Epistasis within the MHC contributes to the genetic architecture of celiac disease

Benjamin Goudey[1,2], Gad Abraham[3], Eder Kikianty[4], Qiao Wang[1,2], Dave Rawlinson[1,2], Fan Shi[1,2], Izhak Haviv[5], Linda Stern[2], Adam Kowalczyk[1,2,6,*], Michael Inouye[3*]

[1]NICTA Victoria Research Lab, The University of Melbourne, Parkville, Victoria 3010, Australia
[2]Department of Computing and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia
[3]Medical Systems Biology, Department of Pathology and Department of Microbiology & Immunology, The University of Melbourne, Parkville, Victoria 3010, Australia
[4]Department of Mathematics, University of Johannesburg, PO Box 524, Auckland Park 2006, South Africa
[5]Bar Ilan University, Safed, Israel
[6]Center for Neural Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia

[*]These authors contributed equally

Correspondence should be addressed to Michael Inouye (minouye@unimelb.edu.au) and Adam Kowalczyk (kowa@unimelb.edu.au)

Abstract

Epistasis has long been thought to contribute to the genetic aetiology of complex diseases, yet few robust epistatic interactions in humans have been detected. We have conducted exhaustive genome-wide scans for pairwise epistasis in five independent celiac disease (CD) case-control studies, using a rapid model-free approach to examine over 500 billion SNP pairs in total. We found 5,359 significant epistatic pairs within the HLA region which achieved stringent replication criteria across multiple studies. These pairs comprised 20 independent epistatic signals which were also independent of known CD risk HLA haplotypes. The strongest independent CD epistatic signal corresponded to variants in the HLA class III region, in particular *PRRC2A* and *GPANK1/C6orf47* which are known to contain variants for non-Hodgkin's lymphoma and early menopause, co-morbidities of celiac disease. Replicable evidence for epistatic variants outside the MHC was not observed. Both within and between European populations, we observed striking consistency of epistatic models and epistatic model distribution. Within the UK population, models of CD based on both epistatic and additive single-SNP effects increased explained CD variance by approximately 1% over those of single SNPs. Models of only epistatic pairs or additive single-SNPs showed similar levels of CD variance explained, indicating the existence of a substantial overlap of additive and epistatic components.

Introduction

The limited success of genome-wide association studies (GWAS) to identify common variants that substantially explain the heritability of many complex human diseases and traits has led researchers to explore other potential sources of heritability (in the wide sense), including the low/rare allele frequency spectrum as well as epistatic interactions between genetic variants [1,2]. Many studies are now leveraging high-throughput sequencing with initial findings beginning to elucidate the effects of low frequency alleles [3-6]. However, the characterization of the epistatic component of complex human disease has been limited, despite the availability of a multitude of statistical approaches for epistasis detection [7-13]. Large-scale systematic research into epistatic interactions has been hampered by several computational and statistical challenges mainly stemming from the huge number of variables that need to be considered in the analysis (>100 billion pairs for even a small SNP array), the subsequent stringent statistical corrections necessary to avoid being swamped by large number of false positive results, and the requirement of large sample size in order to achieve adequate statistical power.

The strongest evidence for wide-ranging epistasis has so far come from model organisms [14,15], and recent evidence has demonstrated that epistasis is pervasive across species and is a major factor in constraining amino acid substitutions [16]. Motivated by the hypothesis that epistasis is commonplace in humans as well, recent studies have begun providing evidence for the existence of epistatic interactions in several human diseases, including psoriasis [17], multiple sclerosis [18], Behçet's disease[19], type 1 diabetes[20], and ankylosing spondylitis [21], as well as the effects of epistasis on expression levels of multiple genes in human peripheral blood [22]. While these studies have been crucial in demonstrating that epistasis does indeed occur in human disease, several questions remain including how wide-ranging epistatic effects are, how well epistatic pairs replicate in other datasets, how the discovered epistatic effects can be characterized in terms of previously hypothesized models of interaction [23,24], and how much (if at all) epistasis contributes to disease heritability [25].

Celiac disease (CD) is a complex human disease characterized by an autoimmune response to dietary gluten. CD has a strong genetic component largely concentrated in the MHC region, due to its dependence on the HLA-DQ2/DQ8 heterodimers encoded by the HLA class II genes *HLA-DQA1* and *HLA-DQB1* [26]. The genetic basis of CD in terms of individual SNP associations has been well-characterized in several GWAS [27-30], including the additional albeit smaller contribution of non-HLA variants to disease risk [31]. The success of GWAS for common variants in CD has recently been emphasized by the development of a genomic risk score that could prove relevant in the diagnostic pathway of CD [32]. Autoimmune diseases have so far yielded the most convincing evidence for epistatic associations[33], potentially due to power considerations since these diseases usually tend to depend on common variants of moderate to large effect within the MHC. Given these findings in conjunction with recent observations that rare coding variants may play a negligible role in common autoimmune diseases [3], we sought to determine whether robust epistasis is detectable in CD and whether it accounts for some of the unexplained disease variance.

Here, we present a large-scale exhaustive study of pairwise epistasis in celiac disease. Leveraging GWIS, a highly efficient approach for epistasis detection [34], we conduct genome-wide scans for all epistatic pairs across five separate CD case/control datasets of European descent, finding thousands of statistically significant pairs despite stringent multiple testing corrections. Next, we show a high degree of concordance of these interactions across the datasets, demonstrating that they are highly robust and replicable. We characterize the common epistatic models found and compare them to previously proposed theoretical models. Further, given complex linkage disequilibrium patterns, we distil the epistatic pairs down to those which are independent of known HLA risk haplotypes and independent of other epistatic pairs. Finally, we examine whether epistatic pairs add more predictive power and explain more disease variation than additive effects of single SNPs.

Results

Datasets are summarized in **Table 1**, these include five independent, previously published GWAS datasets of CD with individuals genotyped from four different European ethnicities: United Kingdom (UK1 and UK2), Finland (FIN), The Netherlands (NL) and Italy (IT) [28,29]. To limit the impact of genotyping error and other sources of non-biological variation, we implemented three stages of validation and quality control (QC): (i) standard QC within each dataset, (ii) independent exhaustive epistatic scans within each of the five datasets, and (iii) derivation of a validated list of epistatic interactions based on UK1. The study workflow is shown in **Figure S1**.

Exhaustive epistatic scans and replication

For each dataset, we implemented stringent sample and SNP level quality control (**Methods**), and then conducted an exhaustive analysis of all possible SNP pairs using the GWIS methodology [34]. Each pair was tested using the GSS statistic, which determines whether a pair of SNPs in combination provides significantly more discrimination of cases and controls than either SNP individually (**Methods**). Forty-five billion pairs were evaluated in the UK1 study (Illumina Hap300/Hap550) and 133 billion SNP pairs were evaluated in each of the four remaining cohorts (Illumina 670Quad and/or 1.2M-DuoCustom). Given this multiple testing burden, we adopted stringent Bonferroni-corrected significance levels of $P = 1.1 \times 10^{-12}$ for the UK1 and $P = 3.75 \times 10^{-13}$ for the remaining datasets.

