Genotype Imputation

Biostatistics 666
Previously

• Hidden Markov Models for Relative Pairs
  – Linkage analysis using affected sibling pairs
  – Estimation of pairwise relationships

• Identity-by-Descent
  – Relatives share long stretches of chromosome
  – Sharing at some markers can be used as surrogate for sharing at unobserved markers
Today

• Genotype Imputation / “In Silico” Genotyping
  – Use genotypes at a few markers to infer genotypes at other unobserved markers

• Closely related individuals
  – Long segments of identity by descent

• Distantly related individuals
  – Shorter segments of identity by descent
Given the above pedigree, what are the likely values of the genotype marked ?/? ...?
In Silico Genotyping For Family Samples

• Family members will share large segments of chromosomes

• If we genotype many related individuals, we will effectively be genotyping a few chromosomes many times

• In fact, we can:
  – Genotype a few markers on all individuals
  – Identify shared segments of haplotypes
  – Genotype additional markers on a subset of individuals
  – Fill in missing genotypes that fall in shared segments

  – Even without information on shared segments, it may be possible to learn about genotypes of relative members
Genotype Inference
Part 1 – Observed Genotype Data
Genotype Inference
Part 2 – Inferring Allele Sharing
Genotype Inference
Part 3 – Imputing Missing Genotypes
Genotype Imputation in Families

- Suppose a particular genotype \( g_{ij} \) is missing
  - Genotype for person \( i \) at marker \( j \)

- Consider full set of observed genotypes \( G \)

- Evaluate pedigree likelihood \( L \) for each combination of \( \{G, g_{ij} = x\} \)

- Posterior probability that \( g_{ij} = x \) is

\[
P(g_{ij} = x | G) = \frac{L(G, g_{ij} = x)}{L(G)}
\]

- For pairs, same HMM as for linkage analysis or checking relatedness.
Standard Linear Model for Genetic Association

- Model association using a model such as:

  \[ E(y_i) = \mu + \beta_g g_i + \beta_c c_i + \cdots \]

- \( y_i \) is the phenotype for individual \( i \)
- \( g_i \) is the genotype for individual \( i \)
  - Simplest coding is to set \( g_i = \) number of copies of the first allele
- \( c_i \) is a covariate for individual \( i \)
  - Covariates could be estimated ancestry, environmental factors...

- \( \beta \) coefficients are estimated covariate, genotype effects
- Model is fitted in variance component framework
Model With Inferred Genotypes

• Replace genotype score $g$ with its expected value:

$$E(y_i) = \mu + \beta_g \bar{g} + \beta_c c + \cdots$$

• Where $\bar{g}_i = 2P(g_i = 2|G) + P(g_i = 1|G)$

• Association test can then be implemented in variance component framework, just as before

• Alternatives would be to
  – (a) impute genotypes with large posterior probabilities; or
  – (b) integrate joint distribution of unobserved genotypes in family
Example I

- Assumptions:
  - Two alleles per marker
  - Equal allele frequencies
  - $\Theta = 0$

- $L(G) = .0061$
- $L(G, g_{22} = 1/1) = .00494$
- $L(G, g_{22} = 1/2) = .00110$
- $L(G, g_{22} = 2/2) = .00006$

- $P(g_{22} = 1/1 | G) = 0.81$
- $P(g_{22} = 1/2 | G) = 0.18$
- $P(g_{22} = 2/2 | G) = 0.01$
- $\bar{g} = 1.80$
Example II

• Assumptions:
  – Two alleles per marker
  – Equal allele frequencies
  – Θ = 0

• $L(G) = 0.000244$
• $L(G, g_{22} = 1/1) = 0.000061$
• $L(G, g_{22} = 1/2) = 0.000122$
• $L(G, g_{22} = 2/2) = 0.000061$

• $P(g_{22} = 1/1 | G) = 0.25$
• $P(g_{22} = 1/2 | G) = 0.50$
• $P(g_{22} = 2/2 | G) = 0.25$
• $\bar{g} = 1.00$
Example III

• Assumptions:
  – Two alleles per marker
  – Equal allele frequencies
  – \( \Theta = 0.10 \)

• Assumptions:
  – Two alleles per marker
  – Equal allele frequencies
  – \( \Theta = 0.10 \)

• \( L(G) = .0054 \)
• \( L(G, g_{22} = 1/1) = .00392 \)
• \( L(G, g_{22} = 1/2) = .00136 \)
• \( L(G, g_{22} = 2/2) = .00012 \)

• \( P(g_{22} = 1/1|G) = 0.73 \)
• \( P(g_{22} = 1/2|G) = 0.25 \)
• \( P(g_{22} = 2/2|G) = 0.02 \)
  
• \( \bar{g} = 1.70 \)
Example IV

Assumptions:
- Two alleles per marker
- Equal allele frequencies
- $\Theta = 0.10$

- $L(G) = .000121$
- $L(G, g_{22} = 1/1) = .000033$
- $L(G, g_{22} = 1/2) = .000061$
- $L(G, g_{22} = 2/2) = .000028$

- $P(g_{22} = 1/1 | G) = 0.273$
- $P(g_{22} = 1/2 | G) = 0.499$
- $P(g_{22} = 2/2 | G) = 0.227$
- $\bar{g} = 1.05$
Power in Sibships of Size 6
Without Parental Genotype Data

T is the number of genotyped offspring.
QTL explains 5% of variance, polygenes explain 35%,
250 sibships, α = 0.001.
Application: Gene Expression Data

• Cheung et al (2005) carried out a genome wide association with 27 expression levels as traits

• Measured in grandparents and parents of CEPH pedigrees and took advantage of HapMap I genotypes

• SNP consortium genotypes also available for ~6000 SNPs in the offspring of each CEPH family
Example: Gene Expression Data

- Panels show GWA scan with CTBP1 expression as outcome
  - Gene is at start of chromosome 4

- Using observed genotypes, most significant association maps in \textit{cis} for 15/27 traits
  - 12 of these reach $p < 5 \times 10^{-8}$

- Using inferred genotypes, most significant association maps in \textit{cis} for 19/27 traits
  - 15 of these reach $p < 5 \times 10^{-8}$

- Data from Cheung et al. (2005)
Point of Situation...

• When analyzing family samples ...

• FOR INDIVIDUALS WITH KNOWN RELATIONSHIPS
  – Impute genotypes in relatives
  – Imputation works through long shared stretches of chromosome

• But the majority of GWAS that use “unrelated” individuals...

• FOR INDIVIDUALS WITH UNKNOWN RELATIONSHIPS
  – Impute observed genotypes in relatives
  – Imputation works through short shared stretches of chromosome
In Silico Genotyping For Unrelated Individuals

• In families, long stretches of shared chromosome

• In unrelated individuals, shared stretches are much shorter

• The plan is still to identify stretches of shared chromosome between individuals...

• ... we then infer intervening genotypes by contrasting samples typing at a few sites with those with denser genotypes
Observed Genotypes

Reference Haplotypes

Study Sample

HapMap
Identify Match Among Reference

Observed Genotypes


Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T T A T G C
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A A G A T C T C C C G A A C C T T G T G C
C G A A G A C T C T T T T C T T T T G T A C
C G A A G A C T C T C C G A C C T C G T G C
C G A A G A C T C C T T T T T C T T C G T G C
C G A A G A T C T C T T T T T C T T C G T G C
C G A A G A C T C T T T T T C T T C G T G C
C G A A G C T C T T T T C T T C G T G C
C G A A G C T C T T T T T C T T C G T G C
C G A A G C T C T T T T T C T T C G T G C
Phase Chromosome, Impute Missing Genotypes

Observed Genotypes

c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t t c A t g g

Reference Haplotypes

C G A G A T C T C C T T C T T C T T C T G T G C
C G A G A T C T C C T T C T T C T T C T G T G C
C C A A G C T C T T T T C T T C T T C T G T G C
C G A A G C T C T T T T C T T C T T C T G T G C
C G A A G C T C T C C G A A C C T T A T G C
C G A G A C T C T C C G A A C C T T A T G C
T G G G G A T C T C C C C G A C C T C A T G G
C G A G A T C T C C C C G A C C T T G T G C
C G A G A T C T C C C C G A C C T T G T G C
C G A G A T C T C C C C G A C C T T G T G C
C G A G A T C T C C C C G A C C T T G T G C
C G A A G C T C T T T T T C T T C T G T G C
Implementation

- Markov model is used to model each haplotype, conditional on all others

- At each position, we assume that the haplotype being modeled copies a template haplotype

- Each individual has two haplotypes, and therefore copies two template haplotypes
The final ingredient connects template states along the chromosome …
Possible States

• A state $S$ selects pair of template haplotypes
  – Consider $S_i$ as vector with two elements $(S_{i,1}, S_{i,2})$

• With $H$ possible haplotypes, $H^2$ possible states
  – $H(H+1)/2$ of these are distinct

• A recombination rate parameter describes probability of switches between states
  – $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a, S_{i+1,2} = b))$ \hspace{1cm} (1-\theta)^2$
  – $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a^*, S_{i+1,2} = b))$ \hspace{1cm} (1-\theta)\theta/H$
  – $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a^*, S_{i+1,2} = b^*))$ \hspace{1cm} (\theta/H)^2$
Emission Probabilities

• Each value of $S$ implies expected pair of alleles

• Emission probabilities will be higher when observed genotype matches expected alleles

• Emission probabilities will be lower when alleles mismatch

• Let $T(S)$ be a function that provides expected allele pairs for each state $S$
Emission Probabilities

\[ P(G_j|S_j) = \begin{cases} 
(1-\varepsilon_j)^2 + \varepsilon_j^2, & T(S_j) = G_j \text{ and } G_j \text{ is heterozygote,} \\
2(1-\varepsilon_j)\varepsilon_j, & T(S_j) \neq G_j \text{ and } G_j \text{ is heterozygote,} \\
(1-\varepsilon_j)^2, & T(S_j) = G_j \text{ and } G_j \text{ is homozygote,} \\
(1-\varepsilon_j)\varepsilon, & T(S_j) \text{ is heterozygote and } G_j \text{ homozygote,} \\
\varepsilon_j^2, & T(S_j) \text{ and } G_j \text{ are opposite homozygotes.} 
\end{cases} \]
Does This Really Work?

Preliminary Results

- Used 11 tag SNPs to predict 84 SNPs in CFH
- Predicted genotypes differ from original ~1.8% of the time
- Reasonably similar results possible using various haplotyping methods

Comparison of Test Statistics, Truth vs. Imputed
Does This Really Work?

- Used about ~300,000 SNPs from Illumina HumanHap300 to impute 2.1M HapMap SNPs in 2500 individuals from a study of type II diabetes

- Compared imputed genotypes with actual experimental genotypes in a candidate region on chromosome 14
  - 1190 individuals, 521 markers not on Illumina chip

- Results of comparison
  - Average $r^2$ with true genotypes 0.92 (median 0.97)
  - 1.4% of imputed alleles mismatch original
  - 2.8% of imputed genotypes mismatch
  - Most errors concentrated on worst 3% of SNPs

Scott et al, Science, 2007
Does this really, really work?

• 90 GAIN psoriasis study samples were re-genotyped for 906,600 SNPs using the Affymetrix 6.0 chip.

• Comparison of 15,844,334 genotypes for 218,039 SNPs that overlap between the Perlegen and Affymetrix chips resulted in discrepancy rate of 0.25% per genotype (0.12% per allele).

• Comparison of 57,747,244 imputed and experimentally derived genotypes for 661,881 non-Perlegen SNPs present in the Affymetrix 6.0 array resulted in a discrepancy rate of 1.80% per genotype (0.91% per allele).

• Overall, the average $r^2$ between imputed genotypes and their experimental counterparts was 0.93. This statistic exceeded 0.80 for >90% of SNPs.

LDLR and LDL example

## Impact of HapMap Imputation on Power

Simulations ensure equal power for directly genotyped SNPs.

<table>
<thead>
<tr>
<th>Disease SNP MAF</th>
<th>tagSNPs</th>
<th>Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5%</td>
<td>24.4%</td>
<td>56.2%</td>
</tr>
<tr>
<td>5%</td>
<td>55.8%</td>
<td>73.8%</td>
</tr>
<tr>
<td>10%</td>
<td>77.4%</td>
<td>87.2%</td>
</tr>
<tr>
<td>20%</td>
<td>85.6%</td>
<td>92.0%</td>
</tr>
<tr>
<td>50%</td>
<td>93.0%</td>
<td>96.0%</td>
</tr>
</tbody>
</table>

Power for simulated case control studies. Simulations ensure equal power for directly genotyped SNPs.

Simulated studies used a tag SNP panel that captures 80% of common variants with pairwise $r^2 > 0.80$. 

Combined Lipid Scans

• SardiNIA (Schlessinger, Uda, et al.)
  – ~4,300 individuals, cohort study
• FUSION (Mohlke, Boehnke, Collins, et al.)
  – ~2,500 individuals, case-control study of type 2 diabetes
• DGI (Kathiresan, Altshuler, Orho-Mellander, et al.)
  – ~3,000 individuals, case-control study of type 2 diabetes

• Individually, 1-3 hits/scan, mostly known loci

• Analysis:
  – Impute genotypes so that all scans are analyzed at the same “SNPs”
  – Carry out meta-analysis of results across scans

Willer et al, Nature Genetics, 2008
Combined Lipid Scan Results
18 clear loci!

New LDL Locus, Previously Associated with CAD
**Comparison with Related Traits:**

**Coronary Artery Disease and LDL-C Alleles**

<table>
<thead>
<tr>
<th>Gene</th>
<th>LDL-C p-value</th>
<th>Frequency CAD cases</th>
<th>Frequency CAD ctrls</th>
<th>CAD p-value</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE/C1/C4</td>
<td>3.0x10^-43</td>
<td>.209</td>
<td>.184</td>
<td>1.0x10^-4</td>
<td>1.17 (1.08-1.28)</td>
</tr>
<tr>
<td>APOE/C1/C4</td>
<td>1.2x10^-9</td>
<td>.339</td>
<td>.319</td>
<td>.0068</td>
<td>1.10 (1.02-1.18)</td>
</tr>
<tr>
<td>SORT1</td>
<td>6.1x10^-33</td>
<td>.808</td>
<td>.778</td>
<td>1.3x10^-5</td>
<td>1.20 (1.10-1.31)</td>
</tr>
<tr>
<td>LDLR</td>
<td>4.2x10^-26</td>
<td>.902</td>
<td>.890</td>
<td>6.7x10^-4</td>
<td>1.29 (1.10-1.52)</td>
</tr>
<tr>
<td>APOB</td>
<td>5.6x10^-22</td>
<td>.830</td>
<td>.824</td>
<td>.18</td>
<td>1.04 (0.95-1.14)</td>
</tr>
<tr>
<td>APOB</td>
<td>8.3x10^-12</td>
<td>.353</td>
<td>.332</td>
<td>.0042</td>
<td>1.10 (1.03-1.18)</td>
</tr>
<tr>
<td>APOB</td>
<td>3.1x10^-9</td>
<td>.536</td>
<td>.520</td>
<td>.028</td>
<td>1.07 (1.00-1.14)</td>
</tr>
<tr>
<td>PCSK9</td>
<td>3.5x10^-11</td>
<td>.825</td>
<td>.807</td>
<td>.0042</td>
<td>1.13 (1.03-1.23)</td>
</tr>
<tr>
<td>NCAN/CILP2</td>
<td>2.7x10^-9</td>
<td>.922</td>
<td>.915</td>
<td>.055</td>
<td>1.11 (0.98-1.26)</td>
</tr>
<tr>
<td>B3GALT4</td>
<td>5.1x10^-8</td>
<td>.399</td>
<td>.385</td>
<td>.039</td>
<td>1.07 (0.99-1.14)</td>
</tr>
<tr>
<td>B4GALT4</td>
<td>1.0x10^-6</td>
<td>.874</td>
<td>.865</td>
<td>.051</td>
<td>1.09 (0.98-1.20)</td>
</tr>
</tbody>
</table>

Comparison to data from WTCCC (Nature, 2007) was made possible by imputation.
Does This Work Across Populations?

- Conrad et al. (2006) dataset
- 52 regions, each ~330 kb
- Human Genome Diversity Panel
  - ~927 individuals, 52 populations
- 1864 SNPs
  - Grid of 872 SNPs used as tags
  - Predicted genotypes for the other 992 SNPs
  - Compared predictions to actual genotypes

Tag SNP Portability
(Evaluation Using ~1 SNP per 10kb in 52 x 300kb regions For Imputation)
Summary

• Genotype imputation can be used to accurately estimate missing genotypes

• Genotype imputation is usually implemented through using a Hidden Markov Model

• Benefits of genotype imputation
  – Increases power of genetic association studies
  – Facilitates analyses that combine data across studies
  – Facilitates interpretation of results
2017 Imputation Accuracy: Europeans (Complete Genomics as Truth)
Imputation
https://imputationserver.sph.umich.edu
Recommended Reading


• Li et al (2010) Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**:816-834