Haplotype phasing in large cohorts: Modeling, search, or both?

Po-Ru Loh

Harvard T.H. Chan School of Public Health
Department of Epidemiology

Broad MIA Seminar, 3/9/16
Overview

• Background: Haplotype phasing

• New long-range phasing method: Eagle
  – Results on N=150K UK Biobank samples
  – Algorithm overview

• Future directions
Humans have diploid genomes

- **Pairs** of chromosomes
  = Paternally-derived + Maternally-derived
Genotyping arrays combine paternal and maternal contributions

- **Paternal haplotype:**
  
  A C A T G G C

- **Maternal haplotype:**
  
  A T A T G G A

- **Genotype calls (measured):**
  
  AA CT AA TT GG GG AC
**Problem:** Phase information is lost!

- **Paternal haplotype:**
  
```
  A   C   A   T   G   G   C
```

- **Maternal haplotype:**
  
```
  A   T   A   T   G   G   A
```

- **Genotype calls (measured):**
  
```
  AA  CT  AA  TT  GG  GG  AC
```

- **Haplotypes could have been:**
  
```
  A   C   A   T   G   G   A
  +  A   T   A   T   G   G   C
  ```
**Problem**: Phase information is lost!

- **Paternal haplotype**: A C A T G G C

- **Maternal haplotype**: A T A T G G A

- **Genotype calls (measured)**: AA CT AA TT GG GG AC

- **Haplotypes could have been**: A C A T G G C + A T A T G G A

*Switch error!*
Problem: Phase information is lost... and phase information is useful

- Uses of phased haplotypes:
  - Genotype **imputation** (e.g., for GWAS)
  - Population genetic analyses
    - Identity-by-descent (IBD) detection
    - Demographic history inference

Marchini et al. 2007
Howie et al. 2012
Problem: Phase information is lost... and phase information is useful

So how can we recover it? (and how accurately can we do it?)
Problem: Phase information is lost... and phase information is useful

So how can we recover it? (and how accurately can we do it?)

Wet lab techniques exist, but they’re expensive (e.g., long read sequencing, Hi-C)
Problem: Phase information is lost... and phase information is useful

So how can we recover it computationally?

(and how accurately can we do it?)
Problem: Phase information is lost... and phase information is useful

So how can we recover it **computationally**? (and how accurately can we do it?)

**Teaser:** It depends on your sample size, but for \( N=150K \) UK samples...

We can phase chr20 **perfectly(!)** in \(~20\%\) of samples... and with \( \leq 2 \) switches in \( >50\% \) of samples.
Phasing (for computer scientists)

Question:
Given a diploid genotype sequence, can we recover its parental haplotypes?

Binary encoding:
- Haploid 0=A, 1=G (at A/G SNP)
- Diploid 0=AA, 1=AG, 2=GG
Question:
Given a diploid genotype sequence and diploid parental genotypes, can we recover its parental haplotypes?
Trio phasing

Answer:
Yes (for the most part):
1. Fill in homozygous sites
Trio phasing

Answer:
Yes (for the most part):
1. Fill in homozygous sites
2. Propagate info...
**Answer:**
Yes (for the most part):
1. Fill in homozygous sites
2. Propagate info... until stuck
Trio phasing

Result:
• Perfect phase* in child at sites homozygous in \( \geq 1 \) individual
• Almost-perfect phase* in parents (up to recombination)
• No phase at all-het sites; need to use LD from a panel of reference haplotypes

* up to genotype error + de novo mutation
Pedigree phasing

- Extension of trio phasing to extended families
  - Additional relatives => more Mendelian constraints (i.e., fewer all-het sites)
- MERLIN software (Abecasis et al. 2002 *Nat Genet*)
Statistical phasing

Question:
Given a large number of diploid genotype sequences, can we recover their parental haplotypes?
Statistical phasing

**Question:**
Given a **large number of** diploid genotype **sequences**, can we recover their parental haplotypes?

**Answer:**
Yes, using **linkage disequilibrium (LD)** = statistical correlation between nearby sites
General statistical phasing approach: Hidden Markov models (HMMs)

**Answer:** Various methods based on probabilistic modeling of population LD work well:
- PHASE (Stephens et al. 2001 AJHG)
- ...
- Beagle (Browning & Browning 2007 AJHG)
- HAPI-UR (Williams et al. 2012 AJHG)
- SHAPEIT2 (Delaneau et al. 2013 Nat Meth)
Limitation of existing methods: Accuracy

Answer:
Various HMM-based methods work well...

albeit not as well as pedigree-based methods:

• 2-5% switch error rates on \( N \approx 10K \) samples
  • Phase switches occur every 20-50 hets (~1-2 cM)
• Accuracy improves with \( N \)

Williams et al. 2012 *AJHG*; Delaneau et al. 2013 *Nat Meth*
Limitation of existing methods: Speed

**Answer:**
Various HMM-based methods work well... albeit not as well as pedigree-based methods... and much more slowly:
- Several days to phase a single chromosome for $N=16K$ samples

HMMs are efficient, but genomic data is big!