To further ensure that the downstream results were robust to technological artefact and population stratification, we took two additional steps: (a) utilizing the raw genotype intensity data available for UK1 for independent SNP cluster plot inspection (performed by Karen A. Hunt, QMUL), and (b) replicating the epistatic interactions of the SNPs passing cluster plot inspection, where replication is defined as a SNP pair exhibiting Bonferroni-adjusted significance both in UK1 and in at least one additional study. Using these criteria, we found that 5,359 SNP pairs (581 unique SNPs) from the UK1 dataset passed both (a) and (b) above. We denote these pairs as 'validated epistatic pairs' (VEPs) below. The full list of VEPs is given in **Supplementary Table 1**. Notably, all VEPs fulfilling these robustness criteria were within the MHC.

More than 126,000 unique pairs achieved Bonferroni-adjusted significance across all five studies, with the vast majority lying within the extended MHC region of chr 6 (**Figure 1** and **Table S2**). Of the 34 epistatic pairs outside the MHC that were significant in at least one study, none passed Bonferroni-adjusted significance in at least one other study and were thus deemed not replicated. As expected, the number and significance of epistatic interactions increased with sample size. Interestingly, some of the strongest epistatic interactions tended to be in close proximity though few SNPs were in LD with only 1% of pairs having $r^2 > 0.5$ (**Figure S2)**. The heatmaps in **Figure 1** also showed that epistasis was widely distributed with distances of >1Mb common between epistatic pairs. While epistatic interactions were consistently located in and around HLA class II genes, further examination of the VEPs found that many of the strongest epistatic pairs were in HLA class III loci, >1Mb upstream of *HLA-DQA1* and *HLA-DQB1* (**Figure S3**).

The extent of replication of the epistatic pairs was apparent from the high degree of similarity in the rankings when pairs were sorted by GSS significance (**Figure 2**), with the top 10,000 pairs exhibiting ~70-80% overlap between the UK1 and UK2 datasets and 40-60% overlap of the UK1 with the pairs found in the NL and FIN datasets. Such high degrees of overlap have essentially zero probability of occurring by chance. The pairs found in the IT dataset showed lower levels of consistency with those detected in the UK1 dataset but overall were still far more than expected by chance with ~30% overlap at ~30,000 pairs ($P < 10^{-1000}$, hypergeometric test).

Empirical epistatic model distributions

The epistatic model provides insight into how disease risk is distributed across the nine pairwise genotype combinations. Following the conventions of Li and Reich [23], we discretized the models for the VEPs to use fully-penetrant values where each genotype combination implies a complete susceptibility or protective effect on disease (**Methods**), simplifying the comparison of models between different SNP pairs.

To establish model consistency, we first replicated the most frequent full penetrance VEP models in the other datasets (**Figure 3**). When considering the distribution of epistatic models we found striking consistency of the UK1 models with those from UK2 and the other Northern European populations (Finnish and Dutch) (**Figure 3**). Only four models from the possible 50 classes [23] occurred with >5% frequency in the Northern European studies, and there was substantial variation in epistatic model as a function of the strength of the interaction. Amongst all VEPs in UK1, the four models corresponded to the threshold model (T; 38.3% frequency), jointly dominant-dominant model (DD; 31.1%), jointly recessive-dominant model (RD; 16.5%), and modifying effect model (Mod; 1.0%) [23,35]. The DD and RD models are considered multiplicative, the Mod model is conditionally dominant (i.e. one variant behaves like a dominant model if the other variant takes a certain genotype), and the T model is recessive. The T model was the most frequent model, especially amongst the strongest pairs.

Interestingly, despite the consistency of MHC epistasis, the VEPs showed noticeable differences in epistatic model distribution in the IT population. This was in contrast to the other Northern European populations but consistent with the different ranking in GSS significance observed above. In the IT population, the distribution of models was altered such that there was a more even distribution. The four most frequent models were still the T model (16.2%), modifying effects (16.5%), DD (15.8%), and RD model (11.3%). But, we also observed that many of the strongest pairs within the IT cohort followed the M86 model, though M86 represented only a small proportion of models overall (1.2%). The other VEP models overall were relatively uniform amongst the remaining models.

The cause(s) of the differences in epistatic model distribution for the IT data are unclear. While cryptic technical factors cannot be ruled out at this stage, we speculate that there may be population specific epistatic variation that follows the known North/South European genetic gradient [36].

Independence of Validated Epistatic Pairs and known HLA risk haplotypes

The VEPs demonstrate that thousands of epistatic SNP pairs can be found in the HLA region. However, due to their co-localization within a region of complex linkage disequilibrium and the presence of known risk haplotypes, it is important to estimate the number of epistatic signals which are (a) independent of the risk haplotypes, and (b) independent of other epistatic signals.

To determine the VEPs that were independent of HLA risk haplotypes, we utilized a likelihood ratio test with the threshold for significance defined using a false discovery rate (FDR) of 5% (P value < 0.044) (**Methods**). After maintaining the VEPs that were independent of risk haplotypes, we used an LD pruning based approach based on Hill's Q statistic, a normalized chi-squared statistic for multi-allelic loci [37], to estimate the number of independent epistatic pairs.

From the 5,359 VEPs, we found that 4,744 pairs (>88% of VEPs) were independent of the known CD risk haplotypes (*DQB1*0302, DQB1*0301, DQB1*0202, DQB1*0201 and DQA1*0301, DQA1*0505, DQA1*0501, DQA1*0201*), indicating that the HLA restricted epistasis is largely independent of known CD risk haplotypes. From the 4,744 VEPs independent of risk HLA haplotypes, LD pruning using a cutoff of Hill's $Q = 0.3$ identified 20 VEPs in UK1 which represent independent epistatic signals (**Methods** and **Table 2**). The diversity of CD risks conferred by the 20 independent VEPs together with HLA risk haplotypes is represented in **Figure S4**.

Contribution of epistatic pairs to celiac disease variance

We next sought to estimate the CD variance explained by the VEPs and single SNPs. To do this, we utilized a multivariable model framework which accounted for all SNPs and/or VEPs at once. To assess the contribution of epistatic pairs to CD prediction and thus genetic variance explained, we employed L1 penalized linear support vector machines (SVM, see **Methods**), an approach which models all variables concurrently (single SNPs and/or pairs) and which has been previously shown to be particularly suited for maximizing predictive ability from SNPs in CD and other autoimmune diseases [32,38]. We have previously shown that additive models of single SNPs explain substantially more CD variance than haplotype-based models [30], thus we employ only the former to estimate the gain in CD variance explained here.

We assessed CD variance explained by constructing three separate models: (a) genome-wide single SNPs only, using the 290,277 SNPs present across all datasets, (b) the VEPs only, i.e. the 5,359 VEPs encoded as 48,231 indicator variables, and (c) a 'combined' model of both single SNPs and VEPs together. The models were evaluated in cross-validation on the UK1 dataset, and the best models in terms of Area Under the Curve (AUC) were then taken forward for external validation in the other four datasets without further modification.

