A Nonparametric Multipoint Screening Method for QTL Mapping

Tian Zheng
Department of Statistics
Columbia University

August 8, 2005

This is a joint work with Professor Shaw-Hwa Lo (Dept. Stat., Columbia Univ.) and Dr. Hui Wang.
Talk outline

Background on QTL mapping

Main results

QBGTA: a multipoint screening algorithm for QTLs

Simulation results

Future efforts
Complex quantitative traits

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (such as weight).
- Most human disorders have multiple quantitative traits as their symptoms. Thus, understanding the genetics behind these traits can further our understanding of the mechanism of these disorders.
- It is believed that the variation in trait such as blood pressure, cholesterol level, height, etc. can be attributed to the joint action of multiple genes (called quantitative trait loci—QTLs), each of only a modest effect marginally (individually).
Complex quantitative traits

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (such as weight).

- Most human disorders have multiple quantitative traits as their symptoms. Thus, understanding the genetics behind these traits can further our understanding of the mechanism of these disorders.

- It is believed that the variation in trait such as blood pressure, cholesterol level, height, etc. can be attributed to the joint action of **multiple** genes (called quantitative trait loci—QTLs), each of only a modest effect marginally (individually).
Complex quantitative traits

- **Quantitative traits**: as opposed to dichotomous traits, have continuous trait values (such as weight).
- Most human disorders have multiple quantitative traits as their symptoms. Thus, understanding the genetics behind these traits can further our understanding of the mechanism of these disorders.
- It is believed that the variation in trait such as blood pressure, cholesterol level, height, etc. can be attributed to the joint action of **multiple** genes (called quantitative trait loci—QTLs), each of only a modest effect marginally (individually).
Challenges of complex quantitative traits

- Searching for QTL’s in a genome scan:
  - A genome scan with markers of an average 1cM inter-marker distance would require 3,500 markers.
  - Not even a fine mapping yet, since
    \[1cM = 100,000 \text{ base pairs (bits)};\]

- Complexity—one needs to consider interactions among different markers to study such a trait.
Challenges of complex quantitative traits

- Searching for QTL’s in a genome scan:
  - A genome scan with markers of an average 1cM inter-marker distance would require 3,500 markers.
  - Not even a fine mapping yet, since
    \[ 1\text{cM} = 100,000 \text{ base pairs (bits)}; \]
  - Complexity—one needs to consider interactions among different markers to study such a trait.
Challenges of complex quantitative traits

- Searching for QTL’s in a genome scan:

  - A genome scan with markers of an average 1cM inter-marker distance would require 3,500 markers.
  - Not even a fine mapping yet, since

    
    \[1\text{cM} = 100,000 \text{ base pairs (bits)};\]

- Complexity—one needs to consider interactions among different markers to study such a trait.
Challenges of complex quantitative traits

- Searching for QTL’s in a genome scan:
  - A genome scan with markers of an average 1cM inter-marker distance would require 3,500 markers.
  - Not even a fine mapping yet, since
    
    \[ 1\text{cM} = 100,000 \text{ base pairs (bits)}; \]

- Complexity—one needs to consider interactions among different markers to study such a trait.
Current efforts in association mapping of QTL

- In most current genome scans, association statistics are usually calculated at individual marker loci. Then detailed modelling of “significant” markers is carried out.
- Due to their weak marginal effects, important QTL’s might not be detected using such strategies.
Current efforts in association mapping of QTL

- In most current genome scans, association statistics are usually calculated at individual marker loci. Then detailed modelling of “significant” markers is carried out.
- Due to their weak marginal effects, important QTL’s might not be detected using such strategies.
An original and powerful Transmission Disequilibrium Test (TDT) was proposed by Spielman et al. (1993) for dichotomous traits.
QTDT by Xiong et al. (1998)

- Extension to quantitative traits

\[
QTDT = (\bar{Y}_M - \bar{Y}_m)^2 / \left( \frac{1}{n_M} + \frac{1}{n_m} \right) S^2
\]

where

- \( \bar{Y}_M (\bar{Y}_m) \) — trait mean among progenies who have inherited allele \( M (m) \) from heterozygous parents;
- \( n_M (n_m) \) is the count of transmissions for allele \( M (m) \);
- and \( S^2 = \sum_k [(Y_{Mk} - \bar{Y}_M)^2 + (Y_{mk} - \bar{Y}_m)^2] / (n_M + n_m - 2) \).

