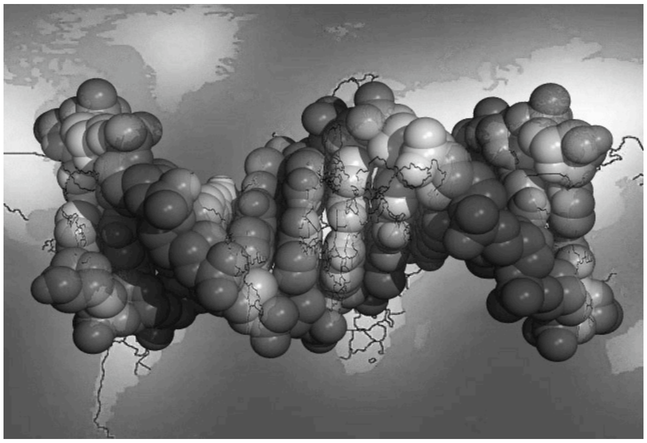
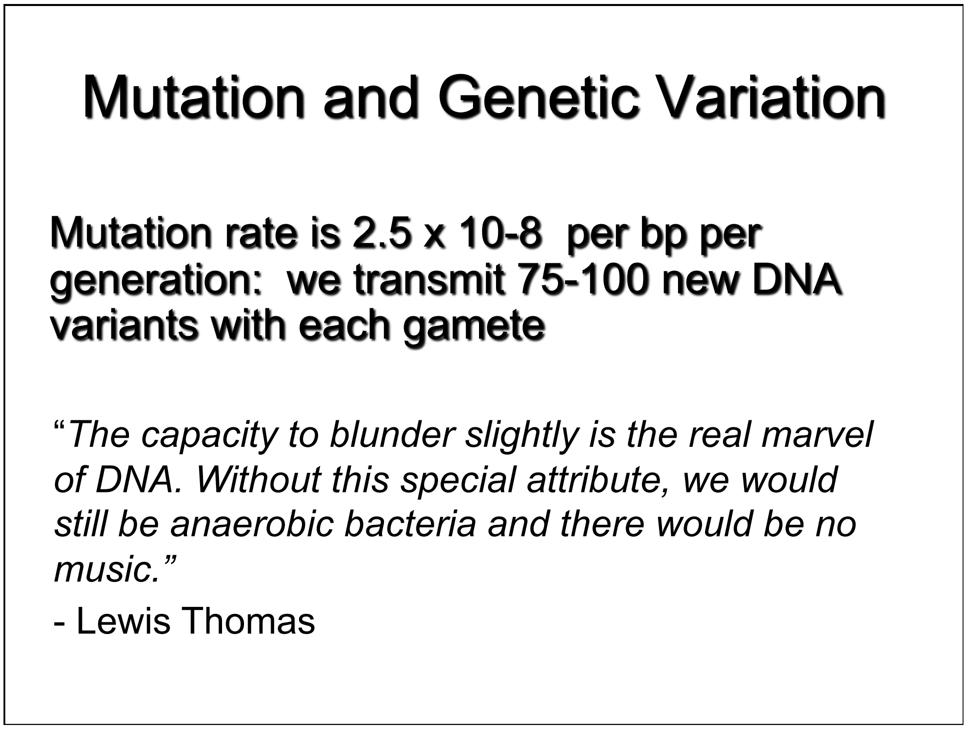
Introduction to Population Genetics



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

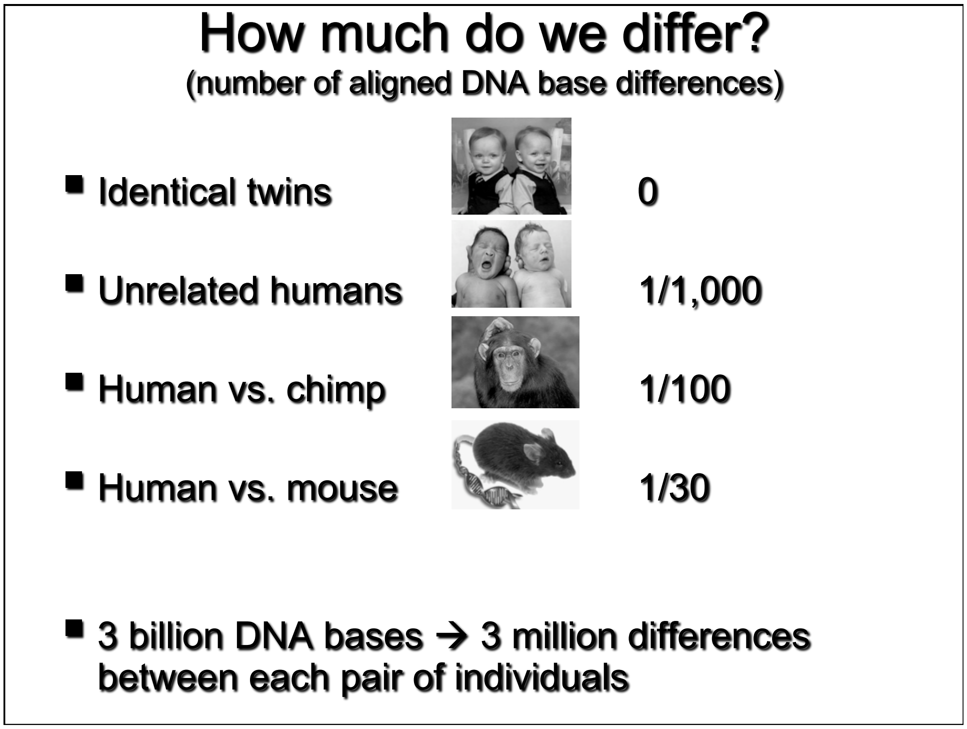


Mutation and Genetic Variation

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

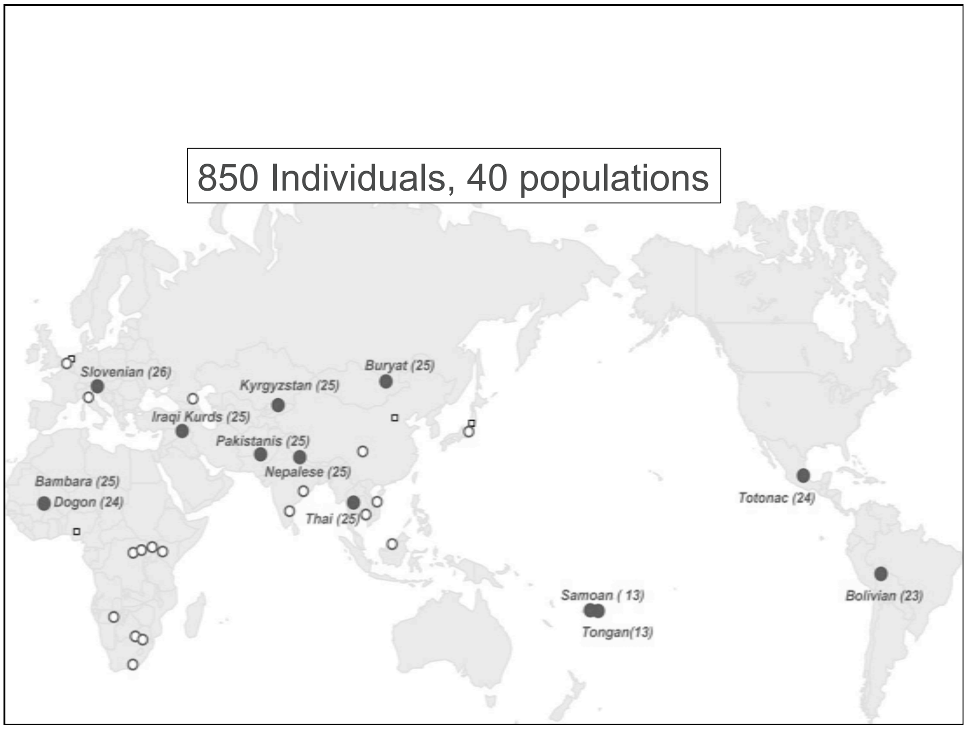


How much do we differ?

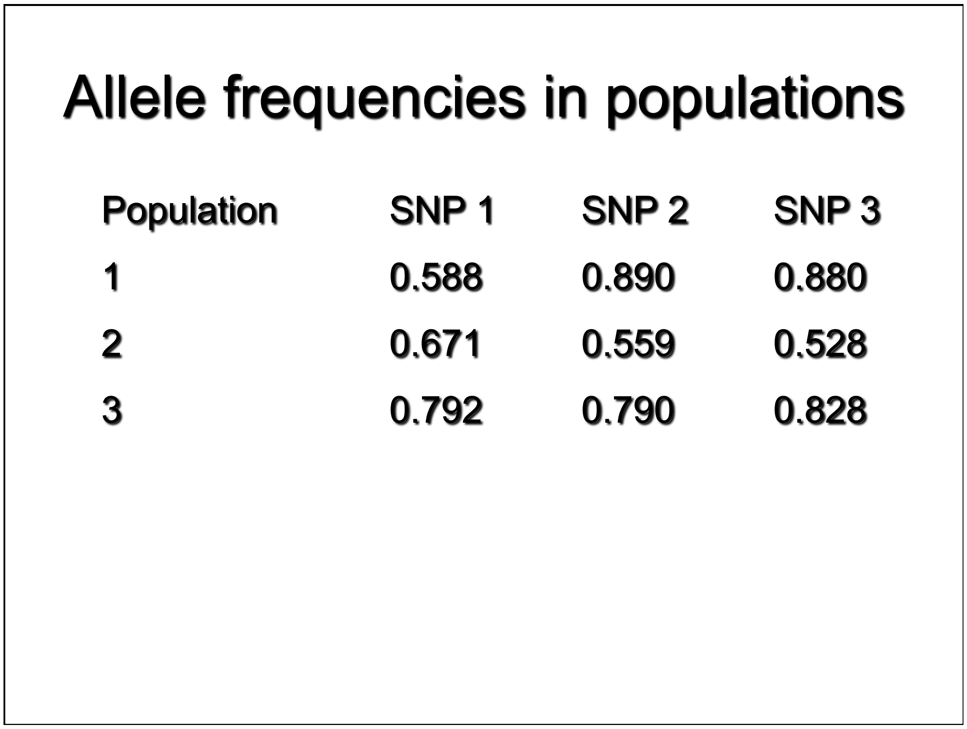
(number of aligned DNA base differences)

Identical twins	0
Unrelated humans	1/1,000
Human vs. chimp	1/100
Human vs. mouse	1/30

- **3 billion DNA bases → 3 million differences between each pair of individuals**



850 Individuals, 40 populations



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?

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T}$$

F_{ST} is the amount of genetic variation that is due to population differences

H_T is the total heterozygosity (variation) in the sample

H_S is the average heterozygosity within each population (continent)

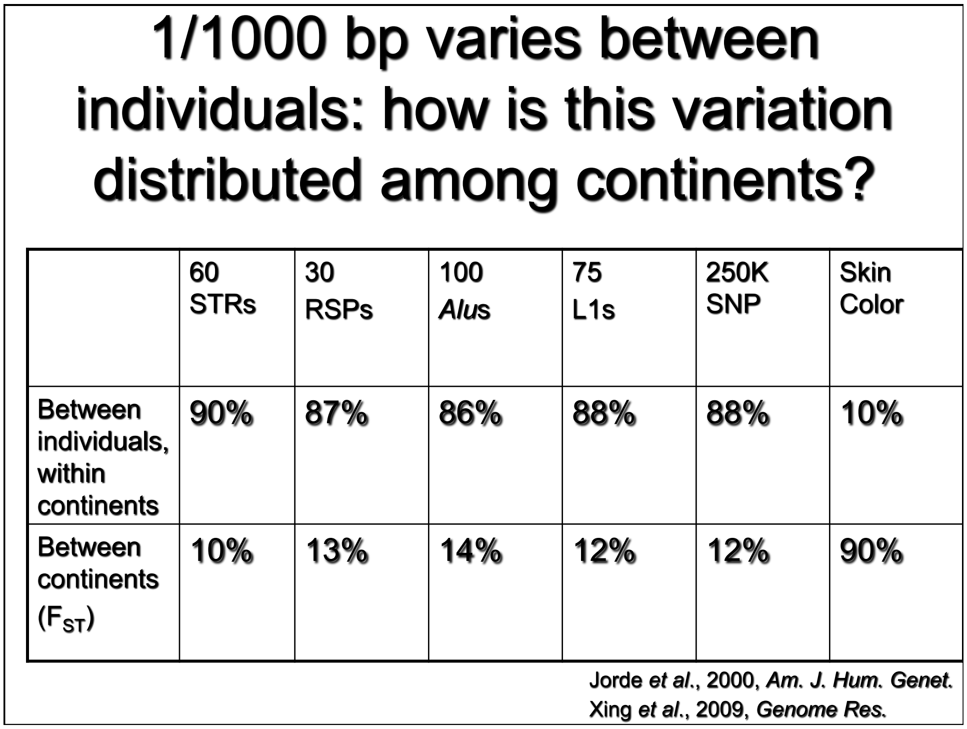
$F_{ST} = 0$: All variation exists within populations; none exists between

$F_{ST} = 1$: All variation exists between populations

1/1000 bp varies between individuals: how is this variation distributed among continents?

	60 STRs	30 RSPs	100 *Alu*s	75 L1s	250K SNP	
Between individuals, within continents	90%	87%	86%	88%	88%	
Between continents (F_{ST})	10%	13%	14%	12%	12%	

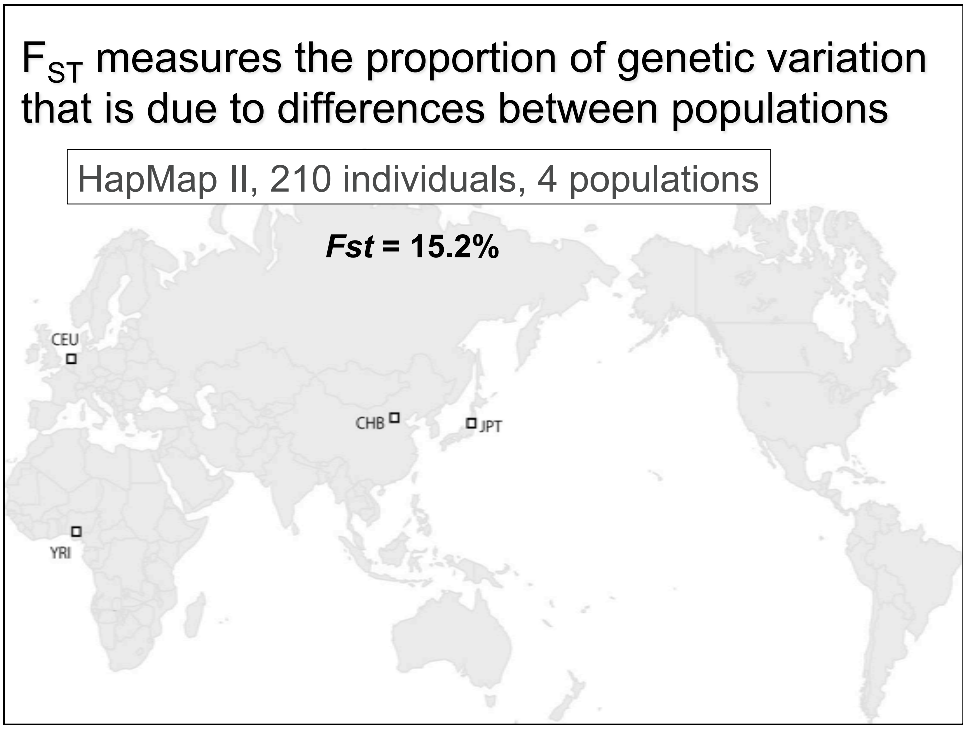
Jorde *et al*., 2000, *Am. J. Hum. Genet.*
Xing *et al*., 2009, *Genome Res.*



1/1000 bp varies between individuals: how is this variation distributed among continents?

	60 STRs	30 RSPs	100 *Alu*s	75 L1s	250K SNP	Skin Color
Between individuals, within continents	90%	87%	86%	88%	88%	10%
Between continents (F_{ST})	10%	13%	14%	12%	12%	90%

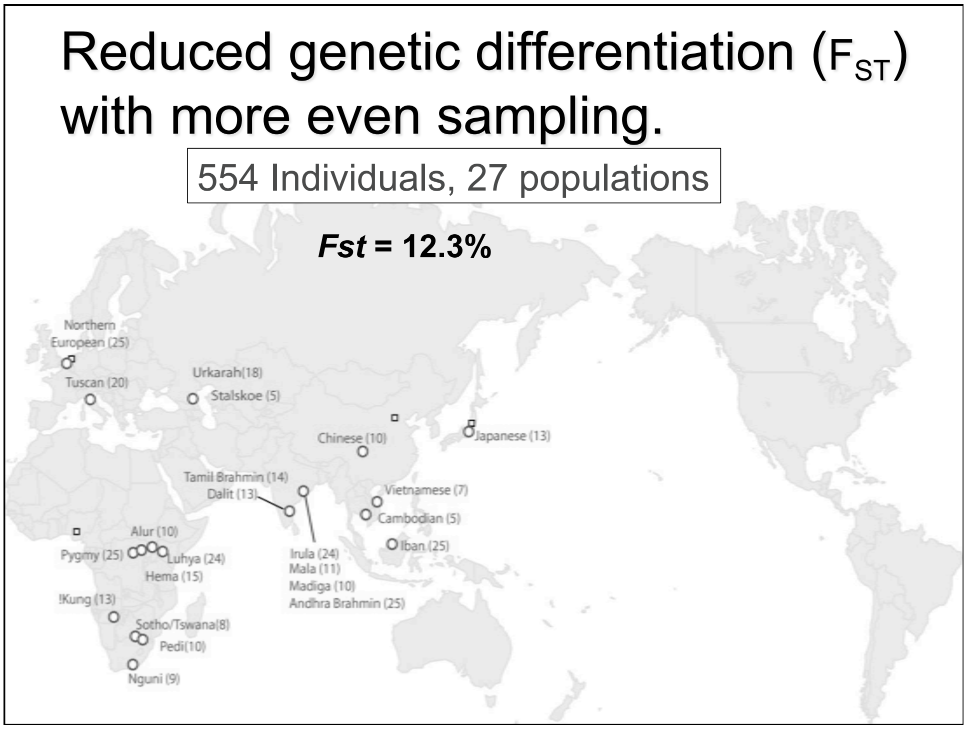
Jorde *et al*., 2000, *Am. J. Hum. Genet.*
Xing *et al*., 2009, *Genome Res.*



F_{ST} measures the proportion of genetic variation that is due to differences between populations

HapMap II, 210 individuals, 4 populations

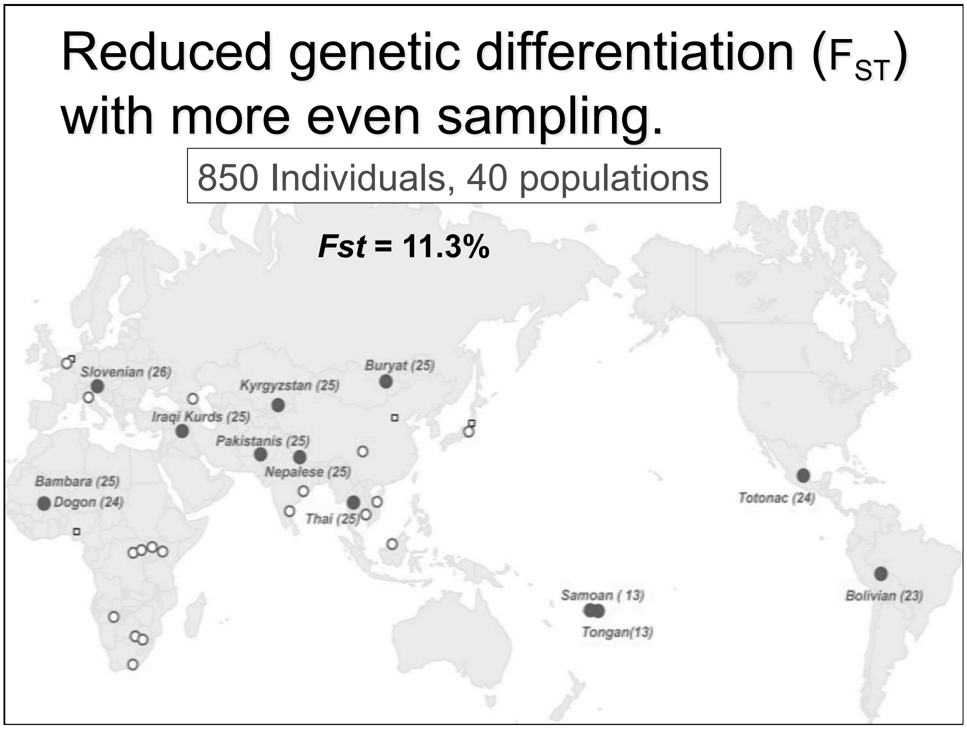
Fst **= 15.2%**



Reduced genetic differentiation (F_{ST}) with more even sampling.

554 Individuals, 27 populations

***Fst* = 12.3%**



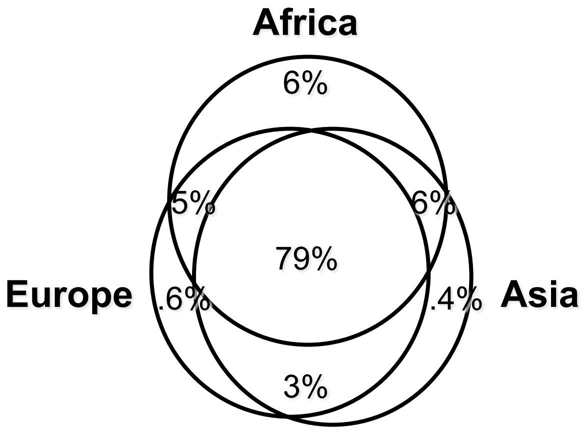
Reduced genetic differentiation (F_{ST}) with more even sampling.

850 Individuals, 40 populations

***Fst* = 11.3%**

Most common genetic variants are shared among populations:

7,742 SNPs >.05 in ENCODE database



Whole-genome sequence comparisons

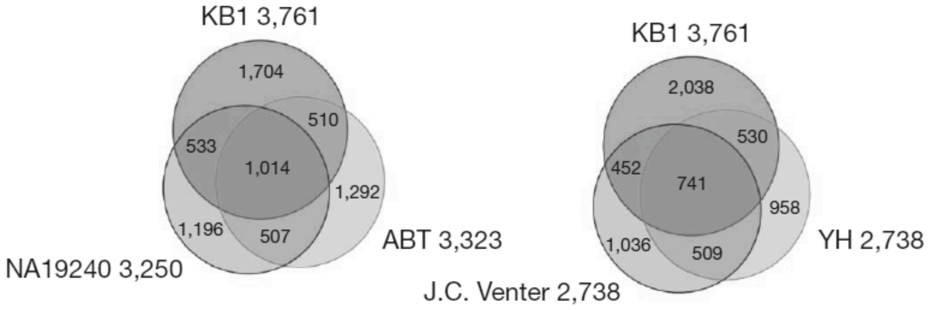


Figure 2 | Three-way relationships among SNPs. SNPs from KB1 are compared with those of the Yoruban NA19240 and ABT (left panel), and with an American of European descent (J. C. Venter) and a Chinese individual (YH) (right panel). Numbers are given in thousands.

Schuster et al., 2010, *Nature*

A simple genetic distance measure

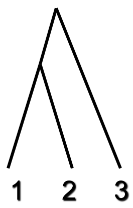
$$D_{ij} = |p_i - p_j|$$

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)

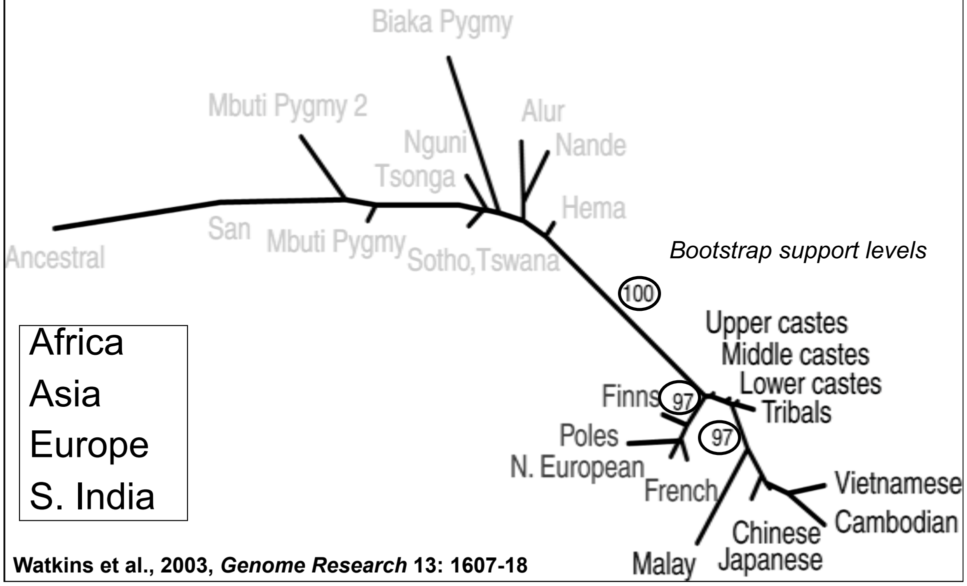
Building a population network



Pop.	SNP 1
1	0.588
2	0.671
3	0.792

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

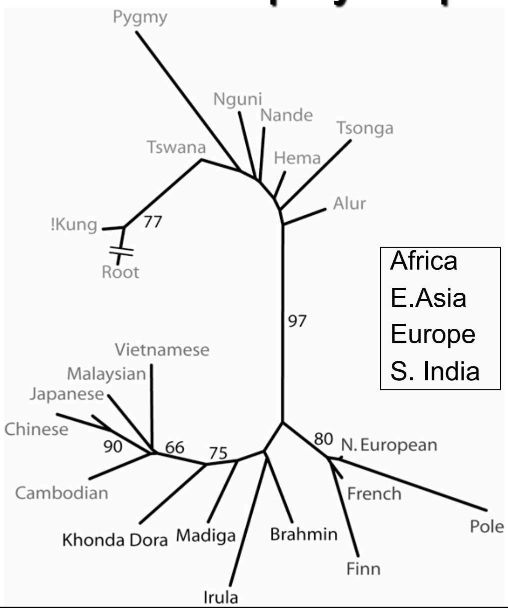
Genetic relationships based on 100 autosomal *Alu* polymorphisms



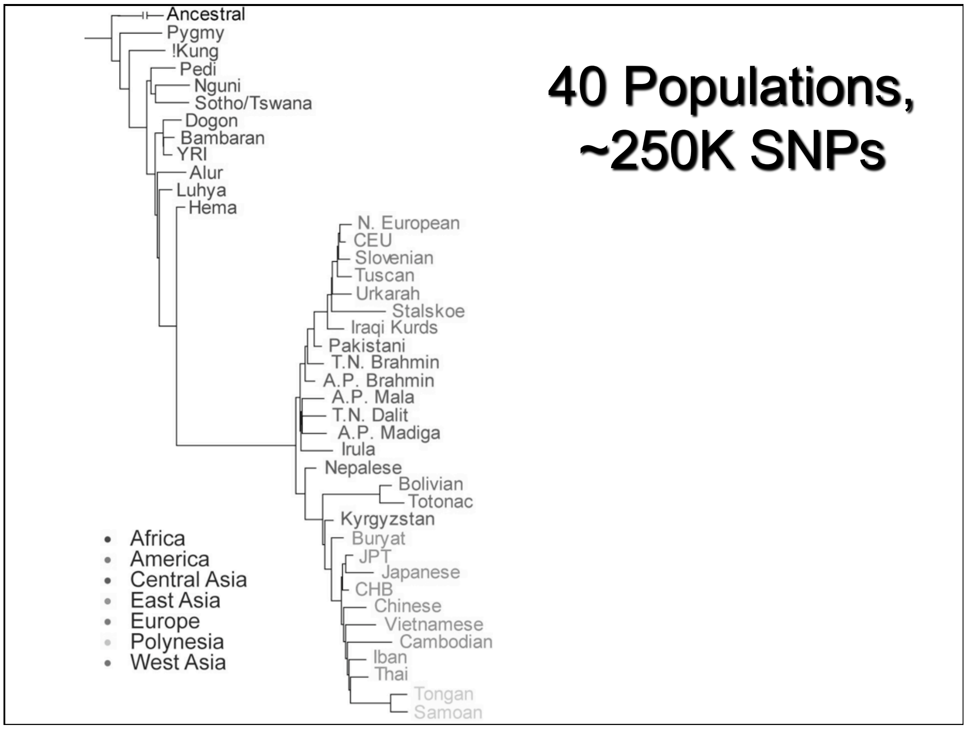
Bootstrap support levels

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

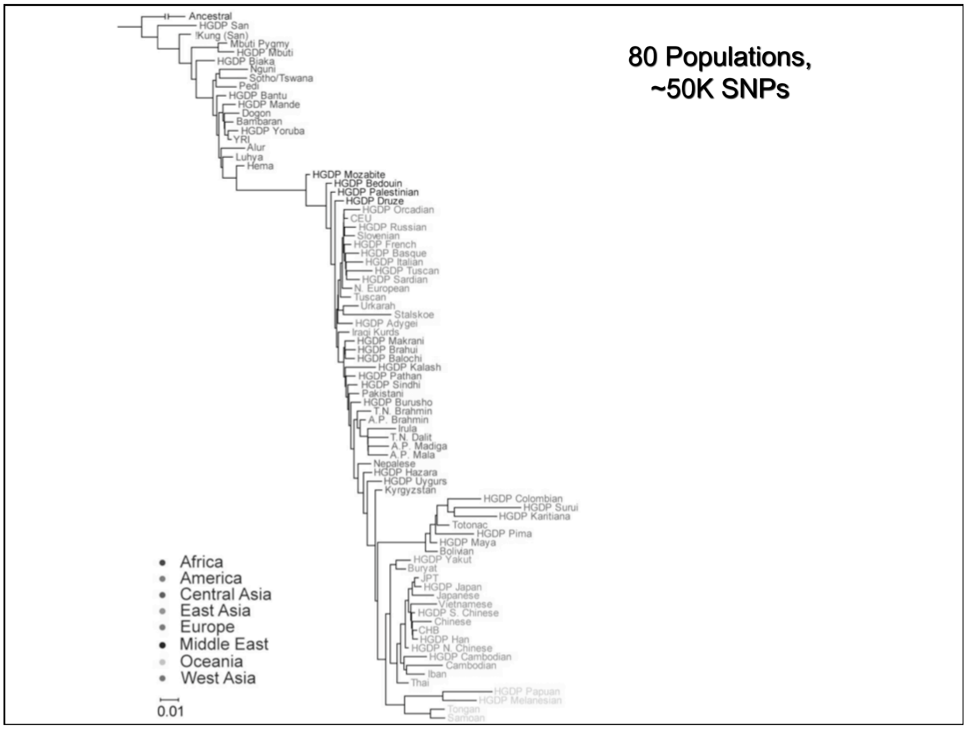
Genetic relationships based on 75 autosomal L1 polymorphisms



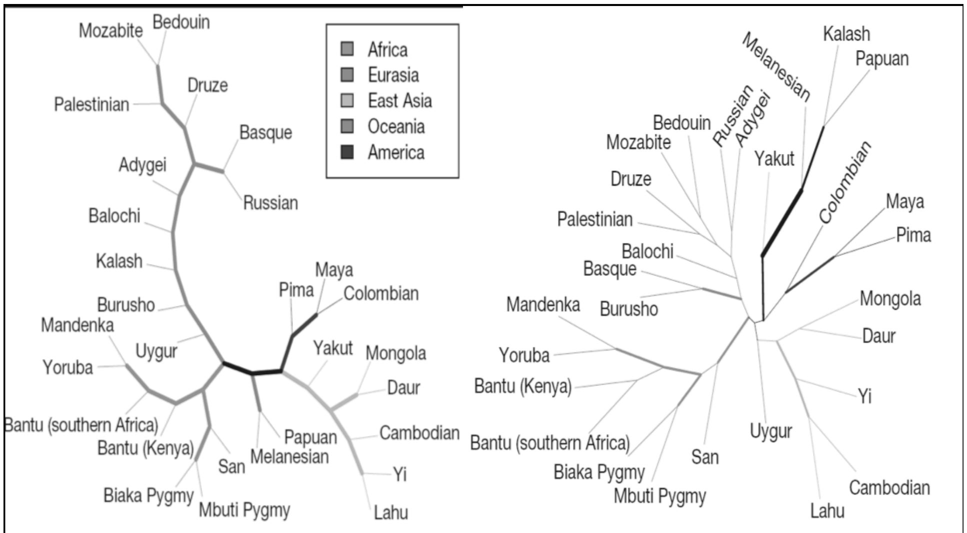
Witherspoon et al., 2006, *Hum. Hered.*, 62: 30-46



