A genetic test for differential causative pathology in disease subgroups

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

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

Abstract

Many common diseases show wide clinical or phenotypic variation. We present a statistical method for determining whether phenotypically defined subgroups of disease cases represent different genetic pathophysiologies, in which disease-associated variants have different effect sizes. Our method models the genomewide distributions of genetic association statistics with mixture Gaussians. We test for differential genetic bases without requiring explicit identification of disease-associated variants, maximizing power compared with standard variant-by-variant analyses. Where evidence for genetic subgroups is found, we present methods for subsequent identification of the contributing genetic variants.

We demonstrate the method on 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 basis with thyroid peroxidase antibody positivity.

Our method determines the existence of a genetic basis for disease heterogeneity, enabling genomics to inform the development of precision medicine.

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 within-disease heterogeneity. This clinical and phenotypic diversity may represent multiple causative pathophysiological processes. Because it is desirable that therapeutic approaches target disease-causative pathways, understanding this phenotypic complexity is valuable for further development of treatments, and the progression towards personalized medicine. Accounting for phenotypic substructures can also improve our ability to detect causative variants by refining phenotypes into subgroups in which causative variants have larger effect sizes [1].

Identification of patient subgroups characterized by different clinical features may aid directed therapy [2]. However, such subgroups may arise from geographic, socioeconomic or environmental influences, or represent different forms of the underlying disease (the latter being of the greatest interest). In this latter case, we expect that the genetic architecture of the disease will differ between subgroups. We present a statistical method for assessing whether this is the case. 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 diverse ethnic patient cohorts with diabetes, most of these differences will be unrelated to the pathogenesis of disease.

We proceed by joint consideration of allelic differences between cases and controls, and

allelic differences between case subgroups independent of controls. If the genetic basis of the disease is identical in both subgroups, these two sets of allelic differences vary independently. We test for the presence of a subset of SNPs with allelic differences between subgroups which additionally show evidence for association with the disease as a whole. The assumption that such a subset exists has been used previously to limit multiple hypothesis testing in single-SNP genetic discovery [3], but has not been formally tested.

Rather than attempting to analyses SNPs individually, a task for which GWAS are typically underpowered, we model allelic differences across all SNPs using multivariate normal models, and analyses their overall distribution. 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 with more confidence, in an approach similar to the empirical Bayes method.

We establish properties of our method by a range of simulations, and demonstrate its use by establishing differentiation between type 1 diabetes (T1D), type 2 diabetes (T2D), and rheumatoid arthritis (RA) datasets, considered as subgroups of generic disease phenotypes. We further demonstrate the method by analyzing differences between the autoimmune thyroid diseases (ATD) Graves' disease (GD) and Hashimoto's thyroiditis (HT). Finally, we apply the method to a T1D dataset partitioned by positivity of of disease-associated autoantibodies. Throughout, we use the term 'subgrouping' to mean a division of a case group into subgroups.

Results

Summary of proposed method

Model of SNP effect sizes We fit two bivariate Gaussian mixture models, corresponding to null and alternative hypotheses, to summary statistics derived from SNP data. We assume genetic data exists at the same SNPs for one group of control samples and two non-intersecting samples of case subgroups. For SNP i, we denote by $\mu_1(i)$, $\mu_2(i)$, $\mu_{12}(i)$ and $\mu_c(i)$ the population minor allele frequencies for the two case subgroups, the whole case group and the control group respectively, and $P_d(i)$, $P_a(i)$ GWAS p-values for comparisons of allelic frequency between case subgroups and between cases and controls, tests of the null hypotheses $\mu_1(i) = \mu_2(i)$ and $\mu_{12}(i) = \mu_c(i)$ respectively. We derive absolute Z scores $|Z_d(i)|$ and $|Z_a(i)|$ from these p-values and consider these as observations of random variables (Z_a, Z_d). Further details are given in the methods and supplementary material, section 1.

We consider each SNP to fall into one of three categories, each 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), the marginal standard deviation of Z_a is 1 and $cor(Z_d, Z_a) = 0$ for these SNPs. If

SNPs in category 3 are also associated with the disease as a whole (alternative hypothesis, H_1), the joint distribution of $(Z_d\, Z_a)'$ for such SNPs will have marginal variances greater than 1, and may have non-zero correlation. Our test is therefore focused on the form of the joint distribution of (Z_d, Z_a) in category 3 (see figure 1).

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 probability density function (*pdf*) of each observation (Z_d, Z_a) is given by

$$\begin{aligned} PDF_{Z_d,Z_a|\Theta}(d,a) =& \pi_0 N_{\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)}(d,a) && \text{(category 1)}\\ &+\pi_1 N_{\left(\begin{smallmatrix}1 & 0\\ 0 & \sigma_1^2\end{smallmatrix}\right)}(d,a) && \text{(category 2)}\\ &+\pi_2 \left(\frac{1}{2} N_{\left(\begin{smallmatrix}\tau^2 & \rho\\ \rho & \sigma_2^2\end{smallmatrix}\right)}(d,a) + \frac{1}{2} N_{\left(\begin{smallmatrix}\tau^2 & -\rho\\ -\rho & \sigma_2^2\end{smallmatrix}\right)}(d,a)\right) && \text{(category 3)} \end{aligned} \quad (1)$$

where $N_\Sigma(d,a)$ denotes the *pdf* of the bivariate normal distribution centered at $\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)$ with covariance matrix Σ at (d,a). Θ is the vector of values $(\pi_0, \pi_1, \tau, \sigma_1, \sigma_2, \rho)$. Under H_0, we have $\rho = 0$ and $\sigma_2 = 1$. The weights (π_0, π_1, π_2) represent the proportion of SNPs in each category, with $\Sigma\pi_i = 1$.

We use the product of values of the above *pdf* for a set of observed $(Z_d\, Z_a)'$ as an objective function to estimate the values of parameters. This is not a true likelihood as observations are not independent due to linkage disequilibrium (LD) between SNPs. We instead use the term 'pseudo-likelihood' (PL). We minimize the degree of LD between SNPs using the LDAK method [4], so the PL is close to a true likelihood.

Model fitting and significance testing We fit parameters π_0, π_1, π_2 $(= 1 - \pi_0 - \pi_1)$, σ_1, σ_2, τ and ρ under both H_1, in which all parameters may vary, and H_0, where (ρ, σ_2) are

set to $(0, 1)$. We compare the fit of the two models using the log-ratio of pseudo-likelihoods. We apply a correcting factor to avoid differences being driven by Z_a values alone, and a term to ensure the model is identifiable (maximum-PL estimators are unique) [5]. We term the resultant test statistic the PLR, or pseudo-log likelihood ratio (methods section, equations 5, 7, 8, 9).

For independent observations (SNPs in linkage equilibrium), the PLR corresponds to a log-likelihood ratio and has a mixture χ^2 distribution by an extension of Wilk's theorem [6]. For randomly-chosen subgroups of a case group (under H_0), the asymptotic distribution of PLR approximately satisfies

$$PLR \sim \begin{cases} \gamma\chi_1^2 & p = \kappa \\ \gamma\chi_2^2 & p = 1 - \kappa \end{cases} \tag{2}$$

for fixed values γ, κ which may be fitted by resampling random subgroups. Although the distribution of PLR can be established empirically without any assumption on its form, the use of distribution 2 enables estimation with fewer simulations (we suggest a minimum of 1000; see supplementary material, section 7), and extrapolation to high PLR values.

