Selection and recombination
- Data
- Results

Are we still evolving?

- Since split with chimps, there has been rapid evolution in proteins expressed in brain and in sperm.

Are we still evolving?

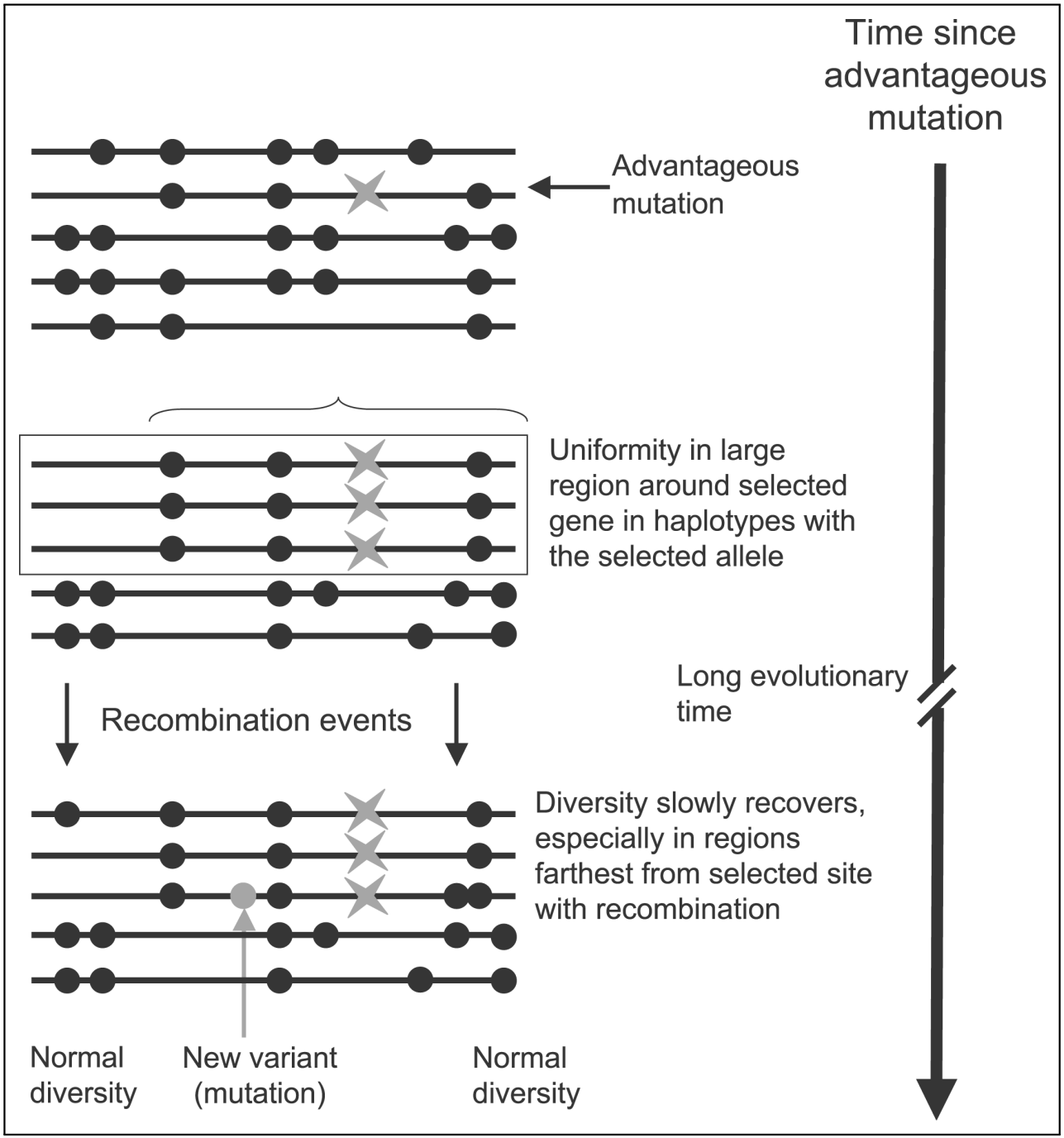
- Since split with chimps, there has been rapid evolution in proteins expressed in brain and in sperm.
- Recent selection at various loci: lactase, DRD4, etc

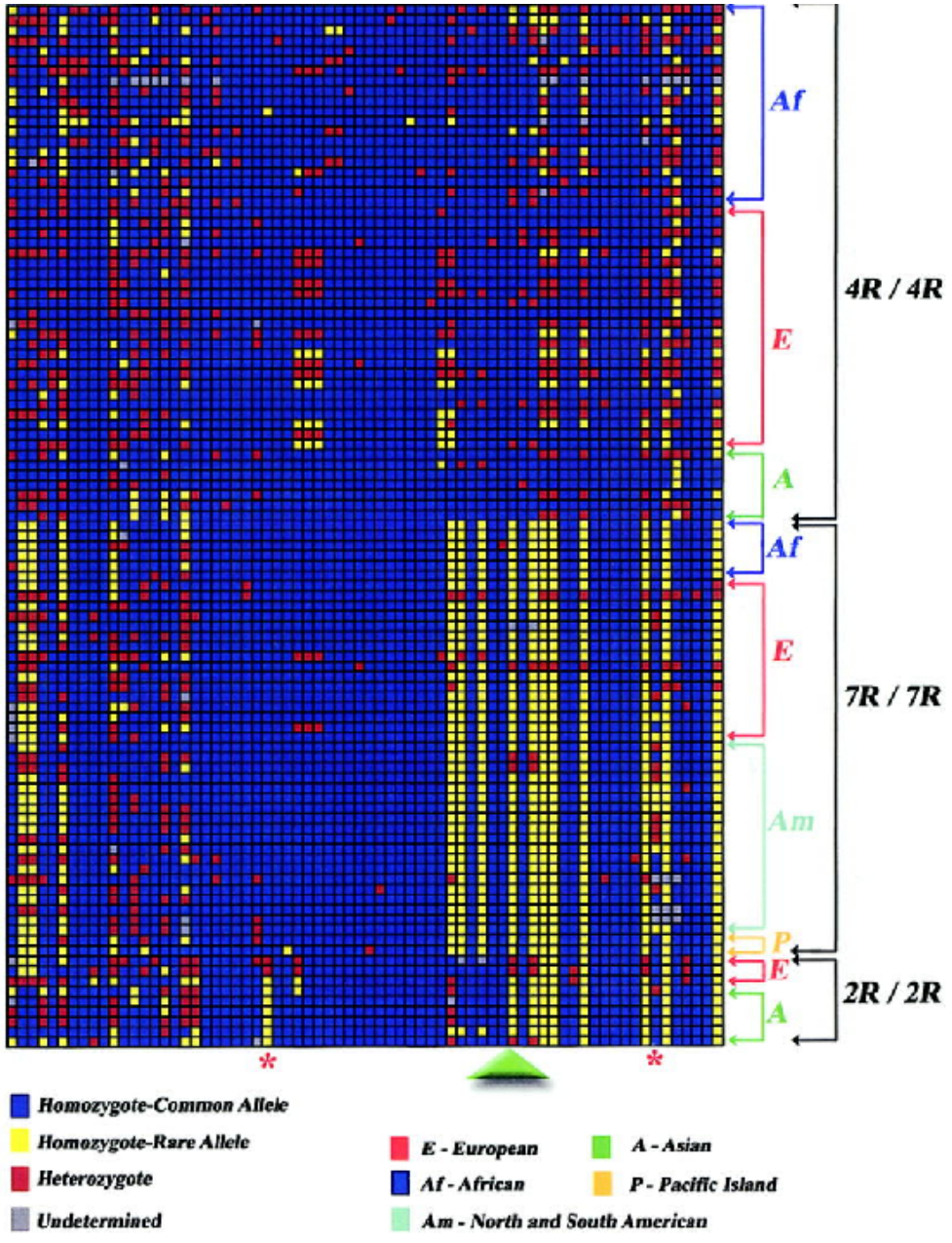
Are we still evolving?

- Since split with chimps, there has been rapid evolution in proteins expressed in brain and in sperm.
- Recent selection at various loci: lactase, DRD4, etc
- How common are such loci in the human genome?

Are we still evolving?

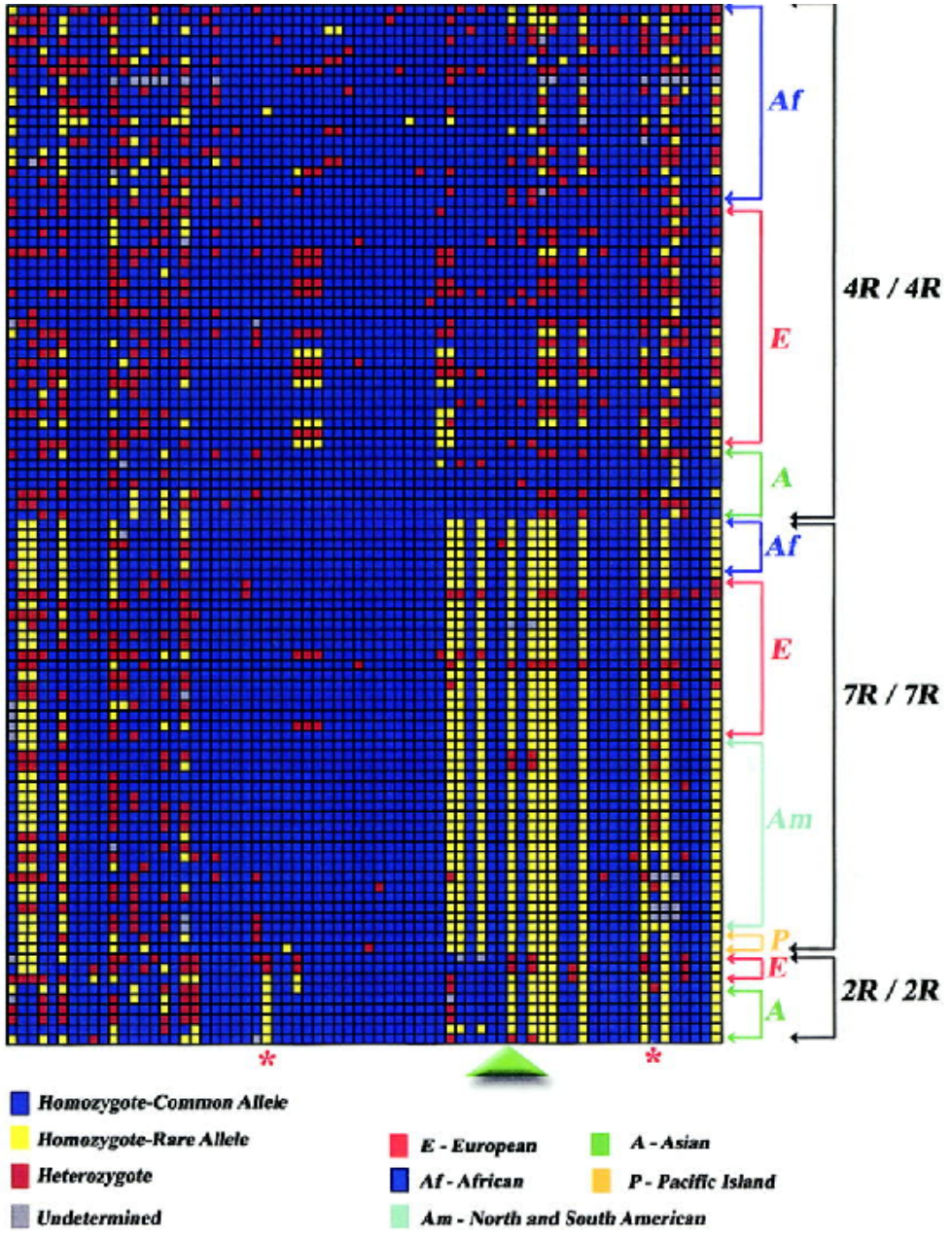
- Since split with chimps, there has been rapid evolution in proteins expressed in brain and in sperm.
- Recent selection at various loci: lactase, DRD4, etc
- How common are such loci in the human genome?
- How can we tell?


