StrVCTVRE: A supervised learning method to predict the pathogenicity of human structural variants

Andrew G. Sharo[1], Zhiqiang Hu[2], Steven E. Brenner[2]*

[1]*Biophysics Graduate Group, University of California, Berkeley, California*

[2]*Department of Plant and Microbial Biology, University of California, Berkeley, California*

**Correspondence:* brenner@compbio.berkeley.edu

Abstract

Whole genome sequencing resolves clinical cases where standard diagnostic methods have failed. However, preliminary studies show that at least half of these cases still remain unresolved, even after whole genome sequencing. Structural variants (genomic variants larger than 50 base pairs) of uncertain significance may be the genetic cause of a portion of these unresolved cases. Historically, structural variants (SVs) have been difficult to detect with confidence from short-read sequencing. As both detection algorithms and long-read/linked-read sequencing methods become more accessible, clinical researchers will have access to thousands of reliable SVs of unknown disease relevance. We show that filtering these SVs by overlap with cataloged SVs is an imperfect solution. Innovative methods to predict the pathogenicity of these SVs will be needed to realize the full diagnostic potential of long-read sequencing. To address this emerging need, we developed StrVCTVRE (Structural Variant Classifier Trained on Variants Rare and Exonic), a classifier that can be used to distinguish pathogenic SVs from benign SVs that overlap exons. We made use of features that capture gene importance, coding region, conservation, expression, and exon structure in a random forest classifier. We found that some features, such as expression and conservation, are important but are absent from SV classification guidelines. Although databases of SVs reflect size biases from sequencing techniques, we leveraged multiple databases to construct a size-matched training set of rare, putatively benign and pathogenic SVs. In independent test sets, we found our method performs accurately across a wide SV size range, which will allow clinical researchers to eliminate nearly 60% of SVs from consideration at an elevated sensitivity of 90%. However, our method and its assessment are still constrained by a small training dataset and acquisition bias in databases of pathogenic variants. StrVCTVRE fills an empty niche in the clinical evaluation of SVs of unknown significance. We anticipate researchers will use it to

prioritize SVs in patients where no variant is immediately compelling, empowering deeper investigation into novel SVs and disease genes to resolve cases.

Introduction

Whole genome sequencing (WGS) can identify causative variants in clinical cases that elude other diagnostic methods[1]. As the price of WGS falls and it is used more frequently, researchers and clinicians will increasingly observe structural variants (SVs) of unknown significance. SVs are a heterogeneous class of genomic variants that include copy number variants such as duplications and deletions, rearrangements such as inversions, and mobile element insertions. While a typical short-read WGS study finds 5,000–10,000 SVs per human genome, long-read WGS is able to identify more than 20,000[2]. This is two orders of magnitude fewer than the ~3 million single nucleotide variants (SNVs) identified in a typical WGS study. Still, despite their relatively small number, SVs play a disproportionately large role in genetic disease and are of great interest to clinical geneticists and researchers.

Clinically, SVs are of interest because they are causal in many rare diseases. Most SVs identified by WGS are benign, but on average, a given SV is more damaging than an SNV due to its greater size and ability to disrupt multiple exons, create gene fusions, and change gene dosage. In a study of 119 probands who received a molecular diagnosis from short-read WGS, 13% were due to a causal SV[3]. This is roughly consistent with an earlier study that found 7% of congenital scoliosis cases are caused by compound heterozygotes comprised of at least one deletion[4]. This suggests that in many rare disorders, SVs constitute a minor yet appreciable fraction of causal variants. Yet, since SVs continue to be challenging to identify, these figures may underestimate the true causal role that SVs play in rare disease. Indeed, in some rare diseases, the majority of cases are caused by SVs. For example, deletions cause most cases of Smith-Magenis syndrome, and duplications cause most cases of Charcot-Marie-Tooth disease type 1A[5].

To continue discovering disease genes in which SVs are causal, researchers face a daunting challenge: prioritizing the thousands of SVs found in WGS. Best practices for SV prioritization are evolving, and generally mirror steps used to prioritize SNVs, because very few SV-tailored impact predictors have been developed. A small number of published studies have focused on identifying pathogenic SVs from WGS[3, 6, 7] and have identified a handful of important steps. Removing low-quality SV calls is essential, as short-read SV callers rarely achieve precision above 80% for deletions and 50% for duplications, even at low recall[8]. Most studies remove SVs

seen at high frequency in population databases or internal controls[9, 10]. Moreover, many studies only investigate variants that overlap an exonic region, as non-coding SVs remain difficult to interpret. Depending on its sensitivity, a pathogenic SV discovery pipeline may produce tens to hundreds of rare exonic variants per proband to be investigated. These values are consistent with a recent population-level study that estimates SVs comprise at least 25% of all rare predicted Loss of Function (pLoF) events per genome[11]. Prioritizing SVs will be necessary for the majority of probands, as shown by a study of nearly 500 unresolved cases that found one or more SVs that warranted further investigation in 60% of cases[3]. Clinically validating all SVs of uncertain significance is currently infeasible, and cohort size for rare diseases will likely never reach a scale sufficient to statistically associate these variants with disease. Therefore, computational tools are needed to prioritize and predict the pathogenicity of rare SVs.

Methods are needed to prioritize SVs for manual investigation, but most tools instead annotate SVs with customized features. General-purpose annotation frameworks such as Ensembl's Variant Effect Predictor (VEP)[12] and SnpEff[13] both annotate SVs with broad consequences based on sequence ontology terms (e.g., transcript_ablation), which we show are not sufficient for prioritization. One standalone annotator, SURVIVOR_ant, annotates SVs with genes, repetitive regions, SVs from population databases, and user defined features. This and similar tools put the onus on researchers to provide informative features and determine how to consider these features in combination, a difficult challenge. A complementary approach is to annotate SVs using cataloged SVs known to be pathogenic or benign. One such SV annotator, AnnotSV, ranks SVs into five classes based on their overlap with known pathogenic or benign SVs and genes associated with disease. This approach can be successful when the causal SV has previously been seen in another proband and was cataloged as pathogenic, but we show it has significant limitations when the causal SV is novel. To effectively prioritize every SV, a method must summarize annotations to provide a quantitative score of SV deleteriousness.

Few methods score SVs by deleteriousness. One standalone impact predictor, SVScore, calculates a pathogenicity score for all possible SNVs within each SV (using CADD scores by default), while considering variant type and gene truncation. SVScore then applies an aggregating operation across the CADD scores (the default is the mean of the top 10%), and this approach has shown promise in identifying SVs under purifying selection[14]. Another stand-alone predictor, SVFX, integrates multiple features, but focuses on somatic SVs in cancer and germline SVs in common diseases[15]. The majority of SVs identified by sequencing are rare (AF < 1%), so the salient challenge in resolving undiagnosed cases is to distinguish rare pathogenic

variants from rare benign variants[11]. Since these existing SV predictors have been trained and assessed on common benign variants[14, 15], they may rely on features that separate common from rare SVs and be less clinically useful[16]. In this manuscript, we introduce StrVCTVRE, the first method to provide a score based on a machine learning approach trained on sets of rare germline SVs from disease cases.

Results

StrVCTVRE implementation

StrVCTVRE is implemented as a random forest, in which many decision trees 'vote' for whether a given SV is pathogenic. The score reflects the fraction of decision trees that 'voted' that the SV is pathogenic. The decision trees are shaped by a learning algorithm, in which each tree sees thousands of examples of pathogenic and benign SVs, and the decision nodes are optimized for accuracy. To promote diverse trees, each node of the decision tree uses only a random subset of the features. Finally, the algorithm is assessed on a held-out test dataset and independent test datasets.

Characterization of StrVCTVRE features

To accurately classify SVs, StrVCTVRE has 17 features that capture gene importance, conservation, coding sequence, expression, and exon structure of the disrupted region (see Methods for detailed descriptions of the features). We assessed gene importance using two features that summarize the depletion of predicted loss-of-function (pLoF) variants in healthy individuals: pLI[17] and LOEUF[9]. To specifically model coding sequence disruptions, we used three coding features: percentage of the coding sequence (CDS) affected by the SV, distance from the CDS start to the nearest position in the SV, and distance from the CDS end to the nearest position in the SV. To explicitly capture when an important gene is predicted loss-of-function (pLoF) due to an SV, we included two predicted loss of important gene features: pLI of a pLoF gene and LOEUF of a pLoF gene. We included a single conservation feature, phyloP of 100 vertebrates[18], which produced the best classification among the conservation features we investigated (see Methods) and was the most informative conservation feature in a rare missense variant classifier.[19] To capture expression features, we included the average expression across all tissues for each exon, exon inclusion across gene isoforms, and overlap with known topologically associating domain (TAD) boundaries. To be able to model potential differences that drive the pathogenicity of deletions and duplications, we included as a feature whether an SV is a deletion or duplication. The remaining features were exon structure features

including the number of exons in a disrupted gene, the number of exons disrupted, whether any exons were constitutive, whether all exons could be skipped in frame, and the rank of the exon in the transcript. When multiple exons or genes were disrupted, we typically took the maximum or minimum of features, as appropriate (see Methods).

Correlation and importance of StrVCTVRE features

Clusters emerged when we calculated these features for a set of SVs, computed the correlation between each feature, and clustered by correlation (Fig. 1a). The most prominent cluster contains gene importance, conservation, coding features, and one exonic feature, with most correlations above Spearman's $\rho = 0.6$. Since both predicted loss of important gene features are present in this cluster, the other features in this cluster may also capture when an important gene is highly disrupted. Smaller clusters included the gene-importance features and exonic features. Expression and deletion/duplication status were the features least correlated with all other features (all $\rho < 0.27$). This low correlation suggests that these features capture unique information, which is unsurprising for deletion/duplication status, but given the feature importance of expression data (Fig. 1b), suggests expression data contains both orthogonal and valuable information in determining SV pathogenicity. The two predicted loss of important gene features were the features most correlated with each other ($\rho = 0.97$), indicating that pLI and LOEUF are generally interchangeable for assessing the importance of highly disrupted genes.

By training on thousands of example SVs, StrVCTVRE discovers which features are useful for discriminating between pathogenic and benign SVs (Fig. 1b). Using Gini importance (see Methods), we found gene importance to be the most important feature. This was followed by a group of features with similar importance that include the conservation, number of exons in a gene, coding sequence, predicted loss of important gene, and expression. The importance of these features is largely intuitive; gene importance, coding sequence, and conservation are expected to be helpful to assess pathogenicity. In contrast, number of exons in gene is likely important because several pathogenic genes have numerous exons (DMD, NF1, BRCA2) and have many representative SVs in our dataset (Fig. S8). Surprisingly, several exonic features had relatively low importance, which may have been caused by the sparsity of SVs in our dataset that affect just a single exon. The low importance of TAD boundaries is counter to findings from a recent cancer SV impact predictor[15] and may reflect StrVCTVRE's focus on exonic SVs. Additionally, the low importance of deletion/duplication status suggests that on average, for exonic deletions and duplications, the region affected by an SV is more important than whether there was a gain or loss of genome content.

Characterization of StrVCTVRE training and held-out test sets

A total of 7,263 pathogenic deletions and 4,551 pathogenic duplications were collected from ClinVar,[20] a public database of variants cataloged by academic institutions, clinical laboratories, and genetic testing companies. We restricted our data to deletions and duplications, as they are the only SV types with more than 500 pathogenic examples in ClinVar. Additionally, deletions and duplications constitute the vast majority (> 95%) of rare gene-altering SVs[10]. Putatively benign variants were collected from ClinVar, gnomAD-SVs[11], and a recent great ape sequencing study[21]. We retained only rare (allele frequency (AF) < 1% in general population) variants in order to match the challenge faced by SV discovery pipelines. Indeed, 92% of SVs identified by sequencing are rare (AF < 1%), so the salient challenge is to distinguish rare pathogenic variants from rare benign variants[11]. Existing SV predictors have been trained and assessed on common benign variants[14, 15], which may cause them to instead rely on features that separate common from rare SVs and result in lower accuracy in clinical use[16].

By training on rare variants, we intend to achieve better accuracy in the challenge faced in SV discovery. To create a rare benign dataset that matches the size range of our pathogenic dataset, we included SVs common in great apes but not humans, which we assume should be benign in humans due to our recent shared ancestry with great apes. Our benign dataset also included unlabeled rare SVs from gnomAD-SVs. Although a small fraction of these unlabeled SVs may be pathogenic, we made two assumptions that mitigated this issue: (1) pathogenic SVs have been depleted by selection so the large majority of unlabeled SVs are benign, and (2) the fraction of truly pathogenic SVs in the pathogenic and benign training sets is sufficiently different for StrVCTVRE to learn important distinguishing features. By including these additional data sources, we brought the ratio of pathogenic to benign closer to 1:1 in our training set, even at small sizes. This would have been impossible with ClinVar data alone due to the dearth of small benign SVs in ClinVar.

To assess the appropriateness of including SVs from apes and gnomAD in our benign dataset, we explored how performance and feature importance changes with these data included. One predictor was trained only on ClinVar SVs, and another predictor was trained on ClinVar SVs, ape SVs, and gnomAD SVs (Fig. 2a Methods). Using leave-one-chromosome-out cross validation, we found both training sets performed similarly, supporting our theory that rare unlabeled gnomAD SVs and great ape SVs are sufficiently depleted in pathogenic variants to be used as a training set of rare, benign variants. Additionally, the predictor trained on all data showed a distribution of feature importance that is more robust and more evenly distributed

among feature categories. This includes a decrease in importance of gene-importance features, which are likely to be overrepresented in ClinVar data, and an increase in importance in coding sequence features, which are an important line of evidence for assessing SV pathogenicity[22].

Before training, all SVs were extensively cleansed to remove duplicate records within and between datasets, remove common variants, and remove variants larger than 3 Mb (see Methods). Pathogenic deletions and duplications were found to have a large size bias, likely due to the sensitivity of detection methods to specific size ranges (Fig. S1). To avoid training on this acquisition bias, putatively benign variants were sampled to match the pathogenic size distribution (Fig. 3). In our training data, we included only those pathogenic SVs that could be matched to a benign SV of similar size and same type (deletion or duplication). Using this matching strategy, we were able to include nearly all pathogenic deletions and duplications below 1 Mb. By including ape and gnomAD SVs, we were able to utilize pathogenic SVs below 10 kilobases (kb), a range nearly absent in ClinVar benign SVs. In the benign training set, 26% of deletions and 75% of duplications came from ClinVar benign or likely benign variants.

To accurately assess StrVCTVRE's performance, we used a held-out test set of ClinVar SVs on chromosomes 1, 3, 5, and 7 (~20% of the total ClinVar dataset). Only ClinVar SVs were used for testing, given it is the highest confidence dataset. The training set consisted of SVs from all three data sources on all remaining chromosomes. The training set consisted of 2,463 pathogenic SVs and 2,372 benign SVs, and the test set consisted of 244 pathogenic SVs and 334 benign SVs. The test set is of reduced size because pathogenic and benign SVs in the test set were matched on length. None of the SVs in the test set were used to develop the trained algorithm. At a sensitivity of 0.90, StrVCTVRE achieved a specificity of 0.54 (Fig. 2b) on the test set. StrVCTVRE performed equally well or better on test sets in which duplicates and common variants were not removed (Fig. S7).

