Analyzing and Reanalyzing the Genome: Findings from the MedSeq Project

Kalotina Machini,[1,2,3] Ozge Ceyhan-Birsoy,[1,7] Danielle R. Azzariti,[1,4] Himanshu Sharma,[1] Peter Rossetti,[1] Lisa Mahanta,[1] Laura Hutchinson,[1] Heather McLaughlin,[1,8] The MedSeq Project, Robert C. Green,[3,4,5] Matthew Lebo,[1,2,3,4,9] and Heidi L. Rehm[1,2,3,4,6,9,*]

Although genome sequencing is increasingly available in clinical and research settings, many questions remain about the interpretation of sequencing data. In the MedSeq Project, we explored how much effort is required to evaluate and report on more than 4,500 genes reportedly associated with monogenic conditions, as well as pharmacogenomic (PGx) markers, blood antigen serotyping, and polygenic risk scores in 100 individuals (50 with cardiomyopathy and 50 healthy) randomized to the sequencing arm. We defined the quality thresholds for determining the need for Sanger confirmation. Finally, we examined the effort needed and new findings revealed by re-analyzing each genome (6–23 months after initial analysis; mean 13 months). Monogenic disease risk and carrier status were reported in 21% and 94% of participants, respectively. Only two participants had no monogenic disease risk or carrier status identified. For the PGx results (18 genotypes in six genes for five drugs), the identified diplotypes prompted recommendation for non-standard dosing of at least one of the analyzed drugs in 95% of participants. For blood antigen studies, we found that 31% of participants had a rare blood antigen genotype. In the cardiomyopathy cohort, an explanation for disease was identified in 48% of individuals. Over the course of the study, 14 variants were reclassified and, upon reanalysis, 18 new variants met criteria for reporting. These findings highlight the quantity of medically relevant findings from a broad analysis of genomic sequencing data as well as the need for periodic reinterpretation and re-analysis of data for both diagnostic indications and secondary findings.

Introduction

Variant interpretation is a challenging aspect of genomic testing. It involves gathering information that will be used for the formal assessment of the pathogenicity of any given variant. The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published guidelines to aid in the interpretation of sequence variants;[1] these guidelines have helped create consistency and rigor in the assessment of evidence.[2] Annotated variant databases cataloging published and unpublished variant interpretations, often with accompanying citations and/or direct evidence, represent an important resource for variant assessment. These databases include ClinVar,[3] the Human Gene Mutation Database (HGMD),[4] and many locus-specific databases.[5,6] However, in all these databases, there exist conflicting and/or incorrect interpretations, and every entry represents only a point-in-time interpretation, such that manual review of the primary evidence and search for new evidence must be performed before one can reliably report a variant in a clinical context or rigorously inform research studies.[2,7,8] Furthermore, the interpretation of variants in asymptomatic individuals changes the prior probability of pathogenicity so that this interpretation requires much more rigorous examination than has been typical of certain variant classes such as seemingly disruptive variants (e.g., frameshift, nonsense, and canonical splice-site variants).

Another major challenge raised by the incorporation of genomic approaches in medical practice has been deciding which genes and diseases are relevant for reporting incidental or secondary findings (clinically relevant variants in genes unrelated to the patient's indication for testing). ACMG has issued recommendations for the return of secondary findings; however, these are limited to a set of 56 (over time expanded to 59) actionable genes where prevention and surveillance could significantly reduce mortality and morbidity.[9,10] However, most genes and diseases have not been systematically evaluated for actionability (e.g., Clinical Actionability Curations), and gene-disease associations are constantly being discovered.[11,12] The Online Mendelian Inheritance in Man (OMIM) database has been foundational for defining new disease-gene associations; however, the clinical validity of these relationships has only recently begun to be systematically evaluated. A standardized framework for the assessment of the clinical validity of gene-disease relationships has been developed by the Clinical Genome Resource (ClinGen)[13] and is now being applied to many genes, but reviewing all published

[1]Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, Cambridge, MA 02139, USA; [2]Department of Pathology, Brigham & Women's Hospital, Boston, MA 02115, USA; [3]Harvard Medical School, Boston, MA 02115, USA; [4]Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [5]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA; [6]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
[7]Present address: Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[8]Present address: Invitae, San Francisco, CA 94103, USA
[9]These authors contributed equally to this work
*Correspondence: hrehm@mgh.harvard.edu
https://doi.org/10.1016/j.ajhg.2019.05.017.

gene-disease relationships as well as keeping up with the constantly changing literature will take time.

In addition to the interpretation and reporting challenges, differing practices and debate remain over the need for orthogonal confirmatory testing, typically performed by Sanger sequencing, of all next-generation sequencing (NGS) variant calls that would be reported to a patient. Increasing evidence documents the reliability of NGS data when best-practice technologies and analytical tools, along with well-validated pipelines and quality metrics, are used.[14,15]

And finally, recent studies reanalyzing exome and genome data suggest that re-evaluation of sequencing data, including that for newly discovered candidate genes, can increase the diagnostic yield;[16,17] however, there are as yet no specific recommendations around best practices for reinterpretation of variants and reanalysis of genomic sequencing tests in a diagnostic setting, let alone a setting for healthy genome screening. These challenges and the sparse data on systematic and comprehensive genome interpretation have contributed to calls for caution about offering genome sequencing (GS) to healthy people.[18] And to date, only a limited number of research studies have been designed to return genomic sequencing results to apparently healthy participants.[19,20]

We previously reported our approach to defining the content and format of the genome report and the results of the initial 20 genomes.[21] Herein, we present the details of our clinical genomic analysis platform and results for the full set of 100 genomes returned to participants in the MedSeq Project; these details include datasets and approaches to facilitating accurate and efficient analysis of genomic sequence data for Mendelian disease risk, carrier status, pharmacogenomic (PGx) profiles, and blood antigen results. We outline approaches to more effectively evaluating predicted loss-of-function (LoF) variation in the low prior probability of pathogenicity context of secondary findings and healthy genomic screening. We also summarize our data from systematic reanalysis of genomes and Sanger confirmation to aid in developing best practices around these evolving areas of genomic testing.

Material and Methods

Project Design

The MedSeq Project design has been described previously.[21,22] In brief, the MedSeq Project was a randomized trial that compared GS to standard of care in two clinical settings, cardiomyopathy clinics, and primary care. In this framework, cardiologists and primary-care physicians recruited 200 patients altogether: 100 patient participants (50 with cardiomyopathy and 50 from primary care) were randomly assigned to receive family history assessment with GS. The study was approved by the Partners Healthcare institutional review board (IRB protocol #2009P002190), and informed consent was obtained from human subjects. The Partners Human Research Committee approved this study.

Genome Sequencing

GS was performed at the Illumina Clinical Services Laboratory as described previously.[21,22] Genomes were sequenced between 2013 and 2015 on the HiSeq 2000 platform to at least 30× mean coverage, and a minimum of 95% of bases were sequenced to at least 8× coverage. Once sequencing was complete, raw data files were sent to the Laboratory for Molecular Medicine (LMM) at Partners HealthCare Personalized Medicine via an encrypted hard drive.

Sequencing Analysis

Alignment, variant calling, and annotation were performed as previously described.[23] In brief, FASTQ sequences were aligned to the reference hg19 genome using the Burrows-Wheeler Aligner (BWA). Variant calls were made with the Genomic Analysis Toolkit (GATK), and all positions with $\geq$ 8× coverage were considered to have adequate coverage. Variant annotation was derived from ALAMUT HT, Variant Effect Predictor, and the LMM's GeneInsight laboratory database. The Quality by Depth (QD), Fisher Strand Bias (FS), and Mapping Quality (MQ) scores were calculated for all variants. After developing our quality metric thresholds through routine Sanger confirmation, as described in the results, we excluded variants with scores of QD $\leq$ 4 or FS $\geq$ 30 from review as likely false positives.

Gene and Variant Filtration

After annotation, we applied two additional filters to identify (A) variants with an entry in HGMD and a minor-allele frequency (MAF) $<$ 5% in European American (EA) or African American (AA) chromosomes from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) and (B) predicted LoF variants with a MAF $<$ 1% from a list of genes with possible disease associations. The genes were drawn from four sources: HGMD, OMIM, Uniprot, and ClinVar. The content of this list evolved throughout the study as sources were updated. In addition, patients in the cardiomyopathy arm passed through another filter that identified all variants in 102 genes associated with monogenic cardiovascular disease.

