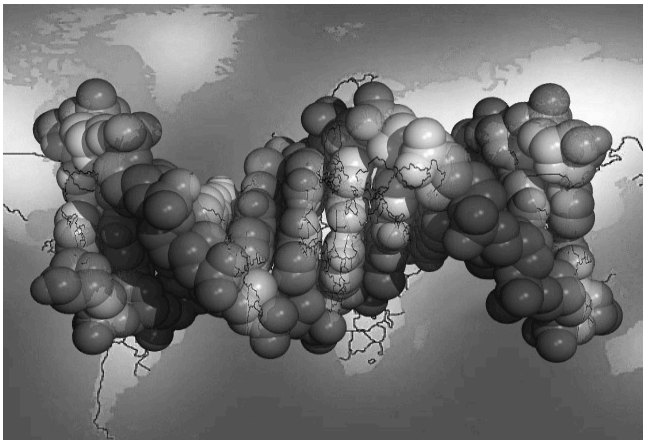
Introduction to Population Genetics



Lynn B. Jorde
Department of Human Genetics
University of Utah School of Medicine
9 April 2014

JOHNS HOPKINS
MEDICINE
CONTINUING MEDICAL EDUCATION

Current Topics in Genome Analysis 2014

Lynn Jorde

No Relevant Financial Relationships with Commercial Interests

Overview

- Patterns of human genetic variation
  - Among populations
  - Among individuals
- "Race" and its biomedical implications
- Linkage disequilibrium and disease-gene identification

Human Genetic Variation: Applications

- Deciphering human history
- Inferring individual ancestry
- Forensics
- Finding and understanding disease-causing genes

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*)

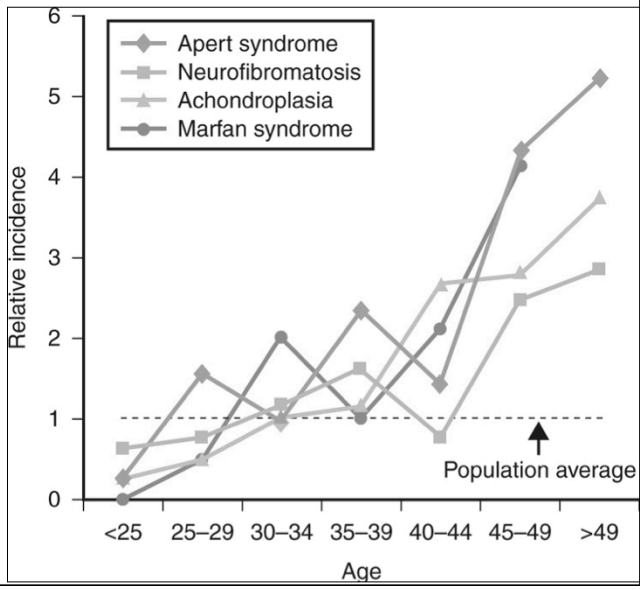
"*The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music."*

- Lewis Thomas

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

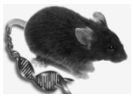
An additional two mutations occur with each year of paternal age (baseline: ~30 mutations in a male aged 30)

(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

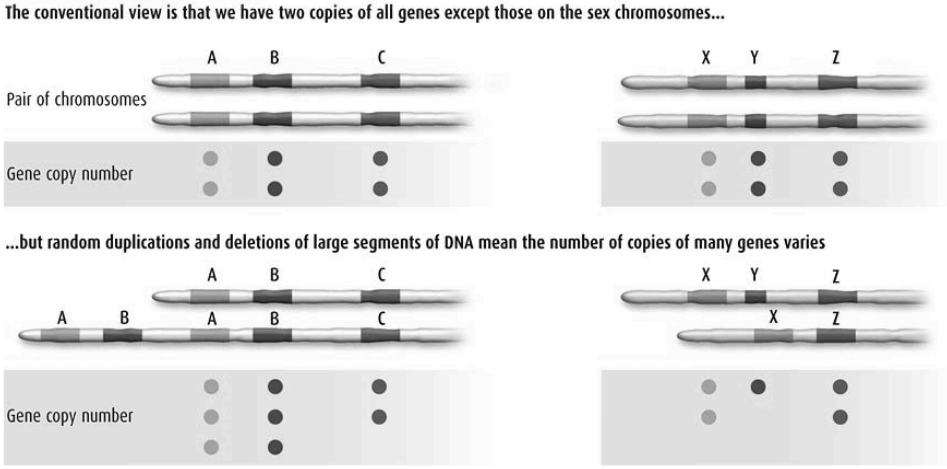
- 3 billion DNA bases → 3 million differences (single nucleotide polymorphisms; SNPs) between each pair of haploid human DNA sequences

Whole-genome sequence diversity in great apes

Species	Sample size	Average number of single nucleotide variants per individual
Homo sapiens	9	3,061,604
Pan troglodytes (common chimpanzee)	24	5,693,903
Gorilla	27	6,492,831
Pongo (orangutan)	10	9,338,148

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

Copy number variants (deletions/duplications > 1000 bp) account for several times more inter-individual variation than do single-nucleotide variants



Each human is heterozygous for at least 100 CNVs

How much do human populations differ?

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.*

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

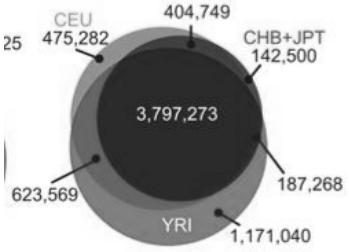
Minor allele present in:	
All 4 groups	78.6%
At least 3 groups	88.0%
At least 2 groups	92.1%
Africa only	7.4%
Any non-African group	0.5%

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

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

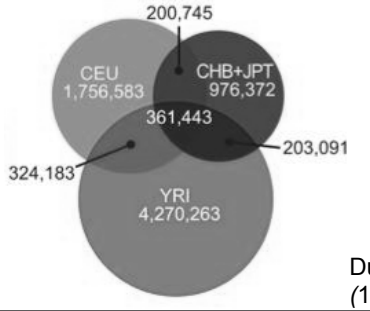
Rare SNVs are much more likely to be population-specific

Common SNVs



Average allele frequency difference between populations: 15%

New rarer SNVs identified by sequencing



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

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 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

