**First program**

Input: One variant, either HGVS g., c., or p.

Output: Precalculatable ACMG evidence scores with reason for score assignment

**Second program**

Input: All ACMG evidence scores (including those manually determined from literature sources)

Output: Final classification (Benign, Likely Benign, VOUS, Likely Pathogenic, Pathogenic)

**Nomenclature**

Preprocess with HGVS package and desired final transcript from MGTL to validate and interconvert g. <-> c. <-> p.

**Population Data**

**Data sources:**

gnomAD variants, (add UK10K, HGVD, CHARGE, ESP6500, Qatar)

gnomAD coverage (+others…)

Gene information table

Non-MVL reputable pathogenic variant reports (ClinVar diagnostic labs (with patients) and manual literature review entries)

**Algorithm:**

Intervar:

Use variant’s highest sub-population MAF.

>5%: BA1

1-5%: BS1

<1%: PM2

PS4 not used

SOP:

“MAF is >=5%, this is considered very strong evidence that the variant is benign… If the variant is <5% but greater than >1%, then a likely benign classification should be considered”

Use variant’s highest sub-population MAF with AC > 2000

Funda:

“No MAF information due to poor sequence quality (<20X)”: skip

Average coverage at position (< 20X): skip (0)

Verify reason behind this and clarify my interpretation… would avg % of reads supporting the variant make more sense?

Can we assume a fair disease allele frequency cutoff is ~0.1% more than most frequent reputable pathogenic variant? (assumes most common disease alleles are already discovered)

For recessive and XL sq root of disease prevalence is disease allele frequency cutoff.

For dominant disease prevalence is disease allele frequency cutoff.

>10 times higher than disease allele frequency: BA1 (-3)

3-10 times higher than disease allele frequency: BS1 (-2)

1-3 times higher than disease allele frequency: BS1 (-1)

<1 disease allele frequency is consistent with a plausible pathogenic variant: PM2 (+0.5)

Should we blacklist certain diseases due to inclusion in gnomAD? Or assemble db of only healthy adults?

Observed in gnomAD for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder AND not adult-onset: BS2 (-0.5)

If (PM2) and number of unrelated patients reported with the variant as pathogenic in non-MVL reputable pathogenic variant reports…

= 2-4: PS4 (+0.5)

>= 5: PS4 (+1)

Zach: Use confirmed SOP-like cutoff until disease allele frequency info is available

**Other variant database**

**Data sources:**

Non-MVL reputable pathogenic variant reports (ClinVar diagnostic labs (with patients) and manual literature review entries)

**Algorithm:**

Pathogenic: PP5 (+0.5)

Benign: BP6 (-0.5)

Intervar:

ClinVar

MGTL:

Non-MVL reputable pathogenic variant reports

Zach:

Non-MVL reputable pathogenic variant reports

**Predictions and computational evidence for deleteriousness/splicing**

**Data sources:**

Gene information table

Reputable pathogenic variant reports (ClinVar diagnostic labs (with patients), MVL, and manual literature review entries)

All pathogenic variant reports (ClinVar, MVL, HGMD, LSDBs, EmoryDB)

Refgene table for lengths and boundaries of exons (Opt: Uniprot or protein sequence db)

Missense predictors (Alamut and dbNSFP)

Splicing predictors (Alamut and dbscSNV)

MultiZ 100 vertebrate alignment (Opt: Conservation predictors)

RepeatMasker

Amino acid substitution matrix (grantham / blossom)

**dbNSFP:**

Xiaoming Liu, Ph.D.

Assistant Professor,

The University of Texas Health Science Center at Houston

20 prediction algorithms (SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster2, MutationAssessor, FATHMM, MetaSVM, MetaLR, CADD, VEST3, PROVEAN, FATHMM-MKL coding, fitCons, DANN, GenoCanyon, Eigen coding, Eigen-PC, M-CAP, REVEL, MutPred)

6 conservation scores (PhyloP x 2, phastCons x 2, GERP++ and SiPhy)

Two branches of dbNSFP v3.5 are provided: dbNSFP3.5a suitable for academic use, which includes all the resources, and *dbNSFP3.5c suitable for commercial use, which does not include Polyphen2, VEST3, REVEL, CADD and DANN*

dbscSNV includes all potential human SNVs within splicing consensus regions (−3 to +8 at the 5’ splice site and −12 to +2 at the 3’ splice site) and two ensemble prediction scores for predicting their potential of altering splicing (ada\_score, rf\_score)

