Gene Clinical Validity Curation Process

Standard Operating Procedure

Version 8
October 2020
The Clinical Genome Resource
Gene Curation Working Group

Table of Contents

BACKGROUND **3**

REQUIRED COMPONENTS **3**

OVERVIEW OF GENE CURATION **4**

Figure 1: GENE CURATION WORKFLOW **5**

CLINICAL VALIDITY CLASSIFICATIONS **6**

DEFINING THE DISEASE ENTITY **9**

LITERATURE SEARCH **13**

GENETIC EVIDENCE **14**

Genetic Evidence Summary Matrix *15*

Case-Level Data *16*

Additional Case-Level Scoring Considerations 20

Segregation Analysis 27

Case-Control Data *36*

Figure 8: Case-control Genetic Evidence Examples 38

EXPERIMENTAL EVIDENCE **39**

Figure 9: Experimental Evidence Summary Matrix *40*

Case-level Variant Evidence vs. Experimental Evidence 43

CONTRADICTORY EVIDENCE **44**

SUMMARY & FINAL MATRIX **46**

RECURATION PROCEDURE **49**

SOP REFERENCES **50**

APPENDIX A: USEFUL WEBSITES FOR CLINGEN GENE CURATORS **50**

APPENDIX B: EXPERIMENTAL EVIDENCE EXAMPLES **54**

APPENDIX C: SEMIDOMINANT MODE OF INHERITANCE OVERVIEW **58**

BACKGROUND

ClinGen's gene curation process is 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 is compiled and used to assign a clinical validity classification per criteria established by the ClinGen Gene Curation Working Group (GCWG) [1]. This protocol details the steps involved in curating a gene-disease relationship and subsequently assigning a clinical validity classification. This curation process is not intended to be 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. While the following protocol provides guidance on the curation process, professional judgment and expertise, where applicable, must be used when deciding on the strength of different pieces of evidence that support a gene-disease relationship.

REQUIRED COMPONENTS

- ClinGen-approved curation training. For training resources please see the ClinGen gene curation website here or contact clingen@clinicalgenome.org.
- The ClinGen Lumping and Splitting guidelines must be consulted to determine the disease entity for curation. Please see guidelines here.
- Access to scientific articles and publications
- Access to the ClinGen Gene Curation Interface (GCI), found here:
    - Access is granted to users that are actively participating on a ClinGen gene curation expert panel (GCEP). Coordinators for the GCEP are responsible for setting up accounts and permissions. If you have trouble accessing the GCI once an account is set up, please contact clingen-helpdesk@lists.stanford.edu.
    - For help with data entry into the Gene Curation Interface, please see the GCI Help document: https://github.com/ClinGen/clincoded/wiki/GCI-Curation-Help or contact clingen-helpdesk@lists.stanford.edu.

Optional: An SOP has been developed to assist in evidence collection through the use of a web-based annotation tool, called Hypothes.is, that allows annotation of web-based publications. Use of this tool has been shown to reduce curation time and facilitate data transfer into the GCI. This is a standalone tool at this time and could be used by the individual or within Expert panels based on forming a group in Hypothes.is. Access to the Hypothes.is Gene Annotation SOP can be found here, or on

the ClinGen website under the Gene Curation Training Materials, Supporting Materials Section.

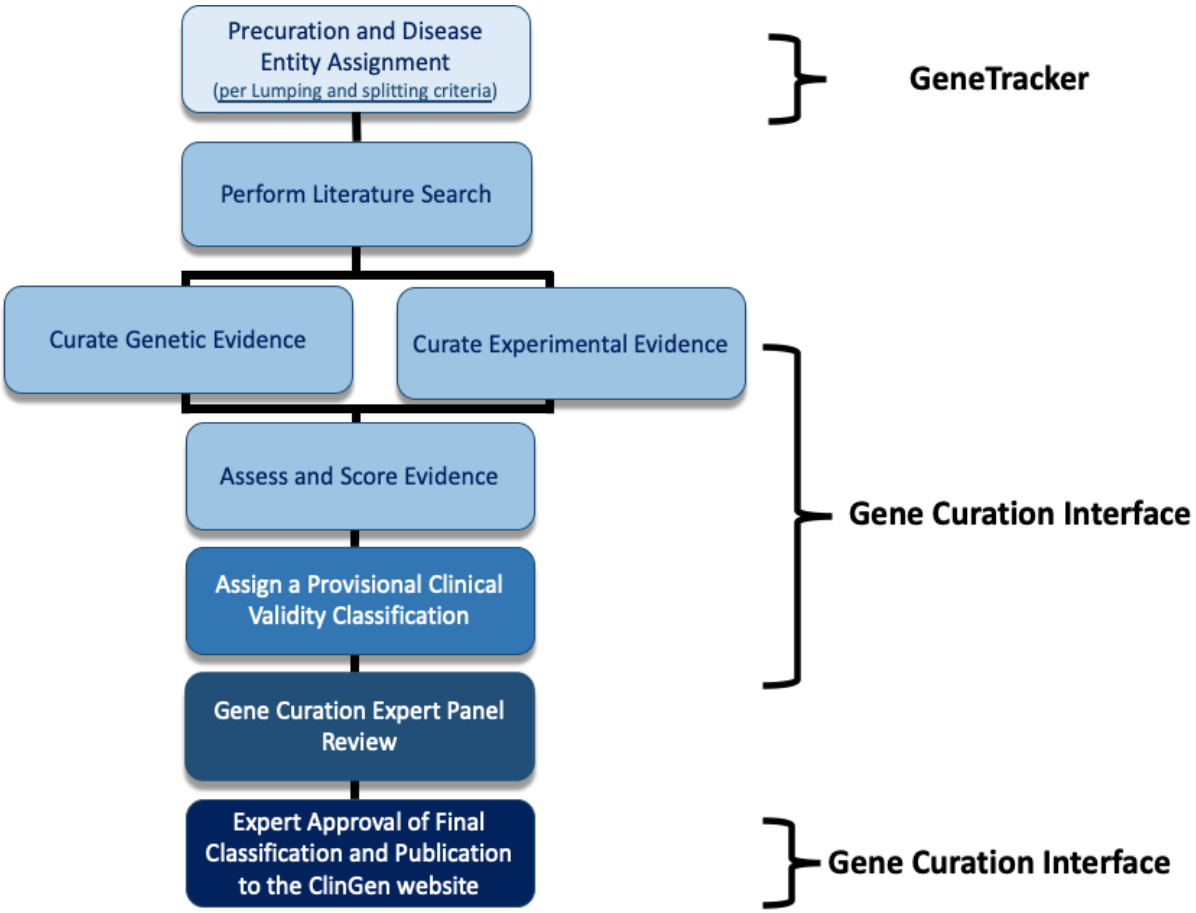
OVERVIEW OF GENE CURATION

The gene curation framework consists of the following essential steps in order to assign a clinical validity classification for a gene-disease relationship (see Figure 1 for a visual representation of the curation workflow):

- Establishing the gene-disease-mode of inheritance to be used in curation
- Evidence collection
- Identification of evidence types
    a. Genetic Evidence
    b. Experimental Evidence
- Evaluation and scoring of evidence
- Expert Review, final classification and approval of a gene-disease relationship

In the subsequent sections of this document, each step will be outlined in detail and general recommendations provided. It is important to note that expert panels may provide specific recommendations for evidence inclusion and scoring for gene-disease relationships under their purview; therefore, final consultation, review, and approval of the evidence with the expert panel is paramount before publishing a clinical validity classification.

Figure 1: GENE CURATION WORKFLOW



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 contradicts any claimed relationship (Figure 2). 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 assertion of a causal, disease related variant in humans is "No Known Disease Relationship" (**NOTE:** prior to August 2019, this category was referred to as "No Reported Evidence"). The level of evidence needed for each supportive gene-disease relationship category builds upon that of the previous category (e.g. "Moderate" builds upon "Limited"). Gene-disease relationships classified as "Contradictory" likely have evidence supporting as well as opposing the gene-disease association. In these cases, the strength of evidence supporting versus opposing the gene-disease relationship should be weighed by the expert panel before a final clinical validity classification is assigned.

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 (at least 2 independent publications over 3 years' time). 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 KNOWN DISEASE RELATIONSHIP[3]		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. A tag designating "animal model only" is applied on clinicalgenome.org for those

		gene-disease pairs in which no human genetic evidence has been asserted, but an animal model exists.
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. Disputing 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. c. While it is nearly impossible to entirely refute a gene's potential role in disease, this category is to be used when all existing data has been fully refuted leaving the gene with essentially no valid evidence remaining, after an original claim.

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.13 for more information.

[2]Examples of appropriate types of supporting experimental data based on those outlined in MacArthur et al. 2014 [2].

[3]As of August 2019, NO REPORTED EVIDENCE has been changed to NO KNOWN DISEASE RELATIONSHIP per the survey results from the Gene Curation Coalition (GenCC). The GCI and website team will facilitate the term change for legacy curations.

ESTABLISHING THE GENE-DISEASE-MODE OF INHERITANCE

Prior to the collection of evidence, it is important to establish the disease entity and mode of inheritance (MOI) that will be curated for the gene in question. Once established, the gene-disease-MOI represents a curation record and allows a curator to begin a curation in the GCI. Once a group has established the appropriate gene-disease-MOI, it should be recorded in the ClinGen GeneTracker before proceeding with curation in the GCI (Figure 1). Contact your GCEP coordinator to understand the responsible party for entering the record for your specific affiliation, as it varies by groups, and if you note any discrepancies between GeneTracker and GCI records. Below are recommendations specific to ascertaining a gene-disease-MOI:

Gene: Gene(s) of interest may be assigned to a curator based on the approved gene list for a GCEP in which they are a member. Only the HGNC approved gene symbol can be used to create a gene-disease-MOI curation record in the GCI. However, use of gene aliases (including previously approved symbols and protein names) may facilitate identification of applicable evidence for inclusion in the curation, including literature and online curatorial resources such as gnomAD, HGNC, NCBI Gene, and Ensembl are a few examples of websites that provide gene aliases and synonyms.

Currently, the GCI will only allow a single record for a given gene-disease-MOI. This is to limit the number of clinical validity classifications assigned to one gene-disease-MOI and reduce redundancy of curations among the various GCEPs. In order to check the current status of a gene-disease-MOI record, curators are directed to search the ClinGen GeneTracker before beginning a curation. Access to the tracking system is determined by your GCEP. Therefore, check with your GCEP coordinator before proceeding with a curation.

DEFINING THE DISEASE ENTITY

Many human genes are implicated in more than one disorder. Therefore, prior to starting a curation and entering details into the GCI, a curator should be absolutely clear on which disease entity is being curated based on the Lumping and Splitting guidelines. A video tutorial on the Lumping and Splitting process is available here. To facilitate defining a disease entity, curators may be asked to perform and present a gene precuration to a GCEP prior to collecting and/or entering evidence into the GCI. After review and discussion, the GCEP will determine which disease entity or entities to move forward with. This can be done offline, or as part of a regularly scheduled meeting at the GCEP's discretion, but should occur before the

curator begins entering information into the GCI. Templates and examples of gene precurations are provided by the Lumping and Splitting Working Group here. Furthermore, all precuration data should be entered into the ClinGen GeneTracker before proceeding with the curation within the GCI (Figure 1). This is to ensure that there is no overlap of curations among different GCEPs as it houses the precuration information, curation status, and expert panel affiliation for all genes in current consideration over all of ClinGen's GCEPs. In addition, entry of this data in the GeneTracker will facilitate display of the lumping and splitting decisions on clinicalgenome.org.

NOTE: Once a curation is started in the GCI, the only mechanism for changing a disease entity is to contact the GCI Help Desk.

Mode of inheritance (MOI): Like disease entities, a gene may also be associated with multiple inheritance patterns. Common MOIs include autosomal dominant, autosomal recessive, X-linked, and semidominant. A list of the MOIs available in the GCI, as well as an outline on the ability to score and/or publish a classification is included in Table 1. Many of the MOIs have associated "adjectives" or distinguishing characteristics, such as imprinting, sex-linked, etc. At this time the use of an "adjective" is optional, and not required to generate a gene-disease-MOI record or a clinical validity classification. Curators may also discuss with the GCEP which MOI is most appropriate during the precuration process.

For genes in which both monoallelic (e.g. autosomal dominant) and biallelic (e.g. autosomal recessive) genetic variation are known to have the same molecular mechanism and result in the same disease entity with varying severity of the phenotype(s), we recommend the use of the semidominant MOI option in the GCI. According to the Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine (2006), semidominance refers to the presentation of phenotypes given the expression of alleles, in which the heterozygous state (A/a) represents an intermediate phenotype (as a/a refers to the wild-type) compared to the homozygous mutant state (A/A), which would be more severe and or earlier onset [3]. An example of semidominance would be the gene-disease relationship between *LDLR* and familial hypercholesterolemia (FHC), in which the autosomal dominant (heterozygous, monoallelic mutant, A/a) form of FHC is adult onset with variable presentation and penetrance of hypercholesterolemia, whereas the autosomal recessive (biallelic mutant form, A/A) form of FHC is severe, childhood onset. Further information on the use of the semidominant MOI can be found in Appendix C. More information on determining disease entities based on inheritance pattern difference, see the Lumping and Splitting guidelines, here.

At this time there are 2 MOIs that cannot be scored in the GCI (Mitochondrial inheritance and Undetermined MOI) (Table 1)**.** For these choices, manual modification of the clinical validity classification in the GCI (on the classification matrix page) is required in order to approve and publish the gene-disease-MOI record to the ClinGen website. In general, gene-disease relationships with a MOI of "Undetermined" should not be classified above "limited," however consulting with the expert panel is encouraged before a final clinical validity classification is assigned. Of note, if "Mitochondrial" or "other", and any adjectives under this choice (including Y-linked, somatic mutation, multifactorial inheritance, and codominance) are the MOIs chosen for a gene-disease relationship, the final clinical validity classification will NOT be permitted to be published on the ClinGen website. Therefore, use caution when making these choices. If you have made an error in the choice of MOI for a gene-disease relationship, please contact the GCI Help Desk.

Table 1. Mode of Inheritance (MOI) choices in the GCI

MOI type	Score in GCI	GCI Calculated classification	GCI Modified classification	Publish to website
Autosomal Dominant (HP:0000006)	✓	✓	✓	✓
Autosomal Recessive (HP:0000007)	✓	✓	✓	✓
Mitochondrial (HP:0001427)	✗	✗	✓	✗
Semidominant (HP:0032113)	✓	✓	✓	✓
X-linked (HP:0001417)	✓	✓	✓	✓
Undetermined MOI (HP:0000005)	✗	✗	✓	✓
Other (includes: Y-linked, Somatic, Multifactorial, and Codominant inheritance)	✗	✗	✗	✗

EVIDENCE COLLECTION

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. At this time only evidence that has an associated PMID can be recorded and scored in the GCI. For larger databases that list multiple variants or case reports, check whether the database includes citations with PMIDs, as this will allow you to score evidence in the GCI. For example, a well-known database called DECIPHER houses a collection of case-level evidence for individuals with genetic conditions. The DECIPHER website contains a section entitled "Citing DECIPHER" that provides a link to the seminal paper, which has a PMID associated (PMID: 19344873). Should a curator choose to use evidence from this database (i.e., the evidence has been deemed appropriate for inclusion by the GCEP), the curator could use this PMID to enter the applicable information on a gene-disease relationship of interest, given further guidance provided below in the Genetic Evidence section. *Check with your GCEP(s) to determine well-known and trusted public databases containing clinical data pertinent to your group, and to determine in which circumstances these cases may be used.* In the event that case report(s) from a database are used for a curation, it is recommended that the identifying case report number is used as the "Individual Label" in the GCI. Furthermore, if applicable, add the URL from the database on any individual, family, or group evidence in the appropriate "Explanation" section when applying a score for the evidence. For a list of general databases of interest and associated PMIDs for scoring, please see Appendix A.

Useful publication search engines: There are several web-based scholarly search engines, and a few of the most widely used for gene curation include:

- PubMed
  - PubMed tutorial
- Google Scholar
  - Has a full-text search feature
  - Google Scholar search tips
- LitVar
  - Allows searching by a variant RefSeq number
- GeneCards
  - Search by gene name
  - Under the "Publications" section
- Mastermind
  - Can search by gene and variant (free version)

        - Standard version is free. Professional version requires a subscription, and only this version can search by disease and in the supplemental data.
    - In general, advanced searches on many of these databases are more informative.

NOTE: One need not comprehensively curate all evidence for a gene-disease relationship (particularly for "Definitive" associations), but instead focus on curating and evaluating the relevant pieces of evidence described in this protocol. **Once you have reached the maximum number of points for a given category, it is not necessary to document further evidence within that category.**

LITERATURE SEARCH

- 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 aliases and synonyms before you search (see "Gene" section above for recommended sites for gene aliases).
    - NOT all search results will be relevant, thus it is important to examine the search results for pertinent information.

- Curating primary literature is encouraged, but if a gene-disease relationship has abundant information (i.e. >100 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]).**
    - Curation may occur from that publication **ONLY** when sufficient details are included in the review article.
    - If sufficient details are **NOT** included in the review article, then the curator will need to return to each original citation to curate the information.

- 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 assertion of the gene-disease relationship in humans was first made.
    - Search PubMed for experimental data (Examples below)
        - [gene] AND [gene function] e.g. [KCNQ1] AND [potassium channel]

    - [protein] AND [function] e.g. [neurofibromin] AND [tumor suppressor]
    - [gene] AND [animal] e.g. [ACTN2] and [mouse OR zebrafish OR xenopus OR drosophila]
  - Additional information may also be available in OMIM in the **"Gene function"** or **"Biochemical Features"** or **"Animal Model"** sections.
  - GeneReviews often has information in the **"Molecular Genetics"** section of the disease entries that may be useful.
  - Other databases such as UniProt , MGI , etc. may also be useful, provided that primary references (and PMIDs) are given that can be curated. For a list of databases that may be helpful for the curation process, see Appendix A.
  - GeneRIFs (Gene Reference Into Function), 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.

- An additional component of the curation process is to determine if evidence supporting the original gene-disease relationship has been replicated; therefore, it is critical to find the **original paper** initially asserting the proposed relationship, as well as others, ideally from independent groups. 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.

Once all of the relevant literature about the gene-disease relationship has been assembled, curation of the different pieces of evidence can begin.

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 for 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 for detailed discussion within the publication, and familial inheritance and pedigree information were provided, this case could be evaluated as case-level data, or the larger data set could be evaluated as case-control data. The curator, in conjunction with their GCEP, should determine which is the stronger piece of evidence, and include that in the curation. The family should not be scored twice (once under case-level data, once within the case-control study).

Genetic Evidence Summary Matrix

A matrix used to categorize and quantify the genetic evidence curated for a gene-disease relationship is provided below (Figure 3).

Figure 3: Genetic Evidence Summary Matrix.

	Evidence Type	Case Information Type (Suggested Starting Score)		Suggested Upgrades		Scoring Range	Points Given	Max Score
				Functional Data	*De Novo*			
Case-Level Data	Variant Evidence*	Predicted or proven null variant (1.5 points)		+0.5 points	+0.5 points	0-3 points (per variant)		12 points
		Other variant type (0.1 points)**		+0.4 points	+0.4 points	0-1.5 points (per variant)		
	Segregation Evidence	Evidence of segregation in one or more families		Sequencing Method		0-3 points		3 points
			Total LOD Score	Candidate Gene Sequencing	Exome/Genome or all genes sequenced in linkage region			
			2-2.99	0.5 points	1 point			
			3-4.99	1 point	2 points			
			≥5	1.5 points	3 points			
Case-Control Data	Case-Control Study Type	Case-Control Quality Criteria		Suggested Points/Study			Points Given	Max Score
	Single Variant Analysis	• Variant Detection Methodology • Power • Bias and Confounding Factors • Statistical Significance		0-6 points				12 points
	Aggregate Variant Analysis			0-6 points				
							Total Allowable Points for Genetic Evidence	12 points

**In the case of AR conditions, evaluate each variant (in trans) independently, then combine for the final score.*
***For 2 missense variants without functional data, score at 0.25. For all other scenarios, round to the nearest 0.5 point. See text below for additional explanatory details.*

Scoring Genetic Evidence

Case-Level Data

Assessing case-level data requires knowledge of the disease entity and inheritance pattern for the gene-disease relationship in question, and careful interrogation of the individual genetic variants identified in each case. Within this framework, a case should only be counted towards supporting evidence if:

- The authors provide sufficient evidence to document the diagnosis, to the extent that the GCEP feels comfortable that the proband truly has the diagnosis in question. Clinical information should be collected in the form of Human Phenotype Ontology (HPO) codes and/or free text. HPO terms are strongly preferred. Free text may be used to augment information captured by HPO terms, or in the event that no appropriate HPO terms exist to describe the phenotype. Sufficient detail should be collected to support the diagnosis. For rare and newly reported conditions, it is strongly recommended that as much clinical detail as possible is captured.

- The variant identified in that individual is a plausible cause for disease (e.g. frequency in the general population is consistent with what is known about penetrance/prevalence of the disease, variant consequence is consistent with disease mechanism (if known), etc. Ideally, the variant will have 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) and segregation analysis (see pp. 27-35 for details) if applicable.

Each genetic evidence type has a suggested default starting score per case.

- The default score is intended to provide an initial suggestion for scoring, given that the evidence for each case meets the minimum criteria described above.

- The default scores assume that the variant type is consistent with the expected disease mechanism.

- If this is not the case, downgrade or do not score unless there is compelling rationale to do so and document this rationale in the Gene Curation Interface (GCI).
    - For example, if the disease mechanism is known to be gain-of-function, do not score null variants.

The suggested default starting score can be up- or downgraded as applicable based on the strength of evidence in a given case.

- Some commonly encountered reasons for upgrade (i.e., the variant is de novo and/or the variant has supportive functional information) and suggested point values for each are included in the scoring matrix above.

- Variants may be up- or downgraded beyond the **values** suggested here (but within the scoring range) based on quality of evidence (or lack thereof) demonstrating its role in disease.
    - For example, a single missense variant with supporting functional evidence (score = 0.5, per Figure 3) may score at the top of its range if that functional evidence is robust and demonstrates that the missense is acting in a manner consistent with the expected disease mechanism.

- Further, variants may be up- or downgraded for **other reasons** beyond those listed in the scoring matrix at the discretion of the GCEP.
    - Other potential reasons to upgrade include: consistency and/or specificity of the phenotype, the likelihood a putative null variant actually leads to loss of function, missense variants within the functional domain associated with the disease, missense variants clustering within the same region in a gene, etc. Discuss with your GCEP what constitutes an upgrade within your particular disease area.
    - Other potential reasons to downgrade include: a nonspecific and/or genetically heterogeneous phenotype, insufficient prior testing to rule out other potential causes of disease, a putative null variant unlikely to result in nonsense-mediated decay (e.g., occurring in the last exon), parental relationships have not been confirmed for *de novo* variants, etc. Discuss with your GCEP what constitutes a downgrade within your particular disease area.
    - Always document the rationale for up- or downgrading variants in the GCI.

A range that indicates both the minimum (i.e., 0 points) and maximum score allowed per case is also included.

- A minimum score of “0” is included to remind GCEPs that just because a variant has been observed does not mean it needs to be scored, particularly if it is of dubious quality/relevance.

    - For example, if a variant has been reported in older literature as being "pathogenic" and causative of the proband's phenotype, but that same variant was later found to be observed in high frequencies in controls, the variant can receive a score of "0" instead of the default for that variant type.

- Expert panels may specify the criteria required to meet default and/or maximum scores based on qualities of the gene(s) or disease entity under their purview, as long as the score does NOT go above the stated maximums.

- Expert panels may find it useful to document any specifications they have set for upgrading or downgrading from default for consistency across curations and a resource for new GCEP members.
    - Check with your GCEP coordinator for availability and access of this specification document within your group

- **Please note that the gene curation interface (GCI) allows scoring in the following increments: 0, 0.1, 0.25, 0.5, 1.0, and in half-point increments thereafter.**
    - Because of this, though the total for two missense variants without supporting functional information is technically 0.2 per the scoring matrix above, please score these at 0.25.
    - For all other scenarios, round to the nearest 0.5 point.
        - For example: one missense variant with supporting functional information (0.1 + 0.4 = 0.5) observed *in trans* with one otherwise plausible missense variant without functional information (0.1) is equal to 0.6, but should be rounded down to 0.5 for GCI scoring purposes.

