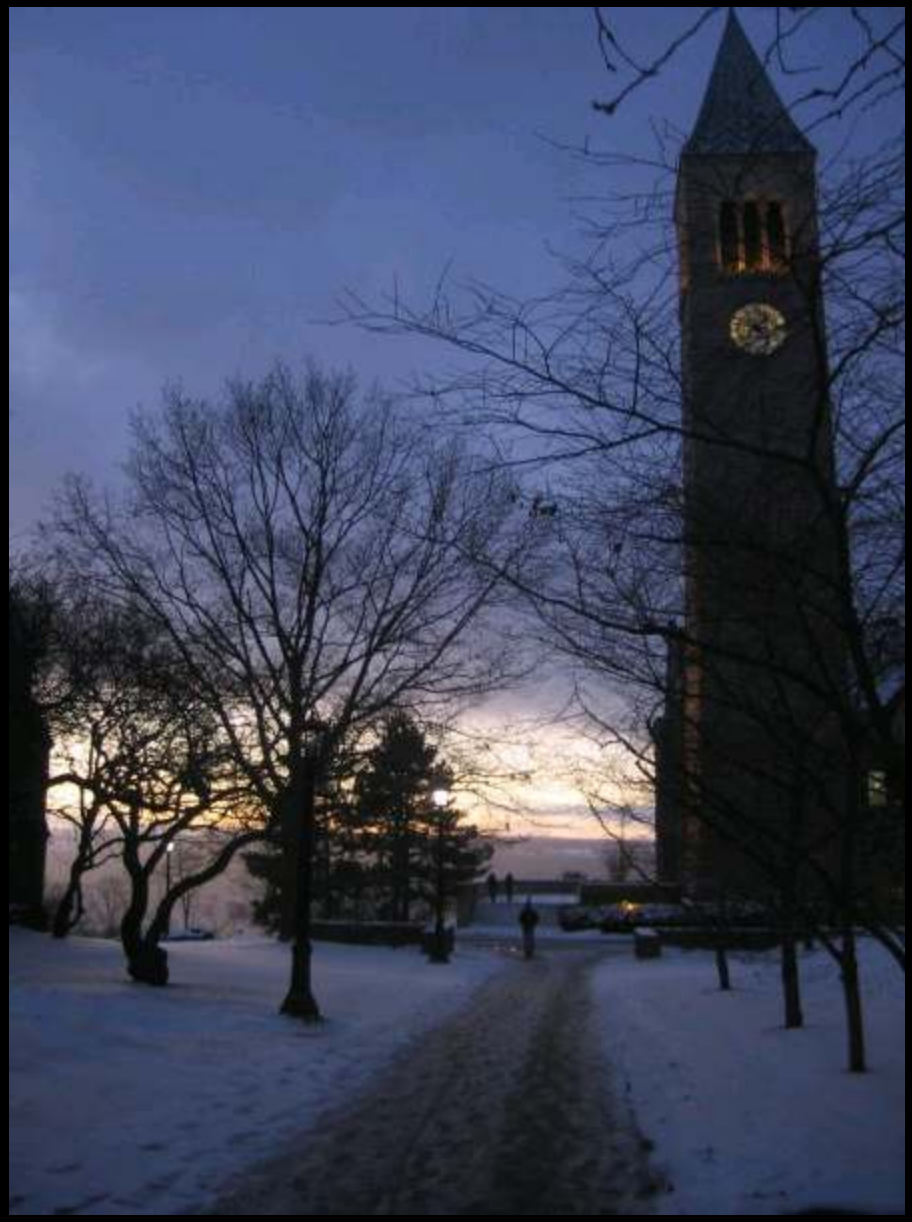
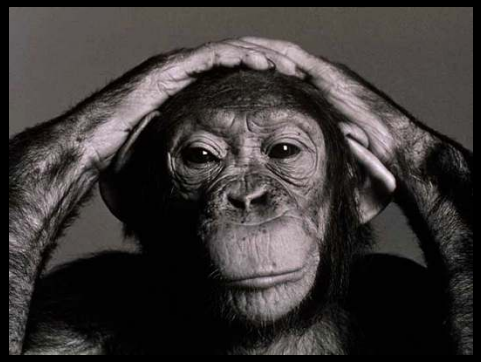


Population Genomics



Andrew Clark
Cornell University

Primary questions of population genomic data

- How much genetic variation is there within species?
- What evolutionary forces have exerted change to the variation?
- What can be inferred about the population's past history from genetic variation?
- What are the limits to the resolution of these methods?

Primary questions of population genomic data

- How much genetic variation is there within a species?
  - How is it quantified?
  - How do we deal with sequencing error?
  - How is the variation structured?

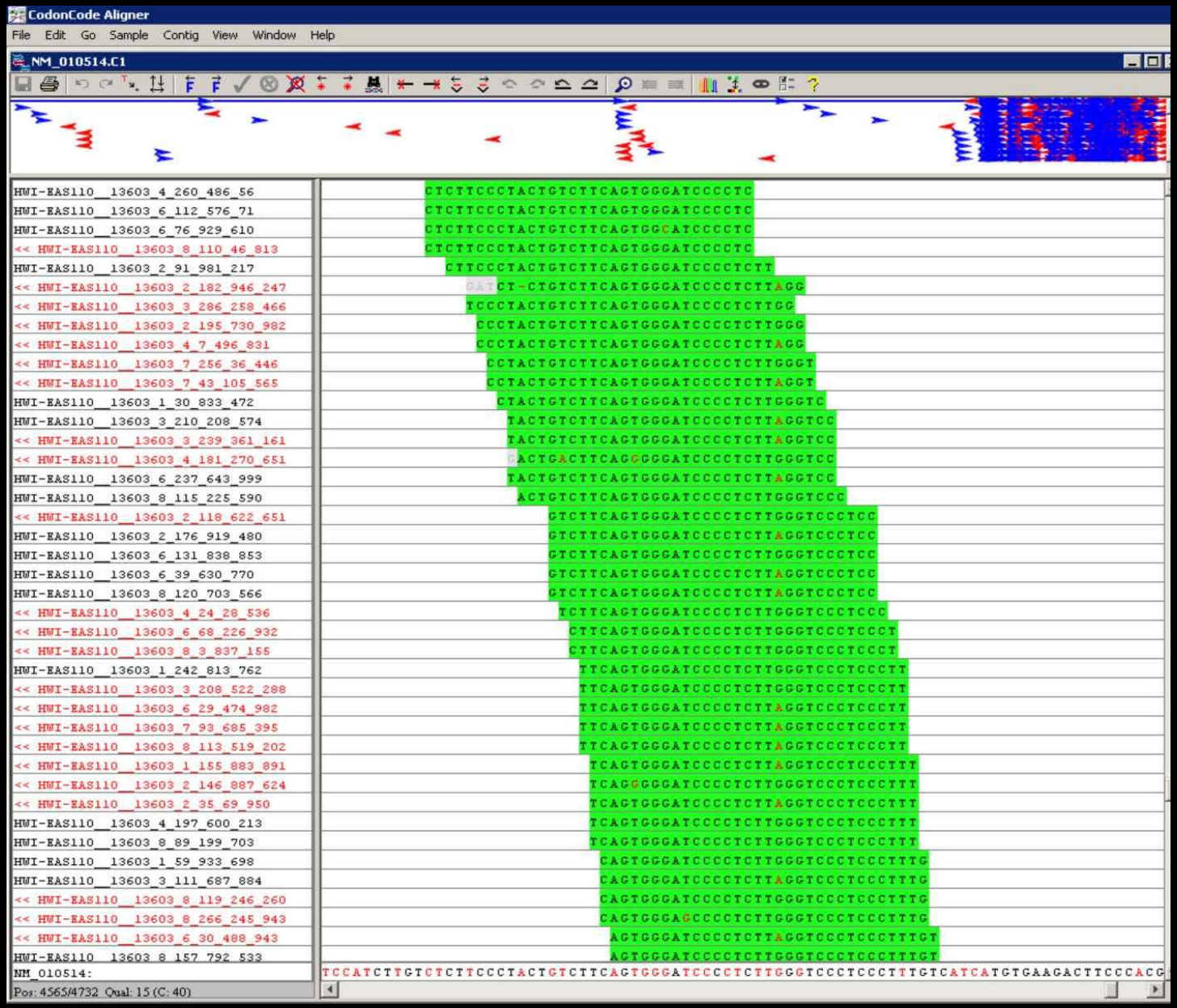
Identifying genetic variation from next gen data

The data consist of many many reads from several individuals.

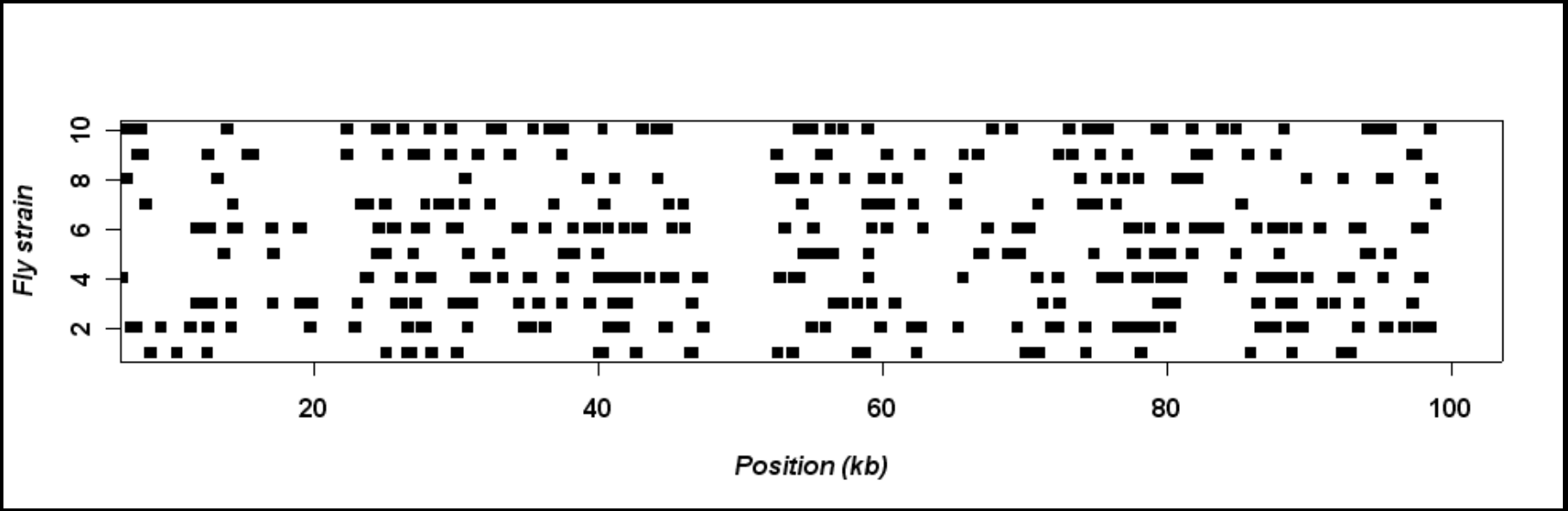
A major decision:

Do we work with SNP calls for each individual (with estimates of confidence) or instead do we make population genetic inferences directly from the alignment of sequence reads?

Multialignments of next-gen (Illumina) reads



Often it is useful to make inference from shallow-coverage data



Sackton et al. 2009

A serious issue of bias

- Insisting on calling each individual's SNP genotype, and setting a high threshold of confidence in SNP calls, means that we undercall rare variants, and many parameters will be mis-estimated.
- Meanwhile, nearly every error will be a false positive SNP call
- For data with thin coverage per individual, you are better off NOT calling individual SNP genotypes.

Johnson and Slatkin 2008 25(1):199-206

DOI: 10.1534/genetics.109.100479

Estimation of Allele Frequencies From High-Coverage Genome-Sequencing Projects

Michael Lynch[1]

Department of Biology, Indiana University, Bloomington, Indiana 47405

Manuscript received January 6, 2009
Accepted for publication March 6, 2009

ABSTRACT

A new generation of high-throughput sequencing strategies will soon lead to the acquisition of high-coverage genomic profiles of hundreds to thousands of individuals within species, generating unprecedented levels of information on the frequencies of nucleotides segregating at individual sites. However, because these new technologies are error prone and yield uneven coverage of alleles in diploid individuals, they also introduce the need for novel methods for analyzing the raw read data. A maximum-likelihood method for the estimation of allele frequencies is developed, eliminating both the need to arbitrarily discard individuals with low coverage and the requirement for an extrinsic measure of the sequence error rate. The resultant estimates are nearly unbiased with asymptotically minimal sampling variance, thereby defining the limits to our ability to estimate

Lynch 2009 Genetics 182(1):295-301.

Lynch (2009) – assuming HW and an error model, estimate allele frequencies from alignments

$$L(p) = K[p^2]^{N_{11}}[2p(1-p)]^{N_{12}}[(1-p)^2]^{N_{22}},$$

$$\begin{aligned} &P(n_1,\, n_2,\, n_3 \,|\, n,\, p,\, \varepsilon) \\ &\quad = p^2\Phi_e\left(n_2,\, n_3;\, n,\, \frac{\epsilon}{3},\, \frac{2\epsilon}{3}\right) + (1-p)^2\Phi_e\left(n_1,\, n_3;\, n,\, \frac{\epsilon}{3},\, \frac{2\epsilon}{3}\right) \\ &\qquad + 2p(1-p)\Phi_e(n_3;\, n,\, \frac{2\epsilon}{3})p(n_1;\, n_1+n_2,\, 0.5). \end{aligned}$$

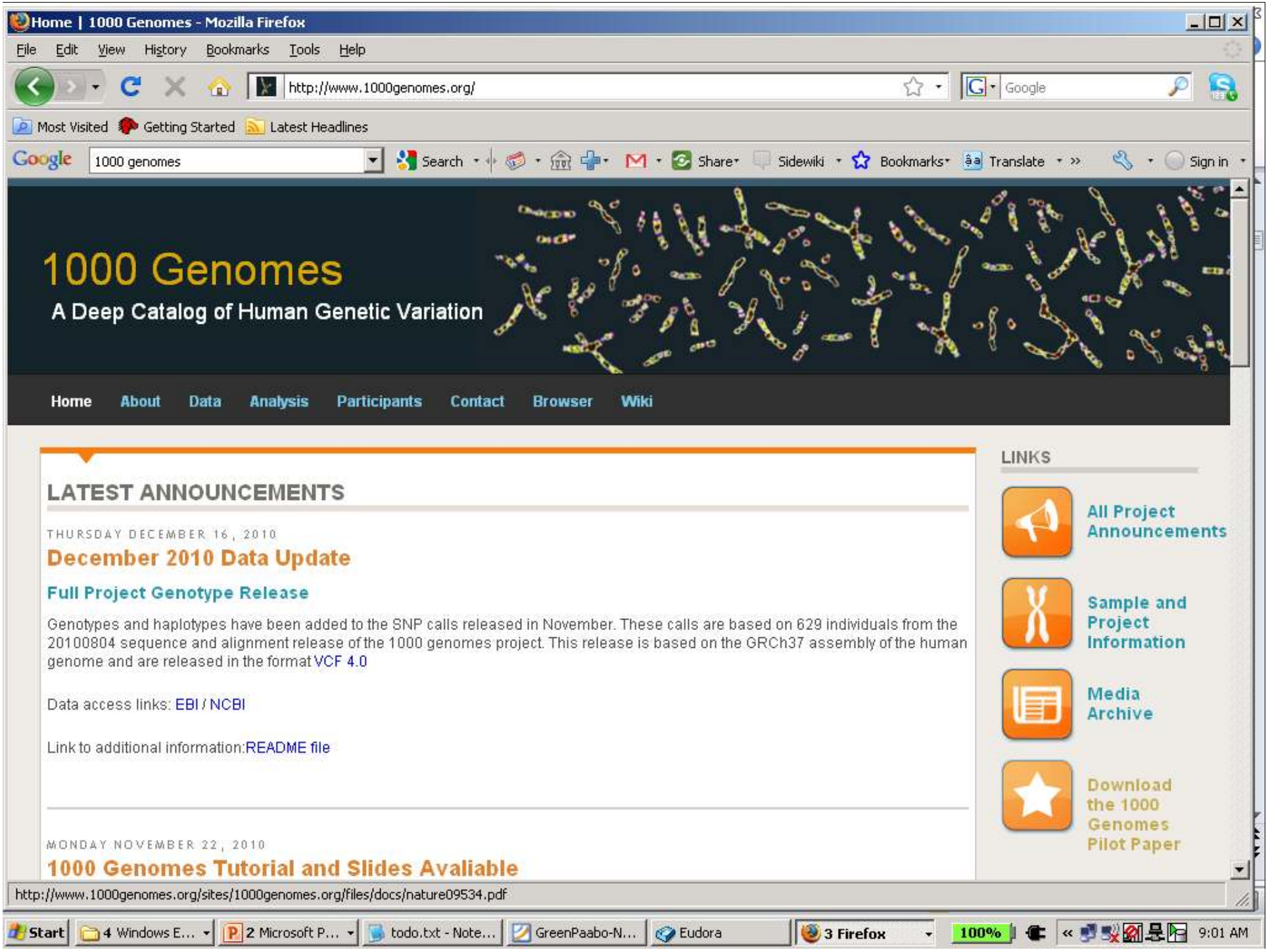
$$L = \sum N(n_1,\, n_2,\, n_3) \cdot \ln[P(n_1,\, n_2,\, n_3 \,|\, n,\, p, \epsilon)], \quad (3)$$

θ from low-coverage, short read data

- Shallow coverage (many gaps)
- High error rate; site-specific error
- Sampling one or two alleles per individual

$$\hat{\theta} = \frac{S_T - \lambda \sum_{r=1}^{v} L_r m_r I(m_r > 1)}{\sum_{r=1}^{v} L_r \left(\sum_{j=k_{\min},r}^{k_{\max},r} p(k_r = j) \sum_{i=1}^{j-1} \frac{1}{i} \right)}$$

Hellmann et al., 2008 Genome Research 18:1020-1029
Pool et al. 2010 Genome Research 20: 291-300.


Home | 1000 Genomes - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.1000genomes.org/
Most Visited Getting Started Latest Headlines
1000 genomes
Search Share Sidewiki Bookmarks Translate Sign in
1000 Genomes
A Deep Catalog of Human Genetic Variation
Home About Data Analysis Participants Contact Browser Wiki
LATEST ANNOUNCEMENTS
THURSDAY DECEMBER 16, 2010
December 2010 Data Update
Full Project Genotype Release
Genotypes and haplotypes have been added to the SNP calls released in November. These calls are based on 629 individuals from the 20100804 sequence and alignment release of the 1000 genomes project. This release is based on the GRCh37 assembly of the human genome and are released in the format VCF 4.0
Data access links: EBI / NCBI
Link to additional information:README file
MONDAY NOVEMBER 22, 2010
1000 Genomes Tutorial and Slides Avaliable
LINKS
All Project Announcements
Sample and Project Information
Media Archive
Download the 1000 Genomes Pilot Paper
http://www.1000genomes.org/sites/1000genomes.org/files/docs/nature09534.pdf
Start 4 Windows E... 2 Microsoft P... todo.txt - Note... GreenPaabo-N... Eudora 3 Firefox 100% 9:01 AM

What evolutionary forces have exerted change to the genetic variation?

- Mutation
- Random genetic drift
- Migration
- Natural selection

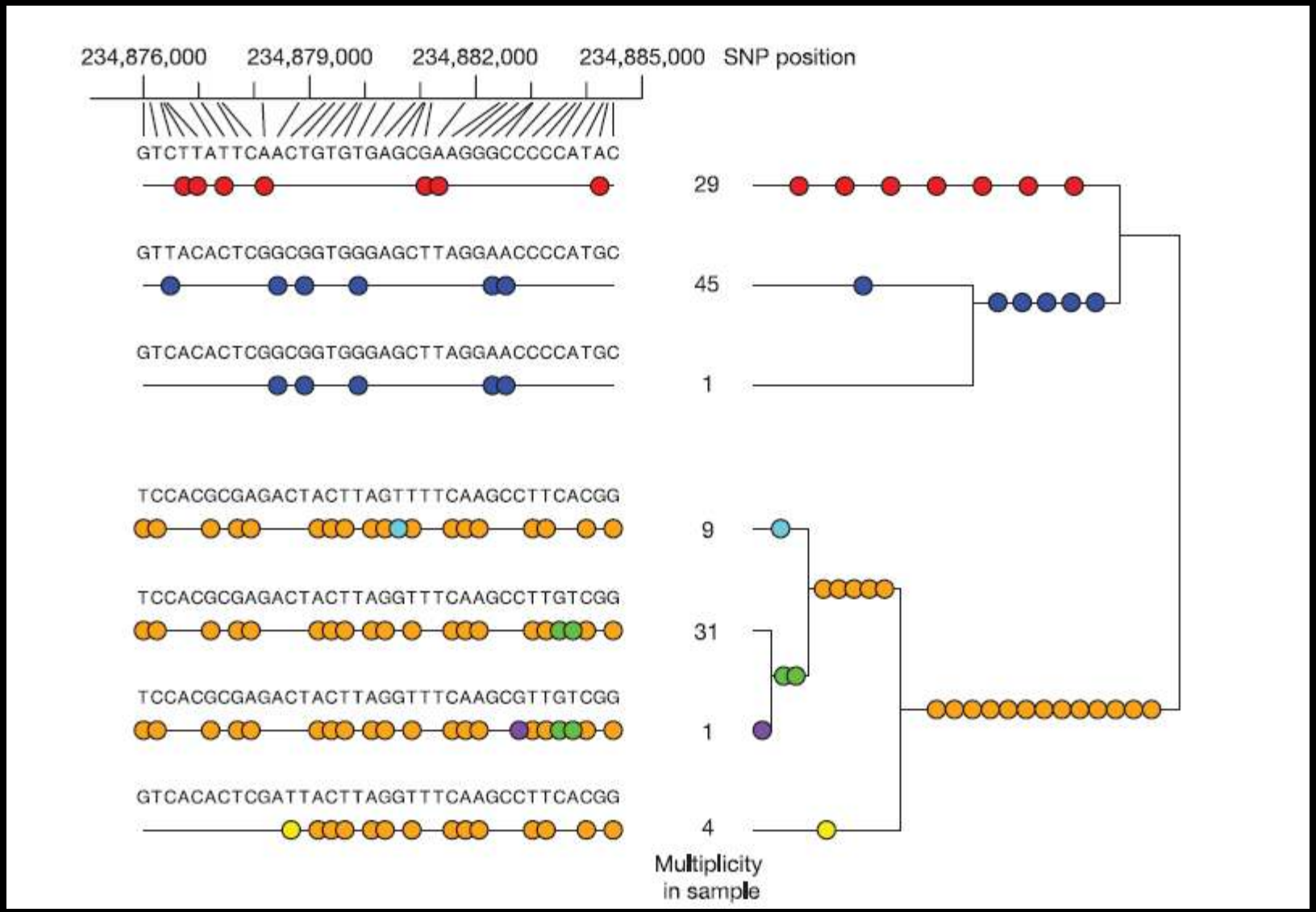
How does whole genome sequence help with inferences about mutation?

Generally the data come from resequencing of mutation –accumulation lines

- Rates of all types of events
- Fits to specific mutation rate models
- Context dependence
  - Neighboring bases
  - Local recombination rate
  - TE proximity
- Separate estimation of N and μ if sample size is large enough.

Mutation, drift and linkage disequilibrium

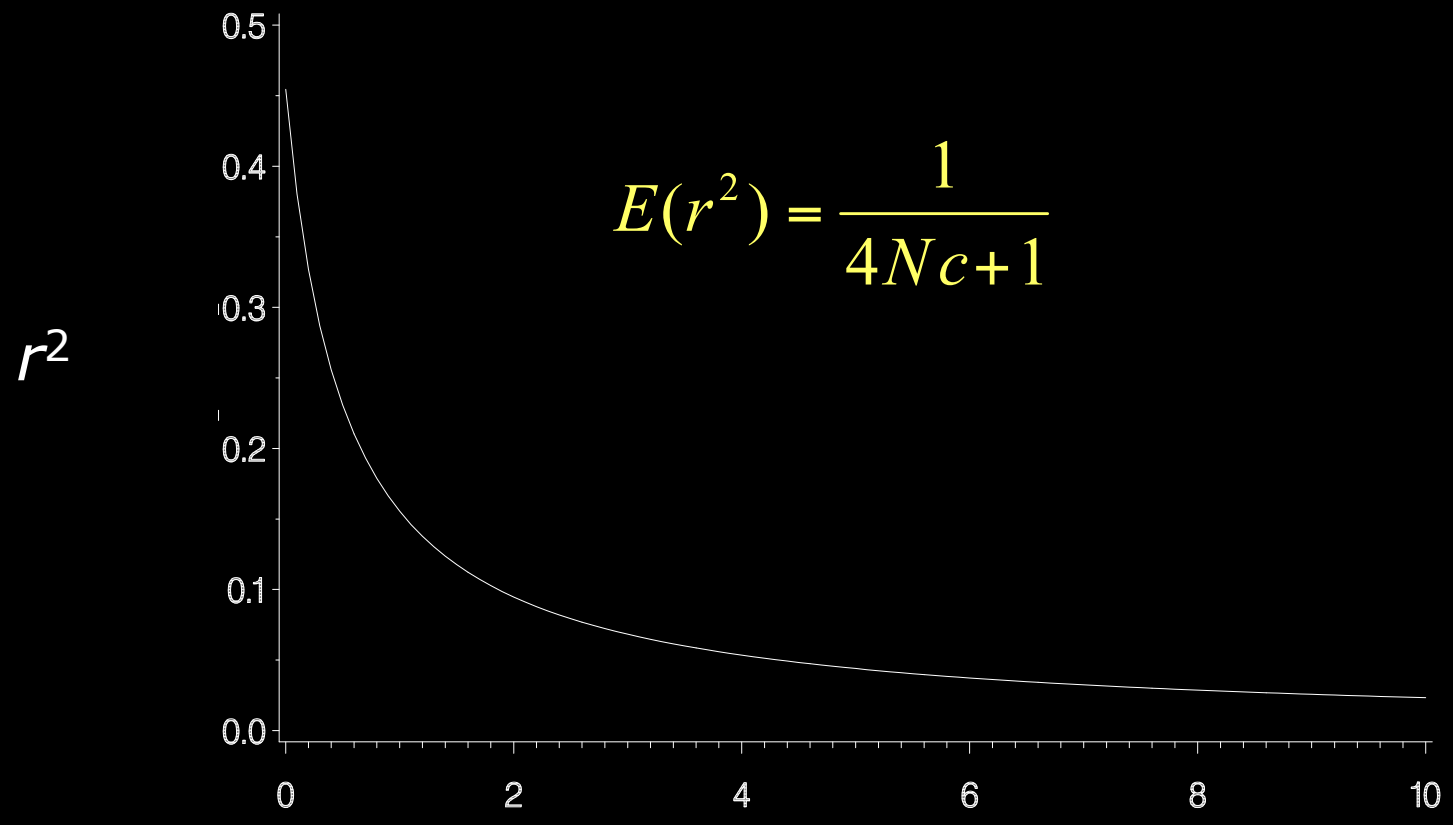
How LD arises as a process of ancestral history of mutations



Chromosome 2 ENCODE region

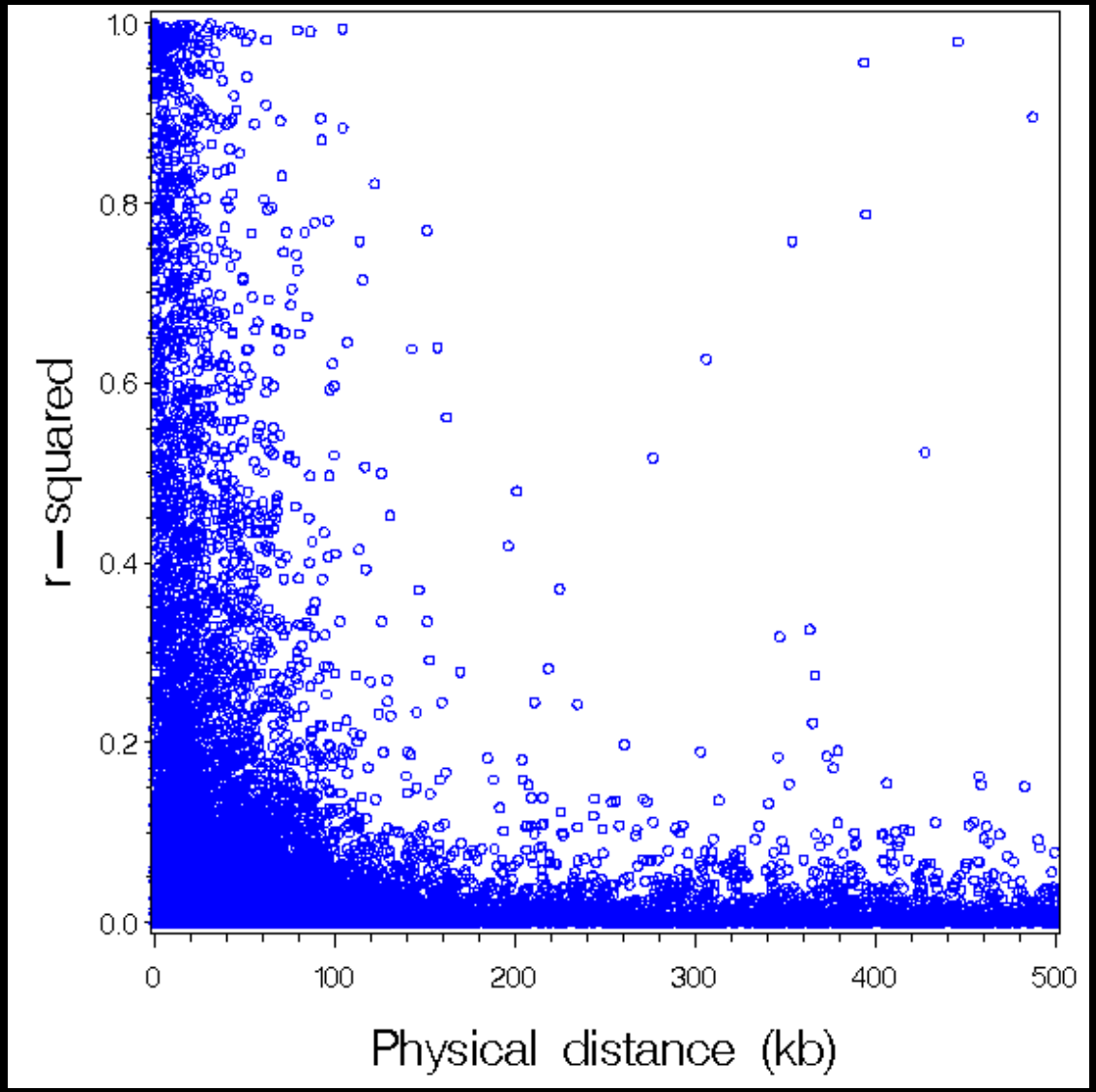
r^2 should be high only for SNPs near disease mutations

$$E(r^2) = \frac{1}{4Nc+1}$$

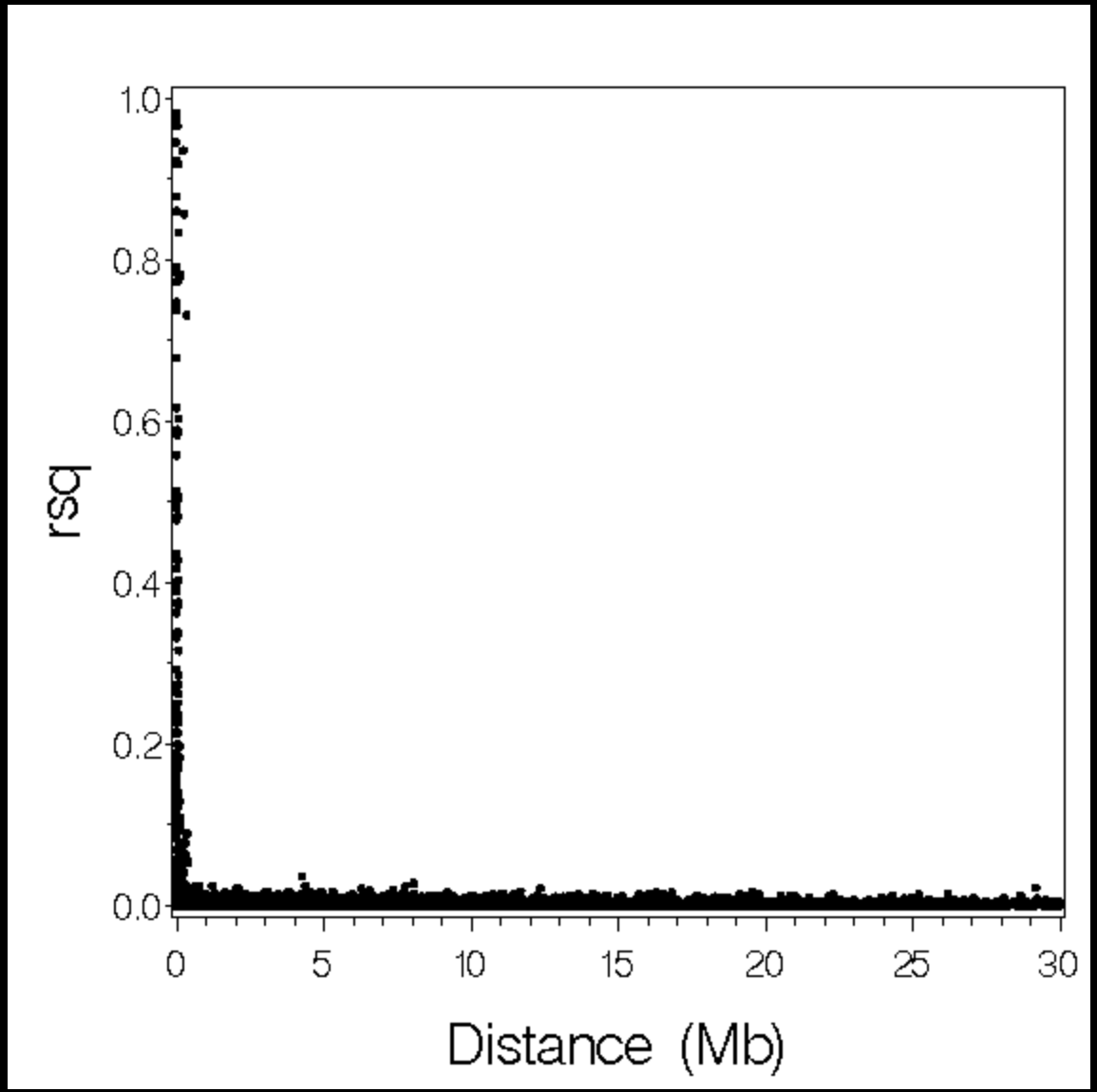


Sved 1971; Kimura and Ohta 1971

r^2 measures Linkage Disequilibrium (LD): LD decays with distance



…so observing high r^2 means the SNP is near the disease gene.



How has next gen sequencing modernized our views about random genetic drift?

- Massive sampling all over the genome.
- Wide variation in time to common ancestry (coalescence) across the genome.
- Wide variation in effective population size across the genome.
- Departures of site frequency spectrum from neutrality lead to inference of past demography.
- X vs autosome differences in effective size – differences in sex-specific migration rates, sex differences in variance in family size, etc..

