Database analysis figures from chanty_GH comments_MP responses

BOLDcaution/7-vep-annotations

EmBold project

Data sources

- ClinVar 04/30/2023 release of variant_summary.txt
  - annotated by VEP

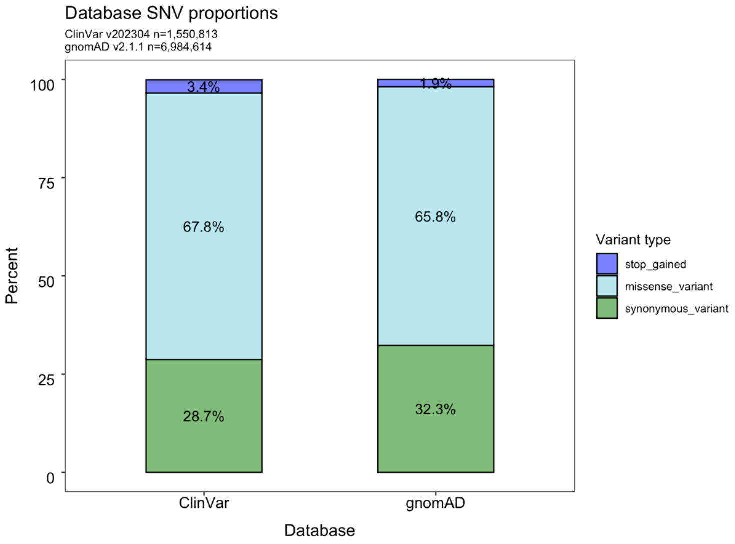
- gnomAD v2.1.1 WGS: whole genome sequencing exome call set data
  - v1 posted to chanty
  - slides include comments by GH

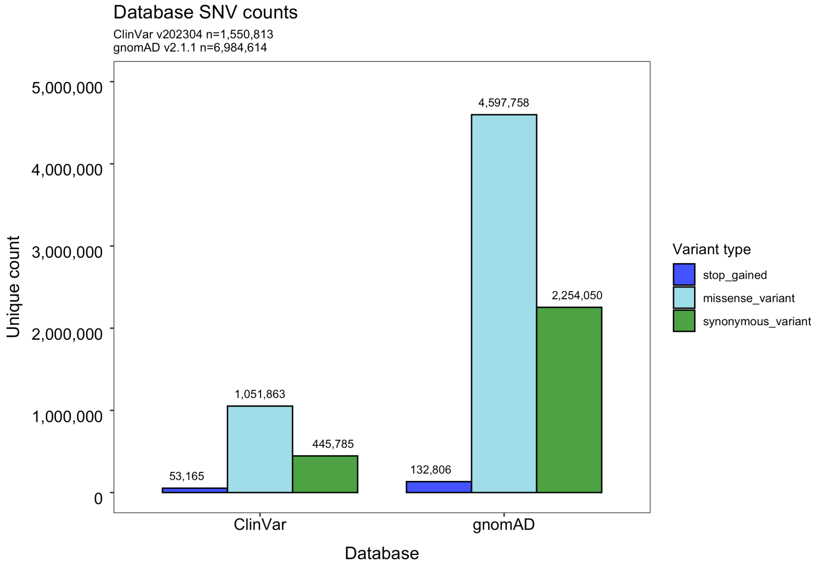
- gnomAD v2.1.1 WES: whole exome sequencing data
  - figures not shown

Combination of WGS & WES: summed AC, AN, nhomalt columns and recalculated AF

- whole genome sequencing exome call set data plus whole exome sequencing data
  - recalculated AN, AC, and AF available for:
    - all
    - afr
    - amr
    - eas
    - non_neuro_nfe
    - fin
    - asj
    - sas*

I first made figures using protein coding SNVs from gnomad v2.1.1 whole genome (WGS) data and recent clinvar release data. Im going to make the equivalent for combo of gnomad whole exome & whole genome seq but to test figure code I just went with smaller dataset (takes less time)..

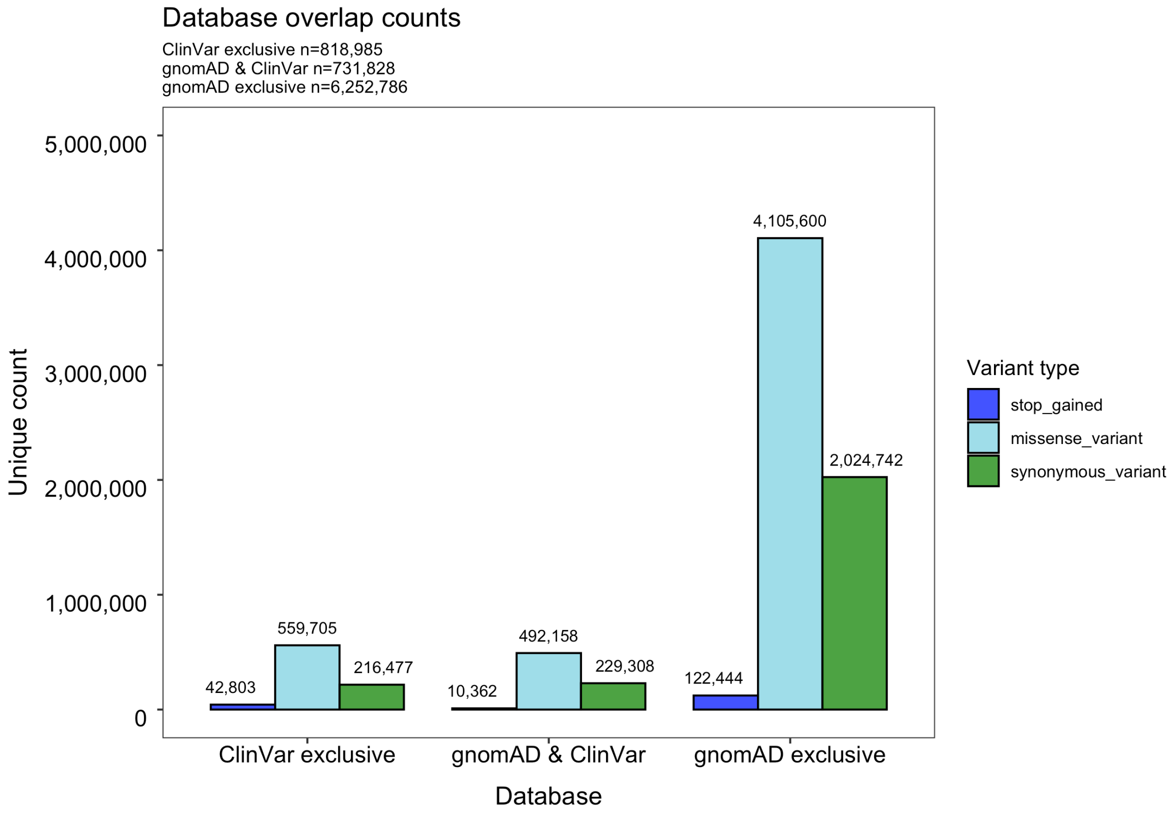




result 1- gnomad has higher % silent and lower % nonsense than clinvar.

