ARTICLE | Genetics inMedicine

Open

The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system

Kenneth D. Mandl, MD, MPH[1,2,3], Tracy Glauser, MD[4,5], Ian D. Krantz, MD[6,7], Paul Avillach, MD, PhD[1,2], Anna Bartels, MScPH[8], Alan H. Beggs, PhD[3,9,10], Sawona Biswas, MS, CGC[6], Florence T. Bourgeois, MD, MPH[1,3,10], Jeremy Corsmo, MPH[5,11], Andrew Dauber, MD, MMSc[5,12], Batsal Devkota, PhD[13], Gary R. Fleisher, MD[3,10], Allison P. Heath, PhD[13], Ingo Helbig, MD[7,14,15], Joel N. Hirschhorn, MD, PhD[3,16,17], Judson Kilbourn[1], Sek Won Kong, MD[1,3], Susan Kornetsky, MPH[18], Joseph A. Majzoub, MD[3,10,16], Keith Marsolo, PhD[19], Lisa J. Martin, PhD[5,20], Jeremy Nix, BS[21], Amy Schwarzhoff, MS, MBA[23], Jason Stedman, BA[2], Arnold Strauss, MD[5,24], Kristen L. Sund, PhD, MS[20], Deanne M. Taylor, PhD[7,14], Peter S. White, PhD[5,21,22], Eric Marsh, MD[25], Adda Grimberg, MD[26], Colin Hawkes, MD, PhD[26], the Genomics Research and Innovation Network

Purpose: Clinicians and researchers must contextualize a patient's genetic variants against population-based references with detailed phenotyping. We sought to establish globally scalable technology, policy, and procedures for sharing biosamples and associated genomic and phenotypic data on broadly consented cohorts, across sites of care.

Methods: Three of the nation's leading children's hospitals launched the Genomic Research and Innovation Network (GRIN), with federated information technology infrastructure, harmonized biobanking protocols, and material transfer agreements. Pilot studies in epilepsy and short stature were completed to design and test the collaboration model.

Results: Harmonized, broadly consented institutional review board (IRB) protocols were approved and used for biobank enrollment, creating ever-expanding, compatible biobanks. An open source federated query infrastructure was established over genotype–phenotype databases at the three hospitals. Investigators securely access the GRIN platform for prep to research queries, receiving aggregate counts of patients with particular phenotypes or genotypes in each biobank. With proper approvals, de-identified data is exported to a shared analytic workspace. Investigators at all sites enthusiastically collaborated on the pilot studies, resulting in multiple publications. Investigators have also begun to successfully utilize the infrastructure for grant applications.

Conclusions: The GRIN collaboration establishes the technology, policy, and procedures for a scalable genomic research network.

Genetics in Medicine (2020) 22:371–380; https://doi.org/10.1038/s41436-019-0646-3

Keywords: genomic medicine; federated networks; electronic health records; biobanking; information technology

INTRODUCTION

Clinicians and researchers interpreting variants must contextualize a patient's genetic variants against population-based references along with detailed phenotyping to meaningfully assess the significance and impact on prognosis based on the care trajectories and outcomes of other patients with the same or related variants. The benefits of incorporating genomics into clinical care can only be realized through the interpretation of large-scale data as applied to the individual. While for cancer, genomics guided care is becoming more commonplace, for many heritable conditions, the necessary empiric data are lacking. Any one hospital or academic center may struggle to identify sufficient numbers of cases to enable discovery, especially for rare conditions. These challenges necessitate large-scale collaborations so that clinicians and researchers can meaningfully assess the

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA; [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; [3]Department of Pediatrics, Harvard Medical School, Boston, MA, USA; [4]Division of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; [5]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA; [6]Division of Human Genetics at the Children's Hospital of Philadelphia, Philadelphia, PA, USA; [7]Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA; [8]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; [9]The Manton Center for Orphan Disease Research, Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA; [10]Department of Pediatrics, Boston Children's Hospital, Boston, MA, USA; [11]Office of Research Compliance and Regulatory Affairs, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; [12]Division of Endocrinology, Children's National Health System, Washington, DC, USA; [13]Center for Data-Driven Discovery in Biomedicine, The Children's Hospital of Philadelphia, Philadelphia, PA, USA; [14]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA; [15]Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [16]Division of Endocrinology, Boston Children's Hospital, Boston, MA, USA; [17]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [18]Research Administration, Boston Children's Hospital, Boston, MA, USA; [19]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, USA; [20]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; [21]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; [22]Department of Biomedical Informatics, University of Cincinnati College of Medicine, Cincinnati, OH, USA; [23]Children's Hospital of Philadelphia, Philadelphia, PA, USA; [24]Division of Cardiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; [25]Division of Neurology, The Children's Hospital of Philadelphia, The Perelman School of Medicine at The University of Pennsylvania, Philadelphia, PA, USA; [26]Division of Endocrinology and Diabetes, The Children's Hospital of Philadelphia, The Perelman School of Medicine at The University of Pennsylvania, Philadelphia, PA, USA. Correspondence: Kenneth D. Mandl (kenneth_mandl@harvard.edu)

These three authors contributed equally: Kenneth D. Mandl, Tracy Glauser and Ian D Krantz

Genomics Research and Innovation Network Members are listed at the end of the paper.

Submitted 27 April 2019; accepted: 20 August 2019
Published online: 4 September 2019

significance and impact on prognosis and selection of therapeutic regimens.

To catalyze research and advance clinical care for both rare and common diseases, three of the leading US children's hospitals have partnered to create the Genomics Research and Innovation Network (GRIN). This network leverages a combined patient population of diverse backgrounds with unparalleled representation across the spectrum of pediatric diseases, along with the ability to consent participants through uniform biobanking protocols. GRIN members have designed a scalable system equipping hospitals to contribute to a distributed genotype–phenotype database of sufficient size and variability to address a broad range of clinical and research questions. While the main focus of GRIN in its initial iteration is to establish a superlative pediatric data resource, it will not be limited to pediatric diseases or centers in its broader scope. A desired end state is to allow hospitals to leverage their existing care delivery processes and information technology (IT) structures to acquire and share digital health record data, biospecimens, and a range of omics measurements that are made during the course of clinical care or under research protocols.

