© American College of Medical Genetics and Genomics

ARTICLE Genetics inMedicine

Open

Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing

Wenbo Mu, MS[1], Bing Li, PhD[1], Sitao Wu, PhD[1], Jefferey Chen, MS[1], Divya Sain, PhD[1], Dong Xu, PhD[1], Mary Helen Black, PhD[1], Rachid Karam, MD, PhD[1], Katrina Gillespie, BSc[1], Kelly D. Farwell Hagman, MS[1], Lucia Guidugli, PhD[1], Melissa Pronold, PhD[1], Aaron Elliott, PhD[1] and Hsiao-Mei Lu, PhD[1]

Purpose: Structural variation (SV) is associated with inherited diseases. Next-generation sequencing (NGS) is an efficient method for SV detection because of its high-throughput, low cost, and base-pair resolution. However, due to lack of standard NGS protocols and a limited number of clinical samples with pathogenic SVs, comprehensive standards for SV detection, interpretation, and reporting are to be established.

Methods: We performed SV assessment on 60,000 clinical samples tested with hereditary cancer NGS panels spanning 48 genes. To evaluate NGS results, NGS and orthogonal methods were used separately in a blinded fashion for SV detection in all samples.

Results: A total of 1,037 SVs in coding sequence (CDS) or untranslated regions (UTRs) and 30,847 SVs in introns were detected and validated. Across all variant types, NGS shows 100% sensitivity and 99.9% specificity. Overall, 64% of CDS/UTR SVs were classified as pathogenic/likely pathogenic, and five deletions/duplications were reclassified as pathogenic using breakpoint information from NGS.

Conclusion: The SVs presented here can be used as a valuable resource for clinical research and diagnostics. The data illustrate NGS as a powerful tool for SV detection. Application of NGS and confirmation technologies in genetic testing ensures delivering accurate and reliable results for diagnosis and patient care.

Genetics in Medicine (2019) 21:1603–1610; https://doi.org/10.1038/s41436-018-0397-6

Keywords: structural variation; CNV; next-generation sequencing; aCGH; genetic diagnostic testing

INTRODUCTION

Structural variations (SVs) are defined as deletions, duplications, inversions, insertions, and translocations of DNA segments ranging from single exons to larger genomic regions. Deletions and duplications with >1 kb DNA are also known as copy-number variations (CNVs).[1] Because SV modulates cellular activities by changing gene copy number, transcription activity, and chromatin structure,[2] and increases cancer risk,[3–5] clinical diagnostic laboratories perform SV detection in a wide range of hereditary cancer susceptibility genes.[6–9] Traditional methods such as microarray-based comparative genomic hybridization (aCGH) and multiplex ligation-dependent probe amplification (MLPA) have been used for SV detection for the past two decades. However, both methods have limitations. aCGH relies on hybridization of genomic DNA with short probes. Repetitive elements, GC-rich regions, and pseudogenes interfere with hybridization and subsequently reduce SV detection accuracy.[10] In addition, the accuracy, resolution, coverage, and cost of aCGH largely rely on probe density. High-density aGCH array has better performance but is much more expensive. MLPA is a low-throughput multiplex polymerase chain reaction (PCR)-based method, which is not suitable for high-throughput screening.[11]

In recent years, development of target enrichment methods and bioinformatics solutions has made NGS a better choice for high-throughput genetic testing. Concurrently analyzing all types of genomic variants on a single platform at a reduced cost provides clinical laboratories with the opportunity to evaluate NGS performance in SV detection. Many bioinformatics algorithms, such as paired-end mapping (PEM) based, split read (SR), read depth (RD), and de novo assembly (AS) methods, have been developed to identify SVs.[12] Several methods such as XHMM,[13] CoNIFER,[14] cnvCapSeq,[15] VisCap,[16] and DECoN[17] have been specifically implemented to analyze target captured or exome sequencing data. However, some of these methods have limited applications for clinical

[1]Ambry Genetics, Aliso Viejo, CA 92656, USA. Correspondence: Hsiao-Mei Lu (hmlu@ambrygen.com)
These authors contributed equally: Wenbo Mu, Bing Li

Submitted 26 May 2018; accepted: 28 November 2018
Published online: 19 December 2018

Table 1 NGS multigene cancer panel gene lists

Cancer panel	# of genes	Gene list
Breast cancer	17	*ATM*, *BARD1*, *BRCA1*, *BRCA2*, *BRIP1*, *CDH1*, *CHEK2*, *MRE11A*, *MUTYH*, *NBN*, *NF1*, *PTEN*, *RAD50*, *RAD51C*, *RAD51D*, *TP53*, *PALB2*
Colorectal cancer	17	*APC*, *BMPR1A*, *CDH1*, *CHEK2*, *EPCAM*, *GREM1*, *MLH1*, *MSH2*, *MSH6*, *MUTYH*, *PMS2*,[a] *POLD1*, *POLE*, *PTEN*, *SMAD4*, *STK11*, *TP53*
Paragangliomas/ pheochromocytomas	12	*FH*, *MAX*, *MEN1*, *NF1*, *RET*, *SDHA*, *SDHAF2*, *SDHB*, *SDHC*, *SDHD*, *TMEM127*, *VHL*
Renal cancer	19	*MLH1*, *MSH2*, *MSH6*, *PMS2*,[a] *PTEN*, *TP53*, *VHL*, *EPCAM*, *FLCN*, *TSC2*, *TSC1*, *SDHB*, *MET*, *MITF*, *SDHC*, *SDHD*, *SDHA*, *FH*, *BAP1*
Pancreatic cancer	13	*APC*, *ATM*, *BRCA1*, *BRCA2*, *CDKN2A*, *EPCAM*, *MLH1*, *MSH2*, *MSH6*, *PMS2*,[a] *STK11*, *TP53*, *PALB2*
Ovarian cancer/uterine cancer	24	*ATM*, *BARD1*, *BRCA1*, *BRCA2*, *BRIP1*, *CDH1*, *CHEK2*, *EPCAM*, *MLH1*, *MRE11A*, *MSH2*, *MSH6*, *MUTYH*, *NBN*, *NF1*, *PMS2*,[a] *PTEN*, *RAD50*, *RAD51C*, *RAD51D*, *STK11*, *TP53*, *PALB2*, *SMARCA4*
Melanoma	7	*BAP1*, *BRCA2*, *CDK4*, *CDKN2A*, *MITF*, *PTEN*, *TP53*

