Published online 7 April 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 2 **1**
https://doi.org/10.1093/nargab/lqac025

Machine-learning of complex evolutionary signals improves classification of SNVs

Sapir Labes[1], Doron Stupp[1,†], Naama Wagner[2,†], Idit Bloch[1], Michal Lotem[3], Ephrat L. Lahad[4,5], Paz Polak[6], Tal Pupko[2] and Yuval Tabach[1,*]

[1]Department of Developmental Biology and Cancer Research, Institute for Medical Research Israel-Canada, Faculty of Medicine, and Hadassah University Medical School, The Hebrew University of Jerusalem, Jerusalem9112001, Israel, [2]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel, [3]Sharett Institute of Oncology, Hadassah University Medical Center, The Hebrew University of Jerusalem, Jerusalem9112001, Israel, [4]Medical Genetics Institute, Shaare Zedek Medical Center, Jerusalem9103102, Israel, [5]Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem9112001, Israel and [6]Oncological Sciences, Icahn School of Medicine at Mount Sinai, NY10029, USA

Received October 24, 2021; Revised February 08, 2022; Editorial Decision February 28, 2022; Accepted March 28, 2022

ABSTRACT

Conservation is a strong predictor for the pathogenicity of single-nucleotide variants (SNVs). However, some positions that present complex conservation patterns across vertebrates stray from this paradigm. Here, we analyzed the association between complex conservation patterns and the pathogenicity of SNVs in the 115 disease-genes that had sufficient variant data. We show that conservation is not a one-rule-fits-all solution since its accuracy highly depends on the analyzed set of species and genes. For example, pairwise comparisons between the human and 99 vertebrate species showed that species differ in their ability to predict the clinical outcomes of variants among different genes using conservation. Furthermore, certain genes were less amenable for conservation-based variant prediction, while others demonstrated species that optimize prediction. These insights led to developing Evo-Diagnostics, which uses the conservation against each species as a feature within a random-forest machine-learning classification algorithm. EvoDiagnostics outperformed traditional conservation algorithms, deep-learning based methods and most ensemble tools in every prediction-task, highlighting the strength of optimizing conservation analysis per-species and per-gene. Overall, we suggest a new and a more biologically relevant approach for analyzing conservation, which improves prediction of variant pathogenicity.

INTRODUCTION

The revolution of high-throughput sequencing and next-generation sequencing (NGS) generated massive human and vertebrate genomic data. This, in turn, promoted countless discoveries that improved healthcare personalization (1). For example, it enabled detection of alleles associated with genetic diseases, thus promoting accurate diagnostics and personalized medicine (2).

However, NGS frequently recognizes human variants with undetermined effects on health, denoted as variants of uncertain significance (VUSs) (3). Since the clinical implications of VUSs determine the risk for genetic diseases and influence patient management (4), it is essential to predict the deleteriousness of these VUSs. For example, pathogenic mutations in the genes BRCA1 and BRCA2 are associated with 45–65% risk of breast cancer and 11–39% risk of ovarian cancer (5). Patients at high risk for hereditary breast and ovarian cancer (HBOC) are treated with preventive surgeries, chemotherapy or radiological follow ups (6–8). Hence, estimating the pathogenicity of VUSs in BRCA genes helps prevent unnecessary invasive treatments and emotional distress (9). The importance of correctly characterizing the effects of VUSs extends to many genes in addition to BRCA1 and BRCA2, for example genes involved in Lynch syndrome (10).

Numerous algorithms have been developed to address the challenge of predicting the clinical implications of VUSs

*To whom correspondence should be addressed. Tel: +972 543973641; Email: yuvaltab@ekmd.huji.ac.il
†The authors wish it to be known that, in their opinion, these authors should be regarded as Joint Second Authors.
Present addresses:
Doron Stupp, Currently at Google, Tel Aviv, 6789141, Israel.
Paz Polak, C2i Genomics, New York, NY 10014, USA.

(11,12). These algorithms estimate the significance of variants either by the degree of conservation (13–15), function prediction (16–20), genomic annotations (21,22), or by an ensemble of these approaches (23–26). Among them, conservation-based algorithms were estimated to be the best individual prediction approach in most variant prediction models (23,27).

Evolutionary conservation of specific loci within a sequence is the result of purifying selection on functionally and structurally important sites. Other sites evolve neutrally or nearly neutral, and mutations in these sites have negligible effects on fitness. Therefore, the functional importance of genomic regions can be inferred from their conservation across species. As such, the increase in genomic data has the potential to improve conservation-based prediction tools (28).

It is nontrivial to translate conservation into accurate clinical prediction of variants. In general, a human VUS that is found in the wild-type sequence of chimpanzees is likely not to be of clinical importance. However, if one finds the human VUS in the wild-type sequence of a more distant species, such as zebrafish, the likelihood that this VUS has clinical importance increases. Thus, comparing human VUSs to different species and the inference of clinical outcomes from such pairwise comparisons may highly depend on the level of evolutionary divergence between the compared sequences. The level of divergence, in turn, depends on the time of divergence among the species and on species-to-species differences in nucleotide substitution rates (29,30). Ideally, a comparison of the human VUSs to multiple organisms should be analyzed simultaneously. However, finding the best way to integrate conservation information across the tree of life is a challenging task. For example, several methodologies have integrated the entire conservation profile to a single conservation score using the same algorithm across all genes (13–15). Such a one-rule-fits-all approach may be problematic given that genes differ in their conservation degree across the evolution, and thus the association between conservation of variants and their pathogenicity may vary among genes. Moreover, the evolutionary rate of a gene may itself vary in different parts of the phylogenetic tree. In addition, even within a specific gene, certain sites may have an accelerated or decelerated evolutionary rate in several clades. One model that accounts for lineage-specific evolutionary rates is the covarion model (29,30). PhyloP, a conservation-based prediction method, partially integrates the covarion model by allowing a single lineage to have a different evolutionary rate than the rest of the phylogenetic tree (15). Even when a conservation score is defined, one should still determine how to transform this score to predictions regarding the clinical outcomes of VUSs, e.g. by defining cutoffs for considering positions as conserved.

Phylogenetic profiling is another common method used to study complex evolutionary information at the levels of genes and proteins (31–35). Phylogenetic profiles, which represent the presence or absence of entire genes (or proteins) among a set of species, can be used to identify genes (or proteins) with analogous functions based on the similarity of their profiles. The rationale of this concept is that genes that share a function, as well as genes that are essential for survival under similar environmental conditions, have been co-evolving (i.e. coordinately changed, or lost and retained) across the tree of life. These profiles are utilized for drug repositioning (36,37) and for predicting protein–protein interactions (38–40), protein complexes (38), genes that participate in common-pathways (32–34,41,42), functional sites (31) and novel disease genes (32,33,41). In recent studies (41,43,44), we showed that in some cases, co-evolution of genes is better captured when focusing on clade-wise phylogenetic profiles. Equivalently, several studies found that the co-evolution of single RNA and DNA positions can be used to predict the function and structure of RNAs and RNA-protein complexes (45–47), and to predict gene annotations, CpG islands, repeat families etc. (48). Another study developed M-CAP (25), a method that predicts the pathogenicity of rare missense variants by integrating 16 preexisting prediction scores and annotations with the patterns of amino acid conservation across 100 species. These studies suggest that phylogenetic profiling at the nucleotide level (i.e. studying the patterns of conservation of single nucleotides across species) may reveal associations between co-evolving nucleotides and the clinical implications of single-nucleotide variants (SNVs). These associations may increase the information extracted from conservation data, improve classification of VUSs and highlight the importance of studying the co-evolution of single nucleotides.

Here, we studied the distribution of signals present in cross-species evolution at the nucleotide level in order to increase the information extracted from conservation data for the sake of predicting variants' outcomes. Our aims were to explore the variation in the ability to predict SNVs pathogenicity in multiple genes using conservation among individual vertebrates, and then to apply our findings for optimizing variant prediction per-gene and per-species. We first focused on studying signals present in genes related to HBOC, which constitute the largest annotated SNVs dataset in ClinVar (49), and then widened our analysis to 115 genes related to various human diseases. To analyze the evolution of nucleotides among these genes, to search for locally co-evolved positions, and to study how species diverge in their utility to predict pathogenic and benign variants among genes, we used the multiple sequence alignment of 100 vertebrates (50,51). To utilize our insights, we developed EvoDiagnostics, a machine-learning based variant prediction model that studies gene-specific conservation patterns of nucleotides (i.e. their nucleotide-level phylogenetic profiles) and optimizes the weights of each species accordingly. We found that EvoDiagnostics outperforms conservation methods in predicting the clinical significance of SNVs. EvoDiagnostics can be used as a stand-alone prediction tool or as a supplement for prediction tools that rely on conservation.

MATERIALS AND METHODS

Downloading and filtering ClinVar data

We downloaded two versions of ClinVar dataset on two different dates: 19 December 2017 and 13 May 2019, (https://www.ncbi.nlm.nih.gov/clinvar/docs/ftp_primer/#variant_summary) (52) (see Data availability). For both

versions, we kept only pathogenic and benign SNVs of assembly version GRCh37 (i.e. we filtered out the likely benign and likely pathogenic variants). We then focused on genes that had at least 50 pathogenic or benign SNVs reported in the 2017 version. This resulted in a total of 13 336 and 17 033 SNVs in the 2017 and 2019 versions, which belonged to a total of 115 genes. We then filtered the SNVs data as follows: we removed SNVs positioned outside of the coordinates of the genes as specified in Ensembl (downloaded using biomaRt (53)). We removed four additional samples that appeared in both ClinVar 2017 and 2019 datasets, two of which were deletion variants falsely assigned as SNVs and two of which were assigned with 'na' instead of a reported human variant. This resulted in a total of 13 250 and 16 944 SNVs in 115 genes, curated by the 2017 and 2019 ClinVar databases, respectively. Every gene in the 2017 and 2019 database had over 50 SNVs, apart from GRIN2A, which had 41 and 47 SNVs, respectively. A total of 1343 SNVs were removed from the 2019 update as compared to the 2017 version, while a total of 5037 SNVs were added. We refer to the added SNVs as the third ClinVar dataset, of 2017–2019 (Supplementary Table S1).

Mapping ClinVar's SNVs to the Multiz alignment of 100 vertebrates

To initially map BRCA1, BRCA2 and PALB2 SNVs reported in ClinVar, we converted the SNVs into BED format using rtracklayer (54), GenomicRanges (55) and Iranges (55) R packages. We then uploaded the files to the GALAXY tool, Extract MAF blocks (56), to map the variants to the 100-way Multiz alignment database (51), which was available through the Locally Cashed Alignment MAF source option. We chose not to split blocks by species.

To map SNVs in the entire pool of 115 genes reported in ClinVar, we used the UCSC hgTables webtool (50,51,57) (see Data availability). We selected the following options—clade: Mammals; genome: Human; assembly: Feb. 2009 (GRCh37/hg19); group: Comparative Genomics; track: Conservation; table: Multiz Align (multiz100way); region: Defined regions (where we inserted the coordinates of all SNVs); output format: MAF—multiple alignment format.

