3

Powerful Testing via Hierarchical Linkage Disequilibrium in Haplotype Association Studies [1]

Summary

Marginal tests based on individual SNPs are routinely used in genetic association studies. Studies have shown that haplotype-based methods may provide more power in disease mapping than methods based on single markers when, for example, multiple disease-susceptibility variants occur within the same gene. A limitation of haplotype-based methods is that the number of parameters increases exponentially with the number of SNPs, inducing a commensurate increase in the degrees of freedom and weakening the power to detect associations. To address this limitation, we introduce a hierarchical linkage disequilibrium model for disease mapping, based on a re-parametrization of the multinomial haplotype distribution, where every parameter corresponds to the cumulant of each possible subset of a set of loci. This hierarchy present in the parameters enables us to employ flexible testing strategies over a range of parameter sets: from standard single SNP analyses through the full haplotype distribution tests, reducing degrees of freedom and increasing the power to detect associations. We show via extensive simulations that our approach maintains the type I error at nominal level and has increased power under many realistic scenarios, as compared to single SNP and standard haplotype-based studies. To evaluate the performance of our proposed methodology in real data, we analyze genome-wide data on rheumatoid arthritis from the Wellcome Trust Case-Control Consortium. The method is publicly available at `https://github.com/BrunildaBalliu/HierarchicalLD`.

[1] Submitted for publication.

3.1 Introduction

Marginal tests based on individual single nucleotide polymorphisms (SNPs) have dominated association analyses in the past decade. Although single SNP analyses have led to the identification of hundreds of genetic variants associated with many complex diseases [Hindorff et al., 2009], greater power might be achieved by using haplotype-based approaches, analyzing multiple markers simultaneously. Haplotype-based association methods incorporate linkage disequilibrium (LD) information from multiple markers and can be more powerful for gene mapping than methods based on single SNPs [Akey et al., 2001; Zaykin et al., 2002; Epstein and Satten, 2003]. For example, haplotype-based methods will be more powerful when multiple disease-susceptibility variants, each with an independent effect, occur within the same gene [Morris and Kaplan, 2002]. Moreover, haplotype-based methods could be preferable to single SNP-based association methods when diseases arise from the interaction of multiple cis-acting susceptibility variants found within a gene, forming a 'super-allele' [Joosten et al., 2001; Tavtigian et al., 2001; Hollox et al., 2001; Clark et al., 1998; Drysdale et al., 2000], since haplotype based methods allow for super-additivity of multiple genetic variants, whereas marginal tests do not [Epstein and Satten, 2003].

Standard haplotype association methods test for differences in haplotype distributions between cases and controls or perform regression analyses in which haplotypes are treated as categorical variables [Schaid et al., 2002; Zaykin et al., 2002; Epstein and Satten, 2003; Spinka et al., 2005; Lin and Zeng, 2006; Boehringer and Pfeiffer, 2009]. Two detailed reviews on existing methods for haplotype-based association analysis are provided by Schaid [2004] and Liu et al. [2008]. Moving from single-SNP to haplotype-based analyses results in a considerable increase in polymorphism and in a commensurate increase in the number of association parameters and therefore the degrees of freedom (d.f.) of the association tests. As a result, the global score or likelihood ratio test statistics will be weakly powered. Moreover, when the haplotype data is sparse, the χ^2 approximation of the distribution of the test statistics might be invalid. An additional difficulty is the ambiguity in haplotype phase when only genotype data are observed. Ambiguity can be handled using an expectation-maximization (EM) algorithm [Dempster et al., 1977; Excoffier and Slatkin, 1995], however, the additional assumption of Hardy-Weinberg equilibrium (HWE) is needed. The d.f. problem and the problem due to many rare haplotypes remain a limitation and force to employ heuristic methods, such as grouping of rare haplotypes [Schaid, 2004]. Due to these limitations of the haplotype-based methods and the myriad possible genetic architectures of complex human diseases, the relative efficiency of using haplotypes versus single markers remains largely unexplored and is often decided by practical considerations.

In this work, we introduce a hierarchical LD model for trait mapping that enables us to employ flexible testing strategies over a range of parameter sets: from standard single SNP analyses through the comparison of full haplotype distributions, thereby allowing to reduce d.f. and increase the power to detect associations. Our model is based on a re-parametrization of the multinomial haplotype distribution, where every parameter corresponds to the joint cumulant of each possible subset of a set of loci [Thiele, 1899; Brillinger, 1991]. For M SNPs, the new parametrization consists of allele frequencies of each SNP, standard pairwise LD parameters (i.e. D'), and higher-

order $(3, \ldots, M)$ LD parameters, corresponding to generalization of the pairwise LD to multiple SNPs. The proposed method is applicable to phased and unphased data and is particularly useful for detecting SNP-SNP interaction effects, long range differences in LD, the presence of 'super-alleles', and all situations where standard haplotype analysis would be considered. Moreover, due to properties of the hierarchy, direct optimization procedures can be constructed, rather than EM-based estimation. Higher order LD among alleles at more than two loci has been suggested in the past by Bennett [1952] and described in Weir [1990] for the case of three and four SNPs. However, to the best of our knowledge, a full parametrization of the haplotype distribution in terms of LD parameters, for an arbitrary number of SNPs, has not yet been provided.

In the following sections, we develop the re-parametrization of the multinomial haplotype distribution, describe estimation procedures and statistical tests with reduced d.f. for inference, and provide guidelines on how our method can be used. A simulation study, based on realistic haplotype distribution from the Wellcome Trust Case Control Consortium (WTCCC) [Burton et al., 2007] and different disease generating models show that the procedure maintains the type I error rate at nominal level and has increased power over the standard single SNP or haplotype based association methods for a variety of realistic scenarios. We apply our method to unphased SNP genotype data from the WTCCC data on rheumatoid arthritis (RA) and identify several new associations.

3.2 Material and methods

3.2.1 Basic notation and assumptions

Consider the case of genotype measurements of M bi-allelic loci. Let $h \in H$ be a haplotype at these loci, with $H = \{0,1\}^M$ the set of possible haplotypes, $|H| = 2^M$. We assume that $h \sim Mult(1, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_h)_{h \in H}$ the parameter vector of the haplotype frequencies, $\boldsymbol{\theta} \in \Theta$ and $\Theta = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \in (0,1)^{2^M}, \sum_{h \in H} \theta_h = 1\}$.

For the situation when genotypes instead of haplotypes are observed, let $\mathbf{G} = (G_1, \ldots, G_N)$ denote genotypes of N individuals; $D = (h_1, h_2)$ denotes a diplotype, *i.e.* an ordered haplotype pair, and $S(g)$ denotes the set of diplotypes that are consistent with genotype g. By assuming HWE, we can model the diplotype distribution using the product distribution. Then, the likelihood of the data can be expressed as [Schaid, 2004]

$$L_0(\mathbf{G}; \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{(h_1,h_2) \in S(G_i)} \theta_{h_1} \times \theta_{h_2}.$$

In the following, we consider case-control studies, with N_1 controls, N_2 cases and sample size $N = N_1 + N_2$. For genotypes $G = (G^{ca}, G^{co})$ the likelihood becomes

$$L(G, \theta) = L_0(G^{ca}, \boldsymbol{\theta}^{ca}) L_0(G^{co}, \boldsymbol{\theta}^{co}),$$

where $\boldsymbol{\theta}^{ca}$ and $\boldsymbol{\theta}^{co}$ are haplotype frequencies for cases and controls, respectively. Standard haplotype testing compares haplotype frequencies of cases and controls as

follows:

$$H_0 : \Theta^0 = \left\{(\boldsymbol{\theta}^{ca}, \boldsymbol{\theta}^{co}) \in \Theta^2 \mid \boldsymbol{\theta}^{ca} = \boldsymbol{\theta}^{co}\right\}, H_1 : \Theta^1 = \left\{(\boldsymbol{\theta}^{ca}, \boldsymbol{\theta}^{co}) \in \Theta^2\right\}. \quad (3.1)$$

Under the null hypothesis, parameters for cases and controls are constrained to be equal, while under the alternative any parameter component can differ between the groups. The EM algorithm can be used to maximize the log-likelihood and compute the maximum likelihood estimates under both the null and alternative hypothesis. The LR-statistic is then

$$LR = 2\left[\log L\left(\mathbf{G}; \hat{\boldsymbol{\theta}}^{\mathbf{1}}\right) - \log L\left(\mathbf{G}; \hat{\boldsymbol{\theta}}^{\mathbf{0}}\right)\right],$$

where $\hat{\boldsymbol{\theta}}^{\mathbf{0}} = \operatorname{argmax}_{\theta \in \Theta^0} L(\mathbf{G}; \boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}^{\mathbf{1}} = \operatorname{argmax}_{\theta \in \Theta^1} L(\mathbf{G}; \boldsymbol{\theta})$. It follows from standard likelihood theory that LR is asymptotically $\chi^2_{2^M-1}$ distributed.

3.2.2 Re-parametrization of the multinomial haplotype distribution

In order to achieve our goal of reducing the d.f., we present a hierarchical model of LD. To this end, Lemma 1 establishes a re-parametrization $\boldsymbol{\delta}$ of the multinomial haplotype frequencies $\boldsymbol{\theta}$, where every parameter corresponds to the joint cumulant of each possible subset of a set of M loci. We start by defining the joint cumulant.