<u>StrVCTVRE eliminates more than half of candidate SVs at 90% sensitivity</u>

In discriminating between pathogenic and putatively benign ClinVar SVs in the test dataset, StrVCTVRE performed substantially better than existing methods. Performance was measured using the area under the receiver operating characteristic curve (AUC). The AUC for StrVCTVRE was 0.823. The next best-performing method was SVScore (AUC = 0.710). StrVCTVRE improved notably in the classification of large duplications and deletions (> 1 MB), a regime in which SVScore by default classifies all SVs as pathogenic (lower left corner of Fig. 2b). Transcript consequence predicted by VEP (AUC = 0.431) performed worse than

random. This poor performance was largely due to VEP annotating more benign SVs than pathogenic SVs with its most deleterious sequence ontology term, transcript ablation (Fig. S3). The poor performance of transcript consequence from VEP reinforces the known limitations of prioritizing variants using sequence ontology terms in isolation. As we intend StrVCTVRE to be used to prioritize SVs seen in clinical cases, it needs to perform well in clinically relevant regimes. To avoid overlooking causal variants, clinical researchers must minimize false negatives, which requires high sensitivity. When compared to existing methods, StrVCTVRE makes substantial improvements in the high-sensitivity regime, as it is able to capture 90% of causal variants at a 46% false positive rate. This suggests that when used clinically, StrVCTVRE may identify 90% of causal SVs while reducing the candidate SV list by 54%.

Performance of StrVCTVRE on an independent test set

All held-out test SVs, and a large fraction of training SVs, come from a single database: ClinVar. To independently test StrVCTVRE, we collected pathogenic and benign SVs from DECIPHER, a public database to which clinical scientists submit SVs seen in patients with developmental disorders[23]. Because there is some overlap between training ClinVar SVs and DECIPHER SVs, we tested on DECIPHER using a leave-one-chromosome-out approach, in which 24 separate StrVCTVRE classifiers were developed, one for each chromosome. Specifically, DECIPHER variants on chromosome 1 were predicted by a StrVCTVRE classifier trained on chromosomes 2, 3, 4, etc. Additionally, we considered only DECIPHER SVs with a reciprocal overlap of less than 10% with any SV used in training StrVCTVRE. This strategy effectively removes concerns regarding training and testing on the same or similar variants. Because StrVCTVRE was trained on variants less than 3 Mb, and few benign variants larger than 3 Mb have been observed[24], all SVs larger than 3 Mb were scored as pathogenic (given a score of 1). Compared to its performance on the held-out test set, StrVCTVRE performed similarly well on the DECIPHER test set, although performance varied across SV size (Fig. 4a). On large SVs (> 500 kb), StrVCTVRE performed very well, partially because most of the SVs larger than 3 Mb are pathogenic. StrVCTVRE performed similarly well on small SVs (< 10 kb), likely due to the large amount of small SVs present in the training data. StrVCTVRE performed modestly on mid-length SVs, and is only slightly more accurate than SVScore.

StrVCTVRE eliminates the most benign variants seen in 221 individuals

Typically, patients with a rare disorder caused by SVs have one or two pathogenic SVs in their genome, and the remainder can be classified as benign. An ideal impact predictor would

prioritize the causal variants and eliminate from consideration as many of the benign variants as possible. To evaluate StrVCTVRE's performance in this scenario, we applied it to SVs called in 2,504 genomes identified by the 1000 Genomes Project phase 3[25] (Fig. 4b). Each genome can be treated as if it came from a proband with a rare disorder whose causal variants have been removed. 221 of these genomes had 1 or more homozygous rare exonic SVs, almost all of which should be benign. Using our leave-one-chromosome-out predictor at a 90% sensitivity threshold, StrVCTVRE identified an average of 59% of these putatively benign variants as benign, compared to 43% when SVScore was used at the same sensitivity (Wilcoxon paired-rank p = 8.06e-06). In a clinical setting, StrVCTVRE classifies more benign variants as benign, allowing clinicians and researchers to eliminate the most benign homozygous variants from consideration at 90% sensitivity.

StrVCTVRE performance improves when data are more reliable

There were undoubtably some mislabeled SVs in the ClinVar dataset used for training and testing. We expect that certain subsets of the data are more reliably labeled, and identified two subsetting methods: classification and supporting evidence. Consider, SVs classified as 'pathogenic' or 'benign' must meet stricter standards, and nominally indicate 99% confidence, compared with 'likely pathogenic' or 'likely benign' SVs, which indicate just 90% confidence. We therefore expected a subset containing only 'pathogenic' and 'benign' SVs would be more reliably labeled than the full dataset. Additionally, we expected that SVs that were submitted with a minimum level of supporting evidence would be more reliably labeled than those submitted with none. As shown in Fig. 5a, we found a consistent shift towards more accurate StrVCTVRE classification in subsets that are more reliably classified. Since StrVCTVRE's performance improves on presumably more reliable data, we have reason to believe it is an accurate method.

In uncataloged variants, SV impact predictors are more accurate than SV annotators

Since probands with the same disorder often have SVs affecting the same genome element, and even recurrent pathogenic de novo SVs are known to occur[26], one strategy used to prioritize SVs is to annotate them with SVs of known pathogenicity. AnnotSV is a popular method used to identify pathogenic SVs based on their overlap with both known pathogenic SVs in the National Center for Biotechnology Information's dbVar, as well as genes associated with disease. This method is more accurate than SV impact predictors for a proband whose causal SV overlaps a cataloged pathogenic dbVar SV (Fig. S6). Yet, many probands have novel causal

SVs that are not cataloged. To compare the performance of AnnotSV and SV impact predictors on novel variants, we created a dataset of DECIPHER SVs that do not overlap dbVar SVs used by AnnotSV, and we recorded the prediction accuracy of each method (Fig. 5b). As expected, AnnotSV performed little better than chance. Yet both StrVCTVRE and SVScore showed significant predictive power, since they consider the properties of an SV and its genomic context, rather than just overlap with cataloged SVs. Researchers searching for SVs in novel disease genes will significantly improve their predictive power by using an SV predictor like StrVCTVRE instead of an SV annotator.

Interpreting StrVCTVRE scores

The StrVCTVRE score ranges from 0 to 1, reflecting the proportion of decision trees in the random forest that classify an SV as pathogenic. Note that StrVCTVRE scores are not probabilities. We recommend against using StrVCTVRE scores as a threshold, and instead recommend that greater consideration be given to SVs with greater StrVCTVRE scores. Yet, many variants are classified using the guidelines for sequence variant interpretation recommended by the American College of Medical Genetics and Genomics (ACMG)[27], in which thresholds are currently required for computational tools. Following work done by Tavtigian et al.[28] to model the ACMG guidelines in a Bayesian framework, we provide a table of StrVCTVRE scores corresponding to different strengths of evidence (Table 1).

	Benign			Pathogenic			
ACMG Evidence	Stand-alone	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Odds Pathogenic	0	0.053	0.48	2.1	4.3	19	350
StrVCTVRE Score	NR	NR	≤0.322	≥0.642	≥0.852	NR	NR

Table 1. StrVCTVRE scores corresponding to ACMG evidence strength.
*NR = Not recommended

Discussion

As exome sequencing continues to expand, clinical researchers face a challenge in identifying one or two causal SVs in the dozens identified by sequencing. The ACMG recently offered

guidelines for classifying SVs, acknowledging that classification is complex and many pathogenic SVs will be classified as variants of uncertain significance due to incomplete knowledge[22]. SV impact predictors can address this challenge, but few SV impact predictors exist. Although SVs comprise a significant fraction of the loss-of-function mutations that cause rare disease, fewer than 10,000 have been cataloged in ClinVar, which leads to acquisition bias that hinders predictor development. Additionally, it's not clear which features are most useful when classifying SVs and how to address the large size range of SVs. StrVCTVRE is the first method to address these problems by predicting the impact of exonic deletions and duplications in rare genetic disorders. We overcame data limitation and bias by both combining SVs across multiple data sources as well as matching pathogenic and benign variants by size. Since clinical researchers must recognize causal SVs among dozens of rare SVs detected in a proband, we trained only on rare SVs and showed that StrVCTVRE outperforms other methods in this regime.

Determining whether a single SV is pathogenic requires consideration of numerous features in combination, as demonstrated by the recent ACMG SV guidelines. Independent of these guidelines, our method identified important features in cataloged SVs. Our findings reinforce the clinical intuition that researchers bring, while also highlighting new areas to explore. Both StrVCTVRE and the ACMG guidelines found gene importance and coding-region disruptions to be critical for SV interpretation. Additionally, StrVCTVRE highlighted two features not discussed in the guidelines: conservation and expression. We found expression in particular is both algorithmically important and uncorrelated with all other features, suggesting it captures unique information for determining pathogenicity. More widespread consideration of expression features could be beneficial for SV classification.

