Standard Operating Protocol: Curation of gene-disease associations (version 1)

This SOP outlines curation steps guiding evaluation of evidence for genes-disease associations. It is based on the CLINGEN's *Gene Clinical Validity Curation Process SOP* version 8 and adapted to the needs and resources available in the PDL. This SOP is accompanied by several modified Clingen decision matrices (Appendix 1, Excel/Drive Path when validated) that allow interactive and efficient scoring of different evidence categories (Genetic, Experimental, Replication, and Contradictory evidence).

Where published data in human studies for a given gene-disease pair are available, PDL stratifies collated evidence into three primary association levels: Strong, Moderate, and Inconclusive. The gene-disease pairs with contradictory evidence will also be included in the Inconclusive category. Additionally, the genes for which a direct association with human disease has not been reported will be categorized as having 'No Known Relationship with Human Disease'.

_____While our evaluation of the available gene-disease association evidence will be primarily based on the quantitative guidelines published by The Clinical Genome Resource Gene Curation Working Group (see Appendix 1 and ADAPTED SCORING GUIDELINES section below), we will also incorporate more general considerations listed here:

GENERAL EVALUATION CONSIDERATIONS:

- **Strong Association**:
    - equivalent to Clingen's Definitive and Strong categories.
    - criteria:
        - At least 3 probands with expected phenotype with **>=3 variants with** compelling evidence for causality, as follows:
            - >=2 independent publications AND **strong** for association in **BOTH of** the below:
                - Variant-level evidence with numerous unrelated probands carrying variants with strong support for pathogenicity:
                    - e.g., de novo, LOF/functional disruption concordant with the mechanism of disease, absence in controls, strong linkage to a small interval
                - Gene-level evidence from different experimental supporting data
                    - Usually well-understood penetrance and/or expressivity
        - No contradictory evidence

- **Moderate/Suggestive Association:**
    - equivalent to Clingen's Moderate category
    - criteria:

        - **3** unrelated probands (preferably in 2 publications or one strong publication with sufficient supporting evidence) with a defined phenotype
        - preferably at least 3 variants with sufficient evidence for causality
            - e.g. *de novo*, LOF/functional disruption concordant with the mechanism of disease, absent in controls, strong linkage to a small interval
        - May be missing penetrance and/or expressivity data
        - Moderate experimental data
        - No contradictory evidence
- **Inconclusive Association:**
    - equivalent to Clingen's Limited and Contradictory Evidence categories
    - criteria:
        - < 3 variants in human participants
            - LOF variant or non-LOF variants with limited support for disease causality
        - At least 1 publication with supporting evidence OR
        - At least 2 case-control studies showing statistically significant enrichment in cases vs controls
        - If multiple variants seen in probands:
            - none with sufficient evidence for disease causality:
                - supported with <u>limited</u> experimental data for gene-disease association OR
                - statistically enriched in clinical populations in >=1 large case-control series without a well-described/clear phenotypic association
        - May be multifactorial or monogenic
        - Mode of inheritance may not be known
        - Contradictory along with supportive evidence may be available

<u>ADAPTED SCORING GUIDELINES BASED ON THE CLINGEN's *Gene Clinical Validity Curation Process SOP* (version 8)</u>

A. **Evidence Types:** Genetic

Please note that each case may be given points for both variant evidence and segregation analysis.

I Scoring Case-Level Data

<u>Minimum criteria to score a specific clinical case using default points:</u>

● Sufficient evidence that the case actually has the relevant diagnosis (n.b., capture as much clinical information as possible)

● The variant identified in the clinical case is a plausible cause for the specific disease conditional on the following:

-variant frequency in the general population

- disease mechanism (if known) (n.b.,do not score a null variant if mechanism is GOF)
    - Document a variety of evidence types to reflect the variant spectrum observed in disease
    - If X-linked disorders
        - watch for rare cases of females affected by X-linked recessive disorders (due to chromosomal aneuploidy, skewed X inactivation, or homozygosity)
            - Females may manifest milder and/or later in onset compared to males
        - Watch for males who carry an X-linked variant but are unaffected or mildly affected (due to Klinefelter syndrome, 47, XXY)
- recurrence in affecteds
- functional impairment
- inheritance pattern
- disease penetrance