We ensured that the species were ordered by their evolutionary distance from human, according to the phylogenetic tree that was used to prepare the 100-way Multiz alignment database (see Data availability, below, for link).

Mapping ClinVar's SNVs to the Multiz alignment of 46 vertebrates

To map SNVs in the entire pool of 115 genes reported in ClinVar, we used the UCSC hgTables webtool (50,51,57). We selected the following options—clade: Mammals; genome: Human; assembly: Feb. 2009 (GRCh37/hg19); group: Comparative Genomics; track: Cons 46-Way; table: Multiz Align (multiz46way); region: Defined regions (where we inserted the coordinates of all SNVs); output format: MAF— multiple alignment format.

Converting the mapped alignments into conservation patterns data

We binarized the species alignment: in a pairwise comparison of the human sequence and another species, a species could either present a residue that is identical to the human wild-type, a residue that is identical to the human variant, or a residue that differs from both. We replaced the residue of a species with 0 if the analyzed species shared the same nucleotide with the human variant, and with 1 otherwise. In cases where the compared species had a gap or a deletion instead of a single nucleotide, we replaced the residue of the species with 0. In cases where the compared species had an insertion, we ignored all inserted nucleotides apart from the first one, and replaced it with 0 or 1, according to the aforementioned conditions (Supplementary Figure S1A). Our binarization approach ensured that within a single position, the three possible SNVs might result with a different conservation pattern among species (Supplementary Figure S2). We refer in the text to residues replaced with 1 as conserved, and residues replaced with 0 as non-conserved.

Plotting functions

We used ComplexHeatmap package in R to construct the heatmaps (58). We used the ggplot2 package (59) to construct all scatter plots, box plots, bar plots and density plots. We used the pROC package (60) for plotting ROC curves and calculating AUCs. We used the corrplot package (61) to plot the Pearson correlation between the accuracy profiles of pairs of genes. We used the ape package (62) to read the 100-way Multiz alignment phylogenetic tree and plot the chronogram of selected species.

Calculating species predictive properties across genes

Using the conservation patterns of SNVs reported in ClinVar until 2017, we considered conserved and non-conserved nucleotides in each species as indicators for pathogenicity and benignity of SNVs in each of the genes, respectively. We calculated the positive predictive value (PPV), negative predictive value (NPV) and accuracy in predicting SNVs by each of the 99 vertebrates, according to the described confusion matrix (Supplementary Table S2).

To find the species that were the best and worst predictors for the clinical implications of SNVs in the 115 genes, we calculated the mean accuracy scored by each species in predicting the clinical implications of SNVs in these genes. We compared these values to 1e6 mean accuracy values that were calculated using random bootstraps. We conducted two-sided *P*-value tests, and calculated the corresponding *q*-values using the Benjamini–Hochberg correction (63).

Training EvoDiagnostics models and estimating their performance by cross-validation

We trained three separate random-forest algorithms on the conservation patterns of either BRCA1, BRCA2 or the 115-genes (ClinVar 2017 version), using the caret package (64). Changing the value of the ntree parameter barely affected the estimated accuracy of the BRCA1, BRCA2 and the 115-genes models (Supplementary Figure S3) while substantially affecting their training time. Hence, in each model we

set ntree to 500, and chose the best mtry value (out of 11 values) through 10-repeated 5-fold hyperparameter tuning. We then performed 100 repeats of 5-fold cross-validation for the BRCA1, the BRCA2 and the 115-genes EvoDiagnostics models, using caret's trainControl function. Each repeat of cross-validation randomly divided the conservation patterns of variants into five folds, and the variants of each fold were predicted by a random forest model trained on the variants of the remaining four folds. Thus, each repeat estimated the prediction scores of EvoDiagnostics for every variant in the training sets. We then calculated the mean EvoDiagnostics score for each of the variants across the 100 repeats, computed a ROC curve and compared it to the prediction of three conservation-based methods (Figure 1A,C,E). We performed paired *t*-test comparisons between BRCA1- and 115-genes EvoDiagnostics and the three conservation methods by comparing the AUCs scored when predicting the SNVs in each of the cross-validation folds separately (five folds over 100 repeats, i.e. a total of 500 folds). We also produced a ROC curve for each repeat of the cross-validation of BRCA1-EvoDiagnostics model and compared them with the ROC curve of the repeat that scored the median AUC (Supplementary Figure S4).

Estimating EvoDiagnostics performance by predicting unseen test sets

We assembled test sets for the three EvoDiagnostics models using SNVs that were reported by ClinVar between December 2017 and May 2019. For each EvoDiagnostics model, the test set included prospectively reported variants in the genes that the model was trained on. The performance of BRCA1-EvoDiagnostics was also evaluated by predicting three additional test sets: SNVs in PALB2 that were reported until December 2017 (#1) and until May 2019 (#2), and SNVs in BRCA2 that were reported between 2017 and 2019 (#3). Since all SNVs in PALB2 that were reported between 2017 and 2019 were pathogenic, we were unable to compute the ROC and AUC of these SNVs.

Downloading prediction scores of conservation-based methods

We downloaded the prediction scores of the conservation-based tools: GERP++ (13), PhastCons (100 ways, 46 ways—vertebrates, 46 ways—primates, 46 ways—placental) (14) and phyloP (100 ways, 46 ways—vertebrates, 46 ways—primates, 46 ways—placental) (15), for every SNV predicted by EvoDiagnostics, from the UCSC hgTables webtool (50,57).

Training a random forest model on nine conservation scores

Using the caret package, we trained a random forest algorithm using the nine aforementioned conservation methods as features. We first performed an hyperparameter tuning analysis in which we evaluated five ntree values and eight mtry values (there are only eight possible mtry values when training on nine features). Once again, we saw that the ntree value had minimal effect on the estimated AUC of the model and thus set it to 500. We then trained the final model and used it to predict the prospectively reported SNVs in the 115 genes.

Comparing EvoDiagnostics with ensemble and deep learning-based methods

We downloaded the prediction scores of CADD and REVEL from HgTables webtool (50,57) (by 29 December 2021), of M-CAP from the link: http://bejerano.stanford.edu/mcap/ (by 15 December 2021), of EVE from the link: https://evemodel.org/download/bulk (by 29 December 2021), of Eigen from the link: http://www.funlda.com/toolkit (by 5 January 2022), and of PrimateAI from the link: https://basespace.illumina.com/projects/51955905 (by 19 December 2021). To note, M-CAP only predicts rare missense variants with allele frequency <1%, and thus its developers instructed to assume that missense variants that have no M-CAP prediction score are likely benign (see http://bejerano.stanford.edu/mcap/). Hence, to increase the number of the test set variants predicted by M-CAP from 1010 to 1325, we set the prediction score of missense variants lacking M-CAP score as 0. This is the M-CAP version used in our paper. The AUCs in predicting the test set variants using the published and the imputed M-CAP versions were 0.939 (1010 SNVs in total) and 0.905 (1325 SNVs in total), respectively.

AUCs in predicting SNVs using EvoDiagnostics models with changing train sizes

We predicted the pathogenicity of newly reported SNVs in BRCA1 and BRCA2 using 1000 separate BRCA1- and BRCA2-EvoDiagnostics models, respectively. Each of the 1000 BRCA1 and BRCA2 models was trained on SNVs in BRCA1 or BRCA2 that were randomly sampled (with replacement) from the entire pool of SNVs reported by ClinVar until 2017. To maintain balanced train sets, the sampling process was coerced to maintain percentages of pathogenic SNVs similar to the ones found in the full BRCA1 and BRCA2 train sets (i.e. 53.0% and 52.9%, respectively). The minimal and maximal percentages of pathogenic SNVs in all the sampled train sets were 52.3% and 54.0% in BRCA1, and 52.0% and 53.8% in BRCA2. The models were trained on different train set sizes, ranging from 50 to 1066 (in BRCA1) or to 1207 (in BRCA2).

RESULTS

SNVs may associate with various diseases, especially when located in conserved positions (13–15). However, inferring the clinical relevance of SNVs from conservation is sometimes complicated. For example, in BRCA1 gene (Figure 2), some positions are extremely conserved in mammals and birds (blue arrow), other are only conserved in chimpanzee (purple arrow), and some are conserved in birds and fish while being poorly conserved in mammals (red arrow) (65). Overall, nucleotides in BRCA1 gene show various patterns of conservation across species (i.e. various nucleotide-level phylogenetic profiles) that cannot be explained solely by the evolutionary distance of the species from human.