StrVCTVRE additionally identified features that are not useful, such as TAD boundary strength and whether there is a gain or loss of DNA. This is consistent with the ACMG guidelines, which do not consider TAD boundaries and give very similar scoring metrics for both copy gain and copy loss. Since SVs range from 50 bp to > 100 Mb, it is challenging to accurately classify SVs across this range. Benign SVs in ClinVar are mainly > 10 kb, but accurate prediction requires training on benign variants from a wide range of sizes. We accomplished this by training on small benign SVs from great apes and gnomAD. When tested on an independent test set, StrVCTVRE made significant improvements over current methods in both small and large SVs. To be helpful in a clinical setting, a method must perform well at moderately high sensitivity. StrVCTVRE satisfies this requirement, and is able to remove 57% of homozygous variants from

consideration at a sensitivity of 90% in the 1000 Genomes Project dataset. Overall, we found StrVCTVRE is the best-performing SV impact predictor, outperforming SVScore in most tasks. This is unexpected given that SVScore derives its scores from CADD, a method trained on > 1,000-fold more variants.

StrVCTVRE is accessible to the widest possible audience. There is both a web-based version (forthcoming) and downloadable command line version (see Data Availability). Whereas SVScore requires users to download an 80 gigabyte (Gb) CADD file, StrVCTVRE only requires a 9 Gb phyloP file. While SNV scores can be pre-computed, SVs must be scored anew, and thus require efficient scoring methods. StrVCTVRE runs rapidly and annotates 10,000 SVs in two minutes, while SVScore annotates the same SVs in three hours.

Following existing predictors, StrVCTVRE predicts the pathogenicity of a variant in isolation. Yet human biology complicates this picture through zygosity and dominance. Since zygosity is not reported for most variants in ClinVar, StrVCTVRE is zygosity-naïve. Additionally, StrVCTVRE's pathogenic training dataset consists largely of variants in genes predicted to lead to dominant disorders (Fig. S4). For these reasons, StrVCTVRE may give lower scores to recessive mutations. Researchers who suspect a recessive mode of inheritance may need to consider lower thresholds for SVs scored by StrVCTVRE and consider these scores in tandem with impact scores from SNVs in trans in the same gene. Since StrVCTVRE considers only the genomic context, it may be helpful to consider StrVCTVRE scores in tandem with one of the many methods that assess the match between patient phenotype and known phenotypes for an affected gene[29-31].

A method is only as good as the data it trains on. StrVCTVRE is limited by the relatively small number of identified pathogenic and putatively benign SVs, as well as the over-representation of certain genes in the dataset (Fig. S8). These data limitations almost certainly curtail the ultimate performance of our approach. We concede that our method is unable to classify inversions and insertions due to limited data; however, these have been shown to contribute to a minority of the pLoF events caused by SVs[11]. We are hopeful that additional clinical sequencing studies will identify a more diverse range of SVs, which will be cataloged in open resources such as ClinVar and leveraged to make more accurate models. We look forward to greater non-coding genome annotations, which will expand our understanding and cataloging of pathogenic noncoding SVs, which remain vexing to classify.

Much of the focus in SV algorithms has been on methods to accurately detect SVs. These methods have left clinical researchers awash with new SVs not previously known. As experimental methods and algorithms advance, SV detection will improve, but SV interpretation will increase in difficulty. StrVCTVRE fills this empty niche in the clinical evaluation of SVs. During genome sequencing analysis, some cases contain an SV that matches a cataloged pathogenic SV or clearly satisfies the conditions for pathogenicity set forth in the ACMG SV guidelines. However, in the many cases in which no SV is immediately promising, StrVCTVRE's niche is to aid clinical researchers in rapidly identifying compelling SVs. In these cases, we anticipate StrVCTVRE will first be used to prioritize SVs for manual investigation, saving researchers valuable time. Then, if a case remains unresolved by manual investigation, StrVCTVRE's predictive capacity can be used to prioritize SVs in novel disease genes for experimental validation. This will empower researchers to identify novel disease genes where haploinsufficiency and triplosensitivity were not previously known causes of disease. Rapid adoption of variant impact predictors will enable researchers to make the most of these new data to improve both patient care and our understanding of basic biology.

Methods

Training, validation, and test datasets

All variants were retrieved in GRCh38 or converted using the University of California, Santa Cruz (UCSC) liftover tool.

All ClinVar SVs were downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz on January 21, 2020. Variants were retained if they fulfilled all the following requirements: clinical significance of pathogenic, likely pathogenic, pathogenic/likely pathogenic, benign, likely benign, or benign/likely benign; not somatic in origin; type of copy number loss, copy number gain, deletion, or duplication; > 49 bp in size; at least 1 bp overlap with exon.

Great ape SVs were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/dgva/estd235_Kronenberg_et_al_2017/vcf/ on April 8, 2019. Deletions were retained if they were absent in humans and homozygous in exactly two of the following species: chimpanzee, gorilla, or orangutan. Only exonic deletions > 49 bp were retained. These deletions are subsequently referred to as *apes.*

gnomAD 2.1.1 SVs (build GRCh37) were downloaded from https://storage.googleapis.com/gnomad-public/papers/2019-sv/gnomad_v2.1_sv.sites.vcf.gz on June 28, 2019. Only duplications and deletions with PASS Filter were retained. Only exonic SVs > 49 bp were retained. gnomAD variants were broken into three categories: SVs with a global minor allele frequency (MAF) > 1% (*gnomAD common*), SVs with a global MAF < 1% with at least one individual homozygous for the minor allele (*gnomAD rare benign*), and SVs with a global MAF < 1% with no individuals homozygous for the minor allele (*gnomAD rare unlabeled*).

Release 2016-05-15 of GRCh38 "DGV Variants" from Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/downloads) was downloaded on April 08, 2019. Allele frequency of each deletion was calculated as 'observedlosses' / (2 * 'samplesize'). Allele frequency of each duplication was calculated as 'observedgains' / (2 * 'samplesize'). Only exonic SVs > 49 bp were retained. Those SVs with an allele frequency greater than 1% are subsequently referred to as *DGV common.*

DECIPHER CNVs (build GRCh37) were downloaded on Jan 27, 2020 from http://sftpsrv.sanger.ac.uk/. Only exonic variants 50 bp or larger with pathogenicity of "pathogenic", "likely pathogenic", "benign", or "likely benign" were retained. Identical variants with conflicting pathogenicity were removed.

1000 Genomes Project merged SVs were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/GRCh38_positions/ on Oct 22, 2019. Only exonic deletions and duplications with a global allele frequency less than 1% were used for testing.

We used exon boundaries from Ensembl biomart, genes v96, Human genes GRCh38.p12, limited only to genes with HGNC Symbol ID(s) and APPRIS annotation. For each gene, a single principal transcript was used, based on the highest APPRIS annotation. For transcripts that tied for highest APPRIS annotation, the longest transcript was used. Exon overlap was determined using bedtools intersect.

Extensive deduplication of data was performed as follows. Benign SVs were ordered (ClinVar benign, ClinVar likely benign, apes, gnomAD rare benign, gnomAD rare unlabeled) and duplicates (reciprocal overlap of 90% or greater) were removed, keeping the first appearance of an SV. Deletions and duplications were considered separately. This removed 577 SVs from ClinVar benign/likely benign, 5 SVs from apes, and 408 SVs from gnomAD. The retained data are subsequently referred to as *benign.* To deduplicate pathogenic SVs, exact matches

between ClinVar pathogenic and ClinVar likely pathogenic were removed from likely pathogenic. SVs were then sorted by size, ascending. SVs with > 90% reciprocal overlap were removed, keeping the smallest variant. Deletions and duplications were considered separately. This removed 2,421 pathogenic SVs. The retained data are subsequently referred to as *pathogenic.* Next, exact matches between benign and pathogenic datasets were removed from both datasets. Finally, duplicates between pathogenic and benign (reciprocal overlap of 90% or greater) were removed from the pathogenic dataset. This removed 3 benign SVs and 82 pathogenic SVs.