NGS Quality Metrics and Sanger Confirmation

All variants that had scores of QD $>$ 4 and FS $<$ 30 and that were considered likely to be reported after primary evidence review were sent for Sanger confirmation. The Sanger methodology used has been described previously.[24] Primers were either newly designed or reused if available in the LMM's primer database.

Gene-Disease Validity Assessment

Evidence for gene-disease associations was reviewed in a manner consistent with the ClinGen framework. Although detailed points for each piece of evidence were not tracked, the same principles were applied and each gene-disease association was assigned to the following six categories: (1) strong or definitive, (2) moderate, (3) insufficient (ClinGen categories "limited," "no evidence," "disputed," and "refuted"), (4) only claim is from genome-wide association study (GWAS), (5) pharmacogenetic association only, and (6) trait (only associated to a benign trait or to a biochemical finding without a clinical phenotype). The first three categories—strong or definitive, moderate, or insufficient—represent a condensed version of the six ClinGen categories, and in a recent comparison of these categories to 64 overlapping gene-disease pairs curated by ClinGen curators, we did not find any discordant

classifications that were not explained by more recently published literature (data not shown). We published classifications of 1,514 gene-disease associations by using this approach in the context of sequencing analysis for newborn infants.[25] We reassessed genes with insufficient evidence during the reanalysis to determine whether new data had been published.

Rare-Variant Interpretation

Variant interpretation was carried out as described previously[21,26] and in accordance with the criteria set by the ACMG-AMP guidelines[1] except as noted below and in the Discussion section of this article with regard to more fine-grained assessment of predicted LoF variants. Each variant was classified into one of five categories: pathogenic (P), likely pathogenic (LP), uncertain significance, likely benign (LB), or benign (B). Variants of uncertain significance (VUS) were further classified as VUS-favor pathogenic (VUS-FP), VUS, or VUS-favor benign (VUS-FB).

In addition, we developed a systematic approach to the assessment of predicted LoF variants (i.e., frameshift, nonsense, +/−1, two splice sites without functional evidence of their effect) by addressing each of the questions listed below.

1. Is there sufficient evidence for the gene's role in human disease (meets ClinGen criteria for definitive/strong/moderate)?
2. Is LoF a well-established mechanism of disease?
3. Are the constraint score and frequency of heterozygous and homozygous predicted LoF variants in large population datasets such as gnomAD consistent with the reported prevalence of the disease and mechanism of pathogenicity?
4. Is the exon present in all biologically relevant transcripts?
5. Is the transcript predicted to undergo nonsense-mediated decay? (Is the new stop codon in the last exon or in the last 50 bases of the penultimate exon?) Are there other pathogenic variants in the same exon, thus supporting the functional importance of the exon?
6. Are there pathogenic LoF variants downstream or upstream of the variant in question?
7. For canonical splice site variants, does the alteration leave the transcript or reading frame intact?

If the answers to all questions led to a confident prediction of LoF, then the variant was generally reported as likely pathogenic if not yet observed in affected individuals or pathogenic if previously observed in one or more affected individuals.

Return of Results Criteria

Variants in the categories P, LP, and VUS-FP were considered for return if the gene met at least a strong level of evidence for association with a monogenic disease. Variants in genes with moderate evidence were also chosen for return if agreed upon after discussion with the broader MedSeq team. Genes that did not meet criteria for return were placed on an exclusion list to eliminate review of other subsequent variants identified in those genes. Variants in the category of VUS were also returned if they were relevant to the specific indication in cardiomyopathy patients. During Mendelian disease gene assessment, prevalence and inheritance (AD, AR, XL) of associated diseases, as well as whether LoF was or was not an established disease mechanism, were recorded.

A set of 18 variants with PharmGKB clinical annotation levels of evidence class I and class II and associated with the metabolism of five drugs (PGx ariants for metformin [C11orf65 rs11212617], clopidogrel [CYP2C19 rs12248560, rs4244285, rs4986893, rs28399504, rs41291556, rs72552267, rs72558186, and rs56337013], warfarin [CYP2C9 rs1057910, rs1799853, rs7900194, rs9332131, rs28371685, and rs28371686, and VKORC1 rs9923231], simvastatin [SLCO1B1 rs4149056], and digoxin [ABCB1 rs1045642] metabolism) commonly used in the treatment of primary care and cardiology patients were pre-selected for return at the beginning of the study. Pharmacogenetic variants other than those listed and traits (as defined above) were also excluded from reporting. In addition, we did not report on rare variants in genes with gene-disease links established only through GWAS, although we did return polygenic risk scores for eight cardiovascular phenotypes (abdominal aortic aneurysm, atrial fibrillation, coronary heart disease, type 2 diabetes, hypertension, obesity, platelet aggregation, and QT prolongation). In brief, we estimated multiplicative polygenic risk scores by using 3–111 published risk alleles per condition and then normalized the scores by using the population median estimated from the 1000 Genomes Project, as described previously.[27]

Report Content

The format and content of the MedSeq Genome report has been described.[21,22] In brief, the reports contained findings related to (1) diagnostic indication (for cardiomyopathy patients; VUS, VUS-FP, LP, P), (2) monogenic disease risk (heterozygous for dominant disorders, homozygous or compound heterozygous for recessive disorders, and hemizygous for X-linked disorders; VUS-FP, LP, P), (3) carrier status for recessive disorders (VUS-FP, LP, P), (4) PGx results for five drugs, (5) blood group antigens,[28] and (6) complex-trait analyses.[27] PGx results encompassed 18 genotypes in six genes (*ABCB1, C11orf65, CYP2C9, CYP2C19, SLC01B1,* and *VKORC1*) associated with the metabolism of five drugs (clopidogrel, digoxin, metformin, simvastatin, and warfarin) commonly used in the treatment of primary-care and cardiology patients (see Table S1). The approach used for predictions of red blood cells and human platelet antigen has been previously described.[28,29]

Report Delivery and Knowledge Updates

MedSeq reports were communicated to providers through an electronic pdf document as well as through our GeneInsight Clinic (GIC) software, which was designed to support physician access to structured genetic data and updates to knowledge over time.[30–32] The GeneInsight software is connected to the LMM's GeneInsight variant knowledgebase such that variant classification updates are automatically delivered to the ordering physician (via email with a link to the GIC) to enable return to patients.[30,31,33,34] Provider access is supported by GIC, which exists both as a standalone software system and through integration with Epic, the Partners Healthcare electronic health record (EHR) in a single sign-on environment.[30,31,34,35] Notifications to providers regarding reclassified variants, as well as updated reports with newly identified or removed variants after GS reanalysis, were delivered via GIC.

Results

Filtration Approach and Gene Curation

To identify the monogenic disease risk and carrier status variants for possible reporting, we relied on two strategies: the identification of variants previously reported as

