# CADD score has limited clinical validity for the identification of pathogenic variants in non-coding regions in a hereditary cancer panel

**Cheryl A Mather**[1,†], **Sean D Mooney**[2], **Stephen J Salipante**[1], **Sheena Scroggins**[1], **David Wu**[1], **Colin C Pritchard**[1], and **Brian H Shirts**[1,*]

[1]Department of Laboratory Medicine, University of Washington, Seattle, Washington, USA

[2]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA

## Abstract

**PURPOSE**—Several *in silico* tools have been shown to have reasonable research sensitivity and specificity for classifying sequence variants in coding regions. The recently-developed Combined Annotation Dependent Depletion (CADD) method generates predictive scores for single nucleotide variants (SNVs) in all areas of the genome, including non-coding regions. We sought to determine the clinical validity of non-coding variant CADD scores.

**METHODS**—We evaluated 12,391 unique SNVs in 624 patient samples submitted for germline mutation testing in a cancer-related gene panel. We compared the distributions of CADD scores of rare SNVs, common SNVs in our patient population, and the null distribution of all possible SNVs stratifying by genomic region.

**RESULTS**—The median CADD scores of intronic and nonsynonymous variants were significantly different between rare and common SNVs (p<0.0001). Despite these different distributions, no individual variants could be identified as plausibly causative among rare intronic variants with the highest scores. The ROC AUC for non-coding variants is modest, and the positive predictive value of CADD for intronic variants in panel testing was found to be 0.088.

**CONCLUSION**—Focused *in-silico* scoring systems with much higher predictive value will be necessary for clinical genomic applications.

## Keywords

Combined Annotation-Dependent Depletion; CADD score; *in silico* predictor; predictive algorithm; non-coding sequences; introns

*Corresponding Author:** Brian Shirts, University of Washington, Department of Laboratory Medicine, NW120, Box 357110, 1959 NE Pacific Street, Seattle, WA 98195-7110, Tel: 206.598.0557, Fax: 206.598.6189, shirtsb@uw.edu.
†Present address: Department of Pathology, University of California, San Francisco, 185 Berry Street, San Francisco, CA 94007

## Introduction

Multi-gene testing of cancer susceptibility is widely applied in clinical practice to attempt to predict the risk of developing cancer. Estimating the effect of DNA variants in these large gene panels is a major clinical challenge. As it is impractical to functionally classify every variant identified, several *in silico* tools have been developed to predict the pathogenicity of single nucleotide variants (SNVs). Many of these tools focus on protein-coding regions of the genome (summarized in 1). However, the number of non-coding variants far outstrips coding variants in the human genome (2, 3), and approximately 88% of trait/disease-associated SNVs in collective genome wide association studies are in intronic or intergenic regions (4). The Combined Annotation-Dependent Depletion (CADD) method is designed to predict the pathogenicity of SNVs at any location in the genome. Kircher et al. (5) described the receiver operating characteristics (ROC) curves of CADD scores for curated, pathogenic mutations defined by the ClinVar database, and showed that a CADD score has a greater area under the curve (AUC) than GerpS, PhCons and phyloP scores for a set of defined variants. They also examined two enhancers and one promoter in which saturation mutagenesis had been previously performed, and showed that CADD had the highest Spearman rank correlation between the predictive score and the observed changes in protein expression (5, Supplementary Figure 17 of that reference), stating that CADD provides, "in principle, a genome-wide, data-rich, functionally generic and organismally relevant estimate of variant effect" (5). Based on claims of genome-wide relevant estimates of variant effect, we sought to test the clinical validity of CADD scores by comparing their distributions in common and rare variants identified in 624 patients tested in our large cancer-risk gene panel, with specific attention on non-exonic variants that did not alter protein coding or canonical splice sites. We evaluated the rare variants with the highest CADD scores in these non-coding regions where CADD score distributions were significantly different than expected. We also explored the hypothetical sensitivity and specificity cutoffs that would be required to achieve meaningful clinical positive or negative predictive values.

