GenIA, the Genetic Immunology Advisor database for inborn errors of immunity

Andrés Caballero-Oteyza, PhD,[a,b,c] Laura Crisponi, PhD,[d]* Xiao P. Peng, MD, PhD,[e]* Kevin Yauy, MD,[f] Stefano Volpi, MD, PhD,[g] Stefano Giardino, MD,[h] Alexandra F. Freeman, MD,[i] Bodo Grimbacher, MD,[c,j,k,l,m] and Michele Proietti, MD, PhD[a,b,c] *Hanover and Freiburg, Germany; Cagliari and Genova, Italy; Baltimore and Bethesda, Md; and Montpellier, France*

Background: To date, no publicly accessible platform has captured and synthesized all of the layered dimensions of genotypic, phenotypic, and mechanistic information published in the field of inborn errors of immunity (IEIs). Such a platform would represent the extensive and complex landscape of IEIs and could increase the rate of diagnosis in patients with a suspected IEI, which remains unacceptably low.
Objective: Our aim was to create an expertly curated, patient-centered, multidimensional IEI database that enables aggregation and sophisticated data interrogation and promotes involvement from diverse stakeholders across the community.
Methods: The database structure was designed following a subject-centered model and written in Structured Query Language (SQL). The web application is written in Hypertext Preprocessor (PHP), Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript. All data stored in the Genetic Immunology Advisor (GenIA) are extracted by manually reviewing published research articles.
Results: We completed data collection and curation for 24 pilot genes. Using these data, we have exemplified how GenIA can provide quick access to structured, longitudinal, more thorough, comprehensive, and up-to-date IEI knowledge than do currently existing databases, such as ClinGen, Human Phenotype Ontology (HPO), ClinVar, or Online Mendelian Inheritance in Man (OMIM), with which GenIA intends to dovetail.
Conclusions: GenIA strives to accurately capture the extensive genetic, mechanistic, and phenotypic heterogeneity found across IEIs, as well as genetic paradigms and diagnostic pitfalls associated with individual genes and conditions. The IEI community's involvement will help promote GenIA as an enduring resource that supports and improves knowledge sharing, research, diagnosis, and care for patients with genetic immune disease. (J Allergy Clin Immunol 2024;153:831-43.)

Key words: *Inborn error of immunity, immune disease, immunogenetics, genotype-phenotype, genetic paradigms, natural history, curation, database, resource, patient-centered*

Since the first description of X-linked agammaglobulinemia in 1952,[1] our understanding of the clinical spectrum of genetic immune disease has expanded beyond primary immunodeficiency to encompass other facets of dysregulated immunity, such as autoimmunity, autoinflammation, allergy/atopy, and increased malignancy risk. In parallel, combining technologic and functional advances has significantly expanded the genetic landscape underlying these clinical presentations. From approximately 100 known monogenic inborn errors of immunity (IEIs) at the turn of the millennia, almost 500 genes and even more associated conditions have now been described.

However, current diagnostic yields remain disappointing for many patients who are strongly suspected of having a genetically driven immune disease. This can be attributed to, among other factors, the difficulty in keeping up with the rapidly growing body of IEI knowledge (no current resource comprehensively catalogs its expanding genotype-phenotype relationships), along with important associated diagnostic paradigms and management details.

Phenotype-driven databases such as Online Mendelian Inheritance in Man (OMIM [www.omim.org]) or the Human Phenotype Ontology (HPO)[2] sacrifice the granularity of allele-specific clinical and mechanistic data, whereas variant- and gene-focused databases such as ClinVar[3] or UniProt (www.uniprot.org) sacrifice many phenotypic dimensions of data. Additionally, the nuances and variety of IEI clinical presentations are such that standard clinical coding systems such as the *International Classification of Diseases, 10th revision* (https://icd.who.int/browse10/2019/en), or HPO[2] remain insufficient for conveying the full extent of phenotypic complexity, even at an individual patient level, much less at the level of cells or families. The availability of detailed immunophenotypic and functional data, in addition to standard clinical and genetic information, adds to the axes of

From [a]the Clinic for Immunology and Rheumatology and [b]the RESiST-Cluster of Excellence 2155, Hanover Medical School; [c]the Institute for Immunodeficiency, Center for Chronic Immunodeficiency, University Hospital Freiburg; [d]the Institute for Genetic and Biomedical Research, The National Research Council, Monserrato, Cagliari; [e]the Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore; [f]the University of Montpellier, LIRMM, CNRS, Reference Center for Congenital Anomalies, Clinical Genetic Unit, Montpellier University Hospital Center; [g]the Center for Autoinflammatory Diseases and Immunodeficiencies, Pediatric Rheumatology Clinic, IRCCS Istituto Giannina Gaslini, Genova, and DINOGMI, Università degli Studi di Genova; [h]the Hematopoietic Stem Cell Transplantation Unit, IRCCS Istituto Giannina Gaslini, Genova; [i]the Laboratory of Clinical Immunology and Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda; [j]the Clinic of Rheumatology and Clinical Immunology, Center for Chronic Immunodeficiency, Medical Center, Faculty of Medicine, and [m]the Centre for Integrative Biological Signalling Studies, Albert-Ludwigs University of Freiburg; and [k]the RESiST-Cluster of Excellence 2155, Hanover Medical School, and [l]the German Center for Infection Research, Satellite Center Freiburg.
*These authors contributed equally to this work.
Received for publication June 27, 2023; revised October 23, 2023; accepted for publication November 15, 2023.
Available online November 30, 2023.
Corresponding author: Michele Proietti, MD, PhD, or Andrés Caballero, PhD, Center for Chronic Immunodeficiency, Breisacher Str 115, University Hospital Freiburg, 79106 Freiburg, Germany E-mail: michele.proietti@uniklinik-freiburg.de. Or: andres.caballero@uniklinik-freiburg.de.
The CrossMark symbol notifies online readers when updates have been made to the article such as errata or minor corrections
0091-6749

https://doi.org/10.1016/j.jaci.2023.11.022

Abbreviations used
GOF: Gain of function
HIES1: Hyper-IgE syndrome 1
HPO: Human Phenotype Ontology
HSCT: Hematopoietic stem cell transplantation
IEI: Inborn error of immunity
IMAD1: Infantile-onset multisystem autoimmune disease 1
IUIS: International Union of Immunological Societies
LOF: Loss of function
OMIM: Online Mendelian Inheritance in Man

data that must be coherently organized. Moreover, IEIs are often associated with unique genetic principles and assumptions that distinguish them from other mendelian conditions. Disease-specific diagnostic- and treatment-related paradigms and pitfalls are often buried in the literature or transmitted by word of mouth but are not widely available in a single searchable repository. Providers are currently faced with the time-intensive and challenging endeavor of gathering and synthesizing information about clinical, genetic, and mechanistic heterogeneity from across scattered resources.

To address these challenges, we designed the Genetic Immunology Advisor (GenIA), a comprehensive, centralized, routinely updated, user-friendly catalog of IEI-focused variant-, gene-, patient-, and disease-specific information. GenIA uses a patient-centered model to link diverse data sets and provide structured, harmonized, and longitudinal information. This model facilitates a more precise, unbiased understanding of the natural history of each IEI without losing sight of each family's or individual's unique form of disease, helping to illuminate observations worthy of further investigation. By piloting the comprehensive curation of 24 genes representing diverse IEIs, we have demonstrated the feasibility, versatility, and value of GenIA by showing how this resource can be used to answer various questions of interest to the IEI community.

METHODS

Database design

Our database is structured around patients as the central node connecting all other data modules (Fig 1) and written using Structured Query Language. It is hosted in an Apache web server with MariaDB, version 10.5.18, installed. Different data sets and data types are currently distributed across 68 interconnected tables, with each of them representing an object or the relationship between objects. Constraints and keys were installed to ensure the compliance and uniqueness of records. We attempted to minimize free-text fields: most columns and attributes in the database tables are designed to hold predefined terms or numbers (integers or decimals) to maximize the query capabilities of the database for future analyses. Whenever possible, we use terms compatible with preexisting ontologies such as those of the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC [www.genenames.org]), Disease Ontology (DO),[4] Experimental Factor Ontology (EFO [www.ebi.ac.uk/efo/]), HPO,[2] the *International Classification of Diseases, 10th revision*, Medical Subject Headings (MeSH [www.nlm.nih.gov/mesh/]), Mondo Disease Ontology (MONDO [mondo.monarchinitiative.org]), National Cancer Institute Thesaurus (NCIT [ncithesaurus.nci.nih.gov/]), Ontology of Adverse Events (OAE),[5] Ontology for Biomedical Investigations (OBI),[6] OMIM, Orphanet Rare Disease Ontology (ORDO), or SNOMED Clinical Terms (SNOMEDCT [bioportal.bioontology.org/ontologies/SNOMEDCT]).

