



Launch of the Gene Curation Coalition (GenCC) Database

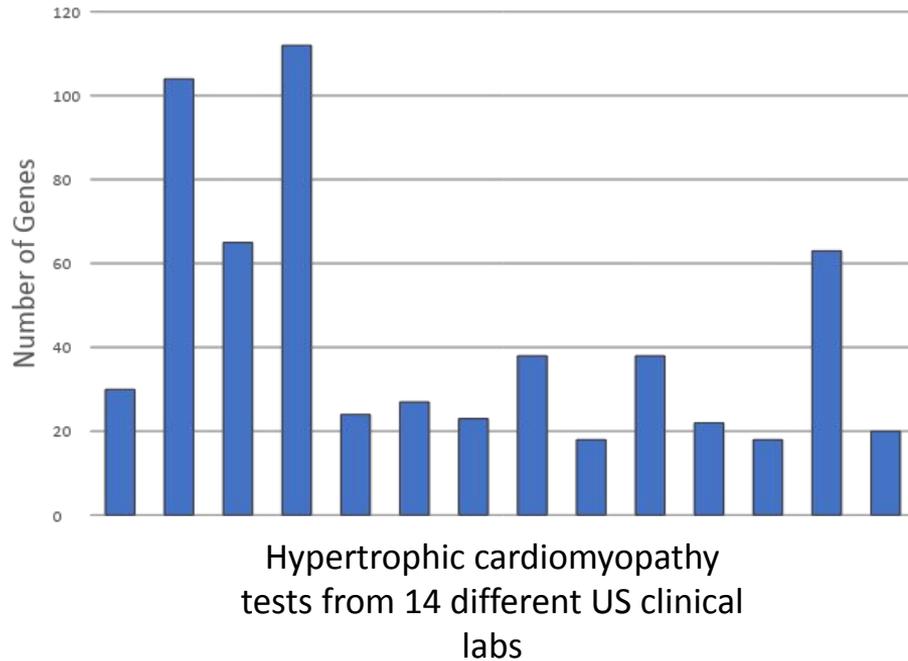
Marina DiStefano
Biocurator Working Group
3.11.21

Overview

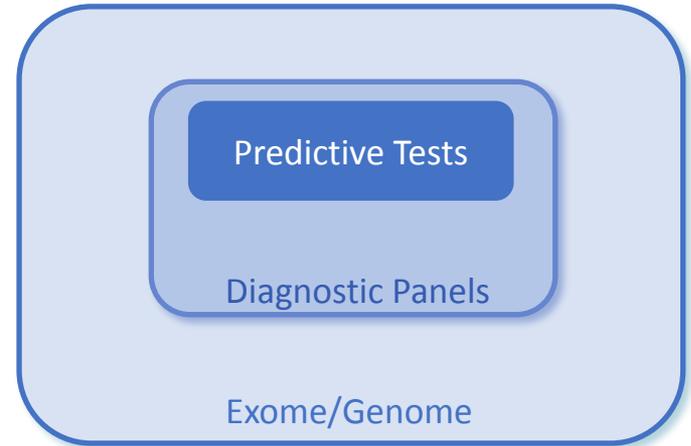
- What is the GenCC and why do we need it?
- A Delphi survey to harmonize clinical validity terms across GenCC
- Live demo of the GenCC database/tips and tricks to use it for ClinGen curation

What is the Gene Curation Coalition (GenCC)
and why do we need it?

The Need for Curated Gene Databases of Gene-Disease Validity



Standards and consensus are needed for which genes are valid disease genes that are ready for clinical testing



Different levels of evidence are needed for different clinical uses

GenCC Members

Public Gene-Level Resources



*Genetics
Home
Reference*

Private Gene-Level Resources



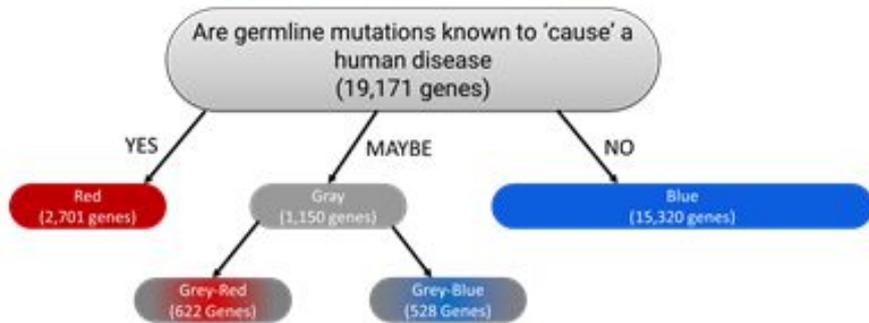
Goals of GenCC

- Understand the approaches and classification systems of the different curation efforts
- Develop consistent terminology for validity assessment as well as inheritance, allelic requirement, mechanism of disease
- Post gene curations from any group willing to share with the public
 - **The GenCC DB is like a ClinVar for genes!**
- Resolve differences and collaborate on gene curation projects



