VarSome ACMG Implementation

varsome.com/about/acmg-implementation

version: 8.1.6, dated: Thu Feb 20 08:37:55 CET 2020

Introduction

The ”Standards and guidelines for the interpretation of sequence variants” was published in 2015 by Sue Richards et al. in their seminal paper (ACMG Guidelines), from which our implementation is derived.

The standards were very much written for interpretation by humans, *not machines*, they assume the clinician has a deep knowledge of the domain and relevant papers and conditions. Automating these standards is a matter of interpretation, we have opted to statistically quantify terms such as “hot-spot” or “well known” resulting in many thresholds that are tuned via our calibration process.

Our guiding principle throughout has been to implement the best algorithms we could, following the advice from our clinical advisors, feedback from the VarSome user community, and using statistically justified thresholds. All the rules provide clear natural language explanations of why they were triggered and which evidence was used, or indeed, a full explanation of why the criteria were not met (this is currently only visible in VarSome).

We also strive to continuously improve our implementation, adjusting rules or thresholds, incorporating new data sources, and adding refinements as new publications and methodology changes are suggested.

Key Databases

The VarSome ACMG annotation relies on vast quantities of accurate curated data from the following databases (in no particular order):

1. **UniProt Variants**, provided by UNIPROT, version **16-Mar-2016** (circa 90.3k records)
2. **UniProt Regions**, provided by UNIPROT, version **07-Jan-2020** (circa 199k records)
3. **RefSeq**, provided by NCBI, version **98**
4. **Mitomap**, provided by CHOP, version **27-Nov-2019** (circa 27.5k records)
5. **dbscSNV**, provided by dbNSFP, version **v1.1** (circa 15.0M records)
6. **dbNSFP genes**, provided by dbNSFP, version **v3.4**
7. **dbNSFP-c**, provided by dbNSFP, version **4.0** (circa 82.8M records)
8. **DANN SNVs**, provided by UCI, using version **2014** (circa 9.41G records) for hg19, **unavailable for hg38**

9. **Cosmic Licensed**, provided by Sanger, version **v90**
10. **CGD**, provided by NHGRI, version **11-Feb-2020**
11. **ClinVar**, provided by NCBI, version **11-Feb-2020** (circa 623k records)
12. **Ensembl**, provided by EMBL, version **99**
13. **ExAC genes**, provided by Broad, version **18-Sep-2018**
14. **GERP**, version **2010**
15. **gnomAD exomes**, provided by Broad, using version **2.1.1** (circa 17.2M records) for hg19, and using version **2.1.1** (circa 17.2M records) for hg38
16. **gnomAD exomes coverage**, provided by Broad, version **2.1**
17. **gnomAD genomes**, provided by Broad, using version **2.1.1** (circa 262M records) for hg19, and using version **3** (circa 708M records) for hg38
18. **gnomAD genomes coverage**, provided by Broad, using version **2.1** (circa 3.14G records) for hg19, and using version **3.0** (circa 3.21G records) for hg38
19. **HGNC**, provided by HUGO, version **12-Feb-2020**
20. Papers & classifications contributed by the VarSome community.

(Version information subject to change at any time).

Clinical Evidence

Clinical Evidence is the foundation stone of our ACMG evaluation, we currently source this from:

- ClinVar
- UniProt
- MitoMap
- Publications linked by VarSome users
- VarSome user classifications

The VarSome options allow the user to specify a minimum number of stars to filter ClinVar, so entries with fewer stars will be ignored, or similarly disable clinical classifications from UniProt.

