A method for identifying genetic heterogeneity within phenotypically-defined disease subgroups

James Liley[1,2], John A Todd[1], and Chris Wallace[1,2,3]

[1] *JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, NIHR Cambridge Biomedical Research Centre, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK*

[2] *Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 0SP, UK*

[3] *MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, CB2 0SR, Cambridge, UK*

Abstract

Many common diseases show wide phenotypic variation. We present a statistical method for determining whether phenotypically defined subgroups of disease cases represent different genetic architectures, in which disease-associated variants have different effect sizes in the two subgroups. Our method models the genome-wide distributions of genetic association statistics with mixture Gaussians. We apply a global test without requiring explicit identification of disease-associated variants, thus maximising power in comparison to a standard variant by variant subgroup analysis. Where evidence for genetic subgrouping is found, we present methods for post-hoc identification of the contributing genetic variants.

We demonstrate the method on a range of simulated and test datasets where expected results are already known. We investigate subgroups of type 1 diabetes (T1D) cases defined by autoantibody positivity, establishing evidence for differential genetic architecture with thyroid peroxidase antibody positivity, driven generally by variants in known T1D associated regions.

Introduction

Analysis of genetic data in human disease typically uses a binary disease model of cases and controls. However, many common human diseases show extensive clinical and phenotypic diversity which may represent multiple causative pathophysiological processes. Because therapeutic approaches often target disease-causative pathways, understanding this phenotypic complexity is valuable for further development of treatment, and the progression towards personalised medicine. Indeed, identification of patient subgroups characterised by different clinical features can aid directed therapy [1] and accounting for phenotypic substructures can improve ability to detect causative variants by refining phenotypes into subgroups in which causative variants have larger effect sizes [2].

Such subgroups may arise from environmental effects, reflect population variation in non-disease related anatomy or physiology, correspond to partitions of the population in which disease heritability differs, or represent different causative pathological processes. Our method tests whether there exist a subset of disease-associated SNPs which have different effect sizes in case subgroups, determining whether heterogeneity corresponds to differential genetic pathology.

Our test is for a stronger assertion than the question of whether subgroups of a disease group exhibit any genetic differences at all, as these may be entirely disease-independent: for example, although there will be systematic genetic differences between Asian and European patient cohorts with type 1 diabetes (T1D), these differences will not generally relate to the pathogenesis of disease.

Rather than attempting to analyse SNPs individually for differences between subgroups, a task for which GWAS are typically underpowered, we model allelic differences across all SNPs using mixture multivariate normal models. This can give insight into the structure of the genetic basis for disease. Given evidence that there exists some subset of SNPs that both differentiate controls and cases and differentiate subgroups, we can then reassess test statistics to search for single-SNP effects.

Results

Summary of proposed method

We jointly consider allelic differences between the combined case group and controls, and allelic differences between case subgroups independent of controls. Specifically, we establish whether the data support a hypothesis (H_1) that a subset of SNPs associated with case-control status have different underlying effect sizes (and hence underlying allele frequencies) in case subgroups. This assumption has been used previously for genetic discovery [3].

H_1 encompasses several potential underlying mechanisms of heterogeneity. A set of SNPs may be associated with one case subgroup but not the other; the same set of SNPs may have different relative effect sizes in subgroups, or heritability may differ between subgroups. These scenarios are discussed in supplementary note 1.

Our overall protocol is to fit two bivariate Gaussian mixture models, corresponding to null and alternative hypotheses, to summary statistics (Z scores) derived from SNP data. We assume a group of controls and two non-intersecting case subgroups, and jointly consider allelic differences between the combined case group and controls, and allelic differences between case subgroups independent of controls (figure 1). Heterogeneity in cases can also be characterised by a quantitative trait, rather than explicit subgroups.

For a given SNP we denote by μ_1, μ_2, μ_{12} and μ_c the population minor allele frequencies for each of the two case subgroups, the whole case group and the control group respectively, and P_d, P_a GWAS p-values for comparisons of allelic frequency between case subgroups and between cases and controls, under the null hypotheses $\mu_1 = \mu_2$ and $\mu_{12} = \mu_c$ respectively (or similarly for quantitative heterogeneity). We then derive absolute Z scores $|Z_d|$ and $|Z_a|$ from these p-values (see figure 1). We consider the values $|Z_d|, |Z_a|$ as absolute values of observations of random variables (Z_d, Z_a) which are samples from a mixture of three bivariate Gaussians. Further details are given in supplementary note 2.

We consider each SNP to fall into one of three categories, with each category corresponding to a different joint distribution of Z_d, Z_a:

1. SNPs which do not differentiate subgroups and are not associated with the phenotype as a whole ($\mu_c = \mu_1 = \mu_2$)

2. SNPs which are associated with the phenotype as a whole but which are not differentially associated with the subgroups ($\mu_c \neq \mu_{12}$; $\mu_1 = \mu_2 = \mu_{12}$)

3. SNPs which have different population allele frequencies in subgroups, and may or may not be associated with the phenotype as a whole ($\mu_1 \neq \mu_2$)

If the SNPs in category 3 are not associated with the disease as a whole (null hypothesis, H_0), we expect Z_d, Z_a to be independent and the variance of Z_a to be 1. If SNPs in category 3 are also associated with the disease as a whole (alternative hypothesis, H_1), the joint distribution of (Z_d, Z_a) will have both marginal variances greater than 1, and Z_a, Z_d may co-vary. Our test is therefore focussed on the form of the joint distribution of (Z_d, Z_a) in category 3. Importantly, we allow that the correlation between Z_d and Z_a may be simultaneously positive at some SNPs and negative at others. This allows for a subset of SNPs to specifically alter risk of one subgroup, and another subset to alter risk for the other subgroup. To accommodate this, we only consider absolute Z scores and model the distribution of SNPs in category 3 with two mirror-image bivariate Gaussians.

Amongst SNPs with the same frequency in disease subgroups (categories 1 and 2), Z_a and Z_d are independent and the expected standard deviation of Z_d is 1. We therefore model the overall joint distribution of (Z_d, Z_a) as a Gaussian mixture in which the *pdf* of each observation (Z_d, Z_a) is given by

$$
\begin{aligned}
PDF_{Z_d,Z_a|\Theta}(d,a) =& \pi_1 N_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)}(d,a) && \text{(category 1)} \\
&+\pi_2 N_{\left(\begin{smallmatrix} 1 & 0 \\ 0 & \sigma_2^2 \end{smallmatrix}\right)}(d,a) && \text{(category 2)} \\
&+\pi_3 \left(\frac{1}{2} N_{\left(\begin{smallmatrix} \tau^2 & \rho \\ \rho & \sigma_3^2 \end{smallmatrix}\right)}(d,a) + \frac{1}{2} N_{\left(\begin{smallmatrix} \tau^2 & -\rho \\ -\rho & \sigma_3^2 \end{smallmatrix}\right)}(d,a) \right) && \text{(category 3)} \qquad (1)
\end{aligned}
$$

where $N_\Sigma(d,a)$ denotes the density of the bivariate normal *pdf* centered at $\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$ with covariance matrix Σ at (d,a). Θ is the vector of values $(\pi_1, \pi_2, \tau, \sigma_2, \sigma_3, \rho)$. Under H_0, we have $\rho = 0$ and $\sigma_3 = 1$. The values (π_1, π_2, π_3) represent the proportion of SNPs in each category, with $\Sigma \pi_i = 1$ (see table 1). Patterns of (Z_d, Z_a) for different parameter values are shown in supplementary table 1.

We use the product of values of the above *pdf* for a set of observed Z_d, Z_a as an objective function ('pseudo-likelihood', PL) to estimate the values of parameters. This is not a true likelihood as observations are dependent due to linkage disequilibrium (LD), although because we minimise the degree of LD between SNPs using the LDAK method [4], the PL is similar to a true likelihood.

	Model	Interpretation
π_1	H_0/H_1	Proportion of SNPs **not** associated with case/control status and **not** associated with subgroup status (category 1)
π_2	H_0/H_1	Proportion of SNPs associated with case/control status but **not** subgroup status (category 2)
π_3	H_0/H_1	Proportion of SNPs associated with subgroup status (category 3)
τ	H_0/H_1	Standard deviation of observed Z_d scores (effect sizes for subgroup status) in category 3
σ_2	H_0/H_1	Standard deviation of observed Z_a scores (effect sizes for case/control status) in category 2
σ_3	H_1 only	Standard deviation of observed Z_a scores (effect sizes for case/control status) in category 3
ρ	H_1 only	'Absolute covariance' between Z_d scores (effect sizes for subgroup status) and Z_a scores (effect sizes for case/control status) in category 3

Table 1: Interpretation of parameter values in the fitted model. Parameters τ, σ_2 and σ_3 are dependent on sample sizes, but can be converted to sample-size independent forms (see supplementary note, section 3.3)

Model fitting and significance testing

We fit parameters π_1, π_2, π_3 $(= 1 - \pi_1 - \pi_2)$, σ_2, σ_3, τ and ρ under H_1 and H_0. Under H_0, $(\rho, \sigma_3) = (0, 1)$.

We then compare the fit of the two models using the log-ratio of PLs, giving an unadjusted pseudo-likelihood ratio (uPLR). We subtract a term depending only on Z_a to minimise the influence of the Z_a score distribution, and add a term $log(\pi_1\pi_2\pi_3)$ to ensure the model is identifiable [5]. We term the resultant test statistic the pseudo-likelihood ratio

(PLR). The distribution of the PLR is minorised by a distribution of the form:

$$PLR|H_0 \sim \begin{cases} \gamma\chi^2_1 & \text{prob} = \kappa \\ \gamma\chi^2_2 & \text{prob} = 1-\kappa \end{cases}. \tag{2}$$

The value γ arises from the weighting derived from the LDAK procedure causing a scale change in the observed PLR. The mixing parameter κ corresponds to the probability that $\rho = 0$, (approximately $\frac{1}{2}$).

We estimate γ and κ by sampling random subgroups of the case group. Such subgroups only cover the subspace of H_0 with $\tau = 1$ (no systematic allelic differences between subgroups), causing the asymptotic approximation of PLR by equation 2 to be poor. We thus estimate γ and κ from the distribution of a similar alternative test statistic, the cPLR (see methods section and supplementary note, section 2.5.1), which is well-behaved even when $\tau \approx 1$ and which majorises the distribution of PLR.

A natural next step is to search for the specific variants contributing to the PLR. An effective test statistic for testing subgroup differentiation for single SNPs is the Bayesian conditional false discovery rate (cFDR) [6, 7] applied to Z_d scores 'conditioned' on Z_a scores. However, this statistic alone cannot capture all the means by which the joint distribution of (Z_a, Z_d) can deviate from H_0, and we also propose three other test statistics, each with different advantages, and compare their performance (supplementary note, section 5.1).

Power calculations, simulations, and validation of method

We tested our method by application to a range of datasets, using simulated and resampled GWAS data. First, to confirm appropriate control of type 1 error rates across H_0, we simulated genotypes of case and control groups under H_0 for a set of 5×10^5 autosomal SNPs in linkage equilibrium (supplementary note 3). Quantiles of the empirical PLR distribution were smaller than those for the empirical cPLR distribution and the asymptotic mixture-χ^2, indicating that the test is conservative when $\tau > 1$ (estimated type 1 error rate 0.048, 95% CI 0.039-0.059) and when $\tau \approx 1$ (estimated type 1 error rate 0.033, 95% CI 0.022-0.045) as expected; see figure 2. The distribution of cPLR closely approximated the asymptotic mixture-χ^2 distribution across all values of τ (supplementary note, section 3.1).

We then established the suitability of the test when SNPs are in LD and when there exist genetic differences between subgroups that are independent of disease status overall. First, we used a dataset of controls and autoimmune thyroid disease (ATD) cases and repeatedly choose subgroups such that several SNPs had large allelic differences between subgroups. We found good FDR control at all cutoffs (supplementary note, figure 3.2) and the overall type 1 error rate at $\alpha = 0.05$ was 0.041 (95% CI 0.034-0.050). Second, we analysed a dataset of T1D cases with subgroups defined by geographical origin. Within the UK, there is clear genetic diversity associated with region [9]. As expected, Z_d scores for geographic subgroups showed inflation compared to for random subgroups (supplementary figure 1). None of the derived test statistics reached significance at a Bonferroni-corrected $p < 0.05$ threshold (min. corrected p value > 0.8, supplementary figure 2).

To examine the power of our method, we used published GWAS data from the Wellcome Trust Case Control Consortium [10] comprising 1994 cases of Type 1 diabetes (T1D), 1903 cases of rheumatoid arthritis (RA), 1922 cases of type 2 diabetes (T2D) and 2953 common controls. We established that our test could differentiate between any pair of diseases, considered as subgroups of a general disease case group (all $< 1 \times 10^{-8}$, table 2).

T1D and RA have overlap in genetic basis [10, 11, 7], as well as non-overlapping associated regions. T1D and T2D have less overlap [11] and T2D and RA less still. This was reflected in the fitted values (table 2, figure 3). The fitted values parametrizing category

		π_1	π_2	π_3	σ_2	σ_3	τ	ρ	p-val
T1D/RA	H_1	0.997	5.69×10^{-4}	2.06×10^{-3}	2.76	1.39	1.74	1.815	3.2×10^{-12}
	H_0	0.997	6.26×10^{-4}	2.48×10^{-3}	2.71	-	1.67	-	
T1D/T2D	H_1	0.573	0.426	9.63×10^{-4}	1.00	2.03	2.25	1.68	1.6×10^{-9}
	H_0	0.578	0.421	8.91×10^{-4}	1.00	-	2.21	-	
T2D/RA	H_1	0.573	0.426	8.71×10^{-4}	1.00	2.23	1.75	1.69	5.1×10^{-9}
	H_0	0.91	8.05×10^{-4}	0.0892	2.25	-	0.97	-	
GD/HT	H_1	0.506	0.487	0.007	1.12	2.90	1.65	2.61	2.2×10^{-15}
	H_0	0.493	0.079	0.428	1.68	-	1.03	-	

Table 2: Fitted parameter values for models of T1D/RA, T1D/T2D, T2D/RA, and GD/HT. H_1 is the null hypothesis (under which $\sigma_3 = 1$, $\rho = 0$) that SNPs differentiating the subgroups are not associated with the overall phenotype; H_1 is the alternative (full model). p values for pseudo-likelihood ratio tests are also shown.

2 in the full model for T1D/RA (π_2, σ_2) were consistent with a subset of SNPs associated with case/control status (T1D+RA vs control) but not differentiating T1D/RA. By contrast, the parametrization of category 2 for T1D/T2D and T2D/RA had marginal variance σ_2 approximately 1, suggesting that a subset of SNPs associated with case/control status but not with 'subgroup' status did not exist in these cases. The rejection of H_0 for the comparisons entails the existence of a set of SNPs associated both with case/control and subgroup status. The H_0 model does not allow such a set of SNPs, forcing the parametrisation of Z_d, Z_a scores for such SNPs to be 'squashed' into a category shape permitted under H_0, with one marginal variance being 1: either category 2 (as happens in T2D/RA since $\pi_2|H_0 \approx \pi_3|H_1$, $\sigma_2|H_0 \approx \sigma_3|H_1$ in T2D/RA) or category 3 (as in T1D/T2D, where $\pi_3|H_0 \approx \pi_3|H_1$, $\tau|H_0 \approx \tau|H_1$).

To determine the power of our test more generally, we showed that power depends on the number of SNPs in category 3 and on the underlying parameters of the true model, depending on the number of samples through the fitted model parameters (Supplementary Note 3.3). We therefore estimated the power of the test for varying numbers of SNPs in category 3 and for varying values of the parameters σ_3, τ, and ρ. (Figure 4; Supplementary Figure 3). As expected, power increases with an increasing number of SNPs in category 3, reflecting the proportion of SNPs which differentiate case subgroups and are associated with the phenotype as a whole. Power also increases with increasing τ, σ_3, and absolute correlation ($\rho/(\sigma_3\tau)$) as high values enable better distinction of SNPs in the second and third categories.

We explored the dependence of power on sample size by sub-sampling the WTCCC data for RA and T1D (figure 4) and compared the power of the PLR with the power to find any single SNP which differentiated the two diseases in several ways (see figure legend). Although the power of the PLR-based test was limited at reduced sample sizes, it remained consistently higher than the power to detect any single SNP which differentiated the two diseases. We then repeated the analysis removing the known T1D- and RA- associated SNP rs17696736. The power to detect a SNP with significant Z_d score (Bonferroni-corrected) amongst SNPs with GW-significant Z_a score dropped dramatically, though the power of PLR was only slightly reduced. This illustrated the robustness of the PLR test to inclusion or removal of single SNPs with large effect sizes, a property not shared by single-SNP approaches.

Estimating power requires an estimate of the underlying values of several parameters: the expected total number of SNPs in the pruned dataset with different population MAF in case subgroups, and the distribution of odds-ratios such SNPs between subgroups and between cases/controls. With sparse genome-wide cover, such as that in the WTCCC study, > 1250 cases per subgroup are necessary for 90% power (discounting MHC region). If SNPs with greater coverage for the disease of interest are used (such as the ImmunoChip for autoimmune diseases) values of π_3, σ_3 and τ are correspondingly higher, and around 500-700 cases per subgroup may be sufficient.

Application to autoimmune thyroid disease and type 1 diabetes

Autoimmune thyroid disease (ATD) takes two major forms: Graves' disease (GD; hyperthyroidism) and Hashimoto's Thyroiditis (HT; hypothyroidism). Differential genetics of these conditions have been investigated. Detection of individual variants with different effect sizes in GD and HT is limited by sample size (particularly HT); however, the $TSHR$ region shows evidence of differential effect [12]. T1D is relatively clinically homogenous with no major recognised subtypes, although heterogeneity arises between patients in levels of disease-associated autoantibodies, and disease course differs with age at diagnosis [3]. We analysed both of these diseases.

For ATD, we were able to confidently detect evidence for differential genetic bases for GD and HT ($p = 2.2\times10^{-15}$). Fitted values are shown in table 2. The distribution of cPLR statistics from random subgroups agreed well with the proposed mixture χ^2 (supplementary figure 4b).

For T1D, we considered four subgroupings defined by plasma levels of the T1D-associated autoantibodies thyroid peroxidase antibody (TPO-Ab, n=5780), insulinoma-associated antigen 2 antibody (IA2-Ab, n=3197), glutamate decarboxylase antibody (GAD-Ab, n=3208) and gastric parietal cell antibodies (PCA-Ab, n=2240). A previous GWAS study on autoantibody positivity in T1D identified only two non-MHC loci at genome-wide significance: 1q23/*FCRL3* with IA2-Ab and 9q34/*ABO* with PCA-Ab [3].

We tested each of the subgroupings retaining and excluding the MHC region. Fitted values for models with and without MHC are shown in supplementary table 2, and plots of Z_a and Z_d scores are shown in supplementary figure 5. Retaining the MHC region, we were able to confidently reject H_0 for subgroupings based on TPO-Ab, IA-2Ab and GAD-Ab (all p-values $< 1.0 \times 10^{-20}$). Although there was evidence that SNPs in the dataset were associated with PCA-Ab level ($\tau \approx 2.5$, null model), the improvement in fit in the full model was not significant, and we conclude that such SNPs determining PCA-Ab status are not in general T1D-associated. This can be seen by in the plot of Z_a against Z_d (supplementary figure 5) where SNPs with high Z_d values do not have higher than expected Z_a values.

With MHC removed, the subgrouping on TPO-Ab was significantly better-fit by the full model ($p = 1.5 \times 10^{-4}$). There was weaker evidence to reject H_0 for GAD-Ab ($p = 0.002$) and IA2-Ab ($p = 0.008$) (Bonferroni-corrected threshold at $\alpha < 0.05$: 0.006). Fitted values of τ in both the full and null models for GAD-Ab were ≈ 1, indicating absence of evidence for a category of non-MHC T1D-associated SNPs additionally associated with GAD-Ab positivity. Collectively, this indicates that differential genetic basis for T1D with GAD-Ab and IA2-Ab positivity is driven principally by the MHC region, and although PCA-Ab status is partially genetically determined, the set of causative variants is independent of T1D causative pathways.

The variation in genetic architecture of T1D with age is not fully understood, but previous studies have suggested larger observed effects at known loci in patients diagnosed at a younger age [13, 14, 15, 16]. We investigated whether these differences were indicative of widespread differences in variant effect sizes with age-at-diagnosis, possibly due to differential heritability (see supplementary note 1). We applied the method to T1D dataset with Z_d defined by age at diagnosis (quantitative trait). Fitted values are shown in supplementary table 3 and Z_a and Z_d scores in supplementary figure 6. The hypothesis H_0 could be rejected confidently when retaining or removing the MHC region (p values $< 1.0 \times 10^{-20}$ and 0.007 respectively). Signed Z_d and Z_a scores for age at diagnosis showed a visible negative correlation ($p = 0.002$) amongst Z_d and Z_a scores for disease-associated SNPs (r_g method 2, figure 5). This is consistent with a higher genetic liability with lower age at diagnosis.

Assessment of individual SNPs

Many SNPs which discriminated subgroups were in known disease-associated regions (Supplementary Tables 4, 5, and 6). In several cases, our method identified disease-associated SNPs which have reached genome-wide significance in subsequent larger studies but for which the Z_a score in the WTCCC study was not near significance. For example, the SNP rs3811019, in the *PTPN22* region, was identified as likely to discriminate T1D and T2D ($p = 3.046 \times 10^{-6}$; supplementary table 5), despite a p value of 3×10^{-4} for joint T1D/T2D association.

For GD and HT, SNPs near the known ATD-associated loci *PTPN22* (rs7554023), *CTLA4* (rs58716662), and *CEP128* (rs55957493) were identified as likely to be contributing to the difference (see supplementary table 7). The SNPs rs34244025 and rs34775390 are not known to be ATD-associated, but are in known loci for inflammatory bowel disease and ankylosing spondylitis, and our data suggest they may differentiate GD and HT (FDR 0.003).

We searched for non-MHC SNPs with differential effect sizes with TPOA positivity in T1D, the subgrouping of T1D for which we could most confidently reject H_0. Previous work [3] identified several loci potentially associated with TPO-Ab positivity by restricting attention to known T1D loci, enabling use of a larger dataset than was available to us. We list the top ten SNPs for each summary statistic for TPO-Ab positivity in supplementary table 8. Subgroup-differentiating SNPs included several near known T1D loci: *CTLA4* (rs7596727), *BACH2* (rs11755527), *RASGRP1* (rs16967120) and *UBASH3A* (rs2839511) [17]. These loci agreed with those found by Plagnol et al [3], but our analysis used only available genotype data, without external information on confirmed T1D loci. We were not able to replicate the same p-values due to reduced sample numbers.

Finally, we analysed non-MHC SNPs with varying effect sizes with age at diagnosis in T1D (supplementary table 9). This implicated SNPs in or near *CTLA4* (rs2352551), *IL2RA* (rs706781), and *IKZF3* (rs11078927).

Discussion

The problem we address is part of a wider aim of adapting GWAS to complex disease phenotypes. As the body of GWAS data grows the analysis of between-disease similarity and within-disease heterogeneity has led to substantial insight into shared and distinct disease pathology [6, 7, 2, 20, 21]. We seek in this paper to use genomic data to infer whether such disease subtypes exist. Our problem is related to the question of whether two different diseases share any genetic basis [18] but differs in that the implicit null hypothesis relates to genetic homogeneity between subgroups rather than genetic independence of separate diseases.

Our test strictly assesses whether a set of SNPs have different effect sizes in case subgroups. We interpret this as 'differential causative pathology', which encompasses several disease mechanisms, discussed in supplementary note 1. In some cases, if subgroups are defined on the basis of the presence or absence of a known disease risk factor, the heritability of the disease will differ between subgroups, with corresponding changes in variant effect sizes.

We use 'absolute covariance' ρ preferentially (see supplementary table 1) because we expect that Z_a and Z_d will frequently co-vary positively and negatively at different SNPs in the same analysis; for instance, if some variants are deleterious only for subgroup 1 and others only for subgroup 2. A potential advantage of our symmetric model is the potential to generate Z_d scores from ANOVA-style tests for genetic homogeneity between three or more subgroups, in which case reconstructed Z scores would be directionless.

Aetiologically and genetically heterogeneous subgroups within a case group correspond to substructures in the genotype matrix. Information about such substructures is lost in a standard GWAS, which only uses the column-sums (MAFs) of the matrix (linear-order information). Data-driven selection of appropriate case subgroups and corresponding analyses of these subgroups can use more of the remaining quadratic-order information the matrix contains. Indeed a 'two-dimensional' GWAS approach (using Z_a and Z_d) instead of a standard GWAS (using only Z_a) may improve SNP discovery, as we found for *PTPN22* in RA/T2D. However, this can only be the case if the subgroups correspond to different variant effect sizes; for other subgroupings, a two-dimensional GWAS will only add noise.

While it seems appealing to use this method to search for some 'optimal' partition of patients, we prefer to focus on testing subgroupings derived from independent clinical or phenotypic data. Firstly, it is difficult to characterise subgroupings as 'better' or 'worse', and no one parameter can parametrise the degree to which two subgroups differ; parameters π_3, τ, and ρ all contribute, and attempts to test the hypothesis using a single measure such as genetic correlation have serious shortcomings (supplementary note, 4). Secondly, even if subgroups could meaningfully be ranked, the search space of potential subgroupings of a case group is prohibitively large (2^N for N cases), making exhaustive searches difficult.

We demonstrated that effect sizes of T1D-causative SNPs differ with age at disease diagnosis. The strong negative correlation observed (figure 5) was consistent with an increased total genetic liability in samples with earlier age of diagnosis, a finding supported by candidate gene studies [14, 15, 16] and epidemiological data [13]. Such a pattern arises naturally from a liability threshold model where total liability depends additively on both genetic effects and environmental influences which accumulate with age (supplementary note 1).

Our method necessarily dichotomises the multitude of mechanisms of heterogeneity, although there are many diverse forms (supplementary table 1, supplementary note 1).

There is potential to further dissect the mechanisms of disease heterogeneity by incorporating estimations of genetic correlation [18] or assessing evidence for liability threshold models [22]. Similar mixture-Gaussian approaches may also be adaptable to this purpose, by assessing other families of effect size distributions.

Our method adds to the current body of knowledge by extracting additional information from a disease dataset over a standard GWAS analysis, and determines if further analysis of disease pathogenesis in subgroups is justified. Our approach is analogous to the intuitive method of searching for between-subgroup differences in SNPs with known disease associations [3] but does not restrict attention to strong disease associations, enabling use of information from disease-associated SNPs which do not reach significance. Our parametrisation of effect size distributions allows insight into the structure of the genetic basis of the disease and potential subtypes, improving understanding of genotype-phenotype relationships.

Methods

Ethics Statement

This paper re-analyses previously published datasets. All patient data were handled in accordance with the policies and procedures of the participating organisations.

Joint distribution of variables Z_a, Z_d

We assume that SNPs may be divided into three categories, as described in the results section (figure 1). Under these assumptions, Z_a and Z_d scores have the joint *pdf* given by equation 1. We define Θ is the vector of values $(\pi_1, \pi_2, \pi_3, \tau, \sigma_2, \sigma_3, \rho)$. Z scores Z_a and Z_d are reconstructed from GWAS p-values for SNP associations. In practice, since our model is symmetric, we only require absolute Z scores, without considering effect direction.

For sample sizes n_1, n_2 and 97.5% odds-ratio quantile α, the expected observed standard deviation of Z scores (that is, σ_2, σ_3, and τ) is given by

$$E\{SD(Z)\} = \sqrt{1 + \frac{log(\alpha)^2 n_1 n_2}{12(n_1 + n_2)}} \tag{3}$$

(supplementary note, section 3.3).

Definition and distribution of PLR statistics

For a set of observed Z scores (Z_a, Z_d) we define the joint unadjusted pseudo-likelihood $PL_{da}(Z|\Theta)$ as

$$log\{PL_{da}(Z_d, Z_a|\Theta)\} = \sum_{Z_d^{(i)} \in Z_d, Z_a^{(i)} \in Z_a} w_i PDF_{Z_d, Z_a|\Theta}(Z_d^{(i)}, Z_a^{(i)}) + C\, log(\pi_1 \pi_2 \pi_3) \tag{4}$$

where the term $C\, log(\pi_1\pi_2\pi_3)$ is included to ensure identifiability of the model [5] and weights w_i are included to adjust for LD (see below).

We now set

$$\begin{aligned} \widehat{\theta_1} &= arg\, max_{\theta \in H_1}\, PL_{da}(Z_d, Z_a|\theta) \\ \widehat{\theta_0} &= arg\, max_{\theta \in H_0}\, PL_{da}(Z_d, Z_a|\theta) \\ uPLR(Z) &= \log \left(\frac{PL_{da}(Z|\widehat{\theta_1})}{PL_{da}(Z|\widehat{\theta_0})} \right) \end{aligned} \tag{5}$$

recalling that H_0 is the subspace of the parameter space H_1 satisfying $\sigma_3 = 1$ and $\rho = 0$.

If data observations are independent, $uPLR$ reduces to a likelihood ratio. Under H_0, the asymptotic distribution of $uPLR$ is then

$$uPLR \sim \frac{1}{2}\begin{cases} \chi^2_1 & p = 1/2 \\ \chi^2_2 & p = 1/2 \end{cases} \tag{6}$$

according to Wilk's theorem extended to the case where the null value of a parameter lies on the boundary of H_1 (since $\rho = 0$ under H_0) [23].

The empirical distribution of $uPLR$ may substantially majorise the asymptotic distribution when $\tau \approx 1$. In the full model, the marginal distribution of Z_a has more degrees of freedom (four; π_1, π_2, σ_2, σ_3) than it does under the null model (two; π_2, σ_2; as $\sigma_3 \equiv 1$). This can mean that certain distributions of Z_a can drive high values of $uPLR$ independent of the values of Z_d (supplementary note 3), which is unwanted as the values Z_a reflect only case/control association and carry no information about case subgroups. If observed $uPLR$s from random subgroups (for which $\tau = 1$ by definition) are used to approximate the null $uPLR$ distribution, this effect would lead to serious loss of power when $\tau >> 1$.

This effect can be managed by subtracting a correcting factor based on the pseudo-likelihood of Z_a alone, which reflects the contribution of Z_a values to the uPLR. We define

$$PL_a(Z_a|\Theta) = \prod_{Z_a^{(i)} \in Z_a} \left(\pi_1 N_{0,1}(Z_a^{(i)}) + \pi_2 N_{0,\sigma_2^2}(Z_a^{(i)}) + \pi_3 N_{0,\sigma_3^2}(Z_a^{(i)}) \right) \tag{7}$$

that is, the marginal likelihood of Z_a. Given $\widehat{\theta_1}$, $\widehat{\theta_0}$ as defined above, we define

$$f(Z_a) = min\left(log\frac{PL_a(Z_a|\widehat{\theta_1})}{PL_a(Z_a|\widehat{\theta_0})}, 0 \right) \tag{8}$$

We now define the PLR as

$$PLR = uPLR - f(Z_a) \tag{9}$$

The action of $f(Z_a)$ leads to the asymptotic distribution of PLR slightly minorising the asymptotic mixture-χ^2 distribution of uPLR, to differential degrees dependent on the value of τ (see supplementary note 3).

We define the similar test statistic $cPLR$:

$$
\begin{aligned}
cPL(Z_d|Z_a,\theta) &= \frac{PL_{da}(Z_a,Z_d|\theta)}{PL_a(Z_a|\theta)} \\
\widehat{\theta}_1^c &= arg\,max_{\Theta\in H_1}\, cPL(Z_d|Z_a,\theta) \\
\widehat{\theta}_0^c &= arg\,max_{\Theta\in H_0}\, cPL(Z_d|Z_a,\theta) \\
cPLR &= log\left(\frac{cPL(Z_d|Z_a,\widehat{\theta}_1^c)}{cPL(Z_d|Z_a,\widehat{\theta}_0^c)}\right)
\end{aligned}
\tag{10}
$$

noting that the expression $\frac{PL_{da}(Z_a,Z_d|\theta)}{PL_a(Z_a|\theta)}$ can be considered as a likelihood conditioned on the observed values of Z_a. Now

$$
\begin{aligned}
PLR &= \log\left(\frac{PL_{da}(Z_d,Z_a|\widehat{\theta}_1)}{PL_{da}(Z_d,Z_a|\widehat{\theta}_0)}\right) - log\left(\frac{PL_a(Z_a|\widehat{\theta}_1)}{PL_a(Z_d|\widehat{\theta}_0)}\right) \\
&= log\left(\frac{cPL(Z_d|Z_a,\widehat{\theta}_1)}{cPL(Z_d|Z_a,\widehat{\theta}_0)}\right)
\end{aligned}
\tag{11}
$$

The empirical distribution of cPLR for random subgroups majorises the empirical distribution of PLR (supplementary note 3). Furthermore, the approximation of the empirical distribution of cPLR by its asymptotic distribution is good, across all values of τ; that is, across the whole null hypothesis space.

Our approach is to compare the PLR of a test subgroup to the cPLR of random subgroups, which constitutes a slightly conservative test under the null hypothesis (see supplementary note 3).

Allowance for linkage disequilibrium

The asymptotic approximation of the pseudo likelihood-ratio distribution breaks down when values of Z_a, Z_d are correlated due to LD. One way to overcome this is to 'prune' SNPs by hiererarchical clustering until only those with negligible correlation remain. A disadvantage with this approach is that it is difficult to control which SNPs are retained in an unbiased way without risking removal of SNPs which contribute greatly to the difference between subgroups.

We opted to use the LDAK algorithm [4], which assigns weights to SNPs approximately corresponding to their 'unique' contribution. Denoting by ρ_{ij} the correlation between SNPs i, j, and $d(i,j)$ their chromosomal distance, the weights w_i are computed so that

$$w_i + \sum_{i \neq j} w_j \rho_{ij}^2 e^{-\lambda d(i,j)} \tag{12}$$

is close to constant for all i, and $w_i > 0$ for all i. The motivation for this approach is that $\sum_{i \neq j} \rho_{ij}^2$ represents the replication of the signal of SNP i from all other SNPs.

This approach has the advantage that if n SNPs are in perfect LD, and not in LD with any other SNPs, each will be weighted $1/n$, reducing the overall contribution to the likelihood to that of one SNP. In practice, the linear programming approach results in many SNP weights being 0. Using the LDAK algorithm therefore allows more SNPs to be retained and contribute to the model than would be retained in a pruning approach.

A second advantage of LDAK is that it homogenises the contribution of each genome region to the overall pseudo-likelihood. Many modern microarrays fine-map areas of the genome known or suspected to be associated with traits of interest [24] which could theoretically lead to peaks in the distribution of SNP effect sizes, disrupting the assumption of normality. LD pruning and LDAK both reduce this effect by homogenising the number of tags in each genomic region.

We adapted the pseudo-likelihood function to the weights by multiplying the contribution of each SNP to the log-likelihood by its weight (equation), essentially counting the ith SNP w_i times over. Adjusting using LDAK was effective in enabling the distributions of PLR to be well-approximated by mixture-χ^2 distributions of the form 2 (supplementary plots 4a, 4b, 4c).

E-M algorithm to estimate model parameters

We use an expectation-maximisation algorithm [25, 26] to fit maximum-PL parameters. Given an initial estimate of parameters $\Theta_0 = (\pi_1^0, \pi_2^0, \tau^0, \sigma_2^0, \sigma_2^0, \rho^0)$ we iterate three main steps:

1. Define for SNP s with Z scores $Z_d^{(s)}$,$Z_a^{(s)}$

$$\begin{aligned}
\zeta_g^{(s)} &= Pr(s \in \text{ category } g|\Theta_i) \\
&\propto \begin{cases} \pi_1^i N_{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}(Z_d^{(s)}, Z_a^{(s)}) & (g=1) \\ \pi_2^i N_{\begin{pmatrix} 1 & 0 \\ 0 & (\sigma_2^i)^2 \end{pmatrix}}(Z_d^{(s)}, Z_a^{(s)}) & (g=2) \\ \pi_3^i \left(\frac{1}{2} N_{\begin{pmatrix} (\tau^i)^2 & \rho^i \\ \rho^i & (\sigma_3^i)^2 \end{pmatrix}}(Z_d^{(s)}, Z_a^{(s)}) + \frac{1}{2} N_{\begin{pmatrix} (\tau^i)^2 & -\rho^i \\ -\rho^i & (\sigma_3^i)^2 \end{pmatrix}}(Z_d^{(s)}, Z_a^{(s)}) \right) & (g=3) \end{cases}
\end{aligned} \tag{13}$$

2. For $g \in (1,2,3)$ and LDAK weight w_s for SNP s set

$$\pi_g^{i+1} = \frac{\sum w_s \zeta_g^{(s)}}{\sum w_s} \tag{14}$$

3. Set

$$(\tau^{i+1}, \sigma_2^{i+1}, \sigma_3^{i+1}, \rho^{i+1}) = \text{arg } \max_{(\tau,\sigma_2,\sigma_3,\rho)} PL(Z_d, Z_a|\pi_1^{i+1}, \pi_2^{i+1}, \tau, \sigma_2, \sigma_3, \rho) \quad (15)$$

Step 3 is complicated by the lack of closed form expression for the maximum likelihood estimator of ρ (because of the symmetric two-Gaussian distribution of category 3), requiring a bisection method for computation. The algorithm is continued until $|PLR(Z_d, Z_a|\Theta_i) - PLR(Z_d, Z_a|\Theta_{i-1})| < \epsilon$; we use $\epsilon = 1 \times 10^{-5}$.