result 2- same as result 1 data just with raw counts of the variant types for each database

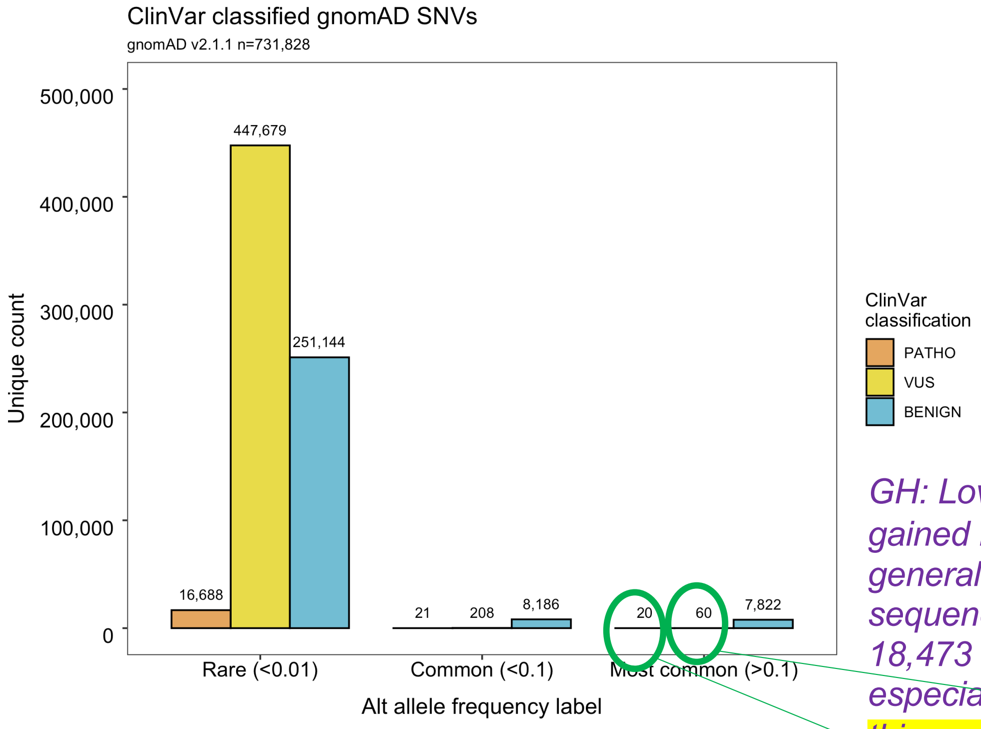
GH: The numbers make sense. If my analysis is correct, random mutation would predict ~26% of mutations to be silent, 70% missense and 4% stop-gained. That would be based on stochastic amino acid distribution, which we know is not the case. There could be an argument that missense and stop-gained mutations are "selected against", but it is probably in the noise.



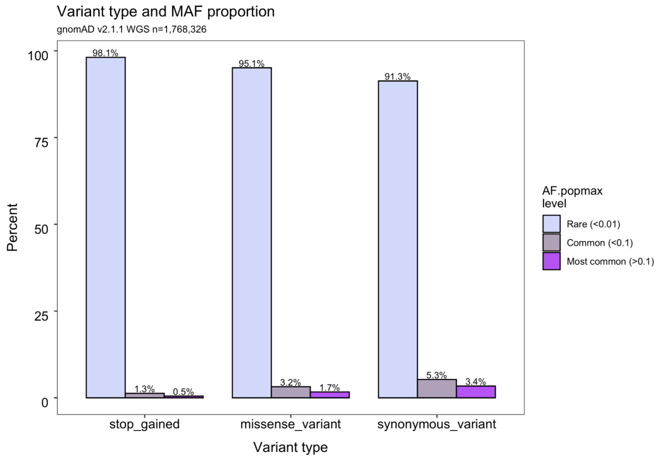
result 3- there is overlap between both databases. For prevalence analysis I have to use only the gnomAD & ClinVar variants and gnomAD exclusive variants because they have the allele freq columns needed for the calc.

GH: Very nice! The ratio of missense/synonymous is 2.5 for ClinVar and 1.7 for gnomAD. Given that both are biased, either for or against disease, a difference probably makes sense.

result 4- majority of all variant types are rare. very few stop gained variants are common. missense and synon have similar counts expect in rare category where missense are far more abundant.



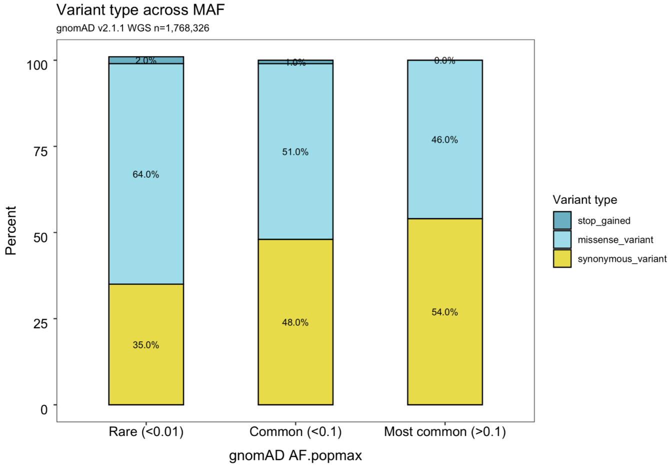
result 5- same data as result 4 but plotted so you can clearly see % of MAF levels represented in each variant type (x axis). synon variants are more likely to be common and be under less selective pressure than missense and stop gain in general.