Time since advantageous mutation
Advantageous mutation
Uniformity in large region around selected gene in haplotypes with the selected allele
Long evolutionary time
Recombination events
Diversity slowly recovers, especially in regions farthest from selected site with recombination
Normal diversity
New variant (mutation)
Normal diversity



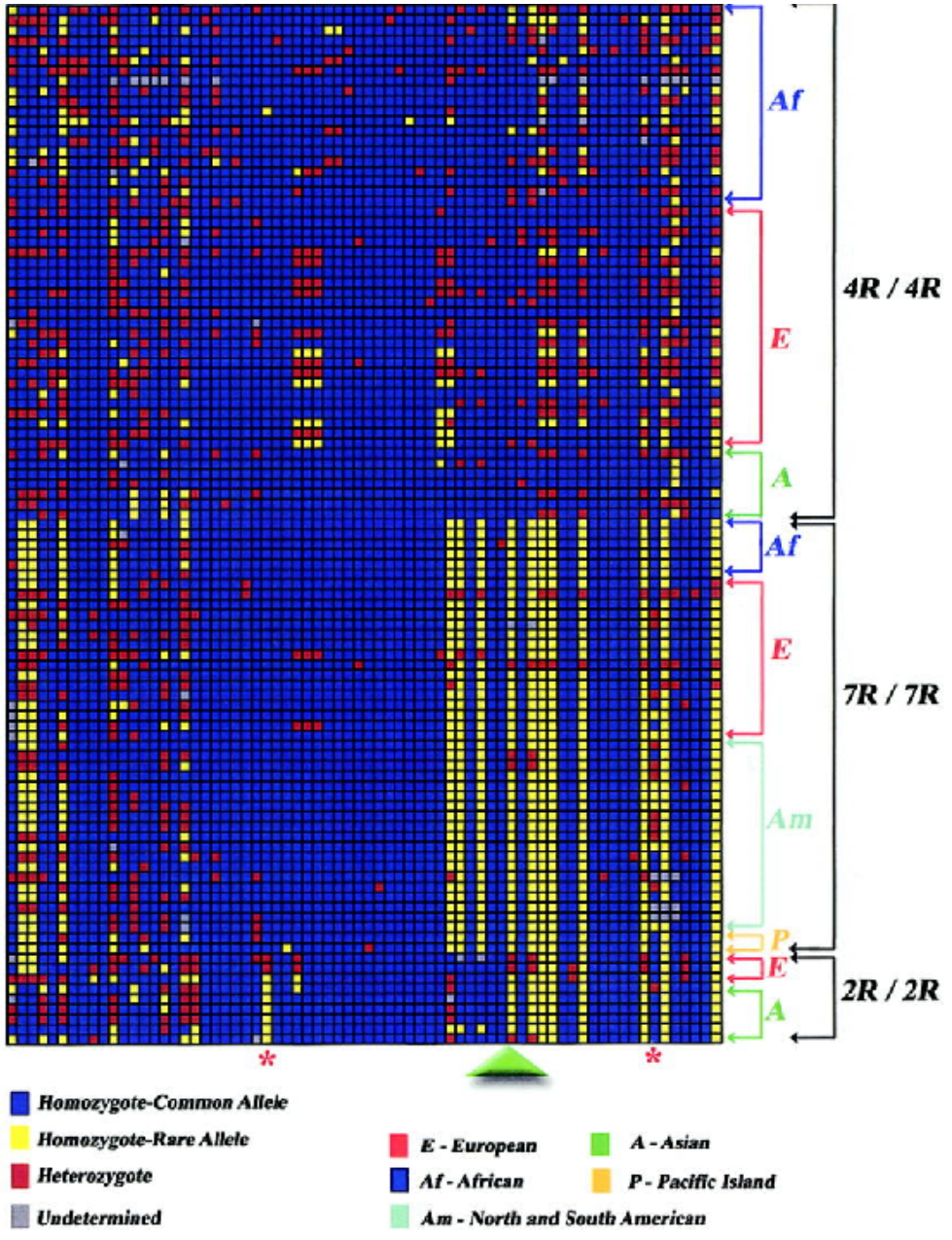
Signature of an ongoing selective sweep at DRD4

- Sweeping allele is
  - common



Signature of an ongoing selective sweep at DRD4

- Sweeping allele is
  - common
  - has low diversity over large region



Signature of an ongoing selective sweep at DRD4

- Sweeping allele is
  - common
  - has low diversity over large region
- High LD over large region

Finding sweeping alleles

- Extended haplotype homozygosity (EHH): Voight et al 2006

Idea behind Extended Haplotype Homozygosity (EHH)

```
   ctaaacagaccaacgtaAgggtacaatgcctaacccagacgttt
20 .........................................
21 .........................................
22 .........................................
23 .........................................
24 .........................................
25 .........................................
26 .........................................
27 t........................................
28 t........................................
29 ...............................c.........
37 ................G..a.gt.....t.........gac.c
38 t..gg..c.....tc.gGaaa.g..ccttt...tg......c..
39 t..gg..c.....tc.gGaaa.g..ccttt...tg......c..
40 t..g.....t..ttccgG..a.gt.....t.........gac.c
41 t..g.....t.gttccgG..a.gt.....t.........gac.c
44 t..g.....t..ttc.gG..acgt.....t.........gac.c
45 t..g.....t.gttc.gG..a.gt.....t.........gac.c
46 t..gg..c.....tc.gGaaa.g..ccttt...tg......cg.
47 t..g.....t.gttccgG..a.gt.....t.........gac.c
48 t..g.....t.gttccgG..a.gt.....t.........gac.c
49 t..g.....t.gttccgG..a.gt.....t.........gac.c
50 tcgg.tc.g.tg.tc.gG..a.g.g....tg....ggt...cg.
51 tcgg.tc.g.tg.tc.gG..a.g.g....tg....ggt...cg.
```

Upper half: pairs of chromosomes are identical at most sites

Idea behind Extended Haplotype Homozygosity (EHH)

```
   ctaaacagaccaacgtaAgggtacaatgcctaacccagacgttt
20 .........................................
21 .........................................
22 .........................................
23 .........................................
24 .........................................
25 .........................................
26 .........................................
27 t........................................
28 t........................................
29 ...............................c.........
37 ................G..a.gt.....t.........gac.c
38 t..gg..c.....tc.gGaaa.g..ccttt...tg......c..
39 t..gg..c.....tc.gGaaa.g..ccttt...tg......c..
40 t..g.....t..ttccgG..a.gt.....t.........gac.c
41 t..g.....t.gttccgG..a.gt.....t.........gac.c
44 t..g.....t..ttc.gG..acgt.....t.........gac.c
45 t..g.....t.gttc.gG..a.gt.....t.........gac.c
46 t..gg..c.....tc.gGaaa.g..ccttt...tg......cg.
47 t..g.....t.gttccgG..a.gt.....t.........gac.c
48 t..g.....t.gttccgG..a.gt.....t.........gac.c
49 t..g.....t.gttccgG..a.gt.....t.........gac.c
50 tcgg.tc.g.tg.tc.gG..a.g.g....tg....ggt...cg.
51 tcgg.tc.g.tg.tc.gG..a.g.g....tg....ggt...cg.
```

Upper half: pairs of chromosomes are identical at most sites

Lower half: pairs identical at fewer sites

Idea behind Extended Haplotype Homozygosity (EHH)

```
   ctaaacagaccaacgtaAgggtacaatgcctaacccagacgttt
20 .........................................
21 .........................................
22 .........................................
23 .........................................
24 .........................................
25 .........................................
26 .........................................
27 t........................................
28 t........................................
29 ...............................c.........
37 ................G..a.gt.....t.........gac.c
38 t..gg..c.....tc.gGaaa.g..ccttt...tg......c..
39 t..gg..c.....tc.gGaaa.g..ccttt...tg......c..
40 t..g.....t..ttccgG..a.gt.....t.........gac.c
41 t..g.....t.gttccgG..a.gt.....t.........gac.c
44 t..g.....t..ttc.gG..acgt.....t.........gac.c
45 t..g.....t.gttc.gG..a.gt.....t.........gac.c
46 t..gg..c.....tc.gGaaa.g..ccttt...tg......cg.
47 t..g.....t.gttccgG..a.gt.....t.........gac.c
48 t..g.....t.gttccgG..a.gt.....t.........gac.c
49 t..g.....t.gttccgG..a.gt.....t.........gac.c
50 tcgg.tc.g.tg.tc.gG..a.g.g....tg....ggt...cg.
51 tcgg.tc.g.tg.tc.gG..a.g.g....tg....ggt...cg.
```

