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

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

Mutation and Genetic Variation

Human mutation rate is $1.0 - 1.5 \times 10^{-8}$ 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

<table>
<thead>
<tr>
<th>Species</th>
<th>SNVs per Individual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orangutans</td>
<td>9.3 million</td>
</tr>
<tr>
<td>Gorillas</td>
<td>6.5 million</td>
</tr>
<tr>
<td>Chimpanzees</td>
<td>5.7 million</td>
</tr>
<tr>
<td>Humans</td>
<td>3-4 million</td>
</tr>
</tbody>
</table>

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...

...but random duplications and deletions of large segments of DNA mean the number of copies of many genes varies.

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?
Allele frequencies in populations

<table>
<thead>
<tr>
<th>Population</th>
<th>SNV 1</th>
<th>SNV 2</th>
<th>SNV 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.588</td>
<td>0.890</td>
<td>0.880</td>
</tr>
<tr>
<td>2</td>
<td>0.671</td>
<td>0.559</td>
<td>0.528</td>
</tr>
<tr>
<td>3</td>
<td>0.792</td>
<td>0.790</td>
<td>0.828</td>
</tr>
</tbody>
</table>

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?

\[ F_{ST} = \frac{H_T - \bar{H}_S}{H_T} \]

- \( F_{ST} \) is the amount of genetic variation that is due to population differences
- \( H_T \) is the total heterozygosity (variation) in the sample
- \( \bar{H}_S \) 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?

<table>
<thead>
<tr>
<th></th>
<th>60 STRs</th>
<th>100 Alus</th>
<th>75 L1s</th>
<th>250K SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between individuals, within continents</td>
<td>90%</td>
<td>86%</td>
<td>88%</td>
<td>88%</td>
</tr>
<tr>
<td>Between continents ($F_{ST}$)</td>
<td>10%</td>
<td>14%</td>
<td>12%</td>
<td>12%</td>
</tr>
</tbody>
</table>

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

---

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

<table>
<thead>
<tr>
<th>Minor allele present in:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>All 4 groups</td>
<td>78.6%</td>
</tr>
<tr>
<td>At least 3 groups</td>
<td>88.0%</td>
</tr>
<tr>
<td>At least 2 groups</td>
<td>92.1%</td>
</tr>
<tr>
<td>Africa only</td>
<td>7.4%</td>
</tr>
<tr>
<td>Any non-African group</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

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

Average allele frequency difference between populations: 15%

<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)

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 \( j \).

<table>
<thead>
<tr>
<th>Pop.</th>
<th>SNV 1</th>
<th>SNV 2</th>
<th>SNV 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.588</td>
<td>0.890</td>
<td>0.880</td>
</tr>
<tr>
<td>2</td>
<td>0.671</td>
<td>0.559</td>
<td>0.528</td>
</tr>
<tr>
<td>3</td>
<td>0.792</td>
<td>0.790</td>
<td>0.828</td>
</tr>
</tbody>
</table>

\[ D_{12} = |0.588 - 0.671| = 0.083 \text{ (avg. over all SNVs)} \]
Building a population network

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

Percent agreement between Supreme Court justices *(New York Times, 2014)* – analogous to % alleles shared among individuals
Neighbor-joining network of Supreme Court justices’ decisions

Principal components analysis (PCA): a multidimensional regression technique

Thanks to: Steve Guthery, MD
Principal components analysis (PCA): a multidimensional regression technique
Genetic similarity between two people can be completely described with a line

% alleles shared = 90

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

% alleles shared = 90
% alleles shared = 92
% alleles shared = 97
Principal components analysis of Supreme Court decision-making agreement

Population relationships based on 100 autosomal Alu polymorphisms

Thanks to: Julie Feusier, PhD-to-be

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

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)

Genetic distance analysis: 15 loci

J Novembre et al. 2008 Nature

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)

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 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 high-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], 3.4 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

