Population Genetics: Practical Applications

Lynn B. Jorde
Department of Human Genetics
University of Utah School of Medicine

Overview

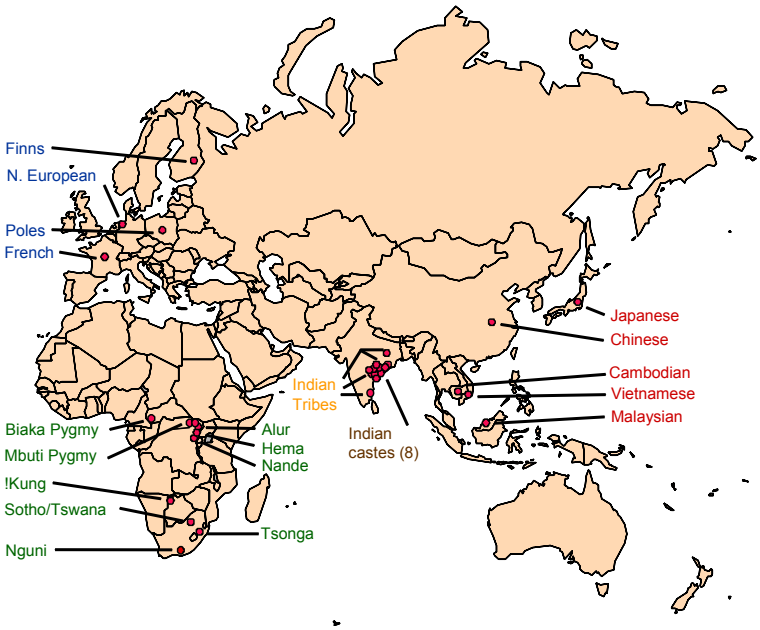
- Patterns of human genetic variation
  - Among populations
  - Among individuals
- "Race" and its biomedical implications
- Linkage disequilibrium, the HapMap, and the search for complex disease genes

Mutation and Genetic Variation

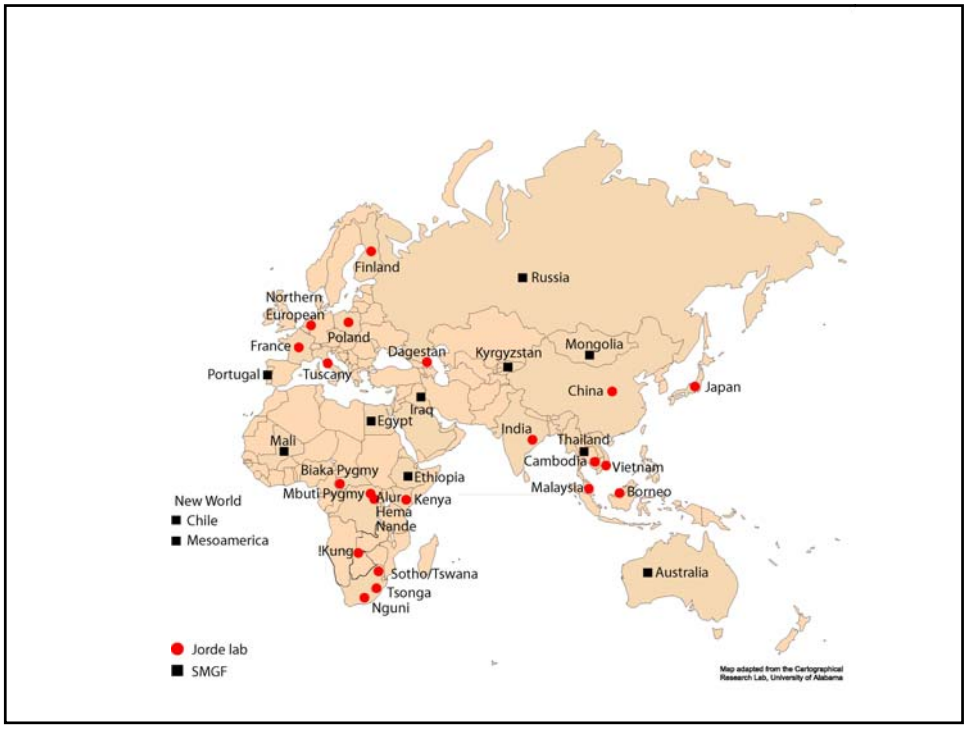
Mutation rate is 2.5 x 10-8 per bp per generation: we transmit 75-100 new DNA variants with each gamete

"The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music."

- Lewis Thomas



> 250 noncoding loci: *Alu,* LINE1, STR, restriction site polymorphisms, mitochondrial DNA



Allele frequencies in populations

Population	SNP 1	SNP 2	SNP 3
1	0.588	0.890	0.880
2	0.671	0.559	0.528
3	0.792	0.790	0.828

1/1000 bp varies between a pair of individuals: how is this variation distributed between continents?

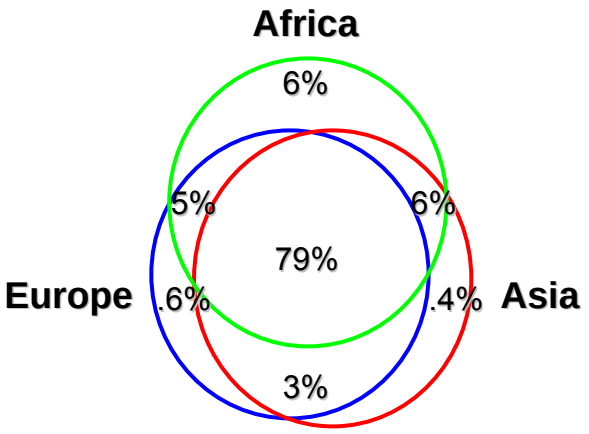
$$F_{ST} = \sum_{i}^{N} \frac{(p_{ik} - \bar{p}_k)^2}{2\bar{p}_k(1-\bar{p}_k)} \Big/ N = \frac{H_T - \bar{H}_S}{H_T}$$

	60 STRPs	*30 RSPs*	*100 Alus*	*75 L1s*
Between individuals, within continents	90%	87%	86%	88%
Between continents (F_{ST})	10%	13%	14%	12%

Jorde et al., 2000, *Am. J. Hum. Genet.* 66: 979-88

Most genetic variants are shared among populations:

7,742 SNPs >.05 in ENCODE database



A simple genetic distance measure

$$D_{ij} = |p_i - p_j|$$

D_{ij} is the genetic distance between populations i and j; p_i and p_j are the allele frequencies of a SNP in populations i and j.

Pop.	SNP 1	SNP 2	SNP 3
1	0.588	0.890	0.880
2	0.671	0.559	0.528
3	0.792	0.790	0.828

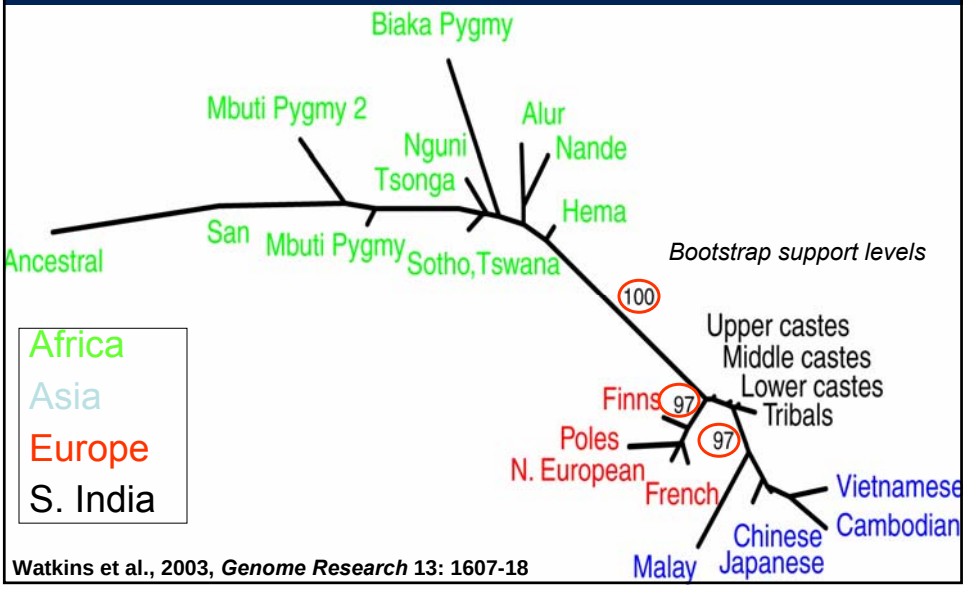
$D_{12} = |0.588 – 0.671| = 0.083$ (avg. over all SNPs)

Building a population network

Pop.	SNP 1
1	0.588
2	0.671
3	0.792

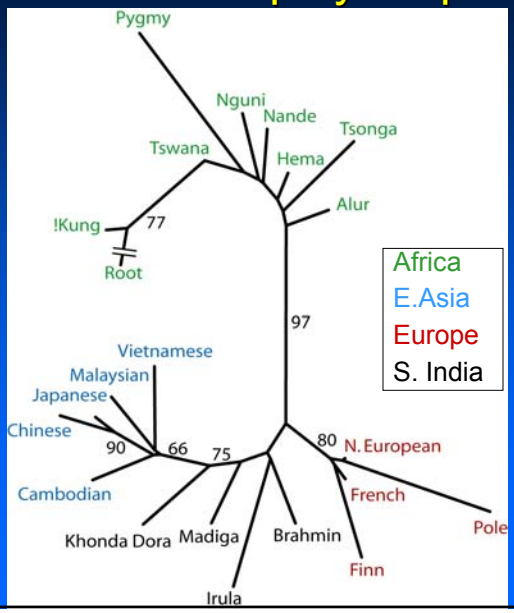
$|p_1 – p_2|$ $| p_3 – (p_1 + p_2)/2 |$

Genetic relationships based on 100 autosomal *Alu* polymorphisms



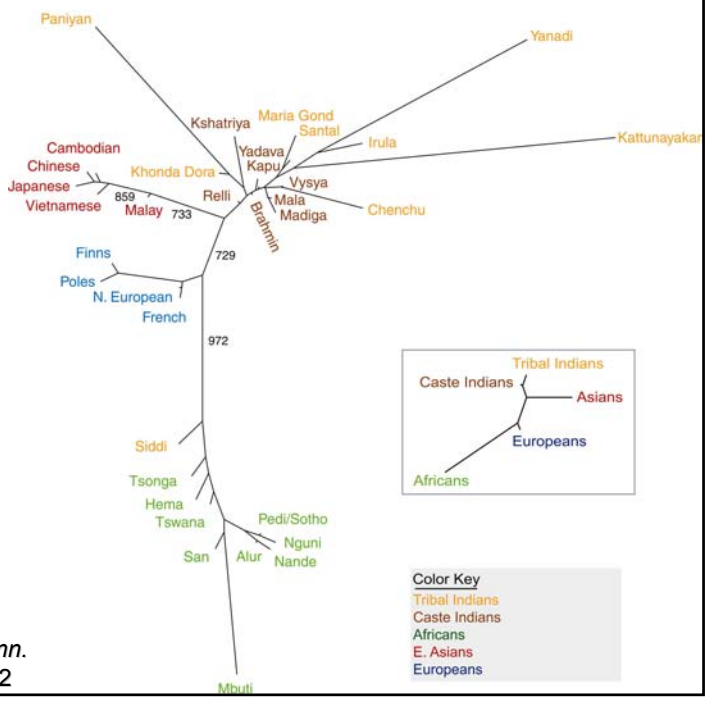
Watkins et al., 2003, *Genome Research* 13: 1607-18

Genetic relationships based on 75 autosomal L1 polymorphisms



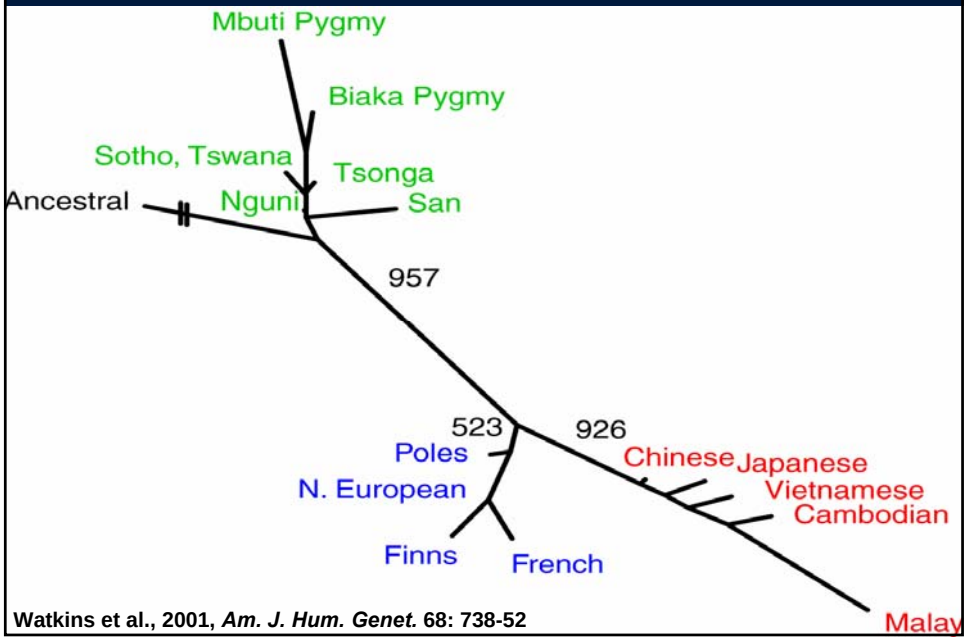
Witherspoon et al., 2006, *Hum. Hered.*, 62: 30-46

