From Multilocus Measures to Tag SNP Selection

STAT8750.03 Bio-informatics

Hengrui Luo

Department of Statistics
The Ohio State University

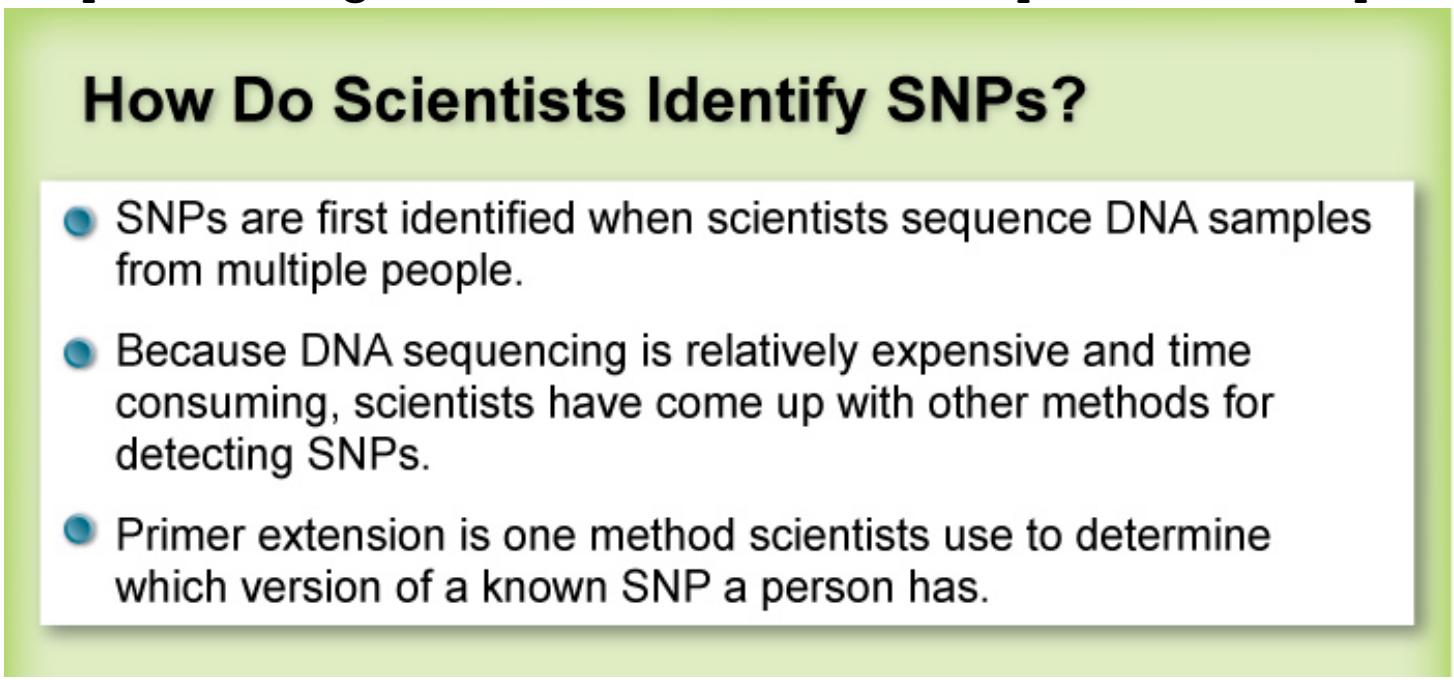
January 9, 2018

Linkage Disequilibrium and SNPs I

- **Linkage Disequilibrium**
  - In Hardy-Weinberg equilibrium, a pair of alleles at different loci are independently inherited in every generation.
    If we use one random variable to represent the status of alleles at this loci then these random variables are **pairwisely independent**.
  - However, genetics study showed that Linkage Equilibrium (LE) is not usually the case, when the frequency of association of their different alleles at these two loci is higher or lower than what would be expected if the loci were independent then this pair of loci are said to be in linkage disequilibrium (LD).
    - ⋆ E.g. AB,Ab,aB,ab should all have $\frac{1}{4}$ chance of being inherited if LE is true; otherwise LD should be considered.
  - LD may be due to
    - ⋆ Actual genetic linkage. The genes in these two loci are closely located on the same chromosome.
    - ⋆ Functional interaction. Some combinations of the pair of alleles at the two loci affect the viability of potential offspring.

