Gene Clinical Validity Curation Process

Standard Operating Procedure

Version 6
August 2018
The Clinical Genome Resource
Gene Curation Working Group

TABLE OF CONTENTS

BACKGROUND..... 3
REQUIRED COMPONENTS..... 3
OVERVIEW OF GENE CURATION..... 3
GENE CURATION WORKFLOW (Fig.1)..... 5
CLINICAL VALIDITY CLASSIFICATIONS (Fig. 2)..... 6
DECIDING ON THE DISEASE ENTITY..... 6
LITERATURE SEARCH..... 8
GENETIC EVIDENCE..... 10
Case-level data..... 11
Genetic Evidence Summary Matrix (Fig. 3)..... 11
Variant Evidence..... 13
Segregation analysis (Fig. 4-6)..... 15
Case-control data..... 23
Examples (Fig. 7)..... 25
EXPERIMENTAL EVIDENCE..... 26
Experimental Evidence Summary Matrix (Fig. 8)..... 26
Case-level vs Experimental Evidence..... 29
CONTRADICTORY EVIDENCE..... 31
SUMMARY & FINAL MATRIX (Fig. 9)..... 32
REFERENCES..... 34
APPENDIX A..... 35
APPENDIX B..... 38

BACKGROUND:

ClinGen's gene curation process is the method designed to aid in evaluating the strength of a gene-disease relationship based on publicly available evidence. Information about the gene-disease relationship, including genetic, experimental, and contradictory evidence curated from the literature is compiled and used to assign a clinical validity classification per criteria established by the ClinGen Gene Curation Working Group (GCWG)[1]. This protocol details the steps involved in curating a gene-disease relationship and subsequently assigning a clinical validity classification. This curation process is not intended to be a systematic review of all available literature for a given gene or condition, but instead an overview of the most pertinent evidence required to assign the appropriate clinical validity classification for a gene-disease relationship at a given time. While the following protocol provides guidance on the curation process, professional judgment 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 (https://www.clinicalgenome.org/working-groups/gene-curation/ and https://www.clinicalgenome.org/curation-activities/gene-disease-validity/educational-and-training-materials/ or contact clingen@clinicalgenome.org
- The ClinGen Lumping and Splitting guidelines must be consulted to determine the disease entity for curation: https://www.clinicalgenome.org/site/assets/files/9703/lumping_and_splitting_guidelines_gene_curation_final.pdf
- Publication access
- Access to the ClinGen Gene Curation Interface (GCI), https://curation.clinicalgenome.org/ .
  Contact clingen-helpdesk@lists.stanford.edu to set up an account. Please send your preferred email address, preferred display name (first and last name), and any ClinGen affiliations (i.e. the expert panels you are planning to work with). Please cc the coordinator of any expert panels that you will be working with so that they know you have asked for the account and affiliations to be set up).
  For further information see the GCI Help document at https://github.com/ClinGen/clincoded/wiki/GCI-Curation-Help
  The GCI Help document also provides information on inputting curation data into the GCI.

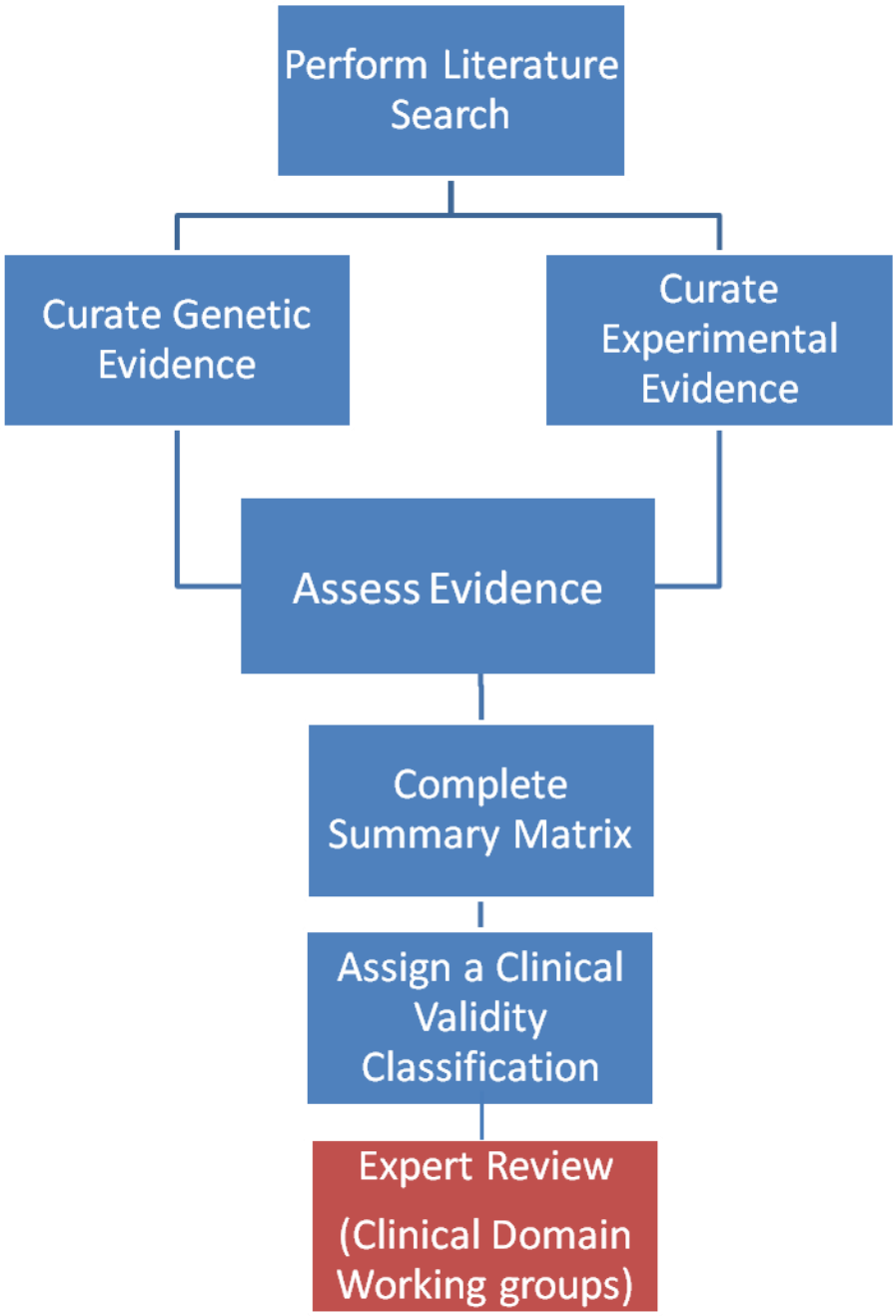
Optional: Microsoft Office (Word, Excel, or Powerpoint to record your data from curation)

OVERVIEW OF GENE CURATION:

The gene curation frame work consists of the following steps.

