nature genetics

M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity

Karthik A Jagadeesh[1,5], Aaron M Wenger[2,5], Mark J Berger[1], Harendra Guturu[2], Peter D Stenson[3], David N Cooper[3], Jonathan A Bernstein[2] & Gill Bejerano[1,2,4]

Variant pathogenicity classifiers such as SIFT, PolyPhen-2, CADD, and MetaLR assist in interpretation of the hundreds of rare, missense variants in the typical patient genome by deprioritizing some variants as likely benign. These widely used methods misclassify 26 to 38% of known pathogenic mutations, which could lead to missed diagnoses if the classifiers are trusted as definitive in a clinical setting. We developed M-CAP, a clinical pathogenicity classifier that outperforms existing methods at all thresholds and correctly dismisses 60% of rare, missense variants of uncertain significance in a typical genome at 95% sensitivity.

Whole-exome sequencing for the diagnosis of Mendelian diseases is now a routine part of medical genetics clinical practice[1–3], but variant interpretation remains a serious challenge[4,5]. It is common practice to classify variants as benign purely on the basis of allele frequency in a control population in comparison to disease frequency[5]. Filtering candidate variants using allele frequency in controls has been shown to be very powerful[6–8]. The American College of Medical Genetics (ACMG) recommends that any variant with an allele frequency greater than 5% in controls be classified as benign[5], and it is common practice to tighten the threshold to 1% or lower to reduce the number of variants of uncertain significance to a manageable number[9,10].

Even after such frequency-based filtering, probands generally have between 200 and 500 missense and truncating variants that are not present in databases of control individuals[9–11]. Most of these variants do not satisfy the criteria proposed by the ACMG for classification as definitively benign or pathogenic, and they are hence termed variants of uncertain significance (VUS)[5]. Further filtering approaches are used to reduce the list of VUS to a number that is tractable for manual expert review. It is critical to avoid both overfiltering, which excludes causative variants, and underfiltering, which leaves an impractical number of variants to review[12]. Both problems can lead to missed diagnoses: overfiltering because the causative variant is never reviewed and underfiltering because the expert is overwhelmed with too many variants to assess them properly.

Many computational methods have been developed to assist with the interpretation of variants. These methods all adopt the same fundamental form: evaluation of variant features to derive a 'pathogenicity likelihood score' from which a label of 'benign', 'pathogenic', or 'uncertain' (or a related term) is assigned. Methods like SIFT[13], PolyPhen-2 (ref. 14), and CADD[15] have been widely adopted and are often consulted by clinicians[5]. Most recently, the dbNSFP ensemble logistic regression[16] (MetaLR) score combined multiple scores to outperform any one individual method. Existing methods, as will be shown, have unacceptably high rates of false negatives (calling a pathogenic variant as benign) and false positives (calling a benign variant as pathogenic)[5]. None achieves the 90% classification accuracy proposed for a clinical classification as 'likely' benign or pathogenic[5]. As a result, clinicians cannot fully rely on the classification of variants by these pathogenicity likelihood scores, and the scores are used as only one of many factors in interpreting a variant[5].

We introduce here a proper complement to the standard allele frequency filter. The Mendelian Clinically Applicable Pathogenicity (M-CAP) score is a pathogenicity likelihood score that aims to misclassify no more than 5% of pathogenic variants while aggressively reducing the list of variants of uncertain significance. Much like allele frequency, M-CAP is readily interpreted; if it classifies a variant as benign, then that variant can be trusted to be benign with high confidence. M-CAP uses gradient boosting trees, a supervised learning classifier that excels at analyzing nonlinear interactions between features, and has state-of-the-art performance in a variety of classification tasks[17–19]. The features M-CAP uses for classification are based on both existing pathogenicity likelihood scores and direct measures of evolutionary conservation, the cross-species analog to frequency within the human population. We provide both (i) a novel method that combines amino acid conservation features with gradient boosting trees that can be applied to any variant training set and (ii) computed scores trained on mutations linked to Mendelian diseases that can be directly used by clinicians to interpret variants of uncertain consequences.

RESULTS

Derivation of the M-CAP score

Evaluating the pathogenicity of missense variants as a supervised machine learning task requires (i) features by which to classify variants,

[1]Department of Computer Science, Stanford University, Stanford, California, USA. [2]Department of Pediatrics, Stanford University, Stanford, California, USA. [3]Department of Medical Genetics, Cardiff University, Heath Park, Cardiff, UK. [4]Department of Developmental Biology, Stanford University, Stanford, California, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to G.B. (bejerano@stanford.edu).

Received 30 June; accepted 26 September; published online 24 October 2016; doi:10.1038/ng.3703

(ii) disjoint labeled training and test sets of both pathogenic and benign missense variants, and (iii) a learning algorithm that attempts to optimally separate pathogenic variants from benign ones using the features provided.

