SOPHiA DDM™ VCF Universal Pipelines

Requirements for Submitting VCF Files

The SOPHiA DDM™ VCF Universal Pipelines enable the analysis and interpretation of variant call format (VCF) files generated by external sequencing and bioinformatics workflows. To ensure consistent and reliable downstream interpretation within the platform, submitted VCF files must adhere to a defined set of formatting and content standards.

This document describes the requirements that VCF files must meet in order to be processed successfully within the SOPHiA DDM™ VCF Universal Pipelines. Please review the specifications below and confirm that your files comply with these requirements before submission.

Table of Contents

1. Supported Variant Types

A single VCF may contain:

- **SNV/INDEL**
- **CNV**
- **Fusions**

A single VCF may contain any combination of the variant types; for instance only SNV/INDELS, SNV/INDELS and CNVs etc.

2. File Format and Compression

All uploaded VCF files must adhere to the VCF 4.2 standard; files that do not conform may not be parsed or fully parsed by the platform. For full details, please refer to the official specification.

2.1. Supported compression

- Uncompressed (flat) VCF
- .gz
- .bz

- Gzip/bzip-compressed files **without** the compression suffix

2.2. Not supported

- .zip archives

3. Reference Genome Assemblies

Supported reference assemblies:

- **hg19** (GCF_000001405.25)
- **hg38** (GCF_000001405.26)

4. VCF Structure Requirements

4.1 Required VCF columns

As per VCF specification, a minimal accepted VCF structure consists of:

- A column name header **#CHROM POS REF ALT QUAL FILTER INFO**
- At least the 7 columns (CHROM, POS, REF, ALT, QUAL, FILTER, INFO)
- Additionally, FORMAT and single-sample name is supported (CHROM, POS, REF, ALT, QUAL, FILTER, INFO, FORMAT, [SAMPLE_NAME])
- All columns included in VCF and defined in the header need to be populated, i.e. if the field is empty is has to be populated with a '.' in accordance with standard VCF specification.

4.2 Chromosome naming (CHROM field)

CHROM must match one of the accepted representations below:

- **seqNum:** 1–25
- **seqName**: 1–22, X, Y, MT
- **fullSeqName:** chr1–chr22, chrX, chrY, chrMT
- Mix of the above conventions in a single VCF are supported

Records with other chromosome representations are treated as unrecognized and will be written to unknown.vcf (see below).

4.3 Multi-sample VCF handling

The pipeline does not provide a full support for multi-sample VCF. Multi-sample VCF should be **split the VCF by sample** and uploaded each as a separate sample.

5. Variant Type Classification

The pipeline assigns each record to exactly one variant type: SHORT, CNV, FUSION, or UNK (unknown). Each variant type is displayed in a separate tab in the SOPHiA DDM™ Platform. Variant type UNK stores all records that could not be assigned to other variant types. Variants assigned to UNK are not visible in the platform interface, but are easily accessible through a downloadable file stored together with other outputs of the pipeline.

5.1 SNV/INDEL

Requirements to correctly identify SNV/INDEL:

- Field TYPE has one of the following values: INDEL, DELIN, SNP, SNV, MNP, SHORT, IVS8-polyT and Column ALT matches standard sequence like A, T, ACG, etc (exact regular expression: r'[actgnACTGN]+').
- Additional SOPHiA DDM™ output-specific flags are also supported: BOLAND or MSH2_AUSTRALIAN, PolyTGT (exact regular expression: r'^\d+\|\d+\,\d+\|\d+$')
- Following cases are not supported for SNV/INDEL:
    - Variants with special tags in REF or ALT such as <NON_REF>
    - Records with N in REF (these are not supported for consequence calculation)

Some downstream computations (e.g., zygosity inference) may require depth-related INFO/FORMAT tags. If present, they should follow common conventions (DP, AD, DP4).

Recommendation: Encode ref and alt counts in DP4 to support robust zygosity inference in SOPHiA DDM™ Platform.

Example SNV/INDEL entries:

#CHROM POS ID REF ALT QUAL FILTER INFO

1 17023743 1 A G 2975.00 PASS DP=107;DP4=1,0,106,0;AD=1,106

1 17027493 2 T G 2227.00 PASS DP=78;DP4=0,0,78,0;AD=0,78

5.2 CNV

Requirements to correctly identify CNVs:

- Field TYPE has one of the following values: CNV, <CNV>, MCNV, DUP, <DUP>, INS, <INS>, DEL, <DEL>, CNV:ROH, <CNV:ROH>, INV, <INV> .
- Column ALT has one of the following values: CNV, <CNV>, MCNV, DUP, <DUP>, INS, <INS>, DEL, <DEL>, CNV:ROH, <CNV:ROH>, INV, <INV> .

In addition, the following INFO field keys are required:

- END=

Example CNV entries:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample
1	820966	.	N	<DEL>	11.04	.	SVTYPE=DEL;SVLEN=-2095;CHREND=1;END=823061	VAF	1.0
1	1327000	.	N	<DUP>	292.6	.	IMPRECISE;CHREND=1;END=1345000;SVLEN=18000	VAF	1.0

5.3 Fusions

Requirements to correctly identify Fusions:

- Field TYPE has one of the following values: BND, FUSION, gene.fusion
- Column ALT has one of the following values: BND, FUSION, gene.fusion (matching regular expression: r'\[|\]'\)

In addition, the following INFO field keys are required:

- MATEID= (each variant needs the two mates .0 and .1 - see required notation in the example)

Example Fusion entries:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO

1	2561480	BND_15072531752972661137.0	N	[6:168911195[N	0.000474594	PASS	SVTYPE=BND;MATEID=BND_15072531752972661137.1
1	7564821	BND_8110842468990017744.0	N	N]15:43441575]	0.000880755	PASS	SVTYPE=BND;MATEID=BND_8110842468990017744.1

5.4 Unknown / Unrecognized variants

A record is routed to **Unknown/Unrecognized** when it meets any of these conditions:

- **Cannot be classified into appropriate variant type:**
  - Does not match any assignment rule for SNV/INDEL, CNV, or FUSION
- **Fails required-field validation for its matched type:**
- Examples (non-exhaustive):
  - CNV-like record missing **INFO/END**
  - Fusion-like record missing **INFO/MATEID** or missing a mate record
  - SNV/INDEL-like record using symbolic ALT such as <NON_REF>
  - REF contains N for SNV/INDEL variants

Unknown variants are **not processed** in the standard annotation pipeline and not displayed in the UI. They are collected into **unrecognized.vcf**, which is downloadable from the UI for the analysis run.