- **Collection of evidence:** The evidence is collected primarily from published peer-reviewed literature, but can also be present in publicly accessible resources, such as variant databases, which can be used with discretion. Literature searches can be conducted using PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and/or Google Scholar (http://scholar.google.com/) (which has a full-text search feature). Advanced searches are generally more informative.
    - PubMed tutorial: https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/cover.html
    - PubMed presentation to ClinGen Biocurator working group: http://tinyurl.com/ydfd826m
    - Google Scholar search help: (https://scholar.google.com/intl/en/scholar/help.html#searching)
- One need not comprehensively curate all evidence for a gene-disease pair (particularly for "Definitive" associations), but instead focus on curating and evaluating the relevant pieces of evidence described in this protocol.
- **Identifying different evidence types:** The curator needs to identify and curate genetic and experimental evidence separately (details are defined later in "Genetic Evidence" and Experimental Evidence" sections). Genetic evidence is divided into two categories: case-level data and case-control data. Typically studies describing individuals or families with variants in the gene of interest will be scored as case-level data, while studies using statistical analysis to determine the enrichment of variants in case and control groups will be scored as case-control data. The gene-level experimental data used in this framework to assess a gene-disease relationship are *in vitro* and *in vivo* functional studies that implicate the causative role of a gene in disease. These are based on MacArthur and colleagues and described in detail below [2].
- **Assignment of clinical validity classification using gene curation matrix:** Next the curator evaluates the evidence and assigns points to the evidence using the scoring matrices provided below (Fig. 3,8). This information is then summarized and tallied to generate a total score and calculated clinical validity classification, which will be reviewed by a committee of appropriate disease experts. While all evidence will be reviewed by disease experts, curators should flag any evidence that requires further discussion with an expert in the disease area.

Figure 1: Gene Curation Workflow



CLINICAL VALIDITY CLASSIFICATIONS:

The gene curation working group members have developed a method to qualitatively define the "clinical validity" of a gene-disease relationship using a classification scheme based on the strength of evidence that supports or refutes any claimed relationship (Figure 2). This framework allows the "clinical validity" of a gene-disease relationship to be transparently and systematically evaluated. These classifications can then be used to prioritize genes for analysis in various clinical contexts. The suggested minimum criteria needed to obtain a given classification are described for each clinical validity classification. These criteria include both genetic and experimental evidence, which are described below in this document. The default classification for genes without an identified variant in humans is "No Reported Evidence." The level of evidence needed for each supportive gene-disease relationship category builds upon the previous category (e.g. "Limited" builds upon "Moderate"). Gene-disease relationships classified as "Contradictory" likely have evidence supporting as well as opposing the gene-disease association, but are described separately from the classifications for supportive gene-disease relationships.

DECIDING ON THE DISEASE ENTITY:

Many human genes are implicated in more than one disorder. Therefore, prior to starting a curation and entering the details into the Gene Curation Interface (GCI), the curator should be absolutely clear on which disease entity is being curated based on the Lumping and Splitting guidelines (https://www.clinicalgenome.org/site/assets/files/9703/lumping_and_splitting_guidelines_gene_curation_final.pdf). Once a curation is started in the GCI, the only mechanism for changing a disease entity is to contact the GCI Help Desk. If needed, the expert group can give guidance on which disease entity should be curated.

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

Contradictory Evidence	**CONFLICTING EVIDENCE REPORTED**	Although there has been an assertion of a gene-disease association, conflicting evidence for the role of this gene in disease has arisen since the time of the initial report indicating a disease association. Depending on the quantity and quality of evidence disputing the association, the association may be further defined by the following two sub-categories: **1. Disputed** a. Convincing evidence *disputing* a role for this gene in this disease has arisen since the initial report identifying an association between the gene and disease. b. Disputing evidence need not outweigh existing evidence supporting the gene-disease association. **2. Refuted** a. Evidence *refuting* the role of the gene in the specified disease has been reported and significantly outweighs any evidence supporting the role. b. This designation is to be applied at the discretion of clinical domain experts after thorough review of available evidence c. While it is nearly impossible to entirely refute a gene's potential role in disease, this category is to be used when all existing data has been fully refuted leaving the gene with essentially no valid evidence remaining, after an original claim.
NOTES		
[1]Variants that disrupt function and/or have other strong genetic and population data (e.g. *de novo* occurrence, absence in controls, strong linkage to a small genomic interval, etc.) are considered convincing of disease causality in this framework. See "Variant Evidence" on p.13 for more information. [2]Examples of appropriate types of supporting experimental data based on those outlined in MacArthur et al. 2014[2].		

LITERATURE SEARCH:

1. The initial search should be **broad** and **inclusive**. A good way to start is by searching **"gene symbol/name AND disease"** (in some cases it may be sufficient to search for the gene name/symbol alone). Ensure that you have looked up gene symbol/name alternatives before you search.
    a. Check HGNC (www.genenames.org/) for old gene symbols and aliases
    b. NCBI Gene (www.ncbi.nlm.nih.gov/gene) also lists gene aliases
    c. NOT all search results will be relevant, thus it is important to examine the search results for pertinent information
2. Curating primary literature is encouraged, but if a gene-disease pair has abundant information (i.e. >50 relevant results returned in a search), review articles may be sufficient. To find reviews, search PubMed with **"gene AND disease AND (review [Publication Type] OR "review literature as topic"[MeSH Terms]).**
    a. Curation may occur from that publication **ONLY** when sufficient details are included in the review article.

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

3. Additional searches are often necessary to identify sufficient gene level experimental evidence. Note that additional gene level experimental evidence may exist in publications **BEFORE** the gene:disease association was first made.
   a. Search PubMed for experimental data (Examples below)
      1. [gene] AND [function] e.g. "KCNQ1" AND "potassium channel"
      2. [protein] AND [function] e.g. "neurofibromin" AND "tumor suppressor"
      3. [gene] AND [animal] e.g. "ACTN2" and "mouse OR zebrafish OR xenopus OR drosophila"
   b. Additional information may also be available in OMIM (www.OMIM.org) in the **"Gene function"** or **"Biochemical Features"** or **"Animal Model"** sections.
   c. GeneReviews (http://www.ncbi.nlm.nih.gov/books/NBK1116/) often has information in the **"Molecular Genetics"** section of the disease entries that may be useful.
   d. Other databases such as UniProt (www.uniprot.org/), MGI (www.informatics.jax.org/), etc. may also be useful, provided that primary references are given that can be curated. For a list of databases that may be helpful for the curation process, see Appendix A.
   e. GeneRIFs (Gene References Into Functions), within NCBI Gene (https://www.ncbi.nlm.nih.gov/gene), lists article links that summarize experimental evidence for a given gene. The link itself leads to an article in PubMed and can serve as an additional source for experimental evidence.

