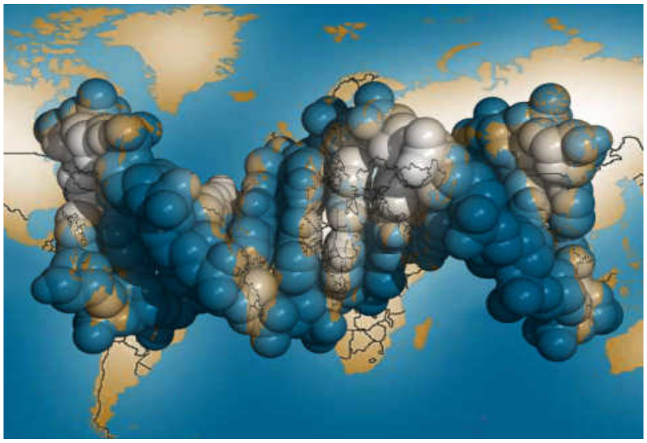
Introduction to Population Genetics



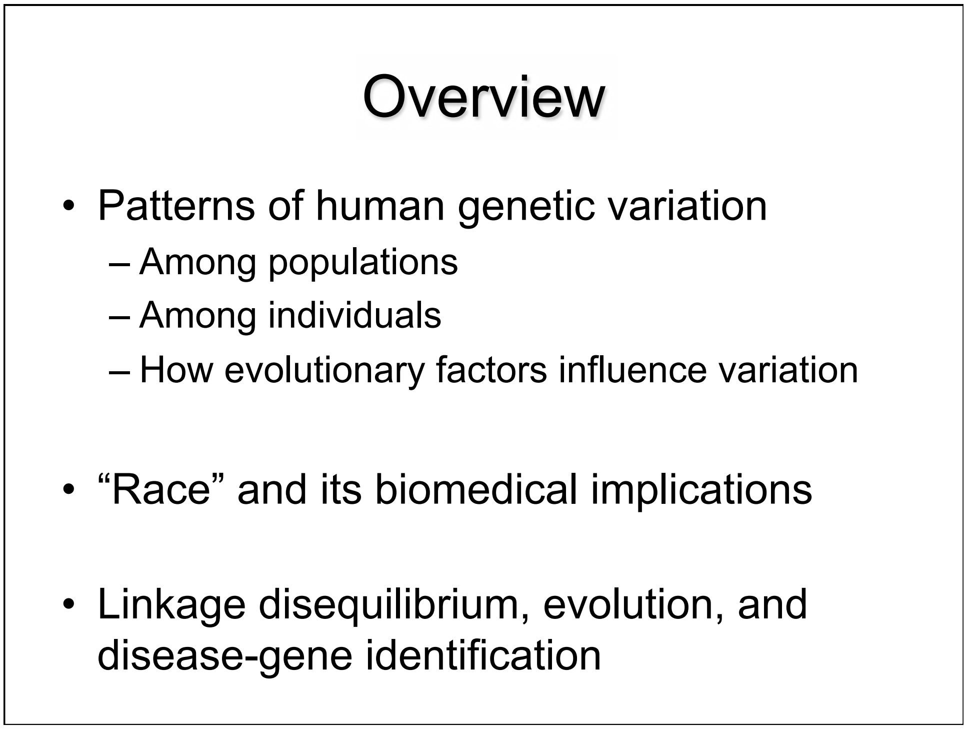
Lynn B. Jorde
H.A. and Edna Benning Presidential Professor and Chair
Department of Human Genetics
University of Utah School of Medicine
6 April 2016

JOHNS HOPKINS
MEDICINE
CONTINUING MEDICAL EDUCATION

Current Topics in Genome Analysis 2016

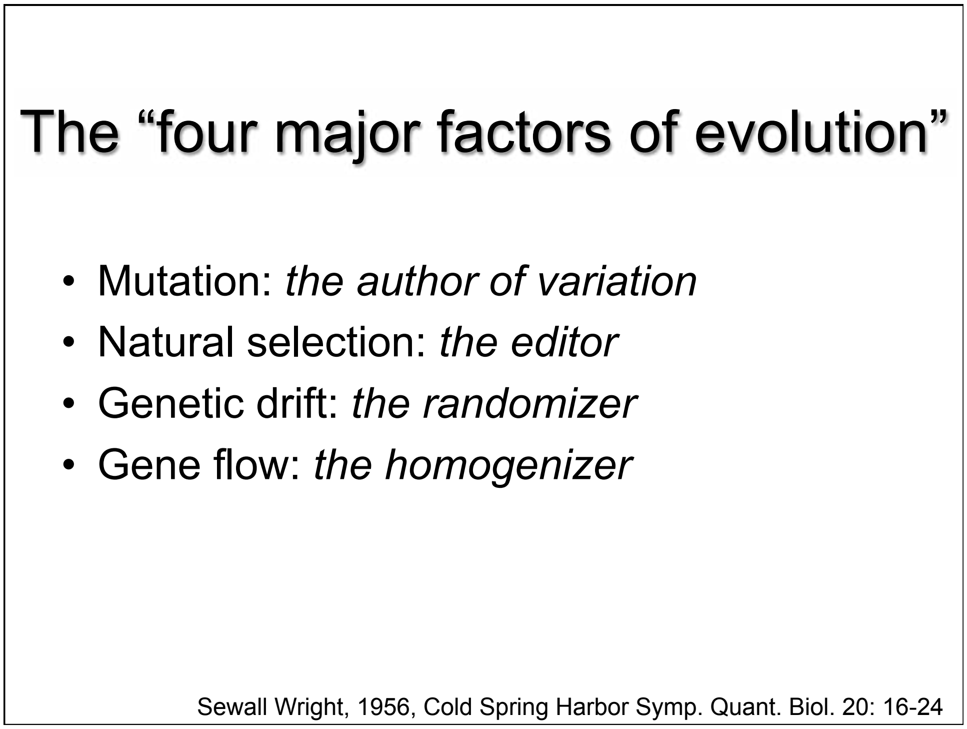
Lynn Jorde

No Relevant Financial Relationships with Commercial Interests



Overview

- Patterns of human genetic variation
  - Among populations
  - Among individuals
  - How evolutionary factors influence variation
- “Race” and its biomedical implications
- Linkage disequilibrium, evolution, and disease-gene identification



The “four major factors of evolution”

- Mutation: *the author of variation*
- Natural selection: *the editor*
- Genetic drift: *the randomizer*
- Gene flow: *the homogenizer*

Sewall Wright, 1956, Cold Spring Harbor Symp. Quant. Biol. 20: 16-24

Mutation and Genetic Variation

Human mutation rate is 1.0 – 1.5 x 10^{-8} per bp per generation: we transmit ~30 new DNA variants with each gamete

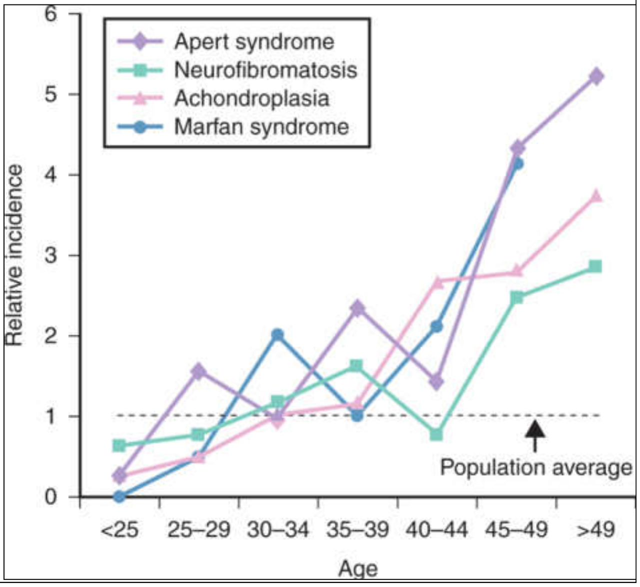
(J. Roach *et al*., 2010, *Science;* D. Conrad *et al*., 2011, *Nature Genetics*)

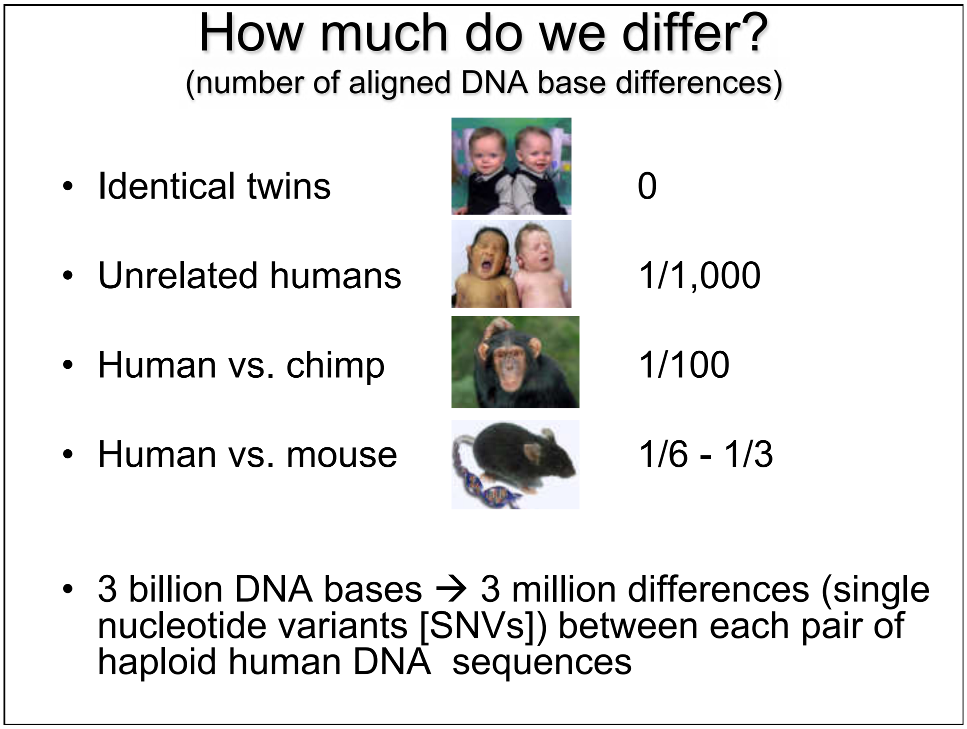
> "*The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music.*"
>
> - Lewis Thomas

Single-gene mutations increase with paternal age: at least 75% of new mutations occur in male germline

An additional two mutations occur with each year of paternal age (baseline: ~30 mutations in a male aged 30)

(Kong et al., 23 Aug. 2012, *Nature*)



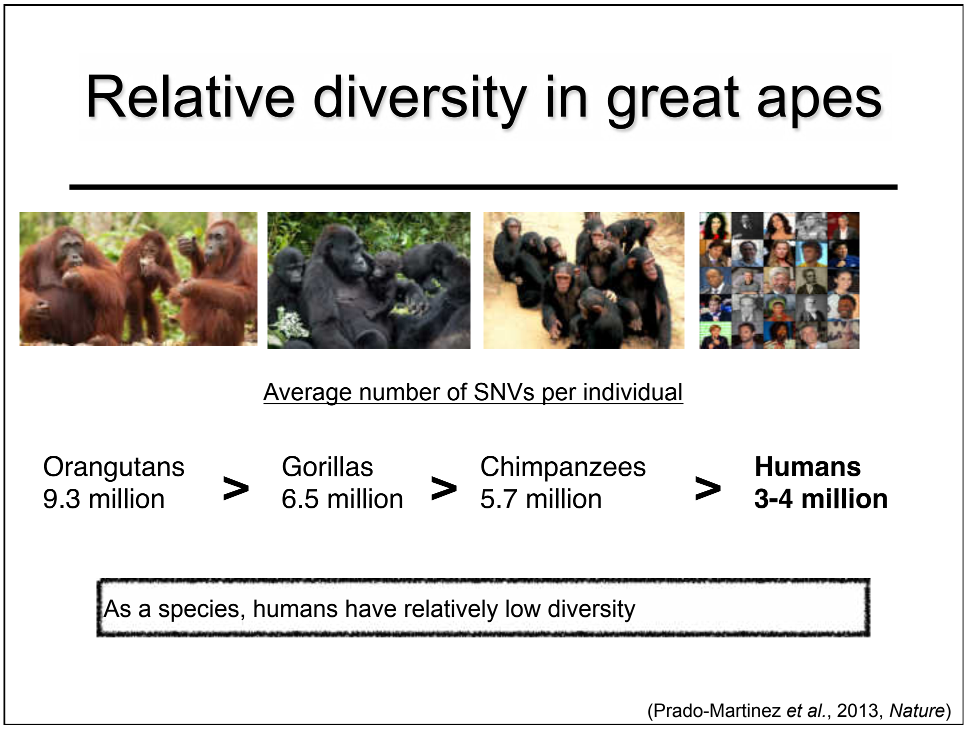


How much do we differ?

(number of aligned DNA base differences)

Identical twins	0
Unrelated humans	1/1,000
Human vs. chimp	1/100
Human vs. mouse	1/6 - 1/3

- 3 billion DNA bases → 3 million differences (single nucleotide variants [SNVs]) between each pair of haploid human DNA sequences



Relative diversity in great apes

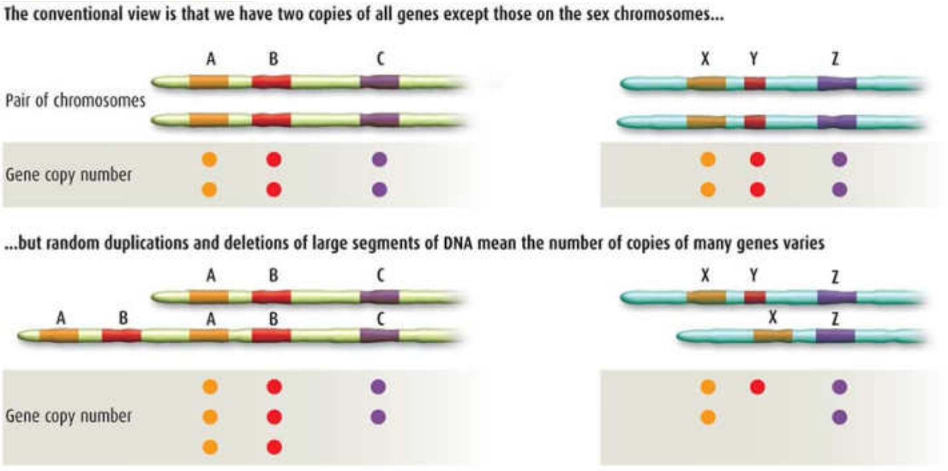
Average number of SNVs per individual

Orangutans 9.3 million > Gorillas 6.5 million > Chimpanzees 5.7 million > **Humans 3-4 million**

As a species, humans have relatively low diversity

(Prado-Martinez *et al.*, 2013, *Nature*)

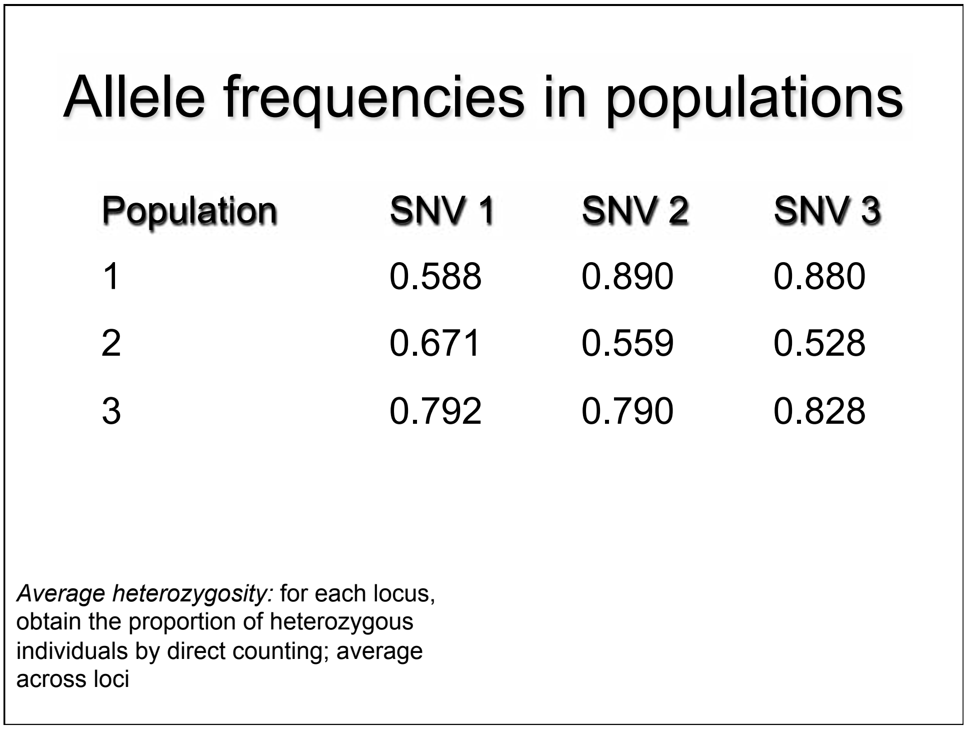
Copy number variants (deletions/duplications > 50 bp) account for more inter-individual variation than do single-nucleotide variants


The conventional view is that we have two copies of all genes except those on the sex chromosomes...
A
B
C
X
Y
Z
Pair of chromosomes
Gene copy number
...but random duplications and deletions of large segments of DNA mean the number of copies of many genes varies
A
B
C
A
B
A
B
C
X
Y
Z
X
Z
Gene copy number

In an average haploid human sequence, ~9 Mb are affected by structural variants; 3.6 Mb are affected by SNVs; on average, humans are heterozygous for ~150 CNVs (Sudmant *et al*., 2015, *Nature)*

How much do human populations differ?