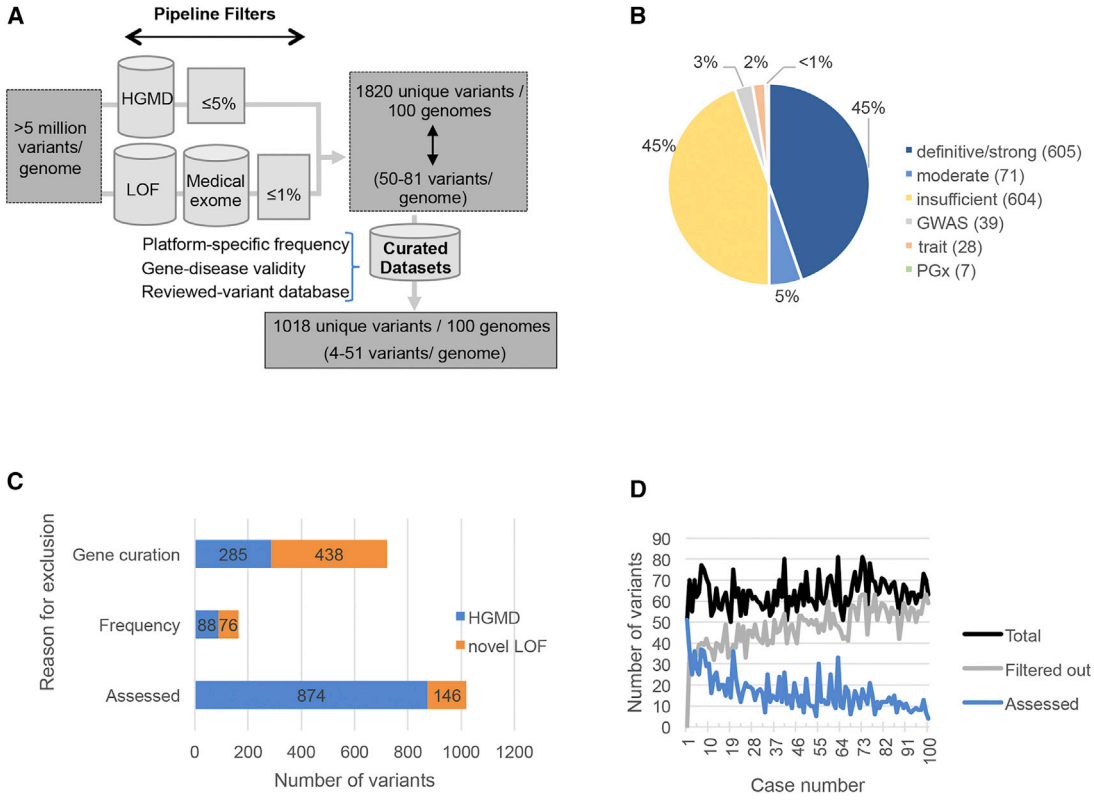
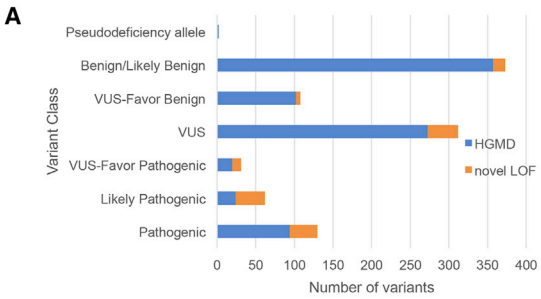
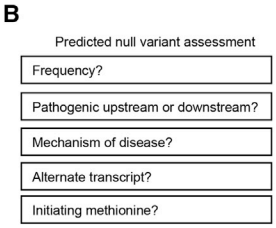


Figure 1. General Filtration Approach and Overall Outcome
(A) Schematic representation of the filters used for rare-variant interpretation of the MedSeq genomes. HGMD = Human Gene Mutation Database; LoF = loss of function.
(B) Pie chart showing the results of gene curation performed for all genes that had at least one variant that came through the MedSeq genome filter.
(C) Bar graph showing the breakdown of unique variants queued for review and their sources (HGMD versus LoF). Variants were excluded either for insufficient gene-disease validity (gene disease validity) or for high MedSeq allele frequency (platform-specific frequency). Please note that several variants were excluded on the basis of both gene disease validity and platform-specific frequency.
(D) Graph showing the reduction of variants that required review as a function of the number of genomes reviewed.

disease-associated (and MAF < 5%) and the identification of novel variants with a predicted LoF effect (and MAF < 1%) in any of 4,631 genes (over time expanded to 5,860) with a reported association with disease (Figure 1A). After reviewing variants for the first 40 cases, we developed a GS platform-specific variant frequency database (primarily populated by MedSeq cases) and excluded 164 variants that had a prior case frequency above 10% regardless of their reported population frequency (or their absence from population frequency databases). Such variants were often technical sequencing artifacts or real variants (often indels) with nomenclature descriptions that did not align with the naming provided by population allele frequency databases. We also began a list of known pathogenic variants with high allele frequencies (often with lower penetrance) to ensure that these variants would not be inadvertently filtered out (e.g., *HFE* c.187C>G [p.His63Asp] and *HFE* c.845G>A [p.Cys282Tyr]; factor V Leiden c.1601G>A [p.Arg534Gln]; *BTD* c.1330G>C [p.Asp444His]; *SERPINA1* Z c.1096G>A [p.Glu366Lys] and *SERPINA1* S c.863A>T [p.Glu288Val]; *CFTR* c.1521_1523delCTT [p.Phe508del]; *CHEK2* c.1100delC [p.Thr367fs]; and *GJB2* c.35delG [p.Gly12Valfs], *GJB2* c.101T>C [p.Met34Thr], and *GJB2* c.109G>A [p.Val37Ile]).

For every variant identified after filtration, the gene-disease evidence was first reviewed or pulled in from a previous case. Variants went on to variant-level review if the genes met at least a strong or moderate level of evidence. Genes that did not meet this evidence level were placed on an exclusion list so that review of other variants subsequently identified in those genes would be eliminated. Of the 4,631 genes (over time expanded to 5,860) designated for possible return, 1,354 genes had one or more variants detected by our filter and were therefore reviewed during the course of the project. Of those 1,354 genes, 44.7% (605/1,354) were classified as definitive or strong, 5.2% (71/1,354) as moderate, 44.6% (604/1,354) as insufficient, and 2.1% (28/1,354) as traits, and the remaining 3.4% (46/1,354) were excluded on the basis of one of the criteria outlined in the methods (e.g., identified only through GWAS studies, variants in pharmacogenetic loci other than the ones selected in the beginning of the study, or non-disease traits) (Figure 1B and Table S2). Furthermore, for 190 of these 1,354 genes, the only variants identified





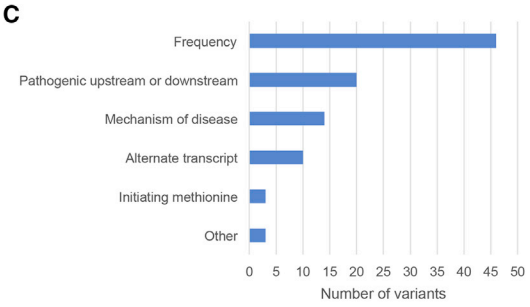


Figure 2. Framework for Predicted LoF Variant Assessment and Results of Manual Variant Curation

(A) Final variant classifications with filter source noted. Blue bars represent variants from the HGMD filter, whereas orange bars illustrate novel predicted LoF variants. (B) Schematic depiction of the LoF checklist designed to assist in the rapid and accurate classification of predicted LoF variants. (C) Application of the predicted LoF checklist on novel as well as previously reported predicted LoF variants.

did not meet our quality thresholds (QD > 4, FS < 30). Of the remaining 1,164 genes whose variant calls met our quality thresholds, 457 of them were in the insufficient gene-level evidence category, and therefore variants were not reviewed for variant-level evidence. Of the 457 genes in the insufficient category, 411 were from annotated gene-disease association sources, including 128 (28%) from the OMIM Morbid Map and 209 (45.7%) having at least one disease-causing mutation (DM) in HGMD.

An average of 5.3 million variants were identified per subject, and the filtration output generated an average of 67 (range 50–81) variants per subject (6,700 variant observations total, 1,820 unique variants). In terms of unique variants, 601 variants came exclusively through the LoF filter (novel LoF), and 1,219 variants came from the HGMD filter. This latter number included 130 previously reported LoF variants (i.e., identified by both filters). During variant triage, 39.7% (723/1,820) of the variants were excluded from assessment, which was based on gene curation. Those variants excluded from assessment included 23.4% (285/1,219) of HGMD variants and 72.9% (438/601) of novel LOF variants. It is worth noting that 91 variants were excluded based on platform-specific frequency in addition to gene curation (Figure 1C). The integration of three curated databases, namely the two exclusion filters (platform-specific allele frequency and gene-disease validity) and the knowledge base of previously reviewed variants, permitted us to gradually decrease the number of variants that required manual review over the course of the study (Figure 1D).

