26Q1 DepMap Mutation Pipeline Documentation

Open source package and documentation

Our mutation pipeline and other omics pipeline has been actively productionalized to be usable as deployable WDL scripts and python packages with docker containers on Terra or with miniwdl. It is open source on GitHub https://github.com/broadinstitute/depmap-omics-wgs.

Since we are no longer sequencing WES, mutation calls for WES samples are not receiving tool or dataset upgrades. Please refer to the 25q3 mutation pipeline documentation for information about WES mutation calling.

DepMap WGS Mutation Calling

Our cancer cell lines mutations are analyzed using mutect2 (M2) with the following specific parameters:

- Default format of **VCF** and **HG38**
- Gatk_docker: **broadinstitute/gatk:4.6.1.0**
- M2_extra_args: **--genotype-germline-sites true --genotype-pon-sites true** for getting access to germline calls.
- Pon: **gs://gatk-best-practices/somatic-hg38/1000g_pon.hg38.vcf.gz**
- un_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-wgs/tree/main/workflows/call_mutations

During the run, an initial filtering step based on mutation quality and the like (FilterMutectCalls – GATK) and mutation annotation (using several tools) occurs.

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 in the prep_annotations workflow.

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.

We also normalize left-align indels and split multiallelic variants to separate rows.

Variant Annotation

We annotate the VCF in the annotate_mutations workflow as follows:

Tool	**Annotation(s)**	**Data source**
bcftools	Segmental duplication regions	UCSC genomicSuperDups track (2025-05-05)
	RepeatMasker	UCSC rmsk track (2025-05-05)
	Hess drivers	Hess et al.
	Oncogenic, hotspot, mutation effect	OncoKB annotator API
	CMC tier	COSMIC Cancer Mutation Census (v99)
	CIViC evidence score	CIViC variant summaries and molecular profile summaries (2025-04-29)
	Oncogene/tumor suppressor	OncoKB cancer gene list (2025-04-29)

	HGNC gene and gene group names	genenames.org (2025-04-29)
	GC proportion	Custom app
snpEff	snpEff	snpEff_v5_2_GRCh38.mane.1.2.ensembl
SnpSift	SnpSift	ClinVar (2025-04-21)
Open-Cravat	**Enabled OC modules:** vcfreporter brca1_func_assay ccre_screen gtex gwas_catalog pharmgkb provean revel spliceai	Open-Cravat (2.12.0)
VEP	(various) **Additional plugins:** AlphaMissense Loftool pLI	VEP (v113) with indicated plugins Genome assembly: GRCh38.p14 MANE Version: v1.3 GENCODE: 47 RefSeq: GCF_000001405.40-RS_2023_10 Regulatory build: 1.0 PolyPhen: 2.2.3 SIFT: 6.2.1 dbSNP: 156 COSMIC: 99 HGMD-PUBLIC: 2020.4 ClinVar: 2024-04 1000 Genomes: Phase 3 (remapped) gnomAD exomes: v4.1 gnomAD genomes: v4.1

These annotations are primarily used for variant filtering/selection, and the most relevant ones are propagated to the final mutation data products (see **Mutation Matrices** below), which can be downloaded on the portal. (Some annotations, like those from the OncoKB API, cannot be included due to licensing restrictions.)

Variant Selection

Refer to the workflow and Python module select_somatic_variants.

1. Filtering

We first filter out low-quality variants, i.e. those annotated by GATK FilterMutect as any of these:

- base_qual
- map_qual
- multiallelic
- slippage
- strand_bias
- weak_evidence

Then, we remove variants with AF less than 0.15 and/or DP less than 5.

2. Variant selection

After filtering, there are two ways variants could be selected for the final dataset.

High quality non-synonymous somatic variants

A variant is selected by this method if **all** of the following are true:

- Impactful splice event or protein sequence changed
    - OR
        - Impactful splice event
            - AND
                - VEP consequence contains the string “splice”
                - VEP impact is “HIGH” or “MODERATE”
        - Protein sequence changed (per VEP HGVSp)
- Not part of a clustered event (per Mutect2)
- Not in a segmental duplication region
- Not in a repeat masker region
- Low population prevalence
    - AND
        - VEP gnomADe_AF <= 1e-05 (imputed with 0)
        - VEP gnomADg_AF <= 1e-05 (imputed with 0)
        - Not in mutect2 PoN (1000 Genomes)
        - Appear in less than 10% of DepMap samples separately for WGS or WES

“Rescued” variants

We also “rescue” certain categories of variants that might not meet all of the strict criteria of “high quality non-synonymous somatic variants”, but are nevertheless known to be relevant in cancer genomics. A variant is selected by this method if **all** of the following are true:

- VEP gnomADe_AF <= 1e-03 and gnomADg_AF <= 1e-03 (note: this cutoff value is different than the one used in the other method)

- Cancer-relevant variant
    - OR
        - Oncokb mutation is “Loss-of-function” or “Gain-of-function”
        - Oncokb oncogenic is “Oncogenic”
        - Oncokb hotspot is TRUE
        - Cosmic CMC tier is 1
        - BRCA1 functional assay score <= -1.328
        - Is high impact oncogene
            - AND
                - Locus is within oncogene (per OncoKB)
                - VEP impact is “HIGH”
        - Is high impact TSG
            - AND
                - Locus is within TSG (per OncoKB)
                - VEP impact is “HIGH”
        - Variant is in Hess driver list
        - Variant is a driver noncoding mutation in TERT promoter (per COSMIC)
        - Variant is in the polypyrimidine track of intron 13 of MET

Mutation Matrices:

We currently release two mutation matrices:

- OmicsSomaticMutationsMatrixDamaging: A variant is considered a damaging mutation if LikelyLoF == True.
- OmicsSomaticMutationsMatrixHotspot: A variant is considered a hot spot if it's present in one of the following: Hess et al. 2019 paper, OncoKB hotspot, COSMIC mutation significance tier 1, TERT promoter mutations (C228T or C250T), mutations in the polypyrimidine track in intron 13 of MET. HLA genes are excluded.

The values in the matrices can be 0, 1, or 2: 0 == no mutation; If there is one or more damaging mutations in the same gene for the same cell line, the allele frequencies are summed, and if the sum is greater than 0.95, a value of 2 is assigned and if not, a value of 1 is assigned.

Important changes for users as of 26Q1

- Mutations in HLA-A and HLA-B are now removed from the hotspot list
- GATK upgraded from 4.2 to 4.6 (includes Mutect2, FilterMutect, others)
- Upgrades to Mutect2 have obviated previous workflow code that manually corrected the “clustered_event” annotation added by FilterMutect
- Annotation by Funcotator removed

- All external datasets used for annotation (e.g. COSMIC CMC, OncoKB oncogenes and tumor suppressor genes, etc.) were upgraded to the latest available versions as of May 2025

Columns

*columns used in filtering/rescuing

	Description	Examples
ModelID/ ModelConditionID		ACH-000001/MC-000001-ZU8p
Chrom	Chromosome	chr1, chr2, chr3, chr4, ...
Pos	Position	14711, 69270, 69511, 69897, ...
Ref	Reference allele	G, A, T, C, ...
Alt	Alternative allele	A, G, C, T, ...
AF	Allele fractions of alternate alleles in the tumor	0.416, 0.969, 0.983, 0.833, ...
DP	Approximate read depth; some reads may have been filtered	26, 30, 106, 4, ...
RefCount	Number of reads with ref variant	1, 0, 10, ...
AltCount	Number of reads with alt variant	12, 3, 103, ...
GT	Genotype	0/1, 1|1, 0|1, 1|0
PS	Phasing set (typically the position of the first variant in the set)	827209, 835538, 873542, 874301, ...
VariantType	Variant class from VEP	SNV, deletion, insertion, substitution
VariantInfo*	Consequence from VEP ordered by impact Please refer to this table to understand the sequence ontology https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html	missense_variant, stop_gained, missense_variant&splice_region_variant
DNAChange	Code of DNA change of the variant	ENST00000454784.10:c.19097_19098delinsAA, ENST00000333244.6:c.6426_6427inv, ENST00000463781.8:c.8223_8224delins

		CG
ProteinChange	Change of the variant if it's in an exon else empty	p.D47E, p.R344PfsTer17, K424_K425insPS, ...,
HugoSymbol*	Hugo Symbol from VEP	CAPN10, SLC6A6, LINC01811, ...,
EnsemblGeneID	Ensembl gene ID annotated from VEP	ENSG00000181143, ENSG00000155657, ENSG00000205592, ENSG00000196218, ...
EnsemblFeatureID	Ensembl transcript ID annotated from VEP	ENST00000397910, ENST00000589042, ENST00000454784, ENST00000359596, ...
HgncName		calpain 10, solute carrier family 6 member 6, golgin A4, ...,
HgncFamily		Calpains, Synapsins, Classical arrestins, ...,
UniprotID	Uniprot database protein ID	Q9C0A6, P40337, Q92777, ...
DbsnpRsID	dbSNP ID (i.e. rs number)	62635655.0, 121434228.0, 121434224.0, 121434227.0, ...
GcContent	GC content	0.3840399002493765, 0.3117206982543641, 0.6683291770573566,...
NMD	Predicted nonsense mediated decay effects for this variant from ClinVar. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene | Percent_of_transcripts_affected'	(MEGF6|ENSG00000162591.17|1|1.00), (AADACL3|ENSG00000188984.12|1|1.00), (CDCA8|ENSG00000134690.11|1|1.00), (VANGL1|ENSG00000173218.15|1|1.00), ...
MolecularConsequence	Comma separated list of molecular consequence from ClinVar in the form of Sequence Ontology ID|molecular_consequence	SO:0001623|5_prime_UTR_variant, SO:0001627|intron_variant, SO:0001819|synonymous_variant, SO:0001583|missense_variant, ...
Exon	Exon number	4/4, 20/20, 26/34, ...
Intron	Intron number	4/19, 3/7, 39/44, ...
VepImpact*	Impact prediction from VEP	MODIFIER, LOW, MODERATE, HIGH
VepBiotype	Biotype information from VEP	lncRNA, protein_coding, unprocessed_pseudogene, ...
VepHgncID	HGNC ID from VEP	HGNC:14825, HGNC:55080, HGNC:53981, ...

VepExistingVariation	Existing Variation from VEP	rs868589190, rs201219564&COSV58736820, rs2691305&COSV58736924, rs200676709&COSV58736747, ...
VepManeSelect	Mane transcript selected by VEP	NM_001005484.2, NM_001385641.1, NM_015658.4, ...
VepENSP	Ensembl protein ID from VEP	ENSP00000493376, ENSP00000478421, ENSP00000317992, ...
VepSwissprot	SwissProt ID from VEP	Q9Y3T9.189, Q6TDP4.147, Q5SV97.109, ...
Sift	SIFT score	tolerated(0.92), tolerated(0.55), deleterious(0), ...
Polyphen	PolyPhen score	benign(0), possibly_damaging(0.767), benign(0.249), ...
GnomadeAF*	Frequency of existing variant in gnomAD exomes combined population	0.0, 3.977e-06, 4e-06, 7.967e-06, ...
GnomadgAF*	Frequency of existing variant in gnomAD genomes combined population	0.0, 6.575e-06, 6.569e-06, ...
VepClinSig	ClinVar clinical significance of the dbSNP variant	benign, conflicting_interpretations_of_pathogenicity, not_provided, ...
VepSomatic	Somatic status of existing variant(s); multiple values correspond to multiple values in the Existing_variation field, this provides the index for VepExistingVariation	0&1, 0&1&1, 0&0&1, ...
VepPliGeneValue	Probabililty of a gene being loss-of-function intolerant (pLI)	0.03, 0.0, 0.36, 0.17, ...
VepLofTool	LoFtool score	0.679, 0.922, 0.164, ...
OncogeneHighImpact*	If variant is an Oncogene in OncoKB and has VepImpact == HIGH	True, False
TumorSuppressorHighImpact*	If variant is a Tumor suppressor in OncoKB and has VepImpact == HIGH. Tumor suppressor list	True, False

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;, ...,
Brca1FuncScore*	Function scores from BRCA1 functional assay	-0.2882635071, -1.899596881, -0.1421652642, ...
CivicID	ID in the CIVIC database	110.0, 908.0, 81.0, ...
CivicDescription	Description from the CIVIC database	MAP2K1 Q56P is a recurrent mutation in melanoma and gastric cancer. This mutation has been shown to confer considerable resistance to AZD6244 treatment of melanoma cell lines.', 'MAP2K1 P124S is a recurrent mutation in melanoma, and is seen in bladder and colon cancer to a lesser degree. The P124S mutation has been shown to contribute to AZD6244 resistance in melanoma cell lines, but considerably less so than its Q56P counterpart.',
CivicScore	Variant score from CIVIC	19.0, 7.5, 104.5, 162.0, ...
LikelyLoF	If variant's mutation effect is "Likely Loss-of-function" or "Loss-of-function" in OncoKB or its VEP impact is "HIGH"	True, False
HessDriver*	Whether a driver hotspot according to Hess et al. 2019	True, False
HessSignature	The mutational signature according to Hess et al if driver hotspot	CpG:1, UV:1, POLE:6, Misc:3, UV:4, APOBEC:1, ...
RevelScore	Functional effect score from revel	0.214, 0.013, 0.095, 0.363, ...
Pharmgkbld	ID in PharmGKB	PA166154663, ...
GwasDisease	Name of the associated disease in a the GWAS-catalog study	Post bronchodilator FEV1, Coronary artery disease (myocardial infarction, percutaneous transluminal coronary angioplasty, coronary artery bypass grafting, angina or chromic ischemic heart disease), ...
GwasPmID	ID in the GWAS-catalog	26634245.0, 28714975.0, 30038396.0,

	database	30595370.0, ...
GtexGene	Associated with a gene expression in gtex (shows that gene's name)	ENSG00000173805.11|ENSG00000173805.11|ENSG00000173805.11, ...
ProveanPrediction	Variant effect prediction from Provean	Neutral, Damaging
AMClass	AlphaMissense class	Likely benign, Likely pathogenic, ambiguous
AMPathogenicity	AlphaMissense pathogenicity	0.1437, 0.7034
Hotspot	Whether variant satisfies at least one of the hotspot criteria above	True, False
Rescue*	Whether variant satisfies at least one of the rescue criteria above	True, False
RescueReason	Supporting evidence for why variant is rescued	"TS_high_impact", "OncoKB, Cosmic, Hess", "Cosmic, TS_high_impact, Hess", ...
EntrezGeneID		673, 3895, 57412, ...