One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation

Stephen E. Lincoln1, Tina Hambuch2,3, Justin M. Zook4, Sara L. Bristow5, Kathryn Hatchell5, Rebecca Truty5, Michael Kennemer5, Brian H. Shirts3, Andrew Fellowes4, Shimul Chowdhury6, Eric W. Klee6, Shazia Mahamdallie7,10, Megan H. Cleveland2, Peter M. Vallone2, Yan Ding5, Sheila Seal7, Wasanthi DeSilva4, Farol L. Tomson8,11, Catherine Huang8, Russell K. Garlick8, Nazneen Rahman7, Marc Salit2,9, Stephen F. Kingsmore5, Matthew J. Ferber6, Swaroop Aradhya1 and Robert L. Nussbaum1

PURPOSE: To evaluate the impact of technically challenging variants on the implementation, validation, and diagnostic yield of commonly used clinical genetic tests. Such variants include large indels, small copy-number variants (CNVs), complex alterations, and variants in low-complexity or segmentally duplicated regions.

METHODS: An interlaboratory pilot study used synthetic specimens to assess detection of challenging variant types by various next-generation sequencing (NGS)-based workflows. One well-performing workflow was further validated and used in clinician-ordered testing of more than 450,000 patients.

RESULTS: In the interlaboratory study, only 2 of 13 challenging variants were detected by all 10 workflows. One well-performing workflow was further evaluated using additional synthetic, ordered testing of more than 450,000 patients.

CONCLUSION: The analytic and clinical sensitivity of NGS workflows can vary considerably, particularly for prevalent, statistically challenging variants. This can have important implications for the design and validation of tests (by laboratories) and the selection of tests (by clinicians) for a wide range of clinical indications.

Genetics in Medicine (2021) 23:1673–1680; https://doi.org/10.1038/s41436-021-01187-w

INTRODUCTION
Clinical genetic tests based on next-generation sequencing (NGS) are increasingly used to aid diagnosis and inform patient care. Specific guidelines are available for the clinical application of this technology and many NGS-based tests that detect single-nucleotide variants (SNVs) and small insertions and deletions (indels) in relatively accessible parts of patients’ genomes have been implemented. NGS has also been extended to detect copy-number variants (CNVs), also called del/dup events, although these methods are less uniformly implemented.

However, conventional short-read NGS methods have well-known limitations that can allow certain technically challenging variants—such as large indels, small CNVs, and complex alterations—to remain undetected, at least without the application of specialized bioinformatic and biochemical methodologies. Even simple SNVs and indels in segmentally duplicated (segdup) or low-complexity genomic regions can present substantial challenges.

The impact of such technically challenging variants on the implementation, validation, and diagnostic yield of commonly used clinical genetic tests has not yet been thoroughly described. In this study, we examined the spectrum of pathogenic germline variants uncovered during germline genetic testing of specific genes or multigene panels for hereditary cancer; cardiovascular, neurological, and pediatric disorders; reproductive carrier screening; and other clinical indications. To accomplish this, we first conducted an interlaboratory pilot study, evaluating ten different NGS laboratory workflows using synthetic positive controls containing variant types known to be challenging for NGS. One of these workflows was implemented and scaled up in a high-volume clinical testing laboratory, and the sensitivity of this workflow was further evaluated using additional synthetic, reference, and clinical specimens. Finally, we examined the attributes of pathogenic variants reported using this workflow in daily practice.

MATERIALS AND METHODS
Pilot study
Synthetic plasmids containing variants of interest were constructed as described previously by SeraCare (Gaithersburg, MD). These plasmids included pathogenic variants in seven hereditary cancer genes. These variants were previously observed in patients and presented specific technical challenges for NGS (Fig. 1, Table S1). The plasmids were titrated into genomic DNA from the Genome in a Bottle (GIAB) GM24385 (HG002) cell line at appropriate concentrations for the synthetic variants to appear heterozygous. These samples were provided to seven collaborating laboratories.

