Haplotype Analysis and Genotype Imputation

Xiaowei Zhan
April 28, 2017
Xiaowei.Zhan@UTSouthwestern.edu

(Slides courtesy of Goncalo Abecasis)
Concepts from yesterday’s lecture

• Haplotype and Linkage

Linkage Disequilibrium (LD):
Genotypes at two loci are not independent

\[
\text{Pr}(AB) \neq \text{Pr}(A) \text{ Pr}(B)
\]
Concepts from yesterday’s lecture

• Haplotype

Related individuals can share long stretches of sequences (haplotype)

from last lecture under the context of *Identity by Descent*
Today’s Outline: Haplotype and Imputation Analysis

- Haplotype analysis
  - Understand the motivation of haplotype analysis
  - Statistical method to infer haplotypes from genotype data
    - Clark’s Greedy algorithm
    - E-M algorithm
    - Hidden Markov Model (HMM)
  - Haplotype association analysis

- Imputation analysis
  - Understand the concept of imputation
  - How to impute genotypes from familial samples
  - How to imputation genotypes from unrelated individuals
    - Hidden Markov Model (HMM)
  - How to use imputed genotypes in association analysis
  - Examples of imputation analyses in GWAS studies
How haplotype analysis can be useful?

Assume we know the haplotype-level genetic data, how can haplotypes be useful?

• Linkage disequilibrium studies  
  (Recall how to calculate D, D’ and r^2)  
  Genetic variations in the haplotype level.  
  e.g. population specific haplotype

• Select markers to genotype  
  Select tag SNP based on haplotypes

• Candidate gene studies  
  Interpret association results  
  Capture the effect of ungenotyped alleles
Haplotype cannot be easily observed

• Biological measurement of haplotypes can be challenging
  
  X-chromosome in males
  Sperm typing
  Hybrid cell lines
  Other molecular techniques
  …

  We only observe genotype data, how to obtain haplotypes?

• *Statistical* approaches to infer haplotypes from genotypes
Observed genotypes $\Leftrightarrow$ possible haplotypes

- Two alleles for each individual  
  (Genotypes observed)

- Multiple haplotypes are compatible with observed genotype  
  (4 haplotype combinations)
Family information can be helpful

- From pedigree, we can phase many markers

- But still, there can be many ambiguities that cannot be resolved

Large number of markers $\Rightarrow$ less proportion of known haplotypes
When there are no relatives…

- Rely on linkage disequilibrium
- Assume the number of haplotypes in a population is small
- Haplotypes tend to be similar
Phasing algorithms

Several milestone methods to infer haplotypes

Clark’s haplotyping method

- Very computationally efficient
- Widely used in the 1990’s
- Clark’s Algorithm

**Figure:**

Inference of Haplotypes from PCR-amplified Samples of Diploid Populations

*Andrew G. Clark*

Department of Biology and Genetics Program, Pennsylvania State University

Start from unambiguous haplotypes

Sequentially resolve haplotypes

Randomly guess the rest haplotypes
Limitation: failed to start

- What kind of genotype/haplotype do we need to have to get started?

- What is the probability of failed start?

\[
\Pr(\text{failure}) \approx \left[1 - \frac{1}{1+\theta} - \frac{\theta}{(1+\theta)^2}\right]^n
\]

\[
\theta: \text{number of marker}
\]

\[
n: \text{sample size}
\]
Pro and Cons

• Andrew Clark’s method

  Very fast

  May failed to start with small sample size

  May leave unresolved haplotypes
E-M method

- E-M (expectation maximization)

  Capable to handle missing genotypes

  Consider allele frequencies

  When there is m unphased genotypes, there are $2^{m-1}$ possible haplotypes => computationally expensive (>25 markers)

*Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population*

*Laurent Excoffier* and *Montgomery Slatkin*†

*Departments of Anthropology and Ecology, University of Geneva and †Department of Integrative Biology, University of California, Berkeley*
Stephen’s method

- Improve the previous EM method by reusing similar haplotypes
- Consider *genealogical* information
Reuse similar haplotypes

- Individual 1: use known haplotypes
- Individual 2: re-use known haplotypes and allow mismatches
Similar haplotypes have more recent common ancestor
MCMC method with Gibbs sampler

