Published online 9 June 2017

Nucleic Acids Research, 2017, Vol. 45, No. 14 ***e131***
doi: 10.1093/nar/gkx505

Prioritizing tests of epistasis through hierarchical representation of genomic redundancies

Tyler Cowman[1] and Mehmet Koyutürk[1,2,*]

[1]Electrical Engineering & Computer Science, Case Western Reserve University, Cleveland, OH 44106, USA and
[2]Center for Proteomics & Bioinformatics, Case Western Reserve University, Cleveland, OH 44106, USA

Received August 09, 2016; Revised May 22, 2017; Editorial Decision May 24, 2017; Accepted May 29, 2017

ABSTRACT

Epistasis is defined as a statistical interaction between two or more genomic loci in terms of their association with a phenotype of interest. Epistatic loci that are identified using data from Genome-Wide Association Studies (GWAS) provide insights into the interplay among multiple genetic factors, with applications including assessment of susceptibility to complex diseases, decision making in precision medicine, and gaining insights into disease mechanisms. Since the number of genomic loci assayed by GWAS is extremely large (usually in the order of millions), identification of epistatic loci is a statistically difficult and computationally intensive problem. Even when only pairwise interactions are considered, the size of the search space ranges from hundreds of millions to billions of locus pairs. The large number of statistical tests performed also makes sufficient type one error correction imperative. Consequently, efficient algorithms are required to filter the tests that are performed and evaluate large GWAS data sets in a reasonable amount of computation time. It has been observed that many pairwise tests are redundant due to correlations in their genotype values across samples, known as linkage disequilibrium. However, algorithms that have been developed for efficient identification of epistatic loci do not systematically exploit linkage disequilibrium. Here, we propose a new algorithm for fast epistasis detection based on hierarchical representation of linkage disequilibrium (LinDen). We utilize redundancies in genotype patterns between neighboring loci to generate a hierarchical structure and execute a branch-and-bound search to prioritize loci testing based on approximations of a test statistic for pairs of locus groups. The hierarchical organization of tests performed by LinDen allows for efficient scaling based on the screened loci. We test LinDen comprehensively on three data sets obtained from the Wellcome Trust Case Control Consortium: type two diabetes, psoriasis, and hypertension. Our results show that, as compared other state-of-the-art tools for fast epistasis detection, LinDen drastically reduces the number of tests performed while discovering statistically significant locus pairs. LinDen is implemented in C++ and is available as open source at http://compbio.case.edu/linden/.

INTRODUCTION

Genome-Wide Association Studies (GWAS) have been celebrated as a comprehensive way of assessing the statistical association between specific genomic variants and a given phenotype within a population. An application of the common-disease common-variant hypothesis (1), GWAS initially focused on single-locus associations (2), and have been successful in identifying significant associations between various genomic variants and many complex diseases (3–5). However, as the limitations of GWAS are now recognized, research has shifted toward multi-locus, epistatic interactions which show more promise in capturing the interplay among multiple variants in their association with complex traits (6).

Epistasis

Epistasis can be viewed as a functional or statistical interaction between two or more genomic variants in the context of a specific phenotype (7). Functional epistasis refers to the notion that one genetic variant can modify the effect of another genetic variant on phenotype (8). Recessive alleles under a Mendelian inheritance model would be an example of this. Statistical epistasis is concerned with finding quantifiable statistically significant relationships between two or more variants in their association with phenotype (9). Ideally, data on statistical epistasis helps inform the inference of the biological underpinnings of functional epistasis. However, it is important to note that the existence of statistical epistasis does not necessarily imply a useful bi-

*To whom correspondence should be addressed. Tel: +1 216 368 2963; Fax: +1 216 368 6888; Email: mehmet.koyuturk@case.edu

ological link. As a data-driven approach, GWAS are usually concerned with detecting statistical epistasis. For this purpose, epistasis is tested using various methods, including multiplicative and genotype-based models (10). In this study, we focus on testing epistasis with genotype-based models, performed through a χ^2 test on the contingency tables associated with pairs of loci.

Computational and statistical challenges

Two of the most significant challenges in detecting epistatic loci both stem from the size of the search space. Exhaustive evaluation of all marginal effects alone can require hundreds of thousands of statistical tests. This number then grows exponentially based on the order of interactions considered, and higher order interactions quickly result in a highly under-determined system. Consequently, most epistatic search algorithms focus only on pairwise interactions. Even so, the sheer number of locus pairs to be tested necessitates significant computational power to calculate all of the test statistics. Furthermore, a large number of independent statistical tests greatly reduces the statistical power of those tests. This makes correction for multiple hypothesis testing rather challenging. For this reason, many methods have been developed to reduce the number of tests necessary for detecting epistasis.

Existing methods

Existing approaches to detecting epistasis can be roughly classified into three categories: Exhaustive methods, filtering-based methods, and heuristic methods. Exhaustive methods examine all pairwise combinations and will not miss any significant interactions, as defined by their chosen measure of significance. These methods achieve improvements in computational performance through use of carefully designed index structures that exploit the patterns in GWAS data (11,12). However, these methods are likely to be significantly slower, and potentially intractable for sufficiently large datasets. Filtering-based methods reduce the search space of possible locus pairings by incorporating prior biological knowledge (13,14). These methods commonly utilize mappings of the loci to known pathways, functional categories, or proteins in protein interaction networks, and limit the search space to pairs that are mapped to the same pathway or interacting proteins. This approach is capable of drastically reducing the time complexity as well as improving the statistical power of the epistasis tests. However, these methods ignore the vast majority of loci pairings available in GWAS data as most loci are neither within coding nor known regulatory regions. A broader class of approaches that include incorporation of prior knowledge comprises heuristic methods in general. These methods aim to strike a balance between the accuracy of exhaustive methods and reduced computational complexity of filtering (15,16) and usually either prescreen loci or prioritize pairs of loci for testing epistasis. (17,18). The tool-set PLINK (19,20) contains what is to our knowledge, the fastest general purpose pairwise epistasis scan currently available, *fast-epistasis*. PLINK achieves this speed through an extremely efficient implementation. Consequently, it still requires a quadratic number of statistical tests to be performed and will not be suited to very high density datasets.

Linkage disequilibrium

Linkage disequilibrium (LD) (21) refers the existence of statistical associations between the genotypes of different loci, and are most commonly observed between loci that are physically close to each other in the genome. This is also known as gametic phase disequilibrium and occurs due to the way in which chromosomes are copied and recombined during meiosis. During prophase, homologous chromosomes exchange segments with one another in a process known as cross over. This results in a statistical linkage between the genotypes of nearby loci as only those on opposing sides of a crossover event will have had their association broken. The closer two loci are, the less likely it is that an event will occur between them as there are less potential locations for one to occur.

LD offers many challenges in the detection of epistatic loci, since loci that are in LD can appear to be statistically interacting due to the patterns induced in the loci that are in disequilibrium (22). However, LD also offers opportunities in the detection of epistasis. In particular, it is common practice to identify groups of loci with strong LD, and use only a single locus from each group for testing epistasis. Since the genotypes of different loci in the same LD group are correlated, it is sufficient to identify the interaction between a representative locus from that group and other loci in the genome (23). This is also useful in integrating data from multiple cohorts, since different studies can genotype different loci, and a missing locus may have an LD partner that is genotyped.

Contributions of this study

In this study, we propose a framework for systematically exploiting LD to reduce the number of tests performed in epistasis detection. Observing that LD is a quantitative concept, and the definition of 'LD groups' relies on arbitrary statistical thresholds, we develop a hierarchical representation of LD groups. In the proposed framework, rather than relying on available LD information derived from the general population, we use the redundancies in the GWA data set that is analyzed. The proposed method, LINDEN, uses correlations between the genotypes of neighboring loci to construct groups that hierarchically represent LD trees and derives representative genotypes for these LD groups. Subsequently, it uses these representative genotypes to score the potential interaction between any pair of loci in the respective groups and filter out pairs of loci groups that are not promising.

We test LINDEN comprehensively on three different GWAS datasets obtained from the Wellcome Trust Case Control Consortium (WTCCC). We observe that LINDEN can substantially reduce the number of tests required without significantly compromising the ability to detect true epistatic pairs. Finally, we demonstrate that the number of tests performed by LINDEN grows sub-quadratically when the growth of the genotyped loci is based on an increase in

density. This result shows that LINDEN can be very useful as GWAS evolve into whole-genome association studies.

METHODS

In this section, we first describe the structure of genome wide association data. We then define epistasis and discuss the challenges involved in identifying epistatic loci. Subsequently, we present an algorithmic framework that exploits redundancies in the genotypes of neighboring loci to efficiently identify epistatic pairs of genomic loci. The workflow of the proposed framework is shown in Figure 1.

Problem description

Genome wide association data. The input to our problem consists of a genome-wide association (GWA) dataset $D = (C, S, g, f)$, in which C refers to the set of loci genotyped for the set of samples S. In this dataset, $g(c, s)$ denotes the genotype of locus $c \in C$ for sample $s \in S$, $f(s)$ the phenotype of sample $s \in S$. We consider instances in which the phenotype can be formulated as a binary trait of interest, e.g., a sample either has type II diabetes or does not. A sample s that has/does not have the trait of interest (i.e., $f(s) = 1/f(s) = 0$) is called a *case*/*control* sample.

Genotype coding. For each locus, the variant that occurs most commonly in the population is called the *major allele*, and the less frequent variant is called the *minor allele*. The genotype for a given locus is determined by the combination of alleles in the two chromosomes. When genotype phasing information is not available, there are three distinct possible genotypes for each loci. Therefore, given the major allele for each locus, a genotype can be coded as a $|C| \times |S|$ matrix G such that $G(c, s)$ denotes the presence or absence of the minor allele in locus c, i.e.:

$$G(c,s) = \begin{cases} 2 & \text{if } g(c,s) \text{ is Homozygous minor} \\ 1 & \text{if } g(c,s) \text{ is Heterozygous} \\ 0 & \text{if } g(c,s) \text{ is Homozygous major} \end{cases} \quad (1)$$

We call each row of this matrix the genotype vector G_c, the *genotype vector* of c.

Epistasis. It has been repeatedly observed that the genotype of a locus may alter the effect of a different locus on an organism's phenotype (8). Such interactions are captured statistically by comparing the association between the phenotype and the combined genotype of a pair of loci with the association between the phenotype and the genotypes of the individual loci. There are multiple models for testing epistasis statistically, including multiplicative and genotype-based models (10). Here, we focus on genotype-based epistasis, i.e., our objective is to identify pairs of loci such that specific combinations of the genotypes of these loci are significantly associated with the phenotype. A common statistical test used for this purpose, and what is used by LINDEN, is the χ^2 test. The χ^2 statistic assesses the strength of the potential association with phenotype based on the distribution of case and control samples in the contingency table that represents the genotype combinations of the two loci.

Exhaustive testing for epistasis. The most straightforward method for identifying epistatic loci is to test all, or selected pairs, of loci to identify locus pairings whose significance of association exceeds a certain threshold (after correction for multiple hypothesis testing). In order to exhaustively test all pairs of loci for epistasis, $\binom{|C|}{2}$ statistical tests must be performed. Since the number of genotyped loci is in the order of hundreds of thousands, this is computationally expensive. Furthermore, testing many pairs greatly reduces statistical power, due to the large number of independent hypotheses tested. To alleviate these problems, we focus on an alternate formulation of the problem and propose a framework for organizing the input loci into *linkage disequilibrium trees*. This enables early identification and filtering of non-significant locus pairs without explicitly testing those pairs for epistasis.

Proposed formulation

We formulate the problem as one of finding the most significant epistatic interaction for each genomic locus.

Definition 1.

(Most Significant Epistatic Partner for a Locus): Let C be the set of loci genotyped in a GWAS. For two genomic loci c_i and $c_j \in C$, let $X^2(c_i, c_j)$ denote the X^2 statistic for the contingency table of the genotypes of c_i and c_j. Then, for each $c_i \in C$, the most significant epistatic partner for c_i is the locus $c_j \in C - \{c_i\}$ such that $X^2(c_i, c_j) > X^2(c_i, c_k)$ for any $c_k \in C - \{c_i, c_j\}$.

The motivation for this approach is that correction for multiple hypotheses renders statistical evaluation highly conservative, making biological interpretations more dependent on arbitrary statistical thresholds. In contrast, ranking pairs of loci based on test scores and focusing on the highest ranked interaction for each locus provides a representative view of the possible biologically relevant interactions of all individual loci that are genotyped. This formulation also enables the use of branch-and-bound algorithms to prune out pairs that are relatively less interesting without performing explicit statistical tests.

