Gene Clinical Validity Curation Process

Standard Operating Procedure

Version 4
May 2017
The Clinical Genome Resource
Gene Curation Working Group

TABLE OF CONTENTS

BACKGROUND:

ClinGen's gene curation process is the method designed to aid in evaluating the strength of a gene-disease relationship based on publicly available evidence. Information about the gene-disease relationship, including genetic, experimental, and contradictory evidence curated from the literature are compiled and used to assign a clinical validity classification per criteria established by the ClinGen Gene Curation Working Group (GCWG). This protocol details the steps involved in curating a gene-disease relationship and subsequently assigning a clinical validity classification. This curation process is not a systematic review of all available literature for a given gene or condition, but instead an overview of the most pertinent evidence required to assign the appropriate clinical validity classification for a gene-disease relationship at a given time.

REQUIRED COMPONENTS:

-ClinGen-approved curation training. For training resources please see the ClinGen gene curation website (https://www.clinicalgenome.org/working-groups/gene-curation/) or contact Erin Riggs (eriggs@geisinger.edu)
-Internet browser
-Publication Access
-Microsoft Office (Word, Excel, or Powerpoint to record your data from curation)

OVERVIEW OF GENE CURATION:

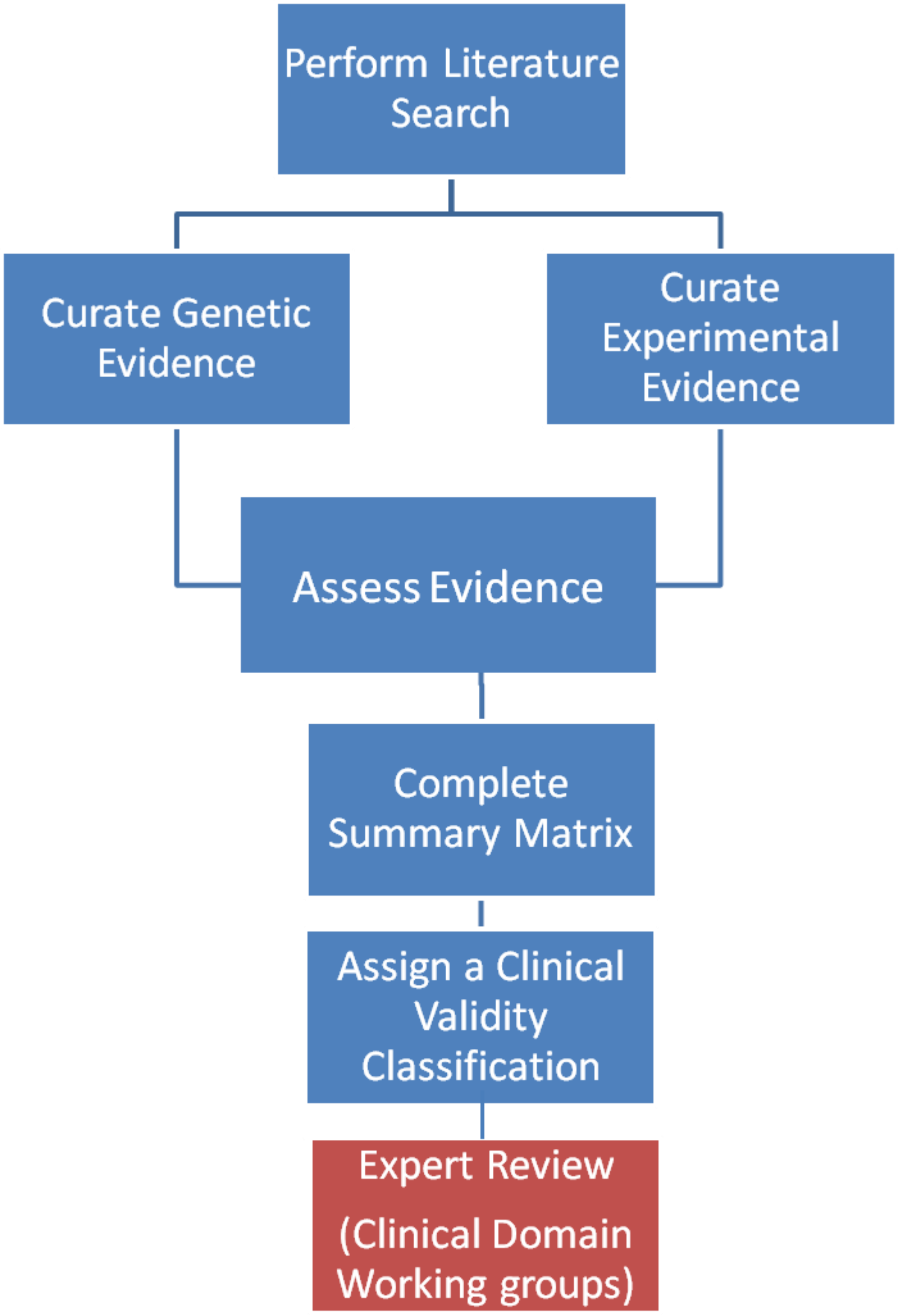
The gene curation frame work consists of the following steps.

- **Collection of evidence:** The evidence is collected primarily from published peer-reviewed literature, but can also be present in publicly accessible resources, such as variant databases, which can be used with discretion. Literature searches can be conducted using PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and/or Google Scholar (http://scholar.google.com/) (which has a full-text search feature). Advanced searches are generally more informative. Google Scholar search help: (https://scholar.google.com/intl/en/scholar/help.html#searching)
- One need not comprehensively curate all evidence for a gene-disease pair (particularly for "Definitive" associations), but instead focus on curating and evaluating the relevant pieces of evidence described in this protocol.
- **Identifying different evidence types:** The curator needs to identify and curate genetic and experimental evidence separately (details are defined later in "Genetic Evidence" and Experimental Evidence" sections). Genetic evidence is divided into two categories: case-level data and case-control data. Typically studies describing individuals or families with variants in the gene of interest

will be scored as case-level data, while studies using statistical analysis to determine the enrichment of variants in case and control groups will be scored as case-control data. The gene-level experimental data used in this framework to assess a gene-disease relationship are *in vitro* and *in vivo* functional studies that implicate the causative role of a gene in disease. These are based on MacArthur and colleagues (PMID:24759409) and described in detail below.

- **Assignment of clinical validity classification using gene curation matrix:** Next the curator evaluates the evidence and assigns points to the evidence using the scoring matrices provided below (Fig. 3,8). This information is then summarized and tallied to generate a total score and calculated clinical validity classification, which will be reviewed by a committee of appropriate disease experts.

