Estimates of Genetic Ancestry

Chaolong Wang

Sequence Analysis Workshop
December 2014 @ University of Michigan
Outline

• Background
  – Population structure: causes and consequences
  – Population stratification in genetic association studies
  – Existing methods to estimate genetic ancestry

• Our approaches to estimate genetic ancestry
  – How to infer individual ancestry from small amounts of sequencing or genotyping data?
  – Simulations and empirical examples
  – Potential applications
Population structure: Different populations differ in allele frequencies at loci across the genome.

Microsatellite examples:

Rosenberg (2011) *Human Biology*
Causes of population structure

Human migration:

National Geographic

Henn et al. (2012) PNAS
Causes of population structure

**Isolation by distance:** gene flow occurs more frequently between neighboring groups.

Genetic similarity decreases as geographic distance increases.

*Figure: Michael DeGiorgio*
Causes of population structure

**Assortative mating:** mating occurs more often between individuals from similar “classes” (color, education, social stratification).

People from the same “class” are genetically more similar. Example: upper caste and lower caste in India

Figure: Michael DeGiorgio
Population stratification in genetic association studies

**Population stratification**: systematic ancestry differences between subjects with different phenotypes, leading to spurious association.

Price et al. 2010, Nat Rev Genet
Consequence of population structure in association studies

**Population stratification:** systematic ancestry differences between subjects with different phenotypes, leading to spurious association.

**Case-control:**

**Quantitative trait:** (e.g. height)


Methods to estimate population structure

• High-dimensional genotype data (>1K samples, >100K loci)

<table>
<thead>
<tr>
<th>Sample</th>
<th>SNP 1</th>
<th>SNP 2</th>
<th>SNP 3</th>
<th>SNP 4</th>
<th>SNP 5</th>
<th>...</th>
<th>SNP L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>...</td>
<td>2</td>
</tr>
<tr>
<td>Sample 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Sample N</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>...</td>
<td>2</td>
</tr>
</tbody>
</table>

• Model-based clustering methods
  – STRUCTURE/ADMIXTURE/FRAPPE
  – Model allele frequencies of (pre-specified) $K$ discrete clusters
  – Computationally challenging for large datasets
  – Not suitable for continuous population structure

Example for 29 worldwide populations (Jakobsson et al. 2008, Nature)
Methods to estimate population structure

- **High-dimensional genotype data (>1K samples, >100K loci)**

<table>
<thead>
<tr>
<th></th>
<th>SNP 1</th>
<th>SNP 2</th>
<th>SNP 3</th>
<th>SNP 4</th>
<th>SNP 5</th>
<th>...</th>
<th>SNP L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>...</td>
<td>2</td>
</tr>
<tr>
<td>Sample 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Sample N</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>...</td>
<td>2</td>
</tr>
</tbody>
</table>

- **Multivariate dimension reduction methods**
  - Principal components analysis (PCA)
  - Multidimensional scaling (MDS)
PCA in decomposing population structure

- **Worldwide population structure**
  - Human Genome Diversity Panel (HGDP, 53 worldwide populations)

PCA in decomposing population structure

- **European population structure**
  - Population Reference Panel (POPRES, 37 European populations)

Title of the paper: *Genes mirror geography in Europe*
Control of population stratification in association studies

- **Control for stratification using estimated ancestry:**
  - Stratified analysis of subgroups followed by meta-analysis
  - Regression on ancestry principal components
  - Genetic matching of study subjects

- **Other approaches without explicitly estimating ancestry:**
  - Linear mixed models
  - Genomic control

- **These approaches require high-quality genotype data across the genome.**
  - GWAS array genotyping data
  - Whole genome sequencing (when genotypes can be accurately estimated)
Targeted sequencing experiments

• **Targeted sequencing focuses on specific regions of interests.**
  - Our AMD study: 10 candidate regions (2MB in total)
  - Goal: search for additional high-risk (rare) variants that provide functional information about the disease

• **Large sample size is required to provide statistical power to detect association signal for rare variants.**
  - Many studies now include >10,000 individuals
  - Likely to include samples of different ancestry background

• **Correcting for population stratification is difficult for targeted sequencing experiments.**
  - Too few variant loci within targeted regions
Estimating ancestry for targeted sequencing data

What happens when we apply PCA to targeted sequencing data?

