Creating an open-source gene curation database from the Gene Curation Coalition (GenCC)

Marina T. DiStefano1, Joanna Amberger2, Christina Austin-Tse3, 4, Marie Balzotti5, Jonathan S. Berg1, Ewan Birney6, Ozge Ceyhan-Birsoy 4, Carol Bocchini2, Elspeth Bruford7, Alison Coffey8, Heather Collins9,10, Fiona Cunningham6, Louise C. Daugherty11, Paul Denny7, Helen Firth12, David Fitzpatrick12, Rebecca E. Foulger11, Jenny Goldstein1, Ada Hamosh2, Matthew Hurles12, Sarah E. Leigh11, Ivone U.S. Leong11, Antoine Marmignon13, Christa L. Martin1, Kalotina Machini 15, Ellen M. McDonagh11, Annie Olry13, Kelly Radtke14, Erin M. Ramos15, Erin Rooney Riggs1, Charlotte Rodwell13, Zornitza Stark16, Jackie Tahiliani17, James S. Ware18,19, Eleanor Williams11, Caroline Wright12, and Heidi L. Rehm1

Please list additional authors from the new participating groups below and we will combine with the list above. Please also remove any authors from your group that did not contribute to the work described below.

Affiliations:

1Clinical Genome Resource (ClinGen), USA

2Online Mendelian Inheritance in Man (OMIM), Johns Hopkins University School of Medicine, USA

3Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

4Partners Healthcare Laboratory for Molecular Medicine, Cambridge, MA, USA

5Myriad Women's Health, San Francisco, CA, USA

6European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

7HUGO Gene Nomenclature Committee (HGNC), European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

8Illumina Clinical Services Laboratory, Illumina Inc., San Diego, CA, USA

9Genetics Home Reference, National Library of Medicine, Bethesda, MD, USA

10ICF, 9300 Lee Highway, Fairfax, VA, 22031 USA

11Genomics England, Queen Mary University of London, Dawson Hall, Charterhouse Square

London, EC1M 6BQ, UK

12Transforming Genetic Medicine Initiative, Wellcome Genome Trust, England, UK

Deciphering Developmental Disorders/Gene2Phenotype, England, UK

13Inserm,US14-Orphanet, France

14AmbryGenetics, Aliso Viejo, CA, USA

15National Human Genome Research Institute, National Institutes of Health, USA

16Australian Genomics, Australia

17Invitae, San Francisco, CA, USA

18National Heart & Lung Institute & MRC London Institute of Medical Sciences, Imperial College London, London, UK

19Royal Brompton & Harefield Hospitals NHS Trust, London, UK

The assessment of the evidence that a gene is linked to a particular disease is critical for variant classification and genomic interpretation. Unless a gene is convincingly linked to disease, the clinical impact of a variant cannot be fully interpreted. Thus, curation of gene-disease validity is a fundamental prerequisite for classifying variants identified in a variety of contexts, such as diagnostic testing, screening, and genomic analysis. The Gene Curation Coalition (GenCC) (thegencc.org) was formed to bring together all groups engaged in the evaluation of gene-disease validity with a willingness to share data publicly, to develop consistent terminology for gene curation activities and to facilitate the consistent assessment of genes that have been reported in association with disease. The GenCC initially included members of groups providing public gene-disease resources including the Clinical Genome Resource (ClinGen), Deciphering Developmental Disorders/Gene2Phenotype (DDD/G2P), Genetics Home Reference (GHR), Genomics England PanelApp, Online Mendelian Inheritance in Man (OMIM), Orphanet, and the Transforming Genetic Medicine Initiative (TGMI). The GenCC drafted standardized definitions of evidence levels for gene disease validity and surveyed the genetics community for their term preferences using the Delphi survey method. Standardized terms were chosen factoring in the 241 community responses and the 71 GenCC member responses received across a series of three rounds of surveying (Definitive, Strong, Moderate, Limited, Animal model only, No known disease relationship, Disputed evidence, and Refuted evidence), and all GenCC member groups agreed to directly use or map to these harmonized terms on their respective curation websites.

Other parties engaged in gene curation and willing to share that data publicly (Ambry, Australian Genomics, Illumina, Invitae, Myriad Women's Health, Partners Laboratory for Molecular Medicine) have now joined the GenCC and efforts are underway to develop an online community resource that aggregates and provides access to gene curations from all sources. To support this, the Genomics England team is setting up a separate instance of their PanelApp software that will display gene-disease validity results in a disease-centric manner for access by the community. GenCC is working with the Monarch Initiative to map the different disease terms used across curation efforts to the MONDO disease ontology which will enable automated aggregation of results. The new GenCC PanelApp instance will contain the following work: ~25,000 OMIM entries, 817 classifications from ClinGen, 314 panels encompassing >12,000 gene disease relationships from the Genomics England PanelApp, ~7400 curations from Orphanet, 5095 gene disease pairs curated using Ambry's published curation method (Farwell Hagmen et al 2016), ~4,000 gene disease relationships from the Australian Genomics PanelApp, 650 gene disease pairs curated by Illumina using the ClinGen framework, ~14,000 gene disease pairs curated by Invitae using their curation method, 208 gene disease pairs curated by Myriad Women's Health, formerly Counsyl, using the ClinGen framework, and ~2230 genes curated by Partners Laboratory for Molecular Medicine using an abbreviated version of the ClinGen framework (Ceyhan-Birsoy et al 2017, Machini et al 2019). All curation results from the participating groups will be mapped to the validity terms that were chosen by the genetics community and then displayed side by side in the GenCC PanelApp. Display of curations across member groups will allow for valuable evidence sharing by the GenCC member groups. For example, ClinGen has 148 curations with limited evidence that could change classification if more evidence were available for evaluation. Additionally, comparison analysis could be done across efforts. For example, when Austrailian Genomics and Genomics England PanelApp curations were compared

there were 1052 discrepancies which resulted in classification changes to 250 gene disease pairs. Comparisons and data sharing such as this will help to resolve any differences in gene classification.