23Q4 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.

DepMap WGS and WES 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:4.2.6.1**
- M2_extra_args: **--genotype-germline-sites true --genotype-pon-sites true** for getting access to germline calls.
- Filter_funcotations to **False**
- 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 variants defined in **Gencode (v34), HGNC (Nov302017), & NCBI (hg38)**.

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

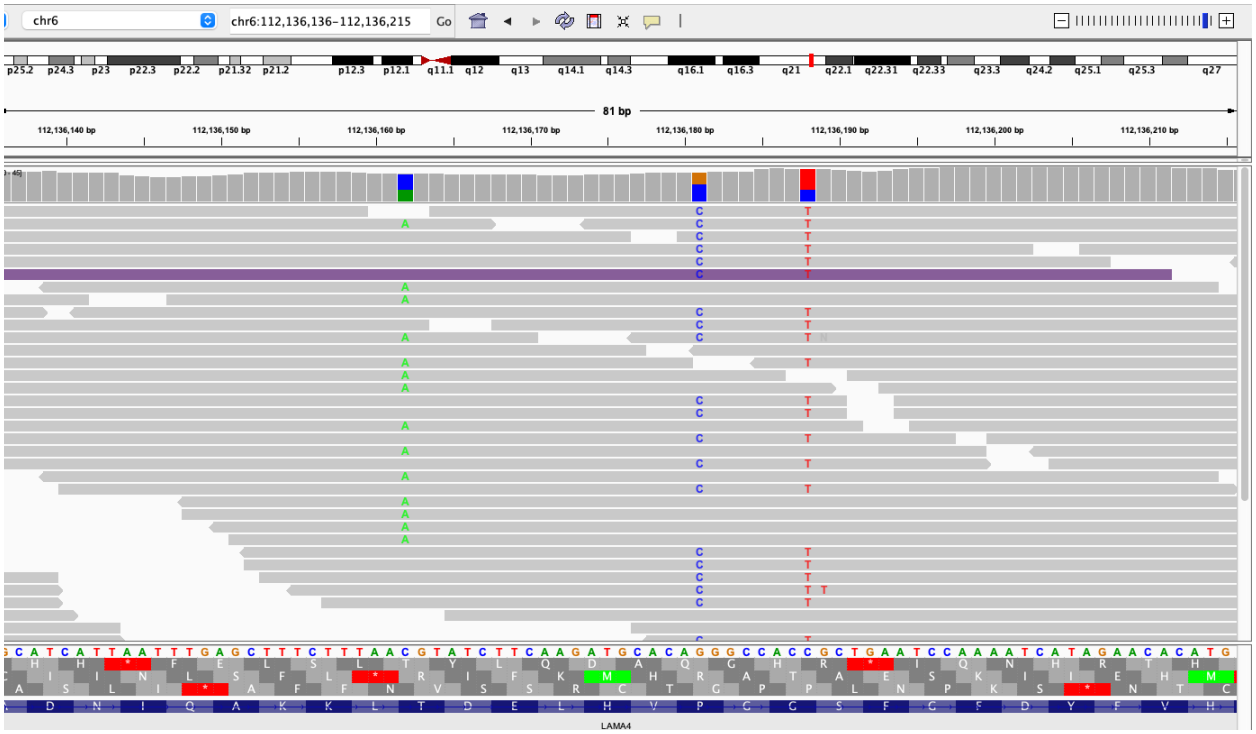
work well, as the pseudo count makes for allele frequencies that never quite equal 1 and the **GT** is always **0/1**. Therefore, when **depth >= 8** and **allele frequency > 0.9,** we use bcftools to change an initial **0/1** to **1|1.**

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. The script removes the clustered_event flag if less than 2 events in the 100 bp region are somatic.

