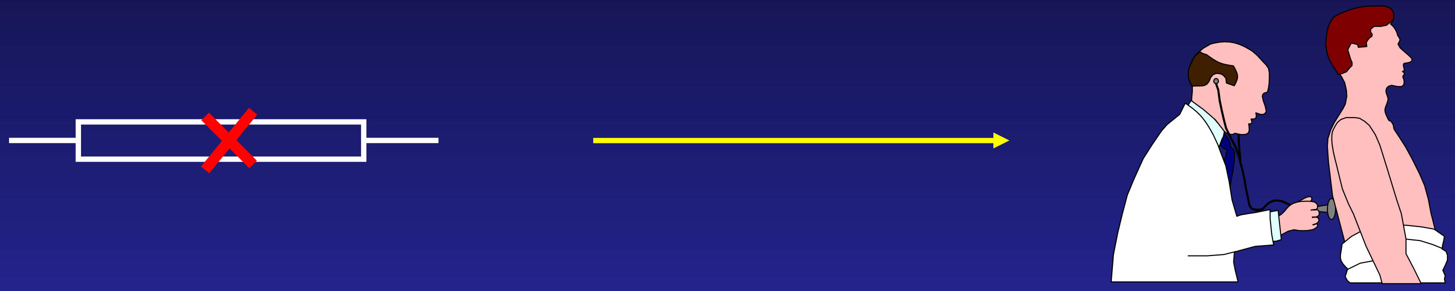
AGAGTTCTGCTCG
AGGGTTATGCGCG

Simple inheritance



Defective Gene

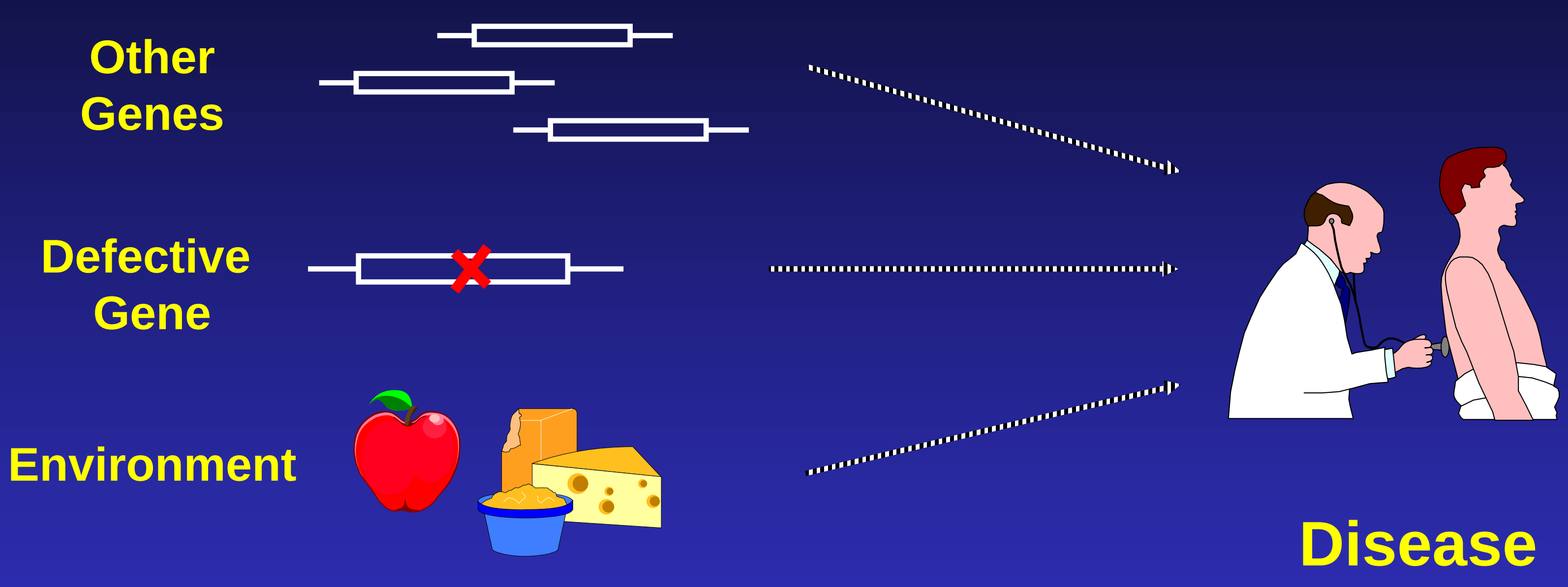
Disease

Cystic Fibrosis
Huntington’s Disease
Muscular Dystrophy

Hemochromatosis
Neurofibromatosis
Ataxia Telangiectasia

Achondroplasia
Fanconi Anemia
Werner Syndrome

Complex inheritance



Other Genes

Defective Gene

Environment

Disease

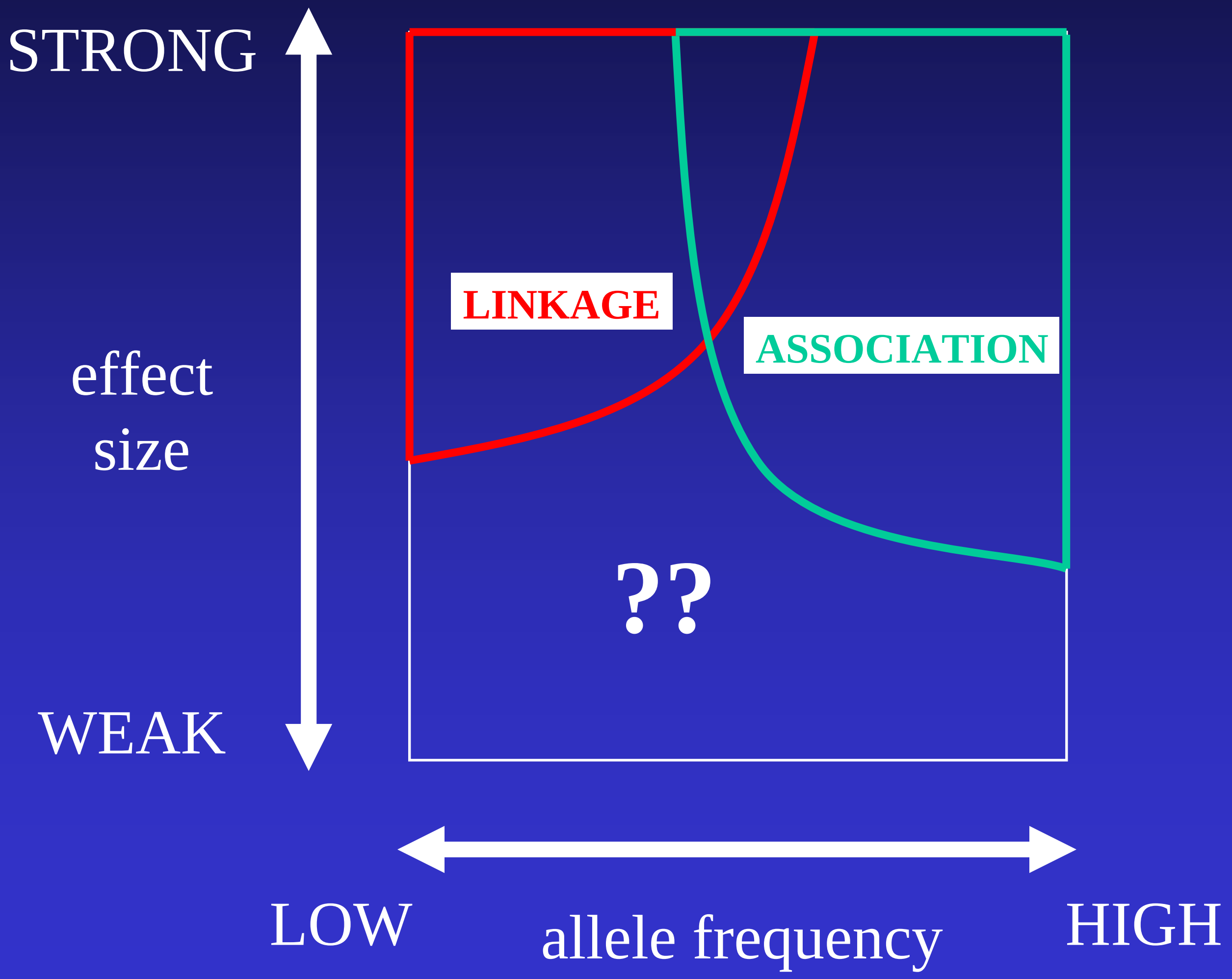
Diabetes
Obesity
Cancer

Heart Disease
Multiple Sclerosis
Asthma

Schizophrenia
Celiac Disease
Autism

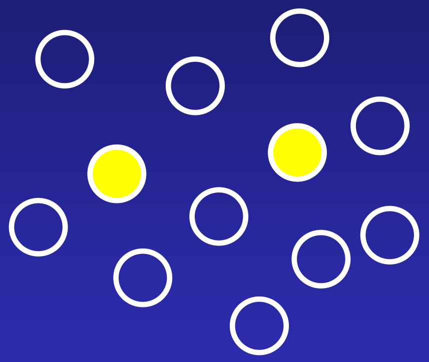
Susceptibility to infectious disease

Genetic model determines search strategy



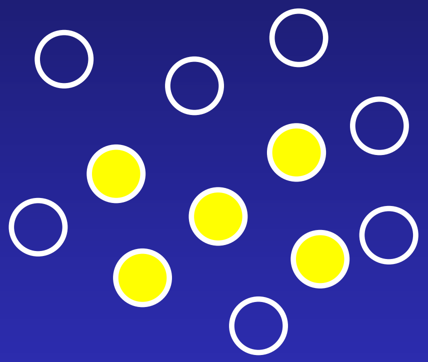
Association studies: common variants

General Population



APOE 2/ 3:85%
APOE 4: 15%

Alzheimer’s Patient



APOE 2/ 3: 60%
APOE 4: 40%

Total sequence variation in humans

Population size:	6x10^{9} (diploid)
Mutation rate:	2x10^{-8} per bp per generation
Expected “hits”:	240 for each bp

∴Every variant compatible with life exists in the population

BUT: Most are vanishingly rare

Compare 2 haploid genomes: 1 SNP per 1331 bp*

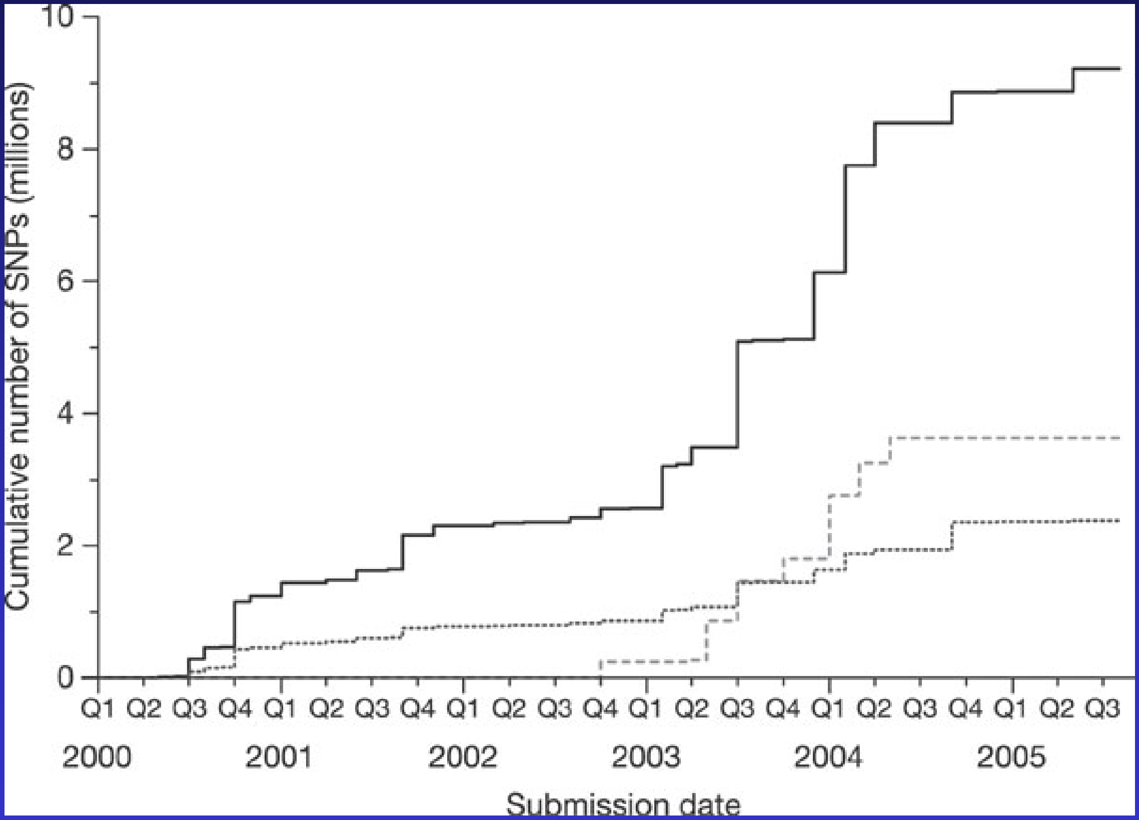
*The International SNP Map Working Group, *Nature* **409**:928 - 933 (2001)