Reciprocally significant epistatic pairs. In practice, the list of identified epistatic pairs are dominated by some 'hub' loci. The hubs emerge even when the search is limited to the most significant epistatic partner for each locus, since the marginal effect of a locus shows its interaction with many other loci as the most significant interaction for each of the other loci (24,25). This leads to the identification of many redundant and uninteresting pairs. Motivated by this observation, we propose to limit the search for *reciprocally significant epistatic pairs*, defined as follows.

Definition 2.

(Reciprocally Significant Epistatic Pairs): Two loci c_i and $c_j \in C$ are said to be reciprocally epistatic if c_i is the most significant epistatic partner for c_j and c_j is the most significant epistatic partner for c_i.

Based on this definition, we define the problem of epistasis detection as one of identifying and ranking all reciprocally significant epistatic pairs of loci. This ensures that each locus appears at most once in the final output. Constraining the output to reciprocal pairs reduces the noise from loci

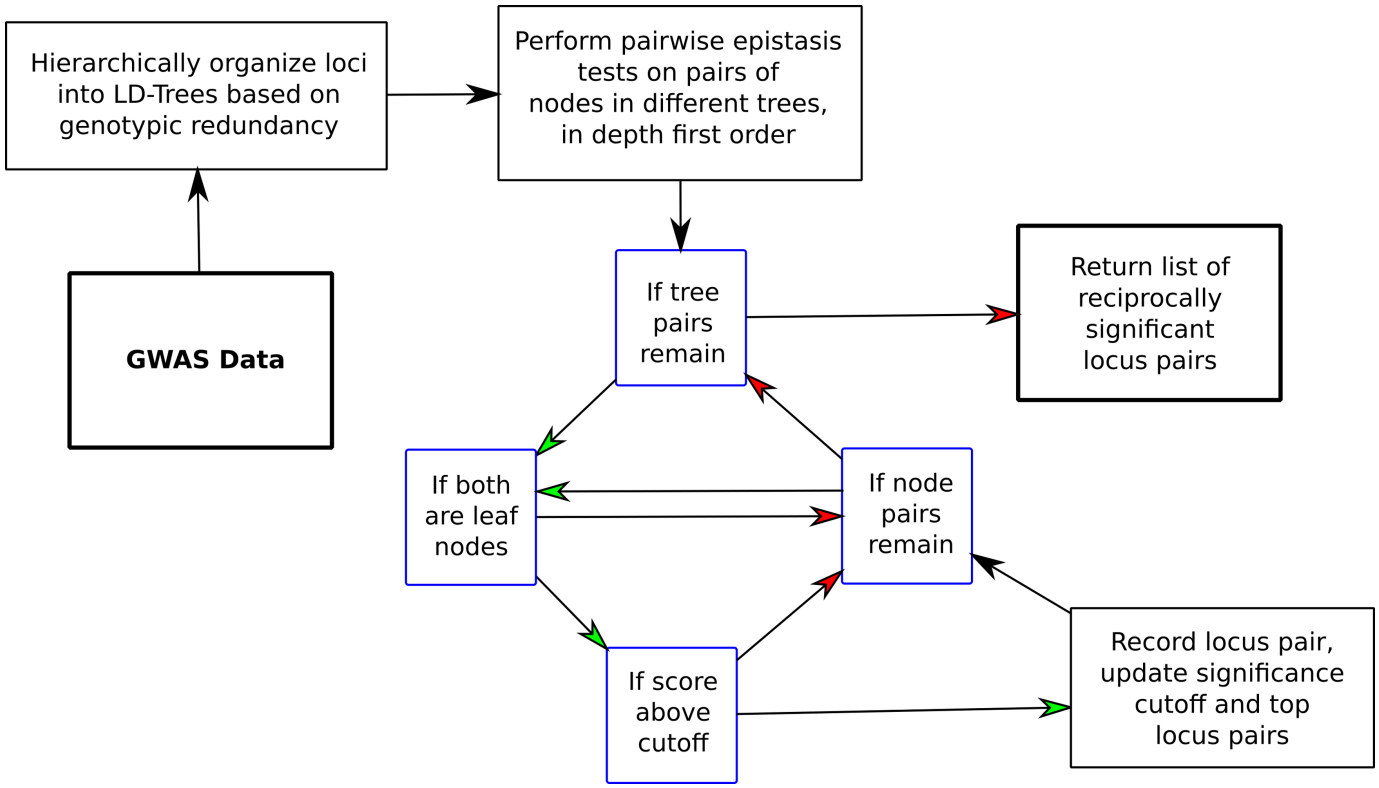

Figure 1. Work-flow of the proposed framework for efficient detection of epistasis. Decision nodes are denoted by blue borders and green and red arrows respectively represent yes and no. The input loci can be filtered to remove those with very low minor allele frequency and those with significant marginal effect. Pairs of loci within a specified number of base pairs are not tested, since linkage disequilibrium may give rise to statistical interactions that are not functionally relevant.

that may have large marginal effects, thus leaving those locus pairs that are most likely to have epistatic interactions with biological significance.

Clearly, *P*-values are tools for assessing whether a finding is interesting from a statistical perspective, but the thresholds used to interpret p-values are rather arbitrary. For this reason, we believe that the qualitative information on being the 'best pair' and 'reciprocity' can be useful for the user who will be interpreting these results. To this end, the labeling of this approach as 'filtering' or 'prioritization' can be helpful but it may also be misleading. To this end, LINDEN can best be described as a heuristic all pairs analysis.

Linkage disequilibrium trees

LINDEN utilizes a tree-based representation of loci to exploit the redundancies in the genotypes of loci that are in linkage disequilibrium (LD). It is possible to reduce the number of tested locus pairs by grouping genomic loci that are in high LD, and using one representative locus from each LD group for testing epistasis. The PLINK clumping tool offers a greedy method for grouping and choosing a representative locus in this manner (20). However, this approach can lead to many false negatives for loci that are not in perfect disequilibrium, since the genotypes of the representative loci may not be sufficient to capture the interactions of other loci in the group. Representation of LD groups using a tree structure better captures the nature of linkage disequilibrium as a continuum, by providing a way to hierarchically represent the degree of LD between different loci. This allows systematic testing of epistasis for multiple loci in a hierarchical manner, thereby reducing the number of statistical tests performed while missing fewer (or no) statistically significant associations.

LD-tree. We define an LD-tree as a full binary tree T, in which each node t represents a set $L(t) \subset C$ of genomic loci and is associated with a representative genotype vector, V_t. Each leaf node represents exactly one individual genomic locus, where the representative genotype vector is the genotype vector of that locus, i.e., $V_t = G_c$ for leaf node t with $L(t) = \{c\}$. For an internal node t, let t_ℓ and t_r denote the children of t. Then, the representative genotype vector V_t is defined as:

$$V_t(s) = \begin{cases} V_{t_\ell}(s), & \text{if } V_{t_\ell}(s) = V_{t_r}(s) \\ \texttt{NIL}, & \text{otherwise} \end{cases} \text{ for all } s \in S. \qquad (2)$$

An example LD-Tree demonstrating the notion of representative genotype vectors is shown in Figure 2.

Construction of an LD-forest

Formulation. The key idea of the proposed algorithm is to generate a set of LD-Trees that enables bounding or approximating the test score for epistasis between pairs of loci based on the comparison of the representative genotype

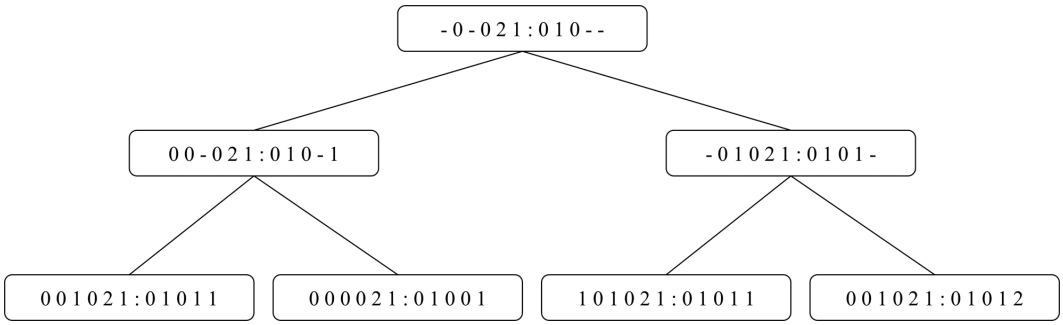

Figure 2. Illustration of the concepts of LD-Tree and representative genotypes. Each representative genotype vector is shown in their respective box. Here a colon denotes the separation between case and control samples, and a—is used to represent the samples that do not overlap . Each leaf node represents a single locus, therefore there are no unknown genotypes at this level. A node that is higher in the tree represents a larger number of loci and therefore has more samples with unknown genotypes in its representative vector.

vectors of the roots (or internal nodes) of these trees. We call such a collection of LD-Trees an LD-Forest and denote it by R.

In this setup, it is clear that the nodes that are closer to the root represent a larger number of loci. However tests involving these nodes are also less informative, since their representative genotype vectors constitute a sparser representation. Recognizing this trade-off, we can formulate LD-Forest construction as an optimization problem. Namely, given genome-wide association data $D = (C, S, g, f)$ and a parameter d, our objective is to compute the minimum number of LD-Trees such that all loci in C are represented by a leaf node in one of the trees and the number of `NIL` values in the representative genotype vector of any root node does not exceed $d|S|$. We call d the 'threshold on the fraction of ambiguous samples'. In this formulation, d is a user-defined parameter that is used to balance the trade-off between 'compression' and accuracy of approximation. It can be shown that this formulation leads to a generalization of the bin-packing problem (26), and is therefore NP-hard. Furthermore, this formulation ignores the accuracy of approximation at each internal node, which is rather important for our end goal. Motivated by these observations, we construct the LD-Forest using an agglomerative heuristic that takes advantage of the knowledge of genomic proximity, and incorporates parameter d in tree construction, utilizing a dynamically adjusted threshold.

Procedure. Our forest construction algorithm starts with a collection of $|C|$ trees, where each tree corresponds to a single locus. Subsequently, it performs iterative scans through the list of all trees and merges pairs of trees such that the representative genotype of the root of the new tree contains at most d^* `NIL`s. The parameter $d^* \leq d$ is a dynamically adjusted threshold that grows with increasing number of iterations. Merging of LD-Trees is done greedily. Over a scan iteration, each tree is compared to the b closest trees, where b is a parameter that defines the extent of the utilization of genomic proximity in deciding which loci can be in LD. The individual LD-Trees are stored sequentially, when two trees are merged the position of the resulting tree within this sequential data structure is set to the previous location first constituent tree. This maintains the relative positions of all of the other trees. Recall that the before the merging procedure, the forest of LD-Trees contains each locus represented by a single LD-Tree arranged in genomic order. Thus the genomic distance between two LD-Trees at any iteration during merging is not explicitly calculated and b refers to the closest Trees within this sequential data structure. When b is very small, only loci that are next to each other are merged. As b gets larger, more and more of the genomic redundancy information provided by the GWAS data is utilized. We use $b = 10$ in our experiments.

The first merger that satisfies the bound on d^*, if any, is performed. In the first iteration, $d^* = 0$ in order to merge loci with identical genotype vectors, and in each successive iteration d^* is incremented by 1%. Thus later iterations allow a greater degree of uncertainty in newly formed root nodes. The merging procedure terminates when $d^* = d$ and no further merging is possible. In practice, the runtime of this algorithm is linear in $|C|$. The parameter b defines a constant maximum number of comparisons for a locus during a merge iteration. The number of merge iterations does not depend on the number of loci provided. Furthermore, the runtime of LD-forest construction is negligible as compared to the testing step, which is described in the next section.

Identification of epistatic pairs

The objective of LINDEN is to produce the correct set of reciprocally significant pairs. LINDEN implements a heuristic that starts by testing all pairs of the roots of LD-trees. This entails at a minimum, one test per pair of trees (between both root nodes). Subsequently, pairs of nodes that are children of promising pairs of parents are recursively tested. This strategy enables pruning out subtrees that are deemed as unlikely to contain any significant interactions between the leaves of the corresponding subtrees. The likelihood of the existence of a significant interaction between leaves is assessed using an estimation function that operates on the representative genomic profiles of the respective root nodes. As mentioned in Figure 1, we do not test loci within 1 Mb of one another. This is done in order to avoid false positives arising only from linkage disequilibrium rather than an interesting functional relationship. (27)

Significance estimation. There are two distinct types of tests between nodes in LD-Trees: tests between two leaf

nodes, and tests that involve internal nodes. Since leaf nodes each represent one locus where all sample genotypes are known, a test between two leaf nodes is a standard χ^2 test between a pair of loci. Naturally a test between two leaf nodes is the only type of test that can result in LINDEN discovering an epistatic interaction, as this is the only type that explicitly tests a pair of individual loci. For tests that involve internal nodes, our purpose is to estimate the likelihood that the corresponding subtrees contain a pair of leaves with significant epistasis. For this purpose, we use an estimation function to account for the fact that the genotype vectors are not perfectly representative of all child nodes. The choice of this function is important in trading off computational savings and accuracy.