Over the past four and a half years, through a process of intensive focus, alignment of cultures and approaches, and commitment of internal resources, GRIN has enabled the production of a generalizable approach to collaboration—from regulatory, technical, and cultural perspectives. Our efforts have also been designed to surmount typically encountered institutional barriers to large-scale data democratization. As such, this initiative aligns well with the FAIR data principles (findability, accessibility, interoperability, reproducibility) recently promoted by the National Institutes of Health and other funding agencies.[1] We describe the structure of this collaboration and the current state of our information technology and regulatory infrastructure, which has been constructed with an eye toward interoperability and scalability. We believe GRIN is ready for others to join and build upon. GRIN central access is available through a web portal.[2]

MATERIALS AND METHODS

Collaboration structure and commitment

Launched as an investigator-driven initiative at the three hospitals, with funding provided by the three institutions' CEOs, GRIN takes a policy and governance approach that builds on local frameworks but ensures a cohesive model that supports deep interinstitutional integration at the level of data sharing and collaboration. A first step was to design the agreements to facilitate efficient and scalable collaboration. The establishment of an organizational structure across the three founding GRIN institutions included six working groups (Biomedical Informatics, Regulatory, Scientific, Legal, Sustainability, and Participant Engagement) that allowed not only an efficient compartmentalized approach to tackling the tasks at hand but also engendered a culture of trust and mutual respect between key investigators and stakeholders across the three institutions through close collaborative working relationships (Fig. 1). Key drivers included calls with institutional leadership, and annual meetings that often included the three Department of Pediatrics chairs.

Biobanking IRB protocol harmonization

Each institution already had an established biobank for sample collection from broadly consented participants under a local institutional review board (IRB) protocol, but with different restrictions on what types of phenotypic data could be used for research, and with whom it could be shared. With the end goal of biobank harmonization across the network, the local biobank protocols and their associated consents were modified to maximize broad sharing of data and samples. This included using specific key elements that enabled seamless integration, including (1) obtaining participant consents empowering investigators to access all clinical information in the medical record, to use samples for genomic research, and to use samples and data for any type of research, regardless of the participant's underlying condition; (2) securing permission to share de-identified and limited data sets of phenotypic and genomic data and to share biospecimens across the institutions and, under acceptable governance, with other outside collaborators (e.g., pharmaceutical companies); (3) getting permission to recontact participants to request additional samples and/or data collection for extended phenotype and to offer enrollment in additional studies; (4) ensuring ability to recontact participants for medically actionable results, in accordance with local procedures.

Federated, queryable data infrastructure

We approached IT design with several desiderata in mind. First, we sought to have the majority of data, initially genomic and electronic health record (EHR) data, remain local at the sites, but combinable upon demand under proper authority.[3] Secondly, we wanted all sites to participate without being compelled to abandon legacy infrastructure. We surmised that building upon local investment in technology, policies, and workflow optimization at each of the three hospitals to build a common infrastructure would lower barriers to joining the network. Thirdly, to promote maintainability, self-direction, and scaling, we sought to use a mesh of modular open source and free software components, each with vibrant user and software developer communities, that have spread virally and scaled across heterogeneous systems.

To date, there has been federal investment of hundreds of millions of dollars in distributed research networks, including SHRINE,[4,5] Sentinel,[6,7] and PCORNet,[8] and in genomics consortia using EHR data, including the Electronic Medical Records and Genomics Network (eMERGE).[9,10] There has not yet been comparable investment in a genomic research data network, connected to health system data, that is distributed, queryable, API-driven, and readily extensible, and none support local processes and innovation. Federation