M-CAP features

M-CAP uses both preexisting and new features for each missense variant. It uses nine established pathogenicity likelihood scores: SIFT[13], PolyPhen-2 (ref. 14), CADD[15], MutationTaster[20], MutationAssessor[21], FATHMM[22], LRT[23], MetaLR[16], and MetaSVM[16]. It also incorporates seven established measures of base-pair, amino acid, genomic region, and gene conservation: RVIS[24], PhyloP[25], PhastCons[26], PAM250 (ref. 27), BLOSUM62 (ref. 27), SIPHY[28], and GERP[29]. In addition, M-CAP introduces 298 new features derived from multiple-sequence alignment of 99 primate, mammalian, and vertebrate genomes to the human genome[30]. Three vectors of 99 binary features indicate whether each species has the human reference amino acid, the alternative amino acid, or no amino acid aligning. One integer feature counts the number of different amino acids observed at the codon across the 100 species. The new features are ideal for machine learning; they delegate the task of deriving a sophisticated measure of evolutionary constraint to the learning model (Online Methods and **Supplementary Tables 1** and **2**).

M-CAP training and test sets

Pathogenic single-nucleotide variants were obtained from the Disease Mutation (DM) class of variants in Human Gene Mutation Database Pro version 2015.2 (HGMD)[31]. Predominantly benign variants were obtained from Exome Aggregation Consortium data set version 0.3 (ExAC), which combines data from exome and genome sequencing studies worldwide for over 60,000 individuals[10]. All individuals in the ExAC database are known not to have severe childhood Mendelian disorders, and, although the database may include some deleterious variants, it is thought to be extremely depleted for highly penetrant pathogenic variants[10]. Overlap between HGMD and ExAC consists mostly of recessive mutations that are tolerated in the heterozygous state and misannotations[10]. These variants were conservatively removed from the benign set, resulting in two disjoint sets.

It is standard practice in the clinical evaluation of exome sequencing results to focus on variants that are rare in the population, as common variants are considered most unlikely to cause rare Mendelian disorders[5]. To best represent this practice, only rare variants (allele frequency ≤1% in all ExAC and 1000 Genomes Project superpopulations) were included in the pathogenic and benign training and test sets. After filtering for rare, missense variants, the sets comprised 63,418 pathogenic variants from HGMD and 3,268,665 predominantly benign variants from ExAC (**Table 1**).

Many of the 63,418 HGMD variants have been used as training data in previous methods that M-CAP uses as features and/or should be compared against. Of the nine preexisting pathogenicity scores (and seven conservation metrics), PolyPhen-2, CADD, MutationTaster, FATHMM, and MetaLR/MetaSVM have all been trained using human mutation data (**Supplementary Table 3**). Similarly, some of the ExAC benign variants have been used as training data for preexisting pathogenicity scores. To avoid bias, we have removed all previously seen variants from the M-CAP training and test sets. The final sets of previously unseen variants consisted of 12,418 rare, missense pathogenic variants and 3,137,919 rare, missense benign variants (**Table 1**).

For computational efficiency, approximately 100,000 variants were randomly subsampled from the full set of benign variants. Then, approximately 1,000 randomly selected variants were held out for each class as a test set to evaluate the final model. The remaining variants were used as a training set to learn the model.

Table 1 Variants in the M-CAP labeled data source sets and patient exomes

	M-CAP source sets		
	HGMD Pro 2015.2 (pathogenic)	ExAC v3 (benign)	Patient exomes (median)
Single-nucleotide variants[a]	104,282	9,291,050	165,934
Missense variants[b]	70,192	3,428,124	9,724
Rare (≤1%), missense variants[c]	63,418	3,287,432	305
Variants after removal of overlap with HGMD[d]	N/R	3,268,665	N/R
Variants after exclusion of those used for training[e]	12,418	3,137,919	259

[a]Single-nucleotide variants in each resource. [b]Missense single-nucleotide variants (altering an amino acid). [c]Missense variants that are rare in the human population (frequency ≤1%). [d]Variants remaining after removing ExAC variants that overlap with HGMD variants. [e]Rare missense variants that have not previously been seen during the training phase for any of the metrics used in evaluation or as features in M-CAP (**Supplementary Table 3**, which also explains why previously trained-on benign data overlap proportionally more with individual patient data than with ExAC, a union of data from many individuals).

M-CAP learning algorithm

M-CAP uses a gradient boosting tree classifier[17], which learns a function of the input features as a linear combination of decision trees, each derived iteratively to correct previously misclassified elements. The model hyperparameters were optimized with a systematic grid search (Online Methods).

M-CAP outperforms existing pathogenicity scores

The pathogenic and benign test sets held out during the learning phase were used to evaluate M-CAP against the most popular pathogenicity scores (SIFT[13], PolyPhen-2 (ref. 14), and CADD[15]), the current state of the art in supervised learning (MetaLR[16]) that combines the above classifiers and additional ones (**Supplementary Table 3**), and a recently published unsupervised approach (Eigen[32]). Performance was measured using the area under receiver operating characteristic (AUC) curves, where a random classifier would obtain an AUC of 0.5 and a perfect classifier would obtain an AUC of 1.0. The methods fell into three clear categories of performance (**Fig. 1a**). CADD, SIFT, Eigen, and PolyPhen-2 all performed much better than a random classifier (AUC = 0.717–0.736) but were all surpassed by MetaLR (AUC = 0.844). M-CAP in turn was superior to all other methods, at all thresholds (AUC = 0.907), owing to its better ability to separate pathogenic from benign variants (**Fig. 1c**).