40 Populations, ~250K SNPs



80 Populations, ~50K SNPs

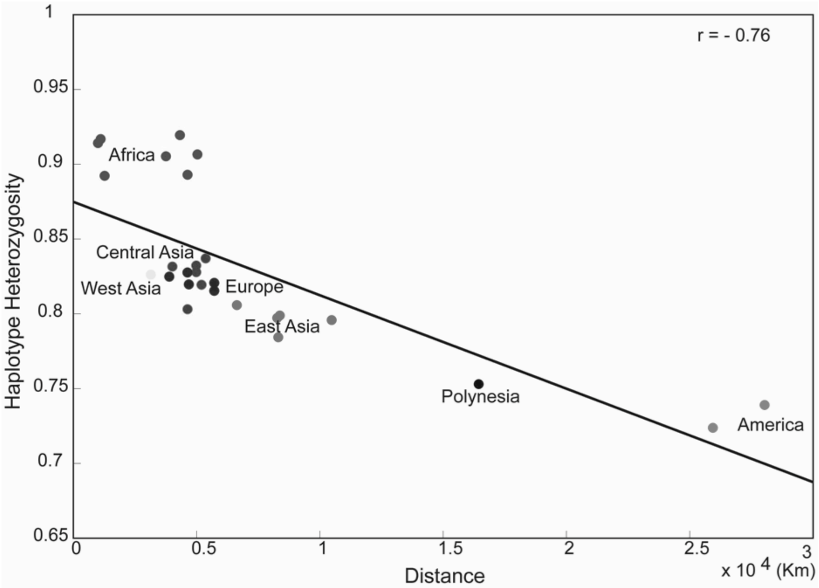


525,910 SNPs

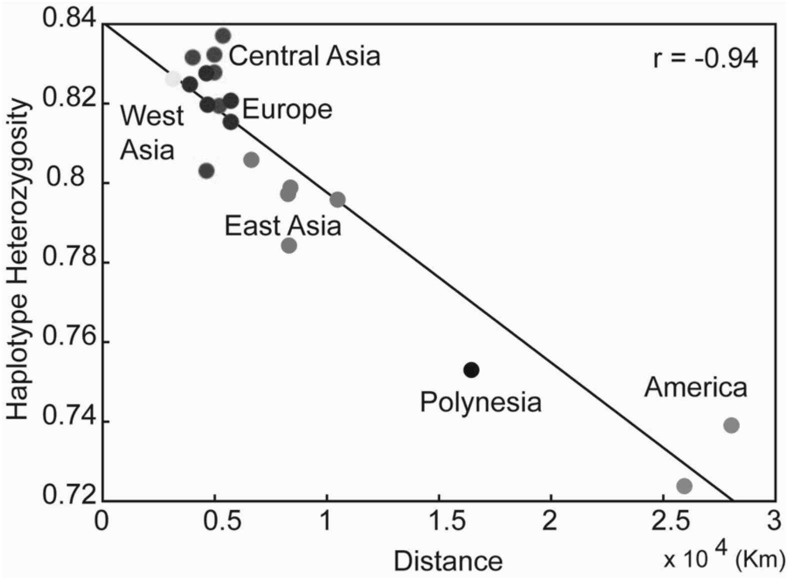
396 copy number variants (CNVs)

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

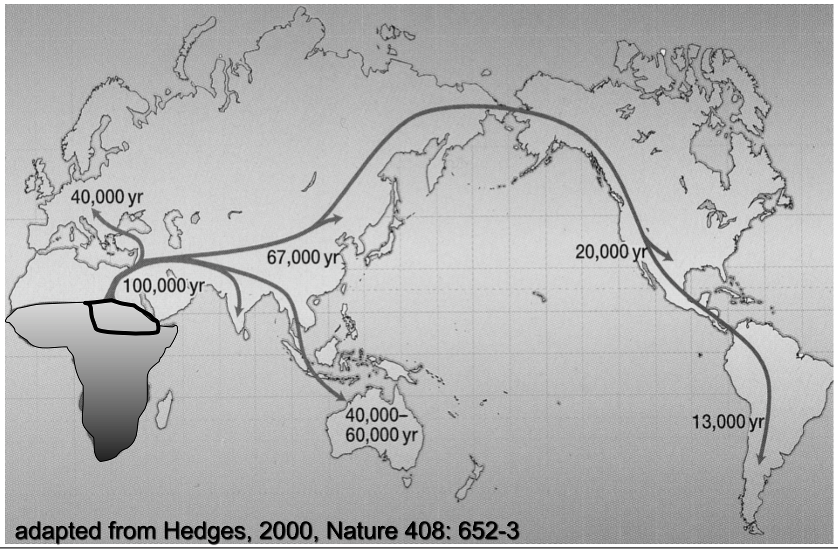
Haplotype diversity declines with geographic distance from Africa



Haplotype diversity declines with geographic distance from Africa



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.*

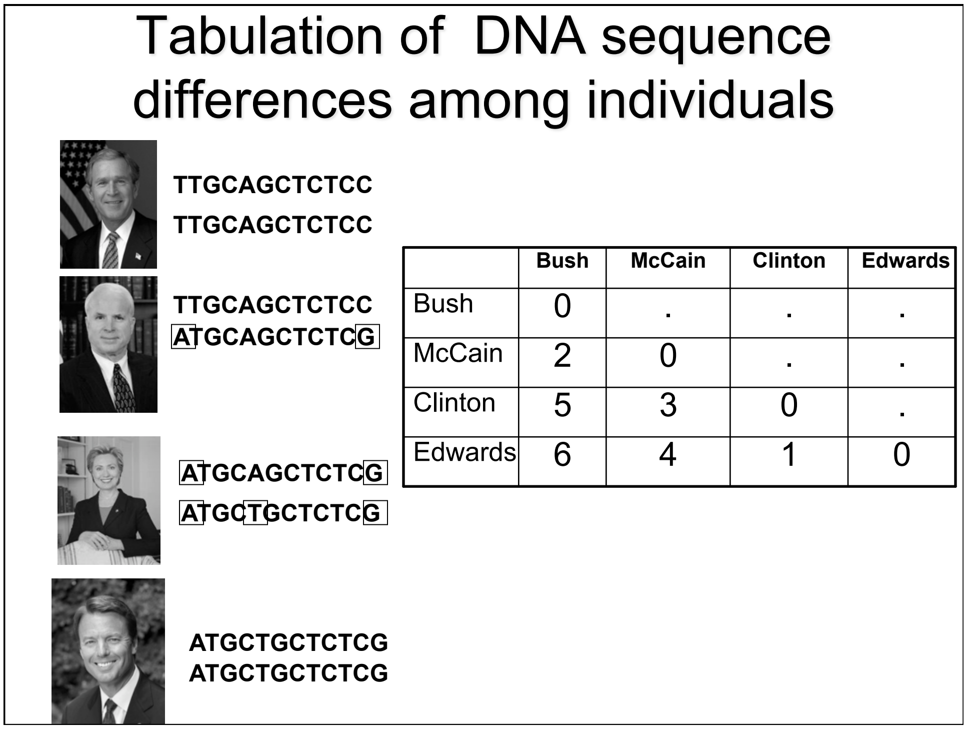
"I am a racially profiling doctor"
-- Satel, May 5, 2002, *New York Times*

"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

NEW TWISTS ON DNA • 100 YEARS AFTER THE WRIGHT BROTHERS
SCIENTIFIC AMERICAN
Tech Leaders of 2003:
The Scientific American 50
DECEMBER 2003
$4.95
WWW.SCIAM.COM
Science Has the Answer:
DOES RACE EXIST?
Genetic Results May Surprise You
The Day the Earth Burned
Reasons to Return to the Moon
COPYRIGHT 2003 SCIENTIFIC AMERICAN, INC.

Bamshad and Olson, 2003

Tabulation of DNA sequence differences among individuals



Tabulation of DNA sequence differences among individuals

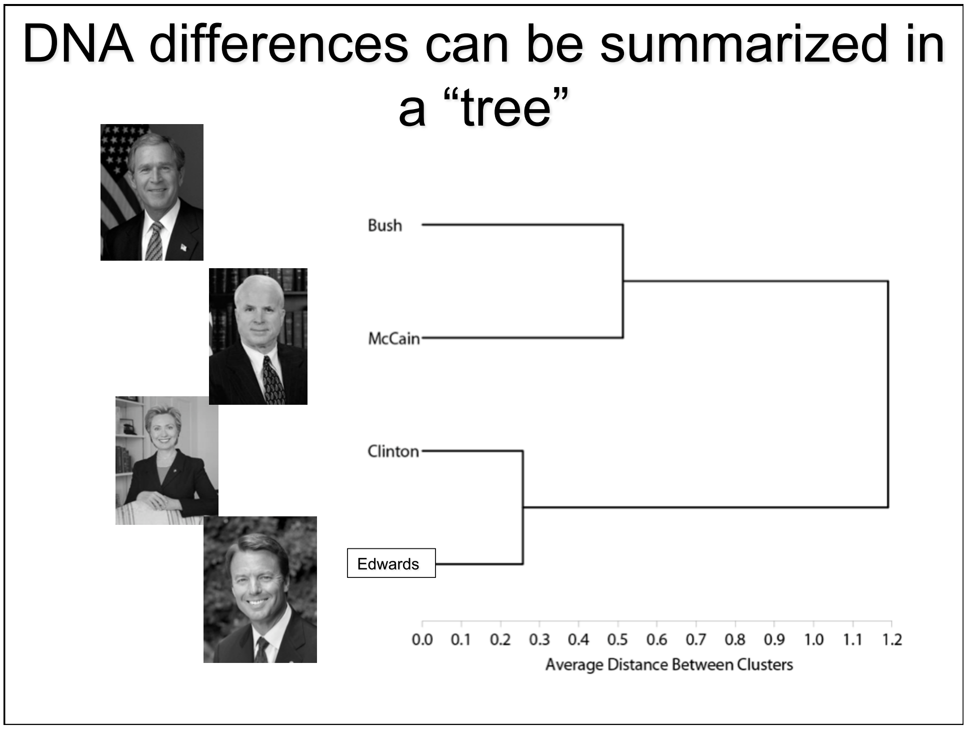
TTGCAGCTCTCC
TTGCAGCTCTCC

TTGCAGCTCTCC
ATGCAGCTCTCG

ATGCAGCTCTCG
ATGCTGCTCTCG

ATGCTGCTCTCG
ATGCTGCTCTCG

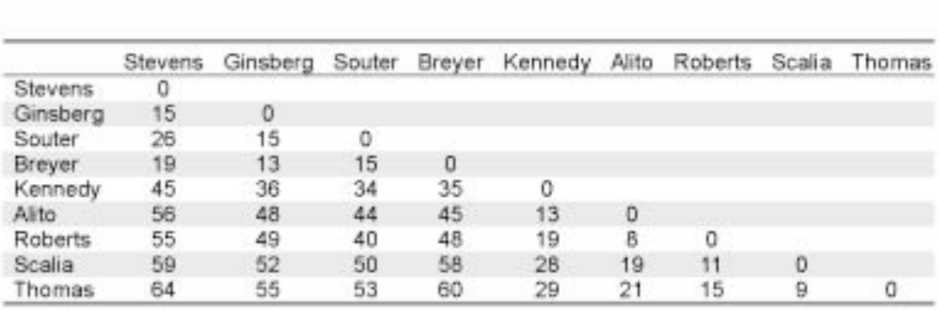
	Bush	McCain	Clinton	Edwards
Bush	0	.	.	.
McCain	2	0	.	.
Clinton	5	3	0	.
Edwards	6	4	1	0



DNA differences can be summarized in a “tree”

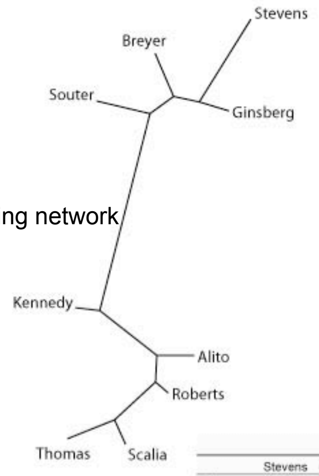
A distance matrix based on Supreme Court decisions

Distance matrix: % disagreement



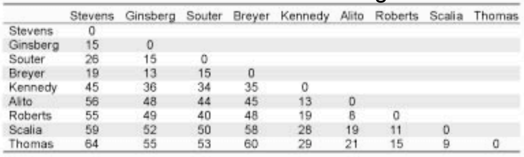
	Stevens	Ginsberg	Souter	Breyer	Kennedy	Alito	Roberts	Scalia	Thomas
Stevens	0								
Ginsberg	15	0							
Souter	26	15	0						
Breyer	19	13	15	0					
Kennedy	45	36	34	35	0				
Alito	56	48	44	45	13	0			
Roberts	55	49	40	48	19	8	0		
Scalia	59	52	50	58	28	19	11	0	
Thomas	64	55	53	60	29	21	15	9	0

Thanks to: Steve Guthery, MD



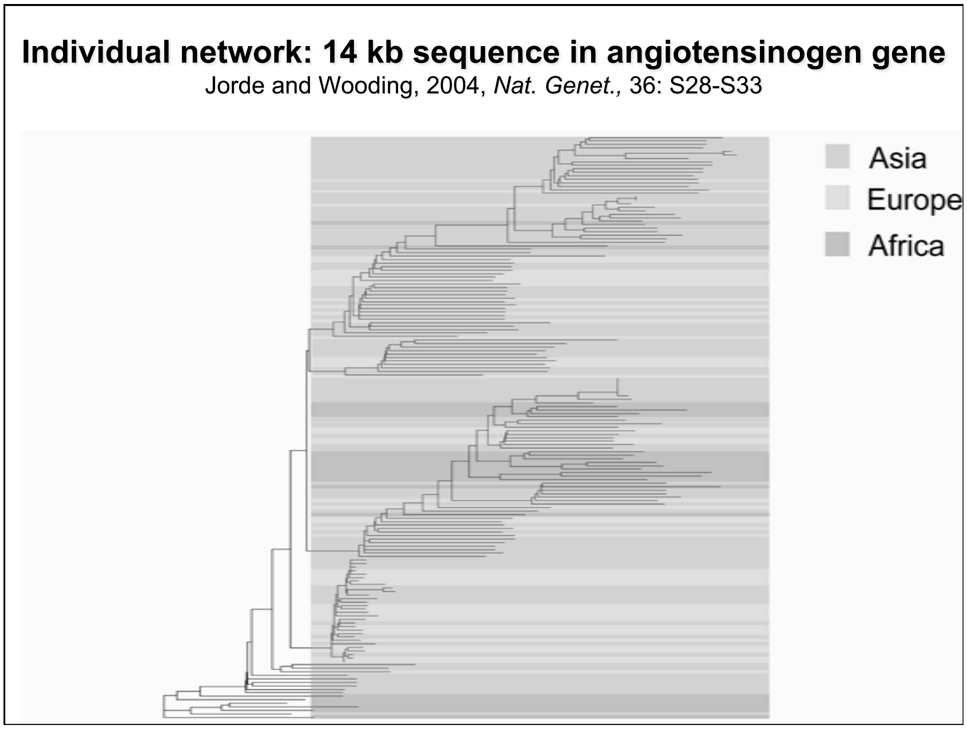
Neighbor-joining network

Distance matrix: % disagreement



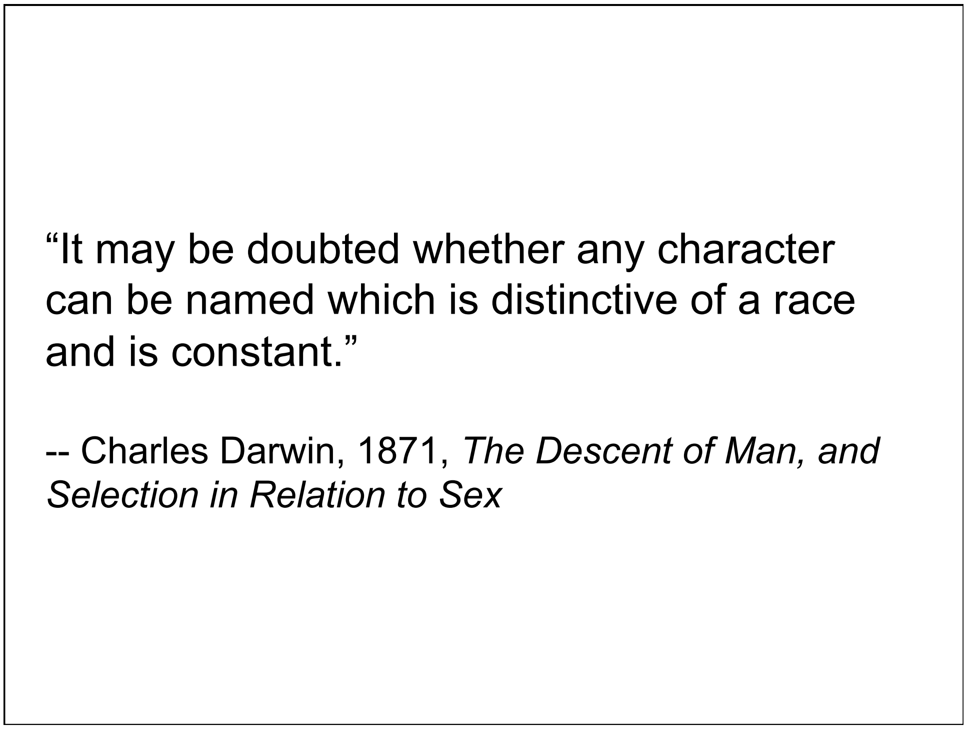
	Stevens	Ginsberg	Souter	Breyer	Kennedy	Alito	Roberts	Scalia	Thomas
Stevens	0								
Ginsberg	15	0							
Souter	26	15	0						
Breyer	19	13	15	0					
Kennedy	45	36	34	35	0				
Alito	56	48	44	45	13	0			
Roberts	55	49	40	48	19	8	0		
Scalia	59	52	50	58	28	19	11	0	
Thomas	64	55	53	60	29	21	15	9	0