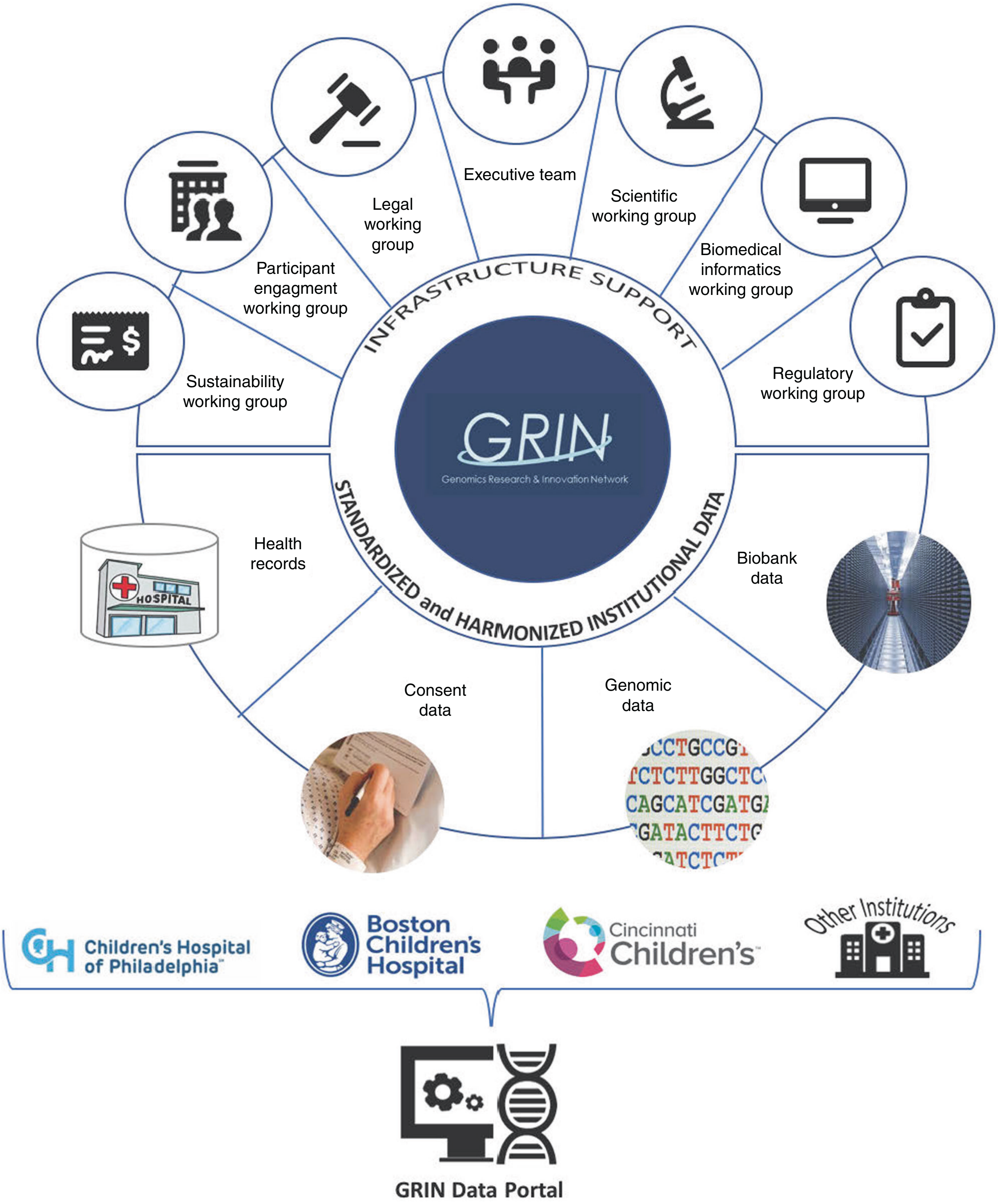

Executive team
Legal working group
Scientific working group
Participant engagment working group
Biomedical informatics working group
Sustainability working group
Regulatory working group
INFRASTRUCTURE SUPPORT
GRIN
Genomics Research & Innovation Network
STANDARDIZED and HARMONIZED INSTITUTIONAL DATA
Health records
HOSPITAL
Biobank data
Consent data
Genomic data
Children's Hospital of Philadelphia
Boston Children's Hospital
Cincinnati Children's
Other Institutions
GRIN Data Portal

Fig. 1 Collaboration structure for the Genomics Research and Innovation Network.

allows data from disparate databases and other sources to be aggregated ad hoc as a virtual database that can be used for aggregated and patient-level analysis. Factors arguing in favor of federated databases include the need for recognition of creators of data and databases, valid limitations on data sharing, a desire to encourage widespread innovation, and the value of domain-specific experts close to data curation and management processes.[11] Importantly, this approach capacitates a rapidly scaling, federated network despite wide variability in IT at each institution.

Pilot studies

To drive system design, we solicited pilot proposals involving investigators from all three hospitals. These pilot studies served as proof of concept for the network design and infrastructure, and also served as examples of how a collaborative network structure can positively influence cooperation among researchers with closely shared interests. The projects were chosen to exemplify the kind of studies that GRIN was designed to support—namely, medical conditions with a strong expected genetic component, relative rarity but sufficient cohort sizes at each institution, availability of relevant clinical data in the EHRs, and project leaders at each of the three institutions willing to collaborate and bring complementary expertise to the table.

RESULTS

Sharing agreements

A set of guiding principles was established setting forth the agreed-upon relationship among the existing GRIN institutions. A basic GRIN tenet is to establish process reciprocity and interoperability whenever possible. Rather than executing separate material transfer agreements (MTAs) to facilitate appropriate sharing of biospecimens consented for research, and phenotypic and genomic data for each research project and participating institution, a single, overarching MTA need only be signed once by participating investigators when joining GRIN. Later, transferred specimens and data are electronically tracked across all participating institutions by GRIN-authorized project management and informatics staff. We further executed an IRB reliance agreement across the three hospitals, enabling a single IRB review for protocols. This early agreement has been superseded by the National Center for Advancing Translational Sciences (NCATS) SMART IRB[12] policies and procedures.

We note that a technical framework to support collaboration is necessary but not sufficient—an academic framework that enables and sustains collaboration is also essential. To this end, we also created a sharing document that serves as a memorandum of understanding of basic rules and principles of collaboration. Several of these were adapted from successful academic consortia doing genetic studies such as the Genetic Investigation of Anthropometric Traits (GIANT) consortium. One of the key principles, “no surprises,” requires tracking of proposed projects to provide the opportunity for GRIN investigators to know what projects are ongoing within GRIN. This tracking allows at a minimum close coordination between related projects, and full collaboration is encouraged wherever possible. A second related principle laid out in the sharing document is “don’t use other people’s data against them to gain a competitive advantage.” The sharing document offers guidance on structure of projects, approaches to developing analysis plans, data sharing and access, publication and authorship, interinstitutional communication, and conflict resolution. To date the network has been accessible to a core group of investigators at each institution, working on the pilot projects. As we open the network widely, access is becoming available to investigators employed at one of the three hospitals who have completed appropriate Collaborative Institutional Training Initiative (CITI) IRB training. Upon first accessing the web portal, authorized users are asked to agree to terms of service. Enforcement of the collaboration principles relies on the capacity to audit system access and upon the institutional obligations of employed investigators.

IRB protocols

We have incorporated the key elements for seamless integration into each institutional biobank consent form using locally acceptable formats and language. Our aligned consenting approach is designed such that de-identified clinical data and associated biospecimens and data can be shared within and across GRIN sites, as well as with industry, commercial collaborators, and other outside collaborators. All data sharing and research use is fully compliant with Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rules. Specifically, participants consent to research involving their protected health information (PHI) within their institution. Data sharing across GRIN sites involves either aggregate results, de-identified data, or limited data sets accompanied by a data use agreement (DUA). Currently, sharing limited data sets requires a project-specific DUA attached to a single human subjects protocol with a lead IRB and reliance agreement. All data sharing requires data use agreements. The consent form may be administered using electronic platforms and signatures, in which case a written copy of the consent is provided to the participant who provides a signature for consent. Research use of the broadly consented EHR, genomic or sample data, if identified, requires a subsequent IRB protocol and approval by local data and sample access committees.

The consent is fully compliant with requirements under the revised NIH Common Rule[13] with respect to handling of personal health information, use of biospecimens for different research types, and provision to participants of clinically relevant results. In the future, to further ensure participant understanding of the research, a concise summary with key study information will be added to the beginning of the consent.