Linkage Disequilibrium and SNPs II

- **Single-nucleotide Polymorphism (SNP)**
  - A single-nucleotide polymorphism (SNP) is one kind of variations in a single nucleotide which creates different alleles at a specific locus. SNPs have been used in genome-wide association studies (GWAS) as high-resolution markers allowing scientists to identify the location of a specific gene which is associated with disease of interest. Other useful biological markers includes microsatellite DNA markers.
    `http://learn.genetics.utah.edu/content/precision/snips/`



Statistical modeling of genetic mapping I

- **Statistical modeling**
  - The trait or diseases are measured quantitatively or qualitatively, we are to model it as a response Y_i, it could be discrete or continuous.
  - But if we consider from the biological view, then there are too many SNPs, or covariates X_i, that could be used to explain the response.
  - Statistically we must consider a model with a lot of covariates, especially in complex diseases.
    - ⋆ Genotyping all related SNP markers is expensive.
    - ⋆ Fitting a model of a lot of covariates has many problems.
- **Dimension Reduction**

  Can we use some sort of dimension reduction techniques to alleviate the problem caused by the large number of covariates?

  Select a small number of characteristic SNPs to represent the remaining SNPs, called **tag SNPs**.
  - The existence of LD/correlation among SNPs make a small fraction of the tag SNPs to be useful.

Statistical modeling of genetic mapping II

- **Choice of tag SNPs**
  How to choose tag SNPs from a lot of SNPs that have been sequenced?
  The idea of choosing tag SNPs is to choose all those SNPs that can represent the most LD information, so first of all we must find a way of measuring the extent of LD between two loci. There are basically two categories of LD measures: [1]
  - Haplotype-block based Methods.
    Such a kind of method focus on the haplotype patterns in a specific population, requires division of chromosomes into blocks.
  - Genome-wide Methods.
    Such a kind of method consider the correlation among all SNP markers across the entire genome to represent genome-wise associations measured by **pairwise LD measures** at two given loci M_1, M_2.

$$D' = \begin{cases} \dfrac{P(M_1M_2)-P(M_1)P(M_2)}{max\left\{-P(M_1)P(M_2), -P(\overline{M_1})P(\overline{M_2})\right\}} & \text{if } P(M_1M_2)-P(M_1)P(M_2) < 0 \\ \dfrac{P(M_1M_2)-P(M_1)P(M_2)}{max\left\{P(M_1)P(\overline{M_2}), P(\overline{M_1})P(M_2)\right\}} & \text{if } P(M_1M_2)-P(M_1)P(M_2) > 0 \end{cases}$$

$$r^2 = \frac{[P(M_1M_2)-P(M_1)P(M_2)]^2}{\left[P(M_1)P(M_2)P(\overline{M_1})P(\overline{M_2})\right]}$$

Statistical modeling of genetic mapping III

The idea of choosing tag SNPs for genetic mapping modeling is very similar to select a representative variable among each of those clusters of similar/correlated covariates, and then evaluate the reduced model by how well it predicts in terms of accuracy. In fact, as we can see from the description of the problem, the problem of choosing tag SNPs is essentially the same as variable selection problem, but here the biological consideration dominates and the measure of similarity becomes the LD measures. (Not exactly PCA)

- *Proposing a LD measure could give us a way of choosing tag SNPs, and conversely an algorithm/method of choosing a tag SNP model could induce a LD measure.*

Clustering components $\Longleftarrow$	clustering criterion
$\downarrow$	*LD measure*
Component selection $\Longleftarrow$	selection criterion
$\downarrow$	*one tag for each cluster*
Reduced model evaluation $\Longleftarrow$	evaluation criterion

$$\begin{cases} \textit{number of tag SNPs} \\ \textit{quality of tag SNPs} \\ (\textit{predictive accuracy}) \end{cases}$$

- **Problem of measuring LD among SNP markers**
  1. One-to-one: the LD between two SNP markers A, B.
  2. Many-to-one: the LD between three SNP markers A, B, C. If A, B are already known to be in *strong* LD, then they are considered as a whole M. After this combination we consider the LD between M, C instead.
  3. Many-to-many: the LD between two SNP sets $A_1, A_2, \cdots A_n$ and $B_1, B_2, \cdots, B_m$ where $m, n \geq 2$.
- **Existing multilocus LD measures**
  - Index of association I_A [2].
    Based on generalization of correlation coefficient, no haplotype information is needed.