In cases where a heterozygous or hemizygous variant causes disease, score based on the characteristics of the single variant observed.

- Example 1: A single rare missense variant (starting score = 0.1 point) with supportive functional information (+ 0.4 point upgrade) would be scored at 0.5 points.
- Example 2: A single rare missense variant (starting score = 0.1 point) with supportive functional information (+0.4 point upgrade) found to be de novo (additional + 0.4 upgrade) would be scored at 1 point after rounding up to the nearest 0.5 (for GCI scoring).
- Example 3: A single null variant (starting score = 1.5 points) found to be *de novo* (+0.5 point upgrade) would be scored at 2 points.

In cases where biallelic variants (*in trans*) cause disease, evaluate each variant independently, then sum for the final score. Some caveats to the evaluation of biallelic variants:

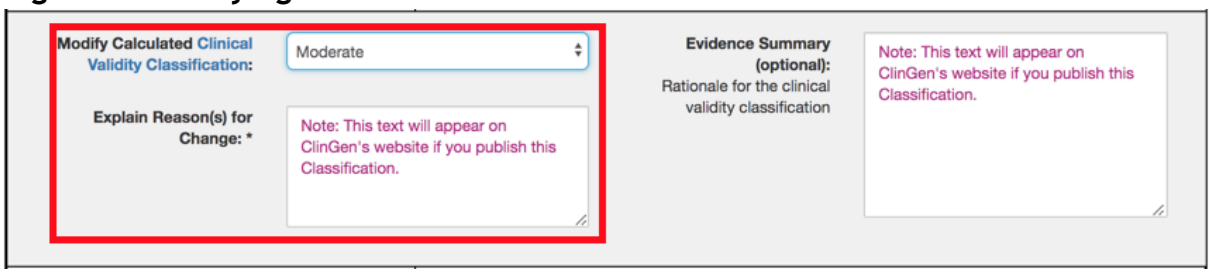
- In general, both variants should be identified (and have some evidence to suggest that they are *in trans*) in the observed case in order to score. In certain scenarios, however, it may be appropriate to score cases where only a single variant has been identified; for example, in the context of diseases in which there is substantial evidence to suggest that biallelic variants cause disease (as opposed to new gene-disease relationships where it may be unclear if the MOI is AR vs. AD), and/or scenarios where there is an alternative method of confirmation that the patient does in fact have the disease in question (e.g., metabolic disorders with diagnostic biochemical profiles). Always discuss with your GCEP whether scoring cases in an AR condition when only one variant has been identified is appropriate.
- For homozygous variants, consider downgrading and/or reducing the maximum number of points such cases could receive. Consider requiring homozygous missense variants to have supporting functional evidence before scoring. The exact parameters surrounding this recommendation should be determined by the GCEP in the context of their specific gene(s)/disease-area.
- Examples of scoring biallelic variants:
    - Example 1: 1 missense variant without supporting functional evidence (0.1) and 1 LOF variant (1.5) *in trans* would equal 1.6, but would be rounded down to 1.5 for GCI scoring purposes.
    - Example 2: 2 de novo missense variants, one with supportive functional evidence (0.1+0.4+0.4 = 0.9) and one without (0.1+0.4 = 0.5) in trans would equal 1.4, but would be rounded up to 1.5 for GCI scoring purposes.

When collecting genetic evidence, the curator is encouraged to document a variety of evidence types to reflect the variant spectrum observed in disease. For example, if a disease is caused by both LOF and missense variants, please include examples of both types in the curation. If a disease is caused exclusively by gain-of-function missense variants, however, there is no need to try to identify other variant types.

To override a calculated classification, the curator should record case information and score it as usual. The classification matrix in the GCI will show the total number of points awarded. However, when calculating the classification, the GCI will automatically cap the points at the stated maximum (10 points for "predicted

or proven null variants" and 7 points for "other variant types"). Therefore, in order to assign the classification approved by the experts, the curator may manually update the classification in the GCI using the dropdown menu on the "classification matrix" tab (Figure 4, red box). If the classification is manually modified e.g. from Moderate to Definitive, rationale for this decision must be given in the free text box under the drop-down menu.

Figure 4. Modifying a Calculated Classification in the GCI



<u>Additional Case-Level Scoring Considerations</u>

De novo variants:

- A variant is considered *de novo* when one of the following scenarios apply:
    - The variant is present in an individual with the disorder but was not found in either parent. In order for a variant to be considered *de novo*, parents must be appropriately tested to show that they do not carry the variant. For individuals with variants in autosomal genes and females heterozygous for an X-linked variant, both parents must be tested. For males who are hemizygous for an X-linked variant, only the mother needs to be tested to investigate *de novo* status.
    - One of the parents of an affected individual is found to have the variant in some cells i.e. is a mosaic. In other words, the variant has arisen "*de novo*" in the parent. The phenotypic features of the parent will depend on the proportion of cells with the variant, and which cell types have the variant.

- When applying an upgrade to the starting default variant score because the variant is found to be *de novo*, consider the following:
    - Is the statistical expectation of *de novo* variation in the gene in question known? In some cases, this can be found in the literature and should be noted (See "literature search" p. 12). Experts in the field should also be consulted. If evidence suggests that de novo variation in this gene is rare, consider upgrading. If the gene is known to have a high rate of de novo variation (e.g., TTN), use caution with scoring or consider not scoring.

- Consider downgrading if parental relationships (i.e., both maternity and paternity of the proband) have not been confirmed. Note that confirmation of parental relationships can be achieved using different methodologies (e.g. short tandem repeat analysis, trio-based exome sequencing).

Predicted or proven null variants:

- This category includes nonsense, frameshift, canonical +/- 1 or 2 splice site variants, single or multi-exon deletions, whole gene deletions, etc. Other variant types, such as missense, may be included in this category if there is sufficient evidence demonstrating complete loss of function.

- Consider downgrading if there is alternative splicing, if the putative null variant is near the C terminus, and/or nonsense mediated decay (NMD) is not predicted (NOTE: NMD is not expected to occur if the stop codon is downstream of the last 50 bp of the penultimate exon).

- Consider downgrading if a gene product is still made, albeit altered. For example, cDNA analysis and/or Western blot from an individual with a canonical splice site change show that an exon is skipped but that the reading frame is maintained, and a protein is produced.

- Individuals with large deletions, duplications, and other chromosomal rearrangements encompassing genetic material outside the gene of interest should not be counted because the impact of the loss/gain for the additional material cannot be assessed.
    - However, if large structural rearrangements represent a significant part of the variant spectrum, it is appropriate to mention these types of variants in the evidence summary.
    - If these types of variants constitute the majority of the variant spectrum (e.g., duplications at 17p12.2 involving the PMP22 gene in Charcot-Marie-Tooth disease), such that the curator is limited in the types of other genetic evidence that may be entered into the GCI, the GCEP may decide to override the calculated classification to account for this type of evidence. In this situation, enter any appropriate single gene variants that can be found, then document the reason for the altered classification, including references to evidence involving large structural variants in the evidence summary. See Figure X for further instruction.

- Intragenic deletions and duplications, depending on their size, may not be able to be formally entered into the GCI as evidence due to technical limitations.
    - ClinGen is actively working on this issue and it will be resolved in the future.
- In the interim, if this type of evidence needs to be used, follow the strategy outlined above for large structural variants: adjust the final classification as necessary, and reference the cases that would have been documented as evidence within the free text evidence summary. See the curation for NRXN1 as an example of this (https://search.clinicalgenome.org/kb/gene-validity/14bb833a-eefd-4f02-b112-c74b6866d5be--2019-07-10T16:00:00).

Other variant with gene impact:

- This category includes, for example, missense variants, and small in-frame insertions and deletions, in addition to variants of any type that result in gain of function or dominant-negative impact.

- As stated above, these types of variants must be at least plausible causes of disease in order to be given the suggested starting default points.

- Some functional impact of the variant to the gene product must be demonstrated for the case to be given upgraded points. Examples of functional impact include reduced activity of an enzyme in cells expressing a variant in the gene of interest, or reduced expression of a gene product when expressed in a heterologous cell system.

- *In silico* predictions do not provide strong evidence for functional impact and are therefore not typically counted as supportive functional data. It may be appropriate to award some upgrade over the default starting points if in-depth *in silico* modeling studies (e.g. impact on 3D structure) have been performed, but this requires discussion with an expert.

Recurrent variants:

Deciding how to score multiple patients with the same variant can be challenging and requires careful consideration. Observations of multiple cases with the same variant(s) can arise from:

- A single patient reported more than once in the literature. The details of each case should be carefully assessed to ensure that the cases are different from each other. If there is any concern that the same case has

been published in multiple papers, the case should be counted only one time.

- Recurrent *de novo* variant. If the variant has occurred *de novo* in multiple patients (with *de novo* status proven by parental testing), score each individual as outlined on page 18.
    - Of note, the same variant arising as *de novo* in multiple individuals with similar phenotypes supports pathogenicity of the variant, as it indicates a hot spot mutation. In these cases, each independent observation may be scored (though we recommend documenting the contribution of other variants if they are available).

- If there is evidence to suggest that a variant has arisen more than once in different populations (e.g. the same variant is present in individuals with different haplotypes), but there is no evidence to indicate that the variant is *de novo* in the patient(s), score each case individually according to the variant type.

- In the event that insufficient or no evidence is available to support that the variant has arisen in different populations and neither case is related, consider downgrading points from the default or not scoring the subsequent cases after the first case, as a conservative measure to reduce overscoring. Consultation with experts within the group is encouraged to guide appropriate scoring given the specific gene and disease of interest.

Founder variants:

- Some genes include known, well-studied pathogenic founder variants, such as *BRCA1* c.68_69delAG, *BRCA1* c.5266dupC, and *BRCA2* c.5946delT, which together account for up to 99% of pathogenic variants identified in individuals of Ashkenazi Jewish ancestry with hereditary breast and ovarian cancer (HBOC), or *GAA* p.Arg854* in African Americans with Pompe disease [4, 5]. If a valid case-control study is available for the variant in question, use this data preferentially and score accordingly. For case-level data, a range of variants in addition to the known founder variant should be curated, if available. This ensures that the classification is not based on one, or a limited number of variants. It may be appropriate to include additional cases with pathogenic founder variants at the discretion of the experts. However, avoid double counting any cases that may have been included in case control studies (see pp.

36-39). Well-known founder variants should be noted either in the curation, or in the curation summary.

- For variants that are reported to be more common in specific populations, which are not well-known pathogenic founder variants, any evidence for the role of the variant in disease must be carefully assessed to avoid over-scoring a variant that is simply common in the population but has little evidence for causing disease. Functional data should be heavily relied upon to ensure that the variant is functionally abnormal and not a benign variant in linkage disequilibrium with the causative genetic change. As above, if a valid case-control study is available for the variant in question, use this data preferentially and score accordingly. After scoring any available case-control studies, curate case-level evidence by including cases with a range of different variants. If all of the genetic evidence has been curated in this manner and the classification has not reached a strong or definitive classification as expected by the expert panel, it may be appropriate to score additional cases with the same variant(s), at the discretion of the GCEP experts. Adjust the case-level scoring as necessary. Alternatively, modification of the calculated clinical validity classification can be made manually within the GCI, providing the inclusion of rationale for the change. Segregation data should be scored as normal (see pp. 27-35). As with all aspects of the gene curation process, the curator should raise any questions with the expert panel.

NOTE: In addition to meeting the above criteria, the variant should not have data that contradicts a pathogenic role, such as an unexplained non-segregation, etc.

General Considerations for Variant Evidence Scoring:

Mode of Inheritance related:

- In X-linked disorders, affected probands will often be hemizygous males and/or manifesting heterozygous females. Curators must be aware of the nuances of interpretation of individual cases and X-linked pedigrees; 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), or males who carry an X-linked variant but are unaffected or mildly affected (due to Klinefelter syndrome, 47, XXY) . Points can be assigned at the discretion of the expert panel and by considering the available evidence. Furthermore, there are known cases of

female carriers of X-linked recessive conditions manifesting symptoms that are milder and/or later in onset compared to males, and scoring of genetic evidence in these examples should be subject to expert review.

Computational and population frequency related:

- 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 (by themselves) 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.

