Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity

Graphical abstract

Authors
Mathieu Quinodoz, Virginie G. Peter, Katarina Cisarova, ... , Sheila Unger, Andrea Superti-Furga, Carlo Rivolta

Correspondence
carlo.rivolta@iob.ch
Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity

Mathieu Quinodoz, Virginie G. Peter, Katarina Cisarova, Beryl Royer-Bertrand, Peter D. Stenson, David N. Cooper, Sheila Unger, Andrea Superti-Furga, and Carlo Rivolta

Summary

We used a machine learning approach to analyze the within-gene distribution of missense variants observed in hereditary conditions and cancer. When applied to 840 genes from the ClinVar database, this approach detected a significant non-random distribution of pathogenic and benign variants in 387 (46%) and 172 (20%) genes, respectively, revealing that variant clustering is widespread across the human exome. This clustering likely occurs as a consequence of mechanisms shaping pathogenicity at the protein level, as illustrated by the overlap of some clusters with known functional domains. We then took advantage of these findings to develop a pathogenicity predictor, MutScore, that integrates qualitative features of DNA substitutions with the new additional information derived from this positional clustering. Using a random forest approach, MutScore was able to identify pathogenic missense mutations with very high accuracy, outperforming existing predictive tools, especially for variants associated with autosomal-dominant disease and cancer. Thus, the within-gene clustering of pathogenic and benign DNA changes is an important and previously underappreciated feature of the human genome, which can be harnessed to improve the prediction of pathogenicity and disambiguation of DNA variants of uncertain significance.

Introduction

It has been previously noted that mechanisms associated with the pathogenesis of missense variants often correlate with the three-dimensional structure of proteins and that, for some disease-associated genes, mutations appear to cluster within specific regions. More systematic analyses have identified DNA sub-regions intolerant to missense variants and domains in protein families enriched in variants associated with disease. In addition, within many genes, pathogenic missense variants tend to cluster within specific domains or regions of the encoded proteins, whereas most loss-of-function variants do not, with the exception of the penultimate and last exons where premature termination codons can escape nonsense-mediated decay (NMD). This information has been used empirically to estimate the pathogenic potential of newly detected variants for individual genes. Surprisingly perhaps, the same information has never been systematically considered to determine clusters of benign and deleterious variants across the entire human exome, or to score the pathogenicity of DNA changes on the same scale.

Identifying variants that underlie Mendelian phenotypes and cancer represents a major challenge in genetic medicine. Achieving this goal relies critically on developing the capability to recognize the one or a few mutations that have a clinical impact among the hundreds of thousands of benign variants which are normally present in an individual’s genome. Next generation sequencing (NGS) is now being applied routinely in clinical diagnostic settings, making the prediction of pathogenic DNA variants, and particularly of missense variants, a crucial element not only for medical research, but also in the context of patient diagnosis and disease management. The American College of Medical Genetics and Genomics (ACMG) has provided guidelines for the classification of variants into five categories: pathogenic, likely pathogenic, variants of uncertain significance (VUSs), likely benign, and benign, thereby establishing a terminology that has been widely adopted by the molecular genetics community. Recently this approach has been validated by using a Bayesian classification framework. However, these criteria still leave many rare missense variants classified as VUSs, resulting in an impasse at the diagnostic level. Many in silico tools have been developed over the past few years to help with this problem, delineating features that are typical of variants with benign versus pathogenic outcomes. Such features are generally uni-dimensional and include evolutionary conservation at the nucleotide or amino acid sequence levels, the structure of the protein, and the severity of the amino acid substitution in terms of change in hydrophobicity, charge, and size. Some of these tools exploit these features per se, whereas others take advantage of machine learning approaches and are trained on sets of pathogenic versus benign DNA changes.

1Institute of Molecular and Clinical Ophthalmology Basel, 4031 Basel, Switzerland; 2Department of Ophthalmology, University of Basel, 4031 Basel, Switzerland; 3Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK; 4Institute of Experimental Pathology, Lausanne University Hospital (CHUV), 1011 Lausanne, Switzerland; 5Division of Genetic Medicine, University of Lausanne and Lausanne University Hospital (CHUV), 1011 Lausanne, Switzerland; 6Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

*Correspondence: carlo.rivolta@iob.ch

https://doi.org/10.1016/j.ajhg.2022.01.006.
© 2022 The Authors. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Finally, meta-predictors offer an optimized combination of existing tools, such as those mentioned above, although they sometimes suffer from circularity issues that are prone to producing falsely optimistic results (e.g., Grimm et al.56).

