Received: 6 January 2022 | Revised: 23 September 2022 | Accepted: 30 September 2022

DOI: 10.1002/mgg3.2072

ORIGINAL ARTICLE

Molecular Genetics & Genomic Medicine Open Access WILEY

Scalable detection of technically challenging variants through modified next-generation sequencing

Susan Rojahn | Tina Hambuch | Jessika Adrian | Erik Gafni | Alex Gileta | Hannah Hatchell | Britt Johnson | Ben Kallman | Kate Karfilis | Curtis Kautzer | Michael Kennemer | Lloyd Kirk | Daniel Kvitek | Jessica Lettes | Fenner Macrae | Fernando Mendez | Joshua Paul | Maurizio Pellegrino | Ronny Preciado | Jan Risinger | Matthew Schultz | Lindsay Spurka | Sajani Swamy | Rebecca Truty | Nathan Usem | Andrea Velenich | Swaroop Aradhya

Invitae, San Francisco, California, USA

Correspondence
Swaroop Aradhya, Invitae, 1400 16th
Street, San Francisco, CA 94103, USA.
Email: swaroop.aradhya@invitae.com

Abstract

Background: Some clinically important genetic variants are not easily evaluated with next-generation sequencing (NGS) methods due to technical challenges arising from high- similarity copies (e.g., *PMS2*, *SMN1/SMN2*, *GBA1*, *HBA1/HBA2*, *CYP21A2*), repetitive short sequences (e.g., *ARX* polyalanine repeats, *FMR1* AGG interruptions in CGG repeats, *CFTR* poly-T/TG repeats), and other complexities (e.g., *MSH2* Boland inversions).

Methods: We customized our NGS processes to detect the technically challenging variants mentioned above with adaptations including target enrichment and bioinformatic masking of similar sequences. Adaptations were validated with samples of known genotypes.

Results: Our adaptations provided high-sensitivity and high-specificity detection for most of the variants and provided a high-sensitivity primary assay to be followed with orthogonal disambiguation for the others. The sensitivity of the NGS adaptations was 100% for all of the technically challenging variants. Specificity was 100% for those in *PMS2*, *GBA1*, *SMN1/SMN2*, and *HBA1/HBA2*, and for the *MSH2* Boland inversion; 97.8%–100% for *CYP21A2* variants; and 85.7% for *ARX* polyalanine repeats.

Conclusions: NGS assays can detect technically challenging variants when chemistries and bioinformatics are jointly refined. The adaptations described support a scalable, cost-effective path to identifying all clinically relevant variants within a single sample.

KEYWORDS

bioinformatics, genetic testing, molecular genetics, next-generation sequencing

[Correction added on November 19, 2022 after first online publication. The figure 2 has been updated in the article.]

1 | INTRODUCTION

Targeted next-generation sequencing (NGS) is routinely used by clinical laboratories to analyze hundreds of disease-related genes in a single test with high accuracy, affordability, and speed. Still, most standard gene-targeted NGS workflows based on short-read chemistry cannot reliably assess certain clinically important variants due to limitations in assay chemistry, sample-to-sample variability, or bioinformatic processes. Classes of such technically challenging variants include single-exon copy number variants (CNVs) (i.e., intragenic deletions and duplications) (Abou Tayoun et al., 2016; Mandelker et al., 2016; Truty et al., 2019) and variants within repetitive sequence tracts (Li & Freudenberg, 2014), regions of low-complexity, or genes that have highly similar copies of their sequences elsewhere (e.g., pseudogenes, paralogs) (Mandelker et al., 2016). These technically challenging variants need to be reliably assessed, as recent estimates suggest they represent at least 14% of clinically significant genetic test results (Lincoln et al., 2021).

Many labs use ancillary non-NGS assays to address these regions, but these methods are often laborious, costly, and thus not well suited for large-scale and cost-effective clinical testing. Further, because multiple assay types are needed to detect all categories of technically challenging variants, referrals to multiple testing labs are sometimes needed to ensure that all relevant variants can be detected in an individual. Although NGS has afforded patients and healthcare providers with significantly better genetic testing opportunities, the capabilities and limitations of short-read NGS technologies must be carefully considered. A primary concern is that most clinical NGS platforms use Illumina instruments and currently produce only contiguous raw sequence data in short segments typically 150 bp–300 bp long. This is problematic when tackling repetitive or duplicated sequences, as long stretches of sequence may have nearly 100% identity with other regions in the genome, hindering accurate short-read alignment and subsequent variant calling (Aziz et al., 2015; Li & Freudenberg, 2014; Mandelker et al., 2016; Rehm et al., 2013). Although alternative sequencing technologies are capable of producing much longer reads (e.g., ~10,000 bp long), these methods are not currently in routine and high-throughput use (Mantere et al., 2019). Another consideration for NGS-based genetic testing is that even when labs use the same NGS technology, many variables can affect test performance, such as differences in personnel training and procedures (Lincoln et al., 2021) or in assay design (and thus assay capabilities). To optimize test performance, bioinformatic analyses must also be customized for each assay. For example, stringently controlled, highly automated sample processing enables reliable high-resolution NGS-based detection of single-exon CNVs by read-depth comparisons instead of requiring the separate use of exon-focused array comparative genomic hybridization (Chong et al., 2014; Johansson et al., 2016; Lincoln et al., 2015; Truty et al., 2019). Similar advances in laboratory setups and controls can unlock other useful NGS capabilities that may not be feasible with generic off-the-shelf workflows.

By refining our targeted NGS workflow from sample accessioning through bioinformatic analysis, we have incorporated testing for many disease genes and specific variants that are typically not interrogated with standard NGS assays. Here, we describe the design and validation of our adapted NGS-based approaches for assessing technically challenging variants in five genes or gene regions that have segmentally duplicated copies with very high sequence similarity (i.e., *PMS2* [OMIM: *600259], *SMN1* [OMIM: *600354], *GBA1* [OMIM: *606463], *HBA1/HBA2* [OMIM: *141800/OMIM: *141850], *CYP21A2* [OMIM: *613815]), in three genes affected by repetitive or low-complexity sequence tracts (i.e., AGG interruptions within the CGG repeats of *FMR1* [OMIM: *309550], polyalanine repeats in exon 2 of *ARX* [OMIM: *300382], and poly-T/TG repeats in *CFTR* [OMIM: *602421]), and in one gene associated with a recurrent copy-number neutral structural change known as a Boland inversion (i.e., *MSH2* [OMIM: *609309]).

2 | METHODS

2.1 | Editorial compliance

This study presents the methodology and validation of adaptations to an NGS-based genetic testing workflow and presents concordance of testing outcomes using a combination of commercial and non-commercial reference materials. All assays and validations complied with Clinical Laboratory Improvement Amendments (CLIA) regulatory guidance as well as professional guidelines and best practice standards for validations (Aziz et al., 2015; Centers for Disease Control and Prevention, 2021).

2.2 | NGS chemistry and bioinformatics

Genomic DNA was isolated from whole blood or saliva by magnetic bead extraction on a Hamilton Microlab STAR liquid handling system (Norcross, Georgia, United States). DNA libraries were prepared by shearing the isolated genomic DNA to an average fragment length of 350 bp, followed by end repair and addition of adapters for paired-end sequencing. Hybridization capture of disease genes was performed using an iteratively optimized pool of oligonucleotide baits (Roche, Pleasanton, CA; Integrated DNA Technologies, Coralville, IA; Twist Bioscience, South San Francisco, CA) targeted to exons, +/− 10–20 bases of flanking intronic sequences, and certain noncoding regions of clinical interest (Lincoln et al., 2015). Samples were sequenced to an average of 350X depth-of-sequence read coverage (and a minimum of >50X for 99% of genes) using HiSeq, NovaSeq, or NextSeq instruments (Illumina, San Diego, CA).

In the bioinformatics pipeline, both community standard and custom algorithms were used to identify single-nucleotide variants (SNVs), small and large insertions and deletions (indels), structural variants with breakpoints in target sequences, and exon-level CNVs (Kurian et al., 2014; Lincoln et al., 2015). High-quality sequence reads were selected according to Picard metrics and aligned with NovoAlign (Novocraft Technologies, Selangor, Malaysia) to a customized version of the GRCh37 reference genome (Church et al., 2011). The GenBank reference sequence for the genes described in this report are as follows: NG_008466.1 (*PMS2*); NC_000007.14 (*PMS2CL*); NG_008691.1 (*SMN1*), NG_008728.1 (*SMN2*), NG_009783.1 (*GBA1*); NG_000006.1 and NG_059186.1 (*HBA1*), NG_000006.1 and NG_059271.1 (*HBA2*); NG_007941.3 (*CYP21A2*); NG_008281.1 (*ARX*); NG_016465.4 (*CFTR*); NG_007529.2 (*FMR1*); and NG_007110.2 (*MSH2*). For genes of clinical interest with nearly identical extra sequence copies (e.g., *PMS2*, *SMN1, HBA1/HBA2*), the paralogous regions were "masked" in the reference genome (converted into Ns or otherwise obscured) to force NGS reads to map to a single location (Figure 1). Sequence variants (i.e., SNVs and indels) were first identified by the Genome Analysis Toolkit (GATK) HaplotypeCaller (Poplin et al., 2017). In *PMS2* and *SMN1*, each of which have highly similar gene copies, SNVs were called by Freebayes to account for copy number >2 caused by reads from multiple loci aligning to a single locus in the reference genome as a result of masking (Garrison & Marth, 2012). For *CYP21A2*, we adjusted the allele balance threshold in our variant caller.

Four custom algorithms were developed and used to support the detection of specific targeted variants or classes of variants. First was a custom-developed CNV caller (CNVitae) that identified single-exon and larger CNVs with a read count-based method. A baseline of read counts associated with normal copy number was set within each sequencing run. Using a statistical mixture model of read counts within target regions (typically exons), the most likely copy number of each segment was estimated. Each called segment was then assigned a quality score indicating the degree of confidence in the copy number determination. CNVitae was initially validated against similar algorithms (publicly available tools) and has been prospectively validated with improvements over time. Second was a custom split-read detection algorithm developed to detect the precise breakpoints of inversions, copy-number changes, large indels, and genomic rearrangements. It was used to identify and score loci with an accumulation of split-read signals above the baseline level. (A split read, also known as a soft-clipped read, partially maps to two *cis* locations: one partially downstream and the other partially upstream of a breakpoint.) Third was a custom variant caller (Coalgen) that identified variants embedded within simple repeat sequences (e.g., homopolymers). Reads aligned to a locus known to harbor simple repeats were computationally extracted and then separately aligned to known haplotypes. The haplotype combination with the maximum likelihood was identified as the genotype. A fourth custom algorithm was used to infer the number and position of AGG interruptions within CGG-repeat alleles in the gene *FMR1*.

2.3 | Disambiguation and orthogonal confirmation

Due to reference genome masking, some variants were called within the primary NGS sequencing data in an ambiguous state, meaning there was uncertainty as to whether they were located within the disease gene or its paralogous copy. These ambiguous variants were then disambiguated (i.e., assigned to their true location) with orthogonal methods including successive long-range PCR (LR-PCR) and nested PCR reactions coupled with long-read sequencing (Pacific Biosciences, Menlo Park, CA, USA), Sanger sequencing, multiplex ligation-dependent probe amplification (MLPA) (MRC Holland, Amsterdam, Netherlands), and MLPA-seq (i.e., an internally developed adaptation of conventional MLPA in which MLPA ligation products are combined and analyzed by NGS instead of capillary electrophoresis).

2.4 | Validation methods

Reference samples with known variant status as determined by conventional orthogonal methods were sourced from the Coriell Institute (Camden, New Jersey), the Associated Regional and University Pathologists (Salt Lake City, Utah), and internally. Features of individual validation samples, including source and known genotype, are provided in the Supporting Information (Tables S1–S11). Validations were guided by recommendations by the American College of Medical Genetics and Genomics (ACMG) and the College of American Pathologists (Aziz

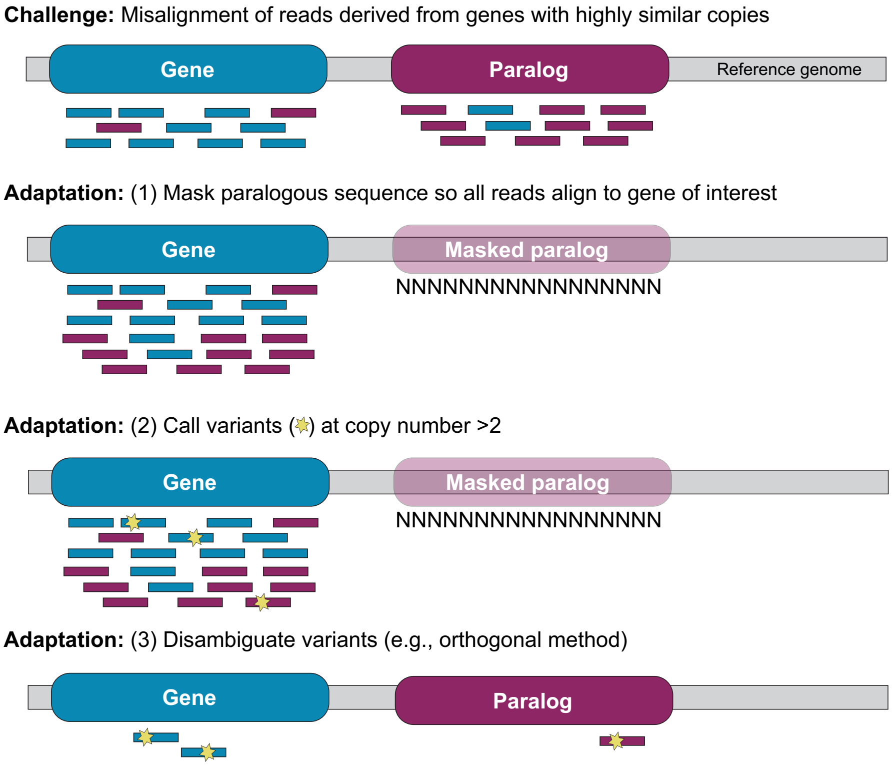
Challenge: Misalignment of reads derived from genes with highly similar copies
Gene
Paralog
Reference genome
Adaptation: (1) Mask paralogous sequence so all reads align to gene of interest
Gene
Masked paralog
NNNNNNNNNNNNNNNNN
Adaptation: (2) Call variants (★) at copy number >2
Gene
Masked paralog
NNNNNNNNNNNNNNNNN
Adaptation: (3) Disambiguate variants (e.g., orthogonal method)
Gene
Paralog

FIGURE 1 Disease genes that share highly similar sequences with other genes or pseudogenes are difficult to analyze with next-generation sequencing methods because of frequent misalignment of short sequencing reads. This schematic provides a high-level overview of this challenge and the steps taken to accurately identify variants in these genes. (1) Mask the sequence of the highly similar copy within the reference genome by replacing reference bases with ns or otherwise computationally obscuring the locus. Sequencing reads from the gene of interest and its similar copy will align only to the gene-of-interest locus. (2) Call variants at copy number >2 to account for the single-locus alignment. (3) Determine the true locus of a variant (i.e., disambiguate the variant) by assigning it to its true locus with orthogonal methods

et al., 2015; Rehm et al., 2013). All validations were also performed to meet CLIA program requirements and New York state standards for clinical genetic testing (Centers for Disease Control and Prevention, 2021; New York State Department of Health, 2014).

Sensitivity was determined as the ratio of all true positives over the sum of all true positives and all false negatives. Specificity was determined as the ratio of all true negatives over the sum of all true negatives and all false positives. Confidence intervals (CIs) were also calculated for both sensitivity and specificity of variant detection.

3 | RESULTS

The technically challenging variants described in this study required nine customized assays and bioinformatic approaches (one for each gene) to be incorporated into a high throughput, multi-gene sequencing workflow. The development and validation of these customizations are described separately by gene below. An overview of the adaptations is also provided in Table 1.

3.1 | Genes with highly similar pseudogenes or paralogous copies

3.1.1 | *PMS2*

The tumor suppressor gene *PMS2* is associated with Lynch syndrome, which confers an increased risk for colorectal, stomach, and other types of cancer (Cerretelli et al., 2020). While the sequences of *PMS2* exons 1–11 are sufficiently unique to allow accurate mapping of sequence reads to the reference genome, variants in the last four exons of this oncogene can be difficult to detect because they share high sequence similarity with a nearby truncated pseudogene, *PMS2CL*. Gene conversion can occur between the similar sequences of *PMS2* and *PM2CL* (Hayward et al., 2007). The final four exons of *PMS2* (exons 12–15) and those of *PMS2CL* are nearly identical, precluding unambiguous alignment of short-read sequences originating from the two loci (Hayward et al., 2007). To circumvent these difficulties, we computationally modified the reference genome sequence to mask the final four exons of *PMS2CL*, forcing reads from the corresponding exons in *PMS2CL*

TABLE 1 Technically challenging variants integrated into next-generation sequencing workflow