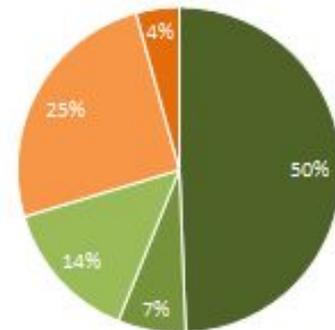
Curation efforts differ in gene curation results

Rahman Gene Disease Map

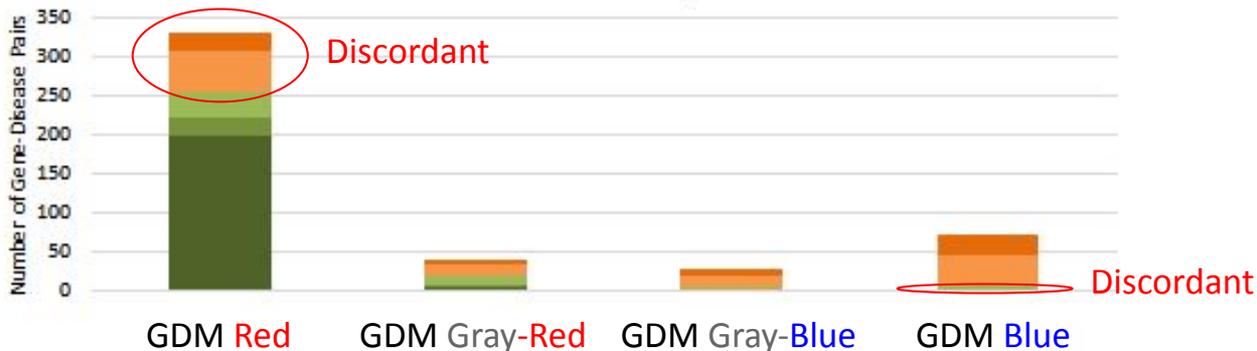


ClinGen Gene-Disease Validity (599 Genes)

- Definitive
- Strong
- Moderate
- Limited
- Refuted or Disputed



ClinGen-GDM Comparison



PanelApp Comparisons Between GEL and AG

content

Australian Genomics
Health Alliance

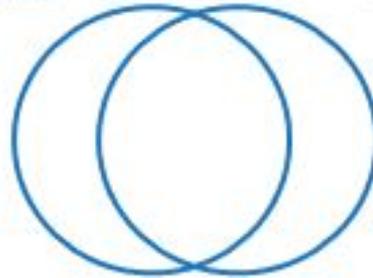
Genomics
england



Global Alliance
for Genomics & Health

ID

1,829 genes



2,720 genes

Epilepsy

1,052 gene discrepancies
reviewed

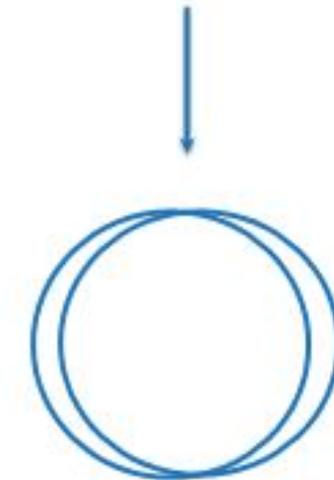
PID

802 new genes added
to panels

Mito

250 existing genes
up/downgraded

Leuko-
dystrophy



GenCC Gene Curation Bakeoff

	Orphanet	OMIM	PanelApp	ClinGen	GraemeBlack	G2P	GDM
ARHGEF6	Confirmed	One family/Possible	Amber/Probable	Limited/Unlikely	Candidate/Possible	Possible/Unlikely	Blue/Unlikely
DAO	Confirmed	Possible/Pending Confirmation	Red/Unlikely	Limited/Possible	Candidate/Unlikely	Not disease causing/Unlikely	Blue/Unlikely
HRG	Confirmed	Disease Gene/Possible	Green/Confirmed	Moderate/Possible	Possible	Possible	Blue/Unlikely
LIG1		VUS	Amber/Probable	Limited/Possible	Unlikely	Unlikely	Blue/Unlikely
NLRP12	Confirmed	Disease Gene	Green/Confirmed	Disputed/Unlikely	Possible	Possible	Blue/Unlikely
PIGM	Confirmed	Disease Gene	Amber/Probable	Limited/Possible	Candidate/Possible	Unlikely	Blue/Unlikely
PTCH2	Confirmed	Disease Gene	Amber/Probable	Limited/Possible	Unlikely	Unlikely	Blue/Unlikely
RNF135	Confirmed				Unlikely	Unlikely	Blue/Unlikely
ROBO2	Confirmed	Susceptibility Gene	Amber/Probable	Limited/Possible	Possible	Unlikely	Blue/Unlikely
SPRX2	Confirmed, Possible	One family/Possible	Amber/Probable		Candidate/Possible	Unlikely	Blue/Unlikely
ZNF81	Confirmed	Unlikely	Red/Unlikely	Disputed/Unlikely	Unlikely	Unlikely	Blue/Unlikely