$$I_A = \frac{\left(\sum_j Var_j + 2\sum_{j,k} Cov_{j,k}\right)}{\sum_j Var_j} - 1 = \frac{2\sum_{j,k} Cov_{j,k}}{\sum_j Var_j}$$

    where Var_j is the variance of samples at locus $j \in A$ and $Cov_{j,k}$ is the covariance of samples between locus j, k .
  - Homozygosity of haplotypes [3, 4].
    Based on generalization of LD measure D, requires haplotype information.

$$H_d = H(M_1 \cdots M_d) - \prod_{i=1}^{d} H(M_i).$$

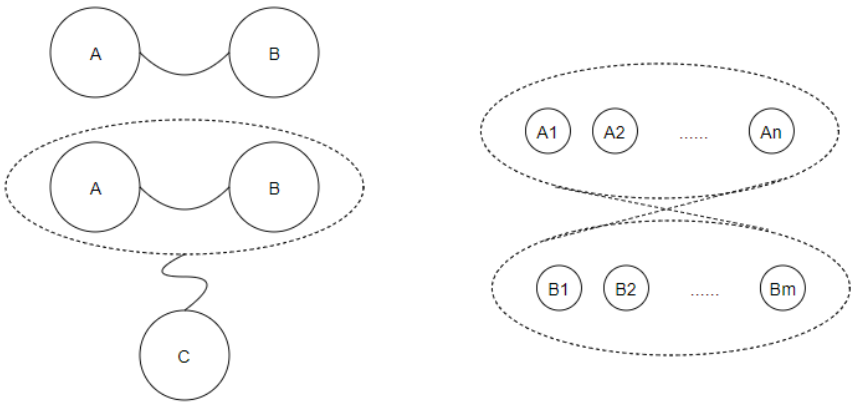
where $M_i, i = 1, 2, \cdots d$ are loci and $H(M_1 \cdots M_d), H(M_i)$ are the haplotype homozygosity and minor allele probability at loci i respectively [4].

- Relative entropy-based LD measure proposed by [5].
  Based on the idea of quantifying the information-theoretic Kullback-Leibler divergence between observed haplotype distribution and haplotype distribution when there is no LD.
  Assume that there are m observed haplotypes at d-loci. With the empirical allele frequency we can compute haplotype frequency $p(X_i)$ and yield following relative entropy LD measure and its normalized version

$$E_d = \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log_2 \frac{p(\boldsymbol{x})}{\prod_{i=1}^{d} p_i(\boldsymbol{x})}$$

$$RE_d = \frac{E_d}{max_{\boldsymbol{x}} E_d(\boldsymbol{x})}$$

where $\boldsymbol{x}$ sums over all observed haplotypes among all these d-loci.



Information Theoretic Motivation I

The multilocus LD measures we introduced above solve those three problems described above, but as pointed out by [1, 5], the problem shall be regarded in the flow chart above. For a locus X with k alleles of frequencies $p(x_i)$, then the uncertainty at the locus can be measured by **entropy** $H(X) = -\sum_{i=1}^{k} p(x_i)\log_2 p(x_i)$, the **joint entropy** for two loci X, Y can also be defined similarly as $H(X,Y) = -\sum_{i=1}^{k}\sum_{j=1}^{l} p(x_i, y_j)\log_2 p(x_i, y_j)$ with $H(X,Y) \leq H(X) + H(Y)$. The **conditional entropy** $H(Y \mid X) := H(X,Y) - H(X)$ measures the randomness from X given Y and **mutual information** that the randomness commonly shared by both $I(X;Y) := H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y)$ with $I(X;Y) = I(Y;X) \in [0, \infty]$.A high value of mutual information means there is a high correlation between these two loci X, Y.

Information Theoretic Motivation II

Can we compare $I(S_1, X), I(S_2, X)$ directly?
No, because if S_1 contains more SNPs than S_2 then $I(S_1, X)$ will usually be larger. Intuitively, when there are more explanatory variables in the model, the model can have more predictive power. Including all the SNP markers will attain the highest mutual information but that goes against our purpose of choosing tag SNPs.
See following figure from [1] for illustration.

