Gene-Disease Validity Curation Process
Standard Operating Procedure

Version 11
September 2024
The Clinical Genome Resource
Gene Curation Working Group

Table of Contents

BACKGROUND

ClinGen's gene curation process is designed to aid in evaluating the strength of a monogenic gene-disease relationship based on publicly available evidence. Information about the gene-disease relationship, including genetic, experimental, and contradictory evidence curated from publicly available sources is compiled and used to assign a gene-disease 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 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 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.
    - o The ClinGen Lumping and Splitting guidelines must be consulted to determine the disease entity for curation. Please see guidelines here.
        - ▪ Publication: Thaxton et al, 2022 PMID: 35754516
- Guidance on disease naming can be found on the Disease Naming Advisory Committee page here on the ClinGen website
    - o If you need assistance with naming consider emailing diseasenaming@clinicalegenome.org
- Access to scientific articles and publications.
    - o Note: valid evidence may be present in pre-publication articles, such as bioRxiv, medRxiv, etc. 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; while some pre-publication articles do have PubMed IDs that could technically be entered into the GCI, the preference at this time is to document them and their impact on scoring in the evidence summary only.
- Access to the ClinGen Gene Curation Interface (GCI), found here:
    - o Access is granted to users that are actively participating on a ClinGen gene curation expert panel (GCEP). Users may register themselves for GCI access, but coordinators for the GCEP are responsible for confirming affiliation access. If you have trouble accessing the GCI once an account is set up, please contact clingen-helpdesk@lists.stanford.edu.
    - o 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.
- o For more information on the GCI, please see the following manuscript PMID: 38663031
- Access to the ClinGen GeneTracker (GT), found here:
  - o Access is granted to users that are actively participating on a ClinGen gene curation expert panel (GCEP) on an as needed basis. Access must be confirmed and approved by the GCEP Coordinator. To set up an account (if needed), please email clingentrackerhelp@unc.edu and cc your GCEP coordinator, stating your preferred email for login and the GCEP with which you are participating. Please confer with your GCEP coordinator on whether or not access to GeneTracker is necessary.

Optional: An SOP has been developed to assist in evidence collection through the use of Hypothes.is, a tool 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.
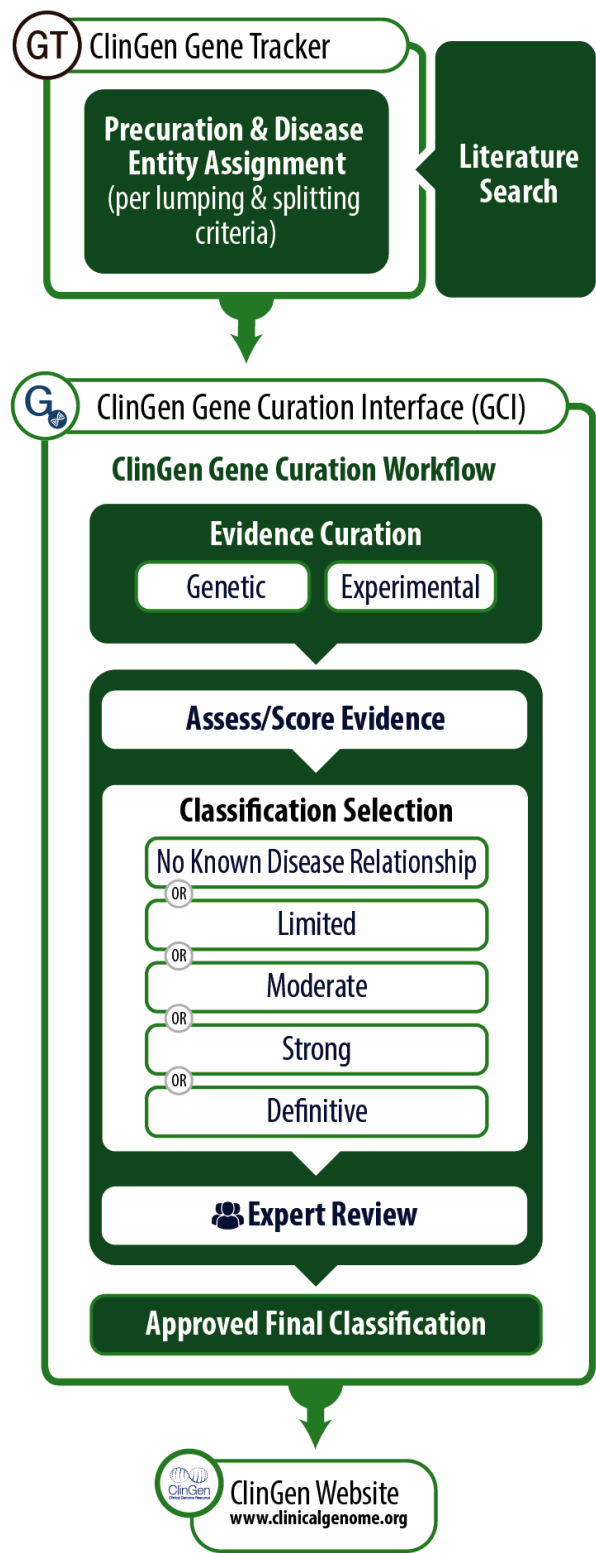
OVERVIEW OF GENE CURATION

The gene curation framework consists of the following essential steps in order to assign a 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 (GDM) to be used in curation
- Evidence collection
  - a. Genetic Evidence
  - b. Experimental Evidence
- Evaluation and scoring of evidence
- Expert Review, final classification and approval of a gene-disease relationship
- Publication of final classification to www.clinicalgenome.org

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 gene-disease validity classification.

Figure 1: GENE CURATION WORKFLOW

GT ClinGen Gene Tracker

Precuration & Disease Entity Assignment (per lumping & splitting criteria)

Literature Search

G ClinGen Gene Curation Interface (GCI)

ClinGen Gene Curation Workflow

Evidence Curation

Genetic | Experimental

Assess/Score Evidence

Classification Selection

No Known Disease Relationship
OR
Limited
OR
Moderate
OR
Strong
OR
Definitive

Expert Review

Approved Final Classification

ClinGen Website
www.clinicalgenome.org

GENE-DISEASE VALIDITY CLASSIFICATIONS

The [ClinGen Gene Curation Working Group](#) has developed a method to qualitatively define the validity of a gene-disease relationship using a classification scheme based on the strength of evidence that supports or contradicts the claimed relationship. This framework allows the 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 evidence level. 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 with contradictory evidence likely also have evidence supporting the gene-disease relationship. In these cases, the strength of evidence supporting versus opposing the gene-disease relationship should be weighed by the expert panel before a final classification is assigned.

SUPPORTIVE EVIDENCE

The links below are intended to provide examples of curations with the specific, denoted classification, however, classifications may change with time so please check the website for the latest classifications.

Definitive
The role of this gene in this particular disease has been repeatedly demonstrated in both the research and clinical diagnostic settings, and has been upheld over time (in general, at least 2 independent scored publications documenting human genetic evidence over at least 3 years' time). 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 page 25 for more information. As with the "strong" category, different types of supporting experimental data is typically also present, but is not required to reach this designation if substantial convincing genetic evidence is present. Examples of appropriate types of supporting experimental data are based on those outlined in MacArthur et al. 2014. No convincing evidence has emerged that contradicts the role of the gene in the specified disease. Definitive curation examples (as of August 2023) are below.