4. An additional component of the curation process is to determine if the original gene-disease association has been replicated; therefore, it is critical to find the **original paper** with the proposed relationship. OMIM and GeneReviews often cite the first publication and should be cross-referenced. Additionally, a recent review article may be helpful in ruling out any contradictory evidence that may have been reported since the original publication.
   a. The **"Allelic Variants"** section of OMIM and the **"Molecular Genetics > Pathogenic allelic variants"** section of GeneReviews may have relevant information.
   b. Be sure to extract information from the **original publication**, NOT directly from these websites.

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

GENETIC EVIDENCE

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

Genetic Evidence Summary Matrix

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

Case-Level Data

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

1) The authors provide sufficient evidence to document the diagnosis. 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 supplement information captured by HPO terms, or in the event 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.

2) The variant identified in that individual has some indication of a potential role in disease (e.g. impact on gene function, recurrence in affected individuals, etc.). Each case may be given points for both variant evidence (see below for details on interpretation) and segregation evidence (see pp. 15-22 for details).

Figure 3: Genetic Evidence Matrix

GENETIC EVIDENCE SUMMARY								
	Evidence Type		Case Information (type of variant identified in proband)	Suggested Points/Case		Points Given	Max Score
				Default	Range		
Case-Level Data	Variant Evidence	Autosomal Dominant OR X-Linked Disorder A	Variant is *de novo*	C 2	0-3	H	M 12
			Predicted or proven null variant	D 1.5	0-2	I	N 10
			Other variant type (not predicted/proven null) with some evidence of gene impact	E 0.5	0-1.5	J	O 7
		Autosomal Recessive Disorder B	Two variants in *trans* and at least one *de novo* or predicted/proven null variant	F 2	0-3	K	P 12
			Two variants (not predicted/proven null) with some evidence of gene impact in *trans*	G 1	0-1.5	L	
	Segregation Evidence		Evidence of segregation in one or more families	Sequencing Method			
			Total LOD Score	Candidate Gene Sequencing	Exome/ Genome or all genes sequenced in linkage region	0-3	Q	R 3
			2-2.99	0.5	1			
			3-4.99	1	2			
			≥5	1.5	3			
Case-Control Data	Case-Control Study Type		Case-Control Quality Criteria	Suggested Points/Study		Points Given	Max Score
	Single Variant Analysis		• Variant Detection Methodology • Power • Bias and Confounding Factors • Statistical Significance	0-6		S	T 12
	Aggregate Variant Analysis			0-6			
					TOTAL ALLOWABLE POINTS for Genetic Evidence		U 12

General Notes for variant scoring:

1. When curating an autosomal dominant disease or an X-linked disorder consider the evidence types in row "A". If you are curating an autosomal recessive disease, consider the evidence types in row "B". In X-linked disorders, affected probands will often be hemizygous males and/or manifesting heterozygous females. Recognizing that there can be rare cases of females affected by X-linked recessive disorders (due to chromosomal aneuploidy, skewed X inactivation, or homozygosity for a sequence variant), or males who carry an X-linked variant but are unaffected or mildly affected (due to Klinefelter syndrome, 47, XXY) evaluators must be aware of the nuances of interpretation of individual cases and X-linked pedigrees. Points can be assigned at the discretion of the expert reviewer taking into account the available evidence. Furthermore, there are known cases of female carriers of X-linked recessive conditions manifesting symptoms that are milder or later in onset compared to males, and scoring of genetic evidence in these examples should be subject to expert review with regard to the assigned gene/disease/inheritance combination.
2. Computational scores (such as conservation scores, constraint scores, *in silico* prediction tools, variation intolerance scores, etc) are often disease and context-dependent and should not be considered as strong pieces of evidence for variant pathogenicity. However, they can be recorded during curation and used as supporting evidence for variant scoring to be confirmed by expert review.
3. For a variant to be considered potentially disease-causing, its frequency in the general population should be consistent with phenotype frequency, inheritance pattern, disease penetrance, and disease mechanism (if known). These pieces of information can often be located in the literature (See "Literature Search" p. 9), 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; http://gnomad.broadinstitute.org) 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. Gene curation committees may find it helpful to set a maximum 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.
4. For each case information category, a suggested number of points per case is provided. However, the points may be altered, within a defined range, to account for the strength of evidence available to indicate that a variant is deleterious (see Figure 3). Within each range, the curator may choose one of the following scores: 0.1, 0.25, 0.5, followed by 0.5 point increments up to the maximum possible score

for that category. However, the curator should always document reasons for any deviation in suggested scores for expert review.

5. When scoring variants for autosomal recessive disorders in individuals who are compound heterozygotes, there should be some evidence to suggest that the variants are in *trans* in order to be scored. For example, for an individual who is compound heterozygous for two variants in the gene of interest, both parents should be tested to show that the variants are in *trans*. Molecular methods showing that variants are in *trans* are also acceptable. For individuals who appear to be homozygous for a variant, testing of the parents is not required in order to count the case.

Variant Evidence:

De novo variants:

- These can be any type of variant, but should be given points depending on statistical expectation of *de novo* variation in the gene in question, if known. In some cases, this can be found in the literature and should be noted if found (See "literature search" p. 9). However, the curator may also leave this to be supplied by experts during curation review.
- There are two scenarios in which points can be awarded for a "*de novo*" variant. For both of these scenarios, the same points range should be used (see Figure 3):
    a. 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. Consider awarding default points to null variants (e.g. nonsense, frameshift, canonical splice site) that appear to be *de novo* based on testing parents for the variant, but award fewer points to missense variants and small, in-frame deletions. The scores can be increased if the maternity and paternity of the proband are confirmed e.g. by short tandem repeat analysis or trio whole exome sequencing (WES). For example, a case with a missense variant could receive default points if maternity and paternity are confirmed. Additional points can be added for any variant if functional evidence supports a deleterious impact for the variant.
    b. One of the parents of an affected individual is found to have to 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.

- Sum up the number of points. The suggested points per case can be found in column "C" (autosomal dominant or X-linked disorders) and "F" (autosomal recessive). Total up all of the variant evidence points and place them in "H" or "K" respectively. If you are using the GCI, it will add the points for you.

