

SNP data analysis in genome-wide association studies

Can Yang

Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology

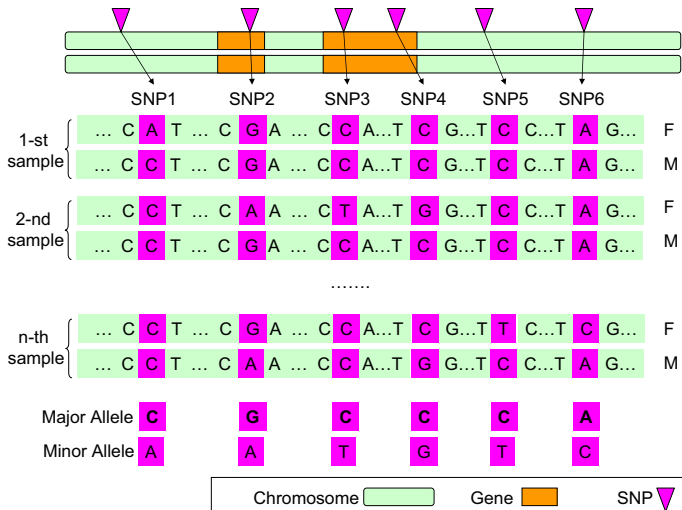
Outline

- 1 Introduction
 - Background
 - Problem description
- 2 Methods
 - BOOST and its extension
 - Identifying disease-associated SNP clusters via CODE
- 3 Conclusion and Acknowledgement
 - Conclusion
 - Acknowledgement

Outline

- 1 Introduction
 - Background
 - Problem description
- 2 Methods
 - BOOST and its extension
 - Identifying disease-associated SNP clusters via CODE
- 3 Conclusion and Acknowledgement
 - Conclusion
 - Acknowledgement

Single Nucleotide Polymorphism (SNP)



SNP Data

- About 10-20 million of SNPs occur in the human genome.
- The Number of genotyped SNPs: $\mathcal{L} \approx 500,000$.
- SNPs are bi-allele markers in the human genome.
- Major alleles: Capital letters (e.g. A, B, ...).
- Minor alleles: Lowercase letter (e.g. a, b, ...).
- Three genotypes for each SNP: AA–0, Aa–1, aa–2.

SNP_1	SNP_2	...	$SNP_{\mathcal{L}}$	disease Status
0	0	...	1	0
1	2	...	0	0
...
2	0	...	1	1

Table: SNP Data: 0–AA; 1–Aa; 2–aa. disease Status: 1–case, 0–control

Single-SNP-based Tests

Hypothesis testing

H_0 : The genotype distribution is the same in cases and controls.

SNP_1	AA	Aa	aa	SNP_2	AA	Aa	aa
Case	480	425	95	Case	600	300	100
Control	490	420	90	Control	640	320	40

Table: SNP_1 : $\chi^2=0.2177$, P -value= 0.8969; SNP_2 : $\chi^2=27.6498$, P -value= 9.907e-07.

- Do it for all SNPs, we get ...

T2D: an example of the GWAS findings

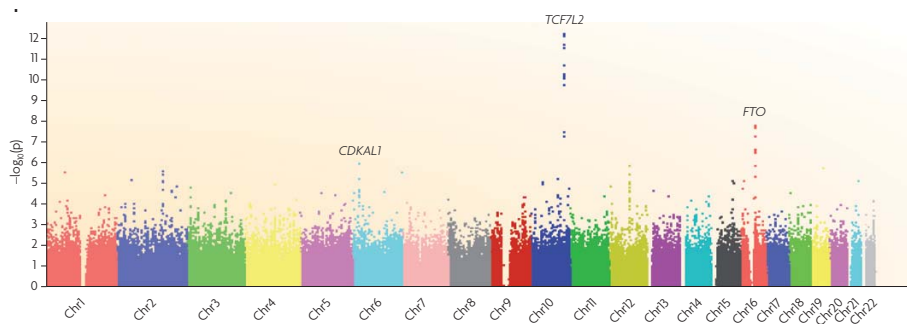


Figure: The GWAS findings of the type 2 diabetes. The $-\log_{10}(P)$ -values of the single-SNP based tests are displayed w.r.t. their genomic positions.



