Gene Curation Working Group
Training Workshop

8/10/16
Utilizing the Clinical Validity Framework

Updated 2/12/18
Important Resources

- All consortium call recording (6/18/16)
- Gene Curation SOP
- Gene Curation manuscript
- Clinical Validity Classifications
- Genetic Evidence Matrix
- Functional Evidence Matrix
- Final Summary Matrix
- GCI SOP
- Biocurator Training Modules
Importance of Expert Reviewers

• Experts may refine scoring in the matrices to better suit the complexities of a given disease area

• Any of the following are appropriate:
  – Adjust default and range scores for any evidence category
    • NOT the maximum scores or overall classification ranges
  – Define features of acceptable case-control studies within their domain
  – Define acceptable functional assays within their domain

• Any changes made within a CDWG need to be documented and consistently applied

• Any **MAJOR** changes need to be reviewed by the GCWG
Curation Workflow Overview

1. Perform Literature Search
   - Curate Genetic Evidence
   - Curate Experimental Evidence

2. Assess Evidence

3. Complete Summary Matrix

4. Assign a Clinical Validity Classification

5. Expert Review (Clinical Domain Working groups)
LITERATURE SEARCH
Finding Relevant Information

- **Broad and inclusive** initial search
  - “gene symbol/name AND disease”
  - Check HGNC ([www.genenames.org/](http://www.genenames.org/)) for old gene symbols and aliases

Broader Search
Lots of results returned
Finding Relevant Information

• **Broad and inclusive initial search**
  - “gene symbol/name AND disease”
  - Check HGNC ([www.genenames.org/](http://www.genenames.org/)) for old gene symbols and aliases

• **Identifying additional relevant information**
  - Search PubMed for experimental data (Examples below)
    • “gene AND function”
    • “protein AND function”
    • "gene AND animal”
  - OMIM ([www.OMIM.org](http://www.OMIM.org)) in the “Gene function" or “Biochemical Features” sections
  - Other databases such as
    • UniProt ([www.uniprot.org/](http://www.uniprot.org/))
    • MGI ([www.informatics.jax.org/](http://www.informatics.jax.org/))
    • GeneRIFs (Gene References Into Functions)
Evaluating Search Results

• What should I curate?
  – Curating primary literature is encouraged
  – NOT all search results will be relevant
    • Curate ALL genetic evidence until the maximum score is achieved
    • Curate across the breadth of experimental evidence available until reaching max
  – Review articles
    • Gene-disease pairs with abundant information (i.e. >50 relevant search results)
      – “gene AND disease AND (review [Publication Type] OR
      – "review literature as topic"[MeSH Terms])
    • When sufficient detail is included in the review article it may be curated otherwise
      use information from the cited publication

• Replication over time
  – Need to find the original paper with the proposed relationship
  – Cross-reference OMIM and GeneReviews
    • "Allelic Variants” section of OMIM
    • “Molecular Genetics > Pathogenic allelic variants” section of GeneReviews
    • Extract information from the original publication NOT directly from these websites
  – Use a recent review article to rule out any contradictory evidence
CASE-LEVEL DATA
Case-Level Data

- **Group/Family/Individual Evidence**
  - On the Gene Curation Interface (GCI), enter information about a group first (if available), followed by family (if available), and then an individual (proband).
    - Note: You cannot assign an individual to a group or family after adding the individual. You must create the group or family first and then add the individual.
  - If the paper uses IDs to differentiate between groups, families, or individuals, use those.
  - If not, use “First Author Year Proband/Family #”
    - Ex: Au 2015 Family 1, Au 2015 Family 2, etc.
Case-Level Data: Segregation

• Document the number of segregations in each family
  – Details begin on page 15 of the SOP
  – Number of segregations may be used in a simplified LOD score calculation documented in SOP
Case-Level Data: Phenotype

- Document the phenotype of your probands in sufficient detail for reviewers to determine their similarity

- Ability to enter HPO terms and free text
  - Use of HPO terms now will allow us to mine this data in the future: expanded phenotypes, genotype-phenotype correlations, etc.
Finding HPO terms