- *OCA2* - oculocutaneous albinism type 2 (AR)
- *HTT* - Huntington disease (AD)
- *LDLR* - hypercholesterolemia, familial, 1 (SD)

- *KDM6A - Kabuki syndrome (XL)*

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. Gene-disease pairs with strong evidence demonstrate considerable genetic evidence (numerous unrelated probands harboring variants with sufficient supporting evidence for disease causality). Compelling gene-level evidence from different types of supporting experimental data is typically also present, but is not required to reach this designation if substantial convincing genetic evidence is present. In addition, no convincing evidence has emerged that contradicts the role of the gene in the noted disease. Evidence should total ≥12 points per the SOP to reach this designation. Strong curation examples (as of August 2023) are below.

- *FNIP1 - FNIP1-associated syndrome (AR)*
- *RAC2 - immunodeficiency 73b with defective neutrophil chemotaxis and lymphopenia (AD)*
- ANKRD17 - syndromic intellectual disability (AD)

Moderate

There is **moderate** evidence to support a causal role for this gene in this disease. Gene-disease pairs with moderate evidence typically demonstrate some convincing genetic evidence (probands harboring variants with sufficient supporting evidence for disease causality with or without moderate experimental data supporting the gene-disease relationship). 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. Evidence should be between 7-11 points per the SOP to reach this designation. Moderate curation examples (August 2023) are below.

- *RPL10 - X-linked syndromic intellectual disability (XL)*
- *MACF - lissencephaly spectrum disorder with complex brainstem malformation (AD)*
- *RFT - RFT1-congenital disorder of glycosylation (AR)*
- *JPH2 - dilated cardiomyopathy (SD)*

Limited

In general, the category of limited should be applied when experts consider the gene-disease relationship to be plausible, but the evidence is not sufficient to score as Moderate. Example scenarios where a classification of "Limited" may be warranted include **(but are not limited to):**

- A moderate number of cases with a *consistent but not highly specific phenotype*. The variants have some support for pathogenicity, but there is little to no functional evidence to support variation.

- A small number of cases with *well-defined, consistent phenotypic presentations*. The variants are plausible causes of disease given the prevalence of the condition and the inheritance pattern.
- A single case with a *rare and distinct phenotype* and a *de novo* occurrence in a highly constrained gene.
- A single case with a *rare and distinct phenotype* and biallelic, loss of function variants.

The Limited category should NOT be applied in circumstances where none of the presented evidence is compelling; in these circumstances, the Disputed category should be considered. Limited curation examples (as of August 2023) are below.

- *SNAI2*- Waardenburg's syndrome (AR): Example in which there is limited genetic data in the presence of a supportive animal model
- *TWIST1* - Sweeney-Cox syndrome (AD): Example in which there is limited genetic data in the presence of a supportive animal model
- *LAS1L* - X-linked syndromic intellectual disability (XL): Example in which some of the reported genetic evidence was not scored, experimental evidence was documented but not scored due to unclear relationship to the disease

NO KNOWN DISEASE RELATIONSHIP

Evidence for a causal role in the monogenic disease of interest (determined using ClinGen lumping and splitting guidance) has not been reported within the literature (published, prepublished and/or present in public databases [e.g. ClinVar, etc.]). These genes might be "candidate" genes based on linkage intervals, animal models, implication in pathways known to be involved in human disease, etc., but no reports have directly implicated the gene in the specified disease. If a claim of a relationship with the specified disease has been reported, but the evidence is minimal or not compelling, consider Limited, Disputed, or Refuted. 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. No known disease relationship curation examples (as of August 2023) are below.

- *ACAT2* - acetyl-CoA acetyltransferase-2 deficiency (Undetermined) Animal model only: Example in which individuals have been reported with an enzyme deficiency but no variants in the gene have been reported.
- *PEX11A* - peroxisome biogenesis disorder (AR): Example in which the gene is a member of a family in which other genes have previously been implicated in disease, but no variants in this gene in affected individuals have been reported.

NOTE: 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.

CONTRADICTORY EVIDENCE

Although there has been an assertion of a gene-disease relationship, the initial evidence is not compelling from today's perspective and/or conflicting evidence has arisen. **Example scenarios include (but are not limited to):**

Disputed

- Only a few cases with non-specific, genetically heterogeneous phenotypes and missense variants; no convincing experimental data available.
- All reported cases have been scored at 0 (or the sum of genetic evidence is below 1) after GCEP review.
- The initially reported variants have now been identified as having a population frequency too high to be consistent with disease.

Disputed curation examples (as of August 2023) are below.

- *DPP6 - complex neurodevelopmental disorder (AD)*: Numerous variants have been reported in this gene in individuals with seemingly disparate phenotypes, inherited from reportedly unaffected parents, or observed in control populations. Gene is not highly constrained for either protein truncating variants or missense variation, calling into question the relevance of previously reported variants in affected individuals.
- *INO80 - immunodeficiency, common variable, 1 (AR)*: Example in which the only reported variants have been ruled out as plausible causes of disease (in this case, due to population frequency). Gene-level experimental evidence provided minimal support for the possibility of a gene-disease relationship.
- *ZNF674 - X-linked intellectual disability (XL)*: Example in which the only reported variants have been ruled out as plausible causes of disease (in this case, a combination of individuals with other possible causative variants, population frequency, and individuals with whole gene deletions without the phenotype). No gene-level experimental data has been reported.

Refuted

Evidence *refuting* the initial reported evidence for the role of the gene in the specified disease has been reported and significantly outweighs any evidence supporting the role. This designation is to be applied at the discretion of clinical domain experts after thorough review of available data. **Example scenarios include (but are not limited to):**

- All existing genetic evidence has been ruled out, leaving the gene with essentially no valid evidence remaining after an original claim.
- Initially reported probands were later found to have an alternative cause of disease.
- Initially reported probands were later determined NOT to have the disease in question.

- Statistically rigorous case-control data demonstrate no enrichment in cases vs. controls.

Refuted curation examples (as of August 2023) are below.

- *GJB6 - nonsyndromic genetic deafness (AR)*: Example in which the variants that were originally reported were not specific to the gene under evaluation (large deletions), and includes a regulatory region controlling another gene known to cause the phenotype (GJB2).
- *RYR2 - arrhythmogenic right ventricular cardiomyopathy (AD)*: All originally reported probands were later determined to have a different disease.
- *BLK - monogenic diabetes (AD)*: In this curation, there is some scored genetic evidence, however, the experts felt that available contradictory evidence outweighed this information. Several of the originally reported variants have since been found to be too common in the general population and present in normoglycemic individuals. A high prevalence of loss of function variants in this gene has been reported in the general population, and there has not been a demonstrated over-representation of rare variants in this gene in monogenic diabetes cohorts.

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 precuration and curation records and/or obtaining the precuration ID 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. GCEP Coordinators are the only individuals that can assign a gene and precuration record to a curator in the GeneTracker, so please check with your GCEP coordinator on your GeneTracker access. For additional questions or concerns on the GeneTracker please email clingentrackerhelp@med.unc.edu.

Currently, the GCI will only allow a single record for a given gene-disease-MOI. This is to reduce redundancy of curations among the various GCEPs. In order to check whether a given gene is of potential interest to other GCEPs, curators are directed to search the ClinGen GeneTracker before beginning a curation.

DEFINING THE DISEASE ENTITY

Many human genes are implicated in more than one disorder. 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 (35754516). 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 GeneTracker and GCI per the defined workflow for gene curation as represented in Figure 1. After review and discussion, the GCEP will determine which disease entity or entities to curate. 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 (under Precuration section).

Precuration identifiers: As of June 1, 2023, a precuration identifier (precuration ID) issued from the GeneTracker is required to start all new GCI records. The HGNC gene symbol, Mondo Disease Ontology identifier (MONDO ID), mode of inheritance (MOI), affiliation (i.e., GCEP) and precuration ID must match between the GeneTracker and GCI to proceed with starting a GCI GDM, adding evidence and generating a final classification. This step is to ensure the linking of these two critical pieces of data to the final curation record, which are both published to the ClinGen website. It is critical to fill out all precuration records accordingly as the information is published to the ClinGen website.

Mondo Identifiers: All gene-disease-mode of inheritance records require the use of a Monarch Disease ontology identifier (Mondo ID). If there is not an appropriate, current Mondo identifier (Mondo ID), or the name and/or definition is not accurate, you can create and/or update the Mondo ID by contacting Mondo. Directions on how to do this can be found in the GCEP protocol (section 4.5: Assist Biocurators with updating and/or creating new Mondo request: Disease nomenclature and or requesting Mondo Identifiers) which can be found at the link here. For more information on the current recommendations for disease naming please see https://www.clinicalgenome.org/working-groups/disease-naming-advisory-committee/.

Mode of inheritance (MOI): Like disease entities, a gene may also be reported 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 are described with "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 (which may vary by severity), 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) typically represents an intermediate phenotype (as a/a refers to the wild-type) compared to the homozygous mutant state (A/A), which may 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, with 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](#).

At this time there is one MOI that cannot be scored in the GCI (Undetermined MOI) (Table 1)**.** For this choice, manual modification of the gene-disease 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 classification is assigned. Of note, if "other", and any adjectives under this choice (including Y-linked, somatic mutation, multifactorial inheritance, and codominance) is the MOI chosen for a gene-disease relationship, the final classification will NOT be permitted to be published on the ClinGen website. Therefore, use caution when making this choice.

If you have made an error in the choice of MOI for a gene-disease relationship, please contact the GCI Help Desk, as only a limited number of MOIs can be updated for a record, and in general, making changes to MOI are not possible by the curator. Table 1 below describes the ability to update a GDM record MOI, which is restricted to updates allowed for MOIs that are monoallelic (e.g. Autosomal Dominant, X-linked). MOI changes that are not possible include autosomal dominant/X-linked to autosomal recessive and semidominant to either autosomal dominant/X-linked or autosomal recessive, so choose the MOI carefully at the precuration stage and before creating a GCI record. If a mistake has been made between one of these MOIs, a new gene-disease-MOI record may need to be created.

Table 1. Mode of Inheritance (MOI) choices in the GCI					
MOI type	Score in GCI	GCI Calculated classification	GCI Modified classification	Ability to change MOI	Publish to website
Autosomal Dominant (HP:0000006)	✓	✓	✓	only to X-linked	✓
Autosomal Recessive (HP:0000006)	✓	✓	✓	✕	✓
Mitochondrial (HP:0001427)	✓	✓	✓	only to autosomal dominant or X-linked	✓
Semidominant (HP:0032113)	✓	✓	✓	✕	✓
X-linked (HP:0001417)	✓	✓	✓	only to autosomal dominant	✓
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.

Check with your GCEP(s) to determine well-known and trusted public databases (e.g. ClinVar, DECIPHER) containing clinical data pertinent to your group, and to determine in which circumstances these cases may be used. When determining whether a case is appropriate for use, consider the following:

- Case must be publicly accessible. For example, do not include cases from DECIPHER that are only available to authorized users.
- Case is well-described with appropriate phenotype, testing, and other variant information.
- Case is not otherwise believed to be described in the literature.
- Evidence for why the variant classification was made is present. A case annotated with having a "pathogenic" variant and no other supporting information may not be sufficient for use.

At this time only evidence that has either an associated PMID or a ClinVar SCV (Submitted record in ClinVar) number can be recorded and scored in the GCI.

Instructions for adding a PMID to the GCI can be found here:
https://vci-gci-docs.clinicalgenome.org/vci-gci-docs/gci-help/adding-pubmed-articles

Instructions for adding a ClinVar SCV can be found [here](#).

Instructions for adding other types of evidence (e.g. cases from other public databases, preprints, etc.): Other databases that may include relevant curation information may have flagship papers that can be used as a proxy to enter the information. For example, [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 (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.

If there is no publication reported in the database, or if you would prefer not to utilize the general paper, describe the case in the free-text evidence summary, and manually adjust the classification if necessary. When describing such cases, please include the database identifier (for example, DECIPHER ID) and relevant links to the information where possible. For a list of general databases of interest and associated PMIDs for scoring, please see Appendix A.

If relevant information is contained within a preprint (e.g., an article posted on BioRxiv or MedRxiv), please document it in a similar manner, by describing it in the free-text evidence summary and manually adjusting the classification if necessary. Even though some preprints are associated with PMIDs, we ask that you NOT use these to enter the articles at this time, as the article will be associated with a different PMID upon publication. Development is underway to support the entry of pre-prints into the GCI; users will be notified via email when this feature is available. Until that time, enter this information as described [here](#) in the GCI Help Documentation.

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.
- [GenCC](#)

- ○ The Gene Curation Coalition (GenCC) is like a ClinVar for gene curations. Multiple submitters submit assertions for gene-disease relationships.
  - ○ Users can filter by gene, condition, submitter, or clinical validity (clinical validity terms used are harmonized with ClinGen terms for direct comparison). Searching by gene is the most inclusive and recommended search.
  - ○ Please use this database for its citation of primary evidence or literature to add to curations. Contact information for submitters can be found on their [submitter pages](#) if more information is needed.
- ● 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" classifications), 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.**

Gene Curation Expert Panels (GCEPs) may find it helpful to develop scoring recommendations for their group in order to apply consistent changes across curations. This is especially helpful for groups that have extensive lists and/or GCEPs that accept new members regularly. While the Gene-Disease Validity SOP describes the types of evidence to score and sets a default for each evidence type, it also describes a range of points. GCEPs may find it useful to set some criteria on when to increase from default (upscore) or decrease from default (downscore) points given the strength of evidence provided.

The GCEP Scoring recommendations are not required, and should not override guidance put forth in the official SOP, rather, they are used to complement the existing official Gene-Disease Validity SOP, providing additional specification where necessary. If GCEP are interested in developing scoring recommendations please see the following for more information:
- The current version of the GCEP Protocol
- Internal GCEP Scoring Recommendation README document
  - Provides information on how to develop scoring recommendations including template document and a folder containing example GCEP Scoring Recommendations

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. Please designate which paper is the "original" utilizing the checkbox feature in the GCI.
    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. 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 2).

	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	Total LOD Score	Sequencing Method		0-3 points		3 points
				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 Study Type	**Case-Control Quality Criteria**	**Suggested Points/Study**				**Points Given**	**Max Score**
Case-Control Data	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**

Figure 2: Genetic Evidence Summary Matrix.

In the case of AR conditions, evaluate each variant (in trans) independently, then combine for the final score.

