Bioinformatics, 34(13), 2018, 2254–2262
doi: 10.1093/bioinformatics/bty079
Advance Access Publication Date: 14 February 2018
Original Paper

OXFORD

Data and text mining

pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion

Ajay Anand Kumar[1,2], Lut Van Laer[1], Maaike Alaerts[1], Amin Ardeshirdavani[3,4], Yves Moreau[3,4], Kris Laukens[2,5], Bart Loeys[1] and Geert Vandeweyer[1,2,*]

[1]Center of Medical Genetics, University of Antwerp and Antwerp University Hospital, Antwerp 2650, Belgium, [2]Biomedical Informatics Research Network Antwerp (biomina), University of Antwerp, Antwerp 2020, Belgium, [3]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven 3001, Belgium, [4]imec, Leuven 3001, Belgium and [5]ADReM Data Laboratory, University of Antwerp, Antwerp 2020, Belgium

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on September 16, 2017; revised on January 25, 2018; editorial decision on February 8, 2018; accepted on February 12, 2018

Abstract

Motivation: Computational gene prioritization can aid in disease gene identification. Here, we propose pBRIT (prioritization using Bayesian Ridge regression and Information Theoretic model), a novel adaptive and scalable prioritization tool, integrating Pubmed abstracts, Gene Ontology, Sequence similarities, Mammalian and Human Phenotype Ontology, Pathway, Interactions, Disease Ontology, Gene Association database and Human Genome Epidemiology database, into the prediction model. We explore and address effects of sparsity and inter-feature dependencies within annotation sources, and the impact of bias towards specific annotations.

Results: pBRIT models feature dependencies and sparsity by an Information-Theoretic (data driven) approach and applies intermediate integration based data fusion. Following the hypothesis that genes underlying similar diseases will share functional and phenotype characteristics, it incorporates Bayesian Ridge regression to learn a linear mapping between functional and phenotype annotations. Genes are prioritized on phenotypic concordance to the training genes. We evaluated pBRIT against nine existing methods, and on over 2000 HPO-gene associations retrieved after construction of pBRIT data sources. We achieve maximum AUC scores ranging from 0.92 to 0.96 against benchmark datasets and of 0.80 against the time-stamped HPO entries, indicating good performance with high sensitivity and specificity. Our model shows stable performance with regard to changes in the underlying annotation data, is fast and scalable for implementation in routine pipelines.

Availability and implementation: http://biomina.be/apps/pbrit/; https://bitbucket.org/medgenua/pbrit.

Contact: geert.vandeweyer@uantwerpen.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Whole-exome sequencing (WES), the current standard approach to identify causal variants in genes underlying human genetic disorders, returns a large number of variants. Databases of known variants such as gnomAD (Lek *et al.*, 2016) provide a powerful first filter. However, identifying the true causal variant often remains a time consuming and challenging task, involving manual evaluation of functional and phenotypical gene information available in literature and biological databases, which is unfeasible without computational tools for larger datasets.

The core principle of computational gene prioritization is to rank candidate genes based on annotation patterns using a discriminatory statistical model. Additionally, these methods can generate hypotheses for novel gene functions. The predictive ability heavily depends on the choice of annotation sources and the technique used to mine the patterns.

Tranchevent *et al.* (2011) presented an overview of existing gene prioritizers classified with respect to integrated annotation sources. Based on the presence or absence of a training set (Moreau and Tranchevent, 2012), these tools are broadly classified as supervised [e.g. Endeavour (Aerts *et al.*, 2006; Tranchevent *et al.*, 2016); ToppGene (Chen *et al.*, 2009)] or unsupervised models [e.g. Biograph (Liekens *et al.*, 2011)].

Next to the learning approach, prioritization results depend on two other aspects: annotation sources can be integrated using early, intermediate and late integration (Pavlidis *et al.*, 2002), and a wide range of statistical methods can be used as the underlying model to rank the genes. Network-based prioritization tools (Lage *et al.*, 2007; Li and Patra, 2010; Kohler *et al.*, 2008; Wu *et al.*, 2008; Zhang *et al.*, 2011), incorporating both protein–protein interaction and phenome networks, are examples of early integration based approaches. Among these, Random Walk with Restart (RWR) gives robust performance with higher predictive accuracy, but it is typically only applicable to single networks and often incorporates only direct neighbourhood information. For multiple networks, Direct Integration of Ranks (DIR) (Chen *et al.*, 2011) and Markov Random Field (MRF) (Chen *et al.*, 2014) were proposed which automatically assign weights to different networks for integration. Recently, a new version the RWR algorithm was proposed that also incorporates multiple heterogeneous networks (RWR-M) (Valdeolivas *et al.*, 2017). Chen *et al.* (2015) proposed a logistic regression based model that utilizes direct and higher-order neighbourhood information in the network for prioritization, together with pathway and expression profiles.

Early integration based approaches can represent topological relationship of entities, but often require complex feature construction during data fusion. In contrast, late integration approaches compute ranks on individual annotation sources and then integrate them to obtain an overall ranking. Rank fusion can become computationally challenging when the number of annotation sources and genes to be prioritized is large. Recently, Zitnik *et al.* (2015) proposed a midway approach, termed intermediate data integration. The main idea is to fuse annotation sources while retaining the overall data structure, thereby capturing internal structures and latent dependencies. Despite the broad range of available methods, most current implementations ignore these internal structural representations (like hierarchical ontologies) and latent inter-feature dependencies during fusion.

It should be noted that updates to annotation sources can eventually alter biological meanings associated with the functionality of any gene. Furthermore, Schnoes *et al.* (2009) pointed out that the advent of next generation sequencing created a large gap between computationally predicted annotations and their experimental validation. For example, three studies (Gillis and Pavlidis, 2013; Groß *et al.*, 2012; Kumar *et al.*, 2013) discussed how changes in the internal directed acyclic graph structure of Gene Ontology (GO) terms over an interval of 10 years can impact subsequent functional analyses. The dynamic nature of biological annotation sources will thus inevitably lead to annotation errors, with a significant potential impact on downstream analysis (Wadi *et al.*, 2016). Although gene-by-gene proximity profiles are at the core of all available prioritization tools, the uncertainty on the proximity scores related to these changes is typically not taken into account, which might impact the prioritization results and lead to less stable ranking.

Another important aspect that should be addressed is the issue of annotation sparsity. Annotation features describing gene functionalites are typically sparsely distributed when considering genome wide data, making feature mining computationally intensive. Moreover, current regression based methods (Wu *et al.*, 2008; Zhang *et al.*, 2011) assume there is no multi-collinear effect of the independent variables (training genes) in the analysis. When multicollinearity is present however, this might lead to inflated values for the regression coefficient estimates, which might in turn lead to over-fitting.