Figure 1: Gene Curation Workflow


Perform Literature Search
Curate Genetic Evidence
Curate Experimental Evidence
Assess Evidence
Complete Summary Matrix
Assign a Clinical Validity Classification
Expert Review
(Clinical Domain Working groups)

CLINICAL VALIDITY CLASSIFICATIONS:

The gene curation working group members have developed a method to qualitatively define the "clinical validity" of a gene-disease relationship using a classification scheme based on the strength of evidence that supports or refutes any claimed relationship. This framework allows the "clinical validity" of a gene-disease relationship to be transparently and systematically evaluated. These classifications can then be used to prioritize genes for analysis in various clinical contexts. The suggested minimum criteria needed to obtain a given classification are described for each clinical validity classification. These criteria include both genetic and experimental evidence, which are described below in this document. The default classification for genes without an identified variant in humans is "No Reported Evidence." The level of evidence needed for each supportive gene-disease relationship category builds upon the previous category (i.e. "Limited" builds upon "Moderate"). Gene-disease relationships classified as "Contradictory" likely have evidence supporting as well as opposing the gene-disease association, but are described separately from the classifications for supportive gene-disease relationships.

Evidence Level		Figure 2: Clinical Validity Classifications (Evidence Description)
Supportive Evidence	DEFINITIVE	The role of this gene in this particular disease has been repeatedly demonstrated in both the research and clinical diagnostic settings, and has been upheld over time (in general, at least 3 years). No convincing evidence has emerged that contradicts the role of the gene in the specified disease.
	STRONG	The role of this gene in disease has been independently demonstrated in at least two separate studies providing **strong** supporting evidence for this gene's role in disease, including both of the following types of evidence: • Strong variant-level evidence demonstrating numerous unrelated probands harboring variants with sufficient supporting evidence for disease causality[1] • Compelling gene-level evidence from different types of supporting experimental data[2]. In addition, no convincing evidence has emerged that contradicts the role of the gene in the noted disease.
	MODERATE	There is **moderate** evidence to support a causal role for this gene in this disease, including both of the following types of evidence: • At least 3 unrelated probands harboring variants with sufficient supporting evidence for disease causality [1] • Moderate experimental data[2] supporting the gene-disease association The role of this gene in disease may not have been independently reported, but no convincing evidence has emerged that contradicts the role of the gene in the noted disease.
	LIMITED	There is **limited** evidence to support a causal role for this gene in this disease, such as: • Fewer than three observations of variants with sufficient supporting evidence for disease causality [1] OR • Variants have been observed in probands, but none have sufficient evidence for disease causality. • Limited experimental data[2] supporting the gene-disease association The role of this gene in disease may not have been independently reported, but no convincing evidence has emerged that contradicts the role of the gene in the noted disease.
NO REPORTED EVIDENCE		Evidence for a causal role in disease has not been reported. These genes might be "candidate" genes based on linkage intervals, animal models, implication in pathways known to be involved in human diseases, etc., but no reports have directly implicated the gene in human disease cases.

Contradictory Evidence	**CONFLICTING EVIDENCE REPORTED**	Although there has been an assertion of a gene-disease association, conflicting evidence for the role of this gene in disease has arisen since the time of the initial report indicating a disease association. Depending on the quantity and quality of evidence disputing the association, the association may be further defined by the following two sub-categories: **1. Disputed** a. Convincing evidence *disputing* a role for this gene in this disease has arisen since the initial report identifying an association between the gene and disease. b. Refuting evidence need not outweigh existing evidence supporting the gene-disease association. **2. Refuted** a. Evidence *refuting* the role of the gene in the specified disease has been reported and significantly outweighs any evidence supporting the role. b. This designation is to be applied at the discretion of clinical domain experts after thorough review of available evidence
NOTES		
[1] Variants that disrupt function and/or have other strong genetic and population data (e.g. *de novo* occurrence, absence in controls, strong linkage to a small genomic interval, etc.) are considered convincing of disease causality in this framework. See "Variant Evidence" on p.12 for more information. [2] Examples of appropriate types of supporting experimental data based on those outlined in MacArthur et al. 2014.		

LITERATURE SEARCH:

1. The initial search should be **broad** and **inclusive**. A good way to start is by searching **"gene symbol/name AND disease"** (in some cases it may be sufficient to search for the gene name/symbol alone). Ensure that you have looked up gene symbol/name alternatives before you search.
    a. Check HGNC (www.genenames.org/) for old gene symbols and aliases
    b. NCBI Gene (www.**ncbi**.nlm.nih.gov/**gene**) also lists gene aliases
    c. NOT all search results will be relevant, thus it is important to examine the search results for pertinent information

2. Curating primary literature is encouraged, but if a gene-disease pair has abundant information (i.e. >50 relevant results returned in a search), review articles may be sufficient. To find reviews, search PubMed with **"gene AND disease AND (review [Publication Type] OR "review literature as topic"[MeSH Terms]).**
    a. Curation may occur from that publication **ONLY** when sufficient details are included in the review article.

b. If sufficient details are **NOT** included in the review article then the curator will need to return to each individual publication to curate the information.

3. Additional searches are often necessary to identify sufficient gene level experimental evidence. Note that additional gene level experimental evidence may exist in publications **BEFORE** the gene:disease association was first made.

a. Search PubMed for experimental data (Examples below)
    - “gene AND function”
    - “protein AND function”
    - "gene AND animal”

b. Additional information may also be available in OMIM (www.OMIM.org) in the **“Gene function"** or **“Biochemical Features”** sections

c. GeneReviews (http://www.ncbi.nlm.nih.gov/books/NBK1116/) often has information in the **“Molecular Genetics”** section of the disease entries that may be useful.

d. Other databases such as UniProt (www.**uniprot**.org/), MGI (www.informatics.jax.org/), etc. may also be useful, provided that primary references are given that can be curated.

e. GeneRIFs (Gene References Into Functions) within NCBI Gene lists article links that summarize experimental evidence for a given gene. The link itself leads to an article in Pubmed and can serve as an additional source for experimental evidence.

4. An additional component of the curation process is to determine if the original gene-disease association has been replicated; therefore, it is critical to find the **original paper** with the proposed relationship. OMIM and GeneReviews often cite the first publication and should be cross-referenced. Additionally, a recent review article may be helpful in ruling out any contradictory evidence that may have been reported since the original publication.

a. The **"Allelic Variants”** section of OMIM and the **“Molecular Genetics > Pathogenic allelic variants”** section of GeneReviews may have relevant information.

b. Be sure to extract information from the **original publication**, NOT directly from these websites.

Now that all of the relevant literature about the gene-disease relationship has been assembled, you can start to curate the different pieces of evidence.