Web application

GenIA can be publicly accessed via a user-friendly web interface (www.geniadb.net) that is programmed using the PHP (Hypertext Preprocessor), HTML (Hypertext Markup Language), CSS (Cascading Style Sheets), and some JQuery libraries from JavaScript for dynamic filtering of records in tables and forms.

Authorized curators log into a separate password-protected portal and navigate a series of structured forms and pull-down menus guiding data entry. Unlike the rest of the web application, this interface was programmed using Bootstrap, version 5.0.2, which is a CSS framework.

Manual curation process

Public online resources and search tools such as PubMed (pubmed.ncbi.nlm.nih.gov), OMIM, or Google Search are used to identify potentially relevant primary publications featuring variants as well as patients and families with IEIs. Searches combine keywords such as the gene or protein name with terms such as *mutation*, *variant*, *patient*, PID, immune dysregulation", "autoimmunity, inflammation, or *immunodeficiency*. Articles in which specific genes or conditions are reviewed or functionally characterized are also considered. We strive to avoid redundant counting of recurrently reported patients. Candidate articles are screened for relevance by scanning the abstract and/or main text.

Inspection is performed gene by gene, article by article, in chronologic order, according to year and month of publication, by a primary curator and a different reviewer. Therefore, patients are generally registered chronologically according to the article in which they were first described. The inspection process (Fig 2) involves the following: (1) registering all index cases and family members to generate family pedigrees; (2) noting ages at study or genetic diagnosis and first manifestation if available; (3) assigning reported genotypes to proband and family members; (4) incorporating all phenotypic (clinical and laboratory) and functional data; and (5) storing the names of corresponding, first, and last authors of the inspected references. Disease entities are cross-checked with the Clinical Genomic Database (CGD [research.nhgri.nih.gov/CGD/]),[7] OMIM, PubMed, and International Union of Immunological Societies (IUIS) nosology.[8] Articles are classified as "inspected" if curation is considered complete or as "inspection pending" if additional information/clarification is needed.

Variant annotation

Before article inspection, all variants are aggregated from OMIM, ClinVar,[3] and original publications found via the aforementioned search algorithms. These variants are imported into the database and annotated using Ensembl's Variant Effect Predictor (VEP) software,[9] version 110. For each variant, curators

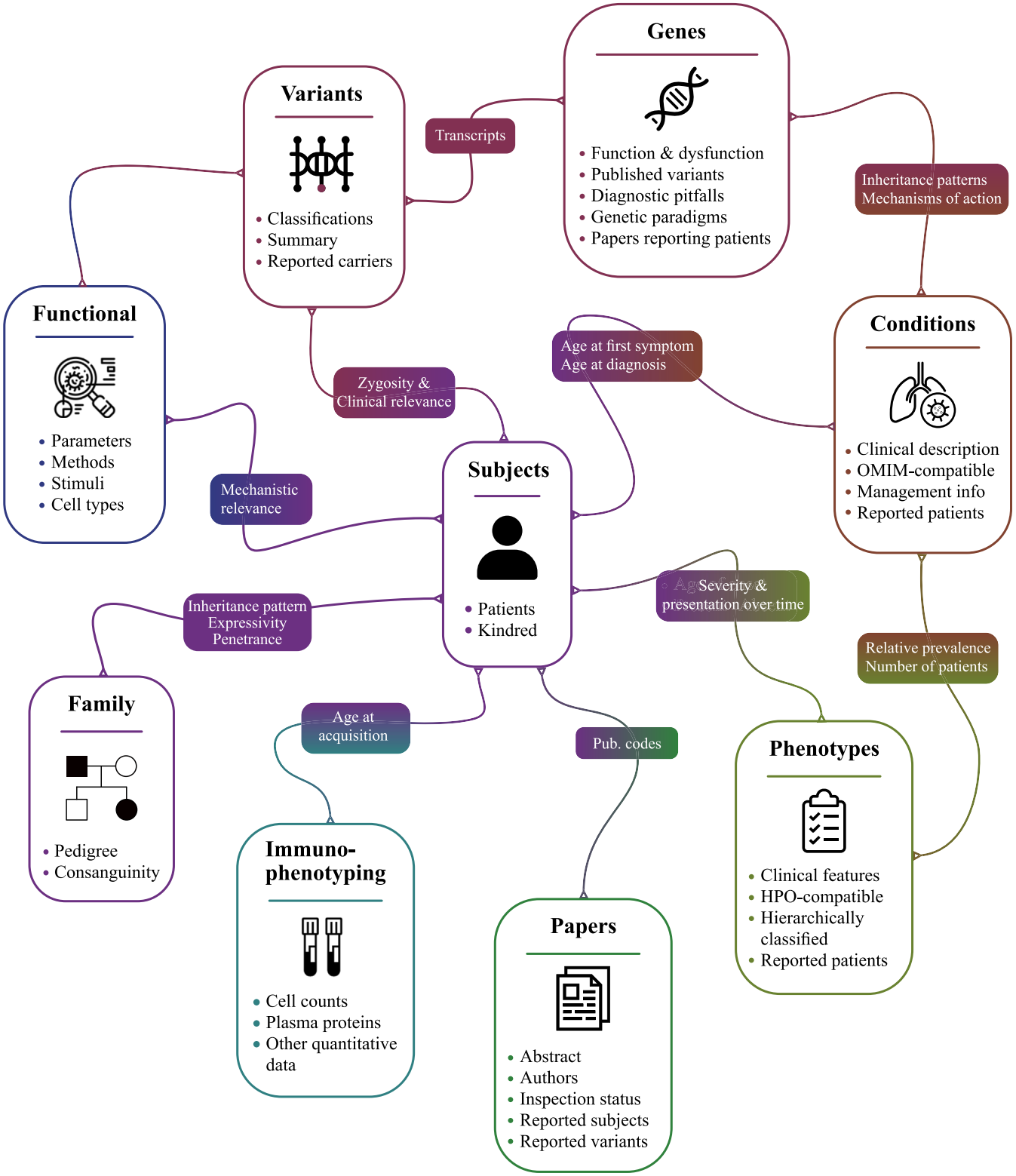

FIG 1. Database design and structure. Simplified graphical representation of the subject-centered model of GenIA's design, showing main database modules and objects and how they are interconnected.

craft a word-limited statement summarizing the variant's predicted pathogenicity classification with associated criteria used, including a detailed description of any functional assay(s) used, based on modified American College of Medical Genetics guidelines. Summaries are updated on a rolling basis whenever new relevant information is published.

