Enhanced disease scoring, imputation and fine-mapping for GWAS in admixed populations: assessment of increased power using African Americans from CARe

Bogdan Pasaniuc

Harvard School of Public Health
Admixed populations

- Most studies performed in homogeneous populations (Europeans)
- Admixed Populations (African-American, Latinos)
  - Additional genetic diversity
  - LD – fine scale in ancestral populations
  - Admixture-LD – segments of distinct ancestry
African American from CARe

Data from CARe Consortium
Case Control Association Studies

- Use LD between typed SNPs and unobserved causal variants
- Methods: Regression, Armitage Trend Test
- Account for false positives:
  - EMMAX [Kang et al ‘10], EIGENSTRAT [Price et al ‘06]
- Standard test in African-Americans:
  - Armitage Trend Test (ATT) with adjustment for genome-wide ancestry

Admixture Mapping: Finding regions with disproportional local ancestry.

Admixture association **signal arises** when the causal variant has **large allele frequency difference** between ancestral populations.

Examples:
- Multiple Sclerosis [Reich et al’05], Hypertension [Zhu ’05], prostate cancer [Freedman ‘06], obesity[Cheng’09], end stage kidney disease [Shlush’10]...
Local ancestry inference

- Traditionally inferred using AIMS
- GWAS dense genotyping ➔ many methods
  - **Ancestral Haplotypes:** HAPMIX [Price et al ‘09], GEDI-ADMX [Pasaniuc et al ‘09], HAPAA [Sundquist et al ‘08] ...
  - **Ancestral allele frequencies:** ANCESTRYMAP[Patterson et al ‘04], LAMP-ANC [Sankararaman et al’08], WINPOP[Pasaniuc et al ‘09] ...
  - **No ancestral information:** LAMP[Sankararaman et al’08]
- Distant ancestral populations ➔ high accuracy
  - **African-Americans** simulated data:
    - No ancestral information ➔ ~94% (LAMP)
    - Ancestral haplotypes ➔ ~99%, $r^2$~0.98 (HAPMIX)
- No artifacts ➔ no false peaks!
HAPMIX local ancestry in 6,209 CARe AA shows no spurious peaks
GWAS in admixed populations

- Either admixture or case-control signals but not both
  - Armitage Trend Test with adjustment for genome-wide ancestry (ATT)
  - Admixture association using cases only (ADM)
- Case-control signal vs. Admixture signal
- Combined approach to increase power?
Case-Control signal

NULL:

\[ p \]

CAUSAL:

\[ p, R \]

Allele frequency

Odds Ratio

LLRT:

\[ \chi^2 \text{ (1 dof)} \]
Association conditioned on local ancestry

NULL:
- African: $p_A$
- European: $p_E$

CAUSAL:
- $p_A$, $R$
- $p_E$, $R$

LLRT:
- $\chi^2$ (1 dof)

Same odds ratio $R$ in Europeans and Africans!

[Teslovich et.al. Nature 2010]
Case only admixture association

**NULL:**

Genome-wide ancestry

\[ \theta_i \]

**CAUSAL:**

\[ \theta_i, \Omega \]

Ancestry odds ratio

**LLRT:**

\[ \chi^2 (1 \text{ dof}) \]
Combining Admixture and Case-Control Signals

**NULL:** $p_A, p_E$

**CAUSAL:** $p_A, p_E, R$

Case-Control

Admixture

SUM:

$\chi^2$ (2 dof)
Combining Admixture and Case-Control Signals

**NULL:**
- Case-Control: $p_A, p_E$
- Admixture: $\theta_i$

**CAUSAL:**
- Case-Control: $p_A, p_E, R$
- Admixture: $\theta_i, \Omega(R)$

**MIX:**
- $\chi^2 (1 \text{ dof})$

Write $\Omega$ as function of $R$ and $p_A, p_E$!
Highly differentiated SNPs
MIX outperforms all other scores on simulations

Multiple Sclerosis, Prostate cancer, End-stage kidney disease …

**Increase** in power at high delta SNPs $\text{abs}(f_{CEU}-f_{YRI})>0.4)$:

- **24%** for MIX over ATT at $R=1.5$

<table>
<thead>
<tr>
<th></th>
<th>R=1.2</th>
<th>R=1.5</th>
<th>R=2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATT</td>
<td>0.0026</td>
<td>0.5533</td>
<td>0.9769</td>
</tr>
<tr>
<td>SUM</td>
<td>0.0028</td>
<td>0.624</td>
<td>0.9874</td>
</tr>
<tr>
<td>MIX</td>
<td><strong>0.0046</strong></td>
<td><strong>0.6899</strong></td>
<td><strong>0.9907</strong></td>
</tr>
</tbody>
</table>

- Scores computed at simulated causal
- 100,000 SNPs, 1,000 cases, 1,000 controls over 6209 AA CARe individuals
- Power at $P<5e-08$
Random SNPs
MIX outperforms all other scores on simulations

Rheumatoid arthritis, Schizophrenia, T2D, Crohn's disease, …

**Increase** in power at random SNPs:

- **8%** for MIX over ATT at R=1.5

<table>
<thead>
<tr>
<th></th>
<th>R=1.2</th>
<th>R=1.5</th>
<th>R=2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATT</td>
<td>0.0017</td>
<td>0.3803</td>
<td>0.8351</td>
</tr>
<tr>
<td>SUM</td>
<td>0.0012</td>
<td>0.3555</td>
<td>0.8287</td>
</tr>
<tr>
<td>MIX</td>
<td><strong>0.0021</strong></td>
<td><strong>0.4131</strong></td>
<td><strong>0.8486</strong></td>
</tr>
</tbody>
</table>

- Scores computed at simulated causal
- 100,000 SNPs, 1,000 cases, 1,000 controls over 6209 AA CARe individuals
- Power at P<5e-08
Genotype Imputation in African Americans

- Many methods for genotype imputation: MACH[Li et al. ‘09], IMPUTE[Howie et al’09], Beagle[Browning&Browning’09]...