- For a variant to be considered potentially disease-causing, its frequency in the general population should be consistent with phenotype frequency, inheritance pattern, disease penetrance, and disease mechanism (if known). These pieces of information can often be located in the literature (See "Literature Search," p. 12), but may also be contributed by experts. If such information is available, the prevalence of the variant in affected individuals should be enriched compared to controls. The Genome Aggregation Database (gnomAD) provides a reference set of allele frequencies for various populations and can be used to assess whether the frequency of the variant in question is consistent with the prevalence of the disease. GCEPs may find it helpful to set a minor allele frequency (MAF) above which a variant would be considered benign. Generally, MAF thresholds will vary as a function of disease prevalence. This MAF threshold is specific to the disease and should apply to all variants being evaluated, in the context of that disease.

Mechanism and phenotype related:

- **Known disease mechanism:** If the mechanism of disease is known, take this into consideration when scoring individual variants; curators should not feel obligated to award a particular variant a default score (or any score at all) if the variant does not align with the known disease mechanism. For example, if the known mechanism of disease is loss of function (LOF), consider awarding default *de novo* points to putative LOF variants (e.g. nonsense, frameshift, canonical splice site) that are shown to be *de novo* based on parental testing for the variant; consider downgrading *de novo* missense variants that do not have evidence supporting LOF or a deleterious effect to the gene of interest.

Conversely, if the mechanism of disease is known to be gain of function (GOF), consider awarding default points to *de novo* missense variants shown to be causing a gain of function of the gene, downgrading missense variants with unclear function, and awarding 0 points to *de novo* putative LOF variants.

- **Constraint metrics:** Constraint metrics provide an estimate of how tolerant a gene is to particular types of variation, such as loss of function or missense variants. This type of information (and documentation on how these estimates were obtained, how to interpret them, etc.) can currently be found on each gene page on the gnomAD website. In general, if population data suggest that a gene may be tolerant of a particular type of variation, consider this information when deciding how to score that type of variation. Constraint information can be helpful if the disease mechanism is unknown, and the condition is one that is expected to be depleted in population databases (such as severe, early-onset conditions). For example, when evaluating a *de novo* missense variant in the context of an unknown disease mechanism, evidence that missense variants are common in the general population may warrant downgrading from default point values. However, this can be context-specific given that the constraint score in gnomAD looks at the gene level. It may be useful to look at pathogenicity predictors for the variant in the case of missense variants, or discuss with experts. When deciding to use constraint metrics as part of a gene-disease validity curation, keep in mind that constraint scores must be interpreted in the context of the gene-disease relationship in question. For example, if the gene is associated with multiple diseases, LOF constraint could be associated with a disease other than the one being curated. In addition, genes associated with severe, pediatric-onset disorders may appear to be more constrained than adult-onset conditions where overall fitness is not impacted. Furthermore, it is important to consider the gene transcript(s) implicated in the disease of interest. By default, gnomAD returns constraint scores based on the longest transcript in Ensembl; however, this may not be the canonical transcript associated with the disease of interest. Therefore, a curator may need to choose the appropriate transcript within gnomAD to assess the appropriate constraint metrics. Also, constraint metrics are currently restricted to dominant disease, therefore there are no metrics to measure constraint in the context of autosomal recessive inheritance. When in doubt, consult with an expert.

- **Specificity of phenotype and extent of previous testing:** When curating for relatively non-specific and/or genetically heterogeneous conditions (e.g., intellectual disability and/or autism, etc.), consider how confident one can be that alternative genetic causes of disease have been ruled out through previous testing. For example, if a variant was identified in a gene during the course of single gene-sequencing (i.e. candidate sequencing) in an individual with autism and no previous testing, consider downgrading from default points, as other genetic etiologies have not been ruled out; consider awarding default points if the variant was identified on whole exome or whole genome sequencing. If the phenotype is highly specific and/or has limited genetic heterogeneity, a single gene test or a limited multi-gene panel may be sufficient to warrant default points. For example, if an enzyme assay has shown deficiency in an enzyme known to be associated with a single gene (and other genetic etiologies are unlikely), then sequencing of that gene alone may be sufficient to award default points. The GCEP may be consulted to outline preferred previous testing for the group.
    - Alternatively, curators may choose to document (but not score) various pieces of evidence if they do not provide compelling supporting or contradicting refuting evidence; just because a particular type of evidence is available does not mean it is required to receive a default score for a given category. However, the curator should always document reasons for any deviation in suggested scores for expert review. To document in the GCI, a curator must at least mark the evidence as "Review" in order for it to show in the final Evidence Summary.

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 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). We suggest awarding different amounts of points depending on the methods used to investigate the linkage interval. For this reason, it is critical that the curator make a note of testing methodologies in families counted towards the segregation score. See below for a) instructions how to count segregations and calculate a simplified LOD score and b) how to evaluate the sequencing methods for the linkage interval and award points accordingly. Note that these are general guidelines; if you encounter cases where

you are unsure how to evaluate/score segregation, please discuss with your expert group and/or the ClinGen Gene Curation working group.

Counting Segregations and Calculating Simplified LOD Scores

If a LOD score has been calculated by the authors of a paper (i.e. published LOD/pLOD):

This LOD score should be documented and may be used to assign segregation points (according to the sequencing methods used to investigate the linkage region and identify the variants) in the scoring matrix (see Fig 6 for scoring suggestions). If a LOD score is 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. See below for more guidance on scoring. If a LOD score has NOT been calculated by the authors of a paper (i.e. estimated LOD/eLOD):

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 3 affected individuals in the pedigree**, including the proband). Genotypes must be specified for all affected and unaffected individuals counted; specifically, parents of affected individuals must be genotyped, or other methods must be used to show that the variants are in *trans* if the affected individuals are noted to be compound heterozygotes.
- Families included in the calculation must not demonstrate any unexplainable non-segregations (for example, 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.

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. An obligate carrier is an individual who has not been tested for the variant in question but who is inferred to carry the variant by virtue of their position in the pedigree (for example, an individual with a parent with the variant and a child with the variant, an individual with a sibling with the variant and a child with the variant, etc.).

For the purposes of counting segregations, dizygotic (fraternal) twins count as two separate individuals and monozygotic (identical) twins count as one individual. For example, if an affected proband has dizygotic twin siblings, both of whom are affected and have the variant, two segregations can be counted. If an affected proband has affected monozygotic twin siblings with the variant, one segregation can be counted.

Within a given gene-disease curation, if more than one family meets the criteria above for scoring segregation information, the LOD scores are summed to assign a final segregation score (using Figures 5 or 6). For example, if Family A has an estimated LOD score of 1.2 and Family B has an estimated LOD score of 1.8, the summed LOD score will equal 3. See the discussion on sequencing method below for guidance on assigning segregation points to the LOD score.

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.

NOTE: Segregation implicates a locus in a disease, NOT a variant. Therefore, all linkage studies should be carefully assessed to ensure that appropriate measures have been taken to rule out other possible causative genes within the critical region (see guide on point assignment based on methods to investigate a linkage region below).

For dominant/X-linked diseases:

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

Figure 5: Dominant/X-linked LOD score table

Dominant Segregations	15	14	13	12	11	10	9	8	7	6	5	4
Estimated LOD	4.5	4.2	3.9	3.6	3.3	3.0	2.7	2.4	2.1	1.8	1.5	1.2

For recessive diseases:

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

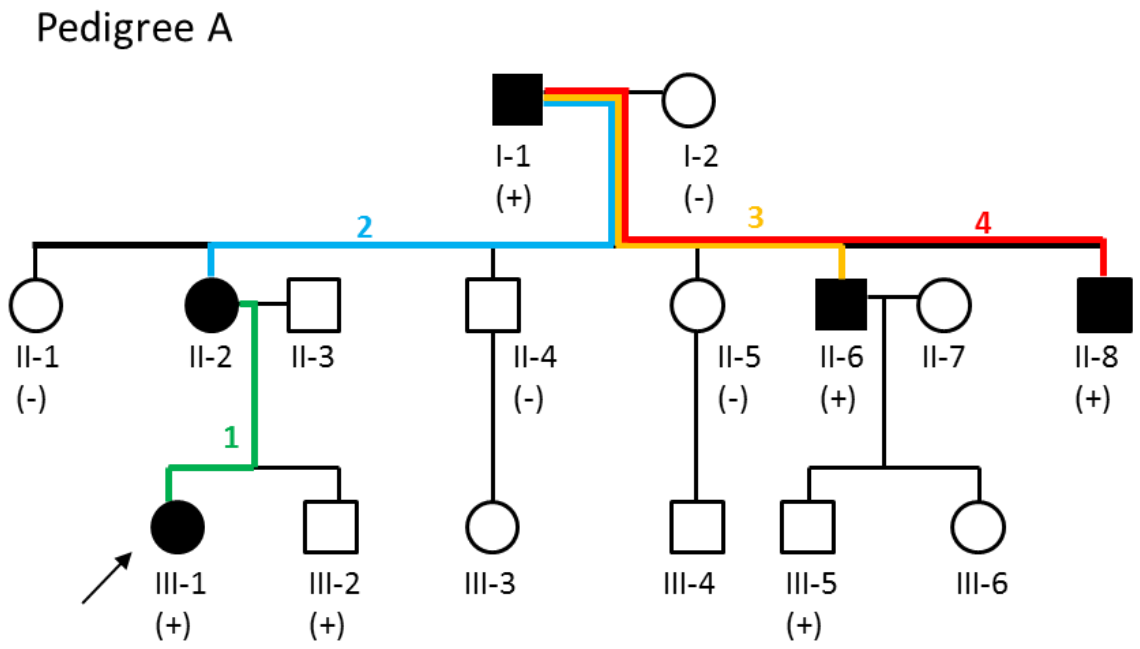
NOTE: In general, the number of affected individuals - 1 is equal to the number of affected segregations from the proband, and can be used interchangeably in this equation. The base numbers, “0.25” and “0.75”, used in this equation represent the risk of being affected vs. unaffected in a classic AR disease model in which both parents are carriers. The eLOD scores provided in Figure 6 refer only to the classic AR disease model. If a pedigree differs from this situation, please adjust the base numbers in the equation above to reflect the risk of inheritance, and use the equation to estimate the LOD score. For example, if one parent is affected with an autosomal recessive condition and the other is a carrier, replace both “0.25” and “0.75” with 0.5.

Figure 6: Recessive estimated LOD (eLOD) score table

		Unaffecteds										
		0	1	2	3	4	5	6	7	8	9	10
Affecteds	3	1.20	1.32	1.45	1.50	1.70	1.82	1.95	2.07	2.20	2.33	2.45
	4	1.81	1.93	2.06	2.18	2.31	2.43	2.56	2.68	2.81	2.93	3.06
	5	2.41	2.53	2.66	2.78	2.91	3.03	3.16	3.28	3.41	3.53	3.66
	6	3.01	3.14	3.26	3.39	3.51	3.63	3.76	3.88	4.01	4.13	4.26
	7	3.61	3.74	3.86	3.99	4.11	4.24	4.36	4.49	4.61	4.74	4.86
	8	4.21	4.34	4.46	4.59	4.71	4.84	4.96	5.09	5.21	5.34	5.46
	9	4.82	4.94	5.07	5.19	5.32	5.44	5.57	5.69	5.82	5.94	6.07
	10	5.42	5.54	5.67	5.79	5.92	6.04	6.17	6.29	6.42	6.54	6.67