In this work, we address the topic of within-gene distribution of pathogenic and benign variants, revealing significant clustering of missense changes across the entire coding genome, and use this information to build a positional score. This value is then integrated with other unsupervised independent features (to avoid any circularity issues) to yield a pathogenicity score (MutScore) that has high discriminatory power for variants of uncertain significance.

Material and methods

Computation of the positional score

First, we extracted missense variants reported in ClinVar as of November 21, 2020 as pathogenic and likely pathogenic (PLP) and as benign and likely benign (BLB). Since some genes have a reduced number of BLB missense variants within ClinVar, we first computed the maximal allele frequency (AF) in gnomAD (maximum of gnomAD_exomes_AF and gnomAD_genomes_AF from dbNFSP4.0) of all reported PLP variants (missense, LoF, and others) for every gene. We then termed this value max-AF-PLP (maximum allelic frequency of PLP variants per gene) and extracted variants having a higher frequency than this value within the gnomAD dataset (r.2.1.1, PASS variants only), since variants occurring more frequently than max-AF-PLP can be considered as likely benign (ACMG criteria BS1). This allowed the addition of 354,888 BLB variants to our analysis and the building of the positional score as well as the amino acid change score (see below, Table S1).

To build our positional score, we used a random forest approach for each transcript, considered individually, having at least ten PLP (arbitrary threshold) and one BLB (the random forest needs at least one negative observation) variants in the training set (see below for details of the set). We used the randomForest function from the R package randomForest (v.4.6.14), with default parameters, except for the ntree value, which was set to 1,000. We considered as positive cases the amino acid positions of PLP missense variants and as negative cases all BLB missense variants (including common gnomAD variants with AF higher than max-AF-PLP) from the training set. Then, for every variant, we selected the score from the isoform in which the highest number of PLP variants was found and defined this value as the positional score. The positional score was set to zero for variants that were present only in transcripts with fewer than ten missense PLP variants in the training set.

Cluster analysis

We computed clusters of missense PLP variants in every transcript of every annotated gene by taking regions of consecutive positional scores higher than 0.05 and containing 5 or more PLP missense variants. A clustering score was also computed, defined as the minimum value between the fraction of PLP missense variants located inside detected clusters (over the total number of PLP variants; this is indicative of the precision of the clustering) and the fraction of the transcript not covered by clusters (over the total length of the transcript; this is indicative of the density of the clustering). Hence, the resulting score ranges between 0 and 1, with higher values indicating higher clustering. We then performed a permutation test for every transcript to evaluate the significance of the clustering score. More precisely, for each transcript, we ran 1,000 simulations by randomly assigning new amino acids positions for PLP and BLB and computing the resulting clustering score. With the resulting scores of these simulations, we could derive a p value and considered as significant those transcripts for which a p value below 0.05 with false discovery rate (FDR) correction was obtained.

Genes were also classified according to their clustering scores. Specifically, they were defined as having low, medium, or high PLP clustering if their best-performing isoform had scores below 0.33, between 0.33 and 0.66, or above 0.66, respectively.

A similar analysis was performed for the cluster analysis of BLB variants, by taking regions of consecutive positional scores lower than 0.05 and containing 10 or more BLB missense variants.

Finally, genes were classified according to the mode of inheritance of their associated phenotype, according to the OMIM database.57 More precisely, genes were included in the autosomal-dominant class if all of their associated phenotypes were labeled in OMIM as “autosomal dominant (AD)” and not “autosomal recessive (AR)” or “somatic mutation (SMu),” whereas they were included in the autosomal-recessive class if all their associated phenotypes were labeled as “autosomal recessive.” For the somatic class, we selected genes with at least one phenotype linked to the label “somatic mutation” and no occurrences of “autosomal recessive.” Genes linked to phenotypes having non-Mendelian or uncertain inheritance (OMIM entries beginning with a question mark or within curly brackets) were excluded from our analyses.

Selection of existing features

To build our model, we first selected all unsupervised scores available through dbNFSP4.028 that were also relevant to coding variants, namely SIFT,24 SIFT4G,59 LRT,25 PROVEAN,25 GERP++,RS,27 phyloP100way, phyloP30way, and phyloP17way,26 phastCons100way, phastCons30way, and phastCons17way,28 Siphy29way,30 dbscSNV-ADA, and dbscSNV-RF.60 Proportion expressed across transcripts (pext score)45 were downloaded separately from the gnomAD database, since they are not available within dbNFSP4.0. The maximal and the mean score for every variant was then computed, yielding two features: pext-max and pext-mean. In order to replace missing values, we computed the median value of all missense mutations from dbNFSP4.0 for each feature, except for dbscSNV-ADA, dbscSNV-RF, and pext scores, for which missing values were simply zeroed.