Pilot comparisons identified reasons for gene curation discordance

- Definition of disease (high penetrance versus inclusion of lower penetrance phenotypes)
- Purpose of curation (validity versus panel inclusion based on phenotype match)
- Differences in disease/phenotype assignment (genes often have claims for multiple diseases)
- Date of evidence evaluation
- Understanding the differences and focusing on resolving them requires harmonization of terms and definitions

Comparison of Terms

ClinGen	GDM	G2P	Orphanet	OMIM	PanelApp
Definitive	Red	Confirmed	Present	Yes	Green
Strong	Red	Confirmed	Present	Yes	Green
Moderate	Grey-Red	Probable	Absent	Yes	Amber
Limited	Grey-Blue / Grey-Red?	Possible	Candidate	?Disease	Red
No evidence	Blue	Absent	Absent	No disease claim	Red
Disputed	Grey-Blue	Absent	Candidate	?Disease	Red/Amber
Refuted	Blue	Absent	Absent (Suppressed)	Reclassified-VUS	Red

A Delphi Survey to Harmonize Clinical Validity Terms

GenCC Term Delphi Survey Process

Draft harmonized definitions for gene curation categories

```
graph TD; A[Draft harmonized definitions for gene curation categories] --> B[Solicit term suggestions from the GenCC members and draft Delphi Survey]; B --> C[Round 1: Survey GenCC members (N=33)  
Round 2: Survey extended membership of GenCC groups (N=38)  
Round 3: Survey the international genetics community (N=241)];
```

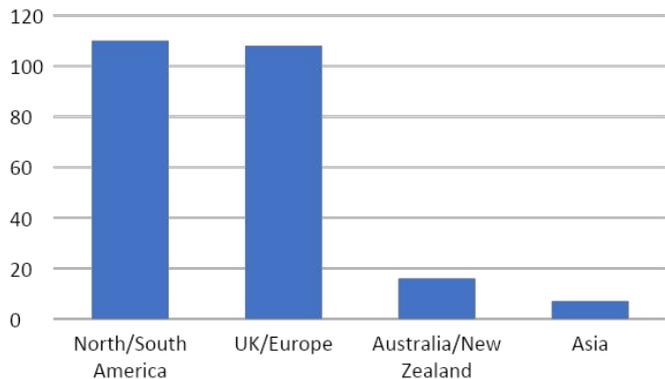
Solicit term suggestions from the GenCC members and draft Delphi Survey

Round 1: Survey GenCC members (N=33)
Round 2: Survey extended membership of GenCC groups (N=38)
Round 3: Survey the international genetics community (N=241)

