Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases

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

Rare genetic disorders, which can now be studied systematically with affordable genome sequencing, are often caused by high-penetrance rare variants. Such disorders are often heterogeneous and characterized by abnormalities spanning multiple organ systems ascertained with variable clinical precision. Existing methods for identifying genes with variants responsible for rare diseases summarize phenotypes with unstructured binary or quantitative variables. The Human Phenotype Ontology (HPO) allows composite phenotypes to be represented systematically but association methods accounting for the ontological relationship between HPO terms do not exist. We present a Bayesian method to model the association between an HPO-coded patient phenotype and genotype. Our method estimates the probability of an association together with an HPO-coded phenotype characteristic of the disease. We thus formalize a clinical approach to phenotyping that is lacking in standard regression techniques for rare disease research. We demonstrate the power of our method by uncovering a number of true associations in a large collection of genome-sequenced and HPO-coded cases with rare diseases.

Introduction

There is widespread interest in the study of rare diseases as a way of understanding the genetic architecture of biological processes. Consequently, tens of thousands of subjects are being phenotyped extensively and enrolled to genome-sequencing studies worldwide. To discover the cause of disease, these subjects would ideally be grouped a priori into clusters with a shared (though unknown) genetic etiology, but this is often hindered by extensive phenotypic and genetic heterogeneity (see Web Resources and examples[1–9]). Rare variant association tests, even those accounting for some degree of genetic heterogeneity, typically summarize the clinical manifestations of a disease with a single variable,[10] which can limit power when multiple phenotypic traits contain complementary information about the same causal genotype. Methods for modeling pleiotropy have proven successful in the context of genome-wide association studies[11,12] but they are ill suited for rare disease studies in which the phenotype data are typically of mixed type and collected with variable detail and completeness.

The Human Phenotype Ontology (HPO)[13] addresses the need for a standardized vocabulary for rare disease phenotypes and is being used to code patients in several large international projects[14–16] (see also Web Resources). The HPO is a directed acyclic graph representing more than 10,000 phenotypic abnormalities in which the nodes (HPO terms) are connected to each other through "is-a" relations, represented as edges. The HPO was created with the support of experts in many areas of medicine to accommodate coding of phenotypic data derived from diverse sources, such as laboratory assays, images, graphs, and clinical interpretation. Methods exist that compare patient HPO data with HPO-coded profiles corresponding to known diseases for the purpose of differential diagnosis.[17,18] The HPO-coded profiles can be supplemented with functional gene-specific information to prioritize genes.[19,20] If genotype data are available, these and other methods[21,22] can be used to prioritize variants and potentially to suggest new causes of disease.[19,20,23] However, the existing approaches do not share information between individually coded patients and as such are not statistical association methods.

Here, we present a regression-based method for discovering associations between arbitrarily diverse sets of HPO-coded phenotypes and genotypes at rare variant sites. To overcome the difficulty of modeling sparse and ontologically structured phenotype data, we treat the HPO-coded phenotypes of the subjects as the explanatory variables and their corresponding genotypes as the response. This is an example of "inverse regression" and is adequate in our setting because we are not interested in interpreting the regression coefficients per se but only in evaluating the probability of association. We define a subject's "genotype" y as a binary label that can take on the values "rare" (1) or "common" (0) according to a pre-specified function of the genetic data. For example, we could define the label "rare genotype" to mean that there is at least one rare variant in a particular gene (dominant inheritance) or at least two rare variants in a particular gene (recessive inheritance).

Our method then seeks to compare two models for the data, indexed by γ. Under the baseline model ($\gamma = 0$), the probability of observing the rare genotype is the same for each case. Under the alternate model ($\gamma = 1$), the probability of observing the rare genotype depends on the "phenotypic similarity" S (to be defined later) of the case to a latent *characteristic* HPO phenotype ϕ.

[1]Department of Haematology, University of Cambridge, NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK; [2]Medical Research Council Biostatistics Unit, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK
[3]These authors contributed equally to this work
*Correspondence: et341@cam.ac.uk
http://dx.doi.org/10.1016/j.ajhg.2016.01.008.

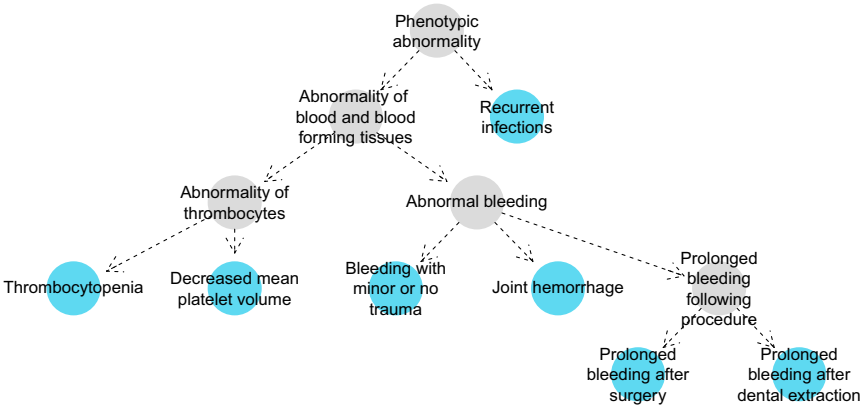

Figure 1. Example HPO Coding of a Subject with Wiskott-Aldrich Syndrome
The nodes in blue imply the presence of the more general ancestral phenotypes depicted as gray nodes. No blue node has a directed path to any other, which means that the blue nodes comprise a *minimal* set of HPO terms. The graph has been simplified by removing nodes that link together only two other nodes.

We adopt a Bayesian inference framework, where the model selection indicator γ and characteristic phenotype ϕ are estimated through their posterior distributions. Of particular interest is the posterior mean of γ, which represents the probability that $\gamma = 1$, thus indicating the strength of evidence for an association.

A crucial element of our approach is the construction of an appropriate function for quantifying the semantic similarity of the characteristic phenotype ϕ to the phenotypes of the subjects. The choice of function is motivated by the need to optimally discriminate between subjects having clinical features that are pertinent to a disorder from those having overlapping or unrelated phenotypes due to a different disorder. To achieve this, we have chosen a function that accounts for the ontological structure of the HPO and induces a parsimonious characteristic phenotype: it selects the required terms to distinguish patient groups while avoiding overfitting and is robust to coding of patients with spurious or sporadic terms. Importantly, the function is flexible with respect to the phenotypic variability of disease and robust to the HPO coding practices of clinicians.

Our Bayesian approach provides a natural means of incorporating information from the scientific literature into our prior belief about the characteristic phenotype. In this work, we focus on gene-specific inference and upweight the prior probability of characteristic phenotypes that are similar to clinical[23] and murine phenotypes[24] relevant to the gene.

We demonstrate the effectiveness of our method in identifying associations between genotype and phenotype through a simulation study, whereby phenotypes are simulated given genotypes in such a way as to emulate the effect of a hypothetical set of pathogenic variants. We go on to apply our inference procedure to a real dataset of more than 2,000 unrelated individuals enrolled to a variety of rare-disease sequencing studies under the auspices of the BRIDGE projects run by the NIHR BioResource – Rare Diseases (Web Resources). We show that our method, implemented in the SimReg software package, can identify genes with rare variants responsible for a diverse set of pathologies in a single application and can estimate recognized disease phenotypes.

Material and Methods

Model Specification

We use a logistic regression framework to specify the two models under comparison:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_i), \\ \gamma = 0: \quad & \log\left(\frac{p_i}{1-p_i}\right) = \alpha, \\ \gamma = 1: \quad & \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta S(\phi, x_i). \end{aligned} \qquad \text{(Equation 1)}$$

Here, $y_1, \ldots, y_N$ are the genotypes of the N subjects in the collection, where $y_i = 1$ if subject i possesses the rare genotype and $y_i = 0$ if subject i possesses the common genotype. $x_1, \ldots, x_N$ are the corresponding phenotypes of the subjects, where x_i comprises the minimal set of HPO terms required to describe the phenotypic abnormalities of subject i. Loosely speaking, a set of terms is minimal if it describes a patient's phenotype without redundancy (e.g., it does not include both "Abnormal bleeding" and "Joint hemorrhage"). More formally, a set of terms is said to be minimal if and only if it lacks elements implied by other terms in the set through directed edges in the HPO. The terms highlighted blue in Figure 1 comprise such a set, because there is no directed path between any pair of blue nodes.

The term $S(\phi, x_i)$ denotes a chosen measure of phenotypic similarity between the characteristic phenotype and subject i's phenotype. Note that our response and predictor are inverted compared to classical regression methods to avoid having to treat sparse and structured HPO data as the response. Under the baseline model, the intercept α is the global rate of rare genotypes. Under the alternate model, there is an additional parameter β, which is strictly positive and captures the effect of a unit increase in phenotypic similarity to the characteristic phenotype ϕ on the log odds of having the rare genotype. Thus, the probability that $\gamma = 1$ is greater in expectation when $S(\phi, x_i)$ is larger if $y_i = 1$ than if $y_i = 0$.

Similarity Measure

Our chosen similarity measure S is built with consideration for (1) quantification of the similarity of terms, (2) quantification of the similarity of a patient phenotype x_i to the characteristic phenotype ϕ, and (3) flexible transformation of the similarity between phenotypes.

Consistent with the ontological literature, we base our measure for the similarity of terms on the information content (IC) of each individual term,

$$\mathrm{IC}(t) \;=\; -\log(\mathrm{frequency}(t)),$$

where the frequency of term t can be derived from its appearance in the case collection, including instances in which this is implied by the presence of more specific terms in the ontology.

We use Lin's[25] similarity function to compare two different terms:

$$s(t_1, t_2) = \frac{2 \times \max_{t \in \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2)} \mathrm{IC}(t)}{\mathrm{IC}(t_1) + \mathrm{IC}(t_2)},$$

where anc(t) denotes the union of term t and its ancestral nodes in the HPO graph. For example, for the hypothetical subject shown in Figure 1, the expression $\max_{t \in \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2)} \mathrm{IC}(t)$ if t_1 were "Thrombocytopenia" and t_2 were "Joint hemorrhage" would correspond to the IC of "Abnormality of blood and blood-forming tissues." Because terms cannot have a higher IC than their descendants, the similarity s between two terms can range between zero and one. Next, we consider asymmetric measures of similarity between a case phenotype and ϕ inspired by the best-match-average (BMA) function,[17] which computes the best match for each term and takes the mean:

$$S_\phi(\phi \rightarrow x_i) = \frac{1}{|\,\phi\,|} \sum_{t_\phi \in \phi} \max_{t_x \in x_i} s(t_\phi, t_x) \mathbf{1}_{t_\phi \in \mathrm{anc}(t_x)},$$

$$S_x(x_i \rightarrow \phi) = \frac{1}{|\,x_i\,|} \sum_{t_x \in x_i} \max_{t_\phi \in \phi} s(t_x, t_\phi) \mathbf{1}_{t_\phi \in \mathrm{anc}(t_x)}.$$

The standard BMA function does not include the indicator variable above, which evaluates to 1 only if the node in ϕ is among the ancestors of the node in x_i. We prefer to include this restriction, which penalizes similarity to ϕ when it includes over-specific terms, in order to concentrate the posterior weight of ϕ preferentially on nodes that are present among the subjects.

The presence of a term in ϕ that is absent from x_i has the effect of lowering S_ϕ, whereas the presence of a term in x_i that is absent from ϕ has the effect of lowering S_x. Summation of two asymmetric similarities, as used in BMA, would allow reasonably high overall similarities to be obtained even when one of the two asymmetric similarities is close to zero. We prefer to multiply rather than add up the two similarity measures to obtain an expression for the overall similarity function used in Equation 1 because it ensures that the overall similarity can be high only when there is a high asymmetric similarity in both directions. However, because the values of S_x and S_ϕ are influenced by factors such as how frequent terms are in the reference database (which affects nodal IC) and the structure of the HPO graph, there is no guarantee that a linear function of their product optimally distinguishes subjects with objectively distinct clinical features. To ensure the model is robust to the choice of S, we allow modulation of the shapes of the similarity parameters, S_ϕ and S_x, through transformations f and g, respectively. A reasonable choice for f and g is the beta cumulative distribution function (CDF), because it maps [0,1] to [0,1] monotonically and allows a wide variety of shapes:

$$f(z,\, a_f,\, b_f) = I_z(a_f,\, b_f),$$

$$g(z,\, a_g,\, b_g) = I_z(a_g,\, b_g),$$

where I_z is the regularized incomplete beta function (see Supplemental Note) and a_f, a_g, b_f, and b_g are unknown parameters to be estimated.

Finally, the overall similarity function is given by

$$S(\phi, x_i) = f\big(S_\phi(\phi \rightarrow x_i),\, a_f,\, b_f\big) \cdot g\big(S_x(x_i \rightarrow \phi),\, a_g,\, b_g\big). \quad \text{(Equation 2)}$$

Priors

We propose the following prior distributions for the model indicator and the regression parameters:

$$\gamma \sim \mathrm{Bernoulli}(\pi),$$

$$\alpha \sim \mathrm{Normal}(\mathrm{mean} = 0, \mathrm{sd} = 5),$$

$$\log \beta \sim \mathrm{Normal}(\mathrm{mean} = 2, \mathrm{sd} = 1).$$

The value of π indicates how likely we believe a priori that there is a true association. All of the analyses in this paper assume $\pi = 0.05$. We place a vague prior on α around 0. Additionally, we include an offset on α by a constant $\widehat{h}_i$ for each individual that can take into account batch effects and factors affecting the background rate of rare genotypes (not shown in Equation 1 for clarity of exposition, see Supplemental Note). The prior distribution on β is positive because the probability of $y_i = 1$ increases with $S(\phi, x_i)$, given $\gamma = 1$. The prior variance of β allows for a wide range of effect sizes given the range of S. The priors on the beta CDF transformations are discussed in the Supplemental Note. In brief, the choice of prior for f favors parsimonious characteristic phenotypes and the prior for g allows for an indeterminate number of nodes appearing sporadically among patients.