Predicted or proven null variants:

- This category includes nonsense, frameshift, canonical +/-1 or 2 splice site variants, single or multi-exon deletions, whole gene deletions, etc. Other variant types, such as missense, may be included in this category if there is sufficient evidence for complete loss of function.
- Assign fewer points if there is alternative splicing or if the null variant is near the C terminus and/or nonsense medicated 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 assigning fewer points if a gene product is still made, albeit altered. For example, cDNA analysis and Western blot for 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.
- Gene constraint scores can be helpful when assessing disease mechanism. For example, the disease mechanism could be assumed to be loss of function (LOF) if the gene is LOF constrained. Constraint scores can be found by searching the gene in ExAC (exac.broadinstitute.org) and viewing the "constraint metric" at the top right of the page. The closer the probability of LOF intolerance (pLI) is to 1, the more LOF-constrained the gene. However, constraint scores must be interpreted in the context of the gene and disease in question. For example, if the gene is associated with multiple diseases, LOF constraint could be associated with a disease other than the one being curated. In addition, genes associated with severe, pediatric-onset disorders may appear to be more constrained than adult-onset conditions where overall fitness is not impacted.
- 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.
- Sum up the number of points. The suggested points per case can be found in column "D" (autosomal dominant or X-linked) and "F" (autosomal recessive). Total up all of the variant evidence points and place them in "I" or "K" respectively. If you are using the GCI, it will add the points for you.

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 have a dominant-negative impact.
- Some functional impact to the gene product must be demonstrated for the case to be given default points. Examples of functional impact include reduced activity of an enzyme in cells expressing a variant in that gene, or reduced expression of a gene product in cells from an individual with a variant(s) in the gene. Impact based on functional validation can score 0.5 or above (up to 1.5/case) depending on the validation quality and disease relevance of the functional assay.
- *In silico* predictions do not provide strong evidence for functional impact and therefore, impact based on *in silico* predictions only would score less than the default 0.5 points. It may be appropriate to award default points if in-depth *in silico* modeling studies e.g. based on impact on 3D structure, have been done, but this requires discussion with an expert.
- Sum up the number of points. The suggested points per case can be found in column "E" (autosomal dominant or X-linked) and "G" (autosomal recessive). Total up all of the variant evidence points and place them in "J" or "L" respectively. If you are using the GCI, it wil add the points for you.

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. If the points given above for the summary matrix exceed the max score, use the Max score found in "M-P" for the summary matrix.

Segregation Analysis:

The use of segregation studies in which family members are genotyped to determine if a variant co-segregates with disease can be a powerful piece of evidence to support 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 that 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:

<u>If a LOD score has been calculated by the authors of a paper:</u>
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. Fill out the "Segregation evidence" portion of the matrix. The number of points should be recorded in "Q".

<u>If a LOD score has NOT been calculated by the authors of a paper:</u>
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 non-explainable non-segregations (for example, a genotype-/phenotype+ individual in a family affected by a disorder with no known phenocopies). Families with non-explainable non-segregations should not be used in LOD score calculations.

If any of the previous conditions are not met, do not use the formula(s) below to estimate a LOD score.

To be conservative in our simplified LOD score estimations, for autosomal dominant or X-linked disorders, only affected individuals (genotype+/phenotype+ individuals) or obligate carriers (regardless of phenotype) should be included in calculations. An

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

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

Within a given gene-disease curation, if more than one family meets the criteria above for scoring segregation information, sum their LOD scores to assign a final segregation score (using the tables in Figures 4 or 5). For example, if Family A has an estimated LOD score of 1.2 and Family B has an estimated LOD score of 1.8, the summed LOD score will equal 3. See the discussion on sequencing method below for guidance on assigning segregation points to the LOD score.

Expert reviewers may choose to specify the most appropriate way to approach segregation scoring within their disease domain, including enacting more formal, rigorous LOD score calculations.

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

For dominant/X-linked diseases:

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

Figure 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

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 estimate the LOD score. For example, if one parent is affected with an autosomal recessive condition and the other is a carrier, replace "0.25" and "0.75" with 0.5.

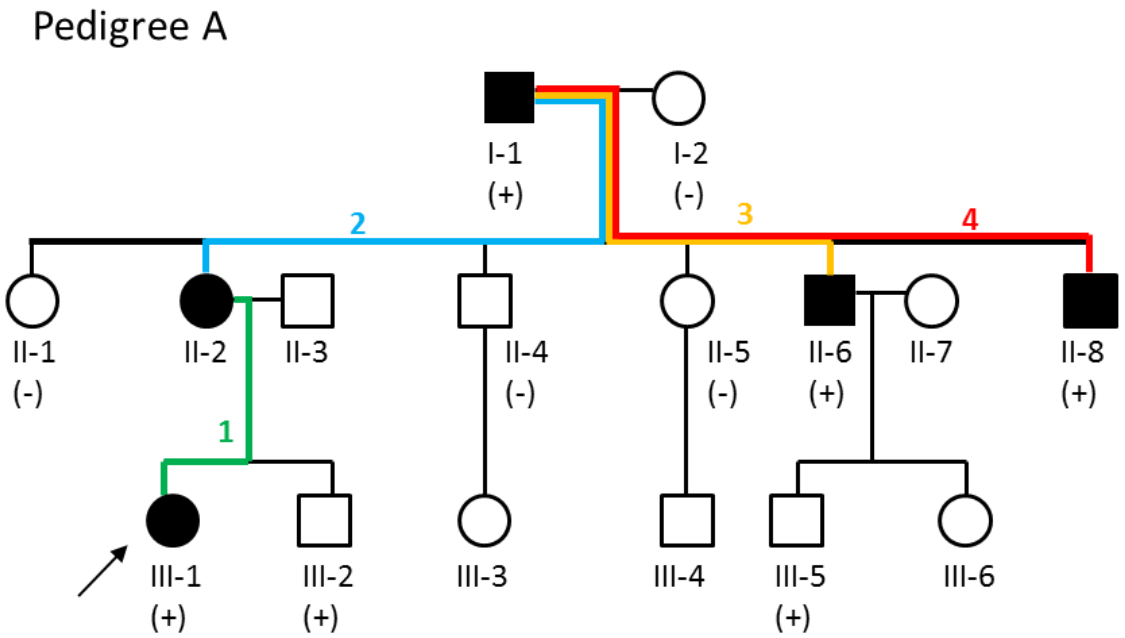
Figure 5: Recessive estimated LOD (eLOD) score table

