ARTICLE

A Fast Association Test for Identifying Pathogenic Variants Involved in Rare Diseases

Daniel Greene,[1,2,3,*] NIHR BioResource, Sylvia Richardson,[3] and Ernest Turro[1,2,3]

We present a rapid and powerful inference procedure for identifying loci associated with rare hereditary disorders using Bayesian model comparison. Under a baseline model, disease risk is fixed across all individuals in a study. Under an association model, disease risk depends on a latent bipartition of rare variants into pathogenic and non-pathogenic variants, the number of pathogenic alleles that each individual carries, and the mode of inheritance. A parameter indicating presence of an association and the parameters representing the pathogenicity of each variant and the mode of inheritance can be inferred in a Bayesian framework. Variant-specific prior information derived from allele frequency databases, consequence prediction algorithms, or genomic datasets can be integrated into the inference. Association models can be fitted to different subsets of variants in a locus and compared using a model selection procedure. This procedure can improve inference if only a particular class of variants confers disease risk and can suggest particular disease etiologies related to that class. We show that our method, called BeviMed, is more powerful and informative than existing rare variant association methods in the context of dominant and recessive disorders. The high computational efficiency of our algorithm makes it feasible to test for associations in the large non-coding fraction of the genome. We have applied BeviMed to whole-genome sequencing data from 6,586 individuals with diverse rare diseases. We show that it can identify multiple loci involved in rare diseases, while correctly inferring the modes of inheritance, the likely pathogenic variants, and the variant classes responsible.

Introduction

Hundreds of thousands of individuals with rare disorders are undergoing whole-genome sequencing in an effort to identify novel disease etiologies, increase our understanding of biological processes, and improve clinical care.[1] Thanks to the affordability of DNA sequencing, population association study designs for diseases affecting fewer than 1 in 2,000 people are now possible. However, the statistical association methods required to identify relevant loci need to fulfil several criteria in order to be well-powered, particularly when the number of cases with a particular disease is small. First, they need to allow some sharing of information across variants because rare diseases are often genetically heterogeneous. Second, they need to account for the presence of pathogenic rare variants that act upon disease risk in a dominant or a recessive manner alongside benign rare variants that do not affect disease risk. Third, they must be capable of integrating prior information into the inference regarding the plausibility of a locus being implicated in a disease and variant-level co-data on pathogenicity. Such co-data can be derived from population allele frequency databases, consequence predictions, conservation-based predictions, or genomic datasets, for example. Lastly, methods need to have efficient implementations that enable fast application across a large number of regions in the genome.

Frequentist association tests for rare variants include the Burden test and the sequence kernel association test (SKAT).[2] The Burden test regresses the phenotype on a genetic score obtained by summing allele counts across all rare variants in a locus. The cohort allelic sums test (CAST)[3] uses a genetic score that is equal to 1 if an individual harbors at least 1 (or 2) variants under a dominant (or recessive) inheritance model, and 0 otherwise, but is statistically equivalent to the Burden test in other respects. SKAT specifies a random effect for each variant and performs a score test under the null hypothesis that the variance of the random effects is zero. The variance-covariance structure of the random effects under the alternative hypothesis is determined by a kernel function, which would typically be a weighted genetic correlation across the variants in the locus. SKAT can incorporate nuisance covariates, accounts for linkage disequilibrium between variants under consideration, and is well-powered for traits whereby many different variants in a locus with varying effect sizes and allele frequencies contribute to the phenotype.

For scientific follow-up, it is important to infer which variants are likely to be pathogenic, conditional on an association being present in a given locus. The backward elimination[4] procedure removes individual variants iteratively from the predictors as long as this increases a test statistic of association (either Burden or SKAT). The adaptive combination of p values (ADA)[5] algorithm ranks variants by p value obtained using Fisher's exact test and selects a threshold on p value that maximizes an aggregate test statistic. As these algorithms prune variants in a stepwise fashion, they do not explore the full space of possible combinations of pathogenic variants. It is also important that inference can be performed sufficiently quickly to enable applications across tens of thousands of regions, with tens to hundreds of variants in each one. The methods above, however, rely on permutations to obtain empirical p values, rendering them computationally expensive.

[1]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge CB2 0XY, UK; [2]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK; [3]Medical Research Council Biostatistics Unit, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK
*Correspondence: dg333@cam.ac.uk
http://dx.doi.org/10.1016/j.ajhg.2017.05.015.

In principle, Bayesian inference lends itself well to rare variant association analysis because it provides a coherent framework for sharing information across variants and provides a natural way of incorporating prior information on variant pathogenicity. The variational Bayes discrete mixture method (vbdm),[6] the Bayesian risk index,[7] and the Bayesian rare variant detector (BRVD)[8] all model a mixture of pathogenic and non-pathogenic variants in a locus, but they employ additive models of disease risk or severity more suited to complex rather than rare diseases caused by dominant or recessive inheritance of one or two pathogenic alleles.

Here we present a Bayesian model in which disease risk depends on the genotypes at rare variants in a locus, a latent mode of inheritance, and a latent partition of variants into pathogenic and non-pathogenic subsets. Different modes of inheritance are modeled by conditioning the probability of case status on the number of pathogenic alleles and the ploidy for each individual at the variants. Thus, disease risk due to compound heterozygosity or X-linked inheritance is explicitly accommodated. Prior knowledge concerning variant pathogenicity can be incorporated in the form of shifts in the log odds of pathogenicity relative to a global mean. By placing a vague prior distribution on the scale of these shifts, the usefulness or otherwise of these co-data are accounted for flexibly to maximize power.

For a given set of variants, inference is performed by comparing the model described above with a baseline model in which disease risk is independent of the genotypes. The mode of inheritance and the pathogenicity of each variant, conditional on an association, can be inferred through the posterior distributions of parameters in the model. Particular classes of variants in a locus may be the only ones that confer disease risk. For example, only variants in the 5′ UTR region or only high-impact coding variants may be involved. Our method can compare models fitted to different classes of variants in order to infer which ones are responsible for disease. Typically the inference process would be repeated over many sets of variants selected from different loci throughout the genome. The procedures are implemented in an efficient R package called BeviMed, which stands for Bayesian evaluation of variant involvement in Mendelian disease.

Material and Methods

Model Specification

Let y be a binary vector of length N indicating whether individual i is a case ($y_i = 1$) or a control ($y_i = 0$) subject with respect to a particular disease. Suppose k rare variants are under consideration (typically in a particular genomic region) and the genotype for individual i at variant j is coded in the ith row and jth column of the genotype matrix G. A genotype of 0 or 2 denotes homozygosity for the common or minor allele, respectively, and a genotype of 1 denotes heterozygosity. Under a baseline model, labeled $\gamma = 0$, y is independent of G and all individuals have a probability of being a case τ_0. Under the association model, labeled $\gamma = 1$, individuals either have or do not have a pathogenic configuration of alleles and have probabilities of being a case subject π and τ, respectively. Whether or not an individual has a pathogenic configuration of alleles depends on a function f of the genotypes G_i of that individual, a latent binary vector z indicating which of the k variants are pathogenic, a value s_i equal to the ploidy of the individual at the variant sites, and a variable m representing the mode of inheritance governing the disease etiology though the k variants:

$$\gamma = 0 : \mathbb{P}(y_i = 1) = \tau_0,$$

$$\gamma = 1 : \mathbb{P}(y_i = 1) = \begin{cases} \tau & \text{if } f(G_{i\cdot}, z, s_i, m) = 0, \\ \pi & \text{if } f(G_{i\cdot}, z, s_i, m) = 1. \end{cases} \quad \text{(Equation 1)}$$

The function f can represent a dominant inheritance model or a recessive inheritance model that accounts for sex-dependent differences in ploidy on the X chromosome (i.e., X-linked recessive inheritance), depending on variable $m \in \{m_{\text{dom}}, m_{\text{rec}}\}$:

$$f(G_{i\cdot}, z, s_i, m_{\text{dom}}) = 1_{\sum_j G_{ij} z_j \geq 1},$$

$$f(G_{i\cdot}, z, s_i, m_{\text{rec}}) = 1_{\sum_j G_{ij} z_j \geq s_i}.$$

Thus, the interpretation of z depends on the mode of inheritance. In order to have a pathogenic allele configuration, individual i requires at least one allele at a variant for which $z_j = 1$ under a dominant model, but s_i alleles under a recessive model. If genotypes are phased, then a requirement that the s_i pathogenic alleles are on different haplotypes can be imposed. Recent relatedness is a potential confounder because it is correlated with both case/control status and genotype and, therefore, only unrelated individuals should be included in the model.

