Lumping vs. Splitting: Defining the disease entity for gene:disease curations

Courtney Thaxton, Ph.D.
on behalf of the Lumping and Splitting Working Group

Lumping and splitting working group members

Chair

Courtney Thaxton, Ph.D.

Co-chair:

Jenny Goldstein Ph.D., CGC

- ClinGen Biocuration supervisor

Nosology and Ontology expert(s):

- Melissa Haendel, Ph.D.
  - PI Monarch Initiative (MonDO)
- Ada Hamosh, M.D., M.P.H
  - director of OMIM

Clinicians/ Clinical molecular geneticists:

- Jonathan Berg, M.D., Ph.D.
  - ClinGen PI
- Heidi Rehm, Ph.D.
  - ClinGen PI

Biocurators:

- Marina DiStefano, Ph.D.
- Kathleen Wallace
- Dane Witmer, Ph.D.
- Erin Rooney Riggs, M.S., CGC (former)
- Raj Ghosh, Ph.D. (former)

BRCA2 GENE; BRCA2

Alternative titles; symbols

FANCD1 GENE; FANCD1

HGNC Approved Gene Symbol: ***BRCA2***

Cytogenetic location: ***13q13.1*** ***Genomic coordinates (GRCh38):*** 1

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number
13q13.1	Fanconi anemia, complementation group D1	605724
	Wilms tumor	194070

MYOSIN, HEAVY CHAIN 7, CARDIAC MUSCLE, BETA; MYH7

Alternative titles; symbols

MYOSIN, CARDIAC, HEAVY CHAIN, BETA

HGNC Approved Gene Symbol: ***MYH7***

Cytogenetic location: ***14q11.2*** ***Genomic coord***

Gene-Phenotype Relationships

Location	Phenotype
14q11.2	Cardiomyopathy, dilated, 1S
	Cardiomyopathy, hypertrophic, 1
	Laing distal myopathy

al dominant
al recessive
ic type

otype
number

FIBRILLIN 1; FBN1

Alternative titles; symbols

FIBRILLIN; FBN

HGNC Approved Gene Symbol: ***FBN1***

Cytogenetic location: ***15q21.1*** ***Genomic coordinates (GRCh38):*** ***15:48,408,305-48,645,787*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
15q21.1	Acromicric dysplasia	102370	AD	3
	Ectopia lentis, familial	129600	AD	3
	Geleophysic dysplasia 2	614185	AD	3
	Marfan lipodystrophy syndrome	616914	AD	3
	Marfan syndrome	154700	AD	3
	MASS syndrome	604308		3
	Stiff skin syndrome	184900	AD	3
	Weill-Marchesani syndrome 2, dominant	608328	AD	3

METHYL-CpG-BINDING PROTEIN 2; MECP2

HGNC Approved Gene Symbol: ***MECP2***

Cytogenetic location: ***Xq28*** ***Genomic coordinates (GRCh38):*** ***X:154,021,799-154,097,730*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
Xq28	Encephalopathy, neonatal severe	300673	XLR	3
	Mental retardation, X-linked syndromic, Lubs type	300260	XLR	3
	Mental retardation, X-linked, syndromic 13	300055	XLR	3
	Rett syndrome	312750	XLD	3
	Rett syndrome, atypical	312750	XLD	3
	Rett syndrome, preserved speech variant	312750	XLD	3
	{Autism susceptibility, X-linked 3}	300496	Mu, XL, IC	3

	Pancreatic cancer	260350	SMu, Mu, AD	3
	{Basal cell carcinoma 7}	614740		3
	{Glioma susceptibility 1}	137800	SMu, AD	3

ATAXIA-TELANGIECTASIA MUTATED GENE; ATM

HGNC Approved Gene Symbol: ***ATM***

Cytogenetic location: ***11q22.3*** ***Genomic coordinates (GRCh38):*** ***11:108,222,483-108,369,101*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
11q22.3	Ataxia-telangiectasia	208900	AR	3
	Lymphoma, B-cell non-Hodgkin, somatic			3
	Lymphoma, mantle cell, somatic			3
	T-cell prolymphocytic leukemia, somatic			3
	{Breast cancer, susceptibility to}	114480	AD	3

When to lump, or when to split?

