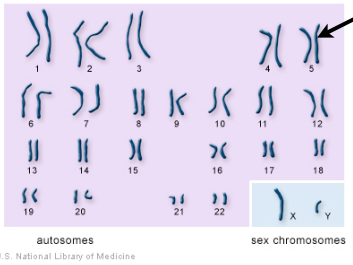
HC70A & SAS70A Winter 2010
Genetic Engineering in Medicine, Agriculture, and Law

Tracking Human Ancestry

Professor John Novembre

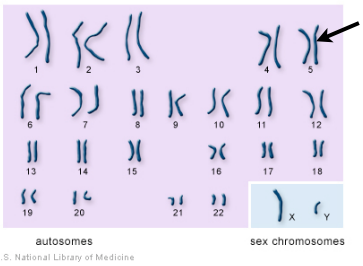
A question



Consider a single portion of your genome.

At that locus, are you homozygous or heterozygous?

Intuitive answers...



Why might you be homozygous?

1) Your mother and father both gave you identical copies they received from a very recent common ancestor (identity-by-descent - inbreeding)

2) There is only one version of the DNA that is viable (purifying selection)

Why might you be heterozygous?

1) Mutation rates at that locus are relatively high

2) Your parents are very distantly related (migration or large population size)

3) Being heterozygous for that locus is common because it is selectively favored (heterozygote advantage)

These are just some of the possible answers...

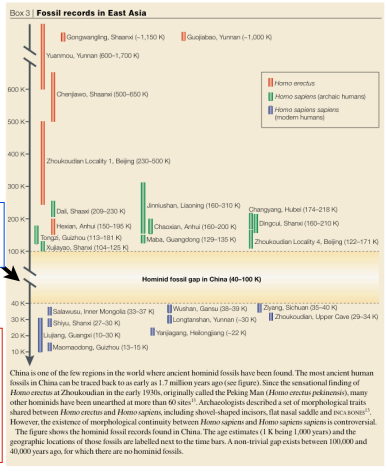
Themes

- Global patterns of human genetic diversity
  - Tracing our ancient ancestry
  - Clines versus clusters debate
- Within-continent patterns
- Personalized genomic ancestry inference
  - What really is ancestry?
  - Admixture and Chromosome painting
- Natural selection and patterns of human genetic diversity
  - Salivary Amylase
  - Eye color (OCA2)

Population genetics

- When and why should we expect to see diversity among copies of DNA sampled from a single population?
- What can we learn about a population from observing patterns of genetic diversity?
- Fundamental processes:
  - Genetic drift / inbreeding
  - Mutation
  - Natural selection
  - Migration
  - Recombination
- A field that integrates multiple scales of biology: molecular-level processes (mutation and recombination) and population-level (migration, genetic drift, natural selection) processes

A puzzle of human ancestry



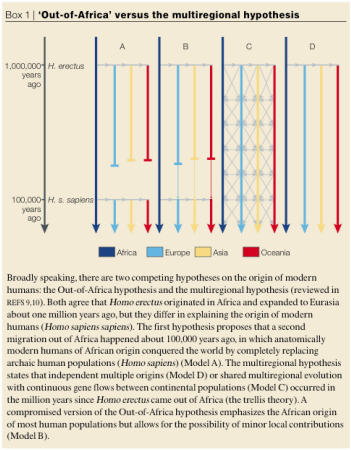
Homo erectus

The fossil record in East Asia contains a 60,000 year gap

Did archaic humans outside-of-Africa die out, only to be replaced later on or is this just an incomplete fossil record?

Jin and Su, Nature Reviews Genetics (2000)

Competing models of human origins



Which of these is the best supported hypothesis for origins of humans around the globe?

(A) Recent Out-of-Africa origin

(B) Recent Out-of-Africa origin with minor contribution from ancestral archaic humans

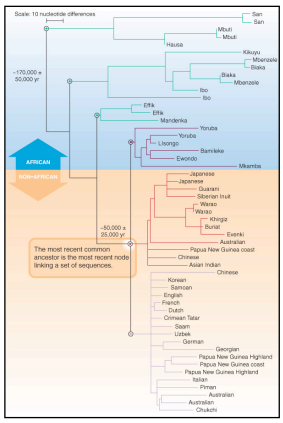
(C) Shared multi-regional evolution

(D) Independent multiple origins

Jin and Su, Nature Reviews Genetics (2000)

All human mitochondrial DNA sequences have a common ancestor "Eve" 120-220k years ago

Non-African sequences have a common ancestor at 25-75k years ago



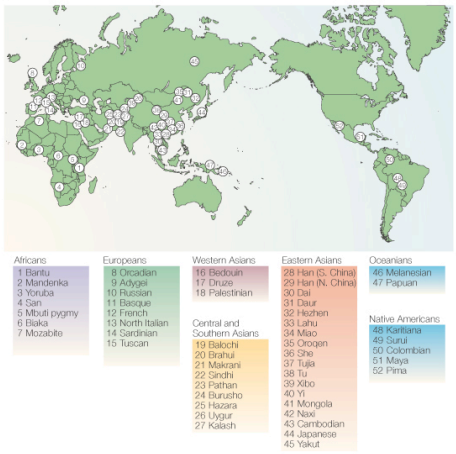
Tree shape is consistent with Out-of-Africa origin

Mitochondrial "Eve" existed in the recent past!