• Click on the hyperlink in the GCI:

• Start typing in phenotype term of interest

• HPO browser will start suggesting terms based on what you have typed:
Finding HPO Terms

Atria septal defect

Copy this number
Primary ID
HP:0001631
Alternative IDs
HP:0001630
PURL
http://purl.obolibrary.org/obo/HP_0001631

Not EXACTLY the term you searched? Check synonyms here
Synonyms
Atrial septal defect (ASD)
Defect in the atrial septum
Atrial septal defect

Textual definition
Atrial septal defect (ASD) is a congenital abnormality of the interatrial septum that enables blood flow between the left and right atria via the interatrial septum.

Logical definition
- has part' some
Intersection of
- closure incomplete
- 'inheres in' some
- interatrial septum
- 'has modifier' some
- abnormal

Broader terms
Superclasses
Abnormality of cardiac atrium
Abnormality of the atrial septum

More specific terms
Subclasses
Primum atrial septal defect
Patent foramen ovale
Secundum atrial septal defect
Sinus venosus atrial septal defect
Case-Level Data: Variant

- Add variants associated with individuals or families by using the ClinVar ID or the ClinGen Allele Registry ID

Check ClinVar for the variant ID
- Searching by HGVS is the most direct way
- If this is unavailable, be sure to include gene name in search
- Tip: If your variant is described in OMIM, you can link directly to its ClinVar page

Allelic Variants (2 Selected Examples):

- Au1-Kline Syndrome

In a 17-year-old boy with Au-Kline syndrome (AUKS; 616580), Au et al. (2015) identified a de novo heterozygous 1-bp duplication (c.953+1dupC, NM_002140.3) in the HNRNPC gene between the +1 and +2 splice sites, which was predicted to alter gene expression either through nonsense-mediated mRNA decay or a frameshift and premature termination (Gly319ArgfsTer6). The mutation was found by exome sequencing, confirmed by Sanger sequencing, and filtered against the dbSNP database.
ClinVar Variant ID

The variant ID can be located here and here:


NM_002140.4(HNRNPK):c.953+1dupG

Variation ID: 212775
Review status: ★★★★☆ (0/4) no assertion criteria provided

Interpretation
Clinical significance: Pathogenic/Likely pathogenic
Last evaluated: Oct 1, 2015
Number of submission(s): 2
Condition(s):
• AU-KLINE SYNDROME [MedGen - OMIM]
See supporting ClinVar records
ClinGen Allele Registry

• If the variant is not in ClinVar, register the allele with the ClinGen Allele Registry:
  – http://reg.clinicalgenome.org/site/cg-registry

• Must be in HGVS nomenclature

• Contact Ronak Patel for a log-in
  – Ronak.Patel@bcm.edu
Copy this number

**Canonical Allele Identifier**

CA501143

**Gene:** HNRNPK

**Identifiers and link-outs to other resources**

- ClinVar Variation Id: 221225
- ClinVar RCV Id: RCV000239991
- dbSNP Id: rs879255263

**Genomic Alleles**

<table>
<thead>
<tr>
<th>HGVS</th>
<th>Genome Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_000009.12:g.83971903_83971904insAA , CM0000671.2:g.83971903_83971904insAA</td>
<td>GRCh38</td>
</tr>
<tr>
<td>NC_000009.10:g.85776638_85776639insAA</td>
<td>NCBIdv3</td>
</tr>
<tr>
<td>NC_000009.11:g.86556810_86556819insAA , CM0000671.1:g.86556810_86556819insAA</td>
<td>GRCh38</td>
</tr>
<tr>
<td>NG_029577.1:g.13751_13752insTT</td>
<td>GRCh57</td>
</tr>
</tbody>
</table>
Case-Level Data: Variant Evidence

Gene-Disease Entry on the GCI

- Document the general category of evidence available demonstrating that the variant has some impact on gene function

You may find more information about a variant here (in silico predictions, ExAC frequency, etc)
CASE-CONTROL DATA
Case-Control Data

