JAMIA

Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants.

Journal:	*Journal of the American Medical Informatics Association*
Manuscript ID:	amiajnl-2011-000309.R1
Article Type:	Brief Communication
Keywords:	gene variant classification, phenotype prediction, machine learning, amino acid properties, missense variants

Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants.

David K. Crockett[1,2], Elaine Lyon[2], Marc S. Williams[1,3], Scott P. Narus[1], Julio C. Facelli[1,4], Joyce A. Mitchell[1]

[1] Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT

[2]Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT

[3]Intermountain Healthcare Clinical Genetics Institute, Salt Lake City, UT

[4] University of Utah Center for High Performance Computing, Salt Lake City, UT

Running Title: Utility of gene-specific predictors

Correspondence:

David K. Crockett

ARUP Laboratories

500 Chipeta Way

Salt Lake City, Utah 84132

Tel: 801-583-2787

Fax: 801-584-5109

Email: david.crockett@aruplab.com

Keywords: amino acid properties, gene variant classification, machine learning, phenotype prediction

Word count: 2122

Abstract

The rapid advance of gene sequencing technologies has produced an unprecedented rate of discovery for genome variation in humans. A growing numbered of authoritative clinical repositories archive gene variants and disease phenotype, yet there are currently many more gene variants that lack clear annotation or disease association. To date, there has been very limited coverage of gene-specific predictors in the literature. Here we present the evaluation of "gene-specific" predictor models based on a Naïve Bayesian classifier for 20 gene-disease data sets, containing 3,986 variants with clinically characterized patient conditions. Utility of gene-specific prediction is then compared "all-gene" generalized prediction and also to existing popular predictors. Gene-specific computational prediction models derived from clinically curated gene variant disease data sets often outperform established generalized algorithms for novel and uncertain gene variants.

Background and Significance

Personalized medicine implies that all relevant clinical information is available on demand for effective patient treatment. Proper interpretation of gene test results is a key component in customizing patient therapy. Efforts such as the Human Variome Project, 1000 Genomes and NCBI Genetic Testing Registry highlight a growing interest in annotation and clinical interpretation of gene variants in human disease.(1-3) As genetic information is incorporated into the electronic medical record, new decision support approaches are needed to provide clinicians with a preferred course of treatment.(4) For decision support rules to add value, the clinical relevance of laboratory information must be well understood.(5, 6)

Furthermore, with rapidly evolving technologies such as SNP chip genome wide association studies and next-generation sequencing, genomic analysis is trending faster and cheaper and yielding much larger data sets. As such, gene variants are being discovered at an almost astronomical pace, with one recent report finding an average of 3 million variants per personal genome.(7) More importantly, for genomic variation to be of real clinical utility, laboratory interpretation and disease association must be well understood for each new gene variant found. (8, 9)

Unfortunately, an increasingly apparent gap exists between rapidly growing collections of genetic variation and practical clinical implementation. Although collections of human genome variation have been underway for years, authoritative repositories of gene variants with clear association to disease phenotype are only now beginning to emerge.(10-14) This is in contrast to existing collections of genome-wide mutations such as dbSNP(15) or OMIM(16) that are not curated using consistent, systematic or transparent methods. Focusing computer predictive algorithms on authoritative and specific gene-disease settings has the potential to bridge this knowledge gap.

Prediction algorithms for computing mutation severity have been used for many years.(17-20) Despite their use in laboratories, they do not have sufficient accuracy to predict disease phenotype to the degree necessary to be clinically applicable. This prompts opportunities to explore the application of advanced informatics approaches to this problem.(21-23) This study expands the recently reported Primary Sequence Amino Acid Properties (PSAAP) algorithm (24, 25), which uses a gene-specific classification approach utilizing amino acid physicochemical properties of the primary amino acid sequence to predict pathogenicity of novel and/or uncertain gene variants. To date, gene-specific approaches have been applied only to the *RET* proto-oncogene and hypertrophic cardiomyopathy.(25, 26)

To evaluate the generalizability of our gene-specific PSAAP algorithm, we extend its use to a set of 20 genes with clinically curated disease variants (Table 1). The analyses also compare the effectiveness of generic gene versus gene specific approaches using a minimum (non-redundant) set of amino acid properties to describe exonic non-synonymous variants coupled with evaluation of overlap and/or trends of biochemical properties of mutation.

Methods

Gene variant data relating well-characterized patient condition to genotype (genotype-phenotype) were assembled from multiple sources including: cystic fibrosis mutation database curated by Ruslan Dorfman (Hospital for Sick Children, Toronto)(27); BioPKU database curated by Nenad Blau (University Children's' Hospital, Zurich)(28); neurofibromatosis type 1 database curated by Ophélia Maertens (Center for Medical Genetics, University Hospital, Ghent) and Collagen, type IV, alpha 5 (*COL4A5*) Mental Retardation Database curated by Judy Savige (Department of Medicine, University of Melbourne) as

hosted by Leiden Open Source Variation Database (LOVD)(29-31); biotinidase (*BTD*) curated by Barry Wolf (Medical Genetics, Henry Ford Hospital, Detroit)(32); aryl hydrocarbon receptor interacting protein (*AIP*) curated by Rodrigo Toledo (Endocrine Genetics Unit, University of Sao Paulo Medical School) (personal communication); Disease Databases hosted by Department of Pathology, University of Utah School of Medicine(33) and genetic testing results archived at ARUP Laboratories (Salt Lake City). The clinically curated gene-disease data sets (n=20) containing some 3986 curated variants are summarized in Table 1.