Human genome diversity panel

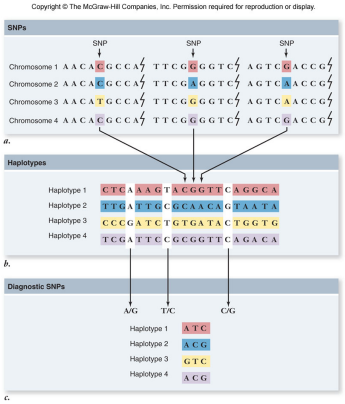
A global-scale sample of human genetic diversity

Not a perfect sampling, but the best studied yet...

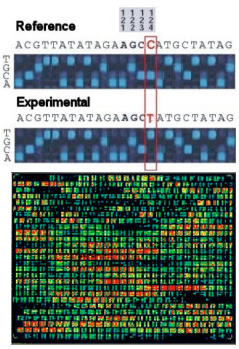


Nature Reviews | Genetics

Review: Single nucleotide polymorphisms (SNPs) and haplotypes



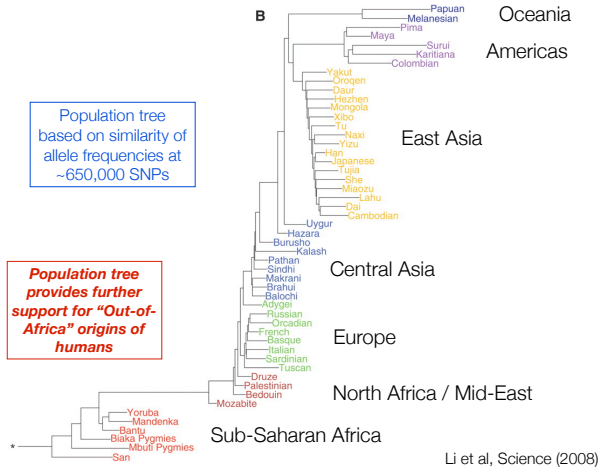
DNA Chips Can Detect SNP Genotypes (Or Haplotypes) Across An Individual's Genome



This Can Then Be Correlated With Diseases &/or Geographical Associations

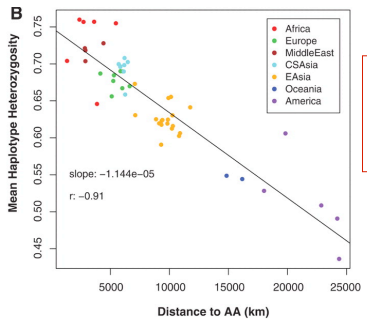
Population tree based on similarity of allele frequencies at ~650,000 SNPs

Population tree provides further support for "Out-of-Africa" origins of humans



Li et al, Science (2008)

Global patterns of haplotype diversity

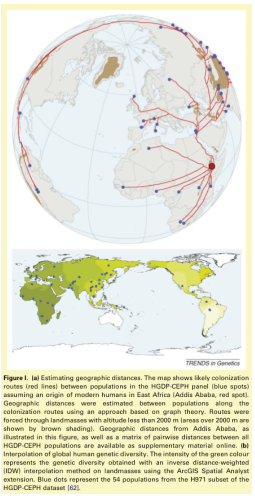


Decay of haplotype heterozygosity is consistent with a *"serial bottlenecks"* during Out-of-Africa expansion

Distance ***via waypoints*** to:
Addis Ababa, Rift Valley, Ethiopia
(Putative origin of early modern humans)

Examples of paths from East Africa to each of the HGDP populations

Interpolated map of genetic diversity



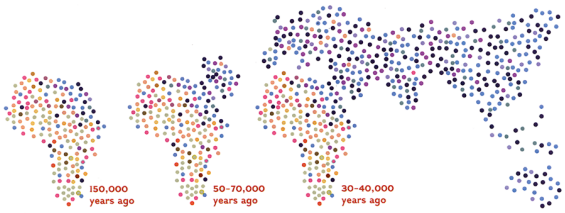
Decay of diversity is consistent with Out-of-Africa expansion

Figure I. (a) Estimating geographic distances. The map shows likely colonization routes (red lines) between populations in the HGDP-CEPH panel (blue spots) assuming an origin of modern humans in East Africa (Addis Ababa, red spot). Geographic distances were estimated between populations along the colonization routes using an approach based on graph theory. Routes were forced through landmasses with altitude less than 2000 m (areas over 2000 m are shown by brown shading). Geographic distances from Addis Ababa, as illustrated in this figure, as well as a matrix of pairwise distances between all HGDP-CEPH populations are available as supplementary material online. (b) Interpolation of global human genetic diversity. The intensity of the green colour represents the genetic diversity obtained with an inverse distance-weighted (IDW) interpolation method on landmasses using the ArcGIS Spatial Analyst extension. Blue dots represent the 54 populations from the H971 subset of the HGDP-CEPH dataset [62].