Variant Annotation

Before the variant annotation, we normalize all the variants by a) indel left alignment; b) multi-allelic split into multiple rows for each allele. These steps are achieved by the bcftools command:
bcftools norm -m- -w 10000 -f ~{fasta} -O z -o ~{sample_id}.norm.vcf.gz ~{input_vcf}

In addition to Funcotator, we have added **SnpEff** (5.1d, 2022-04-19), **ClinVar** (2023-01), and ensembl **VEP** (version 110.1) to further annotate variants, with VEP as the major annotator The VEP reference databases are downloaded as used as a offline mode (homo_sapiens_vep_110_GRCh38.tar.gz from the web page https://useast.ensembl.org/info/docs/tools/vep/script/vep_cache.html#cache_content). The variant annotation pipeline can be found at https://github.com/broadinstitute/depmap_omics/blob/master/sandbox/hgvs/hgvs.wdl.
The cache database in VEP include:

Source	Version (GRCh38)
Ensembl database version	110
Genome assembly	GRCh38.p14
GENCODE	44
RefSeq	110 (GCF_000001405.40_GRCh38.p14_genomic.gff)
Regulatory build	1.0
PolyPhen	2.2.3
SIFT	6.2.1
dbSNP	154
COSMIC	97
HGMD-PUBLIC	2020.4
ClinVar	2023-01
1000 Genomes	Phase 3 (remapped)
gnomAD exomes	r2.1.1, exomes only
gnomAD genomes	r3.1.2, genomes only

We also incorporated OpenCravat 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/

Main annotators used in OpenCravat are:

- Driver/hotspot: **CIViC, hess_et_al, Funcotator, OncoKB, COSMIC**
- Functional: **Revel, Provean, Funcotator**
- Population: **Gtex, Funcotator**
- Pharmacological: **Pharmgkb, CIViC**
- Disease: **Dida, Gwas_catalog, Funcotator**

List of annotators and description:

1. **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 .
2. **GWAS_Catalog**: list if mutation is EBI GWAS catalog repository https://www.ebi.ac.uk/gwas/docs/api .
3. **PharmGKB**: whether that mutation is known to be associated with a drug https://www.pharmgkb.org/ .
4. **CIViC**: clinical interpretation of cancer related SNPs https://civicdb.org/home. Annotation uses data exported on Sep 22, 2022.
5. **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/ .
6. **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
7. **GTEX:** the eQTL SNP annotation from gtex study. The annotation contains 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
8. **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/
9. **OncoKB***: mutation effect, hotspot status, and list of oncogenes and tumor suppressor genes from https://www.oncokb.org/.
10. **COSMIC*:** mutation significance from COSMIC's Cancer Mutation Census https://cancer.sanger.ac.uk/cmc/. Annotation uses data exported on May 9, 2023.

*Under our current license agreement, we cannot release OncoKB or COSMIC annotations as individual columns in the mutation file on the DepMap portal. However, they are used to inform some filtering/rescuing decisions explained below.

Filtering

We remove variants that belong to any of the following categories, with the exception of those marked (*) which can be rescued if a variant is considered important. Criteria for rescue can be found in the following “Rescuing” section.

1) Low quality: We define as low quality any variant that is any of:
    a) Map_qual
    b) Slippage
    c) Strand_bias
    d) Weak_evidence
    e) Clustered_events
    f) Base_qual
    g) AF < 0.15
    h) DP < 2

See FilterMutectCalls – GATK for more details. We removed the above flags from Mutect2 using bcftools:

bcftools view --exclude FILTER~"weak_evidence"||FILTER~"map_qual"||FILTER~"strand_bias"||FILTER~"slippage"||FILTER~"clustered_events"||FILTER~"base_qual"

2) Germline: We define as germline anything that:
    a) Is in our Panel of Normals (Pon) according to mutect2 (*)
    b) Has a gnomADg or gnomADe allele frequency above 10^(-5) (*)

3) Non-coding and non-splicing: we remove the following variants that are non-coding and non-splicing:
    a) VariantInfo contains “splice” and VepImpact is not “MODERATE” or “HIGH” (*)
    b) “synonymous” is the leftmost variant classification in VariantInfo
    c) No HugoSymbol found
    d) Variant_Classification (column in .maf file only) is one of the following: 'Silent', 'RNA', 'Intron', "5'UTR", "3'Flank", 'Splice_Region', "5'Flank" (*)

4) Variants in repeated or segmentally duplicated regions (*). Bed files containing these regions can be found in https://github.com/broadinstitute/depmap_omics/tree/master/data/repeatMasker_max10_noAlt_merged.bed and https://github.com/broadinstitute/depmap_omics/tree/master/data/segDup_majorAllele_withAltContigs_98pcFracMatch_merged.bed, respectively. For details on how these files were generated, see

https://github.com/broadinstitute/depmap_omics/blob/master/docs/source/dna.md#masking.