1Invitae, San Francisco, CA, USA. 2National Institute of Standards and Technology, Gaithersburg, MD, USA. 3Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA. 4Peter MacCallum Cancer Centre, Melbourne, Australia. 5Rady Children’s Institute for Genomic Medicine, San Diego, CA, USA. 6Mayo Clinic, Rochester, MN, USA. 7The Institute of Cancer Research, London, UK. 8SeraCare Life Sciences, Gaithersburg, MD, USA. 9Joint Initiative for Metrology in Biology, Stanford University, Stanford, CA, USA. 10Present address: Great Ormond Street Hospital for Children, London, UK. 11Present address: Meso Scale Diagnostics, Gaithersburg, MD, USA. ✉email: tina.hambuch@invitae.com
NGS methods for detecting a broad spectrum of genetic variation

In our sensitivity and prevalence studies, NGS was performed as described previously with improvements (Fig. 2). In brief, targeted-capture libraries were created using customized probes from Integrated DNA Technologies (Coralville, IA), Roche Sequencing Solutions (Pleasanton, CA), or Twist Bioscience (South San Francisco, CA). Extra probes were included for sites that otherwise had low coverage. Paired-end sequencing (2 × 150) was performed on the Illumina NovaSeq 6000 to at least 300× average depth per sample (typically much higher) and at least 50× at each targeted region. In these cases, heterozygous variants from either locus would appear at an average allele fraction of 25%. Examples include PMS2 (exons 12–14), NEB (exons 83–103), and SMN1/SMN2. When a potential pathogenic variant was suspected in these cases, a secondary assay using long-range polymerase chain reaction (LR-PCR) and long-read sequencing (Pacific Biosciences, Menlo Park, CA) was used as needed to determine in which locus a variant was present.

Variant in or near a homopolymer, STR, or low-complexity region. Variant may or may not be a simple length polymorphism.

- **Examples:**
  - MSH2:c.942+3A>T
  - SCN1A:c.3724_3725dup
  - PRSS1
  - DHCR7:

High sequence homology between genomic regions that prevent short NGS reads from mapping uniquely.

- **Examples:**
  - HBA2:c.427T>C
  - PMS2: Deletion exons 12-14

Large indels

- Insertion, deletion, duplication, or inversion sequence variants 15–49 bp in size.

- **Examples:**
  - CDK9N2A:c.9–32dup
  - DHCR7:c.385–412+5del

Complex rearrangements

- Insertion, deletion, duplication, or inversion sequence variants larger than 50 bp in size. Includes mobile insertion elements.

- **Examples:**
  - BRC42:c.3083, 3084insAlu
  - ABCB11: Complex rearrangement exons 23-28

Small CNVs

- A CNV spanning a single exon, part of one exon, or a single plus a partial exon.

- **Examples:**
  - DMD: Gain exon 22
  - FLCN: Deletion exon 14

Low-complexity repeat associated

- Variant in or near a homopolymer, STR, or low-complexity region. Variant may or may not be a simple length polymorphism.

- **Examples:**
  - MSH2:c.942+3A>T
  - SCN1A:c.3724_3725dup
  - PRSS1

Segdup/Pseudogene associated

- High sequence homology between genomic regions that prevent short NGS reads from mapping uniquely.

- **Examples:**
  - HBA2:c.427T>C
  - PMS2: Deletion exons 12-14

Putative mosaic

- CNVs or SNVs at low variant allele fraction that may or may not be germline mosaic or a result of clonal hematopoiesis.

- **Example:**
  - TP53c:626_626del

Fig. 1 Technically challenging variant types. Variants were categorized as being technically challenging, or not, based on these six criteria. Note that some variants could be considered challenging for multiple reasons (e.g., a single-exon deletion within a segmentally duplicated region). Examples provided are variants observed in the prevalence analysis. Detailed criteria are provided in the Supplemental Methods. CNV copy-number variant, indels, insertions or deletions, NGS next-generation sequencing, Segdup, segmental duplication, SNVs single-nucleotide variants, STR short tandem repeat.
RESULTS
Interlaboratory pilot study
An interlaboratory study was conducted both to reinforce our understanding of the impact of different NGS methodologies on challenging variant types, and to evaluate whether synthetic positive controls are a useful tool for the development and validation of methods to detect such variants. In this study, all 10 NGS workflows at collaborating laboratories were able to sequence and analyze the synthetic control mixtures, demonstrating compatibility of the synthetic approach with various NGS biochemistries. However, only 2 of the 13 challenging variants (as defined in Fig. 1) were detected by all 10 workflows, and just three workflows detected all 13 (Tables 1, S1, S2). Additionally, 3 of the 11 other, less-challenging indels were missed by some workflows.