Computation of the amino acid score

The amino acid change score was computed as follows. First, for every possible missense variant (e.g., Arg > Trp), we computed the number of genes containing this change as PLP in at least one instance (value a) as well as the number of genes containing this change as BLB, again at least once (value b). We then defined the amino acid score as being \(a/(a+b)\), for all missense changes. This approach was adopted so as to avoid any potential bias from genes with numerous identical changes, such as for instance Gly > Xaa substitutions in collagen triple helices.62
Annotation of variants
All variants from the training set (see below), the testing sets (see below), gnomAD, and ClinVar were annotated using ANNOVAR\textsuperscript{53} (based on RefSeq genes) with data from the dbNSFP v.4.0 database and with custom-made tables for scores that were not present in this dataset. These tables included data from ClinPred,\textsuperscript{52} dbSNV-ADA and -RF,\textsuperscript{60} as well as CONDEL.\textsuperscript{48}

Training set
For the training set, we used variants reported in ClinVar as of November 21, 2020. The VCF file was downloaded from the ClinVar website. PLP variants were defined as annotated either as pathogenic or likely pathogenic and BLB variants as annotated benign or likely benign (Table S1). Variants with conflicting interpretations (CI) and variants of uncertain significance (VUSs) were discarded from the training set.

We excluded ClinPred from comparison with MutScore for the training set and testing sets since it uses directly allelic frequency (AF) from gnomAD as a feature, and therefore it is biased against AF, since most BLB variants have higher AF than PLP variants. By stratifying the analysis by AF, we could indeed show that the performance of ClinPred is lower for very rare variants (Figure S1).

Computation of the prediction score (MutScore)
MutScore was built using a random forest approach (randomForest function from the R package randomForest, default parameters, except ntree = 1,000; increasing the number of trees did not improve the performance of the model) with selected existing and novel features, as described above (n = 18; SIFT, SIFT4G, LRT, PROVEAN, GERP++, phylolP100way, phylolP30way, phylolP17way, phastCons100way, phastCons30way, phastCons17way, Siphy29way, dbSNV-ADA, dbSNV-RF, pext-mean, pext-max, amino acid change score, and positional score) on the training set. The ratio of pathogenic out-of-box (OOB) votes over the total number of votes outputted by the random forest model was taken as the score for the variants of the training set for all further analysis. The score for other variants was obtained as the resulting probability of the model (R function “predict”).

10-fold cross-validation on the training set
The training set used to build MutScore (PLP and BLB variants from ClinVar) was split randomly into ten equal parts. Iteratively, nine parts were considered. These parts constituted the training subset of the cross-validation, whereas the remaining tenth was taken as a validation subset. This subdivision was used to compute the amino acid change score, the positional score, and finally the “MutScore.” Subsequently, performance was assessed on both the ten training and the ten validation subsets. AUCs were computed for all permutations, yielding ten values for the training subsets and ten values for the validation subsets. These two groups of AUCs were then compared using an unpaired t test (R function “t.test”), with unequal variance (Figure S2).

Testing sets
For testing set 1, we used ClinVar PLP and BLB variants that were present in the database on September 19, 2021. We excluded variants that were present in the training set (entries present in ClinVar up to November 21, 2020), that were used to train the positional score, or that were present in testing set 3 (see below). This procedure yielded 1,867 PLP and 459 BLB variants (Table S1).

We further stratified PLP variants from the testing set 1 according to the star-based classification system in ClinVar: zero stars (no assertion criteria provided), one star (criteria provided, single submitter), two stars (criteria provided, multiple submitters, no conflicts), three stars (reviewed by expert panel), and four stars (practice guideline). An additional stratification included their molecular origin, i.e., germline only (CLNORIGIN = 1) or de novo (CLNORIGIN = 32 or 33). To establish the performance of the tested tools on these subsets of variants, we used only BLB variants for genes also carrying retained PLP variants, to avoid type 2 circularity.\textsuperscript{56}

For testing set 2, we used as PLP variants all disease-causing missense mutations (DM) from the HGMD database (v.2020.2) that were added since 2017 and were neither in ClinVar nor were included in the training set or another testing set (n = 14,327). Since the HGMD database does not contain BLB variants, we used all missense variants from gnomAD that were (1) absent from the training set, (2) not used to train the positional score, (3) absent from both HGMD and ClinVar, and (4) present only in genes for which we defined at least one PLP variant. In order to have a similar number of BLB and PLP variants, we used an AF threshold of >0.000177 in gnomAD (maximum value of exome and genome subsets), resulting in the selection of 13,248 BLB variants (Table S1). For the analysis of AUCs as a function of time (Figure S3B), we used only HGMD variants published during the year considered.