Data were processed as follows to assure we trained only on rare SVs. Pathogenic and benign SVs that exactly matched a DGV common SV were removed. Pathogenic and benign SVs that were very similar (reciprocal overlap > 90%) to an SV in gnomAD common were removed. This removed 30 benign SVs and 1 pathogenic SV. SVs between 50 bp and 3 Mb were retained, all others were removed.

To assure StrVCTVRE was not learning acquisition bias caused by the sensitivity of different sequencing methods (Fig. S1), the size distribution of benign and pathogenic variants were matched with the following procedure. Benign variants were organized into five tiers: ClinVar likely benign; ClinVar benign; apes; gnomAD rare benign; and gnomAD rare unlabeled. Each pathogenic variant was then matched by size and type (DEL or DUP) to a benign variant, iterating through each tier. Specifically, each pathogenic variant of size *N* seeks a benign variant of the same type in the bin [N - (N / α + 20), N + (N * α + 20)] where α = $\sqrt[101]{10^6}$ (this bin size derived from Ganel et al.[14]). A pathogenic SV first seeks a benign variant in the first tier. If matched, the pathogenic and benign SVs are included in the training set, and the benign variant cannot match any further pathogenic SVs. If no match is found in the first tier, the same process is repeated while progressing through further tiers. Pathogenic variants that do not find a match in any tier are not included in the final training set. This process was continued for all pathogenic SVs and the resulting data are shown in Fig. 3.

After SVs were annotated with features (see below), SVs with identical features were removed from the training set. For feature-identical SVs that were present in both the pathogenic and the benign datasets, all feature-identical SVs were removed. For feature-identical SVs that were present in only the benign dataset or only the pathogenic dataset, a single SV was retained, and duplicates were removed.

<u>Structural variant impact predictors</u>

VEP v96 was downloaded on April 16, 2019, and used to annotate SVs with transcript consequence sequence ontology terms. SVScore v0.6 was downloaded on June 16, 2019. It was run using CADD v1.3, downloaded on June 16, 2019, using all default settings. AnnotSV v2.3.2 was downloaded on Feb 27, 2020. AnnotSV was run using human annotation and default settings.

Structural variant features

All gene and exon boundaries used to determine features came from Ensembl Genes v96 as described above. Each SV was annotated with the following 17 features:

Feature description	**Data type**	**Aggregation method for multiple genes**
SV is deletion or duplication	boolean	NA
Average phyloP score of the 400 most conserved overlapping nucleotides	float	NA
Number of exons SV overlaps by 1 or more bp	integer	max
Minimum exon transcript rank*	integer	min
Number of exons in canonical transcript of gene	integer	min
All overlapped exons can be skipped in frame	boolean	NA
Any overlapped exon is constitutive	boolean	NA
pLI of gene	float	max
LOEUF of gene	float	min
Exon expression (see below)	float	NA
Exon inclusion (see below)	float	NA
TAD boundary strength (according to Gong et al[32])	float	max
Fraction of CDS adjacent to start codon that is not disrupted by SV	float	min
Fraction of CDS adjacent to stop codon that is not disrupted by SV	float	min
Fraction of CDS overlapping SV	float	max
pLI of gene where start codon overlaps SV or >50% of CDS overlaps SV	float	max
LOEUF of gene where stop codon overlaps SV or >50% of CDS overlaps SV	float	min

*The exon transcript rank was defined as the number of exons preceding a given exon in a gene.

Expression features were derived from transcript data downloaded from the GTEx Portal v7[33]. Exon expression was calculated for each nucleotide as the sum of the transcripts per million (TPM) of fragments that map to that nucleotide. Exon inclusion estimated the proportion of transcripts generated by a gene that include a given nucleotide and was calculated for each nucleotide as the TPM of fragments that map to that nucleotide, divided by the sum of TPM that

map to the gene containing that nucleotide. For both features, adjacent base pairs with the same value were merged together into genomic intervals. For SVs that overlapped more than one of these genomic intervals, exon expression was calculated by averaging the 400 highest exon expression genomic intervals contained in that SV. The same was done for exon inclusion. All GTEx tissues were used in this analysis.

To determine which conservation feature to use, we assessed the accuracy of both PhastCons[34] and PhyloP[18] in discriminating between pathogenic and benign SVs using the average of the highest-scoring 200, 400, 600, and 800 nucleotides (Fig. S5). The test set consisted of 200 small (< 800 bp) SVs. We found the mean PhyloP score of the 400 most conserved nucleotides in an SV gave the best accuracy. For both conservation and expression features, if the total overlap between SV and exons was less than 400 intervals, then the remaining overlapped intervals were averaged together to generate the feature. Median imputation was used to fill in missing feature annotations.

In Fig. 1a, features were clustered by correlation using the linkage and fcluster functions from SciPy's hierarchical clustering package. Values for some features were reversed to ensure most matrix correlations are positive.

Random forest classification

StrVCTVRE was implemented as a random forest classifier in Python with scikit-learn v0.17[35], using class RandomForestClassifier. A grid search was performed to find the optimal hyperparameters by using a leave-one-chromosome-out cross validation strategy and validation only on ClinVar data, as described previously. The hyperparameters searched included: the max depth of a tree (5, 10, 15, No limit), max features considered at each split (1, 2, 3, 4), the minimum samples at each leaf node (1, 2, 4), the minimum samples required to split a node (2, 4), the number of trees generated (500, 1000, 3000), and whether to use out-of-bag samples to estimate accuracy (True, False). Several combinations of features performed similarly well, and we chose one that performed well while unlikely to over-fit to the training data—max depth: 10, max features considered at each split: 1, minimum samples at each leaf node: 2, minimum samples required to split a node: 4, number of trees: 1,000. Feature importance used in figures is also known as Gini importance[36], and was calculated using the feature_importances_ attribute of RandomForestClassifier.

Figures

In Fig. 4b, 90% sensitivity thresholds were taken from StrVCTVRE and SVScore performance on the held-out test set (Fig. 2b). In Fig 5a, testing datasets were derived from ClinVar variants as described above, except no size matching was performed. To maintain a balanced ratio of pathogenic to benign variants, variants were oversampled from the smaller dataset.

Calculation of the StrVCTVRE threshold corresponding to each ACMG evidence category

The odds pathogenic for each ACMG evidence category were derived from Tavtigian et al.[28]. Corresponding StrVCTVRE scores were calculated from performance on DECIPHER database SVs. StrVCTVRE score thresholds were determined by sorting test SVs by StrVCTVRE score, calculating the odds pathogenic for SVs in a size 100 sliding window, and using the most conservative threshold at which the sliding window odds pathogenic fell below the ACMG odds pathogenic.

Data availability

StrVCTVRE command line tool can be downloaded from https://github.com/andrewSharo/StrVCTVRE along with all source code for the classifier. A job submission version of StrVCTVRE will be made available at http://compbio.berkeley.edu/proj/strvctvre/.

Acknowledgements

A.G.S. was supported by a National Science Foundation Graduate Research Fellowship under Grand No. DGE 1752814. We thank Dr. Nilah Ioannidis for helpful feedback during the project and for valuable comments on the manuscript. We thank Dr. Aashish Adhikari for insightful comments in the initial planning of the project. We thank Azza Althagafi for thorough testing of our github resources, as well as Lindsey Guan, Reet Mishra, and Ashish Ramesh for early script testing. This study makes use of data generated by the DECIPHER Consortium. A full list of centers who contributed to the generation of the data is available from http://decipher.sanger.ac.uk and via email from decipher@sanger.ac.uk. Funding for the DECIPHER project was provided by the Wellcome Trust. Dr. Steven Brenner was unable to review this manuscript in its entirety due to a medical leave.

