The Decay of Disease Association with Declining Linkage Disequilibrium: A Fine Mapping Theorem

Mehdi Maadooliat[1], Naveen K. Bansal[1], Jiblal Upadhya[1], Manzur R. Farazi[1], Zhan Ye[2], Xiang Li[2], Steven J. Schrodi[3,4]

[1]Statistics and Computer Science, Department of Mathematics, Marquette University, Milwaukee, WI, USA
[2]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, USA
[3]Center for Human Genetics, Marshfield Clinic Research Foundation, Marshfield, WI, USA
[4]Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, Madison, WI, USA

Correspondence:
Steven J. Schrodi
schrodi.steven@mcrf.mfldclin.edu

Abstract

Several important and fundamental aspects of disease genetics models have yet to be described. One such property is the relationship of disease association statistics at a marker site closely linked to a disease causing site. A complete description of this two-locus system is of particular importance to experimental efforts to fine map association signals for complex diseases. Here, we present a simple relationship between disease association statistics and the decline of linkage disequilibrium from a causal site. A complete derivation of this relationship from a general disease model is shown for very large sample sizes. Quite interestingly, this relationship holds across all modes of inheritance. Extensive Monte Carlo simulations using a disease genetics model applied to chromosomes subjected to a standard model of recombination are employed to better understand the variation around this fine mapping theorem due to sampling effects. We also use this relationship to provide a framework for estimating properties of a non-interrogated causal site using data at closely linked markers. We anticipate that understanding the patterns of disease association decay with declining linkage disequilibrium from a causal site will enable more powerful fine mapping methods.

Keywords

Disease association, linkage disequilibrium, haplotypes, two-locus system, causal site, fine-mapping, complex disease genetics

Introduction

Genetic markers closely linked to disease-causing sites will exhibit association with disease through linkage disequilibrium.[1-4] This is the central idea behind population-based association mapping of disease genes using high density SNP arrays.[5,6] However, the decay of disease association with declining linkage disequilibrium from a disease-predisposing, functional site has not yet been completely described even though this is a fundamental property of disease genetics. Doing so will provide much needed information concerning the properties of disease genetics and greatly aid experimental designs and statistical methods for identifying functional variants in regions that exhibit disease association.

Although many have argued that genome-wide association studies have been largely unsuccessful in that they have not revealed a large proportion of the heritability from most complex diseases,[7] it is certainly clear that numerous loci with impressive statistical evidence for correlation with a wide variety of complex diseases have been identified and replicated.[8] In a number of instances, these results have provided much needed insight into the biochemical pathways and cellular mechanisms responsible for increasing disease risk.[9-12] However, the functional variants underlying the majority of these disease-associated regions have yet to be identified and fully described.[13] The dearth of information concerning functional variants obviously presents a sizable impediment to further dissection of complex disease etiologies. If genetic and statistical methods can aid in generating either supporting or opposing evidence for the role of functional motifs within a region of association, then the progression of human genetics studies can be made much more efficient and potent.

When designing fine mapping genotyping experiments, it is important to select genetic variants and subregions so that adequately cover two types of disease models are adequately covered (i.e. the fine mapping design is well-powered to discover the functional variants). The first class of model that should be covered by such efforts would be models of a causal variant driving a portion, or perhaps all of the disease association within a region. Under this model, varying levels of association signal at different sites are explained by different levels of linkage disequilibrium with causal variants. Hence, given allele frequencies and linkage disequilibrium patterns, one can, in principle, back-calculate the properties of putative functional variants that could be driving an initially observed disease association within the region of interest. Known variants, including those that were not initially interrogated, fulfilling these calculated allele frequency and linkage disequilibrium properties with the initial markers should then be included in a fine-mapping panel. The second model to be covered by a fine-mapping panel of markers is one of allelic heterogeneity at a functional motif (e.g., a gene) that was originally found to exhibit a disease association signal. Empirical data tends to strongly favor this type of model over an individual variant serving as the sole driving allele within a region.[14-18] Indeed, it is quite typical for studies aiming to fine map regions harboring a GWAS-significant SNP to reveal multiple disease-correlated variants within the same gene. This is not terribly surprising as the site frequency spectrum is expected to contain vast numbers of rare variants in outbred populations, which is accentuated in rapidly expanding demographics.[19-21] Even if there is a small likelihood of any one of these rare variants to exhibit pathogenic effects, the sheer number of variants segregating at a gene trends to produce multiple functional alleles. To cover this class of disease models, one would want to reliably identify the functional motifs tagged by an initial association signal and proceed by exhaustively interrogating variants within those functional motifs. In practice, this two-model approach guiding fine mapping was employed successfully to identify alleles segregating at the *TRAF1-C5* region conferring susceptibility to rheumatoid arthritis.[22,23]

To address the statistical aspects of prioritizing potentially causal variants within a fine-mapped region, several methods have been developed including a useful Bayesian method was developed by Maller and colleagues,[24] which uses Bayes Factor for each variant in the region and calculates the proportion of the total sum of Bayes Factors in the region that is attributable to that variant, producing a relative ranking of the strength of evidence for each variant within the disease-associated region being causal. These calculations allow for the determination of a credible set of highest ranked variants that explains the large majority of the statistical evidence of disease association within the region of interest. The Maller et al. method has been applied to fine mapping data for complex diseases, such as type 1 diabetes.[25] Other important developments in fine mapping approaches include another Bayesian approach, Bim-Bam[26], methods which determine subsets of variants that likely contain causal sites, CAVIAR[27] and CAVIARBF[28], coalescent-based methods[29-31] and PAINTOR[32], which incorporates functional annotation data in a probabilistic manner.

Here, building upon previous work,[3,33-36] we prove a simple, analytic relationship between case/control association statistics at two closely-linked sites and the linkage disequilibrium between the two sites under a generalized disease genetics model. The result holds for very large sample sizes. Interestingly, the result is invariant with mode of inheritance parameters. Further, we posit that concurrently considering the patterns of disease-association and the genetic architecture within a region of interest may strengthen the ability to assess the likelihood that a particular variant is indeed causal with regard to inflating the risk of disease. By doing so, one may be better able to prioritize variants for functional follow-up studies. For finite sample sizes, dispersion around this relationship is expected and we therefore explore this variation in the result through the use of a Monte Carlo simulation.