MGTL:

Alamut (SSF, NNsplice, MES, GS, HSF), SIFT, PolyPhen2, MutationTaster2 –**individual methods**

Intervar:

metaSVM, metaLR, dbscSNV (ada\_score, rf\_score) –**ensemble predictors**

SIFT, GERP++

**Nonsense and Frameshift**

**Algorithm:**

Intervar:

“Major” isoform (designated by knownGene)

LOF disease mechanism (2000 genes with pathogenic LOF alleles, 3000 HI genes from ExAC)

NMD (New stop codon not in last exon, nor 50bp at 3’ end of penultimate exon)

: PVS1

MGTL:

Funda: inconsistent with SOP

LOF disease mechanism

Not alternatively spliced exon OR reputable pathogenic variants in same exon

Pathogenic LOF variants have been reported 3’

: PVS1 (+2)

LOF disease mechanism

Alternatively spliced exon AND NO reputable pathogenic variants in same exon

OR Pathogenic LOF variants have NOT been reported 3’

No NMD (New stop codon in last exon, or 55bp at 3’ end of penultimate exon)

: PVS1 (+1)

Not LOF disease mechanism

: PVS1 (+0.5)

SOP: inconsistent with in-frame InDel > 15AA = PVS1

LOF disease mechanism

‘Relevant’ transcript

IF any 1

NMD (New stop codon NOT in last exon, or 55bp at 3’ end of penultimate exon)

>= 1 reputable pathogenic variants with PTC same codon (encompasses PS1), 3’, or 2% of transcript length 5’

>= 2 pathogenic variants with PTC same codon (encompasses PS1), 3’, or 2% of transcript length 5’

> 10% of transcript is missing/frameshifted

> 5% extension of protein length

: PVS1

Zach: Ask curators (variant cards inconsistent)

*ACMG paper: “PVS1… complete absence of the gene product by lack of transcription or nonsense-mediated decay”*

Above being the ideal definition but later implied to potentially tolerate existing protein product but with *negligible functionality*

*“[CSS variants could lead to an…] in-frame deletion/insertion, which could retain the critical domains of the protein and lead to a minor length change (PM4)”*

Gray area being what is ‘minor’ and should there be outlet for PM4…

LOF disease mechanism

Not alternatively spliced exon OR pathogenic variants reported same exon OR relevant transcript

IF any 1

NMD (New stop codon NOT in last exon, or 55bp at 3’ end of penultimate exon)

>= 1 reputable pathogenic variants with PTC same codon (encompasses PS1), 3’, or 2% of transcript length 5’

>= 2 pathogenic variants with PTC same codon (encompasses PS1), 3’, or 2% of transcript length 5’

> 10% of transcript is missing/frameshifted

> 5% extension of protein length

net change >= 15 amino acids (15 AA for PVS1 seems not very stringent)

: PVS1

ELSE:

PM4

Additional info to display (separate from calculator):

Quantify portion of protein affected (% protein intact, % protein frameshifted, # AA WT protein)

**Canonical Splice Site (CSS)**

**Algorithm:** Assumes exon-skipping consequence

Intervar:

“Major” isoform (designated by knownGene)

LOF disease mechanism (2000 genes with pathogenic LOF alleles, 3000 HI genes from ExAC)

NMD (New stop codon not in last exon, nor 50bp at 3’ end of penultimate exon)

: PVS1

MGTL:

Funda:

LOF disease mechanism

Not alternatively spliced exon

NMD (New stop codon not in last exon, nor 55bp at 3’ end of penultimate exon)

Pathogenic LOF variants have been reported 3’

: PVS1 (+2)

LOF disease mechanism

Alternatively spliced exon

**In-frame exon**

OR Pathogenic LOF variants have NOT been reported 3’

No NMD (New stop codon in last exon, or 55bp at 3’ end of penultimate exon)

: PVS1 (+1)

Not LOF disease mechanism

: PVS1 (+0.5)

SOP: rather simple since they use single-exon deletion = PVS1 which is the minimal expected consequence

*“For canonical splice variants, do not use the 10% rule or 2% rule… these variants have protein effects that may be difficult to predict (ie, entire exon skipping)”*

*“PVS1 (…single or multi-exon deletion…)”*

LOF disease mechanism

‘Relevant’ transcript

No high AF variants within the same dinucleotide or its pair

: PVS1

Zach:

IF single-exon deletion = PVS1:

LOF disease mechanism

Not alternatively spliced exon OR pathogenic variants reported same exon OR relevant transcript

: PVS1

ELSE:

LOF disease mechanism

Not alternatively spliced exon OR pathogenic variants reported same exon OR relevant transcript

IF any 1

NMD (New stop codon NOT in last exon, or 55bp at 3’ end of penultimate exon)

>= 1 reputable pathogenic variants with PTC same codon (encompasses PS1), 3’, or 2% of transcript length 5’

>= 2 pathogenic variants with PTC same codon (encompasses PS1), 3’, or 2% of transcript length 5’

> 10% of transcript is missing/frameshifted

> 5% extension of protein length

net change >= 15 amino acids (15 AA for PVS1 seems not very stringent)

: PVS1

ELSE:

PM4

Additional info to display (separate from calculator):

* High AF variants within the same canonical splice site and its pair
* “Donor/acceptor of exon X/total”
  + Flag if last acceptor site (“Variants in the canonical splice acceptor site of the last exon are usually considered as VS1 in our laboratory due to the expectation that the poly-A tail will not be present, leading to an unstable transcript”)
  + Flag if first donor (for intron retention consideration)
* dbscSNV scores and Alamut window
* Quantify portion of protein affected (% protein intact, % protein frameshifted, # AA WT protein)
* If NMD is expected

**Start-loss**

**Algorithm:**

Intervar: None

Funda: None

*SOP: Does not specifically suggest corresponding ACMGs*

*LOF disease mechanism*

*Reported start-loss variants*

*# transcripts not using this Met*

*Next Met in same exon*

*Distance/percent of transcript to next Met?*

*Are there pathogenic variants in the region lost?*

*Kozak consensus sequence for next Met? (gcc)gccRcc****AUG****G (I don’t think this is worth checking, literally first two genes I checked don’t follow it (CLUAP1, TTN)*

*Evidence that next Met is naturally used as a start codon? (RNA evidence)*

Zach:

LOF disease mechanism

Not alternatively spliced exon OR pathogenic variants reported same exon OR relevant transcript

OR >10% protein affected

Pathogenic missense variants reported between next Met

: PVS1

<10% AND no pathogenic missense variants reported between next Met: PM4

Other start-loss variants in the same codon reported as pathogenic: PS1

Additional info to display (separate from calculator):

Kozak consensus matching (or atleast logos plot from wikipedia)

High AF start-loss variant

**Stop-loss**

**Algorithm:**

Intervar:

PM4

MGTL:

Funda:

PM4 (+1.5)

SOP:

> 5% extension of protein length: PVS1

Janet: PM4

Zach:

> 5% extension of protein length and >= 15 amino acids: PVS1

Else: PM4

Other stop-loss variants in the same codon reported as pathogenic: PS1

Additional info to display (separate from calculator):

High AF stop-loss variants

If Janet: size of protein affected

Number of protein-extending variants and type

**In-frame InDel**

**Algorithm:**

Intervar:

If not in repeat by repeatmasker: PM4

MGTL: significant differences

SOP:

Not in repeat

Net change >= 15 amino acids

: PVS1

Funda:

Large or drastically changes protein length: PM4 (+1.5)

Small in-frame InDel with an impact on protein function: PM4 (+1)

Small in-frame InDel with an uncertain impact OR is within a repeat: BP3 (-0.5)

Janet: PM4

Zach: Ask curators/compare variant cards (15 AA for PVS1 seems not very stringent)

Not in repeat by repeatmasker

Net change >= 15 amino acids: PVS1

Net change < 15 amino acids: PM4

In repeat by repeatmasker: BP3

Additional info to display (separate from calculator):

Size of protein being affected

Report pathogenic variants within the range of deleted AAs

High AF in-frame InDels within 15bp

**Nonsynonymous**

**Algorithm:**

Intervar:

metaSVM > 0

GERP++\_RS > 2

OR

dbscSNV ada\_score OR rf\_score >= 0.6

: PP3

metaSVM < 0

GERP++\_RS < 2

dbscSNV ada\_score < 0.6

dbscSNV rf\_score < 0.6

: BP4

Intervar’s two-fold problem being it ignores variant type and ignores missing data

Intervar PP3 example problems:  
44 synonymous SNV PP3=1 BP4=0 GERP++\_RS=4.2 metaSVM=1.061 ada\_score=. rf\_score=.