GENETIC EVIDENCE

Genetic evidence may be derived from **case-level data** (studies describing individuals or families with variants in the gene of interest) and/or **case-control data** (studies in which statistical analysis is used to evaluate enrichment of variants in cases compared to controls). While a single publication may include both case-level and case-control data, individual cases should NOT be double-counted (e.g., an individual case that is part of a case-control cohort should not be given points from both the "case-level data" and "case-control data" categories). **For example**, although this would be an unlikely situation, if a case from a case-control study were singled out and a pedigree was provided, this case could be evaluated with case-level data and segregations counted, but should be removed from the case count of the case-control data. In this scenario, a note should be made for expert review.

Genetic Evidence Summary Matrix

A matrix used to categorize and quantify the genetic evidence curated for a gene-disease pair is provided below. **NOTES:** All variants under consideration should be rare enough in the general population to be consistent with prevalence of disease.

Figure 3: Genetic Evidence Matrix

GENETIC EVIDENCE SUMMARY									
		Evidence Type	Case Information			Suggested Points/Case		Points Given	Max Score
						Default	Range		
Case-Level Data	Variant Evidence	Autosomal Dominant OR X-Linked Disorder A	Variant is *de novo*			C 2	0-3	H	M 12
			Proband with predicted or proven null variant			D 1.5	0-2	I	N 10
			Proband with other variant type with some evidence of gene impact			E 0.5	0-1.5	J	O 7
		Autosomal Recessive Disorder B	Two variants in *trans* and at least one *de novo* or a predicted/proven null variant			F 2	0-3	K	P 12
			Two variants (not predicted/proven null) with some evidence of gene impact in *trans*			G 1	0-1.5	L	
	Segregation Evidence		Evidence of segregation in one or more families	LOD Score Examples	3	5	0-7	Q	R 7
					2	4			
					1.5	3			
					1	1.5			

	Case-Control Study Type	Case-Control Quality Criteria	Suggested Points/Study	Points Given	Max Score
Case-Control Data[6]	Single Variant Analysis	• Variant Detection Methodology • Power • Bias and Confounding Factors • Statistical Significance	0-6	S	T 12
	Aggregate Variant Analysis		0-6		
TOTAL ALLOWABLE POINTS for Genetic Evidence					U 12

Case-Level Data

Assessing case-level data requires knowledge of the inheritance pattern of the disease in question and careful interrogation of the individual variants identified in each case. Within this framework, a case should only be counted towards supporting evidence if the variant identified in that individual has some indication of a potential role in disease (e.g. impact on gene function, recurrence in affected individuals, etc.). Each case may be given points for both variant evidence (see below for details on interpretation) and segregation evidence (see p. 13 for details on calculation).

General Notes for variant scoring:

1. When curating an autosomal dominant disease or an X-linked disorder consider the evidence types in row "A". If you are curating an autosomal recessive disease, consider the evidence types in row "B". In X-linked disorders, affected probands will often be hemizygous males (in the case of truly "recessive" disorders) and/or heterozygous females (in the case of "dominant" disorders). Recognizing that there can be rare cases of females affected by X-linked recessive disorders (due to chromosomal aneuploidy, skewed X inactivation, or homozygosity for a sequence variant) evaluators must be aware of the nuances of interpretation of individual cases and X-linked pedigrees. Points can be assigned at the discretion of the expert reviewer taking into account the available evidence. Furthermore, there are known cases of female carriers of X-linked recessive conditions manifesting symptoms that are milder or later in onset compared to males, and scoring of genetic evidence in these examples should be subject to expert review with regard to the assigned gene/disease/inheritance combination.
2. Computational scores (such as conservation scores, constraint scores, *in silico* prediction tools, variation intolerance scores, etc) are often disease and context-dependent and should not be considered as strong pieces of evidence for variant pathogenicity. However, they can be recorded during curation and used as supporting evidence for variant scoring to be confirmed by expert review.

3. For a variant to be considered as disease-causing, its frequency in the general population should be consistent with phenotype frequency, inheritance pattern, and disease penetrance. These can often be located in the literature (See "Literature Search" p. 8), but may also be contributed by experts. The prevalence of the variant in affected individuals should be higher compared to controls.
4. Different scoring from the suggested points/case may be necessary depending on specific case details. Some suggestions are detailed below. For each case information category, a range of points can be given to account for the strength of evidence available (see Figure 3). Within each range, the curator may choose one of the following scores: 0.1, 0.25, 0.5, followed by 0.5 point increments up to the maximum possible score for that category. However, the curator should always document reasons for any deviation in suggested scores for expert review.
5. When scoring variants for Autosomal Recessive disorders, variants should have some evidence to suggest that they are *in trans* in order to be scored.

Variant Evidence:

1. Other variant with gene impact (Missense variants, small in-frame indels, etc.):
   a. At least some impact to gene function must be demonstrated for the case to count. Thus, impact based on predictions only would score less than the default 0.5 points and impact based on functional validation can score 0.5 or above (up to 1.5/case) depending on the validation quality and disease relevance of the functional assay.
   b. Sum up the number of points. The suggested points per case can be found in column "E" (dominant) and "G" (recessive). Total up all of the variant evidence points and place them in "J" (dominant) or "L" (recessive), as appropriate.
2. Predicted or observed null variants (nonsense, frameshift, +/-1,2 splice, whole gene deletion, truncating CNV, etc.):
   a. Assign fewer points if there is alternative splicing or if the LOF variant is near the C terminus and/or NMD is not predicted (**NOTE:** NMD will not occur if the stop codon is downstream of the last 50 bp of the penultimate exon).
   b. Disease mechanism can be assumed loss of function (LOF) if the gene is LOF constrained. However, LOF constraint scores must be interpreted in the context of the gene or disease in question – genes associated with severe, pediatric-onset disorders may appear to be more constrained than adult-onset conditions where overall fitness is not impacted. One place to find a constraint score is ExAC. The LOF constraint score can be found by searching the gene in ExAC (exac.broadinstitute.org) and viewing the

"constraint metric" at the top right of the page for the "LOF" row. The closer the probability of LOF intolerance (pLI) is to 1, the more LOF-constrained the gene.

   c. Sum up the number of points. The suggested points per case can be found in column "D". Total up all of the variant evidence points and place them in "I".

3. *De novo* variants:
   a. These can be any type of variant, but should be given points depending on statistical expectation of de novo variation in the gene in question. In some cases, this can be found in the literature and should be noted if found (See "literature search" p. 8). However, the curator may also leave this to be supplied by experts during curation review.
   b. In order for a variant to be considered truly *de novo*, both parents must be sequenced. Consider increasing the score for the *de novo* variant if maternity and paternity of the proband are confirmed.
   c. Sum up the number of points. The suggested points per case can be found in column "C". Total up all of the variant evidence points and place them in "H".