5) Recurring variants that appear in more than 10% of DepMap samples (*) separately for either WGS or WES

Rescuing

There are variants that might be filtered out according to some of the previous criteria but are meaningful in the context of cancer. Therefore, we gathered the following annotations to construct a list of variants to rescue. **Note that a rescued variant still needs to have both gnomADg and gnomADe allele frequency < 10^(-3) in order to be rescued due to the observed germline mutations in these databases.**

- "Loss-of-function", "Gain-of-function", Oncogenic, or Hotspot according to OncoKB
- Tier 1 in COSMIC's mutation significance
- Brca1FuncScore <= -1.328 according to the BRCA1 Functional Assay
- OncogeneHighImpact or TumorSuppressorHighImpact is True
- HessDriver is True
- TERT promoter mutations rescued using driver noncoding mutations of TERT in the NCV_CDS_syntax_mapping file downloaded from the COSMIC database

More information here:
https://github.com/broadinstitute/depmap_omics/blob/master/depmapomics/tasks/vcf_to_depmap.py

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.

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 23Q4

- Less germline contamination with the latest gnomad database
- Updated set of hotspots
- New annotator for the standard DNA change and protein change
- Assign mutations to isoform with MANE transcript
- Mask artifactual mutations by deleting repeat masker and segmental duplication regions

Columns

*columns used in filtering/rescuing

	Description	**Examples**
ModelID/ProfileID		ACH-000001/PR-dga1pX
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_8224delinsCG

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,...
LofGeneName	Gene name from Predicted loss of function effects from VEP	MEGF6, AADACL3, PRAMEF13, HSPG2, ...
LofGeneId	Gene ID from Predicted loss of function effects from VEP	ENSG00000162591.17, ENSG00000188984.12, ENSG00000279169.3, ENSG00000142798.20, ...
LofNumberOfTranscriptsInGene	Number of Transcripts affected in gene from Predicted loss of function effects from VEP	1, 2, ...
LofPercentOfTranscriptsAffected	Percentage of transcripts affected from Predicted loss of function effects from VEP	1.00, 0.50, ...
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, ...
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. Oncogene list: https://github.com/broadinstitute/depmap_omics/blob/master/depmapomics/tasks/oncokb_dm/data/onocogene_oncokb.txt	True, False
TumorSuppressorHighImpact*	If variant is a Tumor suppressor in OncoKB and has VepImpact == HIGH. Tumor suppressor list: https://github.com/broadinstitute/depmap_omics/blob/master/depmapomics/tasks/oncokb_dm/data/tumor_suppressor_oncokb.txt	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: https://github.com/broadinstitute/depmap_omics/blob/master/depmapomics/tasks/civic_export_09212022.csv	110.0, 908.0, 81.0, ...
CivicDescription	Description from the CIVIC database: https://github.com/broadinstitute/depmap_omics/blob/master/depmapomics/tasks/civic_export_09212022.csv	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: https://github.com/broadinstitute/depmap_omics/blob/master/depmapomics/tasks/civic_export_09212022.csv	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 and 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, ...
PharmgkbId	ID in PharmGKB	PA166154663, ...
DidaID	ID in the DIDA (digenic disease) database	dd172, dd032, ...
DidaName	Name of the digenic disease	MODY, Familial dysfibrinogenemia, ...
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 database	26634245.0, 28714975.0, 30038396.0, 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
Rescue*	Whether variant satisfies at least one of the rescue criteria above	True, False

EntrezGeneID		673, 3895, 57412, ...