

Information-driven screening strategies for complex traits

Tian Zheng
Department of Statistics
Columbia University

November 11, 2005

Acknowledgements

- ▶ <http://statgene.stat.columbia.edu>
- ▶ Professor Shaw-Hwa Lo
- ▶ Graduate students: Iuliana Ionita, Yuejing Ding
- ▶ Our programmer: Lei cong
- ▶ Former members: Hui Wang, Xin Yan
- ▶ We also thank the Whitehead institute, MIT for allowing us to use their data for method demonstrations.
- ▶ The research presented is, in part, supported by NIH and NSF.

The challenges of Complex Traits

- ▶ Complex traits – “... are caused by multiple genes *interacting* with each other and with environmental factors to create a *gradient of genetic susceptibility* to disease.” (Weeks and Lathrop 1995)
- ▶ The dilemma:
 - ▶ Detailed study including interactions among the markers.
 - ▶ Moderate size of individuals.
- ▶ One solution: **marker selection** based on association information before any detailed study on selected important markers.
- ▶ For studies on complex traits, such marker selection should take into consideration possible interaction among loci.

Information-driven evaluation strategy

A backward screening algorithm that

- ▶ applies an association information measure with respect to the disease on a set of markers under study;
- ▶ reduces the set by deleting an unimportant marker that contributes the least to the association information contained in the set;
- ▶ stops when deleting any of the remaining markers will result in a substantial loss of information.
- ▶ The algorithm is ran on **random marker subsets** of a size that can be handled by the size of patients.
- ▶ Final selection is based on the aggregated screening results on a large number of random subsets.

Multi-locus information

- ▶ In a mapping study, genetic information is not collected independently. Instead, for one individual, the collected data contains genotypic information from hundreds of markers on the genome simultaneously.
- ▶ For example, in the case-parent trio design, the differences between the transmitted haplotypes and untransmitted haplotypes contains information that can reveal possible epistatic structure among the disease loci.
- ▶ For population-based case-control studies, unphased multi-locus genotypes should be used to extract interaction information across difference genomic loci.

Case study: BGTA

- ▶ In the following, I will use case-control studies as an example.
- ▶ We have proposed the backward genotype-trait association (BGTA) algorithm.

Genotype Data

For example, 4 diallelic markers (each with alleles, a_1/b_1 , a_2/b_2 , a_3/b_3 and a_4/b_4) are genotyped for one individual participating in the study. An observed four-marker genotype above can be represented by a four-dimension vector with each entry value as the number of allele b_i , $i = 1, 2, 3, 4$.

$$\text{unphased genotype} \implies \text{coded genotype data}$$

$$\begin{pmatrix} \{a_1, a_1\} \\ \{a_2, b_2\} \\ \{a_3, b_3\} \\ \{b_4, b_4\} \end{pmatrix} \implies \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \end{pmatrix}$$

Notation

Consider k diallelic markers M_1, M_2, \dots, M_k . k diallelic markers generate a total of 3^k possible genotypes. Define

$$\mathbb{G}^{(k)} = \{g_1^{(k)}, g_2^{(k)}, \dots, g_{3^k}^{(k)}\}.$$

$g_i^{(k)}$: the i^{th} genotype on k markers;

$n_{d,i}^{(k)}$: the count of $g_i^{(k)}$ among cases;

$n_{u,i}^{(k)}$: the count of $g_i^{(k)}$ among controls;

$P(g_i^{(k)}|D)$ or p_i^d : the frequencies (proportions) of $g_i^{(k)}$ among cases;

$P(g_i^{(k)}|U)$ or p_i^u : the frequencies (proportions) of $g_i^{(k)}$ among controls;

Association between marker genotypes and disease trait

- ▶ Consider N disease trait loci and the **disease genotypes** at these loci, g_l^D , $l = 1, 2, \dots, L$.
- ▶ Further define their population frequency $\Pr(g_l^D)$ and disease penetrance $\Pr(D|g_l^D)$
- ▶ The population disease prevalence is then $\Pr(D) = \sum_{l=1}^L \Pr(D|g_l^D)\Pr(g_l^D)$.
- ▶ For **marker genotype** $g_i^{(k)} \in \mathbb{G}$,
 $\Pr(D|g_i^{(k)}) = \sum_{l=1}^L \Pr(D|g_l^D)\Pr(g_l^D|g_i^{(k)})$.