Total SNPs and fraction in dbSNP (theory)

minimal allele frequency	expected SNPs (millions)	expected SNP frequency (bp)	expected % in database
1%	11.0	290	11-12
5%	7.1	450	15-17
10%	5.3	600	18-20
20%	3.3	960	21-25
30%	2.0	1570	23-27
40%	0.97	3280	24-28

L. Kruglyak and D. Nickerson, *Nat Genet* 27:234-236 2001

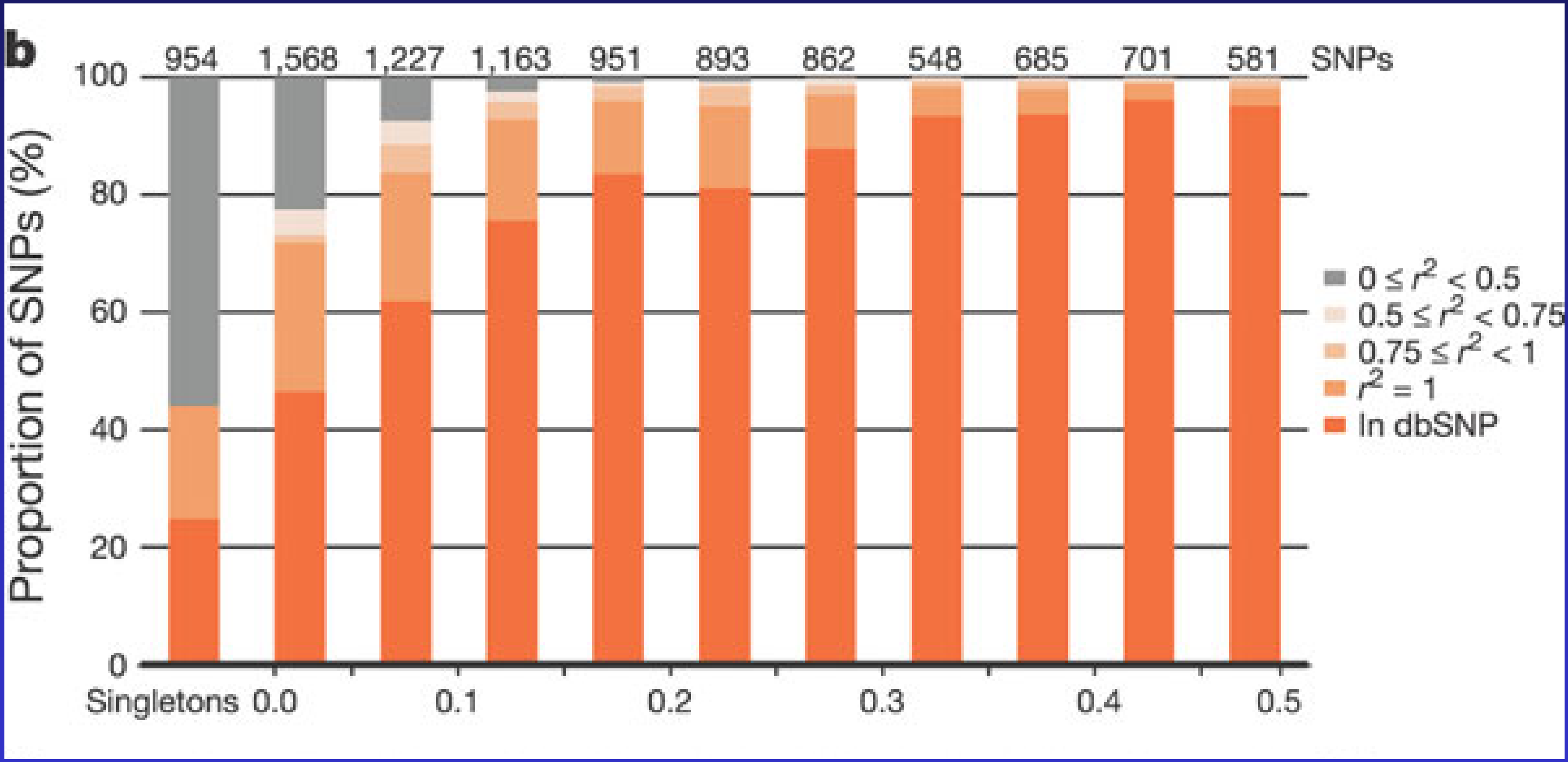
Number of SNPs in dbSNP



SNP detection rates

n	1%	5%	10%	20%	30%	40%
2	.21	.30	.36	.43	.47	.49
3	.32	.46	.55	.65	.71	.74
4	.39	.56	.66	.77	.83	.86
5	.44	.62	.73	.84	.90	.93
6	.48	.68	.78	.89	.94	.96
7	.52	.72	.83	.92	.96	.98
8	.55	.75	.86	.94	.98	.99
9	.57	.78	.88	.96	.98	.99
10	.59	.80	.90	.97	.99	-
16	.69	.89	.96	.99	-	-
24	.76	.95	.99	-	-	-
48	.87	.99	-	-	-	-
96	.95	-	-	-	-	-
192	.99	-	-	-	-	-

Completeness of dbSNP



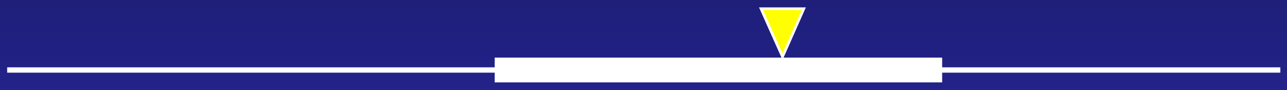
Toward comprehensive association studies

- 7 million common variants exist in genome
- Testing all for association is impractical today
- Can the list be reduced w/o loss of power?
  - Function
  - Linkage disequilibrium

Whole-genome association studies

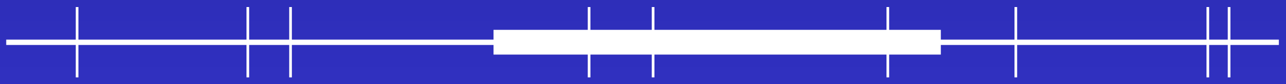
(1) Direct:

Catalog and test all functional variants for association



(2) Indirect:

Use dense SNP map and test for linkage disequilibrium



Collins, Guyer, Chakravarti (1997). Science 278:1580-81

How many functional variants?

1. CODING

Human genes:	30,000
cSNPs per gene*:	4
Amino acid changes*:	40%
Nonconservative:	16%

50,000 nonsynonymous cSNPs

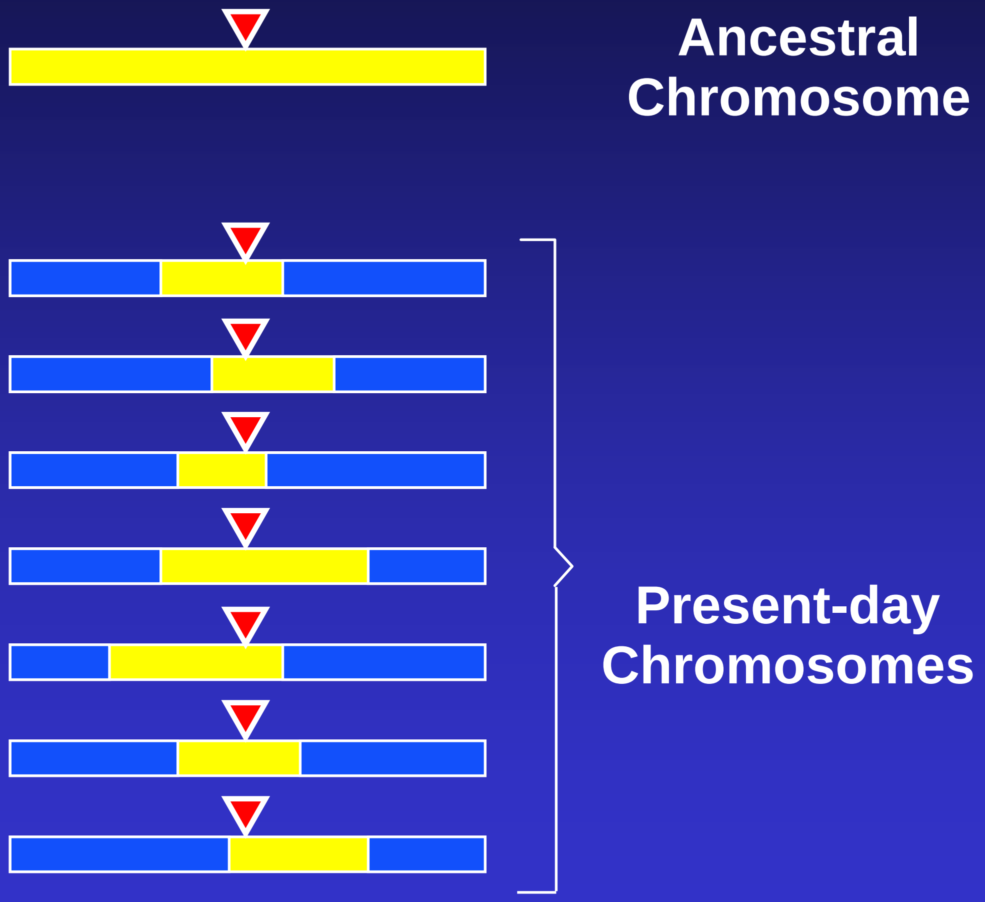
20,000 nonconservative cSNPs

prioritize based on structure, conservation

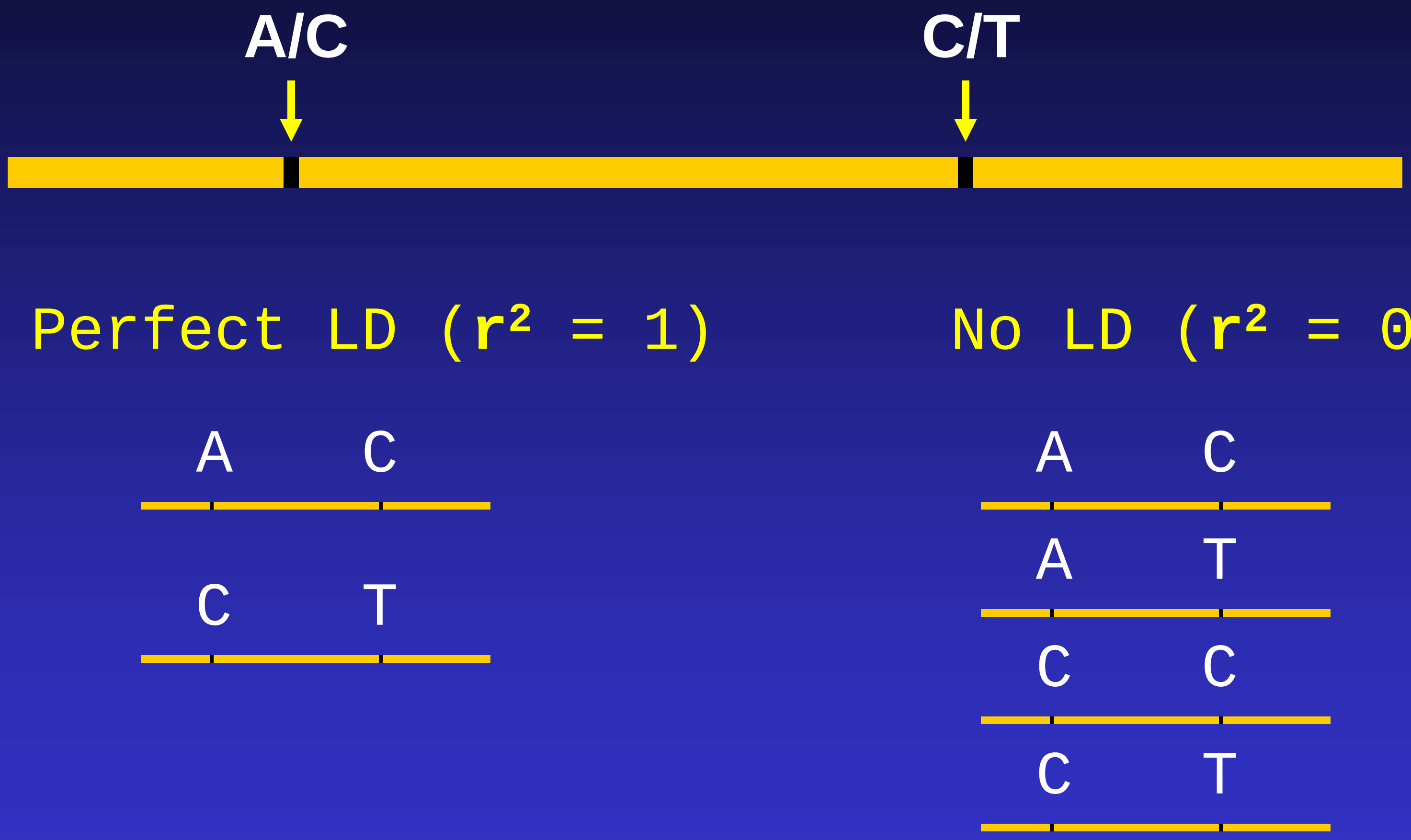
2. NONCODING/REGULATORY ???

