ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

Open

Next-generation carrier screening

Mark A. Umbarger, PhD[1], Caleb J. Kennedy, PhD[1], Patrick Saunders, BS[1], Benjamin Breton, BS[1], Niru Chennagiri, MS[1], John Emhoff, BS[1], Valerie Greger, PhD[1], Stephanie Hallam, PhD[1], David Maganzini, BS[1], Cynthia Micale, PhD[1], Marcia M. Nizzari, MS[1], Charles F. Towne, PhD[1], George M. Church, PhD[2] and Gregory J. Porreca, PhD[1]

Purpose: Carrier screening for recessive Mendelian disorders traditionally employs focused genotyping to interrogate limited sets of mutations most prevalent in specific ethnic groups. We sought to develop a next-generation DNA sequencing–based workflow to enable analysis of a more comprehensive set of disease-causing mutations.

Methods: We utilized molecular inversion probes to capture the protein-coding regions of 15 genes from genomic DNA isolated from whole blood and sequenced those regions using the Illumina HiSeq 2000 (Illumina, San Diego, CA). To assess the quality of the resulting data, we measured both the fraction of the targeted region yielding high-quality genotype calls, and the sensitivity and specificity of those calls by comparison with conventional Sanger sequencing across hundreds of samples. Finally, to improve the overall accuracy for detecting insertions and deletions, we introduce a novel assembly-based approach that substantially increases sensitivity without reducing specificity.

Results: We generated high-quality sequence for at least 99.8% of targeted base pairs in samples derived from blood and achieved high concordance with Sanger sequencing (sensitivity >99.9%, specificity >99.999%). Our novel algorithm is capable of detecting insertions and deletions inaccessible by current methods.

Conclusion: Our next-generation DNA sequencing–based approach yields the accuracy and completeness necessary for a carrier screening test.

Genet Med advance online publication 13 June 2013

Key Words: carrier screening; next-generation DNA sequencing

INTRODUCTION

Carrier screening is performed either preconception or during pregnancy to determine a couple's risk of having a child with a recessive genetic disorder. The number of individuals who could benefit from such screening is substantial because roughly 2 million women give birth to their first child each year in the United States[1]. The disorders for which testing is recommended vary based on a number of different patient-specific factors. For instance, the American Congress of Obstetricians and Gynecologists recommends that screening for cystic fibrosis be offered to all women of reproductive age[2] and that testing be performed for additional disorders if indicated by family history, partner's carrier status, or ethnicity.[3–5]

Today, carrier screening is typically performed using focused genotyping technologies that are designed to interrogate specific mutations within a gene of interest. However, because of cost and complexity, these tests often do not include all known disease-causing mutations. By contrast, next-generation DNA sequencing (NGS) can comprehensively genotype a set of genes in a cost-efficient manner and is therefore poised to supplant current technologies for routine, high-volume carrier screening.

For NGS to be used for carrier screening in a clinical setting, it must satisfy at least three requirements. First, analytical accuracy must be both high and well characterized within the clinically relevant genes or regions. Previous reports have demonstrated a broad range of accuracy values, and in some cases it is unclear whether these values hold within the relevant regions of the genome.[6–9] In addition, accuracy for insertions and deletions is generally either substantially lower or uncharacterized, and measured to lower precision. Second, the NGS workflow employed should yield data sufficient to cover the vast majority of targeted bases at a depth sufficient to make high-quality genotype calls (both variant and nonvariant). It has been noted, however, that the percentage of bases callable at a given depth varies widely with both the sample preparation workflow and the total amount of sequencing.[8,10] Finally, the workflow must be highly robust and reproducible, which can often be achieved through automation. However, typical NGS sample preparation workflows are not amenable to high-throughput automation because of rate-limiting mechanical shearing, reaction purifications, size selections, and kitted reagent costs (typically $50–200 per sample).