Definition. *Let $A = \{A_1, A_2, \ldots, A_M\}$ be a set of random variables. Let P_A refer to the set of partitions of set A into nonempty subsets (blocks). So, for $p \in P_A$, each $b \in p$ is a block. Then, the joint cumulant of the set of random variables A is given as*

$$\kappa(A) = \kappa(A_1, A_2, \ldots, A_M) = \sum_{p \in P_A} (-1)^{|p|-1}(|p|-1)! \prod_{b \in p} E\left(\prod_{A \in b} A\right),$$

where $|p|$ denotes the cardinality of set p.

We also use M-th order cumulant to denote $\kappa(A)$. The joint cumulant is a measure of how far random variables are from independence [Ahlbach et al., 2012]. Notice that if M = 1 or M = 2, the joint cumulant reduces to the expected value and covariance, namely $\kappa(A_1) = \mathsf{E}(A_1), \kappa(A_1, A_2) = \mathsf{E}(A_1 A_2) - \mathsf{E}(A_1)\mathsf{E}(A_2)$.

Lemma 1. *Let $A = \{A_1, A_2, \ldots, A_M\}$ a set of M random variables with $A_j \in \{0, 1\}$. For each $s \in S = 2^A \setminus \emptyset$, let $\delta_s = \kappa(s)$, i.e. the joint cumulant of random variables s. Then $\boldsymbol{\delta} = (\delta_s)_{s \in S}$ is a re-parametrization of $\boldsymbol{\theta}$.*

Here 2^A denotes the power set of A. We interpret A_i as a bi-allelic locus and get that the haplotype distribution can be described by a set of cumulants for which each cumulant uniquely corresponds to a subset of the M loci. Note that first order cumulants correspond to allele frequencies and second order cumulants correspond to standard pairwise LD. Thus, in cases of two SNPs, the re-parametrization reduces to the standard decomposition into allele frequencies and pairwise LD parameters [Weir, 1990]. A proof of Lemma 1 is given in appendix A.1. For a set $\{A_1, A_2, A_3\}$

of random variables, we will write δ_{123} as a shorthand of $\delta_{\{A_1,A_2,A_3\}}$ and η_{123} for $\mathsf{E}(A_1A_2A_3)$. η_{123} is the haplotype frequency for loci 1,2, and 3 with allele 1 chosen at each locus.

As an example to illustrate the lemma, consider the case of three loci. The eight haplotype frequencies $\boldsymbol{\theta} = (\theta_{000}, \theta_{100}, \theta_{010}, \theta_{001}, \theta_{110}, \theta_{101}, \theta_{011}, \theta_{111})^\mathsf{T}$ can be re-parametrized into three allele frequencies, denoted by δ_1, δ_2, and δ_3, three pairwise LD parameters, denoted by δ_{12}, δ_{13}, and δ_{23}, and one third order LD parameter, denoted by δ_{123}, that is $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_{12}, \delta_{13}, \delta_{23}, \delta_{123})^\mathsf{T}$. The pairwise LD parameters for all pair (j, k) of SNPs are given as

$$\delta_{jk} = \mathsf{E}(A_jA_k) - \mathsf{E}(A_j) \times \mathsf{E}(A_k) = \eta_{jk} - \delta_j \times \delta_k. \tag{3.2}$$

As in the case of pairwise LD, higher order LD parameters express the difference between observed and expected haplotype frequencies, when expected frequencies are computed under the assumption of independence, with a value of zero indicating that at least two disjoint subsets of SNPs are independent of each other, and any cumulant involving two (or more) independent SNPs will be zero [Ahlbach et al., 2012]. This becomes apparent from the third order LD parameter:

$$\delta_{123} = \eta_{123} - \delta_1\eta_{23} - \delta_2\eta_{13} - \delta_3\eta_{12} + 2\delta_1\delta_2\delta_3. \tag{3.3}$$

3.2.3 Parameter estimation

The re-parametrization of the haplotype frequencies into allele frequencies and different order LD parameters introduces a hierarchy in the parameters. Specifically, higher order parameters (corresponding to singletons, pairs, triples, etc) only depend on lower order parameters and are independent of same or higher order parameters, given the lower order ones. This hierarchical structure enables us to construct direct optimization procedures avoiding the need for an EM algorithm.

As an example, consider again the case of three SNPs. In the first step we estimate the allele frequencies $\hat{\delta}_j$, $j = 1, 2, 3$. In the second step we estimate the pairwise LD parameters, denoted by $\hat{\delta}_{jk}, j \neq k$, for all pairs j, k of SNPs. Notice that in (3.2) each δ_{jk} depends only on allele frequencies δ_j and δ_k, which we have estimated in the first step, and a single parameter η_{jk} involving a one-dimensional optimization. Similarly, δ_{123} is estimated by a one-dimensional optimization over η_{123} as all other terms in (3.3) can be recovered by applying Lemma 1 from the parameters already estimated. The whole algorithm starts with allele frequencies and performs $2^M - 1 - M$ ensuing single-parameter optimizations.

3.2.4 Standardized LD parameters

LD parameters have the disadvantage of depending on allele frequencies [Hedrick, 1987]. For the two locus case, Lewontin [1964] suggested normalizing the pairwise LD parameter by dividing it by achievable extremes for fixed allele frequencies:

$$\delta_{jk}^{max} = \begin{cases} \min(\delta_j, \delta_k) - \delta_j\delta_k, & \text{if } \delta_{jk} \geq 0 \text{ and} \\ \left|\max(0, \delta_j + \delta_k - 1) - \delta_j\delta_k\right|, & \text{if } \delta_{jk} < 0. \end{cases} \tag{3.4}$$

We suggest to generalize this concept to establish a standardized LD measure for an arbitrary number of loci. Recall that δ_A can be written as

$$\delta_A = \eta_A - \sum_{p \in P_A \setminus A} (-1)^{|p|}(|p|-1)! \prod_{b \in p} \eta_b = \eta_A - \sum_{p \in P_A \setminus A} R_\delta(p),$$

where $R_\delta(p)$ are terms depending on loci $b \in p$ with $|b| < M$. These rest terms $R_\delta(p)$ are considered fixed and bounds for η_A are to be determined completely analogous to the two locus case. Then

$$\delta_A^{\max} = \begin{cases} \eta_A^{\max} - R_\delta, & \text{if } \delta_A \geq 0 \text{ and} \\ \left|\eta_A^{\min} - R_\delta\right|, & \text{if } \delta_A < 0, \end{cases} \tag{3.5}$$

where $R_\delta = \sum_{p \in P_A \setminus A} R_\delta(p)$, and $\eta^{\max}$ and $\eta^{\min}$ are the upper and lower bound for η_A and are defined in appendix A.2. The standardized version of δ_A is then given as follows

$$\delta_A^{'} = \frac{\delta_A}{\delta_A^{\max}}$$

A value of 1 or -1 indicates that the examined loci have not been exposed to all possible recombinations and at least one of all possible haplotype is not present in the population. $\eta_A^{\min}$ and $\eta_A^{\max}$ can be used to define the parameter space in the LD-parametrization which we denote with Δ in the following.

3.2.5 Parameter testing

The hierarchy present in our parametrization enables us to focus on certain orders in the the hierarchy, thus sparing d.f. as compared to testing the full distribution. We start by re-formulating the global haplotype test in terms of LD parameters. Let $\boldsymbol{\delta}^{ca} = (\delta_s^{ca})_{s \in S}$ and $\boldsymbol{\delta}^{co} = (\delta_s^{co})_{s \in S}$ be parameter vectors for cases and controls, respectively. Then (3.1) can be restated as follows

$$H_0 : \Theta_\delta^0 = \left\{(\boldsymbol{\delta}^{ca}, \boldsymbol{\delta}^{co}) \in \Delta^2 \mid \boldsymbol{\delta}^{ca} = \boldsymbol{\delta}^{co}\right\}, H_1 : \Theta_\delta^1 = \left\{(\boldsymbol{\delta}^{ca}, \boldsymbol{\delta}^{co}) \in \Delta^2\right\} \tag{3.6}$$

Again, $LR = 2\left(\log L\left(\mathbf{G}; \hat{\boldsymbol{\delta}}^{\mathbf{1}}\right) - \log L\left(\mathbf{G}; \hat{\boldsymbol{\delta}}^{\mathbf{0}}\right)\right) \sim \chi^2_{2^M-1}$ where $\hat{\boldsymbol{\delta}}^{\mathbf{0}}$, $\hat{\boldsymbol{\delta}}^{\mathbf{1}}$ are ML estimates under the null and alternative. We will refer to (3.6) as a *Full* test because we are testing all orders of LD parameters.

We now consider two families of tests with reduced d.f. The first family consists of tests that involve only lower order LD parameters. We will refer to them as *Bottom-Up* tests. Let P be the set containing the orders for which we would like to test for differences, *e.g.* $P = \{1, 2\}$ if we consider both allele frequencies and pairwise LD. The corresponding null and alternative hypotheses for any such set P is:

$$\begin{aligned} H_0 :& \Theta_{BU,P}^0 = \left\{(\boldsymbol{\delta}^{ca}, \boldsymbol{\delta}^{co}) \in \Delta^2 \middle| \quad \forall s \in S : |s| \in P \Rightarrow \delta_s^{ca} = \delta_s^{co}\right\} \\ H_1 :& \Theta_{BU,P}^1 = \Theta_\delta^1 \end{aligned} \tag{3.7}$$

Under H_0 we only constrain parameters of orders contained in P to be equal.

The second family consists of tests that involve only higher order LD parameters, *e.g.* for $M = 3$, $P = \{2, 3\}$ focuses only on second and third order LD parameters.

