DepMap Omics: 22Q4 Mutation Pipeline Update

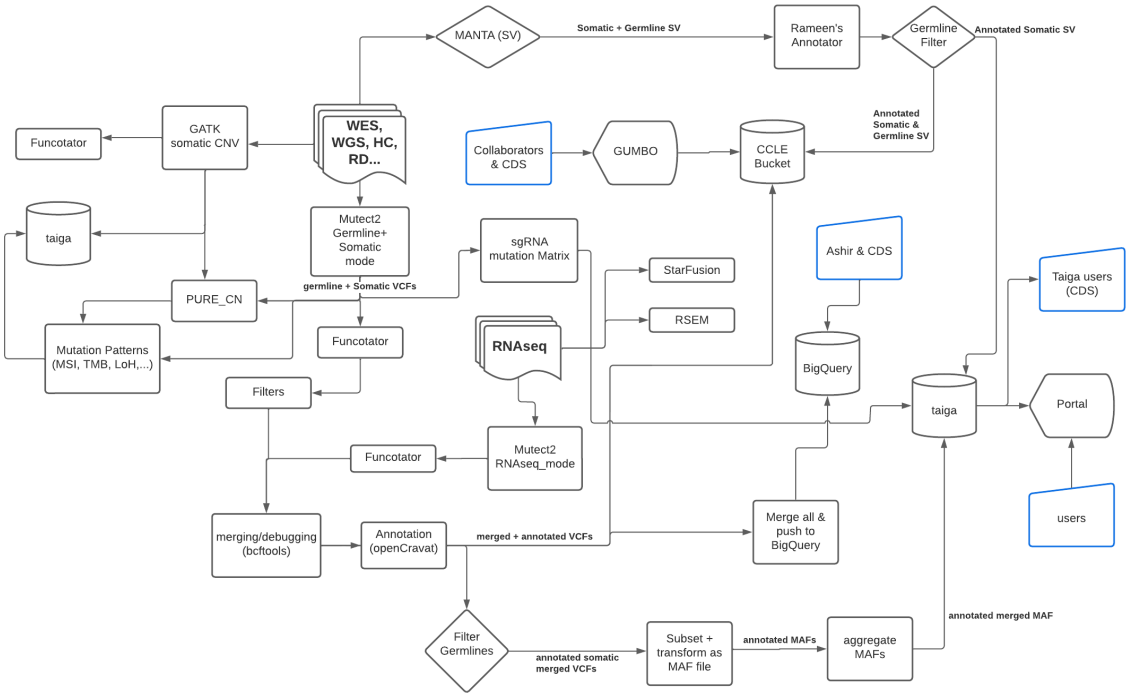
This document presents the changes & updates in the mutations pipeline. It explains why we made those and how it improves and changes the output data.
Here we are talking specifically about the SNPs and shorts INDELs mutation pipeline.
More information about the maf / vcf column names are available in our appendix.

Key Advantages

The previous pipeline had issues with filtering, germline calling, simplicity, scalability and maintainability. In the new pipeline, we have focused on addressing these issues.

1. **Simplicity**: the pipeline is easy to understand, use and debug
2. **Reproducibility**: anyone can now run our pipeline (see links below) whereas the previous pipeline is often too complicated to adapt to someone's workflow.
3. **Maintenance**: We are using a pipeline that is up to date with the community's standards and will continue to be so at every new release.
4. **Germline mutation calling**: integrated in the mutect2 pipeline.
5. **Whole genome mutation calling**: the previous method was not efficient enough to be run on non coding regions of the genome.
6. **Better annotations:** these annotations are now genome wide, using multiple state of the art predictive methods and knowledge bases (more in #Annotation).
7. **Ownership:** we can easily change / add tools and methods at every step of the pipeline.

Mutation Pipeline



Here is the full [WDL script] for the DNAseq pipeline.