45 Autosomal STRs



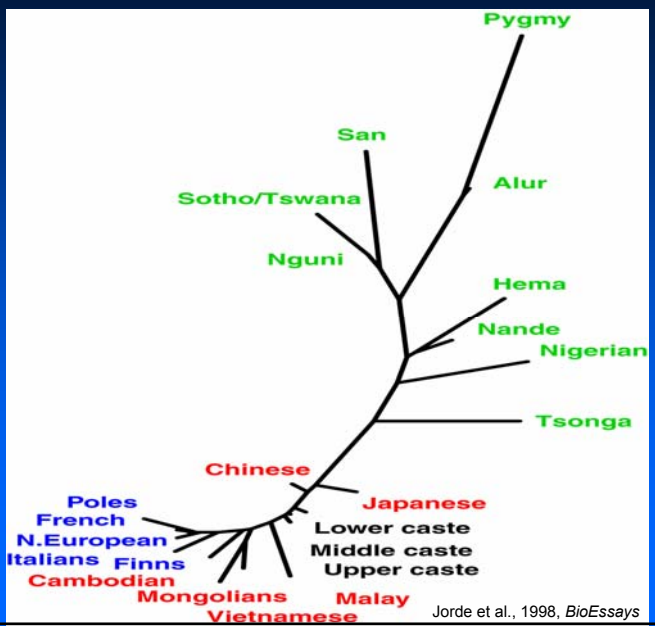
Watkins et al., 2005, *Ann. Hum. Genet.* 69: 680-92

Rooted RSP Tree (30 loci)



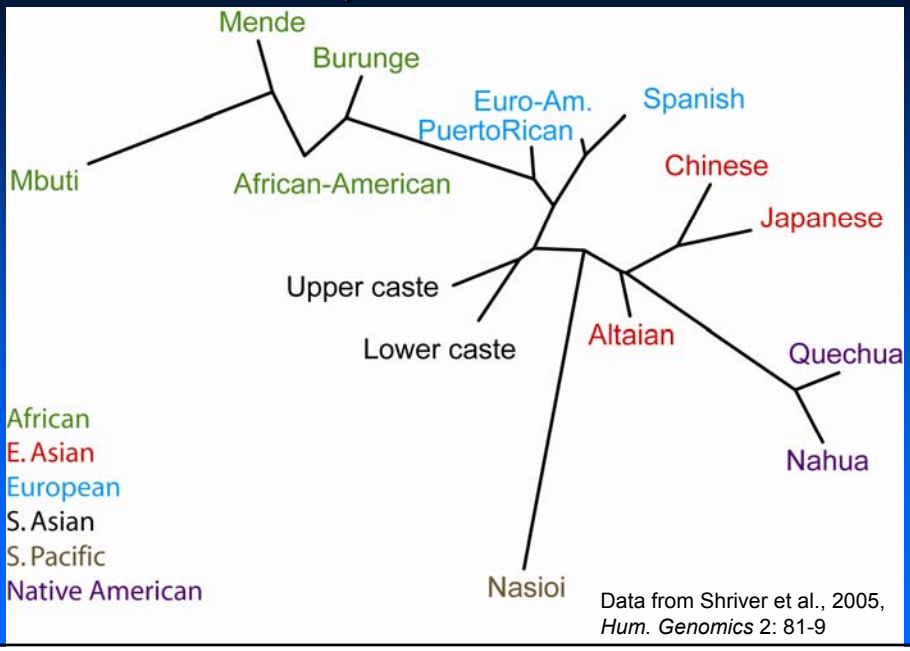
Watkins et al., 2001, *Am. J. Hum. Genet.* 68: 738-52

Mitochondrial DNA (HVS1)

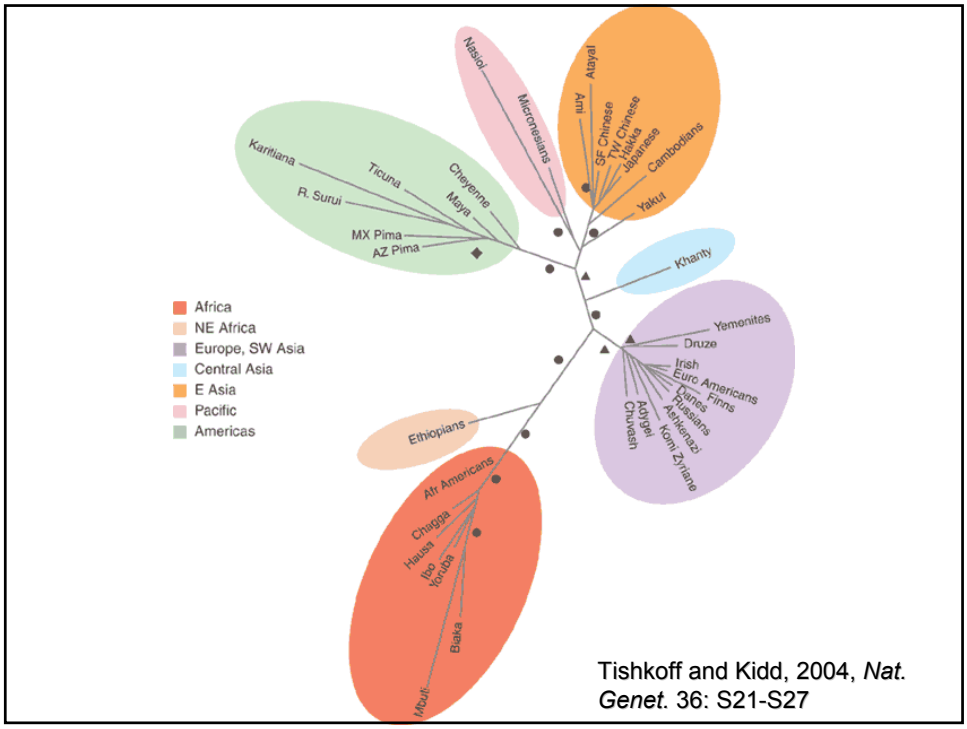


Jorde et al., 1998, *BioEssays*

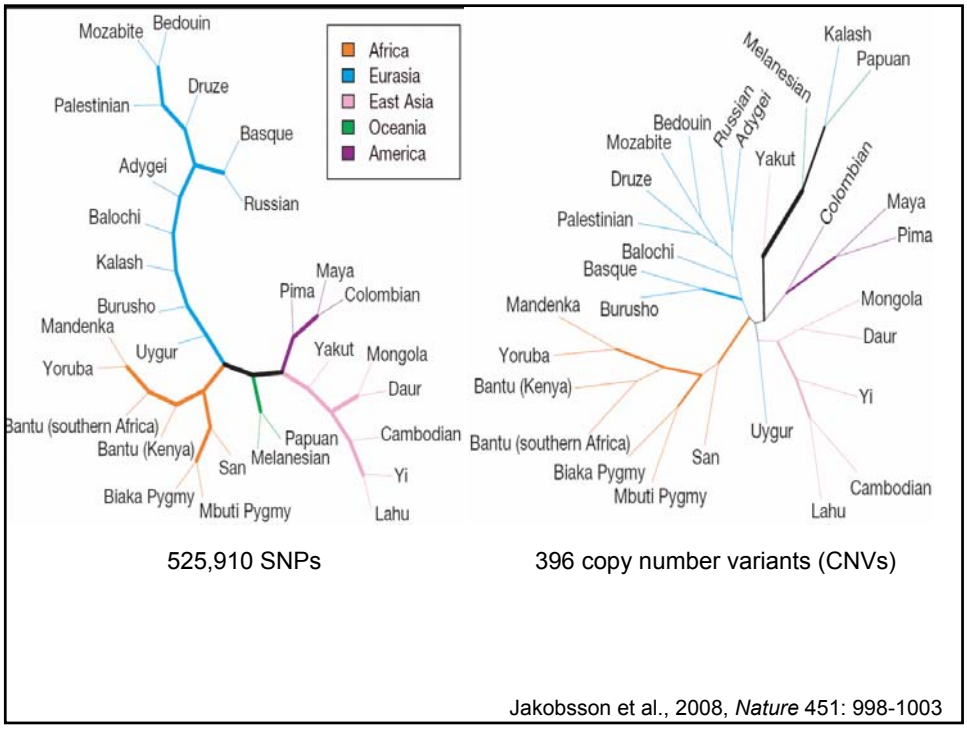
11,078 SNPs



Data from Shriver et al., 2005, *Hum. Genomics* 2: 81-9



Tishkoff and Kidd, 2004, *Nat. Genet.* 36: S21-S27

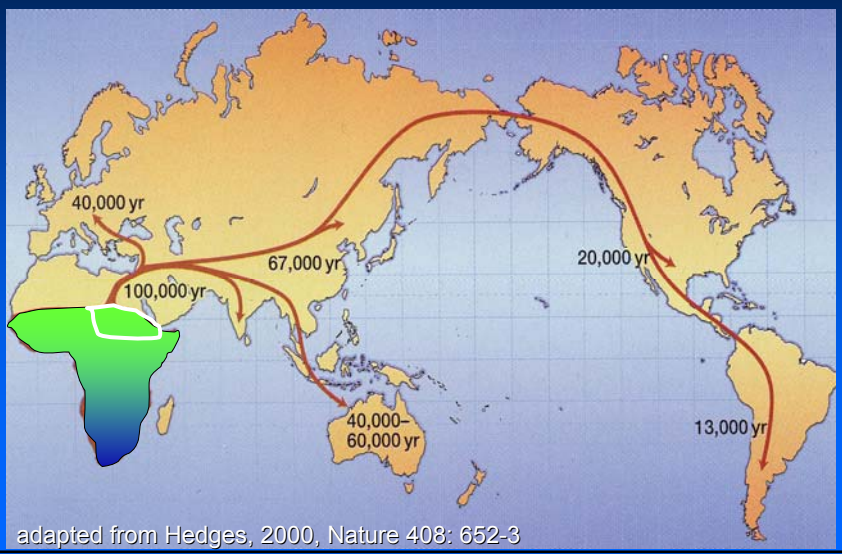


525,910 SNPs

396 copy number variants (CNVs)

Jakobsson et al., 2008, *Nature* 451: 998-1003

Recent African origin of anatomically modern humans



adapted from Hedges, 2000, Nature 408: 652-3

"Race" and genetic variation among individuals (and why does race matter?)

- Prevalence of many diseases varies by population (hypertension, prostate cancer)
- Some common disease-predisposing variants vary among populations
  - Clotting Factor V Leiden variant: 5% of Europeans, < 1% of Africans and Asians
- Responses to some drugs may vary among populations
  - African-Americans may be, on average, less responsive to ACE inhibitors, beta-blockers for lowering blood pressure
- Race is commonly used to design forensic databases (e.g., "Caucasian", African-American, Hispanic)

Recent comments on race

"'Race' is biologically meaningless"
-- Schwartz, 2001, *N. Engl. J. Med.*

"I am a racially profiling doctor"
-- Satel, May 5, 2002, *New York Times*

"These [genetic] data also show that any two individuals within a particular population are as different genetically as any two people selected from any two populations in the world."
-- American Anthropological Association, 1997

Tabulation of DNA sequence differences among individuals

TTGCAGCTCTCC
TTGCAGCTCTCC

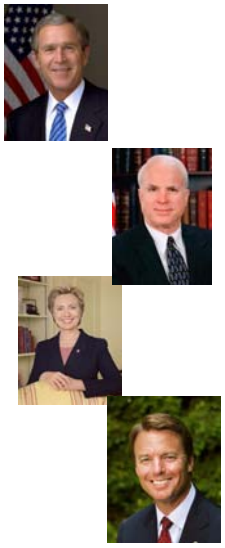
TTGCAGCTCTCC
ATGCAGCTCTCG

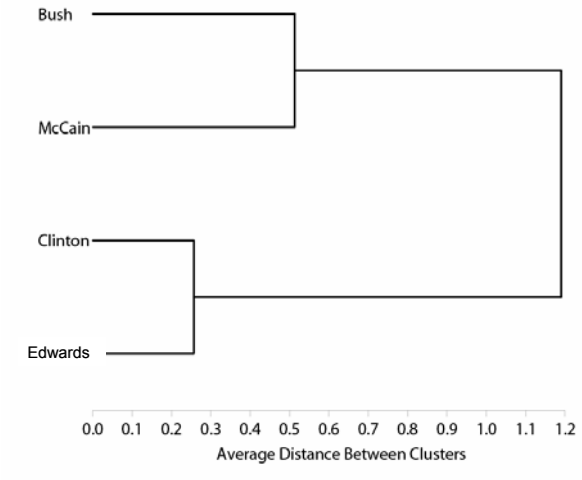
ATGCAGCTCTCG
ATGCTGCTCTCG

ATGCTGCTCTCG
ATGCTGCTCTCG

	Bush	McCain	Clinton	Edwards
Bush	0	.	.	.
McCain	2	0	.	.
Clinton	5	3	0	.
Edwards	6	4	1	0