Approximation

Under the Pritchard- Przeworski derivation[33], the power to detect disease association at a causal site and marker site were found to be approximately the same if the sample size at a marker site is increased by a factor of $(r^2)^{-1}$ over that used in interrogating the causal site. r^2 is the standard measure of linkage disequilibrium between the causal site and the marker site. While certainly an intriguing relationship between sample sizes, as it is, the finding may not always have utility in fine mapping applications as most association studies use the same number samples at all sites interrogated. That said, this relationship can be used to motivate related and illuminating properties regarding how fast the disease association signal can be expected to decay as a function of declining linkage disequilibrium from a causal site. Equating the power at the disease-predisposing site to that at the marker site, it follows that

$$\Phi\left(Z_D\sqrt{r^2} - Z_{1-\alpha/2}\right) \approx \Phi\left(Z_M - Z_{1-\alpha/2}\right); \qquad (1)$$

where Z_D and Z_M are the normally-distributed Z-scores for testing disease-association at the causal site and marker site, respectively; and α is the significance level. Taking the inverse functions and squaring yields the provocative approximation

$$\chi_M^2 \approx r^2 \chi_D^2\ ; \tag{2}$$

where χ_D^2 and χ_M^2 are the chi-squared statistics for disease association at the disease and marker sites, respectively. Plotting this approximation with the χ^2 disease-association statistic on the ordinate and $1-r^2$ on the abscissa is a simple method of displaying the expected linear decay in the χ^2 values as the linkage disequilibrium with a causal site declines at different marker sites. **Figure 1** shows this relationship. This decay pattern was first used empirically in 2008 to fine map the *IL23R* region in psoriasis[37] and has subsequently been used in other applications.[38] Although this approximation is very useful in understanding the decay of disease association with declining linkage disequilibrium from a causal site, several simplifying assumptions were made in the original Pritchard-Przeworski derivation and therefore it is not known how violations of the original assumptions might produce departures from **Eqn 2**. Hence, an exact relationship between disease association statistics and r^2 values with a causal site would aid in clarifying this relationship and motivate statistical approaches to harnessing this pattern for the purpose of fine-mapping functional alleles. Additionally, the allele frequencies are treated as parameters instead of random variables with sampling variances. So, understanding the dispersion around the decay patterns for finite sample sizes would further elucidate the relationships studied.

Full Derivation

Defining the Chi-Squared test statistics following the Pritchard-Przeworski derivation,

$$\chi_D^2 = \frac{[p_D - p_C]^2 \left[2n\left(\frac{n_D}{n_D+n_C}\right)\left(\frac{n_C}{n_D+n_C}\right)\right]}{p(1-p)}, \tag{3}$$

$$\chi_M^2 = \frac{[q_D - q_C]^2 \left[2n\left(\frac{n_D}{n_D+n_C}\right)\left(\frac{n_C}{n_D+n_C}\right)\right]}{q(1-q)}; \tag{4}$$

Where p, p_D, and p_C are the frequencies of the A_1 allele in the combined population, disease-affected population, and the control population, respectively and where q, q_D, and q_C are the frequencies of the B_1 allele in the combined population, disease-affected population, and the control population, respectively. n_D and n_C are the sample sizes for diploid cases and controls, respectively, and $n = n_D + n_C$. For this work, we will assume that cases and controls are drawn from the general population such that the cases and controls are drawn with probabilities corresponding to the disease and healthy proportions.

$$\frac{\chi_M^2}{\chi_D^2} = \frac{p(1-p)(q_D-q_C)^2}{q(1-q)(p_D-p_C)^2} \tag{5}$$

Noting that

$p = p_D K + p_C(1-K)$ and $q = q_D K + q_C(1-K)$,

we can substitute $p_C = \frac{p-Kp_D}{1-K}$ and $q_C = \frac{q-Kq_D}{1-K}$ into **Eqn (5)**, resulting in

$$\frac{\chi_M^2}{\chi_D^2} = \frac{p(1-p)\left(q-q_D\right)^2}{q(1-q)\left(p-p_D\right)^2} \tag{6}$$

The next aim in the derivation is to substitute quantities for the allele frequencies in the affected population at both sites in terms of penetrances, disease prevalence, and general population allele frequencies,

$$p_D = \frac{p}{K}[f_{11}p + f_{12}(1-p)] \tag{7}$$

$$q_D = \frac{P_{11}}{K}[f_{11}p + f_{12}(1-p)] + \frac{P_{21}}{K}[f_{12}p + f_{22}(1-p)]; \tag{8}$$

where $f_{11}, f_{12},$ and f_{22} are the prevalences of the $A_1A_1, A_1A_2,$ and A_2A_2 genotypes, respectively, such that $f_{ij} = P\left(Case \middle| A_iA_j\right)$; $K = P(Case)$, which, under this monogenic model and assuming Hardy-Weinberg Equilibrium in the general population, $K = f_{11}p^2 + 2f_{12}p(1-p) + f_{22}(1-p)^2$; and haplotype frequencies $P_{11} = P(A_1B_1)$, and $P_{21} = P(A_2B_1)$. Applied to complex diseases, it may be useful to think of this disease model as the subset of individuals with a common disease that is primarily driven by a particular locus. With the substitution into **Eqn 6**,

$$\frac{\chi_M^2}{\chi_D^2} = \frac{p(1-p)\left\{q-\frac{P_{11}}{K}\left[f_{11}p+f_{12}(1-p)\right]-\frac{P_{21}}{K}\left[f_{12}p+f_{22}(1-p)\right]\right\}^2}{q(1-q)\left\{p-\frac{p}{K}\left[f_{11}p+f_{12}(1-p)\right]\right\}^2} \tag{9}$$

In **Eqn 9**, the R.H.S. numerator can be simplified to

$$p(1-p)\left(\frac{1}{K^2}\right)(P_{11}-pq)^2[f_{11}p+f_{12}(1-2p)-f_{22}(1-p)]^2,$$

whereas the denominator in **Eqn 9** can be simplified to

$$q(1-q)\left(\frac{1}{K^2}\right)p^2[K-f_{11}p-f_{12}(1-p)]^2$$

Hence, **Eqn 9** can be written as

$$\frac{\chi_M^2}{\chi_D^2} = \frac{D^2(1-p)}{pq(1-q)}\frac{\left[f_{11}p+f_{12}(1-2p)-f_{22}(1-p)\right]^2}{\left[K-f_{11}p-f_{12}(1-p)\right]^2}; \tag{10}$$

where $D = P_{11}P_{22} - P_{12}P_{21} = P_{11} - pq$.

