Computational Problems in Haplotype Recognition

by

Ali Katanforoush

Under supervision of
Dr Hamid Pezeshk and Dr Mehdi Sadeghi

A thesis submitted to the Graduate Studies Office of
University of Tehran

In partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Bioinformatics

Institute of Biochemistry and Biophysics

November 1, 2009
Outlines

- Haplotype basis and terminology
- Haplotype inference
- Haplotype block partitioning
- Assessment of haplotype blocks
  - Common haplotype coverage and tagSNP coverage
  - Robustness of partitioning method
  - Application to recombination hotspot detection
  - Application to disease association studies
**Single Nucleotide Polymorphism; SNP**

- *A genetic variation in a single nucleotide that is sometimes observed among population; not too rare.*

- **SNPs are usually bi-allelic.**

```
AGGACTAGATAATAGACCG
AGGACCACATTATAGTCGG
AGGACCAGATAATAGTCGG
ATGACCACTATTATAGTCGG
ATGACTACATAATAGACCG
```

```
  0 1 1 0 0
  0 0 0 1 1
  0 0 1 0 1
  1 0 0 1 1
  1 1 0 0 0
```
Single Nucleotide Polymorphism; SNP

- **SNP is the result of a substantiated single site mutation in population.**

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>SNP₁</th>
<th>SNP₂</th>
<th>SNP₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haplotype₁</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Haplotype₂</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Haplotype₃</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Haplotype₄</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Single Nucleotide Polymorphism; SNP

- **SNPs** are the most common form of genetic polymorphism in genomes.
- Each new cell contains ~3 new mutations.
- Each new “child” ~20 new mutations.
- Currently more then 4.3 million SNPs have been reported to dbSNP; (0.1% of whole genome).
Define patterns of genetic variation across human genome.
Guide selection of SNPs efficiently to “tag” common variants.
Public release of all data (assays, genotypes).
Haplotype Map of the Human Genome

- Phase I: 1.3 M SNPs in 269 people.
- Phase II: +2.8 M SNPs in 270 people;
  - 30 parent-parent-offspring trios from Nigeria (YRI)
  - 30 trios of European descent from Utah (CEU)
  - 45 unrelated individuals from Beijing (CHB)
  - 45 unrelated individuals from Tokyo (JPT)
- Phase III: 1.3 M SNPs in 1184 people (10 panels).
The first problem; **Genotype Phasing**

- Every genotype can be considered as sum of two **unknown** haplotypes.

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Genotyping</th>
</tr>
</thead>
<tbody>
<tr>
<td>G/T T/T C/G A/A A/A</td>
<td>1 2 1 0 0</td>
</tr>
<tr>
<td>G/G C/C C/G A/T T/T</td>
<td>0 0 1 1 2</td>
</tr>
<tr>
<td>T/T C/T C/C A/T A/T</td>
<td>2 1 0 1 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Real haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A T G A C T A C A T A A T A G A C C G</td>
</tr>
<tr>
<td>A G G A C T A G A T A A T A G A C C G</td>
</tr>
<tr>
<td>A G G A C C A C A T T A T A G T C C G</td>
</tr>
<tr>
<td>A G G A C C A G A T A A T A G T C C G</td>
</tr>
<tr>
<td>A T G A C C A C A T T A T A G T C C G</td>
</tr>
<tr>
<td>A T G A C T A C A T A A T A G A C C G</td>
</tr>
</tbody>
</table>
The first problem; **Genotype Phasing**

- **Given a set of genotype samples of unrelated individuals, determine pairs of haplotypes adding up into given genotypes.**

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Real haplotypes</th>
<th>Inferred haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>G/T</td>
<td>ATGAC T A CAT A ATAG A CCG</td>
<td>1 1 0 0 0</td>
</tr>
<tr>
<td>T/T</td>
<td>ATGAC T A CAT A ATAG A CCG</td>
<td>0 1 1 0 0</td>
</tr>
<tr>
<td>C/G</td>
<td>AGGAC T AGAT AATAG A CCG</td>
<td>0 1 0 0 1</td>
</tr>
<tr>
<td>A/A</td>
<td>AGGAC T AGAT AATAG A CCG</td>
<td>0 1 0 0 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Genotyping</th>
<th>Computational phasing</th>
</tr>
</thead>
<tbody>
<tr>
<td>G/T</td>
<td>1 2 1 0 0</td>
<td>0 1 0 0 0 0</td>
</tr>
<tr>
<td>T/T</td>
<td>1 2 1 0 0</td>
<td>0 1 1 1 0 0</td>
</tr>
<tr>
<td>C/G</td>
<td>0 0 1 1 2</td>
<td>0 0 0 0 1 1</td>
</tr>
<tr>
<td>A/A</td>
<td>1 2 1 0 0</td>
<td>0 0 1 1 1 1</td>
</tr>
<tr>
<td>A/T</td>
<td>2 1 0 1 1</td>
<td>1 0 0 0 0 0</td>
</tr>
<tr>
<td>T/T</td>
<td>2 1 0 1 1</td>
<td>1 1 0 1 1 1</td>
</tr>
</tbody>
</table>