Federated infrastructure

To enable a distributed query of local databases, we adapted the PIC-SURE application programming interface (API).[14] The PIC-SURE API was funded by the NIH’s Big Data to Knowledge Program (BD2K) and, as of 2019, is the only meta-API and variable-level API in the NIH Data Commons and DATA stage effort.[15] The API exposes patient-level and aggregate clinical and genomic data via query abstraction that is structurally consistent across resources. A meta-API blends inconsistency across various APIs by providing use case–focused and descriptive metadata and interactions. PIC-SURE was initially implemented to align all available biomedical data per individual. We subsequently extended the API to address common genotypic and phenotypic

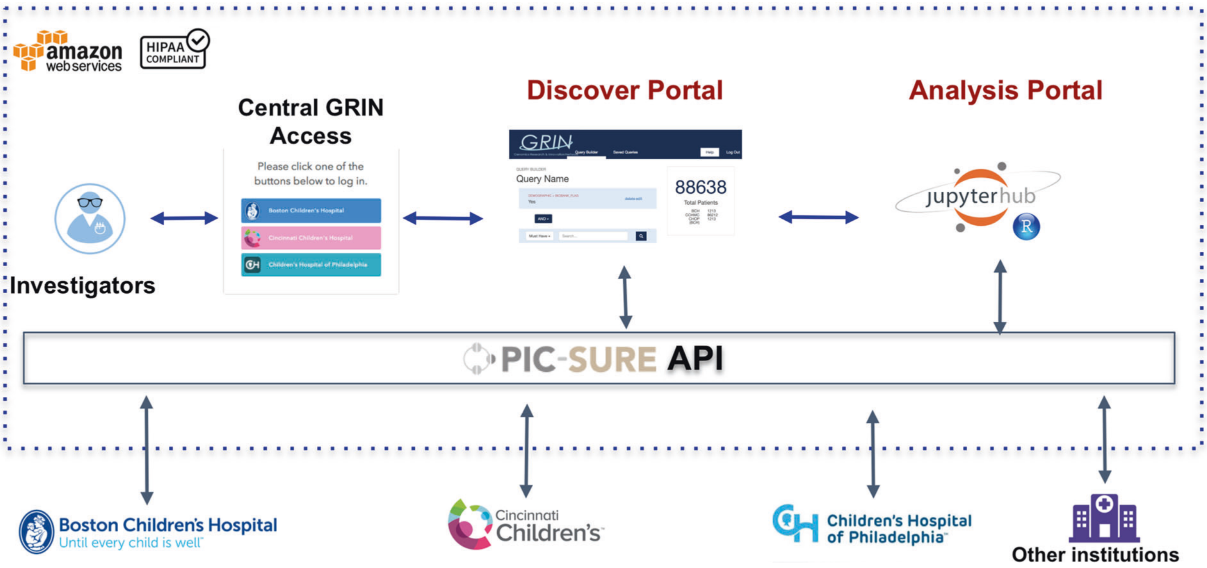

amazon web services
HIPAA COMPLIANT
Central GRIN Access
Please click one of the buttons below to log in.
Boston Children's Hospital
Cincinnati Children's Hospital
Children's Hospital of Philadelphia
Discover Portal
GRIN
Query Name
88638
Total Patients
Analysis Portal
jupyterhub
R
Investigators
PIC-SURE API
Boston Children's Hospital
Until every child is well
Cincinnati Children's
Children's Hospital of Philadelphia
Other institutions

Fig. 2 The PIC-SURE application programming interface (API) is used to access genotype and phenotype data from databases at each site of care. Currently the databases are i2b2/TranSMART instances, but the API is agnostic. Authorized investigators from the three institutions can log in with their standard hospital credentials at the Genomics Research and Innovation Network (GRIN) Central Access Portal. Investigators can interrogate data all three hospitals using the discover portal, which returns aggregate counts by institution. With proper institutional review board (IRB) authorization, they can access line-level de-identified data for exploratory analyses using i2b2/TranSMART, or export line-level data to the analysis portal, an Amazon Web Services (AWS)–hosted environment shared across the three institutions.

queries against each institution's EHR and exome/genome variant sequence repositories (Fig. 2).

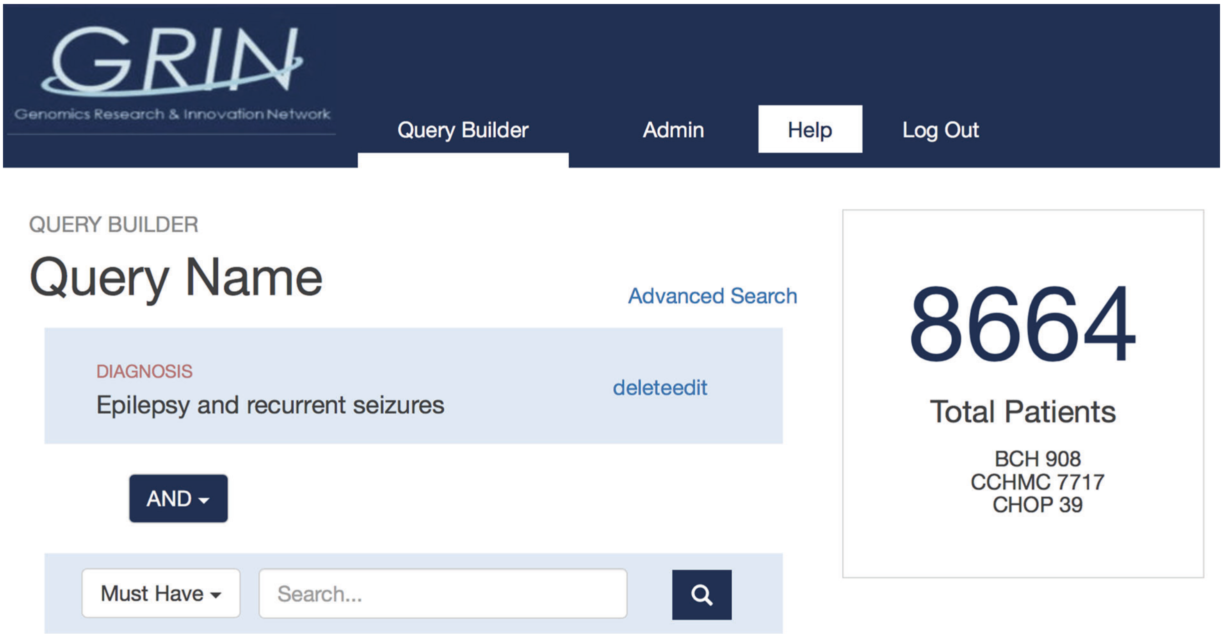
PIC-SURE is designed to integrate patient-level genetic, environmental, behavioral, imaging, and clinical data from distributed sources, and has been recently interfaced over multiple dedicated web portals.[16–20]