*Cargill *et al.*, Halushka *et al.*, *Nat. Genet.* 1999

Linkage disequilibrium around variant



Pairwise linkage disequilibrium between SNPs

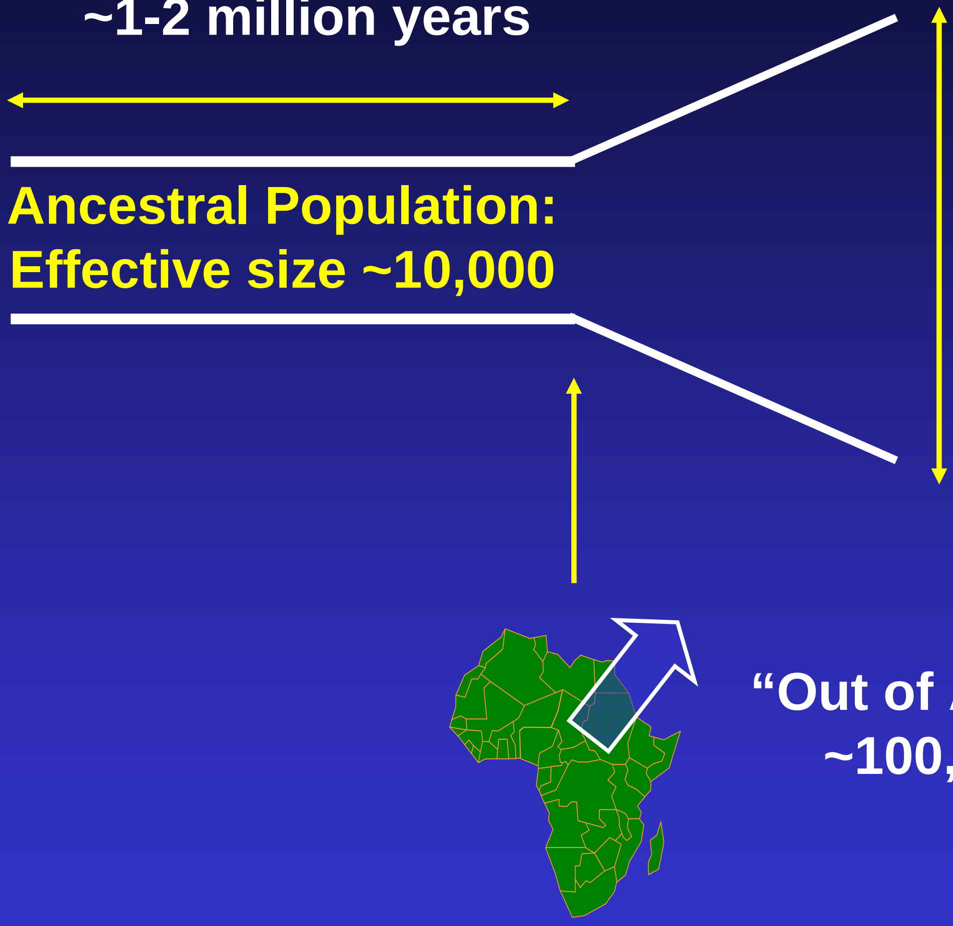


Perfect LD ($\mathbf{r}^2 = 1$)

No LD ($\mathbf{r}^2 = 0$)

Sample size to detect indirect association scale

Global model of human demographic history



~1-2 million years

Ancestral Population:
Effective size ~10,000

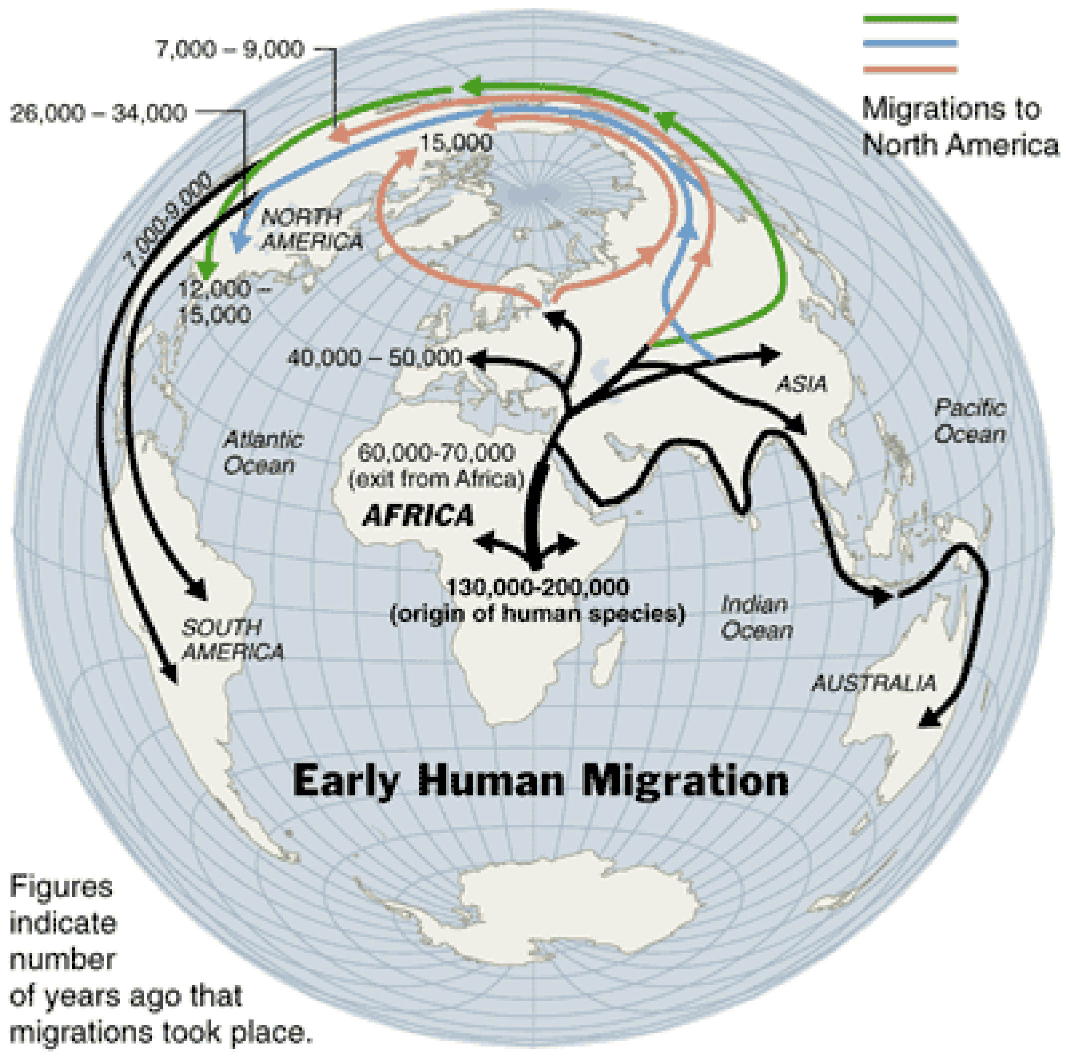
Today's
Population:
~6,000,000,000

"Out of Africa" migration
~100,000 years ago

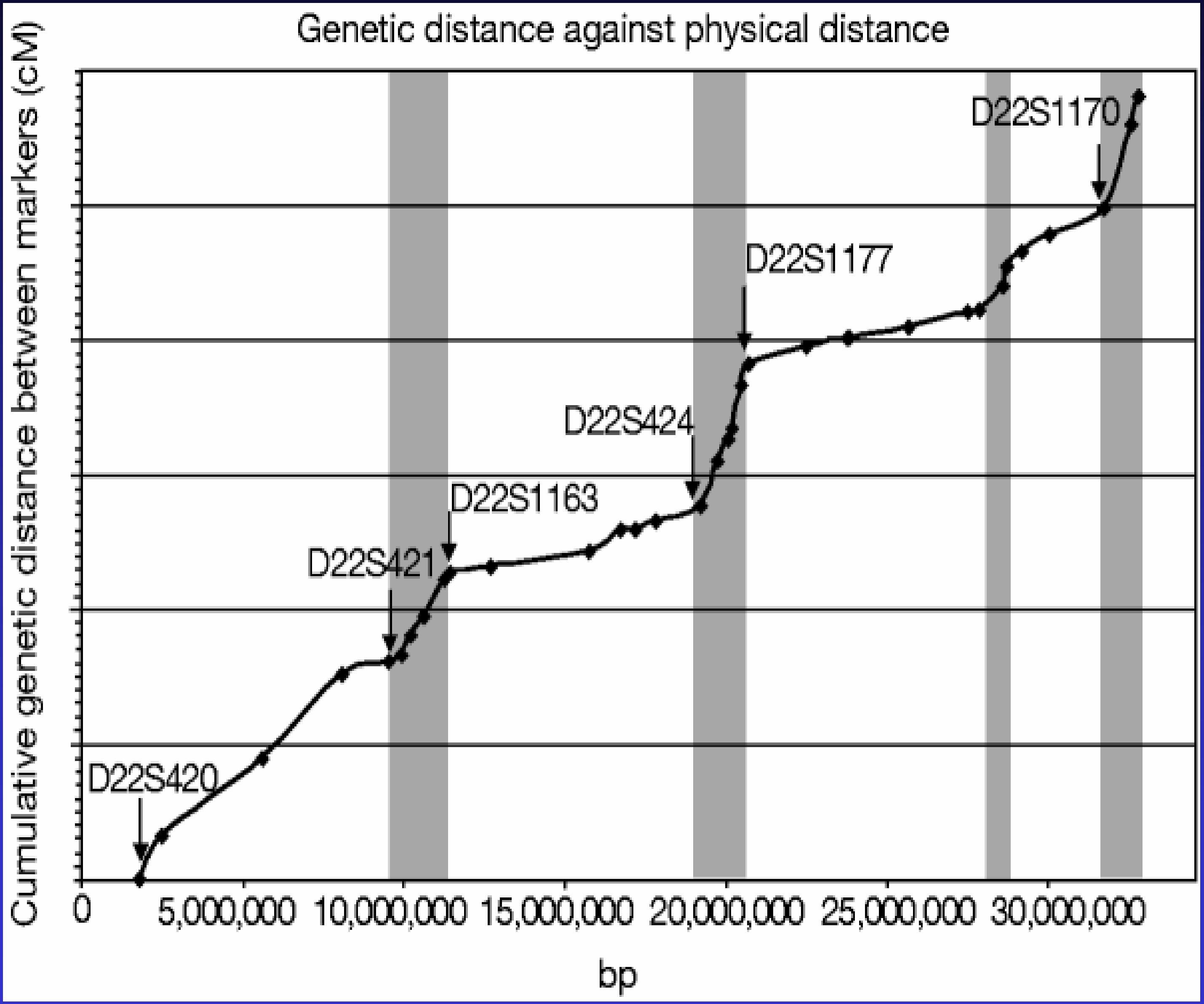
Under this model, useful average values of r^2 exte