We will refer to them as *Top-Down* tests. The corresponding null and alternative hypotheses for any such set P is given by:

$$\begin{aligned} H_0 :&\Theta^0_{TD,P} = \Theta^0_\delta \\ H_1 :&\Theta^1_{TD,P} = \left\{(\boldsymbol{\delta}^{ca}, \boldsymbol{\delta}^{co}) \in \Delta^2 | \quad \forall s \in S : |s| \notin P \Rightarrow \delta^{ca}_s = \delta^{co}_s \right\} \end{aligned} \tag{3.8}$$

Here, parameters are constraint to be equal between cases and control both under H_0 and H_1 except for higher order parameters under the alternative. Both families of tests allow to employ direct optimization both under the null and the alternative. Since lower order parameters are estimated first, higher order parameters, which depend on the lower order parameters, will automatically be estimated to honor these constraints. On the other hand, had we constrained higher order parameters, lower order parameters would have to change once higher order constraints are considered. In these cases ML estimates would have to be found by joint optimization of parameters.

Top-Down tests can be interpreted as performing interaction tests without correcting for main effects. Uncorrected main effect can induce apparent interactions thereby allowing to reject some hypotheses where all differences come from main effects (or orders not included). For these reasons we will interpret these tests as global tests.

3.3 Simulation study

To evaluate the finite sample properties of the proposed re-parametrization and the association tests, we performed a simulation study. In the first part, we investigated type I error and power of the tests in data simulated based on real three-SNP haplotype frequencies from the WTCCC RA study. Here, we focus on the four most significant associations identified from the WTCCC data analysis. In the second part, we study the performance of the tests under several disease generating models, *e.g.* SNPs with main effects only, interacting pairs of SNPs and 'super-alleles'.

In each simulated data set, all tests described in the previous section were applied. For comparison purposes we also list results on the single SNP tests and score test performed using the `R` package `haplo.stats` [Sinnwell and Schaid, 2013]. For the scenarios under the null hypothesis, 10^3 data sets were simulated, each consisting of 2000 cases and 3000 controls. For the scenarios under the alternative hypothesis, 1000 data sets were simulated, also consisting of 2000 cases and 3000 controls.

3.3.1 Data simulation and results using real haplotype frequencies

For each of the four triplets identified as significant from the analysis of the WTCCC data, we estimated the haplotype frequencies in the sample of cases, the sample of controls and the pool of samples. We list these values in Table 3.1. The LD parameters to which these frequencies correspond are listed in Table A.1 of appendix A.4. In order to simulate data under the null hypothesis, we draw random samples from a multinomial distribution using the frequencies estimated from the pool of samples. In order to simulate data under the alternative hypothesis, we draw random

Table 3.1: Estimated haplotype frequencies in the cases (Ca), controls (Co) and pool (P) of cases and controls samples for each of the four triplets identified from the WTCCC data analysis.

	Triplet 1			Triplet 2		
	P	Ca	Co	P	Ca	Co
θ_{000}	.596	.569	.613	.499	.479	.512
θ_{001}	.059	.063	.056	.200	.189	.208
θ_{010}	.104	.098	.107	.015	.015	.015
θ_{011}	.003	.006	.002	.047	.054	.043
θ_{100}	.192	.211	.180	.147	.165	.135
θ_{101}	.028	.037	.022	.062	.059	.064
θ_{110}	.017	.014	.019	.025	.033	.020
θ_{111}	.002	.002	.001	.004	.006	.003

	Triplet 3			Triplet 4		
	P	Ca	Co	P	Ca	Co
θ_{000}	.477	.464	.486	.358	.340	.370
θ_{001}	.135	.172	.110	.088	.115	.071
θ_{010}	.029	.031	.028	.240	.228	.247
θ_{011}	.010	.008	.011	.101	.084	.112
θ_{100}	.115	.112	.115	.148	.166	.137
θ_{101}	.067	.060	.071	.004	.004	.004
θ_{110}	.132	.116	.142	.054	.058	.052
θ_{111}	.036	.035	.037	.006	.006	.006

samples separately for the group of cases and controls from a multinomial distribution using the frequencies estimated in the sample of case and controls, respectively.

Results on type I error rate for all tests and triplets are listed in Table 3.2. At the nominal level, type I error should lie in the interval (4.68, 5.31) for a test to properly maintain type I error. In general, the type I error rate is well maintained. For Triplet 1, the *Bottom-Up* test for allele frequencies and one of the single SNP tests is slightly deflated, while for Triplet 3, both tests are slightly inflated. For Triplet 2 and 4, the *Top-Down* test for third order LD is slightly inflated. Moreover, the *Full* test, is deflated for Triplet 2, type I error rate. All reject rates lie between 4.51 and 5.54.

The power for all tests and triplets is also listed in Table 3.2. For an association to be considered significant in the genome-wide associations study setting, the p-value of the test should be smaller than 5×10^{-8}. In all triplets the single SNP test and both *Top-Down* tests reach power below 80%. Regarding the other tests, different tests seem to be more powerful in each triplet with the *Bottom-Up* test for $P = \{1, 2\}$ being the one with the most consistent power across all triplets. In all triplets the score test from `haplo.stats` performs comparable to the *Full* test or the *Bottom-Up* test for $P = \{1, 2\}$.

Table 3.2: Result on type I error rate (%) and power (%) for the scenarios simulated based on parameters from significant findings from the WTCCC data. The parameter values for each scenario are listed in Table A.1. The *Bottom-Up* tests with $P = \{1\}$ and $P = \{1, 2\}$ test only for differences in allele frequencies and in allele frequencies and pairwise LD parameters; the *Full* test tests for differences in all parameters; the *Top-Down* tests with $P = \{3\}$ and $P = \{2, 3\}$ test only for differences in third order LD parameters and in second and third order LD parameters; the Single SNP tests are three separate one d.f. tests and `haplo.stats` is a score test from package `haplo.stats`. *The score tests from `haplo.stats` can have different d.f. in each data set because the package automatically groups rare haplotypes.

Test		d.f.	Triplet 1	Triplet 2	Triplet 3	Triplet 4
				Type I Error Rate (%)		
Bottom-Up	$P = \{1\}$	3	4.56	5.18	5.47	4.91
	$P = \{1, 2\}$	6	5.05	4.62	5.11	4.92
Full		7	5.13	4.51	5.18	5.16
Top-Down	$P = \{3\}$	1	4.93	5.54	5.36	5.97
	$P = \{2, 3\}$	4	4.98	4.87	5.03	5.35
Single SNP	SNP 1	1	5.1	5.14	5.54	5.26
	SNP 2	1	4.52	5.03	5.17	5.12
	SNP 3	1	4.99	4.71	5.28	4.78
`haplo.stats`		7*	5.29	4.93	5.17	4.86
				Power (%)		
Bottom-Up	$P = \{1\}$	3	66.90	74.80	89.30	71.30
	$P = \{1, 2\}$	6	71.43	70.00	94.80	97.90
Full		7	66.30	65.60	96.70	97.30
Top-Down	$P = \{3\}$	1	0.00	0.00	0.00	0.00
	$P = \{2, 3\}$	4	0.20	0.00	4.40	24.10
Single SNP	SNP 1	1	20.10	23.80	9.80	10.40
	SNP 2	1	0.00	15.70	1.60	9.80
	SNP 3	1	15.20	0.00	47.30	0.00
`haplo.stats`		7*	70.50	69.00	96.80	97.30

3.3.2 Data simulation and results under different disease generating models

In this section we further study the type I error rate and power properties of each test under different disease models and different LD structures. In all scenarios, we considered four SNPs with allele frequencies equal to .05, .18, .31 and .45, respectively. Two structures of LD among the SNPs are considered. In Scenario 1, the SNPs were in equilibrium, thus all second, third and fourth LD parameters were equal to zero. In Scenario 2, the second order standardized LD parameters were set to .4, the third order LD standardized parameters were set to .1 and the fourth order LD parameter was set to zero. In both cases, we mapped the LD parameters to haplotype frequencies, which are listed in Table A.2 of appendix A.4, and used those frequencies to generate haplotype data for a large population of individuals. The LD parameters in Scenario 1 correspond to frequencies in which 11 out of 16 haplotypes had frequencies below 5% and six had frequencies below 1%. On the other hand, in Scenario 2 only four haplotypes had frequencies below 5%.

Using different disease models, we generate the disease status Y of each individual and then sampled 2000 individual from the population of cases and 3000 individuals from the population of controls. For each disease model the following logistic model was used

$$\text{logit}(\text{P}(Y = 1 \mid \mathbf{D})) = \alpha_0 + \sum_{j=1}^{4} \alpha_j G_j + \sum_{j,k=1, j\neq k}^{4} \alpha_{ij} G_j \times G_k + \sum_{s \in S} \gamma_s SA_s \quad (3.9)$$

where α_0 is the intercept; $\alpha_j, j = 1, \ldots, 4$ are the main effect odds ratios of each SNP, $\alpha_{jk}, j, k = 1, \ldots, 4, j \neq k$ are the interaction effect for each pair of SNP; γ_s are the main effects of the 'super-allele' at loci $s \in S$, with $S = \{\{2, 3\}, \{1, 2, 3\}, \{1, 2, 3, 4\}\}$ and

$$SA_{23} = \begin{cases} 0 & \text{if both } h_1 \text{ and } h_2 \notin D_{23} \\ 1 & \text{if one of } h_1, h_2 \in D_{23} \\ 2 & \text{if both } h_1 \text{ and } h_2 \in D_{23} \end{cases}, SA_{123} = \begin{cases} 0 & \text{if both } h_1 \text{ and } h_2 \notin D_{123} \\ 1 & \text{if one of } h_1, h_2 \in D_{123} \\ 2 & \text{if both } h_1 \text{ and } h_2 \in D_{123} \end{cases},$$

$$\text{and } SA_{1234} = \begin{cases} 0 & \text{if } h_1 \neq \text{`1111'} \text{ and } h_2 \neq \text{`1111'} \\ 1 & \text{if } h_1 = \text{`1111'}, h_2 \neq \text{`1111'} \text{ or } h_1 \neq \text{`1111'}, h_2 = \text{`1111'} \\ 2 & \text{if } h_1 = \text{`1111'} \text{ and } h_2 = \text{`1111'} \end{cases},$$

where $D_{23} = \{\text{`0110'}, \text{`1110'}, \text{`0111'}, \text{`1111'}\}$, i.e. all haplotypes that contain the '1' allele at loci 2 and 3 and $D_{123} = \{\text{`1110'}, \text{`1111'}\}$ the haplotypes that contain the '1' allele at loci 1, 2 and 3.