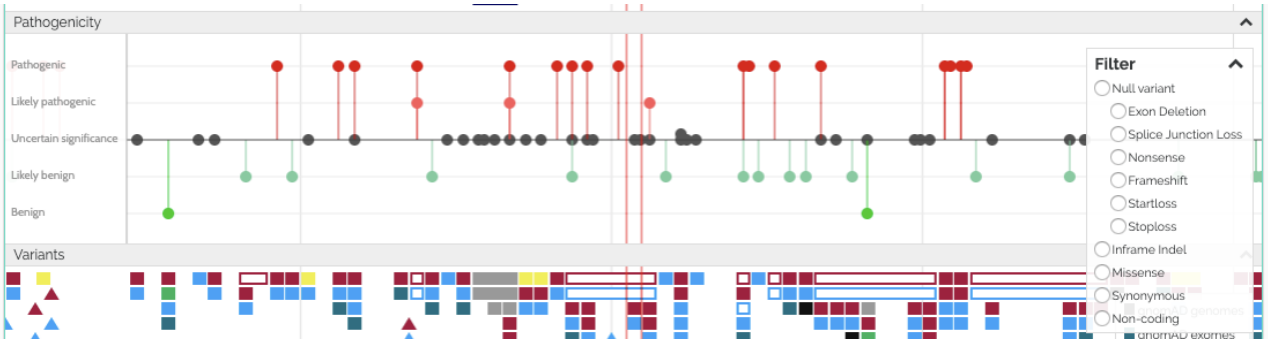
Known Pathogenic

On a daily basis, we re-annotate all the variants from the clinical evidence sources listed above, this data is then used for all the rules that require clinical evidence or derived statistics. For each variant we record its original “source” classification, allele frequency, coding impact, and we also compute its ACMG classification with the clinical evidence rules disabled (PP5 & BP6).

Currently using version **20-Feb-2020** (circa 721k records)

The strengths of rules such as PS1 and PM5 will be downgraded if a variant has been reported pathogenic but that it is not confirmed through the independent ACMG re-annotation.

This database is displayed in VarSome as a “lollipop graph” in the genome browser:



The graph can be filtered by coding impact, or various types of null variants.

Gene Statistics

This database is also rebuilt daily, derived from the known pathogenic variants: it keeps track of how many variants are benign/pathogenic for each gene, along with their coding impacts - these are used in rules PP2 and BP1 for example.

The gene statistics are displayed in the VarSome “gene” page:

We derive a “benign cut-off frequency” from these variant classifications & their allele frequencies for use in rule BS1.

Splice-Site Prediction

We use the scSNV database for splice-site prediction. This is only available for single-nucleotide variants. We use both the 'ADA Boost Splicing' threshold (**0.708**) and 'Random Forest Splicing' threshold (**0.515**) to identify potentially splicing variants for rule BP7.

Conservation

We use GERP++ for conservation tests, this is available for nearly all positions in the hg19 genome.

A position will be considered highly conserved for rules BP7 and BP4 if GERP Rejected Subsitutions (GERP_RS) is greater than **6.8**.

Note however that GERP is not available for hg38 and we therefore skip conservation tests - one of the reasons for which hg38 annotations may be slightly less accurate than for hg19.

Transcript Selection

All the ACMG rules are evaluated against a single transcript. Selecting this transcript is clearly of **critical importance** and can modify the outcome of the classification. Transcripts are prioritised according to the following criteria:

1. Most severe coding impact, or within +/- 2 bases of the splicing site,
2. Canonical,
3. Longest transcript.

The above criteria can be overridden by users as follows:

- Selecting a different transcript in the VarSome UI.
- Configuring transcripts to be used for specific genes in VarSome Clinical.

The Ensembl Transcript Support Level (TSL) is a method to highlight the well-supported and poorly-supported transcript models for users, based on the type and quality of the alignments used to annotate the transcript. We disqualify Ensembl transcripts that have a TSL with a value different from 1.

Note: some variants can be in multiple transcripts associated with multiple genes, although it is rare for a variant to be coding in multiple genes. The rules above will first determine the transcript to use, from which the gene is then derived.

Allele Frequency

VarSome currently uses GnomAD exomes & genomes to evaluate allele counts and frequencies, it uses both the frequency data and the coverage data reported for both these databases.