Manuel review using IGV demonstrated that evidence of the missed variants was visible in most of the raw data sets, indicating that the sensitivity limitations were largely bioinformatic in nature. IGV review of data from synthetic controls and patient specimens demonstrated compatibility of the synthetic approach with various NGS biochemistries. However, only 2 of the 13 challenging variants (as defined in Fig. 1) were detected by all 10 workflows, and just three workflows detected all 13 (Tables 1, S1, S2). Additionally, 3 of the 11 other, less-challenging indels were missed by some workflows.

Prevalence analysis
A consecutive series of patients receiving gene panel testing by Invitae (San Francisco, CA) between June 2018 and March 2020 was retrospectively analyzed. Exome sequencing based tests were excluded owing to methodological differences with panel testing. The specific genes analyzed for each patient were a subset of those assayed, chosen by ordering clinicians. Variants were clinically classified using Sherloc, a framework based on the American College of Medical Genetics and Genomics and Association for Molecular Pathology guidelines. 23,24 Only pathogenic and likely pathogenic variants were included in our prevalence study—variants of uncertain significance and benign variants were excluded. Variants were categorized as technically challenging, or not, using specific criteria (Fig. 1, Supplemental Methods).

Table 1. Proof of concept and sensitivity study results.

| Study                  | Interlaboratory pilot study | Sensitivity study |
|------------------------|-----------------------------|------------------|
| WorkFlow               | 1A                          | 1B               |
| Sequencing             | IPE                         | IPE              |
| Targeting              | Hyb                         | Hyb              |
| Informatics            | CS                          | CS               |
| Samples                | Syn (n = 2)                 | Syn (n = 2)      |
| SNVs                   | 15/15                       | 15/15            |
| Short Indel            | 11/11                       | 11/11            |
| Large Indel            | 4/4                         | 4/4              |
| Segdup-associated      | 3/3                         | 3/3              |
| Low complexity         | 4/4                         | 4/4              |
| Complex and small CNV  | 2/2                         | N/A              |
| CNV ≥2 exon            | N/A                         | N/A              |

Amp, amplicon sequencing; CNV, copy-number variant; CS, custom software; EV, software provided by the sequencing equipment vendor; GIAB, Genome in a Bottle; GS, genome sequencing; Hyb, hybridization capture; IPE, Illumina Paired-end; indel, insertion or deletion; Ion, Ion Torrent; N/A, not applicable; NGS, next-generation sequencing; REF, reference specimens from public biobanks; Segdup, segmental duplication; SNVs, single-nucleotide variants; Syn, synthetic controls; TP, third-party software.

*For the interlaboratory pilot study, the performance of 10 NGS workflows used by seven collaborating laboratories is shown for variants in two synthetic control specimens. In each data cell, the denominator is the number of variants within each assay's target regions and the numerator is the number of these variants that were detected. Normal font indicates 100% observed sensitivity. Bold font indicates an observed limitation. Italics indicate that no study variants were present in regions interrogated by the assay. Details of each of the 10 workflows and the variants are provided in Tables S1 and S2. All workflows included bioinformatics methods to detect SNVs and small indels, and half (workflows 1A, 1B, 2, 6, and 7B) included additional methods to improve sensitivity for large indels and complex variants. SNVs were not included in this study. Workflow 1A had previously detected all of these variants in patients and was included primarily to validate the construction of the synthetic controls and to allow the comparison of synthetic and patient specimen data for the same variants (Figure S1). Workflow 1B corresponds to Fig. 2 and was an evolution of 1A, albeit with substantial differences in the variant calling algorithms used (Table S2, Fig. 2).

