Variant Data Model

Warning

This is a work in progress documentation for v2.0.0

Table of Contents:

Overview

Genomic variant data model plays a crucial role not only in OpenCGA but also in OpenCB suite. Variant data model provides a generic way of representing any variant with any other interesting information associated with it. Variant data model is heavily used in OpenCGA when loading VCF files or when exporting query results. Variant data model is implemented in OpenCB Biodata project, this allows the rest of OpenCB projects such as CellBase to use it.

Goals

Main goals of variant data model include:

- To represent any type of variant (SNV, INDEL) or structural variant (INSERTION, DELETION, CNV, TRANSLOCATION, ...)
- To support *phased* variants and *non-diploid* organisms.
- To provide a file-format agnostic model of storing genomic variant data from VCF, gVCF, microarrays, ...
- To support rich variant annotations for researchers and clinicians
- To store sample data and stats from different datasets
- To allow customisation of variant scores and annotation

Main Features

Some of the main features of the variant data model include:

- support for any type of genomic variant
- rich and customisable annotation integrated

Data Model Overview

A high level representation of the variant looks like this, only main categories are shown:

Variant Data Model Overview

```
{
    "id": "1:69511:A:G",
    "names": ["rs75062661"],
    "chromosome": "1",
    "start": 69511,
    "end": 69511,
    "strand": "+",
    "length": 1,
    "type": "SNV",
    "reference": "A",
    "alternate": "G",
    "studies": [
        {
            "studyId": "demo@family:corpasome",
                       "secondaryAlternates": [],
            "files": [
                {
                    "fileId": "quartet.variants.annotated.vcf.gz"
                    "call" : {},
                    "data": {
                        "FILTER": "PASS",
                                                "QUAL": "4293.01",
                        "HaplotypeScore": "2.4399",
                        "MQ": "15.47",
                                                ...
```

```
                }
            },
                            ...
        ],
        "sampleDataKeys": ["GT", "AD", "DP", "GQ", "PL"],
        "samples": [
            {
                "sampleId": "ISBM200170115",
                "fileIndex": 0,
                "data": ["1/1", "2,171", "173", "99", "2218,228,0"]
            },
            ...
        ],
        "stats": [
                            {
                                    "cohortId": "ALL",
                                    "sampleCount": 4,
                                    "fileCount": 1,
                                    "alleleCount": 8,
                                    "refAlleleCount": 0,
                                    "refAlleleFreq": 0.0,
                                    "altAlleleCount": 8,
                                    "altAlleleFreq": 1.0,
                                    "genotypeCount": {
                                            "0/0": 0,
                                            "0/1": 0,
                                            "1/1": 4
                                    },
                                    "genotypeFreq": {
                                            "0/0": 0.0,
                                            "0/1": 0.0,
                                            "1/1": 1.0
                                    },
                                    "missingAlleleCount": 0,
                                    "missingGenotypeCount": 0,
                                    "maf": 0.0,
                                    "mafAllele": "A",
                                    "mgf": 0.0,
                                    "mgfGenotype": "0/0",
                                    "filterCount": {
                                            "PASS": 0,
                                            "VQSRTrancheSNP99.90to100.
00": 1
                                    },
                                    "filterFreq": {
                                            "PASS": 0.0,
                                            "VQSRTrancheSNP99.90to100.
00": 1.0
                                    },
                                    "qualityCount": 1,
                                    "qualityAvg":
4293.01
                            },
                            ...
                    ],
        "scores": [],
                    "issues": []
            }
    ],
    "annotation": {
            "id": "rs2691305",
            "chromosome": "1",
            "start": 69511,
            "reference": "A",
            "alternate": "G",
            "hgvs": ["ENST00000335137(ENSG00000186092):c.
421A>G"],
            "displayConsequenceType": "missense_variant",
            "consequenceTypes": [
                    {
                            "geneName": "OR4F5",
```

```
                        "ensemblGeneId": "ENSG00000186092",
                "ensemblTranscriptId": "ENST00000335137",
                        "biotype": "protein_coding",
                "cdnaPosition": 421,
                "cdsPosition": 421,
                "codon": "Aca/Gca",
                        "strand": "+",
                        "transcriptAnnotationFlags": ["CCDS",
"basic"],
                        "exonOverlap": [
                                {
                                        "number": "1/1",
                                        "percentage": 0.108932465
                                }
                        ],
                        "proteinVariantAnnotation": {
                                "uniprotAccession": "Q8NH21",
                                "position": 141,
                    "reference": "THR",
                                "alternate": "ALA",
                                "features": [
                                        {
                                                "id": "IPR017452",
