Efficient algorithms for ascertaining markers for controlling for population substructure

Oscar Lao
Department of Forensic Molecular Biology
Erasmus Medical Center (Rotterdam)
New Jersey 2009
Workflow

1. Human population substructure
   - How to detect it?
   - How much?
   - Where does it come from?

2. Why does it matter?

3. Ancestry Sensitive Markers (ASMs) / Ancestry Informative Markers (AIMs)
   - Hypothesis driven. Particular individual clusters are preferred
     - ASMs
     - PhenoASMs
How much there is and how much can be detected. The two sides of the same coin

Plato’s cave myth
Population substructure

DETECTION

- STRUCTURE
- BAPS
- FRAPPE
- GENELAND
- PCA/MDS + K means
- Neural Networks
- ...

Sometimes results are NOT reproducible
Population substructure

HOW MUCH?

Which type?
- Phenotype
- Genotype
  - Y chromosome
  - mtDNA
  - Autosomal markers

Where?
- Worldwide
- Regional (I will focus on Europe)
Phenotypic substructure

Light Eyes Map
according to
Carleton S. Coon
“The Races of Europe”

- 75% - 95%
- 55% - 75%
- 35% - 55%
- 25% - 35%
- 1% - 25%
Y chromosome

The data in this map is supposed to represent the situation before the recent European expansion beginning about 1500 AD. In some cases such as some Native American tribes and the Maori this can be done relatively because STR typing was done. In other cases, especially in America, it is guesswork. The “Other” sectors in America indicate this. Native American groups are labeled by language group as Athabascan, Na-Dene (N-C), and Eskimo. F, K, L, and P are in some cases “catchall” groups because some researchers did not use enough markers for a full haplogroup determination.
mtDNA

Specific tribes or locations are shown at left. Unlabelled pies are for general population in the area. African, American, and especially Polynesian areas are very large. The data in this chart is supposed to represent the situation before the recent European expansion beginning about 1500 AD. Assignments in Australia are somewhat off.
Classical markers

Cavalli-Sforza et al. 1994
1064 samples
51 human populations of global distribution
Autosomal SNPs

650,000 SNPs (FRAPPE)
Africa          Mid.East          Europe          C.S.Asia         E.Asia

550,000 SNPs (STRUCTURE)

Haplotypes

Li et al Science 2008

Jakobsson et al Nature 2008
A set of European populations

23 populations

500 Affy Array
300,000 SNPs

Lao and Lu et al Current Biology 2008
Autosomal SNPs in Europe

Novembre et al Nature 2008
K = 2; Admixture

Correlation with latitude

$R^2 = 0.86$

Correlation with longitude

$R^2 = 0.01$
World

Anayet peak (2574 m), Pyrenees

Europe

Keukenhof garden (-2 m), Netherlands
Non random distribution of population substructure

Chr2. Comparison CEPH Europeans vs CHB Asians

EDAR (positive selection in Asians)
Non random distribution of population substructure

Chromosome 2

LCT (positive selection in North European populations)

Lao and Lu et al Current Biology 2008
Demography shapes the population substructure

Cavalli-Sforza & Feldman Nature Genetics 2003

Simoni et al AJHG 2000
• Selective pressures within the species (locus specific)

Lactose tolerance
Malaria resistance
Human pigmentation

...
• Population substructure & pigmentation (5 SNPs)

MDS

- Europe
- Africa
- Middle East
- Oceania
- America
- C/SAsia
- EAsia
- N Africa

Lao et al Ann Hum Genet 2007
Why population substructure: Confounding factor

CASES

A

G

CONTROLS

A

G

Lung cancer (ICD-9 162)
Smoothed RR, 1986-1998

- 1.5 and more: 43
- 1.3 - 1.5: 180
- 1.1 - 1.3: 559
- 1.05 - 1.1: 263
- 0.95 - 1.05: 596
- 0.85 - 0.95: 356
- 0.77 - 0.91: 1564
- 0.67 - 0.77: 1663
- 0 - 0.67: 3026
Population substructure: improving the detection

Plato’s cave myth

CHANGE THE ALGORITHM FOR DETECTING POPULATION SUBSTRUCTURE
Plato’s cave myth

INCREASE THE RESOLUTION TO SEE THE OBJECTS
• Markers that capture most of the genetic ancestry
  – Estimate ancestry
  – Reduce the number of markers to test for genetic homogeneity
    • Time cost (clustering algorithms can be extremely computational intensive)
    • Economical cost (i.e. exclude individuals BEFORE doing the GWA)
• Based on the existing diversity between individuals (i.e. Paschou et al 2008)

• Based on predefined groups of individuals
  – No phenotype linked
    • Large Genetic distances
    • Signals of positive selection
  – Phenotype linked
    • Covariates with the phenotype of interest
A basic algorithm to ascertain ASMs

• Use a statistic to quantify the amount of differentiation between populations

• Compute the OVERAL non-redundant amount of In between set of SNPs

• Take the best combination of markers from all the possible combinations

• Repeat the process until the information of the set of markers is maximum
A statistic to ascertain ASMs

**informativeness for assignment**

\[
I_n(Q; J) = \sum_{j=1}^{N} \left( -p_j \log p_j + \sum_{i=1}^{K} \frac{p_{ij}}{K} \log p_{ij} \right)
\]

• How much information a marker contains about the ancestry of one individual (measured in \textit{nats})

• Ranges from 0 to the natural logarithm of the number of clusters and it is proportional to the number of differentiated clusters
A statistic to ascertain ASMs

- Computes the non-redundant amount of information when considering more than one marker
- Requires computing the frequency of **ALL** the allelic combinations when considering more than 1 locus
• Problem: The number of combinations increases exponentially with the number of markers.
  – Number of allelic combinations considering 50 SNPs:

\[ 2^{50} = 1,125,899,906,842,624 \]
A way to compute In

\[ I_n(Q; J) = \sum_{j=1}^{N} \left( -p_j \log p_j + \sum_{i=1}^{K} \frac{p_{ij}}{K} \log p_{ij} \right) \]

\[ I_n(Q; J) = \sum_{j=1}^{N} \left( \overline{H_j} - \sum_{i=1}^{K} \frac{H_{ij}}{K} \right) \]

\[ H \approx \frac{1}{N} \sum_{i=1}^{N} \ln(p) \]

By applying the Asymptotic Equipartition Property of Entropy
• Problem: Considering 8,000 markers, ascertaining the best set of 50 markers requires computing:

\[
N_{\text{combinations}} = \frac{8,000!}{50!(8,000 - 50)!} \approx 4 \times 10^{130}
\]
A method to ascertain ASMs

- Population of answers
- Select best answers (>In)
- Random mating
- Recombination/Mutation

Next generation
10k Affymetrix Array (~9000 SNPs after excluding X-SNPs & missing SNPs)

YCC-panel
76 human individuals
21 sampling localities

SNP ascertainment (10 SNPs)

CEPH-HGDP panel
1064 samples
51 human populations of global distribution

Perlegen Database
3 Human populations
~1,500,000 SNPs
(most informative
5 SNPs)

Reproducibility of geographic structure in a different dataset

Test for signatures of positive selection (EHH test)

The genetic algorithm was applied increasing every time the number of selected SNPs.

---

**Selected SNPs in the final 10 SNPs run**

<table>
<thead>
<tr>
<th>Marker name</th>
<th>Chromosome</th>
<th>Gene name</th>
<th>$I_N$ (%) from 4 groups YCC panel</th>
<th>$I_N$ (%) from 7 groups CEPH-HGDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs722869</td>
<td>14</td>
<td>VRK1</td>
<td>29.066</td>
<td>7.960</td>
</tr>
<tr>
<td>rs1858465</td>
<td>17</td>
<td>LOC442008</td>
<td>25.637</td>
<td>9.228</td>
</tr>
<tr>
<td>rs1876482</td>
<td>2</td>
<td>LOC442008</td>
<td>24.589</td>
<td>10.290</td>
</tr>
<tr>
<td>rs1344870</td>
<td>3</td>
<td></td>
<td>22.810</td>
<td>11.074</td>
</tr>
<tr>
<td>rs1363448</td>
<td>5</td>
<td>PCDHGB1</td>
<td>19.418</td>
<td>4.552</td>
</tr>
<tr>
<td>rs952718</td>
<td>2</td>
<td>ABCA12</td>
<td>18.739</td>
<td>9.472</td>
</tr>
<tr>
<td>rs2352476</td>
<td>7</td>
<td></td>
<td>18.317</td>
<td>5.603</td>
</tr>
<tr>
<td>rs714857</td>
<td>11</td>
<td></td>
<td>18.031</td>
<td>6.157</td>
</tr>
<tr>
<td>rs1823718</td>
<td>15</td>
<td>RYR3</td>
<td>17.845</td>
<td>5.451</td>
</tr>
<tr>
<td>rs735612</td>
<td>15</td>
<td>RYR3</td>
<td>14.315</td>
<td>5.530</td>
</tr>
</tbody>
</table>