<u>Upgrade from default points:</u>

- Variant is *de novo* and/or has supportive functional information
- High quality evidence for disease association, examples:
    - A single missense with supporting functional data (score = 0.5 in the Genetic Evidence Matrix), may be upgraded if functional evidence is robust and consistent with the expected disease mechanism
    - Phenotype is consistent and/or specific:
        - a single gene test or a limited multi-gene panel may be sufficient for default points, eg, if an enzyme assay has shown deficiency known to be associated with a particular gene (and other genetic etiologies are unlikely), then sequencing of that gene alone is enough for default points
    - High likelihood that a null variant actually leads to LOF or that missense in a functional domain is associated with disease

<u>Downgrade from default points:</u>

- insufficient/limited prior testing performed in order to rule out other causes of disease
- nonspecific and/or genetically heterogeneous phenotype
    - eg, if a variant identified in a gene after single gene-sequencing in an individual with autism/ID and no previous testing, downgrade as other genetic etiologies not ruled out
- a putative null variant unlikely to result in nonsense-mediated decay (e.g., if new stop codon downstream of the last 50 bp of the penultimate exon)
- parental relationships not confirmed for *de novo* variants.

<u>Do not score a variant if of dubious quality/relevance</u>

- E.g. variant in older literature classified “pathogenic” but found at high frequencies in controls

a. Scoring Biallelic Variants:

- In cases where biallelic variants (in trans) cause disease, evaluate each variant independently, then sum their scores
  - both variants should be identified (with some evidence to suggest they are *in trans*) in the observed case
    - Exceptions to scoring both variants, if the following are present:
      - substantial evidence that biallelic variants cause disease (as opposed to <u>new</u> gene-disease relationships where it is unclear if AR or AD)
      - there is an alternative method of confirmation that the patient has the disease of interest (e.g., metabolic disorders with diagnostic biochemical profiles).
- If homozygous variants, consider downgrading and/or reducing the maximum points for such cases
  - Consider requiring homozygous missense variants to also have supporting functional evidence before scoring

b. Scoring de novo variants:

<u>Minimum criteria to score a specific clinical case using default points:</u>

- Confirmed absence in both parents
  - If in autosomal genes and females heterozygous for an X-linked variant, both parents must have been tested
  - For males who are hemizygous for an X-linked variant, only the mother needs to be tested

○ One of the parents of an affected individual is found to be a mosaic for the “de novo” variant

- if the mechanism of disease is GOF, award default points to *de novo* missense variants shown to be GOF
- if the mechanism of disease is LOF, award default points to likely *de novo* LOF variants (e.g. nonsense, frameshift, canonical splice site) shown to be *de novo* based on parental testing for the variant

<u>Upgrade from default points:</u>

- If literature/evidence suggests that the statistical likelihood of seeing *de novo* variation in the given gene is low
  - otherwise, (e.g., TTN), do not score or use caution

<u>Downgrade from default points:</u>

- if parental relationships (i.e., both maternity and paternity) not confirmed
- downgrade conditional on disease mechanism:
  - if a *de novo* missense without evidence of LOF, where LOF is expected mechanism
  - If a missense with unclear function, where GOF is the mechanism of disease
  - score 0 points if *de novo* putative LOF variants in a gene with known GOF mechanism

c. Scoring Null variants:

Minimum criteria to score a specific clinical case using default points:

- nonsense, frameshift, canonical splice site, single/multi-exon or whole gene deletions
- If other variant types (small in-frame indels, missense), need convincing evidence of complete LOF to award default points

Downgrade from default points:

- if alternative splicing, if the putative null variant is near the C end, and if NMD is not predicted
- if a gene product is still made, but altered
    - Eg., cDNA analysis and/or Western blot from a case with a canonical splice variant shows residual expression despite exon skipping