For testing set 3, we used variants from the DoCM database as PLP variants, excluding variants from the training set and variants used to build the positional and amino acid change features (n = 205). Since the DoCM database only contains pathogenic variants, BLB variants were selected according to the same criteria described for testing set 2 (n = 207, Table S1).

For all testing sets, AUCs of MutScore were compared to other tools using the DeLong test in the “roc.test” function from the package pROC (v1.17.0.1) with default arguments, except for “method=delong.”

Analysis of VUS and CI variants
VUSs and CI variants (ClinVar, dataset of November 21, 2020) in the 3,663 genes with at least one PLP variant in our training set were stratified using the same software.

10-fold cross-validation on the training set
The training set used to build MutScore (PLP and BLB variants from ClinVar) was split randomly into ten equal parts. Iteratively, nine parts were considered. These parts constituted the training subset of the cross-validation, whereas the remaining tenth was taken as a validation subset. This subdivision was used to compute the amino acid change score, the positional score, and finally the “MutScore.” Subsequently, performance was assessed on both the ten training and the ten validation subsets. AUCs were computed for all permutations, yielding ten values for the training subsets and ten values for the validation subsets. These two groups of AUCs were then compared using an unpaired t test (R function “t.test”), with unequal variance (Figure S2).

Testing sets
For testing set 1, we used ClinVar PLP and BLB variants that were present in the database on September 19, 2021. We excluded variants that were present in the training set (entries present in ClinVar up to November 21, 2020), that were used to train the positional score, or that were present in testing set 3 (see below). This procedure yielded 1,867 PLP and 459 BLB variants (Table S1).

We further stratified PLP variants from the testing set 1 according to the star-based classification system in ClinVar: zero stars (no assertion criteria provided), one star (criteria provided, single submitter), two stars (criteria provided, multiple submitters, no conflicts), three stars (reviewed by expert panel), and four stars (practice guideline). An additional stratification included their molecular origin, i.e., germline only (CLNORIGIN = 1) or de novo (CLNORIGIN = 32 or 33). To establish the performance of the tested tools on these subsets of variants, we used only BLB variants for genes also carrying retained PLP variants, to avoid type 2 circularity.\textsuperscript{56}

For testing set 2, we used as PLP variants all disease-causing missense mutations (DM) from the HGMD database (v.2020.2) that were added since 2017 and were neither in ClinVar nor were included in the training set or another testing set (n = 14,327). Since the HGMD database does not contain BLB variants, we used all missense variants from gnomAD that were (1) absent from the training set, (2) not used to train the positional score, (3) absent from both HGMD and ClinVar, and (4) present only in genes for which we defined at least one PLP variant. In order to have a similar number of BLB and PLP variants, we used an AF threshold of >0.000177 in gnomAD (maximum value of exome and genome subsets), resulting in the selection of 13,248 BLB variants (Table S1). For the analysis of AUCs as a function of time (Figure S3B), we used only HGMD variants published during the year considered.

For testing set 3, we used variants from the DoCM database as PLP variants, excluding variants from the training set and variants used to build the positional and amino acid change features (n = 205). Since the DoCM database only contains pathogenic variants, BLB variants were selected according to the same criteria described for testing set 2 (n = 207, Table S1).

For all testing sets, AUCs of MutScore were compared to other tools using the DeLong test in the “roc.test” function from the package pROC (v1.17.0.1) with default arguments, except for “method=delong.”

Analysis of VUS and CI variants
VUSs and CI variants (ClinVar, dataset of November 21, 2020) in the 3,663 genes with at least one PLP variant in our training set were stratified using the same software.

10-fold cross-validation on the training set
The training set used to build MutScore (PLP and BLB variants from ClinVar) was split randomly into ten equal parts. Iteratively, nine parts were considered. These parts constituted the training subset of the cross-validation, whereas the remaining tenth was taken as a validation subset. This subdivision was used to compute the amino acid change score, the positional score, and finally the “MutScore.” Subsequently, performance was assessed on both the ten training and the ten validation subsets. AUCs were computed for all permutations, yielding ten values for the training subsets and ten values for the validation subsets. These two groups of AUCs were then compared using an unpaired t test (R function “t.test”), with unequal variance (Figure S2).

Testing sets
For testing set 1, we used ClinVar PLP and BLB variants that were present in the database on September 19, 2021. We excluded variants that were present in the training set (entries present in ClinVar up to November 21, 2020), that were used to train the positional score, or that were present in testing set 3 (see below). This procedure yielded 1,867 PLP and 459 BLB variants (Table S1).
(v.1.2.0), wairter (v.0.2.0), and caTools (v.1.18.1). We computed the data obtained from ClinVar, gnomAD, conservation scores within dbNSFPv4.0, as well as pext scores from gnomAD to build a graphical representation of the mutational landscape for genes with at least one PLP variant in ClinVar (n = 3,663). UniProt information about regions and domains was obtained with the UniProt.ws and drawProteins packages.