For the sensitivity study, performance is shown for variants in samples from each source. Positive controls with technically challenging variants were difficult to obtain, requiring a large number of reference specimens and additional synthetic controls. A list of samples used in this study is provided in Table S5.
targeted indels were false negatives because (1) the variant altered a PCR primer binding site, (2) the variant was near an amplicon boundary, interfering with alignment, or (3) the variant caused a substantial increase in amplicon size, which the biochemistry could not accommodate. This sequencing platform also exhibited its characteristic limitation with both of the biochemistry could not accommodate. This sequencing platform then applied. Follow-up assays are used to confirm potential false positives, to determine the exact sequence of complex variants, and to resolve the location of variants within segmental duplications. For details see "Materials and Methods." CNVs copy-number variants, GATK Genome Analysis Toolkit, Indels insertions or deletions, LR-PCR long-range polymerase chain reaction, MEIs mobile element insertions, NGS next-generation sequencing, QC quality control, Segdup segmental duplication, SNVs single-nucleotide variants.

Fig. 2 Variant calling process used in prevalence study. A next-generation sequencing (NGS) workflow designed to detect a wide variety of variant types was used in our prevalence study, sensitivity study, and also our pilot study (workflow 1B). NGS reads are aligned to a modified reference genome and multiple variant callers are then applied. Follow-up assays are used to confirm potential false positives, to determine the exact sequence of complex variants, and to resolve the location of variants within segmental duplications. For details see "Materials and Methods." CNVs copy-number variants, GATK Genome Analysis Toolkit, Indels insertions or deletions, LR-PCR long-range polymerase chain reaction, MEIs mobile element insertions, NGS next-generation sequencing, QC quality control, Segdup segmental duplication, SNVs single-nucleotide variants.

Sensitivity study
One NGS workflow (Fig. 2) from the pilot study was used in our prevalence analysis (below). Its sensitivity was further evaluated using a methods-based approach23-4,26 rather than a gene- or variant-based approach, which was the only practical option considering the large number of genes and variants targeted. In such studies, positive control specimens containing a diversity of variants are obtained, and the ability to detect these variants is measured by class. This particular study utilized 94 specimens containing 601 independently characterized positive control variants in 47 genes of interest. All 601 were correctly detected, demonstrating 100% observed sensitivity (Table 1). We found that the manner in which the various specimen types contributed to this study varied considerably, with significant implications for the evaluation of methods to detect challenging variants which we elaborate on here.

The seven GIAB samples, for example, contributed most (470/601, 78.2%) of the study variants, although they had limited clinical or methodological relevance. The vast majority (92%) were SNVs, while pathogenic variants in these 47 genes are often indels or CNVs.27 Moreover, none of the 36 GIAB indels were greater than 5 base pairs (bp) in size, and many challenging genomic locations (e.g., the pseudogene-associated exons of PMS2) did not have any high-confidence calls in the GIAB 3.3.2 data (newer GIAB data sets may improve this particular limitation, however, as shown below). The GIAB samples do contain additional variant types (i.e., CNVs and structural variants),28 but these were not located in or near our targeted genes and were not useful for measuring sensitivity of this assay. Indeed, none of the 470 GIAB variants met our definition of technically challenging.

To increase the number of clinically important variant types, we included 58 additional reference samples (Table S5) and 26 clinical specimens. Unlike GIAB, each of these contributed only one or two independently characterized variants to our study. Nevertheless, this set added 60 indels, 10 of which were larger than 5 bp, and 9 CNVs. Most importantly, it provided 14 variants that met our definition of technically challenging (Fig. 1).

However improved, these variant counts remained small, particularly for the diverse challenging variant subtypes. We thus added synthetic controls to our study, which have the advantage of including multiple variants of interest in each DNA sample. Just three specimens added another 18 heterogeneous, technically challenging variants (more than half of the 32 total), and also added 23 additional indels (Tables S1, S6). Importantly, most of these variants were unique, in contrast to variants in the reference and GIAB samples that were often (21% and 74%, respectively) repeated in multiple specimens (repeated variants may be more useful in demonstrating reproducibility compared to sensitivity). Multiple gene panels using different hybridization assays, but an otherwise common workflow, were developed, validated, and used in our prevalence study, below.