Variant Classification

Of the initial 1,820 unique variants, after excluding 802 variants for gene-disease validity and/or platform frequency, the remaining 1,018 underwent variant classification resulting in 373/1,018 (36.6%) of the variants classified as B or LB, 451 (44.3%) classified as of uncertain significance (including 31 classified as VUS-FP, 312 classified as VUS, and 108 classified as VUS-FB), 62 (6.1%) classified as LP, and 130 (12.8%) classified as P; two variants were pseudodeficiency alleles (Figure 2A). Of 217 reportable variants (P, LP, and a subset of VUS-FP), 134 (61.7%) were previously reported as DM or probable/possible disease-causing mutation (DM?) in HGMD (these included 50 predicted or demonstrated LoFs and six variants affecting non-canonical splice or regulatory regions), and 83 (38.2%) were novel predicted LoF variants. 12 reported variants were in genes with moderate evidence for association with disease, and 10 of these variants were LoF. The majority of variant calls that met criteria for return were detected once (89.6%), and the remaining 10.4% were detected between two and 25 times.

It should be noted that out of 461 variants that were categorized as DM? in HGMD, only 5/461 (1.1%) met our reporting criteria, and of those categorized as DM in HGMD, only 134/758 (17.7%) met our reporting criteria (Table S3).

Predicted LoF Variant Assessment

As part of the variant assessment, we performed a systematic analysis of the 731 predicted LoF variants (frameshift, nonsense, and canonical splice site) identified in 589 genes, and we showed that 502/731 (68.7%) were excluded from the report because the gene met one of the exclusion criteria listed previously (e.g., there was insufficient [416] or moderate [six] evidence for a causal role in Mendelian disease, the gene was responsible for a non-disease trait [26], the only claim was from GWAS [46], or there was a PGx association only [eight]). We then applied a series of questions to systematically evaluate the remaining 229 predicted LoF variants in order to provide more detailed guidance than the ACMG-AMP guidelines.[1] The approach is described in detail in the methods section and shown in Figure 2B.

Of the 229 predicted LoF variants passing gene-disease validity assessment, 96 were not reported to participants

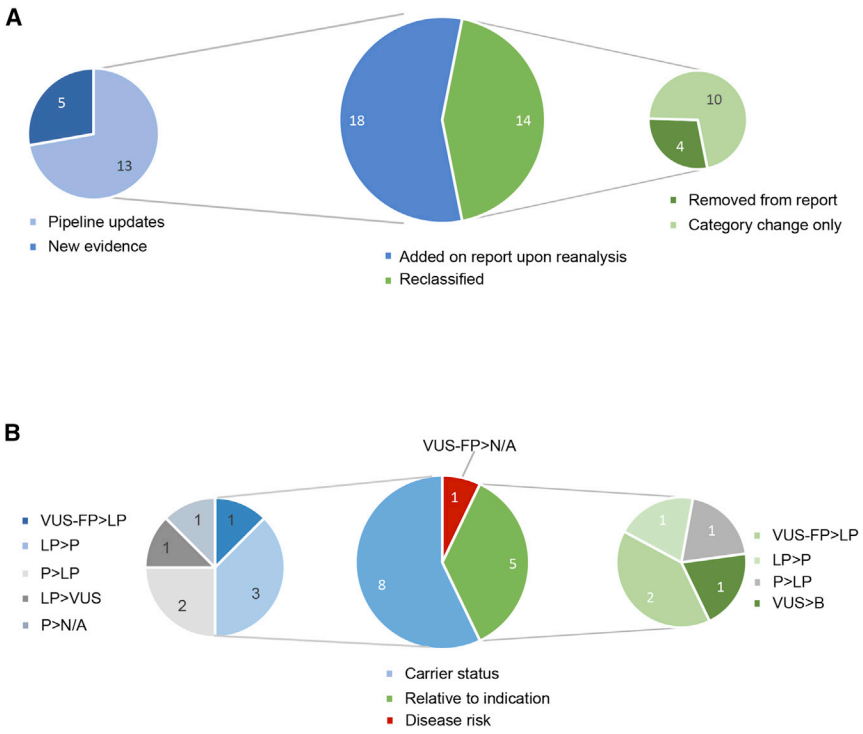


Figure 3. Summary of Genome Reanalysis Findings and Variant Reclassification
(A) The middle pie chart depicts the overall findings, including the number of variants added onto the report upon genome reanalysis (shown in blue), and variants reclassified over the course of the study (shown in green). The pie chart on the left represents reportable variants newly discovered as a result of pipeline updates (light blue) or existing variants newly reported based on new evidence during variant reanalysis (blue). The pie chart on the right shows the consequences of reclassification of previously reported variants; variants removed from reports are shown in dark green, and variants with a category change only (but still reportable) are shown in light green.
(B) Schematic illustration of variant reclassification categories. The middle pie chart shows the number of variants reclassified (relevant to indication in green, monogenic disease risk in red, and carrier status in blue). The pie chart on the left illustrates the various classification changes in carrier status variants, whereas the pie chart on the right focuses on variant reclassifications relative to cardiomyopathy indication. Please note that one monogenic disease variant was reclassified (from VUS-FP to N/A) and was removed from the report because new evidence disputed the gene's association with disease.

for the following reasons: the gene did not have an established LoF mechanism of disease (14); the exon in which the variant was found was excluded in alternate transcripts (10); the variant was located downstream of the most 3′ pathogenic variant (18) or early in the coding region and upstream of the most 5′ variant (2) or the variant affected the initiating methionine (3); the variant's frequency in large population datasets (24) or our platform-specific MedSeq cohort (22) was higher than expected given disease prevalence; and/or other reasons (i.e., the variant was a GWAS variant [1] or trait [2]) (Figure 2C).

Reanalysis

Upon completion of the initial analysis, we sought to determine the rate of new reportable findings through systematic reanalysis of the genomes. The initial analyses happened between July 2013 and February 2015, and during that time, significant changes in our genome interpretation pipeline occurred, i.e., changes included updated versions of HGMD; expansion of the medical exome gene list; updates in ESP, Alamut, and dbSNP; and the addition of and ongoing updates to ClinVar. Using an updated pipeline, we reanalyzed the variant cell format (vcf) files of all MedSeq genomes between August and September 2015 (mean period lapsed between initial and repeat analysis: 13 months, range 6–23 months). This resulted in the identification of 315 additional variants, and upon review, a total of 4% (13/315) met criteria for return. Furthermore, variant reassessment showed that five additional variants from the original 1,820 that were reviewed for return in the initial analysis, but that were excluded from initial reports, were returnable. It should be noted that in addition to thirteen new variants that were reported, two reported variants were removed from reports because of new evidence disputing an association between the gene and disease in one case or highlighting concerns over the technical validity of the variant in another case (Figure 3A).

A deeper analysis of the origin of the thirteen newly returned variants determined that four were returned as a result of new variants reported in online databases, showing the value of periodic reanalysis. An additional four newly added variants were newly returned as a result of pipeline annotation limitations in the initial analysis. For example, two variants had been missed because of incomplete annotation at multiallelic sites, and two were found as a result of improvements in determining predicted LoF variants. The overall rate of returnable variants (13/315; 4.4%) from the additional newly identified variants was statistically significantly lower than the 11.9% returnable rate in the initial analysis (217/1820; $p < 0.0001$). This could be explained if reanalysis often identifies newer variants without accumulated evidence of pathogenicity or genes that have a lower evidence base and are less likely to meet criteria for return.

In addition to the two variants removed from reports during full genomic reanalysis, over the course of the

MedSeq Project, 12 other variants in 13 patients were reclassified: five of these variants were related to the patients' cardiomyopathy diagnoses (two VUS-FP > LP, one LP > P, one P > LP, and one VUS > B), and seven were recessive carrier-status variants (3 LP > P, 2 P > LP, 1 LP > VUS, and 1 VUS-FP > LP). These updates were the result of variant reassessment triggered by the detection of the variant in other cases (MedSeq or non-MedSeq) (Figure 3B).

In addition, the 50 cardiomyopathy genomes were reanalyzed for new causes of cardiomyopathy, and two cases received updates with variants in *ALPK3* (MIM: 617608), a more recently discovered cause of cardiomyopathy[36] (one bi-allelic variant explaining disease and one variant that was heterozygous and therefore inconclusive in the absence of a variant on the second allele). In summary, a total of 22% (22/100) of cases were updated; nine received new variants, ten received updated variant classifications, and three received both types of updates. All updated reports were delivered via the GeneInsight Clinic system.