GH: Love this analysis! Looking at the "most common" variants, 137 stop gained mutations are most common – one would expect these to be generally benign, either because the mutations are late in the codon sequence (C-term deletions), there are paralogs, or…. Similarly, for the 18,473 missense mutations in this group, many may be inconsequential, especially for major disease, but perhaps not for health, or health span. Is this our treasure trove? The diversity, i.e., sequence landscape this represents is humongous.

Can you summarize in which genes these mutations are? Maybe a table of the genes that we can then group by pathway?

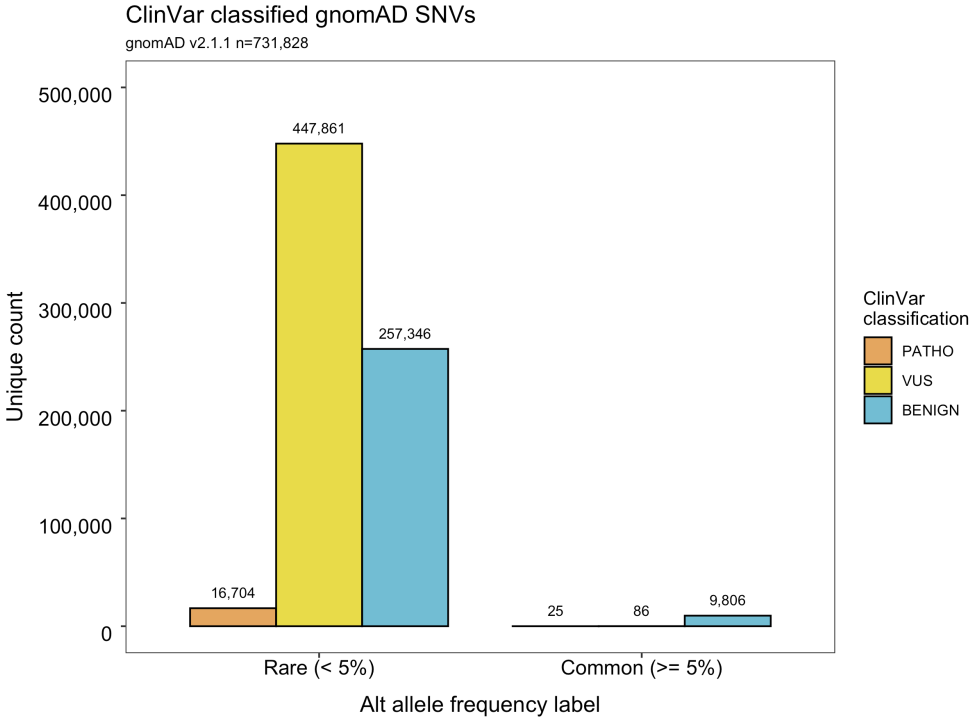
(could one plot a GWAS type output with all genes on the x-axis and MAF on the y-axis?)



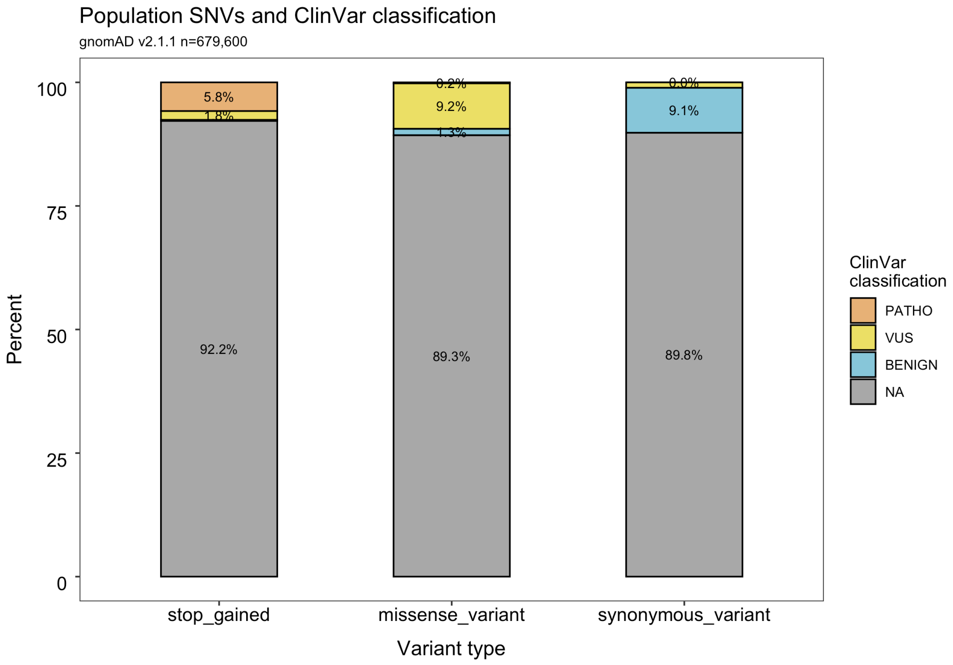
result 6- this should actually follow result 4. 46% of variants over 10% AF are missense.

GH: not sure what to make of the percentages (the absolute numbers on the previous slide are a good story – 18,000 is a lot and suggests that on average about every gene has significant variation for at least one allele).