Disease gene	Type of method adapted	Adaptations	Variants reported
Variants in genes with high-similarity copies			
PMS2	Bioinformatic	Pseudogene masked	SNVs, indels, and CNVs
		Variant calling at > copy number 2	
SMN1/*SMN2*	Bioinformatic	Pseudogene masked	SNVs and indels in *SMN1* terminal exon (exon 8)
		Variant calling at > copy number 2	*SMN1* and *SMN2* copy number
			Suspected pathogenic SNVs and indels in ambiguous *SMN1*/*SMN2* exons 1–7
GBA1	Bioinformatic	Nanomasking of polymorphic paralogous sequence variants	Nineteen specific sequence variants: c.84dupG; c.115 + 1G > A; c.222_224delTAC; c.475C > T; c.595_596delCT; c.680A > G; c.721G > A; c.754 T > A; c.1226A > G; c.1246G > A; c.1297G > T; c.1263_1317del; c.1342G > C; c.1343A > T; c.1448 T > C; c.1504C > T; c.1505G > A; c.1603C > T; c.1604G > A
		Gene conversion/fusion events identified by copy number analysis	
HBA1/ *HBA2*	Assay and bioinformatic	More gene-targeting probes	CNVs in *HBA1* and *HBA2*
		Paralog masked	"Constant Spring" SNV (*HBA2*:C.427 T > C)
		Haplotype-based copy number calling in exons, introns, and intergenic regions	
CYP21A2	Bioinformatic	Gene-specific variant calling (adjusted thresholds)	Select SNVs and indels
		Gene conversion/fusion events identified by copy number analysis	Gene fusions and large gene conversions
			Compensating full gene duplications in the presence of pathogenic variants
Repetitive sequence variants			
ARX (polyalanine)	Assay	Increased DMSO	SNVs, indels, and CNVs
		Increased probes for *ARX* exon 2	
CFTR (poly-T/TG)	Bioinformatic	Gene-specific variant calling	Repeat alleles (5Ts with 11, 12, or 13 TG repeats)
			(SNVs, indels, and CNVs outside the poly-T/TG repeat region are also reported)
FMR1 (AGG interruptions in CGG repeats)	Assay	Screen samples for risk of expansion (based on CGG repeat number)	Number of AGG interruptions
		PacBio sequencing to identify CGG tract sizes and locations	
Structural variants			
MSH2 (Boland inversion)	Bioinformatic	Targeted sequencing probes designed over expected breakpoints	Inversion variant
		Split-read analysis	(SNVs, indels, and CNVs outside the inversion region are also reported)

Abbreviations: CNV, copy number variant; DMSO, dimethyl sulfoxide; indels, insertions and deletions; NGS, next-generation sequencing; LR-PCR, long-range PCR; SNV, single nucleotide variant.

(in addition to *PMS2* reads) to align to the *PMS2* locus (Figure 1). Sequence variants were called by Freebayes using an expected copy number of 4. This method is agnostic to a sequence variant's locus of origin (*PMS2* or *PMS2CL*). In parallel, CNVs were called by CNVitae, which was adapted to accurately call CNVs at this locus relative to an expected copy number of 4. Subsequently, to disambiguate sequence variants, we performed locus-specific nested LR-PCR by adapting previously described methods (Vaughn et al., 2011) and sequenced the resulting amplicons with PacBio sequencing. CNVs that included any of *PMS2* exons 12–15 were disambiguated using MLPA-seq, nested LR-PCR, and PacBio sequencing. CNVs in *PMS2* exons 1–11 did not require disambiguation, as reads were reliably aligned to the correct locus.

To validate the methods for variant detection in *PMS2*, we used two reference sets: one containing sequence variants and the other containing CNVs (Table S1). The sequence variant reference set comprised 32 unique samples carrying 33 orthogonally identified sequence variants in *PMS2* or *PMS2CL*. The CNV reference set consisted of 28 samples carrying 21 orthogonally identified single-exon CNVs (though 7 samples had no *PMS2* or *PMS2CL* CNVs). In a validation experiment with replicates, our custom methods correctly identified all true positive sequence variants (205/205) and no false positives (0/34,876 true negatives). Thus, the NGS-based adaptations had 100% sensitivity (95% CI, 98.2%–100.0%) and 100% specificity (95% CI, 99.9%–100.0%) for sequence variant detection. For CNVs, the adaptations correctly identified all of the true positives (90/90) and no false positives (0/50), yielding 100% sensitivity (95% CI, 95.9%–100.0%) and 100% specificity (95% CI, 92.9%–100.0%) (Table 2).

TABLE 2 Summary of validation results for adapted NGS assays for technically challenging variants

Gene	Validation sample type[a]	No. of orthogonally known variants[b]	Concordant validation results, No. (%)
Variants in genes with highly similar copies			
PMS2	Sequence variant positive	205	205 (100)
	Sequence variant negative	34,876	34,876 (100)
	CNV positive	90	90 (100)
	CNV negative	50	50 (100)
SMN1/*SMN2*	*SMN1* exon 8 CNV	16	16 (100)
	SMN2 exon 8 CNV	14	14 (100)
GBA1	Sequence variant positive	34	34 (100)
	Sequence variant negative	12	12 (100)
HBA1/*HBA2*	CNV positive	44	44 (100)
	CNV negative	25	25 (100)
CYP21A2	Sequence variant positive	68	68 (100)[c]
	Sequence variant negative	1390	1359 (97.8)[c]
	CNV	96	96 (100)
Repetitive sequence variants			
ARX (polyalanine)	Expansion positive	5	5 (100)
	Expansion negative	112	96 (85.7)[d]
CFTR (poly-T/TG)	Variant positive	22	22 (100)
FMR1 (AGG interruptions in CGG repeats)	AGG interruptions length and position	27	27 (100)
Structural variants			
MSH2 (Boland inversion)	Inversion positive	8	8 (100)
	Inversion negative	51,575	51,575 (100)

[a] Additional information regarding validation samples can be found in Tables S1–S12.

[b] May include replicates.

[c] These results were based on the primary assay. A separate validation of the confirmation method (PacBio sequencing) demonstrated a final sensitivity and specificity of 100% for 20 clinical samples of known variant status (Table S8).

[d] These results were based on the primary assay. Samples identified as positive by the primary NGS assay undergo orthogonal confirmation.

3.1.2 | *SMN1* and *SMN2*

Spinal muscular atrophy (SMA) is a degenerative neuromuscular disorder that can cause complete paralysis and death in infancy, if not treated. Almost all cases of SMA are caused by a recurrent pathogenic deletion in the *SMN1* gene. Identifying the common deletion in *SMN1* is challenging because of a nearly identical paralog, *SMN2*. The paralogous copy differs from *SMN1* by a single nucleotide (the gene-determining nucleotide, c.840) in the terminal coding exon and encodes a partially functioning protein (Prior et al., 2000). Further complicating genetic testing for SMA is that *SMN2* copy number varies among individuals. Because high *SMN2* copy number can mitigate the most severe effects of *SMN1* loss through some functional compensation by the SMN2 protein (Kariyawasam et al., 2019), accurate reporting of both *SMN1* and *SMN2* copy number is critical for determining a prognosis for a patient. In addition, *SMN2* copy number can determine whether certain treatments are recommended for SMA-positive individuals (Glascock et al., 2018, 2020).

Our approach to *SMN1* and *SMN2* testing using short-read NGS required multiple adaptations to our bioinformatics methods. First, we masked *SMN2* in the reference genome to force all *SMN2*-derived sequence reads to align to the *SMN1* locus. Sequence variants were then called at a copy number of 4. We also adapted CNVitae to identify CNVs relative to an expected copy number of 4 to account for the co-alignment of the *SMN1* and *SMN2* reads. The independent copy numbers of *SMN1* and *SMN2* were determined based on the allele balance of the gene-determining nucleotide in the clinically relevant eighth exon (conventionally referred to as exon 7), which is the terminal coding exon. All sequence variants and any CNVs that did not involve the terminal coding exon were not disambiguated due to 100% sequence identity; however, deleterious variants in these regions were reported to clinicians and their patients to guide further testing, as necessary.

For carrier screening, samples with two copies of *SMN1* were also examined for the presence of an intronic SNV in *SMN1* (g.27134T > G) that is associated with silent carrier status (i.e., two copies of *SMN1* on one chromosome and no copies on the other) in some ancestral backgrounds (Hendrickson et al., 2009; Luo et al., 2014).

We validated our *SMN1* and *SMN2* copy number calling approach using samples with known *SMN1* and *SMN2* copy number genotypes provided by the biobank, including *SMN1* heterozygous deletions, *SMN2* duplications, and *SMN2* with copy numbers of 3 or more (Table S2). In a validation experiment with replicates, we observed 100% concordance with both *SMN1* (16/16) and *SMN2* (14/14) genotypes (Table 2), yielding a specificity of 100% for copy number detection (95% CI, 79.4%–100.0% for *SMN1* and 95% CI, 76.8%–100.0% for *SMN2)*. Separately, we performed a validation experiment comparing NGS results to droplet digital PCR on 55 clinical samples and found 100% concordance for *SMN1* copy numbers 0–2 and *SMN2* copy numbers 0–4 (Table S3).