- Given that a gene can be associated with several phenotypes and/or disease entities, how can a curator assess if the condition they are curating is the most appropriate?
  - If all curations are split for each phenotypic feature reported in the literature, we risk ***ad infinitum*** curations on ClinGen that might not reflect the true nature of each condition or association, nor the mutational spectrum of the gene.
  - If we lump all of the existing phenotypic features asserted for a gene into a greater known syndrome, we risk ***losing the intricacies*** of each condition and/or phenotype (phenotypic feature).

General principle:

Genes associated with a single published disease entity should only be curated for that condition (i.e. lumped) unless there are indications to split specific phenotypic features of a syndrome or variable phenotype into separate curations based on the guidance provided by our criteria.

Criteria

- **Assertion/ Defining the disease entity**:
  - What has the literature reported, as well as other nosological authorities, about a gene's association with disease?
- **Molecular Mechanism:**
  - Are there differences in the molecular mechanism(s) underlying each asserted disease (and/or condition)?
- **Phenotypic Variability:**
  - Does a phenotype(s) segregate consistently within a pedigree, or present similarly in two or more unrelated probands with the same gene variant?
  - Or, is there variable expressivity of phenotype between family members and/or an unrelated proband harboring the same gene variant?
- **Inheritance Pattern:**
  - Are there differences in inheritance patterns between the disease entities associated with the gene of interest?
  - Do they represent distinct disease entities, or a continuum of disease?

Has an **ASSERTION** for a distinct condition or isolated phenotype been made in the literature or reputable source for disease entities?

NO

YES

NO

Criteria Development

Overview of criteria with examples

Assertion/ Defining a disease entity

- Assess if one or more disease entities have been reported as being associated with a gene by nosological authorities or in the literature.
  - Check OMIM, MonDO (Monarch Initiative), Orphanet, GeneReviews, supplemented by the primary literature.

* 304040

GAP JUNCTION PROTEIN, BETA-1; GJB1

Alternative titles; symbols

GAP JUNCTION PROTEIN, 32-KD
CONNEXIN 32; CX32
LIVER CONNEXIN

HGNC Approved Gene Symbol: *GJB1*

Cytogenetic location: *Xq13.1* ***Genomic coordinates (GRCh38):*** *X:71,215,211-71,225,214* (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
Xq13.1	Charcot-Marie-Tooth neuropathy, X-linked dominant, 1	302800	XLD	3

+601728

Table of Contents

- Title
- Gene-Phenotype Relationships
- Text
  - Description
  - Cloning and Expression
  - Biochemical Features
  - Gene Structure
  - Mapping
  - Gene Function
  - Molecular Genetics
  - Genotype/Phenotype Correlations
  - Animal Model
- Allelic Variants
  - Table View
- References
- Contributors
- Creation Date
- Edit History

PTEN HAMARTOMA TUMOR SYNDROME, INCLUDED; PHTS, INCLUDED
PTEN HAMARTOMA TUMOR SYNDROME WITH GRANULAR CELL TUMOR, INCLUDED

HGNC Approved Gene Symbol: *PTEN*

Cytogenetic location: *10q23.31* ***Genomic coordinates (GRCh38):*** *10:87,863,437-87,971,929* (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
10q23.31	Bannayan-Riley-Ruvalcaba syndrome	153480	AD	3
	Cowden syndrome 1	158350	AD	3
	Endometrial carcinoma, somatic	608089		3
	Lhermitte-Duclos syndrome	158350	AD	3
	Macrocephaly/autism syndrome	605309	AD	3
	Malignant melanoma, somatic	155600		3
	PTEN hamartoma tumor syndrome			3
	Squamous cell carcinoma, head and neck, somatic	275355		3
	VATER association with macrocephaly and ventriculomegaly	276950	AR	3
	{Glioma susceptibility 2}	613028		3
	{Meningioma}	607174	AD	3
	{Prostate cancer, somatic}	176807		3

Molecular Mechanism

- Assess whether differences in molecular mechanism(s) underlie each asserted disease entity.
  - Molecular mechanism(s) includes loss of function (LOF), gain of function (GOF), domain specific effects, isoform specificity, etc.