Most Genetic Diversity Originated in the Founder Populations to Modern Humans!

Diverse From the Start

The diversity of genetic markers is greatest in Africa (multicolored dots in map), indicating it was the earliest home of modern humans. Only a handful of people, carrying a few of the markers, walked out of Africa (center) and, over tens of thousands of years, seeded other lands (right). "The genetic makeup of the rest of the world is a subset of what's in Africa," says Yale geneticist Kenneth Kidd.



Summary: Human origins

- At a global scale, genome-wide diversity patterns broadly consistent with a single, recent origin of modern humans in Africa
  - Further support available from recent Y-chromosome and autosomal gene TMRCA dates
  - Note:
    - A very small number of loci show very ancient TMRCA dates (e.g. 2 million years old).
    - ***Open question:*** Are these evidence for rare, ancient contributions from archaic humans?

Continental-scale clusters of human variation?

Results of an ***unsupervised*** clustering algorithm on microsatellite diversity

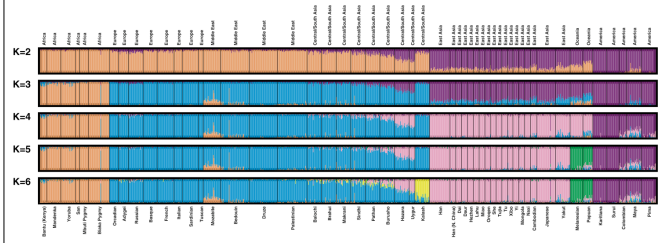


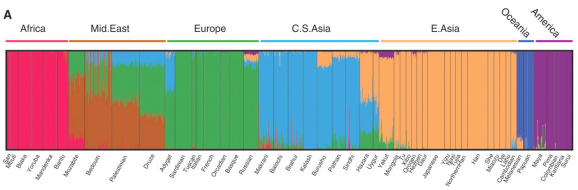
Fig. 1. Estimated population structure. Each individual is represented by a thin vertical line, which is partitioned into K colored segments that represent the individual's estimated membership fractions in K clusters. Black lines separate individuals of different populations. Populations are labeled below the figure, with their regional affiliations above it. Ten *structure* runs at each K produced nearly identical individual membership coefficients, having pairwise similarity coefficients above 0.97, with the exceptions of comparisons involving four runs at $K = 3$ that separated East Asia instead of Eurasia, and one run at $K = 6$ that separated Karitiana instead of Kalash. The figure shown for a given K is based on the highest probability run at that K.

Genetic "clusters" approximate continental-scale regions

Rosenberg et al (2002) Science

Continental-scale clusters of human variation?

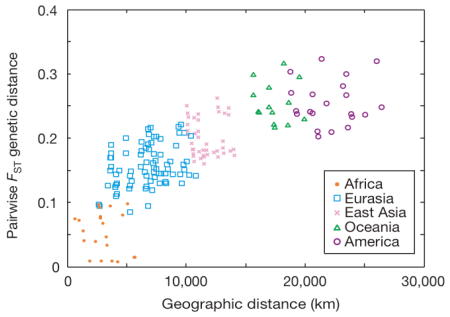
Results of ***unsupervised*** clustering algorithm on ~***650,000 SNPs***



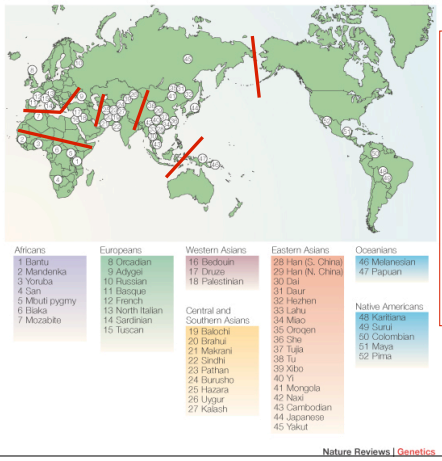
Clusters are now more detailed
Europe, Middle East, and Central South Asia distinguishable

Li et al (2008) Science

Or does differentiation increase smoothly with geographic distance ("clines")?



Clines vs. clusters debate



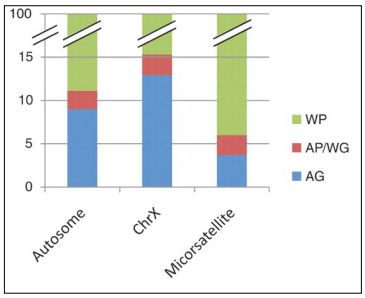
Nature Reviews | Genetics

Perhaps clusters are due to impact of geographic barriers?