M-CAP excels at clinically relevant thresholds

A clinically relevant pathogenicity score must first and foremost limit the fraction of cases in which a pathogenic variant is misclassified as benign to 5% or less. At author-recommended default thresholds, existing pathogenicity scores misclassified over five times as many (26–38%) HGMD pathogenic mutations (**Table 2**). We propose to measure classifier performance in the clinically relevant domain using only the high-sensitivity regional AUC (hsr-AUC), which measures the area under the receiver operator characteristic curve only for a true positive rate of 95 to 100% (Online Methods). M-CAP achieved an hsr-AUC of 0.411 in comparison to hsr-AUC values of 0.257 for MetaLR and 0.079–0.119 for the other methods, offering a 60% improvement over MetaLR and a 245–420% improvement over the other scores (**Fig. 1b**).

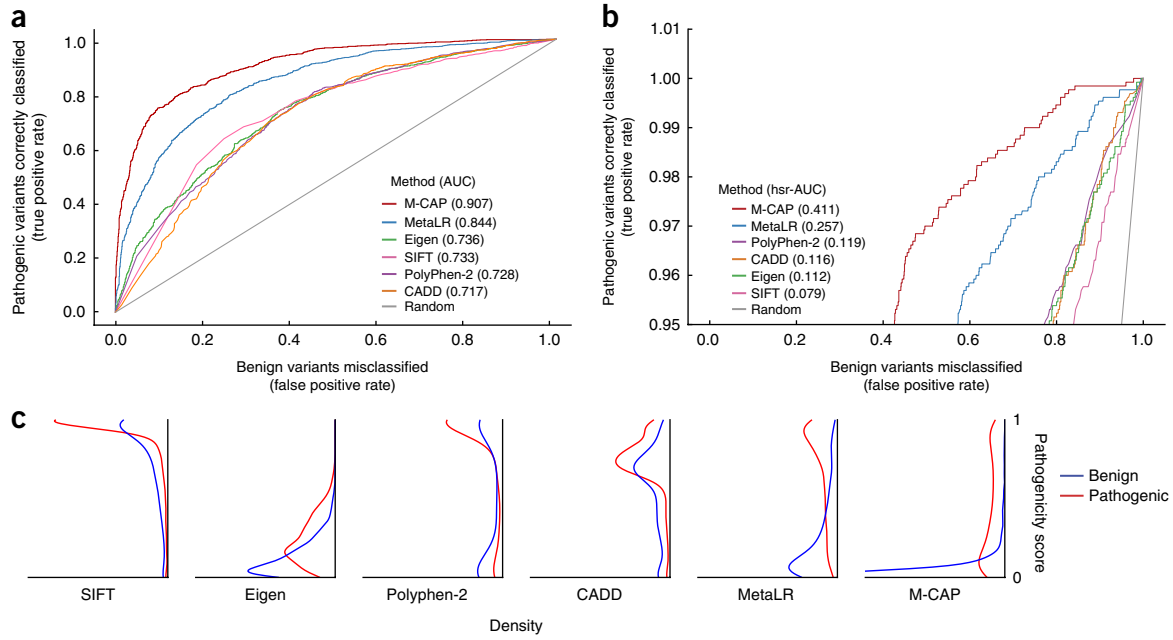

Figure 1 M-CAP outperforms existing pathogenicity likelihood metrics, particularly at the high sensitivity levels required for clinical applications. (**a**) Receiver operating characteristic curves for M-CAP and five popular pathogenicity metrics for missense variants shown with AUC values. (**b**) Enlarged view of the plot in **a** focused on the high-sensitivity (true positive rate ≥ 95%) region showing hsr-AUC, the normalized AUC within the high-sensitivity region. The hsr-AUC for M-CAP is 60% better than that for the current best (retuned) method and 245–420% better than the hsr-AUC values for the widely used SIFT, PolyPhen-2, and CADD scores and the recently released Eigen method. (**c**) Distributions of predicted pathogenicity likelihood scores for pathogenic and benign variants. All test set variants—benign and pathogenic—were combined, scored by each method, sorted, and assigned a linearly scaled pathogenicity score from 0 (lowest pathogenicity score predicted for each tool) to 1 (highest pathogenicity score predicted for each tool). The *y* axis plots the predicted pathogenicity, and the *x* axis shows the fraction of benign (blue) and pathogenic (red) variants in each bin. M-CAP scores show the best separation of predicted values for (truly) benign and pathogenic variants (Mann–Whitney *U* test $P < 4.3 \times 10^{-285}$ compared to $P < 1.8 \times 10^{-204}$ for the second best method, MetaLR).

M-CAP generalizes to Mendelian mutations not found in HGMD

To further demonstrate that M-CAP is not overfit to HGMD and maintains performance on unseen data, we evaluated all methods on a smaller set of rare, missense pathogenic or likely pathogenic mutations[32]. The set included 17 *BRCA1*, 10 *BRCA2*, 39 *CFTR*, and 19 *MLL2* (also known as *KMT2D*) mutations that are not in HGMD. M-CAP (at a high-sensitivity threshold >0.025) successfully classified 100% of these 85 variants as pathogenic. Even after adjusting to 95% sensitivity thresholds (SIFT, <0.49; PolyPhen-2, >0.022; CADD, >10.37; Eigen, >1; MetaLR, >0.106), all other methods correctly classified a smaller fraction of variants: SIFT (96.5%), PolyPhen-2 (88.2%), CADD (90.6%), Eigen (84.7%), and MetaLR (94.1%) (Online Methods and **Supplementary Tables 4** and **5**).