In order to address the above issues, we propose a new computational gene prioritization tool named pBRIT, which applies an Information-Theoretic approach for effective feature mining and Bayesian Ridge Regression (BRR), leading to an intermediate data integration based prioritization model. In this study, we explore the efficiency of text mining methods such as TF-IDF (Term Frequency-Inverse Document Frequency) and latent semantic models (LSM) (Hofmann, 2004) in gene prioritization. We apply TF-IDF for feature extraction and LSM to address sparsity and feature dependencies. Different aspects of pBRIT were evaluated on two separate tasks. First, we compared pBRIT performance with seven existing methods on their original benchmark datasets. Second, we approximated a prospective evaluation using time-stamped benchmark data derived from HPO and compared performance with two additional recent state-of-the-art methods (Endeavour-v3.71 and RWR-M). Finally, we demonstrate the applicability of pBRIT in result visualization and exploration. pBRIT is implemented on a high-performance computing platform, freely available at http://biomina.be/apps/pbrit.

2 Materials and methods

pBRIT offers a three staged gene prioritization, as represented in Figure 1. Unsupervised feature mining, assigning statistical weights to features in the individual annotation sources, is followed by intermediate data fusion. A Bayesian ridge regression model is then built to prioritize candidate genes under a supervised approach. This framework aids in modelling parameter uncertainties arising due to implicit annotation changes or errors.

2.1 Internal representation of annotation sources

We integrated 10 annotation sources, categorized as phenotypic or functional (Fig. 1A). Phenotypic annotations include human phenotype ontology (HPO), HuGe disease navigator (HuGe), the gene association database (GAD) and the disease ontology (DO). For functional annotations, we incorporated Pubmed abstracts, pathway databases, protein–protein interactions (PPI), protein sequence similarities (BLAST), mammalian phenotype ontology (MPO) and gene

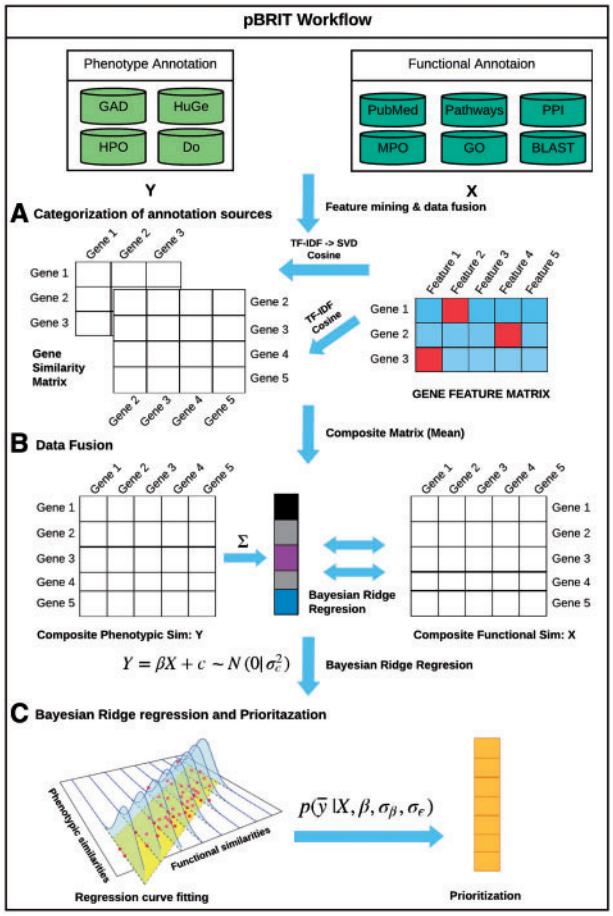

Fig. 1. Schematic workflow of pBRIT. (**A**) Categorization of annotation sources as functional or phenotypic, (**B**) Gene-by-gene proximity profile computation using TF-IDF and TF-IDF→SVD, followed by intermediate data fusion, (**C**) Bayesian ridge regression based candidate gene prioritization

ontology (GO). All annotation sources were downloaded between January 6, 2014 and January 26, 2015 (See Supplementary Section S1 and Table S1.1).

Annotation sources were pre-processed using a generalized version of GOParGenPy (Kumar *et al.*, 2013) to obtain sparse binary matrices with rows representing gene names (mapped to Ensembl ids) and columns representing specific annotation features (Supplementary Fig. S1.3). Entries of 0 and 1 represent feature absence and presence, respectively. For PubMed abstracts, the entries were generalized to the number of feature occurrences per abstract. One exception to the sparse representation was BLAST, for which normalized bit scores from pairwise sequence alignment of all human proteins (available from Uniprot) were used as similarity scores. The matrix is treated as a full matrix (Supplementary Table S1.1).

2.2 Information-theoretic model for feature mining

We computed TF-IDF-based statistical weights for features in the sparse annotation matrices (Equation 1). TF-IDF is based on the relevance and frequency of feature occurrences in the corpus. Features that are less frequent indirectly imply an annotation specific to a gene.

$$\begin{aligned} \mathbf{TF}(f,g) &= 1 + \mathbf{Log}(\mathbf{tf}_{feature,gene}) \\ \mathbf{IDF}(f,G) &= \mathbf{Log}\left(\frac{|G|}{1+|\{g \epsilon G : f \epsilon g\}|}\right) \\ \mathbf{W}(f,G) &= \mathbf{TF} \times \mathbf{IDF} \end{aligned} \tag{1}$$

For all sources except PubMed, the term frequency (*tf*) is equal to one due to the binary data format. IDF(*f*, *G*), or inverse document frequency, denotes the inverse frequency of a particular feature (*f*) across all genes (*G*). Hence, it describes the specificity of a feature. W(*f*, *g*) gives the statistical weight of feature (*f*) for a given gene (*g*). Using TF-IDF, specific features get higher weights, contributing more to the final similarity score used in ranking.

2.3 Modelling feature interdependencies and sparsity

Singular Value Decomposition (SVD) is a matrix factorization technique that reduces the sparsity and can model co-occurrences of the feature concepts (Hofmann, 2004). Through SVD, high dimensional matrices are transformed to a lower dimension, where each original row and column can be represented as a linear combination of latent concepts in the new singular vector space. This linear combination of latent concepts indirectly models any co-occurring or semantically related features. The final number of vectors (*k*) defines both the complexity of the model and the accuracy of representing the original feature space.

Using SVD, each annotation matrix was decomposed in *k* singular values and then projected in those directions. The optimal choice of *k* corresponds to a maximal preservation of variance in the data. Mathematically, this can be expressed as:

$$A_{m\times n} \approx U_{m\times k} D_{k\times k} V_{k\times n}; \tilde{A}_{m\times k} \approx A_{m\times n} V^T_{k\times n} \tag{2}$$

Where, *U* is an $m \times k$ unitary matrix with *k* columns as left singular vectors. *V* is a $k \times n$ unitary matrix with *k* rows as right singular vectors. *D* is a $k \times k$ diagonal matrix holding *k* singular values.