Substituting $K = f_{11}p^2 + 2f_{12}p(1-p) + f_{22}(1-p)^2$,

$$\frac{\chi_M^2}{\chi_D^2} = \frac{D^2(1-p)}{pq(1-q)}\frac{\left[f_{11}p+f_{12}(1-2p)-f_{22}(1-p)\right]^2}{\left[(1-p)\left(-f_{11}p-f_{12}(1-2p)-f_{22}(1-p)\right)\right]^2} \tag{11}$$

$$= \frac{D^2}{pq(1-p)(1-q)}\left[\frac{f_{11}p+f_{12}(1-2p)-f_{22}(1-p)}{f_{11}p+f_{12}(1-2p)-f_{22}(1-p)}\right]^2$$

$$= \frac{D^2}{pq(1-p)(1-q)}$$

$$= r^2 \tag{12}$$

Therefore, we have shown the exact relationship under our model

$$\chi_M^2 = r^2\chi_D^2 \tag{13}$$

Not only is this relationship an exact result under the model employed, but it is universal in that there is no dependence on the penetrances. Thus, we may expect that from a true disease-susceptibility site, that there should be a linear decay in the Chi-squared statistics for disease association with declining r^2 values with the causal site. **Figure 1** shows the expected disease association decay with declining linkage disequilibrium from the causal site for additive, multiplicative, recessive and dominant sets of models. The patterns arising from various relative risks are presented. Similarly, **Figure 2** presents the patterns expected as a function of sample sizes. Aside from **Eqn 13** illuminating a central aspect of disease genetics, we suspect that it carries utility in fine mapping applications – we hypothesize that identifying this type of pattern in fine mapping data will better enable the pinpointing of truly causal sites through harnessing correlated data.

Corollary

Consider the situation where there is a disease-susceptibility site and other sites in differing levels of linkage disequilibrium with the disease-susceptibility site. From large-scale genotyping or sequencing studies, we often know the matrix of pairwise r^2 values, and allele frequencies at each site in the general population, broadly defined. An interesting question arises: If one has genotyped a marker site in a case/control sample set and calculated χ^2_M testing for disease association, can we infer the expected effect size at a non-interrogated causal site? Using **Eqn 13**, and substituting allele frequencies at the causal site,

$$\frac{\chi^2_M}{r^2} = \frac{n_e(p_D - p_C)^2}{2p(1-p)}; \qquad (14)$$

Where $n_e = \frac{2n_D n_C}{n_D + n_C}$, the effective total number of independent diploid samples. For an allelic odds ratio at the causal site, R, the allele frequency in the cases can be written as

$$p_D = \frac{Rp_C}{1 - p_C + Rp_C}$$

Therefore,

$$\left(\frac{Rp_C}{1-p_C+Rp_C}-p_C\right)^2 = 2p(1-p)\frac{\chi_M^2}{n_e r^2} \tag{15}$$

To simplify the derivation, we will assume that the disease studied is not very common such that the allele frequency in controls is well-approximated by the allele frequency in the general population, $p_C \cong p$. This is also true if samples drawn from the general population are serving as the controls. Hence,

$$\frac{Rp}{1-p+Rp} = p + \left(\frac{Z_M}{r}\right)\sqrt{\frac{2p(1-p)}{n_e}} \tag{16}$$

Solving for R,

$$R = \left(\frac{1-p}{p}\right)\left[\frac{p+\sqrt{\frac{2p(1-p)\chi_M^2}{n_e r^2}}}{1-p-\sqrt{\frac{2p(1-p)\chi_M^2}{n_e r^2}}}\right] \tag{17}$$

To illustrate the use and implications of **Eqn 17**, suppose that we have genotyped a site in 500 diploid cases and 500 diploid controls and calculated the test statistic $\chi^2 = 20$, corresponding to p-value = 7.74E-06. Further assume that this region has previously been subjected to next-generation sequencing in individuals derived from the same source population as the cases and controls which has yielded the discovery of numerous additional variants closely linked to the genotyped site, allele frequencies at those variants, and an array of pairwise linkage disequilibrium values across the region of interest. Under that scenario, one would typically have access to good estimates of the general population allele frequencies and r^2 values at sites neighboring the genotyped site that produced the original finding. Suppose that one of these adjacent sites has a general population allele frequency $p = 0.03$ and a linkage disequilibrium value with the genotyped site of $r^2 = 0.2$. Under the two-site model, we would therefore estimate the odds ratio at the putative, non-genotyped, causal site to be 5.17. Put another way, the putative causal site, having the general population allele frequency and linkage disequilibrium values above, would have to have an odds ratio of 5.17 in order to generate a Chi-Squared statistic at the genotyped site of 20 given 500 cases and 500 controls. Indirect inference of the properties of non-interrogated causal sites can be helpful in subsequent experimental efforts to identify disease-predisposing sites in a fine-mapped region. **Figure 3** displays the relationship between the inferred odds ratio at the causal site from disease association data at the marker site as a function of linkage disequilibrium between the two sites. Graphs for various *p*-values at marker site are shown.

The results detailed in **Eqns 1-17** do not treat any of the parameters, such as haplotype frequencies, as random variables. Clearly, haplotype counts in cases and controls should be treated with sampling processes from a larger population. To address this issue, we have constructed a Monte Carlo simulation program to generate haplotypes under a probabilistic model. Under this program we are able to explore the variation around **Eqn 13** and to observe effects that may be produced by different sets of parameters.

Monte Carlo Simulations

In an effort to understand the variation in the patterns of disease association decay as a function of linkage disequilibrium with a causative site, we constructed a Monte Carlo simulation using a generalized disease model (penetrances for each of the three genotypes at the causal site are parameterized) and treating the haplotype counts in cases and controls as random variables. Recombination was introduced between a causal site and a closely linked marker as a realistic method of generating different sets of 2-site haplotypes for the general population.[39] For a rate of recombination, c, and generation time t, we used the following set of recursions (Haldane model of recombination):

$$P_{11,t} = P_{11,t-1}(1-c) + cpq \tag{18}$$

$$P_{12,t} = P_{12,t-1}(1-c) + cp(1-q) \tag{19}$$

$$P_{21,t} = P_{21,t-1}(1-c) + c(1-p)q \tag{20}$$

$$P_{22,t} = P_{22,t-1}(1-c) + c(1-p)(1-q) \tag{21}$$

The corresponding general population allele frequency at the causal site is

$$p_1 = P_{11,t} + P_{12,t} \tag{22}$$

Similarly, the general population allele frequency at the linked marker is

$$q_1 = P_{11,t} + P_{21,t} \tag{23}$$

In the absence of sampling (i.e., for an infinite population size), these will be invariant under the model considered. Assuming Hardy-Weinberg equilibrium in the general population at both sites, the proportion of individuals affected by the disease attributable to this locus, is calculated through the law of total probability,