M. I. McCarthy et al. Genome-wide association studies for complex traits. *Nature Review Genetics*, 356-369. (May, 2008)

Missing heritability

$$h^2 = \frac{\sigma_G^2}{\sigma_{\text{Phenotype}}^2}, \text{ where } \sigma_{\text{Phenotype}}^2 = \sigma_G^2 + \sigma_E^2 \quad (1)$$

- The heritability of the human height is estimated to be around 80%. The identified 40 SNPs can only account for 5%.
- The heritability of the T2D is estimated to be around 30%. The identified 18 SNPs can only account for 6%.
-

Interestingly, a similar concept exists in physics, named “dark matter”.



T. A. Manolio et al.

Finding the missing heritability of complex diseases.

NATURE, 747-753. (Oct., 2009)

Possible explanations of the missing heritability

- **Gene-gene interactions are not thoroughly investigated in current GWAS.**
- **Much larger numbers of common variants of smaller effect have not been found.**

Outline

- 1 Introduction
 - Background
 - Problem description
- 2 Methods
 - BOOST and its extension
 - Identifying disease-associated SNP clusters via CODE
- 3 Conclusion and Acknowledgement
 - Conclusion
 - Acknowledgement

Research objective of finding interacting SNP pairs

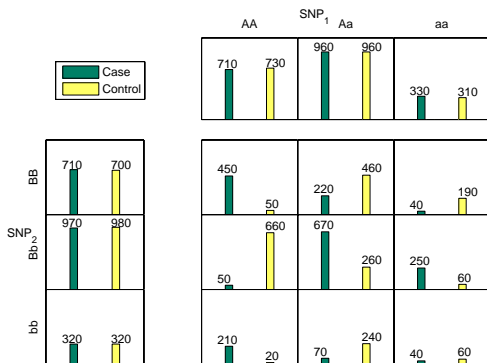
- Input: An $n \times (\mathcal{L} + 1)$ data matrix.

X_1	X_2	...	$X_{\mathcal{L}}$	Y
0	0	...	1	0
1	2	...	0	0
...
2	0	...	1	1

Table: X_{ℓ} : 0-AA; 1-Aa; 2-aa. Disease status Y : 1-case, 0-control

- Output: a set of interacting SNP pairs: $\{(p, q) | X_p \otimes X_q \rightarrow Y\}$.
- The mathematical definition of \otimes will be given later.

Computational challenge



- **Strong interaction but weak marginal effect.** Stepwise strategies fail here.
- Exhaustive search: Evaluate $\mathcal{L}(\mathcal{L} - 1)/2$ SNP pairs for \mathcal{L} SNPs, e.g., $\mathcal{L}(\mathcal{L} - 1)/2 = 1.25 \times 10^{11}$ for $\mathcal{L} = 500,000$.

Outline

- 1 Introduction
 - Background
 - Problem description
- 2 **Methods**
 - **BOOST and its extension**
 - Identifying disease-associated SNP clusters via CODE
- 3 Conclusion and Acknowledgement
 - Conclusion
 - Acknowledgement

Our methods

We have developed some methods to handle genome-wide SNP data:

- MegaSNPHunter, BMC Bioinformatics, 2009.
- SNPHarvester, Bioinformatics, 2009.
- SNPRuler, Bioinformatics, 2010.
- Adaptive GroupLasso, BMC Bioinformatics, 2010.
- **BOOST** (BOolean Operation based Screening and Testing), The American Journal of Human Genetics, Sept., 2010.

Open source software

<http://bioinformatics.ust.hk/BOOST.html>

Computational achievement of BOOST

CPU time efficiency

BOOST can finish the analysis of all pairs of roughly 360,000 SNPs within 60 hours (around 2.5 days) on a standard desktop (3.0 GHz CPU with 4G memory running Windows XP professional x64 Edition system). It is roughly 63 times faster than PLINK.

Data size	BOOST	PLINK
$n = 5000, \mathcal{L} = 1,000$	< 2s	106s
$n = 5000, \mathcal{L} = 5,000$	42s	2,703s
$n = 5000, \mathcal{L} = 10,000$	170s	10,915s

Table: Time comparison of BOOST and PLINK.