## Materials and Methods

### Samples

We evaluated a total of 624 consecutively submitted, unique DNA samples clinically requested for germline cancer susceptibility testing using the University of Washington (UW) BROCA assay (6) between June 2014 and February 2015. All variant data was de-identified prior to release to the investigators in this study. De-identified minimal cancer and family history phenotypes were retained with the data to aid in interpretation of potential variant significance. This project was deemed non-human subjects research consistent with ongoing quality improvement and assurance activities as a component of clinical testing.

### Targeted deep sequencing by BROCA

Library construction, gene capture and massively parallel sequencing were performed using clinical ColoSeq and BROCA assays as previously described (6) and detailed online (http://web.labmed.washington.edu/tests/COLOSEQ, http://tests.labmed.washington.edu/BROCA). Briefly, DNA was sonicated, purified and subjected to end repair, A-tailing and ligation to

Illumina paired-end adapters. The adapter-ligated library was amplified, and individual paired-end libraries were hybridized to a custom design of complementary RNA biotinylated oligonucleotides spanning all exons and non-repetitive intronic regions spanning 49 genes (Supplementary Table 3). The library-bait hybrids were purified and washed. Each library was amplified by PCR using primers with a unique index. After amplification, libraries were quantified, and equimolar concentrations were pooled, denatured, and cluster amplified on a single lane of an Illumina flow cell. Sequencing was performed with $2 \times 101$-bp paired-end reads and a 7-bp index on a HiSeq 2000 (Illumina Inc, San Diego, CA). Mean sequencing depth was over 100 for all samples.

We used a custom targeted sequencing bioinformatics pipeline (7). Reads were mapped to human reference genome 19 (hg19, GRCh37), and alignment was performed using BWA and SAMtools. SNV calling was performed with GATK and VarScan. The entire pipeline was validated and shown to have >99.9% accuracy for single nucleotide changes (7).

### Variant curation

Variant evaluation was limited to probable germline mutations, defined as SNVs with variant read fraction >30%. For this project, rare variants were defined as those identified at a minor allele frequency of less than 1% by the 1000 Genomes Project (1KG; 8). All variants with computed CADD scores were included in the analysis.

### Statistical analysis

Distribution of variant scaled CADD scores was compared for three variant types: rare variants in patient samples, common variants in patient samples, and all possible variants as defined by Kircher et al in Supplementary Table 8 (5). We further grouped variants by genomic region to determine if CADD performed effectively in different genomic contexts. Genomic regions were defined using ANNOVAR (9) as: downstream, intronic, intergenic nonsynonymous, splice site, synonymous, stopgain, upstream, 3' untranslated region (UTR) and 5' UTR. We compared the sample median using the Wilcoxon Rank-Sum test to evaluate the significance of differences.

As we were testing 30 comparisons, we chose a p value of 0.001 as our cutoff for significance. We calculated ratios between the proportion of variants in a group with a given CADD score and plotted these to visually evaluate differences in CADD score for different groups. Statistical tests were performed using built-in R functions (10).

### Evaluation of validity of CADD scores for non-coding variants

To evaluate possible causative variants we used several criteria to narrow the list of variants of interest (VOI) for further analysis. 1) Rare variants had to be in genes broadly consistent with the patient phenotype. For example, we excluded rare variants in known breast cancer risk genes if they were only seen in patients with personal history of colorectal cancer (or patients with a family history excluding breast cancer for those without a personal cancer history). For SNVs that were present in multiple samples, variants were considered if the majority of those patients had cancer phenotypes consistent with the gene mutated. 2) If a pathogenic mutation consistent with patient phenotype was present, other rare variants for

that patient were considered unlikely to be causative and excluded. 3) The variant base was compared to the reference base in up to 100 vertebrate species (the default of the UCSC genome browser [11]). If the variant base was present as the reference base in any of the species for which data was available, the variant was excluded. Remaining variants after this step were considered VOI.

In order for CADD scores to be clinically useful in non-coding regions, we first expect the distribution of CADD scores for rare variants to be different from the null distribution of CADD scores, particularly for variants with high CADD scores. We evaluated rare variants with the highest 10% of CADD scores for intronic variants to determine if these variants might possibly explain patient disease phenotypes. We used the pROC program in R (12) to create a ROC curve for the results of this analysis. The PhyloP score (13) for pairwise alignment of 100 vertebrate species (the default of the UCSC genome browser) was also calculated for the 10% of intronic variants with the highest CADD scores (PhyloP, Supplementary Table 3). Each VOI was evaluated along with the 50 bases proceeding and the 50 bases following the variant base using the Berkeley Drosophila Genome Project (14), Human Splicing Finder 3.0 (15), and NetGene (16) splice site prediction algorithms to predict changes in splice sites along the transcribed strand (Supplementary Table 4).