The algorithm can converge to local rather than global minima of the likelihood. We overcome this by initially computing the pseudo-likelihood of the data at 1000 points throughout the parameter space, retaining the top 100, and dividing these into 5 maximally-separated clusters. The full algorithm is then run on the best (highest-PL) point in each cluster

An appropriate choice of Θ_0 can speed up the algorithm considerably; for simulations, we begin the model at previous maximum-PL estimates of parameters for earlier simulations.

Maximum-cPL estimations of parameters were made using generic numerical optimisation with the *optim* function in R. Prior to applying the algorithm, parameters π_2 and σ_2 are estimated as maximum-PL estimators of the objective function

$$g(Z_a|\pi_2, \sigma_2) = \sum w_i log\{(1 - \pi_2)N_{0,1}(Z_a^{(i)}) + \pi_2 N_{0,\sigma_2^2} Z_a^{(i)}\} \quad (16)$$

where w_i is the weight for SNP i (see supplementary note 3 for rationale). The conditional pseudo-likelihood was maximised over the remaining parameters.

The algorithm and other processing functions are implemented in an R package available at `https://github.com/jamesliley/subtest`

Properties and assumptions of the PLR test

Our assumption that (Z_a, Z_d) follows a mixture Gaussian is generally reasonable for complex phenotypes with a large number of associated variants [8] and our adjustment for the distribution of Z_a (essentially conditioning on observed Z_a) reduces reliance on this assumption. If subgroup prevalence is unequal between the study group and population, our method can still be used with adaptation (supplementary note, section 2.4).

Our test is robust to confounders arising from differential sampling to the same extent as conventional GWAS. For example, if subgroups were defined based on population structure, and population structure also varied between the case and control group, SNPs which differed by ancestry would also appear associated with the disease, leading to a loss of control of type-1 error rate. However, the same study design would also lead to identification of spurious association of ancestry-associated SNPs with the phenotype in a conventional GWAS analysis. As for GWAS, this effect can be alleviated by including the confounding trait as a covariate when computing p-values (Supplementary Note 2).

Prioritisation of single SNPs

An important secondary problem to testing H_0 is the determination of which SNPs are likely to be associated with disease heterogeneity. Ideally, we seek a way to test the association of a SNP with subgroup status (ie, Z_d), which gives greater priority to SNPs potentially associated with case/control status (ie, high Z_a).

An effective test statistic meeting these requirements is the Bayesian conditional false discovery rate (cFDR) [6]. It tests against the null hypothesis H_0' that the population minor allele frequencies of the SNP in both case subgroups are equal (ie, that the SNP does not differentiate subgroups), but responds to association with case/control status in a natural way by relaxing the effective significance threshold on $|Z_d|$. This relaxation of threshold only occurs if there is systematic evidence that high $|Z_d|$ scores and high $|Z_a|$ scores typically co-occur. The test statistic is direction-independent.

Given a set of observed Z_a and Z_d values $Z_a^{(i)}$, $Z_d^{(i)}$, with corresponding two-sided p values p_{ai}, p_{di}, the cFDR for SNP j is defined as

$$\begin{aligned} X_4 &= p_{dj} \frac{\left|\{i : p_{ai} \leq p_{aj} \wedge p_{di} \leq p_{dj}\}\right|}{\left|\{i : p_{di} \leq p_{dj}\}\right|} \\ &\approx Pr(H_0' \big| P_a \leq p_{aj}, P_d \leq p_{dj}) \end{aligned} \tag{17}$$

The value gives the false-discovery rate for SNPs whose p-values fall in the region $[0, p_{dj}] \times [0, p_{aj}]$; this can be converted into a false-discovery rate amongst all SNPs for whom X_4 passes some threshold [7].

We discuss three other single-SNP test statistics in supplementary note 5.1, which test against different null hypotheses. If the hypothesis H_0' is to be tested, then we consider the cFDR the best of these.

Contour plots of the test statistics for several datasets are shown in supplementary figures 7,8, and 9.

Genetic correlation testing

Given the correlation between Z_d and Z_a in the age-at-diagnosis analysis, methods to estimate narrow-sense genetic correlation (r_g) [18, 19] may be adaptable to the subgrouping question by estimating r_g across a set of SNPs between case/control traits of interest, with the potential advantage of characterising heterogeneity using a single widely-interpretable metric. This may be between Z scores derived from comparing the control group to each case subgroup, testing under the null hypothesis $r_g = 1$ (method 1); or between the familiar Z_a and Z_d, under the null hypothesis $r_g = 0$ (method 2).

We explored these methods in supplementary note 4. We show that method 1 leads to systematically high false positive rates, as r_g is also reduced from 1 in subgroupings that are independent of the overall disease process (e.g. hair colour in T2D). We show that method 2 is considerably less powerful than our method because it tests a narrower definition of H_1 which does not take account of the marginal variances of the distribution of Z_d, Z_a in category 3, and requires that correlation between Z_d and Z_a be always positive or always negative, in contrast to our symmetric model (Figure 1). Indeed, parameter ρ estimates an analogue of r_g accounting for simultaneous correlation and anticorrelation.

Methods to compute r_g were not explicitly proposed as a method for subgroup testing, and our analysis does not indicate any general shortcomings. However, comparison with r_g based approaches places our method in the context of established methodology, demonstrating the necessity of considering both variance parameters (τ, σ_3) and covariance parameters (ρ) in testing a subgrouping of interest.

Description of GWAS datasets

ATD samples were genotyped on the ImmunoChip [24] a custom array targeting putative autoimmune-associated regions. Data were collected for GWAS-like analyses of dense SNP data [12]. The dataset comprised 2282 cases of Graves' disease, 451 cases of Hashimoto's thyroiditis, and 9365 controls.

T1D samples were genotyped on either the Illumina 550K or Affymetrix 500K platforms, gathered for a GWAS on T1D [17]. We imputed between platforms in the same way as the original GWAS. The dataset comprised genotypes from 5908 T1D cases and 8825 controls, of which all had measured values of TPO-Ab, 3197 had measured IA2-Ab, 3208 had measured GAD-Ab, and 2240 had measured PCA-Ab. Comparisons for each autoantibody were made between cases positive for that autoantibody, and cases not positive for it. We did not attempt to perform comparisons of individuals positive for different autoantibodies (for instance, TPO-Ab positive vs IA2-Ab positive) because many individuals were positive for both.

To generate summary statistics corresponding to geographic subgroups, we considered the subgroup of cases from each of twelve regions and each pair of regions against all other cases (78 subgroupings in total). To maximise sample sizes, we considered T1D cases as 'controls' and split the control group into subgroups.

Quality control

Particular care had to be taken with quality control, as Z-scores had to be relatively reliable for all SNPs assessed, rather than just those putatively reaching genome-wide significance.. For the T1D/T2D/RA comparison, which we re-used from the WTCCC, a critical part of the original quality control procedure was visual analysis of cluster plots for SNPs reaching significance, and systematic quality control measures based on differential call rates and deviance from Hardy-Weinberg equilibrium (HWE) were correspondingly loose [10]. Given that we were not searching for individual SNPs, this was clearly not appropriate for our method.

We retained the original call rate (CR) and MAF thresholds (MAF $\geq$ 1%, CR $\geq$ 95% if MAF $\geq$ 5%, CR $\geq$ 99% if MAF <5%) but employed a stricter control on Hardy-Weinberg equilibrium, requiring $p \geq 1 \times 10^{-5}$ for deviation from HWE in controls. We also required that deviance from HWE in cases satisfied $p \geq 1.91 \times 10^{-7}$, corresponding to $|z| \leq 5$. The looser threshold for HWE in cases was chosen because deviance from HWE can arise due to true SNP effects [27]. We also required that call rate difference not be significant ($p \geq 1 \times 10^{-5}$) between any two groups, included case-case and case-control differences. Geographic data was collected by the WTCCC and consisted of assignment of samples to one of twelve geographic regions (Scotland, Northern, Northwestern, East and West Ridings, North Midlands, Midlands, Wales, Eastern, Southern, Southeastern, and London [10]). In analysing differences between autoimmune diseases, we stratified by geographic location; when assessing subgroups based on geographic location, we did not.

For the ATD and T1D data, we used identical quality control procedures to those employed in the original paper [12, 17]. We applied genomic control [28] to computation of Z_a and Z_d scores except for our analysis of ATD (following the original authors [12]) and our geographic analyses (as discussed above). In all analyses except where otherwise indicated we removed the MHC region with a wide margin ($\approx 5Mb$ either side).

Acknowledgments

We acknowledge the help of the Diabetes and Inflammation Laboratory Data Service for access and quality control procedures on the datasets used in this study. The JDRF/Wellcome Trust Diabetes and Inflammation Laboratory is in receipt of a Wellcome Trust Strategic Award (107212, JAT) and receives funding from the JDRF (5-SRA-2015-130-A-N, JAT) and the NIHR Cambridge Biomedical Research Centre. The research leading to these results has received funding from the European Unions 7th Framework Programme (FP7/2007-2013, JAT) under grant agreement no. 241447 (NAIMIT). JL is funded by the NIHR Cambridge Biomedical Research Centre and is on the Wellcome Trust PhD programme in Mathematical Genomics and Medicine at the University of Cambridge. CW is funded by the Wellcome Trust (089989,107881, CW) and the MRC (MC_UP_1302/5, CW). The Cambridge Institute for Medical Research (CIMR) is in receipt of a Wellcome Trust

Strategic Award (100140). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

The JDRF/Wellcome Trust Diabetes and Inflammation Laboratory receives funding from Hoffmann La Roche and Eli-Lilly and Company.

Author contributions

AJL: conceived the statistical methods, wrote the software, performed the analyses, analysed the data, and wrote the manuscript. JAT: analysed the data and edited the manuscript. CW: conceived the study, analysed the data, and wrote the manuscript

Code availability

Code available from `https://github.com/jamesliley/subtest` (R package)

Data availability

This paper re-analyses previously published datsets. WTCCC data access for T1D/T2D/RA and controls [10] is described at `https://www.wtccc.org.uk/info/access_to_data_samples.html`. ATD data are available on request to the original study authors [12]. T1D genetic data from [17] is available at `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000180.v3.p2` which we combined with autoantibody data available from study authors [3]

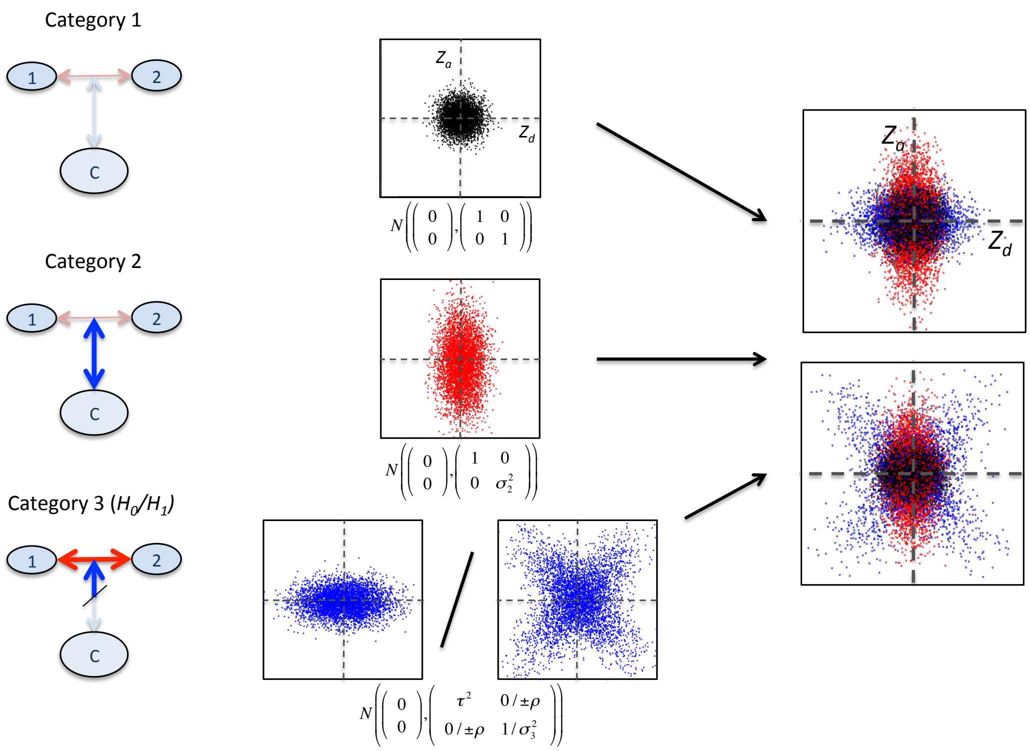


Figure 1: Overview of three-categories model. Z_d and Z_a are Z scores derived from GWAS p-values for allelic differences between case subgroups (1 vs 2), and between cases and controls (1+2 vs C) respectively (left). Within each category of SNPs, the joint distribution of (Z_d, Z_a) has a different characteristic form. In category 1, Z scores have a unit normal distribution; in category 2, the marginal variance of Z_a can vary. The distribution of SNPs in category 3 depends on the main hypothesis. Under H_0 (that all disease-associated SNPs have the same effect size in both subgroups), only the marginal variance of Z_d may vary; under H_1 (that subgroups correspond to differential effect sizes for disease-associated SNPs), any covariance matrix is allowed. The overall SNP distribution is then a mixture of Gaussians resembling one of the rightmost panels, but with SNP category membership unobserved. Visually, our test determines whether the observed overall Z_d, Z_a distribution more closely resembles the bottom rightmost panel than the top.

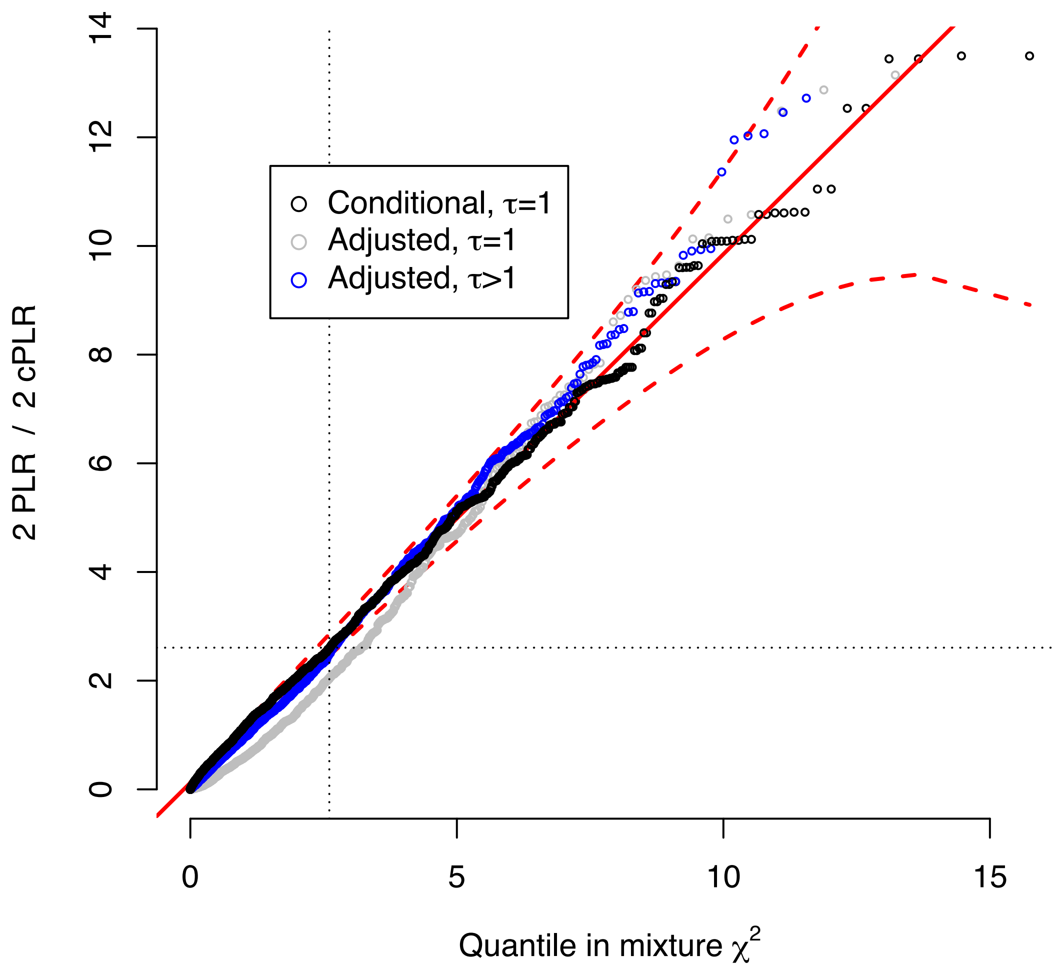


Figure 2: QQ plot from simulations demonstrating type 1 error rate control of PLR test. PLR values for test subgroups under H_0 with either $\tau = 1$ (random subgroups; grey) or $\tau > 1$ (genetic difference between subgroups, but independent of main phenotype; blue) with cPLR values for random subgroups (black) and against proposed asymptotic distribution under simulation ($\frac{1}{2}\left(\chi_1^2 + \chi_2^2\right)$; solid red line; 99% confidence limits dashed red line). The distribution of cPLR for random subgroups majorises the distribution of PLR, meaning the PLR-based test is conservative. Further details are shown in supplementary note, section 3.

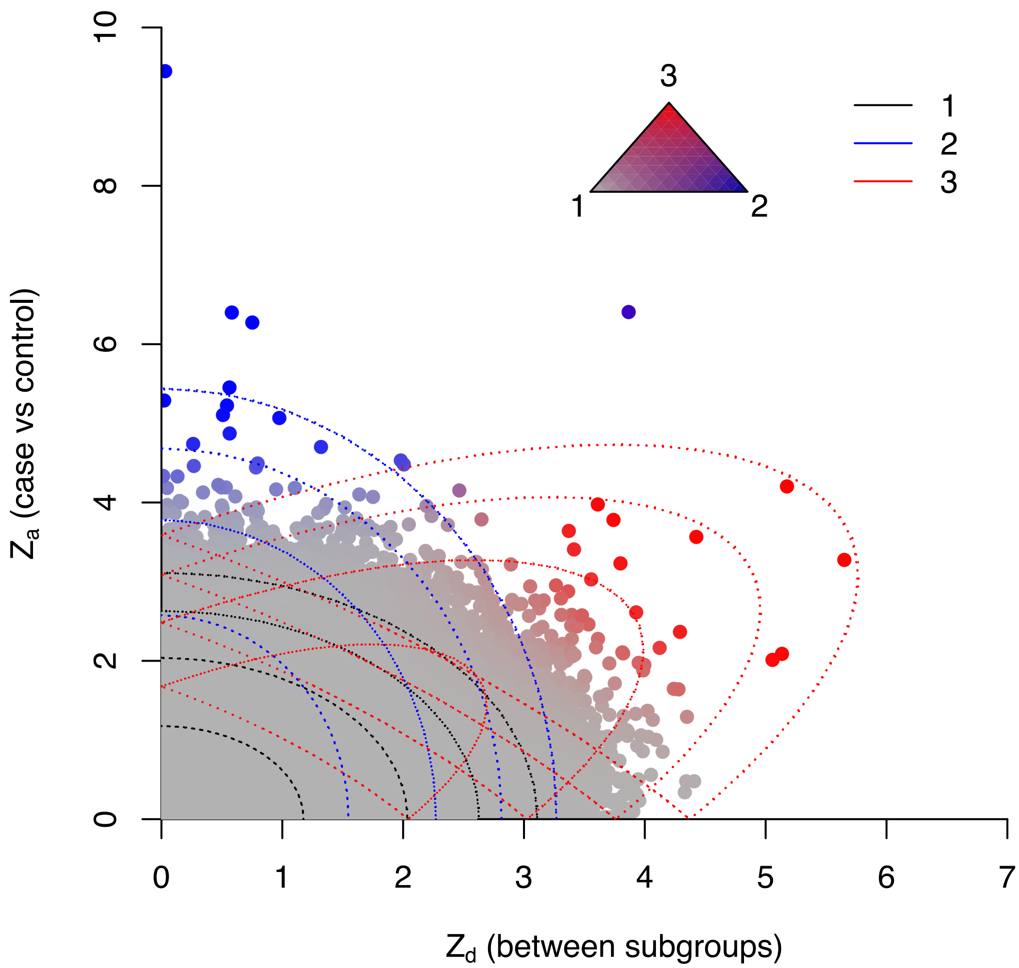


Figure 3: Observed absolute Z_a and Z_d for T1D/RA. Colourings correspond to posterior probability of category membership under full model (see triangle): grey - category 1, blue - category 2, red -category 3. Contours of the component Gaussians in the fitted full model are shown by dotted lines.

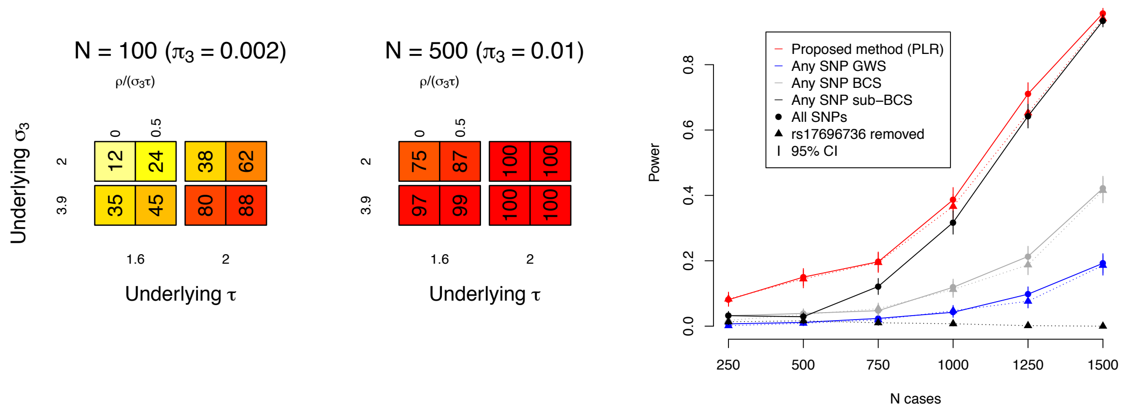


Figure 4: Power of PLR to reject H_0 (genetic homogeneity between subgroups) depends on the number of SNPs in category 3 and the underlying values of model parameters σ_2, σ_3, τ, ρ. Dependence on number of case/control samples arises through the magnitudes of σ_3 and τ (supplementary note, section 3.3). Leftmost figure shows power estimates for various values of π_3, σ_3, τ, ρ. Value N is the approximate number of SNPs in category 3, ($\propto \pi_3$). Each simulation was on 5×10^4 simulated autosomal SNPs in linkage equilibrium. Value $\rho/(\sigma_3 \tau)$ is the absolute correlation between Z_d and Z_a in category 3. Also see supplementary figure 3. Rightmost figure shows power of PLR to detect differences in genetic basis of T1D and RA subgroups of a combined autoimmune dataset, downsampling to varying numbers of cases (X axis). PLR is compared with: power to find ≥ 1 SNP with Z_d score reaching genome-wide significance (GWS, blue; $p \leq 5 \times 10^-8$) or Bonferroni-corrected significance (BCS, green; $p \leq$ 0.05?(total # of SNPs)); and power to detect any SNP with Za score reaching genome-wide significance and Zd score reaching Bonferroni-corrected significance (sub-BCS, grey; p0.05?(total # of SNPs with Za reaching GWS)). Error bars show 95% CIs. Circles/solid lines for each colour show power for all SNPs, triangles/dashed lines for all SNPs except rs17696736. Power for sub-BCS drops dramatically but power for PLR is not markedly affected, indicating relative robustness of PLR to single-SNP effects.

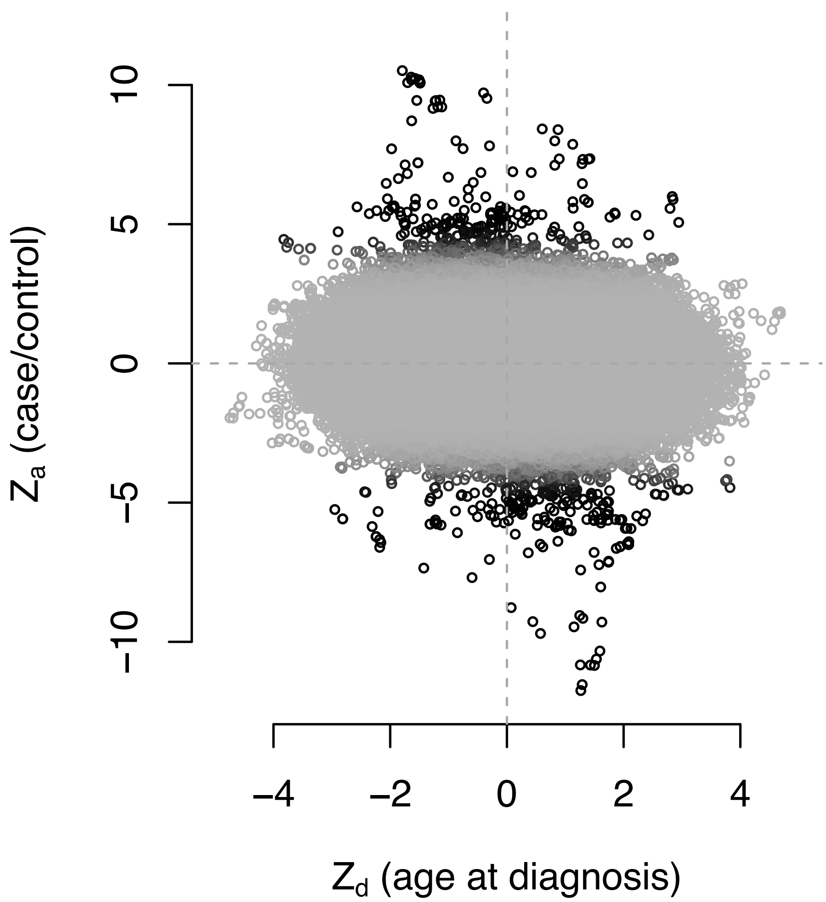


Figure 5: Z_a and Z_d scores for age at diagnosis in T1D, excluding MHC region. Colour corresponds to posterior probability of category 2 membership in null model (since categories in full model are assigned on the basis of correlation), with black representing a high probability. Z_d and Z_a are negatively correlated ($p = 8.7 \times 10^{-5}$ with MHC included, $p = 0.002$ with MHC removed) after accounting for LD using LDAK weights, and weighting by posterior probability of category 2 membership in the null model, to prioritise SNPs further from the origin

References

[1] Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, et al. (2015) Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science translational medicine 7: 311ra174–311ra174.

[2] Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, et al. (2009) A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genetic Epidemiology 34: 335-343.

[3] Plagnol V, Howson JMM, Smyth DJ, Walker N, Hafler JP, et al. (2011) Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLOS Genetics 7.

[4] Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. American Journal of Human Genetics 91: 1011-1021.

[5] Chen H, Chen J, Kalbfleisch JD (2001) A modified likelihood ratio test for homogeneity in finite mixture models. Journal of the Royal Statistical Society, series B (methodological) 63: 19-29.

[6] Andreassen OA, Thompson WK, Schork AJ, Ripke S, Mattingsdal M, et al. (2013) Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. PLOS Genetics 9(4).

[7] Liley J, Wallace C (2015) A pleiotropy-informed bayesian false discovery rate adapted to a shared control design finds new disease associations from gwas summary statistics. PLOS Genetics .

[8] Lo PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, et al. (2015) Efficient bayesian mixed model analysis increases association power in large cohorts. Nature Genetics 47: 284-90.

[9] Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, et al. (2015) The fine-scale genetic structure of the british population. Nature 519: 309-314.

[10] The Wellcome trust case control consortium (2007) Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. Nature 447: 661-678.

[11] Fortune MD, Guo H, Burren O, Schofield E, Walker NM, et al. (2015) Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. Nature Genetics 47: 839-846.

[12] Cooper JD, Simmonds MJ, Walker NM, Burren O, Brand OJ, et al. (2012) Seven newly identified loci for autoimmune thyroid disease. Human Molecular Genetics 21: 5202-5208.

[13] Hyttinen V, Kaprio J, Kinnunen L, Koskenvuo M, Tuomilehto J (2003) Genetic liability of type 1 diabetes and the onset age among 22, 650 young finnish twin pairs in a nationwide follow up study. Diabetes 52: 1052-1055.

[14] Howson JMM, Walker NM, Smyth DJ, Todd JA (2009) Analysis of 19 genes for association with type 1 diabetes in the type 1 diabetes genetics consortium families. Genes and Immunity 10: S74-S84.

[15] Howson JM, Rosinger S, Smyth DJ, Boehm BO, Todd JA, et al. (2011) Genetic analysis of adult-onset autoimmune diabetes. Diabetes 60: 2645–2653.

[16] Howson JM, Cooper JD, Smyth DJ, Walker NM, Stevens H, et al. (2012) Evidence of gene-gene interaction and age-at-diagnosis effects in type 1 diabetes. Diabetes 61: 3012–3017.

[17] 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.

[18] Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. (2015) An atlas of genetic correlations across human diseases and traits. bioRxiv .

[19] Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics 28: 2540–2542.

[20] Traylor M, Bevan S, Rothwell PM, Sudlow C, 2 WTCCC, et al. (2013) Using phenotypic heterogeneity to increase the power of genome-wide association studies: Application to age at onset of ischaemic stroke subphenotypes. Genetic Epidemiology 37: 495-503.

[21] Wen Y, Lu Q (2013) A multiclass likelihood ratio approach for genetic risk prediction allowing for phenotypic heterogeneity. Genetic epidemiology 37: 715–725.

[22] Chatterjee N, Carroll RJ (2005) Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92: 399–418.

[23] Self SG, Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82: 605-610.

[24] Cortes A, Brown MA (2011) Promise and pitfalls of the ImmunoChip. Arthritis Research and Therapy 13.

[25] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, series B (methodological) 39: 1-38.

[26] Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer Series in Statistics. Springer.

[27] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, et al. (2010) Data quality control in genetic case-control association studies. Nature protocols 5: 1564-1573.

[28] Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theoretical Population Biology 60: 155-166.

A method for identifying genetic heterogeneity within phenotypically-defined disease subgroups
Supplementary Note

James Liley, John A Todd and Chris Wallace

November 8, 2016

Contents

1 Disease models in H_1 and H_0 5
1.1 Disease models in H_1 5
1.2 Disease models in H_0 6
1.3 Subgrouping by a risk factor 6

2 Distribution of Z scores 9
2.1 Definitions 9
2.1.1 Unstratified groups 9
2.1.2 Adjustment for strata 10
2.1.3 Adjustment for covariates 11
2.2 Z_d and Z_a are conditionally independent in categories 1 and 2 13
2.2.1 Unstratified or stratified groups 13
2.2.2 Adjustment for covariates 13
2.3 SNPs in category 3 13
2.3.1 Unstratified groups 14
2.3.2 Adjustment for strata 15
2.3.3 Adjustment for covariates 16
2.4 Unequal subgroup prevalences 16
2.4.1 Motivation 16
2.4.2 Behaviour of standard approach 16
2.4.3 Adaptation 18
2.4.4 No adjustment - unbiased sampling 18
2.4.5 Adjustment for strata 19
2.4.6 Adjustment for covariates 19
2.5 Testing procedure 20
2.5.1 Algorithm 20
2.5.2 Rationale 21

3 Details of simulations 25
3.1 Simulations of random genotypes 25
3.2 Simulation on GWAS case group subgroups 26
3.3 Distributions of parameter values for simulation and power calculations 27

4 Genetic correlation as an alternative to PLR test 31
4.1 Overview 31
4.2 Method 1: control-subgroup 1 vs control-subgroup 2 31
4.2.1 Expected behaviour 31
4.2.2 Simulations 32
4.2.3 Application to real data 32
4.3 Method 2: Z_d (case vs control) vs Z_a (subgroup 1 vs 2) 33

4.3.1 Expected behaviour, and relation of ρ_g to ρ 33
4.3.2 Simulations 35
4.3.3 Application to real data 35

5 Other **41**
5.1 Alternative test statistics for retrospective single-SNP analysis 41
5.2 Independence of PLR distribution on subgroup sizes 42
5.3 Number of simulations necessary to fit null distribution 42

Note 1

Disease models in H_1 and H_0

We define ‘differential causative pathology’ (our alternative hypothesis, H_1) to mean that some subset of disease-associated variants have different population effect sizes in the case subgroups in question. Our method tests against the null hypothesis H_0 that all disease associated variants have the same effect sizes in both subgroups. An equivalent formulation of H_0 is that the (possibly empty) sets of SNPs which have different minor allele frequencies in case and control groups and which have different minor allele frequencies in case subgroups are non-intersecting.

The multitude of potential causes for disease heterogeneity necessitate that both H_0 and H_1 encompass a range of such causes. We list several below, with illustration in supplementary table 1.

We define the 'genetic architecture' of a trait as a set of variants and corresponding effect sizes (log-odds ratios or asymptotically similar statistics) between populations with and without the trait. In general, most effect sizes are zero or negligibly small.

1.1 Disease models in H_1

The simplest model of disease heterogeneity in H_1 is the scenario in which some variants are associated with one case subgroup, but not the other. For such a variant the effect size in one subgroup is zero, and in the other nonzero. This would be expected to arise if some of the pathological processes giving rise to the disease were specific to one case subgroup.

A second potential model in H_1 is when the same variants are associated with both subgroups, but the relative effect sizes differ. This may arise in a situation where pathological processes differ in relative impact between subgroups. For instance, if two pathological processes may lead to a disease of interest, and one process is likely to occur during the neonatal period while the another is likely to occur during adolescence, a division of a case group into neonatal-onset and adolescent-onset would likely show variants associated with the first process as being more important in the first subgroup, and variants associated with the second process as being more important in the second, although the set of associated variants may be the same in both subgroups. The scenario may also arise if the cases can be split into subgroups like those described in the first paragraph, but the subgrouping criterion is only an approximation to this split.

A third model is when the same variants are associated with both subgroups with but where the effect sizes in one subgroup are a constant factor larger than in the other subgroup. This corresponds to differential heritability between subgroups, with the same pathological processes present. In a liability threshold model where some environmental variable has an additive effect with genetic risk, we would expect that defining subgroups based on the environmental variable would lead to this scenario (figure 1.1). In this case, the environment modulates the effect of the genetic risk. As an example, under the assumption that a dietary risk factor has an additive effect with genetic risk factors in type 2 diabetes, a disease subgroup with the dietary risk factor would be expected to have lower disease heritability than a subgroup without it.

1.2 Disease models in H_0

Under H_0, all disease associated variants have the same effect size in both subgroups. This may take the form of an absence of any systematic genetic difference between case subgroups, in which case the population allelic frequencies of disease-associated SNPs, and hence the effect sizes of such SNPs between controls and each case subgroup, are equal.

Hypothesis H_0 also allows the presence of genetic differences between subgroups at different SNPs to those associated with the disease. This may be particularly prominent if variation in the disease depends on how the disease process acts on different individual physiologies, in which case genetic variation between subgroups is at different SNPs to those involved in disease causality.