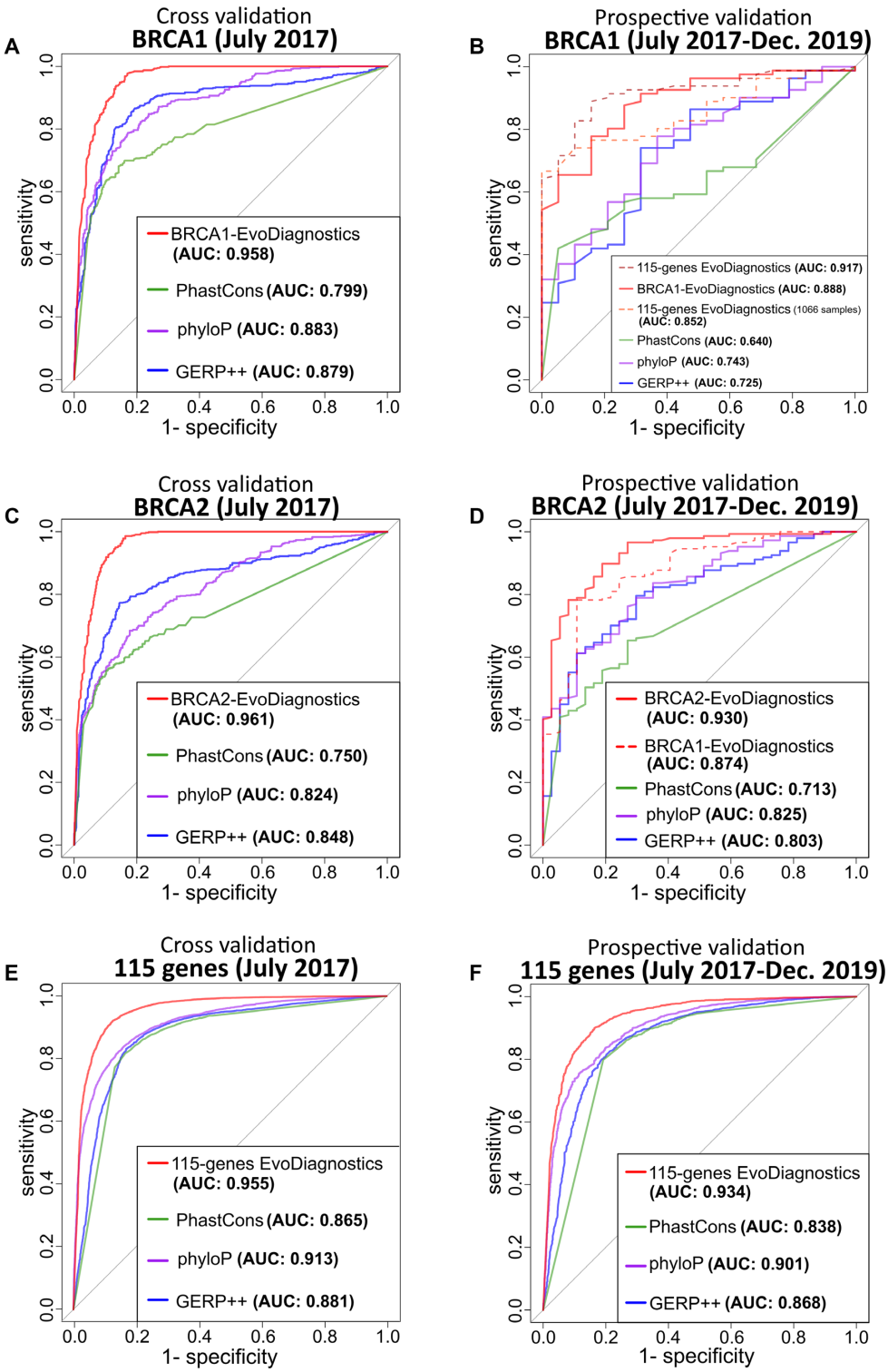

Figure 1. Comparing EvoDiagnostics predictions with conservation methods. ROC curves and AUCs of EvoDiagnostics models (BRCA1-EvoDiagnostics; BRCA2- EvoDiagnostics; 115-genes EvoDiagnostics) were compared to the predictions of the conservation methods: GERP++, phyloP and PhastCons. (**A**) The estimated prediction performances of BRCA1-EvoDiagnostics model, calculated by 100 repeats of 5-fold cross-validation of 1066 SNVs in BRCA1 that were reported by December 2017. (**B**) Temporal validation of the BRCA1-EvoDiagnostics model, calculated by predicting 100 SNVs in BRCA1 reported between 2017 and 2019. (**C**) The estimated prediction performances of BRCA2-EvoDiagnostics model, calculated by 100 repeats of 5-fold cross-validation of 1207 SNVs in BRCA2 that were reported by December 2017. (**D**) Temporal validation of the BRCA2-EvoDiagnostics model, calculated by predicting 184 SNVs in BRCA2 reported between 2017 and 2019. (**E**) The estimated prediction performances of 115-genes EvoDiagnostics model, calculated by 100 repeats of 5-fold cross-validation of 13 250 SNVs in the 115 genes that were reported by December 2017. (**F**) Temporal validation of the 115-genes EvoDiagnostics model, calculated by predicting 5037 SNVs in the 115 genes reported between 2017 and 2019; AUC, area under the curve; ROC, receiver operating characteristic; SNV, single-nucleotide variants.

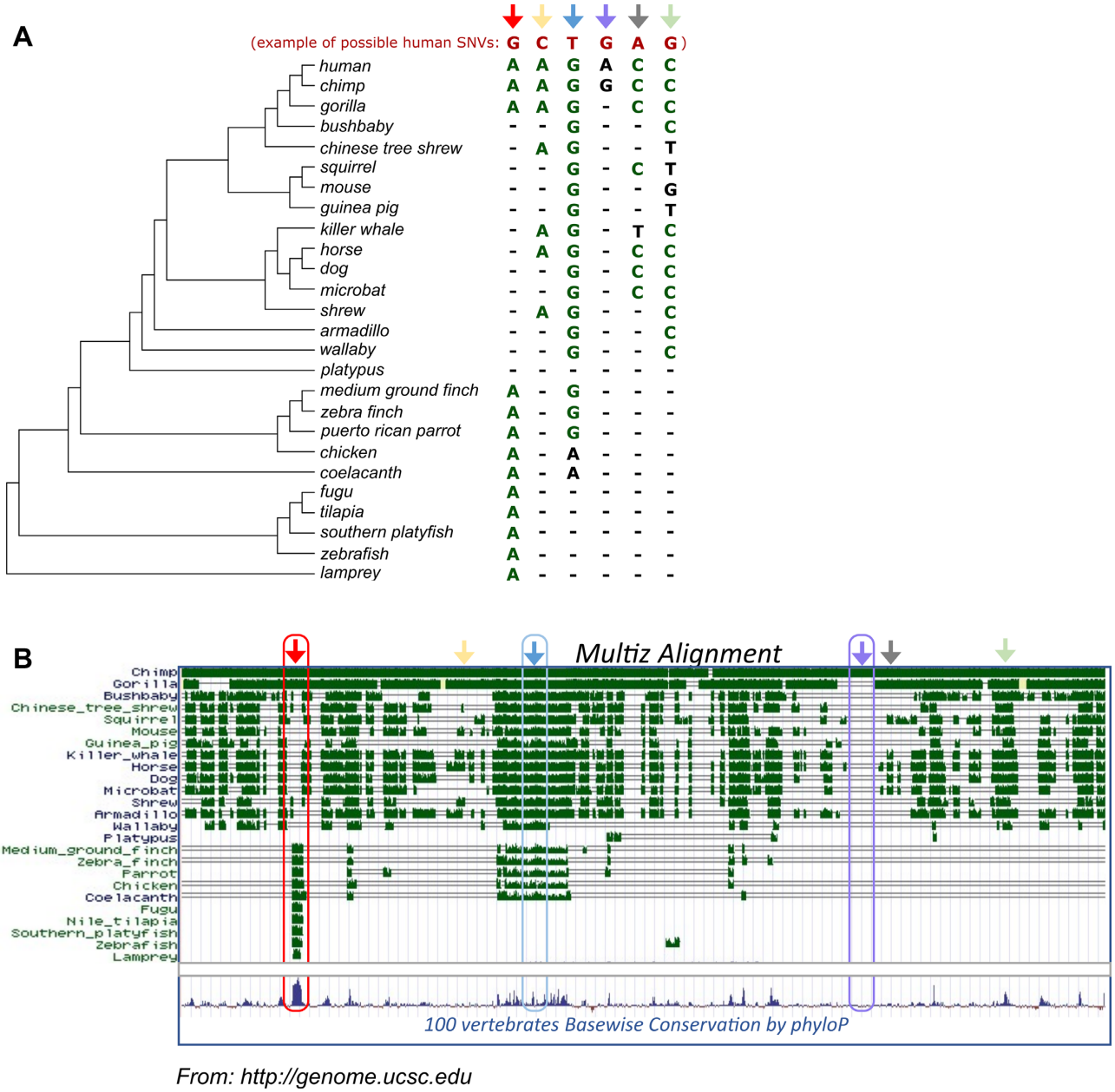

Figure 2. Conservation patterns of nucleotides (nucleotide-level phylogenetic profiles) in BRCA1. (**A**) The phylogenetic tree of selected species from the Multiz Alignment of 100 Vertebrates (50,51). Next to each species are the nucleotides located in 6 coordinates in BRCA1 that presented different conservation patterns across species. Conserved nucleotides are marked in green. Gaps are represented by hyphens. The nucleotides in red represent examples of possible SNVs in these positions. The specific coordinates are specified in Supplementary Data S4. (**B**) A capture from UCSC genome browser (65) of the conservation of BRCA1 in the coordinates chr17:41,225,000–41,277,500 (GRCh37 assembly) across selected species of the Multiz Alignment of 100 Vertebrates. Conserved positions are marked in green. A graphical representation of the conservation of each position across these 100 vertebrates (by phyloP) (15) is presented below. The colored arrows point on the conservation patterns presented in (A).

In the following sections, we studied the conservation patterns across species at the nucleotide level (i.e. the phylogenetic profiles of single nucleotides), and the species-to-species variation in their capability to predict clinical outcomes of SNVs using pairwise conservation comparisons. We then developed machine-learning classifiers that predicted which SNVs were of clinical importance, based on all pairwise comparisons, i.e. a comparative-genomics approach.

Living fossil fish are the best predictors for pathogenicity, while mammals are the best predictors for benignity in BRCA1

Since we suspected that species differ in their utility to predict pathogenicity or benignity of variants (66), we examined how well the clinical implications of SNVs in BRCA1 could be predicted by the conservation among individual organisms. We chose to analyze BRCA1 since the BRCA genes had, by far, the largest number of benign

and pathogenic SNVs reported on ClinVar (version of December 2017) (52) (Supplementary Table S1). We analyzed a total of 1066 pathogenic and benign SNVs in BRCA1. For each variant, we downloaded the multiple sequence alignment of 100 vertebrates and determined whether these positions were conserved in each of the species (see Materials and Methods). We first relied on a naïve classifier that predicts an SNV as pathogenic if the position is conserved and non-pathogenic if otherwise. We defined positive predictions as SNVs that were pathogenic, and thus true positive predictions are pathogenic SNVs that were predicted correctly. We then calculated three performance measurements for each species: (i) Positive predictive value (PPV)—whether the conservation of a nucleotide was a good predictor for pathogenicity, (ii) Negative predictive value (NPV)— whether the non-conservation of a nucleotide was a good predictor for benignity and (iii) Accuracy—The fraction of cases in which the conservation status matched the phenotype (see Methods).

Our analysis revealed that primates such as chimpanzee and orangutan are the best predictors for benignity of variants in BRCA1 (highest NPVs) (Figure 3A). In other words, human variants located in positions that are not conserved among primates are most likely benign. However, comparisons with primate sequences are not highly informative regarding the pathogenicity of variants when the position is conserved among primates. In contrast, a comparison with birds is useful for predicting pathogenic variants but less so for predicting benign ones. The turtles and the alligator resembled birds in their ability to predict pathogenicity but were more successful than birds in predicting benignity. Interestingly, while some fish were incapable of predicting pathogenicity of BRCA1 SNVs, other fish, such as the coelacanth and the spotted gar fish—both considered as living fossils (67,68)—were the best predictors for pathogenicity out of all vertebrates. Overall, in BRCA1 the prediction accuracy of conservation varied among species in a way that did not always associate with their evolutionary distance from human (Figure 3B). Since genes differ in their conservation across the tree of life (some genes are highly conserved across eukaryotes, while others show a high level of divergence even within mammals), we examined whether the species that were found to be good predictors in BRCA1 would also be good predictors in other genes, that is, whether the ability to predict variants using conservation is similar across different genes.