Overall Findings

Altogether, a total of 100 MedSeq patients received results spanning monogenic disease, carrier status, PGx, blood-antigen predictions, and complex-trait analyses. For the cardiomyopathy cohort, an explanation for disease was identified in 48% (24/50) during the initial analysis (Table S4). If diagnostic cardiomyopathy results are excluded, monogenic disease risk findings were reported in 21% (21/100) of participants: 14 with P or LP variants, four with Factor V Leiden risk alleles, and three with variants categorized as VUS-FP (One participant had both a VUS-FP variant conferring monogenic disease risk and a Factor V Leiden risk allele.) (Table S5). Of these results, all were for dominantly inherited diseases except that three individuals had homozygous or compound heterozygous recessive *HFE* variants and one hemizygous male had an X-linked variant. Nearly all patients (94%) had carrier status reported, and the number of heterozygous recessive variants ranged from 0–7 per patient (average 2.6 carrier status variants per person) (Table S6). Only two MedSeq participants had no reported monogenic disease risk or carrier-status variants identified. For the PGx results, across the 18 loci and five drugs, diplotype analysis prompted recommendation for non-standard drug dosing in at least one drug in 95% (95/100) of participants (Table S1). For the polygenic risk results addressing eight cardiovascular phenotypes (abdominal aortic aneurysm, atrial fibrillation, coronary heart disease, type 2 diabetes, hypertension, obesity, platelet aggregation, and QT prolongation), 58% (29/50) of cardiomyopathy patients and 60% (30/50) of primary-care participants were at the 90^{th}–100^{th} percentile rank of relative risk for at least one of the above-mentioned phenotypes (range 0–3). Most additional laboratory and cardiac tests in the MedSeq cohort were prompted by polygenic risk estimates for cardiometabolic traits or *HFE* carrier variant status.[37] For blood-antigen studies, we identified 31% (31/100) of patients with presence of a rare (less than 5% population frequency) blood-antigen genotype.[29]

Comparison of WGS Versus NGS Panel

All 50 cardiomyopathy cases were tested by both a disease-targeted cardiomyopathy panel and GS. Both the panel testing and indication-based analysis of GS data for causes of cardiomyopathy were performed with the tester blinded to the other results.[38] In one patient, an 18 bp duplication was identified by panel testing and not by GS. On the basis of manual review of GS data showing one out of 12 reads with the variant, we believe this discrepancy was most likely due to reduced sequencing coverage and alignment impacts given the size of the duplication. Four patients were found to have additional findings on GS in genes not included in the panel tests. One finding offered a definitive diagnosis: a pathogenic variant in the *PTPN11* gene (MIM: 176876), which is not consistently included in HCM gene panels, and only upon re-examination in light of the molecular findings was the patient's phenotype recognized as consistent with Noonan syndrome with multiple lentigines (NSML, formerly known as LEOPARD syndrome, [MIM: 151100]). Three additional patients had variants with an uncertain role in disease identified by genome but not panel testing: a heterozygous LoF mutation in *ILK* (MIM: 602366), a gene implicated in cardiomyopathy but with only limited evidence to date;[39] a heterozygous VUS in *FLNC* (MIM: 102565) implicated in adult-onset myopathy but with only limited evidence for a role in cardiomyopathy; and a VUS in *RBM20* (MIM: 613171), a known dilated cardiomyopathy gene that had not been included in the patient's original panel test. It is worth noting that this patient also carried a VUS in *ACTN2* (MIM: 102573). In addition, reanalysis identified variants in a novel cardiomyopathy gene, *ALPK3* (MIM: 617608), in two cases as described above.

Sanger Confirmation

Prior to running the first MedSeq cases, we established minimum quality thresholds for determining false positive calls versus potentially true positive calls. Based upon a small set of preliminary data from the reference sample NA12878, we set initial minimum quality thresholds (QD > 4 and/or FS < 30) with an aim of maintaining high sensitivity.[23] We then evaluated these criteria with Sanger confirmation of results from MedSeq genomes to see how they performed in practice, using data from 487 genomic variants. In total, 487 variant observations (407 unique, 30 detected in more than one case) underwent Sanger confirmation. For 12 of these variants, Sanger confirmation could not be performed because they had repeated failed reactions, most likely due to difficult-to-sequence genomic regions. Of the remaining variants, 5.1% (25/475) of the variants were not confirmed, with the majority falling below our subsequently set minimum quality thresholds (18/25 had

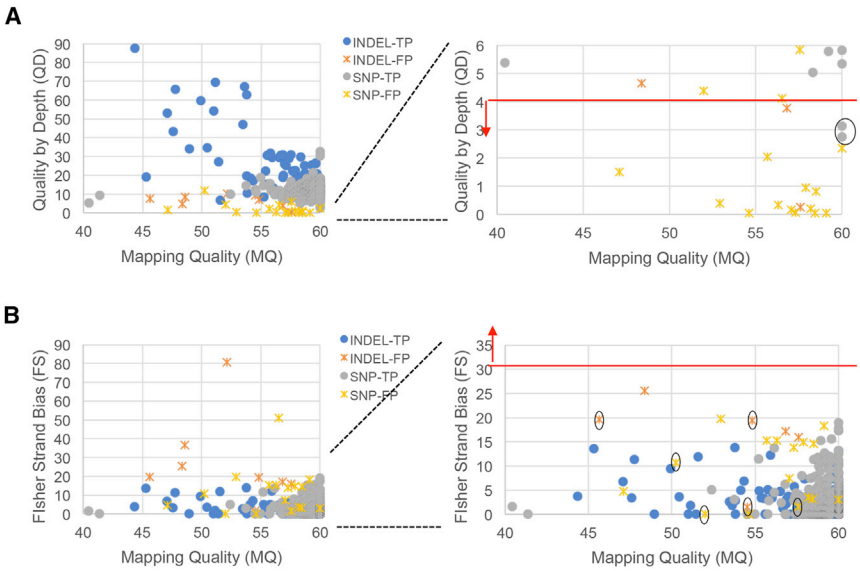


Figure 4. Results of Sanger Confirmation with Scatterplot of Variants That Underwent Sanger Testing

(A and B) Blue circles represent true positive (TP) indels and gray circles represent TP SNPs. Orange crosses represent false positive (FP) indels and yellow crosses represent FP SNPs.

(A)The y axis corresponds to the quality by depth (QD) and the x axis to the mapping quality (MQ). The right plot is a magnification showing variants with QD < 6. The red line indicates the minimum QD threshold, below which we stopped follow-up (QD < 4). The two TP variants with QD < 4 are circled in black.

(B) The y axis corresponds to the Fisher strand bias (FS) and the x axis to the MQ. The right plot is a magnification showing variants with FS < 35. The red line indicates the minimum FS threshold, above which we stopped follow-up (FS > 30). The six FP variants that also had QD > 4 are circled in black.

QD < 4 and/or FS > 30). All of the seven variants (four indels and three single-nucleotide polymorphisms [SNPs]) that were not confirmed despite having QD and FS metrics within our quality thresholds had suboptimal mapping quality (MQ) (mean 51.9, range 45.6–57.6); this emphasizes the need for additional metrics before Sanger confirmation is no longer necessary. Sanger-confirmed variants had an average MQ of 58.5 (range 40.4–61.2). Of the 450 that were confirmed, 442 variants had correct nomenclature, whereas eight variants were insertion or deletion variants annotated with incorrect nomenclature by the variant caller. These eight variants were primarily complex and/or large events that the variant caller split into multiple variants and for which manual correction based on the NGS data was sufficient for correction of the mis-annotation. In one instance, this correction led to the identification of a called frameshift variant being as an in-frame deletion, whereas in the remaining seven cases, the final interpretation did not change. Interestingly, two of the confirmed variants had quality metrics that did not meet our minimum quality thresholds for follow-up: NKX2-5, c.65A>G (p.Gln22Arg), QD = 3.13, FS = 4.42, MQ = 60; and MYH7, c.2609G>A (p.Arg870His), QD = 2.75, FS = 4.51, MQ = 60 (also identified in the individual through panel testing); these results highlight the challenge of balancing sensitivity and specificity when one uses genomic sequencing technologies that need to cover the entire genome (Figure 4).