Upper half: pairs of chromosomes are identical at most sites

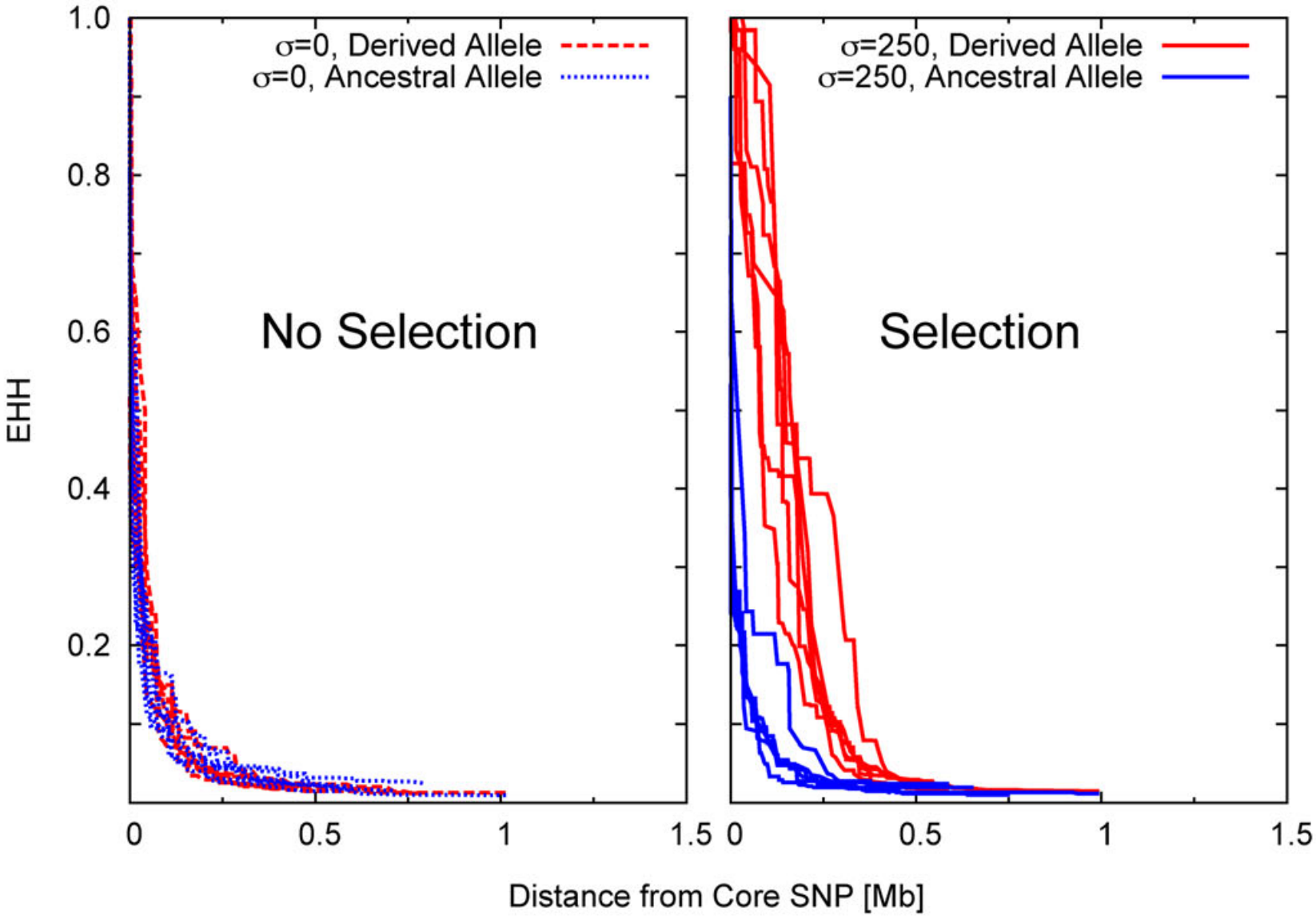
Lower half: pairs identical at fewer sites

This idea underlies EHH.

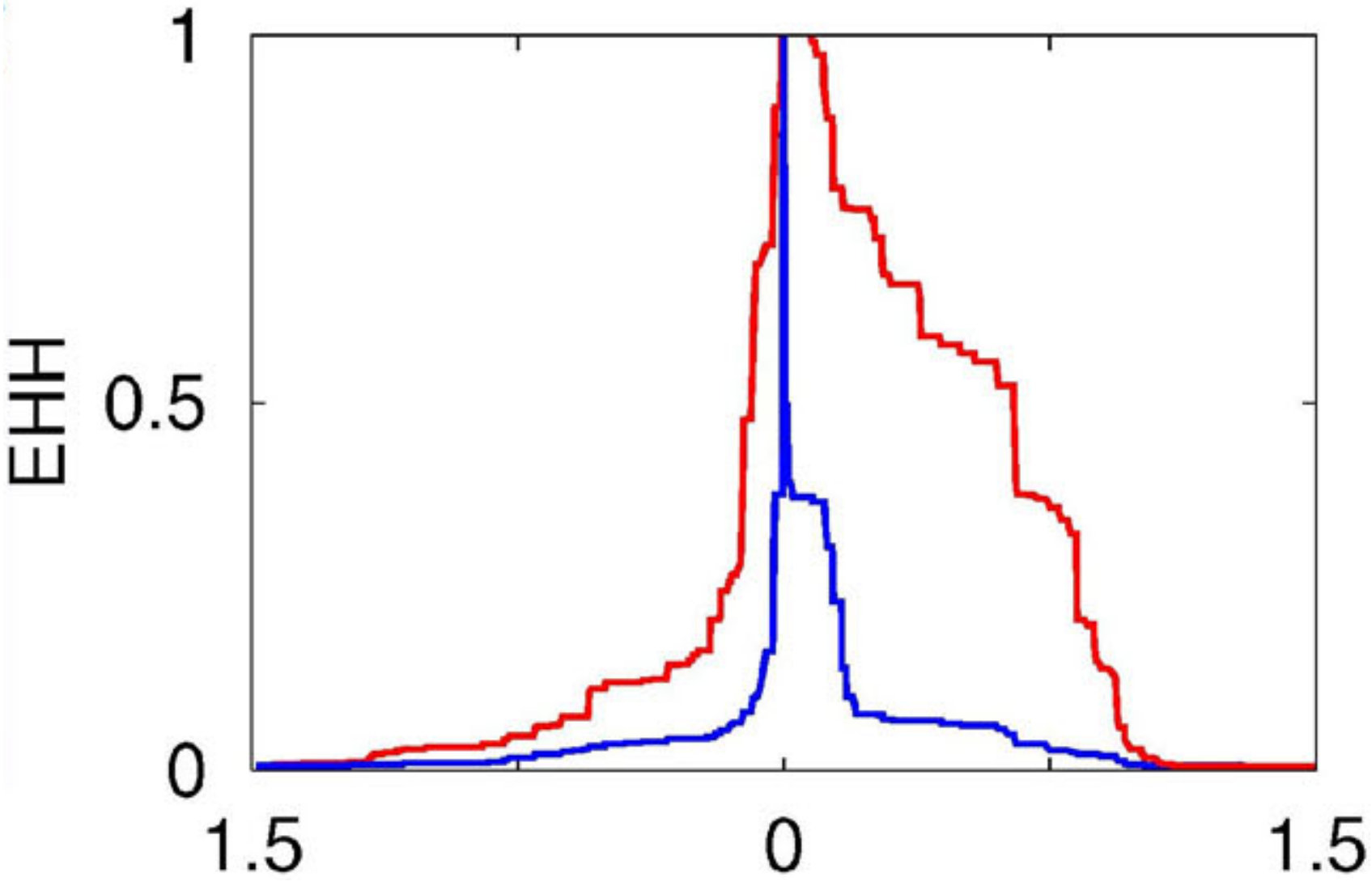
Extended haplotype homozygosity (EHH)

- Select chromosomes that carry allele A at focal site.
- Within this set, calculate the fraction of pairs that are identical at another site, x base-pairs away. This is EHH(x).
- Do the same for chromosomes that *don't* carry A.
- relative EHH is the ratio of the two.

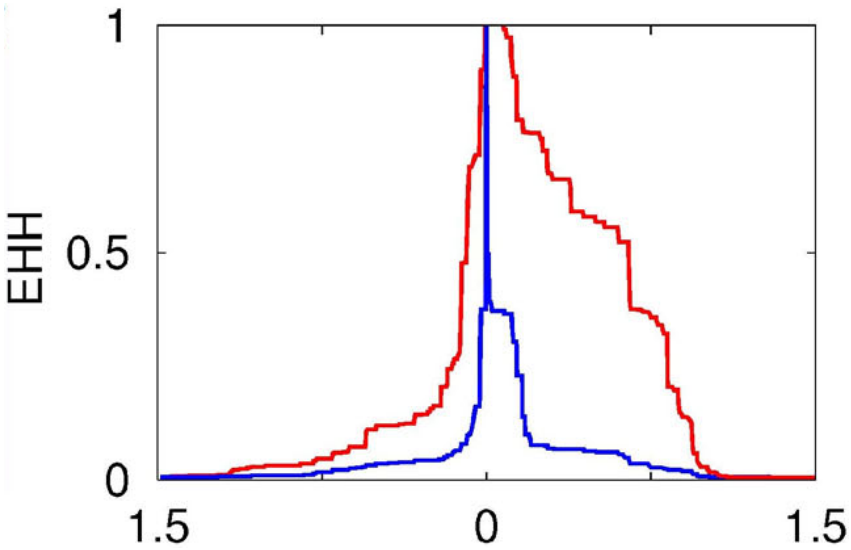
Decay of EHH with distance along chromosome



