Haplotype Transmission Association (HTA) — An "Importance" Measure for Selecting Genetic Markers

Tian Zheng
Department of Statistics
Columbia University

This is a joint work with Professor Shaw-Hwa Lo in the Department of Statistics at Columbia University.
Outline

- Marker selecting problem in genetic mapping for complex traits
- HTA Statistics
- Marker screening based on HTA
- Example
- Summary
Genetic Mapping:

- **Single locus diseases**: the risk of the disease is decided by the differences on one gene.

Conventional approaches—marker-wise tests

- **Common diseases**: Complex traits of common diseases are usually caused by *multiple genes* with possible interactions among them.
  
  - Marker-wise tests cannot capture presumed interactions among disease genes.
  
  - Analysis that takes into account the possible interactions among markers should be used.
Problem:

**Fine mapping:**
Large number of available genetic markers.

**Complex traits:**
Joint analysis of individual markers and their possible interactions.

The number of markers and possible interactions greatly exceed the number of patients in the study.

Solution:

Pre-select a small set of most “important” markers so that detailed analysis that involves interactions can be carried out using data with moderate size.
Marker selection for complex traits

- Start with a large set of candidate genetic markers.
- Screen out the markers with little information regarding the disease traits.
- Take into account possible interaction among the disease genes.
- Time and memory efficient.
Data: genetic information of a random sample of $n$ patients and their parents.

- $2n$ parent-patient transmission pairs
- Each pair consists of two haplotypes
- one transmitted and the other untransmitted

For $l^{th}$ pair, let $h_t^{(l)}$ be the haplotype transmitted to the diseased child, and $h_u^{(l)}$ be the untransmitted

$$n_t^i = \#(h_t^{(l)} = h_i)$$

$$n_u^i = \#(h_u^{(l)} = h_i)$$
Haplotype transmission disequilibrium (HTD)

is defined to measure the amount of linkage/LD information contained in the set of markers being tested:

$$HTD = \sum_i (n_i^t - n_i^u)^2,$$

whose expectation under the null hypothesis is equal to the trace of the Fisher’s information matrix parameterized by haplotype relative risks.
Assume $m$ markers $S_M = \{M_1, M_2, ..., M_m\}$ are being tested, to evaluate the information contributed by the $r^{th}$ marker $M_r$, which has alleles $a_r$ and $b_r$, consider $S^r_M = S_M / M_r$ ($r^{th}$-deleted marker set).

Let

$$S_r = \{h_1, h_2, ..., h_H\}$$

be the set of haplotypes spanned by $S^r_M$, and the counts $n_i^t$, and $n_i^u$ can be defined as before.
Denote by $n^t_i(a_r)$ and $n^t_i(b_r)$, the number of transmissions of the enlarged haplotypes:

\[
\begin{align*}
 h_i \\ a_r \\
 b_r
\end{align*}
\]

and

\[
\begin{align*}
 h_i \\ a_r \\
 b_r
\end{align*}
\]

respectively.

Similarly, two counts $n^u_i(a_r)$ and $n^u_i(b_r)$ are defined for the non-transmissions of the enlarged haplotypes.

It is easy to observe that

\[
\begin{align*}
 n^t_i &= n^t_i(a_r) + n^t_i(b_r) \\
 n^u_i &= n^u_i(a_r) + n^u_i(b_r).
\end{align*}
\]
**HTD** for *m* markers, \( S_M = \{M_1, M_2, \ldots, M_m\} \):

\[
HTD(m) = \sum_{h_i \in \mathcal{S}_r} (n_i^t (a_r) - n_i^u (a_r))^2 + (n_i^t (b_r) - n_i^u (b_r))^2
\]

**HTD** for the *m-1* markers in *r*-th-deleted marker set \( S_M^r = S_M / M_r \):

\[
HTD^r (m - 1) = \sum_{h_i \in \mathcal{S}_r} (n_i^t - n_i^u)^2
\]

\[
= \sum_{h_i \in \mathcal{S}_r} (n_i^t (a_r) + n_i^t (b_r) - n_i^u (a_r) - n_i^u (b_r))^2
\]

\[
= \sum_{h_i \in \mathcal{S}_r} (n_i^t (a_r) - n_i^u (a_r))^2 + (n_i^t (b_r) - n_i^u (b_r))^2
\]

\[
+ 2 \sum_{h_i \in \mathcal{S}_r} (n_i^t (a_r) - n_i^u (a_r))(n_i^t (b_r) - n_i^u (b_r))
\]

\[
= HTD(m)
\]

\[
+ 2 \sum_{h_i \in \mathcal{S}_r} (n_i^t (a_r) - n_i^u (a_r))(n_i^t (b_r) - n_i^u (b_r))
\]
Thus, the amount of information brought by marker $M_r$ can be evaluated using the $HTD$ difference—the information drop.

$$\Delta HTD^r (m - 1) = HTD^r (m - 1) - HTD(m)$$

$$= 2 \sum_{h_i \in H_r} (n_i^t (a_r) - n_i^u (a_r))(n_i^t (b_r) - n_i^u (b_r))$$

Haplotype Transmission Association (HTA) is

$$HTA^r (m) = \sum_{h_i \in H_r} (n_i^t (a_r) - n_i^u (a_r))(n_i^t (b_r) - n_i^u (b_r)) + \sum_{h_i \in H_r, a_r \parallel b_r} n(h_i \parallel h_i)$$
The properties of HTA statistic:

<table>
<thead>
<tr>
<th>Expectation of $HTA_r(m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Most important</strong></td>
</tr>
<tr>
<td>$M_r$ contributes <strong>important</strong> linkage information to the current marker set.</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
</tr>
<tr>
<td>$M_r$ only contains <strong>no</strong> linkage information and no marker in the data set is associated with any disease susceptibility loci.</td>
</tr>
<tr>
<td><strong>Zero</strong></td>
</tr>
<tr>
<td>$M_r$ contributes <strong>little</strong> linkage information but <strong>noise</strong> to the data, and dilutes the true linkage/association information carried by other markers.</td>
</tr>
<tr>
<td><strong>Positive</strong></td>
</tr>
<tr>
<td><strong>Least important</strong></td>
</tr>
</tbody>
</table>
Marker selection algorithm based on the HTA statistic

Data

\[ S_M = \{M_1, M_2, \ldots, M_K\} \]

\( K \) is the total number of markers.

\( m \) is the number of markers retained in \( S_M \). For each \( r = 1, 2, \ldots, m \), calculate \( HTA^r(m) \) for \( M_r \).

Delete the marker with the highest \( HTA^r (m) \) in \( S_M \) and continue in the loop.

Any non-negative \( HTA \)?

Yes

No

Return \( S_M \) as screening result.
Example:

a complex disease with three susceptibility loci $A$, $B$, and $E$.

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Haplotypic Relative Risk (HRR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ABE$</td>
<td>high</td>
</tr>
<tr>
<td>$ABe$</td>
<td>low</td>
</tr>
<tr>
<td>$AbE$</td>
<td>low</td>
</tr>
<tr>
<td>$Abe$</td>
<td>high</td>
</tr>
<tr>
<td>$aBE$</td>
<td>low</td>
</tr>
<tr>
<td>$aBe$</td>
<td>high</td>
</tr>
<tr>
<td>$aBe$</td>
<td>high</td>
</tr>
<tr>
<td>$abe$</td>
<td>low</td>
</tr>
</tbody>
</table>

20 markers are generated with 3 of them in strong linkage/disequilibrium with the disease loci respectively.
### Average HTA values for linked and unlinked markers

<table>
<thead>
<tr>
<th>loop</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>linked</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>2</td>
<td>14</td>
<td>36</td>
<td>93</td>
<td>174</td>
<td>390</td>
<td>805</td>
<td>1626</td>
<td>2864</td>
<td>5538</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>unlinked</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>-6</td>
<td>-6</td>
<td>-15</td>
<td>-41</td>
<td>-71</td>
<td>-114</td>
<td>-343</td>
<td>-716</td>
<td>-1568</td>
<td>-3317</td>
<td>-6563</td>
<td>-12549</td>
<td>-968</td>
<td>-1296</td>
</tr>
</tbody>
</table>

### HTD values during the screening

**Information flow**

- **+**: linked marker
- **o**: unlinked marker

Screening stopped.
Summary

HTA statistic and BHTA algorithm

- HTA measures the importance of a marker in terms of the amount of information contributed by it.

- Screening algorithm based on HTA is fast, haplotype-based. It is able to handle complicated interaction among disease loci.

Reference paper: