Stat 8625

A Comparative Study of The Performance of Multi-Locus LD Measures on Rare and Common Varaiants

Hengrui Luo and Shuyuan Lou

The Ohio State University

December 4, 2017

Introduction

- Classic LD measures are focused on two SNPs, one being disease locus and the other being marker locus.
- Pairwise version will not be able to account for accumulative effect of multiple locus and also not capable of capturing the association for complex disease which is affected by multiple loci.
- Haplotypes,i.e., multi-locus blocks, are getting more and more concern, due to the observation that nearby SNPs tend to affect the trait jointly.
- Some multi-locus LD measures have been published, but which one can capture the underlying LD structure better remain unknown.
- We introduced a few current pairwise and multi-locus measures that are of our interest and compared their performance.

Pairwise LD measure revisit

Normalized LD coefficient (Weir,et al. 1996):

$$D' = \begin{cases} \frac{P(M_1 M_2) - P(M_1)P(M_2)}{max\left\{-P(M_1)P(M_2), -P(\bar{M_1})P(\bar{M_2})\right\}}, & \text{if } P(M_1 M_2) - P(M_1)P(M_2) < 0 \\ \frac{P(M_1 M_2) - P(M_1)P(M_2)}{max\left\{P(M_1)P(\bar{M_2}), P(\bar{M_1})P(M_2)\right\}}, & \text{if } P(M_1 M_2) - P(M_1)P(M_2) > 0 \end{cases} \tag{1}$$

R-squared coefficient (Lewontin, 1964):

$$R^2 = \frac{\left[P(M_1 M_2) - P(M_1)P(M_2)\right]^2}{\left[P(M_1)P(M_2)P(\bar{M_1})P(\bar{M_2})\right]} \tag{2}$$

M_1, M_2 are SNPs at two given loci, we estimate $\hat{P}\left(M_i\right) = \frac{\#minor\ \ allele\ \ at\ \ i}{\#total\ \ alleles\ \ at\ \ i}, \forall i = 1, 2$ and $\hat{P}\left(M_1 M_2\right)$ being the haplotype frequency.

Multi-locus LD measures: Index of association

For a sample of multi-locus measure, the index of association (Agapow,et al. 2001,Brown, et al 1980) is defined as $I_A = \frac{\sum_j Var_j + 2\sum_{j,k} Cov_{j,k}}{\sum_j Var_j} - 1 = \frac{2\sum_{j,k} Cov_{j,k}}{\sum_j Var_j}$ where Var_j is the variance of samples at locus $j \in A4$ and $Cov_{j,k}$ is the covariance of samples between locus j, k. We use its normalized version for d-loci:

$$\bar{r_d} = \frac{\sum_{j,k}^{d} Cov_{j,k}}{\sum_{j,k}^{d} \sqrt{Var_j Var_k}} \tag{3}$$

Multi-locus LD measures: homozygosity of haplotypes

Generalizing the pairwise LD coefficient D to multi-locus (Mueller, et al. 2004, Sabatti, et al 2006):

$$H_d = H\left(M_1 ... M_d\right) - \prod_{i=1}^{d} H\left(M_i\right) \tag{4}$$

Where $M_i, i = 1, 2, ..., d$ are loci and $H\left(M_1 ... M_d\right), H\left(M_i\right)$ are the haplotype homozygosity and minor allele frequency at loci i respectively. In practice, $H\left(M_1 ... M_d\right)$ can be replaced with $\hat{P}_{1...d} = \hat{P}\left(M_1 ... M_d\right)$; $H\left(M_i\right)$ can be replaced with $\hat{P}_i = \hat{P}\left(M_i\right)$ as we defined these estimates above.

Multi-locus LD measures: Relative Entropy LD measure

Based on the idea of quantifying the information-theoretic Kullback-Leibler divergence between observed haplotype distribution and haplotype distribution when there is no LD. And with the empirical allele frequency we can compute haplotype frequency $p(X_i)$ and yield following relative entropy LD measure and its normalized version (Liu, et al 2005):

$$E_d = \sum_x p\left(x\right) \log_2 \frac{p(x)}{\prod_{i=1}^{d} p_i\left(x\right)} \tag{5}$$

$$RE_d = \frac{E_d}{max_x E_d\left(x\right)} \tag{6}$$

Where x sums over all observed haplotypes among all these d-loci. In practice, $H\left(M_1...M_d\right)$ can be replaced with $\hat{P}_{1...d} = \hat{P}\left(M_1...M_d\right)$; $H\left(M_i\right)$ can be replaced with $\hat{P}_i = \hat{P}\left(M_i\right)$ as we defined these estimates above.

Other multi-locus LD measures

Besides these three multi-locus LD measures we are going to implement and compare, there are other LD measures available. One gaining rising popularity is the chromosome segment homozygosity LD measure (CSH) based on a similar idea of IBD applied onto a segment of chromosome. In this measure, the algorithm for solving CSH is based on the iterative equation(Hayes, et al 2003):

$$CSH(M_i, M_j) = argmin_{csh} H(M_i M_j) - \left[csh + \sum_{l} P(recombination\ probability\ for\ binary\ configuration\ l) \right] \tag{7}$$

Other multi-locus LD measures

The other class consist of entropy-based measures, like entropy-type generalization of the r^2 into higher dimension, modified according to the criticism put forward for simple generalization by (Hill,et al., 1981, Nothnagel,et al., 2002):

$$\epsilon(M_i, i = 1, 2, ..., d) = \frac{\sum_l q_l^E \log_2 q_l^E - \sum_{i=1}^m q_i log_2 q_i}{\sum_l q_l^E log_2 q_l^E} \tag{8}$$

where $q_i, i = 1, 2, ..., m$ are frequencies for observed m haplotypes at all these d loci. And then $q_l^E = \prod_{i=1}^d p_j^{I_l^i}(1 - p_j)^{1 - I_l^i}$ with p_j being the observed frequencies of major allele at loci j and I_l^i being the indicator of the event that the major allele occur at i on haplotype l.

Simulation Study: Generating Simulation Data

Following Turkmen and Lin (2016), we also make use of the simulated program provided by Basu and Pan (2011) which allows us to obtain a set of simulated SNPs discretized from prescribed latent multivariate normal variables with correlation structure of first order autoregressive, that is $Corr(X_i, X_j) = \rho^{|i-j|}$ for some constant $\rho \in (0, 1]$. Each latent variable is mapped to 0 or 1, 1 indicating a minor allele, according to the MAF at that loci. Two sets of such dichotomous value are generated at each loci for all samples. At each loci, 697 discrete variables of values $\{0, 1, 2\}$ are generated representing the minor allele counts.

Simulation Study: Generating Simulation Data

- Two genes, CDC27 and FLT1, from GAW17 (Almasy et al., 2011) are selected to generate simulation data. CDC227 is a gene with 33 variants, 3 rare variants and 30 common variants. FLT1 is a gene with 35 variants, 32 rare variants and 3 common variants.
- When there is all zeros for a rare variant locus we randomly assign an 1 in order to allow the LD measures to be calculated correctly. Since $\frac{1}{2\times 697} \approx 0.000717$, such a random assignment will not affect the result too much. Plus, all MAFs in real data set is greater than 0.
- If two SNPs i and j are associated (or in the same haplotype block), then they shall be in linkage disequilibrium.
- Apply multi-locus LD measure on 3 variants, i.e., $d = 3$.

Simulation Study: Pairwise Measures

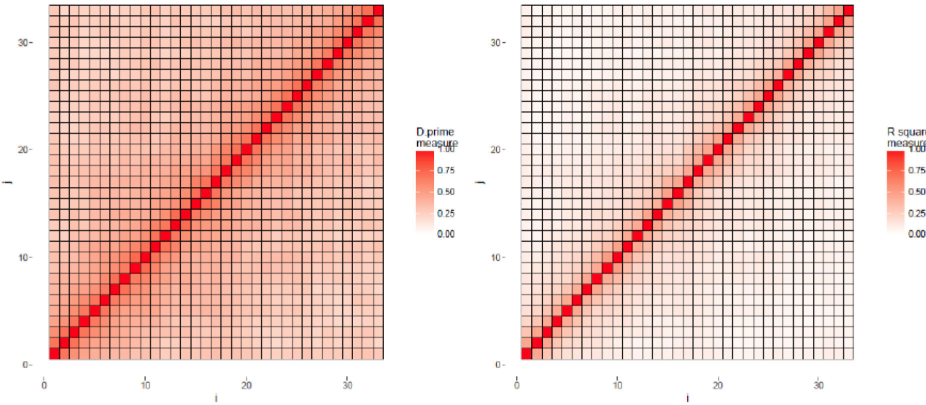


Figure: The heat map of D' and r^2 measures calculated pair-wisely for CDC27 gene in GAW 17 data, which contains 33 variants; 3 rare variants(3,5,19) and 30 common variants