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, as well as 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 (or submitters, in the case of a ClinVar entry) 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.). Curators should consider both the evidence supporting or contradicting the plausibility of the variants' possible role in disease as well as the veracity of the reported clinical diagnosis in order to determine how this evidence should be scored according to this framework regardless of any claims that may have been assigned by the authors or submitters (in the case of a ClinVar variant). Each case may be given points for both variant evidence (see below for details) and segregation analysis (see page 28 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 2) may score at the top of its range (up to 1.5 points) 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, missense variants within the functional domain related to 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, 0.05 (only for homozygous missense variants), 0.1, 0.25, 0.5, 0.75, 1.0, etc. increments after 0.1.**
    - For AD and XL curations, scores are chosen from a dropdown menu with the options described above.
    - For AR curations: scores are chosen from the dropdown menu for each variant, then added together by the GCI. Note that this may result in scores for AR cases having different numerical values than those represented in the dropdown.
        - 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, which is not an option in the typical dropdown menu but is nonetheless an appropriate score for this case.
- **Please note that when entering evidence in the GCI at the individual level, proband labels must be different across publications. If the same proband identifier, such as "Proband 1," is used across several publications, the interface system recognizes this as the same individual which will affect scoring and website display. Please use different labels for probands, for instance adding the first author name followed by the identifier in the paper (i.e. "Wang Proband 1").**

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. For homozygous variants, the variant scored is then doubled because it is present on both alleles (see examples below). Some caveats to the evaluation of biallelic variants include:
- 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 in consanguineous families, consider downgrading the maximum number of points such cases could receive given these probands likely have multiple

homozygous variants due to runs of homozygosity. In these scenarios it is unclear which, if any, of these homozygous variants are causative. This concern may be magnified if targeted or single gene testing was completed. 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 be summed to 1.4 by the GCI.
    - Example 3: homozygous,inherited nonsense variant (1.5*2) =3.0
    - Example 4: homozygous, inherited missense variant with supportive functional evidence ((0.1+0.4)*2) = 1.0
    - Example 5: homozygous nonsense variant with functional evidence (1.5+0.4)*2= 3.8, **GCI will cap at 3 points**

When entering biallelic variants into the GCI, the curator will have to check a box to attest that the two variants are confirmed or suspected *in trans*. If the phase of the variants is entirely unknown, the curator is encouraged to enter the relevant PMID and mark it as nonscorable on the curation panel of the GCI landing page (e.g. "[curation central](#)"). Once the curator confirms that the two variants are confirmed or suspected *in trans,* there will be a second set of checkboxes where the curator has to choose if the variants are confirmed *in trans,* suspected *in trans,* or unknown. The curator should refer to the publication to provide this information.

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.

Additional Case-Level Scoring Considerations

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.
  - Postzygotic mosaicism is also considered as a *de novo* occurrence and should be scored as such. The same caveats to determine this type of mosaicism apply, including that parents are tested and found negative for the variant of interest. Further, the proband being evaluated should have reasonable phenotypes consistent with disease to be scored, as not all instances of postzygotic mosaicism will result in disease onset.

- 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" [page 15](page-15)). 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. As of 2023, single and multi-exon deletions (i.e., intragenic copy number variants [CNVs]) can now be formally entered as evidence into the GCI. If the variant has an existing ClinVar ID, this can be used in the same manner it is for other variants. If the intragenic CNV does not have a ClinVar ID, it can be registered in the ClinGen Allele Registry. Input the corresponding CACN ID into the GCI for documentation and scoring. In general, CNVs used as evidence to support a gene-disease validity curation should be intragenic or involve only a single gene; for CNVs involving multiple genes, it is difficult to determine the effect of the other involved gene(s). Multigenic CNVs or other structural variants should not be included unless the experts on the GCEP feel it is appropriate and explanatory rationale is provided.

- While other variant types, such as missense, may have sufficient evidence demonstrating complete loss of function, we recommend entering those in the "Other Variant with Gene Impact" category and applying upgraded scoring as appropriate. Similarly, some putative null variants have evidence suggesting they do not result in loss of function; for these, we recommend entering those within the "Predicted or Proven Null" category and applying downgraded scoring as appropriate. In either scenario, please detail the rationale behind the non-default scoring in the "Reason for Changed Score" free text box. For example, if there is a missense variant with functional evidence demonstrating that it is acting via loss of function, this variant could be

scored in a similar point range as a predicted/proven null variant, even though it is entered in the "Other" variant category.

- 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 10 for further instruction.

- 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.

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. Consider further upgrading variants with validated functional evidence consistent with a gain of function mechanism.

- 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 (or increased, depending on the mechanism of disease) 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 in general do not provide sufficient evidence for functional impact and are therefore not typically counted as supportive functional data (i.e. upgrades are typically not given for this information). Note that this guidance is distinct from that made by the ClinGen Sequence Variant Interpretation (SVI) group to describe the role of *in silico* predictors in evaluating variants in genes with established gene-disease relationships (Pejaver et al. 2022). In

23

rare circumstances, expert panels may decide to award some upgrade over the default starting points for particularly compelling *in silico* information (e.g. impact on 3D structure)

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 18-22.
  - 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. These variants may be upgraded at the discretion of the GCEP. 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 page 35). 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 page 28). 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 reviewed during curation and used as a check to assess the plausibility of the variant being disruptive. For example, missense variants with a low REVEL score are not particularly suspicious for pathogenicity and therefore may not be scored.

- 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," page 15), 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. 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 related to multiple diseases, LOF constraint could be related to 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. Note the transcript gnomAD returns may not be the most clinically relevant transcript. 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 exome or 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 reported in connection 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 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.). See the X-linked pedigree below for an example:

X-linked Pedigree
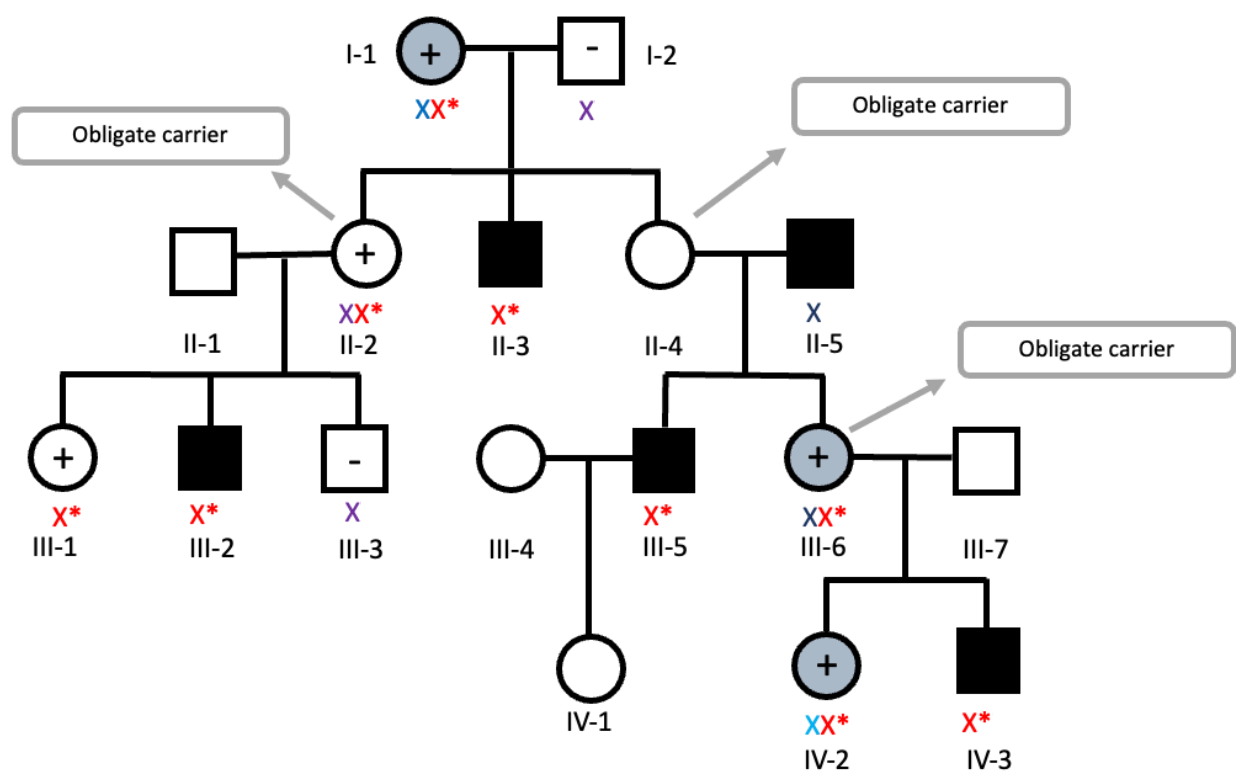
Obligate carriers in an X-linked pedigree. An obligate carrier an individuals suspected of carrying and passing on the suspected pathogenic allele even if they are not molecularly confirmed to have the pathogenic allele, as they are the only person between to affected (genotype+/ phenotype+) individuals. In this pedigree II-4 represents an obligate carrier. as they are the only person between two affected (genotype+/ phenotype+) individuals (represented as I-1 and III-5). Individuals II-2 and III-6 are also obligate carriers, however they have been molecular confirmed to carrier the suspected allele (indicated by the red X with an asterisks). Red X's with the asterisks represent the inherited X-chromosomes with a genetic variation predicted to be pathogenic for a monogenic disease. Blue X's indicated the normal inherited allele. Circles represent females, squares present males. Black and gray shaded circles indicate affected individuals (phenotype+).

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*:
*assuming a carrier mother, not an affected father

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

NOTE: The base number "0.5" used in this equation represents the risk of inheriting a disease allele in the typical autosomal dominant/X-linked disease model (presuming a carrier mother). If the father is the affected individual in an X-linked disorder scenario, the base number should be changed to "1" for any daughters to reflect their risk of inheriting the disease allele.

Figure 4: 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

*Utilizing the formula above as written

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 5 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

30

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, as in the case of the *CRADD*/Syndromic intellectual disability curation (see Harel et al., 2017). The equation below is adjusted to accurately reflect the risk of inheritance.

$$Z \text{ (LOD score)} = \log_{10}\frac{1}{(0.5)^{\text{\# of Affected Individuals-1}}(0.5)^{\text{\# of Unaffected Individuals}}}$$

NOTE: The GCI provides an estimated LOD score utilizing the formula used in a typical AR disease model (assuming both parents are heterozygous carriers). If your situation is different and you need to adjust the denominator, do not rely on the table below (Figure 5) or the GCI-calculated LOD score.

Figure 5: 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.
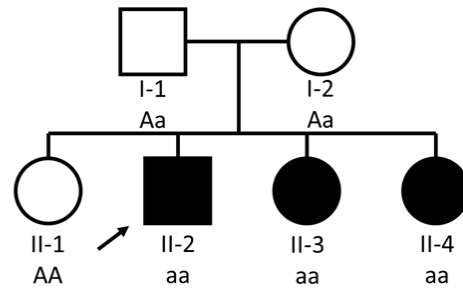
Pedigree A
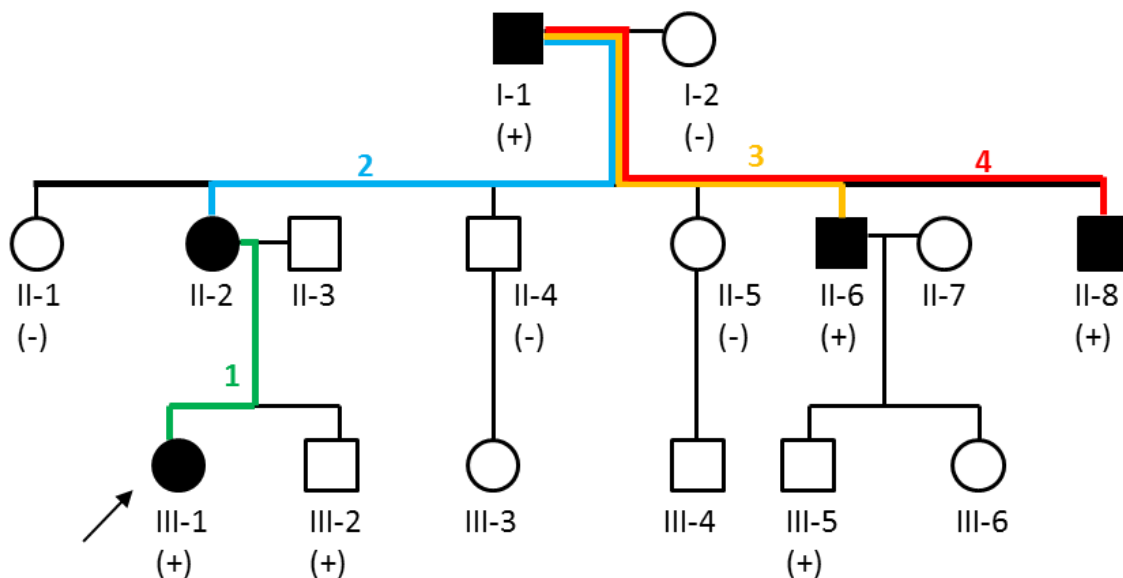
- When estimating LOD scores for autosomal recessive disorders, count unaffected individuals as those who would be at the same risk to inherit two altered alleles as an affected individual, i.e., homozygous normal or heterozygous carrier siblings of a proband. For example, there are two unaffected individuals in Pedigree B, one unaffected individual in Pedigree C, and two unaffected individuals in Pedigree D.

    - If calculating LOD scores for autosomal recessive cases in which a proband is homozygous, variant phasing is not required in order to count appropriate individuals in the family(ies). Parents are not typically counted in the eLOD calculation; only individuals at the same degree of risk as the proband to inherit both variants (e.g. siblings) are considered in the eLOD calculation.

    - If calculating LOD scores for autosomal recessive cases in which the proband has compound heterozygous variants, it is recommended that variants be phased. For example, at a minimum, at least one parent must be genotyped to count appropriate individuals in the family for the eLOD calculation.

- 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.
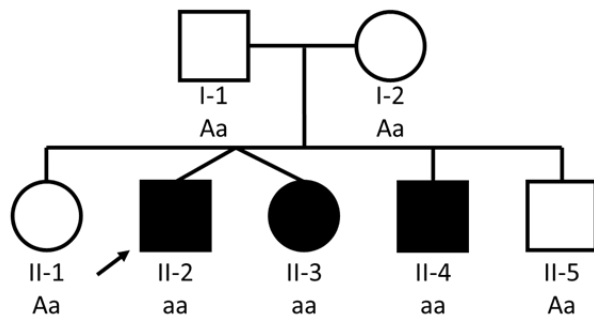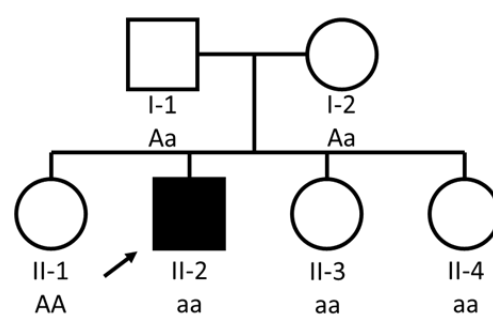
Pedigree B
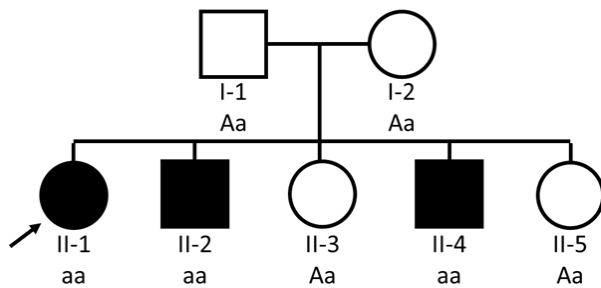
Pedigree C

Pedigree D

Pedigree E

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 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 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 6 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 6 are merely a guide for the curator.

Figure 6: 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.

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 6).