Prevalence of technically challenging variants
In our cohort of 471,591 patients meeting study criteria, 102,085 (21.6%) carried one or more clinically reported pathogenic or likely pathogenic (P/LP) variants in 1,217 distinct genes. This positive rate was expected given the mix of patients’ clinical indications and was not performed. A total of 127,710 P/LP variants were reported, of which indels comprised 31.4%, CNVs 9.7%, and SNVs 58.9%. These variants were confirmed as needed21 and were thus all confidently considered true positives.

Technically challenging variants were prevalent. Of the 127,710 P/LP findings, 17,561 (13.8%; 95% CI 13.6-13.9%) met one or more of our criteria for being technically challenging (Fig. 1). These challenging variants were uncovered in 16,618 patients (i.e., some patients carried more than one) and in 556 genes (46% of those genes with any P/LP findings). Technically
Challenging variants were observed across clinical areas. CNV, copy-number variant; Indel, insertion or deletion. Preventive testing is technically challenging. A total of 102,085 patients with P/LP variants in 1,217 genes are represented in this data set. Two fractions result from a small number of relatively common P/LP variants that are (e.g., in carrier or neurology testing) or are not (e.g., preventive testing) technically challenging. A total of 102,085 patients with P/LP variants in 1,217 genes are represented in this data set. Challenging variants of most types were observed across clinical areas. CNV copy-number variant, Indel insertion or deletion.

Of the 1,194 P/LP variants in MSH2, 20.9% were observed in the neurology test. Other segdup regions, including MYLK, underlying nemaline myopathy, 7.7% of P/LP variants were in the triplicated exons. PMS2, accounted for 22.2% (28,351) of all P/LP findings, for example, and 34% of these (9,683) were considered challenging. These prevalent findings included high, moderate, and low penetrance alleles, and both dominant and recessive modes of inheritance were represented. Excluding these 11 sites, the prevalence of challenging variants remained high (7.9%). Consideration of the rarest findings in our cohort, 18,856 P/LP variants were observed in only a single individual, and 9.2% (2,434) of these were considered challenging. Thus, both rare and relatively common technically challenging variants were often observed.

No single attribute defined all or even most of the technically challenging variants we observed. Rather, a broad spectrum was present. Of the challenging P/LP variants, 42.3% (7,423) were located in low-complexity regions (e.g., homopolymers, short tandem repeats) and 35.0% (6,153) were in segmental duplications (segdups). In addition, 11.4% (1,995) were small CNVs, 6.5% (1,135) large indels, and 1.4% (238) complex rearrangements. Finally, 0.6% of variants (740) were flagged as potentially mosaic based on having an abnormally low NGS allele fraction. Note that not all of these variants were, in fact, mosaic: some may be a result of clonal hematopoiesis, some were apparently within CNVs, but all may warrant investigation. Some variants (118) fell into more than one category (e.g., large indel within a segdup).

A sizable fraction of the low-complexity variants (5,254, 70.8%) were alterations at the CFTR intronic poly-T/poly-TG site, which, depending on diplotype, confer moderate risk for pancreatitis, respiratory disease, and male infertility. Other low-complexity variants were uncovered in 233 different genes, comprising 1.7% of all P/LP findings. Some were particularly challenging for conventional NGS. For example, 91 confirmed findings of MSH2 variant NM_000251.3:c.942+3A>T were observed, which is not a homopolymer length change, but rather an SNV at the end of a 25-bp homopolymer. This single, high-penetrance, pathogenic splicing variant made up 11.0% of all P/LP findings in MSH2, a gene conferring cancer risk (Lynch syndrome) as well as response to certain immuno-oncology (IO) drugs.

An additional 185 premutation and full mutation alleles were observed in FMR1, underlying fragile X syndrome, but were not included in the counts above, owing to methodological differences.

The most common (5,457) findings of P/LP SNVs, indels, and CNVs within segmentally duplicated genes were observed in SMN1/2, GBA, and HBA1/2. All were tested in carrier screening, with SMN1/2 also included in neurology tests. Other segdup regions, including NEX (exons 83–103), PMS2 (exons 12–15), PRSS1, and SDHA, accounted for 358 additional findings within the hereditary cancer, neurology, and pediatric indications. For instance, PMS2 (like MSH2) is involved in Lynch syndrome and IO response, and 20.9% of all 1,194 P/LP findings were located in the four pseudogene-associated exons. In NEB, underlying nemaline myopathy, 7.7% of P/LP variants were in the triplicated exons.

Large indels, small CNVs, and complex rearrangements collectively represented 3,366 P/LP findings, 6.4% of all non-SNPs, affecting 38% of genes and every clinical area (Fig. 4a). More than half (1,836) of these were deletions between 50 bp and one exon in size. Whether such events were considered CNVs or indels was, in practice, defined more by methodology than biology. Mobile elements, sometimes called “jumping genes,” accounted for 128 findings, 58 of which were observed in only a single individual.

Comparisons with public data sets
As expected, most of the P/LP variants we observed were rare, and thus absent from population databases including gnomAD (data not shown), although some of these absences are explained by methodological differences in variant detection between gnomAD and our data. Nevertheless, we examined the gnomAD version 2.1.1 exome sequences as a representative, if heterogeneous, view
of the coverage that exome capture may achieve at the locations of P/LP variants in our study. Although the average coverage among the 125,748 gnomAD exome sequences at these sites is 76x, our P/LP variants had a 5.1% chance of having no coverage in a gnomAD exome and a 10.1% chance of having <20x coverage (Fig. 4b). Even if doubled from 76x to ~150x on average, more typical of clinical exome sequencing, this coverage would likely remain inadequate to detect many of our challenging P/LP variants.

We similarly compared our variant sites with the GIAB benchmark regions for all seven GIAB samples and found that 15.1% of variants were outside of these regions in at least one of the seven, and 9.7% were outside in all seven using version 3.3.2 GIAB data sets. Many (9.7%) of these variants were outside of the benchmark regions in all seven GIAB samples (“Not Any” category) and 15.1% of these variants were outside of these regions in at least one of the seven samples (“Not all”). However, the newer version 4.1 GIAB data, available only for one of the GIAB samples at this time, substantially improves this situation. CNVs were not included in this analysis.

**DISCUSSION**

The comprehensive assessment of pathogenic variants is crucial to diagnosis of hereditary disorders and clinical decision making. Positive findings often suggest specific treatment or management actions, and equally the lack of findings can suggest a different course. Thus, both high sensitivity and its corollary, high negative predictive value (NPV), are valued in genetic testing. As a historical example, molecular testing for severe hemophilia A had limited utility until the discovery of F8 inversions. Subsequently, sensitive tests became available not only for affected patients, but also family counseling, carrier screening, and prenatal diagnosis.

To design and validate (as a lab), or select (as a clinician), the most appropriate tests for any gene or condition, understanding the spectrum of pathogenic variation is important. Our study found that pathogenic variants of technically challenging types for NGS are prevalent across many genes in a large population of patients with diverse clinical indications. These challenging types are heterogeneous and include large indels, small CNVs, mobile element insertions, complex rearrangements, as well as variants within segmental duplications or low-complexity regions. No current NGS platform or current variant calling algorithm can capture all of these variant types. Rather, we found that a battery of algorithms was required and that supplementing short-read NGS with long-read sequencing allowed the resolution of otherwise ambiguous variant calls. Implementing such methods in clinical practice can, however, be complex and require significant expertise and time. Furthermore, we found that NGS methods achieving high sensitivity for some of these variant types also have poor specificity, making confirmatory assays mandatory in some cases, a topic we have detailed separately. Of course, “challenging” is a relative term, and the specific definition we used (Fig. 1) may be too conservative for some laboratory methods, and too permissive for others.

The prevalence rates we observed for challenging variants are undoubtedly underestimates, as we know that the NGS workflow used in our prevalence study cannot achieve perfect sensitivity for
all variants of all types, despite the validation study results described above. Nor have all of the challenging genes or regions relevant to the clinical areas we studied been implemented into this workflow yet, some of which will require additional methodologies (as did FMRI). Furthermore, there are many clinically relevant genes other than the 1,217 described herein, and noncoding regions of clinical impact continue to be uncovered. Further study of pathogenic variation is clearly needed.

Our interlaboratory pilot study highlighted the fact that the sensitivity of different NGS workflows can vary, particularly based on the bioinformatics pipeline used and the types of variants that the pipeline is designed and validated to detect. Off-the-shelf bioinformatics solutions fared least well in this comparison (Table 1), reinforcing the important role of clinical lab directors in thoroughly evaluating tools they use. Indeed, many of the NGS workflows used in this study have improved since, based, in part, on observations from this pilot study, demonstrating the value of comparative benchmarking exercises. While our studies focused on panel testing, many of the issues we identified would equally apply to exome or genome sequencing. Unfortunately, we observed that our study’s pathogenic variants may receive low (or no) coverage in some exome sequencing workflows, and coverage variability in exome sequencing may also affect sensitivity (particularly for small CNVs). Carefully evaluating the coverage and limitations of any NGS methodology for each specific application remains vital.

A recent systematic review of validation studies by Roy et al. found “a clear absence of uniformity” in study design, where (among other issues) highly variable numbers of samples and positive control variants were used to make sensitivity claims. Moreover, some of these studies combined large numbers of SNVs with very few indels, presenting a biased measurement of overall sensitivity. Our results show that this is often inappropriate, given that indel sensitivity can vary greatly among workflows for all but the smallest and simplest indels, and that indels and/or CNVs, including very challenging examples, can be a substantial fraction of pathogenic variants in many genes. As recommended by guidelines, our results demonstrate that reporting analytical sensitivity and statistical confidence by variant type (e.g., SNV, indel, CNV) is critical, as is a breakdown of validation data by variant size and genomic context (e.g., low-complexity and segdup regions). Moreover, because high quality genetic tests should detect clinically prevalent variants and variant types, steps to ensure that validation studies include a representative set of variants should be clearly documented and sensitivity claims linked to what was demonstrated during validation. When a single-sensitivity measurement is required (e.g., in calculating clinical sensitivity) then we believe that it would best be computed on a prevalence-weighted basis, using the known distribution of pathogenic variant types, including the challenging types described herein. Otherwise, sensitivity estimates may be misleading.

Thoroughly validating NGS workflows can be difficult owing to the paucity of readily obtained positive controls containing nontrivial variants in coding regions of clinically relevant genes. This partially explains the overreliance on SNVs for test validation that Roy et al. observed. In the sensitivity study described herein, reference samples from public biobanks complemented the GIAB resources, although many such samples were required. Clinical specimens with independent data were also informative, although these are not replenishable—new specimens would need to be obtained to revalidate processes. We thus developed a set of replenishable synthetic controls containing diverse challenging variants. These accurately mimicked endogenous variants, showed high utility in our interlaboratory study, and greatly improved the breadth of our sensitivity study. Complementary approaches, such as in silico validation, are promising for the future, although ongoing development is required to ensure that these approaches capture all of the complications in NGS data (e.g., Figure S1) particularly for challenging variant types. New validation standards and regulatory science advances will be needed to continue the rapid evolution of clinical genetic testing.

In summary, our results demonstrate that clinically significant but technically challenging variants are prevalent in genes associated with a wide range of clinical indications. As is the case with most genetic diseases, these variants are diverse and often individually rare, but collectively common. Limitations in their detection could lead to appreciable false negative rates. Testing for these variant types is, however, not uniformly implemented, and some validation studies emphasize simpler (and less clinically relevant) variant types. The resources and approaches described herein may help laboratories and clinicians optimize genetic testing to further improve patient care.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon request.