Novembre et al. (2008) *Nature*

2,547 SNPs in POPRES data overlapped with whole exome sequencing
Targeted sequencing data

A lot of sequence reads distribute *randomly* and *sparsely* across the off-target genome!
Workflow of target/exome sequencing

Genomic DNA → Construct shotgun library → Fragments → Hybridization → Pulldown → Wash → Captured DNA → DNA sequencing → Mapping, alignment, variant calling

LASER: Locating Ancestry from SEquence Reads

• Traditional methods such as PCA cannot be directly applied on off-target sequencing data.
  – Genotype uncertainty
  – Large amount of missing data

The LASER method:
• Use off-target sequence reads to place sequenced samples one by one into a reference PCA map of ancestry
  - Directly analyze sequence reads without calling genotypes
  - Analyze each sample with a set of reference individuals

"Ancestry estimation and control of population stratification for sequence-based association studies"

Chaolong Wang1,2,10, Xiaowei Zhan2,10, Jennifer Bragg-Gresham2, Hyun Min Kang3, Dwight Stambolian3, Emily Y Chew4, Kari E Branham5, John Heckenlively5, The FUSION Study6, Robert Fulton7, Richard K Wilson7, Elaine R Mardis7, Xihong Lin8, Anand Swaroop8, Sebastian Zöllner2,9 & Gonçalo R Abecasis2

NATURE GENETICS  VOLUME 46 | NUMBER 4 | APRIL 2014
Data used in LASER

• **Study samples**: low-coverage sequencing reads sparsely distributed across off-target regions.

• **Reference samples with known ancestry**: high-quality genome-wide SNP data.
  
  – Human Genome Diversity Panel (HGDP)
    • 938 individuals from 53 worldwide populations
    • 632,958 autosomal SNPs after QC
    • Li *et al.* (2008) *Science*

  – Population Reference Sample (POPRES)
    • 1,385 individuals from 37 European populations
    • 318,682 autosomal SNPs after QC
    • Novembre *et al.* (2008) *Nature*
Step 1: create a reference map

- Generate a reference map by applying PCA on SNP data of N reference individuals. (Map 0)

- Geographic map
  - 37 populations, 1,385 individuals
  - 318,682 autosomal SNPs
  - Novembre et al. (2008) *Nature*
Step 2: adjust reference to each sample

Given a sample $i$ that was sequenced with coverage $C_{ij}$ at locus $j$, for $j=1,2,\ldots,L$.

Simulate sequence data for all reference individuals with coverage at each locus $j$ equal to $C_{ij}$.

$$P(\text{drawing a read } A) = \begin{cases} 1 - e & \text{if } g_{ij} = AA \\ 0.5 & \text{if } g_{ij} = AB \\ e & \text{if } g_{ij} = BB \end{cases}$$

Ref 1: $\{AB, AA, BB, BB, AB\}$

Ref 2: $\{BB, AB, AA, AA, BB\}$
Step 2: adjust reference to each sample
Step 3: count the variant bases at each locus

Sample i

Ref 1:

Ref 2:
Step 4: construct a sample-specific map

- Perform PCA on combined sequencing data of sample i and N reference individuals. (Map i)
Step 5: find optimal transformation

- **Procrustes analysis**: transform Map $i$ to optimize the similarity to Map 0 based on $N$ reference samples.

Transformations: $f(Y) = \rho YA + B$

Minimize objective: $d(X, f(Y)) = \sum_{i=1}^{N} (x_i - f(y_i))^T (x_i - f(y_i))$

Step 6: apply transformation

Apply the transformation on coordinates of sample i to place it into the reference PCA map.
Step 7: repeat!