Sahara
Tibetan Plateau
Bering Strait
Malay archipelago

Or perhaps clusters are an artifact due to gaps in sampling?

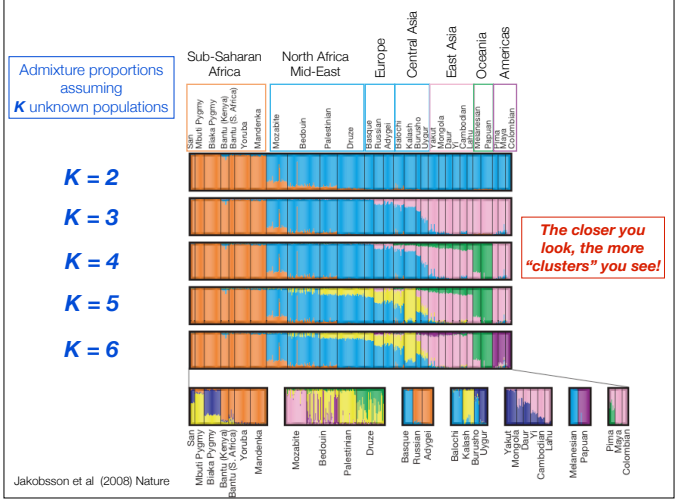
Either way: Variation is mostly found within rather than between groups



Within populations
Among population / within groups
Among groups

Li et al (2008) Science

Admixture proportions assuming ***K*** unknown populations



The closer you look, the more "clusters" you see!

Jakobsson et al (2008) Nature

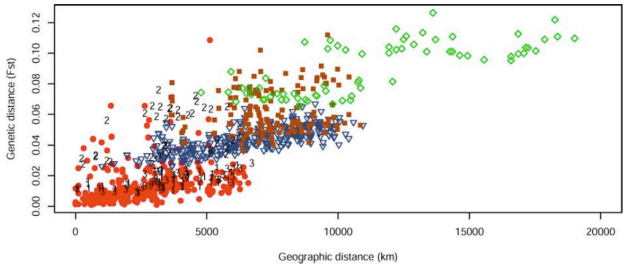


Figure 6. Genetic and Geographic Distance for Pairs of Populations

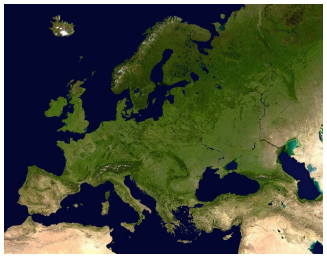
Red circles indicate comparisons between pairs of populations with majority representation in the same cluster in the $K = 5$ plot of Figure 2; blue triangles indicate pairs with one population from Eurasia and one from East Asia; brown squares indicate pairs with one population from Africa and the other from Eurasia; and green diamonds indicate pairs with one population from East Asia and the other from either Oceania or America. Comparisons involving one of Hazara, Kalash, and Uygur and other populations from Eurasia or East Asia are marked 1, 2, and 3, respectively. No comparisons are shown between any of these three groups and any African population.

DOI: 10.1371/journal.pgen.0010070.g006

Patterns of genetic distance as a function of geographic distance reveal cluster and cline patterns

Fine-scale analysis within a continent: Europe as a case study

Image from EU Panorama "27 Europeans"



1400 individuals from 37 unique populations in Europe

Analysis using 197,000 SNP loci

Principal Components Analysis (PCA)=Reduction of dimensions

Single Nucleotide Polymorphisms

Individuals

0 = AA
1 = AB
2 = BB

PC1 PC2

Pop. 1 Pop. 2

Admixed Individuals

PC2 PC1

PCA is often a useful tool to visualize patterns of population structure

Two-dimensional summary of European genetic diversity

PC1 PC2

Subtle differentiation within Switzerland

b

North–south in PC1–PC2 space

East–west in PC1–PC2 space

UK Germany France Spain Portugal Italy

French-speaking Swiss
German-speaking Swiss
Italian-speaking Swiss
French
German
Italian

Genome-wide patterns of variation mirror the geography of Europe!

Differentiation found at small spatial-scale

Novembre et al (2008) *Nature*

Differentiation within Finland

a

Central Lapland
West Lapland
East Lapland
North Oulu
South Oulu
Kainuu

b

Subjects with both parents from the same group

Second coordinate MDS

First coordinate MDS

West Lapland
Central Lapland
East Lapland
South Oulu
North Oulu
Kainuu

Figure 1 Linguistic/geographic groups of Northern Finland and their genetic signature. (**a**) Map of Finland with county boundaries. The subjects in NFBC1966 were all born in the two northern provinces. Counties in Northern Finland are color coded to correspond to the six linguistic/geographical groups that can be identified. (**b**) Scatterplot of the two first components identified by MDS on the matrix of genetic similarity between individuals. Only subjects with both parents born in the same population group are plotted, and they are color coded according to the group of origin.

