Published in final edited form as:

Hum Mutat. 2013 June ; 34(6): 853–859. doi:10.1002/humu.22317.

Prioritization of Retinal Disease Genes: An Integrative Approach

Alex H. Wagner^{1,2,*}, Kyle R. Taylor³, Adam P. DeLuca¹, Thomas L. Casavant^{1,2,3}, Robert F. Mullins², Edwin M. Stone², Todd E. Scheetz^{1,2}, and Terry A. Braun^{1,2}

¹Department of Biomedical Engineering, University of Iowa, Iowa City, Iowa

²Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, Iowa

³Department of Electrical and Computer Engineering, University of Iowa, Iowa City, Iowa

Abstract

The discovery of novel disease-associated variations in genes is often a daunting task in highly heterogeneous disease classes. We seek a generalizable algorithm that integrates multiple publicly available genomic data sources in a machine-learning model for the prioritization of candidates identified in patients with retinal disease. To approach this problem, we generate a set of feature vectors from publicly available microarray, RNA-seq, and ChIP-seq datasets of biological relevance to retinal disease, to observe patterns in gene expression specificity among tissues of the body and the eye, in addition to photoreceptor-specific signals by the CRX transcription factor. Using these features, we describe a novel algorithm, positive and unlabeled learning for prioritization (PULP). This article compares several popular supervised learning techniques as the regression function for PULP. The results demonstrate a highly significant enrichment for previously characterized disease genes using a logistic regression method. Finally, a comparison of PULP with the popular gene prioritization tool ENDEAVOUR shows superior prioritization of retinal disease genes from previous studies.

Keywords

machine learning; data integration; gene prioritization; prediction

Introduction

The retina is an attractive tissue for studying inherited diseases because it can be evaluated by direct observation, functionally by visual acuity and field of vision, structurally by optical coherence tomography and electrically for photoreceptor activity. Example mechanisms for degenerative retinal disease include simple Mendelian inheritance [den Hollander et al., 2006], heterogeneous disorders with multiple genes [Badano et al., 2003; Chiang et al., 2004; 2006; Mykytyn et al., 2001; 2002; Nishimura et al., 2001; 2005], and presumably complex disorders caused by numerous genetic and risk factors [Hageman et al., 2005;

*Correspondence to: Alex H. Wagner, University of Iowa, Biomedical Engineering, 5315 Seamans Center, Iowa City, IA 52242. alex-wagner@uiowa.edu.

Additional Supporting Information may be found in the online version of this article.

Disclosure statement: The authors declare no conflict of interest.

Haines et al., 2005; Jakobsdottir et al., 2005; Klein et al., 2005]. Identifying the genetic variants contributing to the disease phenotype for a patient can lead to a greater understanding of disease progression, heritability, and potential treatments.

For this manuscript, we focused on genes that have variants shown to cause retinal degenerative disorders. The majority of these genes are linked to a group of inherited conditions collectively referred to as retinitis pigmentosa (RP). Patients with the disease gradually develop permanently decreased visual fields and visual acuity as photoreceptors in the retina die. RP is a rare disease, affecting approximately 100,000 people in the United States (Foundation Fighting Blindness; <http://www.blindness.org/>). To date, over 180 retinal disease genes have been identified. These genes account for approximately 50% of known RP cases to date, making gene identification for this disease an area of interest for further investigation.

Exome sequencing allows us to reconstruct a profile of genetic variants in a person's exome—the small percentage of DNA from the genome that is transcribed. A typical exome sequencing yields tens of thousands of variants, from which the number is reduced using several filtering strategies. Thousands of exomes have been sequenced in previous studies [Norton et al., 2012; Tennessen et al., 2012] to identify rare variants and validate their association with diseases. Additional strategies include sequencing multiple exomes to identify variants that cosegregate with the disease state, identifying variants that are likely harmful (missense, nonsense, splice-site), filtering known SNPs, and removing systemic variations (variations that materialize because of the algorithms and parameters that are used). In the absence of thousands of exomes for the phenotype, these strategies fail to reduce the number of variants to a feasible number that can be molecularly confirmed. Here, we explore methods to leverage additional data sources to prioritize variant lists from exome sequencing experiments.

There are several features predicted to be informative in identifying RP and other retinal disorder genes by experts in the field. The cone-rod homeobox transcription factor (CRX) is a well-studied transcription factor that is involved in the differentiation of photoreceptor cells, and therefore regulates genes that likely cause defects in retinal function when disabled [Hennig et al., 2008]. To examine this, ChIP-seq data detailing CRX binding site locations and intensities are included among the features used in this study. Genes that are expressed at higher levels in retinal tissue and are upregulated when compared both with other body tissues and other tissues of the eye are expected to play key roles in retinal function, and are likely candidates for disease-associated mutations. These features are captured in microarray profiles of eye tissues and RNA-seq profiling of retinal tissue and the BodyMap 2.0 profile of human body tissues (see *Systems and Methods* for details). Lastly, photoreceptor-specific genes do not exist in organisms whose evolutionary fitness is determined by genome compactness—thus, on average these genes may be longer or contain more exons than human genes homologous to genes found in single-celled organisms [Eisenberg and Levanon, 2003; Moriyama and Powell, 1998]. To capture this, we calculate the total number of exons and gene length as additional features.