Williams et al. 2012 *AJHG*
Long-range phasing (LRP)

• LRP method: Kong et al. 2008 *Nat Genet*
  – 35K of 316K Icelanders typed (11%)!
  – Identify local “surrogate parents” for each individual; proceed as in trio phasing (fast!)
  – Yield: near-perfect phase calls at 90-95% of het sites

• LRP applications: ~25 subsequent papers in *Nature* or *Nat Genet*

Note: Diagram is an oversimplification; in practice, surrogate parents only match proband in sub-chromosomal segments identical-by-descent (IBD)
Challenge: Recombination breaks down IBD segments

• Shared segment lengths decrease exponentially with # of generations
  – LRP is easy with close relatives … but hard otherwise
Key contrast

HMM vs. LRP: Modeling vs. Search

accurate

fast...

and accurate at
very large sample sizes

“Let’s build a great model!”

“Let’s use lots of data!”
Predictions about LRP

• Kong et al. 2008:
  – “We speculate that having as little as 1% of a population genotyped may be adequate for the method to yield useful results.”

• Browning & Browning 2011:
  – “It is likely that... IBD-based phasing can be extended... by using more sensitive methods for detecting IBD and combining IBD-based phasing with population haplotype frequency models.”
Realization of Predictions about LRP (beyond Iceland)

• Kong et al. 2008:
  – “We speculate that having as little as 1% of a population genotyped may be adequate for the method to yield useful results.”

0.2%, in fact!

• Browning & Browning 2011:
  – “It is likely that... IBD-based phasing can be extended... by using more sensitive methods for detecting IBD and combining IBD-based phasing with population haplotype frequency models.”

Yes!

• Loh, Palamara & Price 2016 Nat Genet (in press)
Overview

• Background: Haplotype phasing

• New long-range phasing method: Eagle
  – Results on $N=150K$ UK Biobank samples
  – Algorithm overview

• Future directions
New LRP hybrid method: Eagle

A. Rapidly identify long (>4cM) IBD segments => make initial phase calls

B. Run two iterations of HMM-based phasing

Note: 4cM = common ancestor ~12 generations ago!
New data: UK Biobank (\(N=150K\) interim release)

- **Samples:**
  - \(N=150K \approx 0.2\%\) of British population
  - 94\% ‘White’ ethnicity
    - 88\% British + 3\% Irish + 3\% Other
  - 72 trios (70 white)
    - Allows benchmarking!

- **Markers:**
  - \(M=650K\) autosomal SNPs after QC
Computational performance:
Eagle achieves \(~14\times\) speedup, \(~3\times\) less RAM

- Benchmark: 6K SNPs \(\approx 1\%\) of genome
- Hardware: up to 10 cores of a Xeon L5640
Benchmarking phase accuracy

- **Switch error**: transition between following paternal haplotype vs. maternal haplotype

  - Switch error rate = (# errors) / (# hets – 1)

![Graph showing genetic position (cM) vs. samples with blue and red bars indicating matches to paternal and maternal haplotypes respectively. Each row represents a haplotype produced by the algorithm.]

- **Graph Details**:
  - Blue = matches paternal haplotype
  - Red = matches maternal haplotype
  - (from trio phasing)

![Each row represents a haplotype produced by the algorithm.](image-url)
Accuracy: Eagle achieves 0.3% switch error rate at $N=150K$

- **Benchmark:**
  - First 40cM of chr10 (~6K SNPs)
  - Trio parents removed
  - Phase accuracy evaluated on 70 trio children (at trio-phased hets)
- 0.3% error rate
  $\approx 1$ switch per 13cM
- 2/3 of Eagle and SHAPEIT2 switch errors for $N=150K$ are minor (1-3 SNPs):
  $\approx 1$ major switch / 40cM

