ClinVar

What is ClinVar

- ClinVar is a freely available, central archive for associating observed variation with supporting clinical and experimental evidence for a wide range of disorders.
- The database can be used interactively or incorporated into variant analysis pipelines using files from the FTP site or APIs.
- ClinVar is submission driven and welcomes submissions from clinical and research laboratories, locus-specific databases and authors of publications.

Why develop ClinVar

- Capacity for identification of human variation has increased
- Cost of manual evaluation of novel variation is not scalable
- Not all variation is published in the literature
- There was a need for
  - a central registry for representation of clinically relevant variation
  - details of how any variation was called, in what population(s), how often it was observed, how often it was not observed
  - links to practice guidelines, research publications, and databases with additional information
  - to update and keep a history of evaluations of clinical validity
  - To not require prior publication

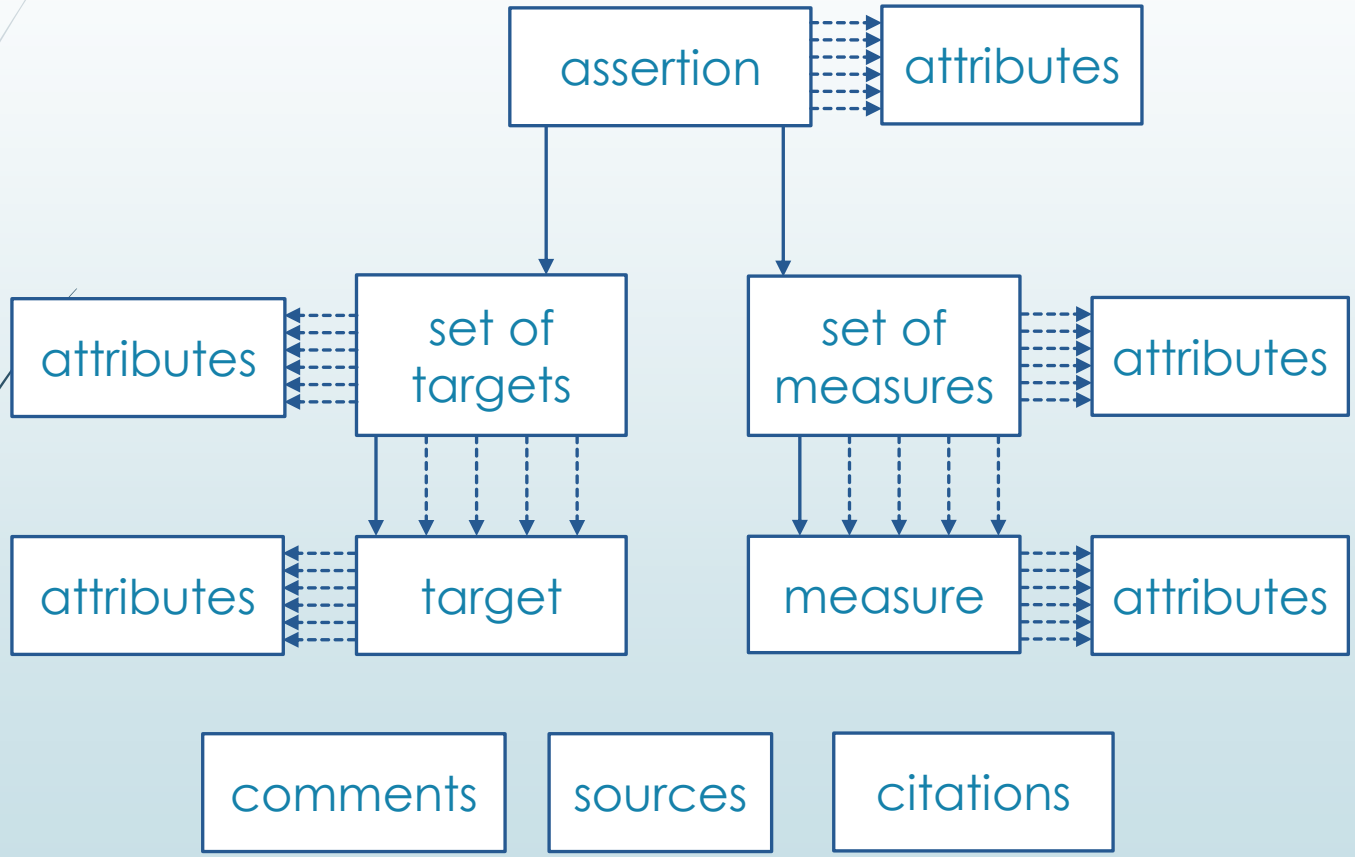
ClinVar functions

- **An archive:** Provides accessioned and versioned reports of the relationships among human variation and human phenotypes with attribution to sources and publications
- **A reporter**:
  - Submitted interpretations (clinical significance, clinical validity)
  - Review status of/confidence in a report of clinical significance
  - Evidence of the genotype/phenotype relationship
  - Standard terminology and usage
- **An aggregator:** Reports related information from Gene, RefSeq, dbSNP, dbVar, Genetic Testing Registry (GTR), *GeneReviews*, phenotype databases and more…
- **A data service**: Provides interactive and programmatic access to current and previous records.