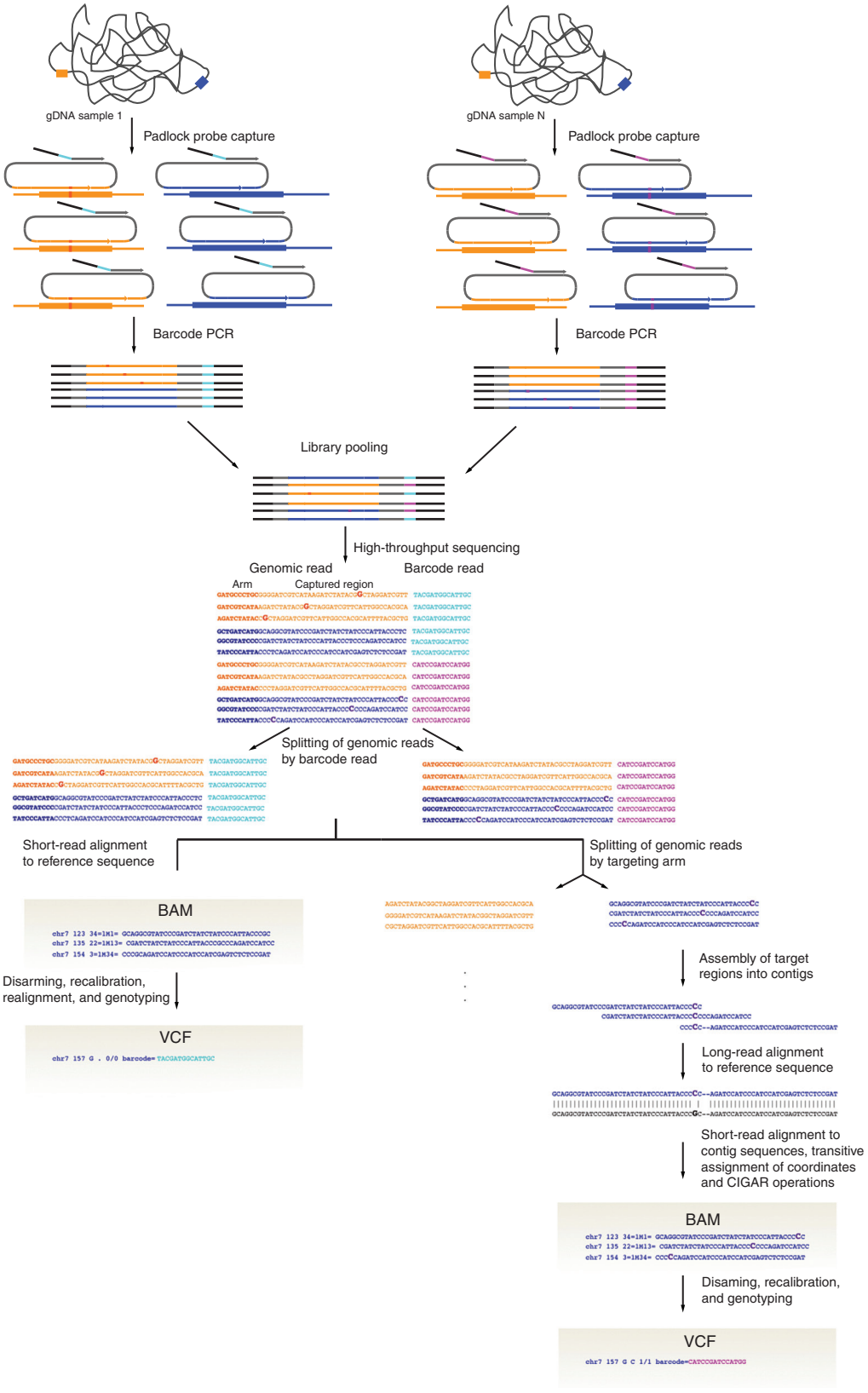
Here, we describe an integrated NGS workflow that meets these requirements for carrier screening. The workflow combines automated, optimized molecular inversion probe target capture with molecular barcoding to maximize the sample throughput of an NGS machine and employs a novel read assembly–based alignment method that enables accurate

The first two authors contributed equally to this work.

[1]Good Start Genetics, Cambridge, Massachusetts, USA; [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. Correspondence: Gregory J. Porreca (gporreca@gsgenetics.com)

Submitted 23 January 2013; accepted 2 May 2013; advance online publication 13 June 2013. doi:10.1038/gim.2013.83

identification of both substitution and insertion/deletion lesions (Figure 1). We apply this workflow to sequence the protein-coding regions of 15 genes in which loss-of-function mutations cause recessive Mendelian disorders often included as part of routine carrier screening, and demonstrate through realistic simulation and comparison with Sanger sequencing data that our approach achieves high accuracies and detects the vast majority of disease-causing mutations.



MATERIALS AND METHODS

Molecular inversion probe design

Molecular inversion probes[11,12] were designed to capture the coding regions and certain well-characterized noncoding regions of 15 genes (see **Supplementary Tables S1 and S2** online). The 5′ and 3′ targeting arms (extension and ligation, respectively) comprised a total of 40 nucleotides and were designed to flank 130-bp target regions. Further details can be found in the **Supplementary Materials and Methods** online.

Target capture, barcoding, and NGS

Genomic DNA was purchased from the Coriell Cell Repositories (Camden, NJ) or isolated from whole blood by the Gentra Puregene method (Qiagen, Gaithersburg, MD) concluding with an overnight incubation at 65 °C. All samples were considered "IRB Exempt" by Liberty IRB, our independent institutional review board. Genomic DNA was subjected to multiplex target capture using molecular inversion probes. Captured product was subjected to PCR to attach molecular barcodes in a manner that allowed sequencing from either end of the captured region.[12] The PCR product was pooled and sequenced on the Illumina HiSeq 2000. Further details can be found in the **Supplementary Materials and Methods** online.

NGS data analysis with alignment-only algorithm

Raw .bcl files were converted to qseq files using bclConverter (Illumina). Fastq files were generated by "debarcoding" genomic reads using the associated barcode reads; reads for which barcodes yielded no exact match to an expected barcode, or contained one or more low-quality base calls, were discarded. The remaining reads were aligned to hg18 on a per-sample basis using Burrows-Wheeler Aligner version 0.5.7 for short alignments,[13] and genotype calls were made using Genome Analysis Toolkit version 1.0.4168 after base-quality score recalibration, realignment (with GATK version 1.0.5083),[14] and targeting arm removal (to prevent synthetic, reference-designed molecular inversion probe arm sequence from interfering with genotype calling). High-confidence genotype calls were defined as having depth ≥50 and strand bias score ≤0. Clinical significance of variant calls was determined by matching against a VCF-formatted database of disease-causing mutations curated from the literature, with equivalent insertion/deletion regions calculated as previously described.[15]

NGS data analysis with Genotyping by Assembly-Templated Alignment algorithm

Debarcoded fastq files were obtained as described above and partitioned by capture region (exon) using the target arm sequence as a unique key. Reads were assembled in parallel by exon using SSAKE version 3.7 with parameters "-m 30 -o 15".[16] The resulting contiguous sequences (contigs) were aligned to hg18 using BWA version 0.5.7 for long alignments[17] with parameter "-r 1". Short-read alignment was performed as described above, except that sample contigs (rather than hg18) were used as the input reference sequence. Software was developed in Java to accurately transfer coordinate and variant data (gaps) from local sample space to global reference space for every BAM-formatted alignment. Genotyping and base-quality recalibration were performed on the coordinate-translated BAM files using GATK version 1.6.5.