In UK1 cross-validation, the combined models led to an increase of ~1.6% in explained CD variance, from 32.6% to 34.2% (respective, AUC of 0.882 and 0.888) (**Table 3**). In external validation, the models based solely on VEPs had overall high predictive ability across all external validation datasets (AUC > 0.83), but slightly less than models based on single SNPs alone. The combined models yielded the highest externally validated AUC of all models, showing gains in AUC over single SNPs of +0.6% AUC in UK2 (Delong's 2-sided test $P = 0.0016$). In the IT dataset, average gains were higher at +1.2% AUC yet marginally significant ($P = 0.0527$), and the differences in FIN and NL were smaller (0.5% and 0.1%, respectively) and not significant.

Combining the UK1 and UK2 into a single dataset (N=7,786 unrelated individuals) and retraining the models in cross-validation showed similar trends with the best models being the combined models (**Table S3**). The combined models from the larger UK1+UK2 dataset also showed higher AUC in external validation than the UK1 only models when validating on the FIN and NL datasets (AUCs +1.3% and +1%; $P = 0.0007083$ and $P = 0.01113$, respectively); however, performance on the IT dataset did not differ significantly (−0.9% AUC, $P= 0.09568$).

Discussion

This study has shown the robust presence of epistasis in celiac disease. The epistatic SNP pairs strongly replicate across cohorts in terms of significance, ranking, and epistatic model. To our knowledge, this level of epistatic signal strength, number of epistatic pairs, and degree of replication has not been previously shown in a complex human disease. We also performed a large-scale empirical characterization of the epistatic models underlying the interactions in CD, with the majority of the VEPs approximately following the threshold model, and a smaller number following dominant-dominant, dominant-recessive, and recessive-recessive models. Further, these patterns were found to be strongly consistent across most of the datasets.

Despite observations that epistatic interactions between SNPs within a locus are enriched for batch effects and poorly clustered genotype clouds [39], the stringent quality control and extensive replication in this study indicate that these SNPs are largely *bona fide* epistatic pairs. A large number of candidate epistatic SNP pairs that did not achieve not achieving Bonferroni significance criteria for replication were still highly statistically associated with CD consistently across datasets, indicating that our estimates of the degree of epistasis in CD may be conservative.

For validated epistatic pairs (VEPs), we found that much of the strongest signal was >1Mb upstream of the well-known *HLA-DQA1* and *HLA-DQB1* risk loci and suggested a potentially important epistatic contribution from HLA class III genes. Indeed the strongest epistatic signal, which was independent of HLA risk haplotypes and other VEPs, was attributable to variants in *PRRC2A* and *GPANK1/C6orf47*. Given that individuals with celiac disease are at elevated risk of non-Hodgkin's lymphoma (NHL), it is intriguing that variants within *PRRC2A* are also associated with NHL [40]. However, for the top *PRRC2A* SNP for NHL (rs3132453) we did not observe a validated epistatic relationship nor linkage disequilibrium between rs3132453 and the epistatic *PRRC2A* SNP (rs2260000), which was low in the HapMap2 CEU ($r^2 = 0.05$). There is also evidence to suggest that women with celiac disease are at increased risk of early menopause [41,42]. A recent genetic association study of menopausal age identified a missense variant within *PRRC2A* (rs1046089) which was predicted as both structurally damaging for PRRC2A as well as an expression QTL for multiple genes [43]. In our study, the *PRRC2A* SNP rs1046089 showed strong epistasis with another proximal variant in *ABHD16A* as well as several other variants in the MHC (**Table S1**) despite low LD with the strongest *PRRC2A* epistatic variant ($r^2 = 0.27$). Overall, our findings indicate that, in addition to the known HLA risk haplotypes for CD, there is epistasis between HLA class III loci, which may have implications for CD co-morbidity.

The epistatic variants were further shown to increase CD variance explained, findings which were replicated in external datasets. Interestingly, models of only epistatic pairs explained nearly as much CD variance as additive models of single SNPs. This observation supports the existence of shared information between additive and epistatic effects [44] which may help to explain some of the controversy between these apparently different classes of effects. Further, our findings imply that determining causal genetic signals of CD, and perhaps other autoimmune/inflammatory diseases with substantial HLA-based effects, will be more difficult than previously thought.

These findings have implications for both the genetic architecture of celiac disease as well as the incorporation of epistasis into genetic models of complex disease. The limitations of the first generation GWAS approach to explain missing heritability has led to the development and application of more sophisticated approaches to resolve this problem, yet success has been elusive. Recent results suggest that rare variants add little to known heritability for a number of autoimmune diseases including celiac disease [26]. Epistasis may offer both additional explained heritability as well as new biology, as evidenced by our findings for HLA class III epistasis. The genetic models of CD generated in this work indicate that while epistatic pairs explain substantial disease variance, overall this variance is largely shared with that of additive effects. Combined epistatic and additive models likely constitute the best solution.

Methods

Quality control

A range of quality control measures were applied to all datasets to limit the impact of genotyping error. For all datasets, we removed non-autosomal SNPs, SNPs with MAF <1%, missingness >1% and those deviating from Hardy Weinberg Equilibrium in controls with $P < 5 \times 10^{-6}$. Samples were removed if data missingness was >1%. Cryptic relatedness was also stringently assessed by examining all pairs of samples using identity-by-descent in PLINK, and removing one of the samples if $\hat{\pi} > 0.05$. The cryptic relatedness filter removed 17 samples within the UK1 cohort that related to other UK1 samples, and 1208 samples from the UK2 cohort which were either related to other UK2 samples or UK1 samples. Dataset sizes in Table 1 are reported after the quality control steps above. Significant epistatic SNP pairs were further assessed by manually inspecting the genotyping cluster plots of both SNPs in the UK1 cohort. Intensity data for the other studies was not available thus epistatic pairs discovered in these datasets were not classified as robust and were not used in disease variance estimates, however the consistency of the statistics and epistatic model across independent datasets indicated that many likely represent bona fide epistasis. Cluster plot inspection removed 115 SNPs with poor genotyping assays.

Statistical tests for epistasis

Here, we summarize the Gain in Sensitivity and Specificity (GSS) test employed to detect epistasis. The test has been presented in detail in [34] and, currently, a web server implementing the GSS test is at http://bioinformatics.research.nicta.com.au/software/gwis/.

There is a long history of discussion around the definition of epistasis, or gene-gene interaction [45]. Here, an epistatic interaction is defined as a significant improvement of a SNP-pair in classifying cases from controls over what is possible using each SNP individually. There are two main differences between our approach and regression-based approaches for detecting epistasis [8,24]. First, our approach is “model-free”, as it makes no assumptions about the way in which genotypes combine to affect the phenotype [7,46], but considers all possible pairwise interactions for each pair, making it potentially more powerful to detect unknown epistatic forms, as empirical knowledge about epistasis in humans is currently lacking. Second, instead of measuring the deviation from additive effects (for example, using a likelihood ratio test), our approach focuses on the utility of the test in case/control classification, quantified using the receiver-operating characteristic (ROC) curve, and measuring the deviation in the curve from that induced by the additive model.