3.1.3 | *GBA1*

Gaucher disease, a potentially fatal lysosomal storage disorder with toxic accumulation of lipids in many tissues, is an autosomal recessive disorder caused by pathogenic variants in *GBA1* (Hruska et al., 2008). Detection of pathogenic *GBA1* variants is complicated by a pseudogene, *GBAP1*. Though highly similar, *GBA1* and *GBAP1* have different polymorphic sequence variants that are not fixed in the population. Gene conversions and gene-pseudogene fusions between the two loci can thus lead to inappropriate alignment of sequence reads during NGS (from *GBAP1*-derived variants aligning to *GBA1* and vice versa) (Figure 2). These alignment artifacts then falsely appear as copy number gains or losses.

To accurately call variants in *GBA1* using NGS, we first modified the reference genome sequence to mask polymorphic sequence variant positions. When no gene conversion events were present, this fine-scale masking of the reference genome enabled accurate alignment of NGS reads to *GBA1* and *GBAP1* and subsequent detection of the 19 most commonly observed pathogenic sequence variants in *GBA1* (15 SNVs and 4 indels) (Table 1). To identify potential gene conversion and fusion events, deletion calls made by CNVitae in *GBA1* were used as a primary assay. Putative rearrangement events were then analyzed by nested LR-PCR and PacBio sequencing to confirm the presence of pathogenic variants in the affected regions.

To validate this approach, we used nine samples known to harbor variants in *GBA1* and two samples with normal *GBA1* (Table S4). In a validation experiment with replicates, we observed 100% concordance for all variants (46/46), representing 100% sensitivity (95% CI, 89.7%–100.0%) and 100% specificity (95% CI, 73.5%–100.0%) (Table 2).

3.1.4 | *HBA1* and *HBA2*

Alpha thalassemia is an inherited autosomal recessive blood disorder most often caused by deletions involving the adjacent, paralogous *HBA1* and *HBA2* genes, which have identical coding sequences but distinct noncoding regulatory regions on chromosome 16. Severity of alpha thalassemia ranges from mild anemia to neonatal death depending on the length, location, and combination

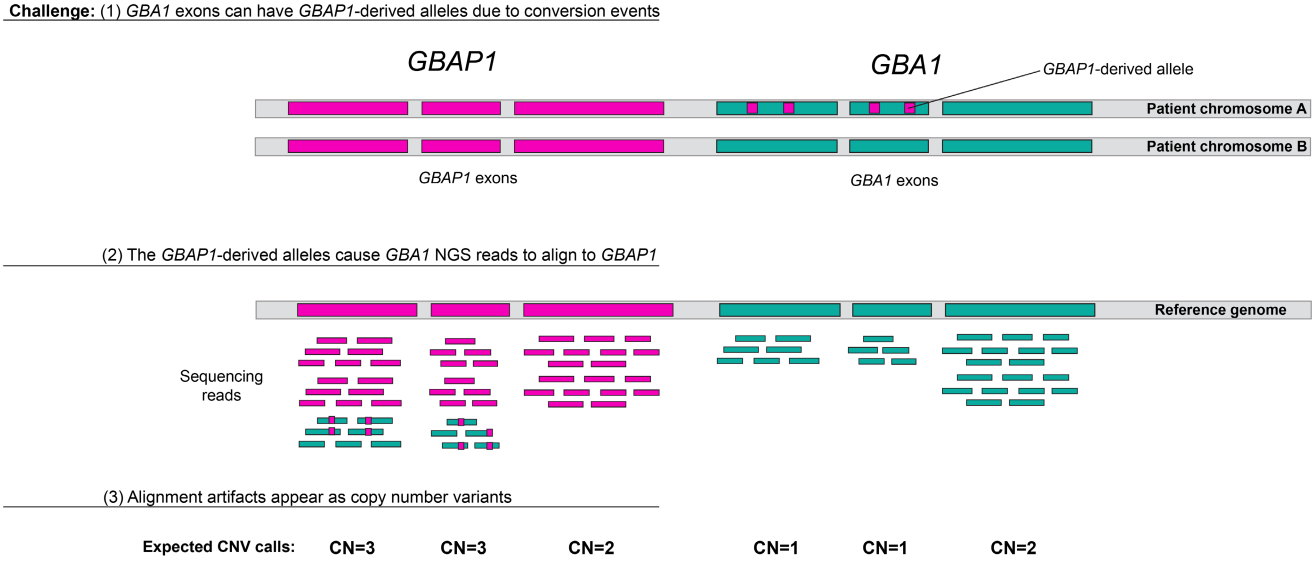
Challenge: (1) GBA1 exons can have GBAP1-derived alleles due to conversion events
GBAP1
GBA1
GBAP1-derived allele
Patient chromosome A
Patient chromosome B
GBAP1 exons
GBA1 exons
(2) The GBAP1-derived alleles cause GBA1 NGS reads to align to GBAP1
Reference genome
Sequencing reads
(3) Alignment artifacts appear as copy number variants
Expected CNV calls:
CN=3
CN=3
CN=2
CN=1
CN=1
CN=2

FIGURE 2 One of the challenges of analyzing *GBA1* by next-generation sequencing is the potential for gene conversion events that cause *GBAP1*-associated alleles to be integrated into *GBA1* exons. In this illustrated example, *GBA1* has undergone a recombination event that results in two converted exons with *GBAP1*-specific alleles. These converted alleles cause next-generation sequencing reads to align to *GBAP1*. When variants are called, these alignment artifacts appear as copy number losses for *GBA1* and copy number gains for *GBAP1*, which are then confirmed orthogonally. CN, copy number; CNV, copy number variant; NGS, next-generation sequencing

of pathogenic deletions. The deletions vary in size and can involve *HBA1, HBA2*, or both genes (Tamary & Dgany, 2005); specific deletions also arose in certain populations and have been propagated widely through those populations. In addition, duplications in the region can complicate CNV calling, especially when a duplication on one chromosome overlaps with a deletion on the other (Figure 3). Thus, these genes present multiple challenges to short-read NGS testing: alignment errors can occur due to the high sequence identity between the two genes, and copy number calls may be ambiguous because the same observed copy number can result from different actual copy number combinations of *HBA1* and *HBA2* (Figure 3).

To test individuals for carrier status for pathogenic CNVs related to alpha thalassemia in a high-throughput NGS workflow, we implemented three modifications. First, we increased the number of targeted capture baits for both *HBA1* and *HBA2* in our assay to obtain deeper sequencing coverage for both genes. Second, we masked the *HBA1* locus and some flanking intergenic sequence in the reference genome to force all reads to align to the *HBA2* locus. Third, to address the region's propensity for overlapping deletion and duplication events, we modified CNVitae in two ways. First, instead of only calling copy numbers within exons, we expanded the analysis to also include introns and intergenic regions of the *HBA1* and *HBA2* loci. Second, instead of calling CNVs at a copy number of 4 (since *HBA1* and *HBA2* reads all align to the reference *HBA2*), CNVitae was adapted to generate copy-number likelihood scores for 325 potential combinations of 25 possible *HBA1* and *HBA2* configurations on each copy of chromosome 16. The resulting CNV likelihoods were compared to known combinations of *HBA1* and *HBA2* configurations associated with alpha thalassemia (including fusions and duplications of *HBA1* and *HBA2* as described in the literature or other external sources). The most concordant combination of *HBA1* and *HBA2* from each copy of chromosome 16 was selected. However, if no match was identified, sequencing data were manually reviewed and, if needed, followed with an orthogonal assay.

To validate this approach, we tested genomic DNA samples with known *HBA1* and *HBA2* genotypes (Table S5). The *HBA1/HBA2* composition was correctly identified in 44/44 positive and 25/25 negative samples, for 100% sensitivity (95% CI, 92.0%–100.0%) and 100% specificity (95% CI, 86.3%–100.0%) (Table 2).