Supplementary Table S1.1 presents the average number of non-zero features per gene in each annotation source used in pBRIT, which ranges from 236 (Pubmed) to 10 (GAD). From Supplementary Figure S1.2, it can be seen that a uniform proportion of variance is explained for all sources with *k* set to 200. Hence, we generalized the choice of *k* equal to 200 for all TF-IDF weighted matrices. Gene-by-gene proximity profiles were obtained using cosine similarity on both TF-IDF and SVD transformed TF-IDF matrices, represented throughout the text as TF-IDF and TF-IDF→SVD, respectively.

2.4 Data fusion

In order to perform Bayesian ridge regression, we compute the composite matrices for the independent and dependent variables in the regression model by averaging the gene-by-gene proximity profiles:

$$X_{composite} = \frac{\sum_f^F X_f}{F}; Y_{composite} = \frac{\sum_p^P Y_p}{P} \tag{3}$$

where, *F* and *P* denote total number of functional and phenotypic annotation sources, respectively. X_f and Y_p represent gene-by-gene proximity profiles for all *f* functional and *p* phenotypic annotations sources, following Equations (1) and (2).

2.5 Prioritization using Bayesian ridge regression model

pBRIT implements the underlying hypothesis that the biological function of a gene is correlated to phenotypic characteristics presented by deregulation of that gene. Mathematically, this can be

formulated by a regression between functional and phenotypic annotations. However, the parameters of such a regression are intrinsically affected by uncertainties in the model arising due to incomplete annotations and changes in the annotation corpus. Regression under a Bayesian framework can model these uncertainties while learning the linear mapping between functional and phenotypic annotation sources. Specifically, we want to model the mean of conditional $E(Y|X)$, i.e. the expected distribution of phenotype similarities given the functional annotation information. This is represented by $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$. For any given n training genes and m test genes which needed to be prioritized, we extract the respective composite matrices for both functional and phenotypic annotations using Equation (3).

The response, or dependent variable vector of the regression model is obtained by $Y_{(n+m)\times 1} = \sum_{j=1}^{n} y_{ij}$. The independent, or predictor variables are the gene-by-gene proximity profiles with respect to n training genes, forming the design matrix $X_{(n+m)\times n}$. The overall regression model is thus given by:

$$\mathbf{Y}_{(n+m)\times 1} = \beta \mathbf{X}_{(n+m)\times n} + \varepsilon; \text{ where, error term } \varepsilon \sim N\left(0, \sigma_\varepsilon^2\right) \quad (4)$$

The unknowns, the regression coefficient β, its corresponding variance σ_β^2 and the residual variance σ_ε^2 can be estimated uniquely from the above regression settings. The regression model of pBRIT uses proximity profiles of both training and test genes in the design matrix. The relatedness of the selected training genes gives a high likelihood of dependencies among the predictor variables. Sometimes, this leads to over-fitting and multi-collinearity of the regression model. Ultimately, multi-collinearity of the predictor variables can lead to inaccurate estimation of regression coefficients, inflated standard error estimates and degradation of model predictability. In order to overcome these problems, we propose a Bayesian ridge regression model. We regularize the estimates by adding a parameter $\tilde{\lambda}$ which is given by the ratio of $\frac{\sigma_\varepsilon^2}{\sigma_\beta^2}$. As the σ_β^2 increases to larger values the solution to find optimal $\widehat{\beta}$ approximates ordinary least squares estimates. Requirements for the optimal choice of $\widehat{\beta}$ are given by:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n+m} \left(y_i - x_i{}^T \boldsymbol{\beta}\right)^2 + \tilde{\lambda} \sum_{j=1}^{n} \beta_j{}^2 \right\} \quad (5)$$

$$E(\boldsymbol{\beta}|y) = \widehat{\boldsymbol{\beta}} = [\mathbf{X}^T\mathbf{X} + \tilde{\lambda}\mathbf{I}]^{-1}\mathbf{X}^T\mathbf{y} \quad (6)$$

In Bayesian setting the likelihood of the model is given by:

$$\text{Likelihood}: \quad p\left(\mathbf{y}|\beta, \sigma_\varepsilon{}^2\right) = \prod_{i=1}^{n+m} N\left[y_i | \sum_{j=1}^{n} x_{ij}\beta_j, \sigma_\varepsilon{}^2\right] \quad (7)$$

$$\text{Prior}: \quad p\left(\beta|\sigma_\beta{}^2\right) = \prod_{i=1}^{n} N\left(\beta_i|0, \sigma_\beta{}^2\right) \quad (8)$$

$$p\left(\sigma_\beta^2\right) = \chi^{-2}\left(\sigma_\beta^2|df_\beta, S_\beta\right) \quad (9)$$

$$p\left(\sigma_\varepsilon^2\right) = \chi^{-2}\left(\sigma_\varepsilon^2|df_\varepsilon, S_\varepsilon\right) \quad (10)$$

We assume NIG (Normal Inverse-Gamma) density priors on unknown regression parameters. The joint posterior distribution of the vector of unknowns, represented by $\theta\epsilon\left(\beta, \sigma_\beta^2, \sigma_\varepsilon^2\right)$ in the model, is proportional to the product of the likelihood and the prior distribution, given by:

$$p(\theta|y) \propto \underbrace{\prod_{i=1}^{n+m} N\left(y_i|\sum_{j=1}^{n} x_{ij}\beta_j\right)}_{\text{Likelihood}} \times \underbrace{\prod_{i=1}^{n} N\left(\beta_i|0, \sigma_\beta^2\right)\chi^{-2}\left(\sigma_\beta^2|df_\beta, S_\beta\right)}_{\text{Prior on } \beta} \times \underbrace{\chi^{-2}\left(\sigma_\varepsilon^2|df_\varepsilon, S_\varepsilon\right)}_{\text{Prior on } \varepsilon} \quad (11)$$

Since the posterior distribution does not have a closed form, a Gibbs sampler was used. Regression analysis was performed using an adapted version of the BLR package (de los Campos *et al.*, 2013) in R. Once the parameters are estimated, the corresponding phenotype concordance score y_{pred} can be predicted by:

$$E\left(X\boldsymbol{\beta}|y, \sigma_\varepsilon^2, \sigma_\beta^2\right) = \mathbf{X}E\left(\boldsymbol{\beta}|y, \sigma_\varepsilon^2, \sigma_\beta^2\right) \quad (12)$$

$$y_{pred} = E\left(X\boldsymbol{\beta}|y, \sigma_\varepsilon^2, \sigma_\beta^2\right) = \mathbf{X}[\mathbf{X}^T\mathbf{X} + \tilde{\lambda}\mathbf{I}]^{-1}\mathbf{X}^T\mathbf{y} \quad (13)$$