By default, our prior distribution on the characteristic phenotype ϕ places a uniform prior probability on all minimal sets of terms of size less than or equal to k. We choose $k = 3$ on the grounds that three nodes should adequately distinguish between the primary features of most rare diseases. If y is set based on variants in a particular feature, such as a gene, then our prior can upweight HPO phenotypes comprising terms annotated to that feature on the basis of reports in the scientific literature. Thus, the prior on ϕ is given by

$$\mathbb{P}(\phi) = \begin{cases} \dfrac{1}{|\,\Phi^{(k)}\,|} & \text{No literature phenotype} \\ \dfrac{S'(M \rightarrow \phi)}{\sum_{\psi \in \Phi^{(k)}} S'(M \rightarrow \psi)} & \text{Literature phenotype } M \end{cases} \quad \text{(Equation 3)}$$

where $\Phi^{(k)}$ denotes the set of all minimal sets of up to k HPO terms and S' is an unstandardized similarity function (see Supplemental Note). In practice, the literature phenotype could be obtained from OMIM or from the Mouse Genome Informatics (MGI) database[24] after mapping murine phenotypes coded using the Mammalian Phenotype Ontology[26] to HPO terms through a cross-species phenotype ontology.[27]

Inference

We perform model comparison using the Markov chain Monte Carlo (MCMC)-based method of Carlin and Chib.[28] We sample the model selection parameter γ from its full conditional distribution while the remaining parameters are sampled using the

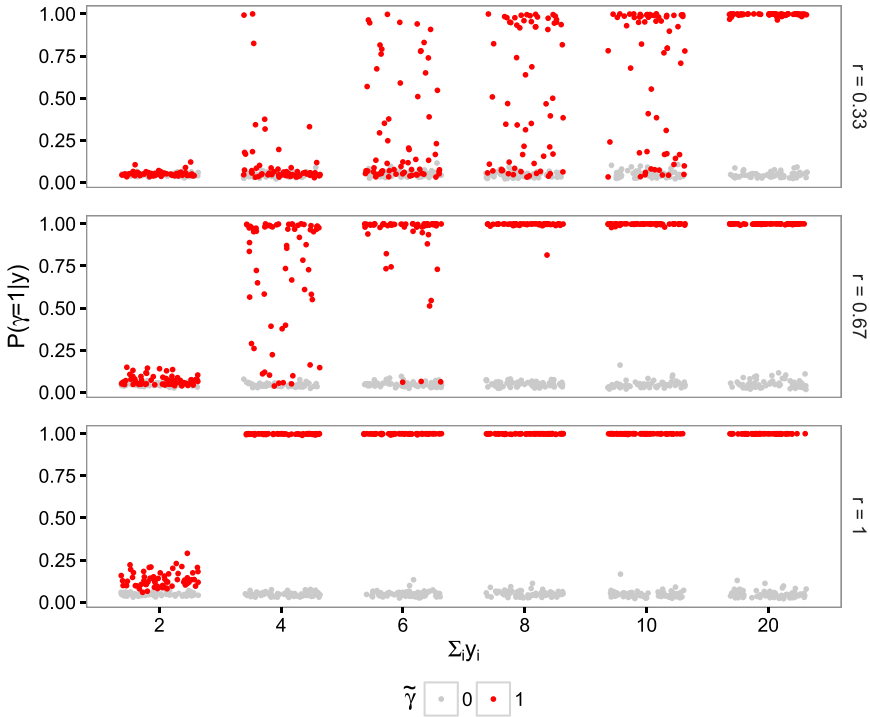

Figure 2. Results of Inference on Simulated Data

Phenotype data were simulated using three levels of expressivity r of the disease terms. The plots within each panel correspond to different frequencies $\sum_i y_i$ of the rare genotype. In each plot, the red dots mark the estimated posterior mean of γ for 64 datasets simulated under $\tilde{\gamma} = 1$ and the gray dots show an equivalent set of estimates for datasets simulated under $\tilde{\gamma} = 0$ (i.e., whereby phenotypes for subjects having $y_i = 1$ are sampled from the same distribution as for subjects having $y_i = 0$). The position of points on the x axis within a plot is arbitrary.

Metropolis-Hastings algorithm or from a pseudoprior distribution, depending on the value of γ at each iteration.

It is not straightforward to sample from the space of minimal sets $\Phi^{(k)}$ when $\gamma = 1$ because not all possible HPO term combinations comprise such a minimal set. To overcome this difficulty, we propose an unrestricted vector of k HPO terms $\tilde{\phi}$ and then derive the associated underlying phenotype ϕ by applying a mapping function υ. We therefore need to impose a prior distribution on the unrestricted space which is compatible with the desired prior for ϕ (Equation 3) on the restricted space. To be precise, the prior on $\tilde{\phi}$ is given by:

$$\mathbb{P}(\tilde{\phi}) = \frac{\mathbb{P}(\upsilon(\tilde{\phi}))}{|\{\tilde{\phi}' \in \mathrm{H}^k : \upsilon(\tilde{\phi}') = \upsilon(\tilde{\phi})\}|},$$

where H^k is the space of all vectors of k HPO terms and υ maps an arbitrary such vector of terms to its corresponding minimal set. The denominator accounts for the number of unrestricted vectors that map to the same minimal set. For further details on the method used to calculate $\mathbb{P}(\phi)$ and $\mathbb{P}(\tilde{\phi})$, the MCMC algorithm, and the tuning of the pseudopriors, refer to the Supplemental Note.

Results

Simulation Study

We assessed the performance of SimReg by analyzing datasets generated under two scenarios, labeled by $\tilde{\gamma}$. Under $\tilde{\gamma} = 1$, the HPO phenotypes $x_{1,\ldots,N}$ were simulated conditional on the genotypes $y_{1,\ldots,N}$ of N individuals whereas under $\tilde{\gamma} = 0$ they were simulated independently of the genotypes. When $\tilde{\gamma} = 1$, phenotypes for all subjects having $y_i = 1$ were formed by selecting terms from an arbitrarily chosen disease template ("Decreased mean platelet volume," "Thrombocytopenia," and "Autism"). Each term was selected with a pre-specified probability r, termed "expressivity," and m further noise terms drawn at random from a set of approximately 1,000 HPO terms were appended, where $m \sim \text{Poisson}(\lambda = 5)$. The set of terms from which the noise terms were drawn was created by selecting 200 HPO terms at random, taking the union with the disease template terms, and then aggregating all the ancestral terms. Phenotypes for subjects having $y_i = 0$ were drawn at random using terms from the above set with $\lambda = 8$ and then mapped to minimal sets. When $\tilde{\gamma} = 0$, all phenotypes were sampled from the noise term set with $\lambda = 8$. This ensures that on average individuals have approximately 8 terms, irrespective of y_i and $\tilde{\gamma}$. The simulation was performed with the set of disease template terms and set of noise terms fixed but with different numbers of individuals carrying the rare genotype ($\sum_i y_i \in \{2, 4, 6, 8, 10, 20\}$ out of $N = 1,000$) and varied levels of expressivity $r \in \{1/3, 2/3, 1\}$. The low expressivity set-ups capture situations in which a fraction of the individuals having a rare genotype can be considered to carry a neutral variant with respect to the disease in question because they have none of the template terms. For the same reason, they capture scenarios of incomplete penetrance of a subset of the underlying rare variants. Furthermore, a degree of genetic heterogeneity is built into our simulation setup, because there is a non-zero probability of a template phenotype term being randomly allocated to an individual with the common genotype.

The results of repeating the simulation 64 times for each value of $\tilde{\gamma}$ and combination of r and $\sum_i y_i$, depicted in Figure 2, show that power to detect a true association, as assessed by the posterior mean of γ, increases with the expressivity of the disease terms r and also with the frequency of the rare genotype in the study sample $\sum_i y_i$ (red dots). Under $\tilde{\gamma} = 0$, the posterior mean of γ remains very close to zero in all circumstances (gray dots). Specifically, we find that 2, 6, and 20 cases out of 1,000 subjects are

Table 1. Studies from which Genetic and Phenotypic Data Were Obtained

Study	Phenotype	Unrelated Subjects	Known Genes
Bleeding and Platelet Disorders (BPD)	detailed patient-specific HPO terms	709	74
Primary ImmunoDeficiency (PID)	Abnormality of the immune system (HP:0002715)	201	131
Pulmonary Arterial Hypertension (PAH)	Pulmonary hypertension (HP:0002092)	422	9
Specialist Pathology Evaluating Exomes in Diagnostics (SPEED)	Retinal dystrophy (HP:0000556)	384	241
	Abnormality of the nervous system (HP:0000707)	215	689
	Abnormality of the nervous system and Retinal dystrophy (HP:0000707, HP:0000556)	7	
	Phenotypic abnormality (HP:0000118)	107	

Note that the SPEED project has a branch dealing with retinal dystrophy and another branch dealing with abnormalities of the nervous system and that 7 individuals are included in both branches. In addition, 107 subjects could not be assigned to a specific sub-project at the time of writing due to lack of information and we assigned them a single abstract HPO term "Phenotypic abnormality" (HP:0000118).

sufficient to obtain perfect or near-perfect discrimination between the two models when the expressivity is 1, 2/3, and 1/3, respectively. When the number of subjects with the rare genotype is equal to 6 and the expressivity is 2/3, which implies that any two individuals with the rare genotype have only a 0.17 chance of having exactly the same template terms, our method can achieve a positive predictive value of 1, even when the negative predictive value is as high as 0.95, by thresholding at $\mathbb{P}(\gamma = 1 \mid y) \geq 0.25$. Under this set-up, we expect 1.78 of the 6 individuals with the rare genotype to have none of the template terms at all, which indicates that the method has some resilience to the presence of $y_i = 1$ induced by neutral rather than pathogenic variants. In order to assess the specificity of the method more accurately, we have simulated 20,000 datasets under the scenario in which there is no association and $\sum_i y_i = 6$ and found that only 7 datasets yield $\mathbb{P}(\gamma = 1 \mid y) > 0.25$, which equates to a specificity of 99.97% for this chosen cut-off (Supplemental Note). We have also extended our simulation study to include a variable controlling genetic heterogeneity, whereby many individuals are drawn from the same template but only a subset have the rare genotype. Power is maintained even in challenging scenarios in which there is substantial genetic heterogeneity and moderate phenotypic expressivity (Supplemental Note). Overall, the results of our simulation study show that our method produces accurate results even in the presence of significant phenotypic or genetic heterogeneity and low expressivity of the rare genotype's characteristic terms. Because these are typical hallmarks of many rare disease studies, our evaluation substantiates the utility of our approach.

Results from Real Data

Our dataset comprises HPO phenotypes and corresponding variant call data for 2,045 unrelated individuals enrolled to a variety of rare-disease sequencing studies (Table 1). Detailed HPO data were available only for subjects enrolled to the Bleeding and Platelet Disorders (BPD) project.[14] BPDs are a heterogeneous group of diseases, including polysymptomatic examples, making them an interesting use-case for the modeling we present. For the other projects, only high-level HPO terms were used (Table 1). A set of genes within which variants are known to be implicated in each class of disorders was provided by BRIDGE collaborators to assess the performance of the model (Supplemental Note).

We used variant call data from 686 sequenced exomes and 1,359 sequenced whole genomes. To account for biases that might alter the baseline rate of rare genotypes (e.g., sequencing platform), we use a plug-in offset in the regression Equation 1, estimated a priori (see Supplemental Note). Variants were retained only if they were predicted to alter protein sequence and were either absent from ExAC (Web Resources) or had an allele frequency therein below 1/1,000 or 1/10,000 when a recessive or dominant mode of inheritance, respectively, was assumed in the analysis. Rare variants were aggregated within genes to account for genetic heterogeneity and increase power. We defined the binary genotypes y based on three different aggregation approaches corresponding to the following hypothetical modes of inheritance: (1) dominant, i.e., presence of at least one rare allele; (2) recessive, i.e., presence of at least two rare alleles; or (3) high-impact dominant, i.e., presence of at least one rare allele predicted[29] to introduce a splice site aberration, frameshift, start loss, or stop gain.

ACTN1 as Exemplar Gene

We now describe the properties of SimReg's output by focusing on a gene, *ACTN1* (MIM: 102575), that has recently been reported to harbor rare variants responsible for reduced platelet number and increased platelet size (macrothrombocytopenia).[30] We note that data for *ACTN1* were used to inform and motivate our choice for the similarity measure given in Equation 2 (Supplemental Note). Once learnt on the *ACTN1* data, this choice has then been used universally for all genes. We observe strong evidence that the rare genotype for *ACTN1* is associated with similarity to a characteristic phenotype ($\mathbb{P}(\gamma = 1 \mid y) = 1$), as expected. The estimated characteristic

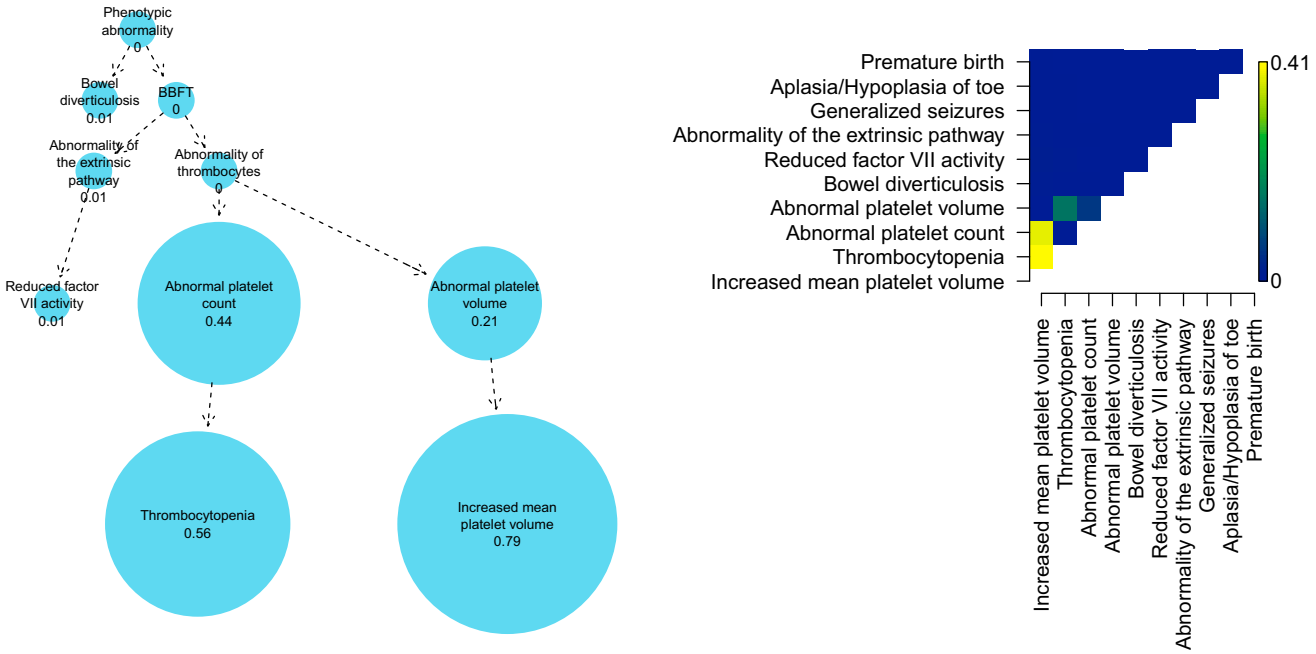

Figure 3. Results for *ACTN1*

The panels show results obtained by applying the SimReg method to phenotype data for all subjects and genotype data for *ACTN1*. There were 43 individuals in our dataset coded with the rare genotype for this gene, of which 22 were coded with "Thrombocytopenia" and "Increased mean platelet volume." The graph shows the estimated probabilities of inclusion of individual terms in ϕ (only the seven terms with the highest probabilities of inclusion and their ancestors are shown). The acronym "BBFT" refers to "Abnormality of blood and blood-forming tissues." The heat map shows the estimated probabilities of pairs of terms co-occurring in ϕ, for pairs composed from the ten most frequently included individual terms.

phenotype focuses primarily on phenotypes that include "Thrombocytopenia" and "Increased mean platelet volume" (Figure 3), which together correspond to macrothrombocytopenia. The slightly more general terms "Abnormal platelet count" and "Abnormal platelet volume" also have substantial marginal posterior weight whereas the rest of the nodes in the HPO have a marginal posterior probability of inclusion less than 0.02. As can be seen in a two-dimensional matrix of the marginal posterior on pairs of terms (Figure 3), there is a high degree of co-occurrence of the two primary terms representing the *ACTN1*-related phenotype, which implies that they are not alternatives but rather complements that together produce a good model fit.