M-CAP eliminates the most variants in patient exomes

A typical exome for a patient with a Mendelian disease contains a very unbalanced set of one or two rare pathogenic mutations and hundreds of rare variants, which are likely benign with respect to the patient's condition (**Table 1**). To represent clinical practice, we evaluated the performance of pathogenicity likelihood metrics in ten diverse patients with a known causative variant (Online Methods and **Supplementary Tables 6–8**). After the standard allele frequency filter (≤1%), the patients had 232–449 (median = 305) rare, missense variants. Next, to ensure accurate evaluation, we filtered out any variants seen during the training phase for M-CAP or any of the other previous metrics used as features or for comparison. The filtered patient test sets had 183–399 (median = 259) rare, missense variants (**Table 1** and **Supplementary Table 8**).

Consistent with their performance on HGMD variants, SIFT, CADD, and PolyPhen-2 each misclassified one or two pathogenic variants in patients as benign and MetaLR misclassified three, whereas M-CAP misclassified none (**Table 2**). When the thresholds for all the methods were modified to achieve consistent 95% sensitivity (**Supplementary Table 6**), SIFT, PolyPhen-2, and MetaLR each misclassified a pathogenic variant in only a single patient. However, the abilities of the different methods to reduce the list of variants of uncertain consequences differed substantially: SIFT, PolyPhen-2, CADD, and Eigen only reduced the list by 20–30% and MetaLR reduced the list by 47%, whereas M-CAP outperformed all other methods and consistently reduced the list by an average of 57% (**Fig. 2a**).

Table 2 Default classifier performance on known disease-causing variants

Method	Authors' recommended threshold	Misclassified HGMD variants (%)	*n* misclassified HGMD variants (*n* = 12,418)	*n* misclassified causative patient variants (*n* = 10)
SIFT	<0.05	38	4,751	1
PolyPhen-2	>0.8	31	3,861	2
CADD	>20	26	3,178	1
MetaLR	>0.5	27	3,405	3
M-CAP	>0.025	5	66 (*n* = 1,300)	0

Performance of the methods was evaluated at the authors' recommended sensitivity thresholds. SIFT, PolyPhen-2, and CADD are commonly used metrics for variant prioritization. MetaLR is a recently published state-of-the-art method that outperforms these scoring measures (**Fig. 1**). The four methods misclassify a substantial fraction of known pathogenic variants from HGMD. M-CAP has been explicitly designed to misclassify only a tolerable 5% of pathogenic variants. The Eigen method, also in **Figure 1**, does not provide default thresholds for classification.