D = Points assigned if total LOD had been obtained only by exome/genome sequencing/all genes in candidate region sequenced approach (see Figure 6).

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 here. 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 4). 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 6, 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 used within the GCI is appropriate for use in the majority of scenarios, there are some situations for which additional logic must be used. For example, in the scenario where one has one LOD score generated with exome data and another LOD score generated with candidate gene sequencing data, the resultant suggested points as calculated by the GCI may be lower for this combined scenario than it may be if only the exome LOD had been entered. To illustrate this, consider the following: For Family 1, an estimated LOD score of 3.1 is obtained from a study involving exome sequencing. 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 6). 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. The candidate LOD case can still be entered, but do not check the "Score?" box in order to exclude it from the final calculation.

We recognize that the methods in each publication vary. Therefore, the suggested points in Figure 6 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. Select status for the case-control studies:
   - Score: Select this option when case-control data is supportive of the gene-disease relationship. The 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.
   - Contradicts: Select this option when a case-control study presents contrary evidence that may bring the gene-disease relationship into question. Note that no score can be assigned if the status is set to "Contradicts," but this will result in a "Yes" in the contradictory evidence field in the final scoring matrix.
   - Review: Select this option if the curator is unsure of the significance of the case-control information and wishes to review it with the expert panel. Note that if "Review" is selected, the curator is unable to assign a score. If the expert panel decides to score this information following discussion, the status will need to be changed to "Score."

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 a gene-disease relationship. **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, sex, self-reported ancestry, 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 7: Case-control Genetic Evidence Examples

Detailed examples and explanations for assigned points are provided in the table below.

Figure 7. 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, ancestry, 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 ancestry	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 ancestry. 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 ancestry. 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
Author E 2021 (Contradicts)	Breast cancer cases:27/322 47 Controls:	Matched by age, ancestry, and location	Cases & controls sequenced by next generation	OR: 1.19 [95% CI: 0.67-2.17; P = 0.55]	Aggregate analysis	Contradicts

	21/32544		sequencing panels			

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 not by ancestry or age) and the p-value is moderately significant. NOTE: Location can be a poor proxy for ancestry in certain cases. If the study is matched by location, but the location is one with extensive migration and/or heterogeneity, the association may be spurious; consider awarding fewer points if that is the case.

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.

Study receiving "Contradicts": In this example, the curator originally selected "Review." After discussion with the GCEP, it was determined that "contradicts" was the most appropriate. This study has large sample sizes, and the cases and controls in the study are appropriately matched. However, the case-control comparisons were not statistically significant (showed no difference in odds ratio, no significant p-value and a 95% CI that crosses 1.0). Here we use "contradicts" to convey that the evidence does not support our hypothesis that a relationship exists between a gene and a disease.

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 related to the disease entity of interest should be scored. Validated functional assays should be identified by expert panels or, if they are curator identified, confirmed by expert review.

Figure 8: Experimental Evidence Summary Matrix

Evidence Category	Evidence Type	Suggested Points/		Points Given	Max Score
		Default	Range		
Function	Biochemical Function	**A** 0.5	0-2	**L**	
	Protein Interaction	**B** 0.5	0-2	**M**	**W 2**
	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**	
	Cell culture model	**G** 1	0-2	**R**	
Rescue	Rescue in human	**H** 2	0-4	**S**	**Y 4**
	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 relationship. 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 relationship. 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. <u>Expression:</u> 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, tissue 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. <u>Functional Alteration</u>: 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. <u>Model System</u>: 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. <u>Rescue</u>: 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**? While the default points and point range are the same for human and non-human model organisms, consider awarding more points if the rescue was in a human.
    c. Was the rescue in a **cell culture model** (*i.e.* a cell culture model engineered to express the variant of interest)?
    d. 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
<u>Distinguishing between functional evidence that supports an individual variant and experimental evidence that supports the gene-disease relationship:</u>
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 co-occurs 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 does not support the gene-disease relationship, this data should be flagged as "Review" or "Contradictory" in the GCI, 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 unaffected 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 (page 38, Fig. 7).**

> **NOTE:** Evidence contradicting a single **variant** as causative for the disease does not necessarily rule out the gene-disease relationship.

2. **Minor allele frequency is too high for the disease:** Many diseases have published prevalence, which can often be found in the GeneReviews entry. If ALL 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.

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. 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 ([page 40](#))**. 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 relationship and the mouse model did not have any phenotype, this could be potentially contradictory evidence.

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.

Acknowledging Secondary Contributors
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. See Appendix D for more information and step by step instructions.

1. The total score within the Genetic Evidence Matrix (Figure 2 "U") is listed in Figure 9 column "A".
2. The total score within the Experimental Evidence Matrix (Figure 8 "Z") is listed in Figure 9 column "B".
3. Figure 9 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 9 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 page 44) is highlighted in the final matrix Figure 9 row "E". Rationale should be provided within the designated sections within the GCI.

NOTE: If there is contradictory evidence present, the final summary matrix will display "Yes" in the field called "contradictory evidence.". The conflicting evidence will be weighed and reviewed by the expert panel, and a final classification reached.

Figure 9: 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 relationship	Gene-level experimental evidence that support the gene-disease relationship	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 9 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.

To override, or modify, 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. The GCI will automatically assign the classification based on the number of points documented and tallied in the system. 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 10, 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. Note, the current recommendation from the Gene Curation Working Group (GCWG) is that a classification can only be modified by 1 level from the calculated classification. For example, if the calculated classification is "Moderate" then an expert panel can choose to either reduce the classification to "Limited" or increase the classification to "Strong." This is the case for all classifications except for "Disputed" and "Refuted" which can be chosen regardless of the calculated classification. Please be conscientious of this recommendation in the GCI.

Figure 10: Modifying a Calculated Classification in the GCI

Reasons for Publishing a Gene-Disease Validity Classification

As of October 2023, curators will be required to select a reason for publishing or republishing a classification. These reasons will be utilized when determining the version number for the curation; version numbers will ultimately be available via the clinicalgenome.org website.

If this is the first time a gene-disease relationship is being published, select "New Curation." If the gene-disease relationship is being recurated or otherwise altered (including for administrative reasons), select one or more of the listed options. Please see the GCI help document (https://vci-gci-docs.clinicalgenome.org/vci-gci-docs/gci-help/publishing-an-approved-gene-disease-record) for a full description of each of the choices. A link to this document can also be found in the top right-hand corner of the GCI (under the "Help" dropdown menu).

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.

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

ClinGen encourages all GCEPs to recurate their own classifications. However, in the event that a GCEP is no longer able to remain active, they may transfer their curations to another GCEP to manage the recuration process. These GCEPs are considered "inactive" and will be designated as such on the ClinGen website. In order for a GCEP to transition to inactive status, they must:
- Confer with their respective Clinical Domain Working Group (CDWG) (or, if there is no overarching CDWG, the ClinGen Gene Curation Working Group) to identify an appropriate GCEP for record transfer.
  - You can email the Gene Curation Working Group at genecuration@clinicalgenome.org.
- Work with the new GCEP to determine a plan for record transfer within the GCI, discuss the status of any outstanding curations, transfer any relevant notes, etc.
- Document this plan on the Inactive GCEP form.
  - Coordinators can access this form on the Group and Personnel Management System, or GPM. Please contact the GPM helpdesk if you need assistance (gpm_support@clinicalgenome.org).