Our technical approach to developing the federated infrastructure included five major steps. Step 1: We agreed to use the model and vocabularies underlying the PCORNet Common Data Model[8] for common clinical data elements. PCORNet was selected because of its wide deployment across over 100 hospitals and institutions. We also developed an agreement to use genomic variant call format (gVCF) for sharing exome and genome sequencing results that are generated using a harmonized pipeline.[21] There is active discussion to incorporate other data models and ontologies, including Health Level Seven's Fast Health Interoperability Resources (FHIR) and the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) data model. Step 2: We exposed the PIC-SURE API on EHR and genomic repositories across the three GRIN hospitals, leveraging resource interfaces built directly into the API codebase.[22] Though our initial local database platforms were based on the i2b2 and i2b2/tranSMART[23,24] open source data analytic platforms, PIC-SURE can be adapted to diverse resources. Step 3: To ensure secure access by only authorized investigators, we implemented authentication and authorization for all API access. Web-based access to this infrastructure is available to all investigators at each site who (1) are employed by the institution, (2) have completed the institutional specific standard courses in human subjects research from CITI,[25] and (3) have signed the GRIN terms of service agreement. Step 4: We designed both a simple and a detailed user interface (UI) to query across the three sites to identify patients and cohorts by phenotype and/or genotype (Figs. 3 and 4). Users must be within one of the three hospital networks to log in and use the query tool. The simple interface was designed using a formal participant-centered design sprint methodology.[26] (The Undiagnosed Disease Network has adapted the GRIN simple interface for its portal). Step 5: Using Amazon Web Services (AWS), we designed a data provisioning workflow so that data could be readily combined across sites, in a common, HIPAA-compliant secure workspace with analysis toolkits including Jupyter Notebooks, R statistical programming tools, and Python.

Pilot studies

To ground development of GRIN in real world driving biology projects, we completed investigations focused on severe epilepsies and extreme short stature. The hospitals funded protocol preparation, enrollment, sample collection, DNA extraction and sequencing, and preparation of EHR data. The epilepsy pilot demonstrates the value of aligned collaborative structures and data types in rapidly generating data and manuscripts. The short stature pilot demonstrates the use of an early version of the IT infrastructure, and informed its subsequent refinement and development, and led to rapid generation of preliminary data for a successful R01.

The epilepsy pilot, which sought to characterize rare and resistant epileptic encephalopathies, served as a catalyst for collaborative studies on genotype–phenotype correlations as well as computational phenotyping efforts enhanced by the



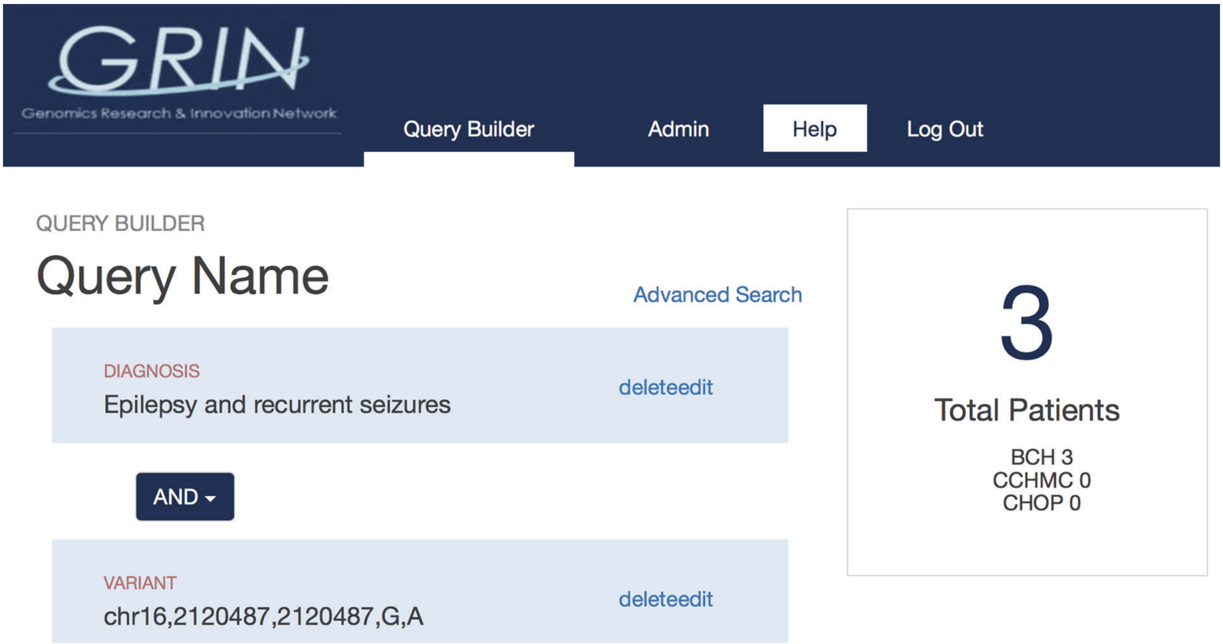
QUERY BUILDER

Query Name

DIAGNOSIS
Epilepsy and recurrent seizures

8664
Total Patients
BCH 908
CCHMC 7717
CHOP 39

Fig. 3 The discovery user interface (UI) finding patients across the hospital with a diagnosis of epilepsy and recurrent seizures. The figure illustrates not only the power of distributed query, but also the nature of a modular, scalable federated network, in that the three hospitals are at different stages of data contribution. The Cincinnati Children's Hospital Medical Center has made its full corpus of electronic health record (EHR) data available for query. The Children's Hospital of Philadelphia and Boston Children's Hospital have only made data available for consented biobank cohorts. Note—both of the latter hospitals are committed to making the full EHR cohort available during 2019.



QUERY BUILDER

Query Name

DIAGNOSIS
Epilepsy and recurrent seizures

VARIANT
chr16,2120487,2120487,G,A

3
Total Patients
BCH 3
CCHMC 0
CHOP 0

Fig. 4 The discovery user interface (UI) finding three patients with a specific variant at one of the hospitals. With proper approvals, samples, sequence, electronic health record data, or recontact can be requested by any investigator at the hospitals.

data standardization efforts within GRIN. The GRIN Consortium funded and facilitated trio exome sequencing for 10 trios per site, which was subsequently used for cohort expansion and harmonization across a larger, joint trio cohort of more than 200 patient–parent trios with existing deep phenotype data. Where possible, future GRIN projects will use a harmonized pipeline to standardize variant calling and annotation.