A reasonable choice for the estimation function would be a function that provides a bound on the χ^2 statistic of any pair of leaves in the subtrees represented by the nodes being tested. Indeed, a provable bound can be obtained by 'filling in' missing genotypes with the values that lead to the largest possible χ^2 statistic. However, in our preliminary experiments, we observed that such a provable bound is too loose to provide any significant earnings in terms of the number of tests avoided. For this reason, we here use an estimation function that does not provide a provable bound, but provides a practically useful approximation to the likelihood that the subtrees contain a significant pair of leaves. It is important that the function we use here to evaluate pairs of internal nodes (sets of loci) is not a bounding function for the significance of pairs of loci in these sets. This function is not anti-monotonic either, thus it serves as a heuristic to prune out the search space. As we show in the result section. this heuristic delivers useful performance in practice. Nevertheless, a tight bounding function that would enable lossless (and potentially more effective) pruning of the search space remains an open problem and an effective solution to this problem may further improve the efficiency of epistasis detection.

When testing between internal nodes, we construct a contingency table from the pair of representative vectors, ignoring samples in which either vector contains a `NIL`. Subsequently, we compute the χ^2 statistic for this contingency table. For nodes x and y, we denote this approximation $F(x, y)$. This process is illustrated in Figure 3. Note that, the χ^2 statistic for any pair of leaves can be higher and lower than this estimation function. However, this estimation function provides a heuristic approximation to the χ^2 statistic for the pairs of leaves that are in the respective subtrees. The accuracy of this approximation depends on the parameter d. For this reason, in the Results section, we comprehensively assess the effect of the parameter d on the accuracy of the resulting algorithm. We also observe that there is an optimal value of d across multiple datasets.

Dynamic significance threshold. As described in the proposed formulation, we are interested in identifying the most significant epistatic partner for each locus. To avoid testing all possible partners for each locus (i.e. performing all of the $\binom{|C|}{2}$ tests), we use a dynamic threshold X^* on the χ^2 statistic. We denote the set of all possible pairs of loci as C_p. We also define a table Y that stores the number of discovered locus pairs for a range of χ^2 values:

$$Y(t) = \begin{cases} |C_p|, & \text{if } t = 0 \\ |\{c_i, c_j \in C_p : \chi^2(c_i, c_j) \geq t\}| & \text{if } 1 \leq t \leq T \end{cases} \quad (3)$$

Here, T is a bound on the maximum achievable χ^2 value by any pair of loci.The dynamic threshold is maintained at the maximum χ^2 for which there have been at least $|C|$ pairs found whose test statistics achieve a χ^2 at least as large, i.e.:

$$X^* = \max_{0 \leq t \leq T} \{Y(t) \geq |C|\} \quad (4)$$

The idea behind this approach is that the number of reciprocally significant pairs can be at most $\frac{|C|}{2}$, thus a pair that is not among the top $|C|$ significant pairs cannot be reciprocally significant. During testing of epistasis, any pair of subtrees that have χ^2 less than X^* are pruned out and none of the children of the respective nodes are tested. The dynamic nature of X^* translates into a branch and bound algorithm in which pruning becomes more aggressive as more significant c_p are found.

Note that the dynamic significance threshold is used to guide LINDEN in determining which LD-tree nodes should be expanded and is based on the locus pairs that have been detected as the algorithm progresses. Thus the dynamic threshold does not necessarily follow the same progression for different values of d. However, due to the large number locus pairs with similar significances in practice it is very close. Once LINDEN concludes, the final list of most significant reciprocal pairings is assessed for statistical significance with one of three methods: the bonferroni corrected p-value based on an exhaustive pairwise analysis, comparison to the background significance after permutation testing with randomized case/control labels, and bonferroni correction based on the number of tests performed by LINDEN. The cutoff obtained through permutation testing is the least conservative of the three but is extremely computationally intensive to calculate. In Figure 12 we show that bonferroni correction based on the tests performed by LINDEN is a good approximation of the permutation testing cutoff.

Pairwise tree test. A simple indexing data structure is used to keep track of the best partner found for each locus and the χ^2 statistic of that interaction. We refer to this structure as W. A stack Q is used to maintain the order of evaluation of node pairs between two LD-trees. To identify epistatic pairs of leaves between two LD-trees, we first test the roots of the two trees. This is achieved by using the previously described estimation function in the proposed formulation. After the roots x and y of two subtrees are tested, $F(x, y)$ is compared against X^*. If $F(x, y) > X^*$, then all pairs of the immediate child nodes of x and y are pushed to Q. Otherwise, all tests involving pairs of descendants of x and y are skipped. If x and y are leaf nodes and the resulting χ^2 statistic exceeds the dynamic threshold X^*, then the χ^2 statistic for x and y is compared to their current most significant discovered partners and W is updated accordingly. The Y table is updated accordingly and X^* increased as necessary. An example illustrating this algorithm is shown in Figure 4.

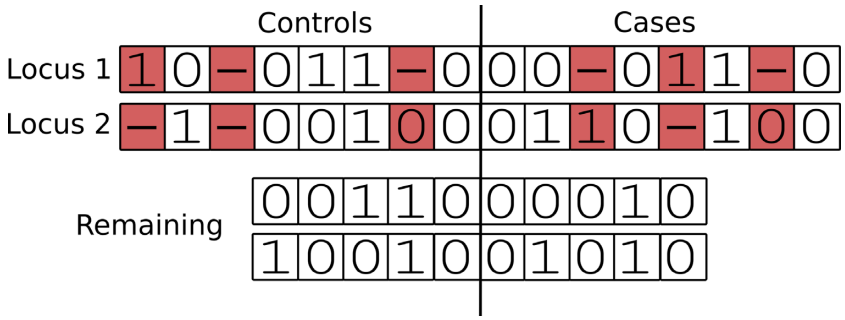

Figure 3. Test statistic approximation. When approximating the χ^2 test between two representative genotype vectors, the samples that do not have a value in the corresponding entries of the representative vectors of both loci are dropped during contingency table creation. Here, the resulting contingency table is constructed from 10 samples.

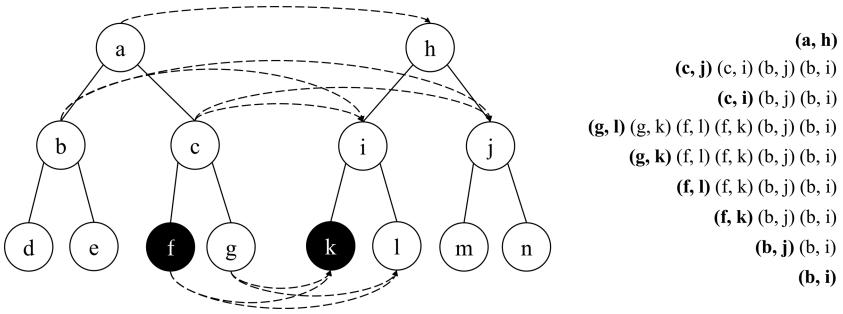

Figure 4. Testing of epistasis between two LD Trees. Each tree contains four loci represented by the leaf nodes. The stack that keeps track of the pairs of nodes to test is shown to the right. Each row corresponds to a different iteration. The pair of nodes that is being tested is shown in bold at each iteration. The black leaf nodes represent two loci that make up a significant pairwise interaction, while the dashed lines represent nodes that are tested against each other. First, node *a* is tested against node *h*, resulting in a value that is above the cutoff, thus all pairwise combinations of their child nodes are added into the stack. Of those, only the test between node *c* and node *i* passes the cutoff, resulting in a total of four leaf node tests performed.

RESULTS

In this section we provide an analysis of LINDEN's ability to reduce the number of statistical tests performed to detect the most reciprocally significant locus pairs, as well as its precision and recall. We next consider LINDEN's performance based on the density of the loci. We compare the performance of our greedy method for LD-Tree construction to a method using Plink's clumping tool. Next we show that LINDEN finds statistically significant locus pairs in all three WTCCC datasets. We provide an empirical comparison of performance between LINDEN, iLoci, and Plink fast-epistasis. Finally, we briefly consider the functional relevance of the locus pairs detected by LINDENon the WTCCC data.

Datasets

We comprehensively test the effectiveness of our method on three datasets obtained from the Wellcome Trust Case Control Consortium. Each set represents, the genotyped loci for patients with a specific disease phenotype and the control samples genotyped at the same loci. We provide a summary of the three datasets in Table 1.

For the comparison of performance between LINDEN, iLoci and Plink we utilize the tool GAMETES (28) to generate simulated GWAS data and implant epistatic loci. We further post process the GAMETES output to simulate linkage disequilibrium and locus density. We describe this procedure in detail later in this section.

Experimental setup

LINDEN is implemented in C++ and is available as open source at http://compbio.case.edu/linden/. To be able to perform comprehensive experiments and characterize the effect of all factors, we run LINDEN on smaller instances obtained by sampling from the entire dataset. Since the method exploits genomic proximity, we subsample contiguous SNPs to create each instance. Namely, for each configuration of parameters and performance criteria, we create five instances by selecting 50k contiguous loci (nearly 10% of the loci in the dataset) for each instance. The set of loci considered in each instance are disjoint from each other. We run the algorithm on each instance and report the mean and the standard deviation of each performance metric across all five instances. Subsequently, in the statistical significance and functional annotation sections, we report results on LINDEN's performance in identifying statistically significant and biologically relevant epistatic pairs on the entire dataset for all of the three diseases.

In the following sections, we use the extent of pruning, accuracy, and efficiency as the main performance criteria. For each parameter combination, we also plot the performance of an algorithm that exploits LD using a flattened tree structure, i.e. for each LD-Tree the leaf nodes are treated as a group. In this algorithm, a test between

Table 1. Description of the three datasets that were used to in computational experiments

Phenotype	#Loci	#Controls	#Cases
Type II Diabetes	495,476	1589	1914
Psoriasis	535,475	2178	5175
Hypertension	454,653	2001	2997

two LD-Groups consists of a random selection of one locus from each group. By comparing the performance of the flattened structure against our proposed hierarchical approach, we investigate whether LINDEN's use of LD-Trees results in a substantial improvement over more naive methods of linkage-disequilibrium prioritization.

Following standard procedure, we filter out all loci that have a minor allele frequency <5%. This is because, if a sufficient number of samples is not available for a given genotype, it becomes too difficult to detect any meaningful statistical association. We also discount interactions between loci that are within one Mbps of each-other. The purpose of this step is to filter out associations that likely stem from linkage disequilibrium between the two loci, rather than indicating a functional relationship. As we show in the statistical significance analysis, even with this standard, fairly conservative screening we are able to detect statistically significant locus pairs in all three datasets.

Reduction in the number of tests performed

We first assess LINDEN's ability to reduce the number of pairwise statistical tests performed as compared to an exhaustive enumeration of all locus pairs. LINDEN performs two types of epistasis tests: (i) tests that involve at least one internal node of an LD-Tree, (ii) tests that involve two leaf nodes (i.e. a test of epistasis between two specific loci). The tests that involve internal nodes represent the 'overhead' introduced by LINDEN in order to reduce the number of tests between the leaves. For this reason, to accurately characterize the reduction in the number of tests performed, we compare the total (internal and leaf) number of tests performed by LINDEN to the total number of pairs of loci in the dataset. In other words, letting z_i and z_ℓ respectively denote the number of internal node tests and number of leaf tests, we quantify LINDEN's performance in reducing the number of performed tests as:

$$\text{Fraction of tests performed} = (z_i + z_l)/\binom{|C|}{2} \qquad (5)$$

Furthermore, to assess the contribution of the tree structure, we also run LINDEN by treating each tree as a flat group of loci, removing the hierarchical information. Namely, after the LD-forest is constructed, we test pairs of randomly selected leaf nodes for each pair of trees. Therefore, the area between the two curves in Figure 5 represents the overhead from evaluating internal nodes in the tree structure. As seen in the figure, this overhead is consistent and effectively negligible. Furthermore, as demonstrated by the results presented in the accuracy section, this small overhead results in a drastic increase in the ability to detect the correct list of most significant reciprocal pairs.