We place beta priors on all three parameters representing risk of disease:

$$\tau_0 \sim \text{Beta}(\alpha_0, \beta_0),$$

$$\tau \sim \text{Beta}(\alpha_\tau, \beta_\tau),$$

$$\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi).$$

The mean risk of disease for individuals without a pathogenic combination of alleles in the variants under consideration is uncertain under both models, and thus we place uniform priors on τ_0 and τ by default. However, as pathogenic combinations of alleles typically confer a high disease risk, we suggest setting the hyperparameters for π to $\alpha_\pi = 6$ and $\beta_\pi = 1$ (i.e., with a mean of 6/7). However, the prior mean could be adapted, for example, to reflect the consistency with which the disease manifests within families.

We adopt a logistic regression framework for the prior probability that the variants are pathogenic. The logit of the prior probabilities are shrunk toward a common mean, ω. If prior information that discriminates between the likely pathogenicity of variants is available, it can be incorporated in the form of a covariate c with regression coefficient ϕ in the regression equation:

$$z_j \sim \text{Bernoulli}\left(p_j\right),$$

$$\text{logit}\, p_j = \omega + \phi c_j.$$

One would typically place a Gaussian prior on the intercept ω but, for computational purposes, we prefer to use a logit-beta prior with hyperparameters α_ω and β_ω (see Appendix A). The prior mean of ω should reflect the expected proportion of variants that are pathogenic, conditional on an association, and may depend on the filtering procedures used to select the variants to include in the model. By default, $p(\omega)$ reflects a prior expectation that 20% of variants are pathogenic and a prior probability of only 0.01 that the proportion of pathogenic variants exceeds 0.54. This prior is well suited to missense variants but a distribution with a higher mean should be specified if most variants are expected to be pathogenic. This would be the case if the variants under consideration are all protein truncating and thought to be functionally equivalent to each other. To ensure that ω can be interpreted as the global mean log odds of pathogenicity, the c are required to sum to zero. Thus, any user-supplied weights, $\tilde{c}_j$, are centered such that $c_j = \tilde{c}_j - (1/k)\sum_l \tilde{c}_l$. We place a log-normal prior on the regression coefficient ϕ to force the effects of the c_j to be the same as their signs. The prior mean of ϕ is set to 1 so that the c_j are interpretable as prior shifts in the log odds of pathogenicity relative to the mean. A prior variance on ϕ of 0.35 ensures that the effect of the co-data can be diminished if the co-data are not informative and increased if they improve the model fit.

Finally, the prior probability on the mode of inheritance parameter m and the model indicator parameter γ need to be specified. By default, we set the prior probabilities for each mode of inheritance given an association to be the same, i.e., $\mathbb{P}(m = m_{\text{dom}} \mid \gamma = 1) = 0.5$, and we assume that there is only a 1% chance a priori of an association, i.e., $\mathbb{P}(\gamma = 1) = 0.01$. However, for a particular set of variants, the choice of values for these parameters could be based on the scientific literature or reference variant databases, for example.

Inference

The principal quantity of interest is the posterior probability of the model indicator γ, which can be derived from a Bayes factor comparing the two models and $\mathbb{P}(\gamma)$. The Bayes factor has two components, the evidence under $\gamma = 0$ and the evidence under $\gamma = 1$. A closed-form expression exists for the evidence under either model and it can be computed rapidly under $\gamma = 0$, irrespective of y. However, the expression for the evidence under $\gamma = 1$ contains a sum over every possible value of z, of which there are 2^k, and k is usually large enough to render this sum computationally intractable.

To tackle this problem, we reviewed various methods for estimating the evidence of a model[9] and chose the method of power posteriors,[10] which enables the evidence to be estimated by Markov chain Monte Carlo (MCMC) sampling. In this method, the MCMC is tempered, which is helpful in a variable selection setting such as ours because it makes exploration of the space of sets of pathogenic variants more efficient. Samples are drawn from a series of related distributions called power posteriors. Each power posterior has a temperature t between 0 and 1 and is proportional to the likelihood of the parameters to the power of t times the prior. These samples can be combined to obtain an estimate of the integrated likelihood (see Appendix A).

Sampling for our model can be done very efficiently because an MCMC update to z_j entails changes only in $f(G_{i\cdot}, z, s_i, m)$ for individuals for whom $G_{ij} > 0$. For convenience, we estimate the evidence conditional on m but we can integrate over it through simple summation. Once the MCMC samples have been collected, the marginal posterior probability of z given γ and m can be obtained directly and used for ranking variants by their likely pathogenicity. The estimated number of pathogenic variants and the expected posterior number of case subjects explained by the pathogenic variants, given $\gamma = 1$, can also be computed (see Appendix A). The posterior probability of γ provides a natural means of ranking sets of variants from different loci across the genome.

The model above assumes that the prior probabilities of variant pathogenicity are conditionally independent. However, particular classes of variants in a locus may confer disease risk, while others may be benign. We can impose a prior correlation structure on the z reflecting these competing hypotheses by fitting a different association model for each class of variant. If one of the hypotheses matches the true etiology of disease, then this modeling approach can improve model fit and thus increase power. Let $\gamma \in \{1, 2, \ldots, g\}$ index the association models and let I_{uv} indicate whether variant v is included in association model u. Then, we can compute the probability of association across the competing models as:

$$\mathbb{P}(\gamma > 0 \mid y, G, c, I) = \frac{\sum_{u=1}^{g} \mathbb{P}\left(y \mid \gamma = u, G^{(u)}, c^{(u)}\right)\mathbb{P}(\gamma = u)}{\sum_{u=0}^{g} \mathbb{P}(y \mid \gamma = u, G^{(u)}, c^{(u)})\mathbb{P}(\gamma = u)},$$

where $G^{(u)} = G_{\cdot\{v:I_{uv}=1\}}$ and $c^{(u)} = c_{\{v:I_{uv}=1\}}$. The prior on the model indicator, $\mathbb{P}(\gamma)$, can be informed by external data. For example, if a gene has a high probability of loss-of-function intolerance,[11] then the prior corresponding to a model of high-impact variants in that gene could be up-weighted relative to competing models. We can also compute the posterior probability of variant pathogenicity averaged over all association models using the following expression:

$$\mathbb{P}(z \mid \gamma > 0, y, G, c, I) = \frac{\sum_{u=1}^{g} \mathbb{P}\left(z \mid \gamma = u, y, G^{(u)}, c^{(u)}\right)\, \mathbb{P}\left(\gamma = u \mid y, G^{(u)}, c^{(u)}\right)}{\sum_{u=1}^{g} \mathbb{P}(\gamma = u \mid y, G^{(u)}, c^{(u)})}.$$

Other quantities of interest, such as the expected posterior number of cases explained by pathogenic variants, can be averaged over models in the same way.