- Under the null hypothesis of no association and normality assumption, QTDT follows a \( \chi^2 \) distribution with 1 d.f.
- A nonparametric extension of this test would be the Wilcoxon signed rank test under the same data setup.
QTDT by Xiong et al. (1998)

- Extension to quantitative traits

\[
\text{QTDT} = \frac{(\bar{Y}_M - \bar{Y}_m)^2}{\left(\frac{1}{n_M} + \frac{1}{n_m}\right)S^2}
\]

where

- \( \bar{Y}_M (\bar{Y}_m) \) — trait mean among progenies who have inherited allele \( M (m) \) from heterozygous parents;
- \( n_M (n_m) \) is the count of transmissions for allele \( M (m) \).
- and \( S^2 = \frac{\sum_k [(Y_{Mk} - \bar{Y}_M)^2 + (Y_{mk} - \bar{Y}_m)^2]}{n_M + n_m - 2} \).

- Under the null hypothesis of no association and normality assumption, QTDT follows a \( \chi^2 \) distribution with 1 d.f.
- A nonparametric extension of this test would be the Wilcoxon signed rank test under the same data setup.
A Nonparametric Multipoint Screening Method for QTL Mapping

- Background on QTL mapping
- Current efforts in association mapping of QTL

**QTDT by Xiong et al. (1998)**

- **Extension to quantitative traits**

\[
\text{QTDT} = (\bar{Y}_M - \bar{Y}_m)^2 / \left( \frac{1}{n_M} + \frac{1}{n_m} \right) S^2
\]

where

- \( \bar{Y}_M (\bar{Y}_m) \) — trait mean among progenies who have inherited allele \( M \) (\( m \)) from heterozygous parents;
- \( n_M (n_m) \) is the count of transmissions for allele \( M \) (\( m \)).
- \( S^2 = \sum_k \left[ (Y_{Mk} - \bar{Y}_M)^2 + (Y_{mk} - \bar{Y}_m)^2 \right] / (n_M + n_m - 2) \).

- Under the null hypothesis of no association and normality assumption, QTDT follows a \( \chi^2 \) distribution with 1 d.f.
- A nonparametric extension of this test would be the Wilcoxon signed rank test under the same data setup.
QTDT by Xiong et al. (1998)

Extension to quantitative traits

\[ \text{QTDT} = (\bar{Y}_M - \bar{Y}_m)^2 / \left( \frac{1}{n_M} + \frac{1}{n_m} \right) S^2 \]

where

- \( \bar{Y}_M (\bar{Y}_m) \) — trait mean among progenies who have inherited allele \( M (m) \) from heterozygous parents;
- \( n_M (n_m) \) is the count of transmissions for allele \( M (m) \).
- and \( S^2 = \sum_k [(Y_{Mk} - \bar{Y}_M)^2 + (Y_{mk} - \bar{Y}_m)^2] / (n_M + n_m - 2) \).

Under the null hypothesis of no association and normality assumption, QTDT follows a \( \chi^2 \) distribution with 1 d.f.

A nonparametric extension of this test would be the Wilcoxon signed rank test under the same data setup.
QTDT by Xiong et al. (1998)

- Extension to quantitative traits

\[ QTDT = \left( \bar{Y}_M - \bar{Y}_m \right)^2 / \left( \frac{1}{n_M} + \frac{1}{n_m} \right) S^2 \]

where

- \( \bar{Y}_M (\bar{Y}_m) \) — trait mean among progenies who have inherited allele \( M (m) \) from heterozygous parents;
- \( n_M (n_m) \) is the count of transmissions for allele \( M (m) \);
- and \( S^2 = \sum_k \left[ \frac{(Y_{Mk} - \bar{Y}_M)^2 + (Y_{mk} - \bar{Y}_m)^2}{n_M + n_m - 2} \right] \).

- Under the null hypothesis of no association and normality assumption, QTDT follows a \( \chi^2 \) distribution with 1 d.f.

- A nonparametric extension of this test would be the Wilcoxon signed rank test under the same data setup.
QTDT by Xiong et al. (1998)

Extension to quantitative traits

\[ \text{QTDT} = (\bar{Y}_M - \bar{Y}_m)^2 / (\frac{1}{n_M} + \frac{1}{n_m}) S^2 \]

where

- \( \bar{Y}_M (\bar{Y}_m) \) — trait mean among progenies who have inherited allele \( M (m) \) from heterozygous parents;
- \( n_M (n_m) \) is the count of transmissions for allele \( M (m) \);
- \( S^2 = \sum_k [(Y_{Mk} - \bar{Y}_M)^2 + (Y_{mk} - \bar{Y}_m)^2] / (n_M + n_m - 2) \).

Under the null hypothesis of no association and normality assumption, QTDT follows a \( \chi^2 \) distribution with 1 d.f.