Single SNP prioritization If evidence for differential genetic basis can be established using the whole SNP set, as above, a natural next step is to search for the specific variants driving the difference. An effective test statistic for testing subgroup differentiation for single SNPs is the Bayesian conditional false discovery rate (cFDR) [7, 8] applied to Z_d scores 'conditioned' on Z_a scores. This enables the association threshold for Z_d scores to be adjusted based on observed Z_a scores. We propose three alternative test statistics, each with different advantages, in the methods section, and compare their performance in supplementary figure 6.

Power calculations and validation of method

We tested our method by application to a range of datasets, proceeding from simulation to real GWAS data. Firstly, we used simulated genotypes of case and control groups for a set of autosomal SNPs satisfying H_0 to establish the suitability of our proposed null distribution of our test statistic. We demonstrated the necessity of adjusting for the distribution of Z_a (supplementary material, section 3.3). After adjustment, the resultant PLR statistics showed good agreement with the expected mixture χ^2 distribution given in equation 2 (figure 2).

Although random subgroups provide a useful indicator of the null distribution of PLR, they do not exhaustively cover H_0, as there is no expected genetic difference between them (the 'true' value of τ is universally 1). To establish the behavior of the test when subgroups have systematic genetic differences that do not correspond to different disease pathologies, we analyzed a dataset of T1D cases with subgroups defined by geographical origin. Within the British Isles, there is clear genetic diversity associated with region [9] and we found Z_d scores for geographic subgroups were inflated compared to random subgroups (supplementary figure 3). No test statistics reached significance at a Bonferroni-corrected $p < 0.05$ threshold (max. corrected p value > 0.3, figure 2, supplementary figure 1), demonstrating that systematic differences between subgroups unrelated to case status do not produce false positive results under our method.

We then tested our method on previously published GWAS datasets [10] to ensure it could correctly distinguish between clinically distinct diseases. We applied it to each pair of three diseases (RA, T1D and T2D), considered in each case as subgroups of a general disease case group and correctly rejected the null hypothesis of genetic homogeneity (all $p < 10^{-6}$, table 1). Empirical distributions of PLR from random subgroups again agreed well with the proposed mixture χ^2 distribution (supplementary figures 9, 11). T1D and RA

have considerable overlap in genetic basis [8, 10, 11], as well as many chromosomal regions associated with only one of the two conditions. T1D and T2D have much less overlap [11] and T2D and RA less still. This was reflected in the fitted values (table 1). The full model for T1D/RA identified three distinctly shaped distributions; the fitted marginal variances of $(1, \sigma_1^2)$ of (Z_d, Z_a) in category 2 (separating case/control, but not separating subgroups) were markedly different from the corresponding marginal variances $(1, 1)$ for category 1 and (τ^2, σ_2^2) for category 3 (separating subgroups), indicating the presence of a subset of SNPs associated with both conditions to a similar extent. By contrast, in the full models for $T1D/T2D$ and $T2D/RA$ the marginal variances for category 2 were close to $(1, 1)$, the same as for category 1, suggesting that essentially all SNPs associated with the two diseases combined are in category 3; that is, have different effect sizes in each disease.

We estimated the power of the method under various scenarios (figure 3, supplementary figure 2). 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 only through the fitted model parameters (supplementary material, section 5). We therefore estimated the power of the test for varying numbers of SNPs in category 3 and for varying values of the parameters σ_2, τ, and ρ. The expected observed standard deviation of Z scores is a function of the numbers of cases and controls and the underlying distribution of effect sizes (methods, equation 3) As expected, power increases markedly 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_2, and correlation $(\rho/(\sigma_2\tau))$ as high values make it easier to distinguish SNPs in the third category from those in the first and second.

We explored the dependence of power on sample size in a disease dataset using the WTCCC cases for RA and T1D, as above. We repeatedly drew subgroups of cases of

varying size (originally c. 2000 cases for each), and tested the null hypothesis of genetic homogeneity. Although power was limited at reduced sample sizes, it remained consistently higher than the power to detect any single SNP which differentiated the two diseases at GWAS or Bonferroni-corrected significance (figure 3).

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 at such SNPs between subgroups and between cases/controls. With sparse genome-wide cover, such as that in the WTCCC study, and discounting genomic regions with very large effect sizes, such as the major histocompatability region (MHC) in autoimmune diseases, broadly 1500 cases per subgroup are necessary for 90% power. If SNPs with greater cover for the disease of interest are used (such as the ImmunoChip for autoimmune diseases) values of π_2, σ_2 and τ are correspondingly higher, and around 500-700 cases per subgroup may be sufficient.

Application to autoimmune thyroid disease and type 1 diabetes

ATD takes two major forms: GD (clinically hyperthyroid) and HT (clinically hypothyroid). The differential genetics of these two conditions have been investigated [12]. We were able to confidently detect evidence for differential genetic basis for GD and HT ($p < 9.9 \times 10^{-6}$ against H_0, table 1).

By contrast, 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]. 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. By restricting attention to known T1D-associated regions, a further 11 autoantibody positivity associa-

tions were found at $FDR \approx 16\%$ [3]. We considered four subgroupings defined by 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). We tested each of the four subgroupings both retaining and excluding the MHC region (supplementary table 1). When retaining the MHC region, we were able to confidently reject H_0 for subgroupings based on TPO-Ab, IA-2Ab and GAD-Ab (p values all $< 10^{-7}$). 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 SNPs determining PCA-Ab status are not in general associated with T1D. This pattern can be seen by in the plot of Z_a against Z_d (supplementary figure 4) in which SNPs with a high Z_d value do not have higher than expected Z_a values.

When the MHC region was removed, the subgrouping based on TPO-Ab remained significant ($p = 6.4 \times 10^{-4}$) but there was weaker evidence for GAD-Ab ($p = 0.031$) and IA2-Ab ($p = 0.034$). The fitted values of τ in both the full and null models for GAD-Ab were ≈ 1, indicating a lack of evidence for a category of T1D-associated SNPs additionally associated with GAD-Ab positivity outside the MHC. Collectively, this indicates that differential genetic basis for T1D with GAD-Ab and IA2-Ab positivity is driven almost entirely by the MHC region, and although PCA-Ab status may be genetically determined, the set of causative variants is independent of T1D causative pathways.

Finally, we applied the method to the same dataset with subgroups defined by age at diagnosis (supplementary table 2, supplementary figure 5). Rather than explicitly splitting cases into two groups, we considered age as an interval variable to compute Z_d. The hypothesis H_0 could be rejected confidently when retaining the MHC region, and less confidently after removing it (p values $< 10^{-7}$ and 0.023 respectively). Plotting signed Z_d

and Z_a scores for age (figure 5) indicated a strong negative correlation ($p < 1 \times 10^{-10}$).

Assessment of individual SNPs

To prioritise single SNPs responsible for the observed differences between T1D, T2D and RA, we used three test statistics: the cFDR, defined earlier, and two alternatives (see Methods) and list the top twenty SNPs for each test statistic in supplementary tables 3, 4, and 5.