EHH at candidate locus: SPAG4

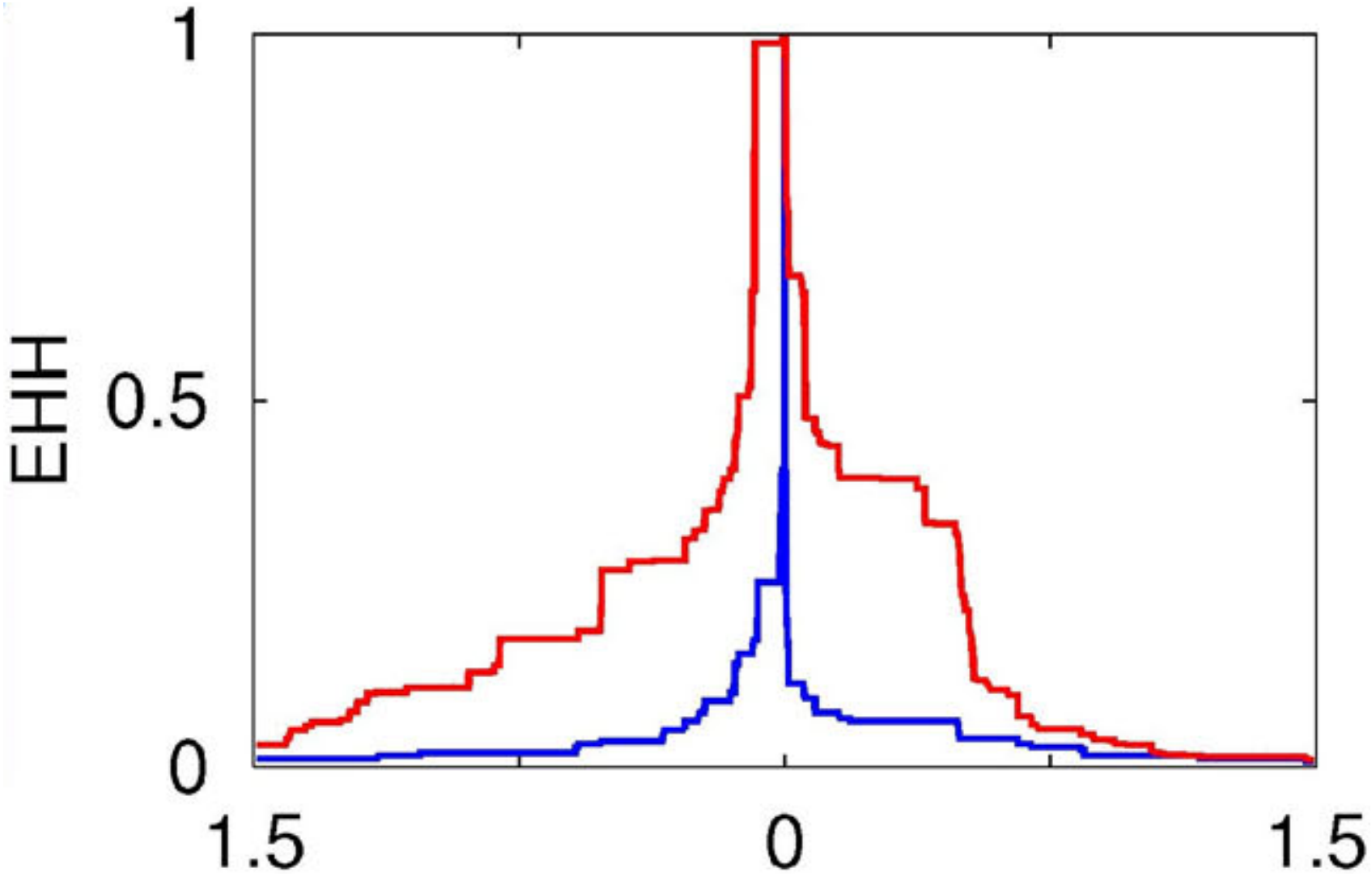


iHS, the integrated haplotype score

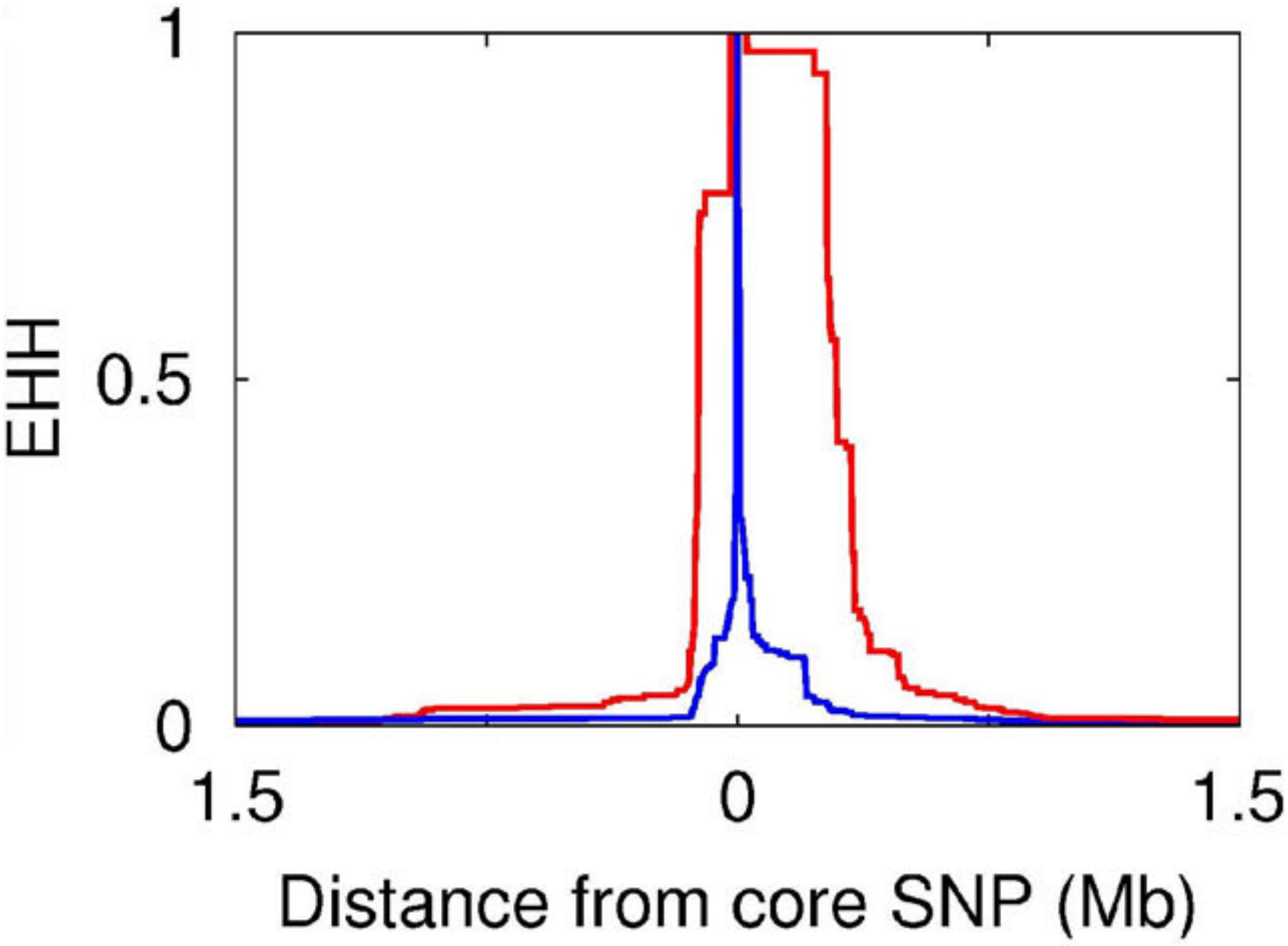


iHS is a normalized measure of the area between the two curves.