		Unaffeteds										
		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

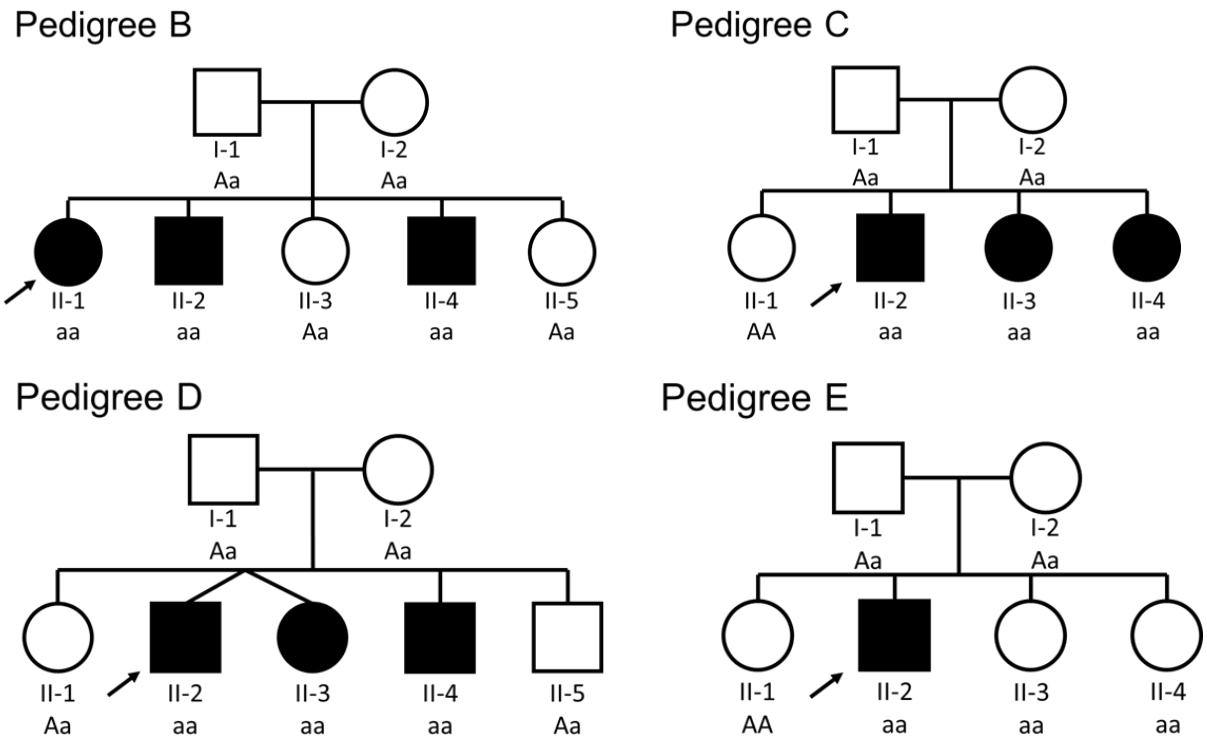
Counting Segregations

1. In general, the number of segregations in the family will be the number of affected individuals minus one, the proband, to account for the proband's genotype phase being unknown. However, as there may be exceptions, segregations should be counted carefully, as outlined below. For example, **pedigree A** shows a family with hypertrophic cardiomyopathy.
   a. There are **four segregations** that can be counted beginning at the proband. This includes the mother (II-2) who is an obligate carrier and can be assumed to be genotype-positive even though she was not tested. Using **four segregations** in the formula above results in an estimated e**LOD score of 1.2**.
   b. 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.



2. 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.
3. 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.



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 in the linkage interval and those that do cannot rule out the effects of thousands of other variants in the interval. **Thus, it is critical for a curator to evaluate the methods used to identify candidate variants.**
Some publications more thoroughly investigate the genes and variants in a linkage interval than others. Accordingly, more points are awarded for segregation evidence in cases where whole exome/genome sequencing was performed or if the entire linkage interval was sequenced. These methods provide more convincing evidence than a candidate gene approach in which only one or a handful of genes in a linkage region are sequenced. See figure 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 taken into account 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 pts
3 - 4.99	1 pts	2 pts
(>/=) 5	1.5 pts	3 pts

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

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

Where:
A = The sum of all LOD scores for candidate gene approach.
B = The sum of all LOD scores for exome sequencing, genome sequencing, and all genes in candidate region sequenced.
C = Points assigned if total LOD had been obtained only by a candidate gene approach (see Figure 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 (link to calculator). The points are rounded to the nearest 0.1. 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 award 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 awarded:

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

Additional logic
While the formula is appropriate for use in the majority of scenarios, there are some situations for which additional logic must be used. This logic, which is coded into the GCI and the calculator, is illustrated by the following example. For Family 1, an estimated LOD score of 3.1 is obtained from a study involving WES. For Family 2, a candidate gene analysis was performed, and a LOD of 1.2 was estimated. In this scenario, 2 points could be awarded for Family 1 alone (as the LOD is between 3-4.99; see Figure 6). The total LOD score for Family 1 and Family 2 is 4.3. If the second piece of evidence were to be taken into account, the points would be reduced to 1.8. In this situation, the formula should not be applied and the maximum number of points (i.e. 2) should be given.

We recognize that the methods in each publication vary, therefore the suggested points in figure 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**.
    a. **Single variant analysis studies** are those in which individual variants are evaluated for statistical enrichment in cases compared to controls. More than one variant may be analyzed, but the variants should be independently assessed with appropriate statistical correction for multiple testing. **For example**, if a study identifies 2 different variants in *MYH7* within a cohort of hypertrophic cardiomyopathy cases, but tests the number of hypertrophic cardiomyopathy cases and unaffected controls that contain only one of the variants and provides a statistic for that variant alone, then the study is classified as a single variant analysis. Similarly, if the same study tests for enrichment of the second variant in the cases and controls and provides a separate statistic for the second variant, this also is a single variant analysis. Often, authors will indicate this either in the article text or in a table of variants.
    b. **Aggregate variant analysis** studies are those in which the statistical enrichment of two or more variants as an aggregate is assessed in cases compared to controls. This comparison could be accomplished by genotyping specific variants or by sequencing the entire gene. For example, if a study identifies 2 different variants in *MYH7,* and then statistically tests the enrichment of both the variants in hypertrophic cardiomyopathy cases over unaffected controls, an aggregate variant analysis was conducted.

2. Case-control studies should be assigned points at the discretion of expert opinion based on the overall quality of each study. Assign each study a number of points between 0-6, then sum the points given to all studies, and fill in "S". NOTE: If the points given exceed the max score, use the Max score found in "T" for the summary matrix.