$$K = p_1^2 f_{11} + 2p_1(1 - p_1)f_{12} + (1 - p_1)^2 f_{22} \tag{24}$$

To calculate the expected haplotype frequencies in cases, we used the above general population frequencies modified through the use of Bayes theorem.[35] Hence, the expected frequency of the A_1B_1 haplotype in cases is

$$V_{11} = \frac{P_{11}}{K}[f_{11}p_1 + f_{12}(1 - p_1)] \tag{24}$$

In an analogous manner, the remaining haplotype frequencies in cases, where the subscript indicates the haplotype, are

$$V_{12} = \frac{P_{12}}{K}[f_{11}p_1 + f_{12}(1 - p_1)] \tag{25}$$

$$V_{21} = \frac{P_{21}}{K}[f_{12}p_1 + f_{22}(1 - p_1)] \tag{26}$$

$$V_{22} = \frac{P_{22}}{K}[f_{12}p_1 + f_{22}(1 - p_1)] \tag{27}$$

The haplotype frequencies in controls are

$$U_{11} = \frac{(P_{11} - V_{11}K)}{1-K} \tag{28}$$

$$U_{12} = \frac{(P_{12} - V_{12}K)}{1-K} \tag{28}$$

$$U_{21} = \frac{(P_{21} - V_{21}K)}{1-K} \tag{28}$$

$$U_{22} = \frac{(P_{22} - V_{22}K)}{1-K} \tag{28}$$

Sampling of the haplotypes from the expected frequencies is accomplished through two independent multinomial variates (one for the cases and one for the controls), such that the joint densities are given by

$$P(X_{11} = x_{11}, X_{12} = x_{12}, X_{21} = x_{21}, X_{22} = x_{22}) = n_D!\left(\frac{U_{11,t}^{x_{11}} U_{12,t}^{x_{12}} U_{21,t}^{x_{21}} U_{22,t}^{x_{22}}}{x_{11}!\, x_{12}!\, x_{21}!\, x_{22}!}\right) \tag{29}$$

$$P(Y_{11} = y_{11}, Y_{12} = y_{12}, Y_{21} = y_{21}, Y_{22} = y_{22}) = n_C!\left(\frac{V_{11,t}^{x_{11}} V_{12,t}^{x_{12}} V_{21,t}^{x_{21}} V_{22,t}^{x_{22}}}{x_{11}!\, x_{12}!\, x_{21}!\, x_{22}!}\right) \tag{30}$$

Hence, the sample frequency of the causal allele in cases and controls, respectively, are

$$\hat{p}_D = (n_D)^{-1}(X_{11} + X_{12}) \text{ and } \hat{p}_C = (n_C)^{-1}(Y_{11} + Y_{12}) \tag{31 and 32}$$

(Plot of mean values)

(Plot of confidence intervals)

Conclusion and Discussion

One of the most fundamental patterns in disease genetics is the nature of the decay of disease association with declining linkage disequilibrium from a causal site. Motivated by the Lai et al and Pritchard-Przeworski derivations for the approximate increase in sample size to attain the equivalent statistical power at a marker site in linkage disequilibrium with a causal site, we first showed how this result could be used to produce an approximation showing a linear relationship in the Chi-Squared association statistics testing disease association at a marker and a causal site and that the ratio of the two was approximately r^2 (**Eqn 2**). Next, using a general two-site model with penetrances, we showed that this is indeed an exact result and invariant to the mode of inheritance model (**Eqn 13**).

Future work focusing on imputing additional properties of a non-interrogated causal variant, or multiple causal variants, within a disease-associated region using the linkage disequilibrium patterns and disease association statistics would provide valuable insights into design and interpretation of fine mapping studies.

Acknowledgments

We are greatly appreciative of the support from Dr. Murray Brilliant, Fritz Wenzel, Steve Ziemba, Terrie Kitchner, Cathy Marx and Marlene Stueland. This study was funded through the generous donations to the Marshfield Clinic Research Foundation and NIMH RO1 MH097464.

