Inferring Ancestry in Admixed Populations using Microarray Probe Intensities

Chen-Ping Fu, Catherine E. Welsh, Fernando Pardo-Manuel de Villena, Leonard McMillan

University of North Carolina at Chapel Hill

ACM-BCB '12, October 8, 2012
Ancestry Inference
Existing Methods: Ancestry Inference w/ Biallelic SNPs
Biallelic SNPs from Genotyping Arrays
Converting Fluorescence into Genotype Calls

B allele (1)

H allele (2)

A allele (0)
Problems with Genotype-based Ancestry Inference

N calls → marker discarded from analysis

Erroneous calls → wrong ancestry inference

Unexpected variation → unexploited useful information
Our Data

- Samples are from the Collaborative Cross (CC)
  - 8 inbred founders
  - Various stages of inbreeding
- Genotyped on the Mouse Universal Genotyping Array (MUGA)
  - 7,854 markers
  - Illumina Infinium platform
  - Designed to discriminate between CC founders
Our approach – use Intensities, not Genotypes
Cluster Similar Strains

- 8-9 replicates of each inbred founder
  - All replicates of the same founder cluster together
- pool together founders that fall in the same cluster
  - Determined by Hotelling's T-squared test with $p \leq 0.001$
- Store cluster means and covariances as homozygous clusters for each SNP
Create Heterozygous Clusters

- Only have 2-4 samples for each of the $\binom{8}{2} = 28$ possible F1 combinations
- Pool together F1s of all founders between pairs of homozygous clusters
- Store cluster means and covariances as heterozygous clusters for each SNP
Problem Statement

• Given:

$m$ possible inbred ancestors generating $m'$ ancestry states per marker, where $m' = m + \binom{m}{2}$. Call this state space $F$.

array with $n$ markers arranged in genomic order

target strain's 2D intensities $x_1...x_i...x_n$ for every marker, where $x_i$ is the 2D intensity at marker $i$

cluster means and covariances for each state in $F$ at every marker

Note: $m' \geq$ number of clusters at each marker (different ancestors may fall within the same cluster)

• Find:

sequence of most likely ancestry states $\{f_1, f_2 \ldots f_i \ldots f_n\}$ at every marker, where $f$ is one of $m'$ states in $F$
Distance Model

- Find the set of ancestor intensities closest to the target sample's intensities across the genome, without excessive transition between ancestor states.

- At each marker, use Mahalanobis Distance
  \[ D_M(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)} \]

  as distance measure from the target intensity \( x \) to each ancestor cluster with mean \( \mu \) and covariance \( S \).

- Over each chromosome, choose \( \{f_1, f_2 \ldots f_i \ldots f_n\} \), \( f \in F \) so that
  \[ D_M(x_1, \text{cluster}(f_1, 1)) + \sum_{i=2}^{n} D_M(x_i, \text{cluster}(f_i, i)) + \text{penalty}(f_{i-1}, f_i) \]
  is minimized,

  where \( D_M(x_i, \text{cluster}(f_i, i)) \) is distance from the target's intensity to state \( f_i \)'s intensity cluster at marker \( i \),

  and \( \text{penalty}(f_{i-1}, f_i) \) is the transition penalty between the ancestry states at markers \( i \) and \( i-1 \).
Dynamic Programming Recurrence

\[ \text{dist}_{f_i=p, f_{i+1}=q} = D_M(x_{i+1}, \text{cluster}(q, i + 1)) + \text{penalty}(p, q) \]
\[ + \min\{\text{dist}_{f_0=r, f_i=p} | \forall r \in F\}, \quad p, q \in F \]

Transition penalties given by the following table:

<table>
<thead>
<tr>
<th>p is homozygous</th>
<th>q is homozygous</th>
<th>p and q share a haplotype</th>
<th>Graphical depiction</th>
<th>penalty(p,q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td></td>
<td>mean $D_M$ between different homozygous clusters</td>
</tr>
<tr>
<td>yes/no</td>
<td>no/yes</td>
<td>yes</td>
<td></td>
<td>1.5* mean $D_M$ between homozygous and heterozygous clusters</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>yes</td>
<td></td>
<td>1.5* mean $D_M$ between different heterozygous clusters</td>
</tr>
<tr>
<td>yes/no</td>
<td>no/yes</td>
<td>no</td>
<td></td>
<td>5.0*mean $D_M$ between homozygous and heterozygous clusters</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>no</td>
<td></td>
<td>5.0*mean $D_M$ between different heterozygous clusters</td>
</tr>
</tbody>
</table>
Results

- We chose to compare with GAIN, a genotype-based inference algorithm designed for the CC
  - We had 6,750 informative markers (GAIN had 5,782)
  - 5,550 markers with 2 homozygous clusters, 1,200 markers with 3 or more homozygous clusters
  - 2.21 homozygous clusters/marker (genotype calls provide 2 – A, B)
  - 3.66 total clusters/marker (genotype calls provide 3 – A, B, H)
Results

- Used whole-genome sequence data for verification
  - DNA sequence data available for 3 CC samples genotyped on MUGA
  - Ran our algorithm and GAIN on these 3 CC samples, then imputed SNPs using the Wellcome Trust's whole-genome sequences
  - When inference between us and GAIN differ, compare all imputed SNPs in the region with sequence data

<table>
<thead>
<tr>
<th>Sample</th>
<th># SNPs where we can GAN differ</th>
<th>SNPs where we agree with sequence</th>
<th>SNPs where GAIN agrees with sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR867m532</td>
<td>33,026</td>
<td>24,092</td>
<td>8,934</td>
</tr>
<tr>
<td>OR1237m224</td>
<td>17,536</td>
<td>14,524</td>
<td>3,011</td>
</tr>
<tr>
<td>OR3067m352</td>
<td>38,621</td>
<td>23,095</td>
<td>15,526</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>89,183</strong></td>
<td><strong>52,144 (69.2%)</strong></td>
<td><strong>27,471 (30.8%)</strong></td>
</tr>
</tbody>
</table>
We can refine breakpoints better
Results – Ancestry Inference

GAIN makes spurious transitions due to erroneous genotype calls, a problem which does not occur in our method.
Conclusions

- We considered other distance measures – Euclidean, Manhattan, etc.
  - Mahalanobis distance most robust, but other distances useful when multiple replicates of ancestors are not available
- We applied our methods to different platforms and populations and found comparable results
- We will extend our model to an HMM – give a vector of probabilities at each marker
- Fluorescence intensity ranges vary between markers → we can move to a per-marker penalty model
- We should explore intensity-based methods for other applications (detecting structural variants, sexing, etc.)
THANK YOU!