Prediction accuracy varies across genes and species

Similar to our analysis of BRCA1, we calculated the PPV (Supplementary Figure S5A), NPV (Supplementary Figure S5B) and accuracy (Figure 4) in predicting the pathogenicity of SNVs in disease-associated genes, using pairwise comparisons against each of the 99 vertebrate species. To eliminate the noise that may result from using limited variant data, we included in our analysis only the genes that had at least 50 benign and pathogenic SNVs reported in ClinVar. These summed up to a total of 115 genes. In 63%, 83% and 64% of the genes, the median PPV, NPV and accuracy, respectively, were >0.7. That is, in most of the genes, the pathogenicity of SNVs was predicted well by pairwise comparisons against most of the species. However, in a small portion of the genes SNVs were poorly predicted, including four genes (CACNA1C, DNAH11, MECP2, USH2A) that their SNVs were predicted with low accuracy by all pairwise comparisons (maximum prediction accuracy of 0.698, 0.667, 0.685, 0.650, and a median accuracy of 0.476, 0.425, 0.523 and 0.506, respectively, over all pairwise comparisons, see Supplementary Data S1).

Analysis of the species showed that their median PPV, NPV and accuracy in predicting variant pathogenicity across the genes ranged from 0.72 to 0.89, from 0.55 to 1.00 and from 0.67 to 0.84, respectively, indicating that species diverged in their prediction properties (Figure 4 and Supplementary Figure S5, lower panels, Supplementary Table S3). Furthermore, each species presented wide ranges of PPV, NPV and accuracy in predicting the SNVs, meaning that each species yielded accurate predictions in some genes, but inaccurate predictions in others. For example, the golden hamster accurately predicted the pathogenicity of SNVs in APC, VHL, CNPA3 and LDLR genes but inaccurately predicted the pathogenicity of SNVs in MSH6, MYH7, KCNQ3 and TNXB genes. Nevertheless, all species yielded at least 0.9 accuracy in predicting at least seven genes. Overall, the green sea turtle and the painted turtle were among the significantly best predictors (mean accuracy: 0.802 and 0.801, respectively, *P*-value: 1.6e-5 and 3e-5, respectively), while the lamprey and the chimp were among the worst (mean accuracy: 0.630 and 0.659, respectively, *P*-value: $>$1e-06 and 4.8e-5, respectively) (Supplementary Figure S6 and Supplementary Data S2).

By ordering species by their mean accuracy in predicting SNVs' pathogenicity and coloring them by the percentage of positions in which their residues were conserved, we pinpointed cases in which the predictiveness was not associated with the evolutionary distance of that species from human (Supplementary Figure S7). Nonetheless, distant vertebrates tended to have higher mean accuracy.

Clustering the genes by their accuracy, PPV and NPV profiles, revealed groups of genes that resemble each other in the capability to use species conservation to predict SNVs (Figure 4 and Supplementary Figure S5).

To better characterize the similarity and differences between the ability to predict SNVs in different genes, we calculated the Pearson correlation coefficient between the accuracy profiles of each pair of genes. Some genes such as DNAH5, FBN2 and TTN, were highly and positively correlated in their species accuracy profiles (all pairwise correlation had Pearson R^2 above 0.804 and *P*-value $<$4.17e-36). Other genes such as PALB2 and ADGRV1 had significant negative correlations (Pearson *R*: -0.813, Pearson R^2: 0.711, *P*-value: 1.59e-24) (Supplementary Figure S8). Clustering analysis grouped the 115 genes into three main clusters (Supplementary Figure S9). The second cluster contained most of the DNA repair genes (i.e. BRCA1, BRCA2, PALB2, MLH1, ATM, MSH2, TP53, LMNA, APC, NF1), which are related to LYNCH, Li-Fraumeni, and hereditary breast and ovarian cancer syndromes. Overall, 8%, 31% and 3% of the genes in the first, second and third cluster, respectively, were DNA repair genes.

These analyses show that the predictive power of conservation differs among species and across genes. Moreover,

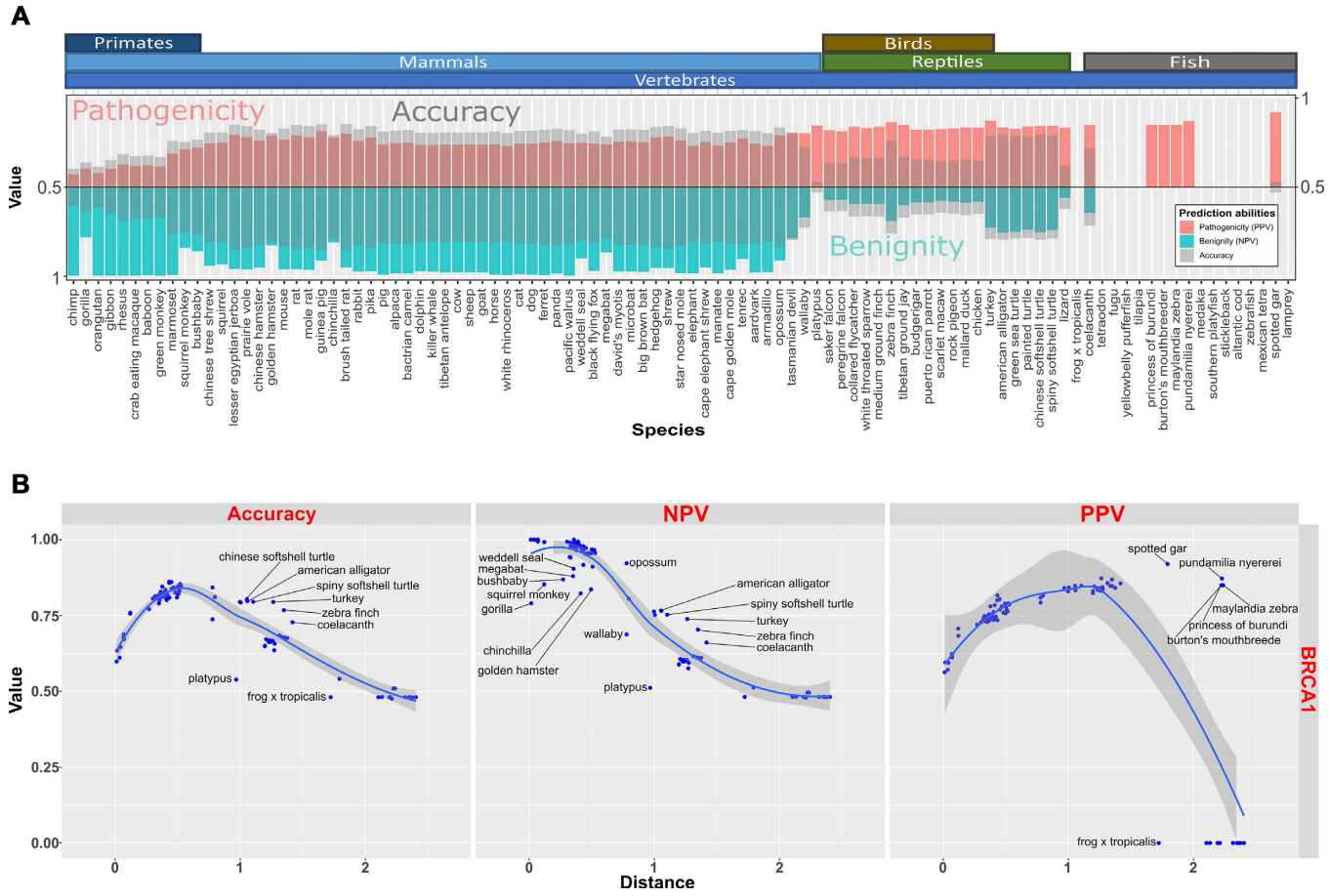

Figure 3. Species properties in predicting SNVs in BRCA1. The PPV ((**A**) red—above the *x*-axis (**B**) third panel), NPV ((**A**) blue—beneath the *x*-axis, (**B**): second panel) and accuracy ((**A**) gray—a mirror image from both sides of the *x*-axis, (**B**): first panel) in predicting pathogenicity of 1066 variants in BRCA1 using the conservation inferred over pairwise comparisons of 99 vertebrate species with human. In (A), clades are ordered from left to right by their estimated evolutionary distance from the human. Both sides of the *y*-axis range from 0.5 to 1, since values of ∼0.5 indicate no predictive power over random predictions. In (B), species are ordered by their estimated distance from human according to the phylogenetic tree used in the 100-ways Multiz alignment. In each panel, the blue line and the gray area represent the loess regression that best fits the scatter plot and its 0.95 confidence interval, respectively. Names of species that deviated from the common trend were added; NPV, negative predictive value; PPV, positive predictive value; SNV, single-nucleotide variants.

genes clustered into groups with similar profiles of species accuracy, PPV and NPV, suggesting that for certain genes some species are better predictors for the clinical outcomes of SNVs than others. After gaining insights regarding the prediction abilities of individual species, we aimed to evaluate the association between SNVs pathogenicity and the cross-species conservation patterns at the single-nucleotide level.

Complex conservation patterns within genes

SNVs in BRCA1 showed a variety of non-random and complex conservation patterns (Figure 2). We studied the association between the pathogenicity of SNVs in BRCA1 and their conservation patterns, i.e. their nucleotide-level phylogenetic profiles (Supplementary Figure S1A,B; see Materials and Methods). Our aim was to move from inference based on pairwise comparisons to a combined analysis of all species together. Analysis of the nucleotide profiles in BRCA1 showed that several exonic and intronic positions, not necessarily in contiguity, had similar profiles (Supplementary Figure S1B). Furthermore, SNVs in BRCA1 clustered by their conservation patterns into 11 sub-clusters that had >10 SNVs each (Figure 5A). We calculated the percentage of pathogenic SNVs in each of the 11 sub-clusters. In five of the sub-clusters, over 78% of the variants were pathogenic, while in five other sub-clusters <9% of the SNVs were pathogenic. That is, pathogenic variants resembled each other in their conservation patterns. Furthermore, in BRCA1 the conservation patterns captured additional information compared to naïve conservation. For example, the conservation patterns of the yellow cluster (123 SNVs, marked with four stars in Figure 5A) resembled the patterns of the red cluster (144 SNVs, marked with one star) up to the wallaby, but lacked conservation along the turtles, lizards and some birds. Despite having similar overall conservation, the yellow cluster was much less pathogenic than the red cluster (64.2% and 81.9%, respectively). In addition, the orange cluster (227 SNVs, marked with two stars) that was more conserved than the red cluster, showed similar pathogenicity to the red cluster (85.5% and 81.9%, respectively), while the turquoise cluster (28 SNVs, marked with

