Clinical Genomics Data Infrastructure and ClinVar

U. Geigenmüller[1], D. Maglott[2], S. Aradhya[3], S. Bale[3], P.R. Billings[4], C. Braastad[5], M. Eisenberg[6], M.J. Ferber[7], K. Fuentes Fajardo[8], M. Hegde[9], B. Kattman[2], S.F. Kingsmore[10], I.S. Kohane[11], D.H. Ledbetter[12], K. Lee[11], E. Lyon[13], C. Lese Martin[9], N.A. Miller[10], J. Ostell[2], J. Paschall[2], H.L. Rehm[14], G.R. Riley[2], C.J. Saunders[10], S.T. Sherry[2], E.D. Trautman[6], V. Zvereff[6], D.M. Margulies[15]

1) Channing Laboratory, Brigham and Women's Hospital, Boston, MA; 2)National Center for Biotechnology Information, NLM, NIH, Bethesda, MD; 3) GeneDx, Gaithersburg, MD; 4) LIFE Technologies, Carlsbad, CA; 5) Athena Diagnostics, Worcester, MA; 6) LabCorp/CMBP, Raleigh, NC; 7) Mayo Clinic, Rochester, MN; 8) NIH/NHGRI/OCD/Undiagnosed Diseases Program, Bethesda, MD; 9) Department of Human Genetics, Emory University School of Medicine, Atlanta, GA; 10) Children's Mercy Hospital, Kansas City, MO; 11) Center for Biomedical Informatics, Harvard Medical School, Boston, MA; 12) Geisinger Health System, Danville, PA; 13) University of Utah Pathology Department/ARUP Laboratories, Salt Lake City, UT; 14) Harvard Medical School, Boston, MA; 15) Correlagen Diagnostics, Waltham, MA.

Problem Statement

Adoption of large-scale sequencing in molecular diagnostics is limited by difficulty of assessing the clinical significance of detected variants.

Limited accessibility of variant observation data

- Primary data are fragmented across publications, locus-specific databases, and proprietary databases and thus hard to find or completely inaccessible.
- Lack of standardization of data structures hinders meta analysis of available data.

Inconsistent variant interpretations

- Varying sets of public data and local primary data used to derive clinical significance of the same variant
- Varying assessment methods used to derive clinical significance from primary data

Variant interpretation not scalable

- Data searches and interpretations largely manual and prohibitively labor intensive

Community experience and expertise not shared

- No mechanism for sharing interpretive decisions widely across laboratories

Project Goal

To develop a centralized, freely accessible, clinical-grade database that …

Increases data accessibility and consistency of use

- Access to all relevant data from both public and private sources

Promotes consistent data use

- Defined data structure enables meta-analysis
- Database checks on accuracy and internal consistency of variant description

Enables scalable variant analysis

- All data on a given variant aggregated in one place
- Data mining by manual community effort and automated text mining
- Data stored and retrievable in computable format

Provides confidence measure for clinical interpretation

- Multiple clinical interpretations stored for each variant, with each interpretation tagged by submitter and algorithm used

  → **Divergent interpretations indicate uncertainty**
- Clinical interpretations classified by curation level

Promotes currency of clinical interpretation

- Alert issued if new data or new interpretations for a variant become available

  → **Revised interpretation can be issued**

Enables development and calibration of interpretive algorithms

ClinVar Database

- ClinVar: created and maintained by NCBI (http://www.ncbi.nlm.nih.gov/clinvar – see Poster *Clinical genetics resources at NCBI: ClinVar and ISCA support evidence-based interpretation of human variation*)
- Controlled vocabulary used for description of variants and associated phenotypes (Table 1).

  Given an identifier, ClinVar will calculate other identifiers as applicable and check submission for internal consistency

Table 1: Variant and phenotype descriptors (by submission)

Variant description	Chromosomal co-ordinates: NCBI identifiers preferred, CCDS, UCSC, or ENSEMBL acceptable
	Coding co-ordinates: HGVS identifiers [1]
	Variant type: Sequence Ontology identifiers [2]
Associated phenotypes	SNOMED CT preferred, UMLS, MeSH, HPO, OMIM, ICD-9, ICD-10, free text acceptable

- Aim is to capture as much data as possible and allow users to make independent judgments on the quality of assertions (Table 2)

Table 2: Data quality (by data source)

Variant detection method	Detection platform
	PHRED score for the variant call and/or metric of technology-specific false positive rate
Phenotype assertion method	Symptoms preferred, test indication acceptable
Type of experimental system	Type of model (animal, cell based, *in-vitro*)

- Observational data will be categorized to allow meta-analysis and use in automated rules-based scoring algorithms (Table 3).