DIAPH1 and *RASGRP2*

Under the high-impact dominant mode of inheritance described above, one of the genes with the highest estimated value of γ that also has a BPD-like inferred phenotype is *DIAPH1* (MIM: 602121) ($\gamma = 0.87$). We recently showed, through an application of our similarity regression approach, that the introduction of a premature stop codon present in two unrelated individuals in the BPD project truncates DIAPH1's 3′ auto-inhibitory domain and causes macrothrombocytopenia, hearing loss, and mild bleeding.[31] As shown in Figure 4 (left), the salient terms in ϕ relate to hearing impairment and abnormality of blood and blood-forming tissues, with the latter driven mainly by thrombocytopenia and bleeding. The high posterior estimate of γ was obtained in part because a sensorineural hearing loss phenotype had previously been reported in the literature,[32] which up-weighted hearing abnormality terms in the prior for ϕ (Table 2). However, even without using an informative prior on ϕ, a high posterior probability of an association ($\gamma = 0.59$) could be found for *DIAPH1*.

RASGRP2 (MIM: 605577) was recently implicated in a new form of Glanzmann's-like thrombasthenia (MIM: 273800) based on data from a single pedigree.[33] Glanzmann's is characterized by impaired platelet aggregation, leading to excessive bleeding. Under a recessive mode of inheritance, our similarity regression successfully detects an association ($\gamma = 0.75$) for *RASGRP2* and estimates a characteristic phenotype concentrated around "Abnormal platelet aggregation" (Figure 4). It is characteristic of Glanzmann's that platelet aggregation is impaired in response to multiple agonists because their common downstream effect—the binding of platelets to fibrinogen—is impeded by the presence of reduced numbers of fibrinogen receptors. Here we also observe this phenomenon but only collagen-induced platelet aggregation carries significant weight in the characteristic phenotype because it is the only specific aggregation term that is shared by all the cases of this recently discovered disorder. There is also a very low probability of inclusion of two rare terms that are not related to the disease—"Atypical scarring of skin" and "Intracranial meningioma"—because of a chance comorbidity in one of the affected cases.

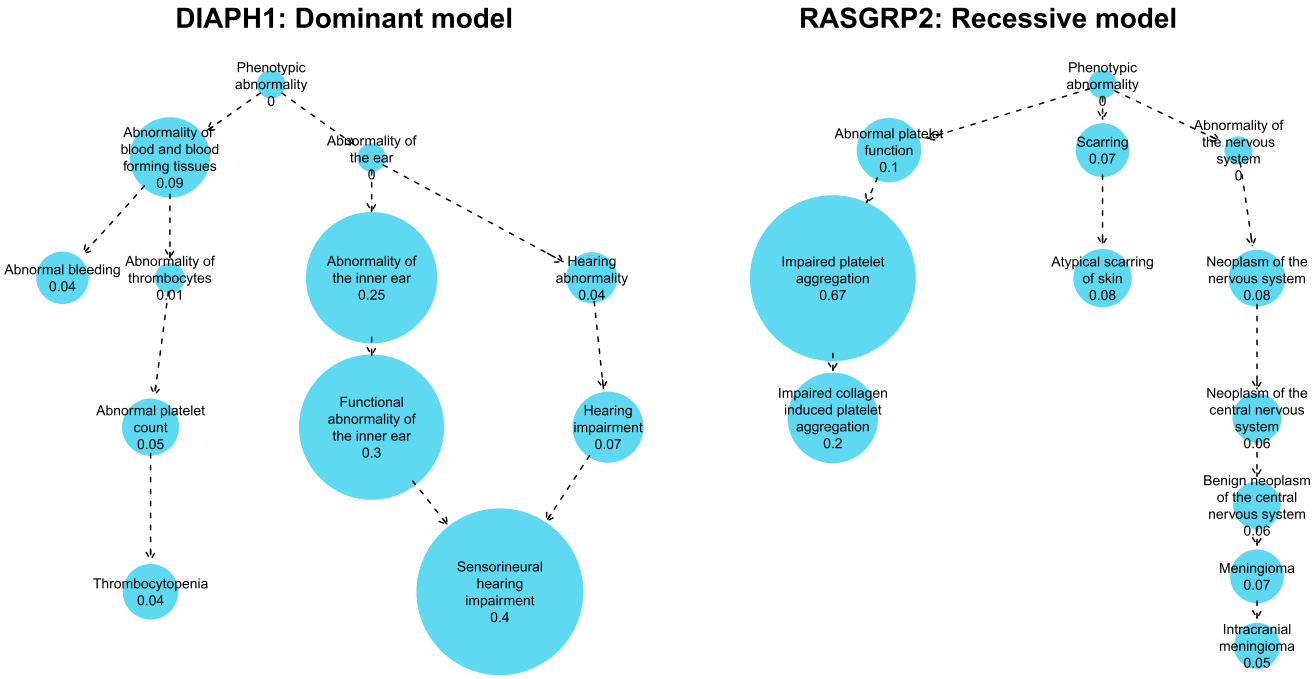

Figure 4. Results for *DIAPH1* and *RASGRP2*
Estimated posterior probabilities of individual terms being included in the characteristic phenotype ϕ using phenotype data for all subjects and variant data for *DIAPH1* ($\sum_i y_i = 2$) encoded under a high-impact dominant model and *RASGRP2* ($\sum_i y_i = 7$) encoded under a recessive model. The ten terms with the highest marginal posterior probability are shown. The estimated posterior probability that $\gamma = 1$ is equal to 0.872 and 0.750 for *DIAPH1* and *RASGRP2*, respectively.

Overall Results

Finally, we turn our attention to the overall results of applying the inference procedure to data for all genes under the three modes of inheritance considered, subject to $\sum_i y_i \geq 2$. In total, we applied the inference to 19,573, 3,134 and 9,733 genes for the dominant, recessive, and high-impact dominant modes of inheritance, respectively. The estimates of $\mathbb{P}(\gamma = 1 \mid y)$ are shown as vertical density plots in Figure 5. For the majority of genes (65%), $\mathbb{P}(\gamma = 1 \mid y) < \mathbb{P}(\gamma = 1) = 0.05$, which implies that no characteristic phenotype can be found that helps distinguish carriers of the rare genotype from other subjects. This result is consistent with the expectation that variants in only a small proportion of genes are implicated in these rare diseases and indicates that specificity is largely controlled.

Strikingly, under all three assumed modes of inheritance, most of the highly confident results (i.e., the genes for which the estimates of $\mathbb{P}(\gamma = 1 \mid y)$ are close to one) are for genes known to be relevant to the pathologies of the patients (indicated by red labels in Figure 5). In all but one case (*KIF1A* [MIM: 601255]), where a gene had $\mathbb{P}(\gamma = 1 \mid y) > 0.25$ and was in one of the projects' set of known genes, a characteristic phenotype similar to the known phenotype was inferred (Table 2). Above a threshold of $\mathbb{P}(\gamma = 1 \mid y) = 0.25$, there was a significant enrichment for known genes (Fisher exact test $p = 2.39 \times 10^{-4}$, 1.98×10^{-4}, and 2.23×10^{-7} for the dominant, recessive, and high-impact dominant modes of inheritance, respectively). Some of the inferred known genes are highlighted more than once across the three modes of inheritance in Figure 5 because there is power to detect the association even when the mode of inheritance is misspecified. For example, *RASGRP2*-related Glanzmann's is recessive, yet $\mathbb{P}(\gamma = 1 \mid y) > 0.25$ even if a high-impact dominant mode of inheritance is assumed.

The black dashes in Figure 5 correspond to unknown genes for which the inferred $\mathbb{P}(\gamma = 1)$ is greater than 0.25, of which there were 8, 1, and 5 found for the dominant, recessive, and high-impact dominant model of inheritance, respectively. These candidates are genes with potentially novel roles in disease and are being actively explored.

Discussion

We have described a method for identifying the genetic determinants of rare diseases that does not require the disease phenotype to be specified a priori. The method uncovers associations between rare genotypes and the similarities between subject phenotypes and a latent characteristic phenotype. Throughout this paper, rare variants have been aggregated within genes according to a hypothesized mode of inheritance in order to define presence or absence of a rare genotype. However, the unit of analysis could be a set of interacting domains or any other arbitrary genomic grouping. During final review of this work, a prioritization

Table 2. Known Genes for which $\mathbb{P}(\gamma=1 \mid y) > 0.25$ and the Inferred Phenotype Was Compatible with the Known Disorder

Gene	MIM No.	Mode of Inheritance	Known Disorder	$\mathbb{P}(\gamma=1 \mid y)$	Highest Marginal Posterior Probability Terms in ϕ
ACTN1	102575	dominant	bleeding and platelet disorder	1.00	increased mean platelet volume (0.79), thrombocytopenia (0.56), platelet count (0.44)
BMPR2	600799	dominant	pulmonary arterial hypertension	1.00	pulmonary hypertension (0.34), elevated pulmonary artery pressure (0.31), pulmonary artery (0.11)
ABCA4	601691	recessive	retinal dystrophy	0.99	retinal dystrophy (0.22), retina (0.22), fundus (0.16)
USH2A	608400	recessive	retinal dystrophy	0.99	retina (0.23), retinal dystrophy (0.2), fundus (0.17)
CRB1	604210	recessive	retinal dystrophy	0.97	retinal dystrophy (0.21), retina (0.18), fundus (0.18)
F11	264900	high-impact dominant	bleeding and platelet disorder	0.95	reduced factor XI activity (0.89), intrinsic pathway (0.11), platelet aggregation (0.07)
RASGRP2	605577	recessive	bleeding and platelet disorder	0.75	platelet aggregation (0.67), collagen-induced platelet aggregation (0.2), platelet function (0.1)
EYS	612424	high-impact dominant	retinal dystrophy	0.70	retinal dystrophy (0.2), retina (0.17), fundus (0.14)
F7	613878	high-impact dominant	bleeding and platelet disorder	0.68	extrinsic pathway (0.5), reduced factor vii activity (0.46), white hair (0.1)
RPGR	312610	high-impact dominant	retinal dystrophy	0.42	retina (0.2), retinal dystrophy (0.17), posterior segment of the eye (0.16)

We display the mode of inheritance under which the association was found, the known disorder, the probability of association, and the top three HPO terms (shown in abbreviated form) in the inferred phenotypes. The marginal posterior probability of inclusion in the characteristic phenotype is shown in brackets next to each term. When an association was found under multiple modes of inheritance, only the true mode is shown. Note that the inferred phenotypes are influenced by prior phenotypic information in the form of OMIM and MGI annotations.

procedure was proposed that combines a standard measure of strength of phenotypic clustering among individuals having two loss-of-function variants in a gene and the probability of the variants appearing in opposite haplotypes in an outbred population.[34] In contrast, our inference procedure is based on statistical principles and the formulation of a model that is flexible with regards to phenotypic expressivity and genetic architecture and robust to noisy clinical coding and moderate genetic heterogeneity. Our Bayesian model naturally accounts for prior evidence of disease phenotypes associated with variants in particular genes by differentially weighting the

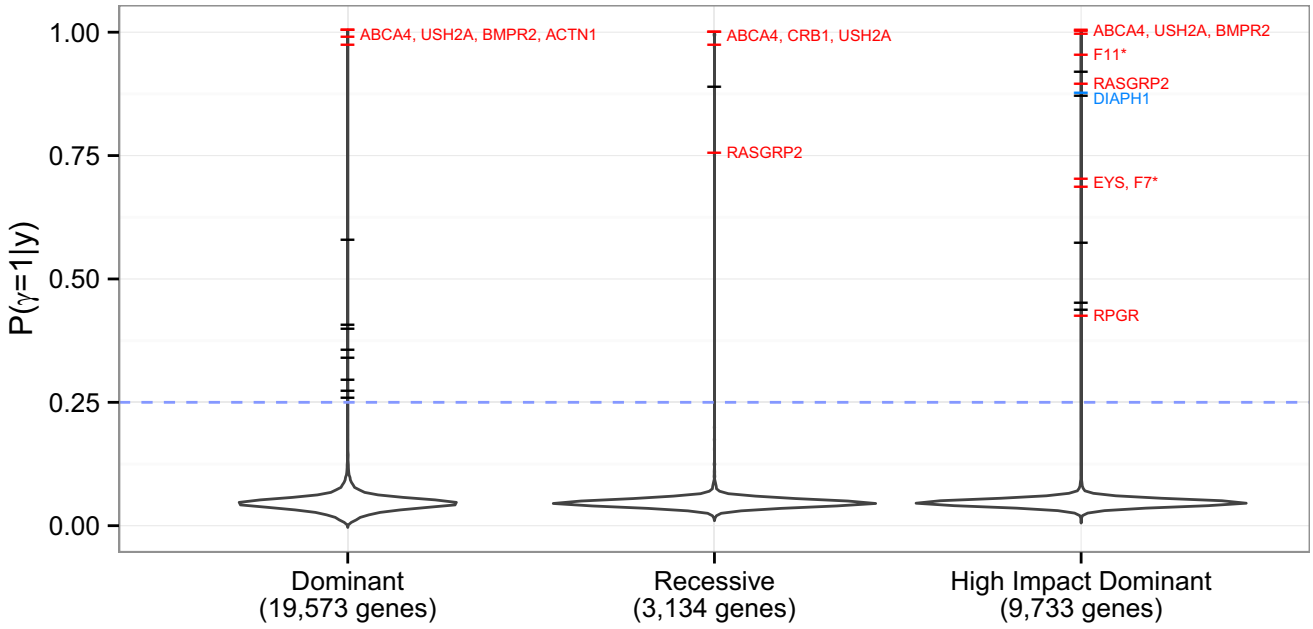

Figure 5. Overall Results
Distributions of the estimated posterior means of γ obtained by applying the SimReg method to each gene under three different modes of inheritance. The tails are truncated at the most extreme values. The dashes indicate values greater than 0.25. The known genes for the BRIDGE project disorders having $\mathbb{P}(\gamma = 1 \mid y) > 0.25$ and a compatible inferred phenotype are labeled and colored in red. An asterisk indicates that a posterior mean of γ greater than 0.25 was estimated only with the use of a prior on ϕ that was informed by the literature of human and murine heritable disorders.

prior probability of inclusion of HPO terms in the characteristic phenotype. Our finding that variants in *DIAPH1* can cause macrothrombocytopenia is an example of how this up-weighting can improve the inference.