---

[Image of a data table showing genotypes and corresponding haplotypes]
Haplotyping inference by maximum parsimony

- **Inferring the set of haplotypes consistent to genotype data requiring to some biological considerations.**
- **Maximum parsimony** is one of the most common models in biology.
- **Other models; Perfect phylogeny, Maximum likelihood, Bayesian model.**
Haplotype inference by maximum parsimony

- Clark's algorithm (1990); greedy algorithm
- 0/1 linear programming, Gusfield (2003)
Methods on other approaches to haplotype inference


- **Inference of haplotype frequencies by maximum likelihood**
  - *Expectation-Maximization*, EM
    Slatkin and Excoffier (1995)
  - *Partition-Ligation*, PL-EM
    Qin et al (2002)

- **Bayesian model**, Smith and Donnelly and Stephens (2001); **PHASE**
Genetic Algorithm; GA

\[
\min f(x) \\
\text{s.t. } P(x) = \text{true}
\]

- Consider \( N \) feasible solutions; each one is represented by a bit string called “chromosome”.
- Select “chromosomes” of highest fitness to produce a new generation.
- Cross-over random pairs of selected “chromosomes” and mutate some bits on other “chromosomes”.
- The optimal solution should be obtained by long repeats.
Genetic Algorithm for haplotype inference with maximum parsimony

- Given $n$ genotypes on $l$ SNPs; $g_1, g_2, \ldots, g_n$
  find

$$\min |H|$$

s.t. $$\exists h_a, h_b \in H : g_i = h_a \oplus h_b, \text{ for } i = 1, 2, \ldots, n$$

- Braaten et al. (2000). *The GA applied to haplotype data at the LDL receptor locus.*
- Tapadar et al. (2000). *Haplotyping in pedigrees via GA.*
- Azuma et al. (2009). *Haplotype frequency estimation by GA.*
A *naive* Genetic Algorithm for MP haplotyping

- “*Chromosome*” representation

<table>
<thead>
<tr>
<th>amb</th>
<th>G</th>
<th>ξ</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12100</td>
<td>1-2--</td>
<td>01100</td>
</tr>
<tr>
<td>2</td>
<td>00112</td>
<td>--34--</td>
<td>11000</td>
</tr>
<tr>
<td>3</td>
<td>21011</td>
<td>-567</td>
<td>00101</td>
</tr>
<tr>
<td>1</td>
<td>10220</td>
<td>8-----</td>
<td>00011</td>
</tr>
</tbody>
</table>

\(\chi\)

\[
\begin{array}{cccccc}
0 & 1 & 1 & 0 & 1 & 0
\end{array}
\]
A naive Genetic Algorithm for MP haplotyping

• “Crossing-over on chromosomes”
A naive Genetic Algorithm for MP haplotyping

- "Mutation on chromosomes"

\[
\begin{align*}
H_{\text{parent}} & = \begin{bmatrix}
0 & 1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1
\end{bmatrix} \\
H_{\text{mutant}} & = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 1
\end{bmatrix}
\end{align*}
\]

\[
\begin{align*}
X_{\text{parent}} & = \begin{bmatrix}
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0
\end{bmatrix} \\
X_{\text{mutant}} & = \begin{bmatrix}
0 & 0 & 1 & 1 & 0 & 1 & 0 & 0
\end{bmatrix}
\end{align*}
\]
A parametric greedy phasing aimed to MP

- **Input:** $n$ genotypes on $l$ SNPs,
- **Algorithm parameters:**
  - a permutation of \( \{1, 2, ..., n\} \), \( \sigma = \langle \sigma_1, \sigma_2, ..., \sigma_n \rangle \)
  - a set of “guide haplotypes” \( \{ h_1, h_2, ..., h_n \} \) where
    \[
    h_i \sim g_i
    \]
- In a greedy manner, it tries to resolve \( g(\sigma_i) \) with one of haplotypes resolving \( g(\sigma_1), g(\sigma_2), ..., g(\sigma_{i-1}) \), but if it fails then applies \( h_i \).
The Genetic Algorithm for MP phasing

- Each “chromosome” ↔ an instances of greedy phasing algorithm
- Various permutations and “guide haplotypes” are encoded by bit-strings.
- Naive procedures for crossing over and mutation are applied on “guide haplotypes”.
- Cross-over and mutation on permutations are also convenient.
Cross-over on permutations

\[ \sigma_1 \begin{array}{cccccccc} 4 & 1 & 7 & 3 & 2 & 6 & 5 & 8 \end{array} \]

\[ \sigma_2 \begin{array}{cccccccc} 2 & 3 & 5 & 6 & 4 & 7 & 8 & 1 \end{array} \]

random merge

cross out repeats

\[ \begin{array}{cccccccc} 4 & 1 & 2 & 7 & 3 & 3 & 2 & 6 & 5 & 6 & 4 & 5 & 7 & 8 & 8 & 1 \end{array} \]

\[ \sigma_{\text{hybrid}} \begin{array}{cccccccc} 4 & 1 & 2 & 7 & 3 & 6 & 5 & 8 \end{array} \]
## Parameter setting for GAhap

<table>
<thead>
<tr>
<th>cr</th>
<th>$cr_{int}$</th>
<th>$mr_{int}$</th>
<th>selection</th>
<th>fitness scaling</th>
<th>successful cases of 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
<td>stochastics</td>
<td>shift linear</td>
<td>18</td>
</tr>
<tr>
<td>0.2</td>
<td>0.5</td>
<td>0.5</td>
<td>stochastics</td>
<td>rank</td>
<td>16</td>
</tr>
<tr>
<td>0.2</td>
<td>0.9</td>
<td>0.1</td>
<td>tournament</td>
<td>linear</td>
<td>16</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.9</td>
<td>uniform</td>
<td>rank</td>
<td>15</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>tournament</td>
<td>rank</td>
<td>14</td>
</tr>
</tbody>
</table>
Effect of “cross over” on convergence

$cr=0$

Best: 31 Mean: 40.4

$cr=0.2$

Best: 29 Mean: 32.2
| method         | framework                      | $|H|$ | haplotype error rate | switch error rate |
|----------------|--------------------------------|-----|----------------------|-------------------|
| HAPLOTYPER     | Bayesian Drichlet prior        | 33  | 5.4                  | 3.0               |
| PHASE          | Bayesian Perfect phylogeny     | 32  | 5.6                  | 3.1               |
| fastPHASE      | Simplified PHASE              | 35  | 7.3                  | 4.5               |
| 2SNP           | 2 SNPs phasing and MST        | 40  | 10.4                 | 5.6               |
| GAhap          | GA and MP                     | 34  | 9.7                  | 5.7               |

Methods have been evaluated with 150 genotypes of GH1 with known phases (Horan et al, 2003)
Before second problem

Generate random haplotype samples under coalescent model

- Simulate a coalescent process.
Before second problem

Generate random haplotype samples under coalescent model

- Determine haplotype frequencies constrained to minor allele frequency.

$$ N^* = \min_f N = \sum_{i=1}^{m} f_i $$

s.t.

$$ \sum_{i=1}^{m} h_{ij} f_i \geq \mu_o N , \quad j = 1, \ldots, l $$

$$ \sum_{i=1}^{m} h_{ij} f_i \leq (1 - \mu_o)N , \quad j = 1, \ldots, l $$

$$ f_i \geq 1 , \quad i = 1, \ldots, m $$
Second problem; *Haplotype block partitioning*

- Genome comprises regions with certain boundaries of which haplotypes are transferred without change through generations.

- *Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.* Patil et al. (2001)

- *Pattern of Linkage Disequilibrium shows a picture of discrete haplotype blocks over genome.* Daly et al. (2001)

- *Haplotype blocks arise in the absence of recombination hot spots.* Wang et al. (2002)
Blocks of limited haplotype diversity