Thanks to: Steve Guthery, MD



Individual network: 14 kb sequence in angiotensinogen gene

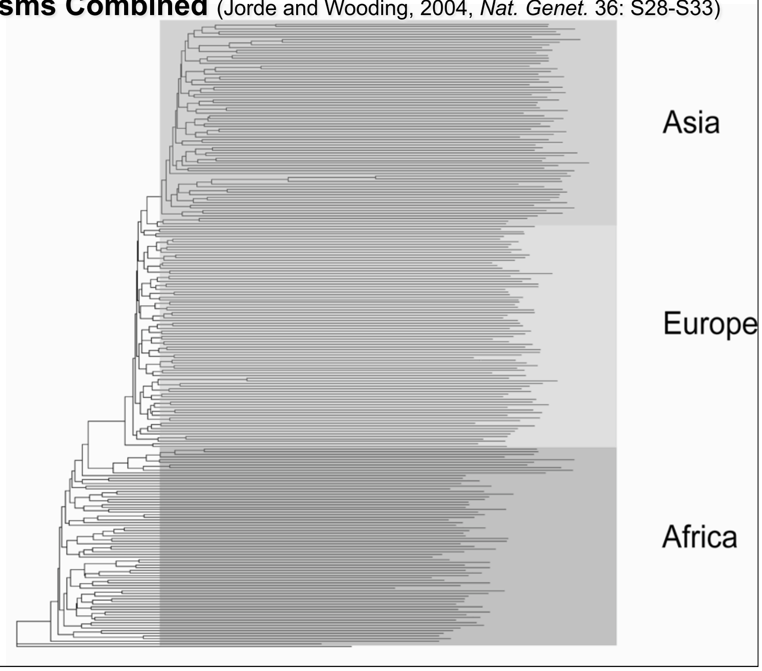
Jorde and Wooding, 2004, *Nat. Genet.,* 36: S28-S33



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

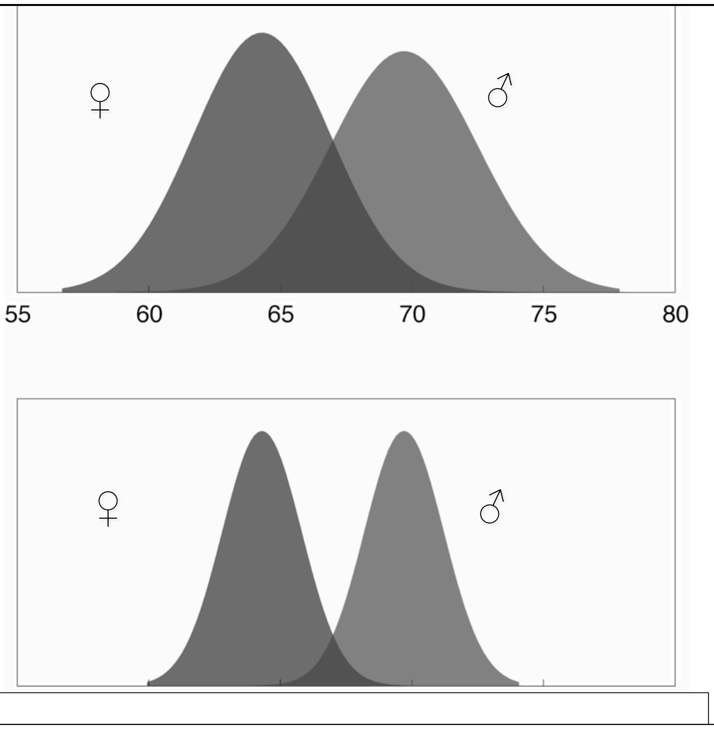
-- Charles Darwin, 1871, *The Descent of Man, and Selection in Relation to Sex*

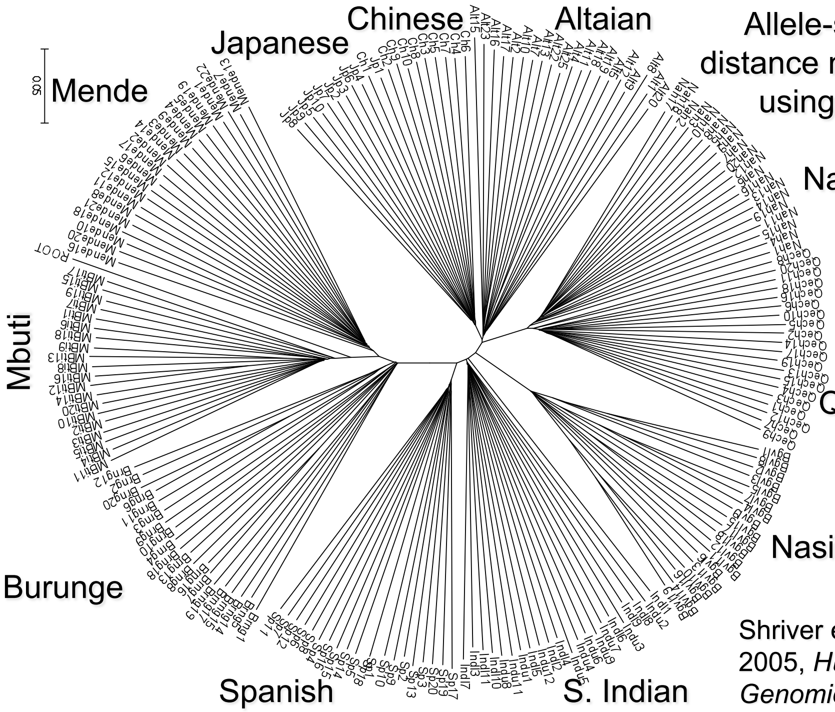
Individual Network: 190 *Alu*, STR, and Restriction Site Polymorphisms Combined (Jorde and Wooding, 2004, *Nat. Genet.* 36: S28-S33)



Height

Height + waist/hip ratio

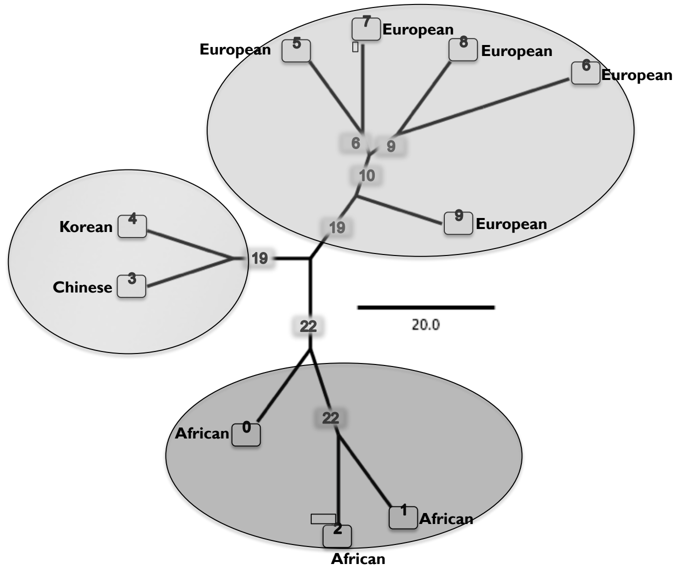




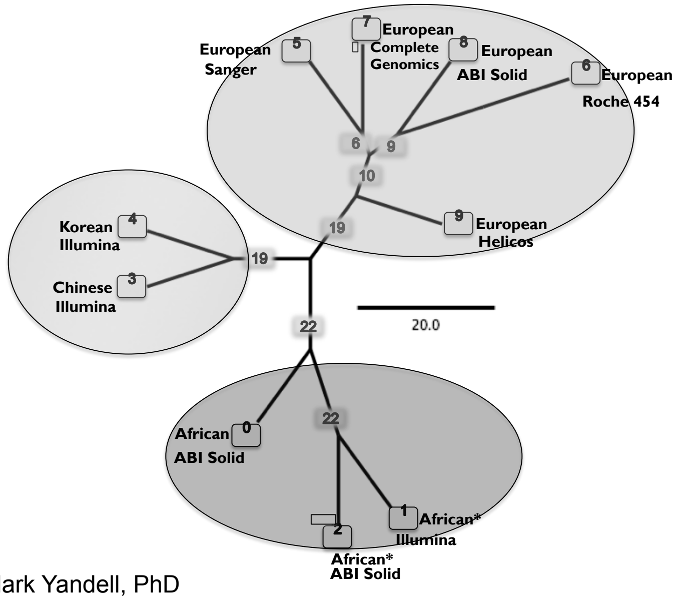
Allele-sharing distance network using 11,078 SNPs

Shriver et al., 2005, *Human Genomics* 2: 81-9

Whole-genome sequence data for 10 humans

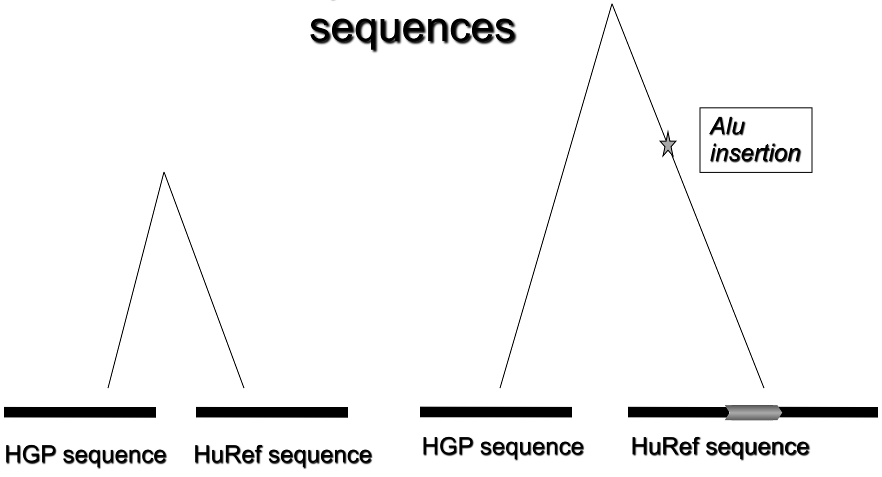


Courtesy of Mark Yandell, PhD



Courtesy of Mark Yandell, PhD

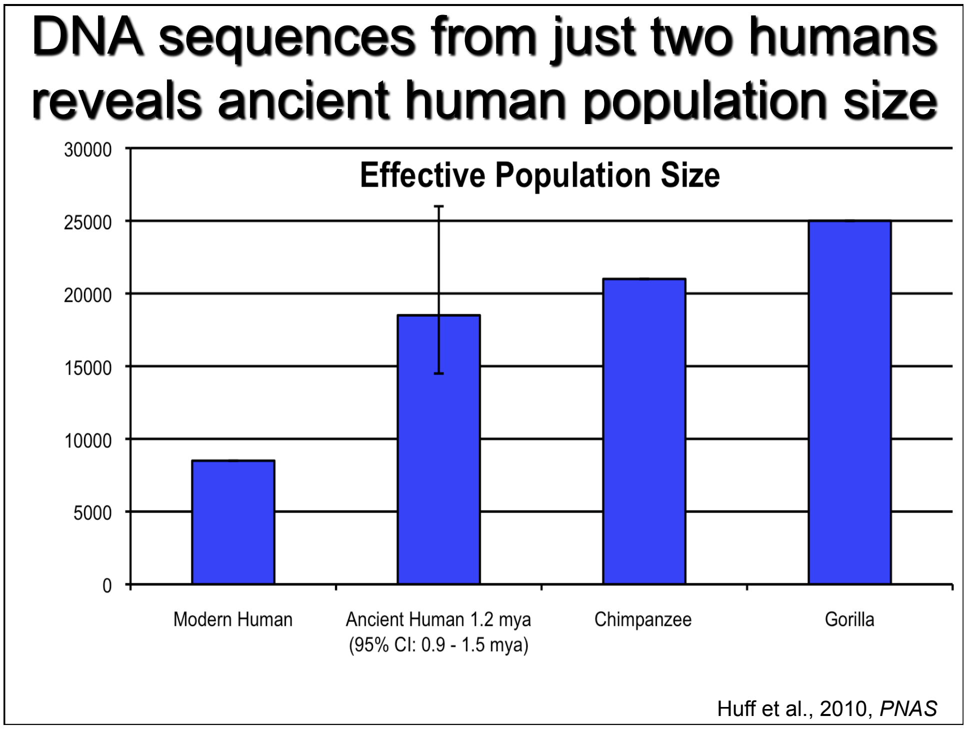
Alu insertions mark "old" genomic regions that can be compared in HuRef and HGP sequences



Average coalescence time = 462,000 years

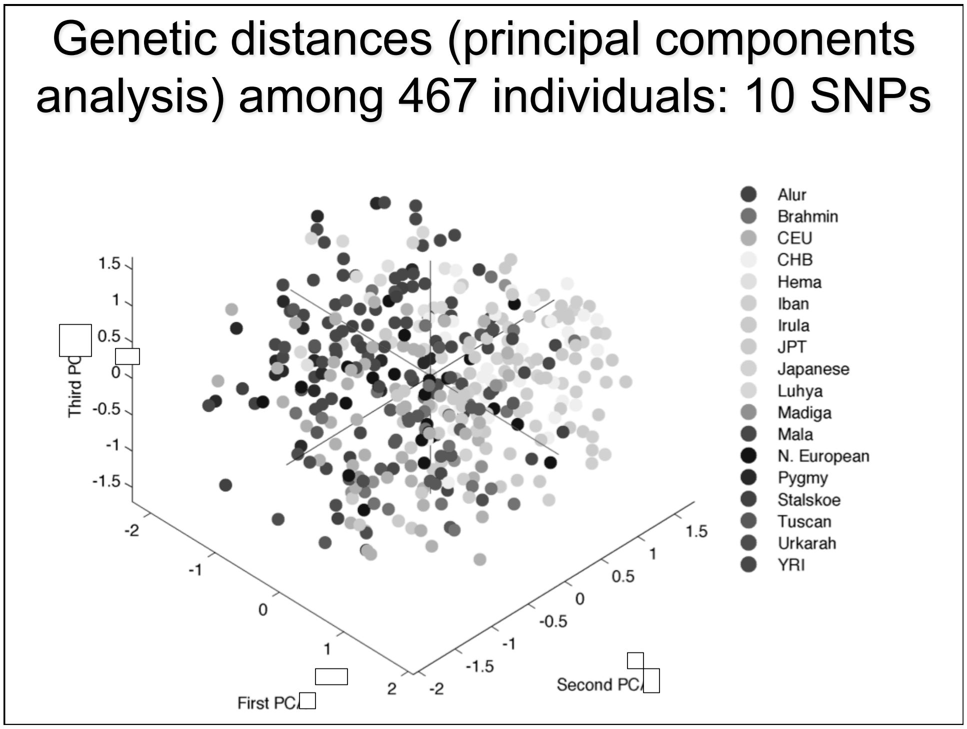
Average coalescence time = 924,000 years