---

**Graph:**

- **Line styles:**
  - Large markers + solid lines: Methods that can phase 150K samples genome-wide in <200 node days
Eagle achieves ~4x lower error than other genome-wide tractable methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Run time (12% of genome)</th>
<th>Switch error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eagle 1x150K</td>
<td>7.9 days</td>
<td>0.31%</td>
</tr>
<tr>
<td>SHAPEIT2 10x15K</td>
<td>24.4 days</td>
<td>1.35%</td>
</tr>
<tr>
<td>HAPI-UR 10x15K</td>
<td>15.8 days</td>
<td>2.20%</td>
</tr>
</tbody>
</table>

• Chromosome-scale analyses: chr1 (short arm), chr10, chr20
• Methods:
  – Eagle, 1 batch of all 150K samples
  – SHAPEIT2 and HAPI-UR, 10 batches of 15K samples
    • Batching is necessary due to computational cost
• Hardware: up to 10 cores of a Xeon L5640
Algorithm overview

1. Detect long matching segments (>4cM IBD); call initial phase
2. Locally refine phase in shorter segments (~1cM IBD/IBS)
3. Run two approximate HMM iterations
Step 1: Rapidly search for long diploid matches (>4cM IBD); make initial phase calls

- **Input:** **Diploid** \(\{0,1,2\}\) data for \(N\) samples
- **Procedure:**
  - Find pairs of diploid segments that share 1 haplotype IBD
  - Combine IBD info to make phase calls
- **Output:** Accurate phase in long chunks of each sample (wherever IBD exists) = **Haploid** \(\{0,1\}\) reference panel of \(2N\) haplotypes
Finding IBD in diploid data

• If two diploid segments share a haplotype, they can’t have opposite homozygous sites (0 vs. 2)

• Basic idea: Scan all pairs of samples for long segments with no opposite homozygotes
  – $O(MN^2)$ time ($N$=#samples, $M$=#markers), but very small constant factor via bit arithmetic

Henn et al. 2012 PLOS ONE
Finding IBD in diploid data

• **Basic idea:** Scan all pairs of samples for long segments with no opposite homozygotes

• **Problem:** Lots of false positives!

• **Solution:**
  – Consider allele frequencies and LD information
    => likelihood ratio score (IBD vs. chance)
  – Check overlapping regions for consistency (i.e., existence of maternal/paternal bicoloring); trust longer regions

Possible haplotype sharing
End of step 1:
Some good chunks, some bad chunks

- We've used **long IBD**
  - Phasing is great where >4cM IBD (detectable in diploids) exists
- How can we use **shorter IBD**?
Step 2: Locally refine phase using dip=hap1+hap2 constraint

- **Input:** Diploid \{0,1,2\} data for \(N\) samples + Haploid \{0,1\} provisional phasing (step 1)

- **Procedure** (for each sample in turn):
  - Find long segments of provisional \{0,1\}-haplotypes consistent with \{0,1,2\} diploid sample
  - In overlapping \(~1\)cM windows, check existence of implied complementary haplotype: \(hap2 = dip - hap1\)

- **Output:** Improved phase calls (switch error rate \(~1.5\)%)

![Step 2: Local phase refinement](image)
Using the dip=hap1+hap2 constraint

• Input: **Diploid** \(\{0,1,2\}\) data for \(N\) samples
  + **Haploid** \(\{0,1\}\) provisional phasing: \(2N\) haplotypes

\[
\begin{array}{c}
???????????????????????????????????????????????????????? (hap2) \\
+ 00010011001000001000001010000101000 & (hap1) \\
00122021002010000101001200000110011 & (dip)
\end{array}
\]

Possible haplotype sharing

• Find long segments of provisional \(\{0,1\}\)-haplotypes consistent with \(\{0,1,2\}\) diploid sample
Using the dip=hap1+hap2 constraint

- Input: **Diploid** \{0,1,2\} data for \(N\) samples
  + **Haploid** \{0,1\} provisional phasing: \(2N\) haplotypes

\[
\begin{array}{c}
?????
\end{array}
\begin{array}{c}
010001010000000100110000001
\end{array}
\begin{array}{c}
?????
\end{array}
\] (hap2)

\[
\begin{array}{c}
+ \begin{array}{c}
00010011001000000100000100000101000
\end{array}
\end{array}
\] (hap1)

\[
\begin{array}{c}
00122021002010000101001200000110011
\end{array}
\] (dip)

Possible haplotype sharing

- Does the implied haplotype (hap2 = dip – hap1) exist among the provisional haplotypes?
  - Need fast, error-tolerant search
  - Locality-sensitive hashing (LSH) overcomes “curse of dimensionality” (work by Indyk, Motwani, et al.)
    - Idea: multiple random hashes
End of step 2: Mostly good chunks; sporadic errors

- We've used long+short IBD
- How can we refine the inference?
Step 3: Run two HMM-like iterations

