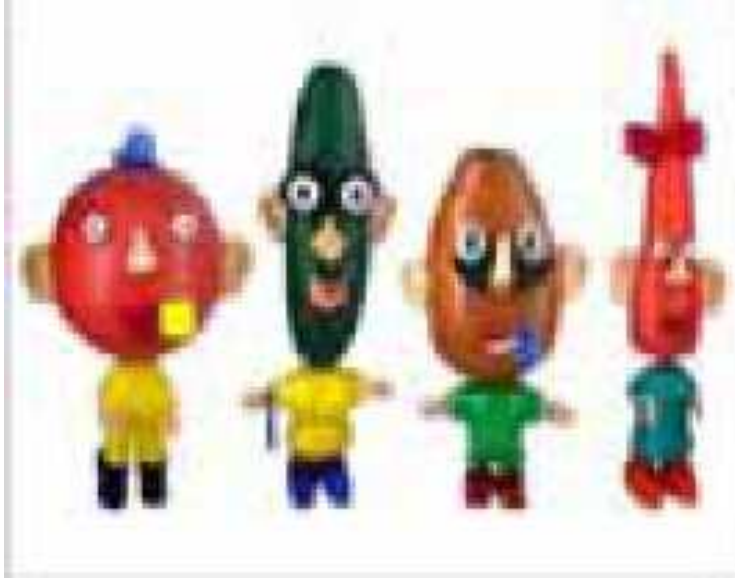
Genome Wide SNP Selection with Entropy Based Methods

Zhenqiu Liu

University of Maryland Greenebaum Cancer Center

The Genetic Diversity in Humane

Any two unrelated people are 99% identical in DNA sequence. The remain 0.1% difference can help explain one person has distinct physical features, is more susceptible to a disease, or responsible differently to a drug or an environmental factor than another person.



Background

- The goal of much genetic research is to find genes that contribute to disease
- Finding these genes should allow an understanding of the disease process, so that methods for preventing and treating the disease can be developed
- For “single-gene disorders”, current methods are usually sufficient

Background

- Most people, however, donŠt have single-gene disorders, but develop common diseases such as heart disease, stroke, diabetes, cancers or psychiatric disorders, which are affected by many genes and environmental factors
- Common-Disease/Common-Variant Theory: The genetic contribution to these diseases is not clear, but many researchers consider common variants to be important

Single Nucleotide Polymorphisms

- A SNP is a single nucleotide site where exactly two (of four) different nucleotides occur in a large percentage of the population
- For example, 30% of the chromosomes may have an A, and 70% may have a G (on a specific site) These two forms, A and G, are called variants or alleles of that SNP
- An individual may have a genotype for that SNP that is AA, AG, or GG.

Genotype and Haplotype

- Diploid populations (e.g., humans) have two copies of each chromosome (one copy inherited from the father, and the other inherited from the mother)
- The collection of SNP variants on a single chromosome copy is a **haplotype**.
- The conflated (mixed) data from the two haplotypes is called a genotype

an example

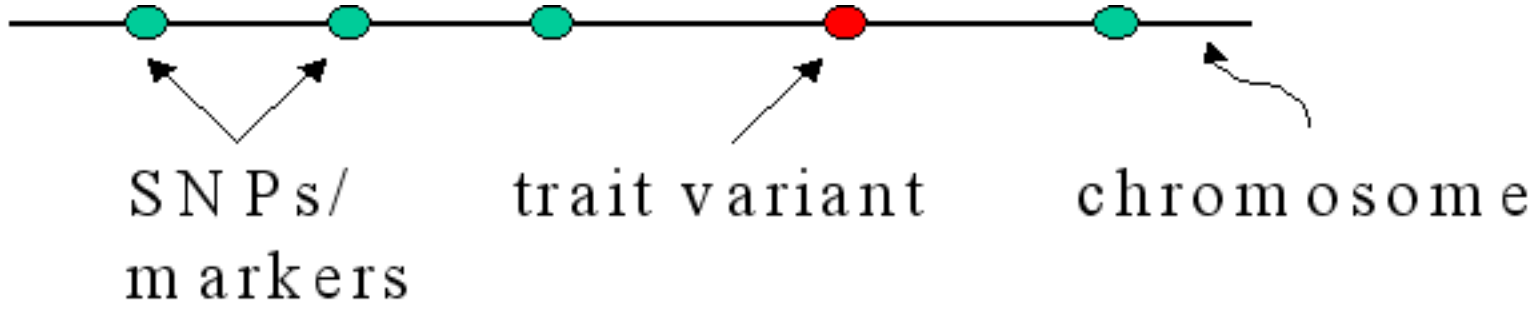
- Each individual has two “copies” of each chromosome.
- At each site, each chromosome has one of two alleles (states) denoted by 0 and 1 (motivated by SNPs)
- Haplotypes for the individual:

$$\begin{array}{ccccccccc} 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \end{array}$$