The approach we have described is a natural and powerful way of modeling many rare disease phenotypes because it accounts for phenotypic abnormalities across all organ systems encoded with variable precision. Studies of syndromic diseases in particular can benefit from this way of uncovering associations. Our model can also be used for predicting the log odds of the rare genotype using solely phenotype data by means of a function implemented in our SimReg software. This could be used to aid diagnosis by indicating which of a patient's genes should be prioritized for sequencing based on his or her HPO terms. Finally, our regression approach might prove useful for performing inference using notions of similarity between terms in other ontologies where a binary response can be encoded.

Although our method improves significantly on modeling of phenotypic heterogeneity, our treatment of genetic heterogeneity can still be refined, because we currently rely on aggregation of genetic information into single binary variables. In the future we will explore improved modeling of genetic heterogeneity, in which the possibility of a mixture of pathogenic and neutral variants is accounted for explicitly. This would be applicable to genes in which different variants can cause drastically different clinical pathologies (e.g., *LMNA* [MIM: 150330]). Allele frequency, conservation, and functional information could also be used to modulate prior distributions.

In summary, our work represents an advancement in the statistical modeling of ontological heterogeneity that might prove useful at a time in which large collections of deeply phenotyped and sequenced cases are being assembled to uncover hitherto elusive causes of rare heterogeneous diseases.

Supplemental Data

Supplemental Data include Supplemental Note (Diagram Representing the $\gamma = 1$ model; Detailed Model Specification; Estimation of the Offset $\widehat{h}_i$; Prior on f and g; Genetic Heterogeneity; Specificity; Inference using Markov Chain Monte Carlo; Calculation of Prior Probability for ϕ and $\bar{\phi}$; Ethics; SimReg Performance; and Lists of Known Genes for the BRIDGE Projects) and one 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.2016.01.008.

Acknowledgments

This work was supported by NIHR award RG65966 (D.G. and E.T.) and the Medical Research Council programme 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 enrollment centers (see Supplemental Note). We are grateful to Dr. William J. Astle for advice on the statistical model and for providing comments on the manuscript. We are particularly thankful to the BPD project members for granting access to detailed HPO terms of subjects.

Received: August 25, 2015
Accepted: January 8, 2016
Published: February 25, 2016

Web Resources

The URLs for data presented herein are as follows:

Care for Rare, http://care4rare.ca
Comprehensive R Archive Network, http://cran.r-project.org
ExAC Browser, http://exac.broadinstitute.org/
Genomics England, http://www.genomicsengland.co.uk
HPO, http://compbio.charite.de/jenkins/job/hpo/
Mouse Genome Informatics, http://www.informatics.jax.org/
NIHR BioResource – Rare Diseases, https://bioresource.nihr.ac.uk/rare-diseases/
OMIM, http://www.omim.org/
RetNet – Retinal Information Network, https://sph.uth.edu/retnet/home.htm
Undiagnosed Diseases Network, http://www.genome.gov/27550959

References

1. Seri, M., Cusano, R., Gangarossa, S., Caridi, G., Bordo, D., Lo Nigro, C., Ghiggeri, G.M., Ravazzolo, R., Savino, M., Del Vecchio, M., et al.; The May-Hegglin/Fechtner Syndrome Consortium (2000). Mutations in MYH9 result in the May-Hegglin anomaly, and Fechtner and Sebastian syndromes. Nat. Genet. *26*, 103–105.
2. Murayama, S., Akiyama, M., Namba, H., Wada, Y., Ida, H., and Kunishima, S. (2013). Familial cases with *MYH9* disorders caused by *MYH9* S96L mutation. Pediatr. Int. *55*, 102–104.
3. Feng, L., Seymour, A.B., Jiang, S., To, A., Peden, A.A., Novak, E.K., Zhen, L., Rusiniak, M.E., Eicher, E.M., Robinson, M.S., et al. (1999). The β3A subunit gene (*Ap3b1*) of the AP-3 adaptor complex is altered in the mouse hypopigmentation mutant pearl, a model for Hermansky-Pudlak syndrome and night blindness. Hum. Mol. Genet. *8*, 323–330.
4. Anikster, Y., Huizing, M., White, J., Shevchenko, Y.O., Fitzpatrick, D.L., Touchman, J.W., Compton, J.G., Bale, S.J., Swank, R.T., Gahl, W.A., and Toro, J.R. (2001). Mutation of a new gene causes a unique form of Hermansky-Pudlak syndrome in a genetic isolate of central Puerto Rico. Nat. Genet. *28*, 376–380.
5. Suzuki, T., Li, W., Zhang, Q., Karim, A., Novak, E.K., Sviderskaya, E.V., Hill, S.P., Bennett, D.C., Levin, A.V., Nieuwenhuis, H.K., et al. (2002). Hermansky-Pudlak syndrome is caused by mutations in *HPS4*, the human homolog of the mouse light-ear gene. Nat. Genet. *30*, 321–324.
6. Zhang, Q., Zhao, B., Li, W., Oiso, N., Novak, E.K., Rusiniak, M.E., Gautam, R., Chintala, S., O'Brien, E.P., Zhang, Y., et al. (2003). Ru2 and Ru encode mouse orthologs of the genes mutated in human Hermansky-Pudlak syndrome types 5 and 6. Nat. Genet. *33*, 145–153.
7. Morgan, N.V., Pasha, S., Johnson, C.A., Ainsworth, J.R., Eady, R.A., Dawood, B., McKeown, C., Trembath, R.C., Wilde, J., Watson, S.P., and Maher, E.R. (2006). A germline mutation in *BLOC1S3/reduced pigmentation* causes a novel variant of

Hermansky-Pudlak syndrome (HPS8). Am. J. Hum. Genet. *78*, 160–166.
8. Li, W., Zhang, Q., Oiso, N., Novak, E.K., Gautam, R., O'Brien, E.P., Tinsley, C.L., Blake, D.J., Spritz, R.A., Copeland, N.G., et al. (2003). Hermansky-Pudlak syndrome type 7 (HPS-7) results from mutant dysbindin, a member of the biogenesis of lysosome-related organelles complex 1 (BLOC-1). Nat. Genet. *35*, 84–89.
9. Cullinane, A.R., Curry, J.A., Carmona-Rivera, C., Summers, C.G., Ciccone, C., Cardillo, N.D., Dorward, H., Hess, R.A., White, J.G., Adams, D., et al. (2011). A BLOC-1 mutation screen reveals that *PLDN* is mutated in Hermansky-Pudlak Syndrome type 9. Am. J. Hum. Genet. *88*, 778–787.
10. 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.
11. O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C., Elliott, P., Jarvelin, M.R., and Coin, L.J. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE *7*, e34861.
12. Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. PLoS ONE *8*, e65245.
13. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. *42*, D966–D974.
14. Westbury, S.K., Turro, E., Greene, D., Lentaigne, C., Kelly, A.M., Bariana, T.K., Simeoni, I., Pillois, X., Attwood, A., Austin, S., et al.; BRIDGE-BPD Consortium (2015). Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med. *7*, 36.
15. Fitzgerald, T.W., Gerety, S.S., Jones, W.D., van Kogelenberg, M., King, D.A., McRae, J., Morley, K.I., Parthiban, V., Al-Turki, S., Ambridge, K., et al.; Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.
16. Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. Hum. Mutat. *36*, 915–921.
17. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am. J. Hum. Genet. *85*, 457–464.
18. Bauer, S., Köhler, S., Schulz, M.H., and Robinson, P.N. (2012). Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics *28*, 2502–2508.
19. Singleton, M.V., Guthery, S.L., Voelkerding, K.V., Chen, K., Kennedy, B., Margraf, R.L., Durtschi, J., Eilbeck, K., Reese, M.G., Jorde, L.B., et al. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am. J. Hum. Genet. *94*, 599–610.
20. Yang, H., Robinson, P.N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat. Methods *12*, 841–843.
21. Robinson, P.N., Köhler, S., Oellrich, A., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., et al.; Sanger Mouse Genetics Project (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. *24*, 340–348.
22. Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci. Transl. Med. *6*, 252ra123.
23. Javed, A., Agrawal, S., and Ng, P.C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat. Methods *11*, 935–937.
24. Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., and Richardson, J.E.; Mouse Genome Database Group (2014). The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res. *42*, D810–D817.
25. Lin, D. (1998). An information-theoretic definition of similarity. In Shavlik, J.W., ed., Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, WI, USA, July 24-27, 1998. (Morgan Kaufmann) pp. 296–304.
26. Smith, C.L., and Eppig, J.T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip. Rev. Syst. Biol. Med. *1*, 390–399.
27. Köhler, S., Doelken, S.C., Ruef, B.J., Bauer, S., Washington, N., Westerfield, M., Gkoutos, G., Schofield, P., Smedley, D., Lewis, S.E., et al. (2013). Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. F1000Res. *2*, 30.
28. Carlin, B.P., and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. J. R. Stat. Soc., B *57*, 473–484.
29. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.
30. 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.
31. Stritt, S., Nurden, P., Turro, E., Greene, D., Jansen, S.B.G., Westbury, S.K., Petersen, R., Astle, W.J., Marlin, S., Bariana, T.K., et al. (2016). A gain-of-function variant in *DIAPH1* causes dominant macrothrombocytopenia and hearing loss. Blood http://dx.doi.org/10.1182/blood-2015-10-675629.
32. Lynch, E.D., Lee, M.K., Morrow, J.E., Welcsh, P.L., León, P.E., and King, M.C. (1997). Nonsyndromic deafness DFNA1 associated with mutation of a human homolog of the *Drosophila* gene *diaphanous*. Science *278*, 1315–1318.
33. Canault, M., Ghalloussi, D., Grosdidier, C., Guinier, M., Perret, C., Chelghoum, N., Germain, M., Raslova, H., Peiretti, F., Morange, P.E., et al. (2014). Human CalDAG-GEFI gene (*RASGRP2*) mutation affects platelet function and causes severe bleeding. J. Exp. Med. *211*, 1349–1362.
34. Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A.F., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T.W., et al.; DDD study (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. Nat. Genet. *47*, 1363–1369.

The American Journal of Human Genetics, Volume *98*

Supplemental Information

Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases

Daniel Greene, Sylvia Richardson, Ernest Turro, and NIHR BioResource

1 Diagram representing the $\gamma = 1$ model

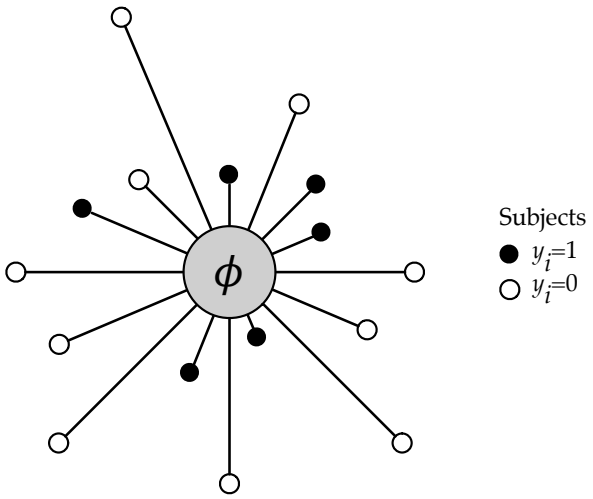

Figure S1: **Cartoon depicting the $\gamma = 1$ model.** Individuals are more likely to carry a rare genotype (indicated by a filled dot) if they are phenotypically similar (as indicated by short edges) to the characteristic phenotype ϕ than if they are dissimilar to it (as indicated by long edges). The angular directions of the edges are purely representational and should not be interpreted. In contrast, under $\gamma = 0$, the rare genotype occurs at a fixed rate irrespective of phenotype.

2 Detailed model specification

The full specification of the two alternative models, $\gamma = 0$ and $\gamma = 1$, described in the main text is given below.

$\gamma = 0$

$$y_i \sim \text{Bernoulli}(p_i),$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \hat{h}_i,$$

with

$$\alpha \sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2),$$

and where $\hat{h}_i$ is an optional plug-in offset parameter (see Section 3).