Sanger sequencing

PCR was carried out with the genomic DNA described in the "Target capture, barcoding, and NGS" section using a modified version of the protocol from Zimmerman et al.[18] and using PCR primers from Jones et al.[19] with M13 tails removed (regions in **Supplementary Table S3** online). More information can be found in the **Supplementary Materials and Methods** online.

Sanger data analysis and cross-validation to NGS

Mutation Surveyor software (MS; Softgenetics, State College, PA) version 4.0.5 was used in batch-mode with default parameters to align ab1 files to target reference sequence and make genotype calls. Positions at which MS base calls did not match in the forward and reverse directions were removed from consideration. All high-quality NGS genotype calls (both reference and non-reference) within 10 bp (inclusive) of target exons were subjected to cross-validation against VCF-converted MS variant calls and orthogonal confirmation, if necessary (see **Supplementary Figure S1** online). A detailed description of cross-validation to NGS is provided in the **Supplementary Materials and Methods** online.

Assessment of detectability of clinical mutations by simulation

A total of 145 Coriell samples were sequenced and analyzed by Genotyping by Assembly-Templated Alignment (GATA, described above). Specific fields (base sequence and qualities) within aligned reads (BAM records) from the Illumina sequencer were manipulated *in silico* to introduce the clinically

Figure 1 Next-generation DNA sequencing workflow. Genomic DNA samples are input to a molecular inversion probe capture reaction. Each target (depicted by blue and orange regions) is captured by multiple probes that anneal to nonoverlapping genomic intervals. PCR is performed using primers containing patient-specific barcodes, yielding barcode libraries (turquoise and purple). Equal volumes of the libraries are pooled and enter the Illumina HiSeq high-throughput sequencing workflow. Following sequencing, reads enter either the alignment only (AO, left) or Genotyping by Assembly-Templated Alignment (GATA, right) analysis pipeline. AO first partitions reads by sample molecular barcode, then in parallel for all samples performs short-read alignment, base-quality recalibration, realignment around putative indels, and genotyping. GATA partitions reads first by sample molecular barcode, then by target. Reads are assembled into contiguous sequences (contigs) that are then aligned to the reference genome. Raw reads are then aligned to the contigs, and raw-read mapping and variant information relative to the reference are determined using reference-contig and read-contig alignments. Finally, base-quality score recalibration and genotyping are performed on the mapped, raw reads. gDNA, genomic DNA.

relevant DNA lesion pattern within the context of empirical sequencing errors and sample-specific proximal variants. To simulate heterozygous carriers, input reads covering the mutation were chosen at random for sequence manipulation with an average probability of 0.5. All reads, whether manipulated or not, were output in fastq format for subsequent GATA analysis as described. This process was repeated for each of 81 clinically significant mutations whereupon genotyped (observed) alleles were cross-referenced back to the original simulated (expected) allele. Samples for which the allele was already present were excluded from simulation (e.g., many Coriell samples in the set contained the common CFTR F508del mutation). Mutations with detection rates <100% between the expected and observed alleles were classified as undetectable by NGS.

Determining clinical significance of variant allele calls

Each NGS-detected variant allele is annotated for functional (clinical) significance by determining its relative position within the corresponding consensus coding sequence. Truncating mutations were considered clinically significant lesions for the purposes of the work presented here. A more detailed description of how clinical significance of variant allele calls was determined can be found in the **Supplementary Materials and Methods** online.

RESULTS

Completeness and reproducibility

We performed automated target capture and molecular barcoding followed by NGS on a set of 194 samples derived from immortalized cell lines (55 containing specific disease-causing mutations and 139 chosen to represent ethnic diversity, see **Supplementary Figure S2** online) and 59 samples derived from whole blood (see **Supplementary Table S4** online). We targeted all exons including 10 nt of flanking intronic sequence, plus additional intronic regions known to contain disease-causing mutations in 15 genes causative of 14 recessive Mendelian diseases (see **Supplementary Table S1** online) using tiling molecular inversion probes (see Materials and Methods section). A total of 25,907,612,945 base pairs (bp) of de-multiplexed sequence were generated, corresponding to an average per-base coverage per sample of 2,399× (minimum: 891×; maximum: 4,000×), see **Supplementary Table S4** online. Of the 42,858 bases targeted for capture in each sample, we made high-confidence genotype calls (both reference and non-reference) at an average of 97.3% (minimum: 92.2%; maximum: 99.8%) for cell line-derived DNA and 99.9% (minimum: 99.8%; maximum: 99.9%) for blood-derived DNA (see **Supplementary Table S2** online).

The DNA extraction protocol used for our blood samples concluded with an overnight incubation at 65 °C in a Tris-based buffer. Subsequent experiments showed that this step reduced the mean size of the purified DNA (see **Supplementary Figure S3** online); shearing was likely caused by acid hydrolysis during a temperature-induced pH shift of the buffer.[20] We hypothesize that lower-molecular-mass genomic DNA is more readily denatured and therefore more accessible to molecular inversion probes, resulting in improved capture performance. Consistent with this hypothesis, we find that reducing the overnight incubation temperature to 25 °C significantly reduces the percentage of target bases that yield high-confidence genotype calls (see **Supplementary Figure S3** online). To assess reproducibility, a subset of 126 samples derived from cell-line DNA (see **Supplementary Table S4** online) was processed twice, each time by a different operator on different liquid-handling equipment. At least 92% of bases were called at ≥50× coverage in all samples, with high agreement between replicates (Pearson correlation coefficient 0.868). Of 5,177,206 total genotype calls compared, 17 were discordant, for a concordance rate of 0.999997. These occurred at only five unique genomic positions, consistent with systematic sequencing error as the primary cause.

