GDC VCF Format

Introduction

The GDC DNA-Seq somatic variant-calling pipeline compares a set of matched tumor/normal alignments and produces a VCF file. VCF files report the somatic variants that were detected by each of the four variant callers. Four raw VCFs (Data Type: Raw Simple Somatic Mutation) are produced for each tumor/normal pair of BAMs. Four additional annotated VCFs (Data Type: Annotated Somatic Mutation) are produced by adding biologically relevant information about each variant.

The GDC VCF file format follows standards of the Variant Call Format (VCF) Version 4.1 Specification (https://samtools.github.io/hts-specs/VCFv4.1.pdf). Raw Simple Somatic Mutation VCF files are unannotated, whereas Annotated Somatic Mutation VCF files include extensive, consistent, and pipeline-agnostic annotation of somatic variants.

VCF file structure

Metadata header

A VCF file starts with lines of metadata that begin with `##`. Some key components of this section include:

- **gdcWorkflow:** Information on the pipelines that were used by the GDC to generate the VCF file. Annotated VCF files contain two *gdcWorkflow* lines, one that reports the variant calling process and one that reports the variant annotation process.
- **INDIVIDUAL:** information about the study participant (`case`), including:
  - *NAME:* Submitter ID (barcode) associated with the participant
  - *ID:* GDC case UUID
- **SAMPLE:** sample information, including:
  - *ID:* NORMAL or TUMOR
  - *NAME:* Submitter ID (barcode) of the aliquot
  - *ALIQUOT_ID:* GDC aliquot UUID
  - *BAM_ID:* The UUID for the BAM file used to produce the VCF
- **INFO:** Format of *additional information* fields

    - NOTE: GDC Annotated VCFs may contain multiple INFO lines. The last INFO line contains information about annotation fields generated by the Somatic Annotation Workflow (../../../Data_Dictionary/viewer/#?view=table-definition-view&id=somatic_annotation_workflow) (see GDC INFO Fields below).
- **FILTER:** Description of filters that have been applied to the variants
- **FORMAT:** Description of genotype fields
- **reference:** The reference genome used to generate the VCF file (GRCh38.d1.vd1.fa)
- **contig:** A list of IDs for the contiguous DNA sequences that appear in the reference genome used to produce VCF files
    - NOTE: Annotated VCFs include contig information for autosomes, sex chromosomes, and mitochondrial DNA. Unplaced, unlocalized, human decoy, and viral genome sequences are not included.
- **VEP:** the VEP command used by the Somatic Annotation Workflow (../../../Data_Dictionary/viewer/#?view=table-definition-view&id=somatic_annotation_workflow) to generate the annotated VCF file.

Column Header Line

Each variant is represented by a row in the VCF file. Below each of the columns are described:

1. **CHROM:** The chromosome on which the variant is located
2. **POS:** The position of the variant on the chromosome. Refers to the first position if the variant includes more than one base
3. **ID:** A unique identifier for the variant; usually a dbSNP rs number if applicable
4. **REF:** The base(s) exhibited by the reference genome at the variant's position
5. **ALT:** The alternate allele(s), comma-separated if there are more than one
6. **QUAL:** Not populated
7. **FILTER:** The names of the filters that have flagged this variant. The types of filters used will depend on the variant caller used.
8. **INFO:** Additional information about the variant. This includes the annotation applied by the VEP.
9. **FORMAT:** The format of the sample genotype data in the next two columns. This includes descriptions of the colon-separated values.
10. **NORMAL:** Colon-separated values that describe the normal sample
11. **TUMOR:** Colon-separated values that describe the tumor sample

See Variant Call Format (VCF) Version 4.1 Specification (https://samtools.github.io/hts-specs/VCFv4.1.pdf) for details.

GDC INFO fields

The following variant annotation fields are currently included in Annotated Somatic Mutation VCF files. Please refer to the DNA-Seq Analysis Pipeline (../../Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/) documentation for details on how this information is generated. VEP Documentation (http://useast.ensembl.org/info/docs/tools/vep/vep_formats.html#output) provides additional information about some of these fields.