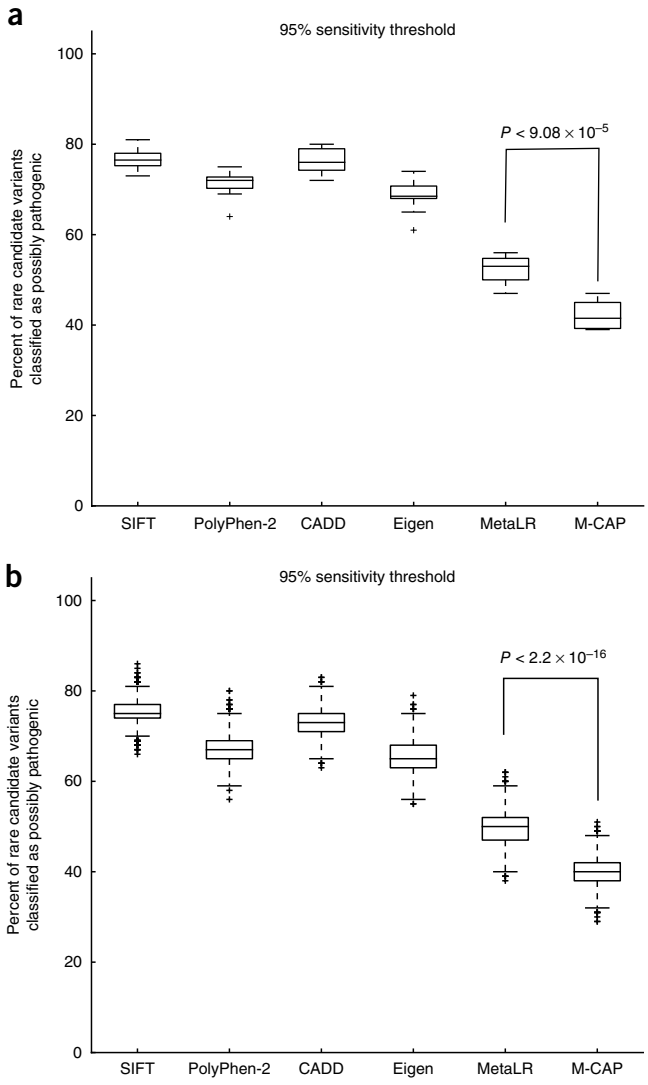

Figure 2 M-CAP correctly eliminates the most variants of uncertain consequences as benign at 95% sensitivity. (**a**,**b**) We use solved exomes from patients ($n = 10$) (**a**) and 1000 Genomes Project exomes ($n = 2{,}504$) (**b**) to examine classifier performance at 95% sensitivity. The *y* axis shows the fraction of variants of uncertain significance that cannot be dismissed as benign by each classifier. SIFT, PolyPhen-2, CADD, and Eigen only reduce the variant list by 20–30% in patient exomes and 22–34% in 1000 Genomes Project exomes. MetaLR reduces the variant list by an average of 47% in patient exomes and 50% in 1000 Genomes Project exomes. M-CAP outperforms all other methods and consistently reduces the variant list by an average of 57% in patient exomes and 60% in 1000 Genomes Project exomes. The whiskers for each box plot stretch beyond the first and third quartiles to 1.5 times the interquartile range. The distribution of variants dismissed by the next best method (MetaLR) differs significantly from that for M-CAP in both patient exomes ($P < 9.08 \times 10^{-5}$) and 1000 Genomes Project exomes ($P < 2.2 \times 10^{-16}$).

M-CAP reduces the number of variants by the most in 2,504 individuals

Each patient with a classic Mendelian disease harbors one or two pathogenic alleles. In the context of solving the genetic basis for these patients, the rest of the variants can be classified as benign. To evaluate the performance of M-CAP on a larger scale, we next applied it to 2,504 individual exomes collected by the 1000 Genomes Project[11]. Each individual can be thought of as a patient with a Mendelian disease for whom the causative variant has been removed from the genome. These individuals had 107–604 (median = 320) rare, missense variants never used for training after filtering using all criteria from **Table 1**.

As with our patients, SIFT, PolyPhen-2, CADD, and Eigen only reduced the 1000 Genomes Project rare variant lists by 22–34%. MetaLR reduced the lists by 50%. M-CAP outperformed all other methods and consistently reduced the lists by an average of 60% (**Fig. 2b**).

Feature contribution to M-CAP

M-CAP uses 318 features as input into the gradient boosting tree model. The model takes advantage of a series of decision rules, incorporating these features across thousands of trees to build the optimal classification boundary. To measure feature weights, we averaged each feature's importance across all trees (Online Methods). The new 'aligned amino acid' and 'reference amino acid match' feature sets had the two highest weights (**Supplementary Table 9**), highlighting that the new features contribute information not available through existing commonly used features.

Synergy between features and the classifier boosts M-CAP performance

M-CAP differs from existing pathogenicity likelihood scores in two key respects: (i) it uses a gradient boosting tree classifier and (ii) it applies new amino acid evolutionary conservation metrics (Online Methods and **Supplementary Tables 1** and **2**). To evaluate the separate contributions of these two changes to M-CAP performance, we compared two different classifiers—logistic regression (used by MetaLR) versus gradient boosting trees (used by M-CAP)—and two sets of features—preexisting conservation and pathogenicity metrics versus preexisting metrics plus our new amino acid conservation metrics (**Supplementary Table 1**). The logistic regression classifier with preexisting metrics served as the baseline.

Individually, the gradient boosting tree classifier and the new conservation-based features boosted performance by 8.49% and 4.25% in hsr-AUC, respectively. Their combination resulted in a synergistic increase of 58.69% in hsr-AUC (**Supplementary Table 10**). This synergistic increase can be attributed to (i) the inability of the logistic regression model to effectively learn patterns using 298 weak amino acid features and (ii) the limited feature space of the 16 preexisting metrics that does not let the gradient boosting tree model reach full potential. Gradient boosting trees introduce weak learners at each stage and are able to take full advantage of the 298 weak amino acid features to build an effective and generalizable model.

The logistic regression classifier that we implement here performed very similarly to the published MetaLR method despite different training data (AUC for our logistic regression = 0.853, MetaLR AUC = 0.844). The MetaLR model is trained using pathogenic mutations from UniProt[33], whereas our logistic regression model was trained with the same HGMD data used to train M-CAP. This suggests that there is no major bias in the training data that we used that gives M-CAP any unfair advantage.

DISCUSSION

Human disease-causing mutations exhibit a range of pathogenicity profiles, from simple (one or two) highly penetrant mutations in Mendelian diseases to mutations with much subtler effects that can increase disease susceptibility additively in complex common diseases. Different pathogenicity classifiers may be designed to focus on these different categories of mutations. SIFT, PolyPhen-2, and MetaLR focus on coding mutations, while CADD and Eigen also

provide scores for noncoding variants. Here we focus explicitly on classifying coding mutations for Mendelian diseases. The need for such classification is acute as the use of exome and genome sequencing in the clinic continues to expand.

Exomes for Mendelian diseases present the challenge of detecting the one or two causative mutations among the hundreds of rare variants not related to a patient's medical condition. In this setting, it is critical to train a classifier first and foremost to seldom misclassify the causative variant as benign. We show here that popular methods misclassify up to 38% of known rare pathogenic mutations as benign at their recommended thresholds. As a result, those interpreting data for Mendelian diseases have become accustomed to treating the scores as a line of evidence in some support of variant classification rather than as a simple rule that enables variant inclusion or exclusion.