Simulation Study: Pairwise Measures

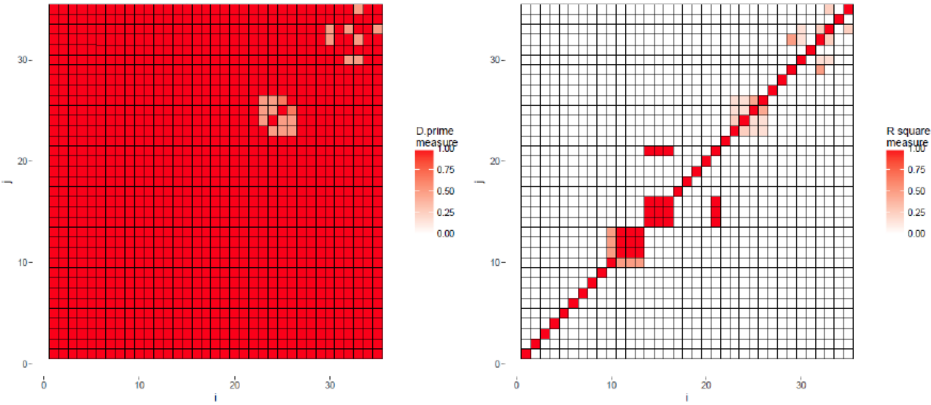


Figure: The heat map of D' and r^2 measures calculated pair-wisely for FLT1 gene in GAW 17 data, which contains 35 variants; 32 rare variants and 3 common variants (3,6,25).

Simulation Study: Multi-locus Measures

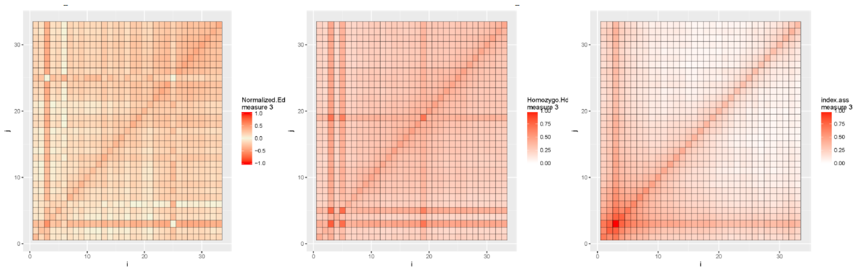


Figure: The heat map of multi-locus measures for M(Xi,Xj , 3) with locus 3, a rare variant on CDC27 fixed. On CDC27 gene with mostly common variants.

Simulation Study: Multi-locus Measures

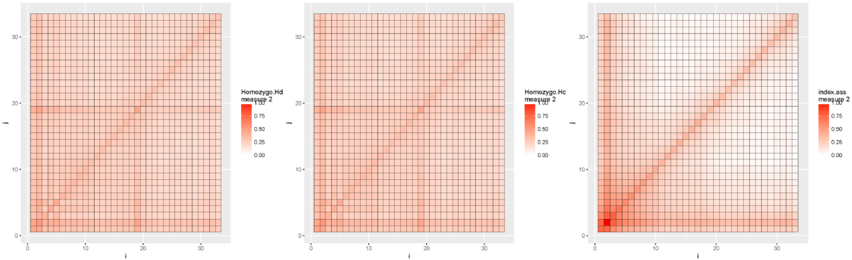
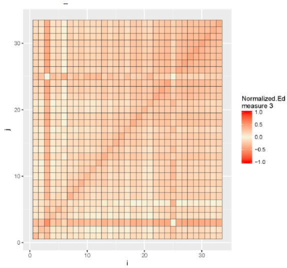
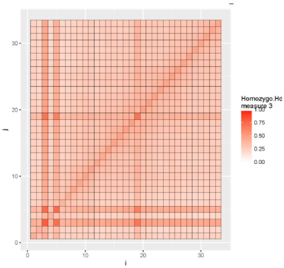
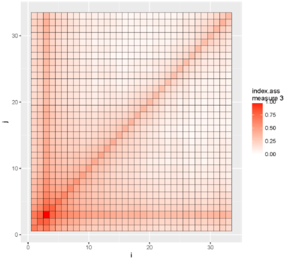