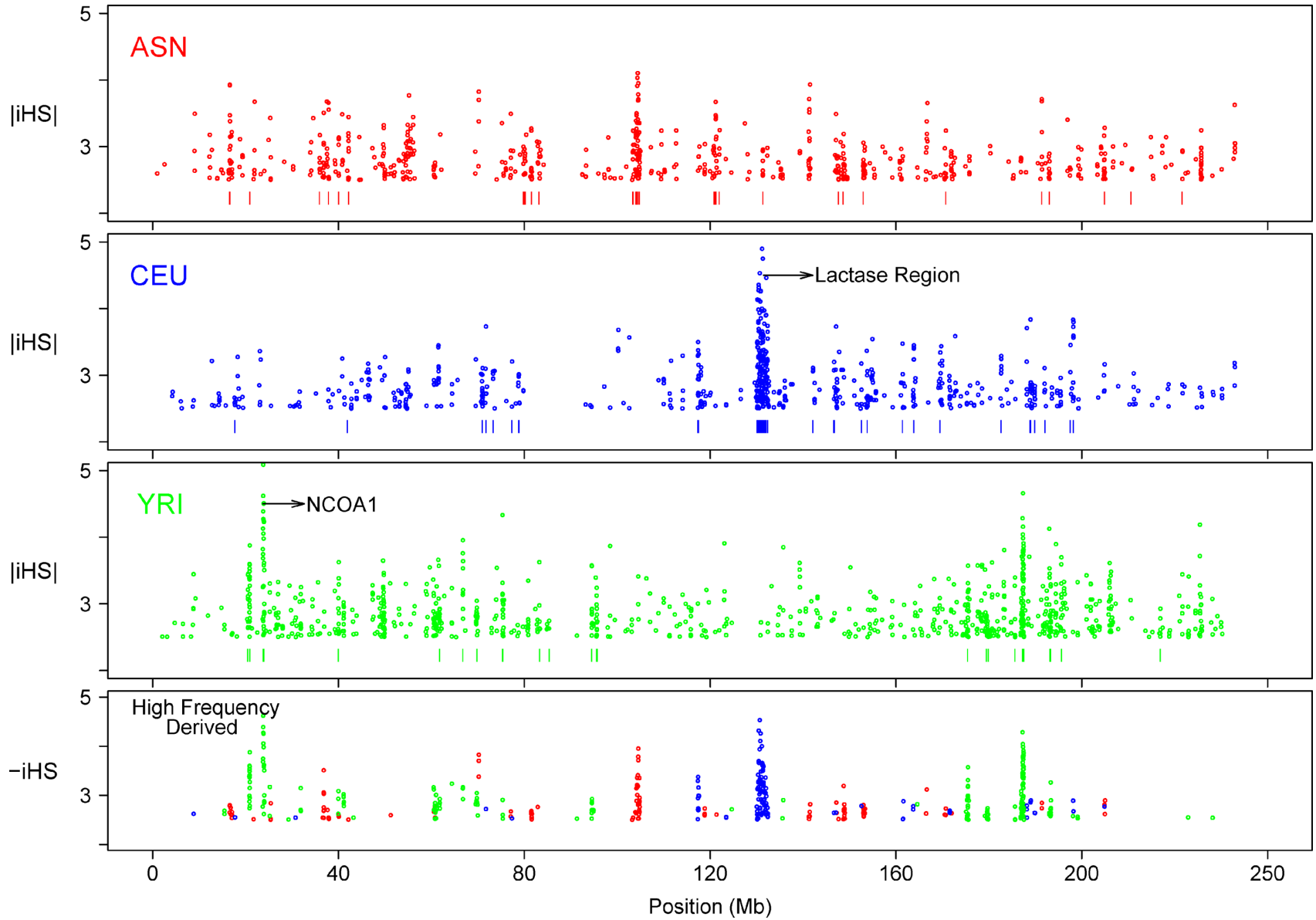
EHH at candidate locus: SNTG1



EHH at candidate locus: NCOA1

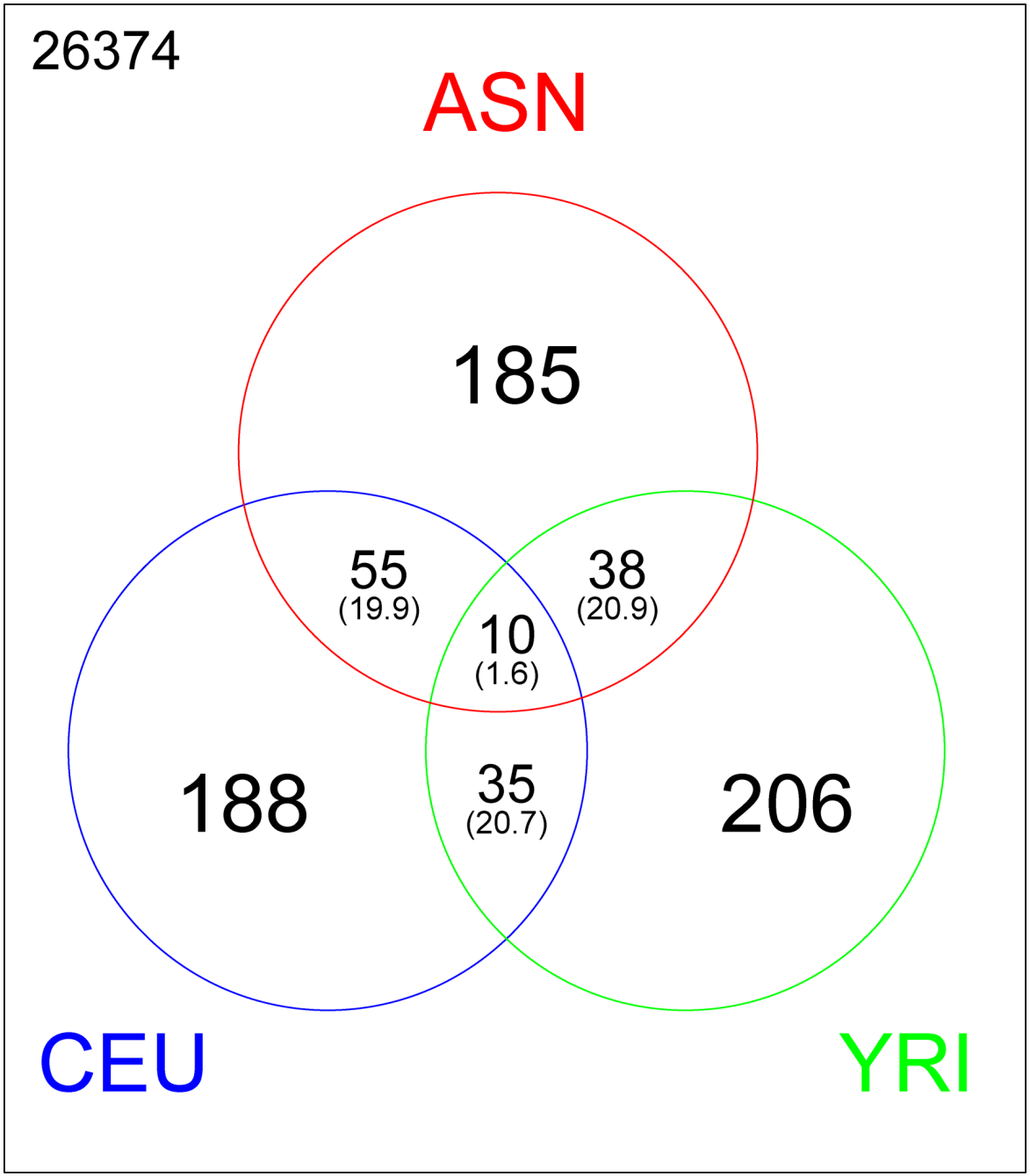


LD on human chromosome 2 (Voight et al 2006)



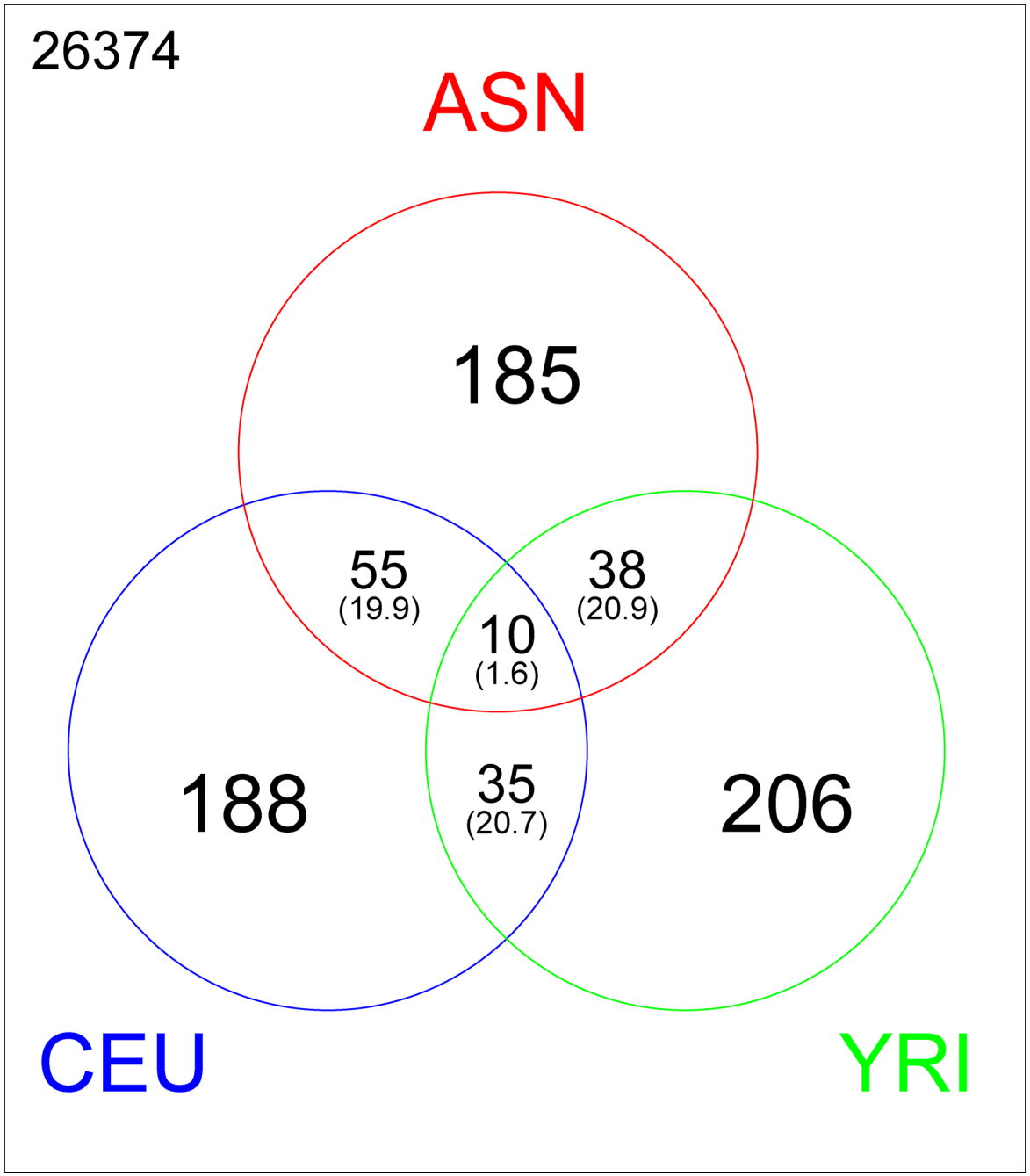
Study of Voight et al (2006)

- 800,000 SNPs in 309 people
- 431 sweeping loci
- Most sweeps started w/i past 10,000 years



Voight et al (2006):
431 sweeping loci.

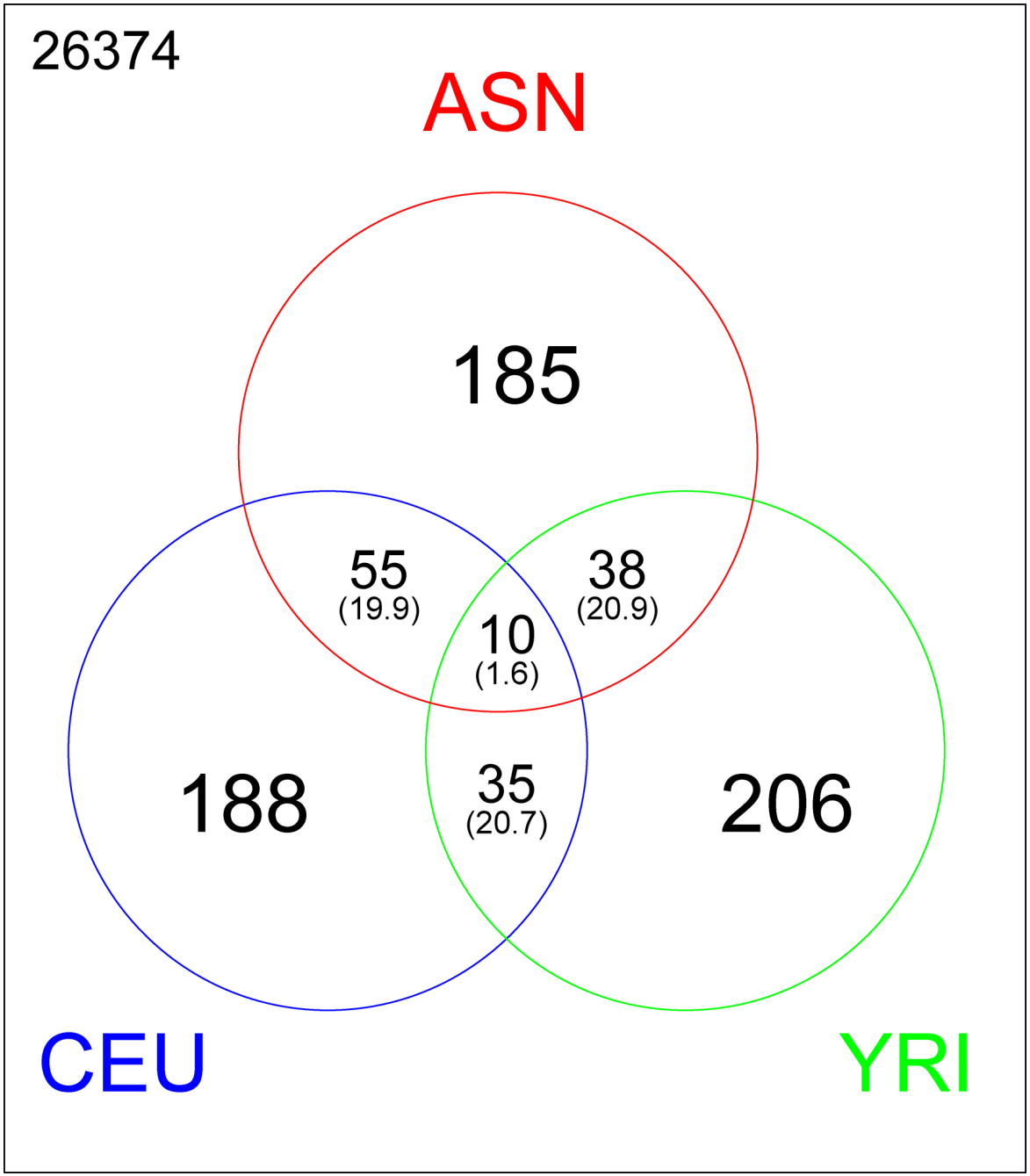
ASN: Asia
YRI: Africa
CEU: Europe.