We introduce the hsr-AUC measure, which measures the AUC only between true positive rates of 95 and 100%. We recommend that future pathogenicity predictors developed with Mendelian diseases in mind be evaluated in this regime.

Our study rethresholds five classifiers to 95% sensitivity and compares them to our new M-CAP algorithm. Three categories of performers emerge. In the first, we find SIFT, PolyPhen-2, CADD, and Eigen, all of which have similar predictive power. Some, like PolyPhen-2 and CADD, can potentially be improved by training on more known pathogenic variants. However, the MetaLR classifier trains on over a dozen pathogenicity classifiers from dbNSFP and is indeed more powerful than any individual classifier (after our retuning). M-CAP is shown to take a major stride in further improving the classification of variants in Mendelian diseases. It improves hsr-AUC performance by 60% over MetaLR and by 245–420% over the other methods. We show that this improvement is a result of two synergistic design choices: a better suited learning algorithm and new genomic features that play to the strength of gradient boosting trees.

Finally, using solved patient exomes and 1000 Genomes Project individuals, we show that M-CAP can be used as a simple rule: it correctly dismisses 60% of the hundreds of variants of uncertain consequences in a typical exome while retaining 95% of known rare pathogenic mutations. Although it only removes three-fifths of likely benign variants, this profile makes M-CAP valuable to clinicians looking to save the precious time of medical geneticists in rapidly detecting a known causative variant or gene in a much reduced list of candidates. The shorter list also greatly aids researchers looking to select candidate variants to study functionally in the 75% of exomes that cannot be solved on the basis of current knowledge[34].

Supervised learning methods generally surpass unsupervised methods when high-quality training data of appropriate type and quantity are available. Here we show that the M-CAP supervised learning method outperforms other approaches at classifying missense variants when trained with a carefully curated set of pathogenic and benign variants. There remains a dearth of training data for noncoding disease-linked mutations, making unsupervised methods such as Eigen attractive for the time being.

When evaluating a supervised machine learning classifier, it is critical for the classifier not to have seen any of the test sets during training. It is straightforward to separate the immediate training and test sets, but it is more difficult to ensure that the test set not be identified indirectly through a feature. M-CAP uses other classifiers as features, and much of the available data have previously been used to train these classifiers. We take great care here to eliminate any of the previously used data from the M-CAP training and test sets. It is important to take this step when evaluating future classifiers, including for unsupervised methods that may indirectly access training sets by using supervised methods as features.

M-CAP continues a line of progress in variant prioritization that illustrates how much room there is to improve on the most commonly used metrics, SIFT, PolyPhen-2, and CADD. While M-CAP offers a clinically relevant classifier, it is important to continue to improve the methods for prioritizing functional variants and to encourage the adoption of superior methods by clinicians and researchers working with exome data.

URLs. M-CAP website, http://bejerano.stanford.edu/MCAP; M-CAP codebase, https://bitbucket.org/bejerano/mcap_public; Eigen data, https://xioniti01.u.hpc.mssm.edu/TrainingTestingDatasets/TestingDatasets/Mendelian/; CADD, http://cadd.gs.washington.edu/download; Eigen, https://xioniti01.u.hpc.mssm.edu/v1.0/EIGEN/; UCSC annotation tracks, http://hgdownload.cse.ucsc.edu/goldenpath.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the members of the Bejerano laboartory, particularly J. Notwell, S. Chinchali, and J. Birgmeier, for technical advice and helpful discussions. P.D.S. and D.N.C. receive financial support from Qiagen through a license agreement with Cardiff University. We thank the PolyPhen-2, CADD, Eigen, FATHMM, MutationTaster, and MetaLR teams for making their training and testing data readily available. This work was funded in part by the Stanford Pediatrics Department, DARPA, a Packard Foundation Fellowship, and a Microsoft Faculty Fellowship to G.B.

AUTHOR CONTRIBUTIONS

K.A.J., A.M.W., M.J.B., and G.B. designed the study and analyzed results. K.A.J. and M.J.B. implemented the model and performed the experiments. K.A.J., A.M.W., and H.G. wrote software tools that were used for analysis. P.D.S. and D.N.C. curated the HGMD data and provided feedback. J.A.B. provided patient exome cases and feedback. K.A.J., A.M.W., and G.B. wrote the manuscript. All authors reviewed and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
2. Iglesias, A. *et al.* The usefulness of whole-exome sequencing in routine clinical practice. *Genet. Med.* **16**, 922–931 (2014).
3. Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *J. Am. Med. Assoc.* **312**, 1880–1887 (2014).
4. Brownstein, C.A. *et al.* An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome Biol.* **15**, R53 (2014).
5. Richards, 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.* **17**, 405–424 (2015).
6. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
7. Simpson, M.A. *et al.* Mutations in *NOTCH2* cause Hajdu–Cheney syndrome, a disorder of severe and progressive bone loss. *Nat. Genet.* **43**, 303–305 (2011).
8. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
9. Taylor, J.C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* **47**, 717–726 (2015).
10. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
11. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