ClinVar database design

- ClinVar design is meta data driven with generic tables that can represent a wide range of data types and relationships.
- Central concepts are the 'target' and 'measure' representing the types of things that may be tested for (e.g. disease) and what is actually tested (e.g. the variant).
- ClinVar uses common tables to manage different types of relationships.
  - Between genes and variation
  - History of merges of data elements
  - Between variants/genes and conditions

ClinVar database concept



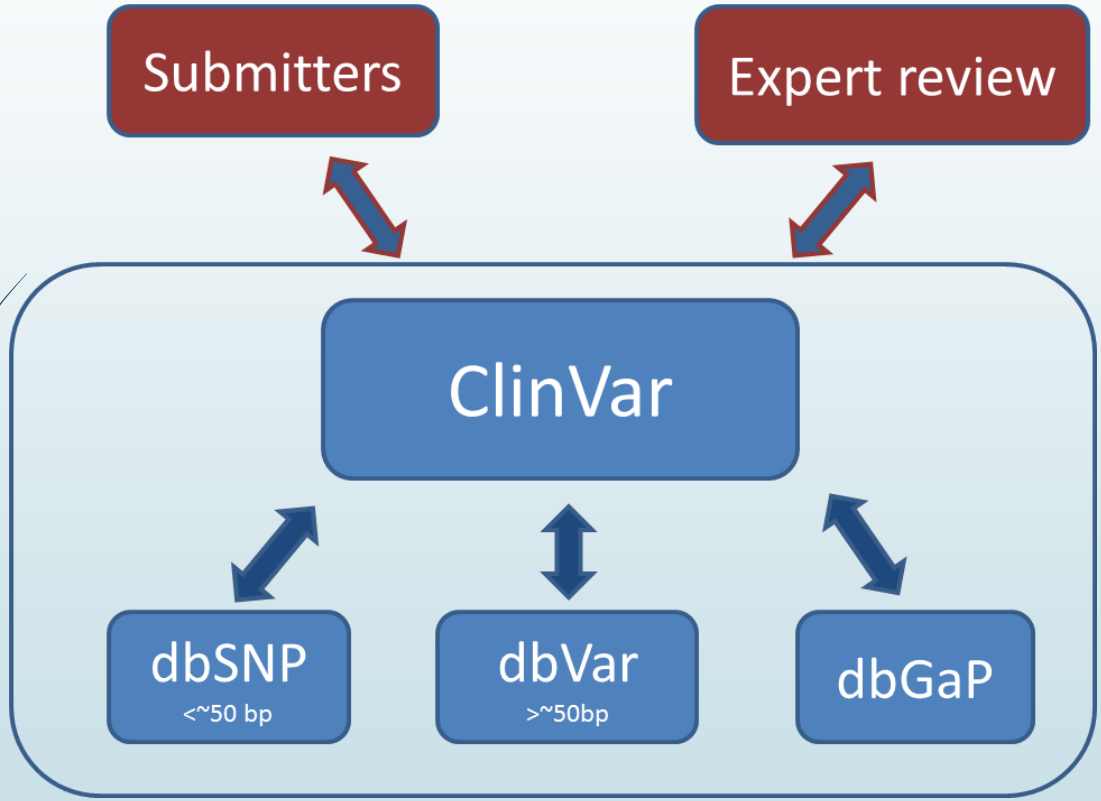
Key elements

- What the data element is
- When it was entered / when it was last touched
- Who entered it / last touched it
- Status (current, secondary, public, etc)
- Relationships
- Identifiers
- If it is a controlled list manage the list separately and use an identifier.
- Use common tables for things like citations, comments, sources.