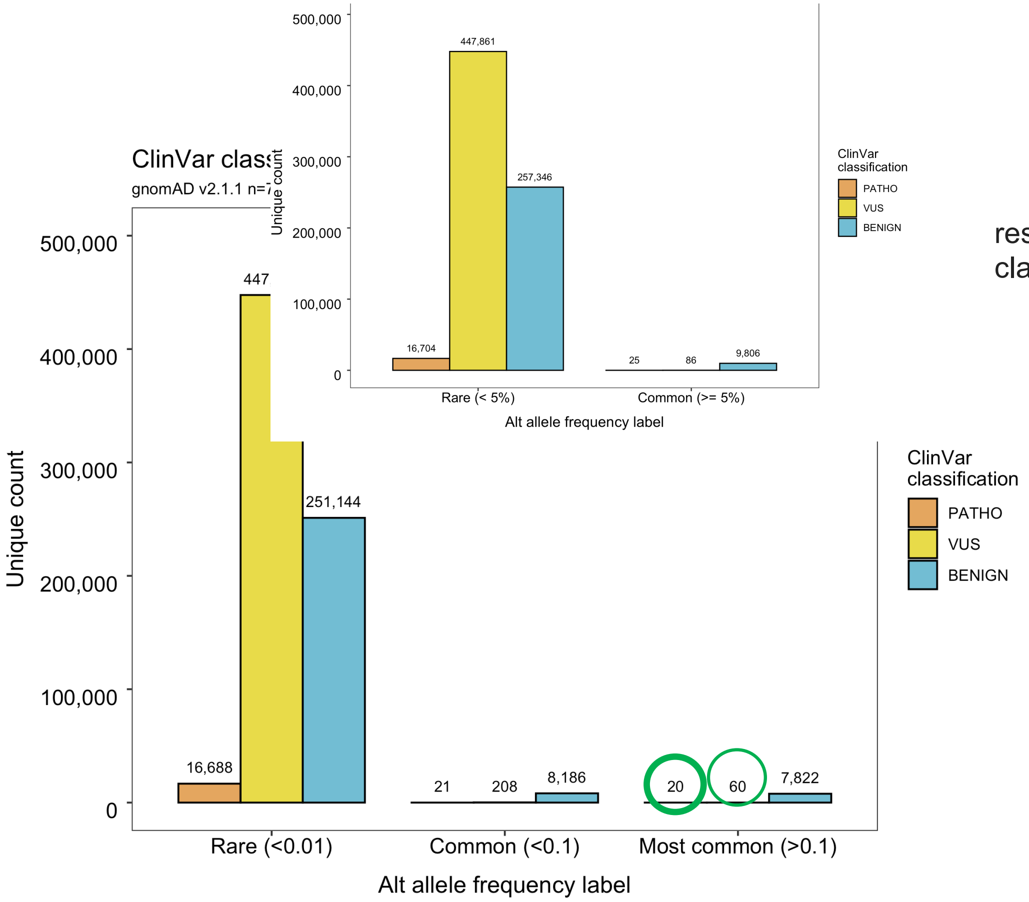
The way the numbers go further away from the expected ratio based on randomness, (4, 70, 26) suggests an evolutionary process.



result 7- first gap figure, shows % of gnomad variants classified in clinvar. large majority have no interpretation available. embold can support shrinking grey area.



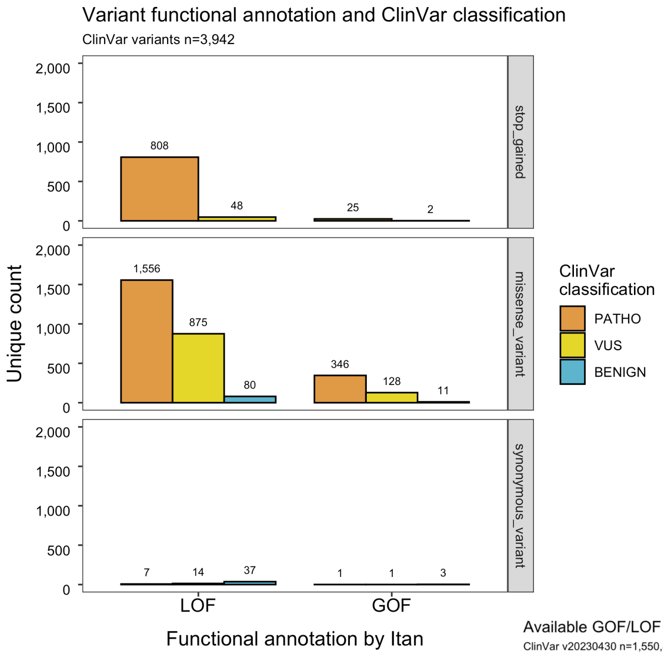
GH: the classification is based on considerable disease, likely based on symptoms seen early in life. We would close both the VUS and NA gap. At the same time, Benign mutations may not be benign from a health span perspective.



result 8- this is the non grey variants in result 7 raw counts. So even for classification of rare variants, most are VUS.

GH: This is also a very interesting group. These are carriers correct? Would this be a group of alleles that in heterozygous form do not have symptoms of disease, or would they be alleles where the homozygous form is only very rarely observed so that oftentimes there is no diseased progeny that would exclude parents from participating? If there is progeny, then such parents would be excluded, however, the would likely have some sort of compensatory mutation.

Most common PATHO or VUS clinvar variants (all SNV types)



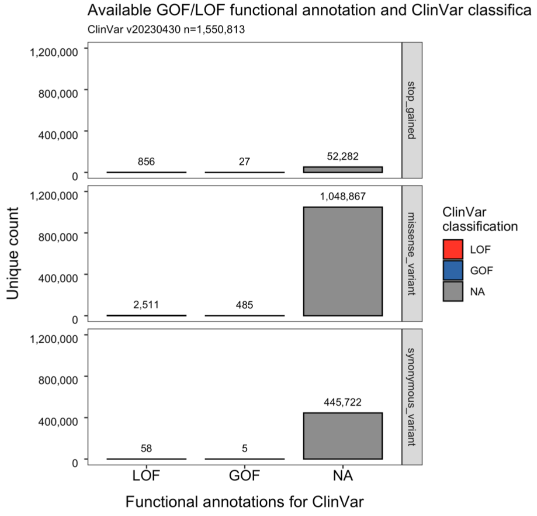
GH: is this Itan (x-axis)

result 9- this is the 4k GOF LOF variants from the largest curation available on SNV and mechanism. This is the second gap figure showing how little mechanism insight is available. Another area where i believe embold and screen platform can have impact since we will be looking into enzyme activity changes and maybe some cell viability (?) or other assay to inform possible Patho, vus, benign interpretation in human cell lines.
another point on result 9- stop gained has less uncertainty than missense, even when both have LOF mechanism annotation.