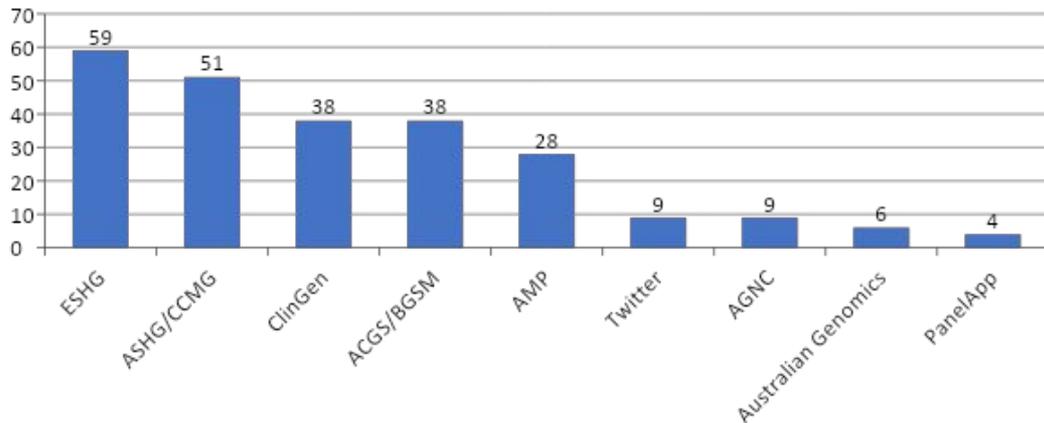
Optional
Explanatory
Video

Survey Round 3 Demographics

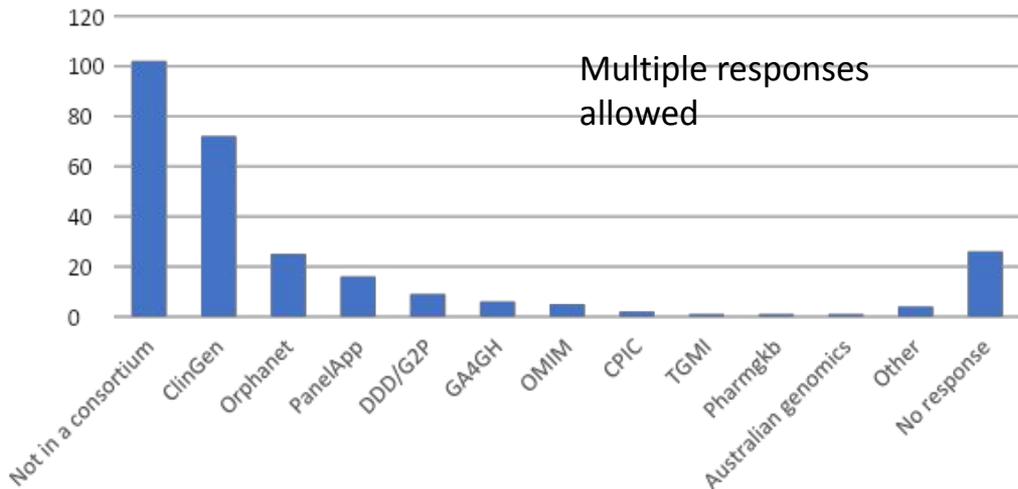
Respondant Location (N=241)



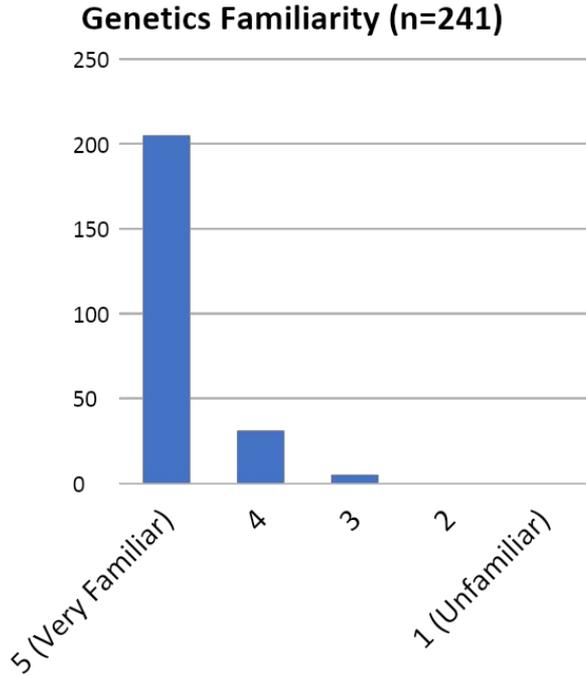
Surveyed Populations (n=241)



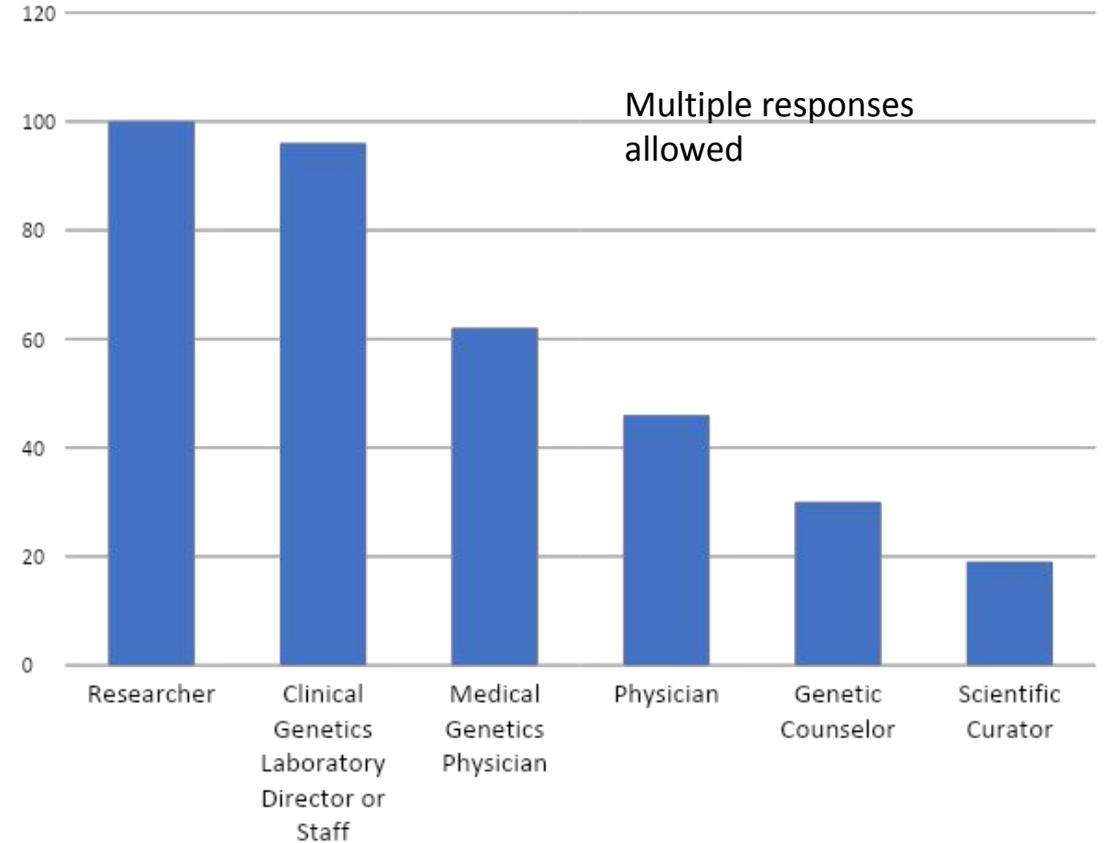
Consortium Participation (n=270)