The main principle behind the GSS is quantification of the gain in predictive power afforded by a putative epistatic pair over and above the predictive power due to each of its constituent SNPs. The difference in predictive power is assessed in terms of the ROC curves induced by the pair and each of the SNPs, more precisely, by assessing the shape of those curves, rather than using the standard metric of the areas under them. The ROC curve is formed by considering each possible genotype (or pair of genotypes), and measuring the sensitivity (true positive rate, TPR) and specificity (1 – false positive rate, FPR) at that point, and ordering them in decreasing order by the ratio TPR/FPR; hence the curve is piecewise linear. Since the two ROC curves induced by the individual SNPs may intersect, we represent their combination by taking their convex hull, which is, simplistically, the best ROC curve that can be produced by a linear combination of the two individual SNPs, and represents a conservative estimate of the predictive power of the individual SNPs. This procedure is equivalent to fitting the optimal (Bayesian) classifier to the given data. The GSS then assigns a p-value to each point in the pair’s ROC curve, based on the probability of observing a combination of genotypes with a higher or equal TPR and a lower or equal FPR, under the null hypothesis that the true TPR and FPR reside below the convex hull. We employ a highly efficient minimax-based implementation, maximizing the probability for each point on the ROC curve (worst case scenario) against all points of the convex hull, and returning the minimum probability over all points [34]; this is done using an exact procedure rather than relying on approximations based on the normal distribution. Finally, the best p-value is assigned as the overall p-value for the pair, allowing the pairs to be ranked and corrected for multiple testing as is standard practice in GWAS. Those SNPs that are significant after multiple testing correction are deemed significant epistatic pairs.

Analogously to odds ratios used for analyses of single SNPs, we can estimate odds ratios for epistatic pairs based on the GSS statistic

$$OR_{GSS} = \frac{(\pi_{\{0,HR\}})(\pi_{\{1,LR\}})}{(\pi_{\{1,HR\}})(\pi_{\{0,LR\}})},$$

where $\pi_{[38]}$ denotes the proportion of samples with phenotype i, 0 for cases and 1 for controls, and carrying genotype combinations which are marked as j with HR (high risk) indicating genotypes which are associated by GSS with cases and LR (low risk) indicating genotypes which are associated with controls. By relying on the model-free GSS approach, this odds ratio can be seen as deriving the specific model maximizing the level of improvement over that of the individual SNPs in the pair.

Representation of the epistatic models

While the GSS approach is the basis for detecting epistatic pairs, the models it produces can be hard to visually interpret and categorize into broad groups. To simplify interpretation, we approximate the models for the statistically significant pairs found via GSS using two representations: balanced penetrance models and full penetrance models.

Balanced penetrance models

Following Li and Reich [23] we employ the penetrance, that is, the probability of disease given the genotype, estimated from the data for each of the nine genotype combinations as (number of cases with combination) / (number of individuals with combination). Representing the epistatic model in terms of penetrance allows us to clearly see which genotype combinations contribute more to disease risk (or conversely, may be protective).

One limitation of the penetrance is that it is typically considered in isolation of the disease background rate (the prevalence), which may be misleading when comparing penetrance levels across datasets with widely varying proportions of cases. For example, a penetrance of 50% for a given SNP would be considered very high in a dataset consisting of 1% cases and 99% controls, but no better than random guessing in datasets with 50%/50% cases and controls. Hence, we employ a standardization to ensure that the penetrance is comparable across datasets, termed *balanced sample penetrance*, and defined as

$$P_{balanced} = \frac{p_{1v}}{p_{1v} + p_{0v}},$$

where p_{iv} refers to the proportional frequency of genotype v in class i, where controls are 0 and cases are 1 (0=controls, 1 = cases). The balanced sample penetrance ranges between 0 and 1, where 0 means that the genotype only occurs in controls, 1 means that the genotype only occurs in cases and 0.5 means the genotype occurs evenly between the two classes. Balanced penetrance can be related to either standard penetrance or relative risk in the data via monotonic transformations. The definition is easily extended to the case of pair of SNPs. The only difference is the use of the $3 \times 3 = 9$ possible genotype combinations from each SNP-pair rather than the 3-value set of genotypes from an individual SNP.

Simplification to full penetrance models

The balanced-penetrance epistatic models provide fine-grained insight into the relative effects of each genotype combination. In addition, we employ a coarse-grain approach where these values are discretized into binary values (0/1), so called "fully penetrant" models, an approach analogous to that of Li and Reich [23]. These binary models forgo some detail but make it easier to categorize epistatic models into broad classes based on their patterns of interaction, such as the classic XOR pattern [8] or the threshold model [24]. Swapping major and minor alleles, and swapping the SNP ordering in the contingency table, can reduce the number of fully penetrant models. Unlike Li and Reich, we do not swap the high and low risk status, as we are interested in distinguishing between protective and deleterious combinations. Furthermore, Li and Reich also excluded models with all high or low risk genotypes. Such models can not exist within the set we are analyzing as they would show no association with disease. Li and Reich were able to show that there are only 51 possible fully penetrant disease models after accounting for symmetries. However, as we do not swap risk status, there will be 100 possible full-penetrance models that can appear within the analysis conducted here [22].

Given that some genotype combinations in certain SNP pairs are rare, there may be insufficient evidence to determine whether they have a substantial effect on disease risk. As such, we have used a simple heuristic for such entries, denoting all cells with a frequency below 1% in both cases and controls as 'low risk'. Experiments with this threshold revealed that altering this cutoff between 0% and 7% made little difference to the overall distribution of our models.

Independence of epistasis from known risk haplotypes

CD strongly depends on specific heterodimers, most notably HLA-DQ2.2, HLA-DQ2.5, and HLA-DQ8, which are in encoded by haplotypes involving the *HLA-DQA1* and *HLA-DQB1* genes, with close to 100% of individuals with CD being positive for one of these molecules. To statistically impute unphased HLA haplotype alleles, we utilized HIBAG [47]. To evaluate whether each VEP was independent of known CD risk haplotypes, we employed the likelihood ratio test, comparing two logistic regression models: (i) a logistic regression of the phenotype on the risk haplotypes (*DQA1*0201*, *DQA1*0501*, *DQA1*0505*, *DQA1*0301*, *DQB1*0201*, *DQB1*0202*, *DQB1*0301*, and *DQB*0302*) and (ii) a logistic regression including both the haplotypes and the VEP. The haplotypes were encoded as allele dosages [38] and the VEP was encoded as 8 binary indicator variables (allowing for an intercept term). We considered an FDR threshold < 0.05, equivalent to $p < 0.044$, as statistically significant, indicating that adding the VEP to the model increased goodness-of-fit over the haplotypes alone (4,744 of the 5,359 tests were FDR-significant).