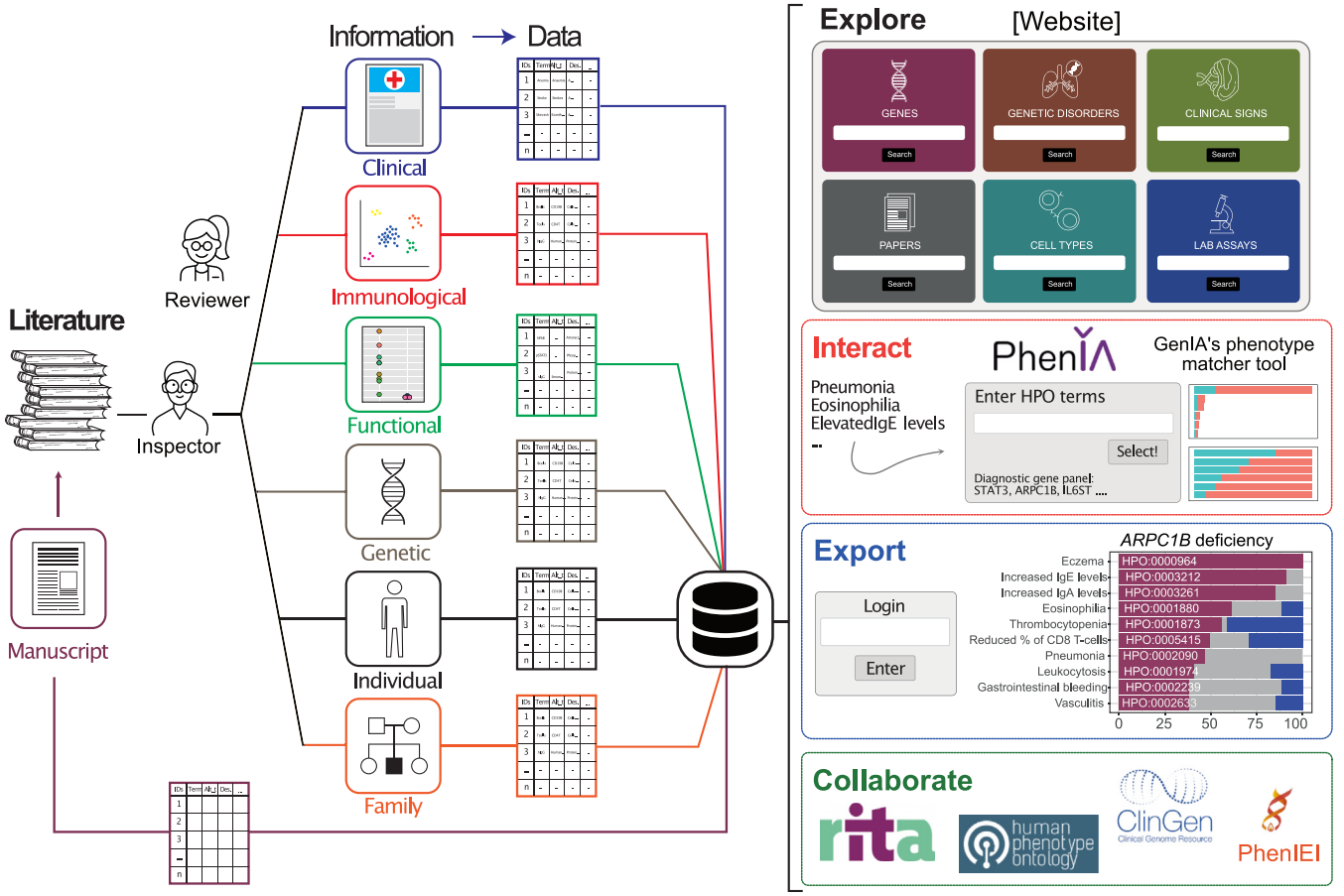

FIG 2. Data inputs and outputs. Schematic representation of the data input and curation process, as represented by the different types of data that are captured, in conjunction with potential outputs and applications of the information collected. All reported data are imported regardless of completeness; we do not exclude published patients on the basis of specific associated data that may be unavailable.

Gene and disease descriptions

Using the aforementioned algorithms, curators craft a word-limited statement summarizing what is currently known about the function of individual genes from the published literature, UniProt, and the Human Protein Atlas. Gene dysfunction in human disease is summarized from the same sources, noting the broad classes of mechanisms and clinical phenotypes linked to each gene. Unique or difficult-to-detect genetic lesions, other known diagnostic challenges, and relevant available immune studies associated with a locus are also searched for and recorded. Each of the clinical phenotypes associated with an IEI gene is also individually assigned a descriptor and acronym, which are currently in the process of being harmonized with those in other databases. Each condition contains a more detailed summary of clinical and laboratory findings for that specific condition, data on all reported patients and families, and any available information on management.

Data mining and extraction

The analysis of the current pilot data was performed using custom commands to query the database and R, version 4.2.1 (2022-06-23), via RStudio for data wrangling and visualization. Graphs were created using the R package ggplot and embedded in figures using Affinity designer software.

RESULTS

Database and graphical user interface development

GenIA was designed to focus on the subject as the central node connecting various data sets (Fig 1), in contrast to the variant-, gene-, or disease-centered model seen in other databases.[3,10] GenIA's graphical user interface is organized into 8 main data modules (Table I[3,11,12]), 6 of which can be accessed directly from the home page (Fig 2) and serve as portals to an interconnected system through which one can reach the others. This modular structure facilitates scalability by allowing the addition of thousands of records per data set (eg, longitudinal instances), the incorporation of data items into existing objects (eg, new laboratory parameter), and the creation of additional modules, such as those on specific diagnosis and management options (currently under development).

GenIA currently contains 68 interconnected tables with more than 500,000 records. Any published subject—proband or affected/unaffected family member—is associated with multiple data types, including genetic data with familial variant segregation, clinical findings, laboratory measurements, and/or functional studies (Fig 1). Additional relevant parameters further strengthen these connections (eg, a genotype is enhanced by the assignment of zygosity and clinical relevance for that specific individual [Fig 1]). In terms of clinical data, we record, whenever available, the age at first clinical manifestation, age at

TABLE I. Detailed description of data modules in GenIA

Module	Type of information included
Genes	● Description of normal gene function ● Description of gene dysfunction associated with human disease states ● Each associated genetic condition is listed with the following: ○ Mode of inheritance ○ Mechanism of action ○ Total number of patients and families reported in the literature and incorporated into GenIA ● All variants in the literature, with the total number of patients carrying each variant ● Gene- and disease-specific diagnostic paradigms and pitfalls to be considered ● References reporting patients and associated relevant or possibly relevant variants in the gene ● Total number of such variants reported per article ● Link to catalog of relevant immune studies available Example: www.geniadb.net/app/gene/info.php?id=36529
Genetic disorders	● Clinical description ● Management options ● Cross-references to other resources and ontologies ● All reported patients and families, with their respective demographics and publication codes ● Presence or absence and number and proportion of reported patients harboring each clinical sign and/or symptom Example: www.geniadb.net/app/disease/info.php?id=141
Clinical signs	● Assigned preferred and alternative clinical terms with associated definitions ● Hierarchically classified using "child" and "parent" terms ● Novel terms introduced if not found in HPO or any of the ontologies mentioned in the Methods section ● All patients in the database reported to present with the indicated clinical finding Example: www.geniadb.net/app/clinterm/info.php?id=25
Articles	● Basic information from the publication (journal, year, abstract, PMID, DOI, etc) ● List of all reported relevant variants ● List of all subjects (probands and affected or unaffected family members) ● List of first, last, and corresponding authors Example: www.geniadb.net/app/ref/info.php?id=607
Cell types	● Preferred and alternative terms, cross-referenced to Cell Ontology ● General description of the cell type ● Hierarchic classification using "child" and "parent" terms ● Normal reference ranges used for cell type of interest in individual research studies, annotated with age, group, sex, population/country Example: www.geniadb.net/app/cell/info.php?id=1
Subjects	● Basic demographic and familial information ● Option to display the subject's family pedigree ● All publication codes used to reference the patient if the patient was reported in the literature more than once ● Molecular diagnosis, age at diagnosis or study, age at the first manifestation ● Reported genetic variants with associated classification, zygosity, and clinical relevance ● Present and absent clinical manifestations, with age at presentation or reporting ● Reported cell counts and immunoglobulin or other plasma protein levels ● Functional assays performed, with associated experimental parameters and outcomes Example: www.geniadb.net/app/subject/info.php?id=101062

(Continued)

TABLE I. *(Continued)*

Module	Type of information included
Variants	● Chromosomal and genetic location, coding, and protein change ● Frequency in healthy population databases such as gnomAD[11] ● Links to external resources such as dbSNP, ClinVar,[3] OMIM, or UniProt ● ACMG-based variant classification ● Variant summary based on synthesis of current literature ● Predicted effects on canonic and noncanonic transcripts ● Reported subjects and families carrying each variant and the variant's zygosity and clinical relevance for each individual ● Functional studies with experimental methods, including assay parameters, cell lines, and stimuli used, and outcomes from each research study in which the variant was tested ● Cell line nomenclature is compatible with Cell Line Ontology[12] Example: www.geniadb.net/app/variant/info.php?id=254&gene=STAT3
Clinical and research studies	● Searchable by gene name, name of test, clinical test code, or parameter to be measured ● Returns a list of all registered clinical- or research-grade descriptive and functional assays ● Each assay is associated with: ○ Description of the methodology used ○ Parameters measured ○ Link to further details on website of laboratory performing test ○ Other available tests for the same gene or condition (Currently populating laboratory studies available in Europe and North America, but with plans to expand to information for labs on other continents) Example: www.geniadb.net/app/assay/info.php?id=8

ACMG, American College of Medical Genetics; *CLO,* Cell Line Ontology; *dbSNP*, Single-Nucleotide Polymorphisms database; *DOI*, digital object identifier; *PMID*, PubMed identifier.

presentation for each sign or symptom, age at genetic diagnosis or age at time of study, age at time of major therapeutic intervention such as hematopoietic stem cell transplantation (HSCT), and age at death. We strive to capture both qualitative and quantitative outcomes of immunophenotypic or functional studies. We also record the age at which a specific test was performed and assay details such as stimuli or cell type(s) used (ie, primary patient cells, cell lines, or model organisms).

