ARTICLE

A Genocentric Approach to Discovery of Mendelian Disorders

Adam W. Hansen,[1,2] Mullai Murugan,[2] He Li,[2] Michael M. Khayat,[1,2] Liwen Wang,[2] Jill Rosenfeld,[1] B. Kim Andrews,[2] Shalini N. Jhangiani,[2] Zeynep H. Coban Akdemir,[1] Fritz J. Sedlazeck,[2] Allison E. Ashley-Koch,[3,4] Pengfei Liu,[1] Donna M. Muzny,[1,2] Task Force for Neonatal Genomics, Erica E. Davis,[5,6] Nicholas Katsanis,[5,6] Aniko Sabo,[1,2] Jennifer E. Posey,[1] Yaping Yang,[1] Michael F. Wangler,[1] Christine M. Eng,[1] V. Reid Sutton,[1,7] James R. Lupski,[1,2,7,8] Eric Boerwinkle,[2,9] and Richard A. Gibbs[1,2,*]

The advent of inexpensive, clinical exome sequencing (ES) has led to the accumulation of genetic data from thousands of samples from individuals affected with a wide range of diseases, but for whom the underlying genetic and molecular etiology of their clinical phenotype remains unknown. In many cases, detailed phenotypes are unavailable or poorly recorded and there is little family history to guide study. To accelerate discovery, we integrated ES data from 18,696 individuals referred for suspected Mendelian disease, together with relatives, in an Apache Hadoop data lake (Hadoop Architecture Lake of Exomes [HARLEE]) and implemented a genocentric analysis that rapidly identified 154 genes harboring variants suspected to cause Mendelian disorders. The approach did not rely on case-specific phenotypic classifications but was driven by optimization of gene- and variant-level filter parameters utilizing historical Mendelian disease-gene association discovery data. Variants in 19 of the 154 candidate genes were subsequently reported as causative of a Mendelian trait and additional data support the association of all other candidate genes with disease endpoints.

Introduction

The foundation of Mendelian disease research is the observation of a direct association between variant alleles affecting the expression of the same gene or perturbing the biological function of its encoded protein product and defined clinical phenotypes in a large enough sample set to satisfy predetermined statistical thresholds.[1–4] For example, cosegregation of specific alleles at a locus with phenotypes in multiple families, or repeated independent occurrences of *de novo* heterozygous (or hemizygous) variants in the same genes, consistent with autosomal-dominant (AD) and X-linked (XL) disease traits, can be the basis of proof establishing association between a Mendelian disorder and a gene. Moreover, bi-allelic pathogenic variants at a locus inherited in *trans* from carrier parents can support an autosomal-recessive (AR) disease trait model. In each case allele segregation with phenotypes can be considered alongside the biological role of the indicated gene/protein together with any other *in silico* prediction or empirical functional data. Although a precise algorithm for "Mendelian causation" has proven elusive, these study components have supported thousands of Mendelian disease-gene associations that have survived the test of time by independent replication.[5]

Mendelian studies often begin with selection of phenotypically homogeneous sets of individuals, followed by systematic genotyping and analysis. This "phenocentric" paradigm, as exemplified by clinical phenotype data aggregated in OMIM[6] (see Web Resources), has contributed most of our current understanding of the genetic basis of human disease. However, its sensitivity is limited by incomplete penetrance, variable expressivity, pleiotropy, locus heterogeneity, ubiquity and non-specificity of certain phenotypic traits, and "granularity" of the semantics of clinical phenotypic descriptions. It generally assumes that enriching a group of individuals for phenotypic homogeneity will also enrich for genetic homogeneity, an assumption that is not always reflected by supporting data.[4,7–10]

Availability of next-generation sequencing (NGS) methods have accelerated the accumulation of gene sequence data from families suspected to harbor Mendelian conditions,[1,3,5,11,12] while at the same time, there have been few advances in high-throughput methods for the study of variant allele function in model systems.[13] Hence, there is an increased utilization of genomic DNA sequencing (ES and whole-genome sequencing [WGS]) of affected and healthy research human subjects and associated clinical samples as a driver for Mendelian disease discovery. This DNA sequence-driven, "genocentric" paradigm has been accelerated by the advent of generalizable *in silico* tools and datasets, such as effective likelihood-based statistical methods to predict potential deleteriousness of missense alleles to protein function[14–16] and

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; [3]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC 27710, USA; [4]Department of Medicine, Duke University Medical Center, Durham, NC 27710, USA; [5]Pediatric Genetic and translational Medicine Center (P-GeM), Stanley Manne Children's Research Institute, Chicago, IL 60611, USA; [6]Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; [7]Texas Children's Hospital, Houston, TX 77030, USA; [8]Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA; [9]School of Public Health, UTHealth, Houston, TX 77030, USA
*Correspondence: agibbs@bcm.edu
https://doi.org/10.1016/j.ajhg.2019.09.027.

nonsense/frameshift alleles to mRNA stability[17] and aggregate databases that report the observed number and class (i.e., missense, loss-of-function, etc.) of variant alleles in large reference datasets (e.g., ExAC, gnomAD, CHARGE, 1000 Genomes).[18–20] Scores derived to reflect the population frequencies of the variant classes observed in such databases have proven of great value in Mendelian disease research, as in general, mutations in genes that exhibit less variability in the population are more likely to result in pathogenic effects—the basic tenet of the Clan Genomics hypothesis[21] and the rare variant family-based genomics approach. Furthermore, recent efforts have applied deep learning to primate-human comparative genomic data to predict variant pathogenicity.[22]

To date, genocentric approaches utilizing these scores have focused on the relatively straightforward interpretation of homozygous predicted loss-of-function (LoF) variation[23] or *de novo* mutation.[7,24] Less effort has been applied to the more challenging exploration of missense variants. Missense variants are challenging because of their abundance, especially when there are no samples available from related individuals to allow exploration of the family-based genomics approach and testing of patterns of segregation (AD, AR, XL) of disease phenotypic traits.

Accumulation of large DNA sequence datasets has been matched by the emergence of sophisticated distributed computing methods that offer high levels of capacity combined with the ability to manipulate large, unstructured datasets. These methods can be deployed locally, or on the cloud, with high dexterity. Apache Hadoop is one such tool, which is becoming increasingly popular in NGS pipeline analysis,[25–33] but has not yet been applied, to our knowledge, to the task of discovering Mendelian disorders and their associated genes. Together with the large amount of sequence data from case subjects and families with suspected Mendelian disorders, these developments provide great opportunity for discovery.

We have accumulated ES data from 18,696 individuals from both gene discovery-focused Mendelian disease research efforts and clinical molecular diagnostic programs at Baylor College of Medicine (see Material and Methods), including 14,755 probands (approximately 30% solved), each with a suspected Mendelian condition, and 3,941 control subjects (either unaffected or part of a Wolff-Parkinson-White syndrome cohort) or family members (affected status unknown). These heterogeneous phenotype and genomics data sample sets had varying amounts of clinical and phenotypic annotation, different representation of information and availability of DNA samples from relatives, and variable consent for use as research subjects or as clinical case subjects where the aim for further analysis was to improve the diagnostic yield (see Material and Methods).[2]

To enable a genocentric analysis, we recorded the variant data from these samples in a single, HIPAA-compliant, secure Hadoop-structured environment, together with appropriate public datasets and computational predictions of variant impact. Data access permissions were carefully managed so as not to inappropriately reveal data from single samples that would compromise privacy agreements. The study revealed an efficient, genocentric pathway to Mendelian discovery and illustrated the power of tools such as Hadoop to enable consolidation of heterogeneously structured genetic data in a single, secure interrogatory environment. Through empirical optimization of search parameters, we identified 154 candidate Mendelian disease-gene associations, 19 of which were reported to OMIM as causative in the months following our initial analysis and discovery. The remaining 135 candidate disease-gene associations are supported by ACMG sequence variant interpretation guidelines, including population and computational predictive data, functional data, and in some instances, *de novo* inheritance.[14,34]

Material and Methods

Samples

Samples were obtained through long-standing research and clinical collaborations. Research samples were collected after written informed consent in conjunction with either the Baylor Hopkins Center for Mendelian Genomics (BHCMG) (H-29697) study with approval by the institutional review board at Baylor College of Medicine or the Task Force for Neonatal Genomics study with approval by the institutional review board at Duke University. Data were also from the Baylor College of Medicine clinical testing laboratories, now incorporated as the Baylor Genetics Laboratories (BG). These data were studied in aggregate for the purpose of improving the diagnostic (protocol H-41191). All genomic studies were performed on DNA extracted from blood or saliva samples. PCA analysis of exome data revealed no significant difference in distribution of ethnicities between case and control samples (Figure S1). Self-reported ethnicity was not tracked.

DNA Sequencing

DNA capture and sequencing of exomes was carried out as previously described by Yang et al.[1] at either the Baylor Genetics (BG) laboratories or at the Baylor College of Medicine Human Genome Sequencing Center (HGSC), through the Baylor-Hopkins Center for Mendelian Genomics initiative. Briefly, using 500 μg of DNA, an Illumina paired-end pre-capture library was constructed according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D) with modifications as described in the *BCM-HGSC Illumina Barcoded Paired-End Capture Library Preparation* protocol. Pre-capture libraries were pooled into 4-plex library pools and then hybridized in solution to the HGSC-designed Core capture reagent[1] (52 Mb, NimbleGen) or 6-plex library pools used the custom VCRome 2.1 capture reagent[1] (42 Mb, NimbleGen) according to the manufacturer's protocol (*NimbleGen SeqCap EZ Exome Library SR User's Guide*) with minor revisions. The sequencing run was performed in paired-end mode using the Illumina HiSeq 2000 platform, with sequencing-by-synthesis reactions extended for 101 cycles from each end and an additional 7 cycles for the index read. With a sequencing yield averaging 8.5 Gb, the sample achieved 93% of the targeted exome bases covered to a depth of 20× or greater. Illumina sequence analysis was performed using the HGSC Mercury analysis pipeline[2,3] (see

Web Resources) which moves data through various analysis tools from the initial sequence generation on the instrument to annotated variant calls (SNPs and intra-read indels). In parallel to the exome workflow, an Illumina Infinium Human Exome v1-2 array was generated for a final quality assessment. This included orthogonal confirmation of sample identity and purity using the Error Rate In Sequencing (ERIS) pipeline developed at the BCM-HGSC. Using an "e-GenoTyping" approach, ERIS screens all sequence reads for exact matches to probe sequences defined by the variant and position of interest. A successfully sequenced sample must meet quality-control metrics of ERIS SNP array concordance (>90%) and ERIS average contamination rate (<5%).

Phenotyping

BG

Unstructured phenotypic data are available for all BG samples. Most of these free text clinical summaries were based on clinical notes and the test requisition, and written by clinical scientists, fellows, and laboratory directors. Test requisitions have evolved over time, but typically consisted of a checklist of symptoms with the ability to write-in additional details, and may have been filled out by MDs, genetic counselors, or nurses. Structured phenotypic data are also available for 9,434 samples, with a mean of 7.76 distinct phenotypic descriptors entered per sample; these data were generated by Codified Genomics (see Web Resources) or other tools, typically mapping test requisition symptoms directly to HPO terms, and subsequently reviewed by clinical scientists, fellows, or laboratory directors.

BHCMG

The BHCMG has developed PhenoDB,[4] a web-based portal for entry of phenotypic and clinical information that is freely available. The 3K features use the preferred term from the Elements of Morphology and are mapped to the Human Phenotype Ontology. A submitter can enter data by family or cohort including information such as phenotypic features, diagnosis, mode of inheritance, clinical history, and upload previous genetic testing results. PhenoDB has several modules allowing for storage of data as well as analysis and GeneMatcher, a tool used to link investigators sharing the same gene of interest.[5]

Computer Infrastructure

HARLEE is a 10-node 280TB Cloudera Hadoop cluster. VCF files for all samples were first annotated with VEP—flagging a most-important transcript per-gene per-variant—then subsequently ingested into HARLEE, together with annotation tables from a variety of sources (Supplemental Material and Methods). Data access is strictly controlled with robust encryption and authentication layers, creating an environment ready to comply with FISMA, HIPAA, Texas Medical Records Privacy Act, and other industry regulations.

Genocentric Query Approach

We performed a series of Impala queries in HARLEE, where each query results in a gene list. Scripts were written and executed with R to handle automation of querying and subsequent statistical analysis and visualization. The commonality across all queries is a search for genes harboring ultra-rare variants—with additional quality control filtering—across at least five case subjects (n = 14,755), absent from all control samples (n = 3,941). With the intent of minimizing false-positive candidate gene volume, controls were broadly defined to include parental samples (n = 3,587) in addition to internal healthy control subjects (n = 42) and a Wolff-Parkinson-White syndrome (WPW) cohort (n = 319).

Specifically, all queries shared the following filter parameters: HARLEE internal allele frequency < 0.01; 1000 Genomes allele frequency < 0.001 or is null; CHARGE consortium (large-scale adult cardiovascular cohort sequenced internally) allele frequency ≤ 0.0001 or is null; gnomAD allele frequency ≤ 0.0001 or is null; variant not cited in PubMed; variant has no dbSNP ID; chromosome name does not start with "GL;" domains field, if not empty, does not start with "low_complexity;" ExAC mu_syn is not null (ExAC did not exclude this gene for constraint score analysis); variant read count ≥ 4, variant allele frequency (VAF) ≥ 0.25. Next, we categorized queries as those looking for "loss-of-function" variants (VEP impact = HIGH) versus those looking for missense variants (VEP impact = MODERATE). Finally, queries were further categorized based on one additional gene-specific or variant-specific bioinformatic score or filtering parameter: for loss-of-function variants, ExAC pLI and loss-of-function intolerance z-score; for missense variants, ExAC loss-of-function z-score and missense intolerance z-score, REVEL, MTR (missense tolerance ratio—a region-specific missense tolerance score), SIFT, and PolyPhen. For each of these scores, we implemented a high-pass (or low-pass, for SIFT and MTR) parameter sweep consisting of up to 1,000 queries, measuring the impact of score-based cutoff filtering on resulting gene list size and OMIM disease annotation over time. For each respective parameter sweep query series, the variable parameter was incremented or decremented by 0.01 across the following score ranges, holding all other filtering criteria constant: 0–1 for pLI, REVEL, SIFT, and PolyPhen; 0–10 for loss-of-function and missense intolerance z-scores; and 0–1.6 for MTR.

To enable validation of our approach and parameter optimization, we annotated genes with OMIM Mendelian disease association data from a freeze of the OMIM data from four different timestamps: early 2013, late 2014, mid 2016, and early 2018. For any given set of genes and a fixed duration of time, we define "discovery" as the number of genes in the set with a disease annotation added to OMIM during the given time span.

Candidate gene lists were identified by selecting a hard-cutoff filter value for each respective variable annotation parameter. The cutoff value was selected for each parameter as the value which optimized "discovery density"—calculated as 2013–2018 discovery divided by the number of genes without OMIM annotations in 2018—with a required minimum output of 20 genes without OMIM annotations in 2018. All genes resulting from a query with a cutoff value maximizing discovery density are considered candidates.

Discovery Density Simulations

To demonstrate the sensitivity of the optimum discovery density metric to input OMIM annotation data—which changes over time as associations between genes and Mendelian phenotypes are published and eventually curated by OMIM—discovery density for all five possible nonoverlapping time intervals was plotted for all queries. These data points were supplemented by a distribution of 1,000 instances of removing 20% of all disease-gene annotations from OMIM 2013, calculating discovery as the number of genes within a given query without an OMIM annotation in a given simulation, with an OMIM annotation in the real 2013 database. Discovery density for each query across this simulated interval was then plotted against 2013–2018 discovery. The average discovery density across all 1,000 simulations was also calculated,

with a linear regression model fitted against average simulated discovery density versus 2013–2018 discovery density.