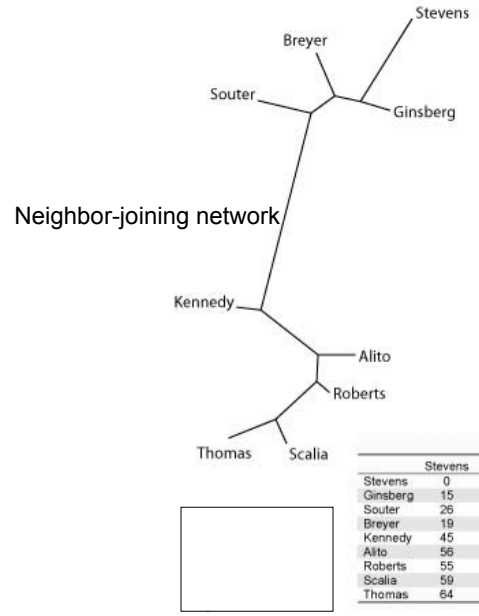
$|p_1 – p_2|$ $| p_3 – (p_1 + p_2)/2 |$

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

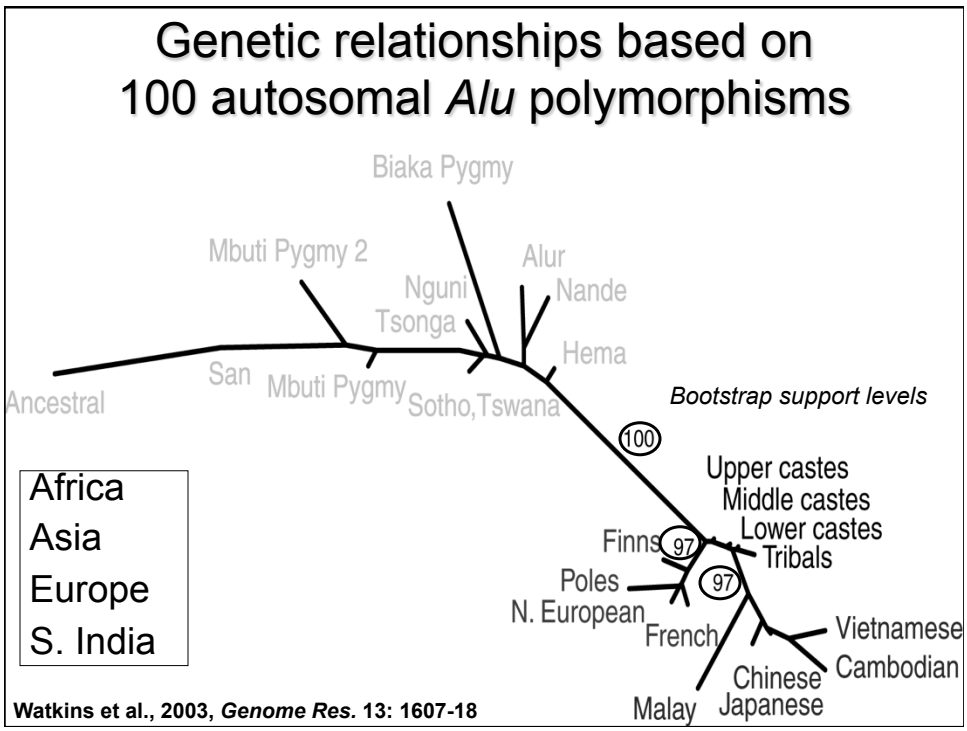
Thanks to: Steve Guthery, MD

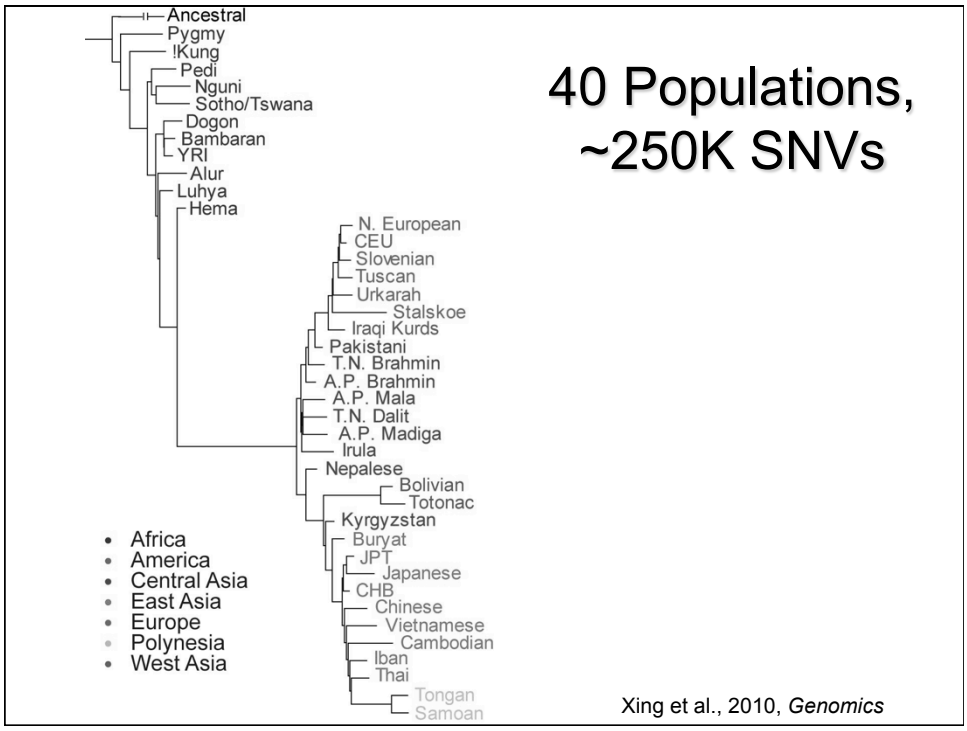


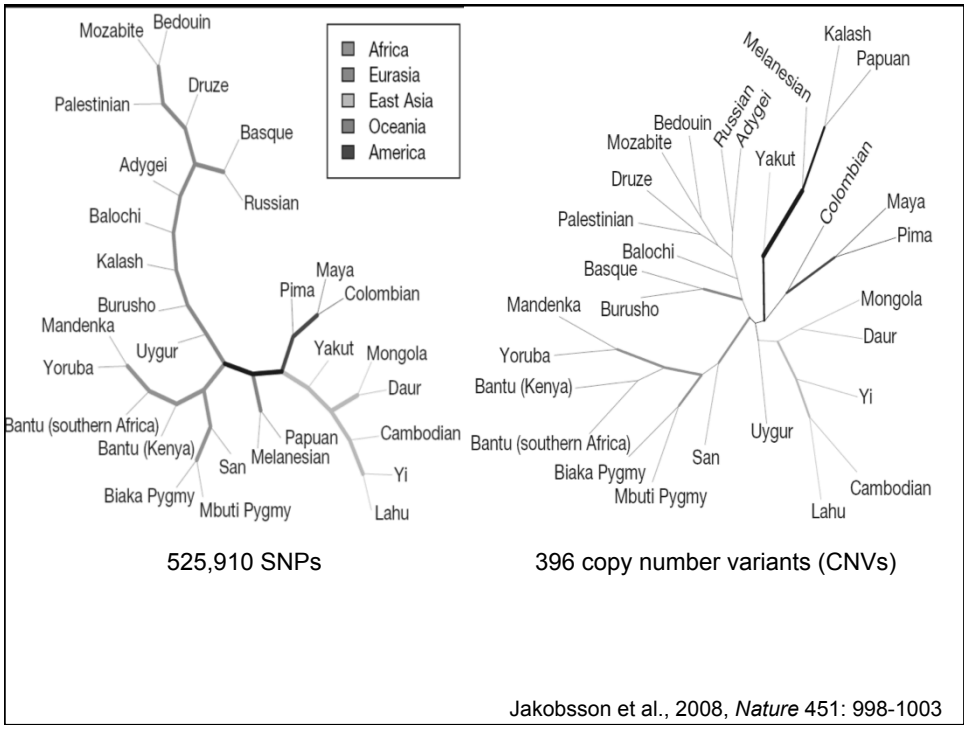
Distance matrix: % disagreement

	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


Genetic relationships based on
100 autosomal *Alu* polymorphisms
Biaka Pygmy
Mbuti Pygmy 2
Alur
Nguni
Nande
Tsonga
Hema
San
Mbuti Pygmy
Sotho,Tswana
Ancestral
Bootstrap support levels
100
Upper castes
Middle castes
Lower castes
Tribals
Finns
97
Poles
97
N. European
French
Vietnamese
Cambodian
Chinese
Japanese
Malay
Africa
Asia
Europe
S. India
Watkins et al., 2003, *Genome Res.* 13: 1607-18


Ancestral
Pygmy
!Kung
Pedi
Nguni
Sotho/Tswana
Dogon
Bambaran
YRI
Alur
Luhya
Hema
N. European
CEU
Slovenian
Tuscan
Urkarah
Stalskoe
Iraqi Kurds
Pakistani
T.N. Brahmin
A.P. Brahmin
A.P. Mala
T.N. Dalit
A.P. Madiga
Irula
Nepalese
Bolivian
Totonac
Kyrgyzstan
Buryat
JPT
Japanese
CHB
Chinese
Vietnamese
Cambodian
Iban
Thai
Tongan
Samoan
Africa
America
Central Asia
East Asia
Europe
Polynesia
West Asia
40 Populations,
~250K SNVs
Xing et al., 2010, *Genomics*

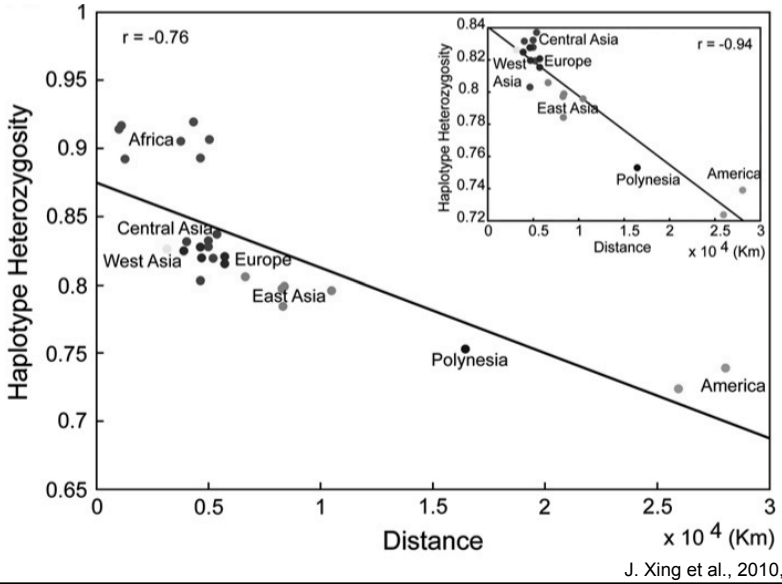


525,910 SNPs

396 copy number variants (CNVs)

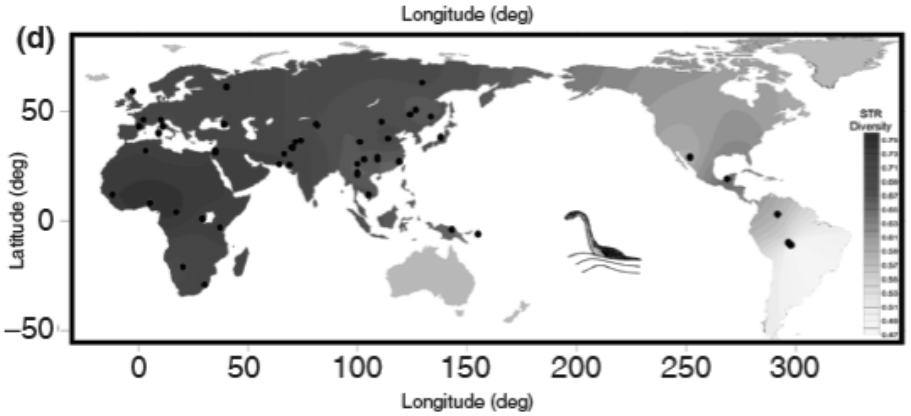
Jakobsson et al., 2008, *Nature* 451: 998-1003

Haplotype diversity declines with distance from Africa



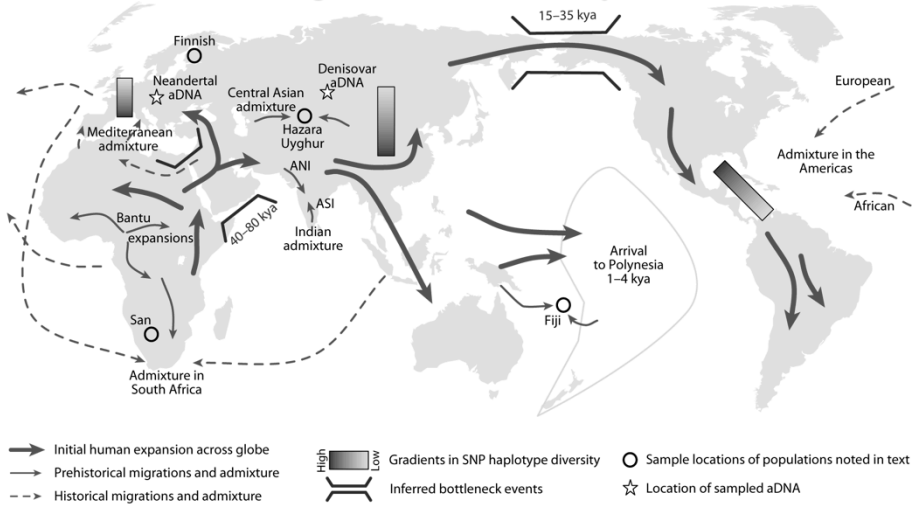
J. Xing et al., 2010, *Genomics*

Serial founder effect



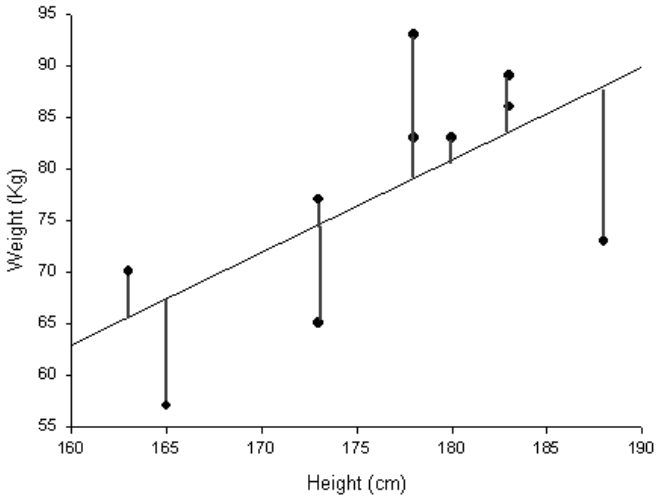
Colonna et al., 2011, *Genome Biol.*

Recent African origin of anatomically modern humans

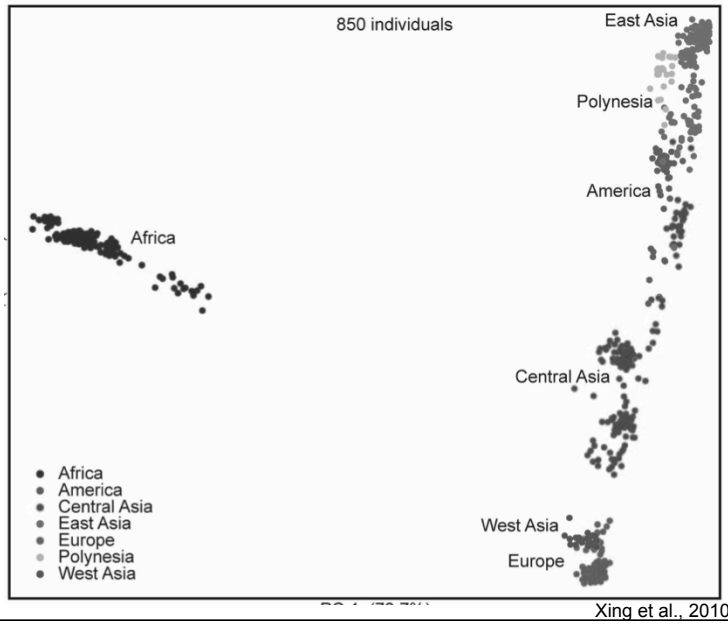


Novembre J, Ramachandran S. 2011.
Annu. Rev. Genomics Hum. Genet. 12:245–74

Principal components analysis: a multidimensional regression technique

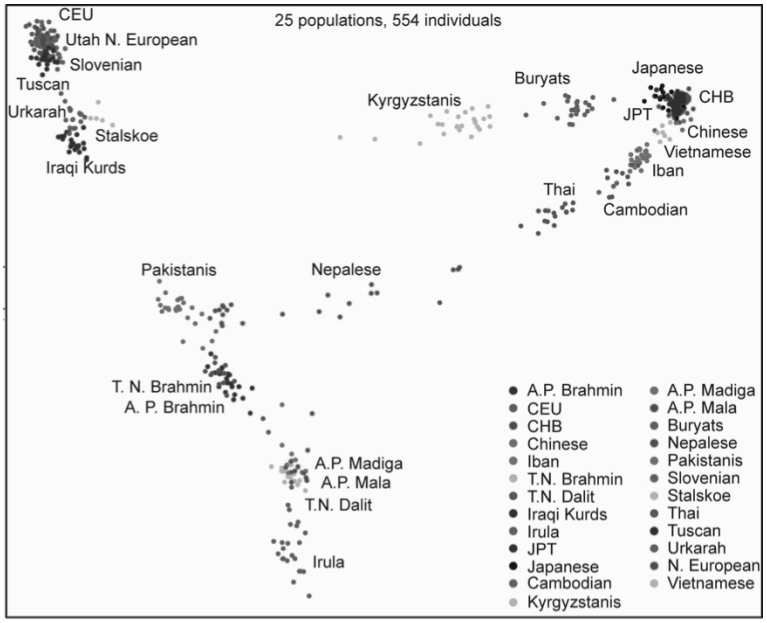


Principal components analysis displays **individual** genetic similarity in 2D: each dot = 1 individual