Under the null hypothesis, all parameters in (3.9), besides the intercept, were zero. Results on type I error rate for all tests and scenarios are listed in Table 3.3. For Scenario 2, in which the four SNPs were in LD, all tests properly control the type I error rate. For Scenario 1, however, some tests are deflated, *Bottom-Up* tests with $P = \{1, 2, 3\}$, the *Full* test and all three *Top-Down* tests, while `haplo.stats` is inflated.

For scenarios under the alternative hypothesis, six different disease models were considered. In Model 1, the four SNPs had only main effects on disease risk. In Model 2, SNP 2 and 3 had main and interaction effects on disease risk. In model 3, SNPs 1, 2 and 3 had only interaction effects. We also studied the power of our approach in the presence of 'super-alleles'. In this case we assumed that the combination of alleles over two, three and

Table 3.3: Result on type I error rate for each test and each scenario listed in Table A.2. The *Bottom-Up* tests with $P = \{1\}$ and $P = \{1, 2\}$ test only for differences in allele frequencies and in allele frequencies and pairwise LD parameters; the *Full* test tests for differences in all parameters; the *Top-Down* tests with $P = \{3\}$ and $P = \{2, 3\}$ test only for differences in third order LD parameters and in second and third order LD parameters; the Single SNP tests are three separate one d.f. tests and `haplo.stats` is a score test from package `haplo.stats`. *The score tests from `haplo.stats` can have different d.f. in each data set because the package automatically groups rare haplotypes.

Test		d.f.	Type I Error Rate	
			Scenario 1	Scenario 2
Bottom-Up	$P = \{1\}$	4	4.67	5.24
	$P = \{1, 2\}$	10	4.86	5.00
	$P = \{1, 2, 3\}$	14	4.43	4.76
Full		15	4.31	4.91
Top-Down Tests	$P = \{4\}$	1	4.22	5.04
	$P = \{3, 4\}$	5	3.93	5.00
	$P = \{2, 3, 4\}$	11	4.36	4.79
Single SNP	SNP 1	1	4.97	5.45
	SNP 2	1	4.59	5.03
	SNP 3	1	4.83	5.13
	SNP 4	1	5.20	5.20
`haplo.stats`		15*	6.09	5.03

four SNPs also had an effect of disease risk. In model 4, SNP 2 and 3 and the haplotype '11' over these two loci had a main effect; in model 5, SNP 1, 2 and 3 and the haplotype '111' had a main effect and in model 6, all four SNPs and the haplotype '1111' had a main effect. Results on power for all tests and models, as well as the exact parameter values for each model, are listed in Table 3.4 for Scenario 1 and in Table 3.5 for Scenario 2.

Based on these results we make the following observations. First, as expected, in the presence only of main effects, i.e. Model 1, for both Scenarios, the most powerful test is the *Bottom-Up* test with $P = \{1\}$. Second, although the *Bottom-Up* test with $P = \{1\}$ does not include second order parameters, its power is comparable to the power of *Bottom-Up* test with $P = \{1, 2\}$ in the presence of both main and interaction effects, i.e. Model 2, or in the presence only of interacting effects, i.e. Model 3. In the presence of 'super-alleles', the power to detect association when the LD among the involved loci is zero and the effect is spread accross three or four loci, i.e. Model 5 and 6 in Scenario 1, is much lower compared to the power in the presence of LD, i.e. Model 5 and 6 in Scenario 2. For both scenarios and all models, except Model 6 for Scenario 1, at least one of the *Bottom-Up* tests is more powerful than `haplo.stats`. The *Bottom-Up* test with $P = \{1, 2\}$ was the one with the most consistent power across all models and scenarios.

Table 3.4: Result on power of each test on Scenario 1. Non-zero parameters for Model 1: $\alpha_i = log(1.2)$, $i = 1, 2, 3, 4$; Model 2: $\alpha_2 = log(1.2), \alpha_3 = log(1.1)$, $\alpha_{12} = log(1.2)$; Model 3: $\alpha_{jk} = log(1.3)$, $j, k = 1, 2, 3$, $j \neq k$; Model 4: $\alpha_1 = \alpha_2 = log(1.1)$, $\gamma_{23} = log(1.5)$; Model 5: $\alpha_i = log(1.1)$,$k = 1, 2, 3$, $\gamma_{123} = log(1.5)$; Model 6: $\alpha_i = log(1.1)$, $i = 1, 2, 3, 4$, $\gamma_{1234} = log(5)$.

Test		d.f.	Power with 95 % CI					
			Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	$P = \{1\}$	4	90.32	90.72	68.67	96.50	32.04	13.18
Bottom-Up	$P = \{1, 2\}$	10	74.82	86.26	89.41	94.52	19.75	10.16
	$P = \{1, 2, 3\}$	14	63.37	79.20	82.89	91.26	15.67	8.52
Full		15	59.57	77.45	81.47	90.24	14.21	8.11
	$P = \{4\}$	1	.00	.00	.00	.00	.00	.00
Top-Down	$P = \{3, 4\}$	5	.00	.00	.00	.00	.00	.00
	$P = \{2, 3, 4\}$	11	.00	.00	1.01	.00	.00	.00
	SNP 1	1	.00	.00	4.10	.00	.00	.00
Single SNP	SNP 2	1	2.70	79.80	19.10	89.40	1.40	.10
	SNP 3	1	10.50	11.00	3.70	16.90	1.20	.40
	SNP 4	1	16.00	.00	.00	.00	1.50	.10
`haplo.stats`		15	65.60	80.10	83.70	90.50	18.32	22.10

Table 3.5: Result on power of each test on Scenario 2. Non-zero parameters for Model 1: $\alpha_1 = \alpha_2 = log(1.2)$, $\alpha_3 = \alpha_4 = log(1.1)$; Model 2: $\alpha_2 = \alpha_3 = log(1.1)$, $\alpha_{12} = log(1.2)$; Model 3: $\alpha_{jk} = log(1.2)$, $j, k = 1, 2, 3$, $j \neq k$; Model 4: $\alpha_1 = \alpha_2 = log(1.1)$, $\gamma_{23} = log(1.3)$; Model 5: $\alpha_i = log(1.1)$,$k = 1, 2, 3$, $\gamma_{123} = log(1.3)$; Model 6: $\alpha_i = log(1.1)$, $i = 1, 2, 3, 4$, $\gamma_{1234} = log(2)$.

Test		d.f.	Power with 95 % CI					
			Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Bottom-Up	$P = \{1\}$	4	75.92	91.67	83.24	77.42	87.37	80.47
	$P = \{1, 2\}$	10	48.37	88.42	79.07	65.59	75.79	71.59
	$P = \{1, 2, 3\}$	14	36.12	81.25	69.48	62.37	64.21	62.60
Full		15	32.96	78.12	68.07	61.29	60.00	60.93
Top-Down	$P = \{4\}$	1	.00	.00	.00	.00	.00	.00
	$P = \{3, 4\}$	5	.00	.00	.00	.00	.00	.00
	$P = \{2, 3, 4\}$	11	.00	.00	.00	.00	.00	.00
Single SNP	SNP 1	1	2.40	.00	32.40	.00	.00	18.30
	SNP 2	1	40.60	85.00	50.40	70.00	50.00	15.20
	SNP 3	1	10.40	65.00	10.40	34.00	31.00	15.60
	SNP 4	1	6.70	.00	.00	1.00	17.00	10.90
`haplo.stats`		15*	37.10	80.00	71.70	61.00	63.00	67.10

3.4 Data example

To illustrate an application of the proposed association tests, we performed an analysis of a data set from the WTCCC, consisting of 1860 cases of RA and 2938 controls. In the initial analysis, single SNP tests were performed and several SNPs, strongly associated with RA, were identified [Burton et al., 2007]. In addition, a list of 59 SNPs, showing 'moderate' association with RA, with nominal significance in the range of 10^{-3} to 10^{-6}, was provided in the initial article. Some of these SNPs map to genes with plausible biological relevance however the single SNP analyses failed to pass the significance threshold.

Here, we investigate possible increase in the significance level of the 59 SNPs when a three SNP haplotype based analysis is used. For each of these SNPs we choose 40 neighboring SNPs that had passed quality control, 20 to the left and 20 to the right side of the SNP and construct all possible triplets between the SNPs that contain the moderately associated SNP. For each of the 59 SNP, 780 triplets were constructed. To avoid problems caused by high LD, we excluded from the analysis all triplets in which at least one of the standardized pairwise LD parameters was above 0.8. For the remaining triplets, the tests mentioned in the previous section were applied. For comparison purposes, we also show results from single-SNP analysis. A triplet of SNPs was considered to be associated with RA if the p-value exceeded the threshold $5 \times 10^{-8} / (N_{tests} \times N_{triplets})$, where $N_{tests} = 5$ is the total number of tests performed on each triplet and $N_{triplets}$ the total number of triplets tested for each 'moderately' associated SNP.

