Imputation & Meta-analysis

Thomas Nichols, PhD
Department of Statistics & Warwick Manufacturing Group
University of Warwick

Big (Slide) Thanks to

Trygve Bakken, PhD
Allen Brain Institute

Goncalo Abecasis, PhD
University of Michigan

Sarah Medland, PhD
QMRI

Jeff Barrett, PhD
Wellcome Sanger Institute

OHBM 2013 - Introduction to Imaging Genetics - 8 June, 2014
Combining Genetics Data

• Need more and more data

• To maximize power, generalizability

• Imaging studies…

• 4-digit sample sizes rare! (But… UK Biobank’s will have 100,000 subjects!)

• Need to combine lots and lots of studies to get sufficient power

ENIGMA study sample sizes (7,795 total n)

78 104
220 221 249 279 327
419 442 485 518 550 558
747
800 871 927
### Combining Genetic Data

- **Problem:** Data missing due to imperfect calling
- **Problem:** Data on different sets of SNPs
- **Solution:** Imputation

<table>
<thead>
<tr>
<th>Subject</th>
<th>SNP 1</th>
<th>SNP 2</th>
<th>SNP 3</th>
<th>SNP 4</th>
<th>SNP 5</th>
<th>SNP 6</th>
<th>SNP 7</th>
<th>SNP 8</th>
<th>SNP 9</th>
<th>SNP 10</th>
<th>SNP 11</th>
<th>SNP 12</th>
<th>SNP 13</th>
<th>SNP 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 7</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 9</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>SNP 1</th>
<th>SNP 2</th>
<th>SNP 3</th>
<th>SNP 4</th>
<th>SNP 5</th>
<th>SNP 6</th>
<th>SNP 7</th>
<th>SNP 8</th>
<th>SNP 9</th>
<th>SNP 10</th>
<th>SNP 11</th>
<th>SNP 12</th>
<th>SNP 13</th>
<th>SNP 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 7</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subj 9</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

• Imputation
  • Haplotype Background
  • Statistics of Imputation

• Meta-Analysis
  • Fixed effects vs. random effects
  • Heterogeneity scores
  • SE weighted vs N weighted
Mitosis separates replicated chromosomes before cell division

- Replicated chromosomes are divided into 2 diploid daughter cells

- Nondisjunction $\rightarrow$ aneuploidy $\rightarrow$ genomic mosaicism
  
  \[\text{chr. fails to split} \quad \rightarrow \quad \text{wrong # of chr.} \quad \rightarrow \quad \text{different cells, different DNA!}\]

- E.g. 4% of healthy neurons have an abnormal number of chr 21

Rehen, JoN, 2005
Meiosis generates haploid cells

- DNA replication and 2 rounds of cell division
- Homologous recombination creates genetic diversity
- 4 haploid daughter cells (sperm or eggs) have a unique set of chromosomes with DNA from both parents
Review: Genetic recombination during meiosis

- Homologous chromosomes align
- Double-stranded DNA break
- DNA from one chromosome ‘crosses over’ and invades sister chromosome
- Enzymes resolve ‘Holliday junction’
Recombination hot spots

- Recombination rate varies across the genome
- Most crossover events cluster into short regions
- DNA sequence motifs contribute to locations of hot spots
Single nucleotide polymorphisms (SNPs)

- DNA sequence variation at a single nucleotide
- Any 2 human individuals will differ at about 1 out of 1000 bases = 3–4 million differences
- SNPs vary between human populations based on ancestry
Hot spots lead to linkage disequilibrium

ACGATCGATGCACGATCGATCGTAGCTAGCCGTATCGTAGCTACGTAGC

Reference Sequence

Person A

10 generations

Mutation!

Recombination hot spots
Hot spots lead to linkage disequilibrium

100–1000s of generations
Linked SNPs are proxies for a causative allele

- **Linkage Disequilibrium**: Non-random association of alleles. They are seen on the same chromosome more frequently than you would expect by chance.

- **Haplotype Block**: A combination of genetic variants which are transmitted together.
Visualization of haplotype blocks

- Red = regions of strong LD; white = little or no LD
- Haplotype blocks are on average 5–20 kilobases long
Haplotype Phasing (1)

• Genotypes don’t tell us how to reconstruct DNA strands!

Observe genotypes at 3 loci

<table>
<thead>
<tr>
<th>Observe genotypes at 3 loci</th>
<th>Have this chromosome pair?</th>
<th>Or this pair?</th>
<th>Or this pair?</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>A</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>CG</td>
<td>C</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>GT</td>
<td>G</td>
<td>T</td>
<td>G</td>
</tr>
</tbody>
</table>

Have this chromosome pair?

Maternal | Paternal

Or this pair?

Maternal | Paternal

Or this pair?

Maternal | Paternal

Observe genotypes at 3 loci

Have this chromosome pair?

Maternal | Paternal

Or this pair?

Maternal | Paternal

Or this pair?

Maternal | Paternal
Haplotype Phasing (2)

• K SNPs have $2^K$ possible haplotypes
  – Here $2^3 = 8$ haplotypes possible

```
A  T  A  T
C  C  G  G
T  T  T  T
A  T  A  T
C  C  G  G
G  G  G  G
```
Haplotype Phasing (3)

- Frequency in population, of course will vary, e.g. ...

18%  
A  
C  
T  

5%  
T  
C  
T  

16%  
A  
G  
T  

25%  
T  
G  
T  

24%  
A  
C  
G  

5%  
T  
C  
G  

0%  
A  
G  
G  

5%  
T  
G  
G
Haplotype Phasing (4)

- Phasing uses LD structure and Haplotype panels to infer
  - Note, the more homogzygosity, the easier it is to phase
Imputation: Haplotype panels

• Haplotype maps
  – Sets of carefully phased haplotypes used as reference

HapMap
30 YRI trios, Yoruba from Ibadan, Nigeria,
30 CEU trios, Utah residents, northern and western European ancestry
44 JPT unrelated Japanese, Tokyo, Japan
45 CHB unrelated Han Chinese, Beijing, China

1000 Genomes Project
1092 individuals, greater coverage
ASW Southwestern US, African ancestry
JPT Japanese in Tokyo
CHB Chinese in Beijing
CHD Chinese in metropolitan Denver
CEU Utah residents, northern and western European ancestry
GIH Gujarati Indians in Houston
LWK Luhya in Webuye, Kenya
MKK Maasai in Kinyawa, Kenya
MXL Los Angeles, Mexican ancestry
PEL Peruvians in Lima, Peru
TSI Toscani in Italy
YRI Yoruba in Ibadan, Nigeria
Imputation

Genotypes with missing data

Matches to panel found (for each chromosome)

Imputed Genotype

Imputation

- In practice, uncertainty on the right haplotypes
- “Genotype” is then probability on \{0,1,2\}, or expected count
- Don’t round! Analysis software uses this information

Imputation: Challenges

• Strand alignment
  – DNA has 2 stands
    • 50% coded forward strand, 50% code on reverse strand
  – All genotypes read off relative to one strand
  – Different platforms use different stands
  – Must start by ensuring all datasets are using same strand, converting when needed
Imputation: Quality

• Imputation doesn’t always work perfectly
• But we have quality scores to warn us
  – Software dependent, but range [0,1]
  – Each attempt to quantify the amount of information in the imputed SNP
• SNPs with low quality scores are omitted
Imputation: Analysis

• Optimal analysis
  – Make use of imputed probability of each genotype
  – But is slow

• Practical Analysis
  – Use expected allele count in usual association/regression
  – Gives good approximation to the slower, fancy method
Methods for Meta-Analysis

• T/Z/P-value based
  – Just uses P-value & sample size from each analysis
  – Not recommended
    • Based on significance, not effects in real units

• Estimate based
  – Requires effect estimates and standard errors in real-data units
    • E.g. change in GM per allelic dose
  – But requires all studies to have consistent units
    • E.g. difficult with BOLD fMRI

Meta-Analysis Methods: Mechanics

For study $i$: $n_i$ sample size, $Z_i$ Z-score (e.g. converted from P-value), $\hat{\beta}_i$ effect estimate, $SE_i$ effect standard error

<table>
<thead>
<tr>
<th>Weights</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_i = \sqrt{n_i}$</td>
<td>$\hat{\beta}_{meta} = \frac{\sum_i w_i \hat{\beta}_i}{\sqrt{\sum_i w_i}}$</td>
<td>$SE_{meta} = \sqrt{1/\sum_i w_i}$</td>
<td>$Z_{meta} = \frac{\hat{\beta}<em>{meta}}{SE</em>{meta}}$</td>
</tr>
<tr>
<td>T/P/Z-based Sample-Size Weighted</td>
<td>Estimate-based Inverse SE Weighted</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Weights

Estimate

Standard Error

Test
Meta-Analysis: Issues

• Which method: Use Estimate-Based if you can
  – Two methods become equivalent when noise variance homogeneous over sites
    • Then SE’s scale with sample size

• Fixed vs random effects...
Fixed-effects vs. Random-effects

• Both these methods are fixed-effect
  – Significance based only intra-study variation

• Random-effects
  – Best practice, preferred
  – But requires enough studies to estimate between-study variation
Heterogeneity Tests

• $I^2$
  – Proportion of variance due to study-to-study variation
  – Ranges $[0,1]$, want it to be small
• Cochran’s Q test

$$Q = \sum_i w_i (\hat{\beta}_i - \hat{\beta}_{meta})^2$$
  – Based on sum-of-squares (about mean)
  – Grows large if variation is larger than due to intra-study variation alone
Conclusions

• Power power power
  – Need to combine as many samples as possible

• Imputation
  – Need to put all genetic data on common basis

• Meta-Analysis
  – Combines evidence for hits, considering precision at each study/site