Estimating the number of independent epistatic signals

To determine the number of independent signals coming from the VEPs, we have used an LD pruning based approach to filter out all SNP pairs that are in disequilibrium with each other. To ensure that the epistatic signals were not caused by haplotype effects, only SNP pairs that were deemed to be independent of HLA haplotypes were examined.

Estimating the linkage disequilibrium between pairs of SNP pairs is more difficult that estimating the LD within SNP pairs, largely because standard measures such as r^2 or D' cannot be used. Traditional measures of LD are designed for examining two binary loci whose frequencies can be reduced to a 2 × 2 table, whereas examining pairs of SNP pairs requires us to examine two multi-allelic loci, whose frequencies in this case can be summarized as a 4 × 4 table.

One widely used method is that of Hill's Q statistic [37], a multi-allelic extension of r^2. It is well known that the r^2 can be derived from the chi-squared test of association, since

$$r^2 = \frac{1}{2n}\chi^2$$

where n is the total number of samples and χ^2 is the chi-squared statistic over the haplotypes formed by the two SNPs. Motivated by this relationship, the Q statistic can be expressed as

$$Q = \frac{1}{2n}\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{1}{2n}\chi^2$$

where O_{ij} and E_{ij} are the observed and expected haplotype counts when there are k and l alleles respectively at the two loci. In the case of examining LD between pairs of SNP pairs $k=l=4$. As with a standard r^2 test, Q ranges between 0 and 1 where 0 indicates no association and 1 indicates perfect correlation between the two loci. Phase information was inferred using SHAPEIT [48], and LD was then computed directly on control samples only. While it is somewhat arbitrary what threshold constitutes independence, given the direct analogy between r^2 and Q we utilized a more conservative threshold for Q ($Q \leq 0.3$) than that commonly used for LD-pruning and tagging procedures with r^2 ($r^2 \leq 0.5$), for example in the PLINK tool. It could be argued that this conservativeness may filter slightly less independent but still informative epistatic signals, thus for the predictive models discussed below, all VEPs were initially allowed to enter the model.

The predictive models

We employed a sparse support vector machine (SVM) implemented in SparSNP [49]. This is a multivariable linear model where the degree of sparsity (number of variables being assigned a non-zero weight) is tuned via penalization. The model is induced by minimizing the L1-penalized squared hinge loss

$$(\beta^*, \beta_0^*) = \arg\min_{\beta,\beta_0} \frac{1}{2N}\sum_{i=1}^{N}\max\{0, 1-y_i(x_i^T\beta+\beta_0)\}^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where β and β_0 are the model weights and the intercept, respectively, N is the number of samples, p is the number of variables (SNPs and/or encoded pairs), x_i is the ith vector of p variables (genotypes and/or encoded pairs), y_i is the ith case/control status $\{+1, -1\}$, and $\lambda \geq 0$ is the L1 penalty. To find the optimal penalty, we used a grid of 100 penalty values within 10 replications of 10-fold cross-validation, and found the model/s that maximized the average area under the receiver-operating characteristic curve (AUC). For models based on single SNPs, we used minor allele dosage $\{0, 1, 2\}$ encoding of the genotypes. For models based on SNP pairs, the standard dosage model is not applicable; hence, we transformed the variable representing each pair (encoded by integers 1 to 9) to 9 indicator variables using the Python library scikit-learn [50], using a consistent encoding scheme across all datasets. The indicator variables were then analyzed in the same way as single SNPs. Results were analyzed in R [51] with the packages ROCR [52] and pROC [53], and plotted using the ggplot2 package [54].

Evaluation of predictive ability and explained disease variance

To maximize the number of SNPs available for analysis, we imputed SNPs in the UK2, FIN, NL, and IT datasets to match those that were in the UK1 dataset but not in former, using IMPUTE v2.3.0 [55]. Post QC this left 290,277 SNPs common to all five datasets. Together with 9×5,359 pairs=48,231 indicator variables, this led to a total of 338,508 variables in the combined singles+pairs dataset. Models trained in cross-validation on the UK1 dataset were then applied without any further tuning to the four other datasets, and the external-validation AUC for these models was then estimated within the validation datasets. To derive the proportion of phenotypic variance explained by the model (on the liability scale), we used the method of Wray et al. [56], assuming a population prevalence of 1%.

Acknowledgements

MI was supported by a Career Development Fellowship co-funded by the NHMRC and Australian Heart Foundation. MI and GA were also supported by University of Melbourne funding. BG, EK, QW, DR, FS, IH and AK were supported by National ICT Australia (NICTA). NICTA is funded by the Australian Government's Department of Communications, Information Technology and the Arts, the Australian Research Council through Backing Australia's Ability, and the ICT Centre of Excellence programs.

We thank the investigators of the van Heel et al., 2007 and Dubois et al., 2010 papers (David van Heel and Cisca Wijmenga) for providing the celiac disease data. We thank Karen A. Hunt (QMUL) for performing cluster plot inspection on the UK1 data. We also thank Rami Mukhtar for useful technical advice regarding implementation of algorithms used here and Andrew Kowalczyk and Leon Gor for assistance with development of software utilized for this work. We also thank Armita Zarnegar for assistance with processing of data and John Markham, Justin Bedo and Geoff Macintyre for insightful discussions and comments.

Tables

Table 1: Datasets

			Celiac cases		**Controls**		
		SNPs[a]	**Samples**[a]	**Platform**[b]	**Samples**[a]	**Platform**[b]	**Ref**
UK1	UK	301,546	763	Illumina Hap300v1-1	1420	Illumina Hap550	(van Heel, et al., 2007)
UK2	UK	515,413	1826	Illumina670-Quad	3777	Illumina 1.2M-Duo	(Dubois, et al., 2010)
FIN	Finland	513,952	647	Illumina670-Quad	1829	Illumina 610-Quad	(Dubois, et al., 2010)
NL	Netherlands	515,169	803	Illumina670-Quad	846	Illumina 670-Quad	(Dubois, et al., 2010)
IT	Italy	515,641	497	Illumina670-Quad	543	Illumina 670-Quad	(Dubois, et al., 2010)
Overlapping SNPs		290,277					

a. The number of samples/SNPs is reported after quality control procedures were applied.
b. All platforms contain a common set of Hap300 markers; the Hap550 and 610-Quad contain a common set of Hap550 markers.