The statistical definition of gene-gene interactions (I)

- Define the indicator function as

$$\mathbb{I}(V = v) = \begin{cases} 1 & \text{If } V = v \text{ is true.} \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

- Consider two SNPs X_p and X_q . The main effect model is

$$\mathcal{M}_M : \log \frac{P(Y = 1 | X_p, X_q)}{P(Y = 0 | X_p, X_q)} = \beta_0 + \underbrace{\beta_{p,0}X_{p,0} + \beta_{p,1}X_{p,1}}_{\text{the effect of } X_p} + \underbrace{\beta_{q,0}X_{q,0} + \beta_{q,1}X_{q,1}}_{\text{the effect of } X_q} \quad (3)$$

Here $X_{p,0} = \mathbb{I}(X_p = 0)$. Similarly for $X_{p,1}, X_{q,0}, X_{q,1}$.

- No interaction exists in this model because log-odds-ratio is **a simply summation of the individual effects of X_p and X_q** .

The statistical definition of gene-gene interactions (II)

- Full model

$$\begin{aligned} \mathcal{M}_F : \log \frac{P(Y = 1 | X_p, X_q)}{P(Y = 0 | X_p, X_q)} = & \beta_0 + \beta_{p,0} X_{p,0} + \beta_{p,1} X_{p,1} + \beta_{q,0} X_{q,0} + \beta_{q,1} X_{q,1} \\ & + \beta_{pq,00} X_{p,0} X_{q,0} + \beta_{pq,01} X_{p,0} X_{q,1} \\ & + \beta_{pq,10} X_{p,1} X_{q,0} + \beta_{pq,11} X_{p,1} X_{q,1} \end{aligned} \quad (4)$$

Measure the interaction effects

The difference of the max log-likelihood of these two models implies interaction effects.

$$\mathcal{I} = 2(\hat{L}_F - \hat{L}_M). \quad (5)$$

Key components of BOOST

- BOOST evaluates all SNP pairs: $\mathcal{L}(\mathcal{L} - 1)/2$ in total.
- For each pair, it makes use of the following components:

Three components of BOOST

- Log-linear models & logistic models.
- Kirkwood Superposition Approximation.
- Boolean operations.

Disadvantage of working with logistic regression

- Finding the MLE of a logistic regression model is a convex optimization problem.
- There are about 10 billions of SNP pairs \implies Solving 10 billions of convex optimization problems!
- It takes about 1.2 years (Ma, 2008).

X_1	X_2	...	X_p	...	X_q	...	X_L	Y
0	0	...	1	...	0	...	1	0
1	2	...	2	...	1	...	2	0
...
2	0	...	0	...	0	...	2	1

Table: n rows

Working with contingency tables

- n_{ijk} denotes the observed count in the cell (i, j, k) , i.e., the number of samples with $(X_p = i) \& (X_q = j) \& (Y = k)$.

Table: The contingency table $[n_{ijk}]_{3 \times 3 \times 2}$ for SNP pair (X_p, X_q) and disease status Y : $Y = 1$ for cases and $Y = 0$ for controls. $\{0, 1, 2\}$ is used to represent $\{AA, Aa, aa\}$

$Y = 0$	$X_q = 0$	$X_q = 1$	$X_q = 2$	$Y = 1$	$X_q = 0$	$X_q = 1$	$X_q = 2$
$X_p = 0$	n_{000}	n_{010}	n_{020}	$X_p = 0$	n_{001}	n_{011}	n_{021}
$X_p = 1$	n_{100}	n_{110}	n_{120}	$X_p = 1$	n_{101}	n_{111}	n_{121}
$X_p = 2$	n_{200}	n_{210}	n_{220}	$X_p = 2$	n_{201}	n_{211}	n_{221}

- Use log-linear models to work with these 18 numbers!