## Modeling of sensitivity and specificity needed to achieve clinically acceptable identification of possible pathogenic variants

For an *in silico* predictive tool to be clinically useful, it must either rule out benign variants with high certainty or identify pathogenic variants with modest certainty to minimize the necessary follow-up functional or co-segregation studies necessary to definitively classify variants. For our practice, we determined that an optimal rule-out predictor would have at least 95% negative predictive value (NPV) consistent with accepted definitions of what constitutes a likely benign variant (17,18). Because of the extensive work that is required to confirm a pathogenic variant, we desire at least a 50% positive predictive value (PPV) to minimize unnecessary follow up of unknown variants.

For given sensitivity and specificity, PPV and NPV vary with the prevalence pathogenic mutations within the set of variants evaluated. We calculated the sensitivity and specificity required to achieve a minimum PPV of 50% and a minimum NPV of 95% using approximate representative mutation prevalence estimates: 50% for evaluation of coding elements in a single gene, 10% for evaluation of coding and non-coding elements of a single gene, 5% for evaluation of coding elements in panel testing, 0.5% for coding and non-coding elements in panel testing, 0.05% for exome testing, and 0.0001% for genome testing (3, 19-24).

## Results

### Comparison of CADD score distribution between rare, common and all possible variants

We identified 12,391 unique SNVs with computed scaled CADD scores in the 624 patient samples. The specific number of variants in downstream, intergenic, intronic,

nonsynonymous, splicing, synonymous, upstream, 3' UTR and 5' UTR regions is summarized in Supplementary Table 1.

We compared rare, common and all possible variants in each category to each other using the Wilcoxon Rank-Sum test. There were statistically significant differences between common and all possible variants for intergenic, nonsynonymous, and upstream SNVs (Supplementary Table 2). As shown in Supplementary Figure 1, when the proportion of common variants at any given CADD score was graphed over the proportion of all possible variants at that score for these significant regions, nonsynonymous variants with CADD scores less than ten were significantly overrepresented in the common variants ($p=4.8\times10^{-14}$). This is consistent with the hypothesis that common nonsynonymous variants have been subject to evolutionary selection and are thus enriched for benign variants. Surprisingly, there was an upward trend from the lowest to the highest CADD scores for the intergenic ($p=2.6\times10^{-5}$) and upstream variants ($p=1.9\times10^{-12}$). This suggests that high CADD score variants in these regions are more likely to occur in our patient samples than would be expected by chance.

There were statistically significant differences between rare and all possible variants for downstream, intergenic, intronic, upstream and 5' UTR SNVs (Supplementary Table 2). As shown in Supplementary Figure 2, when graphed over the proportion of all possible variants at each possible CADD score for these significant regions, we found that rare downstream variants with CADD scores greater than 15 were overrepresented compared to all possible variants ($p=6\times10^{-8}$), as were intronic variants with CADD scores greater than 25 ($p=2.2\times10^{-16}$) and 5' UTR variants with CADD scores greater than 17 ($p=2.2\times10^{-16}$). Rare variants in these regions, therefore, were more likely to have high CADD scores than would be expected by chance. Rare intergenic variants with CADD scores less than four were underrepresented ($p=1.3\times10^{-11}$), as were rare upstream variants with CADD scores less than five ($p=2.2\times10^{-16}$) and rare 5' UTR variants with CADD scores less than six ($p=2.2\times10^{-16}$). This means that rare variants with low CADD scores are statistically less frequent in our patient population than would be expected by chance. These findings are consistent with the hypothesis that rare variants have not been subjected to extensive selective pressure and are more likely to be functionally deleterious. Comparing rare and common variants, there were statistically significant differences for intronic and nonsynonymous variants (Supplementary Table 2). Graphing the proportion of common variants over the proportion of rare variants at each possible CADD score for these significant regions (Figure 1) revealed that SNVs with higher CADD scores were proportionally underrepresented for common intronic variants when compared with rare variants ($p=5\times10^{-6}$). Common SNVs with CADD scores below six were proportionally overrepresented for nonsynonymous variants when compared with rare variants ($p=5\times10^{-11}$). These findings are consistent with the hypothesis that rare variants from these regions are more likely to be deleterious than common ones and thus are more likely to have high CADD scores.