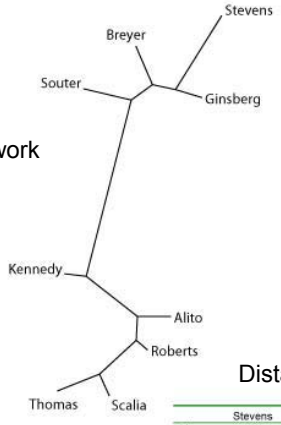
DNA differences can be summarized in a "tree"





A distance matrix based on Supreme Court decisions

Neighbor-joining network



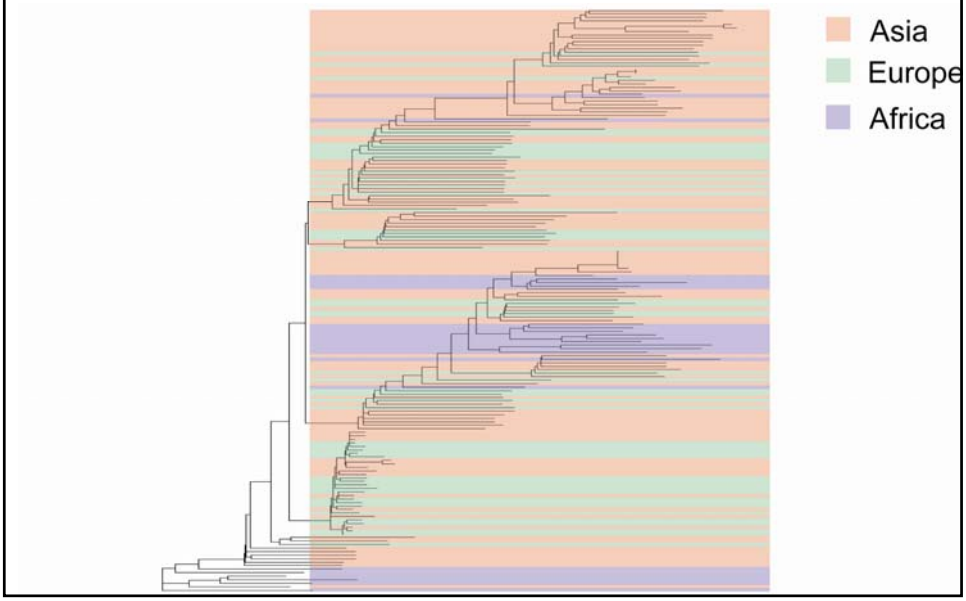
Distance matrix: % disagreement

	Stevens	Ginsberg	Souter	Breyer	Kennedy	Alito	Roberts	Scalia	Thomas
Stevens	0								
Ginsberg	15	0							
Souter	26	15	0						
Breyer	19	13	15	0					
Kennedy	45	36	34	35	0				
Alito	56	48	44	45	13	0			
Roberts	55	49	40	48	19	8	0		
Scalia	59	52	50	58	28	19	11	0	
Thomas	64	55	53	60	29	21	15	9	0

Thanks to: Steve Guthery, MD

Individual network: 14 kb sequence in angiotensinogen gene

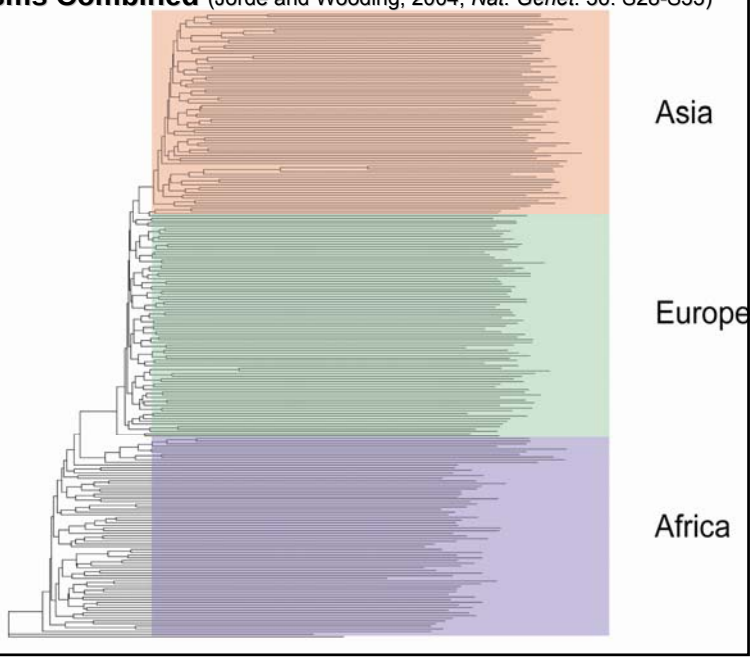
Jorde and Wooding, 2004, *Nat. Genet.*, 36: S28-S33



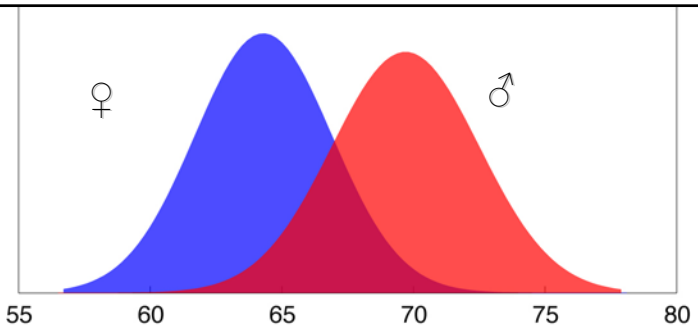
"It may be doubted whether any character can be named which is distinctive of a race and is constant."

-- Charles Darwin, 1871, *The Descent of Man, and Selection in Relation to Sex*

Individual Network: 190 *Alu*, STR, and Restriction Site Polymorphisms Combined (Jorde and Wooding, 2004, *Nat. Genet.* 36: S28-S33)

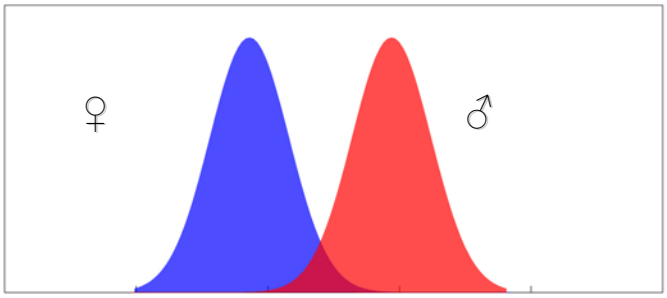


Height

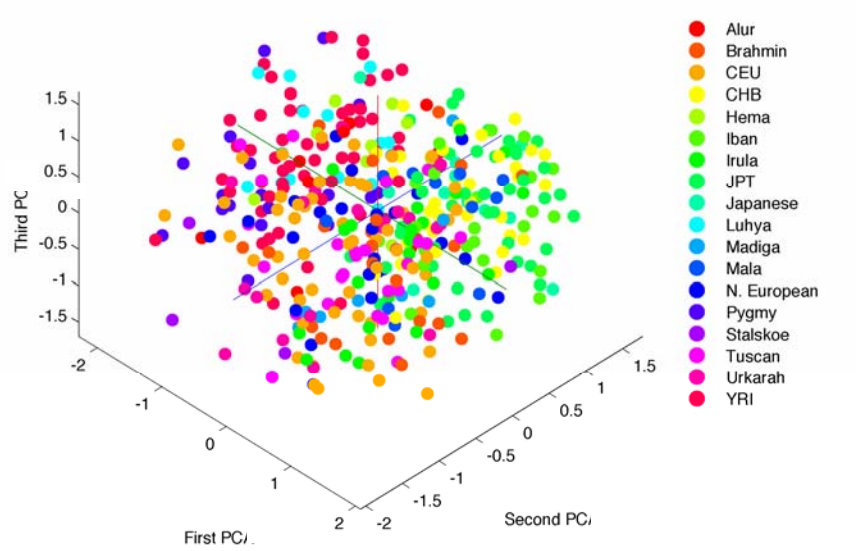


Height +

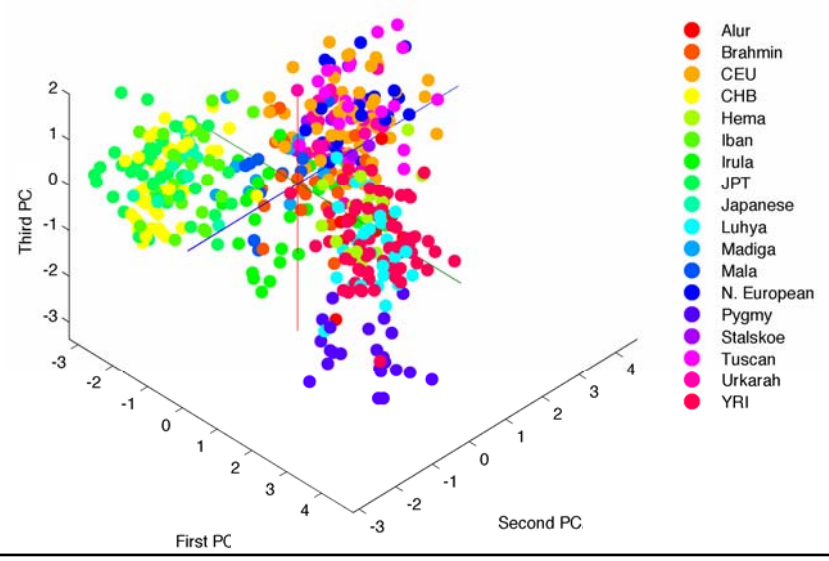
waist/hip ratio



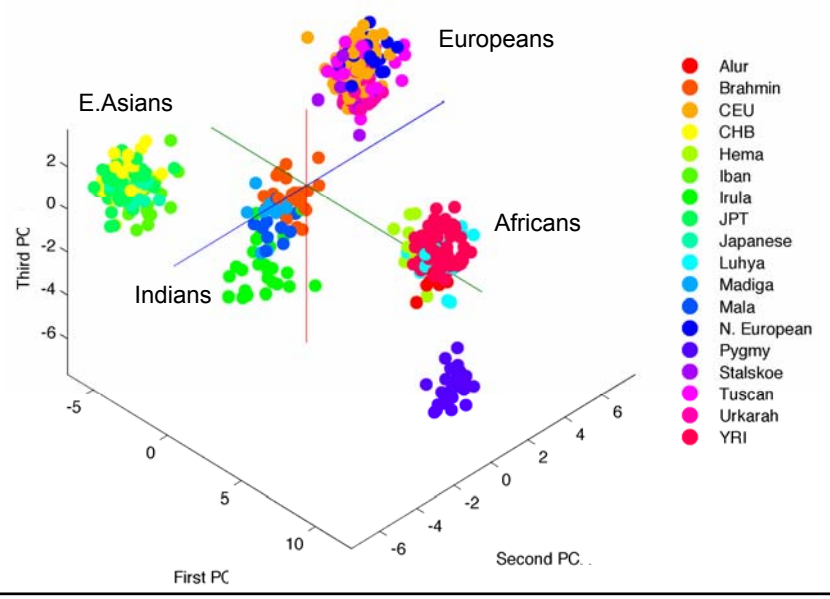
Genetic distances (principal components analysis) among 467 individuals: 10 SNPs



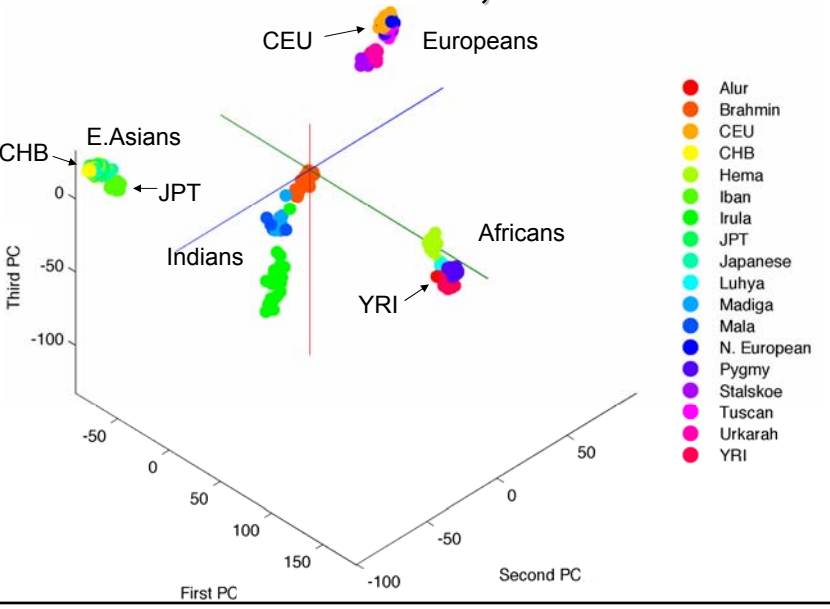
Genetic distances among 467 individuals: 100 SNPs