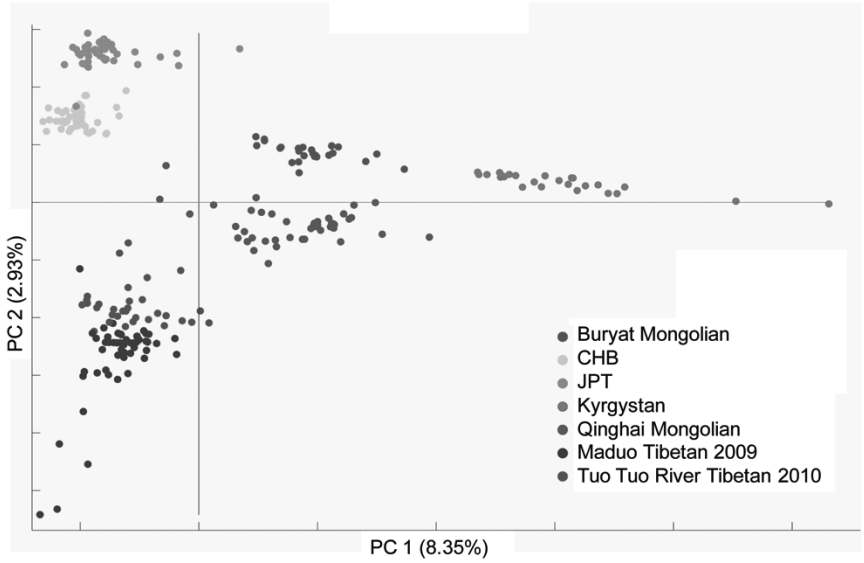
Xing et al., 2010, *Genomics*

PCA: Eurasian Populations

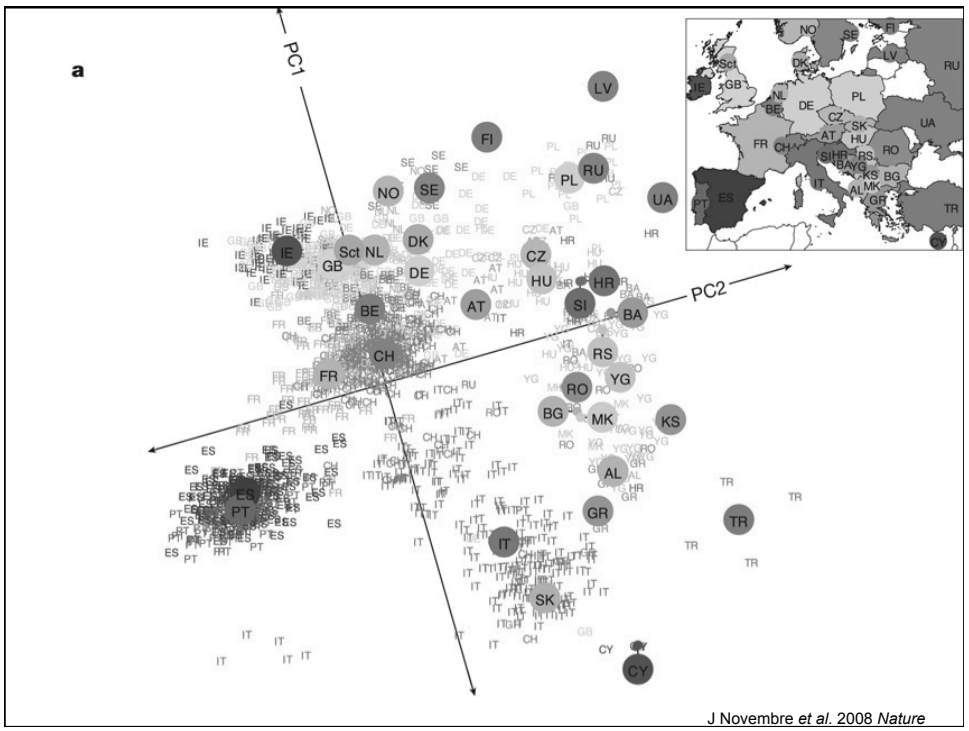


Xing et al., 2010, *Genomics*

PCA can distinguish closely related populations
1 million SNP microarray

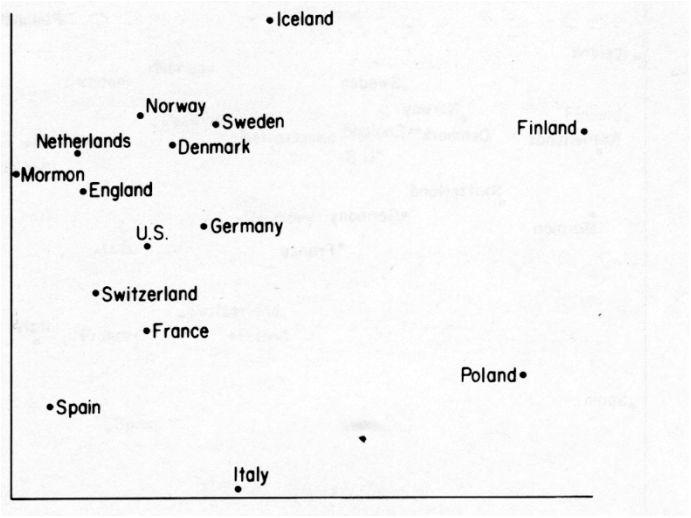


Xing *et al.*, 2013 *PLoS Genetics*



J Novembre *et al.* 2008 *Nature*

Genetic distance analysis: 15 loci

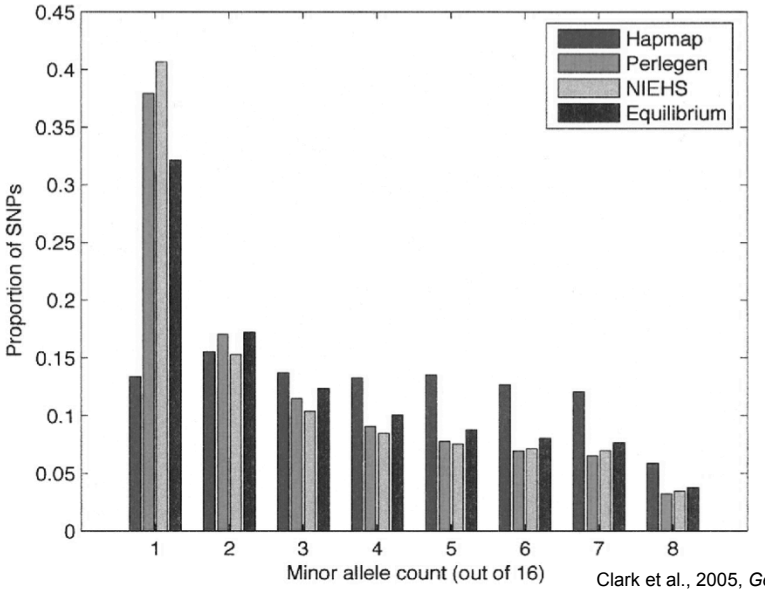


McLellan, Jorde, and Skolnick, 1984, *Am. J. Hum. Genet.*

Sequence data permit more accurate inferences about population history

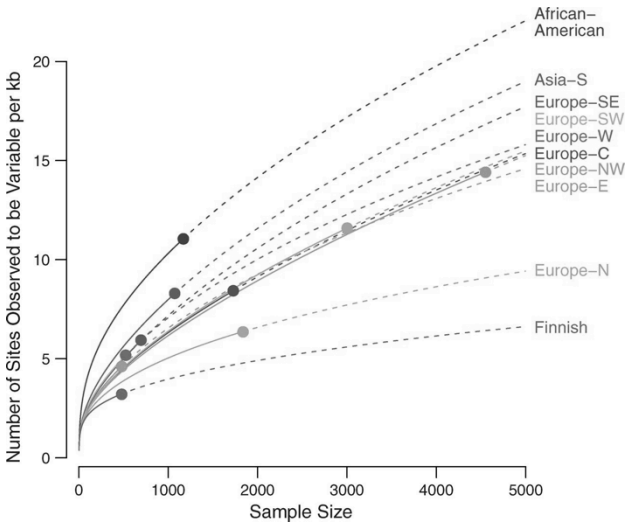
- Microarray SNVs are selected for higher frequency and diversity in Europeans
- Complete DNA sequences are unbiased and include information about rare variants

The effect of ascertainment bias on allele frequencies



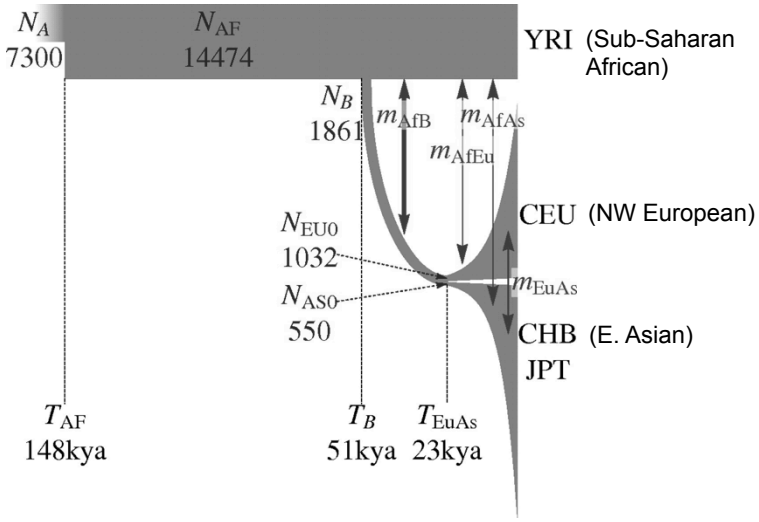
Clark et al., 2005, *Genome Res.* 15: 1496-1502

Number of variants per kilobase of sequence with sample sizes increasing to 5000 people for multiple populations



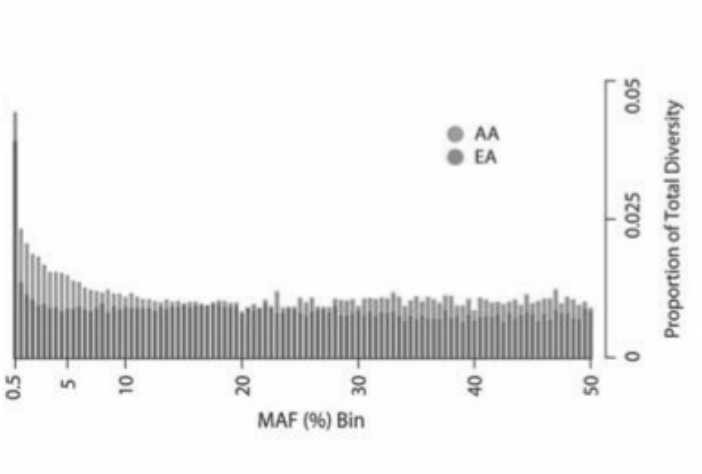
M R Nelson et al. Science 2012;337:100-104

An inferred demographic model, with line width corresponding to population size and time flowing from left to right.



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

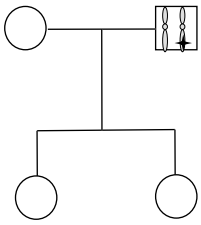
Allele frequency spectrum



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*)*

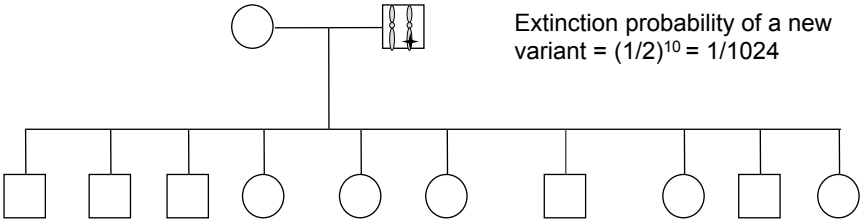
Tennessen *et al.,* 2012, *Science*

Population expansions increase the frequency of rare variants



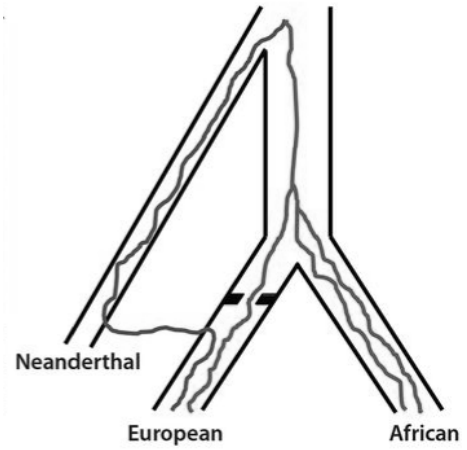
Extinction probability of a new variant = $(1/2)^2 = 1/4$

Population expansions increase the frequency of rare variants



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

Neandertal admixture with anatomically modern humans

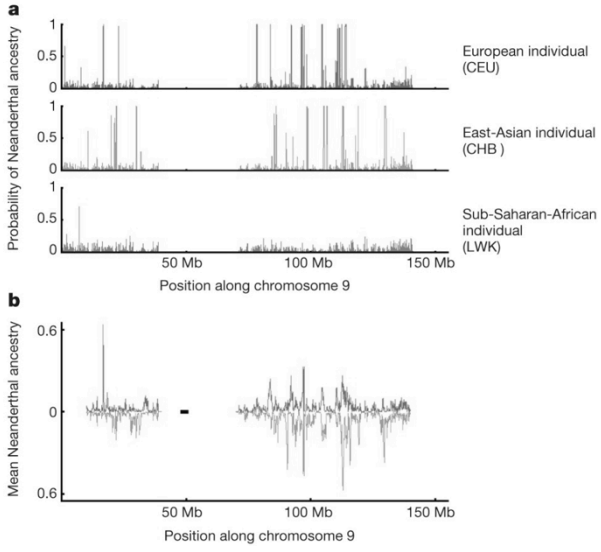


B Vernot and J M Akey, 2014 *Science*

Evidence for mixture between Neandertals and modern humans

- Evidence for mixture from nuclear sequence: 1-4% of modern human DNA has Neandertal origins (Green et al., 2010, *Science*)
- Only non-Africans share DNA with Neanderthals
- Neandertal DNA sharing is seen in all non-African populations
- Could some of the shared sequences have adaptive significance?

Maps of Neandertal ancestry



S Sankararaman *et al.,* 2014 *Nature*

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*

Bamshad and Olson, 2003

Tabulation of DNA sequence differences among individuals

Tabulation of DNA sequence differences among individuals

TTGCAGCTCTCC
TTGCAGCTCTCC

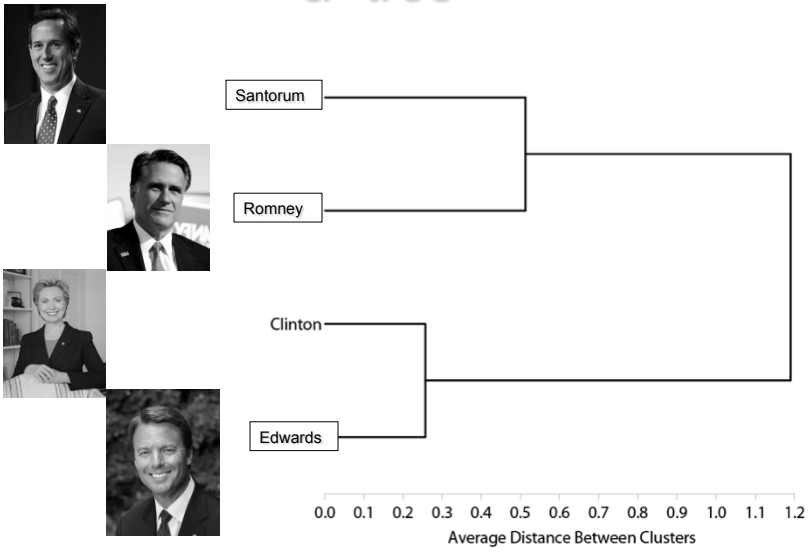
TTGCAGCTCTCC
ATGCAGCTCTCG

ATGCAGCTCTCG
ATGCTGCTCTCG

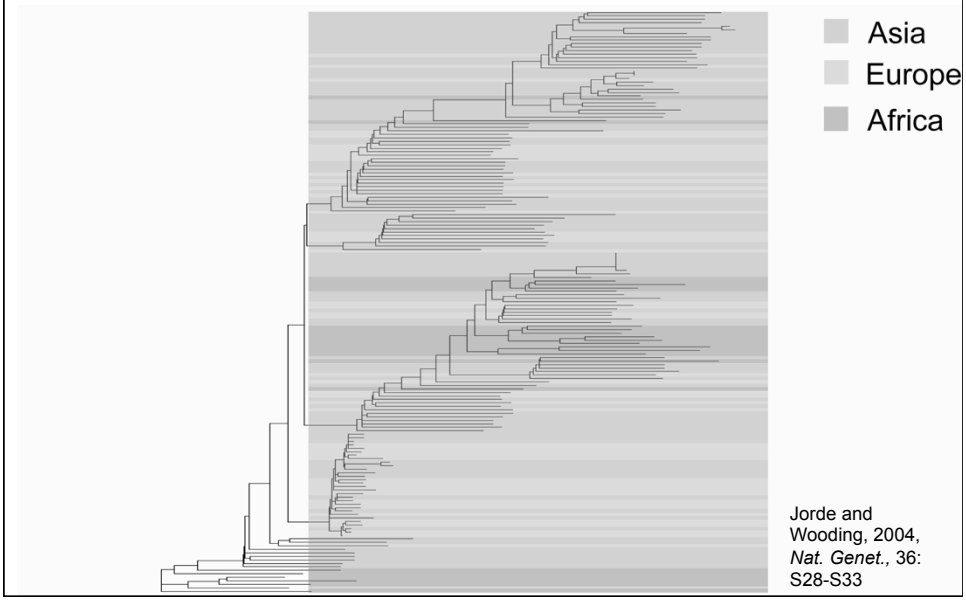
ATGCTGCTCTCG
ATGCTGCTCTCG

	Santorum	Romney	Clinton	Edwards
Santorum	0	.	.	.
Romney	2	0	.	.
Clinton	5	3	0	.
Edwards	6	4	1	0

DNA differences can be summarized in a "tree"



Individual network: 14 kb sequence in angiotensinogen gene

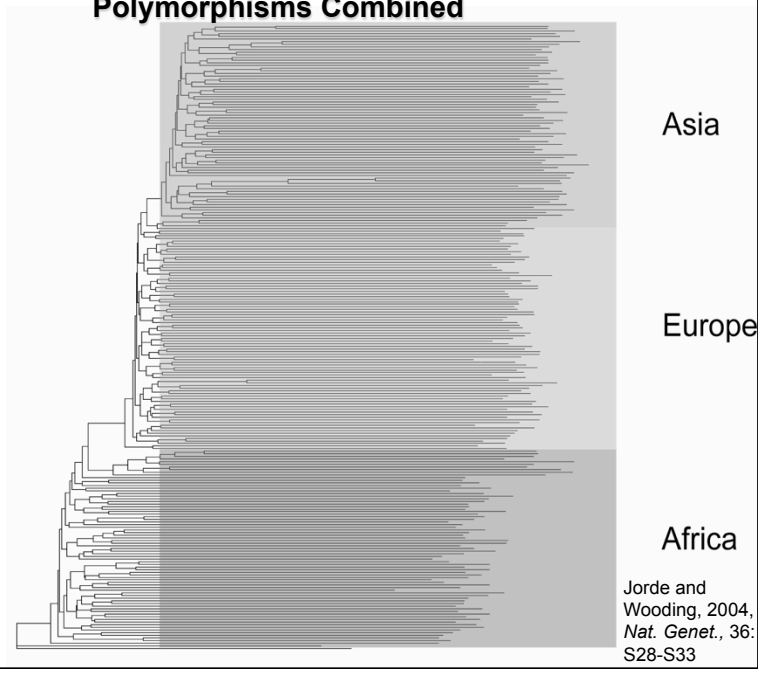


Jorde and Wooding, 2004, *Nat. Genet.*, 36: S28-S33

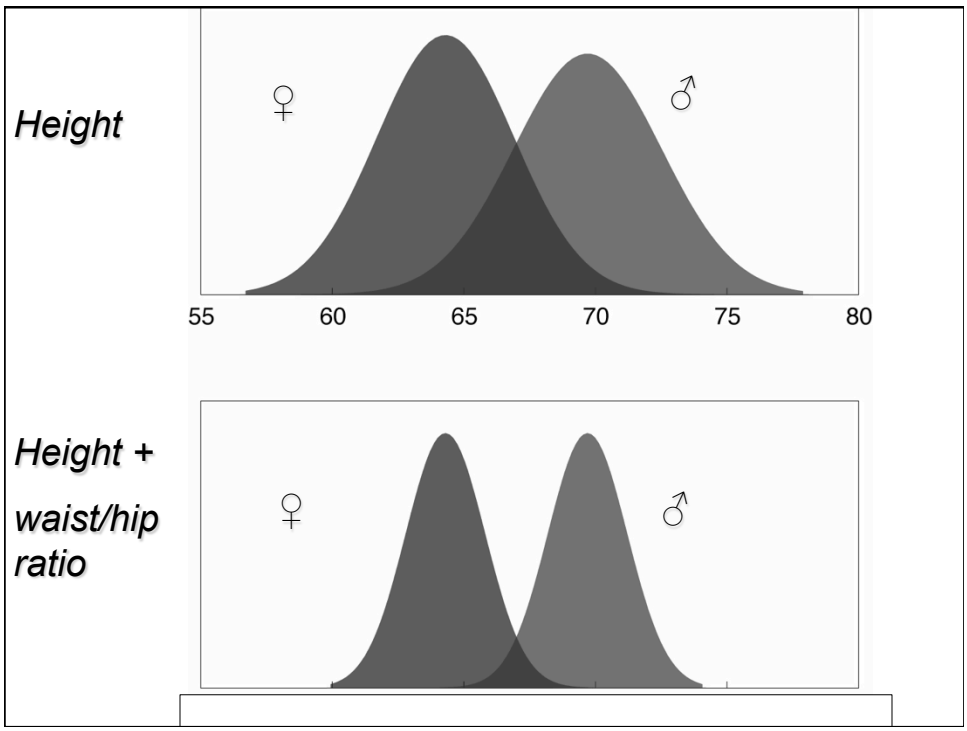
"It may be doubted whether any character can be named which is distinctive of a race and is constant."

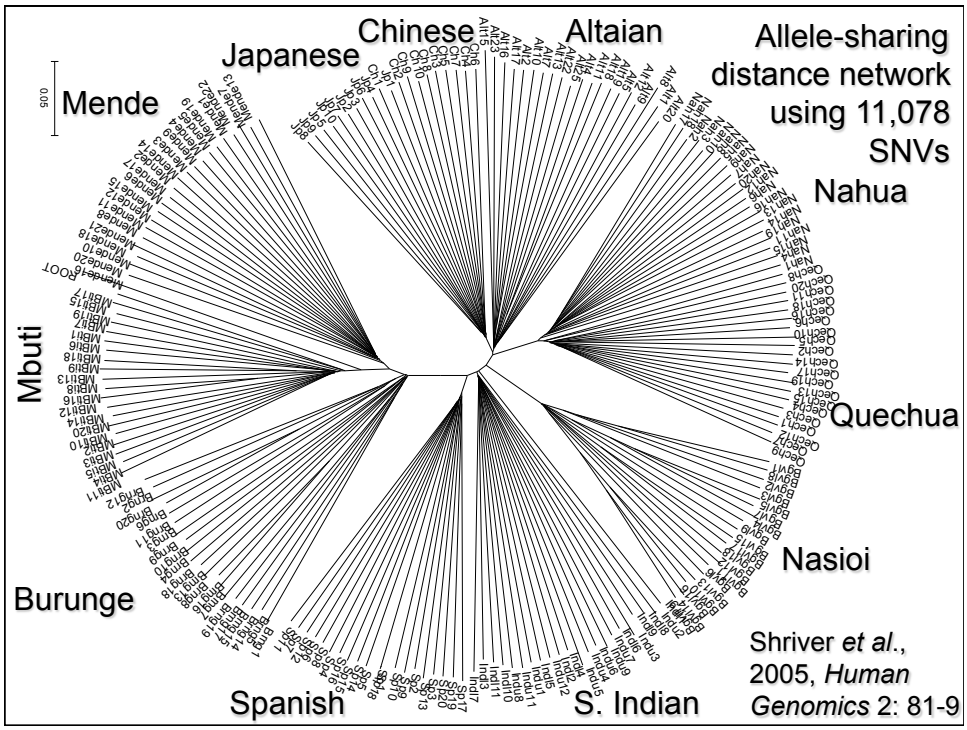
-- Charles Darwin, 1871, *The Descent of Man, and Selection in Relation to Sex*

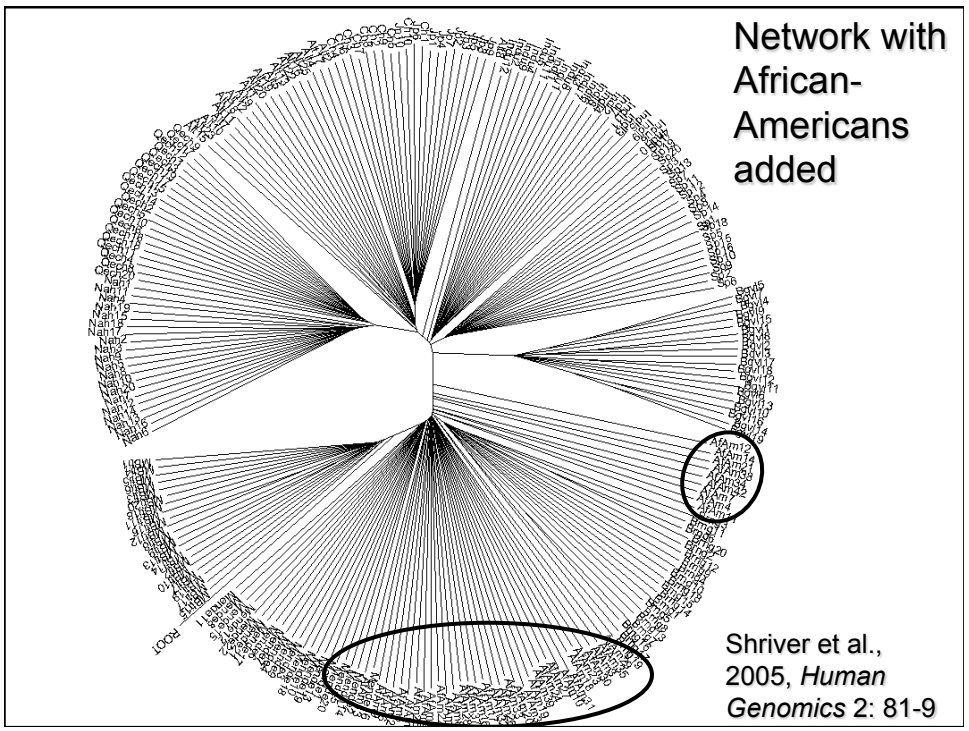
Individual Network: 190 *Alu*, STR, and Restriction Site Polymorphisms Combined

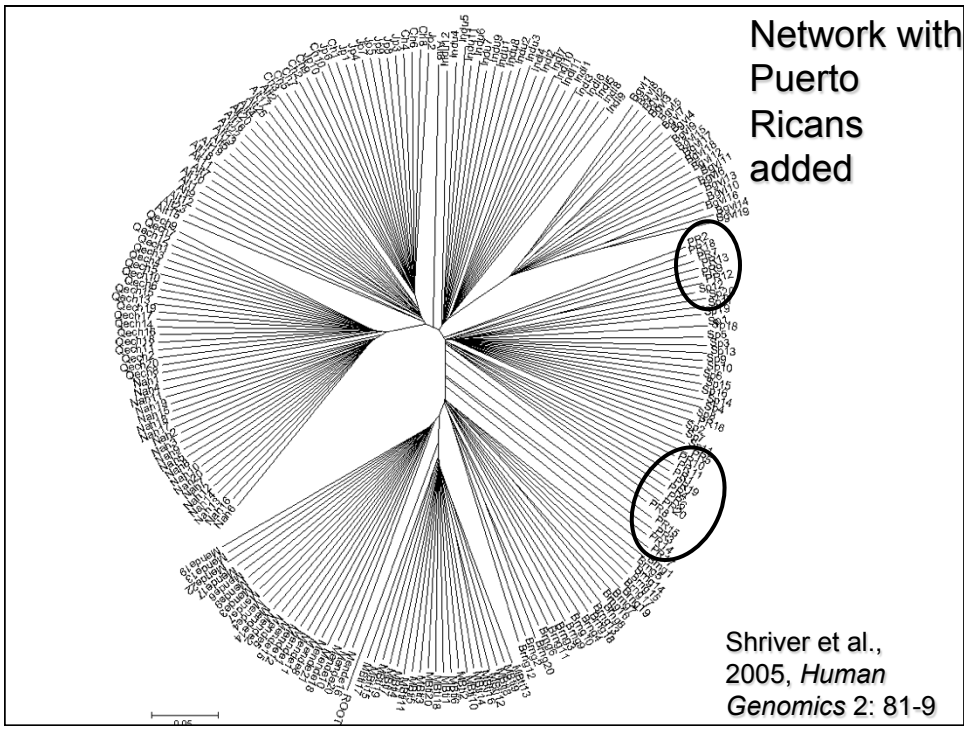


Jorde and Wooding, 2004, *Nat. Genet.*, 36: S28-S33

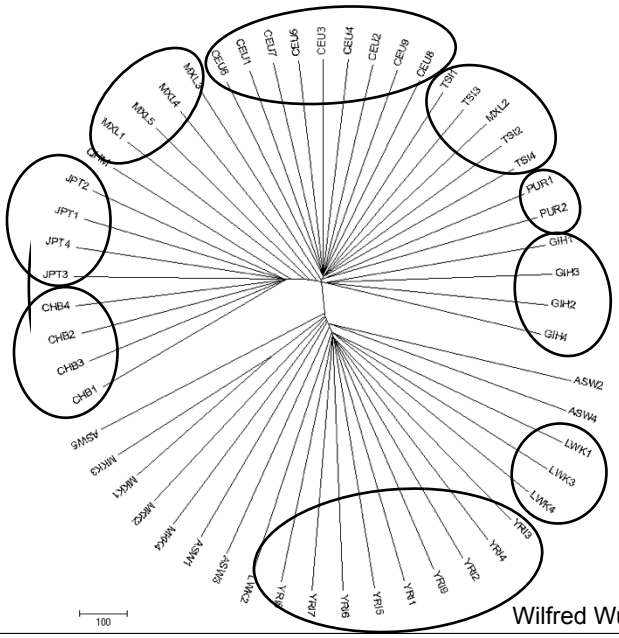

Height
♀
♂
55
60
65
70
75
80
Height +
waist/hip
ratio
♀
♂


Chinese
Altaian
Japanese
Allele-sharing
distance network
using 11,078
SNVs
0.05
Mende
Nahua
Mbuti
Quechua
Nasioi
Burunge
Spanish
S. Indian
Shriver *et al*.,
2005, *Human
Genomics* 2: 81-9


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

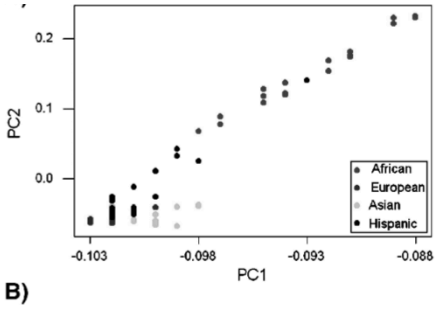
Individual network based on CGI WGS data: 54 individuals



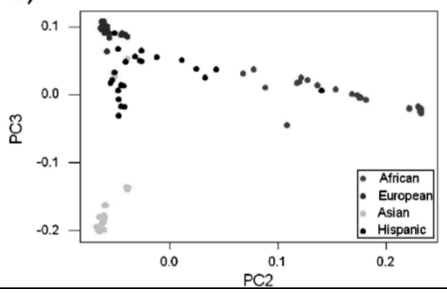
CHB: Chinese
JPB: Japanese
MXL: Mexican
CEU: Utah CEPH
TSI: Tuscan
PUR: Puerto Rican
GIH: Gujarati
ASW: African-American
LWK: Luhya
YRI: Yoruban
MKK: Maasai

Wilfred Wu, MD, PhD

Genetic variation in four American populations (134,000 SNVs)

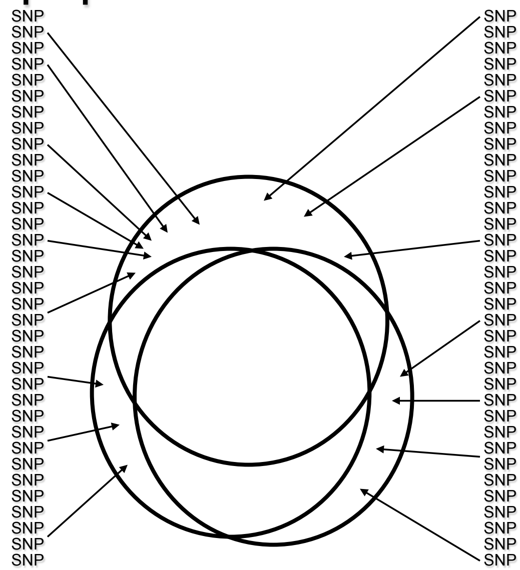


B)



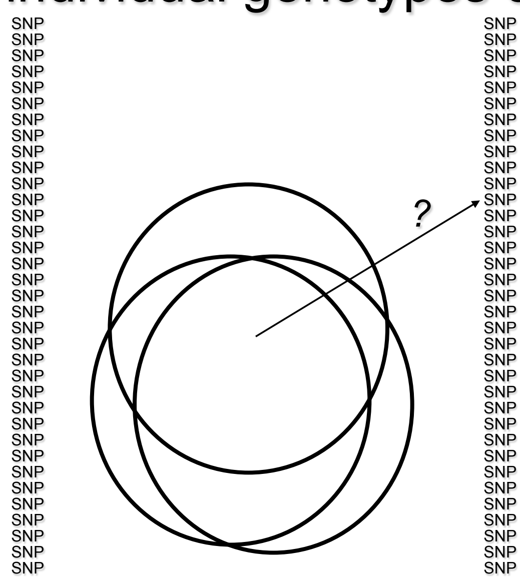
Hughes *et al.*, 2008, *Genomics*

Multiple polymorphisms can predict population affiliation

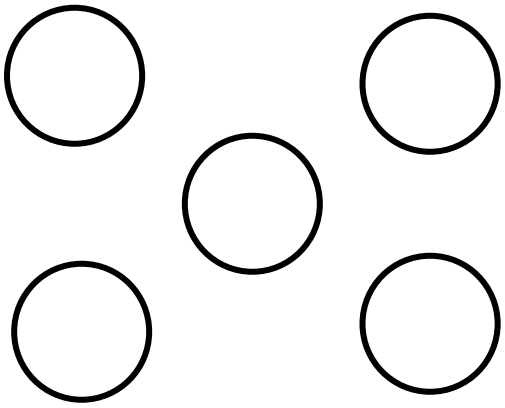


Population affiliation cannot accurately predict individual genotypes or traits

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



The Fallacy of Typological Thinking

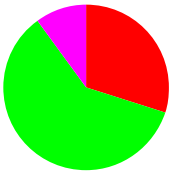


Race as a predictor of ancestry proportions

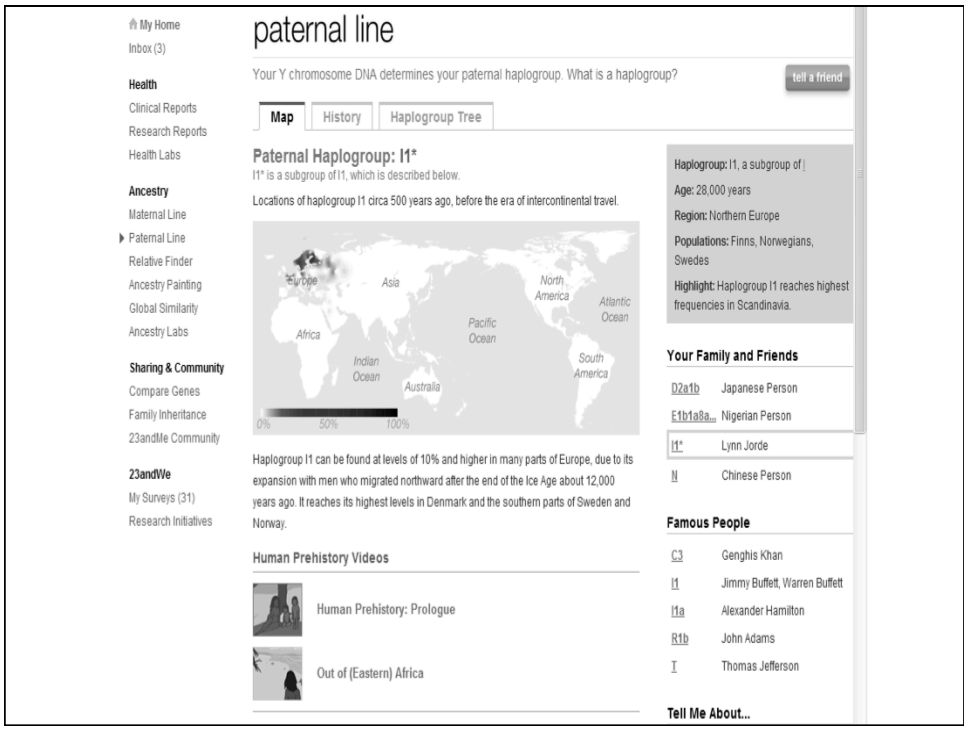
Wayne Joseph

Ancestry vs. Race

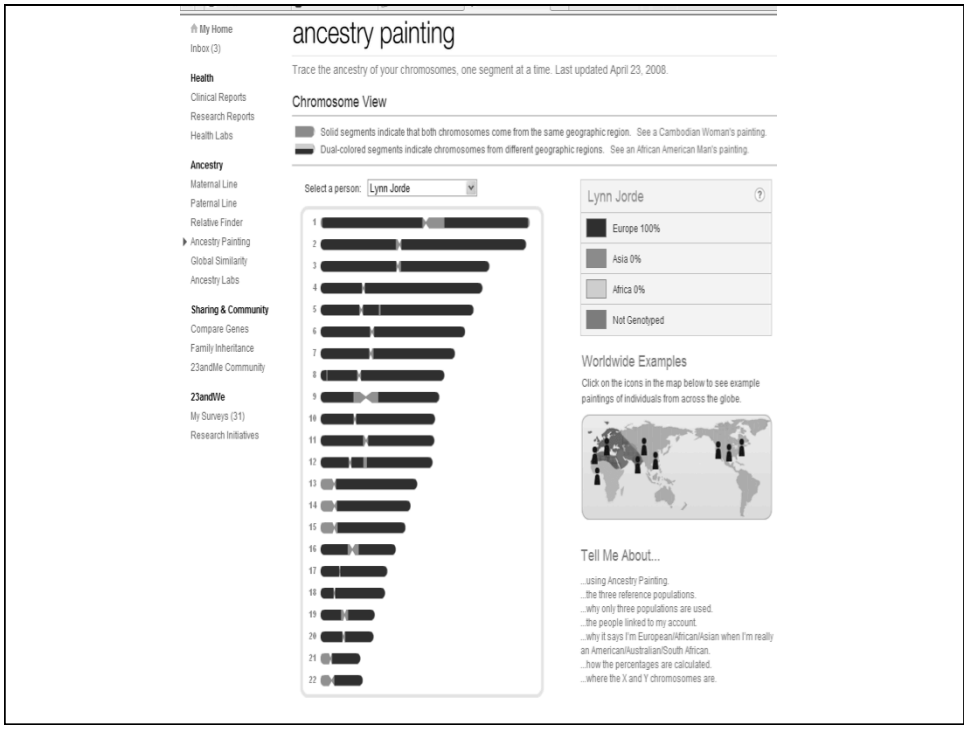
"African-American"

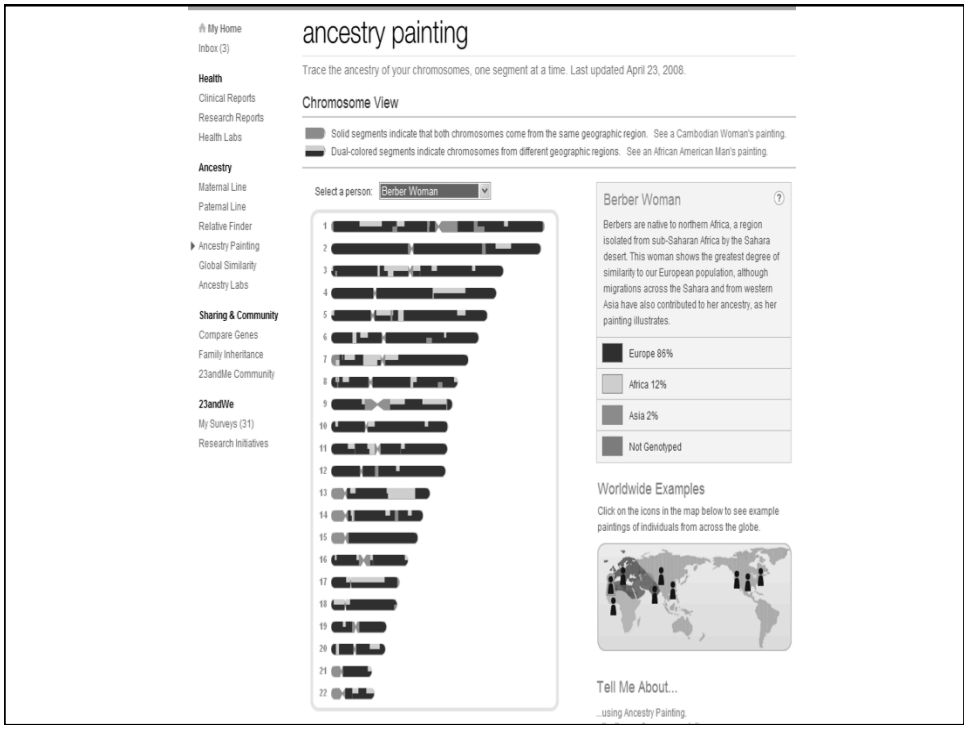


"African-American"




My Home
Inbox (3)
Health
Clinical Reports
Research Reports
Health Labs
Ancestry
Maternal Line
Paternal Line
Relative Finder
Ancestry Painting
Global Similarity
Ancestry Labs
Sharing & Community
Compare Genes
Family Inheritance
23andMe Community
23andWe
My Surveys (31)
Research Initiatives
maternal line
Your mitochondrial DNA determines your maternal haplogroup. What is a haplogroup?
tell a friend
Map
History
Haplogroup Tree
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.
Asia
North America
Atlantic Ocean
Pacific Ocean
Africa
Indian Ocean
Australia
South America
0%
50%
100%
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.
Human Prehistory Videos
Human Prehistory: Prologue
Out of (Eastern) Africa
Recent Posts in Maternal Line
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
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
Tell Me About...
...mitochondrial DNA (mtDNA).
...maternal haplogroups.


My Home
Inbox (3)
Health
Clinical Reports
Research Reports
Health Labs
Ancestry
Maternal Line
Paternal Line
Relative Finder
Ancestry Painting
Global Similarity
Ancestry Labs
Sharing & Community
Compare Genes
Family Inheritance
23andMe Community
23andWe
My Surveys (31)
Research Initiatives
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
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
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.


My Home
Inbox (3)
Health
Clinical Reports
Research Reports
Health Labs
Ancestry
Maternal Line
Paternal Line
Relative Finder
Ancestry Painting
Global Similarity
Ancestry Labs
Sharing & Community
Compare Genes
Family Inheritance
23andMe Community
23andWe
My Surveys (31)
Research Initiatives
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: Berber Woman
Berber Woman
Berbers are native to northern Africa, a region isolated from sub-Saharan Africa by the Sahara desert. This woman shows the greatest degree of similarity to our European population, although migrations across the Sahara and from western Asia have also contributed to her ancestry, as her painting illustrates.
Europe 86%
Africa 12%
Asia 2%
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.

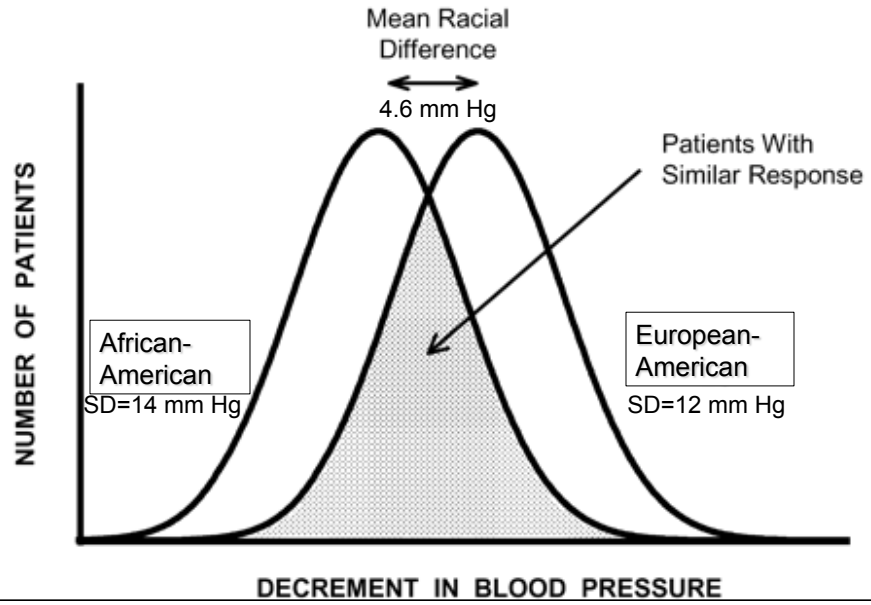

Maternal Line
Paternal Line
Relative Finder
Ancestry Painting
Global Similarity
Ancestry Labs
Sharing & Community
Compare Genes
Family Inheritance
23andMe Community
23andWe
My Surveys (31)
Research Initiatives
Select a person: African American Man
African American Man
Most African Americans today trace 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

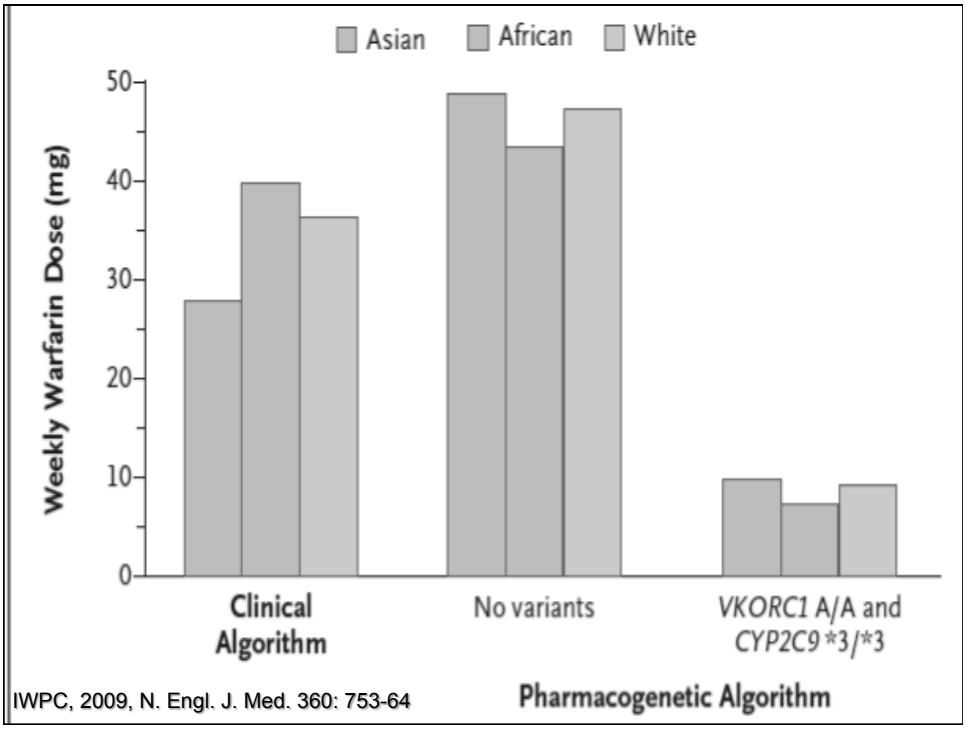
(Sehgal, 2004, *Hypertension* 43: 566-72)



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

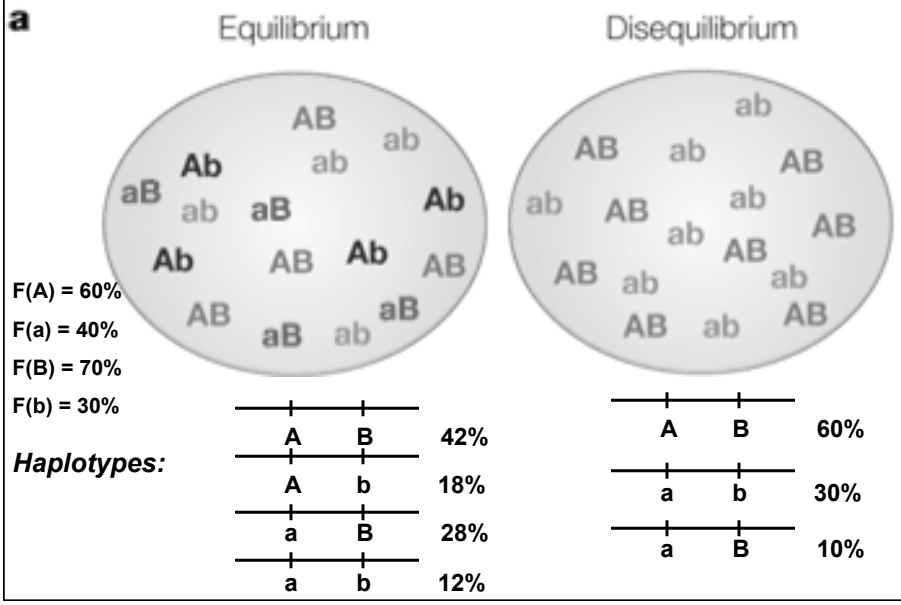


IWPC, 2009, N. Engl. J. Med. 360: 753-64

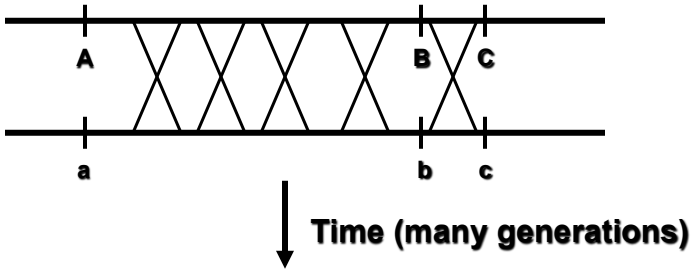
Genetic Variation and "Race"

- 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: nonrandom association of alleles at linked loci



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

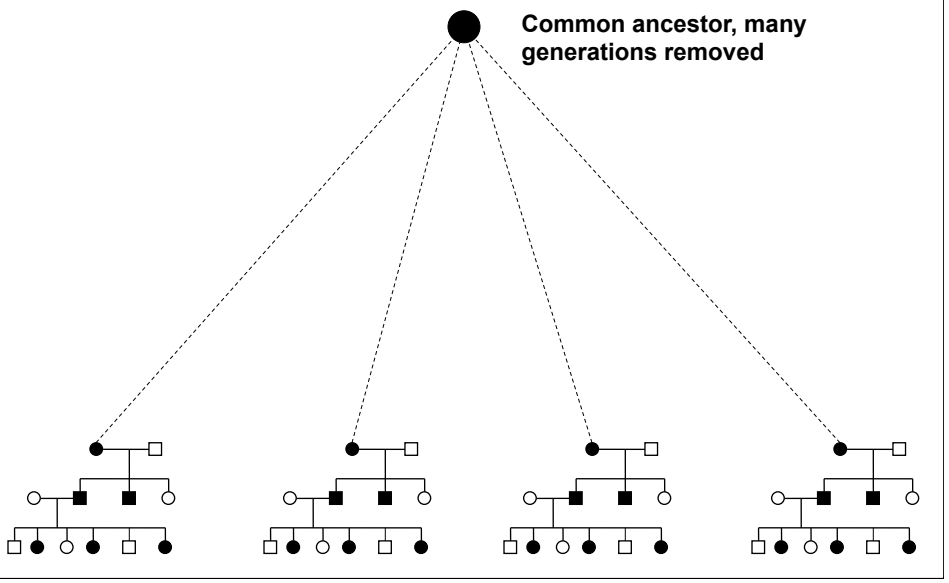


B and C will be found together on the same haplotype more often than A and B: there is more *linkage disequilibrium* between B and C than A and B

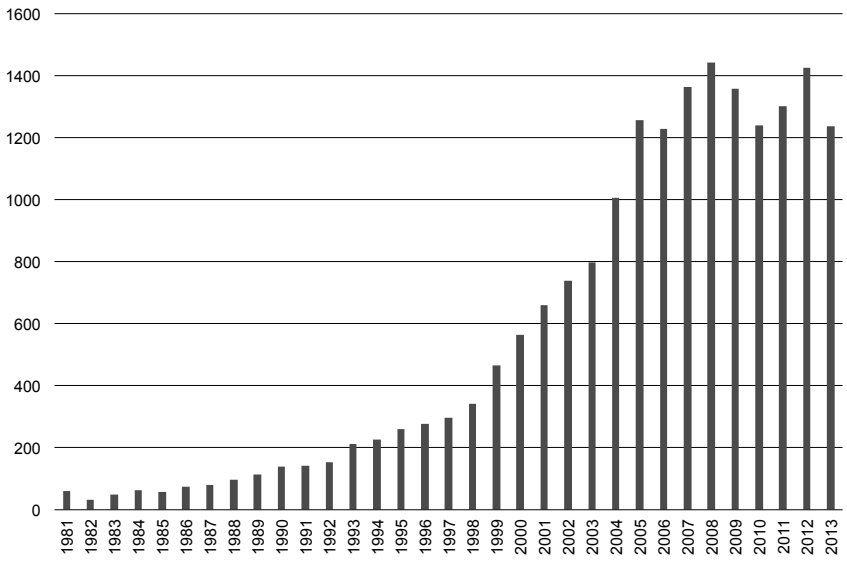
Potential advantages of linkage disequilibrium (LD)

- Family data are *not* necessarily needed
- Microarray technology now exists that allows dense genotype assays (SNVs every 1-3 kb)
- Association studies (linkage disequilibrium) can incorporate many past generations of recombination to narrow the candidate region

Populations are one big (complicated) pedigree



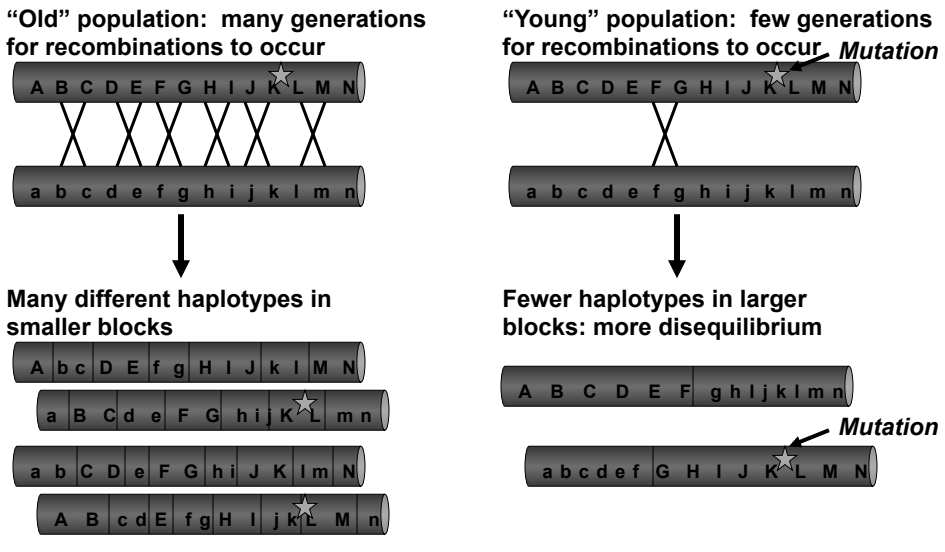
Number of published LD articles per year



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

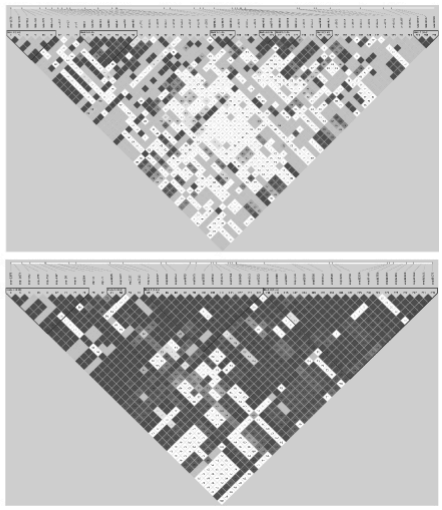
Population "age" can affect haplotype structure



Pairwise LD at the Angiotensinogen locus

Africans

Eurasians



Haploview; red = high LD

How general are these patterns?