3. The quality of each case-control study should be evaluated using the following criteria in aggregate:
   a. **Variant Detection Methodology:** Cases and controls should ideally be analyzed using methods with equivalent analytical performance (e.g. equivalent genotype methods, sufficient and equivalent depth and quality of sequencing coverage).
   b. **Power:** The study should analyze a number of cases and controls given the prevalence of the disease, the allele frequency, and the expected effect size in question to provide appropriate statistical power to detect an association. (**NOTE:** The curator is NOT expected to perform power calculations, but to record the information listed in this section for expert review.)
   c. **Bias and Confounding factors:** The manner in which cases and controls were selected for participation and the degree of case-control matching may impact the outcome of the study. The following are some factors that should be considered:
      i. Are there systematic differences between individuals selected for study and individuals not selected for study (i.e. do the cases and controls differ in variables other than genotype)?
      ii. Are the cases and controls matched by demographic information (e.g., age, ethnicity, location of recruitment, etc.)? Are the cases and controls matched for genetic ancestry, if not did investigators account for genetic ancestry in the analysis?
      iii. Have the cases and controls been equivalently evaluated for presence or absence of a phenotype, and/or family history of disease?
   d. **Statistical Significance**: The level of statistical significance should be weighed carefully.
      i. When an odds ratio (OR) is presented, its magnitude should be consistent with a monogenic disease etiology.
      ii. When p-values or 95% confidence intervals (CI) are presented for the OR, the strength of the statistical association can be weighed in the final points assigned.
      iii. Factors, such as multiple testing, that might impact that interpretation of uncorrected p-values and CIs should be considered when assigning points.

Figure 7: Case-control Genetic Evidence Examples

Detailed explanations for assigned points are provided below the table.

CASE-CONTROL DATA						
Points	**Power**	**Bias/ Confounding**	**Detection Method**	**Statistical Significance**	**Study Type**	**Points (0-6/ study)**
Author A 2015 (Max score)	Breast cancer cases: 100/12,000 Controls: 7/4,500	Matched by age, ethnicity, and location	Cases & controls genotyped for c.1439delA in gene *W*	OR: 5.4 [95% CI: 2.5-11.6; *P* < 0.0001]	Single Variant	6
Author B 2005 (Intermediate score)	HCM Cases: 13/200 Controls: 20/900	Matched by location, but not age or ethnicity	Cases & controls genotyped for p.Arg682Gln in gene *X*	Fisher's exact test *P* = 0.004	Single Variant	4
Author C 2011 (Low score)	Ovarian cancer cases: 11/1,500 Controls: 3/2,000	Matched by ethnicity. Controls from population database (e.g. ExAC)	Cases: sequenced Gene *Y* and counted all cases with null variants. Controls: total individuals from population database with null variants in gene *Y*.	OR of all variants in aggregate: 4.9 (CI: 1.4-17.7; *P* =0.015)	Aggregate analysis	2
Author D 2009 (No case-control score)	Colorectal cancer cases: 11/1,500 Controls: 3/2,000	Matched by ethnicity. Controls from population database (e.g. ExAC)	Cases: sequenced gene *Z* and identified 11 variants in 11 cases. Controls: total individuals from a population database with that were genotyped for the 11 variants identified in controls.	OR of p.Lys342: 4.9 (CI: 1.4-17.7; *P* =0.015)	Not applicable	0

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

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

Study receiving low score (2 points): This study is considered an aggregate analysis since the statistical test analyzed the variants in aggregate across all cases and controls. This study can be assigned 2 points because a population database was used rather than

appropriately-matched controls (i.e. the study is not matched demographically) and the p-value is not very significant. A population database could be used as controls for 2 reasons:

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

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

NOTE: Point totals should NOT exceed the max score. If the totals from "H-Q" exceed the max score, use the max score found in "U" for the genetic evidence portion of the summary matrix. If you are using the GCI, it will add the points for you.

EXPERIMENTAL EVIDENCE

Figure 8: Experimental Evidence Summary Matrix

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

NOTE: Validated functional assays should be identified by expert panels or, if they are curator identified, confirmed by expert review.

Identify the experimental evidence type and assign points according to the following criteria.
For further information and examples see the "Variant evidence vs experimental evidence" section (p28-29) and Appendix B.

1. Biochemical Function: Evidence showing the gene product performs a **biochemical function:** (A) shared with other known genes in the disease of interest, or (B) consistent with the phenotype. NOTE: The biochemical function of both gene products must have been proven experimentally, and not just predicted. When awarding points in this evidence category, the other known gene(s) should have compelling evidence to support the gene-disease association. Consider increasing points based on the strength of the evidence and number of other proteins with the same function that are involved in the same disease. The suggested points/evidence can be found in column "A". Total up all of the experimental points and place them in the points given section found in "L".
2. Protein Interaction: Evidence showing the gene product **interacts** with **proteins previously implicated** in the disease of interest. Typical examples of this data include, but are not limited to: Physical interaction via Yeast-2-Hybrid (Y2H), co-immunoprecipitation (coIP), etc. NOTE: The interaction of the gene products must have been proven experimentally, and not just predicted. Proteins previously implicated in the disease of interest should have compelling evidence to support the gene-disease association. Note: Some studies provide evidence that a variant in the gene of interest disrupts the interaction of the gene product with another protein. In these cases, the positive control, showing interaction between the two wild type proteins, can be counted as evidence of protein interaction. Points can also be awarded to case-level (variant) evidence or functional alteration for the variant disrupting the interaction. The suggested points/evidence can be found in column "B". Total up all of the experimental points and place them in the points given section found in "M".
3. Expression: Summarize evidence showing the gene is expressed in **tissues relevant to the disease of interest** and/or is **altered in expression in patients** who have the disease. Typical examples of this data type are methods to detect a) RNA transcripts (RNAseq, microarrays, qPCR, qRT-PCR, Real-Time PCR), b) protein expression (western blot, Immunohistochemistry). 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.

The suggested points per evidence can be found in column "C". Tally all of the experimental points and place them in the points given section found in "N". **NOTE:** If the sum of all biochemical function, protein interaction, and expression points exceeds the max score of 2, use the Max score found in "W" of the experimental evidence summary matrix.

4. Functional Alteration: Evidence showing that cultured cells, in which the function of the gene has been disrupted, have a phenotype that is consistent with the human disease process. Examples include experiments involving expression of a genetic variant, gene knock-down, overexpression, etc. Divide the evidence according to the following subtypes:
    a. Was the experiment conducted in **patient cells**? The suggested points/evidence can be found in column "D". Total up all of the experimental points and place them in the points given section found in "O".
    b. Was the experiment conducted in **non-patient cells**? The suggested points/evidence can be found in column "E". Total up all of the experimental points and place them in the points given section found in "P".