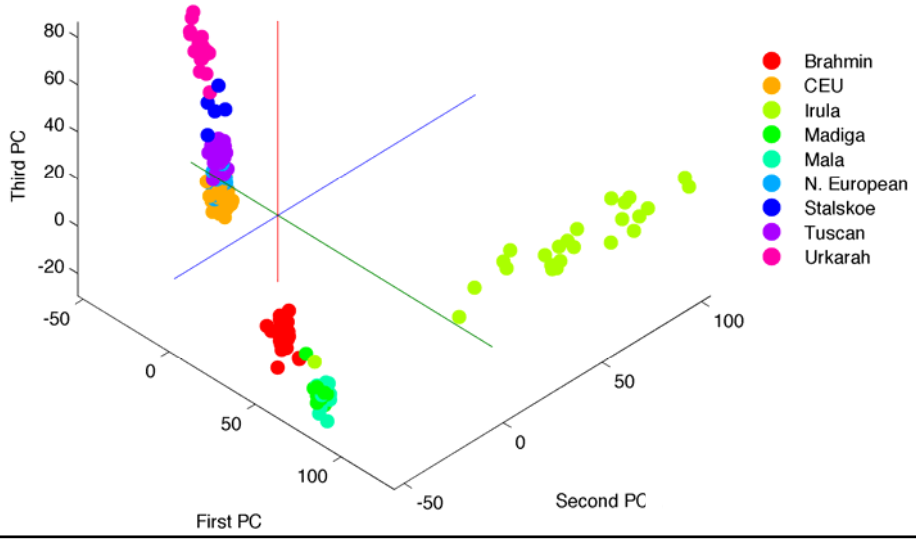
Genetic distances among 467 individuals: 1000 SNPs



Genetic distances among 467 individuals: 261,000 SNPs

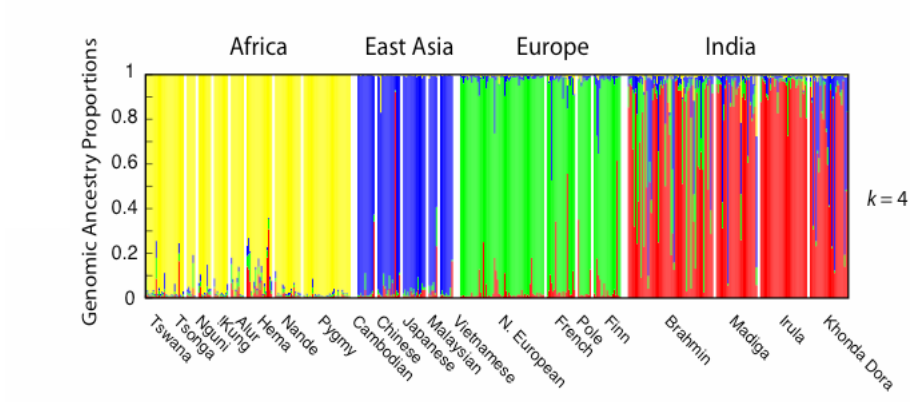


Genetic distance analysis: 205 individuals from Europe and India



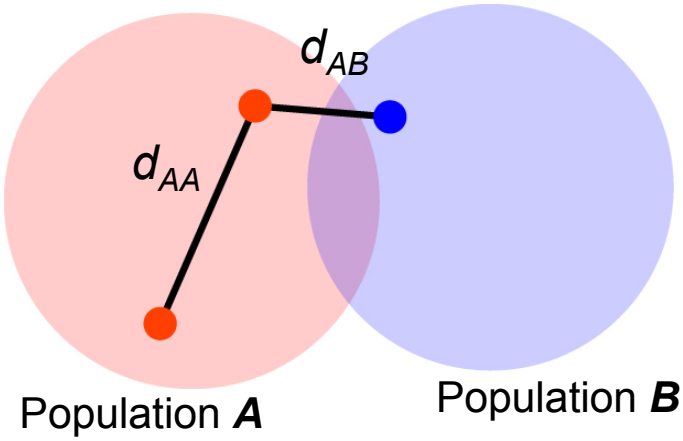
STRUCTURE results: ancestral profiles

Individuals are moved randomly among groups to define k populations in which Hardy-Weinberg and linkage disequilibrium are minimized



Witherspoon et al., 2006, *Hum. Hered.* 62: 30-46

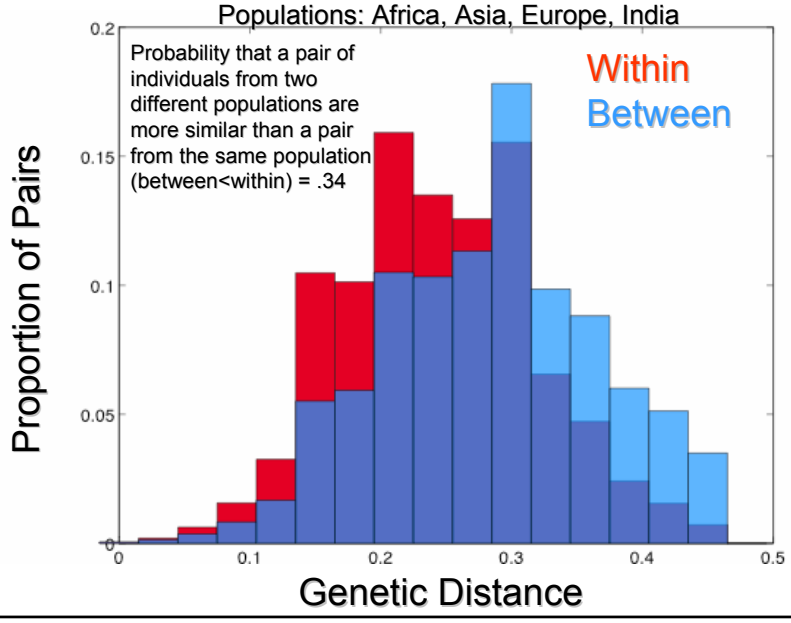
How often are two people from the *same* population **genetically more different** than two people from *different* populations?



Witherspoon et al., 2007, *Genetics* 176: 351-9

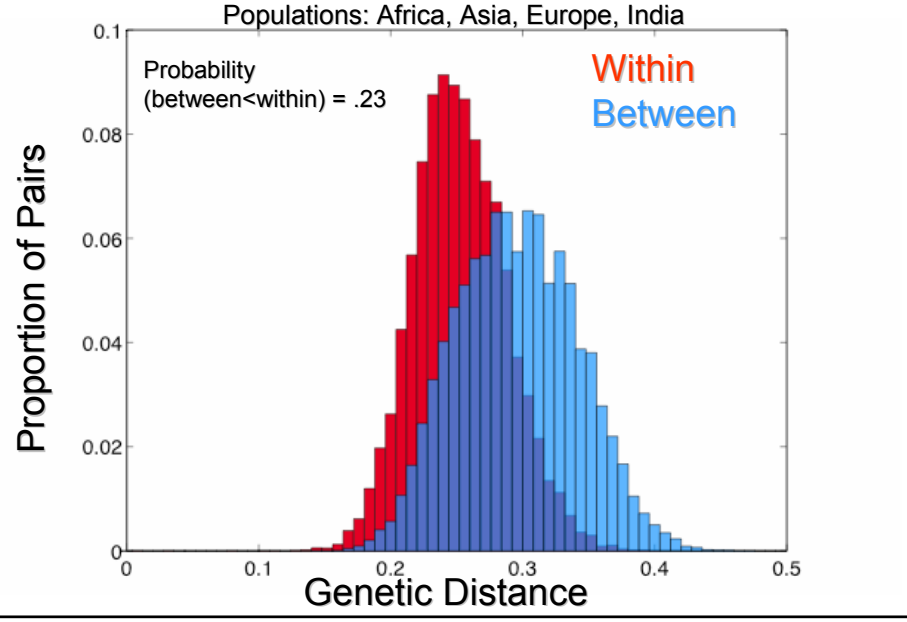
Distribution of individual genetic distances, within and between populations (20 *Alu*s)

Populations: Africa, Asia, Europe, India



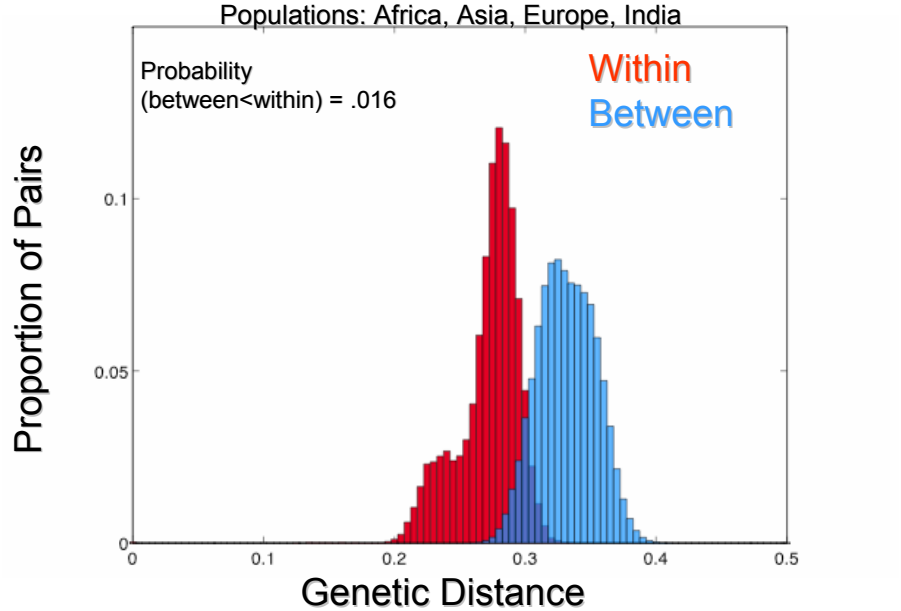
Distribution of individual genetic distances, within and between populations (100 *Alu*s)

Populations: Africa, Asia, Europe, India

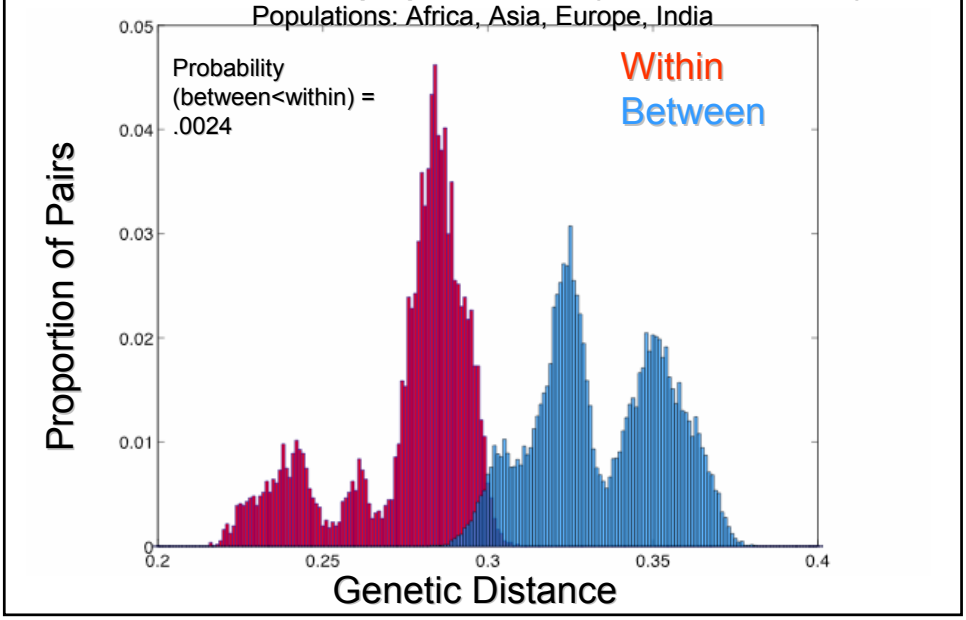


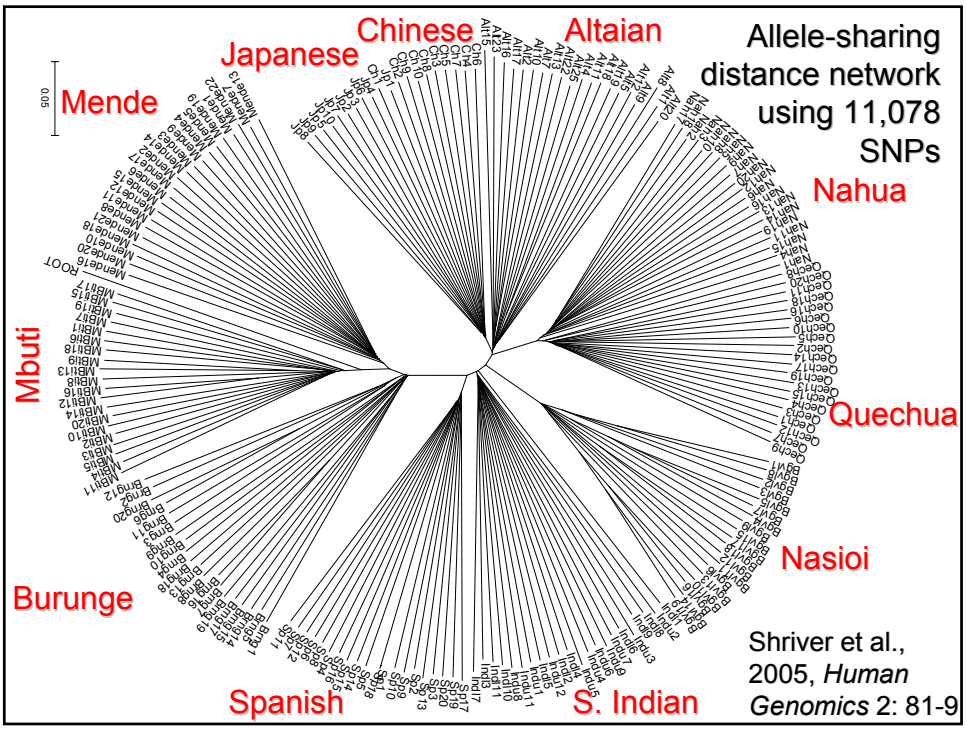
Distribution of individual genetic distances, within and between populations (1000 SNPs)