$\gamma = 1$

$$y_i \sim \text{Bernoulli}(p_i),$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta f\left(S_\phi(\phi \to x_i), a_f, b_f\right) \cdot g\left(S_x(x_i \to \phi), a_g, b_g\right) + \hat{h}_i,$$

with

$$\begin{aligned}
\alpha &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2),\\
\log(\beta) &\sim \text{Normal}(\mu_\beta, \sigma_\beta^2),\\
\log\frac{a_f}{b_f} &\sim \text{Normal}(\mu_f, \sigma_f^2),\\
\log(a_f + b_f) &\sim \text{Normal}(\mu_{f'}, \sigma_{f'}^2),\\
\log\frac{a_g}{b_g} &\sim \text{Normal}(\mu_g, \sigma_g^2),\\
\log(a_g + b_g) &\sim \text{Normal}(\mu_{g'}, \sigma_{g'}^2),\\
\mathbb{P}(\phi) &= \begin{cases} \frac{1}{|\Phi^{(k)}|} & \text{No literature phenotype} \\ \frac{S'(M\to\phi)}{\sum_{\psi\in\Phi^{(k)}} S'(M\to\psi)} & \text{Literature phenotype } M \end{cases}
\end{aligned}$$

where $\Phi^{(k)}$ is the set of all minimal sets of HPO terms of size k. We use the following definitions of S_ϕ, S_x, f, g, S':

$$\begin{aligned}
S_\phi(\phi \to x_i) &= \frac{1}{|\phi|}\sum_{t_\phi\in\phi}\max_{t_x\in x_i} s(t_\phi, t_x)\mathbb{1}_{t_\phi\in\text{anc}(t_x)},\\
S_x(x_i \to \phi) &= \frac{1}{|x_i|}\sum_{t_x\in x_i}\max_{t_\phi\in\phi} s(t_x, t_\phi)\mathbb{1}_{t_\phi\in\text{anc}(t_x)},\\
f(z, a_f, b_f) &= I_z(a_f, b_f),\\
g(z, a_g, b_g) &= I_z(a_g, b_g),\\
S'(M \to \tau) &= \frac{1}{|\tau|}\sum_{t\in\tau}\max_{m\in M}\exp\left(s'(m,t)\right),
\end{aligned}$$

with

$$\begin{aligned}
s'(t_1, t_2) &= \max_{t\in\text{anc}(t_1)\cap\text{anc}(t_2)}\text{IC}(t),\\
s(t_1, t_2) &= \frac{2\times s'(t_1,t_2)}{\text{IC}(t_1)+\text{IC}(t_2)},\\
I_z(a,b) &= \frac{\int_0^z t^{a-1}(1-t)^{b-1}\text{d}t}{\int_0^1 t^{a-1}(1-t)^{b-1}\text{d}t}
\end{aligned}$$

and where $\text{anc}(t)$ is the union of t and all the ancestors of t in the HPO graph and $\text{IC}(t)$ is the information content of term t. Finally, we use the following values for the hyperparameters (see also Section 4):

$$\mu_\alpha = 0,\ \sigma_\alpha^2 = 5,\ \mu_\beta = 2,\ \sigma_\beta^2 = 1,\ \mu_f = 1,\ \sigma_f^2 = 1,\ \mu_{f'} = 2,\ \sigma_{f'}^2 = 1,\ \mu_g = 0,\ \sigma_g^2 = 1.5,\ \mu_{g'} = 2,\ \sigma_{g'}^2 = 1 \quad \text{(S1)}$$

3 Estimation of the offset $\hat{h}_i$

In order to accommodate prior beliefs about the background rate of observing the rare genotype for a particular gene, we obtained point estimates of the effects of gene length and sequencing platform on the log odds of observing the rare genotype. We fitted a generalised linear model linking these variables to the genotype data across all genes for all 2,045 sequenced individuals described in the main text. The model used was:

$$\begin{aligned}
y_{ij} &\sim \text{Bernoulli}(p_{ij}),\\
\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \lambda l_j + \omega^T z_{i\cdot},
\end{aligned}$$

where $y_{ij} = 1$ indicates presence of the rare genotype in gene j for individual i, which occurs with probability p_{ij}, l_j is the length in base pairs of the coding region of gene j and $z_{ik} = 1$ if individual i was sequenced on sequencing platform k and 0 otherwise. Thus, λ is interpretable as the effect size of gene length and $\omega_1, \ldots, \omega_K$ as the effect

sizes of sequencing platforms $1, \ldots, K$. We found that certain sequencing platforms led to gene-specific biases in variant calls. To ensure robustness to these biases, we only used data for genes having a Fisher exact p-value of association between the rare genotype and the sequencing platform greater than 0.05. Under a model of no association, the offset for gene j is given by:

$$y_i \sim \text{Bernoulli}(p_i),$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \hat{h}_i,$$

where $\hat{h}_i = \hat{\lambda} l_j + \hat{\omega}^T z_i$.. The $\hat{h}_i$ was obtained for all genes in all the hypothetical modes of inheritance described in the main text.

4 Prior on f and g

Recall that the overall predictor for the log odds of having a rare genotype under the alternate model is given by

$$f(S_\phi(\phi \to x_i), a_f, b_f) \cdot g(S_x(x_i \to \phi), a_g, b_g).$$

As described in the main text, the presence of a term in the characteristic phenotype ϕ that is absent from the patient phenotype x_i has the effect of lowering S_ϕ, while the presence of a term in x_i that is absent from ϕ has the effect of lowering S_x. For example, if ϕ has one HPO term and it is also present in x_i, then $S_\phi = 1$. However, the presence of one or two additional spurious terms can reduce S_ϕ to as low as 0.5 or 0.33 respectively.

In order to discourage non-parsimonious characteristic phenotypes, we place a high prior weight on f transformations whose corresponding probability density functions have means above 0.5 (i.e. $\frac{a_f}{a_f+b_f} > 0.5$) as this ensures that a good prediction of the log odds cannot be obtained if the absolute value of S_ϕ is low. Specifically, we specify the priors on the parameters of f described in Section 2. The resultant distribution of transformations f and g are represented in Figure S2 (left). However, in order to allow for patients coded with sporadic terms that are not part of the core disease phenotype, we specify a more flexible prior distribution on g than we do on f, as illustrated in Figure S2 (right).

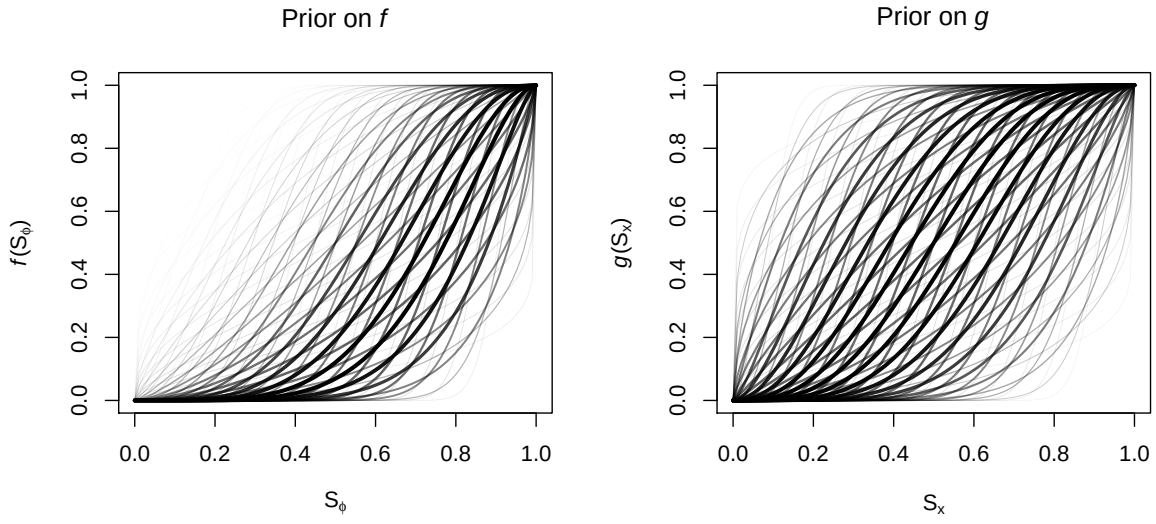

Figure S2: **Prior on f and g.** We show the distribution of shapes for the incomplete beta function transformation of phenotype similarities S_ϕ and S_x for select values of the parameters, given the hyperparameter values in Equation S1. The thickness and opacity of each line is proportional to the prior probability of the corresponding parameterisation of the transformation.

Our choice of hyperparameter values was informed by a sensitivity analysis assessing the model's performance on data for *ACTN1*. We found that not using a transformation at all (i.e. not modulating the similarity with f and

g, which is equivalent to using the identity function obtained by setting $a_f = b_f = a_g = b_g = 1$), or using an overly flexible prior on f, discourages inclusion of the essential 'Thrombocytopenia' term relative to inclusion of spurious alternative terms, conditional on inclusion of the other essential term, 'Increased mean platelet volume'. This occurs because if the value of $\frac{a_f}{a_f+b_f}$ has high posterior weight near 0.5, then spurious terms can be accommodated by mapping values near 0.5 to near 1. As more prior weight is shifted to f transformations with a value of $\frac{a_f}{a_f+b_f}$ greater than 0.5, the probability of joint inclusion of the two key nodes of this disease is increased (Figure S3).

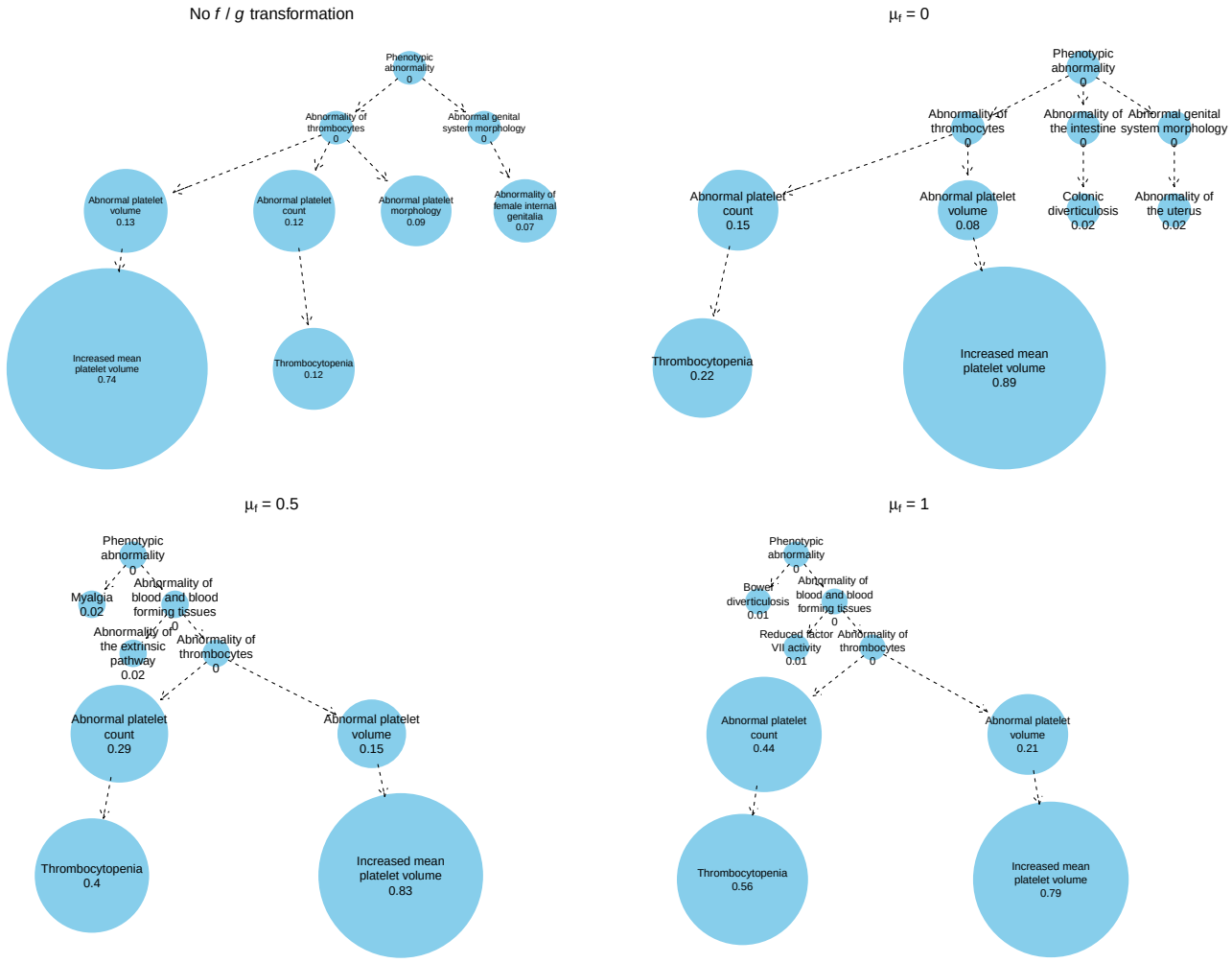

Figure S3: **Inferred ϕ for various parameterisations of the similarity function.** Graphical representation of the posterior distribution of ϕ when no f/g transformations are used and for different values of μ_f (with $\sigma_f^2 = 1, \mu_g = 0, \sigma_g^2 = 1.5$) using the data for *ACTN1*. Each node shows the marginal probability of inclusion in ϕ. Without the f/g transformations, the essential 'Thrombocytopenia' term carries low posterior weight. If the f/g transformations are included, as the value of μ_f is increased, from 0 through 0.5 to 1, the probability of inclusion of the term 'Thrombocytopenia' increases.

Our choice of prior can nevertheless accommodate sporadic absence of disease terms in patients that are part of the characteristic phenotype, provided it can be estimated accurately. Our simulation study (see main text) confirms this because we observe a gradual reduction of the posterior mean value of $\frac{a_f}{a_f+b_f}$ as the expressivity of the terms of the template phenotype for the hypothetical disease phenotype decreases from 1 through $\frac{2}{3}$ to $\frac{1}{3}$ (Figure S4).

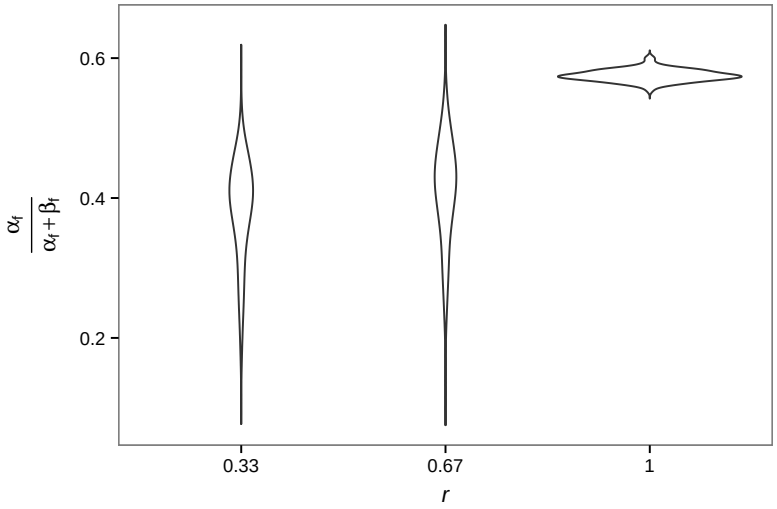

Figure S4: **The posterior mean value of $\frac{a_f}{a_f+b_f}$ for different levels of expressivity.** Distribution of the posterior mean value of $\frac{a_f}{a_f+b_f}$ for three different values of expressivity r. The distributions were obtained over 384 repetitions of our simulation, with $\sum_i y_i = 20$. A decrease in the expressivity, r, of individual terms in the template phenotype results in a decrease in the posterior mean value of $\frac{a_f}{a_f+b_f}$.