Upgrade from default points:

- some functional impact of the variant must be demonstrated
    - E.g., reduced activity of an enzyme in cells expressing the variant
    - this excludes *in silico* predictions, but may cautiously upgrade if in-depth modeling studies (e.g. impact on 3D structure), if convincing

Do not score a CNV/SV/chromosomal rearrangement outside of gene of interest

- Exception: these types of variants are within the gene's variant spectrum and this should be noted
    - E.g., *NRXN1* curation example

d. Scoring recurrent variants:

- Ensure that variant is not from a single patient but reported multiple times independently
- If a recurrent, *de novo* variant:
    - If confirmed with parental testing, score each individual case
    - if similar phenotypes across patients with *de novo variants*, this may indicate a hot spot and supports pathogenicity
        - score each independent observation (along with other variants, if available).
    - If evidence suggests a variant has arisen more than once in different populations but no evidence to indicate the variant is *de novo* in the patient(s), score each case individually according to the variant type
    - if evidence is insufficient to support that the variant has arisen in different populations and cases do not seem to be related, consider downgrading or not scoring the subsequent cases after the first case to reduce overscoring.

e. Scoring Founder Variants:

- prioritize using case-control studies
- the variant should not have data that contradicts a pathogenic role, ie unexplained non-segregation,
- with cases, curate a range of variants in addition to the known founder ones for representativeness
    - avoid double counting any cases that may have been included in case-control studies

- for variants more common in specific populations but that are not well-known pathogenic founder variants, assess role in the disease carefully to avoid over-scoring
    - functional data here are useful to ensure that the common variant is not a benign variant in LD with the causative one

II Case-Control Data

Differentiate between single variant vs aggregate variant studies:

- Single variant studies evaluates individual variants vs aggregated variants for statistical enrichment in cases vs controls.

Assess Quality of the Study & Record in Excel, (Excel file, tab 2):

● Variant Detection Methodology:
- Cases and controls should be analyzed with methods with comparable analytical performance (e.g. sufficient and equivalent depth and coverage).
- E.g. Do not score if controls sequenced & cases genotyped but may still use as case-level data

● Statistical Power: the study considered disease prevalence, allele frequency, and the expected effect size

● Bias and Confounding factors:
    ○ Are cases and controls matched by demographic information, genetic ancestry, other confounders
        - Lower score if population database used rather than matched controls
    ○ Have the cases &controls been equivalently evaluated for the phenotype and/or family history of disease?

● Statistical Significance:
    ○ magnitude of odds ratio (OR) consistent with a monogenic etiology
        - When p-values or 95% confidence intervals (CI) given, consider strength of the stated association in the final points assigned
    ○ multiple testing correction applied, as appropriate

III Segregation Studies (see the Excel file and consult pages 27-35 in the *Gene Clinical Validity Curation Process SOP* v8 for detailed guidance)

Default criteria for using ClinGen's LOD score formula:

- The disorder is rare and highly penetrant
- Phenocopies are rare or absent.
- For dominant or X-linked disorders, the estimated LOD score should be calculated using ONLY families with 4 or more segregations present. The affecteds may be within the same or across generations.
- For recessive disorders, the estimated LOD score should be calculated using ONLY families with at least 3 affected individuals, including the proband
- Genotypes must be specified for all affected and unaffected individuals counted:

    - parents of affected individuals must be genotyped/tested to show that variants are *in trans* if the proband is a compound heterozygote
- Families included in the calculation must not demonstrate any unexplainable non-segregations (e.g, a genotype-/phenotype+ individual in a family affected by a disorder with no known phenocopies).
    - Families with unexplainable non-segregations should **not** be used in LOD score calculations.
- for autosomal dominant or X-linked disorders, only affected individuals (genotype+/phenotype +) or obligate carriers (regardless of the phenotype) should be included
- dizygotic (fraternal) twins count as two **separate** individuals and monozygotic (identical) twins as one
- if more than one family meets the above criteria for scoring segregation, sum the LOD scores
- As segregation implicates a locus NOT a variant, ensure that appropriate measures have been taken to rule out other possible causative genes within the critical region (see detailed guidance in the *Gene Clinical Validity Curation Process SOP* v8, page 30)