3.1.5 | *CYP21A2*

Congenital adrenal hyperplasia (CAH) is an autosomal recessive disorder that affects hormone synthesis and can lead to a range of outcomes including genital dysmorphology, salt wasting, and neonatal death. More than 90% of cases are caused by pathogenic variants in *CYP21A2* (Krone & Arlt, 2009). Identification of *CYP21A2* variants is complicated by the presence of a neighboring pseudogene, *CYP21A1P*, which has 99% sequence similarity

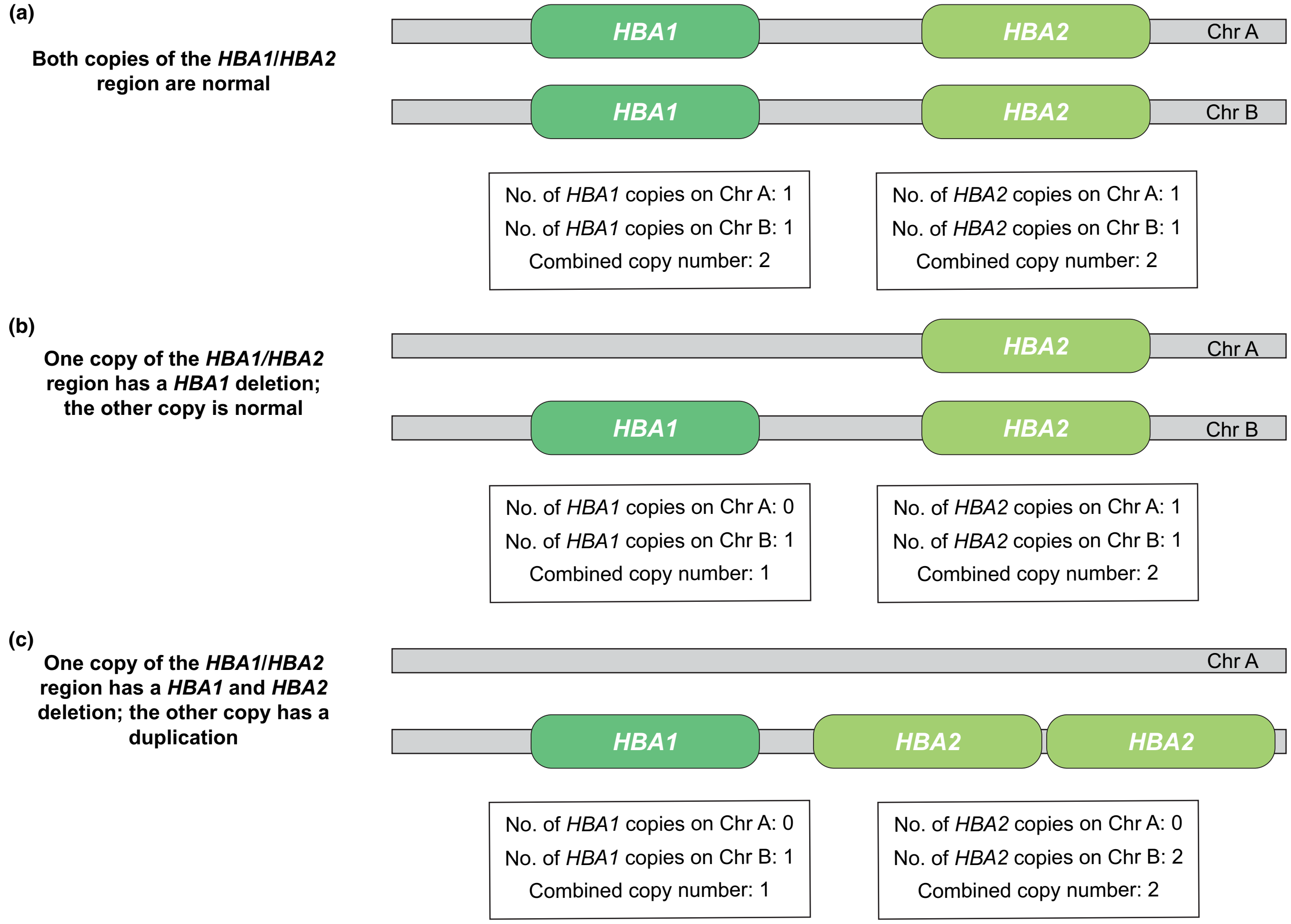
(a)
Both copies of the HBA1/HBA2 region are normal
HBA1
HBA2
Chr A
HBA1
HBA2
Chr B
No. of HBA1 copies on Chr A: 1
No. of HBA1 copies on Chr B: 1
Combined copy number: 2
No. of HBA2 copies on Chr A: 1
No. of HBA2 copies on Chr B: 1
Combined copy number: 2
(b)
One copy of the HBA1/HBA2 region has a HBA1 deletion; the other copy is normal
HBA2
Chr A
HBA1
HBA2
Chr B
No. of HBA1 copies on Chr A: 0
No. of HBA1 copies on Chr B: 1
Combined copy number: 1
No. of HBA2 copies on Chr A: 1
No. of HBA2 copies on Chr B: 1
Combined copy number: 2
(c)
One copy of the HBA1/HBA2 region has a HBA1 and HBA2 deletion; the other copy has a duplication
Chr A
HBA1
HBA2
HBA2
No. of HBA1 copies on Chr A: 0
No. of HBA1 copies on Chr B: 1
Combined copy number: 1
No. of HBA2 copies on Chr A: 0
No. of HBA2 copies on Chr B: 2
Combined copy number: 2

FIGURE 3 Detecting copy number variants involving *HBA1* and *HBA2* is challenging. Different combinations of deletions and duplications in the *HBA1* and *HBA2* locus can produce the same copy number calls in generic next-generation sequencing workflows. For example, in a normal genomic arrangement (a), both genes have 1 copy of each gene on each copy of chromosome 16 (Chr A and Chr B), corresponding to a copy number of 2. In one potential disrupted arrangement (b), a deletion of *HBA1* would result in 0 copies of *HBA1* on Chr A and 1 copy on Chr B, for a combined copy number of 1. In another potential disrupted arrangement (c), a deletion of both *HBA1* and *HBA2* on Chr A and a duplication of *HBA2* on Chr B would result in a combined copy number of 1 for *HBA1* and 2 for *HBA2*. Genetics-informed bioinformatics processes are needed to differentiate between the overlapping copy numbers

to *CYP21A2*. Frequent gene fusion and gene conversion events between the two loci further complicate detection of disease-associated variants in *CYP21A2*.

We developed an NGS assay, optimized for sensitivity, that can detect 60 CAH-associated variants (including the 12 most commonly observed pathogenic variants), potentially compensating *CYP21A2* duplications, and various 30-kilobase deletions in the region. We anticipated ambiguous read alignment and therefore adjusted our caller to consider allele balances below the expectation for standard germline variants, which is typically 50% for a heterozygous variant and 100% for a homozygous variant. Copy number states, as determined by CNVitae, were used to infer the presence of fusions or conversions.

The adapted NGS method for *CYP21A2* variant detection was validated on 24 internal samples known through orthogonal methods to contain sequence variants in clinically relevant *CYP21A2* positions (Table S6). In replicated validation runs for sequence variants, our short-read NGS-based primary assay correctly identified 68/68 positive variants and 1359/1390 negative positions (Table S7), for a sensitivity of 100% (95% CI, 94.7%–100.0%), a specificity of 97.8% (95% CI, 96.9%–98.5%), and an overall concordance of 97.9% (Table 2). All suspected pathogenic variants at reportable loci by primary assay were confirmed with PacBio sequencing, which was separately validated on 20 samples and demonstrated 100% sensitivity (95% CI, 93.0%–100.0%) and 100% specificity (95% CI, 98.1%–100.0%) (Table S8).

The adapted NGS method was also validated on 46 samples confirmed by MLPA to harbor *CYP21A2* duplications and 51 samples with normal copy number (Table S6). The NGS adaptation identified the correct copy number in all validation samples, demonstrating 100% sensitivity (95%

CI, 92.1%–100.0%) and 100% specificity (95% CI, 93.0%–100.0%) for CNVs.

3.2 | Genes affected by low-complexity repeat tracts

3.2.1 | *ARX* (polyalanine repeats)

Pathogenic variants in the X-linked *ARX* gene cause clinically heterogeneous conditions, including a form of early infantile epileptic encephalopathy (EIEE). Because *ARX*-related EIEE arises as a result of polyalanine repeat expansions in exon 2 of this gene (Marques et al., 2015), accurately determining the size of the repeat tract is essential during diagnostic genetic testing in children with epilepsy. Triplet repeats or other low-complexity repeat sequences can be difficult to interrogate by NGS for several reasons, including suboptimal polymerase processivity and fidelity at repetitive sequences and challenges aligning sequence reads to a reference genome.

We adapted our Illumina NGS assay to increase the density of *ARX* exon 2 baits to 5X density, compared to the average 1–2X density for most assayed regions. To increase the efficiency of pre-capture and post-capture amplification steps in the highly GC-rich region, we altered the chemistry by increasing the amount of dimethyl sulfoxide (DMSO). Following NGS and bioinformatic processing, samples with low sequencing depth and/or samples with a split-read signal above background levels at *ARX* exon 2 were identified and further evaluated with an orthogonal method.