Mutation Calling

The mutation calling is done with mutect2 (M2) with the specific parameters:

- Default format of **VCF** and **HG38**
- Interval padding of **100bp**
- Gatk_docker: **broadinstitute/gatk:4.2.6.1**
- M2_extra_args: **--genotype-germline-sites true --genotype-pon-sites true** for getting access to germline calls.
- Filter_funcotations to **False**
- Gnomad: **gs://gatk-best-practices/somatic-hg38/af-only-gnomad.hg38.vcf.gz**
- Pon: **gs://gatk-best-practices/somatic-hg38/1000g_pon.hg38.vcf.gz**
- Run_funcotator to **True**
- Run_orientation_bias_mixture_model_filter to **True**

The pipeline is run such that no bait sets need to be provided.
More information about mutect2: [Mutect2 – GATK],

The full pipeline code for Mutect2 can be accessed at : https://github.com/broadinstitute/depmap_omics/blob/master/WGS_pipeline/mutect2_v4.2.6.1.wdl).

During the run, an initial filtering step based on mutation quality and the like (FilterMutectCalls – GATK) and mutation annotation (Funcotator – GATK, Funcotator Information and Tutorial – GATK) occurs.

From funcotator we annotate variant types defined in **Gencode (v34), CGC (full_2012_03-15), Clinvar (20180429_hg38), Cosmic (v84), HGNC (Nov302017), Uniprot (2014_12), dbSNP (b151) & NCBI (hg38)**. These annotations include deleteriousness, mutation type, its frequency in cancer according to cosmic (per cancer type), fusion association, syndrome association, disease and pharmacogenomic associations, and related gene level information.

VCFs update for issues found in Mutect2

During our QCs and comparisons, we have noted a few issues with the mutect2 output VCF files that must be corrected. One is that the **GT** (genotype: see vcf file format) field does not work well, as the pseudo count makes for allele frequencies that never quite equal 1 and the **GT** is always **0/1**. Therefore, when **depth >= 8** and **allele frequency > 0.9,** we use bcftools to change an initial **0/1** to **1|1.**

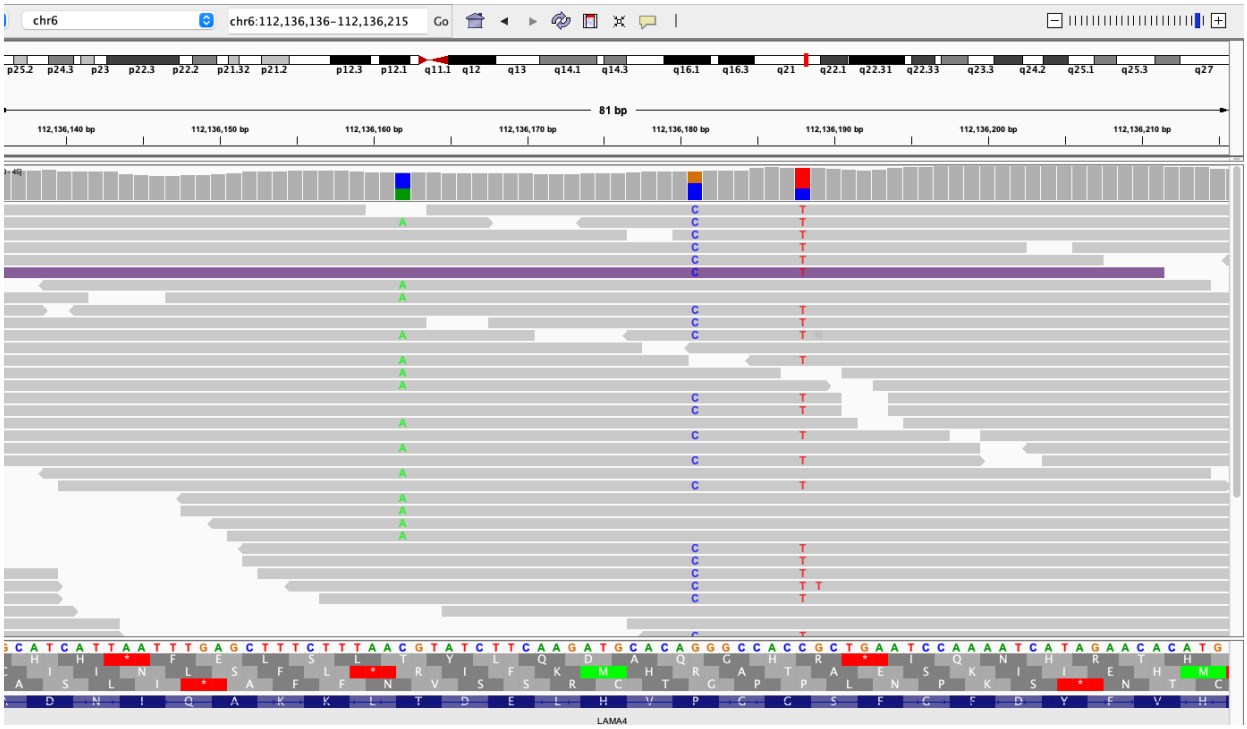
We do the same for a subset of most common multi-allelic sites (which have an allelic depth of 0 for the M (reference) allele):
0/1/2 -> 1/2
0/1/2/3 -> 1/2/3
0/1/2/3/4 -> 1/2/3/4
0/1/2/3/4/5 -> 1/2/3/4/5