Simulation Set-Up

We conducted a simulation study under different scenarios and using different methods in order to evaluate power to detect associations, to assess accuracy in variant pathogenicity classification, and to investigate the effect of integration of variant-level co-data on inference. We generated random allele count matrices for 1,000 individuals at k rare variant sites with allele frequencies of 0.0017 and 0.03 for the dominant and recessive modes of inheritance, respectively. We used $k = 25$ for the main simulation study. We labeled the first five variants pathogenic and the remaining variants non-pathogenic. The case/control labels were simulated using the expression in Equation 1, assuming $s_i = 2$ (i.e., diploidy), a particular mode of inheritance (either dominant or recessive), and a particular combination of values for τ and $\pi \in \{0, (1/10), (2/10), \ldots, 1\}$ such that $\pi > \tau$. Our selection of τ and π is comprehensive but for rare diseases we would expect $\tau < 0.5$ and $\pi \gg 0.5$. For each combination of mode of inheritance, value of τ, and value of π, 5,000 allele count matrices were generated and 5,000 corresponding case/control vectors were generated. The 5,000 datasets were copied and the case/control labels corresponding to the copied set were permuted to break the association between the genotypes and the phenotype. Thus,

under each scenario, we had a pool of 10,000 datasets, of which half were generated under a model of association and half were generated under a model of no association.

In order to assess the performance of different methods in a realistic setting, we evaluated their ability to rank non-permuted datasets among a large set of permuted datasets. Under each simulation scenario, we generated mixtures of 10 non-permuted and 990 permuted datasets selected at random from the corresponding pool. We then applied each method and computed the mean positive predictive value (PPV), over 10,000 repetitions, at 80% power. The PPV, which is equal to one minus the false discovery rate (FDR), is inversely related to power. Thus, a higher PPV for a given power implies greater power for a given FDR. We preferred to evaluate PPV for a given power rather than power for a given FDR because empirical power changes monotonically as the rank threshold for declaring a positive result is lowered, while the empirical FDR does not.

We selected the methods ADA, CAST, and SKAT for comparison as they represent diverse approaches: ADA enables individual variant-level inference, CAST is based on the popular Burden test but can account for either dominant or recessive inheritance modes, and SKAT is a popular and flexible method designed for rare variants affecting complex traits. The default linear kernel function for SKAT is used here. The other methods mentioned above were either inapplicable (e.g., vbdm requires a continuous response), unavailable (BRVD), or shown to be inferior to ADA in a previous publication.[12] Note that ADA p values were computed using 10,000 permutations instead of the default 1,000 in order to reduce the number of ties (parts of the ADA code were re-implemented in C++ in order to complete our simulation study in a reasonable amount of time; modified code available on request).

The results were ranked based on the posterior probability that $\gamma = 1$ for BeviMed and the negative log p value of association for the other methods. Variants were ranked according to BeviMed's marginal posterior probability for the components of z and according to inclusion in ADA's variant selection. The other methods do not provide variant-level inference. Although the backward elimination procedure is implemented for SKAT, it is so slow as to make its use impractical in even a moderately sized study such as this.

To demonstrate the effect of including prior information regarding variant pathogenicity on BeviMed's inference, we conducted a further study whereby we simulated datasets with $m = m_{\text{dom}}$, $\tau = 0.2$, and $\pi = 0.85$ and modified the values of the variant-specific co-data $\tilde{c}_j$ as follows. The values of $\tilde{c}_j$ for all variants was set to either 1 or 0. The number of truly pathogenic variants that were assigned the value 1 was set to 0, 1, 2, 3, 4, or 5, and the number of truly non-pathogenic variants that were assigned the value 1 was set to 0, 4, 8, 12, 16, or 20. Thus, the proportions of correctly and incorrectly up-weighted variants were varied between 0 and 1 in increments of 0.2. In the extreme, the co-data could support the true classification exactly or support the inverted classification exactly. As SKAT can incorporate variant-specific relative weights, we applied it to the same simulated data, setting SKAT's weights for up-weighted variants to twice that of the others. This choice of up-weighting factor was as low as possible while ensuring that, when the weights were perfectly concordant with the true pathogenicity of the variants, the PPV was approximately the same for SKAT as for BeviMed. Expected PPV at 80% power was estimated as described above, based on 5,000 truly associated and 5,000 permuted datasets, for BeviMed and SKAT under each combination of proportions of correctly up-weighted and incorrectly up-weighted variants.

Application to Real Data

The NIHR BioResource–Rare Diseases has generated whole-genome sequencing data for 6,586 unrelated individuals with diverse rare diseases in an effort to identify novel genetic etiologies. We applied BeviMed to the data, setting the case/control status based on two dichotomous phenotypes represented in the study: pathologically low numbers of platelets in the blood stream (thrombocytopenia) with absence of syndromic features ($\sum_i y_i = 184$) and Roifman syndrome ($\sum_i y_i = 3$).

Hereditary thrombocytopenia can be caused by variants in a large number of genes with diverse functions, including genes encoding transcription factors, cytoskeletal proteins, and membrane proteins.[13] Severe thrombocytopenia is typically monogenic and non-syndromic forms are usually dominant. Roifman syndrome (MIM: 616651) is a rare autosomal-recessive disease with symptoms including growth retardation, spondyloepiphyseal dysplasia, and cognitive delay, initially described by Roifman et al.[14] Last year, variants in the non-coding gene *RNU4ATAC* (MIM: 601428) were identified as the cause of this disease on the basis of pedigree studies involving six case subjects.[15] Within the bleeding and platelet disorders branch of the NIHR BioResource dataset, three unrelated case subjects with Roifman were enrolled because they presented with immune thrombocytopenia.

For each gene, we considered single-nucleotide variants (SNVs) and short insertions/deletions (indels) with an allele frequency in ExAC[11] and the whole-genome sequencing component of UK10K[16] less than 1/1,000 and large deletions overlapping exons with an internal frequency less than 1/1,000. SNVs and indels had to have a HIGH or MODERATE Variant Effect Predictor (VEP)[17] impact or they had to have a VEP Sequence Ontology-coded consequence that included non_coding_transcript_exon_variant, 5_prime_UTR_variant, or 3_prime_UTR_variant. If a variant had consequences in relation to multiple transcripts of the same gene, only the highest-impact consequence was retained. In total, we considered 1,338,830 variants in 35,205 gene loci, each containing between 1 and 2,615 variants.

We set $\mathbb{P}(m = m_{\text{dom}} \mid \gamma = 1)$ to 0.8 for thrombocytopenia and 0.1 for Roifman syndrome, to reflect the belief that these diseases tend to be dominantly and recessively inherited, respectively. For each locus, we considered four association models corresponding to four classes of variants:

- High: large deletions and variants with a HIGH impact annotation
- Moderate: variants with a MODERATE or HIGH impact annotation or a consequence including non_coding_transcript_exon_variant but none of the UTR-related consequences
- 5′ UTR: variants without a MODERATE or HIGH impact annotation and a consequence including 5_prime_UTR_variant
- 3′ UTR: variants without a MODERATE or HIGH impact annotation and a consequence including 3_prime_UTR_variant

The hyperparameters were assigned default values except for α_ω and β_ω, which we set to (2,1) instead of (2,8) under the "high" model. This reflects a belief that a greater proportion of variants

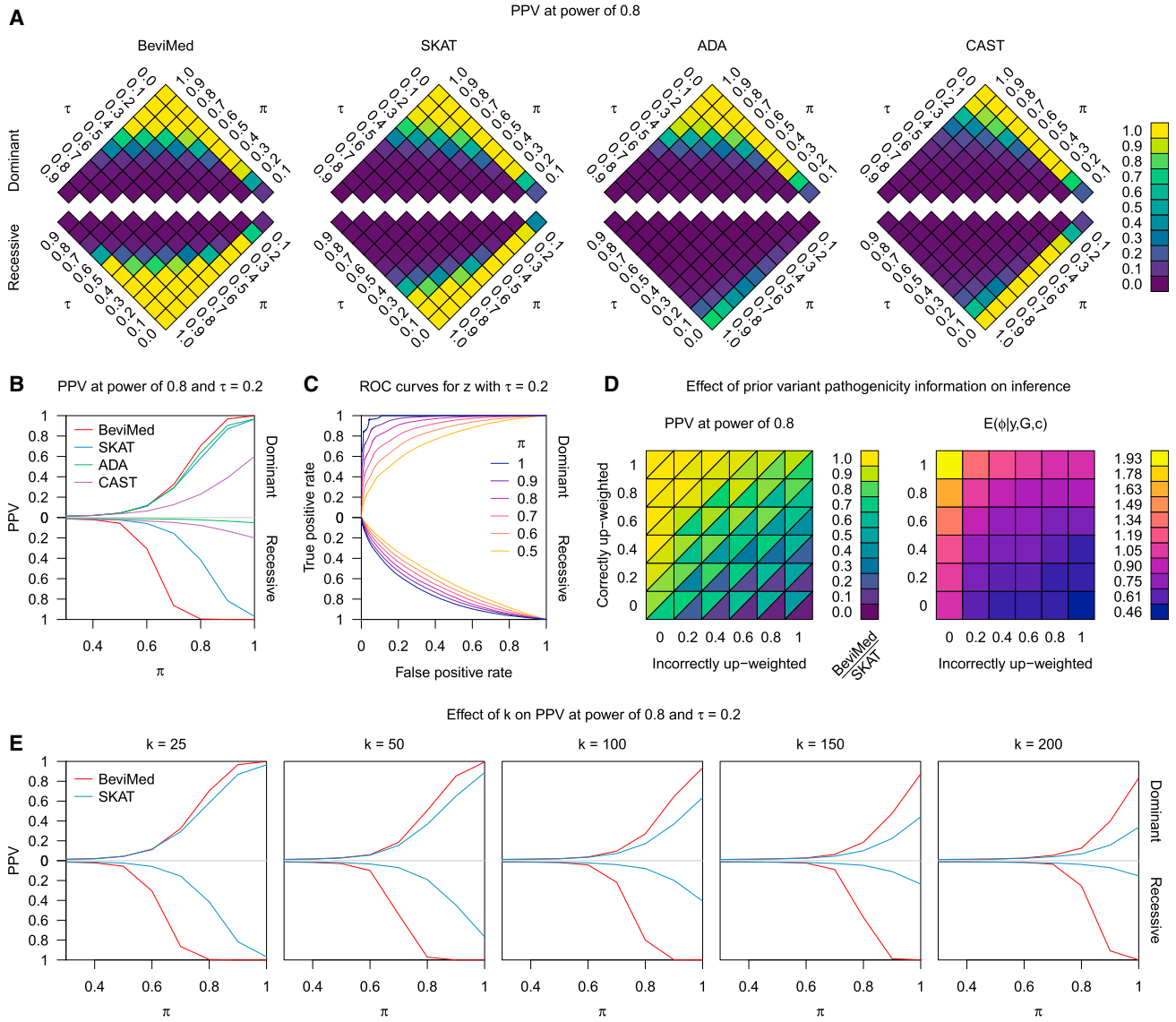

Figure 1. Simulation Study

(A and B) Results of the simulation study. Mean PPV at power of 80% over repeat simulation of the BeviMed, SKAT, ADA, and CAST rare variant association tests for data simulated using the expression in Equation 1 for various combinations of values of τ and π.

(C) Receiver operating characteristic (ROC) curves for the classification of variants as pathogenic by BeviMed for different values of π.

(D) Left: mean PPV at power of 80% for BeviMed and SKAT at $\tau = 0.2$ and $\pi = 0.85$, for varying proportions of pathogenic and non-pathogenic variants being up-weighted in the co-data variables. Right: posterior mean of ϕ corresponding to the applications of BeviMed on the left-hand grid.

(E) Mean PPV at power of 80% over repeat simulation of the BeviMed and SKAT association tests for different values of k.

are likely to be pathogenic under the high model than under the other three models. When we fitted the "moderate" model, we up-weighted the variants that were also included in the high class relative to the others by setting their uncentered weights $\tilde{c}_j$ to 1 rather than 0. For coding loci, we assigned prior probabilities of 0.004, 0.003, 0.002, and 0.001 to the four models above, respectively, in order to reflect the relative biological plausibility of the different classes of variants being involved in disease. For non-coding genes, we assigned a prior probability of the moderate model equal to 0.01. Thus, $\mathbb{P}(\gamma > 0) = 0.01$ for all genes. For completeness, we also applied SKAT to the four classes of variants described above separately, using default settings, and retained the result with the lowest p value for each locus.

Results

Simulation Study

Under a dominant model, BeviMed had a slightly higher PPV than the other methods while, under a recessive model, it greatly outperformed competing methods: when $\pi = 0.8$ and $\tau = 0.2$, BeviMed had a PPV of 100%, while SKAT, CAST, and ADA had a PPV of 42%, 8%, and 2%, respectively (Figures 1A and 1B). This favorable performance was achieved despite using the same priors for model parameters τ and π, irrespective of the values of τ and π used to simulate the data. We note that BeviMed's

Table 1. Performance Comparison

Method	N = 1,000, k = 25	N = 1,000, k = 100	N = 5,000, k = 25	N = 5,000, k = 100	N = 100,000, k = 1,000
Association Tests with Variant Identification					
BeviMed	0.03	0.09	0.07	0.23	38.90
ADA	3.69	11.30	18.76	59.23	–
BE-SKAT	53.46	175.18	137.39	799.80	–
Association Tests without Variant Identification					
CAST	0.01	0.01	0.02	0.05	1.77
SKAT	0.02	0.05	0.09	0.30	140.82
SKAT (IBS)	3.73	4.38	548.47	675.62	–
SKAT (quadratic)	4.08	4.12	571.69	598.96	–
SKAT (2wayIX)	4.09	4.37	580.90	575.67	–

Execution times in seconds of different association tests for datasets with different N and k. BE-SKAT refers to SKAT with backward elimination of variants. SKAT (IBS), SKAT (quadratic), and SKAT (2wayIX) refer to application of SKAT using the weighted identity by state, quadratic, and two-way interactions kernel functions, respectively. The p values for ADA and BE-SKAT were computed using their default number of permutations, respectively 1,000 and 300. Dashes indicate that the method took longer than 1 hr to run.

performance for $\tau = 0.2$ was approximately the same for the following three pairs of values for the hyperparameters α_ω and β_ω: (2,8), which is the default, (1,1), which places a uniform prior on expit(ω), and (2,1), which places higher prior weight on values of expit(ω) near 1.

For $\tau = 0.2$ and high π, BeviMed was able to provide accurate rankings of variants by estimated pathogenicity, particularly under a dominant mode of inheritance (area under the curve = 0.97 at $\pi = 0.9$, Figure 1C). ADA's average classification of variant pathogenicity at $\pi = 0.9$ gave a true positive rate of 0.78 and a false positive rate of 0.063, while BeviMed's true positive rate at that same false positive rate was 0.88.

The results of the simulation study assessing the effect of incorporating variant weights show that BeviMed is substantially more robust to deleterious weightings (Figure 1D, left). When the co-data matched the truth perfectly, the power for BeviMed and SKAT was approximately the same (by design), but when the co-data was entirely counter-productive, BeviMed's PPV was 0.46 and SKAT's PPV was 0.06. BeviMed's advantage was achieved naturally in our Bayesian setting through modulation of ϕ, which had a posterior expectation of 1.93 when the co-data was most useful but only 0.46 when it was least useful (Figure 1D, right).

We evaluated the performance of BeviMed in relation to the most competitive alternative method, SKAT, using the same parameters described above, but increasing the total number variants k to 50, 100, 150, and 200. Power decreased for both methods as the total number of variants increased, but the discrepancy in power between BeviMed and SKAT increased (Figure 1E). For example, under the dominant model, BeviMed's PPV at $k = 200$ and $\pi = 1.0$ was 83% while SKAT's PPV was only 34%.

Computational Performance

We compared the execution times of the different association tests, including SKAT with backward elimination, on simulated datasets generated as described above using $N \in \{1000, 5000, 100000\}$, $k \in \{25, 100, 1000\}$, and allele frequency of 1/1,000. The results, displayed in Table 1, show that CAST is the fastest method, as it uses a straightforward Fisher's exact test. However, CAST is substantially less powerful than BeviMed under both dominant and recessive models (Figure 1). BeviMed has comparable execution time to SKAT for small datasets and surpasses it for large datasets, as BeviMed's complexity scales with $\sum_{i,j} G_{ij} > 0$, which typically increases only linearly with N. SKAT was also run using the other kernels available in the R package: identity by state, quadratic, and two-way interactions. All three modes were substantially slower than BeviMed and linear SKAT (Table 1), less powerful than BeviMed in the simulation study, and less powerful than linear SKAT under at least one mode of inheritance. BeviMed has vastly superior performance to the other methods which can infer the pathogenicity of variants, while also reporting posterior uncertainty in pathogenicity status. The complete set of applications to the data from the NIHR BioResource, which comprises 2 phenotypes, 35,205 loci, 2 modes of inheritance, and 4 variant classes, took 7 hr to complete using 16 CPU cores.

Identifying Associations with Thrombocytopenia

The median value of the posterior probability of association with thrombocytopenia across all gene loci was 0.0064. The independent gene loci for which the posterior probability of association exceeded 0.9 are shown in Table 2. We show the posterior probability of association, the posterior probability of the mode of inheritance parameter, the estimated number of case subjects explained by the pathogenic variants, the estimated number of pathogenic variants that are present in the case subjects, and the total number of variants considered. These results corroborate established knowledge of platelet disorders. *ACTN1* (MIM: 102575)-related macrothrombocytopenia is a dominant bleeding

Table 2. Independent Loci Having $\mathbb{P}(\gamma = 1 \mid \boldsymbol{y}) > 0.9$

Locus	Posterior Probability of Association	Posterior Probability of Dominance	Modal Model	Estimated Number of Explained Case Subjects	Estimated Number of Explaining Variants	Number of Variants
ANKRD26	1.000	1.000	5′ UTR	10.792	7.792	87
RUNX1	1.000	1.000	moderate	8.153	8.191	214
MYH9	1.000	1.000	moderate	10.964	9.116	141
GP1BB	0.999	1.000	moderate	8.223	7.228	69
ACTN1	0.999	1.000	moderate	9.867	7.867	121

Independent loci with a posterior probability of association with thrombocytopenia greater than 0.9.

and platelet disorder recognized since 2013.[18] *GP1BB* (MIM: 138720) has traditionally been linked to a recessive bleeding and platelet disorder called Bernard-Soulier syndrome,[19] but earlier this year we reported a dominant mode of inheritance resulting in a milder platelet phenotype.[20] The posterior on the mode of inheritance parameter strongly favored dominance in this case, which is consistent with an absence of Bernard-Soulier-affected case subjects in our dataset. *RUNX1* (MIM: 151385) encodes a transcription factor linked with a dominant platelet disorder with associated myeloid malignancy. *MYH9* (MIM: 160775) harbors variants responsible for *MYH9*-related disorder, which is characterized by macrothrombocytopenia and occasional Döhle-like inclusion bodies in neutrophils and pathologies of the ear, eye, kidney, or liver. Finally, variants in the 5′ UTR of *ANKRD26* (MIM: 610855) were reported to result in non-syndromic macrothrombocytopenia in 2011[21] after an initial erroneous report that variants in the neighboring gene *ACBD5* (MIM: 616618) were responsible.[22] The association we have identified is driven by variants in the 5′ UTR, despite this class of variants being down-weighted relative to the classes comprising variants with missense or high-impact predicted consequences on the translated gene product. The variant level results of the inference shown in Figure 2 indicate the high posterior probability of association for the first eight variants in the 5′ UTR. It is noteworthy that one of the variants, which encodes c.−113A>C and was reported in a follow-up[23] to the original 2011 paper, does not appear to be pathogenic, as five out of the six individuals with the variant, including one homozygous for the alternate allele, do not have a bleeding or platelet disorder.

There were four additional loci having $\mathbb{P}(\gamma = 1 \mid y) > 0.9$. They all tagged a true association listed in Table 2 but were labeled with the names of neighboring genes and had lower posterior probabilities of association. A missense variant in *WAC* (MIM: 615049) was in linkage disequilibrium with one of the 5′ UTR variants in *ANKRD26*, inducing $\mathbb{P}(\gamma = 1 \mid y) = 0.993$ for the *WAC* locus. The other three results were induced by the presence of deletions in *RUNX1* spanning three neighboring RNA genes or pseudogenes: *AF015262.2* ($\mathbb{P}(\gamma = 1 \mid y) = 0.978$), *RPL34P3* ($\mathbb{P}(\gamma = 1 \mid y) = 0.976$), and *EZH2P1* ($\mathbb{P}(\gamma = 1 \mid y) = 0.974$).

The alternate method that was most powerful based on the results of the simulation, SKAT, did not rank the loci listed above as highly, even when only the variant class with the lowest p value was retained for each gene. *RUNX1*, *ANKRD26*, *MYH9*, *ACTN1*, and *GP1BB* had ranks of 1, 3, 8, 16, and 74, respectively, with none of the other loci in the top 20 ranks being known to be implicated in thrombocytopenia.

Identifying Variants Responsible for Roifman Syndrome

The locus with the highest posterior probability of association with the Roifman syndrome case label was *RNU4ATAC* ($\mathbb{P}(\gamma = 1) = 1.000$), driven by four different single-nucleotide variants in this non-coding gene. Two of the case subjects were compound heterozygous, including for a variant observed in six control subjects, and one was homozygous. As all but one of the variants were seen only in heterozygosity, the posterior probability of variant pathogenicity conditional on recessive inheritance was relatively high across the gene but markedly lower than the causal variants observed in the case subjects, which had a posterior probability of pathogenicity very close to 1 (Figure 3). All other genes had a posterior probability of association less than 0.9 and the expected number of case subjects explained by the variants in other loci was less than 2. SKAT assigned *RNU4ATAC* a p value of 0, but this was also the case for 34 other genes, which were tied in the top rank.

Discussion

We have presented a Bayesian genetic association method for rare diseases that is more powerful than existing methods, particularly for the recessive mode of inheritance, and provides summary statistics on variant-level pathogenicity and mode of inheritance very efficiently. It enables mode of inheritance to be integrated out or inferred from the data. Indeed, we were able to determine a dominant mode of inheritance for variants in a gene, *GP1BB*, that has been associated only with a recessive disorder for more than 30 years. Given an association under a particular mode of inheritance, our method also estimates the number of case subjects explained by pathogenic variants and the number of variants that are pathogenic.

Prior information specific to a particular set of variants under consideration can modulate the evidence of association, which can be critical when the number of case subjects

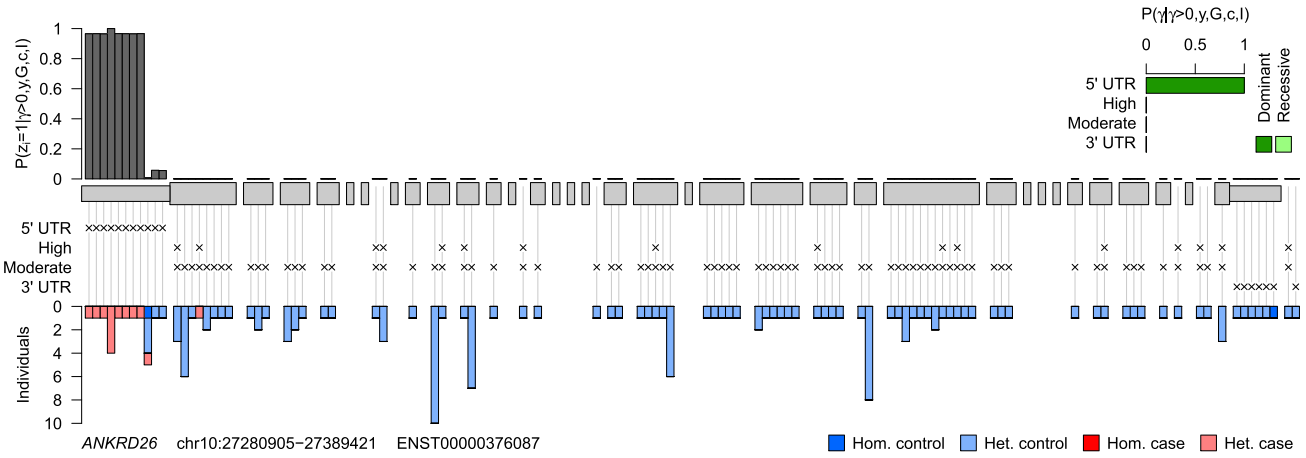

Figure 2. Posterior Probability of Pathogenicity for Rare Variants in *ANKRD26*
Results obtained by applying our inference procedure to rare allele counts in *ANKRD26* against the thrombocytopenia case/control label. Exons are represented by gray blocks starting from the 5′ UTR on the left and ending with the 3′ UTR on the right. The classes that each variant belongs to are indicated by crosses. The bar chart in the top right shows the posterior probability of each association model under each mode of inheritance conditional on an association being present at the locus. The gray bars above show the marginal posterior probabilities of pathogenicity for individual rare variants conditional on an association being present at the locus. The inference algorithm was run with 100,000 iterations instead of the usual 1,000 in order to reduce jitter due to Monte Carlo sampling error. The bar chart beneath shows the breakdown of heterozygous and homozygous carriers of the variants between case and control subjects.

with a shared genetic etiology is small. For example, the prior on the model indicator γ can be adjusted to reflect locus-specific genomic and epigenomic knowledge in order to encourage regions with higher prior plausibility of involvement in the disease phenotype to rank more highly than if the same prior had been used across all regions. The prior probability of pathogenicity for a particular variant, given the association model, can be modulated by knowledge about the variant, such as predicted consequence, allele frequency, or conservation. These variant weights are interpretable as prior shifts in the log odds of pathogenicity, which provides an intuitive basis for assigning particular values to them. In sharp contrast to frequentist approaches, we use a flexible prior on the effect sizes of the weightings that reflects the uncertainty in their utility.

The results of inference on different subsets of rare variants in a locus (selected, for example, on the basis of their predicted consequences) can be interpreted and combined easily in a Bayesian framework using a model selection procedure. The posterior probability of variant pathogenicity and other quantities of interest can be averaged over models. In addition to increasing statistical power if particular classes of variants in a locus are the only ones that confer disease risk, this feature also allows inference of the kind of variants responsible for disease, which may suggest particular genetic etiologies. In our applications, we were able to identify a set of variants in the 5′ UTR of a gene that causes a platelet disorder. The high posterior probability of pathogenicity of variants in the 5′ UTR to the exclusion of coding variants, even those observed only in case subjects, was made possible by our model selection procedure.

Variants highlighted by a method such as ours would usually undergo assessment by a multidisciplinary diagnostic team and it would resolve increasing numbers of case subjects over time. In our application to real data, we have kept the case/control labels the same for each application. However, in the context of genetically heterogeneous diseases, we would recommend relabelling any case subject whose phenotype has been fully accounted for by pathogenic or likely pathogenic variants in a different locus as a control. This boosts specificity as it makes it less likely for a non-pathogenic rare variant carried by a case to induce a high probability of association.

The model assumes that relatedness between individuals is sufficiently low as not to be associated with either case/control status or the genotypes. In practice, we recommend removal of any first, second, or third degree relatives. Our method is designed to be applied to up to thousands of rare variants at a time and efforts should be made to ensure all potentially implicated variants in a locus are included in the model, or the set of models, under comparison. Rare variants would typically be unlinked within a locus but may occasionally be linked across loci. For example, large deletions may span multiple genes and certain pairs of rare variants could be in linkage disequilibrium. In these situations, a non-pathogenic rare variant in one locus linked to a pathogenic variant in another locus could induce a non-causal association. Such associations can either be filtered post hoc through comparison of inference results in nearby loci or avoided altogether by joint modeling of variants across multiple nearby loci.

Although Bayesian inference is typically thought of as slow, our implementation can handle data from more than a million variants spread across tens of thousands of regions called in thousands of samples in a few hours. BeviMed is thus capable of handling with ease the demands of modern genomic datasets in the coding and the regulatory regions of the genome.

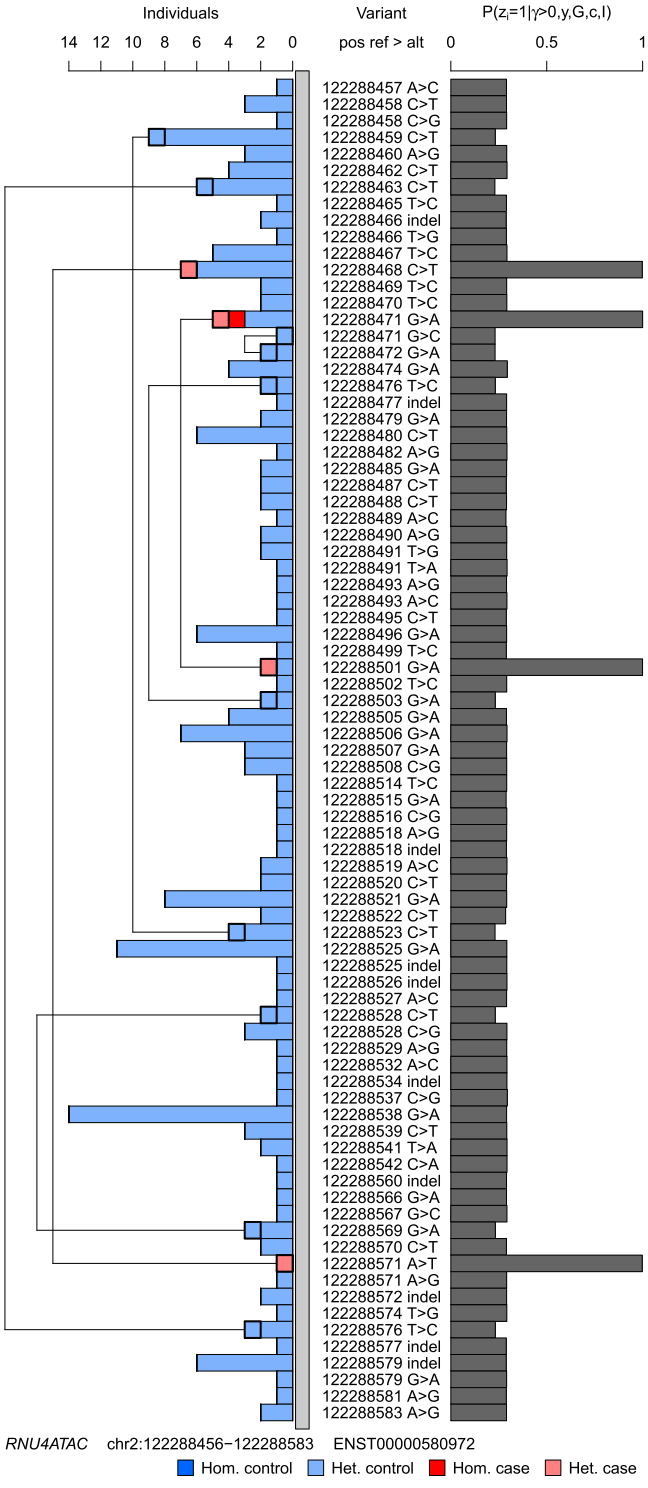

Figure 3. Posterior Probability of Pathogenicity for Rare Variants in *RNU4ATAC*

Results of applying the inference procedure to rare allele counts in *RNU4ATAC* against the Roifman syndrome case label. The bar chart on the right shows the marginal posterior probabilities of pathogenicity for each rare variant conditional on an association being present at the locus. The inference algorithm was run with 100,000 iterations instead of the usual 1,000 in order to reduce jitter due to Monte Carlo sampling error. The bar chart on the left shows the breakdown of heterozygous and homozygous carriers of the variants in case and control subjects. Compound heterozygous individuals with two rare alleles in *RNU4ATAC* were observed, and for each such individual a line is drawn linking the two variants.

Appendix A

Inference on presence of an association is based on the posterior probability of the model indicator γ, which can be derived from the evidence under each model and the prior on γ:

$$\mathbb{P}(\gamma = 1 \mid y) = \frac{\mathbb{P}(y \mid \gamma = 1)\mathbb{P}(\gamma = 1)}{\sum_{u \in \{0,1\}} \mathbb{P}(y \mid \gamma = u)\mathbb{P}(\gamma = u)}.$$

The evidence for the baseline model, $\mathbb{P}(y \mid \gamma = 0)$, can be computed efficiently using the beta function:

$$\frac{B(\alpha_0 + \sum_i y_i, \beta_0 + N - \sum_i y_i)}{B(\alpha_0, \beta_0)}.$$

The evidence for the association model, $\mathbb{P}(y \mid \gamma = 1)$, can be expressed by conditioning on the different modes of inheritance and summing:

$$\begin{aligned}\mathbb{P}(y \mid \gamma = 1) = {} & \mathbb{P}(y \mid \gamma = 1, m = m_{\text{dom}})\mathbb{P}(m = m_{\text{dom}} \mid \gamma = 1) \\ & + \mathbb{P}(y \mid \gamma = 1, m = m_{\text{rec}})\mathbb{P}(m = m_{\text{rec}} \mid \gamma = 1),\end{aligned}$$

where $\mathbb{P}(y \mid \gamma = 1, m)$ is given by:

$$\sum_{z \in \{0,1\}^k} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y \mid z, m)p(z \mid \omega, \phi)p(\omega)p(\phi)\, \mathrm{d}\omega\, \mathrm{d}\phi.$$

For brevity, we have omitted the hyperparameters α_τ, β_τ, α_π, and β_π from the conditioning in $p(y \mid z, m)$ above.

The likelihood $p(y \mid z, m)$ factorizes into two components corresponding to individuals with and without a pathogenic combination of alleles, where the rate parameters τ and π can be integrated out analytically. Thus the likelihood can be expressed in closed form:

$$\begin{aligned}&\frac{B\left(\alpha_\tau + \sum_i y_i\left(1 - x_i^{(z)}\right), \beta_\tau + \sum_i (1 - y_i)\left(1 - x_i^{(z)}\right)\right)}{B(\alpha_\tau, \beta_\tau)} \\ &\quad \times \frac{B\left(\alpha_\pi + \sum_i y_i x_i^{(z)}, \beta_\pi + \sum_i (1 - y_i) x_i^{(z)}\right)}{B(\alpha_\pi, \beta_\pi)},\end{aligned} \qquad \text{(Equation A1)}$$

where $x_i^{(z)} = f(G_i., z, s_i, m)$.

As noted in the main text, we use a logit-beta prior on ω, that is:

$$\text{expit}(\omega) \sim \text{Beta}(\alpha_\omega, \beta_\omega).$$

Thus, when $c_j = 0\ \forall j \in \{1, \ldots, k\}$, z is independent of ϕ, and both ω and ϕ can be integrated out:

$$p(z) = \frac{B\left(\alpha_\omega + \sum_j z_j, \beta_\omega + k - \sum_j z_j\right)}{B(\alpha_\omega, \beta_\omega)}.$$

By default, $\alpha_\omega = 2$ and $\beta_\omega = 8$. The space of z grows exponentially with the number of variants k, which can

run into the dozens or hundreds. Therefore, despite the formulation of the model enabling many of the parameters to be integrated out analytically, the expression for the integrated likelihood cannot be evaluated in practice. Below we describe an alternative method to estimate the integrated likelihood which is computationally tractable.

The method of power posteriors[10] allows us to estimate $\mathbb{P}(y \mid \gamma = 1, m)$ by sampling from $\mathbb{P}(z \mid y, m)^t \mathbb{P}(z)$ at temperatures $t \in \{t_0 = 0, \ldots, t_L = 1\}$ using MCMC. Let $\tilde{z}_l^{(b)}$ be the bth sample drawn at temperature t_l. The log integrated likelihood $\log \mathbb{P}(y \mid \gamma = 1, m)$ can then be estimated by:

$$\sum_{l=0}^{L-1} \log\left(\frac{1}{|\tilde{z}_l|} \sum_{b=1}^{|\tilde{z}_l|} e^{(t_{l+1}-t_l)\ \log\ \mathbb{P}\left(y \mid \tilde{z}_l^{(b)},\ m\right)}\right). \quad \text{(Equation A2)}$$

Running Markov chains at different temperatures concurrently allows exchanges of state between chains at adjacent temperatures, which encourages good mixing. If the Kullback-Leibler (KL) divergence between adjacent power posterior distributions is large, the resulting estimates of $\mathbb{P}(y \mid \gamma = 1, m)$ may be susceptible to substantial numerical error. The method that minimizes this error involves tuning the temperatures using a procedure such as interval bisection[24] and subsequently re-generating the chains to allow mixing between them. By default we use a preselected set of temperatures $t = (l/6)^2$ for $l \in \{0, 1, \ldots, 6\}$, and draw 1,000 samples from each chain. This works well in practice and avoids the need to discard an initial set of MCMC samples for tuning the temperatures.

The use of MCMC to tackle this overall inference problem is in contrast to other methods designed for similar purposes,[12,25] probably because of the stringent requirements for computational speed. However, our algorithm contains features that makes MCMC sampling efficient.

In each chain, Gibbs sampling is used to update each individual component of z in turn. An update to z_j consists of sampling from its full conditional distribution:

$$\text{Bernoulli}\left(\left(1 + \frac{p(y \mid z^{(0)}, m)p(z^{(0)} \mid \omega, \phi)}{p(y \mid z^{(1)}, m)p(z^{(1)} \mid \omega, \phi)}\right)^{-1}\right),$$

where $z_{j'}^{(0)} = z_{j'}^{(1)} = z_{j'}$ for $j' \neq j$ and $z_j^{(0)} = 0, z_j^{(1)} = 1$. During the course of the algorithm, we keep track of $x_i = f(G_{i\cdot}, z, s_i, m)$ for each individual. Given an update of a single component of z, only individuals for whom $G_{ij} > 0$ need to have their corresponding value of x_i updated. G is often sparse as it typically represents rare allele counts, allowing this operation to be performed quickly. If values for c are specified, then ω and ϕ are updated using a Metropolis Hastings within Gibbs.

Averaging over the space of all variant/variable sets using MCMC is a daunting challenge, in particular in circumstances such as these where non-additive models for the interaction effects of the variables are used. However, in practice, there is little collinearity between rare variant allele counts in unrelated individuals, rare allele count matrices are sparse, and the interaction effects in dominant and recessive inheritance are quite simple, leading to low correlation between the elements of z. This means that the sampling procedure can explore the space of z efficiently.

However, when k is large and the mode of inheritance is recessive, with some case subjects being compound heterozygous, mixing of the MCMC sampler could potentially be poor if only one element of z is updated at a time. In particular, it could be very rare for the Markov chain to transition from a state satisfying $z_{j_1} = z_{j_2} = 0$ for some truly pathogenic variants j_1, j_2, to a state where $z_{j_1} = z_{j_2} = 1$, as there may be no intermediate state that would lead to an increase in likelihood. This is particularly problematic if the prior on ω is concentrated near 0. Thus, under $m = m_{\text{rec}}$, we propose updates to elements of z corresponding to variants occurring in the same individuals in tandem, which overcomes the potential rarity of sampling a state with a high likelihood.

The likelihood shown in Equation A1 can be expressed in terms of ratios of gamma functions with arguments that differ by integer amounts less than or equal to the number of individuals. Hence, the differences between all possible return values of $\log \Gamma$ that are required by the procedure can be computed before commencing the sampling and stored to avoid evaluating the $\log \Gamma$ function repeatedly. Evaluating $\log \Gamma$ is the computational bottleneck in the Markov chain updates, and replacing it with look-ups in the pre-computed values tables results in significant speed-ups.

To further improve computational efficiency while maintaining adequate precision, the implementation provides an option to stop sampling once the estimated evidence lies within a given confidence interval, or once there is sufficient confidence that the log evidence is below a given threshold. This behavior is implemented based on the method of consistent batch means.[26] The log evidence is the sum of the logarithms of expectations taken with respect to the power posteriors (Equation A2), so the central limit theorem does not apply and we estimate the confidence interval by simulation.

The samples drawn from the MCMC routine at the temperature $t = 1$ can be used to compute the expected number of case subjects whose risk was due to their pathogenic configuration of alleles, $\mathbb{E}_{z \mid y} \sum_i y_i x_i^{(z)}$, and the expected number of variants involved in the explanation, $\mathbb{E}_{z \mid y} \sum_{j \in \{j' : \sum_i G_{ij'} y_i x_i^{(z)} > 0\}} z_j$.

Supplemental Data

Supplemental Data consists of a table listing additional members and collaborators of the NIHR BioResource and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.05.015.

Acknowledgments

This work was supported by NIHR award RG65966 (D.G. and E.T.) and the Medical Research Council program grant MC_UP_0801/1 (D.G. and S.R.). The NIHR BioResource—Rare Diseases projects were approved by Research Ethics Committees in the UK and

appropriate national ethics authorities in non-UK enrolment centers. We are particularly thankful to the members of the bleeding and platelet disorders project for granting access to detailed phenotype data. We are grateful to Dr. William J. Astle for providing comments on the manuscript.

Received: February 22, 2017
Accepted: May 22, 2017
Published: June 29, 2017

Web Resources

ADA, http://homepage.ntu.edu.tw/~linwy/ADAprioritized.html
BE-SKAT, http://www.columbia.edu/~ii2135/Software_BE_HM.zip
BeviMed, https://cran.r-project.org/web/packages/BeviMed/
OMIM, http://www.omim.org/
SKAT, https://cran.r-project.org/web/packages/SKAT/

References

1. Marx, V. (2015). The DNA of a nation. Nature *524*, 503–505.
2. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
3. Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. *615*, 28–56.
4. Ionita-Laza, I., Capanu, M., De Rubeis, S., McCallum, K., and Buxbaum, J.D. (2014). Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. PLoS Genet. *10*, e1004729.
5. Lin, W.-Y. (2014). Adaptive combination of P-values for family-based association testing with sequence data. PLoS ONE *9*, e115971.
6. Logsdon, B.A., Dai, J.Y., Auer, P.L., Johnsen, J.M., Ganesh, S.K., Smith, N.L., Wilson, J.G., Tracy, R.P., Lange, L.A., Jiao, S., et al.; NHLBI GO Exome Sequencing Project (2014). A variational Bayes discrete mixture test for rare variant association. Genet. Epidemiol. *38*, 21–30.
7. Quintana, M.A., Berstein, J.L., Thomas, D.C., and Conti, D.V. (2011). Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. Genet. Epidemiol. *35*, 638–649.
8. Liang, F., and Xiong, M. (2013). Bayesian detection of causal rare variants under posterior consistency. PLoS ONE *8*, e69633.
9. Friel, N., and Wyse, J. (2012). Estimating the evidence–a review. Stat. Neerl. *66*, 288–308.
10. Friel, N., and Pettitt, A.N. (2008). Marginal likelihood estimation via power posteriors. J. R. Stat. Soc. Series B Stat. Methodol. *70*, 589–607.
11. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
12. Lin, W.-Y. (2016). Beyond rare-variant association testing: pinpointing rare causal variants in case-control sequencing study. Sci. Rep. *6*, 21824.
13. Lentaigne, C., Freson, K., Laffan, M.A., Turro, E., Ouwehand, W.H.; and BRIDGE-BPD Consortium and the ThromboGenomics Consortium (2016). Inherited platelet disorders: toward DNA-based diagnosis. Blood *127*, 2814–2823.
14. Roifman, C. (1997). Immunological aspects of a novel immunodeficiency syndrome that includes antibody deficiency with normal immunoglobulins, spondyloepiphyseal dysplasia, growth and developmental delay, and retinal dystrophy. Can. J. Allergy Clin. Immunol *2*, 94–98.
15. Merico, D., Roifman, M., Braunschweig, U., Yuen, R.K., Alexandrova, R., Bates, A., Reid, B., Nalpathamkalam, T., Wang, Z., Thiruvahindrapuram, B., et al. (2015). Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. Nat. Commun. *6*, 8718.
16. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
17. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol. *17*, 122.
18. Kunishima, S., Okuno, Y., Yoshida, K., Shiraishi, Y., Sanada, M., Muramatsu, H., Chiba, K., Tanaka, H., Miyazaki, K., Sakai, M., et al. (2013). ACTN1 mutations cause congenital macrothrombocytopenia. Am. J. Hum. Genet. *92*, 431–438.
19. Savoia, A., Pastore, A., De Rocco, D., Civaschi, E., Di Stazio, M., Bottega, R., Melazzini, F., Bozzi, V., Pecci, A., Magrin, S., et al. (2011). Clinical and genetic aspects of Bernard-Soulier syndrome: searching for genotype/phenotype correlations. Haematologica *96*, 417–423.
20. Sivapalaratnam, S., Westbury, S.K., Stephens, J.C., Greene, D., Downes, K., Kelly, A.M., Lentaigne, C., Astle, W.J., Huizinga, E.G., Nurden, P., et al. (2017). Rare variants in GP1BB are responsible for autosomal dominant macrothrombocytopenia. Blood *129*, 520–524.
21. Pippucci, T., Savoia, A., Perrotta, S., Pujol-Moix, N., Noris, P., Castegnaro, G., Pecci, A., Gnan, C., Punzo, F., Marconi, C., et al. (2011). Mutations in the 5′ UTR of ANKRD26, the ankirin repeat domain 26 gene, cause an autosomal-dominant form of inherited thrombocytopenia, THC2. Am. J. Hum. Genet. *88*, 115–120.
22. Punzo, F., Mientjes, E.J., Rohe, C.F., Scianguetta, S., Amendola, G., Oostra, B.A., Bertoli-Avella, A.M., and Perrotta, S. (2010). A mutation in the acyl-coenzyme A binding domain-containing protein 5 gene (ACBD5) identified in autosomal dominant thrombocytopenia. J. Thromb. Haemost. *8*, 2085–2087.
23. Noris, P., Perrotta, S., Seri, M., Pecci, A., Gnan, C., Loffredo, G., Pujol-Moix, N., Zecca, M., Scognamiglio, F., De Rocco, D., et al. (2011). Mutations in ANKRD26 are responsible for a frequent form of inherited thrombocytopenia: analysis of 78 patients from 21 families. Blood *117*, 6673–6680.
24. Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. Stat. Comput. *24*, 709–723.
25. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. *95*, 5–23.
26. Flegal, J.M., Haran, M., and Jones, G.L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? Stat. Sci. *23*, 250–260.