We validated this NGS-based modification using five samples in which previous Sanger sequencing had confirmed hemizygous or heterozygous *ARX*-polyalanine expansion genotypes, and 112 normal samples (Table S9). The adapted short-read NGS detected the *ARX*-polyalanine variants in 5/5 positive samples and detected false positives in 16/112 negative samples for a sensitivity of 100% (95% CI, 47.8%–100.0%), a specificity of 85.7% (95% CI, 80.5%–92.7%), and an overall concordance of 86.3% (Table 2). All true positives were confirmed and all false positives were identified using an orthogonal long-read NGS method (PacBio sequencing), ensuring high accuracy.

3.2.2 | *CFTR* (poly-T/TG repeats)

Cystic fibrosis (CF) is a life-threatening disease that affects multiple organs, including the lungs (Elborn, 2016). CF is an autosomal recessive disorder caused by pathogenic variants in the *CFTR* gene, which is also associated with CF-related conditions such as pancreatitis and male infertility. Low-complexity poly-T and TG repeats within intron 9 of the *CFTR* gene can alter splicing of *CFTR* transcripts, and specific combinations of these repeats with other variants affect clinical manifestation of CF-related disorders (Nykamp et al., 2021). Most individuals have poly-T genotypes consisting of 5, 7, or 9 Ts, and the presence of 5 Ts is known to disrupt the function of *CFTR*. Further, the length of the adjacent tract of TG repeats modulates the severity of the effects of the 5 T repeats.

Given the challenge of accurately detecting repeats with standard NGS approaches, we used the custom variant caller Coalgen to determine the composition of the poly-T and TG tracts in *CFTR*. Following alignment to the reference genome, all reads associated with *CFTR* intron 9 were compared against all known combinations of poly-T and TG repeats (specifically, all possible combinations of 2–9 T repeats and 8–13 TG repeats) to identify the best-fitting haplotypes and zygosity according to binomial likelihood. The performance of this method for detecting *CFTR* poly-T and TG repeats was successfully validated using 22 reference samples (Table S10), for which the adapted NGS-based method was 100% sensitive (95% CI, 84.6%–100.0%) with concordant variants called for all 22 samples (Table 2).

3.2.3 | *FMR1* (AGG interruptions in CGG repeats)

Fragile X syndrome, the most common cause of inherited intellectual disability and autism, is primarily caused by expansion of a CGG trinucleotide repeat tract within the 5′ untranslated region of the *FMR1* gene on the X chromosome (Hayward et al., 2017). Normal repeat tracts with less than 40 CGG units are stably transmitted from parent to offspring. Tracts with 55–200 CGG repeats are considered to be in the premutation range, and female individuals carrying these alleles are at significant risk of transmitting a highly expanded full mutation allele (>200 repeats) to offspring, as well as developing premature ovarian failure. Both female and male individuals with premutation repeat alleles are also at risk for Fragile X-associated tremor/ataxia syndrome (FXTAS). For tracts of 55–90 CGG repeats, the risk of expanding to a full mutation is reduced by the presence of interrupting AGGs that stabilize the CGG repeat tract (Nolin et al., 2015). To provide more accurate risk information from carrier screening during reproductive decision-making, correctly identifying the number of AGG interruptions within the *FMR1* CGG-repeat tract can be useful (Nolin et al., 2015).

Since standard short-read NGS cannot reliably interrogate long triplet repeat tracts, and therefore cannot detect AGG interruptions, we developed an approach using orthogonal methods. First, samples from female individuals seeking carrier screening were screened for *FMR1* premutation alleles using repeat-primed PCR and capillary electrophoresis. Samples with 55–90 CGG repeats were then subjected to long-read NGS (PacBio sequencing), and a custom algorithm inferred the number and position of AGG interruptions for each CGG-repeat allele from the resulting sequence data.

We validated this approach with six orthogonally confirmed samples with 55–90 *FMR1* CGG repeats and known AGG profiles (Table S11). In a validation study with replicates, the PacBio-based method successfully detected 27/27 AGG genotypes for 100% concordance with an orthogonal method and 100% sensitivity (95% CI, 87.2%–100.0%) for AGG interruptions in CGG expansion.

3.3 | Genes associated with recurrent structural rearrangements

3.3.1 | MSH2

The tumor suppressor gene *MSH2*, like *PMS2*, is associated with Lynch syndrome (Cerretelli et al., 2020). A recurrent pathogenic variant in *MSH2* involves a copy-neutral paracentric inversion that separates exons 1–7 from the rest of the gene; this is sometimes referred to as the “Boland inversion.” Events such as the Boland inversion are challenging to detect with generic short-read NGS because reads that span the rearrangement breakpoints are soft-clipped and most NGS bioinformatic pipelines are not optimized to harness information to detect this variant type.

To detect the clinically important Boland inversion, we designed oligonucleotide baits to target the known breakpoint regions to ensure that the breakpoints for the inversion could be captured. The custom split-read detection algorithm then identified above-baseline split-read signals resulting from reads partially aligned to intron 7 of *MSH2* and partially aligned ~10 Mb upstream of exon 1 (5′ to the start of *MSH2*). The split-read signals were manually reviewed and confirmed by PacBio sequencing using primers specific to the Boland inversion.

Eight samples known to harbor the Boland inversion and 51,575 samples known to be negative for the inversion were used to validate our methods (Table S12). Our NGS-based approach identified the inversion in all positive samples and no negative samples, indicating 100% sensitivity (95% CI, 63.1%–100.0%) and 100% specificity (95% CI, 99.9–100%) (Table 2).

4 | DISCUSSION

Variants that are technically challenging to detect by standard NGS workflows, and therefore require the use of other methods, account for a considerable proportion of clinically significant variant types (Lincoln et al., 2021). Through appropriate customization, NGS has the capability to capture even these technically challenging variants from a single genetic test with a single sample. Such customization can simplify test selection and test ordering for clinicians and enable laboratories to consolidate various methodologies and workflows into one platform. Here, we have described several NGS-based solutions for addressing technically challenging variants in a high-throughput, cost-effective NGS workflow for targeted gene panels; each of the nine individually developed adaptations included custom adjustments to NGS assay chemistry and/or bioinformatic processing. Five of the adaptations included solutions for assaying technically challenging variants in genes that have copies with high sequence similarity (*PMS2, SMN1*, *GBA1*, *HBA1/HBA2*, and *CYP21A2)*. These variants are typically challenging to analyze in standard NGS workflows because short sequencing reads may align to incorrect locations (e.g., within a pseudogene) or be excluded due to quality control bioinformatic filters, leading to potential false positive or false negative calls. Three of the adaptations addressed variants in genes with repeat sequence tracts (*ARX* polyalanine, *CFTR* poly-T/TG, and *FMR1* AGG interruptions in CGG repeats), which often preclude accurate short-read alignment to the reference genome. Another adaptation enabled accurate identification of the *MSH2* Boland inversion, a copy-neutral sequence inversion that is typically missed by standard NGS due to inadequacies in standard variant callers.

All of these adaptations were validated with known reference samples and demonstrated the desired performance for clinical use across all genes evaluated. In some cases, an NGS adaptation led to better performance than traditional non-NGS methods. For example, in one control sample obtained from a biobank, our NGS-based method for *SMN1*/*SMN2* testing was more precise than the genotype provided by the biobank. The sample was originally described as harboring three or more copies of *SMN2*, but our methods indicated the presence of five copies (confirmed by an external lab) (Stabley et al., 2015) instead of the 3 or more copies indicated by the biobank.

Although the solutions described here were customized for each gene or variant, some principles overlap among our lab-developed NGS assays for gene panels. First, the success of a customized NGS approach depends on the close coupling of consistent, highly automated wet laboratory processes with internally developed bioinformatic analyses. Automation by liquid-handling robots increases the uniformity of sample processing, which creates highly reproducible and stable read depths. This enables important capabilities such as high-resolution intragenic CNV detection with NGS alone (Truty et al., 2019). Second, some adaptations were applied to more than one locus, such as masking highly similar gene copies in the reference genome so that all reads from a disease gene and its segmentally duplicated copy map to a single location (Hogan et al., 2018) and using gene-specific modifications to variant calling that are informed by known distributions of genotypes in the population (Table 1). In addition, bioinformatics solutions such as the split-read caller can be used not only for the Boland inversion but for other technically challenging variants not described here (e.g., disruptive retrotransposon insertions). Third, for some variants, sensitivity was favored over specificity to ensure that a clinically important genotype was not missed, in which case high specificity was ensured through appropriate follow-up orthogonal confirmation via long-read sequencing or other methods. For example, while the validation studies of the NGS-based primary assay for sequence variants in *CYP21A2* demonstrated that the primary assay conservatively overcalls clinically relevant variants, all primary assay-positive samples are subjected to confirmatory analysis by an orthogonal method (PacBio) before the final report is sent to an ordering clinician. The *CYP21A2* confirmation assay was separately validated and demonstrated 100% sensitivity and specificity.