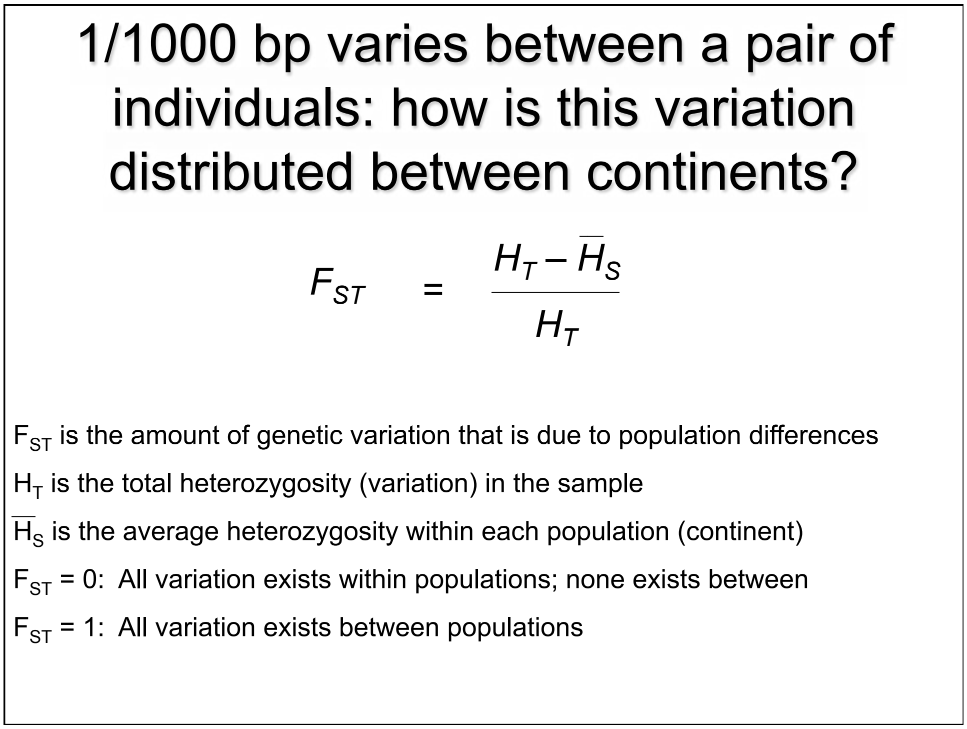
s,
atior
n (26)
Kyrgyzstan (25)
Buryat (25)
Iraqi Kurds (25)
Pakistanis (25)
Nepalese (25)
Thai (25)
Bambara (25)
Dogon (24)
To
Tongan(13)
Bolivian (23)



Allele frequencies in populations

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

Average heterozygosity: for each locus, obtain the proportion of heterozygous individuals by direct counting; average across loci



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

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

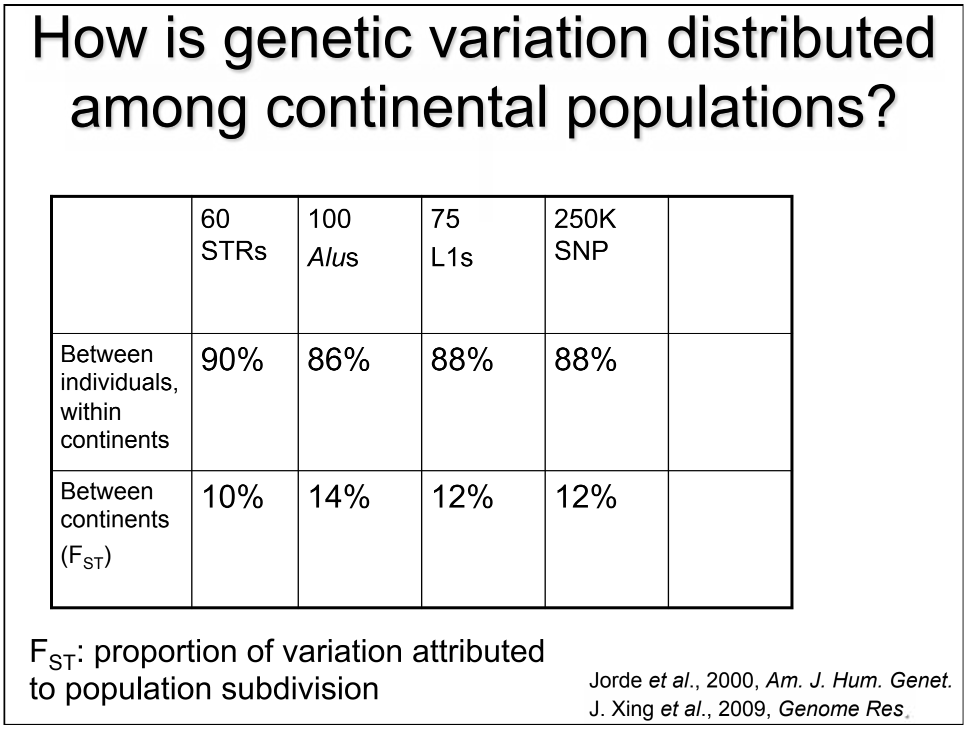
F_{ST} is the amount of genetic variation that is due to population differences

H_T is the total heterozygosity (variation) in the sample

$\overline{H}_S$ 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

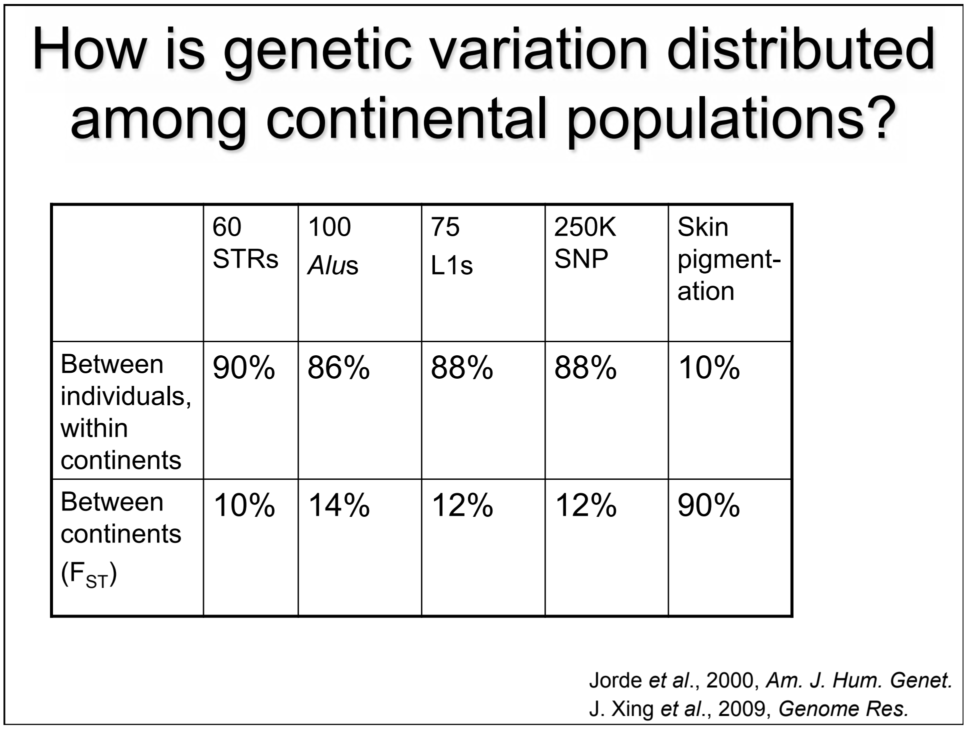


How is genetic variation distributed among continental populations?

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

F_{ST}: proportion of variation attributed to population subdivision

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



How is genetic variation distributed among continental populations?

	60 STRs	100 *Alu*s	75 L1s	250K SNP	Skin pigment-ation
Between individuals, within continents	90%	86%	88%	88%	10%
Between continents (F_{ST})	10%	14%	12%	12%	90%

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

% common SNPs shared among four major regions (Africa, Europe, E. Asia, India): 250K chip results for ~1,000 samples

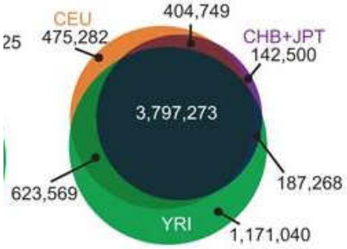
Minor allele present in:	
All 4 groups	78.6%
At least 3 groups	88.0%
At least 2 groups	92.1%
Africa only	7.4%
Any non-African group	0.5%

No SNPs were fixed present in one population, fixed absent in another

J. Xing *et al*., 2010, *Genomics*

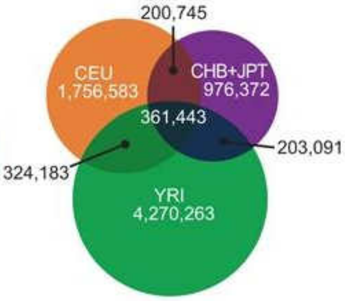
Rare single nucleotide variants (SNVs) are much more likely to be population-specific

Common SNPs



Average allele frequency difference between populations: 15%

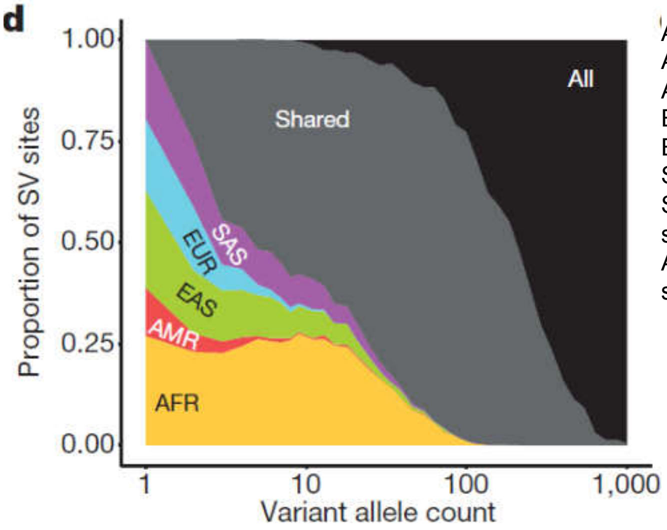
New rarer SNVs identified by sequencing



<5% of alleles with frequency < 2% are shared across continents

Durbin *et al.*, 2010, *Nature*
Auton *et al.*, 2015, *Nature*
(1000 Genomes Project)

Rare copy number variants are population-specific (1000 Genomes data)



AFR=African
AMR=Native American
EAS=East Asian
EUR=European
SAS=South Asian
Shared=present in >1 superpopulation
All=present in all superpopulations

Sudmant et al., 2015, *Nature*

A simple genetic distance to measure population differences

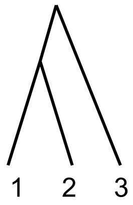
$$D_{ij} = |p_i - p_j|$$

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

Pop.	SNV 1	SNV 2	SNV 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 SNVs)

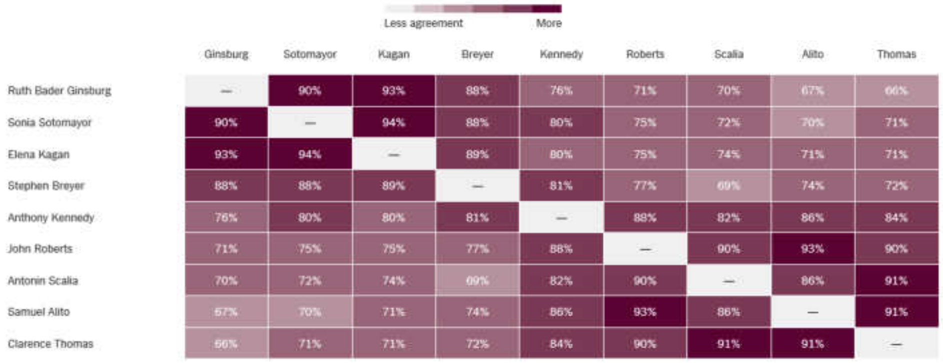
Building a population network



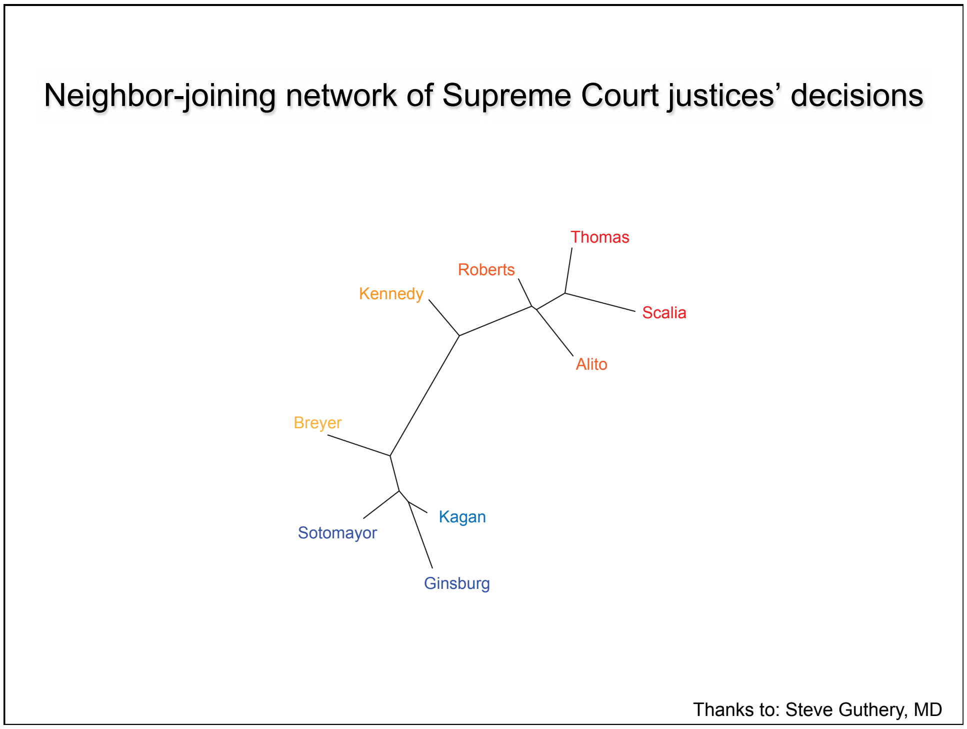
Pop.	SNV 1
1	0.588
2	0.671
3	0.792

$|p_1 - p_2| \quad | p_3 - (p_1 + p_2)/2 |$

Percent agreement between Supreme Court justices (*New York Times,* 2014) – analogous to % alleles shared among individuals

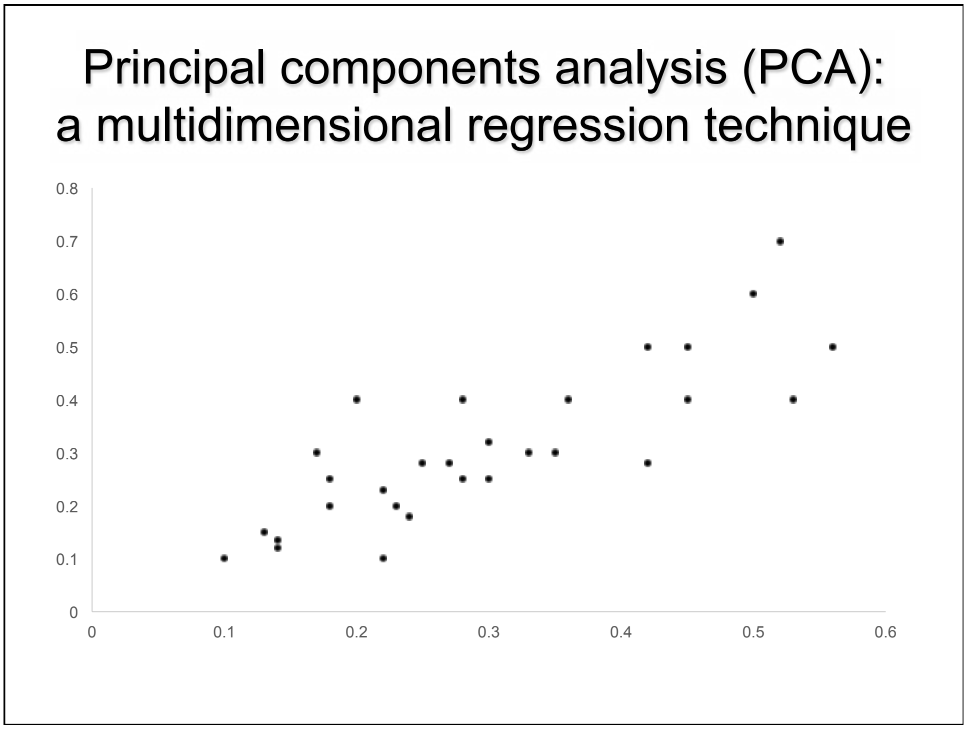

Less agreement
More

	Ginsburg	Sotomayor	Kagan	Breyer	Kennedy	Roberts	Scalia	Alito	Thomas
Ruth Bader Ginsburg	—	90%	93%	88%	76%	71%	70%	67%	66%
Sonia Sotomayor	90%	—	94%	88%	80%	75%	72%	70%	71%
Elena Kagan	93%	94%	—	89%	80%	75%	74%	71%	71%
Stephen Breyer	88%	88%	89%	—	81%	77%	69%	74%	72%
Anthony Kennedy	76%	80%	80%	81%	—	88%	82%	86%	84%
John Roberts	71%	75%	75%	77%	88%	—	90%	93%	90%
Antonin Scalia	70%	72%	74%	69%	82%	90%	—	86%	91%
Samuel Alito	67%	70%	71%	74%	86%	93%	86%	—	91%
Clarence Thomas	66%	71%	71%	72%	84%	90%	91%	91%	—

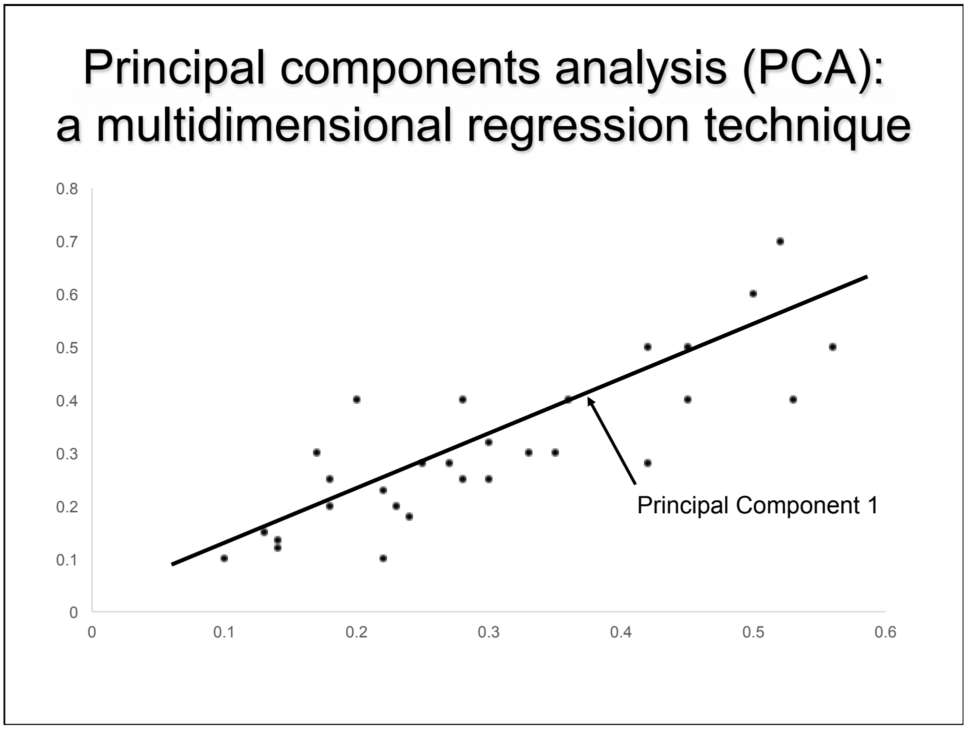


Neighbor-joining network of Supreme Court justices' decisions

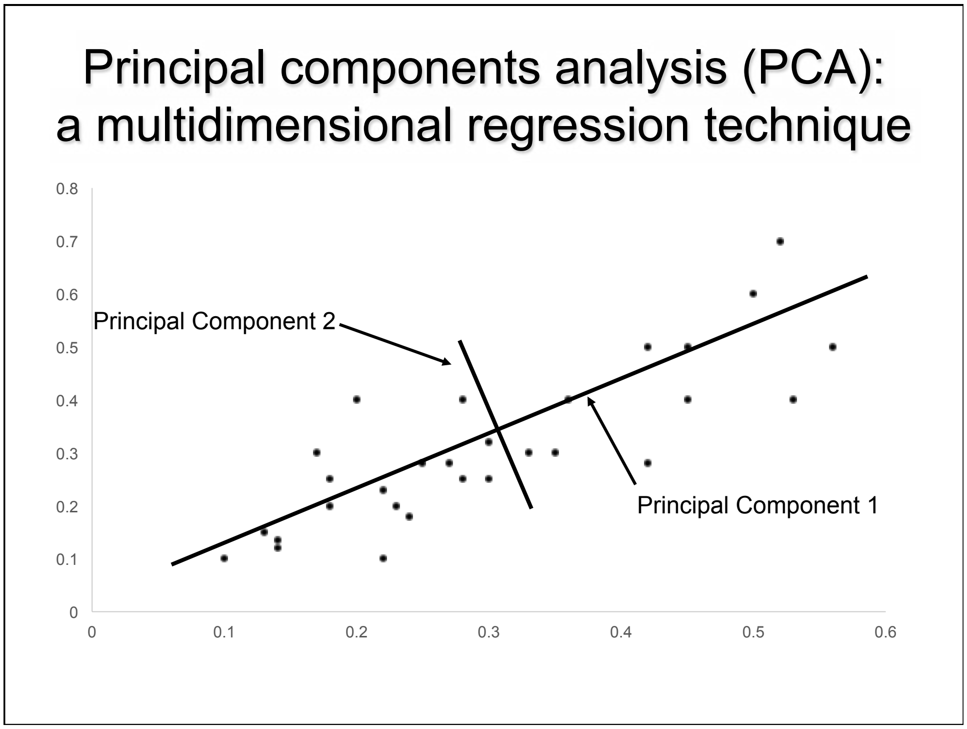
Thanks to: Steve Guthery, MD



Principal components analysis (PCA):
a multidimensional regression technique

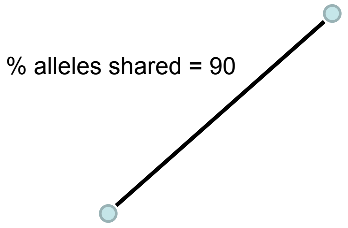


Principal components analysis (PCA): a multidimensional regression technique

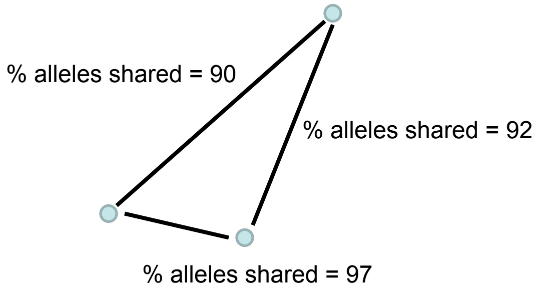


Principal components analysis (PCA): a multidimensional regression technique

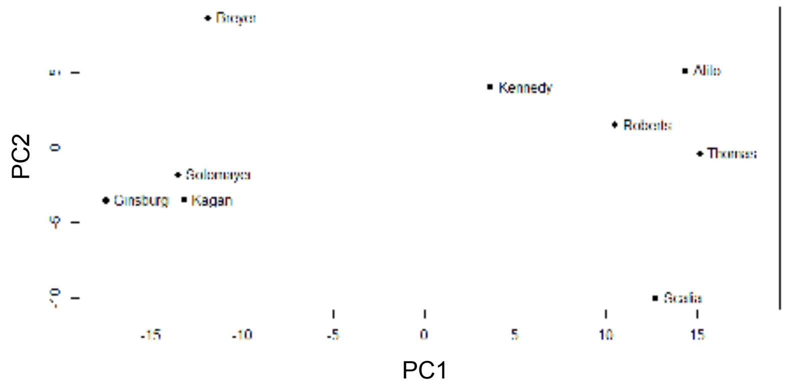
Genetic similarity between two people can be completely described with a line