This is done using the command **bcftools +setGT $FILE -- -t q -n m/m -i'INFO/DP>8 & AF>0.9'**. See further details in our WDL https://github.com/broadinstitute/depmap_omics/blob/master/WGS_pipeline/bcftools.wdl

A second issue is the annotation in the **AS_filter_status** column. This field in the vcf file contains "|" and ",", but their meanings are swapped compared to other columns in the VCF file. We swapped these back everywhere to keep the same meaning and parse the file easily. More in:
https://github.com/broadinstitute/depmap_omics/blob/master/WGS_pipeline/fix_mutect2.wdl &
https://github.com/broadinstitute/depmap_omics/blob/master/WGS_pipeline/fix_mutect2.py

A final issue concerns the clustered event correction. There is a known issue with filtering combined somatic and germline calls from Mutect2:
https://gatk.broadinstitute.org/hc/en-us/community/posts/4404184803227-Mutect2-genotype-germline-sites-filtering-discrepancy- .

Germline variants should not be considered when filtering out clustered events. For example, in the attached screenshot:



The somatic variant on the left is **flagged** as a **clustered_event** because it is near two germline variants. This issue affects about 2.5% of our Mutect2 somatic calls.

Unfortunately, we can't just ignore the clustered events filter since it removes a large number of sequencing and mapping errors. GATK should fix this eventually, but in the near term we have written a script to correct this post-hoc. it removes the clustered_event flag if less than 2 events in 100bp region are somatic.

Annotation

As with variant calling, we looked for an annotation method that would be simple, maintainable, and reproducible. We already have many annotations from funcotator (see above). But we needed more tools and to be able to easily add our own annotations.

Here we decided to go with the openCravat tool which allows user to build/use annotators and other tools to work with mutation files:
https://raw.githubusercontent.com/broadinstitute/depmap_omics/cfbfd3392bcc98ae02d8d971d2ac1d5c28306094/WGS_pipeline/opencravat_dm.wdl
https://open-cravat.readthedocs.io/en/latest/

We have decided on additional annotators based on the literature and available comparisons. Main annotators are defined here:

- Driver/hotspot: **Cscape, Civic, hess_et_al, Funcotator**

- Functional: **Dann, Revel, Provean, Funcotator**
- Splicing: **SpliceAI**
- Population: **Gtex, alfa, Funcotator**
- Expression: **Funseq2, ccre_screen**
- Pharmacological: **Pharmgkb, Civic**
- Disease: **Dida, Gwas_catalog, Funcotator**

List of annotators and description:

1. **ALFA (7)**: list the allele frequency of that mutation from all sequences in dbGAP https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/ .
2. **CScape (50, 11hWGS)**: predicts how likely a SNP is to be a cancer driver https://www.nature.com/articles/s41598-017-11746-4 .
3. **DANN (100, 12hWGS)**: a NN version of CADD, said to be better and works in non coding regions too https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341060/ .
4. **DIDA**: list if two mutations are in a pair listed as creating a digenic disease https://academic.oup.com/nar/article/44/D1/D900/2502599 .
5. **FunSeq2 (32, 8hWGS)**: define non coding mutations of interest based on many heuristics http://info.gersteinlab.org/Funseq2.
   - exp = can give the expression of that sample to associate with the mutation.
6. **GWAS_Catalog**: list if mutation is EBI GWAS catalog repository https://www.ebi.ac.uk/gwas/docs/api .
7. **PharmGKB**: whether that mutation is known to be associated with a drug https://www.pharmgkb.org/ .
8. **Civic**: clinical interpretation of cancer related SNPs https://civicdb.org/home .
9. **REVEL**: ensembl aggregation method of deleteriousness from different methods: MutPred, FATHMM v2.3, VEST 3.0, PolyPhen-2, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP++, SiPhy, phyloP, and phastCons https://pubmed.ncbi.nlm.nih.gov/27666373/ .
10. **Brca1_func_assay:** annotations of mutation functional effect from a large scale functional assay o the BRCA1 gene. https://github.com/KarchinLab/open-cravat-modules-karchinlab/blob/master/annotators/brca1_func_assay/brca1_func_assay.md
11. **ccre_screen:** cCREs are the subset of representative DNase hypersensitivity sites (rDHSs) supported by either histone modifications (H3K4me3 and H3K27ac) or CTCF-binding data in ENCODE.https://github.com/KarchinLab/open-cravat-modules-karchinlab/blob/master/annotators/ccre_screen/ccre_screen.md
12. **gtex:** the eQTL SNP annotation from gtex study. The annotation contain the gene and the tissue this eQTL has been detected for. https://github.com/KarchinLab/open-cravat-modules-karchinlab/blob/master/annotators/gtex/gtex.md
13. **hess_drivers:** annotation from whether the mutation is in the list of drivers from the hess et al paper. This list is trying to remove passenger hotspots but does not contain all cancer types. https://pubmed.ncbi.nlm.nih.gov/31526759/

As part of this annotation, we fix some strings in the header (Specifically, miss-usages of **“** and **‘**).
We also compress as **bgzip** and **reindex** using bcftools.

Improvement

Improvement is a function **depmap_omics.vcf.improve()** (see #subsetting for code availability) which uses the annotations we have to define new sets of columns/annotations that are better defined in our appendix. It also takes care of renaming some important columns and cleaning up column values so that the file has a similar logic / convention throughout.

Subsetting

We define as low quality any variant that is any of:

- map_qual
- slippage
- strand_bias
- weak_evidence
- clustered_events
- base_qual

See FilterMutectCalls – GATK for more details.

We define as germline anything that:

- Is defined as germline by mutect2 (see more about this method here: gatk/mutect.pdf at master · broadinstitute/gatk · GitHub)
- Is in our Panel of Normals (Pon) according to mutect2
- Has a Population Allele Frequency (popaf) above 10^(-3) according to gnomad.

We define as important any mutation that is a

- known driver, gain of function, loss of function.
- putative driver in oncogenes or tumor_suppressors
- putative loss of function in tumor_suppressors

Our list of **tumor suppressors & oncogenes** are available here and have been extracted from OncoKB
These annotations are defined by the annotators we used: Mutect2 Mutation Annotation. Find out more in our WDL file:
https://github.com/broadinstitute/depmap_omics/blob/dev/WGS_pipeline/opencravat_dm.wdl

We are removing all:

- low quality mutations.
- germline mutations except if they are important.
- non coding mutations except if they are important.

More information here:
https://dockstore.org/workflows/github.com/broadinstitute/depmap_omics/filter_to_maf:dev?tab=info
https://github.com/broadinstitute/depmap_omics/blob/master/depmap_omics/vcf.py

Important changes for users:

We will now have 2 MAF files on the portal (both somatic & exonic only):

- One at the model level (ACH-ID) where the best sequencing method is prioritized in the same way as currently done for the CN data (WGS>WES ; DEPMAP>CCLE2>SANGER).
- One at the sequencing profile level (PR-ID) where all the available sequencing methods for a cell line will correspond to a different profile ID.

CCLE_mutations_bool_damaging, CCLE_mutations_bool_hotspot, CCLE_mutations_bool_nonconserving, and CCLE_mutations_bool_otherconserving will no longer be released. Instead, we will now have the following 2 genotyped mutation matrices on the portal. The values in the matrices can be 0 (no mutation), 1 (heterozygous), or 2 (homozygous).

- OmicsSomaticMutationsMatrixDamaging: mutations that are “LikelyLoF” or “CCLEDeleterious” are considered damaging
- OmicsSomaticMutationsMatrixHotspot: mutations that are identified in the hess et al paper according to the “HessDriver” column are considered hotspots

Some cell lines will disappear from the dataset. This corresponds to files that have been lost, or which did not pass our QC.

- The following Sanger lines (missing bam files):
  ACH-002391, ACH-002393, ACH-002394, ACH-002217, ACH-002396, ACH-002390, ACH-002395
- The following lines which have no bam files or failed QC:
  ACH-000003, ACH-000014, ACH-000016, ACH-000033, ACH-000034, ACH-000049, ACH-000057, ACH-000064, ACH-000071, ACH-000084, ACH-000088, ACH-000116, ACH-000170, ACH-000185, ACH-000194, ACH-000229, ACH-000230, ACH-000282, ACH-000299, ACH-000300, ACH-000306, ACH-000333, ACH-000413, ACH-000494, ACH-000526, ACH-000539, ACH-000575, ACH-000578, ACH-000600, ACH-000629, ACH-000642, ACH-000658, ACH-000690, ACH-000710, ACH-000731, ACH-000737, ACH-000742, ACH-000850, ACH-000854, ACH-000904, ACH-000923, ACH-000931, ACH-001042, ACH-001043, ACH-001044, ACH-001047, ACH-001072, ACH-001090, ACH-001091, ACH-001094, ACH-001109, ACH-001142, ACH-001150, ACH-001187, ACH-001207, ACH-001214, ACH-001230, ACH-001234, ACH-001955, ACH-001956, ACH-001957

This also means we are dropping these AF (allele frequency) columns 'CGA_WES_AC', 'HC_AC', 'RD_AC', 'RNAseq_AC', 'SangerWES_AC', 'WGS_AC’ replacing them with one single AF value.

Other changes include many more annotations, more important mutations in key cancer genes and the removal of low AF mutations.

Moreover some additional data will now be publicly available on Terra:

- Raw vcfs (with germline & non coding & QC fail mutations)
- Filtered & Annotated tabular files (with germlines & non coding mutations)

Why germline?

- Reproducibility: other tools (e.g. ancestry bias, PureCN, ethnicity….) need it.
- For more questions: (synthetic lethality, functional matrices, …)

Moreover some mutations will be added and some mutations will be dropped. There are multiple reasons mutations are dropped:

- Some were only being called by other sequencing methods (HC/RD/RNA) ~20%.
- Some are now considered to be of low quality by the mutect2 filters ~5%.
- Some are now considered germlines by the pipeline (see above) ~ 50%
- Some had a low allele frequency (<0.15) ~10%.
- Some have been moved to other locations by the hg38 remapping ~5%
- Some are now described as DNPs or TNPs (dinucleotides/trinucleotides).
- Some have had a change in definition out-of-frame -> in-frame, splice-site -> . , etc..

Appendix

colname	**desc**	**example**	**previous name**
DepMap_ID or ProfileID		ACH-..., PR-....	DepMap_ID
Chrom		chr1, chrMT, chrY, ...	Chromosome
Pos		10234, 1303499, ...	Start_position
Ref		A, T, CTA, ...	Reference_Allele
Alt		TC, T, ACTCCTTTC, ...	Tumor_Allele
AF	allele frequency defined by (alt_count+1)/(ref_count+1)	0.89, 0.1, 0.99, ...	tumor_f
RefCount	number of reads with ref variant	1, 0, 10, ...	t_ref_count
AltCount	number of reads with alt variant	12, 3, 103, ...	t_alt_count

GT	genotype 0/1 means unphased, 1|0 means phased with parentalA being the mutated parent. more info here: https://www.internationalgenome.org/wiki/Analysis/vcf4.0/#:~:text=g.%20H2%3D0.-,Genotype%20fields,-If%20genotype%20information. Defined by mutect2.	0/1, 1|0, 1|1, 0|1	
PS	Phasing set (typically the position of the first variant in the set)	135803987, 410294097, ...,	
VariantType		SNP, INS, DEL, ...	Variant_Type
VariantInfo	gencode_34 variant classification	SILENT, MISSENSE, INTRON, NONSENSE, SPLICE_SITE...	Variant_Classification
DNAChange	code of DNA change of the variant	c.3380A>G, c.e9-7G>A, c.4587G>A, ...	Genome_Change
ProteinChange	code of protein change of the variant if happens in an exon else empty	p.D47E, p.V303I, p.A355A, ...,	Protein_Change
HugoSymbol		CAPN10, SLC6A6, LINC01811, ...,	Hugo_Symbol
HgncName		calpain 10, solute carrier family 6 member 6, golgin A4, ...,	
HgncFamily		Calpains, Synapsins, Classical arrestins, ...,	
Transcript	ENSEMBL transcript ID of the main transcript considered	ENST00000345617.7, ENST00000607357.2, ENST00000256474.3, ...,	Annotation_Transcript
TranscriptExon	exon of the main transcript considered	1, 4, 2, ...,	
TranscriptStrand	strand of the main transcript considered	-, +	Strand
UniprotID		Q9C0A6, P40337, Q92777, ...,	
Str	Variant is a short tandem repeat	Y,	
Popaf	-log10 scaled population allele frequency from gnomad	3.123, 5.123, 1.445, ...	ExAC_AF
DbsnpID		rs150638871, rs778239502, rs1642742, ...,	dbSNP_RS
DbsnpFilter	if it is filtered by dbsnp or not. not excatly the same as previous annotation: dbSNP_Val_Status which was about validation, but seems more stringent.	Y,	dbSNP_Val_Status
Issues	issues can be one of as_specific: assembly specific variant, as_specific, or assembly conflict variant assembly_conflict, between hg19 and hg38	as_specific, assembly_conflict,	
GcContent	percentage of GC content in the viscinity of the mutation	0.83, 0.14, ...	
LineageAssociation	Cancer Gene Census associated TissueType codes from funcotator	"""E, M, O""", E, L, ...,	
CancerMolecularGenetics	if the mutation is known to be recessive or	Rec, Dom,	

	dominant		
CCLEDeleterious	if the mutation is one of DE_NOVO_START_OUT_FRAME DE_NOVO_START_IN_FRAME FRAME_SHIFT_DEL FRAME_SHIFT_INS START_CODON_INS START_CODON_DEL NONSTOP NONSENSE	Y,	isDeleterious
StructuralRelation	if the gene is known to be structurally related to another set of genes by being often fused with or translocated with these genes	"""NPM1, TPM3, TFG, TPM4, ATIC, CLTC, MSN, ALO17, CARS, EML4, KIF5B, C2orf22""", MAST2, ...,	
CosmicOverlappingMutations	number of overlapping mutations from COSMIC in different tissues	1, 10, 0, 30,...,	COSMIChsCnt
CosmicHotspot	if more than 5 times in COSMIC	Y,	isCOSMIChotspot
AssociatedWith	any know relationhsip this mutation can have based on a suite of annotators. usefull for investigating further a specific mutation. know more at https://github.com/broadinstitute/depmap_omics/tree/master/depmapomics/vcf.py 's "improve()" function to know more..E.g: gene_lof: likely lof according to DANN: if likely pathogenic according to revel (score>0.7), cancer; if clinically_significant, hotspot according to cosmic or likely_driver according to cscape ...	gene_function_loss;structural_relation;, gene_function_loss;, gene_function_loss;disease;, expression;chemicals;, ...,	
DannScore	functional effect score from DANN		
RevelScore	functional effect score from revel		
CivicID	ID in the civic database		
CivicDescription	description from the civic database		
CivicScore	score from civic		
HessDriver	whether a driver hotspot according to hess et al.	Y,	
HessSignature	the mutational signature according to hess et al if driver hotspot		
CscapeScore	driver score from the cscape algorithm		
Funseq2Score	funseq2 score representing the functional impact of a non coding mutation		
PharmgkbID	ID in pharmgkb		

DidaID	ID in the DIDA (digenic disease) database		
DidaName	name of the digenic disease		
GwasDisease	name of the associated disease in a the gwas-catalog study		
GwasPmID	ID in the gwas-catalog database		
GTexGene	associated with a gene expression in gtex (shows that gene's name)	RR1,TP53, NMD2,	
LikelyGoF	likely to be gain of function according to some heuristics (if a likley driver in an oncogene)	Y,	
LoF	loss of function according to revel (score >.9) or a set of validated BRCA1 locations	Y,	
LikelyLoF	likely to be LoF driver in a tumor supressor gene or dann_score above .96		
Driver	cancer driver according to Civic	Y,	
LikelyDriver	if mutation has a civic score or is identified in hess et al	Y,	
TranscriptLikelyLoF	list of transcripts that are likely lof according to provean: http://provean.jcvi.org/index.php (value below -2.5)	ENST00000457433;ENST00000271064;ENST00000537531;, ENST00000457433;, ...,	

... means other similar values exist
, at the end means empty value is possible or likely