Frequencies will not be considered valid if:

- Coverage is less than **20**,
- the Allele Number is less than **1000**,
- the GnomAD quality filter is suspect (ie: not PASS).

Rules BA1 and BS1 will iterate through the various ethnicities to see whether the variant is common in a sub-population.

Further databases such as BRAVO and TwinsUK will be incorporated in future.

Calibration

Many of the rules implemented here rely on thresholds, PM1 is a good example where defining a “hot-spot” is clearly a fuzzy measure. In practice we carefully adjust these thresholds through statistical regression against a large population of reliably curated variants. When calibrating, we disable the clinical evidence rules (PP5 and BP6) in order to ensure that the classifier works well in the absence of variant-specific evidence, and thus can be extrapolated reliably beyond the test population. The calibrations are 'fair' in that they do not over-emphasise pathogenic vs benign or uncertain variants, we simply seek to maximise overall accuracy.

Saphetor reserves the right to adjust the implementation of the rules and the calibrated thresholds at any time. In practice this has allowed us to deliver continual improvements in the overall quality of our automated classification - but it also entails that results may change if when re-annotating a variant several months later: methodologies, thresholds, and the clinical data used to calibrate them, may all have changed.

Although we use machine-learning techniques to adjust the thresholds used, we do not use neural-networks in the actual classification itself. We believe it is important to have fully transparent, justifiable and explainable rules, as opposed to inscrutable black-boxes. The 'AI' aspect is also well captured in the computational evidence, DANN being a prime example of how powerful such an approach can be.

Implemented Rules

PVS1

Null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease. (Pathogenic, Very Strong)

The rule first establishes whether this is a null variant by checking its coding impact on the transcript:

- nonsense variant
- frameshift variant
- exon deletion variant
- within ±2 bases of transcript splice site
- start loss variant

We determine that LOF is a “Known Mechanism of Disease” from either:

- The gene statistics: if at least **5** variants in this gene are known to be pathogenic.
- ExAC probability of loss-of-function tolerance is greater than **0.7**.

Purely for information, a list of possible associated diseases is sourced from CGD and reported in the rule explanation.

Note: rule PVS1 disables rule PM4 in order to avoid double-counting the same evidence.

PS1

Same amino acid change as a previously established pathogenic variant regardless of nucleotide change. (Pathogenic, Strong)

This rule only applies to missense variants, it considers all possible **equivalent** amino acid missense variants (ie: resulting in the same amino-acid). If any clinically reported pathogenic variants are identified in the known pathogenic database, we then check whether they are confirmed pathogenic using the ACMG annotation, and the rule triggers with the corresponding strength and explanation.

PS3

Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product. (Pathogenic, Strong)

BS3

Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing. (Benign, Strong)

These two rules leverage the known pathogenic database, looking for papers that refer to in-vitro or functional studies. VarSome user contributions are particularly helpful as users are asked to manually confirm the studies refered to in the paper. For papers linked by ClinVar, UniProt & MitoMap, we automatically scan the title & abstract and look for potential studies.

Ultimately the papers highlighted by this rule must be reviewed by an experienced clinician.

PM1

Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation. (Pathogenic, Moderate)

This rule relies heavily on the known pathogenic database to evaluate how many coding pathogenic variants are found near the variant being considered.

- Hot-Spot: using a region of **30** base-pairs either side of the variant, we check that there are at least **5** pathogenic variants, then weights them by distance to compute a “proximity score”. The rule triggers if the this score is greater than **2.472**.

- Protein Domains: for each UniProt functional domain, the rule will trigger if at least **5** pathogenic variants are found within the domain, and the ratio of pathogenic to non-VUS variants is greater than **0.172**.

The thresholds used by rule PM1 have been established through a careful calibration process and may change over time as further clinical evidence becomes available.

PM2

Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium. (Pathogenic, Moderate)

This rule first uses data from CGD to establish whether the gene is recessive or dominant.