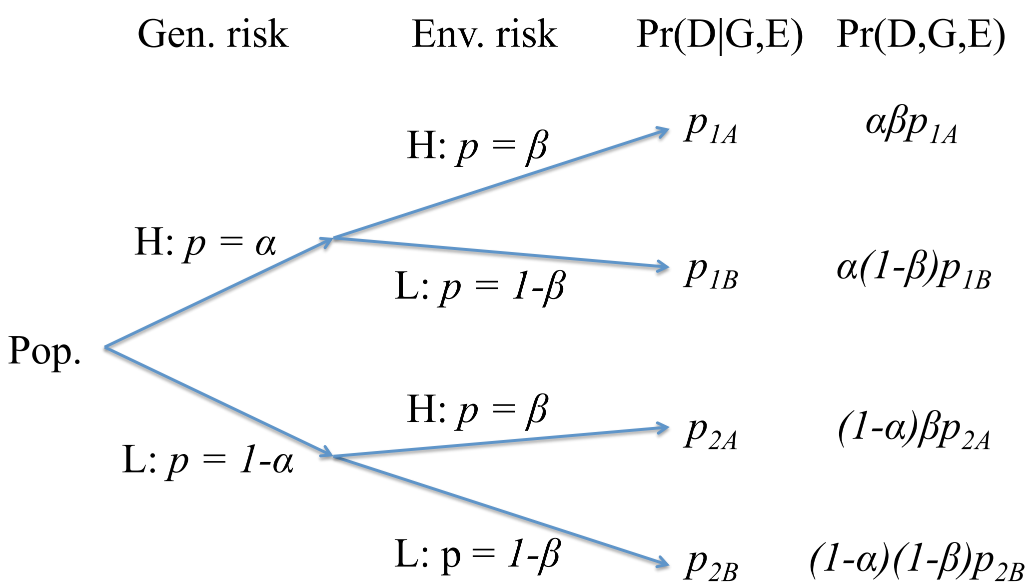


Figure 1.1: In a simplistic disease model, we consider two levels of genetic risk G with frequencies α, $1-\alpha$ and an independent two-level environmental risk factor E with frequencies β, $1-\beta$, and a disease D. In cases with the environmental risk factor, we would expect the ratio of high-genetic risk to low-genetic risk cases to be $\frac{\alpha}{1-\alpha}\frac{p_{1A}}{p_{2A}}$, and in cases without, $\frac{\alpha}{1-\alpha}\frac{p_{1B}}{p_{2B}}$. Assume we define subgroups based on the environmental risk factor. If the risk factor has a multiplicative effect on $Pr(D|G,E)$, so $\frac{p_{1A}}{p_{2A}} = \frac{p_{1B}}{p_{2B}}$, the prevalences of genetic risk groups are identical in the groups, and the heritability of D is the same. If the effect of the environmental risk factor on $Pr(D|G,E)$ changes with G, so the environmental risk factor modulates the genetic risk, this will not hold.

1.3 Subgrouping by a risk factor

Partitioning a case group by a known disease risk factor may lead to subgroupings in either H_0 or H_1 dependent on the interaction between the genetic and environmental risk factors. If the risk factor on which the subgrouping is based has a multiplicative effect on disease risk with genetic factors, then we expect the subgrouping to be in H_0 (figure 1.1). This may take the form of a binary risk factor: if a disease is triggered by an environmental event (for example, a particular mutation driven by environmental mutagens), with susceptibility to that event determined genetically (for instance, impaired ability to repair the mutation), conditioning on environment will not affect the distribution of genetic risk, and the subgrouping will be in H_0. The genetic risk may also be binary; for example, the development of a disease may require the knockout of a particular cellular process, with the genetic risk for the disease solely involved in risk of the knockout.

However, deviation from a locally multiplicative model can also lead to a subgrouping in H_1. One instance this may occur is if disease risk approaches 1. A current model of T1D pathogenesis requires the presence of an environmental insult to trigger genetic susceptibility ([1]), which could be expected to lead to a locally multiplicative relationship between age-at-diagnosis and genetic risk (figure 1.2). However, if genetic risk can be high enough that some individuals are almost sure to get the disease, this will lead to the subgrouping being in H_1 - a potential reason for the observation regarding age-at-diagnosis in T1D in the main text.

Finally, cases may be subgrouped according to non-causative clinical disease associations. Assume some binary clinical marker M has non-zero frequency in healthy individuals and has some set of associated genetic variants G_0. Let D be a genetically homogenous disease with a set of associated variants G_1 such that $G_0 \cap G_1 = \emptyset$ and D (or a necessary precurser of D) probabilistically causes M to occur more often than in the general population. Then when we condition on case status (and hence any necessary precursers of D) the only variants which are associated with M-status in cases will be in G_0, and a subgrouping based on M will be in H_0, despite M being associated with D. If, however, subtypes of D with differential genetic basis induce M to different degrees, and hence M serves as an index of such subtypes of D, then a subgrouping of M will fall in H_1.

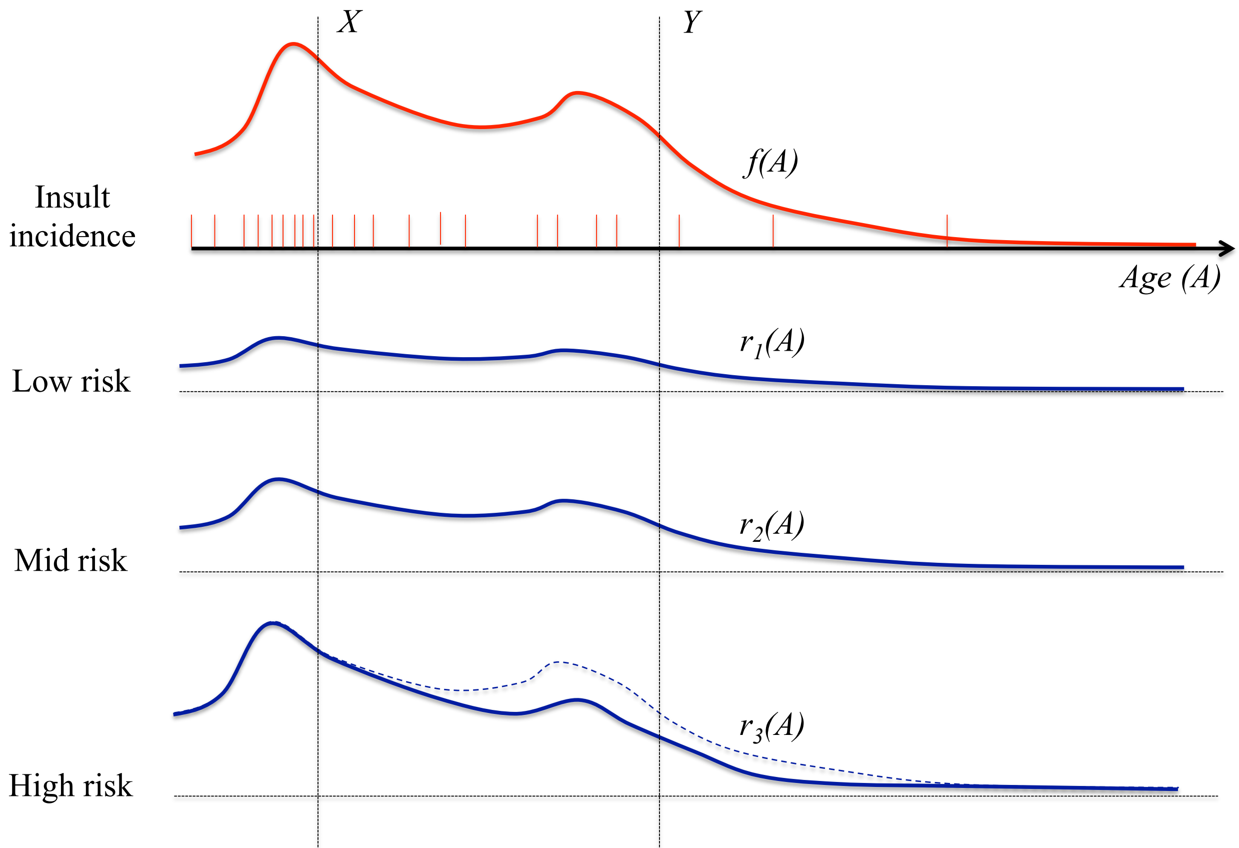


Figure 1.2: In a simplified model of incidence of type 1 diabetes or a similar autoimmune disease, we consider the disease to be triggered by an environmental 'insult'; for instance (eg, a viral illness) and three levels of genetic susceptibility to such insults. Denoting by $f(A)$ the density of such insults at age A (red vertical lines show a possible example for one individual), we expect that for individuals at low or moderate genetic risk the densities $r_1(A)$, $r_2(A)$ of disease incidence are proportional to $f(A)$, with lifetime risk $\int r_1(A)dA$, $\int r_2(A)dA$ respectively. The risk of disease at age A can be considered a product of $f(A)$ and a genetic risk score. In a high-risk group for a disease such as type 1 diabetes, it is possible that the lifetime risk $\int r_3(A)dA$ approaches 1, the high-risk group becomes 'saturated' with disease cases, and there are fewer non-affected individuals in the group at higher age groups, leading to a lower constant of proportionality with $f(A)$ at higher ages (dotted/solid lines). In the absence of the high-risk group, a subgrouping of patients into those with age-at-onset X and those with age-at-onset Y (vertical lines) would be expected to contain the same proportion of low- and mid- genetic risk samples in each subgroup, with correspondingly equal heritability of disease in each subgroup. With the high-risk group, the multiplicative effect of $f(A)$ on disease risk breaks down, inducing an environmental influence on the genetic risk, and changing the heritability between groups.

Note 2

Distribution of Z scores

In this section, we define the test statistics (Z scores) used to characterise allelic differences between groups and describe the rationale for our probabilistic model.

We partition SNPs into three theoretical categories:

1. SNPs which are not associated with case/control status or case subgroup status
2. SNPs which are associated with the main phenotype but have the same effect size in both case subgroups
3. SNPs which are associated with the difference between case subgroups

We consider SNP effect sizes between subgroups and between cases and controls to be realisations of bivariate random variables, which have different distributions in each category.

2.1 Definitions

2.1.1 Unstratified groups

Let x be a random sample of size n_x from patient population X, and y a sample of size n_y from a population Y. Denote by m_x, m_y the allele frequencies of some SNP of interest in x and y, and by μ_x, μ_y the allele frequencies in X and Y. We assume for the moment that x and y are unbiased samples, so $\mu_x = E(m_x)$ and $\mu_y = E(m_y)$.

In general, we compute Z scores from GWAS -defined p-values P_{xy} using the formula

$$Z_{n_x,n_y}(m_x, m_y) = -\Phi^{-1}(P_{xy}/2)sign(m_x - m_y) \tag{2.1}$$

Although there are several ways in which a GWAS p-value may be computed, the resultant Z scores all have several common asymptotic properties. In general, we assume a Z score $Z_{n_x,n_y}(m_x, m_y)$ is a smooth function of allele frequencies m_x, m_y, n_x, n_y with the following properties

1. For fixed observed overall allele frequency $\frac{n_x m_x + n_y m_y}{n_x + n_y}$, $Z_{n_x,n_y}(m_x, m_y)$ is monotonic to the allelic difference $m_x - m_y$
2. Under the null hypothesis $\mu_x = \mu_y$,
   (a) $E(Z_{n_x,n_y}(m_x, m_y)) = 0$
   (b) $var(Z_{n_x,n_y}(m_x, m_y)) = 1$
   (c) $Z_{n_x,n_y}(m_x, m_y) \to_d N(0, 1)$ as $n_x, n_y \to \infty$

These properties imply that the first-order expansion of Z about $(m_x, m_y) = (\mu, \mu)$ is:

$$Z_{n_x,n_y}(m_x, m_y) = \sqrt{\frac{2n_x n_y}{n_x + n_y}} \frac{m_x - m_y}{\sqrt{\mu(1-\mu)}} + O\left((m_x - \mu)(m_y - \mu)\right) \tag{2.2}$$

since

$$\begin{aligned} \sqrt{2n_x}(m_x - \mu_x) &\rightarrow_d N(0, \mu_x(1-\mu_x)) \\ \sqrt{2n_y}(m_y - \mu_y) &\rightarrow_d N(0, \mu_y(1-\mu_y)) \end{aligned} \tag{2.3}$$

and if $\mu_x = \mu_y = \mu$

$$\sqrt{\frac{2n_x n_y}{n_x + n_y}} \frac{m_x - m_y}{\sqrt{\mu(1-\mu)}} \rightarrow_d N(0, 1) \tag{2.4}$$

and only one linear function of m_x, m_y can be asymptotically $N(0, 1)$.

If $\mu_x \neq \mu_y$ and

$$\lambda = \frac{\mu_x - \mu_y}{\sqrt{\frac{\mu_x(1-\mu_x)}{2n_x} + \frac{\mu_y(1-\mu_y)}{2n_y}}} \tag{2.5}$$

remains finite as $n_x, n_y \rightarrow \infty$, we have

$$\begin{aligned} Z_{n_x,n_y}(m_x, m_y) &\approx \frac{m_x - m_y}{\sqrt{\frac{m_x(1-m_x)}{2n_x} + \frac{m_y(1-m_y)}{2n_y}}} \\ &= \frac{(m_x - m_y) - (\mu_x - \mu_y)}{\sqrt{\frac{m_x(1-m_x)}{2n_x} + \frac{m_y(1-m_y)}{2n_y}}} + \frac{\mu_x - \mu_y}{\sqrt{\frac{m_x(1-m_x)}{2n_x} + \frac{m_y(1-m_y)}{2n_y}}} \\ &\rightarrow_d N(0, 1) + \lambda \\ &= N(\lambda, 1) \end{aligned} \tag{2.6}$$

For a randomly chosen SNP, let μ_c be the population allele frequency (AF) in controls, and μ_1, μ_2 the population AFs in case subgroups 1 and 2 respectively, for the same allele. Define ν as the relative prevalence of subgroup 1 and $1 - \nu$ as the relative prevalence of subgroup 2. The population AF across all cases is $\mu_{12} = \nu\mu_1 + (1 - \nu)\mu_2$.

Denote by m_c, m_1, m_2 the corresponding observed AFs in a study with n_c, n_1, n_2 controls and samples in subgroup 1 and subgroup 2 respectively. Define $m_{12} = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}$ as the AF in the whole case group and $n_{12} = n_1 + n_2$. We assume that $\frac{n_1}{n_1 + n_2} \approx \nu$; that is, the case group is an unbiased sample of the case population. We later describe how this assumption can be relaxed.

The values Z_a and Z_d are defined as

$$Z_d = Z_{n_1,n_2}(m_1, m_2) \tag{2.7}$$

$$Z_a = Z_{n_1+n_2,n_c}(m_{12}, m_c) \tag{2.8}$$

2.1.2 Adjustment for strata

If the distribution of some categorical variable (for example, country of origin) associated with allele frequency varies systematically between x and y, stratification may be needed when computing GWAS p-values. This may mean that $E(m_x) \neq E(m_y)$, even if the expected allele frequency is the same in x and y in each stratum.

Assume x is divided into k strata $1..k$, and let n_x^1, n_x^2,..., n_x^k be the number of samples, m_x^1, m_x^2,..., m_x^k the observed allele frequencies and μ_x^1, μ_x^2,..., μ_x^k the expected allele frequencies for a SNP of interest in each stratum (and analagously for y).

We assume the Z score $Z_{\{n_x\},\{n_y\}}(\{m_x\},\{m_y\})$ in this case is a smooth function of $\{n_x^i\}$, $\{n_y^i\}$, $\{m_x^i\}$, $\{m_y^i\}$, which has a first-order expansion about μ^1, μ^2, ... μ^k of the form

$$Z_{\{n_x\},\{n_y\}}(\{m_x\},\{m_y\}) = \frac{1}{\sqrt{\sum_{i\in 1..k} k_i^2 \frac{n_x^i+n_y^i}{2n_x^i n_y^i}\mu^i(1-\mu^i)}} \sum_{i\in 1..k} k_i(m_x^i - m_y^i) + O\left(\sum_{i\in 1..k} k_i(m_x^i - m_y^i)^2\right) \tag{2.9}$$

$$\approx \frac{\sum k_i(m_x^i - m_y^i)}{\sqrt{var\left(\sum k_i(m_x^i - m_y^i)|\mu_x^i = \mu_y^i\right)}} + O\left(\sum k_i(m_x^i - m_y^i)^2\right) \tag{2.10}$$

where coefficients k_i depend only on the values $\{n_x\}$, $\{n_y\}$. For example, if the Cochran-Mantel-Haenszel test is used, $k_i = \frac{2n_x^i n_y^i}{n_x+n_y}$.

Using analogous definitions to section 2.1.1, we now define

$$Z_d = Z_{\{n_1\},\{n_2\}}(\{m_1\},\{m_2\}) \tag{2.11}$$

$$Z_a = Z_{\{n_1+n_2\},\{n_c\}}(\{m_{12}\},\{m_c\}) \tag{2.12}$$

We term the coefficients of the allelic differences $m_1^i - m_2^i$, $m_{12}^i - m_c^i$ in the decomposition of Z_d and Z_a above as k_{di}, k_{ai} respectively.

2.1.3 Adjustment for covariates

If the distribution some continuous confounder associated with allele frequency (for example, height) has a systematically different distribution in x and y, adjustment for covariates may be needed when computing GWAS p-values

We set $G(i)$ as the numerical genotype of sample i (0,1,or 2) and w_i as the covariate value(s) for individual i. We consider w_i to be a sample from a random variable Z with *pdf* f_x in x and f_y in y.

We define the Z score $Z_{x,y}(\{G\},\{w\})$ in this case as a function of observed genotypes which permits a first-order expansion

$$Z_{x,y}(\{G\},\{w\}) = \frac{1}{\sqrt{\bar{m}(1-\bar{m})}}\left(\sum_{i\in x} h_x(w_i)G(i) - \sum_{j\in y} h_y(w_j)G(j)\right) \tag{2.13}$$

where h_x and h_y are functions of covariate scores, depending on the distribution of w in x and y and the relative sizes of n_x and n_y, and parameter $\bar{m}$ is some measure of the overall allele frequency.

The coefficients $h_x(w_i)$, $h_y(w_i)$ can be considered to be 'normalising' the contribution of genotype i to the Z score according to the relative density of covariate w_i in x and y. If the density of some weight w_0 in x is lower than the density in y, then $h_x(w_0)$ should be greater than $h_y(w_0)$ to compensate for this. Indeed, we show that this has to be the case.

The expected genotype of an individual may depend on their covariate value; for an individual i with covariate value(s) w_i in x set $g_x(w_i) = E(G(i))$, and set g_y similarly. Under the null hypothesis, $g_x \equiv g_y$, and the expectation of Z must be 0. We can write the expectation of $Z_{x,y}(\{G\},\{w\})$ as an integral over the domain of w; namely

$$\begin{aligned}
E(\sqrt{\bar{m}(1-\bar{m})}Z_{x,y}(\{G\},\{w\})) &= E\left(\sum_{i\in x} h_x(w_i)G(i) - \sum_{j\in y} h_y(w_j)G(j)\right) \\
&= \sum_{i\in x} h_x(w_i)E(G(i)) - \sum_{j\in y} h_y(w_j)E(G(j)) \\
&= \sum_{i\in x} h_x(w_i)g_x(w_i) - \sum_{j\in y} h_y(w_j)g_y(w_j) \\
&\to n_x \int_{\mathbb{D}(w)} h_x(w)f_x(w)g_x(w)dw - n_y \int_{\mathbb{D}(w)} h_y(w)f_y(w)g_y(w)dw \\
&= \int_{\mathbb{D}(w)} g_x(w)\left(n_x h_x(w)f_x(w) - n_y h_y(w)f_y(w)\right)dw \qquad (2.14)
\end{aligned}$$

Since this must hold for all SNPs and thus for any well-behaved function g_x, we must have

$$n_x h_x f_x \equiv n_y h_y f_y \qquad (2.15)$$

This arises intuitively if we consider adjustment for covariates analogously to adjusting for strata We can rewrite equation 2.9 summing over samples rather than strata (defining $S(i)$ as the stratum of individual i):

$$Z_{\{n_1\},\{n_2\}}(\{m_x\},\{m_y\}) \propto \frac{1}{\sqrt{\bar{m}(1-\bar{m})}}\left(\sum_{i\in x}\frac{c_{S(i)}}{2n_x^i}G(i) - \sum_{j\in y}\frac{c_{S(j)}}{2n_y^j}G(j)\right) + O\left(\sum(m_x^i - m_y^i)^2\right) \qquad (2.16)$$

The values $\frac{c_{S(i)}}{2n_x^i}$, $\frac{c_{S(j)}}{2n_x^j}$ can be considered to be 'normalising' the distribution of strata across x and y by multiply-counting certain individuals in under-represented strata and under-counting individuals in over-represented strata. This is analagous to 'normalising' the contribution of $G(i)$ by h_x according to the population prevalence of covariate value z_i; that is, $f_x(z_i)$.

The sums of genotypes on the right of equation 2.13 can be considered as 'effective' allele frequencies, and we define

$$\begin{aligned}
m'_x &= \sum_{i\in x} h_x(z_i)G(i) \\
m'_y &= \sum_{j\in y} h_y(z_j)G(j) \qquad (2.17)
\end{aligned}$$

with expected values μ'_x, μ'_y respectively. We define 'effective' sample sizes $n'_x = \frac{\mu'_x(1-\mu'_x)}{var(m'_x)}$, $n'_y = \frac{\mu'_y(1-\mu'_y)}{var(m'_y)}$ so that, like allele frequencies, and under appropriate assumptions on the forms of f_x, f_y, g_x, g_y:

$$\frac{m'_x - \mu'_x}{\sqrt{\frac{\mu'_x(1-\mu'_x)}{n'_x}}} \to_d N(0,1) \qquad (2.18)$$

and similarly for m'_y.

We now define

$$\begin{aligned}
Z_d &= Z_{\text{case 1,case 2}}(\{G\},\{w\}) \qquad (2.19) \\
Z_a &= Z_{\text{cases,controls}}(\{G\},\{w\}) \qquad (2.20)
\end{aligned}$$

2.2 Z_d and Z_a are conditionally independent in categories 1 and 2

2.2.1 Unstratified or stratified groups

For SNPs in categories 1 and 2, $\mu_1 = \mu_2$. Hence

$$\begin{aligned} cov(Z_d, Z_a) &\propto cov(m_{12} - m_c, m_1 - m_2) \\ &= cov\left(\frac{n_1 m_1 + n_2 m_2}{n_1 + n_2} - m_c, m_1 - m_2\right) \\ &= \frac{1}{n_1 + n_2}\left(cov(n_1 m_1, m_1) - cov(n_2 m_2, m_2)\right) \\ &= \frac{1}{n_1 + n_2}\left(\mu_1(1-\mu_1) - \mu_2(1-\mu_2)\right) \end{aligned} \tag{2.21}$$

which is 0 under H_0 in categories 1 and 2.

For stratified groups, the same holds for each stratum; that is, $cov(m_{12}^i - m_c^i, m_1^i - m_2^i) = 0$. The independence of Z_d and Z_a follows from the expression of Z_d and Z_a as proportional to sums of allelic differences within strata and independence of the allelic differences in each stratum.

2.2.2 Adjustment for covariates

If we are adjusting for covariates, since $E(Z_d) = E\left(\sum_{i \in c1} h_1(w_i)G(i) - \sum_{j \in c2} h_2(w_j)G(j)\right) = 0$, we have

$$\begin{aligned} cov(Z_d, Z_a | \mu_1' = \mu_2') &\propto cov\left(\sum_{j \in cases} h_{12}(w_j)G(j) - \sum_{i \in ctl} h_c(w_i)G(i), \sum_{i \in c1} h_1(w_i)G(i) - \sum_{j \in c2} h_2(w_j)G(j)\right) \\ &= cov\left(\sum_{i \in c1} h_{12}(w_i)G(i) + \sum_{j \in c2} h_{12}(w_j)G(j), \sum_{i \in c1} h_1(w_i)G(i) - \sum_{j \in c2} h_2(w_j)G(j)\right) \\ &= E\left(\sum_{i \in c1} h_{12}(z_i)h_1(w_i)G(i)(1-G(i)) - \sum_{j \in c2} h_{12}(z_i)h_2(w_j)(G(j)(1-G(j)))\right) \\ &\to \int_{\mathbb{R}} h_{12}(w)\left(n_1 h_1(w) f_1(w) - n_2 h_2(w) f_2(w)\right) g_1(w)(1-g_1(w))dw \\ &= 0 \end{aligned} \tag{2.22}$$

The cancellations are possible because genotypes vary independently in each group; in the second line, $\sum_{i \in ctl} h_c(w_i)G(i) \perp\!\!\!\perp \sum_{i \in c1} h_1(w_i)G(i), \sum_{j \in c2} h_2(w_j)G(j)$, and in the third line, $\sum_{i \in c1} h_{12}(w_i)G(i) \perp\!\!\!\perp \sum_{j \in c2} h_2(w_j)G(j)$ and $\sum_{j \in c2} h_{12}(w_j)G(j) \perp\!\!\!\perp \sum_{i \in c1} h_1(w_i)G(i)$, and $g_1 \equiv g_2$ under H_0. In the fourth line, $n_1 h_1(w) f_1(w) \equiv n_2 h_2(w) f_2(w)$.

2.3 SNPs in category 3

Under H_0, SNPs in category 3 have the same allele frequency in cases and controls but different population allele frequencies between subgroups. Such a set may arise if subgrouping is based on some partially genetically-determined trait which is independent of the main phenotype has the same prevalence in case and control groups. An example may be subgroups defined by heterogeneity in treatment response arising only from individual pharmacokinetic variation. Under this assumption, the marginal variance of the joint distribution of Z_d, Z_a in the direction of Z_a is 1, and Z_d, Z_a are uncorrelated.

Under H_1 we expect SNPs in category 3 to be associated both with case/control status and with subgroup status. We therefore expect the marginal variances of the joint distribution to be greater than 1 in both the Z_a and Z_d directions, and possible correlation/anticorrelation between Z_a and Z_d.

Define $\zeta(\mu_x, \mu_y)$ as the population normalised log odds ratio between μ_x and μ_y:

$$\begin{aligned}\zeta(\mu_x, \mu_y) &= \sqrt{\bar{\mu}(1-\bar{\mu})}\, log\left(\frac{\mu_x(1-\mu_y)}{\mu_x(1-\mu_y)}\right) \\ &= \frac{\mu_x - \mu_y}{\sqrt{\bar{\mu}(1-\bar{\mu})}} + O\left((\mu_x-\mu_y)^2\right)\end{aligned} \tag{2.23}$$

where $\bar{\mu} = \frac{1}{2}(\mu_x + \mu_y)$. For a set of SNPs of interest, we consider μ_1, μ_2, μ_c to be distributed such that $\zeta_d = \zeta(\mu_1, \mu_2)$ and $\zeta_a = \zeta(\mu_{12}, \mu_c)$ can be considered to be random variables with joint *pdf*:

$$F_{\sigma_a^2, \sigma_d^2, \rho_0} = \frac{1}{2}\left(N_{\left(\binom{0}{0}, \begin{pmatrix}\sigma_d^2 & \rho_0 \\ \rho_0 & \sigma_a^2\end{pmatrix}\right)} + N_{\left(\binom{0}{0}, \begin{pmatrix}\sigma_d^2 & -\rho_0 \\ -\rho_0 & \sigma_a^2\end{pmatrix}\right)}\right) \tag{2.24}$$

with σ_d, σ_a, and ρ_0 independent of n_1, n_2, n_c. Under H_0, $\sigma_a = 0$ (same MAFs in cases/controls) and $\rho_0 = 0$. We assume that ζ_d and ζ_a are conserved across strata and covariates.

2.3.1 Unstratified groups

Combining equation 2.2 with the first-order expansion of equation 2.23 about $\bar{\mu}$:

$$\begin{aligned}\zeta(\mu_x, \mu_y) &= \frac{\mu_x - \mu_y}{\sqrt{\bar{\mu}(1-\bar{\mu})}} + O\left((\mu_x-\mu_y)^2\right) \\ &\approx \sqrt{\frac{n_x+n_y}{2n_x n_y}} Z_{n_x,n_y}(\mu_x, \mu_y)\end{aligned} \tag{2.25}$$

so defining $\bar{\mu}_d = \frac{1}{2}(\mu_1 + \mu_2)$ and $\bar{\mu}_a = \frac{1}{2}(\mu_{12} + \mu_c)$, we note (defining c_a and c_d):

$$\begin{aligned}E(Z_d|\bar{\mu}_d, \zeta_d) &= E(Z_d|\mu_1, \mu_2) \\ &= Z_{n_1,n_2}(\mu_1, \mu_2) \\ &= \sqrt{\frac{2n_1n_2}{n_1+n_2}}\zeta_d \\ &\stackrel{\text{def}}{=} c_d\, \zeta_d \\ E(Z_a|\bar{\mu}_a, \zeta_a) &= \sqrt{\frac{2n_{12}n_c}{n_{12}+n_c}}\zeta_a \\ &\stackrel{\text{def}}{=} c_a\, \zeta_a\end{aligned} \tag{2.26}$$

Set $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_c)$. Since m_1, m_2 and m_c are conditionally independent given $\boldsymbol{\mu}$ we have

$$\begin{aligned}cor(Z_a, Z_d|\boldsymbol{\mu}) &= cor(m_{12} - m_c, m_1 - m_2|\boldsymbol{\mu}) \\ &= \frac{cov(\frac{n_1m_1+n_2m_2}{n_1+n_2} - m_c, m_1 - m_2|\boldsymbol{\mu})}{\sigma(m_s - m_c|\boldsymbol{\mu})\sigma(m_1-m_2|\boldsymbol{\mu})} \\ &= \frac{cov(n_1m_1, m_1|\boldsymbol{\mu}) - cov(n_2m_2, m_2|\boldsymbol{\mu})}{(n_1+n_2)\sigma(m_s - m_c|\boldsymbol{\mu})\sigma(m_1-m_2|\boldsymbol{\mu})} \\ &= \frac{\mu_1(1-\mu_1) - \mu_2(1-\mu_2)}{(n_1+n_2)\sigma(m_s - m_c|\boldsymbol{\mu})\sigma(m_1-m_2|\boldsymbol{\mu})} \\ &\approx 0\end{aligned}$$

From equation 2.6, $var(Z_d|\mu_1,\mu_2) = var(Z_a|\mu_1,\mu_2,\mu_c) = 1$. Thus approximately:

$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} |\zeta_d, \zeta_a \sim N\left(\begin{pmatrix} c_d\,\zeta_d \\ c_a\,\zeta_a \end{pmatrix}, I_2\right) \tag{2.27}$$

and the *pdf* of $(Z_a\ Z_d)^T$ at (x,y) has value

$$\begin{aligned} &\iint_{\mathbb{R}^2} N_{(c_d\zeta_d\ c_a\zeta_a)^T, I_2}(x,y) F_{\sigma_a^2,\sigma_d^2,\rho_0}(\zeta_d,\zeta_a)\ d\zeta_d\ d\zeta_a \\ =&F_{1+c_a^2\,\sigma_a^2, 1+c_d^2\,\sigma_d^2, c_a c_d \rho_0}(x,y) \\ =&\frac{1}{2}\left(N_{\left(\binom{0}{0}, \begin{pmatrix} 1+c_d^2\zeta_d^2 & c_a c_d \rho_0 \\ c_a c_d \rho_0 & 1+c_a^2\zeta_a^2 \end{pmatrix}\right)}(x,y) + N_{\left(\binom{0}{0}, \begin{pmatrix} 1+c_d^2\zeta_d^2 & -c_a c_d \rho_0 \\ -c_a c_d \rho_0 & 1+c_a^2\zeta_a^2 \end{pmatrix}\right)}(x,y)\right) \end{aligned} \tag{2.28}$$

which is a symmetric two-Gaussian distribution. Under H_0, the marginal variance in the direction of Z_a (fitted σ_3^2) is 1, and the covariance between Z_d and Z_a is zero.

2.3.2 Adjustment for strata

For stratified groups, we assume ζ_a and ζ_d are conserved across strata, and set $\bar{\mu}_d^i = \frac{1}{2}(\mu_1^i + \mu_2^i)$, $\bar{\mu}_a^i = \frac{1}{2}(\mu_{12}^i + \mu_c^i)$, k_{di} as the coefficient of $m_1^i - m_2^i$ in the first-order expansion of Z_d (equation 2.9), and k_{ai} as the coefficient of $m_{12}^i - m_c^i$ in the first-order expansion of Z_a to find

$$\begin{aligned} E(Z_d|\{\bar{\mu}_d\},\zeta_d) &= E(Z_d|\{\bar{\mu}_1\},\{\bar{\mu}_2\}) \\ &\approx \frac{1}{\sqrt{\sum_{i\in 1..k} k_{di}^2 \frac{n_1^i+n_2^i}{2n_1^i n_2^i}\bar{\mu}_d^i(1-\bar{\mu}_d^i)}} \sum_{i\in 1..k} k_{di}(\mu_1^i - \mu_2^i) \\ &\approx \frac{\sum k_{di}\sqrt{\bar{\mu}_d^i(1-\bar{\mu}_d^i)}}{\sqrt{\sum k_{di}^2 \frac{n_1^i+n_2^i}{2n_1^i n_2^i}\bar{\mu}_d^i(1-\bar{\mu}_d^i)}}\zeta_d \\ &\approx \frac{\sum k_{di}}{\sqrt{\sum k_{di}^2 \frac{n_1^i+n_2^i}{2n_1^i n_2^i}}}\zeta_d \\ &\stackrel{\text{def}}{=} c_d'\zeta_d \end{aligned} \tag{2.29}$$

and

$$\begin{aligned} E(Z_a|\{\bar{\mu}_a^1,\bar{\mu}_a^2,...,\bar{\mu}_a^k\},\zeta_a) &\approx \frac{\sum k_{ai}}{\sqrt{\sum k_{ai}^2 \frac{n_{12}^i+n_c^i}{2n_{12}^i n_c^i}}}\zeta_a \\ &\stackrel{\text{def}}{=} c_a'\zeta_a \end{aligned} \tag{2.30}$$

assuming that for most SNPs the values $\bar{\mu}_d^i$, $\bar{\mu}_a^i$ do not differ markedly across strata. If the Cochran-Mantel-Haenszel test is used,

$$\begin{aligned} c_d' &= \sqrt{\sum k_{di}} \\ &= \sqrt{\sum \frac{2n_1^i n_2^i}{n_1^i + n_2^i}} \\ c_a' &= \sqrt{\sum \frac{2n_{12}^i n_c^i}{n_{12}^i + n_c^i}} \end{aligned} \tag{2.31}$$

and the *pdf* of Z_d, Z_a is then as for equation 2.28 with c_d', c_a' in place of c_d, c_a.

2.3.3 Adjustment for covariates

The expression for $Z_{x,y}(\{G\},\{w\})$ can be rewritten as:

$$Z_{x,y}(\{G\},\{w\}) = \frac{1}{\sqrt{\bar{m}(1-\bar{m})}}(m'_x - m'_y) \tag{2.32}$$

We define the analog of $\zeta(\mu_x,\mu_y)$ given covariate(s) w

$$\zeta(\mu_x,\mu_y)|w = \sqrt{\bar{\mu}(w)(1-\bar{\mu}(w))}\, log\left(\frac{\mu'_x(w)(1-\mu_y(w))}{\mu_x(w)(1-\mu_y(w))}\right) \tag{2.33}$$

and assume that this is independent of w; that is, the effect size is conserved with respect to the covariate. The joint distribution of Z_d and Z_a is then given by the analog of equation 2.28 with appropriate analogues of c_d, c_a.

2.4 Unequal subgroup prevalences

2.4.1 Motivation

The criteria by which subgroups are defined may have a different distribution in the population than in the case group, with the consequence that the disease subtype corresponding to one of the subgroups may be oversampled relative to its true prevalence in the population.

This leads to inaccuracies in the inferred genetic architecture recovered from a case-control study (ie, a typical GWAS), which may take the form of false-positive associations. If there exist variants which differentiate subgroups, oversampling of one subgroup will bias the the observed overall variant effect sizes toward the effect size in the oversampled subgroup, even if the variants are unassociated with the phenotype overall.

In serious cases, this could lead to false identification of variants associated only with subgroup status as associated with the disease as a whole. For example, a GWAS on rheumatoid arthritis (RA) in which the case group had a high prevalence of obesity may identify purely obesity-associated variants as RA-associated.

For stratified and covariate-adjusted analyses, the equivalent problem is failure of population subgroup prevalences to match study subgroup prevalences within each strata or across covariates. This could be a result of ascertainment bias; different geographic locations could report different frequencies of disease subtypes due to differences in clinic specialties.

As well as affecting conventional GWAS analyses, we show below that subgroup oversampling can cause false-positives in our test. We provide a modification to our method to account for this.