Prior to regression, the dependent variable Y and independent variable X were transformed by taking the square root of their values, in order to reduce any non-linearity effects. We follow the BLR guidelines for initializing the priors (de los Campos *et al.*, 2013). The prior on residual variance is indicated by two parameters: Scale, S_ε and degree of freedom, df_ε. The prior variance of the residuals is given by V_ε which is assigned as the variance of the phenotypic concordance score of the training genes. Together, they can be expressed as: $S_\varepsilon = V_{\varepsilon(\text{Train})}(df_\varepsilon + 2)$. Similarly, the prior on the regression coefficient can be expressed as: $S_\beta = \frac{\text{Var}(\text{Y}_{\text{Train}}) \times \left(df_\beta + 2\right)}{\sum_i^n \text{Var}\left(\text{X}_{\text{Train}j}\right)}$. In this study, we chose $df_\varepsilon = df_\beta = 3$. For the Gibbs sampling we chose a total number of iterations of 100 000, a burn-in period of 30 000 and a thinning parameter of 10. The algorithmic details can be found in Supplementary Section S2.

2.6 Cross-validation strategy

The overall performance of pBRIT was evaluated by performing leave one-out cross-validation (LO-OCV) on several benchmark sets. For a given disease, with n known associated genes, we trained our model with $n - 1$ genes and placed the query gene (known gene whose ranking is to be determined) in a list of 99 Test genes randomly selected across the genome. We removed direct contribution of known phenotypic associations of the query gene to the remaining training genes during validation experiments by setting all proximity scores to 'Na' (indicating phenotype information 'Non available'). Hence, the model purely predicts the phenotype concordance score of the query gene, without bias to prior knowledge (See Supplementary Section S2 for details).

We explored the effect of the regression model design on the prediction efficiency in two cases. In the Test.N.Na case (Fig. 2A), the known phenotypic associations of all 99 test genes were taken into account in the regression model, discarding only the known associations of the nth query gene. In contrast, in the Test.ALL.Na case (Fig. 2B), the phenotypic association of all the test genes, along with the query gene, is discarded. Both Test.N.Na and Test.ALL.Na were then combined with either TF-IDF or TF-IDF→SVD based proximity profiles to evaluate the effect of the feature extraction methodology, leading to four analysis scenarios in total. The TF-IDF→SVD_Test.N.Na scenario, reflecting all pBRIT functionality, is referenced as the full pBRIT model hereafter. (See Supplementary Section S2 for algorithmic details).

LO-OCV analysis yields ranks of all the training genes per studied disease. Query gene ranks were normalized to rank-ratios by

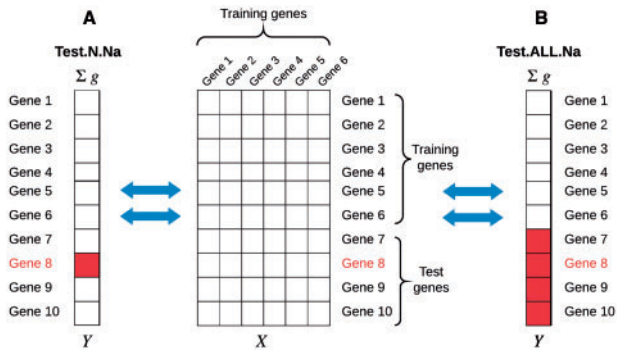

Fig. 2. Bayesian ridge regression. The design matrix (X) contains similarity scores of both training and test genes to training genes. The phenotypic concordance score vector is indicated by Y. For LO-OCV, the summed phenotypic score of the *n*th query gene (**A**. Test.N.Na) or all test genes (**B**. Test.ALL.NA), corresponding to prior phenotypic knowledge, is removed (colored box) during regression parameter estimation (Color version of this figure is available at *Bioinformatics* online.)

dividing them with the total number of test genes (typically $n = 100$) and evaluated by two criteria. First, the mean rank ratio (MRR) of all training genes for a given disease was calculated. The MRR is computed by taking average of rank ratios per disease class and is a metric of efficiency, estimating how many candidates a user must review before the true positive candidate is encountered. Second, the Area Under the Curve (AUC), which measures the prediction accuracy of the model, was obtained from plotting the Receiver Operation Characteristic (ROC) curves. ROC curve analysis measures the trade-off between True positive rate (TPR, sensitivity) and False positive rate (1-specificity). The sensitivity is measured as the percentage of query genes that were ranked above a given threshold. The specificity is defined as the percentage of randomly selected test genes ranked below the threshold (Aerts *et al.*, 2006). Performance differences were evaluated by a two-sided paired Wilcoxon signed-rank test. (For details see Supplementary Section S8). Additionally, we performed a control experiment on the DisGeNET data, replacing the query gene by a random gene not associated with any given *UMLS* class during LO-OCV (Supplementary Material S6: sheet 7 and Section S4.4).

2.7 Validation datasets

As a first benchmark dataset, we obtained 1154 genes associated to 12 disease classes (Goh *et al.*, 2007) used to validate previous prioritization tools (Chen *et al.*, 2015) (Supplementary Material S1: sheet 6–7). The dataset is referenced throughout the text as the Goh *et al.* dataset. Included disease classes are Cardiovascular, Connective tissue, Dermatological, Development, Endocrine, Hematological, Immunological, Metabolic, Muscular, Ophthamalogical, Renal and Skeletal. On average, 100 training genes were available per disease class.

A second benchmark dataset was obtained from the authors of HyDRA (Kim *et al.*, 2015). It consists of eight diseases: Autism, Breast cancer, Colorectal cancer, Endometriosis, Ischaemic stroke, Leukemia, Lymphoma and Osteoarthritis (Supplementary Material S2: sheet 6) and was previously used to evaluate performance of HyDRA against Endeavour and ToppGene. ToppGene and Endeavour are supervised prioritization methods fusing 18 and 20 annotation sources, respectively. In this study, we considered only scores obtained by the respective full annotation models.

Third, we extracted 9414 curated genes, associated with 779 *UMLS* coded diseases from DisGeNET (Pinero *et al.*, 2015) (Supplementary Material S6: sheet 5). Within DisGeNET, we considered only diseases with 4–51 associated genes, resulting in a minimum of 3 and a maximum of 50 training genes during LO-OCV.

Finally, we simulated a prospective benchmark dataset, derived from HPO. For this, we extracted 2025 HPO terms with 2484 novel unique gene-phenotype associations added between January 2015 and February 2017 (Supplementary Material S7: sheet 1–2). For each selected HPO term, we extracted associated genes from the January 2015 release as training genes and performed genome wide prioritization of the novel gene. Similar to the four LO-OCV scenarios, we performed prioritization with and without inclusion of phenotype data from the test genes (labeled Test.Pheno.Include and Test.Pheno.Discard, respectively). Additionally, we extracted a subset of 693 HPO terms having 1111 unique gene associations to evaluate performance of pBRIT in Test.Pheno.Include mode to Endeavour-v3.71 (with usage of 24 annotation sources) and RWR-M (built with four annotation sources).