- Repeat steps 2-6 for all sequenced samples one by one.
LASER: Locating Ancestry from SEquence Reads

Human Genome Diversity Panel (HGDP)
PCA on the HGDP data

A

HGDP samples
- Africa
- Europe
- Middle East
- C/S Asia
- East Asia
- Oceania
- America

B
Simulations based on HGDP

**HGDP:** 938 individuals at 632,958 autosomal SNP loci

**Test set:** 238 individuals;  **Reference set:** 700 individuals
## Simulations based on HGDP

### Sequence-based coordinates vs. SNP-based coordinates

<table>
<thead>
<tr>
<th>Simulated mean coverage</th>
<th>Expected number of loci with reads</th>
<th>Pearson correlation of PC1</th>
<th>Pearson correlation of PC2</th>
<th>Pearson correlation of PC3</th>
<th>Pearson correlation of PC4</th>
<th>Procrustes similarity $t_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>140,010</td>
<td>0.9998</td>
<td>0.9998</td>
<td>0.9996</td>
<td>0.9994</td>
<td>0.9997</td>
</tr>
<tr>
<td>0.20</td>
<td>114,736</td>
<td>0.9998</td>
<td>0.9998</td>
<td>0.9996</td>
<td>0.9993</td>
<td>0.9996</td>
</tr>
<tr>
<td>0.15</td>
<td>88,166</td>
<td>0.9997</td>
<td>0.9998</td>
<td>0.9994</td>
<td>0.9989</td>
<td>0.9995</td>
</tr>
<tr>
<td>0.10</td>
<td>60,234</td>
<td>0.9996</td>
<td>0.9996</td>
<td>0.9991</td>
<td>0.9987</td>
<td>0.9993</td>
</tr>
<tr>
<td>0.05</td>
<td>30,870</td>
<td>0.9994</td>
<td>0.9993</td>
<td>0.9982</td>
<td>0.9973</td>
<td>0.9989</td>
</tr>
<tr>
<td>0.01</td>
<td>6,298</td>
<td>0.9974</td>
<td>0.9966</td>
<td>0.9909</td>
<td>0.9857</td>
<td>0.9949</td>
</tr>
<tr>
<td>0.008</td>
<td>5,043</td>
<td>0.9970</td>
<td>0.9960</td>
<td>0.9891</td>
<td>0.9830</td>
<td>0.9940</td>
</tr>
<tr>
<td>0.006</td>
<td>3,786</td>
<td>0.9948</td>
<td>0.9941</td>
<td>0.9834</td>
<td>0.9791</td>
<td>0.9911</td>
</tr>
<tr>
<td>0.004</td>
<td>2,527</td>
<td>0.9947</td>
<td>0.9941</td>
<td>0.9765</td>
<td>0.9668</td>
<td>0.9887</td>
</tr>
<tr>
<td>0.002</td>
<td>1,265</td>
<td>0.9877</td>
<td>0.9852</td>
<td>0.9468</td>
<td>0.9141</td>
<td>0.9729</td>
</tr>
<tr>
<td>0.001</td>
<td>633</td>
<td>0.9750</td>
<td>0.9689</td>
<td>0.9138</td>
<td>0.8600</td>
<td>0.9508</td>
</tr>
</tbody>
</table>
Simulations based on POPRES

**POPRES:** 1,385 individuals at 318,682 autosomal SNP loci

**Test set:** 385 individuals; **Reference set:** 1,000 individuals
Simulations based on POPRES

Sequence-based coordinates vs. SNP-based coordinates

<table>
<thead>
<tr>
<th>Simulated mean coverage</th>
<th>Expected number of loci with reads</th>
<th>Sequence-based coordinates vs. SNP-based coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pearson correlation of PC1</td>
</tr>
<tr>
<td>0.40</td>
<td>105,063</td>
<td>0.9927</td>
</tr>
<tr>
<td>0.35</td>
<td>94,111</td>
<td>0.9933</td>
</tr>
<tr>
<td>0.30</td>
<td>82,597</td>
<td>0.9906</td>
</tr>
<tr>
<td>0.25</td>
<td>70,492</td>
<td>0.9898</td>
</tr>
<tr>
<td>0.20</td>
<td>57,767</td>
<td>0.9868</td>
</tr>
<tr>
<td>0.15</td>
<td>44,390</td>
<td>0.9825</td>
</tr>
<tr>
<td>0.10</td>
<td>30,327</td>
<td>0.9752</td>
</tr>
<tr>
<td>0.05</td>
<td>15,542</td>
<td>0.9408</td>
</tr>
<tr>
<td>0.01</td>
<td>3,171</td>
<td>0.7541</td>
</tr>
</tbody>
</table>
Homogeneous samples need more markers

- Homogeneous samples need more markers to reveal their geographic structure of genetic variation.
- Europe is the most homogeneous continental group.