Amongst these disease-differentiating SNPs were many in known disease-associated regions (obtained from `http://www.immunobase.org`). For the T1D/RA comparison, these included SNPs in or near the autoimmune-associated loci *PTPN22* (rs12045559), *IL7R* (rs1010599), *INS* (rs6578252), *IKZF4* (rs2292239), *SH2B3* (rs10774613) and *DEXI* (rs12924729). For the T1D/T2D comparison, identified SNPs were in or near the autoimmune-associated loci *PTPN22* (rs6679677), *SH2B3* (rs17696736), *PIK3C2B* (rs12061474), *IKZF4* (rs2292239), and *PTPN2* (rs2542151), and the T2D-associated loci *TCF7L2* (rs7901695) and *FTO* (rs2542151). For the T2D/RA comparison, identified SNPs were in or near autoimmune-associated loci *PTPN22* (rs6679677), *TNFAIP3* (rs11970411), *IL2RA* (rs2104286) and *IKZF3* (rs896136), and T2D-associated loci *TCF7L2* (rs7901695), *TSPAN8* (rs1495377), and *FTO* (rs8050136). Different peak SNPs for *FTO* and *PTPN22* are likely to arise due to noise in Z_d scores; and appropriate fine-mapping analyses will necessary to determine true peak SNPs in all putatively subgroup-differentiating regions.

We analyzed the differences between GD and HT in the same way, and identified SNPs near the known ATD-associated loci *PTPN22* (rs7554023), *CTLA4* (rs58716662), and *CEP128* (rs55957493) as likely to be contributing to the difference (see supplementary table 6). The SNPs rs34244025 and rs34775390 (both 9q34.3) are not known to be associated with ATD, but are in known loci for inflammatory bowel disease [13], and our data suggest

they may differentiate GD and HT (FDR 3×10^{-3}).

We then sought to find the SNPs outside of the MHC region with differential effect sizes with TPO-Ab positivity in T1D, the autoantibody for which we could confidently reject the null hypothesis. 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 7. Subgroup-differentiating SNPs included several near known T1D loci: *CTLA4* (rs7596727), *BACH2* (rs11755527), *RASGRP1* (rs16967120) and *UBASH3A* (rs2839511).

SNPs in the *CTLA4* locus both differentiated GD/HT and were associated with TPO-Ab positivity in T1D.Given that TPO-Ab is known to be associated to different degrees with GH and HT [14], this suggests a possible shared mechanism to TPO-Ab positivity in both diseases. We compared Z_d scores for the TPO-Ab positivity analysis with Z_a scores for ATD, at SNPs common to both datasets. The Z scores were significantly correlated, more pronounced when the MHC region was included ($p < 2 \times 10^{-16}$ with MHC included; $p = 7.5 \times 10^{-8}$ with MHC excluded; see supplementary figure 12). This suggests that the shared mechanism leading to TPO-Ab positivity is driven by multiple regions across the genome.

Finally, we analysed non-MHC SNPs with varying effect sizes with age at diagnosis in T1D (supplementary table 8). This weakly implicated SNPs in or near *CTLA4* (rs2352551), *IL2RA* (rs706781), and *IKZF3* (rs11078927). We demonstrated that effect sizes of T1D-associated SNPs differ generally with age at disease diagnosis, mostly in the opposite direction to association with T1D. 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 [15, 16, 17] and epidemiological data [18].

Discussion

We have developed a method to test whether the genetic basis of a disease differs between two subgroups of a case group, and corresponding methods to prioritise variants contributing to the difference. By way of positive control, we applied the method to diseases with established differences in aetiology. Importantly, the method did not produce undue false positive results when subgroups were defined by genetic differences independent of those involved in disease: that is, from different geographic regions of Great Britain.

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 [7, 8, 1, 19, 20]. We seek in this paper to use genomic data to infer whether such etiologically different disease subtypes exist. Our problem is related to the question of whether two different diseases share any genetic basis (ie [21]) but differs in that the implicit null hypothesis relates to genetic homogeneity between subgroups rather than genetic independence of separate diseases.

Our assumption that (Z_a, Z_d) follows a multivariate mixture Gaussian distribution is reasonable for complex phenotypes with a large number of associated variants [22] but our adjustment of the difference in log-pseudo-likelihood for the distribution of Z_a, essentially equivalent to performing the analysis conditional on observed Z_a, reduces our reliance on this assumption. Because we are comparing two models, the degree to which data conform to a mixture Gaussian distribution is less important than the difference in how well they fit two potential models. If subgroup prevalence is unequal between the study group and population, the expected distribution changes, but our method is still effective (supplementary material, section 2.2).

Further, 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 ethnicity in a phenotype which did not depend on ethnicity, and ethnicity varied between the case and control group, the set of SNPs associated with ethnic differences would also appear associated with the disease, which could lead to an inappropriately high type 1 error rate in our method. However, the same study design would also lead to identification of spurious association of ethnicity-associated SNPs with the phenotype in a conventional GWAS analysis. As for GWAS, this effect can be alleviated by accounting for the confounding trait (ethnicity) as a covariate when computing p-values.

We present three different ways of assessing significance of the PLR test statistic. The first is simply to empirically assess the PLR value against those of random case subgroups, without any assumption on the underlying distribution. The second is to compute p values by assuming that PLR follows a mixture-χ^2 distribution, which is generally justified at significance thresholds above 1×10^{-5}. Finally, we presented a method to derive an upper bound for the p-value based on the fitted values of parameters. The upper bound is very conservative for moderate values of PLR, but it may be more accurate than the mixture-χ^2 approximation at very high PLR values, given the behaviour of PLR in situations with higher LD cutoffs (supplementary material, figure 3, section 4). The mixture-χ^2 method is appropriate for most situations, with the third method only likely to be of use if many subgroupings are to be tested simultaneously, requiring substantial multiple-testing corrections.

Aetiologically and genetically heterogeneous subgroups within a case group correspond to substructures in the genotype matrix. Information about such substructures are lost in a standard GWAS, which only uses the column-sums (MAFs) of the matrix (linear-order information). Phenotype-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. For example, in our analysis of TPO-Ab positivity in T1D, the SNP rs3757247, in the known T1D-associated region containing *BACH2* displayed only $p = 2 \times 10^{-4}$ in a standard T1D vs control comparison, but $p = 1 \times 10^{-6}$ in our 'two-dimensional' analysis of Z_a (T1D vs control) and Z_d (TPO-Ab +ve vs TPO-Ab -ve). However, this can only be the case if the subgroups genuinely correspond to different variant effect sizes; for other subgroupings, a two-dimensional GWAS will only add noise.

Testing whether an individual SNP has different effect sizes in subgroups has lower power than testing association between cases and controls. Because of this, GWAS which are only able to detect a few SNPs at genome-wide significance will frequently be underpowered to detect SNPs with different effect sizes in subtypes. Our method cannot completely counter this problem, but it does enable a prioritization SNPs for follow-up, which can be corroborated with other data - in our case, known genome annotations (`http://www.immunobase.org`) - and can be used to justify candidate-gene association studies on known disease-associated regions.

Our application to T1D subgroups suggested differing effect sizes of T1D-associated SNPs dependent on TPO-Ab, IA2-Ab and GAD-Ab positivity, principally driven by the MHC region, and good evidence for differing non-MHC effect sizes with TPO-Ab positivity, as described previously [23]. This may be reflective of two underlying disease subtypes with differential association with TPO-Ab status. The loci *CTLA4*, *BACH2*, *RASGRP1* and *UBASH3A* indicated by our method agreed with those found previously by restricting attention to T1D-associated SNPs [3], but our analysis used only the available genotype data, without external information on confirmed T1D loci.