Several triplets containing the SNPs rs12723859 and rs12205634 showed a strong association with RA. Specifically, for rs12723859 we identified 40 triplets with 20 unique SNPs, and for rs12205634 we identified 5 triplets with 4 unique SNPs. For rs6920220, 3 triplets consisting of 4 unique SNPs, had p-values smaller than the genome-wide significance threshold 5×10^{-8} but they were no longer significant when adjusting for the multiple number of tests and triplets. For the other 56 SNPs no strong association with RA was identified from the haplotype analysis. In Table 3.6 we list for each of rs6920220, rs12723859, and rs12205634, the p-values of all tests for the two triplets that show the strongest association with RA. For rs6920220 we tested a total of 21 triplets. Only the *Bottom-Up* test for allele frequencies yields a p-value below 5×10^{-8}. If we correct for the number of tests and triplets tested no test yields a significant p-value. For rs12723859 and rs12205634 we tested a total of 144 and 38 triplets respectively. The *Full* test and the *Bottom-Up* tests for $P = \{1\}$ and $P = \{1, 2\}$ yield p-values below 5×10^{-8}. After correcting for the number of tests performed the *Bottom-Up* tests for $P = \{1\}$ no longer gives a significant association, the *Bottom-Up* tests for $P = \{1, 2\}$ is still significant.

Table 3.6: Results on real data

SNPs in the triplet		*Bottom-Up* Test		*Full* Test	*Top-Down* Test		Single SNP tests		
		$P = \{1\}$	$P = \{1, 2\}$		$P = \{3\}$	$P = \{2, 3\}$			
				SNP rs6920220					
rs11961920	rs11970411	2.6e-08	7.9e-08	2.3e-07	.89	.21	5e-06	.16	1.2e-05
rs11970411	rs674451	8.5e-09	9.9e-08	2.8e-07	.81	.56	5e-06	1.2e-05	.25
				SNP rs12723859					
rs12739961	rs1113523	1.8e-10	4.4e-11	5.6e-12	7.78e-03	8.50e-04	3e-05	.0013	2.2e-07
rs12739961	rs17013326	2.4e-10	7.3e-11	7.9e-12	6.40e-03	9.29e-04	3e-05	.0013	3.1e-07
				SNP rs12205634					
rs411136	rs210137	1.9e-08	4.8e-12	1.2e-11	.41	2.26e-05	5.2e-05	4.3e-05	6.9e-02
rs411136	rs210138	2.1e-08	1.1e-11	2.1e-11	3.7e-05	5.1e-05	5.2e-05	4.3e-05	6.9e-02

3.5 Discussion

In this article, we propose a re-parametrization of the multinomial haplotype distribution into allele frequencies, standard pairwise LD parameters, and higher-order LD parameters. Our re-parametrization enables us to employ flexible testing strategies over a range of parameter sets. For example, joint tests of single-SNPs and joint tests of single-SNPs and their pairwise LD showed in both simulated and real data that such tests can often have increased power as compared to the full global haplotype or single-SNP based tests.

In this study, we use rather simplistic multiple testing strategies, namely using a Bonferroni correction for multiple tests performed on the same genotype data. This is certainly not optimal as the performed tests are usually highly correlated. Among our future interests is to develop iterative or sequential testing procedures, *e.g.* [Meinshausen, 2008], which better exhaust the α level. Moreover, we have not focused on the choice of haplotype size or region covered as an optimal strategy. It is likely that the optimal number of SNPs used for haplotype-based approaches will depend on the population history and the genomic region, which is beyond the scope of this report. We are currently working on implementation of the hierarchical LD model in the context of equivalence testing for reconstruction of independent haplotype blocks.

For a case-control sample, population substructure and cryptic relatedness among subjects leads to over-dispersion of the chi-square test statistic for association and causes spurious rejections of the null hypothesis. The data set we are using is known to be fairly homogeneous [Burton et al., 2007] and we do not expect population stratification artifacts. As presented, our method does not allow incorporation of additional covariates and can only handle a binary trait in the present form. One way to deal with covariates at the moment is to perform stratified analyses in a Mantel-Haenszel framework.

To avoid diminished power from the large number of haplotype configurations [Schaid et al., 2002] proposed to either pool rare haplotypes into a single baseline group or to scan a large chromosomal region for sub-segments that may be associated with the trait, starting with single-locus associations, followed by 'sliding' tests for two-locus haplotypes, followed by 'sliding' tests for three-locus haplotypes, and so forth. We saw from our simulation study that, as the number of haplotype configuration increases, pooling rare haplotypes does not avoid the diminished power problem. In addition, analyses involving a series of adjacent markers assume that the most informative markers are the physically closest. However, this is not always the case and tests based on such associations will not always be optimal. Consider for example the case when relatively recent mutations have introduced correlation among two SNPs in a low LD region, with for example 5 SNPs separating them. In order to include the pairwise correlation of the two SNPs of interest, we would have to use a sliding window of size 7 and perform a test with $2^7 - 1 = 127$ d.f.. Given the large number of haplotype configurations, most haplotype frequencies will be very low and pooling most haplotypes would be unavoidable. On the other hand, one could repeat the same procedure, using again a sliding window of 7, but testing only for allele frequencies and pairwise LD parameters. In this case, one would need to perform a test with $7 + \binom{7}{2} = 28$ d.f.. In this study, we followed a similar, heuristic strategy that lead to the identification of novel associations.

In a given population, the mutations that are causal in disease etiology will have arisen on one or more ancestral haplotypes [Degli-Esposti et al., 1992] and thereafter will have spread to other haplotypes by recombination. Early on in this process, very-high-order association will exist, and the most powerful test for association will be a very-high-order association test, since the strength of the high-order effect more than outweighs the large number of d.f. However, this advantage will not survive in perpetuity, since, as shown in Clayton and Jones [1999] high-order effects will be rapidly diluted by recombination, at

progressively more rapid rates than first order association between a single marker or a pair of markers and disease. As a result, tests based on lower order effects will in general be more powerful than the full haplotype tests. This result is also supported by our simulation study, since in the scenarios we considered, *Bottom Up* tests are the most powerful accross all different disease models. Our proposed method allows to flexibly accomodate both higher- and lower-order LD scenarios.

Appendix

A.1. Proof of lemma 1

Consider again the case of genotype measurements of M bi-allelic loci. Let $h \in H$ be a haplotype at these loci, with $H = \{0,1\}^M$ enumerating the 2^M possible haplotypes. Assume that $h \sim Mult(1, \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\theta_h)_{h \in H}$ the parameter vector of the haplotype frequencies of each haplotype, $\boldsymbol{\theta} \in \Theta$ and $\Theta = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \in (0,1)^{2^M}, \sum_{h \in H} \theta_h = 1\}$. Let $A = \{A_1, A_2, \ldots, A_M\}$ a set of M random variables with $A_j \in \{0,1\}$ the indicator random variable for either one of the two alleles at locus $j, j = 1, \ldots, M$. Let $S = 2^A \setminus \emptyset$ be the power set of A, in lexicographical order, without the empty set, that is, the set of all singletons, pairs, triplets, etc of allele indicator random variables.

In order to prove that $\boldsymbol{\delta}$ is a reparametrization of $\boldsymbol{\theta}$, we introduce an intermediate parameterization, denoted as $\boldsymbol{\eta}$. Then the proof goes as follows. First, we show that $\boldsymbol{\eta}$ is a reparametrization of $\boldsymbol{\theta}$. To prove this, we introduce the function $f(\boldsymbol{\theta}, S)$ and prove that f is bijective. Then, we show that $\boldsymbol{\delta}$ is a reparametrization of $\boldsymbol{\eta}$, which implies that $\boldsymbol{\delta}$ is also a reparametrization of $\boldsymbol{\theta}$. Similarly, to prove this, we introduce the function $q(\boldsymbol{\eta}, S)$ and prove that q is bijective.

. *Mapping function g.* *For a set of random variables $s \in S$, let $\tau_s = \{v \in \{0,1\}^M \mid A_j \in s \Leftrightarrow v[j] = 1\}$ a tuple of all haplotypes whose j-th element is 1 if $A_j \in s$. We define g to be a function which takes as an input the parameter vector $\boldsymbol{\theta} = (\theta_h)_{h \in H}$ and outputs the joint expectation of random variables in s, which we denote with η_s. That is,*

$$g(\boldsymbol{\theta}, s) = E\left(\prod_{A \in s} A\right) = \sum_{h \in \tau_s} \theta_h = \eta_s.$$

We illustrate g with the following example. Let $M = 3$ and $s = \{A_1, A_2\}$. Then τ_s will contain two haplotypes, i.e. $\tau_{\{A_1,A_2\}} = \{(1,1,0),(1,1,1)\}$, and $g(\boldsymbol{\theta}, \{A_1, A_2\}) = E(A_1 A_2) = \theta_{(1,1,0)} + \theta_{(1,1,1)}$. We are thus computing the joint expectation of A_1 and A_2 or the haplotype frequency for loci 1 and 2 with allele 1 chosen at each locus.

For a haplotype $(1,1,1)$, we will write θ_{111} as a shorthand of $\theta_{(1,1,1)}$. Similarly, for a set $\{A_1, A_2, A_3\}$ of random variables, we will write η_{123} as a shorthand of $\eta_{\{A_1,A_2,A_3\}}$.