2.8 Implementation of pBRIT

Generation of sparse annotation matrices was done in python using a customized version of GOParGenPy (Kumar *et al.*, 2013). TF-IDF and TF-IDF→SVD computation was done in R using the 'snow' (Tierney *et al.*, 2009) package to parallellize processing and 'irlba' (Baglama and Reichel, 2012) for TF-IDF→SVD computation. The web interface was developed using PHP as front-end and MySQL as back-end, connected to a torque/pbs job manager for prioritization job execution on a high-performance computing cluster.

3 Results

pBRIT was benchmarked against a set of published datasets. The individual datasets were chosen to range from very broad disease categories (Goh *et al.*, HyDRA), often with well known causative genes, to very specific diseases with a minimal number of known involved genes (DisGeNET, HPO). As such, the benchmark data represent an increasingly challenging validation trajectory. Similarly, competing methods were selected to either allow objective comparison on the respective benchmark data (Goh *et al.*, HyDRA), or to represent alternative state of the art methodologies in real life scenario's (Endeavour-v3.71; RWR-M). pBRIT is available as a web-interface and using a command line interface (batch mode). Prioritization of 100 test genes using 30 training genes takes on average 47.8 s using the web-interface. However, using the command line interface, prioritizing 10 similar sets of 100 test genes took approximately 83 s in total. Afterwards, results can be visualized using the web-interface.

3.1 BRR and SVD allows accurate and stable prioritization

LO-OCV on the Goh *et al.* data showed that most of the query genes were ranked among the top 15% highest scoring test genes, with a minimum AUC score of 0.86, under all four analysis scenarios (Supplementary Table S4.1 and Fig. S3A). Despite the broad disease classes and large amount of training genes per disease class, these results already highlight the relevance of different aspects of the pBRIT methodology. First, considering phenotype association scores of random test genes during regression improves AUC scores. This can be seen by comparing Test.N.Na and Test.ALL.Na scenarios, showing effects up to 7%, accompanied by an improvement in MRR from 0.148 to 0.075 (P value = 3.3E-61, Supplementary Material S1: sheet 5). Second, singular value decomposition on the gene-by-feature profiles yields a better resolution of the similarity

profiles, reflected in the slight improvement of AUC and MRR values over all disease classes when changing from TF-IDF to TF-IDF→SVD-based feature extraction. Although the impact of SVD on the final prioritization results is rather limited, the difference is significant (P value = 4.86E-10, Supplementary Material S1: sheet 5). Furthermore, the higher gene-by-feature resolution will also help in the interpretation of the results (see Section 3.6). The dataset was already applied to benchmark four other methods, all applying early or intermediate data integration (Chen *et al.*, 2015). These methods were a) logistic regression based fast F_3PC algorithm b) Markov random field (MRF) c) Random walk with Restart (RWR) based network integration and d) Direct integration ranking (DIR) algorithm. The previously reported maximum AUC score on this dataset was 0.83, achieved by F_3PC. For MRF, RWR and DIR, the AUC scores were 0.731, 0.711 and 0.716, respectively. In our analysis, pBRIT performs better under all scenarios, with a maximum AUC score of 0.94 using the full model (TF-IDF→SVD_Test.N.Na).

Additionally to higher overall AUC scores, they show a lower variance over the individual disease classes compared to the competing methods (Supplementary Fig. S3A and B). The global AUC score standard deviation of 0.015 under the full model indicates that pBRIT is not biased towards specific medical domains. In contrast, the F_3PC algorithm, being the best performing overall method, showed a maximum AUC score of 0.92 under the immunological disease class and a minimum AUC score of 0.68 under the developmental disease class, whereas pBRIT reaches AUC scores of 0.95 and 0.94 for these classes, respectively.

3.2 Intermediate fusion provides uniform prioritization

Subsequently, we wanted to evaluate pBRIT's intermediate data integration against three methods representing late integration. For this, we used another benchmark dataset, previously used to evaluate Endeavour, ToppGene and HyDRA performance. ToppGene and Endeavour integrate ranks computed on individual annotation sources, while HyDRa is an ensemble of rank aggregation methods applied directly on the ranks computed from Endeavour and ToppGene.

The reported AUC score for Endeavour and ToppGene using full annotation models were 0.908 and 0.951, respectively. The best AUC values for HyDRA, using Weighted Kendall, were 0.91 and 0.947, respectively, based on Endeavour and ToppGene ranks. pBRIT has at least similar performance to these late integration methods, with an overall minimal AUC score of 0.93 and a maximum of 0.96 using the full model (see Supplementary Fig. S3B, Table S4.2 and File S2). No significant improvement was observed using SVD for either N.Na or ALL.Na mode (P value = 0.91 and 0.12, respectively). However, there is a significant difference (P value < 0.0002) between N.Na and ALL.Na mode for both feature mining methodologies (See Supplementary Material S2: sheet 5). A more in depth comparison, based on the MRR is available in Supplementary Table S4.2, showing improved MRR values compared to Endeavour for 7/8 included diseases. For 4/8 diseases, the full pBRIT model outperforms both Endeavour, ToppGene and HyDRA based rank aggregation methods. These results indicate that our regression approach after intermediate integration provides a uniform prioritization strategy independent of ensemble methods, with at least similar performance.

3.3 Effect of annotation changes on prioritization

Due to regular updates to the ever expanding biological knowledge base, annotation sources used in gene prioritization are highly dynamic. This is reflected in the monthly archives of ontology based annotation sources such as GO and HPO. Consequently, computing similarity profiles based on these ontologies will also be subjected to changes. As Bayesian Ridge Regression should help in modeling uncertainties related to changing annotations, we explored the potential impact of changing annotations on the prioritization results (Supplementary Section S5). Based on computational feasibility and data availability, we selected GO as part of the functional annotations and HPO as part of phenotypic annotations to construct yearly versions of the annotation framework, ranging from 2009 to 2014, keeping the remaining eight annotation sources stable.

Ranking results of 250 genes from eight disease classes of the HyDRA based dataset are summarized in Supplementary Material S3, S4 and S5, showing a variance of <0.0002 on the overall AUC scores over the included timeframe. Additionally, no significant correlation was observed between annotation changes and the overall change in gene ranking (Supplementary Figs. S5.2.1–S5.2.12).