Results

Detection of intragenic variant clustering
We set out to investigate whether the clustering of missense lesions could be a general feature of the entire human morbid genome, rather than a phenomenon limited to a few specific loci, families of genes/proteins, or conserved domains. We selected all pathogenic or likely pathogenic (PLP) missense DNA variants reported in the ClinVar database,64 the largest public database assessing the pathogenicity of known human variants, as a reference set for disease-causing mutations in the human genome, whereas benign and likely benign substitutions (BLB) were extracted from both ClinVar and gnomAD65 (see Material and methods for details on variant selection). We then determined a “positional score,” based on a random forest model, for each transcript of every gene annotated in the RefSeq database (n = 52,630)66 using the position of amino acids affected by PLP and by BLB missense variants as a feature. This process allowed us to associate a score for the likelihood of pathogenicity for every single codon of every known transcript. Following the selection of disease genes harboring sufficient (10 or more) PLP variants to potentially allow the detection of clustering (840 genes), we identified 3,854 regions in 740 genes with multiple PLP variants and positional scores above a minimal threshold (as defined in the Material and methods). We then computed the clustering score, a parameter assessing both the precision and the density of variant clustering within a given transcript, and subdivided these 740 genes into three classes (with low, medium, or high clustering), as a function of this score (Figures 1A and 1B). The biological significance of such clustering was then assessed by means of a permutation test, and 387 genes were found to have at least one transcript with a p value below 0.05 after FDR correction, indicating that statistically significant clustering of pathogenic missense mutations occurred in more than 40% of all well-characterized disease genes (Figures 1A and 1B, Table S2). In addition, more than 79% of significantly clustered genes had at least one transcript with medium or high clustering (Figure 1B; examples of one gene for each class, Figures 2A–2C; global overview, Figure S4). Interestingly, a large majority of the genes that were associated only with autosomal-dominant phenotypes or with the presence of clinically relevant somatic mutations showed significant clustering (65.9% and 79.6% for the autosomal-dominant and the somatic classes, respectively, Figures SSB and SSC, Table S3), whereas corresponding significant values were obtained for a minority of genes linked to autosomal recessive-only phenotypes (19.7%), with 2.6% of them displaying high clustering (Figure S5A, Table S3).

The same analysis was performed for BLB variants. Within the 851 genes considered (i.e., carrying 10 or more BLB missense changes), 10,136 enriched regions in 490 of them were identified, including 172 (20.2% of the total) displaying a significant clustering score (Figures 1C and 1D, Table S2). However, only 9 genes (1.1%) exhibited high clustering of BLB variants (Figure 1D). The majority of transcripts showed precise but sparse BLB clustering (bottom right, Figure 1C), as for example NSD1, which also displays high PLP clustering (Figure 2D). Sub-classification of BLB variants according to recessive versus dominant and somatic occurrence resulted in a trend similar to that observed for PLP variants (Figures S5D–S5F, Table S3).

Implementation of MutScore
We developed MutScore with the goal of exploiting the biological information contained in this patterning of variants to predict the pathogenicity of DNA changes, and potentially obtaining additional knowledge on the molecular mechanisms underlying pathogenicity. In addition, we aimed at an improved classification of VUSs, which in whole-genome sequencing (WGS) or whole-exome sequencing (WES) studies are mostly represented by low-frequency missense substitutions.

In addition to the positional score described above, we also calculated an “amino acid change score,” which is intended to approximate the likelihood of a particular amino acid substitution (e.g., Arg > Trp) to result in a pathological phenotype (see Material and methods for details on both scores). Moreover, we collected 16 additional existing features from published predictive tools, all of which originated from unsupervised approaches,24–30,59–61 to avoid artificially inflating the resulting predictive power (Figure 3A).

Finally, we built another random forest model that considered these 18 features, trained on a set containing only two discrete classes: PLP (36,966 variants) and BLB (29,066 variants), as assessed by ClinVar in its release of November 21, 2020 (Table S1). The importance of each feature was computed as either mean decrease in accuracy (Figure 3B) or decrease in Gini index (mean decrease in impurity, Figure 3C). In both cases, the positional score as well as the PROVEAN25 and SIFT24 scores were found to be the three most important features (Figures 3B and 3C, Table S4). The least important features were represented by the output of two splicing predictors, probably because only a small minority of missense variants are predicted to alter splicing.

