Variant Data Model Schema

id *String*	Unique variant ID, this consists of chromosome, position, reference and alternate alleles in this format: *chrom:pos:ref:alt*
names *List<String>*	Other IDs found for this genomic variant across all VCF files indexed
chromosome *String*	The chromosome where the genomic variant is located
start *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 *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 *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 *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 *String*	Reference strand for this variant, by default all variants are represented in the positive strand
length *int*	Length of the genomic variation which depends on the variant type
type *VariantType*	Type of variant, the accepted types and Sequence Ontology (SO) terms are: **SNV** SO:0001483 **SNP** SO:0000694 **MNV** SO:0002007 **MNP** SO:0001013 **INDEL** SO:1000032 **INSERTION** SO:0000667 **DELETION** SO:0000159 **TRANSLOCATION** SO:0000199 **INVERSION** SO:1000036 **CNV** SO:0001019 **DUPLICATION** SO:1000035 **BREAKEND** NA **SYMBOLIC** NA

sv

StructuralVariation

Specific information for Structural Variants

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

<table>
<tr><td>secondaryAlternates

List<AlternateCoordinate></td><td>All alternate alleles that have been indexed along with a variant alternate
<table>
<tr><td>chromosome

String</td><td>The chromosome where the genomic variation occurred</td></tr>
<tr><td>start

int</td><td>First position 1-based of the alternate</td></tr>
<tr><td>end

int</td><td>End position 1-based of the alternate</td></tr>
<tr><td>reference

String</td><td>Reference allele</td></tr>
<tr><td>alternate

String</td><td>Alternate allele</td></tr>
<tr><td>type

VariantType</td><td>Type of variant</td></tr>
</table></td></tr>
<tr><td>files

List<FileEntry></td><td>List of files from the study where the variant was present
<table>
<tr><td>fileId

String</td><td>Unique ID of the indexed file</td></tr>
<tr><td>call

OriginalCall</td><td>Original call in the VCF file, this is filled when the variant has been normalised
<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

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

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

List<SampleEntry></td><td>Sample-related data, each element is related to one sample and contains the specific information for one sample
<table>
<tr><td>sampleId

String</td><td>Unique sample ID</td></tr>
<tr><td>fileIndex

int</td><td>The relative index position in files kist where this sample was loaded</td></tr>
<tr><td>data

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

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

List<VariantScore></td><td>Analysis scores such as GWAS precomputed and indexed

<table>
<tr><td>id

String</td><td>Variant score ID</td></tr>
<tr><td>cohort1
String</td><td>The main cohort used for calculating this score</td></tr>
<tr><td>cohort2

String</td><td>The optional secondary cohort used for calculating the score</td></tr>
<tr><td>score

float</td><td>Score value</td></tr>
<tr><td>pValue

float</td><td>Score p-value</td></tr>
</table>
</td></tr>
<tr><td>issues

List<IssueType></td><td>Issues found in this variant for a specific sample in this study

<table>
<tr><td>type

IssueType</td><td>Issues can have one of these types:

<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

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>