ClinVar integration with NCBI resources

- ClinVar is a part of the medical genetics resources at NCBI which includes GTR (Genetic Testing Registry) and MedGen (Information related to human medical genetics).
  - ClinVar and GTR share an SQL database. Many data elements are in common between them and managed in one central location (e.g. organization, diseases, genes, variants, etc.)
  - Data is integrated with MedGen via XML data exchange.
- Variation can be submitted centrally to NCBI. Variants submitted to ClinVar will be sent to dbSNP or dbVar as appropriate and clinically relevant variants from other variation resources are pulled into ClinVar.
- Connections to Gene, Variation Reporter, RefSeqGene…

ClinVar integration dataflow



Data aggregation

- ClinVar collects data from individual experiments and analysis.
- Data is aggregated over both
  - The variant – all information for a specific variant
  - The variant and condition – RCV accession
- Data aggregation includes supplementing the reference accession with additional information from other resources
- Accessions are versioned and archived for access to previously released data.

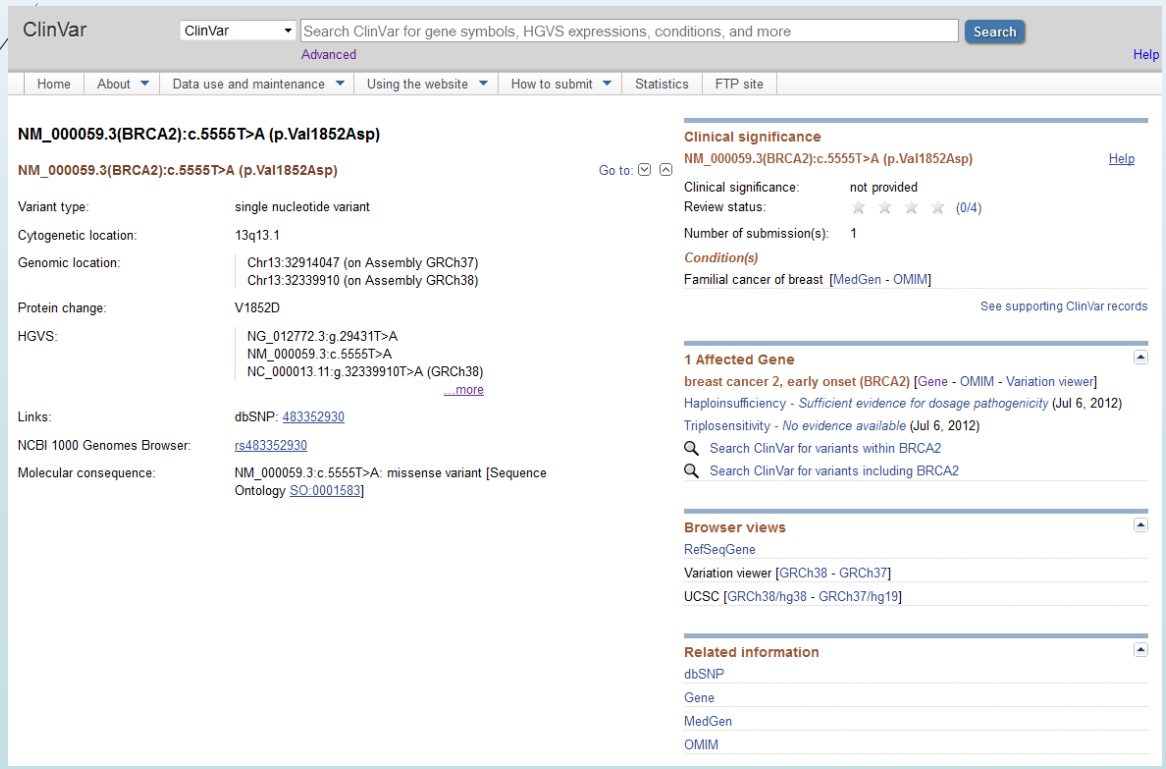
Data aggregation

Accession.version	Example
SCV000000001.1	Submitter A reports NM_000000.1:c5A>C is pathogenic for condition A
SCV000000002.1	Submitter B reports NM_000000.1:c5A>C is a variation of unknown significance for condition A
SCV000000002.2	Submitter B provides an update, now reporting that NM_000000.1:c5A>C is benign with respect to condition A
RCV000000001.1	Combines data from SCV000000001.1 and SCV000000002.2, **indicating that reports of clinical significance are not consistent**
RCV000000002.1	Authoritative group evaluates all evidence and provides the consensus report of the clinical significance of NM_000000.1:c5A>C for condition A