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 and well-applied method of searching for between-subgroup differences in SNPs with known disease associations [3] but we do so without restricting 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 the allelic spectrum of diseases and 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. Under these assumptions, we show in the supplementary material (sections 1,2) that Z_a and Z_d scores are approximately distributed according to model 1; that is,

$$\begin{aligned} PDF_{Z_d,Z_a|\Theta}(d,a) =& \pi_0 N_{\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)}(d,a) && \text{(category 1)}\\ &+\pi_1 N_{\left(\begin{smallmatrix}1 & 0\\ 0 & \sigma_1^2\end{smallmatrix}\right)}(d,a) && \text{(category 2)}\\ &+\pi_2 \left(\frac{1}{2} N_{\left(\begin{smallmatrix}\tau^2 & \rho\\ \rho & \sigma_2^2\end{smallmatrix}\right)}(d,a) + \frac{1}{2} N_{\left(\begin{smallmatrix}\tau^2 & -\rho\\ -\rho & \sigma_2^2\end{smallmatrix}\right)}(d,a)\right) && \text{(category 3)} \end{aligned}$$

where i indexes SNPs and Θ is the vector of values $(\pi_0, \pi_1, \pi_2, \tau, \sigma_1, \sigma_2, \rho)$. 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 use absolute Z scores, without considering effect direction.

We assume log-odds ratios for disease-associated SNPs are normally distributed with variance s^2, so 95% of odds-ratios lie in $[exp(-2s), exp(2s)]$. For sample sizes n_1, n_2 (either number of cases and controls, or number in each subgroup) the expected observed standard deviation of Z scores is

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

This is derived in the supplementary material, section 5

SNP weighting

A difficulty in modelling distributions of SNP effect sizes is that values of Z_a, Z_d may be correlated due to LD. One way to overcome this is to 'prune' SNPs by hierarchical clustering until only those with negligible correlation remain (supplementary material, figure 3). 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 to the difference between subgroups.

We opted to use the LDAK algorithm [4], which assigns weights to SNPs approximately corresponding to their 'unique' contribution. 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 approximately $1/n$, reducing the overall contribution to the likelihood to that of one SNP. In practice, many SNPs weights are 0; however, using the LDAK algorithm allows more SNPs to be retained and contribute to the model than would be retained in a pruning approach. If all SNPs are in linkage equilibrium, all weights are 1.

Definition and distribution of PLR statistics

For a set of observed independent pairs of Z scores $Z = (Z_a, Z_d)$ at a set of SNPs where SNP i has weight w_i, we define the joint unadjusted pseudo-likelihood $PL_{da}(Z|\Theta)$ as per equation 1 as

$$PL_{da}(Z|\Theta) = exp\left\{ C\,log(\pi_0\pi_1\pi_2) + \sum_{(Z_d^{(i)}, Z_a^{(i)}) \in Z} w_i\, log\left(PDF_{Z_d, Z_a|\Theta}(Z_d^{(i)}, Z_a^{(i)}) \right) \right\} \quad (4)$$

and

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

The term $C\,log(\pi_0\pi_1\pi_2)$ $(= C\,log(\pi_0\pi_1(1-\pi_0-\pi_1))$ in equation 4 ensures the maximum pseudo-likelihood estimators are identifiable, a requirement for Wilk's theorem. This is because situations may arise in which the underlying distribution of (Z_d, Z_a) is two-Gaussian rather than three-Gaussian; for instance, if all SNPs which are associated with the phenotype also differentiate subgroups, so there are no SNPs in category 2 (as seen in our comparisons between T1D/T2D and between T2D/RA, table 1). In this case, the distribution of (Z_d, Z_a) could be represented by either $\sigma_1 = 0$, with any values of π_0 and π_1 summing to $1 - \pi_2$, or by $\pi_1 = 0$ with any value of σ_1.

The additional term penalises small category sizes, and in the two-Gaussian case, ensures that the mean maximum pseudo-likelihood estimate of Θ satisfies $\pi_0 = \pi_1$, $\sigma_1 = 0$. It is equivalent to assuming a prior density of $(\pi_0, \pi_1, \pi_2, ...)$ proportional to $\{\pi_0\pi_1\pi_2\}^C$; or assuming the addition of $3C$ further SNPs known to be evenly split between the groups [5].

In order to minimise the bias toward equal category sizes, it is preferable for C to be small; we used $C = 1$.

If data observations are independent, PLR' is analagous to a log-likelihood ratio test statistic D, and under H_0 has the asymptotic distribution

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

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) [6].

Under H_1, the marginal distribution of Z_a has more degrees of freedom (four; π_0, π_1, σ_1, σ_2) than it does under H_0 (three, since $\sigma_2 \equiv 1|H_0$). Certain distributions of Z_a can lead to very high values of PLR' independent of the values of Z_d. This can lead to false positives, as the values Z_a reflect only case/control association and carry no information about case subgroups. This effect is particularly pronounced when the true value of τ is near 1 (supplementary material, section 3.3), in which case the true distribution of PLR' is not well-approximated by the asymptotic distribution.

To allow for this, we subtract a correcting factor $f(Z_a)$ from PLR' based on the pseudo-likelihood of Z_a alone. We define the marginal pseudo-likelihood of Z_a as

$$PL_a(Z_a|\Theta) = \prod_{Z_a^{(i)} \in Z_a} \left(\pi_0 N_{0,1}(Z_a^{(i)}) + \pi_1 N_{0,\sigma_1^2}(Z_a^{(i)}) + \pi_2 N_{0,\sigma_2^2}(Z_a^{(i)})\right)^{w_i} \tag{7}$$

Given $\widehat{\Theta_1}$, $\widehat{\Theta_0}$ as defined in equation 5, we then set

$$f(Z_a) = log\left(\frac{PL_a(Z_a|\widehat{\Theta_1})}{PL_a(Z_a|\widehat{\Theta_0})}\right) \tag{8}$$

and finally

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

For arbitrary $\Theta_1 \in H_1$, $\Theta_0 \in H_0$ we have

$$\begin{aligned} PLR &= \log\left(\frac{PL_{da}(Z|\theta_1)}{PL_{da}(Z|\theta_0)}\right) - log\left(\frac{PL_a(Z_a|\theta_1)}{PL_a(Z_d|\theta_0)}\right) \\ &= log\left(\frac{PL_{da}(Z_a, Z_d|\theta_1) \,/\, PL_a(Z_a|\theta_1)}{PL_{da}(Z_a, Z_d|\theta_0) \,/\, PL_a(Z_a|\theta_0)}\right) \qquad (10) \\ &\approx log\left(\frac{PL(Z_d|Z_a, \theta_1)}{PL(Z_d|Z_a, \theta_0)}\right) \qquad (11) \end{aligned}$$

so the application of the correcting factor $f(Z_a)$ can be considered to be computing and testing a likelihood ratio conditioning on the observed values of Z_a (assuming that $\widehat{\Theta_1} \approx arg\,max_{\Theta \in H_1}\, PL(Z_d|Z_a, \Theta)$, and similar for $\widehat{\Theta_0}$) . This enables us to test the hypothesis while minimising the influence of the distribution of Z_a, and reduces reliance on the assumption that the effect sizes Z_a have a mixture-Gaussian distribution.