Table 2: Independent epistatic signals detected in UK1

			UK1$_{univariate}$		UK1			UK2		FIN		NL		IT	
SNP	Chr	Position (bp)^f	MAF^c	X^2	LD^d	GSS^a	OR^b	GSS^a	OR^b	GSS^a	OR^b	GSS^a	OR^b	GSS^a	OR^b
rs2260000	6	31701455	0.28	40.9	0.59	58.3	14.2	108.4	10.8	95.5	20.6	27.0	10.1	12.6	6.77
rs805262	6	31736712	0.47	24.7											
rs2647050	6	32777745	0.32	16.3	0.22	28.1	7.4	55.8	6.0	7.2	4.3	16.3	4.3	12.9	13.8
rs2856705	6	32778934	0.13	14.7											
rs29232	6	29719410	0.33	7.70	0.33	25.8	4.2	38.8	3.6	32.5	6.6	17.8	4.7	3.9	7.44
rs7776082	6	29775252	0.50	14.7											
rs9268542	6	32492699	0.32	31.9	0.00	22.4	5.6	39.8	5.2	7.9	4.8	9.7	4.5	5.1	7.22
rs2856997	6	32889754	0.31	37.5											
rs1062470	6	31192414	0.29	12.5	0.10	18.7	4.8	20.6	3.5	16.0	5.4	5.8	3.6	0.8	2.1
rs3130712	6	31317489	0.33	17.5											
rs3095352	6	30913900	0.38	16.7	0.24	17.6	5.2	21.5	3.1	22.9	6.3	7.7	4.1	2.8	2.28
rs2250264	6	30929166	0.21	13.2											
rs3130931	6	31242867	0.26	25.7	0.03	17.2	4.3	25.6	3.8	6.8	3.8	9.3	3.6	1.8	1.83
rs3828903	6	31572718	0.27	21.0											
rs2070600	6	32259421	0.05	15.1	0.03	16.6	11.0	24.8	6.0	5.2	6.0	3.2	5.4	0.7	2.35
rs3129871	6	32514320	0.31	18.4											
rs6456785	6	27498378	0.50	7.40	0.41	13.8	2.8	20.4	2.5	11.9	2.8	10.0	2.5	4.6	2.25
rs6918131	6	27588896	0.26	4.40											
rs3948793	6	32867426	0.42	37.6	0.01	13.4	4.7	17.8	3.6	3.5	4.9	4.5	3.8	2.2	2.65
rs2854028	6	33287667	0.21	15.3											
rs3129274	6	33202847	0.37	16.5	0.72	13.0	4.2	37.8	4.3	37.1	8.3	14.5	5.2	8.3	3.85
rs213212	6	33293896	0.27	0.30											
rs3117098	6	32466491	0.24	14.5	0.09	12.6	5.3	25.7	5.6	26.5	10.8	14.4	8.2	6.4	3.81
rs9275390	6	32777134	0.20	22.6											
rs2256965	6	31663109	0.37	13.0	0.02	12.5	4.5	17.1	3.5	2.0	4.1	4.1	3.5	1.1	1.36
rs3830041	6	32299317	0.06	12.3											
rs12660382	6	31551302	0.13	14.6	0.03	12.5	3.3	28.5	Inf	7.4	2.6	8.9	2.9	4.4	3.25
rs2071596	6	31614670	0.14	18.3											
rs3871466	6	31091662	0.11	19.4	0.01	12.4	3.6	19.6	2.8	5.5	2.4	3.0	Inf	2.0	2.08
rs2269425	6	32231617	0.12	15.3											
rs2451741	6	26737383	0.40	6.20	0.41	12.4	2.5	14.3	2.2	5.5	2.2	6.7	2.2	1.5	1.68
rs2494711	6	26757400	0.33	4.60											
rs2535319	6	30822458	0.46	27.5	0.01	12.2	4.1	16.7	3.4	6.5	4.3	5.5	4.9	2.1	4.3
rs241437	6	32905662	0.45	25.7											
rs6900224	6	25545965	0.27	2.30	0.42	12.2	2.6	13.4	2.2	14.4	4.0	8.6	2.8	3.2	2.35
rs12525342	6	25568657	0.18	6.20											
rs241424	6	32912912	0.49	37.3	0.19	12.1	4.7	13.7	4.0	9.1	5.8	5.2	4.7	5.3	3.87
rs2071543	6	32919607	0.12	10.7											
rs2071556	6	33012579	0.39	13.4	0.27	12.0	3.6	26.3	3.1	7.4	3.9	5.7	3.1	1.8	2.48
rs2854028	6	33287667	0.21	15.3											

a. GSS indicates the –log$_{10}$(p-value) of improvement of the pair over each of the SNPs involved measured by the GSS filter described further in the Methods section
b. Odds Ratios are calculated directly from the GSS rather than via logistic regression, discussed further in Methods.
c. Minor Allele Frequency measured in the Control samples in the UK1 cohort
d. r^2 between SNPs constituting the epistatic pair.
e. Each signal represents the strongest of any pairs that show an $r^2 > 0.7$ with both of the SNPs in the pair
f. SNP positions were extracted from build 36
g. X^2 indicates log$_{10}$(p-value) for the standard χ^2 test of association (χ^2 statistics with 2 degrees of freedom).

Table 3: Disease variance explained by models with additive and epistatic genetic effects

		Single SNPs		Combined (single SNPs + pairs)		Validated epistatic pairs	
		Variance explained	AUC (95% CI)	Variance explained	AUC (95% CI)	Variance explained	AUC (95% CI)
Cross validation	UK1	0.326	0.882 [0.880, 0.883]	0.342	0.888 [0.886, 0.890]	0.337	0.886 [0.885, 0.887]
External validation	UK2	0.269	0.855 [0.844, 0.865]	0.282	0.861 [0.851, 0.872]	0.250	0.845 [0.834, 0.856]
	Finn	0.324	0.880 [0.865, 0.896]	0.334	0.885 [0.870, 0.899]	0.293	0.867 [0.851, 0.882]
	IT	0.265	0.853 [0.830, 0.876]	0.290	0.865 [0.843, 0.888]	0.237	0.837 [0.813, 0.862]
	NL	0.274	0.858 [0.839, 0.876]	0.277	0.859 [0.841, 0.877]	0.255	0.847 [0.829, 0.866]

Predictive power and disease variance explained by single SNPs and VEPs in cross-validation and in external validation, using SparSNP models. Models were optimized on the UK1 dataset (n=2183 samples) in cross-validation (290K SNPs), and tested without modification on the other datasets. The proportion of disease variance explained (on the liability scale) assumes a population prevalence of 1%. The 95% CI for AUC in UK1 was computed over the 10x10 cross-validation, and in external validation was computed using DeLong's method (R package pROC). Two-sided DeLong significance tests for AUC of single SNPs+pairs difference from AUC of single SNPs: UK2 P=0.001651, FIN P=0.2743, IT P=0.05271, NL P=0.695.

Figure Legends

Figure 1: Epistatic interactions within the extended MHC region

SNP pairs within 30KB of each other are shown as a single point on each heatmap. The colour of each point represents the most significant -log$_{10}$(P-value) returned by the GSS statistic for SNPs pairs within each point. The -log$_{10}$(P-value) is capped at 30 to increase contrast of lower values.

Figure 2: Replication of epistatic pairs and corresponding epistatic models between datasets and populations

Panel **(a)** shows the overlap of significant epistatic pairs as a percentage between UK1 and remaining cohorts in order of decreasing GSS significance. Vertical dotted lines indicate the Bonferroni-adjusted significance for each study. Panel **(b)** shows the occurrence of genotype combinations for the top pair from UK1. Colouring of cells provides an indication of the epistatic model occurring in each cohort (**Methods**).