Inferring natural selection from DNA sequences

- Substitution rates at syn and nonsyn sites
  - Codon substitution models (PAML)
- Polymorphism vs. divergence (site counts)
  - McDonald-Kreitman test; HKA test
- Local depression in diversity
  - Kim and Stephan; Tajima's D
- High frequency derived alleles
  - Fay and Wu's H
- Exceptionally long haplotypes
  - Extended haplotype homozygosity; iHS
- Excess divergence between populations
  - F_{ST}

Substitution at synonymous and nonsynonymous sites

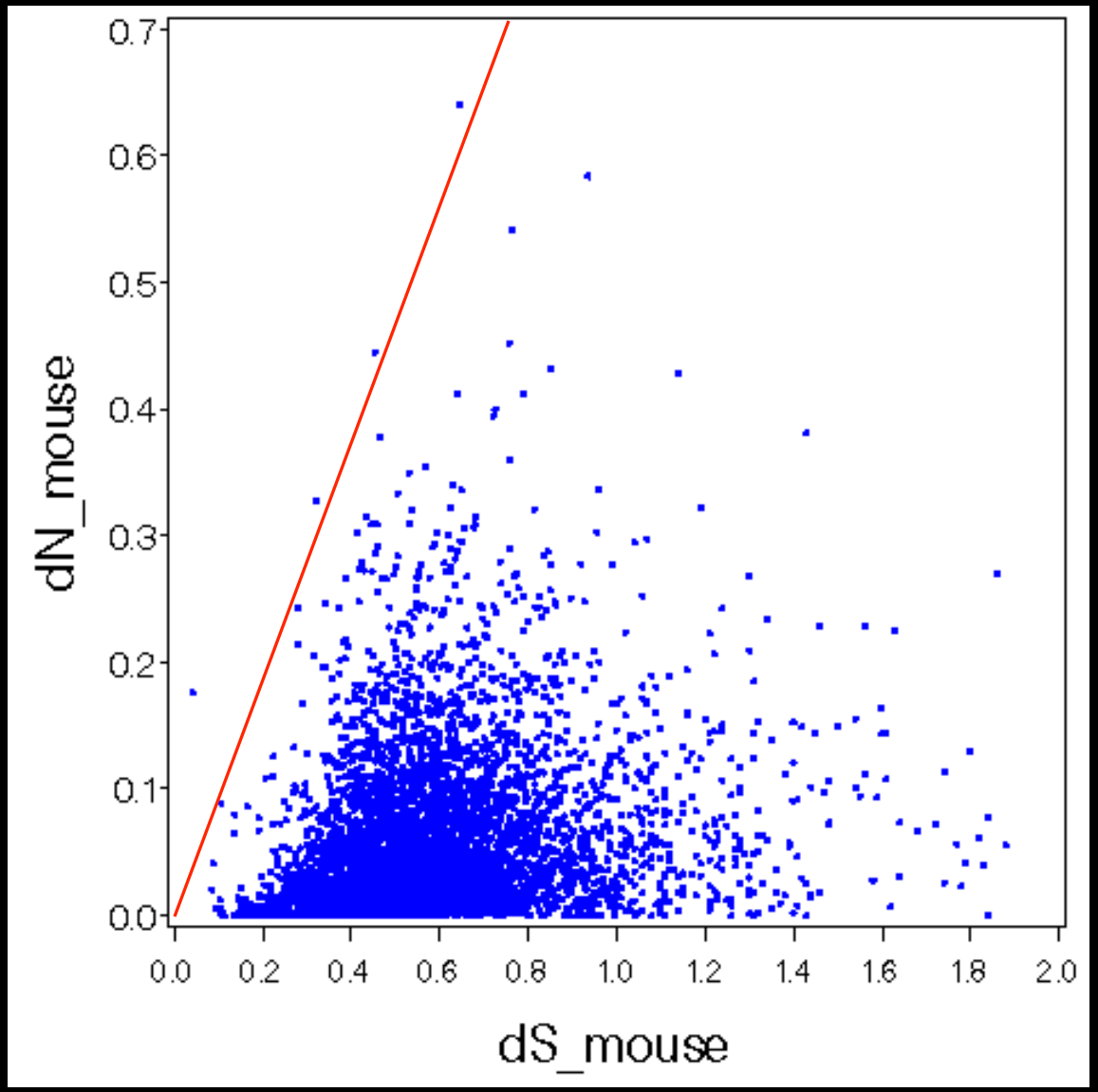
d_N = rate of nonsynonymous substitutions per nonsynonymous site (those that change encoded amino acid)

d_S = rate of synonymous substitution per synonymous site ("silent" sites)

$\omega = d_N/d_S$

Yang and Nielsen 2001

Most genes have $d_N << d_S$ (human-mouse comparison)

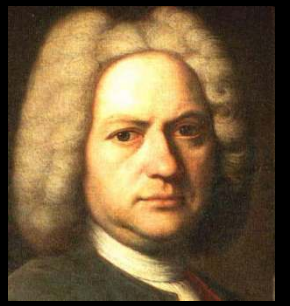


Testing human-specific accelerated divergence by maximum-likelihood.

mouse

d_{N2}

human



d_{N1}

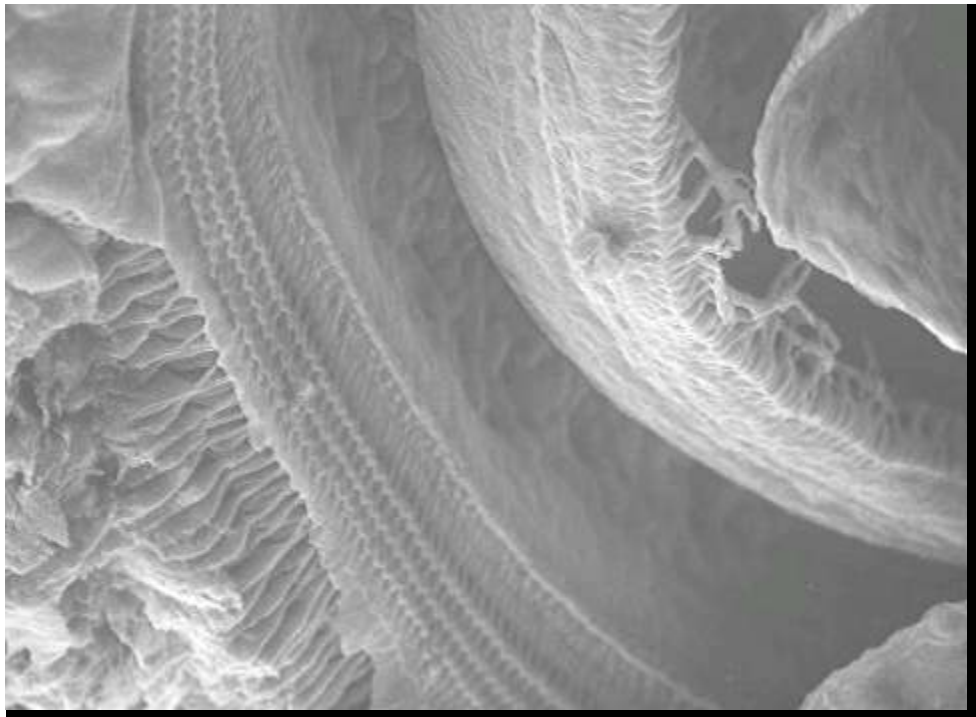
chimp

Mouse-human split: 80-110 MYA
Human-chimp split: 4.6 – 6 MYA

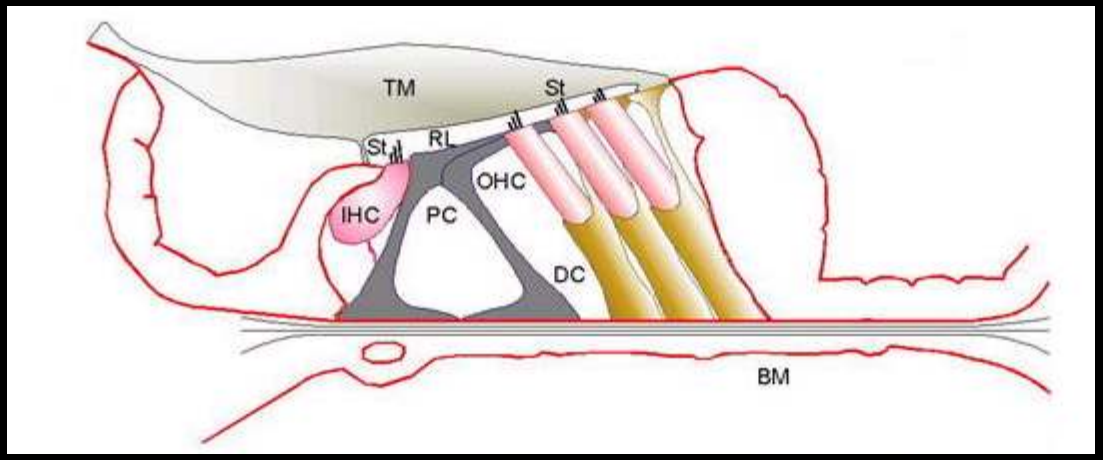
Clark et al. (2003) Science 302:1960-1963

List of most human-accelerated genes

hCG	gene name	Chr	M2_P-value
hCG16247 ✦	tectorin alpha, mucin related	11	0.000016
hCG28335	Olfactory Receptor	11	0.000040
hCG1818624	CUB and Sushi multiple domains 1	8	0.000069
hCG32289	FERM, RhoGEF and pleckstrin domain protein 2	2	0.000084
hCG1730362	olfactory receptor, family 5, subfamily I, member 1	11	0.000151
hCG37074	mannosidase, alpha, class 1A, member 2	1	0.000201
hCG19640	transient receptor potential cation channel, subfamily V-6	7	0.000242
hCG39130	cell adhesion molecule-related	11	0.000293
hCG23342	solute carrier family 6 (neurotransmitter transporter, glycine)	11	0.000337
hCG1788087	Olfactory receptor	11	0.000373
hCG30396 ✦	doublesex and mab-3 related transcription factor 3	9	0.000386
hCG37558	nucleoporin 155kDa	5	0.000444
hCG31797 ✦	winged-helix nude transcription factor	17	0.000557
hCG20073	upstream binding protein 1 (LBP-1a)	3	0.000565
hCG41842	glycine C-acetyltransferase	22	0.000611
hCG24260	TNF receptor-associated factor 5	1	0.000635
hCG1793181	Olfactory receptor	11	0.000695
hCG27397 ✦	pregnancy-associated plasma protein A	9	0.000818
hCG96444	CD2 antigen (cytoplasmic tail) binding protein 2	22	0.000877
hCG39597	thioredoxin-like 2	10	0.000946
hCG39657 ✦	progesterone receptor	11	0.001048



α tectorin defects disrupt tectorial membrane and lead to high frequency hearing loss.

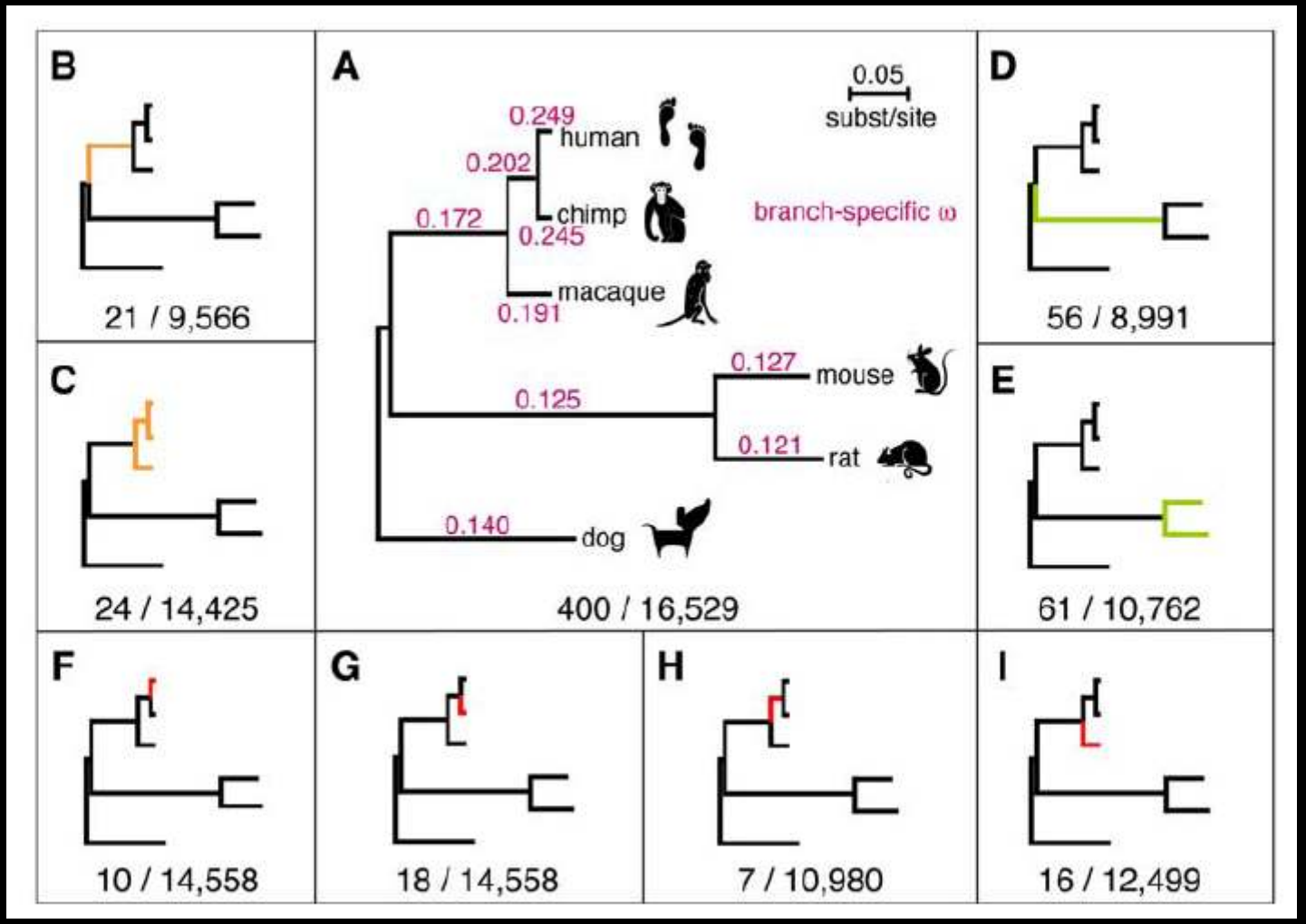


Mustapha et al. 1999. An alpha-tectorin gene defect causes a newly identified autosomal recessive form of sensorineural pre-lingual non-syndromic deafness, DFNB21.
Hum Mol Genet 8:409-412

Growth genes showing human positive selection

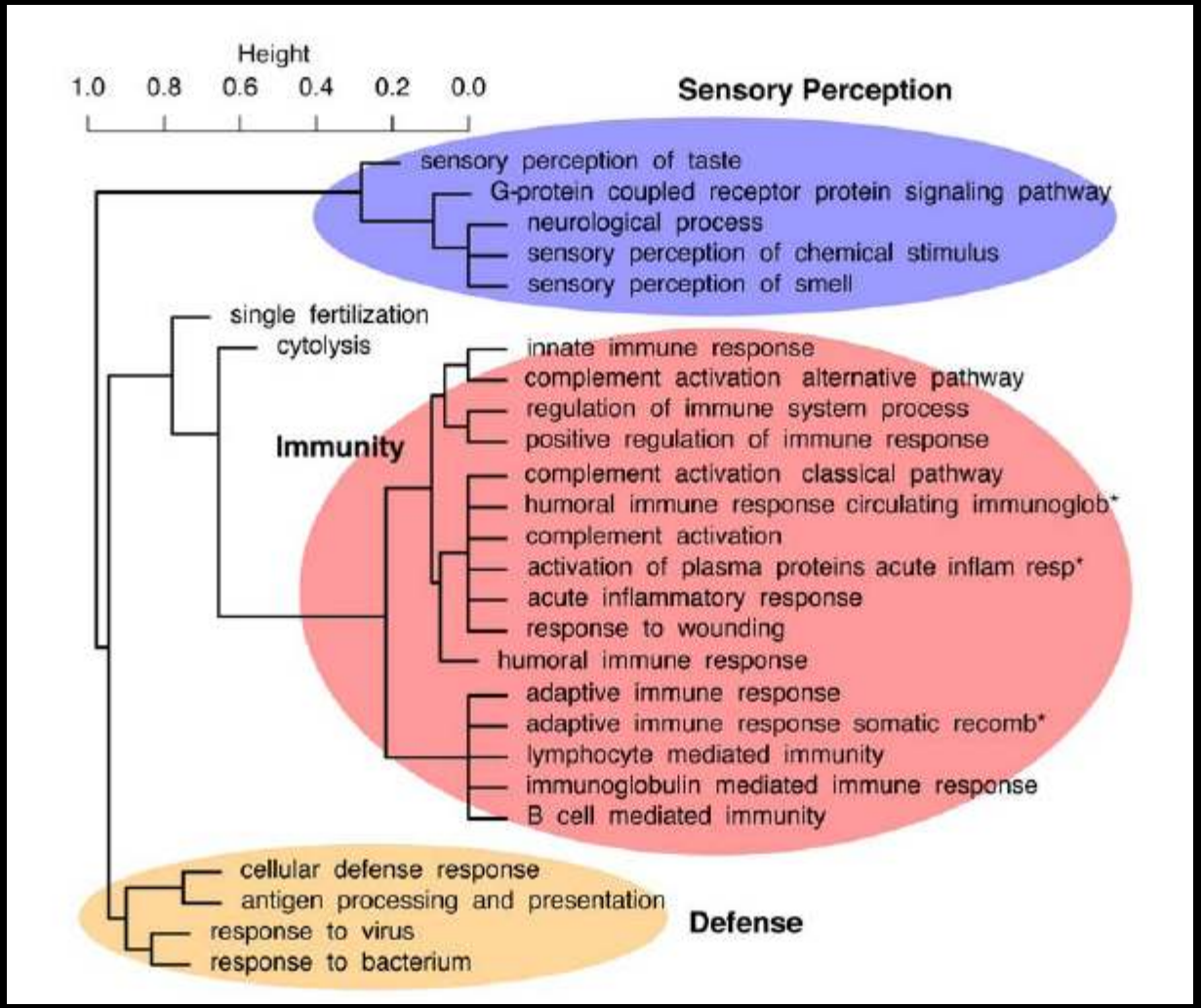
- TSHR - Thyroid stimulating hormone receptor
- GH1 - Growth hormone 1
- GH2 - Growth hormone 2

Inference of selection from branch-specific substitution rate acceleration



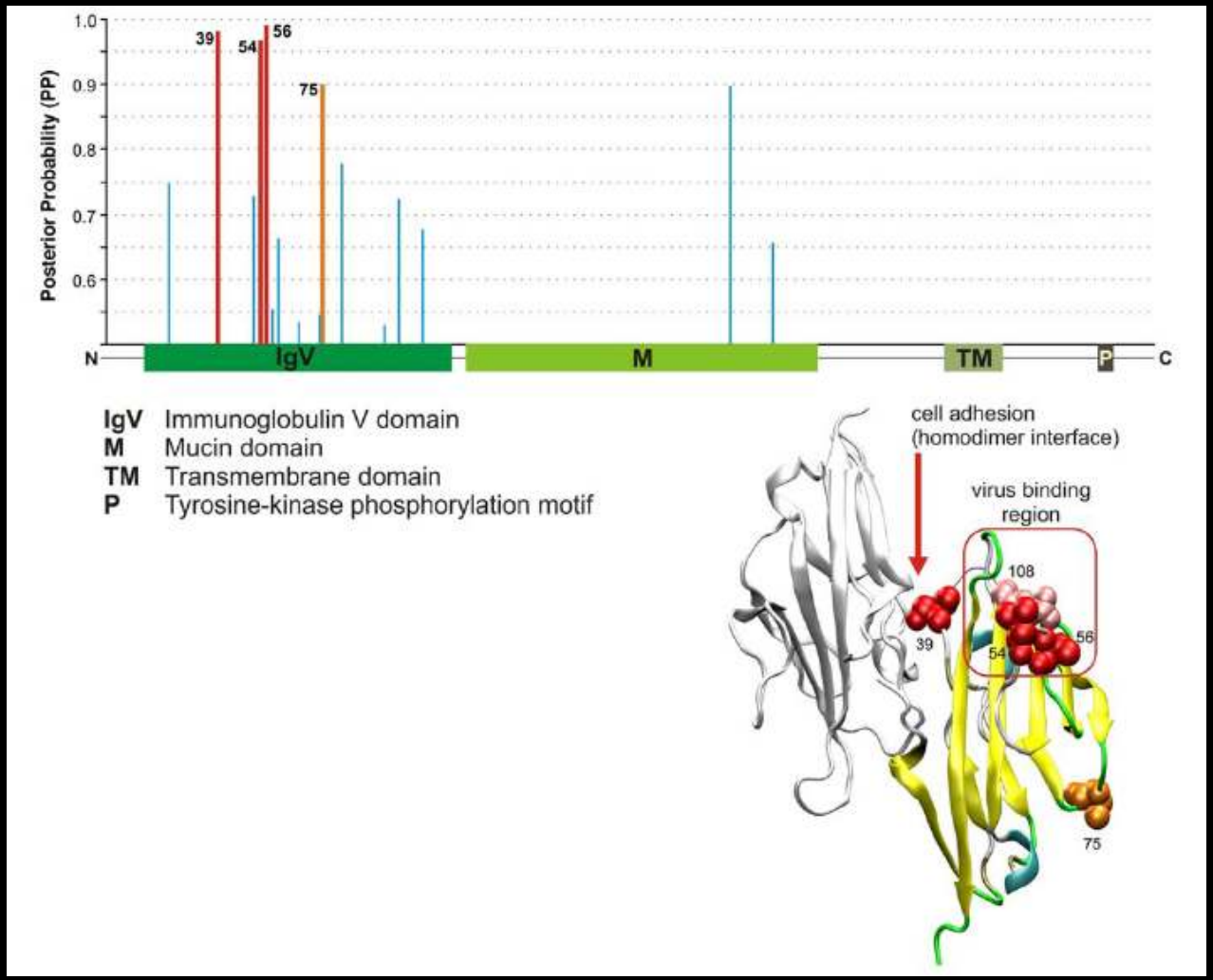
Kosiol et al. (2008) PLosGen 4:e1000144

Gene ontology terms with inflated signs of selection



Kosiol et al. (2008) PLosGen 4:e1000144

Excess amino acid substitutions in Ig variable domain



Kosiol et al. (2008) PLosGen 4:e1000144

Inferring natural selection from DNA sequences

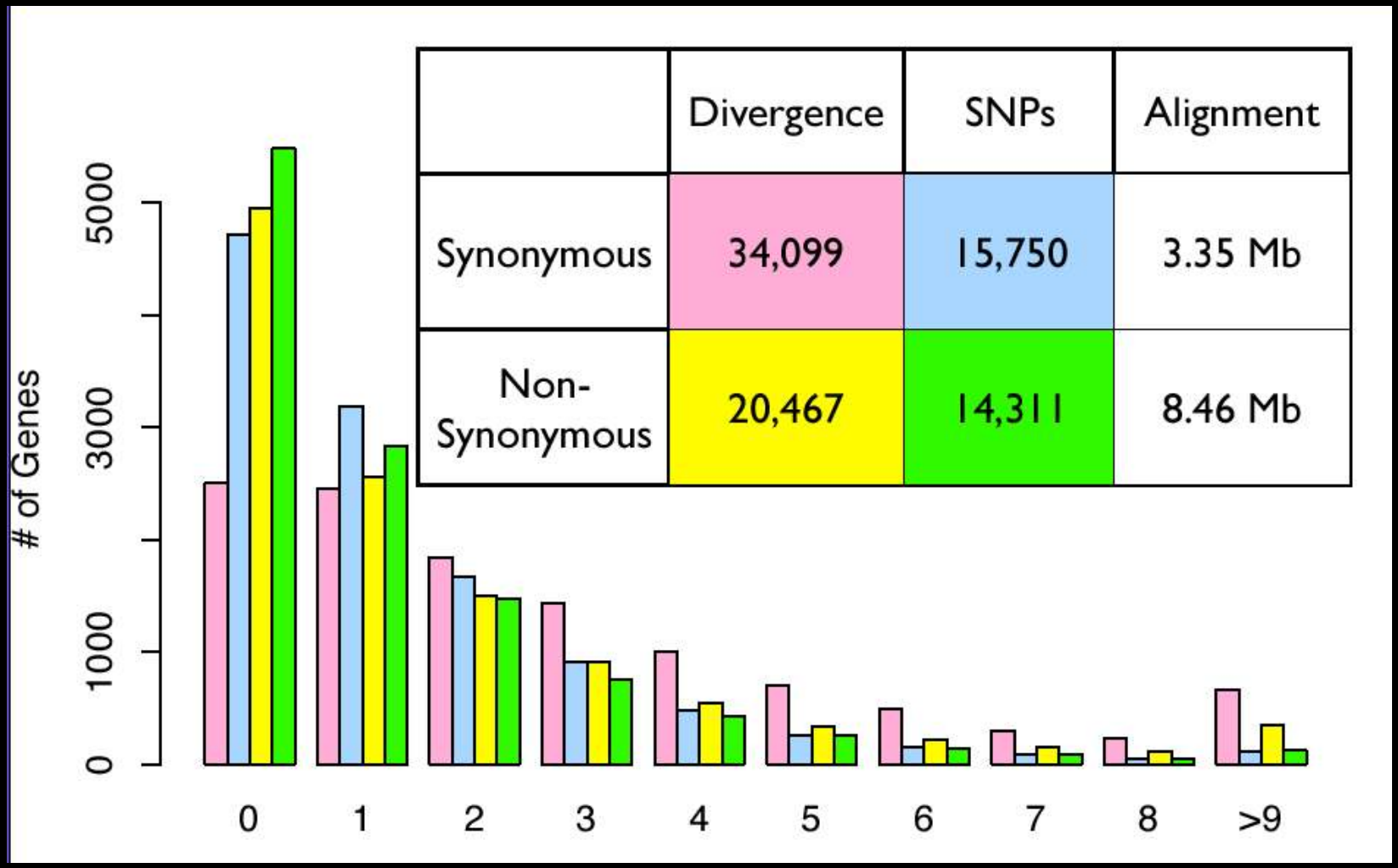
- Substitution rates at syn and nonsyn sites
  - Codon substitution models (PAML)
- **Polymorphism vs. divergence (site counts)**
  - McDonald-Kreitman test; Poisson random field
- Local depression in diversity
  - Kim and Stephan; Tajima's D
- High frequency derived alleles
  - Fay and Wu's H
- Exceptionally long haplotypes
  - Extended haplotype homozygosity; iHS
- Excess divergence between populations
  - Lewontin-Krakauer; F_{ST}

Contrasting polymorphism with interspecific divergence

McDonald-Kreitman Test

	Divergent	Polymorphic
Synonymous	n_{11}	n_{12}
Replacement	n_{21}	n_{22}

Counts of divergent sites and SNPs



$G = 816.03, \quad P < 2 \times 10^{-16}$

Bustamante et al. (2005) Nature 437:1153-1157.

McDonald-Kreitman Poisson random field model

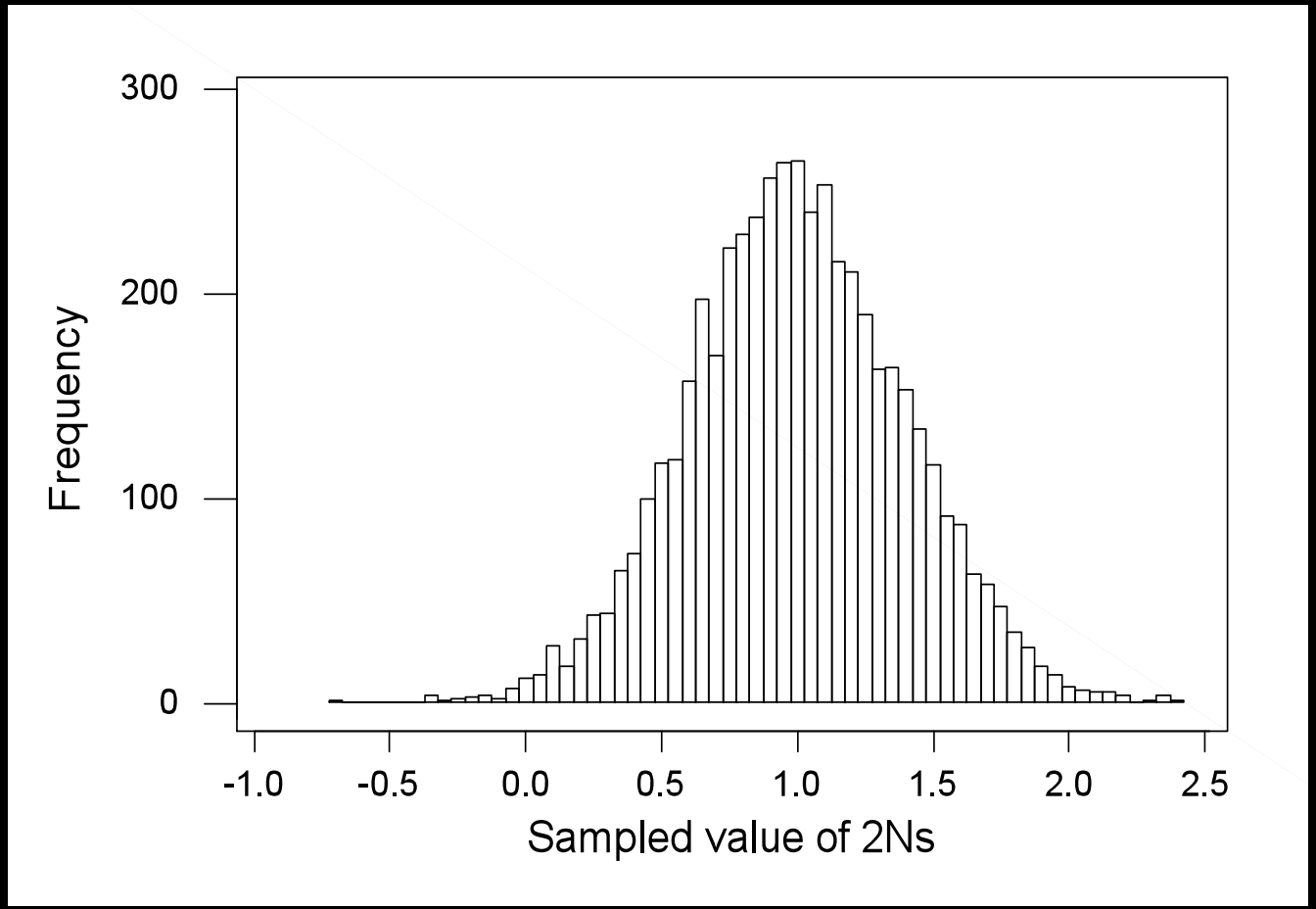
	Divergent	Polymorphic
Synonymous	$2\mu_s\tau$	$4N\mu\Sigma(1/i)$
Replacement	**F(μ, τ, γ=2*Ns*)**	**G(4*N*μ, γ=2*Ns*)**

Carlos Bustamante

Fit the above parameters by MCMC

Sawyer and Hartl (1992) Genetics 132:1161

Posterior density for 2Ns
(obtained for each individual gene)



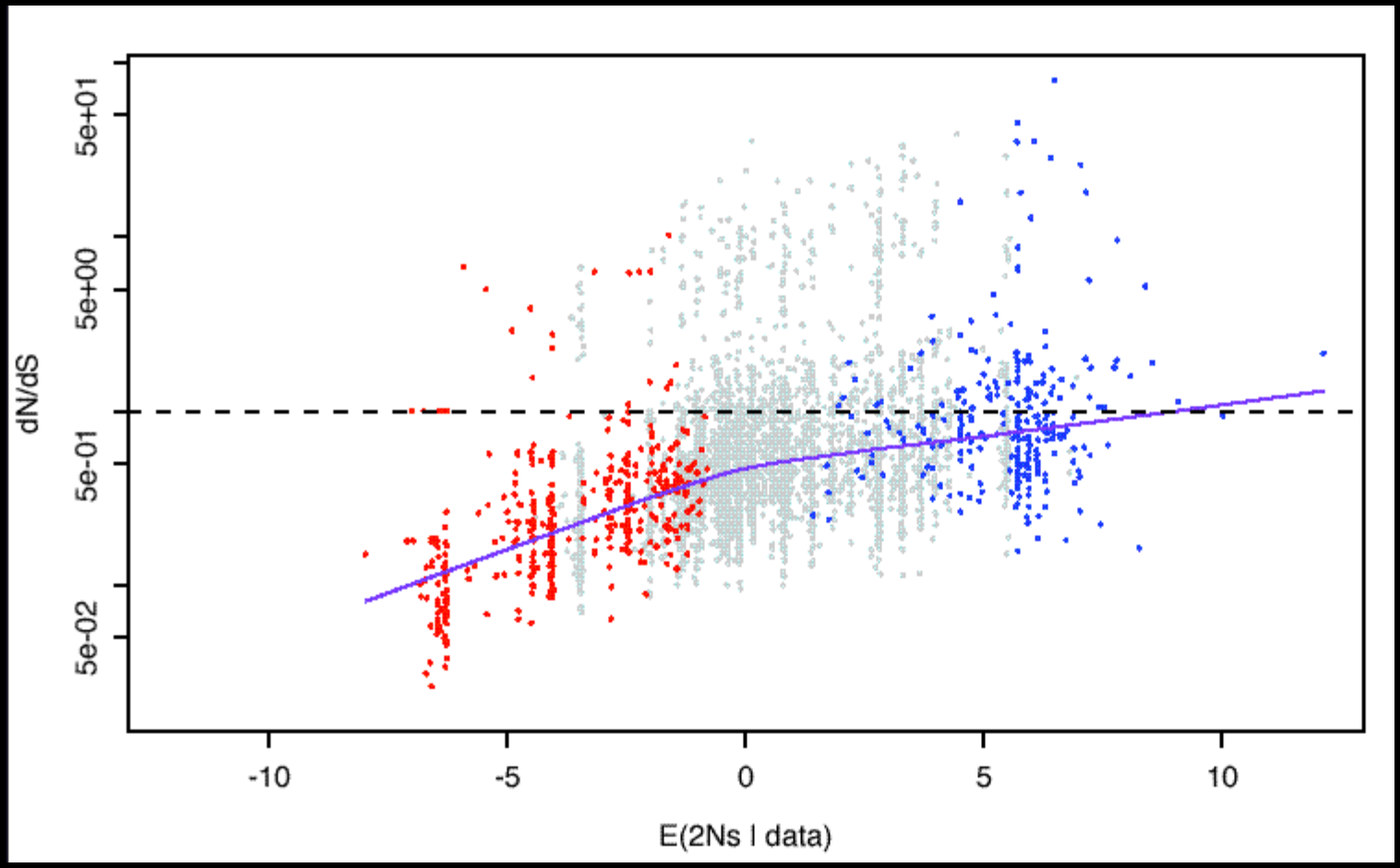
The above indicates that Pr($\gamma > 0$) > 0.97

Inferences from human-chimp tests: outlier processes

Biological process	Number of genes	*p*-value
Immunity and defense	566	0.0000
Sensory perception	274	0.0000
Protein biosynthesis	178	0.0000
Chemosensory perception	173	0.0000
Olfaction	155	0.0000
T-cell mediated immunity	145	0.0000
B-cell- and antibody-mediated immunity	107	0.0000
Gametogenesis	65	0.0002
Spermatogenesis and motility	31	0.0003
Translational regulation	28	0.0003
Biological process unclassified	4127	0.0005
Inhibition of apoptosis	37	0.0044

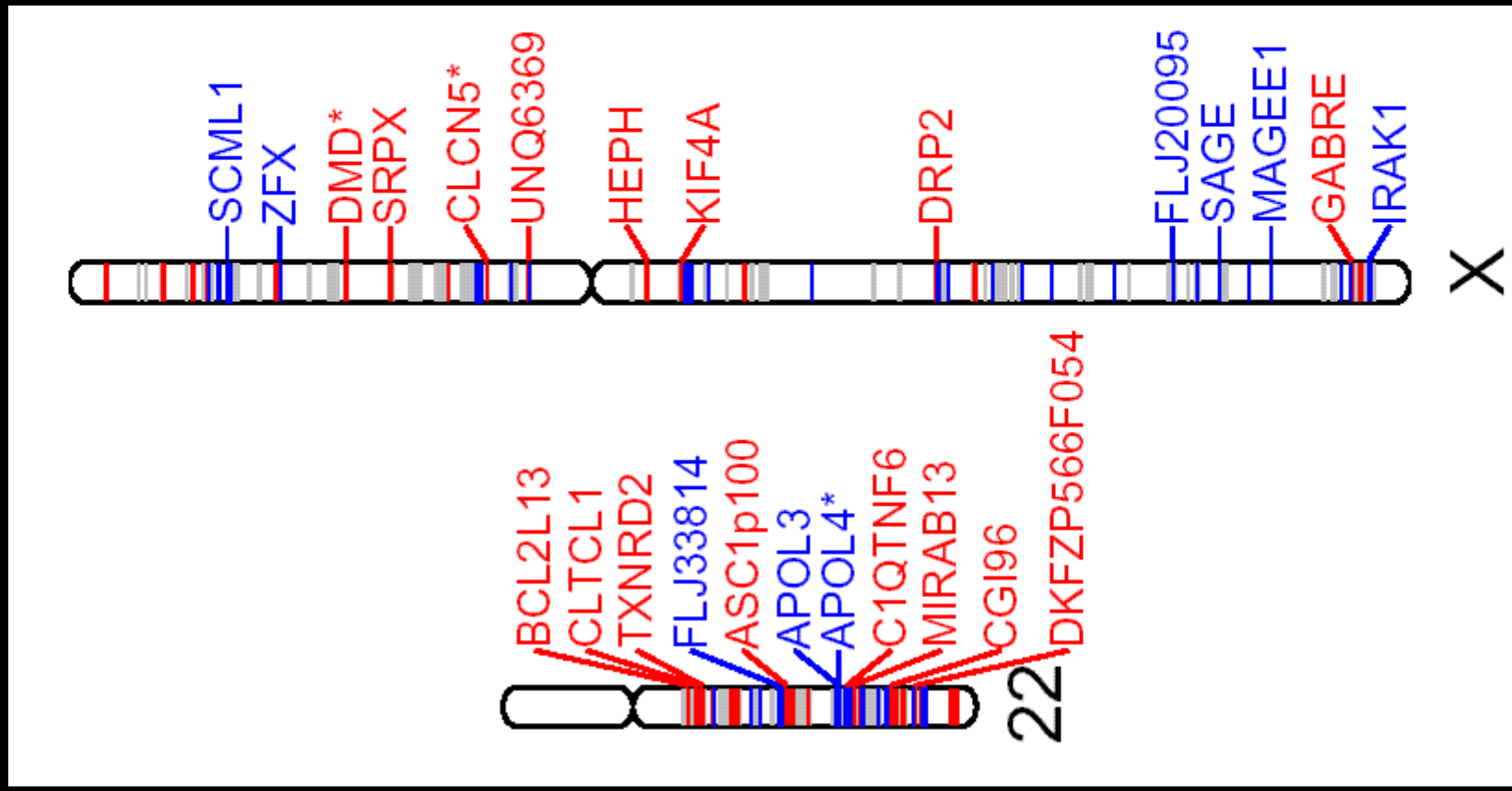
$2Ns$ is correlated with d_N/d_S

(r = 0.377, P < 10^{-9})



Note! Can detect significance even if d_N/d_S << 1.