Huff et al., 2010, *PNAS*

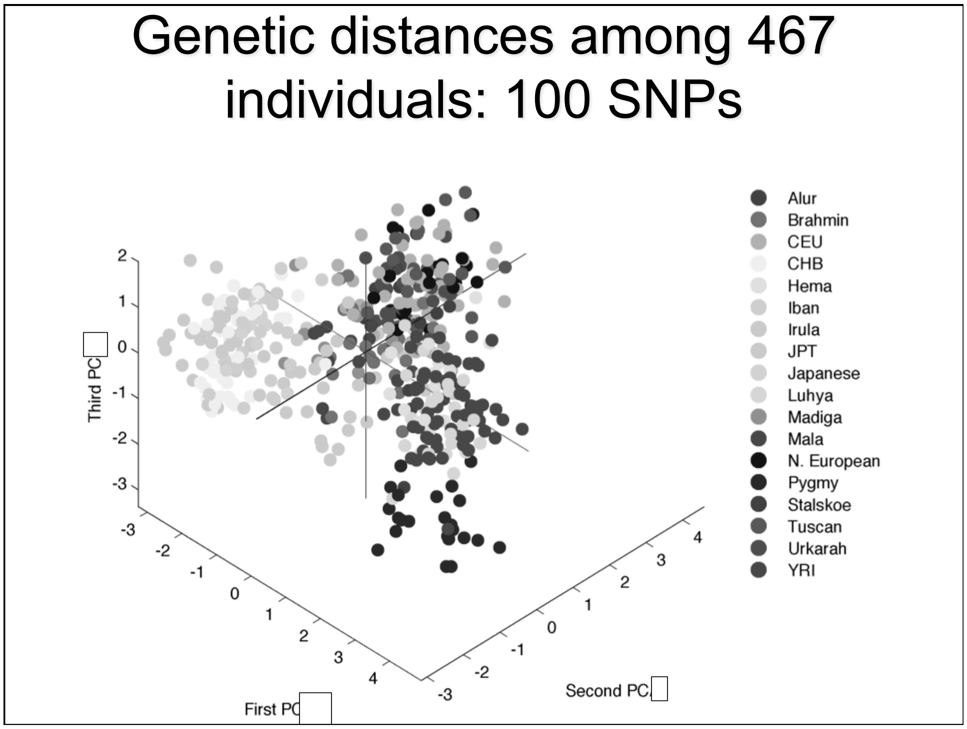


DNA sequences from just two humans reveals ancient human population size

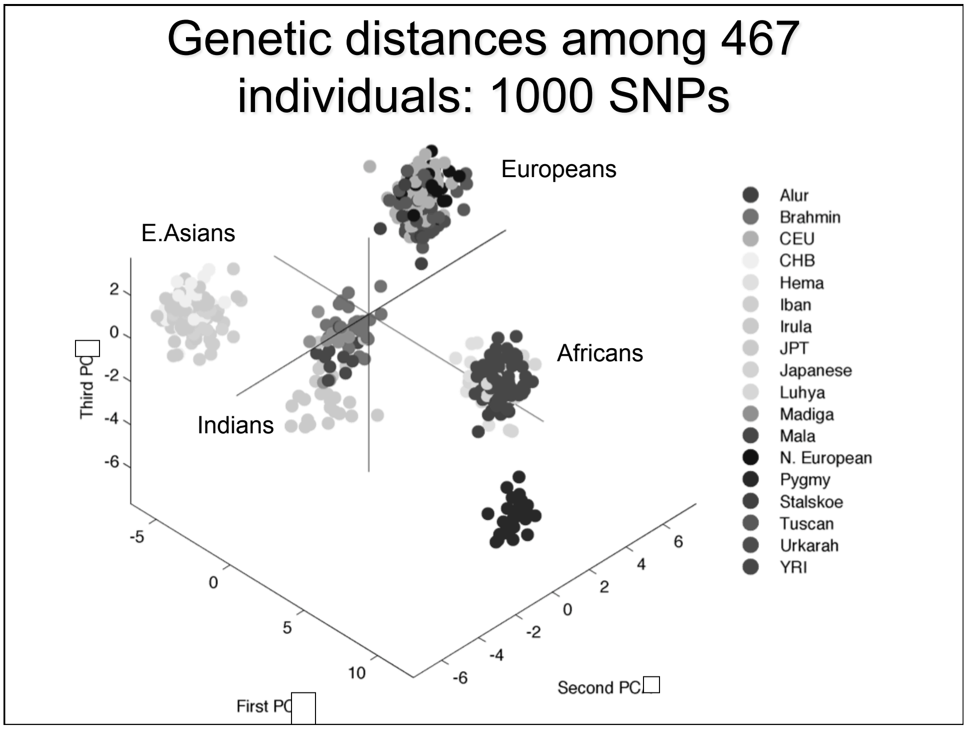
Huff et al., 2010, *PNAS*



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

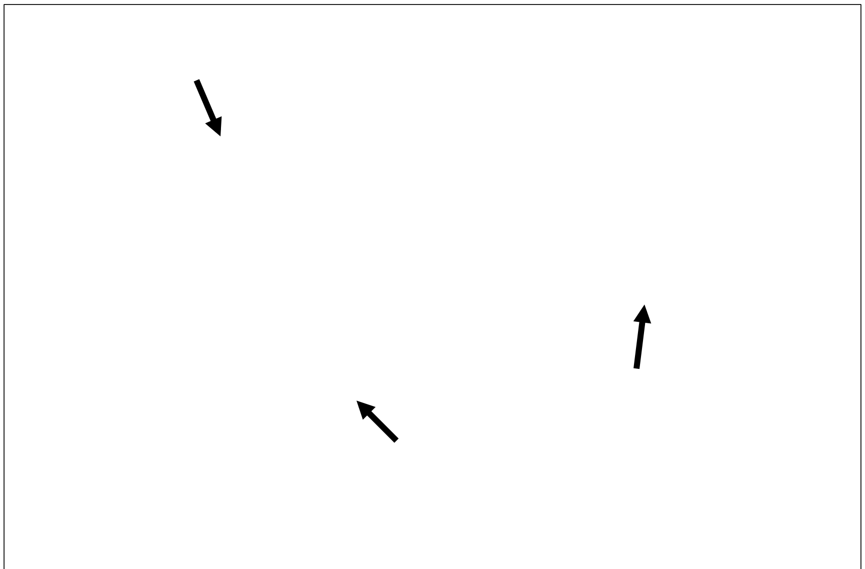


Genetic distances among 467 individuals: 100 SNPs



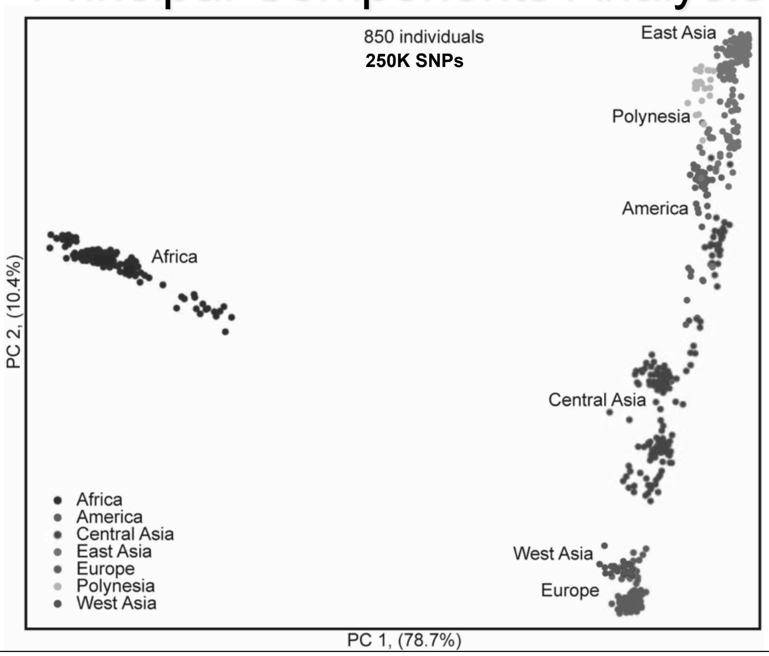
Genetic distances among 467 individuals: 1000 SNPs

Genetic distances: 250K SNPs

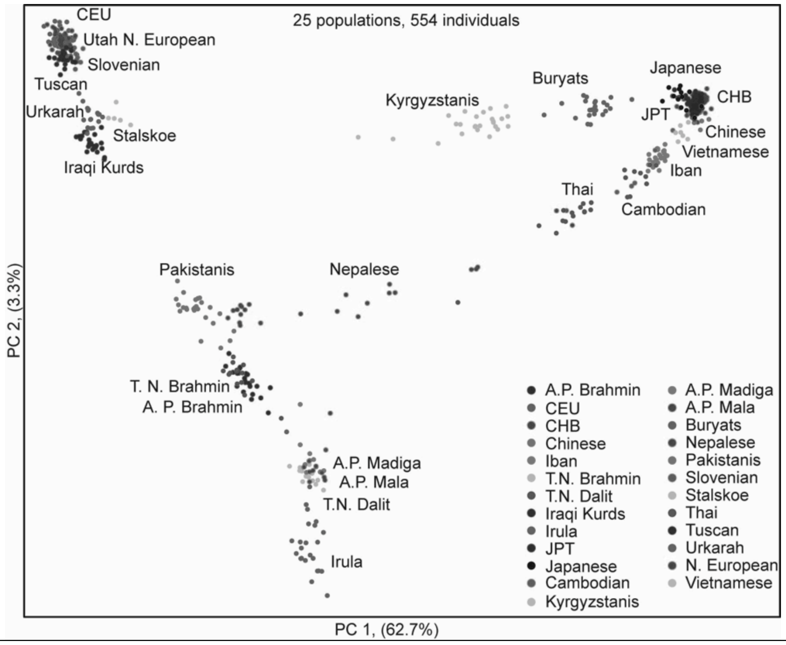


Xing et al., 2009, *Genome Research*

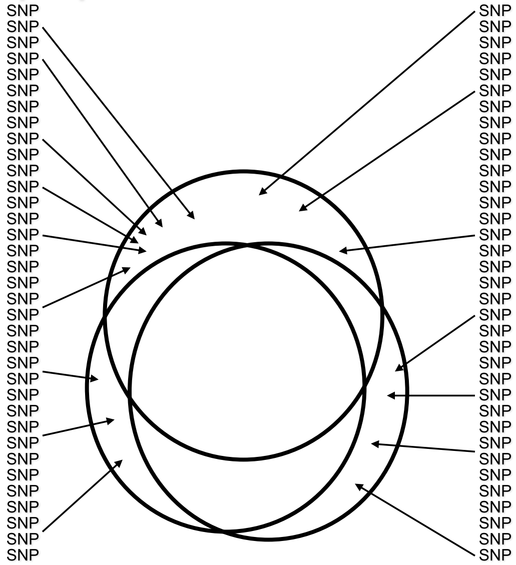
Principal Components Analysis



Eurasian Populations

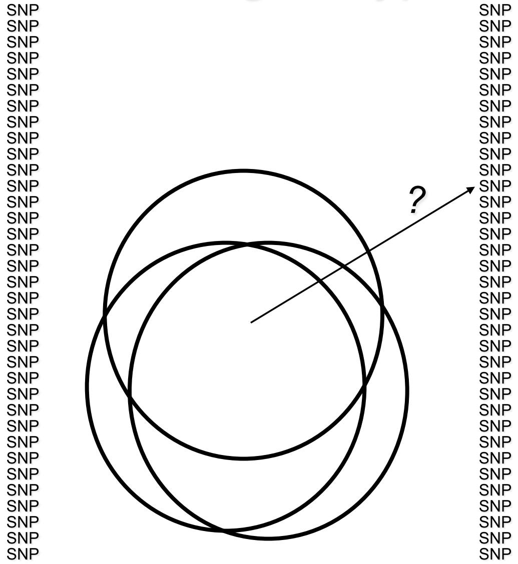


Multiple polymorphisms can predict population affiliation

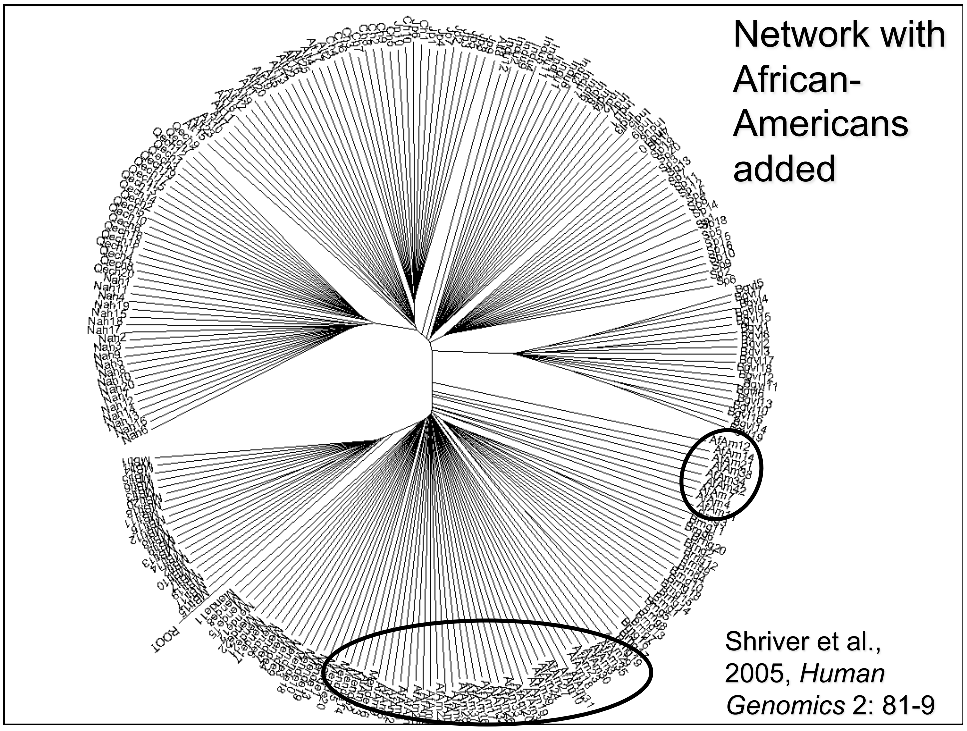


Population affiliation cannot accurately predict individual genotypes or traits

Individual traits or genotypes are shared across populations and differ only in their frequencies

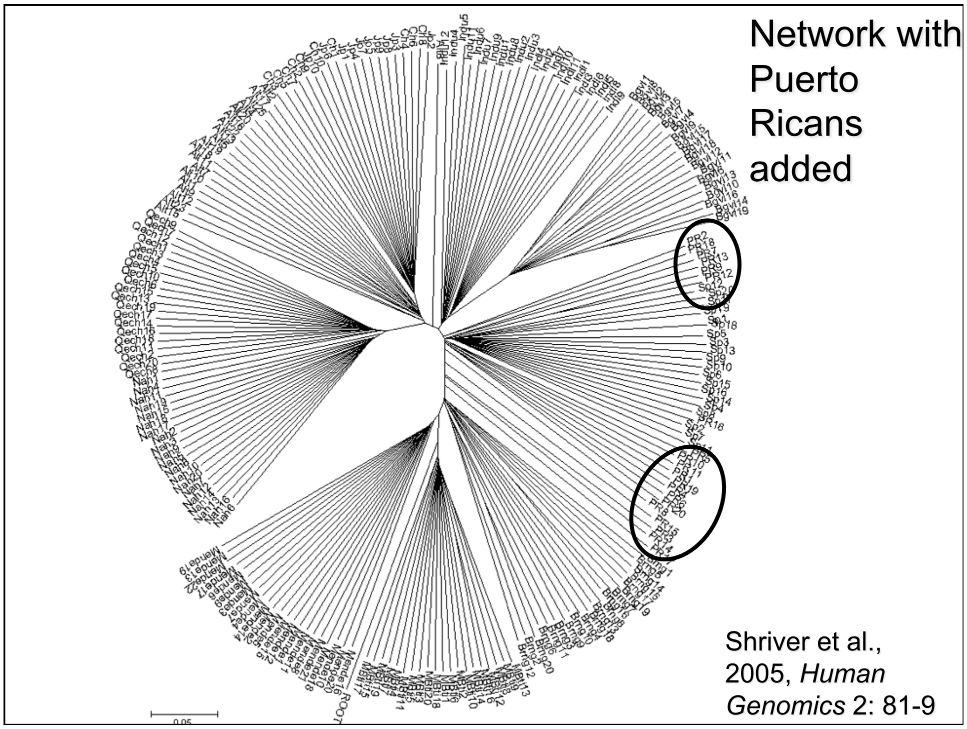


Can we classify everybody?