Previous work in the field of gene prioritization has yielded multiple algorithms and techniques to prioritize variants identified in a number of different disease models [Alsaber et al., 2006; Chang and Lin, 2011; Furney et al., 2008; Gajendran et al., 2007; Huang et al., 2008; Lombard et al., 2007; Rasche et al., 2008]. Most of these algorithms rely heavily on human curated datasets such as protein–protein interaction networks and gene ontologies, resulting in a bias toward well-studied genes [Piro and Di Cunto, 2012]. Our goal is to develop an unbiased technique for the prioritization of variants identified in exome sequencing of patients with retinal diseases.

System and Methods

We created feature vectors using experimental data from ChIP-seq, mRNA-seq, and microarray experiments selected by experts in the field of retinal disease. We used these features to train and test several computational prediction techniques to prioritize variant candidate lists from exome studies. These techniques were scored by a rank-sum metric to compare their effectiveness. A Monte Carlo simulation was used to empirically measure statistical significance of our findings.

Feature Creation

The publicly available ChIP-seq data (CRX vs. IgG control) from mouse retina were obtained online (SRA accession GSE20012). Alignment to the mm9 mouse genome was performed with Bowtie [Langmead et al., 2009]. Peaks were called using MACS and coordinates were translated to the hg19 human genome using the UCSC liftover tool [Zhang et al., 2008]. Mouse peak sequences that did not map to the human genome using this utility due to duplication or partial matches were analyzed with BLAT, and their peak scores were adjusted to account for the percentage of the sequence aligned. After remapping, we recorded three separate features for each gene: the number of peaks within 4 kb of the transcription start site, the summed peak score of those peaks, and the distance to the nearest CRX binding site to the gene.

Publicly available mRNA-seq data from human retina and 16 tissues surveyed in the BodyMap 2.0 experiment were obtained online (SRA accessions GSE22765, ERP000546). Datasets were independently analyzed using Tophat for alignment of mRNA reads to the hg19 human genome and Cufflinks for gene-level expression quantification using RefSeq transcripts as a transcript assembly guide [Trapnell et al., 2009; 2010]. We included estimates from all 17 datasets as independent features, and created a secondary feature that represented the retina expression level in fragments per kilobase exon model per million mapped reads (FPKM) minus the average expression level of the 16 bodymap tissues (also in FPKM).

Available microarray data (GEO accession GSE41102) represent an expression profile of each gene across 10 ocular tissues, including retina [Wagner et al., 2013]. Expression estimates were calculated using the Affymetrix Power Tools PLIER algorithm with GC-content background correction. We created a feature for each of the 10 expression estimates and we generated 10 additional features, each reflecting the z -score of a gene in each tissue

when compared with the other nine tissues. These additional features helped us quantify ocular tissue specificity.

The maximum transcript isoform length of each gene and the exon count for that transcript were recorded for each gene. These values were pulled from RefSeq track data available at the UCSC genome browser.

Feature Scaling

Our experimental data are very heavy tailed, so a logarithmic operation is applied to transform the data to a less extreme distribution. To account for the heterogeneous nature of our training data, it is imperative to perform feature centering and scaling before attempting to generate any models from the following techniques. Centering was performed independently for each feature by mean subtraction and was subsequently scaled by the standard deviation (Fig. 1). Thus, a feature score of 2 after scaling and centering represents a gene that is 2 standard deviations above the mean for that feature.

Known Disease Genes and Candidate Lists

Genes known to be associated with retinal disease mutations were extracted from RetNet, a manually curated list of such genes located at <https://sph.uth.tmc.edu/retnet/> (exact list in Supp. Table S1). All feature vectors were labeled with Human Genome Organization Gene Nomenclature Committee gene symbols, a list we extracted from the *Refseq Genes* track *refFlat* table from the UCSC genome browser located at <http://genome.ucsc.edu/>. All feature vectors with gene symbols identified in the RetNet list were in the “positive” (labeled) class; all other feature vectors were treated as negatives (unlabeled).

Thirteen previous linkage studies for RP as identified in RetNet were studied as a basis for candidate lists in testing algorithm performance (see *Comparing Algorithm Performance*). In each study, all genes in the identified linkage region were assembled as a candidate list. We used these lists to compare the prioritization of the known causative genes (which have been identified since the linkage study was published) using both PULP and the existing ENDEAVOUR gene prioritization tool [Aerts et al., 2006]. Exact gene candidate lists can be found in Supp. Table S2.