Per locus information content extremely low

25 random SNPs

Minor allele frequency

Region of Europe

NW W C SW S SE

At any given SNP, there is little variation among populations

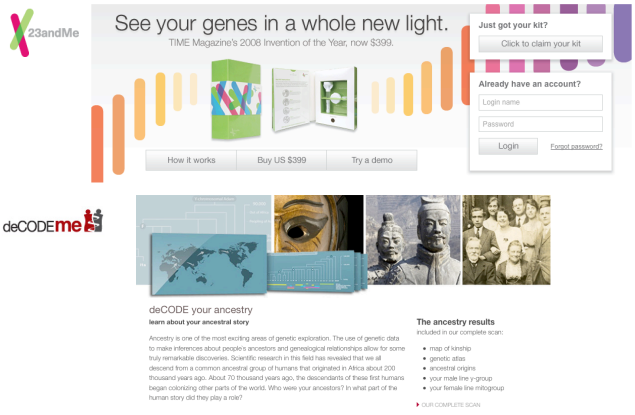
PCA methods pool weak signals across many, many loci to reveal differentiation

- Across loci mean F_{ST}=0.004!

Summary: Clines versus Clusters

- At a global scale, support for clusters and clines can be observed
- With large numbers of SNP markers, patterns of differentiation are detectable even at small-scales within continents (although often more "clinal" than "clustered")
  - Note: Variation is still predominately within vs. between groups
- ***Future directions:*** With whole genomes - we may detect even more subtle patterns of differentiation

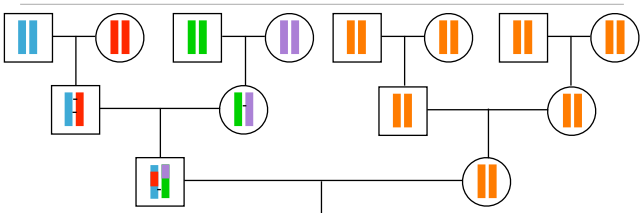
Applications: Ancestry Inference in Personalized Genomics



What do we mean by ancestry?

- My mtDNA test comes back and says I have a Native American mitochondrial haplotype. Does this mean:
  - (a) I am is likely to be 100% Native American
  - (b) My mother's side is likely to be 100% Native American
  - (c) My mother's mother's mother's mother's mother's mother's mother was likely Native American.
  - (d) Not enough information to tell.

What do we mean by ancestry?

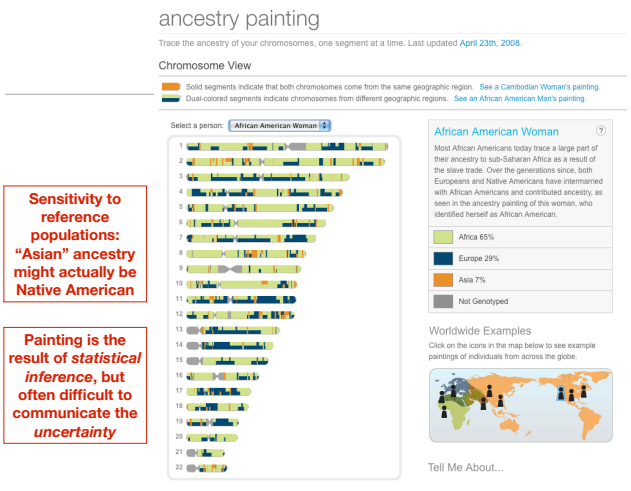


If we have "mixed" ancestry, our chromosomes are a patchwork of DNA with different ancestries

But by chance, some ancestors are "lost" and leave no trace in our genomes at all (Example: Purple ancestor)

"Ancestry" is relative to time: implicit interest in the geographic origins of our *relatively recent* ancestors

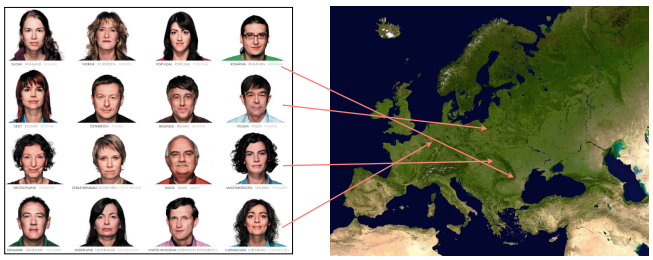
mtDNA and Y-chromsome: Trace only one of your ancestors at any given point back in time



Sensitivity to reference populations: "Asian" ancestry might actually be Native American

Painting is the result of *statistical inference*, but often difficult to communicate the *uncertainty*

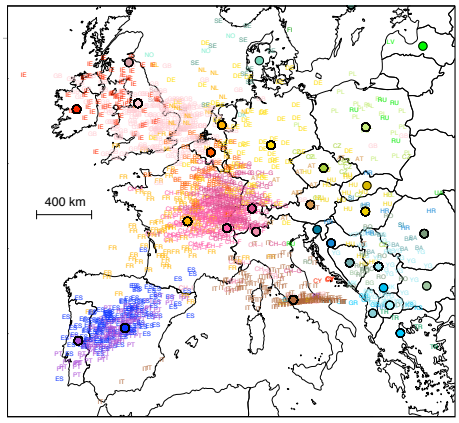
Pushing the limits: Ancestry inference within continents



