# Variant Schema

A high level representation of the variant looks like this:

| | |
|---|---|
| **id**<br>*String* | Unique variant ID, this consists of chromosome, position, reference and alternate alleles in this format: *chrom:pos:ref:alt* |
| **names**<br>*List<String>* | Other IDs found for this genomic variant across all VCF files indexed |
| **chromosome**<br>*String* | The chromosome where the genomic variant is located |
| **start**<br>*int* | The 1-based position where the genomic variant starts. For variants coming from VCF files, this position is likely to be normalised, in this case, the original call in the file is stored in *studies.files.call* (see below) |
| **end**<br>*int* | The 1-based position where the genomic variant ends. For variants coming from VCF files, this position is likely to be normalised, in this case, the original call in the file is stored in *studies.files.call* (see below) |
| **reference**<br>*String* | Reference allele. For variants coming from VCF files, this position is likely to be normalised, in this case, the original call in the file is stored in *studies.files.call* (see below) |
| **alternate**<br>*String* | Alternate allele. For variants coming from VCF files, this position is likely to be normalised, in this case, the original call in the file is stored in *studies.files.call* (see below) |
| **strand**<br>*String* | Reference strand for this variant, by default all variants are represented in the positive strand |
| **length**<br>*int* | Length of the genomic variation which depends on the variant type |
| **type**<br>*VariantType* | Type of variant, the accepted types and Sequence Ontology (SO) terms are:<br><br>**SNV** SO:0001483<br>**SNP** SO:0000694<br>**MNV** SO:0002007<br>**MNP** SO:0001013<br>**INDEL** SO:1000032<br>**INSERTION** SO:0000667<br>**DELETION** SO:0000159<br>**TRANSLOCATION** SO:0000199<br>**INVERSION** SO:1000036<br>**CNV** SO:0001019<br>**DUPLICATION** SO:1000035<br>**BREAKEND** NA<br>**SYMBOLIC** NA |

**sv**

*StructuralVariation*

Specific information for Structural Variants

| | |
|---|---|
| **ciStartLeft**<br>*int* | The confidence interval around START for imprecise variants - left |
| **ciStartRight**<br>*int* | The confidence interval around START for imprecise variants - right |
| **ciEndLeft**<br>*int* | The confidence interval around END for imprecise variants - left |
| **ciEndRight**<br>*int* | The confidence interval around END for imprecise variants - right |
| **copyNumber**<br>*int* | Number of copies for CNV variants |
| **leftSvInsSeq**<br>*String* | Left inserted sequence for long INSERTIONS |
| **rightSvInsSeq**<br>*String* | Right inserted sequence for long INSERTIONS |
| **type**<br>*StructuralVariantType* | Structural variant types and SO terms are:<br>**COPY_NUMBER_GAIN** SO:0001742<br>**COPY_NUMBER_LOSS** SO:0001743<br>**TANDEM_DUPLICATION** SO:1000173 |
| **breakend**<br>*Breakend* | **mate** *BreakendMate*<br>**chromosome** The chromosome of the mate variant<br>**position** The position of the mate variant<br>**ciPositionLeft** The confidence interval around BREAKEND position - left<br>**ciPositionRight** The confidence interval around BREAKEND position - right<br>**orientation** *BreakendOrientation*<br>**SE** Start - End<br>**t[p[** piece extending to the right of **p** is joined after **t**<br>**SS** Start - Start<br>**t]p]** reverse comp piece extending left of **p** is joined after **t**<br>**ES** End - Start<br>**]p]t** piece extending to the left of **p** is joined before **t**<br>**EE** End - End<br>**[p[t** reverse comp piece extending the right of **p** is joined before **t**<br>**insSeq** *String*<br>Sequence inserted between the two breakends |

**studies**

*List<StudyEntry>*

Information specific to each study the variant was read from, such as samples or statistics

| | |
|---|---|
| **studyId**<br>*String* | Unique ID for the study |

<table>
<tr><td>secondaryAlternates<br><br>List<AlternateCoordinate></td><td>All alternate alleles that have been indexed along with a variant alternate<br><table>
<tr><td>chromosome<br><br>String</td><td>The chromosome where the genomic variation occurred</td></tr>
<tr><td>start<br><br>int</td><td>First position 1-based of the alternate</td></tr>
<tr><td>end<br><br>int</td><td>End position 1-based of the alternate</td></tr>
<tr><td>reference<br><br>String</td><td>Reference allele</td></tr>
<tr><td>alternate<br><br>String</td><td>Alternate allele</td></tr>
<tr><td>type<br><br>VariantType</td><td>Type of variant</td></tr>
</table></td></tr>
<tr><td>files<br><br>List<FileEntry></td><td>List of files from the study where the variant was present<br><table>
<tr><td>fileId<br><br>String</td><td>Unique ID of the indexed file</td></tr>
<tr><td>call<br><br>OriginalCall</td><td>Original call in the VCF file, this is filled when the variant has been normalised<br><table>
<tr><td>variantId</td><td>Original call position for the variant, if the file was normalised</td></tr>
<tr><td>alleleIndex</td><td>Alternate allele index of the original multi-allellic variant call</td></tr>
</table></td></tr>
<tr><td>data<br><br>Map<String, String></td><td>File related data that depend on the format of the file the variant was initially read from</td></tr>
</table></td></tr>
<tr><td>sampleDataKeys<br><br>List<String></td><td>Specifies the sample data keys for each sample data (see below). The first key is always genotype (GT).</td></tr>
<tr><td>samples<br><br>List<SampleEntry></td><td>Sample-related data, each element is related to one sample and contains the specific information for one sample<br><table>
<tr><td>sampleId<br><br>String</td><td>Unique sample ID</td></tr>
<tr><td>fileIndex<br><br>int</td><td>The relative index position in files kist where this sample was loaded</td></tr>
<tr><td>data<br><br>List<String></td><td>Sample data, field GT is always the first one. The order and length must match sampleDataKeys field</td></tr>
</table></td></tr>
</table>