[a] *PMS2* is excluded from SV analysis due to pseudogene interference (see "Patient samples" in the "Materials and Methods").

samples due to moderate sensitivities and capabilities for small CNV detection. For example, in a study analyzing 1,017 case–control samples, XHMM achieved 85% sensitivity for CNVs spanning ≥5 exons, and only has 67% sensitivity for CNVs covering <5 exons.[13]

Commercial and academic laboratories that offer germline NGS panel testing have developed and applied SV calling methods. However, reliability, sensitivity and specificity of NGS-based methods have not been adequately investigated due to limited sample size.[6,18] In this study, we evaluated SV detection and interpretation using NGS data from 60,000 clinical samples, and compared the results with aCGH/ MLPA/PCR data. This is by far the largest clinical study of SV detection by NGS. We also investigated false positive (FP) and false negative (FN) SVs to assess the relative advantages and disadvantages of NGS-based approach compared with traditional methods.

MATERIALS AND METHODS

Patient samples

Samples from a consecutive series of 60,000 patients referred to Ambry Genetics (Aliso Viejo, CA, USA) for NGS-based multigene hereditary cancer testing were assessed. Genes in the cancer panels are summarized in Table 1. The covered genomic loci include 795 untranslated and coding regions from 48 genes (Table S1). *PMS2* was excluded from SV analysis due to >99% high sequence similarity in its exons 12–15 with its pseudogene *PMS2CL*, which interferes with NGS capture and sequencing reads alignment. Alternative methods were used to assess single-nucleotide variants and SVs to address this known pseudogene issue. This study was determined to be exempt by the institutional review board.

NGS library preparation and sequencing

Genomic DNA was extracted from whole blood or saliva using the QiaSymphony instrument (Qiagen, Hilden, Germany) according to the manufacturer's instruction. Isolated DNA was quantified using a NanoDrop UV spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and/or Qubit Fluorometer (Thermo Fisher Scientific, Carlsbad, CA, USA). DNA with A260/280 ratio between 1.8 and 2.0, and A260/230 ratio ≥1.6 was used for NGS library preparation. Genomic DNA was mechanically sheared into 150–500 bp fragments with average size of 300 bp with the LE220 focused ultrasonicator (Covaris, Woburn, MA, USA). NGS libraries were prepared with the Freedom EVO100 automated system (Tecan, Mannedorf, Switzerland) using the Kapa library preparation kit (Kapa Biosystems, Wilmington, MA, USA) according to manufacturer's instructions. Libraries were purified using AMPure XP beads (Beckman Coulter, Brea, CA, USA) and quantified with the 2200 TapeStation Instrument (Agilent Technologies, Santa Clara, CA, USA). The customized biotinylated DNA oligonucleotides for bait capture were designed using IDT (Integrated DNA Technologies, Coralville, IA, USA) online design tool. Repeat sequences were masked during probe design. Adapter-ligated DNA was hybridized in solution with the biotinylated DNA oligonucleotides. After hybridization, Streptavidin Dynabeads (Thermo Fisher Scientific, Carlsbad, CA, USA) were used to capture biotinylated DNA. Purified DNA was PCR amplified using the Bio-Rad T100 thermal cycler (Bio-Rad Laboratories, Hercules, CA, USA) with the following conditions: one cycle of 98 °C for 45 seconds, followed by a program of 98 °C for 15 seconds, 65 °C for 30 seconds, and 72 °C for 30 second for 12 cycles, ending with one cycle of extension at 72 °C for 1 minute. Libraries were further purified using AMPure XP beads (Beckman Coulter, Brea, CA, USA), quantified on the 2200 TapeStation (Agilent Technologies, Santa Clara, CA), normalized by DNA concentration, and pooled. Sequencing was conducted on the HiSeq2500 or NextSeq500 (Illumina Inc., San Diego, CA, USA) using 150 bp paired-end sequencing according to the manufacturer's workflow. Sequencing data processing and SV detection algorithm can be found in the Supplementary methods.

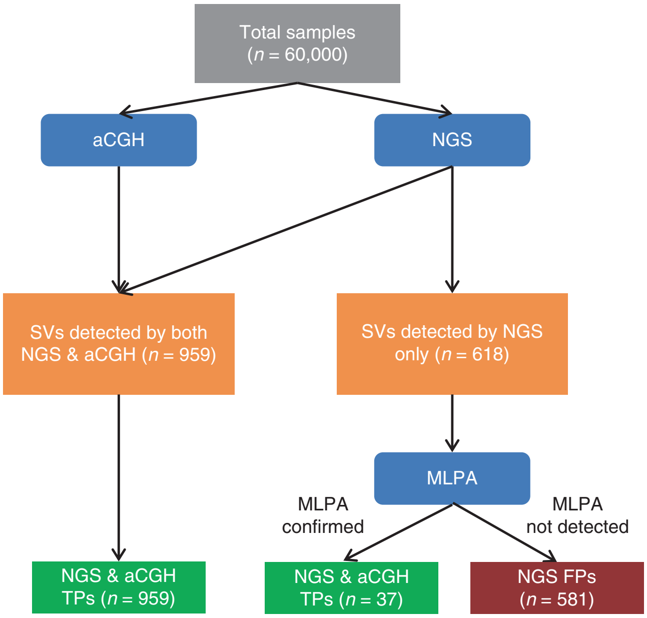
Total samples
(*n* = 60,000)
aCGH
NGS
SVs detected by both NGS & aCGH (*n* = 959)
SVs detected by NGS only (*n* = 618)
MLPA
MLPA confirmed
MLPA not detected
NGS & aCGH TPs (*n* = 959)
NGS & aCGH TPs (*n* = 37)
NGS FPs (*n* = 581)