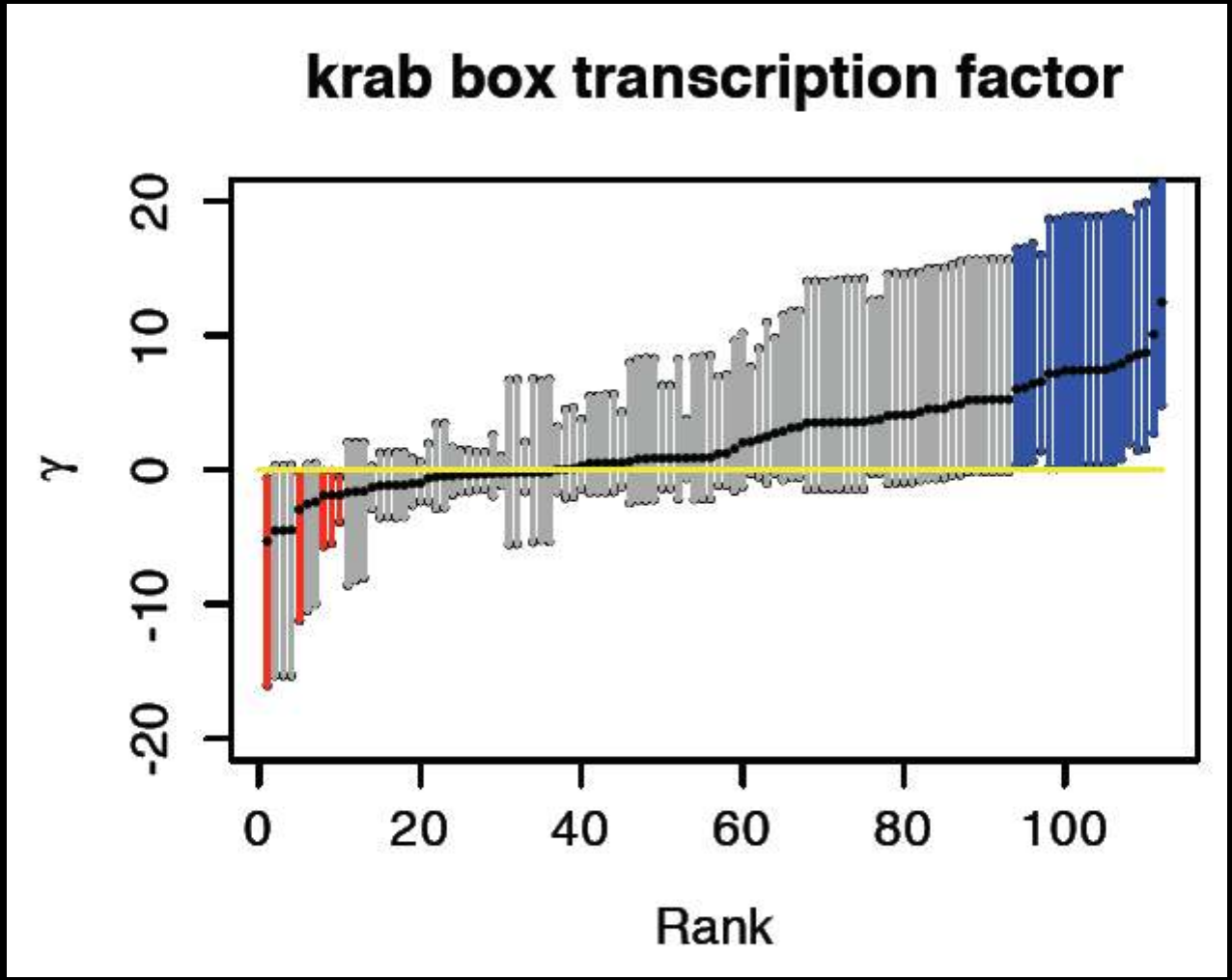
Positively selected Negatively selected



Many negatively selected genes are associated with known Mendelian disorders

Huntingtin has Pr(γ<0) greater than 99%

As a class, transcription factors show strong positive selection

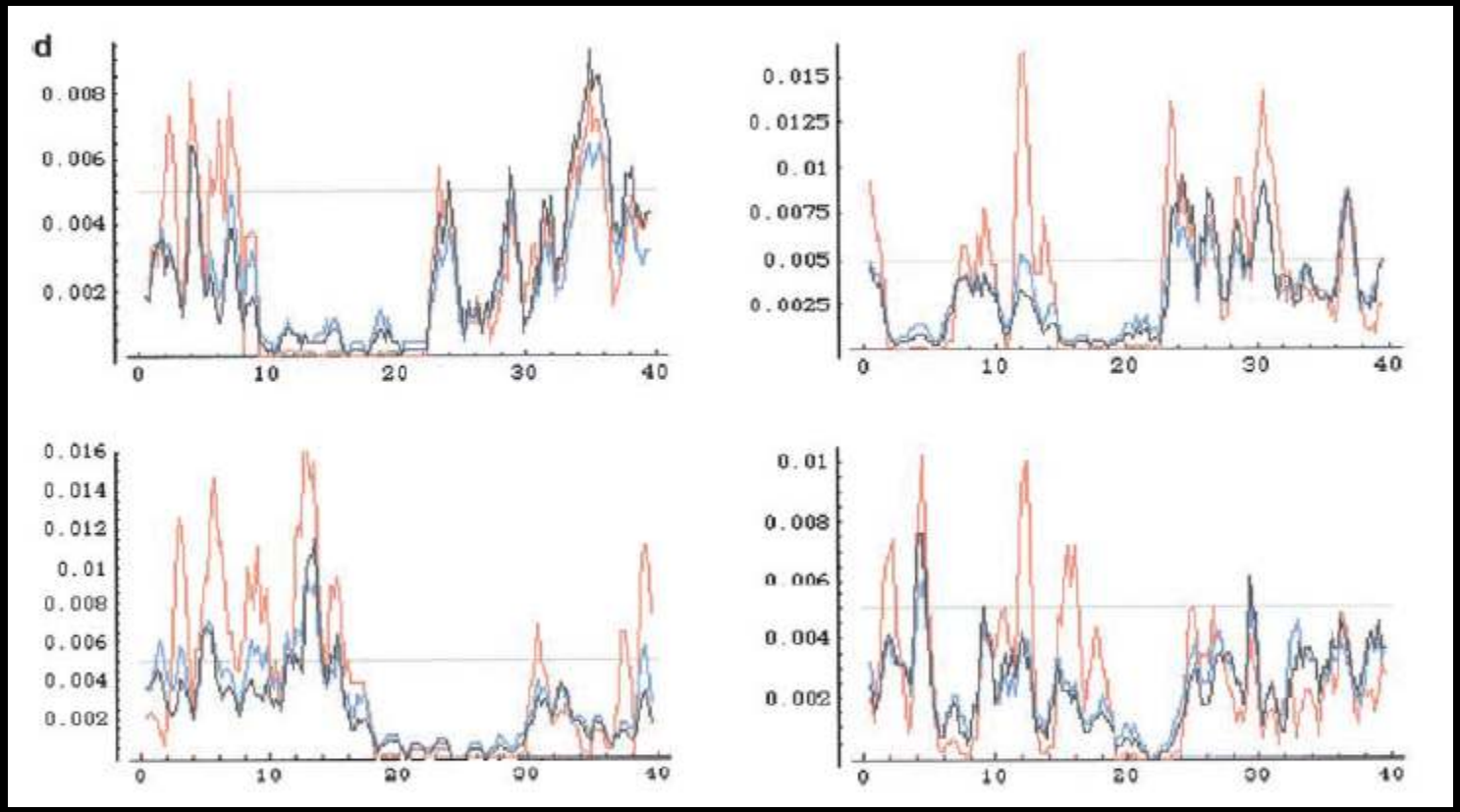


39/240 transcription factors had Pr(γ>0) > 97.5%

Inferring natural selection from DNA sequences

- Substitution rates at syn and nonsyn sites
  - Codon substitution models (PAML)
- Polymorphism vs. divergence (site counts)
  - McDonald-Kreitman test; HKA test
- **Local depression in diversity**
  - Kim and Stephan; Tajima's D
- High frequency derived alleles
  - Fay and Wu's H
- Exceptionally long haplotypes
  - Extended haplotype homozygosity; iHS
- Excess divergence between populations
  - F_{ST}

Selective sweeps leave troughs of diversity



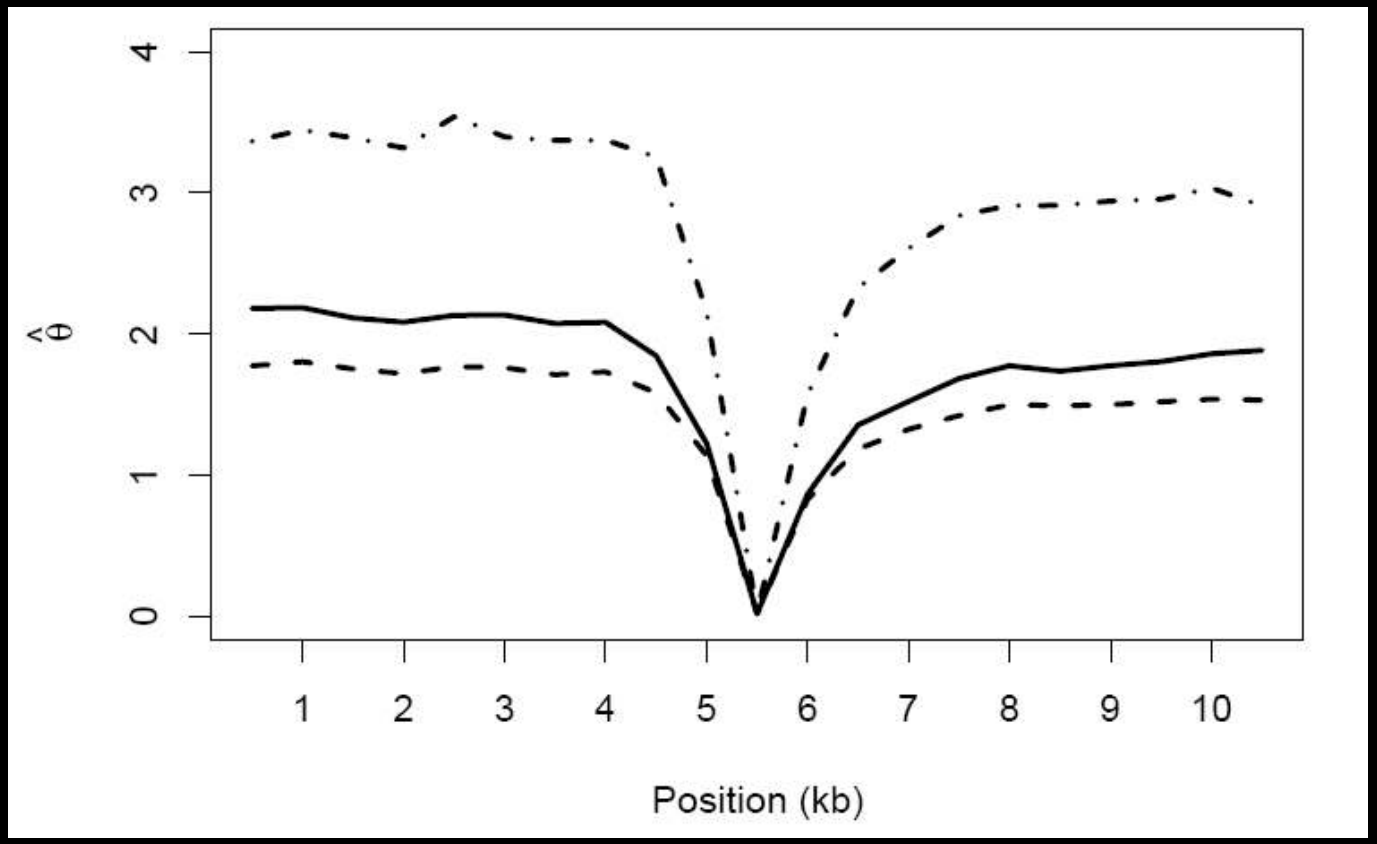
Nucleotide diversity

Position along chromosome

Kim and Stephan (2002) Genetics 160: 765–777

Bottlenecks and pseudoSweeps

- Bottleneck + growth leaves sweep-like troughs of diversity



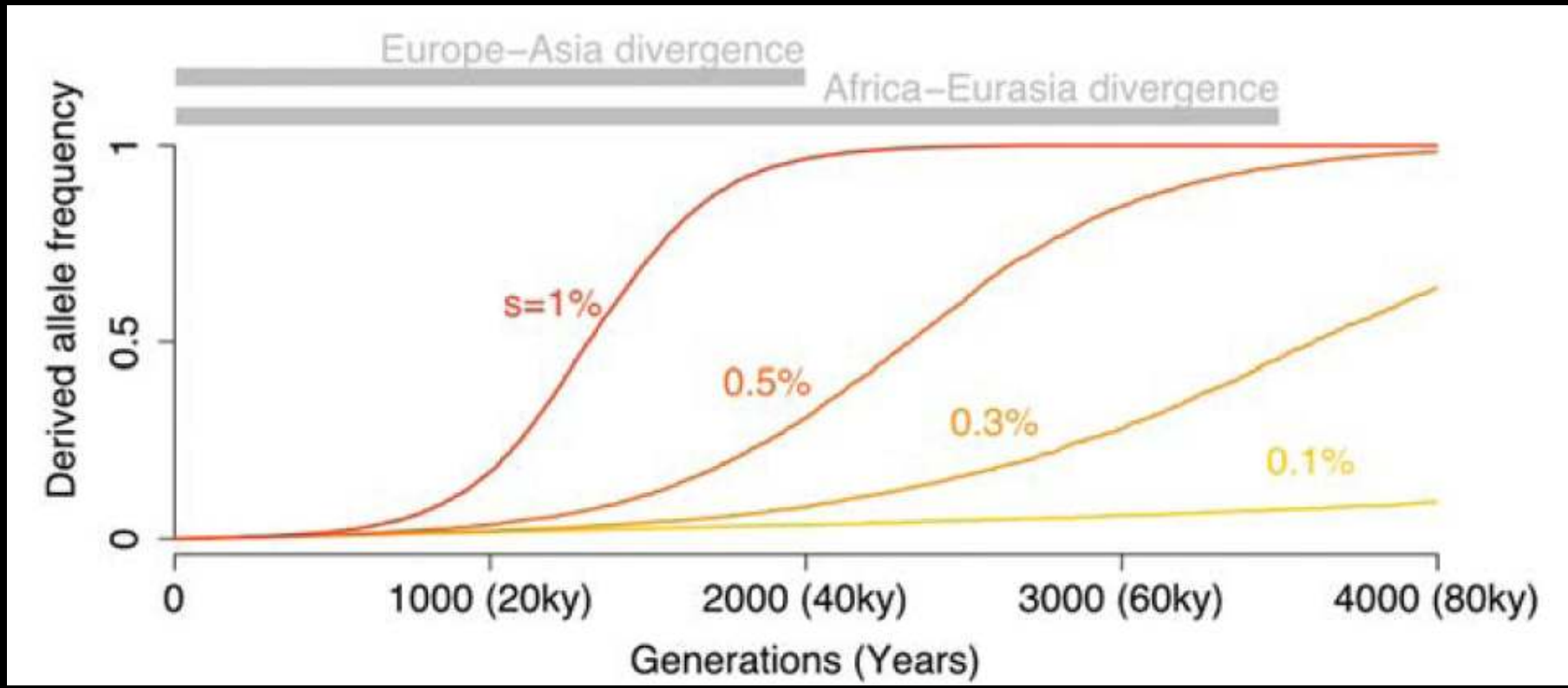
θ_H

θ_π

θ_w

Bottleneck simulation: 10 kb, θ = 0.01/site, ρ = 0.1/site,
t_r = 0.004, d = 0.015 and f = 0.03

Selective sweeps and allele age (different methods have different time-depth)



Coop et al. (2009) Plos Genet 5:e1000500

Using SFS for inference of selection

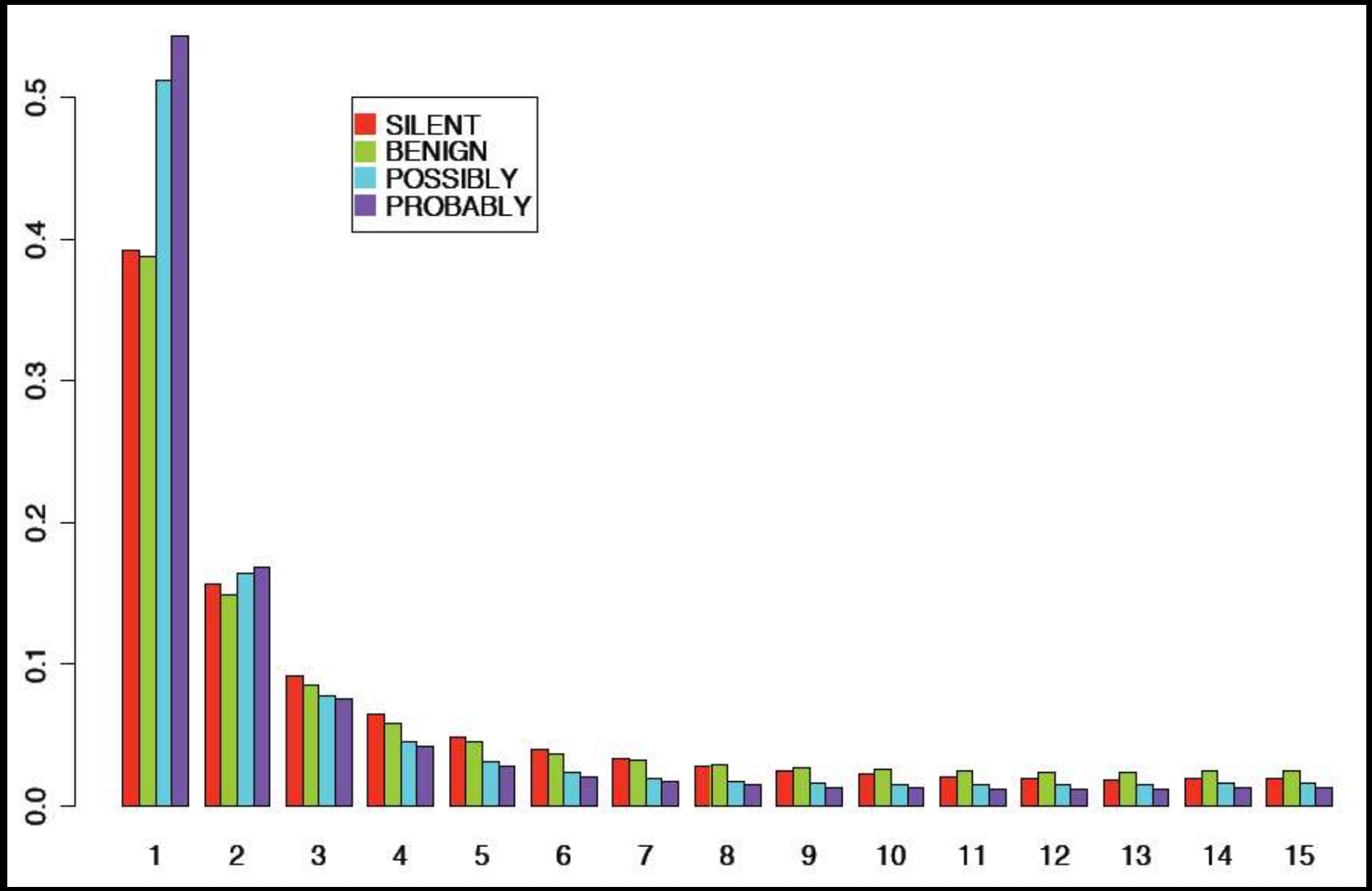
- Can correct for ascertainment bias if ascertainment scheme is thoroughly known.
- Correct for admixture

$$\Pr[x_l^{(i)} = (j,v)] = 2\left(\sum_k p_{klj} q_k^{(i)}\right)\left(\sum_k p_{klv} q_k^{(i)}\right)$$

  - where $q_k^{(i)}$ is the proportion of individual *i*'s genome that came from population *k*, and p_{klj} is the frequency of allele *j* at locus *l* in population *k*
- Correct for demographic history
  - Use composite likelihood to fit growth and bottleneck parameters to silent sites (assumed to be neutral)
  - Use these demographic parameters to correct SFS

Rasmus Nielsen

The Site Frequency Spectrum



Frequency of this SNP class

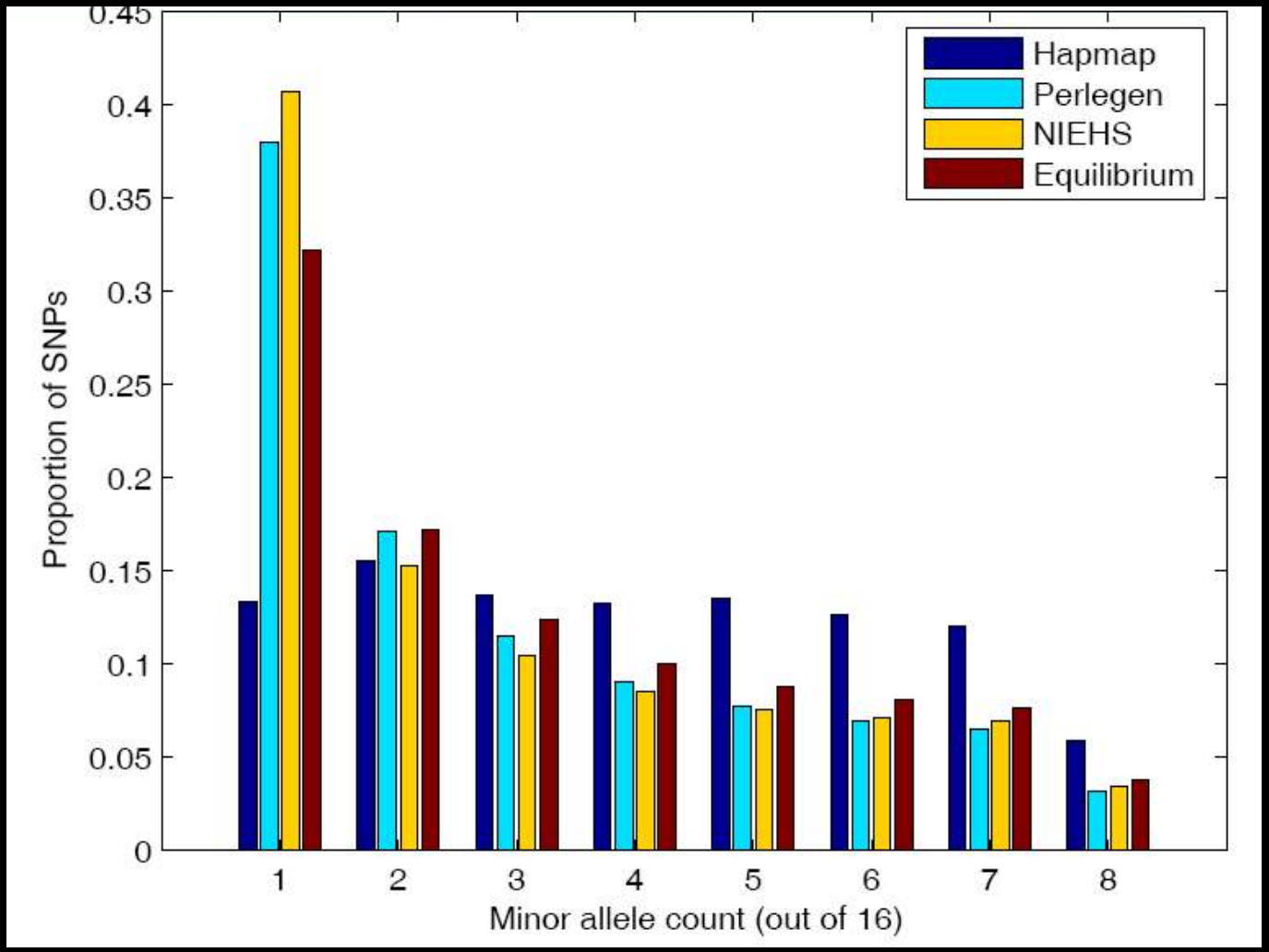
Count of the derived allele

Caveats in using SFS to infer selection

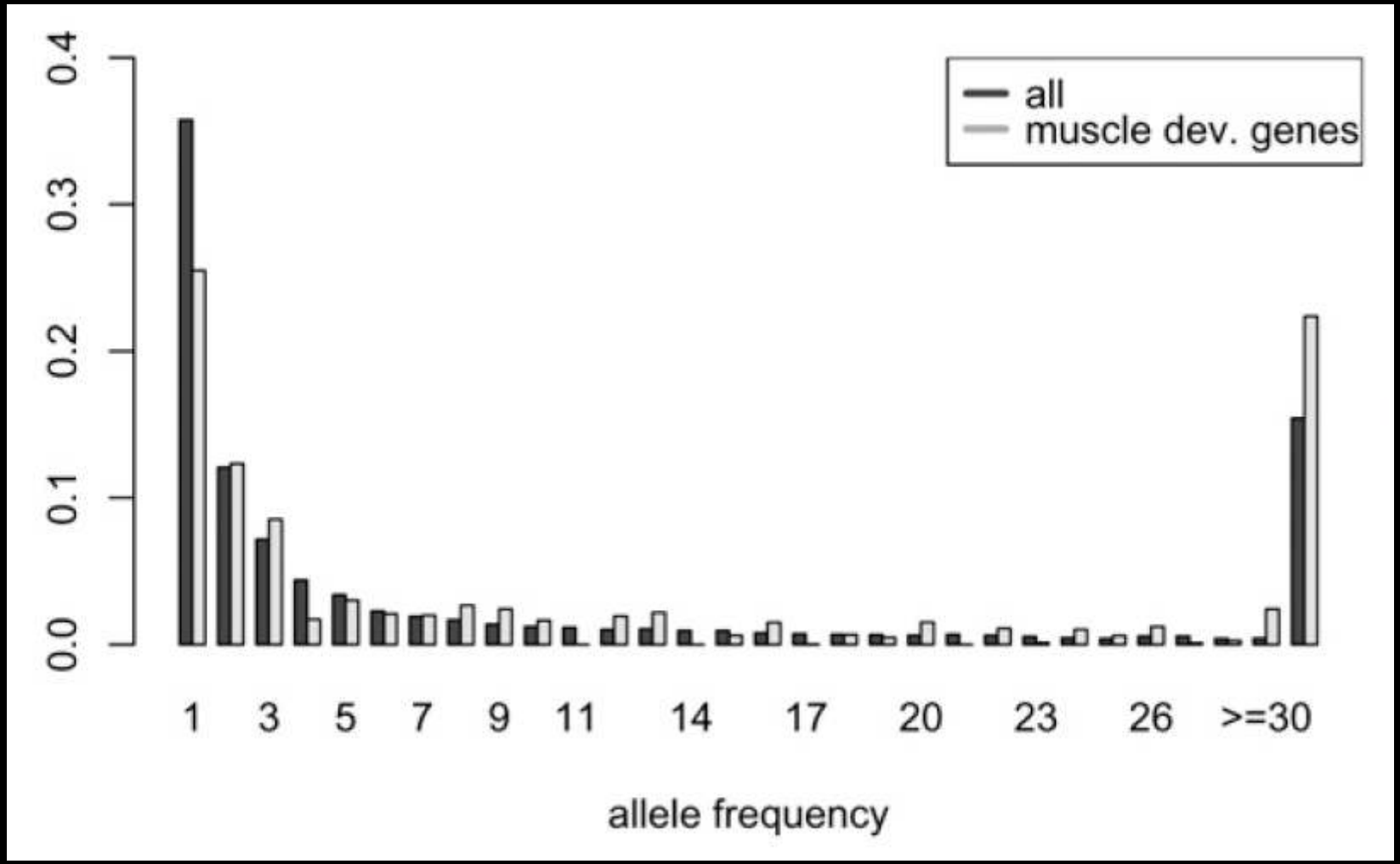
- Ascertainment bias in SNP identification
- Admixture / population stratification
- Demographic history confounds the tests

Ascertainment bias

Difference in SFS of phase I HapMap vs Perlegen SNPs arose from differences in SNP discovery.