A nonparametric extension of this test would be the Wilcoxon signed rank test under the same data setup.
Backward genotype-trait association algorithm for QTLs (QBGTA)

- **Multi-locus genotypes** are used as “study units” to create a global view of the interactions among different regions on the genome.
- Use a measure of trait association information that is robust to departure from normality.
- It screens out markers that do not contain much information of the trait in a **backward** fashion.
- It is proposed for population-based design.
- Can be easily extended to other designs and data types.
Backward genotype-trait association algorithm for QTLs (QBGTA)

- Multi-locus genotypes are used as “study units” to create a global view of the interactions among different regions on the genome.
- Use a measure of trait association information that is robust to departure from normality.
- It screens out markers that do not contain much information of the trait in a backward fashion.
- It is proposed for population-based design.
- Can be easily extended to other designs and data types.
Backward genotype-trait association algorithm for QTLs (QBGTA)

- **Multi-locus genotypes** are used as “study units” to create a global view of the interactions among different regions on the genome.
- Use a measure of trait **association** information that is robust to departure from normality.
- It screens out markers that do not contain much information of the trait in a **backward** fashion.
- It is proposed for population-based design.
- Can be easily extended to other designs and data types.
Backward genotype-trait association algorithm for QTLs (QBGTA)

- **Multi-locus genotypes** are used as “study units” to create a global view of the interactions among different regions on the genome.

- Use a measure of trait **association** information that is robust to departure from normality.

- It screens out markers that do not contain much information of the trait in a **backward** fashion.

- It is proposed for population-based design.

- Can be easily extended to other designs and data types.
Backward genotype-trait association algorithm for QTLs (QBGTA)

- Multi-locus genotypes are used as “study units” to create a global view of the interactions among different regions on the genome.
- Use a measure of trait association information that is robust to departure from normality.
- It screens out markers that do not contain much information of the trait in a backward fashion.
- It is proposed for population-based design.
- Can be easily extended to other designs and data types.
We assume a disease with three susceptibility QTL’s, each with alleles $Q_i$ and $q_i$ (mutated), $i=1,2,3$. An individual needs only any one of these QTLs being homozygous of the mutated $q_i$ to have the alleviated trait values.
Main results

Example

Power comparison

<table>
<thead>
<tr>
<th>Category</th>
<th>Average Power (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave. power in detecting associated markers</td>
<td>80.95% (QBGTA)</td>
</tr>
<tr>
<td></td>
<td>59.52% (QTDT)</td>
</tr>
<tr>
<td>Power in detecting disease loci jointly</td>
<td>42.86% (QBGTA)</td>
</tr>
<tr>
<td></td>
<td>21.43% (QTDT)</td>
</tr>
<tr>
<td>Ave. number of disease loci detected</td>
<td>2.43 (QBGTA)</td>
</tr>
<tr>
<td></td>
<td>1.79 (QTDT)</td>
</tr>
</tbody>
</table>
Consider $k$ markers $M_i$, $i = 1, \ldots, k$, each with 2 alleles, $a_i$ and $b_i$, (possible values of a marker). Each individual has two copies for each marker.

<table>
<thead>
<tr>
<th>Indiv</th>
<th>Trait</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y_1$</td>
<td>$a_1</td>
<td>b_1$</td>
<td>$a_2</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>$n$</td>
<td>$y_n$</td>
<td>$b_1</td>
<td>b_1$</td>
<td>$a_2</td>
</tr>
</tbody>
</table>

**Genotype Coding**

$$
\begin{bmatrix}
  a_1|b_1 \\
a_2|a_2 \\
\vdots \\
b_k|b_k
\end{bmatrix}
\text{def}
\begin{bmatrix}
  1 \\
  2 \\
  \vdots \\
  0
\end{bmatrix}$$
Population-based QTL data

Consider \( k \) markers \( M_i, i = 1, \ldots, k \), each with 2 alleles, \( a_i \) and \( b_i \), (possible values of a marker). Each individual has two copies for each marker.