    NOTE: If the sum of all functional alteration points exceeds the max score of 2, use the Max score found in "X" of the experimental evidence summary matrix.
5. Model System: A **non-human model organism** or **cell culture model** with a disrupted copy of the gene shows a phenotype consistent with the human disease state. Note: Cell culture models should recapitulate the features of the diseased tissue e.g. engineered heart tissue, or cultured brain slices. These results should be summarized accordingly:
    a. Was the gene disruption in a **non-human model organism**? The suggested points/evidence can be found in column "F". Total up all of the experimental points and place them in the points given section found in "Q".
    b. Was the gene disrupted in a **cell culture model**? The suggested points/evidence can be found in column "G". Total up all of the experimental points and place them in the points given section found in "R"
6. Rescue: Summarize evidence showing that the **phenotype in humans** (i.e. patients with the condition), **non-human model organisms, cell culture models,** or **patient cells** can be rescued. If the phenotype is caused by loss of function, summarize evidence showing that the phenotype can be rescued by exogenous wild-type gene, gene product, or targeted gene editing. If the phenotype is caused by a gain of function variant, summarize the evidence

showing that a treatment which specifically blocks the action of the variant (e.g. siRNA, antibody, targeted gene editing) rescues the phenotype. These results should be recorded accordingly:

a. Was the rescue in a **human?** For example, successful enzyme replacement therapy for a lysosomal storage disease. The suggested points/evidence can be found in column "H". Total up all of the experimental points and place them in the points given section found in "S".
b. Was the rescue in a **non-human model organism**? The suggested points/evidence can be found in column "I".
   Note: While the default points and range of points are the same for human and non-human model organism, consider awarding more points if the rescue was in a human. Total up all of the experimental points and place them in the points given section found in "T".
c. Was the rescue in a **cell culture model** (*i.e.* a cell culture model engineered to express the variant of interest). The suggested points/evidence can be found in column "J". Total up all of the experimental points and place them in the points given section found in "U".
d. Was the rescue in **patient cells**? The suggested points/evidence can be found in column "K". Total up all of the experimental points and place them in the points given section found in "V".
   NOTE: If the sum of all models and rescue points exceeds the max score of 4, use the Max score found in "Y" of the experimental evidence summary matrix.

Tally the total number of experimental evidence points from Rows "W-Y" and enter them on Row "Z". If you are using the GCI, it will add the points for you. **NOTE:** If the total experimental evidence points exceed the max score, use the Max score of 6 points for the summary matrix. **Please prioritize curating genetic evidence over experimental evidence to reach a definitive score.**

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

Case level variant evidence vs experimental evidence

Not all 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 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. These decisions should be discussed with experts in the disease area, if needed.

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

- Any model organism with a variant initially identified in a human with the disorder.

CONTRADICTORY EVIDENCE

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

1. **Case-control data is not significant:** As case-control studies evaluate variants in healthy vs affected individuals, if there is no statistically significant difference in the variants between these groups, this should be marked as potentially contradictory evidence for expert review. **See case-control examples above (p.25, Fig. 7) NOTE:** Evidence contradicting a single **variant** as causative for the disease does not necessarily rule out the gene:disease relationship.
2. **Minor allele frequency is too high for the disease:** Many diseases have published prevalence, which can often be found in the GeneReviews entry. If ALL minor alleles in a gene are present in a specific population or the general population (ExAC, ESP, 1000Genomes) at a frequency that is higher than what is estimated for the disease, this could suggest lack of gene-disease relationship and should be marked as potentially contradictory evidence for expert review. **For example,** Adams-Oliver syndrome is an autosomal dominant disease and has a prevalence of 0.44 in 100,000 (4.4e-6) live births. If a new gene were being curated for this disease and supposedly pathogenic variants were identified with an allele frequency in ExAC (or gnomAD) of 0.4882, this could be potentially contradictory evidence. **NOTE:** Evidence contradicting a single **variant** as causative for the disease does not necessarily rule out the gene:disease relationship. Additionally, disease prevalence can vary in different populations, so read the GeneReviews entry thoroughly and keep demographic information in mind during this evaluation.
3. **The gene-disease relationship cannot be replicated:** One measure of a gene-disease relationship is its replication both over time and across multiple studies and disease cohorts. If a study could not identify any variants in the gene being curated in an affected population that was negative for other known causes of the disease, this could be considered potentially contradictory evidence and

should be marked for expert review. **However,** when assigning this designation, a curator 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 association. In any case, if a curator suspects that any evidence supports a lack of gene-disease association, it should be marked for expert review.

4. **Non-segregations:** Non-segregations should be considered carefully, as age-dependent penetrance and phenotyping of relatives could have an impact on the number of apparent non-segregations within a family. Thus, the age of unaffected variant carriers should be of similar age to the affected variant carriers. If a curator suspects non-segregations, these should be noted for expert review.
5. **Non-supporting functional evidence:** The types of different experimental evidence are detailed in the **"Experimental Evidence" Section (p. 26)**. If any of this experimental evidence suggests that variants, although found in humans, do not affect function or that the function is not consistent with the established disease mechanism, this evidence should be marked as potentially contradictory evidence for expert review. **For example**, if a gene were being curated for a disease association and the mouse model did not have any phenotype, this could be potentially contradictory evidence.

SUMMARY AND FINAL MATRIX

A summary matrix was designed to generate a “provisional” clinical validity assessment using a point system consistent with the qualitative descriptions of each classification. This final gene curation matrix and instructions for filling it out can be found below. **If you are using the GCI, it will add the points for you and generate a PDF summary of the evidence and PMIDs used.**

1. Fill in the "Gene/Disease Pair" at the top of the matrix (Figure 9).
2. Enter the score calculated for Genetic Evidence Matrix (Fig. 3 p. 11) in row "A".
3. Enter the score calculated for Experimental Evidence Matrix (Fig. 8 p. 26) in row "B".
4. The sum of A and B is entered in row "C".
5. Refer to the publication date of the original publication of the gene-disease relationship and consider all other literature to complete row "D":
   a. YES if > 3 years have passed since the original publication AND there are >2 publications about the gene-disease relationship

b. NO if >3 years have passed, BUT not >2 publications
c. NO if < 3 years have passed

6. If there is valid contradictory evidence (see p. 31), compile this and briefly describe it (including the PubMed ID numbers) in row "E".
7. Choose the clinical validity classification associated with the value of the total points/replication over time (Row C, Row D) and complete row "F".
   NOTE: No matter the score, if there is contradictory evidence present, the curator classification must be listed as "Conflicting Evidence reported". The conflicting evidence will be weighed and reviewed by a domain expert.
8. When the gene-disease curation is reviewed by an expert, the expert will fill out the final classification in row "G".