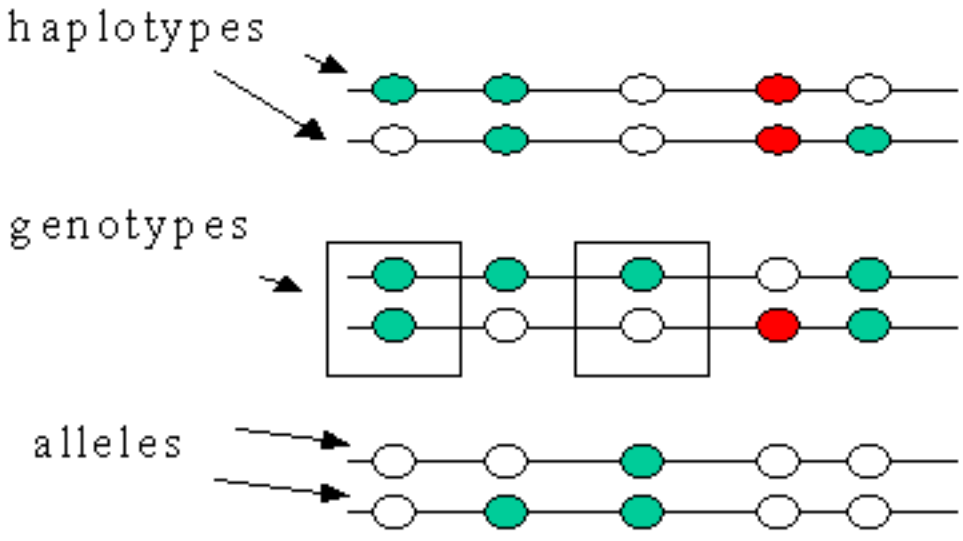
- Genotype for the individual:

$$\begin{array}{ccccccccc} 2 & 1 & 2 & 1 & 0 & 0 & 1 & 2 & 0 \end{array}$$

A Graphic Presentation



Population Data



SNP Association Studies

1. SNP Discovery:
   - Where do I find SNPs to use in my association studies? (e.g. databases, direct resequencing)
2. SNP Selection:
   - How do I choose SNPs that are informative? (i.e. assessing SNP correlation - linkage disequilibrium)
3. SNP Associations:
   - How to find one gene or group of SNP associate with disease?
4. SNP Replication/Function: How is function predicted or assessed

Pairwise LD Measure r^2

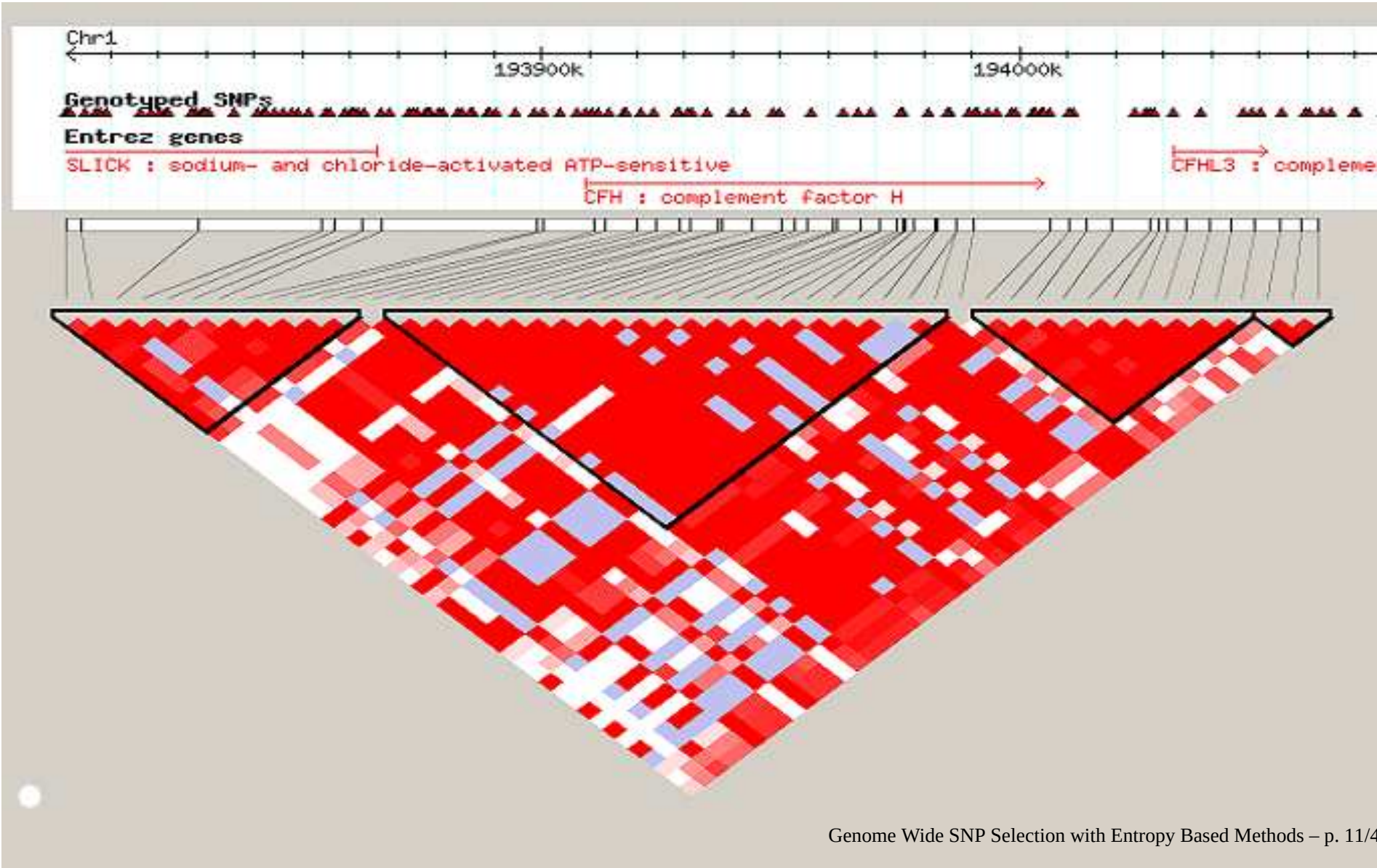
Two bi-allelic markers:

- Locus 1: A, a
- Locus 2: B, b
- Allele frequencies: P_A, P_a, P_B, P_b.
- Haplotype frequencies: P_{AB}, P_{Ab}, P_{aB}, P_{ab},

The r^2 measure is

$$r^2 = \frac{(P_{AB}P_{ab} - P_{aB}P_{Ab})^2}{P_A P_B P_a P_b}$$

Output with Haploview



Objectives

- A multilocus LD measure (ER) with generalized mutual information.

Objectives

- A multilocus LD measure (ER) with generalized mutual information.
- Criteria $\omega(\lambda)$ for tagging SNP selection with joint information and ER.

Objectives

- A multilocus LD measure (ER) with generalized mutual information.
- Criteria $\omega(\lambda)$ for tagging SNP selection with joint information and ER.
- Algorithms for SNP selection (tagging).

Introduction

- Classical LD measures such as $D\prime$ and r^2 are pairwise LD between two loci. They can not provide direct measure of LD for multiple loci.

Introduction

- Classical LD measures such as $D\prime$ and r^2 are pairwise LD between two loci. They can not provide direct measure of LD for multiple loci.
- Multilocus LD measure ε proposed by Nothnagel *et al.* (2002) is useful in many applications. ε is defined as follows:

$$\varepsilon = \frac{H_E - H}{H_E}$$

Definition of ε

Given a chromosomal segment containing n SNPs, let p_j be the frequency of major allele of the jth SNP, $j = 1, \ldots, n$. Suppose there are m observed haplotype with frequency q_i, $i = 1, \ldots, m$, then the entropy of haplotype distribution is defined as

$$H = \sum_{i=1}^{m} q_i \log_2(q_i).$$

Under the assumption of linkage equilibrium, we have

$$q_k^E = \prod_{j=1}^{n} p_j^{I_k^j} (1 - p_j)^{1 - I_k^j},$$

ε Continued

where I_k^j is a index function with value 0 and 1. Then

$$H_E = \sum_{i=1}^{2^n} q_k^E \log_2(q_k^E)$$

and

$$\varepsilon = \frac{H_E - H}{H_E}$$

- $0 \leq \varepsilon < 1$, but can never reach 1.

ε Continued

where I_k^j is a index function with value 0 and 1. Then

$$H_E = \sum_{i=1}^{2^n} q_k^E \log_2(q_k^E)$$

and

$$\varepsilon = \frac{H_E - H}{H_E}$$

- $0 \leq \varepsilon < 1$, but can never reach 1.
- The larger the ε, the greater the LD.

Drawbacks of ε

- The upper bound of ε can never reach 1.

Drawbacks of ε

- The upper bound of ε can never reach 1.
- For a block in which all SNPs are in complete LD, ε's outcome is dependent on the number of SNPs considered.

Drawbacks of ε

- The upper bound of ε can never reach 1.
- For a block in which all SNPs are in complete LD, ε's outcome is dependent on the number of SNPs considered.
- It is computational inefficient.

Our Work

To overcome of the above drawbacks:

- We proposed an ER measure.

Our Work

To overcome of the above drawbacks:

- We proposed an ER measure.
- also proposed a criteria and algorithms for SNP selection using ER measure.

Mulitlocus LD Measure ER

Assume that each haplotype has n marks and there are m haplotype overall and $\mathbf{x}_i$ be the ith haplotype. x_{ij} be the allele at locus j and haplotype i, our LD measure is

$$E = \sum_{i=1}^{m} p(\mathbf{x}_i) \log_2 \frac{p(\mathbf{x}_i)}{\prod_{j=1}^{n} p_j(x_{ij})}. \tag{1}$$

Because of the properties of K-L distance, this LD measure is nonnegative and is zero if and only if the variables are independent. This measure is bounded.

ER Continued

The bound can be found in terms of entropies of component variables.

$$E \leq \sum_{j=1}^{n} H(\mathbf{x}_j) - \max_j H(\mathbf{x}_j) = E_{\max}.$$

Consequently, we can use the normalized LD measure with

$$(2) \qquad ER = \frac{E}{E_{\max}} = \frac{\sum_{i=1}^{m} p(\mathbf{x}_i) \log_2 \frac{p(\mathbf{x}_i)}{\prod_{j=1}^{n} p_j(x_{ij})}}{\sum_{j=1}^{n} H(\mathbf{x}_j) - \max_j H(\mathbf{x}_j)}$$

Properties of ER

1. $0 \leq ER \leq 1$, ER is 0 and 1 when the SNPs are in complete LE and LD respectively.
2. For two loci, $ER \approx r^2$ under certain condition.

Criteria for SNP Selection

The criteria for selecting tagging SNPs is defined as follows:

$$\omega(S, \lambda) = (1 - \lambda)HD(S) + \lambda(1 - ER(S)),$$

where

$$HD(S) = \frac{H(S)}{H(X)}$$

represents the normalized joint information of selected SNPs. $0 \leq \lambda \leq 1$, ER(S) is the multilocus LD measure for selected SNPs.
Obviously with the proposed criteria ω, we can either do the exhaustive search or forward (backward and stepwise) selection for selecting SNPs.

FSA(λ)

1. Set predetermined constants δ_1, δ_2, and λ, and the maximum number of selected SNPs.
2. Choose the first SNPs X_j that maximizes the entropy $H(X_j)$. Then set $t = 1$ and $X_s^t = \{X_j\}$.
3. let $j = \textit{argmax}_k\{\omega_k, k \in X_{-s}^t\}$, where X_{-s}^t contains the remaining SNPs not in X_s^t. If $\frac{H(s)}{H(X)} > \delta_1$ or $ER(S) > \delta_2$ (or $t > N$, an additional criteria if one desires), then the algorithm is terminated and X_s^t is the set of selected SNPs; otherwise, set $X_s^{t+1} = \{X_s^t, X_j\}$ and go back to 3.

Assessment of ER

Example 1: There are only two haplotypes of 1111111111 and 2222222222 with frequency 0.9 and 0.1 respectively. The values of ER , ε and r^2 are given in the following Table.

Table 1: **LD Outputs with various window size**

No. Loci	ER	ε	r^2
2	1.0	0.50	1.0
3	1.0	0.67	-
4	1.0	0.75	-
5	1.0	0.80	-
10	1.0	0.9	-

Example 2

Table 2: **Input Data 2 For LD Measure**

Haplotype	Count (freq.)
2211112211	0.34
1212112112	0.28
1211121221	0.26
1122211112	0.07
2121112111	0.05

Comparison Pairwise LD Measure

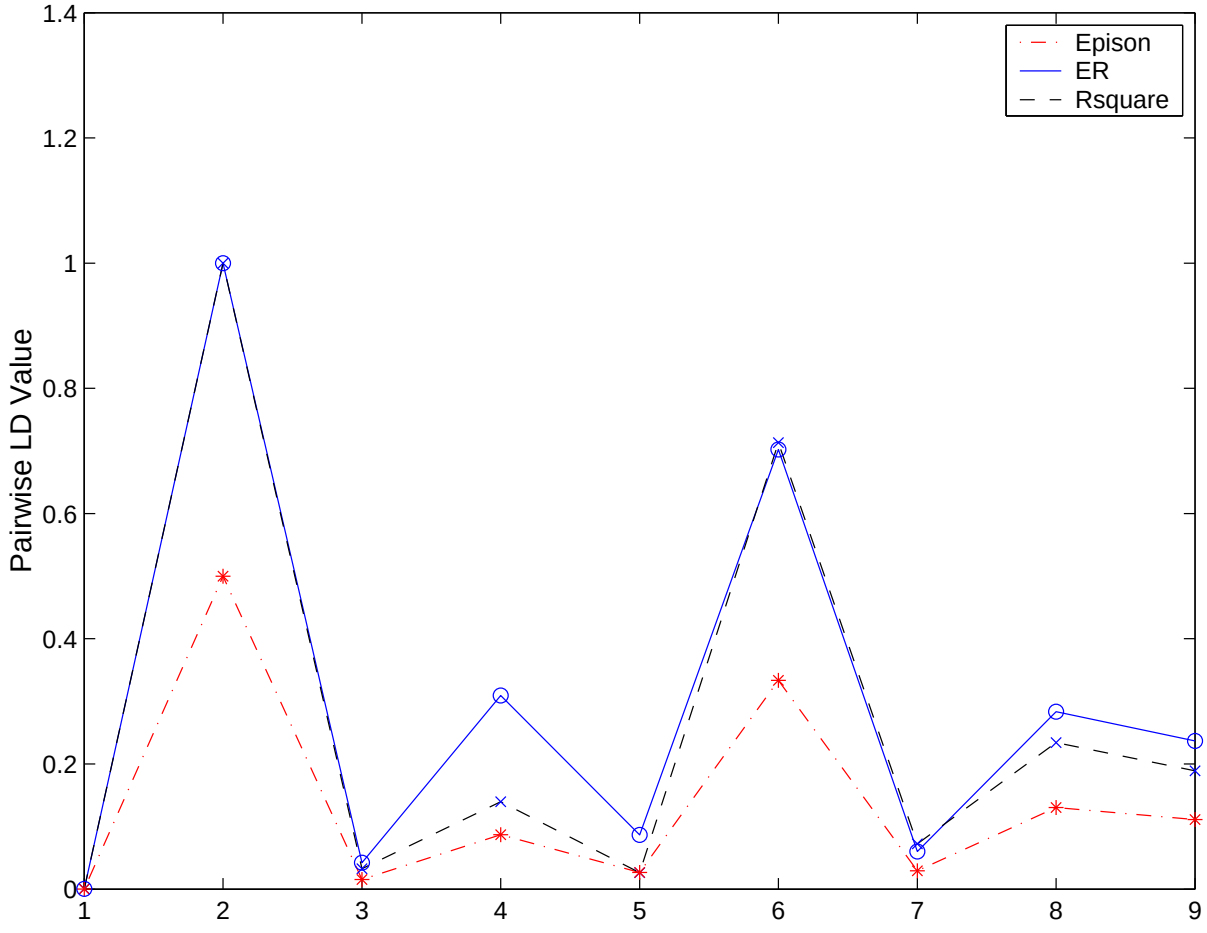


Figure 1: Pairwise LD Comparison: ER , ε, and r^2

Selecting Tagging SNPs

The results of tagging SNPs selection are evaluated with popular criteria: haplotype r^2 (RSQ) and Proportion of Diversity Explained (PDE). The results of RSQ and PDE are based on exhaustive search and our results are based on forward selection. Two haplotype data were used. The first haplotype data is with 20 loci and the second haplotype data is with 51 loci. both data are estimated from Clayton's genotype data.

20 loci evaluated with RSQ

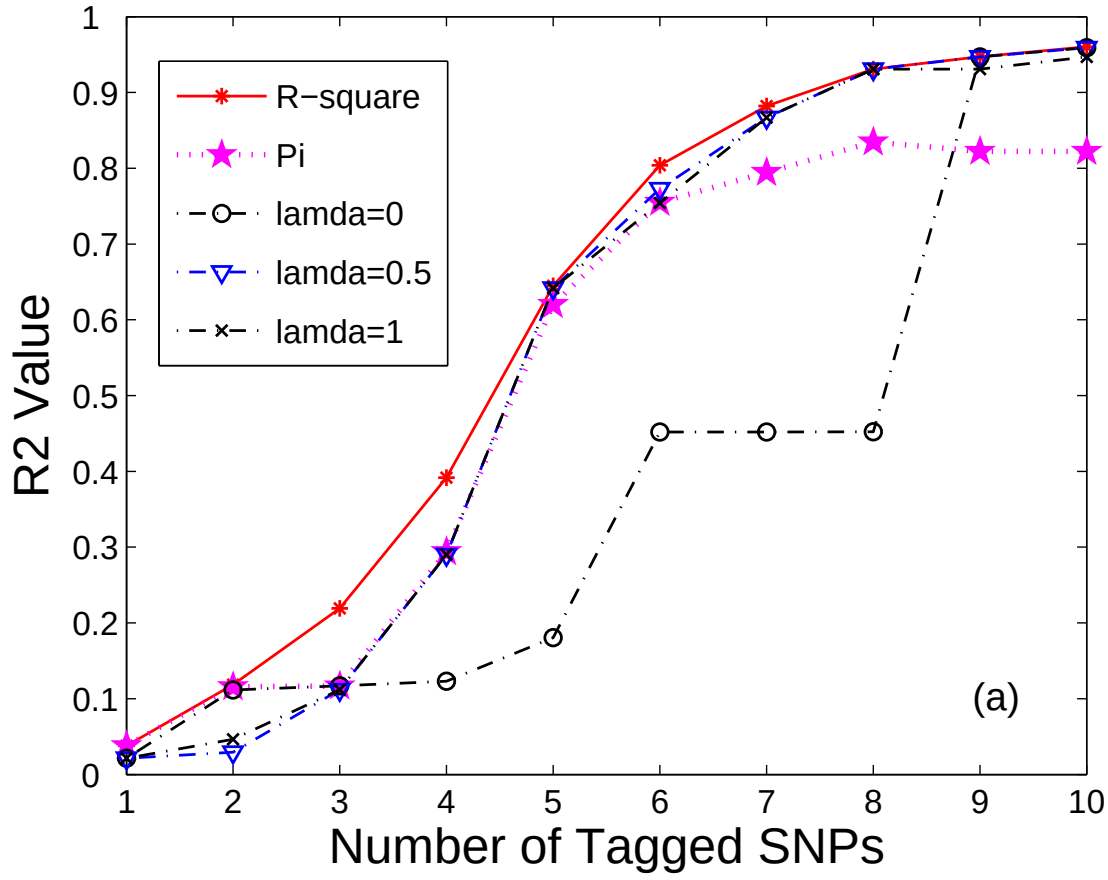


Figure 2: Performance Evaluation with RSQ

20 loci evaluated with PDE

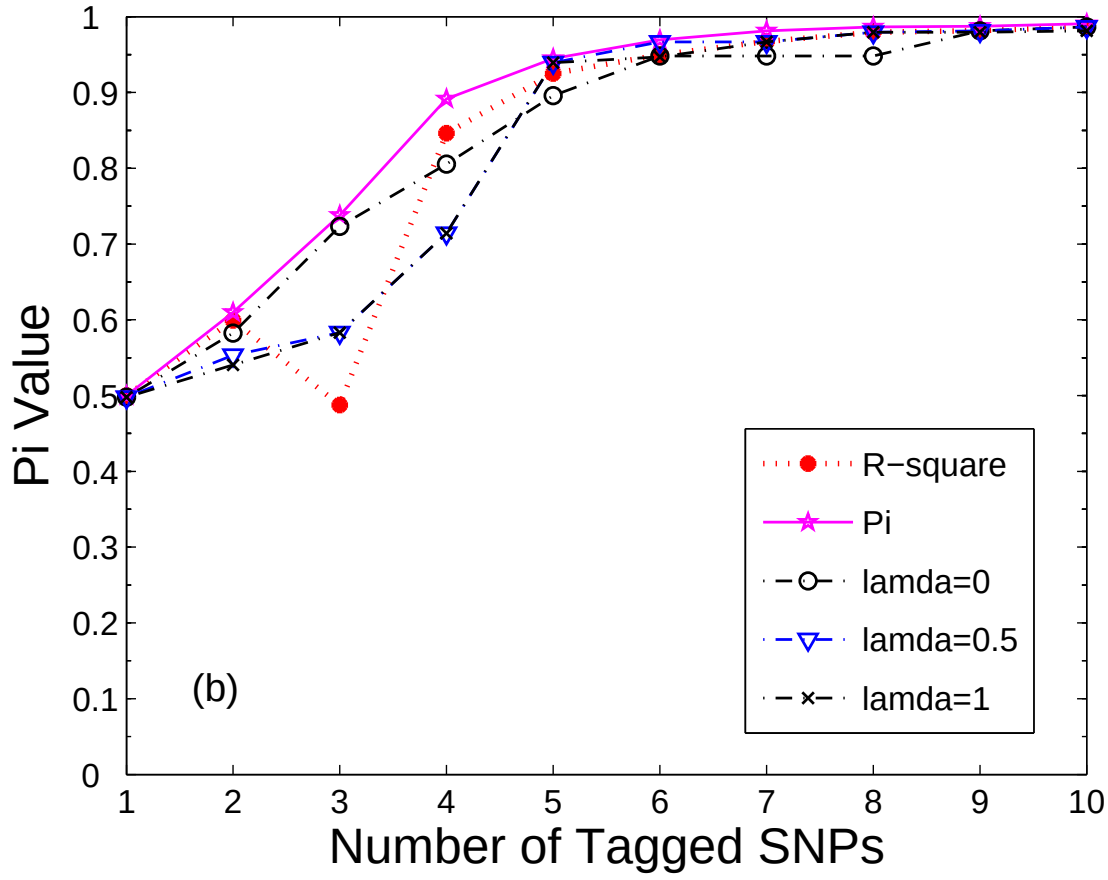


Figure 3: Performance Evaluation with PDE

Haplotype Block Assumption

- Haplotype block is defined as discrete regions of low diversity whose boundaries are conserved across distinct haplotypes. Most algorithms in the literature are haplotype block tagging, that is, grouping SNPs into segments of low haplotype diversity and then tagging a subset of the SNPs within that block.

Haplotype Block Assumption

- Haplotype block is defined as discrete regions of low diversity whose boundaries are conserved across distinct haplotypes. Most algorithms in the literature are haplotype block tagging, that is, grouping SNPs into segments of low haplotype diversity and then tagging a subset of the SNPs within that block.
- However significant theoretical and empirical evidence which shows that conserved substructure may be lost when the data is being fitted to a block structure.

Genome-wide SNP Selection

- Not many block free methods available in the literature.

Genome-wide SNP Selection

- Not many block free methods available in the literature.
- entropy based multilocus LD measure ER was used as the criterion to be optimized.

Genome-wide SNP Selection

- Not many block free methods available in the literature.
- entropy based multilocus LD measure ER was used as the criterion to be optimized.
- Cross entropy Monte Carlo (CEMC), and searches the tagging SNPs from the full set that optimizes a criterion.

The Problem

The objective of SNP tagging is to choose a smallest subset of SNPs that maximizes the ω criterion. Mathematically the problem can be defined as finding S^* such that

$$S^* = \arg\max_S \{\omega(S), \quad S \subset \{1, \ldots, n\}\}. \tag{3}$$

This problem is combinatorial in nature and an exhaustive search requires searching through all subsets of indexes of the SNPs. This is not tractable even for a moderate number of SNPs.

The Algorithm

- Set $\mathbf{p}^0$ with each $p_j^0 \in (0,1)$. For instance, $p_j^0 = 0.5$ indicates that each SNP can be selected with 50% chances. Set t =0.
- Draw a sample $\mathbf{z}_i = (z_{i1}, \cdots, z_{in}),\ i = 1, \cdots, N$, of Bernoulli vectors with success probability $\mathbf{p}^t$. Find the tagging index set $S_i = \{j | z_{ij} = 1, j = 1, \cdots, n\}$ and calculate $\Phi(\mathbf{z}_i) = \omega(S_i)$ for all $i's$ and sort them in ascending order: $\Phi_{(1)} \leq \ldots \leq \Phi_{(N)}$. Let $[(1-\rho)N]$ be the integer part of $(1-\rho)N$, then we have the sample $(1-\rho)$-quantile of the performances: $y^t = \Phi_{([(1-\rho)N])}$, where $\rho < 1$ is a free parameter needed to be specified.

Algorithm Continued . . .

- Use the same sample $\mathbf{z}_i,\ i = 1, \cdots, N$ to update the parameter vector $\mathbf{p}^{t+1} = (p_1^{t+1}, \ldots, p_n^{t+1})$ via

$$p_j^{t+1} = \frac{\sum_{i=1}^{N} I(\Phi(\mathbf{z}_i) \geq y^t) z_{ij}}{\sum_{i=1}^{N} I(\Phi(\mathbf{z}_i) \geq y^t)}, \quad j = 1, \cdots, n.$$

- if $||\mathbf{p}^{t+1} - \mathbf{p}^t|| < \varepsilon_1$ and $|y^{t+1} - y^t| < \varepsilon_2$, then go to Step 5; otherwise set t= t+1 and go back to step 2. Note that $||.||$ denotes a norm such as the sum of squared component distances.
- Output $y = \Psi_{(N)}$ and the corresponding selected SNPs set S, which will be taken as the estimate of our tagging SNP set S^*.

Results: 51 Genotype Data

For CEMC and FSA(ω), we used $\lambda = 0.4$. Furthermore, for CEMC, we set $p_j^0 = 0.5, j = 1, \cdots, n, N = 1000$, $\rho = 0.1$, and $\varepsilon_1 = \varepsilon_2 = 10^{-6}$.

Comparisons of tagging SNP sets and their performances evaluated using the RSQ criterion for three methods on a small dataset.

Algorithm	SNP Index														RSQ
CEMC	1	2	5	7	13	16	17	19	22	25	30	32	41	46	0.953
FSA(ω)	1	5	7	10	13	16	17	22	25	32	39	41	43	46	0.940
FSA(RSQ)	1	5	7	10	13	16	17	22	25	30	31	41	43	46	0.954

Results: 51 Genotype data

We see that CEMC is very close to the gold standard set by FSA(RSQ), while FSA(ω) is slightly lagging behind, according to the RSQ criterion. These results demonstrate that CEMC is indeed a viable alternative for SNP selection, as it is capable of selecting a tagging set that is composed of only 27% of the full set but has retained 95% of the haplotype diversity.

Results: two simulated dataset

Table 2: Sizes of tagging SNP sets and their RSQ values with various λ settings.

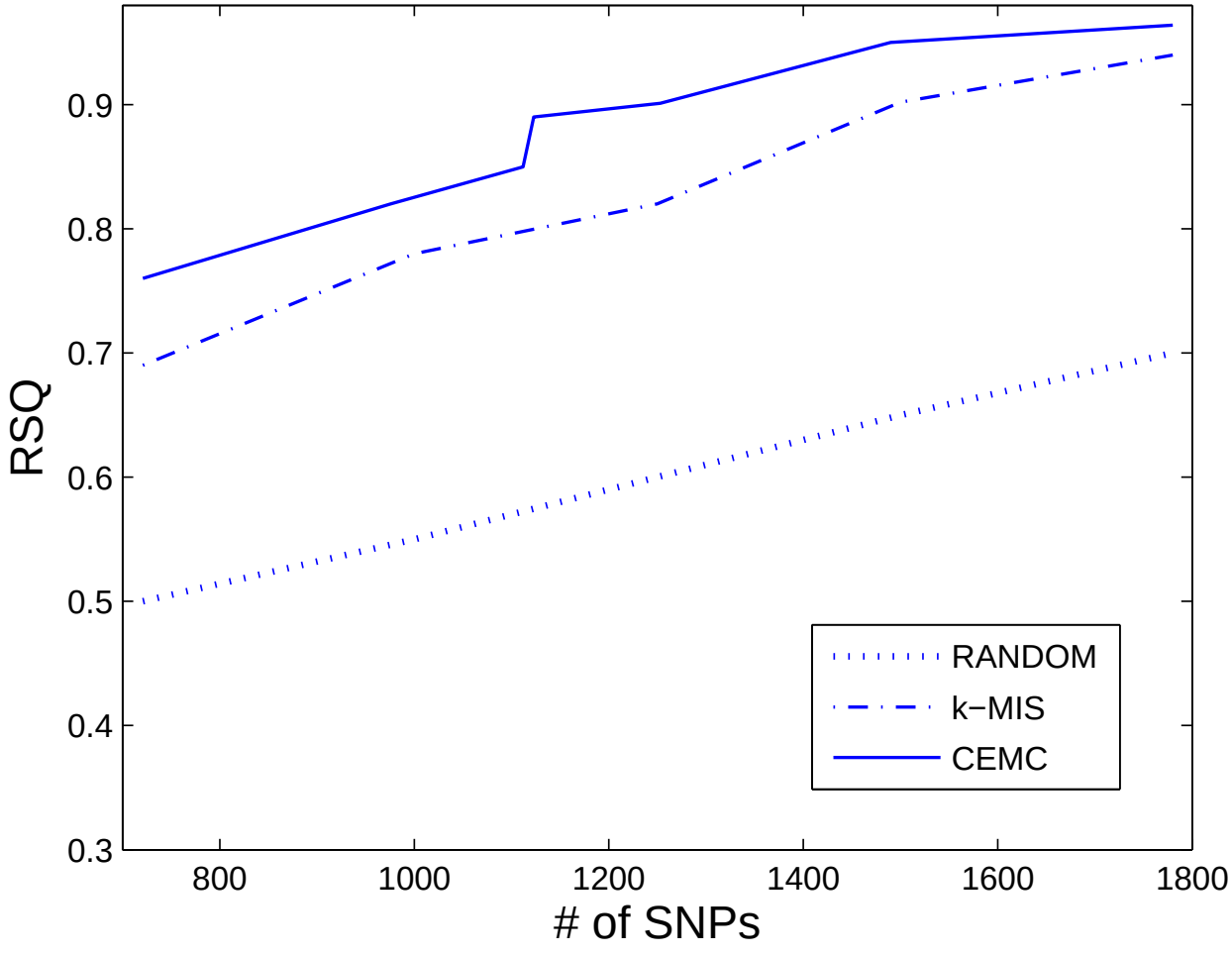
			λ				
Data	Algorithm		0.3	0.4	0.5	0.6	0.7
781-SNP Set	CEMC/RAND	No. SNPs	47	48	47	42	36
	CEMC	RSQ	0.921	0.907	0.903	0.872	0.809
	RAND[a]	RSQ	0.265	0.271	0.265	0.212	0.164
		(SD)	0.081	0.097	0.102	0.087	0.092
1390-SNP set	CEMC/RAND	No. SNPs	69	65	64	60	63
	CEMC	RSQ	0.909	0.854	0.819	0.772	0.749
	RAND[a]	RSQ	0.496	0.463	0.457	0.401	0.445
		(SD)	0.032	0.043	0.048	0.039	0.026

[a]For random selection (RAND), the process was repeated 100 times. The RSQ and SD are the average and standard deviation over the 100 selected SNP sets.

Results: one real dataset

This data set consists of 4120 SNPs distributed along chromosome 22 with a median spacing of 4kb, genotyped by the 5' nuclease assay (de la Vega *et al.* 2002) on 45 DNA samples of Caucasian individuals obtained from the NIGMS Human Variation Panel (Coriell Institute of Medical Research, Camden, NJ). It is particularly interesting to analyze this dataset since its density and sample size are similar to those in the International HapMap Project.

Results: one real dataset



Conclusions

Results with these large scale datasets demonstrate that CEMC is computationally feasible for whole genome SNP selection. Furthermore, the results show that CEMC is significantly better than random selection, and it also outperformed another block-free selection algorithm for the dataset considered.

Collaborators

- Dr. Shili Lin from the Ohio State University.
- Dr. Ming Tan from University of Maryland