Kruglyak *Nature Genetics* **(1999)**

Human migration out of Africa

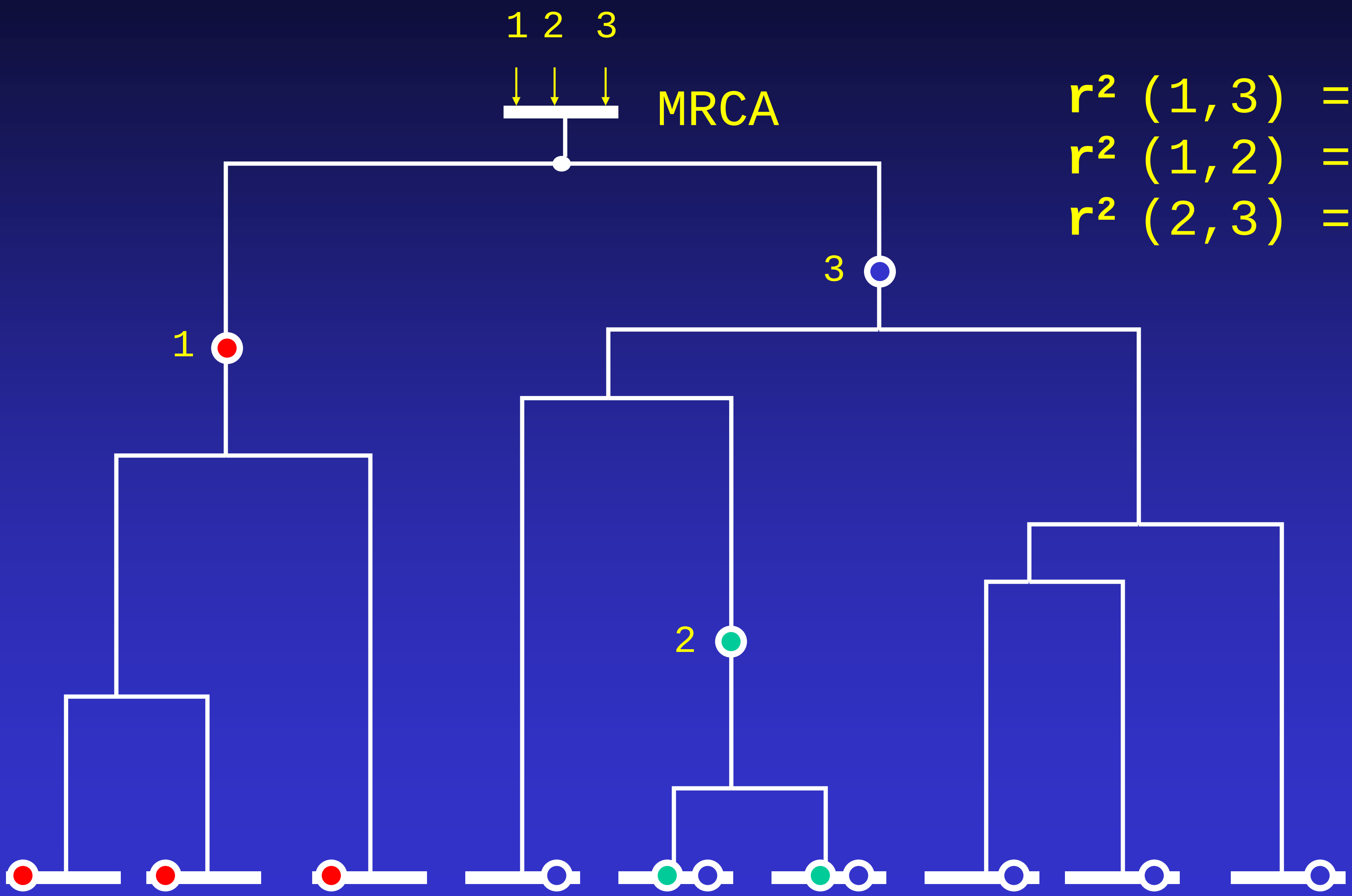


Recombination is not uniform on chromosome 22



From Dunham *et al*., Nature 402:489-495, 1999

Age of mutations and LD



r^2 (1,3) = 1

r^2 (1,2) = 0.1

r^2 (2,3) = 0.1

Need Empirical Mesurement of LD

Across the entire genome

In multiple populations

Kruglyak *Proc Natl Acad Sci USA* **(1999)**

The International HapMap Consortium

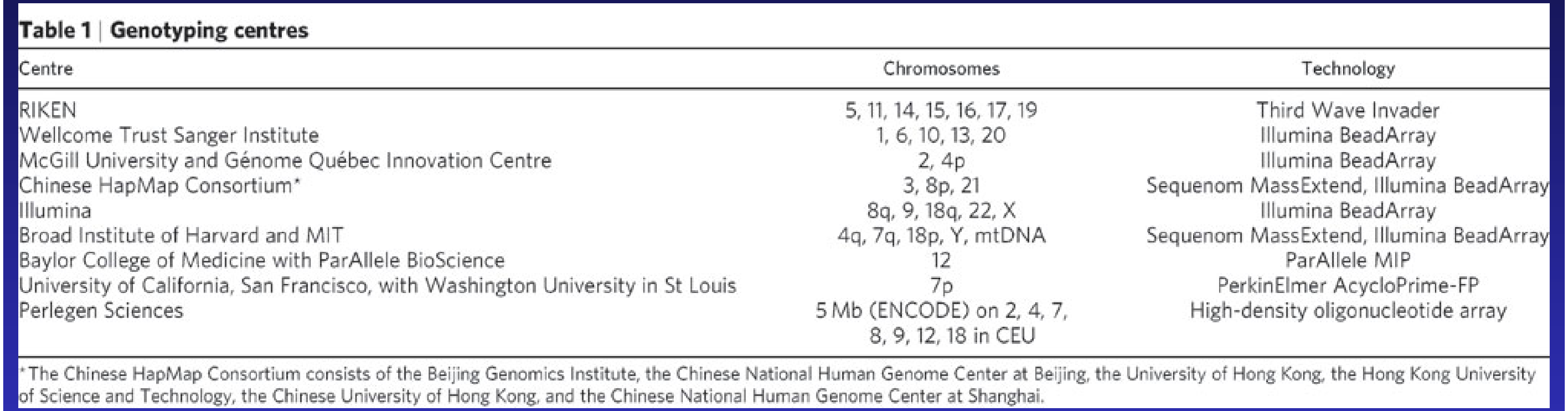


Table 1 | Genotyping centres

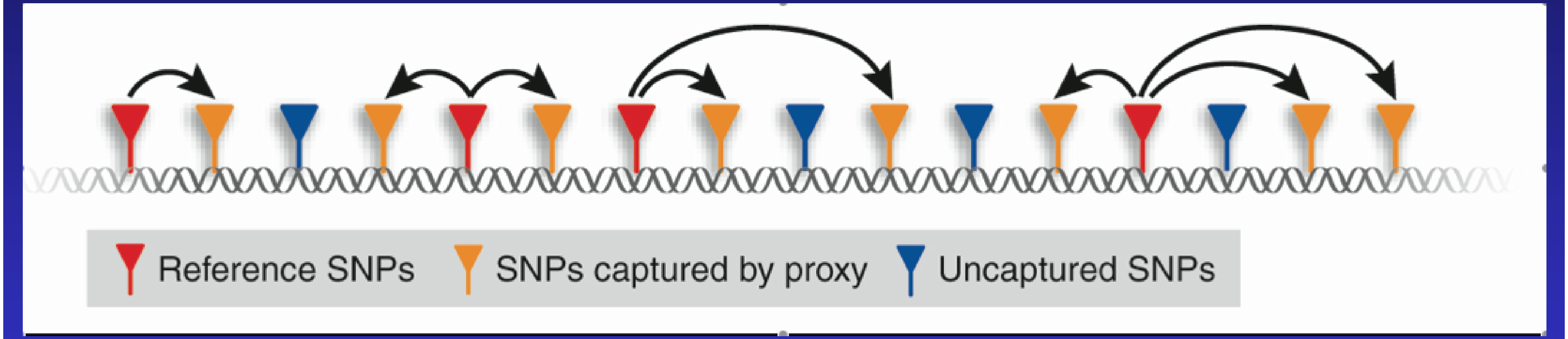
Centre	Chromosomes	Technology
RIKEN	5, 11, 14, 15, 16, 17, 19	Third Wave Invader
Wellcome Trust Sanger Institute	1, 6, 10, 13, 20	Illumina BeadArray
McGill University and Génome Québec Innovation Centre	2, 4p	Illumina BeadArray
Chinese HapMap Consortium*	3, 8p, 21	Sequenom MassExtend, Illumina BeadArray
Illumina	8q, 9, 18q, 22, X	Illumina BeadArray
Broad Institute of Harvard and MIT	4q, 7q, 18p, Y, mtDNA	Sequenom MassExtend, Illumina BeadArray
Baylor College of Medicine with ParAllele BioScience	12	ParAllele MIP
University of California, San Francisco, with Washington University in St Louis	7p	PerkinElmer AcycloPrime-FP
Perlegen Sciences	5 Mb (ENCODE) on 2, 4, 7, 8, 9, 12, 18 in CEU	High-density oligonucleotide array

*The Chinese HapMap Consortium consists of the Beijing Genomics Institute, the Chinese National Human Genome Center at Beijing, the University of Hong Kong, the Hong Kong University of Science and Technology, the Chinese University of Hong Kong, and the Chinese National Human Genome Center at Shanghai.

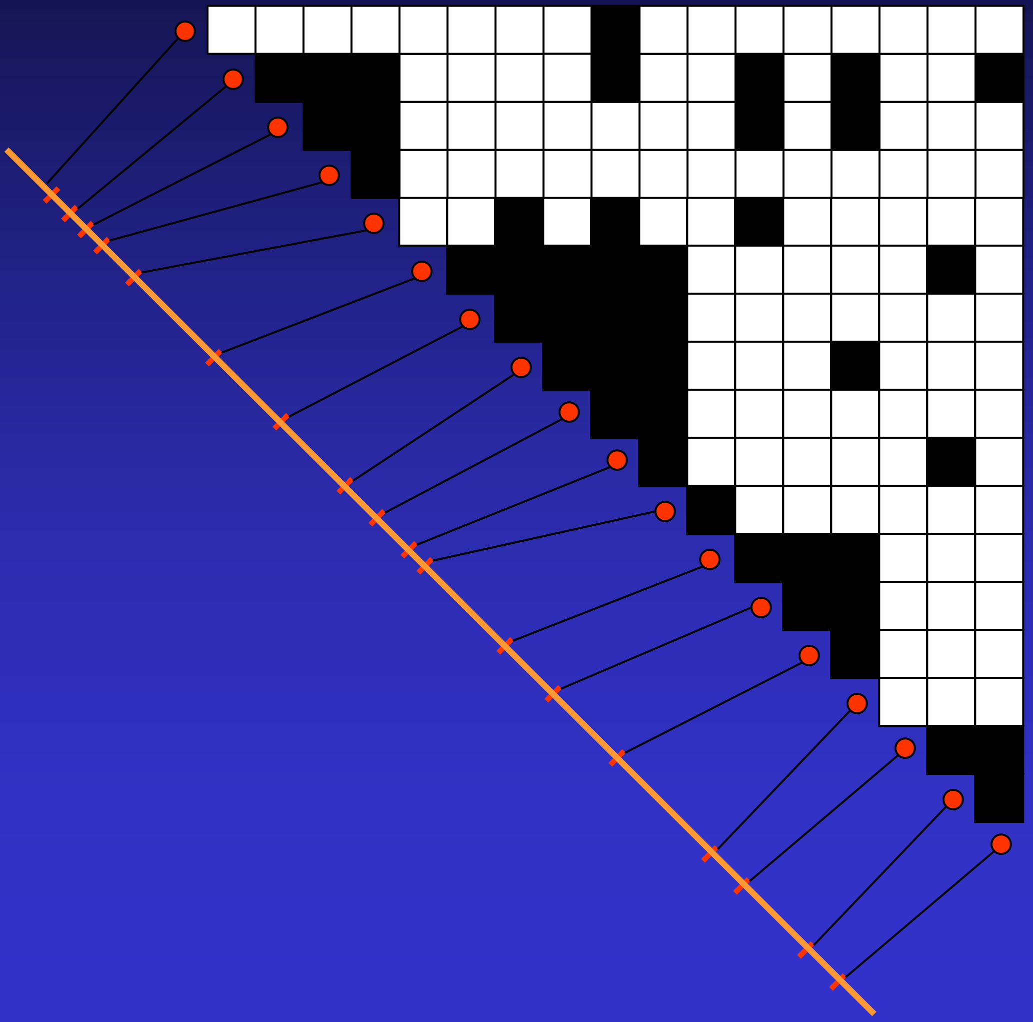
1 million SNPs genotyped in 90 individuals from each of 3 ethnic groups