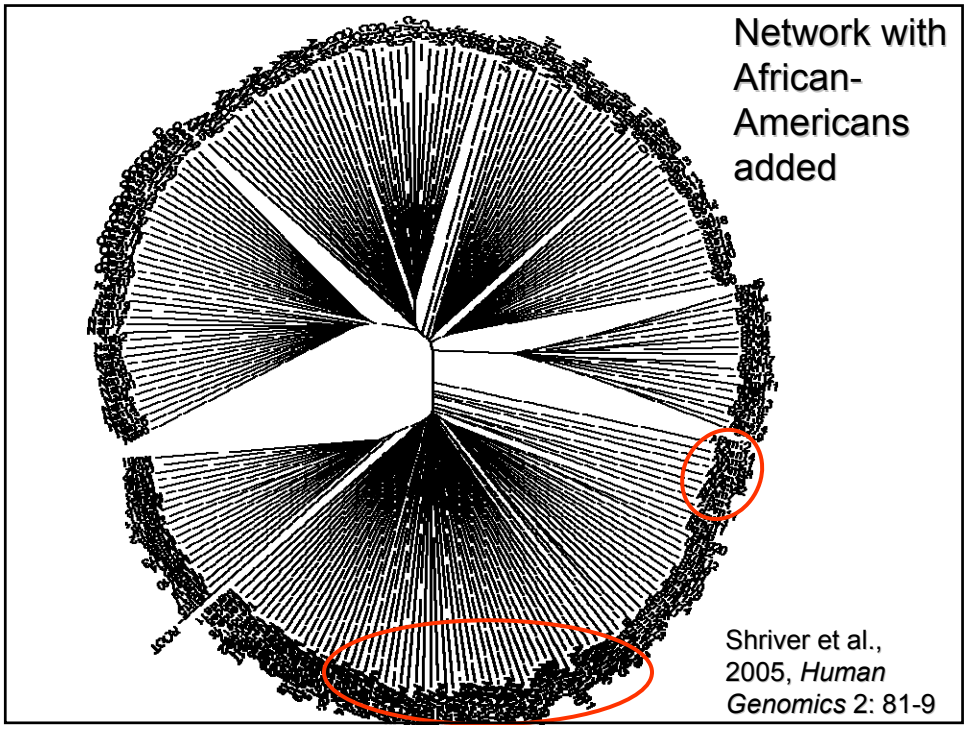
Populations: Africa, Asia, Europe, India

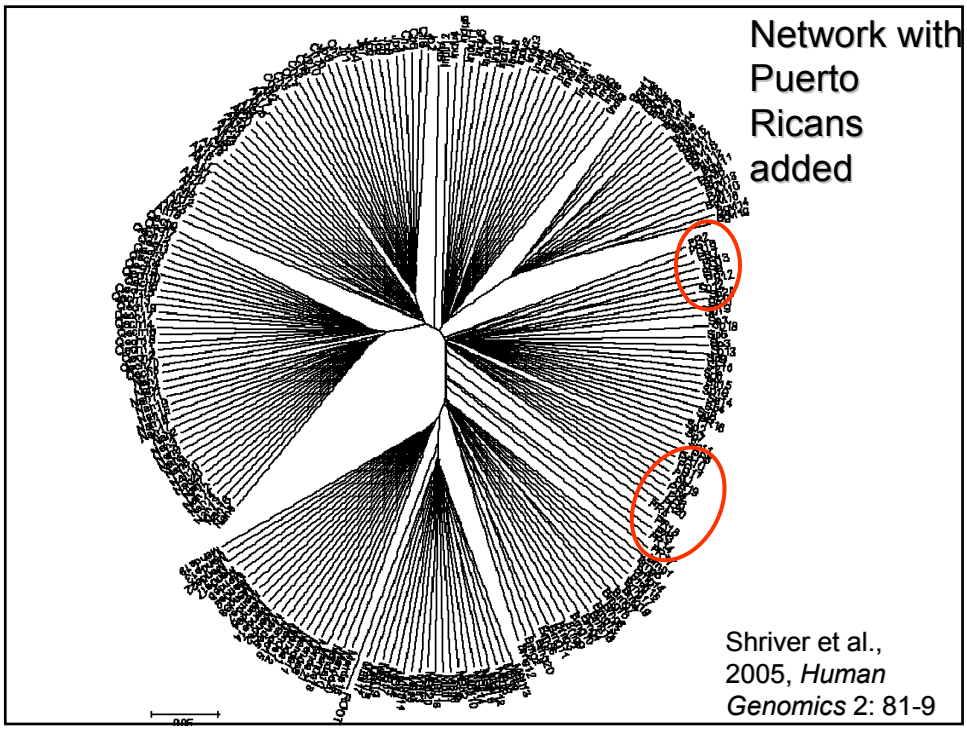


Distribution of individual genetic distances, within and between populations (11,555 SNPs)

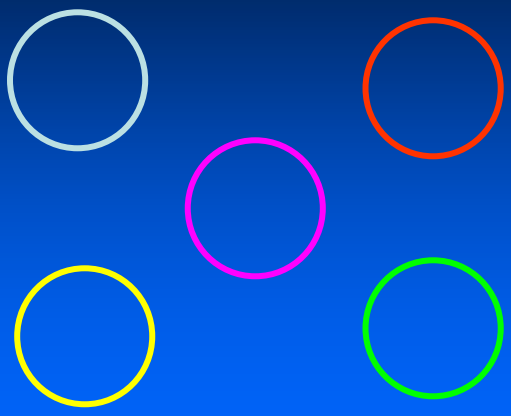





Network with African-Americans added
Shriver et al., 2005, *Human Genomics* 2: 81-9

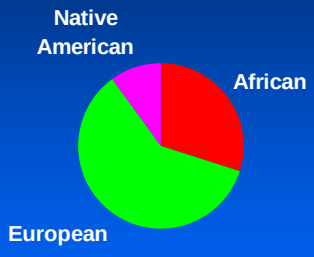

Network with Puerto Ricans added
Shriver et al., 2005, *Human Genomics* 2: 81-9

The Fallacy of Typological Thinking



Ancestry vs. Race

“African-American”



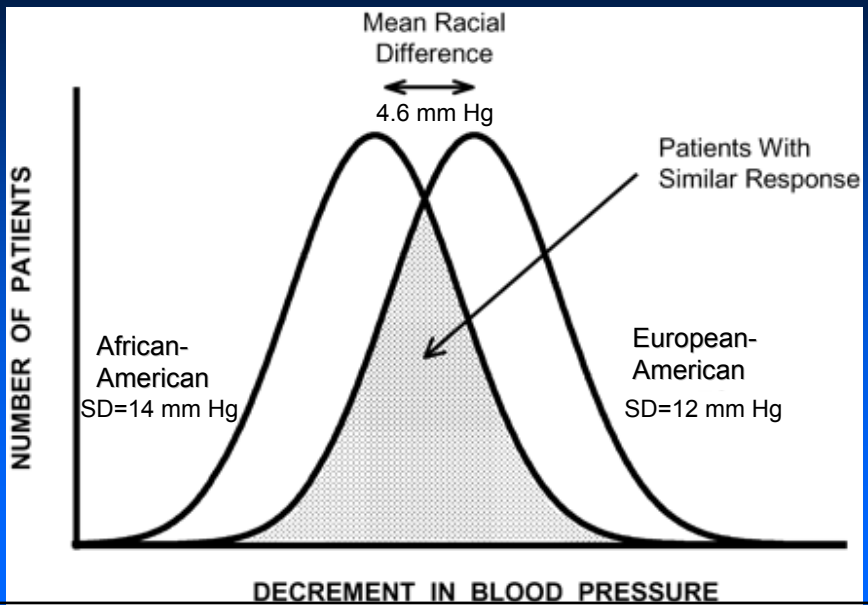
“African-American”

What do these findings imply for biomedicine?

- Large numbers of independent DNA polymorphisms can inform us about ancestry and population history
- Responses to many therapeutic drugs may involve variation in just a few genes (along with environmental variation)
- These variants typically differ between populations only in their *frequency* and imply substantial overlap between populations

Blood pressure response to ACE inhibitors

(Sehgal, 2004, *Hypertension* 43: 566-72)



Frequencies of SNPs associated with response to anti-hypertensives

	CYP11B2 C-344T	*Angiotensin 2 receptor 1 A1166C*	*α-adducin G614T*	*G protein β3 C825T*	*Angio-tensinogen A-6G*
Africa	.20	.02	.07	.72	.98
Asia	.33	.05	.48	.43	.80
Europe	.43	.29	.21	.34	.49

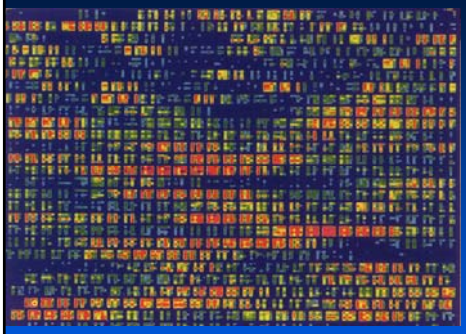
Average allele-frequency difference among major populations is 0.15

Gefitinib (Iressa) and non-small cell lung cancer

- Gefitinib inhibits epidermal growth factor receptor (EGFR) tyrosine kinase activity
- Effective in 10% of Europeans, 30% of Asians (Japanese, Chinese, Koreans)
- Somatic mutations in *EGFR* found in 10% of Europeans, 30% of Japanese
- 80% of those with mutations respond to gefitinib; 10% of those without mutations respond

Johnson and Jänne, 2005, *Cancer Res.* 65: 7525-9

Microarrays and “personalized medicine”



Hundreds of thousands of different DNA sequences can be placed on a single array

These sequences are compared with DNA from a patient to test for mutations

Signals are rapidly processed by a computer

Genetics and Race

- Genetic variation is correlated with geography and tends to be distributed continuously across geographic space
- “Race” may not be biologically meaningless, but it is biologically imprecise; ancestry is more informative
- Personalized medicine, when feasible, will be medically more useful than ethnicity or race
- Genetics provides no evidence that supports racism and much evidence that contradicts it

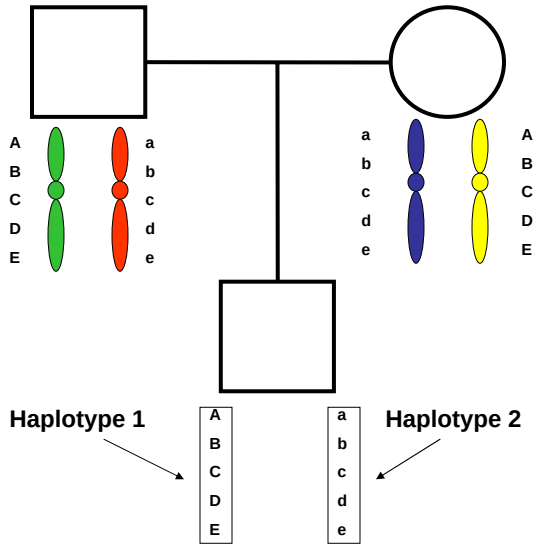
SNPs, haplotypes, linkage disequilibrium, and gene mapping

- A SNP with minor allele frequency (MAF) > 1% is found, on average, at 1/300 bp (roughly 10 million total)
- A “common” SNP (MAF > 5%) is found at about 1/600 bp (roughly 5 million total)
- SNPs have low mutation rates and can be typed by automated methods

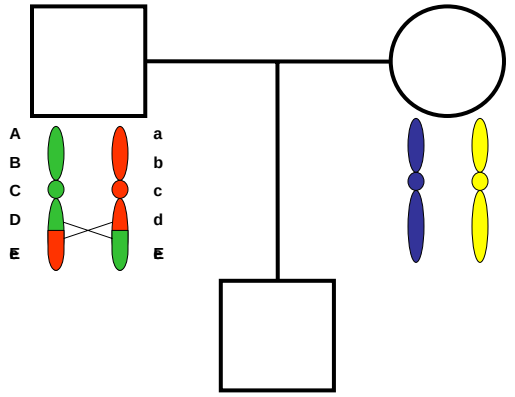
Whole-genome association: the cost problem

- A whole-genome association study seeks any SNP allele that is found with elevated frequency in disease cases
- At $.001 per SNP, genotyping 5 million SNPs costs $5,000 per person
- A study involving 1,000 cases and 1,000 controls would cost $10,000,000
- Will SNP association reveal disease genes, and do we need to test all of these SNPs?