### Evaluation of validity of CADD scores for non-coding variants

If CADD scores are to have clinical validity for the identification of novel pathogenic variants in non-coding regions, then the subset of rare variants with the highest CADD

scores in genomic regions with significantly different CADD scores between rare and common variants should be enriched for pathogenic variants. As the only non-coding region that had statistically different CADD scores between rare and common variants were introns, we specifically looked at the 10% of rare intronic variants with the highest CADD scores to evaluate whether these mutations could possibly cause disease in our patient population. Two hundred eighty-six of 690 variants evaluated were in genes not known to cause the type of cancer found in the patient or patient's family and were thus excluded. Thirty-eight of the 404 remaining rare variants were seen in patients with other known pathogenic variants, and were thus considered unlikely to cause the phenotype in those patients. Three hundred and five of the remaining 366 variants were present as the conserved base in one or more of the vertebrate species evaluated by MULTIZ alignment of up to 100 vertebrate species, which was used as evidence that there was no functional consequence to the variant. This left 61 VOI.

There was no significant enrichment of VOI as the CADD score cutoff increased. Forty-two of 517 variants with CADD scores between 10.51 and 14.99 (8.1%) were VOI. Sixteen of 145 variants with CADD scores between 15 and 19.99 (11%) were VOI, and for variants with CADD scores 20 (28 total), there were 3 VOI (10.7%). We plotted the ROC curve (Figure 2a) of VOI over all rare variants in the CADD score range examined to determine whether there was an optimal cutoff at which CADD score identified the most VOI (highest sensitivity) with the highest specificity. The area under the curve was 0.591 (95% confidence interval 0.516-0.667), and there was no CADD cutoff at which sensitivity and specificity were optimized. The PPV of CADD score to identify VOI at a score 10.51 was 8.8%.

In an effort to identify pathogenic mutations, we used three splice-site prediction algorithms (NNSplice, Human Splice Finder 3.0 and NetGene) to evaluate the possibility that splice-site changes on the transcribed strand caused by VOI introduced alternative splice sites. There were no variants predicted to introduce novel splice sites by all three prediction algorithms tested (Supplementary Table 4). NNSplice and HSF3.0 splice predictions were consistent for five variants, and three of these showed a 15% or greater increase in splicing score for both predictions. However, the frequency of these variants in our overall clinical sample set was not significantly higher than the reported variant frequency in the 1KG dataset, and the clinical histories of other individuals with these variants evaluated outside this study were not consistent with the gene of interest, suggesting that these variants are unlikely to substantially alter disease risk.

No definitively pathogenic variants were present in our cohort, and thus we were unable to robustly measure the false negative rate (FNR). For this reason, we also computed the ROC of non-coding variants using data from Figure 3 of the paper by Kircher et al. (5), a dataset that contained well-characterized pathogenic variants (Figure 2b). The AUC of this ROC curve was 0.663 (95% CI 0.607-0.720), which was not significantly better than the one from our data.

We further evaluated CADD scores for pathogenic intronic variants using 47 deep intronic variants reported to be deleterious in the literature (Supplementary Table 5; 25-40). The CADD scores for these variants ranged from 0.356-19.05, with a median of 3.498. Only five

of the 47 variants (10.6%) had CADD scores >10.51, the cutoff at which we evaluated variants for possible pathogenicity. More than 50% of all rare intronic variants in our dataset had CADD scores greater than the median for known pathogenic variants (3.498), consistent with low AUC. To evaluate the comparative usefulness of CADD scores compared to other *in silico* predictive algorithms for intronic regions, we compared the performance of PhyloP scores from 100 vertebrate species to that of CADD scores for the identification of VOI. The AUC for the ROC curve of the PhyloP analysis was 0.666 (95% CI 0.593-0.739), which is similar to that seen for CADD scores (Figure 2a). However, as our definition of VOI included evaluation of the conservation of the base between species, we cannot exclude ascertainment bias.