Precision and recall

When the threshold on the fraction of ambiguous samples, $d = 0$, our method is equivalent to a standard pairwise exhaustive test of all loci. This is because LD-Trees are only formed between loci that exhibit identical genotype vectors, thus resulting in an LD-Forest containing only either trees of height zero each representing a single locus, or trees with identical genotype vectors at the leaf nodes. Based on this observation, we run LINDEN with $d = 0$ to obtain a list of reciprocally significant locus pairs, and treat this list as the 'ground truth' for reciprocally significant locus pairs. Using this list for different values of d, we assess the precision and recall of LINDEN in identifying reciprocally epistatic locus pairs. Namely, for a given $d > 0$, the recall is defined as the fraction of pairs identified by LINDEN, among those that are identified with $d = 0$. Similarly, precision is defined as the fraction of pairs identified, among those that are identified with threshold $d = 0$ over the total number of pairs identified. The results for this analysis are shown in Figure 6 (recall) and Figure 7 (precision). As seen in both figures, the overall behaviors of precision and recall are consistent, across the three datasets representing three different diseases, as well as between different trials on contiguous groups of loci within a dataset.

It is important to note that precision and recall are rather conservative measures of the performance of LINDEN. In practice, if a true reciprocally epistatic pair is not detected, frequently it is because one of the loci has been substituted for another and the resulting pairwise significance is very close to that of the pair in the exhaustive list. This notion is also supported by the high correlation between precision and recall in all three datasets. Furthermore, any reciprocal pair that is reported is highly significant, regardless of whether it belongs in the set of the reciprocal pairs identified by exhaustive testing.

Effect of the threshold on fraction of ambiguous samples

The choice of value for d represents a trade-off between accuracy and runtime. To provide reasonable guidelines for choosing this parameter, we systematically investigate its effect on performance. First, we observe that any value of $d \geq 0.5$ is not useful. Recall that d represents the maximum fraction of samples that can have an ambiguous genotype in a node of an LD-Tree, and therefore are dropped while testing internal nodes. When performing a pairwise test between two internal nodes, if both nodes drop half of their samples, and those samples are disjoint, then there is no information left to assess significance This becomes very likely when $d \geq 0.5$.

Assessing performance gain. Clearly, there is a trade-off between the reduction in the number of tests and the recall

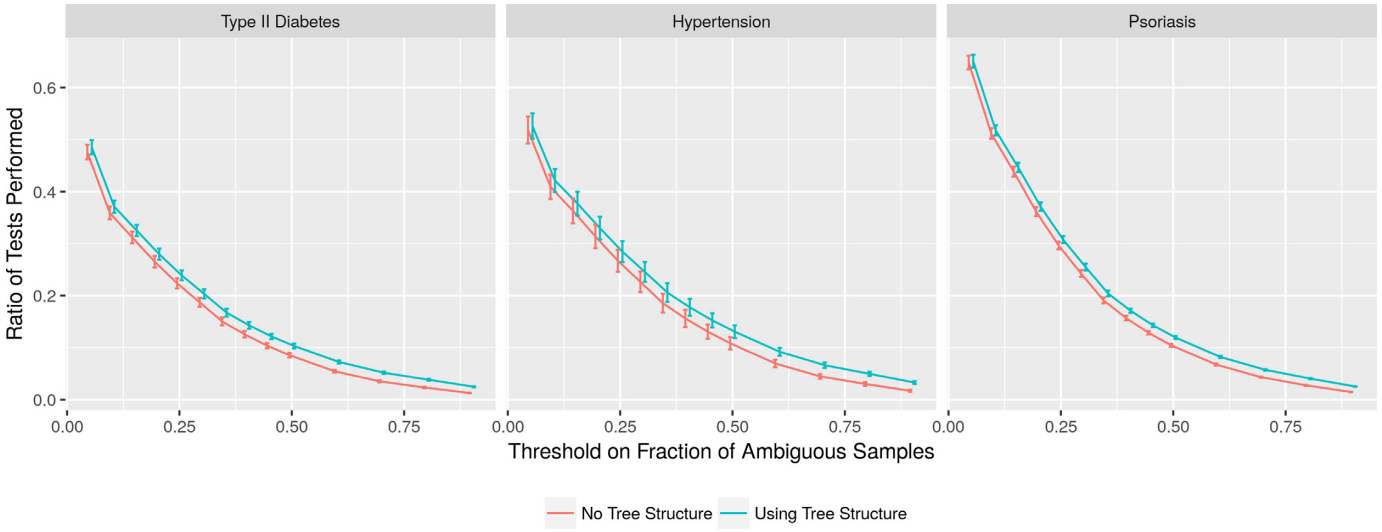

Figure 5. Reduction in the number of tests performed. Fraction of tests performed by LINDEN compared to exhaustive pairwise testing as a function of the threshold on the fraction of ambiguous samples.

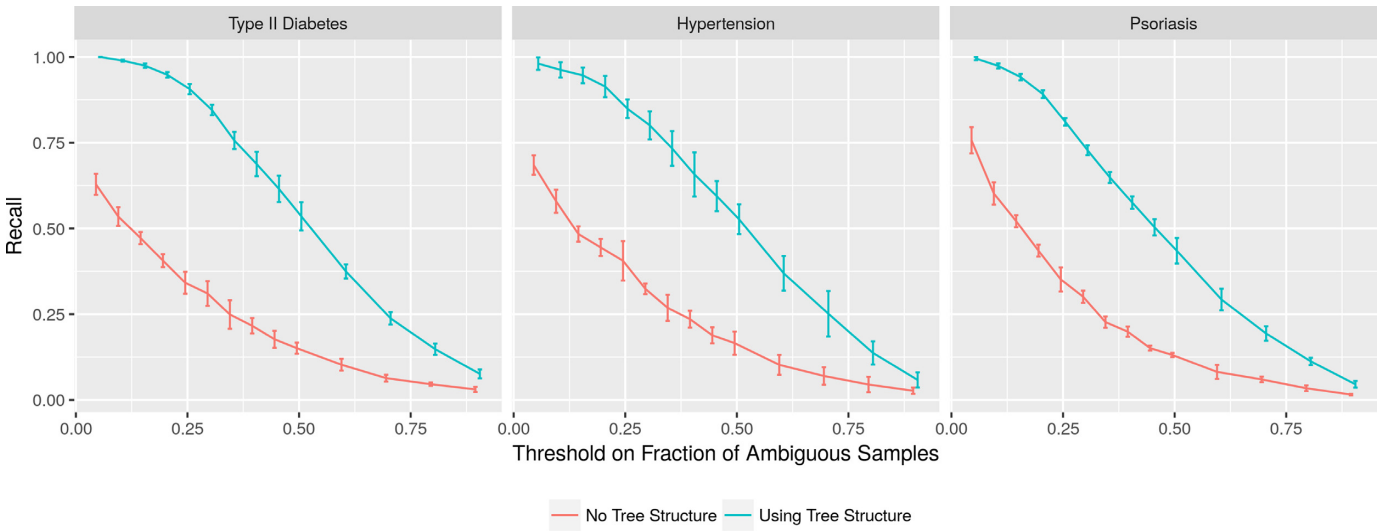

Figure 6. LINDEN's recall in identifying reciprocally significant locus pairs as compared to exhaustive testing. Plots depict recall as a function of the threshold on the fraction of ambiguous samples.

and precision of identified reciprocally significant pairs. To assess LINDEN's performance in resolving this trade-off, we propose a criterion, termed gain, that combines these two metrics. Namely, we define recall/precision gain as the ratio of recall/precision to the reduction in number of tests. To be more precise, we define:

$$\text{recall gain} = \frac{\text{recall}}{\text{fraction of tests performed}} \tag{6}$$

$$\text{precision gain} = \frac{\text{precision}}{\text{fraction of tests performed}} \tag{7}$$

The recall and precision gain as a function of d for all three datasets are shown in Figures 8 and 9 respectively. As seen in both figures, the gain provided by LINDEN grows as a function of the fraction of ambiguous samples (d), peaks around $d = 0.5$, and then sharply goes down with increased variance. The rapid decline in gain, as well as the instability for $d > 0.5$ is expected for the reasons explained previously. The growth in both recall and precision gain for values of d up to 0.5 is consistent across all datasets. This observation suggests that the hierarchical pruning provided by LINDEN is indeed beneficial, in that the more aggressively the loci are merged, the more number of tests are reduced, with tolerable loss of precision and recall. Note that, as seen in the figures, both recall and precision gain remain almost constant when the tree structure is not utilized. This observation demonstrates that the principled way of hierarchically approximating significance for groups of loci adds value to the exploitation of genomic redundancies.

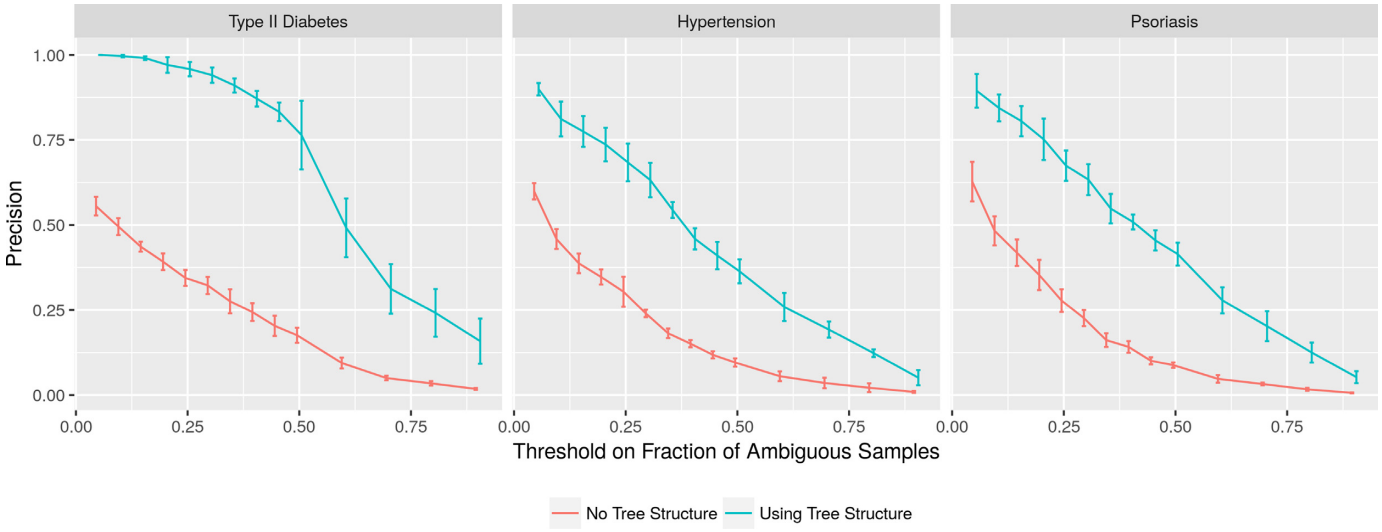

Figure 7. LINDEN's precision in identifying reciprocally significant locus pairs as compared to exhaustive testing. Plots depict precision as a function of the threshold on the fraction of ambiguous samples.

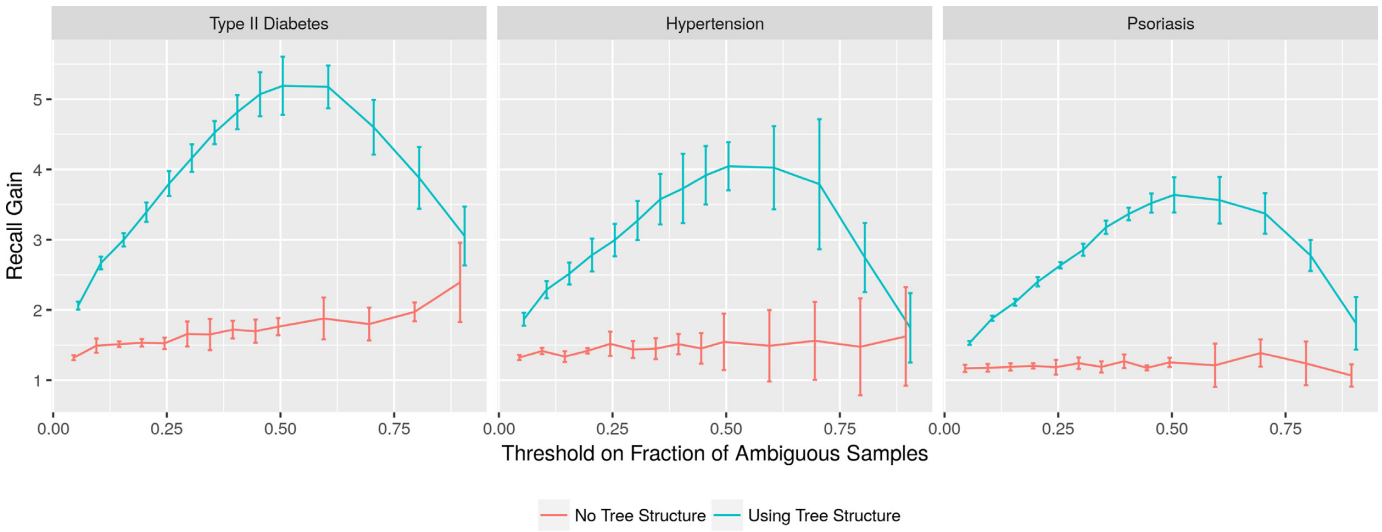

Figure 8. The gain provided by LINDEN in terms of recall in identifying reciprocally significant locus pairs as compared to exhaustive testing. Plots depict the fraction of recall and the reduction in the number of tests as a function of the threshold on the fraction of ambiguous samples.