. *Mapping function f.* *We define f to be a function which takes as input the parameter vector $\boldsymbol{\theta} = (\theta_h)_{h \in H}$ and outputs the joint expectation of random variables in s for all $s \in S$. That is,*

$$f(\boldsymbol{\theta}, S) = \{g(\boldsymbol{\theta}_{\tau_s}, s)\}_{s \in S} = \left(\sum_{h \in \tau_s} \theta_h\right)_{s \in S} = (\eta_s)_{s \in S}.$$

We illustrate f with the following example. For $M = 3$ markers,

$$S = \{\{A_1\}, \{A_2\}, \{A_3\}, \{A_1, A_2\}, \{A_1, A_3\}, \{A_2, A_3\}, \{A_1, A_2, A_3\}\}.$$

Moreover

$$\begin{aligned}
\tau_{A_1} &= \{(1,1,1),(1,1,0),(1,0,1),(1,0,0)\},\\
\tau_{A_2} &= \{(1,1,1),(1,1,0),(0,1,1),(0,1,0)\},\\
\tau_{A_3} &= \{(1,1,1),(0,1,1),(1,0,1),(0,0,1)\},\\
\tau_{\{A_1,A_2\}} &= \{(1,1,1),(1,1,0)\},\\
\tau_{\{A_1,A_3\}} &= \{(1,1,1),(1,0,1)\},\\
\tau_{\{A_2,A_3\}} &= \{(1,1,1),(0,1,1)\}, \text{ and}\\
\tau_{\{A_1,A_2,A_3\}} &= (1,1,1).
\end{aligned}$$

Hence

$$f(\boldsymbol{\theta},S) = \begin{pmatrix} g(\boldsymbol{\theta}_{\tau_{A_1}},\{A_1\})\\ g(\boldsymbol{\theta}_{\tau_{A_2}},\{A_2\})\\ g(\boldsymbol{\theta}_{\tau_{A_3}},\{A_3\})\\ g(\boldsymbol{\theta}_{\tau_{\{A_1,A_2\}}},\{A_1,A_2\})\\ g(\boldsymbol{\theta}_{\tau_{\{A_1,A_3\}}},\{A_1,A_3\})\\ g(\boldsymbol{\theta}_{\tau_{\{A_2,A_3\}}},\{A_2,A_3\})\\ g(\boldsymbol{\theta}_{\tau_{\{A_1,A_2,A_3\}}},\{A_1,A_2,A_3\}) \end{pmatrix} = \begin{pmatrix} \theta_{111}+\theta_{110}+\theta_{101}+\theta_{100}\\ \theta_{111}+\theta_{110}+\theta_{011}+\theta_{010}\\ \theta_{111}+\theta_{011}+\theta_{101}+\theta_{001}\\ \theta_{111}+\theta_{110}\\ \theta_{111}+\theta_{101}\\ \theta_{111}+\theta_{011}\\ \theta_{111} \end{pmatrix} = \begin{pmatrix} \eta_1\\ \eta_2\\ \eta_3\\ \eta_{12}\\ \eta_{13}\\ \eta_{23}\\ \eta_{123} \end{pmatrix}$$

We are thus computing the frequency of the 'marginal' haplotypes over sets of singletons, pairs and triplets of markers.

Lemma 2. *: **Reparametrization** $\boldsymbol{\eta}$. Let $\boldsymbol{\eta} = (\eta_s)_{s\in S} = f(\boldsymbol{\theta},S) = \{g(\boldsymbol{\theta}_{\tau_s},s)\}_{s\in S}$, with $\boldsymbol{\eta}\in\Lambda$, and $\Lambda = \{\boldsymbol{\eta} \mid \boldsymbol{\eta} = f(\boldsymbol{\theta},S), \boldsymbol{\theta}\in\Theta\}$. Then, $\boldsymbol{\eta}$ is a re-parametrization of $\boldsymbol{\theta}$. That is, $f:\Theta\to\Lambda$ is bijective.*

Notice here that we limit $\boldsymbol{\eta}$ to take values in the image of function f. This guarantees that when a bijective function is used to map $\boldsymbol{\eta}$ back to $\boldsymbol{\theta}$'s, those haplotype frequencies will be properly defined, i.e. $\boldsymbol{\theta}\in(0,1)^M$ and $\sum_{h\in H}\theta_h$=1. Before we proceed to prove Lemma 2, we introduce the inverse functions of g and f.

. *Mapping function g^{-1}.* *Let $H^* = \left\{v\in\{0,1\}^M \mid \langle v,v\rangle \neq 0\right\}$ the set of all possible haplotypes over M loci except the haplotype containing only '0' alleles. For a haplotype $h\in H^*$, let $s_h = \{s\in S \mid h[j]=1 \Leftrightarrow A_j\in s\}$ and $\tau_h = \{s\in S \mid s_h\subseteq s\}$. We define g^{-1} to be a function which takes as an input the parameter vector $\boldsymbol{\eta} = \{\eta_s\}_{s\in S}$ and outputs θ_h, the frequency of haplotype h, for $h\in H^*$. That is,*

$$g^{-1}(\boldsymbol{\eta},h) = \sum_{s\in\tau_h}(-1)^{|s_h|+|s|}\eta_s = \theta_h.$$

We illustrate g^{-1} with the following example. Let $M=3$ markers and $h=(1,0,0)$. Then $s_h=\{A_1\}$ and $\tau_h = \{\{A_1,A_2\},\{A_1,A_3\},\{A_1,A_2,A_3\}\}$. Thus

$$\begin{aligned}
\theta_{100} &= (-1)^1\{(-1)^1\eta_1 + (-1)^2\eta_{12} + (-1)^2\eta_{13} + (-1)^3\eta_{123}\}\\
&= \eta_1 - \eta_{12} - \eta_{13} + \eta_{123} = \eta_1 - (\eta_{12}-\eta_{123}) - (\eta_{13}-\eta_{123}) - \eta_{123}\\
&= \eta_1 - \theta_{110} - \theta_{101} - \theta_{111}.
\end{aligned}$$

. *Mapping function $\boldsymbol{f}^{-1}$.* *We define f^{-1} to be a function which takes as input the parameter vector $\boldsymbol{\eta}$ and outputs the haplotype frequencies $\boldsymbol{\theta}$. That is,*

$$f^{-1}(\boldsymbol{\eta},H) = \left[\left\{g^{-1}(\boldsymbol{\eta}_{s_h},h)\right\}_{h\in H^*}, 1-\sum_{h\in H^*} g^{-1}(\boldsymbol{\eta}_{s_h},h)\right].$$

Notice here that the frequency of the haplotype which contains the '0' allele at all the markers is cumputed as one minus the sum of the frequencies of all other haplotypes, i.e. all $h \in H^*$. This guarantees that $\sum_{h \in H} \theta_h = 1$.

We illustrate f^{-1} with the following example. For $M = 3$ markers

$$H^* = \{(1,0,0),(0,1,0),(0,0,1),(1,1,0),(1,0,1),(0,1,1),(1,1,1)\},$$

and

$$\begin{aligned}
&s_{(1,0,0)} = \{A_1\} && \text{and} && \tau_{(1,0,0)} = \{\{A_1\},\{A_1,A_2\},\{A_1,A_3\},\{A_1,A_2,A_3\}\}.\\
&s_{(0,1,0)} = \{A_2\} && \text{and} && \tau_{(0,1,0)} = \{\{A_2\},\{A_1,A_2\},\{A_2,A_3\},\{A_1,A_2,A_3\}\},\\
&s_{(0,0,1)} = \{A_3\} && \text{and} && \tau_{(0,0,1)} = \{\{A_3\},\{A_1,A_3\},\{A_2,A_3\},\{A_1,A_2,A_3\}\},\\
&s_{(1,1,0)} = \{A_1,A_2\} && \text{and} && \tau_{(1,1,0)} = \{\{A_1,A_2\},\{A_1,A_2,A_3\}\},\\
&s_{(1,0,1)} = \{A_1,A_3\} && \text{and} && \tau_{(1,0,1)} = \{\{A_1,A_3\},\{A_1,A_2,A_3\}\},\\
&s_{(0,1,1)} = \{A_2,A_3\} && \text{and} && \tau_{(0,1,1)} = \{\{A_2,A_3\},\{A_1,A_2,A_3\}\},\\
&s_{(1,1,1)} = \{A_1,A_2,A_3\} && \text{and} && \tau_{(1,1,1)} = \{A_1,A_2,A_3\}.
\end{aligned}$$

Hence

$$f^{-1}(\boldsymbol{\eta}, H) = \begin{pmatrix} \theta_{111}\\ \theta_{011}\\ \theta_{101}\\ \theta_{110}\\ \theta_{001}\\ \theta_{010}\\ \theta_{100}\\ \theta_{000} \end{pmatrix} = \begin{pmatrix} \eta_{123}\\ \eta_{23}-\eta_{123}\\ \eta_{13}-\eta_{123}\\ \eta_{12}-\eta_{123}\\ \eta_{3}-\eta_{13}-\eta_{23}+\eta_{123}\\ \eta_{2}-\eta_{12}-\eta_{23}+\eta_{123}\\ \eta_{1}-\eta_{12}-\eta_{13}+\eta_{123}\\ 1-\left\{\sum_{h\in H^*} g^{-1}(\boldsymbol{\eta}_{s_h}, h)\right\} \end{pmatrix}.$$

We now proceed with the proof of Lemma 2. To prove that f is bijective we need to show that f is both injective, i.e. $\forall \boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta, f(\boldsymbol{\theta}, H) = f(\boldsymbol{\theta}^*, H) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^*$, and surjective, i.e. $\forall \boldsymbol{\eta} \in \Lambda, \quad \exists \boldsymbol{\theta} \in \Theta : f(\boldsymbol{\theta}, H) = \boldsymbol{\eta}$. Now that we have defined the inverse function of f, it is easy to show that,

$$f(\boldsymbol{\theta}, H) = f(\boldsymbol{\theta}^*, H) \Rightarrow f^{-1}\{f(\boldsymbol{\theta}, H)\} = f^{-1}\{f(\boldsymbol{\theta}^*, H)\} \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}^*$$

and $\forall$ arbitrary parameter vectors $\boldsymbol{\eta} \in \Lambda$ we can choose $\boldsymbol{\theta} = f^{-1}(\boldsymbol{\eta}, H)$ such that $f(\boldsymbol{\theta}, S) = f\left\{f^{-1}(\boldsymbol{\eta}, H), S\right\} = \boldsymbol{\eta}$. This concludes the proof of the bijectivness of f, which concludes also the proof of Lemma 2.