Nature 2005

Goal: panel of proxy SNPs



Optimal selection of SNPs for LD studies

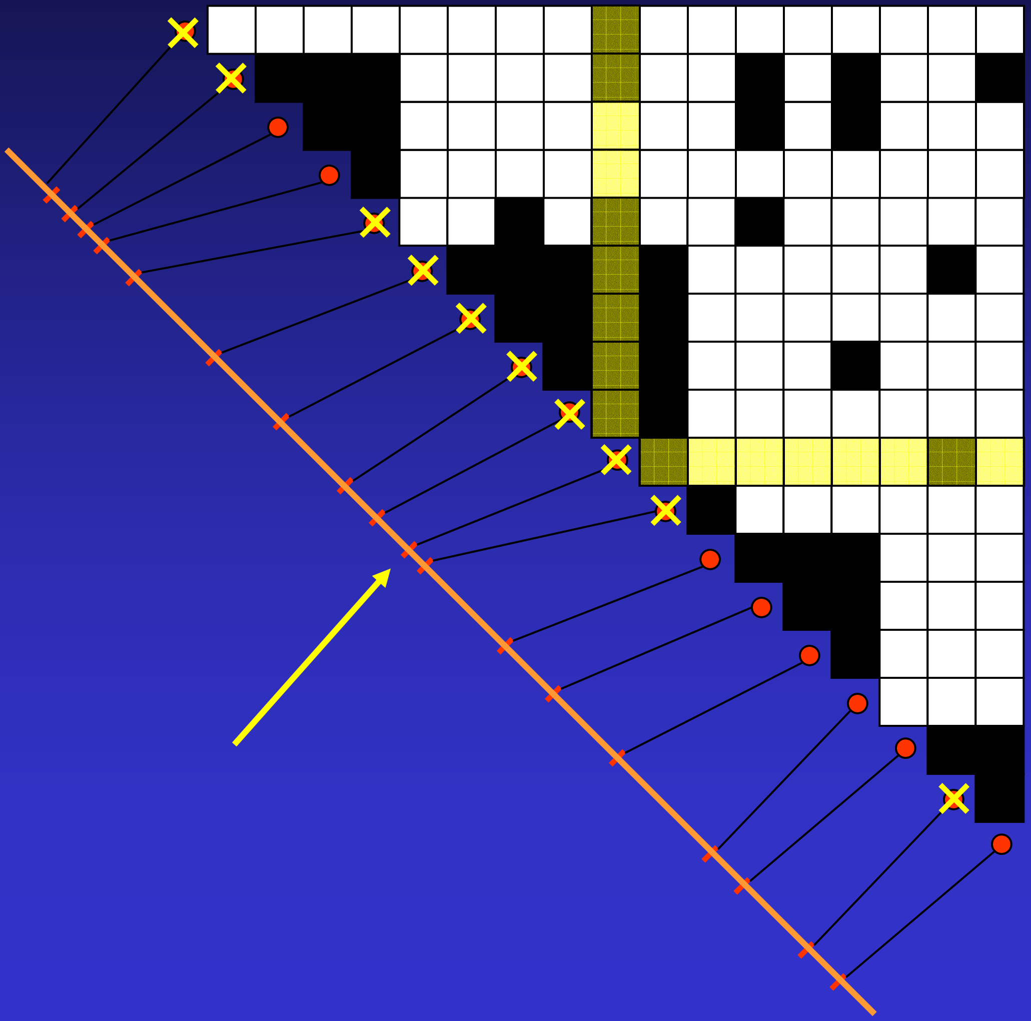


18 SNPs

■ High LD

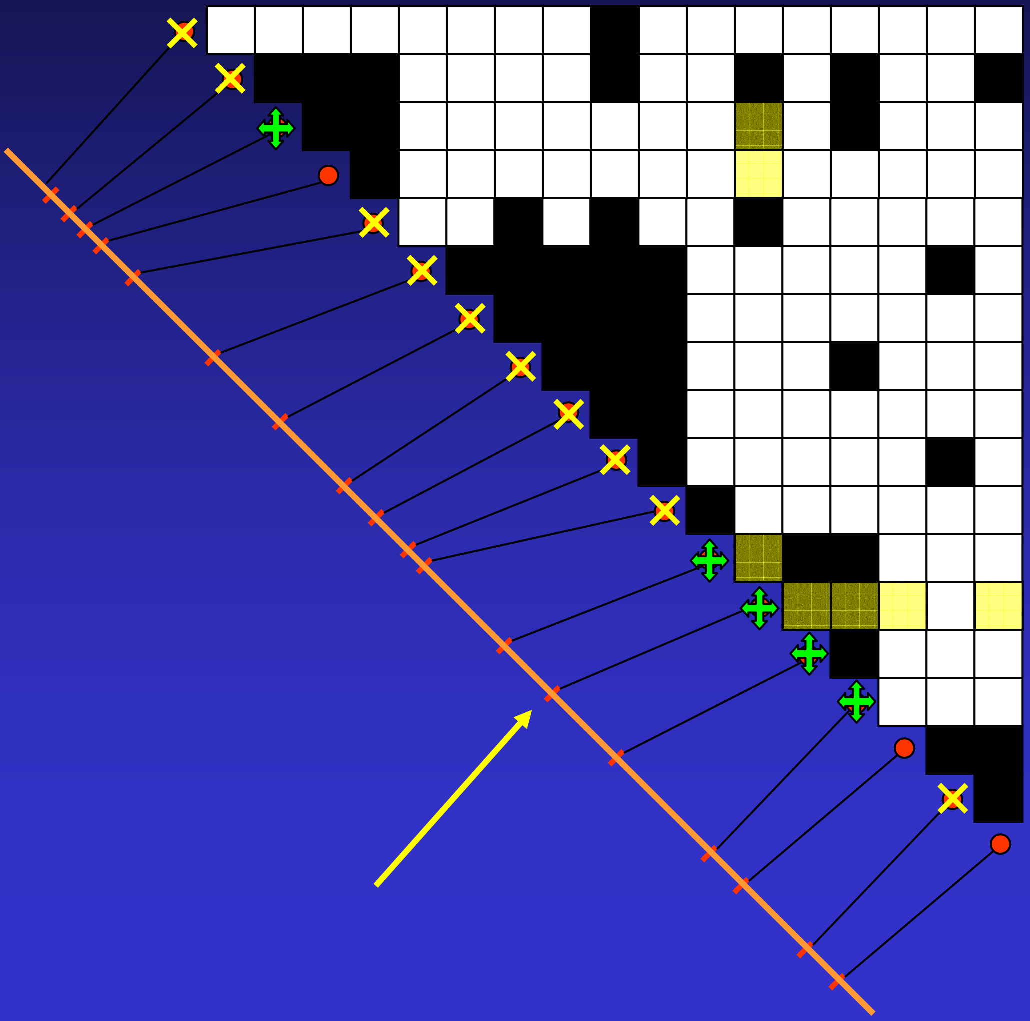
□ Low LD

Optimal selection of SNPs for LD studies



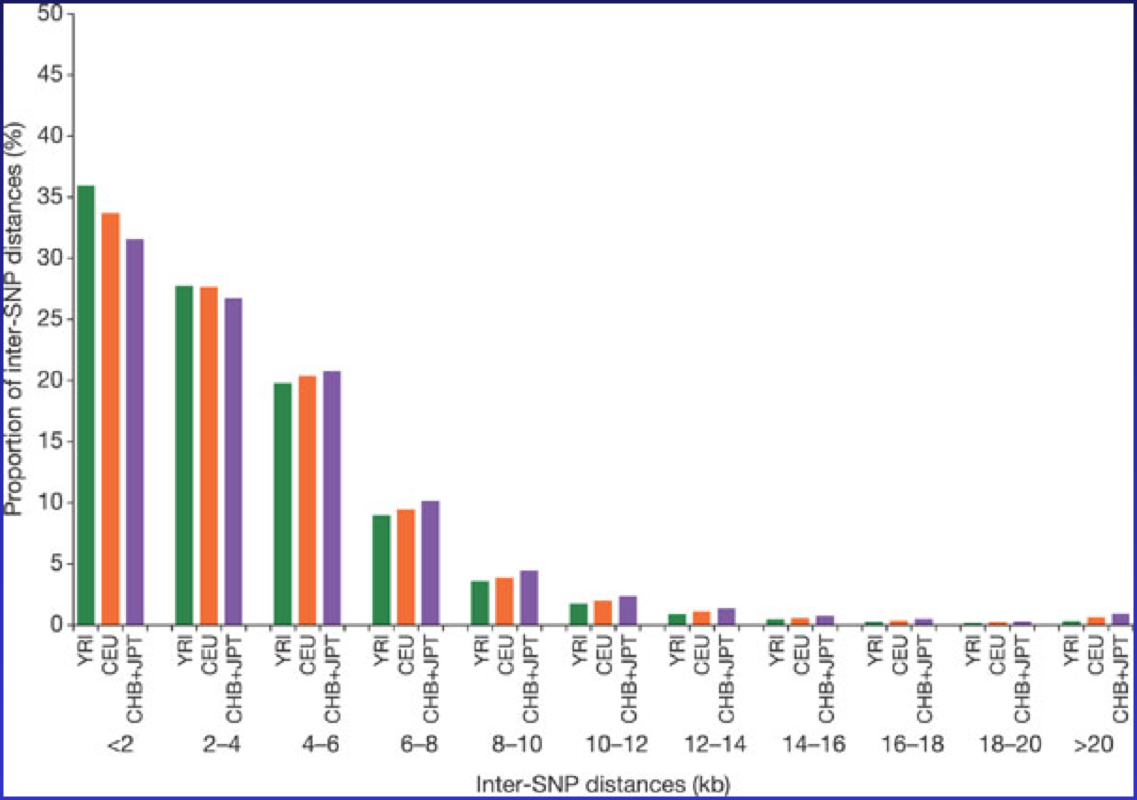
One SNP
assays 10

Optimal selection of SNPs for LD studies

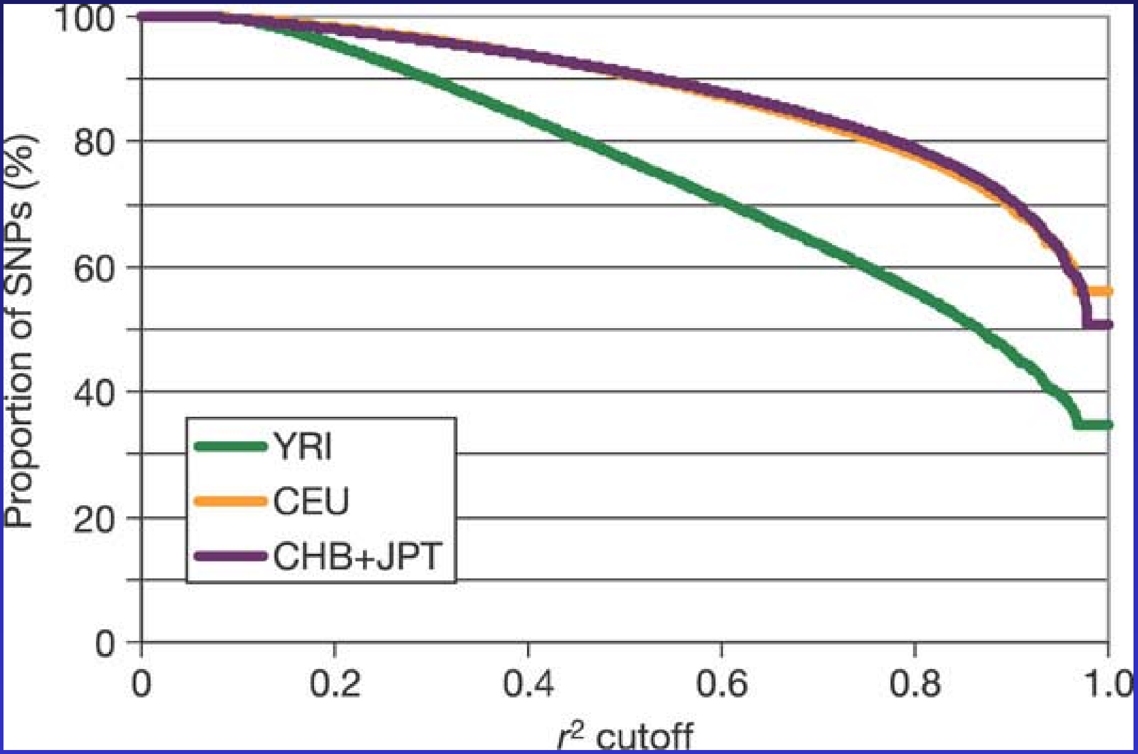


Two SNPs
assay 15

Distances between HapMap SNPs



Phase I SNPs captured by proxies



Required number of proxy SNPs

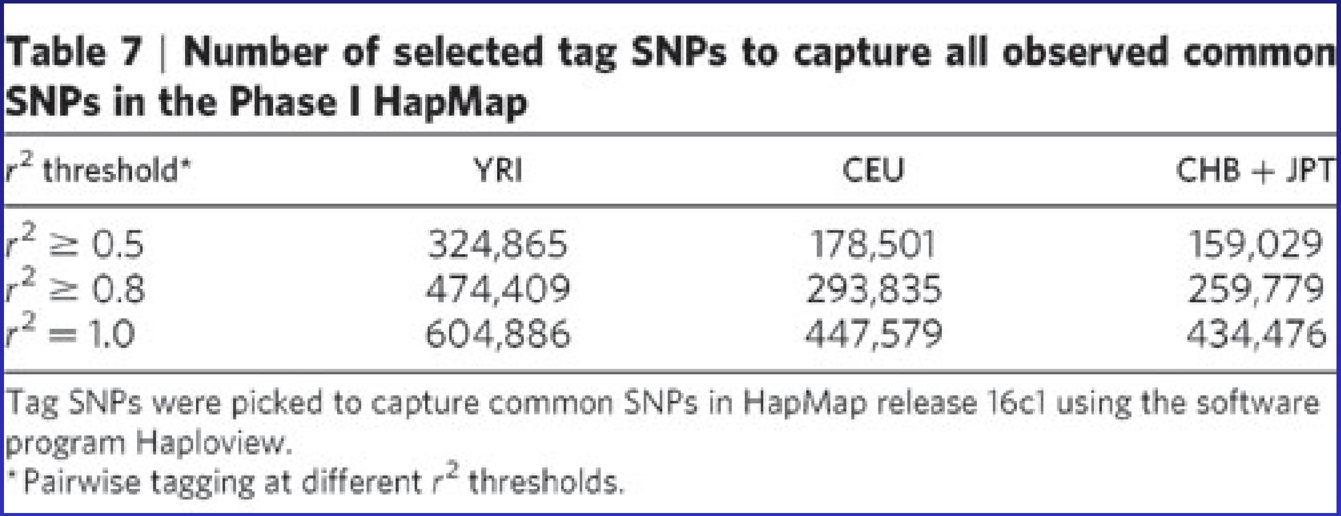


Table 7 | Number of selected tag SNPs to capture all observed common SNPs in the Phase I HapMap

r^2 threshold*	YRI	CEU	CHB + JPT
$r^2 \geq 0.5$	324,865	178,501	159,029
$r^2 \geq 0.8$	474,409	293,835	259,779
$r^2 = 1.0$	604,886	447,579	434,476

Tag SNPs were picked to capture common SNPs in HapMap release 16c1 using the software program Haploview.