Effect of genotype density

The proposed LD-Forest data structure has a uniquely useful property in that it becomes more efficient with increasing density of genotyped loci. A greater density means that the average distance between loci is decreased and thus the likelihood or degree of genotypic redundancy is increased and therefore LINDEN is likely to perform a greater degree of LD-Tree merging. For this reason, for a fixed genome, LINDEN effectively enables performance of a quadratic number of tests (in terms of the number of loci genotyped) in sub-quadratic time. This is particularly relevant in the context of whole-genome association studies as these studies cover a sampling from the entire genome, meaning that a larger input necessarily has a greater density.

In order to understand the effect of increasing density, we fix d to 0.45 (based on the observations reported in the previous subsection) and assess performance as a function of genotype density. Since available WTCCC data is already genotyped and it is not possible to increase its density without genotyping new loci, we generate data for different densities by sub-sampling lower-density loci from available loci. To be more precise, while generating genotype data for density ρ, we randomly select a starting locus and generate a GWAS dataset from the actual dataset by skipping $\lceil \frac{1}{1-\rho} \rceil$ contiguous loci for every locus retained. The results of this analysis are shown in Figure 10. We also show that the ratio of tests performed compared to an exhaustive enumeration decreases substantially as the genotype density increases. However, the accuracy and precision also decrease moderately, and observe that the overall gain provided by LINDEN improves with increased genotype density.

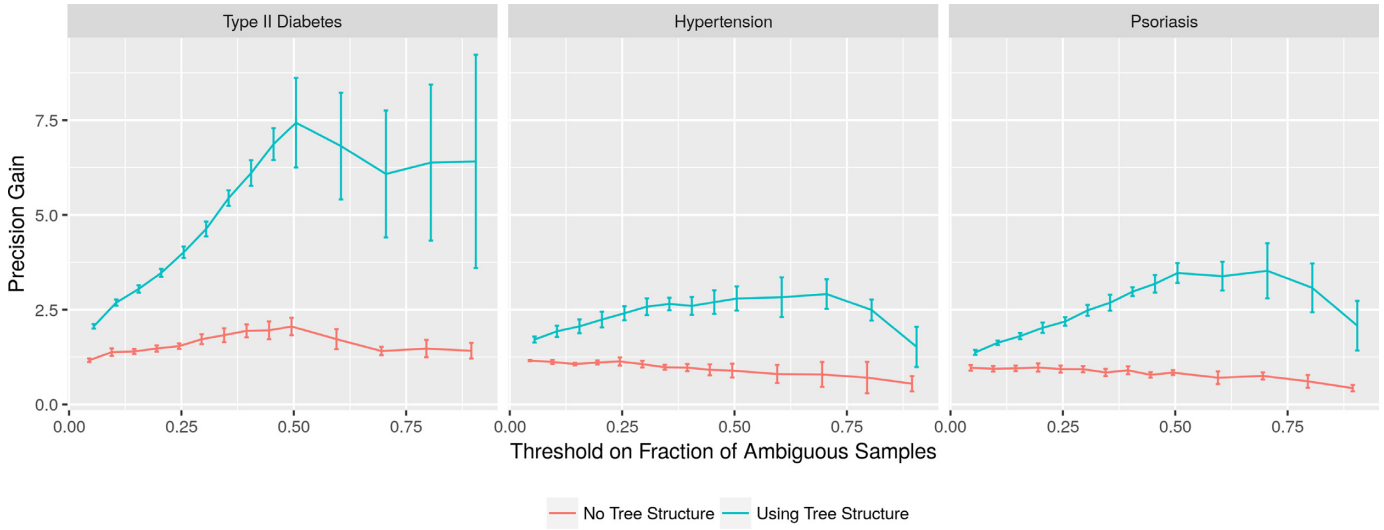

Figure 9. The gain provided by LINDEN in terms of precision in identifying reciprocally significant locus pairs as compared to exhaustive testing. Plots depict the fraction of precision and the reduction in number of tests as a function of the threshold on the fraction of ambiguous samples.

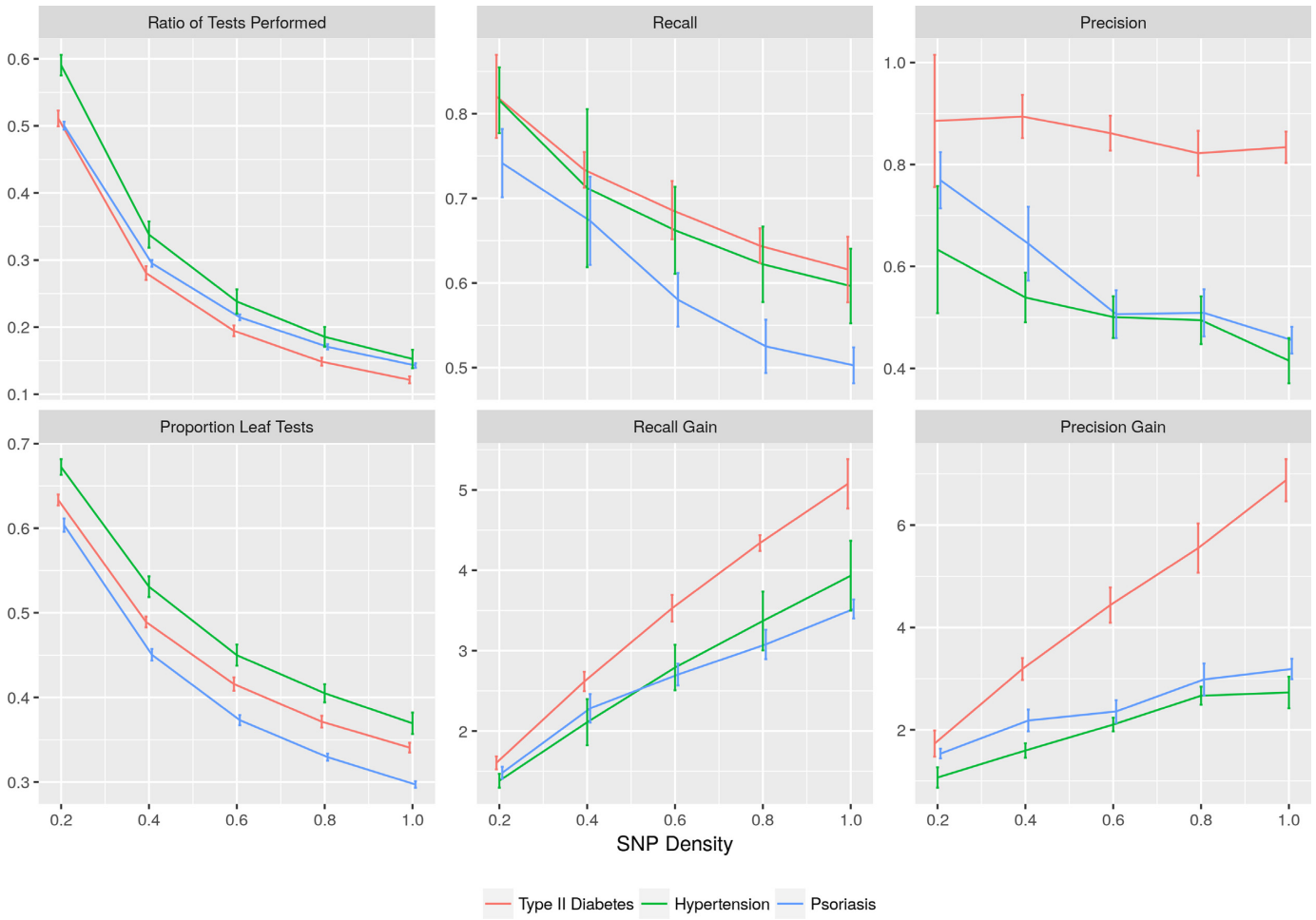

Figure 10. Effect of genotype density on performance. Each panel shows the behavior of a performance figure as a function of SNP density on subsampled loci from the three WTCCC datasets, for LINDEN and exhaustive testing. In these experiments, the threshold on the fraction of ambiguous samples is set to 0.45.

Comparison against established linkage disequilibrium values

Although LINDEN aims to exploit genomic redundancies that arise from linkage disequilibrium (LD), it utilizes genomic redundancy observed in the specific dataset rather then using established knowledge on the LD between different loci. To understand whether this data-oriented approach yields better results than utilizing LD information derived from the general population, we use the PLINK clumping tool (19) to group loci based on the r^2 values. A higher value of r^2 results in more aggressive grouping. We perform an exhaustive pairwise test between the representatives. The results of these analyses are shown in Figure 11. As seen in the figure, grouping based on r^2 yields relatively unfavorable performance as compared to LINDEN. Specifically, while grouping based on r^2 provides reasonable precision and recall if small groups are used, precision and recall rapidly decline as grouping is performed more aggressively. Inspection of the behavior of recall and precision gain as a function of the threshold on r^2 (aggressiveness of grouping) suggests that the gain is minimal and does not improve as grouping becomes more aggressive. We also observe a massive spike in variance near $r^2 = 1.0$ (most aggressive grouping), this is likely an artifact of the low number of reciprocal pairs detected at these high r^2 levels.

Statistical significance

We next test LINDEN's ability to find statistically significant pairs when evaluating the entire set of available loci for each of the three datasets. The results of this analysis are shown in Figure 12. We also provide summary statistics in Table 2. Although LINDEN is quite efficient, performing a complete pairwise search for multiple permutations across the three datasets is computationally intensive. For this reason, we use a 10% significance level rather than 5%, thus requiring a lower number of permutations. Each plot shows the *P*-value for the top one thousand discovered reciprocal locus pairs, as well as the top one thousand reciprocal pairs detected in ten iterations of permutation testing. We show the top one thousand in order to provide a comparison between the background pairwise significances against those pairs that are statistically significant. Only those points above at least one of cutoff lines represent a statistically significant pair based on permutation testing.

We use $d = 0.45$ to remain on the slightly conservative side of the parameter setting that appears to balance the trade-off between efficiency and accuracy. Even though 0.5 appears to be optimum, it is possible, though unlikely, for a test between two internal nodes to contain no usable samples if their sets of unknown genotypes are disjoint. For $d = 0.45$, on the other hand, the number of samples available to any internal node test is at least 10% of the total number of samples.

In permutation testing, we generate the null models by randomizing the case control labels for the samples, thus breaking the association between genotype and phenotype. Each plot also includes three horizontal lines representing three significance thresholds. Bonferroni-Standard refers to a cutoff that corresponds to Bonferroni correction for an exhaustive test of all pairs in the dataset at the 10% significance level. Bonferroni-Reduced-Tests refers to the cutoff calculated from the number of tests performed by LINDEN (where the number of tests performed is calculated as the total number of leaf tests and internal node tests) again at the 10% significance level. Permutation-Testing represents a cutoff chosen conservatively as the point that is above all scores achieved in the permuted datasets. Note that these are not Manhattan plots, the points are spread out to make it easier to see the individual points at the higher significance p-values.

As expected, in all three datasets we can see that the standard Bonferroni correction is too conservative, thus the overall statistical power of a standard exhaustive search is limited. Permutation testing generally produces less conservative, appropriate significance cutoffs with the trade-off that they are much more computationally intensive to calculate. In general, with the exception of specialized methods (11), permutation testing is not feasible in epistasis detection.

LINDEN's estimate of the significance threshold based on the number of tests performed is substantially closer to the cutoff determined by permutation testing while remaining as easy to calculate as a standard Bonferroni correction. This is because it implicitly takes into account the fact that the individual tests are not truly independent, and Bonferroni correction operates under the assumption of independent tests. Furthermore, notice that in all three data sets, there are reciprocal pairs that would not be reported as statistically significant if the standard Bonferroni cutoff were used. There is a clear practical benefit to our adjusted threshold.