<table>
<thead>
<tr>
<th>Region</th>
<th>$F_{ST}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>9.704</td>
</tr>
<tr>
<td>Europe</td>
<td>0.212</td>
</tr>
<tr>
<td>Africa</td>
<td>1.334</td>
</tr>
<tr>
<td>Asia</td>
<td>4.706</td>
</tr>
<tr>
<td>E. Asia</td>
<td>1.874</td>
</tr>
<tr>
<td>C.S. Asia</td>
<td>2.140</td>
</tr>
</tbody>
</table>

Estimate ancestry from SNP genotypes

**Same framework for analyzing genotype data:**

- **Motivations:**
  - Sequence reads might not be available
    - Array-genotyping data
  - Joint analysis of sequencing and array-genotyping data
  - Computational time scales linearly with sample size
    - PCA scales cubically
  - Robust to family structure within the sample
  - Can handle large amounts of missing data
    - Ancient DNA samples
    (Skoglund et al. 2012, *Science*)
Challenges

• Small number of overlapped markers
  – POPRES European reference panel: ~319K SNPs after QC
  – ExomeChip array: ~273K SNPs by design
  – Shared by POPRES and ExomeChip: 3,983 SNPs

• Too expensive to whole-genome sequence a large reference sample
Impute the reference panel

- Use 1000 Genomes data to impute POPRES
  - Imputed POPRES: 4.2 million SNPs after QC
  - Overlapped with the ExomeChip: 19,123 SNPs

- PC1 reflects the north-south population structure.
- PC2 reflects some imputation artifacts.
- The east-west population structure is likely captured by higher order PCs.
Imputation causes artifacts in PCA

PCA on the **original** POPRES data

PCA on the **imputed** POPRES data

Association with PC2 of the imputed POPRES data

Driven by $\sim$9000 SNPs around the centromere of chromosome 11
Project from high-dimensional PC space

**K-dimensional** PCA map $X$ based on the genotyped SNPs of the reference panel. 

**Project Procrustes analysis:**
Search for a set of $K'$-to-$K$ dimensional transformations $f$ in $Y$ such that the similarity between $f(Y)$ and $X$ is maximized.

- **Projection**, rotation, reflection, translation, scaling

**Solution:**
The transformation $f$ does not have close form solution, but can be numerically solved using an iterative algorithm.

**K'-dimensional** PCA map $Y$ based on the SNPs shared by the imputed reference panel and the study sample ($K'\geq K$)
Project from high-dimensional PC space

• Combining imputation and high-dimensional projection can substantially improve the ancestry estimation!

True ancestry

19,123 SNPs, $K' = 20$

$t_0 = 0.9277$ when $K' = 20$
LASER 2.0

- Same strategies can be used to improve ancestry estimation from off-target sequence reads.

Simulation:

1. Take off-target coverage patterns from the Exome Sequencing Project, and down sample to 5% of the original coverage (~0.05X on average).

2. Simulate sequence reads based on the genotypes of 385 POPRES Europeans.
Application to the AMD study

- **Targeted sequencing of 10 AMD risk loci**
  - Sequenced at 127X across 0.97Mb targeted region
  - The off-target region is covered at ~0.2X on average
  - Sequenced 2,348 cases and 789 controls
AMD samples on HGDP reference panel
Genotype vs. sequence for 931 AMD samples

SNP-based coordinates

Sequence-based coordinates
SNP vs. sequencing for 928 AMD samples

A

SNP-based

SNPs

PC1

PC2

B

Sequence-based

Off-target sequencing

PC1

PC2
LASER 1.0 vs. LASER 2.0

3,066 AMD sample with European ancestry

LASER 1.0
K’=2, original POPRES reference

t₀=0.9171

LASER 2.0
K’=20, imputed POPRES reference

t₀=0.9628
**LASER 1.0 vs. LASER 2.0**

*Sample-specific Procrustes similarity score* (partially) reflects estimation accuracy of each sample.

Ancestry estimation improves for all samples, especially for samples that have extremely low coverage off-target data.
Be cautious in choosing reference panel

- **Limitation:** results might be difficult to interpret when the reference panel does not include relevant ancestry groups.