12. Rehm, H.L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* **15**, 733–747 (2013).
13. Ng, P.C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
14. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
15. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
16. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).
17. Hastie, T., Tibshirani, R. & Friedman, J. *Elements of Statistical Learning* (Springer, 2003).
18. Fusi, N., Smith, I., Doench, J. & Listgarten, J. *In silico* predictive modeling of CRISPR/Cas9 guide efficiency. Preprint at *bioRxiv* http://dx.doi.org/10.1101/021568 (2015).
19. Ogutu, J.O., Piepho, H.-P. & Schulz-Streeck, T. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc.* **5** (Suppl. 3), S11 (2011).
20. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).
21. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
22. Shihab, H.A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).
23. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
24. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
25. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
26. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
27. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).
28. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).
29. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
30. Kuhn, R.M., Haussler, D. & Kent, W.J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
31. Stenson, P.D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).
32. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J.D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
33. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
34. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *J. Am. Med. Assoc.* **312**, 1870–1879 (2014).

ONLINE METHODS

Missense variant annotation. ANNOVAR v527 was used to annotate variants with predicted effect on protein-coding genes using gene isoforms from Ensembl gene set version 75 for the hg19/GRCh37 assembly of the human genome[35]. All coding isoforms were used where the transcript start and end sites were marked as complete and the coding span was a multiple of three.

Training and test data sets. Pathogenic variants were obtained from HGMD[31]. Predominantly benign variants were obtained from ExAC data set version 0.3. Because SIFT[13] and PolyPhen-2 (ref. 14) do not score all missense variants, to generate the comparison in **Figure 1**, the test set was refined to only variants with a score from all methods.

Test data from *BRCA1*, *BRCA2*, *CFTR*, and *MLL2*. A separate set of pathogenic mutations was obtained from the Eigen website (see URLs) on 15 April 2016. The data set contains 28 *BRCA1*, 13 *BRCA2*, 92 *CFTR*, and 31 *MLL2* missense mutations. After filtering for rare missense variants not in HGMD, 17 *BRCA1*, 10 *BRCA2*, 39 *CFTR*, and 19 *MLL2* mutations remained (**Supplementary Table 4**).

Existing pathogenicity scores. SIFT v1.03 (ref. 13), PolyPhen-2 (HumVar)[14], and MetaLR[16] scores were obtained from dbNSFP v3.0a[36,37], downloaded on 15 August 2015.

CADD v1.3 (ref. 15) scores were downloaded from the CADD website (see URLs) on 30 August 2015. Eigen v1.0 coding Phred scores were downloaded from the Eigen website (see URLs) on 14 January 2016.

Variant features. The M-CAP classifier uses both preexisting and new variant features (**Supplementary Tables 1** and **2**).

Preexisting features. M-CAP uses other pathogenicity likelihood scores as features. In addition to SIFT, PolyPhen-2, CADD, and MetaLR, M-CAP incorporates MetaLR[16], MutationTaster 2 (ref. 20), MutationAssessor release 2 (ref. 21), FATHMM v2.3 (ref. 22), and LRT[23]. The MetaLR, MutationTaster, MutationAssessor, FATHMM, and LRT scores were obtained from dbNSFP v3.0a[36,37]. These methods have missing values for a subset of variants, in which case we by default set the score to the maximal pathogenic score.

M-CAP also incorporates publicly available features for base-pair, region, gene, and amino acid conservation, all downloaded on 15 August 2015. PhyloP[25] and PhastCons[26] scores are from the UCSC hg19 100-way Multiz Conservation track (see URLs). Residual variation intolerance scores (RVIS) were calculated for all Ensembl genes using the published RVIS methodology[24]. The standard PAM250 (ref. 27) and BLOSUM62 (ref. 27) matrices were used as amino acid substitution scores. GERP++ (ref. 29) and SIPHY[28] were also from dbNSFP v3.0a.

New features. M-CAP also introduces new metrics for evolutionary constraint on amino acid residues in coding genes. The ortholog for each human codon was obtained from the UCSC hg19 100-way Multiz alignment[30] with per-alignment column mapping.

Two hundred and ninety-eight conservation features were assigned to each missense variant (**Supplementary Tables 1** and **2**): 99 'aligned codon' binary features (1 per species) indicate whether the species aligns to the human codon across all three base pairs, 99 'reference amino acid match' binary features indicate whether the species has the same amino acid as the reference human genome, 99 'alternative amino acid match' binary features indicate whether the species has the same amino acid as the variant, and 1 'unique amino acids' integer feature counts the number of different amino acids observed at the codon across the 100 species.

Gradient boosting tree classification. Classification was performed using a gradient boosting tree model[17], which learns a function F_M of the input features x as a linear combination of decision trees $h_1, h_2, \ldots, h_M$. Each decision tree is derived iteratively to correct previously misclassified elements[17] and better predict y, the true labels of the variants. Each tree h_i uses randomly selected features and is derived using the classification and regression tree (CART) algorithm[17]. CART constructs binary decision trees using the feature and threshold that yield the minimum entropy at each node.

The classifier was implemented in Python with sci-kit learn machine learning library v0.16.1 (ref. 38), using the negative binomial log likelihood as the loss function L (where loss is "deviance" in sci-kit). The loss function examines the number of samples that are correctly classified in an iteration to determine the weight γ_m for each tree h_m.