NOTE: In addition to meeting the above criteria, the variant should not have data that contradicts a pathogenic role, such as non-segregation, etc. If the points given above for the summary matrix exceed the max score, use the Max score found in "M-P" for the summary matrix.

Segregation Analysis:

The use of segregation studies in which family members are genotyped to determine if a variant co-segregates with disease can be a powerful piece of evidence to support or refute a gene-disease relationship. For the purposes of this framework, we are employing a simplified analysis in which we assume the recombination fraction (θ) is zero (i.e. non-recombinants are not observed) to estimate a LOD score (see equations below).

If a LOD score has been calculated by the authors of a paper:

- This LOD score should be documented and may be used to assign segregation points in the scoring matrix (see Fig 6 for suggestions). If provided by the authors, the ClinGen curator should not use the formula(s) below to estimate a new LOD score. If for some reason you do not agree with the published LOD score, do not assign any points and discuss the concerns with the expert reviewers.

If a LOD score has NOT been calculated by the authors of a paper:

- Curators may estimate a LOD score using the simplified formula(s) below if the following conditions are met:

- 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 affected individuals may be within the same generation, or across multiple generations.
- For **recessive disorders**, the estimated LOD score should be calculated using **ONLY families with at least two affected individuals** in the pedigree. Genotypes must be specified for all affected and unaffected individuals counted.
- Families included in the calculation must not demonstrate any non-explainable non-segregations (for example, a genotype^{-}/phenotype^{+} individual in a family affected by a disorder with no known phenocopies). Families with non-explainable non-segregations should not be used in LOD score calculations.

- If any of the previous conditions are not met, do not use the formula(s) below to estimate a LOD score.
- To be conservative in our simplified LOD score estimations, for autosomal dominant or X-linked disorders only affected individuals (genotype+/phenotype+ individuals) or obligate carriers (regardless of phenotype) should be included in calculations.
- Within a given gene-disease curation, if more than one family meets the criteria above for scoring segregation information, sum their LOD scores to determine the appropriate number of points to assign (using the tables in Figures 4 or 5). For example, if Family A has an estimated LOD score of 1.2 and Family B has an estimated LOD score of 1.8, award them a total of 5 points (based on a summed LOD score of 3). This is done automatically when using the Gene Curation Interface.
- Expert reviewers may choose to specify the most appropriate way to approach segregation scoring within their disease domain, including enacting more formal, rigorous LOD score calculations.

For dominant/X-linked diseases:

$$Z \text{ (LOD score)} = \log_{10} \frac{1}{(0.5)^{\text{Segregations}}}$$

Dominant Segregations	LOD	Points
15	4.5	6.5
14	4.2	6
13	3.9	5.5
12	3.6	5.5

11	3.3	5
10	3	5
9	2.7	4.5
8	2.4	4
7	2.1	4
6	1.8	3.5
5	1.5	3
4	1.2	1.5

Figure 4: Dominant/X-linked Segregation Table

For recessive diseases:

$$\text{Z (LOD score)} = \log_{10} \frac{1}{(0.25)^{\text{# of Affected Individuals-1}}\,(0.75)^{\text{# of Unaffected Individuals}}}$$

NOTE: In general, the number of affected individuals - 1 is equal to the number of affected segregations and can be used interchangeably in this equation.

Family Scenario	LOD	Points
2 affecteds/1 unaffected	0.72	1
3 affecteds/1 unaffected	1.3	2.5
4 affecteds/1 unaffected	1.9	3.5
2 affecteds/2 unaffected	0.85	1
3 affecteds/2 unaffected	1.45	2.5
4 affecteds/2 unaffected	2.05	4
2 affecteds/3 unaffected	1	1.5
3 affecteds/3 unaffected	1.5	3
4 affecteds/3 unaffected	2.18	4

Figure 5: Recessive LOD ScoreTable

"Segregation evidence" matrix scoring suggestions for LOD ranges can be found below:

LOD Range	Points
0.72 – 0.99	1
1 – 1.24	1.5
1.25 – 1.49	2.5 (10:1)
1.5 – 1.74	3
1.75 – 1.99	3.5
2 – 2.49	4 (100:1)
2.5 – 2.99	4.5
3 – 3.49	5 (1000:1)
3.5 - 3.99	5.5
4 – 4.49	6
4.5 – 4.99	6.5

(>/=) 5	7 (MAX SCORE for segregation)

Figure 6: Proposed Matrix Scoring for LOD Ranges

Counting Segregations

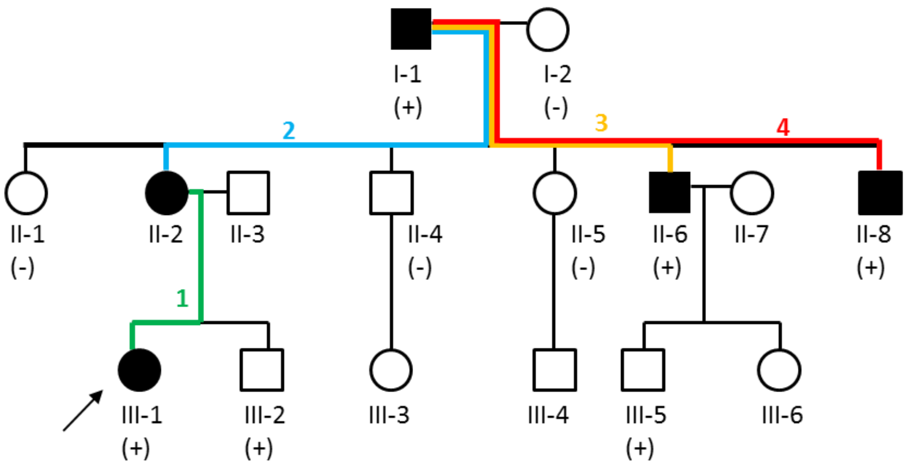
1. In general, the number of segregations in the family will be the number of affected individuals minus one, the proband, to account for the proband's genotype phase being unknown.

For example, **pedigree A** shows a family with hypertrophic cardiomyopathy.

a. There are **four segregations** that can be counted beginning at the proband. This includes the mother (II-2) who is an obligate carrier and can be assumed to be genotype-positive even though she was not tested. Here, there are **four segregations**, resulting in a **LOD score of 1.2** and **1.5 points** can be assigned to this segregation data.

b. For disorders with reduced penetrance such as cardiomyopathy, it is **safest to only use affected genotype + individuals for segregation.** Obligate carriers (individuals who must be carriers, because they have a genotype+ parent and a genotype + child, should also be included, regardless of phenotype)**.** In this case, the absence of a phenotype in two genotype-positive individuals (III-2 and III-5) is considered irrelevant as they can be explained by delayed onset and/or reduced penetrance.