Contrasting SFS by Gene Ontology categories



Nielsen et al. (2009) Genom Res 19:838–849

Tests of selection based on corrected SFS

- Two-dimensional SFS – test homogeneity of SFS in Europeans vs. African
- Mann-Whitney U – compare each gene to the genome-wide frequency spectrum
- HKA – standard Hudson-Kreitman-Aguadé test
- F_{ST} – test significance by Fisher exact test
- Fisher Pooled probability test

Some genes with high divergence and low polymorphism

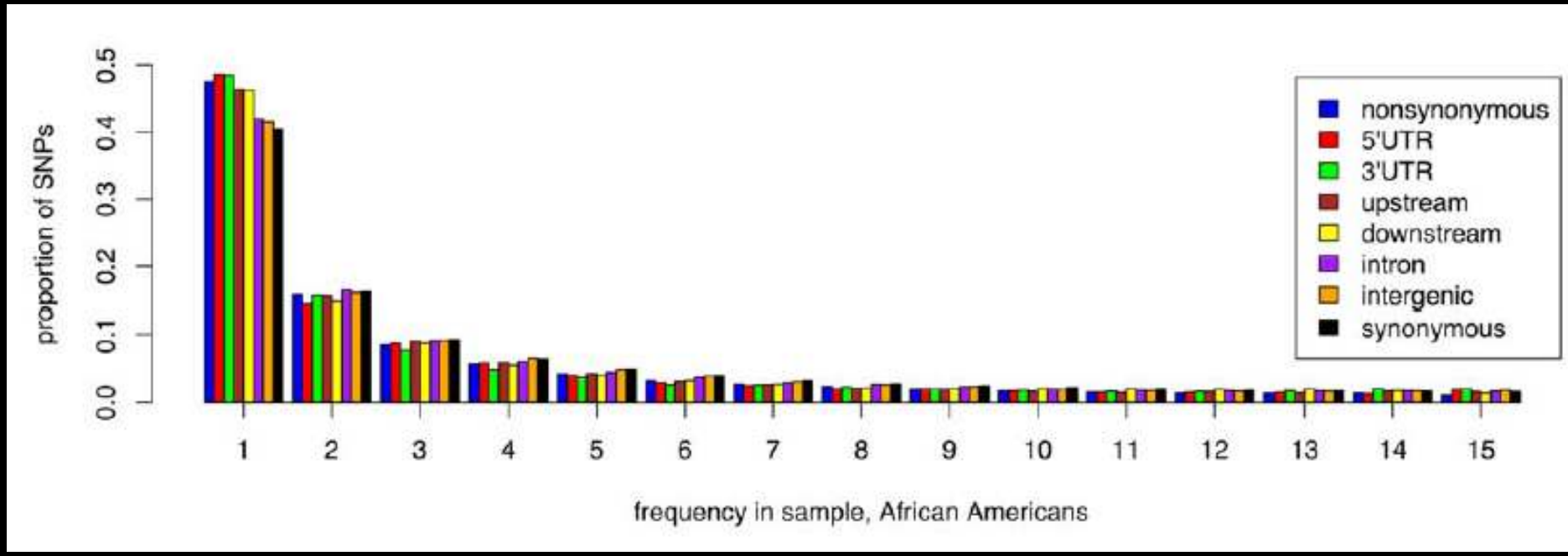
- TCOF1 – Treacher-Collins syndrome
- IRAK1 – kinase associated with interleukin-1 receptor
- ZNF565 – and several other KRAB box TFs
- MAGEC2 – testis-specific cancer antigen
- SAGE1 – X-linked, overexpressed in tumors
- ZFX – sex determination

Noncoding SNPs and regulatory polymorphism

- Celera Genomics sequenced 18 M reads in regions flanking human genes
- Human-mouse conserved regions targeted
- Identified 26,675 SNPs in these regions in 39 humans (with chimp as outgroup)

Dara Torgerson

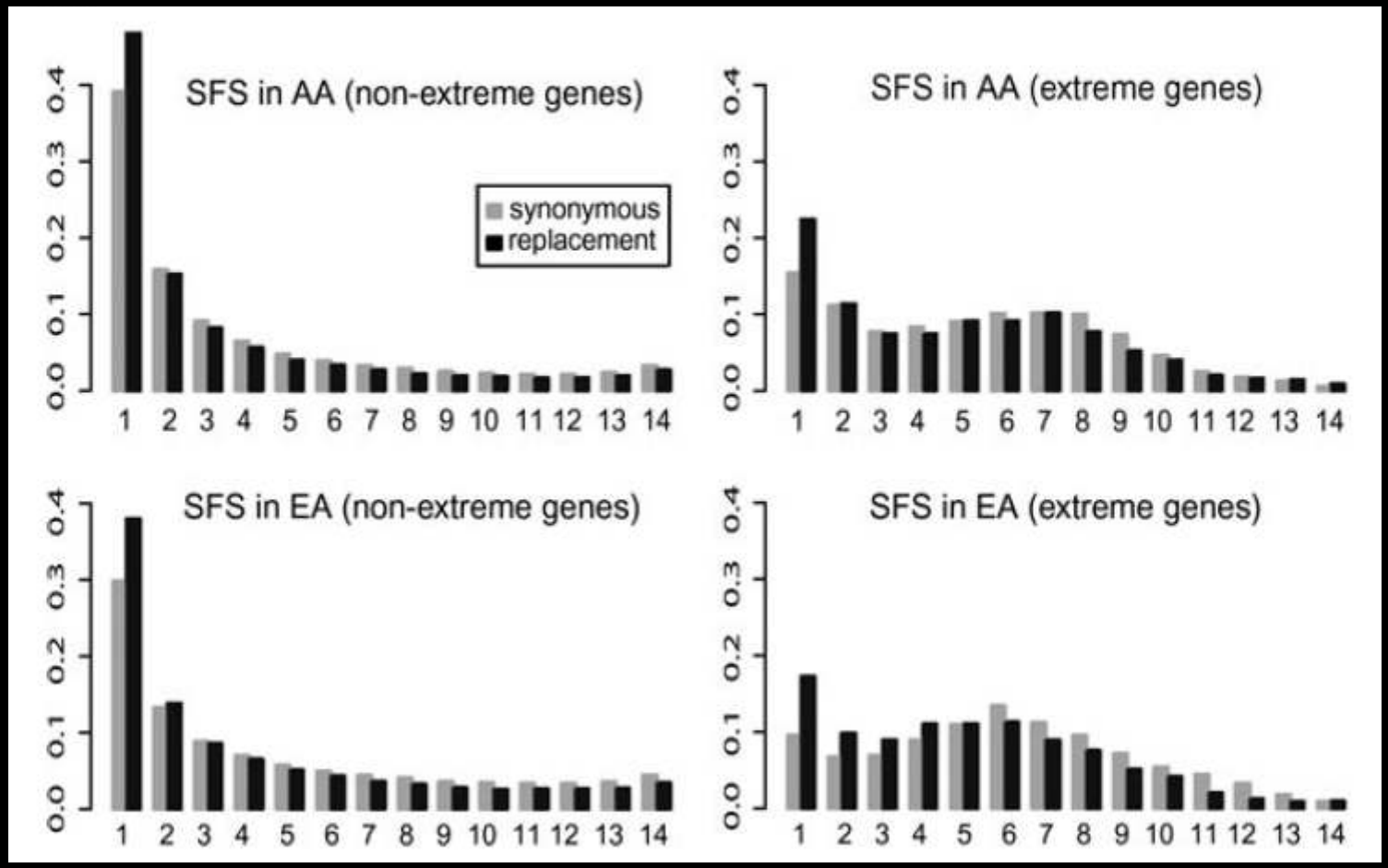
Contrast in SFS in vs. out of *cis* -regulatory modules



Clear signature of purifying selection in UTRs.

Torgerson et al. (2009) Plos Gen 5:e1000592

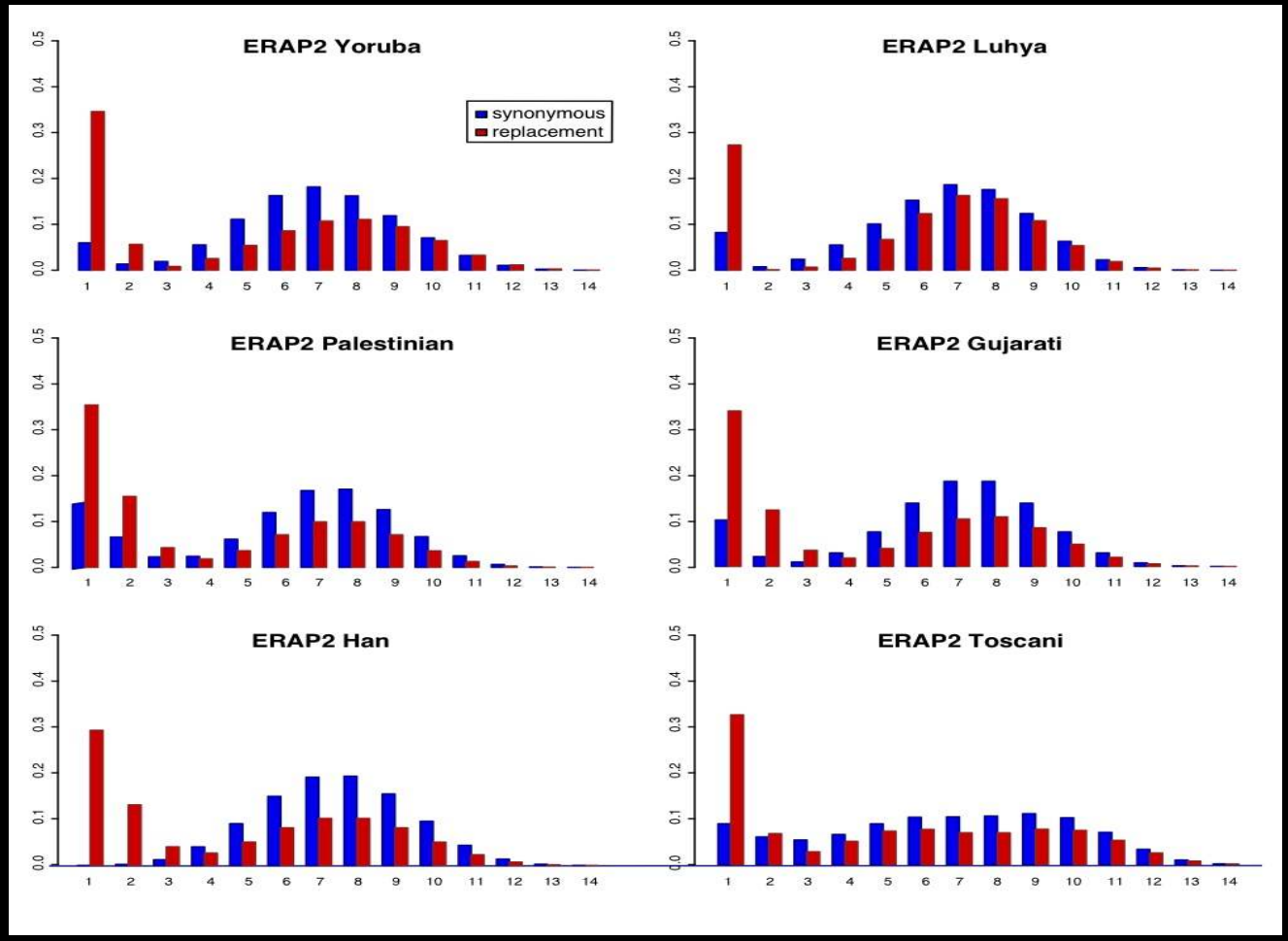
Seeking evidence for balancing selection: Excess intermediate frequency alleles



Aida Andres

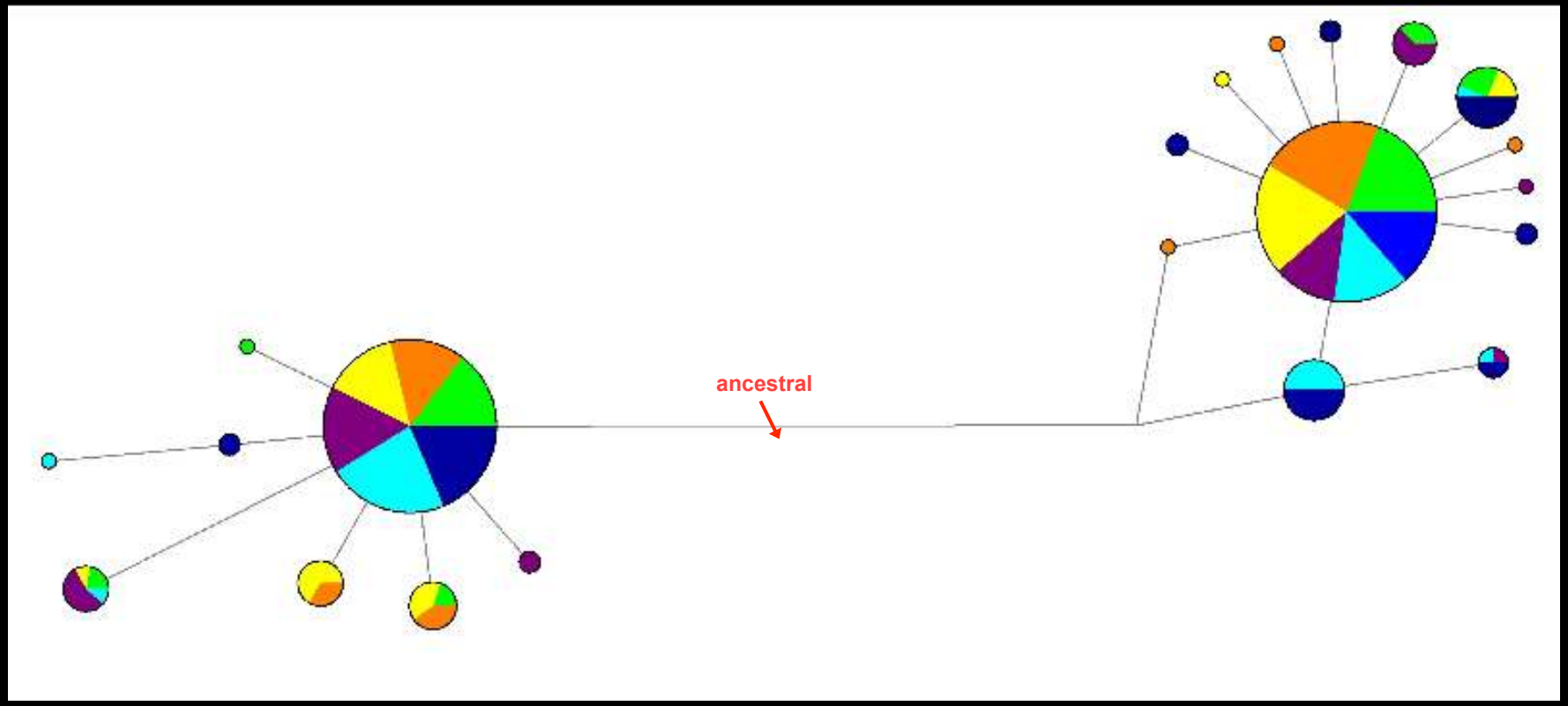
Andres et al. (2009) Mol. Biol. Evol. 26:2755–2764

Site Frequency Spectra for ERAP2



Andres et al. (2009) Mol. Biol. Evol. 26:2755–2764

ERAP2 haplotype network



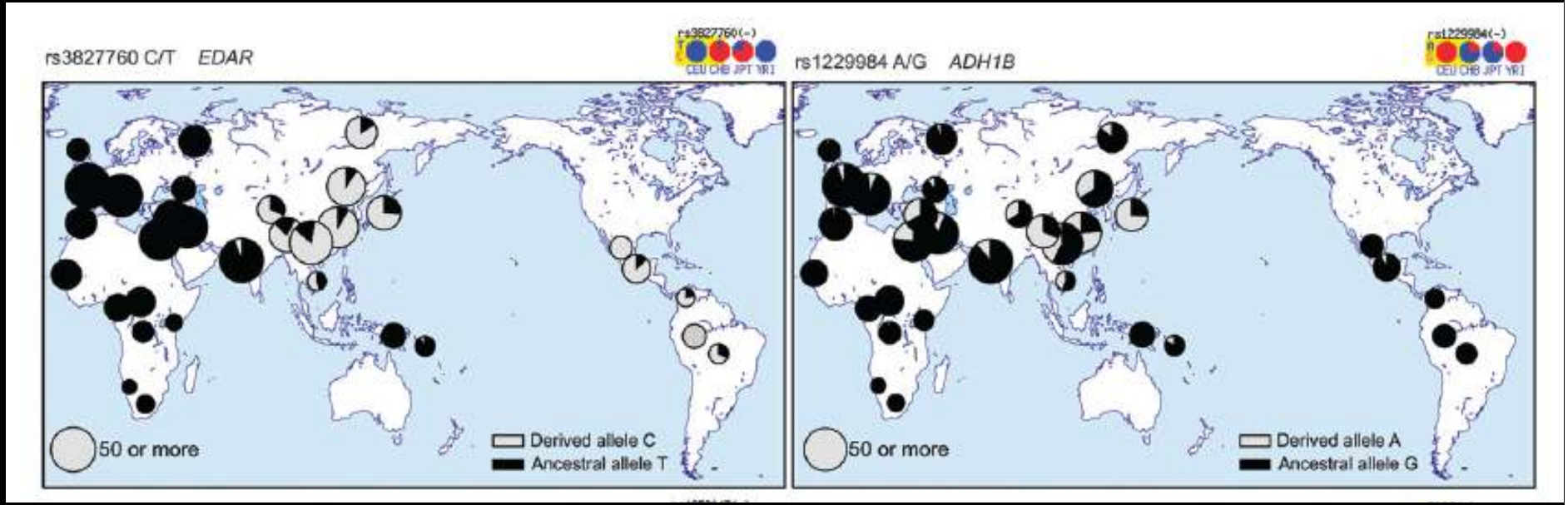
Yoruba
Luhya
Palestinian
Gujarati
Han
Toscani

Andres et al. (2009) Mol. Biol. Evol. 26:2755–2764

Inferring natural selection from DNA sequences

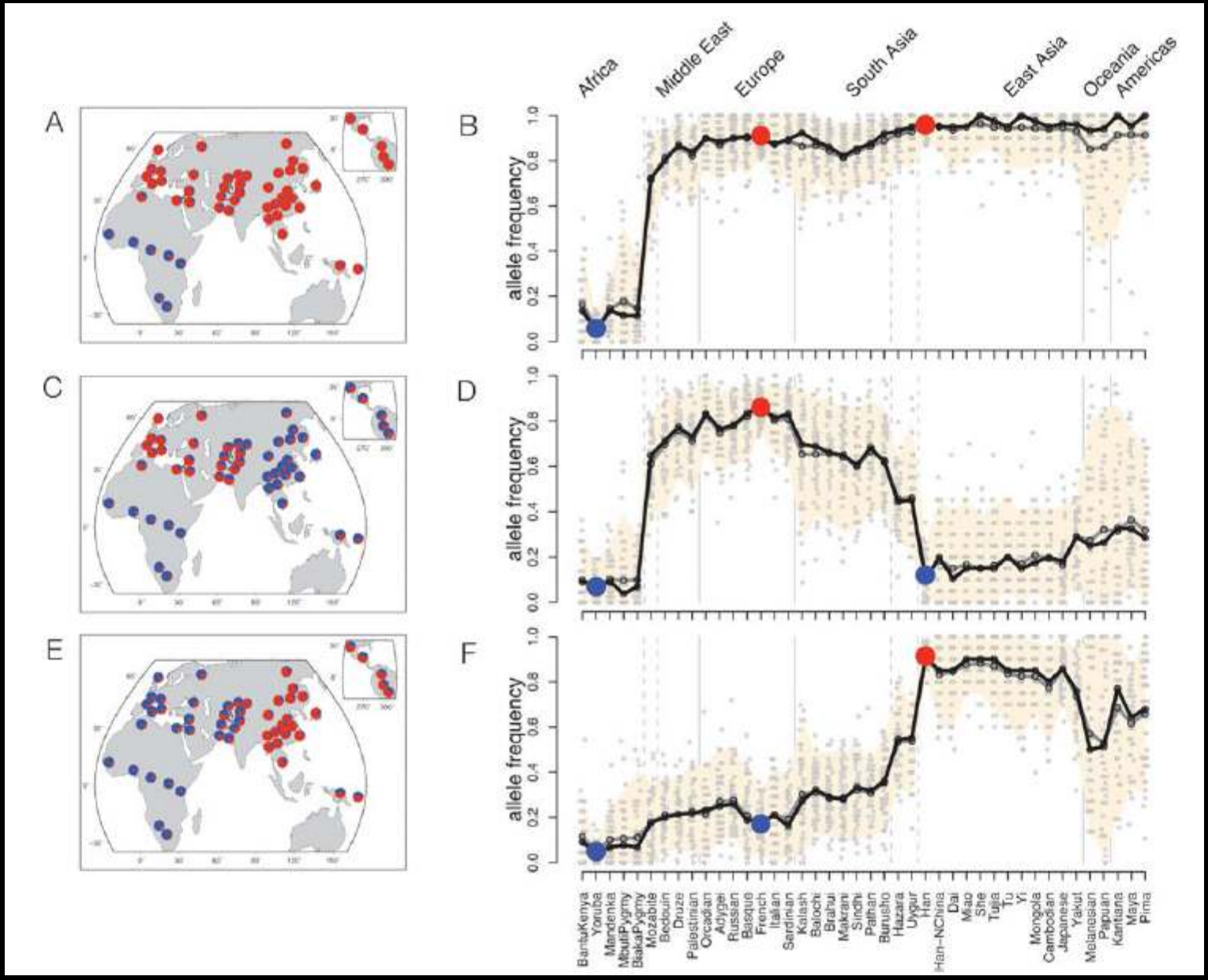
- Substitution rates at syn and nonsyn sites
  - Codon substitution models (PAML)
- Polymorphism vs. divergence (site counts)
  - McDonald-Kreitman test; HKA test
- Local depression in diversity
  - Kim and Stephan; Tajima's D
- High frequency derived alleles
  - Fay and Wu's H
- **Exceptionally long haplotypes**
  - Extended haplotype homozygosity; iHS
- **Excess divergence between populations**
  - F_{ST}

A large proportion of genes with a signature of natural selection have elevated among-population divergence



Xue et al. (2009) Genetics 183: 1065–1077

Wide diversity among geographic patterns of SNPs



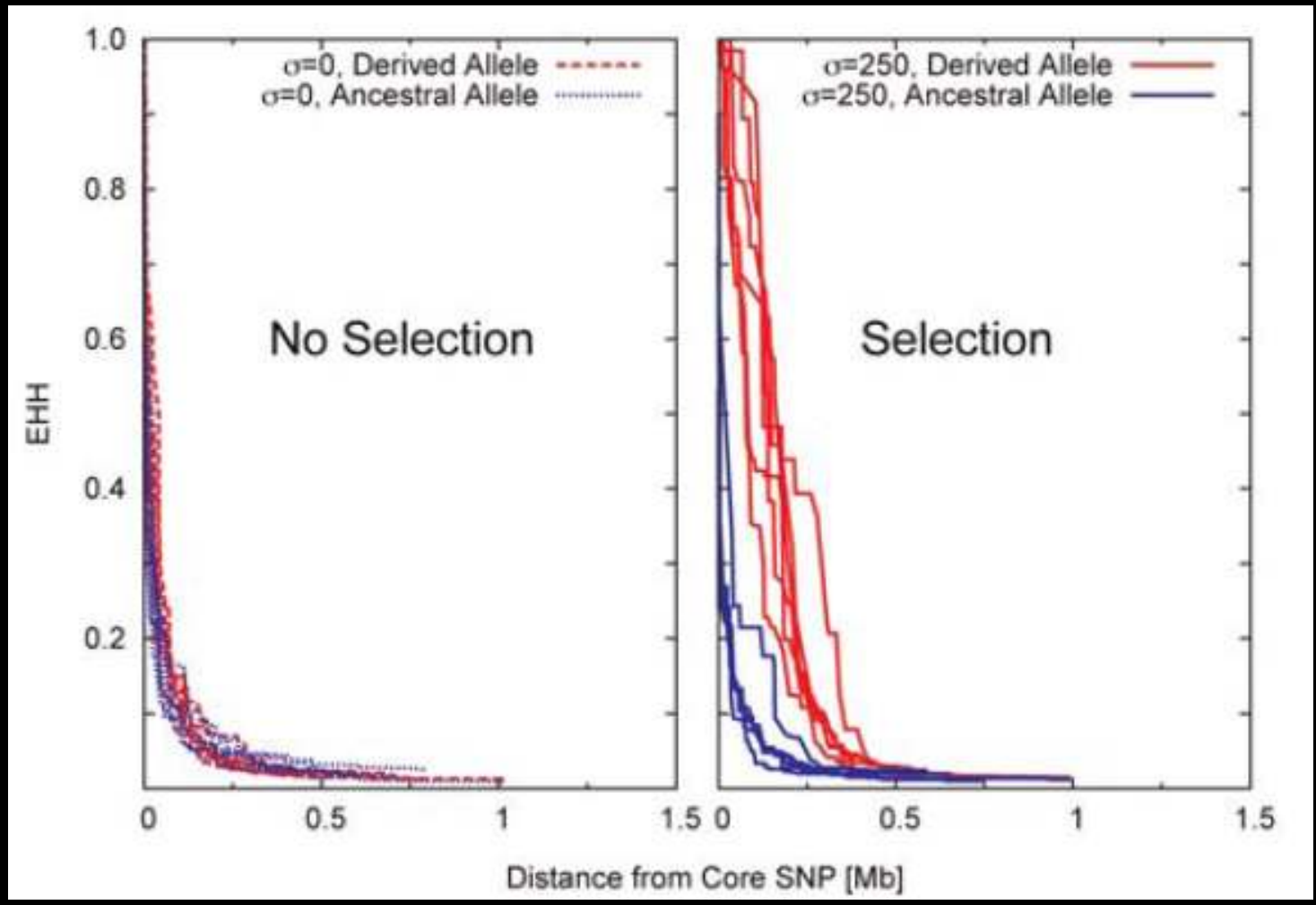
Coop et al. (2009) Plos Genet 5:e1000500

Exceptional divergence in allele frequency found in the 1000 genomes project

SLC24A5	skin pigmentation
SLC45A5	skin pigmentation
DARC	immunity
TLR1	immunity
ADH1B	alcohol metabolism
EDAR	hair morphology
ABCC11	ear wax consistency

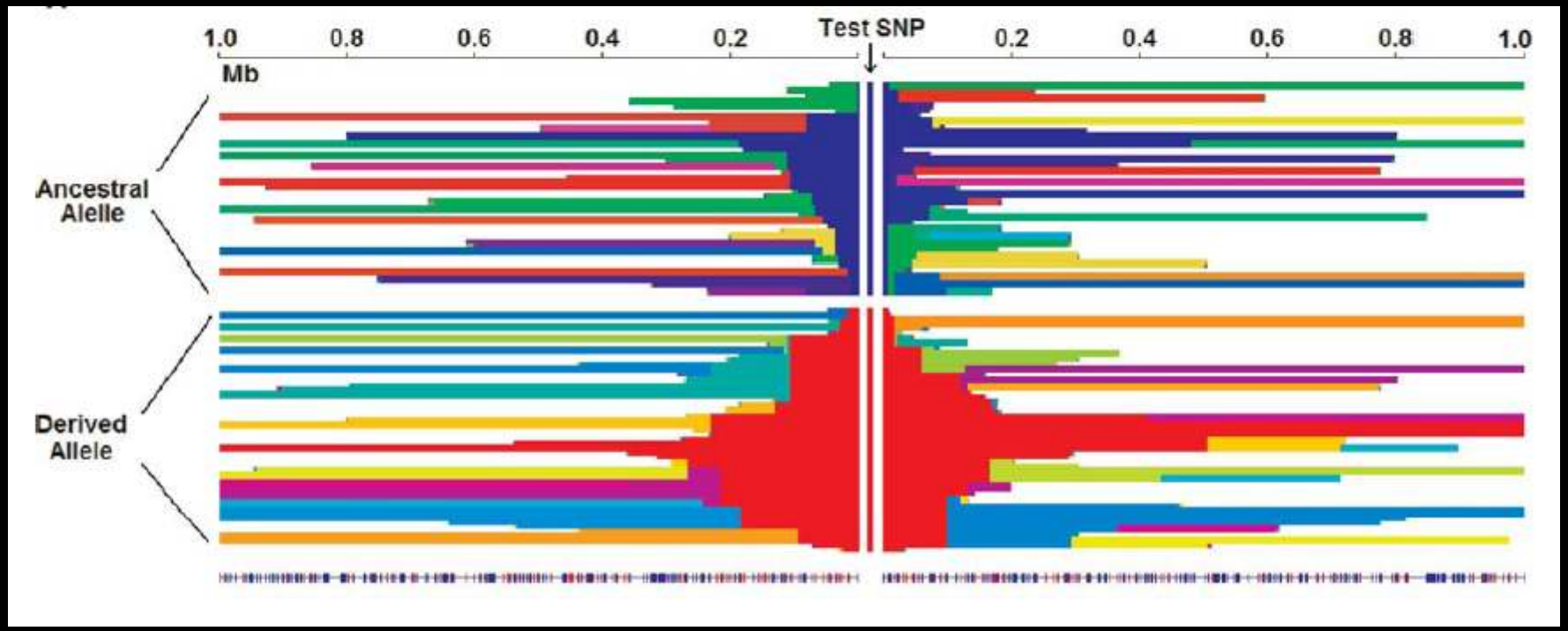
But the 1000 genomes project finds thousands of other SNPs with high divergence between populations

Positive selection drives extended haplotypes



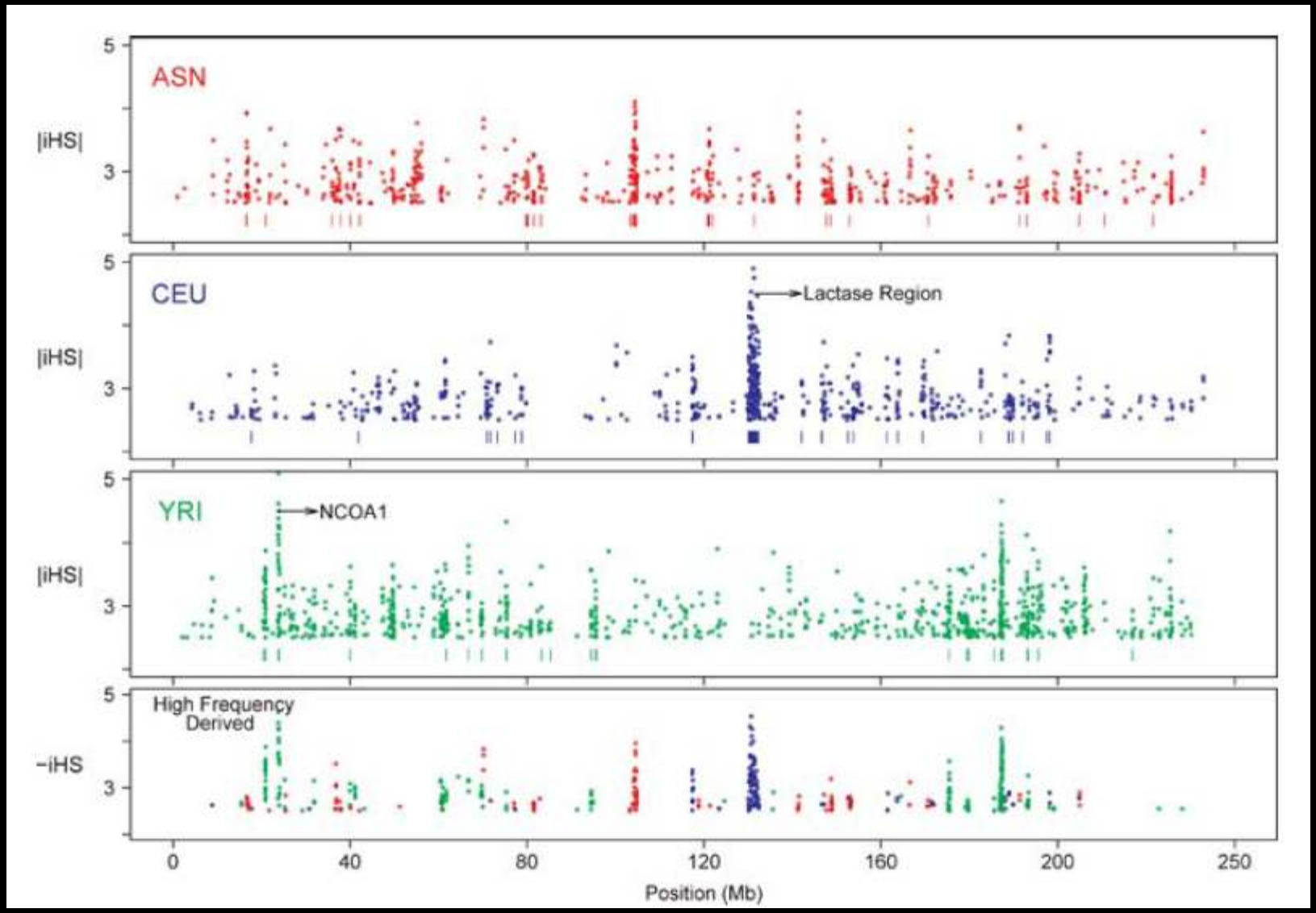
Voight et al. (2006) Plos Biol 4:e72

High frequency derived alleles associated with long haplotypes



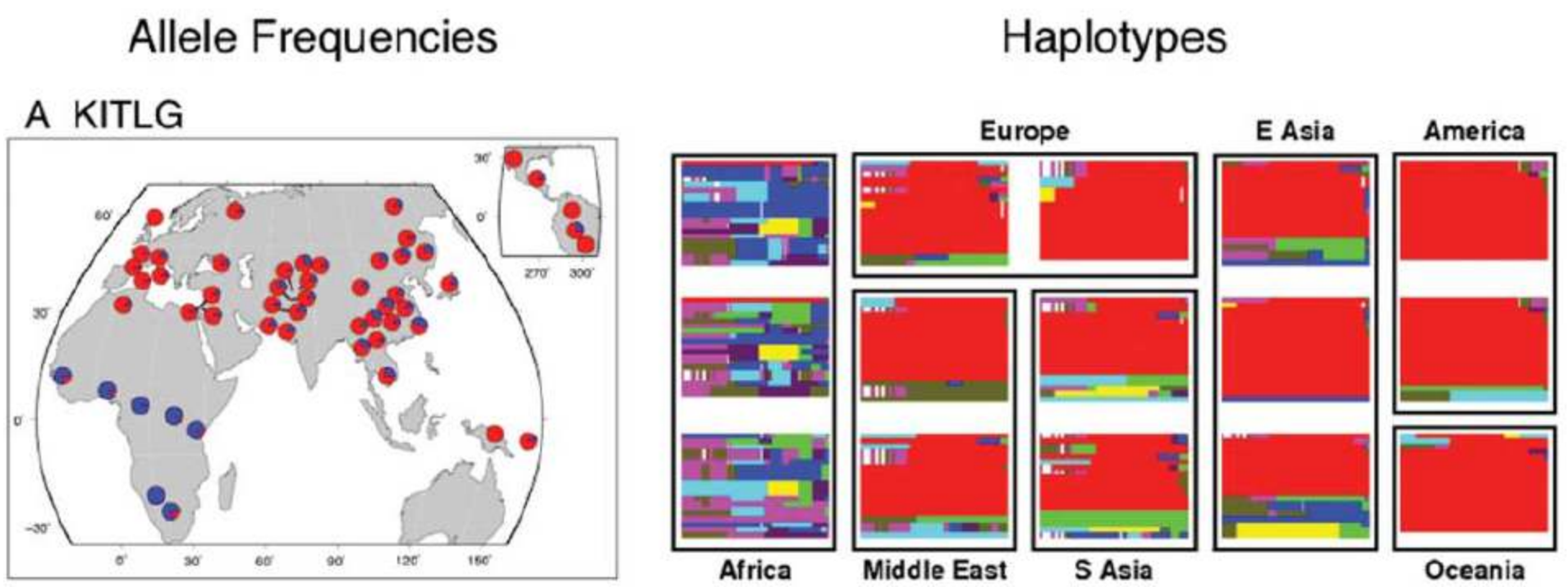
Voight et al. (2006) Plos Biol 4:e72

Adult persistence of Lactase expression



Voight et al. (2006) Plos Biol 4:e72

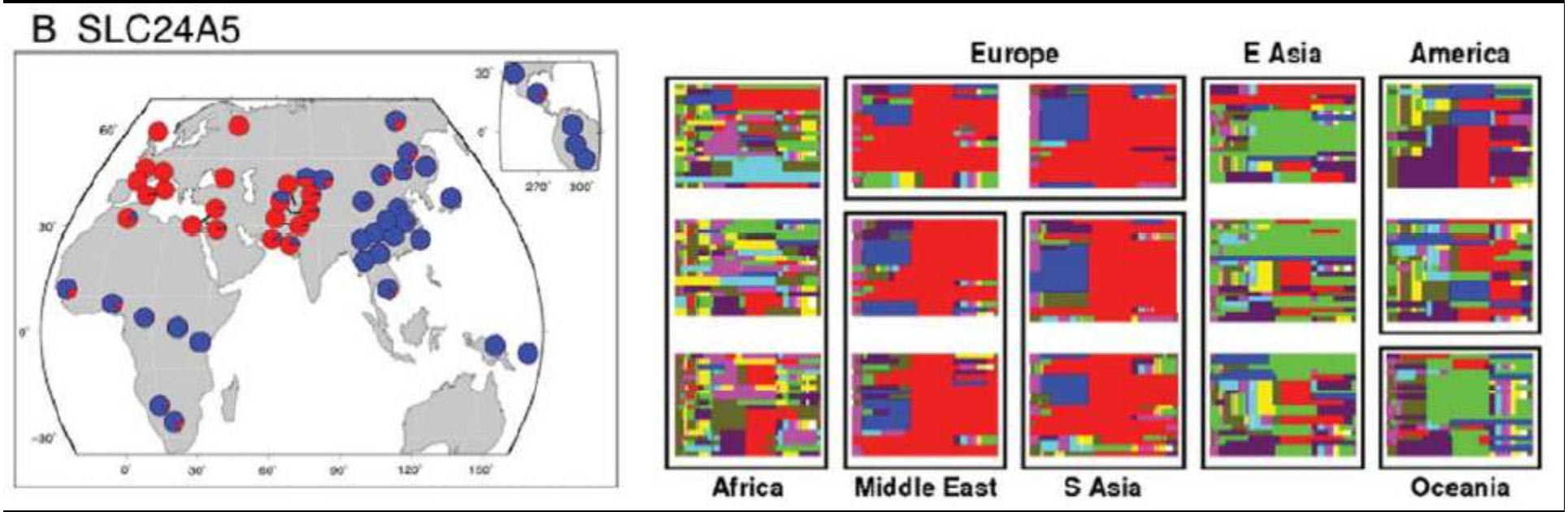
Genes with strong inter-population differentiation and unusually long haplotypes