We now proceed to prove that $\boldsymbol{\delta}$ is a reparametrization of $\boldsymbol{\eta}$ and hence a reparametrization of $\boldsymbol{\theta}$. First, we introduce functions $c(\boldsymbol{\eta}, s)$ and $q(\boldsymbol{\eta}, S)$.

. *Mapping function $\boldsymbol{\kappa}$.* *Let P_s refer to the family of sets of all possible partitions of a set of random variables $s, s \in S$, into nonempty subsets (blocks). So, for $p \in P_s$, each $b \in p$ is a block. Moreover, let $\tau'_s = 2^s \setminus \emptyset$ the power set of s minus the empty set. We define κ to be a function which takes as an input the parameter vector $\boldsymbol{\eta} = (\eta_s)_{s \in S}$ and outputs the joint cumulant of the set of random variables in s, which we denote by δ_s. That is,*

$$\kappa(\boldsymbol{\eta}, s) = \sum_{p \in P_s} (-1)^{|p|-1}(|p|-1)! \prod_{b \in p} E\left(\prod_{A \in b} A\right) = \sum_{p \in P_s} (-1)^{|p|-1}(|p|-1)! \prod_{b \in p} \eta_b = \delta_s.$$

We illustrate function c with the following example. Let $M = 3$ and $s = \{A_1, A_2, A_3\}$. Then,

$$P_s = \Big\{ \{A_1, A_2, A_3\}, \{\{A_1, A_2\}, \{A_3\}\}, \{\{A_1, A_3\}, \{A_2\}\}, \{\{A_2, A_3\}, \{A_1\}\}, \{\{A_1\}, \{A_2\}, \{A_3\}\}\}$$

and $\boldsymbol{\eta}_{\tau'_s} = (\eta_1, \eta_2, \eta_3, \eta_{12}, \eta_{13}, \eta_{23}, \eta_{123})$.

Thus,

$$\begin{aligned}\kappa\left(\boldsymbol{\eta}_{\tau'_s}, s\right) &= (-1)^{1-1}(1-1)!\ \eta_{\{A_1,A_2,A_3\}} + (-1)^{2-1}(2-1)!\ \eta_{\{A_1,A_2\}}\eta_{\{A_3\}} \\ &+ (-1)^{2-1}(2-1)!\ \eta_{\{A_1,A_3\}}\eta_{\{A_2\}} + (-1)^{2-1}(2-1)!\ \eta_{\{A_2,A_3\}}\eta_{\{A_1\}} \\ &+ (-1)^{3-1}(3-1)!\ \eta_{\{A_1\}}\eta_{\{A_2\}}\eta_{\{A_3\}} \\ &= \eta_{123} + \eta_{12}\eta_3 + \eta_{13}\eta_2 + \eta_{23}\eta_1 + 2\,\eta_1\eta_2\eta_3\end{aligned}$$

. *Mapping function q.* *We define q to be a function which takes as input the parameter vector $\boldsymbol{\eta} = (\eta_s)_{s\in S}$ and outputs the joint cumulant of random variables in s for all $s \in S$. That is,*

$$q(\boldsymbol{\eta}, S) = \left\{\kappa\left(\boldsymbol{\eta}_{\tau'_s}, s\right)\right\}_{s\in S} = \{\delta_s\}_{s\in S}.$$

We illustrate q with the following example. For $M = 3$,

$$\begin{aligned}P_{\{A_1\}} &= \{A_1\}, \\ P_{\{A_2\}} &= \{A_2\}, \\ P_{\{A_3\}} &= \{A_3\}, \\ P_{\{A_1,A_2\}} &= \{\{A_1, A_2\}, \{\{A_1\}, \{A_2\}\}, \\ P_{\{A_1,A_3\}} &= \{\{A_1, A_3\}, \{\{A_1\}, \{A_3\}\}, \\ P_{\{A_2,A_3\}} &= \{\{A_2, A_3\}, \{\{A_2\}, \{A_3\}\}, \text{ and} \\ P_{\{A_1,A_2,A_3\}} &= \{\{A_1, A_2, A_3\}, \{\{A_1, A_2\}, \{A_3\}\}, \{\{A_1, A_3\}, \{A_2\}\}, \{\{A_2, A_3\}, \{A_1\}\}, \\ &\quad \{\{A_1\}, \{A_2\}, \{A_3\}\}\}.\end{aligned}$$

Hence $q(\boldsymbol{\eta}, S) = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_{12} - \eta_1\eta_2 \\ \eta_{13} - \eta_1\eta_3 \\ \eta_{23} - \eta_2\eta_3 \\ \eta_{123} - \eta_{12}\eta_3 - \eta_{13}\eta_2 - \eta_{23}\eta_1 + 2\eta_1\eta_2\eta_3 \end{pmatrix}$.

We are thus computing the joint cumulant of all possible sets of singletons, pairs and triplets of markers.

Lemma 3. ***Reparametrization $\boldsymbol{\delta}$*** *Let $\boldsymbol{\delta} = (\delta_s)_{s\in S} = q(\boldsymbol{\eta}, S) = \left(\kappa\left(\boldsymbol{\eta}_{\tau'_s}, s\right)\right)_{s\in S'}$ with $\boldsymbol{\delta} \in \Delta$ and $\Delta = \{\boldsymbol{\delta} \mid \boldsymbol{\delta} = q \circ f(\boldsymbol{\theta}, S), \boldsymbol{\theta} \in \Theta\}$. Then $\boldsymbol{\delta}$ is a re-parametrization of $\boldsymbol{\eta}$. That is, $q : \Lambda \to \Delta$ is a bijective mapping function.*

Notice here that we limit $\boldsymbol{\delta}$ to take values in the image of function q. This guarantees that when a bijective function is used to map $\boldsymbol{\delta}$ back to $\boldsymbol{\eta}$ and then back to $\boldsymbol{\theta}$'s, those haplotype frequencies will be properly defined. Before we proceed to prove Lemma 3, we introduce the inverse functions of c and q.

. Mapping function κ^{-1} *. We define κ^{-1} to be a function which takes as an input the parameter vector $\boldsymbol{\delta} = (\delta_s)_{s\in S}$ and outputs η_s, the joint expectation of the set of random variables in s. That is,*

$$\kappa^{-1}(\boldsymbol{\delta}, s) = \delta_s - \sum_{p\in P_s\setminus s} (-1)^{|p|}(|p|-1)! \prod_{b\in p}\left\{\delta_b - \sum_{p'\in P_b\setminus b}(-1)^{|p'|}(|p'|-1)! \prod_{b'\in p'}\delta_{b'}\right\}.$$

We illustrate function κ^{-1} with the following example. Let $M=3$ and $s=\{A_1,A_2,A_3\}$. Then, $\boldsymbol{\delta}_{\tau_s'} = (\delta_1,\delta_2,\delta_3,\delta_{12},\delta_{13},\delta_{23},\delta_{123})$. Hence,

$$\begin{aligned}
\kappa^{-1}\left(\boldsymbol{\delta}_{\tau_s'}, s\right) &= \delta_{\{A_1,A_2,A_3\}} + (-1)^2(2-1)!\ \left(\delta_{\{A_1,A_2\}} + (-1)^2(2-1)!\,\delta_{A_1}\delta_{A_2}\right)\delta_{A_3} \\
&\quad + (-1)^2(2-1)!\ \left(\delta_{\{A_1,A_3\}} + (-1)^2(2-1)!\,\delta_{A_1}\delta_{A_3}\right)\delta_{A_2} \\
&\quad + (-1)^2(2-1)!\ \left(\delta_{\{A_2,A_3\}} + (-1)^2(2-1)!\,\delta_{A_2}\delta_{A_3}\right)\delta_{A_1} \\
&\quad + (-1)^3(3-1)!\,\delta_{A_1}\delta_{A_2}\delta_{A_3} \\
&= \delta_{\{A_1,A_2,A_3\}} + \left(\delta_{\{A_1,A_2\}} + \delta_{A_1}\delta_{A_2}\right)\delta_{A_3} + \left(\delta_{\{A_1,A_3\}} + \delta_{A_1}\delta_{A_3}\right)\delta_{A_2} \\
&\quad + \left(\delta_{\{A_2,A_3\}} + \delta_{A_2}\delta_{A_3}\right)\delta_{A_1} - 2\delta_{A_1}\delta_{A_2}\delta_{A_3}
\end{aligned}$$

Which is the same expression we would get if we used the definition of $\delta_{\{A_1,A_2,A_3\}}$, and solved for η_{123}, that is

$$\begin{aligned}
\delta_{\{A_1,A_2,A_3\}} &= \eta_{234} - \eta_{12}\eta_3 - \eta_{13}\eta_2 - \eta_{23}\eta_1 + 2\eta_1\eta_2\eta_3 \\
\Rightarrow \eta_{123} &= \delta_{\{A_1,A_2,A_3\}} + \eta_{12}\eta_3 + \eta_{13}\eta_2 + \eta_{23}\eta_1 - 2\eta_1\eta_2\eta_3 \\
&= \delta_{\{A_1,A_2,A_3\}} + \left(\delta_{\{A_1,A_2\}} + \delta_{A_1}\delta_{A_2}\right)\delta_{A_3} + \left(\delta_{\{A_1,A_3\}} + \delta_{A_1}\delta_{A_3}\right)\delta_{A_2} \\
&\quad + \left(\delta_{\{A_2,A_3\}} + \delta_{A_2}\delta_{A_3}\right)\delta_{A_1} - 2\delta_{A_1}\delta_{A_2}\delta_{A_3}
\end{aligned}$$

For a set $\{A_1,A_2,A_3\}$ of random variables, we will write δ_{123} as a shorthand of $\delta_{\{A_1,A_2,A_3\}}$.