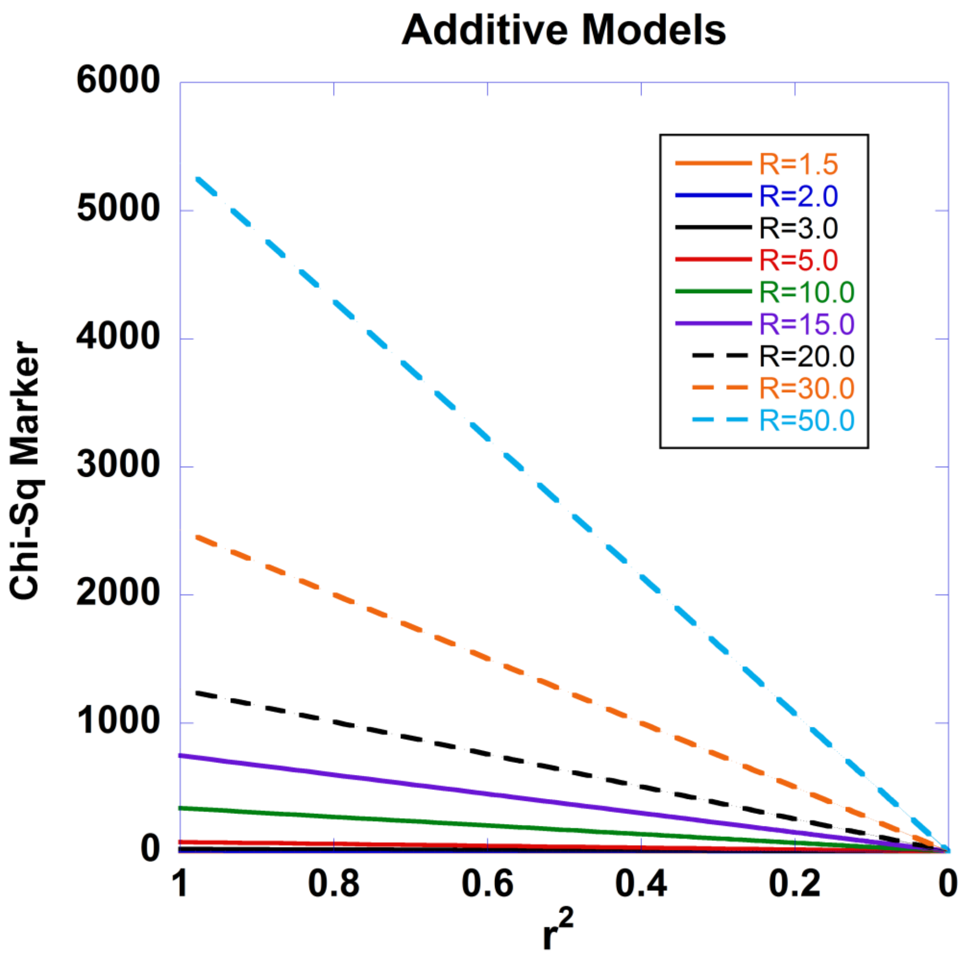
References

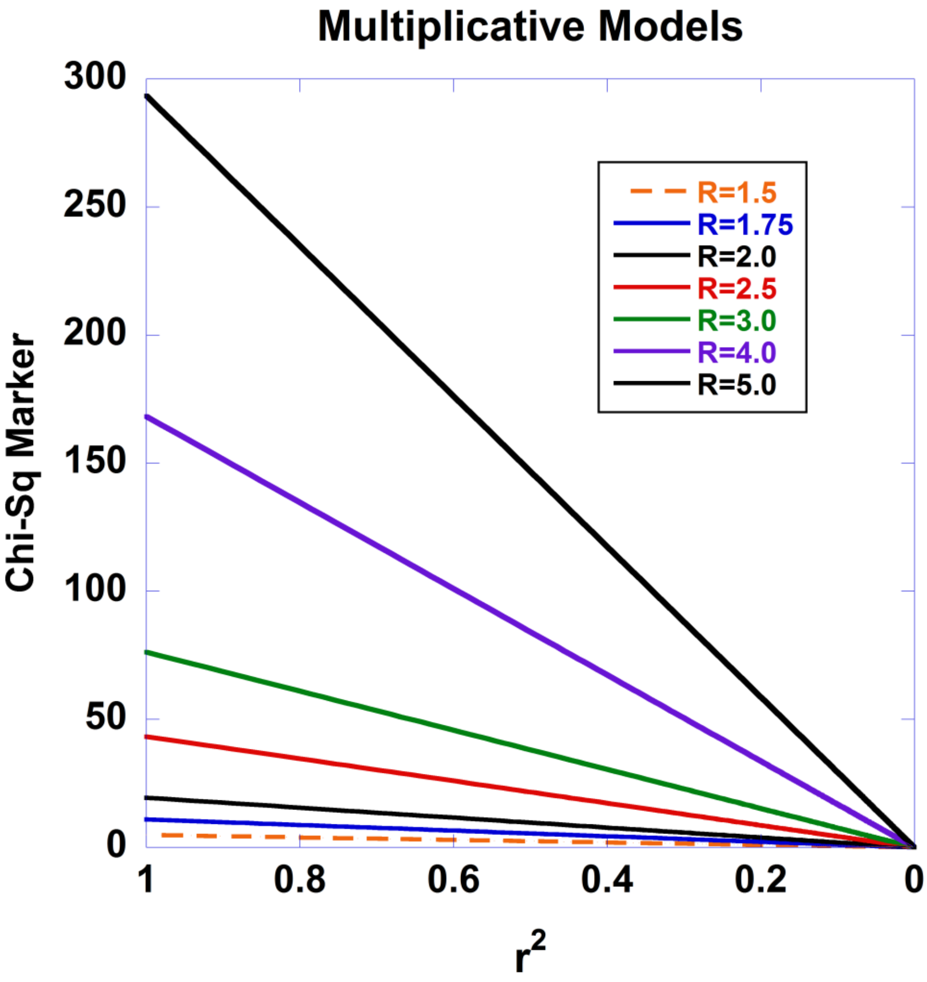
1. Morton NE (2005) Linkage disequilibrium maps and association mapping. J Clin Invest. 115(6):1425-1430.
2. Slatkin M (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477-485.
3. Lai C, Lyman RF, Long AD, Langley CH, Mackay TF (1994) Naturally occurring variation in bristle number and DNA polymorphisms at the scabrous locus of Drosophila melanogaster. Science 266(5191):1697-1702.
4. Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18:19-24.
5. McVean G, Spencer CC, Chaix R (2005) Perspectives on human genetic variation from the HapMap Project. PLoS Genet 1(4):e54.
6. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7(10):781-791.
7. Latham JR (2011) The failure of the genome. The Guardian.
8. Welter D, MacArthur J, Morales J, Burdett T, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42 (Database issue): D1001-D1006.
9. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90(1):7-24
10. Klein RJ, Zeiss C, Chew EY, Tsai JY, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308(5720):385-389.
11. Cargill M, Schrodi SJ, Chang M, Garcia VE, et al. (2007) A large-scale genetic association study confirms *IL12B* and leads to the identification of *IL23R* as psoriasis-risk genes. Am J Hum Genet 80(2):273-390.
12. Xavier RJ, Huett A, Rioux JD (2008) Autophagy as an important process in gut homeostasis and Crohn’s disease pathogenesis. Gut 57(6):717-720.