3.4 Effect of training set size and annotation bias

Although an ongoing debate in the machine learning domain is whether robust prediction requires more training data or better algorithms (Zhu *et al.*, 2012), the amount of training data is important for any supervised learning method. In the above benchmark sets, the number of training genes per disease class was large, especially for the Goh *et al.* data, and often involved well studied disease genes. Here, we evaluated pBRIT performance using limited training sets, targeting individual disease-gene associations extracted from the DisGeNET database (Pinero *et al.*, 2015) According to Figure 3A, Supplementary Table S4.3 and File S6: sheet 6, a small but significant ($P < 0.0005$) improvement in performance is seen between TF-IDF and TF-IDF→SVD feature extraction, using either regression strategy. On the other hand, the results again illustrate the importance of including phenotype association scores of both training and test genes during prioritization, with an overall improvement of over 10% in AUC ($p \simeq 0$). Analysis of MRR values (Fig. 4) shows that this effect flattens out past 25 training genes. This is also reflected in a positive correlation between AUC scores and number of training genes for 'All.NA' setups (Pearson's product-moment correlation, $P < 0.01$, Supplementary Figs. S6.2 and S6.4), which was absent for both 'N.Na' setups (Supplementary Figs. S6.2 and S6.4).

Next, we explored the possibility of annotation bias on our results. Pathway and MPO annotations often contain near perfect matches for the phenotype/disease categories that are getting predicted. As this might lead to biased results, we removed pathway and MPO databases from the model and repeated the analysis. Supplementary Figure S4.2B and Table S4.3B shows a small performance decrease (−0.01 in AUC score) compared to the full annotation set for both N.Na modes. For both ALL.Na modes, the decrease in AUC score is more pronounced. These results demonstrate that the potential effect of annotation bias is minimal in pBRIT.

3.5 Real-world performance evaluation

LO-OCV has long been a standard approach to evaluate gene prioritization tools. Since well characterized genes tend to dominate prioritization results, LO-OCV estimates might be over-optimistic. Therefore, a real test for any prioritization tool should be its capacity to prioritize newly discovered genes with minimal disease association information. To achieve this, we evaluated pBRIT performance on HPO to gene associations assigned after creation of pBRIT's annotation database (January 2015, Supplementary Table S1). pBRIT was used to prioritize genes in the context of

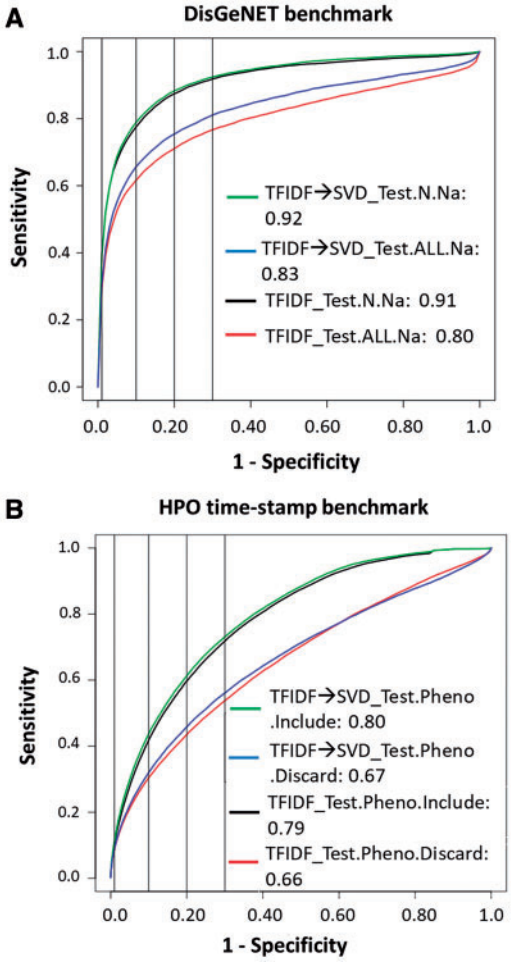

Fig. 3. ROC plot of pBRIT benchmark performance. (**A**) 779 *UMLS*-coded disease classes obtained from DisGeNET and (**B**) 2025 time-stamped HPO terms. The four vertical lines indicate the top1%, top10%, top20% and top30% of query genes which were prioritized

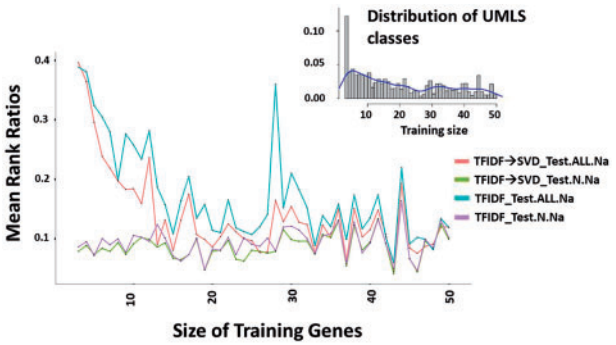

Fig. 4. Impact of training set size. Main: Mean rank ratio (MRR) versus number of training genes. Incorporation of test gene phenotypic information (N.NA) in the regression model results in a low and stable MRR, irrespective of feature extraction methodology. Without phenotypic information (All.Na), MRR decreases with increasing number of training genes. Insert: Distribution of training sizes per disease class

individual HPO phenotypic terms, instead of multi-phenotype diseases. A maximum AUC score of 0.80 and minimal MRR of 0.205 was obtained with the full pBRIT model (Fig. 3B; Supplementary Table S4.4 and File S7). SVD had a small but significant positive effect on prioritization (P value = 2.00E-56). Inclusion of phenotype data during regression again resulted in significantly better results for either feature mining methodology, similar to the retrospective validations. Lastly, pBRIT (with an annotation release updated in December 2016) was directly compared with two recent tools, Endeavour-v3.71 and Random Walk with Restart on multiple networks (RWR-M), which both have internal annotation sources built in or before December 2016. We achieved a maximum AUC score of 0.87 in comparison to 0.85 for Endeavour ($P < 0.0004$) and 0.68 for RWR-M methods ($P < 7.666348e-196$) (See Supplementary Section S4.2.1 and Fig. S4.3B for further details).

3.6 Results exploration and visualization

Researchers designing experiments based on prioritization results need insight into which annotation sources and training genes contribute more towards the ranking of specific genes. Although early and intermediate data fusion can obfuscate interpretation, we provide an interface to intuitively explore and explain these individual contributions.

As an example, prioritization results for *KCNA2* in the context of epileptic encephalopathy (Syrbe *et al.*, 2015) are shown in Figure 5 (For details, see Supplementary Section S7). The heatmap explains the gene-by-gene similarities. Darker shades indicate a larger contribution to the prioritization. *KCNA2* is top ranked mainly because of a higher similarity to *KCNB1*, *HCN1*, *KCNQ2* and *SCN2A*. Despite direct evidence in the literature of disease association for *NECAP1*, functional similarities to *KCNA2* are negligible. Comparison of Figure 5 and Supplementary Figure S7.1 shows that SVD transformation of the gene-by-feature matrices results in visibly more pronounced similarity scores. Second, pBRIT provides heatmaps of similarity scores per individual annotation source (Fig. 5B). These gene-specific plots highlight the training genes and annotation sources contributing most to the ranking of that particular gene. Again, it can be seen from Figure 5 and Supplementary Figure S7.1 that SVD provides more pronounced similarity profiles.

Finally, the pBRIT web-interface provides actual overlapping features between training and test genes, with the corresponding TF-IDF scores.

4 Discussion

We present a novel gene prioritization tool, based on Bayesian Ridge regression and utilizing an information-theoretic approach towards feature extraction followed by intermediate data integration. We compared pBRIT performance to nine current state-of-the-art methods under a variety of conditions, reflecting both different aspects of our methodology and varying degrees of prior evidence.

Although the Goh *et al.* (2007) benchmark set does not represent a typical gene prioritization use case due to extensive and curated gene lists associated to high level disease classes, important conclusions could be drawn from it. First, it provides initial evidence that the implemented TF-IDF approach is feasible, as pBRIT globally outperforms four existing methods using alternative approaches, which were originally benchmarked on this dataset. It thus demonstrates the validity of applying TF-IDF in discriminatory mining of genomic features other than textual information, for which it was originally presented. In our case, these features are structured concepts holding specific details about gene functionality or phenotype associations. Furthermore, leveraging of phenotypic information and performing SVD transformation of the feature-by-gene

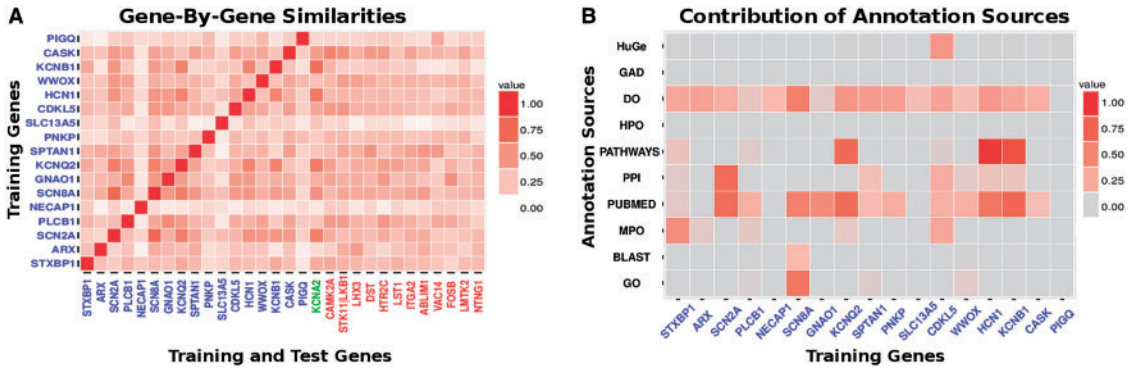

Fig. 5. Exploring prioritization results using heatmap plots. (A). The functional annotation matrix X, illustrating the contribution of individual training genes (Y-axis) during regression, using the full TF-IDF->SVD_Test.Pheno.Include model. Darker shades indicate higher contributions. In this example, the gene to be prioritized was KCNA2. (B) Contribution of individual annotation sources for each training gene to the ranking of KCNA2

matrices, being two of the key characteristics of pBRIT, improves AUC scores by approximately 9%. Finally, stable AUC scores across individual disease classes indicates that prioritization is not biased towards particular disease classes.

The singular value decomposition of TF-IDF weighted gene-by-feature matrices, prior to data integration is a novel characteristic of pBRIT. We applied SVD transformation because TF-IDF based weights show two limitations. First, weights are generally computed for every individual feature assuming feature independence. However, features do co-occur in biological data and as such contain additional functional information about the gene. Second, since most features are rather specific, the binary gene-by-feature matrix holds many zeros. This might impact feature learning and generally leads to less cohesive gene clusters, in turn affecting the overall gene-by-gene similarity profiles used to prioritize candidate genes. SVD allows a reduction in the feature space dimensionality, thereby reducing sparseness and implicitly modelling co-occurrences and latent relationships. Despite a limited 1–4% gain in performance over TF-IDF alone, the benefits are twofold. First, latent relations can be helpful to identify candidate genes in rare diseases having related disorders with overlapping phenotypes. Here, the choice of SVD-based feature mining can improve prioritization based on training genes implicated in those related disorders. Second, as shown in Figure 5A and B, SVD increases the resolution of the gene-by-gene similarity profiles, facilitating result interpretation using the provided visualization tools.

In reality only a limited set of training genes can be defined for most genetic disorders. These genes are often less studied and reflect a subjective measurement of how well they describe the underlying disease etiology. We simulated this by selecting DisGeNET disease classes with maximally 51 associated genes. The obtained results illustrate the power of the second key characteristic of pBRIT. Indeed, we observe a significant improvement of approximately 11–14% (Supplementary Table S4.3) when phenotypic information from the 99 random test genes is taken into account (Test.N.Na) during regression. As this effect is more pronounced for smaller training sets, it makes pBRIT a valuable asset in the study of rare and less studied disorders.

Given the potential impact of annotation changes over time on functional interpretation (Wadi *et al.*, 2016), it should be noted that pBRIT results are not subject to such changes. This is likely attributable to BRR, which implicitly captures the uncertainties in the model arising due to changes, and stabilizes the ranking. Although all annotation sources get updated, our simulation was limited to HPO and GO due to the availability of archived and quantifiable data over a fixed time interval. Therefore, we can not fully exclude that the impact is nullified by the contribution of the eight annotation sources that were kept stable during the experiment.

Overall, pBRIT performs equally well, and often better than competing methods on cross-validation studies. Although we removed prior knowledge on phenotypic association between test and training genes during regression, this knowledge still contributed indirectly to similarity scores through the IDF component of the TF-IDF calculations. To exclude any indirect contribution of prior knowledge, the real test is therefore to prioritize genes that have been published after construction of the internal annotation database. For these genes, the tools by definition lack any prior knowledge of the gene-disease association. Hence, we performed a final validation of pBRIT on 2025 HPO terms having 2484 novel gene-phenotype associations. Interestingly, the obtained maximum AUC score of 0.80 using the full pBRIT model, is lower in comparison to the performed LO-OCV-based analyses, confirming the tendency of LO-OCV to overestimate performance. A subset of 693 HPO terms was also analyzed using Endeavour-v3.71 and RWR-M, with neither method outperforming the full pBRIT model. The inferior performance of RWR-M, using only four annotation sources, mainly demonstrates that integration and fusion of more and relevant annotation sources has a distinctive advantage. Since Endeavour-v3.71 and pBRIT both use approximately similar annotation sources, our results indicate that intermediate integration combined with BRR offers a valuable alternative to late integration and rank fusion while offering superior computation speed.

Additionally, we want to highlight the importance of community driven data competitions in this context [like CAFA (Jiang *et al.*, 2016)] for developing true prospective benchmark datasets, such that future function prediction and prioritization tools can be properly evaluated on high quality and unbiased datasets.

Despite the promising performance of pBRIT, we believe that further improvements are possible. First, alternative to empirical selection of k in SVD, one might apply a probabilistic generative model, either using classical Latent Dirichlet Allocation (Blei *et al.*, 2003) or aspect Bernoulli models (Bingham *et al.*, 2009). Second, we focused on giving equal weights to all annotation sources during the construction of composite similarity matrices. However, data driven approaches for optimal weight selection might further improve performance. Finally, it might be interesting to investigate the influence of incorporating informative priors such as Zellner's g-prior (Hanson *et al.*, 2014) in the Bayesian ridge regression.

5 Conclusion

In conclusion, our results present pBRIT as robust and performant. Its performance was competitive, or better, compared to current state-of-the-art methods when applied to their benchmark datasets. We demonstrated performance of pBRIT both at the level of the information-theoretic model, by evaluating TF-IDF and SVD as feature extraction approaches, and by contrasting intermediate data fusion to other data fusion methodologies, and at the level of the regression model, by evaluating the effect of incorporating phenotypic information from test genes into the model. Additionally, we explored the predictive power of pBRIT to detect novel disease causing genes without prior information in the internal database. We demonstrated that regression under the Bayesian framework has an advantage in handling uncertainties and errors in the annotation sources, while incorporation of a ridge regression model helps in alleviating the problem of over-fitting and multi-collinearity in the model. Ultimately, these aspects lead to a more robust prediction. We can therefore conclude that each aspect of the pBRIT methodology provides distinct and additive benefits, making the TF-IDF→SVD.Pheno.Include approach, referenced as the full model, the method of choice in real-world application. Finally, we extended the prioritization task to provide insight in the resulting gene ranks through visualization. Using heatmap plots showing both pre- and post-integration similarity scores, together with actual feature matches between training and test genes, interpretation of gene ranks becomes intuitive.

Acknowledgements

This research was supported by funding from the University of Antwerp (Lanceringsproject), the Fund for Scientific Research, Flanders (FWO, Belgium, G.0221.12, 1513715N and 12D1717N), The Dutch Heart Foundation (2013T093) and the Fondation Leducq (MIBAVA-Leducq 12CVD03). Work by KULeuven researchers was supported by KU Leuven CELSA/17/032, the Flemish Government (IWT 150865, FWO 06260) and VIB: ELIXIR. B.L. is senior clinical investigator of the Fund for Scientific Research, Flanders and holds a starting grant from the European Research Council (ERC-StG-2012-30972-BRAVE).

Conflict of Interest: none declared.

References

Aerts,S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.

Baglama,J. and Reichel,L. (2012) irlba: Fast Partial SVD by Implicitly-Restarted Lanczos Bidiagonalization. R Package Version 1.0.2.

Bingham,E. *et al.* (2009) The aspect Bernoulli model: multiple causes of presences and absences. *Pattern Anal. Appl.*, **12**, 55–78.

Blei,D.M. *et al.* (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Chen,B. *et al.* (2014) Identifying disease genes by integrating multiple data sources. *BMC Med. Genomics*, **7**, S2.

Chen,B. *et al.* (2015) A fast and high performance multiple data integration algorithm for identifying human disease genes. *BMC Medical Genomics*, **8**, S2.

Chen,J. *et al.* (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.

Chen,Y. *et al.* (2011) In silico gene prioritization by integrating multiple data sources. *PLoS One*, **6**, e21137.

de los Campos,G. *et al.* (2013) Genome-enabled prediction using the BLR (Bayesian Linear Regression) R-package. Genome-Wide Association Studies and Genomic Prediction, Humana Press, Totowa, NJ, pp. 299–320.

Gillis,J. and Pavlidis,P. (2013) Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics*, **29**, 476–482.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, **104**, 8685–8690.

Groß,A. *et al.* (2012) Impact of ontology evolution on functional analyses. *Bioinformatics*, **28**, 2671–2677.

Hanson,T.E. *et al.* (2014) Informative *g*-priors for logistic regression. *Bayesian Anal.*, **9**, 597–612.

Hofmann,T. (2004) Latent semantic models for collaborative filtering. *ACM Trans. Inform. Syst.*, **22**, 89–115.

Jiang,Y. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, **17**, 184.

Kim,M. *et al.* (2015) HyDRA: gene prioritization via hybrid distance-score rank aggregation. *Bioinformatics*, **31**, 1034–1043.

Kohler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes *Am. J. Hum. Genet.*, **82**, 949–958.

Kumar,A.A. *et al.* (2013) GOParGenPy: a high throughput method to generate Gene Ontology data matrices. *BMC Bioinformatics*, **14**, 242.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60 706 humans. *Nature*, **536**, 285–291.

Li,Y. and Patra,J.C. (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**, 1219–1224.

Liekens,A.M. *et al.* (2011) BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol.*, **12**, R57.

Moreau,Y. and Tranchevent,L.C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.

Pavlidis,P. *et al.* (2002) Learning gene functional classifications from multiple data types. *J. Comput. Biol.*, **9**, 401–411.

Pinero,J. *et al.* (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**, bav028.

Schnoes,A.M. *et al.* (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.

Syrbe,S. *et al.* (2015) De novo loss-or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. *Nat. Genet.*, **47**, 393–399.

Tierney,L. *et al.* (2009) Snow: A parallel computing framework for the R system. *International Journal of Parallel Programming*, **37**, 78–90.

Tranchevent,L.C. *et al.* (2011) A guide to web tools to prioritize candidate genes. *Brief. Bioinformatics*, **12**, 22–32.

Tranchevent,L.C. *et al.* (2016) Candidate gene prioritization with Endeavour. *Nucleic Acids Res.*, **44**, W117–W121.

Valdeolivas,A. *et al.* (2017) Random walk with restart on multiplex and heterogeneous biological networks. *bioRxiv*. doi: 10.1101/134734.

Wadi,L. *et al.* (2016) Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, **13**, 705–706.

Wu,X. *et al.* (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.

Zhang,W. *et al.* (2011) DomainRBF: a Bayesian regression approach to the prioritization of candidate domains for complex diseases. *BMC Syst. Biol.*, **5**, 55.

Zhu,X. *et al.* (2012) Do we need more training data or better models for object detection? *BMVC*, **3**, 5.

Zitnik,M. *et al.* (2015) Gene prioritization by compressive data fusion and chaining. *PLoS Comput. Biol.*, **11**, e1004552.