Fig. 1 Validation of NGS-based SV detection method on deletions and duplications covered by aCGH, MLPA and NGS. The UTR deletions and duplications, Alu insertions, and *MSH2* inversions were excluded from the analysis. *FP,* false positive; *TP,* true positive.

Validation of SV by aCGH, MLPA, and PCR

The NGS results were validated with aCGH and MLPA on the same 60,000 samples, respectively (Fig. 1). aCGH analysis was performed as previously described.[19] aCGH data was extracted using Agilent Feature Extraction software v11.0.1.1 and analyzed for copy-number changes using the Agilent Genomic Workbench 7.0 software package (Agilent Technologies, CA, USA) and/or the BioDiscovery Nexus 8.0 (BioDiscovery, CA, USA). aCGH data were reviewed by two independent clinical laboratory scientists. Data with conflict conclusions or genomic regions that had array probes with fluctuation signal were further evaluated using MLPA kit (MRC Holland, Amsterdam, Netherlands) following the manufacturer's instructions. MLPA was also performed if deletions or duplications were detected by NGS but not detected by aCGH. Multiple-step PCR was used to verify *MSH2* inversion and Alu insertion as described previously.[20,21] Intron deletions and duplications were checked with Integrative Genomics Viewer (IGV) (http://www.broadinstitute.org/igv) by experienced analysts. Deletions and duplications detected by aCGH and/or MLPA, and *MSH2* inversions detected by multistep PCR, were considered as true positive SVs, and were used to analyze NGS sensitivity and specificity.

RESULTS

Detection of SVs in 60,000 clinical samples

A total of 31,884 SVs were concordantly detected by multiple platforms with 30,847 (96.9%) intron deletions/duplications and 1,037 (3.1%) SVs covering UTRs and coding regions. Among the 1,037 SVs, there were 15 (1.4%) UTR deletions, 6 (0.6%) UTR duplications, 145 (14.0%) single-exon deletions, 45 (4.3%) single-exon duplications, 378 (36.5%) multiple-exon deletions, 299 (28.8%) multiple-exon duplications, 13 (1.3%) Alu insertions, 7 (0.7%) *MSH2* inversions, and 129 (12.4%) processed pseudogenes (Fig. 2a and Fig. S1).

To investigate the clinical relevance of these findings, the 1,037 SVs were classified as pathogenic, variant likely pathogenic (VLP), variant of unknown significance (VUS), and benign based on an internally developed clinical variant classification scheme, which follows the American College of Medical Genetics and Genomics recommendations and guidelines.[22] Overall, 960 (1.6%) patients had at least one SV, of whom 522 (54%) were found to carry pathogenic/VLP SVs. Among the SVs involving UTR/coding regions, 576 (55.5%) were classified as pathogenic, 16 (1.5%) as VLP, 316 (30.5%) as VUS, and 129 (12.5%) as benign. When examined by SV category, 523 (97.2%) deletions and 49 (14.0%) duplications were classified as pathogenic/VLP (Fig. 2b). As expected, the percentage of pathogenic/VLP SVs was significantly higher in deletions than in duplications ($p <$ 0.001). The pathogenic and VLP SVs were not evenly distributed among all tested genes, with more SVs in *BRCA1* ($p < 0.001$), *MSH2* ($p < 0.001$), and *CHEK2* genes ($p < 0.001$) than in other genes (Fig. 2c). No pathogenic or VLP SV was detected in ten genes (*BAP1*, *CDK4*, *MEN1*, *MET*, *MITF*, *POLD1*, *POLE*, *RET*, *SDHAF2*, and *TSC1*).

Classification is challenging for the 80% of duplications that cover the first or the last coding exon because it is difficult to predict the effect of these alterations on protein function. This led to the high ratio of VUS SVs among duplications (Fig. 2b). Only seven duplications, including five *GREM1* (NM_013372.6) 5'UTR duplications,[23] one *MLH1* (NM_000249.3) CDS6_CDS12 duplication,[24] and one *BRCA1* (NM_007294.3) CDS3_CDS5 duplication,[25] were classified as pathogenic or VLP. There were seven 5'UTR-CDS7 inversions in *MSH2* (NM_000251.1: c.-9509222_1277-3156inv) detected by NGS. Six Alu insertions in the CDS of *BRCA2* (NM_000059.3: c.156_157insAlu or c.8219_8220insAlu), and seven Alu insertions in the CDS49 of *ATM* (NM_000051.3: c.7374_7375insAlu) were also detected by NGS (Table S2). The *BRCA2* c.156_157insAlu has been reported as a Portuguese founder pathogenic variant that promotes breast/ovarian cancer formation.[20] *MSH2* inversions increase the risk for the development of nonpolyposis colorectal cancer.[26] Of the seven *MSH2* 5'UTR-CDS7 inversions, one was initially not confirmed by PCR due to an adjacent indel in the PCR primer binding region (Fig. S2). The inversion was later confirmed by using a redesigned PCR primer. This further indicates that traditional PCR can be confounded by single-nucleotide polymorphisms (SNPs) or short insertions/deletions that are close to primer binding sites, which leads to false negative results.

Polymorphic SVs and processed pseudogenes

The 30,847 (96.9%) intronic deletions/duplications were only detected by NGS and were confirmed by reviewing the data

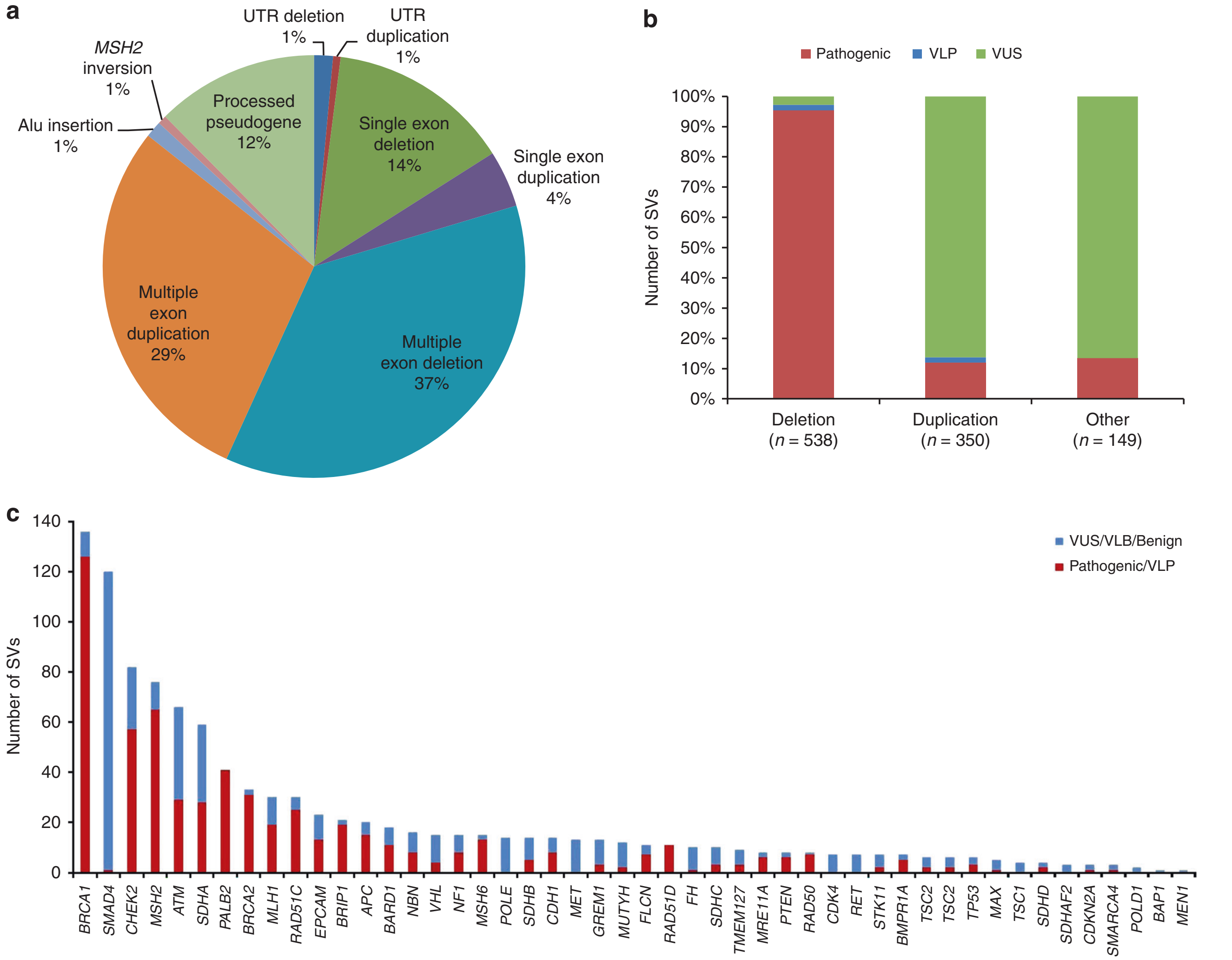

Fig. 2 SV classification. (**a**) Percentage of SVs in each category is shown as a pie chart. Bar chart representation for (**b**) classification and distribution of VUS, VLP and pathogenic deletions, duplications, and other SVs, and (**c**) distribution of SVs in 48 hereditary cancer genes. *VLB,* variant likely benign.

with IGV. The deletions/duplications include 28,216 in *MITF*, 2,560 in *PTEN*, 61 in *CDH1*, 4 in *APC*, 1 in *ATM*, 1 in *MET*, 1 in *MSH6*, 2 in *NF1*, and 1 in *SMARCA4*. The 28,216 deletions in *MITF* were all c.859-521_859-109del413, which is a polymorphic SV found in many samples. Although many intronic SVs may not have direct clinical impact, an 899-bp *PTEN* intronic deletion (NM_000314.4 c.80-956_80-58del899) observed in 4.3% of our samples was also reported by Sandell et al. in 4% of patients with suspected cancer syndromes and 3% of unaffected individuals. While this variant is unlikely to have a direct effect on disease phenotypes, it could be problematic for *PTEN* exon 2 variant analysis if undetected.[27]

From the NGS results, processed pseudogenes appeared as multiple deletions and duplications on their corresponding protein coding genes. This led to manual review of the data with IGV and identification of processed pseudogenes. There were 129 processed pseudogenes identified by NGS including 119 for *SMAD4*, 4 for *MAX*, 5 for *NBN*, and 1 for *SDHB* (Fig. S3–S6 and Table S3). Duplications of ten exons (5'UTR, CDS1, CDS2, CDS4, CDS5, CDS6, CDS8, CDS9, CDS10, and CDS11) and deletions of nine introns (5'UTR-CDS1 intron, CDS1-CDS2 intron, CDS2-CDS4 intron, CDS4-CDS5 intron, CDS5-CDS6 intron, CDS6-CDS8 intron, CDS8-CDS9 intron, CDS9-CDS10 intron, and CDS10-CDS11 intron) were observed in a *SMAD4* transcript (NM_005359.5), which is likely caused by a processed pseudogene from reverse transcription of messenger RNA (mRNA).[28] Sequence analysis of deletions and duplications indicates the pseudogene does not contain CDS3 and CDS7. The frequency of observed *SMAD4* pseudogene (0.2%) is similar to our previous study (0.26%) and is consistent with aCGH/MLPA results.[29] Likewise, the *MAX* (NM_002382.3) pseudogene skips CDS2, the *NBN* (NM_002485.4) pseudogene only contains ten CDSs (CDS2 to CDS11), and the *SDHB* (NM_003000.2) pseudogene contains all eight CDSs. In support of our findings, the *SDHB* and *MAX* pseudogenes were also listed in Pseudogene.org, a comprehensive database that reports ~8,000 high-confidence processed pseudogenes.[30]

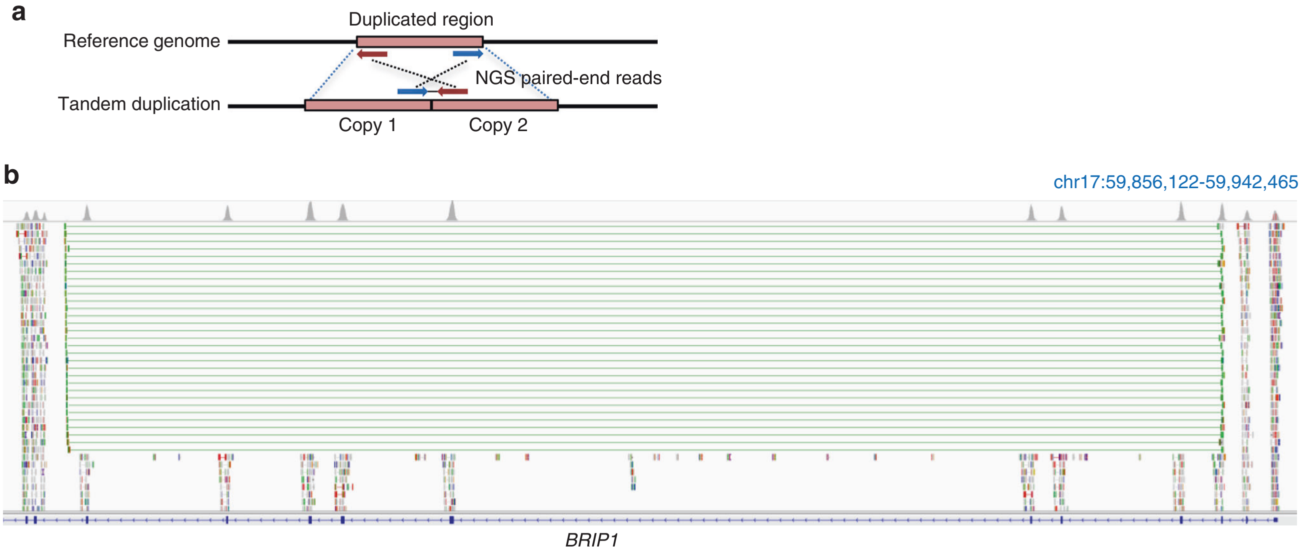

Fig. 3 Tandem duplication detected by NGS paired-end reads. (**a**) Tandem duplication is determined by NGS paired-end reads with long insert size and reversed pair orientation. Blue arrows represent the first read in a pair. Red arrows represent the second read in a pair. (**b**) IGV screenshot of *BRIP1* NM_032043.2 CDS2_CDS10 duplication (c.94-123_1628+1351dup77112). Sequencing read coverage for exons is shown as gray peaks on the top. Peak height represents level of coverage. Green lines represent the linked sequencing reads that belong to the same pair. Green bars connected by green lines represent sequencing reads with long insert size and reversed pair orientation. On the bottom, exons are shown as blue boxes, introns are shown as blue lines, and transcription direction is indicated by blue arrows.

The *NBN* pseudogene was also detected by aCGH and NGS in a previous study with 200 samples.[31] Although pseudogenes interfere with the alignment of sequencing reads to the reference genome and thus contribute to false positive variant calls,[29] the algorithm developed in this study realigns sequencing reads using Pindel 0.2.5[32] to avoid erroneous variant calls and enables detection of processed pseudogenes.

Precise mapping of breakpoints for SVs

We were able to identify precise boundaries for 30 duplications and 55 deletions because their breakpoints were covered by sequencing reads (Table S4). The deletions range from 225 bp to 14,550 bp, and duplications range from 395 bp to 167,490 bp. Obtaining precise locations of deletions/duplications facilitates SV classification and improves clinical interpretation. For example, we were able to reclassify a CDS7 deletion of *MUTYH* (NM_001128425.1) as pathogenic because the deletion covers a portion of the endonuclease domain and the splice donor site. It is difficult to evaluate the pathogenicity of duplications unless they have been studied previously or are tandemly duplicated. To solve this issue, additional NGS assays with extended intron coverage were performed as described.[33] There were four duplications reclassified from VUS to pathogenic after they were confirmed to be in tandem by NGS. A CDS2_CDS10 duplication of *BRIP1* (NM_032043.2) was initially classified as VUS. It was predicted to be an out-of-frame duplication after it was confirmed to be tandem by NGS (Fig. 3), leading to a frameshift of important functional domains. Similarly, two CDS4_CDS11 duplications of *MRE11A* (NM_005591.3) and a CDS24 duplication of *RAD50* (NM_005732.3) were reclassified as pathogenic after they were identified as tandem by NGS. Interestingly, some deletions/duplications from different samples tend to have exactly the same breakpoints, suggesting a common mutation mechanism. For example, a 6-kb CDS3_CDS6 deletion in *BRCA2* (NM_000059.3) was detected in two unrelated samples. A 2.4-kb CDS8_CDS9 duplication in *EPCAM* (NM_002354.2) was identified in four unrelated samples.

NGS detects SVs with high sensitivity and specificity

To evaluate the sensitivity and specificity of the NGS method for SV detection, we split large SV calls into regions and counted false positives by regions (Table S1 and "Materials and Methods"). For example, a CDS16_CDS20 deletion in *MLH1* (NM_000249.3) detected by NGS and confirmed as CDS16_CDS19 deletion by aCGH was counted as four true positives and one false positive, since deletions of CDS16_CDS19 were detected by both methods while the CDS20 deletion was not detected by aCGH. All 7,165 regions in 996 deletions/duplications in coding regions previously detected by aCGH and MLPA were also detected by NGS (Fig. 1). Therefore, NGS achieved 100% sensitivity (excluding UTR deletions/duplications, Alu insertions, and *MSH2* inversions). On the other hand, 581 SVs spanning 1,468 target regions detected by NGS but not confirmed by a secondary assay were identified as false positives. Among 40 of the true positive SVs, we identified 51 regions initially called by NGS that were not detected by MLPA. There were a total of 1,519 false positive regions (1,468 + 51). NGS bait capture probes in this study target UTR regions in 19 of 48 genes on the cancer panel. There were a total of 241 UTR regions detected as deletions or duplications by aCGH and/or

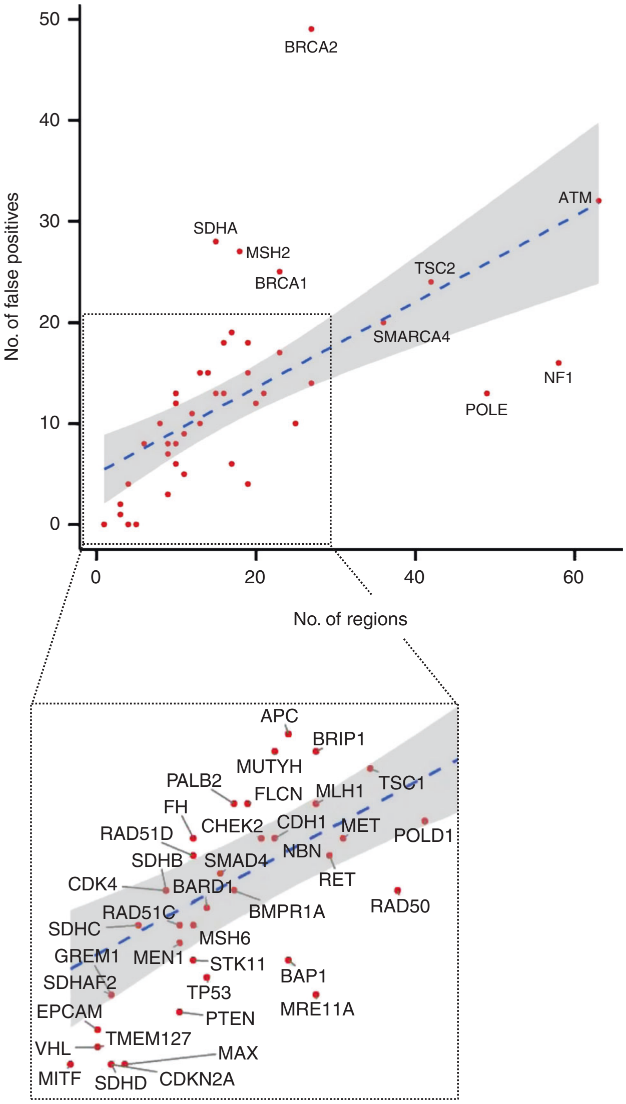

Fig. 4 Linear relationship between the number of false positive SVs and the number of regions in SVs in 48 hereditary cancer genes. Genes (red dots), the linear regression line (blue dotted line), and the standard error (gray shade) are shown in the upper panel. Region in the square is enlarged to show the genes (lower panel).

MLPA in the 60,000 samples. These deletions and duplications in the UTR regions were also detected by NGS. Therefore, the sensitivity of NGS deletions/duplications on UTR regions was 100%.

All 581 false positive SVs were identified only by comparing NGS coverage with the reference coverage data set without supporting junction reads. Further analysis showed 408 (70.2%) false positive SVs were single-exon deletions/duplications. There was no correlation between the occurrence of false positive regions and their corresponding GC content or mappability of sequencing reads, indicating coverage bias due to GC content and mappability was addressed well by the algorithm. The number of false positives was moderately correlated (Pearson correlation coefficient = 0.61, $p < 0.001$) with the number of target regions in the genes (Fig. 4), suggesting the false positives were due to coverage fluctuations caused by experimental variation from sample collection, library preparation, or capture hybridization. There were in total 47,700,000 target regions (795 target regions per sample × 60,000 samples), and 47,692,410 regions (47,700,000 regions – 7,590 positive regions) were SV negative. There were 47,690,891 regions (47,692,410 negative regions – 1,519 false positive regions) correctly called as true negative (TN) SVs by the NGS method. Therefore, the false positive rate [FP/(FP + TN)] is 0.0032% [1,519/(1,519 + 47,690,891)], and the specificity [TN/(TN + FP)] is higher than 99.9% [47,690,891/(47,690,891 + 1,519)].

Among the 996 deletions/duplications that were covered by both NGS and aCGH, 959 (96.3%) deletions/duplications were detected by aCGH. There were 37 (7.6%) deletions/duplications missed by aCGH (Fig. 1 and Table S5). It is known that aCGH is potentially confounded by the presence of genomic duplications in target regions, such as pseudogenes, which interfere with hybridization of DNA to the short length (60-mer) probes. Meanwhile, aCGH also failed to identify a small number of single-exon deletions/duplications.

DISCUSSION

SVs cause large-scale changes in the human genome and can lead to severe genetic disorders and cancer predisposition.[3–5] They are routinely reported by many clinical diagnostic laboratories that perform germline genetic testing. The ACMG published guidelines to provide clinical laboratory geneticists instruction on properly reporting postnatal constitutional CNVs.[22] Rapid and accurate detection of pathogenic variations including SVs is critical for clinicians, who rely on the genetic testing results to make correct clinical decisions. Both false positive and false negative results have adverse health-care impacts on patients and their families.

NGS as a powerful method for SV detection and interpretation

Here, we describe a large-scale study encompassing 60,000 patient samples and over 30,000 SVs analyzed by NGS, aCGH, MLPA, and PCR. Our data not only show the distribution of a variety of SVs in hereditary cancer predisposition genes, but also provide insight into classification of SVs for genetic diagnostic testing. Compared with small indels and single-nucleotide variants (SNVs), interpretation of SVs is more challenging. Breakpoints of SVs detected by NGS can further reduce the number of VUS SVs, especially for duplications. In our study, one deletion and four duplications were reclassified to be pathogenic with breakpoint information. After examination of breakpoints for 53 unique deletions/duplications, we found 18 (34.0%) deletions/duplications had their 3' or 5' ends in the middle of CDS, and 30 (56.6%) deletions/duplications had at least one end conjugated to an Alu element. Enrichment of Alu elements flanking the breakpoints implies Alu-mediated recombination is an important cause for deletions/duplications. For example, among the seven individual deletions/duplications in *BRCA1*/*BRCA2*, breakpoints from two deletions/duplications were within exons, and breakpoints from the other five deletions/duplications were in Alu elements in intronic regions. This is consistent with the findings of Mazoyer et al. that Alu repeats are involved in 79% of *BRCA1*/*BRCA2* rearrangement.[34]

NGS coverage outside of coding regions is essential for detection of known pathogenic SVs. For example, 3'UTR deletions of *EPCAM* were reported to confer Lynch syndrome and are therefore included in routine Lynch syndrome diagnostics.[35] A ~40-kb duplication located upstream of *GREM1* increases ectopic *GREM1* expression and predisposes to hereditary mixed polyposis syndrome (HMPS).[36] Given this information, we extended our NGS coverage into the 5'UTR of *GREM1*, 3'UTR of *EPCAM*, as well as 5'UTR and intron of *MSH2*. As a result, with the combination of expanded NGS coverage and analysis, we detected 3 duplications in *GREM1* 5'UTR, 13 deletions in *EPCAM* 3'UTR, and 7 inversions of *MSH2* 5'UTR-CDS7. Concurrent 3'UTR and partial gene deletions of *EPCAM* and *MSH2*, respectively, were found in two distinct samples, which confounded interpretation of SV effects. *EPCAM* and *MSH2* deletions may be associated with increased risk of colorectal, duodenal, and pancreatic cancers, and decreased risk of endometrial cancer. Individuals with a deletion expanding closer to the *MSH2* gene may have higher risk of developing endometrial cancer. The NGS-based approach reported in this study allows for a more accurate characterization of the extent of SVs, which in turn can be used to better understand genotype–phenotype associations.[37] Moreover, the SVs reported here can serve as a valuable reference for other researchers and clinicians.

aCGH alone may be insufficient for SV detection

Many studies have been performed to improve the quality of detection for SNVs and indels, but the accuracy of SV detection has not been fully evaluated due to the limited number of known SVs. NGS is a high-throughput method that analyzes many genes and all types of variants simultaneously at a reduced cost. With its high resolution, NGS can detect precise breakpoints of SVs in some cases. In this study, we showed that aCGH is potentially confounded by the presence of genomic duplications in target regions, such as pseudogenes, and may also miss single-exon deletion/duplications. Therefore, additional confirmation methods are recommended when only aCGH is used to detect SVs in highly homologous regions. Our NGS assay is less affected by pseudogenes because we use 150 bp paired-end reads to sequence 150–500 bp genomic regions. However, the accuracy of the NGS method is reduced when shorter reads are used or genes with high sequence similarity to their pseudogene region are studied.

Limitations of NGS method for SV detection

We acknowledge the limitations of NGS. SV detection using NGS tends to be affected by coverage variation, which is introduced at experimental steps such as fragmentation, hybridization, PCR amplification, and sequencing. We found that 581 SVs identified by NGS were false positives, and 70% of them were single-exon deletions/duplications. False positive SVs spanning multiple exons were markedly reduced, because coverage variation was generated randomly and thus was unlikely to change the coverage of multiple consecutive exons in the same direction. In addition, large genes tend to have more false positive SVs than small genes in our data cohort, because there are more regions in larger genes possibly affected by coverage fluctuation. To ensure correct and accurate results for clinical reporting, secondary confirmation methods are required to follow up all SVs detected by NGS. Traditional methods such as aCGH, MLPA, and PCR are recommended for such confirmation.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (https://doi.org/10.1038/s41436-018-0397-6) contains supplementary material, which is available to authorized users.