Finally, we introduce two parameters γ and κ into the distribution 6 to give the distribution of PLR in equation 2. The value of ρ in model 1 is necessarily non-negative under H_1 and is 0 under H_0. The mixture χ^2 distribution arises because when the fitted value of ρ is 0 there is only a single degree of freedom difference between H_0 and H_1. Asymptotically, we expect the fitted value of ρ to be 0 in half of all cases if H_0 holds [6]. In practice, a more accurate estimate of the distribution can easily made by allowing the mixing proportions of the distribution to vary, hence the parameter κ in equation 2. The value γ arises from the weighting of the SNPs. The value corresponds to the effective number of independent SNPs.

Allowance for linkage disequilibrium

Adjusting using LDAK and applying the correcting factor $f(Z_a)$ was effective in enabling the distributions of PLR to be well-approximated by mixture-χ^2 distributions of the form 2 (supplementary plots 9, 10, 11). It is impossible to assert the suitability of the distribution for very high PLR scores, and from simulations with some residual LD between SNPs, we believe it is possible that the mixture χ^2 approximation may break down at such scores, leading to inappropriate underestimation of p-values (supplementary material, figure 3). However, we show that

$$\begin{aligned} PLR &\leq A + B\left(\tau^2 - log(\tau^2) - log(1 - \frac{\rho^2}{\sigma_2^2\tau^2})\right) \\ &\stackrel{\text{def}}{=} A + B\,F\left(\tau, \frac{\rho^2}{\sigma_2^2\tau^2}\right) \end{aligned} \tag{12}$$

for some constants A and B, which can be estimated by simulation. The distributions of τ and $\rho/(\tau^2\sigma^2)$ are tractable even when Z_a scores are correlated, and the convolution F can be numerically computed. Since $Pr(PLR > \alpha) \leq Pr(F\left(\tau, \frac{\rho^2}{\sigma_2^2\tau^2}\right) > \alpha)$ this formula can be used to give a bound on the p value for high PLR values. (supplementary material, section 4).

This reflects the effect on the PLR from each of the two additional degrees of freedom in the full model. The use of this inequality can be justified intuitively by considering that correlation between observations can inflate the PLR by two means: firstly, by inflating the value of the fitted parameters ρ and σ_2, and secondly by leading to the data fitting a given parameter set better than it would for independent data. Because LD leads to clusters of observations with similar (Z_a, Z_d) rather than points with large Z_a and Z_d, we expect it to lead to inflation principally by the latter effect, the magnitude of which corresponds to the value of parameter B. The inequality 12 assesses the significance of the PLR based only

on the fitted values ρ and τ, without assessing how well the data fits these values, and thus circumvents the problem of correlation.

E-M algorithm to estimate model parameters

We use an expectation-maximisation algorithm [24, 25] to fit parameters. Given an estimate of parameters $\Theta^i = (\pi_0^i, \pi_1^i, \tau^i, \sigma_1^i, \sigma_2^i, \rho^i)$ at iteration i of the algorithm, we iterate three main steps:

1. Define for SNP s

$$\zeta_g^{(s)} \leftarrow Pr(s \in \text{ category } g|\Theta_i) \tag{13}$$

2. For $g \in (1, 2, 3)$ set

$$\pi_g^{i+1} \leftarrow \overline{\zeta_g^{(s)}} \tag{14}$$

3. Set

$$\begin{aligned}(\tau^{i+1}, \sigma_1^{i+1}, \sigma_2^{i+1}, \rho^{i+1}) &\leftarrow \arg\max_{(\tau,\sigma_1,\sigma_2,\rho)} PL(Z_d, Z_a|\pi_0^{i+1}, \pi_1^{i+1}, \tau, \sigma_1, \sigma_2, \rho) \\ \Theta^{i+1} &\leftarrow (\pi_0^{i+1}, \pi_1^{i+1}, \tau^{i+1}, \sigma_1^{i+1}, \sigma_2^{i+1}, \rho^{i+1})\end{aligned} \tag{15}$$

Step 3 is complicated by the lack of closed form expression for the maximum pseudo-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 $|PLRZ_d, Z_a|\Theta^i) - PLR(Z_d, Z_a|\Theta^{i-1})| < \epsilon$; we use $\epsilon = 1 \times 10^{-5}$.

Depending on the initial estimate of parameters Θ^0, the algorithm can converge to local rather than global minima of the pseudo-likelihood. We overcome this by firstly computing the pseudo-likelihood of the data at 1000 values of Θ^0 throughout the parameter space,

retaining the top 100, and clustering these into five maximally-separated clusters according to an adapted Euclidean distance metric. The algorithm above is then started at the best (highest PL) point in each cluster

Although the algorithm can take several hundred iterations to converge, appropriate choice of Θ_0 can speed up the algorithm considerably; for simulations, we typically begin the model at previous maximum-PL estimates of parameters for earlier simulations. The algorithm and other processing functions are implemented in an R package available at `https://github.com/jamesliley/subtest`

Prioritization of single SNPs

We propose four summary statistics for testing the degree to which single SNPs have differential effect sizes in disease subgroups. 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 $\Theta_1 = \{\pi_0, \pi_1, \pi_2, \sigma_1, \sigma_2, \tau, \rho\}$:

$$\begin{aligned} X_1 &= Pr(\text{SNP} \in \text{category } 3|\Theta_1) \\ &= \frac{\frac{1}{2}\pi_2 \left(N_{\mathbf{0},\begin{pmatrix}\tau^2 & \rho \\ \rho & \sigma_2^2\end{pmatrix}}(z_a, z_d) + N_{\mathbf{0},\begin{pmatrix}\tau^2 & -\rho \\ -\rho & \sigma_2^2\end{pmatrix}}(z_a, z_d) \right)}{PDF_{\Theta_1}(z_a, z_d)} \end{aligned} \tag{16}$$

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 $\sigma_2 >> \sigma_1$, the test statistic X_1 will be high for high values of $|Z_a|$ even when $|Z_d|$ is low (supplementary figures 6). 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

$$X_2 = log\{PL(z_a, z_d|\Theta_1)\} - log\{PL(z_a, z_d|\Theta_0)\} \tag{17}$$

This has the advantage that high values of X_2 directly identify the SNPs contributing to a higher pseudo-log likelihood ratio (PLR). A disadvantage is the sensitivity to the behavior of the fitted parameters under H_0, which may be variable (see results section, table 1, supplementary figures 6), 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 prioprioritizes 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_1 (null

model) and σ_1, σ_2 (full model). Given fitted values τ', σ', we suggest the value

$$\alpha = \frac{log(\sigma')}{log(\tau') + log(\sigma')} \tag{18}$$

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 $N_{\mathbf{0},\left(\begin{smallmatrix}\tau'^2 & 0\\ 0 & 1\end{smallmatrix}\right)}$ or $N_{\mathbf{0},\left(\begin{smallmatrix}1 & 0\\ 0 & \sigma'^2\end{smallmatrix}\right)}$ 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 $\pi_{(0,1,2)}$.

The final test statistic, X_4, is a conditional false discovery rate (cFDR) [7, 8]. 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). Given a set of observed Z_a and Z_d values z_{ai}, z_{di}, $i \in$ SNPs with corresponding two-sided p values p_{ai}, p_{di}, the value of X_4 for SNP j is defined as

$$\begin{aligned} X_4 &= p_{dj} \frac{\left|\{i : p_{ai} \leq p_{aj} \wedge p_{di} \leq p_{dj}\}\right|}{\left|\{i : p_{di} \leq p_{dj}\}\right|} \\ &\approx Pr(H_0' \big| P_a \leq p_{aj}, P_d \leq p_{dj}) \end{aligned} \tag{19}$$

If only the hypothesis H_0' is to be tested, then we consider X_4 the best test statistic, as it takes account of the overall empirical distribution of Z_a, Z_d to assign lower values to X_4 with high $|Z_a|$, if the overall data indicate that this is appropriate [7, 8]. 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 [8].