Comparison with other methods

We compare the statistical power and runtime of LINDEN to iLoci (18) and the Plink (20) fast-epistasis tool. All three methods are intended to provide fast heuristic detection of epistatically interacting SNP pairs. We also considered TEAM (11) for comparison, but found that it does not scale to the number of SNPs we are interested in analyzing. We use the simulation tool GAMETES (28) to generate simulated genotype data and implant epistatic interactions.

Statistical power. For each model generated, we implant a single pair of epistatically interacting loci. GAMETES always places the target pair at the end of the dataset. Thus, we manually post-process the output of GAMETES to randomize the locations of the targets for each run. All trials consist of 4000 loci with 1500 case and control samples. The minor allele frequencies of the background loci range from 5 to 50%. We examine four levels of heritability and the minor allele frequencies of the target pair, resulting in a total of 16 different parameter combinations. For each parameter combination we perform 100 replicates. In Figure 13, we provide the results of these tests.

In general, LINDEN and Plink deliver superior performance as compared to iLoci. For most models, the performance of Plink and LINDEN is similar. However, Plink appears to have have difficulty with datasets that have low heritability and high minor allele frequency. Because Plink is exhaustive, this is probably an unavoidable mathematical property of compression in genotype categories when con-

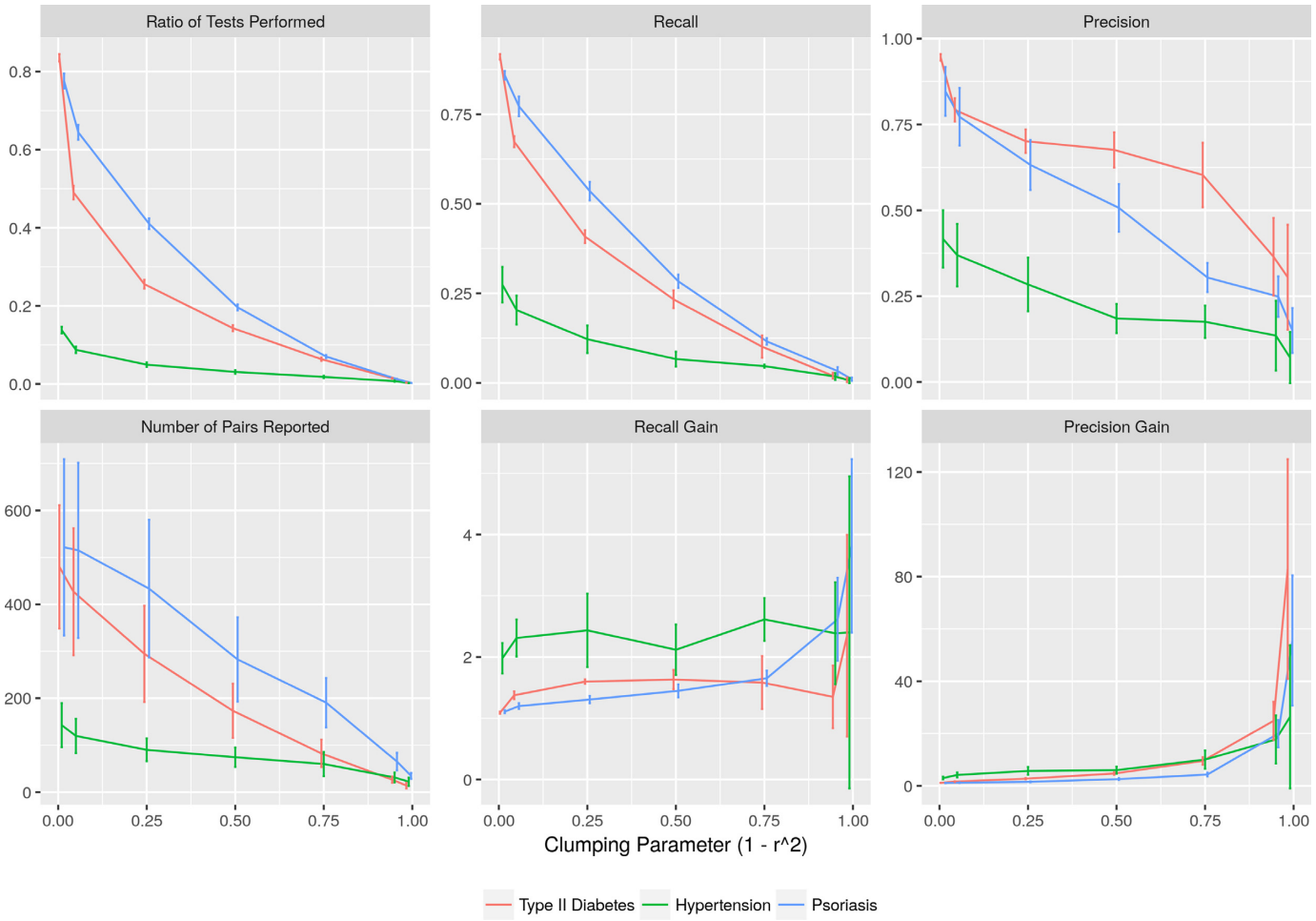

Figure 11. The performance of an algorithm that uses linkage disequilibrium in the general population to exploit genomic redundancies. PLINK's clumping tool is used to group loci based on linkage disequilibrium ($1 - r^2$). Each performance figure is plotted as a function of the minimum correlation ($1 - r^2$) required for clumping. Grouping of loci is more aggressive (more loci are grouped together) for larger values of the clumping parameter.

Table 2. Summary statistics on the experiments on real datasets

Phenotype	# of Tested loci	# of Trees	Proportion of internal tests	Proportion of leaf tests	Test reduction (%)	Fraction filtered	# Pairs reported
Type II Diabetes	375,801	123,297	72.5	27.5	10.7	0.04	2278 (19)
Psoriasis	514,158	186,115	75.1	24.9	13.1	0.02	3985 (21)
Hypertension	298,611	106,581	70.5	29.5	12.7	0.04	2379 (20)

Tested loci refers to the number of loci left after filtering by the minor allele frequency and significance of marginal effect as described in the methods section. This is followed by the number of trees generated by LINDEN and the proportions of internal (between groups of loci) and leaf (between individual loci) node tests. Test reduction is the number of tests performed divided by the number of pairs of loci that are tested by an exhaustive approach. Fraction filtered is the number of pairs that are not tested since they are too proximate on the genome. Finally, # Pairs reported is the total number of reciprocally significant pairs returned by LINDEN. The number in parentheses denotes the pairs passing Bonferroni correction based on the number of tests performed by LINDEN.

structing the odds ratio tables. LINDEN does not have this problem as it calculates the χ^2 statistic for the complete contingency table. In Figure 14, we consider the average rank of detected target pairs. As seen in the figure, when LINDEN detects the implanted pair, it always ranks it as the most significant pair. For Plink and iLoci, the target pair is not always ranked as the top pair when it is detected. This is especially problematic in the simulated data because the only pair that should have an epistatic interaction is the implanted pair. The rest of the potential pairs are background noise. This means that in many instances, Plink and iLoci are effectively unable to distinguish between the implanted pair the background.

Simulating linkage disequilibrium. To simulate linkage disequilibrium, we apply a two parameter transformation to the GAMETES output. For each locus in the original set, we generate a random integer in the interval [*0 .. mean*

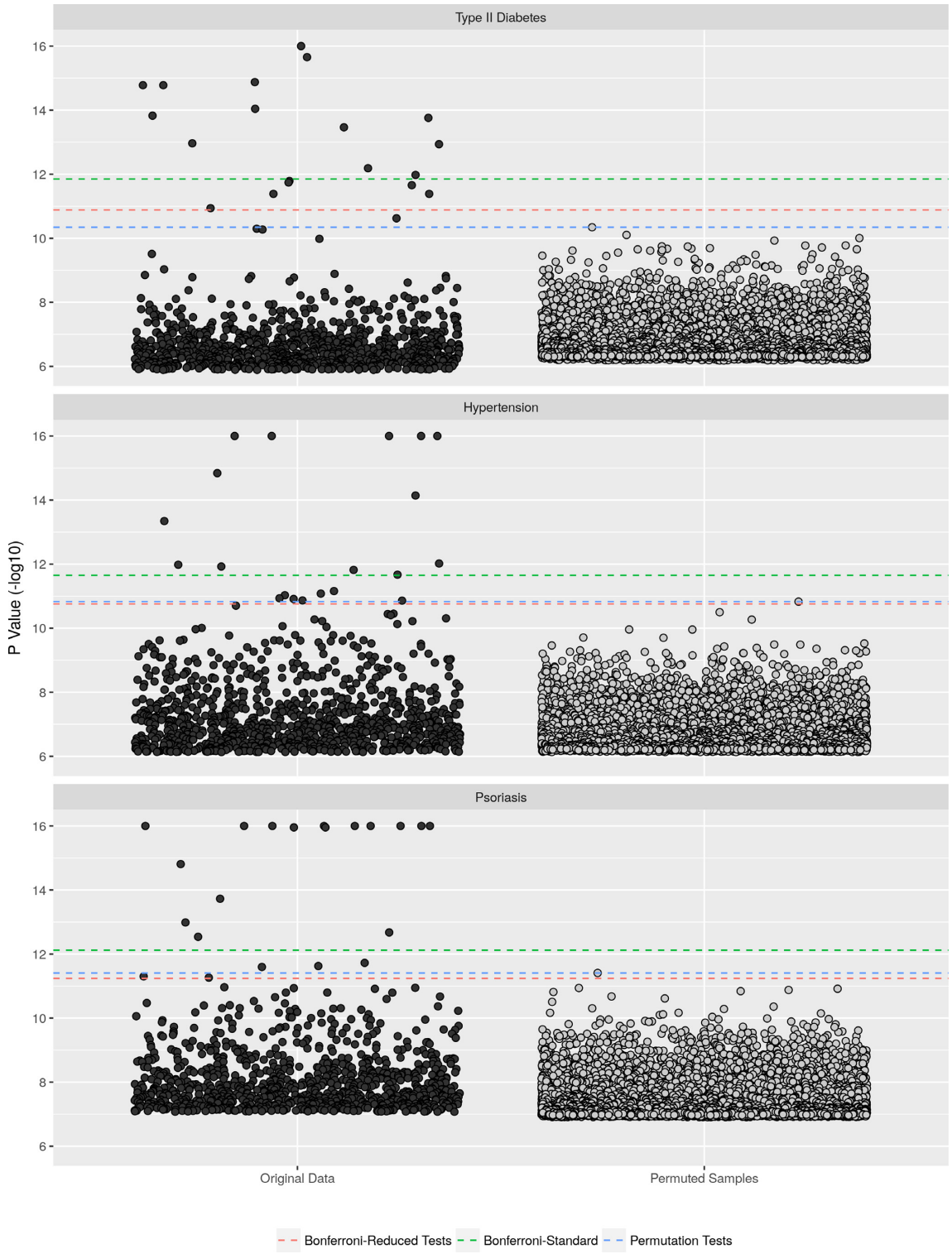

Figure 12. Statistical significance of epistatic loci identified by LINDEN. The reciprocally significant pairs of loci detected on each of the three datasets are shown on the left-hand-side of each panel, where each panel represents a different dataset. The reciprocally significant pairs of loci detected on ten permuted versions of these datasets are shown on the right-hand-side of each panel. For each disease, three different significance threshold are shown; Bonferroni correction considering all possible pairs of loci, Bonferroni correction based on the total number of tests (including leaves and internal nodes) performed by LINDEN, and the be p-value of the most significant pair identified across ten permutations. For all three diseases, the correction provided by the number of tests performed by LINDEN closely matches to the empirical correction provided by permutation tests. Note that, these are not Manhattan plots, the points have been staggered horizontally to aid in visualization.