Performance: Spatial prediction

- Regression-based approach using PC scores
- "Leave-one-out" cross-validation performance:
- For countries with n>15:
  50%: < 240 km
  90%: < 630 km
- Overall:
  50% : < 540 km
  90% : < 840 km

Caveat: With PCA, mixed ancestry leads to an intermediate positioning



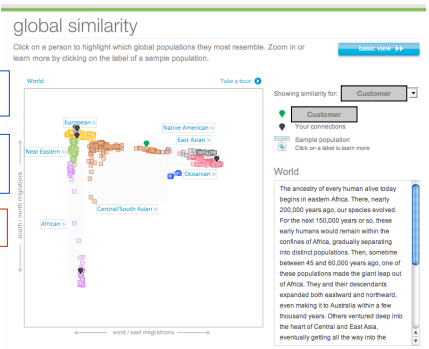
PCA and mixed ancestry

PCA of HGDP populations

Customer with mixed East Asian / European ancestry show in green.

Customer appears to be Central Asian!

Note: 23andme FAQ warns customer of this issue



Summary: Personal ancestry inference

- Immense potential, but numerous challenges:
  - Obtaining the appropriate reference samples
  - Communicating to customer:
    - What we really mean by ancestry
    - Statistical uncertainty in the inferences
    - Limitations of particular analyses
  - Danger of customers forming notions of group membership / race?

What's your opinion:

- Personal ancestry inference is:
  - (a) A waste of money.
  - (b) A harmless hobby if that's what you're in to.
  - (c) Fascinating - sign me up!
  - (d) The first steps towards genetic elitism.

Question:

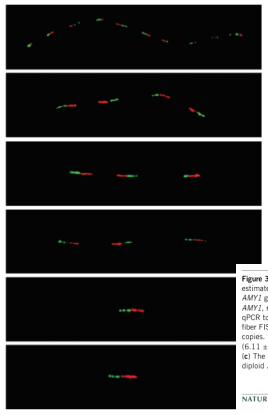
- Have cultural changes had an effect on human evolution?
  - (a) Yes
  - (b) No

Salivary amylase copy number variation

Japanese individual
10 + 4 = 14 copies

Biaka individual
3 + 3 = 6 copies

1 + 1 = 2 copies
Chimpanzee



AMY1 copy number variation as seen by fluorescent in-situ hybridization (FISH)

Copy-number variation is highly variable among human populations

Figure 3 High-resolution fiber FISH validation of *AMY1* copy number estimates. Red (~10 kb) and green (~8 kb) probes encompass the entire *AMY1* gene and a retrotransposon directly upstream of (and unique to) *AMY1*, respectively. (**a**) Japanese individual GM18972 was estimated by qPCR to have 14 (13.73 ± 0.93) diploid *AMY1* gene copies, consistent with fiber FISH results showing one allele with ten copies and the other with four copies. (**b**) Biaka individual GM10472 was estimated by qPCR to have six (6.11 ± 0.17) diploid *AMY1* gene copies, consistent with fiber FISH results. (**c**) The reference chimpanzee (Clint; SO06006) was confirmed to have two diploid *AMY1* gene copies.

NATURE GENETICS VOLUME 39 | NUMBER 10 | OCTOBER 2007

Diet and Variation in salivary amylase copy number

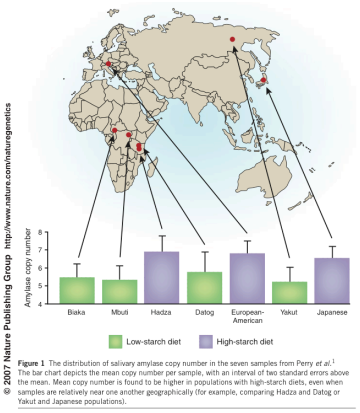
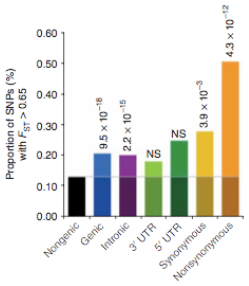


Figure 1 The distribution of salivary amylase copy number in the seven samples from Perry *et al.*[1] The bar chart depicts the mean copy number per sample, with an interval of two standard errors above the mean. Mean copy number is found to be higher in populations with high-starch diets, even when samples are relatively near one another geographically (for example, comparing Hadza and Datog or Yakut and Japanese populations).