52 stopgain SNV PP3=1 BP4=0 GERP++\_RS=5.38 metaSVM=. ada\_score=. rf\_score=.

Intervar BP4 example problem:

166  stopgain       SNV  PP3=0  BP4=1  GERP++\_RS=-1.72   metaSVM=-1.005  ada\_score=.       rf\_score=.

Same amino acid substitution as ClinVar pathogenic variant

dbscSNV ada\_score AND rf\_score < 0.6

: PS1

Same amino acid affected as ClinVar pathogenic variant but different substitution

dbscSNV ada\_score AND rf\_score < 0.6

: PM5

Variant in interpro protein domain which contains only pathogenic or likely pathogenic variants (exact number of patho requirement unclear, 1?) with no ClinVar benign or AF > 5% variants: PM1

>80% ClinVar pathogenic variants are missense and <10% of all missense are benign: PP2

>80% of ClinVar pathogenic variants are LOF: BP1

MGTL:

Funda:

3/5 Alamut splice-affecting: PP3 (+0.5)

Position = first three exonic bp of donor, first exonic bp of acceptor: PP3 (+0.5)

SIFT, PolyPhen2, MutationTaster = deleterious

AA absolutely conserved

: PP3 (+0.5)

SIFT, PolyPhen2, MutationTaster = tolerated: BP4 (-0.5)

AA is reference in >= 3 mammals: BP4 (-0.5)

Same amino acid substitution as pathogenic variant: PS1 (+2)

Same amino acid affected as pathogenic variant but different substitution: PM5 (+0.5)

Funda lacks details for these three

Variant is in mutational hotspot or well-established functional domain lacking benign variants: PM1 (+1)

Missense variant in a gene with low rate of benign missense variation and missense pathogenic variants are common (“Check gene-specific variant spectrum”): PP2 (+0.5)

Missense variant in a gene in which only truncating variants are known to cause disease (“Check gene-specific variant spectrum”): BP1 (-0.5)

SOP: fairly vague as far as ACMG

ExAC missense constraint = significantly constrained: PP2

>= 3 pathogenic missense variants within 5 amino acids: PM1

Zach:

*ACMG paper: “The variant amino acid change being present in multiple nonhuman mammalian species in an otherwise well-conserved region, suggesting the amino acid change would not compromise function,* ***can be considered strong evidence for a benign interpretation****”* clear example where a manual override to benign is recommended which doesn’t necessarily fit into guideline scoring

PP3 BP4 *“Ex: missense/splicing/conservation preds must be 100% concordant”*

PM5 [missense where same AA affected as patho]

PS1 [same (ex: missense) AA sub as patho]

dbscSNV ada\_score > 0.97: PP3

Use SIFT, PolyPhen2, MutationTaster for v1 preds, after performance assessment recommend update.

SIFT, PolyPhen2, MutationTaster = deleterious: PP3

AA absolutely conserved in mammals: PP3

SIFT, PolyPhen2, MutationTaster = tolerated: BP4

Variant amino acid is reference in >= 3 mammals: BP4

Same amino acid substitution as pathogenic variant: PS1

Same amino acid affected as pathogenic variant and substitution is similar or more biochemically different: PM5 (need to decide AA sub matrix and cutoffs [grantham/blossum])

>= 3 pathogenic missense variants within 5 amino acids

OR

Present in manually determined hotspot functional domain from gene information table (GlyXY motif, known disulfide bridge, phosphorylation site, .bed interval of critical residues) and no BEN or high AF variants are present with a AA sub matrix worse than the

: PM1

Gene’s valid disease variant type is “GOF(missense only)”: PP2

I’m hesitant to encourage the use of PP2 based on anything but manually reviewed GOF genes, see variant spectrums for genes given as example by ACMG below with ExAC constraints

>80% of pathogenic variants are LOF (would like to refine this by checking ALMS1 and ASPM spectrums)

OR

Gene’s valid disease variant type is “LOF-only”

: BP1

Additional info to display (separate from calculator):

Alamut scores / window

rf\_score

All variants used for PM5 and PS1

Expanded MSA

High AF nonsynonymous variants affecting same codon

**Synonymous or intronic (non-canonical splicing consensus)**

**Algorithm:**

Intervar:

metaSVM > 0

GERP++\_RS > 2

: PP3

dbscSNV ada\_score OR rf\_score >= 0.6: PP3

dbscSNV ada\_score AND rf\_score < 0.6

Nucleotide GERP++ < 2

: BP7

MGTL:

3/5 Alamut splice-affecting: PP3 (+0.5)

3/5 Alamut not splice-affecting: BP7 (-1.5)

Zach:

dbscSNV ada\_score > 0.97: PP3

dbscSNV ada\_score between 0.6

dbscSNV ada\_score < 0.97: BP7

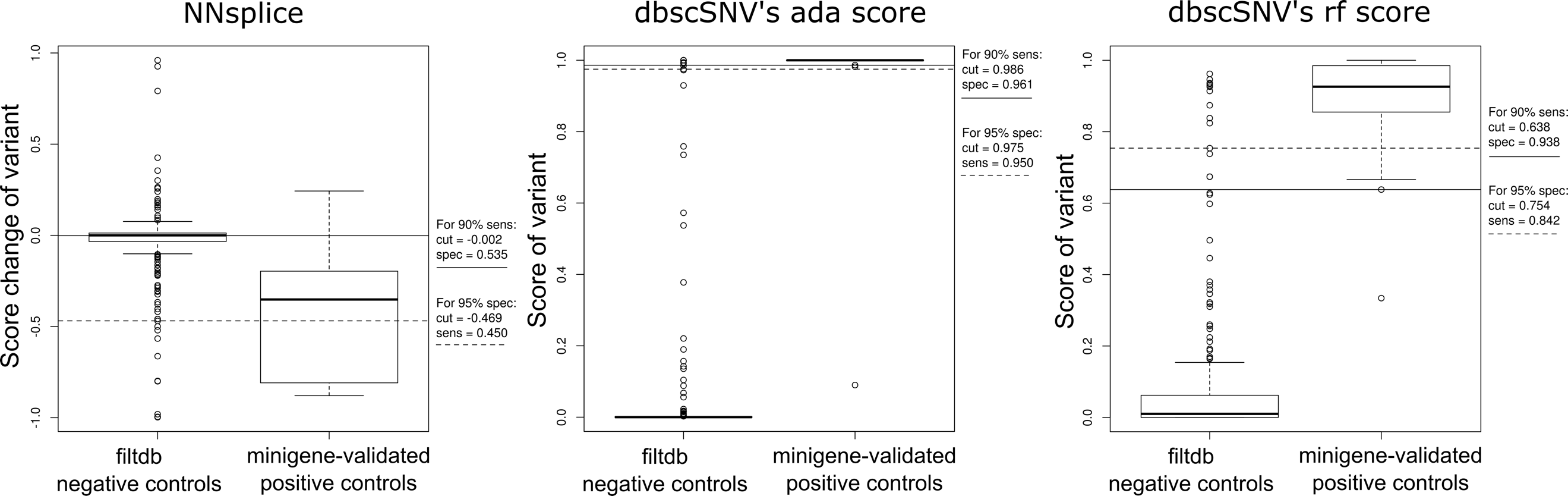
Additional info to display (separate from calculator):

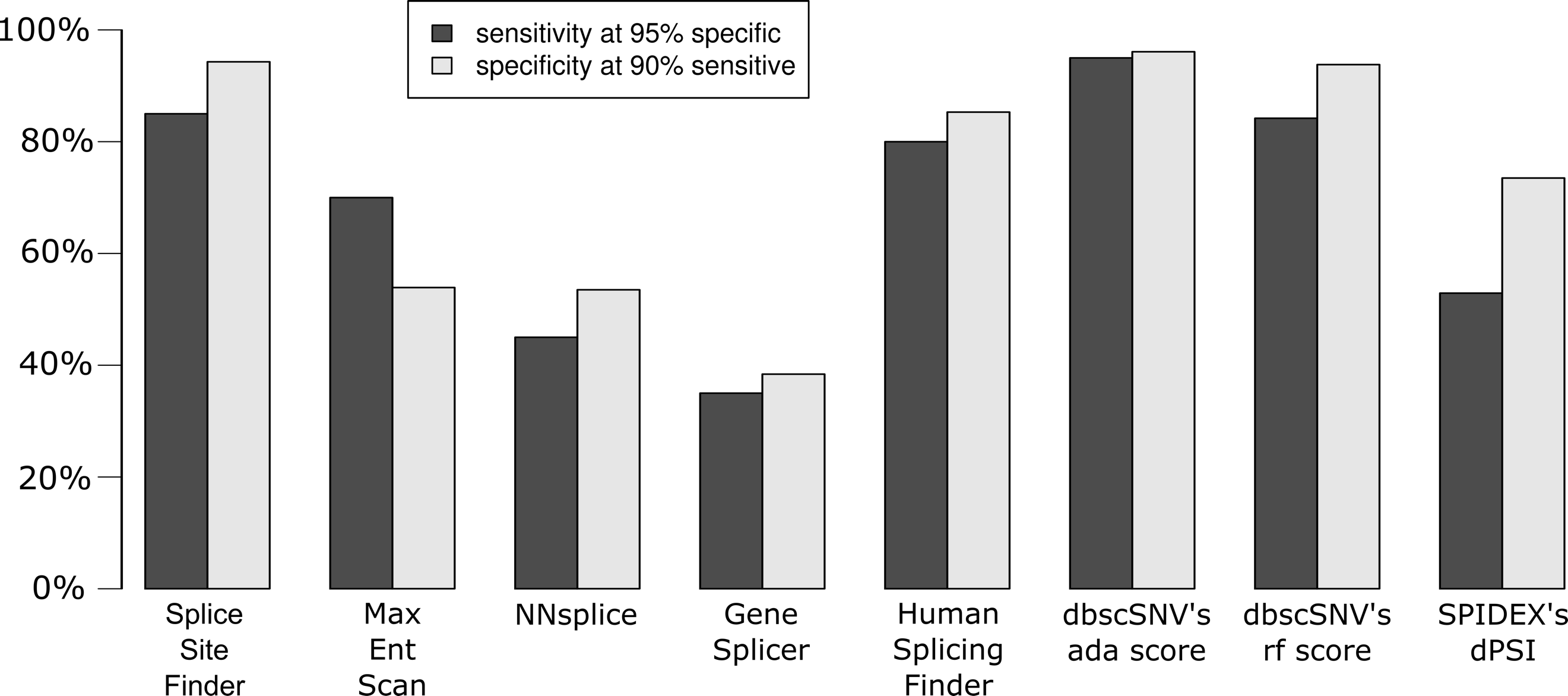
Alamut scores / window

rf\_score

Conservation MSA

**dbscSNV score cutoff reasoning**

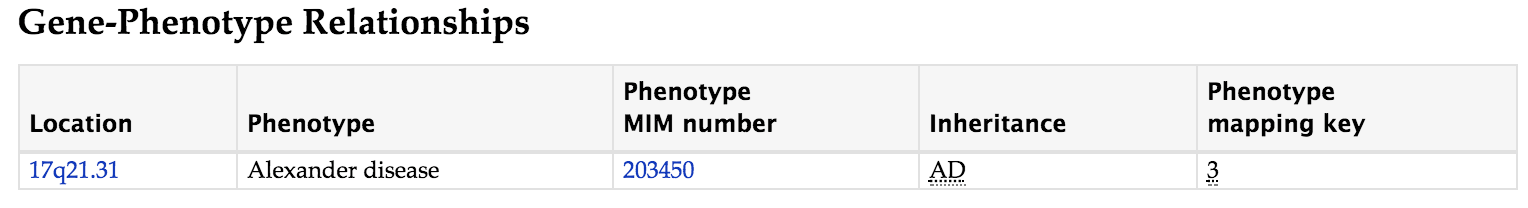
****

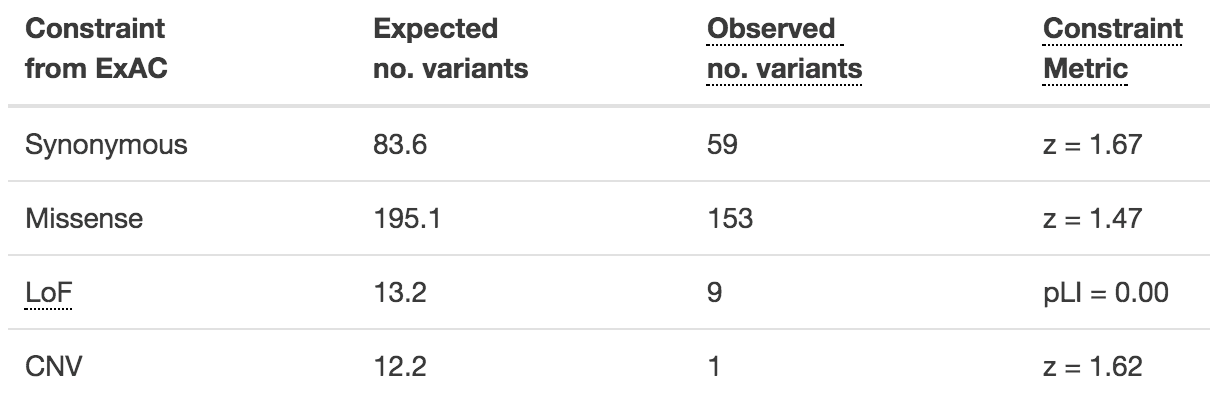
****

Alamut’s predictors

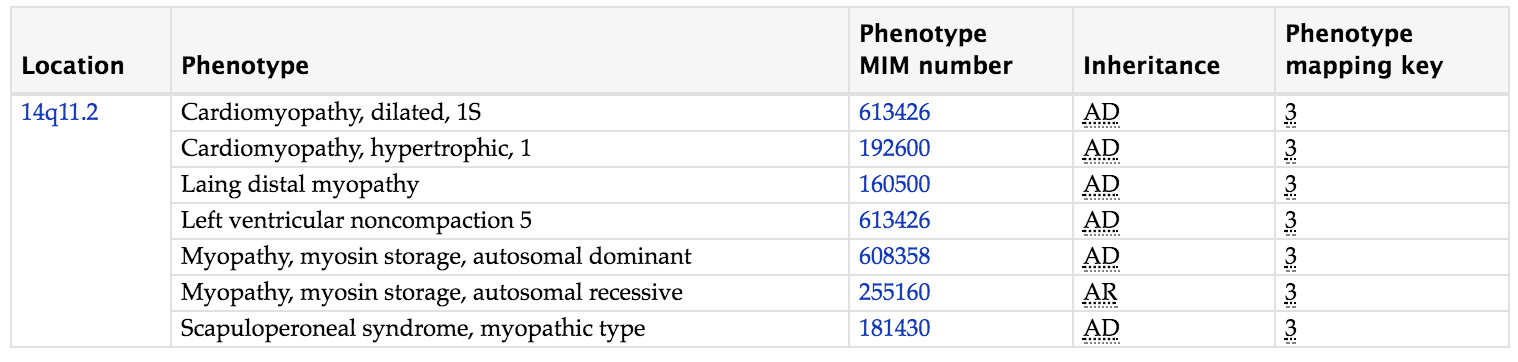
**Why using PP2 is questionable:**

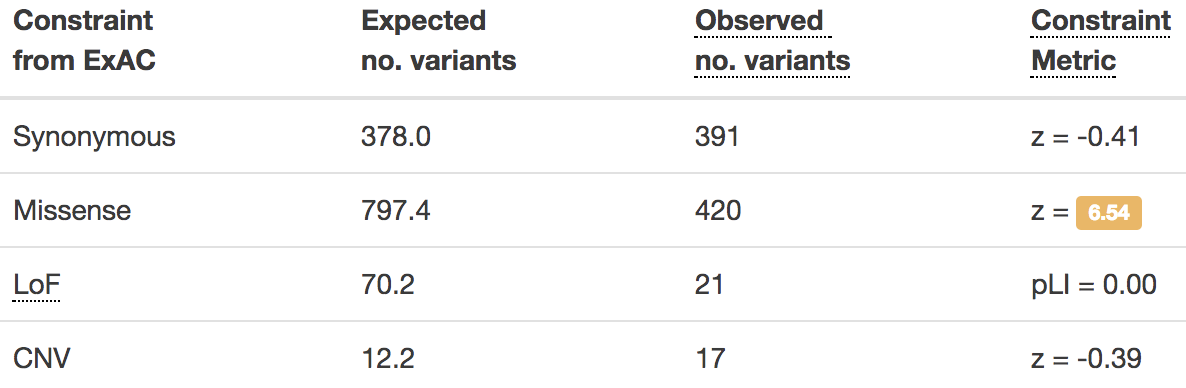
**GFAP – fairly established GOF disease mechanism:**