Derived allele is out-of-Africa

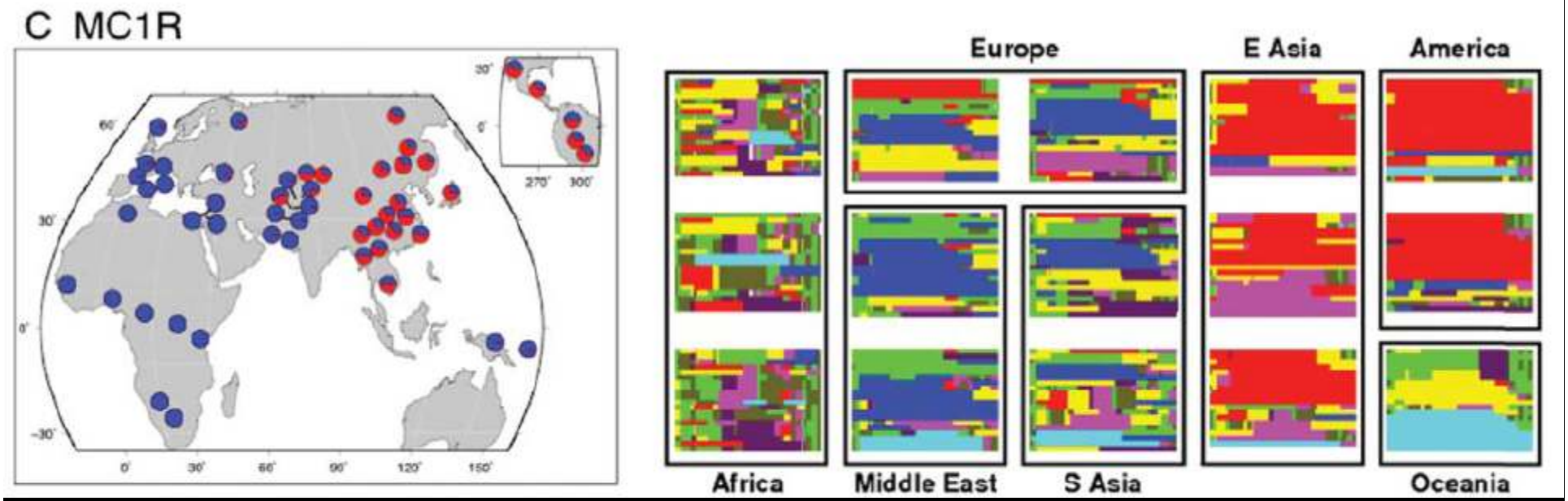
Coop et al. (2009) Plos Genet 5:e1000500

Derived allele is European and central Asian



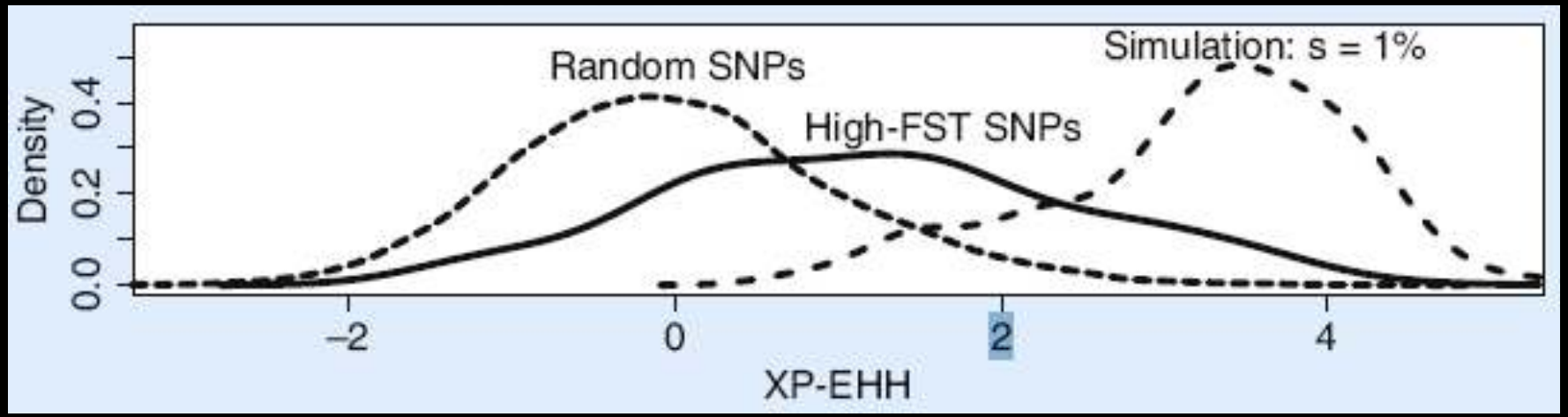
Coop et al. (2009) Plos Genet 5:e1000500

Derived allele is east Asian



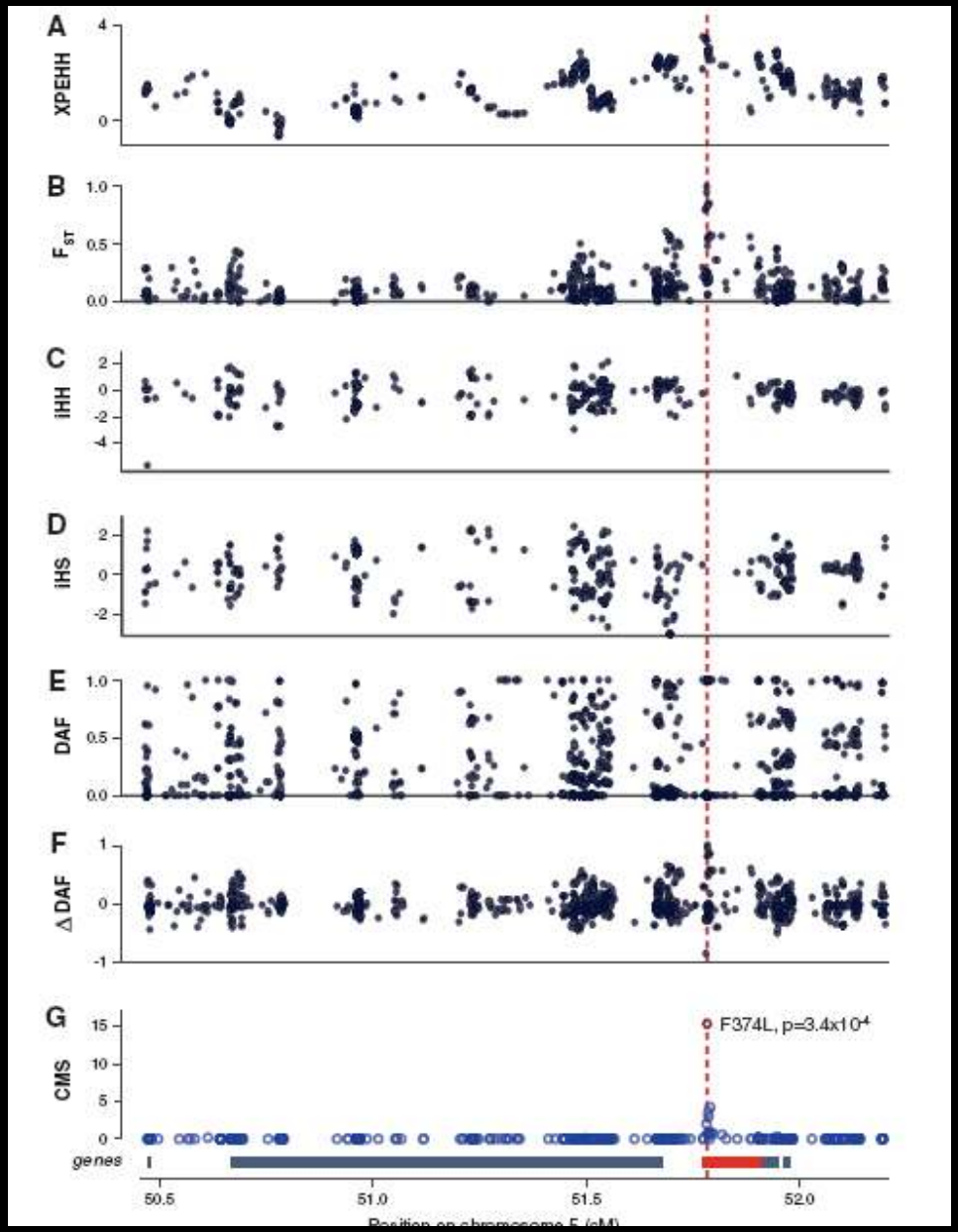
Coop et al. (2009) Plos Genet 5:e1000500

But high FST SNPs are not necessarily on extended haplotypes



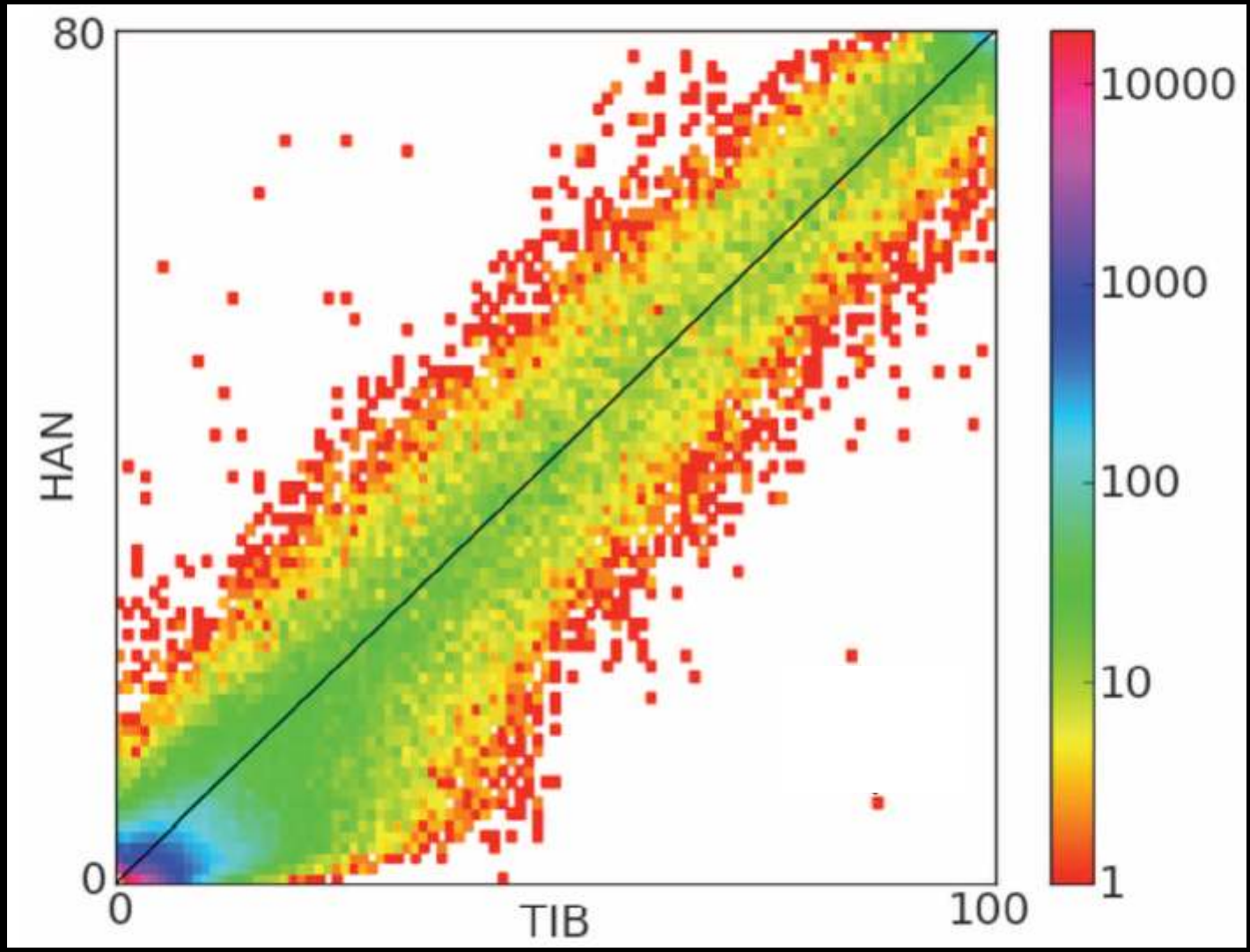
Pritchard et al. (2010) Curr Biol 20:R208–R215

Composite tests likely will be the most powerful



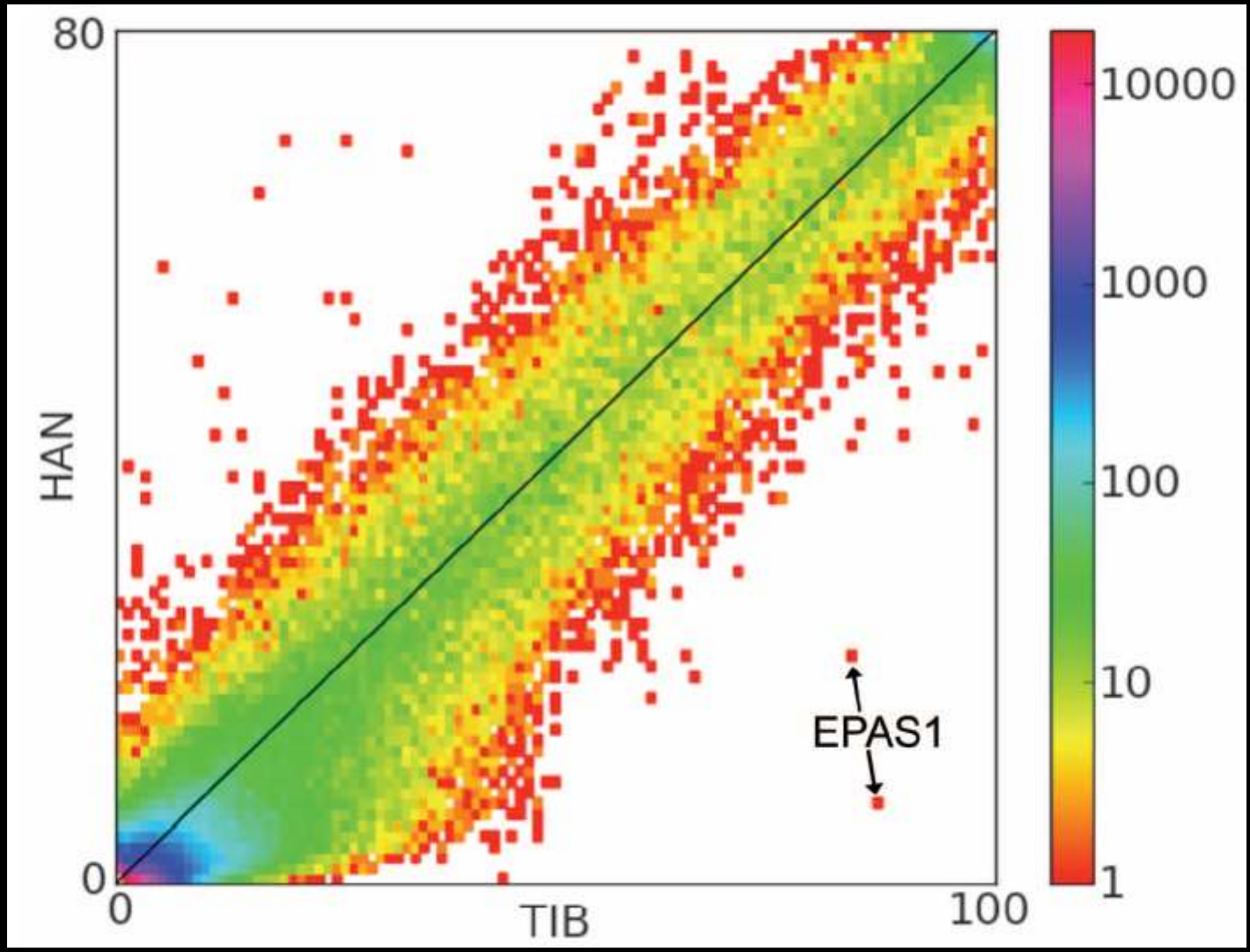
Grossman et al. (2010) Science 327:883-886

Two-dimensional site-frequency spectrum



Yi et al. (2010) Science 329:75-78

Full exome sequencing of 50 Han and 50 Tibetans identifies EPAS1 in altitude acclimation



Yi et al. (2010) Science 329:75-78

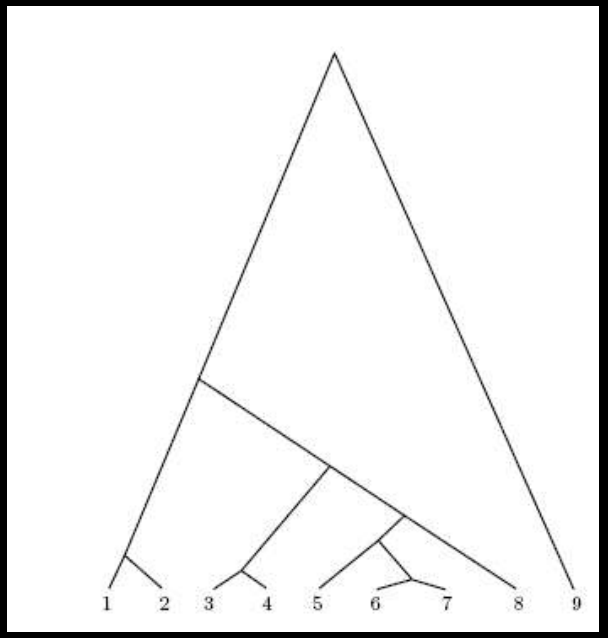
IBD and the neutral coalescent

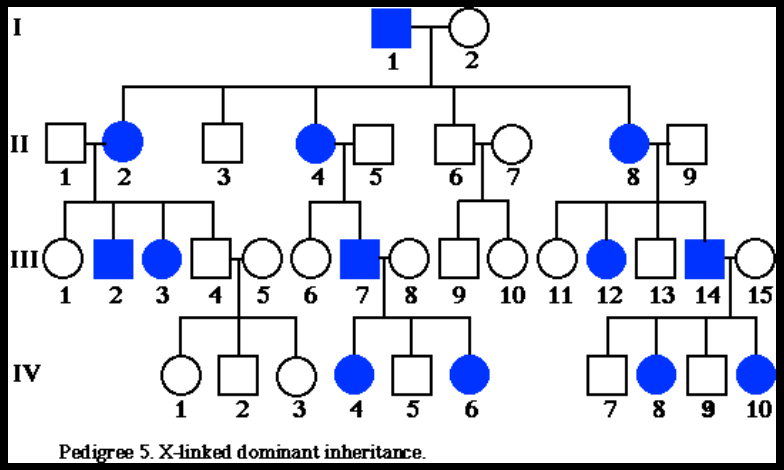
The neutral coalescent assumes the entire sample has common ancestry.

But IBD modeling assumes that at the onset of the studied time (*e.g.* pedigree founders), none of the alleles are IBD.

These two model assumptions would appear to be at odds.

IBD and the neutral coalescent





Pedigree 5. X-linked dominant inheritance.

IBD is for quite recent time scales, whereas the coalescent may span 1 MY or more.

The coalescent produces allelic lineages, and samples from within an allelic lineage need not be IBD.

Textbook recursion for Identity

F is the probability that two alleles drawn from the population are IBD.

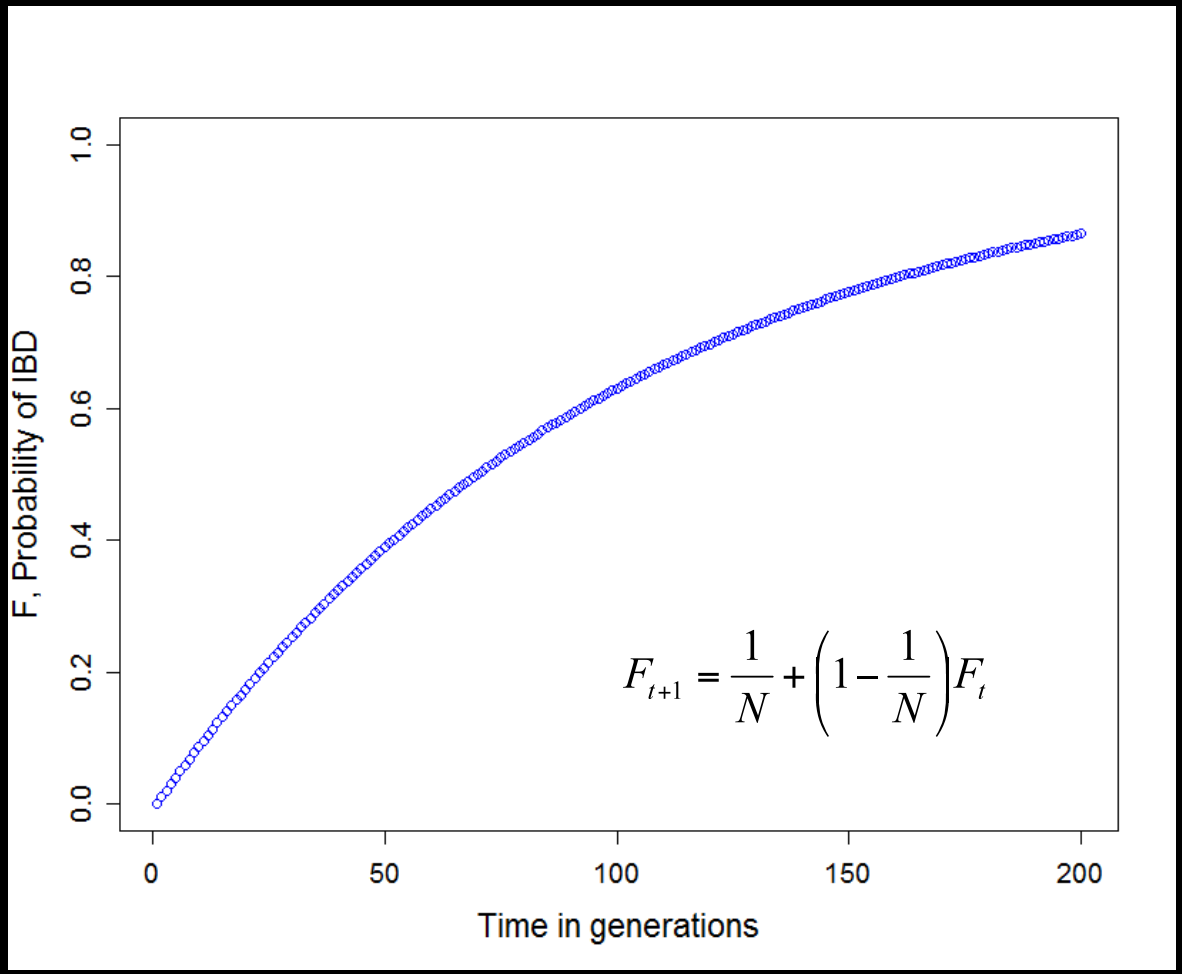
Start with no alleles identical.

If *N* is the number of chromosomes, the chance of drawing the same allele twice (*i.e*. these two copies would be IBD) is 1/*N*.

So the recursion is:

$$F_{t+1} = \frac{1}{N} + \left(1 - \frac{1}{N}\right) F_t$$

IBD increases quickly in a small population (N = 100)



IBD under neutrality

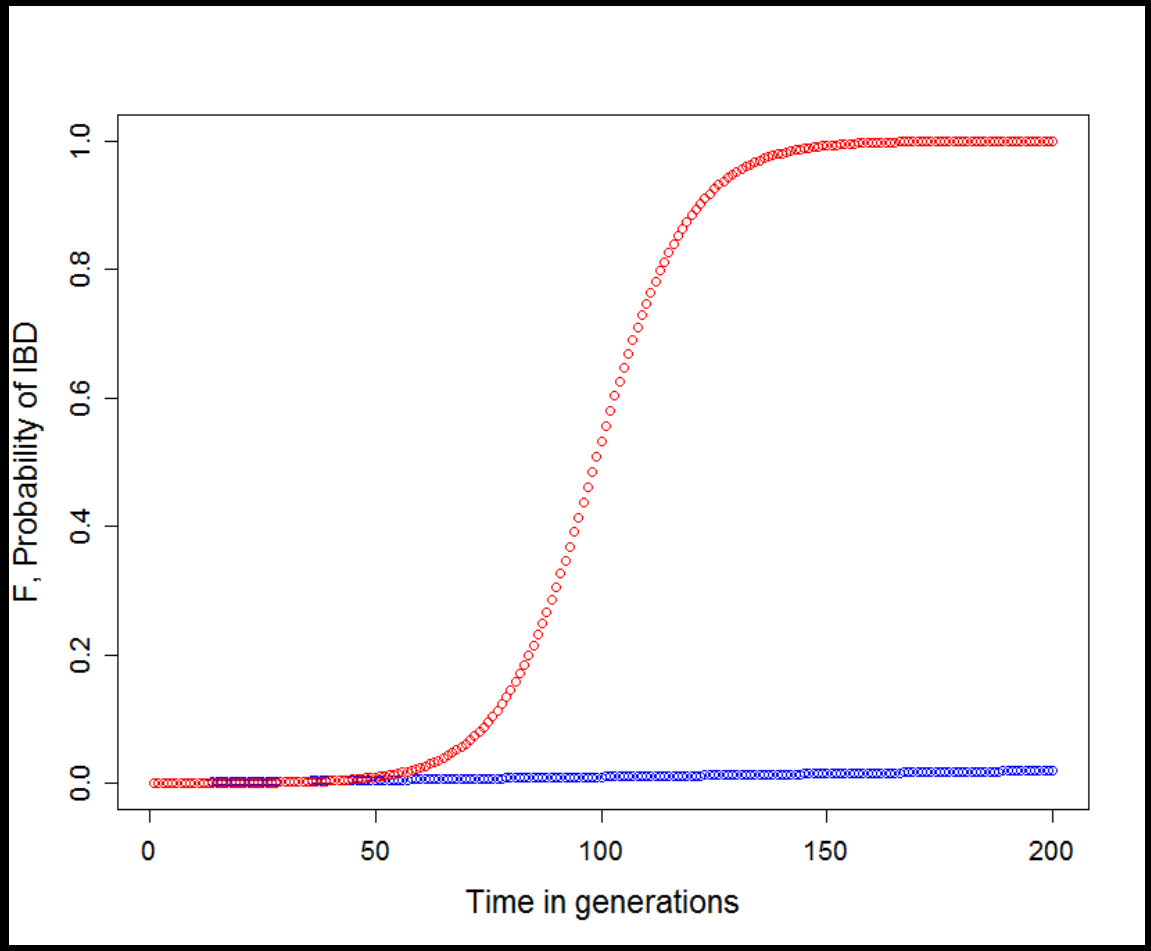
Define F_i as the probability of IBD sharing of allele *i*.

p_i is the frequency of allele A_i.

The overall probability of IBD is then

$$F = \sum_{i=1}^{k} p_i^{\,2} F_i$$

Positive selection increases IBD



selected, s=0.1

neutral (N=10,000)

IBD and natural selection

If allele A_i has fitness w_i (haploid selection) then

$$p_i(t+1) = \frac{w_i p_i(t)}{\overline{w}}$$

where

$$\overline{w} = \sum_{i=1}^{k} w_i p_i(t)$$

IBD and natural selection

Substituting this expression for $p_i(t+1)$ into the expression for $F(t+1)$:

$$F(t+1) = \sum_{i=i}^{k} \left(\frac{w_i p_i(t)}{\overline{w}} \right)^2 \frac{1}{N p_i(t)} = \left(\frac{\overline{w^2}}{\overline{w}^2} \right) \frac{1}{N}$$

IBD and natural selection

Substituting this expression for $p_i(t+1)$ into the expression for $F(t+1)$:

$$F(t+1) = \sum_{i=i}^{k} \left(\frac{w_i p_i(t)}{\overline{w}} \right)^2 \frac{1}{N p_i(t)} = \left(\frac{\overline{w^2}}{\overline{w}^2} \right) \frac{1}{N}$$

So if the alleles differ in fitness, under haploid selection, IBD always increases.

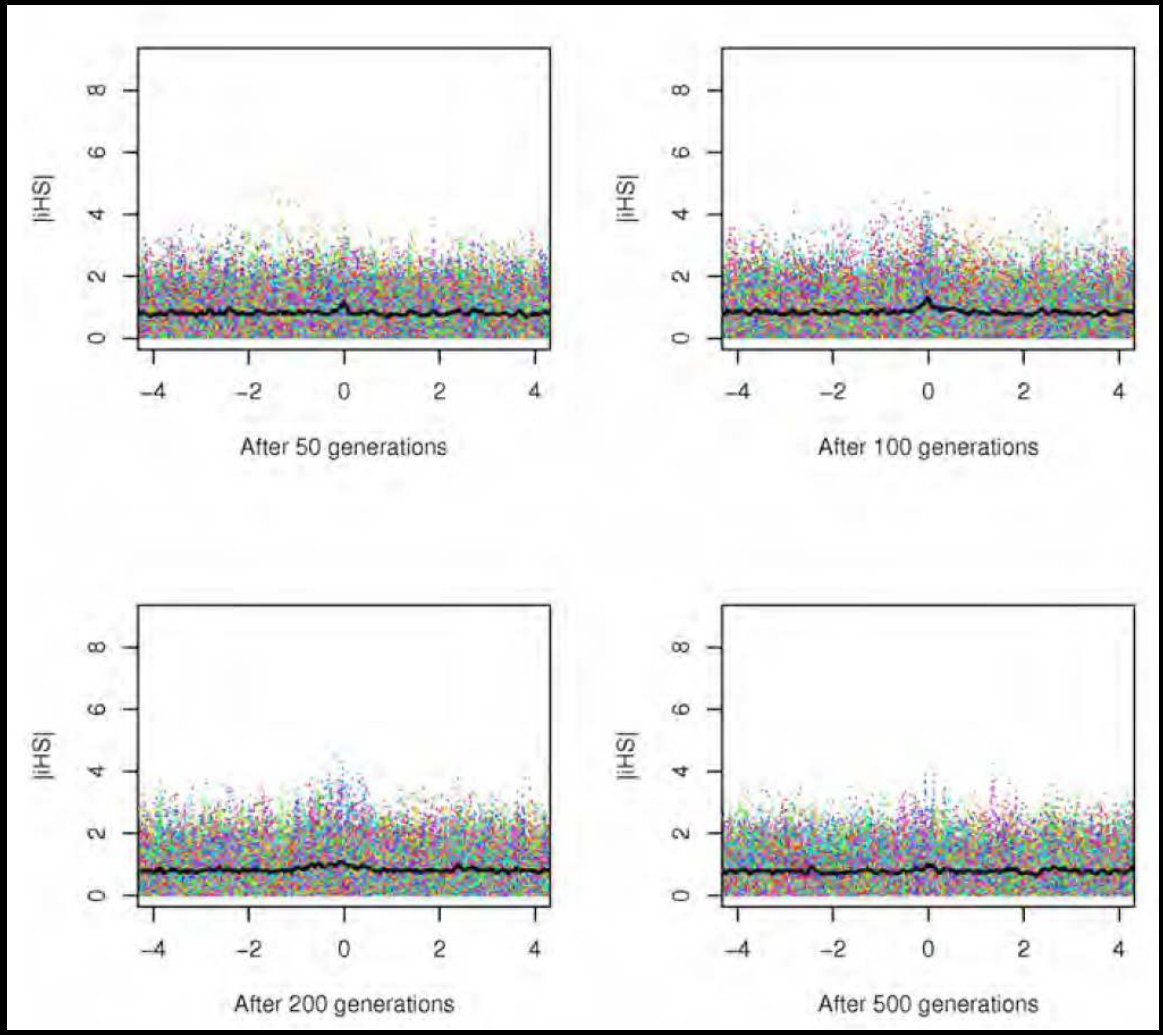
Simulations

1. Used mpop (Pickrell et al. 2009) for forward simulations of 10,000

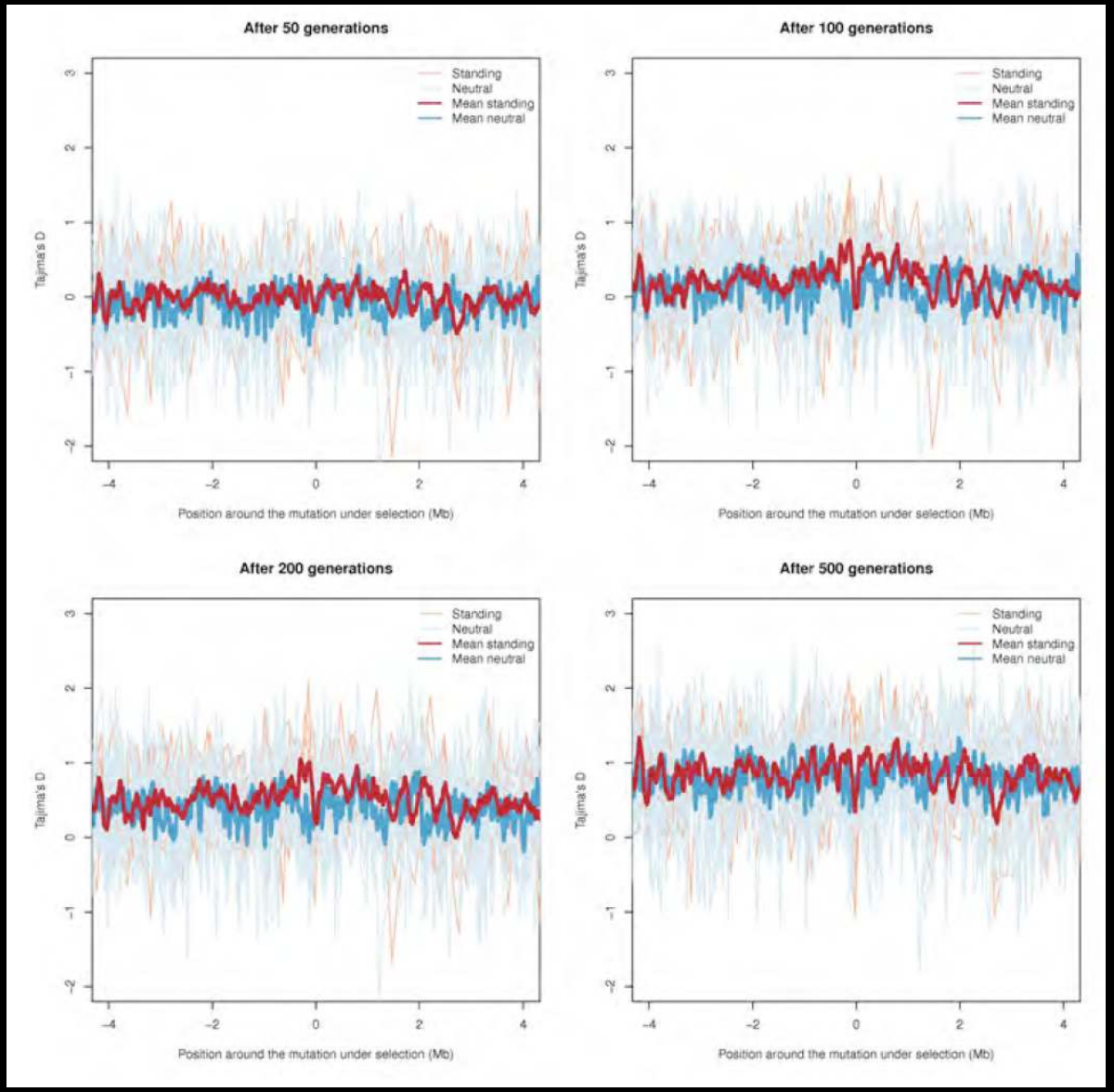
2. For the ’New Allele’ scenario used selective advantage of s=0.1 and s=0.01 to one new allele.