Counting Segregations

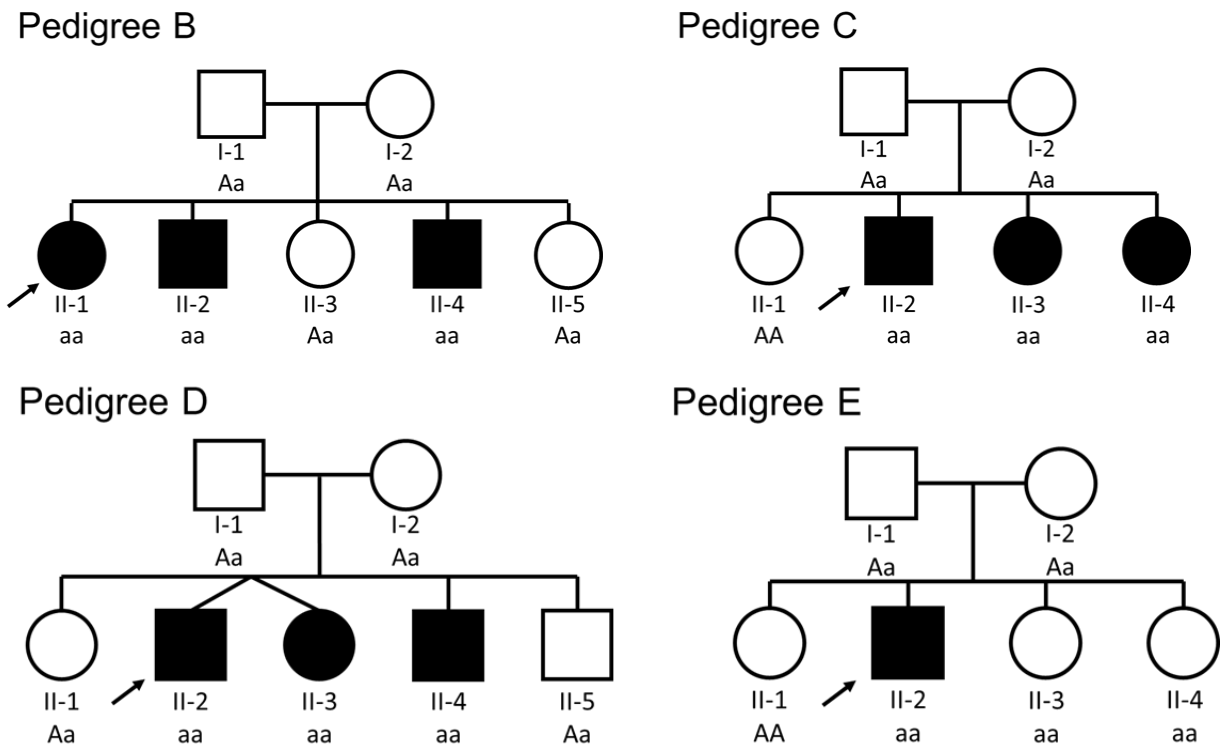
- In general, the number of segregations in the family will be the number of affected individuals minus one, the proband, to account for the proband's genotype phase being unknown. However, as there may be exceptions, segregations should be counted carefully, as outlined below. For example, **pedigree A** shows a family with hypertrophic cardiomyopathy.
    - 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. Using **four segregations** in the formula above results in an estimated e**LOD score of 1.2**.
    - For disorders with reduced penetrance such as cardiomyopathy, it is **safest to only use affected genotype$^+$ (genotype$^+$/phenotype$^+$) individuals for segregation.** Obligate carriers (i.e. any individual who can be definitively inferred to be genotype positive based on the genetic status of other family members, as discussed above) should also be included, regardless of phenotype**.** In this case, the absence of a phenotype in two genotype$^+$ individuals (III-2 and III-5) is considered irrelevant as they can be explained by delayed onset and/or reduced penetrance. However, these individuals are not included in the eLOD calculation because they are unaffected.



- When estimating LOD scores for autosomal recessive disorders, count unaffected individuals as those 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. For example, there are two unaffected individuals in Pedigree B, one unaffected individual in Pedigree C, and two unaffected individuals in Pedigree D.

- For reasonably penetrant Mendelian disorders, a single LOD score can be calculated across multiple families, providing that each family meets the criteria above. For example, in pedigrees B, C and D, each with fully penetrant recessive hearing loss, the LOD scores can be added ((1.45 for B) + (1.32 for C) + (1.45 for D)) to give a total LOD score of 4.22. However, pedigree E cannot be included in this LOD score total because this family does not have enough affected individuals.

- For help with counting segregations, please see the "Interactive Training Modules" section of the Gene-Disease Validity Training page, found here.



Assigning points to LOD scores:
While segregation evidence can be convincing for a particular locus, 10s or even 100s of genes can be within a linkage interval. Thus, segregation does not necessarily implicate a single gene or variant. Many publications do not thoroughly investigate other genes or variants found within the linkage interval and those that do cannot rule out the effects of potentially thousands of other variants in the interval. **Thus, it is critical for a curator to evaluate the methods used to identify candidate variants.**

Some publications more thoroughly investigate the genes and variants in a linkage interval than others. Accordingly, more points are awarded for segregation evidence in cases where whole exome/genome sequencing was performed or if the entire linkage interval was sequenced. These methods provide more convincing evidence than a candidate gene approach in which only one or a handful of genes in a linkage region are sequenced. See Figure 7 below for suggested point ranges for LOD scores.

NOTE: For this scoring matrix, LOD scores from all families meeting size requirements **must be summed before awarding segregation points, regardless of the sequencing methodology used**. Sequencing methodology (e.g., candidate gene sequencing, whole exome sequencing, etc.) should be accounted for when deciding on the most appropriate score for this evidence. See example 2 below for an

example of scoring multiple families with variants ascertained via different methodologies. Note that simply having a single family meeting the minimum size requirements is not necessarily enough to warrant any points. As the methods in each publication vary, the suggested points in Figure 7 are merely a guide for the curator.

Figure 7: Proposed Matrix Scoring for different LOD score ranges

Total summed LOD score across all families	Sequencing method	
	Candidate gene sequencing	Exome/genome or all genes sequenced in linkage region
0-1.99	0 pts	0 pts
2-2.99	0.5 pts	1 pt
3 - 4.99	1 pt	2 pts
(>/=) 5	1.5 pts	3 pts

A formula has been developed to help curators determine the number of points to assign when there are multiple pieces of segregation evidence.

$$\text{Segregation points} = \left(\left(\frac{A}{A+B}\right) * C\right) + \left(\left(\frac{B}{A+B}\right) * D\right)$$

Where:
A = The sum of all LOD scores for candidate gene approach.
B = The sum of all LOD scores for exome sequencing, genome sequencing, and all genes in candidate region sequenced.
C = Points assigned if total LOD had been obtained only by a candidate gene approach (see Figure 7).
D = Points assigned if total LOD had been obtained only by exome/genome sequencing/all genes in candidate region sequenced approach (see Figure 7).

Note: For C and D, these points are derived from the candidate and exome/genome points assigned within the range of the total summed LOD score (A+B).

A calculator using this formula is available (link to calculator). The points are rounded to the nearest 0.1 point. This calculator has been incorporated into the ClinGen Gene Curation Interface (GCI) so that the number of segregation points is automatically calculated, as illustrated in the examples below.

Example Scenarios:

Example 1: Linkage analysis was performed on one large family with autosomal dominant hypertrophic cardiomyopathy (HCM). There are **11 affected individuals** in the pedigree (phenotype$^+$/genotype$^+$), and using our simplified LOD score formula, this corresponds to a **LOD score of 3** (see Figure 5). The linkage region for this family contained 15 genes and the authors **sequenced all of the genes in the linkage interval** and the HCM variant was the only suspicious variant. Looking at Figure 7, you can assign this LOD score **2 points**.

Example 2: Let's return to Pedigrees B, C, and D above, assuming now that we know more about how the linkage intervals were investigated or how the variants were identified.

Pedigree B: LOD Score 1.5, Variants identified using whole exome sequencing
Pedigree C: LOD Score 1.3, Variants identified using whole exome sequencing
Pedigree D: LOD Score 1.5, Variants identified using candidate gene analysis. Only the gene of interest was sequenced.

Using the formula above, 1.7 points would be assigned:

$$\left(\left(\frac{1.5}{4.3}\right) * 1\right) + \left(\left(\frac{2.8}{4.3}\right) * 2\right) = 1.7$$

Additional Logic

While the formula is appropriate for use in the majority of scenarios, there are some situations for which additional logic must be used. This logic, which is coded into the GCI and the calculator, is illustrated by the following example. For Family 1, an estimated LOD score of 3.1 is obtained from a study involving WES. For Family 2, a candidate gene analysis was performed, and a LOD of 1.2 was estimated. In this scenario, 2 points could be awarded for Family 1 alone (as the LOD is between 3-4.99; see Figure 7). The total LOD score for Family 1 and Family 2 is 4.3. If the second piece of evidence were to be included, the points would be reduced to 1.8. In this situation, the formula should not be applied and the maximum number of points (i.e. 2) should be given.

We recognize that the methods in each publication vary. Therefore, the suggested points in Figure 7 are merely a guide for the curator. If curators are unsure of segregation scoring based on genotyping method, please consult experts.

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**.
    - **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.

    - **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 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.

3. The quality of each case-control study should be evaluated using the following criteria in aggregate:

- **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).

- **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.)

- **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:
    - 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)?
    - 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?
    - Have the cases and controls been equivalently evaluated for presence or absence of a phenotype, and/or family history of disease?

- **Statistical Significance**: The level of statistical significance should be weighed carefully.
    - When an odds ratio (OR) is presented, its magnitude should be consistent with a monogenic disease etiology.
    - 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.
    - Factors, such as multiple testing, that might impact that interpretation of uncorrected p-values and CIs should be considered when assigning points.

Figure 8: Case-control Genetic Evidence Examples
Detailed examples and explanations for assigned points are provided in the table below.

Figure 8. 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. ExAC)	Cases: sequenced Gene *Y* and counted all cases with null variants. Controls: total individuals from population database with null variants in gene *Y*.	OR of all variants in aggregate: 4.9 (CI: 1.4-17.7; *P* =0.015)	Aggregate analysis	2
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 11 variants in 11 cases. Controls: total individuals from a population database that were genotyped for the 11 variants identified in controls.	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 specific variants 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.

NOTE: The maximum score for the Case-control category is 12 points, which is the maximum allowable points for the entire Genetic Evidence category.

EXPERIMENTAL EVIDENCE

There are several forms of experimental and functional assays to elucidate gene function. For clinical validity classifications, only evidence that supports the role of a gene in a disease, or phenotypic features associated with the disease entity of interest count as applicable evidence for scoring. Validated functional assays should be identified by expert panels or, if they are curator identified, confirmed by expert review.

Figure 9: 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	L	W 2
	Protein Interaction	B 0.5	0-2	M	
	Expression	C 0.5	0-2	N	
Functional Alteration	Patient cells	D 1	0-2	O	X 2
	Non-patient cells	E 0.5	0-1	P	
Models	Non-human model organism	F 2	0-4	Q	Y 4
	Cell culture model	G 1	0-2	R	
Rescue	Rescue in human	H 2	0-4	S	
	Rescue in non-human model organism	I 2	0-4	T	
	Rescue in cell culture model	J 1	0-2	U	
	Rescue in patient cells	K 1	0-2	V	
Total Allowable Points for Experimental Evidence					Z 6

Identify the experimental evidence type and assign points according to the following criteria. For further information and examples see the "Variant evidence vs experimental evidence" section in **Appendix B**.

1. Biochemical Function: Evidence showing the gene product performs a **biochemical function:** (A) shared with other known genes in the disease of interest, or (B) consistent with the phenotype. NOTE: The biochemical function of both gene products must have been proven experimentally, and not just predicted. When awarding points in this evidence category, the other known gene(s) should have compelling evidence to support the gene-disease association. Consider increasing points based on the strength of the evidence and number of other proteins with the same function that are involved in the same disease.

2. Protein Interaction: Evidence showing the gene product **interacts** with **proteins previously implicated** in the disease of interest. Typical examples of this data include, but are not limited to: Physical interaction via Yeast-2-Hybrid (Y2H), co-immunoprecipitation (coIP), etc.

   NOTE: The interaction of the gene products must have been proven experimentally, and not just predicted. Proteins previously implicated in the disease of interest should have compelling evidence to support the gene-disease association. Note: Some studies provide evidence that a variant in the gene of interest disrupts the interaction of the gene product with another protein. In these cases, the positive control, showing interaction between the two wild type proteins, can be counted as evidence of protein interaction. Points can also be awarded to case-level (variant) evidence or functional alteration for the variant disrupting the interaction.

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). An example scenario to consider for altered expression in patients includes studies in which expression of the gene of interest (and even additional genes) is examined in tissue and/or cell samples obtained from individuals with the disease of interest in which the molecular etiology of the individual is unknown. For instance, tissues samples from 10 individuals diagnosed with hypertrophic cardiomyopathy were examined by western blot analysis and found that gene X was reduced in the heart cells of all patients. Expert reviewers may specify appropriate uses of this category in the context of their particular disease domain. For example, groups may choose to award points based on the specificity of expression in relevant organs.

   NOTE: The sum of all biochemical function, protein interaction, and expression points may not exceed the max score of 2 points.

4. Functional Alteration: Evidence showing that cultured cells, in which the function of the gene has been disrupted, have a phenotype that is consistent with the human disease process. Examples include experiments involving expression of a genetic variant, gene knock-down, overexpression, etc. Divide the evidence according to the following subtypes:

a. Was the experiment conducted in **patient cells**?
b. Was the experiment conducted in **non-patient cells**?