DISCLOSURE

All authors are employed by and receive salaries from Ambry Genetics when they are working on this project. Hereditary cancer panel testing is among Ambry Genetics' commercially available tests.

REFERENCES

1. Freeman JL, Perry GH, Feuk L, et al. Copy number variation: new insights in genome diversity. Genome Res. 2006;16:949–961.
2. Guan P, Sung W-K. Structural variation detection using next-generation sequencing data: a comparative technical review. Methods. 2016;102:36–49.
3. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463:899–905.
4. Shlien A, Malkin D. Copy number variations and cancer susceptibility. Curr Opin Oncol. 2010;22:55–63.
5. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437–455.
6. Feng Y, Chen D, Wang GL, Zhang VW, Wong LJ. Improved molecular diagnosis by the detection of exonic deletions with target gene capture and deep sequencing. Genet Med. 2015;17:99–107.
7. Judkins T, Leclair B, Bowles K, et al. Development and analytical validation of a 25-gene next generation sequencing panel that includes the BRCA1 and BRCA2 genes to assess hereditary cancer risk. BMC Cancer. 2015;15:215.
8. LaDuca H, Stuenkel AJ, Dolinsky JS, et al. Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients. Genet Med. 2014;16:830–837.
9. Pritchard CC, Salipante SJ, Koehler K, et al. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J Mol Diagn. 2014;16:56–67.
10. Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet. 2004;66:488–495.
11. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. 2002;30:e57.
12. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14 suppl 11: S1.

13. Fromer M, Moran JL, Chambert K, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012;91:597–607.
14. Krumm N, Sudmant PH, Ko A, et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012;22: 1525–1532.
15. Bellos E, Kumar V, Lin C, et al. cnvCapSeq: detecting copy number variation in long-range targeted resequencing data. Nucleic Acids Res. 2014;42:e158.
16. Pugh TJ, Amr SS, Bowser MJ, et al. VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. Genet Med. 2016;18:712–719.
17. Fowler A, Mahamdallie S, Ruark E, et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res. 2016;1:20.
18. Retterer K, Scuffins J, Schmidt D, et al. Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. Genet Med. 2015;17:623–629.
19. Chong HK, Wang T, Lu HM, et al. The validation and clinical implementation of BRCAplus: a comprehensive high-risk breast cancer diagnostic assay. PLoS One. 2014;9:e97408.
20. Peixoto A, Santos C, Rocha P, et al. The c.156_157insAlu BRCA2 rearrangement accounts for more than one-fourth of deleterious BRCA mutations in northern/central Portugal. Breast Cancer Res Treat. 2009;114:31–38.
21. Rhees J, Arnold M, Boland CR. Inversion of exons 1-7 of the MSH2 gene is a frequent cause of unexplained Lynch syndrome in one local population. Fam Cancer. 2014;13:219–225.
22. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST, Working Group of the American College of Medical Genetics Laboratory Quality Assurance C. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med. 2011;13:680–685.
23. Davis H, Irshad S, Bansal M, et al. Aberrant epithelial GREM1 expression initiates colonic tumorigenesis from cells outside the stem cell niche. Nat Med. 2015;21:62–70.
24. Baudhuin LM, Ferber MJ, Winters JL, et al. Characterization of hMLH1 and hMSH2 gene dosage alterations in Lynch syndrome patients. Gastroenterology. 2005;129:846–854.
25. Kwong A, Chen J, Shin VY, et al. The importance of analysis of long-range rearrangement of BRCA1 and BRCA2 in genetic diagnosis of familial breast cancer. Cancer Genet. 2015;208:448–454.
26. Wagner A, van der Klift H, Franken P, et al. A 10-Mb paracentric inversion of chromosome arm 2p inactivates MSH2 and is responsible for hereditary nonpolyposis colorectal cancer in a North-American kindred. Genes Chromosomes Cancer. 2002;35:49–57.
27. Sandell S, Schuit RJ, Bunyan DJ. An intronic polymorphic deletion in the PTEN gene: implications for molecular diagnostic testing. Br J Cancer. 2013;108:438–441.
28. Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. 2005;33:2374–2383.
29. Millson A, Lewis T, Pesaran T, et al. Processed pseudogene confounding deletion/duplication assays for SMAD4. J Mol Diagn. 2015;17:576–582.
30. Zhang Z, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. 2003;13:2541–2558.
31. Manchini-Dinardo D, Judkins T, Elias MC, et al. Dosage analysis by next generation sequencing and microarray CGH indicates putative processed pseudogenes in SMAD4 and NBN. Paper presented at: ACMG Annual Clinical Genetics Meeting; March 24–28, 2015; Salt Lake City, UT.
32. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865–2871.
33. Pesaran T, Karam R, Huether R, et al. Beyond DNA: an integrated and functional approach for classifying germline variants in breast cancer genes. Int J Breast Cancer. 2016;2016:2469523.
34. Mazoyer S. Genomic rearrangements in the BRCA1 and BRCA2 genes. Hum Mutat. 2005;25:415–422.
35. Tutlewska K, Lubinski J, Kurzawski G. Germline deletions in the EPCAM gene as a cause of Lynch syndrome—literature review. Hered Cancer Clin Pract. 2013;11:9.
36. Jaeger E, Leedham S, Lewis A, et al. Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1. Nat Genet. 2012;44:699–703.
37. Kempers MJ, Kuiper RP, Ockeloen CW, et al. Risk of colorectal and endometrial cancers in EPCAM deletion-positive Lynch syndrome: a cohort study. Lancet Oncol. 2011;12:49–55.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, and provide a link to the Creative Commons license. You do not have permission under this license to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2018