<table>
<thead>
<tr>
<th>Indiv</th>
<th>Trait</th>
<th>( M_1 )</th>
<th>( M_2 )</th>
<th>( \ldots )</th>
<th>( M_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( y_1 )</td>
<td>( a_1</td>
<td>b_1 )</td>
<td>( a_2</td>
<td>a_2 )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>( b_1</td>
<td>b_1 )</td>
<td>( a_2</td>
<td>b_2 )</td>
</tr>
<tr>
<td>( n )</td>
<td>( y_n )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Genotype Coding

\[
\begin{bmatrix}
  a_1 | b_1 \\
  a_2 | a_2 \\
  \vdots \\
  b_k | b_k \\
\end{bmatrix}
\begin{bmatrix}
  1 \\
  2 \\
  \vdots \\
  0 \\
\end{bmatrix} \overset{\text{def}}{=} \begin{bmatrix}
  1 \\
  2 \\
  \vdots \\
  0 \\
\end{bmatrix}
\]
A statistical view at the problem

- Assume there are \( D \) QTLs affecting the trait.
- Suppose that, totally, there are \( L \) possible genotypes on these QTLs, and denote \( g_{l}^{D} \) as the \( l \)th of them.
- Let \( f(y|g_{l}^{D}) \) denote the trait distribution corresponding to genotype \( g_{l}^{D} \).
- The population distribution of the trait is then

\[
f(y) = \sum_{l=1}^{L} f(y|g_{l}^{D})P(g_{l}^{D}).
\]
A statistical view at the problem

▶ Assume there are $D$ QTLs affecting the trait.
▶ Suppose that, totally, there are $L$ possible genotypes on these QTLs, and denote $g_i^D$ as the $l$th of them.
▶ Let $f(y|g_i^D)$ denote the trait distribution corresponding to genotype $g_i^D$.
▶ The population distribution of the trait is then

$$f(y) = \sum_{l=1}^{L} f(y|g_i^D)P(g_i^D).$$
A statistical view at the problem

- Assume there are $D$ QTLs affecting the trait.
- Suppose that, totally, there are $L$ possible genotypes on these QTLs, and denote $g_i^D$ as the $l$th of them.
- Let $f(y|g_i^D)$ denote the trait distribution corresponding to genotype $g_i^D$.
- The population distribution of the trait is then

$$f(y) = \sum_{l=1}^{L} f(y|g_i^D)P(g_i^D).$$
A statistical view at the problem

- Assume there are $D$ QTLs affecting the trait.
- Suppose that, totally, there are $L$ possible genotypes on these QTLs, and denote $g^D_l$ as the $l$th of them.
- Let $f(y \mid g^D_l)$ denote the trait distribution corresponding to genotype $g^D_l$.
- The population distribution of the trait is then

$$f(y) = \sum_{l=1}^{L} f(y \mid g^D_l)P(g^D_l).$$
A statistical view at the problem

- **k markers**, each with three possible genotypes (0, 1, 2), can generate $3^k$ possible genotypes.
- The number of different genotypes observed, defined as $G_k$, is often much smaller, and bounded by $n$.
- We define $G = \{g_1^{(k)}, g_2^{(k)}, \ldots, g_{G_k}^{(k)}\}$ as the set of observed genotypes spanned by the $k$ underlying markers.
- Conditioning on an observed genotype of an individual, say $g_i^{(k)}$,

$$f(y | g_i^{(k)}) = \sum_{l=1}^{L} f(y | g_l^D, g_i^{(k)})P(g_l^D | g_i^{(k)})$$

$$= \sum_{l=1}^{L} f(y | g_l^D)P(g_l^D | g_i^{(k)}).$$
A statistical view at the problem

- **$k$ markers**, each with three possible genotypes (0, 1, 2), can generate $3^k$ possible genotypes.
- The number of different genotypes observed, defined as $G_k$, is often much smaller, and bounded by $n$.
- We define $\mathbb{G} = \{g_1^{(k)}, g_2^{(k)}, \ldots, g_{G_k}^{(k)}\}$ as the set of observed genotypes spanned by the $k$ underlying markers.
- Conditioning on an observed genotype of an individual, say $g_i^{(k)}$,

\[
 f(y|g_i^{(k)}) = \sum_{l=1}^{L} f(y|g_i^D, g_i^{(k)})P(g_i^D|g_i^{(k)})
\]
\[
 = \sum_{l=1}^{L} f(y|g_i^D)P(g_i^D|g_i^{(k)}).
\]
A statistical view at the problem

- **k markers**, each with three possible genotypes (0, 1, 2), can generate $3^k$ possible genotypes.
- The number of different genotypes observed, defined as $G_k$, is often much smaller, and bounded by $n$.
- We define $G = \{g_1^{(k)}, g_2^{(k)}, \ldots, g_{G_k}^{(k)}\}$ as the set of observed genotypes spanned by the $k$ underlying markers.
- Conditioning on an observed genotype of an individual, say $g_i^{(k)}$,

\[
 f(y|g_i^{(k)}) = \sum_{l=1}^{L} f(y|g_l^D, g_i^{(k)})P(g_l^D|g_i^{(k)})
\]

\[
 = \sum_{l=1}^{L} f(y|g_l^D)P(g_l^D|g_i^{(k)}).
\]
A statistical view at the problem

