ARTICLE Genetics inMedicine

Trajectory of exonic variant discovery in a large clinical population: implications for variant curation

Uyenlinh L. Mirshahi, PhD[1], Jonathan Z. Luo, BS[1], Kandamurugu Manickam, MD, MPH[1], Amr H. Wardeh, BS[1], Tooraj Mirshahi, PhD[1], Michael F. Murray, MD[1] and David J. Carey, PhD[1]

Purpose: Precision health initiatives and reduced sequencing costs are driving large-scale human genome analyses. Genetic variant curation is a bottleneck in clinical applications. The burden of variant curation can be high for newly discovered variants because they are less likely to have undergone previous clinical annotation; the rate of discovery of genetic variants in large clinical populations has not been empirically determined.

Methods: We determined the rate of accrual of unique sequence variants in 90,000 exome sequences. Separate analyses were done for 17,267 autosomal genes and a subset of 74 actionable genes; the effect of relatedness in the cohort was also determined.

Results: Variant discovery showed a nonlinear growth pattern. The rate of unique variant accrual decreased as the database size increased; by 90,000 exomes 97% of all projected coding and splicing variants had been observed. Variants in 74 actionable genes showed a similar pattern. Family relatedness slightly reduced the rate of discovery of unique variants.

Conclusion: The heaviest burden of interpretation for genetic variants occurs early and diminishes as the database size increases. Our data provide a framework for scaling pathogenic genetic variant discovery and curation, a critical element of patient care in the era of precision health.

Genetics in Medicine (2019) 21:1417–1424; https://doi.org/10.1038/s41436-018-0353-5

Keywords: exome sequencing; variant curation; genomic screening; secondary findings; sequence scaling

INTRODUCTION

Any large-scale project that seeks to deliver genomic findings to participants faces bottlenecks associated with variant curation for each newly encountered variant in a gene of interest, where variant curation and interpretation can be rate limiting steps. As described by the FDA draft guidance, variant curation involves generating and maintaining an updated database of variants, and variant interpretation entails assigning a pathogenicity status to a genetic variant with respect to the disease phenotype.[1] Both tasks are interdependent. A false positive assignment of a variant could lead to unneeded and costly screening or procedures, and a false negative assignment could preclude the patient from timely preventive therapy. This impacts not only clinical care but also the economically strained health-care system as a whole. Clinical as well as research labs are increasingly focusing on the time-consuming task of reviewing available data and literature to provide evidence-based classification of variants. The American College of Medical Genetics and Genomics (ACMG), the Association for Molecular Pathology (AMP), and the College of American Pathologists (CAP) jointly published guidelines to classify and name variants using evidence-based criteria.[2,3] Despite routine application of these guidelines in clinical and research settings, discordance in classification remains.[4,5] Tools to streamline variant classification have been developed.[6,7] The FDA recently announced its intention to provide guidelines on and perhaps standardize how variants are generated, curated, and interpreted, underscoring these issues as an emerging public health concern.[1]

The Geisinger Health System (GHS) Genomic Screening and Counseling Initiative (formerly the GenomeFIRST project) aims to identify individuals with clinically actionable genetic variants in 76 genes (Geisinger 76, or G76) that are linked to 27 medical conditions, and to return these results to the participants and their medical providers.[8,9] The starting point for this project is a large database of exome sequences from Geisinger patients who consented to participate in the MyCode Community Health Initiative.[10]

In light of the challenges of clinical variant classification, we hypothesized that the time and effort needed to curate and interpret variants is greatest at the beginning of such a project, and that this effort will be reduced as the cohort size increases. This hypothesis assumes that the rate of discovery of novel variants will decline as the exome sequence database for a population grows. We sought to empirically test this

[1]Geisinger Clinic, Geisinger Health System, Danville, PA, USA. Correspondence: Uyenlinh L. Mirshahi (umirshahi1@geisinger.edu)

Submitted 27 August 2018; accepted: 25 October 2018
Published online: 19 November 2018

hypothesis in a set of 90,000 exome sequences of participants in the Geisinger MyCode project, by determining rates of unique variant discovery in 74 autosomal genes of the G76 clinically actionable gene set. We provide an example of the clinical utility of this concept by examining the rate of growth of *BRCA1* and *BRCA2* variants, where the frequency of new variants requiring pathogenicity assignment is calculated.

MATERIALS AND METHODS

Cohort description

The data for this study came from the DiscovEHR collaboration between Geisinger and the Regeneron Genetics Center. As part of this collaboration, DNA samples from Geisinger patients were used for exome sequencing. The data analyzed consisted of exome variants from the first 90,000 adult participants. The cohort characteristics have been described previously.[9,10] This study was reviewed and does not involve "human subjects" as defined in 45CFR46.102(f); and therefore, is not subject to oversight by the Geisinger Institutional Review Board. The use of genetics data from the MyCode Community Initiatives was approved by the Geisinger Institutional Review Board.

Exome sequencing

Exome sequencing for the DiscovEHR collaboration has been previously described in detail.[9,11] Briefly, DNA samples were exome-sequenced using NimbleGen probe target-capture (SeqCap VCRome, 61,019 exomes) and xGen capture (Integrated DNA Technologies, 31,393 exomes), followed by sequencing on an Illumina v4 HiSeq 2500 to a coverage depth of greater than 20× in 90% of target regions. Genomic variant call format (gVCF) files created by VCRome capture or xGen capture were joint-called separately in groups of 200 individual gVCFs to a prepared pseudosample containing all single-nucleotide variant (SNV) and indel sites to create pVCF files. Two hundred individual pVCFs from VCRome or xGen were combined prior to combining the final union of the VCRome pVCF and xGen pVCF to create a union pVCF. The union pVCF sequence reads were aligned to GRChr38. Variant quality controls included filtering variants for quality by depth (QD) > 5.0 and DP > 10 (indels) or QD > 3 and DP > 10 (missense). Project-level quality controls of the combined data set included genotype and sample call rates >98% and Hardy–Weinberg equilibrium $p > 1\text{e-}06$ (PLINKv1.9) (ref. [12]). Of the resulting 92,297 samples, 90,000 were randomly selected for this study. For *BRCA1* and *BRCA2* variants, samples with alternate allelic balance >15% (missense) or >20% (indels) and at least five or more alternate reads were selected.

Variant Annotation

Sequence variants were annotated to coding DNA and functional proteins using the National Center for Biotechnology Information (NCBI) RefSeq Gene definitions, selecting for the transcript with the longest coding sequence among the transcripts with a Locus Reference Genome (LRG) annotation, and excluding transcripts without annotated start and stop codons (SNP & Variation Suite, Golden Helix, Bozeman, MT).[2,13] We define coding variants as nonsynonymous (missense, insertions, deletions) and synonymous variants; splicing variants as "canonical" splicing variants that are defined by the GT and AG intronic nucleotides 2 base-pairs on the intronic side of a splice junction; putative loss-of-function variants (pLOFs) as canonical splice donor and splice acceptor variants, initiation and stop loss variants, stop-gained variants, and variants causing a shift in the reading frame (frameshift); and missense variants as SNVs with one or more base change that alters amino acid sequence, including in-frame deletions, and in-frame insertions.

To simplify variant annotation, we examined 17,267 autosomal genes in chromosomal regions of monogenic loci and excluded genes that overlap in polygenic loci.[14] For actionable genes, we selected the 74 autosomal genes of the G76 actionable gene list (Supplemental Table S1). The G76 clinically actionable genes include 56 genes for 25 medical conditions recommended by the ACMG for return of secondary findings by the criteria that deleterious variants in these genes would result in highly penetrant disease phenotypes that could be improved through medical interventions.[8] The Geisinger project team reviewed evidence available after the original 56-gene list was developed and added additional genes for the same medical conditions, as well as *ACVRL1* and *ENG* for hereditary hemorrhagic telangiectasia and *OCT* for ornithine transcarbamylase deficiency.[9]

Variants were mapped to ClinVar annotation track (SNP & Variation Suite, last updated in August 2017) and selected for those with germline minor allele origin. Pathogenicity of variants were defined as pathogenic or likely pathogenic variants with review status of two or more stars, denoting multiple submitters with the consensus classification.

Simulation for asymptote of variant accumulation

We modeled the number of unique variants observed in the database with the R "car" package and the function nls for nonlinear least square fit typically used to model population growth to estimate the asymptote. The asymptote is predicted to indicate the number of variants at which no new unique variants will be observed. To extrapolate the numbers of variants in cohort sizes beyond 60,000 or 90,000 exomes, we regressed the number of variants associated with incremental increases of 10,000 exomes, ranging from 100,000 to 360,000 for all samples, and 70,000 to 290,000 for unrelated samples.

Graphs

All data were plotted using GraphPad Prism (La Jolla, CA), except for the variant simulation graphs, which was generated in R using ggplot2 package.[15] Statistical analyses were performed using GraphPad Prism.

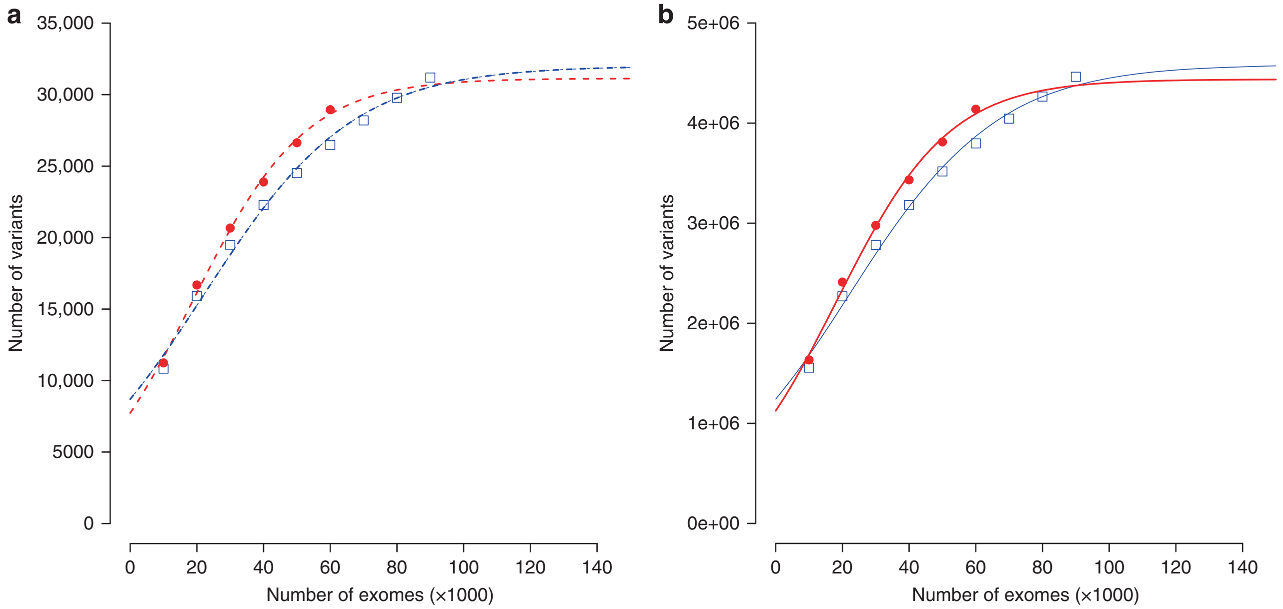

Fig. 1 Accrual of coding and splicing variants in unrelated and related individuals from the DiscovEHR study in 74 actionable genes and all genes. Samples from all 90,000 individuals or from 60,000 unrelated individuals were divided into 10,000 exome increments and accrual of variants per 10,000 exomes is shown. Curves were fit to a nonlinear model to estimate the asymptote. **a** Accrual of variants in 74 genes are shown for all individuals (blue open squares and blue dashed line) and for unrelated individuals (red filled circles and red dashed line). **b** Accrual in 17,267 genes are shown for all individuals (blue open squares and blue solid line) and for unrelated individuals (red filled circles and red solid line).

RESULTS

Coding and splicing variants in 90,000 exomes

We examined variants in 74 clinically actionable autosomal genes selected for return of pathogenic findings. Variants were identified in 90,000 exome sequences available currently from the DiscovEHR collaboration study.[9] We focused on variants in protein coding regions and canonical splice site variants because these are most likely to have functional consequences and be classified as likely pathogenic or pathogenic. The genes, transcripts used for the analysis, and associated clinical conditions are listed in Table S1. Of the 31,194 unique coding regions and splice site variants identified in these 74 genes, 62% were single-nucleotide missense variants, 33% were synonymous variants, and 4.7% were pLOF variants. The majority of coding and splicing variants are low frequency: 97% had minor allele frequencies (MAF) < 1%, and 94% had a MAF < 0.1% (Figure S1). Based on this finding, we used all coding and splicing variants without MAF cut-offs in subsequent analyses.

The number of new variants observed decreases as the sample size increases

The 90,000 exome sequences were divided randomly into nine groups of 10,000, and the number of coding region and splice site variants observed in the 74 genes was determined as a function of database size. As shown in Fig. **1a** (blue open squares), the incremental increase in the number of unique variants decreased with increasing database size and approached a plateau. With each increase of 10,000 exomes, there were successive decrements in the number of new variants observed (see Table S2). Approximately 35% (10,833) of the variants in 90,000 exomes were observed in the first 10,000 exomes, but only 757 (2.4%) of the total variants were observed first in the increment from 80,000 to 90,000 exomes (Table S2). A nonlinear model was used to extrapolate the results to larger database sizes (Fig. **1a**, blue line). This predicted an asymptote of 32,009 ± 946 (SEM) unique variants in a database of 250,000 exomes (r^2 0.996). Beyond 200,000 exomes, the rate of accrual of new coding or splicing variants was predicted to be low and approached zero (Table S2). This analysis suggests that at the current DiscovEHR database size of 90,000 exome sequences, approximately 95% of unique coding and splicing variants in the population have been observed.

As the MyCode cohort consists of a majority of individuals with at least one first- or second-degree relative as determined by genome-wide identity by descent (IBD) estimates,[11] we sought to determine the effects of familial relatedness on the rate of accrual of coding and splicing variants in the 74 genes. Removal of all but 1 individual in the related groups resulted in ~62,200 individuals with no familial relationship up to third degree (PI-HAT > 0.1875), of which 60,000 were randomly selected for analysis. Figure **1a** (red closed symbols) shows that the number of unique coding and splicing variants per 10,000-exome increment was slightly higher when only unrelated individuals were included. Simulation predicted the same number of total variants, with an asymptote of 31,137 ± 1143 (SEM) unique variants in a database of 210,000 exomes (r^2 0.998, Fig. **1a** red line), with 93% of projected variants observed in the first 60,000 exomes (Table S3).

To determine if these observations were unique to genes associated with monogenic disorders, we carried out a similar analysis for the set of 17,267 nonoverlapping autosomal genes (see Materials and Methods). Figure **1b** shows the variant accrual rates for all 90,000 individuals (blue) and for 60,000 unrelated individuals (red). While the numbers of variants were higher, the shapes of the curves (when expressed as percent of predicted asymptotes) are indistinguishable from those determined for the 74 actionable genes (Figure S2, Table **1**, Tables S4, S5), indicating similar patterns of variant accrual.

We also analyzed the trajectory of discovery of variants in functional subclasses (missense, synonymous, and pLOF) in the 74 actionable genes. As shown in Figure S3 the increase in total number of variants in each functional subclass increased nonlinearly as the size database increased, with a progressive decrease in the number of new variants per incremental increase in database size. The rate of discovery of pLOF variants was slightly lower than synonymous or missense variants. Twenty-eight percent of pLOF variants observed in the 90,000 exome database were observed in the first 10,000 exomes, compared with 38% and 34% for synonymous and missense variants.

Singletons become a decreasing proportion of variants

We also determined the number of singletons (variants observed in only one individual in the data set) in the 74 actionable genes and their proportion of total variants as the number of exome sequences increases. Up to a sample size of 30,000 exomes singletons comprised ≥50% of coding and splicing variants (Fig. **2a**). This decreased to 38% as the sample size increased to 90,000 exomes: as the database grew in size some variants that were singletons were identified in additional carriers and the number of newly discovered singletons decreased (Fig. **2b**). The rate of discovery of singletons was higher if related individuals were excluded (Fig. **2b**). The decrease in the rate of discovery of new singletons was observed for all functional classes of variants (Fig. **2c**). Putative LOF variants were more likely than other variant types to remain as singletons (Fig. **2d**). Singletons accounted for 49% of pLOF variants at a database size of 90,000 exomes.

The number of new variants classified in ClinVar decreased as the size of the cohort increased

We also analyzed variants that had been assessed in ClinVar as pathogenic/likely pathogenic (P/LP), benign/likely benign (B/LB), or variants of unknown significance (VUS). Of 31,194 coding region and splice site variants in the 74 actionable genes in 90,000 exomes, 10,440 (33%) had been classified in ClinVar, with varying degrees of evidence support; 1162 variants (3.7%) reached at least 2-star review status, which represents variants classified similarly by multiple submitters. Of these 1162 variants, 51% were classified as B/LB, 8% were P/LP, and 41% were VUS (Table S6). Not surprisingly, the majority of P/LP variants were pLOFs, the majority of B/LB were synonymous variants, and the majority of VUS were missense variants.

We determined the number of missense, pLOF, and synonymous variants in each assertion group as the database size increased. Figure **3** shows that the majority of missense and synonymous variants classified as B/LB were observed in even a relatively small sample size; at 10,000 exomes, approximately 80% of B/LB variants present in 90,000 exomes were observed. This is mainly due to the fact that variants of this type have a higher frequency within the population. Similarly, approximately half of all VUS variants were observed in the first 10,000 exomes. In contrast, variants classified as P/LP continued to accrue, albeit at a progressively slower rate, as the database size grew; at 10,000 exomes, approximately 20% of the variants present in 90,000 exomes were observed; at 30,000 exomes, more than half were observed.

Accrual of variants in *BRCA1* and *BRCA2*

We applied a similar strategy to estimate the expected burden of variant curation for variants in the hereditary cancer genes *BRCA1* and *BRCA2*. We determined the numbers of variant carriers, new variants, singletons, and ClinVar 2-star variants in sequential batches of 50,000 and 40,000 exomes (Table **2**). As expected, fewer new variants and fewer singletons were observed in the second group of exomes. To estimate the number of variants that would require curation, we removed variants previously characterized in ClinVar with 2-star or greater level of assertion and variants observed in a single carrier. This resulted in 556 variants requiring curation in the first 50,000 exomes and 65 in the second 40,000 exomes, consistent with a reduced burden of variant classification as the database grows.

DISCUSSION

The purpose of this study was to provide empirical data on the rate of discovery of unique variants within a single health system cohort as a function of the size of the database of available exome sequences. There were two primary motivations for generating this data. As described previously,[9] this exome sequence database, which can be linked to longitudinal electronic health record data, provides a powerful resource for genomic discovery. Information of the type presented here provides a framework for determining the number of exome sequences needed for such research. In addition, and as has been described,[16] we have developed a program to return "clinically actionable" variant findings to patients and their medical providers as a means to provide information on significant genetic risks for potentially life-threatening conditions. A major bottleneck for such a program is curation of novel variants as they are observed. Our hypothesis was that the rate of new variant accrual would decline as the size of the sequence database increased; if this could be demonstrated, it would suggest that the time and effort required for clinical interpretation of variants would decrease as the database expands. The data presented support this hypothesis: Our results show that accumulation of coding region and splice site variants follows a growth model in which the rate of

Table 1 Summary statistics of variant accumulation analyses

	# Exomes	Asymptote (SD)	# Exomes to reach asymptote	% Asymptote at current database[a]	Slope (SD)	r^2
All genes, all individuals	90,000	4,587,310 (126,700)	360,000	97	22.4 (2)	0.997
All genes, unrelated individuals	60,000	4,437,535 (158,500)	290,000	93	16.9 (2)	0.998
74 genes, all individuals	90,000	32,010 (946)	260,000	95	22.4 (2)	0.996
74 genes, unrelated individuals	60,000	31,137 (1143)	210,000	93	16.9 (2)	0.998

Accrual of coding and splicing variants in 17,267 genes (all genes) or in 74 actionable genes in all individuals and in unrelated individuals were fitted to a nonlinear least square fit model with simulation to obtain the trajectory of variant growth to asymptote (See Fig. 1). The summary statistics of the curve fitting are detailed below. At the current DiscovEHR database size, 93–97% of projected variants have been observed (see Supplemental Tables S2–S5).
SD, standard deviation.
[a] % Asymptote at current database: % of variants attained at the current # exomes in the DiscovEHR cohort from Supplemental Tables S2–S5.

accumulation of previously unobserved variants decreases as the cohort size increases. Approximately one-third of variants were discovered in the first 10,000 exomes, and approximately 95% of variants are discovered in the first 90,000 exomes. While this might be intuitively obvious, to our knowledge these are the first empirical data to describe the asymptotic behavior of variant accumulation. These findings suggest that the heaviest burden of clinical variant interpretation lies at the beginning, i.e., that much of this effort is “front-loaded,” mainly due to repetition of variants, progressively reducing the number of new variants to curate and interpret.

We observed that the accumulation of variants in a subset of clinically actionable genes was similar to that of the entire exome (Fig. 1). Although these genes are linked to autosomal dominant genetic disorders, most of these conditions occur in later stages of life and so may not be subject to strong purifying selection. In addition, it has been shown that individuals who harbor pathogenic variants for *SCN5A* and *KCNH2* presented with similar risks for arrhythmia and other cardiac disease conditions as those who lack the variants,[17,18] underscoring the importance of penetrance of actionable genes.

It should be noted that the specific rates of variant discovery reported here are applicable to the DiscovEHR cohort from which the exome sequence data was generated. Individuals in the DiscovEHR cohort are all patients of Geisinger, a large integrated health system, who live in north central and northeastern Pennsylvania and consented to participate in the MyCode Community Health Initiative. The DiscovEHR cohort, therefore, represents an unselected clinical population that reflects the demographic characteristics of the region. As this population is predominantly (98%) of White European descent with approximately 56% of the participants related to at least one other participant in the cohort by first- or second-degree relationships,[11] the number of unique variants in this database and the frequencies of specific alleles could be different from other cohorts with different racial, ethnic, or familial relationship backgrounds. It will be important to compare these results with other populations. Our findings did show that family relatedness in the population affected the rate of unique variant accrual (Fig. 1, Table 1). This suggests that less related populations may require a smaller cohort to reach an asymptote. Importantly, for other health systems that are considering large-scale clinical sequencing projects, the results presented here provide a framework for scaling the processes of variant discovery and variant curation.

Missense SNVs are typically the hardest to classify because their functional consequences are difficult to predict. Bioinformatic tools that predict “deleteriousness” based on sequence conservation, biochemical considerations, etc. can help, but these alone are usually not sufficient for high-confidence assertions of pathogenicity. This complicates the task of variant classification, because missense variants are the most common type of variant observed in exome sequencing studies. These factors combine to make missense variants the majority of VUS. Our findings highlight the need for comprehensive functional testing of VUS in clinically actionable genes because assessing their pathogenicity using traditional clinical and pedigree-based studies will be difficult, especially for low-frequency variants. Approaches similar to those recently reported by Findlay and colleagues will become instrumental to functional characterization of such rare variants.[19] The encouraging news from our analysis is that the rate of discovery of new missense variants declines as the size of the database increases.

The rate of discovery of unique LOF variants was slightly lower than synonymous or missense variants, and pLOF variants were more likely to remain as singletons as the database size increased, probably reflecting greater negative selection for this class of variants. One of the advantages of exome sequencing of a large population is the ability to uncover pLOFs and other extremely rare variants, including those associated with rare monogenic disorders. In addition to identifying variants in *BRCA1* and *BRCA2*, used as examples here, we have used this exome sequence database to identify variants associated with *DICER1* syndrome and maturity onset diabetes of the young, rare autosomal dominant disorders (U. Mirshahi et al., unpublished data) and cystic fibrosis, a rare recessive disorder (Sugunaraj et al., unpublished data).

Singletons are often interpreted as VUS because they are more likely to be private variants with little or no previous experimental or clinical data for curation. This underscores the need to share discoveries with others in the genomics community, such as curation consortia, to build a database for

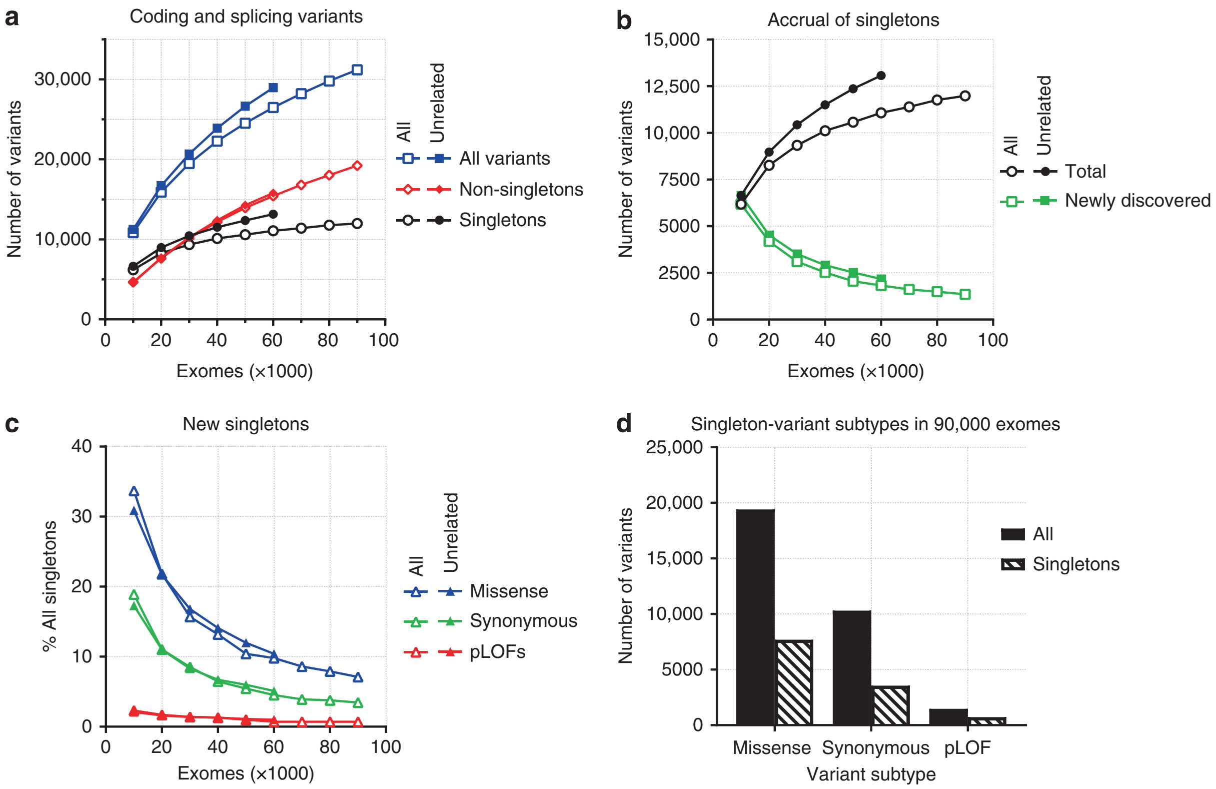

Fig. 2 Singletons make up a large proportion of the coding and splicing region variants of the 74 autosomal actionable genes. a Accrual of all variants (blue), singletons (black), and all variants excluding singletons (red) per 10,000 exome increments are plotted for all 90,000 exomes (open symbols) and for unrelated exomes (closed symbols) as shown. **b** In singletons, the accrual of newly discovered variants is reduced as the database increased for all exomes (open symbols) and for unrelated exomes (closed symbols). **c** Accrual of new singletons observed in each variant subtype in all 90,000 exomes and 60,000 unrelated exomes plotted as % All singletons in each database. **d** Comparison of singletons and all variants in each variant subtype at 90,000 exomes. *pLOF* putative loss of function.

these rare variants. Large databases from comprehensive health systems where genetic information is linked to longitudinal electronic health record (EHR) data (with laboratory and imaging data, diagnoses, medication data, procedure codes, etc.) will assist in determining the pathogenicity of these rare variants. Because the MyCode cohort consists of individuals seeking health care through a large integrated system, more than half of the participants with exome sequence data have one or more first-degree relatives who have also been sequenced through this program. It is possible to calculate degrees of relatedness and infer pedigrees from the exome sequence data.[11] We are exploring the use of this information for *in silico* familial studies that could shed light on the clinical consequences of specific variants, including singletons. Furthermore, an important element of the MyCode Community Health Initiative is the possibility to contact participants to invite them to engage in follow-up studies to provide additional health information, collect family history data, or invite family members to participate in genetic screening.

The technical specification for the exome sequence data used for this study (minimum of 20× coverage for >90% of the target regions) is lower than the read depths for some clinical gene panels. This is a potential limitation for population-based screening efforts where the goal is to generate exome sequence data on a large number of individuals. However, when the quality filters described in the Methods section are applied, the false positive rate for SNVs and indels is very low. It should be noted that any variant result that is returned to a participant is confirmed by orthogonal confirmation. We believe that rigorous quality controls for sequencing, alignment, and variant calling must be applied prior to clinical implementation.[20] It should be noted that in this study we focused on SNV and indel variants. The ability to accurately call copy-number variants and other structural variants from exome sequence data is less reliable, and their prevalence and causal roles in disease phenotypes are less well understood because most are discovered through patient disease cohorts.[21,22] The number of these variants and the trajectory of their accrual are important questions that remain to be answered, but this will require improvements in the technologies used to identify copy-number and other structural variants from exome sequence data.

While our results suggest that the burden of variant interpretation will be reduced as sequence databases expand, they do not mitigate the need to keep improving and standardizing methods used to curate variants and assess their

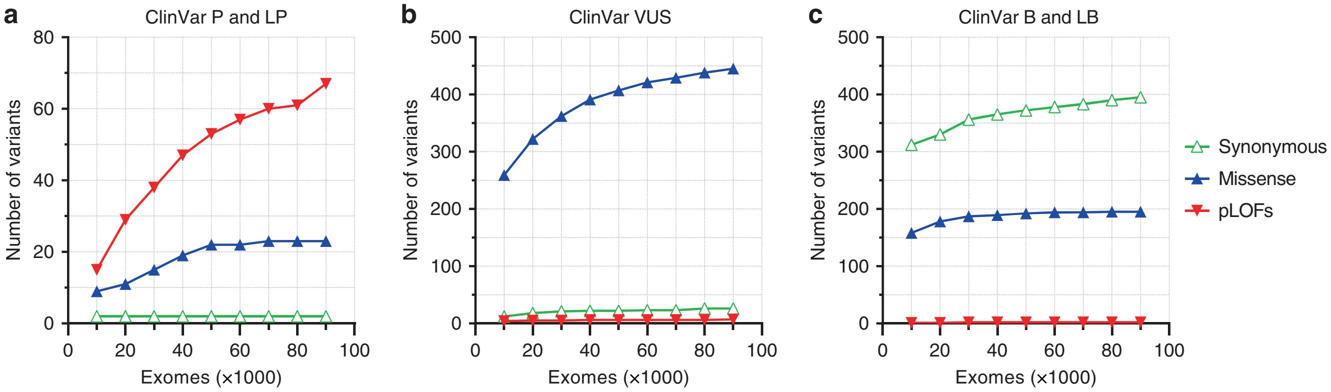

Fig. 3 Accrual of variants classified by ClinVar as a function of database size. DiscovEHR variants previously classified as pathogenic/likely pathogenic, variants of unknown significance, benign/likely benign with consensus review status of multiple submitters in ClinVar were separated by functional class. The number of synonymous variants (green), missense (blue), and pLOFs (red) per 10,000 exome increments are plotted for (**a**) pathogenic/likely pathogenic, (**b**) variants of unknown significance, and (**c**) benign/likely benign variants. Curves are drawn by connecting scatter points on the plots. *B* benign, *LB* likely benign, *LP* likely pathogenic, *P* pathogenic, *pLOF* putative loss of function, *VUS* variant of unknown significance.

Table 2 The burden of new variant curation for *BRCA1* and *BRAC2* is reduced as the database grows

Data sets (exomes)	Variants[a]	Carriers	New variants[b]	New singletons	Variants in ClinVar[c] with >1 carriers	Variants requiring curation
First 50,000	1263	10,564	1263	618	89	556
Last 40,000	1167	7883	473	397	11	65
Total 90,00	1736	15,742				

The data were divided into two sets of 50,000 and 40,000 exomes, and the numbers of carriers and coding and splicing variants in *BRCA1* and *BRCA2* were determined. The number of new variants requiring curation was determined after removing variants previously classified by multiple submitters.

[a] The number of unique variants in the two data sets includes variants found in both.

[b] Includes singletons.

[c] By multiple submitters with consensus (≥2 stars).

pathogenicity. While some progress in this area has been made, a lack of consistency and standardization in these areas has been well documented and must be addressed.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (https://doi.org/10.1038/s41436-018-0353-5) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

The authors thank the participants of the MyCode Community Health Initiative for use of their genomic and electronic health information; without them, this study would not be possible. The DiscovEHR study was funded in part by the Regeneron Genetics Center.

DISCLOSURE

M.F.M. reports having received research grants from Regeneron, and personal fees from InVitae, both outside of the submitted work. The other authors declare no conflicts of interest.

REFERENCES

1. US Food and Drug Administration. Use of public human genetic variant databases to support clinical validity for next generation sequencing (NGS)-based in vitro diagnostics: draft guidance for stakeholders and food and drug administration staff. Center for Biologics Evaluation and Research (CBER); Office of Communication, Outreach, and Development (OCOD), Silver Spring, MD, 2016.
2. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405–424.
3. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med. 2017;19:249–255.
4. Amendola LM, Jarvik GP, Leo MC, et al. Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. Am J Hum Genet. 2016;98:1067–1076.
5. Harrison SM, Dolinsky JS, Knight Johnson AE, et al. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet Med. 2017;19:1096–1104.
6. Kleinberger J, Maloney KA, Pollin TI, Jeng LJ. An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants. Genet Med. 2016;18:1165.
7. Patel RY, Shah N, Jackson AR, et al. ClinGen pathogenicity calculator: a configurable system for assessing pathogenicity of genetic variants. Genome Med. 2017;9:3–11.
8. Dorschner MO, Amendola LM, Turner EH, et al. Actionable, pathogenic incidental findings in 1,000 participants' exomes. Am J Hum Genet. 2013;93:631–640.
9. Dewey FE, Murray MF, Overton JD, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016;354:1549–1561.
10. Carey DJ, Fetterolf SN, Davis FD. et al. The Geisinger MyCode Community Health Initiative: an electronic health record-linked biobank for precision medicine research. Genet Med. 2016;18:906–913.

11. Staples J, Maxwell EK, Gosalia N, et al. Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. Am J Hum Genet. 2018;102:874–889.
12. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–575.
13. Dalgleish R, Flicek P, Cunningham F, et al. Locus reference genomic sequences: an improved basis for describing human DNA variants. Genome Med. 2010;2:24.
14. Thompson JN Jr, Thoday JM. A definition and standard nomenclature for "polygenic loci". Heredity (Edinb). 1974;33:430–437.
15. Wickham H Ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2009. http://ggplot2.org.
16. Abul-Husn NS, Manickam K, Jones LK, et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science. 2016;354:1550–1558.
17. Van Driest SL, Wells QS, Stallings S, et al. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. JAMA. 2016;315:47–57.
18. Smith JL, Tester DJ, Hall AR, et al. Functional invalidation of putative sudden infant death syndrome-associated variants in the KCNH2-encoded Kv11.1 channel. Circ Arrhythm Electrophysiol. 2018;11:e005859.
19. Findlay GM, Daza RM, Martin B, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018 Sept 12; doi: https://doi.org/10.1038/s41586-018-0461-z [Epub ahead of print].
20. Packer JS, Maxwell EK, O'Dushlaine C, et al. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. Bioinformatics. 2016;32:133–135.
21. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16:172–183.
22. Hurles ME, Dermitzakis ET, Tyler-Smith C. The functional impact of structural variation in humans. Trends Genet. 2008;24:238–245.