3. For ’Standing Variation’ scenario used selective advantage of s=0.1 and s=0.01 to a previously neutral allele of frequency 0.1.

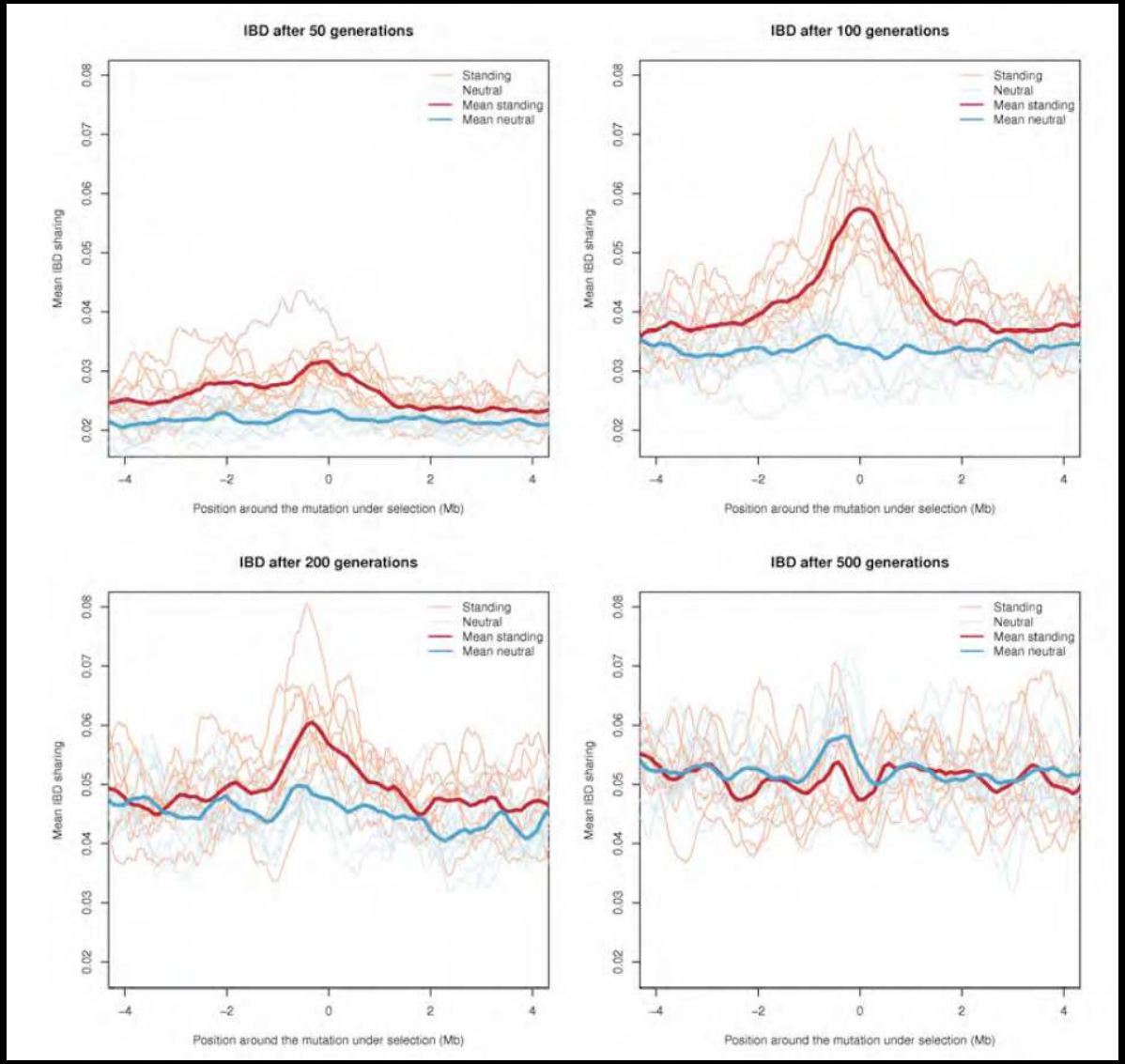
iHS after positive selection (s = 0.1) on standing variation



Tajima's D after positive selection (*s* = 0.1) on standing variation



IBD after positive selection (s = 0.1) on standing variation

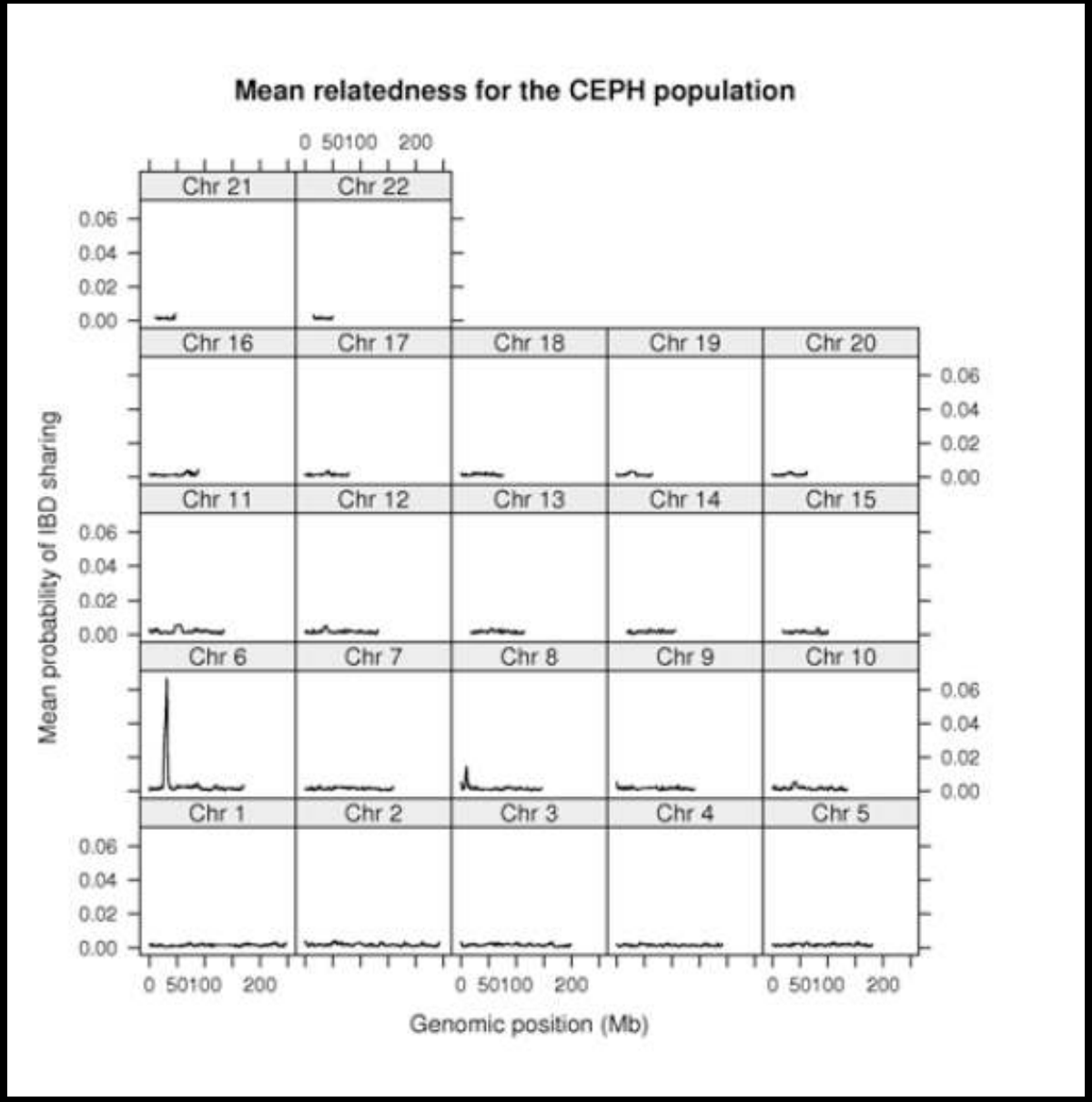


HapMap3 data

	Population	#individuals	#Unrelated	#SNP*
ASW	African ancestry in Southwest USA	90	42	120030
CEU	Northern and Western European ancestry (CEPH)	180	109	105291
CHB	Han Chinese in Beijing, China	90	82	88533
CHD	Chinese in Metropolitan Denver, Colorado	100	70	85901
GIH	Gujarati Indians in Houston, Texas	100	83	106768
JPT	Japanese in Tokyo, Japan	91	82	84557
LWK	Luhya in Webuye, Kenya	100	83	193280
MEX	Mexican ancestry in Los Angeles, California	90	45	94474
MKK	Maasai in Kinyawa, Kenya	180	143	196540
TSI	Toscans in Italy	100	77	104540
YRI	Yoruba in Ibadan, Nigeria	180	108	193373

Table S1: Table of the HapMap phase 3 individuals used in this study. The #SNP* column contains the number of SNPs left after data cleaning and LD removal.

IBD across HapMap CEPH samples



"Hits" of significant excess in IBD

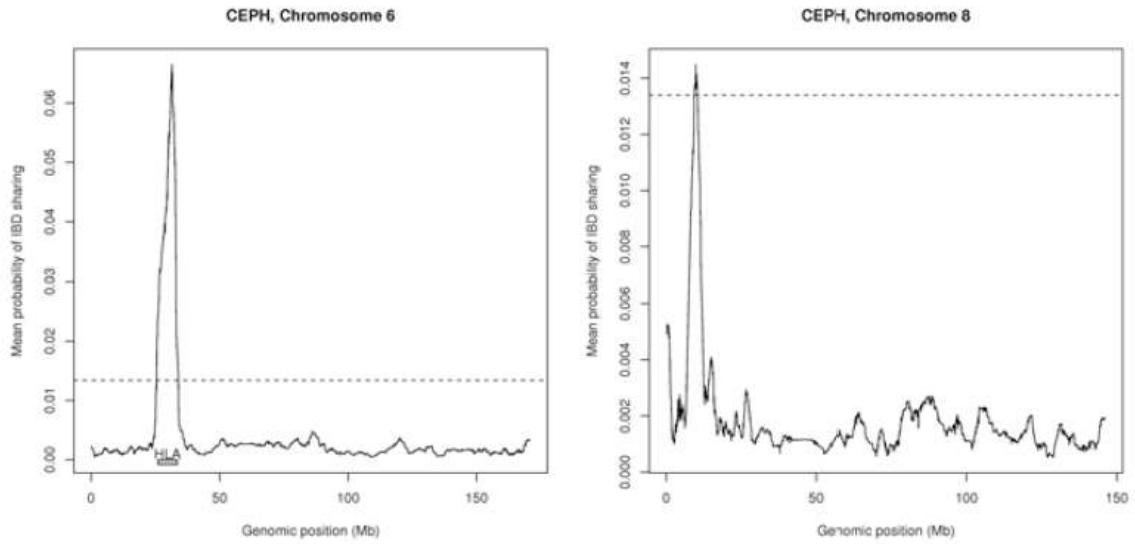


Figure 6: IBD sharing for chromosome 6 and 8 in the CEPH population. Mean probability of IBD sharing among the 109 unrelated individuals in the HapMap phase 3 CEPH sample is shown for chromosome 6 and chromosome 8. The horizontal dashed lines indicate the critical value achieved through coalescent simulations.

The chromosome 6 region spans HLA; chromosome 8 region has defensin genes.

IBD sharing in all samples

Chromosome	position in Mb	populations
1	56.0	CHD*
2	50.9-51.2	GIH*, YRI, CHD
2	52.55	LWK*
3	84.9-85.3	JPT*, CHB
5	153.2-153.8	MEX*, CEPH
6	29.3-31.5, (w.o. MKK 31.2-31.5)	All populations
8	8.4-10.5	CEPH, CHB, CHD, GIH, JPT, MEX, TSI
11	49.3-51.1	CEPH*, CHD*, JPT, MEX
11	55.2-56.7	TSI*, MKK*, YRI*
14	77.7	ASW*

Table 1: Table of the regions corresponding to the top signal for each population. A * after a population name indicates that the region in question contains the highest peak of that population (excluding the HLA regions and the region on chromosome 8 that contains a high peak in CEPH). Included are also the populations that on the same chromosome have their highest peak in close proximity (at most 1Mb away).

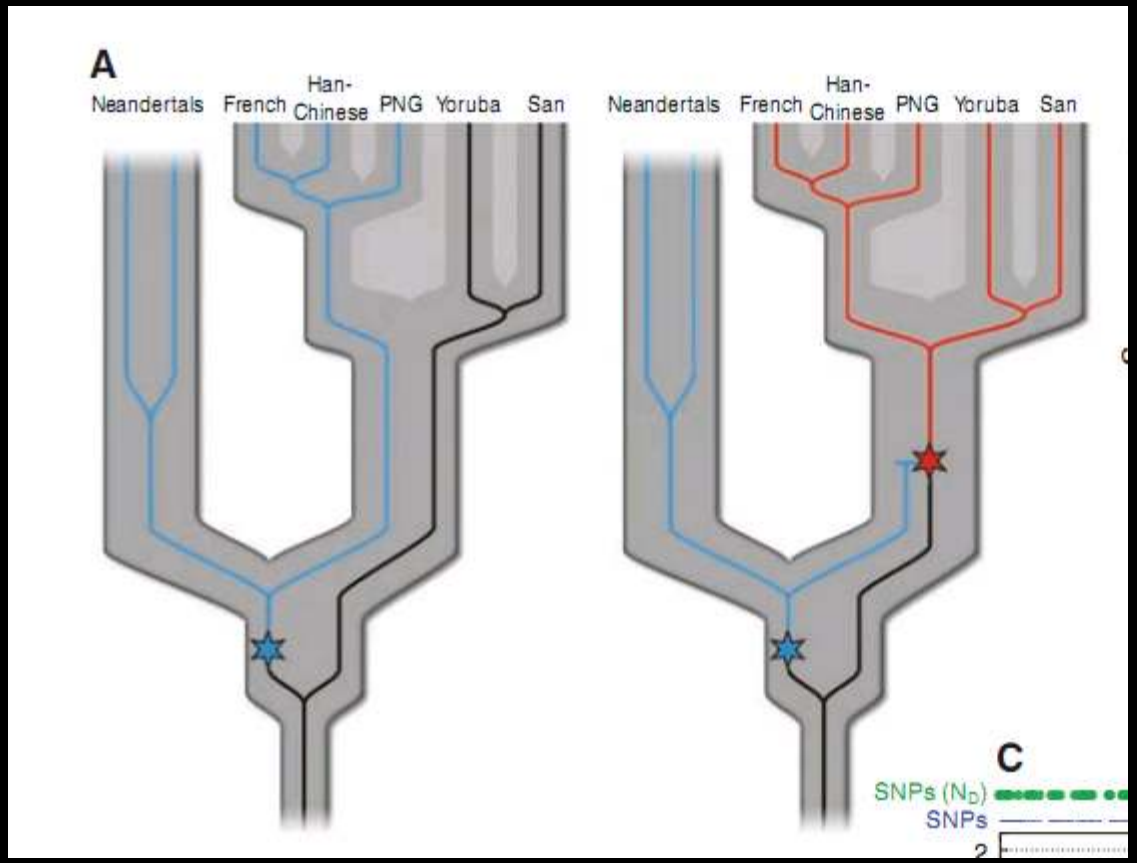
A Draft Sequence of the Neandertal Genome

Richard E. Green,[1*†‡] Johannes Krause,[1†§] Adrian W. Briggs,[1†§] Tomislav Maricic,[1†§] Udo Stenzel,[1†§] Martin Kircher,[1†§] Nick Patterson,[2†§] Heng Li,[2†] Weiwei Zhai,[3†||] Markus Hsi-Yang Fritz,[4†] Nancy F. Hansen,[5†] Eric Y. Durand,[3†] Anna-Sapfo Malaspinas,[3†] Jeffrey D. Jensen,[6†] Tomas Marques-Bonet,[7,13†] Can Alkan,[7†] Kay Prüfer,[1†] Matthias Meyer,[1†] Hernán A. Burbano,[1†] Jeffrey M. Good,[1,8†] Rigo Schultz,[1] Ayinuer Aximu-Petri,[1] Anne Butthof,[1] Barbara Höber,[1] Barbara Höffner,[1] Madlen Siegemund,[1] Antje Weihmann,[1] Chad Nusbaum,[2] Eric S. Lander,[2] Carsten Russ,[2] Nathaniel Novod,[2] Jason Affourtit,[9] Michael Egholm,[9] Christine Verna,[21] Pavao Rudan,[10] Dejana Brajkovic,[11] Željko Kucan,[10] Ivan Gušic,[10] Vladimir B. Doronichev,[12] Liubov V. Golovanova,[12] Carles Lalueza-Fox,[13] Marco de la Rasilla,[14] Javier Fortea,[14¶] Antonio Rosas,[15] Ralf W. Schmitz,[16,17] Philip L. F. Johnson,[18†] Evan E. Eichler,[7†] Daniel Falush,[19†] Ewan Birney,[4†] James C. Mullikin,[5†] Montgomery Slatkin,[3†] Rasmus Nielsen,[3†] Janet Kelso,[1†] Michael Lachmann,[1†] David Reich,[2,20*†] Svante Pääbo[1*†]

Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with

Green et al. 2010 Science 328:710-722

Identifying selection sweeps that occurred since the split with Neanderthal



Left: Neanderthal has derived allele in modern human
Right: Sweep occurred in modern human only.

Genes with excess human-specific fixed differences

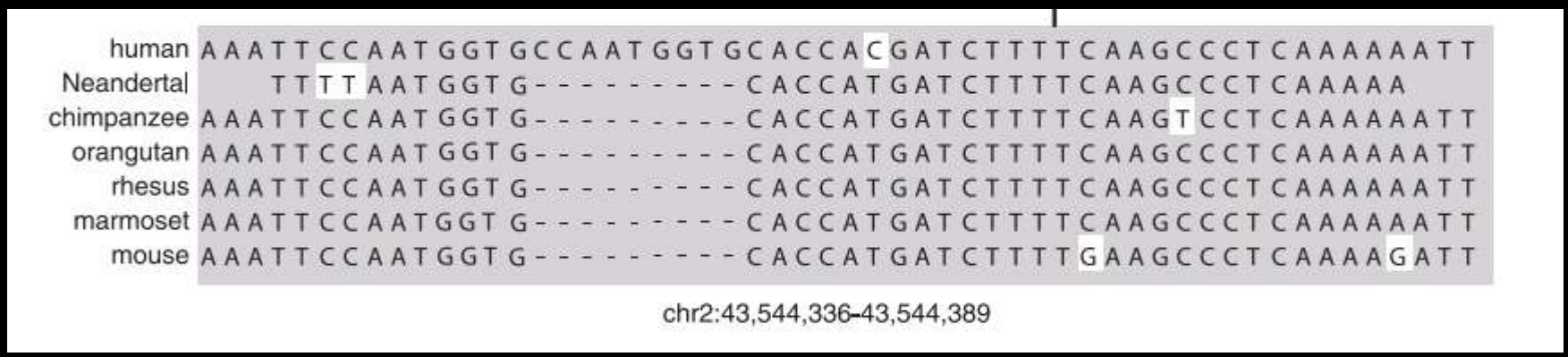


Table 3. Top 20 candidate selective sweep regions.

Region (hg18)	*S*	Width (cM)	Gene(s)
chr2:43265008-43601389	-6.04	0.5726	*ZFP36L2;THADA*
chr11:95533088-95867597	-4.78	0.5538	*JRKL;CCDC82;MAML2*
chr10:62343313-62655667	-6.1	0.5167	*RHOBTB1*
chr21:37580123-37789088	-4.5	0.4977	*DYRK1A*
chr10:83336607-83714543	-6.13	0.4654	*NRG3*
chr14:100248177-100417724	-4.84	0.4533	*MIR337;MIR665;DLK1;RTL1;MIR431;MIR493;MEG3;MIR770*
chr3:157244328-157597592	-6	0.425	*KCNAB1*
chr11:30601000-30992792	-5.29	0.3951	
chr2:176635412-176978762	-5.86	0.3481	*HOXD11;HOXD8;EVX2;MTX2;HOXD1;HOXD10;HOXD13; HOXD4;HOXD12;HOXD9;MIR10B;HOXD3*

Conclusions about Natural Selection

- Evidence for natural selection in humans is rampant.
- Signatures of selection arise from patterns of polymorphism and divergence.
- Various tests probe different times in the past when selection could have acted. Extended haplotype tests are for very recent events.

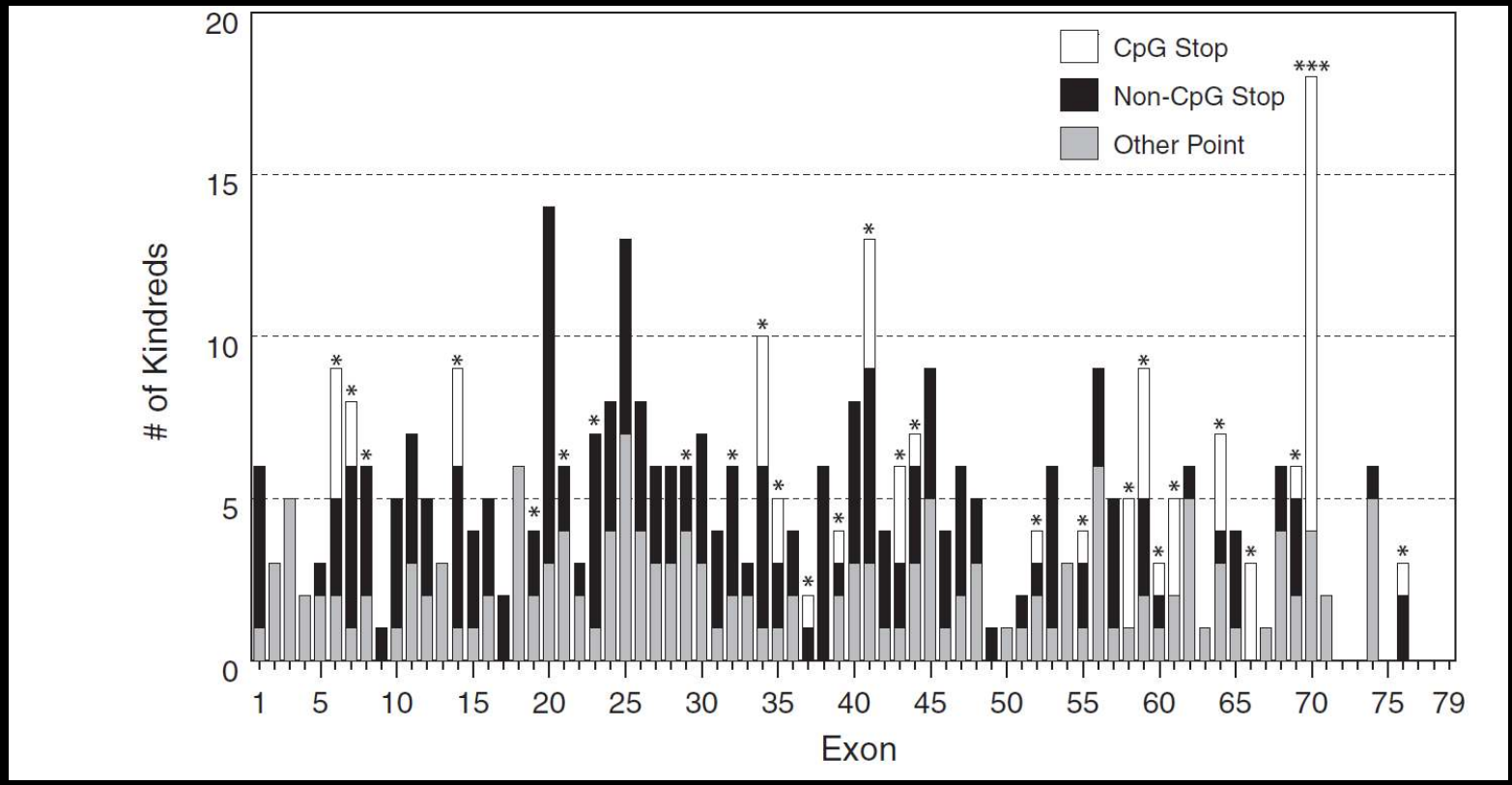
- Positive selection is especially common in genes involved in defense, perception, and reproduction.
- Genes that have a signature of recent selection generally display larger allele frequency differences between populations.
- Power of tests is improved with additional primate genome sequences.

What can be inferred about the population's past demographic history from genetic variation?

- Can we infer past demographic changes?
- Levels of admixture?
- Can we estimate times of common ancestry from genetic data?

What would we learn about rare variation from ultra-deep resequencing?

There are >3000 mutations in Dystrophin



Flanagan et al. 2009 *Human Mutation*

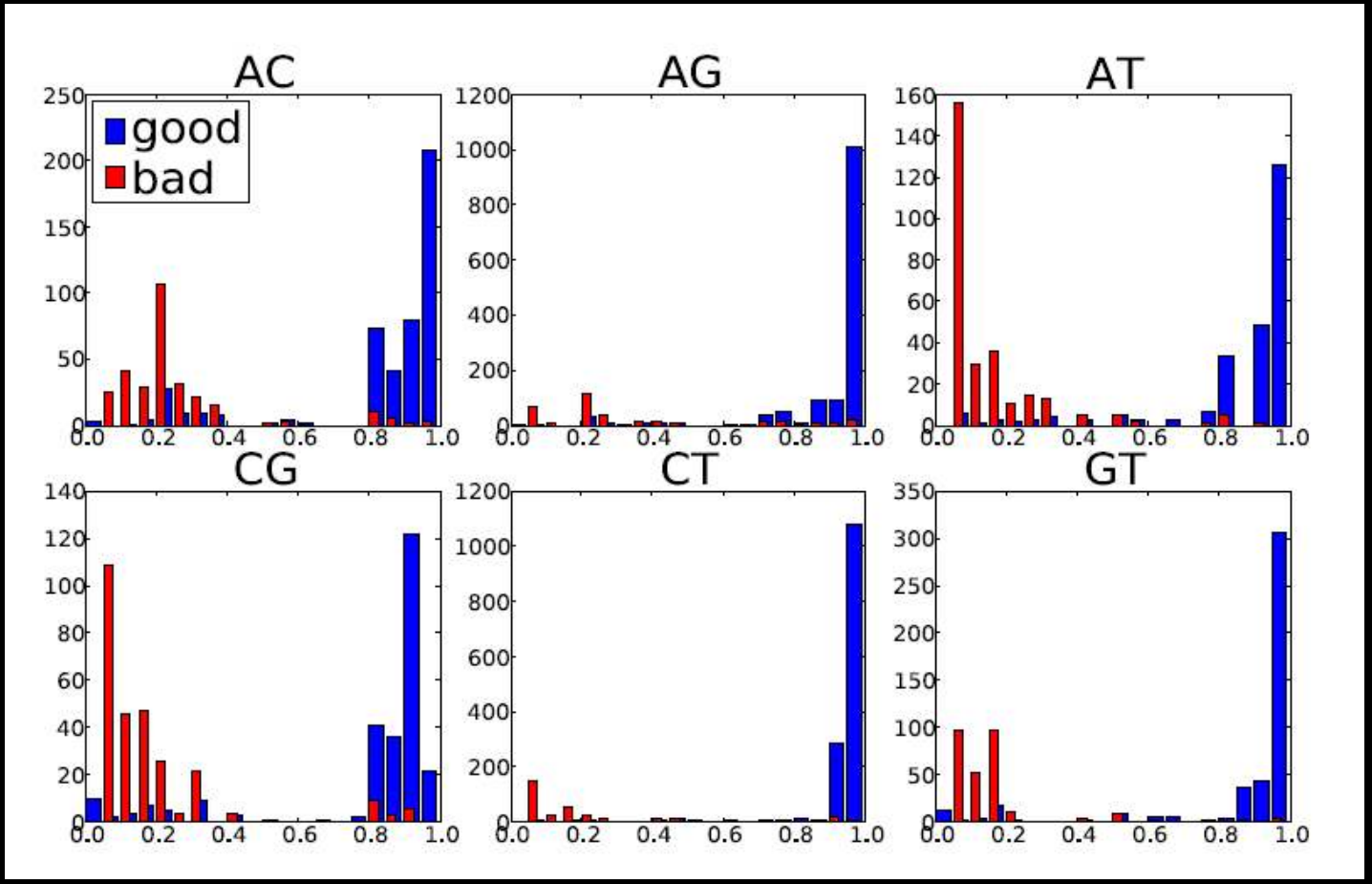
Atherosclerosis Risk in Communities Study
sponsored by the National Heart, Lung and Blood Institute (NHLBI)

- 13,715 people
- Four U.S. cities
- Longitudinal medical records for 20 years.
- BCM performed PCR + Sanger sequencing of *HHEX* and *KCNJ11* in all.

Dealing with sequencing error

- 13.4 kb x 13,715 individuals = 140 Mbp.
- If the error rate per base call is one in 100,000, there would be 1400 false positive SNP calls.
- Standard SNP calling methods are not up to the task.

Developed a discriminator with useful estimates of error rate



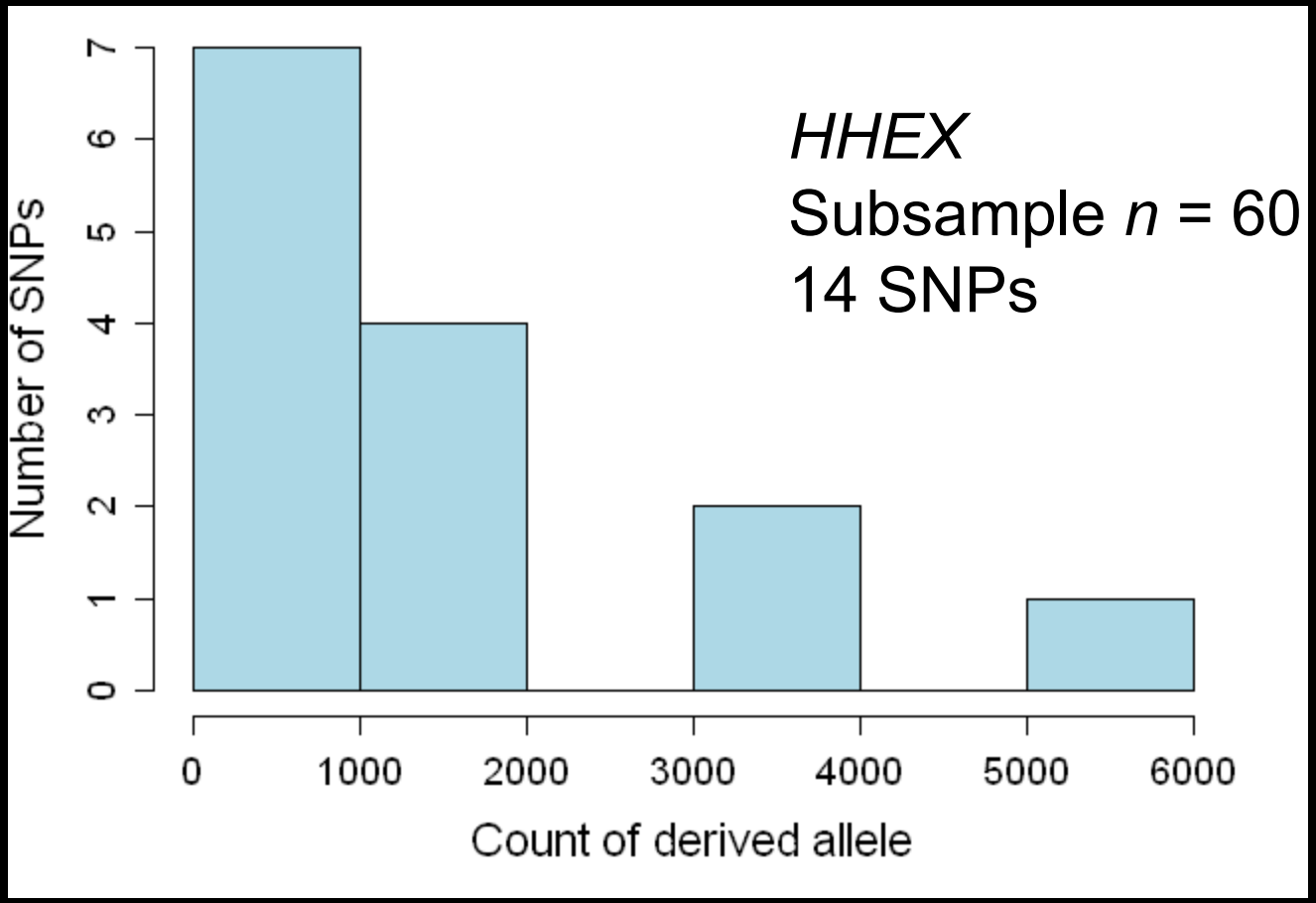
Alex Coventry

Sequencing results

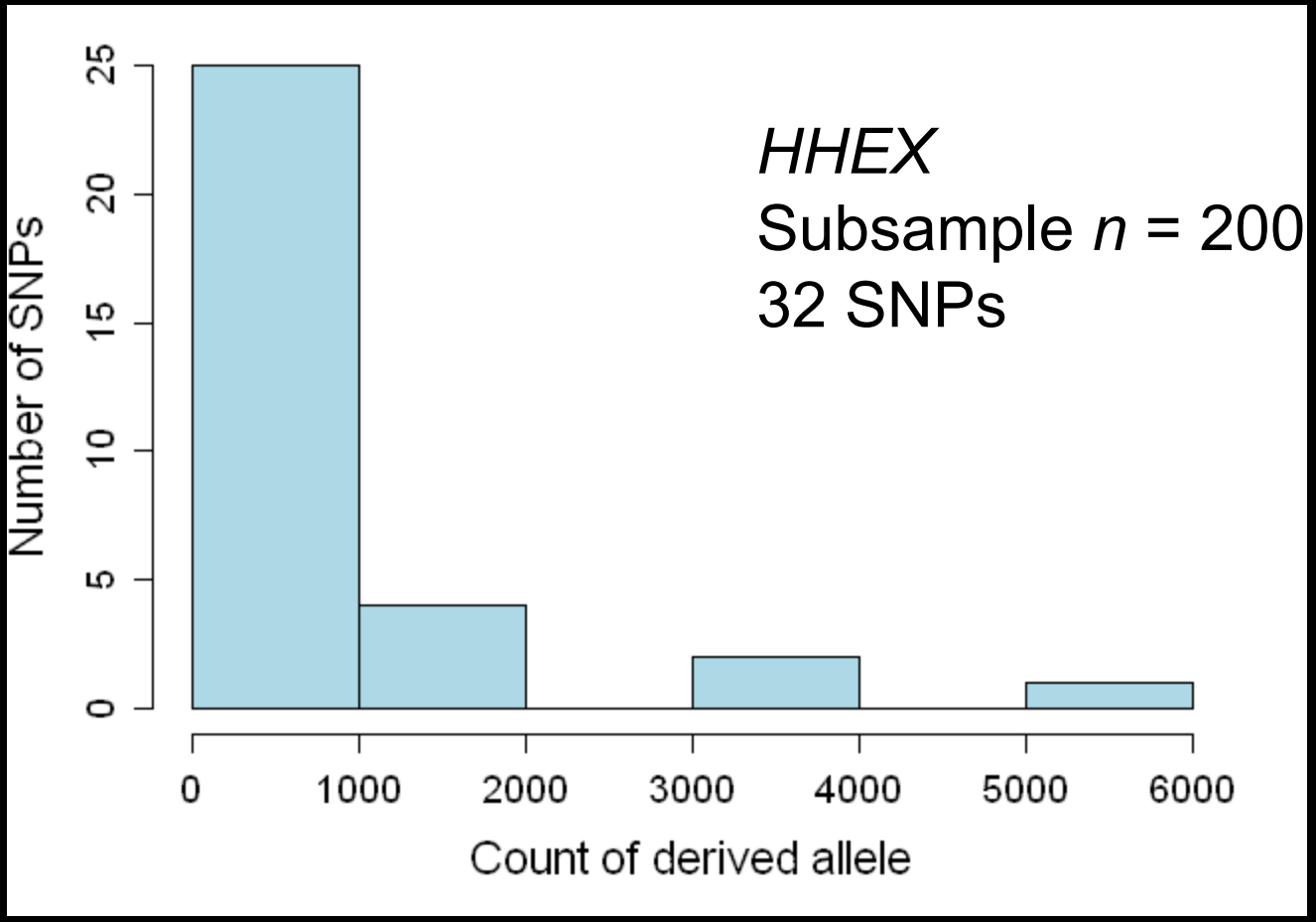
- 52 PCR amplicons in *HHEX* and *KCNJ11* were sequenced in 13,715 people.
- More than 578 SNPs and 52 indels were detected (156 inferred to be “damaging”).
- For validation, amplicons with the called SNPs were re-sequenced by 454 (at the Baylor Genome Center).