Network with African-Americans added

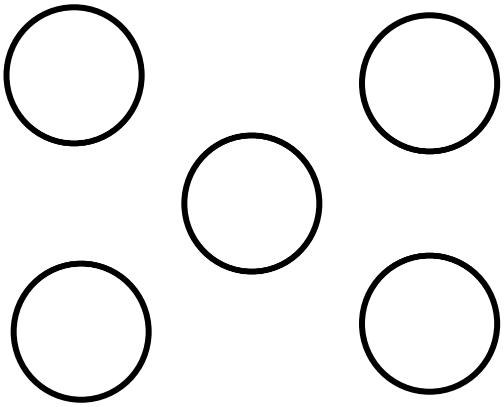
Shriver et al., 2005, *Human Genomics* 2: 81-9



Network with Puerto Ricans added

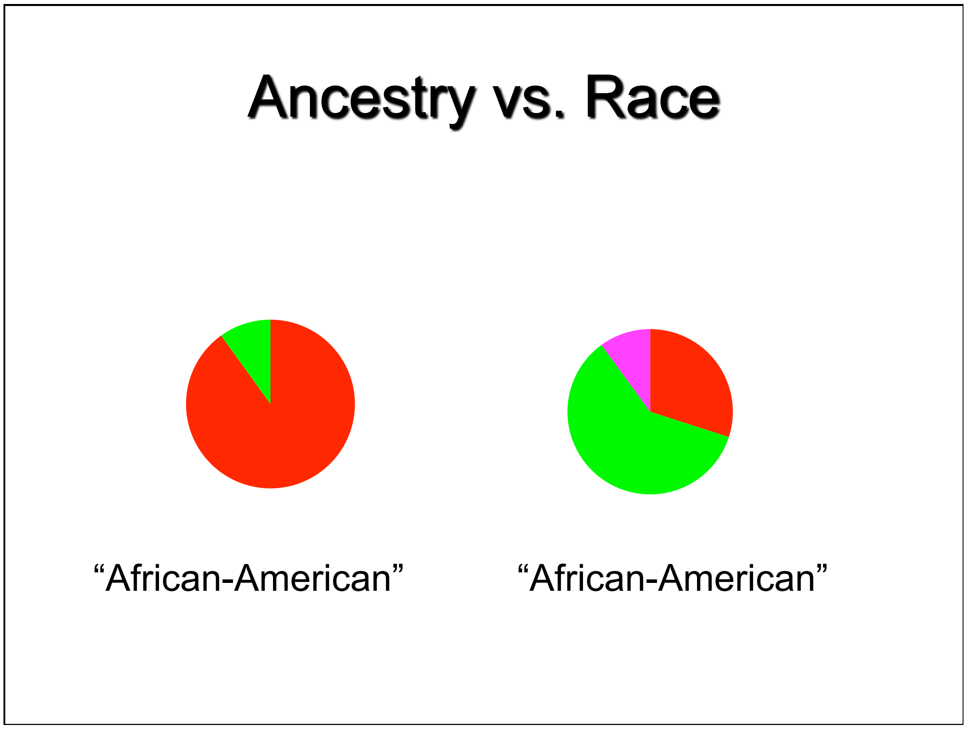
Shriver et al., 2005, *Human Genomics* 2: 81-9

The Fallacy of Typological Thinking



Race as a predictor of ancestry proportions

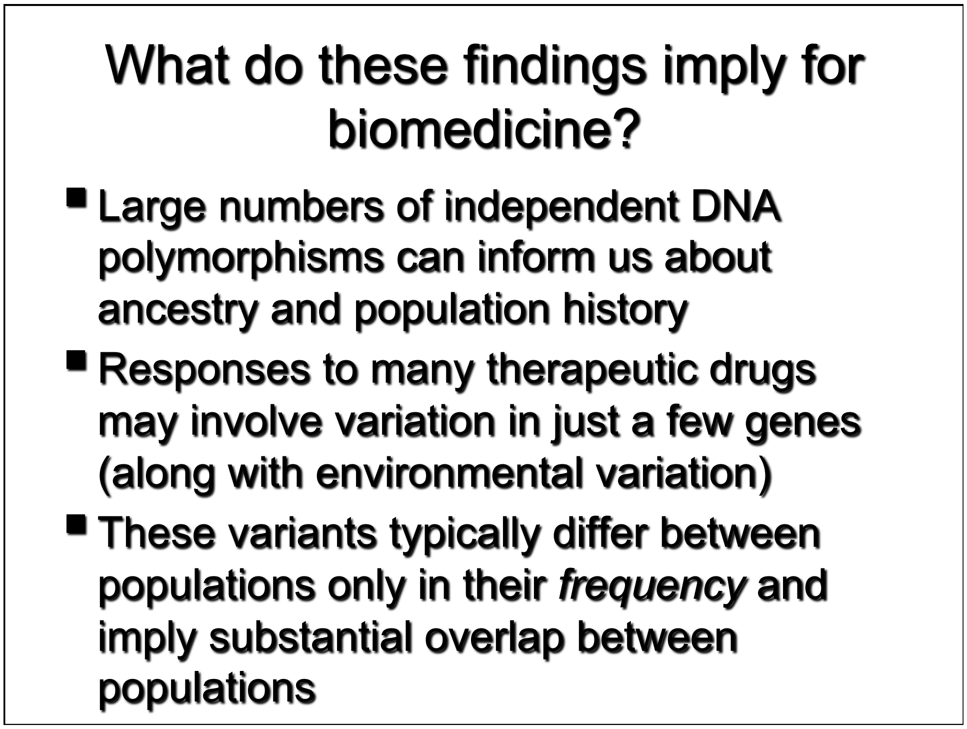
Wayne Joseph



Ancestry vs. Race

“African-American”

“African-American”

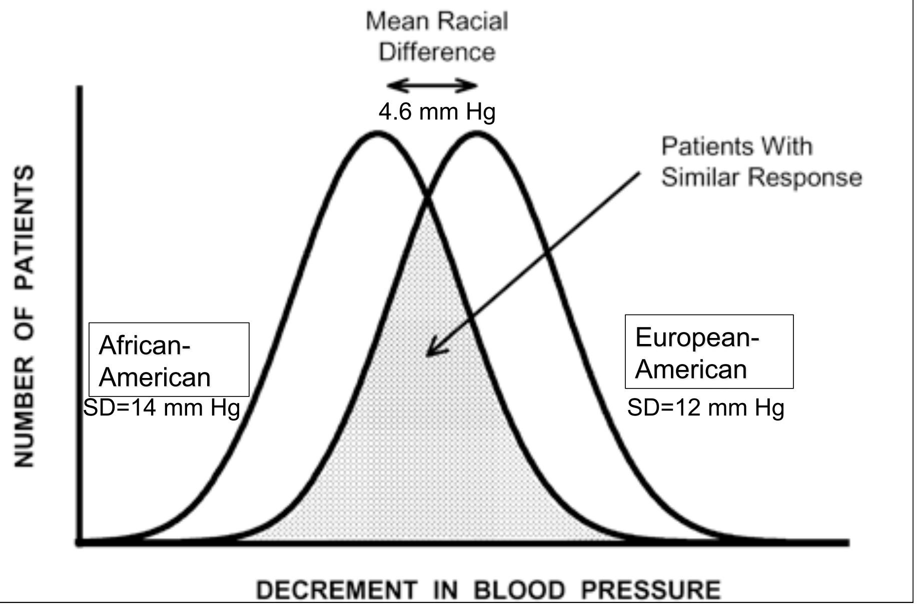


What do these findings imply for biomedicine?

- 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)

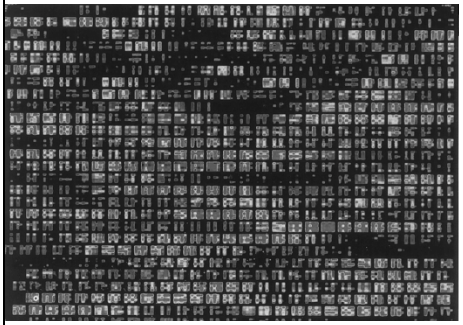


Gefitinib (Iressa) and non-small cell lung cancer

- 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

“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

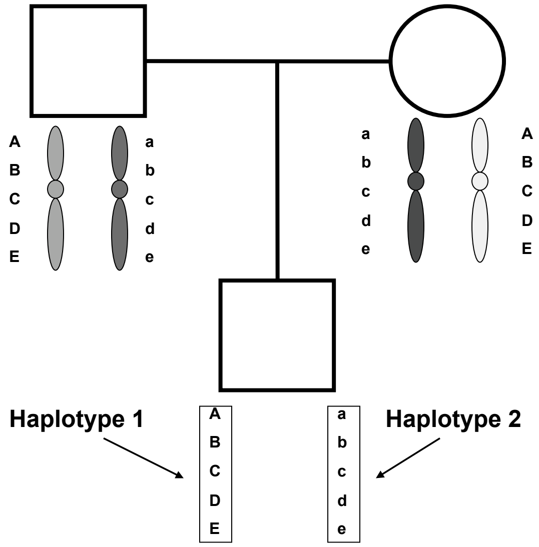
Genetic Variation and “Race”

- 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

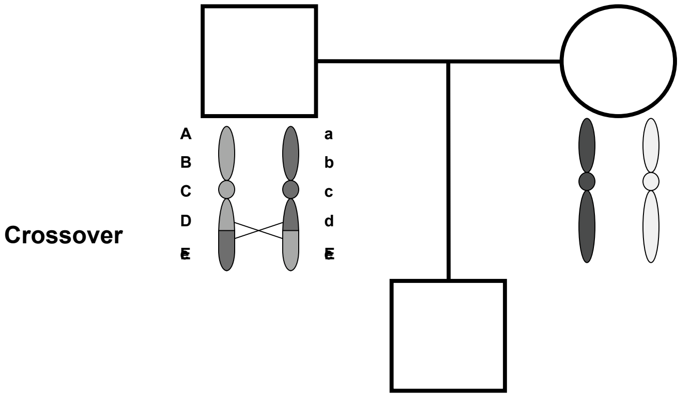
SNPs, haplotypes, linkage disequilibrium, and gene mapping

- 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)
- 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

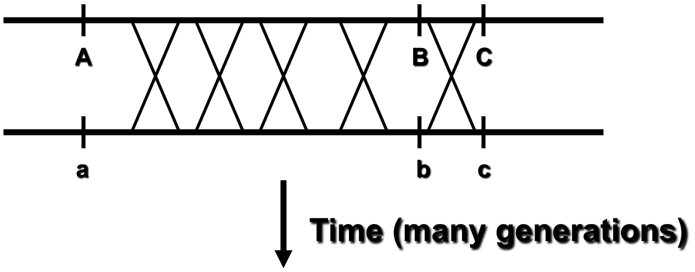


Crossovers during meiosis can create new haplotype combinations



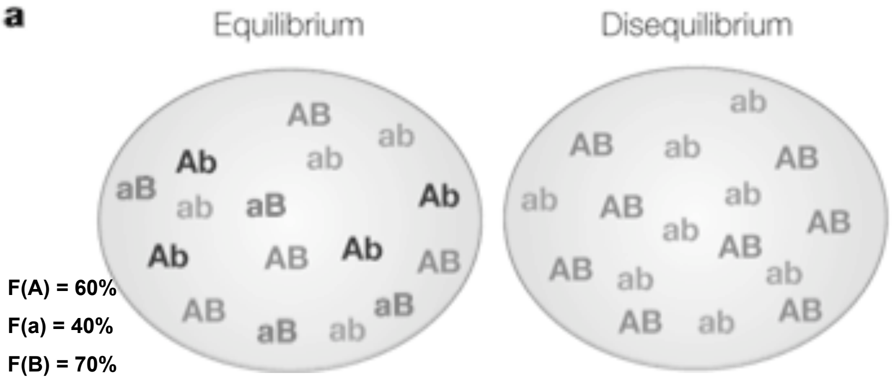
Crossovers **occur during meiosis, producing *recombination* of alleles**

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



F(A) = 60%

F(a) = 40%

F(B) = 70%

F(b) = 30%

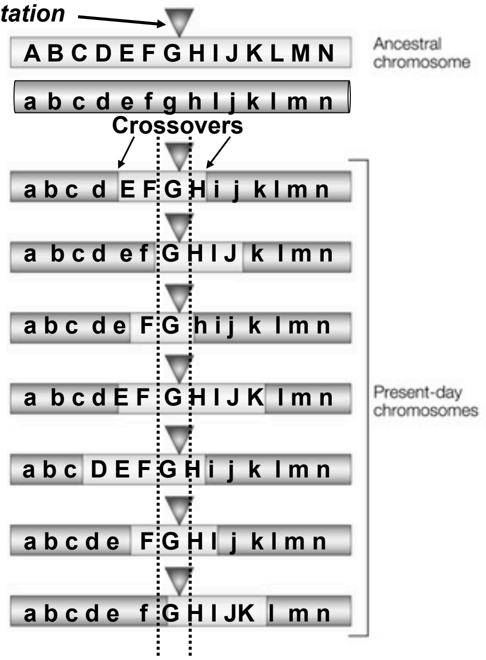
Haplotypes:

A	B	42%
A	b	18%
a	B	28%
a	b	12%

A	B	60%
a	b	30%
a	B	10%

Cystic fibrosis mutation

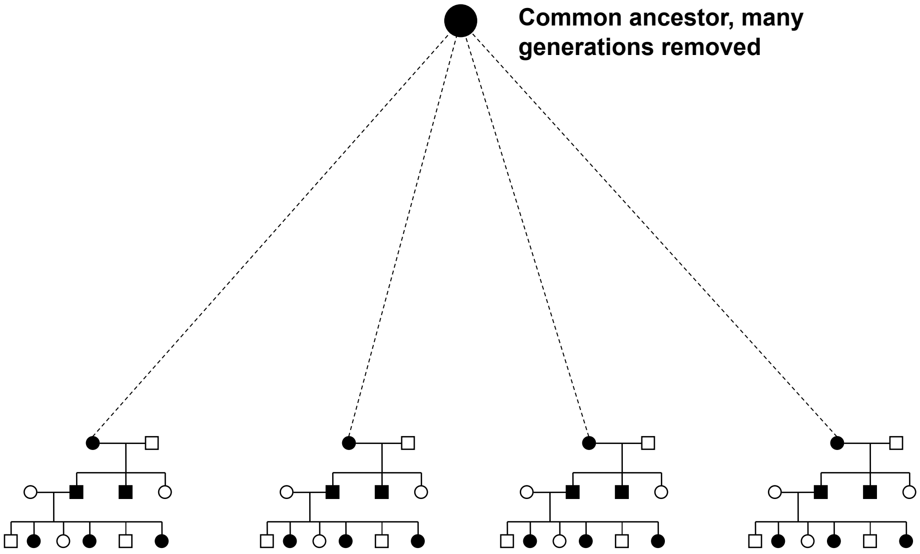
A disease-causing mutation will be *associated* with nearby polymorphisms in a population of individuals


