
• Archive of interpretations of variants relative to conditions
• Variant-level information
• Fully public and freely available
• Submission-driven database
• Curation support from NCBI staff
ClinVar integrates four domains of information

- Variation
- Condition
- Interpretation
- Evidence
Archiving variant interpretations

- Full data archive in XML format
- Released monthly
  - Weekly XML updates are provided but not archived
- Monthly back-up, tape for long term storage
Accessions and versions

- **FAIR principle: Findable**
  - data are assigned globally unique and persistent identifier
  - data are registered or indexed in a searchable resource

- Both submitted and aggregate records have an accession

- Version is incremented when the submitted data is changed
  - e.g. SCV000183753.2, RCV000205831.2

- Each instance of an XML is stored internally, whether the version changed or not
Data standards

- FAIR principle: Interoperable
  - data use vocabularies that follow FAIR principles
- Critical for interoperability and data quality
- Challenge of standards for variation, disease
- Use standards but allow legacy, other names
What is the variant?

- 985A>G
- c.985A>G (p.K304E)
- p.K329E:AAA>GAA
- MC K329e
- ACADM, LYS304GLU
- 607008.0001
- LYS304GLU
- K304E
- K329E
- K333E
- A985G
- c.985A>G
- 985A>G (K304E)
- 985A>G (K329E)
- Analysis of ACADM 985A>G mutation
- K304E (985 A->G)
- K304E (K329E)
- K304E only
- K329E(985A>G)
- Mutation c.985A>G (p.K304E)
- c985A>G
- includes: K304E (985A>G)
- c.985A>G (p.Lys304Glu)
- p.K304E
- p.Lys329Glu
- previously known as p.Lys329Glu)
- Lys304Glu (985A>G)
- NM_000016.5:c.985A>G
- NM_001127328.2:c.997A>G
- NM_000016.4:c.985A>G
- NG_007045.1:g.41804A>G
- NM_001127328.1:c.997A>G
- NP_000007.1:p.Lys329Glu
- NP_001120800.1:p.Lys333Glu
- c.985A>G for p.Lys329Glu
- p.Lys333Glu
- NC_000001.10:g.76226846A>G
- NC_000001.11:g.75761161A>G
What is the disease?

- hereditary breast and ovarian cancer, BROVCA2
- hereditary breast and ovarian cancer, BROVCA1
- Inflammatory breast cancer
- Male breast cancer
- Childhood breast cancer
- Androgen insensitivity, partial, with breast cancer
- Hereditary breast and ovarian cancer syndrome
- Multifocal breast cancer
- breast cancer
- Familial triple-negative breast cancer
- Triple-negative breast cancer
- Hereditary breast and ovarian cancer
- bilateral breast cancer
- lobular breast cancer
- phyllodes breast cancer
- Hereditary breast and ovarian cancer syndrome (HBOC)
## Data standardization

<table>
<thead>
<tr>
<th>Content</th>
<th>Authorities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence variants</td>
<td>HGVS</td>
</tr>
<tr>
<td>Structural variants</td>
<td>ISCN (being developed)</td>
</tr>
<tr>
<td>Accessions for the variant location</td>
<td>dbSNP, dbVar</td>
</tr>
<tr>
<td>Genes</td>
<td>HGNC</td>
</tr>
<tr>
<td>Conditions</td>
<td>Orphanet: group terms</td>
</tr>
<tr>
<td></td>
<td>OMIM: disease-specific terms</td>
</tr>
<tr>
<td></td>
<td>Human phenotype ontology: clinical features</td>
</tr>
<tr>
<td>Reference sequence</td>
<td>Assembly: Genome Reference Consortium (GRC)</td>
</tr>
<tr>
<td></td>
<td>Gene-specific: RefSeqGene/LRG</td>
</tr>
<tr>
<td>Type of variation, location in gene</td>
<td>Sequence ontology</td>
</tr>
<tr>
<td>Variant effects</td>
<td>VAriO, Sequence ontology</td>
</tr>
<tr>
<td>Clinical significance</td>
<td>ACMG</td>
</tr>
</tbody>
</table>
Required data

• FAIR Principle: Reusable
  • meta(data) are richly described with a plurality of accurate and relevant attributes
  • (meta)data are associated with detailed provenance

•Submitter information
• Variant description
  • Assembly
• Interpreted condition
• Interpretation
• Collection method
• Allele origin
• Affected Status

• Date last evaluated
• More evidence - comment, citation, counts of individuals/families
Data access

- **FAIR principle: Accessible**
  - data are retrievable by their identifier using a standardized communications protocol
  - the protocol is open, free, and universally implementable

- **Programmatic**
  - E-utilities
  - Planned development: RESTful APIs

- **Monthly full releases**
  - Comprehensive XML extraction
  - VCF files
  - Tab-delimited summary files, e.g. genes, variants, conflicts

- **Website**
What data to preserve?

• Community input
  • Medical Genetics Working Group
  • User/submitter groups, e.g. ClinGen
  • Informatics users
  • User mail/conferences
  • Surveys

• Usage statistics
Deciding not to continue to preserve data

• SKY-CGH
  • Detailed cytogenetic analyses of tumor-vs.-normal cancer samples
  • Data still available on NCBI’s ftp site, will be available in dbVar at NCBI

• Peptidome
  • Archived tandem mass spectrometry peptide and protein identification data
  • Data still available on NCBI’s ftp site and at the PRoteomics IDEntifications (PRIDE) database at EBI
Acknowledgements

- Mark Benson
- Garth Brown
- Chao Chen
- Shanmuga Chitipiralla
- Baoshan Gu
- Jennifer Hart
- Douglas Hoffman
- Wonhee Jang
- Brandi Kattman
- Ken Katz
- Jennifer Lee

- Zenith Maddipatla
- Donna Maglott
- Adriana Malheiro
- Michael Ovetsky
- George Riley
- Wendy Rubinstein
- Amanjeev Sethi
- Ray Tully
- Ricardo Villamarin
- George Zhou

- Steve Sherry
- Jim Ostell
- David Lipman