B. EXPERIMENTAL EVIDENCE:

1. Biochemical Function studies:
    - Based on empirical evidence & not *in silico* prediction
    - Proteins previously implicated in the disease should have strong evidence for association with disease
    - Show that gene product performs a biochemical function:
        - that is shared with other established genes in the disease of interest OR
        - consistent with the phenotype.
    - Upscore with an increasing number of other proteins with the same function in the same disease.

2. Protein Interaction:
    - Shows that the gene product interacts with proteins previously implicated in the disease of interest
        - E.g., physical interaction via Yeast-2- Hybrid (Y2H), co-immunoprecipitation (coIP)
    - If a study shows that a variant disrupts the gene's interaction with another protein
        - To score, need positive control showing interaction between the two wild type proteins
        - Points can also be awarded to case-level/variant evidence that a variant disrupts interaction.

3. Expression:
    - evidence that gene expressed in tissues relevant to the disease or its expression altered in patients with the disease
    - Methods:

      a) RNA transcripts (RNAseq, microarrays, qPCR, qRT-PCR, Real-Time PCR),

      b) protein expression (western blot, immunohistochemistry)
        - E.g. expression in patient tissues and/or cell samples
    - may choose to award points based on the specificity of expression in relevant organs

4. Functional Alteration:

- cultured cells where gene function is disrupted have phenotype consistent with the human disease
    - Examples: expression of a genetic variant, gene knock-down, overexpression.
- divide the evidence according to whether the experiment conducted in patient vs non-patient cells

5. Model System:

- a non-human model organism or cell culture with a disrupted copy of the gene has a phenotype consistent with the human disease
    - Cell culture models should recapitulate features of the diseased tissue
        - e.g. engineered heart tissue or cultured brain slices.

6. Rescue:

- Summarize rescue evidence
    - If rescued in humans (i.e. patients), non-human model organisms, cell culture models, or patient cells
        - may want to award more points if human vs. non-human model organism
            - E.g. ERT in LSDs
- If LOF, summarize if rescued by exogenous wild-type gene, gene product, or targeted gene editing
- If GOF, summarize treatment that specifically blocks the action of the variant (e.g. siRNA, antibody, targeted gene editing) to rescue the phenotype.

General comments on evaluating experimental evidence:

- prioritize curating genetic evidence over experimental evidence to reach a definitive score
- Ideally find evidence from each category below
- Differentiate between functional evidence for Variant- (case-level) vs Gene (experimental):
- Count as case-level data (count only once to prevent overscoring):
    - For experimental evidence that does not **directly** support the role of the gene in the disease or recapitulation of disease phenotypes, but indicates that variant is damaging to the gene function
        - Immunolocalization: gene product is mislocalized in cells from a patient or in cultured cells.
            - Exception: count as experimental evidence (functional alteration) if mislocalization/accumulation of an altered gene product is a known mechanism of disease
        - Mini-gene splicing assay or RT-PCR shows impacted splicing

        - enzyme activity deficient if a variant studied in cultured cells
        - variant is shown to disrupt the normal interaction of the gene product with another protein
            - If the latter strongly implicated in the same disease, the interaction can be counted in experimental data (Function: protein interaction) and
            - the lack of interaction due to the variant can be counted as case-level evidence
        - Tissue or cells from a variant carrier have altered expression (Western blot).
- Count as experimental evidence:
    - If a signaling pathway known to be involved in the disease mechanism.
        - Expression of a missense variant in cells shows that the gene product can no longer function as part of this pathway.
    - Altered expression of the gene shown in multiple patients with the disease regardless of the causative variant and even if their genotyped unknown
    - The variant is associated with a known feature of the disease e.g. abnormal mislocalization either in patient cells or cultured cells
    - Any model organism with a variant initially identified in a human with the disorder

C. CONTRADICTORY EVIDENCE:

For detailed guidance, see the *Gene Clinical Validity Curation Process SOP* v8 (pp. 44-46). Some examples of contradictory evidence are listed below.