An overarching goal of this work was to develop a more clinically useful, efficient, and cost-effective standard for NGS-based clinical genetic testing with gene panels—one that improves diagnostic yield by capturing more clinically significant variant types using a single sample, thus eliminating the need for sequential testing and likely significantly reducing the time and costs to reach a diagnosis for many patients. These improvements build on the key advantage already conferred by NGS, namely the ability to test many genes and genomic regions in multiple individuals simultaneously. Other clinical labs have also reported NGS adaptations that can streamline clinical diagnostic testing, such as previously reported NGS-based primary assays for *PMS2* variants that can be followed by confirmation of positive findings (Gould et al., 2018; Herman et al., 2018). Such adaptations stand to substantially benefit patients. In addition, novel bioinformatic solutions are expanding the capabilities of NGS analysis. (Hogan et al., 2018; Lincoln et al., 2021; Yu et al., 2019) Future adaptations may enable additional complicated genes to be included in high-throughput NGS workflows based on common structural challenges such as the presence of pseudogenes, short tandem repeats, and others. To achieve this, gene-specific assay development and validation will be needed. While the methods and adaptations we have described here apply to targeted NGS gene panels and not to whole exome or whole genome sequencing, some of these adaptations can be used for multiple sequencing approaches. For example, it is possible to detect exonic copy number variants or other structural rearrangements through whole exome sequencing (Retterer et al., 2015, 2016) although the resolution and sensitivity may be reduced relative to that of a deep coverage gene panel. NGS capabilities are continually evolving, with emerging capabilities including detection of more complicated clinically relevant variant types such as large triplet repeat expansions, mosaic variant burden across the genome, and novel chromosomal rearrangements. Beyond these, more complicated sequence architectures demand the use of multiple next-generation methods such as PacBio sequencing, Bionano optical mapping, and Oxford nanopore sequencing to reliably sequence all types of genomic regions, as elegantly demonstrated in a report of the first "complete sequence" of a human genome (Nurk et al., 2022). All of these advances may eventually provide complete genomic information with every type of variant captured at a low cost and without impact on turnaround time for the majority of individuals tested, further improving the breadth with which genomic information is used in preventive and diagnostic health care.

AUTHOR CONTRIBUTIONS

Susan Rojahn: writing – original draft, writing – review & editing; Tina Hambuch: conceptualization, supervision, validation, writing – original draft, writing – review & editing; Jessika Adrian: validation, writing – review & editing; Erik Gafni: methodology, writing – review & editing; Alex Gileta: methodology, validation, writing – review & editing; Hannah Hatchell: methodology, writing – review & editing; Britt Johnson: supervision, validation, writing – review & editing; Ben Kallman: validation, writing – review & editing; Kate Karfilis: validation, writing – review & editing; Curtis Kautzer: methodology, validation, writing – review & editing; Michael Kennemer: methodology, validation, writing – review & editing; Lloyd Kirk: methodology, writing – review & editing; Daniel Kvitek: methodology, writing – review & editing; Jessica Lettes: methodology, writing – review & editing; Fenner Macrae: methodology, writing – review & editing; Fernando Mendez: methodology, validation, writing – review & editing; Joshua Paul: methodology,

writing – review & editing; Maurizio Pellegrino: methodology, validation, writing – review & editing; Ronny Preciado: methodology, writing – review & editing; Jan Risinger: methodology, validation, writing – review & editing; Matthew Schultz: methodology, validation, writing – review & editing; Lindsay Spurka: methodology, validation, writing – review & editing; Sajani Swamy: methodology, validation, writing – review & editing; Rebecca Truty: methodology, writing – review & editing; Nathan Usem: methodology, validation, writing – review & editing; Andrea Velenich: methodology, validation, writing – review & editing; Swaroop Aradhya: conceptualization, methodology, supervision, writing – original draft, writing – review & editing.

ACKNOWLEDGMENT

The authors thank Marjorie Jody Westbrook of Invitae for her review and feedback on technical aspects of this manuscript and Kerry Aradhya of Invitae for scientific editing.

ETHICS STATEMENT

This study was conducted in compliance with CLIA and CAP standards for clinical molecular genetic testing, including with respect to patient privacy and informed consent, wherever applicable.

CONFLICT OF INTEREST

All authors are current or former employees and shareholders of Invitae.

DATA AVAILABILITY STATEMENT

Swaroop Aradhya had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Clinically reported variants are made publicly available in ClinVar according to the data-sharing preferences of the individual tested: https://www.ncbi.nlm.nih.gov/clinvar/submitters/500031/.

ORCID

Susan Rojahn https://orcid.org/0000-0001-5888-7693
Swaroop Aradhya https://orcid.org/0000-0001-6219-2931

REFERENCES

Abou Tayoun, A. N., Krock, B., & Spinner, N. B. (2016). Sequencing-based diagnostics for pediatric genetic diseases: Progress and potential. *Expert Review of Molecular Diagnostics*, *16*(9), 987–999.

Aziz, N., Zhao, Q., Bry, L., Driscoll, D. K., Funke, B., Gibson, J. S., Grody, W. W., Hegde, M. R., Hoeltge, G. A., Leonard, D. G. B., Merker, J. D., Nagarajan, R., Palicki, L. A., Robetorye, R. S., Schrijver, I., Weck, K. E., & Voelkerding, K. V. (2015). College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Archives of Pathology & Laboratory Medicine*, *139*(4), 481–493.

Centers for Disease Control and Prevention. (2021). *Clinical laboratory improvement amendments (CLIA) laws and regulations*. Division of Laboratory Systems (DLS). https://www.cdc.gov/clia/law-regulations.html

Cerretelli, G., Ager, A., Arends, M. J., & Frayling, I. M. (2020). Molecular pathology of lynch syndrome. *The Journal of Pathology*, *250*(5), 518–531.

Chong, H. K., Wang, T., Hsiao-Mei, L., Seidler, S., Hong, L., Keiles, S., Chao, E. C., Stuenkel, A. J., Li, X., & Elliott, A. M. (2014). The validation and clinical implementation of BRCAplus: A comprehensive high-risk breast cancer diagnostic assay. *PLoS One*, *9*(5), e97408.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biology*, *9*(7), e1001091.

Elborn, J. S. (2016). Cystic fibrosis. *The Lancet*, *388*(10059), 2519–2531.

Garrison, Erik, & Gabor Marth. (2012). *Haplotype-based variant detection from short-read sequencing*. arXiv. http://arxiv.org/abs/1207.3907.

Glascock, J., Sampson, J., Connolly, A. M., Darras, B. T., Day, J. W., Richard Finkel, R., Howell, R., Klinger, K. W., Kuntz, N., Prior, T., & Shieh, P. B. (2020). Revised recommendations for the treatment of infants diagnosed with spinal muscular atrophy via newborn screening who have 4 copies of SMN2. *Journal of Neuromuscular Diseases*, *7*(2), 97–100.

Glascock, J., Sampson, J., Haidet-Phillips, A., Connolly, A., Darras, B., Day, J., Finkel, R., Howell, R. R., Klinger, K., Kuntz, N., Prior, T., Shieh, P. B., Crawford, T. O., Kerr, D., & Jarecki, J. (2018). Treatment algorithm for infants diagnosed with spinal muscular atrophy through newborn screening. *Journal of Neuromuscular Diseases*, *5*(2), 145–158.

Gould, G. M., Grauman, P. V., Theilmann, M. R., Spurka, L., Wang, I. E., Melroy, L. M., Chin, R. G., Hite, D. H., Chu, C. S., Maguire, J. R., Hogan, G. J., & Muzzey, D. (2018). Detecting clinically actionable variants in the 3′ exons of PMS2 via a reflex workflow based on equivalent hybrid capture of the gene and its pseudogene. *BMC Medical Genetics*, *19*(1), 176.

Hayward, B. E., De Vos, M., Valleley, E. M. A., Charlton, R. S., Taylor, G. R., Sheridan, E., & Bonthron, D. T. (2007). Extensive gene conversion at the PMS2 DNA mismatch repair locus. *Human Mutation*, *28*(5), 424–430.

Hayward, B. E., Kumari, D., & Usdin, K. (2017). Recent advances in assays for the fragile X-related disorders. *Human Genetics*, *136*(10), 1313–1327.

Hendrickson, B. C., Donohoe, C., Akmaev, V. R., Sugarman, E. A., Labrousse, P., Boguslavskiy, L., Flynn, K., Rohlfs, E. M., Walker, A., Allitto, B., Sears, C., & Scholl, T. (2009). Differences in SMN1 allele frequencies among ethnic groups within North America. *Journal of Medical Genetics*, *46*(9), 641–644.

Herman, D. S., Smith, C., Liu, C., Vaughn, C. P., Palaniappan, S., Pritchard, C. C., & Shirts, B. H. (2018). Efficient detection of copy number mutations in PMS2 exons with a close homolog. *The Journal of Molecular Diagnostics: JMD*, *20*(4), 512–521.