High-starch diet populations have higher average copy number for than low-starch diet populations

The impact of natural selection on population differences

SNPs in genic regions and especially non-synonymous SNPs are more often extremely differentiated than "non-genic" SNPs



Positive selection in one population but not another can lead to differentiation at SNPs in the region of the gene under selection

The most differentiated SNPs between human groups are likely to be in regions that have undergone recent adaptive evolution

Bariero et al (2008) Nature Genetics

Categories of genes that appear to have been under selection *judging by extreme population differentiation*

Table 1 Genes showing the strongest signatures of positive selection

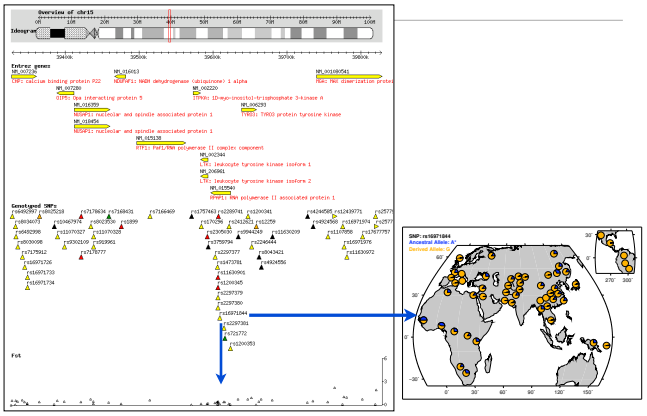
Phenotype category	Genes
Morphological traits (for example, skin pigmentation and hair development)	*ABCC11*, ***EDAR***, *SLC45A2*, *PKP1*, *PLEKHA4*, ***SLC24A5***
Immune response to pathogens	*CEACAM1*, *CR1*, ***DUOX2***, *VAV2*
DNA repair and replication	*MPG*, *POLG2*, *TDP1*
Sensory functions (for example, olfaction and eye development)	*COL18A1*, *OR52K2*, *RP1L1*
Insulin regulation, metabolic syndrome (obesity, diabetes, hypertension)	*ALMS1*, *CEACAM1*, *ENPP1*
Various metabolic pathways (for example, ethanol, intestinal zinc and citrulline)	***ADH1B***, *ASS1*, *SLC39A4*
Miscellaneous	*FBXO31*, *RTTN*, ***SPAG6***
Unknown	*ABCC12*, ***ADAT1***, *AK127117*[a], C17orf46, C8orf14, COLEC11, ***CPSF3L***, *DNAJC5B*, *DNHD1*, *ETFDH*, *EXOC5*, ***FAIM***, ***CCDC142***[b], *FLJ37464*[a], *FXR1*, *GCN5L2*, *KIAA0984*[a], *LAMB4*, *LOC648511*[a], ***LIMCH1***, ***PCGF1***[b], *PLEKHG4*, *POL3S*[b,c], ***RNF135***, ***SLC30A9***, *SYTL3*, *TEX15*, ***TTC31***[b], *VPS33B*, *ZNF646*[c]

These genes contain at least one nonsynonymous or 5′-UTR mutation with $F_{ST} > 0.65$. An exhaustive list of 582 genes containing other classes of genic SNPs with $F_{ST} > 0.65$ is provided in **Supplementary Table 1**. Genes in bold correspond to those also presenting significant long-range haplotypes, as measured by the iHS statistic[4], or defined as top candidates for recent selective sweeps[3].
[a]These genes have not yet been attributed a HUGO-approved symbol. [b]These three genes are located in a linkage-disequilibrium block in chromosome 2. [c]These two genes are located in a linkage-disequilibrium block in chromosome 16.

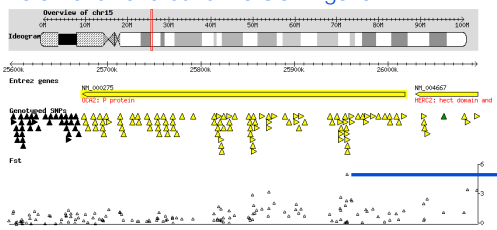
Bariero et al (2008) Nature Genetics

What we see as surface indicators of "race" (skin, hair, eye-color) are some of the most unique regions of the genome

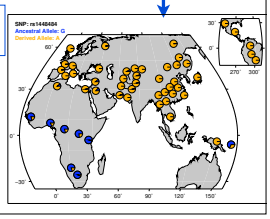
Differentiation at a random region of Chromosome 15



Differentiation at around the OCA2 gene



Differentiation at SNP rs1448484 in the OCA2 gene



OCA2 came up in Lecture 5...

- Was it a gene that is known to affect:
  - (a) Type II diabetes susceptibility
  - (b) Eye-color
  - (c) Intelligence
  - (d) Cystic-fibrosis

SNPs in Human P Protein Gene Lead To Different Eye Colors

OCA2

From SNPedia

OCA2, the oculocutaneous albinism gene (also known as the human P protein gene, or, DN10), is a gene associated with albinism and certain pigmentation effects in general such as eye color, skin color, and hair color.