Voight et al (2006): 431 sweeping loci.

ASN: Asia
YRI: Africa
CEU: Europe.

Most are sweeping w/i only one continent.

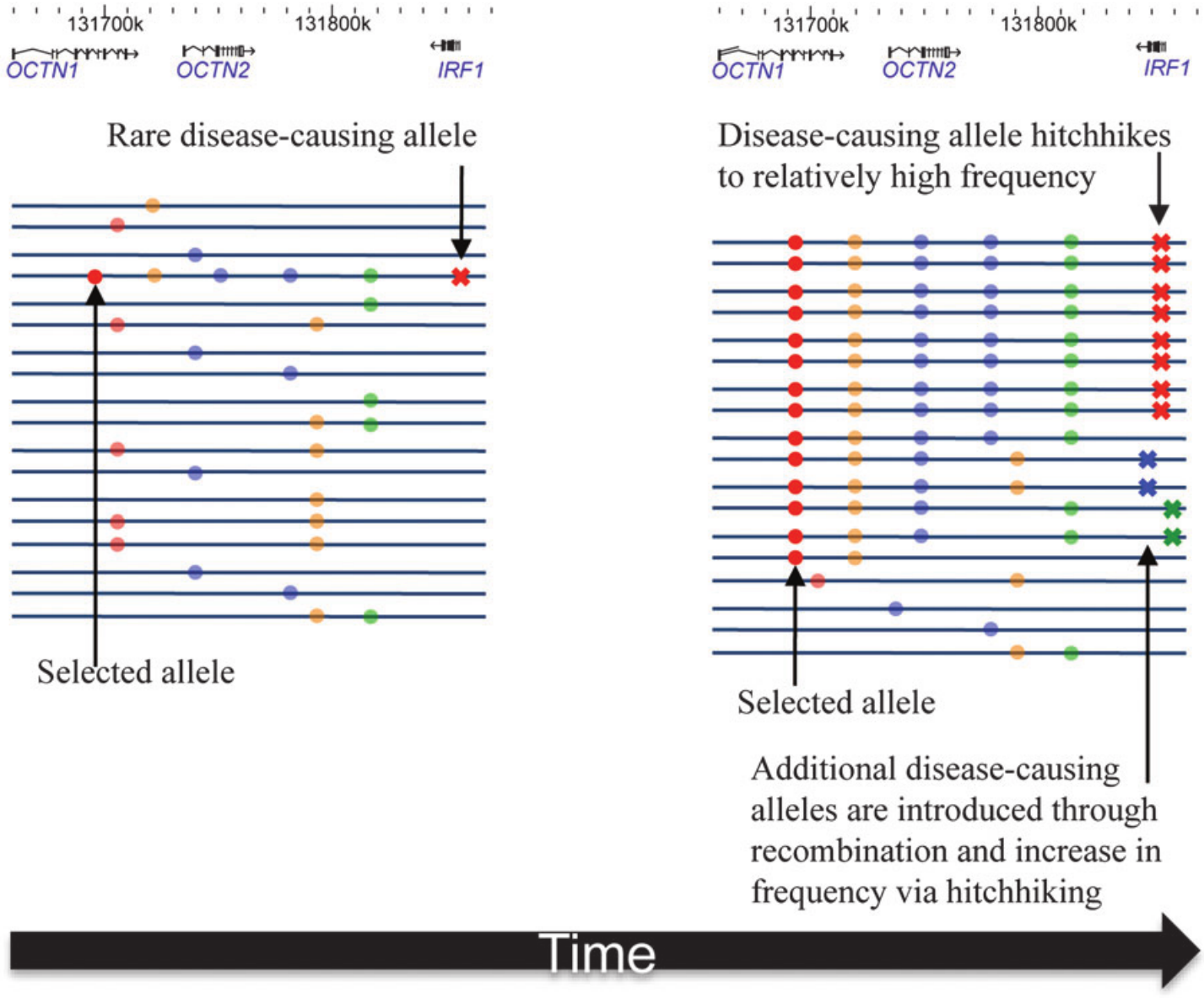


Voight et al (2006): 431 sweeping loci.

ASN: Asia
YRI: Africa
CEU: Europe.

Most are sweeping w/i only one continent.
Also true of Wang et al data.

IBD5



(Huff et al. 2012)

IBD5 and inflammatory bowel disease

- IBD5 is a 250 kb haplotype associated with Crohn's disease.
- Within this region, variant 503F of the OCTN1 gene covaries with Crohn's.
- OCTN1 transports the antioxidant ergothioneine.
- Why should such a gene cause Crohn's disease?

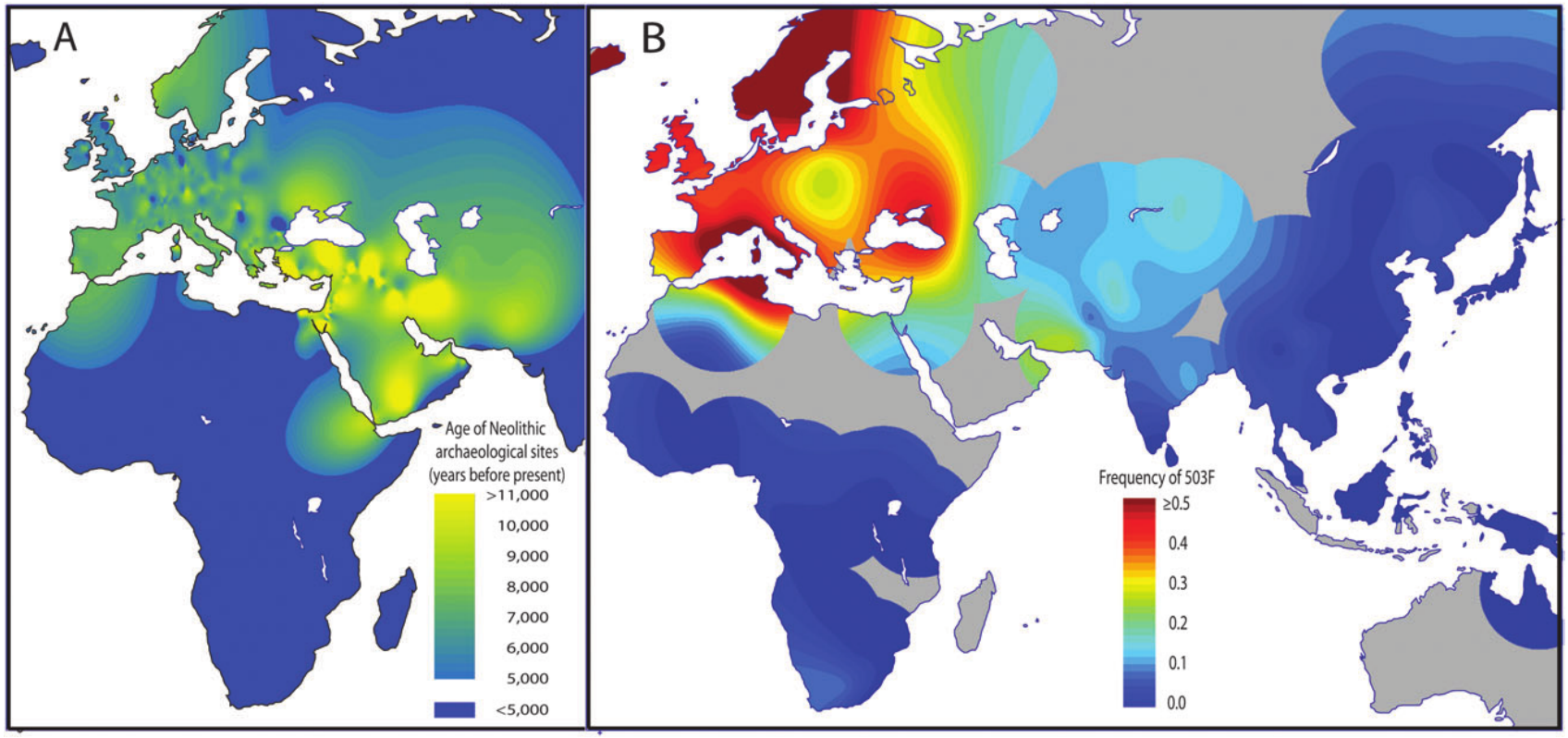
Ergothioneine (ET)

- An antioxidant, synthesized by fungi, and present in most plants and animals.
- Low in wheat, barley, lentils, and peas—foods domesticated early in Middle East.
- Early farmers would have lacked ET.

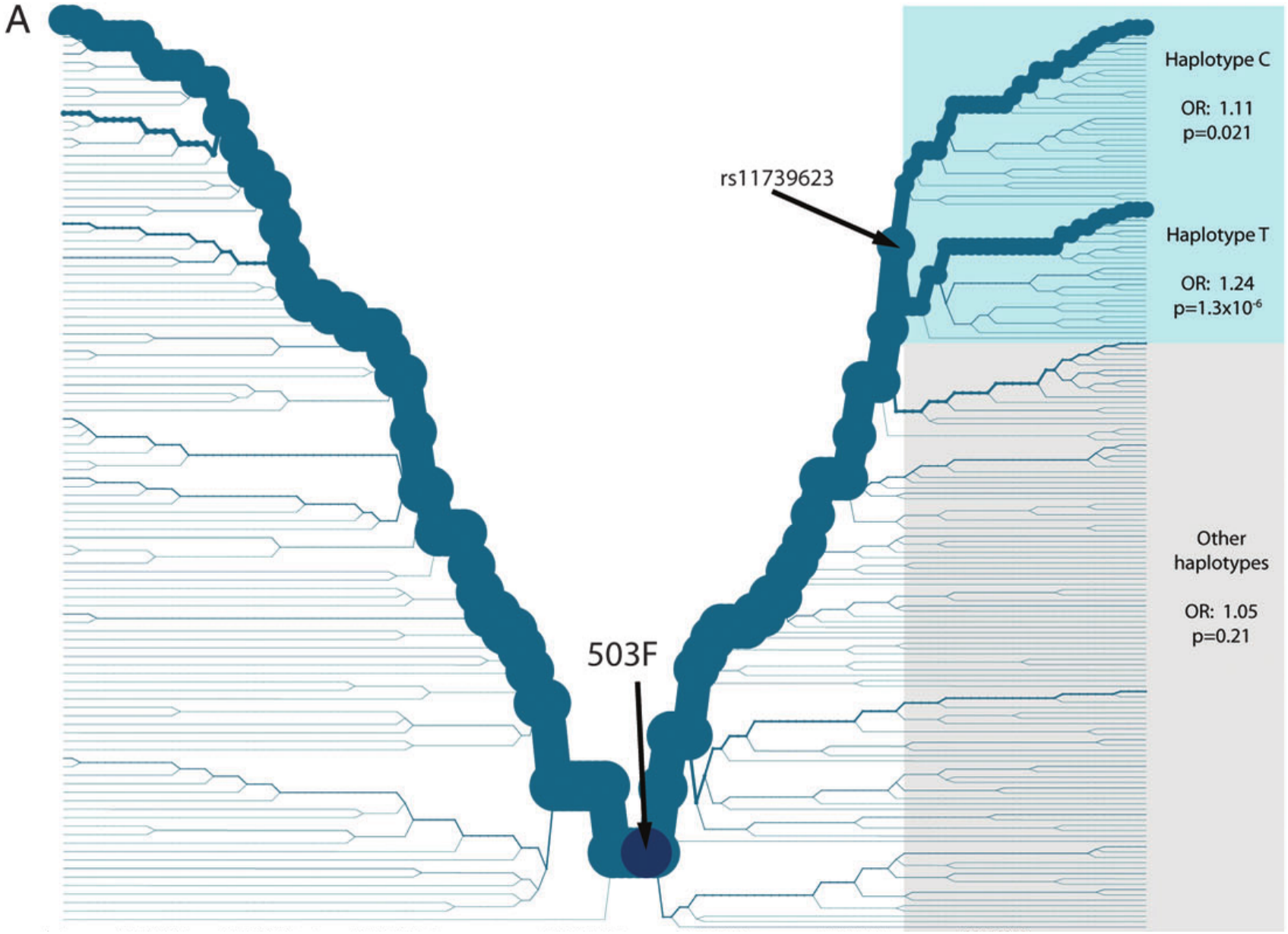
The OCTN1 protein

- Transports ergothioneine (ET).
- The 503F mutation increases rate of ET transport by 50%.
- Highly conserved in evolution, suggesting strong selection.
- Highly specific to ET.
- Common in Middle East and Europe; rare elsewhere.
- LD extent suggests that 503F mutated 7,750–19,025 y ago.

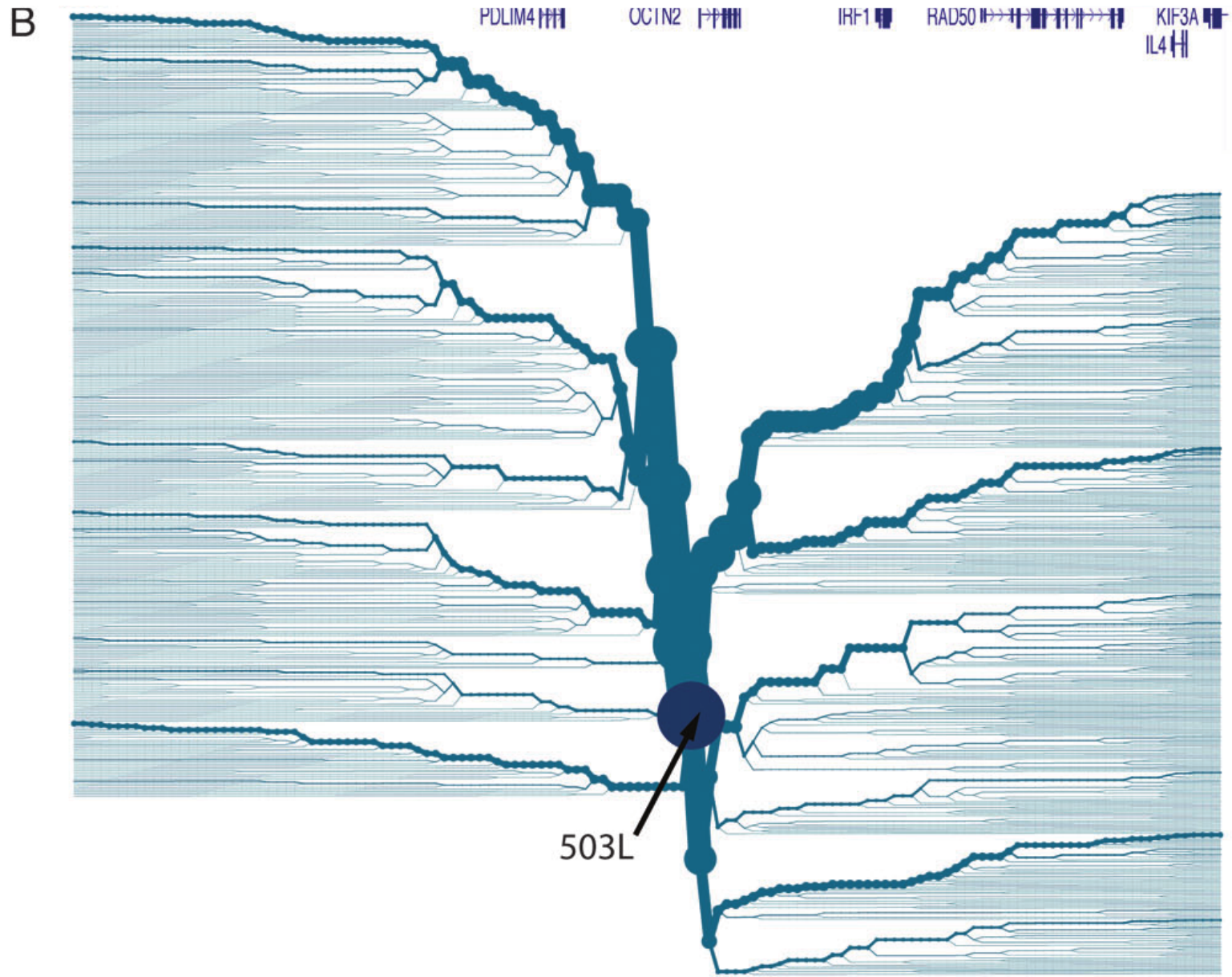
Distribution of early farming (A) and 503F (B)