Phenocentric Query Approach

We also established a methodology for rapidly conducting a large-scale phenocentric analysis for discovery of variation in genes associating with a specific phenotype, assuming a dominant inheritance model. To conduct a large-scale phenocentric analysis, we first counted the number of distinct samples with variants—meeting specific criteria—per gene across all Mendelian exomes in HARLEE. Variant filtering criteria is as follows: single-nucleotide variant; MAF < 0.0001 (gnomAD, CHARGE); MAF < 0.001 (1000 Genomes); MAF < 0.01 (HARLEE); if REVEL score is available, REVEL score ≥ 0.25; remove chromosome names beginning with "GL;" remove variants where VEP domain annotation begins with "Low_complexity;" remove genes where ExAC does not calculate gene-level scores; VEP existing_variation annotation must be empty or null; VEP impact annotation must be "MODERATE" (indicative of missense variation) or "HIGH" (frameshift, start-loss, stop-gain, stop-loss, or canonical splice site disrupting). We then normalized the mutation rate by cohort size, in addition to normalizing by both cohort size and gene cDNA length.

We then replicated this analysis on a phenocentric sub-cohort filtered out of the overall cohort, only including samples annotated with at least one phenotypic term matching a list of provided terms. We then calculated phenotypic enrichment for each gene by dividing the normalized mutation rates by the respective normalized mutation rates from the overall cohort. Focusing on hearing disorders, phenotypic search terms included "middle ear," "hearing," and "deaf." Fisher's exact test was utilized to test for significant enrichment, and p values were false discovery rate-corrected.

Results

Computational Infrastructure

The Hadoop Architecture Lake of Exomes (HARLEE) is a data lake created in a Hadoop environment (powered by Cloudera) for housing and facilitating analysis of next-generation sequencing data. This resource provides a flexible environment for simultaneously housing structured, semi-structured, unstructured and heterogeneous data; SQL-on-Hadoop solutions to perform high-speed simple queries and complex comparison queries of the data; a cost-effective solution that uses commodity hardware; the ability to scale-as-required by adding more nodes; fault tolerance achieved by storing the data in triplicate across the nodes; a secure, compliant-ready environment; and granular control of data access privilege. In the current study, anonymized sample-level data were appropriately protected by master-level password access in order to allow only qualified individuals to access specific data components. Instantiation of a more elaborate tiered access system can easily be imposed upon the current HARLEE and be applied to more outward-facing activities.

Benchmark and stress tests via multiple tools, including TeraGen, TeraSort, TestDFSIO, NNBench, and MRBench showed that the performance of the cluster during data ingestion and querying (Table S1), even with the overhead for encryption/decryption (Table S2), far surpassed the capability of programmatic approaches that provide the same results by parsing and interpreting flat files. The architecture allowed warehousing large volumes (i.e., 30,000+ samples, 6 TB) of heterogeneous data while providing rapid sample-level access on the order of a few seconds.

HARLEE Facilitates Genocentric Mendelian Discovery

HARLEE was loaded with data from VCF files from ES samples that were annotated with transcript effect information (VEP) (Figure 1). Utilizing HARLEE, data from all samples were further annotated with known minor allele frequencies (ExAC, gnomAD, CHARGE, 1000 Genomes), functional predictions from multiple bioinformatic algorithms (SIFT, PolyPhen, REVEL, Missense Tolerance Ratio [MTR]), clinical variant information (ClinVar), and other sources.[15,16,18,35]

To identify annotation features that would facilitate discovery of genes with associated Mendelian disorders, a series of empirical tests were performed on the accrued data, structured on genotypic parameters, without regard for the underlying phenotypic or disease trait inheritance/segregation patterns. As lists of genes associated with known Mendelian phenotypes curated by OMIM were available from different years (2013–2018), a comparison of the yields of genes discovered at different time points was informative. Throughout, the ratio of known disease-associated genes/all genes that were identified by different parameters and cutoffs were used to optimize different parameters and to maximize the likelihood of enrichment for genes associated with undiscovered Mendelian disorders.

This approach distinguished groups of genes with varying levels of enrichment for known Mendelian phenotypic associations. Figure 2 illustrates testing of a single variable, titrating scores that predict intolerance of genes to missense variants (ExAC's missense intolerance z-score). As anticipated, when the z-score increased, the total number of genes that were identified decreased, while the fraction that were already known to associate with Mendelian disease in 2016 increased (Figure 2). This reflects a trend where LoF mutations in highly constrained genes are more likely to be pathogenic.[18] The majority of genes with a missense intolerance z-score higher than 8 were already reported to OMIM as disease causing in 2016, with two exceptions. In bin 8-8.5, *CLTC*—Clathrin, Heavy Chain (MIM: 118955)—was since reported to associate with multiple malformation and developmental delay (MIM: 617854).[36] In bin 8.5-8, *POLR2A*—RNA Polymerase II, subunit A (MIM: 180660)—was recently reported to associate with a neurodevelopmental syndrome with infantile-onset hypotonia.[37] Hence, this straightforward threshold of a high-missense z-score (>8.5) provides high enrichment for genes known to

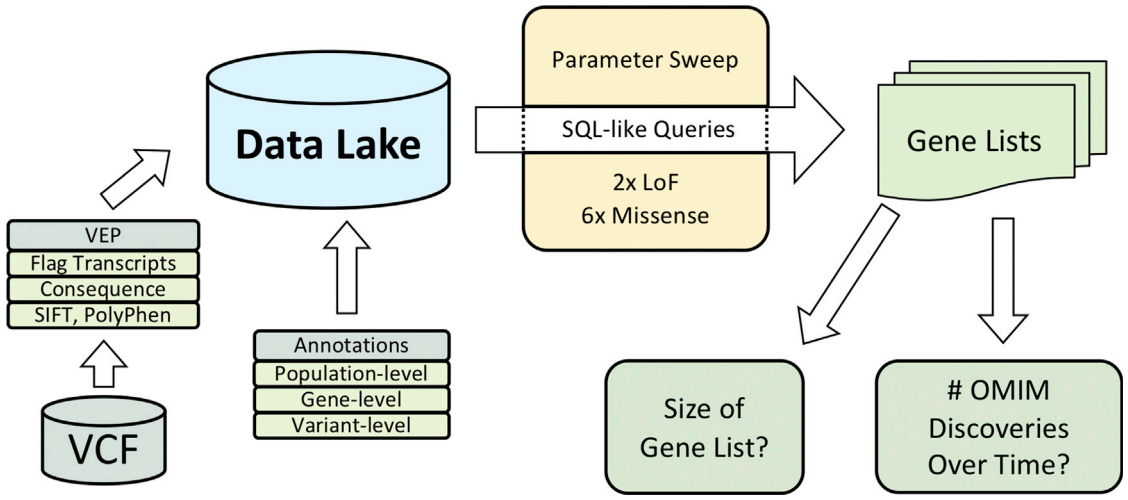


Figure 1. HARLEE Workflow
ES VCF files are first annotated with Variant Effect Predictor (VEP), where one transcript is flagged per variant per gene. Consequence, SIFT, PolyPhen, variant allele frequency from multiple sources, domain information, and other annotations are additionally ascertained by VEP. VEP output is loaded into a Hadoop architecture data lake. Finally, population-, variant- and gene-level annotations from a variety of sources are loaded, allowing for modular, on-demand annotation. After samples and annotations are separately loaded into HARLEE, a series of SQL-like queries generate distinct gene lists. Bioinformatic filtering parameters based on loaded annotations are tuned to optimize discovery density, which takes into account the volume of genes reported to OMIM as disease-associated over time normalized against the number of remaining genes without OMIM disease annotations.

associate with Mendelian disorders, but yields few discoveries.

Interestingly, the change in proportion of genes known to associate with Mendelian phenotypes was not smooth toward the upper end of constraint: at a missense z-score between 6.5 and 7, a bin with a much lower disease gene enrichment than the surrounding bins was observed. Owing to ongoing efforts by OMIM to curate the literature for newly reported Mendelian disease-gene associations, when the analysis was repeated for the list of known Mendelian genes in 2018, 2/9 of those genes (*DHX30* [MIM: 616423], *SMC1A* [MIM: 300040]) were revealed to have associated Mendelian phenotypes by 2018 (MIM: 617804 and 300590, respectively).[38,39] Thus, this empirical strategy also showed that a "parameter sweep" could identify bins containing sets of genes that were enriched at an intermediate level for known Mendelian phenotypes, as well as many strong candidates for future discovery (Figures 2 and S2).

For further analyses, we defined "discovery density" for a given set of genes as the change in number of genes with associated Mendelian phenotypes in OMIM, over time, normalized (divided) by the number of remaining genes without reported Mendelian disease trait associations. For example, if 10 genes out of a set of 20 had a reported disease association in 2013 and 18/20 were reported to associate with a Mendelian trait by 2018, the discovery density would be $(18 - 10)/(20 - 18) = 4$.

The suitability of the use of discovery density to optimize filter parameters for discovery of disease-associated genes was separately tested to ensure that the correlations were robust. Overall, we found that use of this measure based upon almost any combination of available data from different years of OMIM was effective, provided years with low absolute discovery rates were avoided (Figures S4–S7).

Subsequent analyses aimed to identify sets of candidate "Mendelian disease-associated genes"—or genes which can be disrupted by variants pathogenic for Mendelian phenotypes—for each annotation score by identifying parameters that yielded the highest discovery density. To reduce the impact of discovery density outlier values inflated by small gene set sizes, candidate disease-associated gene sets were constrained to a minimum size of 20. In total, eight parameter-sweep query series were performed: two testing putative loss-of-function variants (via pLI and loss-of-function [LoF] intolerance z-score) and six testing missense variants (via LoF intolerance z-score, missense intolerance z-score, REVEL, MTR, SIFT, and PolyPhen2).

Apart from variant consequence and the variable parameters cited above, the tests to identify groups of genes rich in undiscovered Mendelian disease-associated genes maintained constant filter parameters across all queries. We limited analyses to include high-quality, ultra-rare (MAF $<$ 1/10,000) variants. Genes were limited to those harboring such variants across at least five different ES entries in HARLEE, substantiating a minimum potential cohort of individuals for each proposed candidate disease-associated gene. No specific filters to remove particular classes of genotype (i.e., heterozygous or homozygous) were included; however, the stringent allele frequency filter greatly enriches for heterozygous variants (often *de novo* mutations), and thus the subset of true pathogenic variants identified are mostly expected to be dominant-acting alleles for an AD disease trait. Stringent allele

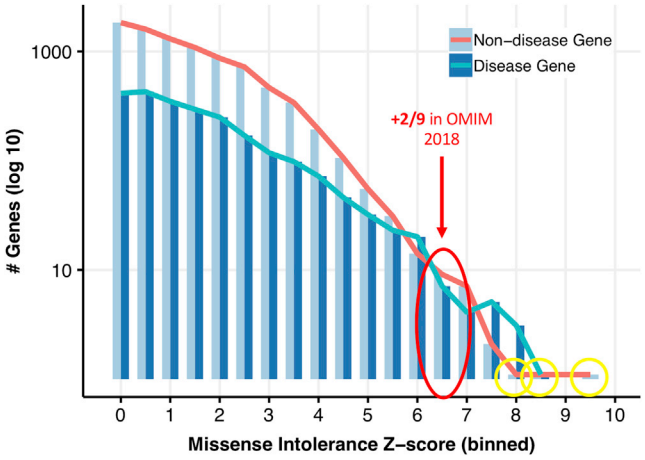


Figure 2. Missense Intolerance Z-Score Pilot Query Series
Query input (missense intolerance z-score range) is plotted against query output (number of resulting genes with and without OMIM disease annotations in 2016). Except for the variable missense intolerance z-score range, all queries were identical, outputting genes within a given z-score range (bin width = 0.5) where at least five case exomes harbored high-quality ultra-rare missense variants (absent from controls). One bin with an intermediate constraint score range (red) had a lower-than-expected proportion of disease-associated genes; 2/9 of these genes were reported as associating to an OMIM phenotype between 2016 and 2018. Outlier genes with extreme constraint scores not known to associate with Mendelian disorders in 2016 (yellow) were all recently reported as disease associated: *CLTC*, *POLR2A*, and *TRRAP*.

frequency filtering appropriately provided a bias toward specificity, rather than sensitivity, as minimizing false positives was an important goal. All variants detected in research samples (see Material and Methods) within candidate disease-associated genes are publicly available (Table S3).

Predicted Loss-of-Function Variants

HARLEE facilitated the identification of 33 candidate disease-associated genes from two distinct loss-of-function variant annotation parameter sweep queries. First, from the loss-of-function intolerance z-score query series, an optimum cutoff value of 7.37 yielded a discovery density of 0.65, while the gene set size was constrained at a minimum of 20 (Table 1). Higher cutoff values could yield a higher discovery density value, but only with a small gene set. Second, the pLI parameter sweep query series identified 29 candidates, including 9 with a pLI score of >0.9999999 that almost certainly constitute Mendelian disease-associating genes.[20] The optimum cutoff value of 0.998 (calculated excluding genes where pLI = 1) yielded an optimum discovery density of 0.6.

In combination, the two methods yielded 33 candidate disease-associated genes with 16 identified by both (Figure 3). Among candidates from these sets, *CHD3* (MIM: 602120), *DOCK3* (MIM: 603123), *KIAA1109* (MIM: 611565), *MYO9A* (MIM: 604875), and *VPS13D* (MIM: 608877) have since been reported to have associated Mendelian phenotypes in OMIM (MIM: 618205, 618292, 617822, 618198, and 607317, respectively), providing evidence in support of our approach.[40–45]

Missense Variants

HARLEE identified 130 candidate disease-associated genes from six distinct missense variant annotation parameter sweeps. From the missense variant loss-of-function intolerance z-score query series, a cutoff value of 9.28 yielded a discovery density of 0.55 with 20 candidate disease-associated genes. From the missense intolerance z-score query series, a cutoff value of 6.23 yielded a discovery density of 0.65, again with a minimal candidate list size of 20 genes. From the MTR query series, an optimum cutoff value of MTR less than 0.42 yielded a discovery density of 0.372 with 43 candidate genes. From the REVEL query series, an optimum cutoff value of REVEL greater than 0.91 yielded a discovery density of 0.2 with 60 candidate genes.

The query series based upon PolyPhen and SIFT produced results that contrasted from the four other methods described above. The PolyPhen query series analysis yielded a discovery density of just 0.109, and an overly large number of 247 candidate genes. Likewise, SIFT analysis yielded a discovery density of 0.086, yielding an unreasonable candidate gene set of 4,837 genes. For both series, the maximum discovery density value occurred when setting the cutoff value to the highest level of constraint for the respective scores (0 for SIFT, 1 for PolyPhen). Because of the excessively large and intractable gene list sets resulting from the PolyPhen and SIFT analyses, combined with their lower discovery density values compared to the six other query series, we did not further utilize these metrics.