Field	Description
Allele	The variant allele used to calculate the consequence
Consequence	Consequence type of this variant
IMPACT	The impact modifier for the consequence type
SYMBOL	The HUGO gene symbol
Gene	Ensembl stable ID of the affected gene
Feature_type	Type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature.
Feature	Ensembl stable ID of the feature
BIOTYPE	The type of transcript or regulatory feature (e.g. protein_coding)
EXON	Exon number (out of total exons)
INTRON	Intron number (out of total introns)
HGVSc	The HGVS coding sequence name
HGVSp	The HGVS protein sequence name
cDNA_position	Relative position of base pair in cDNA sequence
CDS_position	Relative position of base pair in coding sequence
Protein_position	Relative position of the affected amino acid in protein
Amino_acids	Change in amino acids (only given if the variant affects the protein-coding sequence)
Codon	The affected codons with the variant base in upper case
Existing_variation	Known identifier of existing variant; usually a dbSNP rs number if applicable
ALLELE_NUM	Allele number from input; 0 is reference, 1 is first alternate, etc.
DISTANCE	Shortest distance from variant to transcript
STRAND	The DNA strand (1 or -1) on which the transcript/feature lies
FLAGS	Transcript quality flags
VARIANT_CLASS	Sequence Ontology variant class
SYMBOL_SOURCE	The source of the gene symbol
HGNC_ID	HGNC gene ID
CANONICAL	A flag indicating if the transcript is denoted as the canonical transcript for this gene

Field	Description
TSL	Transcript support level
APPRIS	APPRIS isoform annotation
CCDS	The CCDS identifer for this transcript, where applicable
ENSP	The Ensembl protein identifier of the affected transcript
SWISSPROT	UniProtKB/Swiss-Prot identifier of protein product
TREMBL	UniProtKB/TrEMBL identifier of protein product
UNIPARC	UniParc identifier of protein product
RefSeq	RefSeq gene ID
GENE_PHENO	Indicates if the gene is associated with a phenotype, disease or trait
SIFT	The SIFT prediction and/or score, with both given as prediction (score)
PolyPhen	The PolyPhen prediction and/or score
DOMAINS	The source and identifier of any overlapping protein domains
HGVS_OFFSET	Indicates by how many bases the HGVS notations for this variant have been shifted
GMAF	Non-reference allele and frequency of existing variant in 1000 Genomes
AFR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined African population
AMR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined American population
EAS_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined East Asian population
EUR_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined European population
SAS_MAF	Non-reference allele and frequency of existing variant in 1000 Genomes combined South Asian population
AA_MAF	Non-reference allele and frequency of existing variant in NHLBI-ESP African American population
EA_MAF	Non-reference allele and frequency of existing variant in NHLBI-ESP European American population
ExAC_MAF	Frequency of existing variant in ExAC combined population
ExAC_Adj_MAF	Adjusted frequency of existing variant in ExAC combined population
ExAC_AFR_MAF	Frequency of existing variant in ExAC African/American population

Field	Description
ExAC_AMR_MAF	Frequency of existing variant in ExAC American population
ExAC_EAS_MAF	Frequency of existing variant in ExAC East Asian population
ExAC_FIN_MAF	Frequency of existing variant in ExAC Finnish population
ExAC_NFE_MAF	Frequency of existing variant in ExAC Non-Finnish European population
ExAC_OTH_MAF	Frequency of existing variant in ExAC combined other combined populations
ExAC_SAS_MAF	Frequency of existing variant in ExAC South Asian population
CLIN_SIG	Clinical significance of variant from dbSNP
SOMATIC	Somatic status of existing variant(s)
PHENO	Indicates if existing variant is associated with a phenotype, disease or trait
PUBMED	Pubmed ID(s) of publications that cite existing variant
MOTIF_NAME	The source and identifier of a transcription factor binding profile aligned at this position
MOTIF_POS	The relative position of the variation in the aligned TFBP
HIGH_INF_POS	A flag indicating if the variant falls in a high information position of a transcription factor binding profile (TFBP)
MOTIF_SCORE_CHANGE	The difference in motif score of the reference and variant sequences for the TFBP
ENTREZ	Entrez ID
EVIDENCE	Evidence that the variant exists

formatics Pipeline: DNA-Seq Analysis ❯ (../../Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/)

❮ Previous: File Format: MAF (../MAF_Format/)