993 autosomal markers

10 SNPs  No admixture

Admixture

Ameri  O  E-Asia  C/S-Asia  Europe  M. East  Africa

ASM for continental differentiation using HapMap III

K = 6 (1000 (randomly ascertained) markers, Admixture, 10,000 burning, 10,000 retained simulations)

K = 5 (50 markers, Admixture, 500,000 burning, 500,000 retained simulations)

K = 6 (100 markers, Admixture, 100,000 burning, 100,000 retained simulations)
ASMs for continental differentiation using Illumina 650k

25 ascertained markers. PCA
ASMs for continental differentiation using Illumina 650k

- CEPH

K = 5 (50 ascertained markers, Admixture, 500,000 burning, 500,000 retained simulations)
K = 2 (5000 random markers, Admixture, 10,000 burning, 10,000 retained simulations)

K = 3 (500 ascertained markers, Admixture, 10,000 burning, 10,000 retained simulations)
Use of the 500 ASMs for correcting the effect of population substructure

Association between OCA_HERC2 region and iris color adjusted for ancestry sensitive markers

-log10(P-value) vs Physical position (Mb)

-log10(P-value)

Physical position (Mb)

- 500KSNPs
- 498AncestrySensitiveSNPs
- Random498SNPs
- NoAdjustment(Original_AJHG)
• Recall
  – Population substructure is only a problem when PHENOTIPIC and GENOTYPIC variation covariates
  – Why not ascertaining markers that are associated to the particular spatial pattern of the phenotype?

\[
I_n(Q; P \mid J) = I_n(Q; P; J) - I_n(Q; J)
\]

“Amount of information of the phenotype (P) conditional on the genotype (J): How well could we correctly classify one individual given that we know his phenotype if we already know his genotype in a particular locus”
PhenoASMs for Crohn disease

Replication of signals from recent studies of Crohn’s disease identifies previously unknown disease loci for ulcerative colitis

Andre Franke¹,⁵, Tobias Balschun¹,⁵, Tom H Karlsen², Jürgen Hedderich³, Sandra May¹, Tim Lu³, Dörthe Schuldt¹,⁴, Susanna Nikolaus⁴, Philip Rosenstiel¹, Michael Krawczak³ & Stefan Schreiber¹,⁴

Following up on recent genome-wide association studies (GWAS) of Crohn’s disease, we investigated 50 previously reported susceptibility loci in a German sample of individuals with Crohn’s disease (n = 1,850) or ulcerative colitis (n = 1,103) and healthy controls (n = 1,817). Among these loci, we identified variants in 3p21.31, NKX2-3 and CCNY as susceptibility factors for both diseases, whereas variants in PTPN2, HERC2 and STAT3 were associated only with ulcerative colitis in our sample collection.
PhenoASMs: a little bit further

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AB</th>
<th>BB</th>
<th>Marginal phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>$P(AA)P(C</td>
<td>AA)$</td>
<td>$P(AB)P(C</td>
<td>AB)$</td>
</tr>
<tr>
<td>D</td>
<td>$P(AA)P(D</td>
<td>AA)$</td>
<td>$P(AB)P(D</td>
<td>AB)$</td>
</tr>
</tbody>
</table>
• Update $\theta$ with a Metropolis algorithm
• Update the covariance matrix of the proposal distribution by means of a “quasi-perfect adaptive MCMC” (Andrieu and Atchade)
• Compute the harmonic mean of the likelihood in order to obtain a rough estimate of $P(M|D)$
Phenotype-genotype association for eye color
Phenotype-genotype association for bitter taste

TAS2R38

4/27/2009
New Jersey 2009
Phenotype-genotype association for bitter taste
Phenotype-genotype association for bitter taste
Low to moderate human population differentiation
• Mainly associated to geography
• No sharp discontinuities, except in particular genomic regions (selection?)
• Results depend on the clustering algorithm
• ASMs can improve the detection of population substructure
Conclusions

• $I_n$ is a good statistic for ascertaining markers to differentiate predefined populations
• If a prior definition of a population is used, ASMs will tend to differentiate such population, independently of the biological meaning
• PhenoASMs as the next level of ASMs?
In collaboration with

Acknowledgements

Tim Lu

Andreas Wollstein
Petros Drineas
Peristeia Paschou

Manfred Kayser

Michael Krawczak
Thank you very much!