Resource Requirements

Over the course of the initial analysis, the time it took for each case to be reviewed by both a primary reviewer and a geneticist was tracked for 90 cases. The time for the interpretation (including fellow assessment and geneticist review) of the first 20 cases (for which we tracked time) was an average of 8.5 h/case, and this had decreased to an average of 6.25 h/case for the last 20 cases (data not shown).

Discussion

Through the MedSeq Project, we sought to address some outstanding issues hampering the interpretation of genomic sequencing results and to facilitate the transition to the era of large-scale genomic screening.

Besides MedSeq, only a limited number of research studies[19,20] have been designed to return genomic sequencing results to apparently healthy participants, and recent studies highlight the difficulties in identifying and returning potentially actionable genetic variants in healthy individuals.[40,41] We elected to return a broad array of data including monogenic disease risk, carrier status, polygenic risk scores for cardiovascular and related disease, limited PGx information, and red blood cell and platelet antigens.

It is worth noting that none of the 11 different monogenic disease findings we reported to MedSeq participants were in the 59 genes included in the ACMG recommendation. The ongoing advances in medical science make it impossible to know which of any monogenic disease findings will be actionable in the future, and actionability can have very broad meanings; therefore, many individuals are increasingly interested in a broader array of results.

New gene-disease associations, novel disease variants, and new evidence on existing variants continue to be identified, and as a result, 22% of patients in the MedSeq Project had new findings or updated results over the course of the study (average time for reanalysis 13 months, range 6–23 months). These findings (on healthy genomes and/or otherwise unrelated to indication) are consistent with other studies in which re-evaluation of sequencing data

has increased the diagnostic yield,[16,17] and they underscore the critical need to develop sustainable approaches to allow reanalysis of sequencing data and mechanisms for returning updated results to patients in order to enhance the utility of genomic testing. For updates to reported variants, we employed the GeneInsight platform to enable automated delivery of knowledge updates without the need for laboratory staff time to report those updates.

Nearly all patients had carrier status for one or more recessive disorders; however, we found that the vast majority of reportable variants are extremely rare (or absent) in the general population and span a large array of genes and diseases. We identified pathogenic or likely pathogenic variants in >120 genes, and <20 genes had more than one reportable variant (same or different among participants). Had we limited reportable findings to those genes present on commercially offered carrier-screening panels (i.e., Counsyl, Integrated, sema4, and Invitae), only 104/222 (46.8%) of these results would have been returned to participants.

For only half of the genes with variants that passed our basic filters did we find at least moderate evidence for disease association to warrant review of variant evidence. This emphasizes the extensive challenges involved with direct use of the scientific literature and gene-disease databases for selecting genes for clinical testing. One must carefully review the evidence for the gene-disease relationship before one considers returning a variant in such a gene. This is particularly important for the identification of LoF variants. LoF is the most frequent mechanism of disease for both recessive and dominant Mendelian disorders, and therefore it is easy to incorrectly assume pathogenicity if the gene-disease evidence is not reviewed and the established mechanism of disease is not delineated.[42] More recently, several studies have pointed out the plethora of LoF variation in dominant genes in the genomes of healthy individuals and have questioned the penetrance of the associated disorders.[7,43,44] However, our standardized assessment of predicted LoF variants detected in the genomes of MedSeq participants suggests that over half are simply in genes with limited validity for a gene-disease relationship but that others might not actually be found to disrupt gene function upon splicing impact analysis and scrutiny of transcript and location.

This issue highlights the major difference between analyzing genomes from healthy individuals and examining a gene in a diagnostic setting. The prior probability of true LoF of a variant is much higher when observed in a gene already implicated in a patient's phenotype than it is in those genes found in healthy patients. The actual probability relates primarily to the specificity of the gene-phenotype relationship, but nonetheless caution is warranted in the interpretation of apparent LoF variants, such as frameshift, nonsense, and canonical splice-site variants, in healthy genome analysis.

It should be noted that our approach is roughly consistent with the 2015 ACMG-AMP guidelines and even more so with the recent ClinGen PVS1 flowchart,[45] which provides more detailed structure for addressing the caveats stated in the ACMG-AMP guidelines. However, appropriate classification of predicted LoF variants according to the ACMG-AMP guidelines, particularly if PM2 is not met (e.g., absence in the population) yet the allele frequency is still consistent with disease prevalence, is challenging, and many variants do not achieve "likely pathogenic" or "pathogenic" status when this is the professional opinion of the laboratory. We feel that our approach is more consistent with professional opinion on the significance of this class of variants when careful assessment of all caveats is taken into account so that overinterpretation is avoided. These issues have been acknowledged by the ClinGen Sequence Variant Interpretation working group and are slated to be addressed by the new ACMG-AMP-ClinGen working group recently convened to update the 2015 ACMG-AMP guidelines.

To help facilitate gene curation and variant reanalysis in more automated ways, laboratories will need to share interpreted variants and supporting evidence to enable higher-throughput support for both primary analysis and reinterpretation. The ClinVar database has continued to grow at a linear rate, suggesting that we are not close to saturating clinically relevant variant identification and underscoring the importance of robust sharing of all variant evidence so that the community can fully benefit from this resource. ClinGen's ongoing evaluation of gene-disease validity will assist in further focusing the analysis of relevant genes, decreasing VUS rates, and shortening genome analysis times.

As we look to a future where GS is more commonplace, it is used for multiple indications per individual, and the sequence data might reside in the EHR accessible to physicians, new paradigms for how to support reinterpretation as well as interpretation for additional indications will be needed. For reinterpretation, a direct pipeline to ClinVar might be sufficient to trigger updates of previously reported variants, vastly reducing the laboratory resources needed, as with the automated updates we describe for our use of the GeneInsight system, which is directly integrated into our EHR. However, reanalysis that would account for updated variant callers or newly discovered relationships between genes and disease—and determine whether any variants newly identified might explain old or new disease indications—is likely to continue to involve some manual effort on the part of the laboratory. The appropriate frequency of reanalysis remains to be understood, but on the basis of our results, we feel that full reanalysis and reinterpretation on an annual basis is likely to continue to yield new findings for many individuals. The presence of new symptoms or availability of new treatments that are specific to genetic findings will undoubtedly play a factor in the frequency and utility of reanalysis.

Laboratories are increasingly moving diagnostic testing platforms from multi-gene panels to exome and genome approaches. The advantages include a reduced number of tests that must be maintained and validated as well as the ability to rapidly update content as new genes are implicated in disease. However, the extent to which laboratories ensure coverage of all known disease-relevant exons and genomic regions, as well as ensure detection of all types of variation implicated in disease, is quite variable. Therefore, we compared genomic versus panel-based approaches for detecting the causes of disease in our cohort of 50 patients with cardiomyopathy. Given the ability to query new disease genes, the genomes were also reanalyzed for causes of cardiomyopathy at the end of the study. Excluding inconclusive results, this analysis showed that GS missed one variant that was identified by panel testing, whereas two etiologies were found by GS (one at initial analysis and one during reanalysis) given the more limited content on panel-based tests. These results highlight the fact that each method still has some advantages over the other, and we are not yet in a state where a single approach to genetic testing is comprehensive and superior to all others. Clinicians must still weigh the benefits and limitations of each test for each patient's indication, and might need to reflex to a broader test, if starting with a panel, or to a more focused and technically comprehensive test, if starting with genomic approaches when initial testing is negative.

A substantive added cost and increase in turn-around time comes from the inclusion of Sanger confirmation in NGS-based genetic tests, a practice that is still standard in most clinical laboratories, whether a panel, exome, or genome testing platform is used. However, the necessity for Sanger confirmation of variants identified through NGS has been a subject of discussion and analysis.[14,15] Increased understanding of the performance metrics of NGS-based tests is allowing removal of Sanger confirmation, particularly when the opportunities for sample mix-up throughout a testing workflow are reduced through sample identity tracking measures. On the basis of our findings, we could reduce the number of variants confirmed by Sanger if we apply more stringent thresholds and consider a combination of metrics. To further refine these thresholds by using confidence intervals and additional quality metrics, we combined our data to inform a larger study, which has been recently published.[15] Implementation of laboratory-established thresholds of this nature can lead to substantive gains in efficiency and cost reductions without compromising quality. However, setting thresholds to balance sensitivity and specificity can vary on the basis of the clinical context; for example, optimizing sensitivity in diagnostic settings can ensure that variants in genes associated with a patient's phenotype are detected.

In summary, in this study, we implemented a robust workflow for full genomic screening to identify clinically significant findings from over 5,000 genes spanning monogenic disease risk, carrier status, PGx findings, rare blood antigens, and complex-trait risk analysis with high-quality interpretation of both gene- and variant-level evidence. This resulted in monogenic-disease risk findings in 21% (21/100) of individuals. Furthermore, 22% (22/100) received updated findings after reanalysis. Use of an ER-integrated interface enabled automated real-time delivery of updates to physicians on previously reported variants. We share our methods and datasets to enable the community to benefit from these approaches and curated results as clinical genome interpretation continues to gain increasing uptake in both diagnostics and presymptomatic screening.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.05.017.

Acknowledgments

Research reported in this publication was supported by the National Institutes of Health under award numbers U01HG006500 and U41HG006834. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of Interests

Most authors are clinical service providers through their stated affiliations. The following authors have an additional conflict of interest: Robert C. Green receives personal compensation from AIA, Genome Medical, Helix, Prudential, Verily, and Veritas for speaking or consulting.

Received: March 6, 2019
Accepted: May 22, 2019
Published: June 27, 2019

Web Resources

ClinGen, https://clinicalgenome.org/curation-activities/gene-disease-validity/
Clinical Actionability Curations, https://search.clinicalgenome.org/kb/actionability
ESP, https://evs.gs.washington.edu/EVS/
HGVS databases and tools, http://www.hgvs.org/content/databases-tools
OMIM, https://www.omim.org/

References

1. 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.
2. Harrison, S.M., Dolinsky, J.S., Knight Johnson, A.E., Pesaran, T., Azzariti, D.R., Bale, S., Chao, E.C., Das, S., Vincent, L.,

and Rehm, H.L. (2017). Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet. Med. *19*, 1096–1104.
3. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44* (D1), D862–D868.
4. Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D., and Cooper, D.N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum. Genet. *136*, 665–677.
5. Pinard, A., Miltgen, M., Blanchard, A., Mathieu, H., Desvignes, J.P., Salgado, D., Fabre, A., Arnaud, P., Barré, L., Krahn, M., et al. (2016). Actionable genes, core databases, and locus-specific databases. Hum. Mutat. *37*, 1299–1307.
6. Fokkema, I.F., Taschner, P.E., Schaafsma, G.C., Celli, J., Laros, J.F., and den Dunnen, J.T. (2011). LOVD v.2.0: the next generation in gene variant databases. Hum. Mutat. *32*, 557–563.
7. 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.
8. Yang, S., Lincoln, S.E., Kobayashi, Y., Nykamp, K., Nussbaum, R.L., and Topper, S. (2017). Sources of discordance among germ-line variant classifications in ClinVar. Genet. Med. *19*, 1118–1126.
9. Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O'Daniel, J.M., Ormond, K.E., et al.; American College of Medical Genetics and Genomics (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet. Med. *15*, 565–574.
10. Kalia, S.S., Adelman, K., Bale, S.J., Chung, W.K., Eng, C., Evans, J.P., Herman, G.E., Hufnagel, S.B., Klein, T.E., Korf, B.R., et al. (2017). 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. *19*, 249–255.
11. 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.
12. Boycott, K.M., Rath, A., Chong, J.X., Hartley, T., Alkuraya, F.S., Baynam, G., Brookes, A.J., Brudno, M., Carracedo, A., den Dunnen, J.T., et al. (2017). International cooperation to enable the diagnosis of all rare genetic diseases. Am. J. Hum. Genet. *100*, 695–705.
13. Strande, N.T., Riggs, E.R., Buchanan, A.H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S.S., Goldstein, J., Ghosh, R., Seifert, B.A., Sneddon, T.P., et al. (2017). evaluating the clinical validity of gene-disease associations: An evidence-based framework developed by the clinical genome resource. Am. J. Hum. Genet. *100*, 895–906.
14. Baudhuin, L.M., Lagerstedt, S.A., Klee, E.W., Fadra, N., Oglesbee, D., and Ferber, M.J. (2015). Confirming variants in next-generation sequencing panel testing by Sanger sequencing. J. Mol. Diagn. *17*, 456–461.
15. Roy, S., Coldren, C., Karunamurthy, A., Kip, N.S., Klee, E.W., Lincoln, S.E., Leon, A., Pullambhatla, M., Temple-Smolkin, R.L., Voelkerding, K.V., et al. (2018). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. J. Mol. Diagn. *20*, 4–27.
16. Wenger, A.M., Guturu, H., Bernstein, J.A., and Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet. Med. *19*, 209–214.
17. Costain, G., Jobling, R., Walker, S., Reuter, M.S., Snell, M., Bowdin, S., Cohn, R.D., Dupuis, L., Hewson, S., Mercimek-Andrews, S., et al. (2018). Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. Eur. J. Hum. Genet. *26*, 740–744.
18. Lindor, N.M., Thibodeau, S.N., and Burke, W. (2017). Whole-genome sequencing in healthy people. Mayo Clin. Proc. *92*, 159–172.
19. Biesecker, L.G., and Green, R.C. (2014). Diagnostic clinical genome and exome sequencing. N. Engl. J. Med. *371*, 1170.
20. Linderman, M.D., Nielsen, D.E., and Green, R.C. (2016). Personal genome sequencing in ostensibly healthy individuals and the PeopleSeq Consortium. J. Pers. Med. *6*, 14.
21. McLaughlin, H.M., Ceyhan-Birsoy, O., Christensen, K.D., Kohane, I.S., Krier, J., Lane, W.J., Lautenbach, D., Lebo, M.S., Machini, K., MacRae, C.A., et al.; MedSeq Project (2014). A systematic approach to the reporting of medically relevant findings from whole genome sequencing. BMC Med. Genet. *15*, 134.
22. Vassy, J.L., Lautenbach, D.M., McLaughlin, H.M., Kong, S.W., Christensen, K.D., Krier, J., Kohane, I.S., Feuerman, L.Z., Blumenthal-Barby, J., Roberts, J.S., et al.; MedSeq Project (2014). The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. Trials *15*, 85.
23. Tsai, E.A., Shakbatyan, R., Evans, J., Rossetti, P., Graham, C., Sharma, H., Lin, C.F., and Lebo, M.S. (2016). Bioinformatics workflow for clinical whole genome sequencing at Partners HealthCare Personalized Medicine. J. Pers. Med. *6*, 12.
24. Zimmerman, R.S., Cox, S., Lakdawala, N.K., Cirino, A., Mancini-DiNardo, D., Clark, E., Leon, A., Duffy, E., White, E., Baxter, S., et al. (2010). A novel custom resequencing array for dilated cardiomyopathy. Genet. Med. *12*, 268–278.
25. Ceyhan-Birsoy, O., Machini, K., Lebo, M.S., Yu, T.W., Agrawal, P.B., Parad, R.B., Holm, I.A., McGuire, A., Green, R.C., Beggs, A.H., and Rehm, H.L. (2017). A curated gene list for reporting results of newborn genomic sequencing. Genet. Med. *19*, 809–818.
26. Duzkale, H., Shen, J., McLaughlin, H., Alfares, A., Kelly, M.A., Pugh, T.J., Funke, B.H., Rehm, H.L., and Lebo, M.S. (2013). A systematic approach to assessing the clinical significance of genetic variants. Clin. Genet. *84*, 453–463.
27. Kong, S.W., Lee, I.H., Leshchiner, I., Krier, J., Kraft, P., Rehm, H.L., Green, R.C., Kohane, I.S., MacRae, C.A.; and MedSeq Project (2015). Summarizing polygenic risks for complex diseases in a clinical whole-genome report. Genet. Med. *17*, 536–544.
28. Lane, W.J., Westhoff, C.M., Uy, J.M., Aguad, M., Smeland-Wagman, R., Kaufman, R.M., Rehm, H.L., Green, R.C., Silberstein, L.E.; and MedSeq Project (2016). Comprehensive red

blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. Transfusion *56*, 743–754.
29. Lane, W.J., Westhoff, C.M., Gleadall, N.S., Aguad, M., Smeland-Wagman, R., Vege, S., Simmons, D.P., Mah, H.H., Lebo, M.S., Walter, K., et al.; MedSeq Project (2018). Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. Lancet Haematol. *5*, e241–e251.
30. Aronson, S.J., Clark, E.H., Babb, L.J., Baxter, S., Farwell, L.M., Funke, B.H., Hernandez, A.L., Joshi, V.A., Lyon, E., Parthum, A.R., et al. (2011). The GeneInsight Suite: a platform to support laboratory and provider use of DNA-based genetic testing. Hum. Mutat. *32*, 532–536.
31. Aronson, S.J., Clark, E.H., Varugheese, M., Baxter, S., Babb, L.J., and Rehm, H.L. (2012). Communicating new knowledge on previously reported genetic variants. Genet. Med. *14*, 713–719.
32. Rehm, H.L. (2017). Evolving health care through personal genomics. Nat. Rev. Genet. *18*, 259–267.
33. Aronson, S., Mahanta, L., Ros, L.L., Clark, E., Babb, L., Oates, M., Rehm, H., and Lebo, M. (2016). Information technology support for clinical genetic testing within an academic medical center. J. Pers. Med. *6*, 4.
34. Aronson, S.J., and Rehm, H.L. (2015). Building the foundation for genomics in precision medicine. Nature *526*, 336–342.
35. Klinkenberg-Ramirez, S., Neri, P.M., Volk, L.A., Samaha, S.J., Newmark, L.P., Pollard, S., Varugheese, M., Baxter, S., Aronson, S.J., Rehm, H.L., and Bates, D.W. (2016). Evaluation: A qualitative pilot study of novel information technology infrastructure to communicate genetic variant updates. Appl. Clin. Inform. *7*, 461–476.
36. Almomani, R., Verhagen, J.M., Herkert, J.C., Brosens, E., van Spaendonck-Zwarts, K.Y., Asimaki, A., van der Zwaag, P.A., Frohn-Mulder, I.M., Bertoli-Avella, A.M., Boven, L.G., et al. (2016). Biallelic truncating mutations in ALPK3 cause severe pediatric cardiomyopathy. J. Am. Coll. Cardiol. *67*, 515–525.
37. Vassy, J.L., Christensen, K.D., Schonman, E.F., Blout, C.L., Robinson, J.O., Krier, J.B., Diamond, P.M., Lebo, M., Machini, K., Azzariti, D.R., et al.; MedSeq Project (2017). The impact of whole-genome sequencing on the primary care and outcomes of healthy adult patients: A pilot randomized trial. Ann. Intern. Med. *167*, 159–169.
38. Cirino, A.L., Lakdawala, N.K., McDonough, B., Conner, L., Adler, D., Weinfeld, M., O'Gara, P., Rehm, H.L., Machini, K., Lebo, M., et al.; MedSeq Project (2017). A comparison of whole genome sequencing to multigene panel testing in hypertrophic cardiomyopathy patients. Circ Cardiovasc Genet *10*, e001768.
39. Knöll, R., Postel, R., Wang, J., Krätzner, R., Hennecke, G., Vacaru, A.M., Vakeel, P., Schubert, C., Murthy, K., Rana, B.K., et al. (2007). Laminin-alpha4 and integrin-linked kinase mutations cause human cardiomyopathy via simultaneous defects in cardiomyocytes and endothelial cells. Circulation *116*, 515–525.
40. Van Driest, S.L., Wells, Q.S., Stallings, S., Bush, W.S., Gordon, A., Nickerson, D.A., Kim, J.H., Crosslin, D.R., Jarvik, G.P., Carrell, D.S., et al. (2016). Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. JAMA *315*, 47–57.
41. Safarova, M.S., Klee, E.W., Baudhuin, L.M., Winkler, E.M., Kluge, M.L., Bielinski, S.J., Olson, J.E., and Kullo, I.J. (2017). Variability in assigning pathogenicity to incidental findings: insights from LDLR sequence linked to the electronic health record in 1013 individuals. Eur. J. Hum. Genet. *25*, 410–415.
42. Dai, Z., Whitt, Z., Mighion, L.C., Pontoglio, A., Bean, L.J.H., Colombo, R., and Hegde, M. (2017). Caution in interpretation of disease causality for heterozygous loss-of-function variants in the MYH8 gene associated with autosomal dominant disorder. Eur. J. Med. Genet. *60*, 312–316.
43. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al.; 1000 Genomes Project Consortium (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.
44. MacArthur, D.G., and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. Hum. Mol. Genet. *19* (R2), R125–R130.
45. Abou Tayoun, A.N., Pesaran, T., DiStefano, M.T., Oza, A., Rehm, H.L., Biesecker, L.G., Harrison, S.M.; and ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum. Mutat. *39*, 1517–1524.

The American Journal of Human Genetics, Volume *105*

Supplemental Data

Analyzing and Reanalyzing the Genome:

Findings from the MedSeq Project

Kalotina Machini, Ozge Ceyhan-Birsoy, Danielle R. Azzariti, Himanshu Sharma, Peter Rossetti, Lisa Mahanta, Laura Hutchinson, Heather McLaughlin, The MedSeq Project, Robert C. Green, Matthew Lebo, and Heidi L. Rehm

Table S1 Pharmacogenomic results in the MedSeq cohort				
Interpretation Summary	Genotype	Cardiomyopathy	Primary care	Total (%)
Digoxin (Dysrhythmias, heart failure) ABCB1: c.3435T>C				
Increased metabolism and decreased serum concentration of digoxin	Homozygous_alt (CC)	16	16	**32**
Typical metabolism and serum concentration of digoxin	Heterozygous (TC)	22	27	**49**
Decreased metabolism and increased serum concentration of digoxin	Homozygous_ref (TT)	12	7	**19**
Metformin (Type 2 diabetes mellitus) C11orf65: c.175-5285G>A/T				
Increased glycemic response to metformin	Homozygous_alt (TT)	10	12	**22**
Typical glycemic response to metformin	Heterozygous (TG)	24	22	**46**
Decreased glycemic response to metformin	Homozygous_ref (GG)	16	16	**32**
Simvastin (Hyperlipidemia) SLCO1B1: c.521T>C				
Increased risk of simvastatin-related myopathy	Heterozygous (TC)	10	9	**19**
Typical risk of simvastatin-related myopathy	Homozygous_ref (TT)	40	41	**81**
Significantly increased risk of simvastatin-related myopathy	Homozygous_alt (CC)	0	0	**0**

Warfarin (Anti-coagulation) VKORC1: c.1639G>A, CYP2C9: c.[430C>T; 1075A>C]				
Decreased dose requirement	Heterozygous (GA), *2/*3 Homozygous_alt (AA), *1/*3 Homozygous_alt (AA), *2/*3 Homozygous_alt (AA), *2/*2	0 0 0 1	2 1 1 0	**5**
Standard dose requirement	Heterozygous (GA), *1/*3 Heterozygous (GA), *2/*1 Homozygous_alt (AA), *1/*1 Homozygous_alt (AA), *1/*2 Homozygous_ref (GG), *1/*3 Homozygous_ref (GG), *2/*2	2 6 5 1 3 1	2 1 10 4 2 0	**37**
Increased dose requirement	Heterozygous (GA), *1/*1 Homozygous_ref (GG), *1/*2 Homozygous_ref (GG), *1/*1	9 3 19	16 3 8	**58**
Clopidogrel (Anti-coagulation) CYP2C19: c.[-806C>T; 681G>A; 636G>A]				
Increased response to clopidogrel	*1/*17 *17/*17	17 3	11 3	**34**
Typical response to clopidogrel	*1/*1	18	23	**41**
Decreased response to clopidogrel	*1/*2 *2/*17	10 2	10 1	**23**
Significantly decreased response to clopidogrel	*2/*2	0	2	**2**