<table>
<thead>
<tr>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>011010000001010</td>
<td>011010</td>
<td>0000</td>
</tr>
<tr>
<td>000101000000110</td>
<td>011010</td>
<td>0000</td>
</tr>
<tr>
<td>011010000001010</td>
<td>011010</td>
<td>0000</td>
</tr>
<tr>
<td>011010011000110</td>
<td>011010</td>
<td>01010</td>
</tr>
<tr>
<td>0001010111001011</td>
<td>000101</td>
<td>0110</td>
</tr>
<tr>
<td>101000100001011</td>
<td>000101</td>
<td>0110</td>
</tr>
<tr>
<td>00010111000010</td>
<td>000101</td>
<td>0110</td>
</tr>
<tr>
<td>01101011000010</td>
<td>000101</td>
<td>0110</td>
</tr>
<tr>
<td>01101011000010</td>
<td>010100</td>
<td>1100</td>
</tr>
<tr>
<td>101000110010011</td>
<td>101000</td>
<td>0010</td>
</tr>
</tbody>
</table>
An early example of haplotype blocks

- **Block structure and common haplotypes on 5q31**

  ![Diagram showing the block structure and common haplotypes on 5q31](image)

  **SNPs**

  **Block sizes:**
  - block 1: 84 kb
  - block 2: 3 kb
  - block 3: 14 kb
  - block 4: 30 kb
  - block 5: 25 kb
  - block 6: 11 kb
  - block 7: 92 kb
  - block 8: 21 kb
  - block 9: 27 kb
  - block 10: 55 kb
  - block 11: 19 kb

  courtesy Daly et al. (2001)
Bases of haplotype block definition

- **Haplotype diversity**
  - common haplotype
  - minimum number of SNPs to cover information of majority of haplotypes, (Patil et al. 2001, Zhang et al. 2002)

- **Linkage Disequilibrium**
  - point estimation of LD coefficient, $D$, $r^2$, $D'$
  - interval estimation, (Gabriel et al. 2002)

- **Four gamete test**, (Wang et al. 2002)
Local partitioning vs. global partitioning methods

- A local partitioning method defines haplotype blocks in a way that boundaries of each block are determined independent from other blocks.
- By local partitioning usually, a series of separated regions on genome, like “islands” forms blocks.
- A global partitioning method defines a whole partitioning for genome rather defining each block independently.
- By Global partitioning usually, genome is “tiled” by blocks tightly placed next to each other.
## Methods on haplotype block partitioning

<table>
<thead>
<tr>
<th>abbr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOT</td>
</tr>
<tr>
<td>MB</td>
</tr>
<tr>
<td>HB</td>
</tr>
<tr>
<td>MDL</td>
</tr>
<tr>
<td>GAM</td>
</tr>
<tr>
<td>GAB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Partitioning Block definition structure</th>
<th>Block constraint</th>
<th>Software</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotspot</td>
<td>Local*</td>
<td>recombination hotspots</td>
<td></td>
<td>Myers, 05</td>
</tr>
<tr>
<td>Minimum block number</td>
<td>Global</td>
<td>minimum number of blocks</td>
<td>haplotype diversity</td>
<td>Zhang, 05</td>
</tr>
<tr>
<td>HapBlock</td>
<td>Global</td>
<td>minimizing total number of tagSNPs</td>
<td>haplotype diversity</td>
<td>Zhang, 05</td>
</tr>
<tr>
<td>MDBlock</td>
<td>Global</td>
<td>minimum description length</td>
<td>haplotype diversity</td>
<td>Anderson, 03</td>
</tr>
<tr>
<td>Four gamete test</td>
<td>Local</td>
<td>evidence of recombination</td>
<td>fourth gamete</td>
<td>Wang, 02</td>
</tr>
<tr>
<td>Gabriel’s method</td>
<td>Local</td>
<td>evidence of recombination</td>
<td>strongly associated SNPs</td>
<td>Gabriel, 02</td>
</tr>
</tbody>
</table>

Note: Precomputed results available by HapMap.
Application of haplotype block partitioning

- Studies on human origin, history of human migrations and genetic diversity between races
- Genetic mapping and recognizing recombination hotspots
- Genotyping and phasing
- tagSNP selection
- Disease Association Study
Global haplotype partitioning for maximal associated SNP pairs

Outlines

- Categorize SNP pairs into association classes.
- Establish a constrained optimization to find blocks which include the most possible number of “associated” pairs subjected to limited number of “independent” pairs.
- Solve the constrained programming.
Linkage Disequilibrium between two SNPs

\[
P(AB) = P(A) \cdot P(B)
\]

- In the presence of crossing over,
- In the absence of crossing over,

\[
P(AB) \neq P(A) \cdot P(B)
\]