Genetic similarities among three people can be completely described with a plane (two dimensions)

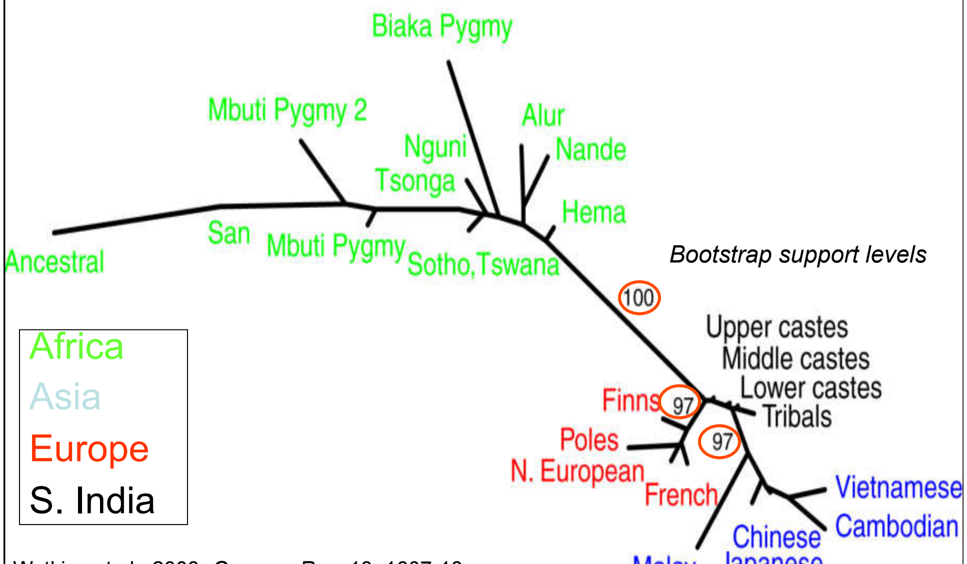


Principal components analysis of Supreme Court decision-making agreement

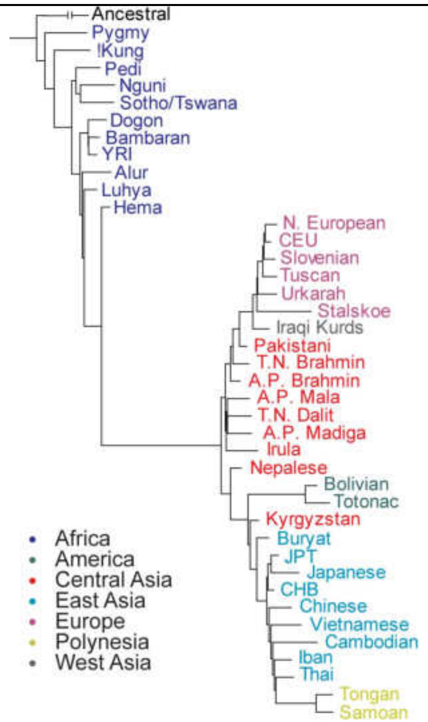


Thanks to: Julie Feusier, PhD-to-be

Population relationships based on 100 autosomal *Alu* polymorphisms



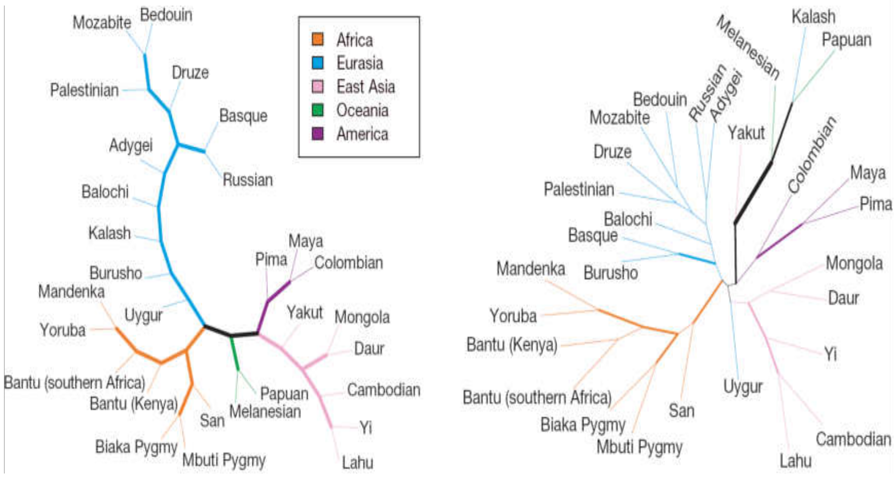
Watkins et al., 2003, *Genome Res.* 13: 1607-18



40 Populations, ~250K SNVs

Xing et al., 2010, *Genomics*

Similar patterns seen in Human Genome Diversity Project data

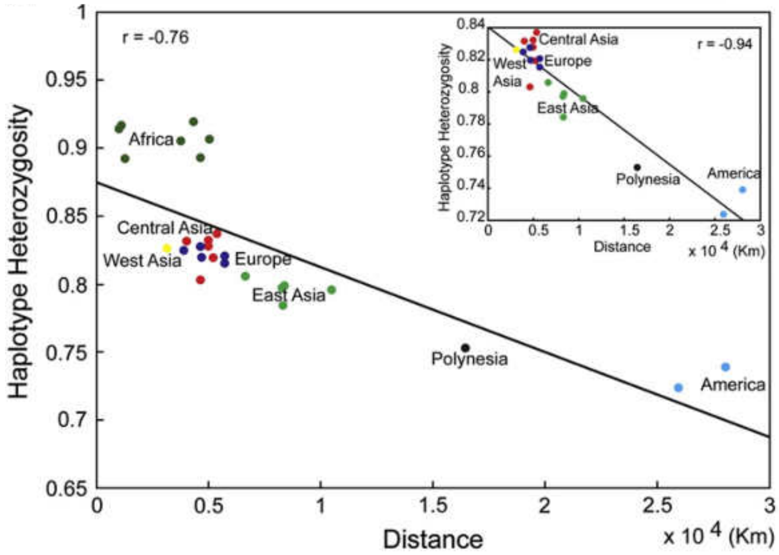


525,910 SNPs

396 copy number variants (CNVs)

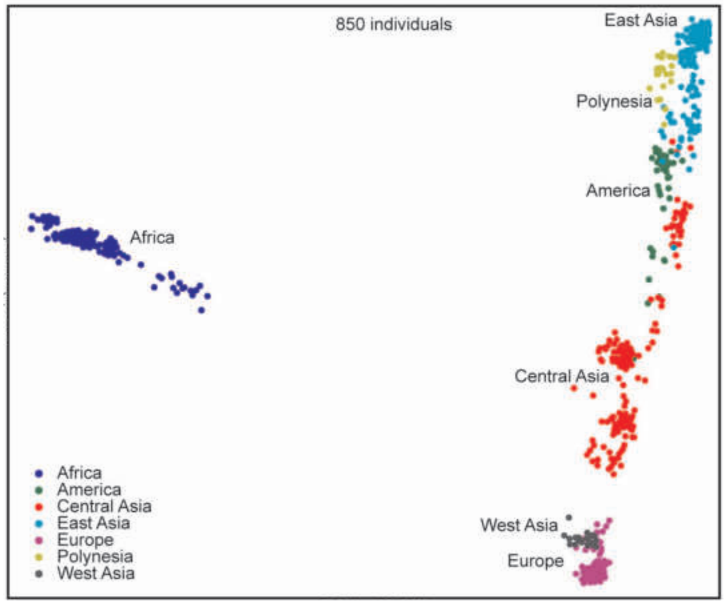
Jakobsson et al., 2008, *Nature* 451: 998-1003

Haplotype diversity declines with distance from Africa

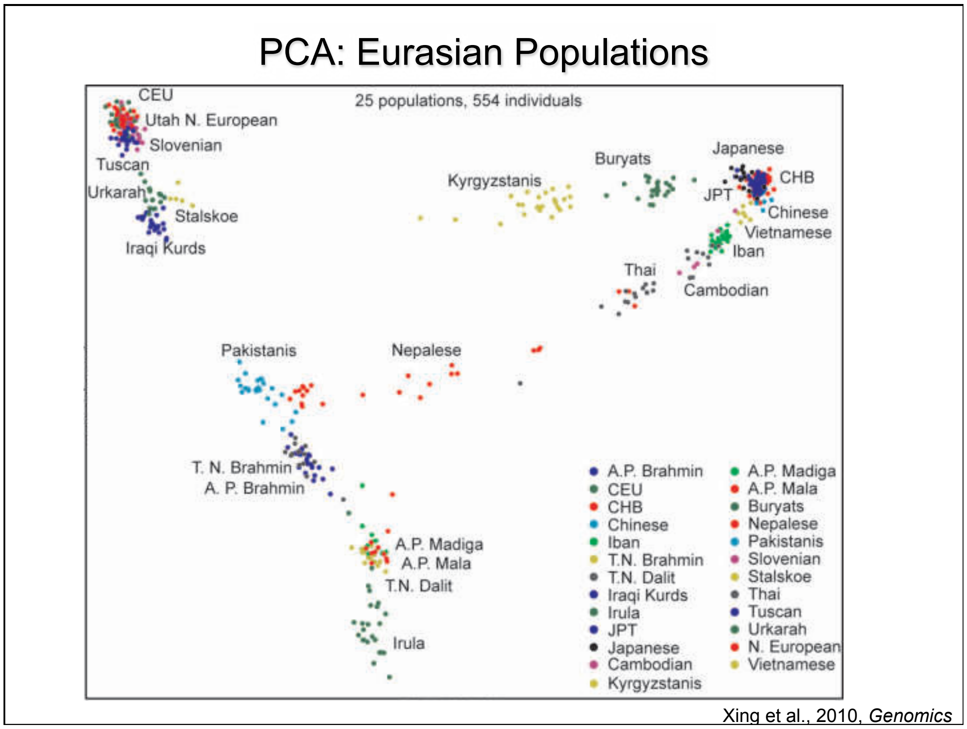


J. Xing et al., 2010, *Genomics*

Principal components analysis (PCA) displays **individual** genetic similarity in 2D: each dot = 1 individual

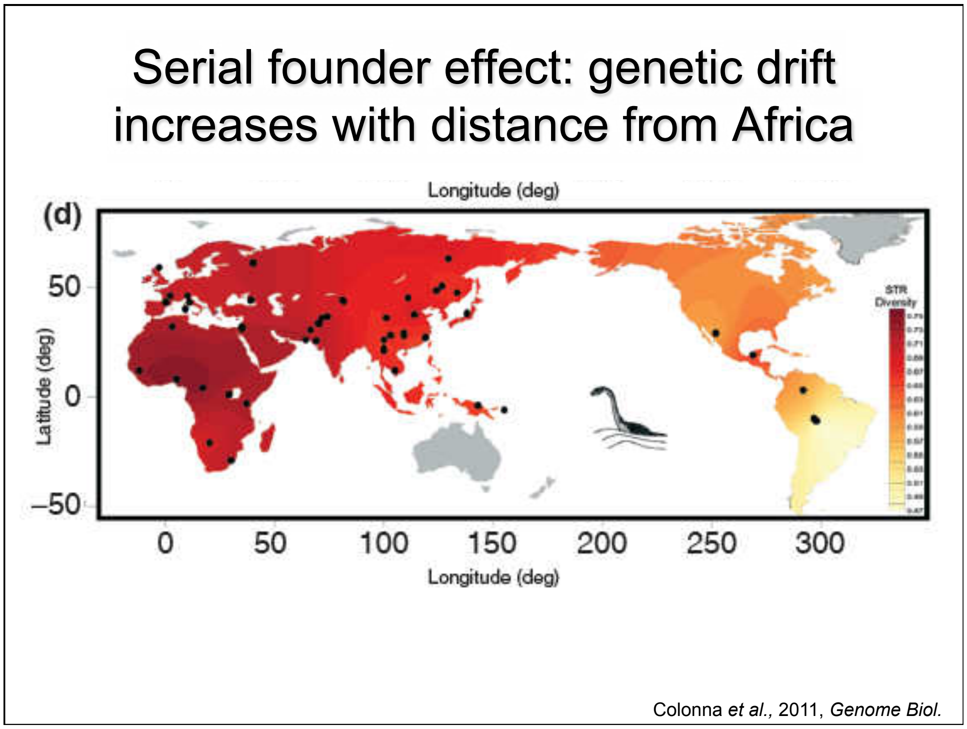


Xing et al., 2010, *Genomics*



PCA: Eurasian Populations

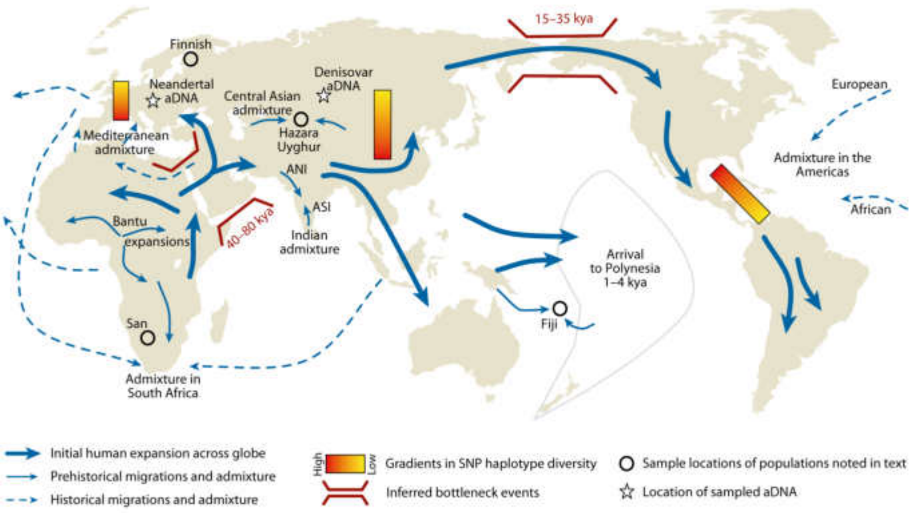
Xing et al., 2010, *Genomics*



Serial founder effect: genetic drift increases with distance from Africa

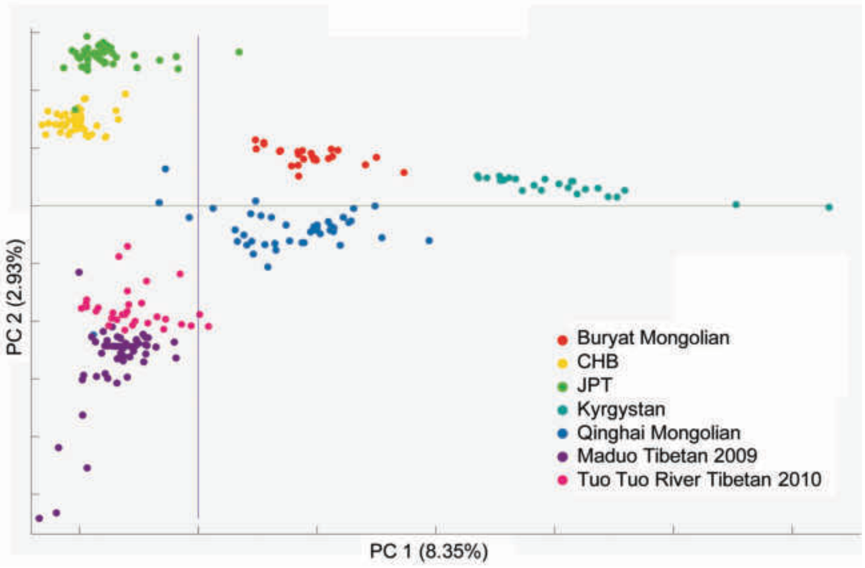
Colonna *et al.*, 2011, *Genome Biol.*

Recent African origin of anatomically modern humans

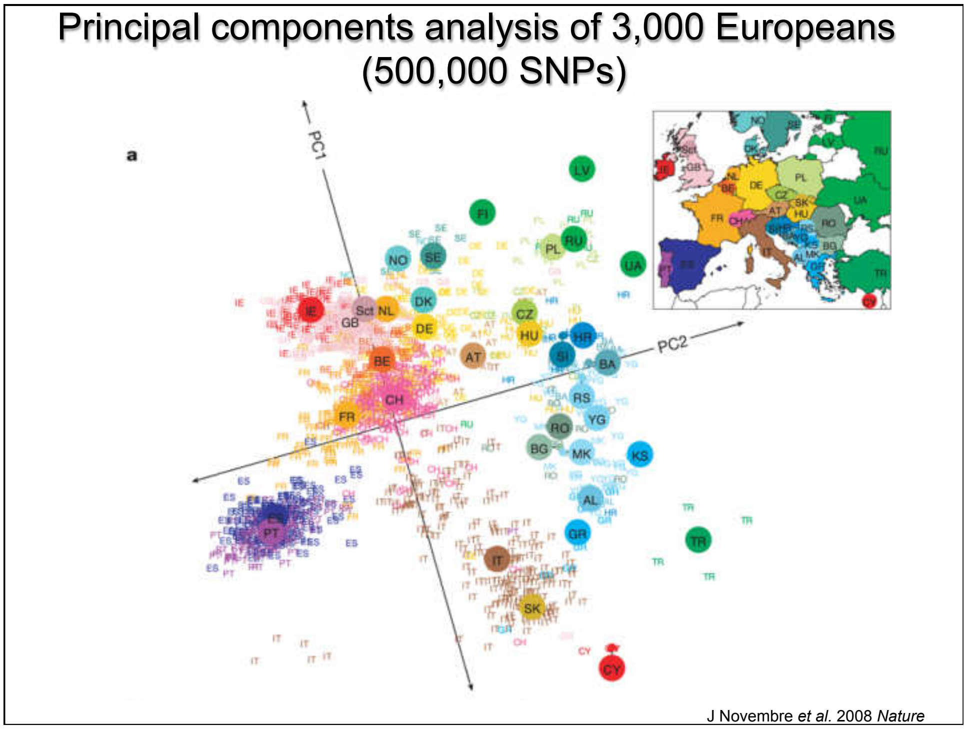

15–35 kya
Finnish
Neandertal aDNA
Denisovar aDNA
Central Asian admixture
Hazara Uyghur
European
Mediterranean admixture
ANI
ASI
Admixture in the Americas
African
Bantu expansions
40–80 kya
Indian admixture
Arrival to Polynesia 1–4 kya
Fiji
San
Admixture in South Africa
Initial human expansion across globe
Prehistorical migrations and admixture
Historical migrations and admixture
High
Low
Gradients in SNP haplotype diversity
Inferred bottleneck events
Sample locations of populations noted in text
Location of sampled aDNA

