HHS Public Access

Author manuscript

Genet Med. Author manuscript; available in PMC 2021 December 10.

Published in final edited form as:

Genet Med. 2021 October ; 23(10): 1993–1997. doi:10.1038/s41436-021-01213-x.

A Framework for Automated Gene Selection in Genomic Applications

L Lazo de la Vega[1,2,3,4], **W Yu**[1], **K Machini**[1,2,3], **CA Austin-Tse**[1,3,5], **L Hao**[1,3], **CL Blout Zawatsky**[6], **H Mason-Suares**[1,2,3], **RC Green**[3,4,6,7], **HL Rehm**[2,3,4,5], **MS Lebo**[1,2,3,4]

[1]Laboratory for Molecular Medicine, Mass General Brigham Personalized Medicine, Cambridge, MA

[2]Department of Pathology, Brigham & Women's Hospital, Boston, MA

[3]Harvard Medical School, Boston, MA

[4]Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA

[5]Center for Genomic Medicine and Departments of Pathology and Medicine, Massachusetts General Hospital, Boston, MA

[6]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

[7]Ariadne Labs, Boston, MA

Abstract

Purpose—An efficient framework to identify disease associated genes is needed to evaluate genomic data for both individuals with an unknown disease etiology and those undergoing genomic screening. Here, we propose a framework for gene selection used in genomic analyses, including applications limited to genes with strong or established evidence levels and applications including genes with less or emerging evidence of disease association.

Methods—We extracted genes with evidence for gene-disease association from the Human Gene Mutation Database, Online Mendelian Inheritance in Man, and ClinVar to build a comprehensive gene list of 6,145 genes. Next, we applied stringent filters in conjunction with computationally curated evidence (DisGeNET) to create a restrictive list limited to 3,929 genes with stronger disease associations.

Results—When compared to manual gene curation efforts, including the Clinical Genome Resource, genes with strong or definitive disease associations are included in both gene lists

AUTHOR CONTRIBUTIONS

The authors report the following contributions: Conceptualization (MSL, KM, CAAT, HMS), Data Curation (WY, LH), Formal Analysis (LLV, WY), Funding Acquisition (RCG), Investigation (LLV, KM), Methodology (MSL, HLR), Project Administration (CLBZ), Software (WY, LH), Supervision (MSL), Visualization (WY, LLV), Writing - Original Draft (LLV, MSL), Writing - Review and Editing (HLR, CLBZ, RCG, CAAT, HMS). All authors reviewed and approved the final manuscript.

ETHICS DECLARATION

This project has been reviewed and approved by the Mass General Brigham IRB. All individuals consented for clinical genomic screening and all individual data was de-identified.

at high percentages, while genes with limited evidence are largely removed. We further confirmed the utility of this approach in identifying pathogenic and likely pathogenic variants in 45 genomes.

Conclusion—Our approach efficiently creates highly sensitive gene lists for genomic applications, while remaining dynamic and updatable, enabling time savings in genomic applications.

INTRODUCTION

As genome and exome sequencing become standard in clinical genetic testing for patients with unknown genetic etiology and in broad genomic screening for population precision health, an efficient framework to identify and capture all known disease associated genes and pathogenic variants is needed. With the scope of analysis in these assays covering over 20,000 genes, it is challenging to rapidly determine which genes have evidence of clinical relevance. To limit the interpretative burden of reviewing variants from all genes, a well-defined “medical exome” is needed, consisting of genes with sufficient levels of evidence to warrant review in a clinical assay.

There have been efforts to establish highly curated lists of gene-disease associations (GDAs), but these are often small, though highly curated. Most notably, the Clinical Genome Resource (ClinGen) has established a robust framework to determine gene-disease validity through manual assessment of strength of evidence that is used within their multiple disease-specific expert panels and working groups[1]. While these GDAs are well-curated, the intense effort required has limited the breadth of genes currently annotated. On the other end of the spectrum, computational tools, such as DisGeNET, attempt to classify the GDAs of all genes by integrating multiple databases into a single GDA score[2–4]. However, the accuracy and validity of this scoring system has not been assessed. Other efforts have taken the approach of crowd-sourcing and/or collating GDAs, such as Genomics England’s PanelApp and the recently launched Gene Curation Coalition (GenCC), which allow diagnostic gene panels to be shared, downloaded, and evaluated by the scientific community, though they may be limited by the interests and thoroughness of the submitters[5,6].