A grid search was used to optimize the gradient boosting tree hyperparameters: the number of decision trees M (from 4,000 to 6,000 in increments of 250); the depth of each tree (constituting {3, 4, 6, 8}), always with a minimum of three samples per leaf node; and the learning rate α (constituting {0.005, 0.01, 0.05, 0.1}). Gradient boosting trees were calculated as

$$F_m(x) = F_{m-1}(x) - \alpha\gamma_m h_m(x)$$

$$\gamma_m = \arg\min_\gamma L(y, F_{m-1}(x) - \alpha\gamma h_m(x)) \quad (1)$$

To avoid overfitting the model, the grid search was performed with fivefold cross-validation. The training data were divided into five equal and independent subsets; the model was trained on four-fifths of the data, a threshold was defined at 95% sensitivity on the training set, and accuracy ((true positives + true negatives)/total) was evaluated on the held-out one-fifth ('validation set'). This was performed five times, once with each fifth of the data set as the validation set.

A model was trained for each combination of hyperparameter values, and the parameter combination that resulted in the highest average accuracy on the validation sets was selected (M = 5,750, tree depth = 6, α = 0.01). The final classifier was retrained on the full training set of 11,118 HGMD variants and 94,934 ExAC variants with the optimal hyperparameters from the grid search.

The asymptotic run time of gradient boosting trees grows linearly with the number of trees, the number of features, and the depth of each tree and grows as $N\log N$ with the size of the training data (N). Actual run time, including grid search, was 3,600 CPU-hours on a six-core large memory Dell PowerEdge server.

Calculating feature weights. In a gradient boosting tree, each feature's weight is calculated by measuring its Gini importance or mean decrease in impurity[17]. The Gini index is calculated for each node (n) over all classes (K). We define p_{nk} as the proportion of class k observations (y_i) in node n. The Gini index for a specific node is the sum of the variance in proportion for all classes (K). The Gini importance for a feature (f) is then defined as the sum of the Gini index values over all nodes incorporating the feature (nodes$_f$) in the gradient boosting tree model. This is the underlying implementation for the default 'feature_importances' method available in the Python sci-kit learn library. We use this 'feature_importances' attribute to recover the weights for our trained gradient boosting tree model

$$p_{nk} = \frac{1}{N_n} \sum_{y_i \in R_n} I(y_i = k)$$

with Gini index (GI)

$$\sum_{k \neq k^\star} p_{nk}\, p_{nk^\star} = \sum_{k=1}^{K} p_{nk}(1 - p_{nk})$$

and Gini importances.

$$\sum_{n \in \text{nodes}_f} \text{GI}_n$$

Thresholds to label variants as benign or pathogenic. Default pathogenicity thresholds for SIFT, PolyPhen-2, CADD, and MetaLR are as recommended by the original authors: <0.05 for SIFT, >0.8 for PolyPhen-2, >20 for CADD, and >0.5 for MetaLR.

The M-CAP classifier aims to achieve high sensitivity, which we define as (true positives)/(true positives + false negatives) ≥ 0.95. This means that at least 95% of HGMD variants are correctly classified as pathogenic. A threshold was determined with the final classifier such that 95% of the pathogenic training set had an M-CAP score above the threshold (were classified as pathogenic). The same approach was used to define the high-sensitivity threshold for the other classifiers (**Supplementary Table 6**).

High-sensitivity regional area under the curve. The hsr-AUC metric quantifies classifier performance in the high-sensitivity region that is most relevant for clinical applications (true positive rate ≥ 95%; **Fig. 1b**). The hsr-AUC metric is defined as the area under the curve within the high-sensitivity region normalized by the area of the entire region (0.05).

Patient data sets. The whole-exome sequences of human patients diagnosed with Mendelian disorders were obtained, after informed consent, from research subjects referred by the Clinical Genetics Service at Stanford Children's Health (P01–P09) and from database of Genotypes and Phenotypes (dbGaP) study phs000204.v1.p1 (ref. 6; P10). Sequencing and diagnosis for Stanford patients were performed by outside clinical laboratories (Baylor Miraca Genetics Laboratory, Ambry Genetics, and UCLA Clinical Genomics Center), which provided the raw sequencing reads upon patient consent. All human subject research was performed under guidelines approved by the Stanford Institutional Review Board.

Sequencing reads were mapped to the GRCh37/hg19 assembly of the human genome using BWA-MEM v0.7.10-r789 (ref. 39). Variants were called using GATK v3.4-46-gbc02625 following the HaplotypeCaller workflow from GATK best practices[40].

For each patient examined, the causative variant was known from publication or a clinical report confirmed by a medical geneticist (**Supplementary Table 6**). Patient phenotype information is listed in **Supplementary Table 7**. The rare, missense variants in each patient are listed in **Supplementary Table 8**. All original raw read files are available upon request or from dbGaP, respectively.

Data availability. M-CAP scores for all rare, missense variants in the human genome, along with the source code, amino acid conservation features, and final trained model for the M-CAP classifier are available through the M-CAP website (see URLs), licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. The M-CAP repository is also available at Bitbucket (see URLs).

35. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
36. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011).
37. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* **34**, E2393–E2402 (2013).
38. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
39. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
40. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

doi:10.1038/ng.3703