Association between marker genotypes and disease trait

- ▶ If *none* of the k markers are in association with the disease loci, i.e.,

$$\Pr(g_l^D | g_i^{(k)}) \equiv \Pr(g_l^D), \quad l = 1, 2, \dots, L, \quad i = 1, 2, \dots, 3^k,$$

- ▶ then

$$\begin{aligned} \Pr(D | g_i^{(k)}) &= \sum_{l=1}^L \Pr(D | g_l^D) \Pr(g_l^D | g_i^{(k)}) \\ &= \sum_{l=1}^L \Pr(D | g_l^D) \Pr(g_l^D) \\ &= \Pr(D). \end{aligned}$$

Association between marker genotypes and disease trait

- ▶ If one or more markers are in association with the trait, such *association* must be reflected in the fact that, for some $g_i^{(k)}$, $\Pr(D|g_i^{(k)}) \neq \Pr(D)$.
- ▶ Detecting association between the markers and the disease can be done through comparing $\Pr(D|g_i^{(k)})$ and $\Pr(D)$.
- ▶ Also, $\frac{\Pr(D|g_i^{(k)})}{\Pr(D)} = \frac{\Pr(g_i^{(k)}|D)}{\Pr(g_i^{(k)})}$, and

$$\Pr(g_i^{(k)}|D) - \Pr(g_i^{(k)}) \propto \left[\Pr(g_i^{(k)}|D) - \Pr(g_i^{(k)}|U) \right].$$

- ▶ Multi-locus genotype-trait association can be studied through the comparison of the genotype distributions among the cases and the controls.

Information measure: genotype-trait disequilibrium (GTD)

Using

$$\hat{P}(g_i^{(k)}|D) - \hat{P}(g_i^{(k)}|U) = \frac{n_{d,i}^{(k)}}{n_d} - \frac{n_{u,i}^{(k)}}{n_u} = C \cdot \left(\omega \cdot n_{d,i}^{(k)} - (1 - \omega) \cdot n_{u,i}^{(k)} \right),$$

$$\text{GTD}^{(k)} = \sum_{i=1}^{3^k} \left(\omega \cdot n_{d,i}^{(k)} - (1 - \omega) \cdot n_{u,i}^{(k)} \right)^2.$$

where $\omega = \frac{n_u}{n_d + n_u}$ and $1 - \omega = \frac{n_d}{n_d + n_u}$ can be considered as weights that adjust for different sizes of the cases and the controls.

Genotype-trait disequilibrium (GTD)

The expectation of GTD is derived as

$$\begin{aligned} E(\text{GTD}^{(k)}) &= \frac{(n_d n_u)^2}{(n_d + n_u)^2} \sum (p_i^d - p_i^u)^2 \\ &\quad + \frac{n_d n_u}{(n_d + n_u)^2} \left[n_u \sum p_i^d (1 - p_i^d) + n_d \sum p_i^u (1 - p_i^u) \right], \end{aligned}$$

where p_i^d is short for $\Pr(g_i^{(k)}|D)$ and p_i^u for $\Pr(g_i^{(k)}|U)$ to simplify the formulations.

Genotype-trait disequilibrium (GTD)

- ▶ $E(\text{GTD})$ has the smallest value under the null hypothesis.
- ▶ If **some** of the current markers are associated with the trait, then $E(\text{GTD}^{(k)})$ will increase as p_i^d 's and p_i^u 's diverge.
- ▶ The stronger the association signal, the more difference between p_i^d 's and p_i^u , and the larger the value of GTD.
- ▶ However, for a given marker set, a large value of $\text{GTD}^{(k)}$ only indicate that **some** of the markers in this set are *important*, or associated with the trait.
- ▶ If these *unimportant* markers are removed from the marker set, the association signals should become stronger.
- ▶ The importance of any marker M_r can then be evaluated by the difference between the GTDs before and after removing this given marker.

Genotype-trait association (GTA)

- ▶ Consider a current set of k markers, $S_k = \{M_1, M_2, \dots, M_k\}$, and S_{k-1}^r denotes the new set less a certain marker M_r .
- ▶ We define a new statistic Genotype-Trait Association (GTA) is defined as

$$\text{GTA}(r) = \frac{1}{2}\Delta\text{GTD} + \tilde{A} \quad (1)$$

where r indicates the underlying marker evaluated by this GTA score.

- ▶ The adjusting term \tilde{A} is added so that $E(\text{GTA}(r))=0$ under the null hypothesis.
- ▶ if M_r is not associated with the trait, $E(\text{GTA}(r)) \geq 0$.
- ▶ and if M_r is associated with the trait, $E(\text{GTA}(r)) < 0$.
- ▶ The magnitude of the value reflects M_r 's importance.

One BGTA screening

1. Start with all markers in the candidate set.
2. Given the current marker set with k markers, calculate the $GTA(r)$ score for each marker, $r = 1, 2, \dots, k$.
 - ▶ If there are non-negative scores, delete the marker with the maximum $GTA(r)$ score;
 - ▶ otherwise stop and return the remaining markers.
3. If there is no marker remains in the set, stop; otherwise set $k=k-1$ and continue to step 2.