                        "start": 34,
                                                "end": 280,
                        "description": "GPCR, rhodopsin-like, 7TM"
                                        },
                                        ...
                                ],
                                "keywords": ["Cell membrane",
"Complete proteome", "Disulfide bond", ...],
                                "substitutionScores": [
                                        {
                                                "description":
"tolerated",
                                                "score": 0.63,
                                                "source": "sift"
                                        },
                                        {
                                                "description":
"benign",
                                                "score": 0.003,
                                                "source":
"polyphen"
                                        }
                                ],
                        },
                        "sequenceOntologyTerms": [
                                {
                                        "accession": "SO:0001583",
                                        "name": "missense_variant"
                                }
                        ]
                }
        ],
        "conservation": [
                {
                        "score": 1.149999976158142,
                        "source": "gerp"
                },
                {
                        "score": 0.1289999932050705,
                        "source": "phastCons"
                },
                {
                        "score": -0.527999997138977,
                        "source": "phylop"
                }
        ],
        "cytoband": [
                {
```

```
                                "name": "p36.33",
                                "chromosome": "1",
                                "start": 1,
                                "end": 2300000,
                                "stain": "gneg"
                        }
                ],
                "functionalScore": [
                        {
                                "score": -0.7899999618530273,
                                "source": "cadd_raw"
                        },
                        {
                                "score": 0.03999999910593033,
                                "source": "cadd_scaled"
                        }
                ],
                "geneDrugInteraction": [],
                "geneTraitAssociation": [],
                "populationFrequencies": [
                        {
                                "altAllele": "G",
                "altAlleleFreq": 0.84222084,
                "altHomGenotypeFreq": 0.77478045,
                "hetGenotypeFreq": 0.1348808,
                "population": "ALL",
                "refAllele": "A",
                "refAlleleFreq": 0.15777917,
                "refHomGenotypeFreq": 0.090338774,
                "study": "GNOMAD_GENOMES"
                        },
                        {
                                "altAllele": "G",
                "altAlleleFreq": 0.9637507,
                "altHomGenotypeFreq": 0.94847214,
                "hetGenotypeFreq": 0.03055722,
                "population": "NFE",
                "refAllele": "A",
                "refAlleleFreq": 0.03624925,
                "refHomGenotypeFreq": 0.02097064,
                 "study": "GNOMAD_GENOMES"
                        }
                ],
                "repeat": [
                        {
                                "chromosome": "1",
                "copyNumber": 2.0,
                "end": 87112,
                "id": "9119",
                "percentageMatch": 0.992904,
                "source": "genomicSuperDup",
                "start": 10001
                        },
                        {
                                "chromosome": "1",
                                "copyNumber": 2.0,
                                "end": 87112,
                                "id": "14903",
                                "percentageMatch": 0.995437,
                                "source": "genomicSuperDup",
                                "start": 18393
                        }
                ],
                "traitAssociation": [
                        {
                                "additionalProperties": [
                                        {
                                                "name":
"mutationSomaticStatus_in_source_file",

"value": "Confirmed "
```

```
"somatic "

"variant"}],
                                                    "alleleOrigin": [],
                                                    "bibliography": [],
                                                    "ethnicity": "Z",
                                                    "genomicFeatures":
[{"featureType": "gene",

"xrefs": {"symbol": "OR4F5"}},

{"featureType": "gene",

"xrefs": {"symbol": "8301"}}],
                                                    "heritableTraits": [],
                                                    "id": "COSM4144171",
                                                    "somaticInformation":
{"histologySubtype": "neoplasm",

"primaryHistology": "other",

"primarySite": "thyroid",

"sampleSource": "",

"tumourOrigin": ""},
                                                    "source": {"name":
"cosmic"},
                                                    "submissions": []}]}
                                  "additionalAttributes": {
                      "opencga": {
                          "attribute": {
                              "annotationId": "CURRENT",
                              "release": "1"
                          }
                      }
                  }
}
```


Implementation

Variant data model is implemented in OpenCB Biodata project, this allows the rest of OpenCB projects such as CellBase, Oskar to