The final set of 130 candidate disease-associated genes from our missense variant query series were therefore the union of the gene sets resulting from the loss-of-function intolerance z-score, missense intolerance z-score, MTR, and REVEL parameter sweeps. This included three genes identified by both loss-of-function and missense intolerance z-scores, three genes identified by both loss-of-function z-score and MTR, two genes identified by both loss-of-function z-score and REVEL, five genes identified by both missense intolerance z-score and MTR, and one gene identified by both missense intolerance z-score and REVEL (Figure 3). Of note, *TRRAP* (loss-of-function z-score, missense z-score, and MTR) (MIM: 603015) and *CACNA1I* (missense z-score, MTR, and REVEL) (MIM: 608230) were each identified by three scores. Subsequently, *TRRAP* was reported as a Mendelian disease-associated gene by our collaborators at BHCMG, independent of this analysis (MIM: 618454).[46] Furthermore, *CACNA1A* (MIM: 601011) and *CACNA1E* (MIM: 601013) have both been reported to be associated with neurodevelopmental disorders (MIM: 617106, 108500, 141500, 183086 for *CACNA1A*; MIM: 618285 for *CACNA1E*), establishing a relationship between voltage-dependent calcium channel dysfunction and neurodevelopmental disease, serving as

Table 1. Summary of Mendelian Discovery Analysis

Variant Category	Parameter Name	Parameter Category	Optimum Value	Discovery Density	Gene List Size	Accepted as Candidates
LoF	pLI	gene-level	>0.998	0.6	29	true
LoF	LoF z-score	gene-level	>7.37	0.65	20	true
Missense	LoF z-score	gene-level	>9.28	0.55	20	true
Missense	Mis z-score	gene-level	>6.23	0.65	20	true
Missense	MTR	variant-level	<0.42	0.372	43	true
Missense	PolyPhen	variant-level	≥1	0.109	247	false
Missense	REVEL	variant-level	>0.91	0.2	60	true
Missense	SIFT	variant-level	≤0	0.086	4,837	false

The cutoff value for each query series was selected to optimize discovery density (with a minimum constrained candidate disease-associating gene list size of 20). Results from PolyPhen and SIFT were not considered candidate disease-associating genes as the scores saturated (a maximum level of constraint yielded a maximum cutoff value) with a relatively large remaining gene list size. For the pLI analysis, genes with pLI = 1.0 were excluded from the discovery density optimization calculations, and those genes without OMIM disease annotations were automatically considered candidate disease-associated genes.

evidence in support of *CACNA1I* as a candidate Mendelian disease-associated gene.[47–49]

Combined Set of Candidate Disease-Associating Genes

In total, we identified 154 distinct candidate disease-associated genes between the loss-of-function and missense variant analyses. On average, these candidates (mean length = 5,147 bp; median length = 3,525 bp) are longer than the average coding gene (mean = 1,649 bp; median = 1,227 bp). This is a shared property of all known Mendelian disease-associated genes and does not reflect a systematic bias that would inherently increase false positives; the set of all coding genes with OMIM disease annotations in 2018 (mean = 2,213 bp; median = 1,554 bp) is also significantly longer than the set of all genes (unpaired t test, $p < 0.0001$).

Multiple lines of qualitative and quantitative evidence support the merits of HARLEE disease-gene association discovery. First, comparing these genes against the current set of OMIM annotations at the time of preparing this manuscript (May 2019) revealed 19 candidates have since been reported to associate with Mendelian phenotypes: *ATP1A1*[50,51] (MIM: 182310), *CACNA1E*,[49,52] *CHD3*,[40] *CLTC*[36] (MIM: 118955), *DOCK3*,[41] *FBXO11*[53,54] (MIM: 607871), *IRF2BPL*[55] (MIM: 611720), *KDM5B*[56] (MIM: 605393), *KIAA1109*,[42] *LINGO1*[57] (MIM: 609791), *MACF1*[58] (MIM: 608271), *MAST1*[59] (MIM: 612256), *MYO9A*,[43] *PDE1C*[60] (MIM: 602987), *SCN3A*[61] (MIM: 182391), *SET*[62] (MIM: 600960), *TBX2*[63] (MIM: 600747), *TCF20*[64] (MIM: 603107), and *VPS13D*.[44,45] Permutation analysis sampling 154 random genes without OMIM annotations (as of January 2018) revealed this to be a highly significant enrichment of recently reported disease-associated genes (expected = 2.29; $p < 0.00001$; n = 100,000 permutations). Furthermore, 9 of the 33 loss-of-function candidate genes intersected with the 130 missense candidates: *CHD3*, *CSMD3* (MIM: 608399), *KIAA1109*, *LRP1B* (MIM: 608766), *MDN1* (MIM: 618200), *MYCBP2* (MIM: 610392), *MYO9A*, *RYR3* (MIM: 180903), and *VPS13D*. Notably, four of these—*CHD3*,[40] *KIAA1109*,[42] *MYO9A*,[43] and *VPS13D*[44,45]—have since had associating Mendelian phenotypes reported to OMIM (MIM: 618205, 617822, 618198, and 607317, respectively). In addition, although not yet reported in OMIM, *RYR3* was recently reported to associate with arthrogryposis.[65]

Similarly, following manual curation of this gene set, searching the OMIM website for reported phenotypic associations, two additional genes were discovered to have associated phenotypes reported in OMIM: *CTD-307407.11* and *SMO*. These genes were not initially recognized by our pipeline as having Mendelian disease associations in OMIM because of a discrepancy in gene symbol. OMIM uses the symbols *BBS1* (MIM: 209901) and *SMOH* (MIM: 601500), respectively, for these genes. Thus, we report a total of 133 candidate Mendelian disease-associated genes without a Mendelian phenotype yet reported in OMIM (Table S4).

Next, we compared these remaining candidate disease-associating genes without OMIM annotations with an alternate set of reported disease-gene associations (UniProt),[66] demonstrating enrichment of UniProt disease associations for genes in this set. Out of 20,382 genes captured in the ES design of the *de novo* enrichment analysis described in this manuscript, 16,568 genes did not have a disease association in OMIM as of May 2019. Of these 16,568 genes, 240 (1.45%) had a disease association in UniProt as of February 2019. Including all UniProt gene symbols—standard and non-standard—in comparison, four, or 2.96% of the 135 candidate disease-associated genes (pre- manual curation), were part of this UniProt disease-associated set: *CELSR1* (MIM: 604523), *MEIS1* (MIM: 601739), *SF1* (MIM: 601516), and *SMO*. (Notably, the "*SF1*" gene in UniProt (MIM: 184757) was different than the "*SF1*" candidate from our analysis.) Thus, 132 candidates remain without any reported disease association in OMIM or UniProt at the time of preparing this manuscript. Permutation analysis sampling 135 random genes without OMIM annotations revealed our candidate set to be significantly enriched for genes with disease

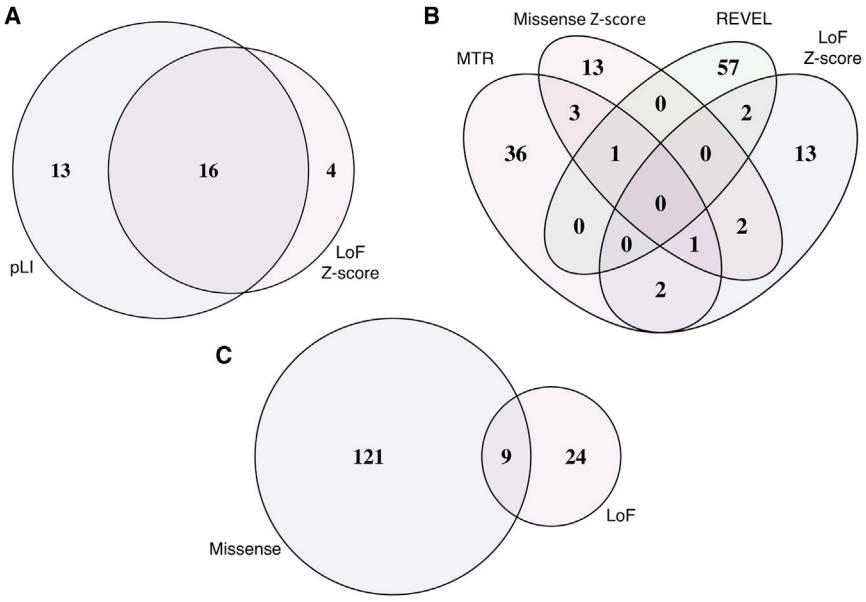


Figure 3. Summary of Candidate Disease-Associated Genes by Category
154 genes flagged as candidate Mendelian disease-associated genes, grouped by constraint-metric query series. Shown are (A and B) variant annotation parameter sweep candidate gene list overlaps: loss-of-function (A) and missense (B); (C) high-priority candidates at the intersection of loss-of-function and missense variant parameter sweep candidate genes.

associations in UniProt, but not in OMIM (again allowing for matching of non-standard gene symbols) (expected = 1.95; p = 0.04482; n = 100,000 permutations).

Finally, we sought replication by intersecting the 154 candidate disease-associated genes with a set of 309 genes harboring 344 *de novo* mutations across a set of 242 ES trios with a wide range of congenital anomalies.[67,68] Of these 309 genes, 216 harbored *de novo* nonsynonymous (missense or stopgain) variants; 78 harbored only synonymous *de novo* variants; 15 genes only carried variants in the 3′ or 5′ UTRs or intronic (including splice-site) variants. Six genes overlapped between the set of 154 candidate disease-associated genes and the 216 genes harboring *de novo* nonsynonymous variants: *AATK* (MIM: 605276), *CELSR1*, *IRF2BPL*, *MYO5C* (MIM: 610022), *ROCK1* (MIM: 601702), and *UBC* (MIM: 191340). Permutation analysis revealed that the set of genes harboring *de novo* nonsynonymous mutations is significantly enriched for genes in our set of 154 candidate disease-associated genes (expected = 1.68; p = 0.005; n = 10,000 permutations). Significantly, variants in *IRF2BPL* were also reported to cause a Mendelian disorder in August 2018, further validating our approach.[55] No genes overlapped between the set of 154 candidate disease-associated genes and the synonymous or noncoding *de novo* variant sets, supporting the model that *de novo* nonsynonymous variants are much more likely to be pathogenic than other *de novo* variants.

HARLEE Facilitates Reverse Genetic Screen Prioritization

HARLEE is also a powerful tool for phenocentric approaches to Mendelian genetic discovery. To illustrate this capability, we sought to identify genes enriched with potentially deleterious genetic variation in individuals with apparent auditory system dysfunction. We first counted the number of samples with ultra-rare variants (see Material and Methods) in each gene across all Mendelian samples in HARLEE, filtering out likely benign variants with a REVEL score less than 0.25, normalizing variant-per-gene count by cohort size. We then repeated this analysis on the subset of all samples whose phenotypic descriptions contain auditory system-related phenotypic keywords. For each gene harboring ultra-rare variants across at least two samples in the auditory phenotype cohort, we then measured cohort-specific enrichment by dividing the ultra-rare variant occurrence rate in the phenocentric cohort by the same rate across all samples in HARLEE.

The top three enriched genes in the auditory phenotype cohort with reported Mendelian phenotypes in OMIM are all directly or indirectly related to hearing loss: (1) variants in the second-most enriched gene overall—*GRHL2* (MIM: 608576), harboring 23× more ultra-rare variants than the background rate in HARLEE—are known to cause autosomal-dominant deafness[69] (MIM: 608641); (2) deficiency of the fifth-most enriched gene—*ECHS1* (MIM: 602292)—is reported to cause deafness in the context of mitochondrial encephalopathy (MIM: 616277);[70] (3) mutations in the thirteenth-most enriched gene—*KDM6A* (MIM: 300128)—cause Kabuki syndrome (MIM: 300867), which leads to hearing loss in approximately 40% of case subjects.[71] These preliminary findings therefore support this strategy of prioritization of genes with phenocentric enrichment for potentially pathogenic variation in HARLEE.[72]

Discussion

We report the application of a Hadoop data lake to Mendelian discovery. Furthermore, we report a large-scale aggregation of 18,696 research and clinical ES data for subjects with suspected Mendelian disease traits. The data were used to discover 132 candidate Mendelian disease-associating genes through an optimization-based genocentric approach. In addition, a phenocentric approach utilized HARLEE to prioritize genes for an ongoing reverse genetic screen for hearing-related genes. These candidates are now available to be studied to further assert proof of Mendelian association.

The methods for identifying the candidates are agnostic to presumed zygosity at a locus and it is likely that the vast majority will display a clinical phenotype with a dominant mode of inheritance—i.e., an AD disease trait. Indeed, of the 19 original candidates recently reported by OMIM to associate with a Mendelian phenotype, 13 have been reported to demonstrate AD inheritance ($p = 0.0835$; binomial probability), and AD inheritance for high-impact variants cannot be ruled out for 2 additional genes (*LINGO1*, *MYO9A*) based on reported cases in OMIM. It also can be anticipated that a significant subset of these genes will eventually reveal more complex architectures such as recessive inheritance or even compound inheritance of coding and non-coding common variant alleles.[73–76] Many case subjects may require extensive follow up, including WGS and scrutiny of databases for genome-wide variant data.

Our approach does not intend to diagnose or solve individual clinical or research case subjects, but rather aims to discover candidate disease-associated genes, constituting cohorts of individuals harboring variants that may or may not be pathogenic in a shared gene. Each of these candidates will ultimately be revealed to either associate with one or more Mendelian disorders or not. For false-positive genes, without a true disease association, none of the individual variants detected in our analysis can be pathogenic. For true-positive genes, only some—but not all—variants in the gene must be pathogenic, notwithstanding the possibility of incidental discovery of a true-positive disease-associated gene where each of the detected variants are actually benign. We anticipate the ratio of pathogenic to benign variants, as well as the ratio of true-positive to false-positive disease-gene associations, to vary across filtering parameters. There may be a negative correlation between maximum discovery density value for a given parameter cutoff value and the associated false-positive rate or benign variant rate. For instance, SIFT and PolyPhen analyses were excluded on the basis of yielding subjectively large candidate disease-associated gene set sizes. Indeed, the discovery density values for SIFT (0.086) and PolyPhen (0.109) are much lower than those of MTR (0.372) or pLI (0.6). However, so long as discovery of Mendelian disease-gene associations and pathogenic variants continues, it is impossible to define true false-positive gene discovery or benign variant rates.

HARLEE is well suited for applications other than the discovery of Mendelian disease-gene associations, including the discovery of previously unrecognized pathogenic variants within genes known to associate with Mendelian disorders and the study of the molecular and genetic models underlying phenotypic expansion.[77,78] In one application, HARLEE was utilized for sample re-analysis and recruitment, identifying three additional individuals with *de novo* variants in *DDX3X* (MIM: 300160), previously missed by experienced geneticists searching the same ES data for the exhaustive set of individuals indicated for the study.[79] The robust yet flexible nature of a Hadoop data lake such as HARLEE is a powerful tool for genocentric reanalysis of individuals sequenced through clinical labs.

An important advantage of the structured HARLEE data management system is the ability to interrogate specific alleles—meeting any possible bioinformatic filtering criteria—in key samples without exposure of individual identifiers. Within the scope of this study, this is achieved by grouping query results into higher-level categories (i.e., domain, gene, or pathway), reporting aggregating variant counts across selected samples. For more outward-facing activities, our approach can be replicated by a front-end web application with permission to access and query full variant-level information, enabling users without sample-level access permission to query genes for recurrence of variants meeting variable bioinformatic filtering criteria in samples with filterable phenotypic descriptors. For example, a query for recurrent loss-of-function variants in *RB1* (MIM: 614041), filtering to include only samples with retinoblastoma, might report a count of two individuals in publicly available ingested datasets. Users who wish to pursue a detailed study involving the resulting individuals could then request to access sample-level information or to be connected with the referring physicians.[80] We anticipate that the flexibility and power allowed by such a framework will accelerate disease-gene association discovery.

The discovery and functional annotation of all ~20–22,000 human genes in the human genome will likely be increasingly dependent upon genocentric analysis. Nevertheless, phenocentric analysis will continue to play an indispensable role in Mendelian disease discovery and clinical re-analysis of extant data. The two approaches are complementary, in the same way that classical reverse genetic approaches complement forward genetic approaches. Ultimately, perhaps the most effective approach for solving the genotype-phenotype puzzle that underlies biological discovery in human genetics and clinical genomics[81] will consist of iterating between genocentric and phenocentric analyses and perspectives. In one hypothetical instance, a disease-gene association may be initially identified through a phenocentric cohort analysis, where multiple individuals in a cohort with highly similar phenotypes share the same type of suspect genetic lesion. Iterating to a genocentric analysis of the candidate gene across large cohorts, utilizing tools such as HARLEE, additional individuals may be revealed to harbor identical genetic lesions with the same or different phenotypes. Re-visiting the phenotypic data of these individuals may reveal molecular explanations of pleiotropy, variable expressivity, or incomplete penetrance[75] unlikely to be ascertained solely through phenocentric approaches.

Other large-scale genocentric projects have been recently reported, each with a unique scope or angle. The ongoing Deciphering Developmental Disorders (DDD) project is an important resource investigating the genotype-phenotype relationship in the context of *de*

novo mutations in developmental disorders.[7] Similarly, the ongoing Human Knockout Project is intended to characterize the extent and impact of homozygous loss-of-function (LoF) variation in populations with elevated rates of consanguinity, such as Pakistani and Finnish populations.[23] Our efforts have built upon this foundation of large-scale genocentric analysis, expanding the paradigm into the area of broadly defined suspected Mendelian disease traits, introducing a high-yield methodology initially agnostic to genotype or *de novo* inheritance status, relevant to both missense and loss-of-function variation.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.09.027.