1. Case-control data show no statistically significant difference

2. Minor allele frequency is too high for the disease

3. The gene-disease relationship cannot be replicated both over time and across multiple cohorts:

- If a study could not identify any variants in the gene being curated in an affected population that was negative for other known causes of the disease, it is potentially contradictory
    - however, take into account that a disease may be too rare or a study too small/underpowered

4. Non-segregations:

- Watch for age-dependent penetrance and appropriate phenotyping of the relatives
    - the age of unaffected variant carriers should be similar to the affected variant carriers.

5. Non-supporting functional evidence or functional evidence not consistent with the disease mechanism

D. REPLICATION CONSIDERATIONS

- Counted if if > 3 years have passed since the original publication AND there are >2 convincing publications about the gene-disease relationship

E. SUGGESTED RECURATION SCHEDULE per CLINGEN's Standard Gene Recuration procedure_V8.

Table 2: Standard Gene-Disease Clinical Validity Recuration Procedure

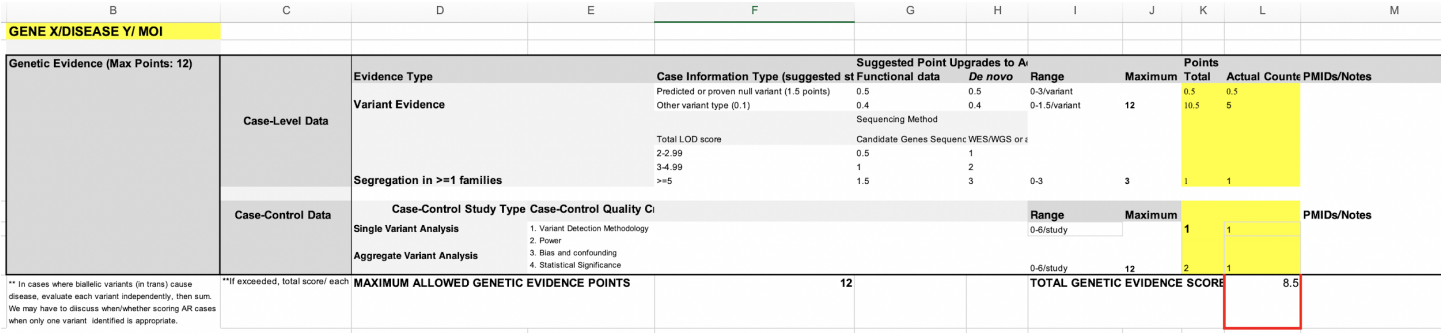
Classification	Interval for re-evaluation	Specifications for Recuration
Strong	**3 years from the original discovery publication date**	- Consider if any contradictory evidence has been published since the last approved classification. - Check if any new disease entities have been asserted. - Update with new pertinent information.
Moderate	2? years after the last approval date, unless new gene evidence comes to light before that, during the course of clinical reporting, or unless automated literature search flags point toward new evidence.	- Consider if any contradictory evidence has been published since the last approved classification - Check if any new disease entities have been asserted. - Update with any new pertinent information
Limited	3? years after the last approval date, unless new gene evidence comes to light before that, during the course of clinical reporting, or unless automated literature search flags point toward new evidence.	- Consider if any contradictory evidence has been published since the last approved classification - Check if any new disease entities have been asserted. - Update with any new pertinent information. - In genes where there may be strong evidence against the asserted association, is there new supporting evidence of the gene-disease relationship?
No Known Human Disease Relationship	# years after the last approval date, unless new gene evidence comes to light before that, during the course of clinical reporting, or unless automated literature search flags point toward new evidence.	- Check if any new disease entities have been asserted.

Appendix 1

Adapted ClinGen Matrices for Gene-Evidence Association Evaluation and Scoring

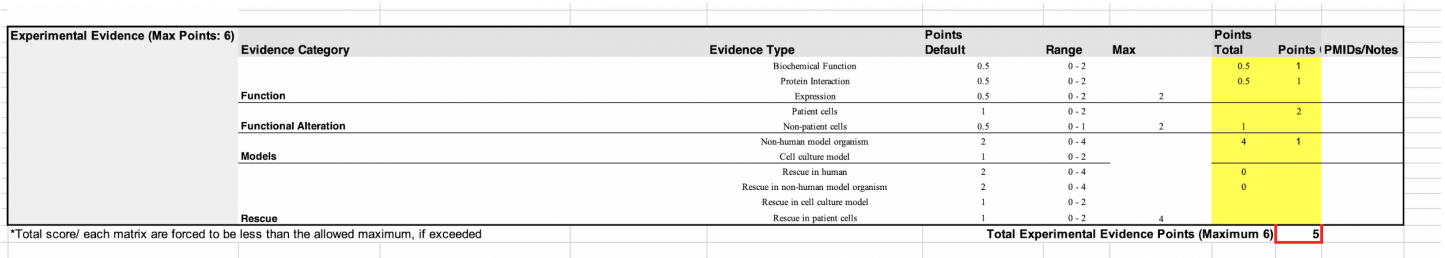
- Yellow cells are to be filled by the curator/ reviewer
- Allowable score maxima are noted in each matrix
- Subtotals (outlined in red) from the Genetic and Experimental metrices are automatically passed to the Summary Matrix where classification category is populated.
- Curation date is populated automatically
- A copy of the score for a given gene-disease association should be saved and enclosed with other records specific to that gene

1. Genetic Evidence Matrix



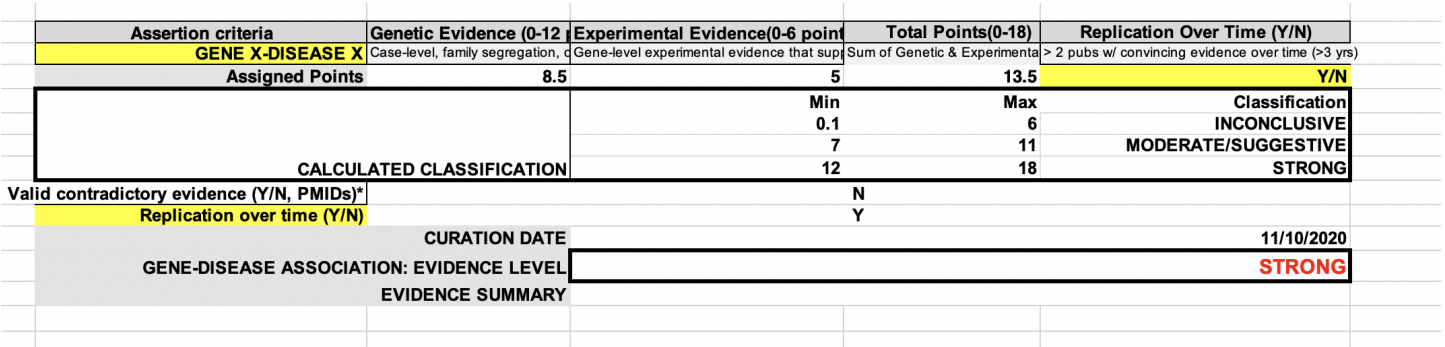
GENE X/DISEASE Y/ MOI												
Genetic Evidence (Max Points: 12)					Suggested Point Upgrades to A					Points		
		Evidence Type	Case Information Type (suggested st		Functional data	*De novo*	Range	Maximum	Total	Actual Counte	PMIDs/Notes
	Case-Level Data	Variant Evidence	Predicted or proven null variant (1.5 points)		0.5	0.5	0-3/variant		0.5	0.5	
			Other variant type (0.1)		0.4	0.4	0-1.5/variant	12	10.5	5	
					Sequencing Method						
			Total LOD score		Candidate Genes Sequenc	WES/WGS or a					
			2-2.99		0.5	1					
			3-4.99		1	2					
		Segregation in >=1 families	>=5		1.5	3	0-3	3	1	1	
	Case-Control Data	Case-Control Study Type	Case-Control Quality Cr								
							Range	Maximum			PMIDs/Notes
		Single Variant Analysis	1. Variant Detection Methodology				0-6/study		1	1	
			2. Power								
		Aggregate Variant Analysis	3. Bias and confounding								
			4. Statistical Significance				0-6/study	12	2	1	
** In cases where biallelic variants (in trans) cause disease, evaluate each variant independently, then sum. We may have to discuss when/whether scoring AR cases when only one variant identified is appropriate.	**If exceeded, total score/ each	MAXIMUM ALLOWED GENETIC EVIDENCE POINTS		12			TOTAL GENETIC EVIDENCE SCORE			8.5	

2. Experimental Evidence Matrix



Experimental Evidence (Max Points: 6)			Points			Points	
	Evidence Category	Evidence Type	Default	Range	Max	Total	Points	PMIDs/Notes
		Biochemical Function	0.5	0 - 2		0.5	1	
		Protein Interaction	0.5	0 - 2		0.5	1	
	Function	Expression	0.5	0 - 2	2			
		Patient cells	1	0 - 2			2	
	Functional Alteration	Non-patient cells	0.5	0 - 1	2	1		
		Non-human model organism	2	0 - 4		4	1	
	Models	Cell culture model	1	0 - 2				
		Rescue in human	2	0 - 4		0		
		Rescue in non-human model organism	2	0 - 4		0		
		Rescue in cell culture model	1	0 - 2				
	Rescue	Rescue in patient cells	1	0 - 2	4			
*Total score/ each matrix are forced to be less than the allowed maximum, if exceeded					Total Experimental Evidence Points (Maximum 6)		5	

3. Summary Matrix



Assertion criteria	Genetic Evidence (0-12 p	Experimental Evidence(0-6 point	Total Points(0-18)	Replication Over Time (Y/N)
GENE X-DISEASE X	Case-level, family segregation, c	Gene-level experimental evidence that sup	Sum of Genetic & Experimenta	> 2 pubs w/ convincing evidence over time (>3 yrs)
Assigned Points	8.5	5	13.5	Y/N
		Min	Max	Classification
		0.1	6	INCONCLUSIVE
		7	11	MODERATE/SUGGESTIVE
CALCULATED CLASSIFICATION		12	18	STRONG
Valid contradictory evidence (Y/N, PMIDs)*		N		
Replication over time (Y/N)		Y		
CURATION DATE				11/10/2020
GENE-DISEASE ASSOCIATION: EVIDENCE LEVEL				STRONG
EVIDENCE SUMMARY				

4. Case-control study summary template (see Excel file, the'case-controls studies template' tab)

ClinGen CASE-CONTROL DATA EVALUATION: TEMPLATE

Points	Power	Bias/Confounding	DetectionMethod	StatisticalSignificance	StudyType	Points(0-6/study)
Author A 2015 (Max score)	Breast cancer cases: 100/1	Matched by age, ethnicity, and l	Cases & controls genot	OR: 5.4 [95% CI: 2.5-11.6; P	Single Variant	6
Author B 2005 (Intermedia	HCM Cases: 13/200 Contro	Matched by location, but not ag	Cases & controls genot	Fisher's exact test P = 0.004	Single Variant	4
Author C 2011 (Low score)	Ovarian cancer cases: 11/1	Matched by ethnicity. Controls fr	Cases:sequenced Gen	OR of all variants in aggregat	Aggregate analysi	2
Author D 2009 (No case- c	Colorectal cancer cases: 11	Matched by ethnicity. Controls fr	Cases: sequenced gen	OR of p.Lys342: 4.9 (CI: 1.4-	Not applicable	0