Ancestral Diversity

Lynn B. Jorde
Department of Human Genetics
University of Utah School of Medicine
29 June 2012
Overview

• Background on human genetic variation

• How sequencing has changed things
How is genetic variation distributed among continental populations?

<table>
<thead>
<tr>
<th></th>
<th>60 STRs</th>
<th>30 RSPs</th>
<th>100 Alus</th>
<th>75 L1s</th>
<th>250K SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>individuals,</td>
<td>90%</td>
<td>87%</td>
<td>86%</td>
<td>88%</td>
<td>88%</td>
</tr>
<tr>
<td>within</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>continents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between</td>
<td>10%</td>
<td>13%</td>
<td>14%</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>continents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(F&lt;sub&gt;ST&lt;/sub&gt;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

J. Xing et al., 2009, Genome Res.
How is genetic variation distributed among continental populations?

<table>
<thead>
<tr>
<th></th>
<th>60 STRs</th>
<th>30 RSPs</th>
<th>100 Alu</th>
<th>75 L1s</th>
<th>250K SNP</th>
<th>Skin pigmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between individuals, within continents</td>
<td>90%</td>
<td>87%</td>
<td>86%</td>
<td>88%</td>
<td>88%</td>
<td>10%</td>
</tr>
<tr>
<td>Between continents ($F_{ST}$)</td>
<td>10%</td>
<td>13%</td>
<td>14%</td>
<td>12%</td>
<td>12%</td>
<td>90%</td>
</tr>
</tbody>
</table>

J. Xing et al., 2009, Genome Res.
% 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*
40 populations, ~250K SNPs

Xing et al., 2010, Genomics
Population relationships in HGDP sample

525,910 SNPs

396 copy number variants (CNVs)

Human population relationships, based on 29 blood group and protein polymorphisms

(Nei and Roychoudhury, 1993, Mol. Biol. Evol.)
Haplotype diversity declines with distance from Africa

J. Xing et al., 2010, Genomics
Recent African origin of anatomically modern humans

Principal components analysis displays individual genetic similarity in 2D: each dot = 1 individual.
Microarray-based SNPs portray population relationships accurately but are biased

- Microarray SNPs 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
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: Luhyia
YRI: Yoruban
MKK: Maasai

Wilfred Wu, MD
Complete Genomics vs. 34 1000 Genomes sequences (Phase 1)

Average between-platform difference = 348,000 variants

Wilfred Wu, MD
Rare SNPs are much more likely to be population-specific

Common SNPs previously identified in dbSNP (build 129)

New rarer SNPs identified by sequencing

Average allele frequency difference between populations: 15%

Durbin et al., 2010, *Nature* (1000 Genomes Project)
Allele age, $t$, as a function of frequency

$$t = \frac{-4Np}{1-p} \ln(p)$$

$N$, effective population size = 10,000

$p = .05$: age = 150,000 years
$p = .01$: age = 47,000 years
We expect many novel variants with each whole-genome sequence

Pelak et al., 2010, *PLoS Genet.*
Number of novel variants in 200 WGS samples

Number of Unique Variants Remaining

Number of Genomes/Exomes Compared

Genome
50Mb Exome
2.1Mb Exome

Marc Singleton
Novel variants in between-population comparisons
Novel variants in between-population comparisons
False-positive results increase dramatically with inaccurate case-control matching: WGS data

Comparison of 30 European disease cases with mixtures of European and African controls: VAAST analysis of WGS data

Yandell, et al., 2011, Genome Res.
LOD score needed for genome-wide significance for detection of GATA4 as a disease-causing mutation

Chad Huff, PhD
LOD score needed for genome-wide significance for detection of GATA4 as a disease-causing mutation

![Graph showing LOD score needed for genome-wide significance for detection of GATA4 as a disease-causing mutation.](image-url)
Conclusions

• Because of population specificity of rare variants, more sequencing in more populations is needed

• We need to do experiments to determine how closely one needs to match control genomes to minimize false positive findings
Acknowledgments

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

LSU: Mark Batzer
Limited sampling produces high $F_{ST}$ values

HapMap II, 210 individuals, 4 populations

$Fst = 15.2\%$
Reduced genetic differentiation ($F_{ST}$) with more even sampling.

554 individuals, 27 populations

$F_{ST} = 12.3\%$
Reduced genetic differentiation ($F_{ST}$) with more even sampling.