Figure 3: Variation in epistatic models within and between populations

Distribution of epistatic models for VEPs in different studies as increasingly less significant SNP pairs are examined. Different colours represent a different subset of epistatic models. The “other” group represents the remaining set of models. Models have been simplified using the rules provided in [23].

Figures

Figure 1: Epistatic interactions within the extended MHC region

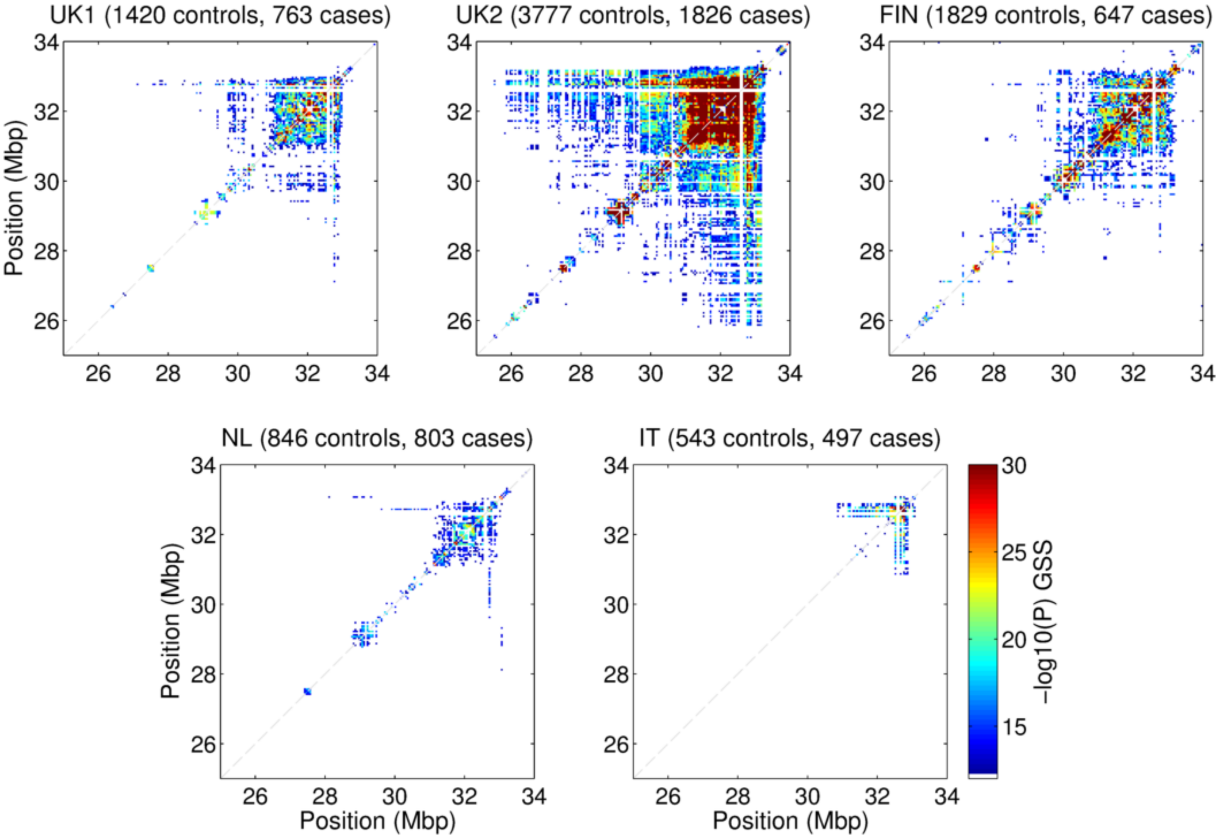


Figure 2: Replication of epistatic pairs and corresponding epistatic models between datasets and populations

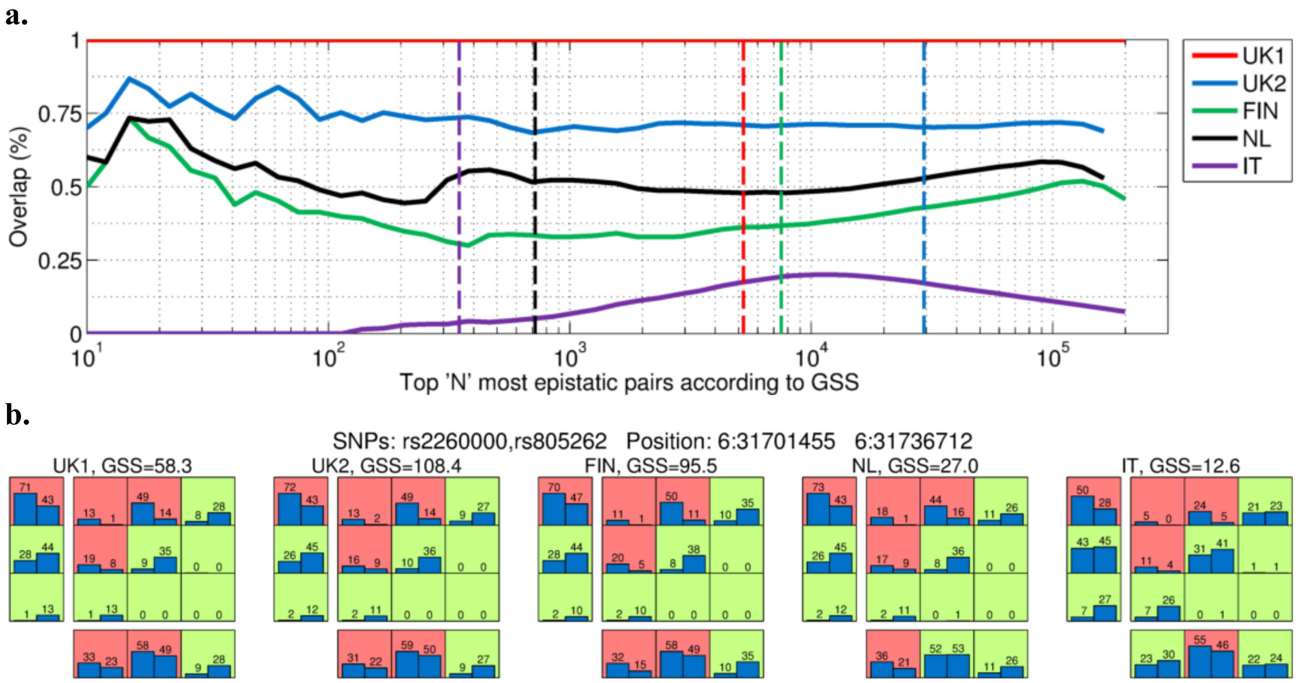
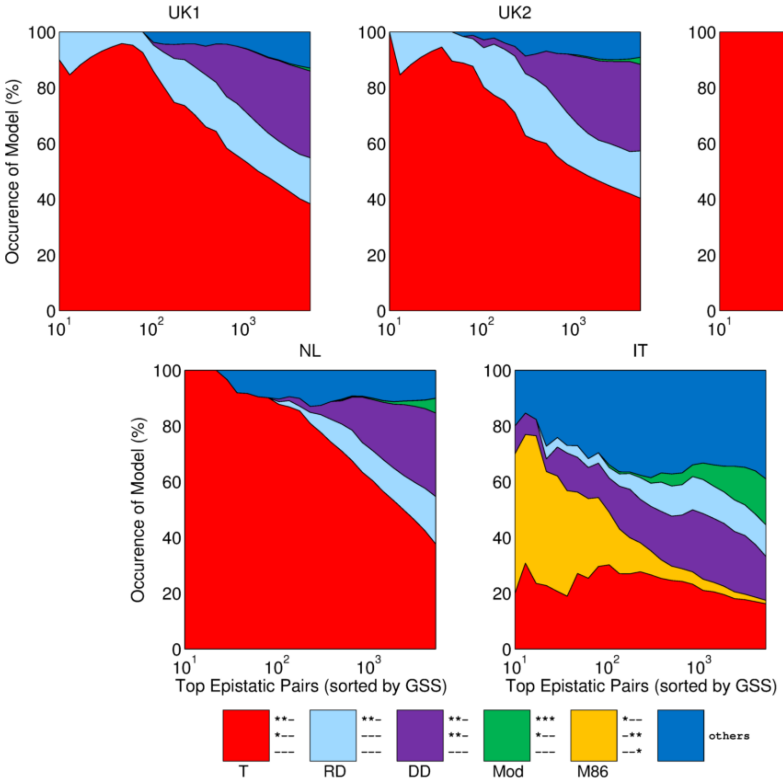


