ClinVar Standards

Overall Goals and Guidelines

- The goal of standardization of data elements is to facilitate data exchange with an emphasis on enabling the exchange of data between computational systems. Data elements will therefore be governed by community accepted ontologies and controlled vocabularies whenever possible.
- ClinVar standards must be rigorous enough to enable data exchange while flexible enough to accommodate ambiguity as well as scientific and technological change.
- While standardization of data elements is highly desired, it is a high priority to allow submissions from groups even if they are not able to express data elements within the constraints of standards.

Data Quality

ClinVar aims to provide a clinical grade mutation database. Efforts should be made to ensure that the data submitted to and presented by the resource of the highest quality. However, it is outside of the scope of this resource to set strict quality requirements for variant observations and clinical assertions submitted to the resource . Rather, ClinVar will aim to capture as much data as possible and make it available to data consumers to allow them to make independent judgments on the quality of assertions.

Variant observation

Variant observations ideally will include a PHRED score for the variant call.
In the absence of a PHRED score, some metric of expected false positive rate inherent to a given detection technology is recommended.

Clinical Assertions

Clinical assertions will be categorized in a way that communicates the relative level of confidence in the assertion:

- Uncurated – large datasets on phenotyped cohorts
- Curated by single entity – clinical labs, ClinVar curators
- Expert curation – model curation projects
- Guideline – practice guidelines

Standards Recommendations

- Variants to be identified using HGVS nomenclature (http://www.hgvs.org/mutnomen/recs.html#general). HGVS guidelines will be followed except where otherwise noted below. Redundantly, some guidelines are:
    - Variants to be reported at a DNA level in relation to a reference sequence. NCBI reference sequence accessions and versions are the preferred reference sequence identifiers, though CCDS, UCSC or ENSEMBL identifiers are acceptable. Genome build must be clearly specified if not directly implied from reference sequence identifier.

    - Genomic coordinates preferred, particularly in cases where a nucleotide variant may have several different consequences due to it appearing in multiple contexts (e.g. multiple splice isoforms)
        - NOTE: this differs from HGVS recommendation to use coding sequence coordinates whenever possible.
    - Genes to be described using HGNC gene symbols
    - Variants to be reported at a DNA coordinates in the genomic, gene or transcript context; protein sequence context may be supplied but must be accompanied by a DNA position
- Sequence Ontology (http://www.sequenceontology.org) terms will be used to specify variation type, molecular consequence, molecular context and functional consequence.
- Phenotypes, diseases and clinical terms will be described using standard ontologies wherever possible, though accommodation must be made for conditions that have no sanctioned term. SNOMED CT is the preferred ontology with acceptable entries coming from UMLS, MeSH, HPO, OMIM, ICD-9 and ICD-10.
    - ClinVar will present data using its preferred vocabulary. Submissions will be preserved in their original form with some translation occurring at NCBI.
- In addition to standardized terms, testing indications may be submitted as free text that is not governed by any vocabulary. This is done in recognition of the high value included in these data.
- Variant classifications
    - Variants will be initially be classified using the following categories:
        - Pathogenic
        - Likely pathogenic
        - Uncertain
        - Likely benign
        - Benign
    - Criteria for membership in these categories will be defined and redefined on an ongoing basis in response to community feedback.