Contour plots of the test statistics in various circumstances are shown in supplementary figures 6,7.

Description of GWAS datasets

ATD samples were genotyped on the ImmunoChip [26] 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 [27]. 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.

We also used published GWAS data from the Wellcome Trust Case Control Consortium [10] comprising 1994 cases of T1D, 1903 cases of RA, 1922 cases of T2D and 2953 common controls after quality control.

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, using WTCCC data [10], 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 [28]. We further required that call rate difference not be significant ($p \geq 1 \times 10^{-5}$) between any two groups, including 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 analyzing differences between autoimmune diseases, we stratified by geographic location for generating our (Z_a, Z_d) statistics; 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, 27]. We applied genomic control [29] 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 of $\approx 5Mb$ either side (chr. 6, 24.69 - 38.40 Mb in NCBI build 36; equivalent in other builds)

Type 1 error rates and power calculations

We simulated genotype data for 2000 controls and 2000 cases divided into two subgroups at 50,000 independent autosomal SNPs in Hardy-Weinberg equilibrium from the following null scenarios in equal proportions:

1. (a) (Z_d, Z_a) under H_0 with $\sigma_2 = 1$, $\rho = 0$.

   (b) (Z_d, Z_a) under H_0 with σ_2 allowed to vary, $\rho = 0$

2. (a) (Z_d, Z_a) under H_1 with $\rho = 0$.

   (b) (Z_d, Z_a) under H_1 with ρ allowed to vary

Full details of the distributions of the underlying parameters are given in the supplementary information, section 3.

For a more structured null hypotheses, we used T1D GWAS data [27] described above and, for each of twelve geographic UK regions and each possible pair of regions, we considered the subgroup of cases from that region or pair against all other cases (78 subgroupings in total). To maximize sample sizes, we considered T1D cases as 'controls' and split the control group into 'disease' subgroups.

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) and receives funding from the JDRF (5-SRA-2015-130-A-N) and the NIHR Cambridge Biomedical Research Centre. The research leading to these results has received funding from the European Union's 7th Framework ProgProgram7/2007-2013) under grant agreement no. 241447 (NAIMIT). JL is funded by the NIHR Cambridge Biomedical Research Centre and is on the Wellcome Trust PhD program Mathematical Genomics and Medicine at the University of Cambridge. CW is funded by the Wellcome Trust (089989). 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.

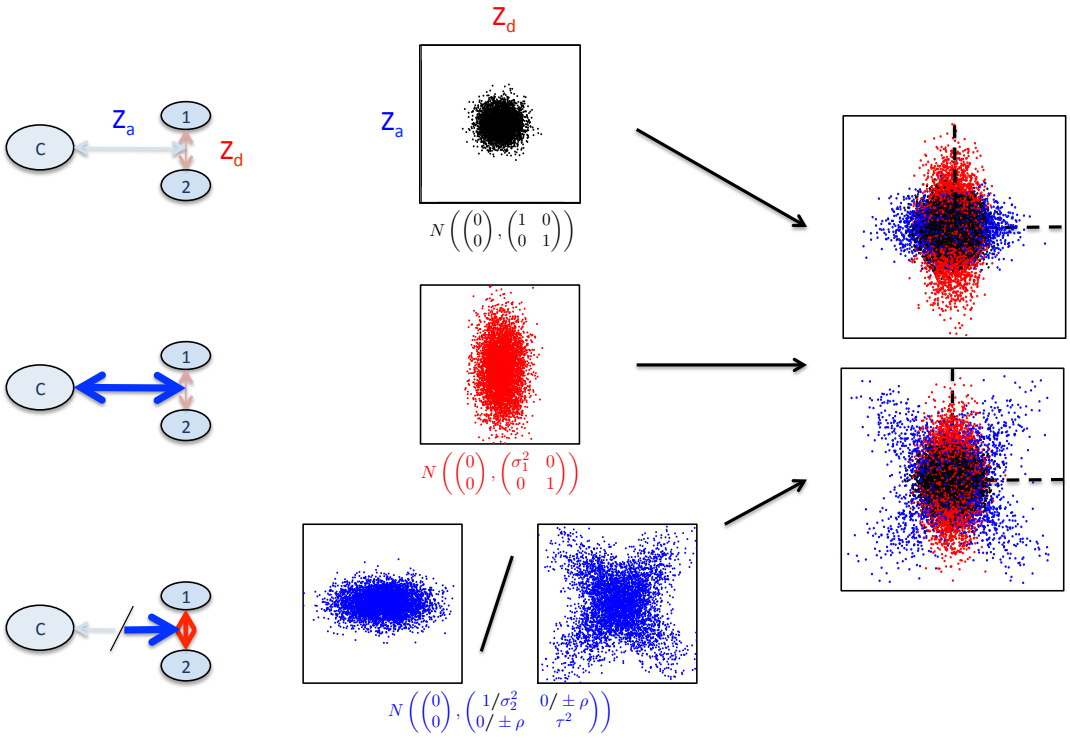

Figure 1: Overview of three-categories model. Z_d and Z_a are Z scores derived from GWAS p-values for allelic differences between case subgroups (1 vs 2, leftmost panels), and between cases and controls ($1+2$ vs C, leftmost panels) respectively (top left). Within each category of SNPs, the joint distribution of (Z_d, Z_a) has a different characteristic form. In the first group, Z scores have a unit normal distribution; in the second, the marginal variance of Z_a can vary. The distribution of SNPs in the third category 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. Because we only consider absolute Z scores (upper right quadrant), we model the distribution of SNPs in category 3 with two mirror-image Gaussians. The overall SNP distribution is then a mixture of Gaussians (rightmost panel). Visually, our test determines whether the observed overall Z_d, Z_a distribution more closely resembles the bottom rightmost panel than the top.