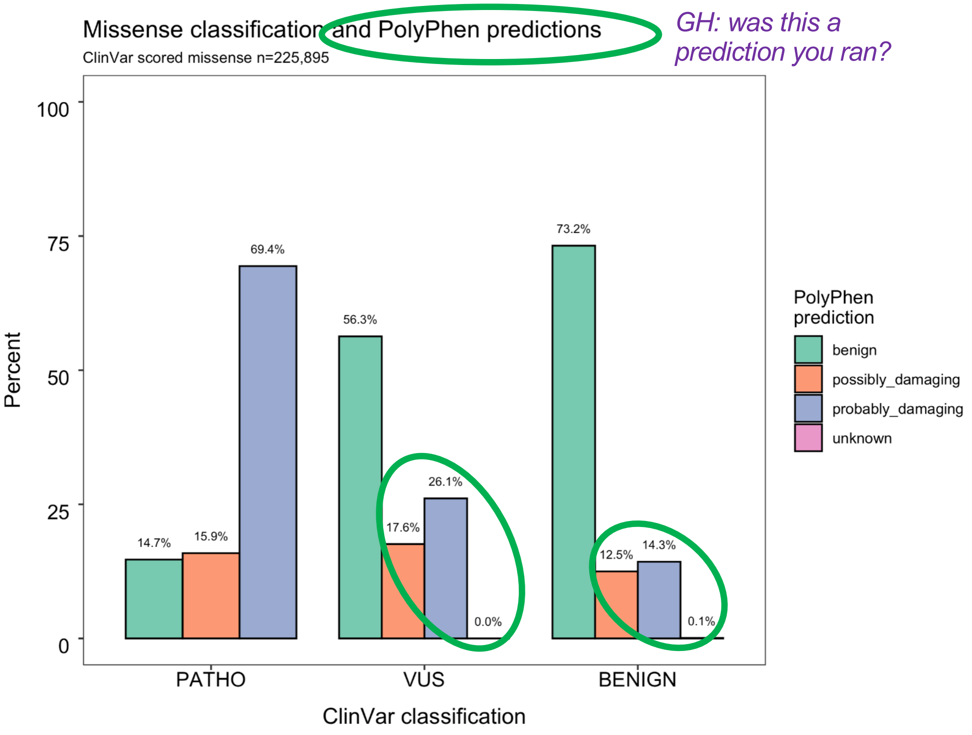
GH: I am missing a few pieces here. What is Itan? Are these rare, common, or most common mutations, or all of them?

I guess an underlying question is which type of mutation (rare, common, or most common) is most relevant for health span. I don't think we can answer it right now and with our process try to figure out where we should look. Most common feels like the simplest approach, and it could be wrong. That's OK for our first steps in this endeavor however and being able to say that based on data is all we need, for now.

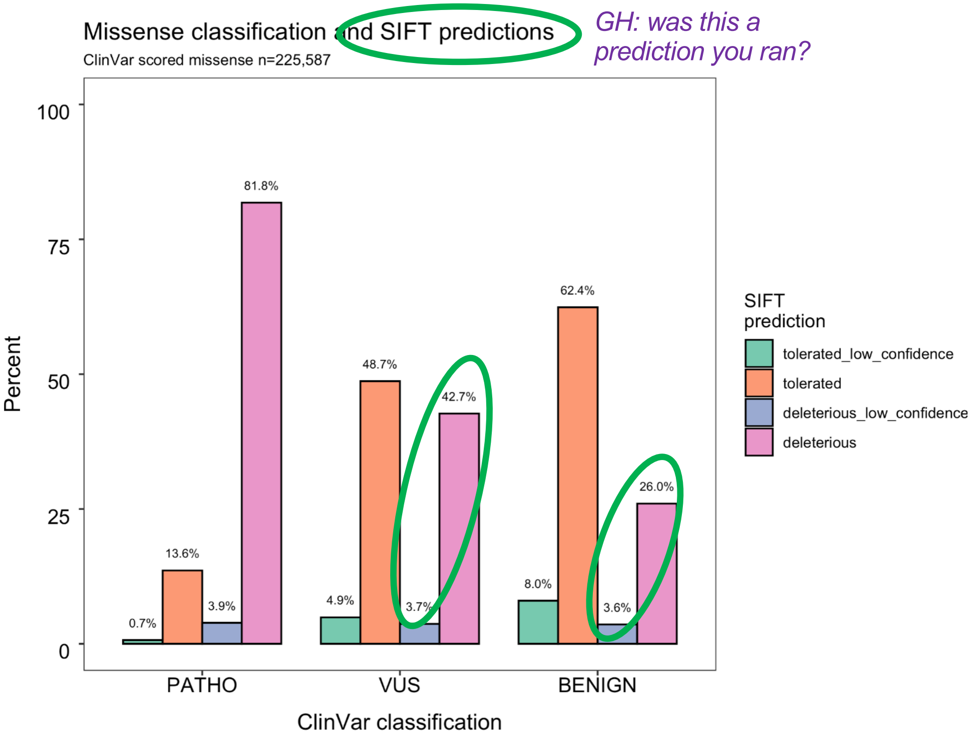
Nonetheless, there may be some interesting data in here, especially from a pathway perspective. Also, details of Pathogenic synonymous mutations would be good to understand.



result 10- to conclude this first phase of variant analysis, i highlight how clinvar classifications and missense pathogenicity scores disagree. I picked SIFT and PolyPhen, the two most widely used metrics, to show that scores are not perfect. Even for pathogenic classified mutations-most are predicted correctly- but still have benign / low confidence / tolerated predictions. basically just a point on predictions are not reliable enough for clinical use..yet.



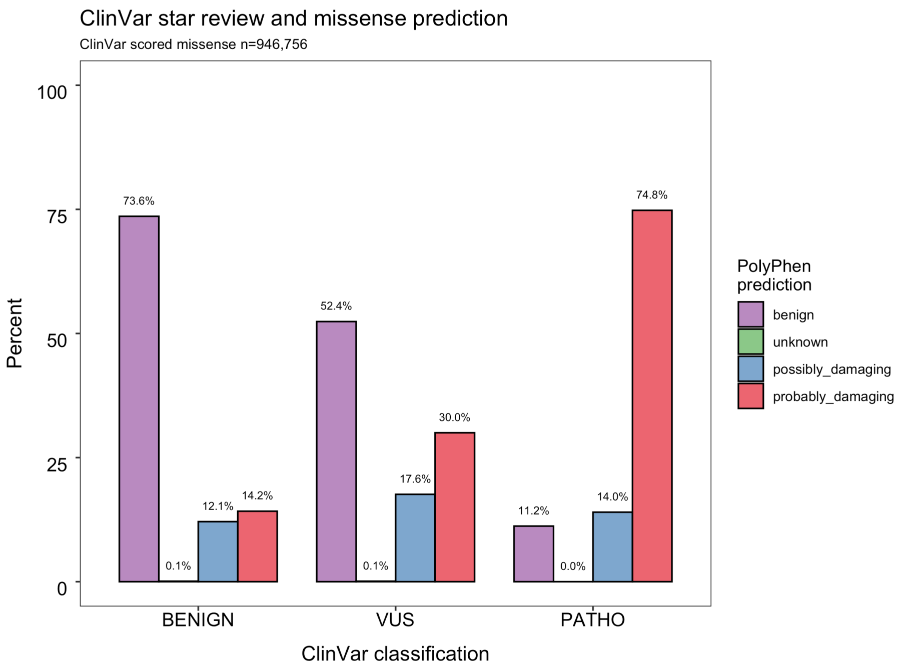
GH: was this a prediction you ran?