- \( k \) markers, each with three possible genotypes (0, 1, 2), can generate \( 3^k \) possible genotypes.
- The number of different genotypes observed, defined as \( G_k \), is often much smaller, and bounded by \( n \).
- We define \( G = \{ g_1^{(k)}, g_2^{(k)}, \ldots, g_{G_k}^{(k)} \} \) as the set of observed genotypes spanned by the \( k \) underlying markers.
- Conditioning on an observed genotype of an individual, say \( g_i^{(k)} \),

\[
f(y|g_i^{(k)}) = \sum_{l=1}^{L} f(y|g_l^D, g_i^{(k)})P(g_l^D|g_i^{(k)})
\]

\[
= \sum_{l=1}^{L} f(y|g_l^D)P(g_l^D|g_i^{(k)}).
\]
A statistical view at the problem

If none of the markers, $M_1, \ldots, M_k$ is associated with the trait QTLs, i.e., $P(g_i^D | g_i^{(k)}) \equiv P(g_i^D)$, $l = 1, \ldots, L$, then

$$f(y | g_i^{(k)}) = \sum_{l=1}^{L} f(y | g_i^D) P(g_i^D | g_i^{(k)})$$

$$= \sum_{l=1}^{L} f(y | g_i^D) P(g_i^D)$$

$$= f(y).$$

This indicates that the conditional distributions, $f(y | g_i^{(k)})$'s, are uniformly equal to the population distribution $f(y)$. 
A statistical view at the problem

If none of the markers, $M_1, \ldots, M_k$ is associated with the trait QTLs, i.e., $\mathbb{P}(g_i^D | g_i^{(k)}) \equiv \mathbb{P}(g_i^D)$, $l = 1, \ldots, L$, then

$$f(y | g_i^{(k)}) = \sum_{l=1}^{L} f(y | g_i^D) \mathbb{P}(g_i^D | g_i^{(k)})$$

$$= \sum_{l=1}^{L} f(y | g_i^D) \mathbb{P}(g_i^D)$$

$$= f(y).$$

This indicates that the conditional distributions, $f(y | g_i^{(k)})$'s, are uniformly equal to the population distribution $f(y)$. 
A statistical view at the problem

The difference between the **conditional trait distribution** given a genotype profile of the markers and the **population trait distribution**

\[
f(y \mid g_{i}^{(k)}) - f(y) = \sum_{l=1}^{L} f(y \mid g_{i}^{D})P(g_{i}^{D} \mid g_{i}^{(k)}) - \sum_{l=1}^{L} f(y \mid g_{i}^{D})P(g_{i}^{D})
\]

\[
= \sum_{l=1}^{L} [P(g_{i}^{D} \mid g_{i}^{(k)}) - P(g_{i}^{D})] \times f(y \mid g_{i}^{D})
\]

captures the trait variation information through the association between the markers and QTL’s.
The statistical problem

The problem of searching for association between the markers and the trait can be translated into a test:

**The hypotheses**

\[ H_0 : f(y | g_1^{(k)}) = f(y | g_2^{(k)}) = \cdots = f(y | g_{G_k}^{(k)}) = f(y) \]

\[ H_a : f(y | g_i^{(k)}) \neq f(y), \text{ for some } i's \in \{1, 2, \ldots, G_k\} \]

Screening of a large number of markers should be viewed as a preliminary inferential process.
The statistical problem

The problem of searching for association between the markers and the trait can be translated into a test:

The hypotheses

\[ H_0 : f(y | g_1^{(k)}) = f(y | g_2^{(k)}) = \cdots = f(y | g_{G_k}^{(k)}) = f(y) \]
\[ H_a : f(y | g_i^{(k)}) \neq f(y), \text{ for some } i's \in \{1, 2, \ldots, G_k\} \]

Screening of a large number of markers should be viewed as a preliminary inferential process.
Rank sums by genotypes

- Assume that a sample of \( n \) individuals is genotyped.
- For the \( j \)th individual, \( \gamma_j^{(k)} \) \((\in \mathbb{G})\) denotes his/her k-locus genotype, and \( y_j \) denotes his/her quantitative trait value.
- Each individual is assigned a rank \( R_j \) based on his/her trait value.
- To evaluate markers \( M_1, \ldots, M_k \), with genotypes \( g_i^{(k)} \), \( i = 1, \ldots, G_k \), we consider the rank sums by genotypes,

\[
W_i^{(k)} = \sum_{\gamma_j^{(k)} = g_i^{(k)}} R_j, \quad i = 1, 2, \ldots, G_k.
\]
A Nonparametric Multipoint Screening Method for QTL Mapping