Demographics



How would you describe yourself? (N= 353)



Finalized Terms

- Delphi survey round 3 had **93% term agreement** (13/14 terms) with round 2, suggesting the genetics community is harmonious with GenCC
- All groups map to these terms on the GenCC website (Or “Supportive” if not enough granularity to do so)

GenCC
Definitive
Strong
Moderate
Limited
Animal Model Only
No Known Disease Relationship
Disputed Evidence
Refuted Evidence

Launch of the GenCC database



Scott Goehringer

Overview of submission pilot

- GenCC member groups used spreadsheet-based submissions (API based submission is a future upgrade)
- Curations are mapped to the harmonized terms generated by the Delphi survey (with the addition of “Supportive” for OMIM and Orphanet)
- Disease terms accepted are OMIM, Orpha, and MONDO (harmonization is performed with MONDO)
- Download available of all data sets but OMIM (must obtain from OMIM’s website)
- 8 groups completed pilots and more are pending

3728

Submitted Classifications

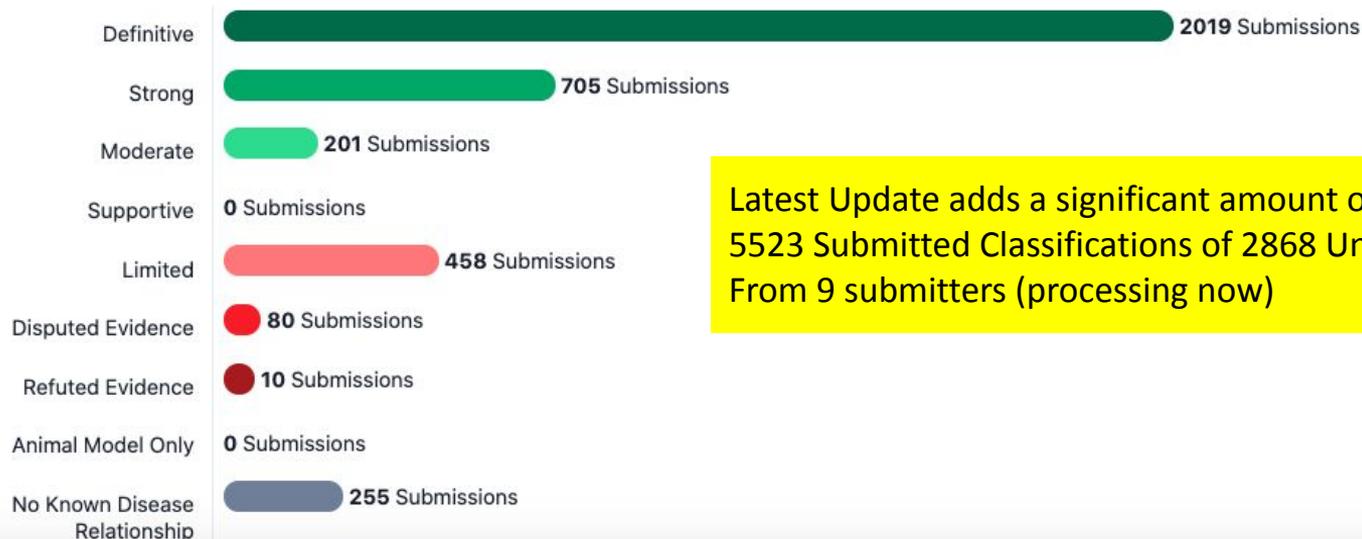
2282

Unique Genes with Submissions

8

Submitters with submissions
[Learn about GenCC's submitters](#)

Classifications Visualized



Latest Update adds a significant amount of data:
5523 Submitted Classifications of 2868 Unique Genes
From 9 submitters (processing now)

Next: Analyze data to assess concordance and begin collaborations to resolve differences

We have not yet analyzed the data but that is next!