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. Bean, L.H., et al., *Diagnostic gene sequencing panels: from design to report-a technical standard of the American College of Medical Genetics and Genomics (ACMG)*. Genet. Med. 22(3): p.453-461
3. Landrum M.J., et al., *ClinVar: Improving Access to Variant Interpretations and Supporting Evidence*. Nucleic Acids Res. 46(D1):D1062-D1067.
4. MacArthur, D.G., et al., *Guidelines for investigating causality of sequence variants in human disease*. Nature. 508(7497): p. 469-76.
5. 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
6. Petrucelli, N., et al., *BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer*. GeneReviews. 1998.
7. 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.
8. Pejaver V., et al.; ClinGen Sequence Variant Interpretation Working Group. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022 Dec 1;109(12):2163-2177.

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 so that it can be used to curate evidence from those sites in the event there is no more specific publication outlining the evidence used. For more instructions on how to enter evidence from databases see page 15. 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 relationships 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 genetic 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.
- **Monarch Disease Ontology (MonDO)**
    - https://www.ebi.ac.uk/ols4
    - 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**
    - https://hpo.jax.org/app/
    - Standardized vocabulary and codes for human phenotypic abnormalities.
- **Monarch Initiative**
    - https://monarchinitiative.org/

        ○   Search for a disease then choose the "phenotypes" tab for a list of related 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**
  - https://www.uniprot.org/
  - Comprehensive resource for protein sequence and functional information.
- **MARRVEL**
  - http://marrvel.org/
  - Resource that aggregates relevant databases including model organism, population, and disease databases.
- **ClinGen Gene Curation FAQ**
  - https://clinicalgenome.org/docs/gene-curation-faq/

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
  - PMID:29165669
- **ClinVar Miner**

- ○ https://clinvarminer.genetics.utah.edu/
  - ○ A web-based tool for filtering and viewing ClinVar data
- **Simple ClinVar**
  - ○ https://simple-clinvar.broadinstitute.org/
  - ○ An interactive web-based tool for exploring and retrieving gene and variant data and summary statistics from ClinVar. Data is not updated on a regular basis.
- **Leiden Open Variation Database (LOVD)**
  - ○ http://www.lovd.nl/3.0/home
  - ○ Listings of variants within human genes and related phenotypes; includes links to locus-specific databases.
- Developmental Brain Disorder Gene Database
  - ○ Geisinger DBD Genes Database (geisingeradmi.org)
  - ○ A curated resource for researchers & clinicians providing genotype and phenotype data from six neurodevelopmental disorders obtained from published literature.

GENE CURATION DATABASE
- **GenCC**
  - ○ https://search.thegencc.org/
  - ○ The GenCC (Gene Curation Coalition) is a global effort to harmonize gene level resources. Submitters submit assertions for gene disease relationships.
  - ○ Curators should use the primary evidence cited in curations (evidence summaries and attached PMIDs), not the assertions of the submitters.

ALLELE FREQUENCIES
- **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 (v2.1.1).

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://apid.dep.usal.es:8080/APID/init.action
  - 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.mmrrc.org/catalog/StrainCatalogSearchForm.php?SourceCollection=KOMP
  - 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 or other database. Some sites may reference if cases have been published in the literature, however many may not.
- **DECIPHER**
  - https://www.deciphergenomics.org/
  - Database that houses over 30,000+ case reports.
  - Seminal paper PMID: 19344873
- **GenomeConnect**
  - **GenomeConnect is the ClinGen patient registry that works to engage individuals in data sharing through its own registry and by working with other**

gene/condition-specific registries. Case-level data is shared with published to ClinVar and includes phenotyping and variants.

- ○ **Search the clinicalgenome.org website for your gene of interest and click into the gene page. If GenomeConnect has submitted variants in that gene with case-level data to ClinVar, that will be displayed as a tab under the gene summary (see red in Figure below). Clicking on that tab will take you to a page that then links to the ClinVar GenomeConnect submissions for that gene. For guidance on searching, view a short video here:** https://m.youtube.com/watch?v=YgQkER2TCz8&list=PLrik3QIJ5Zvu8NCyz0KRIc5_Fhenb9nSw&index=3&pp=iAQB

- ○ Seminal paper PMID: 26178529
- ○ NOTE: Email info@genomeconnect.org to request additional information on participants. GenomeConnect has the ability to recontact participants, and can work with GCEPs to obtain information necessary to support curation.
- ● 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 a definitive relationship with HCM. Therefore, the function of MYH7 is shared with other known genes in the disease of interest. (Default: 0.5 points)
- Example: Biallelic variants in DRAM2 and 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 the

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 similar or 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 semidominant 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 28. 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 28. 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:

A. Family 1

Heterozygous, affected

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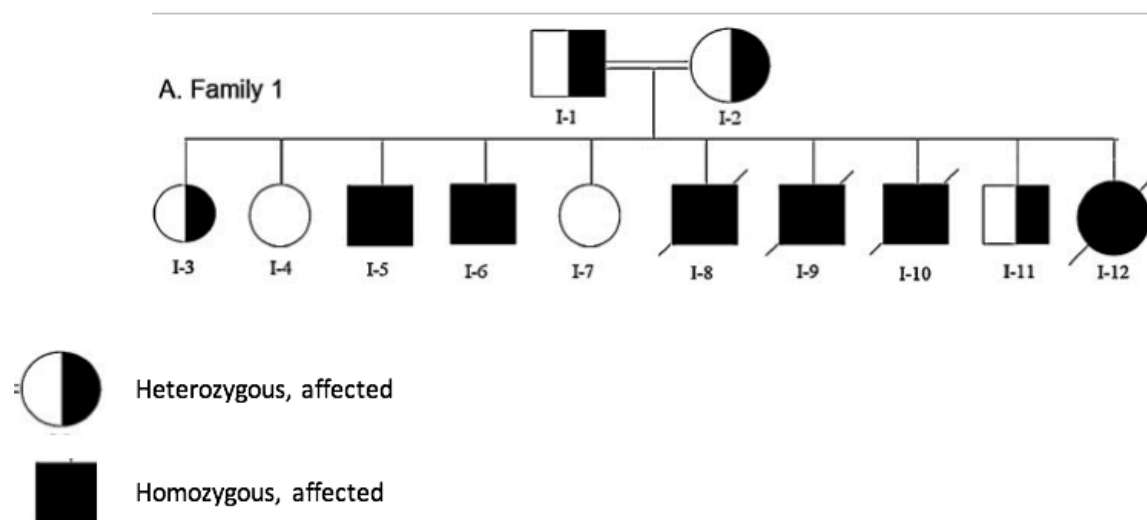
This semidominant family meets the criteria for AR segregation inclusion, as there are 6 affected, genotype positive individuals in the pedigree (I-5, I-6, I-8, I-9, I-10, I-12). Whereas, only 2 segregations are present to an AD MOI, which does not meet the requirement of 4 segregations to include an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

Semidominant Pedigree Example #2:

B. Family 2

Heterozygous, affected

Homozygous, affected
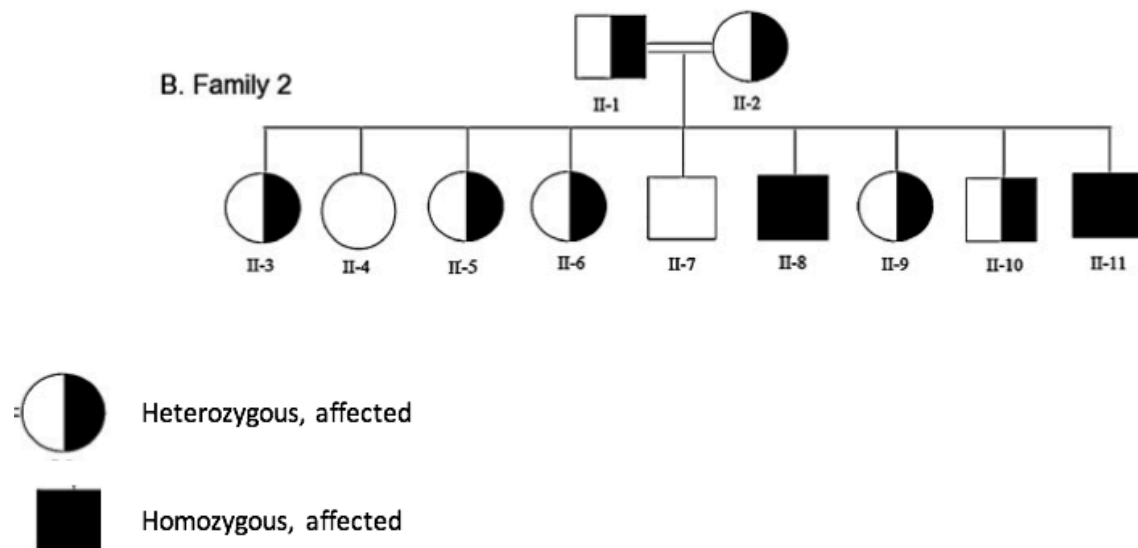
This semidominant family meets the criteria for AD segregation inclusion, as there are 5 segregations among genotype[+]/phenotype[+] individuals (counting from either II-1 or II-2 down to each of the 5 affected children). It does not meet the criteria for AR segregation inclusion, as there are only 2 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:
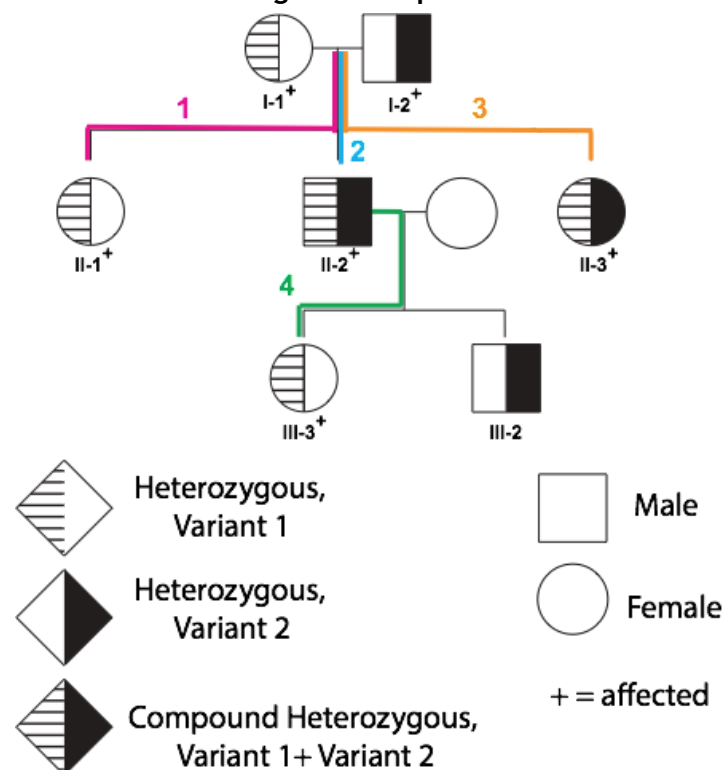
APPENDIX D: ACKNOWLEDGING SECONDARY CONTRIBUTORS OR APPROVERS

For gene curations representing a collaborative effort for a shared gene(s) of interest across multiple expert panels, it is common practice to recognize this effort through the use of the Secondary Contributors function in the GCI.

NOTE: If a GCEP has interest in remarking on a gene curation that has been published by another group for the purposes of a manuscript, and they agree with the published classifications and subsequent data with no further evidence to contribute, the curation can be remarked on in any manuscript as long as the proper attribution of the GCEP that curated the gene-disease relationship is acknowledged.

If your GCEP identifies a gene of shared interest, please reach out to the coordinator of the other GCEP(s) to discuss steps needed for contribution.
- The GCEP that owns the record will typically have to make all necessary edits to the curation record, so coordination among groups is helpful and appreciated. Please work with the GCEP that owns the record to provide help and shared effort to reduce burden of data entry (e.g. provide all curated data to enter for additional cases, provide the final approved evidence summary sentence to copy and paste in the GCI, etc.)
- In general, it is appreciated if these requests are granted as it helps the clinicalgenome.org user identify relevant genes across GCEPs, as some diseases span multiple clinical domains (e.g. syndromic genes).
    - If there is any concern, please reach out to your Clinical Domain Working Group(CDWG) chair(s) and coordinator(s).

Secondary Contributor: For curations in which additional expertise or inclusion of genetic and/or experimental curated data and/or updates to an evidence summary were provided, consider using the Secondary Contributor acknowledgement.

Secondary Approver: For curations in which curated data was provided by an additional GCEP, including reference sources, scoring, and full curation details, such that the secondary GCEP needed to contribute to the final arbitration of the classification for approval, consider using the Secondary Approver acknowledgement.
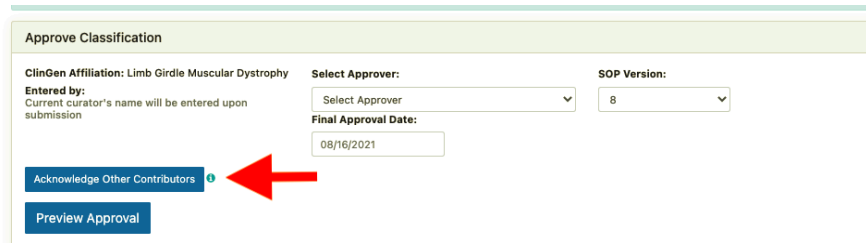
ONLY ONE OF THESE SHOULD BE USED PER CURATION.
- Do not list your own affiliation as one of these. You are acknowledged for your contribution as the owner of the record.

GCI instructions:

Acknowledgement of a secondary contributor(s) or approver(s) happens at the approval stage for any gene-disease clinical validity classification in the Gene Curation Interface (GCI).
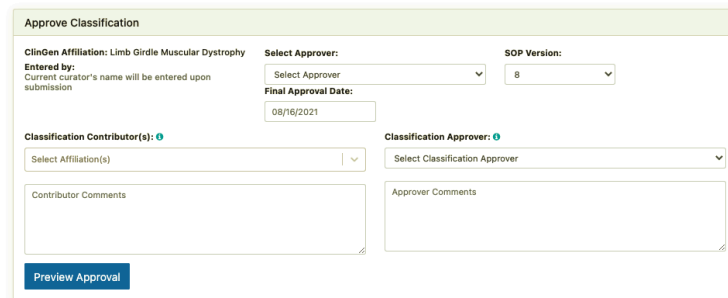
1. At the stage of approving, choose the "Acknowledge Other contributors" button (see red arrow in the figure below).

2. Select from "Classification Contributor" or "Classification Approver". **In general only one of these categories will need to be selected, if only one additional expert panel has contributed.** In other words, you do not need to fill out both, unless you have reason to include more than one EP.

3. Click the down arrow at the right of the "Select affiliation" box (see red arrow below) or begin typing a name into the search box to select the appropriate affiliation(s). More than one expert panel can be selected.

    a. **Note, this list contains GCEPs and VCEPs, therefore it is good practice to check the clinicalgenome.org for the current affiliation name and number (the last two digits listed here corresponded with the URL of the expert panel).**

4. Proceed with the typical steps to complete the approval of a gene curation record which includes selecting the Approver from the affiliation, and clicking "preview approval."

5. Submit the approval and move on to publishing the record. Once published to clinicalgenome.org, the record should display the secondary contributor or approver in the final record (example below).