Measuring interaction effects via log-linear models

Table: Equivalent loglinear and logistic models (Agresti, 2002).

log-linear model	logistic model
M_H	\mathcal{M}_M
M_S	\mathcal{M}_F

Measure interaction effects using log-linear models

According to Equation (5) and the above equivalence,

$$\mathcal{I} = 2(\hat{L}_F - \hat{L}_M) = 2(\hat{L}_S - \hat{L}_H) \quad (6)$$

where \hat{L}_H and \hat{L}_S is the max log-likelihood of M_H and M_S , respectively.

Measuring interaction effects via log-linear models

After some algebra, we have

$$\begin{aligned} 2(\hat{L}_S - \hat{L}_H) &= 2n \sum_{i,j,k} \left[\hat{\pi}_{ijk} \log \frac{\hat{\pi}_{ijk}}{\hat{p}_{ijk}^H} \right] \\ &= 2n D_{KL}(\hat{\pi}_{ijk} || \hat{p}_{ijk}^H). \end{aligned} \quad (7)$$

where $D_{KL}(\hat{\pi}_{ijk} || \hat{p}_{ijk}^H)$ is known as KL divergence of $\hat{\pi}_{ijk}$ and \hat{p}_{ijk}^H .

- $\hat{\pi}_{ijk}$ in Eq. (7) is given by $\frac{n_{ijk}}{n}$.
- \hat{p}_{ijk}^H in Eq. (7) needs iterative estimation.

Kirkwood Superposition Approximation

To further accelerate computation, we propose a non-iterative approximation of $\hat{\rho}_{ijk}^H$:

Kirkwood Superposition Approximation

$$\hat{\rho}_{ijk}^K = \frac{1}{\eta} \pi_{i|j} \pi_{j|k} \pi_{k|i} = \frac{1}{\eta} \frac{\pi_{ij+} \pi_{i+k} \pi_{+jk}}{\pi_{i++} \pi_{+j+} \pi_{+++k}} \quad (8)$$

where $\eta = \sum_{i,j,k} \frac{\pi_{ij+} \pi_{i+k} \pi_{+jk}}{\pi_{i++} \pi_{+j+} \pi_{+++k}}$ is a normalization term.

KSA bound

$$\hat{L}_S - \hat{L}_H \leq \hat{L}_S - \hat{L}_{KSA}. \quad (9)$$

BOOST

Boolean Operation based Screening and Testing

For each SNP pair,

- Step 1: Collect the contingency table using Boolean operations.
- Step 2 (Screening): Check KSA bound: $2(\hat{L}_S - \hat{L}_{KSA}) \leq \tau$.
- Step 3 (Testing): Only compute $2(\hat{L}_S - \hat{L}_H)$ of SNP pairs whose $2(\hat{L}_S - \hat{L}_{KSA}) > \tau$.

Real data description

WTCCC data

We have applied BOOST to analyze data (14,000 cases in total and 3,000 shared controls) from the Wellcome Trust Case Control Consortium (WTCCC) on seven common human diseases:

- bipolar disorder (BD)
- coronary artery disease (CAD)
- Crohn's disease (CD)
- hypertension (HT)
- rheumatoid arthritis (RA)
- type 1 diabetes (T1D)
- type 2 diabetes (T2D)



WTCCC.

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

Nature, 447:661-678, 2007.

T1D v.s. RA

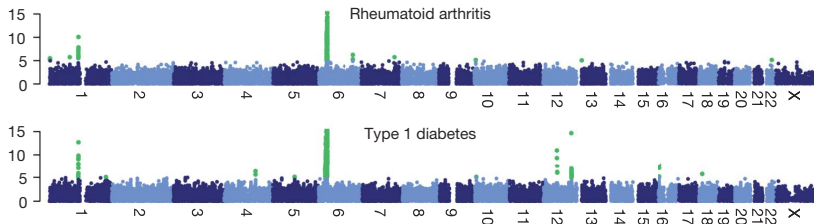


Figure: The single-SNP based analysis results of RA and T1D on the WTCCC data sets. These results do not reveal much difference between RA and T1D. The x-axis is the chromosome position and the y-axis is the $-\log_{10}(P)$.