Novembre J, Ramachandran S. 2011.
Annu. Rev. Genomics Hum. Genet. 12:245–74

PCA can distinguish closely related populations: 1 million SNP microarray

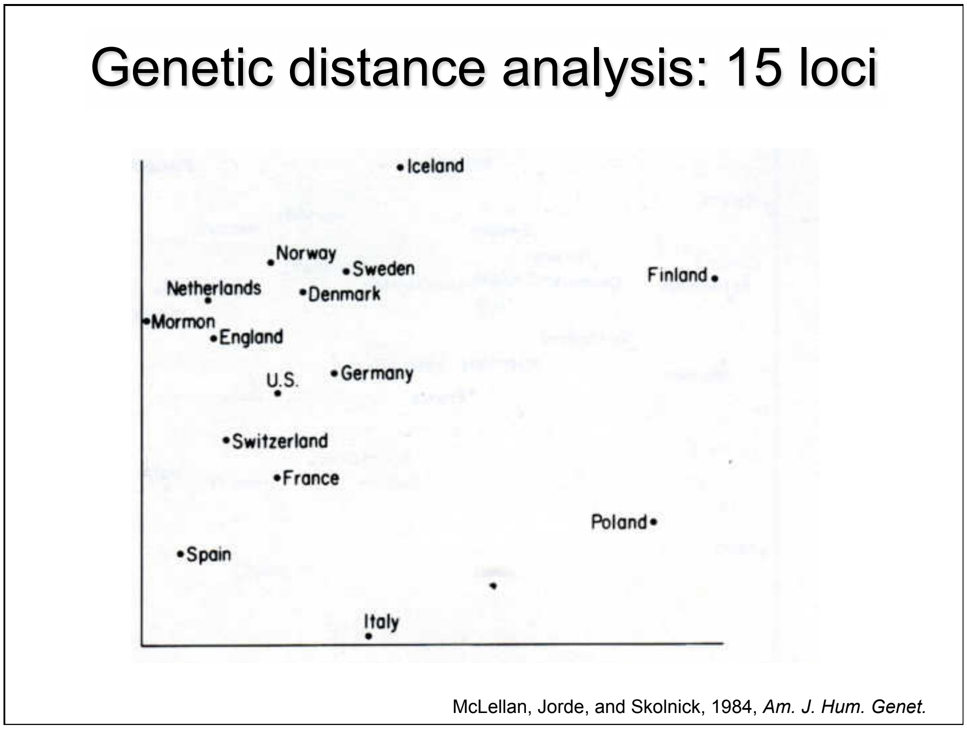


Xing *et al*., 2013 *PLoS Genetics*



Principal components analysis of 3,000 Europeans (500,000 SNPs)

J Novembre *et al.* 2008 *Nature*



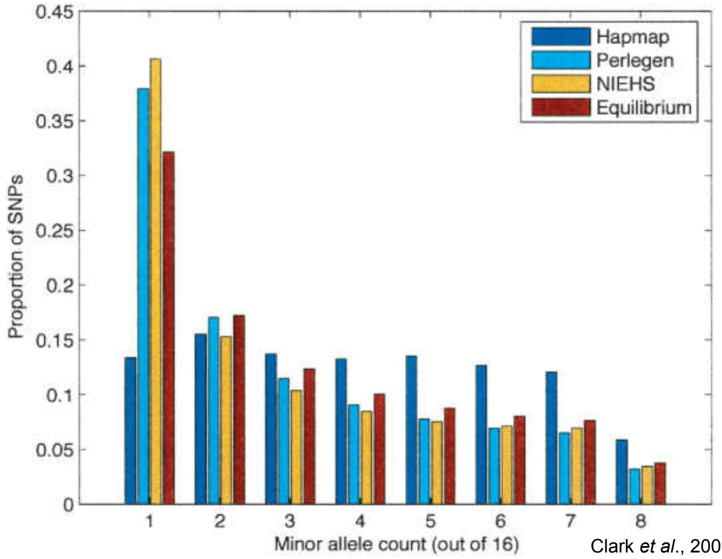
Genetic distance analysis: 15 loci

McLellan, Jorde, and Skolnick, 1984, *Am. J. Hum. Genet.*

Sequence data permit more accurate inferences about population history

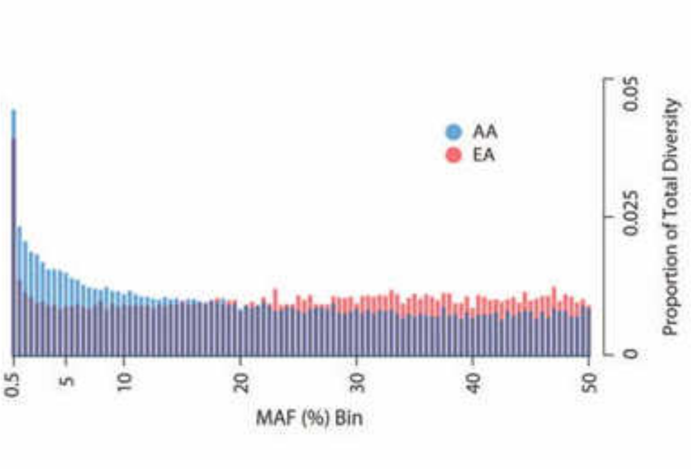
- Microarray SNPs are selected for higher frequency and diversity in Europeans
- Complete DNA sequences are unbiased and include information about rare variants
- Coalescence methods can be used effectively with sequence data

The effect of ascertainment bias on allele frequencies: Microarray data cannot accurately estimate demographic parameters (population size, growth rates)



Clark *et al*., 2005, *Genome Res.* 15: 1496-1502

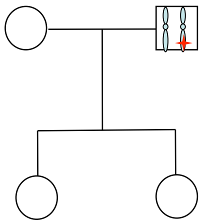
Allele frequency spectrum (2,440 exomes) indicates a recent population expansion



73% of *all* protein-coding SNVs and 86% of deleterious SNVs arose within past 5,000-10,000 years (Fu et al., 2013, *Nature,* 493: 216-20)

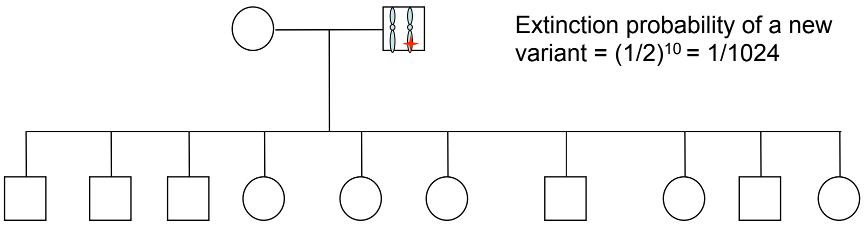
Tennessen *et al.,* 2012, *Science*

Population expansions increase the frequency of rare variants



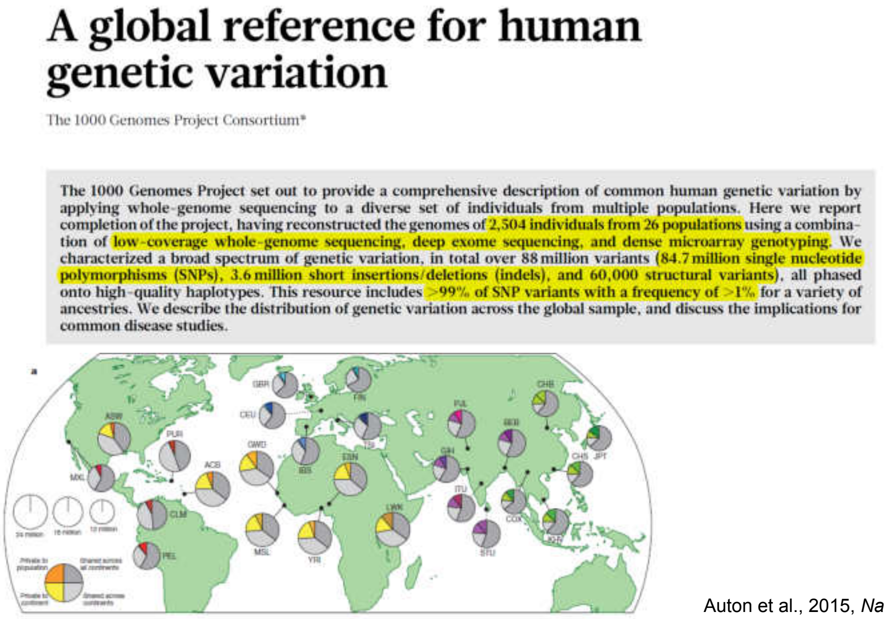
Extinction probability of a new variant = $(1/2)^2 = 1/4$

Population expansions increase the frequency of rare variants



Extinction probability of a new variant = $(1/2)^{10}$ = 1/1024

The 1000 Genomes Project

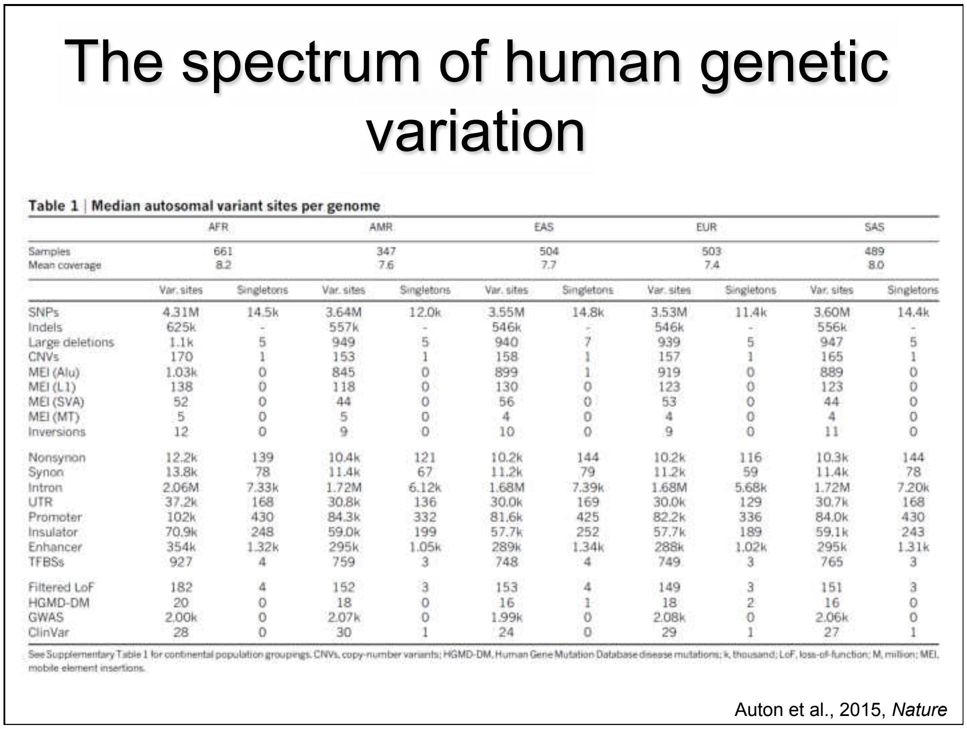


A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

Auton et al., 2015, *Nature*



The spectrum of human genetic variation

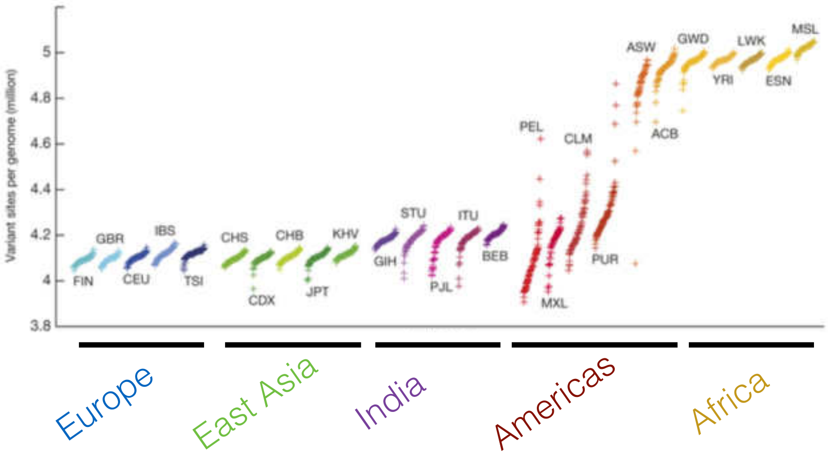
Table 1 | Median autosomal variant sites per genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean coverage	8.2		7.6		7.7		7.4		8.0	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (L1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
Nonsynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBSs	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.

Auton et al., 2015, *Nature*

Variation in individuals: 1000 Genomes Project

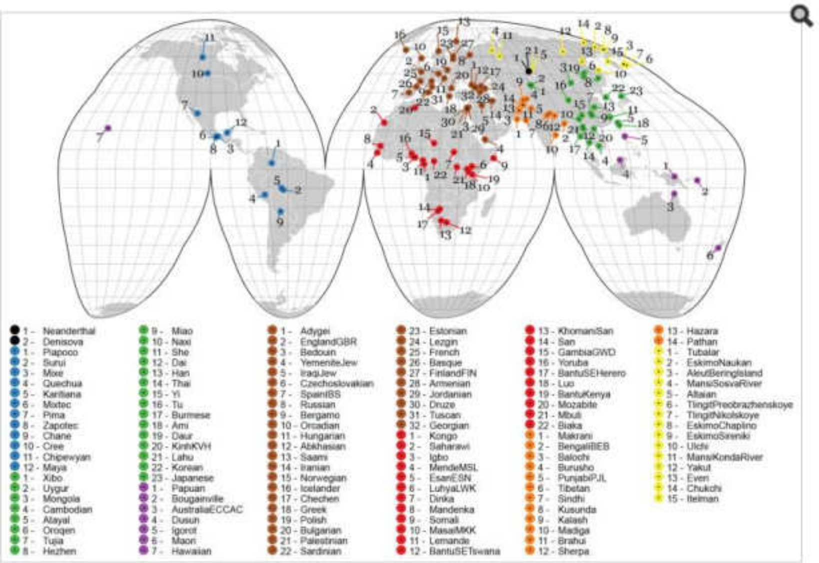


Auton et al., 2015, *Nature*

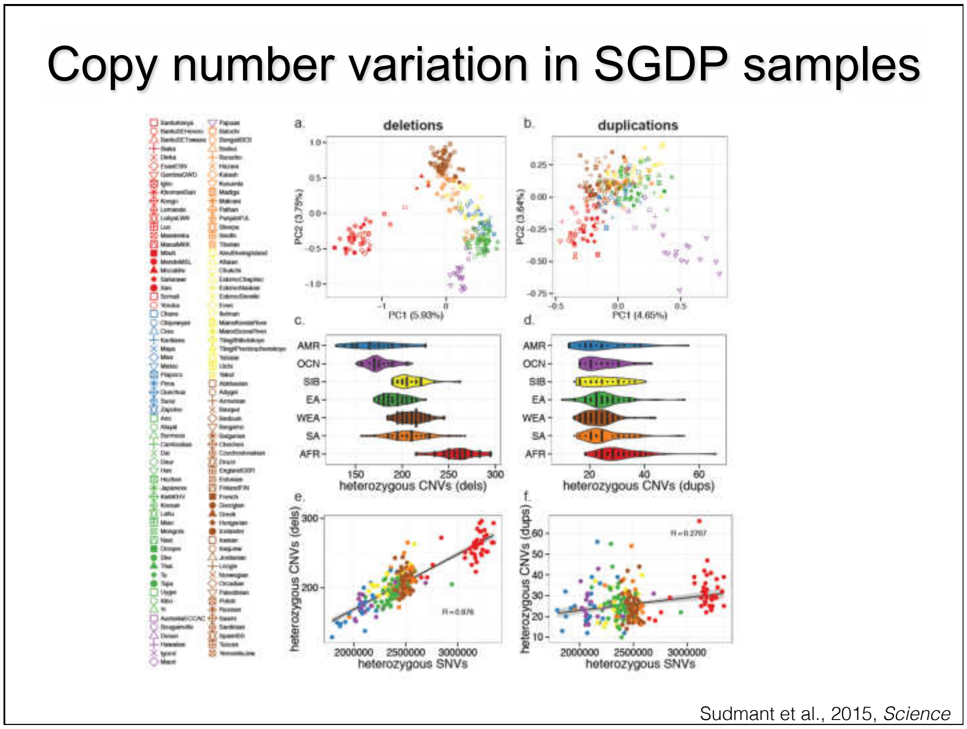
A “typical” human genome

Protein truncating	149 - 182
Peptide altering	10,000 -12,000
Regulatory (UTR, TBS, promoter, etc.)	459,000 - 565,000
Associated with complex trait	~2,000
ClinVar disease causing	24 - 30

Simons Genome Diversity Project (SGDP): 300 individuals in 142 populations; 40x sequencing