- **Recommendation:** start with a worldwide reference panel and gradually narrow down to fine-scale regional panels.
Potential applications

• Control of population stratification in genetic association studies.
  – Regression on ancestry coordinates
  – Match ancestry background of study samples
  – Facilitate modeling of the relationship between phenotypes and ancestry coordinates

• Can apply to study sequencing data of ancient DNA samples, which often have abundant missing data.
  – Skoglund et al. (2012, Science) investigated the genetic relationship of four ancient DNA samples in Europe with modern humans using a similar approach.
Application to the AMD study

• **Targeted sequencing of 10 AMD risk loci (0.97Mb)**
  – To search for additional high-risk (rare) variants that can provide information about function
  – Sequenced 2,348 cases and 789 controls
    • Known high-risk variant R1210C in CFH gene has P=2.6x10^{-3}
    • Not enough sample size for studying rare variants.

• **Expanding our experiment**
  – Identify additional ancestry-matched controls from public resources to augment our sample size
  – Plan
    • Place AMD samples in the worldwide ancestry map
    • Place other sequenced samples in the same map
    • Identify matched controls for all cases
Matching results

• Search for matches from >6,800 samples in the Exome Sequencing Project

• Build matched set
  – 2,268 AMD cases
  – 2,268 matched controls
  – Focused on sites with >10X depth
  – Exclude sites near indels
  – 430 protein changing variants in both ESP and AMD experiments

• R1210C variant now has P=2.9x10^{-6} (initial P=2.6x10^{-3})

• A new rare variant K155Q in the C3 gene: P=2.7x10^{-4} (initial P=6.3x10^{-3}), OR=2.68

Zhan et al. (2013) Nature Genetics
Validation of the K155Q variant

Zhan et al. (2013) Nature Genetics

<table>
<thead>
<tr>
<th>Sample set</th>
<th>Controls</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>MAF</td>
</tr>
<tr>
<td><strong>Discovery sample</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequenced samples ($N = 4,536$)</td>
<td>2,268</td>
<td>0.004</td>
</tr>
<tr>
<td><strong>Follow-up samples</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany: University of Regensburg ($N = 2,976$)</td>
<td>1,147</td>
<td>0.006</td>
</tr>
<tr>
<td>United States: Vanderbilt/Miami ($N = 1,819$)</td>
<td>726</td>
<td>0.004</td>
</tr>
<tr>
<td>Netherlands: Rotterdam Study ($N = 1,409$)</td>
<td>1,280</td>
<td>0.005</td>
</tr>
<tr>
<td>UK: Cambridge AMD Study ($N = 1,279$)</td>
<td>423</td>
<td>0.006</td>
</tr>
<tr>
<td>United States: University of California, Los Angeles/University of Pittsburgh ($N = 830$)</td>
<td>211</td>
<td>0.004</td>
</tr>
<tr>
<td><strong>deCODE study</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deCODE discovery sample ($N = 52,578$)</td>
<td>51,435</td>
<td>0.005</td>
</tr>
<tr>
<td><strong>Meta-analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All follow-up samples ($N = 8,313$)</td>
<td>3,787</td>
<td>0.005</td>
</tr>
<tr>
<td>Discovery and all follow-up samples ($N = 12,849$)</td>
<td>6,055</td>
<td>0.005</td>
</tr>
<tr>
<td>Discovery, all follow-up and deCODE samples ($N = 65,427$)</td>
<td>57,490</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Summary

- A statistical framework to trace individual ancestry in a reference PCA space.
  - Accurate even with small amounts of sequence/genotype data
  - Robust to family structure and sampling distribution
  - Computationally efficient, $\sim O(n)$
  - Easy for parallel computation

- High-dimensional projection and genotype imputation can substantially improve the accuracy of our ancestry estimates.

- Software package and references:
  - **LASER**: [http://www.sph.umich.edu/csg/chaolong/LASER/](http://www.sph.umich.edu/csg/chaolong/LASER/)
  - **Version 2**: Wang et al., Improved ancestry estimation for both genotyping and sequencing data using projection Procrustes analysis and genotype imputation. *AJHG* (under review).
Other resources: admixed samples

- SEQMIX
  - Improve estimation of local ancestry for admixed samples using low-coverage off-target sequence data