T1D v.s. RA

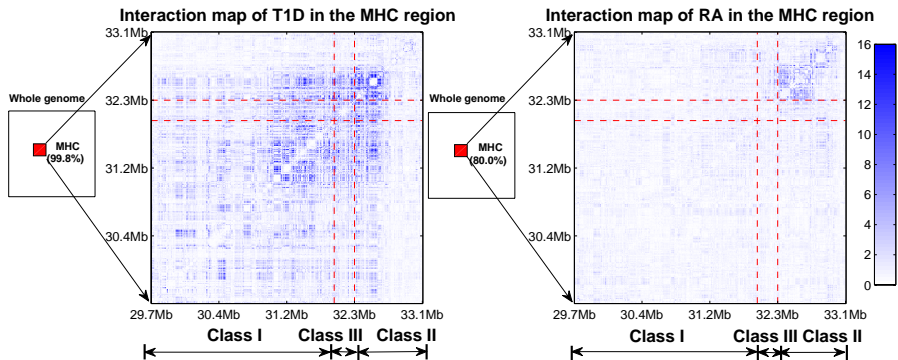


Figure: Interaction pattern of T1D v.s. RA data from WTCCC: MHC region on Chromosome 6.

Interacting SNPs (with weak marginal effects)

SNP 1		SNP 2		Interaction
SNP	Single-locus P -value	SNP	Single-locus P -value	BOOST P -value
rs2524057	4.807×10^{-1}	rs9276448	8.878×10^{-3}	5.362×10^{-14}
rs2524057	4.807×10^{-1}	rs5014418	1.116×10^{-2}	2.738×10^{-13}
rs2853934	8.336×10^{-2}	rs9276448	8.878×10^{-3}	2.507×10^{-13}
rs2524115	1.215×10^{-1}	rs9276448	8.878×10^{-3}	6.456×10^{-13}
rs3873385	3.368×10^{-1}	rs9276448	8.878×10^{-3}	3.186×10^{-14}
rs3873385	3.368×10^{-1}	rs5014418	1.116×10^{-2}	3.841×10^{-14}
rs3873385	3.368×10^{-1}	rs6919798	6.077×10^{-2}	4.257×10^{-13}
rs396038	9.939×10^{-2}	rs9276448	8.878×10^{-3}	5.894×10^{-13}

Table: The interaction SNP pairs in the two regions shown in the previous figure. The SNPs in the column 'SNP 1' reside in the gene HLA-B and The SNPs in the column 'SNP 2' locate at the block across the genes HLA-DQA2 and HLA-DQB2. They show strong interactions without displaying significant main effects.

Extensions of BOOST

- We extend BOOST to handle association allowing for interactions (Chapter 4).
- We find a hidden association pattern in GWAS using log-linear models and Boolean structure (Chapter 5).

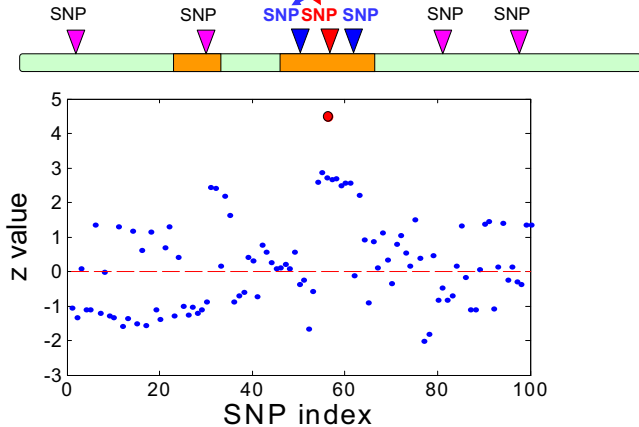
Outline

- 1 Introduction
 - Background
 - Problem description
- 2 **Methods**
 - BOOST and its extension
 - **Identifying disease-associated SNP clusters via CODE**
- 3 Conclusion and Acknowledgement
 - Conclusion
 - Acknowledgement

Motivation

A causal SNP but not directly genotyped

Observed and correlated with the causal SNP



Problem statement

- Input: the z value of each SNP obtained by the trend test and denote them as $\mathbf{z} = \{z_1, z_2, \dots, z_L\}$.
- Output: Partition SNPs into two groups (based on the z values):
 - The null group \mathcal{G}_0 : SNPs are unassociated with the disease.
 - The non-null group \mathcal{G}_1 : SNPs are associated with the disease.