A B C D E F G H I J K L M N
Ancestral chromosome
a b c d e f g h I j k l m n
Crossovers
a b c d E F G H i j k l m n
a b c d e f G H I J k l m n
a b c d e F G h i j k l m n
a b c d E F G H I J K l m n
a b c D E F G H i j k l m n
a b c d e F G H I j k l m n
a b c d e f G H I JK l m n
Present-day chromosomes

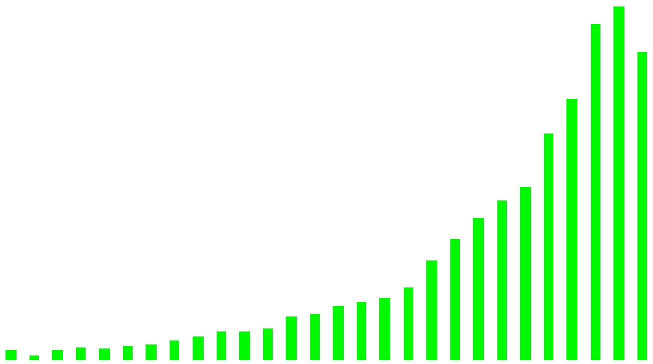
Potential advantages of linkage disequilibrium (LD)

- Family data are *not* necessarily needed
- Microarray technology now exists that allows dense genotype assays (SNPs every 3 kb)
- Association studies (linkage disequilibrium) can incorporate many past generations of recombination to narrow the candidate region

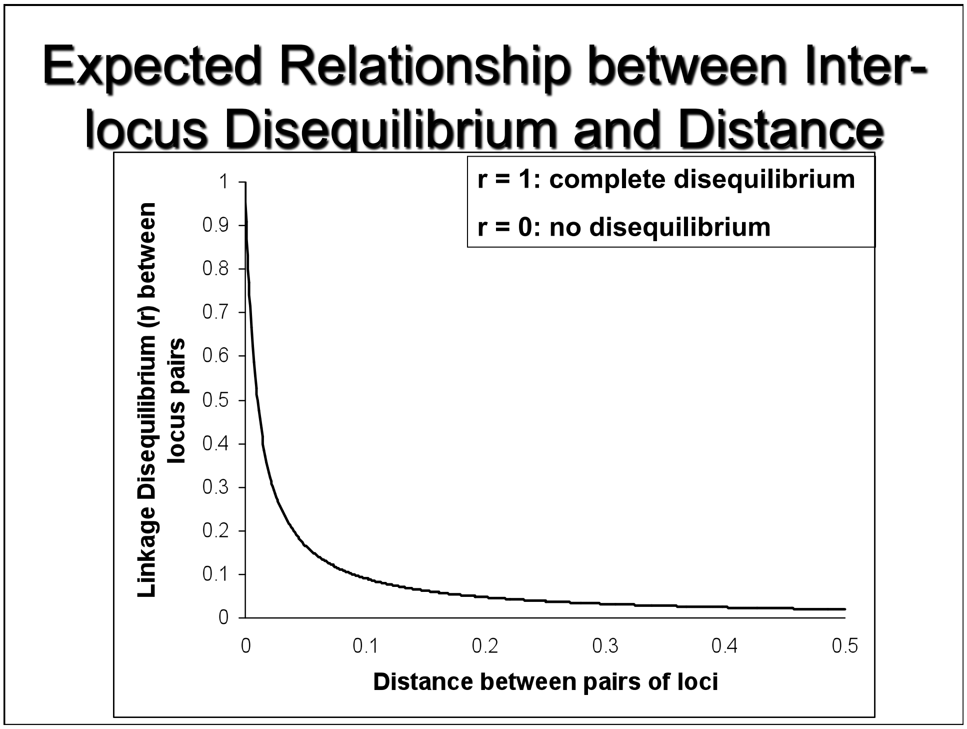
Populations are one big (complicated) pedigree


Common ancestor, many generations removed

Number of published LD articles



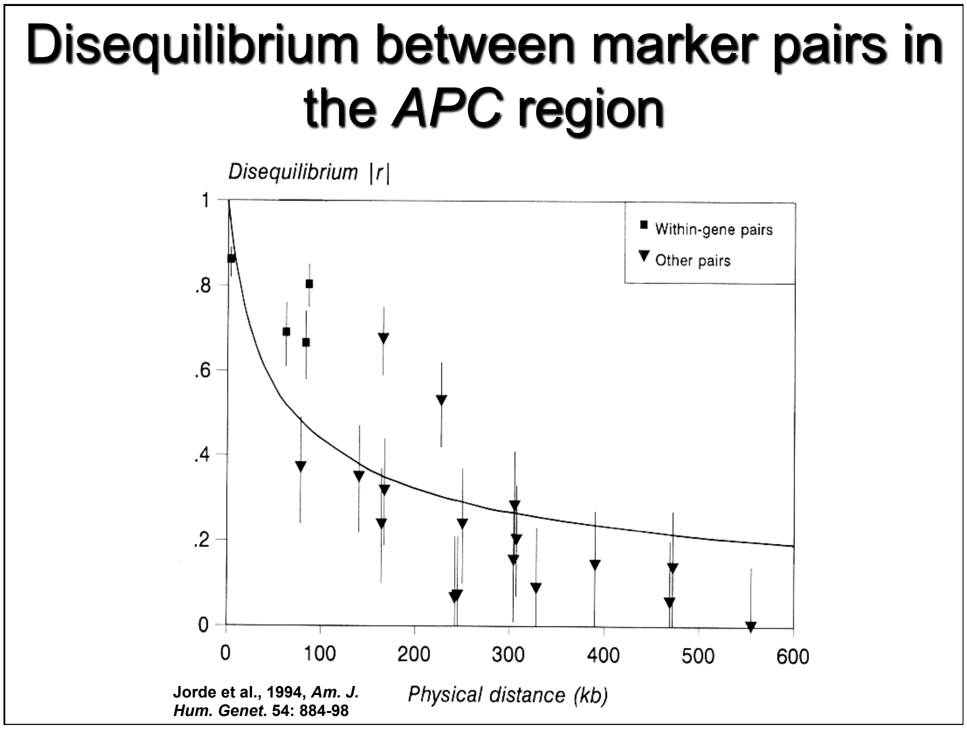
Is there a simple, uniform relationship between inter-locus physical distance and inter-locus linkage disequilibrium?



Expected Relationship between Inter-locus Disequilibrium and Distance

r = 1: complete disequilibrium

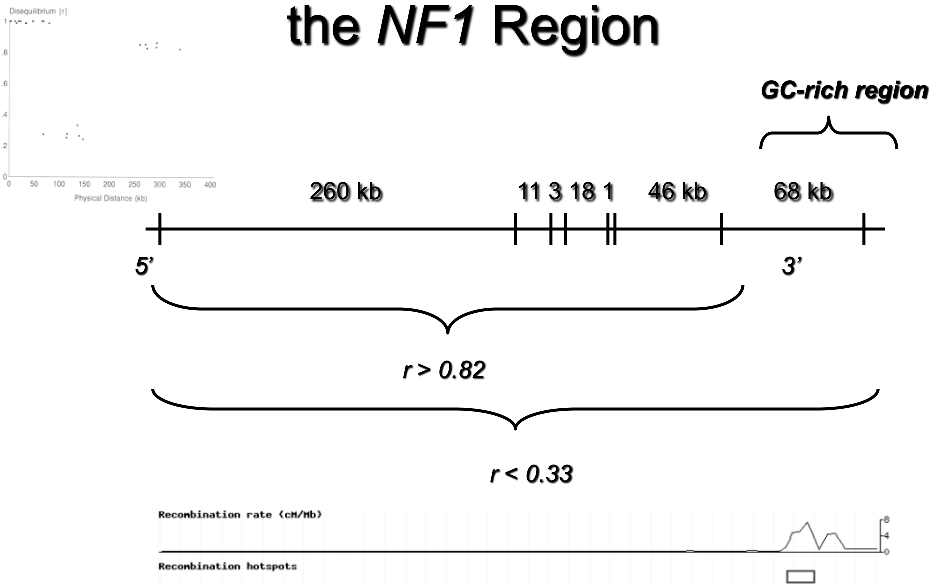
r = 0: no disequilibrium



Disequilibrium between marker pairs in the *APC* region

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

Uneven Disequilibrium Pattern in the *NF1* Region



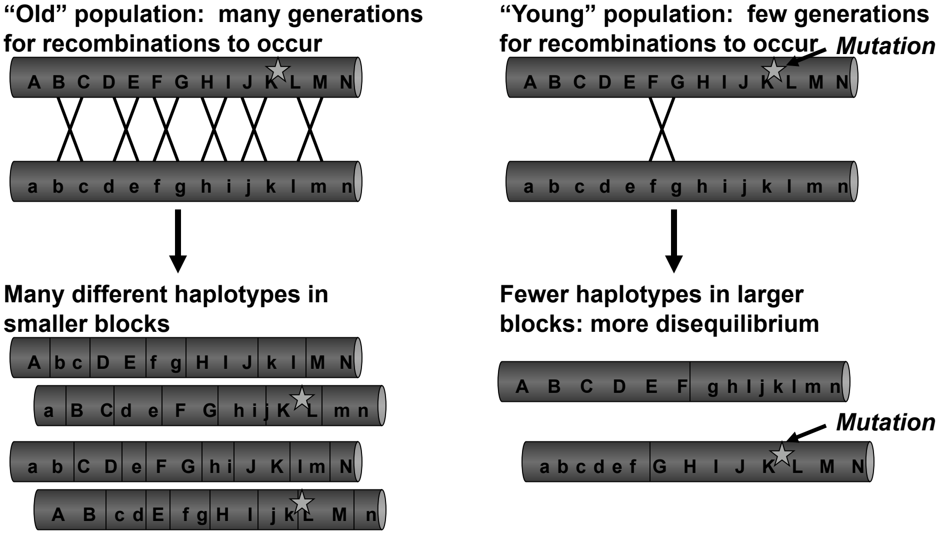
Factors that May Affect Linkage Disequilibrium Patterns

- Chromosome location
  - Telomeric vs. centromeric
  - Intragenic vs. extragenic
- DNA sequence patterns (GC content; presence of *Alu* elements)
- 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



“Old” population: many generations for recombinations to occur

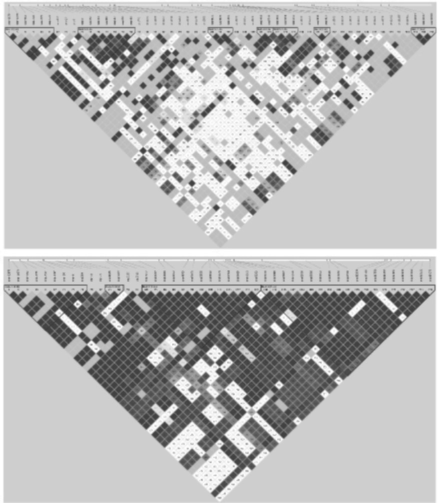
“Young” population: few generations for recombinations to occur

Many different haplotypes in smaller blocks

Fewer haplotypes in larger blocks: more disequilibrium

Mutation
A B C D E F G H I J K L M N
a b c d e f g h i j k l m n
A B C D E F G H I J K L M N
a b c d e f g h i j k l m n
A b c D E f g H I J k l M N
a B C d e F G h i j K L m n
a b C D e F G h i J K l m N
A B c d E f g H I j k L M n
A B C D E F g h I j k l m n
Mutation
a b c d e f G H I J K L M N

Pairwise LD at the Angiotensinogen locus



Africans

Eurasians

Haploview; red = high LD

How general are these patterns?

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

A Map of the World, 1544

CHARTA COSMOGRAPHICA. CVM VENTORVM PROPRIA NATVRA ET OPERATIONE.
Circius, Noordt noordtweſt. SEPTE N Septentrionalis, Noordt. TRIO Aquilo, Noordt noordt ooſt.
Zona frigida
Indie Orientalis pars
Tropicus Cancri
Æquinoctialis
TARTARIA
Scythia
ASIA
Perſi
AFRICA
AMERICA
PERV
Oceanus Indicus Meridionalis
Tropicus Hybernus
Promontorium bone Spei
Strictum Magellanicum
REGIONVM LONGITVDO AD ORTVM.
Argeſtis. Weſt noordtweſt.
OCCI Zephyrus DENS. Weſt.
Libs Weſt zuydtweſt.
Helleſpontius Ooſt noordtooſt.
ORI Subſolanus ENS. Ooſt.
Vulturnus. Ooſt zuydtooſt.
Auſtroafricus, Zuydt zuydtweſt. MERI Auſter, Zuydt DIES. Euroaſter, Zuydt zuydtooſt.
I

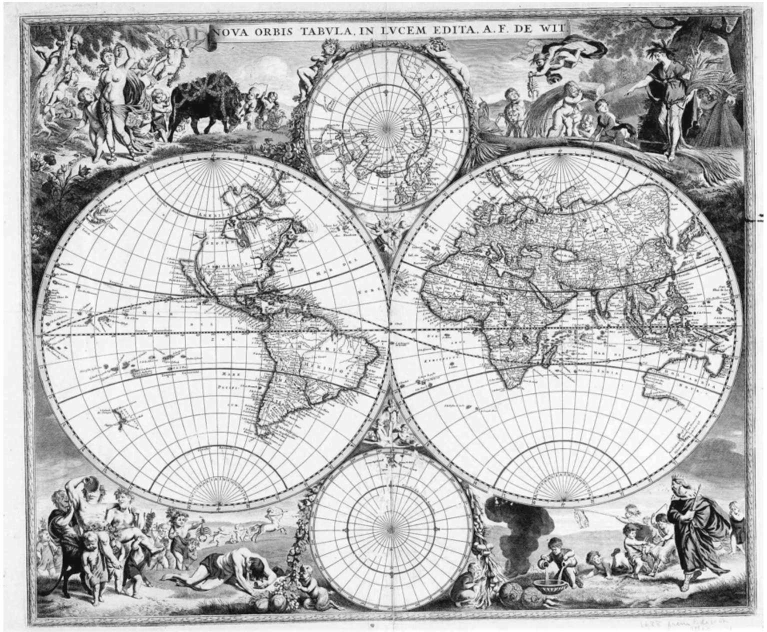
In search of a better map: The International Haplotype Map Project