The GenCC database is currently released under a Fort Lauderdale Agreement for the benefit of the wider biomedical community. You can freely download and search the data, and we encourage the use and publications of validity data for specific targeted sets of genes. However, we ask that you not publish global (site-wide) analyses of these data, or of large gene sets, until after the GenCC flagship paper has been published (estimated to be spring 2021). After the flagship publication, data will be available free of restriction under a CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. The GenCC requests that you give attribution to GenCC and the contributing sources whenever possible and appropriate.

Ongoing work: Identify and resolve conflicts

Conflict identification will be done in phases

Phase 1: exact disease term matches

Phases 2/3: broader disease terms used for conflict calculation

Preliminarily there are 90 conflicts of exact disease term/MOI matches
(Conflicts in this case defined as Definitive/Strong/Moderate v All others)

Conflict Report

Lorem ipsum dolor sit amet

Consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco commodo consequat. Consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco commodo consequat.

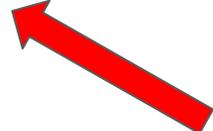
Flag Gene... Disease... MOI... Subm...



Conflicting classifications controls
Group A [slider] Group B [slider] Edit ⚙️

Gene	Submitters	Definitive	Strong	Moderate	Supportive	Limited	Disputed	Refuted	Animal	No Known	Details
ABCA5 HGNC:135	3	2	0	1	0	0	0	0	0	0	Details
BCA1 HGNC:235	2	0	0	1	0	1	0	0	0	0	Details
CA6 HGNC:335	3	0	1	0	2	0	0	0	0	0	Details
XXA1 HGNC:435	3	0	0	0	0	1	0	2	0	0	Details

Genes with conflicts highlighted



Editable conflict parameters

(GenCC will have a recommended setting)

Mock ups, Phase 1

What is this?

Lorem ipsum dolor sit amet
Consectetur adipiscing elit, sed
enim ad minim veniam, quis nostr

Flag Gene... Disease

ABCA5
MONDO
HGNC:135
Submits
Enable

BCA1
MONDO
HGNC:235
Submits
Enable

CA6
MONDO
HGNC:335
Submits
Enable

XXA1
MONDO
HGNC:435
Submits
Enable disease grouping

Classification Conflict Controls

The primary Gene Disease Mode-of-Inheritance (GDM) Trio labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco commodo consequat.

Gene Disease Mode-of-Inheritance (GDM) Trios that have classifications outside of the "include" group defined below will be highlighted with a  icon and unique background color.

	Include	Exclude	
	Group A	Group B	
Definitive	<input checked="" type="radio"/> Include	Default	<input type="radio"/>
Strong	<input checked="" type="radio"/> Include	Default	<input type="radio"/>
Moderate	<input checked="" type="radio"/> Include	Default	<input type="radio"/>
Supportive	<input checked="" type="radio"/> Include	Default	<input type="radio"/>
Limited	<input type="radio"/>	Default	<input checked="" type="radio"/> Exclude
Disputed Evidence	<input type="radio"/>	Default	<input checked="" type="radio"/> Exclude
Refuted Evidence	<input type="radio"/>	Default	<input checked="" type="radio"/> Exclude
Animal Model Only	<input type="radio"/>	Default	<input checked="" type="radio"/> Exclude
No Known Disease	<input type="radio"/>	Default	<input checked="" type="radio"/> Exclude

SAVE

Reset

Close

Definitive Strong Moderate Supportive Disputed Refuted Animal No Known

Inconsistent classifications

0 0 0 0 Details

0 0 0 0 Details

0 0 0 0 Details

0 2 0 0 Details

Phase 2: Database level disease-grouping choices

- Editable conflict parameters, user has log-in and can save searches
- Downloadable conflict report
- Calculating conflicts using broad ontology choices made across the entire dataset (e.g. Treat all disease terms in an OMIM phenotypic series as disease matches)

Phase 3: Ontology tree search/Panel Builder

- MONDO hierarchies will be incorporated into the database
- User will have the ability to calculate conflicts based on their own rule choices (e.g. Disease keyword matches, setting all terms in a specific hierarchy as synonyms)
- The integration of MONDO will allow conflict reports and disease panel building (next slide)

Future Work: Ontology Tree search/Panel Builder

1. Term search

Users can search by any term in the MONDO hierarchy, HPO phenotypes, disease name keywords

2. List filtration

This list can be filtered by clinical validity, by submitters

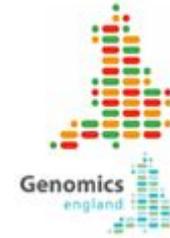
3. Download filtered list

Filtered list can be downloaded and used for panel test design or indication-based analysis of exome/genomes

Acknowledgements

Steering Committee

Heidi Rehm <i>GenCC Chair</i> ClinGen	Helen Firth DECIPHER
Marina DiStefano <i>GenCC Coordinator</i> ClinGen	Ada Hamosh OMIM
Christina Austin-Tse Laboratory for Molecular Medicine (LMM)	Annie Olry Orphanet
Marie Balzotti Myriad Women's Health	Kelly Radtke Ambry
Alison Coffey Illumina	Catherine Snow Genomics England PanelApp
Heather Collins Genetics Home Reference	Zornitza Stark PanelApp Australia
Fiona Cunningham EMBL-EBI	Jackie Tahiliani Invitae
Yaron Einhorn Franklin by Genoox	James Ware Gene2Phenotype (G2P) from TGM



*Genetics
Home
Reference*

Scott Goehringer - Website & Database
Phil Weller

Live Database Demo

search.thegenc.org

Hidden Database Screenshots

Classifications Sorted By Validity

GJB3

[Return to list](#)

[Help](#)

Gene Symbol: **GJB3**
HGNC:4285

Locus Group: **Protein-Coding Gene**

Locus Type: **Gene With Protein Product**

By Classification By Disease By Submitter

Filters: Classifications **3** Diseases **3** MOI **1** Submitters **2**

Strong classifications

Strong **GJB3** **erythrokeratoderma variabilis** **AD** **03/05/2018** **ClinGen**
HGNC:4285 [MONDO:0017851](#) [Public Report](#) [Assertion Criteria](#) [More Details](#)
Evaluated
10/27/2020
Submitted

Limited classifications

Limited **GJB3** **autosomal dominant nonsyndromic deafness 2B** **AD** **10/09/2020** **PanelApp Australia**
HGNC:4285 [MONDO:0012976](#) [Public Report](#) [Assertion Criteria](#) [More Details](#)
[Submitted as: OMIM:612644](#)
Evaluated
11/09/2020
Submitted

Disputed Evidence classifications

Disputed Evidence **GJB3** **nonsyndromic genetic deafness** **AD** **08/23/2018** **ClinGen**
HGNC:4285 [MONDO:0019497](#) [Public Report](#) [Assertion Criteria](#) [More Details](#)
Evaluated
10/27/2020
Submitted

Classifications Sorted By Disease

GJB3

[Return to list](#)

[Help](#)

Gene Symbol: **GJB3**
HGNC:4285

Locus Group: Protein-Coding Gene

Locus Type: Gene With Protein Product

MONDO is used as the mapping ontology, but we accept OMIM, Orpha, and MONDO IDs

By Classification **By Disease** By Submitter

Filters: Classifications 3 Diseases 3 MOI 1 Submitters 2

Autosomal dominant nonsyndromic deafness 2B classifications

Limited **GJB3** autosomal dominant nonsyndromic deafness 2B AD ⓘ
HGNC:4285 MONDO:0012976 ⓘ
Submitted as: OMIM:612644 ⓘ
10/09/2020 Evaluated PanelApp Australia
11/09/2020 Submitted
Public Report ⓘ Assertion Criteria ⓘ More Details ⓘ

Erythrokeratoderma variabilis classifications

Strong **GJB3** erythrokeratoderma variabilis AD ⓘ
HGNC:4285 MONDO:0017851 ⓘ
03/05/2018 Evaluated ClinGen
10/27/2020 Submitted
Public Report ⓘ Assertion Criteria ⓘ More Details ⓘ

Nonsyndromic genetic deafness classifications

Disputed Evidence **GJB3** nonsyndromic genetic deafness AD ⓘ
HGNC:4285 MONDO:0019497 ⓘ
08/23/2018 Evaluated ClinGen
10/27/2020 Submitted
Public Report ⓘ Assertion Criteria ⓘ More Details ⓘ

Individual Submission Page

Submission Details

Submitted by [ClinGen](#)

Submitter: **ClinGen**
GENCC:000102

Classification: Definitive
GENCC:100001

Gene: [COL4A5](#)
HGNC:2207

Disease: [Alport syndrome](#)
MONDO:0018965

Mode Of Inheritance: [X-linked](#)
HP:0001417

Classification Date: 11/03/2020

Evidence/Notes: COL4A5 was first reported in relation to X-linked Alport syndrome in 1990 (Myers et al., 2339699). Eight missense, frameshift, nonsense and splice-site variants reported in humans were included in this curation, however approximately 1000 pathogenic or likely pathogenic variants in this gene have been reported in ClinVar. Evidence supporting this gene-disease relationship includes case-level data, segregation data and experimental data. Variants in this gene have been reported in at least 8 probands in 7 publications (10563487, 14993485, 23085274, 25572247, 27725546, 28542346, 27796712). Variants in this gene segregated with disease in 18 additional family members. More evidence is available in the literature, but the maximum score for genetic evidence (12 pts.) has been reached. This gene-disease association is supported by knock-in mouse models and expression studies in humans (15153557, 30582011, 16301374). In summary, COL4A5 is definitely associated with X-linked Alport syndrome. This has been repeatedly demonstrated in both the research and clinical diagnostic settings, and has been upheld over time. This classification was approved by the ClinGen Hearing Loss Working Group on 3/19/19 (SOP Version 6).