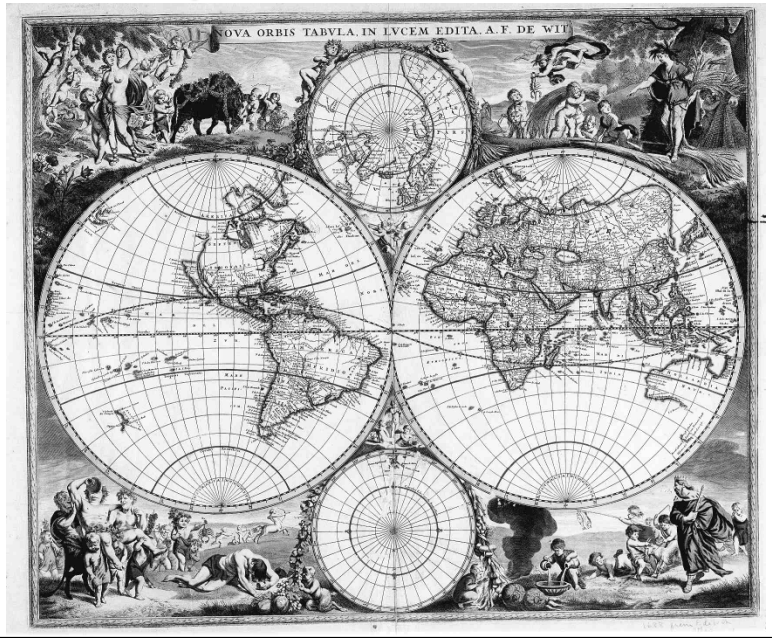
To what extent does LD vary with genomic location and population?

A Map of the World, 1544

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

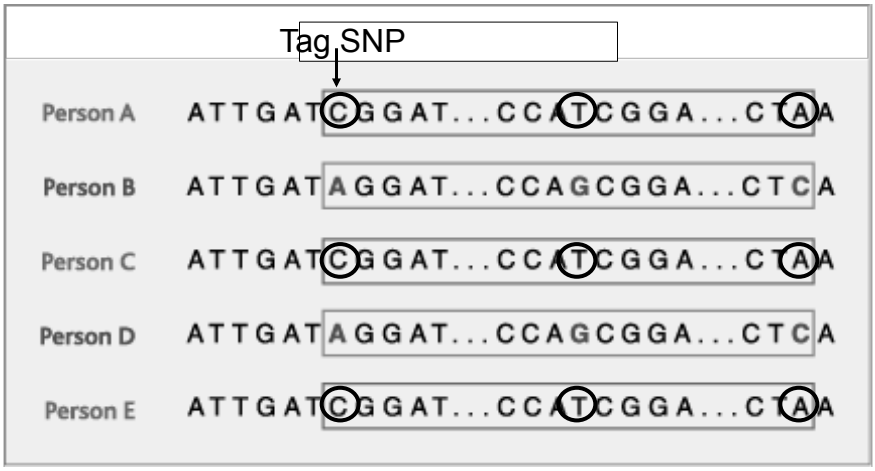
A Map of the World, 1688



Genetic applications of HapMap

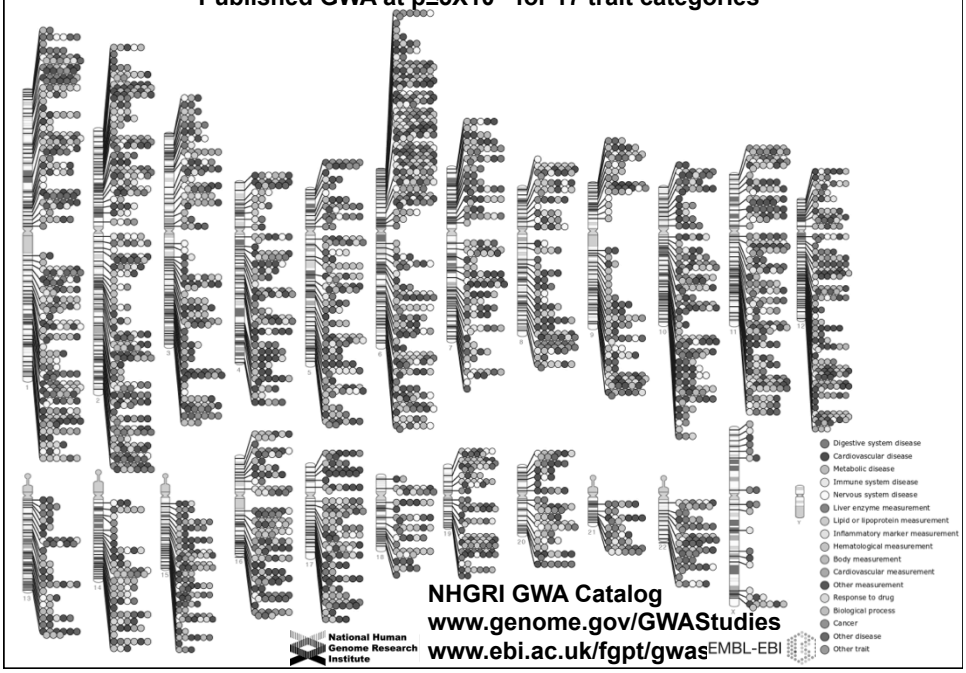
- Understanding human genome-wide haplotype diversity
- Detection of recombination hotspots
- Detection of genes that have experienced strong natural selection
- Detection of disease-causing mutations

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

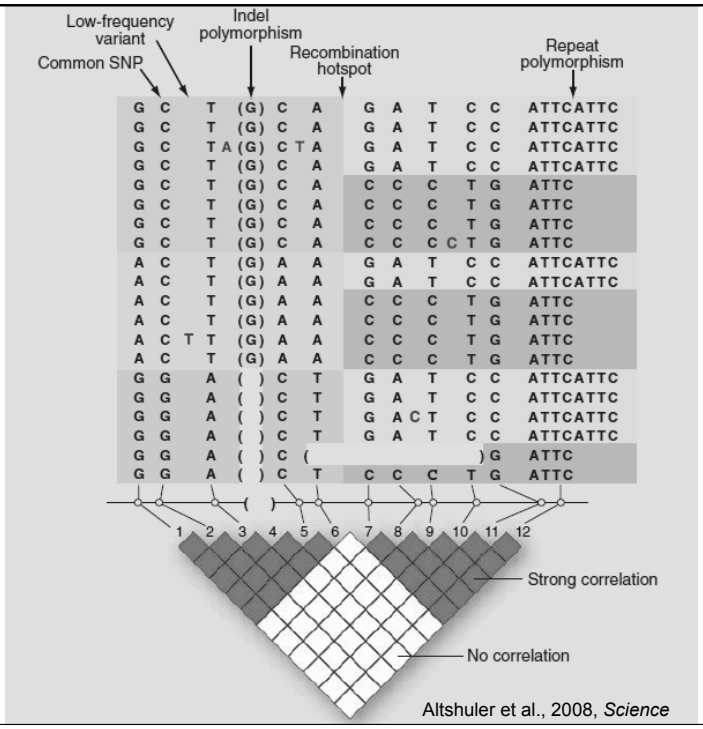


For whole-genome association studies, "complete" coverage is given by about 1.6 million SNPs for African populations, 600,000 to 1M SNPs for non-African populations

Published Genome-Wide Associations through 12/2012
Published GWA at p≤5X10^{-8} for 17 trait categories

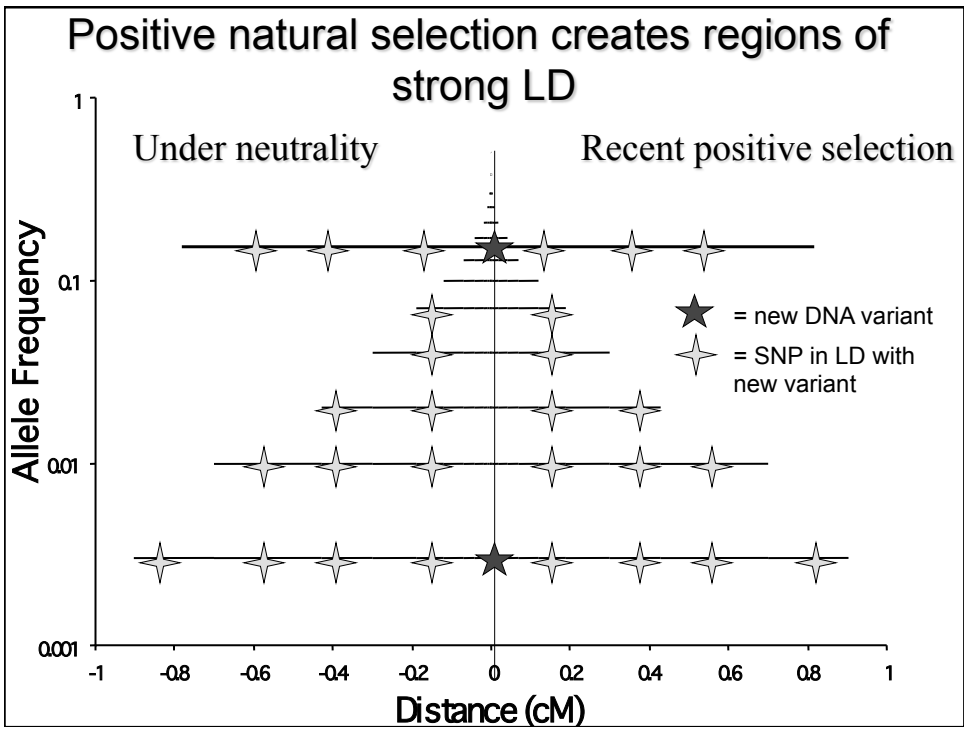


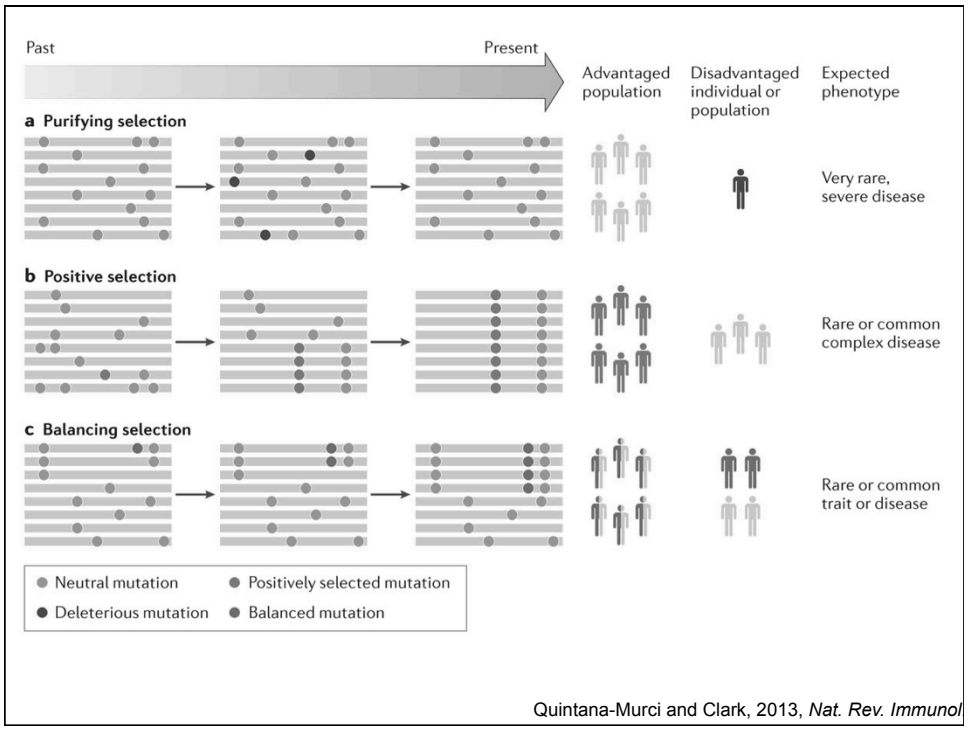
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas



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





Quintana-Murci and Clark, 2013, *Nat. Rev. Immunol*

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*

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*!