Genoxus Annotation

Genoxus Labs Team: genoxuslabs@gmail.com

Version 1.0
Mar 31, 2026

Introduction

Genoxus Annotation is a harmonized and curated collection of human genetic variant databases designed to support accurate and salable variant annotation. Variant annotation following genetic testing such as whole genome sequencing (WGS) or whole exome sequencing (WES) is a critical step in identifying and interpreting disease-associated genetic factors. As sequencing technologies continue to generate large volumes of genomic data, robust and well-structured annotation resources are essential for translating raw variant calls into clinically meaningful insights.

Genoxus Annotation v1.0 integrates data from NCBI ClinVar. ClinVar provides curated information on the clinical significance of a broad spectrum of genetic variants—including single nucleotide variants (SNVs), insertions (INS), deletions (DEL), INDELs, copy number variations (CNVs), and structural variants (SVs)—along with their associated diseases and traits. The GWAS Catalog complements this by focusing primarily on SNVs identified through genome-wide association studies, linking common variants to complex diseases and phenotype traits.

By harmonizing variant representations, standardizing disease terminology, and consolidating evidence across sources, Genoxus Annotation provides a unified framework that streamlines variant interpretation for research and clinical applications.

Data Organization

Genoxus Annotation as Open Data on AWS is hosted in a S3 bucket. Multiple folders are created here, each mapping to one chromosome, for example **chr01**, or **chr10**, etc. Within each chromosome folder, there are multiple folders mapped to that chromosome's base pairs(bp) locations. For example, "1-1000000" is mapped to the first 1 million bp, and subsequent folders are mapped with 1 million bp increments.

In each of these folders there is a file called **clinvar.parquet**. This file contains all variants documented by ClinVar for the given bp range. These variants are synchronized with update from NCBI ClinVar.

Instead of describing the ClinVar data schema, a user may simply open any **clinvar.parquet** file to see the data structure. The usage instruction section below also contains 1 data sample to illustrate the data schema.

Usage Instruction

To query ClinVar data, a user should first determine which chromosome and bp location the target variant targets.

For example, for the following variant:

chromosome: 10 bp starts at: 320129 bp ends at: 735607

The target folder is: chr10/1-1000000.

To query variants with possible overlapping in this region, the user may use the following python code:

```
    import duckdb
    parquet_file = "clinvar.parquet"
    conn = duckdb.connect(database=':memory:')

    chromosome = 10
    start_pos = 320129
    stop_pos = 735607
    var_chunk = stop_pos - start_pos
    if var_chunk <= 0:
        print ("Bad variation info, exit...")
        return

    start_abs = start_pos - var_chunk if start_pos > var_chunk else 0
    stop_abs = stop_pos + var_chunk

    print(f"--- Searching for records where the first assembly's Chromosome
is {chromosome} ---")

    # DuckDB uses 1-based indexing for lists/arrays.
    # We also check if the list is not empty to prevent errors.
    query = f"""
    SELECT *
    FROM '{parquet_file}',
        UNNEST(Host_Genes) AS t(gene),
        UNNEST(gene.Location) AS t2(loc)
    WHERE loc.GRCh37_assembly IS NOT NULL
        AND loc.GRCh37_assembly.Chromosome = '{chromosome}'
        AND loc.GRCh37_assembly.Start >= '{start_abs}'
        AND loc.GRCh37_assembly.Stop <= '{stop_abs}';
    """

    print (query)

    # Using fetch_df() is often more convenient for inspection
    results = conn.execute(query).fetchall()

    print (f"Found {len(results)} records")
    for r in results:
        print (r)

    conn.close()
```