Notation

- $\|\mathbf{x}\|_0$ denotes the ℓ_0 -norm, which counts the number of nonzero entries.
- $\|\mathbf{x}\|_1 = \sum_i |x_i|$ denotes the ℓ_1 -norm.
- $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ and $\|\mathbf{x}\|_2^2 = \sum_i x_i^2$ denote the ℓ_2 -norm and the squared ℓ_2 -norm, respectively.

Formulation (I)

- An outlier detection model

$$\mathbf{z} = \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad (12)$$

where $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \dots, \gamma_L\}$ and $\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, \dots, \epsilon_L\}$.

- $i \in \mathcal{G}_0 : z_i = \epsilon_i;$
- $i \in \mathcal{G}_1 : z_i = \gamma_i.$

$$\min_{\boldsymbol{\gamma}} \frac{1}{2} \|\mathbf{z} - \boldsymbol{\gamma}\|_2^2 + \lambda_1 \|\boldsymbol{\gamma}\|_0. \quad (13)$$

- If $\gamma_i \neq 0$, we have $\gamma_i = z_i$ to minimize (13)

$$\min_{\boldsymbol{\gamma}} \frac{1}{2} \sum_{i:\gamma_i=0} z_i^2 + \lambda_1 \|\boldsymbol{\gamma}\|_0. \quad (14)$$

Formulation (II)

- We introduce $\mathbf{s} = (s_1, \dots, s_{\mathcal{L}})$ as the support of γ :

$$s_i = \begin{cases} 0, & \text{if } \gamma_i = 0 \\ 1, & \text{if } \gamma_i \neq 0. \end{cases} \quad (15)$$

- Equation (14) can be written as

$$\min_{\gamma} \frac{1}{2} \sum_i [z_i^2 (1 - s_i)] + \lambda_1 \|\mathbf{s}\|_0. \quad (16)$$

$$\text{s.t. } s_i \in \{0, 1\}.$$

- Contiguous Outlier DEtection (CODE)

$$\min_{\mathbf{s}} \frac{1}{2} \sum_i [z_i^2 (1 - s_i)] + \lambda_1 \|\mathbf{s}\|_0 + \lambda_2 \sum_{i=1}^{\mathcal{L}-1} w_i |s_i - s_{i+1}|. \quad (17)$$

$$\text{s.t. } s_i \in \{0, 1\}, i = 1, \dots, \mathcal{L}.$$

Stability selection

We use sub-sampling and average the results to produce a stable solution.

- For each subsampling round, we randomly sample half of the cases and half of the controls from the entire data set.
- Let \mathbf{z}_b^* denote the set of z values for the b -th subsampling.
- We run our model on \mathbf{z}_b^* and obtain the support \mathbf{s}_b^* .

The probability of the i -th SNP being detected can be obtained by

$$\pi_i = \frac{\sum_{b=1}^B s_{i,b}^*}{B}, \quad (18)$$

where B is the number of subsampling.

Illustration of sub-sampling and the resulting support

(sub-sampling and cut)

Threshold τ

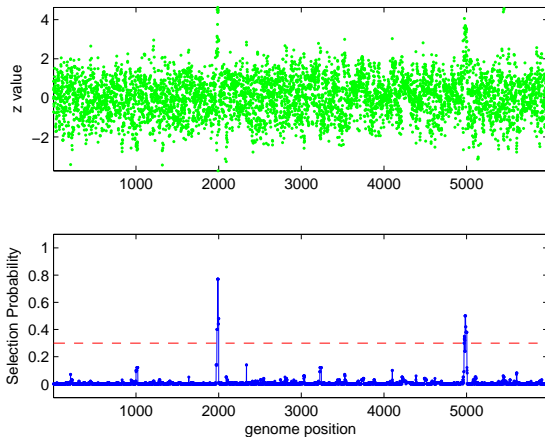


Figure: $\tau = 0.3$. $\mathcal{A}_\tau = \{i : \pi_i \geq \tau\}$ denotes the selected SNPs.