The rule will then trigger if the allele frequency is:

- Not found in GnomAD, with valid GnomAD coverage,
- Less than **0.0001** (see ACMG Guidelines) for recessive genes,
- The allele count less than **5** for dominante genes.

The mode of inhertiance and GnomAD coverage are provided in the rule's explanation.

PM4

Protein length changes as a result of in-frame deletions/insertions in a non-repeat region or stop-loss variants. (Pathogenic, Moderate)

This rule only applies to in-frame indels or stop-loss variants that cause the length of the protein to change. The rule will not fire if the variant is in a repeat region as reported by UniProt or by checking for short repetitive regions in the DNA itself.

In order to avoid double-counting the same evidence, rule PM4 will not be applied if rule PVS1 was triggered.

PM5

Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before. (Pathogenic, Moderate)

This rule is a weaker version of PS1, it similarly only applies to missense variants, but considers all **possible** amino acid missense variants in the same codon. If any clinically reported pathogenic variants are identified in the known pathogenic database, we then check whether they are confirmed pathogenic using the ACMG annotation, and the rule triggers with the corresponding strength and explanation.

PP2

Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease. (Pathogenic, Supporting)

BP1

Missense variant in a gene for which primarily truncating variants are known to cause disease. (Benign, Supporting)

These two “variant spectrum” rules are very similar: they only apply to missense variants and rely heavily on the gene statistics for the relevant gene:

- PP2 checks that the ratio of pathogenic missense variants over all non-VUS missense variants is greater than **0.51**, with a secondary requirement that the ratio of pathogenic variants over all clinically reported variants is greater than **0.12**,
- BP1 conversely checks that the ratio of benign missense variants over all non-VUS missense variants is greater than **0.51**, with a secondary requirement that the ratio of benign variants over all clinically reported variants is greater than **0.24**.

The calibration section explains how these thresholds are established.

PP3

Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.) (Pathogenic, Supporting)

BP4

Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.) (Benign, Supporting)

These two rules use a very similar implementation. The data-sets used are static, many are sourced from dbNSFP which covers all non-synonymous coding single-nucleotide variants.

Based on combined accuracy tests, we have selected the following sub-set of in-silico prediction tools:

- DANN
- DEOGEN2
- FATHMM-MKL
- M-CAP
- Mutation Assessor
- Mutation Taster

- Primate AI
- Polyphen 2 (only for users of VarSome Clinical)
- MVP
- EIGEN
- REVEL
- SIFT
- GERP: a simple conservation test is used if no other data is available.

As more tools become available this list will change. Some tools have far greater coverage than others, for example DANN is available for all SNVs, where most other tools are only available for non-synonymous coding SNVs. The GERP score is available for nearly all positions and is used to establish whether the position is conserved.

Wherever possible we use the default pathogenic/benign predictions from each tool, however for some tools (DANN, SIFT and GERP) we use internally calibrated thresholds.

The algorithm counts the number of pathogenic & benign predictions, and will trigger if the ratio of pathogenic classifications to total classifications (respectively benign) exceeds the **0.53**. We have found this to be significantly more accurate than the unanimous verdict strictly required by the ACMG Guidelines.

In order to avoid double-counting, rule BP4 will not be evaluated if rule BP7 was triggered. It also explicitly checks for conservation itself rather than relying on the in-silico tools alone.

Statistically, if a variant is not found in any static in-silico database, it is most likely to be pathogenic. To refine this somewhat extreme prediction, we use GERP as a simple fall-back in the absence of any other prediction, returning a pathogenic prediction if GERP_RS is greater than **3.597**, or benign if the variant is non-truncating and GERP_RS is less than **3.561** (note that these thresholds are lower than that used for conservation).

Note: we have also developed a naïve Bayes classifier that significantly improves the accuracy of the combined verdict from the various in-silico prediction tools, however this feature is not currently enabled.

PP5

Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation. (Pathogenic, Supporting)

BP6

Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation. (Benign, Supporting)

Similarly to rules PS3 and BS3, these two rules leverage the known pathogenic database to report whether the variant has been clinically reported (see clinical evidence, but without any reference to in-vitro or functional studies).

The default strength for these rules is **Supporting**, per ACMG Guidelines, however our implementation will use stronger rule strengths if borne out by the available. Whilst this is not strictly in-line with guidelines, it does allow us to highlight clinical evidence, and users are always free to manually change the strength used when reviewing the verdict.

Strength **Supporting** is used by default, but for the following exceptions:

- ClinVar
  - **Very Strong** if 'practice guideline' = 4 stars, or 'reviewed by expert panel' = 3 stars,
  - **Moderate** if consistent submissions from multiple sources = 2 stars,
- VarSome user entries
  - **Very Strong** if more than 3 VarSome users have linked publications and classified the variant,
  - **Strong** if only 2 publications linked by users,
  - **Moderate** if only 1 publication linked by users,

In order to avoid double-counting, these rules will not be evaluated if rules PP3 or BS3 have triggered.

BA1

Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium. (Benign, Stand Alone)

Rule BA1 is applied if the allele frequency is greater than the threshold **0.05**. This is in strict concordance with the ACMG Guidelines and determines a variant to be stand-alone benign for Mendelian disease.

The BA1 Exceptions have also been implemented, as recommended by ClinGen.

Note that rules BS1 and BS2 may trigger at much lower frequency thresholds.

BS1

Allele frequency is greater than expected for disorder. (Benign, Strong)

Here we find the highest GnomAD allele frequency for the variant across the main population ethnicities and compare this to the benign cut-off frequency derived from the gene statistics. If there are too few known variants (fewer than **5**), we use a much higher default threshold, **0.015**, for rare diseases.

In order to avoid double-counting, rule BS1 is not evaluated if either rules BA1 or PM2 were triggered first.

BP3

In-frame deletions/insertions in a repetitive region without a known function. (Benign, Supporting)

This rule is closely related to PM4: it uses UniProt to ensure the variant isn't in a domain with a known function, checks that the variant is indeed in a repeat region, and verifies that there are no known clinically reported pathogenic variants within **3** base-pairs of the repeat region under consideration.

BP7

A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved. (Benign, Supporting)

This rule applies to synonymous variants only that are not deemed highly-conserved in GERP (see conservation).

Splicing is checked as follows:

- the variant is found further than **3** bases from a splice site, and there is no scSNV splice-site prediction
- not predicted splicing using scSNV (see splice-site prediction for more information).

Unimplemented Rules

The following rules are not implemented or not currently available to VarSome users - in most cases this is because the necessary data required to evaluate the rules is not in the public-domain, or the rules require patient-specific information, sometimes on a per-variant basis. Should they have more evidence, users can manually toggle rules on or off in VarSome, or adjust the strength used, and the resulting classification will be re-evaluated immediately.

PS2

De novo (both maternity and paternity confirmed) in a patient with the disease and no family history. (Pathogenic, Strong)

We have a prototype implementation, but is not integrated into VarSome.

PS4

The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls. (Pathogenic, Strong)

This rule has not been implemented.

PM3

For recessive disorders, detected in trans with a pathogenic variant (Pathogenic, Moderate)

This rule has not been implemented.

PM6

Assumed de novo, but without confirmation of paternity and maternity. (Pathogenic, Moderate)

We have a prototype implementation, but is not integrated into VarSome.

PP1

Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease. (Pathogenic, Supporting)

We have a prototype implementation, but is not integrated into VarSome.

PP4

Patient's phenotype or family history is highly specific for a disease with a single genetic etiology. (Pathogenic, Supporting)

This rule has not been implemented.

BS4

Lack of segregation in affected members of a family. (Benign, Strong)

We have a prototype implementation, but is not integrated into VarSome.

BP2

Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern. (Benign, Supporting)

This rule has not been implemented.

BP5

Variant found in a case with an alternate molecular basis for disease. (Benign, Supporting)

This rule has not been implemented.