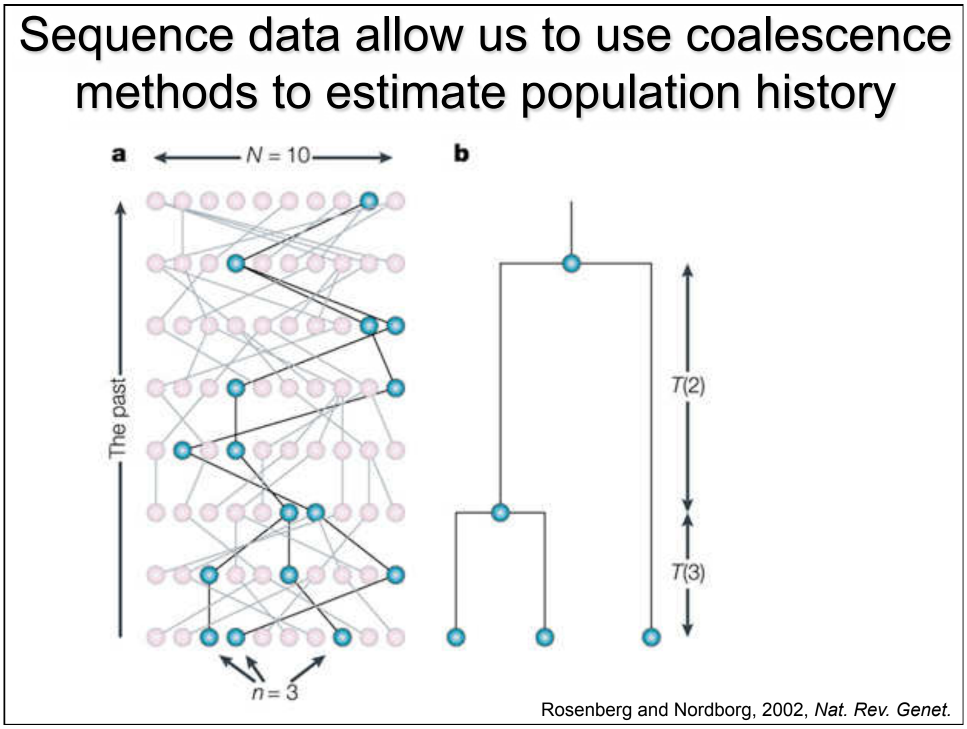


Sudmant et al., 2015, *Science*



Copy number variation in SGDP samples

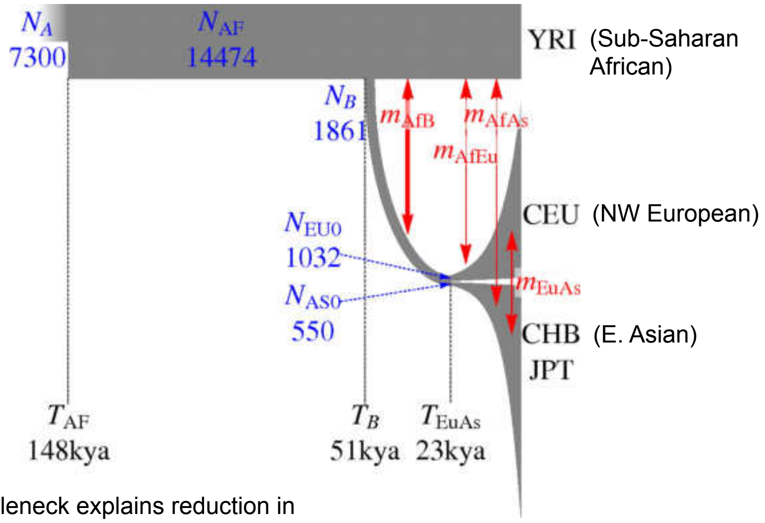
Sudmant et al., 2015, *Science*



Sequence data allow us to use coalescence methods to estimate population history

Rosenberg and Nordborg, 2002, *Nat. Rev. Genet.*

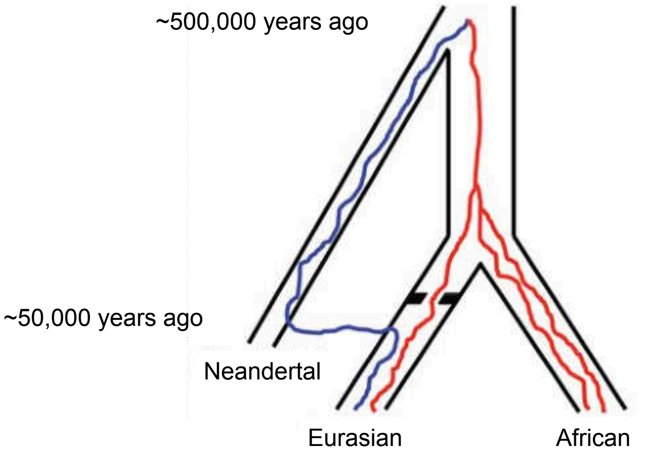
A sequence-based demographic model, with line width corresponding to population size and time flowing from left to right



Population bottleneck explains reduction in human genetic diversity
Recent expansion explains excess of rare alleles

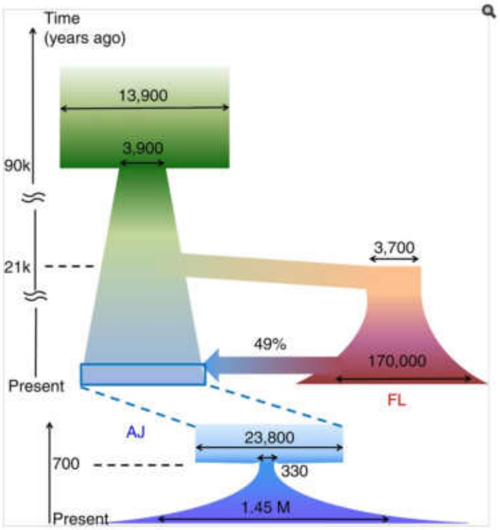
Gravel S et al. 2011 *PNAS* 108:11983-11988

Neandertal admixture with anatomically modern humans: On average, non-Africans have 1-4% Neandertal DNA



B Vernot and J M Akey, 2014 *Science*

Sequence-based reconstruction of Ashkenazi Jewish demographic history



Carmi et al., 2014, *Nat. Comm.*

Drift has increased the frequencies of several disease-causing mutations

- Three founder mutations in *BRCA1* or *BRCA2* are seen in 2.5% of Ashkenazi Jews (1/200 in general population)
- *APC* mutation predisposing to colorectal cancer is seen in 6% of Ashkenazi population
- Several lysosomal storage disorders (Gaucher, Niemann-Pick, Tay-Sachs) are relatively common

What can genetics tell us about “race”?

“’Race’ is biologically meaningless”
-- Schwartz, 2001, *N. Engl. J. Med.*

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

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

Bamshad and Olson, 2003

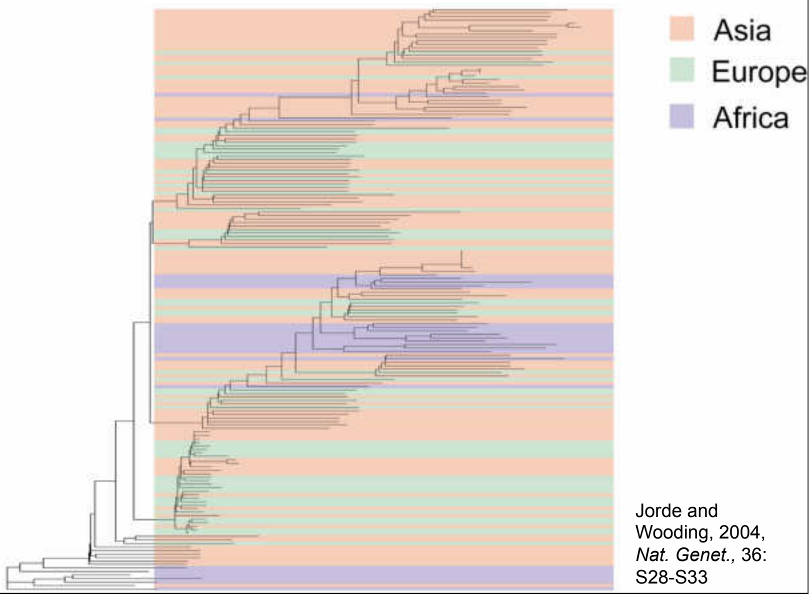
SCIENCE AND SOCIETY

Taking race out of human genetics

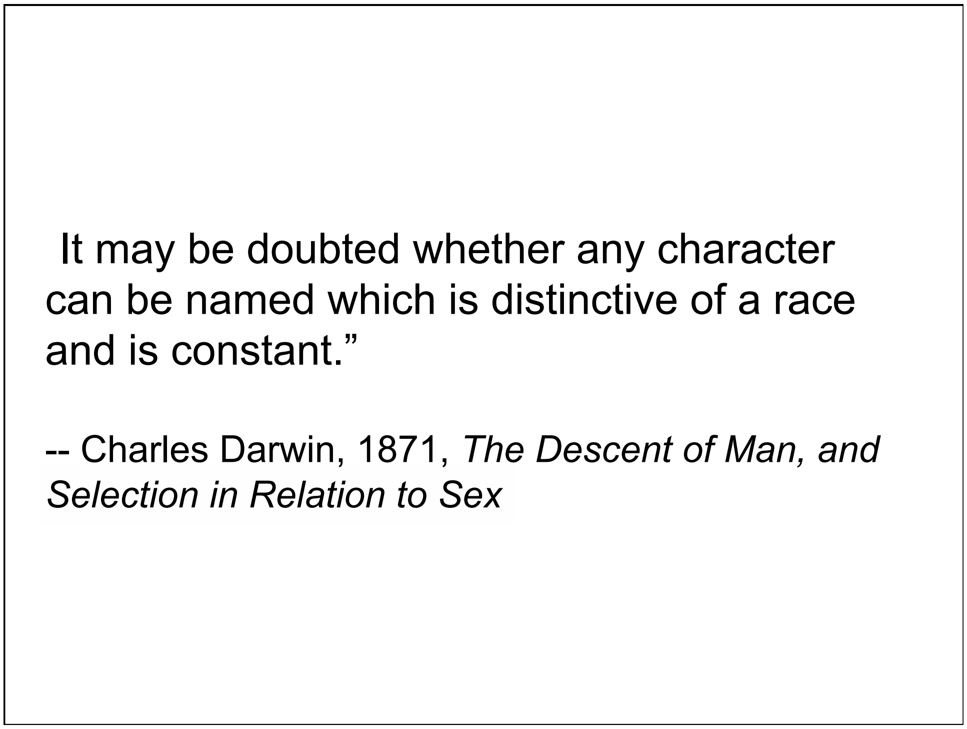
Engaging a century-long debate about the role of race in science

-- Yudell *et al*., 2016, *Science*

Individual network: 14 kb sequence in angiotensinogen gene

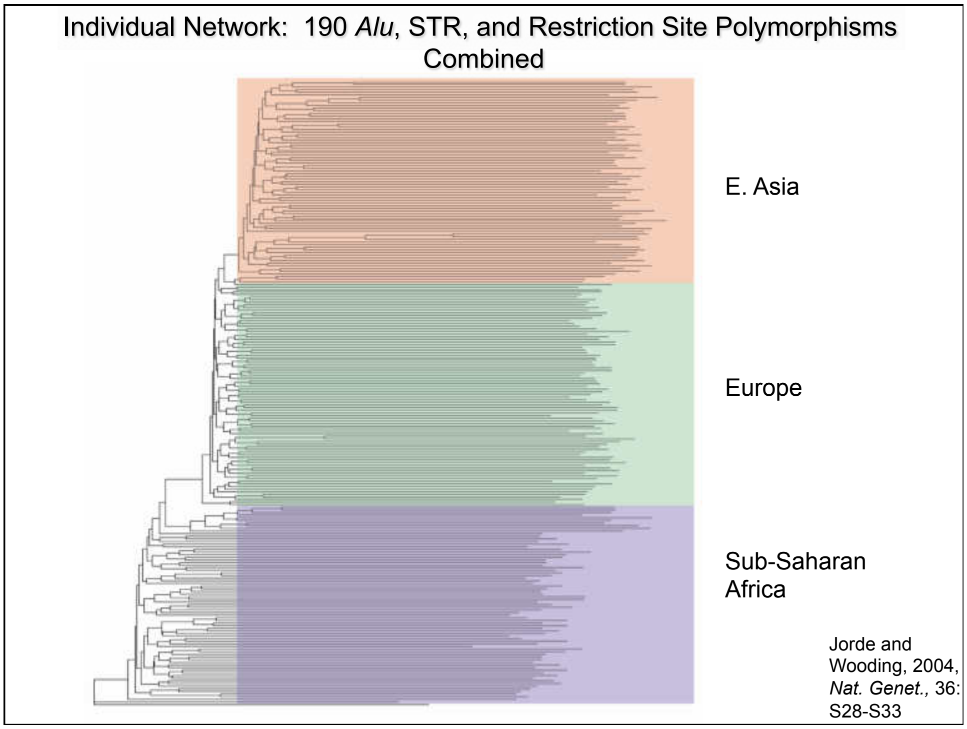


Jorde and Wooding, 2004, *Nat. Genet.*, 36: S28-S33



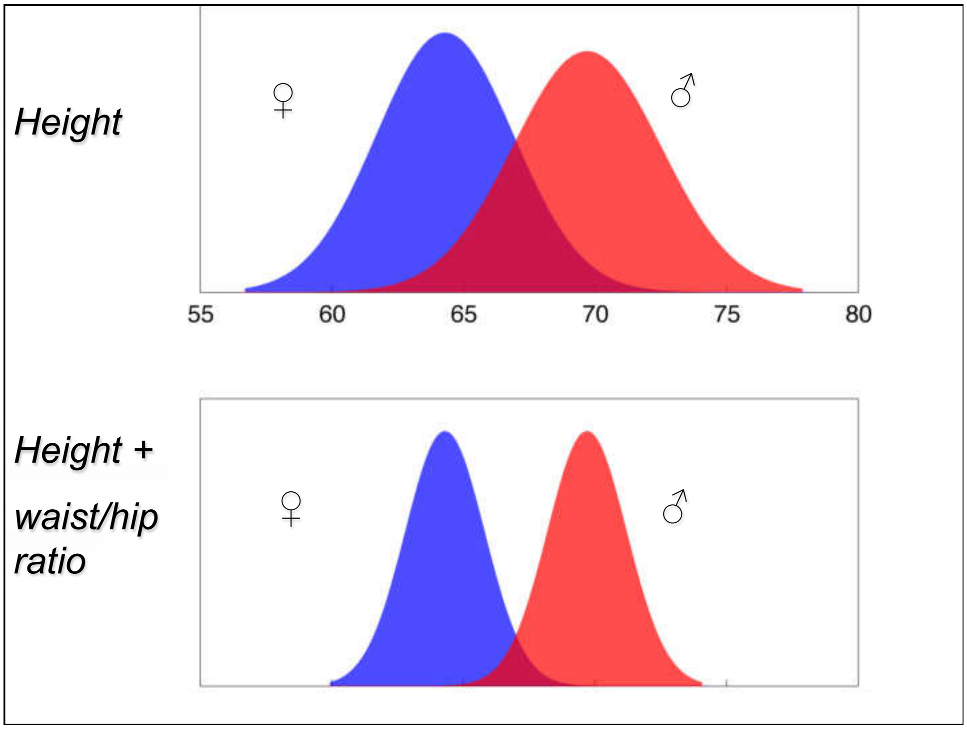
It may be doubted whether any character can be named which is distinctive of a race and is constant.”

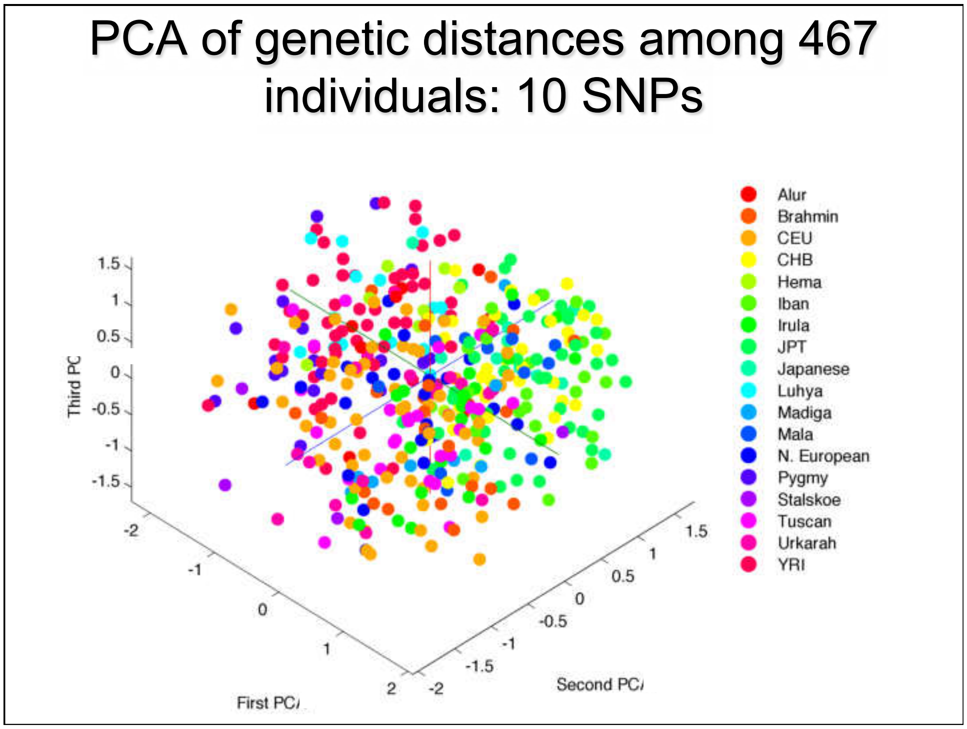
-- Charles Darwin, 1871, *The Descent of Man, and Selection in Relation to Sex*

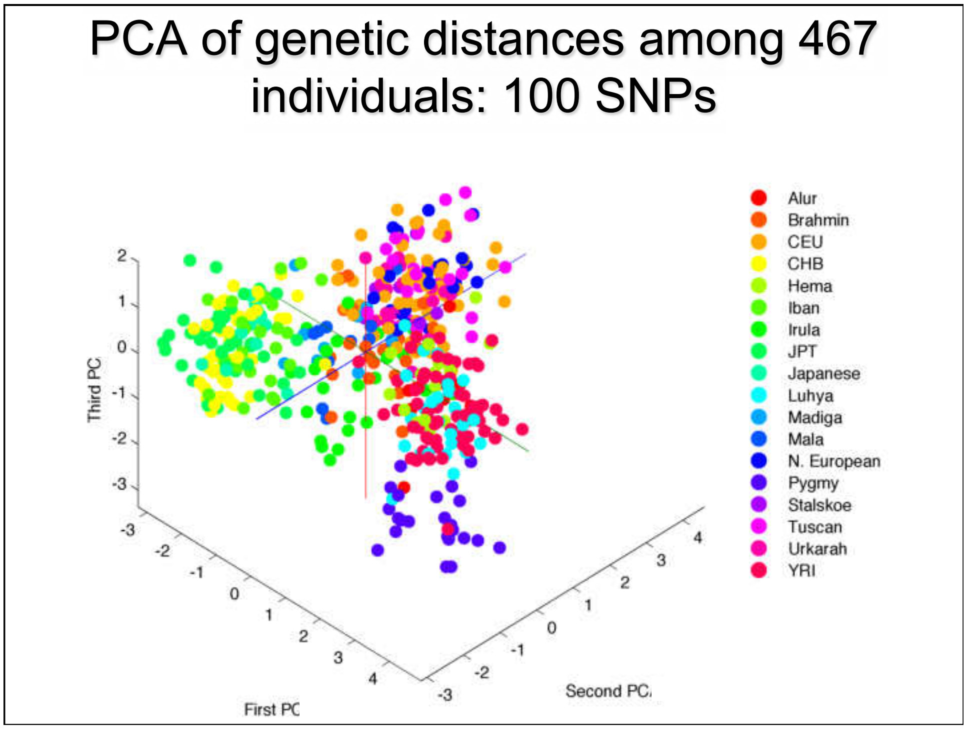


Individual Network: 190 *Alu*, STR, and Restriction Site Polymorphisms Combined

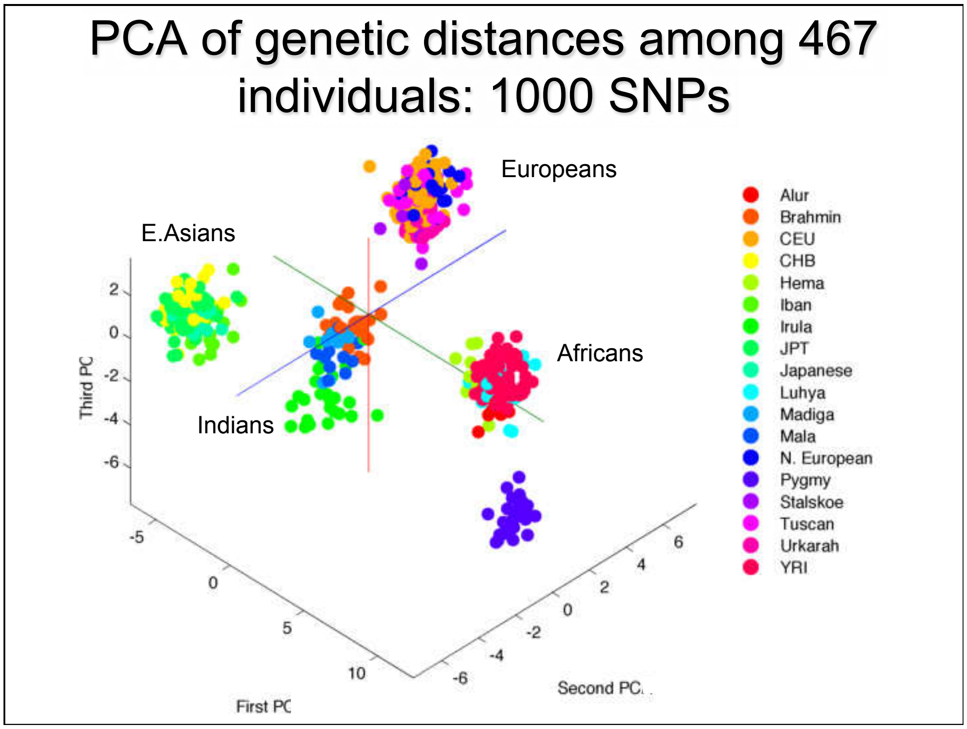
Jorde and Wooding, 2004, *Nat. Genet.*, 36: S28-S33