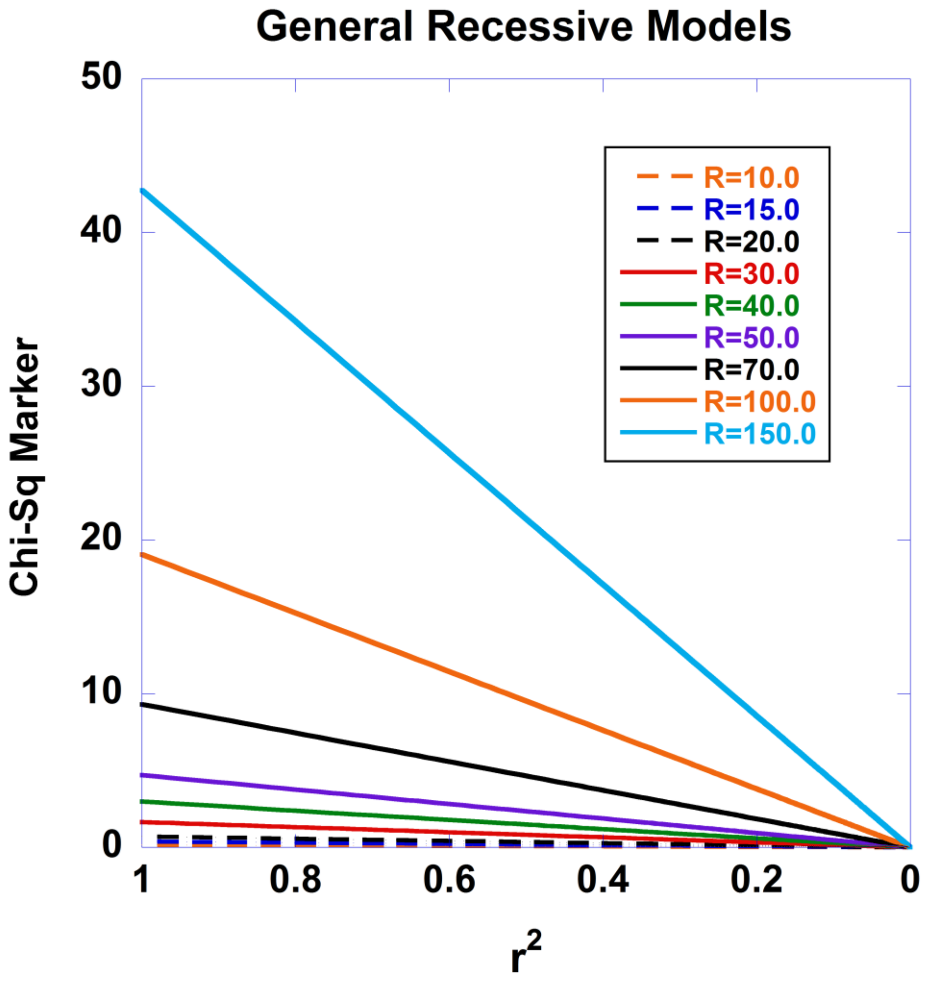
13. McClellan J, King MC (2010) Genetic heterogeneity in human disease. Cell 141:210-217.
14. Raychaudhuri S, Iartchouk O, Chin K, Tan PL, et al. (2011) A rare penetrant mutation in CFH confers high risk of age-related macular generation. Nat Genet 43:1232-1236.
15. Seddon JM, Yu Y, Miller EC, Reynolds R, et al. (2013) Rare variants in *CFI*, *C3* and *C9* are associated with high risk of advanced age-related macular degeneration. Nat Genet 45:1366-1370.
16. Nelson MR, Wegmann D, Ehm MG, Kessner D, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337(6090):100-104.
17. Kim-Howard X, Sun C, Molineros JE, Maiti AK, et al. (2013) Allelic heterogeneity in *NCF2* associated with systemic lupus erythematosus (SLE) susceptibility across four ethnic populations. Hum Mol Genet 23(6):1656-1668.
18. Rivas MA, Beaudoin M, Gardet A, Stevens C, et al. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 43:1066-1073.
19. Wright S (1931) Evolution in Mendelian Populations. Genetics 16:97-159.
20. Coventry A, Bull-Otterson LM, Liu X, Clark AG, et al. (2010) Depp resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Communications 1: Article 131.
21. Keinan A, Clark AG (2012) Recent explosive human population growth resulted in an excess of rare genetic variants. Science 336(6082):740-743.
22. Schrodi SJ, Chang M, Ardlie K, Amos CI, et al. (2007) A large-scale rheumatoid arthritis genetic study identifies TRAF1 variants on chr 9q33.2. ASHG Abstract 2007.
23. Chang M, Rowland CM, Garcia VE, Schrodi SJ, et al. (2008) A large-scale rheumatoid arthritis genetic study identifies association at chromosome 9q33.2. PLoS Genet 4(8):e1000107.
24. Wellcome Trust Case Control Consortium, Maller JB, McVean G, Byrnes J, et al. (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet 44(12):1294-1301.
25. Onengut-Gumuscu S, Chen W-M, Burren O, Cooper NJ, et al. (2015) Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet 47(4):381-386.
26. Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Gene 3(7):e114.
27. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E (2014) Identifying causal variants at loci with multiple signals of association. Genetics 198(2):497-508
28. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, et al. (2015) Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetics 200(3):719-736.
29. Graham J (1998) Disequilibrium fine-mapping of a rare allele via coalescent models of gene ancestry. UMI Dissertation Services. MI, USA.
30. Morris AP, Whittaker JC, Balding DJ (2002) Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am J Hum Genet 70:686-707.
31. Zollner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169(2):1071-1092.
32. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, et al. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet 10(10):e1004722.
33. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69(1):1-14.

34. Zaykin DM, Meng Z, Ehm MG (2006) Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Hum Genet 78:737-746.
35. Schrodi SJ, Garcia VE, Rowland C, Jones HB (2007) Pairwise linkage disequilibrium under disease models. Eur J Hum Genet 15(2):212-220.
36. Schrodi SJ, Garcia VE, Rowland CM (2009) A fine mapping theorem to refine results from association genetic studies. ASHG Abstract.
37. Garcia VE, Chang M, Brandon R, Li Y, et al. (2008) Detailed genetic characterization of the interleukin-23 receptor in psoriasis. Genes and Immun 9:546-555.
38. Farh KK, Marson A, Zhu J, Kleinewietfeld M, et al. (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518(7539):337-343.
39. Hartl DL, Clark AG (1989) *Principles of Population Genetics 2nd Ed*., Sinauer Associates, Inc. Sunderland, Massachusetts.

Figure 1. The Expected Decay of Disease Association with Declining Linkage Disequilibrium for Four Modes of Inheritance. The standard recursive haplotype frequencies under recombination were used to generate a series of haplotype combinations. The disease-predisposing allele at the causal site was set at a general population frequency of 0.01. The penetrance f_{22} was set to 0.001 and the remaining two penetrances varied according to the modes of inheritance examined and the relative risks cited in the Figures. Sample sizes were set at n_D=2000 and n_C=2000. **Fig. 1a** displays the results for an additive model, such that f_{12} is the arithmetic mean of f_{22} and f_{11}. **Fig. 1b** shows the results under a multiplicative model. **Fig 1c** shows the results under a general recessive model. **Fig 1d** shows the results under a general dominant model.