To generate a final score for pathogenicity (MutScore), the model was applied to all single-nucleotide substitutions resulting in a missense change, for all possible RefSeq genes (range: 0–1, representing the likelihood for a given missense variant to be pathogenic). A 10-fold cross-validation step was also performed to check
whether overfitting had occurred during training. The areas under the curve (AUCs) from the validation sets from this cross-validation procedure (average \(= 0.949\)) were not statistically different from the corresponding training sets (average \(= 0.950\)) (two-tailed unpaired t test, \(p = 0.701\), Figure S2), showing no overfitting of the model on the training data.

**Testing sets**

To determine the real predictive power of our tool, we evaluated its performance with respect to different sets of data that were not used in the training process. Specifically, we considered recent ClinVar entries, logged between November 22, 2020 and September 19, 2021 (testing set 1), the HGMD database \(^{67}\) (testing set 2), and the database of cancer variants DoCM\(^{68}\) (testing set 3).

For testing set 1, including 5,021 PLP and 2,035 BLB variants, MutScore had an AUC of 0.937 (Figure 4A) which was the highest value, at a statistically significant level, across all tools tested (\(p < 0.05\), DeLong test). The closest competitors were REVEL and VEST4, with AUCs of 0.924 and 0.919, respectively. For testing set 2, PLP variants were selected to correspond to all disease-causing mutations (DM) from the HGMD database that were present in neither testing set 1 nor in the training set. In addition, to avoid using variants that were used to train other tools (such as REVEL and VEST4), we deliberately selected only PLP variants from HGMD that were added after such tools.
were published, i.e., in 2017 or at a later date. For BLB entries, we selected variants from the gnomAD database that did not overlap with previous analyses (see Material and methods for details). This resulted in the selection of 14,327 PLP and 13,248 BLB missense variants in 2,603 different genes. Again, MutScore exhibited a significantly higher performance than other tools ($p < 0.05$, DeLong test, Figure 4B, Table S5) with an AUC of 0.915 compared to 0.904 for REVEL, the second-best predictor. No other predictor achieved an AUC above 0.900. For testing set 3, we selected somatic cancer variants from the DoCM database, a curated set of somatic variants with established relevance to cancer biology. More precisely, we considered all variants from this database that were not present in the training set and were not used to build the model as PLP entries, whereas BLB entries were selected according to the same procedures described for testing set 2. This resulted in the identification of 205 PLP and 207 BLB variants. Once more, MutScore had the highest AUC value, in a statistically significant way (AUC = 0.960, $p < 0.05$, DeLong test), followed by M-CAP (AUC = 0.943, Figure 4C, Table S5).

It is interesting to note that the performance of some predictors appeared to decrease when evaluating more recent variants, and in particular when HGMD data prior to 2017 versus post-2017 (testing set 2) were used (Figure S3A). This could be due to certain tools having been overfitted on data used to train them versus data that were completely naïve to them (Figure S3B), a conclusion supported by the observation that untrained tools (SIFT, PROVEAN, GERP, PhyloP, etc.) retained their power on old versus new entries (Figure S3A). More specifically, the performance bias toward older entries by trained versus untrained tools was significant (average differences between new and old entries: $-0.0024$ and $-0.0170$ for untrained and trained tools, respectively; $p$ value $= 5.8 \times 10^{-5}$, by t test, bilateral with unequal variance).

Performance on subsets of testing set 1 from ClinVar
We wished to test whether MutScore performance would reflect the actual pathogenicity of variants as assessed by curated experimental evidence. ClinVar attributes a score ranging from zero to four stars to every entry, depending upon the number of submissions and on data supporting...
its pathogenicity. For instance, it was recently shown that the tool Rhapsody classified PLP variants with zero stars (i.e., no assertion criteria provided) less accurately than other variants. MutScore's AUCs correlated well with the number of stars attributed by ClinVar to PLP variants, with values of 0.908 for variants with zero stars, 0.946 for variants with one star (i.e., criteria provided, single submitter), 0.963 for variants with two stars (i.e., criteria provided, multiple submitters, no conflicts), and 0.967 for variants with three stars (i.e., reviewed by expert panel). Since there were too few missense variants with four stars (i.e., practice guideline) to assess, we could not compute performance for such a small dataset. Other prediction tools displayed the same trend (Figure S6, Table S5). MutScore had a significantly higher AUC with 0.946 for the PLP variants with zero stars, which represents more than 66% of all PLP variants.