Acknowledgments

This work was supported in part by grants UM1 HG008898 from the National Human Genome Research Institute (NHGRI) to the Baylor College of Medicine Center for Common Disease Genetics; UM1 HG006542 from the NHGRI/National Heart, Lung, and Blood Institute (NHLBI) to the Baylor Hopkins Center for Mendelian Genomics; R01 NS058529 and R35 NS105078 (J.R.L.) from the National Institute of Neurological Disorders and Stroke (NINDS); and P50 DK096415 (N.K.) from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This work was also supported in part by the Baylor College of Medicine President's Circle Precision Medicine/Population Health Initiative. A.W.H. was supported in part by NIH T32 GM08307-26 and The Cullen Foundation. J.E.P. was supported by NHGRI K08 HG008986.

We thank Huda Y. Zoghbi and Joshua M. Shulman for their insight and feedback as related to genocentric and phenocentric studies of human disease. We thank Jeremy Easton-Marks, Simon White, Joshua Traynelis, Piyushkumar Panchel, and Brian Palazzo for assistance with data architecture, data wrangling, and systems administration. We thank Stephen Wilson for sharing archived OMIM database downloads.

Declaration of Interests

J.R.L. has stock ownership in 23andMe and Lasergen, is a paid consultant for Regeneron Pharmaceuticals, and is a coinventor on multiple US and European patents related to molecular diagnostics for inherited neuropathies, eye diseases, and bacterial genomic fingerprinting. The Department of Molecular and Human Genetics at Baylor College of Medicine derives revenue from the chromosomal microarray analysis and clinical exome sequencing offered in the Baylor Genetics Laboratory (http://baylorgenetics.com).

Received: June 6, 2019
Accepted: September 27, 2019
Published: October 24, 2019

Web Resources

Codified Genomics, https://www.codifiedgenomics.com
HARLEE analysis scripts and notebooks, https://github.com/BCM-HGSC/HARLEE_analysis
Mercury, https://www.hgsc.bcm.edu/software/mercury
OMIM, https://www.omim.org

References

1. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N. Engl. J. Med. *369*, 1502–1511.
2. Bamshad, M.J., Shendure, J.A., Valle, D., Hamosh, A., Lupski, J.R., Gibbs, R.A., Boerwinkle, E., Lifton, R.P., Gerstein, M., Gunel, M., et al.; Centers for Mendelian Genomics (2012). The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am. J. Med. Genet. A. *158A*, 1523–1525.
3. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al.; Centers for Mendelian Genomics (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am. J. Hum. Genet. *97*, 199–215.
4. Posey, J.E., Rosenfeld, J.A., James, R.A., Bainbridge, M., Niu, Z., Wang, X., Dhar, S., Wiszniewski, W., Akdemir, Z.H.C., Gambin, T., et al. (2016). Molecular diagnostic experience of whole-exome sequencing in adult patients. Genet. Med. *18*, 678–685.
5. Posey, J.E., O'Donnell-Luria, A.H., Chong, J.X., Harel, T., Jhangiani, S.N., Coban Akdemir, Z.H., Buyske, S., Pehlivan, D., Carvalho, C.M.B., Baxter, S., et al.; Centers for Mendelian Genomics (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet. Med. *21*, 798–812.
6. McKusick, V.A. (2007). Mendelian Inheritance in Man and its online version, OMIM. Am. J. Hum. Genet. *80*, 588–604.
7. McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al.; Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.
8. White, J., Beck, C.R., Harel, T., Posey, J.E., Jhangiani, S.N., Tang, S., Farwell, K.D., Powis, Z., Mendelsohn, N.J., Baker, J.A., et al. (2016). POGZ truncating alleles cause syndromic intellectual disability. Genome Med. *8*, 3.
9. Stessman, H.A.F., Willemsen, M.H., Fenckova, M., Penn, O., Hoischen, A., Xiong, B., Wang, T., Hoekzema, K., Vives, L., Vogel, I., et al. (2016). Disruption of POGZ Is Associated with Intellectual Disability and Autism Spectrum Disorders. Am. J. Hum. Genet. *98*, 541–552.
10. Dentici, M.L., Niceta, M., Pantaleoni, F., Barresi, S., Bencivenga, P., Dallapiccola, B., Digilio, M.C., and Tartaglia, M. (2017). Expanding the phenotypic spectrum of truncating POGZ mutations: Association with CNS malformations, skeletal abnormalities, and distinctive facial dysmorphism. Am. J. Med. Genet. A. *173*, 1965–1969.
11. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. *2012*, 251364.
12. Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. Trends Genet. *24*, 133–141.

13. Austin, C.P., Battey, J.F., Bradley, A., Bucan, M., Capecchi, M., Collins, F.S., Dove, W.F., Duyk, G., Dymecki, S., Eppig, J.T., et al. (2004). The knockout mouse project. Nat. Genet. *36*, 921–924.
14. Ghosh, R., Oak, N., and Plon, S.E. (2017). Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. *18*, 225.
15. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.
16. Traynelis, J., Silk, M., Wang, Q., Berkovic, S.F., Liu, L., Ascher, D.B., Balding, D.J., and Petrovski, S. (2017). Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. Genome Res. *27*, 1715–1729.
17. Coban-Akdemir, Z., White, J.J., Song, X., Jhangiani, S.N., Fatih, J.M., Gambin, T., Bayram, Y., Chinn, I.K., Karaca, E., Punetha, J., et al.; Baylor-Hopkins Center for Mendelian Genomics (2018). Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. Am. J. Hum. Genet. *103*, 171–187.
18. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
19. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
20. Kosmicki, J.A., Samocha, K.E., Howrigan, D.P., Sanders, S.J., Slowikowski, K., Lek, M., Karczewski, K.J., Cutler, D.J., Devlin, B., Roeder, K., et al. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat. Genet. *49*, 504–510.
21. Lupski, J.R., Belmont, J.W., Boerwinkle, E., and Gibbs, R.A. (2011). Clan Genomics and the Complex Architecture of Human Disease. Cell *147*, 32–43.
22. Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y., Kosmicki, J.A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nat. Genet. *50*, 1161–1170.
23. Saleheen, D., Natarajan, P., Armean, I.M., Zhao, W., Rasheed, A., Khetarpal, S.A., Won, H.-H., Karczewski, K.J., O'Donnell-Luria, A.H., Samocha, K.E., et al. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature *544*, 235–239.
24. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.
25. Taylor, R.C. (2010). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics *11* (*Suppl 12*), S1.
26. Niemenmaa, M., Kallio, A., Schumacher, A., Klemelä, P., Korpelainen, E., and Heljanko, K. (2012). Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. Bioinformatics *28*, 876–877.
27. O'Driscoll, A., Daugelaite, J., and Sleator, R.D. (2013). 'Big data', Hadoop and cloud computing in genomics. J. Biomed. Inform. *46*, 774–781.
28. Zou, Q., Li, X.B., Jiang, W.R., Lin, Z.Y., Li, G.L., and Chen, K. (2014). Survey of MapReduce frame operation in bioinformatics. Brief. Bioinform. *15*, 637–647.
29. Siretskiy, A., Sundqvist, T., Voznesenskiy, M., and Spjuth, O. (2015). A quantitative assessment of the Hadoop framework for analyzing massively parallel DNA sequencing data. Gigascience *4*, 26.
30. Hodor, P., Chawla, A., Clark, A., and Neal, L. (2016). cl-dash: rapid configuration and deployment of Hadoop clusters for bioinformatics research in the cloud. Bioinformatics *32*, 301–303.
31. O'Driscoll, A., Belogrudov, V., Carroll, J., Kropp, K., Walsh, P., Ghazal, P., and Sleator, R.D. (2015). HBLAST: Parallelised sequence similarity–A Hadoop MapReducable basic local alignment search tool. J. Biomed. Inform. *54*, 58–64.
32. de Castro, M.R., Tostes, C.D.S., Dávila, A.M.R., Senger, H., and da Silva, F.A.B. (2017). SparkBLAST: scalable BLAST processing using in-memory operations. BMC Bioinformatics *18*, 318.
33. Yin, Z., Lan, H., Tan, G., Lu, M., Vasilakos, A.V., and Liu, W. (2017). Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges. Comput. Struct. Biotechnol. J. *15*, 403–411.
34. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al.; ACMG Laboratory Quality Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.
35. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.
36. DeMari, J., Mroske, C., Tang, S., Nimeh, J., Miller, R., and Lebel, R.R. (2016). CLTC as a clinically novel gene associated with multiple malformations and developmental delay. Am. J. Med. Genet. A. *170A*, 958–966.
37. Haijes, H.A., Koster, M.J.E., Rehmann, H., Li, D., Hakonarson, H., Cappuccio, G., Hancarova, M., Lehalle, D., Reardon, W., Schaefer, G.B., et al. (2019). De Novo Heterozygous POLR2A Variants Cause a Neurodevelopmental Syndrome with Profound Infantile-Onset Hypotonia. Am. J. Hum. Genet. *105*, 283–301.
38. Lessel, D., Schob, C., Küry, S., Reijnders, M.R.F., Harel, T., Eldomery, M.K., Coban-Akdemir, Z., Denecke, J., Edvardson, S., Colin, E., et al.; DDD study; and C4RCD Research Group (2017). De Novo Missense Mutations in DHX30 Impair Global Translation and Cause a Neurodevelopmental Disorder. Am. J. Hum. Genet. *101*, 716–724.
39. Deardorff, M.A., Kaur, M., Yaeger, D., Rampuria, A., Korolev, S., Pie, J., Gil-Rodríguez, C., Arnedo, M., Loeys, B., Kline, A.D., et al. (2007). Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of cornelia de Lange syndrome with predominant mental retardation. Am. J. Hum. Genet. *80*, 485–494.
40. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al.; DDD study (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with

macrocephaly and impaired speech and language. Nat. Commun. *9*, 4619.
41. Helbig, K.L., Mroske, C., Moorthy, D., Sajan, S.A., and Velinov, M. (2017). Biallelic loss-of-function variants in DOCK3 cause muscle hypotonia, ataxia, and intellectual disability. Clin. Genet. *92*, 430–433.
42. Gueneau, L., Fish, R.J., Shamseldin, H.E., Voisin, N., Tran Mau-Them, F., Preiksaitiene, E., Monroe, G.R., Lai, A., Putoux, A., Allias, F., et al.; DDD Study (2018). KIAA1109 Variants Are Associated with a Severe Disorder of Brain Development and Arthrogryposis. Am. J. Hum. Genet. *102*, 116–132.
43. O'Connor, E., Töpf, A., Müller, J.S., Cox, D., Evangelista, T., Colomer, J., Abicht, A., Senderek, J., Hasselmann, O., Yaramis, A., et al. (2016). Identification of mutations in the MYO9A gene in patients with congenital myasthenic syndrome. Brain *139*, 2143–2153.
44. Seong, E., Insolera, R., Dulovic, M., Kamsteeg, E.-J., Trinh, J., Brüggemann, N., Sandford, E., Li, S., Ozel, A.B., Li, J.Z., et al. (2018). Mutations in VPS13D lead to a new recessive ataxia with spasticity and mitochondrial defects. Ann. Neurol. *83*, 1075–1088.
45. Gauthier, J., Meijer, I.A., Lessel, D., Mencacci, N.E., Krainc, D., Hempel, M., Tsiakas, K., Prokisch, H., Rossignol, E., Helm, M.H., et al. (2018). Recessive mutations in >VPS13D cause childhood onset movement disorders. Ann. Neurol. *83*, 1089–1095.
46. Cogné, B., Ehresmann, S., Beauregard-Lacroix, E., Rousseau, J., Besnard, T., Garcia, T., Petrovski, S., Avni, S., McWalter, K., Blackburn, P.R., et al.; CAUSES Study; and Deciphering Developmental Disorders study (2019). Missense variants in the histone acetyltransferase complex component gene TRRAP cause autism and syndromic intellectual disability. Am. J. Hum. Genet. *104*, 530–541.
47. Damaj, L., Lupien-Meilleur, A., Lortie, A., Riou, É., Ospina, L.H., Gagnon, L., Vanasse, C., and Rossignol, E. (2015). CACNA1A haploinsufficiency causes cognitive impairment, autism and epileptic encephalopathy with mild cerebellar symptoms. Eur. J. Hum. Genet. *23*, 1505–1512.
48. Travaglini, L., Nardella, M., Bellacchio, E., D'Amico, A., Capuano, A., Frusciante, R., Di Capua, M., Cusmai, R., Barresi, S., Morlino, S., et al. (2017). Missense mutations of CACNA1A are a frequent cause of autosomal dominant nonprogressive congenital ataxia. Eur. J. Paediatr. Neurol. *21*, 450–456.
49. Heyne, H.O., Singh, T., Stamberger, H., Abou Jamra, R., Caglayan, H., Craiu, D., De Jonghe, P., Guerrini, R., Helbig, K.L., Koeleman, B.P.C., et al.; EuroEPINOMICS RES Consortium (2018). De novo variants in neurodevelopmental disorders with epilepsy. Nat. Genet. *50*, 1048–1053.
50. Schlingmann, K.P., Bandulik, S., Mammen, C., Tarailo-Graovac, M., Holm, R., Baumann, M., König, J., Lee, J.J.Y., Drögemöller, B., Imminger, K., et al. (2018). Germline De Novo Mutations in ATP1A1 Cause Renal Hypomagnesemia, Refractory Seizures, and Intellectual Disability. Am. J. Hum. Genet. *103*, 808–816.
51. Lassuthova, P., Rebelo, A.P., Ravenscroft, G., Lamont, P.J., Davis, M.R., Manganelli, F., Feely, S.M., Bacon, C., Brožková, D.Š., Haberlova, J., et al. (2018). Mutations in ATP1A1 Cause Dominant Charcot-Marie-Tooth Type 2. Am. J. Hum. Genet. *102*, 505–514.
52. Helbig, K.L., Lauerer, R.J., Bahr, J.C., Souza, I.A., Myers, C.T., Uysal, B., Schwarz, N., Gandini, M.A., Huang, S., Keren, B., et al.; Task Force for Neonatal Genomics; and Deciphering Developmental Disorders Study (2018). De Novo Pathogenic Variants in CACNA1E Cause Developmental and Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. Am. J. Hum. Genet. *103*, 666–678.
53. Gregor, A., Sadleir, L.G., Asadollahi, R., Azzarello-Burri, S., Battaglia, A., Ousager, L.B., Boonsawat, P., Bruel, A.-L., Buchert, R., Calpena, E., et al.; University of Washington Center for Mendelian Genomics; and DDD Study (2018). De Novo Variants in the F-Box Protein FBXO11 in 20 Individuals with a Variable Neurodevelopmental Disorder. Am. J. Hum. Genet. *103*, 305–316.
54. Fritzen, D., Kuechler, A., Grimmel, M., Becker, J., Peters, S., Sturm, M., Hundertmark, H., Schmidt, A., Kreiß, M., Strom, T.M., et al. (2018). De novo FBXO11 mutations are associated with intellectual disability and behavioural anomalies. Hum. Genet. *137*, 401–411.
55. Marcogliese, P.C., Shashi, V., Spillmann, R.C., Stong, N., Rosenfeld, J.A., Koenig, M.K., Martínez-Agosto, J.A., Herzog, M., Chen, A.H., Dickson, P.I., et al.; Program for Undiagnosed Diseases (UD-PrOZA); and Undiagnosed Diseases Network (2018). IRF2BPL Is Associated with Neurological Phenotypes. Am. J. Hum. Genet. *103*, 245–260.
56. Faundes, V., Newman, W.G., Bernardini, L., Canham, N., Clayton-Smith, J., Dallapiccola, B., Davies, S.J., Demos, M.K., Goldman, A., Gill, H., et al.; Clinical Assessment of the Utility of Sequencing and Evaluation as a Service (CAUSES) Study; and Deciphering Developmental Disorders (DDD) Study (2018). Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. Am. J. Hum. Genet. *102*, 175–187.
57. Ansar, M., Riazuddin, S., Sarwar, M.T., Makrythanasis, P., Paracha, S.A., Iqbal, Z., Khan, J., Assir, M.Z., Hussain, M., Razzaq, A., et al. (2018). Biallelic variants in LINGO1 are associated with autosomal recessive intellectual disability, microcephaly, speech and motor delay. Genet. Med. *20*, 778–784.
58. Dobyns, W.B., Aldinger, K.A., Ishak, G.E., Mirzaa, G.M., Timms, A.E., Grout, M.E., Dremmen, M.H.G., Schot, R., Vandervore, L., van Slegtenhorst, M.A., et al.; University of Washington Center for Mendelian Genomics; and Center for Mendelian Genomics at the Broad Institute of MIT and Harvard (2018). MACF1 Mutations Encoding Highly Conserved Zinc-Binding Residues of the GAR Domain Cause Defects in Neuronal Migration and Axon Guidance. Am. J. Hum. Genet. *103*, 1009–1021.
59. Tripathy, R., Leca, I., van Dijk, T., Weiss, J., van Bon, B.W., Sergaki, M.C., Gstrein, T., Breuss, M., Tian, G., Bahi-Buisson, N., et al. (2018). Mutations in MAST1 Cause Mega-Corpus-Callosum Syndrome with Cerebellar Hypoplasia and Cortical Malformations. Neuron *100*, 1354–1368.e5.
60. Wang, L., Feng, Y., Yan, D., Qin, L., Grati, M., Mittal, R., Li, T., Sundhari, A.K., Liu, Y., Chapagain, P., et al. (2018). A dominant variant in the PDE1C gene is associated with nonsyndromic hearing loss. Hum. Genet. *137*, 437–446.
61. Zaman, T., Helbig, I., Božović, I.B., DeBrosse, S.D., Bergqvist, A.C., Wallis, K., Medne, L., Maver, A., Peterlin, B., Helbig, K.L., et al. (2018). Mutations in SCN3A cause early infantile epileptic encephalopathy. Ann. Neurol. *83*, 703–717.
62. Stevens, S.J.C., van der Schoot, V., Leduc, M.S., Rinne, T., Lalani, S.R., Weiss, M.M., van Hagen, J.M., Lachmeijer, A.M.A., Stockler-Ipsiroglu, S.G., Lehman, A., Brunner, H.G.; and CAUSES Study (2018). De novo mutations in the SET nuclear proto-oncogene, encoding a component of the inhibitor of histone acetyltransferases (INHAT) complex in patients with

nonsyndromic intellectual disability. Hum. Mutat. *39*, 1014–1023.

63. Liu, N., Schoch, K., Luo, X., Pena, L.D.M., Bhavana, V.H., Kukolich, M.K., Stringer, S., Powis, Z., Radtke, K., Mroske, C., et al.; Undiagnosed Diseases Network (UDN) (2018). Functional variants in TBX2 are associated with a syndromic cardiovascular and skeletal developmental disorder. Hum. Mol. Genet. *27*, 2454–2465.
64. Vetrini, F., McKee, S., Rosenfeld, J.A., Suri, M., Lewis, A.M., Nugent, K.M., Roeder, E., Littlejohn, R.O., Holder, S., Zhu, W., et al.; DDD study (2019). De novo and inherited TCF20 pathogenic variants are associated with intellectual disability, dysmorphic features, hypotonia, and neurological impairments with similarities to Smith-Magenis syndrome. Genome Med. *11*, 12.
65. Pehlivan, D., Bayram, Y., Gunes, N., Coban Akdemir, Z., Shukla, A., Bierhals, T., Tabakci, B., Sahin, Y., Gezdirici, A., Faith, J.M., et al. (2019). The Genomics of Arthrogryposis, a Complex Trait: Candidate Genes and Further Evidence for Oligogenic Inheritance. Am. J. Hum. Genet. *105*, 132–150.
66. Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., et al.; The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45* (D1), D158–D169.
67. Katsanis, N., Cotten, M., and Angrist, M. (2012). Exome and genome sequencing of neonates with neurodevelopmental disorders. Future Neurol. *7*, 655–658.
68. Katsanis, S.H., Minear, M.A., Sadeghpour, A., Cope, H., Perilla, Y., Cook-Deegan, R., Duke Task Force for Neonatal Genomics, Katsanis, N., Davis, E.E., and Angrist, M. (2018). Participant-Partners in Genetic Research: An Exome Study with Families of Children with Unexplained Medical Conditions. J. Participat. Med. *10*, e2.
69. Vona, B., Nanda, I., Neuner, C., Müller, T., and Haaf, T. (2013). Confirmation of GRHL2 as the gene for the DFNA28 locus. Am. J. Med. Genet. A. *161A*, 2060–2065.
70. Haack, T.B., Jackson, C.B., Murayama, K., Kremer, L.S., Schaller, A., Kotzaeridou, U., de Vries, M.C., Schottmann, G., Santra, S., Büchner, B., et al. (2015). Deficiency of ECHS1 causes mitochondrial encephalopathy with cardiac involvement. Ann. Clin. Transl. Neurol. *2*, 492–509.
71. Tekin, M., Fitoz, S., Arici, S., Cetinkaya, E., and Incesulu, A. (2006). Niikawa-Kuroki (Kabuki) syndrome with congenital sensorineural deafness: evidence for a wide spectrum of inner ear abnormalities. Int. J. Pediatr. Otorhinolaryngol. *70*, 885–889.
72. Gonzaga-Jauregui, C., Harel, T., Gambin, T., Kousi, M., Griffin, L.B., Francescatto, L., Ozes, B., Karaca, E., Jhangiani, S.N., Bainbridge, M.N., et al.; Baylor-Hopkins Center for Mendelian Genomics (2015). Exome Sequence Analysis Suggests that Genetic Burden Contributes to Phenotypic Variability and Complex Neuropathy. Cell Rep. *12*, 1169–1183.
73. Wu, N., Ming, X., Xiao, J., Wu, Z., Chen, X., Shinawi, M., Shen, Y., Yu, G., Liu, J., Xie, H., et al. (2015). TBX6 null variants and a common hypomorphic allele in congenital scoliosis. N. Engl. J. Med. *372*, 341–350.
74. Liu, J., Zhou, Y., Liu, S., Song, X., Yang, X.Z., Fan, Y., Chen, W., Akdemir, Z.C., Yan, Z., Zuo, Y., et al.; DISCO (Deciphering disorders Involving Scoliosis and COmorbidities) Study (2018). The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. Hum. Genet. *137*, 553–567.
75. Liu, J., Wu, N., Yang, N., Takeda, K., Chen, W., Li, W., Du, R., Liu, S., Zhou, Y., Zhang, L., et al.; Deciphering Disorders Involving Scoliosis and COmorbidities (DISCO) study; Japan Early Onset Scoliosis Research Group; and Baylor-Hopkins Center for Mendelian Genomics (2019). TBX6-associated congenital scoliosis (TACS) as a clinically distinguishable subtype of congenital scoliosis: further evidence supporting the compound inheritance and TBX6 gene dosage model. Genet. Med. *21*, 1548–1558.
76. Yang, N., Wu, N., Zhang, L., Zhao, Y., Liu, J., Liang, X., Ren, X., Li, W., Chen, W., Dong, S., et al. (2019). TBX6 compound inheritance leads to congenital vertebral malformations in humans and mice. Hum. Mol. Genet. *28*, 539–547.
77. Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H., Walkiewicz, M., Bi, W., Xiao, R., Ding, Y., et al. (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. N. Engl. J. Med. *376*, 21–31.
78. Karaca, E., Posey, J.E., Coban Akdemir, Z., Pehlivan, D., Harel, T., Jhangiani, S.N., Bayram, Y., Song, X., Bahrambeigi, V., Yuregir, O.O., et al. (2018). Phenotypic expansion illuminates multilocus pathogenic variation. Genet. Med. *20*, 1528–1537.
79. Wang, X., Posey, J.E., Rosenfeld, J.A., Bacino, C.A., Scaglia, F., Immken, L., Harris, J.M., Hickey, S.E., Mosher, T.M., Slavotinek, A., et al.; Undiagnosed Diseases Network (2018). Phenotypic expansion in *DDX3X* - a common cause of intellectual disability in females. Ann. Clin. Transl. Neurol. *5*, 1277–1285.
80. Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. Hum. Mutat. *36*, 915–921.
81. Veltman, J.A., and Lupski, J.R. (2015). From genes to genomes in the clinic. Genome Med. *7*, 78.

The American Journal of Human Genetics, Volume *105*

Supplemental Data

A Genocentric Approach

to Discovery of Mendelian Disorders

Adam W. Hansen, Mullai Murugan, He Li, Michael M. Khayat, Liwen Wang, Jill Rosenfeld, B. Kim Andrews, Shalini N. Jhangiani, Zeynep H. Coban Akdemir, Fritz J. Sedlazeck, Allison E. Ashley-Koch, Pengfei Liu, Donna M. Muzny, Task Force for Neonatal Genomics, Erica E. Davis, Nicholas Katsanis, Aniko Sabo, Jennifer E. Posey, Yaping Yang, Michael F. Wangler, Christine M. Eng, V. Reid Sutton, James R. Lupski, Eric Boerwinkle, and Richard A. Gibbs

Supplemental Figures

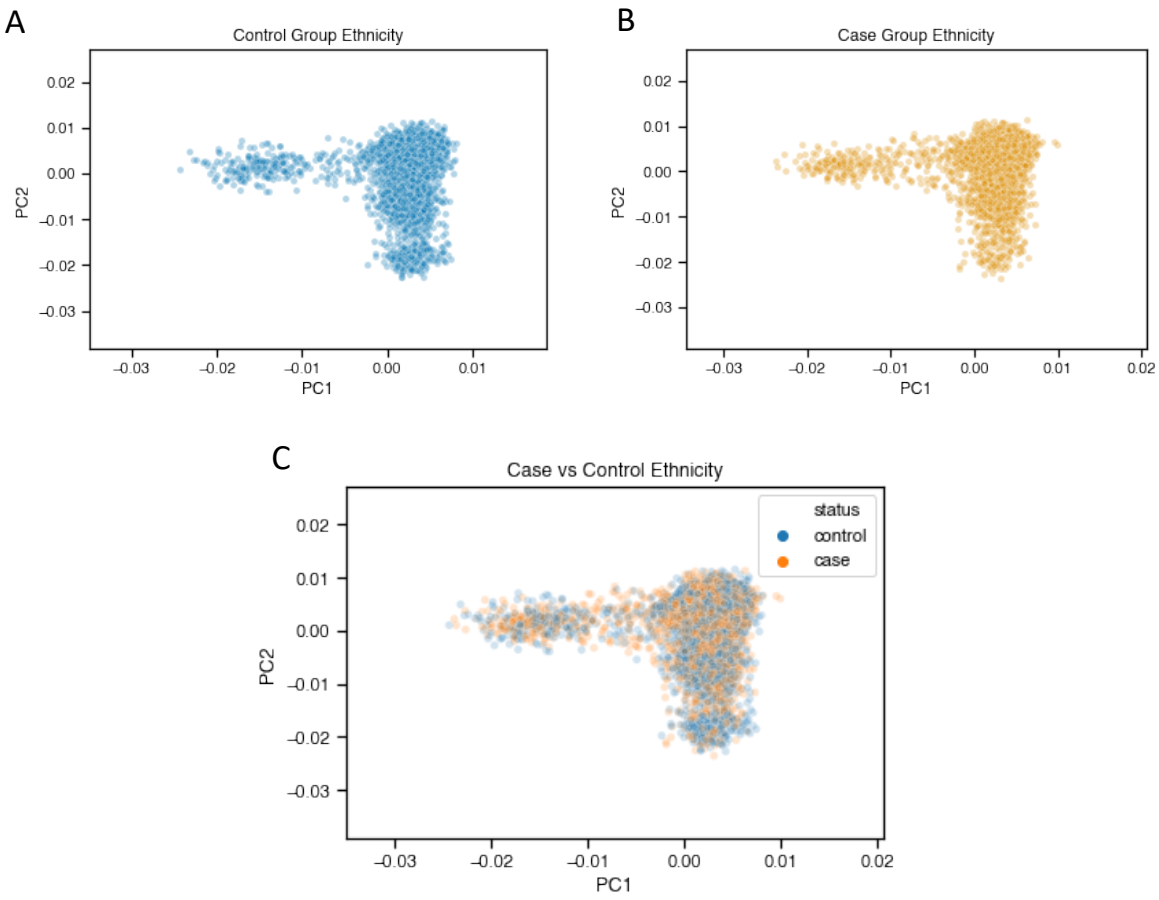


Figure S1 – Case and control ethnicities. Genomic PCA plots showing ethnicity distribution of case vs control samples: A) Control group; B) Random sampling of the case group equal in size to the control group; C) An overlay of plots A and B.

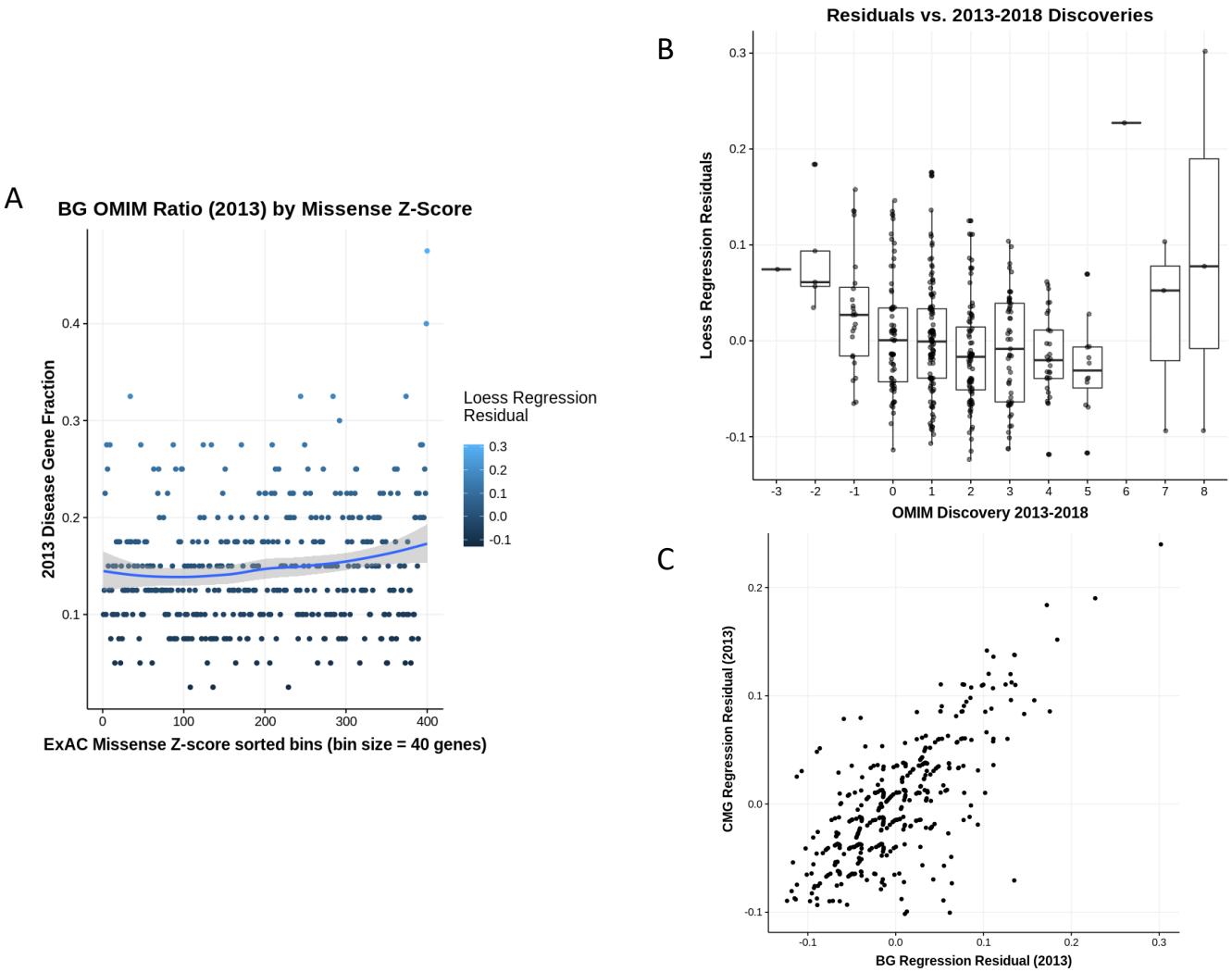


Figure S2 – OMIM ratio outlier regression residual analysis. A) All gene hits are sorted by missense z-score then binned into groups of a consistent, tractable size (40 genes). 2013 OMIM disease gene fraction and a loess regression curve are plotted (95% confidence interval shaded in gray). B) OMIM 2018 vs 2013 disease annotations are compared to quantify disease gene discovery. Gene lists with the higher discovery tend to have the lowest residuals. However, the highest discovery is observed in outliers with the most extreme constraint scores. C) Loess regression residuals of missense z-score bin index vs. 2013 OMIM ratio correlate strongly across independent BG and CMG data sets.

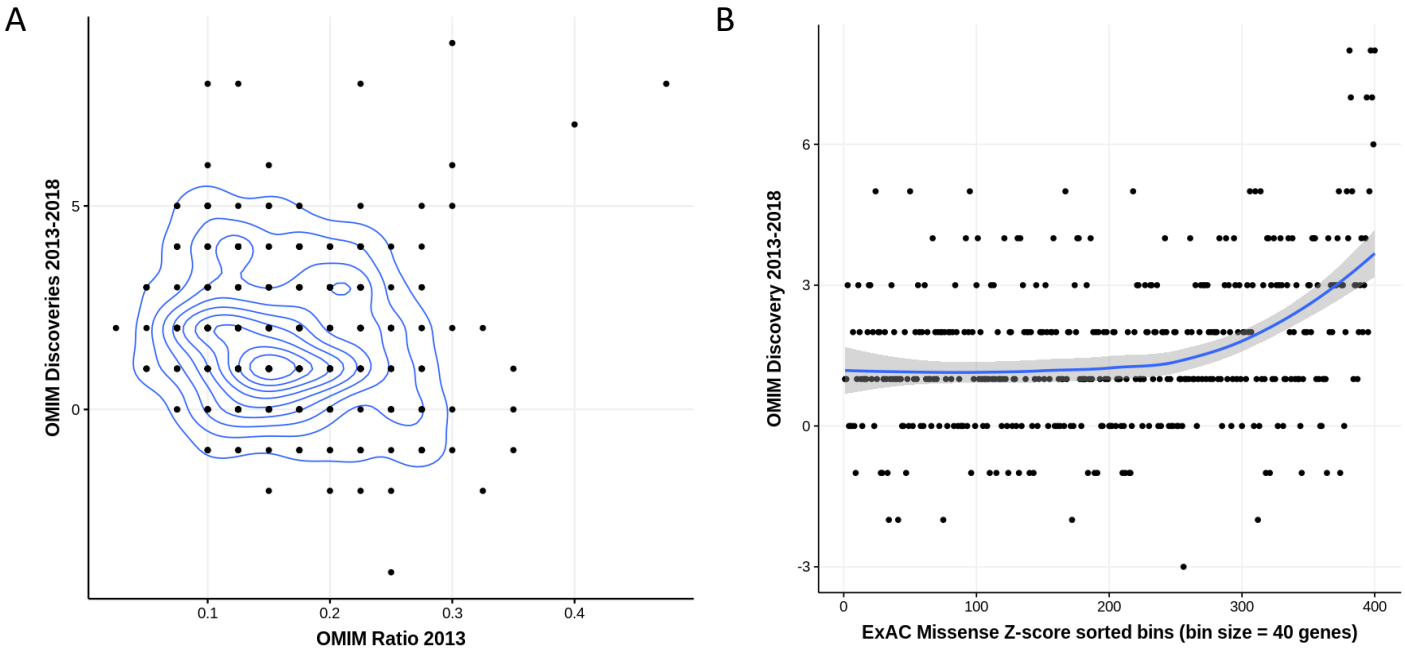


Figure S3 – OMIM 2013-2018 discoveries by other gene set features. In addition to regression residuals (Figure S2), other features of the missense intolerance z-score query series were qualitatively tested for association with discovery through visualization. A) OMIM ratio high outliers (≥ 0.4) had a striking association with discovery. B) Missense intolerance z-score also correlated with discovery, with loess regression plotted in blue, with 95% confidence intervals shaded gray.

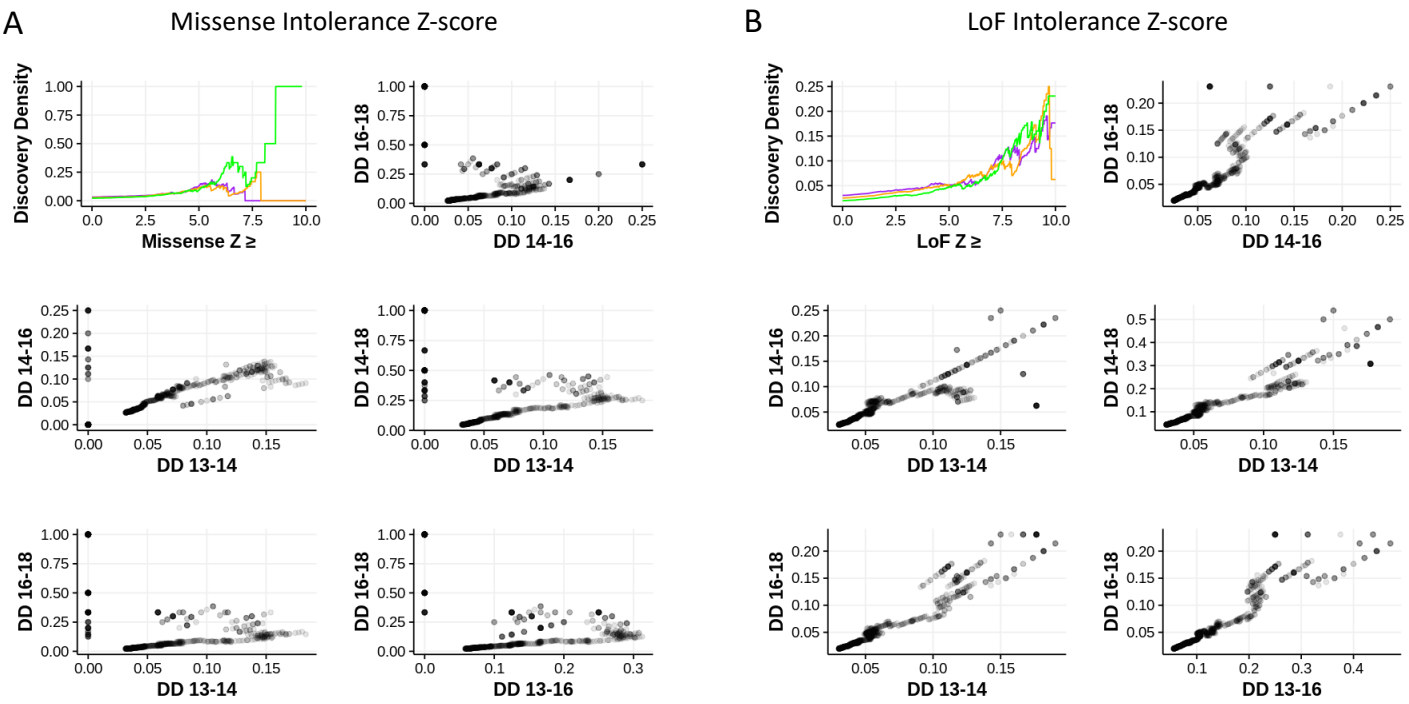


Figure S4 – Past vs. future OMIM discovery density for missense variant parameter sweep query series over time. Past discovery density (DD) consistently correlates with future DD. A) Missense Intolerance Z-score, B) LoF Intolerance Z-score. For the upper-left panels, purple is 2013-2014 discovery density, orange is 2014-2016 discovery density, and green is 2016-2018 discovery density. For the other panels, each point represents a gene query at a fixed z-score cutoff.

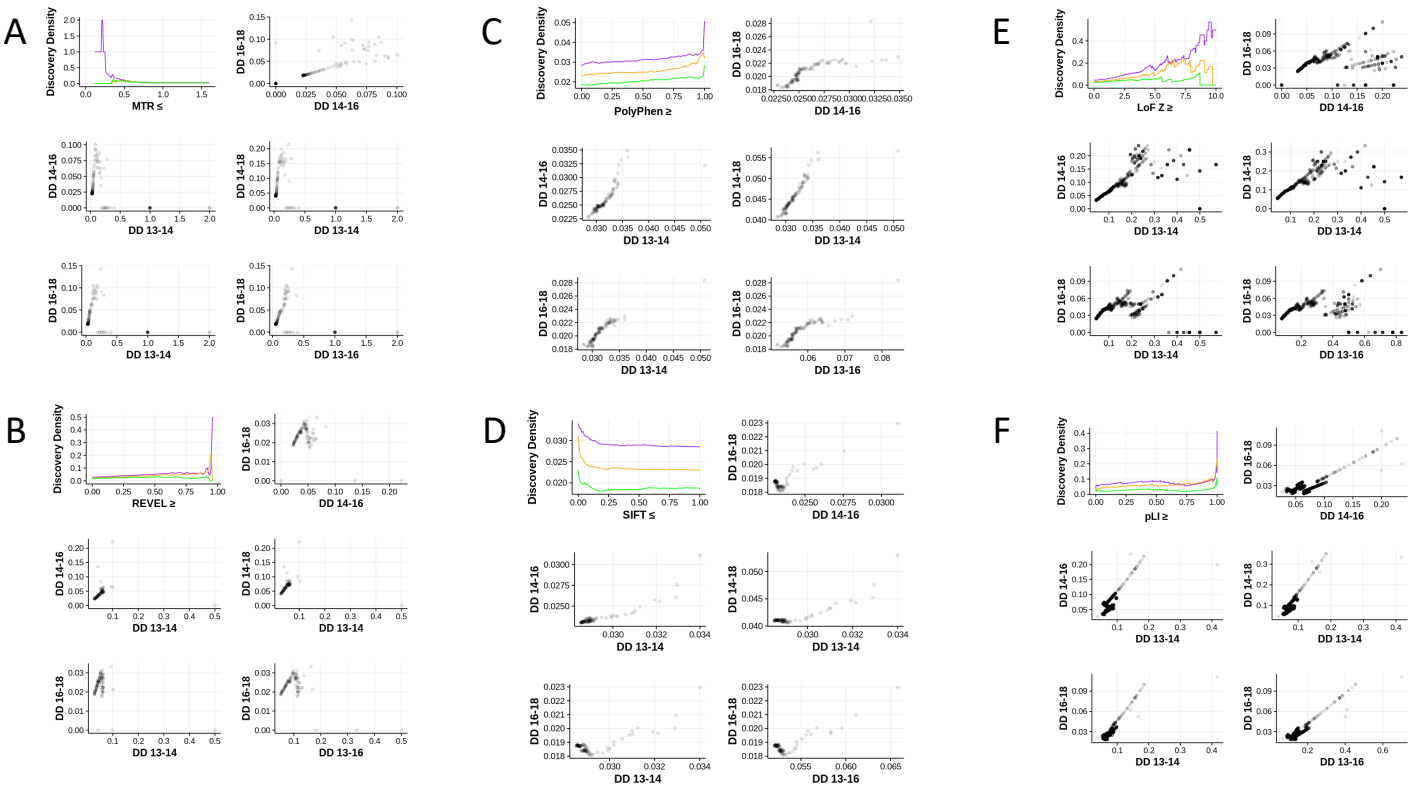
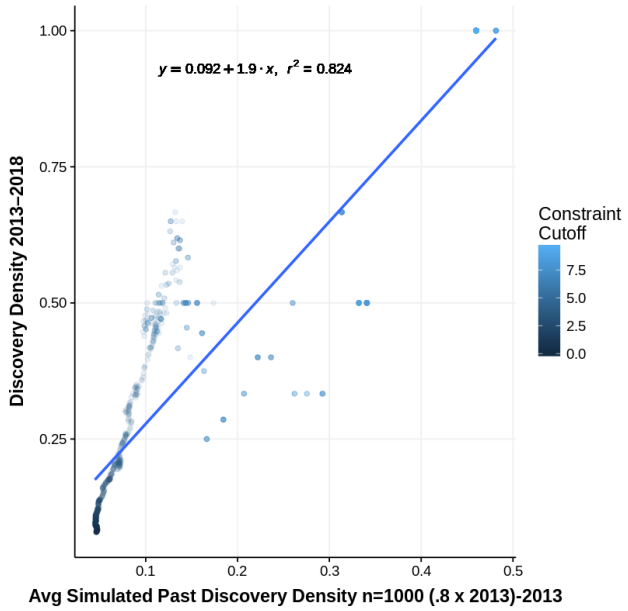


Figure S5 – Past vs. future OMIM discovery density for each annotation parameter sweep query series over time. Past discovery density (DD) correlates with future DD, supporting the strategy of selecting DD-optimizing queries and associated gene lists as candidate disease-associating genes. A-D) Variant-level missense variant parameter sweeps: MTR (A), REVEL (B), PolyPhen (C), and SIFT (D). E-F) LoF variant parameter sweeps: LoF Intolerance Z-score (E) and pLI (F).

A



B

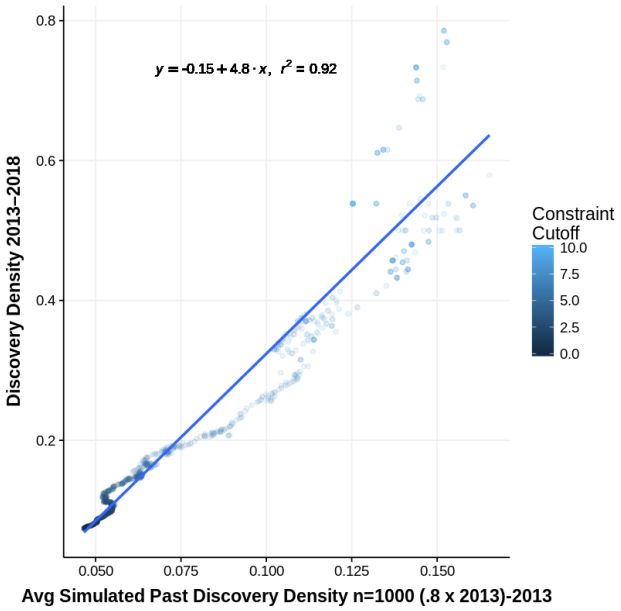


Figure S6 – Simulated (n=1,000) pre-2013 OMIM - 2013 vs. OMIM 2013 - 2018 DD. Simulated past DD correlates with 2013-2018 DD, supporting the strategy of selecting discovery density-optimizing queries and associated gene lists as candidate disease genes. A) Missense Intolerance Z-score, B) LoF Intolerance Z-score (for missense variants).

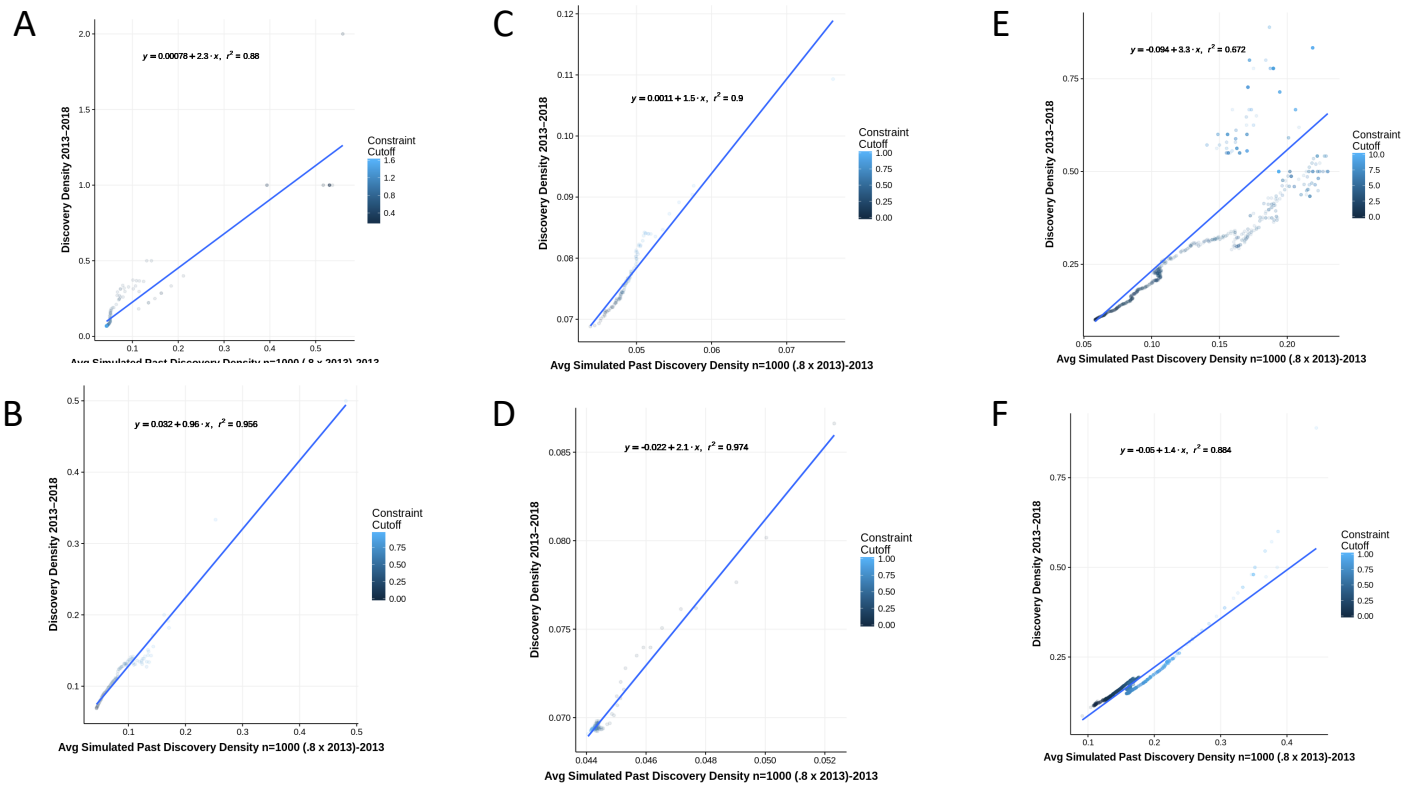


Figure S7 – Simulated (n=1,000) pre-2013 OMIM - 2013 vs. OMIM 2013 - 2018 discovery density. Simulated past DD correlates with 2013-2018 DD, supporting the strategy of selecting discovery density-optimizing queries and associated gene lists as candidate disease genes. A-D) Variant-level missense variant parameter sweeps: MTR (A), REVEL (B), PolyPhen (C), and SIFT (D). E-F) LoF variant parameter sweeps: LoF Intolerance Z-score (E) and pLI (F).

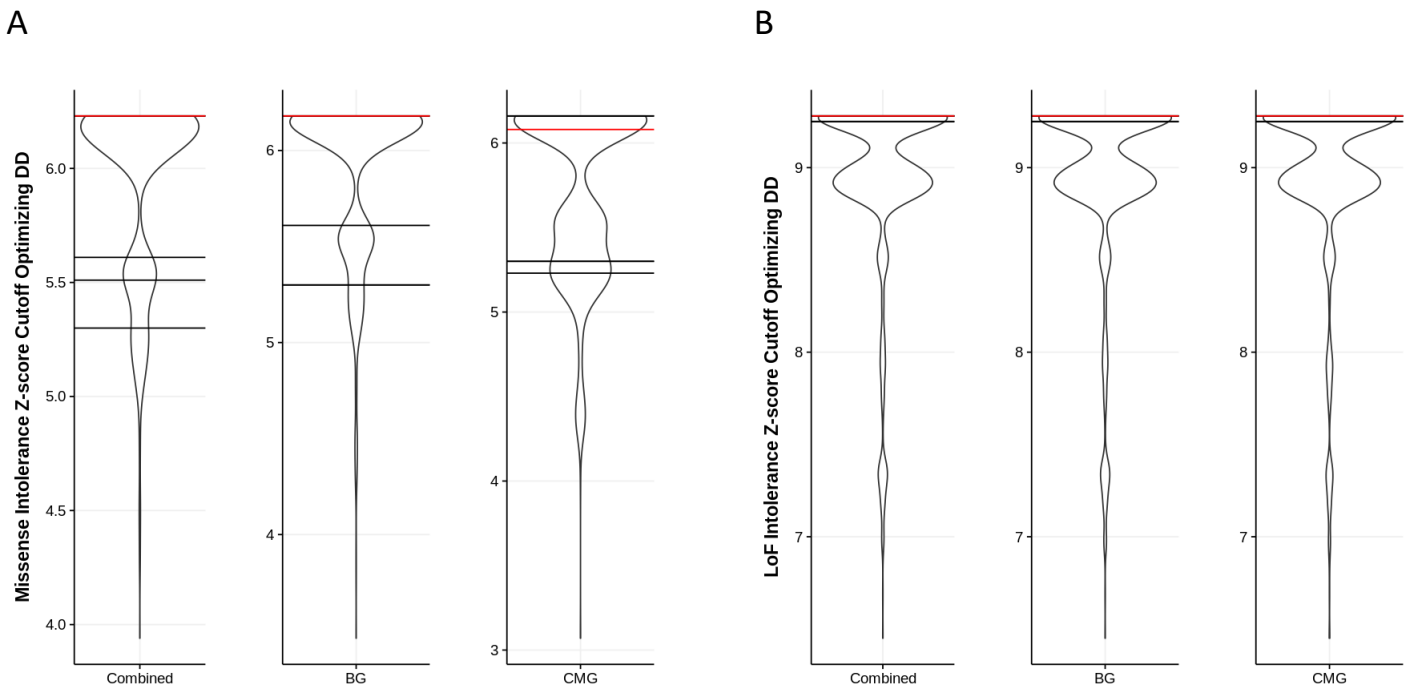


Figure S8 – Discovery-density optimizing filter cutoff variance across OMIM years and input datasets for missense variant parameter sweep query series. A) Missense Intolerance Z-score, B) LoF Intolerance Z-score. Cutoff values for real OMIM data are indicated with horizontal lines. Red horizontal lines indicate 2013-2018 DD-optimizing cutoff values, which for the combined data were used to define candidate novel disease gene lists. For the set of simulated (n=1,000) random downsamplings (.8x) of the 2013 OMIM annotation set (2013_{sim}), the distribution of cutoff values optimizing each 2013_{sim}-2013 DD is indicated by the violin plot width. Real cutoff values optimizing 2013-2018 DD (red) are remarkably stable across independent and aggregate data sets. Real cutoff values for shorter time intervals (black) are less stable across data sets, likely due to noise associated with a smaller change in OMIM annotation volume. For these two query series, many real and simulated DD-optimizing cutoff values are saturated at the minimal constrained candidate disease gene size of 20, evidenced by the appearance of a truncated upper distribution in the simulated data violin plots.

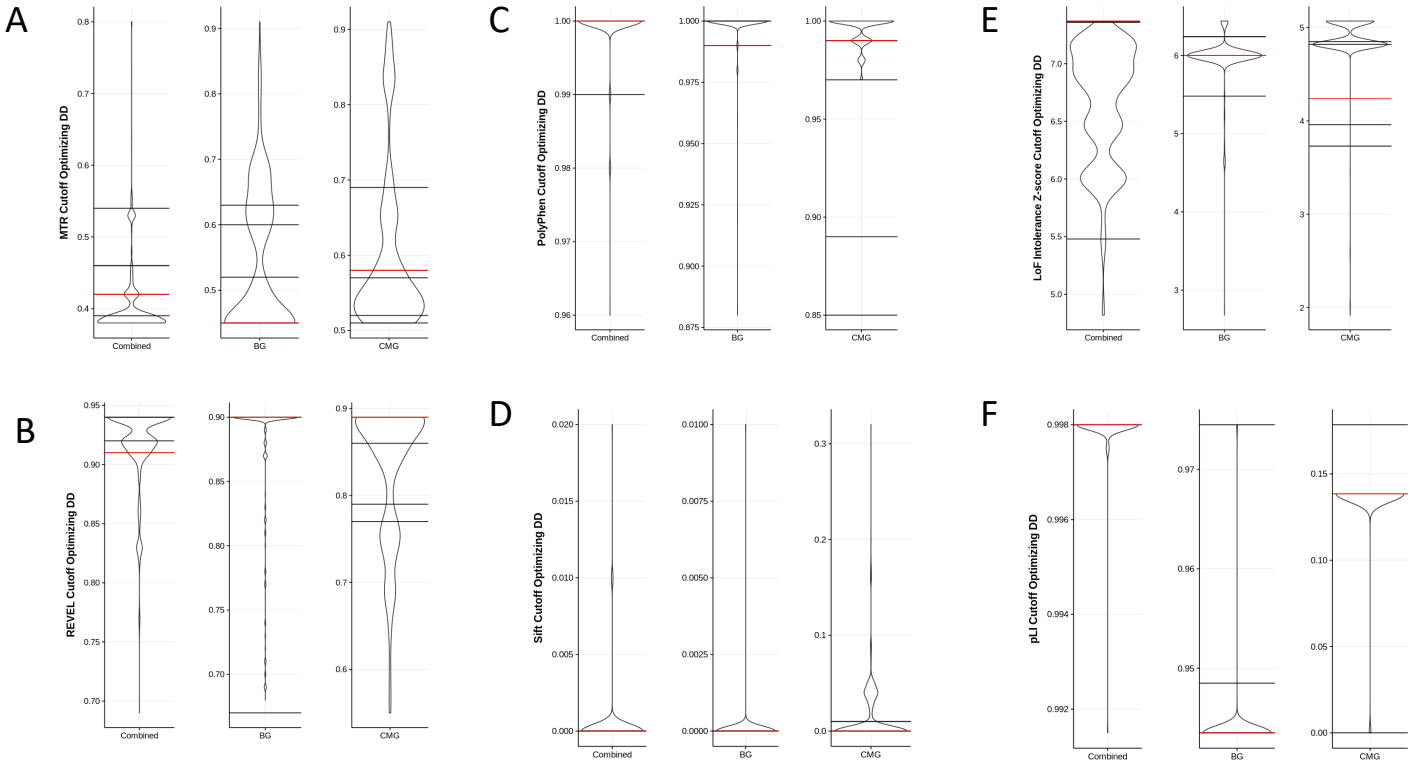


Figure S9 – Discovery-density optimizing filter cutoff variance across OMIM years and input datasets. A-D) Variant-level missense variant parameter sweeps: MTR (A), REVEL (B), PolyPhen (C), and SIFT (D). E-F) LoF variant parameter sweeps: LoF intolerance z-score (E) and pLI (F).

Supplemental Tables

	Avg. Total Time	# of Mappers	# of Reducers	Avg. Map Time	Avg. Shuffle Time	Avg. Merge Time	Avg. Reduce Time
TeraGen	0:17:56	180	NA	0:16:35	NA	NA	NA
TeraSort	0:15:09	7,650	96	0:00:11	0:03:28	0:00:03	0:02:12
TeraValidate	0:01:45	96	1	0:01:30	0:00:02	NA	NA

Table S1 – HARLEE performance benchmarks. Generating, sorting, and validating a terabyte of data with tools TeraGen, TeraSort, and TeraValidate, respectively, *without* encrypting data.

	Avg. Total Time	# of Mappers	# of Reducers	Avg. Map Time	Avg. Shuffle Time	Avg. Merge Time	Avg. Reduce Time
TeraGen	0:22:06	180	NA	0:17:37	NA	NA	NA
TeraSort	0:17:07	7,650	96	0:00:12	0:04:21	0:00:03	0:02:44
TeraValidate	0:01:41	96	1	0:01:15	0:00:02	NA	NA

Table S2 – HARLEE performance benchmarks with data encryption. Generating, sorting, and validating a terabyte of encrypted data with tools TeraGen, TeraSort, and TeraValidate, respectively.

Supplemental Methods

OMIM ratio outlier hypothesis is supported by OMIM ratio regression analysis

When we have finally catalogued all Mendelian disease-gene associations, we can anticipate that associations between OMIM disease enrichment for gene sets filtered against variable bioinformatic filtering parameters across a range of stringencies will reveal generalizable trends (albeit with a certain degree of noise). Until then, at any given point in time, OMIM represents an incomplete set of Mendelian disease-genes associations–an arbitrary subset of the eventual complete set of associations. Thus, as new disease-gene associations are reported to OMIM over time, OMIM annotations for a given parameter value should gradually transition from a stochastic subset into the true complete set of associations. Accordingly, as the curves of plots of OMIM ratio vs. variable bioinformatic filtering parameters gradually take shape, they should reveal an actual relationship between a given parameter and its differential enrichment for proportion of genes associated with Mendelian disorders across a range of filter values applied to a given cohort of affected individuals.

In line with this logic, together with our qualitative observations from the aforementioned preliminary missense intolerance z-score analysis, we hypothesized that, when binning gene lists based on filter parameter value (ie. missense intolerance z-score between 6.5 and 7.0), outlier lists with lower-than-expected disease gene enrichment will be ‘corrected’ as more disease genes are reported. To the extent this hypothesis holds true, we can accordingly prioritize gene lists for discovery efforts as those with lower-than-expected disease gene enrichment.

To test this hypothesis, we sorted all genes harboring recurrent ultra-rare variants across at least five samples in HARLEE by missense intolerance z-score, then grouped them into bins of equal size (40 genes). We then defined an expected degree of enrichment for Mendelian phenotypes by fitting a loess regression curve to the plot of OMIM disease gene fraction by gene-bin missense intolerance z-score rank (Figure S2). We show that the residuals of this plot correlate inversely with disease gene discovery reported in OMIM between 2013-2018. Furthermore, the residuals of loess regression models fitted on corresponding features from independent data sets (BG vs CMG) closely mirror each other, suggesting either robustness of this approach across Mendelian genocentric cohorts or a high degree of similarity of the BG and CMG data sets.

This evidence in support of the OMIM outlier hypothesis–that future discovery can be informed by smoothing a curve of current OMIM ratios and flagging outlier gene sets–supports a generalizable principle that future discovery can be enriched for by identifying gene sets with 'lower-than-expected' disease gene enrichment in OMIM at a given point in time. However, we have not pursued this approach for discovery due to the following rationale: 1) our models lack the degree of refinement and precision that we feel would be ideal for this type of prioritization; 2) regardless, the models tend to predict that genes with extreme filter parameter values are most enriched for discovery; and 3) it follows reason that the most likely of all candidate disease genes should be those with the most extreme constraint–given that we do observe loss-of-function or likely pathogenic missense variants in these genes (Figure S3).

Past discovery density correlates with future discovery density

The strategy of focusing gene discovery efforts on gene sets with a high past discovery density inherently assumes that, within the context of a fixed set of genes resulting from a constant set of query parameters, past discovery density correlates with future discovery density. Thus, in order to evaluate the general correlation between past and future discovery density, discovery density for all five possible nonoverlapping time intervals–based on the OMIM data available–was compared for each parameter sweep query series. Specifically, discovery density over the following combinations was compared: 2013-2014 vs. 2014-16; 2013-2014 vs. 2016-2018; 2013-2014 vs. 2014-2018; 2013-2016 vs. 2016-2018; and 2014-2016 vs 2016-2018. In all instances–except for three of the five REVEL series–a strong positive correlation between past and future discovery density was observed (Figures S4-S5).

We supplemented these data by repeatedly (n=1,000) removing 20% of the 2013 OMIM annotations, effectively 'simulating' a pre-2013 OMIM annotation set. For each query of HARLEE we calculated discovery density for each simulation–as well the mean discovery density across all simulations–as the number of genes with OMIM annotations in 2013, without annotations in a given pre-2013 simulation (Figures S6-S7). For each parameter sweep series, this simulated discovery density was then compared against actual 2013-2018 discovery density, creating a set of discovery density comparisons less prone to the noisy effects of small discovery volume.

For each simulation for a given parameter sweep, we also calculated the discovery density-optimizing filter cutoff value. For each score, the optimal cutoff parameters calculated from the

actual OMIM data points fell within the distribution of the 1,000 simulated optimal cutoff parameters. To test the sensitivity of the optimum discovery density metric to input data, we further calculated these optimum discovery metrics–for both real and simulated OMIM data–separately for independent BG and CMG sample sets. Here we observed more deviation across input data sets than across OMIM annotation period for a fixed input data set (Figures S8-S9).

Significant deviations from the real data correlation trendlines typically occur where discovery density is zero or low for one of the two time periods. We believe this is an artifact of small sample size–or number of discoveries for a given period of time for a given query–as the simulated data contains no instances of a zero-discovery density value. Furthermore, the REVEL simulated data (Figure S5B) does not show the non-linear artifacts seen in three of the real OMIM data REVEL comparisons (Figure S4B upper-right, lower-left, and lower-right panels). Indeed, the average simulated discovery density exhibits a strong positive correlation with actual 2013-2018 discovery density for each of the parameter sweeps tested.

Taken together, we believe these data comparing past vs. future discovery density–with both real and simulated data–validate the assumption that past discovery density correlates with future discovery density. We reasoned that a strategy selecting gene sets with high past discovery density should generally increase the probability that genes in the set without a current reported Mendelian disease association are in fact true, unreported Mendelian disease genes, thus accelerating the overall rate of future gene discovery.

Computer Infrastructure:

Hadoop/Cloudera cluster

HARLEE is a 10-node 280TB Cloudera Hadoop cluster. Each node has a 24-core CPU, 256GB memory and 10x4TB storage drives. Benchmarking and stress testing of the environment was performed with TeraGen, TeraSort, TestDFSIO, NNBench and MRBench. Variant data is stored using two complimentary Hadoop technologies–HBase and Parquet–which provide both rapid sample-level access to the raw variant data and a framework for complex SQL-like querying across the entire data set.

Variant/data ingestion

We have created an extract-transform-load (ETL) process that first ingests raw VCF files into HBase using the HBase Java API. The entire VCF file, including the header, is stored with the body of the VCF being converted to JavaScript Object Notation (JSON) key-value pairs; the INFO field tags are easily encoded in this way as are the sample and format felids. Columns that contain single values such as “filter” and “qual” are given the column name as the key. Unique row keys are created by concatenating a sample unique identifier with the chromosome, start position, end position, ref and alt alleles. Variants are stored in one HBase column and range features such as gVCF blocks, or structural variants are stored in a separate column for convenience when retrieving data. The ingest is secure, pushing VCF data directly into encrypted HBase tables. It is also rapid, a single process easily parsing over 20,000 variants per second, allowing us to store a typical clinical exome in 20-30 seconds; a whole genome gVCF with over 25M lines will be ingested in around 15 minutes. The ingested data is triplicated

within HDFS for redundancy but also compressed using SNAPPY compression–the Hbase footprint is therefore roughly equivalent to the uncompressed VCF. A whole genome gVCF takes up approximately 5Gb per sample, thus giving a theoretical maximum capacity of 56,000 whole genomes with current hardware.

Once the VCF data is encoded and encrypted within HBase, a map-reduce process is used to import the variant data into flat Parquet tables. A simple VCF-like schema is employed, using Apache HIVE user-defined functions (UDFs) to parse JSON data. Non-standard key-value pairs such as INFO field data remain JSON encoded, allowing us to store VCFs from multiple variant callers within the same schema.

Variant annotation

One advantage of the HARLEE is the ability to annotate 'on demand', or as data is queried, by joining annotations to variant data based on a common variant ID. This high degree of modularity is contrasted with a typical bioinformatic variant analysis pipeline, where specific annotation features are more or less hard-coded into an at-scale analysis effort. Modular annotation is ideal for a research environment, where questions to be asked or annotation features to be utilized are not known in advance. In order to streamline this modular annotation, we are creating an "annotation database" within the HGSC Data Lake. We are depositing both gene- and variant-level annotations into this database, including annotations from the following sources: ClinVar, dbSNP, ExAC, gnomAD, MTR (SOURCE), REVEL (SOURCE), 1,000 genomes project, dbNSFP, and other useful annotations downloaded from BioMart and the UCSC Genome Browser.

One difficulty with variant annotation is that many annotation features–such as SIFT and PolyPhen–are transcript-specific. Furthermore, there is no well-established or standardized approach for systematically mapping variants of interest to a single transcript. To resolve this issue for the time being, we have chosen to annotate all distinct HGSC variants with Ensembl Variant Effect Predictor (VEP), storing the output as the "variant" table within the annotation database. As part of its annotation script, VEP has the option of flagging a transcript of interest per variant per gene. Because no well-justified solution to transcript selection been published and extensively adopted to our knowledge, we have chosen to utilize this VEP option, with transcripts flagged by the default multi-tiered VEP logic: 1) canonical status of transcript; 2) APPRIS isoform annotation; 3) transcript support level; 4) biotype of transcript (protein_coding preferred); 5) CCDS status of transcript; 6) consequence rank–according to a table published on the VEP website; 7) translated, transcript or feature length (longer preferred). To maintain the flexibility of this resource, for a given variant we store all possible transcript annotations, flagging selected transcripts as opposed to removing all unselected transcripts. Furthermore, because we are dependent on VEP for transcript flagging, we have decided to obtain several additional annotations from VEP for convenience.

VEP command:

```
perl [VEP_path]/vep -i "$infile" -o "$outfile"
--dir_cache [VEP_path]/cache_files
--dir_plugins [VEP_path]/Plugins
--fork 8
--buffer_size 1000
--merged
--cache
--offline
--force_overwrite
```

```
--stats_text
--json
--assembly GRCh37
--everything
--total_length
--nearest gene
--hgvsg
--check_existing
--flag_pick_allele_gene
--fasta [VEP_path]/fasta_files/
1>process_1.vepRS.log
2>process_1.vepRS.err
```

Thus, following the parquet ingestion, the genotype data is joined with the annotated variant table–which is also mirrored in HBase and Parquet. Any novel variants are first annotated using VEP as described above, then all variants are subsequently queried against other useful annotations from our annotations database. The variant data is organized in this way so as to make obtaining all the relevant information about a given variant across multiple annotation sources a very fast and simple lookup, typically on the order of milliseconds.

Data access

Access to clinical data is strictly controlled. Data is encrypted at rest and in motion and tiered access is provided via Sentry and Kerberos authentication and authorization. Analysts are given access to subsets of the data based on requirements for specific projects. Within their sphere of data access, multiple options are available for analysts to interface with the data. Users with command line computing experience may prefer to utilize the Hadoop File System (HDFS) command line interface, where Apache Hive or Apache Impala shells may be launched. Alternately, the Hadoop User Experience (HUE) provides a user-friendly web app enabling visual

browsing of databases and tables, querying with Hive and Impala, workflow management, and a job browser detailing status of current and past jobs. However, while HUE is a more user-friendly interface than the command line, data querying in HUE still requires experience with SQL-like syntax. Thus, to facilitate data mining for users with limited or no SQL experience, further access is provided through Pentaho web portals which facilitate query building and data visualization without requiring any SQL, command line, or programmatic interface. Furthermore, Pentaho enables a framework for clinical reporting functions and complex analytics/visualization. Finally, data may also be accessed through the Java Database Connectivity (JDBC) API, which in turn allows for direct querying of the data from scripting languages such as R.

Security and compliance

We have implemented a multi-faceted and multi-layered security system to ensure the security and privacy of the data on Hadoop and to provide a compliance ready environment to comply with FISMA, HIPAA, Texas Medical Records Privacy Act and other industry regulations. The five pillars of our security implementation are 1) authentication, 2) authorization, 3) auditing, 4) data protection and 5) perimeter security.

Authentication is implemented by verifying the identity of the entity (user or service) trying to access the data with a strong Kerberos-enabled mechanism, specifically Microsoft Active Directory (AD) with Kerberos authentication. User and service principals are created and authenticated in Active Directory with passwords and keytab files, respectively, before they can

interact with the Hadoop cluster. With Kerberos enablement, users must first authenticate themselves to the Active Directory Kerberos Key Distribution Center (KDC) to obtain a valid Ticket-Granting-Ticket (TGT). The TGT is then used by Hadoop services to verify the user's identity. With Kerberos, a user is not only authenticated on the system they are logged into, but they are also authenticated to the network. Any subsequent interactions with other services that have been configured to allow Kerberos authentication for user access are also secured. (All Hadoop projects we are utilizing i.e. HDFS, MapReduce, HBase, Hive, HUE, Impala, Sentry etc. are all Kerberos enabled.) With this level of Kerberos enabled authentication, we have ensured that only legitimate AD users can authenticate to the system and have virtually eliminated any threat of user impersonation.

The next pillar is authorization. With authorization, we define the access or control an entity has over a given resource. To eliminate the overhead associated with managing access at the user level, we have created groups in Microsoft Active Directory and have assigned users to these groups. These groups are then mapped to roles in Apache Sentry. Sentry implements role-based access control (RBAC) with its roles being mapped to permissions. We have implemented granular permissions at the file-, directory-, database- and table-level in Apache Sentry and have mapped these permissions to the roles we have created. With this level of permission granularity, we can control and manage access to user groups efficiently thereby ensuring while public data is accessible to all groups, sensitive research/clinical data is accessible only to users/groups with access, and PHI data is accessible only to an authorized clinical group with HIPAA training. All the Hadoop projects other than HBase grant access to the users via the groups they are assigned to in Active Directory using Sentry. Sentry plugins are

added to the Hadoop projects during installation (supplementary figure 3); as the entity tries to access a given resource–e.g. Hive–the resource accesses the Sentry service via the plugin and verifies access. HBase utilizes independent Access Control Lists (ACL) for managing access; though access can be configured at the global (all databases), namespace, table or cell level, we have started with granting access at the namespace level to our user groups with the intention of implementing table, column and even cell level security as we start including sensitive data in HBase and need to partition user groups by table or columns. The combination of Cloudera and Hadoop technologies will allow us to manage access as granularly as possible thereby enabling us to house public, private, semi-private and sensitive data together while still providing access to multiple users and groups, both internal and external. We believe this will be greatly beneficial to both researchers and clinicians who hitherto have been only able to access portions of the data with difficulty through different means.

Auditing is a pillar that is critical for managing the compliance and data governance requirements of the Hadoop cluster. Without auditing, all of the other security pillars have limited effect because of a lack of visibility. We have implemented auditing for the Hadoop cluster using Cloudera Navigator and Cloudera Manager. With auditing, we are able to easily keep track of who is doing what on the cluster and when; this includes both positive events–actions that are successful and allowed–and negative events–actions that are unsuccessful and not allowed. In addition to centralized auditing–with Cloudera Navigator and Manager–we are also able to peruse detailed audit reports, giving us a quick and easy overview on who did what and when on the cluster. We are able to view the data lineage, which is helpful in identifying multiple key data attributes: the origin of the data; whether the data can be trusted for the

required analysis; and whether the data is being used by other users. With metadata tagging and indexing we are able to locate and track data easily and subsequently analyze user activity. Ultimately, with the auditing capabilities we have implemented on our Hadoop cluster, we are able to comply relatively easily with compliance and governance rules.

Encryption is a pillar, which generally speaking relates to data protection. With data protection–particularly encryption–becoming mandatory to comply with federal and state regulations, we have implemented both over-the-wire encryption and at-rest encryption. With over-the-wire encryption, we protect data while it is in transit over network channels and with at-rest encryption, we protect data when it is persisted to disk. The data in the Hadoop cluster, stored in HDFS, is protected end-to-end both during transfer and at rest via transparent data encryption (TDE). The performance overhead associated with encrypting/decrypting data is about 10-14%. We decided to encrypt all our data in HDFS as the need to maintain the security and privacy of the data overshadows the performance overhead. We implemented this by establishing encryption zones in HDFS and storing all our data in these zones. The directories and files in these encryption zones will be transparently encrypted upon write and transparently decrypted upon read (TDE). The encryption zone is associated with a key and each file/directory also has its own encrypted key and is managed using the Key Management Service (KMS) and a Key Store. The default KMS implementation combines the KMS and key store functions into a single service. As this implementation should not be used in a production environment with sensitive data, our KMS implementation uses the Cloudera Navigator Key Trustee as the key store. This separates the KMS and key store roles and allows them to be separated on different servers, which in turn provides better key protection. Additionally, we

have also utilized Cloudera Navigator to encrypt areas outside HDFS–i.e. log directories and database storage directories–to protect sensitive data in these locations.

The final security pillar is perimeter security. With perimeter security we provide guarded access to the Hadoop environment. A Cisco ASA firewall with an intrusion detection system tightly controls access to the HGSC network from outside the network. Additionally, the Hadoop cluster and associated ecosystem is designed to run on distinct network VLAN (Virtual LAN) that segregates the Hadoop cluster network traffic from other unencrypted network traffic. The Hadoop cluster and ecosystem reside on an independent, secure demilitarized zone with regulated external access. The security and compliance readiness of the Hadoop environment is ensured through a perimeter fence that houses the servers, network switches and infrastructure rack in a physically secure data center.

Supplemental Note

Task Force for Neonatal Genomics Consortium

Alexander Allori[2], Misha Angrist[3], Patricia Ashley[4], Margarita Bidegain[4], Brita Boyd[5], Eileen Chambers[6], Heidi Cope[1,7], C. Michael Cotten[4], Theresa Curington[1], Erica E. Davis[1], Sarah Ellestad[5], Kimberley Fisher[8], Amanda French[9], William Gallentine[10,11], Ronald Goldberg[4], Kevin Hill[12], Sujay Kansagra[10], Nicholas Katsanis[1], Sara Katsanis[3], Joanne Kurtzberg[13], Jeffrey Marcus[2], Marie McDonald[14], Mohammed Mikati[10], Stephen Miller[12], Amy Murtha[5], Yezmin Perilla[1], Carolyn Pizoli[10], Todd Purves[15], Sherry Ross[15,16], Azita Sadeghpour[1], Edward Smith[10], John Wiener[15]

[1]Center for Human Disease Modeling, Duke University Medical Center, Durham, NC USA
[2]Department of Surgery, Division of Plastic Maxillofacial and Oral Surgery, Duke University Medical Center, Durham, NC USA
[3]Science and Society, Duke University School of Medicine, Durham, NC USA
[4]Department of Pediatrics, Division of Neonatology, Duke University Medical Center, Durham, NC USA
[5]Department of Obstetrics and Gynecology, Division of Maternal-Fetal Medicine, Duke University Medical Center, Durham, NC USA
[6]Department of Pediatrics, Division of Pediatric Nephrology, Duke University Medical Center, Durham, NC USA
[7]Department of Medicine, Duke University Medical Center, Durham, NC, USA
[8]Neonatal Perinatal Research Unit, Duke University Medical Center, Durham, NC USA
[9]Fetal Diagnostic Center, Duke University Medical Center, Durham, NC USA
[10]Department of Pediatrics, Division of Pediatric Neurology, Duke University Medical Center, Durham, NC USA
[11]Present address: Department of Neurology, Division of Pediatric Neurology, Stanford University Lucile Packard Children's Hospital, Palo Alto, CA
[12]Department of Pediatrics, Division of Pediatric Cardiology, Duke University Medical Center, Durham, NC USA
[13]Department of Pediatrics, Division of Pediatric Blood and Marrow Transplantation, Duke University Medical Center, Durham, NC USA
[14]Department of Pediatrics, Division of Medical Genetics, Duke University Medical Center, Durham, NC USA
[15]Department of Surgery, Division of Pediatric Urology, Duke University Medical Center, Durham, NC USA
[16]Present address: Department of Urology, University of North Carolina, Chapel Hill, NC USA