Multiplicative Models
300
250
200
150
100
50
0
Chi-Sq Marker
1
0.8
0.6
0.4
0.2
0
r²
R=1.5
R=1.75
R=2.0
R=2.5
R=3.0
R=4.0
R=5.0


General Recessive Models
Chi-Sq Marker
50
40
30
20
10
0
1
0.8
0.6
0.4
0.2
0
r^2
R=10.0
R=15.0
R=20.0
R=30.0
R=40.0
R=50.0
R=70.0
R=100.0
R=150.0

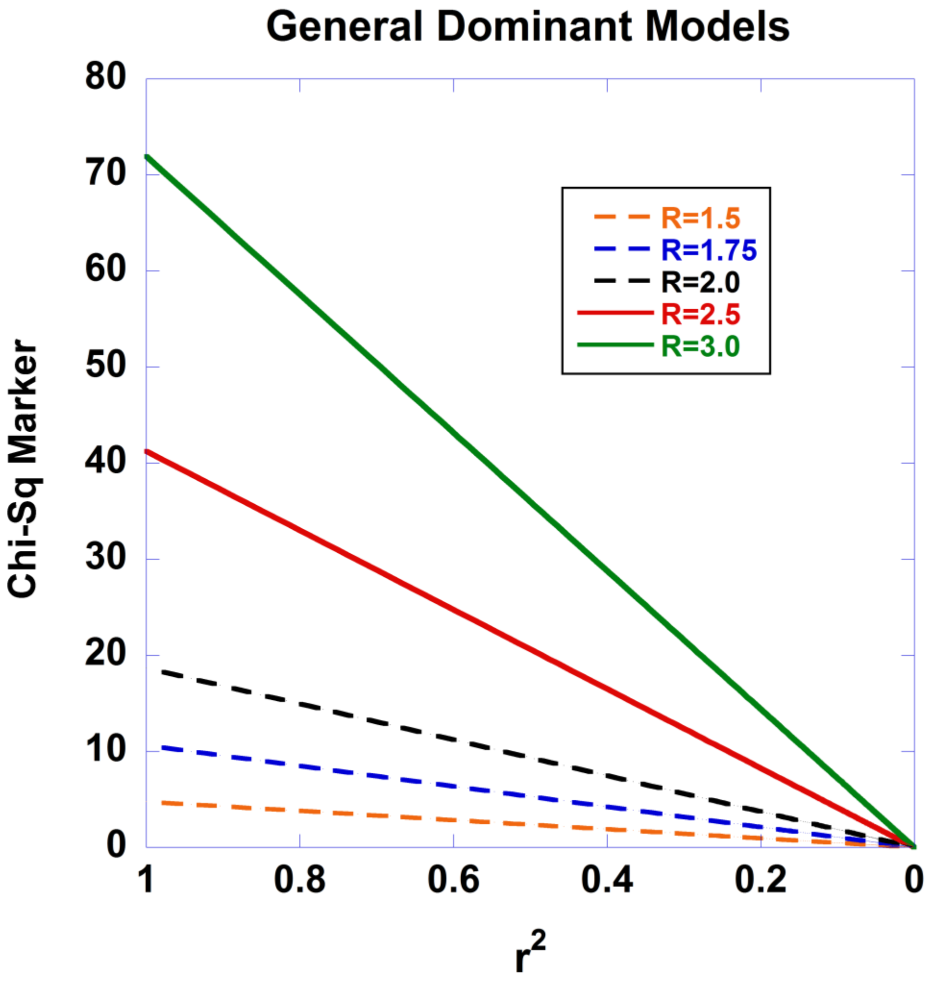

General Dominant Models
Chi-Sq Marker
80
70
60
50
40
30
20
10
0
1
0.8
0.6
0.4
0.2
0
r^2
R=1.5
R=1.75
R=2.0
R=2.5
R=3.0

Figure 2. Effect of Sample Size on the Expected Decay of Disease Association with Declining Linkage Disequilibrium.