Generating and maintaining up-to-date gene lists remains challenging since assessing all GDAs is prohibitively time-consuming and evidence supporting new and existing GDAs is continuously generated. Previously published projects from our group, BabySeq and MedSeq, required manual curation resulting in a list of 1,514 and 1,489 GDAs, respectively. In both projects, this was a labor intensive and time-consuming process that is not easily replicated in an efficient manner[7,8]. Therefore, a balance between efficiency and thoroughness is required to make the analysis of genomic data more feasible.

Here, we propose a framework to create gene lists for genomic analyses that balances efficiency, robustness, and accuracy with the ability to be routinely updated with new genes as associations emerge from the literature. This approach generates two lists of disease associated genes based on different levels of evidence (comprehensive and restrictive) to be used in genomic applications.

METHODS

Data Sources Used to Generate the Comprehensive and Restrictive Gene Lists

Extensive databases of gene and/or variant associations, including the Human Gene Mutation Database (HGMD), ClinVar, Online Mendelian Inheritance in Man (OMIM), and DisGeNET, were used to identify genes with any reported GDA [2–4,9–11]. Each data source was also parsed to identify, when applicable, the number of classified variants and their review date, publications, and gene identifiers (Supplemental Methods).

Data Sources Used for Validation of Comprehensive and Restrictive Gene Lists

Data sources incorporated for gene list validation included: 1) 1,490 GDAs evaluated in MedSeq[8]; 2) 1,514 GDAs evaluated in BabySeq[7]; 3) 1,212 gene curations in 995 genes captured by ClinGen as of March 14, 2021[1]; 4) 4,884 GDAs in the Incidentalome and Mendeliome panel from PanelApp Australia (accessed February 26, 2021)[5]; 5) 6,378 GDAs in the Paediatric panel from GenomicsEngland PanelApp (accessed February 26, 2021)[5]; and 6) 2,187 GDAs across 5 laboratories in GenCC (accessed March 13, 2021)[6]. Each dataset included a list of GDAs and their strength of evidence. Classifications used in each dataset and how they map to an overall strength of evidence are provided in Table S1 and defined in Supplementary Methods.

Genome Sequencing and Analysis

Genome sequencing data was generated from 45 individuals undergoing non-indication based genomic screening (Supplemental Methods) with >30X mean coverage and a minimum completeness of >95% of all bases at 15X or higher. Variants were filtered to the comprehensive or restrictive gene lists to identify Pathogenic (P) or Likely Pathogenic (LP) variants (Supplementary Methods). Only genes mapping to GRCh37 were analyzed (Table S2). Evidence for GDAs were manually curated and each GDA was assigned one the following categories: (1) Definitive, (2) Strong, (3) Moderate, or (4) Limited using ClinGen criteria for gene-disease association. Following gene and variant curation using 2015 American College of Medical Genetics and Genomics/Association of Molecular Pathology guidelines[12] with ClinGen rule specifications, only P/LP variants in genes with a strong or definitive GDA were considered reportable.

RESULTS

Generation of Comprehensive and Restrictive Gene Lists

To build a comprehensive gene list for clinical genomic applications, we extracted all genes from extensive datasets meeting any of the following criteria: 1) ≥1 P/LP variant in ClinVar, excluding CNVs overlapping multiple genes; 2) ≥1 variant classified as pathogenic (disease-causing mutation; DM) in HGMD; or 3) listed in Morbid Map from OMIM, excluding susceptibility and non-disease genes (Figure 1A). Following these filters, the comprehensive list included 6,145 genes that have been implicated in Mendelian disease. Of note, 3,825 genes were present in all 3 datasets, with HGMD contributing the most unique genes (Figure 1B).

For many genomic applications, restricting the analysis to genes with stronger disease associations is preferable to reduce the burden on the laboratory. We, thus, further limited the comprehensive list by applying criteria using the number of P/LP variants, the recency of interpretation, and computational predictions for GDAs from DisGeNET. Specifically, only genes fulfilling any of the following criteria were retained: 1) ≥4 P/LP variants in ClinVar evaluated within the last 6 years (2015 or more recently) by any submitter; 2) ≥1 2-star P/LP variant in ClinVar; 3) mitochondrial genes with ≥1 P/LP variant in ClinVar; 4) ≥4 DMs in HGMD with supporting publications within the last 6 years (2015 or more recently); 5) genes with a DisGeNET GDA score ≥0.7. To add more stringency, we filtered this intermediate list to remove genes with lower levels of evidence, only keeping genes that met at least one of the following criteria 1) ≥1 DM in HGMD with a supporting publication within the last 2 years (2019 or more recently), 2) ≥1 P/LP variant with a last evaluated date in ClinVar within the last 2 years (2019 or more recently), or 3) genes with a DisGeNET GDA score ≥0.3. All mitochondrial genes in the intermediate list were also kept at this stage. After applying both sets of filters, a restrictive gene list of 3,929 genes remained, with 3,427 genes present in all original data sources (Figure 1).

Comparing Gene Lists to Previous Curations

To determine the utility of the gene lists and specificity of the filtering strategy, we compared the comprehensive and restrictive lists to manual gene curations, including rigorous expert curations in ClinGen, manual assessments by an individual lab for BabySeq and MedSeq, crowdsourced approaches in PanelApp Australia and GenomicEngland PanelApp, and a consensus-based method from GenCC. When both lists were compared to the 995 genes from ClinGen, we observed that all definitive (655 genes) or strong (20 genes) gene-disease pairs in ClinGen were captured by both lists except for one definitive GDA missing from the restrictive list: the *CD79B* gene associated with Agammaglobulinemia 6. This gene only had 2 P/LP variants in ClinVar and 3 DMs in HGMD. The latest ClinVar submission date was in 2007 and there were no publications after 2015 in HGMD (Table S2). Some gene-disease pairs with limited, disputed, refuted, or no evidence were removed from the comprehensive list (6.2%; 13/210), while many more were removed from the restrictive list (30%; 63/210) (Figure 2A).

Comparing the gene lists to the more rapid assessments of genes in MedSeq or BabySeq[7,8], we observed that all definitive or strong gene-disease pairs classified in both studies (603 genes and 951 genes, respectively) were captured by both lists, except for the strong *RPS15* association with Diamond-Blackfan anemia curated in BabySeq that was not included in either gene list. The GDA between *RPS15* and Diamond-Blackfan anemia was reassessed by the BabySeq team and downgraded to limited due to lack of supporting evidence. The comprehensive and restrictive gene lists also removed 51% (347/680) and 76.3% (519/680), respectively, of genes with insufficient or other classifications in MedSeq (Figure 2B) and 6.2% (13/211) and 29.4% (76/211), respectively, of genes with limited or other classifications in BabySeq (Figure 2C).

Additional analyses were performed using 1) a consensus interpretation from the largest panels of PanelApp Australia and GenomicsEngland PanelApp and 2) a consensus GDA

from GenCC. For the PanelApp analysis, most green-rated genes (1,956 genes) were captured by both lists, except for 41 genes (2.1%) removed from the restrictive list. The relatively high number of green-rated genes excluded from PanelApp in our restrictive list is expected as PanelApp is primarily focused on gene panels for in-depth diagnostic analysis and have not necessarily undergone extensive GDAs using rigorous criteria, such as is used in ClinGen. The gene lists also removed 12.9% (9/70) and 68.6% (48/70) of red genes from the comprehensive list and restrictive list, respectively (Figure 2D). In the GenCC comparison, all Definitive/Strong genes were captured in both lists, except for *SMOC2*, associated with Dentin Dysplasia Type I, that was missing from the restrictive list (Figure 2E). This gene only had 2 DMs with the most recent publication in 2013 and 3 P/LP variants in ClinVar with the most recent evaluation date in 2017.

Genome Sequencing Results Using Different Gene Lists

To determine the performance of the gene lists in practice, genomic data from 45 individuals were screened for reportable variants using both the comprehensive and restrictive gene lists. Following variant filtration for putative P/LP variants, a total of 1,287 variants were identified in the comprehensive list, while only 1,096 were present in the restrictive list, a removal of 191 variants (15%; Figure S1A). Per individual, this equated to an average of 29 (min=14; max=43) and 24 (min=12; max=35) variants in the comprehensive and restrictive lists, respectively. While 58% (402/696) of the genes in the comprehensive list met Strong or Definitive disease association after manual review, this ratio increased to 73% in the restrictive list (402/551). After variant assessment, all reportable variants from the comprehensive list – defined as P/LP associated to a strong or definitive GDA – were also identified in the restrictive list (an average of 3 variants per individual; min=0; max=7) (Figure S1B).

DISCUSSION

Part of an effective and efficient strategy for exome and genome analyses includes defining an appropriate list of genes to interrogate for pathogenic variants. All genes with evidence for a disease association are needed for expanded analyses. However, in different contexts, the level of evidence required for the GDA may vary. For instance, genes with less or emerging evidence of disease association may be useful in a settings where additional familial studies can help determine the likelihood of the gene's responsibility for the individual's disease. However, lists including limited evidence genes will have less utility in the context of genomic screening where the asymptomatic individual will not contribute evidence to the GDA and there is no or very limited utility of returning the result.

Here, we provide a framework that utilizes available databases to efficiently generate both a comprehensive (6,145 genes) and a restrictive list (3,929 genes) of disease-associated genes (Figure 1; Table S1). Compared to ClinGen expert panels, the restrictive gene list excluded 30% of genes with lower levels of evidence, while maintaining all strong or definitively associated genes, aside from one gene with older and borderline evidence (Figure 2). Additionally, using the restrictive gene list in 45 genomes captured all reportable variants, while reducing the number of variants needed to be reviewed by 15%.

Further refinements to this approach can help further reduce the burden of genomic analyses, including utilizing more variant level information in the approach, such as handling variants with discordant classifications and variants whose population frequencies suggest they are too common to be associated with Mendelian disease. However, our current approach is easily implemented and updatable, shows high performance when compared to manually curated datasets, and can provide increased efficiency as genomic applications become more routine.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

Funding support was partly provided by grant 5R01HL143295 from the National Institutes of Health/National Heart, Lung, and Blood Institute (LLV, CLZ, RCG, HLR, MSL). The authors would like to thank the Gene Curation Coalition (GenCC) for generating curated content used in this project. GenCC's curated content was obtained at www.thegencc.org [March 13th, 2021] and includes contributions from the following organizations: Invitae, Illumina, Myriad Women's Health, Ambry Genetics, and TGMI/G2P.

Disclosure:

Dr. Lebo, Dr. Lazo de la Vega, and Ms. Blout Zawatsky reports grants from NIH during the conduct of the study. Dr. Rehm reports grants from NIH during the conduct of the study; she also reports personal fees from Genome Medical outside the submitted work. Dr. Green reports grants from NIH during the conduct of the study; he also reports personal fees from AIA, SavvySherpa, Verily, and Wamberg, all outside the submitted work. Dr. Austin-Tse, Dr. Mason-Suares, Dr. Machini, Dr. Hao and Dr. Yu have nothing to disclose.

DATA AVAILABILITY

The gene lists and data used to develop the lists can be found at https://Broad.io/genelist.

REFERENCES

1. Strande NT, Riggs ER, Buchanan AH, et al. Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am J Hum Genet. 2017;100(6):895–906. [PubMed: 28552198]
2. Piñero J, Queralt-Rosinach N, Bravo À, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015;2015:bav028. [PubMed: 25877637]
3. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833–D839. [PubMed: 27924018]
4. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48(D1):D845–D855. [PubMed: 31680165]
5. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51(11):1560–1565. [PubMed: 31676867]
6. The Gene Curation Coalition. Accessed March 13, 2021. https://thegencc.org/
7. Ceyhan-Birsoy O, Machini K, Lebo MS, et al. A curated gene list for reporting results of newborn genomic sequencing. Genet Med. 2017;19(7):809–818. [PubMed: 28079900]
8. Machini K, Ceyhan-Birsoy O, Azzariti DR, et al. Analyzing and Reanalyzing the Genome: Findings from the MedSeq Project. Am J Hum Genet. 2019;105(1):177–188. [PubMed: 31256874]

9. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–D1067. [PubMed: 29165669]
10. OMIM - Online Mendelian Inheritance in Man. Accessed August 12, 2020. https://www.omim.org/
11. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Hum Genet. 2020;139(10):1197–1207. [PubMed: 32596782]
12. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24. [PubMed: 25741868]

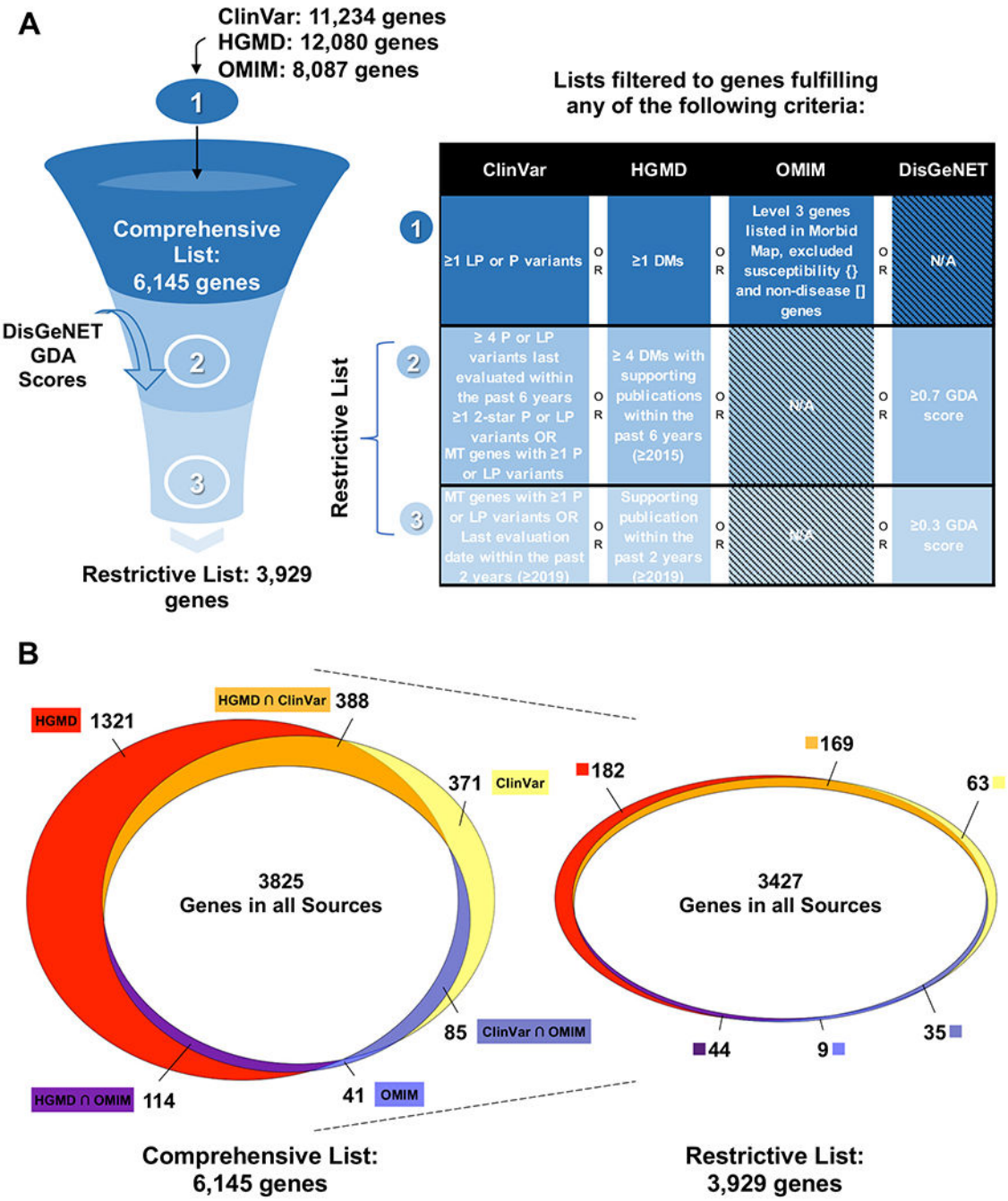

Figure 1:

(A) Schematic of the criteria fulfilled at each stage of the gene filtration process. Genes with entries in ClinVar (11,234 genes), OMIM Morbid Map (8,087 genes), and HGMD (12,080 genes) were integrated to generate the comprehensive and restrictive gene lists. Filtration parameters for each stage are presented in the right panel. (B) Venn diagram of the comprehensive (left) and restrictive (right) gene lists, including the number of genes meeting criteria in the initial databases.

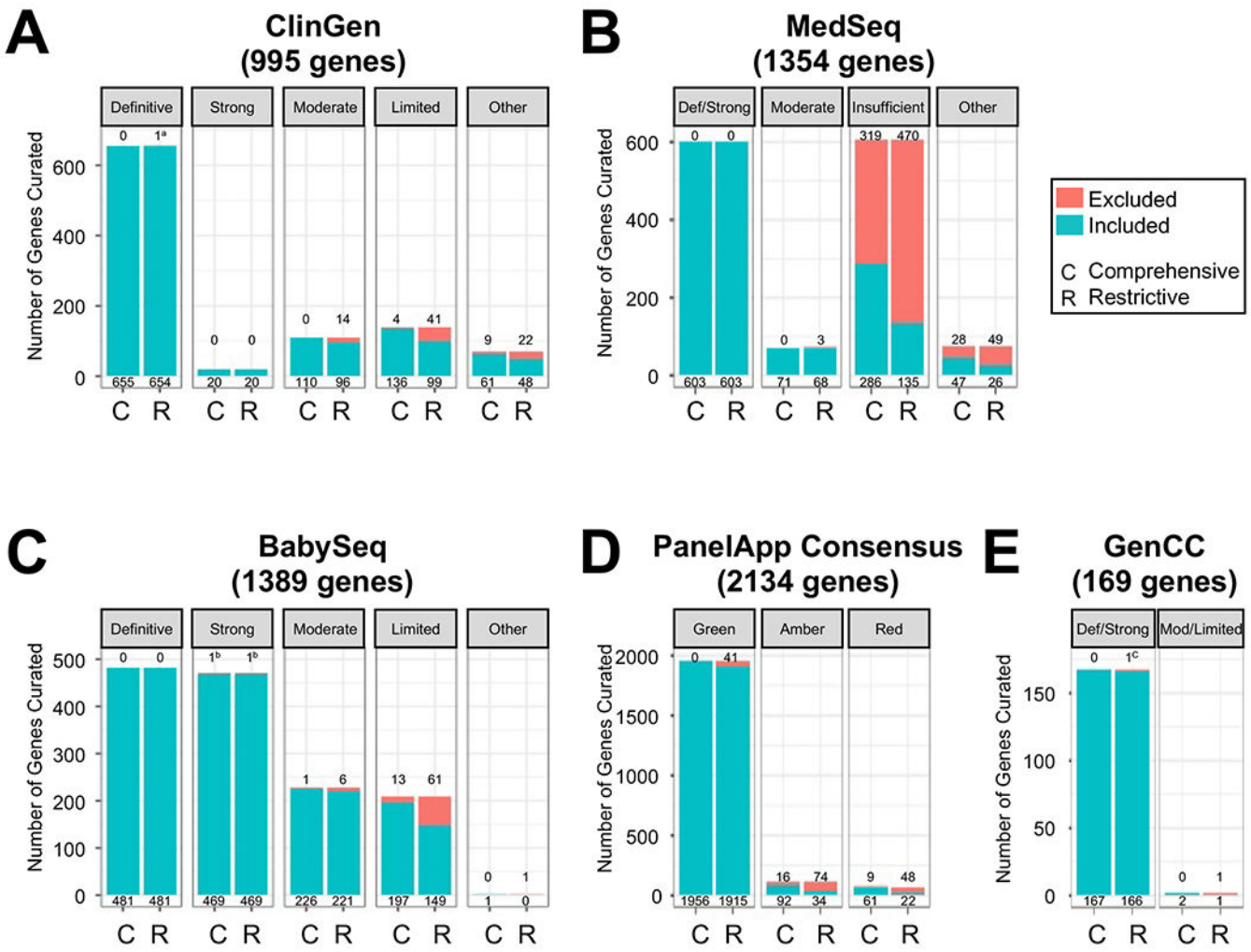

Figure 2:

Comprehensive and restrictive gene lists were compared to the GDA classifications assigned by 6 resources (A) ClinGen, (B) MedSeq, (C) BabySeq, (D) consensus of Australian PanelApp (Incidentalome and Mendeliome panel) and GenomicsEngland PanelApp (Paediatric Panel), and (E) consensus from GenCC. Numbers below the bar represent the number of genes included and numbers above the bar are the number of genes excluded in the respective list. Other: conflicting, refuted, disputed, no reported evidence, trait, pharmacogenomic association, only claim is from GWAS, and does not meet criteria; C: Comprehensive Gene List; R: Restrictive Gene List; [a]*CD79B*; [b]*RPS15*; [c]*SMOC2*