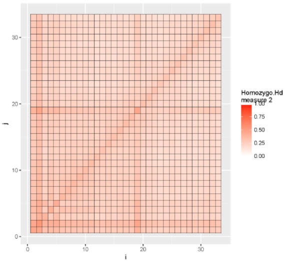
Figure: The heat map of multi-locus measures for M(Xi,Xj , 2) with locus 2, a common variant on CDC27 fixed.

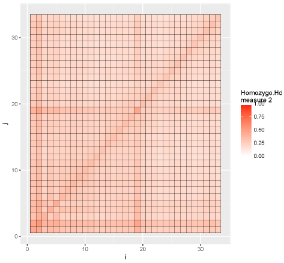
Simulation Study: Multi-locus Measures

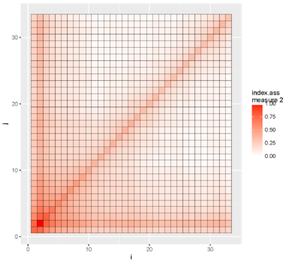












Simulation Study: Multi-locus Measures

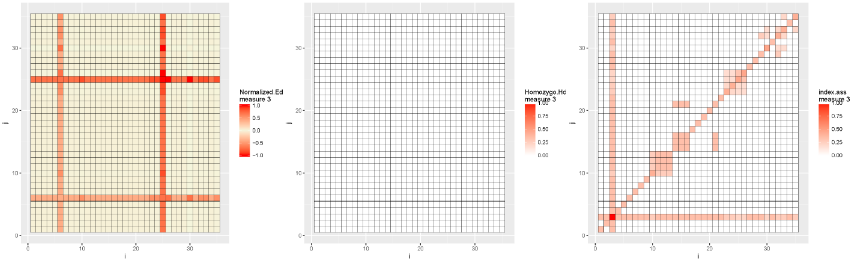


Figure: The heat map of multi-locus measures for M(Xi,Xj , 3) with locus 3, a rare variant on FLT1 fixed. On FLT1 gene with mostly common variants.

Simulation Study: Multi-locus Measures

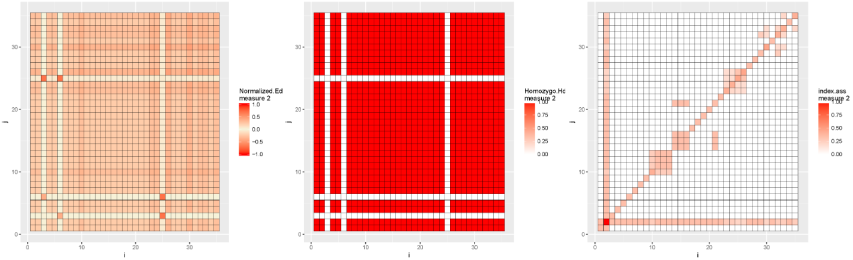
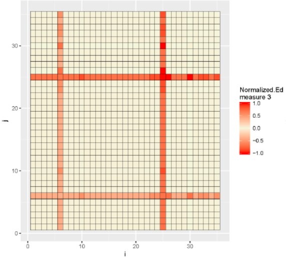
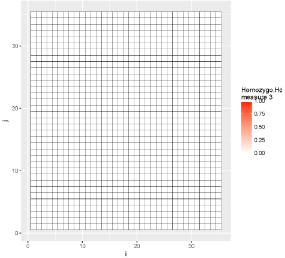
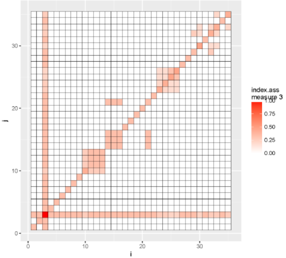


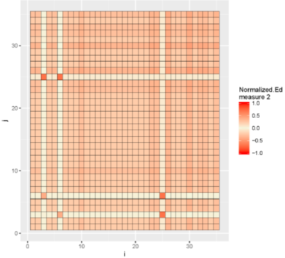
Figure: The heat map of multi-locus measures for M(Xi,Xj , 2) with locus 2, a common variant on FLT1 fixed.