5 Genetic heterogeneity

We performed a different version of the simulation study in the main text to assess the performance of our method when genetic heterogeneity is controlled explicitly. Here, we vary a parameter representing the extent of genetic heterogeneity, v, so that for each individual having $y_i = 1$, there were v additional individuals with phenotypes simulated from the same distribution but having $y_i = 0$.

We applied the inference to data sets generated with $v \in \{0, 1, 3, 9\}$. Thus, the simulations where $v = 0$ correspond to the scenario of the simulations described in the main text, and those where $v = 9$ represent situations where only one tenth of the cases having a phenotype arising from the disease template have $y_i = 1$.

The results of the simulation, given as box plots of the estimated posterior means of γ under the various scenarios (Figure S5), demonstrate that although power goes down as genetic heterogeneity increases, the sensitivity of the method, thresholding on $\gamma > 0.25$, approaches 100% when expressivity r is 1 and $\sum_i y_i$ is at least 6, and also when expressivity r is $\frac{2}{3}$ and $\sum_i y_i$ is at least 10, irrespective of v. When $v = 3$ and $r = \frac{1}{3}$, which means the HPO terms have very low expressivity and only a quarter of individuals drawn from the template phenotype carry the rare genotype, γ exceeded 0.25 in 87.5% of our simulations as long as 20 out of $1,000$ individuals carried the genotype. Thus we conclude that our method is powerful even in challenging scenarios in which there is substantial genetic heterogeneity and low phenotypic expressivity.

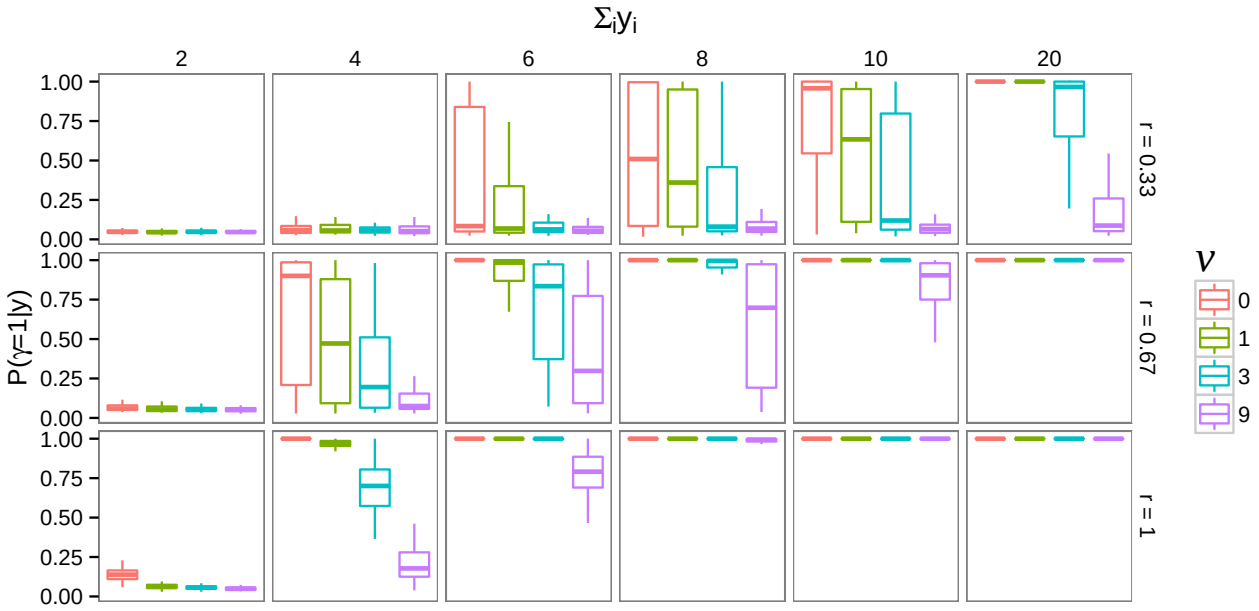

Figure S5: **Relationship between genetic heterogeneity and power.** Box plots showing the distribution of the posterior estimates of γ at various levels of phenotypic expressivity, r, for various sample frequencies of the rare genotype, $\sum_i y_i$ and with different levels of genetic heterogeneity as captured by v. The boxes contain the inter-quartile range, with whiskers extending to the extreme values up to 1.5 times the inter-quartile range from the box.

6 Specificity

The simulation study presented in the main text shows that if the phenotypes are homogeneously selected from a wide range of HPO nodes, then our method is unlikely to produce high posterior estimates of γ. However, the simulation in the main text is based on only 64 repetitions for each simulation set-up (shown as 64 grey dots in each panel). In order to more accurately assess the specificity of our method we simulated 20,000 independent sets of phenotypes, simulated with a total of 6 cases having the rare genotype. The distribution of the posterior mean values of γ inferred for the data sets are shown in the left panel of Figure S6. There were a total of 7 simulated data sets for which the value was greater than 0.25, with the highest estimate being equal to 0.86, which equates to a specificity of 99.97%. The data set for which the highest value was obtained contained four (out of six) individuals with the rare genotype, labelled, 3–6, who had a high mean posterior similarity (> 0.3) to the characteristic phenotype (middle panel of Figure S6). By chance, these four individuals had been assigned highly specific terms relating to bone ossification, the toe and long bone morphology (right panel of Figure S6). This coincidental sharing of HPO terms by these individuals who also carried the rare genotype led to the abnormally high posterior estimate of γ. However, this is a desirable property of our method because in practice it is not possible to know whether such a correlation is causal or spurious.

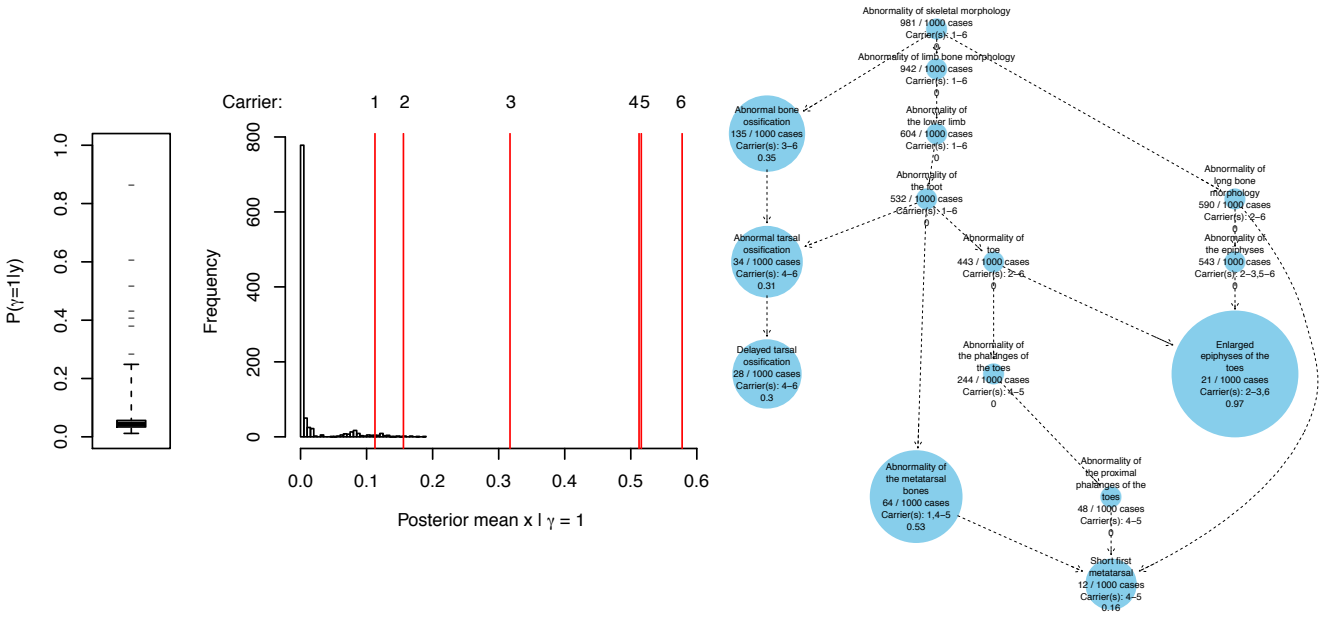

Figure S6: **Evaluation of the specificity of the inference procedure.** The distribution of posterior γ for applications of the inference to the 20,000 repeats of the simulation is shown as a box plot. The box contains the inter-quartile range, with whiskers extending between the lowest posterior γ obtained and 0.25. For the simulated data set for which the highest posterior mean value of γ was inferred, the posterior mean similarities to ϕ, x_i, for the 1,000 simulated patient phenotypes are shown as a histogram, with those of the individuals with the rare genotype, $i|y_i = 1$, marked by red lines. The inferred characteristic phenotype ϕ for this data set is shown as a graph. Each node is labelled with a) the HPO term b) the number of simulated individuals out of 1,000 who had the term c) which individuals with the rare genotype had the term (as labelled in the middle panel) and d) the posterior probability of inclusion in ϕ conditional on $\gamma = 1$ (also represented by node size).

7 Inference using Markov chain Monte Carlo (MCMC)

7.1 Carlin and Chib method

The method of Carlin and Chib is a means of inferring the parameters in two models and computing a Bayes factor comparing them. Instead of targeting the posterior distribution of each model individually, the following function is targeted:

$$(\gamma, \theta^{(0)}, \theta^{(1)}) \mapsto (1-\gamma)L_y^{(0)}(\theta^{(0)})p_0(\theta^{(0)})f_1(\theta^{(1)}) + \gamma L_y^{(1)}(\theta^{(1)})p_1(\theta^{(1)})f_0(\theta^{(0)}).$$

Here, $\theta^{(0)}$ and $\theta^{(1)}$ are vectors of the parameters of models 0 and 1 respectively and $p_0(\theta^{(0)})$ and $p_1(\theta^{(1)})$ are their respective priors. The likelihood functions under model 0 and 1 are given by $L_y^{(0)}(\theta^{(0)})$ and $L_y^{(1)}(\theta^{(1)})$ respectively. The functions $f_0(\theta^{(0)})$ and $f_1(\theta^{(1)})$ are arbitrary probability density functions called 'pseudopriors' representing the conditional probability distributions of the parameters of one model given the alternate model is true, i.e. $\theta^{(0)}|\gamma = 1$ and $\theta^{(1)}|\gamma = 0$ respectively. The conditional posterior distributions of the parameters $\theta^{(0)}|\gamma = 0$ and $\theta^{(1)}|\gamma = 1$ can be estimated from MCMC samples made at iterations when $\gamma = 0$ and $\gamma = 1$, respectively, and the posterior probability that model 1 is true can be estimated from the proportion of iterations in which $\gamma = 1$.

Let α^* be the intercept parameter under $\gamma = 0$ and α be the intercept parameter under $\gamma = 1$ so that they may be distinguished. For convenience, we perform inference of $\tilde{\phi}$, which is on the unrestricted space of vectors of HPO terms, rather than ϕ, because it is difficult to propose uniformly from the space of minimal sets $\Phi^{(k)}$. However, ϕ can be recovered from $\tilde{\phi}$ easily by mapping to the corresponding minimal set. The MCMC algorithm proceeds to target the following distribution:

$$\begin{aligned}
&\mathbb{P}(\gamma, \alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi}|y) \propto \\
&(1-\gamma)L_y^{(0)}(\alpha^*)p_0(\alpha^*)f_1(\alpha)f_1(\beta)f_1(a_f)f_1(b_f)f_1(a_g)f_1(b_g)f_1(\tilde{\phi}) \\
&+\gamma L_y^{(1)}(\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi})p(\alpha)p(\beta)p(a_f)p(b_f)p(a_g)p(b_g)p(\tilde{\phi})f_0(\alpha^*).
\end{aligned}$$

For optimal mixing of the Markov chain, the pseudopriors should approximate the respective conditional posterior distribution given the model, that is, $f_0(\theta^{(0)}) \propto L_y^{(0)}(\theta^{(0)})p_0(\theta^{(0)})$ and $f_1(\theta^{(1)}) \propto L_y^{(1)}(\theta^{(1)})p_1(\theta^{(1)})$. To achieve this, we tune the pseudopriors using empirical summary statistics obtained by running initial Markov chains under each model separately. For parameters α^* and α, Normal pseudopriors are used, while for the strictly positive parameters β, a_f, b_f, a_g and b_g, Log-Normal pseudpriors are used. The hyperparameters of these pseudopriors are obtained using maximum likelihood estimation based on the MCMC samples. We compose a pseudoprior for $\tilde{\phi}$ by counting the number of appearances of HPO terms in any of the k slots of $\tilde{\phi}$ throughout the tuning iterations:

$$\mathbb{P}(t) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{k}\mathbb{1}(\tilde{\phi}_{ij} = t) + \epsilon}{Ik + \epsilon|H|}, \tag{S2}$$

where I is the number of MCMC tuning iterations, t is a term in the set of HPO terms H and $\tilde{\phi}_{ij}$ is the j^{th} element of $\tilde{\phi}$ in the i^{th} iteration. We allow a non-zero probability of inclusion of terms which have not been sampled at all during the tuning batch by setting $\epsilon = 1$. Using the above expression, we define the pseudoprior on $\tilde{\phi}$ as

$$f_1\left(\tilde{\phi}\right) = \prod_{j=1}^{k}\mathbb{P}\left(\tilde{\phi}_j\right).$$

7.2 MCMC updates

Each iteration of the MCMC algorithm comprises the following steps:

1. An update of α^*:

   $\gamma = 0$ Propose an update of α^* by drawing from

   $$\alpha^{*'} \sim \text{Normal}(\alpha^*, s_\alpha^2)$$

   and accepting with probability

   $$\min\left(1, \frac{L_y^{(0)}(\alpha^{*'})p(\alpha^{*'})}{L_y^{(0)}(\alpha^*)p(\alpha^*)}\right).$$

   $\gamma = 1$ Sample $\alpha^{*'}$ from the pseudoprior distribution for α^*:

   $$\alpha^{*'} \sim \text{Normal}(\hat{\mu}_{\alpha^*}, \text{sd} = \hat{\sigma}_{\alpha^*}^2).$$

2. An update of α:

   $\gamma = 0$ Sample α' from the pseudoprior distribution for α:

   $$\alpha' \sim \text{Normal}(\hat{\mu}_\alpha, \hat{\sigma}_\alpha^2).$$

   $\gamma = 1$ Propose an update of α by drawing from

   $$\alpha' \sim \text{Normal}(\alpha, s_\alpha^2)$$

   and accepting with probability

   $$\min\left(1, \frac{L_y^{(1)}(\alpha', \beta, a_f, b_f, a_g, b_g, \tilde{\phi})p(\alpha')}{L_y^{(1)}(\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi})p(\alpha)}\right).$$

3. An update of β:

   $\gamma = 0$ Sample $\log \beta'$ from the pseudoprior distribution for $\log \beta$

   $$\log \beta' \sim \text{Normal}(\hat{\mu}_\beta, \hat{\sigma}^2_\beta).$$

   $\gamma = 1$ Propose an update of $\log \beta$ by drawing from

   $$\log \beta' \sim \text{Normal}(\beta, s^2_\beta)$$

   and accepting with probability

   $$\min\left(1, \frac{L^{(1)}_y(\alpha, \beta', a_f, b_f, a_g, b_g, \tilde{\phi}) p(\beta')}{L^{(1)}_y(\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi}) p(\beta)}\right).$$

4. An update of the shape parameters a_f, b_f, a_g and b_g analagously as is done for β.

5. An update of $\tilde{\phi}$:

   $\gamma = 0$ Sample $\tilde{\phi}$ from the pseudoprior distribution for $\tilde{\phi}$ by sampling all k terms independently from the distribution described in Equation S2.