* Pairwise tagging at different r^2 thresholds.

HapMap performance

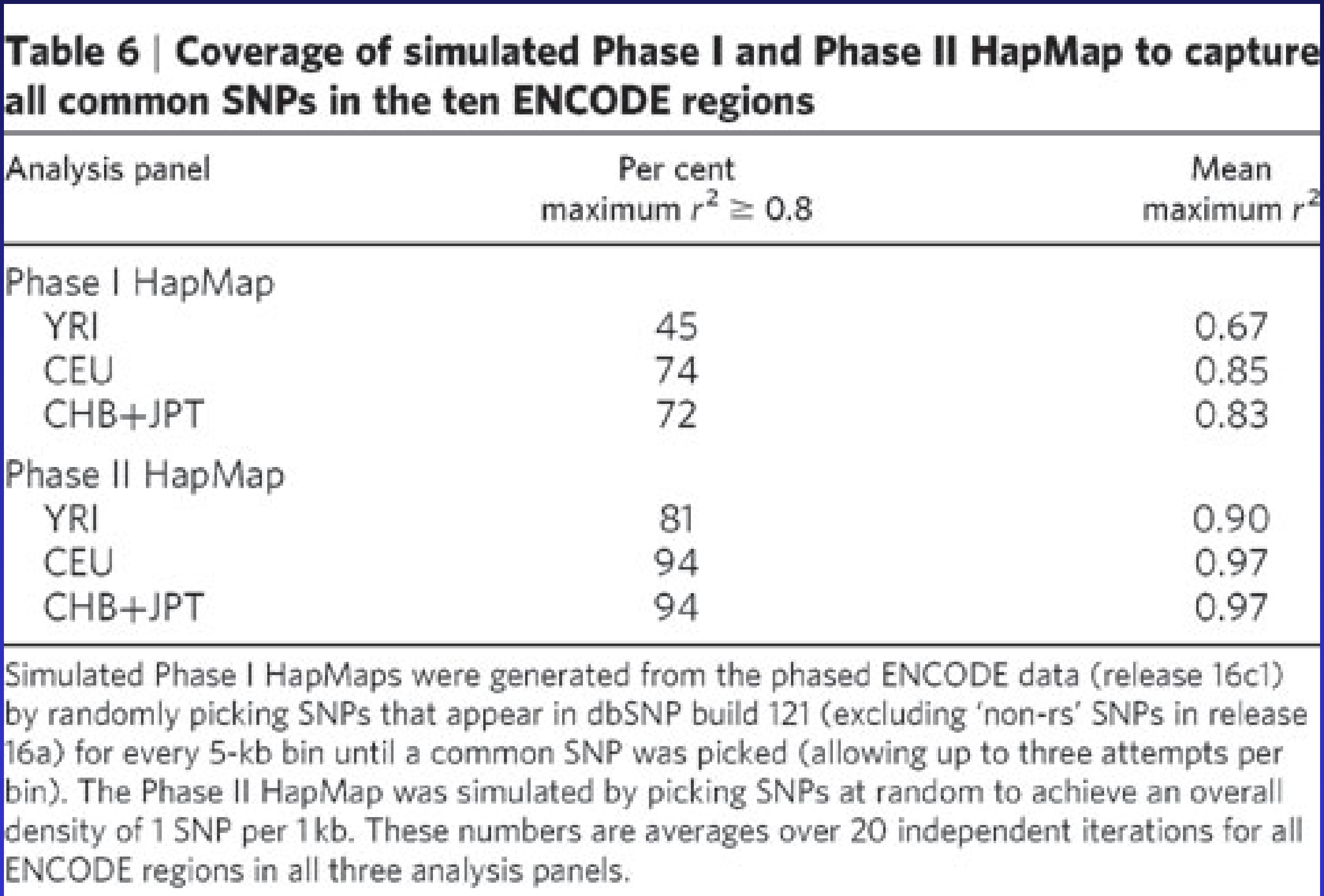
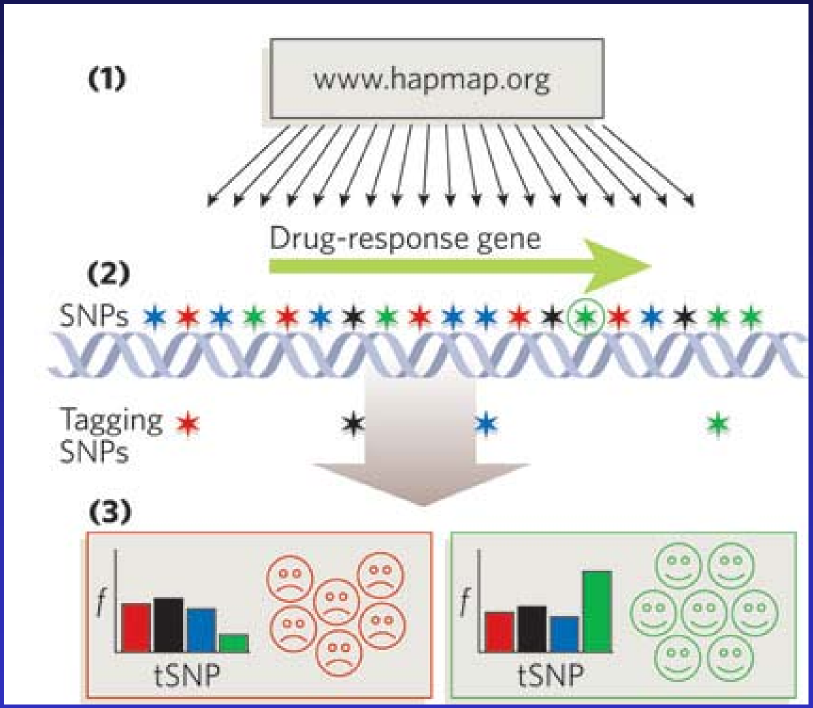


Table 6 | Coverage of simulated Phase I and Phase II HapMap to capture all common SNPs in the ten ENCODE regions

Analysis panel	Per cent maximum $r^2 \geq 0.8$	Mean maximum r^2
Phase I HapMap		
YRI	45	0.67
CEU	74	0.85
CHB+JPT	72	0.83
Phase II HapMap		
YRI	81	0.90
CEU	94	0.97
CHB+JPT	94	0.97

Simulated Phase I HapMaps were generated from the phased ENCODE data (release 16c1) by randomly picking SNPs that appear in dbSNP build 121 (excluding 'non-rs' SNPs in release 16a) for every 5-kb bin until a common SNP was picked (allowing up to three attempts per bin). The Phase II HapMap was simulated by picking SNPs at random to achieve an overall density of 1 SNP per 1 kb. These numbers are averages over 20 independent iterations for all ENCODE regions in all three analysis panels.