A *haplotype* is the DNA sequence found on one member of the chromosome pair

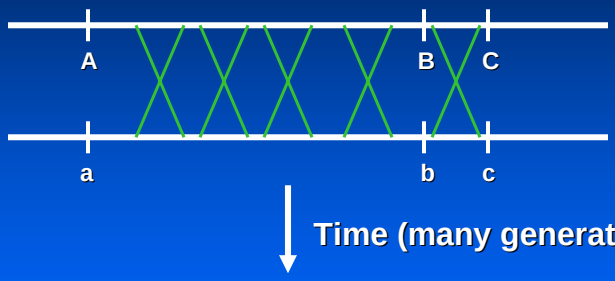


Crossovers during meiosis can create new haplotype combinations



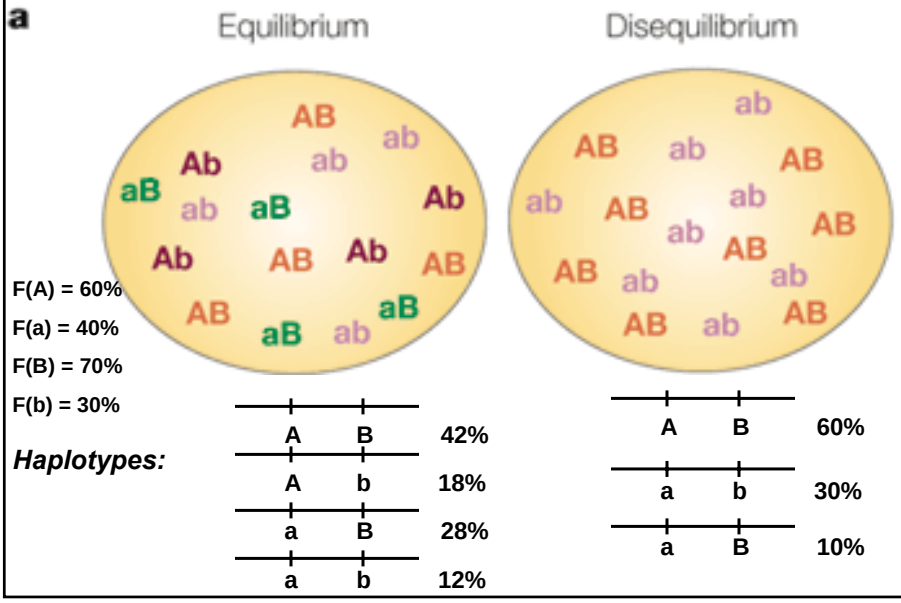
Crossovers **occur during meiosis, producing** ***recombination*** **of alleles**

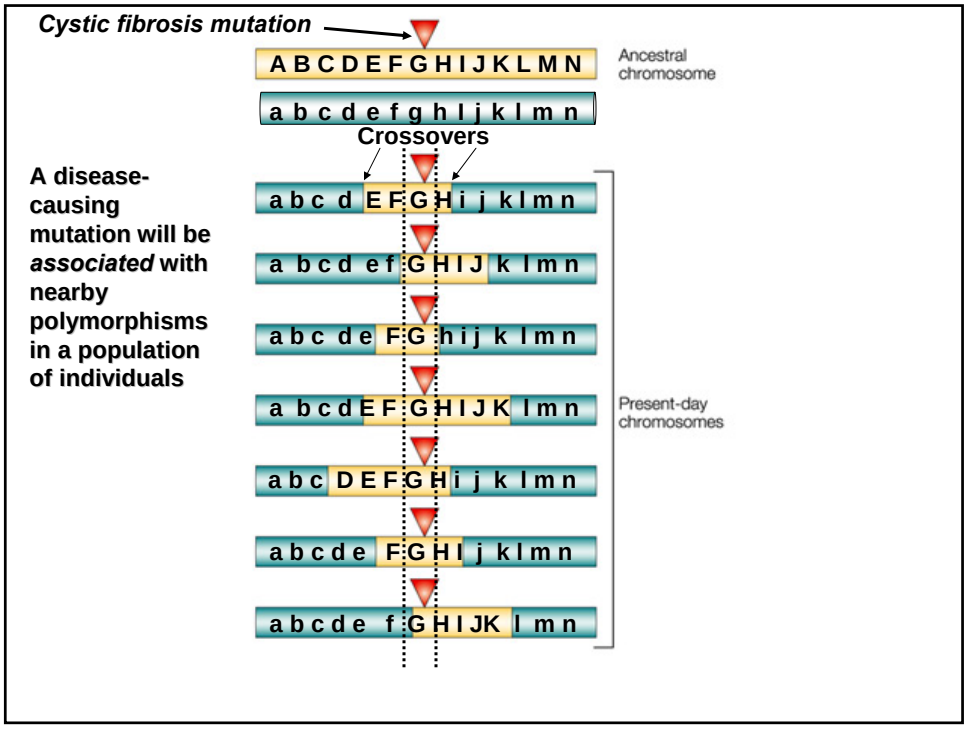
Over time, more crossovers will occur between loci located further apart



B and C will be found together on the same haplotype more often than A and B: there is more *linkage disequilibrium* between B and C than A and B

Linkage disequilibrium: nonrandom association of alleles at linked loci

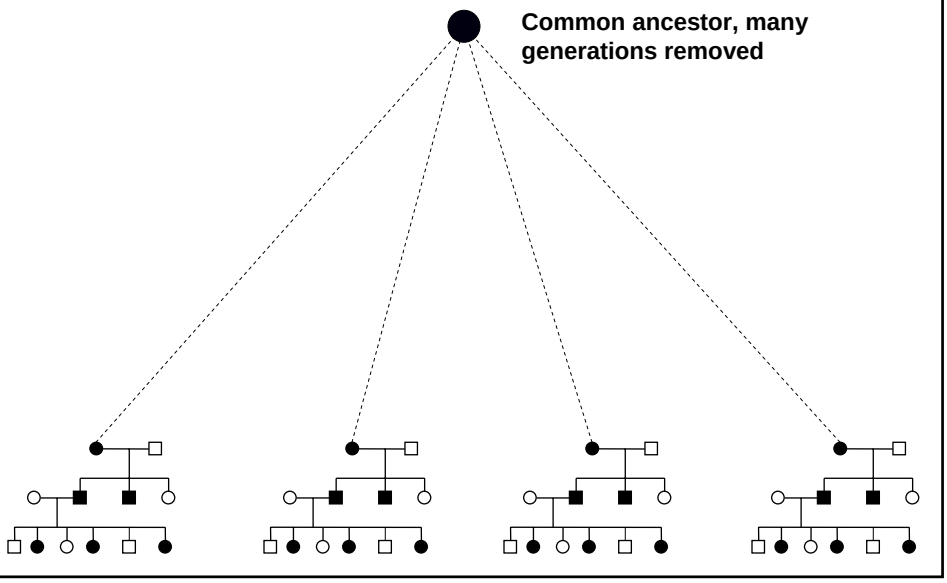




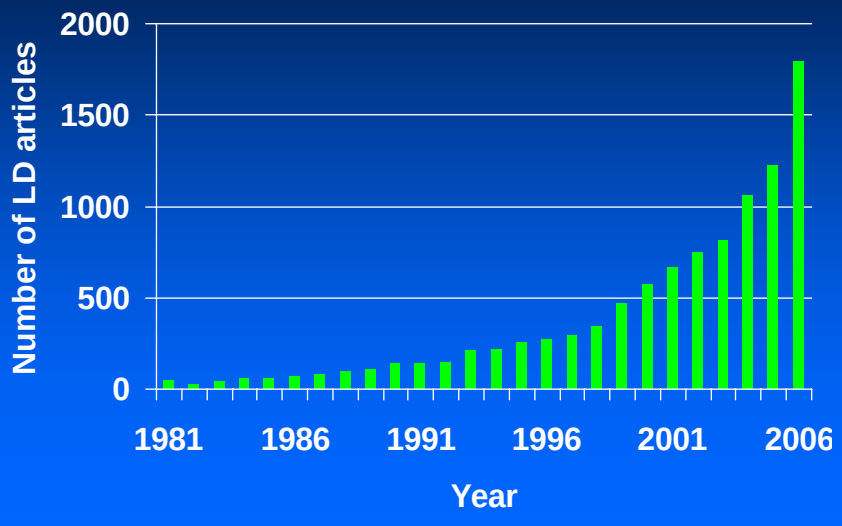
Potential advantages of linkage disequilibrium (LD)

- Family-based linkage studies of complex diseases often yield large candidate regions (~10-20 million base pairs)
- Association studies (linkage disequilibrium) can incorporate many past generations of recombination to narrow the candidate region
- Family data are *not* necessarily needed

Populations are one big (complicated) pedigree

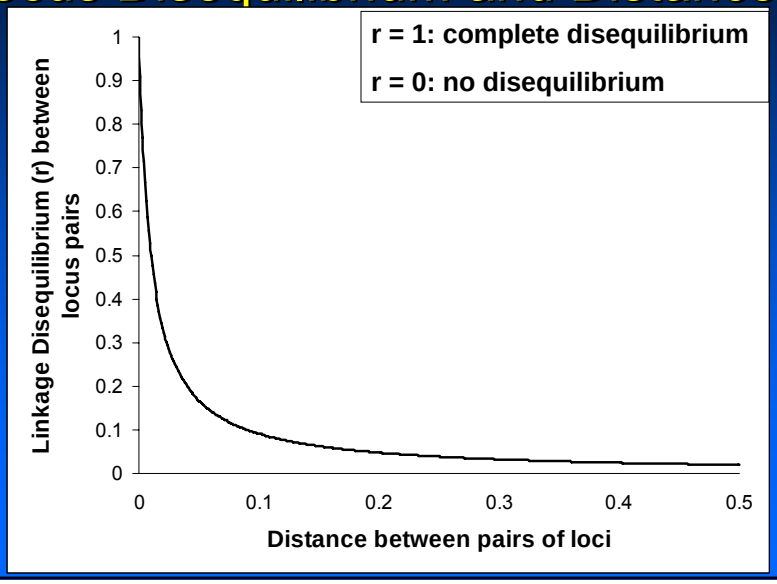


Number of published LD articles

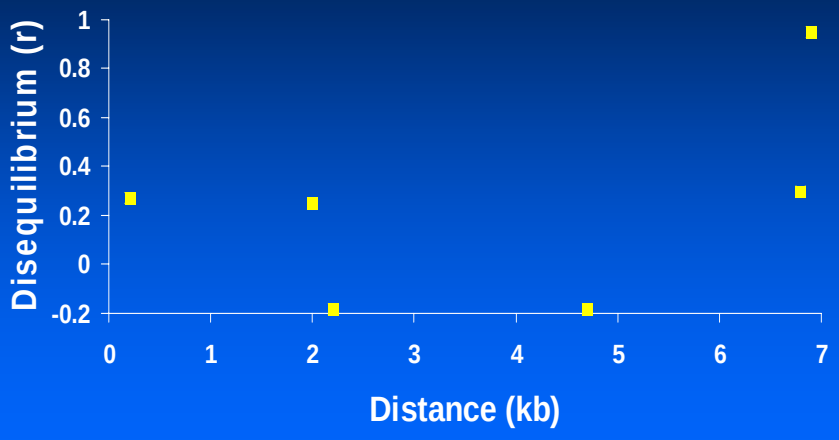


Is there a simple, uniform relationship between inter-locus physical distance and inter-locus linkage disequilibrium?