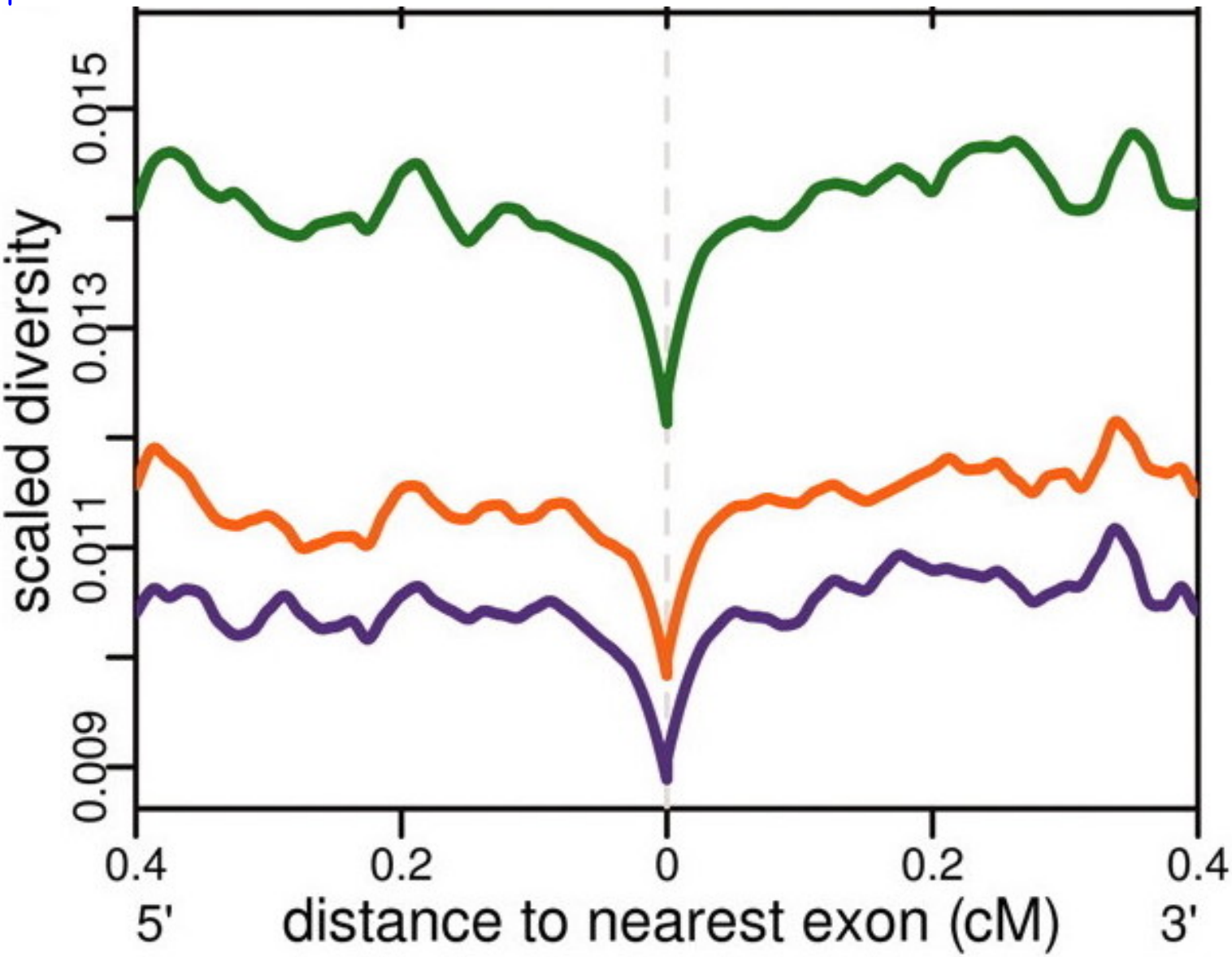
503F sits on a long LD block



503L sits on a short LD block



Diversity trough around exons in YRI, CEU, and CHB+JPT



Selective sweep predicts a trough in diversity

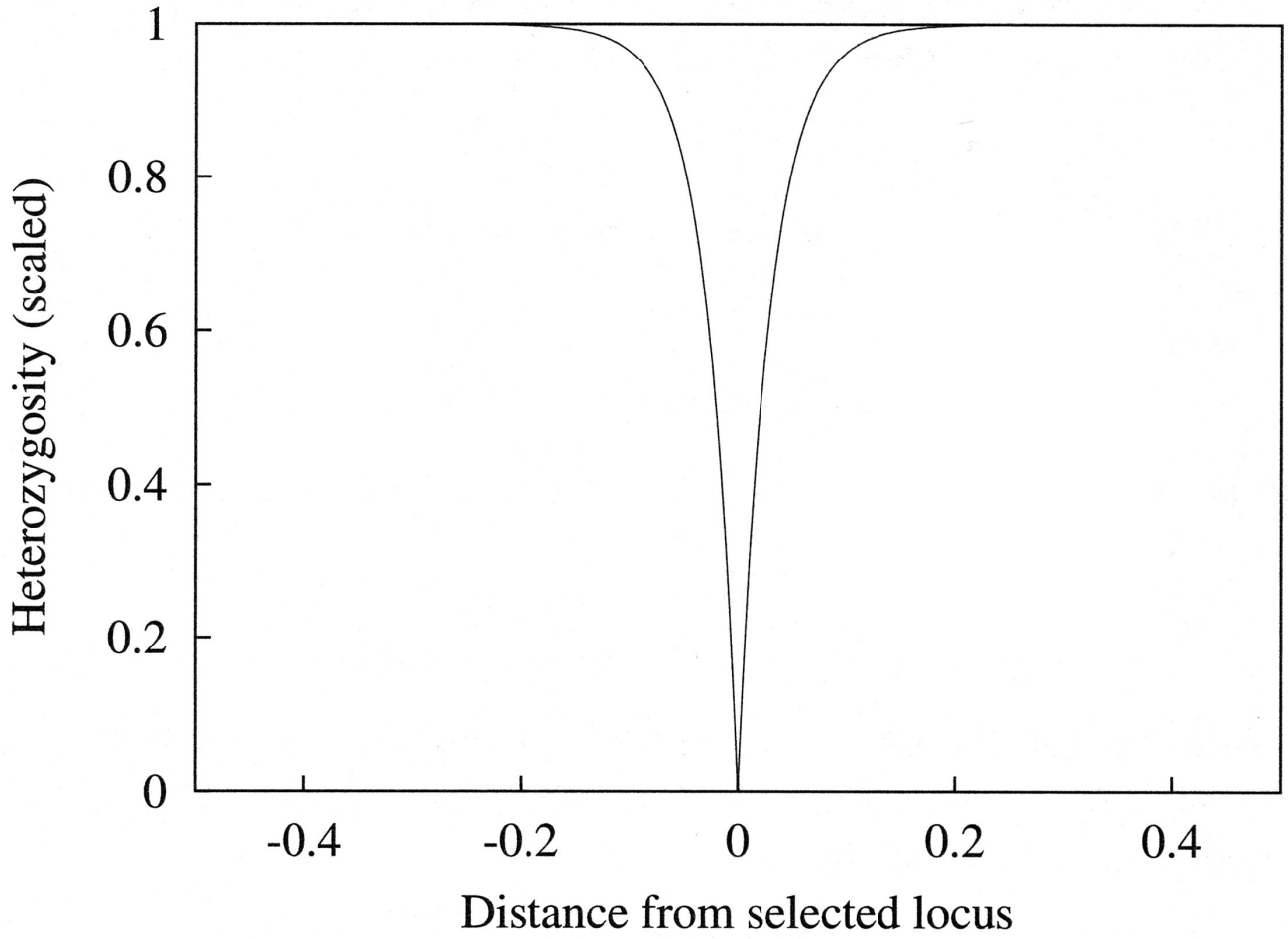


Figure 4.4: The ratio of the final to initial heterozygosity at a neutral locus as a function of the distance from the selected locus as measured by r/s. Negative values of r/s are left of the selected locus, positive values are to the right.

Why the trough?

As the favored allele sweeps to fixation, alleles at linked loci 'hitch-hike" to higher frequency.

Reduces diversity at linked loci.

Effect declines with distance from sweeping site.

Purifying selection also predicts a diversity trough

Mildly deleterious mutations may drift to moderate frequencies.

Are eventually removed by selection.

Any mutations at linked sites are also removed, lowering diversity at linked sites. This is called "background selection," as Jon will soon explain.

Hard sweeps versus soft sweeps

Hard sweep

- originates as new mutation on a single chromosome.
- mutant is in LD with linked sites
- selection reduces variation at linked sites

Soft sweep

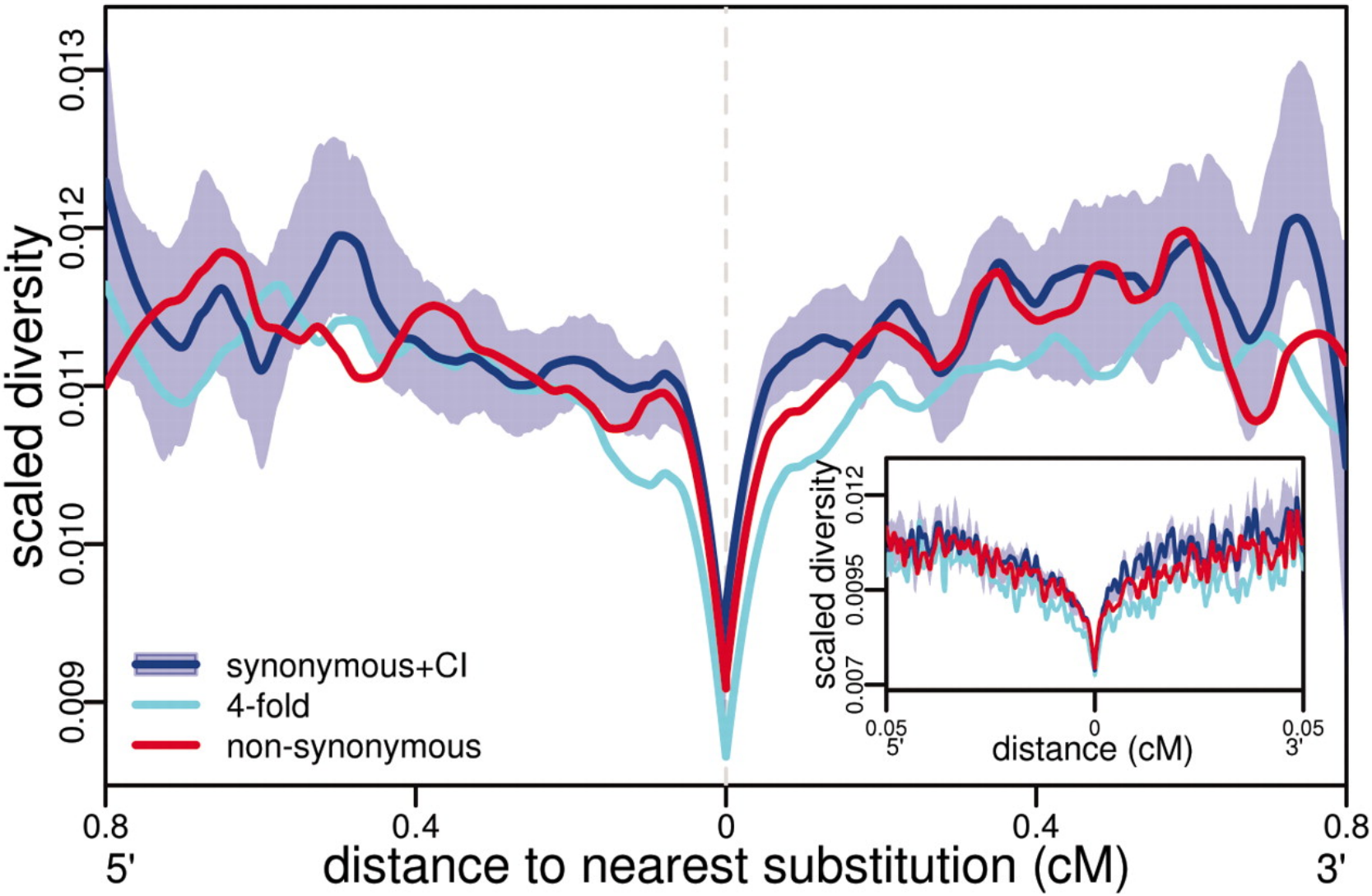
- an environmental change (or increase in population size) gives adaptive value to a variant that was previously neutral.
- there may be many initial copies of favored allele
- weak LD with linked sites
- selection has smaller effect on linked variation

Study of Hernandez et al (2011)

If the diversity trough results from classic sweeps, it should be deepest around functional changes. [Why?]

Hernandez et al compared two types of trough: (1) those around human-specific amino-acid fixations, and (2) those around synonymous substitutions.

They expect the troughs to be deepest around amino-acid fixations.


scaled diversity
0.009
0.010
0.011
0.012
0.013
synonymous+CI
4-fold
non-synonymous
0.8
5'
0.4
0
0.4
0.8
3'
distance to nearest substitution (cM)
scaled diversity
0.007
0.0095
0.012
0.05
5'
0
0.05
3'
distance (cM)

No difference in diversity troughs around synonymous and non-synonymous fixations

Hernandez et al (2011): Recent adaptive evolution in humans is mostly *not* the result of hard selective sweeps. This paper has been viewed as support for the importance of soft sweeps.

I don't buy it. Soft sweeps should not generate pronounced diversity troughs.

I think the paper tells us that diversity troughs are mainly the result of background selection—not adaptive evolution