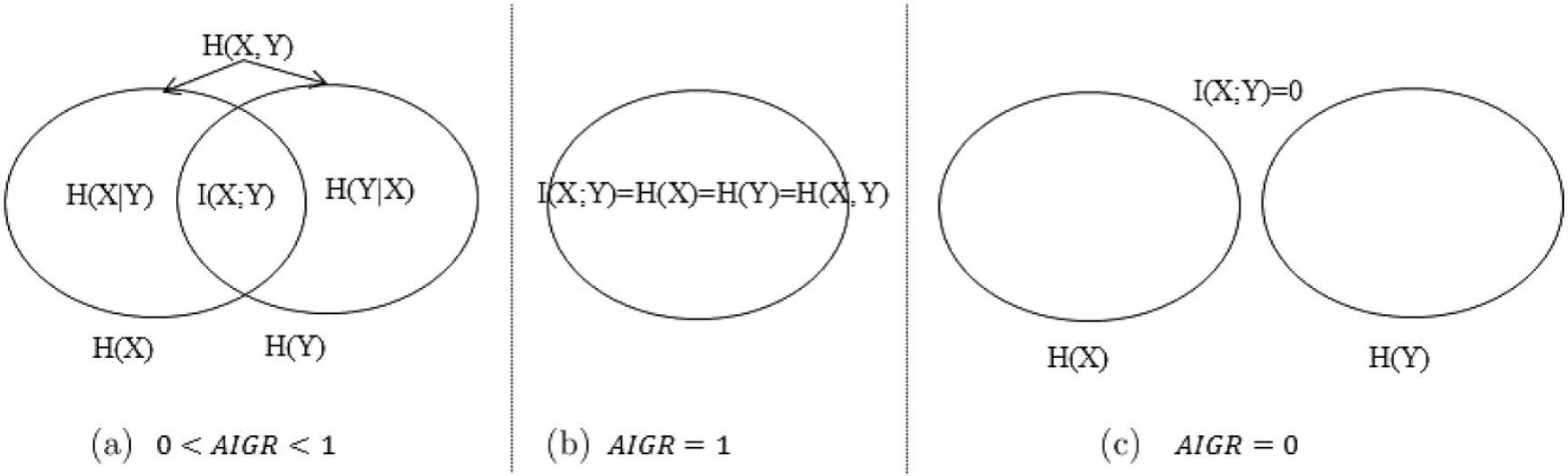


Fig. 2. AIGR relationship of X and Y. (a) General LD relationship of X and Y. (b) Complete LD of X and Y. (c) Complete LE of X and Y.

Definition of Average Information Gain Ratio (AIGR)

- Extension of entropy to $H(S_1, S_2)$. For two sets of SNPs S_1, S_2 we have $H(S_1, S_2) := \sum_{i=1}^{n} \sum_{j=1}^{m} p(s_{1_i}, s_{2_j}) \log_2 p(s_{1_i}, s_{2_j})$ where s_{1_i}, s_{2_j} are the i, j-th loci from set 1 and 2 respectively.
- Formal definition of Average Information Gain Ratio (AIGR).
  $AIGR(S_1, S_2) = \frac{1}{2}\left(\frac{I(S_1;S_2)}{H(S_1)} + \frac{I(S_1;S_2)}{H(S_2)}\right) \in [0, 1]$
- Bounds of AIGR.
  - $AIGR = 0$ when X, Y are in complete LE.
  - $AIGR \in (0, 1)$ when X, Y are inbetween LE and LD.
  - $AIGR = 1$ when X, Y are in complete LD, a high correlation between these two sets S_1, S_2.