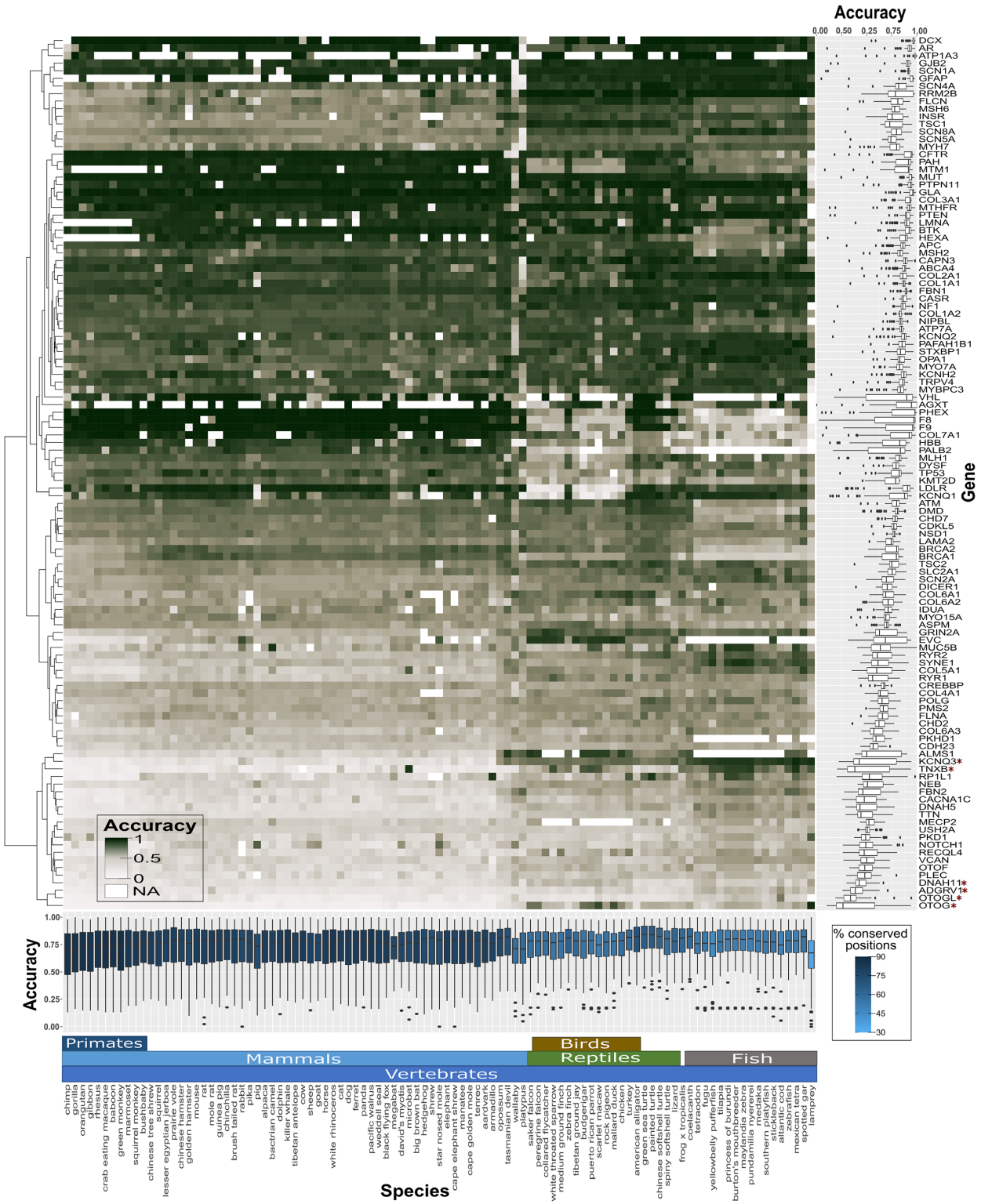

Figure 4. Species accuracies in predicting SNVs in 115 disease-associated genes. The heatmap presents the accuracies of 99 species (*x*-axis) in predicting SNVs in each of the 115 diseases-associated genes (*y*-axis). For each species, the accuracy in predicting variants in each gene was calculated as the fraction of pathogenic and benign SNVs that were conserved and non-conserved among the species, respectively. Dark and bright colors represent high and low accuracies, respectively. We colored white (NA) the cases in which a species was entirely conserved or non-conserved across all SNVs in a gene, and in the cases in which all SNVs of a gene were either benign or pathogenic. The right panel presents the distribution of the accuracies in predicting SNVs in each of the genes. Red asterisks mark the six genes that had the lowest median accuracy (all had medians below 0.426). The lower panel shows the distribution of the accuracies scored by each of the species, and the colors represent the percentage of conserved positions in each species across the entire pool of variants. The genes were clustered by their accuracy profiles across the species. Species were ordered by their estimated evolutionary distance from the human; SNV, single-nucleotide variants.

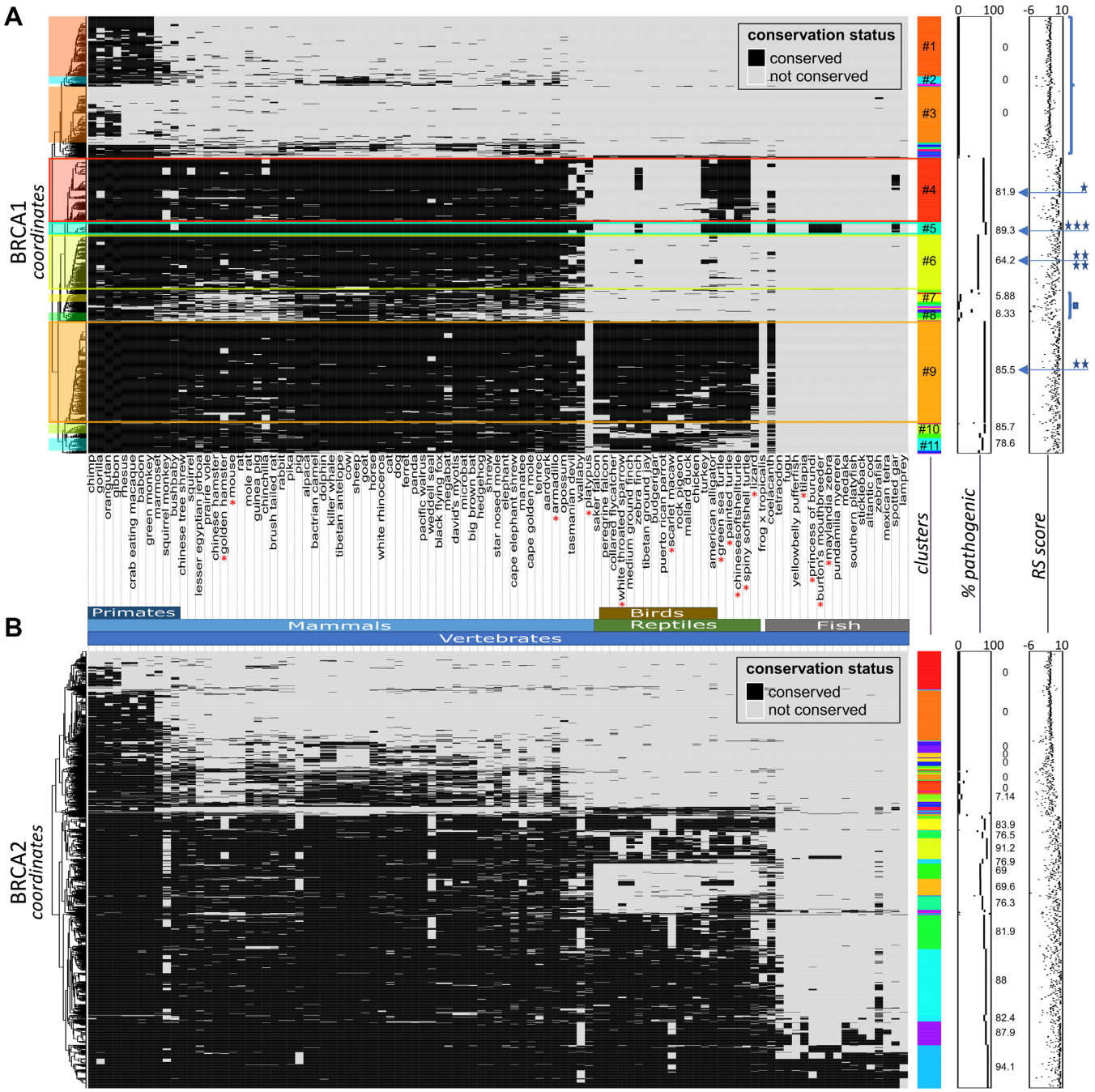

Figure 5. Conservation patterns (nucleotide-level phylogenetic profiles) of SNVs in BRCA along 99 vertebrates. The heatmaps show the conservation patterns of benign and pathogenic SNVs in BRCA1 (**A**, 1066 SNVs) and BRCA2 (**B**, 1207 SNVs) along 99 vertebrates. SNVs data were downloaded from ClinVar. Conserved and non-conserved alleles in each species were colored black and grey, respectively. The *x*-axis represents the 99 aligned vertebrates, ordered from left to right by their increasing evolutionary distance from human. Annotations on the right, from left to right: color annotation for each cluster (37 clusters in (A) and 59 in (B)). In (A), the 11 clusters that had at least 10 SNVs were numbered; the percentage of pathogenic SNVs in each cluster (exact percentages were printed for clusters containing >10 SNVs); GERP++ prediction scores ('RS score') for each SNV (13), ranging from -10 to + 6. The blue arrows, stars and braces, and the red asterisk in (A), point to specific clusters and species discussed in the results section. In (A), the portion in the dendrogram that corresponded with the 11 clusters was colored. The colored rectangles in (A) mark the clusters discussed in the results section.

three stars) that had similar conservation patterns to the red cluster, apart from some fish, had the highest pathogenicity (89.3%).

We then tested whether these results could be generalized to additional genes. Repeating the above analyses for BRCA2, another DNA repair gene (69), led to similar results (Figure 5B, Supplementary Figures S10 and 1C): SNVs in BRCA2 presented a variety of non-random complex conservation patterns, which were associated with the pathogenicity of these SNVs.

These results indicate that conservation patterns add information over naïve conservation, and that pathogenic and benign variants differ in their conservation patterns. This suggests that these patterns can be used to detect pathogenic SNVs.

Machine-learning algorithm predicts SNVs pathogenicity from conservation patterns

We developed the EvoDiagnostics algorithm to classify SNVs as either pathogenic or benign. EvoDiagnostics is a machine-learning classifier that utilizes the information in nucleotide-level phylogenetic profiles both for learning and for testing. In the learning phase, the algorithm learns the associations between the conservation patterns of variants and their pathogenicity. Each feature of the classifier corresponds to a pairwise comparison between a specific vertebrate species and the human variant. The value of each feature is 0 if the analyzed species share the same nucleotide with the human variant, and is 1 otherwise (see Materials and Methods for more details). EvoDiagnostics is based on the random forest classification algorithm, which outperformed alternative tested classification algorithms such as SVM with various kernels, KNN, LDA, logistic regression and naïve Bayes (Supplementary Data S3, Supplementary Table S4).