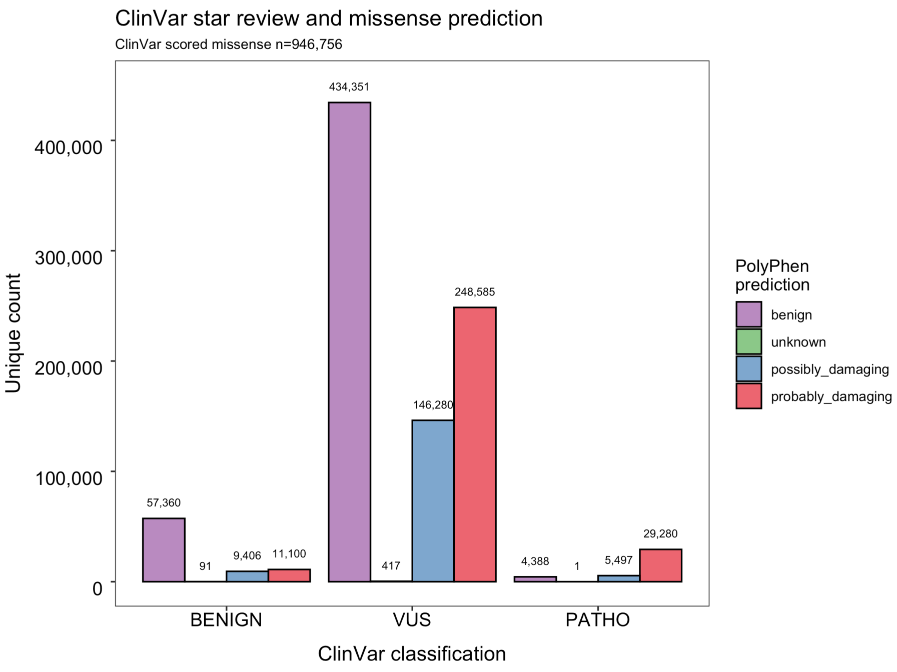


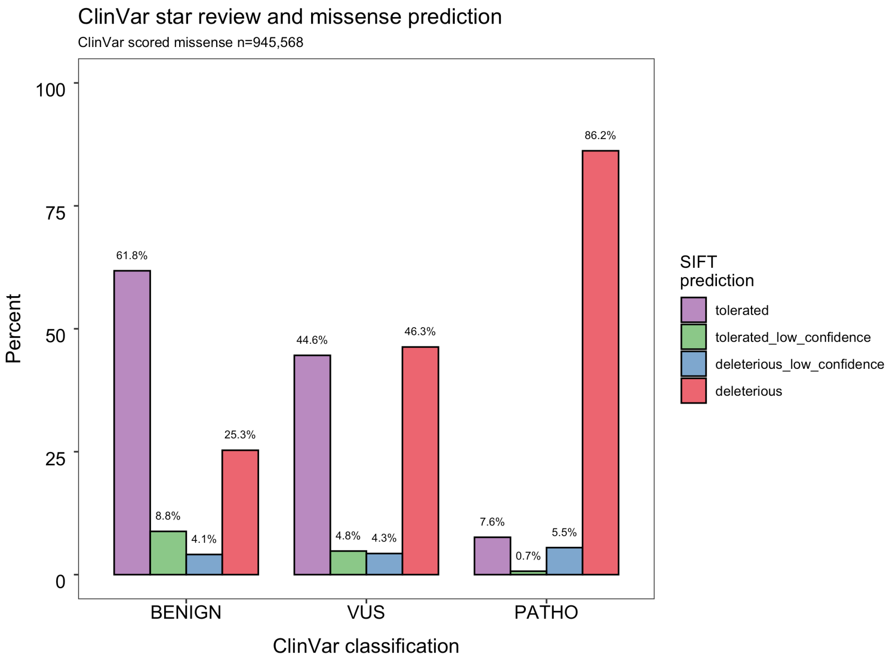
GH: was this a prediction you ran?

GH: Also intriguing! What are the absolute numbers, especially for the most common (>0.1) alleles.
Would PolyPhen and SIFT be able to predict health span impacting mutations, or only disease causing?

In other words, how does their scaling work? Circled missense alleles that are classified as benign or VUS could be interesting.

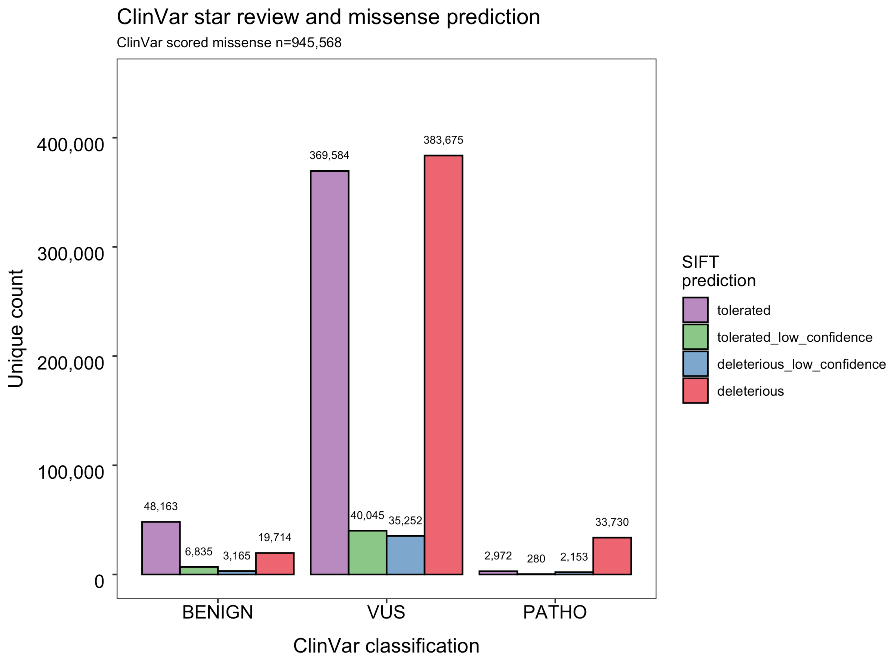

ClinVar star review and missense prediction
ClinVar scored missense n=946,756
Percent
100
75
50
25
0
73.6%
0.1%
12.1%
14.2%
52.4%
0.1%
17.6%
30.0%
11.2%
0.0%
14.0%
74.8%
BENIGN
VUS
PATHO
ClinVar classification
PolyPhen prediction
benign
unknown
possibly_damaging
probably_damaging