Public Report: [Click here to view the public report](#)
<https://search.clinicalgenome.org/kb/gene-validity/0407dc2e-1cab-4043-889d-4695b043d7b3--2019-03-19T16:00:00>

Assertion Criteria: [Click here to view assertion criteria](#)

Submitter Page

ClinGen

GenCC Ref: GENCC:000102

[Return to list](#)

[Help](#)

ClinGen

This page is a summary of pilot submissions provided by ClinGen. [Click here](#) to be notified about GenCC updates.

ClinGen is a National Institutes of Health (NIH)-funded resource dedicated to building a central resource that defines the clinical relevance of genes and variants for use in precision medicine and research.

<https://www.clinicalgenome.org>

Website

<https://www.clinicalgenome.org>

Personnel

Marina DiStefano, Coordinator

Email: mdistefa@broadinstitute.org

Assertion Criteria

<https://clinicalgenome.org/curation-activities/gene-disease-validity/training-mate...>

Classifications Visualized



Submissions

1101 total number of submissions

Summary Statistics Page

3728

Submitted Classifications

2282

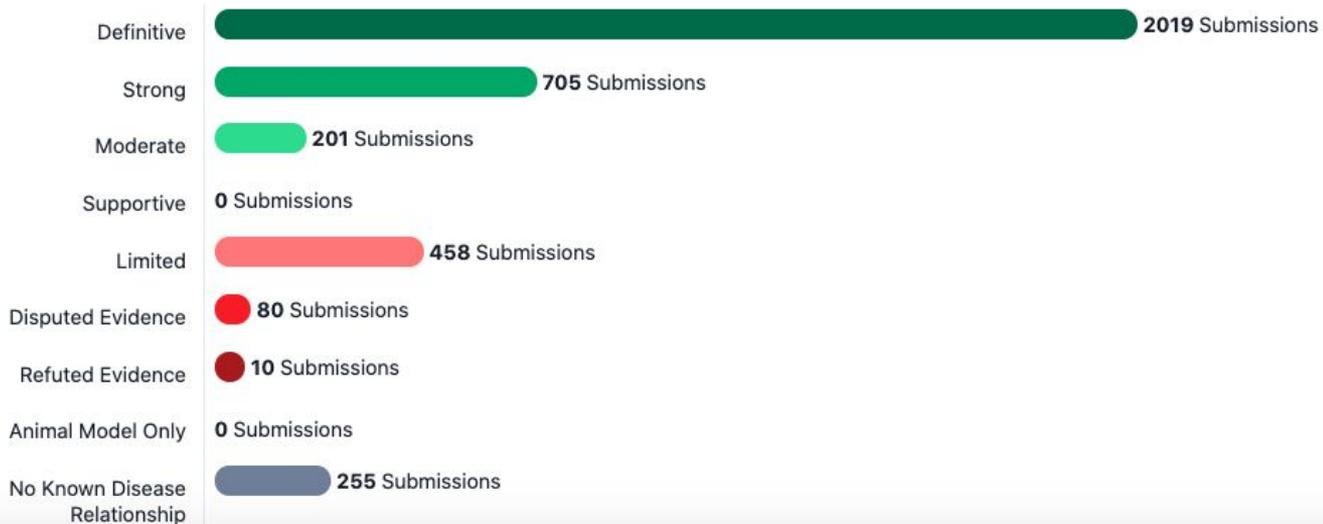
Unique Genes with Submissions

8

Submitters with submissions

[Learn about GenCC's submitters](#)

Classifications Visualized



Ability to Download Data

GenCC Data

 Help

Data Download

Member groups submit assertions about gene-disease relationships. Each entry will be an assertion for a gene, disease, and a mode of inheritance with a link to evidence supporting that assertion. Member groups have submitted pilot sets of data and will continue to add to the data set over time. Due to licensing restrictions, a GenCC download does not include OMIM data. OMIM data can be accessed and downloaded through <https://www.omim.org/>

 Submissions (xlsx)

 Submissions (xls)

 Submissions (tsv)

 Submissions (csv)

API Access

Access to GenCC through and API will be released in early 2021. We recommend anyone interested in access to the API, or would like to be notified when early access is available, to [signup](#) so the GenCC to keep you informed.

 [Stay Informed About GenCC's New Features](#)