• MCMC method can iteratively improve solutions
  1. Initialize haplotypes
  2. Sample haplotypes of one individual given other’s haplotypes
  3. Update the estimated haplotypes for one individual
  4. Repeat the above millions of times

• MCMC method will converge to an optimal solution
• The result is equivalent to EM algorithm
Stephen’s algorithm

- Improve the update step by incorporate genealogical information (coalescent theory)

\[
\Pr(h \mid H) = \sum_{\alpha} \sum_{S} \frac{n_{\alpha}}{n} \left( \frac{\theta}{n + \theta} \right)^{S} \frac{n}{n + \theta} \left( P^S \right)_{\alpha h}
\]
ShapeIT/MaCH software

- Based on Stephen’s model, modern phasing software optimizes computational efficiency
  - Blockwise computation (ShapeIT)
  - Hidden Markov Model (MaCH)
    - Markov haplotying
    - Same model can be easily adapted for imputation

MaCH: Li, Yun, et al. "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." *Genetic epidemiology* 34.8 (2010): 816-834. (cited by 1373)
Summary on haplotype inference

• Three classic statistical approaches to infer haplotypes from genotype data
  
  Clark’s greedy algorithm
  E-M algorithm
  Stephen’s genealogical approach

• Lab: practical workshop

  Phase one sample from HapMap3 project using ShapeIT
Association models for haplotype analysis

- Association tests
  are haplotype frequencies the same in two populations
e.g. case vs. control, population 1 vs. population 2

- The simplistic approach to compare haplotypes reconstructions
  - Calculate haplotype frequency for each group
  - Find mostly likely haplotype for each individual
  - Compare haplotype frequency between the two groups

Question: any caveat in this approach?

NOT RECOMMENDED!!!
Observe genotypes in CASE

The phase reconstruction in the five ambiguous individuals will be driven by the haplotypes observed in individual 1.
Inferred haplotypes for CASE

This kind of phenomenon will occur with nearly all population based haplotyping methods!
Observe genotypes for CONTROL

Note these are identical, except for the single homozygous individual
Inferred haplotypes for CONTRL

Ooops The difference in a single genotype in the original data has been greatly amplified by estimating haplotypes
Inferred haplotypes for CASE and CONTRL

Problematic to conclude haplotype frequencies differs between groups, although or frequencies differ.
Lesson learned

- Do NOT treat case/control in different haplotype inference procedures
- Treat case/control **together**
- Or use likelihood approaches

Estimated haplotype frequencies, imply a likelihood for the observed genotypes

\[
L = \prod_i \sum_{H \sim G_i} P(H)
\]

- individuals
- possible haplotype pairs, conditional on genotype
- haplotype pair frequency
Likelihood approaches

- Test two sets of models

  - Calculate 3 likelihoods:
    - Maximum likelihood for combined sample, $L_A$
    - Maximum likelihood for control sample, $L_B$
    - Maximum likelihood for case sample, $L_C$

$$2 \ln \left( \frac{L_B L_C}{L_A} \right) \sim \chi^2_{df}$$

$df$ is hard to obtain.
Use permutations.
Hypothesis testing

- Previously: test haplotype frequency between two populations
- Often, we want to test
  - Are haplotypes different between two populations?
- Note: this is different than single marker test
ACE gene example


• 10 di-allelic polymorphism
  Spanning 26k region
  Common markers
Haplotype analysis

- 3 ACE haplotype clades.
- Clade “B” = Clade “C” 
  Similar phenotypic effect
- Interpretations

Functional variants on the right

Think: if functional variants on the left, which two clades have similar phenotypes?
Regression models

- Predictor
  Haplotype counts

- Regression parameters
  Phenotypic effect of each haplotype

- Response
  Phenotype values
Exemplar design matrix

\[
E \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1/2 & 1/2 \\ 1 & 1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}
\]

Hypothetical set-up when observed haplotypes are:
- \( h_1/h_1 \) for individual 1
- \( h_2/h_3 \) for individual 2
- \( h_1/h_3 \) for individual 3

Zaykin et al, 2002
When haplotype is unknown

- Use Baye’s rule to calculate:

\[
\Pr(h_2, h_3 \mid G_i) = \frac{\Pr(G_i \mid h_2, h_3) p_{h_2} p_{h_3}}{\sum_{u,v} \Pr(G_i \mid h_u, h_v) p_{h_u} p_{h_v}}
\]

- Use partial counts in the design matrix
Is haplotype test more powerful?