Comprehensive immune gene-disease curation

To maximize our ability to capture human disorders associated with IEI genes or considered IEI phenocopies, we combine ongoing manual literature review with automated PubMed searches and cross-check findings against other resources such as OMIM and IUIS nosology.[8] As an ongoing effort, we have already identified more than 490 genes associated with 597 immune-related disorders, including many emerging N-of-1 cases. This significantly exceeds what has been cataloged by OMIM (with 433 genes associated with 516 disorders) or the IUIS (453 genes associated with 507 disorders in 2022) (Fig 3, *A* and see Table E1 in the Online Repository at www.jacionline.org).

Our research shows that dysfunction of these 490 genes is associated with at least 686 different conditions, a minority of which are not known to feature immune phenotypes. The range of clinical conditions associated with an IEI gene is not comprehensively cataloged by either OMIM or IUIS (Fig 3, *B* and see Table E1). Our current data suggest that single genotype-phenotype relationships are predominantly autosomal recessive (76.8%), but monoallelic phenotypes in those same genes may be awaiting discovery (Fig 3, *B* and see Table E1).

GenIA also enables longitudinal visualization of IEI genetic conditions discovered over time, based on incipient publication dates (Fig 3, *C*). This process has now been completed for significantly more than the 24 pilot genes fully curated for this study (see Fig 3, *C* and *D* and see Table E1).

Pilot IEI gene curation

As proof of concept, we completed the curation of 140 publications associated with 24 pilot genes chosen to represent diverse mechanisms and phenotypes (Fig 3, *D*). Some genes are currently associated with only a single publication, whereas others such as *STAT3* are associated with so many genes that we chose to limit curation to the 38 earliest and/or major articles during this pilot study (curation will continue thereafter). This may reflect differences in available evidence for genotype-phenotype relationships, prevalence and incidences of associated conditions, awareness or diagnostic challenges, and/or timing of discoveries. Currently, an additional 50 genes are undergoing curation.

In addition to summarizing what is currently known about gene function(s), mechanisms of gene dysfunction(s) leading to human disease, clinical and cellular phenotypes, potential or proven management strategies, relevant diagnostic challenges (cryptic splicing, structural and noncoding variant detection), and genetic paradigms (digenic/oligogenic inheritance, incomplete penetrance, somatic mosaicism and reversion) are thoroughly investigated for each gene and recorded as well. For example, 4 of our 24 pilot genes (*ARPC1B,*[13] *IL6ST,*[14] *STAT3,*[15] and *TLR8*[16]) were found to be associated with "somatic mosaicism and/or reversion."

We imported data from 899 affected individuals from 607 families for the 24 completely curated genes. Genetic information was available and recorded for 878 individuals (98%), clinical

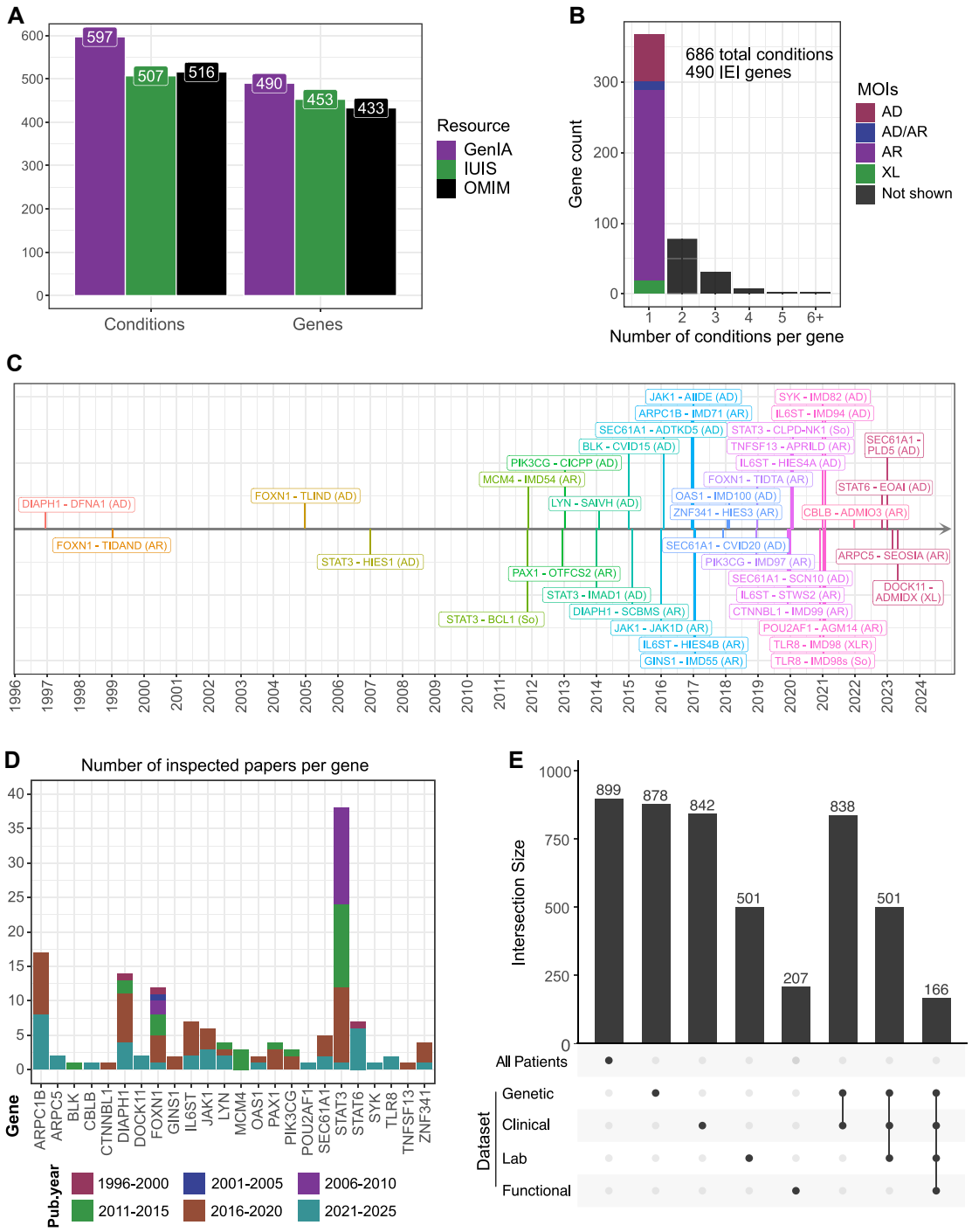

FIG 3. Pilot gene curation. **A,** Comparison of the total number of genes and associated IEI conditions cataloged by GenIA, the 2022 IUIS classification, and OMIM. **B,** Modes of inheritance (MOIs) and total number of genetic conditions (non-IEI disorders included) associated with 490 "IEI" genes. Disctint MOIs are color-coded, AD (autosomal dominant), AR (autosomal recessive), XL (X-linked). **C,** Time line showing the year in which a specific genetic condition for the 24 currently curated genes set was first described in a research article. Conditions are colored according to discovery year. The same gene may appear more than once if associated with multiple conditions. **D,** Number of research articles that have been fully inspected, grouped by gene and colored by publication year, for the set of 24 currently curated genes. **E,** Upset plot showing the overlap between available data sets (genetic, clinical, laboratory, and functional) for more than 850 imported patients harboring mutations in any of the 24 pilot genes.