References

1. Clark MM, Stark Z, Farnaes L et al. 2018. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ genomic medicine 3*.
2. Lappalainen T, Scott AJ, Brandt M, Hall IM. 2019. Genomic analysis in the age of human genome sequencing. *Cell 177*:70-84.
3. Holt JM, Birch CL, Brown DM et al. 2019. Identification of Pathogenic Structural Variants in Rare Disease Patients through Genome Sequencing. *bioRxiv*:627661.
4. Wu N, Ming X, Xiao J et al. 2015. TBX6 null variants and a common hypomorphic allele in congenital scoliosis. *New England Journal of Medicine 372*:341-350.
5. Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annual Review of Medicine 61*:437-455.
6. Wright CF, McRae JF, Clayton S et al. 2018. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genetics in Medicine 20*:1216.
7. Sanchis-Juan A, Stephens J, French CE et al. 2018. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short-and long-read genome sequencing. *Genome medicine 10*:95.
8. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome biology 20*:117.
9. Karczewski KJ, Francioli LC, Tiao G et al. 2019. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*:531210.
10. Abel HJ, Larson DE, Chiang C et al. 2018. Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *bioRxiv*:508515.
11. Collins RL, Brand H, Karczewski KJ et al. 2019. An open resource of structural variation for medical and population genetics. *bioRxiv*:578674.
12. McLaren W, Gil L, Hunt SE et al. 2016. The ensembl variant effect predictor. *Genome Biology 17*:122.
13. Cingolani P, Platts A, Wang LL et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly 6*:80-92.
14. Ganel L, Abel HJ, Consortium F, Hall IM. 2017. SVScore: an impact prediction tool for structural variation. *Bioinformatics 33*:1083-1085.
15. Kumar S, Harmanci A, Vytheeswaran J, Gerstein MB. 2019. SVFX: a machine-learning framework to quantify the pathogenicity of structural variants.
16. Li M-X, Kwan JS, Bao S-Y et al. 2013. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genetics 9*:e1003143.
17. Lek M, Karczewski KJ, Minikel EV et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature 536*:285.
18. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research 20*:110-121.
19. Ioannidis NM, Rothstein JH, Pejaver V et al. 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics 99*:877-885.
20. Landrum MJ, Lee JM, Benson M et al. 2017. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research 46*:D1062-D1067.

21. Kronenberg ZN, Fiddes IT, Gordon D et al. 2018. High-resolution comparative analysis of great ape genomes. *Science 360*:eaar6343.
22. Riggs ER, Andersen EF, Cherry AM et al. 2019. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genetics in Medicine*:1-13.
23. Firth HV, Richards SM, Bevan AP et al. 2009. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics 84*:524-533.
24. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research 42*:D986-D992.
25. Sudmant PH, Rausch T, Gardner EJ et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature 526*:75.
26. Sanders SJ, Ercan-Sencicek AG, Hus V et al. 2011. Multiple recurrent de novo CNVs, including duplications of the 7q11. 23 Williams syndrome region, are strongly associated with autism. *Neuron 70*:863-885.
27. Richards S, Aziz N, Bale S et al. 2015. 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. *Genetics in Medicine 17*:405.
28. Tavtigian SV, Greenblatt MS, Harrison SM et al. 2018. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine 20*:1054.
29. Köhler S, Schulz MH, Krawitz P et al. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics 85*:457-464.
30. Singleton MV, Guthery SL, Voelkerding KV et al. 2014. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *The American Journal of Human Genetics 94*:599-610.
31. Zemojtel T, Köhler S, Mackenroth L et al. 2014. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science translational medicine 6*:252ra123-252ra123.
32. Gong Y, Lazaris C, Sakellaropoulos T et al. 2018. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nature communications 9*:542.
33. Lonsdale J, Thomas J, Salvatore M et al. 2013. The genotype-tissue expression (GTEx) project. *Nature Genetics 45*:580.
34. Siepel A, Bejerano G, Pedersen JS et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research 15*:1034-1050.
35. Pedregosa F, Varoquaux G, Gramfort A et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research 12*:2825-2830.
36. Hastie T, Tibshirani R, Friedman J. 2009. *The elements of statistical learning: data mining, inference, and prediction*: Springer Science & Business Media.

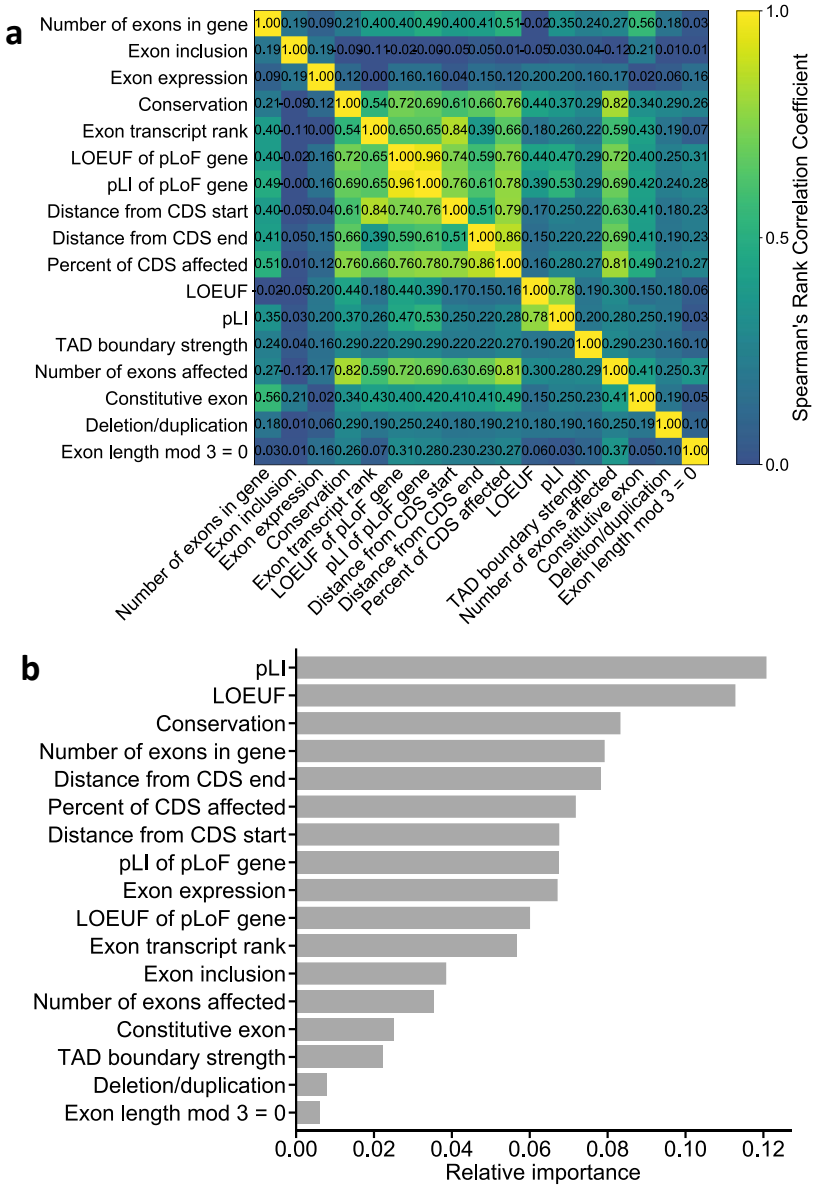

Fig. 1. By considering feature clustering and importance, we can identify features providing unique and important information. **a,** Correlation matrix of StrVCTVRE features. Features were ordered by hierarchical clustering, and some values were reversed to reduce negative correlation between features. Float values represent Spearman's rank correlation between features. **b,** Feature importance of StrVCTVRE features. Feature importance was estimated using mean decrease in impurity (Gini importance). Note that expression features had high importance yet were uncorrelated with all other features, suggesting they capture unique and useful information.