The work resulted in the first description of epilepsy due to missense variants in *GABRG2*,[27] and the first description of de novo variants of *CACNA1E* as a cause of severe developmental and epileptic encephalopathies.[28] In addition, the harmonized GRIN IRB protocol enabled contribution to the largest genetic studies in the field, including the most recent genome-wide association study (GWAS) by the International League Against Epilepsy (ILAE)[29] and a large burden analysis performed on more than 17,000 individuals with epilepsy. The GRIN support for these contributions removed institutional barriers for data sharing and provided infrastructure for effective patient recruitment and phenotyping. Challenges faced in sharing data for joint analyses by investigators across the three institutions led to the design of the current HIPAA-compliant, cloud-hosted platform.

In addition to the collaborative studies catalyzed by GRIN, the data standardization allowed the epilepsy pilot project investigators to systematically develop and apply concepts for data harmonization for epilepsy phenotypes, overcoming the "phenotypic bottleneck" that has long prevented the inclusion of deep phenotypic data into genetic studies.[30] The relative lack of phenotypic information stems from the fact that while genetic information can be generated at scale through exome and genome sequencing efforts, phenotyping remains a manual, nonscalable task. This has resulted in genetic studies that are relatively impoverished in phenotypic information compared with the depth of genotyping. The GRIN epilepsy pilot investigators went on to lead international efforts for harmonization of gene curation criteria with respect to phenotypic information and the expansion of common phenotypic formats such as the Monarch ontology for common presentations of neurodevelopmental disorders.[31] Finally, the GRIN epilepsy pilot study contributed to the first gene discovery in neurodevelopmental disorders based on harmonized and standardized phenotypic information, identifying de novo variants in *AP2M1* through a phenotypic similarity analysis based on Human Phenotype Ontology (HPO) terms across the entire GRIN epilepsy cohort and further cohorts.[32]

The goal of the short stature pilot project was to identify specific rare subphenotypes of short stature based on clinical characteristics that are readily identifiable as discrete data elements within the EHR. As a proof of principle, we sought to enroll a cohort of patients with apparent resistance to insulin-like growth factor 1 (IGF-1), a key mediator of growth hormone action. We hypothesized that patients with apparent IGF-1 resistance represent a cohort of individuals that is enriched for monogenic causes of short stature. To identify these individuals, we generated queries of the EHR looking for patients with heights below -2 standard deviations and IGF-1 levels greater than the 90th percentile when adjusted for age and sex. In total, we were only able to successfully identify and recruit ten patients who met these criteria across all three sites pointing to the extreme rarity of this condition and the need for collaborative recruitment across multiple sites. Using exome sequencing, we did identify a genetic etiology in three of the ten subjects.

Challenges arising during the short stature pilot project informed the design of the IT infrastructure, which was being simultaneously developed. At the outset of the pilot, a single query was not yet able to be run at all three institutions. Two of the sites did use the same EHR platform (i2b2), which facilitated more rapid implementation of the query across those two sites as opposed to the third site, which used a different platform. Additionally, we found that using a lab value as the basis for a query also posed numerous challenges. While the IGF-1 result is a discrete data field, the normal ranges of IGF-1 vary tremendously by assay and there were multiple assays used even at individual institutions. A substantial amount of work was required to transform each individual assay's results into *Z*-scores. Finally, despite very strictly defined auxological and lab-based criteria for this subcohort, it still ended up including a very heterogeneous group of patients, which was not ideal for gene discovery. Further work is underway to commoditize the development of computational phenotypes prior to recruitment of new cohorts, including incorporation of natural language processing techniques. This pilot project demonstrated the feasibility of recruiting a very rare clinical presentation across multiple sites using EHR-based searches. The challenges that arose informed the design of workflows and IT systems. The experience generated preliminary data for a successful R01 application to continue this work in the GRIN Consortium (A Multicenter Collaborative Clinical Study to Identify Novel Causes of Severe Pediatric Growth Disorders, 1 R01 HD093622 01A1).

Beyond the initial pilots, the GRIN infrastructure has supported multiple NIH grant submissions involving investigators with diverse specialties and backgrounds across the three institutions.

Scalability

In a test of scalability, two of the sites joined the federated query and tested and refined the integration. The third site was able to install a local database, upload their data, and connect to the federated query in less than 48 hours. Subsequently, as of July 2019, the three institutions contributed data for over 1,016,337 patients into the GRIN network. Of those patients, 11,897 have the GRIN consent enabling recontact, and 4860 of these patients have biobanked GRIN-consented samples. A total of 95,803 have any biobank sample at all. All this information is now available for controlled access querying across the network and for export to a secure shared workspace.

The preliminary success of GRIN resulted in a funded cooperative agreement (U01TR002623) from the NIH/NCATS to scale the approach to new institutions across the Clinical and Translational Science Awards (CTSA) Program Consortium, with a goal of building a genomic information commons with distributed data and federated governance.

DISCUSSION

The American College of Medical Genetics and Genomics (ACMG) 2017 position statement[33] "Laboratory and clinical genomic data sharing is crucial to improving genetic health care" enumerates benefits of genotype–phenotype data sharing, including (1) key clinical attributes of the phenotype of those with genetic diseases can be described, (2) qualitative strength of the association between genetic diseases and the underlying causative genes can be established, (3) classification of genomic variants across the range of benign to pathogenic can be established, (4) differences in variant interpretation among laboratories can be reconciled, and (5) expensive duplication of previously resolved, but unpublished, research cohorts can be reduced.

GRIN uniquely addresses the ACMG's call to action. It is unusual for three leading and often competing institutions to broadly and deeply collaborate and share data—particularly sensitive genetic data, processes, and patient populations. The GRIN initiative has not just been an investigator-driven collaboration, but truly an institutional one, with buy-in and active participation initiated at the CEO level and shared by department chairs, IRB leads, offices of general counsel, compliance officers, commercialization offices, biobank directors, informatics leads, and both clinical and basic science investigators.

Our approach is complementary and additive to other prominent initiatives developing population-level genotype–phenotype reference data. These tend to fall into four categories. First are databases of genotype–phenotype relationships as observed and submitted by researchers, such as OMIM,[34] ClinVar,[35] and the National Human Genome Research Institute (NHGRI) Genome-Wide Association Study (GWAS) Catalog.[36–38] Second are databases such as the Genome Aggregation Database (gnomAD)[39]—the next iteration of the ExAC database[18] and the 1000 Genomes Project[40]—that aggregate sequences collected from other studies for secondary use. Third, patients and other study participants are invited to donate data to registries like GenomeConnect[41] or enroll in cohorts like the NIH All of Us initiative,[42] which is consenting one million participants to contribute biological samples and EHR data for research. Fourth, the Global Alliance for Genomics and Health (GA4GH) is establishing an open specification for an API supporting federated query across myriad disparate variant databases.[43]

We contrast our approach to these important and complementary efforts. Specifically, GRIN seeks to create a definitive and representative reference cohort and associated phenotypic and genomic data set, in full control of hospitals and health-care systems, that links the genome to diagnosis, clinical progression, and therapeutic response, establishing a valuable resource of linked biospecimens. This strategy is especially important for rare disorders, where meaningful research requires expansion of cohorts across multiple collection centers, granular observational data that are consistent across contributing sites, direct and collaborative interactions between interested experts with complementary expertise at multiple sites, and representation of rare variants that might not be present in sequence aggregations derived from large generalized, disease-specific, or ethnically homogeneous populations.