We trained EvoDiagnostics on the conservation patterns of the 1066 pathogenic and benign BRCA1 SNVs (reported in Clinvar by December 2017, see Materials and Methods). We estimated EvoDiagnostics prediction properties using 100-repeated 5-fold cross-validation. The prediction values calculated via cross-validation separated pathogenic variants from benign, with an AUC of 0.959±0.012 (Figure 1A, Supplementary Figures S4 and S11A–H). On the same data, three conservation-based methods: GERP++ (13), phyloP (15) and PhastCons (14) scored significantly lower AUCs of 0.879±0.024, 0.883±0.021 and 0.799±0.028, respectively (all paired *t*-test *P*-values < 5e-290). We then tested the accuracy of EvoDiagnostics in predicting an independent test set of 100 prospectively reported variants in BRCA1 (reported between December 2017 and May 2019). On this test set, EvoDiagnostics scored an AUC of 0.888, while GERP++, phyloP and PhastCons scored AUCs of 0.725, 0.743 and 0.640, respectively (Figure 1B).

Similar results were obtained upon training our model on the 1207 pathogenic and benign BRCA2 SNVs reported in Clinvar, scoring AUCs of 0.961±0.011 in cross-validation and 0.930 when predicting 184 prospectively reported SNVs, i.e. on an independent test data (Figure 1C,D and Supplementary Figure S11I–P). This suggested that our prediction strategy could be relevant for additional genes. Our next step was to study the relationship between the amount of data available for each studied gene and the performance of the machine-learning classifier.

EvoDiagnostics accurately predicts variants in genes with limited data

Optimizing predictions per gene is challenging when SNVs data are scarce (< ∼300 SNVs). This is highlighted by the association between the AUC score and the number of SNVs used for training BRCA1 and BRCA2 EvoDiagnostics models (Supplementary Figure S12). We tested the hypothesis that learning from one gene (with ample data) can be informative for predicting the pathogenicity of SNVs in other genes (with limited data). To this end, we tested the ability of BRCA1-EvoDiagnostics, trained on BRCA1, to predict prospectively reported SNVs in BRCA2 and PALB2 (see Materials and Methods). We focused on the associations between BRCA1, BRCA2 and PALB2 since the three genes co-evolved in animals (41), they all participate in the HRR-pathway (i.e. share a biological function) (70–72) and positively correlate in their species-accuracy profiles (Figure 4 and Supplementary Figure S9). BRCA1-EvoDiagnostics predicted variants in PALB2 (reported until 2017 and reported until 2019) and in BRCA2 (reported between 2017 and 2019) with AUCs of 0.935, 0.929 and 0.874 respectively. GERP++, phyloP, and PhastCons scored AUCs of 0.800, 0.778 and 0.777, respectively, when predicting SNVs in PALB2 (until 2017), AUCs of 0.842, 0.829 and 0.835, respectively, when predicting SNVs in PALB2 (until 2019), and AUCs of 0.803, 0.825 and 0.713, respectively, when predicting SNVs in BRCA2 (Figure 1D; Supplementary Figure S13 and Supplementary Table S5). As expected, BRCA2-EvoDiagnostics outperformed BRCA1-EvoDiagnostics in predicting the unseen BRCA2 SNVs, scoring an AUC of 0.930 (Figure 1D and Supplementary Table S5).

Next, we examined the ability of an EvoDiagnostics model trained on the 115 genes (denoted 115-genes EvoDiagnostics) to predict variants in these genes. 115-genes EvoDiagnostics scored an AUC of 0.955±0.004 in cross-validation, while GERP++, phyloP and PhastCons scored AUCs of 0.881±0.006, 0.913±0.005 and 0.865±0.006, respectively (all paired *t*-test *P*-values < 3.29e-246) (Figure 1E). When predicting 5037 prospectively reported SNVs in the 115 genes, 115-genes EvoDiagnostics scored an AUC of 0.934, while GERP++, phyloP and PhastCons scored AUCs of 0.868, 0.901 and 0.838, respectively (Figure 1F and Supplementary Table S5). When focusing on BRCA1's 100 newly reported variants, the 115-genes EvoDiagnostics model scored an AUC of 0.917, outperforming the gene-specific BRCA1-EvoDiagnostics model, which scored an AUC of 0.888 (Figure 1B). However, BRCA1-EvoDiagnostics was trained on 1066 SNVs while 115-genes EvoDiagnostics was trained on ∼13× more data (13 250 SNVs). As expected, this shows that large data from other genes compensate for the lack of data in a specific gene. We next compared the performance of BRCA1-EvoDiagnostics model with a 115-genes EvoDiagnostics model trained on an equally sized train set (i.e. 1066 SNVs randomly sampled from the 115 genes' variants-pool, excluding SNVs in BRCA1). In this case, the 115-genes Evo-

Diagnostics (1066 samples) model underperformed in predicting the newly reported BRCA1 variants, scoring an AUC of 0.852 (Figure 1B and Supplementary Table S5). This shows that both the train-set size and the per-gene optimization are important for predictions.

It is important to compare the performance of the 115-genes EvoDiagnostics model with gene-specific models trained on each of the 115 genes, since such comparisons may provide a better understanding of the interplay between train size, per-gene optimization, and prediction accuracy. However, apart from BRCA1 and BRCA2 genes, none of the 115 genes had a balanced dataset (i.e. with similar pathogenic and benign SNVs counts) with at least 300 reported SNVs in ClinVar 2017 (Supplementary Table S6) hindering such an analysis.

Using a larger set of species from different clades improves variant prediction

We next examined the effect the number and divergence of species has on prediction accuracy. We trained an EvoDiagnostics model on the conservation patterns of SNVs in the 115 genes across 46 vertebrates (51) (denoted EvoDiagnostics46). We compared EvoDiagnostics46 with the 115-genes EvoDiagnostics100 model, GERP++, and four different versions of phyloP and PhastCons: calculated based on 100 vertebrates, 46 vertebrates, 46 placental species and 46 primates (14,15) (Supplementary Figure S14A and Supplementary Table S7A). EvoDiagnostics100 and EvoDiagnostics46 scored the highest AUCs of 0.934 and 0.918, respectively. PhastCons and phyloP scored AUCs of 0.838 and 0.901, respectively, while the three kinds of 46-species PhastCons and 46-species phyloP scored AUCs of 0.814 and less, and 0.891 and less, respectively. Across all methods, using 100 species from different clades improved the prediction accuracy, suggesting that it increased the information captured by conservation analysis. Furthermore, analyzing only primates, as done by a certain version of PhastCons and phyloP, scored the lowest AUCs (Supplementary Figure S14A and Supplementary Table S7A). This is congruent with our finding that primates are of the worst predicting species (Supplementary Data S2). In addition, we compared EvoDiagnostics to a random forest model trained on all the nine mentioned conservation scores as the model's features, to which we denoted RF conservation (see Materials and Methods, Supplementary Figure S14A, Supplementary Table S7A). EvoDiagnostics100, EvoDiagnostics46 and conservation RF scored AUCs of 0.934, 0.918 and 0.910, respectively.

Comparing EvoDiagnostics with ensemble and deep learning-based models

To this point, we compared EvoDiagnostics to long-known and well-established conservation methods that enhanced prediction accuracy thanks to the strong fidelity of traditional nucleotide-level conservation analyses (27). Next, we evaluated more recently developed methods, which were based on either using cutting-edge deep-learning algorithms on amino acid sequences, or on integrating numerous annotations additional to conservation. This included two deep learning based models that classify missense SNVs based on amino acid sequences of over 140K species (EVE (73)), and of six non-human primate species populations (PrimateAI (74)), and four different ensemble tools [CADD (23,75), M-CAP (25), REVEL (24) and Eigen (76)].

We downloaded each tool's most updated version of prediction scores of the prospectively reported SNVs in the 115 genes (i.e. the test set, see Materials and Methods). Some tools had not provided predictions for the entire test set (Supplementary Table S7B). For each method, we calculated the AUCs in predicting the missense, noncoding, nonsense and nonsynonymous SNVs in separate and overall (see Supplementary Table S7B, Figure 6). In predicting the entire test set, EvoDiagnostics outperformed all methods besides for CADD, an ensemble-based tool that uses three conservation-based features (e.g. phyloP, PhastCons and GERP) and 57 non-conservation based features. In predicting missense SNVs, EvoDiagnostics significantly outperformed all conservation-based methods. In predicting noncoding SNVs, EvoDiagnostics performed as conservation-based methods, which were significantly more accurate than all ensemble and deep-learning based methods, besides for CADD. We also compared the prediction of the test set SNVs that overlapped across all evaluated methods (Supplementary Figure S14B). Although the overlap spanned only ∼10% of the test set (537 out of 5,037 SNVs) and included only missense variants, EvoDiagnostics was not inferior to the performance of ensemble and deep-learning tools (EvoDiagnostics100 scored an AUC of 0.906 while M-CAP, PrimateAI, EVE ASM, EVE BPU, CADD, REVEL, Eigen.raw, Eigen.phred, Eigen.PC.raw and Eigen.PC.phred scored AUCs of 0.939, 0.891, 0.897, 0.901, 0.919, 0.936, 0.908, 0.908, 0.895 and 0.895, respectively).

DISCUSSION

It is currently a common practice to treat conservation as a one-rule-fits-all concept and analyze conservation globally across genes and species. In this functional genomics study, we suggested a new approach for analyzing conservation data, and demonstrated its benefits in variant prediction. We optimized conservation analysis per-species and per-gene and showed that the conservation patterns at the levels of nucleotides, genes and species, capture new insights overlooked by traditional conservation.

From the nucleotides' perspective, we showed that the complex conservation patterns of nucleotides are more informative than naïve conservation computations that assign a single score for each site. For example, nucleotides with identical conservation percentages diverged into different groups based on their conservation patterns across species. These groups demonstrated various associations with pathogenicity. To our knowledge, we are the first to use nucleotide-level phylogenetic profiles to directly estimate clinical outcomes of SNVs, and the first to characterize its substantial contribution to prediction accuracy compared to traditional conservation analysis. Our findings are supported by a recent work showing that the conservation of nucleotides can be characterized using 100 different states that associate with the functional importance of nucleotides (48). Furthermore, M-CAP tool uses 16 genomic