SCN8A

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
12q13.13	?Cognitive impairment with or without cerebellar ataxia	614306	AD	3
	Epileptic encephalopathy, early infantile, 13	614558	AD	3
	Seizures, benign familial infantile, 5	617080	AD	3

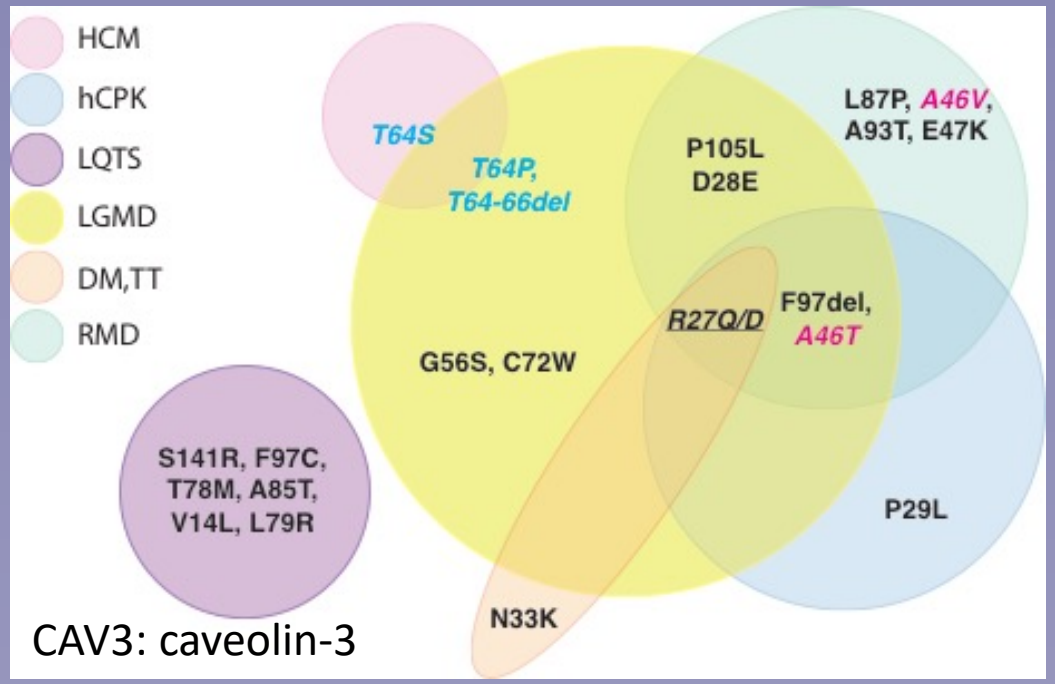
→ LOF

⇒ GOF

Phenotypic Variability

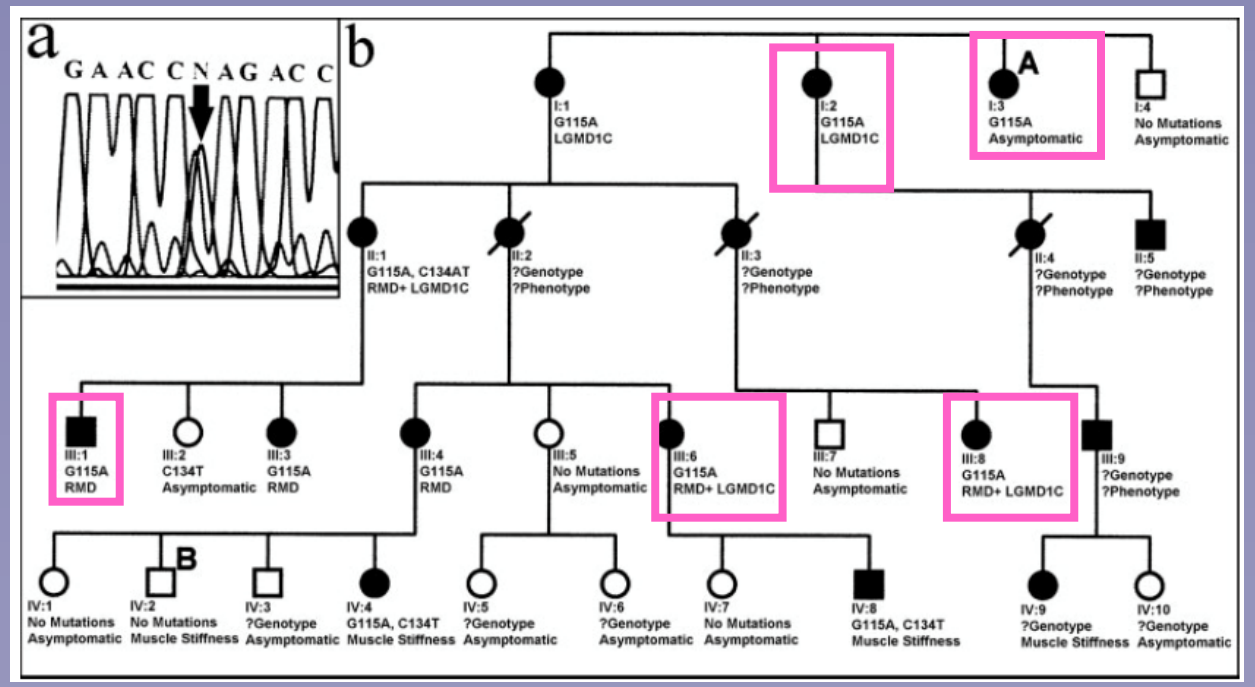
➢ Assess how the gene, the variant(s), and the phenotypic feature(s) segregate throughout a single pedigree (intrafamilial expressivity), or between two or more unrelated probands (interfamilial expressivity).

Overlapping variants



CAV3: caveolin-3

Intrafamilial Variability of R27Q variant



Fee et al., 2004

Inheritance Pattern

- Assess if the disease entities asserted for the gene follow different inheritance patterns.
  - It is important to determine if the disease entities are distinct (multiple varying phenotypes between them), or part of a continuum of disease (same phenotypes, differing severity).

* 607585

ATAXIA-TELANGIECTASIA MUTATED GENE; ATM

HGNC Approved Gene Symbol: ATM

Cytogenetic location: 11q22.3 *Genomic coordinates (GRCh38): 11:108,222,483-108,369,101* (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
11q22.3	Ataxia-telangiectasia	208900	AR	3
	Lymphoma, B-cell non-Hodgkin, somatic			3
	Lymphoma, mantle cell, somatic			3
	T-cell prolymphocytic leukemia, somatic			3
	{Breast cancer, susceptibility to}	114480	AD	3

Ataxia-telangiectasia = Characterized by cerebellar ataxia, telangiectasias, immune defects, and a *predisposition to malignancy (breast cancer and lymphomas).*

Susceptibility to breast cancer = characterized by susceptibility to develop mammary carcinoma.

Reasons to LUMP