Coventry et al, 2010 *Nature Communications*

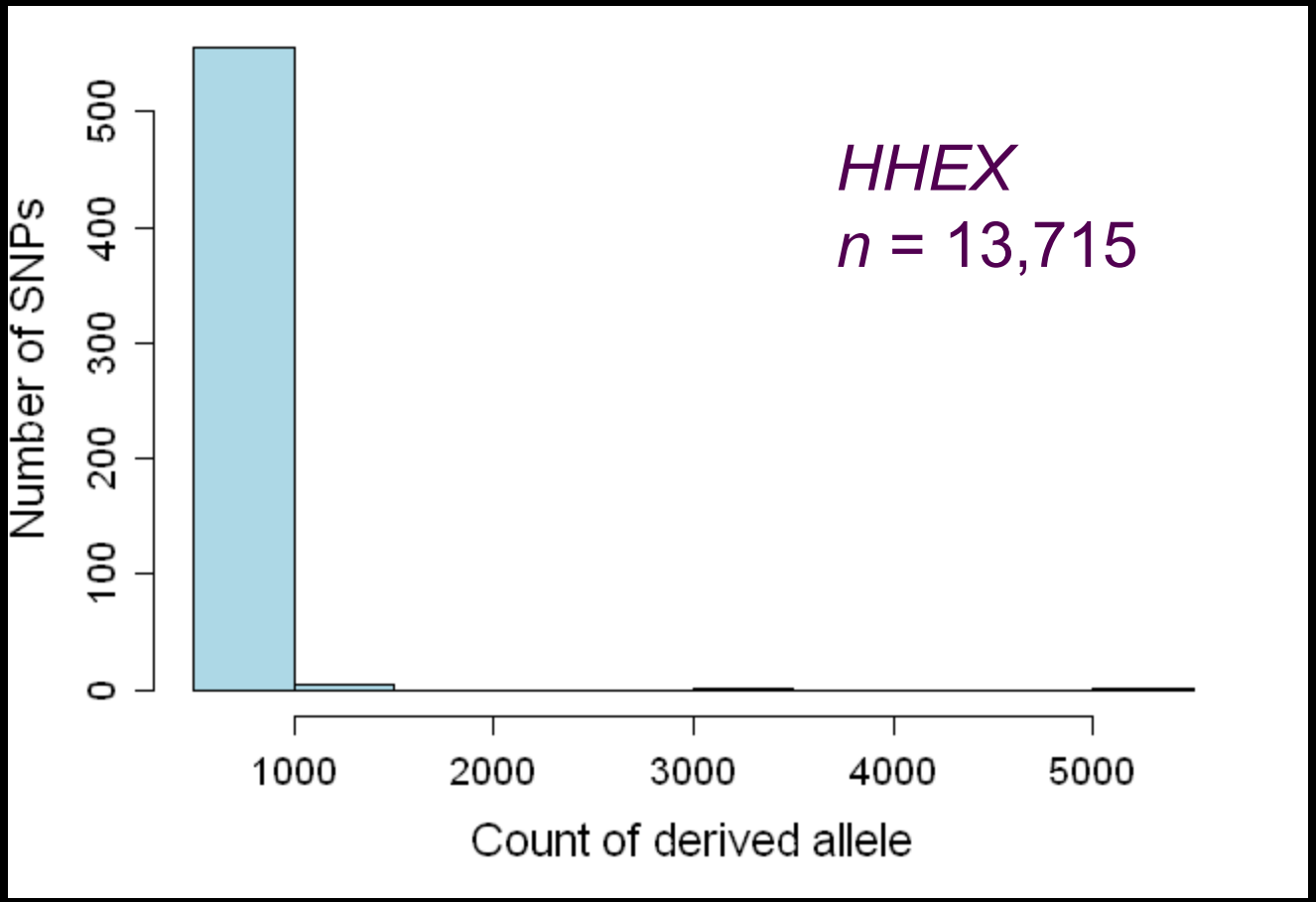
Site frequency spectrum from a small subsample



Site frequency spectrum from a slightly larger sample

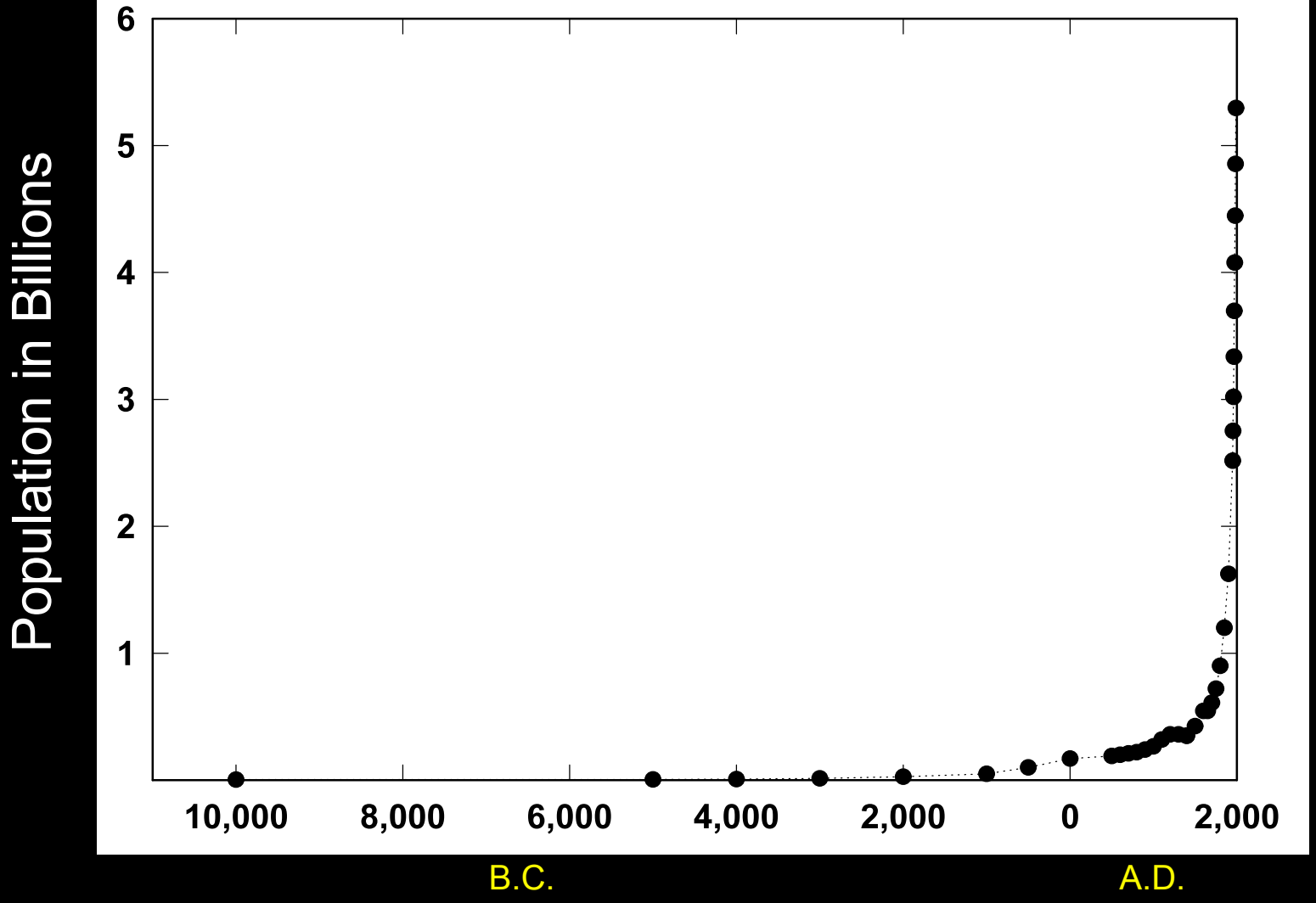


The full sample reveals a vast inflation of rare variants



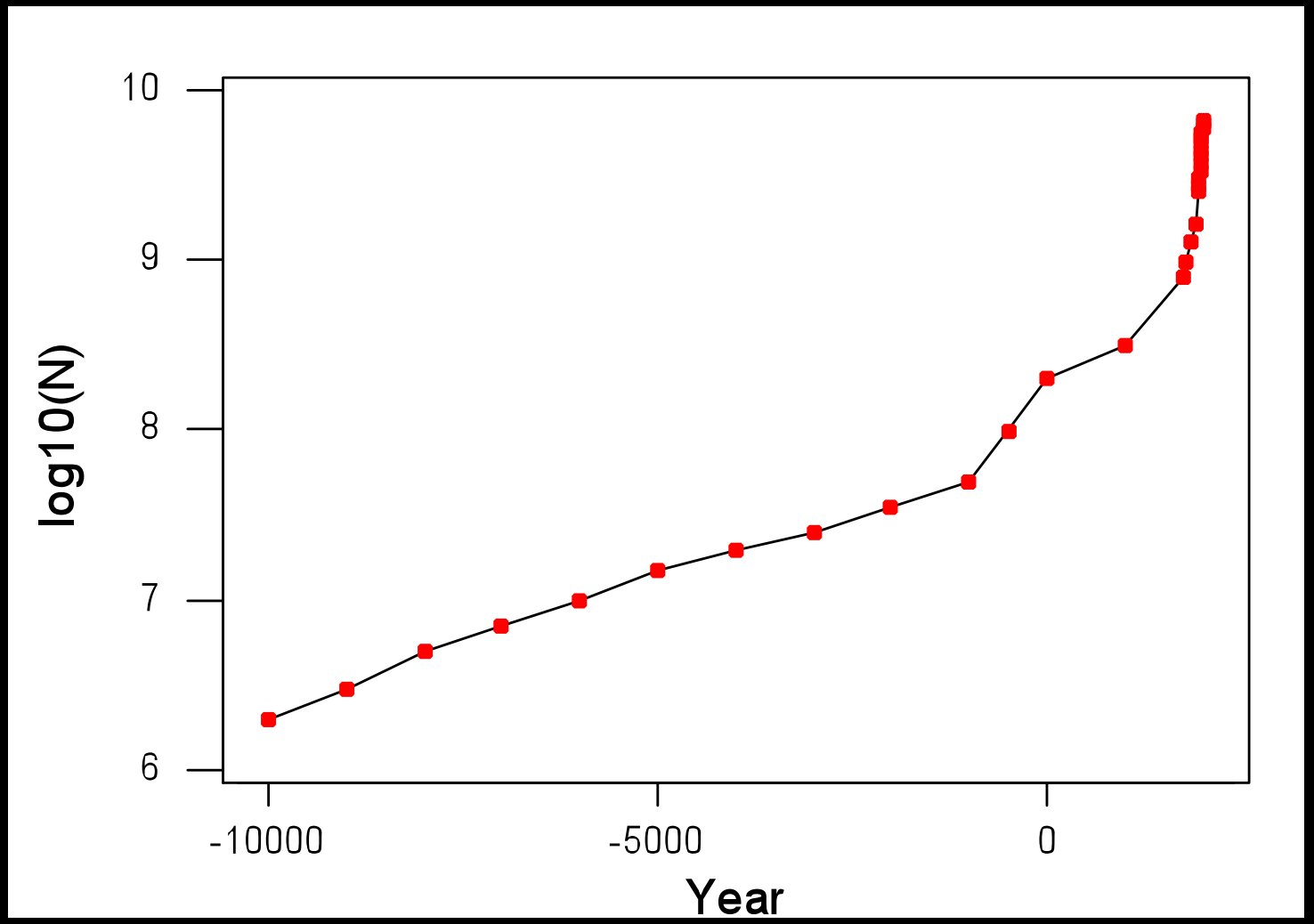
Why are there so many rare variants in humans?

World Population Size, 10,000 B.C. to Present



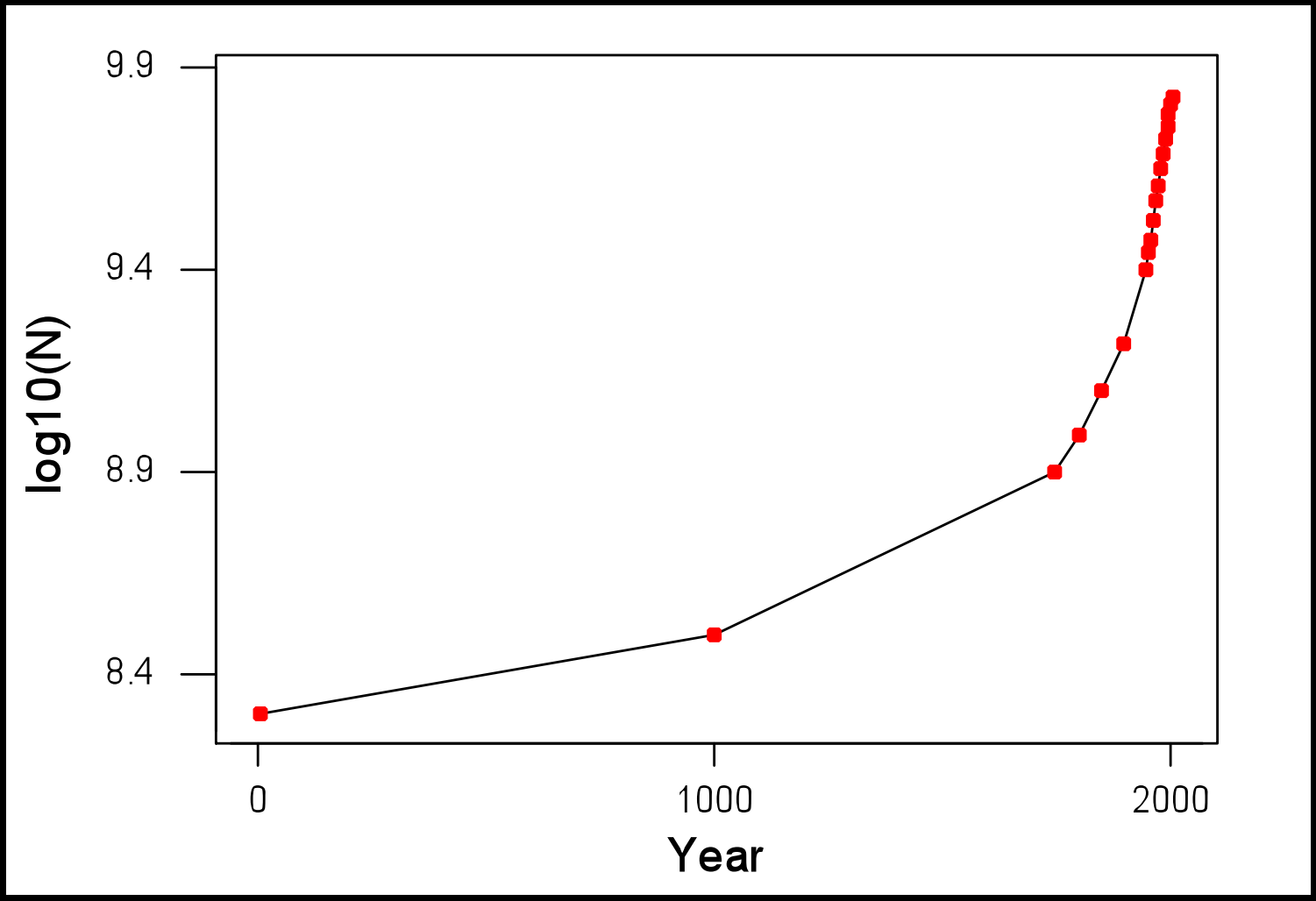
Cohen J 1995 *How Many People can the Earth Support?*

Super-exponential growth of human population



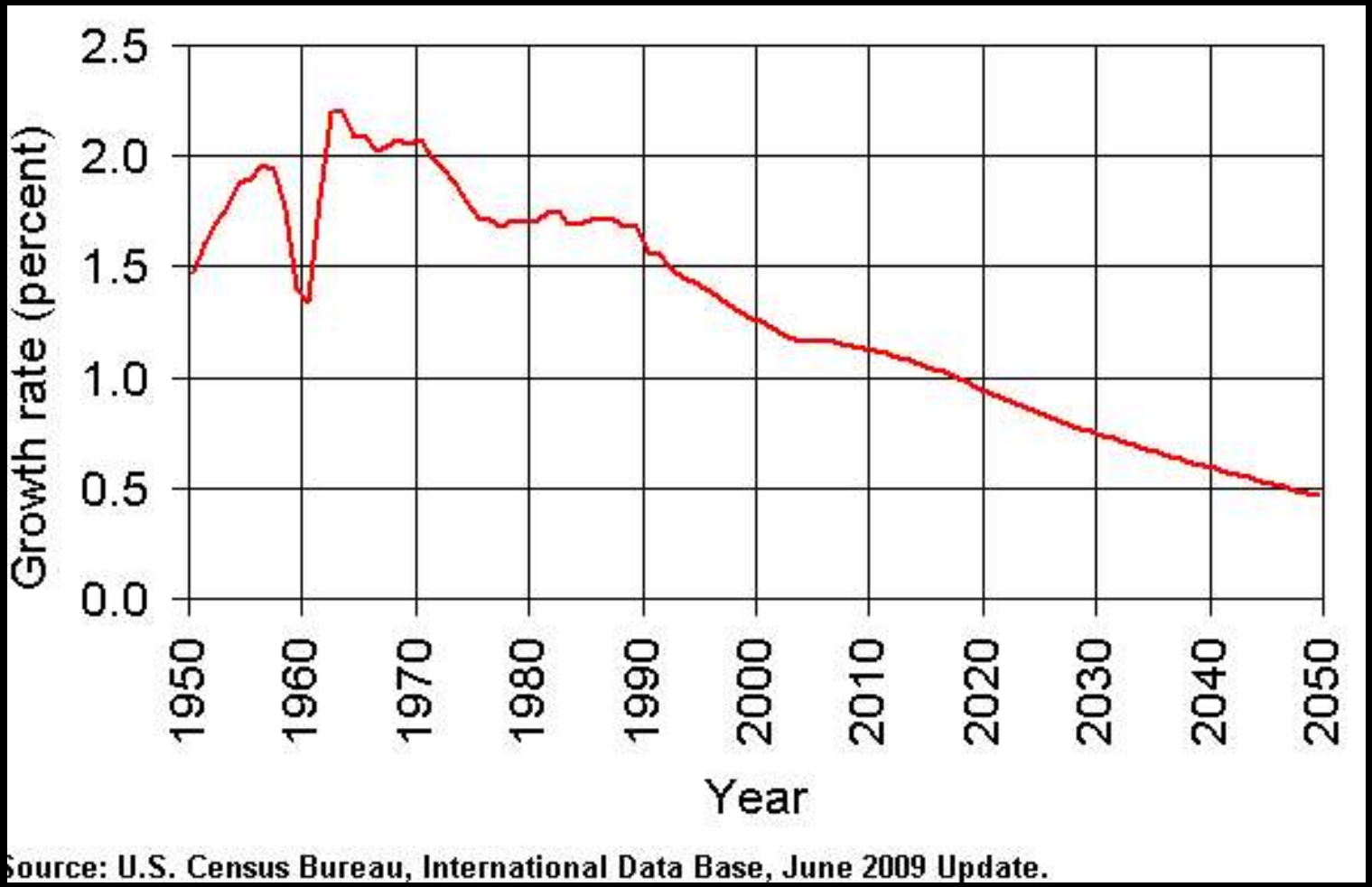
Cohen J 1995 *How Many People can the Earth Support?*

Recent super-exponential growth



Cohen J 1995 *How Many People can the Earth Support?*

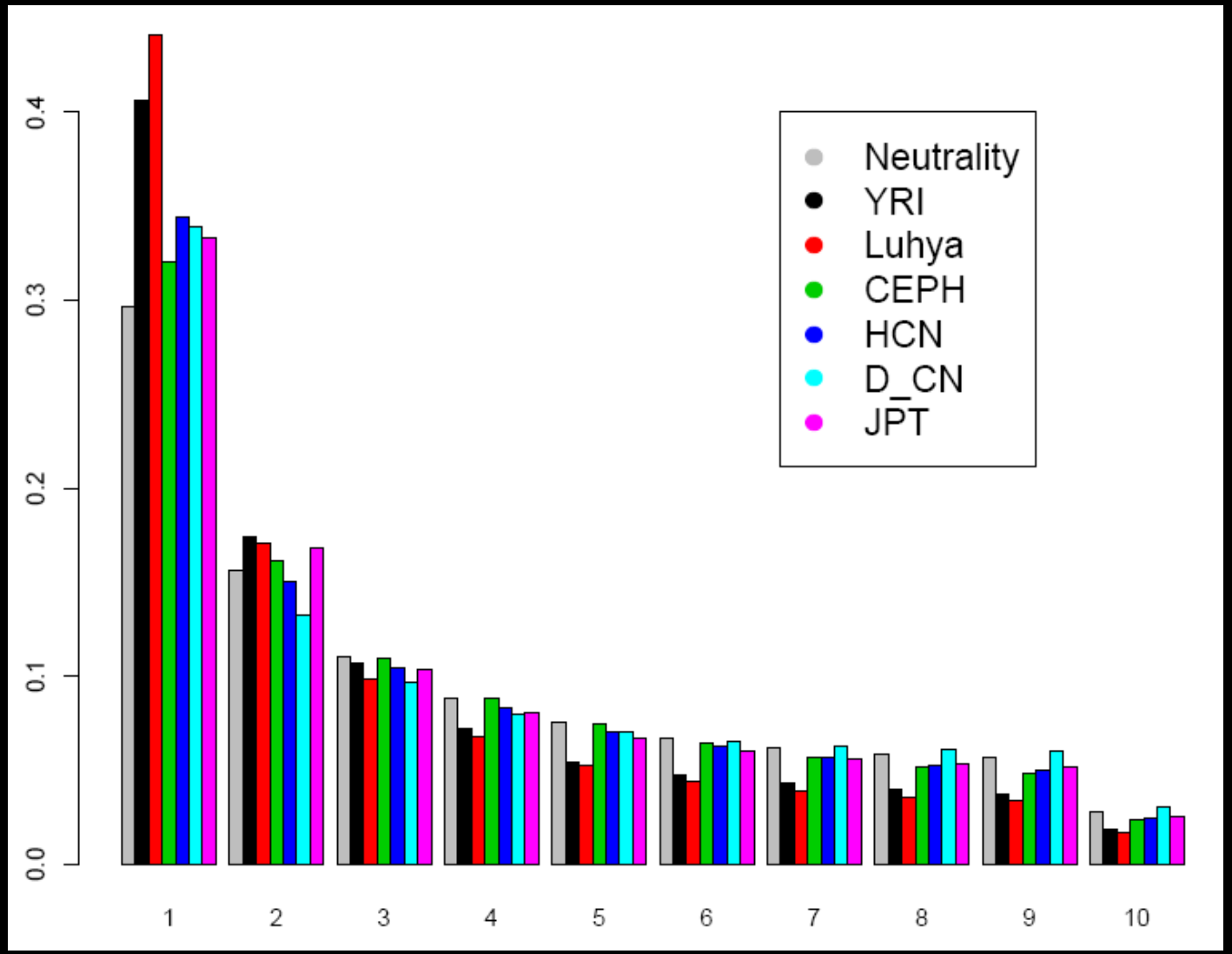
World Population Growth Rate



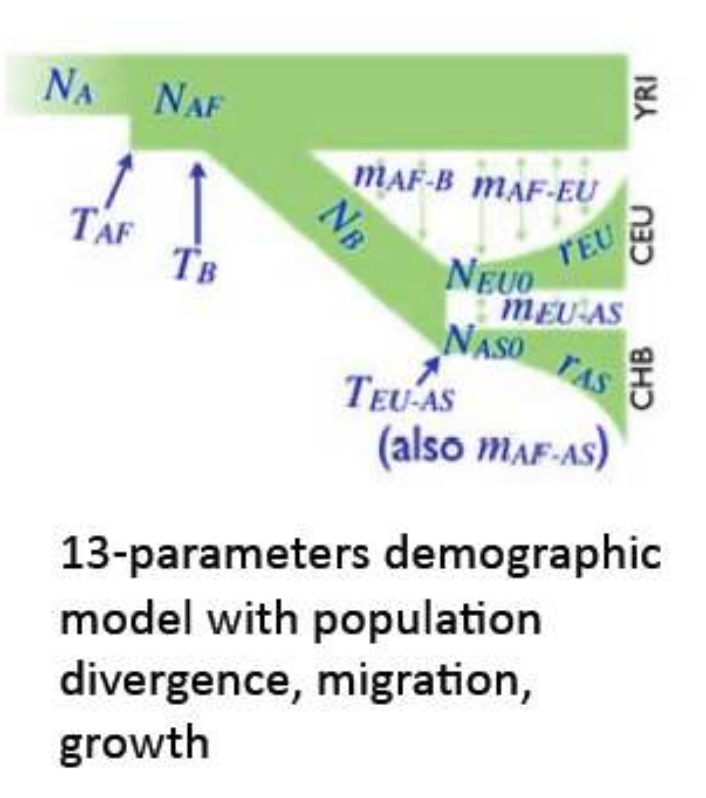
Source: U.S. Census Bureau, International Data Base, June 2009 Update.

An expanding population distorts the genealogy, giving an excess of rare variants.