We also investigated the performance of MutScore according to ClinVar's defined origin of PLP missense variants: strictly germline versus de novo. In this test, MutScore had the highest AUCs for germline and de novo variants (0.938 and 0.908, respectively, Figures S7A and S7B, Table S5). Again, all differences were statistically significant (p < 0.05, DeLong test), except for the comparison with VEST4 on de novo variants. MutScore also displayed a lower decrease in AUC for de novo variants compared to germline variants when compared to VEST4 and REVEL (Table S5). This can be explained by the fact that both de novo and somatic missense mutations tend to exert their influence via dominant gain-of-function mechanisms and hence usually only affect specific portions of a protein (e.g., a kinase domain). Traditional tools consider amino acid residue conservation, but usually do not take into account mutational clustering or positional information, and therefore their predictive power may be less efficient with respect to that of MutScore.

Finally, since the positional score appears to be the most important feature of our model (Figure 3B), we evaluated MutScore's performance for genes that were previously well characterized from a mutational standpoint (HCGs [highly characterized genes], with a positional score > 0) versus genes that were not (PCGs [poorly characterized genes]).

Figure 3. Outline of the procedures used to build MutScore and MutLand and importance of different features within the training set
(A) Framework representing the different steps followed to generate MutScore and MutLand (dark green, database; light green, variant set; light blue, computation; dark blue, scores; orange, graphical output).
(B) Ranking of features in the model based on mean decrease in accuracy.
(C) Ranking of features in the model based on mean decrease in Gini index.
genes], positional score = 0). MutScore displayed the highest AUC for both HCGs and PCGs for testing set 1 (ClinVar, Figures S8A and S8B) as well as for testing set 2 (HGMD, Figures S8C and S8D), although the difference with respect to other top-performing tools for PCGs was only marginal (≤ 0.006 units, overall) and statistically not significant (Table S5). Testing set 3 was not used since it did not contain sufficient PLP variants in PCGs. As expected, performance for HCGs was higher than for PCGs in all cases.

Variants of uncertain significance (VUSs) and with conflicting interpretation (CI)

As a next step, we investigated the ability of MutScore, as well as of the two predictors that displayed the best
performance in previous tests (VEST4 and REVEL), to reclassify ambiguous variants. Following the reassessment of all VUSs and conflicting interpretation (CI) variants from ClinVar, MutScore succeeded in reclassifying 54.7% of VUSs and 61.7% of CI variants, compared to 35.5% and 33.9% of VUSs and 40.7% and 37.3% of CI variants for VEST4 and REVEL, respectively (Figures 5 and S9). In particular, MutScore had an edge in redirecting VUSs and CI variants toward BLB variants (ACMG classes 2 and 1), compared to other predictors (Figure 5, left column and Figure S9). Again, this was probably a consequence of the positional score, which allows for a better assessment of variants with respect to their presence within mutational clusters or outside of them, whereas existing algorithms evaluate variants independently of such regional information.70 We can assume that many variants were reclassified as BLB by virtue of their presence outside pathogenic clusters.

**MutLand and MutScore-batch**

To provide a visual representation of our results in individual genes, as well as to facilitate the scoring of newly identified variants, we created an interactive, web-accessible interface displaying data from ClinVar, gnomAD, and UniProt, conservation scores from other tools, PLP and BLB clusters, as well as the output from MutScore. As an example, Figure 6 shows the MutLand output for KCNQ2, which is known to harbor clusters of PLP missense variants in specific transmembrane segments, in the pore loop, and in some intracellular helices.10,71 In the MutLand representation of this gene, clustering of pathogenic missense variants can be easily identified, whereas PLP LoF variants do not cluster. Many VUSs are scattered along the entire protein sequence. Of these, those affecting amino acids in the C-terminal portion, located outside of PLP clusters and with a lower MutScore value, might then be reclassified as likely benign. To allow the user to interrogate multiple MutScore values simultaneously, we also created a separate web interface, MutScore-batch.

**Discussion**

It has long been known that the pattern of pathogenic variants within a limited number of genes72 or gene classes follows a non-random and biological function-driven distribution,9,10,73 such as variants in the triple-helical region of collagen74,75 or in the BRCT region of BRCA1.76,77 Here we find that clustering of pathogenic missense variants occurs in almost half of all human genes, genome-wide, and that for about 18% of them such clustering is highly delimited. This also applies to benign missense changes, irrespective of their allelic frequency, as clusters of BLB variants are detectable in approximately 20% of all genes associated with a hereditary condition. Based on the distribution of MutScore values in relation to protein regions and protein classes, it is reasonable to assume that this clustering of DNA variants is mainly shaped by pathogenic mechanisms occurring at the protein level. For instance, in humans the majority of dominant mutations lead to gain-of-function or dominant-negative events, usually affecting amino acid residues located within specific domains or regions of a given protein.15,78,79 This phenomenon is clearly reflected in our finding that PLP variants associated with dominant conditions are mostly identified in clusters, and is even more pronounced for somatic variants involved in cancer (Figure S5). By contrast, most recessive missense mutations act via loss-of-function (or reduced
function) mechanisms and in general this type of DNA changes is more dispersed along the protein sequence.\textsuperscript{72} In particular, recessive PLP missense variants tend to cluster less and, if they do, to generate larger and more loosely defined regions. The mirror effect of these events is often visible for BLB variants: in genes associated with dominant conditions, BLB variants tend to cluster anywhere in the gene other than in regions with PLP clusters. Conversely, in genes linked to recessive conditions, BLB clustering is in general absent.