1. An assertion for only one disease entity has been made in the literature.
2. No difference in molecular mechanism is observed among the disease entities.
3. Intrafamilial phenotypic variability is as, or more, pronounced than interfamilial variability.
4. The difference in the inheritance pattern for the disease entities is representative of a continuum of disease, i.e. mild carrier phenotypic features are observed in recessive disease or dosage impacts are observed for dominant disease (more severe phenotype in homozygotes).
   a. **Note:** curate for the well-established inheritance pattern and note the additional manifestations in carrier state or homozygous state in the GCI.
5. The disease entities in question are seemingly part of a variable phenotype observed within a single organ system and there is insufficient evidence for any single phenotype.
   a. If variants for each entity are VUSs, and no distinguishing phenotype is observed, then lump for a broader phenotype.

Reasons to SPLIT

1. An assertion for more than one distinct disease entity has been made in the literature.
2. A well-established difference in molecular mechanism(s) between two or more disease entities is observed.
3. The representative disease entities between differing inheritance patterns are distinguishable, with notable varying phenotypes and/or clinical management distinctions.
4. To dispute a disease entity asserted for the gene in question.
   a. Must have convincing evidence to dispute or refute.
   b. This would be a very rare occurrence, and the isolated disease entity being disputed or refuted cannot be included as part of the phenotypic spectrum observed in a syndrome associated with the gene of interest.