- **Cohort Descriptions** - Disease, phenotypes, ethnicity, age, sex, etc.
- **Variant detection methodology** - Previous testing (e.g. negative for \textit{BRCA1}/\textit{BRCA2}), methods for variant detection.
- **Power** - # of cases and controls with variant out of total number tested.
- **Bias and confounding factors** - e.g. were cases and controls matched?
- **Statistical significance** - e.g. OR or HR or p-value from Fischer's test
Evaluating Case-Control Study Quality

• Comments (Free text)
  – Any additional important information
  – Variant(s) Found- HGVS name for each variant OR # variants & type (e.g. 12 truncating variants)

• Points given
  – Evaluate the evidence as a whole to assign points
  – Take into account each of the 4 categories (variant detection methodology, power, bias and confounding factors, statistical significance)
  – Points may be altered at the discretion of the Clinical experts
  – (*See SOP for more detail)
EXPERIMENTAL EVIDENCE
Experimental Data

- **Evidence ID**
  - Unique identifier for curator’s ref.
  - Name: Author yr Experimental Evidence

- **Assign points as directed in the experimental evidence section of the SOP (p. 24).**

- **Utilize default scores where appropriate.**

- **When deviating from default, then comment.**
# Genetic Evidence Summary

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Case Information Type</th>
<th>Suggested points/case</th>
<th>Points given</th>
<th>Max Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Default</td>
<td>Range</td>
<td></td>
</tr>
<tr>
<td><strong>Case Level Data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variant Evidence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autosomal Dominant OR X-Linked Disorder</td>
<td>Variant is <em>de novo</em></td>
<td>2</td>
<td>0-3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Proband with predicted or proven null variant</td>
<td>1.5</td>
<td>0-2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Proband with other variant type with some evidence of gene impact</td>
<td>0.5</td>
<td>0-1.5</td>
<td>7</td>
</tr>
<tr>
<td>Autosomal Recessive Disease</td>
<td>Two variants in <em>trans</em> and at least one <em>de novo</em> or a predicted/proven null variant</td>
<td>2</td>
<td>0-3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Two variants (not predicted/proven null) with some evidence of gene impact in <em>trans</em></td>
<td>1</td>
<td>0-1.5</td>
<td></td>
</tr>
<tr>
<td><strong>Segregation Evidence</strong></td>
<td>Evidence of segregation in one or more families</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total LOD Score</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Candidate Gene Sequencing</td>
<td>0.5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exome/Genome or all gene sequenced in linkage region</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>≥5</td>
<td>1.5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>Case Control Data</strong></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td><strong>Case-Control Study Type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Single Variant Analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Variant Detection Methodology</td>
<td>0-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Power</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Bias and Confounding Factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Statistical Significance</td>
<td>0-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Aggregate Variant Analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL ALLOWABLE POINTS for Genetic Evidence</td>
<td>12</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Experimental Evidence Summary

## Experimental Evidence Summary Table

<table>
<thead>
<tr>
<th>Evidence Category</th>
<th>Evidence Type</th>
<th>Suggested Points/Assay</th>
<th>Points Given</th>
<th>Max Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Default</td>
<td>Range</td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td>Biochemical Function</td>
<td>0.5</td>
<td>0.5-2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Protein Interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expression</td>
<td></td>
<td>0.5-2</td>
<td>2</td>
</tr>
<tr>
<td>Functional Alteration</td>
<td>Patient cells</td>
<td>1</td>
<td>1-2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Non-patient cells</td>
<td>0.5</td>
<td>0.5-1</td>
<td>2</td>
</tr>
<tr>
<td>Models &amp; Rescue</td>
<td>Animal model</td>
<td>2</td>
<td>2-4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Cell culture model system</td>
<td>1</td>
<td>0.5-2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Rescue in animal model</td>
<td>2</td>
<td>2-4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Rescue in engineered equivalent</td>
<td>1</td>
<td>0.5-2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Total Allowable Points for Experimental Evidence**: 6
## Final Summary Matrix

<table>
<thead>
<tr>
<th>GENE/DISEASE PAIR:</th>
<th>Genetic Evidence (0-12 points)</th>
<th>Experimental Evidence (0-6 points)</th>
<th>Total Points (0-18)</th>
<th>Replication Over Time (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assertion criteria</strong></td>
<td>Case-level, family segregation, or case-control data that support the gene-disease association</td>
<td>Gene-level experimental evidence that support the gene-disease association</td>
<td>Sum of Genetic &amp; Experimental Evidence</td>
<td>&gt; 2 pubs w/ convincing evidence over time (&gt;3 yrs)</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Assigned Points</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CALCULATED CLASSIFICATION</strong></td>
<td>LIMITED</td>
<td>1-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MODERATE</td>
<td>7-11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STRONG</td>
<td>12-18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DEFINITIVE</td>
<td>12-18 &amp; Replicated Over Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Valid contradictory evidence (Y/N)</strong></td>
<td>List PMIDs and describe evidence:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CURATOR CLASSIFICATION</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FINAL CLASSIFICATION</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Final Summary Matrix - GCI

### COL3A1 - Ehlers-Danlos syndrome, vascular type

**Classification**
- Curator: Jen McGlaughon
- Calculated Classification: 18 (Definitive)
- Modified Classification: No Modification
- Last Saved: 2017 Mar 22, 10:31 am

**COL3A1**
- HGNC Symbol: COL3A1
- NCBI Gene ID: 1281

**Ehlers-Danlos syndrome, vascular type**
- Disease ID: MONDO:0017314
- OMIM ID: 120180

**Creator:** Jen McGlaughon – 2017 Mar 07, 1:17 pm
**Contributors:** Jen McGlaughon
**Last edited:** Jen McGlaughon – 2017 Apr 05, 3:43 pm

### Calculated Classification Matrix

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Count</th>
<th>Total Points</th>
<th>Points Counted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autosomal Dominant OR X-linked Disorder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proband with other variant type with some evidence of gene impact</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Proband with predicted or proven null variant</td>
<td>8</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Variant is de novo</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Two variants (not predicted/proven null) with some evidence of gene impact in trans</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Two variants in trans and at least one de novo or a predicted/proven null variant</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Segregation</td>
<td>1</td>
<td>0.5(1.2)</td>
<td>0.5</td>
</tr>
<tr>
<td>Case-Control</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Genetic Evidence Total</strong></td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Functional Evidence

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Count</th>
<th>Total Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biochemical Functions</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>Protein Interactions</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Expression</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

### Functional Alteration

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Count</th>
<th>Total Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient cells</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Non-patient cells</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Models

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Count</th>
<th>Total Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-human model organism</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Cell culture model</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rescue in human</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rescue in non-human model organism</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rescue in cell culture model</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rescue in patient cells</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Experimental Evidence Total

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Count</th>
<th>Total Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Points</td>
<td>18.00</td>
<td></td>
</tr>
</tbody>
</table>

*Combined LOD Score
Final Summary Matrix - GCI

<table>
<thead>
<tr>
<th>Gene/Disease Pair</th>
<th>Genetic Evidence (0-12 points)</th>
<th>Experimental Evidence (0-6 points)</th>
<th>Total Points (0-18 points)</th>
<th>Replication Over Time (Yes/No)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assigned Points</td>
<td>12</td>
<td>6</td>
<td>18.00</td>
<td>☑</td>
</tr>
</tbody>
</table>
| Calculated
Classification |

- LIMITED 0.1-6
- MODERATE 7-11
- STRONG 12-18
- DEFINITIVE 12-18 & Replicated Over Time

Contradictory Evidence? Proband: No Experimental: No

Modify Calculated Clinical Validity Classification: No Modification

Mark status as "Provisional Classification" (optional): ☑

Evidence Summary: Summary of the evidence and rationale for the clinical validity classification (optional).

Last Saved Summary Classification

Definitive
(2017 Mar 22, 10:31 am)