This 20 gene collection contained 1639 exonic non-synonymous SNP's (nsSNP) with known outcomes of benign (n=607) and pathogenic (n=1032). The gene variants were characterized using physicochemical properties of the substituted amino acid as recently reported.(24, 25) Briefly, gene-specific clinically curated missense variants (nsSNP's) were characterized using a Naïve Bayes classification scheme of primary amino acid sequence only and delta differences in physical, chemical, conformational, or energetic properties between the amino acid present in the wild type and the variant. Descriptors were attributes derived from 544 amino acid properties archived in AAindex v9.4.(34) AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids. For each gene variant, vectors of delta values for each biochemical property of the substituted amino acid were calculated and the resulting mutation described by an array of variables, corresponding to the absolute value of the difference between wild type and mutant – as trained in a gene-specific setting.

Based on curated clinical outcomes of benign or pathogenic, the minimum (non-redundant) set of amino acid properties needed to describe pathogenicity of gene variants was investigated using various attribute selection methods such as correlation-based feature subset selection, SVM-RFE and Relief-F and various classifiers. Thresholds of 95% (or 0.95) for Greedy-Stepwise and Ranker were used during

this analysis. The best performing correlation-based feature subset selection and Naïve Bayes classification was implemented using the Weka software package.(35)

For each of the 20 genes, random selection was used to build a 2/3 training and a 1/3 test sets with known class labels (benign, pathogenic). Training and test sets were to keep the original ratio of benign and pathogenic constant, but without regard to functional motif or protein location. Next, based on curated clinical classification of benign or pathogenic, algorithm training and pathogenicity prediction was performed gene-by-gene. Gene-specific models were also tested for prediction of other gene-disease outcomes, by using the training set of one gene and a test set from a second gene. In a similar fashion, an "all-gene" model was constructed using all the available training sets. This "all-gene" model was then tested by making gene-by-gene predictions. Due to a low number of nsSNP exonic substitution variants, five genes (*MECP2*, *MSH2*, *MSH6*, *PLOD1* and *SPINK1*) were only included in the all-gene training set, and not used for gene-specific training. Algorithm performance was evaluated using each gene test set, with sensitivity (true positive rate), specificity (true negative rate), and positive predictive value (PPV or precision) calculated for each classifier algorithm and gene-specific and all-gene permutations.

Well established prediction tools such as PolyPhen (18) and SIFT (17) are primarily based on multiple alignment and amino acid substitution penalties have been available for many years. More recently, MutPred (20) which calculates probability of deleterious mutations by disrupted molecular mechanism. Additionally, PMut (19) is neural net based and trained on human mutations. (A more detailed description of each prediction algorithm is given in Supplementary Data.) Lastly, gene-specific algorithm performance was compared to well established prediction algorithms such as SIFT(17), PolyPhen(18),

PMUT(19) and MutPred(20). Comparison of established prediction tools with gene-specific trained algorithms may increase our understanding of predicting mutation status.

For all genes, the full length protein isoform was used for this study. Splice variants were not considered. All gene variants were mapped to their reference amino acid sequence from UniProtKB (http://www.uniprot.org). Protein reference sequences are summarized in Supplementary Table 1.

Results and Discussion

Overall, the performance of the gene-specific trained algorithm was significantly better (8% to 13%) than the "all-gene" model, with p values of 0.00001 (sensitivity), 0.00113 (specificity) and 0.00012 (PPV) as shown in Figure 1. For the genes evaluated, the PPV of our gene-specific PSAAP algorithm averaged 89% (82% to 94%). This was on average 11% higher than the "all-gene" model where PPV ranged from 62% to 86%. The one exception was *SLC22A5*, where PPV remained constant. Sensitivity averaged 13% higher than the "all-gene" model, except for *SPRED1* which was 6% decreased. Specificity was also generally improved (9% average) for all but *PMS2* (no increase) and *NF1*, which was 5% decreased.

For the genes studied here, the PSAAP gene-specific prediction performs well. PPV values are displayed in Supplementary Table 2. The self against self is plotted on the diagonal in blue with ppv>80 bolded. Other gene predictor performance with PPV above 80 is shaded in orange. Interestingly, gene-specific prediction models do not seem to generalize well – even across similar protein functional families. For instance, Supplementary Table 2 shows that the *RET* kinase trained model (94% PPV) performed lower for the *ACVRL1* kinase (84% PPV) while the *ACVRL1* trained predictor (88% PPV) only predicted *RET* with 80% PPV. Additionally, the carboxylase enzyme *BTD* (91% PPV) only predicted the hydroxylase *PAH* gene variant outcome with 76% PPV, while the *PAH* trained predictor (89% PPV) only predicted *BTD* with

59% PPV. It is notable however, that 3 out of 15 genes (*SPRED1*, *NF1 and GALT*) yielded comparable numbers for predicting disease association across other genes.