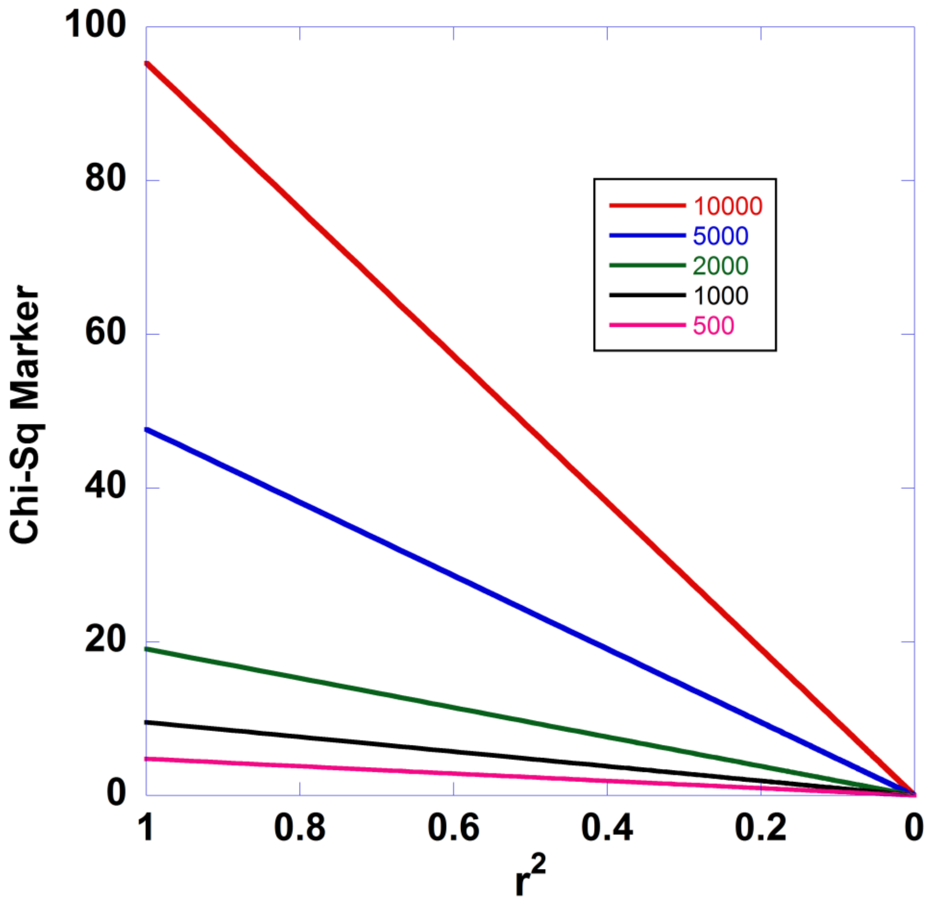
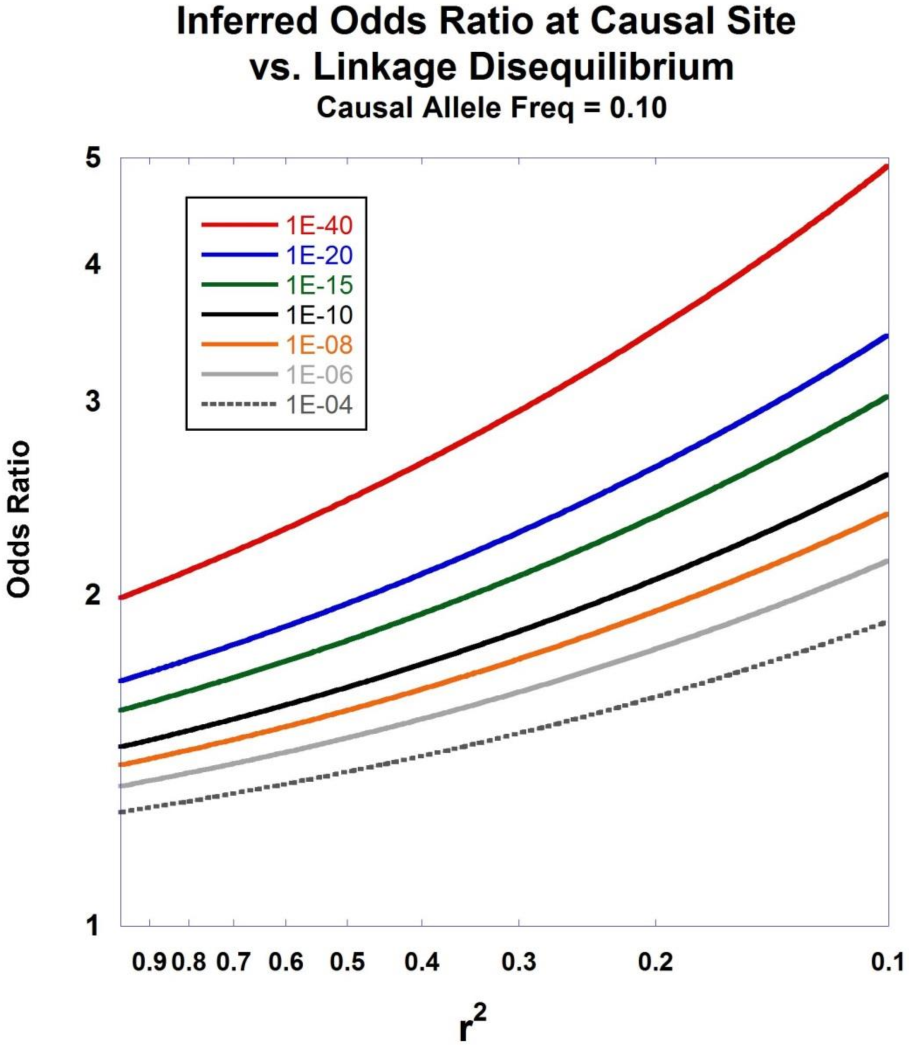
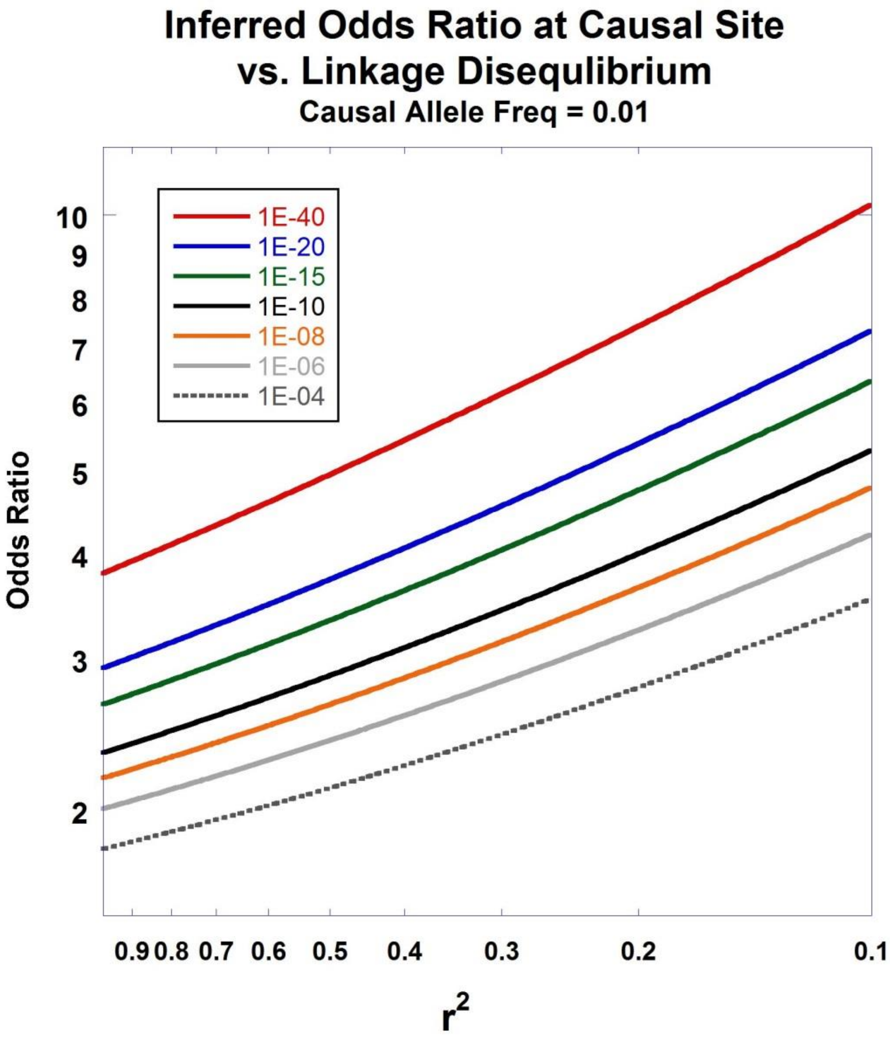


Figure 3. Inferred Odds Ratio. The relationship between the inferred odds ratio at a causal site and the level of linkage disequilibrium with an interrogated marker is presented in **Fig. 3a** and **Fig. 3b**. **Eqn 17** is used for the calculations. The seven curves show the patterns of expected odds ratios for disease association at the causal site under different observed p-values calculated at the marker site. Sample size was set at n_e=5000. **Fig. 3a** shows results assuming that the disease-predisposing allele at the causal site has frequency of 0.10 in the general population, whereas **Fig. 3b** sets that frequency at 0.01.




Inferred Odds Ratio at Causal Site
vs. Linkage Disequlibrium
Causal Allele Freq = 0.01
1E-40
1E-20
1E-15
1E-10
1E-08
1E-06
1E-04
Odds Ratio
10
9
8
7
6
5
4
3
2
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
r2