Sanger concordance

To assess the overall accuracy of our NGS genotype calls (reference and non-reference) on a set of 194 samples (see **Supplementary Table S4** online), we compared genotype calls for the target region from the NGS pipeline with those generated by automated analysis (Mutation Surveyor) of bidirectional Sanger sequence of PCR amplicons. Within a total of 6,997,906 bp of sequence called by both methods, we observed 3,973 concordant and 1,220 discordant single-nucleotide variant (SNV) genotype calls. We performed a manual review of discrepant calls and ultimately determined that nine high-quality SNV calls were true discrepancies, corresponding to eight NGS false positives and one NGS false negative (Table 1, see **Supplementary Note 1** online). We observed a total of 4,000 true-positive SNV calls and 6,992,746 true-negative SNV calls (Table 1). The NGS SNV false-positive rate was 1.14×10^{-6} (95% Wilson binomial confidence interval

Table 1 Comparison of NGS genotype calls (alignment-only algorithm) to Sanger-derived genotype calls

			TP	FP	FN	TN
SNV	Heterozygous	dbSNP	2,495	0	1	
		Not dbSNP	247	8	0	
	Homozygous	dbSNP	1,245	0	0	6,992,746
		Not dbSNP	13	0	0	
	Unique		231	3	1	
Indel	Total		61	396	3	
	Unique		17	27	2	6,992,358
	Known		31	–	0	

FN, false-negative calls (reference NGS, non-reference Sanger); FP, false-positive calls (non-reference NGS, reference Sanger); Indel, insertion and deletion; NGS, next-generation DNA sequencing; SNV, single-nucleotide variant; TN, true-negative calls (reference NGS, reference Sanger); TP, true-positive calls (non-reference NGS, non-reference Sanger).

Sanger genotype calls were considered truth. dbSNP membership was determined relative to version 129. Indel calls were considered unique if they differed by sequence pattern or equivalence region. Known indels are disease-causing mutations present in previously annotated samples.

[5.80×10^{-7} – 2.26×10^{-6}]). Positive predictive value for common (i.e., in dbSNP) SNV calls was 100%; it was 97% for novel (i.e., not in dbSNP) calls. The false-positive calls occurred at five unique genomic loci, three of which were at adjacent positions in a single exon of the gene *MCOLN1* and were caused by GATK realignment.

The NGS SNV false-negative rate was 2.50×10^{-4} (95% Wilson binomial confidence interval [1.28×10^{-5}–1.41×10^{-3}]. Sensitivity for common SNV calls was 99.97% (95% CI: 99.9–100%), and sensitivity was 100% (95% CI: 98.6–100%) for novel calls. The false-negative call observed occurred in chromosome 11 of a sample previously characterized as aneuploid.[21] Of 473 NGS reads covering the false-negative locus, 9.5% supported the correct heterozygous A/C genotype call (**Figure 2a**), with Sanger sequencing showing low peak height for the alternate A allele (**Figure 2b**). Shotgun full-genome sequencing of this sample demonstrated a bimodal distribution of allele ratios for heterozygous calls in chromosome 11 (**Figure 2c**) and illustrated variable chromosome copy numbers (**Figure 2d**), supporting the conclusion that this sample was aneuploid.

For insertions and deletions (indels), we observed a total of 61 true positives, 394 false positives (27 unique alleles), and 3 false negatives (2 unique alleles, both in exon 1 of *SMPD1*), for a sensitivity of 95.3% overall (95% CI: 86.9–99.0%), and 100% if exon 1 of *SMPD1* is excluded from analysis (95% CI: 94.1–100%). Of 31 clinically relevant disease mutations, we detected all 31, for a sensitivity at disease-causing loci of 100% (95% CI: 88.8–100%).

Detection of pathogenic mutations

We next sought to assess our ability to detect variants that cause the Mendelian diseases targeted by our panel (see **Supplementary Table S1** online) in a set of 194 cell line–derived samples. Of these samples, 55 were derived from individuals who were either carriers of or affected by one of the diseases being assayed and collectively contained a total of 95 previously characterized disease mutations. During the design of our NGS workflow, we determined that three of these lesions would be inaccessible by our approach—two were large deletions spanning multiple exons, and one was within a region of *CFTR* exon 10 that is paralogous to other genomic regions (see **Supplementary Table S5** online). Of the 92 mutations we could expect to detect by NGS, we detected all 92 (see **Supplementary Table S5** online), and coverage-based analyses of the regions harboring the two large deletions illustrated that evidence of these lesions is present in the sequencing data (see **Supplementary Figure S4** online). We also identified truncating (and likely disease-causing) mutations in two affected samples in which previously only one mutation was known (see **Supplementary Figure S5** online and **Supplementary Table S5** online), as well as 9 carriers in the set of 139 previously uncharacterized