<table>
<thead>
<tr>
<th></th>
<th>AFR</th>
<th>AMR</th>
<th>EUR</th>
<th>SAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td>681</td>
<td>82</td>
<td>347</td>
<td>7.6</td>
</tr>
<tr>
<td>Mean coverage</td>
<td>6.9</td>
<td>7.3</td>
<td>7.7</td>
<td>7.4</td>
</tr>
<tr>
<td>Var sites</td>
<td>4.31M</td>
<td>4.15k</td>
<td>3.64M</td>
<td>12.0k</td>
</tr>
<tr>
<td>Singletons</td>
<td>3.95M</td>
<td>3.68k</td>
<td>3.45M</td>
<td>11.4k</td>
</tr>
<tr>
<td>Large deletions</td>
<td>0.54k</td>
<td>0.46k</td>
<td>0.46k</td>
<td>0.46k</td>
</tr>
<tr>
<td>CNVs</td>
<td>0.14k</td>
<td>0.15k</td>
<td>0.14k</td>
<td>0.14k</td>
</tr>
<tr>
<td>MIs (L)</td>
<td>0.03k</td>
<td>0.04k</td>
<td>0.03k</td>
<td>0.03k</td>
</tr>
<tr>
<td>MIs (GV)</td>
<td>0.02k</td>
<td>0.02k</td>
<td>0.02k</td>
<td>0.02k</td>
</tr>
<tr>
<td>Inversions</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>Non-syn</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>Synon</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>Intron</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>UTR</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>Promotor</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>Insulator</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>Enhancer</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>TFBSs</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>Filled Lof</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>HGVDR DM</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>GWAS</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
<tr>
<td>ClinVar</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
<td>0.01k</td>
</tr>
</tbody>
</table>

See Supplementary Table S1 for continental population groups. CNVs, copy number variants; HGVDR DM, Human Genome Variation Database disease mutations; Lof, loss of function; M, millions; ML, mosaic element insertions.

Variation in individuals: 1000 Genomes Project

Auton et al., 2015, Nature
A “typical” human genome

<table>
<thead>
<tr>
<th>Functional class</th>
<th>Average number/genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein truncating</td>
<td>149 - 182</td>
</tr>
<tr>
<td>Peptide altering</td>
<td>10,000 - 12,000</td>
</tr>
<tr>
<td>Regulatory (UTR, TBS, promoter, etc.)</td>
<td>459,000 - 565,000</td>
</tr>
<tr>
<td>Associated with complex trait</td>
<td>~2,000</td>
</tr>
<tr>
<td>ClinVar disease causing</td>
<td>24 - 30</td>
</tr>
</tbody>
</table>

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

Sudmant et al., 2015, Science
Copy number variation in SGDP samples

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

Sudmant et al., 2015, Science

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

- Three founder mutations in \textit{BRCA1} or \textit{BRCA2} are seen in 2.5% of Ashkenazi Jews (1/200 in general population)

- \textit{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”  

“I am a racially profiling doctor” 

Bamshad and Olson, 2003

**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

<table>
<thead>
<tr>
<th>Asia</th>
<th>Europe</th>
<th>Africa</th>
</tr>
</thead>
</table>

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*
Height

Height + waist/hip ratio

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

- African Americans (ASW)
- CEPH (CEU)
- Chinese (CHB)
- Finns (FIN)
- Japanese (JPT)
- Kenyans (LWK)
- Nigerians (ESN)
- Nigerians (YRI)
- Tuscans (TSI)
- Vietnamese (KHV)
The Fallacy of Typological Thinking

Ancestry vs. Race

“African-American”

“African-American”
What do these findings imply for biomedicine?

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

- 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)

- Mean Racial Difference
- 4.6 mm Hg

Patients With Similar Response

- African-American
  - SD=14 mm Hg
- European-American
  - SD=12 mm Hg

DECREMENT IN BLOOD PRESSURE

---

**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

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

![Diagram showing crossovers between loci A, B, and C over time](image)

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.

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

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.
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

Under neutrality

Recent positive selection

= new DNA variant

= SNP in LD with new variant

Examples of genes in which elevated LD indicates recent positive selection

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>G6PD</td>
<td>Malaria protection</td>
</tr>
<tr>
<td>CYP3A5</td>
<td>Sodium retention</td>
</tr>
<tr>
<td>LCT (lactase enhancer)</td>
<td>Lactase persistence</td>
</tr>
<tr>
<td>SLC24A5</td>
<td>Skin pigmentation</td>
</tr>
<tr>
<td>EPAS1, EGLN1</td>
<td>High-altitude hypoxia response</td>
</tr>
</tbody>
</table>

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 O₂ 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

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!