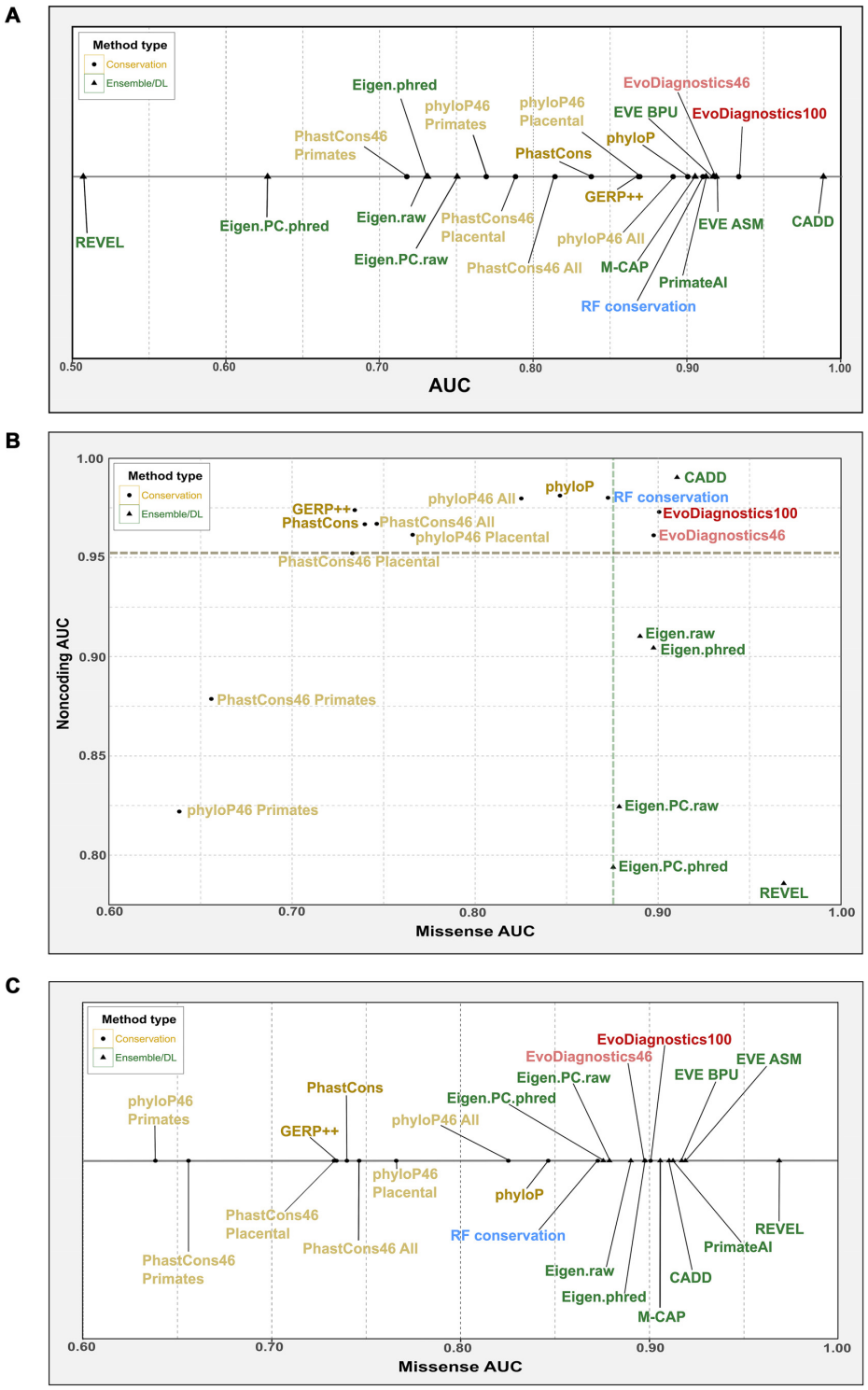

Figure 6. AUCs of EvoDiagnostics and other methods in predicting different types of SNVs. AUCs in predicting the test set SNVs (i.e. prospectively reported pathogenic and benign SNVs in the 115 genes) by different prediction methods. These AUCs were calculated using each method's most updated version of pre-calculated prediction scores of the test set SNVs (see Materials and Methods). Exact counts of SNVs predicted by each method is described in Supplementary Table S7. Ensemble and deep learning (DL) based models are marked with black triangles and green text. The remaining methods are marked with black circles and either blue, red or yellow text. RF conservation (blue) refers to the random forest model trained on the nine mentioned conservation scores (see Materials and Methods). EvoDisgnostics models (red) and conservation tools (yellow) that were based on 100 species alignments are marked with darker colored text compared to their equivalent 46 species-based versions. (**A**) One dimensional representation of the AUCs in predicting all types of variants in the test set. (**B**) AUCs in predicting the missense (*x*-axis) and noncoding (*y*-axis) test set SNVs. Methods that provided no predictions of noncoding SNVs are missing from this panel. The vertical and the horizontal dashed lines represent the minimal AUCs scored by all ensemble/DL tools in predicting missense variants, and by most conservation-based tools in predicting noncoding variants, respectively. (**C**) One dimensional representation of the AUCs in predicting the missense SNVs; AUC, area under the curve; DL, deep learning; SNV, single nucleotide variant.

annotations in combination with conservation patterns of rare amino acid variants across species (represented by three features per-species) to improve prediction of rare missense variants (25).

At the genes' perspective, genes vary in their evolutionary history and conservation. Different positions within each gene show a variety of complex conservation patterns, which capture gene-specific associations with the pathogenicity of variants. Furthermore, we showed that genes cluster by their accuracy, NPV and PPV profiles, demonstrating that some genes resemble in the capability to have their variants predicted using conservation, while others differ. For example, BRCA1 and BRCA2 share similar conservation patterns and their accuracy profiles are positively correlated, while their accuracy profiles are negatively correlated with the accuracy profiles of TTN and FBN2. Thus, tailored conservation analysis that considers the variation among genes is likely to be beneficial, yet most current prediction tools evaluate conservation identically across genes.

At the species perspective, our results clearly demonstrated that species differ in their ability to predict the clinical outcomes of variants. These species-to-species differences exist not only between distinct clades (e.g. mammals versus birds), but also within clades (e.g. coelacanth versus lamprey). Furthermore, Malhis *et al*. (66) showed that the NPV in predicting variants using species conservation decreases with evolutionary distance from human. Sundaram *et al*. (74) demonstrated that amino acid variants at high allele frequencies in chimpanzee populations indicate benign consequences in human. Our analysis not only supported these findings but also showed that the PPV increased when using evolutionary distant species, while the accuracy did not necessarily associate with evolutionary distances. Comparing between conservation-based prediction tools that analyzed 46 and 100 species, emphasized the effect the analyzed set of species has on prediction accuracy. Furthermore, we showed that the ability of species to predict the clinical outcomes of variants differed across genes. This means that conservation among a certain species could be informative in one gene but non-informative in another (e.g. golden hamster in APC versus in MSH6). Hence, it is likely beneficial to integrate the nucleotide conservation patterns across species using species-weights that are optimized for the prediction task.

To apply our findings and validate them, we developed EvoDiagnostics, a prediction model that classifies SNVs using random forest based on conservation patterns. EvoDiagnostics was optimized according to the utility of each species to predict variants in the genes of interest, i.e. the best combination of features was computed to maximize performance on the training data.

EvoDiagnostics outperformed conservation tools when prospectively predicting SNVs in 115 disease-related genes, especially when predicting missense variants. Furthermore, we showed that EvoDiagnostics optimized for BRCA1 outperformed the 115-genes EvoDiagnostics model (that was trained on an equally sized train set) in predicting SNVs in BRCA1. EvoDiagnostics also predicted variants in genes with limited data better than naïve conservation.

We expect to increase the accuracy of EvoDiagnostics with the accumulation of additional data from genes currently not included in our analysis. Furthermore, as data accumulate, we expect to optimize EvoDiagnostics per type of variants (e.g. missense, noncoding, nonsense and synonymous). This would be accomplished by characterizing associations between conservation patterns, variant type and variant pathogenicity, while providing insights regarding the ability of different species to predict different types of variants across the genes. Expanding the pool of analyzed species may also improve the performance of our model. One limitation of EvoDiagnostics is its current inability to predict the pathogenicity of insertions and deletions mutations. Other important future directions include testing additional coding schemes for converting alignments to features, and taking amino acid attributes into account. Ensemble prediction tools such as REVEL (24), M-CAP (25), CADD (23), Eigen (76) and LINSIGHT (77) use conservation within their set of features. Thus, implementing EvoDiagnostics as a feature in ensemble-based prediction methods would potentially increase their prediction accuracies.

Overall, we expect EvoDiagnostics to promote personalized-based medicine by improving VUSs predictions and patient management. We believe that a deeper understanding of the crosstalk between the nucleotide-level conservation patterns, the genes and the species, could shed light on the evolutionary processes the genes went through and promote applications of evolutionary concepts into medicine and biology.

DATA AVAILABILITY

The datasets generated during the current study and the code used to generate EvoDiagnostics models are available in: https://figshare.com/account/home#/projects/134078.

The datasets analyzed during the current study are available in the ClinVar repository (52) [https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/archive/2017/variant_summary_2017--12.txt.gz, https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/archive/2019/variant_summary_2019--05.txt.gz], in the 100-way MultiZ alignment repository [http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/hg19.100way.commonNames.nh], and in the UCSC HgTables webtool (57) [http://genome.ucsc.edu/cgi-bin/hgTables] (see Materials and Methods for the exact attributes used in this study for downloading data from HgTables).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dolev Rahat, Menachem Isseroff and Daniel Gelman for their insightful comments on the manuscript.

FUNDING

Israel Innovation Authority [0395084]; Israel Science Foundation [1591/19, 802/16]; The Alex U Soyka Pancreatic

Cancer Research Project (RIA, HH); Edmond J. Safra Center for Bioinformatics, Tel Aviv University (to N.W., in part); Ariane de Rothschild Woman Doctoral Program (to S.L.).
Conflict of interest statement. None declared.