Data integration

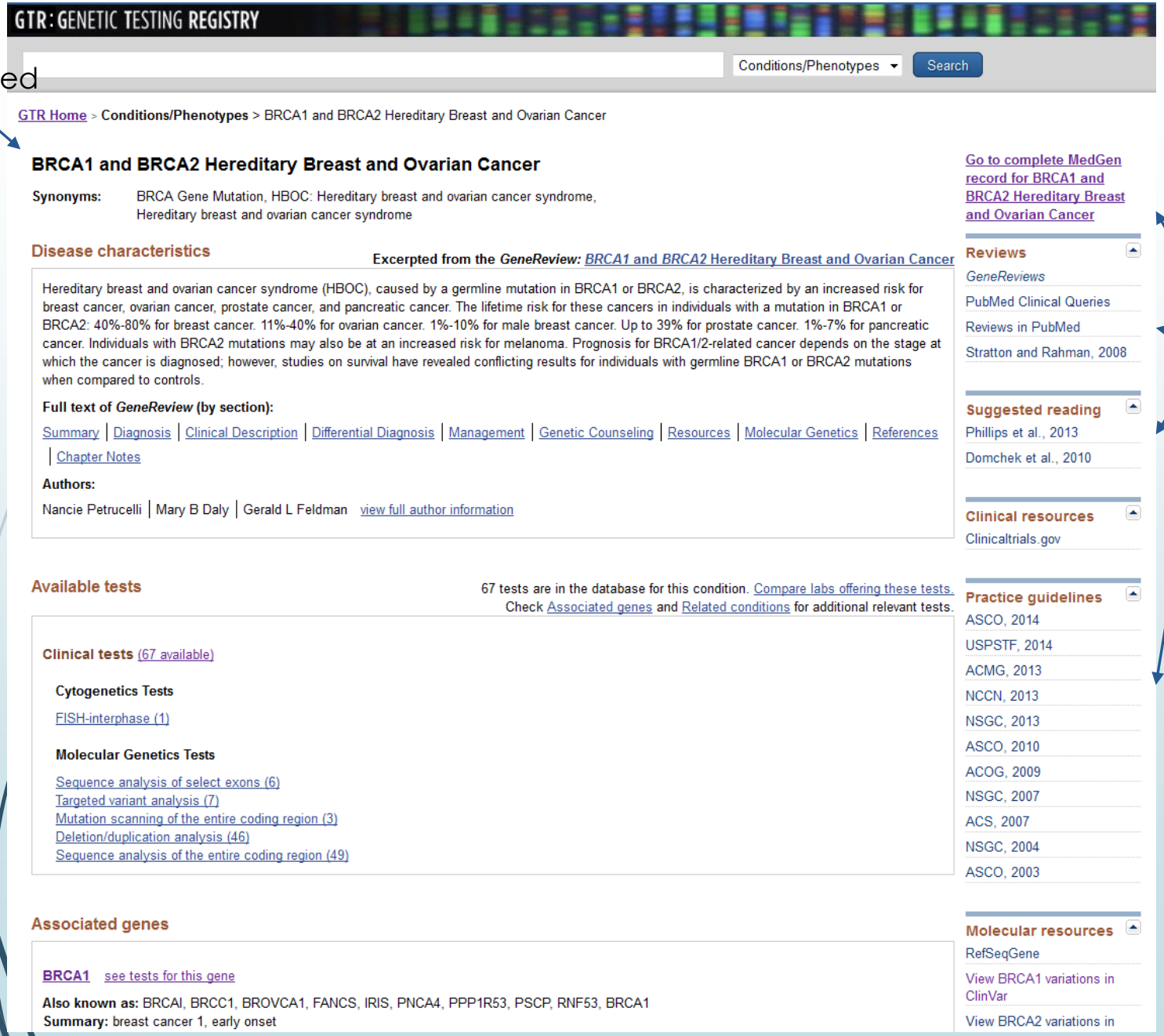
- Agree on identifiers to be used for data exchange with other resources.
- Make data available via API's or via standard data formats.
  - Include versioning of API's and notification of new versions
- Discovery links and panels.

You searched for:



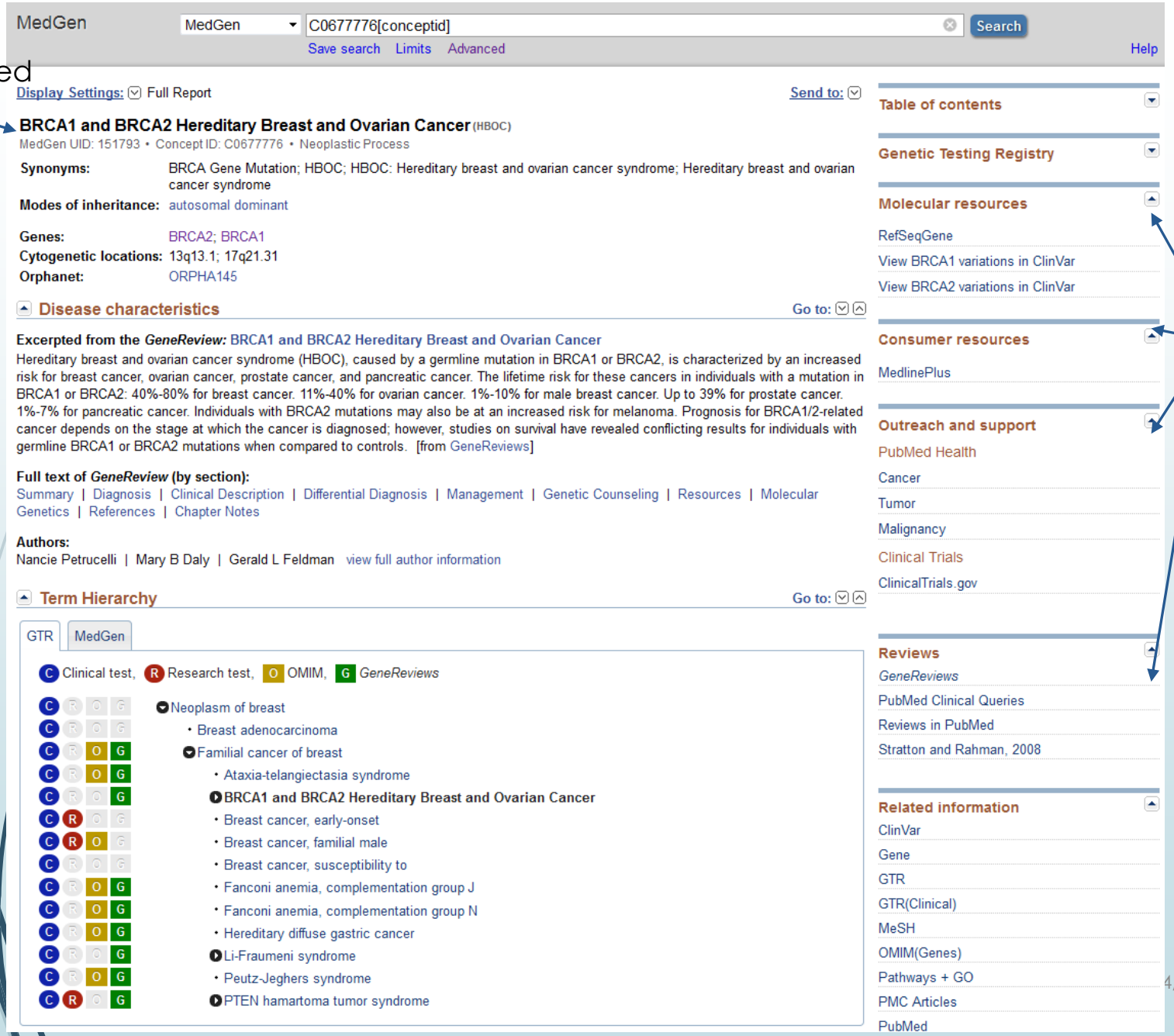
You might be interested In:

You searched for:



You might be interested In:

You searched for:



You might be interested In:

Content integrity

- Curation is done in ClinVar and GTR only in a limited capacity.
- Review data regularly programmatically to try and determine what needs curation.
- Automatically correct when possible

Gaging 'success'

- Log analysis
  - NCBI tracks use of both submission and search sites
    - How many visitors
    - Where to people go
    - What links do they use

Community relevance and participation

- GTR
  - Community relevance
    - Submitted PRA to OMB and received comments on burden and fields to be collected.
    - Held usability sessions throughout development to attempt to gage what was useful in the design
    - Redesigned the All GTR search based on community feedback
  - Community participation
    - Motivated by labs tests being discoverable
    - There was not a central genetic testing repository
- ClinVar
  - Community relevance
    - Engaged with ClinGen and large submitter groups to gage community interest
    - Have done re-designs of the website (including releasing the variant report) based on community feedback
  - Community participation
    - Accessions required by journals for publication
    - Automatic dataflows with large repositories (OMIM, GeneReviews, etc.)

Lessons learned from development of complex DBs - ClinVar and others

- Capture data rather than interpretation. Interpretation can be stored for ease of access but it is important to capture the raw data so if methods for data interpretation change this can be re-done.
- Extensive database normalization is not always the most efficient storage approach when trying to store complex data sets.
- Interaction and integration with other systems. Use of identifiers rather than strings is best if at all possible.
- Community participation is always a challenge.
- Design to multiple species even if the intention is for one. NCBI has a taxonomy database to facilitate specifying species.