- **Input:** Diploid \{0,1,2\} data for \(N\) samples
  + Haploid \{0,1\} provisional phasing (step 2)

- **Procedure** (for each sample in turn):
  - Build small sets of locally best reference haplotypes
  - Find approximate max-likelihood (Viterbi) path
    - Instead of full dynamic programming, use beam search: branch and prune
LRP vs. HMM-based phasing

• LRP: “top-down”
  – **Search** for long stretches of IBD; use homozygous sites to call phase
  – Don’t worry about all-het sites; clean up later

• HMM: “bottom-up”
  – **Model** haplotype structure
  – Get short-range phasing right first; then improve phase accuracy at longer scales
  – Well-engineered HMMs (SHAPEIT2) implicitly make use of long IBD
    • O’Connell et al. 2014 *PLOS Genet*
Future direction: Best of both worlds?

HMM vs. LRP: Modeling vs. Search

- accurate
- fast...
  and accurate at very large sample sizes

- How about both?
- Positional Burrows-Wheeler Transform (PBWT, Durbin 2014): efficient data structure for haplotype storage/search
Eagle-PBWT (work in progress)

Series of haplotype trees

$M$
001000010100101000101110010001
010010100100100100011010010010
...
000100100100000100100100010000

$N$

$\Theta$

$1$

$\Theta$

$\Theta$

$1$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$

$\Theta$
Other future directions

• Reference-based phasing
  – Haplotype Reference Consortium (HRC) has $N=32K$ samples for phasing + imputation

• Applications of high-accuracy phasing
  – Detecting allele-specific expression
  – Detecting clonal mosaicism
  – Inferring demography
Acknowledgments

- Alkes Price
- Pier Francesco Palamara
- Gaurav Bhatia
- Sasha Gusev
- Mark Lipson
- Bogdan Pasaniuc
- Nick Patterson
- Noah Zaitlen

Eagle software:  http://hsph.harvard.edu/alkes-price/software/

In-sample imputation accuracy:

\[ R^2 \geq 0.75 \] down to 0.1% MAF

• Benchmark:
  - First 40cM of chr10 (~6K SNPs)
  - 2% of genotypes randomly masked and imputed
  - In-sample imputation \( R^2 \) measured on 120K British-ancestry samples

• Note: Eagle and SHAPEIT2 hard-call imputed haplotypes; dosages allow higher \( R^2 \)
In-sample imputation accuracy remains high for non-British samples

• Benchmark:
  – First 40cM of chr10 (~6K SNPs)
  – 2% of genotypes randomly masked and imputed
  – In-sample imputation $R^2$ stratified by self-reported ancestry (including minority populations):
    • 5K “any other white”
    • 1K Indian
    • 1K Caribbean
Eagle pre-phasing improves downstream imputation accuracy

<table>
<thead>
<tr>
<th>MAF bin</th>
<th>SHAPEIT2 10x15K</th>
<th>Eagle 1x150K</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1–0.2%</td>
<td>0.574 (0.012)</td>
<td>0.594 (0.012)</td>
<td>0.020 (0.002)</td>
</tr>
<tr>
<td>0.2–0.5%</td>
<td>0.665 (0.010)</td>
<td>0.679 (0.010)</td>
<td>0.013 (0.002)</td>
</tr>
<tr>
<td>0.5–1%</td>
<td>0.753 (0.009)</td>
<td>0.765 (0.009)</td>
<td>0.012 (0.001)</td>
</tr>
<tr>
<td>1–2%</td>
<td>0.786 (0.008)</td>
<td>0.798 (0.008)</td>
<td>0.012 (0.001)</td>
</tr>
<tr>
<td>2–5%</td>
<td>0.812 (0.007)</td>
<td>0.822 (0.007)</td>
<td>0.010 (0.001)</td>
</tr>
<tr>
<td>5–10%</td>
<td>0.881 (0.007)</td>
<td>0.888 (0.006)</td>
<td>0.007 (0.000)</td>
</tr>
<tr>
<td>10–50%</td>
<td>0.924 (0.004)</td>
<td>0.928 (0.004)</td>
<td>0.004 (0.000)</td>
</tr>
</tbody>
</table>

- Chromosome-scale analyses: chr1 (short arm), chr10, chr20
- Pre-phasing:
  - Eagle, 1 batch of all 150K samples
  - SHAPEIT2, 10 batches of 10K samples (due to computational cost)
- Imputation: Haplotype Reference Consortium (r1)
  - PBWT imputation algorithm on Sanger Imputation Service