This query returns multiple records, for example one of them is the following:

```
{
  "Variant_Identifications": {
    "ID": "57851",
    "Name": "GRCh38/hg38 10p15.3(chr10:180143-550594)x3",
    "VCV_accession": "VCV000057851"
  },
  "Variant_Type": {
    "Type": "copy number gain"
  },
  "Record_Timeline": {
    "Date_Created": "2015-06-28T00:00:00",
    "Date_LastUpdated": "2024-05-08T00:00:00",
    "Date_MostRecentSubmission": "2015-06-28T00:00:00",
    "Current_Status": "current"
  },
  "Reference_Allele": "",
  "Alternate_Allele": "",
  "Host_Genes": [
    {
      "Symbol": "DIP2C",
      "FullName": "disco interacting protein 2 homolog C",
      "ID": "22982",
      "OMIM": "611380",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 274201,
            "Stop": 689668,
            "Strand": "-",
            "Target_Length": 415468
          },
          "GRCh37_assembly": null
        },
        {
          "GRCh38_assembly": null,
          "GRCh37_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 320129,
            "Stop": 735607,
            "Strand": "-",
            "Target_Length": 415479
          }
        }
      ]
    },
    {
      "Symbol": "LOC106783507",
      "FullName": "nonconserved acetylation island sequence 51 enhancer",
      "ID": "106783507",
      "OMIM": "",
```

```
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 470408,
            "Stop": 470923,
            "Strand": "+",
            "Target_Length": 516
          },
          "GRCh37_assembly": null
        }
      ]
    },
    {
      "Symbol": "LOC126860802",
      "FullName": "BRD4-independent group 4 enhancer GRCh37_chr10:294268-
295467",
      "ID": "126860802",
      "OMIM": "",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 248328,
            "Stop": 249527,
            "Strand": "+",
            "Target_Length": 1200
          },
          "GRCh37_assembly": null
        }
      ]
    },
    {
      "Symbol": "LOC126860803",
      "FullName": "CDK7 strongly-dependent group 2 enhancer
GRCh37_chr10:310208-311407",
      "ID": "126860803",
      "OMIM": "",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 264268,
            "Stop": 265467,
            "Strand": "+",
            "Target_Length": 1200
          },
          "GRCh37_assembly": null
        }
      ]
    },
```

```
    {
      "Symbol": "LOC126860804",
      "FullName": "P300/CBP strongly-dependent group 1 enhancer
GRCh37_chr10:369502-370701",
      "ID": "126860804",
      "OMIM": "",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 323562,
            "Stop": 324761,
            "Strand": "+",
            "Target_Length": 1200
          },
          "GRCh37_assembly": null
        }
      ]
    },
    {
      "Symbol": "LOC126860805",
      "FullName": "BRD4-independent group 4 enhancer GRCh37_chr10:390524-
391723",
      "ID": "126860805",
      "OMIM": "",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 344584,
            "Stop": 345783,
            "Strand": "+",
            "Target_Length": 1200
          },
          "GRCh37_assembly": null
        }
      ]
    },
    {
      "Symbol": "LOC126860806",
      "FullName": "MED14-independent group 3 enhancer GRCh37_chr10:409736-
410935",
      "ID": "126860806",
      "OMIM": "",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 363796,
            "Stop": 364995,
            "Strand": "+",
```

```
            "Target_Length": 1200
          },
          "GRCh37_assembly": null
        }
      ]
    },
    {
      "Symbol": "LOC126860807",
      "FullName": "CDK7 strongly-dependent group 2 enhancer
GRCh37_chr10:461462-462661",
      "ID": "126860807",
      "OMIM": "",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 415522,
            "Stop": 416721,
            "Strand": "+",
            "Target_Length": 1200
          },
          "GRCh37_assembly": null
        }
      ]
    },
    {
      "Symbol": "LOC130003153",
      "FullName": "ATAC-STARR-seq lymphoblastoid active region 2891",
      "ID": "130003153",
      "OMIM": "",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 283813,
            "Stop": 283882,
            "Strand": "+",
            "Target_Length": 70
          },
          "GRCh37_assembly": null
        }
      ]
    },
    {
      "Symbol": "LOC130003154",
      "FullName": "ATAC-STARR-seq lymphoblastoid active region 2892",
      "ID": "130003154",
      "OMIM": "",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
```

```
            "Cytogenetic_Location": "10p15.3",
            "Start": 307147,
            "Stop": 307196,
            "Strand": "+",
            "Target_Length": 50
          },
          "GRCh37_assembly": null
        }
      ]
    },
    {
      "Symbol": "ZMYND11",
      "FullName": "zinc finger MYND-type containing 11",
      "ID": "10771",
      "OMIM": "608668",
      "Location": [
        {