   $\gamma = 1$ Propose updating $\tilde{\phi}$ to $\tilde{\phi}'$ by setting component $t = \left[\tilde{\phi}\right]_j$ (where j is chosen at random from $1, \ldots, k$) to a random term t', selected with probability π_t. Hence $\tilde{\phi}'$ can be specified as

   $$\left[\tilde{\phi}'\right]_h = \begin{cases} t' & h = j \\ \left[\tilde{\phi}\right]_h & \text{otherwise.} \end{cases}$$

   The proposal is accepted with probability

   $$\min\left(1, \frac{L^{(1)}_y(y|\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi}') p(\tilde{\phi}') \pi_t}{L^{(1)}_y(\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi}) p(\tilde{\phi}) \pi_{t'}}\right).$$

   We set the proposal distribution of the new term $\{\pi_t : t \in \text{H}\}$ to equal that of the individual components of $\tilde{\phi}$ under its pseudoprior (see Equation S2). An alternative approach that does not rely on a tuning chain is to propose a new term proportionally to the number of subjects having $y_i = 1$ whose phenotypes include the term or one of its descendants in the HPO:

   $$\pi_t \propto \sum_{i=1}^{N} \mathbb{1}_{y_i = 1} \mathbb{1}_{t \in \bigcup_{t' \in x_i} \text{anc}(t')}.$$

6. An update of γ by Gibbs sampling:

   $$\gamma' \sim \text{Bernoulli}\left(\frac{\omega^{(1)}}{\omega^{(0)} + \omega^{(1)}}\right),$$

   where

   $$\omega^{(0)} = (1 - \pi)\, L^{(0)}_y(\alpha^*) f_1(\alpha) f_1(\beta) f_1(a_f) f_1(b_f) f_1(a_g) f_1(b_g) f_1(\tilde{\phi}) p(\alpha^*),$$
   $$\omega^{(1)} = \pi\, L^{(1)}_y(\alpha, \beta, a_f, b_f, a_g, b_g, \tilde{\phi}) p(\alpha) p(\beta) p(a_f) p(b_f) p(a_g) p(b_g) p(\tilde{\phi}) f_0(\alpha^*),$$

   where π is the prior probability that $\gamma = 1$.

8 Calculation of prior probability for ϕ and $\tilde{\phi}$

In order to calculate $p(\phi)$ when using a uniform distribution over $\Phi^{(k)}$, we need to calculate the number of distinct minimal sets $\left|\Phi^{(k)}\right|$. This is trivial when $k = 1$, as $\left|\Phi^{(k)}\right| = |\mathrm{H}|$. However it becomes more computationally intensive as k increases, so in our implementation we use the approximation $\binom{|\mathrm{H}|}{k}$. This approximation works well in practice when k is small. It has no effect on the update of the $\tilde{\phi}$ parameter, as the $\left|\Phi^{(k)}\right| = |\mathrm{H}|$ expression cancels out in the acceptance probability for $\tilde{\phi}'$, but it does affect the update of γ as it penalises the model $\gamma = 1$ slightly by overestimating the size of $\left|\Phi^{(k)}\right|$.

When using an informative prior distribution, weighted by similarity to the literature phenotype as described in the main text, we need to calculate $\sum_{\psi \in \Phi^{(k)}} S'(M \to \psi)$. In order to avoid having to sum over the entire space $\Phi^{(k)}$, we employ the approximation $\left|\Phi^{(k)}\right| \times k \times \mathrm{mean}_{\psi \in \mathrm{H}} S'(M \to \psi)$.

Finally, to compute $p(\tilde{\phi})$, we also need to calculate the number of alternative unrestricted vectors that map to the same minimal set, i.e. $\left|\left\{\tilde{\phi}' \in \mathrm{H}^k : \upsilon(\tilde{\phi}') = \upsilon(\tilde{\phi})\right\}\right|$, where ν maps an unrestricted vector to a minimal set. We use the following expression for the number of representations:

$$\left|\bigcup_{t \in \upsilon(\tilde{\phi})} \mathrm{anc}(t)\right|^k + \sum_{i=1}^{|\upsilon(\tilde{\phi})|} (-1)^i \binom{|\upsilon(\tilde{\phi})|}{i} \left(\left|\bigcup_{t \in \upsilon(\tilde{\phi})} \mathrm{anc}(t)\right| - i\right)^k$$

9 Ethics

Table S1 lists the ethics authorities for which the NIHR BioResource – Rare Diseases study has approval. All study procedures were performed after the participants provided informed written consent and were in accordance with the Declaration of Helsinki.

Name of national ethics authority	Ethics approval number	Country
Cambridgeshire 1 Research Ethics Committee	10/H0304/66	United Kingdom
East of England – Cambridge Central	13/EE/0325	United Kingdom
Institut National de La Santé et de la Recherche Médicale	RBM-01-14	France
Sir Charles Gairdner Group Human Research Ethics Committee	2012-095	Australia
Ethics Committee of the University Hospital Leuven	ML3580	Belgium
Ethics Board of the University of Greifswald	n/a	Germany
Ethics Board 2 at Campus Virchow – Klinikum, Charité University Hospital, Berlin	EA2/170/05	Germany
Children's Hospital of Philadelphia Institutional Review Board	IRB#12-008603	USA
Beth Israel Deaconess Medical Center IRB	Protocol #: 2011P000337	USA

Table S1: **Ethics approval information.** Names, ethics approval numbers and countries of ethics authorities approving the NIHR BioResource – Rare Diseases study.

10 SimReg performance

We applied the inference procedure to simulated data sets to assess the performance of SimReg. We varied the number of phenotyped individuals and the number of terms (sampled from a preset collection of approximately 1,000 terms) allocated to each individual, and programmed the algorithm to generate 20,000 MCMC samples (of which 10,000 are tuning iterations). The results of the performance test are shown in Table S2.

N	2 terms	4 terms	8 terms
100	7.31	9.03	11.55
1,000	47.38	65.70	91.74
10,000	451.54	627.75	879.78

Table S2: **Computational performance.** Completion times in seconds for applications of the SimReg procedure. The rows indicate the total number of individuals included in the inference, and the columns indicate the number of HPO phenotype terms allocated to each individual. These results were obtained by running SimReg on a single CPU of a computer with 2.40GHz processors.

11 Lists of known genes for the BRIDGE projects

Genes for which variants are known to underlie a BRIDGE project disorder are listed below:

Bleeding and Platelet Disorders (BPD)
ACTN1, *ANKRD26*, *ANO6*, *AP3B1*, *BLOC1S3*, *BLOC1S6*, *CYCS*, *DTNBP1*, *ETV6*, *F10*, *F11*, *F13A1*, *F13B*, *F2*, *F5*, *F7*, *F8*, *F9*, *FERMT3*, *FGA*, *FGB*, *FGG*, *FLI1*, *FLNA*, *GATA1*, *GFI1B*, *GGCX*, *GNE*, *GP1BA*, *GP1BB*, *GP6*, *GP9*, *HOXA11*, *HPS1*, *HPS3*, *HPS4*, *HPS5*, *HPS6*, *HRG*, *ITGA2B*, *ITGB3*, *LMAN1*, *LYST*, *MCFD2*, *MPL*, *MYH9*, *NBEA*, *NBEAL2*, *ORAI1*, *P2RY12*, *PLA2G4A*, *PLAT*, *PLAU*, *PLG*, *PROC*, *PROS1*, *RASGRP2*, *RBM8A*, *RUNX1*, *SERPINC1*, *SERPIND1*, *SERPINE1*, *SERPINF2*, *STIM1*, *STXBP2*, *TBXA2R*, *TBXAS1*, *THBD*, *THPO*, *VIPAS39*, *VKORC1*, *VPS33B*, *VWF*, *WAS*

Pulmonary Arterial Hypertension (PAH)
ACVRL1, *BMPR2*, *CAV1*, *EIF2AK4*, *ENG*, *KCNK3*, *SMAD1*, *SMAD4*, *SMAD9*

Primary Immune Disorders (PID)
ADA, *AICDA*, *AIRE*, *AK2*, *AP3B1*, *ATM*, *BLM*, *C1QC*, *C2*, *C4B*, *C5*, *C6*, *C7*, *C8A*, *C8B*, *C8G*, *C9*, *CARD11*, *CARD9*, *CASP10*, *CD19*, *CD27*, *CD3D*, *CD3E*, *CD40*, *CFD*, *CFH*, *CFI*, *CFP*, *CHD7*, *CIITA*, *CORO1A*, *CTLA4*, *CXCR4*, *CYBA*, *CYBB*, *DCLRE1C*, *DKC1*, *DNMT3B*, *DOCK8*, *ELANE*, *F12*, *FAS*, *FERMT3*, *FOXP3*, *G6PC3*, *GATA2*, *HAX1*, *IFNGR1*, *IFNGR2*, *IKBKB*, *IKBKG*, *IL10*, *IL10RA*, *IL10RB*, *IL12B*, *IL12RB1*, *IL2RA*, *IL2RG*, *IL7R*, *IRAK4*, *IRF8*, *ISG15*, *ITK*, *JAGN1*, *JAK3*, *KRAS*, *LCK*, *LIG1*, *LIG4*, *LRBA*, *LYST*, *MAGT1*, *MBL2*, *MEFV*, *MPO*, *MRE11A*, *MVK*, *MYD88*, *NBN*, *NCF1*, *NCF2*, *NCF4*, *NFKB2*, *NFKBIA*, *NHEJ1*, *NHP2*, *NLRP3*, *NOP10*, *ORAI1*, *PGM3*, *PIK3CD*, *PNP*, *PRF1*, *PRKCD*, *PSMB8*, *RAB27A*, *RAG1*, *RAG2*, *RBCK1*, *RFX5*, *RFXANK*, *RFXAP*, *RPSA*, *RTEL1*, *SERPING1*, *SH2D1A*, *SLC29A3*, *SMARCAL1*, *STAT1*, *STAT3*, *STAT5B*, *STIM1*, *STK4*, *STX11*, *STXBP2*, *TAP1*, *TAP2*, *TAPBP*, *TBX1*, *TCN2*, *TERT*, *TINF2*, *TNFRSF1A*, *TTC7A*, *UNC13D*, *VPS45*, *WAS*, *XIAP*, *ZAP70*, *ZBTB24*