GRIN manages a cohort of participants who can be recontacted at any time. GRIN creates local data sets and processes that directly benefit local investigators and encourage collaboration across the network, while also enabling them to readily collaborate with investigators globally. Another important advantage of the federated approach is that the software and processes can be used for local purposes. For example, at Boston Children's Hospital, the same software instance is being used both to run the local Precision Link Biobank[44] as well as to participate in GRIN.

Notably, GRIN provides a particularly meaningful demographic complement to the All of Us initiative with several distinct features: (1) a continuously updating longitudinal phenotype; (2) a distributed approach to combining data from each site of care following previously articulated key principles of stakeholder engagement in federated networks,[3] including local benefit to participating sites and parsimony of data standards; and (3) participating sites and all of their investigators have complete access to their local data, query access to aggregate data across participating centers, and protocol-by-protocol access to the full data set for investigator-initiated studies. The modularity and scalability of GRIN are now ready for testing. We are evaluating interoperation with other open source efforts, such as the Broad Institute's Hail Variant Store,[45] components of the Gabriella Miller Kids First Pediatric Data Resource,[46] and the gNOME variant pipeline.[47] Given the ease of onboarding among the first three sites, we anticipate that it will be straightforward for additional sites to join this federated genotype–phenotype research network. The policies and procedures for collaboration and interoperable biobank enrollment will be sharable as well.

GRIN also shares features with the Accrual to Clinical Trials (ACT) network, which includes sites across the Clinical and Translational Science Center consortium, funded by NCATS/NIH. ACT uses a similar federated query technology[5,48] that interrogates EHR data across multiple sites to identify and match potential participants for clinical trials.

Conclusions

The Genomics and Research Innovative Network has successfully developed a unique, scalable, and federated approach that addresses bioinformatic, regulatory, biobank, and collaboration barriers that have previously slowed the rate of pediatric genomic discovery. GRIN's next phase will focus on widespread adoption within its three founding institutions to accelerate cohort identification, variant discovery, and validation studies. Simultaneously, we expect that GRIN's approach to standardization and harmonization will spread to include other leading research institutions, within pediatrics, and beyond. GRIN's ultimate goal is to transform health care through collaborative genomic discovery and implementation. The model relies not only on distributed and federated data, but also on decentralized governance by participating institutions.

ACKNOWLEDGEMENTS

We gratefully acknowledge financial support from Boston Children's Hospital, Children's Hospital of Philadelphia, and Cincinnati Children's Hospital Medical Center. This work was supported by cooperative agreement U01TR002623 from NCATS/NIH. K.D.M. was also supported in part by R01GM104303 from the NIH. A.H.B. was supported by the Manton Center for Rare Disease Research, NIH U54HD090255 and R01HD075802.