QBGTA: a multipoint screening algorithm for QTLs

The main statistics

Rank sums by genotypes

- Assume that a sample of $n$ individuals is genotyped.
- For the $j$th individual, $\gamma_j^{(k)} \in \mathbb{G}$ denotes his/her $k$-locus genotype, and $y_j$ denotes his/her quantitative trait value.
- Each individual is assigned a rank $R_j$ based on his/her trait value.
- To evaluate markers $M_1, \ldots, M_k$, with genotypes $g_i^{(k)}$, $i = 1, \ldots, G_k$, we consider the rank sums by genotypes,

$$W_i^{(k)} = \sum_{\gamma_j^{(k)} = g_i^{(k)}} R_j, \quad i = 1, 2, \ldots, G_k.$$
The main statistics

Rank sums by genotypes

- Assume that a sample of \( n \) individuals is genotyped.
- For the \( j \)th individual, \( \gamma_j^{(k)} \) (\( \in \mathcal{G} \)) denotes his/her \( k \)-locus genotype, and \( y_j \) denotes his/her quantitative trait value.
- Each individual is assigned a rank \( R_j \) based on his/her trait value.
- To evaluate markers \( M_1, \ldots, M_k \), with genotypes \( g_i^{(k)} \), \( i = 1, \ldots, G_k \), we consider the rank sums by genotypes,

\[
W_i^{(k)} = \sum_{\gamma_j^{(k)}=g_i^{(k)}} R_j, \quad i = 1, 2, \ldots, G_k.
\]
Rank sums by genotypes

- Assume that a sample of \( n \) individuals is genotyped.
- For the \( j \)th individual, \( \gamma_j^{(k)} (\in G) \) denotes his/her k-locus genotype, and \( y_j \) denotes his/her quantitative trait value.
- Each individual is assigned a rank \( R_j \) based on his/her trait value.
- To evaluate markers \( M_1, \ldots, M_k \), with genotypes \( g_i^{(k)} \), \( i = 1, \ldots, G_k \), we consider the rank sums by genotypes,

\[
W_i^{(k)} = \sum_{\gamma_j^{(k)} = g_i^{(k)}} R_j, \quad i = 1, 2, \ldots, G_k.
\]
Information measure

Quantitative Genotype-Trait Disequilibriums (QGTD)

\[ QGTD^{(k)} = \sum_{i=1}^{G_k} \left( W_i^{(k)} - \frac{n + 1}{2} n_i^{(k)} \right)^2, \quad (n = \sum_{i=1}^{G_k} n_i^{(k)}) \]

where the superscript \( k \) indicates the size of current marker set on which \( QGTD^{(k)} \) is defined.
Similar to the "rank-sum variance", the score \( QGTD^{(k)} \) reflects the variation of the phenotype trait across different genotypes. The stronger the association is between the QTLs and the markers, the larger the value of \( QGTD^{(k)} \) is.
Importance measure

▶ The information content of the set \{M_1, \ldots, M_k\} measured by QGTD will be high as long as some of them are in strong association with the trait QTLs.
▶ We call these markers “important markers”.
▶ To identify these “important markers”, one needs to screen out unimportant markers.

To evaluate the importance of the \(r\)th marker, \(M_r\), given the rest of the marker set, we define:

Quantitative Genotype-Trait Association (QGTA)

\[
QGTA(r) = \frac{1}{2} \left( QGTD^{(k-1)}(M_r \text{ removed}) - QGTD^{(k)} \right) + \tilde{A}
\]
Importance measure

- The information content of the set \( \{ M_1, \ldots, M_k \} \) measured by QGTD will be high as long as some of them are in strong association with the trait QTLs.

- We call these markers "important markers".

- To identify these "important markers", one needs to screen out unimportant markers.

To evaluate the importance of the \( r^{th} \) marker, \( M_r \), given the rest of the marker set, we define:

**Quantitative Genotype-Trait Association (QGTA)**

\[
QGTA(r) = \frac{1}{2} \left( QGTD^{(k-1)}(M_r \text{ removed}) - QGTD^{(k)} \right) + \tilde{A}
\]
A Nonparametric Multipoint Screening Method for QTL Mapping

QBGTA: a multipoint screening algorithm for QTLs

The main statistics

Importance measure

- The information content of the set $\{M_1, \ldots, M_k\}$ measured by QGTD will be high as long as some of them are in strong association with the trait QTLs.
- We call these markers “important markers”.
- To identify these “important markers”, one needs to screen out unimportant markers.