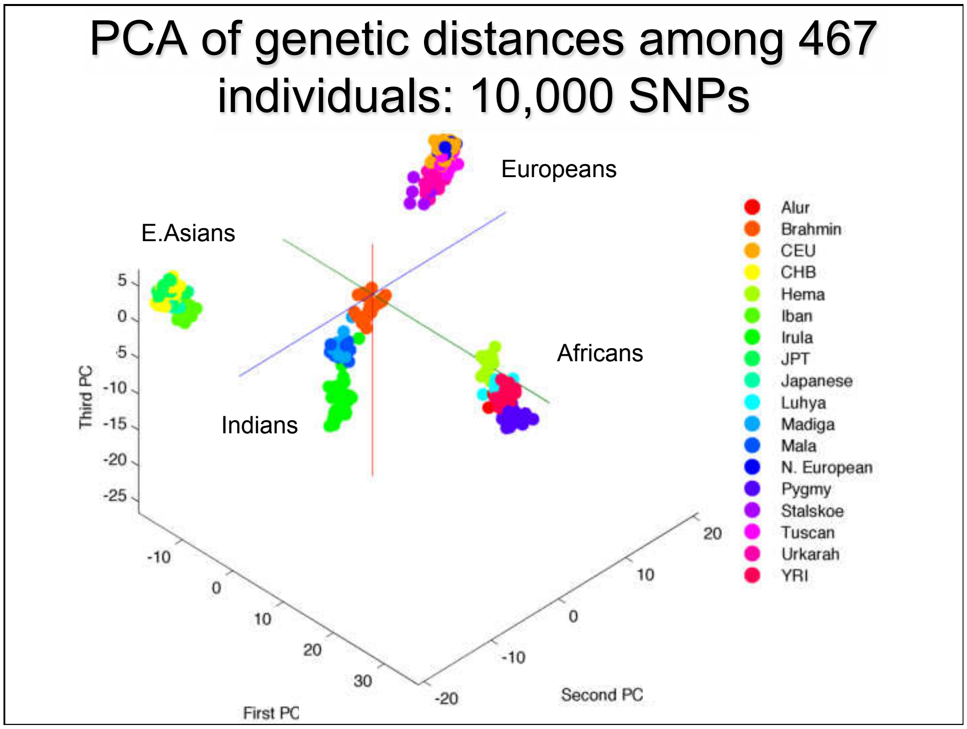

PCA of genetic distances among 467 individuals: 10 SNPs



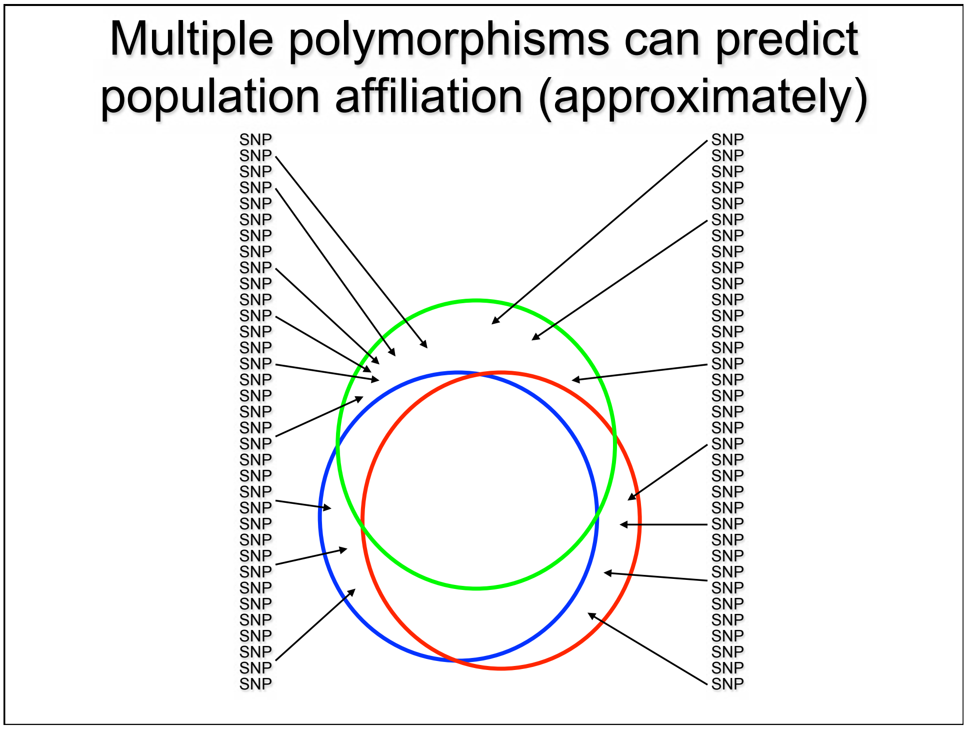
PCA of genetic distances among 467 individuals: 100 SNPs



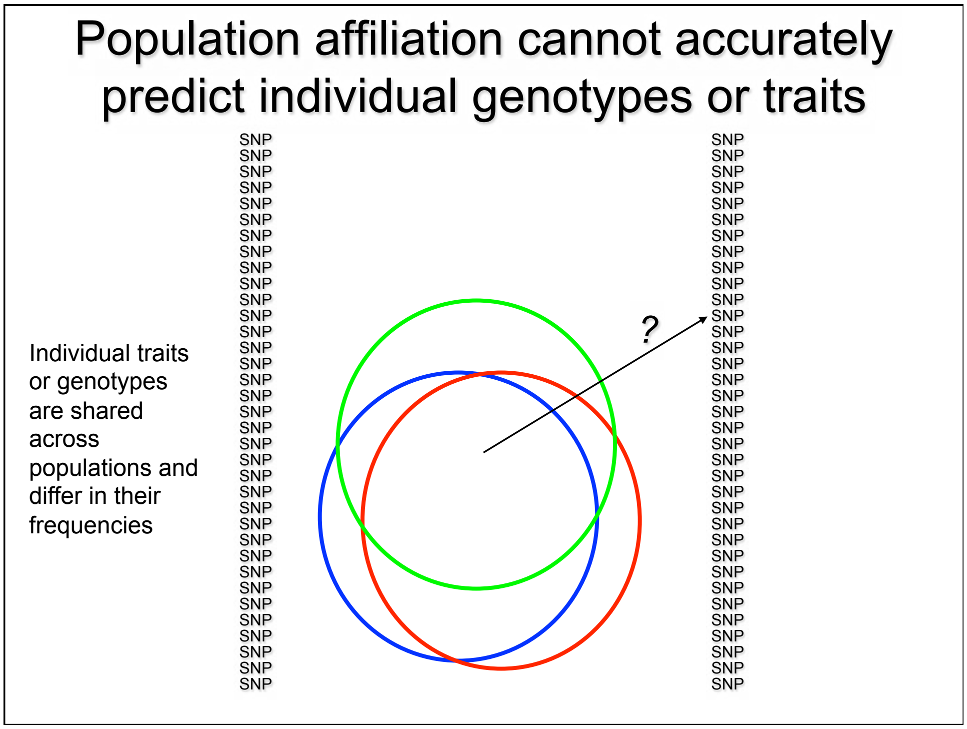
PCA of genetic distances among 467 individuals: 1000 SNPs



PCA of genetic distances among 467 individuals: 10,000 SNPs

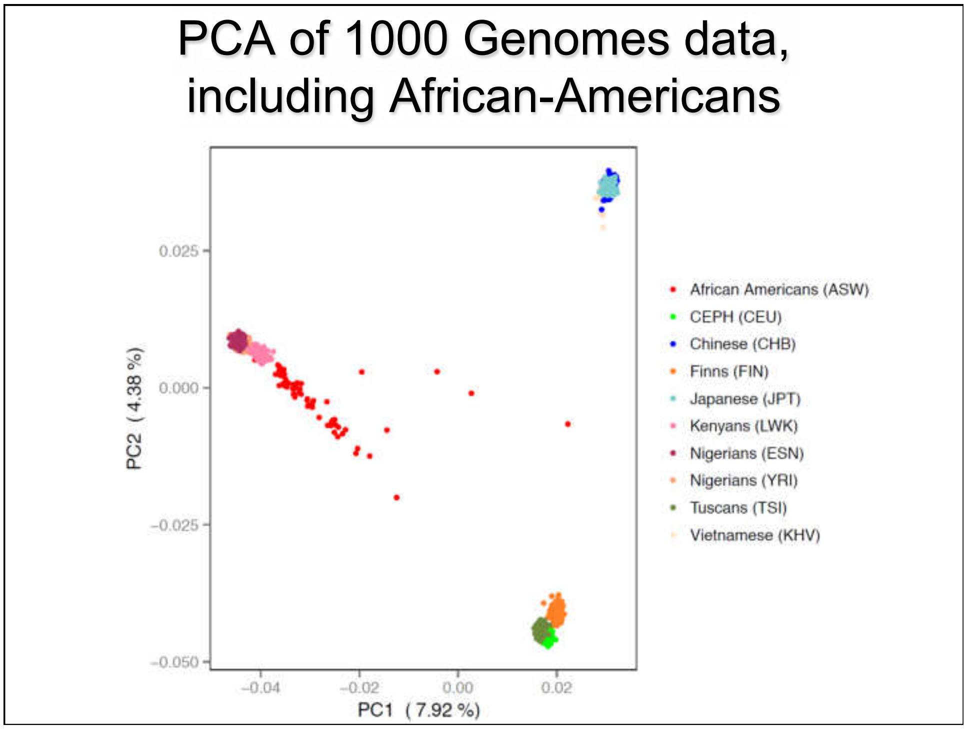


Multiple polymorphisms can predict population affiliation (approximately)



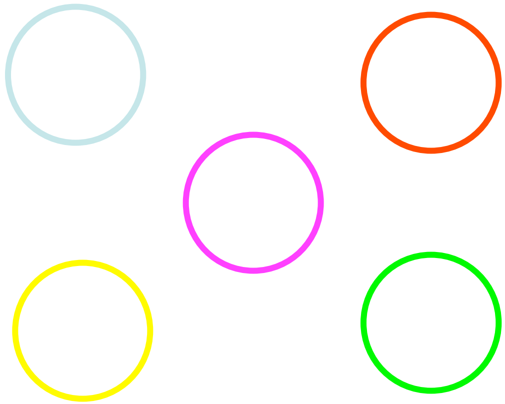
Population affiliation cannot accurately predict individual genotypes or traits

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

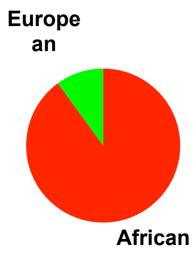


PCA of 1000 Genomes data, including African-Americans

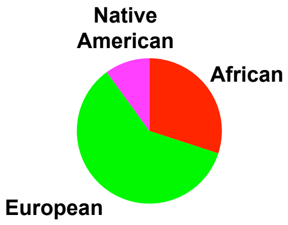
The Fallacy of Typological Thinking



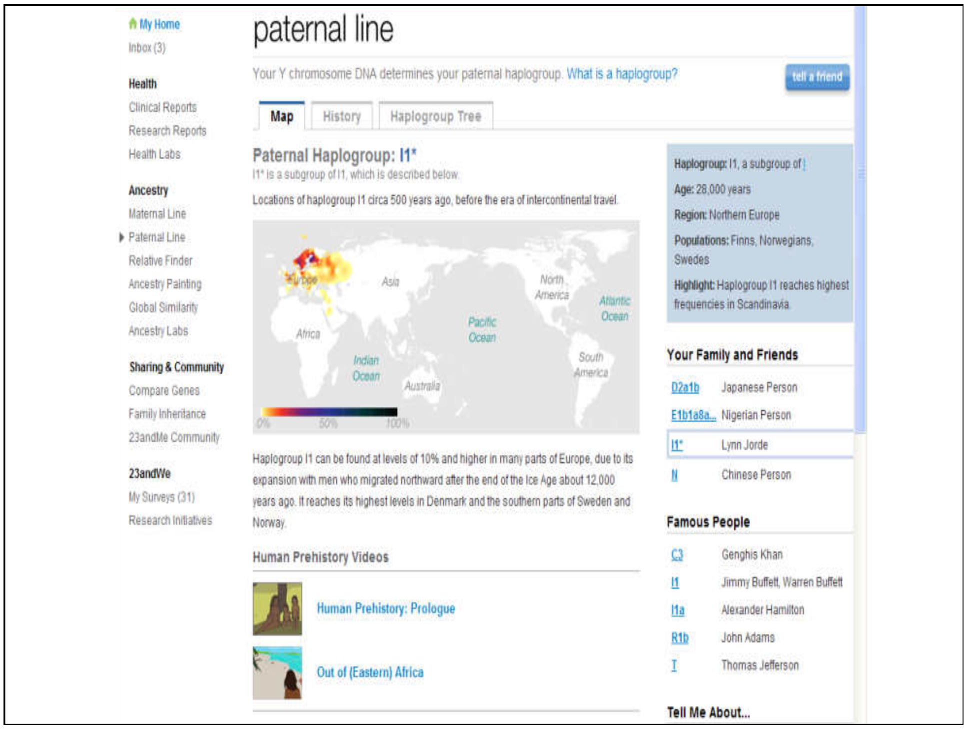
Ancestry vs. Race



“African-American”



“African-American”



paternal line

Your Y chromosome DNA determines your paternal haplogroup. What is a haplogroup?

Paternal Haplogroup: I1*

I1* is a subgroup of I1, which is described below.

Locations of haplogroup I1 circa 500 years ago, before the era of intercontinental travel.

Haplogroup: I1, a subgroup of I

Age: 28,000 years

Region: Northern Europe

Populations: Finns, Norwegians, Swedes

Highlight: Haplogroup I1 reaches highest frequencies in Scandinavia.

Haplogroup I1 can be found at levels of 10% and higher in many parts of Europe, due to its expansion with men who migrated northward after the end of the Ice Age about 12,000 years ago. It reaches its highest levels in Denmark and the southern parts of Sweden and Norway.

Your Family and Friends

D2a1b	Japanese Person
E1b1a8a...	Nigerian Person
I1*	Lynn Jorde
N	Chinese Person

Famous People

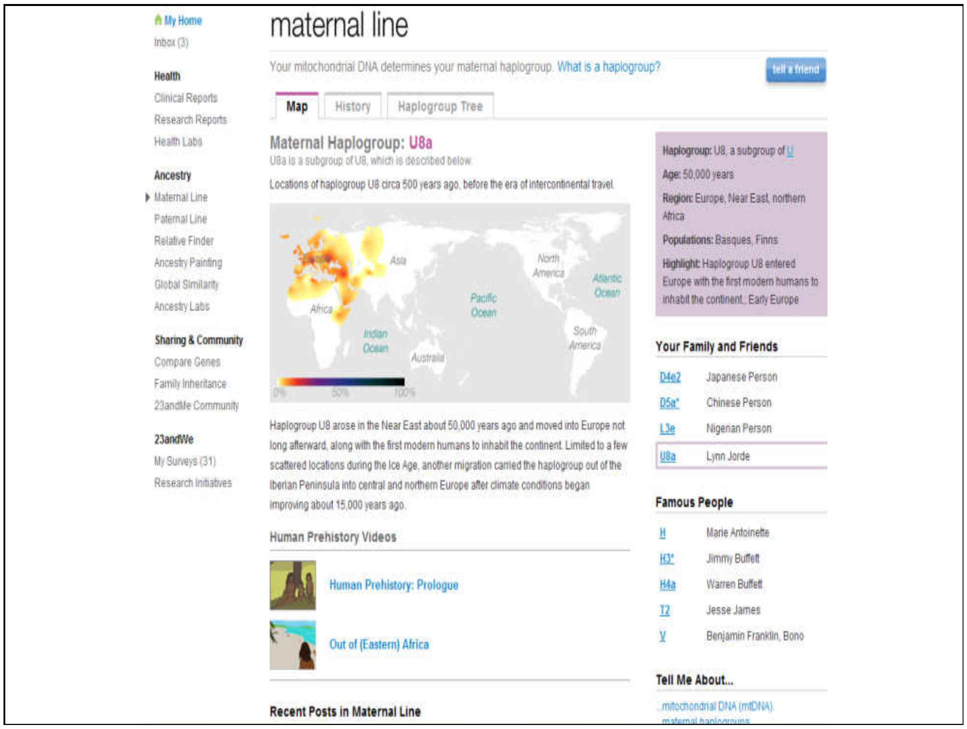
C3	Genghis Khan
I1	Jimmy Buffett, Warren Buffett
I1a	Alexander Hamilton
R1b	John Adams
T	Thomas Jefferson

Human Prehistory Videos

Human Prehistory: Prologue

Out of (Eastern) Africa

Tell Me About...



maternal line

Your mitochondrial DNA determines your maternal haplogroup. What is a haplogroup?

Maternal Haplogroup: U8a

U8a is a subgroup of U8, which is described below.

Locations of haplogroup U8 circa 500 years ago, before the era of intercontinental travel.

Haplogroup: U8, a subgroup of U

Age: 50,000 years

Region: Europe, Near East, northern Africa

Populations: Basques, Finns

Highlight: Haplogroup U8 entered Europe with the first modern humans to inhabit the continent.; Early Europe

Haplogroup U8 arose in the Near East about 50,000 years ago and moved into Europe not long afterward, along with the first modern humans to inhabit the continent. Limited to a few scattered locations during the Ice Age, another migration carried the haplogroup out of the Iberian Peninsula into central and northern Europe after climate conditions began improving about 15,000 years ago.

Your Family and Friends

D4e2	Japanese Person
D5a*	Chinese Person
L3e	Nigerian Person
U8a	Lynn Jorde

Famous People

H	Marie Antoinette
H3*	Jimmy Buffett
H4a	Warren Buffett
T2	Jesse James
V	Benjamin Franklin, Bono

Human Prehistory Videos

Human Prehistory: Prologue

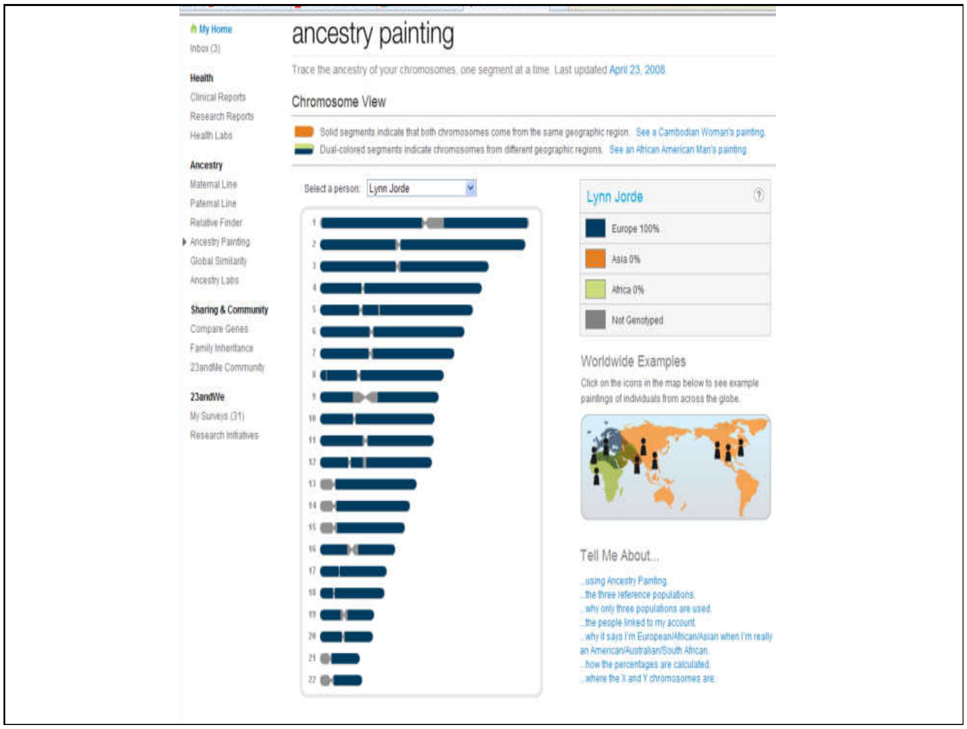
Out of (Eastern) Africa

Tell Me About...

...mitochondrial DNA (mtDNA).

...maternal haplogroups

Recent Posts in Maternal Line



ancestry painting

Trace the ancestry of your chromosomes, one segment at a time. Last updated April 23, 2008

Chromosome View

🟧 Solid segments indicate that both chromosomes come from the same geographic region. See a Cambodian Woman's painting.

▨ Dual-colored segments indicate chromosomes from different geographic regions. See an African American Man's painting.

Select a person: Lynn Jorde

Lynn Jorde

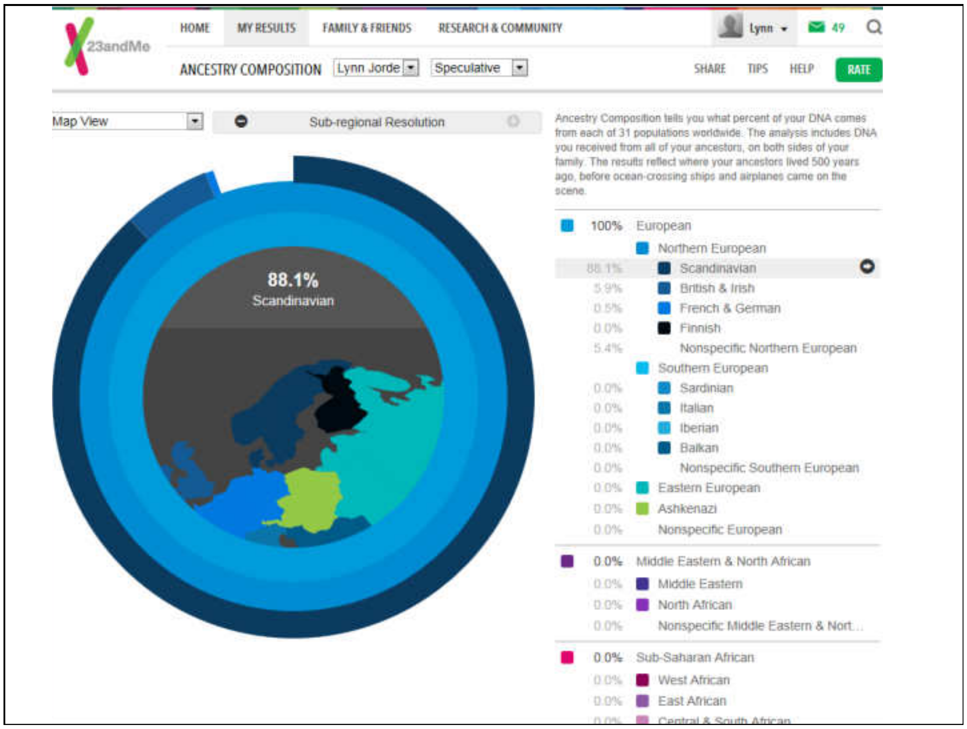
■	Europe 100%
🟧	Asia 0%
🟩	Africa 0%
■	Not Genotyped

Worldwide Examples

Click on the icons in the map below to see example paintings of individuals from across the globe.

Tell Me About...

...using Ancestry Painting.
...the three reference populations.
...why only three populations are used.
...the people linked to my account.
...why it says I'm European/African/Asian when I'm really an American/Australian/South African.
...how the percentages are calculated.
...where the X and Y chromosomes are.

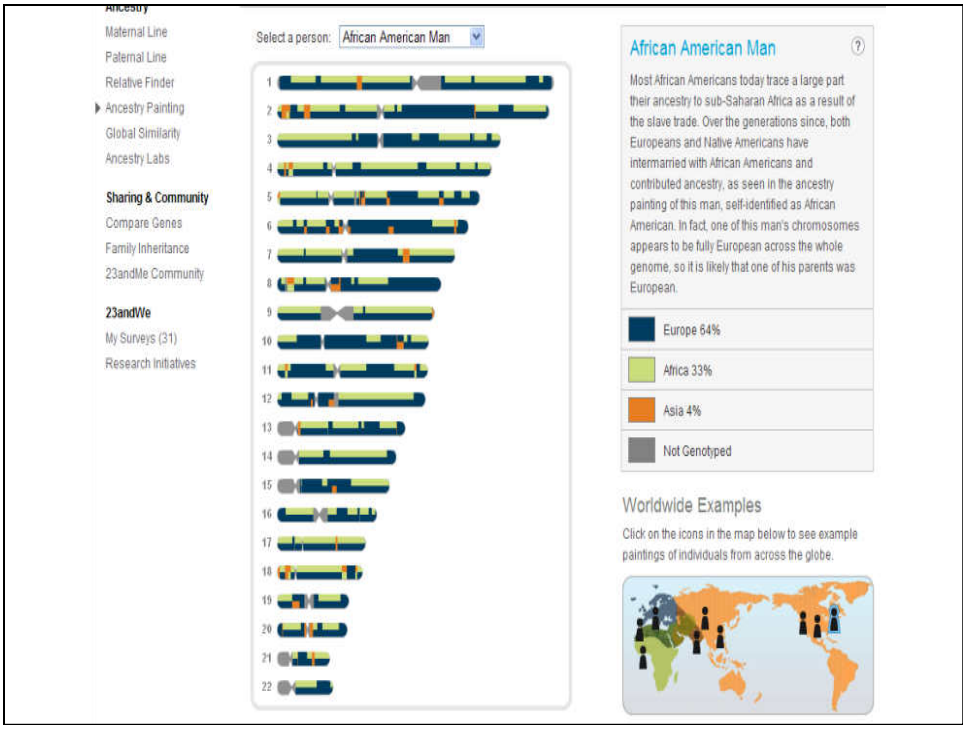


ANCESTRY COMPOSITION Lynn Jorde Speculative

Map View Sub-regional Resolution

Ancestry Composition tells you what percent of your DNA comes from each of 31 populations worldwide. The analysis includes DNA you received from all of your ancestors, on both sides of your family. The results reflect where your ancestors lived 500 years ago, before ocean-crossing ships and airplanes came on the scene.

■	100%	European
		■ Northern European
	88.1%	■ Scandinavian
	5.9%	■ British & Irish
	0.5%	■ French & German
	0.0%	■ Finnish
	5.4%	Nonspecific Northern European
		■ Southern European
	0.0%	■ Sardinian
	0.0%	■ Italian
	0.0%	■ Iberian
	0.0%	■ Balkan
	0.0%	Nonspecific Southern European
	0.0%	■ Eastern European
	0.0%	■ Ashkenazi
	0.0%	Nonspecific European
■	0.0%	Middle Eastern & North African
	0.0%	■ Middle Eastern
	0.0%	■ North African
	0.0%	Nonspecific Middle Eastern & Nort...
■	0.0%	Sub-Saharan African
	0.0%	■ West African
	0.0%	■ East African
	0.0%	■ Central & South African



Select a person: African American Man

African American Man

Most African Americans today trace a large part their ancestry to sub-Saharan Africa as a result of the slave trade. Over the generations since, both Europeans and Native Americans have intermarried with African Americans and contributed ancestry, as seen in the ancestry painting of this man, self-identified as African American. In fact, one of this man's chromosomes appears to be fully European across the whole genome, so it is likely that one of his parents was European.

■ Europe 64%

■ Africa 33%

■ Asia 4%

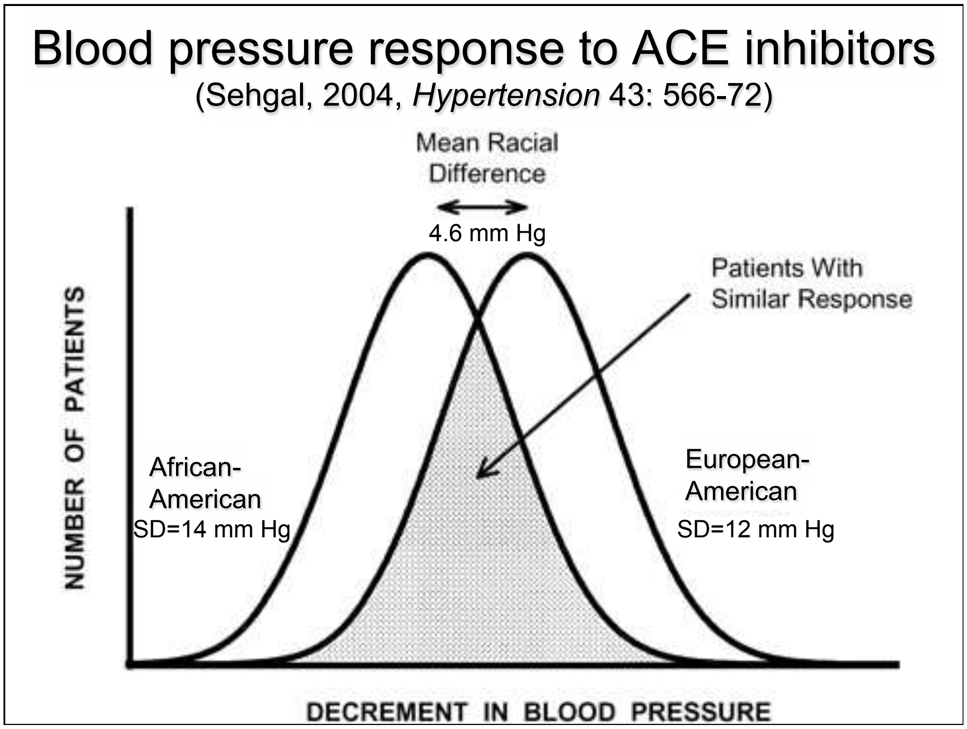
■ Not Genotyped

Worldwide Examples

Click on the icons in the map below to see example paintings of individuals from across the globe.

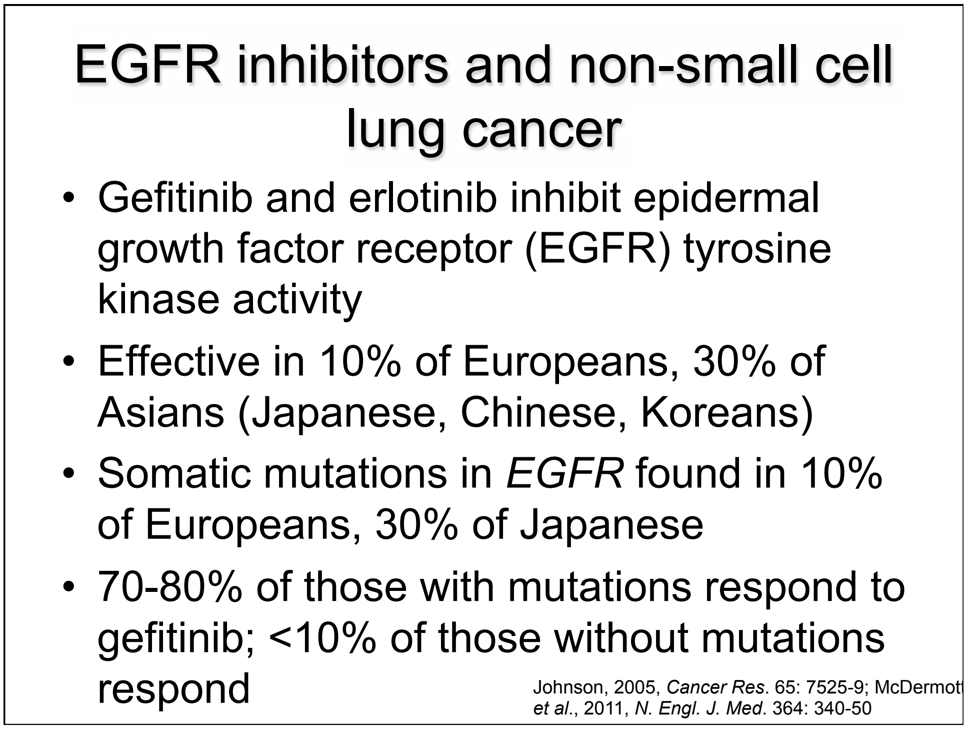
What do these findings imply for biomedicine?

- Large numbers of independent DNA polymorphisms can inform us about ancestry and population history
- 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)



EGFR inhibitors and non-small cell lung cancer

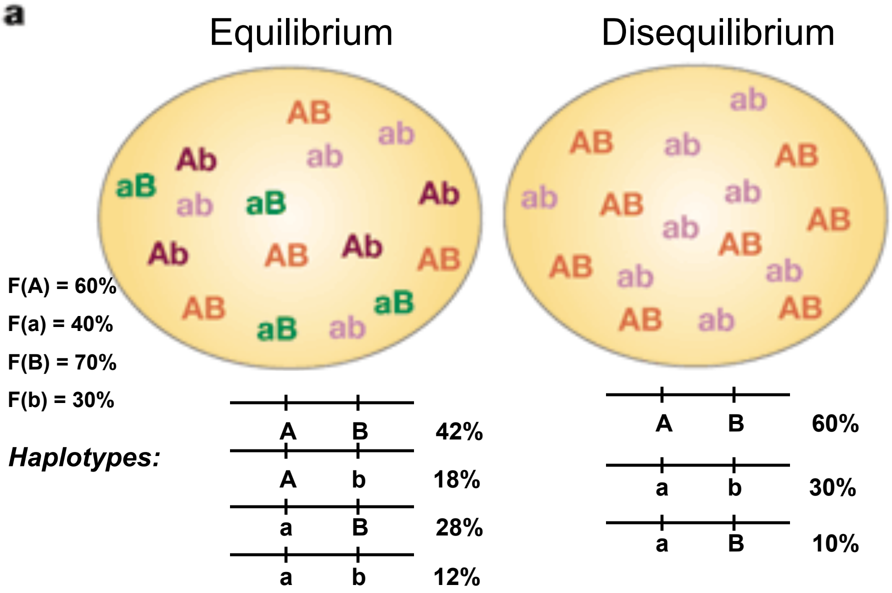
- Gefitinib and erlotinib inhibit epidermal growth factor receptor (EGFR) tyrosine kinase activity
- Effective in 10% of Europeans, 30% of Asians (Japanese, Chinese, Koreans)
- Somatic mutations in *EGFR* found in 10% of Europeans, 30% of Japanese
- 70-80% of those with mutations respond to gefitinib; <10% of those without mutations respond

Johnson, 2005, *Cancer Res*. 65: 7525-9; McDermott *et al*., 2011, *N. Engl. J. Med*. 364: 340-50

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
- Individual ancestry provides more medically useful information

Linkage disequilibrium and disease-gene mapping: nonrandom association of alleles at linked loci


a
Equilibrium
Disequilibrium
AB ab Ab ab aB ab aB Ab Ab AB Ab AB AB aB ab aB
ab AB ab AB ab AB ab AB ab AB AB ab ab AB ab AB

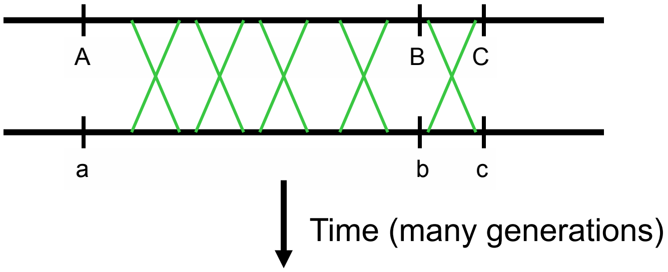
F(A) = 60%
F(a) = 40%
F(B) = 70%
F(b) = 30%

Haplotypes:

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

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

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

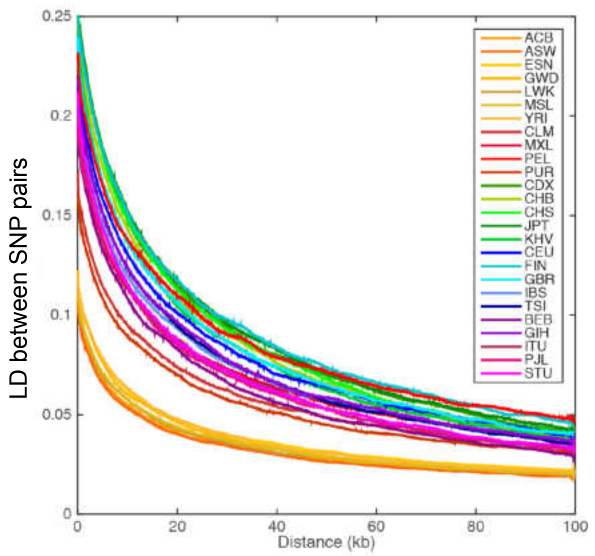


B and C will be found together on the same haplotype more often than A and B: there is more *linkage disequilibrium* between B and C than A and B

Factors that May Affect Linkage Disequilibrium Patterns

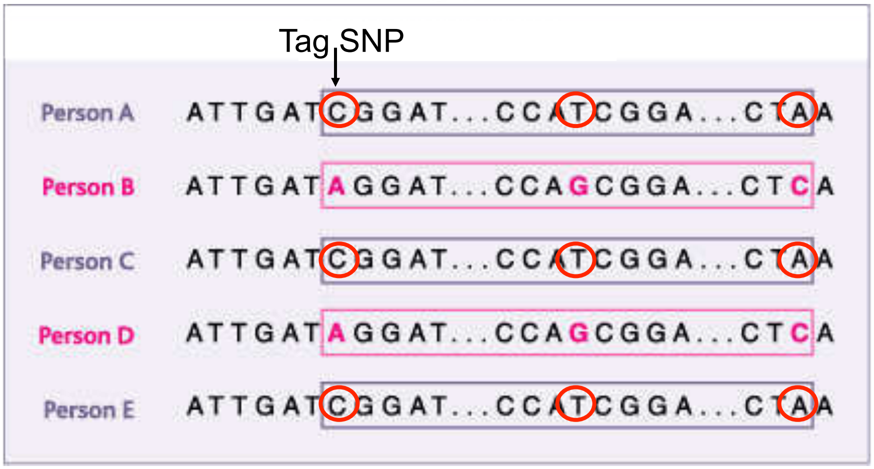
- Chromosome location
  - Telomeric vs. centromeric
  - Intragenic vs. extragenic
- DNA sequence patterns (GC content; presence of *Alu* elements)
- Recombination hotspots (1 every 50-100 kb)
  - 13-mer bound by *PRDM9* associated with 40% of hotspots
- Evolutionary factors: LD varies among populations
  - Natural selection
  - Gene flow
  - Mutation, gene conversion
  - Genetic drift
  - Time elapsed since founding of population

Linkage disequilibrium (LD) decays with physical distance more quickly in "older" populations