ClinVar star review and missense prediction
ClinVar scored missense n=946,756
Unique count
400,000
300,000
200,000
100,000
0
57,360
91
9,406
11,100
434,351
417
146,280
248,585
4,388
1
5,497
29,280
BENIGN
VUS
PATHO
ClinVar classification
PolyPhen prediction
benign
unknown
possibly_damaging
probably_damaging



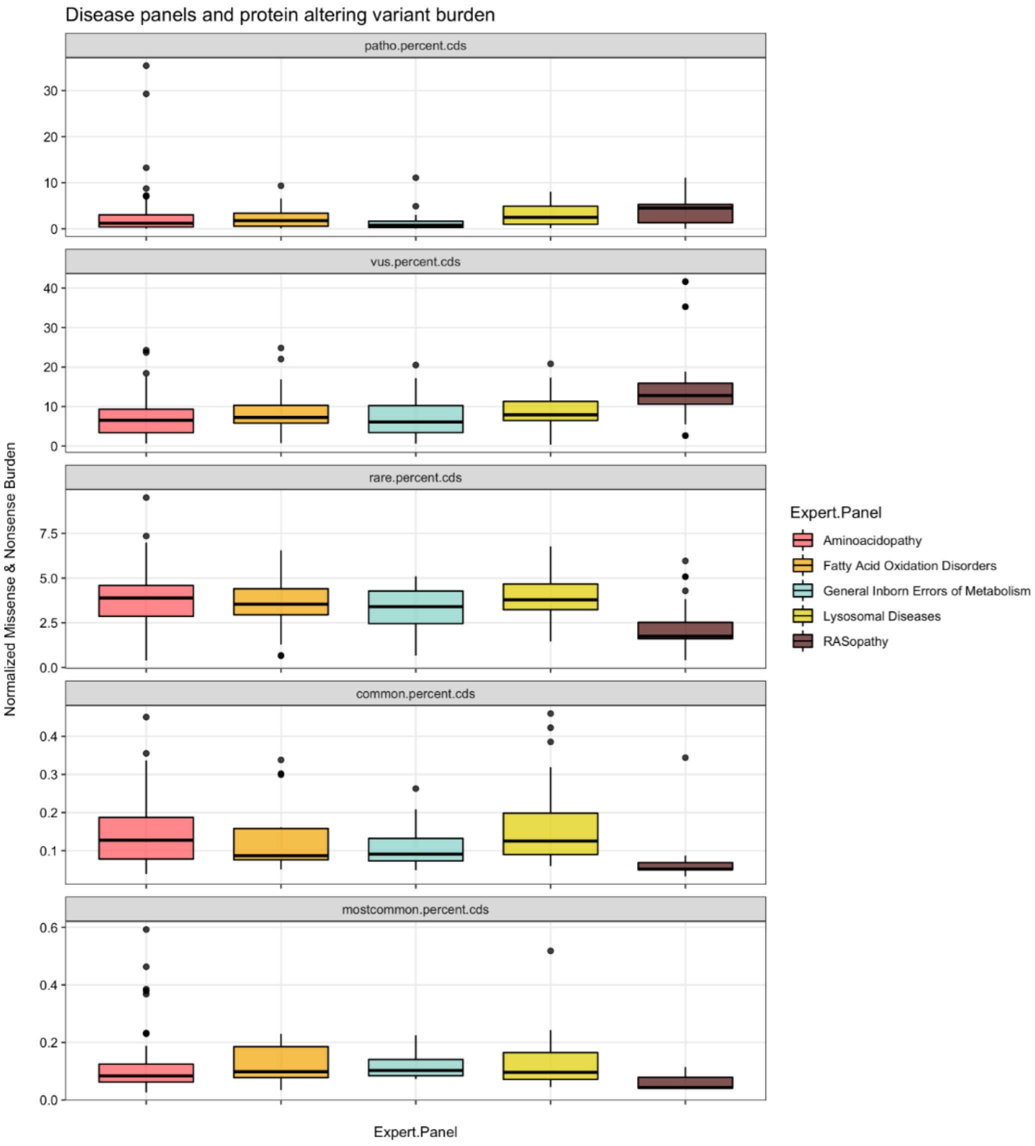
ClinVar star review and missense prediction