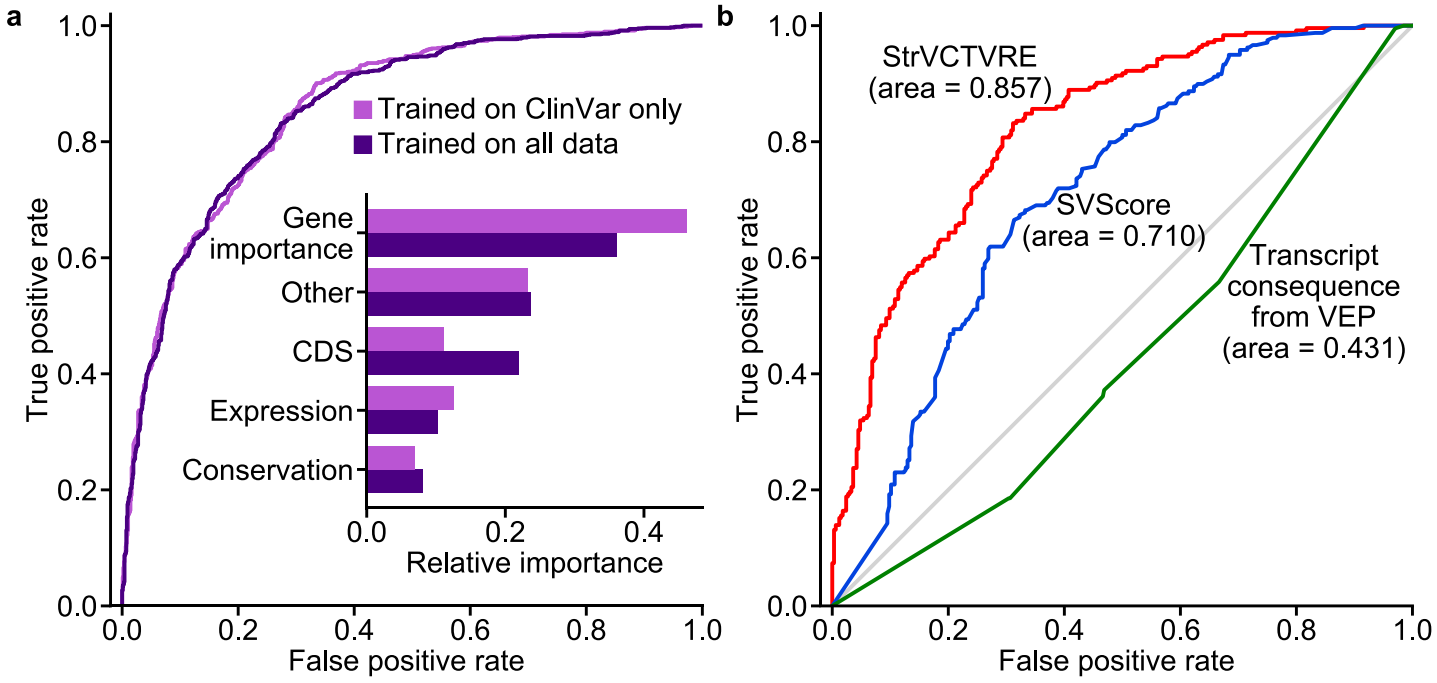

Fig. 2. By training on multiple datasets, StrVCTVRE learned more diverse feature importances and performed well on a held-out test set. **a,** Receiver operative characteristic comparing models trained on two different benign datasets: ClinVar in light purple, and all data (ClinVar, SVs common to apes but not humans, and rare gnomAD SVs) in dark purple. When validated only on ClinVar data, performance is not significantly different. However the feature importances (inset) of the classifier trained on all data were signficantly more evenly distributed among feature categories. This suggests that unlabeled rare SVs and common ape SVs are a suitable benign training set. **b,** Receiver operative characteristic comparing StrVCTVRE (red) to other methods on a held-out test set comprised of ClinVar SVs on chromosomes 1, 3, 5, and 7.

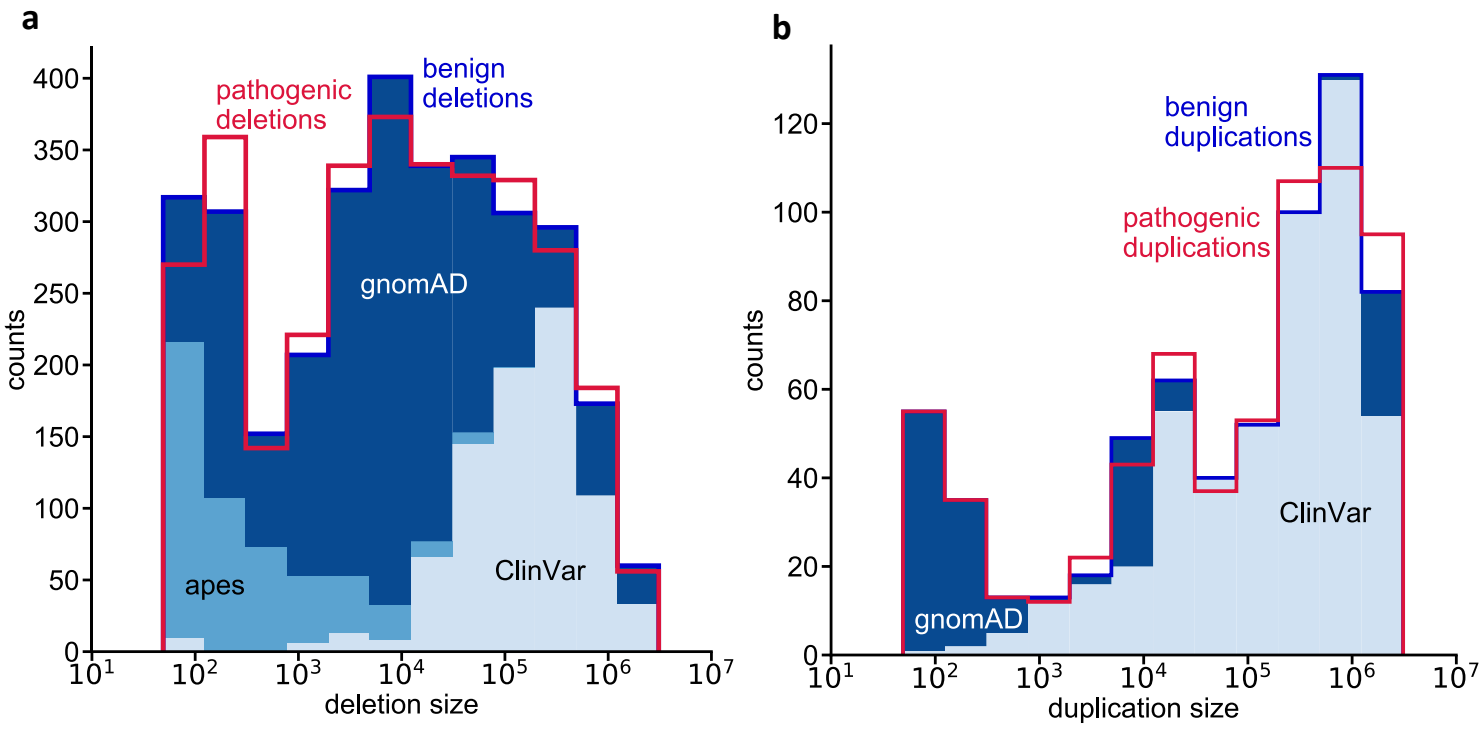

Fig. 3. Benign training SVs closely match the size distribution of pathogenic SVs and were drawn from multiple datasets. Histogram of pathogenic and benign (**a**) deletions and (**b**) duplications. **a,** Benign deletions are composed of 26% ClinVar, 16% apes, and 58% gnomAD. **b,** Benign duplications are composed of 75% ClinVar and 25% gnomAD. Use of apes and gnomAD SVs allowed us to include more small pathogenic SVs in our training data.