Expected Relationship between Inter-locus Disequilibrium and Distance

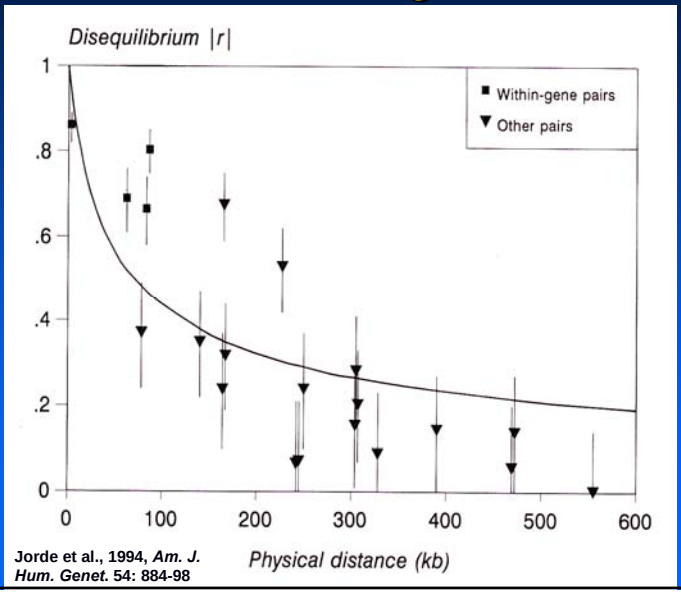


Linkage disequilibrium vs. physical distance on chromosome 11p



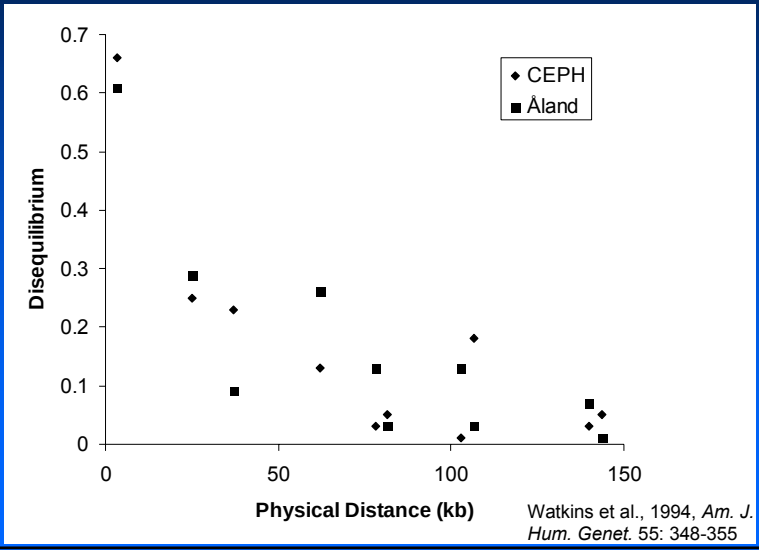
Barker et al., 1984, Am. J. Hum. Genet. 36: 1159-71

Disequilibrium between marker pairs in the *APC* region



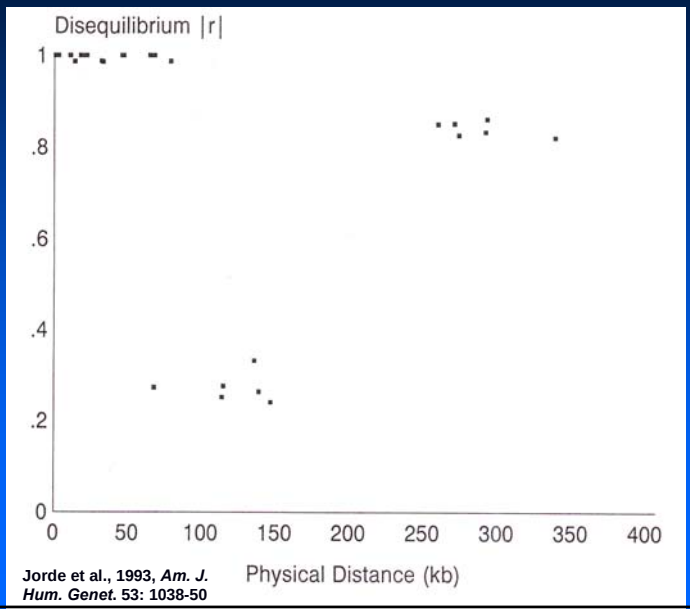
Jorde et al., 1994, *Am. J. Hum. Genet.* 54: 884-98

Linkage Disequilibrium and Physical Distance: *vWF* Region



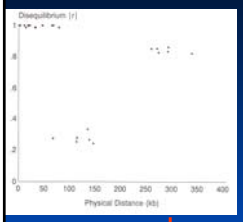
Watkins et al., 1994, *Am. J. Hum. Genet.* 55: 348-355

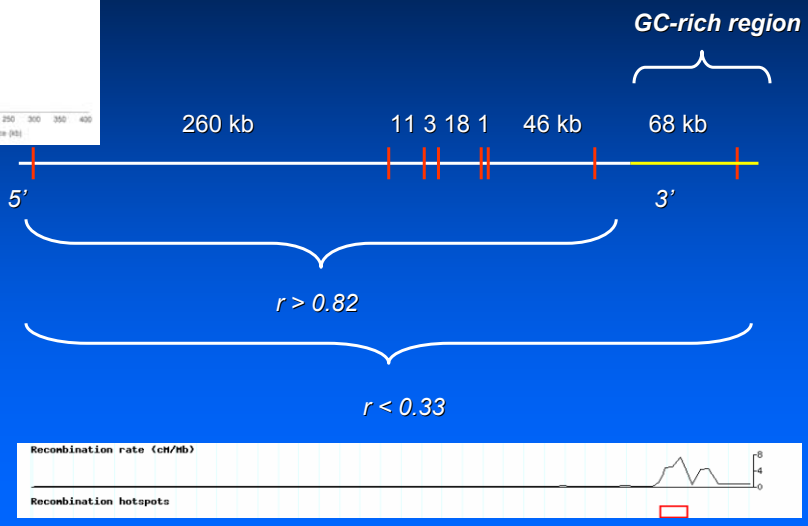
Disequilibrium in the *NF1* region



Jorde et al., 1993, *Am. J. Hum. Genet.* 53: 1038-50

Uneven Disequilibrium Pattern in the NF1 Region





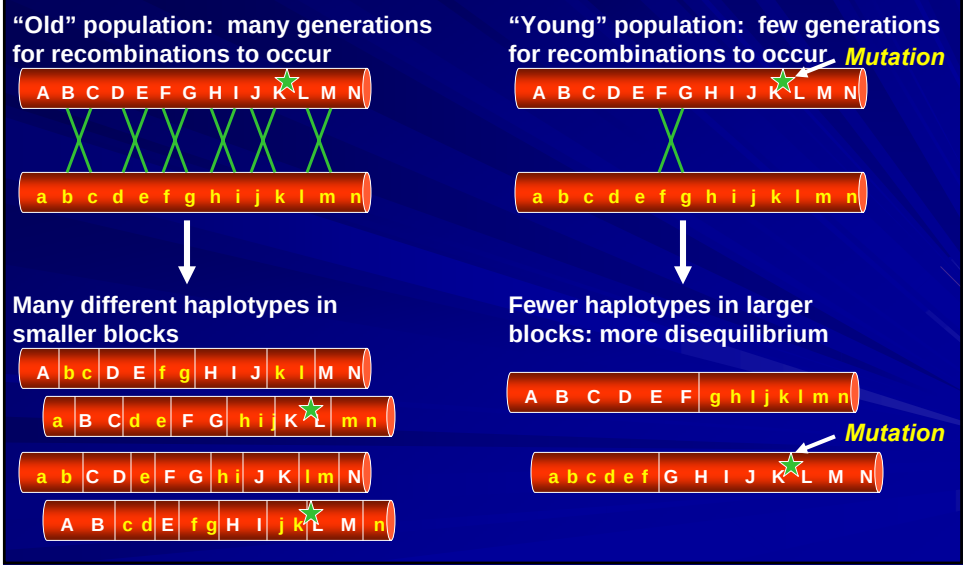
Factors that May Affect Linkage Disequilibrium Patterns

- Chromosome location
  - Telomeric vs. centromeric
  - Intragenic vs. extragenic
- DNA sequence patterns (GC content)
- Recombination hotspots (1 every 50-100 kb)
- Evolutionary factors: LD varies among populations
  - Natural selection
  - Gene flow
  - Mutation, gene conversion
  - Genetic drift

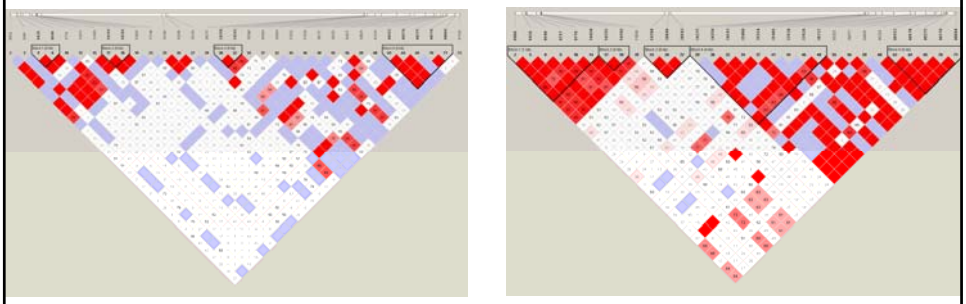
Patterns of genetic variation: implications for disequilibrium

- Continental variation patterns affect stratification and admixture LD mapping design
- Greater "age" of African populations: LD persists over shorter physical distances
- Greater divergence of African populations: LD patterns more likely to differ from other populations: African-American populations especially useful for admixture LD mapping
- Common alleles and haplotypes are likely to be shared across populations: association patterns may be shared

Population "age" can affect haplotype structure



Linkage disequilibrium: *CD4* region



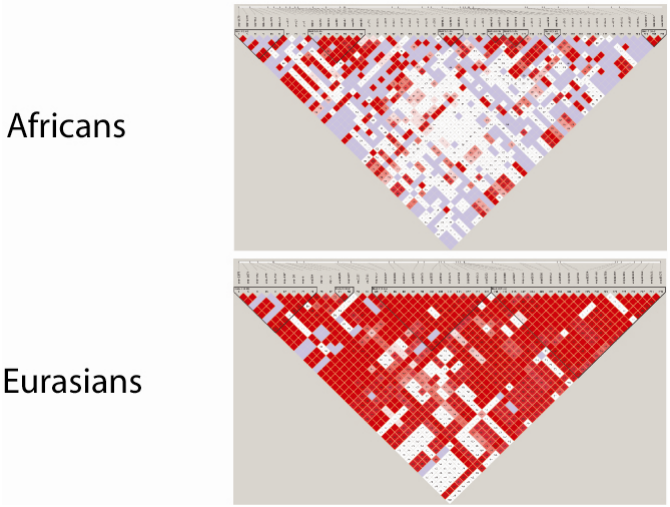
African

Non-African

Haploview

Prahalad et al., submitted

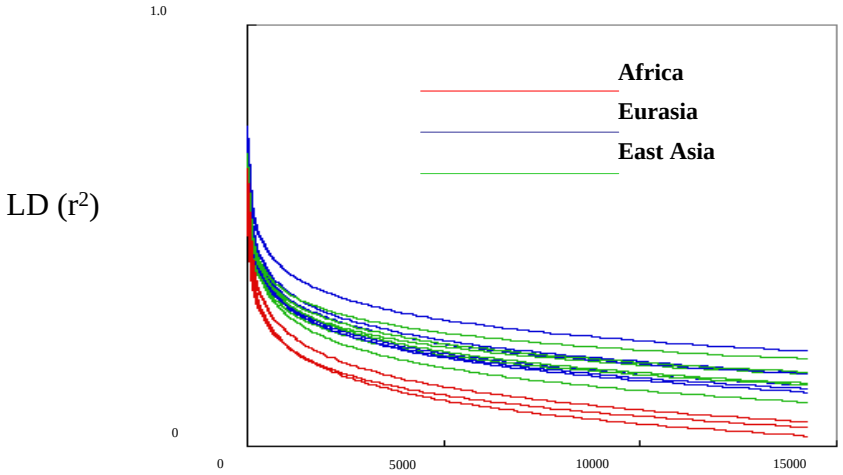
Pairwise LD at the Angiotensinogen locus



Africans

Eurasians

Population variation in *AGT* disequilibrium



Nakajima et al., 2004, *Am. J. Hum. Genet.* 74: 898-916

How general are these patterns?

To what extent does LD vary with genomic location and population?