Figure 3: Variation in epistatic models within and between populations



References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
2. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446-450.
3. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, et al. (2013) Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 498: 232-235.
4. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stancakova A, et al. (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45: 197-201.
5. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, et al. (2012) A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488: 96-99.
6. Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, et al. (2013) Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature 497: 517-520.
7. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11: 2463-2468.
8. Wan X, Yang C, Yang Q, Xue H, Fan X, et al. (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 87: 325-340.
9. Hu X, Liu Q, Zhang Z, Li Z, Wang S, et al. (2010) SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. Cell Res 20: 854-857.
10. Hemani G, Theocharidis A, Wei W, Haley C (2011) EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics 27: 1462-1465.
11. Prabhu S, Pe'er I (2012) Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. Genome Res 22: 2230-2240.
12. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, et al. (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 241: 252-261.
13. Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39: 1167-1173.
14. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5: 618-625.
15. Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF (2013) Genetic incompatibilities are widespread within species. Nature.
16. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. Nature 490: 535-538.
17. Genetic Analysis of Psoriasis C, the Wellcome Trust Case Control C, Strange A, Capon F, Spencer CC, et al. (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet 42: 985-990.
18. Lincoln MR, Ramagopalan SV, Chao MJ, Herrera BM, Deluca GC, et al. (2009) Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility. Proc Natl Acad Sci U S A 106: 7542-7547.
19. Kirino Y, Bertsias G, Ishigatsubo Y, Mizuki N, Tugal-Tutkun I, et al. (2013) Genome-wide association analysis identifies new susceptibility loci for Behcet's disease and epistasis between HLA-B [ast] 51 and ERAP1. Nature genetics 45: 202-207.
20. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature genetics 41: 703-707.
21. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, et al. (2011) Interaction between ERAP1

and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat Genet 43: 761-767.
22. Hemani G, Theocharidis A, Wei W, Haley C (2011) EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics 27: 1462-1465.
23. Li W, Reich J (2000) A complete enumeration and classification of two-locus disease models. Hum Hered 50: 334-349.
24. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37: 413-417.
25. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4: e1000008.
26. van Heel DA, Hunt K, Greco L, Wijmenga C (2005) Genetics in coeliac disease. Best Pract Res Clin Gastroenterol 19: 323-339.
27. Trynka G, Hunt Ka, Bockett Na, Romanos J, Mistry V, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics 43: 1193-1201.
28. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nature Genetics 39: 827-829.
29. Dubois PCA, Trynka G, Franke L, Hunt Ka, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics 42: 295-302.
30. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 40: 395-402.
31. Romanos J, Rosen A, Kumar V, Trynka G, Franke L, et al. (2014) Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. Gut 63: 415-422.
32. Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, et al. (2014) Accurate and robust genomic prediction of celiac disease using statistical learning. PLoS Genet 10: e1004137.
33. Mitchison NA, Rose AM (2011) Epistasis: The key to understanding immunological disease? European journal of immunology 41: 2152-2154.
34. Goudey B, Rawlinson D, Wang Q, Shi F, Ferra H, et al. (2013) GWIS--model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. BMC Genomics 14 Suppl 3: S10.
35. Neuman RJ, Rice JP (1992) Two-locus models of disease. Genet Epidemiol 9: 347-365.
36. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98-101.
37. Hill WG (1975) Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor Popul Biol 8: 117-126.
38. Abraham G, Kowalczyk A, Zobel J, Inouye M (2013) Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genetic Epidemiology 37: 184-195.
39. Lee SH, Nyholt DR, Macgregor S, Henders AK, Zondervan KT, et al. (2010) A simple and fast two-locus quality control test to detect false positives due to batch effects in genome-wide association studies. Genet Epidemiol 34: 854-862.
40. Nieters A, Conde L, Slager SL, Brooks-Wilson A, Morton L, et al. (2012) PRRC2A and BCL2L11 gene variants influence risk of non-Hodgkin lymphoma: results from the InterLymph consortium. Blood.
41. Sher KS, Mayberry JF (1994) Female Fertility, Obstetric and Gynaecological History in Coeliac Disease. Digestion 55: 243-246.
42. Sher KS, Mayberry JF (1996) Female fertility, obstetric and gynaecological history in coeliac disease: a case control study. Acta Pædiatrica 85: 76-77.
43. Stolk L, Perry JRB, Chasman DI, He C, Mangino M, et al. (2012) Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. Nat Genet 44: 260-268.
44. Mackay TF (2014) Epistasis and quantitative traits: using model organisms to study gene-gene

interactions. Nat Rev Genet 15: 22-33.

45. Phillips PC (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9: 855-867.
46. Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Konig IR, et al. (2009) Machine learning in genome-wide association studies. Genet Epidemiol 33 Suppl 1: S51-57.
47. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, et al. (2014) HIBAG--HLA genotype imputation with attribute bagging. Pharmacogenomics J 14: 192-200.
48. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9: 179-181.
49. Abraham G, Kowalczyk A, Zobel J, Inouye M (2012) SparSNP: Fast and memory-efficient analysis of all SNPs for phenotype prediction. BMC Bioinformatics 13: 88.
50. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12: 2825-2830.
51. R Core Team (2012) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
52. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940-3941.
53. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12: 77.
54. Wickham H (2009) ggplot2: elegant graphics for data analysis. New York: Springer.
55. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.
56. Wray NR, Yang J, Goddard ME, Visscher PM (2010) The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genetics 6: e1000864.