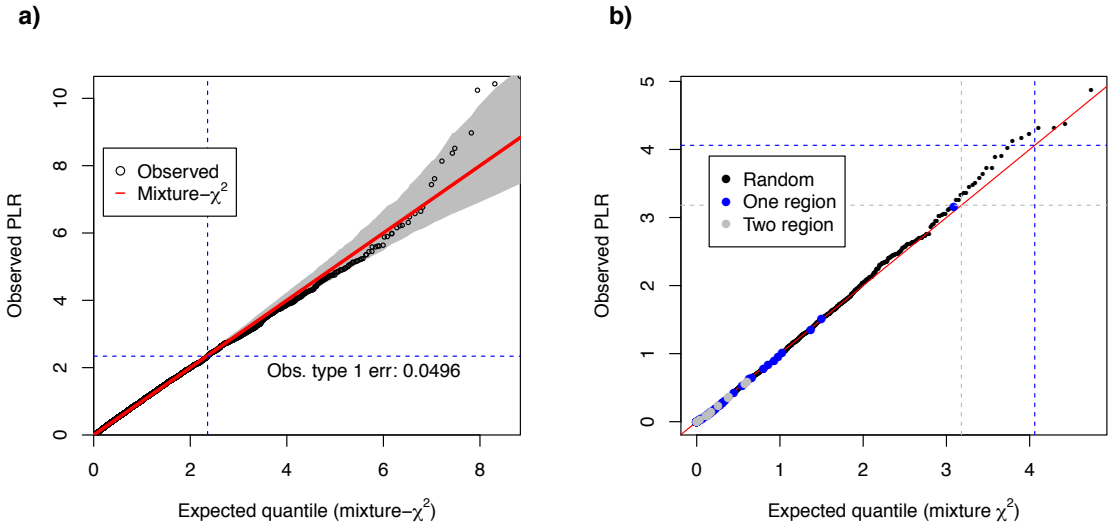

Figure 2: Type 1 error rate control. QQ plots comparing observed and proposed asymptotic distribution of PLR. X-axis shows quantiles of the best-fit scaled mixture-χ^2 distribution (eqn. 2) and Y-axis the observed PLR statistics. **Left panel; a):** simulated GWAS data, on 5×10^4 autosomal SNPs in linkage equilibrium; total of 1×10^4 simulation runs. The critical PLR value corresponding to $p = 0.05$ is marked with blue dotted lines. 95% confidence limits are shown in grey. **Right panel; b):** subgroups of T1D GWAS samples. Points are coloured according to whether they come from a random subgrouping (black), or geographical subgrouping based on one (grey) or two (blue) UK regions. Critical PLR values corresponding to Bonferroni-corrected $p = 0.05$ are marked with blue (two-region; 66 comparisons) or grey (one region; 12 comparisons). 4.5×10^3 simulations runs shown.

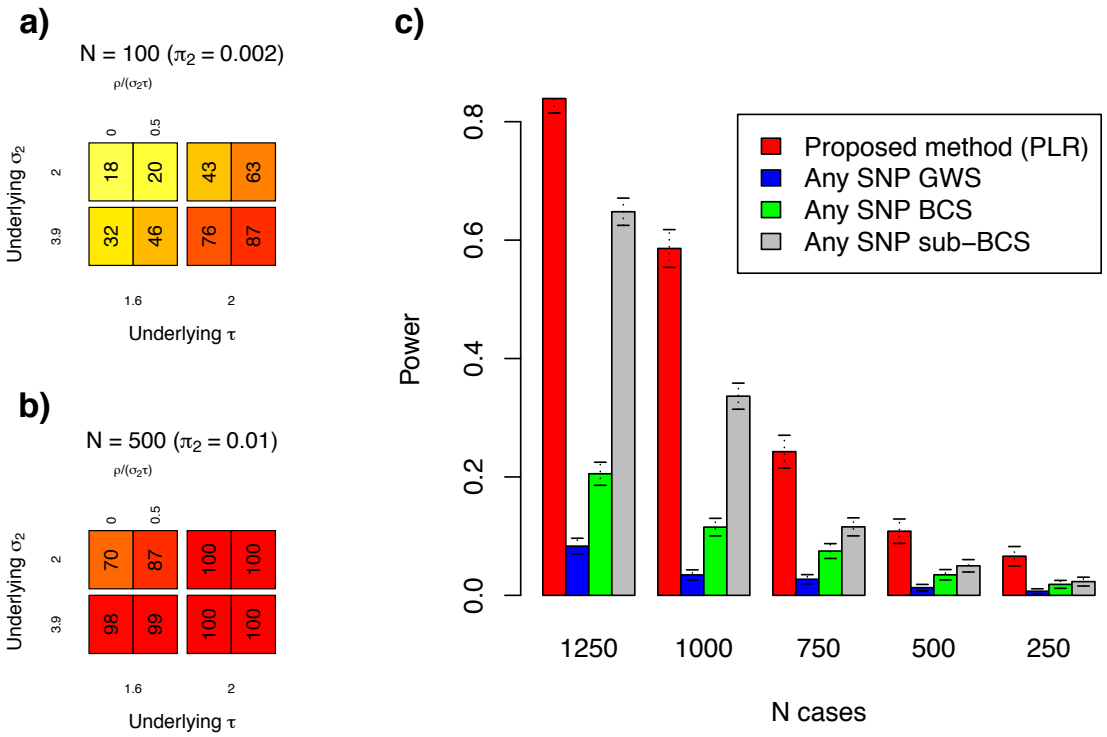

Figure 3: Power to reject the null hypothesis (genetic homogeneity between subgroups) depends on the number of SNPs in category 3 and the true underlying values of the model parameters σ_1, σ_2, τ, and ρ. Dependence on the number of case/control samples arises through the magnitudes of σ_2 and τ (supplementary material, section 5). **Left panel; a), b)**: estimates of power for various values of π_2, σ_2, τ, and ρ. Each simulation included 5×10^4 simulated autosomal SNPs in linkage equilibrium. N is the approximate number of SNPs in category 3 (corresponding π_2). $\rho/(\sigma_2\tau)$ is the correlation between Z_a and Z_d in category 3. More extensive plots of power are shown in supplementary figure 2. **Right panel; c):** power to detect differential genetic bases of T1D and RA subgroups of an autoimmune disease, down-sampling to varying numbers of cases (columns), for our proposed method; for testing whether the most significant Z_d reaches genome-wide significance ("Any SNP GWS", $p \leq 5 \times 10^{-8}$) or Bonferroni-corrected significance ("Any SNP BCS"; $p \leq 0.05/(\textit{total \# of SNPs})$); and for testing whether any SNP with genome-wide significant Z_a has Z_d reaching Bonferroni-corrected significance ("Any SNP sub-BCS"; $p \leq 0.05/(\textit{total \# of SNPs with } Z_a \textit{ reaching GWS})$). The height of each bar indicates the proportion of times the null hypothesis is rejected and error bars indicate 2 standard errors.

		π_0	π_1	π_2	σ_1	σ_2	τ	ρ	PLR	p-value	p-value bound
T1D/RA	H_1	0.997	5.69×10^{-4}	2.06×10^{-3}	2.76	1.39	1.74	1.815	15.0	1.9×10^{-9}	1.4×10^{-5}
	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	11.5	2.07×10^{-7}	1.3×10^{-4}
	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	10.8	4.93×10^{-7}	2.0×10^{-4}
	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	33.5	1.87×10^{-14}	9.9×10^{-6}
	H_0	0.493	0.079	0.428	1.68	-	1.03	-			

Table 1: Parameters of models fitted to T1D/RA, T1D/T2D, T2D/RA, and GD/HT. H_0 is the null hypothesis (under which $\sigma_2 = 1$, $\rho = 0$) that SNPs differentiating the subgroups are not associated with the overall phenotype; H_1 is the alternative (full model). Note the different forms the maximum pseudo-likelihood estimators of parameters under H_0 may take. P-values are computed by extrapolating the best-fit mixture-χ^2 distribution and upper bounds on the p-value were obtained from our alternative method. For the first three analyses, the PLR values were not markedly larger than those we were able to simulate, and the mixture χ^2 approximation was good at all simulated values (supplementary figures 9) so the upper bounds are likely to be very conservative. For the GD/HT comparison, a large extrapolation is necessary from simulated values (supplementary figure 10) so the upper bound for the p-value may be a better approximation than that from the mixture-χ^2. Because PLR values are dependent on SNP-specific LDAK weights, they are not readily comparable between different case groups, although they may be compared between different subgroupings of the same case group.

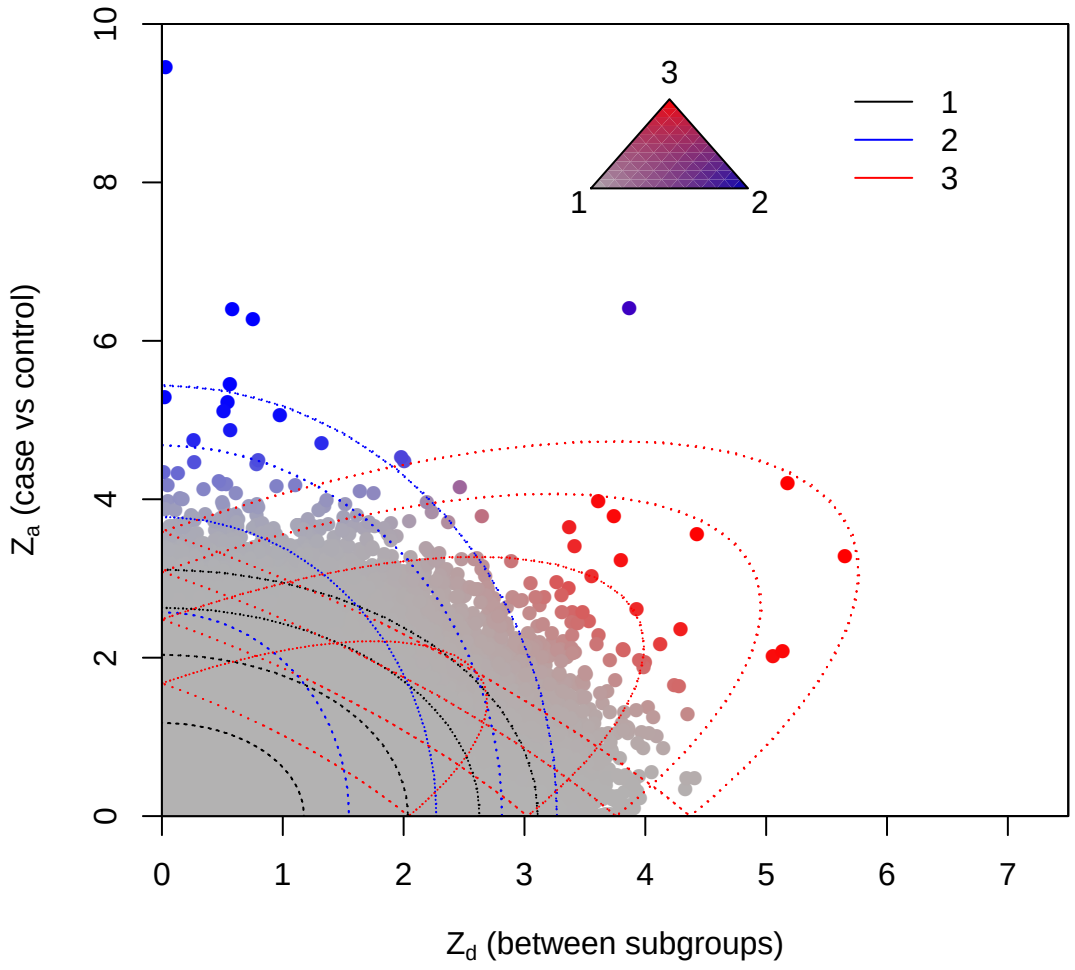

Figure 4: Observed absolute Z_a and Z_d for T1D/RA. Colourings correspond to posterior probability of category membership under full model (see triangle): grey - category 1, blue - category 2, red -category 3. Contours of the component Gaussians in the fitted full model (at are shown by dotted lines.

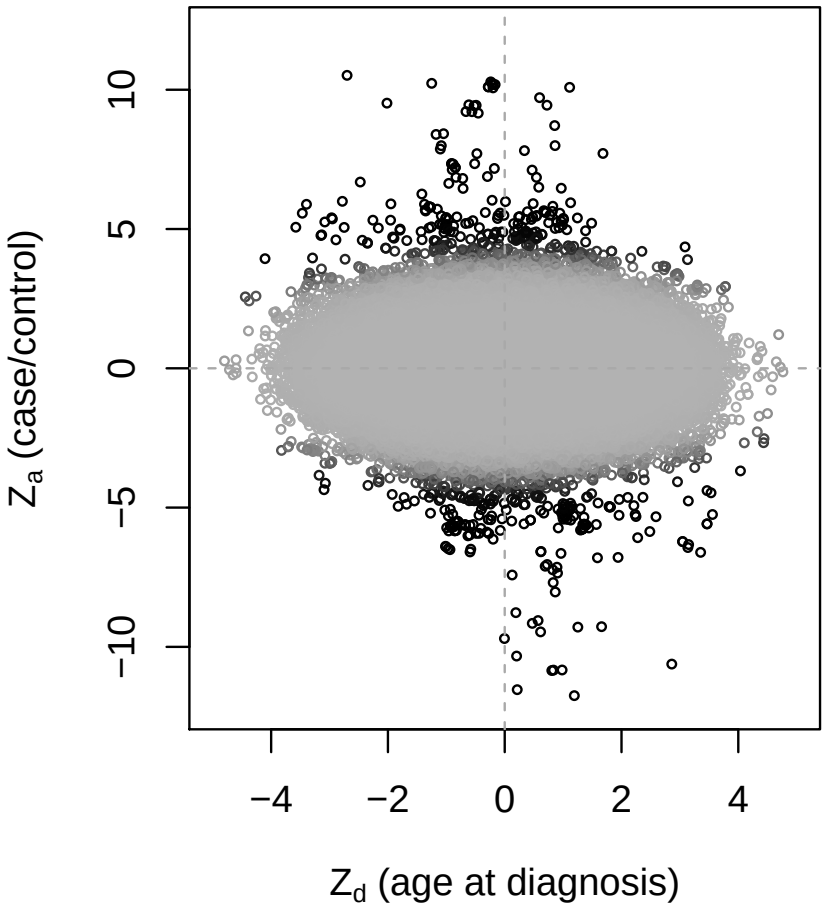

Figure 5: Z_a and Z_d scores for age at diagnosis in T1D, excluding MHC region. Colour corresponds to posterior probability of category 3 membership (X_1), with black representing a high probability. Data are negatively correlated (p value $< 2 \times 10^{-16}$ incl. MHC, $p = 1.3 \times 10^{-5}$ with MHC removed)

References

[1] 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.

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

[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] 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.

[7] 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).

[8] 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 .

[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] Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DPB, et al. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491: 119-124.

[14] Chardres T, Chapal N, Bresson D, Bes C, Giudicelli V, et al. (2002) The human anti-thyroid peroxidase autoantibody repertoire in Graves' and Hashimoto's autoimmune thyroid disease. Immunogenetics 54: 141-157.

[15] 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.

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

[17] 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.

[18] 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.

[19] Traylor M, Bevan S, Rothwell PM, Sudlow C, The Wellcome Trust Case Control Consortium 2, 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.

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

[21] Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. (2015) An atlas of genetic correlations across human diseases and traits. Nature Genetics 47.

[22] Loh 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.

[23] Howson JM, Dunger DB, Nutland S, Stevens H, Wicker LS, et al. (2007) A type 1 diabetes subgroup with a female bias is characterised by failure in tolerance to thyroid peroxidase at an early age and a strong association with the cytotoxic t-lymphocyte-associated antigen-4 gene. Diabetologia 50: 741–746.

[24] 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.

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

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

[27] 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.

[28] 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.

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