Due to the backward fashion of BGTA, those markers that are of strong interactions tend to return together.

One BGTA screening

1. Start with all markers in the candidate set.
2. Given the current marker set with k markers, calculate the $\text{GTA}(r)$ score for each marker, $r = 1, 2, \dots, k$.
 - ▶ If there are non-negative scores, delete the marker with the maximum $\text{GTA}(r)$ score;
 - ▶ otherwise stop and return the remaining markers.
3. If there is no marker remains in the set, stop; otherwise set $k=k-1$ and continue to step 2.

Due to the backward fashion of BGTA, those markers that are of strong interactions tend to return together.

Limitation of individual BGTA screenings

- ▶ Why: the performance of BGTA algorithm is limited by the dimension of the data.
- ▶ It is difficult for BGTA to screen informatively, in one iteration, all candidate markers.
- ▶ A small number of observations (cases and controls here) only allow one to informatively study the interactions among a small number of dimensions.
- ▶ Such a *scope of inspection* is decided by the size of data, and there is no statistical or mathematical tactic can overcome this limitation.

Random subsets BGTA screenings

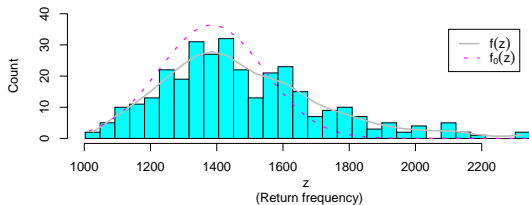
1. **Randomly** select a **subset** of k out of all the markers in a genome scan, where k is relatively small, say 10, so that the screening on these k markers is more informative.
2. Repeat BGTA screening on a large number B , say 10,000, of random subsets and record the screening result of each repetition.
3. Calculate the number of times a marker is returned by BGTA, or so called **return frequency** for each marker.
4. Markers are then ranked by their return frequencies. Important markers (significantly more frequently returned) are selected based on the distribution of the return frequencies.
5. Joint returning patterns observed in such random subset screenings are also to be shown to contain information on inter-locus interactions.

Random subsets BGTA screenings

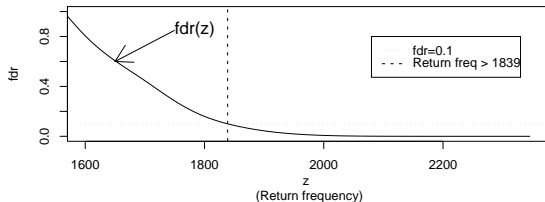
- ▶ Random subsetting the markers and carrying out a detailed screening allows one to explore the large candidate marker set using the largest scope of inspection allowed by the data.
- ▶ Marker selection criterion
 - ▶ An ad hoc way: return frequencies more than 3rd quartile plus 1.8 times IQR (inter-quartile range).
 - ▶ Based on the local FDR rule proposed by Efron (2004).

Marker selection: local FDR

Histogram of return frequencies



False Discovery Rate

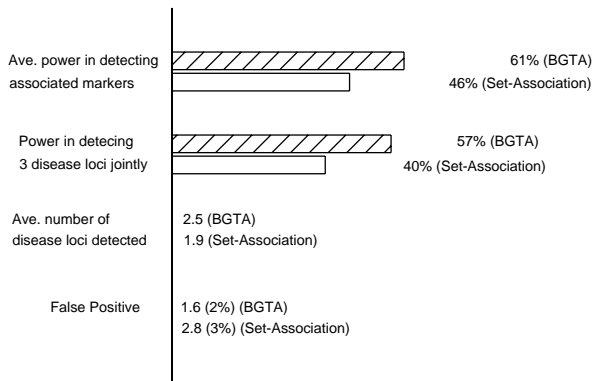


Example 1: genetic heterogeneity model

- ▶ Three susceptibility loci, which are physically independent. Each with a normal allele and a mutation allele.
- ▶ Being homozygous of mutated allele at any of these three disease loci causes an elevated risk of the disease.
- ▶ 80 diallelic markers are simulated.
- ▶ There are 2 markers in linkage/association with each of the disease genes. $\theta = 0.01$ (recombination fraction) and $\Delta = 0.8$ (standard LD). (A total of 6 associated markers.)
- ▶ 150 cases and 150 controls.
- ▶ 300 independent data sets are simulated.
- ▶ For each data set, 5000 random subset screenings are used.
- ▶ The results are compared with the set association method proposed in Hoh et al. 2001.