The improved performance of gene-specific algorithms may be explained in part by an important observation that biochemical and/or structural characteristics of mutation specific to one disease may be lost or diluted when combined with large genome-wide data sets for algorithm development. This can be illustrated by plotting non-synonymous variants specific to a gene-disease condition as compared to random amino acid substitutions. When 1000 random amino acid changes were plotted (Supplementary Figure 1A), a wide distribution evenly covers the entire range of possible substitutions. In contrast, when 1000 pathogenic mutations are graphed, characteristic trends of specific residues and frequency of substitution are readily seen (Supplementary Figure 1B). More importantly, disease-specific examples of this concept are shown in Figure 2. In the *RET* proto-oncogene (associated with medullary thyroid cancers), some 79% of all pathogenic changes were found to involve cysteine (C) to some other residue (X) as displayed in Figure 2A. In the *COL4A5* gene (associated with Alport syndrome), 84% of pathogenic changes involve glycine (G) to other residues (X) as shown in Figure 2B. To confirm this trend, further experiments should be performed as additional curated gene-disease collections become available.

Although the majority of the PSAAP models did not perform as well for predicting pathogenicity in other genes-diseases, most still outperformed established algorithms. As shown in Table 2, a majority of genes (13 out of 15) analyzed using the gene-specific PSAAP trained algorithm had improved PPV as compared to other algorithms, with the overall PPV increasing 8.8% to 22.0%. For example, the PSAAP model specific for *SPRED1* (93% PPV as seen in Table 2), when analyzed using established prediction algorithms yielded precision scores from 56% to 71%. As mentioned above, the PSAAP model specific

for *RET* kinase (94% PPV) underperformed for the *ACVRL1* kinase (84% PPV), however, both still outperformed established algorithms, where on-line predictions for *ACVRL1* only ranged from 57% to 81% PPV. Two exceptions to this trend were *GALT* and *SMAD*, in which MutPred and/or PMut scored slightly higher as shown bolded/underlined in Table 2.

It is important to note that the all-gene trained Bayes predictor also compares favorably to established algorithms, with the average, minimum and maximum PPV for each predictor also summarized in Table 2. For instance, although the gene-specific trained PSAAP model yielded the best PPV, the all-gene trained model outscores 3 of 4 established predictors, with MutPred being the exception. This observation may highlight the importance of authoritative variant data and amino acid physicochemical properties being used to develop/train algorithms. It also demonstrates that primary acid sequence only, when coupled with amino acid properties, can be successfully used to develop predictor algorithms.

Finally, a minimum attribute set of amino acid properties seems specific to each gene-disease, with overlap found among different genes using three feature selection methods ranging from 11% to 80% as summarized in Supplementary Table 3. Representative examples are shown in Figure 3. Interestingly, the gene models with more shared amino acid attributes (*GALT*, 80%; *NF1*, 62%; *SPRED1*, 60%) also had the best generalizability. Of note, both *SMAD4* and *GALT* did well using the established on-line prediction tools, where *SMAD4* also had 58% overlap. Without considering the above mentioned 4 genes, the overlap ranged from only 11% to 37%. Overlap for the all gene data set follows this same trend, showing only 38% overlap between the feature selection methods.

Conclusion

The number of authoritative disease and locus specific gene variant collections in use for clinical diagnostics is rapidly growing. These clinically-curated gene variant data sets, with reliable genotype-phenotype association, can readily be utilized for training and test set performance of machine classifiers. The generalizability of classification rules across multiple genes and diseases may be strengthened as the number of curated disease variants continues to increase, although our analysis suggests that gene-specific approaches will, with few exceptions, outperform generic approaches. Nonetheless, the recognition that the proposed classifier outperforms existing tools is important, given that it will take time for disease-specific curated genotype-phenotype databases to be developed and for some ultra-rare diseases such databases may never be realistic.

For machine learning classifiers, amino acid attributes characteristic of substitution mutations for a given disease may be lost or diluted when combined with multiple genes and diseases. A key distinguishing feature of this gene-specific classifier methodology is that algorithms are trained explicitly to curated monogenic disease outcomes. While this methodology is complementary to established generalized prediction tools, algorithms should take advantage of authoritative (clinically-curated) gene variant collections where they exist. This is especially important when pathologic variants exhibit characteristic trends or properties specific to a given disease.

This study included only gene variant collections with clearly documented disease association and known to the authors – and represents the largest collections to-date of clinically curated gene-disease results as used for diagnostic and gene test reporting purposes. Although correlation of genotype-phenotype offers therapeutic options that would otherwise remain hidden and may lead to disease specific mutation-guided management strategies, appropriate caution is justified when clinicians are asked to trust computational outcomes for determining patient care.(36) Continued interaction

between clinicians and laboratorians to refine mutation-specific clinical classification is imperative to optimal patient care.(5, 6)

Acknowledgements

The authors gratefully acknowledge the extensive disease curation of gene variants by Drs. Dorfman, Blau, Maertens, Savige, Wolf, Toledo and others.

Competing Interests

Authors have no competing interests to declare.

Funding

This work has been supported by ARUP Institute for Clinical and Experimental Pathology, National Library of Medicine Training grant [grant number #LM007124] and NCRR Clinical and Translational Science Award [grant number #1KL2RR025763-01].

References

1. Cotton RG, Al Aqeel AI, Al-Mulla F, et al. Capturing all disease-causing mutations for clinical and research use: toward an effortless system for the Human Variome Project. Genet Med. 2009 Dec;**11**(12):843-9.

2. Durbin RM, Abecasis GR, Altshuler DL, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;**467**(7319):1061-73.

3. Javitt G, Katsanis S, Scott J, Hudson K. Developing the blueprint for a genetic testing registry. Public Health Genomics. 2010;**13**(2):95-105.

4. Hoffman MA. The genome-enabled electronic medical record. J Biomed Inform. 2007 Feb;**40**(1):44-6.

5. Hoffman MA, Williams MS. Electronic medical records and personalized medicine. Hum Genet. 2011 Jul;**130**(1):33-9.

6. Marshall E. Human genome 10th anniversary. Human genetics in the clinic, one click away. Science. 2011 Feb 4;**331**(6017):528-9.

7. Moore B, Hu H, Singleton M, Reese MG, Yandell M. Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole-genome-based clinical diagnostics Genet Med. 2011:In Press.

8. Li C. Personalized medicine - the promised land: are we there yet? Clin Genet. 2010 Dec 13.

9. Guttmacher AE, McGuire AL, Ponder B, Stefansson K. Personalized genomic information: preparing for the future of genetic medicine. Nat Rev Genet. 2010 Feb;**11**(2):161-5.

10. Thony B, Blau N. Mutations in the BH4-metabolizing genes GTP cyclohydrolase I, 6-pyruvoyl-tetrahydropterin synthase, sepiapterin reductase, carbinolamine-4a-dehydratase, and dihydropteridine reductase. Hum Mutat. 2006 Sep;**27**(9):870-8.

11. Calderon FR, Phansalkar AR, Crockett DK, Miller M, Mao R. Mutation database for the galactose-1-phosphate uridyltransferase (GALT) gene. Hum Mutat. 2007 Oct;**28**(10):939-43.

12. Margraf RL, Crockett DK, Krautscheid PM, et al. Multiple endocrine neoplasia type 2 RET protooncogene database: repository of MEN2-associated RET sequence variation and reference for genotype/phenotype correlations. Hum Mutat. 2009 Apr;**30**(4):548-56.

13. Crockett DK, Pont-Kingdon G, Gedge F, Sumner K, Seamons R, Lyon E. The Alport syndrome COL4A5 variant database. Hum Mutat. 2010 Aug;**31**(8):E1652-7.

14. Li W, Sun L, Corey M, et al. Understanding the population structure of North American patients with cystic fibrosis. Clin Genet. 2011 Feb;**79**(2):136-46.

15. Single Nucleotide Polymorphism Database. [ncbi.nlm.nih.gov/projects/SNP/].

16. Online Mendelian Inheritance in Man. [ncbi.nlm.nih.gov/omim/].

17. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;**31**(13):3812-4.

18. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002 Sep 1;**30**(17):3894-900.

19. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics. 2005 Jul 15;**21**(14):3176-8.

20. Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009 Nov 1;**25**(21):2744-50.

21. Tavtigian SV. Comparison of programs for in silico assessment of missense substitutions. Hum Mutat. 2011 Jun;**32**(6):v.

22. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat. 2011 Apr;**32**(4):358-68.

23. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011 Apr 8;**88**(4):440-9.

24. Crockett DK, Piccolo SR, Narus SP, Mitchell JA, Facelli JC. Computational feature selection and classification of *RET* phenotypic severity. J Data Mining in Genom Proteomics 2010 16 Dec **1**(2):1-4.

25. Crockett DK, Piccolo SR, Ridge PG, et al. Predicting phenotypic severity of uncertain gene variants in the *RET* proto-oncogene. PLoS One. 2011;**6**(3):e18380.

26. Jordan DM, Kiezun A, Baxter SM, et al. Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. Am J Hum Genet. 2011 Feb 11;**88**(2):183-92.

27. Cystic Fibrosis Mutation Database. [http://www.genet.sickkids.on.ca/cftr/app].

28. BIOPKU:International Database of Patients and Mutations causing BH4-responsive HPA/PKU. [http://www.biopku.org/biopku/] Sep 2010.

29. Neurofibromatosis Type 1 Database - germline. [http://medgen.ugent.be/LOVD2/home.php?select_db=NF1_germline].

30. Neurofibromatosis Type 1 Database - somatic. [http://medgen.ugent.be/LOVD2/home.php?select_db=NF1_somatic].

31. Collagen, type IV, alpha 5 (COL4A5) Mental Retardation Database [http://www.LOVD.nl/COL4A5].

32. Biotinidase Deficiency and BTD database. [http://www.arup.utah.edu/database/BTD/BTD_welcome.php].

33. ARUP online scientific resource, disease databases. [http://www.arup.utah.edu/database/index.php].

34. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res. 2000 Jan 1;**28**(1):374.

35. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004 Oct 12;**20**(15):2479-81.

36. Tchernitchko D, Goossens M, Wajcman H. In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. Clin Chem. 2004 Nov;**50**(11):1974-8.

Figure Legends

Figure 1. Performance of the gene-specific PSAAP algorithm as compared to all-gene algorithm plotted to show A) sensitivity, B) specificity and C) positive predictive value (PPV). Significance was calculated using a 2 tailed paired t-test.

Figure 2. Disease specificity of pathogenic mutations demonstrated by plotting A) the *RET* proto-oncogene variants where 79% of pathogenic changes are cysteine [C] to another residue [X] and B) *COL4A5* where 84% pathogenic changes are glycine [G] to another residue [X] again showing characteristic trends of specific residues and frequency of substitution that may be lost when diluting gene-specific data into genome wide computational methods.

Figure 3. Venn diagram showing overlap of amino acid properties to characterize benign and pathogenic gene variants using three feature selection methods (CfsSubset, Relief-F, SVM-RFE). Overlap for A) *RET* with only 14% shared attributes, B) *GALT* with a much higher 80% overlap and C) the all-gene data set with only 38% shared attributes.

Table 1. Summary of clinically-curated gene variant data sets (n=20) with known disease association.

Gene Symbol *Biological Function*	Gene Name *Disease Association*	Curated Variants	Exonic nsSNPs
ACVRL1 activin receptor activity, type 1	activin A receptor type II-like 1 *hereditary hemorrhagic telangiectasia*	332	192
AIP transcription coactivator activity	aryl hydrocarbon receptor interacting protein *familial pituitary adenoma*	102	84
BTD biotin carboxylase activity	biotinidase *biotinidase deficiency*	155	105
CFTR chloride channel regulator activity	cystic fibrosis transmembrane conductance regulator *cystic fibrosis*	252	121
COL4A5 extracellular matrix structural constituent	collagen, type IV, alpha 5 *X-linked Alport syndrome (hereditary nephritis)*	600	266
ENG TGF β receptor activity	endoglin *hereditary hemorrhagic telangiectasia*	397	124
GALT uridylyltransferase activity	galactose-1-phosphate uridylyltransferase *galactosemia*	247	168
GJB2 gap junction channel activity	gap junction protein, beta 2 (connexin 26) *hereditary sensorineural hearing loss*	61	43
MECP2 transcription co-repressor activity	methyl CpG binding protein 2 *Rett syndrome*	26	14
MSH2 guanine/thymine mispair binding	mutS homolog 2 *hereditary nonpolyposis colonrectal cancer*	89	8
MSH6 guanine/thymine mispair binding	mutS homolog 6 *hereditary nonpolyposis colonrectal cancer*	34	10
NF1 Ras GTPase activator activity	neurofibromin 1 *neurofibromatosis type 1*	125	121
PAH phenylalanine catabolism	phenylalanine hydroxylase *phenylketonuria (PKU)*	730	126
PLOD1 procollagen-lysine-dioxygenase activity	procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 1 *Ehlers-Danlos syndrome type VI*	34	12
PMS2 mismatched DNA binding	postmeiotic segregation increased 2 *hereditary nonpolyposis colorectal cancer*	348	45
RET transmembrane receptor kinase activity	ret proto-oncogene *multiple endocrine neoplasia, medullary thyroid carcinoma*	146	97
SLC22A5 carnitine transporter activity	solute carrier family 22, member 5 *primary carnitine deficiency*	95	57
SMAD4 transcription activator activity	SMAD family member 4 *juvenile polyposis syndrome, pancreatic cancer*	86	23
SPINK1 endopeptidase inhibitor activity	serine peptidase inhibitor, Kazal type 1 *hereditary pancreatitis*	73	5
SPRED1 inactivation of MAPK activity	sprouty-related, EVH1 domain containing 1 *Legius syndrome (neurofibromatosis type-like syndrome)*	54	18

Table 2. Gene-specific and all-gene algorithm PPV as compared to established algorithms.

Gene	PSAAP[a]	All-gene[b]	SIFT[c]	PolyPhen[d]	PMut[e]	MutPred[f]
ACVRL1	**88**	77	57	67	69	81
AIP	**91**	71	71	73	80	79
BTD	**91**	79	77	72	71	87
CFTR	**90**	63	68	74	70	89
COL4A5	**88**	82	58	74	62	73
ENG	**92**	83	62	64	73	65
GALT	86	77	66	65	58	**87**
GJB2	**87**	77	69	74	67	83
NF1	**89**	70	64	70	70	84
PAH	**89**	80	59	76	77	84
PMS2	**88**	63	64	74	74	72
RET	**94**	84	78	54	72	84
SLC22A5	**90**	82	74	76	53	82
SMAD4	84	82	71	70	85	**86**
SPRED1	**93**	86	71	65	56	71
(avg	**89.3**	77.1	67.3	69.9	69.1	80.5)
(min	**84.0**	63.0	57.0	54.0	53.0	65.0)
(max	**94.0**	86.0	78.0	76.0	85.0	89.0)

[a] Primary Sequence Amino Acid Properties (PSAAP) algorithm, gene-specific trained.
[b] Primary Sequence Amino Acid Properties (PSAAP) algorithm, all-gene (n=20) trained.
[c] Analyzed with default settings at http://sift.jcvi.org.
[d] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[e] Analyzed with default settings at http://mmb.pcb.ub.es/PMut.
[f] Analyzed with default settings at http://mutdb.org/mutpred.

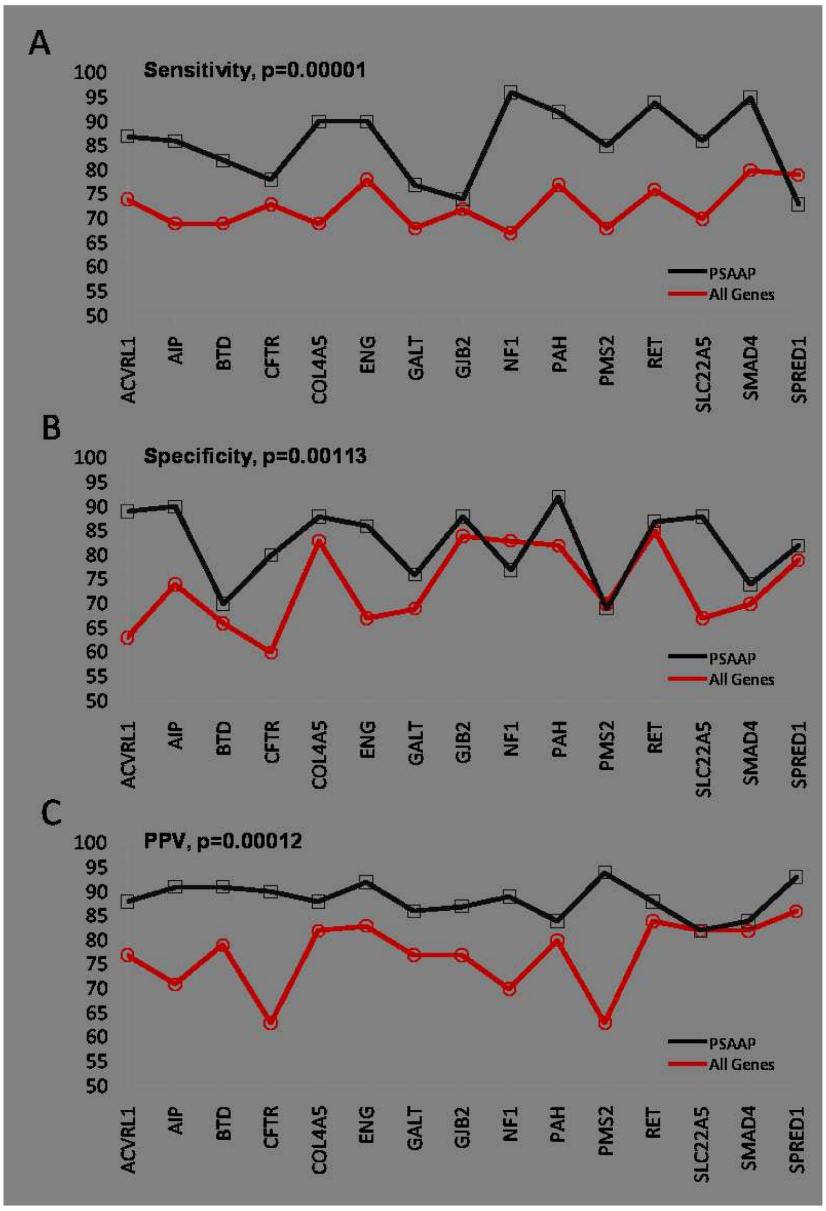


Figure 1

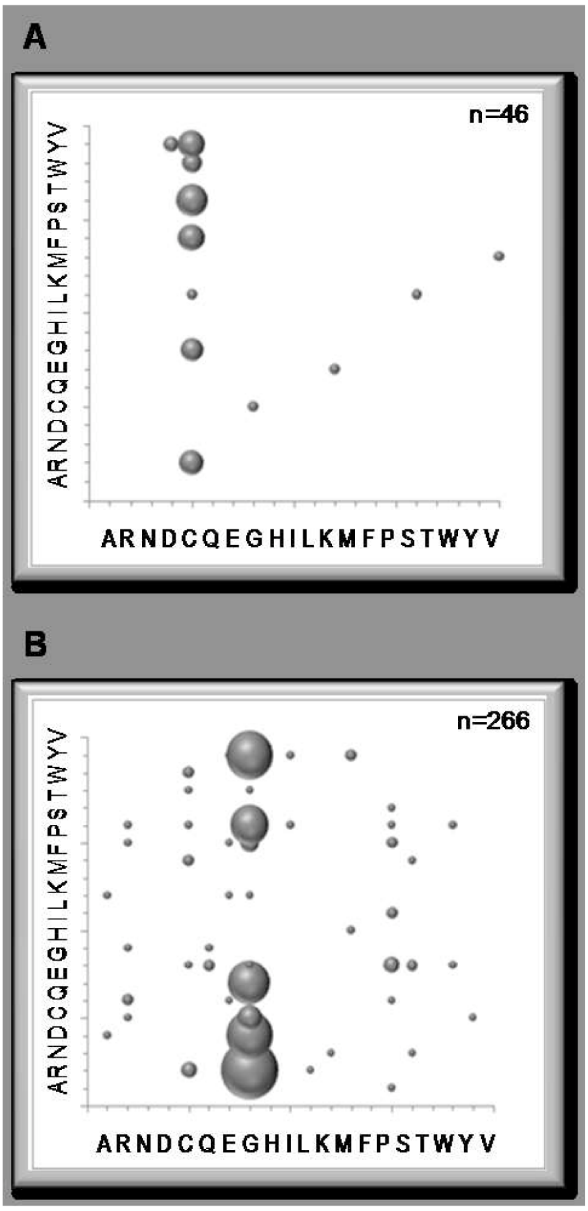


Figure 2

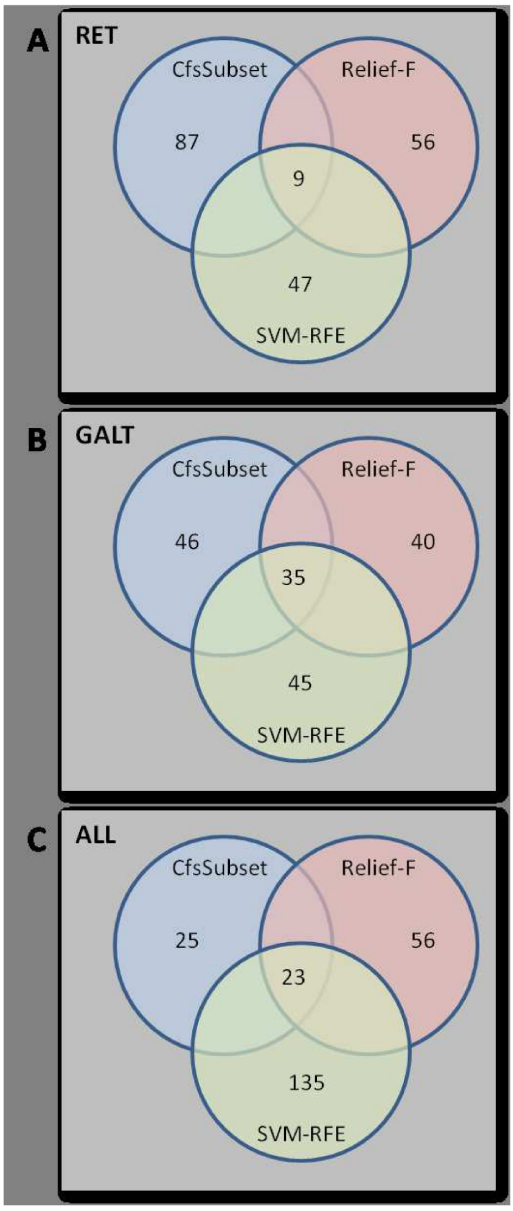


Figure 3

Supplementary Table 1. Reference amino acid sequence from UniProtKB[a].

Gene symbol	UniProt #	Protein name	AA length	Date accessed
ACVRL1	P37023	ACVL1_HUMAN	503	December 6, 2010
AIP	O00170	AIP_HUMAN	330	January 5, 2011
BTD	P43251	BTD_HUMAN	543	December 6, 2010
CFTR	P13569	CFTR_HUMAN	1480	December 6, 2010
COL4A5	P29400	CO4A5_HUMAN	1685	December 7, 2010
ENG	P17813	EGLN_HUMAN	658	December 7, 2010
GALT	P07902	GALT_HUMAN	379	December 7, 2010
GJB2	P29033	CXB2_HUMAN	226	December 7, 2010
MECP2	P51608	MECP2_HUMAN	486	December 7, 2010
MSH2	P43246	MSH2_HUMAN	934	December 8, 2010
MSH6	P52701	MSH6_HUMAN	1360	December 8, 2010
NF1	P21359	NF1_HUMAN	2839	January 5, 2011
PAH	P00439	PH4H_HUMAN	452	January 6, 2011
PLOD1	Q02809	PLOD1_HUMAN	727	December 9, 2010
PMS2	P54278	PMS2_HUMAN	862	December 9, 2010
RET	P07949	RET_HUMAN	1114	December 9, 2010
SLC22A5	O76082	S22A5_HUMAN	557	December 9, 2010
SMAD4	Q13485	SMAD4_HUMAN	552	January 7, 2011
SPINK1	P00995	ISK1_HUMAN	79	December 9, 2010
SPRED1	Q7Z699	SPRE1_HUMAN	444	December 9, 2010

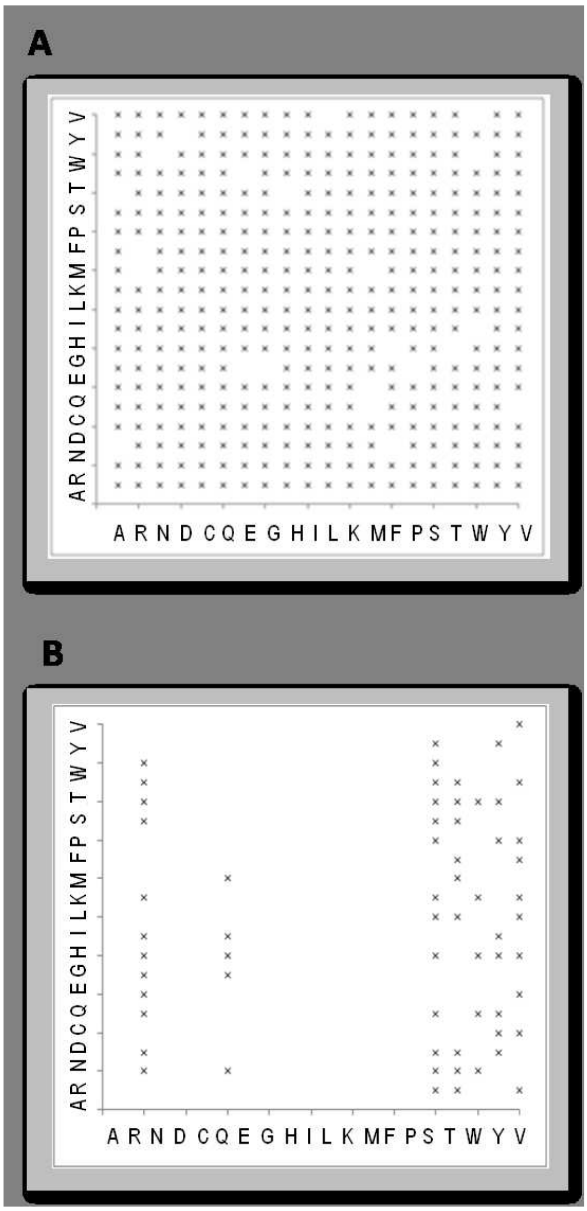
[a] http://www.uniprot.org.

Supplementary Table 2. PPV of gene-specific algorithms to predict pathogenicity in other genes.

	ACVRL1	AIP	BTD	CFTR	COL4A5	ENG	GALT	GJB2	NF1	PAH	PMS2	RET	SLC22A5	SMAD4	SPRED1
ACVRL1	**88**	**83**	74	70	**84**	77	79	79	**85**	74	76	**80**	**81**	72	78
AIP	72	**91**	62	62	69	59	66	55	68	57	65	63	62	58	62
BTD	77	79	**91**	77	**85**	73	**82**	**81**	**85**	76	70	70	71	**81**	**85**
CFTR	53	62	56	**90**	56	54	59	55	51	54	47	60	53	57	61
COL4A5	47	58	62	51	**88**	**83**	55	61	52	57	46	56	57	56	50
ENG	48	47	62	57	**84**	**92**	49	55	51	56	50	60	54	60	61
GALT	**83**	**82**	**85**	**80**	77	74	**86**	77	**80**	**81**	**85**	**80**	**81**	77	**84**
GJB2	67	56	73	54	56	70	73	**87**	55	66	69	64	62	56	71
NF1	**90**	76	**84**	75	**90**	**89**	75	79	**89**	**83**	75	73	78	**81**	**84**
PAH	62	74	59	55	63	58	64	60	**82**	**89**	58	71	65	60	59
PMS2	66	62	63	61	61	70	55	69	62	71	**88**	66	70	63	56
RET	**84**	69	62	42	64	57	46	72	66	72	45	**94**	49	68	59
SLC22A5	74	66	63	73	72	71	69	68	73	70	68	72	**82**	71	**81**
SMAD4	49	53	65	61	49	64	47	53	67	67	56	52	64	**84**	67
SPRED1	**82**	**85**	**85**	**87**	**87**	**87**	**80**	**84**	**81**	**83**	77	**86**	**84**	**80**	**93**

Supplementary Table 3. Overlap of minimum set of amino acid properties describing disease association.

	CfsSubset	Relief-F	SVM-RFE	Overlap
ACVRL1	7	39	49	7
AIP	90	29	117	25
BTD	41	20	39	8
CFTR	19	161	139	12
COL4A5	63	65	88	21
ENG	13	82	59	9
GALT	46	40	45	35
GJB2	11	37	145	11
NF1	28	20	39	18
PAH	29	73	129	24
PMS2	13	58	107	11
RET	87	56	47	9
SLC22A5	76	96	87	13
SMAD4	63	65	88	42
SPRED1	59	44	31	27
All GENE	**25**	**56**	**135**	**23**



Supplementary Figure 1. Specificity of pathogenic mutations demonstrated by plotting A) simulated random amino acid substitutions (n=1000) showing a wide distribution that evenly covers the entire range of possible substitutions and B) known pathogenic mutations (n=1000) showing characteristic trends of specific residues and frequency of substitution.