Implementation of lumping and splitting criteria

Pre-curation and Binning of gene:disease relationships

Pre-curation strategy

- Involves assessment of the **gene** in question and the conditions/phenotypes associated with any gene and include:
  1. Assessing the disease entities/ conditions/ phenotypic features associated with the gene of interest.
  2. Defining the molecular mechanism(s) for the gene and disease entities.
     1. ***Specifically, look at the genetic variants associated with each disease entity.***
  3. Assessing the variability of phenotypic presentations.
  4. Noting inheritance pattern differences between disease entities.
- Can also address disease onset and severity, and the age at phenotypic presentation.

Pre-curation: Important sites

- Recommended sites to use for pre-curation:
  - OMIM
  - Orphanet
    - Another ontological source is Monarch Initiative (MonDo)
  - ClinVar
  - GeneReviews
    - If available
  - Pubmed
    - specifically in reference to using a nice review, instead of primary literature and case reports
- This is not meant to be an exhaustive search, but curators may find when performing the full curation, once delving into the primary literature, that lumping and splitting issues arise
  - If this occurs, it maybe necessary to revisit the lumping and splitting criteria, and to speak with an expert.

Pre-curation: Template

Gene	Phenotype/Syndrome(s)	Associated MIM	Orphanet number	Inheritance Pattern	Mechanism	Variants (protein unless stated with a c. for codon)	Presenting phenotype (s)

The core template should include, the (1) asserted disease entities, (2) inheritance patterns, (3) molecular mechanisms (more specifically the variants asserted for each disease entity), and (4) presenting phenotypes..

Disease Onset	Are learning disabilities noted?	Is developemetnal delay noted?	PMID	Additional info	Lump or split	* = gnomAd frequency

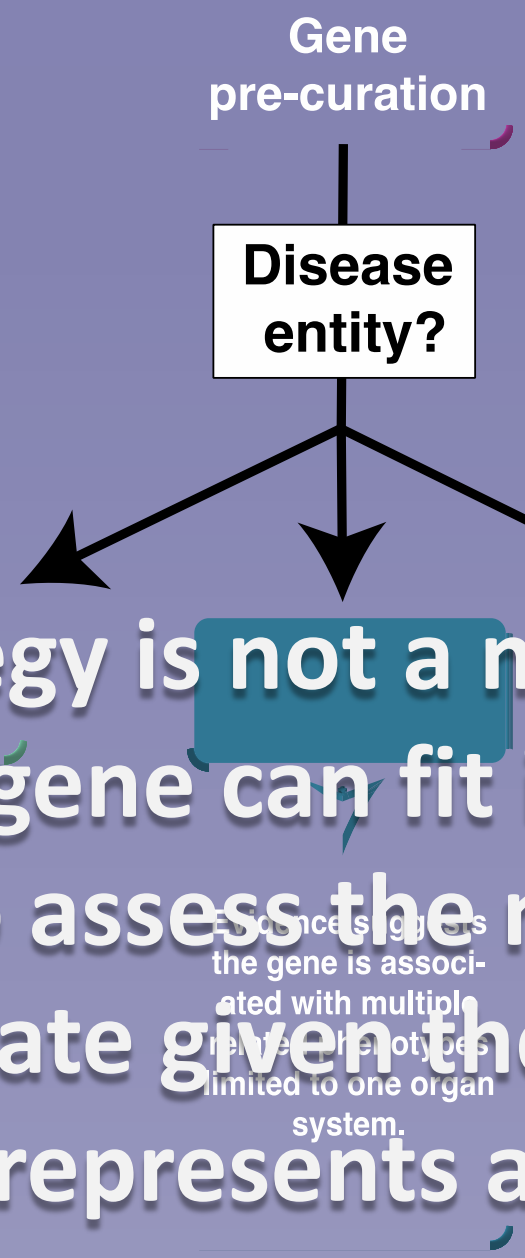
The remaining template can include information that the specific Expert Panel deems important for delineating the disease entity: clinical management, age of disease onset or phenotypic presentation, population frequencies, etc.