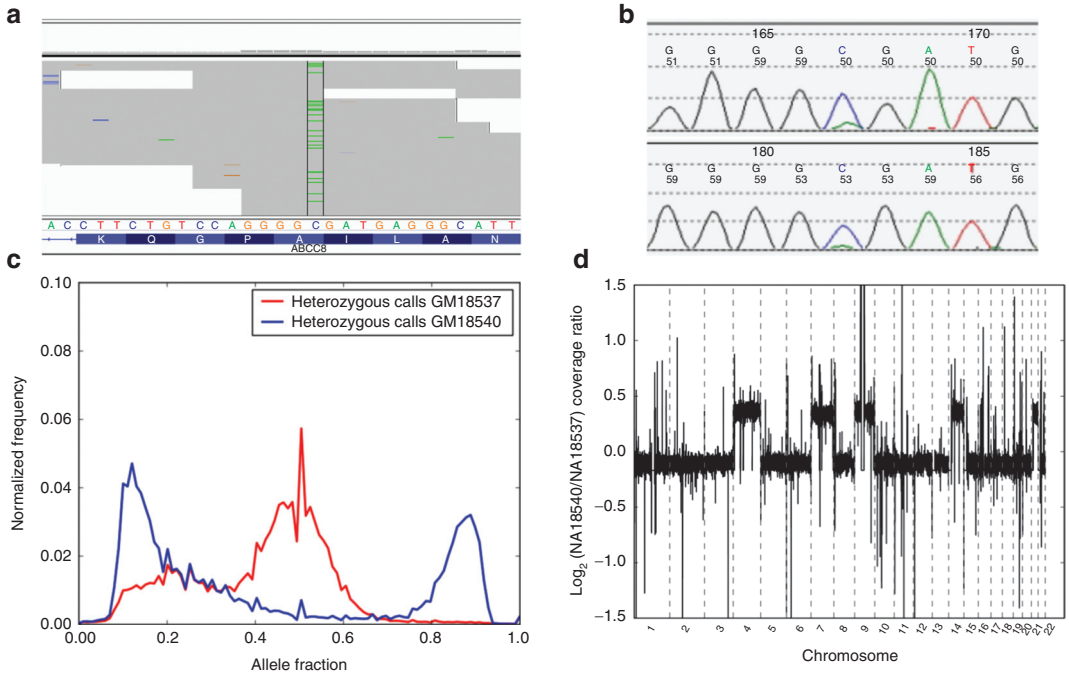


Figure 2 Skewed allelic fractions in aneuploid cell-line GM18540. (**a**) IGV view of NGS data from GM18540 for the non-reference genotype call of interest (shown between vertical lines). (**b**) Bidirectional Sanger data for the variant-containing region. (**c**) Histogram of allele ratios for all non-reference genotype calls in chromosome 11 derived from whole-genome shotgun sequencing (WGSS) of GM18540 and control sample GM18537. (**d**) Genome-wide relative coverage for GM18540. WGSS coverage data for each of the autosomes was binned into 50 kb intervals and the log-ratios of the per-sample mean normalized values were plotted versus chromosome position. Dashed vertical lines denote chromosome boundaries; within a chromosome the ratios are arranged according to genomic position. IGV, integrative genomics viewer; NGS, next-generation DNA sequencing.

HapMap, 1000 Genomes Project, and Human Diversity Panel samples (see **Supplementary Table S5** online).

GATA

Although substitutions comprise the majority of coding variation in the human genome, indels are often clinically relevant. Indels, especially when large or present *in cis* with substitutions, are notoriously difficult to detect with short NGS reads. Assembly of short reads can improve indel detection sensitivity, but this is often at the cost of decreased SNV and indel specificity due to the presence of spurious contigs. We devised an algorithm, termed GATA, that first forms an assembly from reads partitioned into subsets by targeting arm sequence and then performs base quality– and coverage-informed genotyping by alignment of raw reads back to the assembled contigs (**Figure 1**). We compared the performance of GATA for indel genotyping to the more conventional genotyping-by-alignment-only (AO) algorithm used in the Sanger concordance studies. Across a set of 147 samples analyzed, both indel sensitivity and specificity were increased with GATA relative to AO (**Table 2**). GATA detected 23 unique insertions and deletions, which were confirmed by manual review of Sanger traces. Of these, nine (39%) were not detected by AO in one or more samples, including *BLM* c.2207_2212delinsTAGATTC—the most common disease-causing mutation for Bloom syndrome in people of Ashkenazi Jewish descent[22]—as well as several alleles in *SMPD1* (see **Supplementary Table S6** online), the gene associated with Niemann–Pick disease (**Figure 3**). Performance for substitutions was identical for both detection methods (AO and GATA). GATA and AO both utilize GATK; however because versions and functionality differ, we assessed whether the newer (GATA) version of GATK, when used with AO, would improve indel performance. Although this eliminates false-positive indel calls, sensitivity is only marginally improved (**Table 2**).

Table 2 Genotyping by Assembly-Templated Alignment (GATA) algorithm improves detection of insertions and deletions

	AO	AO/GATK165	GATA
TP	104	116	211
FP	28	0	0
FN	47	46	0
Uncalled[a]	70	59	10
Sensitivity	0.689	0.716	1.0
Precision	0.788	1.0	1.0

FN, false-negative; FP, false-positive; NGS, next-generation DNA sequencing; TP, true-positive.

Raw variant alleles (positive calls) from 147 samples were filtered by depth and strand bias (for AO/GATK165 and GATA) and categorized according to NGS data analysis method, alignment only (AO), AO/GATK165, which utilized version 1.6.5 of GATK that was also used for GATA, or GATA. We first classified variant calls using calls automatically generated by Mutation Surveyor (MS) from Sanger traces. Next, for calls that were discordant between MS and either AO or GATA, we manually reviewed the Sanger data to resolve these calls as AO or GATA TP/FP/FN shown below. Variant calls flagged as low confidence (depth <50 or strand bias <0) were considered uncalled.

[a]Polymorphisms in the first exon of *SMPD1* accounted for the majority of uncalled and discordant alleles, which were not considered in accuracy calculations.