. Mapping function q^{-1}. *We define q^{-1} to be a function which takes as input the parameter vector $\boldsymbol{\delta} = (\delta_s)_{s\in S}$ and outputs $\boldsymbol{\eta} = (\eta_s)_{s\in S}$, the joint expectation of the set of random variables in s for all $s\in S$. That is,*

$$q^{-1}(\boldsymbol{\delta}, S) = \left\{\kappa^{-1}\left(\boldsymbol{\delta}_{\tau_s'}, s\right)\right\}_{s\in S} = (\eta_s)_{s\in S}.$$

We illustrate q^{-1} with the following example. For $M=3$,

$$q^{-1}(\boldsymbol{\delta}, S) = \begin{pmatrix}
\delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_{12}+\delta_1\delta_2 \\ \delta_{13}+\delta_1\delta_3 \\ \delta_{23}+\delta_2\delta_3 \\ \delta_{123}+\delta_{12}\delta_3+\delta_{13}\delta_2+\delta_{23}\delta_1-2\delta_1\delta_2\delta_3
\end{pmatrix} = \begin{pmatrix}
\eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_{12} \\ \eta_{13} \\ \eta_{23} \\ \eta_{123}
\end{pmatrix}.$$

We now proceed with the proof of Lemma 3. To prove that q is bijective we need to show that q is both injective, i.e. $\forall\boldsymbol{\eta},\boldsymbol{\eta}^*\in\Lambda, q(\boldsymbol{\eta},S)=q(\boldsymbol{\eta}^*,S)\Rightarrow\boldsymbol{\eta}=\boldsymbol{\eta}^*$, and surjective, i.e. $\forall\boldsymbol{\delta}\in\Delta,\quad \exists\boldsymbol{\eta}\in\Lambda : q(\boldsymbol{\eta},S)=\boldsymbol{\delta}$. Now that we have defined the inverse function of q, it is easy to show that, $q(\boldsymbol{\eta},S)=q(\boldsymbol{\eta}^*,S)\Rightarrow q^{-1}\{q(\boldsymbol{\eta},S)\}=q^{-1}\{q(\boldsymbol{\eta}^*,S)\}\Rightarrow\boldsymbol{\eta}=\boldsymbol{\eta}^*$ and for all arbitrary parameter vectors $\boldsymbol{\delta}\in\Delta$, we can choose $\boldsymbol{\eta}=q^{-1}(\boldsymbol{\delta},S)$ such that $q(\boldsymbol{\eta},S)=q\left\{q^{-1}(\boldsymbol{\delta},S),S\right\}=\boldsymbol{\delta}$. This concludes the proof of the bijectivness of q, which concludes also the proof of Lemma 3 and thus of Lemma 1.

A.2. Standardized parameters

Recall that δ_A can be expressed as

$$\delta_A = \eta_A - \sum_{p \in P_A \setminus A} (-1)^{|p|}(|p|-1)! \prod_{b \in p} \eta_b = \eta_A - \sum_{p \in P_A \setminus A} R_\delta(p) = \eta_A - R_\delta$$

where $R_\delta(p)$'s depend on loci $b \in p$ with $|b| < M$. These rest terms $R_\delta(p)$ are considered fixed and bounds for δ_A are to be determined completely analogous to the two locus case based on η_A.

First, η_A is upper bound by all lower-order η_s and lower bound by 0. That is

$$\eta_A \leq U_1(A) := \min\{\eta_s | s \in S \setminus A\}.$$
$$\eta_A \geq L_1(A) = 0$$

Second, further constraints are imposed by the relationship between η_A and lower order haplotype frequencies η_s. It is straightforward to see that η_s can be restated as:

$$\eta_s = g(\theta, s) = \theta_s + \sum_{t \in S, t \supset s} (-1)^{|t|-|s|-1} \eta_t$$

Here, θ_{h_s} is the frequency for haplotype $h_s = \left\{v \in \{0,1\}^M \mid v[j] \Leftrightarrow A_j \in s\right\}$, the haplotype with M loci with 1-alleles at loci s and 0-alleles elsewhere. Note, that all the sums above include η_A Solving for η_A gives us:

$$\begin{aligned}
\eta_A &= (-1)^{|A|-|s|-1} \left\{ \eta_s - \theta_{h_s} - \sum_{t \in S \setminus A, t \supset s} (-1)^{|t|-|s|-1} \eta_t \right\} \\
&= (-1)^{|A|-|s|} \left\{ \theta_{h_s} - \eta_s + \sum_{t \in S \setminus A, t \supset s} (-1)^{|t|-|s|-1} \eta_t \right\} \\
&= (-1)^{|A|-|s|} \left\{ \theta_{h_s} - \eta_s - \sum_{t \in S \setminus A, t \supset s} (-1)^{|t|-|s|} \eta_t \right\} \\
&= (-1)^{|A|-|s|} \left[\theta_{h_s} - \left\{ \eta_s + \sum_{t \in S \setminus A, t \supset s} (-1)^{|t|-|s|} \eta_t \right\} \right] \\
&= (-1)^{|A|-|s|} \left\{ \theta_{h_s} - \sum_{t \in S \setminus A, t \supseteq s} (-1)^{|t|-|s|} \eta_t \right\} \\
&= \sigma_s \left(\theta_{h_s} - R_s \right),
\end{aligned}$$

where $\sigma = (-1)^{|A|-|s|}$ and $R_s = \sum_{t \in S \setminus A, t \supseteq s} (-1)^{|t|-|s|} \eta_t$. Each η_s therefore contributes an upper and lower bound to η_A by choosing $\theta_{h_s} = 0$ or $\theta_{h_s} = 1$:

$$\eta_A \leq U_s := \begin{cases} \max(1 - R_s, 0) & \text{if } \sigma \geq 0, \\ \min(-R_s, 0) & \text{if } \sigma < 0 \end{cases},$$

$$\eta_A \geq L_s := \begin{cases} \max(-R_s, 0) & \text{if } \sigma \geq 0, \\ \min(1 - R_s, 0) & \text{if } \sigma < 0 \end{cases}.$$

With $U_2(A) := \min\{U_s | s \in S \setminus \{A\}\}$ and $L_2(A) := \max\{L_s | s \in S \setminus \{A\}\}$, we get

$$\eta_A^{\max} := \min\{U_1(A), U_2(A)\},$$
$$\eta_A^{\min} := \max\{L_1(A), L_2(A)\}.$$

Then $\eta_A^{\max}$ and $\eta_A^{\min}$ can be used as above to standardize δ_A.

A.4. Additional Tables

Table A.1: Linkage disequilibrium parameters in the cases (Ca), controls (Co) and pool (P) of cases and controls samples for each of the four triplets identified from the WTCCC data analysis.

	Triplet 1			Triplet 2		
	P	Ca	Co	P	Ca	Co
δ_1	.239	.264	.223	.239	.264	.223
δ_2	.126	.120	.130	.091	.108	.081
δ_3	.091	.108	.081	.314	.308	.319
δ_{12}	-.374	-.502	-.279	.111	.135	.081
δ_{13}	.111	.135	.081	-.112	-.199	-.046
δ_{23}	-.544	-.382	-.663	.363	.351	.377
δ_{123}	.172	.084	.229	-.697	-.665	-.716

	Triplet 3			Triplet 4		
	P	Ca	Co	P	Ca	Co
δ_1	.350	.324	.366	.213	.234	.199
δ_2	.207	.191	.218	.401	.376	.417
δ_3	.247	.276	.229	.199	.209	.193
δ_{12}	.712	.696	.720	-.297	-.268	-.306
δ_{13}	.103	.033	.170	-.759	-.790	-.739
δ_{23}	-.105	-.174	-.040	.227	.088	.335
δ_{123}	-.160	-.119	-.190	.203	.521	-.126

Table A.2: Corresponding haplotype frequencies for SNPs in linkage equilibrium (Scenario 1), and SNPs in LD (Scenario 2).

Haplotype	Scenario 1	Scenario 2
θ_{0000}	.292	.358
θ_{0001}	.239	.088
θ_{0010}	.135	.240
θ_{0011}	.111	.101
θ_{0100}	.065	.148
θ_{0101}	.054	.004
θ_{0110}	.030	.053
θ_{0111}	.025	.007
θ_{1000}	.015	.358
θ_{1001}	.013	.088
θ_{1010}	.007	.240
θ_{1011}	.006	.101
θ_{1100}	.003	.148
θ_{1101}	.003	.004
θ_{1110}	.002	.053
θ_{1111}	.001	.007