Pre-curation Strategy

Gene	Disease Entity(s)	MIM #	Inheritance Pattern	Gene variant(s)	Presenting phenotype(s)
ACTN2	Dilated cardiomyopathy (DCM +/- LVNC)	**612158**	AD	Q9R, **A119T**, N841Ifs, A732T, V115M	left ventricular enlargement, systolic dysfunction, atrial fibrillation, ventricular arrhythmia
	Hypertrophic cardiomyopathy (HCM +/- LVNC)	**612158**	AD	G111V, T495M, R759T, **A119T**, E628G, M228T	asymmetric septal hypertrophy, atrial hypertrophy, arrhythmia
CAV3	Hypertrophic cardiomyopathy (HCM)	192600	AD	**T64S**	left ventricular hypertrophy (LVH), asymmetric septal hypertrophy, arrhythmias
	HyperCKemia	123320	AD	**A46T**, **R27Q** , **F97del**, P29L	***elevated creatine kinase (CK) levels***, normal muscle biospy
	Long QT Syndrome	611818	AD	S141R, F97C, T78M, A85T, V14L, L79R	prolonged Qt interval, polymorphic ventricular arrhythmias (torsade de pointes)
	Limb-girdle muscular dystrophy	607801	AD, **AR** (Semi-dominance)	**P105L**, **T64_T66del**, **T64P**, G56S*, C72W* **A46T**, **R27Q**, **D28E**, **F97del**	adult onset proximal muscle weakness, elevated creatine kinse in serum, abnormal protein aggregtion in sarcomere(s), ***can present with cardiomyopathy (HCM and DCM)***
	Distal myopathy, Tateyama type	614321	AD	**R27Q**, N33K	muscle weakness, heart palpitations, Pes cavus, hand and calf atrophy, ***elevated CK levels***
	Rippling muscle disease, type 2	606072	AD	**P105L**, **A46T**, *A64V*, **R27Q**, **D28E**, **F97del** L87P, A93T, E47K	muscle weakness and cramps, pain and stiffness with exercise, muscle hyperirritability, hypertrophy, mounding. ***Elevated CK levels***. ***Cardiomyopathy (DCM and/or HCM) and sudden cardiac death.***
MYL2	Hypertrophic cardiomyopathy (HCM)	608758	AD	A13T, E22K, P94R, R58Q,	left ventricular hypertrophy (LVH), asymmetric septal hypertrophy, arrhythmias

Used a combination of OMIM, Orphanet, GeneReviews, and literature reviews

Binning Strategy



The binning strategy is not a measure to restrict the curations, as a gene can fit into more than one bin. It is a way to assess the most appropriate conditions to curate given the presentation of phenotypes, and represents a form of lumping.

Other Considerations

Sometimes a GCEP may wish to curate genes for their potential to be associated with a phenotypic feature (or phenotype) that has special testing, treatment or management distinctions, but which may not represent a truly distinct condition; i.e. the phenotypic feature is part of a known syndrome.

In these cases, the gene should be assessed for the syndrome and not for the isolated phenotype (or phenotypic features) UNLESS the criteria above are met and indicate an appropriate split curation.

In order to display the significance of a subset of isolated features of a syndrome, GCEPs may find it useful to generate a table to depict the possibility of presenting as an isolated phenotype as well as the presence, absence, or likelihood of individual features of interest for publication and testing purposes. **This data can be displayed simply as an annotated table in publications, without requiring a formal splitting for gene curation or use of the gene-disease Clinical Validity Classifications**.