Simulation to assess detectability of rare pathogenic mutations

Although we were able to empirically demonstrate detectability for all disease-causing mutations present in our sample set, there exist a number of disease-causing mutations for which samples cannot be readily obtained. To assess whether our NGS workflow can detect these additional mutations, we sought to perform simulations *in silico*. Because detectability can be affected by any element of the workflow, we implemented a simulator that employed read sets from actual samples rather than model reads derived from the reference genome at uniform coverage. This allowed for realistic representation of target abundance distribution, neighboring *in cis* variants, as well as cycle- and context-dependent sequencing errors. Disease-causing variants were introduced into raw reads by a Bernoulli process, with an average 0.5 probability of introducing the lesion, to simulate the heterozygous genotypes carrier screening aims to detect. A total of 81 heterozygous variants were simulated in a read set of at least 144 samples, with the exception of *CFTR* c.1521_1523delCTT (F508del), the most common disease-causing mutation for cystic fibrosis in Caucasian populations[23,24] (see **Supplementary Table S7** online). This mutation was present in several samples, which were removed from simulation analysis (see Materials and Methods section). Of the simulated variants, 67 (83%) were correctly genotyped in all (generally 145 of 145) samples and only four relatively large (>7 bp) deletions were undetected in one or more samples (see **Supplementary Table S7** online). We were unable to make high-confidence genotype calls for the remaining 10 variants. We did not find any variants to be undetectable in all samples (see **Supplementary Table S7** online).

DISCUSSION

Robustness, completeness, and accuracy are three of the main factors that define the utility of a genetic carrier testing workflow in a clinical laboratory. By utilizing a target enrichment methodology that is performed in a single tube and requires no mechanical shearing or purifications of individual samples, we have developed an automated NGS workflow that yields highly reproducible results across samples and operators. This reproducibility ensures that samples will not have to be rerun frequently, minimizing both turnaround time and per-sample cost.

Because each clinically meaningful base pair must be sequenced before an actionable medical report can be generated, a high level of completeness minimizes the amount of costly rework necessary for a sample. We have demonstrated completeness consistent with low to no rework for the samples and target set studied here, and substantially better than other previously reported methods using multiplex target capture or PCR with NGS.[8,11,12,25] This improvement is probably the result of a number of optimizations