To evaluate the importance of the $r^{th}$ marker, $M_r$, given the rest of the marker set, we define:

Quantitative Genotype-Trait Association (QGTA)

$$QGTA(r) = \frac{1}{2} \left( QGTD^{(k-1)}(M_r \text{ removed}) - QGTD^{(k)} \right) + \tilde{A}$$
Importance measure

- The information content of the set \( \{M_1, \ldots, M_k\} \) measured by QGTD will be high as long as some of them are in strong association with the trait QTLs.
- We call these markers “important markers”.
- To identify these “important markers”, one needs to screen out unimportant markers.

To evaluate the importance of the \( r \)th marker, \( M_r \), given the rest of the marker set, we define:

\[
QGTA(r) = \frac{1}{2} \left( QGTD^{(k-1)}(M_r \text{ removed}) - QGTD^{(k)} \right) + \tilde{A}
\]
Importance measure

- The information content of the set \( \{ M_1, \ldots, M_k \} \) measured by QGTD will be high as long as some of them are in strong association with the trait QTLs.

- We call these markers “important markers”.

- To identify these “important markers”, one needs to screen out unimportant markers.

To evaluate the importance of the \( r \)th marker, \( M_r \), given the rest of the marker set, we define:

Quantitative Genotype-Trait Association (QGTA)

\[
QGTA(r) = \frac{1}{2} \left( QGTD^{(k-1)}(M_r \text{ removed}) - QGTD^{(k)} \right) + \tilde{A}
\]
Importance measure

Quantitative Genotype-Trait Association (QGTA)

\[ \text{QGTA}(r) = \frac{1}{2} \left( \text{QGTD}^{(k-1)}(M_r \text{ removed}) - \text{QGTD}^{(k)} \right) + \tilde{A} \]

Properties of QGTA

QGTA\((r)\) has expectation zero under the null hypothesis.

If one or more of the current markers is in association with the trait,

- the expected value of QGTA\((r)\) will be negative if \(M_r\) is important i.e., associated with the QTLs;
- the expected value of QGTA\((r)\) will be positive if \(M_r\) is unimportant i.e., not associated with the QTLs;
Importance measure

Quantitative Genotype-Trait Association (QGTA)

\[ QGTA(r) = \frac{1}{2} \left( QGTD^{(k-1)}(M_r \text{ removed}) - QGTD^{(k)} \right) + \tilde{A} \]

Properties of QGTA

QGTA(r) has expectation zero under the null hypothesis. If one or more of the current markers is in association with the trait,

- the expected value of QGTA(r) will be **negative** if \( M_r \) is **important** i.e., associated with the QTLs;
- the expected value of QGTA(r) will be **positive** if \( M_r \) is **unimportant** i.e., not associated with the QTLs;
A Nonparametric Multipoint Screening Method for QTL Mapping

QBGTA: a multipoint screening algorithm for QTLs

The algorithm

**QBGTA algorithm**

1. Evaluate all current markers based on their QGTA(r) scores;
2. Stop if all the current markers have negative QGTA(r); Otherwise, delete the marker with the highest non-negative QGTA(r);
3. Stop if no marker remains, otherwise return to 1.

**Issue of sparseness**

When the number of possible genotypes is much larger than the number of individuals, QGTA is zero due to the sparseness of the rank-genotype table.
A Nonparametric Multipoint Screening Method for QTL Mapping

- QBGTA: a multipoint screening algorithm for QTLs
- The algorithm

**QBGTA algorithm**

1. Evaluate all current markers based on their QGTA(r) scores;
2. Stop if all the current markers have negative QGTA(r); Otherwise, delete the marker with the highest non-negative QGTA(r);
3. Stop if no marker remains, otherwise return to 1.

**Issue of sparseness**

When the number of possible genotypes is much larger than the number of individuals, QGTA is zero due to the sparseness of the rank-genotype table.
QBGTA: a multipoint screening algorithm for QTLs

The algorithm

1. Evaluate all current markers based on their QGTA(r) scores;
2. Stop if all the current markers have negative QGTA(r); Otherwise, delete the marker with the highest non-negative QGTA(r);
3. Stop if no marker remains, otherwise return to 1.

Issue of sparseness

When the number of possible genotypes is much larger than the number of individuals, QGTA is zero due to the sparseness of the rank-genotype table.
A Nonparametric Multipoint Screening Method for QTL Mapping

QBGTA: a multipoint screening algorithm for QTLs

The algorithm

**QBGTA algorithm**

1. Evaluate all current markers based on their QGTA(r) scores;
2. Stop if all the current markers have negative QGTA(r); Otherwise, delete the marker with the highest non-negative QGTA(r);
3. Stop if no marker remains, otherwise return to 1.