- Enough long time after being settled
- no admixture
- no selection

- In the presence of crossing over,
  \[P(AB) = P(A) \cdot P(B)\]
- In the absence of crossing over,
  \[P(AB) \neq P(A) \cdot P(B)\]
Standardized coefficient of LD

\[ D = D_{XY} = P(X = 1, Y = 1) - P(X = 1).P(Y = 1) \]

\[ D' = D'_{XY} = \frac{D_{XY}}{D_{MAX}} \]

\[ D_{MAX} = \begin{cases} 
\min(P(X = 0).P(Y = 1), P(X = 1).P(Y = 0)), & \text{if } D_{XY} > 0 \\
\min(P(X = 0).P(Y = 0), P(X = 1).P(Y = 1)), & \text{if } D_{XY} < 0 
\end{cases} \]
Assessment of LD estimation

- **Confidence interval**, (Gabriel et al, 2002)
  - Apply thresholds on confidence interval of $|D'|$
  - Each SNP pair is then categorized into three classes; “strongly associated”, “recombinant” and “uninformative”

- **Fisher’s exact test and p-value**, (present work)

<table>
<thead>
<tr>
<th></th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
<th>$n - n_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>$n_{00}$</td>
<td>$n_{01}$</td>
<td>$n - n_a$</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>$n_{10}$</td>
<td>$n_{11}$</td>
<td>$n_a$</td>
</tr>
</tbody>
</table>

$$F_{ex} = \frac{n_a!n_b!(n - n_a)!(n - n_b)!}{n!n_{00}!n_{01}!n_{10}!n_{11}!}$$
An association index for SNP pairs based on Fisher’s Exact Test

\[
p_{\text{left}} = \sqrt{\frac{1}{n_{11}}} F_{\text{ex}}(n_{11}) + \sum_{i<n_{11}} F_{\text{ex}}(i),
\]

\[
p_{\text{right}} = \sqrt{\frac{1}{n_{11}}} F_{\text{ex}}(n_{11}) + \sum_{i>n_{11}} F_{\text{ex}}(i),
\]

\[
p_{\text{value\_one\_tailed}} = \begin{cases} 
  p_{\text{left}}, & \text{if } n_{11} < n_{\text{max}} \\
  p_{\text{right}}, & \text{if } n_{11} > n_{\text{max}} \\
  \max(p_{\text{left}}, p_{\text{right}}), & \text{if } n_{11} = n_{\text{max}}
\end{cases}
\]
An association index for SNP pairs based on Fisher’s Exact Test

- Estimate value of $\mid D' \mid$ on given sample.
- Compute $p$-value of Fisher’s exact test.
- Apply thresholds on $p$-value results in a three state association index; “associated”, “independent” and “not statistically significant”.

Notion of Fisher’s Exact Test

| \( n_{11} \) | 5 | 4 | 3 | 2 | 2 | 2 | 2 | 3 | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 4 | 3 | 2 | 3 | 3 | 4 | 4 | 3 | 3 | 3 |
| \( F_{ex} \) | 1 | 9 | 30 | 38 | 38 | 38 | 30 | 3 | 38 | 38 | 38 | 38 | 38 | 38 | 19 | 19 | 38 | 19 | 38 | 19 | 9 | 30 | 38 | 30 | 30 | 9 | 9 | 30 | 30 | 30 | 30 |
| \( r^2 \) | .41 | .16 | .03 | 0 | 0 | 0 | 0 | .03 | .27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .08 | .08 | 0 | .08 | .16 | .03 | 0 | .03 | .03 | .16 | .16 | .03 | .03 | .03 |
| \( |D'| \) | 1 | .64 | .27 | .11 | .11 | .11 | .11 | .27 | 1 | .11 | .11 | .11 | .11 | .11 | .56 | .56 | .11 | .11 | .11 | .56 | .56 | .11 | .56 | .64 | .27 | .11 | .27 | .27 | .64 | .64 | .27 | .27 | .27 | .27
Global haplotype partitioning for maximal associated SNP pairs

- Establish a constrained optimization ...
Global haplotype partitioning for maximal associated SNP pairs

- Establish a constraint optimization ...

\[
\max \sum_{i=1}^{k} B[s_{i-1}, s_i]
\]

\[
s.t. \quad \sum_{i=1}^{k} A[s_{i-1}, s_i] < \alpha N_{ind}
\]
Solve the constrained programming

- Convert into an unconstrained optimization using a Lagrange multiplier;
  \[
  \max_S \sum_{i=1}^{k} B[s_{i-1}, s_i] - \lambda A[s_{i-1}, s_i]
  \]