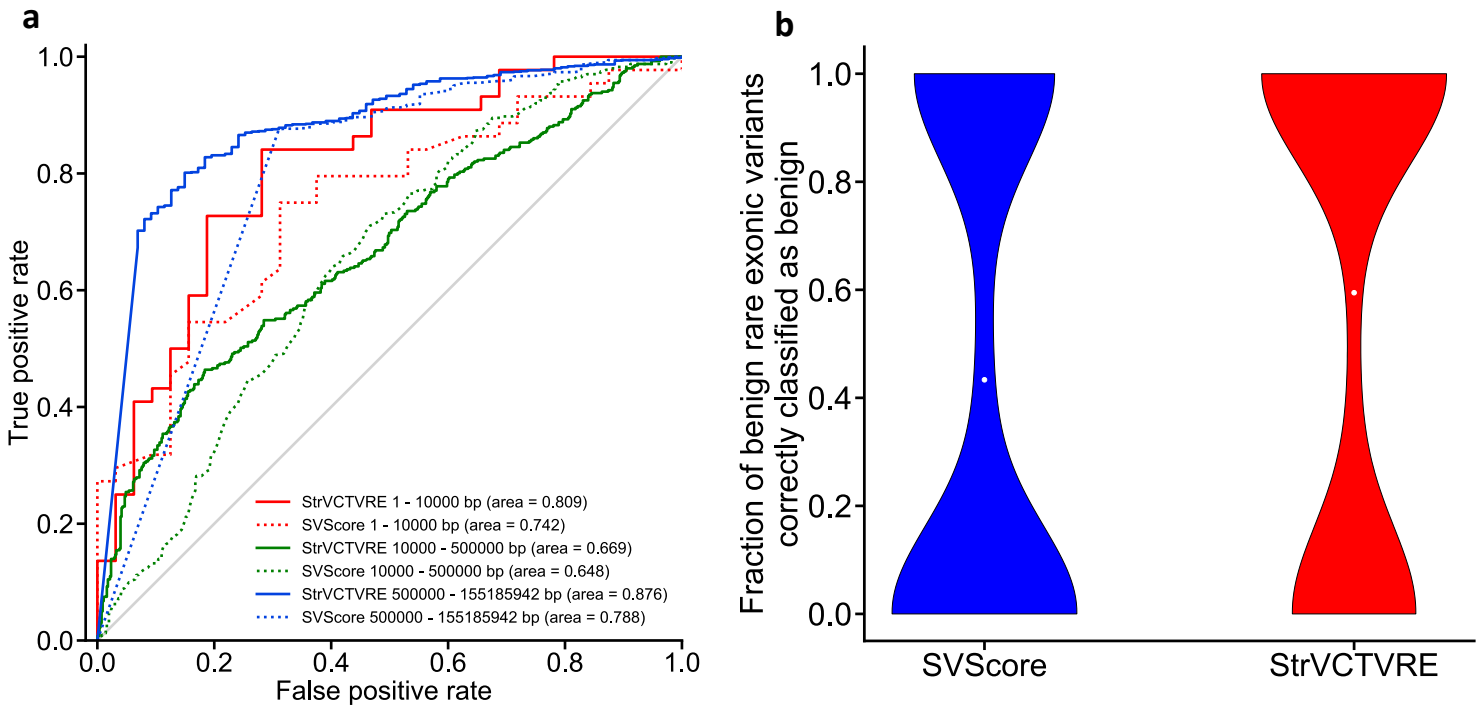

Fig. 4. a, Across three size ranges, StrVCTVRE accurately classified variants in an independent test set. In this AUC comparison of StrVCTVRE and SVScore, three size ranges of SVs were considered. StrVCTVRE performed best on the large (blue) and small (red) SVs, while performing comparable to SVScore on mid-sized variants (green). **b**, StrVCTVRE eliminated a significantly larger fraction of benign SVs from consideration. When tested on rare exonic SVs from 221 putatively healthy individuals, StrVCTVRE was able to correctly classify the most benign variants. White dots represent mean values. For both methods, the threshold for variant consideration was at 90% sensitivity.

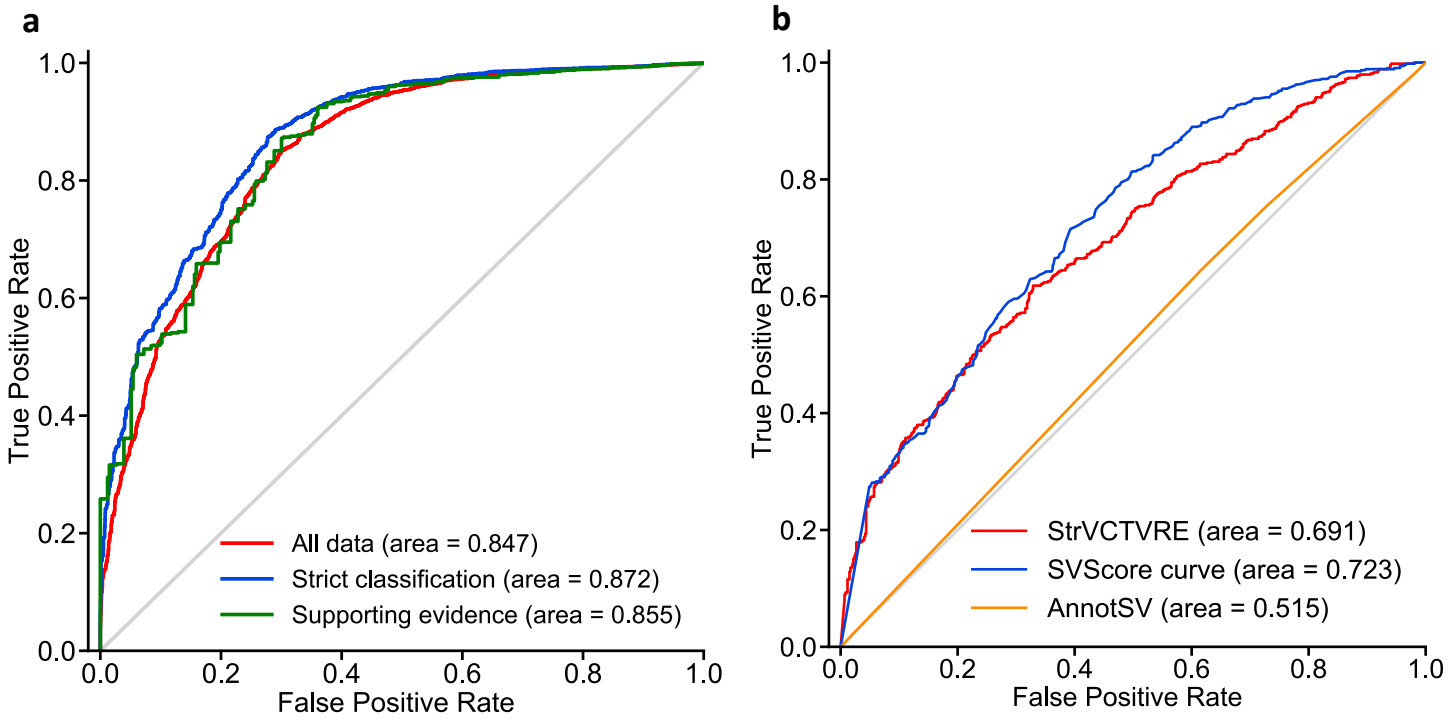

Fig. 5. a, When presented with data that is more reliably classified, StrVCTVRE's performance improved. ROC plot showing StrVCTVRE's performance on all ClinVar data (red) compared to performance on two datasets: (1) only ClinVar SVs classified as 'Pathogenic' or 'Benign', and (2) only ClinVar SVs with supporting evidence provided. **b**, ROC comparing two prediction-based methods (StrVCTVRE and SVScore) to one annotation-based method (AnnotSV). AnnotSV ranks an SV as 'pathogenic' or 'likely pathogenic' when the SV overlaps a catalogued pathogenic SV or known disease gene. To generate this figure, all SVs overlapping one of AnnotSV's catalogued pathogenic SVs were removed from the dataset. As expected, AnnotSV performs barely better than chance. In contrast, the prediction-based methods perform well because they consider the genomic context of the variants.