DISCLOSURE

T.G. is on the Scientific Advisory Board of Myriad Neuroscience and Clarigent Health. He has been an advisor/consultant to Supernus, Neurelis, ucb Pharma, Eisai, Sunovion, and SK Life Science. He receives royalties from Myriad Neuroscience and McGraw Hill Education. J.N.H. declares that he is on the Scientific Advisory Board of Camp4 Therapeutics. S.W.K. receives sponsored research support from Pfizer and Quest Diagnostics. K.D.M. receives sponsored research support from Quest Diagnostics. K.M. has received compensation as a consultant for Novartis. The other authors declare no conflicts of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
2. Genomics Research and Innovation Network (GRIN). GRIN portal. 2019. https://www.grinnetwork.org/. Accessed 2 September 2019.
3. Mandl KD, Kohane IS. Federalist principles for healthcare data networks. Nat Biotechnol. 2015;33:360–363.
4. Mandl KD, et al. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture. J Am Med Inform Assoc. 2014;21:615–620.
5. McMurry AJ, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS ONE. 2013;8:e55811.
6. Carnahan RM, Bell CJ, Platt R Active Surveillance: The United States Food and Drug Administration's Sentinel Initiative. In Andrews EB, Moore N (eds.) Mann's Pharmacovigilance 2014 https://doi.org/10.1002/9781118820186.ch27. Accessed 2 September 2019.
7. Platt R, et al. The FDA Sentinel Initiative—an evolving national resource. N Engl J Med. 2018;379:2091–2093.
8. Fleurence RL, et al. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014;21:578–582.
9. Fullerton SM, et al. Return of individual research results from genome-wide association studies: experience of the Electronic Medical Records and Genomics (eMERGE) Network. Genet Med. 2012;14:424–431.
10. Ritchie MD, et al. Electronic medical records and genomics (eMERGE) network exploration in cataract: several new potential susceptibility loci. Mol Vis. 2014;20:1281–1295.
11. Brookes AJ, Robinson PN. Human genotype–phenotype databases: aims, challenges and opportunities. Nat Rev Genet. 2015;16:702–715.
12. National Center for Advancing Translational Sciences. NCATS SMART IRB Platform. 2018. https://ncats.nih.gov/ctsa/projects/smartirb. Accessed 2 September 2019.
13. American Association for Cancer Research. Revised common rule allows broad consent. Cancer Discov. 2017;7:346.
14. Harvard Medical School Department of Biomedical Informatics. BD2K PIC-SURE RESTful API. 2018. http://bd2k-picsure.hms.harvard.edu/. Accessed 2 September 2019.
15. National Institutes of Health. DCPPC APIs overview. 2018. https://tinyurl.com/DCPPC-APIs-overview. Accessed 2 September 2019.
16. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey. 2019. https://www.cdc.gov/nchs/nhanes/index.htm. Accessed 2 September 2019.
17. Funk LM, Shan Y, Voils CI, Kloke J, Hanrahan LP. Electronic Health Record Data Versus the National Health and Nutrition Examination Survey (NHANES): a comparison of overweight and obesity rates. Med Care. 2017;55:598–605.
18. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–291.
19. McQuillan GM, McLean JE, Chiappa M, Corporation H, Lukacs SL. National Health and Nutrition Examination Survey Biospecimen Program: NHANES III (1988–94) and NHANES 1999-2014. Vital Health Stat. 2015;2:1–14.
20. Simons Foundation. Simons Foundation Austism Research Initiative (SFARI). 2018. https://www.sfari.org/. Accessed 2 September 2019.
21. Broad Institute. What is a GVCF and how is it different from a 'regular' VCF?. 2014. https://software.broadinstitute.org/gatk/documentation/article.php?id=4017. Accessed 2 September 2019.
22. Harvard Medical School Department of Biomedical Informatics. ExAC Browser. 2018. http://exac.hms.harvard.edu/. Accessed 2 September 2019.
23. Murphy SN, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17:124–130.
24. Scheufele E, et al. tranSMART: an open source knowledge management and high content data analytics platform. AMIA Jt Summits Transl Sci Proc. 2014;2014:96–101.
25. Collaborative Institutional Training Initiative. CITI Program Biomedical (Biomed) Basic Human Subject Research Course. 2019. https://about.citiprogram.org/en/course/biomedical-biomed-basic/. Accessed 2 September 2019.
26. Google Ventures. Design Sprint. 2019. http://www.gv.com/sprint/. Accessed 2 September 2019.
27. Shen D, et al. De novo GABRG2 mutations associated with epileptic encephalopathies. Brain. 2017;140:49–67.
28. Helbig KL, et al. De novo pathogenic variants in CACNA1E cause developmental and epileptic encephalopathy with contractures, macrocephaly, and dyskinesias. Am J Hum Genet. 2018;103:666–678.
29. International League Against Epilepsy Consortium on Complex Epilepsies. Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. Nat Commun. 2018;9:5269.
30. Helbig I, Lindhout D. Advancing the phenome alongside the genome in epilepsy studies. Neurology. 2017;89:14–15.
31. Helbig I, et al. The ClinGen Epilepsy Gene Curation Expert Panel—bridging the divide between clinical domain knowledge and formal gene curation criteria. Hum Mutat. 2018;39:1476–1484.
32. Helbig I, et al. A recurrent missense variant in AP2M1 impairs clathrin-mediated endocytosis and causes developmental and epileptic encephalopathy. Am J Hum Genet. 2019;104:1060–1072.
33. ACMG Board of Directors. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. Genet Med. 2017;19:721–722.
34. McKusick-Nathans Institute of Genetic Medicine. OMIM: Online Mendelian Inheritance in Man. 2019. https://www.omim.org/. Accessed 2 September 2019.
35. National Center for Biotechnology Information. ClinVar. 2019. https://www.ncbi.nlm.nih.gov/clinvar/. Accessed 2 September 2019.
36. National Human Genome Institute. A catalog of published genome-wide association studies. 2015. https://www.genome.gov/catalog-of-published-genomewide-association-studies. Accessed 2 September 2019.
37. Buniello A, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–d1012.

38. European Molecular Biology Laboratory—European Bioinformatics Institute. GWAS catalog. 2019. https://www.ebi.ac.uk/gwas/. Accessed 2 September 2019.
39. gnomAD Consortium. gnomAD browser beta. 2019. http://gnomad.broadinstitute.org/. Accessed 2 September 2019.
40. 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
41. Clinical Genome Resource (ClinGen). GenomeConnect. 2019. https://www.genomeconnect.org/. Accessed 2 September 2019.
42. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372:793–795.
43. Global Alliance for Genomics and Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science. 2016;352:1278–1280.
44. Bourgeois FT. et al. Development of the Precision Link Biobank at Boston Children's Hospital: challenges and opportunities. J Pers Med. 2017;7:E21.
45. Broad Institute of MIT and Harvard. Hail. 2016. https://hail.is/docs/0.1/overview.html. Accessed 2 September 2019.
46. Gabriella Miller KidsFirst Research Program. KidsFirst dashboard. 2018. https://portal.kidsfirstdrc.org/. Accessed 2 September 2019.
47. Lee IH, et al. Prioritizing disease-linked variants, genes, and pathways with an interactive whole-genome analysis pipeline. Hum Mutat. 2014;35:537–547.
48. Weber GM, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16:624–630.

the Genomics Research and Innovation Network

Darlene Barkman[27], Erin M. Borglund[28], Ramkrishna Chakrabarty[29], Alka Chandel[30], Anil Kumar Degala[28,31], Thomas DeSain[31], Philip Dexheimer[30], Parth Divekar[30], Alyssa Ellis[28], Mike Furgason[28,34], Christopher Geehan[32], Andrew Joseph Guidetti[31], Alba Gutierrez[31], Barbara Hallinan[33,34], Becca Harper[35], Niloofar Jalali[31], Jaspreet Khanna[28,31], Christopher Kirby[32], Gabor Korodi[31], Michal Kouril[30], Amy Kratchman[27], Ranjay Kumar[31], Guillaume Labilloy[30], In-Hee Lee[28], Bria Morgan[31], James Morgan[30], Louis J Muglia[34,36], Aleksandr Nikitin[28,31], Mike Pistone[37], Anna Poduri[38], Andrew Rupert[30], Kristen Safier[39], Piotr Sliz[29], Gelvina Stevenson[27], Joseph St. GemeIII[27], Vidhu Thaker[40,41], Simone Temporal[27], Prakash Velayutham[30], Julie Wijesooriya[34], Bryan Wolf[27], Andrew Wooten[36], Alan Yen[42] and Yu Zhang[31]

[27]Children's Hospital of Philadelphia, Philadelphia, PA, USA. [28]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [29]Research Computing, Boston Children's Hospital, Boston, MA, USA. [30]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [31]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [32]Office of General Counsel, Boston Children's Hospital, Boston, MA, USA. [33]Division of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [34]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. [35]Clinical Translational Research Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [36]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [37]Innovation Ventures, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [38]Department of Neurology, Boston Children's Hospital, Boston, MA 02115, USA. [39]Legal Department, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [40]Department of Pediatrics, Columbia University Medical Center, New York, NY, USA. [41]Division of Endocrinology, Boston Children's Hospital, Boston, MA, USA. [42]Technology & Innovation Development Office, Boston Children's Hospital, Boston, MA, USA