Received: 8 July 2020; Revised: 8 April 2021; Accepted: 11 April 2021; Published online: 18 May 2021

REFERENCES

1. Korf, B. R. & Rehm, H. L. New approaches to molecular diagnosis. JAMA. 309, 1511–1521 (2013).
2. Rehm, H. L. et al. ACMG clinical laboratory standards for next-generation sequencing. Genet. Med. 15, 733–747 (2013).
3. Roy, S. et al. 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 (2018).
4. Santani, A. et al. Designing and implementing NGS tests for inherited disorders: a practical framework with step-by-step guidance for clinical laboratories. J. Mol. Diagn. 21, 369–374 (2019).
5. Nord, A. S., Lee, M., King, M.-C. & Walsh, T. Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genomics. 12, 184 (2011).
6. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 14, S1 (2013).
7. Lincoln, S. E. et al. A systematic comparison of traditional and multigene panel testing for hereditary breast and ovarian cancer genes in more than 1000 patients. J. Mol. Diagn. 17, 533–544 (2015).
8. Truty, R. et al. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. Genet. Med. 21, 114–123 (2019).
9. Telenius, A. et al. Deep sequencing of 10,000 human genomes. Proc. Natl. Acad. Sci. U.S.A. 113, 11901–11906 (2016).
10. Eichler, E. E. Genetic variation, comparative genomics, and the diagnosis of disease. N. Engl. J. Med. 381, 64–74 (2019).
11. Torene, R. I. et al. Mobile element insertion detection in 89,874 clinical exomes. Genet. Med. 22, 974–978 (2020).
12. Goldfelder, R. L. et al. Medical implications of technical accuracy in genome sequencing. Genome Med. 8, 24 (2016).
13. Mandelker, D. et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genet. Med. 18, 1282–1289 (2016).
14. Herman, D. S. et al. Efficient detection of copy number mutations in PMS2 exons with a close homolog. J. Mol. Diagn. 20, 512–521 (2018).
15. Ebbert, M. T. W. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. 20, 97 (2019).
16. Kudalkar, E. M. et al. Multiplexed reference materials as controls for diagnostic next-generation sequencing: a pilot investigating applications for hypertrophic cardiomyopathy. J. Mol. Diagn. 18, 882–889 (2016).
17. Zook, J. M. et al. An open resource for accurately benchmarking small variant and structural variations from targeted high-throughput sequence data. BMC Genomics. 12, S1 (2013).
18. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).
19. Church, D. M. et al. Modernizing reference genome assemblies. PLoS Biol. 9, e1001091 (2011).
20. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
AUTHOR CONTRIBUTIONS
Conceptualization: S.E.L., T.H., R.K.G., R.L.N. Data curation: S.E.L., K.H., R.T. Formal analysis: S.E.L., J.M.Z., K.H., R.T., M.K. Investigation: J.M.Z., B.H.S., A.F., S.C., E.W.K., S.M., M.H.C., P.M.V., Y.D., S.S., W.D., N.R., M.S., F.K., J.F., S.A., R.L.N.

ETHICS DECLARATION
Our study protocol was approved by the Western Institutional Review Board. Reference specimens were used under the terms of their respective material transfer agreements.

COMPETING INTERESTS
S.E.L., S.L.B., R.L.N., S.A., K.H., T.H., R.T., M.K. are employees of and own stock and/or stock options in Invitae. S.E.L. also owns stock in Illumina, Thermo Fisher, and Pacific Biosciences. R.L.N. owns stock in Maze Therapeutics, Genome Medical and Personalis, and consults for Pfizer Pharmaceuticals. R.K.G. is an employee of and owns stock in SeraCare. C.H. is an employee of SeraCare. F.L.T. is a former employee of SeraCare and current employee of Meso Scale Diagnostics. E.W.K. receives royalties from SoftGenetics. M.F. receives royalties from SoftGenetics and serves as a consultant to OneOme. N.R. serves as a Non-Executive Director of AstraZeneca. The other authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41436-021-01187-w.

Correspondence and requests for materials should be addressed to T.H.

Reprints and permission information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.