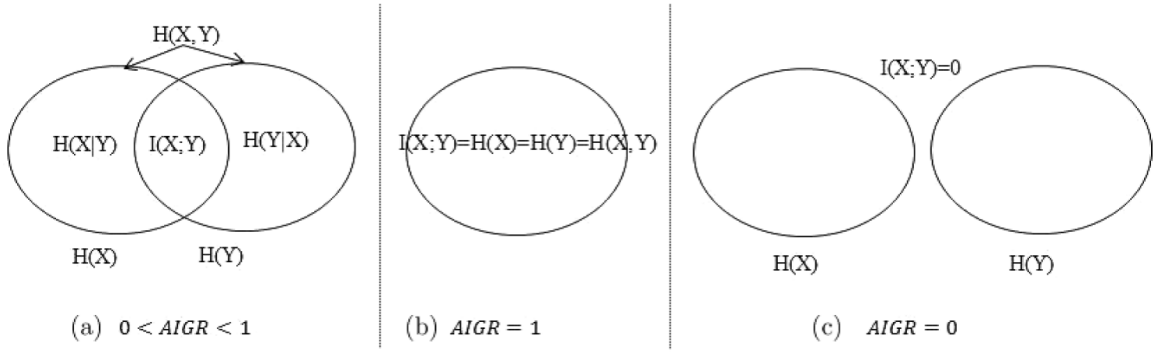


Fig. 2. AIGR relationship of X and Y. (a) General LD relationship of X and Y. (b) Complete LD of X and Y. (c) Complete LE of X and Y.

Tag SNPs selection using AIGR I

- **Tag SNPs selection**
  - Threshold selection
    - ⋆ A threshold selection will calculate the value of a chosen dissimilarity measure (in our case *AIGR*) between all SNPs and cluster them by a prescribed threshold θ. In this case we aggregate the individual SNPs by following conditions $\max_{i,j}\{AIGR(C_i, C_j)\} > \theta$ where $C_\bullet$ is the cluster of SNPs.
    - ⋆ Within each cluster, the SNPs have a very strong correlation with each other in sense that they have a very high value of *AIGR* (and also the interpretation of *AIGR*).
    - ⋆ Now we want to select one representative among all the SNPs for each one group. we can perform
      Deterministic selection.
      We choose the maximizer $arg\max_k \frac{1}{r}\sum_{l=1,l\neq k}^{r} AIGR(SNP_l, SNP_k)$ where SNP_l, SNP_k are two distinct SNPs from **the same cluster** obtained above, this cluster has r SNPs in total. Intuitively, we choose the SNP that has the highest total correlation with all other SNPs in **this cluster**.
      Random selection.

Tag SNPs selection using AIGR II

- **Tag SNPs evaluation**
  - Quality measure $\mathfrak{r}^2$ [7]. Based on a split of diploid data ("Haploidizing" diploid data) and obtain haplotype from genotype data.
  - Haplotype diversity measure H_p [8]. Based on total number of bases among different haplotypes for different loci combination.
  - Leave-one-out cross-validation. This is the evaluation method that is chosen by [1].

Comparison of Multilocus Measures and Tagging Algorithms I

There are advantages and drawbacks for each of these multilocus LD measures. Some are applicable to haplotype data only; others are applicable to allele data as well as haplotype data.

LD Measures	Multilocus	Genotype (Allele) data	Haplotype data	Range of measure
r^2	No	Yes	No	/
D	No	Yes	No	/
I_A	Yes	Yes	No	$[-1, 1]$
H_d	Yes	No	Yes	$[0, 1]$
RE_d	Yes	Yes	Yes	$[0, 1]$
$AIGR$	Yes	Yes	Yes	$[0, 1]$

It is also of interest to compare the tagging algorithm based on the AIGR (AIGR-Tagger) with other tagging algorithms. Suppose we have n samples with m SNPs.

tag SNP algorithm	Complexity	Additional parameters
Fast Tagger	$\mathcal{O}\left(n \cdot m^k\right)$	k is the number of SNPs in a tagging rule
FSFS	$\mathcal{O}\left(k \cdot m \cdot n^2\right)$	k is the parameter of KNN
AIGER-Tagger	$\mathcal{O}\left(n \cdot m^2\right)$	

Comparison of Multilocus Measures and Tagging Algorithms II

also a comparison between algorithms in terms of compression ratios $R = \frac{N_{\text{tag SNPs}}}{N_{\text{total SNPs}}}$ are given in [1]

Table 3. Summary of the number of selected tag SNPs, prediction accuracy and compression ratio on the haplotype datasets of AIGR-Tagger, Fast Tagger and FSFS.