Using the HapMap

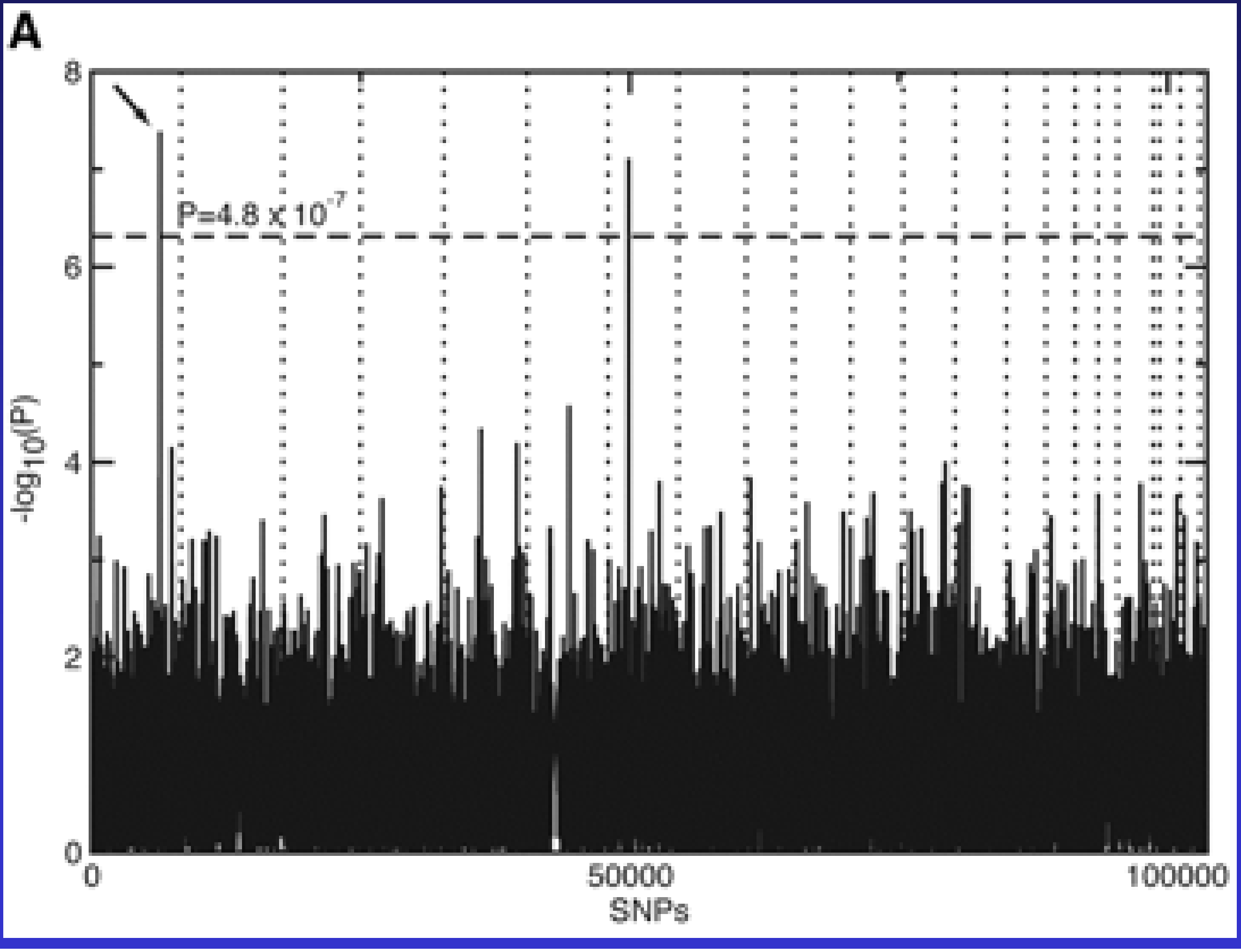


Goldstein & Cavalleri *Nature* 2005

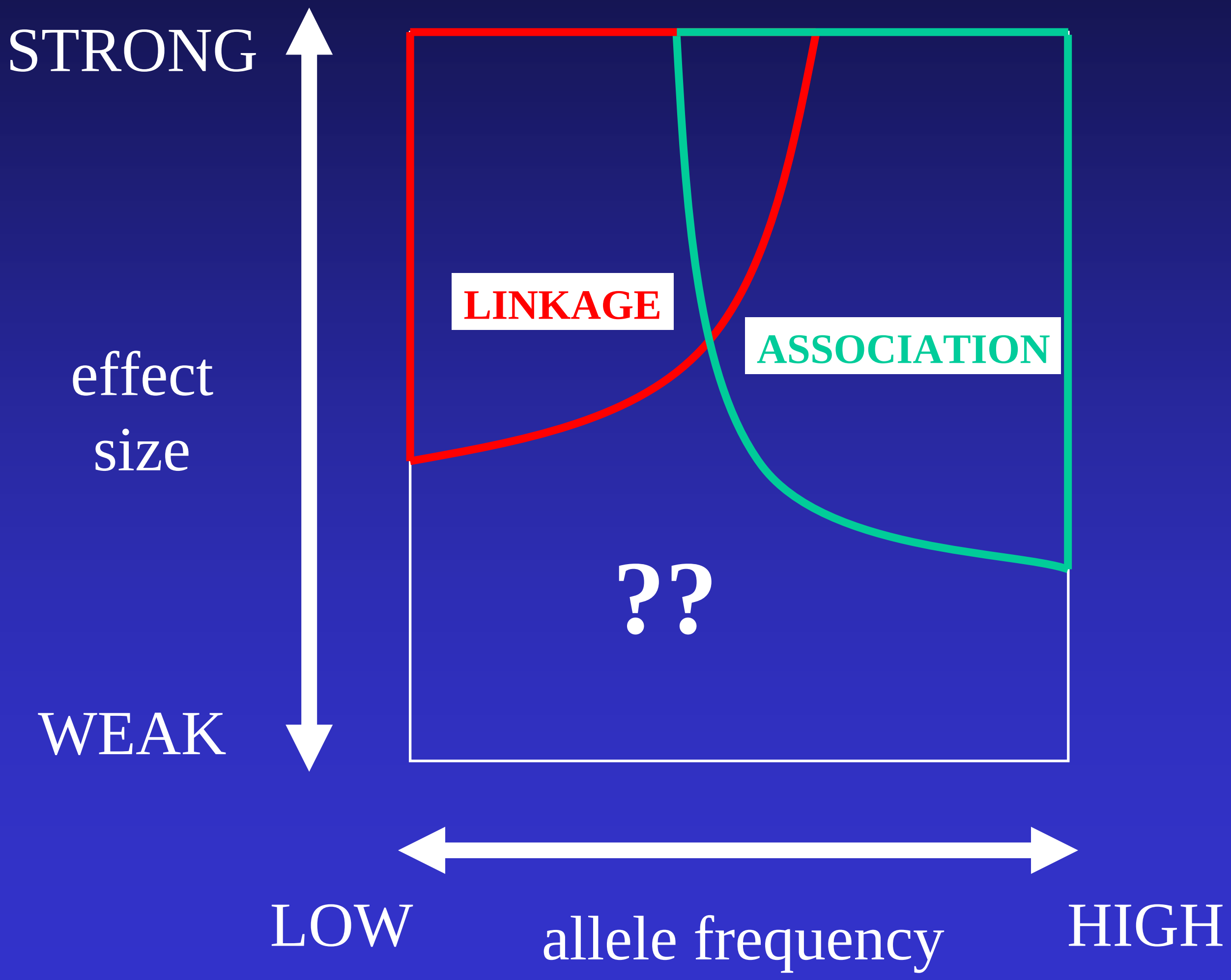
Science 2005 308:385-9

Complement factor H polymorphism in age-related macular degeneration.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J.



Distribution of genetic effects will determine success rate



Next phase: genome re-sequencing

Genome sequencing in microfabricated high-density picolitre reactors.
Margulies, M. et al. *Nature* 437, 376-80 (2005).

Accurate multiplex polony sequencing of an evolved bacterial genome.
Shendure, J. et al. *Science* 309, 1728-32 (2005)

Goal: sequence an individual person’s genome
for under $1000

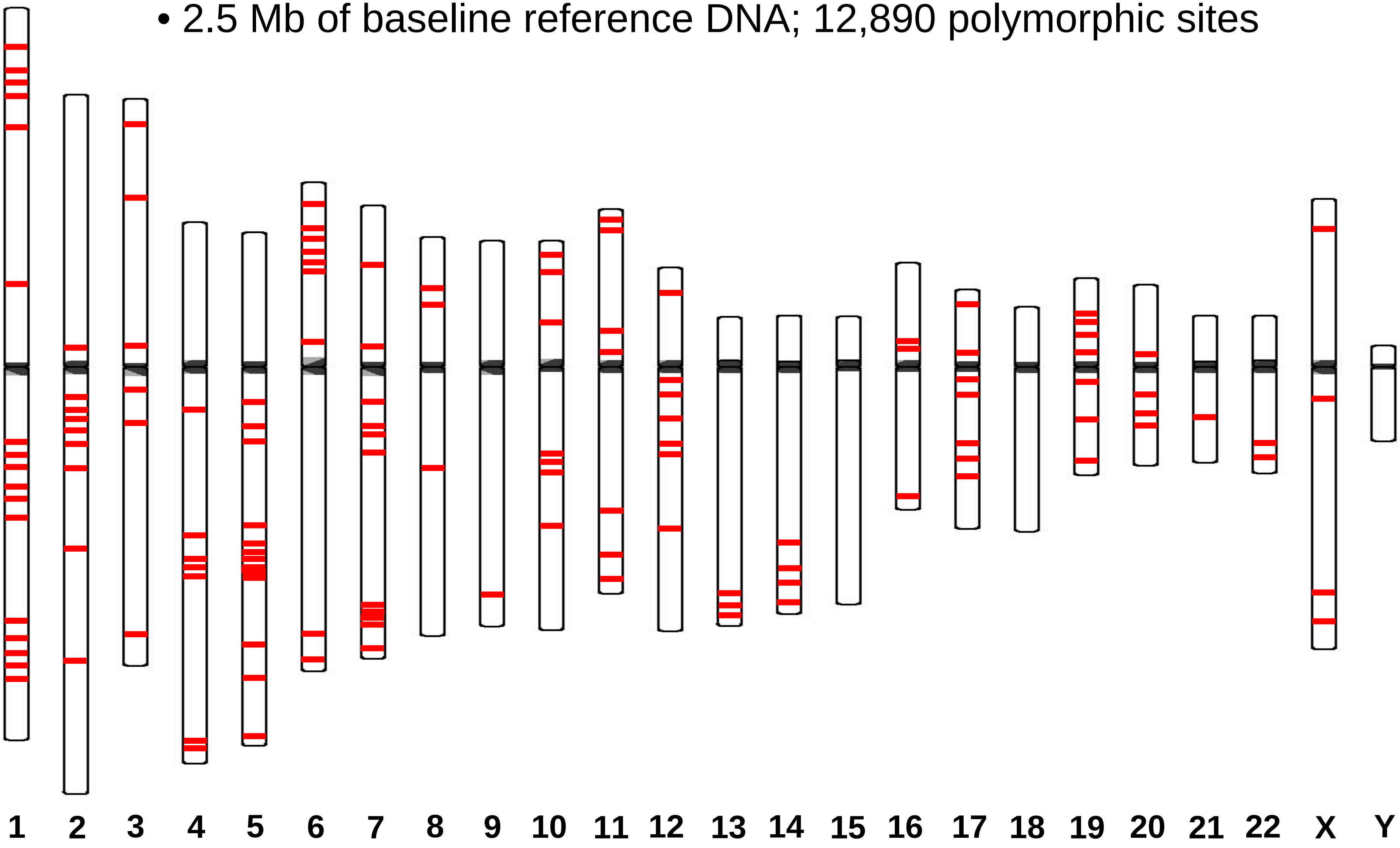
Signatures of selection in human genes

What is the role of selection in human variation?

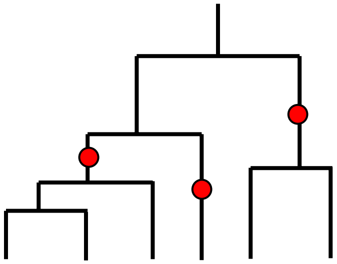
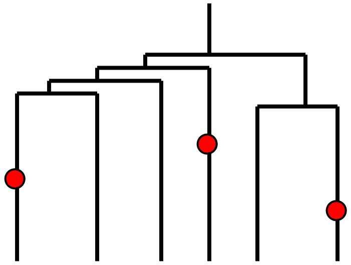
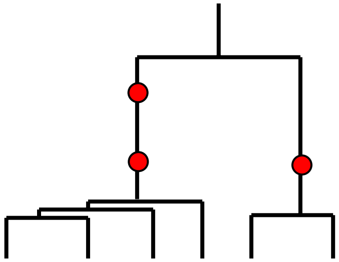
- Neutral theory (Kimura 1968; King & Jukes 1969) provides the null hypothesis
- Can we detect signatures of selection against background variability caused by genetic drift?
- How prevalent are selective events and what can they teach us about human evolution?

SeattleSNPs data

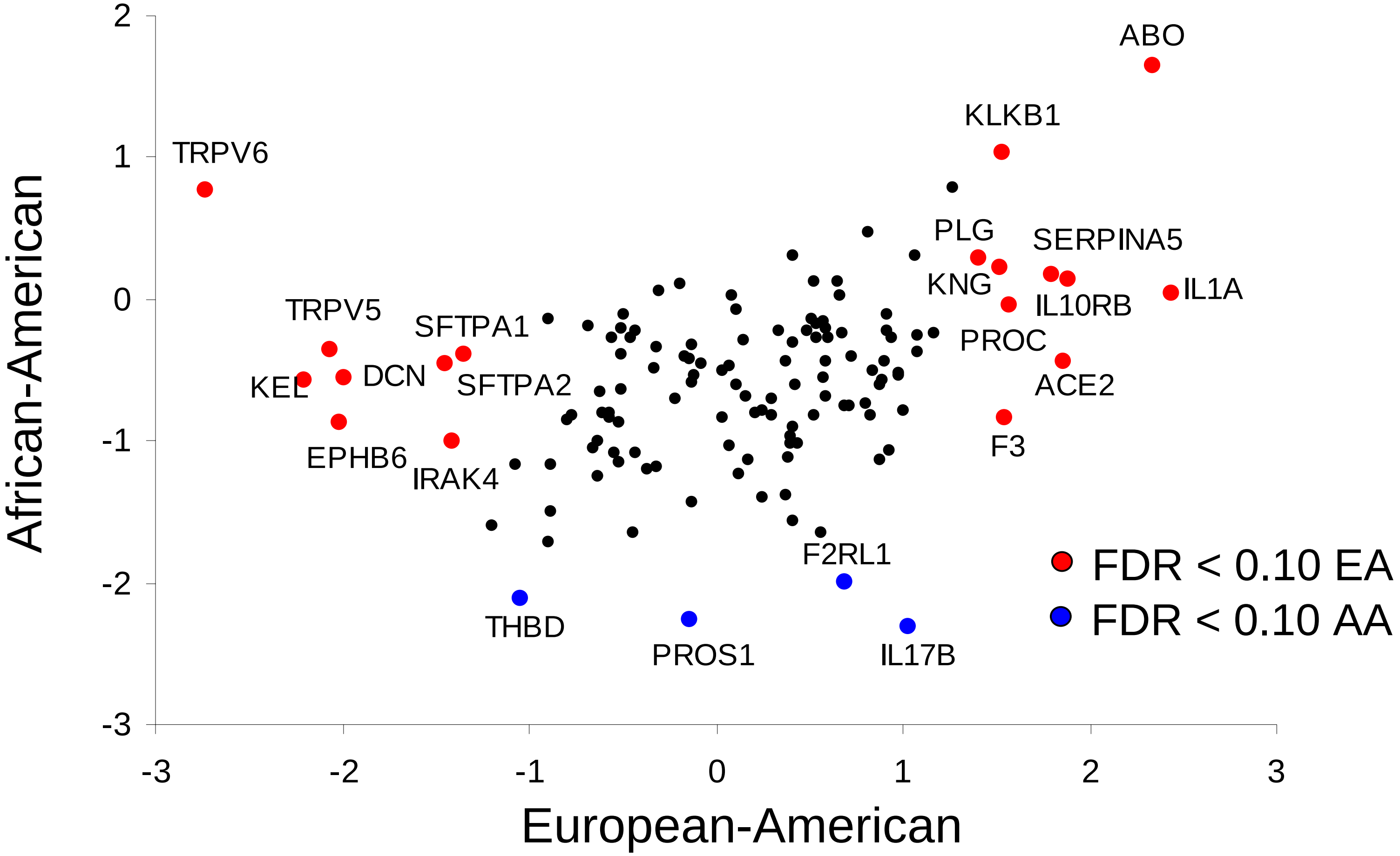
- 132 genes sequenced in 47 individuals from two populations
- 2.5 Mb of baseline reference DNA; 12,890 polymorphic sites