A large (>3,000 individuals) study of Caucasians indicates that the following **OCA2** variants, all located in the first intron of the gene, are preferentially linked to blue eye color inheritance; together, they form haplotypes that (in some cases at least) predict eye color with greater than 50:50 odds. [PMID 17236130; OMIM 203200.0013 (http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=203200&a=203200_AllelicVariant0013)]

- rs7495174
- rs6497268
- rs11855019

The haplotypes are defined in order as listed above for these 3 SNPs, so, for example, the TGT haplotype refers to rs745174(T)-rs6497268(G)-rs11855019(T). The correspondence between diplotypes (the two haplotypes in one individual) and the % of individuals with blue/gray, green/hazel/ and brown eye color , respectively, was reported as follows for the most common diplotypes[PMID 17236130]:

- TGT/TGT: 62.5, 28.0, 9.5
- TGT/***TTC***: 47.1, 20.3, 32.6
- TGT/***CGT***: 28.6, 14.3, 57.1
- TGT/***TGC***: 27.9, 22.1, 50.0
- ***TGC***/***TTC***: 25.0, 8.3, 66.7
- ***TTT***/***TGC***: 20.7, 31.0, 48.3
- TGT/***TTT***: 17.6, 38.5, 44.0
- TGT/***CTC***: 7.9, 23.3, 68.8

The haplotypes shown in ***bold italics*** represent the ones reported by the authors of this study to be most associated with brown eye color. Furthermore, the haplotypes shown above are as published, and the associated SNPs - which have since changed # as well - are not in the orientation shown in dbSNP.

More recently, a study of a large Danish family led to associations with 2 SNPs in a different region of **OCA2** as linked to blue or brown eye color:

- rs12913832
- rs1129038

Earlier studies found different regions of the **OCA2** gene to also be predictive of eye color;

- **OCA2** SNP rs1800401 helps predict brown eye color. [PMID 12163334, PMID 15889046; OMIM 203200.0011 (http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=203200&a=203200_AllelicVariant0011)]
- **OCA2** SNP rs1800407 may be associated with green/hazel eye color in some populations, but not others. [PMID 12163334, PMID 15889046; OMIM 203200.0012 (http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=203200&a=203200_AllelicVariant0012)]

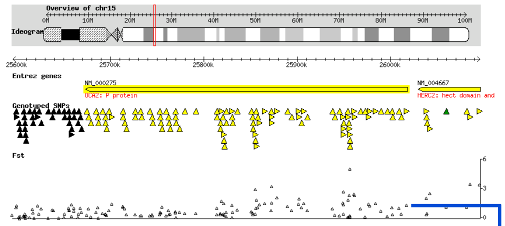
is a	gene
is	mentioned by
wikipedia	OCA2 (http://en.wikipedia.org/wiki/OCA2)
google	OCA2 (http://www.google.com/search?hl=en&q=OCA2)
gopubmed	OCA2 (http://www.gopubmed.org/search?q=OCA2)
23andMe	OCA2 (https://www.23andme.com/you/explorer/gene/?gene_name=OCA2)
GeneRIF	4948 (http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=4948&ordinalpos=1&itool=EntrezSystem2.PEntrez.Gene.Gene_ResultsPanel.Gene_F
dbSNP	4948 (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=4948&chooseRs=all)
PubMed	4948 (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=Link&LinkName=gene_pubmed&from_uid=4948)
HugeNav	4948 (http://hugenavigator.net/HuGENavigator/huGEPedia.do?firstQuery=OCA2}&geneID=4948&typeSubmit=GO&check=y&typeOption=gene&which=2&pubOrderType=pubD)

	Chromosome position
Rs1129038	26,030,454
Rs11631797	26,175,874
Rs12593929	26,032,853
Rs1800401	25,933,648
Rs1800407	25,903,913
Rs2238289	26,126,810
Rs2240203	26,167,797
Rs28934272	25,903,842
Rs3935591	26,047,607
Rs3940272	26,142,318
Rs4778241	26,012,308
Rs7170852	26,101,581
Rs7183877	26,039,328
Rs7495174	26,017,833
Rs8028689	26,162,483
Rs916977	26,186,959

Human Eye Color

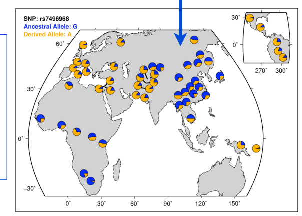
NOTE: OCA2 also affects skin color and hair color

Differentiation at around the OCA2 gene



Differentiation at SNP rs7496968 (known to affect eye-color) in the OCA2 gene

G allele carried on haplotype known to predispose individuals to brown eyes



Conclusions:

- Humans have been recently, and continue to be, evolving
- Patterns of genetic variation in humans point towards recent common ancestry in Africa
- With modern large-scale data sets, we can identify the personal ancestry of individuals to a fine spatial scale
- Many of the most differentiated regions of the genome seem to be the results of selection related to:
    - Novel diets
    - External morphology in response to different climates
    - Immune system / disease evasion
- Beyond these few differentiated regions (some of which are relevant to medicine), most variation is found globally (and most of the genome doesn't even vary!)
- Intelligent views about our common genetic heritage and diversity will be crucial in our post-genomic world