A Map of the World, 1544

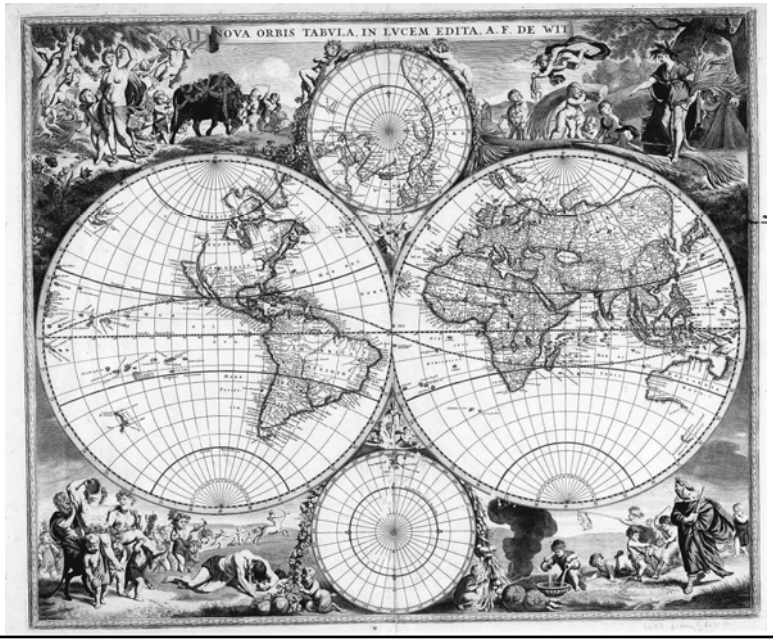
In search of a better map: The International Haplotype Map Project

- 600,000 SNPs (1 per 5 kb) genotyped in 270 individuals
  - 90 CEPH Utah individuals (30 trios)
  - 90 Yoruban from Nigeria (30 trios)
  - 90 East Asians (45 Chinese, 45 Japanese)
- Evaluate patterns of linkage disequilibrium and haplotype structure
  - Variation in different genomic regions
  - Variation in different populations

Some of the issues surrounding HapMap

- Choice of populations
  - How best to *sample* human diversity
  - Families vs. unrelated individuals
  - Sample size
- SNP ascertainment and density
- ELSI
  - Informed consent (individual consent and community consultation)
  - Avoidance of stigmatization

A Map of the World, 1688



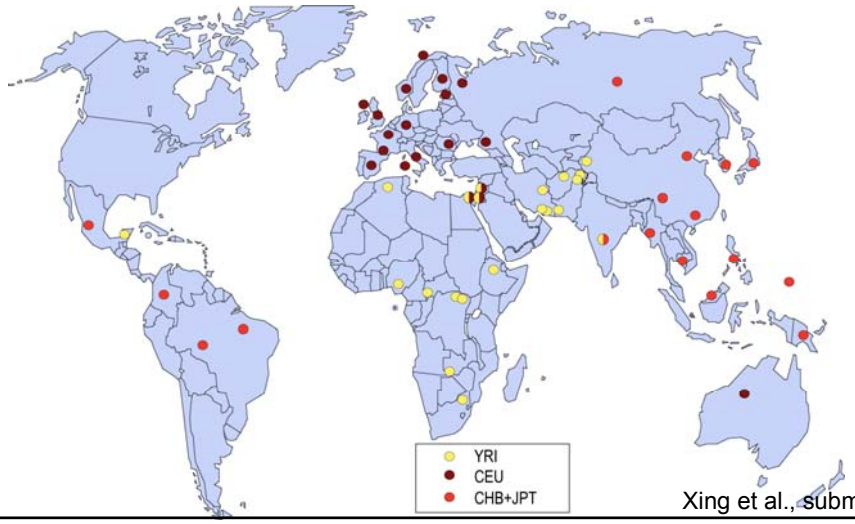
Genetic applications of HapMap

- Understanding human genome-wide haplotype diversity
- Detection of recombination hotspots
- Detection of genes that have experienced strong natural selection
- Detection of disease-causing mutations

SNPs in disequilibrium are redundant: we don't need to type all of them

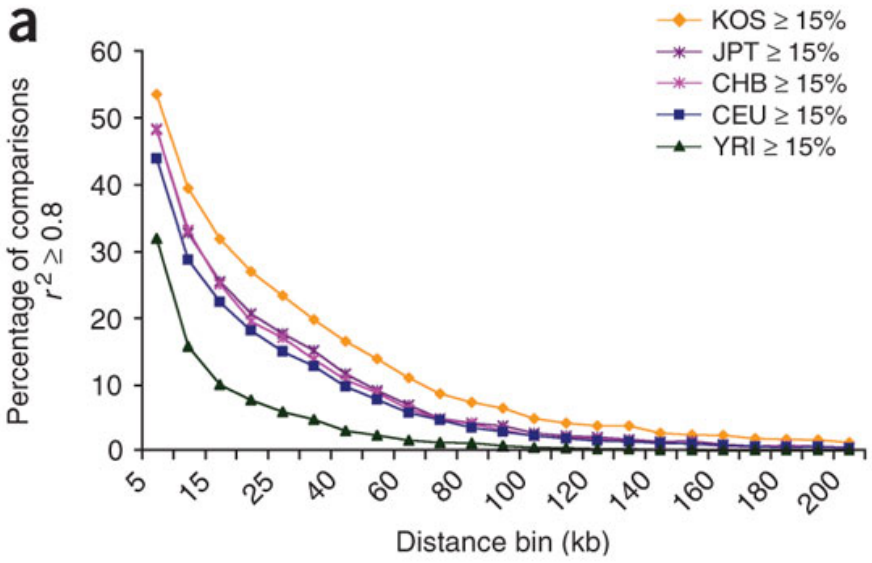
For whole-genome association studies, "complete" coverage is given by about 1.6 million SNPs for African populations, 1,000,000 SNPs for non-African populations

Portability of HapMap tag SNPs: HapMap SNPs recover 80-90% or more of SNP variation in other populations



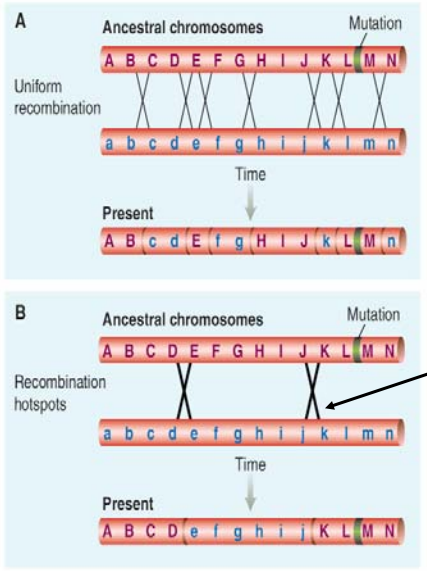
Xing et al., submitted

LD decline in Kosrae, an isolate, compared to HapMap samples



Bonnen et al., 2006, *Nat. Genet.* 38: 214-7

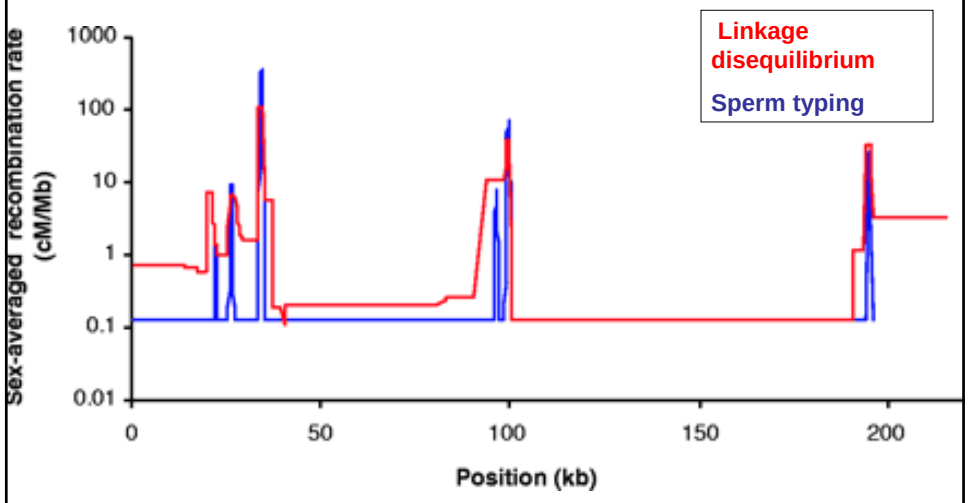
Recombination hotspots and haplotype blocks



Recombination hotspots

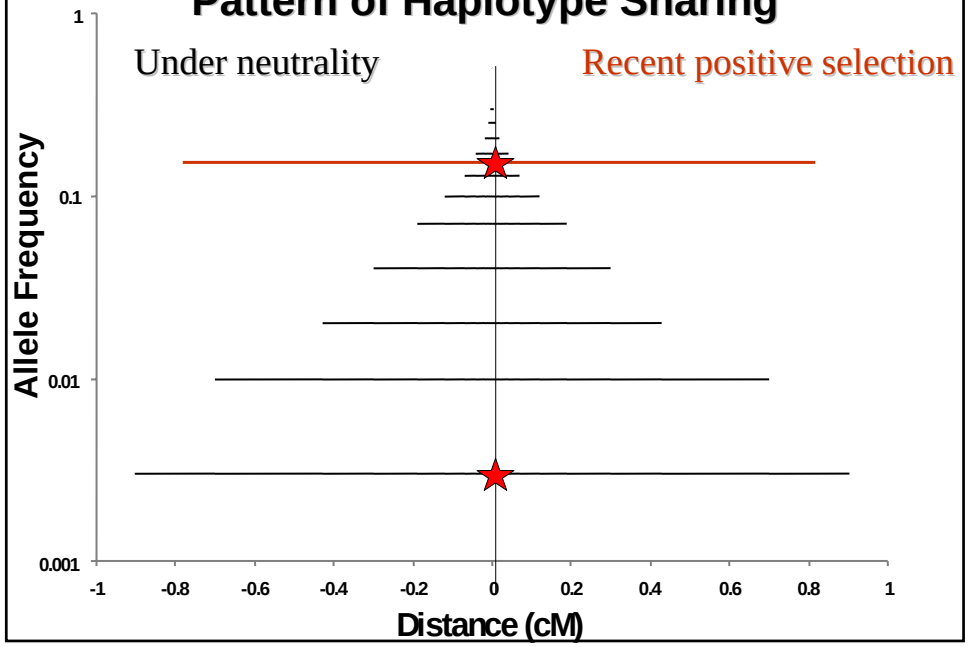
- LD patterns indicate 25,000 - 50,000 hotspots in human genome (1 every 50 – 100 kb) (Myers et al., 2005, *Science* 310: 321-4)
- 80% of recombination occurs in ~15% of the genome (60% occurs in 6% of genome)
- Hotspots are not congruent in human and chimpanzee, despite 99% sequence identity: suggests hotspots evolve rapidly and may not be sequence-dependent

Linkage disequilibrium detects true recombination hotspots accurately



McVean et al., 2004, *Science* 304: 581-4

LD and Natural Selection: Hypothetical Pattern of Haplotype Sharing



Examples of genes in which elevated LD indicates recent natural selection

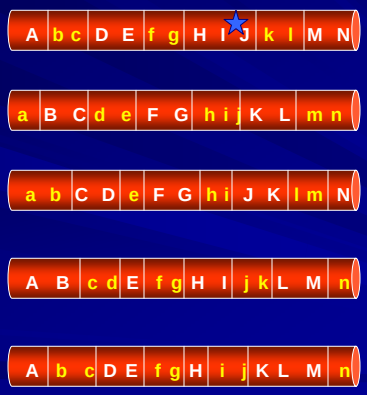
Gene	Phenotype
G6PD	Malaria protection
Hemochromatosis	Iron absorption
CYP3A5	Sodium retention
Lactase	Lactose tolerance
SLC24A5	Skin pigmentation
Alcohol dehydrogenase	Ethanol metabolism

Voight et al., 2006. *PLOS Biology* 4: 446-458

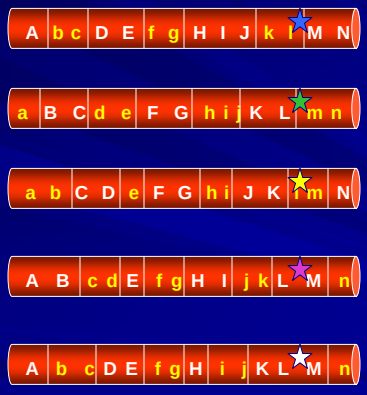
Linkage disequilibrium and single-gene diseases: many successes

- Cystic fibrosis
- Hemochromatosis
- Wilson disease
- Friedreich's ataxia
- Bloom syndrome
- Werner syndrome
- Progressive myoclonus epilepsy
- Torsion dystonia
- Diastrophic dysplasia (and many other "Finnish" diseases)

Association (linkage disequilibrium) studies are most successful when the disease is (mostly) caused by a single mutation



Multiple disease-causing mutations can pose problems for association analysis



How can we reduce heterogeneity?

- Define the trait consistently and accurately
- Identify subtypes
  - Early onset
  - Severe expression
  - Atypical expression
- Use strict, narrow population definitions

Linkage disequilibrium and complex diseases: some recent successes

- *NOD2* (*CARD15*), *IL23R* and Crohn's disease
- *ADAM33, GPRA,* and asthma
- Neuregulin and schizophrenia
- Complement factor H and age-related macular degeneration
  - *HapMap data used to define a 41 kb block to focus mutation search*

Population genetics and genome analysis

- Genetic variation contains useful information about population history
- Genetic variation provides a more informed view of “race” and its relevance to medicine
- Population genetic analysis has been critical in understanding linkage disequilibrium
- Population genetics is *fun*!