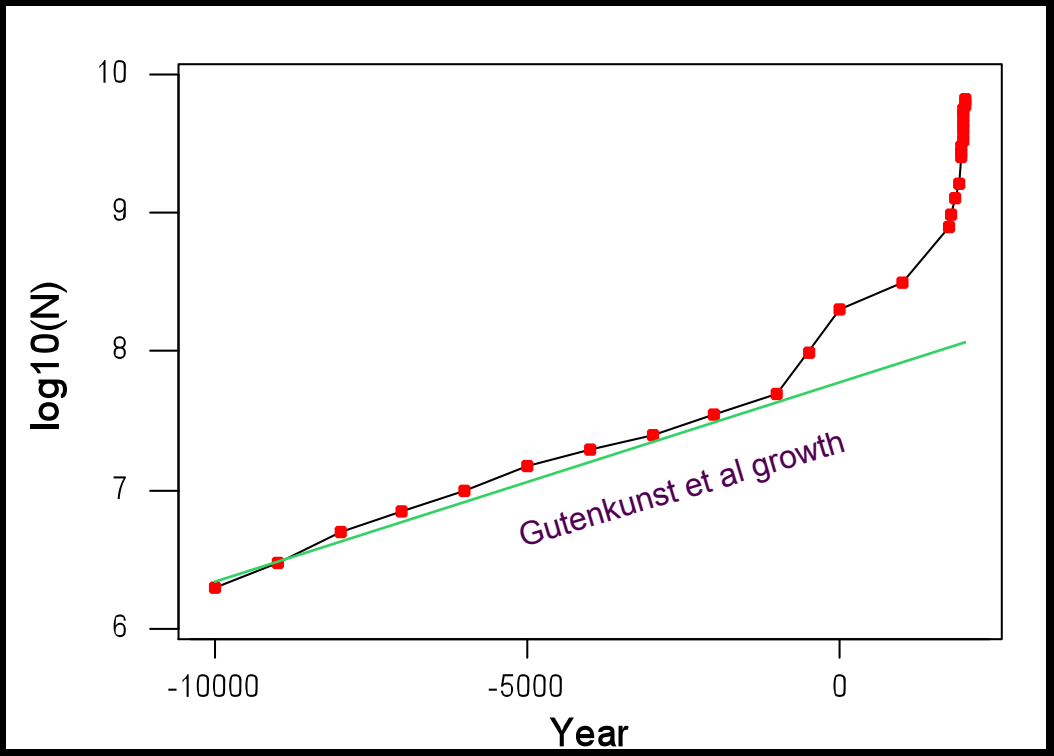
Observed site frequency spectra



Human demographic models: fitted to older, common SNPs

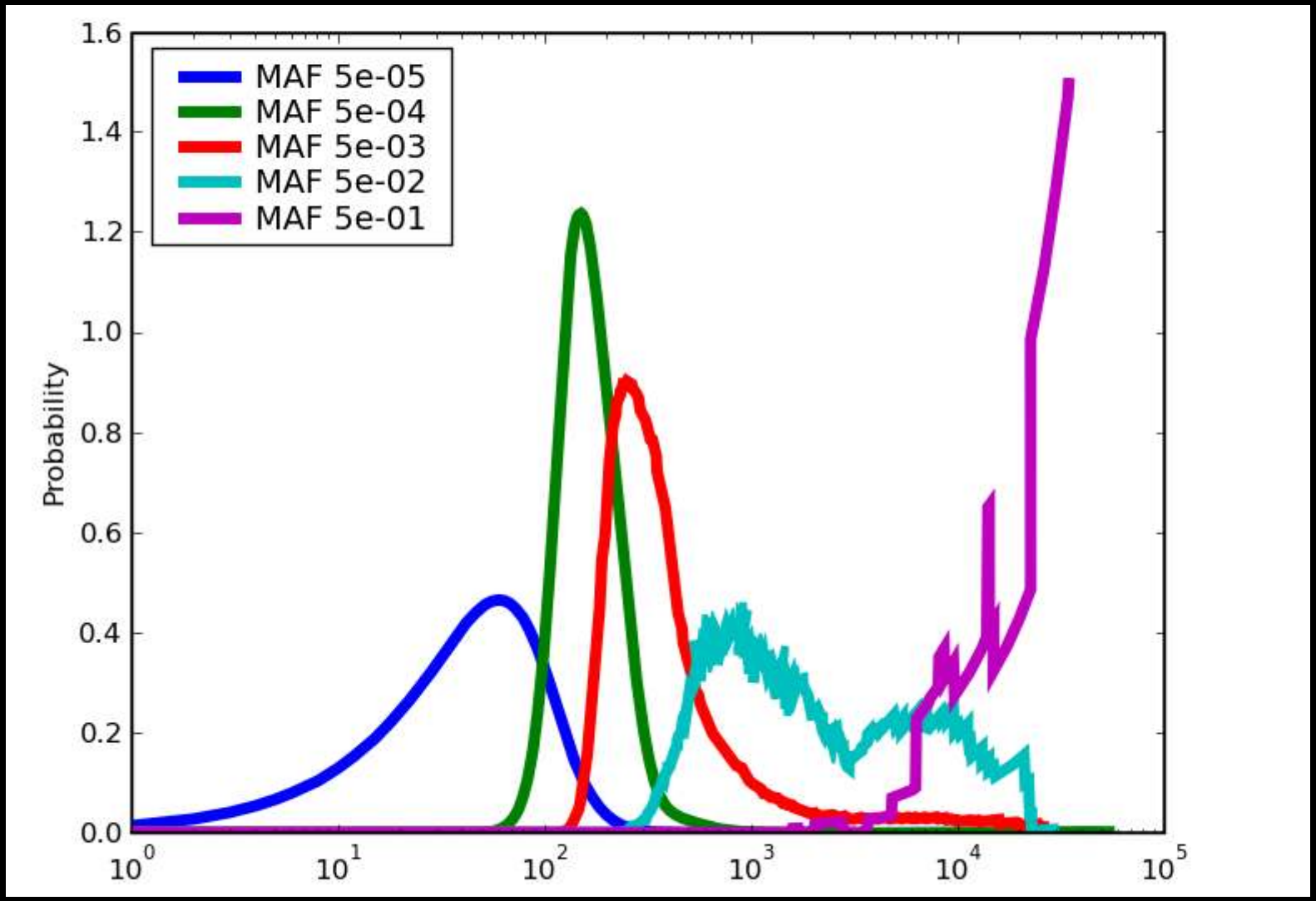


13-parameters demographic model with population divergence, migration, growth



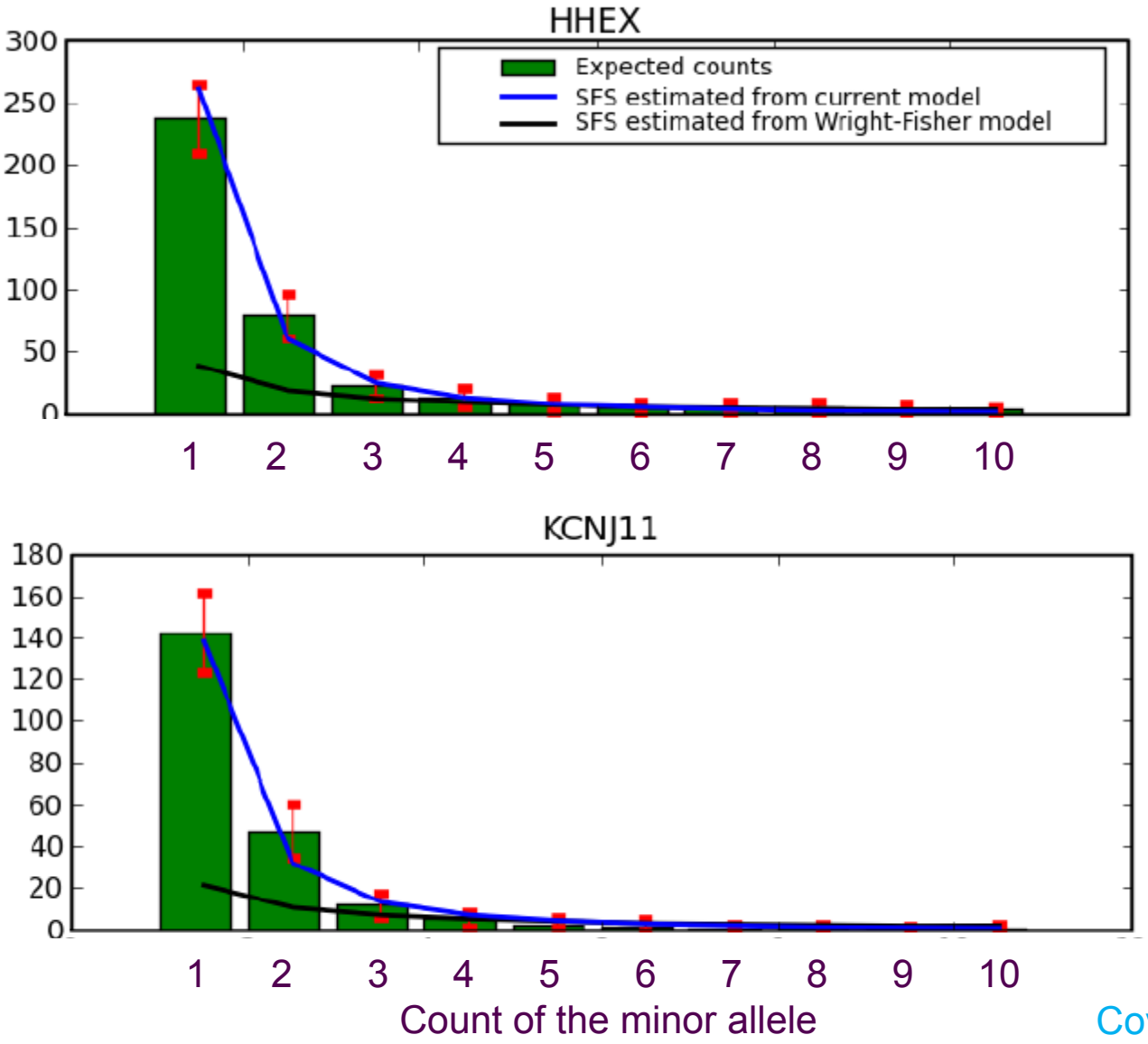
Gutenkunst *et al*. 2009 PLos Genet

Age distribution of alleles given their frequency



Allele age in generations

Recent explosive population growth fits the site frequency spectrum well



Coventry et al. 2010

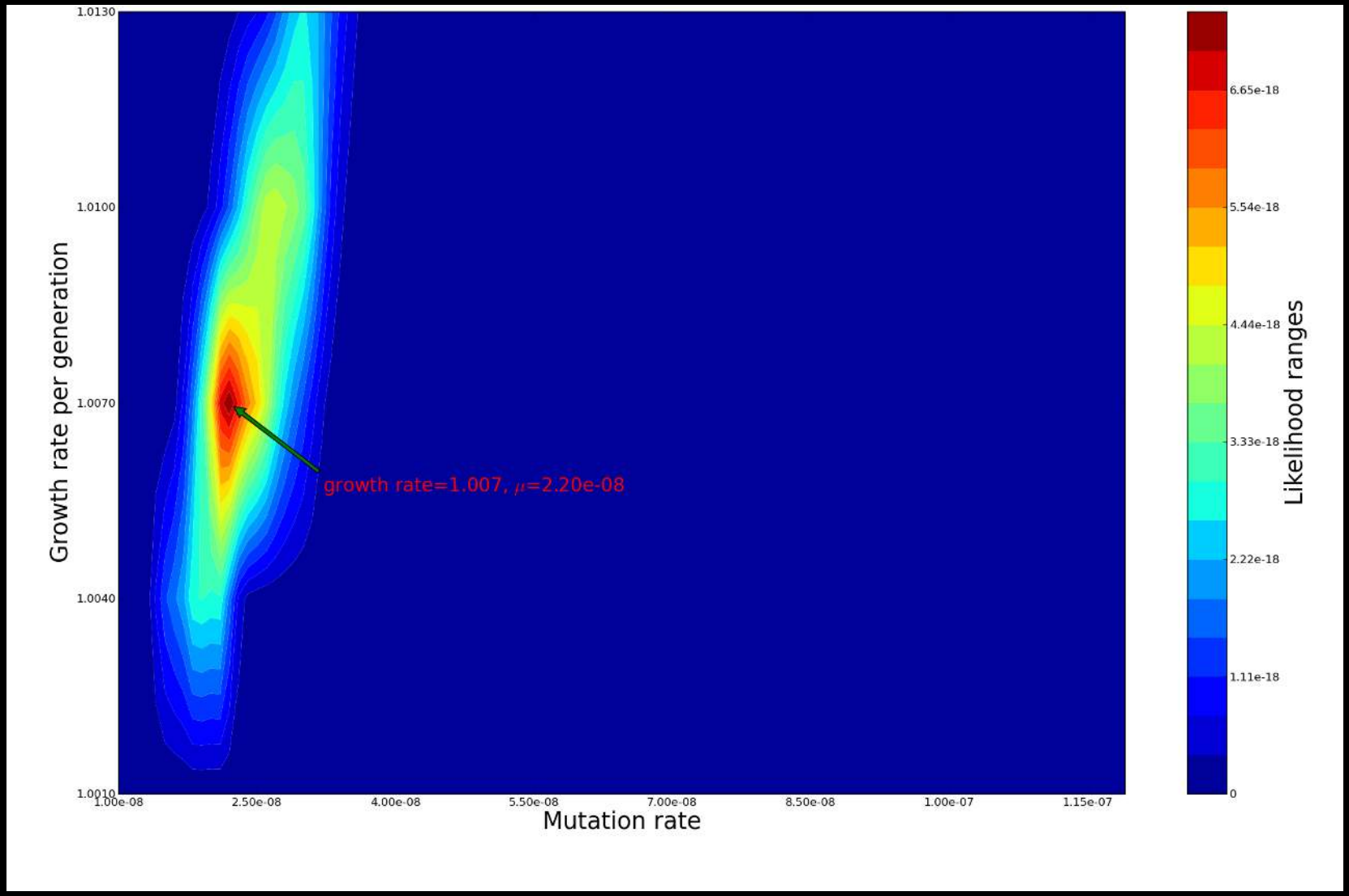
Decoupling of N_e and μ in very large samples

- When the tree is sufficiently tip-heavy, closely related lineage accumulate differences at a Poisson rate.
- This allows nearly direct estimation of the mutation rate, μ.

Wakeley and Takahashi 2003 MBE

Joint likelihood surface of mutation and growth rates (HHEX)

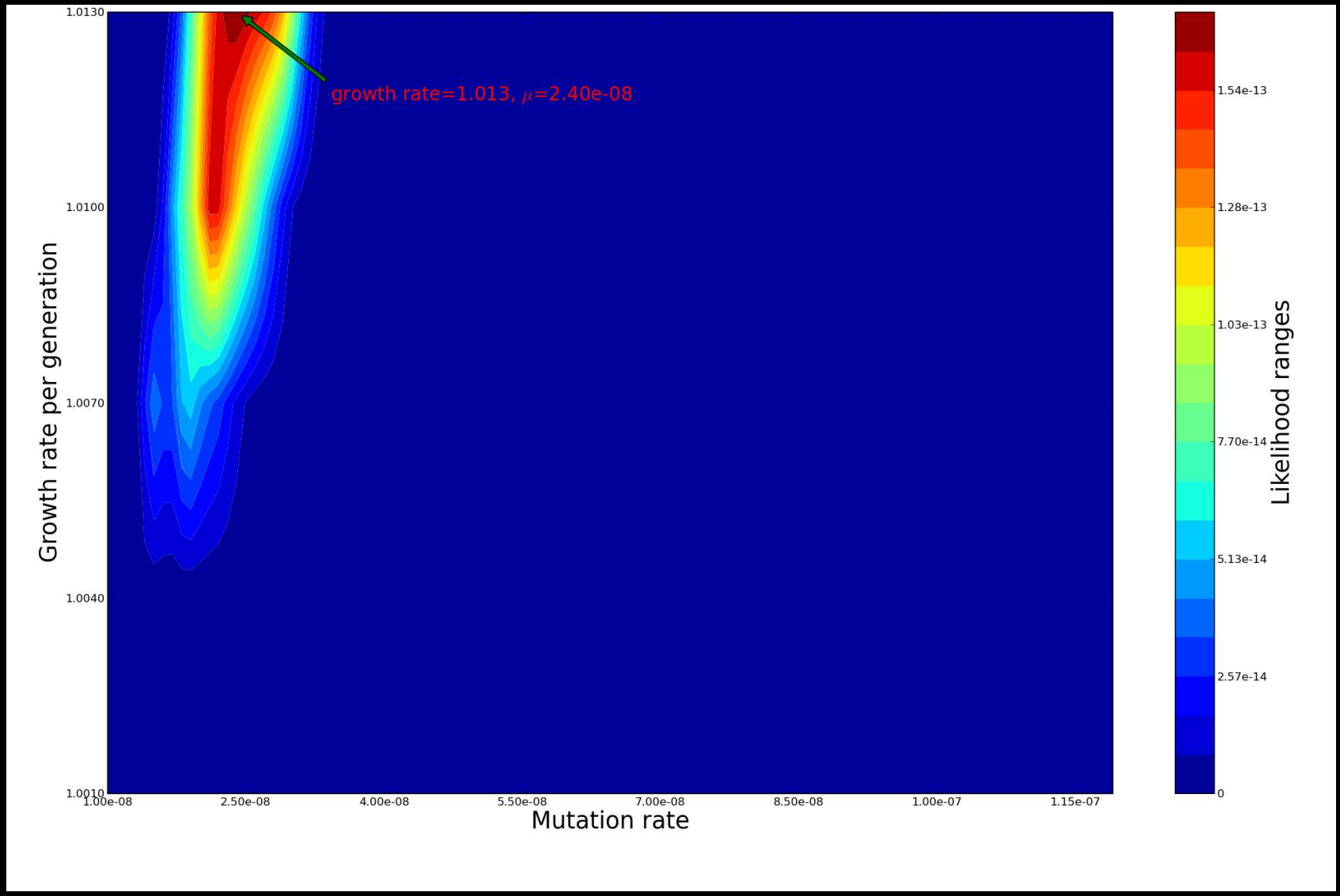
$r = 0.7\%$, $\mu = 4.90 \times 10^{-8}$



Coventry et al. 2010

Joint likelihood surface of mutation and growth rates (KCNJ11)

r = 1.3%, μ = 5.10 x 10^{-8}



Coventry et al. 2010

What proportion of SNPs found in each new full genome sequence do we expect to be novel?

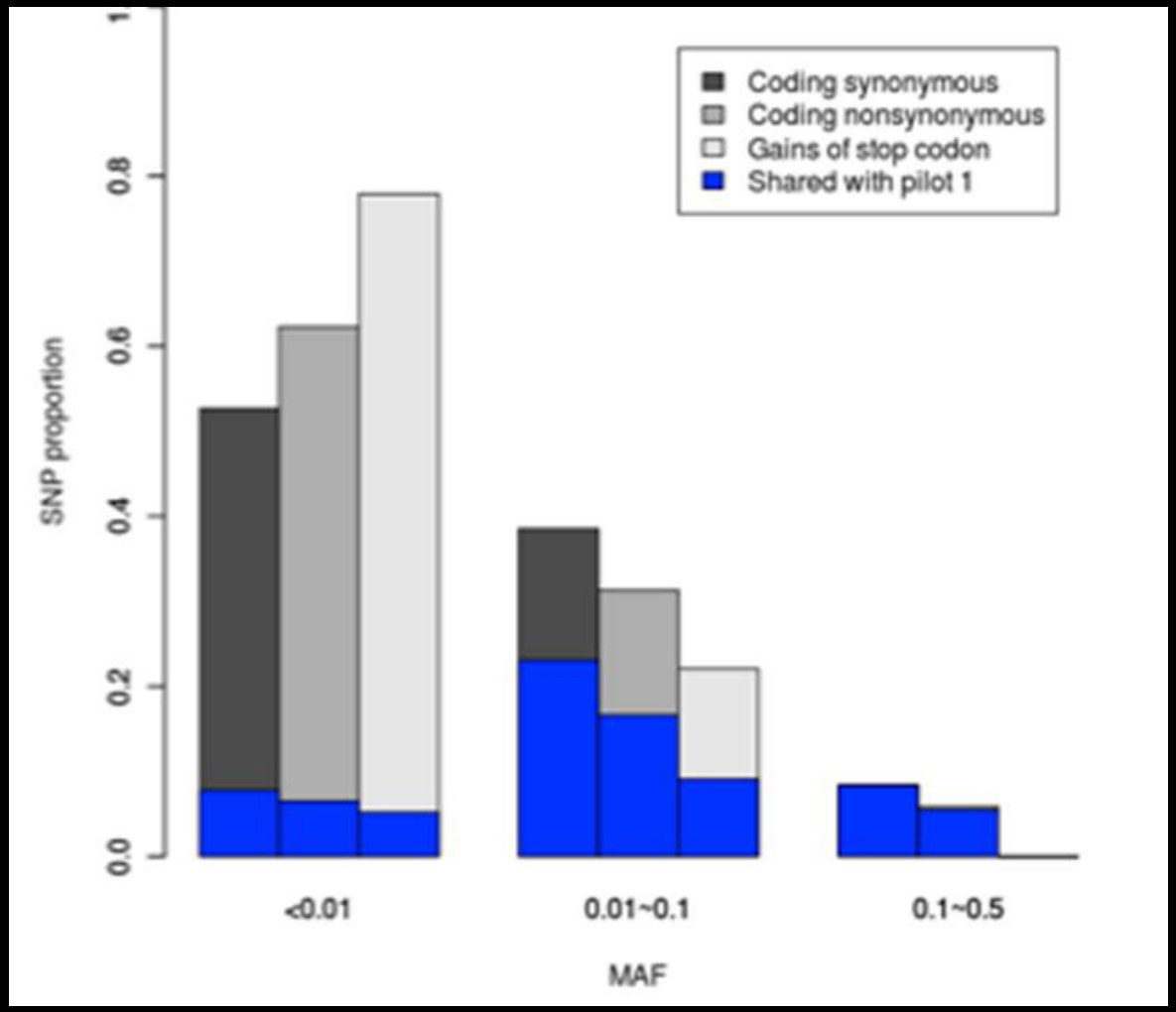
Sample	Ethnicity	Total SNPs*	Novel SNPs	Percent not found in 1000 Genomes
NA19730	Mexican-American	2,720,959	171,333	6.3%
By ancestry	Native American	1,652,924	106,325	6.4%
	Not Assigned	486,671	30,185	6.2%
SJK	Korean	2,772,777	83,902	3.0%
YH	Han Chinese	2,723,366	93,059	3.4%
ABT	Bantu (West African)	3,189,934	335,003	10.5%
KB1	San (S. African)	2,932,030	507,009	17.3%
NA18507	Yoruba (West African)	3,122,493	211,942	6.8%
NA19240 (SOLiD only)	Yoruba (West African)	2,887,093	161,833	5.6%
NA12878 (SOLiD only)	European	2,244,606	120,314	5.4%
Venter	European	2,406,043	147,132	6.1%
Watson	European	2,555,732	182,504	7.1%

Bryc *et al*., submitted (Jeff Kidd)

Expected discovery of novel SNPs

	ENCODE3		Constant population size	Model of European history	Model of European history with explosive growth	Just recent explosive growth
N = Diploid Sample size	European /E. Asian	African				
60	CEU: 3.6%±2.6% CHB: 4.1%±2.7% JPT: 3.6%±2.7%	YRI: 3.7%±2.5%	1.67% (1.77% in simulations)	1.88%	3.71%	3.41% (1.93-fold from constant)
89	CEU: 2.9%±2.4% CHB: 3.5%±2.4% JPT: 2.8%±1.9%	YRI: 2.7%±2.0%	1.12% (1.18%)	1.35%	3.32%	2.83% (2.40-fold)
118	CEU: 2.5%±2.2%	YRI: 2.2%±1.9%	0.85% (0.91%)	1.06%	3.11%	2.53% (2.78-fold)
1000			0.10% (0.10%)	0.14%	1.85%	1.57% (15.7-fold)

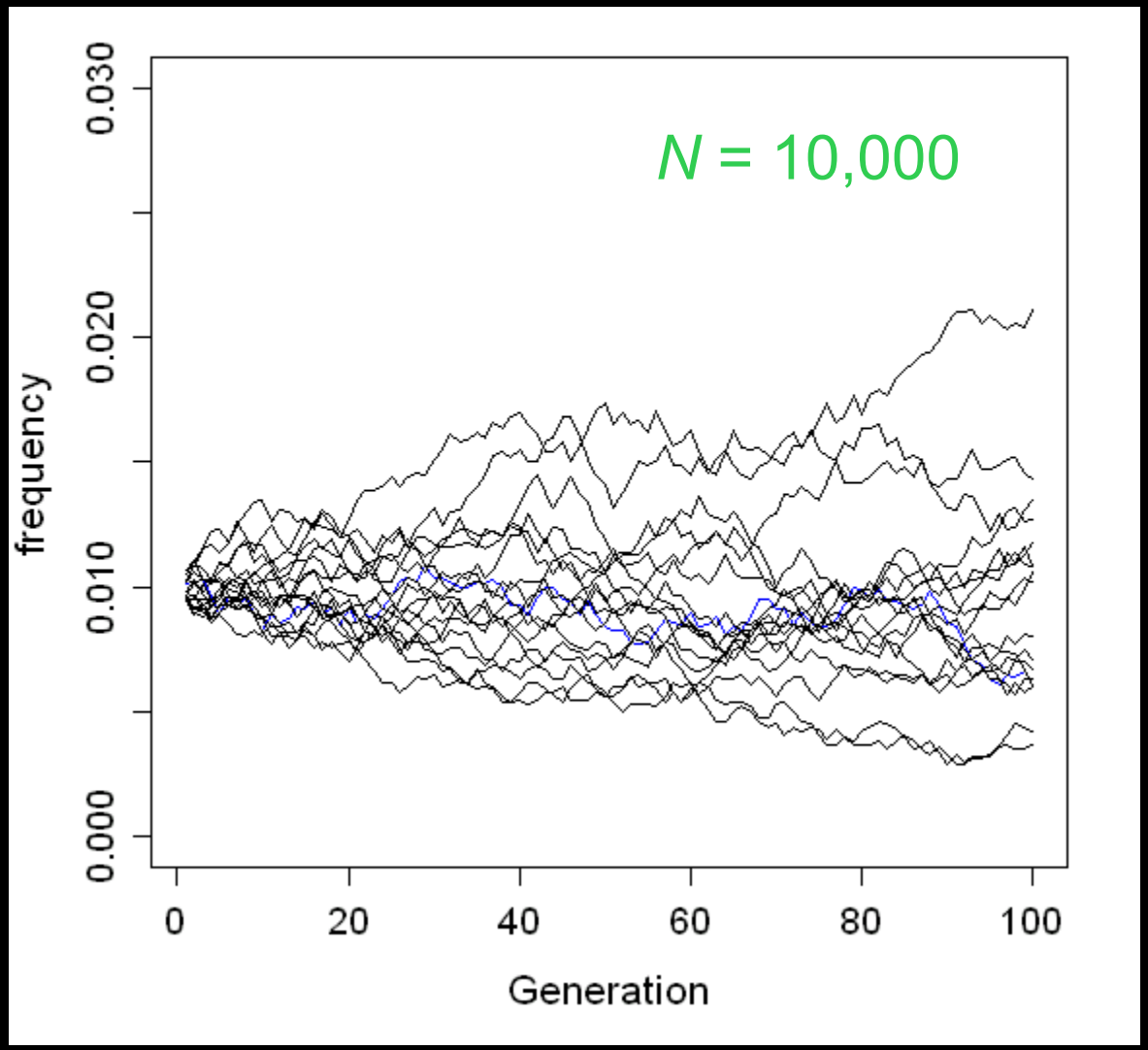
Rare variants are enriched for nonsynonymous and premature terminations



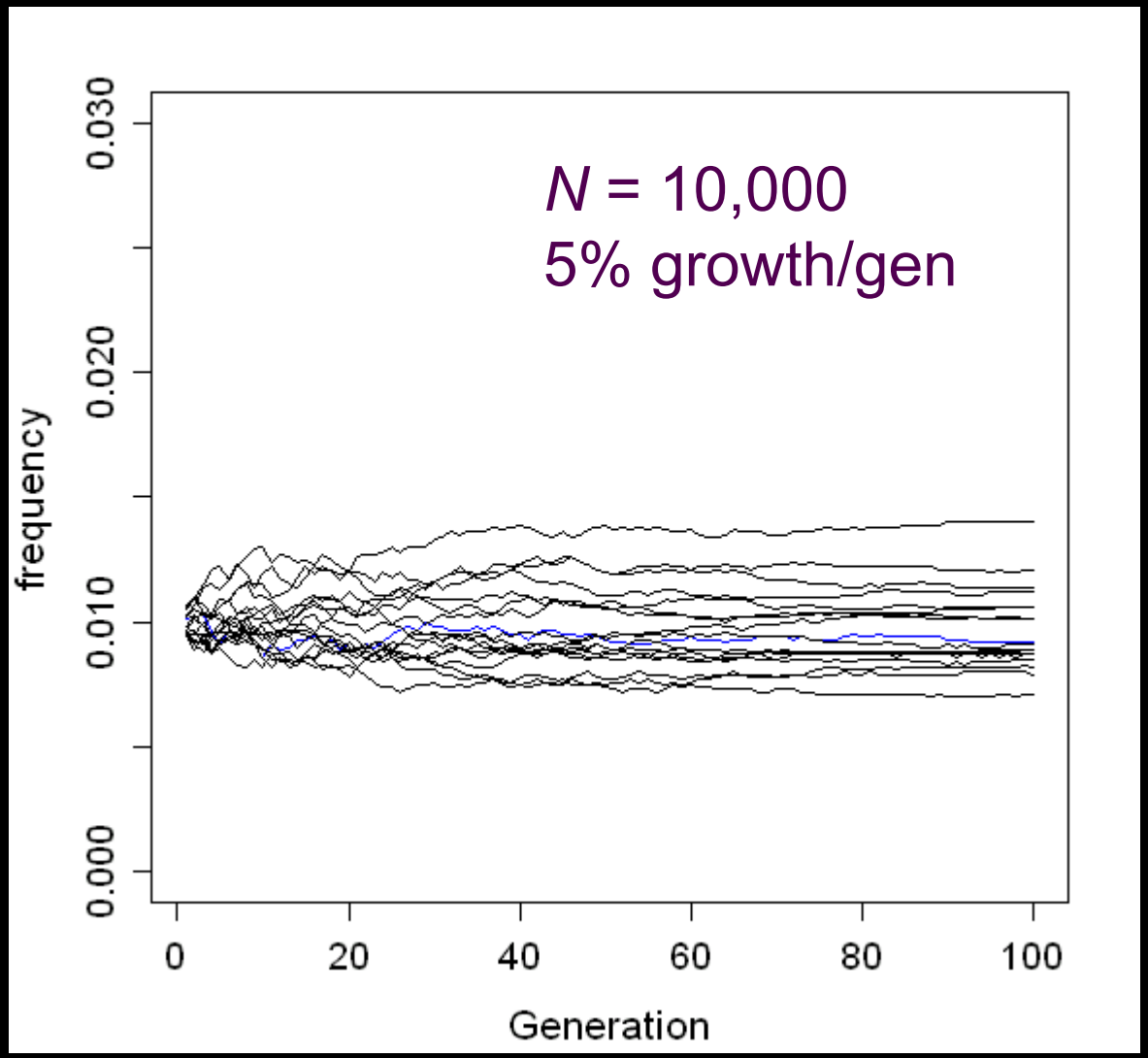
Marth et al. *submitted*

Explosive population growth inflates the expected number of deleterious mutations within each individual.

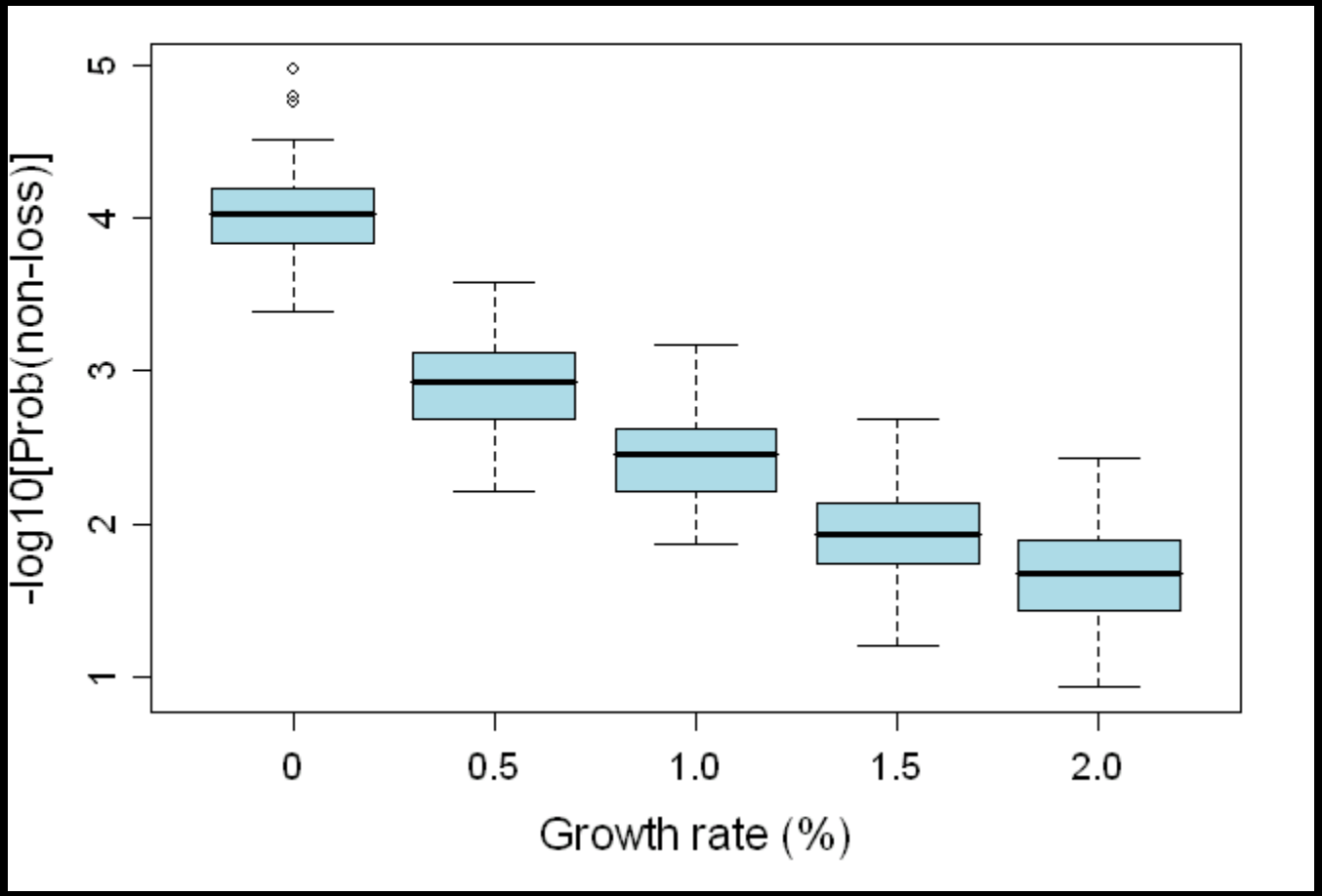
Standard Wright-Fisher drift



Wright-Fisher drift in a growing population



Probability of loss of new mutations declines with population growth



Conclusions about inference of past demography from genetic data

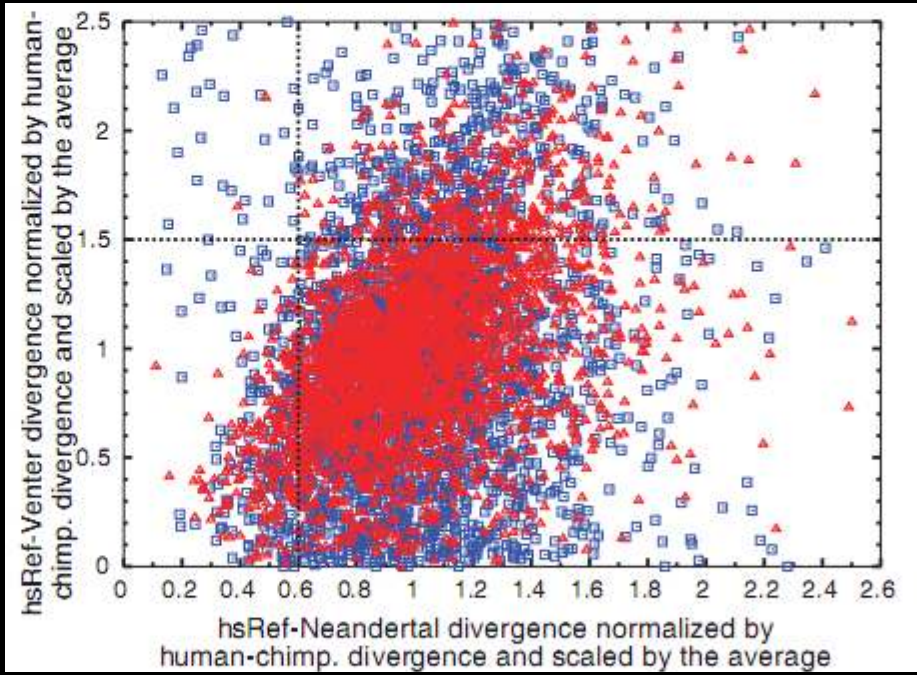
- The site frequency spectrum can reveal a lot about the past 2N generations or so.
- A huge excess of rare variants means that the population has expanded in the past.
- Huge samples must be sequenced to detect population growth in the last 100 generations.

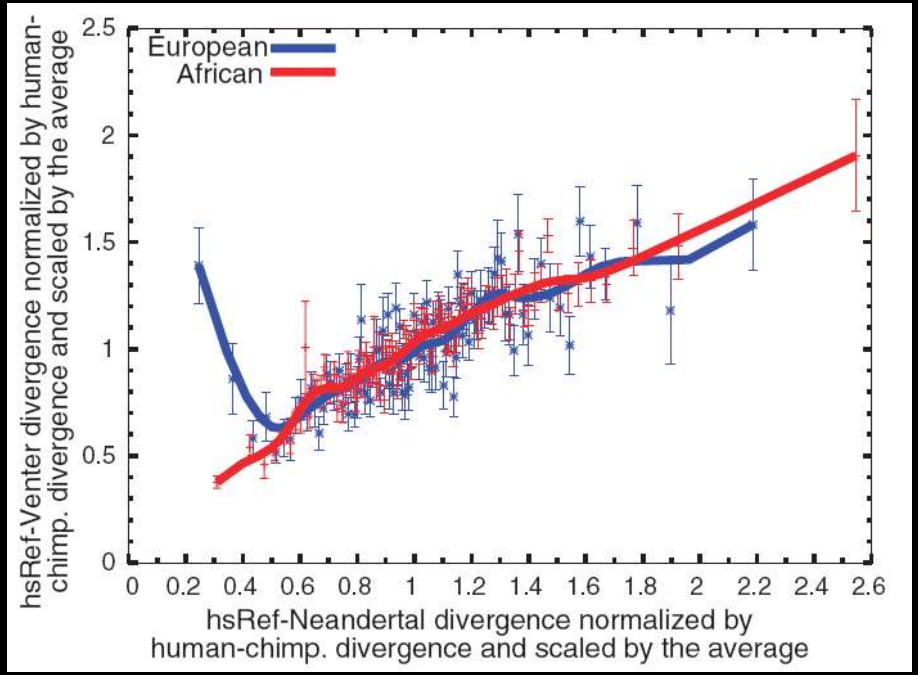
How can we determine whether modern humans admixed with Neanderthal?

Finding Neanderthal admixture

- Identify regions of the genome in which modern humans have deeper common ancestry than do Africans.
- Of those regions, ask whether there is excess sharing of variants in Neanderthal.
- If so, then these are candidates for genome regions from Neanderthal that were admixed into modern humans.

Regions with high Eur-Afr divergence, with super-deep Eur coalescence, often share close similarity with Neanderthal ((Eur, Neander),Afr)

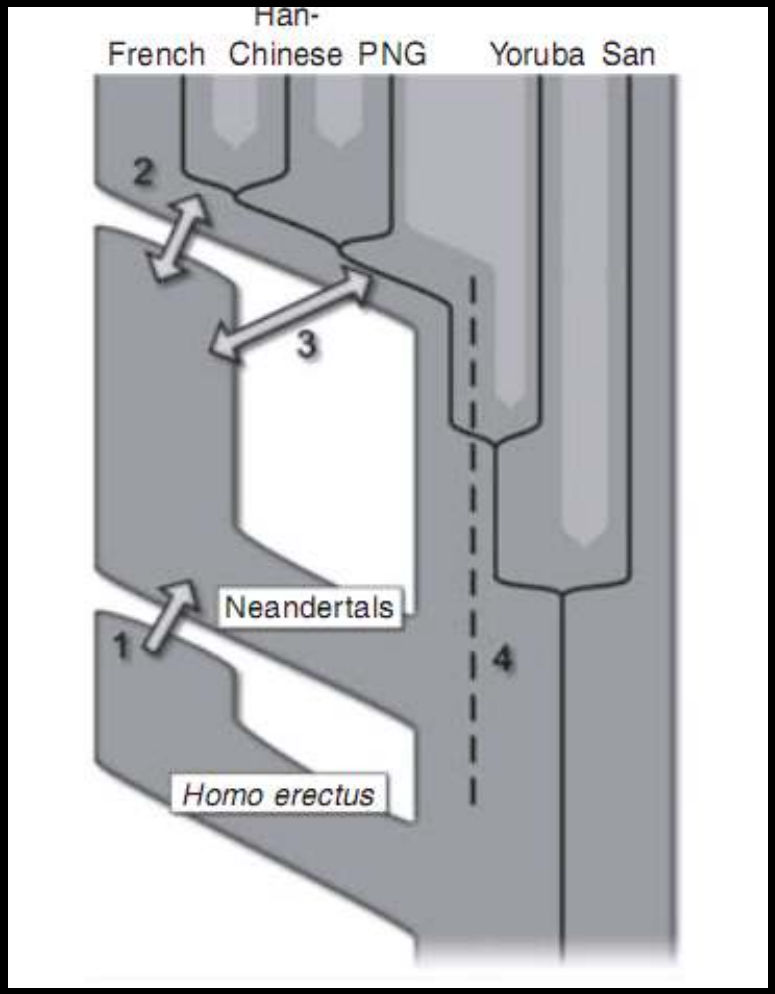




hsRef (human reference sequence) is African American, which means it is a mosaic of regions of African and European ancestry.

Green et al. (2010) Science 328:710-722

Models of Neanderthal admixture



1. Archaic admixture (high divergence of Neanderthal vs. modern human.
2. Admix to Modern European only (no support).
3. Data consistent with older admixture with out-of-Africa.
4. Deep population structure in Africa also consistent.

Primary questions of population genomic data

- How much variation is there?
- What evolutionary forces have exerted change to the variation?
- What can be inferred about the population's past demographic history from genetic variation?
- What are the limits to the resolution of these methods?

What are the limits to the resolution of these methods?

- Extremely powerful methods are extremely good at identifying artifacts in the data.
- Evolution happened only once, so it is a challenge to infer significance of some observations.
- Data sharing – how to assess quality of science without full data sharing?
- Data quality – How do we assess it? What keeps it from degenerating?
- Data overload – will we continue to have means to share petabytes of data?