          "GRCh38_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 130088,
            "Stop": 254637,
            "Strand": "+",
            "Target_Length": 124550
          },
          "GRCh37_assembly": null
        },
        {
          "GRCh38_assembly": null,
          "GRCh37_assembly": {
            "Chromosome": "10",
            "Cytogenetic_Location": "10p15.3",
            "Start": 180404,
            "Stop": 300576,
            "Strand": "+",
            "Target_Length": 120173
          }
        }
      ]
    }
  ],
  "Protein_Changes": [],
  "Classifications": {
    "GermlineClassification": [
      {
        "Conditions": [
          {
            "Type": "PhenotypeInstruction",
            "Traits": [
              {
                "Name": "See cases",
                "Label": "Preferred"
              }
            ],
```

```
            "Affected_Genes": [],
            "Citations": []
          }
        ],
        "Date_LastEvaluated": "2011-08-12T00:00:00",
        "Date_Created": "2015-06-28T00:00:00",
        "Date_MostRecentSubmission": "2015-06-28T00:00:00",
        "Review_Status": "",
        "Description": "",
        "Clinical_Assertion": {
          "Date_LastUpdated": "2015-06-28T00:00:00",
          "Assertion": "variation to disease",
          "Review_Status": "criteria provided, single submitter",
          "Classification": {
            "GermlineClassification": "Uncertain significance"
          },
          "Reference": {
            "Method": "Kaminsky et al. (Genet Med. 2011)",
            "Citation": [
              "Type",
              "Journals",
              "URLs",
              "Citation_Text"
            ]
          },
          "Asserted_Traits": [
            {
              "Trait": "See cases",
              "Label": "Preferred"
            }
          ]
        }
      }
    ],
    "SomaticClinicalImpact": [
      {
        "Conditions": [
          {
            "Type": "PhenotypeInstruction",
            "Traits": [
              {
                "Name": "See cases",
                "Label": "Preferred"
              }
            ],
            "Affected_Genes": [],
            "Citations": []
          }
        ],
        "Date_LastEvaluated": "2011-08-12T00:00:00",
        "Date_Created": "2015-06-28T00:00:00",
        "Date_MostRecentSubmission": "2015-06-28T00:00:00",
        "Review_Status": "",
        "Description": "",
```

```
        "Clinical_Assertion": {
          "Date_LastUpdated": "2015-06-28T00:00:00",
          "Assertion": "variation to disease",
          "Review_Status": "criteria provided, single submitter",
          "Classification": {
            "GermlineClassification": "Uncertain significance"
          },
          "Reference": {
            "Method": "Kaminsky et al. (Genet Med. 2011)",
            "Citation": [
              "Type",
              "Journals",
              "URLs",
              "Citation_Text"
            ]
          },
          "Asserted_Traits": [
            {
              "Trait": "See cases",
              "Label": "Preferred"
            }
          ]
        }
      ],
      "OncogenicityClassification": [
        {
          "Conditions": [
            {
              "Type": "PhenotypeInstruction",
              "Traits": [
                {
                  "Name": "See cases",
                  "Label": "Preferred"
                }
              ],
              "Affected_Genes": [],
              "Citations": []
            }
          ],
          "Date_LastEvaluated": "2011-08-12T00:00:00",
          "Date_Created": "2015-06-28T00:00:00",
          "Date_MostRecentSubmission": "2015-06-28T00:00:00",
          "Review_Status": "",
          "Description": "",
          "Clinical_Assertion": {
            "Date_LastUpdated": "2015-06-28T00:00:00",
            "Assertion": "variation to disease",
            "Review_Status": "criteria provided, single submitter",
            "Classification": {
              "GermlineClassification": "Uncertain significance"
            },
            "Reference": {
              "Method": "Kaminsky et al. (Genet Med. 2011)",
```

```
            "Citation": [
              "Type",
              "Journals",
              "URLs",
              "Citation_Text"
            ]
          },
          "Asserted_Traits": [
            {
              "Trait": "See cases",
              "Label": "Preferred"
            }
          ]
        }
      }
    ]
  },
  "md5sum": "3528b0661a7ee28249896ed9f0be83ce",
  "Genomic_Location": {
    "Chromosome": "10",
    "Cytogenetic_Location": "10p15.3",
    "Start": 226083,
    "Stop": 596534,
    "Strand": "",
    "Target_Length": 370452
  }
}
```


Roadmap

The following datasets are planned in future releases:

GWAS data is planned to be included in v2 release.

GnomAD data is planned to be included in v3 release.