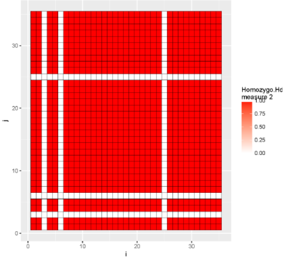
Simulation Study: Multi-locus Measures

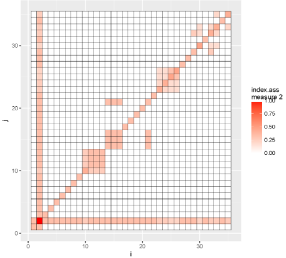












Conclusion and Discussions

- Classical two-locus measures will have relatively reasonable performance when the dominant SNPs on the same gene are common variants, but there are not obvious difference when they are calculated for common-common, common-rare, rare-rare variant pairs.
- Multi-locus measures seem to be quite sensitive to whether a variant is actually rare or not. When the variant is rare, the multi-locus measures are performing much better than two-locus pair-wise LD measures.
- When there are both common and rare variants existing among a collection of SNPs a more careful examination should be done before LD measures could be interpreted directly as an indicator of association between variants.

Conclusion and Discussions

- Within a domain consisting of mostly rare variants, the index association seem to be the optimal choice of quantifying the linkage disequilibrium;
- Within a domain consisting of mostly common variants, the entropy base multi-locus measures seem to be the optimal choice because it reveals more structures on the heat map
- When the haplotype frequencies do not have an appropriate estimator to estimate, H_d behaves not quite good (which is often the case for rare variants).

On-going and Future Work

- With more information, it is expected that H_d can our performs E_d because it incorporate more information in its calculation.
- Working on a software named Haploview (Barrett,et al., 2004) that provide an estimation of Haplotype Frequency.
- Exploring other methods of estimating Haplotype Frequency
- Looking for a better way to generate simulation data.

Reference

- Weir, B. S., and C. Cockerham. "Genetic data analysis II: Methods for discrete population genetic data. Sinauer Assoc." Inc., Sunderland, MA, USA (1996).
- Lewontin, R. C. "The interaction of selection and linkage. I. General considerations; heterotic models." Genetics 49.1 (1964):49.
- Agapow, Paul-Michael, and Austin Burt. "Indices of multilocus linkage disequilibrium." Molecular Ecology Resources 1.1-2(2001): 101-102.
- Brown, A. H. D., M. W. Feldman, and E. Nevo. "Multilocus structure of natural populations of Hordeum spontaneum." Genetics 96.2 (1980): 523-536.
- Mueller, Jakob C. "Linkage disequilibrium for different scales and applications." Briefings in bioinformatics 5.4 (2004):355-364.
- Sabatti, Chiara, and Neil Risch. "Homozygosity and linkage disequilibrium." Genetics 160.4 (2002): 1707-1719
- Liu, Zhenqiu, and Shili Lin. "Multilocus LD measure and tagging SNP selection with generalized mutual information." Genetic epidemiology 29.4 (2005): 353-364.

Reference

- Hayes, Ben J., et al. "Novel multilocus measure of linkage disequilibrium to estimate past effective population size." Genome research 13.4 (2003): 635-643.
- Hill, William G. "Estimation of effective population size from data on linkage disequilibrium." Genetics Research 38.3(1981): 209-216.
- Nothnagel, M., R. Frst, and K. Rohde. "Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks." Human heredity 54.4 (2002): 186-198.
- Turkmen, Asuman, and Shili Lin. "Are rare variants really independent?." Genetic Epidemiology 41.4 (2017): 363-371.
- Almasy, Laura et al. Genetic Analysis Workshop 17 Mini-Exome Simulation. BMC Proceedings 5.Suppl 9 (2011): S2. PMC. Web. 4 Dec. 2017.
- Basu, Saonli, and Wei Pan. "Comparison of statistical tests for disease association with rare variants." Genetic epidemiology 35.7 (2011): 606-619.

Reference

- Barrett, Jeffrey C., et al. "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics 21.2 (2004): 263-265.