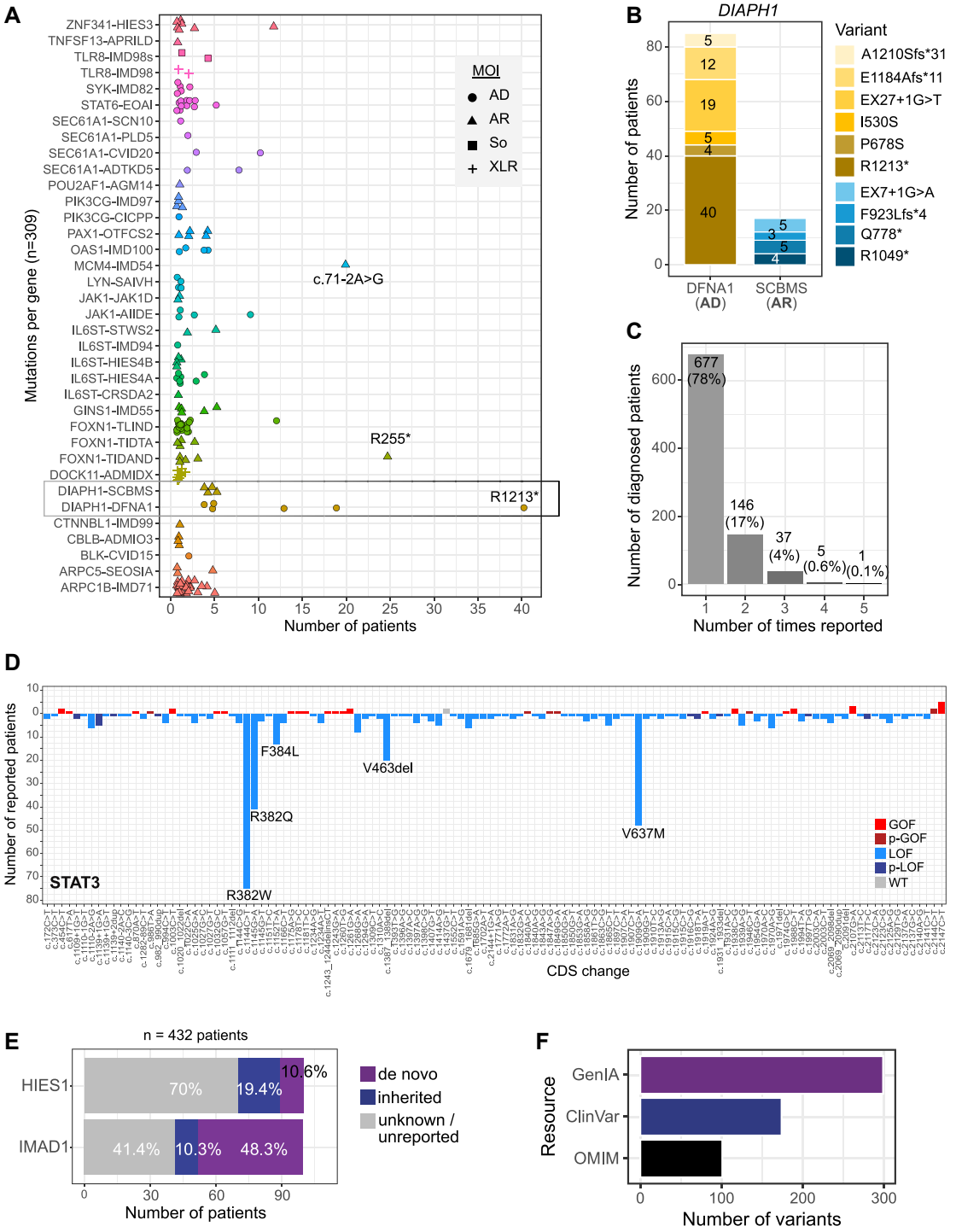

FIG 4. Curation of pilot gene variants. **A,** Number of patients reported per published variant grouped and colored by gene or genetic condition for the 24 pilot genes (see Table E2); *STAT3* is shown separately in (**D**). **B,** Relative frequency of each variant for the gene *DIAPH1* grouped by the associated genetic condition (DFNA1 and SCBMS). **C,** Number of times the same genetically diagnosed patient is reported in the literature (using only patients reported in the inspected articles shown in Fig 3, *D*). **D,** Total patient distribution for all reported *STAT3*-GOF (experimentally confirmed [*light red*]) or *STAT3*-p-GOF (presumed GOF [*dark red*]) variants versus *STAT3*-LOF (confirmed [*light blue*]) or p-LOF (presumed LOF [*dark blue*]) variants versus isomorphic variants (*gray*). LOF includes hypomorphic or amorphic dominant negative or potential haploinsufficiency mutations (see Table E3). **E,** Relative distribution of 432 patients with *STAT3* disease-causing variants by mode of inheritance, grouped by genetic condition (HIES1 and IMAD1). **F,** Total number of variants published for the 24 pilot genes and registered in GenIA compared with those registered in ClinVar or OMIM (see Table E2). *AD*, Autosomal dominant; *AR*, autosomal recessive; *p-GOF*, predicted GOF; *p-LOF*, predicted LOF; *WT*, wild type.

information for 842 (94%), immunophenotypic information for 501 (56%) and functional data for 230 (23%) (Fig 3, *E*). Genetic and clinical information is available for 838 patients (93%), whereas detailed immunophenotypic data are additionally available for 501 (56%), with 166 (18%) also having functional data (Fig 3, *E*).

Variant curation for probands and families

All variants identified as disease-causing or disease-associated variants in a gene are first registered in the database, then annotated and classified as described in Methods, and ultimately associated with patients and publications. This enables us to accurately calculate the total number of reported individuals carrying each variant (Fig 4, *A*, *B*, and *D*). To avoid overestimation, we investigate and note any patients who are reported in the literature more than once. Of the 866 patients with molecular diagnoses recorded during the pilot gene curation process, we identified at least 189 (22%) who have been reported at least twice, with some found up to 4 or 5 times in different articles (Fig 4, *C*).

Approximately 93% of all reported variants in the 24 pilot genes (288 of 309) are found in 5 or fewer patients each, but recurrent variants are seen more frequently in some genes (eg, *STAT3, DIAPH1*, *FOXN1*) (Fig 4, *A*, *B*, and *D* and see Table E2 in the Online Repository at www.jacionline.org). Some of these may be founder mutations, such as *FOXN1* p.R255* in Italian patients,[17] *MCM4* c.71-2A>G in the Irish Traveller community,[18-20] and *DIAPH1* p.R1213* (reported in European, Japanese, and North American populations); others may localize to protein motifs or domains that are mutational hotspots (Fig 4, *D*). Of the 192 distinct *STAT3* variants from the 38 inspected articles currently registered in GenIA, 32 are associated with gain-of-function (GOF) effects leading to infantile-onset multisystem autoimmune disease 1 (IMAD1), and 156 are associated with loss-of-function (LOF) effects predominantly leading to hyper-IgE syndrome (HIES)-like phenotypes (Fig 4, *D* and see Table E3 in the Online Repository at www.jacionline.org). We recognize the ongoing work aimed at clarifying potential haploinsufficient versus dominant negative LOF mutations and the possibility that these represent distinct clinical entities, but we have grouped them here to avoid confusion. Notably, 5 pathogenic variants (p.R382W, p.R382Q, p.F384L, p.V463del, and p.V637M) account for nearly half of all individuals with *STAT3* HIES (197 of 425), with multiple substitutions at residues R382, F384, and V637 known to be disease causing.

After the probands, we then register all reported family member data, using pedigrees to depict the familial segregation of genotypes with phenotypes (see Fig E1 in the Online Repository at www.jacionline.org). This also provides information regarding intrafamilial phenotypic variability, incomplete penetrance, or prevalence of inherited versus *de novo* conditions (as shown for *STAT3*, the latter is significantly more common for GOF than DN mutations [Fig 4, *E*]). Lastly, a comparative analysis shows that our dedicated resource holds more reported variants than do other commonly used resources with a more generalist approach (Fig 4, *F*).

Longitudinal clinical, immunophenotypic, and functional data

Patient-specific data curation involves extraction and assignment of clinical terms, genetic data, laboratory data (including immunophenotyping), and results from functional assays, along with age obtained if available (Fig 3, *E*). Wherever possible, we capture the following: (1) age at first clinical manifestation (age of onset), (2) age at clinical and/or molecular/genetic diagnosis or age at time of study if the former is unknown, (3) age at major therapeutic interventions such as HSCT, and (4) age at death (Fig 5, *C*). All data points captured per patient thus far are summarized in Fig E2, *A-D* (see the Online Repository at www.jacionline.org), with longitudinal data plotted for a patient with *SYK* GOF (see Fig E2, *E*).

This enables us to compare expected ages of onset for different genetic conditions (Fig 5, *A*), as well as times to diagnosis and treatment (Fig 5, *B*). Very broad ranges of diagnostic delays can be seen for some genetically heterogeneous conditions such as HIES. We hypothesize that this is partly influenced by the timing of a condition's discovery and would thus expect to see "time to diagnoses" downtrending over time. Interestingly, this appears to be true for patients with HIES 1 (HIES1), but with the current data, it is not seen as clearly for *STAT3* GOF (IMAD1) (see Fig E3, *A* in the Online Repository at www.jacionline.org). Delays in diagnosis and treatment—and whether the condition required HSCT or led to early death—can also be appreciated at the individual patient level for each condition (taking ARPC1B deficiency as an example) (Fig 5, *C*). Using the age at death of all deceased patients, we could also estimate the life expectancy distribution for patients with a specific genetic condition (see Fig E3, *B*).

We also distinguish between features reported as absent versus those not queried, enabling us to more accurately estimate prevalences for individual clinical findings and even potentially draw conclusions about age-dependent penetrance. In the case of ARPC1B deficiency, of the more than 180 distinct clinical terms that we identified, many are reported by OMIM but without information on the frequency of occurrence, whereas most are simply absent from HPO (Fig 6, *A* and *B*). For genes associated with more than 1 genetic condition (such as *STAT3*), it is possible to compare the relative distribution of clinical manifestations across distinct conditions to show where phenotypes overlap or diverge (Fig 6, *C-E*).

Clinical diagnostic applications

To help clinicians use GenIA for triaging differential diagnoses for patients whose condition has not been definitively diagnosed, we developed a Shiny app, GenIA PhenoMatcher (https://geniadb.shinyapps.io/phenomatcher/), which generates lists of candidate genes, genetic conditions, and the number and percentage of known patients for a given input of clinical manifestations (Fig 2). For instance, among the 24 pilot genes, *STAT3, IL6ST, ARPC1B, ZNF341, STAT6, SYK,* and *PAX1* are the genes recommended for consideration if a patient presents with the combined features of "recurrent pneumonia plus elevated IgE levels plus peripheral eosinophilia" (Fig 6, *F*). Additionally, given the significant incomplete and age-dependent penetrance associated with IEIs, the longitudinal data found in GenIA can help providers evaluate whether the age at which their patient presented with specific symptoms is consistent with the known natural history for the suspected condition (Fig 6, *G*).

GenIA can be further exploited to enhance the utility of other applications, such as the PhenIEI (https://github.com/kyauy/PhenIEI) phenotype matching system. Preliminary analysis using our pilot gene data shows that GenIA significantly augments the symptom-gene associations stored in this database (1288

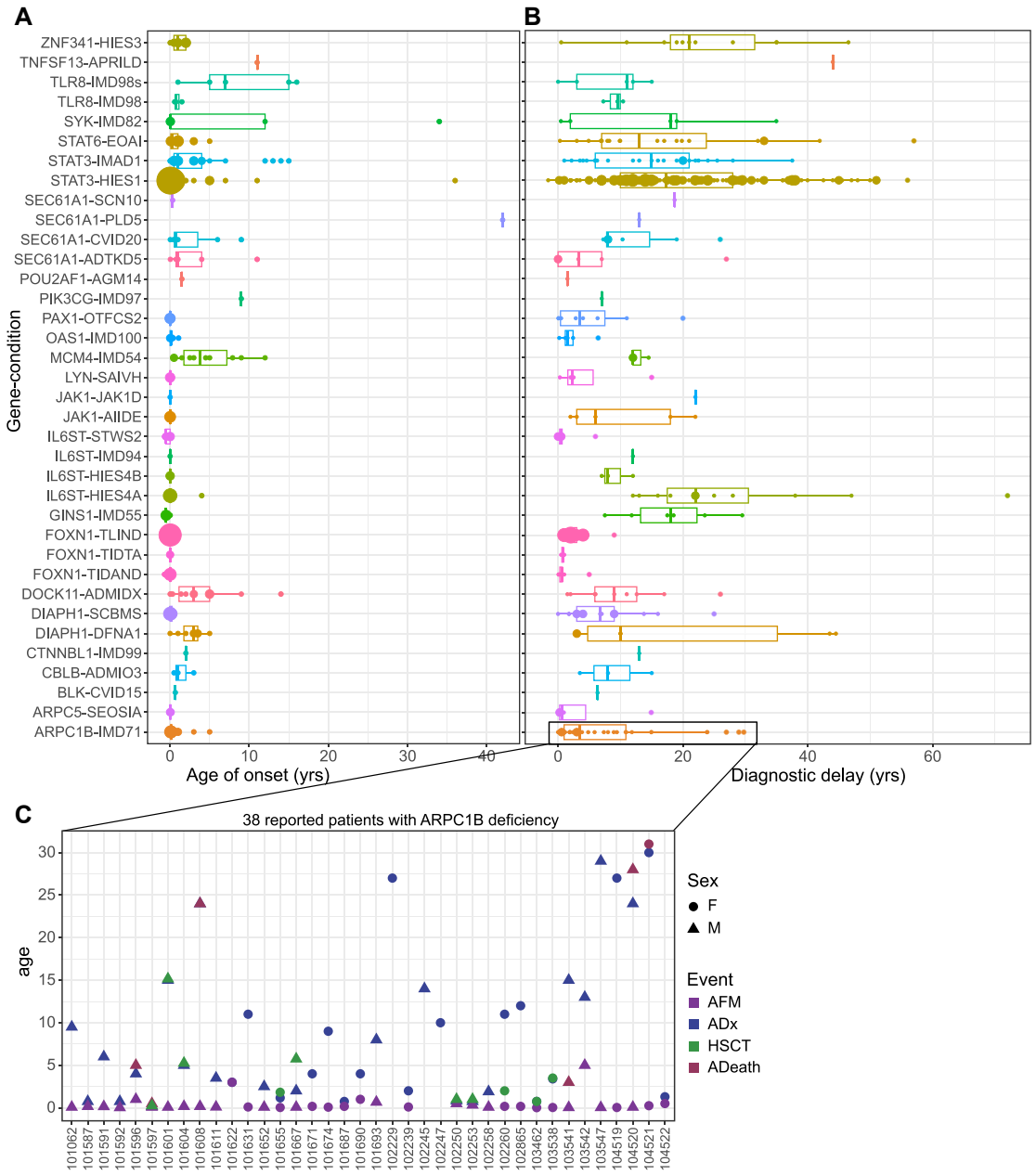

FIG 5. Age of onset, diagnostic delay, and inheritance of genetic conditions. **A,** Box plots showing ages of onset for patients grouped and colored by genetic condition (showing only conditions associated with the 24 pilot genes). **B,** Box plots showing the diagnostic delay (age at study/genetic diagnosis minus the age at first manifestation) for patients with different genetic conditions. Dot size is correlated with the number of patients. **C,** Age at first manifestation (AFM), molecular diagnosis (ADx), HSCT (hematopoietic stem cell transplant), and/or death (ADeath) for patients with ARPC1B deficiency. *F,* Female; *M,* male.

previously unreported associations from IUIS and HPO) (see Fig E4, *A* in the Online Repository at www.jacionline.org). GenIA's contribution of these additional clinical inputs improves PhenIEI-based differential diagnosis triage, as shown with use of the large IEI cohort reported in 2022 by Similuk et al[21] as a test set (see Fig E4, *B*).

DISCUSSION

Among the challenges that we face in IEI research and clinical care is the extensive and expanding genotype-phenotype landscape with its wealth of associated information. Thus far, this has not been systematically cataloged in a single, easily curated and searchable database; therefore, we developed GenIA as a resource to tackle this challenge.

The flexibly of GenIA's unique subject-centered model accommodates various data types and classes (Fig 1). For example, comprehensive genotype-phenotype data are not typically found in phenotype-driven catalogs, whereas in-depth information about zygosity, clinical relevance, phenotypes, mechanisms, and management is typically unavailable from large variant databases. GenIA integrates these and other forms of data to help

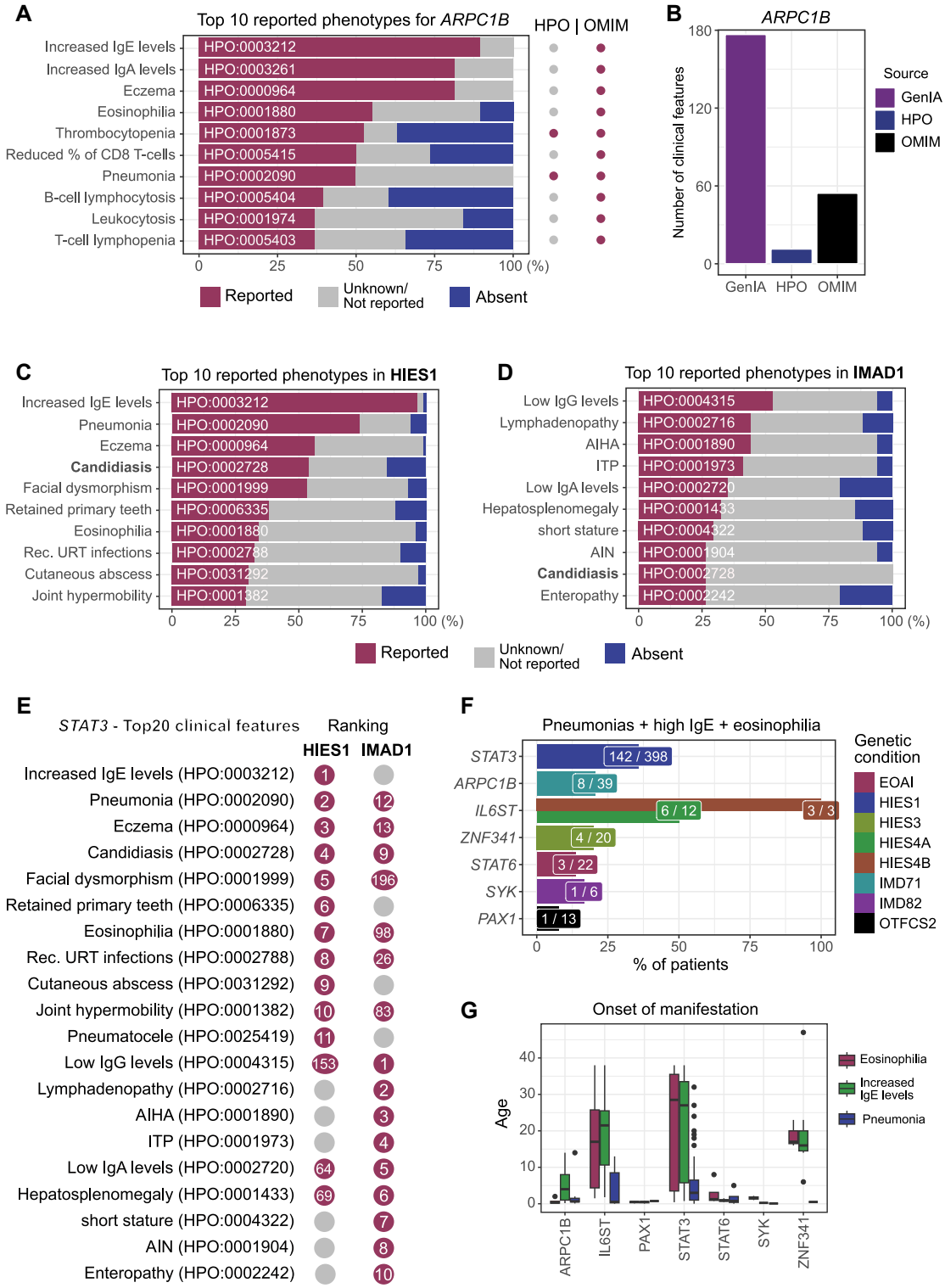

FIG 6. Phenotypic curation and data mining. **A,** Top 10 clinical terms (phenotypes) reported for 38 patients with ARPC1B deficiency with relative frequency in GenIA compared to the HPO database or OMIM. **B,** Total number of clinical terms reported for all patients with ARPC1B deficiency in GenIA versus in HPO or OMIM. **C** and **D,** Comparison of top 10 clinical manifestations of patients with *STAT3* mutations associated with either HIES1 or IMAD1. **E,** Ranking of the most commonly reported phenotypes in patients with STAT3-associated HIES1 or IMAD1, including the top 11 features for each condition. **F,** Different genetic conditions of the patients found in GenIA and matching the combination of the 3 indicated clinical manifestations. **G,** Age distribution at which patients were reported to show the same manifestations as in (**F**), grouped by gene defect. *AIHA,* Autoimmune hemolytic anemia; *AIN,* autoimmune neutropenia; *ITP,* immune thrombocytopenia; *URT,* upper respiratory tract.

providers draw valuable conclusions about a condition's clinical and genetic landscape, diagnostic and management challenges, transmission, penetrance, and expressivity. The intention of creating GenIA is not to replace but rather to dovetail and harmonize with existing resources and endeavors, such as OMIM, ClinVar, HPO, and ClinGen.[22] Some preexisting resources curate more multifaceted and detailed data but focus only on 1 or a few conditions (ie, CFTR2 [https://cftr2.org/], Infevers [https://infevers.umai-montpellier.fr]). Although we acknowledge the challenges of extending this level of detail and completeness to a much larger group of disorders, we strongly believe that this can be accomplished—particularly with involvement from the IEI community. We are also investigating the possibility of increasing efficiency by adapting emerging machine learning and artificial intelligence strategies for use in some modules, but without losing the nuances of human biology or responsible scrutiny.

By achieving complete curation of 24 pilot IEI genes in less than 1 year, with each curator working on a volunteer basis, we have demonstrated how our current curation interface and protocol renders our goals for GenIA realistic and achievable. Indeed, while preparing this article, we completed curation of 4 of the 24 genes (*ARPC5*, *DOCK11, LYN,* and *STAT6*). We have used these pilot genes to demonstrate how the data stored in GenIA can be mined and synthesized to ask important clinical and biologic questions or identify specific concerns to be addressed. Incomplete and unequal data availability across articles is unavoidable, even for our pilot cohort (Fig 3, *E*). Not only are publications limited by content regulations but they may also lack patient-specific comprehensive examination or longitudinal follow-up data. Moreover, some currently used assays may not yet have been developed at the time of study. Fortunately, we were able to extract detailed clinical data for the vast majority of genetically confirmed patients registered thus far, enabling us to generate fairly accurate calculations for the prevalence of clinical features associated with specific genes (Fig 6, *A* and *B*) and conditions (Fig 6, *C* and *D*).

The richness and quality of data in GenIA are useful to diverse stakeholders in the IEI community. Its integrated design facilitates detailed natural history studies of disease progression or investigations into novel therapeutics. GenIA can also serve as a powerful tool for those interested in elucidating genotype-phenotype correlations, conducting mechanistic studies, curating gene-disease relationships, developing variant classification guidelines, or performing diagnostic studies (Fig 2, *A*). GenIA also supports the work of laboratories developing diagnostic assays or performing genetic analysis and/or reanalysis (Fig 3, *C*) by providing a reliable and comprehensive resource for organizing and triaging IEI conditions, genes, and variants. Clinical providers can use GenIA's catalog of IEI-specific diagnostic studies and challenges to strategize immunophenotypic workup or further evaluate variants identified on genetic testing. Additionally, GenIA has the ability to enhance the power of other applications such as PhenIEI, and at the same time, our user-friendly PhenoMatcher tool facilitates efficient formulation of differential diagnoses. Finally, GenIA can also be used as a resource to promote collaborations within the IEI community by identifying researchers working on specific genes and conditions associated with the curated publications (see Table E4 in the Online Repository at www.jacionline.org).

Most importantly, we invite other stakeholders in the field to join our endeavor so that it remains truly a collaborative resource made for and by the IEI community. There are many ways to participate; they include but are not limited to (1) curating a gene of interest, (2) collaborating to reduce redundant curation when writing articles, (3) informing us about available functional studies and/or assay data for resolution of VUSes (variants of uncertain significance), (4) sharing published data before an article becomes publicly available, (5) alerting us to missing genotype-phenotype relationships, (6) offering critical expertise and insight on what has already been curated, or (7) even just replying to our e-mail inquiries seeking collateral information on publications. We are actively in discussions with other groups such as the Clinical Immunology Society, European Society of Immunodeficiencies, ClinGen, and HPO about how to synergize efforts, but we would greatly welcome involvement from other institutions, societies, and individuals. Our collective effort can build an enduring resource that supports and improves knowledge sharing, research, diagnosis, and care for patients with genetic immune disease.

DISCLOSURE STATEMENT

This work was supported by the Center for Chronic Immunodeficiency, Freiburg Center for Rare Diseases, as well as by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy — EXC 2155 (project 390874280 [to M.P. and B.G.]), and the German Federal Ministry of Education and Research through a grant to the German Auto-Immunity Network (grant 01GM1910A [to B.G.]) M.P. also receives funding from the Deutsche Forschungsgemeinschaft (under Transregio 359 PILOT (to M.P.) and the Fritz Thyssen Foundation (grant 10.18.1.039MN [to M.P.]), B.G. also receives support by the Deutsche Forschungsgemeinschaft CRC grant IMPATH SFB1160/2_B5 and under Germany's Excellence Strategy (CIBSS 2189 project ID 390939984 [to B.G.]); the E-rare program of the European Union, managed by the Deutsche Forschungsgemeinschaft (grant GR1617/14-1/iPAD [to B.G.]), and the European Union EU-H2020-MSCA-COFUND EURIdoc programme (project 101034170 [to B.G.]). We acknowledge the RESIST Cluster of Excellence Medizinische Hochschule Hannover for their full support for the Open Access Publication of this article.

Disclosure of potential conflict of interest: The authors declare that they have no relevant conflicts of interest.

We thank Dr Ivan Chinn, Dr Sergio Rosenzweig, and the Clinical Immunology Society (CIS), as well as Jessica Quinn, scientific director at the Jeffrey Modell Foundation, for their advice, encouragement, and support. We thank Dr James Verbsky and the CIS Diagnostic Lab Committee for choosing to use GenIA as the platform for storing and updating information about clinically available diagnostic immune studies in North America. We thank Dr Emmanuelle Jouanguy for her help clarifying the genetic diagnosis of a French family with GINS1 deficiency that was initially suspected to have MCM4 deficiency. We thank Dr Jacques G. Rivière, Dr Laura Alonso García, Dr Maruša Debeljak, and Dr Gašper Markelj for their help clarifying information on families with ARPC1B deficiency. We thank Dr Cristina Glocker (née Woellner) and Dr Ayse Metin for their help with regard to the patients with *STAT3* who were included in her thesis and their publications. We thank Dr Sara Sebnem Kilik and Dr Ben-Zion Garty for their input regarding reference values used for immunoglobulin and lymphocyte subpopulations in Turkish and Israeli patients with *ZNF341* mutations.

Clinical implications: GenIA is a collaborative, user-friendly public resource that centralizes and integrates all aspects of genetic immune disease information with the goal of helping to improve diagnosis and management.

REFERENCES

1. Buckley CR. Agammaglobulinemia, by Col. Ogden C. Bruton, MC, USA, Pediatrics, 1952;9:722-728. Pediatrics 1998;102:213-5.
2. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res 2021;49:D1207-17.
3. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42:D980-5.
4. Schriml LM, Munro JB, Schor M, Olley D, McCracken C, Felix V, et al. The human disease ontology 2022 update. Nucleic Acids Res 2022;50:D1255-61.
5. He Y, Sarntivijai S, Lin Y, Xiang Z, Guo A, Zhang S, et al. OAE: The ontology of adverse events. J Biomed Semantics 2014;5:29.
6. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The ontology for biomedical investigations. PLoS One 2016;11:e0154556.
7. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. Proc Natl Acad Sci U S A 2013;110:9851-5.
8. Tangye SG, Al-Herz W, Bousfiha A, Cunningham-Rundles C, Franco JL, Holland SM, et al. Human inborn errors of immunity: 2022 update on the classification from the International Union of Immunological Societies expert committee. J Clin Immunol 2022;42:1473-507.
9. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol 2016;17:122.
10. Fokkema IFAC, Kroon M, López Hernández JA, Asscheman D, Lugtenburg I, Hoogenboom J, et al. The LOVD3 platform: efficient genome-wide sharing of genetic variants. Eur J Hum Genet 2021;29:1796-803.
11. Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat Commun 2020;11:1-13.
12. Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, et al. CLO: the cell line ontology. J Biomed Semantics 2014;5:37.
13. Brigida I, Zoccolillo M, Cicalese MP, Pfajfer L, Barzaghi F, Scala S, et al. T-cell defects in patients with germline mutations account for combined immunodeficiency. Blood 2018;132:2362-74.
14. Materna-Kiryluk A, Pollak A, Gawalski K, Szczawinska-Poplonyk A, Rydzynska Z, Sosnowska A, et al. Mosaic IL6ST variant inducing constitutive GP130 cytokine receptor signaling as a cause of neonatal onset immunodeficiency with autoinflammation and dysmorphy. Hum Mol Genet 2021;30:226-33.
15. Hsu AP, Sowerwine KJ, Lawrence MG, Davis J, Henderson CJ, Zarember KA, et al. Intermediate phenotypes in patients with autosomal dominant hyper-IgE syndrome caused by somatic mosaicism. J Allergy Clin Immunol 2013;131:1586-93.
16. Aluri J, Bach A, Kaviany S, Chiquetto Paracatu L, Kitcharoensakkul M, Walkiewicz MA, et al. Immunodeficiency and bone marrow failure with mosaic and germline TLR8 gain of function. Blood 2021;137:2450-62.
17. Bosticardo M, Yamazaki Y, Cowan J, Giardino G, Corsino C, Scalia G, et al. Heterozygous FOXN1 variants cause low TRECs and severe T cell lymphopenia, revealing a crucial role of FOXN1 in supporting early thymopoiesis. Am J Hum Genet 2019;105:549-61.
18. Gineau L, Cognet C, Kara N, Lach FP, Dunne J, Veturi U, et al. Partial MCM4 deficiency in patients with growth retardation, adrenal insufficiency, and natural killer cell deficiency. J Clin Invest 2012;122:821-32.
19. Hughes CR, Guasti L, Meimaridou E, Chuang CH, Schimenti JC, King PJ, et al. MCM4 mutation causes adrenal failure, short stature, and natural killer cell deficiency in humans. J Clin Invest 2012;122:814-20.
20. Casey JP, Nobbs M, McGettigan P, Lynch S, Ennis S. Recessive mutations in MCM4/PRKDC cause a novel syndrome involving a primary immunodeficiency and a disorder of DNA repair. J Med Genet 2012;49:242-5.
21. Similuk MN, Yan J, Ghosh R, Oler AJ, Franco LM, Setzer MR, et al. Clinical exome sequencing of 1000 families with complex immune phenotypes: toward comprehensive genomic evaluations. J Allergy Clin Immunol 2022;150:947-54.
22. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen – the clinical genome resource. N Engl J Med 2015;372:2235-42.