Example 1

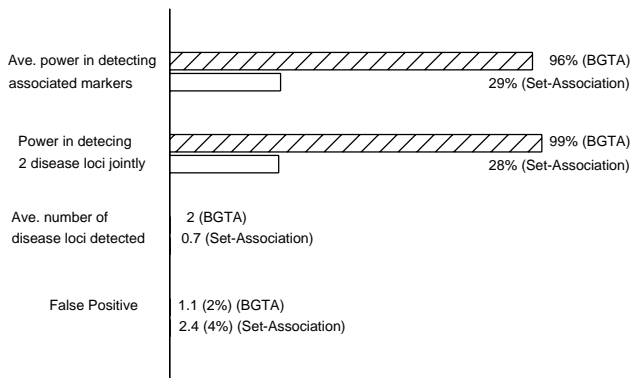
Comparison between Set-Association and BGTA
under genetic heterogeneity model (example 1)



Example 2: Epistasis model I

Genotype at locus A	Genotype at locus B			Marginal effect
	<i>b/b</i> (1/4)	<i>b/B</i> (1/2)	<i>BB</i> (1/4)	
<i>a/a</i> (1/4)	0	0	1	0.25
<i>a/A</i> (1/2)	0	1	0	0.5
<i>AA</i> (1/4)	1	0	0	0.25
Marginals	0.25	0.5	0.25	

Example 2

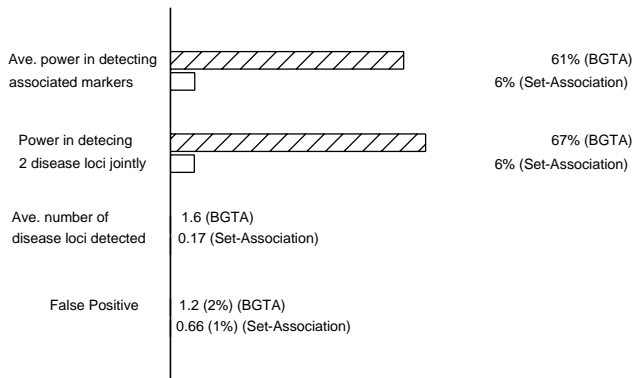
Comparison between Set-Association and BGTA
under epistasis model I (example 2)

Example 3: Epistasis model II

Genotype at locus A	Genotype at locus B			Marginal effect
	<i>b/b</i> (1/4)	<i>b/B</i> (1/2)	<i>BB</i> (1/4)	
<i>a/a</i> (1/4)	0	0	1	0.25
<i>a/A</i> (1/2)	0	0.5	0	0.25
<i>AA</i> (1/4)	1	0	0	0.25
Marginals	0.25	0.25	0.25	

Example 3

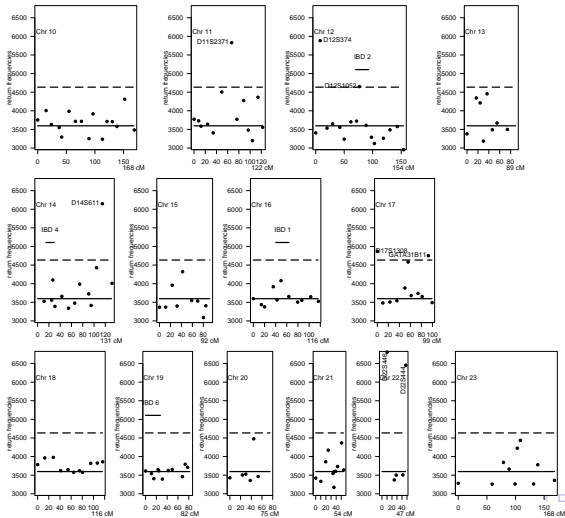
Comparison between Set-Association and BGTA
under epistasis model II (example 3)



Inflammatory bowel disease

- ▶ Received from the Whitehead Institute, MIT.
- ▶ The original data were collected on ASP.
- ▶ For demonstration purpose, 112 pedigrees are divided into two samples.
- ▶ A patient is selected from each pedigrees in sample 1.
- ▶ A control is selected from each pedigrees in sample 2.
- ▶ Not exactly population based but the cases and controls are independent.
- ▶ A total of 402 markers are used in this example.
- ▶ 10 imputations.
- ▶ A total of 500,000 BGTA screenings (40 hours).

IBD example



Current and future efforts

- ▶ Backward Haplotype Transmission Association (BHTA) methods (Lo and Zheng 2002; 2004) for case-parent triads.
- ▶ Multilocus Linkage methods (Ionita and Lo 2005, to appear, Human Heredity)
- ▶ We have also developed similar strategies for QTL mapping.
- ▶ Ideas presented here are being extended to gene expression-based classifications.
- ▶ We are developing software on these methods.
- ▶ More theoretical research is under way on the general idea of these information measures.