Population genetic and admixture analyses

Eimear Kenny
Assistant Professor
Institutes of Personalized Medicine and Multiscale Biology at Mount Sinai School of Medicine

Representing and presenting slides from the 1000 Genomes Project Consortium
Phase1: 1092 samples from 14 populations
Motivation for Local Ancestry Inference

Using 1000 Genomes admixed populations to:

- study impact of recent admixture on patterns of genomic variation
- understand pre-historical demography
- inform variant discovery for medical genomics and personalized medicine
Admixture

Recombination creates mosaic of ancestry
Phase 1 AMR Samples

Native Americans (Mao et al.)

PC2

PC1
Accuracy assessment using simulated haplotypes
Simulated ASW

Simulated MXL

1. Generate ancestry tract locations with Wright-Fisher model

   - CEU
   - YRI
   - NatAm

2. Add bases to tracts using phased reference haplotypes from HapMap + Mao et al.
Simulated ASW

Simulated MXL

reference panel

CEU | YRI | NatAm

LAMP | HapMix | MultiMix | RFMix

Per site diploid calls

Compare ancestry calls

Provide reference panel from other HapMap samples + Mao et al

Each group runs local ancestry estimation

3- or 6-way diploid call per site
Example miscall

Genetic Position (cM)

Truth

Called

miscalloing wrong ancestry

Called

micalling wrong ancestry

Called

Genetic Position (cM)

0 50 100 150

CEU

YRI

NAT
Results from the simulated data: some methods perform better than others

Simulated admixed individuals

- HAPMIX
- MULTIMIX
- RFMIX
- LAMP

**Accuracy**

Simulated ASW

Simulated MXL
HapMix MXL ($K = 3$) Results
Strategy to obtain highest confidence calls for downstream analysis

Consensus calls for **97.4%** MXL and **99.3%** ASW genomes
Marginal improvement for ASW

Accuracy

HAPMIX LAMP MULTIMIX RFMIX

Majority Vote

Simulated ASW
For downstream analysis select high confidence regions

Simulated MXL

Majority Vote

Average Accuracy
Sequence diversity in admixed TGP samples
Inferred local ancestry calling in admixed samples

Select Affy6.0 sites

NatAm Affy 6.0 (Mao et al) CEU + YRI TGP

reference panel

ASW

CLM

PUR

MXL

LAMP

HapMix

MultiMix

RFMix

Diploid consensus calls

Phase Omni + Sequence data

OR

Used Omni data for trio phasing

Phased unrelated (10 CLM + 1 MXL)

*Marchini (Shapeit) approach used for phasing all samples together
Good agreement
Not so good agreement

(Probably best to “mask” and not make call…)
Summary

• Admixture deconvolution produced for all Phase 1 samples and pushed to DCC
  – Merged calling improved accuracy
  – Estimated 99%+ accuracy for ASW and 96-97% for MXL, CLM, PUR samples

• Local ancestry calls are available as bed files per individual on the DCC
Some cool applications
Inferring “ancestral” allele frequencies

What are the allele frequencies in each population?

-Bayesian answer

$$P(f|D) = \frac{P(D|f)P(f)}{\int df' P(D|f')P(f')}$$
Inferring “ancestral” allele frequencies

What are the allele frequencies in each population?

- Bayesian answer

\[
P(f|D) = \frac{P(D|f)P(f)}{\int df' P(D|f')P(f')}
\]

Calculate across all sites
Confidence interval width varies by population

<table>
<thead>
<tr>
<th>Pop</th>
<th>Nat Hap per site</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUR</td>
<td>20</td>
</tr>
<tr>
<td>CLM</td>
<td>60</td>
</tr>
<tr>
<td>MXL</td>
<td>110</td>
</tr>
</tbody>
</table>

Counts

Width of CI
Summary

- SNP diversity, novelty rate, and Non-Syn/Syn ratios mirror demography
- Recover allele frequencies in ancestral populations
- Samples and pipelines are useful beyond TGP
Admixture Working Group

• Bustamante lab (Eimear Kenny, Simon Gravel, Fouad Zakharia, Brian Maples)
• Marchini lab (Claire Churchouse)
• Halperin lab (Yael Baran)
• Myers lab (Anjali GuptaHinch)
• Burchard (Chris Gignoux)
• Abigail Bigham and Mark Shriver (Mao et al. Affy 6.0 data)
Credits

More information at www.1000genomes.org