Table 3: Observational data (by data source)

Unrelated unaffected individuals AND Unrelated affected individuals	# individuals tested
	Ethnicity of individuals tested
	Age group of individuals tested
	# variant occurrences
Affected families	# affected with variant
	# related affected without variant
	# variant occurrences in affected
	# *de-novo* variant occurrences in affected
	# observed transmissions of variant allele from het parent
	# related unaffected with variant (zygosity consistent with mode of inheritance), by age group
Experimental system	Effect size

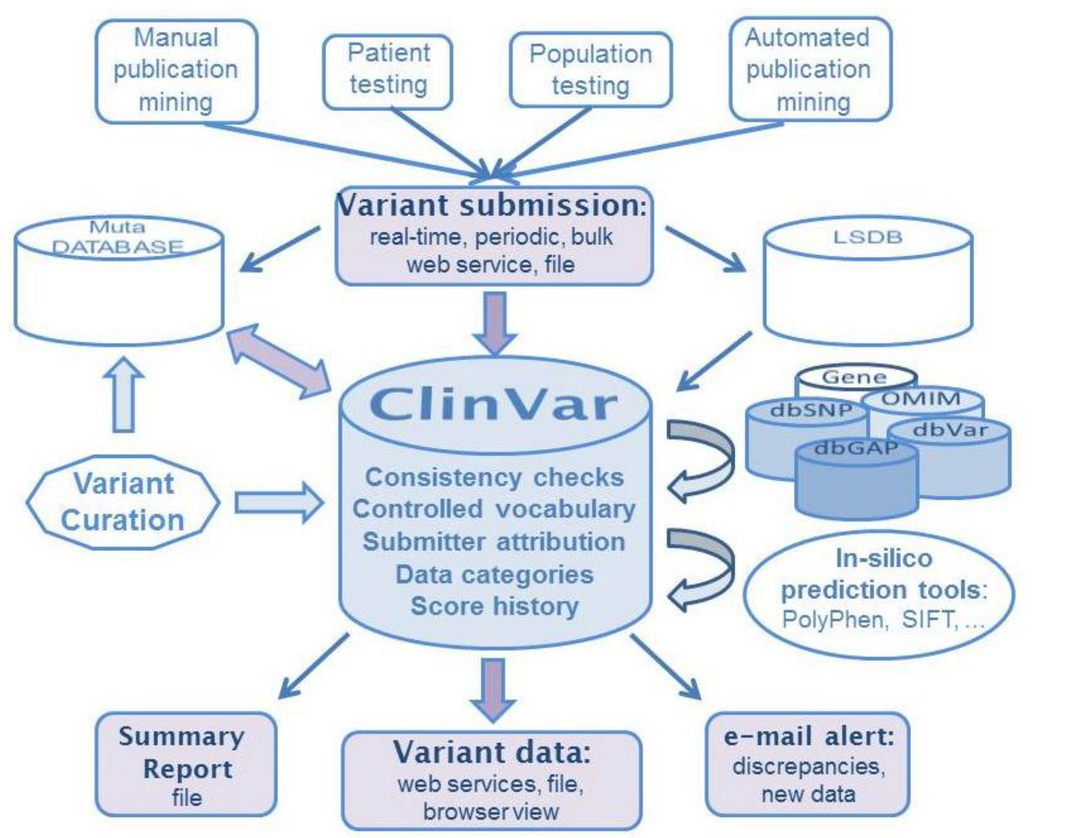
- Clinical assertions will be categorized as uncurated or curated by registered submitter, by expert-panel, or by practice guideline

Table 4: Clinical assertion (by submission)

Clinical assertion	Variant classification category
	Method of variant classification
	Curation level

Variant Data Flow

Integration of ClinVar with other databases and tools



Status

- 36 labs that have agreed to contribute sequence-level data
- 20,000 variants loaded
- Position papers about policies for participation, sustainability model, ClinVar data elements, standards, and technical methods available on http://www.ncbi.nlm.nih.gov/clinvar/community/

Mailing List Address: clinvar@ncbi.nlm.nih.gov

References

1. Human Genome Variation Society: Nomenclature for the description of sequence variants http://www.hgvs.org/mutnomen/
2. The Sequence Ontology Project. http://www.sequenceontology.org/