- 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

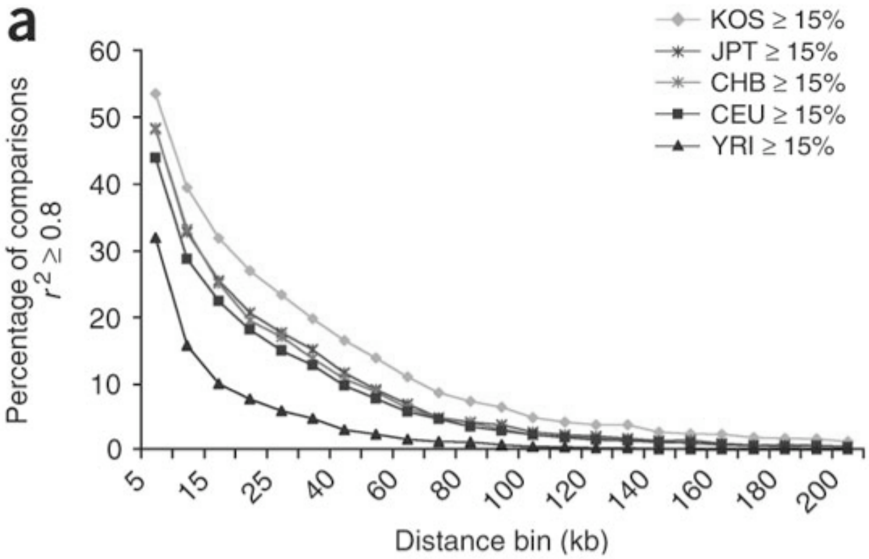
A Map of the World, 1688



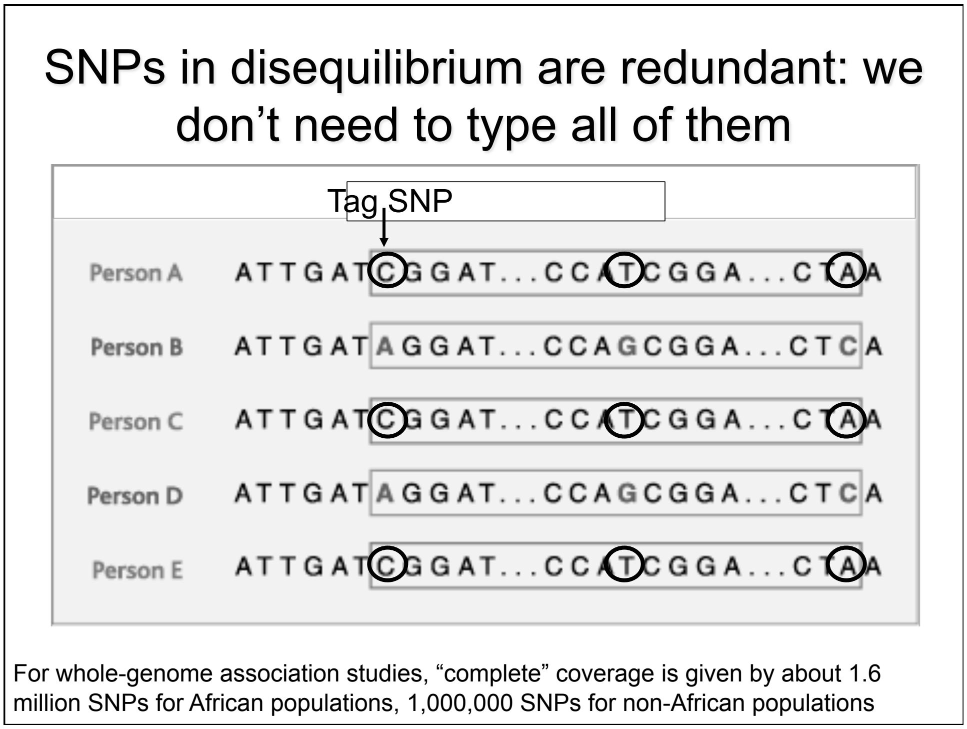
Genetic applications of HapMap

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

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



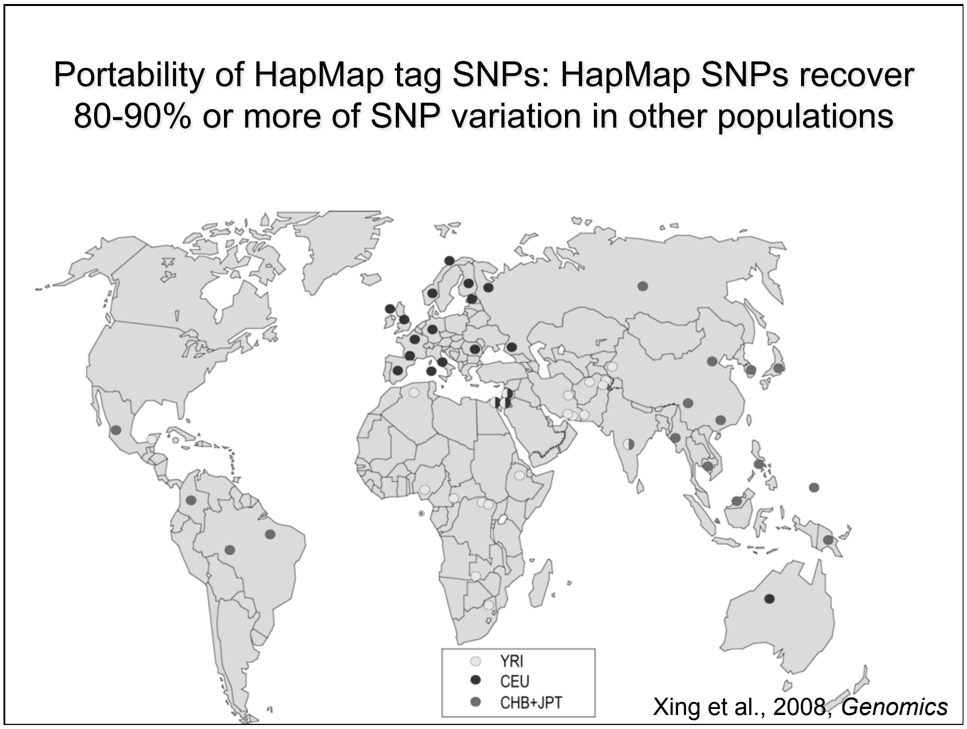
Bonnen et al., 2006, *Nat. Genet.* 38: 214-7



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

```
Person A   ATTGATCGGAT...CCATCGGA...CTAA
Person B   ATTGATAGGAT...CCAGCGGA...CTCA
Person C   ATTGATCGGAT...CCATCGGA...CTAA
Person D   ATTGATAGGAT...CCAGCGGA...CTCA
Person E   ATTGATCGGAT...CCATCGGA...CTAA
```

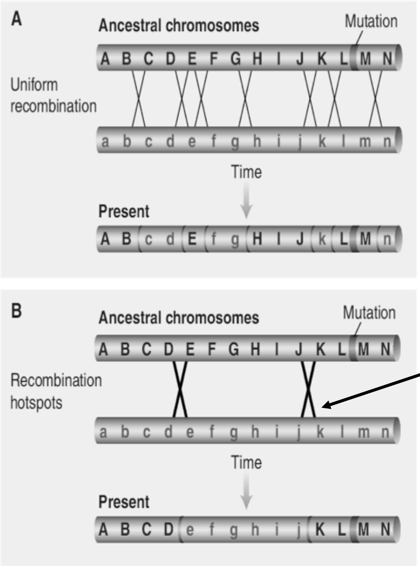
For whole-genome association studies, "complete" coverage is given by about 1.6 million SNPs for African populations, 1,000,000 SNPs for non-African populations



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

Xing et al., 2008, *Genomics*

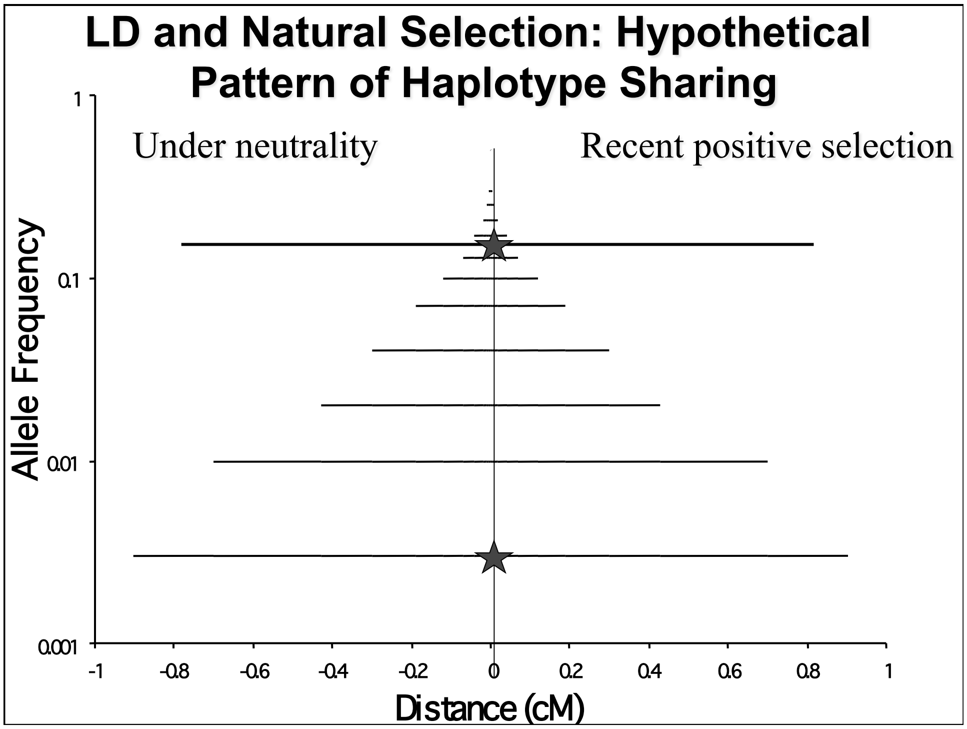
Recombination hotspots and haplotype blocks


A
Ancestral chromosomes
Mutation
A B C D E F G H I J K L M N
Uniform recombination
a b c d e f g h i j k l m n
Time
Present
A B c d E f g H I J k L M n
B
Ancestral chromosomes
Mutation
A B C D E F G H I J K L M N
Recombination hotspots
a b c d e f g h i j k l m n
Time
Present
A B C D e f g h i j K L M N

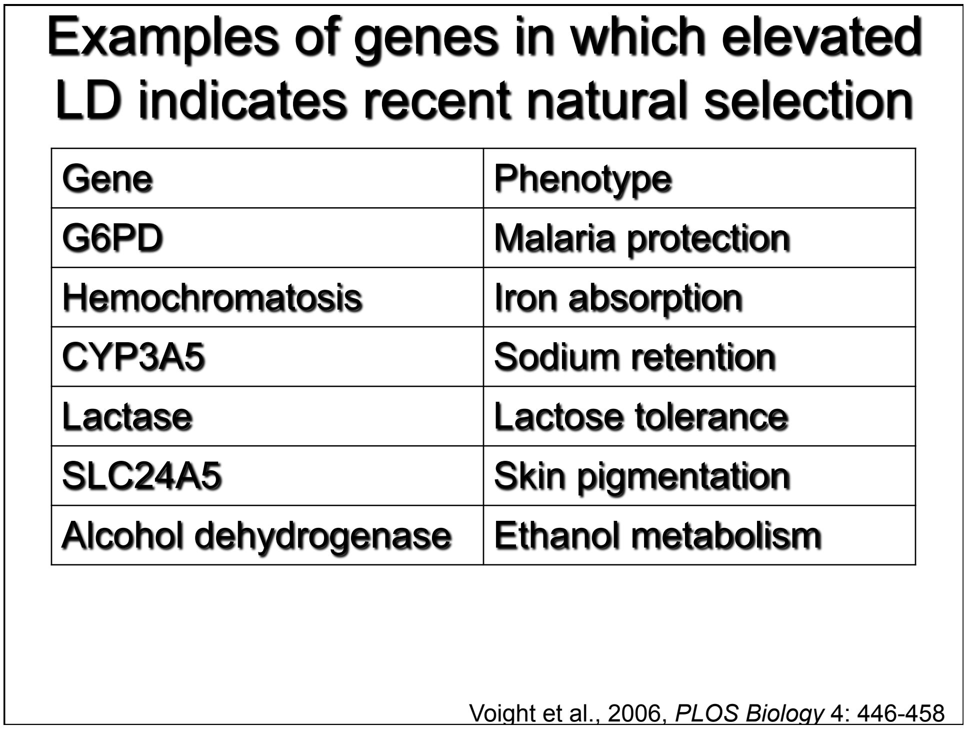
Recombination elevated at least 10X in a 1-2 kb region

Recombination hotspots

- LD patterns indicate 25,000 - 50,000 hotspots in human genome (1 every 50 – 100 kb) (Myers et al., 2005, *Science;* Coop et al., 2008, *Science*)
- 60% of crossovers occur in only 10% of the genome
- Hotspots are not congruent in human and chimpanzee, despite 99% sequence identity: suggests hotspots evolve rapidly and may not be sequence-dependent



LD and Natural Selection: Hypothetical Pattern of Haplotype Sharing



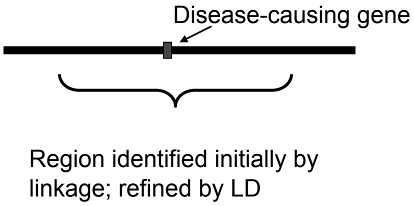
Examples of genes in which elevated LD indicates recent natural selection

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

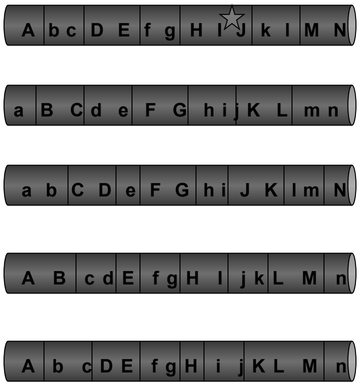
Voight et al., 2006, *PLOS Biology* 4: 446-458

Linkage disequilibrium and single-gene diseases: many successes in fine-mapping disease-causing genes

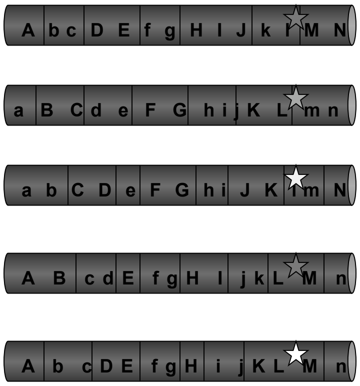
- Cystic fibrosis
- Hemochromatosis
- Wilson disease
- Friedreich's ataxia
- Bloom syndrome
- Werner syndrome
- Progressive myoclonus epilepsy
- Torsion dystonia
- Diastrophic dysplasia (and many other "Finnish" diseases)



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

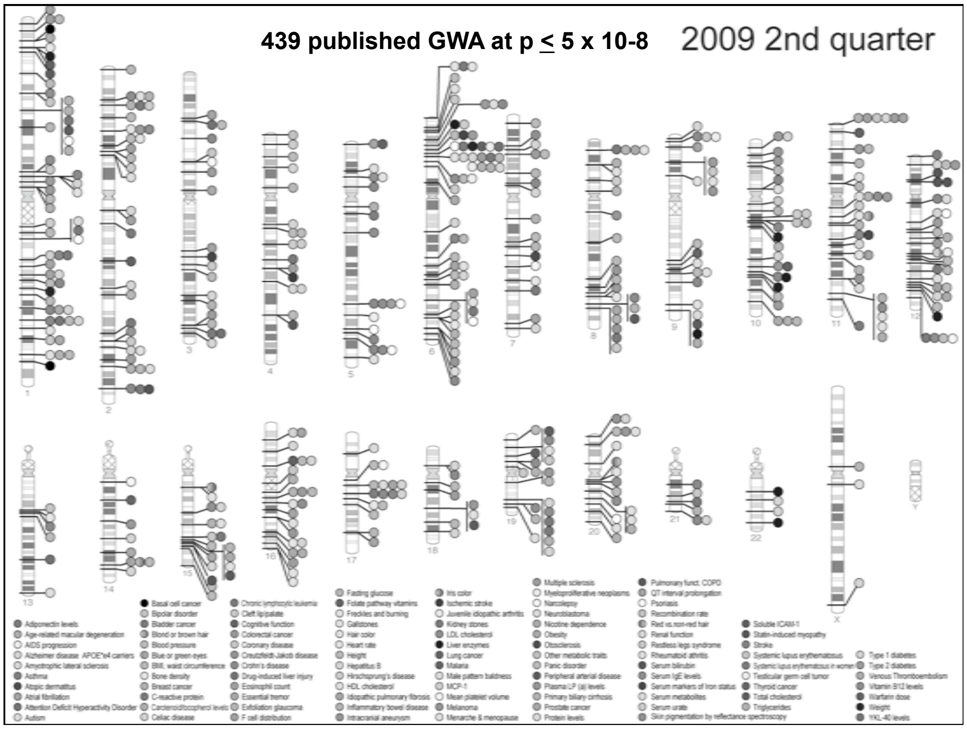

A b c D E f g H I J k l M N
a B C d e F G h i j K L m n
a b C D e F G h i J K l m N
A B c d E f g H I j k L M n
A b c D E f g H i j K L M n

Multiple disease-causing mutations can pose problems for association analysis


A b c D E f g H I J k l M N
a B C d e F G h i j K L m n
a b C D e F G h i J K l m N
A B c d E f g H I j k L M n
A b c D E f g H i j K L M n

How can we reduce heterogeneity and enhance a genetic signal?

- Define the trait consistently and accurately
- Identify subtypes
  - Early onset
  - Severe expression
  - Atypical expression
- Use strict, narrow population definitions, based on known evolutionary history
  - Population isolates may have reduced haplotype diversity and environmental heterogeneity



439 published GWA at p ≤ 5 x 10-8 2009 2nd quarter

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 and its application in disease-gene mapping
- Population genetics is *fun*!

Acknowledgments

University of Utah: Jinchuan Xing, Dave Witherspoon, Chad Huff, Tatum Simonson, Steve Guthery, Scott Watkins, Yuhua Zhang, Liz Marchani, Bob Weiss, Steve Wooding, Alan Rogers

LSU: Mark Batzer

Sorenson Molecular Genealogy Foundation: Scott Woodward, Edgar Gomez