850 individuals, 40 populations

$F_{st} = 11.3\%$

J. Xing et al., 2010, Genomics
Genetic distance analysis: 15 loci

Proportion of shared alleles between pairs of individuals, relative to a single panmictic population.

- Populations from the same continent:
  - CEU-TSI
  - CHB-CHD
  - LWK-YRI
  - CHB-JPT
  - CEU-CHB
  - CEU-YRI
  - CHB-YRI

- Populations from different continents:

MAF = 2%
Examples of genes in which elevated LD indicates recent natural selection

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>G6PD</em></td>
<td>Malaria protection</td>
</tr>
<tr>
<td><em>HFE</em> (hemochromatosis)</td>
<td>Iron absorption</td>
</tr>
<tr>
<td><em>CYP3A5</em></td>
<td>Sodium retention</td>
</tr>
<tr>
<td><em>LCT</em> (lactase enhancer)</td>
<td>Lactase persistence</td>
</tr>
<tr>
<td><em>SLC24A5</em></td>
<td>Skin pigmentation</td>
</tr>
<tr>
<td>Alcohol dehydrogenase</td>
<td>Ethanol metabolism</td>
</tr>
<tr>
<td><em>EPAS1, EGLN1</em></td>
<td>Hypoxia response</td>
</tr>
</tbody>
</table>

Voight et al., 2006, *PLOS Biology* 4: 446-458
Simonson et al., 2010, *Science*
503F Variant of *OCTN1*

- Arose approximately 12,000 years ago; freq. 30-50% in Europe.
- 503F is a gain-of-function mutation that increases ergothioneine substrate efficiency by 300%. 

# Recent Positive Selection at IBD5

<table>
<thead>
<tr>
<th>Sample</th>
<th>iHS</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HapMap CEU</td>
<td>-3.1</td>
<td>0.0007</td>
</tr>
<tr>
<td>HGDP Russian</td>
<td>-2.75</td>
<td>0.0044</td>
</tr>
<tr>
<td>HGDP Sardinian</td>
<td>-2.76</td>
<td>0.0075</td>
</tr>
<tr>
<td>HGDP French</td>
<td>-2.64</td>
<td>0.0076</td>
</tr>
<tr>
<td>HGDP Basque</td>
<td>-2.37</td>
<td>0.0128</td>
</tr>
</tbody>
</table>

The *OCTN1* association could be explained by genetic hitchhiking.

*IRF1* is involved in innate immunity and clearance of intracellular bacteria.
Disease association in 1868 cases and 5540 controls

**IRF1** is expressed 72% more highly in Crohn disease intestinal tissue than in control tissue; no other gene in *IBD5* region shows expression differences.

Linkage disequilibrium* increases with distance from Africa

*We can think of linkage disequilibrium as a measure of multi-locus homozygosity.

Whole-genome sequence data give results congruent with array SNPs

B. Moore et al., 2011, Genetics in Medicine
But results vary by platform

B. Moore et al., 2011, Genetics in Medicine
Ancestral profiles: 250K SNPs

Structure analysis

$K = 4$

Xing et al., 2009, Genome Research
An inferred demographic model, with line width corresponding to population size and time flowing from left to right (1000 Genomes data)

Gravel S et al. PNAS 2011;108:11983-11988
The age, $t$, of a neutral allele can be estimated by its frequency

$$t = \frac{-4Np}{1 - p} \ln(p)$$

(\text{where } N \text{ is effective population size})

Kimura and Ohta, 1973, *Genetics*
Allele sharing and allele frequency for 3,228 \textit{Alu} insertion polymorphisms

<table>
<thead>
<tr>
<th>Allele frequency, p:</th>
<th>(0 &lt; p &lt; 0.05)</th>
<th>(0.05 &lt; p &lt; 0.10)</th>
<th>(p &gt; 0.10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of observing \textit{Alu} outside Africa, given ascertainment in Africa</td>
<td>0.09</td>
<td>0.25</td>
<td>0.80</td>
</tr>
<tr>
<td>Probability of observing \textit{Alu} in Africa, given ascertainment outside Africa</td>
<td>0.41</td>
<td>0.76</td>
<td>0.97</td>
</tr>
</tbody>
</table>

53 African samples (Bantu and Pygmy)
49 non-African samples (Tuscan and Brahmin)