****

****

**MYH7 – mechanism remains largely unknown but constraints imply GOF:**

****

****

Need to verify reputable diagnostic labs to search for patient reported variants

**>90% of entries from top 30 submitters**

**Reject**

**CLIA certified**

**TBD**

|  |  |
| --- | --- |
| Illumina Clinical Services Laboratory,Illumina | 94375 |
| GeneDx | 31467 |
| Emory Genetics Laboratory,Emory University | 27974 |
| Invitae | 22043 |
| OMIM | 21020 |
| Ambry Genetics | 19239 |
| Laboratory for Molecular Medicine,Partners HealthCare Personalized Medicine | 16193 |
| PreventionGenetics,PreventionGenetics | 15707 |
| Genetic Services Laboratory, University of Chicago | 10257 |
| Counsyl | 5670 |
| GeneReviews | 5253 |
| Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) | 4810 |
| Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA), c/o University of Cambridge | 3006 |
| ITMI | 2686 |
| Sharing Clinical Reports Project (SCRP) | 2231 |
| Biesecker Lab/Human Development Section,National Institutes of Health | 2214 |
| InSiGHT | 2105 |
| Breast Cancer Information Core (BIC) (BRCA2) | 1872 |
| Breast Cancer Information Core (BIC) (BRCA1) | 1653 |
| Division of Genomic Diagnostics,The Children's Hospital of Philadelphia | 1638 |
| Tuberous sclerosis database (TSC2) | 1559 |
| Retina International | 1524 |
| LabCorp | 1520 |
| Cardiovascular Biomedical Research Unit,Royal Brompton & Harefield NHS Foundation Trust | 1511 |
| LDLR-LOVD, British Heart Foundation | 1394 |
| ARUP Institute,ARUP Laboratories | 1220 |
| Blueprint Genetics | 1195 |
| Epithelial Biology, Institute of Medical Biology, Singapore | 1078 |
| Center for Pediatric Genomic Medicine,Children's Mercy Hospital and Clinics | 1053 |
| Systems Biology Platform Zhejiang California International NanoSystems Institute | 1024 |

**Closing summary of remaining questions, comments, ideas**

-Need table of genes with **disease allele frequency cutoff, inheritance pattern, age of onset, if dominant the penetrance, and valid disease variant types GOF(missense only) vs Reduction-in-Function(LOF+Missense) vs LOF-only gene list, hotspot mutational regions**

This is relevant to multiple ACMG scores including population database, PVS1, PP2

Book with carrier testing panel info should have carrier frequencies for disease allele frequency cutoffs

Questions:

1. What is the exact requirements for the utilization of pathogenic variants from non-MVL sources?

Rong has suggested I follow Ashley’s definition since if curators feel they should manually review a variant to trust it, this variant will then be entered into the MVL and therefore trusted anyway. But might require a refresh/rerun of the precalc suggestions to update the suggestion with the now reviewed variant

1. What proportion of the protein being deleted is satisfactory for PVS1 versus PM4 (10% rule, 15AA rule, single exon deletion)? (Will affect nonsense, frameshift, CSS, nonframeshift, startloss, stoploss and overlaps with the desired functionality of having modifiers for the ACMG scores)
2. Can we apply PS1 to all variants?
3. What is the preferred method for determining the importance of the exon affected by the variant (and should this be noted for EVERY variant type including missense?):
   1. Pathogenic variants in exon
   2. “Constitutive” exon? (In every transcript)?
   3. Not alternatively spliced? (not in every transcript, but in all isoforms transcribing the exon)?
   4. Majority of transcripts? (83% or at least 5/6 … 85% or more than 5/6,)?
   5. “Relevant” transcript

Rong has suggested I assume the “relevant” transcript model but provide majority of transcript type information

1. Should we display MVL VOUSs when looking at nearby variant reports to alert curators that both variants’ conclusions may be affected by the VOI and need review?
2. “No MAF information due to poor sequence quality (<20X)”: skip OR Average coverage at position (< 20X): skip (0)… Verify reason behind this and clarify my interpretation… would avg % of reads supporting the variant make more sense?
3. Can we assume a fair disease allele frequency cutoff is ~0.1% more than most frequent reputable pathogenic variant? (assumes most common disease alleles are already discovered)

Some things it appears curators would like to see for every variant to facilitate manual overrides/comments:

* Snapshot of UCSC GB indicating variant’s location on gene-level view with all RefGene transcripts
  + With position of predicted PTC
* MSA on protein-sequence/structure level showing AA and nucleotide differences for surrounding 10AAs
* If near an annotated splice site, show Alamut zoom-in of exon-intron boundary with splicing preds and nucleotide MSA
* All benign (“high” AF) and pathogenic variants within +/-15bp, and any MVL variants of the same variant type (missense+inframe indels // LOF variants)