Auton et al., 2015, *Nature*
1000 Genomes data

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



Person A ATTGATCGGAT...CCATCGGA...CTAA

Person B ATTGATAGGAT...CCAGCGGA...CTCA

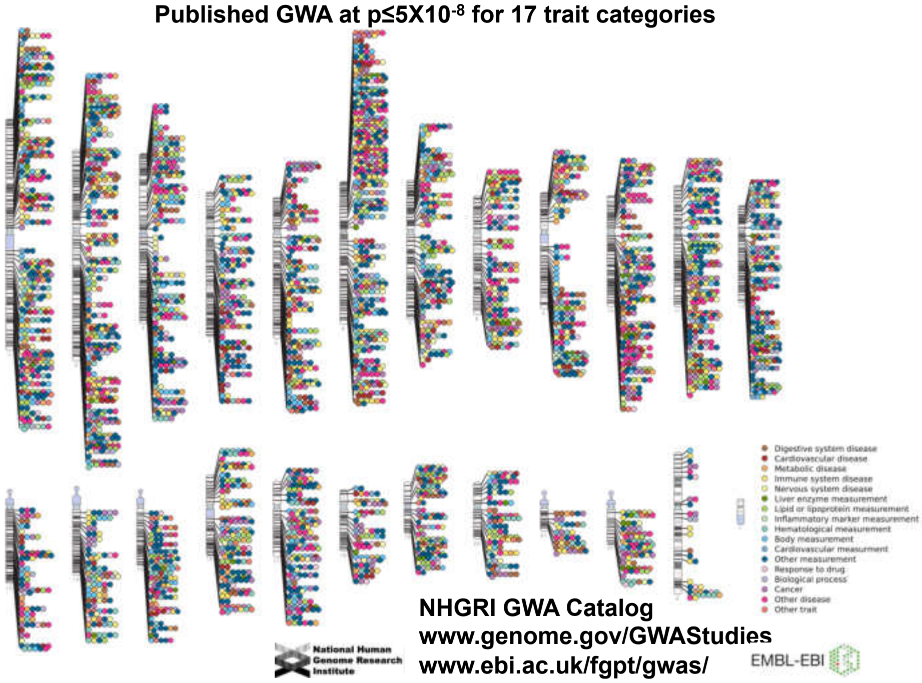
Person C ATTGATCGGAT...CCATCGGA...CTAA

Person D ATTGATAGGAT...CCAGCGGA...CTCA

Person E ATTGATCGGAT...CCATCGGA...CTAA

For genome-wide association studies, "complete" coverage is given by about 1.6 million SNPs for African populations, 600,000 to 1M SNPs for non-African populations

Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories



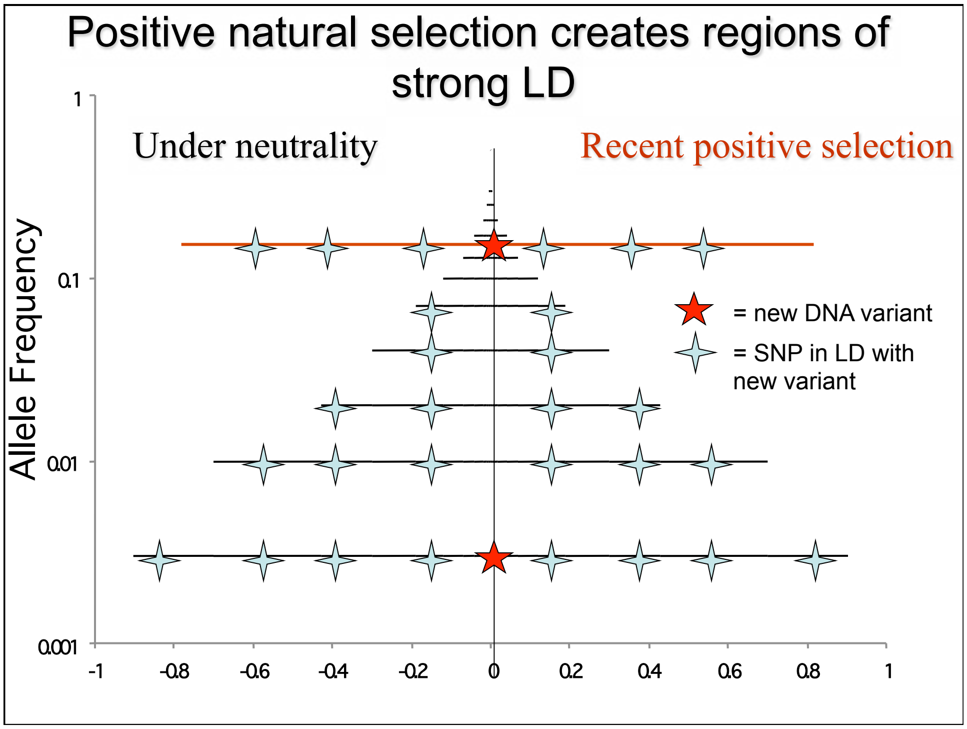
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

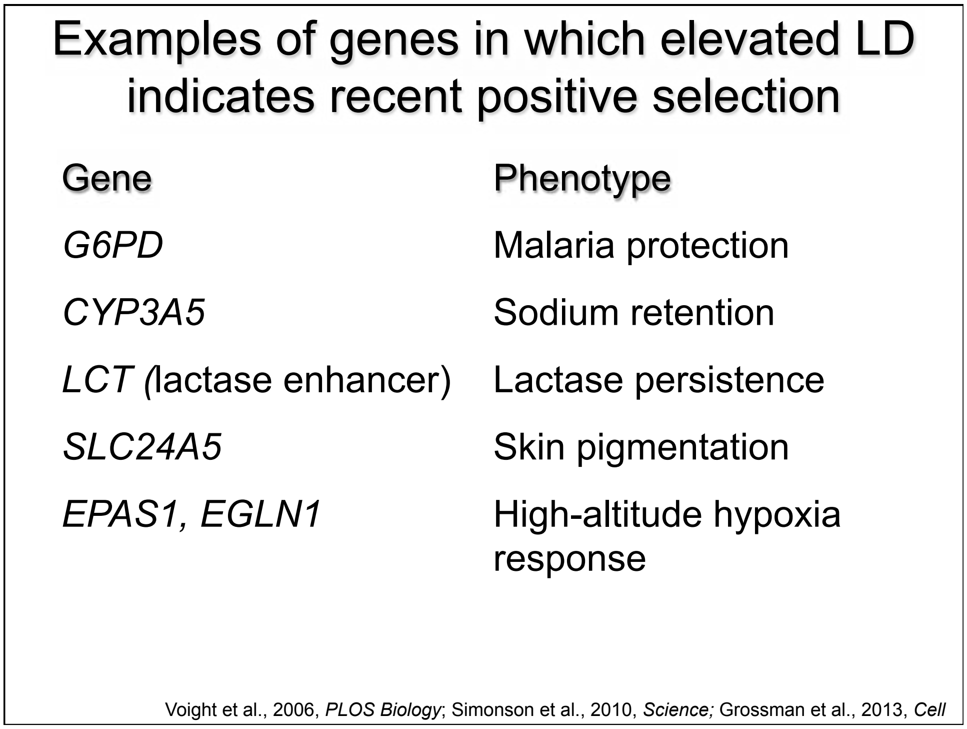
EMBL-EBI

Recombination hotspots

- LD patterns indicate 25,000 - 50,000 hotspots in human genome (1 every 50 – 100 kb) (Myers et al., 2005, *Science*)
- 60% of all recombination occurs in 6% of genome) (Coop et al., 2008, *Science* 319: 1395-8)
- Hotspots are not congruent in human and chimpanzee and vary among human populations



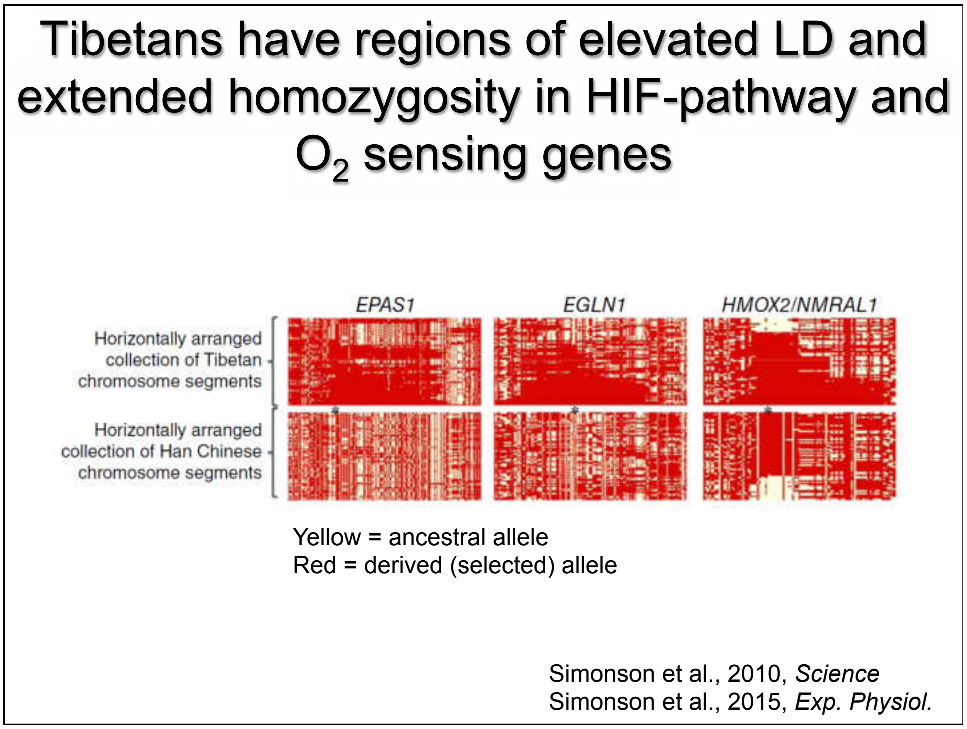
Positive natural selection creates regions of strong LD



Examples of genes in which elevated LD indicates recent positive selection

Gene	Phenotype
G6PD	Malaria protection
CYP3A5	Sodium retention
LCT (lactase enhancer)	Lactase persistence
SLC24A5	Skin pigmentation
EPAS1, EGLN1	High-altitude hypoxia response

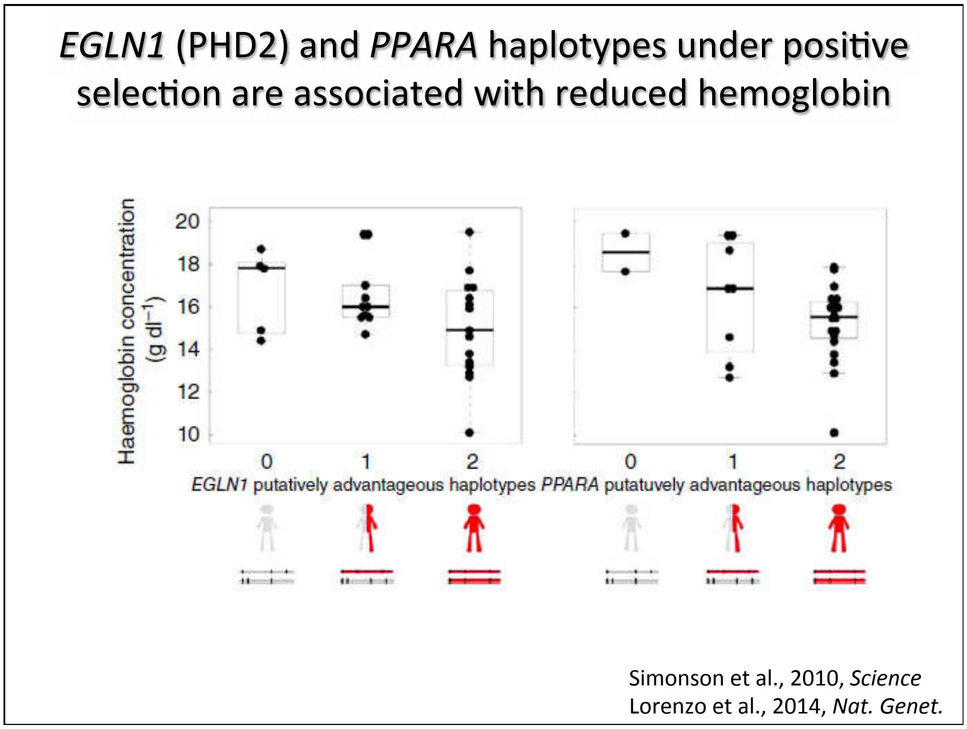
Voight et al., 2006, *PLOS Biology*; Simonson et al., 2010, *Science;* Grossman et al., 2013, *Cell*



Tibetans have regions of elevated LD and extended homozygosity in HIF-pathway and O_2 sensing genes

Yellow = ancestral allele
Red = derived (selected) allele

Simonson et al., 2010, *Science*
Simonson et al., 2015, *Exp. Physiol.*

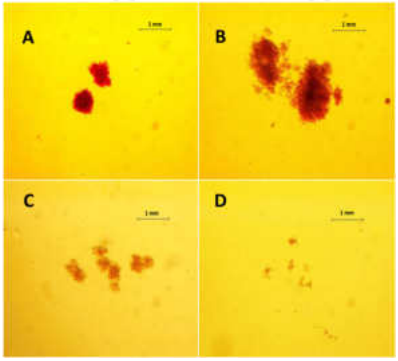


EGLN1 (PHD2) and *PPARA* haplotypes under positive selection are associated with reduced hemoglobin

Simonson et al., 2010, *Science*
Lorenzo et al., 2014, *Nat. Genet.*

Erythroid progenitor cells produce the Tibetan phenotype under hypoxia

Wild type
under
normoxia



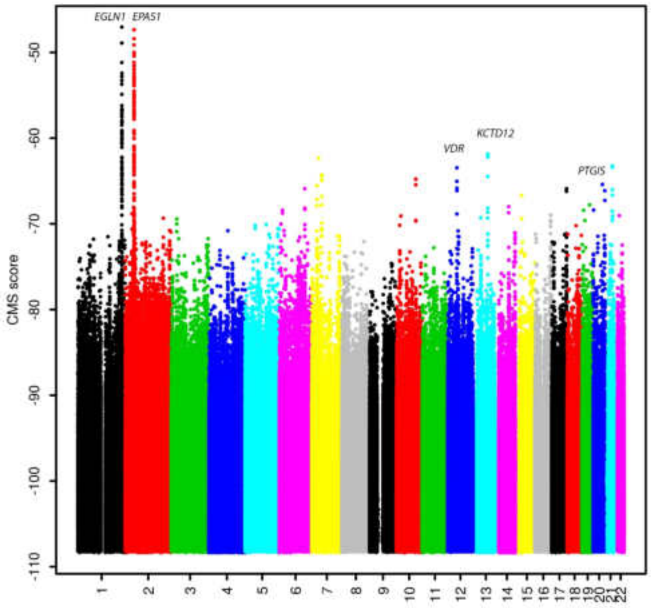
Wild type
under hypoxia
(5% O2)

PHD2^{D4E,C127S}
mutants under
normoxia

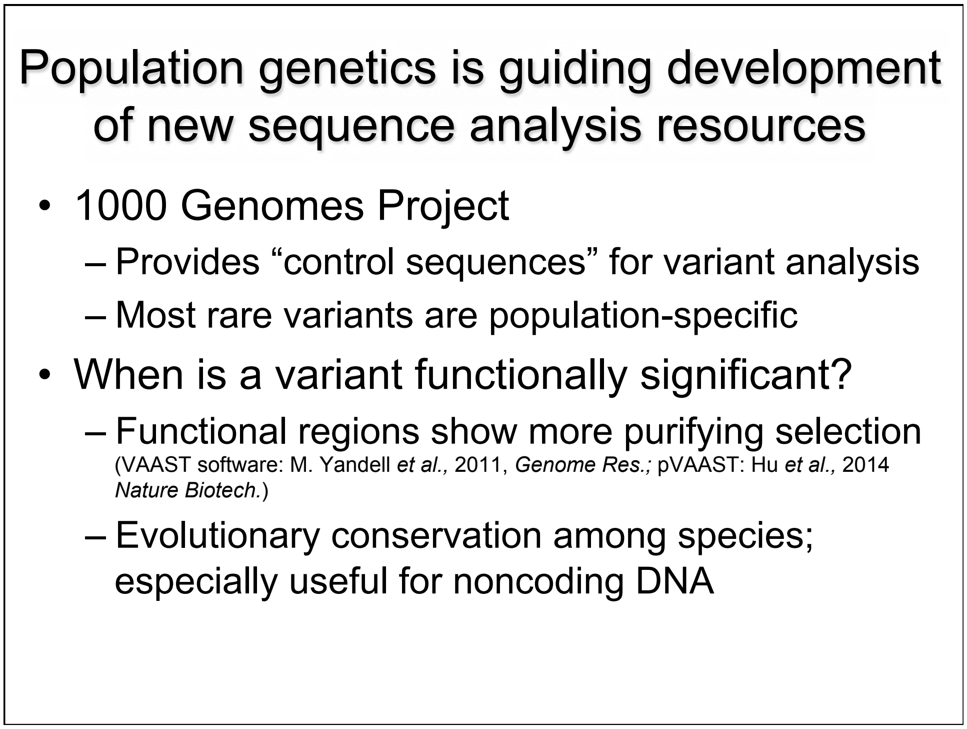
PHD2^{D4E,C127S}
mutants under
hypoxia

PHD2^{D4E,C127S} produces a gain of function under hypoxic conditions, reducing hemoglobin concentration and providing protection from polycythemia .

Composite of Multiple Signals (CMS) test for recent positive selection

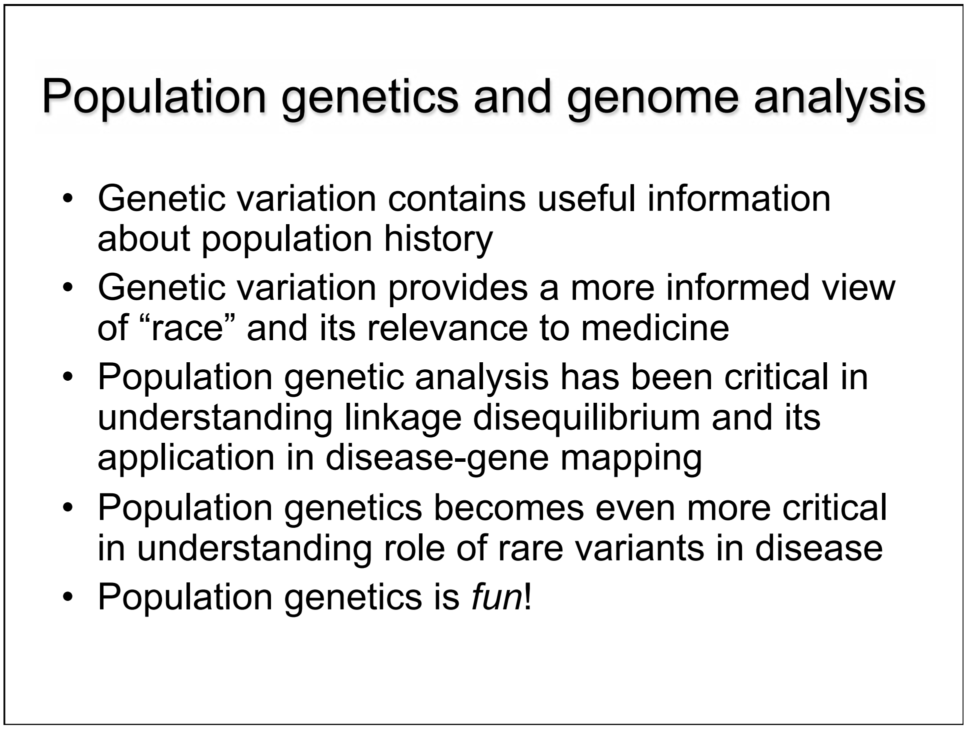


Hu *et al*., *Genome Research* (under review)



Population genetics is guiding development of new sequence analysis resources

- 1000 Genomes Project
  - Provides “control sequences” for variant analysis
  - Most rare variants are population-specific
- When is a variant functionally significant?
  - Functional regions show more purifying selection (VAAST software: M. Yandell *et al.,* 2011, *Genome Res.;* pVAAST: Hu *et al.,* 2014 *Nature Biotech.*)
  - Evolutionary conservation among species; especially useful for noncoding DNA



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 becomes even more critical in understanding role of rare variants in disease
- Population genetics is *fun*!