Figure 9: Clinical Validity Summary Matrix

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

References:

1. **Strande, N.T., et al., *Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource*. Am J Hum Genet. 100(6): p. 895-906.**
2. **MacArthur, D.G., et al., *Guidelines for investigating causality of sequence variants in human disease*. Nature. 508(7497): p. 469-76.**

APPENDIX A: USEFUL WEBSITES FOR CLINGEN GENE CURATORS

The following websites are free and publically 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.
If there are additional websites that you think curators should be aware of, please contact Jenny Goldstein (jennifer.goldstein@unc.edu)

LITERATURE SEARCHES

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

REVIEWS/DISEASE ENTITIES

- **Online Mendelian Inheritance in Man (OMIM)**
  - http://www.ncbi.nlm.nih.gov/omim
    A comprehensive compendium of human genes and phenotypes that is updated regularly. Summaries of gene-disease associations and references to primary literature can be found here.
- **GeneReviews**
  - http://www.ncbi.nlm.nih.gov/books/NBK1116/
    Provides clinically relevant information for hundreds of different inherited conditions. The "Molecular Genetics" section of each entry may be useful for information on common variants for a gene. The "Establishing the Diagnosis" section typically contains a summary of the genetic testing options, including the different genes involved and proportion of cases caused by variants in each gene.
- **Monarch Disease Ontology (MonDO)**
  - https://www.ebi.ac.uk/ols/ontologies/mondo
    Human disease ontology merging information from multiple disease resources.
- **ORPHANET**
  - http://www.orpha.net
    Online inventory of human diseases.

PHENOTYPES

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

GENES AND GENE PRODUCTS

- **HUGO Gene Nomenclature Committee (HGNC)**
  - http://www.genenames.org
    An online repository of approved gene nomenclature.

- **National Center for Biotechnology Information (NCBI) gene**
  - http://www.ncbi.nlm.nih.gov/gene
    Integrates information from a wide range of species. Includes gene nomenclature, reference sequences, maps, expression, protein interactions, pathways, variations, phenotypes, functional evidence (in GeneRIFs) links to locus-specific resources.

GENES AND GENE PRODUCTS

- **Ensembl**
  - http://www.ensembl.org/index.html
    Nomenclature, splice variants, references sequences, maps, variants, expression, comparative genomics, ontologies, and function.
- **UCSC Genome Browser**
  - https://genome.ucsc.edu/
    Genome browser with access to genome sequence data from a range of species.
- **UniProt**
  - www.uniprot/org
    Comprehensive resource for protein sequence and functional information.

VARIANT DATABASES

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

ALLELE FREQUENCIES

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

GENE EXPRESSION

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

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

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.
- **Agile Protein Interactomes DataServer (APID)**
  - http://cicblade.dep.usal.es:8080/APID/init.action#tabr2
    Comprehensive collection of protein interactions from over 400 organisms.
- **STRING database**
  - http://string-db.org/
    Database of known and predicted protein interactions.

MOUSE MODELS

- **Mouse Genome Informatics**
  - https://www.jax.org/jax-mice-and-services
    Database of laboratory mice, providing integrated genetic, genomic, and biological data.
- **Knockout Mouse Project (KOMP)**
  - https://www.komp.org/
    Initiative to generate a public resource of **mouse** embryonic stem cells containing a null mutation in every gene in the **mouse** genome.

APPENDIX B: EXPERIMENTAL EVIDENCE EXAMPLES

FUNCTION

Biochemical function:

- ***Example: MYH7 and hypertrophic cardiomyopathy (HCM)***
  Variants in *MYH7* have been identified in patients with HCM. *MYH7* encodes the beta-myosin heavy chain, the major protein comprising the thick filament of the cardiac sarcomere. Genes encoding other thick filament cardiac sarcomeric proteins, including MYBPC3, MYL2, MYL3, have been definitively associated with HCM. Therefore, the function of MYH7 is shared with other known genes in the disease of interest. (Default: 0.5 points)
- ***Example: Biallelic mutations in DRAM2 cause retinal dystrophy.***
  El-Asrag et al, 2016 [1] report variants in *DRAM2* in patients with retinal dystrophy. The authors review 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 (Brownstein et al, 2013 [2]; Cooper et al, 2014, [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***
  Li et al, 2015 [4] used qPCR to demonstrate that TMEM132E is highly expressed in the cochlea and the brain, two tissues that can be affected by hearing loss. 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***
  Mencacci et al, 2016 [5] reported variants in PDE10A in individuals with childhood onset chorea. They report that microarray data from post-mortem brain tissue showed exceptionally high expression in the putamen, consistent with data in the Allen Mouse Brain Atlas and previous

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

- **Example: Leptin and Severe early-onset obesity**
  Leptin is a hormone secreted by adipose tissue that signals satiety. Montague et al 1997 [8], examined two severely obese children from a consanguineous Pakistani family. 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, 2016 [9] studied smooth muscle cells derived from isolated pluripotent stem cells from patients with Marfan syndrome and variants in *FBN1* (p.Cys1242Tyr and p.Gly880Ser). 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. Destablization 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 and 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, 2012[10] transduced neonatal murine cardiomyocytes with AAV constructs with FHL1 p.Lys45Serfs and p.Cys276Ser variants. Variant FHL1 proteins were mislocalized and did not incorporate into the sarcomere. Localization and incorportation 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, 2015 [4] knocked down TMEM132E in zebrafish using antisense morpholino oligos. 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, 2012 [10] measured contraction in AAV transduced **rat engineered heart tissue** (rEHT) expressing FHL1 variants. 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 (Farooqi et al 1999)[12], hyperphagia ceased and the amount of body fat lost was 15.6kg (accounting for 95% of the weight lost).(Default: 2 points)
- **Example: Rescue in an animal model**
  TMEM132E and autosomal recessive sensorineural hearing loss
  Li et al, 2015 [4] injected human TMEM132E mRNA into antisense oligo knockdown zebrafish. This partially rescued the hearing defects in those fish. (1 point was given instead of 2, default, 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, 2012 [12] studied cultured fibroblasts from a patient with EDS type IV who was heterozygous for p.Gly252Val in COL3A1 and from a healthy control. 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 the 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, 2006 [13] report clinical improvements in 8 patients with infantile-onset Pompe disease who received a weekly intravenous infusion of recombinant GAA for 52 weeks. 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)

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.