1 Marker

---

*Fig. 1.* Sample $-\log(p\text{ values})$ against the marker map plots for window size of 1 using p values from the asymptotic $F$ test.

*Zaykin et al, 2002*
Is haplotype test more powerful?

3 Markers

*Fig. 2.* Sample $-\log(p \text{ values})$ against the marker map plots for window size of 3 using $p$ values from the asymptotic $F$ test.

*Zaykin et al., 2002*
Is haplotype test more powerful?

5 Markers

Higher than 3-marker and 1-marker test

Fig. 3. Sample \(-\log(p\text{ values})\) against the marker map plots for window sizes of 5 using \(p\) values from the asymptotic \(F\) test.

Zavkin et al. 2002

University of Texas Southwestern Medical Center
Summary for haplotype association tests

- Testing haplotypes can improve power
- Testing one haplotype is usually not enough
- The significance value need to be empirically evaluated
  
  e.g. permute case/control labels.
Imputation

- Genotype Imputation / “In silico” Genotyping

- Related individuals
  - Share segments of identity by descent
Intuition

• What is genotype of ?/?
Imputation for family samples

• Family samples share large segments of chromosomes

• Use this information, we can design a cost-effective way for genotyping

1. Genotype a few markers for all samples
2. Infer shared segments of haplotypes
3. Genotype additional markers
4. Fill in the missing genotypes
Imputation for family samples

[Genetic diagram showing family relationships and genetic markers]
Infer allele sharing
Impute missing genotype
Genotype imputation for family samples

• A particular genotype \( g_{ij} \) is missing (true value is 0, 1 or 2).

• Observed genotypes are \( G \)

• Using a pedigree likelihood (Lander-Green algorithm or Elston-Stewart algorithm), we can calculate

\[
P(g_{ij} = 0, 1 \text{ or } 2 \mid G)
\]

and

\[
\bar{g}_{ij} = E(P(g_{ij} = 0, 1 \text{ or } 2 \mid G)) = 2P(g_{ij} = 2 \mid G) + P(g_{ij} = 1 \mid G)
\]
Association test of *observed* genotypes

- Model association using a model such as:
  \[ E(y_i) = \mu + \beta_g g_i + \beta_c c_i + \ldots \]

- \( y_i \) 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
Association test of *imputed* 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
Imputation for unrelated individuals

- Family samples share longer segments of chromosome
- Unrelated individual share much short segments
- It is still possible to infer stretches of sharing between unrelated individuals
- Then the study design of unrelated individuals can be similar to family samples
Study design

Observed Genotypes

Reference Haplotypes

Study Sample

HapMap
Identify stretches

Observed Genotypes

. . . . A . . . . . . . . . A . . . . . .

Reference Haplotypes
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 c A t g g

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

University of Texas Southwestern Medical Center
Implementation

- Hidden Markov Model – a very useful but complex model
- Observe $X$ (observed genotypes)
- Goal is to infer the hidden state, $S$ (reference haplotypes)
Commonly used imputation software

1. **MaCH** – classic, highly accurate
   http://csg.sph.umich.edu/abecasis/mach/download/

2. **Minimac** – efficient, good for sequence data
   http://genome.sph.umich.edu/wiki/Minimac3

3. **Michigan Imputation Server** - cloud-based imputation service, large panel of reference haplotypes
   https://imputationserver.sph.umich.edu/index.html

4. **IMPUTE/IMPUTE2** – classic, similar to MaCH, Minimac
   https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
Will imputation work

- 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
### Imputation improve study power

<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 $r^2 > 0.80$. 
Enable combination of studies

- New LDL loci, previously associated with CAD

NOTE: Imputed SNP is denser than FUSION, SardiNIA, DGI alone.
Boost GWAS signal

- LDLR example


Summary

- Genotype imputation (in silico genotyping) can estimate missing genotypes accurately
- Genotype imputation are implemented in Hidden Markov Model
- Benefits of imputation includes:
  - Increase power of GWAS study
  - Facility combination on studies (different platform, QC et al)
  - Better interpretation of GWAS results
- Lab
  - Use MiniMac to impute artificially masked genotypes of one sample in the HapMap3 project