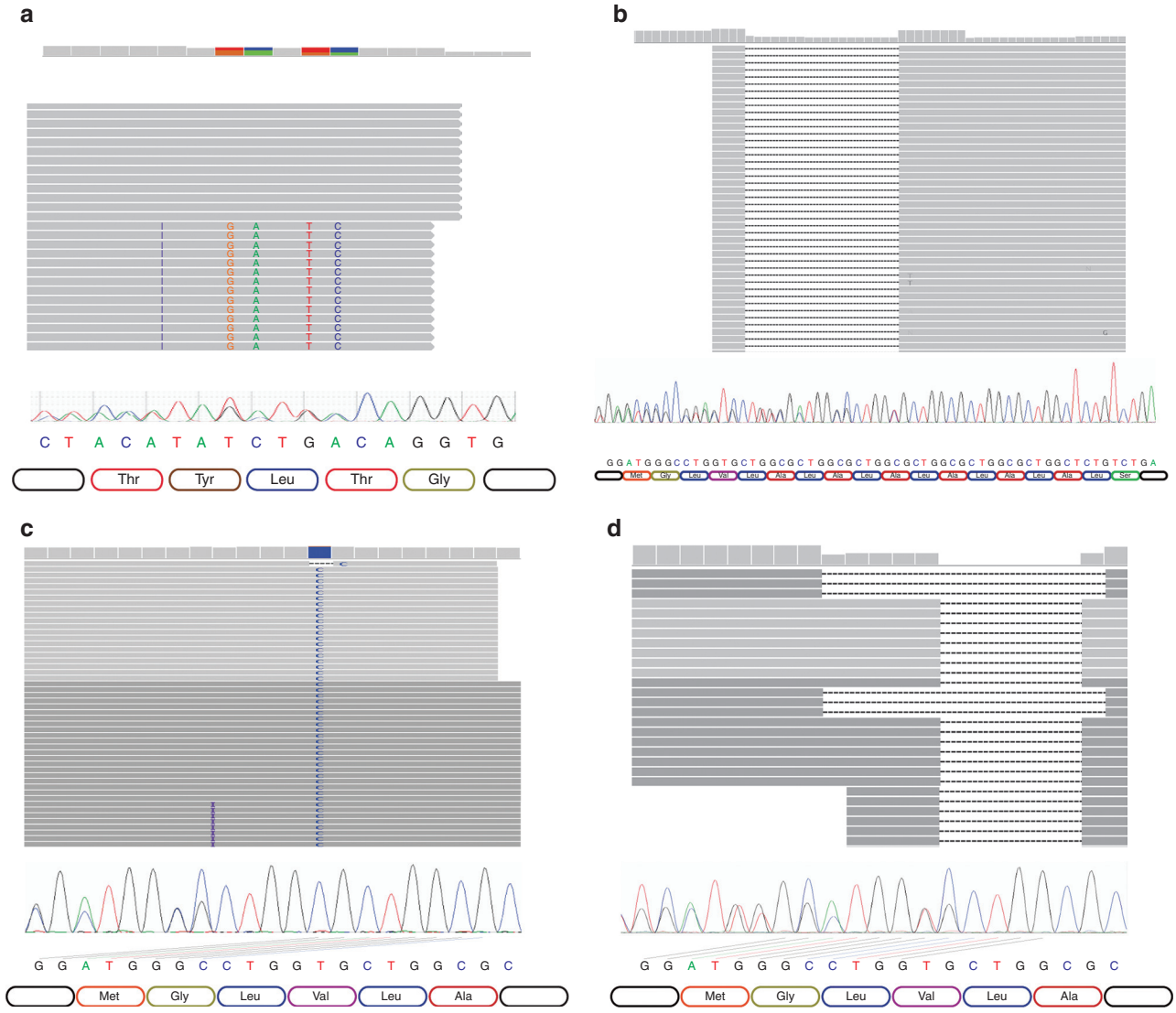


Figure 3 Genotyping by Assembly-Templated Alignment correctly genotypes insertions and deletions that are undetectable by the alignment-only method. Read from top to bottom, each panel provides tracks for cumulative depth of coverage (vertical gray bars); representative MIP alignments (horizontal gray bars) with mismatches (colored letters), insertions (purple bars), and gaps (dashed lines); chromatogram; reference DNA and amino acid sequence for (**a**) heterozygous *BLM* c.2207_2212delinsTAGATTC in sample GM04408 as well as several alleles in the first exon of *SMPD1* including (**b**) a heterozygous 18-bp deletion in sample GM20342 (minus strand), (**c**) a heterozygous 12-bp insertion and homozygous substitution in sample GM17282 (plus strand), and (**d**) compound heterozygous 6- and 12-bp deletions in sample GM00502 (minus strand). Chromatogram trace offsets corresponding to specific heterozygous insertion and deletion patterns are indicated with slanted lines color coded by reference base. For clarity, offsets are shown for (**c**) and (**d**) only. MIP, molecular inversion probe.

we have made, including the use of a tiling MIP design that ensures multiple probes capture each base, thereby reducing the probability of allele dropout and systematic sequencing error, and the use of a DNA extraction protocol that effortlessly shears the DNA to a lower molecular mass. It should be noted, however, that performance with other sets of genes may vary due to interfering effects such as extreme GC content (see **Supplementary Figure S6** online), repetitive or low-complexity regions, and paralogous sequence.

Regarding accuracy, the only SNV false negative that we observed was in a sample that exhibited skewed allele ratios along the chromosome, which should not commonly occur when testing for germline mutations in clinical specimens derived from whole blood. In addition, the SNV false-positive rate of ~1.1 per million base pairs corresponds to a low confirmation burden for clinical testing and surpasses values previously reported. Given our small target set and the rare nature of indels, it is difficult to provide a precise measurement of our accuracy for indels, although our data do suggest that the use of GATA substantially improves our ability to detect small lesions. The high level of coverage used in this study probably had a positive effect on both the accuracy and completeness achieved. Although lower levels of coverage could be used, it would be reasonable to expect that this

would lead to either lower accuracy (ability to detect heterozygous lesions) or completeness, and might be unnecessary because list sequencing reagent costs here approximate those for genotyping arrays, although capital costs may be substantially different.

It is worth noting that measuring accuracy to a sufficient level of precision and generality can be challenging within conserved coding regions because selective pressure limits the spectrum of variation present. Although we sequenced a large number of samples, the relatively small size of our target limited the number of unique alleles observable and meant that ~90% of such variants were common (i.e., present in dbSNP). Nonetheless, there is no *a priori* reason to believe that our measured accuracy will not generalize to other rare and private mutations present in the targeted loci. Supporting this point, our simulations using real data and controlled for sample-to-sample variability indicate that we can detect a number of very rare disease-causing alleles of different types and sequence contexts, including insertions (up to 12 bp), deletions (up to 22 bp), and complex combinations thereof.

The reference standard one considers ground truth can impose a ceiling on measurable accuracy. We employed a large-scale automated analysis of what is widely deemed the "gold standard" for DNA sequencing: bidirectional Sanger traces derived from PCR amplicons.[26] The NGS workflow detected allele dropout in the Sanger data, a known limitation of that technology (see **Supplementary Figure S7** online)[27] and not surprising because each base sequenced by NGS was captured by multiple probes with independent targeting arms. Had we instead employed the less laborious and more commonly used reference of HapMap Project genotyping data, we would have observed 12 NGS false negatives and 7 false positives in the subset of samples characterized by this approach (see **Supplementary Table S8** online). Because these were all shown by our Sanger analysis to be HapMap Project genotyping errors, this would have underestimated both sensitivity and specificity.

Indel detection methods that only employ gapped alignment of short reads to reference are often limited by false positives introduced by systematic, context-dependent sequencing error, and false negatives introduced by failure of the aligner to open or extend gaps. An assembly-based paradigm would address these limitations, but raw contigs do not always carry base-quality and coverage information. The GATA algorithm combines these approaches to deliver sensitive and specific indel detection with SNV performance on par with a traditional AO pipeline.

Many alleles detected exclusively by GATA were from a short-tandem-repeat region encoding the N-terminal signal peptide in *SMPD1* (see **Supplementary Table S6** online). Consistent with previous reports, GATA detected non-reference alleles in 96% of samples, a rate that is strikingly high because hg18 contains a minor allele that is frequently substituted (V36A). Although common hexanucleotide indels at this locus are clinically benign,[28,29] any pathogenic mutation present *in cis* would probably be missed using a conventional approach for variant detection. Indeed, when reads were aligned independently, several genomic positions in this region consistently fell below our specified coverage threshold. GATA therefore should yield higher sensitivity for rare mutations linked to polymorphisms in the first exon of *SPMD1* and potentially other short-tandem-repeat loci as well.

The simulation methodology applied here attempts to assess detectability of rare pathogenic mutations in a highly realistic manner. Simply deriving reads from a reference genome modified to include the mutation of interest can overestimate the detection probability because of real-world factors that would otherwise render the mutation undetectable. In addition, we are able to determine whether a mutation is sometimes, rather than always or never, detectable because it is simulated in the read sets of hundreds of samples; for example, this could occur in a particular genetic background with a low-frequency *in cis* variant that interferes with alignment of reads containing the mutation. Nonetheless, certain mutation types, particularly large deletions, are still not amenable to this paradigm because they could fundamentally alter the distribution of reads generated across the relevant region. In these cases, either human samples or synthetic templates remain the only way to assess detectability. Furthermore, the outcome of this analysis can feed back into the probe design phase if it identifies genes or regions that harbor variants that will be difficult to accurately detect.

We have presented an automated, integrated workflow that converts human genomic DNA isolated from blood or cell lines into clinically relevant variant calls within 6 days. We achieved high genotype concordance with conventional electrophoretic sequencing across a set of 15 genes, and demonstrated the ability to detect a range of important disease-causing mutations. Our new analysis pipeline allows for sensitive and specific detection of indels, while simultaneously incorporating raw base quality and coverage into SNV (non-reference) genotype calls. Realistic simulation on actual run data indicates that a number of pathogenic mutations undetectable by a traditional alignment-based genotyping approach are accessible by GATA. Collectively, the data presented here indicate that our workflow has met our three requirements for a carrier screening test, and hence is ready for clinical use.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

DISCLOSURE

M.A.U., C.J.K., P.S., B.B., N.C., J.E., V.G., S.H., D.M., C.M., M.M.N., C.F.T., and G.J.P. are or were employees of, and shareholders in, Good Start Genetics. M.A.U., C.J.K., B.B., V.G., M.M.N., C.F.T., G.M.C., and G.J.P. are listed as inventors on patents or patent applications related to this work. G.M.C. is an adviser to, and shareholder in, Good Start Genetics.

REFERENCES

1. Central Intelligence Agency. *The World Factbook 2012.* Central Intelligence Agency: Washington, DC, 2012.
2. ACOG Committee Opinion No. 486. Update on carrier screening for cystic fibrosis. *Obstet Gynecol* 2011;117(4):1028–1031.
3. ACOG committee opinion No. 432. Spinal muscular atrophy. *Obstet Gynecol* 2009;113(5):1194–1196.
4. ACOG Committee Opinion No. 442. Preconception and prenatal carrier screening for genetic diseases in individuals of Eastern European Jewish descent. *Obstet Gynecol* 2009;114(4):950–953.
5. ACOG Committee Opinion No. 469. Carrier screening for fragile X syndrome. *Obstet Gynecol* 2010;116(4):1008–1010.
6. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53–59.
7. Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 2010;327:78–81.
8. Bell CJ, Dinwiddie DL, Miller NA, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* 2011;3(65):65ra4.
9. Rothberg JM, Hinz W, Rearick TM, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011;475:348–352.
10. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010;7:111–118.
11. Porreca GJ, Zhang K, Li JB, et al. Multiplex amplification of large sets of human exons. *Nat Methods* 2007;4:931–936.
12. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods* 2009;6:315–316.
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
14. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–498.
15. Krawitz P, Rödelsperger C, Jäger M, Jostins L, Bauer S, Robinson PN. Microindel detection in short-read sequence data. *Bioinformatics* 2010;26:722–729.
16. Warren RL, Sutton GG, Jones SJ, Holt RA. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 2007;23:500–501.
17. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589–595.
18. Zimmerman RS, Cox S, Lakdawala NK, et al. A novel custom resequencing array for dilated cardiomyopathy. *Genet Med* 2010;12:268–278.
19. Jones S, Zhang X, Parsons DW, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 2008;321:1801–1806.
20. Tris-Cl. *Cold Spring Harbor Protocols*. 2006;2006(1):pdb.rec8063. http://cshprotocols.cshlp.org/content/2006/1/pdb.rec8063.short. Accessed 1 January 2012.
21. Locke DP, Sharp AJ, McCarroll SA, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet* 2006;79:275–290.
22. Li L, Eng C, Desnick RJ, German J, Ellis NA. Carrier frequency of the Bloom syndrome blmAsh mutation in the Ashkenazi Jewish population. *Mol Genet Metab* 1998;64:286–290.
23. Estivill X, Bancells C, Ramos C. Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. The Biomed CF Mutation Analysis Consortium. *Hum Mutat* 1997;10:135–154.
24. Watson MS, Cutting GR, Desnick RJ, et al. Cystic fibrosis population carrier screening: 2004 revision of American College of Medical Genetics mutation panel. *Genet Med* 2004;6:387–391.
25. Gowrisankar S, Lerner-Ellis JP, Cox S, et al. Evaluation of second-generation sequencing of 19 dilated cardiomyopathy genes for clinical applications. *J Mol Diagn* 2010;12:818–827.
26. Schrijver I, Aziz N, Farkas DH, et al. Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the Association for Molecular Pathology. *J Mol Diagn* 2012;14:525–540.
27. Quinlan AR, Marth GT. Primer-site SNPs mask mutations. *Nat Methods* 2007;4:192.
28. Wan Q, Schuchman EH. A novel polymorphism in the human acid sphingomyelinase gene due to size variation of the signal peptide region. *Biochim Biophys Acta* 1995;1270:207–210.
29. Dastani Z, Ruel IL, Engert JC, Genest J Jr, Marcil M. Sphingomyelin phosphodiesterase-1 (SMPD1) coding variants do not contribute to low levels of high-density lipoprotein cholesterol. *BMC Med Genet* 2007;8:79.