- Given a fixed \( \lambda \), the partitioning can be obtained via a dynamic programming procedure;
\[
S^{\text{opt}}(\cdot) = \cdot,
\]
\[
S^{\text{opt}}(i) = \max_{1 \leq d \leq \min(w, i)} \left\{ S^{\text{opt}}(i - d) + S(i; d) \right\}, \quad \text{for } i = 1, \ldots, l
\]
Global haplotype partitioning for maximal associated SNP pairs

<table>
<thead>
<tr>
<th>abbr.</th>
<th>Method</th>
<th>Partitioning Block definition structure</th>
<th>Block constraint</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPG</td>
<td>GPMAP** based on Gabriel’s index</td>
<td>Global maximizing associated SNPs</td>
<td>independent SNPs</td>
<td>Haplovie+GPMAP</td>
</tr>
<tr>
<td>GPF</td>
<td>GPMAP based on Fisher’s exact test</td>
<td>Global maximizing associated SNPs</td>
<td>independent SNPs</td>
<td>Haplovie+GPMAP</td>
</tr>
</tbody>
</table>
Method evaluation

• General features of haplotype blocks,
  – block length and block distribution
  – coverage of “common haplotypes”
  – consistency with LD pattern
  – the number of minimum tagSNP and coverage
  – similarity between different partitioning methods

• Robustness of partitioning method.

• Performance on identification recombination hotspots

• Performance on case-control association study.
Evaluation on ENCODE haplotypes

● The Encyclopedia of DNA Elements (ENCODE)

● Ten regions have been selected by ENCODE project as the pilot phase to identify the functional elements of human genome.

● There are about 2000 SNPs assayed by the HapMap Project in each ENCODE region (CEU panel).

● We reduced SNPs to those which are commonly ascertained for all three HapMap panels.

● Moreover, we drew out the top 400 SNPs ordered by heterozygosity out of each region.
## General features of haplotype blocks

<table>
<thead>
<tr>
<th></th>
<th>HOT</th>
<th>MB</th>
<th>HB</th>
<th>MDL</th>
<th>GAM</th>
<th>GAB</th>
<th>GPG</th>
<th>GPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average block length (kbp)</td>
<td>68.9</td>
<td>46.8</td>
<td>36.2</td>
<td>17.1</td>
<td>13.3</td>
<td>23.3</td>
<td>35.7</td>
<td>39.7</td>
</tr>
<tr>
<td>Average block length (SNP)</td>
<td>52</td>
<td>36</td>
<td>27</td>
<td>13</td>
<td>10</td>
<td>18</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>Total run time (sec.)</td>
<td>N/A</td>
<td>22</td>
<td>743</td>
<td>3295</td>
<td>112</td>
<td>143</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Common haplotype coverage</td>
<td>0.67</td>
<td>0.89</td>
<td>0.91</td>
<td>0.96</td>
<td>0.96</td>
<td>0.93</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>Hole freq.</td>
<td>0.50</td>
<td>0.23</td>
<td>0.14</td>
<td>0.06</td>
<td>0.04</td>
<td>0.04</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>Island freq.</td>
<td>0.08</td>
<td>0.10</td>
<td>0.09</td>
<td>0.17</td>
<td>0.19</td>
<td>0.11</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Consistency with LD pattern
The number of haplotype tagging SNPs

<table>
<thead>
<tr>
<th>region name</th>
<th>chromosome band</th>
<th>HOT</th>
<th>MB</th>
<th>HB</th>
<th>MDL</th>
<th>GAM</th>
<th>GAB</th>
<th>GPG</th>
<th>GPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENr112</td>
<td>2p16.3</td>
<td>37</td>
<td>35</td>
<td>33</td>
<td>51</td>
<td>96</td>
<td>77</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>ENr131</td>
<td>2q37.1</td>
<td>32</td>
<td>48</td>
<td>40</td>
<td>60</td>
<td>101</td>
<td>79</td>
<td>47</td>
<td>42</td>
</tr>
<tr>
<td>ENr113</td>
<td>4q26</td>
<td>34</td>
<td>37</td>
<td>30</td>
<td>46</td>
<td>67</td>
<td>47</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>ENm010</td>
<td>7p15.2</td>
<td>36</td>
<td>37</td>
<td>34</td>
<td>49</td>
<td>87</td>
<td>78</td>
<td>37</td>
<td>36</td>
</tr>
<tr>
<td>Enm013</td>
<td>7q21.13</td>
<td>17</td>
<td>16</td>
<td>15</td>
<td>34</td>
<td>69</td>
<td>29</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>Enm014</td>
<td>7q31.33</td>
<td>38</td>
<td>27</td>
<td>25</td>
<td>48</td>
<td>67</td>
<td>47</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>ENr321</td>
<td>8q24.11</td>
<td>27</td>
<td>35</td>
<td>26</td>
<td>49</td>
<td>63</td>
<td>49</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td>ENr232</td>
<td>9q34.11</td>
<td>51</td>
<td>47</td>
<td>42</td>
<td>63</td>
<td>70</td>
<td>75</td>
<td>49</td>
<td>52</td>
</tr>
<tr>
<td>ENr123</td>
<td>12q12</td>
<td>14</td>
<td>33</td>
<td>29</td>
<td>48</td>
<td>79</td>
<td>59</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>ENr213</td>
<td>18q12.1</td>
<td>31</td>
<td>36</td>
<td>30</td>
<td>52</td>
<td>69</td>
<td>46</td>
<td>28</td>
<td>33</td>
</tr>
</tbody>
</table>

- The minimum number of htSNPs for each haplotype block has been obtained using htSNPer (Ding et al. 2005)
htSNP coverage
Similarity of blocks between different methods

<table>
<thead>
<tr>
<th>Method</th>
<th>HOT</th>
<th>MB</th>
<th>HB</th>
<th>MDL</th>
<th>GAM</th>
<th>GAB</th>
<th>GPG</th>
<th>GPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOT</td>
<td>1.00</td>
<td>0.39</td>
<td>0.37</td>
<td>0.20</td>
<td>0.18</td>
<td>0.32</td>
<td>0.45</td>
<td>0.46</td>
</tr>
<tr>
<td>MB</td>
<td>0.39</td>
<td>1.00</td>
<td>0.67</td>
<td>0.36</td>
<td>0.31</td>
<td>0.58</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>HB</td>
<td>0.37</td>
<td>0.67</td>
<td>1.00</td>
<td>0.42</td>
<td>0.36</td>
<td>0.60</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>MDL</td>
<td>0.20</td>
<td>0.36</td>
<td>0.42</td>
<td>1.00</td>
<td>0.54</td>
<td>0.44</td>
<td>0.31</td>
<td>0.29</td>
</tr>
<tr>
<td>GAM</td>
<td>0.18</td>
<td>0.31</td>
<td>0.36</td>
<td>0.54</td>
<td>1.00</td>
<td>0.44</td>
<td>0.26</td>
<td>0.24</td>
</tr>
<tr>
<td>GAB</td>
<td>0.32</td>
<td>0.58</td>
<td>0.60</td>
<td>0.44</td>
<td>0.44</td>
<td>1.00</td>
<td>0.57</td>
<td>0.53</td>
</tr>
<tr>
<td>GPG</td>
<td>0.45</td>
<td>0.60</td>
<td>0.58</td>
<td>0.31</td>
<td>0.26</td>
<td>0.57</td>
<td>1.00</td>
<td>0.89</td>
</tr>
<tr>
<td>GPF</td>
<td>0.46</td>
<td>0.57</td>
<td>0.55</td>
<td>0.29</td>
<td>0.24</td>
<td>0.53</td>
<td>0.89</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Robustness of block partitioning

How many times a certain method reproduce the same boundaries when applied to simulated recombinant haplotypes?

Boundaries of haplotype blocks in 9q34.11 obtained by different methods.
Application to recombination hotspots detection

• Generate random haplotype samples under coalescent model with recombination using msHOT (Hellenthal & Stephens 2007);

• Two simulated haplotype set, each one with 100 samples

• Each sample contains 40/100 haplotypes on 300 SNPs

• Six 2kb regions are considered as hotspots regions, in random

• Recombination rate is chosen 50-400 times higher than background for hotspots.
**Application to recombination hotspots detection**

- **Total error rate on detection of recombination hotspots**

<table>
<thead>
<tr>
<th></th>
<th>MB</th>
<th>HB</th>
<th>MDL</th>
<th>GAM</th>
<th>GAB</th>
<th>GPG</th>
<th>GPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>false positive rate</td>
<td>2.6</td>
<td>15.4</td>
<td>3.8</td>
<td>5.3</td>
<td>3.0</td>
<td>1.9</td>
<td>1.6</td>
</tr>
<tr>
<td>false negative rate</td>
<td>2.5</td>
<td>0.9</td>
<td>1.2</td>
<td>0.7</td>
<td>3.1</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>total error rate</td>
<td>5.1</td>
<td>16.3</td>
<td>5.0</td>
<td>6.0</td>
<td>6.1</td>
<td>2.8</td>
<td>2.6</td>
</tr>
<tr>
<td>N=100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>false positive rate</td>
<td>3.8</td>
<td>12.2</td>
<td>2.6</td>
<td>4.9</td>
<td>3.4</td>
<td>1.1</td>
<td>1.0</td>
</tr>
<tr>
<td>false negative rate</td>
<td>2.2</td>
<td>1.1</td>
<td>0.9</td>
<td>0.9</td>
<td>2.2</td>
<td>0.8</td>
<td>1.1</td>
</tr>
<tr>
<td>total error rate</td>
<td>6.0</td>
<td>13.3</td>
<td>3.5</td>
<td>5.8</td>
<td>5.6</td>
<td>1.9</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Application to recombination hotspots detection

(a) $n = 40$
(b) $n = 100$
Application to disease association study

- Single site association test;

\[
\begin{array}{c|cc|}
 & X = 0 & X = 1 \\
\hline
\text{case} & n_{0,\text{case}} & n_{1,\text{case}} \\
\text{control} & n_{0,\text{control}} & n_{1,\text{control}} \\
\hline
n_0 & n_1 & n
\end{array}
\]

\[
\chi^r_{ss} = n r^r = \frac{n \left( n_{0,\text{case}} n_{1,\text{control}} - n_{1,\text{case}} n_{0,\text{control}} \right)}{n_0 n_1 n_{\text{case}} n_{\text{control}}}
\]

- Haplotype-based association test;

Chi-squared test on a hierarchical clustering of case/control haplotypes
Application to disease association study

- Simulate random case-control samples under various multiplicative models;
  - $GRR_1$ (first genotype relative risk ratio) = 3, 5
  - $DAF$ (disease allele frequency) = 0.05-0.15, 0.20-0.30
  - The sample generator, $gs$ (Li & Chen 2008) simulates the pattern of LD in real haplotypes.
  - 500 sample sets of 50 cases / 50 controls for each ENCODE regions have been produced.
  - The causative SNP has been removed from samples before assessment.
**Type I error in the disease association study**

<table>
<thead>
<tr>
<th>Disease model parameters</th>
<th>SS*</th>
<th>MB</th>
<th>HB</th>
<th>MDL</th>
<th>GAB</th>
<th>GPG</th>
<th>GPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAF = 5% - 15% , GRR1 = 3</td>
<td>0.26</td>
<td>0.20</td>
<td>0.17</td>
<td>0.16</td>
<td>0.17</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>DAF = 5% - 15% , GRR1 = 5</td>
<td>0.32</td>
<td>0.24</td>
<td>0.21</td>
<td>0.20</td>
<td>0.22</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>DAF = 20% - 30% , GRR1 = 3</td>
<td>0.29</td>
<td>0.21</td>
<td>0.18</td>
<td>0.17</td>
<td>0.19</td>
<td>0.14</td>
<td>0.15</td>
</tr>
<tr>
<td>DAF = 20% - 30% , GRR1 = 5</td>
<td>0.34</td>
<td>0.25</td>
<td>0.22</td>
<td>0.22</td>
<td>0.23</td>
<td>0.17</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Marker selection

- Uniform marker selection;
  the first SNP out of every k consecutive SNPs is selected as marker.

- Prioritized marker selection;
  ranking each SNPs based on its “informativeness”, then select markers with respect to the ranking in each haplotype block.
Effect of marker selection on performance of disease gene identification

- **Uniform marker selection**

![Graphs showing the effect of marker selection on performance](image)
Effect of marker selection on performance of disease gene identification

- Prioritized marker selection
Conclusion

Genotype phasing with maximum parsimony

- Incorporating a parametric greedy phasing into GA made a considerable improvement in results.

- Yet, the search space of most parsimonious haplotypes is rather complicated to be tractable by Genetic Algorithm.

- It seems that the most parsimonious haplotypes are not necessarily near to actual haplotypes, in practice.
Conclusion

Haplotype block partitioning using the global partitioning for maximal associated SNP pairs

- Methods of pairwise analysis of SNPs find blocks of limited haplotype diversity.
- There is not any general concordance among block boundaries with different methods.
- By permutation re-sampling it has been shown that the Gabriel’s method and its association index are highly robust. Our algorithm is also relatively robust.
Conclusion

- The global block partitioning methods performed best in identification of recombination hotspots.
- The block-based association test is considerably more efficient than the conventional single site association test, in case-control study.
- Our block partitioning method performed best accuracy for the case-control study, even when a low marker density is available.