Region	SNP numbers (MAF > 1%)	Number of tag SNPs			Prediction accuracy			Compression ratio		
		AIGR-Tagger	Fast Tagger	FSFS	AIGR-Tagger	Fast Tagger	FSFS	AIGR-Tagger	Fast Tagger	FSFS
ENm010	322	**79**	95	110	**0.9902**	0.9574	0.9489	**0.2453**	0.2950	0.3416
ENm013	483	**37**	57	92	0.9689	**0.9900**	0.9887	**0.0766**	0.1180	0.1905
ENm014	611	**72**	101	172	**0.9846**	0.9755	0.9408	**0.1178**	0.1653	0.2815
ENr112	751	**83**	114	230	**0.9820**	0.9581	0.9429	**0.1105**	0.1518	0.3063
ENr113	772	**70**	96	175	**0.9710**	0.9698	0.9592	**0.0907**	0.1244	0.2267
ENr123	772	**88**	120	223	**0.9784**	0.9720	0.9599	**0.1140**	0.1554	0.2889
ENr131	835	**124**	153	310	**0.9829**	0.8983	0.8567	**0.1485**	0.1832	0.3713
ENr213	529	**72**	86	185	**0.9915**	0.9726	0.9607	**0.1361**	0.1626	0.3497
ENr232	390	**69**	86	133	**0.9833**	0.9491	0.9230	**0.1769**	0.2205	0.3410
ENr321	484	**72**	88	128	**0.9806**	0.9690	0.9591	**0.1488**	0.1818	0.2645
Overall	5949	**766**	996	1758	**0.9813**	0.9612	0.9440	**0.1288**	0.1674	0.2955

Comparison of Multilocus Measures and Tagging Algorithms III

which is performed on a data set with following summary

Table 2. Datasets of ENCODE region SNPs.

Region name	Chromosome band	Genomic interval	Genotype SNP numbers (MAF > 1%)	Haplotype SNP numbers (MAF > 1%)	Genotyping group
ENm010	7p15.2	Chr7:26924045..27424045	756	322	UCSF-WU, Perlegen
ENm013	7q21.13	Chr7:89621624..90121624	1053	483	Broad, Perlegen
ENm014	7q31.33	Chr7:126368183..126865324	1135	611	Broad, Perlegen
ENr112	2p16.3	Chr2:51512208..52012208	1273	751	McGill-GQIC, Perlegen
ENr113	4q26	Chr4:118466103..118966103	1401	772	Broad, Perlegen
ENr123	12q12	Chr12:38626477..39126476	1312	772	BCM, Perlegen
ENr131	2q37.1	Chr2:234156563..234656627	1335	835	McGill-GQIC, Perlegen
ENr213	18q12.1	Chr18:23719231..24219231	882	529	Illumina, Perlegen
ENr232	9q34.11	Chr9:130725122..131225122	742	390	Illumina, Perlegen
ENr321	8q24.11	Chr8:118882220..119382220	877	484	Illumina, Perlegen

Comparison of Multilocus Measures and Tagging Algorithms IV

Further questions to be explored include (but not limited to)

- How will the new multilocus measure AIGR behaves on the SNPs sets S_1, S_2 consisting of rare-rare/rare-common/common-common(The variants in the paper [1] always have some sort of) variants?
- How will the new AIGER-Tagger algorithm improve the detection/simulation of LD haplotypic data?
- Since the AIGR is based on mutual information, could we improve the multilocus LD measure further when we already know some information about the individuals? For example, if the SNPs data are from a family or a sub-population.

Time for question and thanks for your attention!

References I

[1] Liao, Bo, et al. "New multilocus linkage disequilibrium measure for tag SNP selection." Journal of Bioinformatics and Computational Biology 15.01 (2017): 1750001.

[2] Agapow, Paul-Michael, and Austin Burt. "Indices of multilocus linkage disequilibrium." Molecular Ecology Resources 1.1-2 (2001): 101-102.

[3] Mueller, Jakob C. "Linkage disequilibrium for different scales and applications." Briefings in bioinformatics 5.4 (2004): 355-364.

[4] Sabatti, Chiara, and Neil Risch. "Homozygosity and linkage disequilibrium." Genetics 160.4 (2002): 1707-1719.

[5] Liu, Zhenqiu, and Shili Lin. "Multilocus LD measure and tagging SNP selection with generalized mutual information." Genetic epidemiology 29.4 (2005): 353-364.

[6] Lewontin, R. C. "The interaction of selection and linkage. I. General considerations; heterotic models." Genetics 49.1 (1964): 49.

References II

[7] Stram, Daniel O., et al. "Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study." Human heredity 55.1 (2003): 27-36.

[8] Carlson, Christopher S., et al. "Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans." Nature genetics 33.4 (2003): 518-521.