Simulation

- The *locfdr* method (Efron, 2010).
- Fused Lasso (Tibshirani, 2008).

$$\min_{\mu} \sum_{i=1}^{\mathcal{L}} (z_i - \mu_i)^2 + \lambda_1 \sum_{i=1}^{\mathcal{L}} |\mu_i| + \lambda_2 \sum_{i=1}^{\mathcal{L}-1} w_i |\mu_i - \text{sign}(r_{i,i+1}) \mu_{i+1}| \quad (19)$$

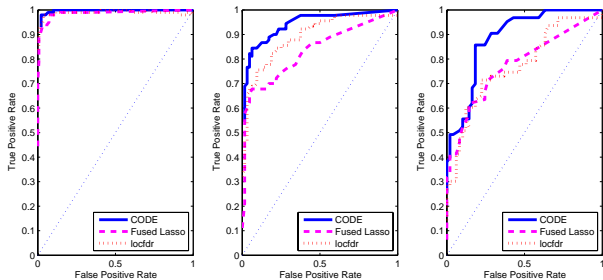


Figure: Performance comparison of CODE, the modified fused Lasso and the *locfdr* method. From left to right, OR are 1.49, 1.35, 1.28, respectively.

Two independent real data sets

SNP name	z	π_j	Location	Gene	Chr
rs2304773*	-0.40343	0.56	234235820	SAG	2q37.1
rs894100	-2.8705	0.56	234237765	SAG	2q37.1
rs3792097	-2.5395	0.56	234238784	SAG	2q37.1
rs3792096	-2.7996	0.56	234238900	SAG	2q37.1
rs2241874	-3.7749	0.56	234247627	SAG	2q37.1
rs2241873	-3.4969	0.56	234247924	SAG	2q37.1
rs4785433	-1.3691	0.44	50586941	NKD1	16q12
rs933566	2.2137	1.00	50642201	NKD1	16q12
rs8047222	4.4660	1.00	50660962	NKD1	16q12

SNP name	z	π_j	Location	Gene	Chr
rs1000141	-4.3470	0.59	234242347	SAG	2q37.1
rs3792091	-2.9304	0.47	234251322	SAG	2q37.1
rs4785220	3.3683	0.34	50627378	NKD1	16q12
rs4785437	-3.8216	0.35	50598798	NKD1	16q12

Table: The WTCCC CD data set and Duerr's CD data set.

Outline

- 1 Introduction
 - Background
 - Problem description
- 2 Methods
 - BOOST and its extension
 - Identifying disease-associated SNP clusters via CODE
- 3 Conclusion and Acknowledgement**
 - Conclusion**
 - Acknowledgement

Conclusion

- We have developed a fast approach, named BOOST, to detecting gene-gene interactions in genome-wide scale.
- We extend BOOST to handle some related topics.
- We also develop a method, named CODE, to identify disease-associated SNP clusters.

Outline

- 1 Introduction
 - Background
 - Problem description
- 2 Methods
 - BOOST and its extension
 - Identifying disease-associated SNP clusters via CODE
- 3 Conclusion and Acknowledgement
 - Conclusion
 - Acknowledgement

Acknowledgement

- I wish to express my great appreciation to my advisors, Prof. Yu and Prof. Yang.
- Special thanks go to my colleague Dr. Xiang Wan. We work together and frequently discuss with each other. The idea of Boolean operations should be attributed to him.
- Thank my parents and my wife Ju Wang for their deepest love.

Publication list (First author or Joint-First author)



Identifying disease-associated SNP clusters via contiguous outlier detection
submitted to Bioinformatics, 2011.



The choice of null distributions for detecting gene-gene interactions in genome-wide association studies.
BMC Bioinformatics, 2011.



A hidden two-locus disease association pattern in genome-wide association studies.
BMC Bioinformatics, 2011.



BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies.
The American Journal of Human Genetics, 2010.



Detecting two-locus associations allowing for interactions in genome-wide association studies.
Bioinformatics, 2010.



Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso.
BMC Bioinformatics, 2010.



SNPHarvester: a filtering based approach for detecting epistatic interactions in genome-wide association studies.
Bioinformatics, 2009.