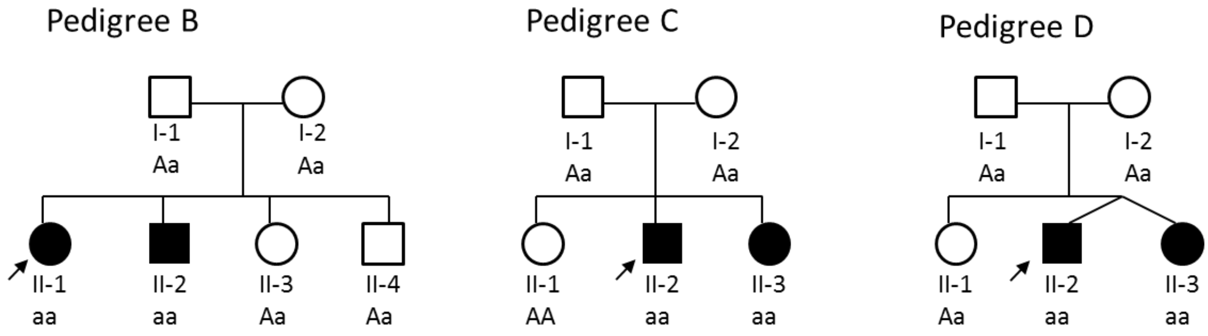
Pedigree A



2. For reasonably penetrant Mendelian disorders, a single LOD score can be calculated across multiple families by counting the number of times a variant segregates with affected individuals to increase the LOD score.

For example, in pedigrees B, C and D, each with fully penetrant recessive hearing loss, there are:

a. 3 affected recessive segregations (one in each family)
b. 4 unaffected individuals (**NOTE:** Count unaffected individuals who would be at the same risk to inherit two altered alleles as an affected individual, i.e. homozygous normal or heterozygous carrier siblings of a proband.)
c. These can be added together across all 3 pedigrees to calculate the LOD score.
d. Here, **3 affected** segregations and **4 unaffected** individuals result in a **LOD score of 2.3**, and **4 points** can be assigned to this segregation data.



3. Fill out the "Segregation evidence" portion of the matrix. If a LOD score is not provided in a publication, a LOD score can be calculated as described in the Segregation Analysis section above and the number of points corresponding to that particular score are recorded in "Q". See proposed scoring ranges in Figure 6 above for guidance. NOTE: If the points given exceed the max score, use the Max score found in "R" for the summary matrix.

Case-Control Data:

Case-Control studies are those in which statistical analysis is used to evaluate enrichment of **variants in cases compared to controls**. Each case-control study should be independently assessed based on the criteria outlined in this section to evaluate the quality of the study design. Consensus with a clinical domain expert group is highly recommended.

1. Case-control studies are classified based on how the study is designed to evaluate variation in cases and controls: **single variant analysis** or **aggregate variant analysis**.
   a. **Single variant analysis studies** are those in which individual variants are evaluated for statistical enrichment in cases compared to controls. More

than one variant may be analyzed, but the variants should be independently assessed with appropriate statistical correction for multiple testing. **For example**, if a study identifies 2 different variants in *MYH7* within a cohort of hypertrophic cardiomyopathy cases, but tests the number of hypertrophic cardiomyopathy cases and unaffected controls that contain only one of the variants and provides a statistic for that variant alone, then the study is classified as a single variant analysis. Similarly, if the same study tests for enrichment of the second variant in the cases and controls and provides a separate statistic for the second variant, this also is a single variant analysis. Often, authors will indicate this either in the article text or in a table of variants.

b. **Aggregate variant analysis** studies are those in which the statistical enrichment of two or more variants as an aggregate is assessed in cases compared to controls. This comparison could be accomplished by genotyping specific variants or by sequencing the entire gene. For example, if a study identifies 2 different variants in *MYH7,* and then statistically tests the enrichment of both the variants in hypertrophic cardiomyopathy cases over unaffected controls, an aggregate variant analysis was conducted.

2. Case-control studies should be assigned points at the discretion of expert opinion based on the overall quality of each study. Assign each study a number of points between 0-6, then sum the points given to all studies, and fill in "S". NOTE: If the points given exceed the max score, use the Max score found in "T" for the summary matrix.
3. The quality of each case-control study should be evaluated using the following criteria in aggregate:
    a. **Variant Detection Methodology:** Cases and controls should ideally be analyzed using methods with equivalent analytical performance (e.g. equivalent genotype methods, sufficient and equivalent depth and quality of sequencing coverage).

    b. **Power:** The study should analyze a number of cases and controls given the prevalence of the disease, the allele frequency, and the expected effect size in question to provide appropriate statistical power to detect an association. (**NOTE:** The curator is NOT expected to perform power calculations, but to record the information listed in this section for expert review.)

    c. **Bias and Confounding factors:** The manner in which cases and controls were selected for participation and the degree of case-control matching may impact the outcome of the study. The following are some factors that should be considered:

i. Are there systematic differences between individuals selected for study and individuals not selected for study (i.e. do the cases and controls differ in variables other than genotype)?

ii. Are the cases and controls matched by demographic information (e.g., age, ethnicity, location of recruitment, etc.)? Are the cases and controls matched for genetic ancestry, if not did investigators account for genetic ancestry in the analysis?

iii. Have the cases and controls been equivalently evaluated for presence or absence of a phenotype, and/or family history of disease?

d. **Statistical Significance**: The level of statistical significance should be weighed carefully.

i. When an odds ratio (OR) is presented, its magnitude should be consistent with a monogenic disease etiology.

ii. When p-values or 95% confidence intervals (CI) are presented for the OR, the strength of the statistical association can be weighed in the final points assigned.

iii. Factors, such as multiple testing, that might impact that interpretation of uncorrected p-values and CIs should be considered when assigning points.

NOTE: Point totals should NOT exceed the max score. If the totals from "H-Q" exceed the max score, use the max score found in "U" for the genetic evidence portion of the summary matrix. Please prioritize curating genetic evidence over experimental evidence to reach a definitive score.

Figure 7: Case-control Genetic Evidence Examples

Detailed explanations for assigned points are provided below the table.

CASE-CONTROL DATA						
Points	**Power**	**Bias/ Confounding**	**Detection Method**	**Statistical Significance**	**Study Type**	**Points (0-6/ study)**
Author A 2015 (Max score)	Breast cancer cases: 100/12,000 Controls: 7/4,500	Matched by age, ethnicity, and location	Cases & controls genotyped for c.1439delA in gene *W*	OR: 5.4 [95% CI: 2.5-11.6; *P* < 0.0001]	Single Variant	6
Author B 2005 (Intermediate score)	HCM Cases: 13/200 Controls: 20/900	Matched by location, but not age or ethnicity	Cases & controls genotyped for p.Arg682Gln in gene *X*	Fisher's exact test *P* = 0.004	Single Variant	4
Author C 2011 (Low score)	Ovarian cancer cases: 11/1,500 Controls: 3/2,000	Matched by ethnicity. Controls from population database (e.g.	Cases: sequenced Gene *Y* and counted all cases with null variants.	OR of all variants in aggregate: 4.9 (CI: 1.4-17.7; *P* =0.015)	Aggregate analysis	2

		ExAC)	Controls: total individuals from population database with null variants in gene *Y*.			
Author D 2009 (No case-control score)	Colorectal cancer cases: 11/1,500 Controls: 3/2,000	Matched by ethnicity. Controls from population database (e.g. ExAC)	Cases: sequenced gene *Z* and identified p.Lys342Ter in 11 cases. Controls: total individuals from population database with p.Lys342Ter in gene Z.	OR of p.Lys342: 4.9 (CI: 1.4-17.7; *P* =0.015)	Not applicable	0

Study receiving the max score (6 points): This single-variant analysis could receive the full 6 points based on the number of appropriately matched (i.e. no Bias or Confounding factors in study design) cases and controls analyzed (i.e. Power was sufficient given the prevalence of breast cancer as a disease) and the OR was highly statistically significant (P<0.0001) with a 95% CI that did not cross 1.0.

Study receiving intermediate score (4 points): This single-variant analysis could receive 4 points since the controls were not appropriately matched to the cases (i.e. by location alone and neither by ethnicity nor age) and the p-value is moderately significant.

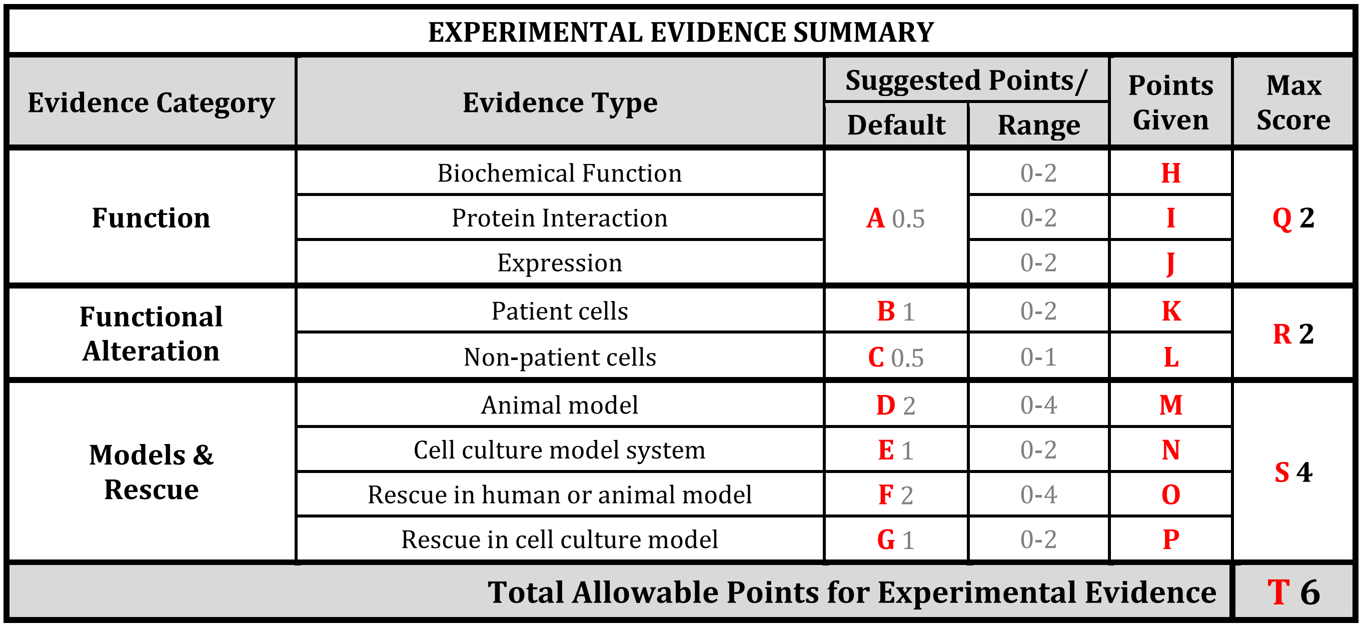
Study receiving low score (2 points): This study is considered an aggregate analysis since the statistical test analyzed the variants in aggregate across all cases and controls. This study can be assigned 2 points because a population database was used rather than appropriately-matched controls (i.e. the study is not matched demographically) and the p-value is not very significant. A population database could be used as controls for 2 reasons:

a. Both the cases and controls were sequenced for the entire gene *Y*.
b. The total number of individuals with null variants (i.e. nonsense, canonical splice-site, and frameshift) was compared between cases and controls.

Study receiving no score (0 points): While this study is similar to the study receiving 2 points, the detection method differed between cases and controls (i.e. cases were sequenced, controls were genotyped). In the cases, gene *Z* was sequenced. However, only the controls with a specific variant were used for comparison to the cases. Although this study cannot be counted as case-control data, it can be counted as case-level data.

EXPERIMENTAL EVIDENCE

Figure 8: Experimental Evidence Summary Matrix



EXPERIMENTAL EVIDENCE SUMMARY					
Evidence Category	Evidence Type	Suggested Points/		Points Given	Max Score
		Default	Range		
Function	Biochemical Function	A 0.5	0-2	H	Q 2
	Protein Interaction		0-2	I	
	Expression		0-2	J	
Functional Alteration	Patient cells	B 1	0-2	K	R 2
	Non-patient cells	C 0.5	0-1	L	
Models & Rescue	Animal model	D 2	0-4	M	S 4
	Cell culture model system	E 1	0-2	N	
	Rescue in human or animal model	F 2	0-4	O	
	Rescue in cell culture model	G 1	0-2	P	
Total Allowable Points for Experimental Evidence					T 6

NOTE: Validated functional assays should be identified by expert panels or, if they are curator identified, confirmed by expert review. Identify the experimental evidence type and assign points according to the following criteria:

1. Biochemical Function: Summarize evidence showing the gene product performs a **biochemical function** shared with other known genes in the disease of interest, or consistent with the phenotype. The suggested points/evidence can be found in column "A". Total up all of the experimental points and place them in the points given section found in "H".
2. Protein Interaction: Summarize evidence showing the gene product **interacts** with **proteins previously implicated** (genetically or biochemically) in the disease of interest. Typical examples of this data are: Physical interaction via Yeast-2-Hybrid (Y2H) and/or co-immunoprecipitation (coIP). The suggested points/evidence can be found in column "A". Total up all of the experimental points and place them in the points given section found in "I".
3. Expression: Summarize evidence showing the gene is expressed in **tissues relevant to the disease of interest** and/or is **altered in expression in patients**

who have the disease. Typical examples of this data type are: methods to detect a) RNA transcripts (RNAseq, microarrays, qPCR, qRT-PCR, Real-Time PCR) b) protein expression (western blot, Immunohistochemistry). The suggested points per evidence can be found in column "A". Total up all of the experimental points and place them in the points given section found in "J".

NOTE: If the sum of all biochemical function, protein interaction, and expression points exceeds the max score of 2, use the Max score found in "Q" of the experimental evidence summary matrix.

4. Functional Alteration: Summarize evidence showing the gene and/or gene product **function** is demonstrably altered in cell culture models and/or patients carrying candidate mutations. For instance, does disrupting the gene in cells have a phenotype similar to that in human patients? Examples include experiments involving gene knock-down, overexpression, etc. Divide the evidence according to the following subtypes:
    a. Was the experiment conducted in **patient cells**? The suggested points/evidence can be found in column “B”. Total up all of the experimental points and place them in the points given section found in "K".
    b. Was the experiment conducted in **non-patient cells**? The suggested points/evidence can be found in column “C”. Total up all of the experimental points and place them in the points given section found in "L".

    NOTE: If the sum of all functional alteration points exceeds the max score of 2, use the Max score found in "R" of the experimental evidence summary matrix.
5. Model System: A non-human **animal** or **cell culture model** with a disrupted copy of the gene shows a phenotype consistent with the human disease state. These results should be summarized accordingly:
    a. Was the gene disruption in a non-human **animal model**? The suggested points/evidence can be found in column “D”. Total up all of the experimental points and place them in the points given section found in "M".
    b. Was the gene disrupted in a **cell culture model**? The suggested points/evidence can be found in column “E”. Total up all of the experimental points and place them in the points given section found in "N”
6. Rescue: Summarize evidence showing the **phenotype in cell culture models** or in **humans** or in **non-human animal models** can be rescued by exogenous wild-type gene product. These results should be recorded accordingly:

a. Was the rescue in a **human** or in a **non-human animal model**? The suggested points/evidence can be found in column “F”. Consider awarding more points if the rescue was in a human (for example, successful enzyme replacement therapy for a lysosomal storage disease). Total up all of the experimental points and place them in the points given section found in "O”.
b. Was the rescue in a **cell culture model** (*i.e.* a cell culture model engineered to express the variant of interest) The suggested points/evidence can be found in column “G”. Consider awarding more points if the cells are patient-derived. Total up all of the experimental points and place them in the points given section found in "P”.
NOTE: If the sum of all models and rescue points exceeds the max score of 4, use the Max score found in "S" of the experimental evidence summary matrix.

Total up the total number of experimental evidence points from Rows “H-P” and enter them on Row “T”. **NOTE:** If the total experimental evidence points exceed the max score, use the Max score of 6 points for the summary matrix. **Please prioritize curating genetic evidence over experimental evidence to reach a definitive score.**

Experimental Evidence Examples:

A. Lan et al 2013; PMID: 23290139

Abbreviated Abstract: Familial hypertrophic cardiomyopathy (HCM) is a prevalent hereditary cardiac disorder linked to arrhythmia and sudden cardiac death. While the causes of HCM have been identified as genetic mutations in the cardiac sarcomere, the pathways by which sarcomeric mutations engender myocyte hypertrophy and electrophysiological abnormalities are not understood. To elucidate the mechanisms underlying HCM development, we generated patient-specific induced pluripotent stem cell cardiomyocytes (iPSC-CMs) from a ten-member family cohort carrying a hereditary HCM missense mutation (Arg663His) in the MYH7 gene.

1) Functional Alteration, patient cells (1 point): iPSCs were generated from dermal fibroblasts from a family of 10 individuals with HCM. They were differentiated into cardiomyocytes and cells with the variant of interest (Arg663His in MYH7) recapitulated the HCM phenotype, including cellular enlargement (measured by imaging) and contractile arrhythmia (measured by whole-cell patch clamping).

2) Function, Biochemical function (1 point was given instead of 0.5 because two methods were used to demonstrate altered biochemical function): Calcium handling is critical for excitation-contraction coupling, which can be altered in HCM. A) Cells expressing the Arg663His MYH7 variant demonstrated calcium handling abnormalities, which were measured by using fluorescent calcium dye. B) Overexpression of the variant in healthy cells also produced these abnormalities.

B. van de Laar et al 2010; PMID: 21217753

Abbreviated Abstract: Thoracic aortic aneurysms and dissections are a main feature of connective tissue disorders, such as Marfan syndrome and Loeys-Dietz syndrome. We delineated a new syndrome presenting with aneurysms, dissections and tortuosity throughout the arterial tree in association with mild craniofacial features and skeletal and cutaneous anomalies. In contrast with other aneurysm syndromes, most of these affected individuals presented with early-onset osteoarthritis. We mapped the genetic locus to chromosome 15q22.2-24.2 and show that the disease is caused by mutations in SMAD3.

1) Function, Expression (0.5 points): 99 individuals with thoracic aortic aneurysms were sequenced for alterations in SMAD3. Variants were identified in 2 out of 99 cases. One of the variants was a frameshift the resulted in reduced mRNA expression, likely due to nonsense- mediated RNA decay (NMD), which was evaluated by cDNA sequencing.

2) Functional alteration, non-patient cells (0 points because data were not shown): Fibroblasts expressing the SMAD3 frameshift variant were treated with cyclohexamide and mRNA levels of SMAD were restored to nearly wild type levels, suggesting NMD occurs.

C. Li et al 2015; PMID: 25331638

Abbreviated Abstract: Autosomal-recessive nonsyndromic hearing loss (ARNSHL) features a high degree of genetic heterogeneity. Many genes responsible for ARNSHL have been identified or mapped. We previously mapped an ARNSHL locus at 17q12, herein designated DFNB99, in a consanguineous Chinese family. In this study, whole-exome sequencing revealed a homozygous missense mutation (c.1259G>A, p.Arg420Gln) in the gene-encoding transmembrane protein 132E (TMEM132E) as the causative variant.

1) Function, Expression (0.5 points): qPCR was used to demonstrate that TMEM132E was highly expressed in the cochlea and the brain, two tissues that can be affected by hearing loss. Western blotting confirmed that the protein also followed this pattern.

2) Model Systems, animal model (1 point): TMEM132e was knocked down in zebrafish using antisense morpholino oligos. The morpholino animals displayed delayed startle response and reduced extracellular microphonic potentials, suggesting hearing loss.

3) Rescue, animal model (1 point was given instead of 2 because the gene only partially rescues the phenotype): Human TMEM132E mRNA was injected into the antisense oligo knockdown animals. This partially rescued the hearing defects in those fish.

CONTRADICTORY EVIDENCE

NOTE: This designation is to be applied at the discretion of clinical domain experts after thorough review of available evidence. The curator will collect the contradictory evidence and the classification (Disputed/Refuted) is to be determined by the clinical domain experts. Below are a few categories of the most common types of contradictory evidence. Note that this list is not all-inclusive and if the curator feels that a piece of evidence offers evidence that does not support the gene-disease relationship, this data should always be recorded (Summary and PMIDs) and pointed out for expert review.

1. **Case-control data is not significant:** As case-control studies evaluate variants in healthy vs affected individuals, if there is no statistically significant difference in the variants between these groups, this should be marked as potentially contradictory evidence for expert review. **See case-control examples above (p.19, Fig. 7) NOTE:** Evidence contradicting a single **variant** as causative for the disease does not necessarily rule out the gene:disease relationship.

2. **Minor allele frequency is too high for the disease:** Many diseases have published prevalence, which can often be found in the GeneReviews entry. If ALL minor alleles in a gene are present in a specific population or the general population (ExAC, ESP, 1000Genomes) at a frequency that is higher than what is estimated for the disease, this could suggest lack of gene-disease relationship and should be marked as potentially contradictory evidence for expert review. **For example,** Adams-Oliver syndrome is an autosomal dominant disease and has a prevalence of 0.44 in 100,000 (4.4e-6) live births. If a new gene were being curated for this disease and supposedly pathogenic variants were identified with an allele frequency in ExAC of 0.4882, this could be potentially contradictory evidence. **NOTE:** Evidence contradicting a single **variant** as causative for the disease does not necessarily rule out the

gene:disease relationship. Additionally, disease prevalence can vary in different populations, so read the GeneReviews entry thoroughly and keep demographic information in mind during this evaluation.

3. **The gene-disease relationship cannot be replicated:** One measure of a gene-disease relationship is its replication both over time and across multiple studies and disease 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, this could be considered potentially contradictory evidence and should be marked for expert review. **However,** when assigning this designation, a curator need consider disease prevalence. If a disease is rare, a small study may not identify any variants in the curated gene. **For example,** Perrault syndrome is characterized by hearing loss in males and ovarian dysfunction in females and only 100 cases have been reported. Thus, if a study with a small cohort does not identify any variants in a gene being curated for this syndrome, this may not necessarily be evidence against gene-disease association. In any case, if a curator suspects that any evidence supports a lack of gene-disease association, it should be marked for expert review.

4. **Non-segregations:** Non-segregations should be considered carefully, as age-dependent penetrance and phenotyping of relatives could have an impact on the number of non-segregations within a family. Thus, the age of unaffected variant carriers should be of similar age to the affected variant carriers. If a curator suspects non-segregations, these should be noted for expert review.

5. **Non-supporting functional evidence:** The types of different experimental evidence are detailed in the **"Experimental Evidence" Section (p. 21)**. If any of this experimental evidence suggests that variants, although found in humans, do not affect function or that the function is not consistent with the established disease mechanism, this evidence should be marked as potentially contradictory evidence for expert review. **For example**, if a gene were being curated for a disease association and the mouse model did not have any phenotype, this could be potentially contradictory evidence.

SUMMARY AND FINAL MATRIX

A summary matrix was designed to generate a “provisional” clinical validity assessment using a point system consistent with the qualitative descriptions of each classification. This final gene curation matrix and instructions for filling it out can be found below.

Fill in the "Gene/Disease Pair" at the top of the matrix.

1. Enter the score calculated for Genetic Evidence Matrix (Fig. 3 p. 10) in row "A".
2. Enter the score calculated for Experimental Evidence Matrix (Fig. 8 p. 21) in row "B".
3. The sum of A and B is entered in row "C".
4. Refer to the publication date of the original publication of the gene-disease relationship and consider all other literature to complete row "D":
    a. YES if > 3 years have passed since the original publication AND there are >2 publications about the gene-disease relationship
    b. NO if >3 years have passed, BUT not >2 publications
    c. NO if < 3 years have passed
5. If there is valid contradictory evidence (see p. 25), compile this and briefly describe it (including the PubMed ID numbers) in row "E".
6. Choose the clinical validity classification associated with the value of the total points/replication over time (Row C, Row D) and complete row "F".
   NOTE: No matter the score, if there is contradictory evidence present, the curator classification must be listed as "Conflicting Evidence reported". The conflicting evidence will be weighed and reviewed by a domain expert.
7. When the gene-disease curation is reviewed by an expert, the expert will fill out the final classification in row "G".

Figure 9: Clinical Validity Summary Matrix

<table>
<tr><th colspan="5">GENE/DISEASE PAIR:</th></tr>
<tr><th>Assertion criteria</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>
<tr><td><b>Description</b></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 & Experimental Evidence</td><td>> 2 pubs w/ convincing evidence over time (>3 yrs)</td></tr>
<tr><td><b>Assigned Points</b></td><td>A</td><td>B</td><td>C</td><td>D</td></tr>
<tr><td colspan="2" rowspan="4"><b>CALCULATED CLASSIFICATION</b></td><td><b>LIMITED</b></td><td colspan="2"><b>1-6</b></td></tr>
<tr><td><b>MODERATE</b></td><td colspan="2"><b>7-11</b></td></tr>
<tr><td><b>STRONG</b></td><td colspan="2"><b>12-18</b></td></tr>
<tr><td><b>DEFINITIVE</b></td><td colspan="2"><b>12-18
& Replicated Over Time</b></td></tr>
<tr><td><b>Valid contradictory evidence (Y/N)*</b></td><td colspan="4"><b>List PMIDs and describe evidence:</b>
E</td></tr>
<tr><td colspan="2"><b>CURATOR CLASSIFICATION</b></td><td colspan="3">F</td></tr>
<tr><td colspan="2"><b>FINAL CLASSIFICATION</b></td><td colspan="3">G</td></tr>
</table>