Tests of allele frequency distribution

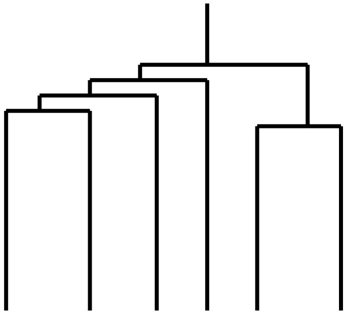
	Neutral	**Positive Selection**	**Balancing Selection**
			
		Excess of Low Frequency Alleles	Excess of Intermediate Frequency Alleles
Tajima's D $\hat{\pi} - \hat{\theta}_W$	D = 0	D < 0	D > 0
Fu and Li's D* $S - \eta_S$	D* = 0	D* < 0	D* > 0
Fay and Wu's H $\hat{\pi} - \overset{\frown}{\theta}_H$	H = 0	H < 0	NA

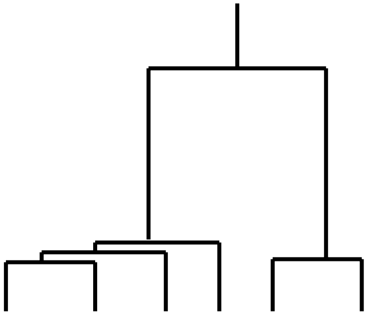
Tajima’s D



	EA	AA
genes with evidence for selection by one or more tests:	37	4

Disentangling selection from demographic history





Selection	Positive	Balancing
Neutral Deviation	Expansion	Bottleneck or Structure
Neutrality Tests	D, D*, F* < 0	D, D*, F* > 0

- Demographic history affects all loci, while selection is locus-specific
- Use empirical distribution over all loci to infer demographic history
- Test whether evidence for selection of specific loci is robust

Models of demographic history

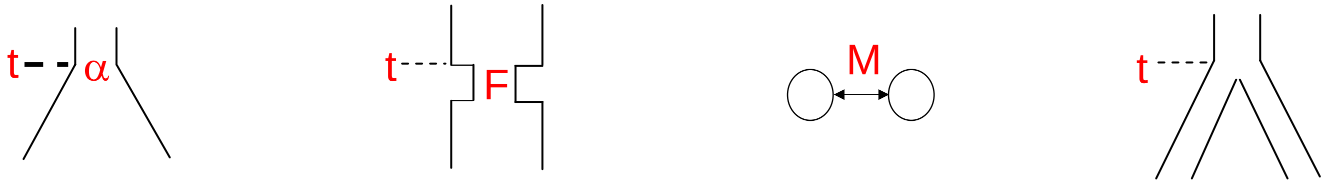
Models

Expansion

Bottleneck

Structure (Island Model)

Structure (Splitting)



Simulations 300 parameter combinations; 10,000 coalescent simulations each

Model Selection For each model identify “best-fit” parameter values

Results

AA	t = 50 Kyr
	$\alpha = 1 \times 10^{-3}$/gen
		t = 100 Kyr
		F = 0.375
			M = 4	t = 70 Kyr

EA	t = 10 Kyr
	$\alpha = 5 \times 10^{-4}$/gen
		t = 40 Kyr
		F = 0.175
			M = 4	t = 70 Kyr

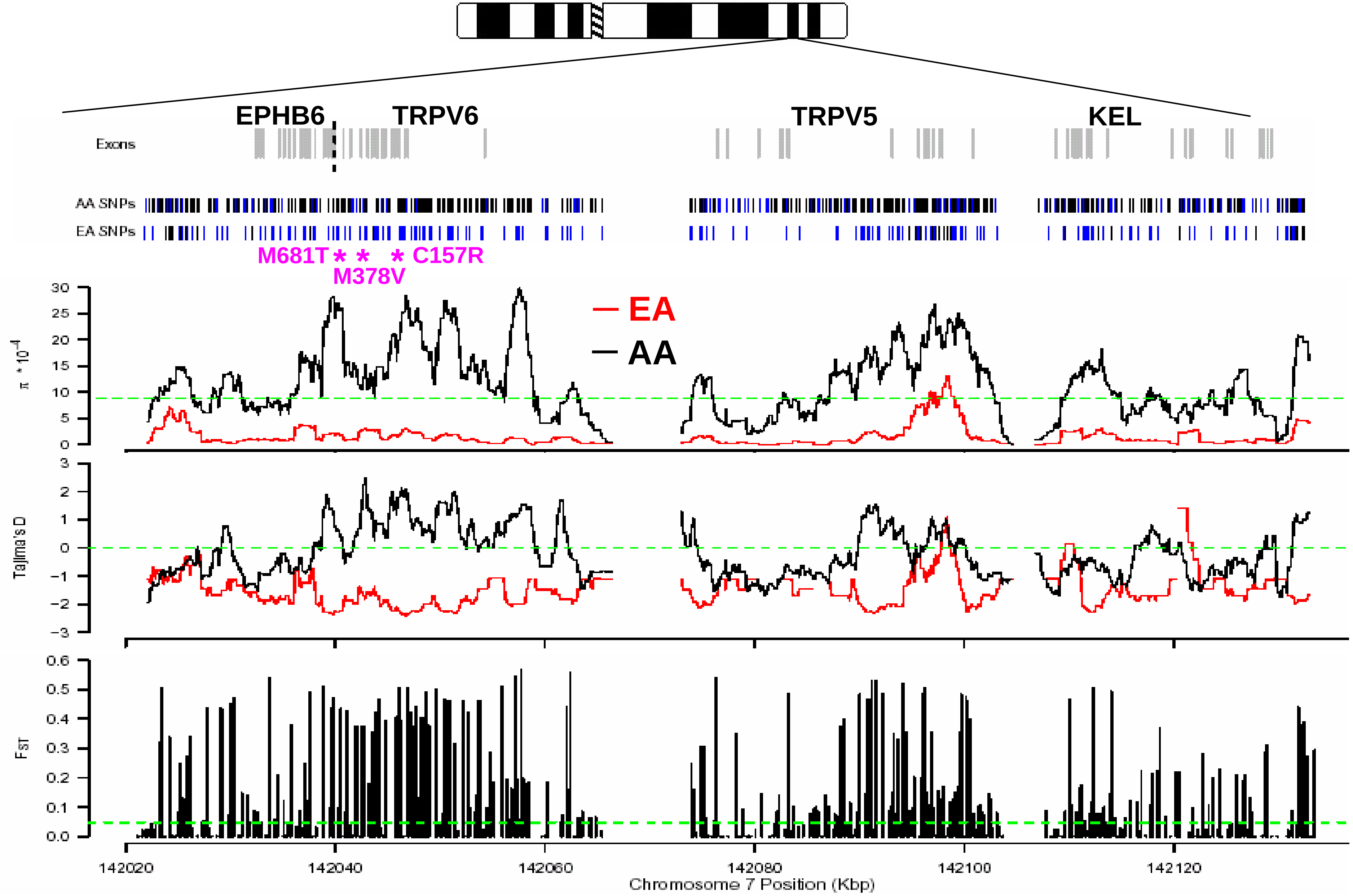
Demographically robust selection genes

Gene	Chromosome	Type of Selection	Panther Process
CYP4A11	1	Positive	Lipid, fatty acid and steroid metabolism
TNFRSF1B	1	Positive	Immunity and defense
IL1A	2	Balancing	Immunity and defense
EPHB6	7	Positive	Signal transduction
KEL	7	Positive	Protein metabolism and modification
TRPV5	7	Positive	Transport
TRPV6	7	Positive	Transport
ABO	9	Balancing	Protein metabolism and modification
IL10RA	11	Positive	Immunity and defense
DCN	12	Positive	Signal transduction
IRAK4	12	Positive	Immunity and defense
VTN	17	Positive	Immunity and defense
CEBPB	20	Positive	Immunity and defense
ACE2	X	Balancing	Protein metabolism and modification
IL24	1	Positive	Immunity and defense
IL17B	5	Positive	Immunity and defense

🟨 **European-American (14/37)**

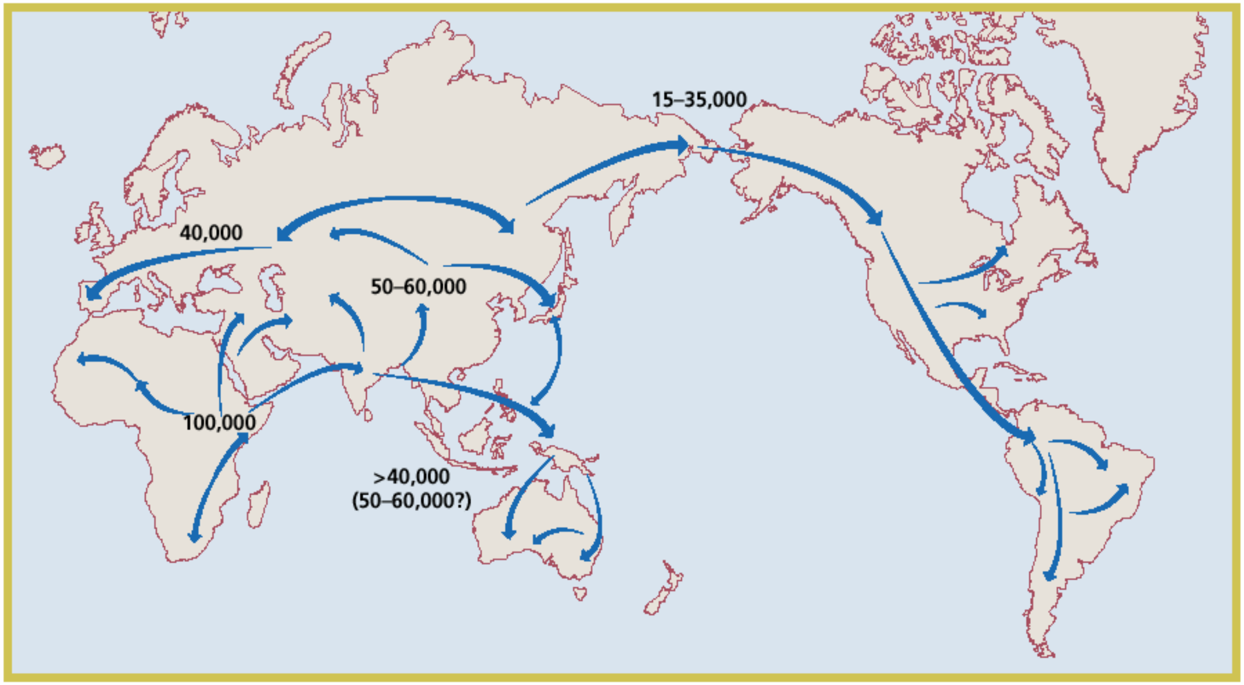
🟦 **African-American (2/4)**

Chromosome 7q: Recent selective sweep



Implications: Recent human evolution

- Signatures of selection are population-specific
- Many more signatures in European-derived sample
- Examples of local adaptation?



BOB CRIMI

Selective Pressures:

- Climate
- Dietary
- Pathogens
- Cultural