- Extensions to African-Americans
  - Use CEU+YRI (w/o weights) as reference panel [HapMap3 ‘10], [Shriner et al ’10],[Palmer&Hirschhorn unpublished]
  - Joint imputation and local ancestry GEDI-ADMX  [Pasaniuc et al ’09]
  - Coalescent based local haplotype weighting [Pasaniuc et al ’10]

- Use local ancestry as guide:
  - 2 Eur ➔ only CEU as reference
  - 0 Eur ➔ only YRI as reference
  - 1 Eur ➔ both YRI +CEU haplotypes

Reference panel 1

Reference panel 2

Haplotype from admixed individual

Identify ancestry breakpoints

Perform locally optimal imputation

Locally imputed segments

Assemble imputed haplotype

Nature Reviews | Genetics
Rosenberg et al. Nature Reviews Genetics 2010
Accounting for local ancestry improves MACH imputation in African Americans from CARe
MIX outperforms other scores in the presence of imputation

Increase in power at R=1.5 of:
- 11% (97%, delta>0.4) for MIX over ATT
- 6% MIX over ATT-dose

<table>
<thead>
<tr>
<th></th>
<th>R=1.2 random</th>
<th>R=1.2 Δ&gt;0.4</th>
<th>R=1.5 random</th>
<th>R=1.5 Δ&gt;0.4</th>
<th>R=2.0 random</th>
<th>R=2.0 Δ&gt;0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATT</td>
<td>0.001</td>
<td>0.0008</td>
<td>0.2871</td>
<td>0.2988</td>
<td>0.762</td>
<td>0.7762</td>
</tr>
<tr>
<td>ATT-dose</td>
<td>0.001</td>
<td>0.0008</td>
<td>0.3009</td>
<td>0.3134</td>
<td>0.7775</td>
<td>0.7938</td>
</tr>
<tr>
<td>SUM</td>
<td>0.0007</td>
<td>0.002</td>
<td>0.2668</td>
<td>0.5086</td>
<td>0.7567</td>
<td>0.9729</td>
</tr>
<tr>
<td>MIX</td>
<td><strong>0.0013</strong></td>
<td><strong>0.0034</strong></td>
<td><strong>0.3184</strong></td>
<td><strong>0.5915</strong></td>
<td><strong>0.778</strong></td>
<td><strong>0.9786</strong></td>
</tr>
</tbody>
</table>

- Scores computed at imputed simulated causal
- 100,000 SNPs, 1,000 cases, 1,000 controls over 6209 AA CARe individuals
- Power at P<5e-08 for all scores except ADM (P<1e-05)
Causal SNP absent: not typed or imputed

- Maximum statistic in a window of 40 SNPs centered on simulated causal
- R=1.5
- Bolded results denote results obtained by removing the simulated causal

<table>
<thead>
<tr>
<th>Score</th>
<th>Proportion of regions that are genome wide significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATT</td>
<td>0.3834 0.1752</td>
</tr>
<tr>
<td>SUM</td>
<td>0.3571 0.1675</td>
</tr>
<tr>
<td>MIX</td>
<td>0.4158 0.1988</td>
</tr>
</tbody>
</table>
Coronary heart disease (CHD) and type 2 diabetes (T2D)

We evaluated the scores at loci previously associated to these phenotypes

CARe is a cohort study

- 929 cases, 4150 controls for CHD
- 179 cases, 3328 controls for T2D

Small number of cases ➔ potential advantage from case-only admixture signal marginal
MIX and ATT show similar performances on real CARe phenotypes

Data from CARe Consortium
FGFR2 locus in a breast cancer GWAS in African Americans

- 3153 cases and 2831 controls (C. Haiman et al.)
- FGFR2 shown to be associated to breast cancer [Hunter et al, NG ‘07]
- Fine-mapping efforts point to SNP rs2981578 as being the most likely causal [Udler et al HMG ‘10]
- Imputed all HapMap2 SNPs [CEU+YRI]
- Applied scores to all SNPs within 100kb of rs2981578
MIX(SUM) outperforms ATT at FGFR2 locus in a breast cancer GWAS in African Americans

Data from AABC study [C. Haiman&G.Chen et al.]
Recommendations for GWAS in African Americans

- Admixture and case-control signals ➔ improved power for association

- MIX: single causal with same effect across populations
  - Various scenarios (multiple causal variants, heterogeneous effects) may lead to suboptimal performance

- MIX performs well when causal imputed or not present in data

- Local-ancestry aware framework independent of imputation method

- Continuous phenotypes: Armitage Trend Test

- Fine-mapping for causal variants
  - SUM-MIX: test whether the admixture signal is fully explained by SNP OR
  - Beta release: http://www.hsph.harvard.edu/faculty/alkes-price/software/
Acknowledgments

- Harvard School of Public Health & Broad Institute
  - A. Price, S. Pollack, N. Zaitlen

- Harvard Medical School & Broad Institute
  - D. Reich, N. Patterson, A. Tandon

- Oxford University
  - S. Myers

- CARe Consortium
  - G. Lettre, J. Wilson et al.

- African American Breast Cancer data
  - C. Haiman, G. Chen et al.