ClinVar scored missense n=945,568



ClinVar star review and missense prediction

ClinVar scored missense n=945,568



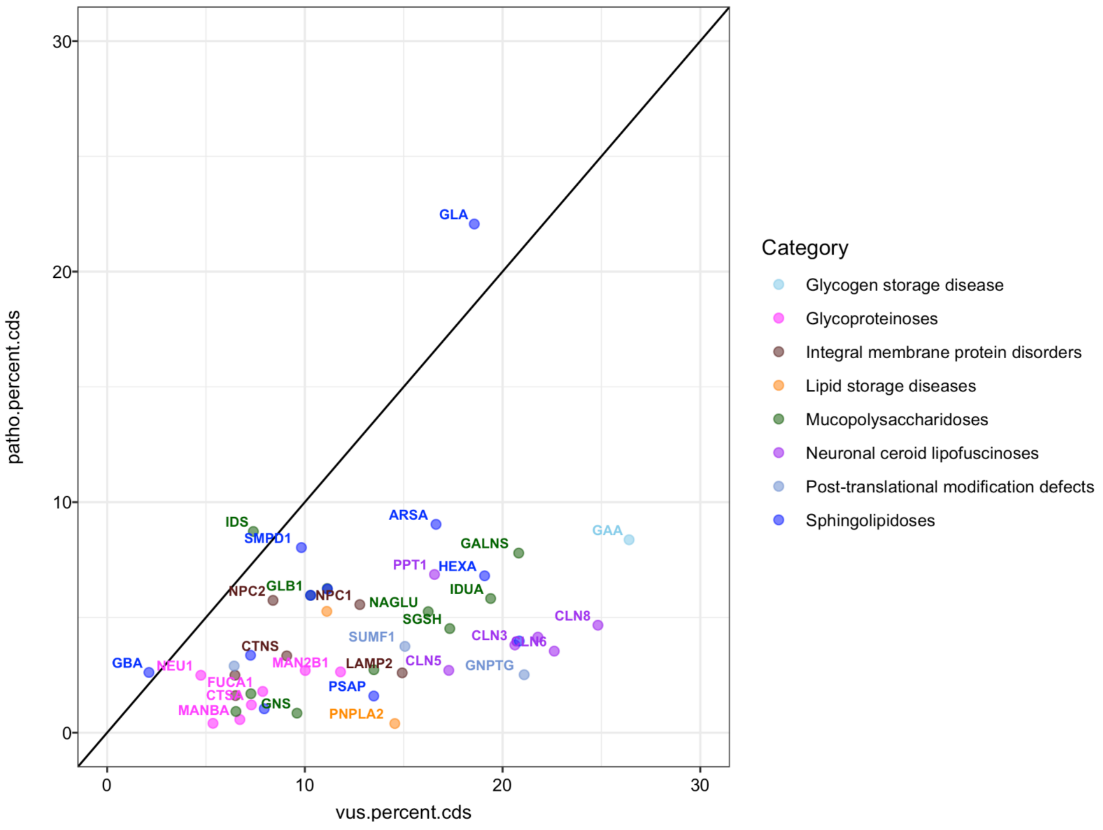
Next batch of analysis goes more into the gene counts and gene variant burden using counts of clinvar and gnomad variants that were normalized by gene CDS length. I use CDS length over protein length following examples in previous papers + CDS is at the same nucleotide level as variants.

GH: What is size of the panel?

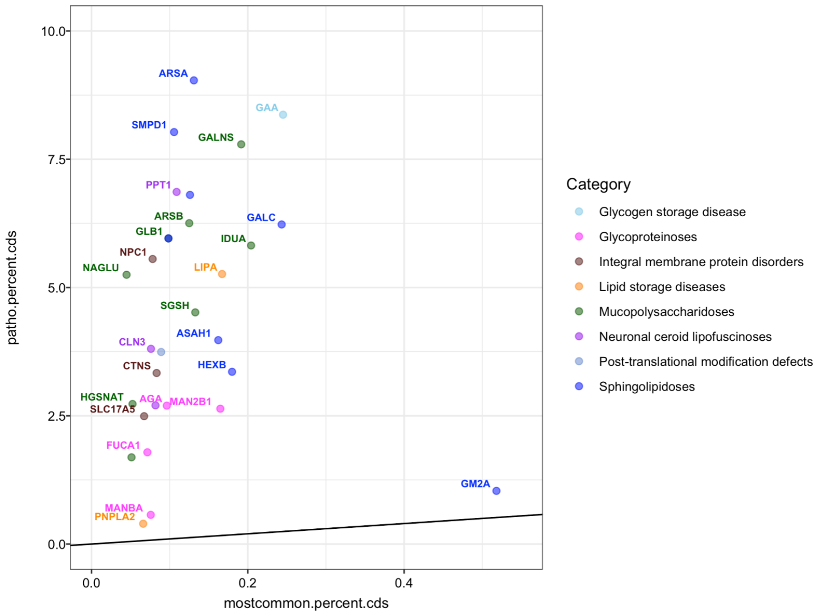
I am not sure I fully understand the data. My interpretation (which may very well be wrong) is that for instance for the aminoacidopathy related genes in this panel, the average number of pathogenic alleles that has been identified, is about 0.5-2.

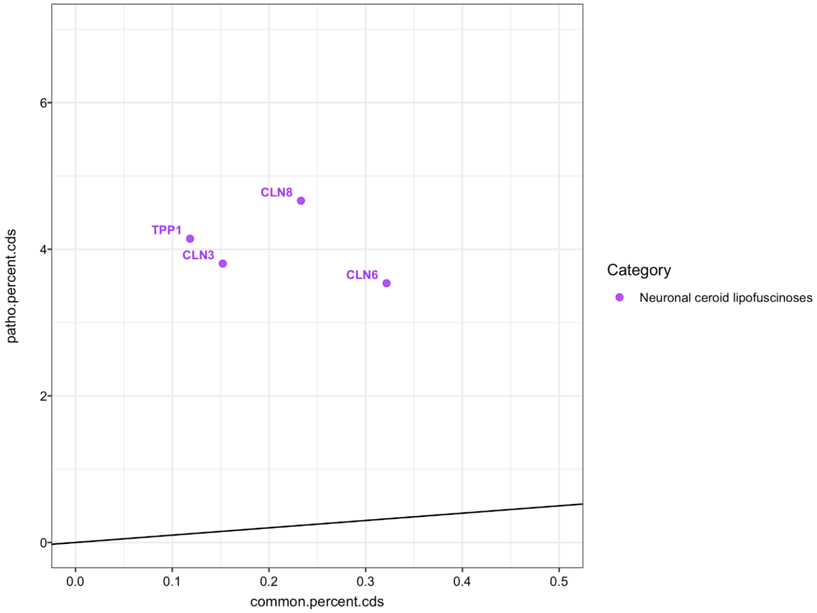
PAH (PKU) would be an outlier where there are a lot more pathogenic mutations known.

For the LSD it would be 1-4 with no outliers. Is the most informative data in the outliers? Not sure what we do with it, please help me.
What does intrigue me is the outliers in the mostcommon graph. Can you list them?



GH: What does this tell us?
From a medical perspective, we would like to have as many proteins/genes on the other side of the diagonal, to indicate the "state of knowledge". For instance, a protein with coordinates (0,30) would have no VUSs and we would know the pathogenic mutations. Is that true?





GH: Does this present the relationship between a mutation property (the most common allele) with a gene property (% of mutations that are pathogenic?

I think plots like these can be very informative and guide us to our pipeline. Maybe it is MAF (>0.1) on one axis or PolyPhen or SIFT on another? How would the MTHFR minor alleles look in such a plot (or COMT, GBA^{PD}*, etc.)*

Combination of WGS & WES: summed AC, AN, nhomalt columns and recalculated AF