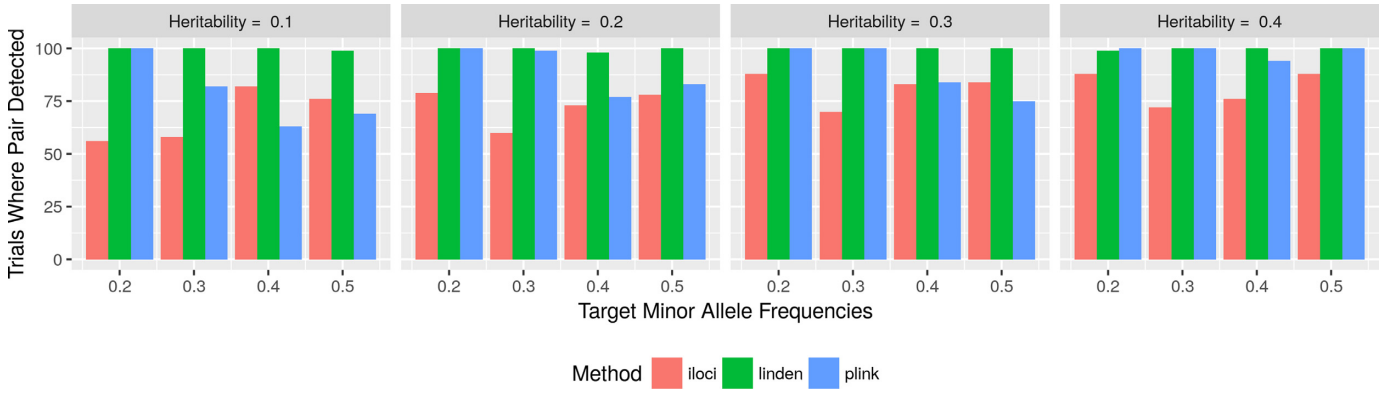

Figure 13. Power to detect target pair on simulated data. The figures show the number of trials in which iLoci, Plink, and LINDEN were able to detect the implanted epistatic pair in 100 trials, as a function of minor allele frequency and heritability.

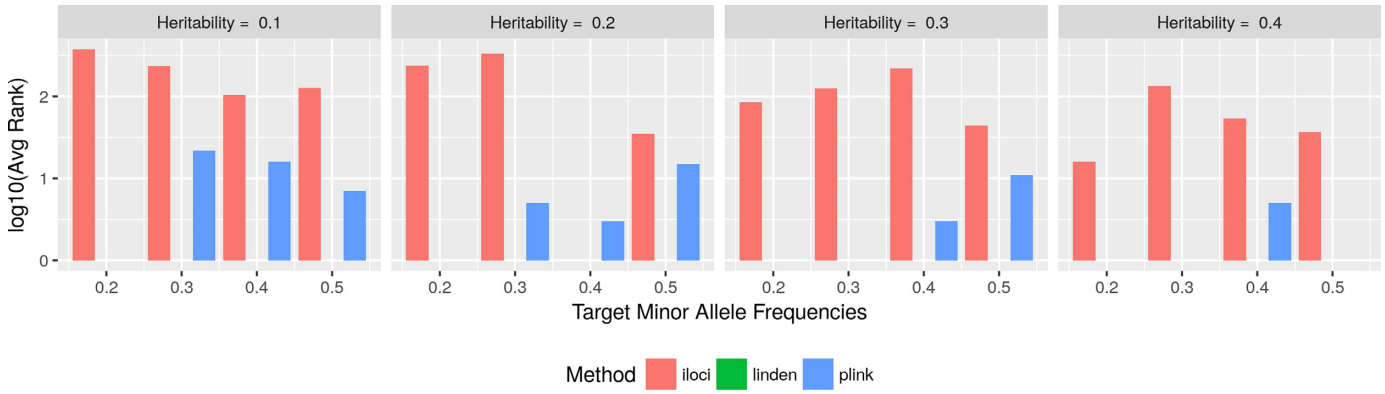

Figure 14. Ranking of target pair. For the trials in Figure 13 in which the target pair was detected, these figures show the average rank of the target pair in the list of identified interactions. Notice that this is in log scale, for all trials LINDEN reported the target pair as the most significant when detected.

Table 3. Gene Ontology functional enrichment for Type II Diabetes based on cellular component

GO cellular component	Expected	Enrichment	*P*-value
Synapse	3.32	3.92	3.37E–02
Unclassified	15.26	0.39	0.00E00

Table 4. Gene ontology functional enrichment for hypertension based on cellular component

GO cellular component	Expected	Enrichment	*P*-value
Cell projection	8.04	2.61	4.52E–02
Cell periphery	22.43	1.92	3.57E–03
Plasma membrane	21.96	1.91	5.46E–03
Membrane	40.45	1.51	2.09E–02
Unclassified	14.21	.21	0.00E00

Table 5. Gene Ontology functional enrichment for hypertension based on biological process

GO biological process	Expected	Enrichment	*P*-value
Anatomical structure morphogenesis	10.68	2.71	2.64E–03
System development	18.31	2.08	1.94E–02
Anatomical structure development	21.45	2.05	2.25E–03
Single-organism developmental process	23.78	1.89	1.61E–02
Developmental process	24.15	1.86	2.53E–02
Single-organism process	58.22	1.36	2.50E–02
Unclassified	19.50	0.31	0.00E00

Table 6. Gene ontology functional enrichment for psoriasis based on cellular component

GO cellular component	Expected	Enrichment	*P*-value
MHC class I protein complex	0.06	81.82	7.29E–06
MHC class II protein complex	0.10	68.73	2.01E–08
MHC protein complex	0.14	63.12	4.99E–11
Integral component of lumenal side of ER membrane	0.14	42.08	1.07E–05
Lumenal side of ERM	0.14	42.08	1.07E–05
Lumenal side of membrane	0.15	40.63	1.31E–05
ER to Golgi transport vesicle membrane	0.21	28.74	9.98E–05
ER to Golgi transport vesicle	0.26	23.10	3.55E–04
Transport vesicle membrane	0.43	14.03	6.18E–03
Integral component of ERM	0.63	11.17	4.36E–03
Intrinsic component of ERM	0.65	10.74	5.64E–03
Endocytic vesicle membrane	0.75	9.29	1.44E–02
Integral component of organelle membrane	1.31	6.85	9.11E–03
Intrinsic component of organelle membrane	1.38	6.55	1.30E–02
Endosome membrane	1.96	5.11	3.44E–02
Plasma membrane protein complex	2.60	5.00	2.63E–03
Endosome	3.88	3.61	4.08E–02
Membrane protein complex	5.57	3.05	4.56E–02
Unclassified	15.85	.50	0.00E00

Table 7. Gene ontology functional enrichment for psoriasis based on biological process

GO biological process	Expected	Enrichment	*P*-value
AP and PEPA via MHC I via ERP, TAP-independent	0.02	100	4.44E−03
AP and PEPA via MHC I via ERP	0.03	100	2.61E−04
AP and PEPA via MHC I	0.06	71.40	2.89E−03
AP and PEPA	0.06	65.45	4.07E−03
AP and presentation of endogenous antigen	0.07	56.10	7.48E−03
Interferon-gamma-mediated signaling pathway	0.39	20.40	5.96E−05
AP and POP or polysaccharide antigen via MHC II	0.50	13.88	6.92E−03
Response to interferon-gamma	0.74	13.45	3.87E−05
Cellular response to interferon-gamma	0.65	12.37	2.65E−03
AP and POP antigen	0.96	10.44	4.07E−04
AP and presentation of exogenous peptide antigen	0.87	10.33	2.13E−03
Positive regulation of T cell activation	0.98	10.17	5.18E−04
Positive regulation of homotypic cell-cell adhesion	1.00	9.97	6.25E−04
AP and presentation of exogenous antigen	0.91	9.93	2.96E−03
Positive regulation of leukocyte cell-cell adhesion	1.01	9.92	6.55E−04
AP and presentation	1.15	9.60	1.98E−04
Positive regulation of cell-cell adhesion	1.18	8.50	2.66E−03
Positive regulation of lymphocyte activation	1.27	7.89	5.23E−03
Regulation of T cell activation	1.43	7.71	1.75E−03
Regulation of leukocyte cell-cell adhesion	1.47	7.50	2.30E−03
Immune response-activating cell surface RSP	1.49	7.37	2.72E−03
Regulation of homotypic cell-cell adhesion	1.51	7.27	3.11E−03
Positive regulation of leukocyte activation	1.39	7.19	1.19E−02
Positive regulation of cell activation	1.44	6.96	1.58E−02
Regulation of lymphocyte activation	1.88	5.84	2.57E−02
Regulation of cell-cell adhesion	1.88	5.84	2.57E−02
Immune response-activating signal transduction	2.06	5.83	9.03E−03
Positive regulation of immune response	2.98	5.37	3.94E−04
Activation of immune response	2.30	5.22	2.78E−02
Regulation of cell activation	2.31	5.19	2.97E−02
Positive regulation of immune system process	4.41	4.53	1.15E−04
Regulation of immune response	4.74	4.22	3.68E−04
Innate immune response	5.15	3.88	1.43E−03
Immune response	7.28	3.30	1.42E−03
Regulation of immune system process	7.50	3.07	9.38E−03
Defense response	7.67	3.00	1.38E−02
Positive regulation of response to stimulus	10.35	2.80	1.77E−03
Cell surface RSP	11.10	2.70	2.22E−03
Immune system process	11.02	2.63	6.41E−03
Positive regulation of biological process	27.00	1.93	1.21E−03
Unclassified	21.76	.41	0.00E00

Additional abbreviations are as follows: AP (antigen processing), PEPA (presentation of endogenous peptide antigen), POP (presentation of peptide), RSP (receptor signaling pathway), ERP (endoplasmic reticulum pathway).

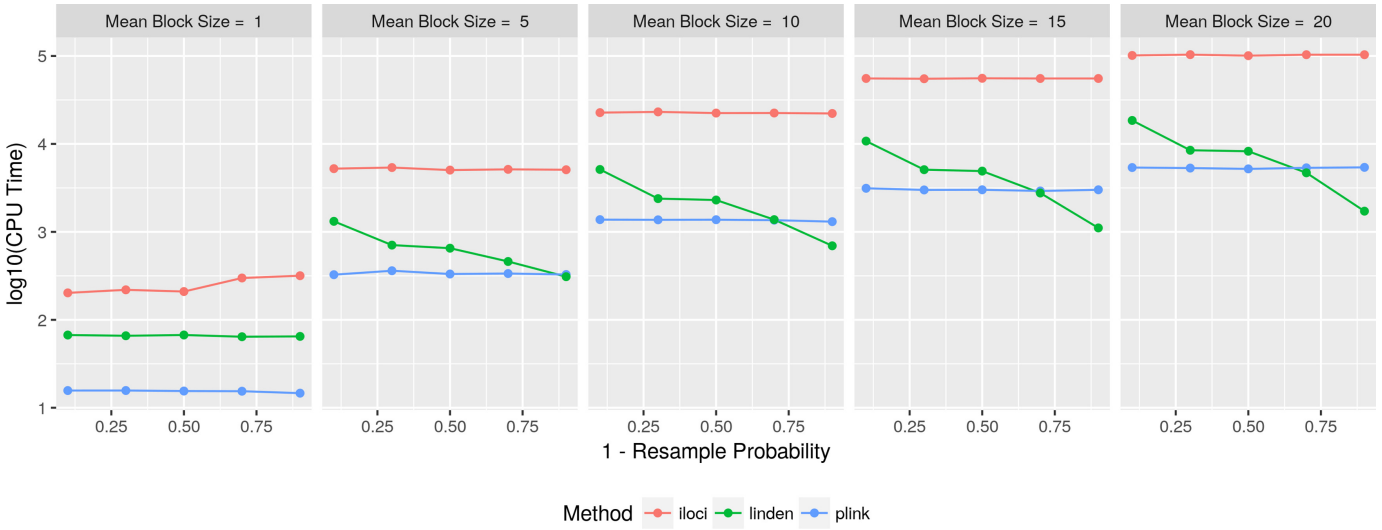

Figure 15. Runtime with simulated LD. Comparison of the CPU runtime required for iLoci, Plink and LINDEN, based on the simulation of linkage disequilibrium by replicating loci and adding noise in the data generated by GAMETES. The run-times are shown as a function of mean block size and re-sample probability. Here, mean block size refers to the average number of copies of loci and re-sample probability refers to the fraction of loci that have a different genotype as compared to their block. Thus, as the figures move from left to the right, the genotyped loci get denser and as we move to the right on the x axis, linkage disequilibrium gets stronger.