| | |
|---|---|
| **stats**<br><br>*List<VariantStats>* | Variant stats for each variant in the different cohorts, it contains the following fields: |

| | |
|---|---|
| **cohortId**<br>*String* | Unique cohort identifier within the study. |
| **sampleCount**<br>*int* | Count of samples with non-missing genotypes in this variant from the cohort.<br>This value is used as denominator for genotypeFreq. |
| **fileCount**<br>*int* | Count of files with samples from the cohort that reported this variant.<br>This value is used as denominator for filterFreq. |
| **alleleCount**<br>*int* | Total number of alleles in called genotypeCounters. It does not include missing alleles.<br>This value is used as denominator for refAlleleFreq and altAlleleFreq. |
| **refAlleleCount**<br>*int* | Number of reference alleles found in this variant. |
| **refAlleleFreq**<br>*float* | Reference allele frequency calculated from refAlleleCount and alleleCount, in the range [0,1] |
| **altAlleleCount**<br>*int* | Number of main alternate alleles found in this variants. It does not include secondary alternates. |
| **altAlleleFreq**<br>*float* | Alternate allele frequency calculated from altAlleleCount and alleleCount, in the range [0,1] |
| **missingAlleleCount**<br>*int* | Number of missing alleles. |
| **missingGenotypeCount**<br>*int* | Number of genotypes with all alleles missing (e.g. ./.). It does not count partially missing genotypes like "./0" or "./1". |
| **genotypeCount**<br><br>*Map<String, int>* | Number of occurrences for each genotype.<br>This does not include genotype with all alleles missing (e.g. ./.), but it includes partially missing genotypes like "./0" or "./1".<br>Total sum of counts should be equal to the count of samples. |
| **genotypeFreq**<br>*Map<String, float>* | Genotype frequency for each genotype found calculated from the genotypeCount and samplesCount, in the range [0,1] |
| **maf**<br>*float* | Minor allele frequency. Frequency of the less common allele between the reference and the main alternate alleles.<br>This value does not take into acconunt secondary alternates. |
| **mafAllele**<br>*String* | Allele with minor frequency. |
| **mgf**<br>*float* | Minor genotype frequency. Frequency of the less common genotype seen in this variant.<br>This value takes into account all values from the genotypeFreq map. |
| **mgfGenotype**<br>*String* | Genotype with minor frequency. |
| **filterCount**<br>*Map<String, int>* | The number of occurrences for each FILTER value in files from samples in this cohort reporting this variant.<br>As each file can contain more than one filter value (usually separated by ';'), the total sum of counts could be greater than to the count of files. |
| **filterFreq**<br>*Map<String, float>* | Frequency of each filter calculated from the filterCount and filesCount, in the range [0,1] |
| **qualityCount**<br>*int* | The number of files from samples in this cohort reporting this variant with valid QUAL values.<br>This value is used as denominator to obtain the qualityAvg |
| **qualityAvg**<br>*float* | The average Quality value for files with valid QUAL values from samples in this cohort reporting this variant.<br>Some files may not have defined the QUAL value, so the sampling could be less than the filesCount. |

<table>
<tr><td></td><td>
<table>
<tr><td>scores<br><br>List<VariantScore></td><td>Analysis scores such as GWAS precomputed and indexed<br><br>
<table>
<tr><td>id<br><br>String</td><td>Variant score ID</td></tr>
<tr><td>cohort1<br>String</td><td>The main cohort used for calculating this score</td></tr>
<tr><td>cohort2<br><br>String</td><td>The optional secondary cohort used for calculating the score</td></tr>
<tr><td>score<br><br>float</td><td>Score value</td></tr>
<tr><td>pValue<br><br>float</td><td>Score p-value</td></tr>
</table>
</td></tr>
<tr><td>issues<br><br>List<IssueType></td><td>Issues found in this variant for a specific sample in this study<br><br>
<table>
<tr><td>type<br><br>IssueType</td><td>Issues can have one of these types:<br><br>
<table>
<tr><td>DUPLICATION</td></tr>
<tr><td>DISCREPANCY</td></tr>
<tr><td>MENDELIAN_ERROR</td></tr>
<tr><td>DE_NOVO</td></tr>
</table>
</td></tr>
<tr><td>sample<br><br>SampleEntry</td><td>The sample information containing sampleId, fileIndex and data (see above)</td></tr>
</table>
</td></tr>
</table>
</td></tr>
<tr><td>annotation</td><td>Variant Annotation object, this is a large data model and is documented independently</td></tr>
</table>

In the next section you can find the variant annotation schema