REFERENCES

1. Rabbani,B., Nakaoka,H., Akhondzadeh,S., Tekin,M. and Mahdieh,N. (2016) Next generation sequencing: implications in personalized medicine and pharmacogenomics. *Mol. Biosyst.*, **12**, 1818–1830.
2. Richards,S., Aziz,N., Bale,S., Bick,D., Das,S., Gastier-Foster,J., Grody,W.W., Hegde,M., Lyon,E., Spector,E. *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. *Genet. Med.*, **17**, 405–424.
3. Vears,D.F., Sénécal,K. and Borry,P. (2017) Reporting practices for variants of uncertain significance from next generation sequencing technologies. *Eur. J. Med. Genet.*, **60**, 553–558.
4. Maxwell,K.N., Hart,S.N., Vijai,J., Schrader,K.A., Slavin,T.P., Thomas,T., Wubbenhorst,B., Ravichandran,V., Moore,R.M., Hu,C. *et al.* (2016) Evaluation of ACMG-guideline-based variant classification of cancer susceptibility and non-cancer-associated genes in families affected by breast cancer. *Am. J. Hum. Genet.*, **98**, 801–817.
5. Antoniou,A., Pharoah,P.D.P., Narod,S., Risch,H.A., Eyfjord,J.E., Hopper,J.L., Loman,N., Olsson,H., Johannsson,O., Borg,Å. *et al.* (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.*, **72**, 1117–1130.
6. Rebbeck,T.R., Friebel,T., Lynch,H.T., Neuhausen,S.L., Veer,L., Garber,J.E., Evans,G.R., Narod,S.A., Isaacs,C., Matloff,E. *et al.* (2004) Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: the PROSE study group. *J. Clin. Oncol.*, **22**, 1055–1062.
7. Rebbeck,T.R., Lynch,H.T., Neuhausen,S.L., Narod,S.A., Veer,L., Garber,J.E., Evans,G., Isaacs,C., Daly,M.B., Matloff,E. *et al.* (2002) Prophylactic oophorectomy in carriers of BRCA1 or BRCA2 mutations. *N. Engl. J. Med.*, **346**, 1616–1622.
8. Rebbeck,T.R., Levin,A.M., Eisen,A., Snyder,C., Watson,P., Cannon-Albright,L., Isaacs,C., Olopade,O., Garber,J.E., Godwin,A.K. *et al.* (1999) Breast cancer risk after bilateral prophylactic oophorectomy in BRCA1 mutation carriers. *JNCI J. Natl. Cancer Inst.*, **91**, 1475–1479.
9. Kurian,A.W., Hare,E.E., Mills,M.A., Kingham,K.E., McPherson,L., Whittemore,A.S., McGuire,V., Ladabaum,U., Kobayashi,Y., Lincoln,S.E. *et al.* (2014) Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. *J. Clin. Oncol.*, **32**, 2001–2009.
10. Solomon,I., Harrington,E., Hooker,G., Erby,L., Axilbund,J., Hampel,H., Semotiuk,K., Blanco,A., Klein,W.M.P., Giardiello,F. *et al.* (2017) Lynch syndrome limbo: patient understanding of variants of uncertain significance. *J. Genet. Couns.*, **26**, 866–877.
11. Li,J., Zhao,T., Zhang,Y., Zhang,K., Shi,L., Chen,Y., Wang,X. and Sun,Z. (2018) Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res.*, **46**, 7793–7804.
12. Hassan,M.S., Shaalan,A.A., Dessouky,M.I., Abdelnaiem,A.E. and ElHefnawi,M. (2019) A review study: computational techniques for expecting the impact of non-synonymous single nucleotide variants in human diseases. *Gene*, **680**, 20–33.
13. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
14. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
15. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
16. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
17. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
18. Schwarz,J.M., Rödelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
19. Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
20. Shihab,H.A., Gough,J., Cooper,D.N., Stenson,P.D., Barker,G.L.A., Edwards,K.J., Day,I.N.M. and Gaunt,T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models. *Hum. Mutat.*, **34**, 57–65.
21. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, **12**, 931–934.
22. Ritchie,G.R.S., Dunham,I., Zeggini,E. and Flicek,P. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.
23. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
24. Ioannidis,N.M., Rothstein,J.H., Pejaver,V., Middha,S., McDonnell,S.K., Baheti,S., Musolf,A., Li,Q., Holzinger,E., Karyadi,D. *et al.* (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.*, **99**, 877–885.
25. Jagadeesh,K.A., Wenger,A.M., Berger,M.J., Guturu,H., Stenson,P.D., Cooper,D.N., Bernstein,J.A. and Bejerano,G. (2016) M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.*, **48**, 1581–1586.
26. Quang,D., Chen,Y. and Xie,X. (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–763.
27. Zeng,Z. and Bromberg,Y. (2019) Predicting functional effects of synonymous variants: a systematic review and perspectives. *Front. Genet.*, **10**, 914.
28. McGuire,A.L., Gabriel,S., Tishkoff,S.A., Wonkam,A., Chakravarti,A., Furlong,E.E.M., Treutlein,B., Meissner,A., Chang,H.Y., López-Bigas,N. *et al.* (2020) The road ahead in genetics and genomics. *Nat. Rev. Genet.*, **21**, 581–596.
29. Tuffley,C. and Steel,M. (1998) Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.*, **147**, 63–91.
30. Galtier,N. (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, **18**, 866–873.
31. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci.*, **96**, 4285–4288.
32. Tabach,Y., Billi,A.C., Hayes,G.D., Newman,M.A., Zuk,O., Gabel,H., Kamath,R., Yacoby,K., Chapman,B., Garcia,S.M. *et al.* (2013) Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature*, **493**, 694–698.
33. Tabach,Y., Golan,T., Hernández-Hernández,A., Messer,A.R., Fukuda,T., Kouznetsova,A., Liu,J., Lilienthal,I., Levy,C. and Ruvkun,G. (2013) Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Mol. Syst. Biol.*, **9**, 692.
34. Sadreyev,I.R., Ji,F., Cohen,E., Ruvkun,G. and Tabach,Y. (2015) PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic Acids Res.*, **43**, W154–W159.
35. Bloch,I., Sherill-Rofe,D., Stupp,D., Unterman,I., Beer,H., Sharon,E. and Tabach,Y. (2020) Optimization of co-evolution analysis through phylogenetic profiling reveals pathway-specific signals. *Bioinformatics*, **36**, 4116–4125.
36. Braun,M., Sharon,E., Unterman,I., Miller,M., Shtern,A.M., Benenson,S., Vainstein,A. and Tabach,Y. (2020) ACE2 co-evolutionary pattern suggests targets for pharmaceutical intervention in the COVID-19 pandemic. *Iscience*, **23**, 101384.

37. Unterman,I., Bloch,I., Cazacu,S., Kazimirsky,G., Ben-Zeev,B., Berman,B.P., Brodie,C. and Tabach,Y. (2021) Expanding the MECP2 network using comparative genomics reveals potential therapeutic targets for rett syndrome. *Elife*, **10**, e67085.
38. Szurmant,H. and Weigt,M. (2018) Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr. Opin. Struct. Biol.*, **50**, 26–32.
39. Croce,G., Gueudré,T., Ruiz Cuevas,M.V., Keidel,V., Figliuzzi,M., Szurmant,H. and Weigt,M. (2019) A multi-scale coevolutionary approach to predict interactions between protein domains. *PLOS Comput. Biol.*, **15**, e1006891.
40. Sferra,G., Ponzi,M. and Pizzi,E. (2018) Molecular interplay between organisms by phylogenetic profiling. *PeerJ Prepr.*, **6**, e27373v1.
41. Sherill-Rofe,D., Rahat,D., Findlay,S., Mellul,A., Guberman,I., Braun,M., Bloch,I., Lalezari,A., Samiei,A., Sadreyev,R. *et al.* (2019) Mapping global and local coevolution across 600 species to identify novel homologous recombination repair genes. *Genome Res.*, **29**, 439–448.
42. Li,Y., Calvo,S.E., Gutman,R., Liu,J.S. and Mootha,V.K. (2014) Expansion of biological pathways based on evolutionary inference. *Cell*, **158**, 213–225.
43. Stupp,D., Sharon,E., Bloch,I., Zitnik,M., Zuk,O. and Tabach,Y. (2021) Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat. Commun.*, **12**, 6454.
44. Tsaban,T., Stupp,D., Sherill-Rofe,D., Bloch,I., Sharon,E., Schueler-Furman,O., Wiener,R. and Tabach,Y. (2021) CladeOScope: functional interactions through the prism of clade-wise co-evolution. *NAR Genomics Bioinforma.*, **3**, lqab024.
45. Hopf,T.A., Colwell,L.J., Sheridan,R., Rost,B., Sander,C. and Marks,D.S. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.
46. Marks,D.S., Hopf,T.A. and Sander,C. (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–1080.
47. Weinreb,C., Riesselman,A.J., Ingraham,J.B., Gross,T., Sander,C. and Marks,D.S. (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell*, **165**, 963–975.
48. Arneson,A. and Ernst,J. (2019) Systematic discovery of conservation states for single-nucleotide annotation of the human genome. *Commun. Biol.*, **2**, 248.
49. Harrison,S.M., Riggs,E.R., Maglott,D.R., Lee,J.M., Azzariti,D.R., Niehaus,A., Ramos,E.M., Martin,C.L., Landrum,M.J. and Rehm,H.L. (2016) Using clinvar as a resource to support variant interpretation. *Curr. Protoc. Hum. Genet.*, **89**, 8.16.1–8.16.23.
50. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC genome browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
51. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F.A., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
52. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
53. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The biomart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
54. Lawrence,M., Gentleman,R. and Carey,V. (2009) rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, **25**, 1841–1842.
55. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
56. Afgan,E., Baker,D., Batut,B., van den Beek,M., Bouvier,D., Čech,M., Chilton,J., Clements,D., Coraor,N., Grüning,B.A. *et al.* (2018) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.
57. Karolchik,D. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, 493–496.
58. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
59. Gómez-Rubio,V. (2017) ggplot2 - elegant graphics for data analysis (2nd edition). *J. Stat. Softw.*, **77**, 1–3.
60. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.-C. and Müller,M. (2011) pROC: an open-source package for r and S+ to analyze and compare ROC curves. *BMC Bioinform.*, **12**, 77.
61. Wei,T.S.V. (2021) R package 'corrplot': visualization of a correlation matrix (version 0.84). https://github.com/taiyun/corrplot(27 January 2021, date last accessed).
62. Paradis,E. and Schliep,K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.
63. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
64. Kuhn,M. (2008) Building predictive models in r using the caret package. *J. Stat. Softw.*, **28**, 1–26.
65. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
66. Malhis,N., Jones,S.J.M. and Gsponer,J. (2019) Improved measures for evolutionary conservation that exploit taxonomy distances. *Nat. Commun.*, **10**, 1556.
67. Amemiya,C.T., Powers,T.P., Prohaska,S.J., Grimwood,J., Schmutz,J., Dickson,M., Miyake,T., Schoenborn,M.A., Myers,R.M., Ruddle,F.H. *et al.* (2010) Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. *Proc. Natl. Acad. Sci.*, **107**, 3622–3627.
68. Braasch,I., Gehrke,A.R., Smith,J.J., Kawasaki,K., Manousaki,T., Pasquier,J., Amores,A., Desvignes,T., Batzel,P., Catchen,J. *et al.* (2016) The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.*, **48**, 427–437.
69. Patel,K.J., Yu,V.P.C.C., Lee,H., Corcoran,A., Thistlethwaite,F.C., Evans,M.J., Colledge,W.H., Friedman,L.S., Ponder,B.A.J. and Venkitaraman,A.R. (1998) Involvement of BRCA2 in DNA repair. *Mol. Cell*, **1**, 347–357.
70. Moynahan,M.E., Pierce,A.J. and Jasin,M. (2001) BRCA2 is required for homology-directed repair of chromosomal breaks. *Mol. Cell*, **7**, 263–272.
71. Moynahan,M.E., Chiu,J.W., Koller,B.H. and Jasin,M. (1999) BRCA1 controls homology-directed DNA repair. *Mol. Cell*, **4**, 511–518.
72. Moynahan,M.E. and Jasin,M. (2010) Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nat. Rev. Mol. Cell Biol.*, **11**, 196–207.
73. Frazer,J., Notin,P., Dias,M., Gomez,A., Min,J.K., Brock,K., Gal,Y. and Marks,D.S. (2021) Disease variant prediction with deep generative models of evolutionary data. *Nature*, **599**, 91–95.
74. Sundaram,L., Gao,H., Padigepati,S.R., McRae,J.F., Li,Y., Kosmicki,J.A., Fritzilas,N., Hakenberg,J., Dutta,A., Shon,J. *et al.* (2018) Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.*, **50**, 1161–1170.
75. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
76. Ionita-Laza,I., McCallum,K., Xu,B. and Buxbaum,J.D. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.
77. Huang,Y.-F., Gulko,B. and Siepel,A. (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.