### Sensitivity and specificity needed to achieve clinically acceptable identification of possible pathogenic variants

For maximal clinical validity, a predictive tool should identify benign variants with confidence while minimizing the number of variants that require further work-up. We calculated the minimum sensitivity and/or specificity needed for a predictor to achieve a positive predictive value of at least 50% and a negative predictive value of at least 95% (Table 1). If pathogenic mutations represent 0.5% of all rare mutations (our estimate for the prevalence of pathogenic mutations in coding and non-coding regions of genes in panel testing), then the required minimum specificity for an *in silico* tool to identify pathogenic mutations at this level of confidence is 99.5%, though the sensitivity of the tool is less critical at that mutation prevalence (Table 1). The sensitivity required for an *in silico* prediction tool to meet clinically meaningful negative predictive value requirements increases as the prevalence of pathogenic mutations increases, whereas the specificity become less critical for overall performance. On the other hand, if the number of genes and rare variants tested increases, the specificity required to achieve a clinically meaningful positive predictive value increases and the sensitivity becomes less critical for overall performance.

## Discussion

The comparison of the CADD scores of common and rare variants in different genomic areas with the CADD scores of all possible variants in these areas as defined by Kircher et al. (5) suggest that CADD scores may have modest predictive power for nonsynonymous variants. We found that in a clinical population, common nonsynonymous variants have significantly lower CADD scores than those produced by random mutations, whereas the distribution of CADD scores for nonsynonymous rare variants is no different than the null distribution. This suggests that CADD scores may correlate with functional consequences, as common nonsynonymous variants, which are likely to be functionally benign, having gone through extensive natural selection, have lower CADD scores.

For both intergenic and upstream variants, common variants with low CADD scores were proportionally underrepresented while those with higher CADD scores were proportionately overrepresented compared to all possible variants in these regions, which was unexpected. If CADD scores are representative of evolutionary selection, this suggests evolutionary

pressure supporting promoter diversity. Alternatively, this could represent a statistical anomaly due to the fact that our panel does not cover a large proportion of the intergenic and upstream regions present in the human genome. For downstream, intronic, intergenic, upstream and 5' UTR regions, the rare variants with higher CADD scores are overrepresented compared to those expected by chance, which is the pattern one would expect if these variants were more likely to be functionally deleterious, having not undergone extensive selection.

Differences in observed distributions between CADD scores for common and rare intronic and nonsynonymous variants suggest that CADD scores may be useful for identification of pathogenic intronic or nonsynonymous variants in targeted testing situations when used in combination with other data. However, our data suggests that CADD scores are unlikely to be useful for identifying disease causing mutations in other non-coding regions in cancer risk genes. Evaluation of the 10% of rare intronic variants with the highest CADD scores revealed 61 variants in genes consistent with patient presentation out of 690 examined. Additional investigation suggested that these variants are unlikely to cause substantial disease risk (Supplementary Table 3). The absence of any convincing pathogenic or likely pathogenic variants in our clinical dataset was a major limitation of our analysis. For this reason, we also evaluated the non-coding data from the original paper used to describe CADD scoring (5), which had known positive variants as well as a set of 47 previously reported pathogenic deep intronic variants. The ROC curve for non-coding variants from the original Kircher et al. dataset and the one generated from our dataset are similar, supporting our conclusion about the very low positive predictive value of CADD score for non-coding variants. This conclusion is further supported by our evaluation of known pathogenic intronic variants from the literature.

For an *in silico* predictive tool to be useful in clinical interpretation of unique variants, should have high negative predictive value (NPV) to avoid missing truly pathogenic variants and moderate positive predictive value (PPV) to minimize further clinical workup. Our analysis of PPV and NPV in different clinical situations suggests that for a single gene test in which 50% of the identified rare variants are pathogenic , the required sensitivity of a predictor must be very high (94.8%) to achieve an appropriately high NPV, but the required specificity is low. Given the reported sensitivity and specificity of CADD in this scenario from the work of Kircher et al. (5), it is possible that a CADD score cutoff value for nonsynonymous mutations could approach this level of sensitivity. The number of variants in non-coding regions is higher, however, and there is a lower density of pathogenic mutations in non-exonic regions. In our patient population, for example, there were more than eight times as many rare variants in non-coding regions as there were in coding regions and splice sites (Supplementary Table 1). Evaluating the ROC curves generated for non-coding variants using the data from Kircher et al. (5, Figure 2b) in the context of PPV, it becomes clear that there is no cutoff at which CADD score is clinically useful for non-exonic variants. Additionally, if more genes are added to a panel (thereby increasing the number of rare coding or non-coding variants to evaluate), the gap between the current performance of CADD and the performance required for clinical usefulness increases. We thus conclude that while CADD scores are "in principle, a genome-wide, data-rich, functionally generic and organismally relevant estimate of variant effect" (5), in clinical practice for hereditary cancer

panels (or, likely, to any larger genomic test) lack predictive power. There may be situations where CADD scores or other *in silico* scores can be combined with other predictors to produce clinically useful predictions; these combined analysis situations will need to be evaluated separately to determine how much CADD scores independently improve predictions.

Another issue in interpreting CADD scores (or other predictive scores) is the distinction between changes that are functionally deleterious and that are clinically pathogenic. The underlying data used for CADD scores are evolutionary and functional predictors. There are many situations where a deleterious variant does not cause clinical phenotype. This separation between functional prediction and clinical consequence reduces the real-world predictive value of predictive scores.

Although sensitivity and specificity of CADD have been shown to be high in datasets balanced for known pathogenic and benign variants, sensitivity and specificity are test values that are agnostic to population prevalence. The real-world positive predictive value of CADD score and other *in silico* tests is not high enough to effectively classify individual non-exonic variants or reduce the number of potential pathogenic variants to those that could be efficiently followed up in the context of a hereditary cancer panel. This finding supports the idea that currently available *in silico* predictive scores should be used at most as supporting evidence of pathogenicity, as is currently recommended by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (Richards, 2015).

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

## References

1. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011; 12(9):628–40. [PubMed: 21850043]

2. 1000 Genomes Project Consortium. Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–73. [PubMed: 20981092]

3. Pelak K, Shianna KV, Ge D, et al. The characterization of twenty sequenced human genomes. PLoS Genet. 2010; 6(9):e1001111. [PubMed: 20838461]

4. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009; 106(23):9362–7. [PubMed: 19474294]

5. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46(3):310–5. [PubMed: 24487276]

6. Pritchard CC, Smith C, Salipante SJ, et al. ColoSeq provides comprehensive lynch and polyposis syndrome mutational analysis using massively parallel sequencing. J Mol Diagn. 2012; 14(4):357–66. [PubMed: 22658618]

7. Pritchard CC, Salipante SJ, Koehler K, et al. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J Mol Diagn. 2014; 16(1):56–67. [PubMed: 24189654]

8. 1000 Genomes Project Consortium. Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. [PubMed: 23128226]

9. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164. [PubMed: 20601685]

10. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2014. URL http://www.R-project.org/

11. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res. 2002; 12(6):996–1006. [PubMed: 12045153]

12. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77. [PubMed: 21414208]

13. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2009; 20:110–21. [PubMed: 19858363]

14. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. Comp Biol. 1997; 4(3):311–23.

15. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acid Res. 2009; 37(9):e67. 2009. [PubMed: 19339519]

16. Brunak S, Engelbrecht J, Knudsen S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. Journal of Mol Biol. 1991; 220:49–65. [PubMed: 2067018]

17. Plon SE, Eccles DM, Easton D, et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat. 2008; 29(11):1282–91. [PubMed: 18951446]

18. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015; 17(5):405–24. [PubMed: 25741868]

19. Newman B, Mu H, Butler LM, Millikan RC, Moorman PG, King MC. Frequency of breast cancer attributable to BRCA1 in a population-based series of American women. JAMA. 1998; 279(12):915–21. [PubMed: 9544765]

20. Ng PC, Levy S, Huang J, et al. Genetic variation in an individual human exome. PLoS Genet. 2008; 4(8):e1000160. [PubMed: 18704161]

21. Ramus SJ, Gayther SA. The contribution of BRCA1 and BRCA2 to ovarian cancer. Mol Oncol. 2009; 3(2):138–50. [PubMed: 19383375]

22. Kim H, Choi DH. Distribution of BRCA1 and BRCA2 mutations in Asian patients with breast cancer. J Breast Cancer. 2013; 16(4):357–65. [PubMed: 24454456]

23. Foley SB, Rios JJ, Mgbemena VE, et al. Use of Whole Genome Sequencing for Diagnosis and Discovery in the Cancer Genetics Clinic. EBioMedicine. 2015; 2(1):74–81. [PubMed: 26023681]

24. Tung N, Battelli C, Allen B, et al. Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. Cancer. 2015; 121(1):25–33. [PubMed: 25186627]

25. Hübner CA, Utermann B, Tinschert S, et al. Intronic mutations in the L1CAM gene may cause X-linked hydrocephalus by aberrant splicing. Hum Mutat. 2004; 23(5):526. [PubMed: 15108295]

26. Gámez-Pozo A, Palacios I, Kontic M, et al. Pathogenic validation of unique germline intronic variants of RB1 in retinoblastoma patients using minigenes. Hum Mutat. 2007; 28(12):1245.

27. Takeshima Y, Yagi M, Okizuka Y, et al. Mutation spectrum of the dystrophin gene in 442 Duchenne/Becker muscular dystrophy cases from one Japanese referral center. J Hum Genet. 2010; 55(6):379–88. [PubMed: 20485447]

28. Castoldi E, Duckers C, Radu C, et al. Homozygous F5 deep-intronic splicing mutation resulting in severe factor V deficiency and undetectable thrombin generation in platelet-rich plasma. J Thromb Haemost. 2011; 9(5):959–68. [PubMed: 21320286]

29. Kulseth MA, Lyle R, Rødningen OK, Sorte H, Prescott T. Exon trapping analysis of c.301-19G > A in intron 1 of the SHH gene in a patient with a microform of holoprosencephaly. Eur J Med Genet. 2011; 54(2):130–5. [PubMed: 21044704]

30. Meeths M, Chiang SC, Wood SM, et al. Familial hemophagocytic lymphohistiocytosis type 3 (FHL3) caused by deep intronic mutation and inversion in UNC13D. Blood. 2011; 118(22):5783–93. [PubMed: 21931115]

31. Richards AJ, McNich A, Whittaker J, et al. Splicing analysis of unclassified variants in COL2A1 and COL11A1 identifies deep intronic pathogenic mutations. Eur J Hum Genet. 2012; 20(5):552–8. [PubMed: 22189268]

32. Spier I, Horpaopan S, Vogt S, et al. Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. Hum Mutat. 2012; 33(7):1045–50. [PubMed: 22431159]

33. Cavalieri S, Pozzi E, Gatti RA, Brusco A. Deep-intronic ATM mutation detected by genomic resequencing and corrected in vitro by antisense morpholino oligonucleotide (AMO). Eur J Hum Genet. 2013; 21(7):774–8. [PubMed: 23211698]

34. Constantino L, Rusconi D, Soldá G, et al. Fine characterization of the recurrent c.1584+18672A>G deep-intronic mutation in the cystic fibrosis transmembrane conductance regulator gene. Am J Respir Cell Mol Biol. 2013; 48(5):619–25. [PubMed: 23349053]

35. Steele-Stallard HB, Le Quesne Stabej P, Lenassi E, et al. Screening for duplications, deletions and a common intronic mutation detects 35% of second mutations in patients with USH2A monoallelic mutations on Sanger sequencing. Orphanet J Rare Dis. 2013; 8:122. [PubMed: 23924366]

36. Bholah Z, Smth MJ, Byers HJ, Miles EK, Evans DG, Newman WG. Intronic splicing mutations in PTCH1 cause Gorlin syndrome. Fam Cancer. 2014; 13(3):477–80. [PubMed: 24659465]

37. Bonifert T, Karle KN, Tongael F, et al. Pure and syndromic optic atrophy explained by deep intronic OPA1 mutations and an intralocus modifier. Brain. 2014; 137(8):2164–77. [PubMed: 24970096]

38. Bach JE, Wolf B, Oldenburg J, Müller CR, Rost S. Identification of deep intronic variants in 15 haemophilia A patients by next generation sequencing of the whole factor VIII gene. Thromb Haemost. 2015; 114(4):757–67. [PubMed: 25948085]

39. Liquori A, Vaché C, Baux D, et al. Whole USH2A Gene Sequencing Identifies Several New Deep Intronic Mutations. Hum Mutat. 2015 epub ahead of print.

40. Palagano E, Blair HC, Pangrazio A, et al. Buried in the Middle but Guilty: Intronic Mutations in the TCIRG1 Gene Cause Human Autosomal Recessive Osteopetrosis. J Bone Miner Res. 2015; 30(10):1814–21. [PubMed: 25829125]
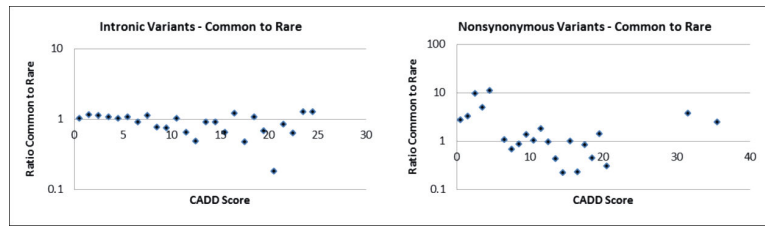
**Figure 1.**
Ratio of common to rare variants with significant differences by Wilcoxin Rank Sum test. The proportion of common variants at any given CADD score was compared to that of rare variants at the same CADD score (rounded to the nearest 1). Only genomic regions with significant differences by Wilcoxin Rank Sum test were evaluated graphically.
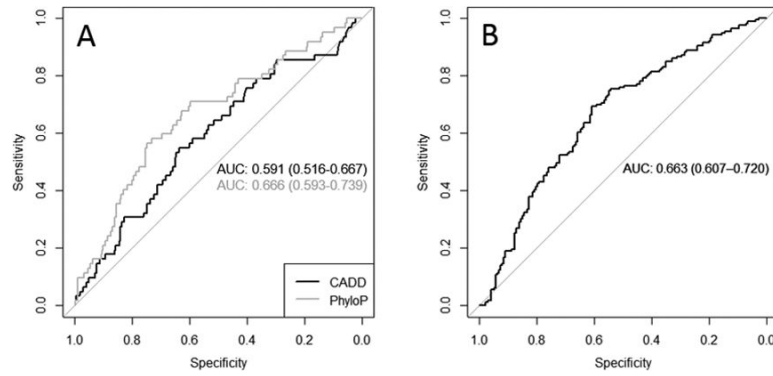
**Figure 2. Receiver-operating characteristics (ROC) curves for non-coding variants**
**A.** ROC curve for CADD score (black) and 100 vertebrate PhyloP score (grey) for variants of interest in the top 10% of rare intronic variants. **B.** ROC curve for CADD score for non-coding variants from Kircher et al (2014) source data.

**Table 1**

Minimum sensitivity and specificity of an *in silico* predictive tool needed for clinical validity.

| Estimated percent of rare variants that are pathogenic | Clinical test example | Minimum sensitivity | Minimum specificity |
|---|---|---|---|
| 50% | Coding sequence, single gene | 0.948 | NA |
| 10% | Coding and non-coding sequences, single gene | 0.526 | 0.888 |
| 5% | Coding sequence, 25-50 gene panel | NA | 0.947 |
| 0.5% | Coding and non-coding sequences, 25-50 gene panel | NA | 0.995 |
| 0.05% | Exome | NA | 0.9995 |
| 0.001% | Whole genome | NA | 0.999999 |

The minimum sensitivity and/or specificity to achieve a positive predictive value (PPV) of at least 50% and a negative predictive value (NPV) of at least 95% was calculated depending on the number of clinically important rare variants as a fraction of the total rare variants (prevalence of clinically important rare variants). For each prevalence value, we have listed an example of a clinical test type that might be expected to have clinically important variants at that prevalence.