2.4.2 Behaviour of standard approach

We mathematically demonstrate the effect of mismatched sample and population subgroup frequencies in the scenario where no strata or covariates are used. The extension to the generalised cases is similar.

Assume that in the disease population, the 'true' prevalences of subgroups 1 and 2 are ν, $1-\nu$, and define $\mu_{12} = \nu\mu_1 + (1-\nu)\mu_2$ as the underlying MAF across all cases in the population. In the hypothesis test to compute P_a, the hypothesis $H_a : \mu_c = \frac{n_1 m_1 + n_2 m_2}{n_1+n_2}$ is not equivalent to $H : \mu_c = \mu_{12}$.

Since $E(m_{12}) = E\left(\frac{n_1 m_1 + n_2 m_2}{n_1+n_2}\right) = \frac{n_1\mu_1+n_2\mu_2}{n_1+n_2} \neq \mu_{12}$, equation 2.27 becomes

$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} |\boldsymbol{\mu} \sim N\left(\begin{pmatrix} Z_{n_1,n_2}(\mu_1,\mu_2) \\ Z_{n_{12},n_c}(\frac{n_1\mu_1+n_2\mu_2}{n_1+n_2},\mu_c) \end{pmatrix}, I_2\right) \tag{2.34}$$

Now

$$Z_{n_{12},n_c}\left(\frac{n_1\mu_1+n_2\mu_2}{n_1+n_2},\mu_c\right)\approx\frac{c_a}{\sqrt{\bar{\mu}(1-\bar{\mu})}}\left(\frac{n_1\mu_1+n_2\mu_2}{n_1+n_2}-\mu_c\right)$$
$$=\frac{c_a}{\sqrt{\bar{\mu}(1-\bar{\mu})}}\left((\mu_{12}-\mu_c)+\left(\frac{n_1}{n_1+n_2}-\nu\right)(\mu_1-\mu_2)\right)$$
$$\approx c_a(\zeta_a+k\zeta_d) \tag{2.35}$$

where $k=(\frac{n_1}{n_1+n_2}-\nu)$, so the unconditional distribution of $(Z_a\ Z_d)^T$ in this case is given by

$$\iint_{\mathbb{R}^2} N_{(c_d\zeta_d\ c_a(\zeta_a+c\zeta_d))^T,I_2}(x,y)F_{\sigma_a^2,\sigma_d^2,\rho_0}(\zeta_d,\zeta_a)\ d\zeta_d\ d\zeta_a$$
$$=\frac{1}{2}\left(N_{(\binom{0}{0},\sigma_2)}(x,y)+N_{(\binom{0}{0},\sigma_3)}(x,y)\right) \tag{2.36}$$

where

$$\sigma_2=\begin{pmatrix}1+c_d^2\zeta_d^2 & c_ac_d(\rho_0+k\zeta_d^2)\\ c_ac_d(\rho_0+k\zeta_d^2) & 1+c_a^2(\zeta_a^2+k^2\zeta_d^2+2k\rho_0)\end{pmatrix}$$
$$\sigma_3=\begin{pmatrix}1+c_d^2\zeta_d^2 & c_ac_d(-\rho_0+k\zeta_d^2)\\ c_ac_d(-\rho_0+k\zeta_d^2) & 1+c_a^2(\zeta_a^2+k^2\zeta_d^2-2k\rho_0)\end{pmatrix} \tag{2.37}$$

Distribution 2.36 consists of the sum of two Gaussians which are not mirror images in the x and y axes. Conceptually, the aberrance between prevalences of subgroups in the population and in the study induces a bias in Z_a toward either Z_1 or Z_2, whichever is comparatively over-represented in the study compared to the population.

This effect is demonstrated in figure 2.1, with simulated data and approximate distribution as per 2.36. As the discrepancy between the relative proportions grows, the distributions precess around the origin. Importantly, under H_0 ($\sigma_a=0$, $\rho_0=0$) the distribution of Z_d, Z_a will not satisfy $\sigma_3=1$, $\rho=0$, and our standard approach is inappropriate.

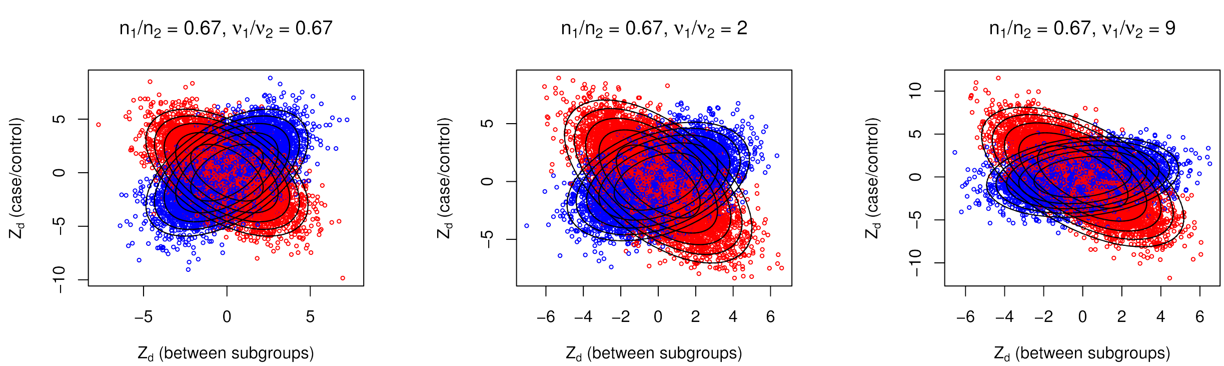


Figure 2.1: Distribution of (Z_a, Z_d) for SNPs in category 3 when observed subgroup frequency (n_1/n_2) does not match underlying subgroup frequency in the population $(\nu_1/\nu_2=\nu/(1-\nu))$. Red and blue points correspond to the two Gaussian distributions comprising the underlying distribution of effect sizes. Contour lines of distributions are shown. Note the precession in the axes of the distributions as the difference between ν_1/ν_2 and n_1/n_2 increases, and loss of symmetry when $\nu_1/\nu_2\neq n_1/n_2$

2.4.3 Adaptation

If the true proportion of case subgroups in the population are known, the problem of oversampled subgroups can be overcome by a recalculation of Z_a. The problem broadly arises because the expected value of the observed allele frequency in cases, $E(m_{12})$, is different from the true allele frequency μ_{12} in cases in the population, for SNPs in category 3.

This can be addressed by using an unbiased estimate of the true population allele frequency $m'_{12} = \nu m_1 + (1-\nu)m_2$ in place of m_{12}. The resultant Z score, Z'_a, is obtained by adjusting Z_a by subtracting a multiple of Z_d:

$$Z'_a = \frac{1}{\sqrt{1+\beta^2}}\left(Z_a - \beta Z_d\right) \tag{2.38}$$

so, given a between-subgroup effect size ζ_d, $var(Z'_a|\zeta_d) = 1$. We choose β so that $E(Z'_a) = 0$ for SNPs in category 3 (see below). The adjustment leads to systematic covariance between Z_d and Z'_a.

Z_a and Z_d are independent conditioned on ζ_a, ζ_d and $\bar{\mu}$. Thus under H_0 and conditioning on $\bar{\mu}$

$$\begin{aligned}
cov(Z'_a, Z_d|\zeta_a, \zeta_d) &= \frac{1}{\sqrt{1+\beta^2}} E(Z_d(Z_a - \beta Z_d)|\zeta_a, \zeta_d) \\
&= \frac{1}{\sqrt{1+\beta^2}}\left(E(Z_d Z_a|\zeta_a, \zeta_d) - \beta E(Z_d^2|\zeta_a, \zeta_d)\right) \\
&= \frac{-\beta}{\sqrt{1+\beta^2}} var(Z_d^2|\zeta_d) \\
&= \frac{-\beta}{\sqrt{1+\beta^2}} \qquad (2.39)
\end{aligned}$$

and because ζ_d and ζ_a are independent under H_0, $cov(Z'_a, Z_d) = \frac{-\beta}{\sqrt{1+\beta^2}}$ in every category. We denote this consistent covariance by ρ_c

Hence the overall model for Z_d, Z_a changes to

$$\begin{aligned}
PDF_{Z_d,Z_a|\Theta}(d,a) =& \pi_1 N_{\begin{pmatrix} 1 & \rho_c \\ \rho_c & 1 \end{pmatrix}}(d,a) && \text{(category 1)} \\
&+ \pi_2 N_{\begin{pmatrix} 1 & \rho_c \\ \rho_c & \sigma_2^2 \end{pmatrix}}(d,a) && \text{(category 2)} \\
&+ \pi_3 \left(\frac{1}{2} N_{\begin{pmatrix} \tau^2 & \rho+\rho_c \\ \rho+\rho_c & \sigma_3^2 \end{pmatrix}}(d,a) + \frac{1}{2} N_{\begin{pmatrix} \tau^2 & -\rho+\rho_c \\ -\rho+\rho_c & \sigma_3^2 \end{pmatrix}}(d,a)\right) && \text{(category 3)} \qquad (2.40)
\end{aligned}$$

where, under H_0, $\rho = 0$ and $\sigma_3 = 1$. This requires a slight modification of the fitting algorithm. Our R package at `https://github.com/jamesliley/subtest` contains an implementation.

2.4.4 No adjustment - unbiased sampling

If no strata nor covariates are used, we set

$$\begin{aligned}
\beta &= \left(\frac{n_1}{n_1+n_2} - \nu\right)\frac{c_a}{c_d} \\
&\stackrel{\text{def}}{=} k\frac{c_a}{c_d} \qquad (2.41)
\end{aligned}$$

recalling the definitions of c_a and c_d from equation 2.26, and that ν is the proportion of cases of subgroup 1 in the population while $\frac{n_1}{n_1+n_2}$ is the proportion in the study. The value $k = \left(\frac{n_1}{n_1+n_2} - \nu\right)$ thus corresponds to the dissimilarity between subgroup prevalences in the case group and in the population.

Under H_0, for SNPs in category 3 we have

$$
\begin{aligned}
E(Z'_a|\zeta_d) &\propto E\left(Z_a - k\frac{c_a}{c_d}Z_d\right) \\
&= \frac{c_a}{\sqrt{\bar{m}(1-\bar{m})}}E\left(\left(\frac{n_1m_1+n_2m_2}{n_1+n_2} - m_c\right) - \left(\frac{n_1}{n_1+n_2} - \nu\right)(m_1-m_2)\right) \\
&= \frac{c_a}{\sqrt{\bar{m}(1-\bar{m})}}E\left(\nu m_1 + (1-\nu)m_2 - m_c\right) \\
&= 0
\end{aligned}
\tag{2.42}
$$

since $E(\nu m_1 + (1-\nu)m_2) = \nu\mu_1 + (1-\nu)\mu_2 = \mu_c = E(m_c)$ for all SNPs under H_0.

2.4.5 Adjustment for strata

In the equivalent adjustment for stratified groups, we define

$$
\beta = \sqrt{\frac{\sum k_{di}^2 \frac{n_1^i+n_2^i}{2n_1^in_2^i}}{\sum k_{ai}^2 \frac{n_{12}^i+n_c^i}{2n_{12}^in_c^i}}}\frac{\sum k_{ai}\left(\frac{n_1^i}{n_1^i-n_2^i} - \nu\right)}{\sum k_{di}}
\tag{2.43}
$$

so, assuming $\mu_1^i - \mu_2^i$ are conserved and $\bar{\mu}_a^i$, $\bar{\mu}_d^i$ are close to conserved across strata, and given $\bar{\mu}_a^i \approx \bar{\mu}_d^i|H_0$:

$$
\begin{aligned}
E(Z'_a|H_0) &= \frac{\sum k_{ai}(\mu_{12}^i - \mu_c^i)}{\sqrt{\sum k_{ai}^2\frac{n_{12}^i+n_c^i}{2n_{12}^in_c^i}\bar{\mu}_a^i(1-\bar{\mu}_a^i)}} + \beta\frac{\sum k_{di}(\mu_1^i - \mu_2^i)}{\sqrt{\sum k_{di}^2\frac{n_1^i+n_2^i}{2n_1^in_2^i}\bar{\mu}_d^i(1-\bar{\mu}_d^i)}} \\
&\approx \frac{\sum k_{ai}(\mu_{12}^i - \mu_c^i)}{\sqrt{\bar{\mu}_a(1-\bar{\mu}_a)}\sqrt{\sum k_{ai}^2\frac{n_{12}^i+n_c^i}{2n_{12}^in_c^i}}} + \beta\frac{\sum k_{di}(\mu_1^i - \mu_2^i)}{\sqrt{\bar{\mu}_d(1-\bar{\mu}_d)}\sqrt{\sum k_{di}^2\frac{n_1^i+n_2^i}{2n_1^in_2^i}}} \\
&= \frac{1}{\sqrt{\bar{\mu}_a(1-\bar{\mu}_a)}\sqrt{\sum k_{ai}^2\frac{n_{12}^i+n_c^i}{2n_{12}^in_c^i}}}\left(\sum k_{ai}\left(\frac{n_1^i\mu_1^i+n_2^i\mu_2^i}{n_1^i+n_2^i} - \mu_c^i\right) - \left(\frac{n_1^i}{n_1^i+n_2^i} - \nu\right)(\mu_1^i - \mu_2^i)\right) \\
&= \frac{1}{\sqrt{\bar{\mu}_a(1-\bar{\mu}_a)}\sqrt{\sum k_{ai}^2\frac{n_{12}^i+n_c^i}{2n_{12}^in_c^i}}}\left(\sum k_{ai}((\nu\mu_1^i + (1-\nu)\mu_2^i) - \mu_c^i)\right) \\
&= 0
\end{aligned}
\tag{2.44}
$$

2.4.6 Adjustment for covariates

If covariates are used, we define the functions h_{12}, h_1, f_1, f_2 as per section 2.1.3 and set

$$
\beta = \frac{\int_{\mathbb{D}(w)} h_{12}(w)\left(n_1(1-\nu)f_1(w) - n_2\nu f_2(w)\right)dw}{\int_{\mathbb{D}(w)} n_1h_1(w)f_1(w)dw}
\tag{2.45}
$$

so

$$
\begin{aligned}
E(Z_a') \propto & \sqrt{\bar{\mu}(1-\bar{\mu})} E(Z_a - \beta Z_d) \\
= & \left(\sum_{i \in c1} h_{12}(w_i) G(i) + \sum_{i \in c2} h_{12}(w_i) G(i) - \sum_{i \in controls} h_c(w_i) G(i) \right) \\
& - \beta \left(\sum_{i \in c1} h_1(w_i) G(i) - \sum_{i \in c2} h_2(w_i) G(i) \right) \\
\to & \int_{\mathbb{D}(w)} h_{12}(w) \left(n_1 f_1(w) g_1(w) + n_2 f_2(w) g_2(w) \right) - n_c h_c(w) f_c(w) g_c(w) dw \\
& - \beta \left(g_1(w) - g_2(w) \right) \int_{\mathbb{D}(w)} n_1 h_1(w) f_1(w) dw \\
= & \int_{\mathbb{D}(w)} h_{12}(w) \left(n_1 f_1(w) + n_2 f_2(w) \right) \left(\nu g_1(w) + (1-\nu) g_2(w) \right) - n_c h_c(w) f_c(w) g_c(w) dw \\
= & \int_{\mathbb{D}(w)} n_{12} h_{12}(w) f_{12}(w) \left(\nu g_1(w) + (1-\nu) g_2(w) - g_c(w) \right) dw \\
= & 0
\end{aligned}
\tag{2.46}
$$

since $n_{12}h_{12}f_{12} \equiv n_c h_c f_c$ and $n_1 h_1 f_1 \equiv n_2 h_2 f_2$ from section 2.1.3, $g_1 - g_2$ is constant by assumption, and the expected population genotypes at covariate value w are the same in cases $(\nu g_1(w) + (1-\nu) g_2(w))$ and controls $(g_c(w))$ under H_0.

2.5 Testing procedure

2.5.1 Algorithm

For testing a subgrouping S of interest, we use the following protocol:

1. Compute Z_a scores between cases and controls
2. For the proposed subgrouping S
   (a) Compute scores Z_d^S corresponding to S,
   (b) Fit parameters of full and null models $\Theta_1^S = arg\,max_{\Theta \in H_1} L(Z_d^S, Z_a|\Theta)$, $\Theta_0^S = arg\,max_{\Theta \in H_0} L(Z_d^S, Z_a|\Theta)$
   (c) Compute $uPLR = log\{L(Z_d^S, Z_a|\Theta_1^S)\} - log\{L(Z_d^S, Z_a|\Theta_0^S)\}$ and adjusting factor $f(Z_a|\Theta_1^S, \Theta_0^S) = log\{L(Z_a|\Theta_1^S)\} - log\{L(Z_a|\Theta_0^S)\}$
   (d) Compute $PLR_S = uPLR - f(Z_a|\Theta_1^S, \Theta_0^S)$
3. For > 1000 random subgroups R of the case group
   (a) Compute scores Z_d^* corresponding to R
   (b) Fit parameters $\Theta_1^* = arg\,max_{\Theta \in H_1} L(Z_d^R|Z_a, \Theta)$, $\Theta_0^* = arg\,max_{\Theta \in H_0} L(Z_d^R|Z_a, \Theta)$
   (c) Compute $cPLR = log\{L(Z_d^*|Z_a, \Theta_1^*)\} - log\{L(Z_d^R|Z_a, \Theta_0^*)\}$
4. Estimate parameters γ, κ of the null distribution of $cPLR$ (of the form $\gamma\left(\kappa\chi_1^2 + (1-\kappa)\chi_2^2\right)$), which majorises the null distribution of PLR.
5. Compute p-value for PLR_S using this distribution.

In summary, we compare an adjusted pseudo-log likelihood ratio for a subgrouping of interest to conditional pseudo-log likelihood ratios for randomly-chosen subgroupings.

2.5.2 Rationale

A problem arises with the behaviour of the unadjusted pseudo-log likelihood ratio statistic $uPLR = log\{L(Z_d^S, Z_a|\Theta_1^S)\} - log\{L(Z_d^S, Z_a|\Theta_0^S)$ when the true value of τ (the marginal variance of Z_d in group 3) is near 1, corresponding to an absence of SNPs which differentiate subgroups.

If $\tau = 1$, there can be no differential genetic architecture between the subgroups, as there are no systematic genetic differences between them at all. However, the joint distribution of Z_d, Z_a may still be in H_1; if Z_a has an equally weighted three-Gaussian mixture distribution with variances $1, a^2, b^2$, and $Z_d \sim N(0,1)$, the true parameter values are $(\pi_2, \pi_3, \tau, \sigma_2, \sigma_3, \rho) = (\frac{1}{3}, \frac{1}{3}, 1, a, b, 0) \in H_1 \setminus H_0$ (figure 2.2).

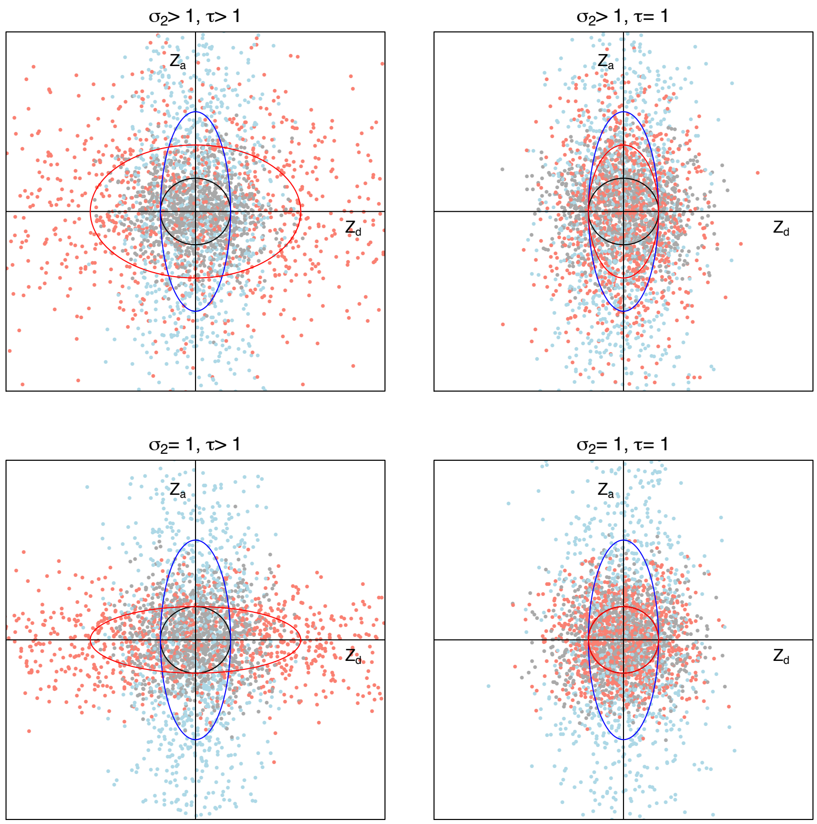


Figure 2.2: Potential for false positives when $\tau = 1$. Black/grey points and contours correspond to category 1, blue to category 2, and red/pink to category 3. Top two figures show potential distributions of Z_d, Z_a with $\sigma_3 > 1$; bottom two figures distributions with $\sigma_3 = 1$. A test based on the unadjusted pseudo-log likelihood ratio $uPLR = log\{L(Z_d^S, Z_a|\Theta_1^S)\} - log\{L(Z_d^S, Z_a|\Theta_0^S)\}$ will reject H_0 for both of the top two scenarios. However, we do not want to reject H_0 for the top right figure, in which $\tau = 1$ (no genetic difference between subgroups). This scenario is possible in real data, as the distribution of Z_a is only approximately normal and may more closely resemble a three-gaussian mixture distribution (where components have variances σ_2^2, σ_3^2 and 1) than a two-Gaussian mixture distribution (where components have variances σ_1^2 and 1).

This problem is particularly prevalent in randomly-chosen subgroups, since $\tau = 1$ by assumption in this case. If the distribution of Z_d, Z_a from a test subgrouping is to be compared against corresponding distributions from random subgroupings, this problem must be addressed.

Consider the function

$$\begin{aligned} K(\boldsymbol{Z},\Theta) &= K(Z_d, Z_a, \pi_1, \pi_2, \tau, \sigma_2, \sigma_3, \rho) \\ &= PL(\boldsymbol{Z}|\Theta) - E\{PL(\boldsymbol{Z}|\Theta \setminus \tau, \tau = 1)\} \\ &= PL(Z_d, Z_a|\Theta) - PL(Z_a|\Theta) + c(Z_a) \end{aligned} \tag{2.47}$$

where $c(Z_a)$ is a constant depending only on the values of Z_a. Because the parameters π_2, σ_2 only describe the distribution of Z_a, we have

$$\begin{aligned} \frac{\partial}{\partial \pi_2} PL(Z_d, Z_a|\Theta) &\approx \frac{\partial}{\partial \pi_2} PL(Z_a|\Theta) \\ \frac{\partial}{\partial \sigma_2} PL(Z_d, Z_a|\Theta) &\approx \frac{\partial}{\partial \sigma_2} PL(Z_a|\Theta) \end{aligned} \tag{2.48}$$

so $\frac{\partial K}{\partial \pi_2} \approx 0$ and $\frac{\partial K}{\partial \sigma_2} \approx 0$, and the value of K changes only slightly with changes in π_2, σ_2. Set

$$\begin{aligned} \Theta_1 &= arg\, max_{\Theta \in H_1} PL(\boldsymbol{Z}|\Theta) \\ \Theta_1^* &= arg\, max_{\Theta \in H_1 | \pi_2 = \hat{\pi_2}, \sigma_2 = \hat{\sigma_2}} PL(\boldsymbol{Z}|\Theta) \end{aligned} \tag{2.49}$$

Under H_0, there is no systematic overlap between SNPs associated with the main phenotype (for which the distribution of effect sizes is parametrised by π_2, σ_2) and with the subgrouping phenotype, so fixing π_2 and σ_2 has minimal effect on the maximum-PL estimates of the other parameters, and hence $K(\boldsymbol{Z}, \Theta_1) \approx K(\boldsymbol{Z}, \Theta_1^*)$. Because

$$K(\boldsymbol{Z}, \Theta_1^*) \leq max_{\Theta \in H_1 | \pi_2 = \hat{\pi_2}, \sigma_2 = \hat{\sigma_2}} K(\boldsymbol{Z}, \Theta) \tag{2.50}$$

we have, setting $\Theta_1^c = arg\, max_{\Theta \in H_1 | \pi_2 = \hat{\pi_2}, \sigma_2 = \hat{\sigma_2}} K(\boldsymbol{Z}, \Theta)$:

$$K(Z, \Theta_1) \leq K(Z, \Theta_1^c) \tag{2.51}$$

Consider the value

$$\Theta_0^c = arg\, max_{\Theta \in H_1 | \pi_2 = \hat{\pi_2}, \sigma_2 = \hat{\sigma_2}} K(\boldsymbol{Z}, \Theta) \tag{2.52}$$

$$= arg\, max_{...} \{PL(Z_d, Z_a|\Theta) - PL(Z_a|\Theta)\} \tag{2.53}$$

Now since σ_3 is fixed at 1 under H_0, and $PL(Z_a|\Theta)$ only depends on π_1, π_3 through the difference between the variances of their associated distribution components (1 and σ_3 respectively), we have

$$PL(Z_a|\Theta) = PL(Z_a|\hat{\pi_2}, \hat{\sigma_2}) \tag{2.54}$$

Thus maximising K in equation 2.52 is analogous to maximising $PL(Z_d, Z_a|\Theta)$. If we choose $\hat{pi_1}$ and $\hat{\sigma_2}$ to be approximately equal to their maximum-PL estimates under H_0, then

$$\begin{aligned} \Theta_0^* &= arg\, max_{\Theta \in H_0 | \pi_2 = \hat{\pi_2}, \sigma_2 = \hat{\sigma_2}} PL(\boldsymbol{Z}|\Theta) \\ &\approx arg\, max_{\Theta \in H_0} PL(\boldsymbol{Z}|\Theta) \\ &= \Theta_1 \end{aligned} \tag{2.55}$$

so $K(\Theta_1) \approx K(\Theta_1^c)$. Thus, under H_0, using equation 2.51

$$\begin{aligned} cPLR &= K(\boldsymbol{Z}, \Theta_1^c) - K(\boldsymbol{Z}, \Theta_0^c) \\ &\geq K(\boldsymbol{Z}, \Theta_1) - K(\boldsymbol{Z}, \Theta_0) \\ &= PLR \end{aligned} \tag{2.56}$$

Under H_0, with $\tau > 1$, the unadjusted PLR (equal to $PL(\boldsymbol{Z}|\Theta_1) - PL(\boldsymbol{Z}|\Theta_2)$) and cPLR both have identical mixture-χ^2 distributions (the scaling factor γ arises from LDAK weights, common to both, and the mixing parameter κ tends to be approximately 1/2). The cPLR has the advantage that the empirical distribution is closely approximated by a consistent mixture-χ^2 distribution for all values of τ. By comparing PLR to this distribution, we produce a conservative test.

Heuristically, contributions to the unadjusted PLR can come from either the distribution of Z_a or the interaction between Z_a and Z_d, and inflation in the unadjusted PLR when $\tau = 1$ arise only from the former. If the former effect is large, the parameters Θ_1 will tend to be values which maximise the former effect, at the expense of the latter. By completely eliminating the former effect, using the adjustment, only this compromised contribution of the latter is allowed to contribute to the adjusted PLR. The distribution is less conservative for larger values of τ, since the presence of SNPs with large Z_a values constricts the fitted distribution of Z_a. By contrast, the values which maximise the $cPLR$ effectively take into account the adjustment for Z_a, and the compromise of the latter effect does not occur.

If we were to use the adjusted $uPLR$ to generate the null distribution using random subgroups, the majorisation of the observed distribution by the mixture-χ^2 may lead to loss of FDR control in test subgroups with $\tau > 1$. However, using the slightly anti-conservative distribution of $cPLR$ to fit the null distribution overcomes this problem. Indeed, some conservatism is desirable when $\tau = 1$ as a double guard against rejecting H_0. The power of $cPLR$ to reject H_0 is, however, somewhat lower than the power of the PLR, so we test using adjusted $uPLR$ and fit the null distribution with $cPLR$.

Note 3

Details of simulations

3.1 Simulations of random genotypes

Firstly, we simulated genotypes at independent SNPs to establish the distributions of PLR and $cPLR$ under H_0 with $\tau = 1$ and $\tau > 1$.

We simulated the following scenarios:

1. (a) (Z_d, Z_a) under H_0 with $\tau = 1$
   (b) (Z_d, Z_a) under H_0 with τ allowed to vary
2. (Z_d, Z_a) under H_1

In each case, Z_a and Z_d were calculated from simulated genotypes at 5×10^4 independent autosomal SNPs in Hardy-Weinberg equilibrium. Because the sample size only affects PLR through the size of the fitted parameters (supplementary material, section 3.3) we fixed the sample size at 2000 controls and 1000 cases of each subgroup and varied the underlying effect size distribution. Larger sample sizes correspond to larger deviations of underlying values of σ_2, σ_3, τ from 1 (table 3.1).

For all simulations, we computed the uPLR and PLR (with adjustment $f(Z_a)$). For scenario 1a ($\tau = 1$, corresponding to random subgroups) we additionally computed the cPLR. Simulations 2 functioned as power calculations; the results from these are shown in the main text.

We tested over values of π_3 from $\{10^{-3}, 10^{-2}, 0.1, 0.2\}$. Values of σ_2, σ_3, τ were chosen corresponding to 97.5% quantiles of odds ratios in $\{1.5, 2, 2.5\}$ for case/control comparison (Z_a) or $\{1, 1.2, 1.5, 2\}$ for between-subgroups comparison (Z_d), table 3.1.Values of ρ were chosen corresponding to correlations in $\{0, 0.1, 0.5\}$.

	97.5% quantile of odds ratios			
n_1, n_2	1.2	1.5	2	2.5
500, 500	1.20	1.75	2.66	3.41
1000, 500	1.25	1.94	3.02	3.89
1000, 1000	1.36	2.27	3.62	4.71

Table 3.1: Approximate expected standard deviations of observed Z scores for given odds-ratio distributions at various study sizes. For instance, if a study had 500 cases of each subgroup, and 95% of 'true' odds ratios (corresponding to population MAFs) for SNPs in category 3 were less than 1.5, the expected value of τ (the standard deviation of Z scores for SNPs in category 3) would be 2.66.

We compared the observed distributions of PLR from simulations 1a,1b with the observed distribution of cPLR from simulation 1a. Q-Q plots are shown in figure 3.1. The distribution of cPLR agrees well with a mixture-χ^2 distribution, as does the distribution of PLR for simulation 1b. The distributions of PLR for simulations 1a,1b are minorised by the distribution of cPLR, more so for simulations 1a ($\tau = 1$), leading to a conservative test overall. Using $cPLR$ to fit a null distribution, and using a significance cutoff $p < 0.05$, leads to a false-discovery rate of 0.048 (95% CI 0.039-0.059) in subgroups with $\tau > 1$ and 0.033 (95% CI 0.022-0.045) in subgroups with $\tau = 1$.

We also show the distribution of unadjusted PLR (uPLR) for simulations 1a and 1b. The distribution for 1a markedly majorises the mixture-χ^2 distribution, and has a very different distribution to that for 1b. Thus, if a test subgroup with $\tau >> 1$ was compared to random subgroups using unadjusted PLR, the test would have very low power to reject H_0. Finally, we plotted the estimated null distribution for all tests of real disease datasets, and found that the empirical distributions of cPLR from random subgroups agreed well with the proposed mixture χ^2 distribution (Supplementary Figures 4a, 4b, 4c).

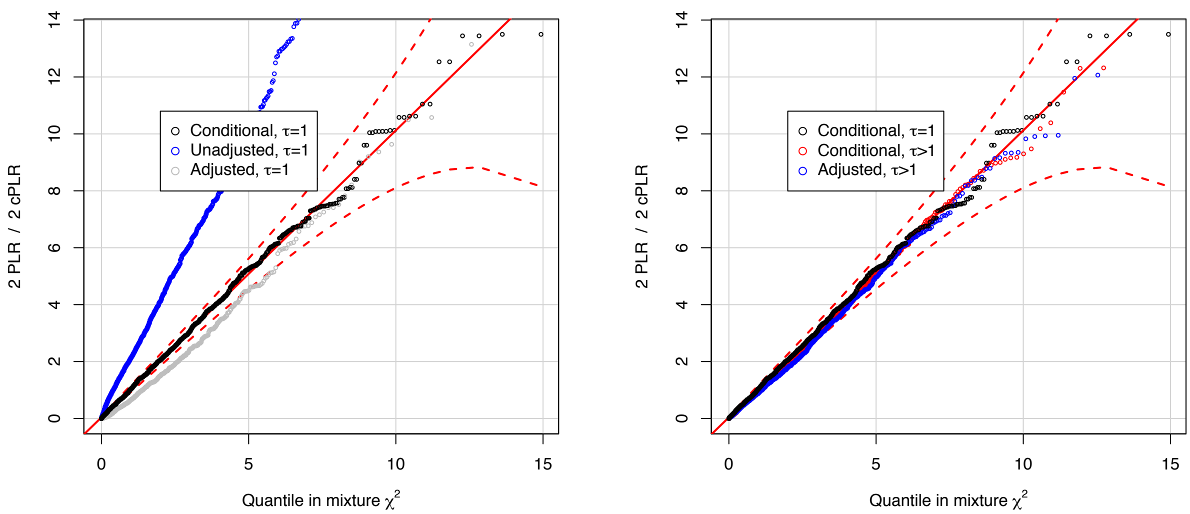


Figure 3.1: Q-Q plots comparing distributions of PLR and $cPLR$ for subgroups based on simulated genotypes with a random variable distributed as $\frac{1}{2}(\chi^2_1 + \chi^2_2)$ (that is, $\gamma = \kappa = \frac{1}{2}$). In both plots, the black points correspond to conditional PLR (cPLR) values for 'random' subgroups ($\tau = 1$). The observed distribution is well-approximated by the asymptotic mixture-χ^2. The left-hand plot shows the distributions of unadjusted and adjusted PLR for subgroups with $\tau = 1$. The distribution of unadjusted PLR markedly majorises the mixture-χ^2, but the adjustment largely fixes this. The right-hand plot compares the distribution of cPLR for random subgroups with PLR for subgroups with $\tau > 1$. The distribution of cPLR is well-approximated by the mixture-χ^2 whether $\tau = 1$ (black) or $\tau > 1$ (red). In both plots, the distribution of cPLR and the mixture-χ^2 distribution slightly majorise the distribution of PLR, leading to a conservative test.

3.2 Simulation on GWAS case group subgroups

To check the extensibility of these results to real data, we performed a similar set of simulations on data generated from subgroups of an ATD case group. In order to simulate scenarios in which $\tau > 1$, we selected subgroups for which groups of ≈ 50 SNPs differentiated subgroups without being associated with the disease in general.

Specifically, we repeatedly polled the overall dataset for sets of 2000 SNPs in linkage equilibrium, then clustered them hierarchically using a Euclidean distance metric. We then chose the first-appearing cluster of 50 SNPs, and hierarchically clustered the individuals in the case group according to a metric based on similarity across the 50 SNPs. When there were two clusters of individuals left, we denoted the two clusters as subgroup 1 and subgroup 2. The mean resultant fitted value of τ was ≈ 5 and standard deviation of fitted values was ≈ 1.5.

For simulated subgroups with $\tau = 1$ (randomly chosen) and with $\tau > 1$ we computed PLR and $cPLR$. As for simulated genotypes, the resultant distributions showed good agreement with the proposed mixture-χ^2 distributions (figure 3.2), with the approximation of the null distribution of PLR with the distribution of $cPLR$ again leading to a conservative test, as expected. The type 1 error rate corresponding to $\alpha = 0.05$ was 0.52 (95% CI 0.043-0.061) in subgroups with $\tau > 1$ and 0.012 (95% CI 0.007-0.016) in subgroups with $\tau = 1$.