Hogan, G. J., Vysotskaia, V. S., Beauchamp, K. A., Seisenberger, S., Grauman, P. V., Haas, K. R., Hong, S. H., Jeon, D., Kash, S., Lai, H. H., Melroy, L. M., Theilmann, M. R., Chu, C. S., Iori, K., Maguire, J. R., Evans, E. A., Haque, I. S., Mar-Heyming, R., Kang, H. P., & Muzzey, D. (2018). Validation of an expanded carrier screen that optimizes sensitivity via full-exon sequencing and panel-wide copy number variant identification. *Clinical Chemistry*, *64*(7), 1063–1073.

Hruska, K. S., LaMarca, M. E., Ronald Scott, C., & Sidransky, E. (2008). Gaucher disease: Mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Human Mutation*, *29*(5), 567–583.

Johansson, L. F., van Dijk, F., de Boer, E. N., van Dijk-Bos, K. K., Jongbloed, J. D. H., van der Hout, A. H., Westers, H., Sinke, R. J., Swertz, M. A., Sijmons, R. H., & Sikkema-Raddatz, B. (2016). CoNVaDING: Single exon variation detection in targeted NGS data. *Human Mutation*, *37*(5), 457–464.

Kariyawasam, D. S. T., D'Silva, A., Lin, C., Ryan, M. M., & Farrar, M. A. (2019). Biomarkers and the development of a personalized medicine approach in spinal muscular atrophy. *Frontiers in Neurology*, *10*(August), 898.

Krone, N., & Arlt, W. (2009). Genetics of congenital adrenal hyperplasia. *Best Practice & Research. Clinical Endocrinology & Metabolism*, *23*(2), 181–192.

Kurian, A. W., Hare, E. E., Mills, M. A., Kingham, K. E., McPherson, L., Whittemore, A. S., McGuire, V., Ladabaum, U., Kobayashi, Y., Lincoln, S. E., Cargill, M., & Ford, J. M. (2014). Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *32*(19), 2001–2009.

Li, W., & Freudenberg, J. (2014). Mappability and read length. *Frontiers in Genetics*, *5*(November), 381.

Lincoln, S. E., Hambuch, T., Zook, J. M., Bristow, S. L., Hatchell, K., Truty, R., Kennemer, M., Shirts, B. H., Fellowes, A., Chowdhury, S., Klee, E. W., Mahamdallie, S., Cleveland, M. H., Vallone, P. M., Ding, Y., Seal, S., DeSilva, W., Tomson, F. L., Huang, C., … Nussbaum, R. L. (2021). 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. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *23*, 1673–1680. https://doi.org/10.1038/s41436-021-01187-w

Lincoln, S. E., Kobayashi, Y., Anderson, M. J., Yang, S., Desmond, A. J., Mills, M. A., Nilsen, G. B., Jacobs, K. B., Monzon, F. A., Kurian, A. W., Ford, J. M., & Ellisen, L. W. (2015). A systematic comparison of traditional and multigene panel testing for hereditary breast and ovarian cancer genes in more than 1000 patients. *The Journal of Molecular Diagnostics: JMD*, *17*(5), 533–544.

Luo, M., Liu, L., Peter, I., Zhu, J., Scott, S. A., Zhao, G., Eversley, C., Kornreich, R., Desnick, R. J., & Edelmann, L. (2014). An Ashkenazi Jewish SMN1 haplotype specific to duplication alleles improves pan-ethnic carrier screening for spinal muscular atrophy. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *16*(2), 149–156.

Mandelker, D., Schmidt, R. J., Ankala, A., Gibson, K. M. D., Bowser, M., Sharma, H., Duffy, E., Hegde, M., Santani, A., Lebo, M., & Funke, B. (2016). Navigating highly homologous genes in a molecular diagnostic setting: A resource for clinical next-generation sequencing. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *18*(12), 1282–1289.

Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. *Frontiers in Genetics*, *10*(May), 426.

Marques, I., Sá, M. J., Soares, G., do Céu Mota, M., Pinheiro, C., Aguiar, L., Amado, M., Soares, C., Calado, A., Dias, P., & Sousa, A. B. (2015). Unraveling the pathogenesis of ARX polyalanine tract variants using a clinical and molecular interfacing approach. *Molecular Genetics & Genomic Medicine*, *3*(3), 203–214.

New York State Department of Health. (2014). *Clinical laboratory standards of practice. 2014*. Wadsworth Center. https://wadsworth.org/sites/default/files/WebDoc/1184889505/GETE_June2014.pdf

Nolin, S. L., Glicksman, A., Ersalesi, N., Carl Dobkin, W., Ted Brown, R., Cao, E. B., Sah, S., Latham, G. J., & Hadd, A. G. (2015). Fragile X full mutation expansions are inhibited by one or more AGG interruptions in premutation carriers. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *17*(5), 358–364.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53.

Nykamp, K., Truty, R., Riethmaier, D., Wilkinson, J., Bristow, S. L., Aguilar, S., Neitzel, D., Faulkner, N., & Aradhya, S. (2021). Elucidating clinical phenotypic variability associated with the polyT tract and TG repeats in CFTR. *Human Mutation*, *42*(9), 1165–1172.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. https://doi.org/10.1101/201178

Prior, T. W., Leach, M. E., & Finanger, E. (2000). Spinal muscular atrophy. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, G. Mirzaa, & A. Amemiya (Eds.), *GeneReviews®*. University of Washington, Seattle.

Rehm, H. L., Bale, S. J., Bayrak-Toydemir, P., Berg, J. S., Brown, K. K., Deignan, J. L., Friez, M. J., Funke, B. H., Hegde, M. R., Lyon, E., & Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Commitee. (2013). ACMG clinical laboratory standards for next-generation sequencing. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *15*(9), 733–747.

Retterer, K., Juusola, J., Cho, M. T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K. G., McKnight, D., Bai, R., Suchy, S., Friedman, B., Tahiliani, J., Pineda-Alvarez, D., Richard, G., Brandt, T., Haverfield, E., … Bale, S. (2016). Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *18*(7), 696–704.

Retterer, K., Scuffins, J., Schmidt, D., Lewis, R., Pineda-Alvarez, D., Stafford, A., Schmidt, L., Warren, S., Gibellini, F., Kondakova, A., Blair, A., Bale, S., Matyakhina, L., Meck, J., Aradhya, S., & Haverfield, E. (2015). Assessing copy number from exome sequencing and exome Array CGH based on CNV Spectrum in a large clinical cohort. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *17*(8), 623–629.

Stabley, D. L., Harris, A. W., Holbrook, J., Chubbs, N. J., Lozo, K. W., Crawford, T. O., Swoboda, K. J., Funanage, V. L., Wang, W., Mackenzie, W., Scavina, M., Sol-Church, K., & Butchbach, M. E. R. (2015). SMN1 and SMN2 copy numbers in cell lines derived from patients with spinal muscular atrophy as measured by Array digital PCR. *Molecular Genetics & Genomic Medicine*, *3*(4), 248–257.

Tamary, H., & Dgany, O. (2005). Alpha-Thalassemia. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, G. Mirzaa, & A. Amemiya (Eds.), *GeneReviews®*. University of Washington, Seattle.

Truty, R., Paul, J., Kennemer, M., Lincoln, S. E., Olivares, E., Nussbaum, R. L., & Aradhya, S. (2019). Prevalence and properties of intragenic copy-number variation in mendelian disease genes. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *21*(1), 114–123.

Vaughn, C. P., Hart, K. J., Samowitz, W. S., & Swensen, J. J. (2011). Avoidance of pseudogene interference in the detection of 3' deletions in PMS2. *Human Mutation*, *32*(9), 1063–1071.

Yu, A. C.-S., Yim, A. K.-Y., Chan, A. Y.-Y., Yuen, L. Y. P., Wing Chi, A., Cheng, T. H. T., Lin, X., Li, J.-W., Chan, L. W. L., Mok, V. C. T., Chan, T.-F., & Chan, H. Y. E. (2019). A targeted gene panel that covers coding, non-coding and short tandem repeat regions improves the diagnosis of patients with neurodegenerative diseases. *Frontiers in Neuroscience*, *13*(December), 1324.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rojahn, S., Hambuch, T., Adrian, J., Gafni, E., Gileta, A., Hatchell, H., Johnson, B., Kallman, B., Karfilis, K., Kautzer, C., Kennemer, M., Kirk, L., Kvitek, D., Lettes, J., Macrae, F., Mendez, F., Paul, J., Pellegrino, M., Preciado, R. ... Aradhya, S. (2022). Scalable detection of technically challenging variants through modified next-generation sequencing. *Molecular Genetics & Genomic Medicine*, *10*, e2072. https://doi.org/10.1002/mgg3.2072