We reasoned that the clustering or dispersion of missense variants might be used to infer the pathogenicity of newly observed DNA changes, similar to the use of evolutionary conservation across species. For this purpose, we developed MutScore, a predictor tool, and its graphical interface, MutLand, to enable the easy visualization of mutational landscapes. In addition to existing unsupervised dimensions, MutScore integrates two novel features into its final predictive model: a positional score and an amino acid change score. Furthermore, it builds heavily on curated information of clinically relevant DNA variants, as defined by ClinVar. The positional score in particular, essentially defining regions of a gene in which pathogenic mutations or benign variants are likely (or unlikely) to occur, appears to be the most important feature, conferring upon MutScore an edge over existing algorithms. Interestingly, the integration of as many benign variants as possible within the model also appears to be very important, since it allows for the determination of a “harmlessness threshold” that is gene specific and improves prediction even more.

When tested with real data from three independent datasets (HGMD and recent ClinVar data for constitutional disorders, as well as DoCM for cancer), MutScore performed markedly better than existing tools in discriminating between pathogenic and non-pathogenic variants. This performance allowed for a high rate of disambiguation of VUSs, a key issue in current NGS-based genetic diagnostics.

One limitation of our scoring approach lies in the fact that prediction is partly dependent on the positional score, and therefore on information pertaining to existing pathogenic variants, such as the number of entries in ClinVar and their quality and accuracy. This explains why MutScore’s performance is still uneven with respect to different genes in the human genome and essentially matches that of other meta-predictors for poorly characterized genes. As more and more variants are identified and their clinical implications are assessed, both information on variant clustering and MutScore’s performances are likely to increase substantially.

Figure 6. Example of a MutLand graphical output
Here we display the output for KCNQ2 (GenBank: NM_172107.4), listing various attributes of this gene/protein, including MutScore. Variants were extracted from ClinVar, release of November 21, 2020.

Please cite this article in press as: Quinodoz et al., Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves..., The American Journal of Human Genetics (2022), https://doi.org/10.1016/j.ajhg.2022.01.006
over time. Furthermore, integration of data on the tridimensional structure of proteins, e.g., from the AlphaFold project, may help the future development of MutScore and improve even further its predictive power.

In conclusion, our analysis of the regional distribution of missense variants within human disease genes reveals that extensive clustering is common, both for pathogenic and benign DNA changes. This observation may help in clarifying protein structure and function and will hopefully prime further research into mechanisms of selection and evolution. Moreover, the AI-driven integration of clustering information in MutScore allows for more accurate pathogenicity scoring and the disambiguation of variants of uncertain significance. Together with its graphical interface MutLand and with MutScore-batch, this tool promises to become a useful instrument in genetic medicine and could be used as a stepping stone for new research projects aiming to define further key properties of the morbid human genome.

Data and code availability

MutScore values can be retrieved from https://iob-genetic.shinyapps.io/mutscore/ or from https://iob-genetic.shinyapps.io/mutscore-batch and they can be used under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. For commercial uses of MutScore, please contact the authors.

The code used to generate the MutScore and the analyses presented here is available at https://github.com/mquinodo/mutscore, except for the part using ANNOVAR, which can be found at https://annovar.openbioinformatics.org/en/latest/.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.01.006.

Acknowledgments

We would like to thank Ms. Sitta Förhr for her careful revision of this manuscript. This work was supported by the Swiss National Science Foundation (grants #176097 and #204285 to C.R.). Research performed by D.N.C. and P.D.S. was partly supported by QIAGEN Inc.

Declaration of interests

D.N.C. and P.D.S. acknowledge QIAGEN Inc. for their financial support through a License Agreement with Cardiff University. The other authors do not declare any conflicts of interest.

Received: August 20, 2021
Accepted: January 11, 2022
Published: February 3, 2022

Web resources

CONDEL, http://bbuglab.irbbarcelona.org/fannsd
dbNSFP, https://sites.google.com/site/joppen/dbNSFP
gnomAD, (downloads), https://gnomad.broadinstitute.org/downloads
MutScore, https://iob-genetic.shinyapps.io/mutscore

References


prediction of pathogenic point mutations via extended features. Bioinformatics 34, 511–513.