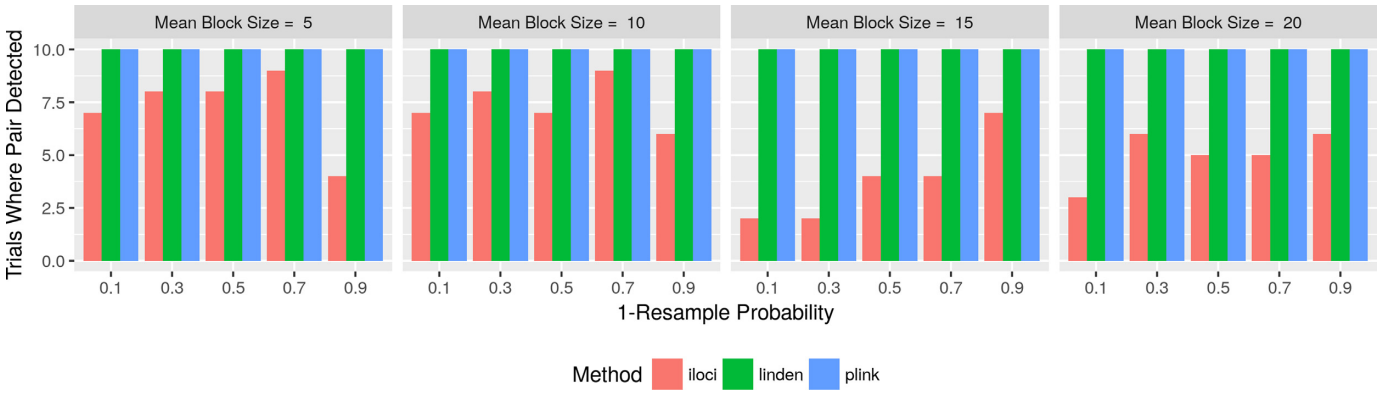

Figure 16. Power to detect epistatic pairs in simulated data with linkage disequilibrium. The figures show the number of trials in which iLoci, Plink and LINDEN are able to detect the implanted epistatic pair in 100 trials, as a function of resample probability and mean block size.

block size]. This specifies the number of extra loci to clone based on the original, both upstream and downstream of the locus. Initially, the genotype of each clone locus is identical to its predecessor. The second parameter refers to the probability that a given genotype of the new locus will be re-sampled. Note that the re-sampling probabilities are set to the frequencies of the predecessor's genotypes. For example, if an original locus is transformed such that the parameters are (2, 0.5) respectively, the original locus will generate a copy of itself upstream and downstream of itself. These copies will then re-sample each of their genotypes with a probability of 0.5 based on the original genotype frequencies. The new upstream locus generates another locus upstream of it, using its own genotype frequencies to re-sample, and likewise for the downstream locus. Under this model, the mean block size parameter represents the size of the haplotype blocks while the re-sample probability represents the overall density and linkage disequilibrium of the dataset.

Runtime. We next compare the runtime of the three methods as a function of the correlation between the genotypes of proximate loci. This is used to investigate the improvement in LINDEN's performance as the density, and thus linkage disequilibrium, of the input loci is increased. For this analysis the original set of loci contains 2000 SNPs, genotyped for 1500 cases and 1500 controls. We set heritability and minor allele frequency to 0.2 for the target pairs. The results are shown in Figure 15. As the mean block size increases each method requires a longer amount of time to complete. This is because the number of loci in each dataset increases, for example with a mean block size of five, there are roughly five times as many loci in the set. Both Plink and iLoci are unaffected by the re-sample probability and experience a constant time to calculate based solely on the number of loci in the dataset. LINDEN however, becomes more efficient as the re-sample probability decreases, where lower re-sample probability corresponds to increased dataset density and linkage disequilibrium. Larger mean block sizes

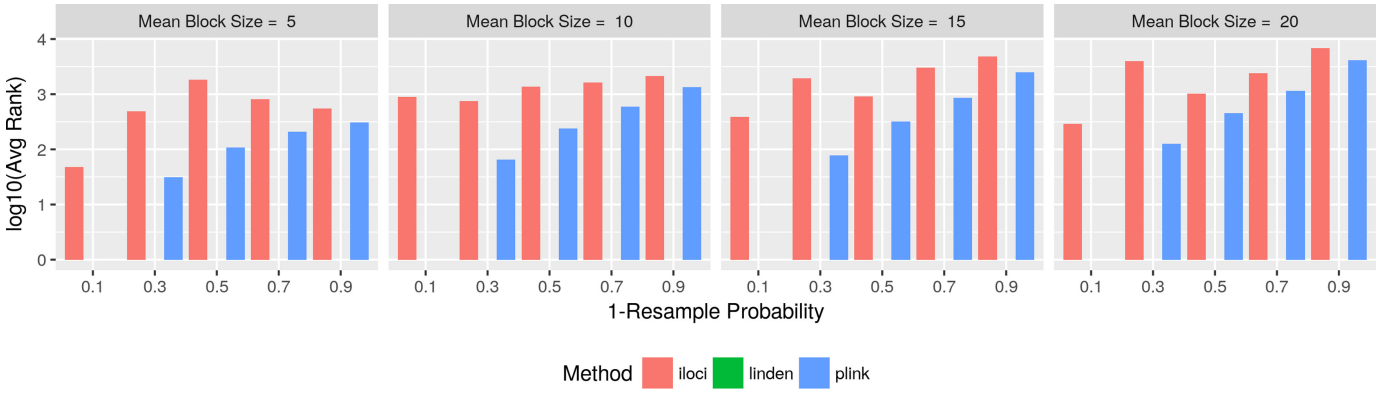

Figure 17. Ranking of target pair with simulated LD. For the trials in Figure 16 in which the target pair was detected, these figures show the average rank of the target pair in the list of pairings returned. Notice that this is in log scale, for all trials LINDEN reported the target pair as the most significant when detected.

also increase the efficiency. We can see that LINDEN begins faster than iLoci and becomes faster than plink with increased, 'density'. We also note that Plink and LINDEN provide speed-up through different means (Plink reduces the complexity of the tests without filtering tests whereas LINDEN reduces the number of tests). For this reason, it is possible to use these two methods in combination to further speed-up the computation.

Statistical power at the presence of linkage disequlibrium. To ensure that the addition of simulated linkage disequilibrium to the GAMETES output does not drastically change the statistical power of LINDEN. we calculate these measures for the same models and iterations shown in Figure 15 and provide these results in Figures 16 and 17. Notice that the value of 0.2 for heritability and minor allele frequency corresponds to an instance on which both Plink and LINDEN were able to detect the target pair without difficulty. Thus, as expected, both methods once again perform well with the simulated LD.

Functional annotation and biological relevance

In order to assess the functional significance of the epistatic locus pairs detected by LINDEN, we perform Gene Ontology enrichment analysis based on cellular component and process for all three phenotype sets. (29). In each case we, consider the top one hundred pairs of reciprocally significant loci. We then map the individual loci to the closest gene within 50kb. This results in 112, 123 and 105 loci mapped in the type II diabetes, psoriasis, and hypertension datasets respectively. The 'unclassified' term refers to genes that Gene Ontology was unable to classify. For all Gene Ontology analyses we show both the expected number of terms for each category and the enrichment based on that expected value. Thus an enrichment score greater than one describes an over-representation of genes for a given term.

For Type II diabetes, we find an over-representation of genes that are associated with the synapse cellular component shown in Table 3. Interestingly, a genetic link between propensity for type II diabetes and Alzheimer's through deficiency in synaptic function has been previously reported (30). Evidence of epigenetic links between synaptic impairments and diabetes have also been found (31).

There is a modest number of enriched terms associated with epistatic pairs for hypertension, as shown in Table 4 and Table 5. Overall, these terms seem to be more general, associated with structure and anatomical development.

The dataset with the most significant gene enrichment, for both process and component, is psoriasis. (14) This is not particularly surprising as psoriasis is an autoimmune disorder well known to have a strong genetic component, particularly in the MHC and HLA regions. From tables 6 and 7 we can see that the most significant over-representation categories include the MHC region, and most other categories involve the immune system in some capacity.

CONCLUSION

We have developed a fast method for the detection of epistatic interactions between pairs of loci in genome wide association data. By hierarchically grouping loci that are in high linkage disequilibrium we are able to reduce the number of statistical tests performed in an all pairs screen. This approach improves statistical power and speed of computation. Our algorithm exhibits sub quadratic complexity in the number of input loci when increasing overall genotype density. LINDEN is implemented in C++ and is available as open source at http://compbio.case.edu/linden/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Marzieh Ayati, Daniel Savel, Mustafa Coskun and Zachary Stanfield for useful discussions.
This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

FUNDING

Wellcome Trust [076113]; National Institute of Health [RO1-LM04127] from the National Libraries of Medicine; National Cancer Institute [U01-CA198941]. Funding for open access charge: National Cancer Institute [U01-CA198941].
Conflict of interest statement. None declared.

REFERENCES

1. Pritchard,J.K. and Cox,N.J. (2002) The allelic architecture of human disease genes: common disease–common variant...or not? *Hum. Mol. Genet.*, **11**, 2417–2423.
2. Visscher,P.M., Brown,M.A., McCarthy,M.I. and Yang,J. (2012) Five years of GWAS discovery. *Am. J. Hum. Genet.*, **90**, 7–24.
3. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
4. Rivas,M.A., Beaudoin,M., Gardet,A., Stevens,C., Sharma,Y., Zhang,C.K., Boucher,G., Ripke,S., Ellinghaus,D., Burtt,N. *et al.* (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, **43**, 1066–1073.
5. Billings,L.K. and Florez,J.C. (2010) The genetics of type 2 diabetes: what have we learned from GWAS? *Ann. N. Y. Acad. Sci.*, **1212**, 59–77.
6. Carlborg,Ö. and Haley,C.S. (2004) Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.*, **5**, 618–625.
7. Wei,W.-H., Hemani,G. and Haley,C.S. (2014) Detecting epistasis in human complex traits. *Nat. Rev. Genet.*, **15**, 722–733.
8. Cordell,H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, **11**, 2463–2468.
9. Marchini,J., Donnelly,P. and Cardon,L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.
10. Risch,N. (1990) Linkage strategies for genetically complex traits. I. Multilocus models.. *Am. J. Hum. Genet.*, **46**, 222.
11. Zhang,X., Huang,S., Zou,F. and Wang,W. (2010) TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, **26**, i217–i227.
12. Wan,X., Yang,C., Yang,Q., Xue,H., Fan,X., Tang,N.L. and Yu,W. (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.
13. McKinney,B. and Pajewski,N.M. (2011) Six degrees of epistasis: statistical network models for GWAS. *Front. Genet.*, **2**, 109.
14. Liu,Y., Maxwell,S., Feng,T., Zhu,X., Elston,R.C., Koyutürk,M. and Chance,M.R. (2012) Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from GWAS data. *BMC Syst. Biol.*, **6**, S15.
15. He,D., Wang,Z. and Parada,L. (2015) MINED: an efficient mutual information based epistasis detection method to improve quantitative genetic trait prediction. In: *Bioinformatics Research and Applications*. Springer, pp. 108–124.
16. Prabhu,S. and Pe'er,I. (2012) Ultrafast genome-wide scan for SNP–SNP interactions in common complex disease. *Genome Res.*, **22**, 2230–2240.
17. Ayati,M. and Koyutürk,M. (2014) Prioritization of genomic locus pairs for testing epistasis. In: *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, pp. 240–248.
18. Piriyapongsa,J., Ngamphiw,C., Intarapanich,A., Kulawonganunchai,S., Assawamakin,A., Bootchai,C., Shaw,P.J. and Tongsima,S. (2012) iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. *BMC Genomics*, **13**, S2.
19. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., De Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
20. Chang,C.C., Chow,C.C., Tellier,L.C., Vattikuti,S., Purcell,S.M. and Lee,J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 1.
21. Reich,D.E., Cargill,M., Bolk,S., Ireland,J., Sabeti,P.C., Richter,D.J., Lavery,T., Kouyoumjian,R., Farhadian,S.F., Ward,R. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
22. Cordell,H.J. and Clayton,D.G. (2005) Genetic association studies. *The Lancet*, **366**, 1121–1131.
23. Weiss,K.M. and Clark,A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.*, **18**, 19–24.
24. Ma,L., Ballantyne,C., Brautbar,A. and Keinan,A. (2014) Analysis of multiple association studies provides evidence of an expression QTL hub in gene-gene interaction network affecting HDL cholesterol levels. *PLoS One*, **9**, e92469.
25. Lippert,C., Listgarten,J., Davidson,R.I., Baxter,J., Poon,H., Kadie,C.M. and Heckerman,D. (2013) An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. *Scientific Rep.*, **3**, 1099.
26. Dyckhoff,H. (1990) A typology of cutting and packing problems. *Eur. J. Oper. Res.*, **44**, 145–159.
27. Wang,X., Elston,R.C. and Zhu,X. (2010) The meaning of interaction. *Hum. Heredity*, **70**, 269–277.
28. Urbanowicz,R.J., Kiralis,J., Sinnott-Armstrong,N.A., Heberling,T., Fisher,J.M. and Moore,J.H. (2012) GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Mining*, **5**, 1.
29. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B., Lewis,S., Group,W. P.W. *et al.* (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.
30. Han,W. and Li,C. (2010) Linking type 2 diabetes and Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6557–6558.
31. Wang,J., Gong,B., Zhao,W., Tang,C., Varghese,M., Nguyen,T., Bi,W., Bilski,A., Begum,S., Vempati,P. *et al.* (2014) Epigenetic mechanisms linking diabetes and synaptic impairments. *Diabetes*, **63**, 645–654.