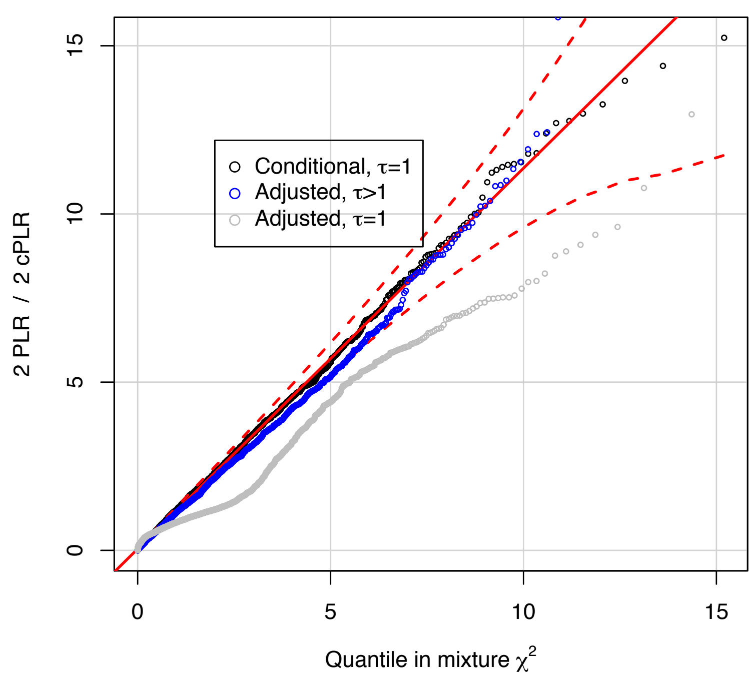


Figure 3.2: Comparison of distributions of PLR and $cPLR$ for subgroups of an ATD case group, chosen so $\tau = 1$ or $\tau > 1$. The distribution of cPLR for random subgroups ($\tau = 1$) and the distribution of PLR for subgroups with $\tau >> 1$ are both well-approximated by a random variable distributed as $\frac{1}{2}(\chi_1^2 + \chi_2^2)$; red dashed lines show 99% pointwise confidence intervals. The distribution of PLR when $\tau = 1$ is minorised by the mixture-χ^2 leading to a conservative test if a subgroup with $\tau = 1$ is tested using PLR against the observed distribution of cPLR for random subgroups. Because $\tau = 1$ implies no genetic difference between subgroups, this is reasonable behaviour for the test.

3.3 Distributions of parameter values for simulation and power calculations

We assume a distribution of summary statistics parametrised by six variables: π_1, π_2, σ_2, σ_3, τ, and ρ (the value of π_3 is determined by π_1 and π_2). The space of all parameter values is too large to meaningfully assess performance of our test across it, so for each simulation, we draw the value of underlying parameters from sets of potential values chosen to reflect values which may arise in real data.

For a SNP S in two groups of size n_1, n_2, denote the population allele frequencies as μ_1, μ_2 and the corresponding observed allele frequencies as m_1, m_2. Set $\mu = \frac{\mu_1 n_1 + \mu_2 n_2}{n_1 + n_2}$ as the overall observed MAF, $r = log\left(\frac{\mu_1(1-\mu_2)}{\mu_2(1-\mu_1)}\right)$ and $R = log\left(\frac{m_1(1-m_2)}{m_2(1-m_1)}\right)$ as the 'underlying' and observed log-odds ratios respectively. To first order

$$\begin{aligned} SE\{R\} &= \sqrt{\frac{1}{2m_1n_1} + \frac{1}{2(1-m_1)n_1} + \frac{1}{2m_2n_2} + \frac{1}{2(1-m_2)n_2}} \\ &\approx \sqrt{\frac{1}{2m(1-m)}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned} \tag{3.1}$$

The observed Z score is, to first order, $Z = \frac{R}{SE(R)}$. Now

$$\begin{aligned} E(Z|\mu, r) &\approx r\sqrt{\frac{2\mu(1-\mu)n_1n_2}{n_1+n_2}} \\ SD(Z|\mu, r) &\approx 1 \end{aligned} \tag{3.2}$$

Consider r as a $N(0, \sigma^2)$ random variable, and fix μ. Now, to first order

$$Z|\mu \sim N\left(0, 1 + \frac{2\mu(1-\mu)\sigma^2 n_1 n_2}{n_1+n_2}\right) \tag{3.3}$$

Assuming μ to have an approximately uniform distribution on $(0, 0.5]$, this gives

$$Z \sim N\left(0, 1 + \frac{\sigma^2 n_1 n_2}{3(n_1+n_2)}\right) \tag{3.4}$$

An interpretable description of the underlying odds-ratio distribution is the 0.975 quantile of 'true' odds ratios (approximately 2 standard deviations). If 97.5% of 'true' odds ratios r fall in $[1/\alpha, \alpha]$, then $\sigma \approx \frac{log(\alpha)}{2}$ and the expected value of the corresponding observed standard deviation of Z (that is, σ_2, σ_3, or τ) is

$$\sqrt{1 + \frac{log(\alpha)^2 n_1 n_2}{12(n_1+n_2)}} \tag{3.5}$$

Some examples are shown in table 3.2:

		Study size (n_1/ n_2)					
α	SD	100/100	100/500	500/500	500/1000	1000/1000	2000/2000
1.1	0.05	1.02	1.03	1.09	1.12	1.17	1.32
1.2	0.09	1.07	1.11	1.30	1.39	1.54	1.94
1.3	0.13	1.13	1.22	1.56	1.71	1.97	2.60
1.5	0.20	1.30	1.46	2.10	2.36	2.80	3.83
2	0.35	1.73	2.08	3.31	3.79	4.58	6.41

Table 3.2: Correspondence between odds-ratio distribution and standard deviation of observed Z score for various study sizes. Column α is the 97.5 % quantile of population odds-ratios for SNPs with non-zero effect sizes (approximately two standard deviations). Column SD is the corresponding standard deviation of the underlying log-odds ratio distribution (assumed to be normal). Entries in the table correspond to expected standard deviations of observed Z scores; that is, σ_2, σ_3 or τ. We allow different odds-ratio distributions between cases and controls for SNPs in categories 2 and 3 (corresponding to σ_2 and σ_3 respectively). For σ_2 or σ_3, n_1 is the number of cases and n_2 the number of controls; for τ; n_1 and n_2 are the number of cases in each disease subgroup.

Note 4

Genetic correlation as an alternative to PLR test

4.1 Overview

The presence of genetic heterogeneity between disease subgroups could be tested for by adapting several known methods, although to our knowledge no specific method has yet been developed. One potential approach is to estimate the narrow-sense genetic correlation (r_g) across a set of SNPs between case/control traits of interest, either between Z scores derived from comparing the control group to each case subgroup, testing under the null hypothesis $r_g = 1$ (method 1); or between the familiar Z_a and Z_d, under the null hypothesis $r_g = 0$ (method 2).

This approach should have the advantage of characterising heterogeneity using a single widely-interpretable metric. However, both methods have, in our naive application, have multiple shortcomings which preclude their general use to subgroup testing. The most important of these are systematic false-positives arising in method 1, and false-negatives arising in method 2. We demonstrate this theoretically and in simulations. In addition, genetic correlation is a signed test statistic; genetic effects in the same direction contribute positively, and opposite directions contribute negatively, causing a loss of power in situations where pleiotropy between the phenotypes involves shared effects of both types. Finally, we found that tests involving r_g were less powerful than the PLR in rejecting the null hypothesis in real genetic data (ATD; GD vs HT).

Genetic correlation is an estimate of the similarity in genetic basis of two traits. A useful formal definition is given by Bulik-Sullivan et al [2]. Let S be a set of SNPs and X denote a vector of additively coded genotypes (0, 1 or 2) for a random individual at the SNPs in S. For traits Y_1, Y_2 set

$$\begin{aligned} \beta &= \arg\max_{\alpha \in \mathbb{R}^{|S|}, ||\alpha||=1} cor(Y_1,\, X^t\alpha) \\ \gamma &= \arg\max_{\alpha \in \mathbb{R}^{|S|}, ||\alpha||=1} cor(Y_2,\, X^t\alpha) \end{aligned} \tag{4.1}$$

where the maximum is taken across the entire population. The genetic correlation between traits across SNPs in S, r_g, is then given by

$$r_g = \frac{\beta^t \gamma}{||\beta||\,||\gamma||} = \sum_{i \in S} \beta_i \gamma_i \tag{4.2}$$

4.2 Method 1: control-subgroup 1 vs control-subgroup 2

4.2.1 Expected behaviour

We firstly consider method 1. In this approach, we consider two case-control comparisons:

1. Case subgroup 1 *vs* control group

2. Case subgroup 2 *vs* control group

We denote Z scores derived from GWAS p-values comparing between controls and subgroup 1 by Z_1 and scores between controls and subgroup 2 by Z_2 (figure 4.1). An estimated genetic correlation significantly less than 1 (or at least significantly less than estimates from random subgroups) may indicate different causative architectures for the subgroups, in the form of differing relative effect sizes for disease-associated variants.

However, using this method will not distinguish between different disease-causative architectures and genetic differences between subgroups unrelated to the overall phenotype. In terms of the parameters of our three-categories model, method 1 will be liable to reject the null whenever $\tau > 1$, regardless of whether $\sigma_3 > 1$ (that is, regardless of whether subgroup-differentiating SNPs are in general disease-associated). Indeed, for a set value of τ, the negative contribution of SNPs in group 3 to the observed r_g will often be maximised when H_0 holds; that is, $\sigma_3 = 1$.

Consider a SNP in category 3. Under a simple model in which case subgroups are the same size, we denote by μ_c the population MAF of the SNP in controls, and μ_1 and μ_2 the population AF of the same allele in cases. To first order $Z_1 \propto \mu_1 - \mu_c$ and $Z_2 \propto \mu_2 - \mu_c$. Assume $\mu_1 - \mu_2$ is set at some constant $m > 0$. Because $m > 0$, the SNP is associated with at least one of the subgroups, and hence contributes to the genetic correlation. The value of this contribution to the correlation is proportional to $Z_1 Z_2$, which is proportional to $(\mu_1 - \mu_c)(\mu_2 - \mu_c)$.

This is minimised when $\mu_c = \frac{1}{2}(\mu_1 + \mu_2)$. This is exactly the scenario in which the genetic subgroup differences are unrelated to the phenotype as a whole. In other words, dividing the case group on an arbitrary genetically-associated phenotype (ie hair colour, ethnicity, presence of a second unrelated disease) would lead to a lowering of r_g *more* than would a differential disease process with the same heritability (figure 4.1).

4.2.2 Simulations

We demonstrated this on our ATD dataset by using the subgroups generated under H_0 as in simulation 1b (see section 3.2). These subgroups had a true value of τ greater than 1, but $\sigma_3 = 1$ and $\rho = 0$.

For each simulated subgroup, we computed the genetic correlation between the two studies using two methods - LD score regression (LDSC) [2] and genome-wide complex trait analysis (GCTA) [3] - and computed our PLR statistic. We also computed genetic correlation and PLR scores for multiple random subgroups of the ATD case group. Significance of the genetic correlation was assessed by either comparing the observed r_g to the values observed in random subgroups (LDSC) or comparing the likelihood of the observed data with an alternative model in which $r_g \equiv 1$.

As expected, r_g estimates using both methods were markedly lower in subgroups with simulated genotypic differences than they were in random subgroups (figure 4.2). In the LDSC method, a cutoff of $p < 0.05$ led to rejecting the null in of 45% (SE 2%) of cases, and in GCTA in in 29% (SE 5%) of cases. The PLR method did not reject the null more often than expected, rejecting the null in 4% (SE 1%) of cases.

4.2.3 Application to real data

We also used both LDSC and GCTA to test the hypothesis of differential genetic architecture in GD and HT. The GCTA method was unable to reject the null hypothesis ($p = 0.217$), using a likelihood ratio test against a null model with $r_g = 1$. The LDSC method was able to reject the null at $p < 0.05$, though not at the same significance as the PLR (LDSC: $p = 0.012$, PLR $p = 2.2 \times 10^{-15}$). This suggests that the r_g based methods are less powerful than the PLR in this context. This is likely due to the PLR responding to an additional degree of freedom (σ_3) between the null and full models.

4.3 Method 2: Z_d (case vs control) vs Z_a (subgroup 1 vs 2)

4.3.1 Expected behaviour, and relation of ρ_g to ρ

In method 2, we consider the two case-control comparisons:

1. Combined case group *vs* control group
2. Case subgroup 1 *vs* case subgroup 2

analogous to our approach in the PLR method, with the two comparisons corresponding to Z_a and Z_d respectively. We estimate r_g between these two traits, and test against the null hypothesis that $r_g = 0$.

The value of r_g relates to the estimated value of ρ_g in our full model. For a set S of disease-associated SNPs with additive (non-epistatic) effects in linkage equilibrium, and a binary trait y, we have

$$cor(y, X^t\alpha) = \sum_{i\in S} cor(y, \alpha_i X_i) = \sum_{i\in S} \alpha_i cor(y, X_i) \tag{4.3}$$

This is maximised when $\alpha_i \propto cor(y, X_i)$. If $\mu_1(i)$ denotes the AF of SNP i in S amongst the population with $y = 1$, $\mu_0(i)$ the corresponding $\mu_c(i)$ the overall AF of SNP i and p the incidence of the trait in the population (that is, $Pr(y = 1)$), we have

$$cor(y, X_i) = \sqrt{2p(1-p)}\frac{\mu_1(i) - \mu_0(i)}{\sqrt{\mu_c(i)(1-\mu_c(i))}} \tag{4.4}$$

Given observed allele frequencies $m_1(i)$, $m_0(i)$ at SNP i in a GWAS between traits 1 and 2 with n_1 and n_0 samples respectively, the Z score for significance of that SNP is

$$\begin{aligned} Z(i) &= \frac{m_1(i) - m_0(i)}{SE(m_1(i) - m_0(i))} + O((m_1(i) - m_0(i))^2) \\ &= \frac{m_1(i) - m_0(i)}{\sqrt{\frac{m_1(i)(1-m_1(i))}{n_1} + \frac{m_0(i)(1-m_0(i))}{n_0}}} + O((m_1(i) - m_0(i))^2) \end{aligned} \tag{4.5}$$

so

$$\lim_{\substack{n_1,n_0\to\infty \\ |\mu_1-\mu_0|\to 0}} \left(\frac{1}{n_1} + \frac{1}{n_0}\right) \frac{Z(i)}{cor(y, X_i)} = \sqrt{p(1-p)} \tag{4.6}$$

Amongst SNPs in LE with small effect sizes ($\mu_1 - \mu_0$ small), expression 4.3 is maximised for $\alpha_i \propto \lim_{n_1,n_0\to\infty} Z(i)$. If we denote by Z_{1i}, Z_{2i} the GWAS Z scores for SNP i in phenotypes 1 and 2 respectively in studies with all group sizes $\Theta(n)$, the genetic correlation between the phenotypes is

$$r_g \approx \lim_{n\to\infty} \frac{\sum_{i\in S} Z_{1i} Z_{2i}}{\sqrt{\sum_{i\in S} Z_{1i}^2 \sum_{i\in S} Z_{2i}^2}} \tag{4.7}$$

The sum is over all SNPs S, but the only SNPs with non-vanishing contributions to r_g are those which are associated with both phenotypes. For the two traits in method 2, these SNPs are exactly those which are in our (idealised) category 3 in our full model. Writing C_i as the category of the SNP i we can rewrite the above as

$$r_g \approx \lim_{n\to\infty} \frac{\sum_{i\in S} I(C_i = 3) Z_{1i} Z_{2i}}{\sqrt{\sum_{i\in S} I(C_i = 3) Z_{1i}^2 \sum_{i\in S} I(C_i = 3) Z_{2i}^2}} \tag{4.8}$$

for which an obvious estimator is

$$\hat{r_g} = \frac{\sum_{i \in S} Pr(C_i = 3) Z_{1i} Z_{2i}}{\sqrt{\sum_{i \in S} Pr(C_i = 3) Z_{1i}^2 \sum_{i \in S} Pr(C_i = 3) Z_{2i}^2}} \tag{4.9}$$

If we were to define our full model such that Z_a, Z_d for SNPs in category 3 were distributed as a single bivariate Gaussian distribution with covariance ρ' (as opposed to our current model of two symmetric Gaussians), the updating step for ρ in the E-M algorithm would have a similar form. Indeed, if Θ_{n-1} is the set of estimates for $\{\pi_1, \pi_2, \sigma_2, \sigma_3, \tau, \rho'\}$ after step $n-1$ of the E-M algorithm, the updating steps for ρ', τ, σ_3 are

$$\begin{aligned}
(\rho')_n &\leftarrow \frac{\sum_{i \in S} Pr(C_i = 3|\Theta_{n-1}) Z_a(i) Z_d(i)}{\sum_{i \in S} Pr(C_i = 3|\Theta_{n-1})} \\
(\sigma_3)_n &\leftarrow \sqrt{\frac{\sum_{i \in S} Pr(C_i = 3|\Theta_{n-1}) Z_a(i)^2}{\sum_{i \in S} Pr(C_i = 3|\Theta_{n-1})}} \\
(\tau)_n &\leftarrow \sqrt{\frac{\sum_{i \in S} Pr(C_i = 3|\Theta_{n-1}) Z_d(i)^2}{\sum_{i \in S} Pr(C_i = 3|\Theta_{n-1})}}
\end{aligned} \tag{4.10}$$

and hence when the E-M algorithm converges, $\rho'/(\sigma_3 \tau)$ is an estimator for r_g. Testing $r_g \neq 0$ in this scenario is broadly equivalent to testing whether $\rho' \neq 0$ in the adapted full model.

When developing the PLR method, we chose not to use this simpler model, opting for a more complex two-Gaussian distribution of (Z_a, Z_d) for SNPs in category 3. There were several reasons for our choice. Importantly, $\rho' \neq 0$ implies $\rho > 0$, so the test $r_g \neq 0$ tests a more specific proposition than the PLR.

Testing for $\rho' \neq 0$ or $r_g \neq 0$ is weakened when Z_a and Z_d are correlated at some group of SNPs and anticorrelated at others. We note that this simultaneous correlation and anticorrelation is likely in many biological scenarios. Given two disease subgroups 1 and 2, deleterious variants associated only with subgroup 1 will have correlated Z_a, Z_d values, whereas deleterious variants associated only with subgroup 2 will have anticorrelated Z_a and Z_d.

In addition, the presence of between-subgroup heterogeneity, as characterised by the presence of SNPs with simultaneously high $|Z_d|$ and $|Z_a|$ values, does not require that Z_a and Z_d have to be correlated or anticorrelated at all. The presence of a set of SNPs whose marginal variances of Z_a and Z_d are simultaneously significantly larger than 1 is sufficient evidence for heterogeneity of disease basis. This was the impetus for including the additional parameter σ_3 in the full model.

Uncorrelated Z_a and Z_d may well occur in situations where the main sources of variation between the subgroups are only weakly associated with the overall phenotype, while less associated variants are strongly associated. This would be expected to occur in situations where the subtypes have known genetic differences. If, for example, a subgrouping phenotype was based on visual acuity in the phenotype of symptomatic Type 2 diabetes, variants associated with general macular degeneration would have large $|Z_d|$ scores with low $|Z_a|$ scores, while variants associated with microvascular glucose sensitivity would have larger $|Z_a|$ scores and smaller (but still overdispersed) $|Z_d|$ scores.

The behaviours of r_g/ρ', ρ, τ and σ in various scenarios are summarised in supplementary table 1. Overall, we consider that while ρ_g is a useful statistic, it does not capture the variety of forms that disease heterogeneity can take.

4.3.2 Simulations

We tested the ability of GCTA to reject the null hypothesis $r_g = 0$ on simulated data. We simulated genotypes for 4000 controls and 2000 cases in each of two subgroups at 10000 SNPs in linkage equilibrium. Genotypes were simulated in such a way that Z_a and Z_d scores would have the distributions

$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \quad \text{at 7000 SNPs } (\pi_1 = 0.7)$$
$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}\right) \quad \text{at 2000 SNPs } (\pi_2 = 0.1)$$
$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & \rho \\ \rho & 4 \end{pmatrix}\right) \quad \text{at } \xi * 1000 \text{ SNPs } (\pi_3 = 0.2)$$
$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & -\rho \\ -\rho & 4 \end{pmatrix}\right) \quad \text{at } (1-\xi) * 1000 \text{ SNPs } (\pi_3 = 0.2)$$

The value ξ represents the degree to which Z_a, Z_d scores can show both correlation and anticorrelation, and ρ represents the extent of the correlation/anticorrelation. We ran simulations at $\rho = 0$ and for $\rho \in \{0, 0.5, 1, 2\}$ for $\xi = 0$ (no anticorrelation), $\xi = 0.2$ (mostly correlation, some anticorrelation) and $\xi = 0.5$ (equal correlation and anticorrelation). The large value of π_3 was to ensure that both PLR and GCTA should be well-powered to reject the null hypothesis where able, but not so well-powered as to be incomparable.

We estimated r_g using the GCTA method [3]. Significance was assessed using the provided likelihood-ratio test comparing the fitted model with a null model in which $r_g = 0$.

We did not test LDSC in this scenario, as it estimates r_g based on phenomena arising from the LD matrix, and simulation would entail setting an inherent effect size for these phenomena through specifying an LD matrix. Since the shortcomings we identify are with the use of r_g itself, rather than the method used to simulate it, we considered this reasonable.

As expected, the test based on $r_g = 0$ was not able to reject the null hypothesis when $\rho = 0$ or $\xi = 0.5$, and power was markedly reduced when some anticorrelation was present, at $\xi = 0.2$ (figure 4.3, table 4.1). While the test was able to systematically reject the null hypothesis when $\xi \in \{0, 0.2\}$, $\rho > 0$, the power was universally lower than that of the PLR test (table 4.1). This was likely due to information gained from the additional degree of freedom (σ_3) between the full and null models in the PLR test. We did not simulate any scenarios where $\sigma_3 = 1$, as this would imply that SNPs in category 3 were not systematically associated with the subgrouping phenotype, and hence correlation with Z_a would be spurious.

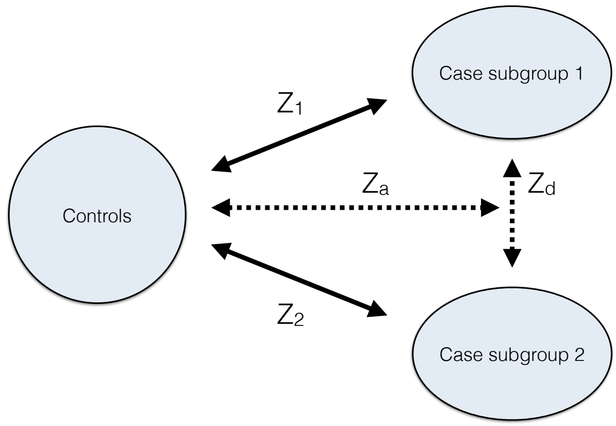
4.3.3 Application to real data

Finally, we assessed whether we could reject H_0 by testing against $r_g = 0$ on our ATD dataset (MHC removed), with subtypes GD and HT. We used both the LDSC and GCTA methods to do this. While both were able to reject the null hypothesis (LDSC: $r_g = -0.579, p = 0.04$, from known null distribution of ρ_g; GCTA: $r_g = -0.580, p = 1 \times 10^{-3}$ from likelihood-ratio test) neither could do so as confidently as the PLR test ($p = 2.2 \times 10^{-15}$).

Our proposed test is complex, and parametrises disease heterogeneity using several variables (namely π_3, σ_3, τ and ρ) rather than providing a single metric. We consider this complexity to be necessary; heterogeneity in a phenotype can arise in many ways and the heterogeneous genetic architecture can take many forms. A test specifically to detect SNPs with large, genome-wide significant effect sizes in one disease subgroup but not the other may miss heterogeneity characterised by subtle effect size differences across many SNPs with small effects. Our method can ideally detect heterogeneity in a general sense in multiple situations, and give insight into the architecture in the form of the fitted parameters.

ρ	ξ	**GCTA**	**PLR**
0	0	0.09 (0.002)	1 (-)
0	0.2	0.12 (0.002)	1 (-)
0	0.5	0.06 (0.002)	1 (-)
0.5	0	0.55 (0.006)	1 (-)
0.5	0.2	0.13 (0.004)	1 (-)
0.5	0.5	0.06 (0.001)	1 (-)
1	0	0.96 (0.002)	1 (-)
1	0.2	0.59 (0.005)	1 (-)
1	0.5	0.07 (0.002)	1 (-)
2	0	1 (-)	1 (-)
2	0.2	1 (-)	1 (-)
2	0.5	0.04 (0.001)	1 (-)

Table 4.1: Power of tests to reject the null hypothesis at $\alpha = 0.05$ in simulated data. Brackets show standard error. Value ρ is the degree of correlation/anticorrelation between Z_d and Z_a. Value ξ is the degree of split between correlation and anticorrelation; $\xi = 0$ corresponds to correlation only, $\xi = 0.2$ to mostly correlation with some anticorrelation, and $\xi = 0.5$ to a half/half mix. Testing for subgroup heterogeneity using GCTA is adequately powerful when correlation ρ is present, but declines markedly when both correlation and anticorrelation are present, and is effectively zero when $p = 0.5$ or $\rho = 0$. The PLR-based test was able to reject H_0 universally in all cases.



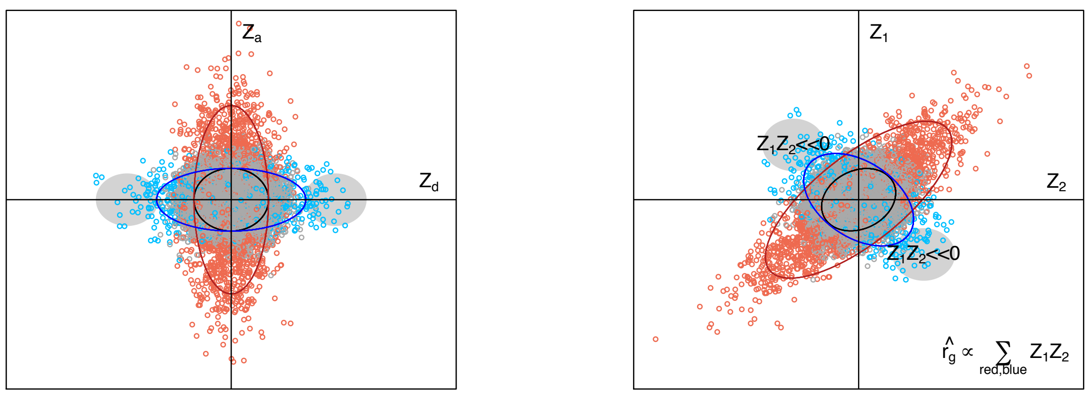


Figure 4.1: One way to test for phenotypic heterogeneity using genetic correlation (r_g) is to estimate r_g for two separate case-control studies; each comparing the control group to one of the disease subgroups, and test whether the estimated r_g is significantly less than 1. We denote by Z_1, Z_2 the sets of Z- scores corresponding to allelic differences between controls and cases of subtype 1 and between controls and cases of subtype 2 respectively (top panel) in contrast to our usual Z_a and Z_d scores. A shortcoming of this method is that r_g is decreased by the presence of SNPs which show allelic differences between subtypes, but are unrelated to the phenotype overall. In this sense, the test $r_g < 1$ is responsive to *any* genetic difference between subtypes - not just those which correspond to differing disease pathology. This scenario would arise if subgroups were defined based on a phenotype with non-zero heritability which was unrelated to the disease; eg, subgroups of T1D defined by hair colouring. The lower two panels demonstrate this scenario. The left panel shows (simulated) Z_a and Z_d scores for a set of SNPs under H_0, where grey corresponds to category 1, red to category 2, and blue to category 3. The right lower panel shows the corresponding sets of Z_1 and Z_2 values. SNPs in the grey circles, and generally SNPs coloured blue, will contribute negatively to the overall genetic correlation, which is asymptotically proportional to the sum of $Z_1 Z_2$ over all SNPs coloured red or blue.

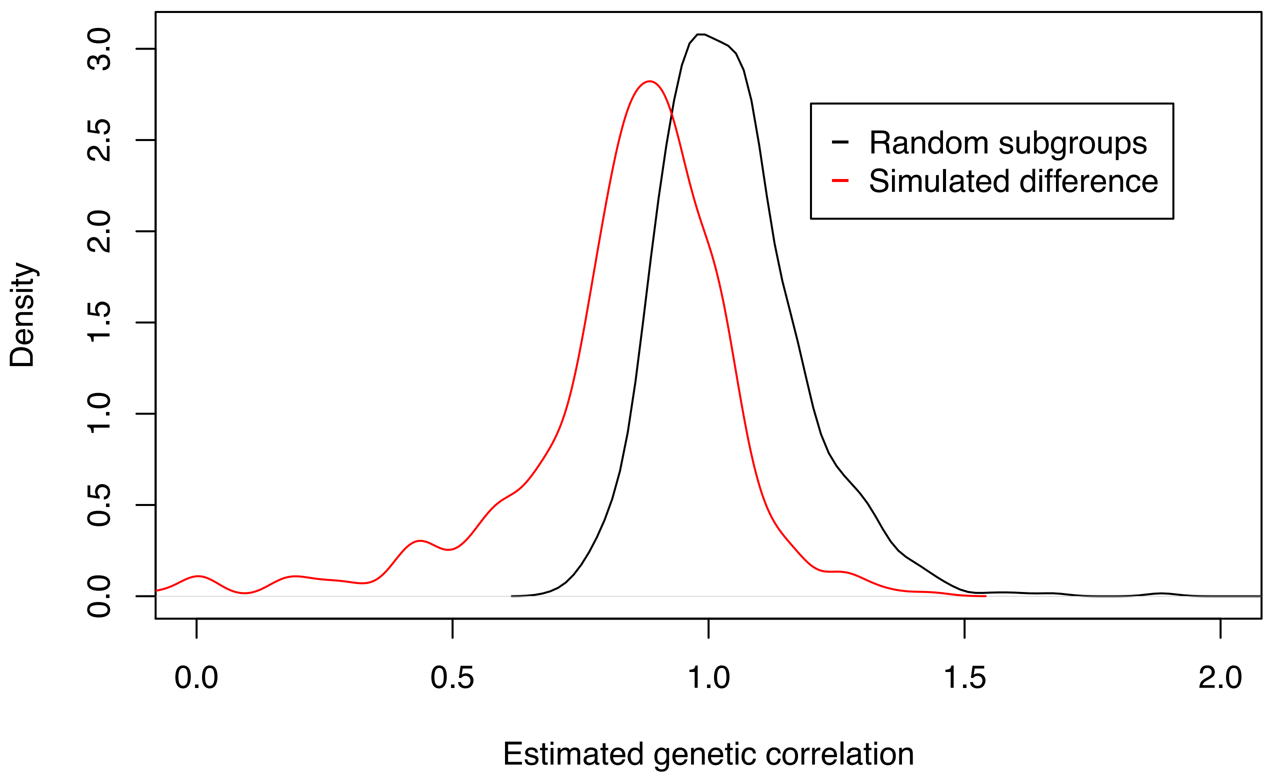


Figure 4.2: Density of estimated r_g (LDSC method) for method 1. Estimates for random subgroups generated under H_0 are shown in black. Estimates for subgroups with a simulated difference ($\tau > 1$) are shown in red. A test based on method 1 would reject H_0 if r_g was significantly less than 1; however, as the plot shows, this would lead to systematic false positives in the scenario where $\tau > 1$. Some estimated values of r_g are greater than 1 due to the way the statistic is estimated under the LDSC method.

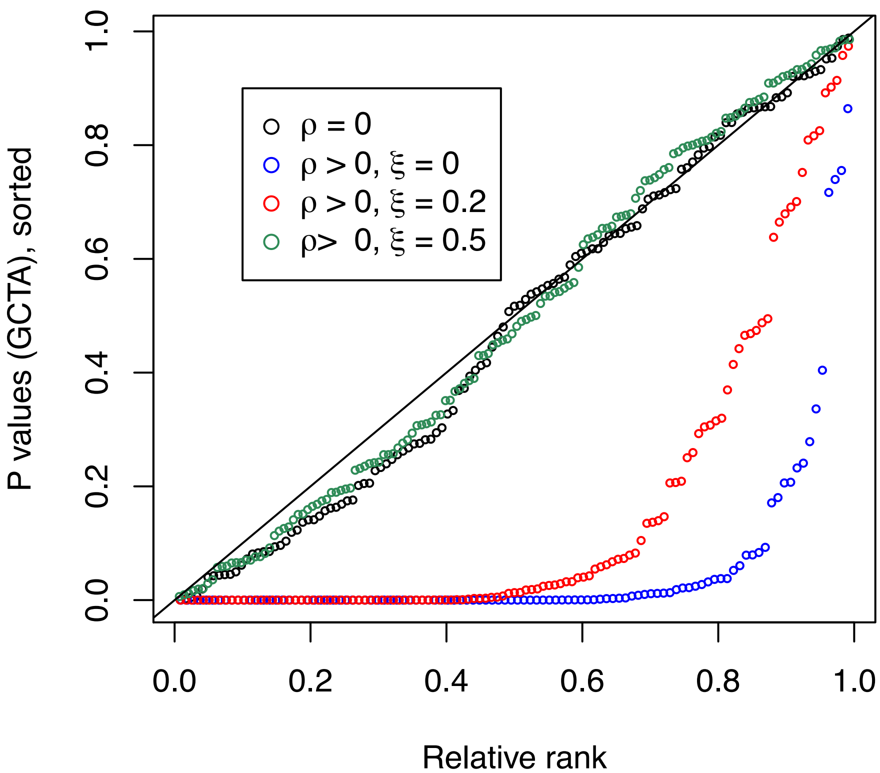


Figure 4.3: Sorted p values from test of null hypothesis $r_g = 0$ under simulations in which $\rho \in \{0, 0.5, 1, 2\}$ and $\xi \in \{0, 0.2, 0.5\}$. In all simulations, H_0 is false (with $\sigma_3 > 0$). GCTA is able to reject the null hypothesis only if $\rho > 0$ and $p \neq 0.5$, and power is reduced (ie, p-values are higher) if $p = 0.2$ compared to $p = 0$. If $\rho = 0$ or $\xi = 0.5$, the p-values show effectively no deviation from $U(0, 1)$. Thus a test based on rejecting $\rho_g = 0$ is not suitable for our purposes.

Note 5

Other

5.1 Alternative test statistics for retrospective single-SNP analysis

We propose four summary statistics for testing the degree to which single SNPs have differential effect sizes in disease subgroups. The fourth of these, the Bayesian conditional false discovery rate (cFDR) is discussed in the methods section of the main text. The three alternative statistics (which we term X_1, X_2, X_3) test against slightly different null hypotheses.

The first, X_1, is the posterior probability of membership of the third category of SNPs under the full model; that is, for a SNP of interest with Z scores z_a, z_d and given fitted parameters $\Theta_1 = \{\pi_1, \pi_2, \pi_3, \sigma_2, \sigma_3, \tau, \rho\}$:

$$\begin{aligned} X_1 &= Pr(\text{SNP} \in \text{category } 3|\Theta_1) \\ &= \frac{\frac{1}{2}\pi_3 \left(N_{\mathbf{0}, \begin{pmatrix} \tau^2 & \rho \\ \rho & \sigma_3^2 \end{pmatrix}}(z_a, z_d) + N_{\mathbf{0}, \begin{pmatrix} \tau^2 & -\rho \\ -\rho & \sigma_3^2 \end{pmatrix}}(z_a, z_d) \right)}{PDF_{\Theta_1}(z_a, z_d)} \end{aligned} \tag{5.1}$$

This test statistic has the advantage of straightforward FDR control against the null hypothesis $H_0 = \{\text{SNP} \in \text{category } 1/2|\Theta_1\}$, assuming the validity of Θ_1. It also reflects the overall shape of the distribution. A disadvantage is the dependence on the model implied by Θ_1; in circumstances where $\sigma_3 >> \sigma_2$, the test statistic X_1 will be high for high values of $|Z_a|$ even when $|Z_d|$ is low (supplementary figures 7). This is a particular problem if tested regions include very strong associations; for example, the MHC region in autoimmune phenotypes.

Our second statistic, X_2, is the difference in pseudo-log likelihood of a given SNP under the full and null models; that is, given fitted parameters Θ_1 under H_1 and Θ_0 under H_0

$$X_2 = log\{PL(z_a, z_d|\Theta_1)\} - log\{PL(z_a, z_d|\Theta_0)\} \tag{5.2}$$

This has the advantage that high values of X_2 directly identify the SNPs contributing to a higher pseudo-likelihood ratio. A disadvantage is the sensitivity to the behaviour of the fitted parameters under H_0, which may be variable (see main paper, results section, page 7 and table 2), and absence of direct FDR control. Because X_1 and X_2 tend to highlight uninteresting SNPs in differing circumstances, we found a combination of both to be useful to find SNPs which are 'unusual' (high X_1) and contribute to the PLR (high X_2).

The third test statistic is defined as $X_3 = z_a^{\alpha} z_d^{1-\alpha}$, $\alpha \in (0, 1)$. We chose this test statistic as we are broadly searching for evidence of correlation between Z_a and Z_d, and SNPs contribute to measures of correlation principally through the value of $Z_a Z_d$. This test statistic identifies SNPs with concurrently high Z_a and Z_d in an obvious way, so is of most use when SNPs which differentiate subgroups are not of interest unless they are also associated with the overall phenotype.

The value of α is set in order to prioritise SNPs with high Z_d over those with high Z_a; for instance, with $\alpha = 0.5$ will give equal weight to a SNP with $Z_a = 10$, $Z_d = 1$ and a SNP with $Z_a = 1$, $Z_d = 10$, but in general the second SNP will be of far greater interest. To determine the best value of α, we consider how much we may expect Z_a and Z_d to deviate from 0, using both the full and null models.

We set τ' as the largest value of τ across both models, and σ' as the largest of σ_2 (null model) and σ_2, σ_3 (full model). Given fitted values τ', σ', we suggest the value

$$\alpha = \frac{log(\sigma')}{log(\tau') + log(\sigma')} \tag{5.3}$$

so that the statistic X_3 has the same value at the points $(1, \tau')$ and $(\sigma', 1)$. The rationale for this is that SNPs which have the true underlying distributions $N_{\mathbf{0},\begin{pmatrix} \tau'^2 & 0 \\ 0 & 1 \end{pmatrix}}$ or $N_{\mathbf{0},\begin{pmatrix} 1 & 0 \\ 0 & \sigma'^2 \end{pmatrix}}$ are uninteresting; we seek deviance from both of these distributions. A hypothesis test for X_3 can then be computed, using the appropriate values of $\pi_{(0,1,2)}$.

Contour plots of the test statistics for several datasets are shown in supplementary figures 7,8.

5.2 Independence of PLR distribution on subgroup sizes

PLR and cPLR values for randomly chosen subgroups are all derived from data with the same Z_a values, with the distribution of Z_d expected to be $N(0, 1)$ and independent of Z_a regardless of the relative sizes of random subgroups. Therefore we expect that the asymptotic distribution (main paper, equation 2 does not depend on relative subgroup size. An important consequence of this is that if several subgroupings of a phenotype are being simultaneously assessed, the empirical distribution of cPLR need only be calculated once.

We demonstrate this assertion by simulation. Using our autoimmune thyroid disease dataset, we simulated random subgroups from the combined case group (GH+HT) for a range of relative sizes, repeating the simulation 1000 times for each subgroup size. Figure 5.1 shows the observed distributions of PLR and cPLR as compared to the overall distribution. These plots are consistent with independence of empirical PLR and cPLR distributions on subgroup size.

5.3 Number of simulations necessary to fit null distribution

We assessed the number of simulated random subgroups required to estimate the parameters γ, κ of the null distribution of the cPLR. We took bootstrap samples of various sizes from our list of simulated random subgroups ($\tau = 1$) of the ATD data. For each sample, we computed the fitted values of γ and κ and the observed p-values associated with observed PLR values of 2, 3, 5, and 10, i.e. expected p values 0.08, 0.03, 0.004 and 1.5×10^{-6} respectively (figure 5.2)

This suggests that 1000 simulations is generally adequate, and it is difficult to improve accuracy markedly past this point. For this number of simulations, 95% of computed values for κ, γ, $Pr(PLR > 2|\kappa, \gamma)$ and $Pr(PLR > 5|\kappa, \gamma)$ were in $[0.44, 0.56]$, $[0.46, 0.72]$, $[0.069, 0.97]$ and $[0.0021, 0.0057]$ respectively. As expected, consistency of p-value estimates is poorer for lower p-values, as these correspond to greater extrapolations of the distribution.

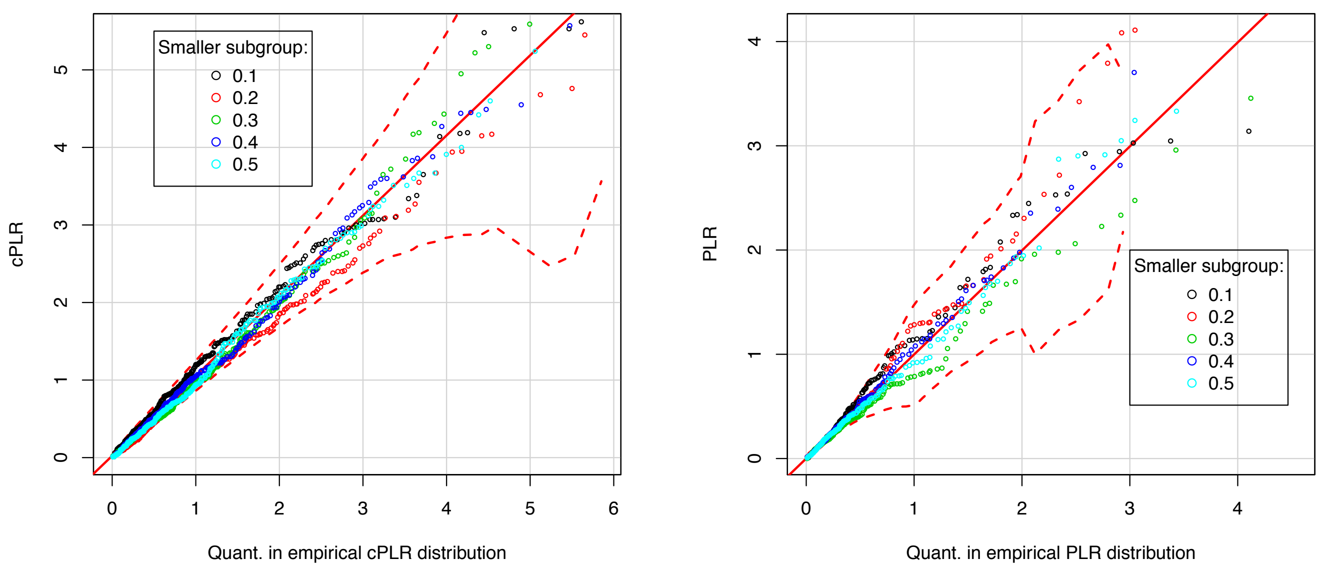


Figure 5.1: Distributions of PLR and cPLR for various relative sizes of subgroups. Simulations are on ATD data. Legend shows the proportion of cases in the smaller subgroup. Leftmost plot shows distribution of observed cPLR, rightmost distribution of PLR. Red dotted lines show empirical 99% confidence limits. Distributions are similar for all relative subgroup sizes.

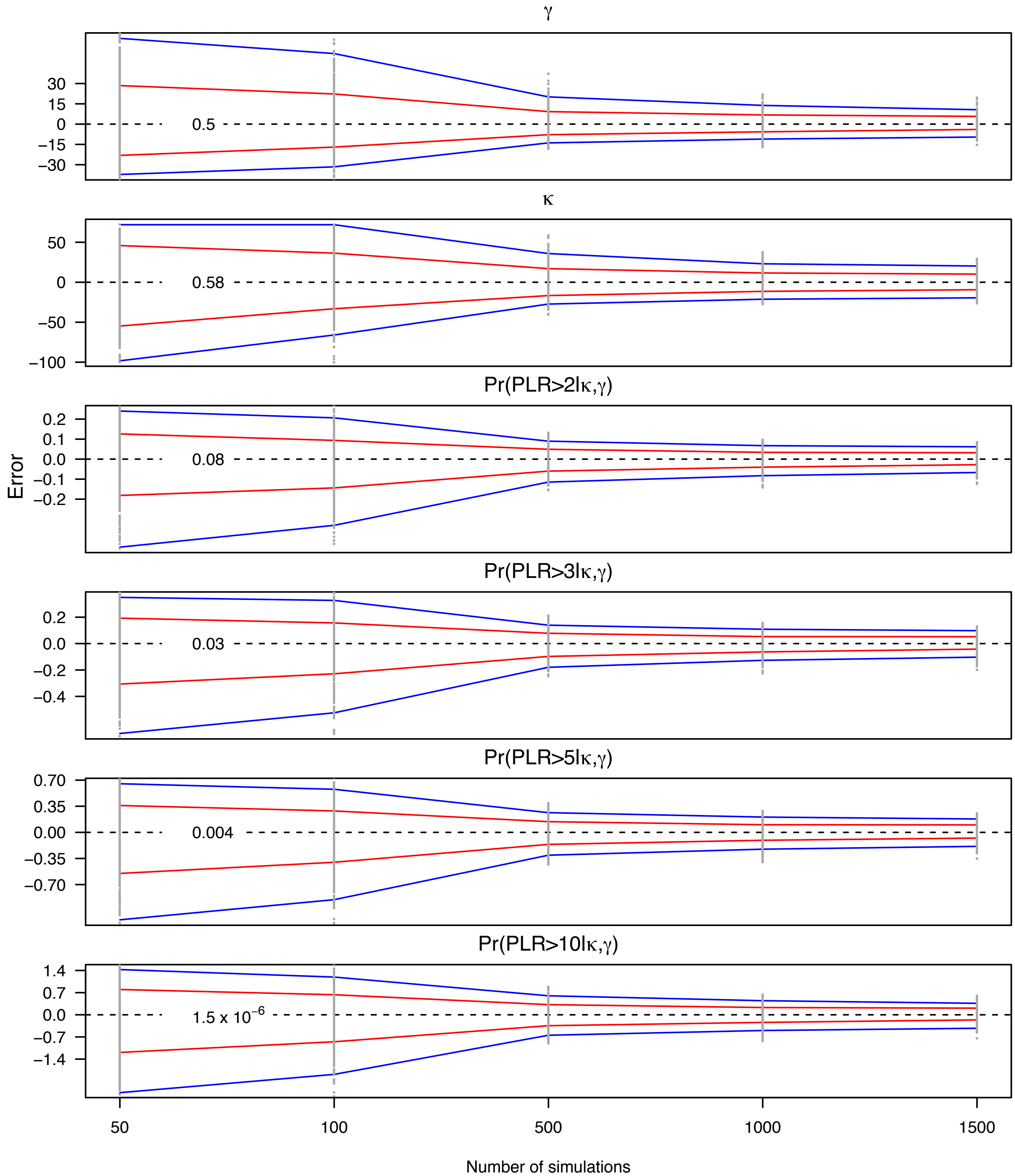


Figure 5.2: Distributions of estimated parameters γ and κ and various corresponding p-values, using various numbers of simulated random subgroups. Blue lines show quantiles of observed distribution corresponding to $\pm 2\sigma$; red lines show quantiles corresponding to $\pm\sigma$. Errors in γ and κ are shown as percentage errors as compared to median. Errors in p-values are shown as log_{10} fold changes from median. Values of the median value of each variable are shown. Observed values are shown as grey dots.

Bibliography

[1] Rodriguez-Calvo T, Sabouri S, Anquetil F, von Herrath MG (2016) The viral paradigm in type 1 diabetes: Who are the main suspects? Autoimmunity Reviews .

[2] Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. (2015) An atlas of genetic correlations across human diseases and traits. bioRxiv .

[3] Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics 28: 2540–2542.

A method for identifying genetic heterogeneity within phenotypically-defined disease subgroups
Supplementary Figures

James Liley, John A Todd and Chris Wallace

November 8, 2016

List of Figures

1 Geographic subgroups 2
2 Test statistics for geographically-defined disease subgroups 3
3 Power of test at a range of values of π_3, σ_3, τ, and ρ 6
4 Distribution of observed cPLRs for random subgroups 8
5 Z scores for autoantibody-based subgroups 10
6 Z scores for age at diagnosis in T1D 11
7 Comparison of test statistics for single-SNP effects (T1D/RA; GD/HT) . . 16
8 Single-SNP effects for TPOAb in T1D 17
9 Single-SNP effects for age at diagnosis in T1D 18

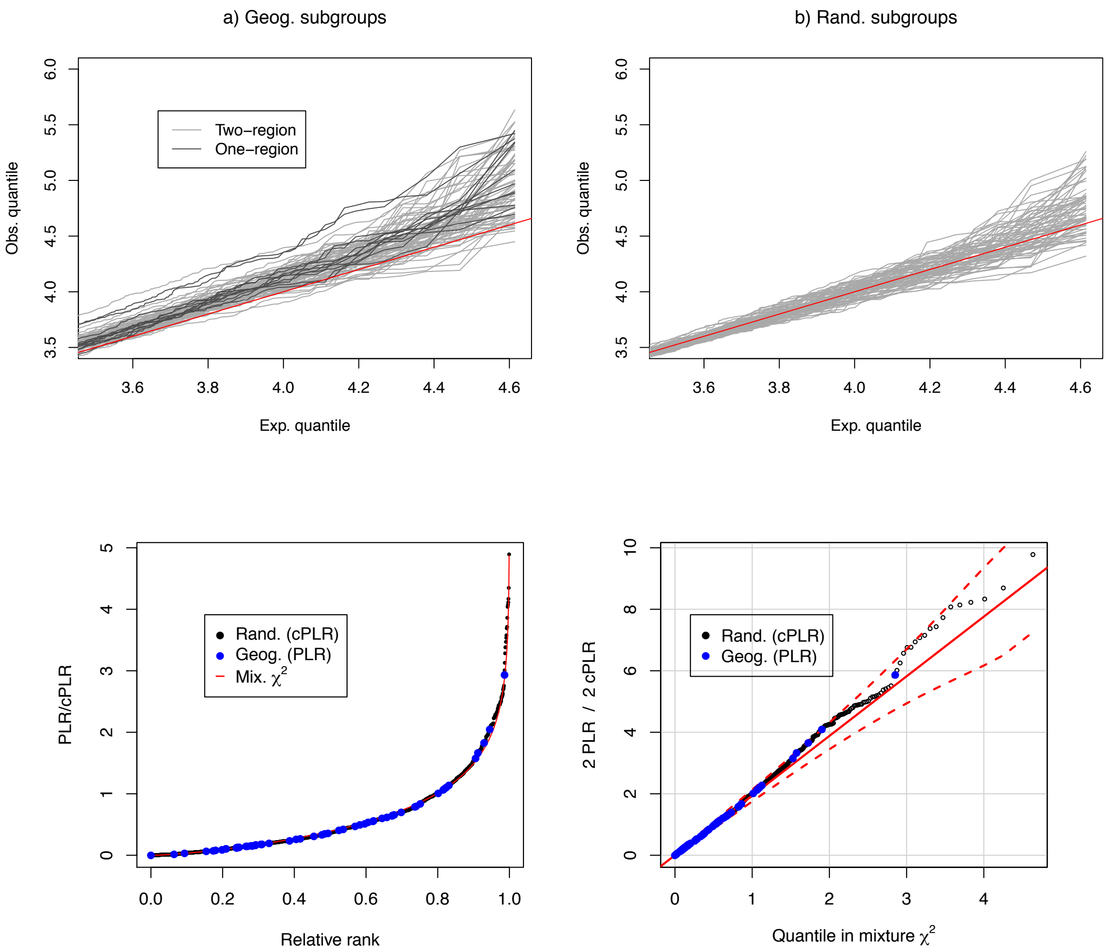


Figure 1: Top plot shows Z_d scores arising from geography-based subgroups compared with expected normal. Leftmost plot shows quantiles of Z scores from geography based subgroups; two-region subgroups in light grey and one-region subgroups in dark grey. Considerable inflation is seen compared to Z-scores arising from random subgroups, in rightmost figure.
Lower plots show distribution of cPLR values from random subgroups against observed PLR values from geographically-defined subgroups. Leftmost plot shows cPLR values from random subgroups plotted in ascending with PLR values from random subgroups shown in blue. Rightmost plot is Q-Q plot comparing null cPLR distribution with the asymptotic mixture-χ^2.

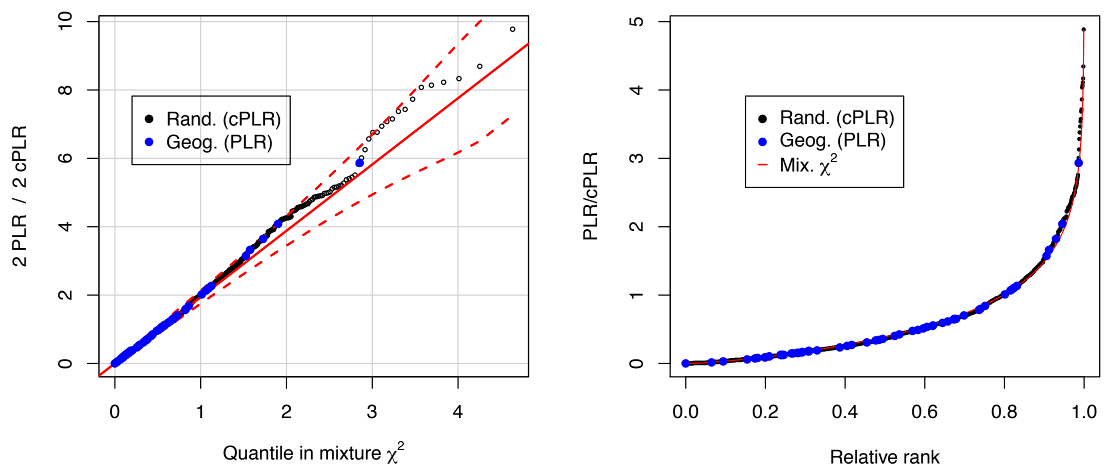
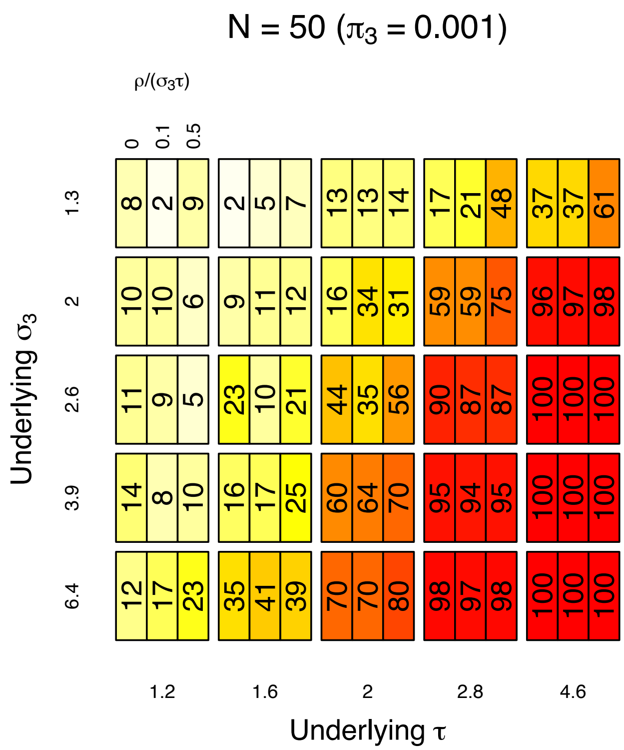
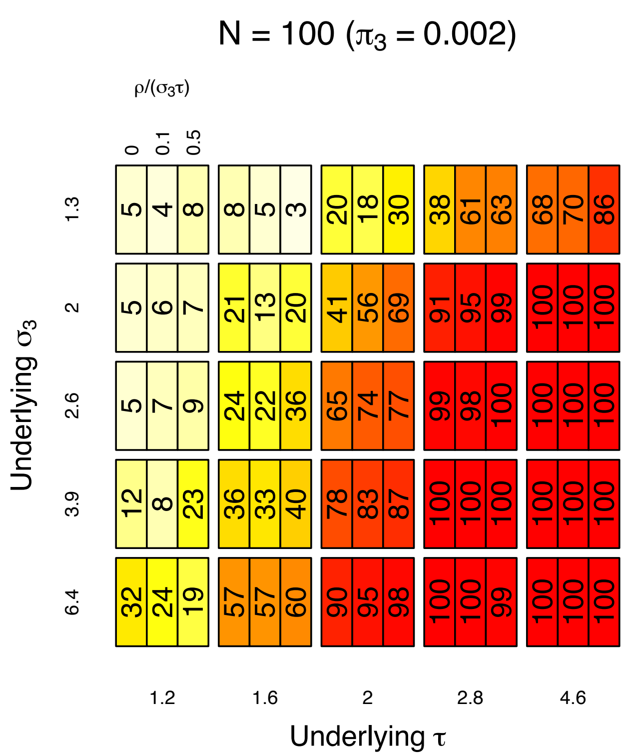
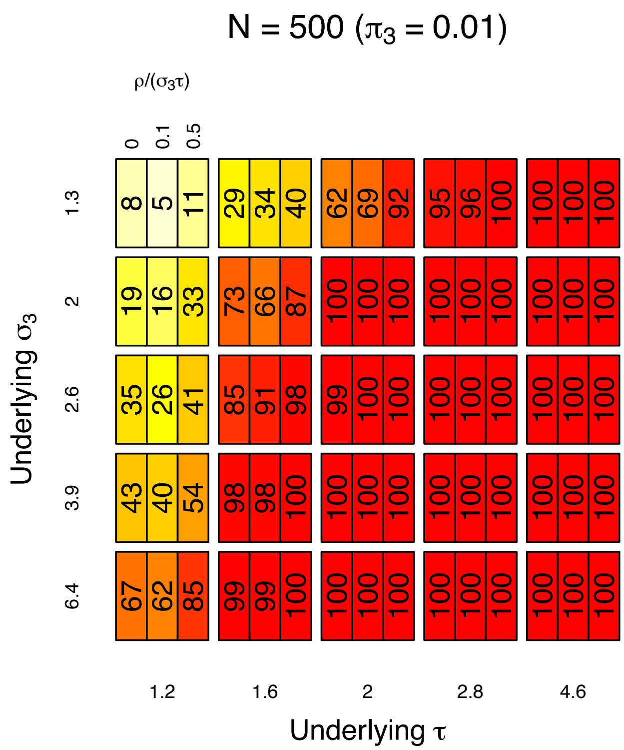
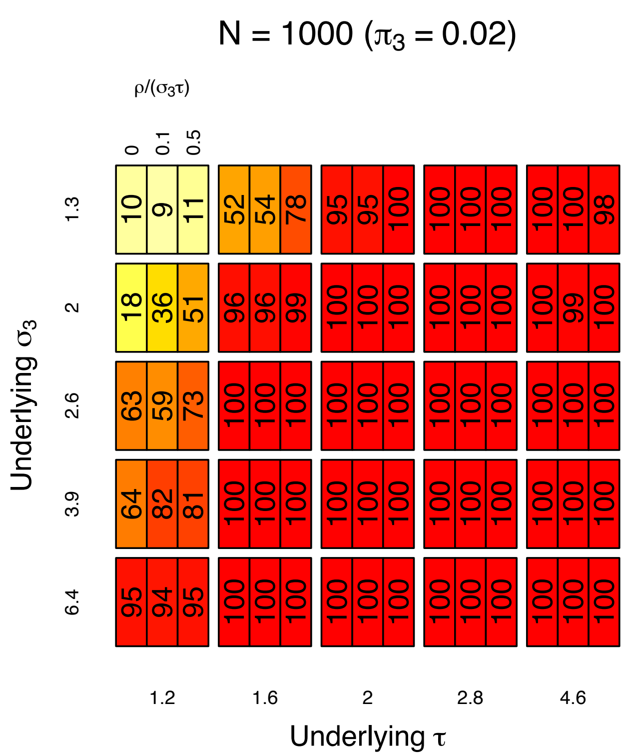


Figure 2: Summary of test statistics (PLR) from geographically-defined subgroups, based on WTCCC data [1] for controls and type 1 diabetes (T1D). In each instance, one subgroup was defined as the controls coming from either one or two geographic regions, and the other subgroup as the controls coming from the remaining nine or ten geographic regions. We also generated > 2000 randomly allocated subgroups and computed the cPLR. The left panel shows a Q-Q plot of cPLR values from random subgroups against the asymptotic mixture-χ^2 distribution, with blue points representing the PLRs of geographic subgroups. The right panel shows cPLR values plotted in ascending order with the PLR values from geographic subgroups included as blue points. The minimum Bonferroni-corrected empirical p value was > 0.5


N = 50 (π_3 = 0.001)
ρ/(σ3τ)
0 0.1 0.5
1.3: 8 2 9 | 2 5 7 | 13 13 14 | 17 21 48 | 37 37 61
2: 10 10 6 | 9 11 12 | 16 34 31 | 59 59 75 | 96 97 98
2.6: 11 9 5 | 23 10 21 | 44 35 56 | 90 87 87 | 100 100 100
3.9: 14 8 10 | 16 17 25 | 60 64 70 | 95 94 95 | 100 100 100
6.4: 12 17 23 | 35 41 39 | 70 70 80 | 98 97 98 | 100 100 100
Underlying σ3
1.2 1.6 2 2.8 4.6
Underlying τ


N = 100 (π_3 = 0.002)
ρ/(σ3τ)
0 0.1 0.5
1.3: 5 4 8 | 8 5 3 | 20 18 30 | 38 61 63 | 68 70 86
2: 5 6 7 | 21 13 20 | 41 56 69 | 91 95 99 | 100 100 100
2.6: 5 7 9 | 24 22 36 | 65 74 77 | 99 98 100 | 100 100 100
3.9: 12 8 23 | 36 33 40 | 78 83 87 | 100 100 100 | 100 100 100
6.4: 32 24 19 | 57 57 60 | 90 95 98 | 100 100 99 | 100 100 100
Underlying σ3
1.2 1.6 2 2.8 4.6
Underlying τ


N = 500 (π_3 = 0.01)
$\rho/(\sigma_3\tau)$
0 0.1 0.5
Underlying σ_3
1.3 2 2.6 3.9 6.4
8 5 11 29 34 40 62 69 92 95 96 100 100 100 100
19 16 33 73 66 87 100 100 100 100 100 100 100 100 100
35 26 41 85 91 98 99 100 100 100 100 100 100 100 100
43 40 54 98 98 100 100 100 100 100 100 100 100 100 100
67 62 85 99 99 100 100 100 100 100 100 100 100 100 100
1.2 1.6 2 2.8 4.6
Underlying τ


N = 1000 (π_3 = 0.02)
$\rho/(\sigma_3\tau)$
0 0.1 0.5
Underlying σ_3
1.3 2 2.6 3.9 6.4
10 9 11 52 54 78 95 95 100 100 100 100 100 100 98
18 36 51 96 96 99 100 100 100 100 100 100 100 99 100
63 59 73 100 100 100 100 100 100 100 100 100 100 100 100
64 82 81 100 100 100 100 100 100 100 100 100 100 100 100
95 94 95 100 100 100 100 100 100 100 100 100 100 100 100
1.2 1.6 2 2.8 4.6
Underlying τ

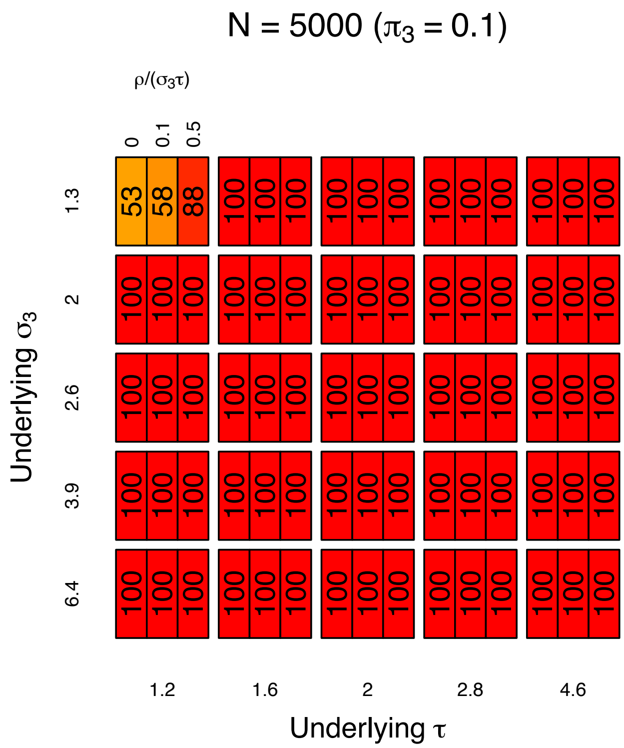
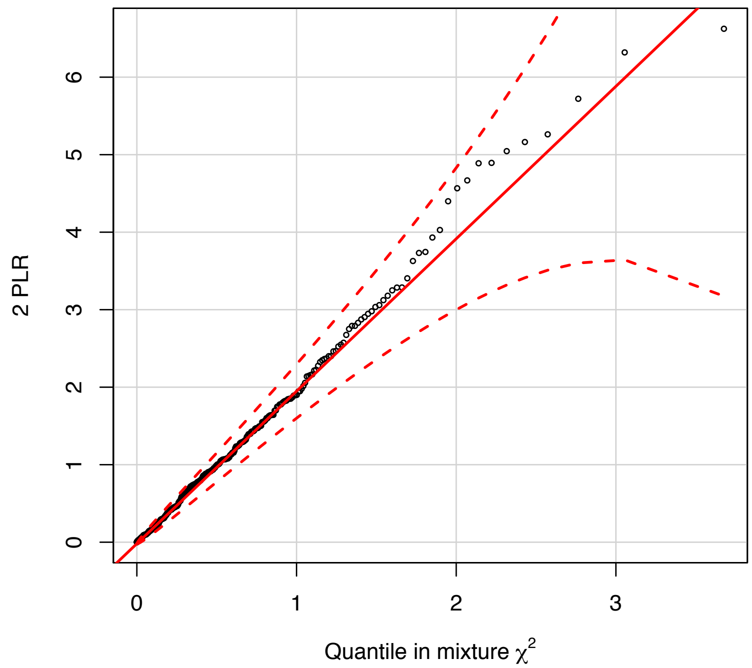
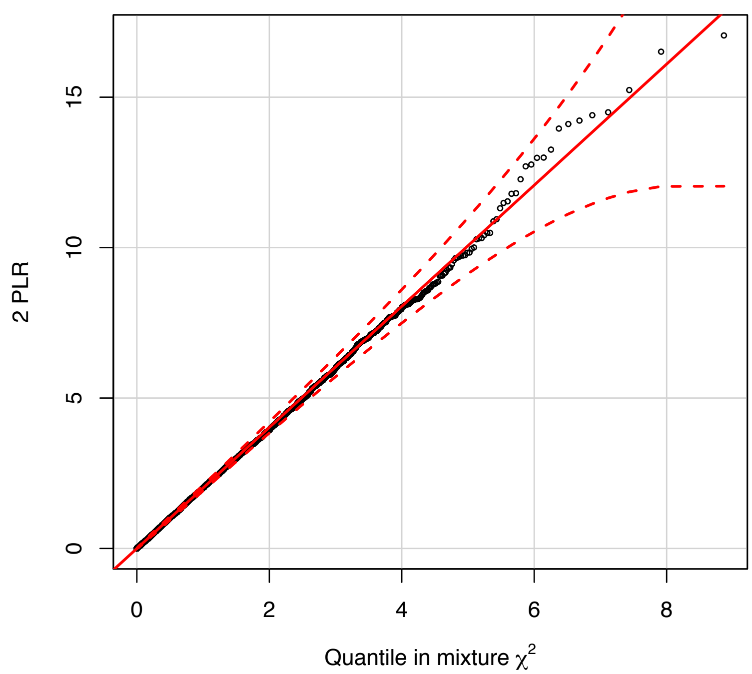


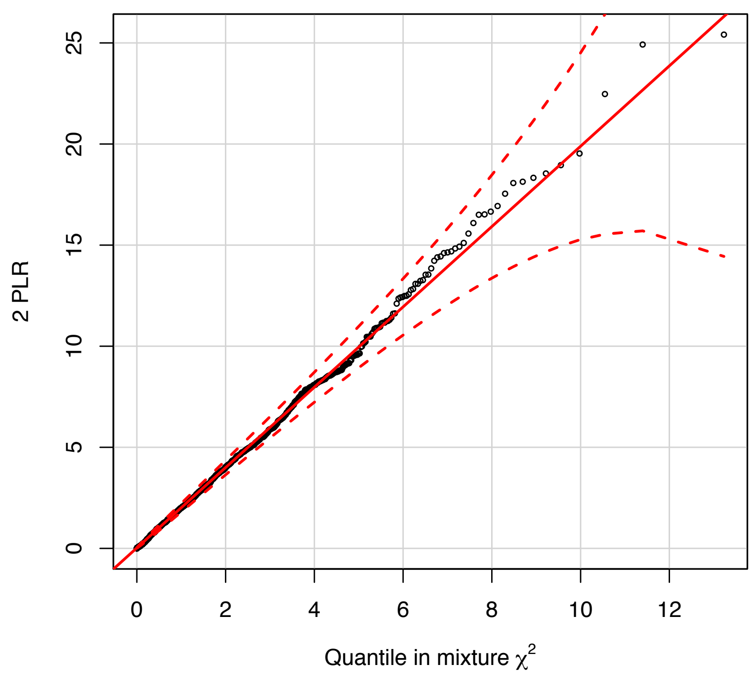
Figure 3: Estimates of power for various values of π_3, σ_3, τ, and ρ. The value N is the approximate number of SNPs in category 3, corresponding to π_3. In total, each simulation was on 5×10^4 simulated autosomal SNPs in linkage equilibrium. The value $\rho/(\sigma_3\tau)$ is the correlation (rather than covariance) between Z_a and Z_d in category 3.



(a) T1D/T2D/RA data

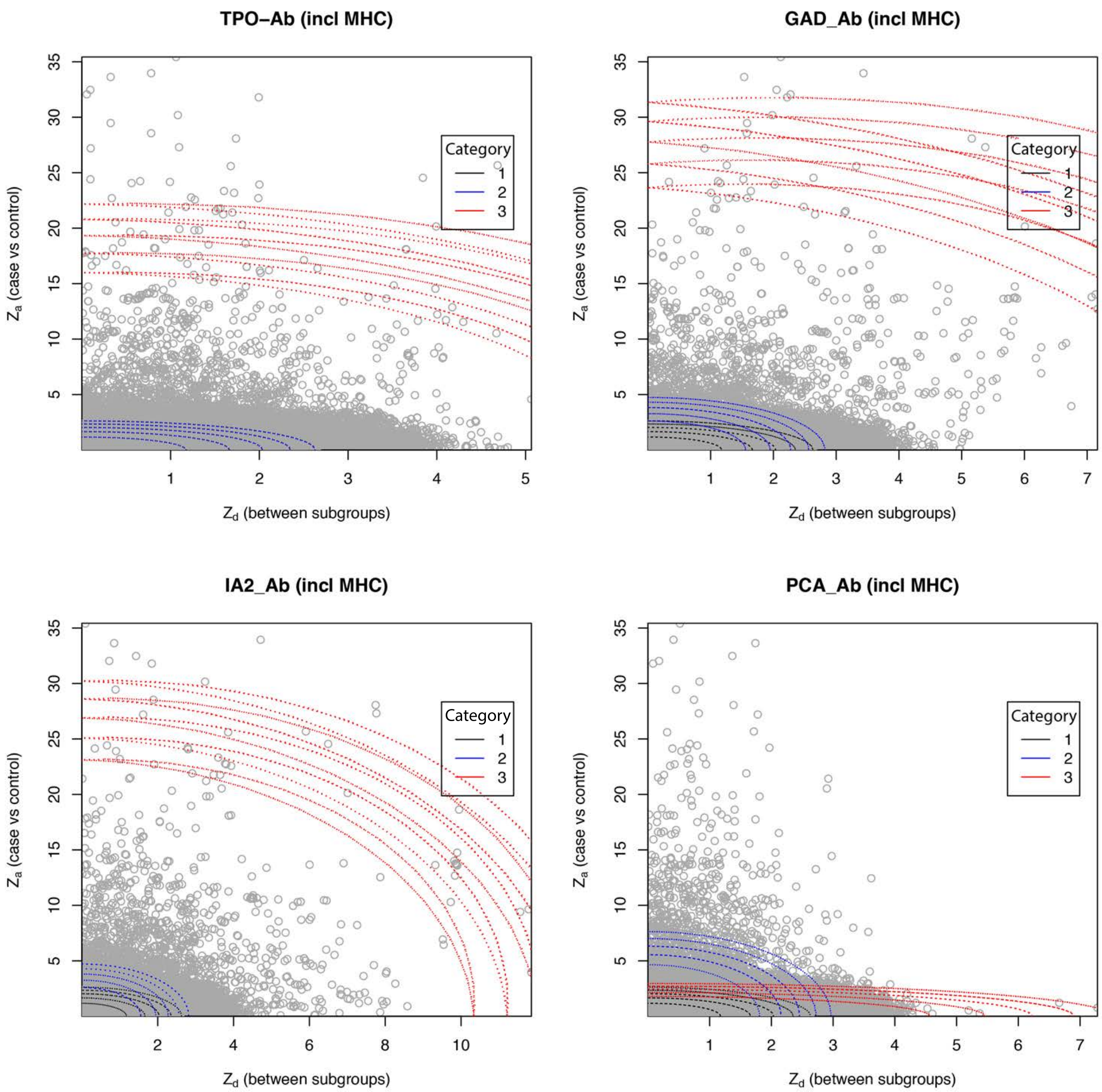


(b) ATD:GH/HT data



(c) T1D (AAB) data

Figure 4: Q-Q plot of the distribution of observed test statistics (cPLR) for random subgroups of tested phenotypes (T1D/RA/T2D combined, GH/HT combined, T1D) against a mixture χ^2 distribution of the form $\gamma * (\kappa\chi_1^2 + (1-\kappa)\chi_2^2)$. A 99% confidence interval is shown by the dashed red lines. The distribution is well-approximated by the asymptotic mixture-χ^2 in all cases.


TPO−Ab (incl MHC)
GAD_Ab (incl MHC)
IA2_Ab (incl MHC)
PCA_Ab (incl MHC)
Category
1
2
3
Za (case vs control)
Zd (between subgroups)

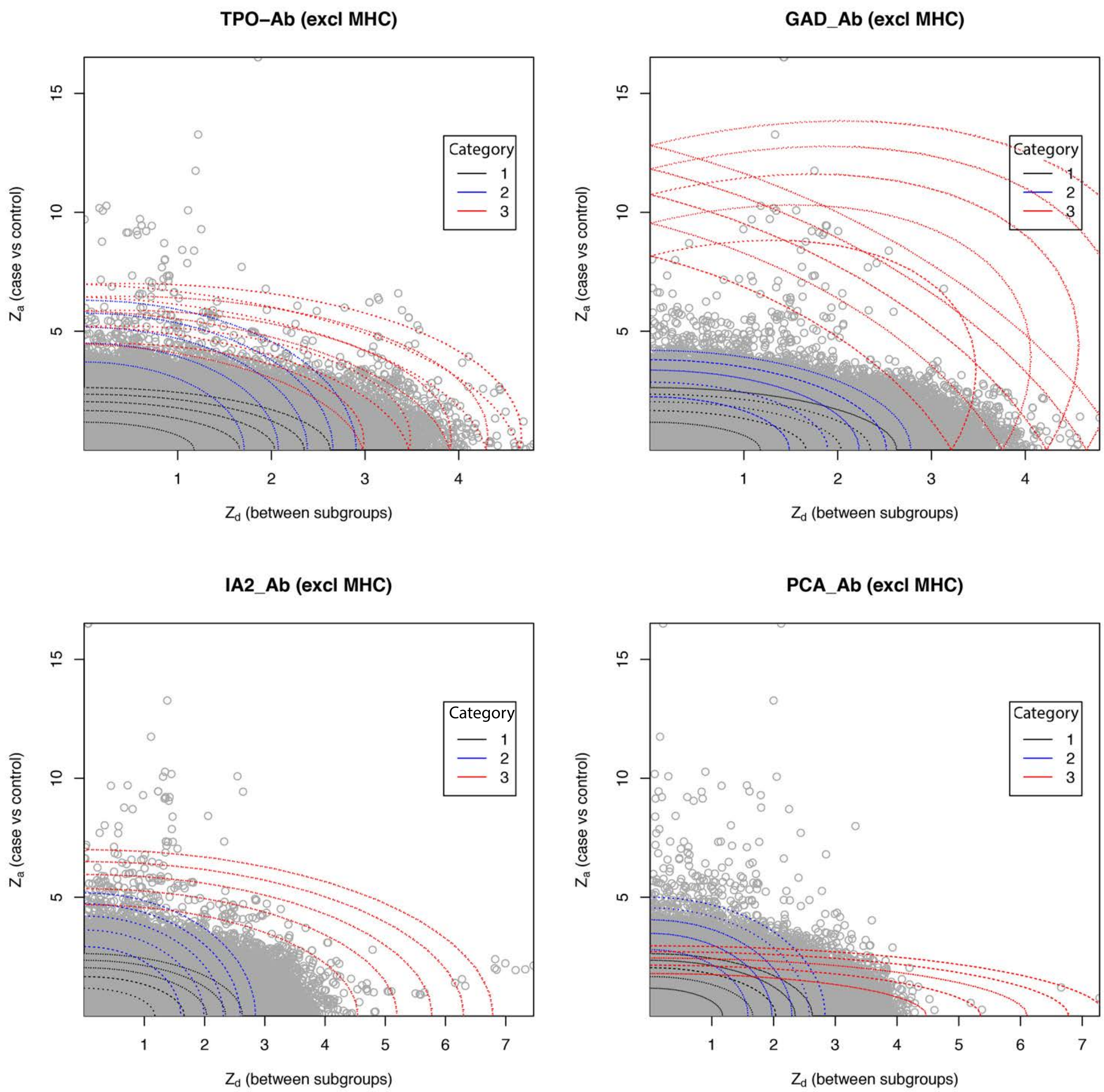


Figure 5: Observed Z_a and Z_d scores (grey) for T1D subtypings based on autoantibody positivity, including or excluding the MHC region, and contours of parameters of fitted models (coloured ellipses). Full models are shown for the comparisons involving TPO-Ab, GAD-Ab, and IA2-Ab, and null models for PCA-Ab (for which the null hypothesis could not be rejected). Note the differing X-axis scales. The plots illustrate the rationale for the three-category model; for TPO-Ab, GAD-Ab and IA2-Ab, a tendency is seen for SNPs associated with autoantibody positivity (high $|Z_d|$) to be associated with T1D also (high $|Z_a|$). This tendency is not seen for PCA-Ab, and is minimal for non-MHC SNPs in GAD-Ab. Further analysis of the plot for TPOAb positivity (top left) is shown below.

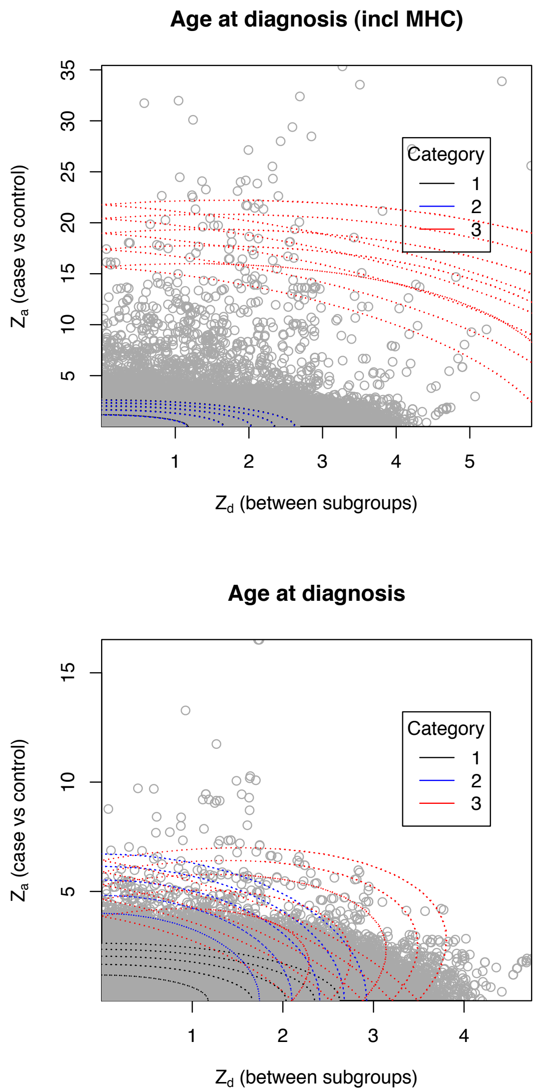
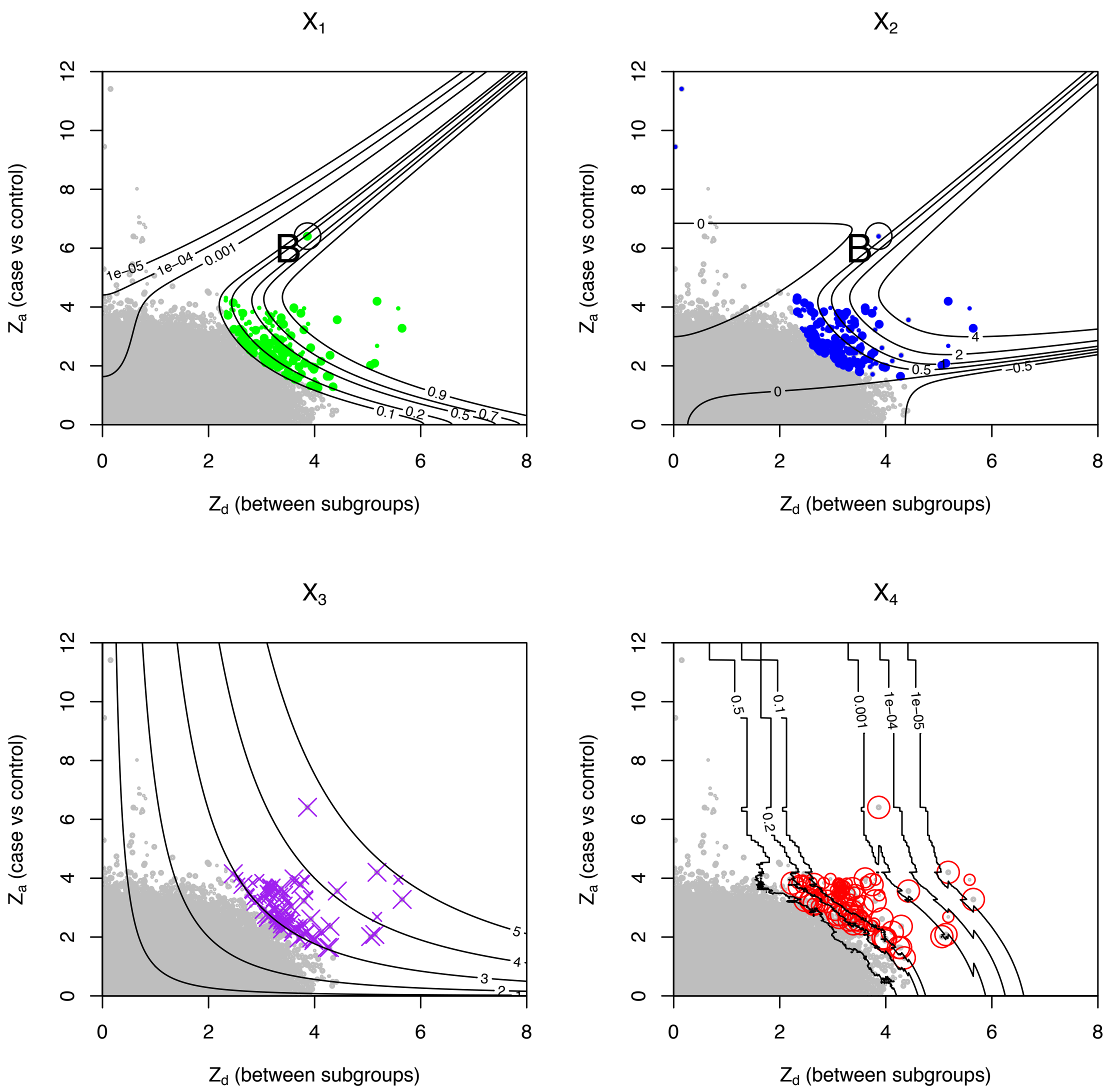
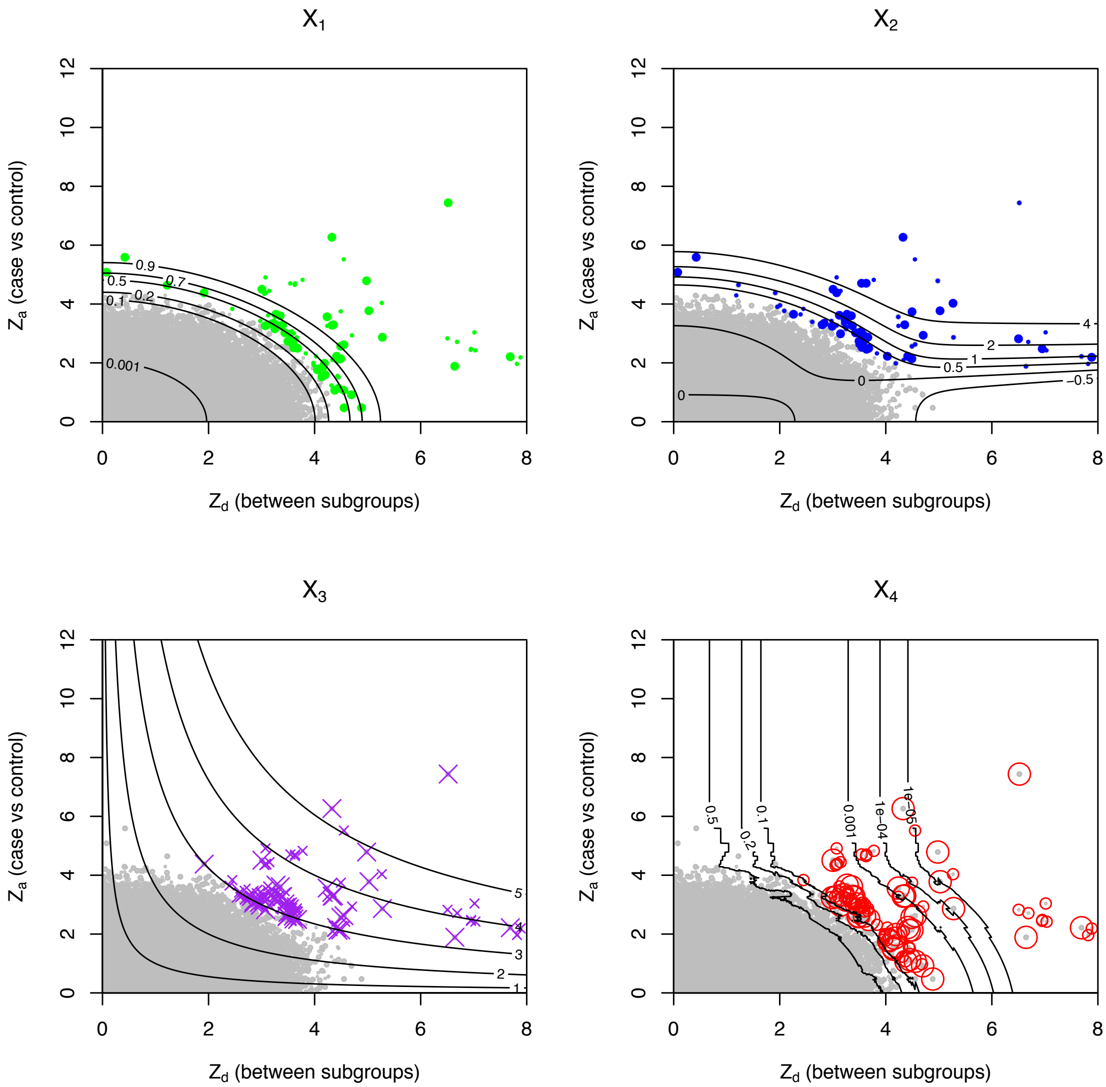


Figure 6: Observed Z_a and Z_d scores for T1D subclassified by age at diagnosis. Non-MHC SNPs are shown in red.

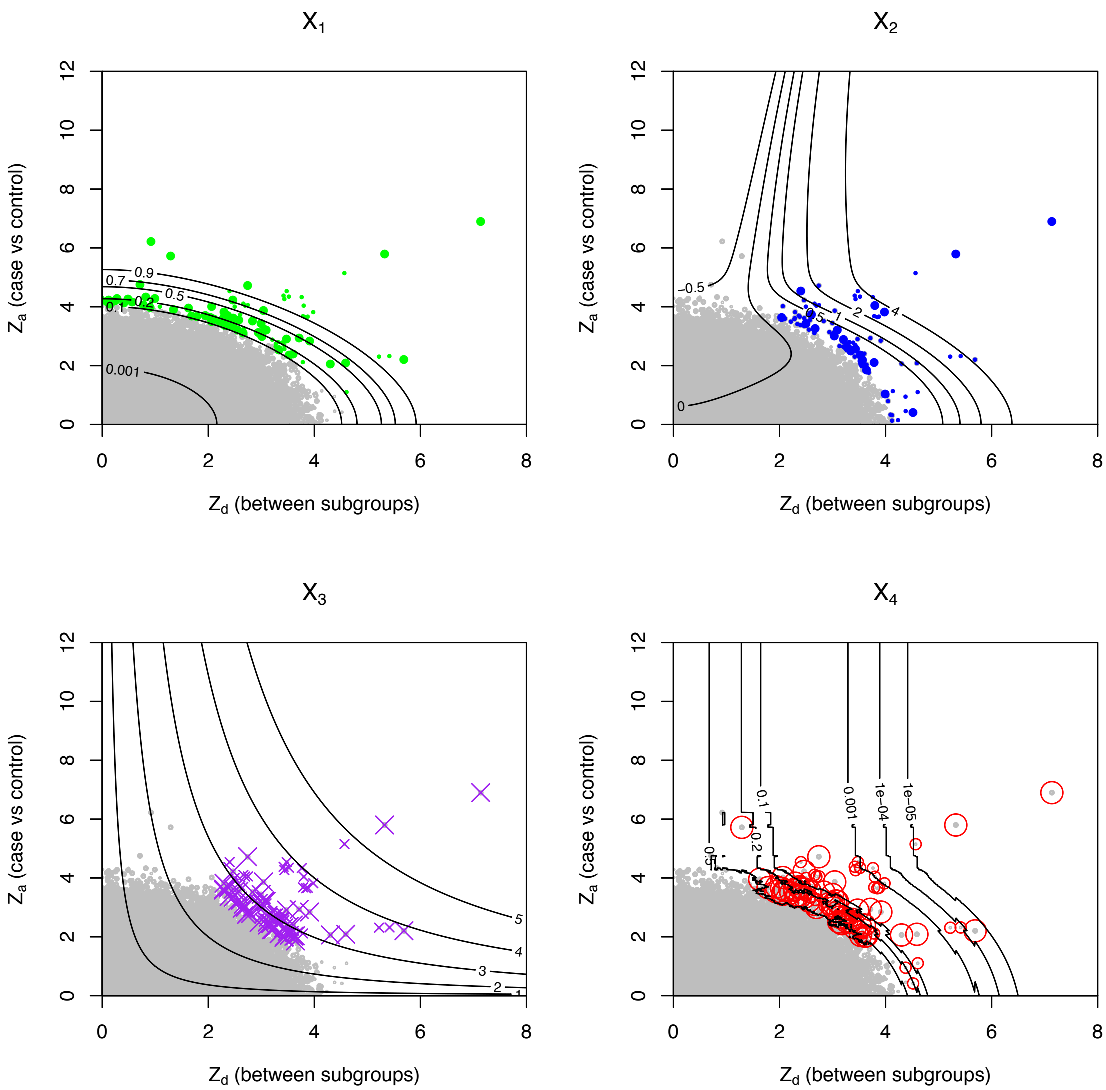
(a) T1D/RA comparison



(b) T1D/T2D comparison



(c) T2D/RA comparison



(d) GD/HT (ATD) comparison

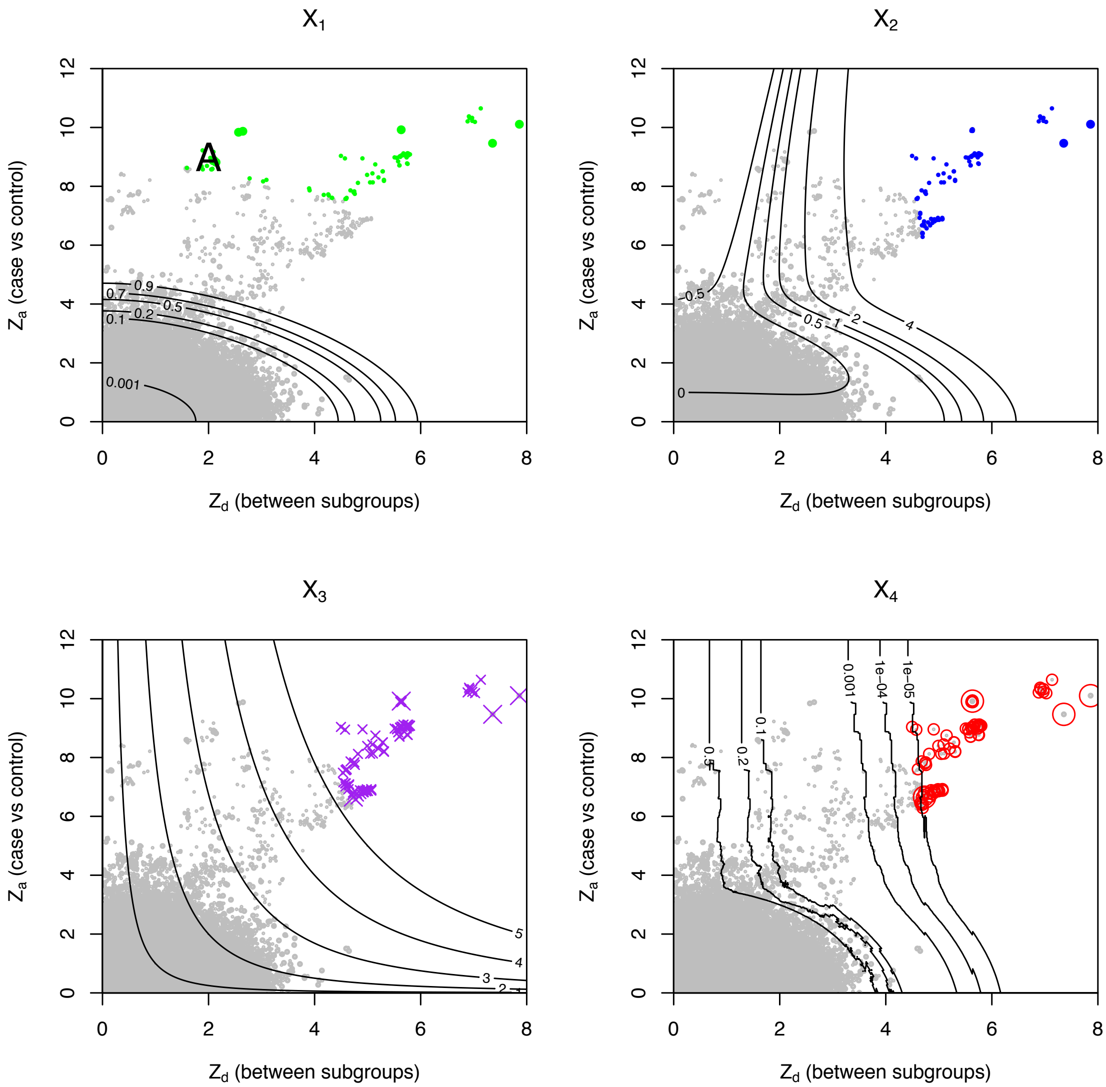


Figure 7: We demonstrate all four test statistics for single-SNP effects in the comparisons betwen T1D/T2D/RA, and between GD and HT (preceding pages). The top 100 SNPs for each test statistic are highlighted, with larger symbols corresponding to SNPs with non-zero weights after applying LDAK [2]; that is, the SNPs which contributed to the model fit. Contours of each test statistic are shown in grey.
Differences are evident in the behaviour of the test statistics X_1 and X_2 between the two datasets; X_3 and X_4 are more robust. The different null hypotheses between X_3 and X_4 are responsible for the difference in shape near the line $Z_a = 0$. Contours of X_4 are jagged due to the dependence of this statistic on the distribution of Z scores.
All methods primarily identified SNPs with both high $|Z_a|$ and $|Z_d|$ scores as contributors. As evident from the comparison between GH and HT, the statistic X_1 is vulnerable to falsely declaring SNPs as subgroup-differentiating despite low $|Z_d|$ scores (labeled 'A', top left panel, GD/HT). This arises due to the full model having a markedly higher value of σ_3 than σ_2, leading to SNPs with very high $|Z_a|$ values having a high posterior probability of category 3 membership.
This is partially able to be overcome by combining the test statistics X_1 and X_2 into one, which we typically do by only considering X_2 scores in SNPs with X_1 greater than some cutoff. However, this is not always effective, as is evident from the above figure for T1D/T2D. In this case, as discussed in the main paper, almost all SNPs with high Z_a also had high Z_d, meaning that the two distributions forming categories 2 and 3 under the null model were essentially the same. This led to the fitted parameters of the null model supporting SNPs falling into two distributions; one with identity covariance matrix, and the other with $var(Z_d) > 1$, $var(Z_a) = 1$ (see fitted parameters).
The different alternative hypothesis for X_4 (different population MAFs in subgroups without requiring association with the phenotype overall) meant that SNPs with low $|Z_a|$ scores may be identified by X_4 in addition to those identified by X_1, X_2 and X_3 (contour lines on bottom right panel, both figures). SNPs which are isolated may be missed by both X_1 and X_2 (label 'B', top two panels, T1D/RA), due to the fitted distribution of SNPs in category 3 tending to be driven by clusters of SNPs.
Given these results, we consider X_3 and X_4 to generally be the most appropriate measure for single SNP effects, although in appropriate circumstances X_2 can be used alone or conditionally on X_1.

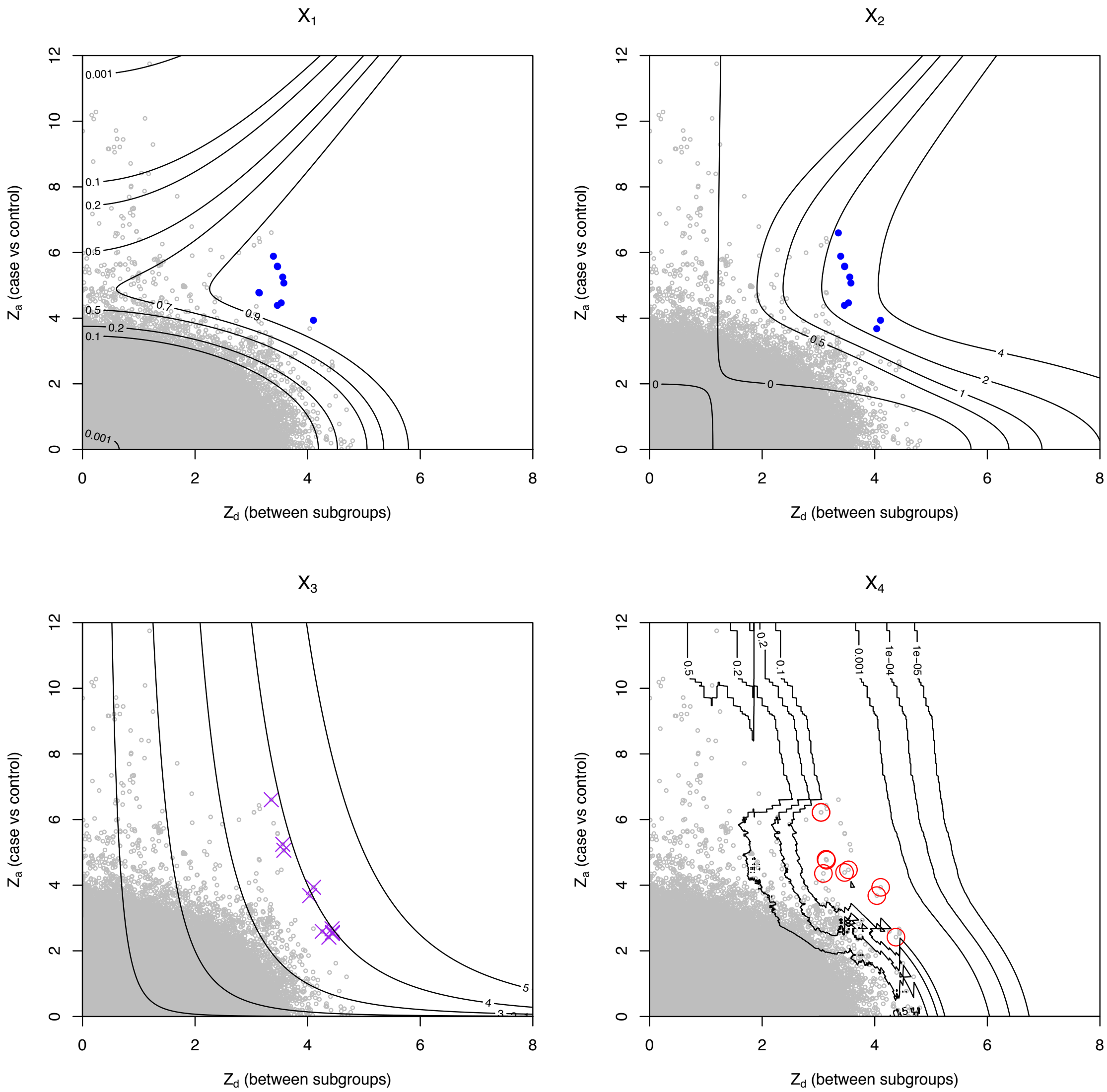


Figure 8: We assessed the SNPs responsible for the observed difference in pseudo-likelihood ratio for our analysis of TPOAb positivity in T1D. SNPs in the MHC region were removed from the analysis (co-ordinates 25-38 Mb, GChR build 37). We combined X_1 and X_2 into a single test statistic, by only considering SNPs with $X_1 > 0.7$ and then considering the top SNPs for X_2. The top ten SNPs for $X_2|X_1 > 0.7$ (blue, top two panels), X_3 (purple, bottom left panel), and X_4 (red, bottom right panel) are shown. Contours of each summary statistic are shown as black lines. Details of SNPs are shown in the supplementary tables.

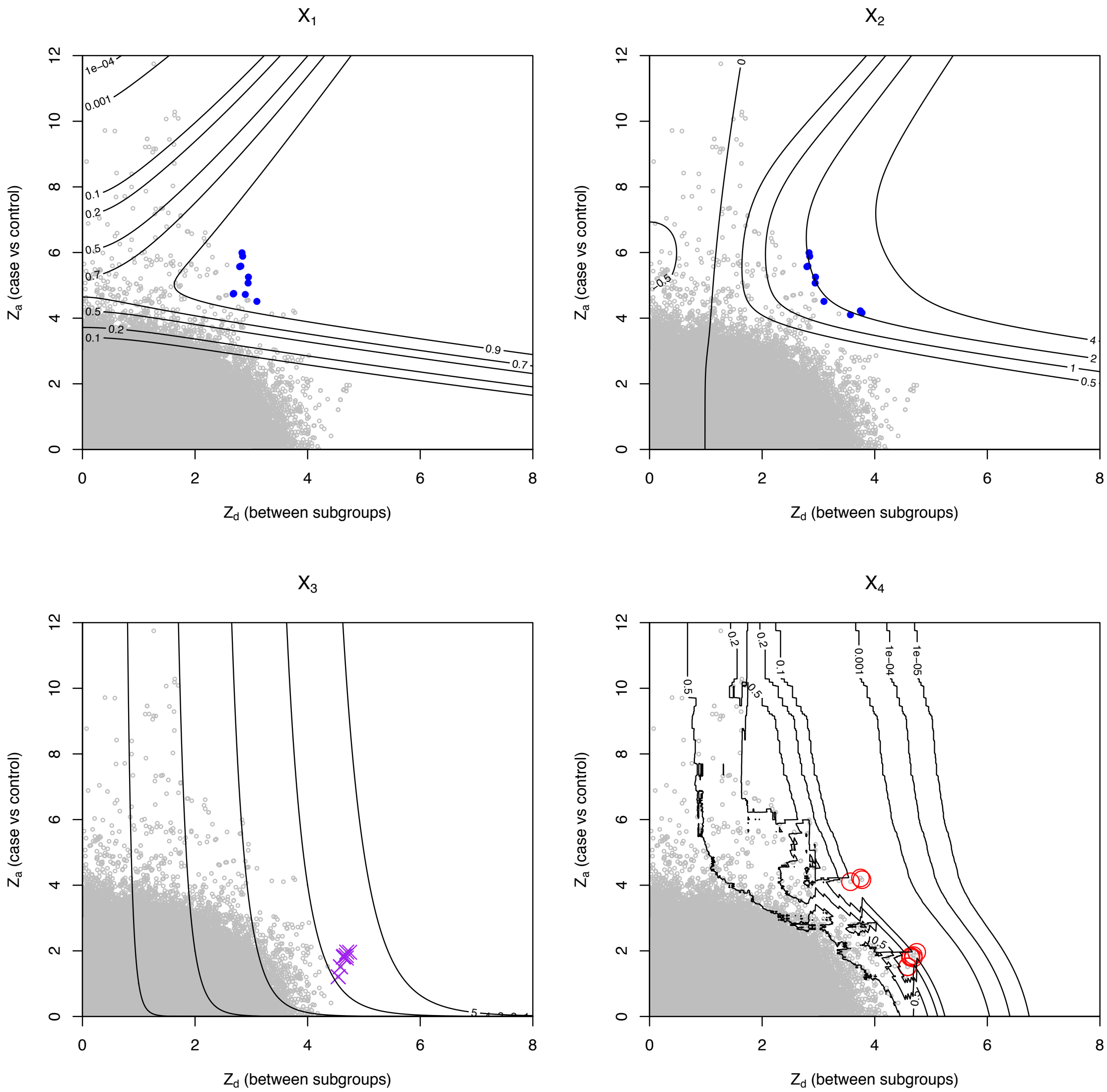


Figure 9: We assessed the SNPs responsible for the observed difference in pseudo-likelihood ratio for our analysis of age at diagnosis in T1D. SNPs in the MHC region were removed from the analysis (co-ordinates 25-38 Mb, GChR build 37). We combined X_1 and X_2 into a single test statistic, by only considering SNPs with $X_1 > 0.7$ and then considering the top SNPs for X_2. The top ten SNPs for $X_2|X_1 > 0.7$ (blue, top two panels), X_3 (purple, bottom left panel), and X_4 (red, bottom right panel) are shown. Contours of each summary statistic are shown as black lines. Details of SNPs are shown in the supplementary tables.

References

[1] The Wellcome trust case control consortium (2007) Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. Nature 447: 661-678.

[2] Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. American Journal of Human Genetics 91: 1011-1021.

A method for identifying genetic heterogeneity within phenotypically-defined disease subgroups

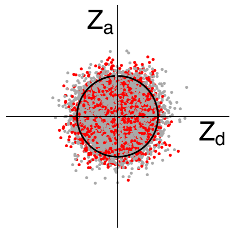
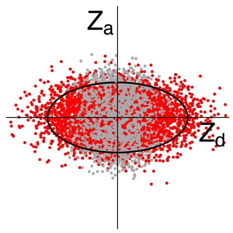
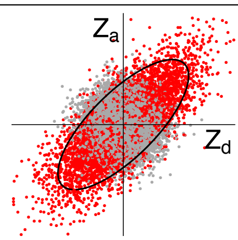
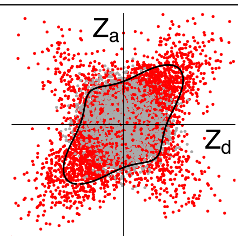
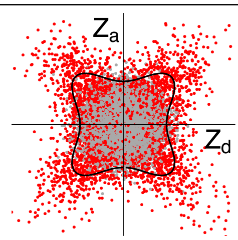
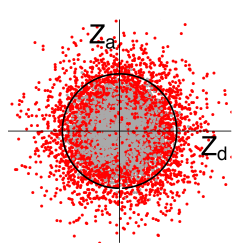
Supplementary Tables

James Liley, John A Todd and Chris Wallace

November 8, 2016

List of Tables

1 Forms of genetic architecture under different causes of heterogeneity . . . 3
2 Model parameters for autoantibody positivity in T1D . . . 4
3 Model parameters for age at diagnosis in T1D . . . 4
4 Top SNPs differentiating T1D and RA . . . 5
5 Top SNPs differentiating T1D and T2D . . . 6
6 Top SNPs differentiating T2D and RA . . . 7
7 Top SNPs differentiating GD and HT subgroups of ATD. . . . 8
8 Top SNPs for TPOAb positivity in T1D . . . 9
9 Top SNPs for age at diagnosis in T1D . . . 10

Form	$r_g^{(1)}$	$r_g^{(2)}$	ρ	τ	σ_3	Phenomenon
	1	0	0	> 1	1	H_0: $Z_d, Z_a \sim N(0, I_2)$; all-environmental cause for subgroup phenotype
	$\ll 1$	0	0	> 1	1	H_0: Z_d, Z_a independent; subgrouping phenotype independent of main phenotype;
	$1/ < 1$	$\gg 0$	$\gg 0$	> 1	> 1	H_1: Z_d, Z_a correlated; eg. same pathways; different heritability (age-of-onset)
	< 1	> 0	$\gg 0$	> 1	> 1	H_1: Z_d, Z_a mostly correlated, some anticorrelation; eg. most variants associated with subgroup 1, some with subgroup 2
	< 1	0	$\gg 0$	> 1	> 1	H_1: Z_d, Z_a both correlated and anticorrelated; eg. variants either associated only with subgroup 1 or only with subgroup 2
	< 1	0	0	> 1	> 1	H_1: $var(Z_d) > 1$ and $var(Z_a) > 1$ but not correlated; general shared genetic architecture between subgrouping phenotype and main phenotype, effect sizes independent

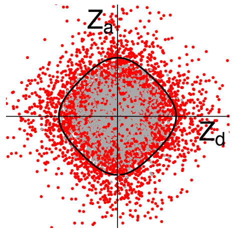
	< 1	0	0	> 1	> 1	H_1: shared genetic architecture between subgrouping phenotype and main phenotype, effect sizes dependent but not correlated or anticorrelated

Table 1: Heterogeneity between case subgroups may arise in multiple ways, some of which are illustrated here. Plots show the distribution of Z_d and Z_a for SNPs in category 3 (those which differentiate subgroups). Column $r_g^{(1)}$ corresponds to genetic correlation in method 1 (between Z scores for control vs subgroup 1 and control vs subgroup 2), and column $r_g^{(2)}$ to genetic correlation in method 2 (between Z_a and Z_d); see supplementary material, section 4. SNPs in category 1 (not differentiating cases/controls and not differentiating subgroups) are shown in grey for reference, and SNPs in category 2 are omitted. In the first two rows, the pathology leading to heterogeneity is genetically independent of the pathology leading to the main phenotype; our null hypothesis. The test $r_g^{(1)} < 1$ will reject H_0 for the scenario in row 2, as well as other scenarios. The test $r_g^{(2)} \neq 0$ rejects H_0 for the scenario in row 3, but is weakened in the scenario in row 4 due to the anticorrelation, and will not be able to reject H_0 for rows 5-7. Since ρ detects correlation and anticorrelation simultaneously, it will additionally reject H_0 for row 4 and will not be weakened in row 3. However, it is necessary to test for $\sigma_3 > 1$ to reject H_0 for rows 5 and 6.

	Model	π_1	π_2	π_3	σ_2	σ_3	τ	ρ	p-value
TPO-Ab	Full	0.511	0.487	2.407×10^{-3}	0.994	6.545	1.552	0.991	$< 1 \times 10^{-20}$
	Null	0.987	2.333×10^{-3}	0.011	6.634	-	1.308	-	
TPO-Ab	Full	0.997	2.898×10^{-4}	3.031×10^{-3}	4.698	2.291	1.497	0.338	1.5×10^{-4}
no MHC	Null	0.989	1.882×10^{-3}	9.087×10^{-3}	3.11	-	1.318	-	
GAD-Ab	Full	0.995	3.557×10^{-3}	1.057×10^{-3}	2.832	8.866	2.295	5.484	$< 1 \times 10^{-20}$
	Null	0.997	2.328×10^{-3}	3.002×10^{-4}	6.639	-	2.153	-	
GAD-Ab	Full	0.997	2.9×10^{-3}	3.434×10^{-4}	2.279	4.531	1.055	3.424	0.002
no MHC	Null	0.792	1.883×10^{-3}	0.206	3.111	-	0.997	-	
IA2-Ab	Full	0.995	3.275×10^{-3}	1.244×10^{-3}	2.804	8.291	3.027	1.575	$< 1 \times 10^{-20}$
	Null	0.997	2.287×10^{-3}	3.805×10^{-4}	6.674	-	3.852	-	
IA2-Ab	Full	0.998	1.362×10^{-3}	7.904×10^{-4}	3.318	2.212	2.145	0	0.008
no MHC	Null	0.998	1.88×10^{-3}	2.073×10^{-4}	3.112	-	2.889	-	
PCA-Ab	Full	0.997	2.336×10^{-3}	3.413×10^{-4}	6.631	0.37	2.097	0.422	> 0.5
	Null	0.998	2.335×10^{-3}	1.276×10^{-4}	6.632	-	2.54	-	
PCA-Ab	Full	0.997	2.759×10^{-3}	1.303×10^{-4}	2.508	5.58	2.256	0	> 0.5
no MHC	Null	0.998	1.884×10^{-3}	1.384×10^{-4}	3.111	-	2.5	-	

Table 2: Parameters of models fitted to T1D autoantibody positivity data. With MHC retained (co-ordinates 25-38 Mb, GChR build 37) all full models fit better than null models with the exception of those fitted to PCA-Ab positivity. With MHC removed, effect sizes were lower, but the null hypothesis could be rejected for TPOA-Ab positivity, with weaker evidence for rejecting the null hypothesis for GAD-Ab and IA2-Ab. In most cases, there was evidence of SNPs differentiating subgroups (typically, fitted $\tau > 1$). There were generally a small number of SNPs which strongly differentiated cases and controls (a small value of π_2, π_3 corresponding to the larger value of σ_2, σ_3). P-values were computed against the null distribution of cPLR for random subgroups, which showed good agreement with the asymptotic mixture-χ^2 distribution (see supplementary figure **??**). P-values shown are unadjusted for multiple testing.

Age	Full	0.898	0.099	2.4×10^{-3}	0.96	6.558	1.601	3.644	4.9×10^{-37}
	Null	0.885	2.338×10^{-3}	0.113	6.631	-	0.945	-	
Age	Full	0.997	1.881×10^{-4}	3.035×10^{-3}	5.257	2.372	1.159	1.315	0.007
no MHC	Null	0.782	1.891×10^{-3}	0.216	3.107	-	0.97	-	

Table 3: Parameters of models fitted to age at diagnosis in T1D, considered as a parameter rather than defining subgroups. The full model fit significantly better than the null model when the MHC region was included or excluded. Plotted Z_a and Z_d scores are shown in supplementary figure **??**. The fitted models show evidence of SNPs associated with age at diagnosis (fitted $\tau > 1$). P-values were computed against the null distribution of cPLR for random subgroups, which showed good agreement with the asymptotic mixture-χ^2 distribution (see supplementary figure **??**).

SNP details				Z scores		Values (rank)				Summary statistics	
SNP	Chr	Pos	Gene	$\|Z_d\|(p)$	$\|Z_a\|$	X_1	X_2	X_3	X_4	p-val (X_3)	FDR (X_4)
rs12045559	1	113708908	*PTPN22*	2.892 (3.8×10^{-3})	3.217	0.332	0.383	3.003	0.061 (15)	5.593×10^{-5}	0.169
rs415024	5	9445358		3.051 (2.3×10^{-3})	2.941	0.344	0.394	3.012	0.083 (20)	5.415×10^{-5}	0.247
rs1010599	5	35944231	*IL7R*	3.367 (7.6×10^{-4})	2.881	0.538	0.72 (15)	3.186 (16)	0.078 (18)	2.092×10^{-5}	0.23
rs4024109	5	35955375	*IL7R*	3.307 (9.4×10^{-4})	2.792	0.456	0.561 (18)	3.115 (19)	0.107	3.072×10^{-5}	0.357
rs17085170	5	95198087		4.291 (1.8×10^{-5})	2.365	0.832	1.353 (9)	3.477 (10)	0.022 (12)	4.695×10^{-6}	0.046
rs3114834	7	109192112		2.649 (8.1×10^{-3})	3.787	0.308	0.351	3.005	0.072 (17)	5.504×10^{-5}	0.213
rs12549890	8	21045174		3.535 (4.1×10^{-4})	2.459	0.449	0.519 (19)	3.11 (20)	0.141	3.154×10^{-5}	0.48
rs16874205	8	107271324		4.428 (9.5×10^{-6})	3.565	0.994	4.48 (3)	4.102 (4)	6.669×10^{-4} (3)	2.768×10^{-7}	3.612×10^{-3}
rs4076319	10	85129122		4.124 (3.7×10^{-5})	2.165	0.656	0.773 (14)	3.285 (15)	0.066 (16)	1.269×10^{-5}	0.186
rs10736277	10	121705898		3.415 (6.4×10^{-4})	3.411	0.775	1.434 (8)	3.413 (12)	0.021 (11)	6.581×10^{-6}	0.045
rs7912574	10	121717404		3.265 (1.1×10^{-3})	2.957	0.499	0.648 (16)	3.153 (17)	0.084	2.528×10^{-5}	0.25
rs2065660	10	121754185		3.557 (3.8×10^{-4})	3.031	0.73	1.23 (11)	3.361 (14)	0.036 (13)	8.529×10^{-6}	0.091
rs6578252	11	2226817	*INS*	3.481 (5×10^{-4})	2.576	0.473	0.572 (17)	3.13 (18)	0.113	2.87×10^{-5}	0.374
rs705698	12	54670954	*IKZF4*	5.058 (4.2×10^{-7})	2.016	0.934	0.871 (13)	3.656 (8)	1.241×10^{-3} (6)	2.016×10^{-6}	6.997×10^{-3}
rs705702	12	54676903	*IKZF4*	5.135 (2.8×10^{-7})	2.086	0.956	1.067 (12)	3.737 (6)	8.403×10^{-4} (4)	1.371×10^{-6}	4.78×10^{-3}
rs2292239	12	54768447	*IKZF4*	5.651 (1.6×10^{-8})	3.278	1	4.991 (2)	4.663 (2)	8.184×10^{-6} (1)	2.35×10^{-8}	2.586×10^{-5}
rs4766443	12	109864518		3.372 (7.5×10^{-4})	3.644	0.813	1.623 (7)	3.466 (11)	0.016 (10)	5.058×10^{-6}	0.069
rs10774613	12	110008885	*SH2B3*	3.929 (8.5×10^{-5})	2.614	0.769	1.279 (10)	3.403 (13)	0.055 (14)	6.902×10^{-6}	0.152
rs1265566	12	110179096	*SH2B3*	3.612 (3×10^{-4})	3.975	0.942	2.764 (5)	3.736 (7)	6.384×10^{-3} (8)	1.373×10^{-6}	0.027
rs17696736	12	110949538	*SH2B3*	3.867 (1.1×10^{-4})	6.409	0.237	0.256	4.622 (3)	9.922×10^{-4} (5)	2.823×10^{-8}	5.809×10^{-3}
rs16961362	15	33731898		3.799 (1.5×10^{-4})	3.23	0.896	2.123 (6)	3.588 (9)	0.011 (9)	2.769×10^{-6}	0.052
rs1711029	15	51491702		5.178 (2.2×10^{-7})	4.201	1	7.503 (1)	4.809 (1)	8.759×10^{-6} (2)	1.187×10^{-8}	2.748×10^{-5}
rs12924729	16	11095284	*DEXI*	3.741 (1.8×10^{-4})	3.784	0.952	2.92 (4)	3.756 (5)	3.989×10^{-3} (7)	1.278×10^{-6}	0.015
rs1942707	18	60768535		4.279 (1.9×10^{-5})	1.642	0.411	0.108	3.052	0.083 (19)	4.376×10^{-5}	0.247

Table 4: Top 20 SNPs differentiating T1D and RA (MHC removed), considered as subgroups of a general autoimmune phenotype, for each of four summary statistics. Positions are in NCBI build 36. Because of the large number of SNPs with evidence for differentiating the subgroups, only SNPs with non-zero weights after applying the LDAK procedure are included in this table. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0' \| |Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0' | X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores		Values (rank)				Summary statistics	
SNP	Chr	Pos	Gene	$\|Z_d\|(p)$	$\|Z_a\|$	X_1	X_2	X_3	X_4	p-val (X_3)	FDR (X_4)
rs17013326	1	113801358	*PTPN22*	3.007 (2.6×10^{-3})	4.509	0.957	3.169 (7)	3.741 (9)	3.431×10^{-3} (13)	9.501×10^{-7}	0.026
rs1230666	1	113885452	*PTPN22*	4.327 (1.5×10^{-5})	6.265	1	14.651 (2)	5.283 (2)	1.512×10^{-5} (6)	9.121×10^{-10}	8.1×10^{-5}
rs6679677	1	114015850	*PTPN22*	6.52 (7×10^{-11})	7.437	1	22.963 (1)	6.999 (1)	7.052×10^{-11} (1)	3.583×10^{-14}	2.658×10^{-10}
rs6661817	1	114159076	*PTPN22*	3.372 (7.5×10^{-4})	3.26	0.573	0.801	3.311 (17)	0.021	7.411×10^{-6}	0.128
rs3811019	1	114183625	*PTPN22*	3.353 (8×10^{-4})	3.609	0.768	1.413 (14)	3.489 (11)	9.448×10^{-3} (17)	3.046×10^{-6}	0.057
rs12061474	1	201120971	*PIK3C2B*	3.252 (1.1×10^{-3})	3.153	0.422	0.511	3.198 (20)	0.036	1.366×10^{-5}	0.234
rs903228	2	53603700		0.077 (0.94)	5.085	0.726	1.425 (13)	0.738	≥ 0.5	0.372	≥ 0.5
rs7666328	4	116140909		0.425 (0.67)	5.593	0.954	3.307 (6)	1.706	≥ 0.5	0.021	≥ 0.5
rs2544677	5	86435018		3.272 (1.1×10^{-3})	3.661	0.753	1.36 (15)	3.476 (12)	9.796×10^{-3} (18)	3.278×10^{-6}	0.059
rs2112168	5	86440646		3.199 (1.4×10^{-3})	3.335	0.503	0.664	3.272 (18)	0.03	9.241×10^{-6}	0.192
rs7917983	10	114722872	*TCF7L2*	4.357 (1.3×10^{-5})	3.303	0.973	2.988 (8)	3.752 (8)	8.059×10^{-4} (9)	9.129×10^{-7}	5.483×10^{-3}
rs7901275	10	114722896	*TCF7L2*	4.331 (1.5×10^{-5})	3.287	0.969	2.895 (9)	3.732 (10)	8.347×10^{-4} (10)	9.996×10^{-7}	5.325×10^{-3}
rs7901695	10	114744078	*TCF7L2*	7.691 (1.5×10^{-14})	2.215	1	1 (16)	3.93 (5)	1.566×10^{-10} (2)	4.229×10^{-7}	5.471×10^{-10}
rs12243326	10	114778805	*TCF7L2*	6.645 (3×10^{-11})	1.889	1	0.447 (20)	3.372 (14)	7.82×10^{-8} (3)	5.464×10^{-6}	3.365×10^{-7}
rs3741939	12	3517792		4.885 (1×10^{-6})	0.473	0.713	0	1.386	0.016 (19)	0.066	0.096
rs705698	12	54670954	*IKZF4*	4.494 (7×10^{-6})	2.569	0.907	1.492 (12)	3.324 (16)	2.223×10^{-3} (12)	7.058×10^{-6}	0.016
rs705702	12	54676903	*IKZF4*	4.554 (5.3×10^{-6})	2.624	0.932	1.673 (11)	3.383 (13)	1.886×10^{-3} (11)	5.177×10^{-6}	0.013
rs2292239	12	54768447	*IKZF4*	5.026 (5×10^{-7})	3.78	1	5.14 (4)	4.31 (4)	1.186×10^{-5} (5)	8.463×10^{-8}	6.252×10^{-5}
rs4766443	12	109864518	*SH2B3*	3.44 (5.8×10^{-4})	3.025	0.474	0.586	3.21 (19)	0.032	1.264×10^{-5}	0.208
rs10774613	12	110008885	*SH2B3*	4.415 (1×10^{-5})	2.223	0.783	0.804 (17)	3.049	6.221×10^{-3} (16)	3.07×10^{-5}	0.042
rs1265566	12	110179096	*SH2B3*	3.398 (6.8×10^{-4})	3.276	0.601	0.867	3.332 (15)	0.021	6.733×10^{-6}	0.126
rs17696736	12	110949538	*SH2B3*	4.981 (6.3×10^{-7})	4.788	1	8.774 (3)	4.876 (3)	2.532×10^{-6} (4)	6.675×10^{-9}	1.263×10^{-5}
rs12924729	16	11095284	*SH2B3*	4.236 (2.3×10^{-5})	3.563	0.98	3.46 (5)	3.859 (6)	4.771×10^{-4} (8)	5.72×10^{-7}	3.029×10^{-3}
rs7193144	16	52368187	*FTO*	4.493 (7×10^{-6})	2.139	0.803	0.747 (18)	3.011	4.761×10^{-3} (14)	3.735×10^{-5}	0.033
rs8050136	16	52373776	*FTO*	4.442 (8.9×10^{-6})	2.127	0.768	0.689 (19)	2.986	5.919×10^{-3} (15)	4.257×10^{-5}	0.041
rs9926289	16	52378004	*FTO*	4.19 (2.8×10^{-5})	1.985	0.519	0.337	2.8	0.018 (20)	1.203×10^{-4}	0.109
rs2542151	18	12769947	*PTPN2*	5.278 (1.3×10^{-7})	2.866	0.998	2.619 (10)	3.797 (7)	7.514×10^{-5} (7)	7.456×10^{-7}	4.5×10^{-4}

Table 5: Top 20 SNPs differentiating T1D and T2D (MHC removed), considered as subgroups of a general diabetic phenotype, for each of four summary statistics. Positions are in NCBI build 36. Because of the large number of SNPs with evidence for differentiating the subgroups, only SNPs with non-zero weights after applying the LDAK procedure are included in this table. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0' \| |Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0' | X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores		Values (rank)				Summary statistics	
SNP	Chr	Pos	Gene	$\|Z_d\|(p)$	$\|Z_a\|$	X_1	X_2	X_3	X_4	p-val (X_3)	FDR (X_4)
rs10858002	1	113794974	*PTPN22*	2.997 (2.7×10^{-3})	3.44	0.413	0.472 (11)	3.171 (11)	0.042 (11)	1.492×10^{-5}	0.27
rs17013326	1	113801358	*PTPN22*	2.465 (0.01)	4.223	0.743	0.933 (6)	3.072 (15)	0.029 (8)	2.656×10^{-5}	0.172
rs1230666	1	113885452	*PTPN22*	5.326 (1×10^{-7})	5.801	1	10.802 (2)	5.515 (2)	1.508×10^{-7} (2)	6.511×10^{-11}	1.391×10^{-6}
rs6679677	1	114015850	*PTPN22*	7.137 (9.5×10^{-13})	6.9	1	19.721 (1)	7.039 (1)	9.563×10^{-13} (1)	3.084×10^{-15}	8.924×10^{-12}
rs3811019	1	114183625	*PTPN22*	2.943 (3.3×10^{-3})	3.132	0.207	0.199	3.019 (19)	0.11	3.64×10^{-5}	≥ 0.5
rs10931347	2	189007813		3.014 (2.6×10^{-3})	2.984	0.17	0.163	3.002 (20)	0.14	3.952×10^{-5}	≥ 0.5
rs6846031	4	178394297		3.474 (5.1×10^{-4})	2.908	0.324	0.379 (13)	3.23 (9)	0.053 (12)	1.041×10^{-5}	0.326
rs11970411	6	138220854	*TNFAIP3*	3.601 (3.2×10^{-4})	2.402	0.161	0.181	3.052 (17)	0.162	2.948×10^{-5}	≥ 0.5
rs3114834	7	109192112		3.092 (2×10^{-3})	3.204	0.305	0.328 (14)	3.137 (12)	0.07 (14)	1.822×10^{-5}	0.452
rs16874205	8	107271324		3.042 (2.4×10^{-3})	3.866	0.748	1.213 (5)	3.355 (7)	0.013 (6)	5.053×10^{-6}	0.077
rs2104286	10	6139051	*IL2RA*	3.911 (9.2×10^{-5})	2.842	0.554	0.804 (8)	3.433 (4)	0.03 (9)	3.19×10^{-6}	0.185
rs7917983	10	114722872	*TCF7L2*	2.296 (0.02)	3.823	0.356	0.283 (16)	2.828	0.08 (16)	1.079×10^{-4}	0.478
rs7901275	10	114722896	*TCF7L2*	2.064 (0.04)	4.016	0.407	0.268 (17)	2.709	0.093	2.092×10^{-4}	≥ 0.5
rs7901695	10	114744078	*TCF7L2*	5.689 (1.3×10^{-8})	2.203	0.987	4.345 (3)	3.861 (3)	6.625×10^{-5} (3)	3.159×10^{-7}	5.521×10^{-4}
rs12243326	10	114778805	*TCF7L2*	4.592 (4.4×10^{-6})	2.087	0.587	0.912 (7)	3.326 (8)	0.01 (4)	6.013×10^{-6}	0.056
rs10736277	10	121705898	*TCF7L2*	2.992 (2.8×10^{-3})	3.064	0.194	0.188	3.021 (18)	0.113	3.562×10^{-5}	≥ 0.5
rs770738	12	10034164	*DEXI*	2.364 (0.02)	3.698	0.298	0.236	2.838	0.084 (19)	1.021×10^{-4}	0.478
rs1495377	12	69863368	*TSPAN8*	4.302 (1.7×10^{-5})	2.063	0.369	0.483 (10)	3.186 (10)	0.032 (10)	1.37×10^{-5}	0.197
rs7961581	12	69949369	*TSPAN8*	3.715 (2×10^{-4})	2.935	0.486	0.655 (9)	3.373 (6)	0.025 (7)	4.576×10^{-6}	0.145
rs551714	13	20436464		2.381 (0.02)	3.659	0.28	0.221	2.838	0.086 (20)	1.021×10^{-4}	0.49
rs1711029	15	51491702		2.742 (6.1×10^{-3})	4.722	0.971	2.145 (4)	3.424 (5)	0.011 (5)	3.406×10^{-6}	0.058
rs1054028	16	22834715		3.063 (2.2×10^{-3})	3.222	0.302	0.322 (15)	3.127 (13)	0.074 (15)	1.903×10^{-5}	0.458
rs7193144	16	52368187	*FTO*	2.469 (0.01)	3.622	0.29	0.245	2.888	0.081 (18)	7.761×10^{-5}	0.491
rs8050136	16	52373776	*FTO*	2.575 (0.01)	3.548	0.287	0.255	2.936	0.081 (17)	5.855×10^{-5}	0.489
rs896136	17	35904973	*IKZF3*	3.02 (2.5×10^{-3})	3.122	0.23	0.231	3.061 (16)	0.094	2.826×10^{-5}	≥ 0.5

Table 6: Top 20 SNPs differentiating T2D and RA (MHC removed), considered as subgroups of a general phenotype, for each of four summary statistics. Positions are in NCBI build 36. Because of the large number of SNPs with evidence for differentiating the subgroups, only SNPs with non-zero weights after applying the LDAK procedure are included in this table. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0' \| |Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0' | X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores		Values (rank)				Summary statistics	
SNP	Chr	Pos	Gene	$\|Z_d\|(p)$	$\|Z_a\|$	X_1	X_2	X_3	X_4	p-val (X_3)	FDR (X_4)
rs6679677	1	114105331	*PTPN22*	2.568 (0.01)	9.84	1	1.994 (8)	4.019 (7)	0.01 (14)	2.433×10^{-6}	0.045
rs2476601	1	114179091	*PTPN22*	2.649 (8.1×10^{-3})	9.88	1	2.224 (7)	4.109 (5)	8.063×10^{-3} (11)	1.784×10^{-6}	0.035
rs7554023	1	160162988	*ATF6??*	3.625 (2.9×10^{-4})	1.855	0.11	0.064	2.899	0.012 (18)	1.529×10^{-4}	0.054
X2-204400444- -CA-DELETION	2	204400444	*CTLA4*	2.142 (0.03)	8.822	1	1.015 (11)	3.435 (10)	0.044	1.803×10^{-5}	0.193
X2-204408002- -CCT-DELETION	2	204408002	*CTLA4*	2.103 (0.04)	8.879	1	0.907 (12)	3.399 (12)	0.041	2.047×10^{-5}	0.179
rs58716662	2	204423821	*CTLA4*	2.447 (0.01)	5.968	1	1.93 (9)	3.294 (14)	0.02	3.026×10^{-5}	0.089
rs78960870	2	204458162	*CTLA4*	2.171 (0.03)	6.143	1	1.335 (10)	3.071 (20)	0.043	7.274×10^{-5}	0.188
rs13030124	2	204402508	*CTLA4*	2.091 (0.04)	8.871	1	0.879 (13)	3.385 (13)	0.042	2.146×10^{-5}	0.184
rs3997876	2	179005067	*PRKRA*	7.863 (3.8×10^{-15})	10.102	1	25.336 (1)	8.548 (1)	3.008×10^{-14} (1)	8.23×10^{-17}	1.046×10^{-13}
rs3997878	2	179004872	*PRKRA*	7.358 (1.9×10^{-13})	9.467	1	22.077 (2)	8.003 (2)	1.777×10^{-12} (2)	5.231×10^{-15}	6.2×10^{-12}
rs6720771	2	154461782		4.091 (4.3×10^{-5})	0.428	0.05	0.051	1.927	0.015 (20)	0.011	0.065
rs6723546	2	154617139		4.137 (3.5×10^{-5})	0.76	0.07	0.065	2.352	8.97×10^{-3} (12)	1.833×10^{-3}	0.039
rs12638263	3	187278549	*BCL6*	3.459 (5.4×10^{-4})	2.263	0.176	0.116	3.003	0.011 (17)	9.761×10^{-5}	0.05
rs34244025	9	138290559	*ESP33*	4.649 (3.3×10^{-6})	1.426	0.414	0.497	3.135 (19)	6.905×10^{-4} (6)	5.625×10^{-5}	2.932×10^{-3}
rs34775390	9	138293196	*ESP33*	4.595 (4.3×10^{-6})	1.508	0.414	0.494	3.169 (18)	6.897×10^{-4} (5)	4.904×10^{-5}	2.923×10^{-3}
rs6582394	12	40972456		3.247 (1.2×10^{-3})	2.504	0.199	0.126	2.977	0.014 (19)	1.093×10^{-4}	0.061
rs10220315	14	80197971	*CEP128*	2.947 (3.2×10^{-3})	4.818	0.999	2.934	3.472 (9)	5.334×10^{-3} (8)	1.576×10^{-5}	0.024
rs10136185	14	80210225	*CEP128*	2.954 (3.1×10^{-3})	4.579	0.996	2.85	3.419 (11)	5.943×10^{-3} (9)	1.912×10^{-5}	0.027
rs78304225	14	80276765	*CEP128*	2.853 (4.3×10^{-3})	4.249	0.977	2.341	3.258 (15)	0.011 (16)	3.453×10^{-5}	0.047
rs327443	14	80291769	*CEP128*	3.331 (8.7×10^{-4})	6.025	1	4.141 (5)	4.059 (6)	1.628×10^{-3} (7)	2.122×10^{-6}	7.152×10^{-3}
rs327465	14	80299793	*CEP128*	4.731 (2.2×10^{-6})	6.653	1	8.77 (4)	5.301 (4)	7.38×10^{-6} (4)	1.677×10^{-8}	2.329×10^{-5}
rs55957493	14	80539807	*CEP128*	5.634 (1.8×10^{-8})	9.916	1	13.509 (3)	6.803 (3)	2.346×10^{-8} (3)	8.75×10^{-12}	7.965×10^{-8}
rs17545310	14	80540892	*CEP128*	2.844 (4.5×10^{-3})	6.223	1	2.882 (6)	3.692 (8)	7.092×10^{-3} (10)	7.401×10^{-6}	0.031
rs2284734	14	80623486	*CEP128*	2.81 (5×10^{-3})	4.358	0.984	2.366	3.253 (16)	0.011 (15)	3.53×10^{-5}	0.047
rs2284735	14	80623539	*CEP128*	2.99 (2.8×10^{-3})	3.781	0.899	1.764	3.234 (17)	9.448×10^{-3} (13)	3.782×10^{-5}	0.041

Table 7: Top 20 SNPs differentiating Graves' disease and Hashimoto's thyroiditis (MHC removed), considered as subgroups of autoimmune thyroid disease, for each of four summary statistics. Positions are in NCBI build 37. Because of the density of the genotyping chip used and the large number of SNPs with evidence of differentiating the subgroups, with non-zero weights after applying the LDAK procedure are included in this table. The column 'LDAK' gives the weight attributed to the SNP by the LDAK procedure. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0'||Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0'|X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores		Values (rank)				Summary statistics	
SNP	Chr	Pos	Gene	$\lvert Z_d\rvert(p)$	$\lvert Z_a\rvert$	X_1	X_2	X_3	X_4	p-val (X_3)	FDR (X_4)
rs231790	2	204408819	*CTLA4*	3.465 (5.3×10^{-4})	5.584	0.979	2.672 (6)	3.825	0.047	2.179×10^{-6}	0.067
rs231797	2	204414352	*CTLA4*	3.466 (5.3×10^{-4})	5.567	0.979	2.677 (5)	3.824	0.047	2.204×10^{-6}	0.069
rs231804	2	204416891	*CTLA4*	3.046 (2.3×10^{-3})	6.22	0.925	1.764	3.531	0.031 (6)	1.07×10^{-5}	0.034
rs11571304	2	204417021	*CTLA4*	3.047 (2.3×10^{-3})	6.22	0.925	1.764	3.532	0.032 (7)	1.064×10^{-5}	0.036
rs3087243	2	204447164	*CTLA4*	3.355 (7.9×10^{-4})	6.606	0.94	2.163 (10)	3.86 (7)	0.036	1.812×10^{-6}	0.044
rs6748358	2	204465150	*CTLA4*	3.395 (6.9×10^{-4})	5.887	0.969	2.471 (9)	3.805	0.044	2.45×10^{-6}	0.062
rs7596727	2	204491827	*CTLA4*	3.578 (3.5×10^{-4})	5.065	0.987	2.958 (2)	3.845 (9)	0.044	1.946×10^{-6}	0.062
rs2352551	2	204503002	*CTLA4*	3.558 (3.7×10^{-4})	5.247	0.986	2.904 (3)	3.856 (8)	0.043	1.845×10^{-6}	0.059
rs3757247	6	91014184	*BACH2*	4.037 (5.4×10^{-5})	3.683	0.952	2.611 (7)	3.961 (4)	0.015 (2)	1.046×10^{-6}	0.017
rs11755527	6	91014952	*BACH2*	4.105 (4×10^{-5})	3.936	0.98	3.265 (1)	4.069 (1)	0.015 (1)	5.814×10^{-7}	0.016
rs619192	6	91025670	*BACH2*	3.135 (1.7×10^{-3})	4.791	0.971	2.124	3.423	0.03 (5)	1.913×10^{-5}	0.032
rs1847472	6	91029880	*BACH2*	3.465 (5.3×10^{-4})	4.389	0.975	2.579 (8)	3.638	0.02 (3)	5.968×10^{-6}	0.024
rs604912	6	91043041	*BACH2*	3.144 (1.7×10^{-3})	4.762	0.971	2.137	3.426	0.033 (9)	1.866×10^{-5}	0.039
rs17251453	12	91026211		4.26 (2×10^{-5})	2.593	0.674	0.831	3.844 (10)	0.071	1.951×10^{-6}	0.12
rs12426486	12	91052228		4.433 (9.3×10^{-6})	2.671	0.788	1.171	3.992 (2)	0.053	8.797×10^{-7}	0.079
rs7334298	13	41359622		4.442 (8.9×10^{-6})	2.566	0.752	1.013	3.965 (3)	0.066	1.017×10^{-6}	0.107
rs9525555	13	41408305		4.44 (9×10^{-6})	2.525	0.735	0.948	3.951 (5)	0.037	1.102×10^{-6}	0.048
rs9532960	13	41443676		4.377 (1.2×10^{-5})	2.418	0.657	0.727	3.871 (6)	0.033 (8)	1.696×10^{-6}	0.038
rs16967120	15	36707739	*RASGRP1*	3.088 (2×10^{-3})	4.359	0.949	1.896	3.317	0.035 (10)	3.361×10^{-5}	0.043
rs2839511	21	42721590	*UBASH3A*	3.533 (4.1×10^{-4})	4.47	0.981	2.76 (4)	3.709	0.029 (4)	4.072×10^{-6}	0.03

Table 8: Top ten SNPs differentiating TPOA positive and negative T1D (MHC removed), for each of four summary statistics. Positions are in NCBI build 36. Only SNPs with positive weights after applying the LDAK procedure (and therefore used in fitting the model) were considered here. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0' \mid |Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0' | X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores		Values (rank)				Summary statistics	
SNP	Chr	Pos	Gene	$\|Z_d\|(p)$	$\|Z_a\|$	X_1	X_2	X_3	X_4	p-val (X_3)	FDR (X_4)
rs231790	2	204408819	*CTLA4*	2.815 (4.9×10^{-3})	5.584	0.971	1.953 (7)	2.978	0.143	7.2×10^{-4}	0.146
rs231797	2	204414352	*CTLA4*	2.795 (5.2×10^{-3})	5.567	0.971	1.924 (8)	2.957	0.117	7.831×10^{-4}	0.117
rs11571293	2	204425958	*CTLA4*	2.834 (4.6×10^{-3})	5.995	0.966	2.044 (4)	3.013	0.271	6.162×10^{-4}	0.447
rs6748358	2	204465150	*CTLA4*	2.847 (4.4×10^{-3})	5.887	0.968	2.048 (3)	3.022	0.282	5.989×10^{-4}	0.494
rs7596727	2	204491827	*CTLA4*	2.944 (3.2×10^{-3})	5.065	0.978	1.977 (6)	3.078	0.204	4.648×10^{-4}	0.276
rs2352551	2	204503002	*CTLA4*	2.949 (3.2×10^{-3})	5.247	0.978	2.042 (5)	3.091	0.364	4.349×10^{-4}	≥ 0.5
rs1560418	3	159972335		4.541 (5.6×10^{-6})	1.21	5.957×10^{-3}	0.11	4.075 (10)	0.082	3.168×10^{-6}	0.275
rs1560417	3	159972476		4.543 (5.5×10^{-6})	1.213	5.988×10^{-3}	0.11	4.078 (9)	0.09	3.108×10^{-6}	0.088
rs511198	4	116234541		4.685 (2.8×10^{-6})	1.858	0.025	0.131	4.344 (3)	0.039 (4)	6.607×10^{-7}	0.137
rs506851	4	116234970		4.746 (2.1×10^{-6})	1.956	0.033	0.141	4.413 (1)	0.07 (10)	4.33×10^{-7}	0.227
rs503256	4	116244902		4.645 (3.4×10^{-6})	1.832	0.023	0.128	4.305 (5)	0.037 (3)	8.292×10^{-7}	0.131
rs473989	4	116246844		4.687 (2.8×10^{-6})	1.781	0.021	0.128	4.33 (4)	0.045 (7)	7.184×10^{-7}	0.166
rs505277	4	116248257		4.634 (3.6×10^{-6})	1.81	0.022	0.127	4.29 (6)	0.033 (1)	9.08×10^{-7}	0.113
rs1507935	4	116368809		4.586 (4.5×10^{-6})	1.514	0.011	0.116	4.188 (7)	0.055 (8)	1.644×10^{-6}	0.168
rs867036	4	116381578		4.581 (4.6×10^{-6})	1.515	0.011	0.115	4.184 (8)	0.056 (9)	1.709×10^{-6}	0.172
rs7694946	4	116413588		4.693 (2.7×10^{-6})	1.963	0.032	0.138	4.37 (2)	0.088	5.593×10^{-7}	0.096
rs706781	10	6126391	*IL2RA*	3.098 (1.9×10^{-3})	4.515	0.964	1.862 (10)	3.195	0.401	2.764×10^{-4}	≥ 0.5
rs907092	17	35175785	*IKZF3*	3.746 (1.8×10^{-4})	4.221	0.958	2.227 (1)	3.783	0.045 (6)	1.565×10^{-5}	0.163
rs11078927	17	35317931	*IKZF3*	3.774 (1.6×10^{-4})	4.168	0.951	2.175 (2)	3.805	0.045 (5)	1.392×10^{-5}	0.161
rs4795400	17	35320546	*IKZF3*	3.569 (3.6×10^{-4})	4.106	0.927	1.868 (9)	3.61	0.035 (2)	3.785×10^{-5}	0.121

Table 9: Top ten SNPs with differing effect sizes with age at diagnosis in T1D (MHC removed), for each of four summary statistics. Positions are in NCBI build 36. Only SNPs with positive weights after applying the LDAK procedure (and therefore used in fitting the model) were considered here. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0'||Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0'|X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.