Population genetic inference in the personal genome era

Carlos D. Bustamante

Department of Genetics

Stanford University School of Medicine
Motivation

- Technological developments have dramatically driven down the cost of sequencing and genotyping
- Large-scale projects underway to document genome-wide variation in many species (e.g., human, dogs, rice, cattle) across individuals
  - Quantify genetic differences within and among populations
  - Map genes for traits of interest (e.g., disease-susceptibility, morphology)
  - Reconstruct demographic history and detect targets of recent natural selection
- Personal genomic sequences opens new opportunities and challenges for population genetics.
Motivation and Objectives

• Genome-wide association mapping has been quite successful in identifying common variation (e.g., MAF > 5%) influencing disease risk in Northern European populations

• However, for many traits, only a small proportion of the expected heritability is explained by current GWAS hits (e.g., height -> 150 hits explain 2-4%) and many groups are under-represented in medical genomics research

• Understanding the contribution of rare and common genetic genetic variants will likely require multi-ethnic and trans-ethnic genome-wide studies that compare completely sequenced genomes of many individuals with and without a particular disease

• It will be critical to account for the role of population stratification at fine scales both in terms of genomic and geographic location in these studies
Population Reference Sample (POPRES)

- Assemble large repository of genetically diverse DNA samples (~6,000 samples)
- Generate dense genotype data using key marker panels (Affy 500K)
- Establish resource for studying human population genetics, recent demography, and admixture
- Demographic and genotype data publicly available

Nelson, et al., AJHG (Sept. 2008)
Questions

• To what degree are modern human populations accessible for medical genetic studies sub-structured?

• What are accurate models for describing human genetic diversity?

• Can we use these kinds of data to reconstruct recent “personal” genetic history?

Nelson, et al., AJHG (Sept. 2008)
PopRes + HapMap Phase II

Nelson, et al., AJHG (Sept. 2008)
Data Management and Visualization is a Challenge!

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>SNP 2</th>
<th>SNP 3</th>
<th>...</th>
<th>SNP 500,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indiv 1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>Indiv 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Indiv n</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>aa</th>
<th>Aa</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Principle Components = major axes of variation
Principle Components = major axes of variation
GSK + Stanford HGDP across 70,000 SNPs

Within Switzerland

Continental PCA and Structure Analysis

Reconstructing Personal Genetic History?
Spatial Predictions of Ancestry

\[
x = \beta_{x1}u_1 + \beta_{x2}u_2 + \beta_{x11}u_1^2 + \beta_{x22}u_2^2 + \beta_{x12}u_1u_2 + \varepsilon
\]

\[
y = \beta_{y1}u_1 + \beta_{y2}u_2 + \beta_{y11}u_1^2 + \beta_{y22}u_2^2 + \beta_{y12}u_1u_2 + \varepsilon
\]

(\text{longitude, latitude})

(\text{PC1, PC2})


John Novembre

Matthew Stephens

(U. Chicago)
African American

European

Proportion of African ancestry

\[ P = \frac{b}{(a+b)} \]

Yoruban (YRI)

Dr. Sarah Tishkoff

Approach

- Run PCA on African Americans and diverse potential ancestral populations
- For each individual $i$ at each 15-SNP window $k$, calculate

$$\text{score}_{ik} = M'_{ik} \times e_k$$

$M'_{ik}$ are the normalized and scaled genotypes of the markers in window $k$ for individual $i$

e_k is the vector of loadings corresponding to the markers in window $k$

- Overlay a Hidden Markov Model to make ancestry assignment calls

HMM for Admixture Estimation in African Americans

Individual 27 admixture estimation, Chromosome 1, 15–SNP windows

Individual 34 admixture estimation, Chromosome 1, 15–SNP windows

Individual ancestry results

Chromosome Number

Base pair position

African ancestry

Heterozygous European/African ancestry

European ancestry

~5 minutes for all AfAm’s in the data
Dataset

- Affymetrix 500K SNP arrays

<table>
<thead>
<tr>
<th>Region</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>203</td>
</tr>
<tr>
<td>African American</td>
<td>365</td>
</tr>
<tr>
<td>Europe</td>
<td>400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population</th>
<th>Language</th>
<th>$F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Igbo</td>
<td>Non-Bantu Niger-Kordofanian</td>
<td>0.074</td>
</tr>
<tr>
<td>Brong</td>
<td>Non-Bantu Niger-Kordofanian</td>
<td>0.077</td>
</tr>
<tr>
<td>Yoruba</td>
<td>Non-Bantu Niger-Kordofanian</td>
<td>0.089</td>
</tr>
<tr>
<td>Kongo</td>
<td>Bantu Niger-Kordofanian</td>
<td>0.112</td>
</tr>
<tr>
<td>Bamoun</td>
<td>Bantu Niger-Kordofanian</td>
<td>0.201</td>
</tr>
<tr>
<td>Xhosa</td>
<td>Bantu Niger-Kordofanian</td>
<td>0.257</td>
</tr>
<tr>
<td>Fang</td>
<td>Bantu Niger-Kordofanian</td>
<td>0.266</td>
</tr>
<tr>
<td>Hausa</td>
<td>Afro-Asiatic</td>
<td>0.325</td>
</tr>
<tr>
<td>Kaba</td>
<td>Nilo-Saharan</td>
<td>0.353</td>
</tr>
<tr>
<td>Mada</td>
<td>Afro-Asiatic</td>
<td>0.97</td>
</tr>
<tr>
<td>Bulala</td>
<td>Nilo-Saharan</td>
<td>1.581</td>
</tr>
<tr>
<td>Fulani</td>
<td>Non-Bantu</td>
<td>2.973</td>
</tr>
</tbody>
</table>

Placing Hispanic/Latino Genomic Diversity on the Map

Dr. Harry Ostrer

Chris Velez

Kasia Bryc
Latin American Diversity

Autosomal PCA

Population
- Dominican
- Puerto Rico
- Colombian
- Ecuador
- Mexican
- Maya
- Karitiana
- Surui
- Pima
- Quechua
- Aymara
- Nahua
- African
- Af. Amer.
- Af. Amer.
- European

Source
- Coriell
- POPRES
- HapMap
- Mao
- NYU Latino
- HGDP
Complexity of Admixture
Take home message:

• Personal ancestry reconstruction (including detection of admixture tracts) is feasible on genome-wide scale

• African-Americans exhibit, on average, ~78% West African and 22% European ancestry from SNP chip data with large variation among individuals and genomic segments within individuals

• Hispanic-Latinos vary tremendously in admixture proportions from European, African, and Native American source populations

• A key question is understanding how full genome data will improve ancestry deconvolution and fine-mapping of ancestry “break points”