NOTE: The sum of all functional alteration points may not exceed the max score of 2 points

5. Model System: A **non-human model organism** or **cell culture model** with a disrupted copy of the gene shows a phenotype consistent with the human disease state. Note: Cell culture models should recapitulate the features of the diseased tissue e.g. engineered heart tissue, or cultured brain slices. These results should be summarized accordingly:
   a. Was the gene disruption in a **non-human model organism**? NOTE: If a gene-disease pair does not have genetic evidence (i.e. classified as No Known Disease Relationship), but a non-human model organism is scored, an "Animal Model Only" tag will appear on this curation when it is published to the ClinGen website.
   b. Was the gene disrupted in a **cell culture model**?

6. Rescue: Summarize evidence showing that the **phenotype in humans** (i.e. patients with the condition), **non-human model organisms, cell culture models,** or **patient cells** can be rescued. If the phenotype is caused by loss of function, summarize evidence showing that the phenotype can be rescued by exogenous wild-type gene, gene product, or targeted gene editing. If the phenotype is caused by a gain of function variant, summarize the evidence showing that a treatment which specifically blocks the action of the variant (e.g. siRNA, antibody, targeted gene editing) rescues the phenotype. These results should be recorded accordingly:
   a. Was the rescue in a **human?** For example, successful enzyme replacement therapy for a lysosomal storage disease.
   b. Was the rescue in a **non-human model organism**?
   c. While the default points and point range are the same for human and non-human model organism, consider awarding more points if the rescue was in a human. Was the rescue in a **cell culture model** (*i.e.* a cell culture model engineered to express the variant of interest)? Was the rescue in **patient cells**?

      NOTE: The sum of all models and rescue may not exceed the max of 4 points.

Experimental Evidence Summary Score: The total experimental evidence points may not exceed the max score of 6, regardless of the individual evidence category or evidence type score tally. It is best practice to prioritize curating genetic evidence over experimental evidence to reach a definitive score, however for cases in which the gene-disease relationship is well-known or has substantial experimental evidence, a curator is encouraged to attempt to curate experimental evidence from each evidence category (i.e. Functional, Functional Alteration, Models and Rescue), where applicable.

For specific examples of different pieces of experimental evidence, please see Appendix B.

Case-level Variant Evidence vs. Experimental Evidence

Distinguishing between functional evidence that supports an individual variant and experimental evidence that supports the gene-disease relationship:

Not all functional evidence supports the role of the gene in the disease. Therefore, the curator must carefully consider whether to count functional evidence in the experimental evidence section or in the case-level data section. Only evidence that supports the role of the gene in the disease should be counted in the experimental evidence section. Experimental evidence that does not directly support the role of the gene in the disease or recapitulation of disease phenotypes, but indicates that the variant is damaging to the gene function can, instead, be used to increase points in the case-level data section. Some very general examples are given below. Please note that these examples are a guide. Each piece of evidence should be carefully considered when deciding on which category to assign points. Furthermore, the piece of evidence should only be counted once, to prevent overscoring of a single piece of evidence. Ultimately, these decisions should be discussed with experts in the disease area.

Case-level variant evidence, general examples:

- Immunolocalization showing that the gene product is mislocalized in cells from a patient or in cultured cells. This would be counted as case-level variant evidence UNLESS mislocalization/accumulation of an altered gene product is a known mechanism of disease, in which case this evidence could be counted as experimental evidence (functional alteration).
- Mini-gene splicing assay or RT-PCR showing that splicing is impacted by a splice-site variant.
- A variant in a gene encoding an enzyme is expressed in cultured cells and enzyme activity is deficient.

- A variant is shown to disrupt the normal interaction of the gene product of interest (protein A) with another protein (protein B). NOTE: If protein B is 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 variant evidence.
- Tissue or cells, from an individual with a variant in the gene of interest, showing altered expression of that gene (e.g. reduced expression shown by Western blot).

Experimental evidence, general examples:

- A signaling pathway is 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 is shown repeatedly in multiple patients with the disease regardless of the causative variant, e.g. altered expression in a group of patients with multiple different variants, or in a group of patients with the disease but for whom the genotype has not been determined. For an example, see Appendix B.
- The variant is shown to be associated with a known hallmark of the disease e.g. abnormal deposition or mislocalization of a gene product, abnormal contractility of cells, etc., either in patient cells or cultured cells expressing the variant.
- Any model organism with a variant initially identified in a human with the disorder.

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 and present the contradictory evidence to experts, while the classification (Disputed/Refuted) is to be determined by the clinical domain experts. Below are a few examples 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 be flagged as “Review” or “Contradictory” in the GC, or otherwise 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.32, Fig. 8).**

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 of the proposed pathogenic variants in a gene are present in a specific population or the general population (ExAC, gnomAD, 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 gnomAD of over 10%, this could be potentially contradictory evidence. **NOTE:** Evidence contradicting a single **variant** as causative for the disease does not necessarily rule out the entire 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 must 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 the gene-disease relationship. In any case, if a curator suspects that any evidence contradicts a gene-disease relationship, 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 apparent 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. 36)**. 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.

NOTE: Contradictory evidence may be present in pre-publication articles, such as BioRxiv. In these cases, consult with the expert panel on the validity and use in the clinical validity classification. If used, note the evidence in the Evidence Summary.

SUMMARY & 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. For ClinGen GCEPs using the GCI, the GCI will automatically tally points, assign a classification within the points range, and generate a PDF summary of the evidence, including the PMIDs and evidence captured. **It is required that expert groups summarize the gene curation evidence used in the "Evidence Summary" box in the GCI, which will be displayed on the website when the final clinical validity classification is published.** The Gene Curation Working Group has provided a document with suggested standardized example text, found here, that can be used to guide gene curation summaries.

If multiple expert groups have contributed to a classification, acknowledgement of the contribution should be made using the Secondary contributors function on the approving page of the GCI. This will allow recognition of the collaboration on the final published summary on the clinicalgenome.org website.

1. The total score within the Genetic Evidence Matrix (Figure 3 "U") is listed in Figure 10 column "A".

2. The total score within the Experimental Evidence Matrix (Figure 9 “Z”) is listed in Figure 10 column "B".
3. Figure 10 column "C" represents the total points for the gene-disease-MOI curation record.
4. Refer to the publication date of the original publication of the gene-disease relationship and consider all other literature when assessing replication over time (Figure 10 column "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. Valid contradictory evidence (see pp. 44-46) is highlighted in the final matrix Figure 10 row "E". Rationale should be provided within the designated sections within the GCI.

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, and a final classification reached.

Figure 10: Clinical Validity Summary Matrix

GENE/DISEASE PAIR:				
Assertion criteria	**Genetic Evidence (0-12 points)**	**Experimental Evidence (0-6 points)**	**Total Points (0-18)**	**Replication Over Time (Y/N)**
Description	Case-level, family segregation, or case-control data that support the gene-disease association	Gene-level experimental evidence that support the gene-disease association	Sum of Genetic & Experimental Evidence	> 2 pubs w/ convincing evidence over time (>3 yrs.)
Assigned Points	A	B	C	D
CALCULATED CLASSIFICATION		**LIMITED**	**0.1-6**	
		MODERATE	**7-11**	
		STRONG	**12-18**	
		DEFINITIVE	**12-18 & Replicated Over Time**	
Valid contradictory evidence (Y/N)*	**List PMIDs and describe evidence:** E			
CURATOR CLASSIFICATION	F			
FINAL CLASSIFICATION	G			

Figure 10 footnotes:

- "Strong" is typically used to describe gene-disease pairs with at least 12 points but no replication over time. However, if the experts feel that there is a compelling reason to classify a gene-disease relationship as "Strong," that is otherwise between "Moderate" and "Definitive," then they should do so, provided that the rationale for this decision is documented in the GCI.
- While the total points guide the provisional classification, they do not determine the final approved classification. Instead, the experts consider the overall evidence, with the points as a guide, to finalize a gene-disease classification. It is within the expert and/or group's purview to upgrade or downgrade a classification; however, documentation of their rationale is required.

RECURATION PROCEDURE

ClinGen has developed recommendations for re-evaluating previously approved gene-disease validity classifications. Requirements for the recommended interval for recuration are listed in Table 2. For more detailed information, refer to the recuration document here.

Table 2: Standard Gene-Disease Clinical Validity Recuration Procedure	
Classification	**Interval for re-evaluation**
Definitive	**No set requirement**
Strong	**3 years from the original discovery publication date**
Moderate	**2 years after the last approval date**
Limited	**3 years after the last approval date**
No Known Disease Relationship	**No set requirement**
Disputed	**3 years after the last approval date**
Refuted	**No set requirement**

SOP REFERENCES

1. Strande, N.T., et al., *Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource*. Am J Hum Genet. 100(6): p. 895-906.
2. MacArthur, D.G., et al., *Guidelines for investigating causality of sequence variants in human disease*. Nature. 508(7497): p. 469-76.
3. Ganten, D. et al. (Ed.), *Semidominant Allele*. Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine (2006 ed.): p.171. https://doi.org/10.1007/3-540-29623-9
4. Petrucelli, N., et al., *BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer*. GeneReviews. 1998.
5. Becker, J.A., et al., *The African origin of the common mutation in African American patients with glycogen-storage disease type II*. Am J Hum Genet. 62(4): p. 991-4.

APPENDIX A: USEFUL WEBSITES FOR CLINGEN GENE CURATORS

The following websites are free and publicly available. While this list is not exhaustive, it includes websites that are often used during the ClinGen gene curation process. A brief description for each website is given below; please go to the websites for more information. In addition, for sites which have an associated publication, we have included the PMID. This PMID can be used as a general ID to curate evidence from these sites. It is strongly encouraged that you specify the use of the site in the curation evidence, including any titles, tags, or other identifiers mentioned.

If there are additional websites that you think curators should be aware of, please contact clingen@clinicalgenome.org.

LITERATURE SEARCHES

- **PubMed**
  - https://www.ncbi.nlm.nih.gov/pubmed

REVIEWS/DISEASE ENTITIES

- **Online Mendelian Inheritance in Man (OMIM)**
  - http://www.ncbi.nlm.nih.gov/omim
  - A comprehensive compendium of human genes and phenotypes that is updated regularly. Summaries of gene-disease associations and references to primary literature can be found here.
- **GeneReviews**
  - http://www.ncbi.nlm.nih.gov/books/NBK1116/
  - Provides clinically relevant information for hundreds of different inherited conditions. The “Molecular Genetics” section of each entry may be useful for information on common variants for a gene. The “Establishing the Diagnosis” section typically contains a summary of the genetic testing options, including the different genes involved and proportion of cases caused by variants in each gene.

    - **Many GeneReviews have an associated PMID, however at this time (July 2019) they do NOT work in the GCI.**
- **Monarch Disease Ontology (MonDO)**
    - https://www.ebi.ac.uk/ols/ontologies/mondo
    - Human disease ontology merging information from multiple disease resources.
- **ORPHANET**
    - http://www.orpha.net
    - Online inventory of human diseases.

PHENOTYPES

- **Human Phenotype Ontology (HPO) Browser**
    - http://www.human-phenotype-ontology.github.io/
    - Standardized vocabulary and codes for human phenotypic abnormalities.
- **Monarch Initiative**
    - https://monarchinitiative.org/phenotype
    - Search for a disease then choose the "phenotypes" tab for a list of associated clinical features which links to the corresponding HPO code.

GENES AND GENE PRODUCTS

- **HUGO Gene Nomenclature Committee (HGNC)**
    - http://www.genenames.org
    - An online repository of approved gene nomenclature.
- **National Center for Biotechnology Information (NCBI) gene**
    - http://www.ncbi.nlm.nih.gov/gene
    - Integrates information from a wide range of species. Includes gene nomenclature, reference sequences, maps, expression, protein interactions, pathways, variations, phenotypes, functional evidence (in GeneRIFs) links to locus-specific resources.
    - Each subcategory may list an associated PMID. For example, under the "Expression" header, each sequencing choice in the drop down has an associated PMID. Choose the correct PMID that goes with the sequencing method cited for expression in the GCI.
- **GeneCards**
    - **https://www.genecards.org/**
    - Integrate information from several sources, and includes a publication section.
- **Ensembl**
    - http://www.ensembl.org/index.html
    - Nomenclature, splice variants, references sequences, maps, variants, expression, comparative genomics, ontologies, and function.
- **UCSC Genome Browser**
    - https://genome.ucsc.edu/Genome browser with access to genome sequence data from a range of species.
- **UniProt**
    - www.uniprot/org
      Comprehensive resource for protein sequence and functional information.