Specialist Pathology: Evaluating Exomes in Diagnostics (SPEED) - Neurological
AAAS, *ABAT*, *ABCB7*, *ABCC9*, *ABCD1*, *ABHD5*, *ACAD9*, *ACADM*, *ACADS*, *ACAT1*, *ACOX1*, *ACTB*, *ACY1*, *ADCK3*, *ADSL*, *AFF2*, *AFG3L2*, *AGA*, *AGK*, *AGL*, *AKT1*, *ALDH18A1*, *ALDH3A2*, *ALDH4A1*, *ALDH5A1*, *ALDH7A1*, *ALDOA*, *ALDOB*, *ALMS1*, *ALPL*, *ALS2*, *ALX1*, *ALX3*, *AMER1*, *AMPD2*, *AMT*, *ANKRD11*, *ANO3*, *AP4B1*, *AP4E1*, *AP4M1*, *AP4S1*, *AP5Z1*, *APOPT1*, *APTX*, *ARG1*, *ARID1A*, *ARL6*, *ARSA*, *ARSE*, *ARX*, *ASAH1*, *ASL*, *ASPA*, *ASPM*, *ASXL1*, *ATIC*, *ATL1*, *ATM*, *ATN1*, *ATP13A2*, *ATP1A3*, *ATP7A*, *ATP7B*, *ATRX*, *ATXN2*, *ATXN3*, *AUH*, *B3GALT6*, *B4GALNT1*, *B4GALT7*, *BBS1*, *BBS10*, *BBS12*, *BBS2*, *BBS4*, *BBS5*, *BBS7*, *BBS9*, *BCKDHA*, *BCKDHB*, *BCOR*, *BICD2*, *BIN1*, *BLM*, *BMP4*, *BMPER*, *BRAF*, *BRAT1*, *BRCA2*, *BRIP1*, *BRWD3*, *BSCL2*, *BSND*, *BTD*, *BUB1B*, *C12orf65*, *C19orf12*, *C2orf71*, *C5orf42*, *C9orf72*, *CA2*, *CA8*, *CASK*, *CBS*, *CC2D1A*, *CC2D2A*, *CCBE1*, *CCND2*, *CCT5*, *CDC6*, *CDH15*, *CDKL5*, *CDON*, *CDT1*, *CENPJ*, *CEP290*, *CEP41*, *CEP57*, *CHD7*, *CHRNA4*, *CHST14*, *CHST3*, *CHUK*, *CIB2*, *CKAP2L*, *CLN3*, *CLN5*, *CLN6*, *CLN8*, *CNTNAP2*, *COL11A2*, *COL1A1*, *COL2A1*, *COL4A1*, *COL4A2*, *COLEC11*, *COQ9*, *COX10*, *COX15*, *COX6B1*, *COX7B*, *CPS1*, *CRB1*, *CREBBP*, *CSF1R*, *CSTB*, *CTC1*, *CTDP1*, *CTNS*, *CTSA*, *CTSD*, *CUL4B*, *CYP27A1*, *CYP2U1*, *CYP7B1*, *DAG1*, *DARS2*, *DBT*, *DCTN1*, *DCX*, *DDC*, *DDHD1*, *DDHD2*, *DDOST*, *DDR2*, *DDX11*, *DHCR7*, *DHFR*, *DIS3L2*, *DLAT*, *DLD*, *DMD*, *DMPK*, *DNMT3B*, *DOCK8*, *DOLK*, *DPAGT1*, *DPM1*, *DRD2*, *DYM*, *EBP*, *EFNB1*, *EFTUD2*, *EGR2*, *EHMT1*, *EIF2AK3*, *EIF4G1*, *ELAC2*, *ELOVL4*, *EP300*, *EPG5*, *ERCC2*, *ERCC3*, *ERCC4*, *ERCC6*, *ERCC8*, *ERLIN2*, *ESCO2*, *ETHE1*, *EVC*, *EVC2*, *EXOSC3*, *EXT1*, *EYA1*, *EZH2*, *FA2H*, *FAM111A*, *FAM126A*, *FAM20C*, *FANCA*, *FANCC*, *FANCD2*, *FANCE*, *FBN1*, *FBN2*, *FBP1*, *FBXO7*, *FGD1*, *FGD4*, *FGF3*, *FGFR1*, *FGFR2*, *FGFR3*, *FH*, *FIG4*, *FKRP*, *FKTN*, *FLNA*, *FLNB*, *FLVCR1*, *FMR1*, *FOLR1*, *FOXG1*, *FOXRED1*, *FRAS1*, *FREM2*, *FTCD*, *FTL*, *FTSJ1*, *FUCA1*, *GABRA1*, *GABRB3*, *GABRG2*, *GAD1*, *GALC*, *GALE*, *GALT*, *GAMT*, *GATA6*, *GATM*, *GBA*, *GBA2*, *GCDH*, *GCH1*, *GDAP1*, *GFAP*, *GFM1*, *GJA1*, *GJB1*, *GJC2*, *GK*, *GLB1*, *GLDC*, *GLI3*, *GLUD1*, *GLUL*, *GM2A*, *GNAL*, *GNAS*, *GNPAT*, *GNPTAB*, *GNPTG*, *GNS*, *GPR56*, *GRIA3*, *GRIK2*, *GRIN2A*, *GRN*, *GTF2H5*, *GUSB*, *HADH*, *HAX1*, *HCCS*, *HCFC1*, *HDAC4*, *HDAC8*, *HEXA*, *HEXB*, *HGSNAT*, *HK1*, *HOXA1*, *HPRT1*, *HRAS*, *HSD17B10*, *HSD17B4*, *HSPD1*, *HSPG2*, *HTT*, *HUWE1*, *IDS*, *IDUA*, *IFT140*, *IGF1*, *IGF1R*, *IGF2*, *IKBKG*, *IL1RAPL1*, *INPP5E*, *IQSEC2*, *ISPD*, *ITGA7*, *IVD*, *KANSL1*, *KARS*, *KAT6B*, *KBTBD13*, *KCNC3*, *KCNJ10*, *KCNQ2*, *KCNT1*, *KCTD7*, *KDM5C*, *KDM6A*, *KIAA0196*, *KIAA1279*, *KIF11*, *KIF1A*, *KIF1C*, *KIF5A*, *KIF7*, *KIRREL3*, *KMT2A*, *KMT2D*, *KRAS*, *KRIT1*, *L1CAM*, *L2HGDH*, *LAMA2*, *LAMC3*, *LAMP2*, *LARGE*, *LEPRE1*, *LHX3*, *LITAF*, *LMBRD1*, *LMNA*, *LRP2*, *LRP5*, *LRPPRC*, *LRRK2*, *LYST*, *MAN2B1*, *MANBA*, *MAOA*, *MAP2K1*, *MAP2K2*, *MAPT*, *MASP1*, *MC2R*, *MCCC1*, *MCOLN1*, *MCPH1*, *MECP2*, *MED12*, *MEF2C*, *MEGF10*,

MEGF8, *MFSD8*, *MGAT2*, *MGP*, *MID1*, *MITF*, *MKKS*, *MKS1*, *MLC1*, *MMAA*, *MMAB*, *MMACHC*, *MMADHC*, *MNX1*, *MOCS2*, *MPLKIP*, *MPV17*, *MPZ*, *MRE11A*, *MT-ATP6*, *MT-ND4*, *MT-TK*, *MTHFR*, *MTMR2*, *MT-PAP*, *MTR*, *MTRR*, *MTTP*, *MUT*, *MYCN*, *MYH3*, *MYO5A*, *MYO7A*, *NAGA*, *NAGLU*, *NAGS*, *NBN*, *NDE1*, *NDP*, *NDRG1*, *NDUFA1*, *NDUFS1*, *NDUFS4*, *NDUFS7*, *NDUFS8*, *NDUFV1*, *NEFL*, *NEU1*, *NF1*, *NFU1*, *NHS*, *NIPA1*, *NIPBL*, *NKX2-1*, *NKX2-5*, *NPC1*, *NPC2*, *NPHP1*, *NRAS*, *NSD1*, *NSDHL*, *NT5C2*, *NTRK1*, *NUBPL*, *OCRL*, *OFD1*, *OPA3*, *OPHN1*, *ORC1*, *ORC4*, *ORC6*, *OTC*, *OTX2*, *PAFAH1B1*, *PAH*, *PAK3*, *PALB2*, *PANK2*, *PARK2*, *PARK7*, *PAX2*, *PAX6*, *PC*, *PCBD1*, *PCCA*, *PCCB*, *PCDH19*, *PCNT*, *PDCD10*, *PDE4D*, *PDGFB*, *PDGFRB*, *PDHA1*, *PDHX*, *PDSS2*, *PEPD*, *PEX1*, *PEX10*, *PEX12*, *PEX13*, *PEX14*, *PEX16*, *PEX19*, *PEX2*, *PEX26*, *PEX3*, *PEX5*, *PEX6*, *PEX7*, *PGAP1*, *PGK1*, *PHF6*, *PHGDH*, *PIGA*, *PIGL*, *PIGO*, *PIGV*, *PIK3CA*, *PIK3R2*, *PINK1*, *PITX3*, *PLA2G6*, *PLEC*, *PLOD1*, *PLP1*, *PMM2*, *PMP22*, *PNKD*, *PNKP*, *PNPLA6*, *PNPO*, *PNPT1*, *POC1A*, *POLG*, *POMGNT1*, *POMGNT2*, *POMT1*, *POMT2*, *PORCN*, *POU1F1*, *PPP2R2B*, *PQBP1*, *PRKAR1A*, *PRKRA*, *PROP1*, *PRPS1*, *PRRT2*, *PRSS12*, *PRX*, *PSEN1*, *PSMB8*, *PSPH*, *PTCH1*, *PTDSS1*, *PTEN*, *PTPN11*, *PTS*, *PYCR1*, *QDPR*, *RAB23*, *RAB39B*, *RAB3GAP1*, *RAB3GAP2*, *RAD21*, *RAF1*, *RAI1*, *RBM8A*, *RECQL4*, *REEP1*, *REEP2*, *RET*, *RNASEH2A*, *RNASEH2B*, *RNASEH2C*, *RNASET2*, *RNU4ATAC*, *ROGDI*, *ROR2*, *RPGRIP1L*, *RPS6KA3*, *RTN2*, *RYR1*, *SACS*, *SALL1*, *SATB2*, *SBF2*, *SC5D*, *SCN1A*, *SCN1B*, *SCN4A*, *SCN8A*, *SCO1*, *SCO2*, *SDHA*, *SDHAF1*, *SETBP1*, *SF3B4*, *SGCE*, *SGSH*, *SH3TC2*, *SHH*, *SHOC2*, *SIGMAR1*, *SIL1*, *SIX3*, *SKI*, *SLC12A6*, *SLC16A2*, *SLC17A5*, *SLC19A3*, *SLC20A2*, *SLC22A5*, *SLC25A15*, *SLC25A20*, *SLC2A1*, *SLC2A10*, *SLC33A1*, *SLC35C1*, *SLC46A1*, *SLC4A4*, *SLC52A3*, *SLC5A5*, *SLC6A1*, *SLC6A17*, *SLC6A19*, *SLC6A3*, *SLC6A5*, *SLC6A8*, *SLC9A6*, *SLX4*, *SMARCA2*, *SMARCA4*, *SMARCAL1*, *SMARCB1*, *SMC1A*, *SMOC1*, *SMPD1*, *SNCA*, *SOX10*, *SOX2*, *SOX3*, *SPAST*, *SPG11*, *SPG20*, *SPG21*, *SPG7*, *SPR*, *SPRED1*, *SRD5A3*, *STRA6*, *STS*, *STXBP1*, *SUMF1*, *SURF1*, *SYNGAP1*, *SYNJ1*, *SYP*, *TAF1*, *TARDBP*, *TAT*, *TAZ*, *TBC1D24*, *TBCE*, *TBP*, *TBX1*, *TCF4*, *TCOF1*, *TECPR2*, *TFAP2A*, *TFAP2B*, *TFG*, *TGFBR1*, *TH*, *THAP1*, *TIMM8A*, *TMCO1*, *TMEM165*, *TMEM237*, *TMEM67*, *TMEM70*, *TOR1A*, *TP63*, *TPP1*, *TRAPPC9*, *TREX1*, *TRIM32*, *TRIM37*, *TSC1*, *TSC2*, *TSPAN7*, *TTC19*, *TTC8*, *TUBA1A*, *TUBA8*, *TUBB2B*, *TUBB4A*, *TUSC3*, *TWIST1*, *TYR*, *UBE3A*, *UBR1*, *UGT1A1*, *UMPS*, *UPF3B*, *UROC1*, *VAMP1*, *VDR*, *VIPAS39*, *VLDLR*, *VPS35*, *WDPCP*, *WDR45*, *WDR62*, *WNT5A*, *XPA*, *ZBTB20*, *ZC4H2*, *ZDHHC9*, *ZEB2*, *ZFYVE26*, *ZIC2*, *ZNF711*

Specialist Pathology: Evaluating Exomes in Diagnostics (SPEED) - Retinal Dystrophy
ABCA4, *ABCC6*, *ABHD12*, *ACBD5*, *ADAM9*, *ADAMTS18*, *AHI1*, *AIPL1*, *ALMS1*, *ARL2BP*, *ARL6*, *ARMS2*, *ATF6*, *ATXN7*, *BBIP1*, *BBS1*, *BBS10*, *BBS12*, *BBS2*, *BBS4*, *BBS5*, *BBS7*, *BBS9*, *BEST1*, *C12orf65*, *C1QTNF5*, *C2*, *C21orf2*, *C2orf71*, *C3*, *C8orf37*, *CA4*, *CABP4*, *CACNA1F*, *CACNA2D4*, *CAPN5*, *CC2D2A*, *CDH23*, *CDH3*, *CDHR1*, *CEP164*, *CEP250*, *CEP290*, *CERKL*, *CFB*, *CFH*, *CHM*, *CIB2*, *CLN3*, *CLRN1*, *CNGA1*, *CNGA3*, *CNGB1*, *CNGB3*, *CNNM4*, *COL11A1*, *COL2A1*, *COL9A1*, *CRB1*, *CRX*, *CSPP1*, *CYP4V2*, *DFNB31*, *DHDDS*, *DHX38*, *DMD*, *DRAM2*, *DTHD1*, *EFEMP1*, *ELOVL4*, *EMC1*, *ERCC6*, *EYS*, *FAM161A*, *FBLN5*, *FLVCR1*, *FSCN2*, *FZD4*, *GDF6*, *GNAT1*, *GNAT2*, *GNPTG*, *GPR179*, *GRK1*, *GRM6*, *GUCA1A*, *GUCA1B*, *GUCY2D*, *HARS*, *HGSNAT*, *HK1*, *HMCN1*, *HMX1*, *HTRA1*, *IDH3B*, *IFT140*, *IFT172*, *IFT27*, *IMPDH1*, *IMPG1*, *IMPG2*, *INPP5E*, *INVS*, *IQCB1*, *ITM2B*, *JAG1*, *KCNJ13*, *KCNV2*, *KIAA1549*, *KIF11*, *KIZ*, *KLHL7*, *LAMA1*, *LCA5*, *LRAT*, *LRIT3*, *LRP5*, *LZTFL1*, *MAK*, *MERTK*, *MFN2*, *MFRP*, *MKKS*, *MKS1*, *MT-ATP6*, *MTTP*, *MVK*, *MYO7A*, *NDP*, *NEK2*, *NEUROD1*, *NMNAT1*, *NPHP1*, *NPHP3*, *NPHP4*, *NR2E3*, *NR2F1*, *NRL*, *NYX*, *OAT*, *OFD1*, *OPA1*, *OPA3*, *OPN1LW*, *OPN1MW*, *OPN1SW*, *OR2W3*, *OTX2*, *PANK2*, *PAX2*, *PCDH15*, *PCYT1A*, *PDE6A*, *PDE6B*, *PDE6C*, *PDE6G*, *PDE6H*, *PDZD7*, *PEX1*, *PEX2*, *PEX7*, *PGK1*, *PHYH*, *PITPNM3*, *PLA2G5*, *PLK4*, *PNPLA6*, *POC1B*, *PRCD*, *PRDM13*, *PROM1*, *PRPF3*, *PRPF31*, *PRPF4*, *PRPF6*, *PRPF8*, *PRPH2*, *PRPS1*, *RAB28*, *RAX2*, *RB1*, *RBP3*, *RBP4*, *RD3*, *RDH11*, *RDH12*, *RDH5*, *RGR*, *RGS9*, *RGS9BP*, *RHO*, *RIMS1*, *RLBP1*, *ROM1*, *RP1*, *RP1L1*, *RP2*, *RP9*, *RPE65*, *RPGR*, *RPGRIP1*, *RPGRIP1L*, *RS1*, *SAG*, *SDCCAG8*, *SEMA4A*, *SLC24A1*, *SLC7A14*, *SNRNP200*, *SPATA7*, *SPP2*, *TEAD1*, *TIMM8A*, *TIMP3*, *TLR3*, *TLR4*, *TMEM126A*, *TMEM237*, *TOPORS*, *TREX1*, *TRIM32*, *TRPM1*, *TSPAN12*, *TTC8*, *TTLL5*, *TTPA*, *TUB*, *TUBGCP4*, *TUBGCP6*, *TULP1*, *UNC119*, *USH1C*, *USH1G*, *USH2A*, *VCAN*, *WDPCP*, *WDR19*, *WFS1*, *ZNF408*, *ZNF423*, *ZNF513*