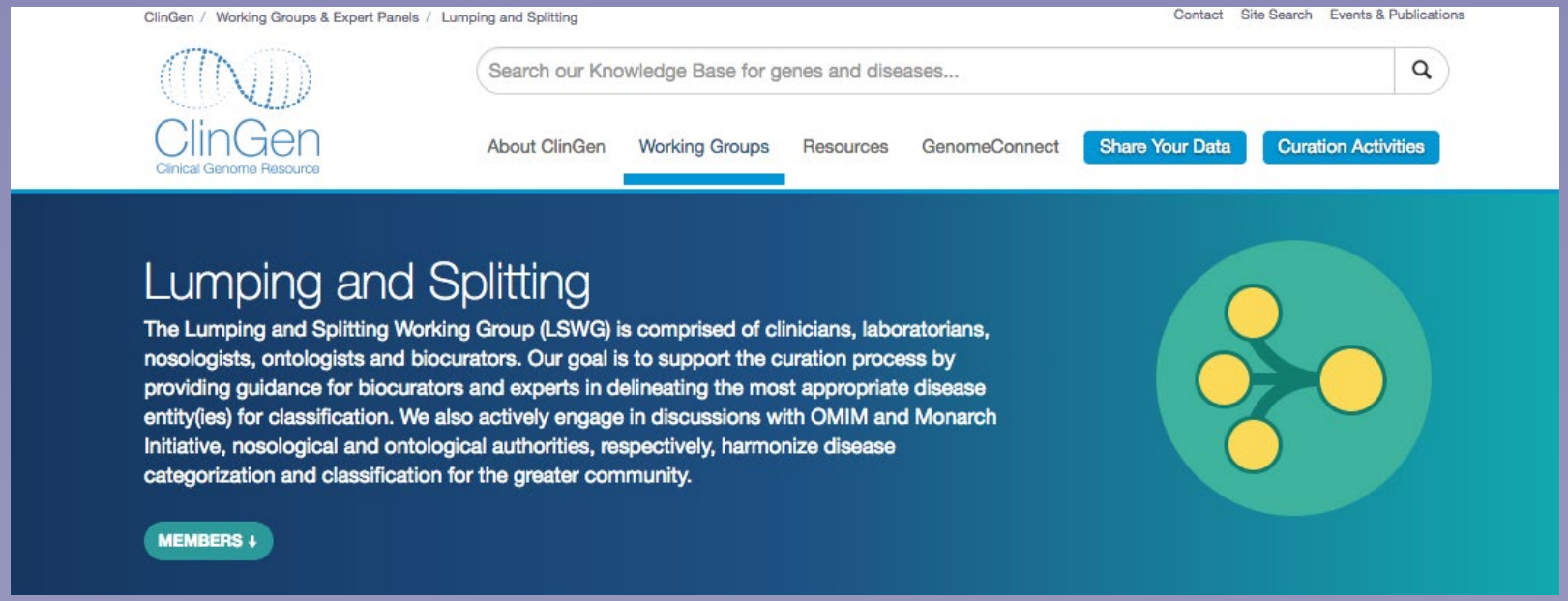
ClinGen Gene:Disease Clinical Validity Classifications for genes causing syndromes which include Hypertrophic Cardiomyopathy as part of the presentation.

Please note: These genes were curated for the syndrome NOT HCM

Gene	Disease	Classification	*Group	Notes and PMIDs
ABCC9	Cantu syndrome	Definitive	I	
BAG3	Myofibrillar myopathy	Definitive	I	
CACNA1C	Timothy syndrome	Definitive		
CAV3	Caveolinopathy	Definitive		
COX15	Cytochrome c oxidase deficiency	Strong	I	
CRYAB	Myofibrillar myopathy	Definitive		
DES	Myofibrillar myopathy	Definitive		
FHL1	Emery Dreifuss muscular dystrophy	Definitive	II	
FLNC	Myofibrillar myopathy	Definitive	II	
FXN	Friedreich's ataxia	Definitive	I	
GAA	Pompe disease	Definitive	II	Rare LOPD cases can present with HCM (PMID 28185884, 25213570, 12162158, 277063)
GLA	Fabry disease	Definitive	II	~1% of individuals with HCM have Fabry disease (PMID 18154965, 19797197, 21890869, 27225851)
LAMP2	Danon disease	Definitive	II	
LDB3	Myofibrillar myopathy	(Moderate – not yet expert-reviewed. Previously curated for HCM but now lumped for syndrome)		
PRKAG2	(TBD)	Definitive		
SLC25A4	Mitochondrial DNA depletion syndrome	Definitive	I	
TTR	Amyloidosis	Definitive	II	

*Group I: Syndromes that include HCM (or HCM-like features) as part of the phenotypic spectrum but for which other phenotypic features are typically present.

*Group II: Syndromes that may present with HCM (or HCM-like features), in the absence of other clinical features, or with mild features that may be missed.

For more information, please see the Lumping and Splitting Working Group page on clinicalgenome.org or email courtney_thaxton@med.unc.edu

https://www.clinicalgenome.org/working-groups/lumping-and-splitting/