VARIANT DATABASES

- **ClinVar**

    - http://www.ncbi.nlm.nih.gov/clinvar/
    - Public archive of human gene variants and phenotypes submitted by clinical and research laboratories, genetics clinics, locus specific databases, expert groups, and OMIM.
- **Leiden Open Variation Database (LOVD)**
    - http://www.lovd.nl/3.0/home
    - Listings of variants within human genes and associated phenotypes; includes links to locus-specific databases.

ALLELE FREQUENCIES

- **Exome Aggregation Consortium (ExAC)**
    - http://www.exac.broadinstitute.org
    - Database with aggregated and harmonized data from over 60,000 human exomes from unrelated individuals. Provides allele frequencies in different major racial and ethnic groups.
- **Genome Aggregation Database (gnomAD)**
    - http://gnomad.broadinstitute.org/
    - Database with aggregated and harmonized data from over 123,000 human exomes and 15,000 human genomes from unrelated individuals. Provides allele frequencies in different major racial and ethnic groups.

GENE EXPRESSION

- See data on individual gene pages on **NCBI Gene** and **Ensembl**
    - https://www.ncbi.nlm.nih.gov/gene
    - http://www.ensembl.org/index.html
- **The Human Protein Atlas**
    - http://www.proteinatlas.org/
    - Seminal paper PMID: 18853439

- **Genotype-Tissue Expression (GTEx) project**
    - https://gtexportal.org/home/
    - Seminal paper PMID: 23715323
- **BioGPS**
    - http://biogps.org/#goto=welcome
    - Seminal paper PMID: 19919682

PROTEIN INTERACTION

- See data on individual gene pages on **NCBI Gene** and **Ensembl**
    - https://www.ncbi.nlm.nih.gov/gene
- **Biological General Repository for Interaction Datasets (BioGRID)**
    - https://thebiogrid.org/
    - Compilation of genetic and protein interaction data from model organisms and humans.
    - Latest publication update PMID: 30476227
- **Agile Protein Interactomes DataServer (APID)**
    - http://cicblade.dep.usal.es:8080/APID/init.action#tabr2
    - Comprehensive collection of protein interactions from over 400 organisms.
    - Reference article PMID: 30715274
- **STRING database**
    - http://string-db.org/

    - Database of known and predicted protein interactions.
    - Associated PMID: 27924014

MOUSE MODELS

- **Mouse Genome Informatics**
    - https://www.jax.org/jax-mice-and-services
    - Database of laboratory mice, providing integrated genetic, genomic, and biological data.
    - Each mouse model will contain a list of "references" that can be used. In addition, a curator may choose to include the URL for the MGI page for the mouse references or mouse model.
- **Knockout Mouse Project (KOMP)**
    - https://www.komp.org/
    - Initiative to generate a public resource of **mouse** embryonic stem cells containing a null mutation in every gene in the **mouse** genome.

- **International Mouse Phenotyping Consortium (IMPC)**
    - https://www.mousephenotype.org/
    - Initiative that is phenotyping numerous mouse model lines.
    - Latest database update article, PMID: 31127358

CASE-LEVEL DATABASES

The following lists public resources containing case report genetic evidence. **Note:** Take caution when using case-level information from these databases and ensure that the individual has not been reported in another publication. Some sites may reference if cases have been published in the literature, however many may not.

- **DECIPHER**
    - https://decipher.sanger.ac.uk/
    - Database that houses over 30,000+ case reports.
    - Seminal paper PMID: 19344873
- **Genome Connect**
    - https://www.ncbi.nlm.nih.gov/clinvar/submitters/506185/
    - ClinGen patient registry. Case-level data is published to ClinVar and includes phenotyping and variants.
    - Seminal paper PMID: 26178529
- **denovo-db**
    - http://denovo-db.gs.washington.edu/denovo-db/
    - Database of *de novo* variation found in the genome.
    - Seminal paper PMID: 27907889
- **MyGene2**
    - https://mygene2.org/MyGene2/
    - Database of case reports.
    - PMID: 27191528

APPENDIX B: EXPERIMENTAL EVIDENCE EXAMPLES

FUNCTION

Biochemical function:

- ***Example: MYH7 and hypertrophic cardiomyopathy (HCM)***
  Variants in *MYH7* have been identified in patients with HCM. *MYH7* encodes the beta-myosin heavy chain, the major protein comprising the thick filament of the cardiac sarcomere. Genes encoding other thick filament cardiac sarcomeric proteins, including *MYBPC3, MYL2, MYL3*, have been definitively associated with HCM. Therefore, the function of *MYH7* is shared with other known genes in the disease of interest. (Default: 0.5 points)
- ***Example: Biallelic mutations in DRAM2 cause retinal dystrophy.***
  Variants in *DRAM2* have been reported by El-Asrag et al. in patients with retinal dystrophy [1]. The authors recap previous experimental evidence suggesting that *DRAM2* is involved in autophagy and discuss the importance of autophagy in normal photoreceptor function. Localization of *DRAM2* in the inner segment of the photoreceptor layer and the apical surface of the retinal pigment epithelium is consistent with a role in photoreceptor autophagy. Therefore, the predicted function of *DRAM2* is consistent with the disease process. (Default: 0.5 points)
- ***Example: GAA and Pompe disease***
  Pompe disease (glycogen storage disease type II) is characterized by accumulation of glycogen in lysosomes. *GAA* encodes acid alpha-glucosidase, a lysosomal enzyme which breaks down glycogen. The function of acid alpha-glucosidase is therefore consistent with the disease process. (Default: 0.5 points)

Protein interaction:

- ***Example: KCNJ8 and Cantu syndrome***
  The products of the *KCNJ8* and *ABCC9* genes interact to form ATP-sensitive potassium channels. Gain of function variants in *ABCC9* were reported in about 30 individuals with Cantu syndrome. Subsequently, gain of function variants in *KCNJ8* were also reported in individuals with Cantu syndrome [2, 3]. Protein interaction points can be awarded to *KCNJ8* due to interaction of the gene product with a protein implicated in the disease (encoded by *ABCC9*). (Default: 0.5 points)

Expression:

- ***Example: TMEM132E and autosomal recessive sensorineural hearing loss***
  Using qPCR, *TMEM132E* has been demonstrated to be highly expressed in the cochlea and the brain, two tissues that can be affected by hearing loss [4]. Western blotting confirmed that the protein is expressed in these tissues. (Default: 0.5 points)
- ***Example: PDE10A and childhood onset chorea with bilateral striatal lesions***
  Variants in *PDE10A* have been reported in individuals with childhood onset chorea [5]. Microarray data from post-mortem brain tissue showed exceptionally high expression in the putamen, consistent with data in the Allen Mouse Brain Atlas and

previous publications showing high and selective *PDE10A* expression in human striatum at both the RNA and protein levels [6, 7]. While PDE10A is transcribed in many tissues, the highest expression is in brain (https://gtexportal.org/home/gene/PDE10A). Points can be awarded because *PDE10A* expression is relevant to the disease of interest. (Default: 0.5 points)

- **Example: Leptin and Severe early-onset obesity**
  Leptin is a hormone secreted by adipose tissue that signals satiety, examined in two severely obese children from a consanguineous Pakistani family [8]. Circulating leptin levels were measured by ELISA and were found to be very low compared with controls and unaffected family members. (Default: 0.5 points)

FUNCTIONAL ALTERATION

- ***Example: Functional alteration, patient cells***
  FBN1 variants in Marfan Syndrome
  Granata et al. studied smooth muscle cells derived from isolated pluripotent stem cells from patients with Marfan syndrome and variants in *FBN1* (p.Cys1242Tyr and p.Gly880Ser) [9]. FBN1 deposition into the extracellular matrix (ECM) and contractility of the differentiated smooth muscle cells in response to carbachol stimulation were measured. Results indicated that the ECM is destabilized for cells with the variant. Destabilization of the ECM in muscle cells is a hallmark of aortic aneurysm. Because aortic aneurysm is a phenotypic feature of Marfan syndrome, changes to ECM organization support the disease mechanism. This evidence can be counted as functional alteration. (Default: 1 point)
- ***Example: Functional alteration, non-patient cells***
  FHL1 and Emery-Dreifuss Muscular Dystrophy (EDMD)
  Some patients with EDMD develop hypertrophic cardiomyopathy. Freidrich et al. transduced neonatal murine cardiomyocytes with AAV constructs with *FHL1* p.Lys45Serfs and p.Cys276Ser variants [10]. Variant *FHL1* proteins were mislocalized and did not incorporate into the sarcomere. Localization and incorporation into the sarcomere for *MYBPC3*, a known causative gene for HCM, was also perturbed. Because *MYBPC3* is known to be involved in HCM, and sarcomere disruption is a hallmark of HCM, the changes in its expression and localization of mutant *FHL1* in cultured non-patient cells is experimental evidence to support the disease mechanism. (Default: 0.5 points)

MODELS AND RESCUE

- ***Example: Animal model***
  TMEM132E and autosomal recessive sensorineural hearing loss
  Li et al. knocked down *TMEM132E* in zebrafish using antisense morpholino oligos [4]. The morpholino animals displayed delayed startle response and reduced extracellular microphonic potentials, suggesting hearing loss. (Default: 2 points)
- ***Example: Cell culture model***
  FHL1 and Emery-Dreifuss Muscular Dystrophy (EDMD)

Some patients with EDMD develop hypertrophic cardiomyopathy. Freidrich et al. measured contraction in AAV transduced **rat engineered heart tissue** (rEHT) expressing *FHL1* variants [10]. rEHT tissue expressing the mutant *FHL1* constructs had significantly altered contraction parameters. Hypercontractility and diastolic dysfunction are hallmarks of HCM, therefore changes to these parameters due to mutant *FHL1* expression support the disease mechanism. (Default: 1 point)

- ***Example: Rescue in human***
  Leptin and Severe early-onset obesity
  The *LEP* gene encodes leptin, a satiety hormone that is secreted by adipose tissue. Montague et al. reported that two severely obese children from a consanguineous Pakistani family had frameshift variants in *LEP* [8]. When one of these children was treated with recombinant Leptin for 12 months, hyperphagia ceased and the amount of body fat lost was 15.6kg (accounting for 95% of the weight lost) [11]. (Default: 2 points)
- ***Example: Rescue in an animal model***
  TMEM132E and autosomal recessive sensorineural hearing loss
  Li et al. injected human *TMEM132E* mRNA into antisense oligo knockdown zebrafish [4]. This partially rescued the hearing defects in those fish. (1 point was given instead of the default 2 because the mRNA only partially rescues the phenotype).
- ***Example: Rescue in patient cells***
  COL3A1 and Ehlers-Danlos, vascular type
  EDS Type IV is caused by dominant-negative mutations in the procollagen type III gene, *COL3A1*. Müller et al. studied cultured fibroblasts from a patient with EDS type IV who was heterozygous for p.Gly252Val in *COL3A1* and from a healthy control [12]. The authors identified a single siRNA that was able to knockdown the mutant *COL3A1* mRNA (>90%) in the patient-derived fibroblasts without affecting wild type *COL3A1*. Prior to treatment with siRNA, the mutant cells showed disorganized bundles of collagen fibers. After treatment with siRNA, the morphology of the extracellular matrix more closely resembled healthy control fibroblasts. (Default: 1 point)
- ***Example: Rescue in humans***
  Pompe disease is caused by deficient activity of acid-alpha glucosidase (*GAA*). Patients with the infantile onset form typically die by one year of age if untreated. Kishnani et al. reported clinical improvements in 8 patients with infantile-onset Pompe disease who received a weekly intravenous infusion of recombinant *GAA* for 52 weeks [13]. Clinical improvements included amelioration in cardiomyopathy, improved growth, and acquisition of new motor skills in 5 patients, including independent walking in three of them. Although four patients died after the initial study phase, the median age at death was significantly later than expected for patients who were not treated. Treatment was safe and well tolerated. (4 points)

APPENDIX B REFERENCES:

1. El-Asrag, M.E., et al., *Biallelic mutations in the autophagy regulator DRAM2 cause retinal dystrophy with early macular involvement*. Am J Hum Genet. 96(6): p. 948-54.
2. Brownstein, C.A., et al., *Mutation of KCNJ8 in a patient with Cantu syndrome with unique vascular abnormalities - support for the role of K(ATP) channels in this condition*. Eur J Med Genet. 56(12): p. 678-82.
3. Cooper, P.E., et al., *Cantu syndrome resulting from activating mutation in the KCNJ8 gene*. Hum Mutat. 35(7): p. 809-13.
4. Li, J., et al., *Whole-exome sequencing identifies a variant in TMEM132E causing autosomal-recessive nonsyndromic hearing loss DFNB99*. Hum Mutat. 36(1): p. 98-105.
5. Mencacci, N.E., et al., *De Novo Mutations in PDE10A Cause Childhood-Onset Chorea with Bilateral Striatal Lesions*. Am J Hum Genet. 98(4): p. 763-71.
6. Fujishige, K., J. Kotera, and K. Omori, *Striatum- and testis-specific phosphodiesterase PDE10A isolation and characterization of a rat PDE10A*. Eur J Biochem, 1999. 266(3): p. 1118-27.
7. Coskran, T.M., et al., *Immunohistochemical localization of phosphodiesterase 10A in multiple mammalian species*. J Histochem Cytochem, 2006. 54(11): p. 1205-13.
8. Montague, C.T., et al., *Congenital leptin deficiency is associated with severe early-onset obesity in humans*. Nature, 1997. 387(6636): p. 903-8.
9. Granata, A., et al., *An iPSC-derived vascular model of Marfan syndrome identifies key mediators of smooth muscle cell death*. Nat Genet. 49(1): p. 97-109.
10. Friedrich, F.W., et al., *Evidence for FHL1 as a novel disease gene for isolated hypertrophic cardiomyopathy*. Hum Mol Genet. 21(14): p. 3237-54.
11. Farooqi, I.S., et al., *Effects of recombinant leptin therapy in a child with congenital leptin deficiency*. N Engl J Med, 1999. 341(12): p. 879-84.
12. Muller, G.A., et al., *Allele-specific siRNA knockdown as a personalized treatment strategy for vascular Ehlers-Danlos syndrome in human fibroblasts*. FASEB J. 26(2): p. 668-77.
13. Kishnani, P.S., et al., *Chinese hamster ovary cell-derived recombinant human acid alpha-glucosidase in infantile-onset Pompe disease*. J Pediatr, 2006. 149(1): p. 89-97.

APPENDIX C: SEMIDOMINANT MODE OF INHERITANCE OVERVIEW

A semidominant mode of inheritance (MOI) is applied to disease entities in which both autosomal dominant (AD) and autosomal recessive (AR) MOIs are observed and represent a continuum of disease (e.g. the same phenotypes are observed for both MOIs at differing severities). See more explanation on page 10. Determination of semidominant inheritance is made according to the ClinGen Lumping and Splitting guidelines.

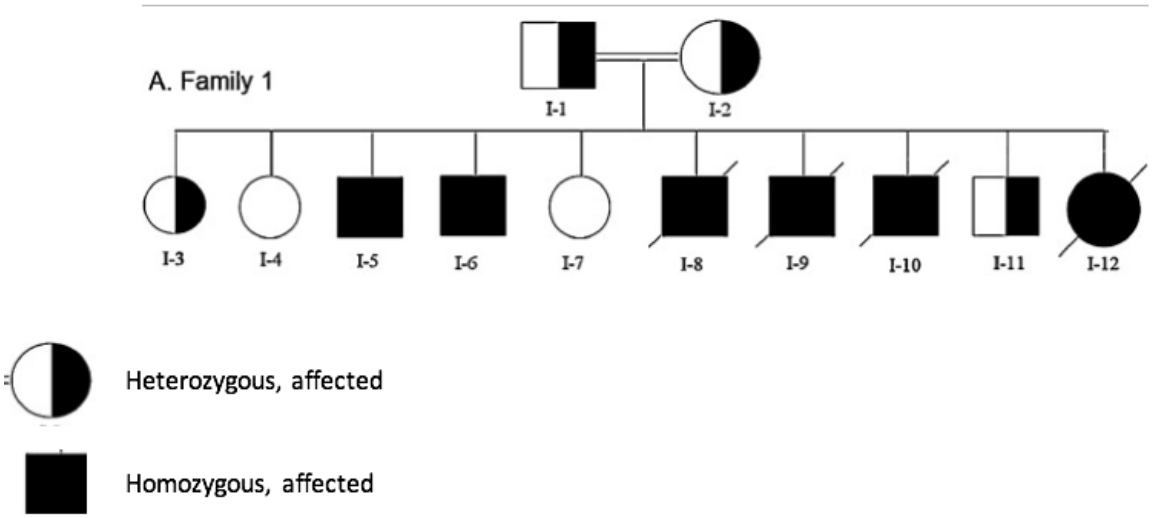
Selection of the semidominant MOI in the GCI allows scoring of individual case reports that have either AD or AR inheritance, as well as inclusion of segregation scoring for pedigrees displaying either AD, AR, or semidominant MOI, in the same gene-disease-MOI record.

When scoring individual case-level evidence in a semi-dominant curation, score each variant in accordance with the context in which it is observed, e.g., heterozygous variants should be scored as a typical heterozygous variant would be scored, and biallelic variants should be scored as typical biallelic variants would be scored. When working within the semidominant MOI in the GCI, all "Case Information Type" options are available for use in the scoring module to accommodate these different scenarios.

For segregation, evaluation and scoring will be prioritized based on the MOI displayed in the family being evaluated, and includes either AD, AR or semidominant MOI, and will follow the specifications and guidelines provided in the Segregation section beginning on page 25. Briefly, if a published LOD (pLOD) score is provided, use this score and indicate the MOI (AD, AR, or semidominant) of the family, as well as the sequencing method to appropriately categorize the evidence for scoring. If no pLOD is provided, a LOD score can be estimated (eLOD). In cases in which a family is either strictly AD or strictly AR, the families must meet the minimum required segregations or affected number of individuals for inclusion. Briefly, for AD this means at least 4 segregations within one pedigree must be present to estimate a LOD score; and for AR, at least 3 affected individuals with the genotype (phenotype^{+}/ genotype^{+}) are required to include an eLOD in the overall genetic evidence score. If using the GCI, the interface will calculate the eLOD based on the logic provided in the Segregation section on page 25. For cases in which a family displays a semidominant MOI, where affected individuals in the family represent both AD and AR inheritance, and a pLOD is not provided, the eLOD is calculated from EITHER the AD individuals OR the AR, whichever group meets the current specifications listed above. Examples of estimating a LOD score from semidominant pedigrees are provided below.

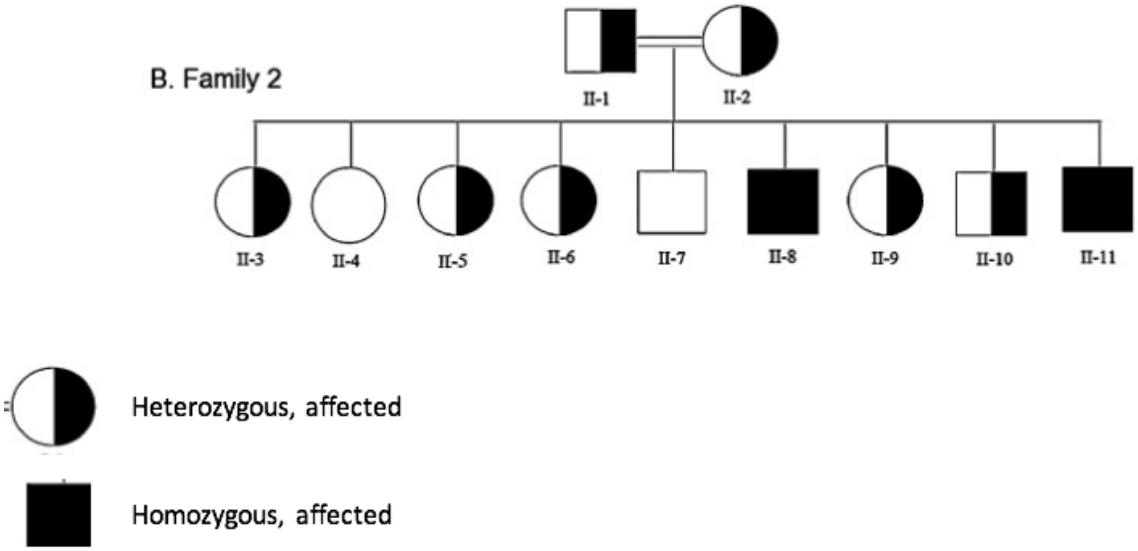
NOTE: The GCI will NOT calculate an appropriate eLOD if you enter in both AR and AD segregation information at the same time. Only one MOI can be used to apply an eLOD.

Semidominant Pedigree Example #1:



This semidominant family meets the criteria for AR segregation inclusion, as there are 6 affected, genotype positive individuals in the pedigree (I-5, I-6, I-8, I-9, I-10, I-12). Whereas, only 2 segregations are present to an AD MOI, which does not meet the requirement of 4 segregations to include an eLOD in the final genetic evidence score.

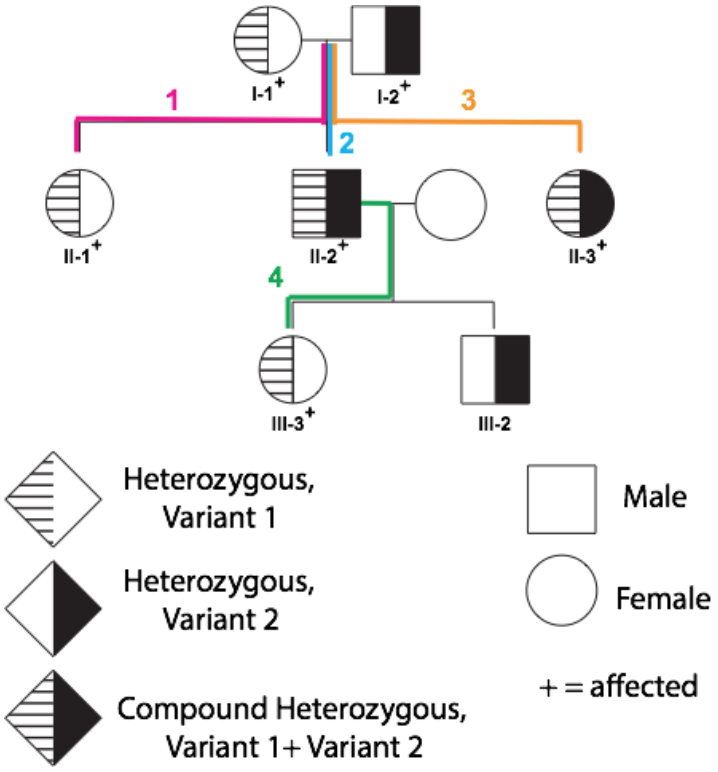
Semidominant Pedigree Example #2:



This semidominant family meets the criteria for AD segregation inclusion, as there are 5 segregations among genotype^{+}/phenotype^{+} individuals (counting from either II-1 or II-2 down to each of the 5 affected children). It does not meet the criteria for AR segregation inclusion, as there are only 2 genotype^{+}/phenotype^{+} individuals within the pedigree.

For semidominant families where two different variants in the same gene of interest are present in the pedigree and AR individuals are compound heterozygous carrying each variant of interest, the same rules apply; however, segregations among AD MOI should be

restricted to one variant of interest. Furthermore, if there are three or more generations present in the pedigree, segregation for AD can include individuals with the variant of interest that are AR. For example, in semidominant pedigree Example #3 below, there are 4 segregations among carriers of Variant 1. In this case AR II-2 can be counted as they are a carrier of Variant 1 and between two AD carriers of the same variant. Variant 2 could not be counted towards segregation points as there are only 3 segregations, therefore it does not meet the minimum 4 segregations required. When scoring segregation from semidominant pedigrees containing AR compound heterozygous cases, please make a note of the variant that met the inclusion criteria in the GCI under the "Additional Segregation Information" section.

Summary of Pedigree #3: Compound heterozygous individuals can only be counted if they have a parent who is affected that is genotype^{+} for at least one variant of interest, and a child that is affected with the same variant of interest in the parent.

Semidominant Pedigree Example #3: