nature genetics

A method for identifying genetic heterogeneity within phenotypically defined disease subgroups

James Liley[1,2], John A Todd[1,3] & Chris Wallace[1,2,4]

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 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 maximizing power in comparison to 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 data sets, for which expected results are already known. We investigate subgroups of individuals with type 1 diabetes (T1D) defined by autoantibody positivity, establishing evidence for differential genetic architecture with positivity for thyroid-peroxidase-specific antibody, driven generally by variants in known T1D-associated genomic regions.

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 toward personalized medicine. Indeed, identification of patient subgroups characterized 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 exists a subset of disease-associated SNPs that 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 the subgroups of a disease group exhibit any genetic differences at all, as such differences may be entirely independent of the disease. For example, although there will be systematic genetic differences between cohorts of Asian and European patients with T1D, these differences will not generally relate to pathogenesis.

Rather than attempting to analyze SNPs individually for differences between subgroups, a task for which genome-wide association studies (GWAS) are typically underpowered, we modeled allelic differences across all SNPs using mixture multivariate normal models. This approach 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 case subgroups, we could then reassess test statistics to search for single-SNP effects.

RESULTS

Summary of proposed method

We jointly considered allelic differences between the combined case group and controls, as well as allelic differences between case subgroups independent of controls. Specifically, we established whether the data supported a hypothesis (H_1) stating that a subset of SNPs associated with case–control status had different underlying effect sizes (and, hence, underlying allele frequencies) in the 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 the subgroups or heritability may differ between the subgroups. These scenarios are discussed in the **Supplementary Note**.

Our overall protocol is to fit two bivariate Gaussian mixture models, which correspond 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 the controls, as well as allelic differences between the case subgroups independent of the controls (**Fig. 1**). Heterogeneity in cases can also be characterized by a quantitative trait rather than by explicit subgroups.

For a given SNP, we denote the population minor allele frequencies (MAFs) for each of the two case subgroups, the whole case group,

[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, UK. [3]Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [4]MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Cambridge, UK. Correspondence should be addressed to J.L. (ajl88@cam.ac.uk) or C.W. (cew54@cam.ac.uk).

Received 2 August; accepted 23 November; published online 26 December 2016; doi:10.1038/ng.3751

and the control group as μ_1 and μ_2, μ_{12}, and μ_c, respectively. GWAS *P* values for comparisons of the allelic frequency between the case subgroups and between cases and controls under the null hypotheses $\mu_1 = \mu_2$ and $\mu_{12} = \mu_c$ are denoted as P_d and P_a, respectively (terms are defined similarly for quantitative heterogeneity). We then derive absolute *Z* scores $|Z_d|$ and $|Z_a|$ from these *P* values (**Fig. 1**). We consider the values $|Z_d|$ and $|Z_a|$ as absolute values of observations of random variables (Z_d and Z_a, respectively), which are samples from a mixture of three bivariate Gaussians. Further details are given in the **Supplementary Note**.

We consider each SNP to fall into one of three categories, with each category corresponding to a different joint distribution of Z_d and Z_a: category 1 comprises SNPs that do not differentiate case subgroups and are not associated with the phenotype as a whole ($\mu_c = \mu_1 = \mu_2$); category 2 comprises SNPs that are associated with the phenotype as a whole but that are not differentially associated with the case subgroups ($\mu_c \neq \mu_{12}$; $\mu_1 = \mu_2 = \mu_{12}$); and category 3 comprises SNPs that have different population allele frequencies in the case subgroups and that 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), then we expect Z_d and Z_a to be independent and the variance of Z_a to be 1. If the SNPs in category 3 are also associated with the disease as a whole (alternative hypothesis, H_1), then the marginal variances for the joint distribution of Z_d and Z_a will both be greater than 1, and Z_a and Z_d may co-vary. Our test is therefore focused on the form of the joint distribution of Z_d and Z_a in category 3. Notably, 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 the risk of one case subgroup and another subset to alter the risk for the other case 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.

Among SNPs with the same frequency in the disease subgroups (categories 1 and 2), Z_a and Z_d are independent and the expected s.d. of Z_d is 1. We therefore model the overall joint distribution of Z_d and Z_a as a Gaussian mixture in which the probability density function (PDF) of each observation (Z_d,Z_a) is given by

$$\begin{aligned}\mathrm{PDF}_{Z_d,Z_a}(d,a) = {} & \pi_1 N_{\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}}(d,a) + \pi_2 N_{\begin{pmatrix}1 & 0\\ 0 & \sigma_2^2\end{pmatrix}}(d,a) \\ & + \pi_3 \left(\frac{1}{2} N_{\begin{pmatrix}\tau^2 & \rho\\ \rho & \sigma_3^2\end{pmatrix}}(d,a) + \frac{1}{2} N_{\begin{pmatrix}\tau^2 & -\rho\\ -\rho & \sigma_3^2\end{pmatrix}}(d,a) \right)\end{aligned} \quad (1)$$

where $N_\Sigma(d,a)$ denotes the density of the bivariate normal PDF centered at $\begin{pmatrix}0\\0\end{pmatrix}$ with covariance matrix Σ at (d,a). For H_0, we have $\rho = 0$ and $\sigma_3 = 1$. The values π_1, π_2, and π_3 represent the proportion of SNPs in each category, with $\Sigma\pi_i = 1$ (**Table 1**). Patterns of Z_d and 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 and Z_a values as an objective function ('pseudo-likelihood', PL) to estimate the values of the parameters. This is not a true likelihood, as observations are dependent owing to linkage disequilibrium (LD), although because we minimize the degree of LD between SNPs using the linkage disequilibrium–adjusted kinships (LDAK) method[4] the PL is similar to a true likelihood.

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 the PLs, giving an unadjusted pseudo-likelihood ratio (uPLR). We subtract a term dependent on only Z_a to minimize the influence of the Z_a score distribution and add the term $\log(\pi_1\pi_2\pi_3)$ to ensure that the model is identifiable[5]. We term the resultant test statistic the pseudo-likelihood ratio (PLR). The distribution of the PLR is minorized by a distribution of the form

$$\mathrm{PLR}\,|\,H_0 \sim \begin{cases} \gamma\chi_1^2 & \text{prob} = \kappa \\ \gamma\chi_2^2 & \text{prob} = (1-\kappa) \end{cases} \quad (2)$$

where χ_n^2 represents the χ^2 distribution with *n* degrees of freedom. 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 0.5).

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 the case 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 (defined in the Online Methods and **Supplementary Note**), which is well-behaved even when $\tau \approx 1$ and which majorizes the distribution of PLR.

A natural next step is to search for the specific variants that contribute 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_d and Z_a can deviate from H_0, and we also propose three other test statistics, each with different advantages, and compare their performance (**Supplementary Note**).

Power calculations, simulations, and validation of method

We tested our method by application to a range of data sets, using simulated and resampled GWAS data. First, to confirm appropriate control of type 1 error rates across H_0, we simulated genotypes for case and control groups under H_0 for a set of 5×10^5 autosomal SNPs in linkage equilibrium (**Supplementary Note**). Quantiles for 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% confidence interval (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 (**Fig. 2**). The empirical distribution of cPLR closely approximated the distribution of the asymptotic mixture χ^2 across all values of τ (**Supplementary Note**).

We then established the suitability of the test when SNPs are in LD and when genetic differences exist between subgroups that are independent of disease status overall. First, we used a data set of controls and autoimmune thyroid disease (ATD) cases and repeatedly chose case subgroups such that several SNPs had large allelic differences between the subgroups. We found good FDR control at all cutoffs (**Supplementary Note**), and the overall type 1 error rate at $\alpha = 0.05$ was 0.041 (95% CI = 0.034–0.050). Second, we analyzed a data set of T1D cases with subgroups that were defined by geographical origin.

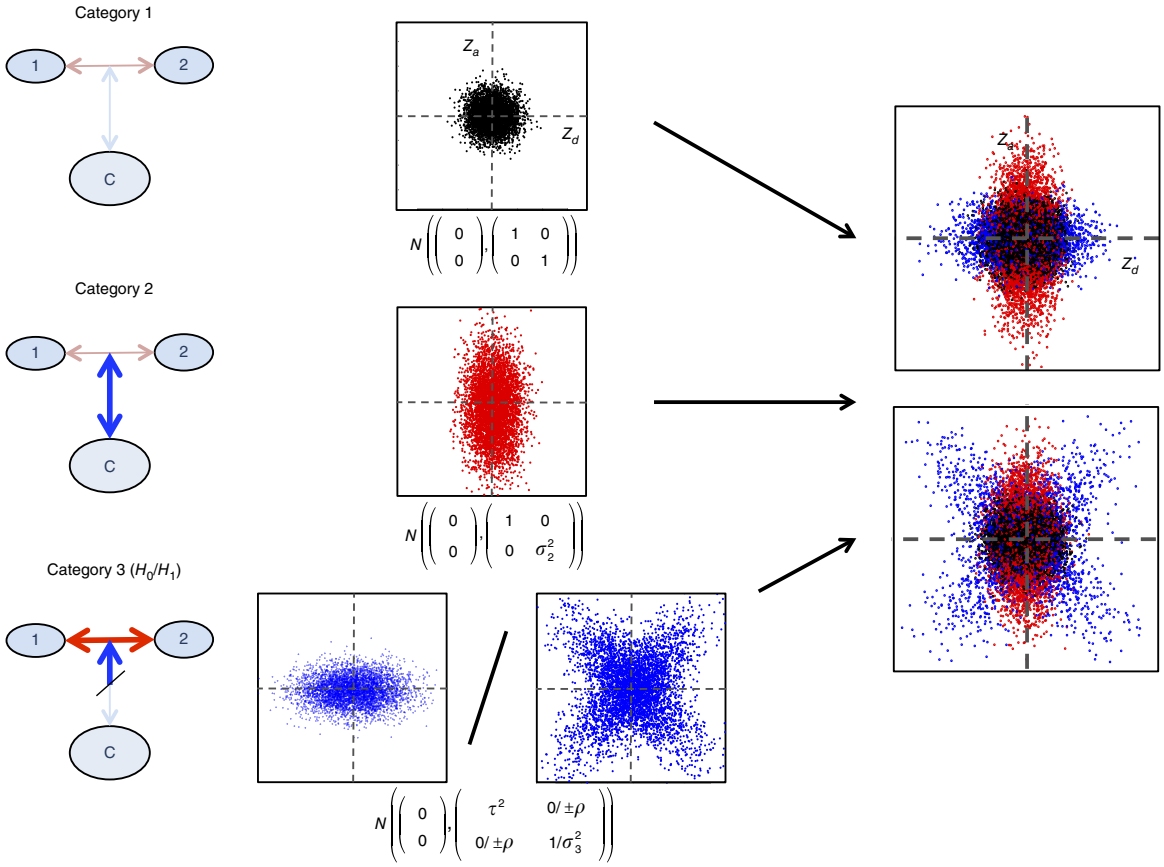

Figure 1 Overview of the three-category model. Left, Z_d and Z_a are Z scores derived from GWAS P values for allelic differences between case subgroups (1 versus 2) and between cases and controls (1 + 2 versus C), respectively. Middle, within each category of SNPs, the joint distribution of Z_d and Z_a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

Within the UK, there is clear genetic diversity associated with region[8]. As expected, Z_d scores for geographical subgroups showed inflation as compared to random subgroups (**Supplementary Fig. 1**). None of the derived test statistics reached significance at a Bonferroni-corrected threshold of $P < 0.05$ (minimum corrected P value > 0.8; **Supplementary Fig. 2**).

To examine the power of our method, we used published GWAS data from the Wellcome Trust Case Control Consortium[9] (WTCCC), which comprised 1,994 cases of T1D, 1,903 cases of rheumatoid arthritis (RA), 1,922 cases of type 2 diabetes (T2D), and 2,953 common controls. We established that our test could differentiate between any pair of diseases considered as subgroups of a general disease case group (all P values $<1 \times 10^{-8}$; **Table 2**).

T1D and RA have overlapping genetic bases[7,9,10], as well as having non-overlapping associated regions. T1D and T2D have less genetic overlap[10], and T2D and RA less still. This was reflected in the fitted values (**Fig. 3** and **Table 2**). The fitted values parametrizing category 2 in the full model for T1D–RA comparison (π_2,σ_2) were consistent with a subset of SNPs that were associated with case–control status (T1D + RA versus controls) but did not differentiate T1D and RA. By contrast, the parametrization of category 2 for the T1D–T2D and T2D–RA comparisons had marginal variance $\sigma_2 \sim 1$, 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 these comparisons entails the existence of a set of SNPs that are associated both with case–control status and subgroup status. The H_0 model does not allow such a set of SNPs, forcing the parametrization of Z_d and Z_a scores for these SNPs to be 'squashed' into a category shape permitted under H_0, with one marginal variance being 1, either in category 2 (as happens in the T2D–RA comparison because $\pi_2|H_0 \approx \pi_3|H_1$ and $\sigma_2|H_0 \approx \sigma_3|H_1$) or category 3 (as in the T1D–T2D comparison, where $\pi_3|H_0 \approx \pi_3|H_1$ and $\tau|H_0 \approx \tau|H_1$).

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**). 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 (**Fig. 4** and **Supplementary Fig. 3**). As expected, power increased with an increasing number of SNPs in category 3, reflecting the proportion of SNPs that differentiate the case subgroups and are associated with the phenotype as a whole. Power also increased with increasing τ, σ_3, and absolute correlation ($\rho/(\sigma_3\tau)$), as high values for these parameters enable better distinction of SNPs in categories 2 and 3.

We explored the dependence of power on sample size by subsampling the WTCCC data for RA and T1D (**Fig. 4**), and we compared

Table 1 Interpretation of parameter values in the fitted model

Parameter	Model	Interpretation
π_1	H_0/H_1	Proportion of SNPs not associated with case–control status (category 1)
π_2	H_0/H_1	Proportion of SNPs associated with case–control status but not with case subgroup status (category 2)
π_3	H_0/H_1	Proportion of SNPs associated with case subgroup status (category 3)
τ	H_0/H_1	s.d. of observed Z_d scores (effect sizes for case subgroup status) in category 3
σ_2	H_0/H_1	s.d. of observed Z_a scores (effect sizes for case–control status) in category 2
σ_3	H_1 only	s.d. 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

Parameters τ, σ_2, and σ_3 are dependent on sample sizes but can be converted to sample-size-independent forms (**Supplementary Note**).

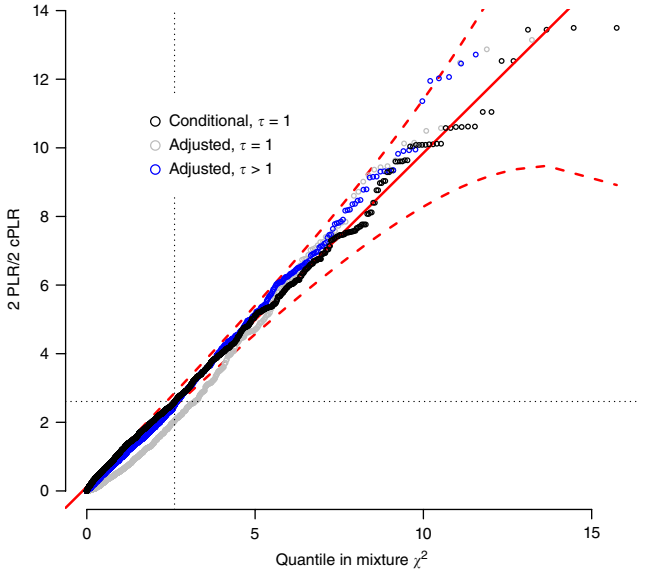

Figure 2 Quantile–quantile plot from simulations demonstrating type 1 error rate control of the PLR test. PLR values for test subgroups under H_0 with either $\tau = 1$ (random case subgroups) or $\tau > 1$ (genetic difference between case subgroups that are independent of the main phenotype) along with cPLR values for random case subgroups are plotted against the proposed asymptotic distribution under simulation ($(\chi_1^2 + \chi_2^2)/2$). The asymptotic distribution is represented by the solid red line, with the dashed red lines corresponding to the 99% confidence limits. The distribution of cPLR for random subgroups majorizes the distribution of PLR, indicating that the PLR-based test is conservative. Further details are given in the **Supplementary Note**.

the power of the PLR-based test with the power to find any single SNP that differentiated the two diseases in several ways (see legend for **Fig. 4**). 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 that 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) among SNPs with genome-wide significant Z_a score dropped dramatically, although the power of the PLR-based test was only slightly reduced. This illustrates the robustness of the PLR test to the 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, such as the expected total number of SNPs in the pruned data set with different population MAFs in case subgroups and the distribution of the odds ratios of such SNPs between case subgroups and between cases and controls. With sparse genome-wide coverage, such as that in the WTCCC study, >1,250 cases per subgroup are necessary for 90% power (discounting the major histocompatibility complex (MHC) region). If SNP arrays with greater coverage for the disease of interest are used (such as the Immunochip for autoimmune diseases), then the 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

ATD takes two major forms, Graves' disease (GD; hyperthyroidism) and Hashimoto's thyroiditis (HT; hypothyroidism). The differential genetic basis of these conditions has been investigated. Detection of individual variants with different effect sizes in GD and HT is limited by sample size (particularly for HT); however, the region around *TSHR* (encoding thyroid-stimulating hormone receptor) shows evidence of a differential effect[11]. T1D is relatively clinically homogenous, with no major recognized subtypes, although heterogeneity arises between patients with respect to the levels of disease-associated autoantibodies and disease course differs with age at diagnosis[3]. We analyzed both of these diseases.

For ATD, we were able to confidently detect evidence for differential genetic bases for GD and HT ($P = 2.2 \times 10^{-15}$). Fitted values are shown in **Table 2**. The distribution of cPLR statistics from random subgroups agreed well with the proposed mixture χ^2 distribution (**Supplementary Fig. 4**).

For T1D, we considered four subgroupings defined by plasma levels of the T1D-associated autoantibodies thyroid peroxidase antibody (TPO-Ab; n = 5,780), insulinoma-associated antigen 2 antibody

Table 2 Fitted parameter values for models of T1D–RA, T1D–T2D, T2D–RA, and GD–HT comparison

Comparison	Model	π_1	π_2	π_3	σ_2	σ_3	τ	ρ	*P* value
T1D–RA	H_1	0.997	5.69×10^{-4}	2.06×10^{-3}	2.76	1.39	1.74	1.82	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	0.910	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	–	

H_1 is the null hypothesis (for which $\sigma_3 = 1$, $\rho = 0$) that SNPs differentiating the case subgroups are not associated with overall phenotype; H_1 is the alternative (full model). *P* values for PLR tests are also shown.

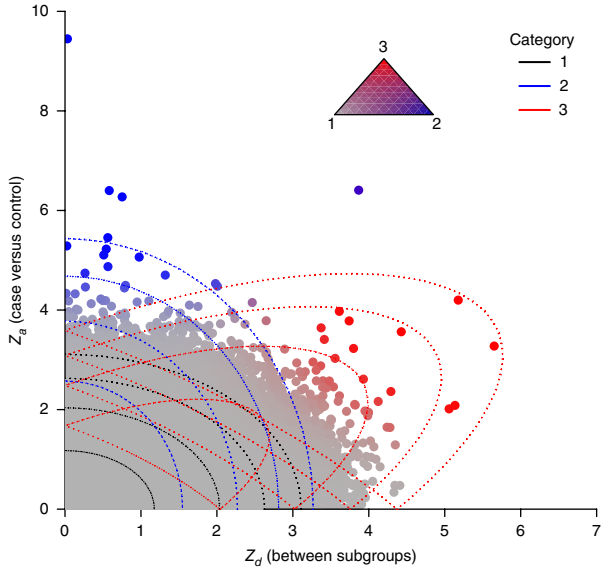

Figure 3 Observed absolute Z_a and Z_d scores for T1D–RA comparison. Colors correspond to the posterior probability of category membership under the full model (see triangle): gray, category 1; blue, category 2; red, category 3. The contours of the component Gaussians in the fitted full model are shown by the dotted lines.

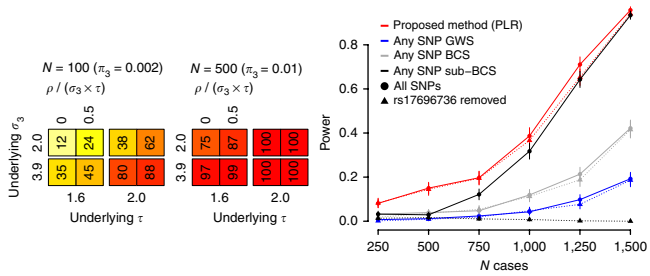

Figure 4 The power of PLR testing to reject H_0 (genetic homogeneity between case subgroups) depends on the number of SNPs in category 3 and the underlying values of model parameters σ_2, σ_3, τ, and ρ. Power is dependent on the number of case and control samples through the magnitudes of σ_3 and τ (**Supplementary Note**). Left, power estimates for various values of π_3, σ_3, τ, and ρ. 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. ρ ($\sigma_3 \times \tau$) is the absolute correlation between Z_a and Z_d in category 3 (also see **Supplementary Fig. 3**). Right, the power of PLR testing to detect differences in the genetic basis of the T1D and RA subgroups of a combined autoimmune data set, downsampling to varying numbers of cases (x axis). The power for PLR testing is compared with the power to find ≥1 SNP with a Z_d score reaching genome-wide significance (GWS; $P \leq 5 \times 10^{-8}$) or Bonferroni-corrected significance (BCS; $P \leq 0.05$/(total number of SNPs)) and the power to detect any SNP with a Z_a score reaching genome-wide significance and a Z_d score reaching Bonferroni-corrected significance (sub-BCS; $P \leq 0.05$/(total number of SNPs with a Z_a score reaching genome-wide significance)). Error bars show 95% confidence intervals. Circles and solid lines for each color show power for all SNPs; triangles and dotted lines show power for all SNPs except rs17696736. Power for the sub-BCS approach drops dramatically but power for PLR testing is not markedly affected when rs17696736 is excluded, indicating the relative robustness of PLR testing to single-SNP effects.

(IA2-Ab; n = 3,197), glutamate decarboxylase antibody (GAD-Ab; n = 3,208), and gastric parietal cell antibody (PCA-Ab; n = 2,240). A previous GWAS on autoantibody positivity in patients with T1D identified only two non-MHC loci at genome-wide significance: 1q23 (*FCRL3*; encoding Fc receptor like 3) associated with IA2-Ab and 9q34 (*ABO*; encoding ABO, α1-3-*N*-acetylgalactosaminyltransferase and α1-3-galactosyltransferase) associated with PCA-Ab[3].

We tested each of the subgroupings while retaining and excluding the MHC region. Fitted values for models with and without the MHC region are shown in **Supplementary Table 2**, and plots of Z_a and Z_d scores are shown in **Supplementary Figure 5**. By retaining the MHC region, we were able to confidently reject H_0 for subgroupings based on positivity for TPO-Ab, IA-2Ab, and GAD-Ab (all P values $<1.0 \times 10^{-20}$). Although there was evidence that SNPs in the data set were associated with PCA-Ab positivity ($\tau \approx 2.5$, 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 associated with T1D. This can be seen in the plot of Z_a against Z_d (**Supplementary Fig. 5**), in which SNPs with high Z_d values do not have higher-than-expected Z_a values.

With the MHC region removed, the subgrouping based on TPO-Ab positivity 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 $\alpha < 0.05$ of 0.006). The fitted values of τ in both the full and null models for GAD-Ab were ~1, indicating an absence of evidence for a category of non-MHC T1D-associated SNPs that are additionally associated with GAD-Ab positivity. Collectively, these findings indicate that the differential genetic basis for T1D with GAD-Ab versus IA2-Ab positivity is driven principally by the MHC region; 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 who were diagnosed at a younger age[12–15]. We investigated whether these differences are indicative of widespread differences in variant effect sizes dependent on age at diagnosis, possibly due to differential heritability (**Supplementary Note**). We applied the method to the T1D data set 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 are shown in **Supplementary Figure 6**. Hypothesis H_0 could be rejected confidently when retaining or removing the MHC region ($P < 1.0 \times 10^{-20}$ and P = 0.007, respectively). Signed Z_d and Z_a scores for age at diagnosis showed a visible negative correlation (r_g method 2; P = 0.002) among Z_d and Z_a scores for disease-associated SNPs (**Fig. 5**). This is consistent with higher genetic liability with lower age at diagnosis.

Assessment of individual SNPs

Many SNPs that discriminated subgroups were in known disease-associated regions (**Supplementary Tables 4–6**). In several cases, our method identified disease-associated SNPs that 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, SNP rs3811019, in the *PTPN22* (protein tyrosine phosphatase, non-receptor type 22) region, was identified as likely to discriminate between 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 contribute to the difference in genetic basis for these diseases (**Supplementary Table 7**). SNPs rs34244025 and rs34775390 are not known to be associated with ATD but are in known loci for inflammatory bowel disease and ankylosing

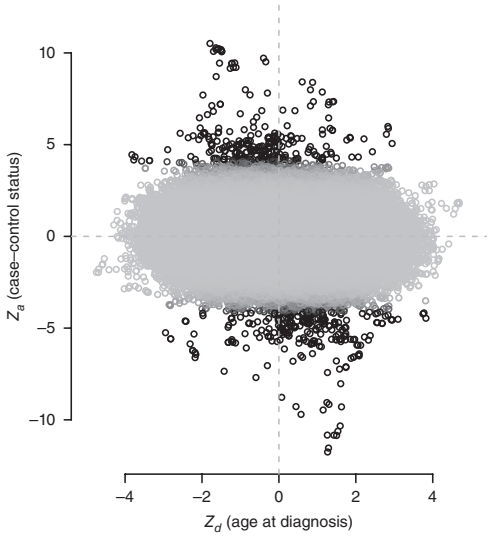

Figure 5 Z_a and Z_d scores for age at diagnosis in T1D, excluding the MHC region. Shading corresponds to the posterior probability of category 2 membership in the null model (as categories in the full model are assigned on the basis of correlation), with black representing high probability. Z_d and Z_a scores are negatively correlated ($P = 0.002$ with the MHC region removed) after accounting for LD using LDAK weights and weighting by the posterior probability of category 2 membership in the null model to prioritize SNPs further from the origin.

spondylitis, and our data suggest that they may differentiate between GD and HT (FDR = 0.003).

We searched for non-MHC SNPs with differential effect sizes with TPO-Ab 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 the focus to known T1D loci, enabling use of a larger data set 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-associated loci, such as *CTLA4* (rs7596727), *BACH2* (rs11755527), *RASGRP1* (rs16967120), and *UBASH3A* (rs2839511)[16]. These loci agreed with those found by Plagnol *et al.*[3], but our analysis used the available genotype data only, without external information on confirmed T1D-related loci. We were not able to replicate the same *P* values owing to reduced sample numbers.

Finally, we analyzed non-MHC SNPs with varying effect sizes with age at diagnosis in T1D (**Supplementary Table 9**). This analysis 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[2,6,7,17,18]. We sought here 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[19] but differs in that the implicit null hypothesis relates to genetic homogeneity between case subgroups rather than to 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 the **Supplementary Note**). In some cases, if subgroups are defined on the basis of the presence or absence of a known disease risk factor, then the heritability of the disease will differ between subgroups, with corresponding changes in variant effect sizes.

We preferentially use 'absolute covariance' ρ (**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, some variants may be deleterious only for subgroup 1 and others may be deleterious only for subgroup 2. A potential advantage of our symmetric model is the potential to generate Z_d scores from analysis of variance (ANOVA)-style tests for genetic homogeneity between three or more subgroups, in which case-reconstructed *Z* scores would be directionless.

Etiologically 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 uses only 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 comparison. 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.

Although 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. First, it is difficult to characterize subgroupings as 'better' or 'worse', and no one parameter can parametrize 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**). Second, even if subgroups could meaningfully be ranked, the search space of the potential subgroupings of a case group is prohibitively large (2^N for *N* cases), making exhaustive searches difficult.

We demonstrated that the effect sizes of T1D-causative SNPs differ with age at disease diagnosis. The strong negative correlation observed (**Fig. 5**) is consistent with an increased total genetic liability in samples with an earlier age of diagnosis, a finding supported by candidate gene studies[13–15] and epidemiological data[12]. Such a pattern arises naturally from a liability-threshold model, in which total liability depends additively on both genetic effects and environmental influences that accumulate with age (**Supplementary Note**).

Our method necessarily dichotomizes the multitude of mechanisms of heterogeneity, although there are many diverse forms (**Supplementary Table 1** and **Supplementary Note**). There is potential to further dissect the mechanisms of disease heterogeneity by incorporating estimations of genetic correlation[19] or assessing evidence for liability-threshold models[20]. 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 data set in comparison to a standard GWAS analysis and determines whether 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 it does not restrict focus to strong disease associations, enabling use of information

from disease-associated SNPs that do not reach significance. Our parametrization 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

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge the help of the Diabetes and Inflammation Laboratory Data Service for access and quality control procedures on the data sets used in this study. The JDRF/Wellcome Trust Diabetes and Inflammation Laboratory is in receipt of a Wellcome Trust Strategic Award (107212; J.A.T.) and receives funding from the JDRF (grant 5-SRA-2015-130-A-N; J.A.T.) and the NIHR Cambridge Biomedical Research Centre. The research leading to these results has received funding from the European Union's Seventh Framework Programme (grant FP7/2007-2013; J.A.T.) under grant agreement 241447 (NAIMIT). J.L. is funded by the NIHR Cambridge Biomedical Research Centre and is on the Wellcome Trust PhD program in Mathematical Genomics and Medicine at the University of Cambridge. C.W. is funded by the Wellcome Trust (grants 089989 and 107881) and the MRC (grant MC_UP_1302/5). The Cambridge Institute for Medical Research (CIMR) is in receipt of a Wellcome Trust Strategic Award (100140). We thank M. Simmonds, S. Gough, J. Franklyn, and O. Brand for sharing their AITD genetic association data set and all patients with AITD and control subjects for participating in this study. The AITD UK national collection was funded by the Wellcome Trust. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

J.L. conceived the statistical methods, wrote the software, performed the analyses, analyzed the data, and wrote the manuscript. J.A.T. analyzed the results and edited the manuscript. C.W. conceived the study, analyzed the data, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).
2. Morris, A.P. *et al.* A powerful approach to subphenotype analysis in population-based genetic association studies. *Genet. Epidemiol.* **34**, 335–343 (2010).
3. Plagnol, V. *et al.* Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet.* **7**, e1002216 (2011).
4. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
5. Chen, H., Chen, J. & Kalbfleisch, J.D. A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. Series B Stat. Methodol.* **63**, 19–29 (2001).
6. Andreassen, O.A. *et al.* Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet.* **9**, e1003455 (2013).
7. Liley, J. & Wallace, C. A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. *PLoS Genet.* **11**, e1004926 (2015).
8. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
9. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
10. Fortune, M.D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* **47**, 839–846 (2015).
11. Cooper, J.D. *et al.* Seven newly identified loci for autoimmune thyroid disease. *Hum. Mol. Genet.* **21**, 5202–5208 (2012).
12. Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M. & Tuomilehto, J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes* **52**, 1052–1055 (2003).
13. Howson, J.M.M., Walker, N.M., Smyth, D.J. & Todd, J.A. Analysis of 19 genes for association with type I diabetes in the Type I Diabetes Genetics Consortium families. *Genes Immun.* **10** (Suppl. 1), S74–S84 (2009).
14. Howson, J.M., Rosinger, S., Smyth, D.J., Boehm, B.O. & Todd, J.A. Genetic analysis of adult-onset autoimmune diabetes. *Diabetes* **60**, 2645–2653 (2011).
15. Howson, J.M. *et al.* Evidence of gene–gene interaction and age-at-diagnosis effects in type 1 diabetes. *Diabetes* **61**, 3012–3017 (2012).
16. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
17. Traylor, M. *et al.* Using phenotypic heterogeneity to increase the power of genome-wide association studies: application to age at onset of ischemic stroke subphenotypes. *Genet. Epidemiol.* **37**, 495–503 (2013).
18. Wen, Y. & Lu, Q. A multiclass likelihood ratio approach for genetic risk prediction allowing for phenotypic heterogeneity. *Genet. Epidemiol.* **37**, 715–725 (2013).
19. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
20. Chatterjee, N. & Carroll, R.J. Semiparametric maximum-likelihood estimation exploiting gene–environment independence in case-control studies. *Biometrika* **92**, 399–418 (2005).

ONLINE METHODS

Ethics statement. This paper reanalyzes previously published data sets. All patient data were handled in accordance with the policies and procedures of the University of Cambridge.

Joint distribution of variables Z_a and Z_d. We assume that SNPs may be divided into three categories, as described in the Results section (**Fig. 1**). Under these assumptions, Z_a and Z_d scores have the joint PDF given by equation (1). We define Θ as the vector of values π_1, π_2, π_3, τ, σ_2, σ_3, and ρ. Z scores Z_a and Z_d are reconstructed from GWAS P values for SNP associations. In practice, because our model is symmetric, we only require absolute Z scores, without considering effect direction.

For sample sizes n_1 and n_2, and 97.5% odds ratio quantile α, the expected observed s.d. of Z scores (i.e., σ_2, σ_3, and τ) is given by

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

(**Supplementary Note**).

Definition and distribution of PLR statistics. For a set of observed Z scores $Z = (Z_a, Z_d)$, we define the joint unadjusted pseudo-likelihood $\mathrm{PL}_{da}(Z \mid \Theta)$ as

$$\log\{\mathrm{PL}_{da}(Z \mid \Theta)\} = \sum_{Z_d^{(i)} \in Z_d, Z_a^{(i)} \in Z_a} w_i \log\left(\mathrm{PDF}_{Z_d, Z_a \mid \Theta}\left(Z_d^{(i)}, Z_a^{(i)}\right)\right) + C\log(\pi_1\pi_2\pi_3) \tag{4}$$

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

We now set

$$\begin{aligned} \widehat{\theta_1} &= \arg\max_{\Theta \in H_1} \mathrm{PL}_{da}(Z_d, Z_a \mid \Theta) \\ \widehat{\theta_0} &= \arg\max_{\Theta \in H_0} \mathrm{PL}_{da}(Z_d, Z_a \mid \Theta) \\ \mathrm{uPLR}(Z) &= \log\left(\frac{\mathrm{PL}_{da}(Z \mid \widehat{\theta_1})}{\mathrm{PL}_{da}(Z \mid \widehat{\theta_0})}\right) \end{aligned} \tag{5}$$

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

For data observations that are independent, uPLR is reduced to a likelihood ratio. Under H_0, the asymptotic distribution of uPLR is then

$$\mathrm{uPLR} \sim \begin{cases} \chi_1^2 & \text{prob} = 1/2 \\ \chi_2^2 & \text{prob} = 1/2 \end{cases} \tag{6}$$

according to Wilk's theorem extended to the case where the null value of a parameter lies on the boundary of H_1 (as $\rho = 0$ under H_0)[21].

The empirical distribution of uPLR may substantially majorize 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, and σ_3) than it does under the null model (two; π_2 and 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**), which is unwanted as the Z_a values reflect only case–control association and carry no information about case subgroups. If observed uPLRs from random subgroups (for which $\tau = 1$ by definition) are used to approximate the null uPLR distribution, then 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

$$\mathrm{PL}_a(Z_a \mid \Theta) = \prod_{Z_a^{(i)} \in Z_a} \left(\pi_1 N_{0,1}(Z_a^{(i)}) + \pi_2 N_{0,\sigma_2^2}(Z_a^{(i)}) + \pi_1 N_{0,\sigma_3^2}(Z_a^{(i)})\right) \tag{7}$$

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

$$f(Z_a) = \min\left(\log\frac{\mathrm{PL}_a(Z_a \mid \widehat{\theta_1})}{\mathrm{PL}_a(Z_a \mid \widehat{\theta_0})}, 0\right) \tag{8}$$

We now define the PLR as

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

The action of $f(Z_a)$ leads to the asymptotic distribution of PLR slightly minorizing the asymptotic mixture χ^2 distribution of uPLR to differential degrees dependent on the value of τ (**Supplementary Note**).

We define the similar test statistic cPLR as

$$\begin{aligned} \mathrm{cPL}(Z_d \mid Z_a, \theta) &= \log\left(\frac{\mathrm{PL}_{da}(Z_d, Z_a \mid \theta)}{\mathrm{PL}_a(Z_a \mid \theta)}\right) \\ \widehat{\theta_0^c} &= \arg\max_{\Theta \in H_0} \mathrm{cPL}(Z_d \mid Z_a, \Theta) \\ \widehat{\theta_1^c} &= \arg\max_{\Theta \in H_1} \mathrm{cPL}(Z_d \mid Z_a, \Theta) \\ \mathrm{cPLR}(Z) &= \log\left(\frac{\mathrm{cPL}\left(Z_d \mid Z_a, \widehat{\theta_1^c}\right)}{\mathrm{cPL}\left(Z_d \mid Z_a, \widehat{\theta_0^c}\right)}\right) \end{aligned} \tag{10}$$

noting that the expression

$$\frac{\mathrm{PL}_{da}(Z_d, Z_a \mid \theta)}{\mathrm{PL}_a(Z_a \mid \theta)}$$

can be considered as a likelihood conditioned on the observed values of Z_a. Now

$$\begin{aligned} \mathrm{PLR} &= \log\left(\frac{\mathrm{PL}_{da}\left(Z_d, Z_a \mid \widehat{\theta_1}\right)}{\mathrm{PL}_{da}\left(Z_d, Z_a \mid \widehat{\theta_0}\right)}\right) - \log\frac{\mathrm{PL}_a\left(Z_a \mid \widehat{\theta_1}\right)}{\mathrm{PL}_a\left(Z_a \mid \widehat{\theta_0}\right)} \\ &= \log\left(\frac{\mathrm{cPL}\left(Z_d \mid Z_a, \widehat{\theta_1}\right)}{\mathrm{cPL}\left(Z_d \mid Z_a, \widehat{\theta_0}\right)}\right) \end{aligned} \tag{11}$$

The empirical distribution of cPLR for random subgroups majorizes the empirical distribution of PLR (**Supplementary Note**). Furthermore, the approximation of the empirical distribution of cPLR by its asymptotic distribution is good across all values of τ; i.e., 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 (**Supplementary Note**).

Allowance for linkage disequilibrium. The asymptotic approximation of the pseudo-likelihood ratio distribution breaks down when values of Z_a and Z_d are correlated because of LD. One way to overcome this is to 'prune' SNPs by hierarchical 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 that contribute greatly to the difference between subgroups.

We opted to use the LDAK algorithm[4], which assigns weights to SNPs that approximately correspond to their 'unique' contribution. Using ρ_{ij} to denote the correlation between SNPs i and j, and using $d(i,j)$ to denote 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 $\sum \rho_{ij}^2$ for $i \neq j$ represents the replication of the signal of SNP i from that of all the other SNPs.

This approach has the advantage that, if n SNPs are in perfect LD and not in LD with any other SNPs, then 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 to contribute to the model than would be retained in a pruning approach.

A second advantage of LDAK is that it homogenizes 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[22], 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 homogenizing 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 (4)), 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 in equation (2) (**Supplementary Fig. 4**).

Expectation–maximization algorithm to estimate model parameters. We use an expectation–maximization algorithm[23,24] 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 $Z_d^{(s)}, Z_a^{(s)}$

$$\zeta_g^{(s)} = \Pr\left(s \in \text{category } g | \Theta_i\right)$$

$$\propto \begin{cases} \pi_1^i N_{\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}}\left(Z_d^{(s)}, Z_a^{(s)}\right) & (g=1) \\ \pi_2^i N_{\begin{pmatrix}1 & 0\\ 0 & (\sigma_2^i)^2\end{pmatrix}}\left(Z_d^{(s)}, Z_a^{(s)}\right) & (g=2) \\ \pi_3^i \left(\frac{1}{2} N_{\begin{pmatrix}(\tau^i)^2 & \rho^i\\ \rho^i & (\sigma_3^i)^2\end{pmatrix}}\left(Z_d^{(s)}, Z_a^{(s)}\right) + \frac{1}{2} N_{\begin{pmatrix}(\tau^i)^2 & -\rho^i\\ -\rho^i & (\sigma_3^i)^2\end{pmatrix}}\left(Z_d^{(s)}, Z_a^{(s)}\right) \right) & (g=3) \end{cases} \quad (13)$$

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

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

3. Set

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

Step 2 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 $|\text{PLR}(Z_d,Z_a \mid \Theta_i) - \text{PLR}(Z_d,Z_a \mid \Theta_{i-1})| < \epsilon$; 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 1,000 points throughout the parameter space, retaining the top 100 and dividing these into five 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 optimization 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

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

where w_i is the weight for SNP i (see the **Supplementary Note** for rationale). The conditional pseudo-likelihood was maximized 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[25], 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, then our method can still be used with adaptation (**Supplementary Note**).

Our test is robust to confounders arising from differential sampling to the same extent as that for conventional GWAS. For example, if subgroups were defined on the basis of population structure and population structure also varied between the case and control groups, then SNPs that differed by ancestry would also appear to be associated with the disease, leading to a loss of control of the 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**).

Prioritization 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 (i.e., Z_d), which gives greater priority to SNPs potentially associated with case–control status (i.e., 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, which states that the population MAFs of the SNP in both case subgroups are equal (that is, that the SNP does not differentiate subgroups), but it 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 independent of direction.

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

$$X_4 = \frac{|\{i : p_{ai} \le p_{aj} \cap p_{di} \le p_{dj}\}|}{|\{i : p_{di} \le p_{dj}\}|} \approx \Pr(H'_0 \mid P_a \le p_{aj}, P_d \le p_{dj}) \quad (17)$$

The value gives the FDR for SNPs whose P values fall in the region $[0,p_{dj}] \times [0,p_{aj}]$; this can be converted into an FDR among all SNPs for whom X_4 passes some threshold[7].

We discuss three other single-SNP test statistics in the **Supplementary Note**, 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 data sets are shown in **Supplementary Figures 7–9**.

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)[19,26] may be adaptable to the subgrouping question by estimating r_g across a set of SNPs between traits from case–control studies of interest, with

doi:10.1038/ng.3751

the potential advantage of characterizing 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 explore these methods in the **Supplementary Note**. 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 (for example, hair color 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 (**Fig. 1**). Indeed, parameter ρ estimates an analog of r_g to account 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 and σ_3) and covariance parameters (ρ) in testing a subgrouping of interest.

Description of GWAS data sets. ATD samples were genotyped on the Immunochip[22], a custom genotyping array targeting putative autoimmunity-associated regions. Data were collected for GWAS-like analyses of dense SNP data[11]. The data set comprised 2,282 cases of Graves' disease, 451 cases of Hashimoto's thyroiditis and 9,365 controls.

T1D samples, which were gathered for a GWAS on T1D[16], were genotyped on either the Illumina 550K or Affymetrix 500K platform. We imputed between platforms in the same way as that for the original GWAS. The data set comprised genotypes from 5,908 T1D cases and 8,825 controls, of which all had measured TPO-Ab values, 3,197 had measured IA2-Ab values, 3,208 had measured GAD-Ab values and 2,240 had measured PCA-Ab values. 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 who were positive for different autoantibodies (for instance, TPO-Ab positive versus IA2-Ab positive) because many individuals were positive for both.

To generate summary statistics corresponding to geographical subgroups, we considered the subgroup of cases from each of 12 regions and each pair of regions against all other cases (78 subgroupings in total). To maximize 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 comparisons between data from T1D, T2D and RA cases, which we reused 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 were correspondingly loose[9]. 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 ≥ 1%, CR ≥ 95% if MAF ≥ 5%, CR ≥ 99% if MAF < 5%) but used a stricter control on Hardy–Weinberg equilibrium, requiring $P \geq 1 \times 10^{-5}$ for deviation from Hardy–Weinberg equilibrium in controls. We also required that deviance from Hardy–Weinberg equilibrium in cases satisfied $P \geq 1.91 \times 10^{-7}$, corresponding to $|z| \leq 5$. The looser threshold for Hardy–Weinberg equilibrium in cases was chosen because deviance from Hardy–Weinberg equilibrium can arise because of true SNP effects[27]. We also required that the call rate difference not be significant ($P \geq 1 \times 10^{-5}$) between any two groups, including case–case and case–control differences. Geographical data were collected by the WTCCC and consisted of assignment of samples to one of 12 geographical regions (Scotland, northern Ridings, northwestern Ridings, east Ridings, west Ridings, North Midlands, Midlands, Wales, eastern England, southern England, southeastern England and London[9]). In analyzing differences between autoimmune diseases, we stratified by geographical location; when assessing subgroups on the basis of geographical location, we did not.

For the ATD and T1D data, we used quality control procedures identical to those used in the original papers[11,16]. We applied genomic control[28] to computation of Z_a and Z_d scores, except for in our analysis of ATD (following the original authors[11]) and our geographical analyses (as discussed above). In all analyses, except where otherwise indicated, we removed the MHC region with a wide margin (~5 Mb on either side).

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

Data availability. This paper reanalyzes previously published data sets. WTCCC data access for T1D, T2D, RA and controls[9] is described at https://www.wtccc.org.uk/info/access_to_data_samples.html. ATD data are available upon request from the authors of the original study[11]. T1D genetic data from ref. 16 are available at the database of Genotypes and Phenotypes (dbGaP) under accession phs000180.v3.p2, which we combined with autoantibody data available from the study authors[3].

21. Self, S.G. & Liang, K.Y. Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**, 605–610 (1987).
22. Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
23. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **39**, 1–38 (1977).
24. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning. Springer Series in Statistics* (Springer, 2001).
25. Loh, P.R. *et al*. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
26. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism–derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
27. Anderson, C.A. *et al*. Data quality control in genetic case–control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
28. Devlin, B., Roeder, K. & Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.* **60**, 155–166 (2001).