Additionally, exome sequencing was performed on a single member of the family sequenced by Zuchner et al. (2011) to discover the *DHDDS* gene variant that causes RP. An Agilent SureSelect 38MB capture was performed and the library sequenced on an Illumina GAIIX. Sequence was aligned with BWA [Li and Durbin, 2009] and variants were called using the Unified Genotyper from the Genome Analysis Toolkit (GATK) [McKenna et al., 2010]. Variants were filtered to remove intronic, intergenic, and synonymous variations. Variants were removed with allele frequencies greater than 1% in 1000 genomes [1000 Genomes Project Consortium, 2010], and 0.6% in the Exome Sequencing Project (ESP) cohorts (Exome Variant Server, NHLBI Exome Sequencing Project, Seattle, WA [URL:<http://evs.gs.washington.edu/EVS/>] Jan 2012). A similar procedure was used to identify variants in a patient sequenced by Tucker, et al. (2011) in a study identifying a *MAK* gene variant as a cause of RP. Exact gene candidate lists can be found in Supp. Table S2.

Nontraditional Classifiers

Traditional supervised machine learning techniques use labeled data to train a model that can then be used to predict the labels for unknown data. The labels are in the form of classes, which can be binary or multivalued. Our problem does not fall into the traditional supervised machine learning category, but is based on predicting which examples are positive or negative from only data points that are labeled or not. The nature of our data defines the labeled instances as positive (i.e., genes known to cause retinal disease) and the unlabeled instances as a mixture of positive and negative (i.e., genes that have not yet been shown to cause retinal disease, some of which may actually cause it).

Previous research on classifying with positive and unlabeled training examples formalized the concept of a nontraditional classifier that predicts the probability a test instance is labeled (and thus positive). Such a classifier returns probabilities that differ from a traditional classifier by only a constant factor [Elkan and Noto, 2008]. Formally, for a traditional classifier $f(x)$ that predicts if an input is positive or negative, a nontraditional classifier $g(x)$ that predicts if an input is labeled (positive) or unlabeled, and an empirically derived constant factor c , there exists a relationship between the classifiers such that:

$$f(x) = g(x) / c \quad (1)$$

Thus, unlabeled examples can be treated as negative examples and traditional machine learning techniques can be trained on these data to predict the class of the unlabeled examples. As a corollary, they also state that if the goal is to rank testing instances, there is no need to estimate the constant factor to transform the output from the nontraditional classifier. Applying these results, we are able to perform leave-one-out testing on our gene set using a variety of nontraditional classifiers based on established machine learning techniques. This was executed by iteratively removing an example from the full gene set, training the classifier on the remaining examples, and then using the model to generate a probability that the withheld gene is a member of the positive class. After repeating this process for all instances in the full gene set, the genes are ranked in order of decreasing probability (Fig. 2). This method is used with all the classifiers described below.

Logistic regression—Logistic regression performs classification by fitting a sigmoid function to the training instances and using the fitted sigmoid to compute the probability that an unknown instance is a positive or negative example [Le Cessie and Van Houwelingen, 1992]. The sigmoid function is fitted to the data such that the decision boundary lies at 0.5, where instances from one class are less than 0.5 and instances of the other class are greater than 0.5. Maximum likelihood estimation is used to estimate the parameters of the logistic regression model. Implementations of logistic regression were used from both R and Weka [Hall et al., 2009] to independently verify results, and rankings were determined as described in the previous section.

Naïve Bayes—The Naïve Bayes classifier uses Bayes' theorem to compute the probability that an instance belongs to a class by using the *naïve* assumption that each feature is independent from each other [John and Langley, 1995]. This assumption simplifies Bayes'

theorem and makes it fast even on large datasets. Probability outputs from this technique were ranked directly, as previously described.

Support vector machines—Support vector machines (SVM) classify by finding the maximum-margin separating plane between two classes, and uses this hyperplane as the decision boundary for unknown points [Vladimir and Vapnik, 1995]. SVMs were trained using Platt's Sequential Minimal Optimization algorithm [Platt, 1998]. Linear SVMs operate in the feature space of the dataset and are therefore limited to only finding hyperplanes in the feature space. Nonlinear SVMs use what is known as the kernel trick to transform the feature space into an inner product space—where the data are ideally linearly separable. In our case, we used the Radial Basis Function kernel that finds an elliptical decision boundary within the feature space.

One important aspect of SVMs is the importance of tuning the parameters, specifically C and gamma. The parameter C controls the cost of placing a training instance on the wrong side of the hyperplane during training. Gamma is a parameter used in the kernel to control the transform function. We used the libsvm package for our implementation of the SVM, along with their grid search utility for parameter selection [Chang and Lin, 2011]. We modified the utility to compute the F_2 score to better tune the parameters by accounting for both recall and precision (favoring recall). This is important because the number of positive examples is only 161, and the number of unlabeled examples is approximately 17,000. This means that an accuracy over 99% is easily obtained by classifying all examples as negative.

Classification for the SVM is determined by which side of the separating hyperplane an unknown datum would be placed. The distance between the unknown data point and the hyperplane can also be computed, and is used in the function for generating probability output from the SVM. To generate probabilities from the trained SVM, a technique called Platt's scaling is used [Platt, 1999].

Hierarchical clustering—We evaluated a ranking based on the principles of hierarchical clustering. For each gene, we calculated the average Pearson correlation distance to every RetNet gene. If the calculation was performed for a RetNet gene, self-comparison was excluded from the average. Low Pearson correlation distance averages were ranked as better than high Pearson correlation distance averages.

Scoring Classifiers

To compare the performance of the probabilistic classifiers described above, we selected from each ranking genes that are in RetNet. Adding the ranks of these genes provides a rank sum, which can be used to compare how highly ranked RetNet genes are among all genes in their respective rankings. This comparison is nonparametric, which allows us to quantitatively evaluate the performance of each classifier without regard to the quality of the calibration of the output probabilities. Sports rankings are used for ease of reference (i.e., 1 is the highest ranked gene).

Significance of these findings was evaluated using a Monte Carlo method. For each trial, we created a ranking of equal size to the classifier rankings, and then selected from it randomly

without replacement a number of times equal to the size of the RetNet set. Summing the selected ranks yields a single rank sum trial selected randomly. We repeated this process for 1 million trials to obtain an empirical distribution of randomly selected rank sums.

ENDEAVOUR Comparison

Using the ENDEAVOUR Web tool (<http://homes.esat.kuleuven.be/~bioiuser/endeavour/tool/endeavourweb.php>), we trained the algorithm with the RetNet gene list, removing any samples from training that were to be prioritized as part of a candidate list. This process was performed thirteen times for the previous gene intervals discussed in *Known Disease Genes and Candidate Lists*. The training data selected for this algorithm were the same as used by Ebermann et al. (2007) in the discovery of an Usher Syndrome mutation.

Performance on Recent Studies

To demonstrate performance on gene lists from recent exome sequencing studies, we tested our algorithm on gene lists from two recent RP causative gene variant papers [Tucker et al., 2011; Zuchner et al., 2011]. We rank each of these gene lists with PULP and report the ranking of the known causative variant.

Results

We compared each of our classifiers against a Monte Carlo simulation of ranking 13 historical candidate lists to observe significance and relative performance. We also compared the performance of the logistic regression algorithm against a state-of-the-art prediction tool—ENDEAVOUR—that uses predefined, general datasets. Finally, we prioritized a candidate list from a recent RP study to observe how our results compared with the published findings.

Comparing Classifier Performance

Using a Monte Carlo method found that all of the classifiers were able to prioritize RetNet genes with a significance of $P < 1e-6$ (Fig. 3). Ordering the ability of the classifiers to prioritize RetNet genes, we observed from best to worst: logistic regression > SVM > hierarchical clustering > naïve Bayes. Rank sum difference (as described in *Systems and Methods*) between these classifiers is visually depicted in Figure 3.

The logistic regression classifier coefficients were analyzed to determine which features were most informative. To obtain a single set of coefficients, we trained a model that did not exclude any genes, and evaluated the model. All of the coefficients contributed toward the regression model (see Supp. Table S3), but two were clearly given more weight than the rest. The largest observed weight was for the “RetDiff” feature, a compound feature that measured the expression in retina RNA-seq subtracted by the average expression in the BodyMap 2.0 RNA-seq data. The second largest weight was for the “Exons” feature, indicating that genes with more exons are more likely to influence retinal degenerative disease. However, these observations must be considered in the context of the feature vector as a whole—many of the other features are better correlated to each other, and thus have reduced weight, despite being good predictors [Tolosi and Lengauer, 2011].

Comparing Algorithm Performance

We selected 13 candidate lists from previous studies to observe algorithm performance. Prioritization of candidates from these lists produced a summed rank of 81 among all top candidates. The Monte Carlo simulation found this result to be highly significant, with a $P = 0.000115$ (Table 1). We also found that in 54% of these cases, the highest ranked gene was a RetNet gene, and in 77% a RetNet gene was within the top five ranked genes.

Additionally, we tracked performance of ENDEAVOUR on the same genetic intervals to compare our performance with that of an existing Web tool for gene prioritization (Table 1). We observed that ENDEAVOUR also significantly prioritizes the candidate lists from these intervals with a $P = 0.00428$. ENDEAVOUR predicted a RetNet gene as the most likely candidate in 8% of cases, and predicted a RetNet gene within the top five in 38% of cases.

Performance on Recent Studies

In a study of a family with autosomal recessive RP, the causative variant was identified in *DHDDS* by exome sequencing of three affected siblings and an unaffected sibling [Zuchner et al., 2011]. Using the filtering criteria described in Systems and Methods on a single exome from the study, we had a resultant list of 20 candidate genes. Ranking these genes with PULP, we identified *DHDDS* as the second most likely causative gene. A second study identifying a *MAK* gene variant associated with RP yielded a list of 348 genes [Tucker et al., 2011]. PULP ranked *MAK* as the second most likely causative gene in this list.

Discussion

Often in exome sequencing experiments, candidate lists are lengthy, and the most exclusive filters can run a risk of erroneously filtering out the candidate of interest, despite being a good filter for the average case. Our strategy allows for the prioritization of variants without using these criteria (e.g., assuming homozygosity, filtering on dbSNP). The results show that there are several methods for prioritizing gene candidates by creating models based on available data beyond the exome sequencing.

Throughout the course of obtaining the results, we ran into several pitfalls that make this problem more computationally demanding than traditional machine learning problems. We briefly discussed the problem of training a nontraditional classifier using only positive and unlabeled data, as well as the problem of a heavily imbalanced training set. We addressed these problems by evaluating candidates nonparametrically and by using appropriate objective function criteria (F_2 score in lieu of accuracy). Additionally, there is the problem of the limited set size—our data are simultaneously both our training and our testing data. We approached this problem by using a leave-one-out approach with our classifiers, which is traditionally used to estimate the performance of the classifier. This allowed us to train similar, unbiased models for computing estimates for each gene.

One of the advantages of our analysis is its generalizability. Through the integration of multiple informative features and the application of multiple logistic regression models, PULP calculates probabilities that any gene is a member of a known disease gene set. Provided are several examples of common datasets that meet these criteria for retinal

disease, though selection of informative features is still a matter of expert knowledge. Using the same technique on high-throughput sequencing and microarray datasets related to ciliopathies, cancers, and other heterogeneous diseases will provide unbiased prioritization of genes for these other disease classes to assist in the pursuit of causative mutations.

Our results demonstrated a marked improvement in retinal disease gene prioritization using expert-selected data versus general datasets by comparison with the ENDEAVOUR Web utility. Although the general approach still significantly prioritizes gene lists, the adjusted rank scores demonstrate a wide margin in the comparative abilities of the two approaches. This is not a novel observation; several disease-specific approaches have been previously developed to improve upon prioritizations by general approaches based on this [Piro and Di Cunto, 2012]. To our knowledge, such a system has not been implemented for retinal diseases before this.

Ultimately, our results show that our selected feature set in conjunction with PULP is highly effective in identifying and prioritizing gene candidates for heritable retinal diseases. One advantage of using a logistic regression classifier with PULP is the intuitive, well-calibrated probabilistic outputs. In the future, we plan to use estimators to scale these probabilities into an informative score for assessing confidence in our results. Another future goal is to improve upon the model with the addition of other informative features, for example, binding site data for other transcription factors associated with photoreceptor development and RNA-seq data for other ocular tissues. Additionally, ensemble methods using the informative power of both logistic regression and other classifier methods may improve future performance. Finally, we plan to develop our technique into a publicly accessible Web tool for use in all avenues of retinal disease research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Mark Christopher, David Dellsperger, and David Thole for their assistance in data collection. This research was supported in part through computational resources provided by The University of Iowa, Iowa City, Iowa.

Acknowledgments

Contract grant sponsor: National Institute of General Medical Sciences Bioinformatics Award (Grant T32GM082729).

References

- 1000 Genomes Project Consortium. Durbin RM, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467:1061–1073. [PubMed: 20981092]
- Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. Gene prioritization through genomic data fusion. *Nat Biotechnol*. 2006; 24:537–544. [PubMed: 16680138]

- Alsaber R, Tabone CJ, Kandpal RP. Predicting candidate genes for human deafness disorders: a bioinformatics approach. *BMC Genomics*. 2006; 7:180. [PubMed: 16854223]
- Badano JL, Ansley SJ, Leitch CC, Lewis RA, Lupski JR, Katsanis N. Identification of a novel Bardet-Biedl syndrome protein, BBS7, that shares structural features with BBS1 and BBS2. *Am J Hum Genet*. 2003; 72:650–658. [PubMed: 12567324]
- Banerjee P, et al. Homozygosity and physical mapping of the autosomal recessive retinitis pigmentosa locus (RP14) on chromosome 6p21.3. *Genomics*. 1998; 48:171–177. [PubMed: 9521870]
- Bareil C, et al. Segregation of a mutation in CNGB1 encoding the beta-subunit of the rod cGMP-gated channel in a family with autosomal recessive retinitis pigmentosa. *Human Genetics*. 2001; 108:328–334. [PubMed: 11379879]
- Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011; 2:27.
- Chen XN, Korenberg JR, Jiang M, Shen D, Fong HK. Localization of the human RGR opsin gene to chromosome 10q23. *Human Genetics*. 1996; 97:720–722. [PubMed: 8641686]
- Chiang AP, Beck JS, Yen HJ, Tayeh MK, Scheetz TE, Swiderski RE, Nishimura DY, Braun TA, Kim KY, Huang J, Elbedour K, Carmi R, et al. Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). *Proc Natl Acad Sci USA*. 2006; 103:6287–6292. [PubMed: 16606853]
- Chiang AP, Nishimura D, Searby C, Elbedour K, Carmi R, Ferguson AL, Secrist J, Braun T, Casavant T, Stone EM, Sheffield VC. Comparative genomic analysis identifies an ADP-ribosylation factor-like gene as the cause of Bardet-Biedl syndrome (BBS3). *Am J Hum Genet*. 2004; 75:475–484. [PubMed: 15258860]
- den Hollander AI, Koenekoop RK, Yzer S, Lopez I, Arends ML, Voeselek KE, Zonneveld MN, Strom TM, Meitinger T, Brunner HG, Hoyng CB, van den Born LI, Rohrschneider K, Cremers FPM. Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am J Hum Genet*. 2006; 79:556–561. [PubMed: 16909394]
- Ebermann I, Scholl H, Charbel Issa P, Becirovic E, Lamprecht J, Jurklics B, Millán J, Aller E, Mitter D, Bolz H. A novel gene for Usher syndrome type 2: mutations in the long isoform of whirlin are associated with retinitis pigmentosa and sensorineural hearing loss. *Hum Genet*. 2007; 121:203–211. [PubMed: 17171570]
- Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet*. 2003; 19:362–365. [PubMed: 12850439]
- Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. *ACM, New York, NY*. 2008:213–220.
- Furney SJ, Calvo B, Larranaga P, Lozano JA, Lopez-Bigas N. Prioritization of candidate cancer genes—an aid to oncogenomic studies. *Nucleic Acids Res*. 2008; 36:e115. [PubMed: 18710882]
- Gajendran VK, Lin JR, Fyhrie DP. An application of bioinformatics and text mining to the discovery of novel genes related to bone biology. *Bone*. 2007; 40:1378–1388. [PubMed: 17320499]
- Gerber S, et al. The photoreceptor cell-specific nuclear receptor gene (PNR) accounts for retinitis pigmentosa in the Crypto-Jews from Portugal (Marranos), survivors from the Spanish Inquisition. *Human Genetics*. 2000; 107:276–284. [PubMed: 11071390]
- Gu SM, et al. Mutations in RPE65 cause autosomal recessive childhood-onset severe retinal dystrophy. *Nature Genetics*. 1997; 17:194–197. [PubMed: 9326941]
- Hageman GS, Anderson DH, Johnson LV, Hancox LS, Taiber AJ, Hardisty LI, Hageman JL, Stockman HA, Borchardt JD, Gehrs KM, Smith RJ, Silvestri G, et al. A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc Natl Acad Sci USA*. 2005; 102:7227–7232. [PubMed: 15870199]
- Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA. Complement factor H variant increases the risk of age-related macular degeneration. *Science*. 2005; 308:419–421. [PubMed: 15761120]
- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explor Newslett*. 2009; 11:10–18.

- Hennig AK, Peng GH, Chen S. Regulation of photoreceptor gene expression by Crx-associated transcription factor network. *Brain Res.* 2008; 1192:114–133. [PubMed: 17662965]
- Huang QY, Li GH, Cheung WM, Song YQ, Kung AW. Prediction of osteoporosis candidate genes by computational disease-gene identification strategy. *J Hum Genet.* 2008; 53:644–655. [PubMed: 18463784]
- Jakobsdottir J, Conley YP, Weeks DE, Mah TS, Ferrell RE, Gorin MB. Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am J Hum Genet.* 2005; 77:389–407. [PubMed: 16080115]
- John, GH.; Langley, P. Estimating continuous distributions in Bayesian classifiers. Morgan Kaufmann Publishers Inc.; San Francisco, CA: 1995. p. 338-345.
- Keen TJ, et al. A YAC contig spanning the dominant retinitis pigmentosa locus (RP9) on chromosome 7p. *Genomics.* 1995; 28:383–388. [PubMed: 7490071]
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. *Science.* 2005; 308:385–389. [PubMed: 15761122]
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10:R25. [PubMed: 19261174]
- Le Cessie S, Van Houwelingen J. Ridge estimators in logistic regression. *Appl Stat.* 1992; 41:191–201.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2009; 25:1754–1760. [PubMed: 19451168]
- Lombard Z, Tiffin N, Hofmann O, Bajic VB, Hide W, Ramsay M. Computational selection and prioritization of candidate genes for fetal alcohol syndrome. *BMC Genomics.* 2007; 8:389. [PubMed: 17961254]
- Martinez-Mir A, et al. A new locus for autosomal recessive retinitis pigmentosa (RP19) maps to 1p13–1p21. *Genomics.* 1997; 40:142–146. [PubMed: 9070931]
- McGee TL, Devoto M, Ott J, Berson EL, Dryja TP. Evidence that the penetrance of mutations at the RP11 locus causing dominant retinitis pigmentosa is influenced by a gene linked to the homologous RP11 allele. *Am J human Genetics.* 1997; 61:1059–1066. [PubMed: 9345108]
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297–1303. [PubMed: 20644199]
- Moriyama EN, Powell JR. Gene length and codon usage bias in *drosophila melanogaster*, *saccharomyces cerevisiae* and *escherichia coli*. *Nucleic Acids Res.* 1998; 26:3188–3193. [PubMed: 9628917]
- Myktyyn K, Braun T, Carmi R, Haider NB, Searby CC, Shastri M, Beck G, Wright AF, Iannaccone A, Elbedour K, Riise R, Baldi A, et al. Identification of the gene that, when mutated, causes the human obesity syndrome BBS4. *Nat Genet.* 2001; 28:188–191. [PubMed: 11381270]
- Myktyyn K, Nishimura DY, Searby CC, Shastri M, Yen HJ, Beck JS, Braun T, Streb LM, Cornier AS, Cox GF, Fulton AB, Carmi R, et al. Identification of the gene (BBS1) most commonly involved in Bardet–Biedl syndrome, a complex human obesity syndrome. *Nat Genet.* 2002; 31:435–438. [PubMed: 12118255]
- Nishimura DY, Searby CC, Carmi R, Elbedour K, Van Maldergem L, Fulton AB, Lam BL, Powell BR, Swiderski RE, Bugge KE, Haider NB, Kwitek-Black AE, et al. Positional cloning of a novel gene on chromosome 16q causing Bardet–Biedl syndrome (BBS2). *Hum Mol Genet.* 2001; 10:865–874. [PubMed: 11285252]
- Nishimura DY, Swiderski RE, Searby CC, Berg EM, Ferguson AL, Hennekam R, Merin S, Weleber RG, Biesecker LG, Stone EM, Sheffield VC. Comparative genomics and gene expression analysis identifies BBS9, a new Bardet–Biedl syndrome gene. *Am J Hum Genet.* 2005; 77:1021–1033. [PubMed: 16380913]

- Norton N, Robertson PD, Rieder MJ, Zuchner S, Rampersaud E, Martin E, Li D, Nickerson DA, Hersherberger RE. Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era. *Circ Cardiovasc Genet*. 2012; 5:167–174. [PubMed: 22337857]
- Ott J, et al. Localizing multiple X chromosome-linked retinitis pigmentosa loci using multilocus homogeneity tests. *Proc Nat Acad Sci USA*. 1990; 87:701–704. [PubMed: 2300556]
- Piro RM, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J*. 2012; 279:678–696. [PubMed: 22221742]
- Platt J. Sequential minimal optimization: a fast algorithm for training support vector machines. 1998
- Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif*. 1999; 10:61–74.
- Rasche A, Al-Hasani H, Herwig R. Meta-analysis approach identifies candidate genes and associated molecular networks for type-2 diabetes mellitus. *BMC Genomics*. 2008; 9:310. [PubMed: 18590522]
- Rebello G, et al. Apoptosis-inducing signal sequence mutation in carbonic anhydrase IV identified in patients with the RP17 form of retinitis pigmentosa. *Proc Nat Acad Sci USA*. 2004; 101:6617–6622. [PubMed: 15090652]
- Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012; 337:64–69. [PubMed: 22604720]
- Thiselton DL, et al. Mapping the RP2 locus for X-linked retinitis pigmentosa on proximal Xp: a genetically defined 5-cM critical region and exclusion of candidate genes by physical mapping. *Genome Research*. 1996; 6:1093–1102. [PubMed: 8938433]
- Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*. 2011; 27:1986–1994. [PubMed: 21576180]
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009; 25:1105–1111. [PubMed: 19289445]
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010; 28:511–515. [PubMed: 20436464]
- Tucker BA, Scheetz TE, Mullins RF, DeLuca AP, Hoffmann JM, Johnston RM, Jacobson SG, Sheffield VC, Stone EM. Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene *male germ cell-associated kinase (MAK)* as a cause of retinitis pigmentosa. *Proc Natl Acad Sci USA*. 2011; 108:E569. [PubMed: 21825139]
- Tuson M, Marfany G, Gonzalez-Duarte R. Mutation of CERKL, a novel human ceramide kinase gene, causes autosomal recessive retinitis pigmentosa (RP26). *Am J Human Genetics*. 2004; 74:128–138. [PubMed: 14681825]
- Vladimir, VN.; Vapnik, V. The nature of statistical learning theory. Springer-Verlag New York, Inc.; New York, NY.: 1995.
- Xu SY, Denton M, Sullivan L, Daiger SP, Gal A. Genetic mapping of RP1 on 8q11-q21 in an Australian family with autosomal dominant retinitis pigmentosa reduces the critical region to 4 cM between D8S601 and D8S285. *Human Genetics*. 1996; 98:741–743. [PubMed: 8931712]
- Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008; 9:R137. [PubMed: 18798982]
- Zuchner S, Dallman J, Wen R, Beecham G, Naj A, Farooq A, Kohli MA, Whitehead PL, Hulme W, Konidari I, Edwards YJ, Cai G, et al. Whole-exome sequencing links a variant in *DHDDS* to retinitis pigmentosa. *Am J Hum Genet*. 2011; 88:201–206. [PubMed: 21295283]

Figure 1.

Box plots of feature values used in our algorithm. **A:** Feature ranges are highly dissimilar before scaling and centering. Some data are not displayed on this figure due to their extreme values. **B:** Features here are centered more evenly, and have similar, smaller ranges. Extreme outliers (over 10 standard deviations from the mean) are not displayed on this figure.

Figure 2.

A workflow for ranking genes according to how similar they are to RetNet genes. **A:** We collect and preprocess quantitative and informative features for each gene from multiple experiments, including RNA-seq, ChIP-seq, and microarrays in retina. **B:** For each gene, we remove that gene from the full set, and (**C**) train a probabilistic classifier on the remaining set. **D:** We then use the trained model to test the removed gene and obtain a probability that it exhibits our criterion. **E:** We collect the output probability for each gene and append it to a list of the probabilities from all N genes. **F:** We sort the probability list in descending order and rank the genes from 1 to N based on this ordering. This is the disease-likelihood ranking for each gene.

Figure 3. The distribution of Monte Carlo trials and the performance of various probabilistic ranking techniques. A smaller rank sum indicates a set with a greater enrichment for low-ranked RetNet genes. The histogram represents the rank sum frequencies of 1 million trials.

Table 1

Comparison of Logistic Regression Technique and ENDEAVOUR

Interval ID	Logistic regression rank	Logistic regression adjusted rank	ENDEAVOUR rank	ENDEAVOUR adjusted rank	Publication
INT000031	1 (1/31)	3.23	3 (1/32)	9.38	Banerjee et al. (1998)
INT000032	1 (2/20)	5.00	11 (2/18)	61.11	Ott et al. (1990)
INT000033	13 (1/24)	54.17	7 (1/23)	30.43	Thiselton et al. (1996)
INT000034	1 (2/67)	1.49	4 (2/67)	5.97	Bareil et al. (2001)
INT000035	23 (1/145)	15.86	6 (1/130)	4.62	Gu et al. (1997)
INT000036	3 (1/21)	14.29	3 (1/22)	13.64	Gerber et al. (2000)
INT000037	1 (2/39)	2.56	12 (1/38)	31.58	Chen et al. (1996)
INT000038	1 (1/50)	2.00	11 (1/46)	23.91	Martinez-Mir et al. (1997)
INT000039	4 (1/10)	40.00	10 (1/10)	100.00	Tuson et al. (2004)
INT000040	1 (2/14)	7.14	5 (2/12)	41.67	Keen et al. (1995)
INT000041	9 (2/41)	21.95	17 (2/34)	50.00	Rebello et al. (2004)
INT000042	1 (1/14)	7.14	1 (1/15)	6.67	Xu et al. (1996)
INT000043	22 (1/39)	56.41	18 (1/41)	43.90	McGee et al. (1997)
Totals	81	231.24	108	422.87	
<i>P</i> values	0.00012		0.00428		

Column rankings of genetic intervals as identified in RetNet. Rank columns are read as: Highest rank RetNet gene (number of RetNet genes in list/ number of genes ranked). Adjusted ranks are calculated as $100 \times (\text{Highest Rank RetNet gene} / \text{Number of Genes Ranked})$. Adjusted ranks are used to normalize the rankings obtained by each algorithm by the number of genes that were prioritized by the algorithm.