**Issue of sparseness**

When the number of possible genotypes is much larger than the number of individuals, QGTA is zero due to the sparseness of the rank-genotype table.
A Nonparametric Multipoint Screening Method for QTL Mapping

QBGTA: a multipoint screening algorithm for QTLs

The algorithm

**QBGTA algorithm**

1. Evaluate all current markers based on their QGTA(r) scores;
2. Stop if all the current markers have negative QGTA(r); Otherwise, delete the marker with the highest non-negative QGTA(r);
3. Stop if no marker remains, otherwise return to 1.

**Issue of sparseness**

When the number of possible genotypes is much larger than the number of individuals, QGTA is zero due to the sparseness of the rank-genotype table.
A Nonparametric Multipoint Screening Method for QTL Mapping

QBGTA: a multipoint screening algorithm for QTLs

The algorithm

QBGTA-based screening algorithm

In most of the computational examples we have run, n=400 individuals can be used to informatively screen no more than 10 markers at a time. In this case, the screening of a large number of markers, say $K = 5000$ should be carried out as follows:

Random subset screening

- Select a random set of $k$ markers (say $k=10$) out of $K$ candidate markers (say $K = 5000$), and run QBGTA algorithm on this subset of markers.
- Repeat the random subset screening a large number $B$ (say $B=100,000$) of times and record the result of each subset screening (returned “important” markers).
- Rank each marker according to the number of times that it is returned as “important”.
QBGTA-based screening algorithm

In most of the computational examples we have run, n=400 individuals can be used to informatively screen no more than 10 markers at a time. In this case, the screening of a large number of markers, say $K = 5000$ should be carried out as follows:

**Random subset screening**

- Select a random set of $k$ markers (say $k=10$) out of $K$ candidate markers (say $K = 5000$), and run QBGTA algorithm on this subset of markers.
- Repeat the random subset screening a large number $B$ (say $B=100,000$) of times and record the result of each subset screening (returned “important” markers).
- Rank each marker according to the number of times that it is returned as “important”.

QBGTA: a multipoint screening algorithm for QTLs

The algorithm
QBGTA-based screening algorithm

In most of the computational examples we have run, \( n = 400 \) individuals can be used to informatively screen no more than 10 markers at a time. In this case, the screening of a large number of markers, say \( K = 5000 \) should be carried out as follows:

Random subset screening

- Select a random set of \( k \) markers (say \( k = 10 \)) out of \( K \) candidate markers (say \( K = 5000 \)), and run QBGTA algorithm on this subset of markers.

- Repeat the random subset screening a large number \( B \) (say \( B = 100,000 \)) of times and record the result of each subset screening (returned “important” markers).

- Rank each marker according to the number of times that it is returned as “important”. 
In most of the computational examples we have run, n=400 individuals can be used to informatively screen no more than 10 markers at a time. In this case, the screening of a large number of markers, say $K = 5000$ should be carried out as follows:

**Random subset screening**

- Select a random set of $k$ markers (say $k=10$) out of $K$ candidate markers (say $K = 5000$), and run QBGTA algorithm on this subset of markers.
- Repeat the random subset screening a large number $B$ (say $B=100,000$) of times and record the result of each subset screening (returned “important” markers).
- Rank each marker according to the number of times that it is returned as “important”.
The selection criterion

- Sorted markers
- Return frequencies

(A) Observed return frequencies
(B) 2-component normal mixture (p=0.215)
- Normal comp I: N(1387, 164)
- Normal comp II: N(1721, 252)
- 99 percentile, comp I: N(1387, 164)
Example 1: Epistatic Genes

This is an extreme model where the marker-by-marker strategy won’t work at all.

<table>
<thead>
<tr>
<th>Genotype at locus B</th>
<th>Genotype at locus A</th>
<th>Marginal effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$A_1A_1$</td>
<td>$A_1A_2$</td>
</tr>
<tr>
<td>$B_1B_1$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$B_1B_2$</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>$B_2B_2$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Marginal effect</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

30 markers are simulated and 2 of them are in association with the two QTLs, respectively ($\theta = 0.01$, and LD=0.95). With sample size of 1000, QBGTA has an average power of 95% in detecting genes A and B jointly.
Example 2: Mendelian Inheritance

Simulation results
Example 2: Mendelian Inheritance

Four simulations in the Mendelian Model
Future efforts

- Combining linkage and association information during the screening.
- Extension to other study design such as affected sib pairs.
Association versus linkage (in genetic sense)

By searching for strong association signals using a fine map of markers, association methods can prove to be more powerful.