Characterizing sensitivity and coverage of clinical WGS as a diagnostic test for genetic disorders

Yan Sun

BGI

Fengxia Liu

BGI

Chunna Fan

BGI

Yaoshen Wang

BGI

Lijie Song

BGI

Zhonghai Fang

BGI

Rui Han

BGI

Zhonghua Wang

BGI

Xiaodan Wang

BGI

Ziying Yang

BGI

Zhenpeng Xu

BGI

Jiguang Peng

BGI

Chaonan Shi

BGI

Hongyun Zhang

BGI

Wei Dong

BGI

Hui Huang

BGI

Yun Li

BGI

Yanqun Le

BGI

Jun Sun (✉ sunjun@bgi.com)

BGI

Zhiyu Peng

<https://orcid.org/0000-0003-2349-750X>

Research article

Keywords: WGS, sensitivity and PPV, DP and breadth of coverage, CNV, clinical diagnosis

Posted Date: May 29th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-30015/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at BMC Medical Genomics on April 13th, 2021. See the published version at <https://doi.org/10.1186/s12920-021-00948-5>.

Abstract

Background

With the reduce of cost and incomparable advantages, WGS is likely to change the way of clinical diagnosis of rare and undiagnosed diseases. However, the sensitivity and breadth of coverage of clinical WGS as a diagnostic test for genetic disorders has not been fully evaluated, especially for CNV detection. All the samples underwent WGS using MGISEQ-2000. The performance of NA12878, YH cell line, and the Chinese trios were measured for sensitivity, PPV, depth and breadth of coverage. We also compared the performance of WES and WGS using NA12878. The sensitivity and PPV were tested using family-based trio design in the Chinese trios. We also developed a systematic WGS pipeline for the analysis of 8 clinical cases with known disease-causing variants.

Results

In general, the sensitivity and PPV for SNV/indel detection increased with mean depth, and reached a plateau at a ~ 40X mean depth using down-sampling samples of NA12878. With a mean depth of 40X, the sensitivity of homozygous and heterozygous SNPs of NA12878 was > 99.25% and > 99.50% respectively, and PPVs were 99.97% and 98.96%. Homozygous and heterozygous indels showed lower sensitivity and PPV. The sensitivity and PPV is still not 100% even with a mean depth of ~ 150X. We also observed a substantial variation in the sensitivity of CNV detection across different tools, especially in CNVs with a size of less than 1 kb. In general, the breadth of coverage for disease-associated genes and CNVs increased with mean depth. The sensitivity and coverage of WGS (~ 40X) is better than WES (~ 120X). Among the Chinese trios with ~ 40X mean depth, the sensitivity in the offspring was > 99.48% and > 96.36% for SNP and indel detection, and PPVs were 99.86% and 97.93%. All the 12 previously validated variants in the 8 clinical cases were successfully detected by our WGS pipeline.

Conclusions

The current standard of a mean depth of 40X may be sufficient for SNV/indel detection and identification of most CNVs. Clinical scientists should know the range of sensitivity and PPV for different classes of variants for a particular WGS pipeline, which would be useful when interpreting and delivering clinical reports.

Background

So far, there are more than 8,000 Mendelian diseases recorded by OMIM (Online Mendelian Inheritance in Man), more than 5,000 of which have phenotype description and molecular basis. The rapid development of massively parallel sequencing (MPS) technology has revolutionized the field of genetic diagnosis in clinical setting, making it possible for MPS to be a routine part of clinical care. Recently, whole-exome

sequencing (WES) and whole genome sequencing (WGS) are routinely used and are gradually being optimized for the detection of rare and common genetic variants in humans [1–4]. Comparing to WES, WGS is more powerful for detecting variants. Theoretically, WGS has the potential to identify nearly all forms of genetic variation [5], including single-nucleotide variants (SNVs) in both the protein-coding and noncoding regions (such as introns and promoters) of the genome, small insertions/deletions (indels), and copy-number variants (CNVs) [6, 7]. Without target region selection, WGS could provide a more uniform DP for the genome, making the detection of CNVs readier. What is more, WGS could provide higher sensitivity and higher yield in variant detection in the coding regions [8–10]. Several studies have demonstrated the advantages of WGS for variant detection [11–14]. For patients who highly suspected with genetic disorder, WGS maybe one and the optimal for further evaluation when the patient remains undiagnosed after clinical WES. With the reduce of the cost in sequencing and its incomparable advantages, WGS is likely to change the way of clinical diagnosis of rare and undiagnosed diseases and is bound to become a routine part of clinical care in the near future.

So far, WGS still follows the three steps since the invention of MPS technology: preparation of template, library construction and sequencing, and data analysis. Factors that influence the three steps (such as quality of genomic DNA [15, 16], library preparation [17–19], sequencing platform [18, 20], bioinformatics analytical pipeline [21]) may influence the results of clinical WGS. Building a standardized wet laboratory and bioinformatics pipeline can improve comparability of WGS. The analysis of a clinical WGS usually starts from quality evaluation. Mean DP is often recognized as a general indicator of overall sensitivity. It has been reported that variant calling is more reliable with increasing DP [22]. As a crucial factor in data quality, most centers conducting MPS technology would determine thresholds for average DP across WES/WGS, and determine the minimum DP that must be achieved for a certain fraction of target bases [23]. The breadth of coverage describes the fraction of the total target genomic region that is sequenced to an adequate depth in a particular assay [4]. The American College of Medical Genetics and Genomics (ACMG) recommends that 90%-95% breadth of coverage above a minimum threshold of 10X should be achieved for exome data with an average depth of 100X [23]. So far, a mean depth of 30-50X is the most widely used mean DP for WGS [24, 25]. However, the sensitivity and breadth of coverage of clinical WGS has not been fully evaluated, especially for CNV detection, some of which are associated with human disease [26–29]. For clinical WGS, the sensitivity and coverage of CNVs has not been comprehensively investigated.

In this study, we performed a systematic analysis of the sensitivity and coverage of clinical WGS using 5 gold standard samples (NA12878, YH, NA24631, NA24694 and NA24695). Then we applied clinical WGS for the reanalysis of 8 clinical cases with known disease-causing variants. The results may provide a reference for the laboratories who perform clinical WGS.

Results

Study design

In order to test the sensitivity and coverage of clinical WGS, Genome in a Bottle (GIAB) sample HG001 (NA12878), HG005 (NA24631)/HG006 (NA24694)/HG007 (NA24695) (known as the Chinese son/father/mother), and YH cell line (a human lymphoblastoid cell line from first Asian genome donor) [30] were collected and sequenced using MGISEQ-2000 platform. All the samples used in this study are listed in Table 1.

Table 1
Sample information

Sample Name	Sequencing	Platform	DNA input	PCR/PCR-free	Read length	Mean depth
NA12878-1	WGS	MGISEQ-2000	1 µg	PCR	PE100	197
NA12878-2	WES	MGISEQ-2000	300 ng	PCR	PE100	220
YH	WGS	MGISEQ-2000	1 µg	PCR	PE100	151
NA24631	WGS	MGISEQ-2000	1 µg	PCR	PE100	111
NA24694	WGS	MGISEQ-2000	1 µg	PCR	PE100	115
NA24695	WGS	MGISEQ-2000	1 µg	PCR	PE100	117

Figure 1 shows the overall design of this study. First, in order to assess the recommended depth for proband-only WGS in clinical diagnostics, the analysis of the sensitivity and positive predictive value (PPV) of high-confidence SNPs/indels, the sensitivity of CNV detection, and the depth and breadth of coverage for disease-associated genes and CNVs were performed using down-sampling samples of NA12878-1. The results of sensitivity and PPV from a single genome may be difficult to generalize to a range of samples [31]. So in this part, similar analyses were also performed for down-sampling samples of another high depth sequencing sample of YH. After determining the recommended mean DP for proband-only WGS, the Chinese trios (NA24631, NA24694 and NA24695) with recommended mean DP were used to test the sensitivity and PPV of trio-based WGS when taking advantage of the family-based trio design in clinical WGS. Using down-sampling samples of NA12878-1 and NA12878-2, we also compared the performance of WES and WGS using the recommended mean DP. Finally, we analyzed 8 clinical cases with known disease causing variants using our WGS pipeline (Fig. 1).

Sensitivity And Positive Predictive Value Of Variant Calls

To evaluate the performance of variant calls in detecting true genotypes, the high-confidence calls (SNPs and indels) for GIAB sample HG001 (NA12878) and HG005/HG006/HG007 (Chinese son/father/mother)

were considered as true-positive calls (v3.3.2) [35]. We restricted the calculation of sensitivity (high-confidence calls detected by our method/all the high-confidence calls in GIAB) and PPV (high-confidence calls detected by our method/all the variants detected by our method) of variant calls to the high confidence region (v3.3.2) [35]. High-confidence variant calls and regions tend to include a subset of variants and regions that are easier to detect [35]. The sensitivity and PPV of variant calls in YH sample were also evaluated. The alleles validated by the Illumina 1M BeadChip were considered as “true-positive” calls for YH [30]. Genotype quality (GQ) and DP were used to filter out variants with erroneous variant calls.

Sensitivity Of Cnv Detection

In order to detect the sensitivity of proband-only WGS for CNV detection. The 12 down-sampling samples of NA12878-1 (10X-150X) with increasing mean DP were evaluated. CNVnator (read depth) [32], BreakDancer (read pair) [33] and LUMPY (read depth and read pair) [34] were used for the detection of CNVs for the 12 down-sampling samples. In this study, a total of 3 CNV call datasets were assessed. The overall sensitivity of CNVnator, BreakDancer and LUMPY of the 12 down-sampling samples for CNV call set 1 (BreakDancer only included the detection of deletions in CNV call set 1), CNV call set 2 and CNV call set 3 was shown in Fig. 3. In general, CNV calling is reliable with increasing DP (Fig. 3a-c). At increasing sequencing depths, the trends of the sensitivity curves for the 3 CNV tools were different from one another. CNVnator showed a wide range in the sensitivity with varying DP, and the sensitivity visibly increased with mean depth, which indicates that the sensitivity of CNV detection is positively correlated with sequencing depth.

We also observed that the size of CNVs may influence the sensitivity of CNV tools. The performance of each tool varied along with the size of CNVs (Supplementary Fig. 1–9). Taking the deletions in CNV call set 1 as an example, the widely-used tool CNVnator may not be suitable for CNV detection when the size of CNV is less than 1 kb. When the CNV size is less than 1 kb, the sensitivity significantly increased with sequencing depth from 10X to 30X but reached a plateau at a ~ 40X depth (Supplementary Fig. 10). This indicated that when the CNV size is less than 1 kb, the detection rate is greatly influenced by sequencing depth, which is less obvious when the CNV size is between 6–70 kb. While, BreakDancer provided a better performance for deletion detection in all size of CNVs. These results suggested that, clinical scientists should pay more attention to the selection of CNV tools when focusing on different CNV sizes.

The selection of CNV call set and CNV detection tools may influence the sensitivity of CNV detection, making the assessment of the recommended depth for CNV detection of proband-only WGS difficult. In order to investigate the minimum requirement of mean DP for CNV detection in proband-only WGS, using 3 CNV call sets (CNV call set 1, 2, 3) and the detection results of 3 CNV tools (CNVnator, BreakDancer and LUMPY), we defined a “miss detection index” (MDI) value in this study. The MDI value for a specific mean DP is defined as the frequency when the specific mean DP shows the “lowest” sensitivity for different

CNV size in a CNV call set. Without regard to selection of CNV call set and CNV detection tools, MID can be used to evaluate the recommended depth for CNV detection of proband-only WGS.

$$MDI_i = \frac{M_i}{N}$$

In the formula, M means the number of times when mean depth i shows the “lowest” sensitivity of CNV detection, N means the total number of times for all the depth that shows the “lowest” sensitivity of CNV detection. To obtain qualified CNV sizes in a CNV call set for evaluation, some criteria need to be fulfilled for a CNV size (Supplementary material). Detailed criteria and examples of the calculation of MDI can be found in Supplementary material.

As a result, the MDI value of a depth of 10X, 20X, 30X and 40X ranked first in the down-sampling samples of NA12878-1 (Fig. 3d-f). 10X-40X account for more than 71.98% of the total depth. Taking together the sensitivity of detecting CNVs and sequencing costs, a sequencing depth of ~ 40X provides the best value for CNV detection, as is indicated by the trends in the sensitivity curves (Fig. 3a-c).

Depth and breadth of coverage for disease-associated genes and CNVs

Although WGS is better than WES for variation detection in patients with genetic disorders, the coverage of coding exons in key disease-associated genes of WGS has not been fully evaluated. In order to investigate the breadth of coverage of proband-only WGS for disease-associated genes, the breadth of coverage of 6 gene sets for the 12 down-sampling samples of NA12878-1 (10X-150X) and YH with increasing mean depth were evaluated. For each exons of the coding genes, we calculated the percent of exonic bases covered at more than 10X depth, which was reported to provide 95% sensitivity for heterozygous SNVs in WES [4]. None of the 12 down-sampling samples of NA12878-1 and YH covered 100% of the coding exons in the 6 gene sets except for the ACMG59 gene set (Supplementary Table 2). The results of down-sampling samples of NA12878-1 seems slightly better than down-sampling samples of YH, probably caused by the total sequencing depth of NA12878-1 (~ 197) and YH (~ 151). Across the 6 gene sets, limited range of variation was found in the down-sampling samples when the mean depth is more than 40X (Supplementary Table 2).

As for the ACMG 59 genes, we also observed a range of variation in the breadth of coverage for the 12 down-sampling samples. As such, a mean depth of more than 70X for YH and 90X for NA12878-1 covered 100% for all the ACMG 59 genes. Mean depth of 30X to 50X were most widely used for WGS [24, 25]. The proportion of genes in ACMG59 gene set covered 100% at $\geq 10X$ was 93.22%, 98.31% and 96.61% for NA12878-1_30X, NA12878-1_40X and NA12878-1_50X (with mean depths of ~ 30X, ~ 40X

and ~ 50X). For mean depth of ~ 30X, ~ 40X and ~ 50X for YH, we observed that 86.44%, 93.22% and 98.31% of genes were covered 100% at $\geq 10X$. The breadths of coverage are significantly better when the average sequencing depth is more than 40X (Fig. 4a). Sites of all genes are covered more than 99.9% when the sequencing depth is ~ 40X. Interestingly, we also observed poorly covered RYR1 and TGFBR1 gene in this study (Fig. 4a) comparing to a previously published paper for measuring sensitivity and coverage of clinical WES [4], indicating that the poor coverage of RYR1 and TGFBR1 might be caused by the features of the gene regions, not the sequencing methods used. Clinical scientists need to pay more attention to these genes when performing clinical WGS. Considering the cost of sequencing, a sequencing depth of ~ 40X provides the best value for the coverage of the ACMG 59 gene set, as is indicated by the trends in breadth of coverage value of the ACMG 59 genes. We observed similar patterns for down-sampling samples of YH (Supplementary Table 3). We also examined the percentage of genes at $\geq 20X$ coverage in the ACMG 59 gene set, which could provide 99% sensitivity for heterozygous SNVs [4]. 81.36% and 59.32% of genes were covered 100% for NA12878-1 and YH when the mean DP is ~ 40X.

CNVs are another major part of human genetic variation, which are often found to be associated with human diseases. For the CNVs in the DECIPHER database, most CNVs can be well covered (more than 95% coverage) at $> 10X$ depth when the sequencing depth is 40X-50X for NA12878-1 (Fig. 4b). Clinical scientists should pay more attention to 22q11.2 distal deletion syndrome, Cat-Eye Syndrome Type I and Miller-Dieker syndrome, which showed a coverage of 88.96% and 4.54% even with sequencing depth of 100X. Similar patterns were also observed for down-sampling samples of YH (Supplementary Table 4).

Sensitivity And Ppv Among Trios

The purpose of the analysis of trios, is to test the sensitivity and PPV when taking advantage of the family-based trio information in clinical WGS. Although the sequencing depth of the Chinese trio is more than 100, here we used down-sampling samples with a mean depth of 44.00X, 43.77X and 43.05X for NA24631, NA24694 and NA24695. The reason why we concentrate on the depth of ~ 40X is that, ~ 40X depth is the most widely used and recommended depth for WGS, which is also consistent with some of our previous results, especially for CNV detection.

In this study, we took advantage of the family-based trio design (Fig. 5a) to calculate the sensitivity and PPV in high confidence regions (NISTv3.3.2/GRCh37) [35]. The analysis was restricted to variants with $DP \geq 10X$ and $GQ \geq 20$ (Fig. 5a). We focused on the loci where one parent was homozygous for the alt allele and the other was homozygous for the reference allele. For these variants, the offspring should be heterozygous. This could provide a new “gold standard set” for NA24631 (the offspring). Comparing to the high-confidence calls of NA24631 provided by NIST (the high-confidence call set), the new gold standard set could be used to test the sensitivity and PPV for trio-based WGS. Figure 5b showed the results of the sensitivity and PPV of the “gold standard set” and the high-confidence call set for NA24631. As a result, the sensitivity of the “gold standard set” for SNP and indel detection was $> 99.48\%$ and $> 96.36\%$ respectively, and PPVs were 99.86% and 97.93%. Trio-based analysis showed great improvement

for the PPV of indel detection (Fig. 5b), PPV for indel detection improved from 89.68% (using the high-confidence call set) to 97.93% (using the “gold standard set”).

Wgs And Wes

Here, we evaluated the performance of WES (MGIEasy Exome FS Library Prep Set) and WGS (MGIEasy PCR-Free DNA Library Prep Set) on DNA samples of NA12878. NA12878-1_40X and NA12878-2_120X were used for evaluation. NA12878-1_40X was the down-sampling sample of NA12878-1 with a mean DP of 40X. NA12878-2_120X was an down-sampling sample of NA12878-2 with a mean DP of 120X, which is typical and current standard for clinical WES [36]. Two quality parameters for variation detection (DP and GQ), sensitivity for SNV/indels detection, breadth of coverage of the list of 8,394 putative disease-associated genes and CNVs were compared in this part.

DP and GQ are two main parameters assessing quality of variant calls, which are often used to filter out variants with erroneous variant calls [30, 37]. First, we investigated the GQ and DP distribution of NA12878-1_40X and NA12878-2_120X in the regions of the human genome covered by WES (59,082,036 bp). We found similar results to those obtained in a previous WES study [8]. The distribution of DP for the variants ranged more widely in NA12878-2_120X than in NA12878-1_40X (Supplementary Fig. 11), with a median at 94X depth but a mode at 63X and 58X for SNPs and indels, indicating low levels of coverage for a substantial proportion of variants. While the distribution of DP was normal-like for NA12878-1_40X, with a median at 42X and coinciding mode at 41X for both SNPs and indels (Supplementary Fig. 11). The vast majority of variants called by NA12878-2_120X had a GQ close to 100, and fluctuated along with the GQ scores. The distribution of GQ for variants in NA12878-1_40X showed a mode in low GQ area, probably caused by insufficient variant calls in the WES regions.

For the detection of SNVs and indels, “true positive” calls were further restricted to the regions of the human genome covered by WES (59,082,036 bp). A total of 44,726 variants were used for evaluation. DP and GQ filtering were not used in this part. In general, the sensitivity and PPV of WGS (NA12878-1_40X) is higher than WES (NA12878-2_120X) for the 44,726 variants (Supplementary Table 5), except for the PPV for homozygous indel detection. For 90.39% of gold SNPs, WES and WGS yielded the same genotype. More than 63.41% of these concordant SNVs were identified as heterozygous, which was similar to those obtained in previous WES studies [4, 15, 25].

Then, we investigated the breadth of coverage of NA12878-1_40X and NA12878-2_120X in the 8,394 putative disease-associated genes and CNVs (DECIPHER database). As a result, WGS showed better coverage of both putative disease-associated genes and CNVs. More than 99.77% sites of the exon regions of the 8,394 putative disease-associated genes were covered with a depth of ≥ 10 for NA12878-1_40X, while NA12878-2_120X covered 99.46% of the exon regions. NA12878-2_120X was poorly covered for CNVs in the DECIPHER database (Supplementary Table 6). More than 69.69% of the CNVs showed a coverage of less than 10%.

Analysis Of 8 Clinical Cases With Known Disease-causing Variants

In this study, samples of 8 clinical cases with known variants of various types were recruited and reanalyzed using our WGS pipeline. All the 12 variants were validated previously by methods other than MPS technology. Focusing on SNVs, indels and CNVs, we applied our method to 8 clinical cases using singleton WGS. Variants were manually assessed for quality and interpreted according to the American College of Medical Genetics and Genomics (ACMG) guidelines for variant classification [27, 28]. All the previously validated variants (Supplementary Table 7) were successfully detected by our WGS pipeline, which further demonstrated the sensitivity of the method.

Based the sequencing DP, the average cost for the WGS approach was ~\$490 per sample for the 8 clinical cases in current study. The overall cost (from DNA extraction to reporting) for a single case with a mean sequencing depth of ~ 40X is approximately \$600, including ~\$280 for chemicals (DNA extraction, library construction and sequencing), ~\$220 for labor, and ~\$100 for depreciation expenses.

Discussion

The emergence of MPS technology makes multigene sequencing, exome level sequencing, and even genome level sequencing possible, which has been more and more widely used in clinical diagnosis for genetic disease. In this study, we performed a comprehensive analysis of sensitivity and coverage of clinical WGS as a diagnostic test for genetic disorders. First, we analyzed the sensitivity and PPV value of high-confidence SNPs/indels, and sensitivity of CNV detection using down-sampled NA12878 and YH. A new MDI value was defined for the evaluation of CNV detection; then we investigated the depth and breadth of coverage for disease-associated genes and CNVs in down-sampling samples of NA12878 and YH. A new gene set and a CNV call set were generated during this process; then we compared the performance of WES and WGS on DNA samples of NA12878; we also tested the sensitivity and PPV of variant calls when taking advantage of the family-based trio design in clinical WGS; finally, we analyzed 8 clinical cases with known disease causing variants using our WGS pipeline. The results suggested that, WGS can be used in the detection of SNVs/indels/CNVs with high sensitivity, the current standard of a mean depth of ~ 40X may be a cost-effective sequencing depth for SNV/indel detection and the identification of most CNVs, WGS is bound to be widely used and become a routine part of clinical care in the near future.

During the analysis of the PPV of indels in the high confidence regions, we detected a slightly decline for heterozygous indels in the 12 down-sampling samples of NA12878 with the increase of DP. One reasonable explanation is that, with the increase of depth, both of the true positives and false positives increases. It turned out that the false positives increased a little bit faster than the true positives, as a result, the PPV declined. Massive scale p population based polymorphism database and further filtering of the variants may be useful to solve this problem.

In addition to SNVs and indels, another major part of human genetic variation is copy number variation. According to the clinical requirements, choosing the suitable methods and tools for accurate and reliable detection of CNVs is important for clinical diagnostics. WGS can detect nearly all known genetic variations. However, our results indicated that, although CNV calling is reliable with increasing DP, the performance of CNV tools varied a lot. Finding the right tool for CNV detection is difficult for clinical scientists. Our results suggested that read pair methods (BreakDancer in particular) were the best performing approaches for the identification of deletions of more than 1 kb. What is more, although some “gold standard” CNV call set is widely used in published papers [38–40], lacking validation of various methods, some CNVs may be false positives with wrong or low resolution boundaries. Factors (such as CNV size and the selected “gold standard” set) may also influence the sensitivity of CNV detection, making it difficult to determine the sufficient DP for CNV detection. In this study, we introduced the concept of MDI to solve this problem. Finally, we found that the current standard of a mean depth of 40X may be sufficient for the identification of most CNVs. However, there are limitations of our analysis: we only sampled a small number of CNV callers (CNVnator, BreakDancer and LUMPY). In real clinical setting, the application of more than one CNV calling algorithm should be considered to improve the sensitivity of CNV detection. Analyzing sensitivity of CNV detection with all available CNV tools would be an interesting research topic. In addition, we found that additional coverage is associated with the overall increases of sensitivity for CNV detection, however, this is less obvious as the CNV size is more than 100 kb, which has also been found in another published paper [41].

There are no perfect “gold standard” CNV dataset for benchmarking. So, in this study we compiled a list of 2,022 “likely true positive” from the 3 most commonly used CNV call sets of NA12878 from published papers [38–40] for benchmarking (Supplementary Table 8), including 1,912 deletions and 110 duplications. This new set is a combination of the 3 CNV call sets after evaluation. For this new CNV call set, we defined true positives as the variants were detected by at least one CNV tool (CNVnator, BreakDancer and LUMPY) with more than 50% reciprocal overlap and confirmed by visualization of copy ratio though an in-house script. This new set is an ideal “gold standard” CNV call set of NA12878 for clinical WGS benchmarking.

Turn-around time and cost of WGS are two key point for clinical application of WGS. The entire workflow of this method lasts approximately 11–12 days from the recruitment of sample to clinical reporting for one sample. BGI produced MegaBOLT (MegaBOLT bioinformatics analysis accelerator) along with the sequence platform MGISEQ-2000, which is an MGI self-developed and MPS-concentrated hardware accelerating system for bioinformatics analysis. MegaBOLT supports the analysis of WGS and WES, and is 20 times faster than the traditional GATK approach, which can be used to shorten the bioinformatics process. Along with the development of automated diagnosis tools [42, 43], which could be used to prioritize patient phenotypes and expedite genetic disease diagnosis, the turn-around time of WGS could be further reduced. The overall cost, including chemicals, labor, and depreciation expenses for the WGS approach was \$600 per sample (~ 40X depth). Sequencing accounted for nearly half of the total cost. In general, variant calling is more reliable with increasing DP. However, there are detection ceiling for some genes and/or regions (such as regions related to Miller-Dieker syndrome), which cannot be solved by

increasing the sequencing depth. The cost and sensitivity of WGS need to be balanced. Our results suggested that the current standard of a mean depth of 40X may be sufficient for the identification of most SNVs and CNVs. The reduce of the cost and turn-around time would further improve the clinical application of WGS.

Conclusions

In summary, the successful application of WGS as a diagnostic test for genetic disorders in real clinical setting requires comprehensive assessment of the depth and breadth of coverage, and sensitivity of WGS. In this study, we observed variation in the detection of SNV/indel/CNV, and substantial variation in coverage of medically implicated genes and CNVs. In real clinical setting, clinical scientists should know the range of sensitivity and PPV for different classes of variants for a particular WGS pipeline, which would be useful when interpreting and delivering clinical reports. We believe that WGS is likely to change the way of clinical diagnosis of rare and undiagnosed diseases in the near future.

Abbreviations

WGS Whole genome sequencing

WES Whole exon sequencing

OMIM Online Mendelian Inheritance in Man

MPS Massively parallel sequencing

SNV Single-nucleotide variant

Indels Small insertions and deletions

CNV Copy-number variant

ACMG The American College of Medical Genetics and Genomics

GIAB Genome in a Bottle

PPV Positive predictive value

MDI Miss detection index

DP Depth of coverage

GQ Genotype quality

GATK Genome Analysis Toolkit

Methods

Sequencing of samples

Based on the study design, DNA samples of NA12878, NA24631, NA24694 and NA24695 were procured from Coriell (Camden, NJ). 1 µg of DNA were used for YH and the GIAB Chinese family to generate paired-end reads of 100 bp. The most widely used NA12878 were sequenced 2 times (WES and WGS) (Table 1). This study and all the protocols were approved by the ethics committee of BGI (NO. BGI-IRB19143).

Alignment And Variant Calling

In this study, a standard bioinformatics pipeline was used for the analysis of all the samples (Supplementary material). In order to eliminate variability due to differences by various bioinformatics tools, we developed a standard bioinformatics pipeline, which included current widely used tools (Supplementary material). In general, after removal of sequencing adapters and low-quality reads, “clean reads” were aligned to the GRCh37 with BWA 0.7.12-r1039 [44]. Genome Analysis Toolkit (GATK)-package-4.0.11.0 [37] MarkDuplicates was used to remove duplicate reads. After realignment around indels and quality scores re-calibration using GATK-package-4.0.11.0 [37], VCF files were then generated using GATK for further analysis. As for the CNV detection, 3 widely used tools (CNVnator [32], BreakDancer [33] and LUMPY [34]) were performed in this study.

After trimming sequencing adapters and consecutive low-quality bases using fastp [45], the clean reads of NA12878-1 were down-sampled by the sequencing depth of 10X (NA12878-1_10X), 20X (NA12878-1_20X), 30X (NA12878-1_30X), 40X (NA12878-1_40X), 50X (NA12878-1_50X), 60X (NA12878-1_60X), 70X (NA12878-1_70X), 80X (NA12878-1_80X), 90X (NA12878-1_90X), 100X (NA12878-1_100X), 120X (NA12878-1_120X) and 150X (NA12878-1_150X) using seqtk (<https://github.com/lh3/seqtk>). The high coverage NA12878-1 sample was sequenced using MGISEQ-2000 platform to an average depth of ~197X (PE100) using PCR-based library construction. The clean reads of YH were also down-sampled by sequencing depth.

Breadth Of Coverage For Disease-associated Genes And Cnvs

In order to assess the recommended depth for proband-only WGS in clinical diagnostics, we collected a total of 6 gene sets for coverage analysis of disease-associated genes: ACMG59 [46], ClinVar [47] (3,824 genes, accessed on 19 February 2019), Genetic Home Reference [48] (1,471 genes, accessed on 2 July 2019), HGMD [49] (8,171 genes, professional March 2018), OMIM [50] (3,835 genes, accessed on 4 April 2018) and Orphanet [51] (2,405 genes, accessed on 2 July 2019). For the annotation of gene regions, we used the information of NCBI annotation release 104. For different transcripts, we first used the transcripts used in the HGMD database. The rest of the gene region was a combination of the regions of

all transcripts. Coverage analysis of the 12 down-sampling samples of NA12878-1 and YH for the 6 gene sets were performed for evaluation.

Genes in a single gene set of the 6 gene sets is incomplete. In order to generalize a new gene list that contains all the putative disease-associated genes, we compiled a list of 8,394 putative disease-associated genes from the 6 gene sets (Supplementary Table 9). This new gene set was generated using the following criteria: 1) a gene was kept when it was recorded with association status of “assessed” in the Orphanet database, and with one of the following association type: a) “Disease-causing germline mutation(s) in”, b) “Disease-causing germline mutation(s) (gain of function) in”, c) “Disease-causing germline mutation(s) (loss of function) in”, d) “Major susceptibility factor in”, e) “Modifying germline mutation in”, f) “Role in the phenotype of”, g) “Candidate gene tested in”; 2) a gene was kept when it was recorded as a disease related gene in the Genetic Home Reference database; 3) a gene was kept when it was recorded as a disease related gene and the molecular basis of the disease is known, unless the inheritance of the disease was recorded as “SMu” (somatic mutation) only in OMIM database; 4) a gene was kept when at least one variant of the gene was recorded as probable/possible pathological mutations in the HGMD database; 5) a gene was kept when at least one variant of the gene was recorded as pathogenic or likely pathogenic in the ClinVar database; 6) a gene was kept unless the gene was not recorded with well-defined gene ID or genome coordinate information in NCBI annotation release 104. This new gene list was used in the comparison of WES and WGS and is ideal for coverage analysis of clinical WGS.

We also performed coverage analysis of the 12 down-sampling samples of NA12878-1 and YH for CNVs described in the DECIPHER database (GRCh37_v9.29). Detailed information can be found in Supplementary Table 4.

Cnv Analysis

Unlike SNP and indel, there are no perfect "gold standard" CNV dataset for benchmarking. In this study, in order to assess the recommended depth for proband-only WGS in clinical diagnostics, we conducted the evaluation of sensitivity of CNV detection using 3 CNV call sets of NA12878 from published papers [38–40]. CNV call set 1 was also used by Haraksingh et al [38] for the benchmarking of CNV detection from 17 commercially available arrays and low-coverage WGS. CNV call set 2 was called by a machine learning based approach (svclassify), and obtained from the GIAB Consortium benchmark SV calls resource [39]. CNV call set 3 included 874 deletions detected by both reference-based (a custom pipeline and PBHoney using both raw and error-corrected reads) and assembly-based analysis via single-molecule technologies [40]. All these 3 CNV call sets have been previously compiled from NA12878 for benchmarking and downloaded for further analysis in current study (Supplementary Table 10).

In the 12 down-sampling samples of NA12878-1, CNVnator (v0.3.2) [32], BreakDancer (v1.4.5) [33] and LUMPY (v0.2.13) [34] were assessed for sensitivity of CNV detection with default or recommended parameters. In a CNV call set, true positives were classified as CNVs with at least a 50% reciprocal overlap

with CNVs in the call set (BEDTools) [52]. For benchmarking, we determined the number of gold standard CNVs detected in the 12 down-sampling samples of NA12878-1 for the 3 CNV call sets.

Sensitivity and PPV of variant calls in the Chinese trios

To test the sensitivity and PPV of variant calls for trio-based WGS, we investigated the sensitivity and PPV when taking advantage of the family-based trio information in the Chinese trios. Using the segregation pattern, we focused on the autosomes and X chromosome of NA24631, NA24694 and NA24695. Taking advantage of the family-based trio design, we analyzed all variants ($DP \geq 10X$) of both parents where one parent was consistently called as homozygous for the reference allele and the other as homozygous for the alt allele. We then used the variant calls (SNPs and indels) in the offspring to test the sensitivity and PPV in these loci.

Declarations

Ethics approval and consent to participate

Written informed consents were obtained from all the participants. This study was approved by the Institutional Review Board of BGI (NO. BGI-IRB19143) and was performed in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and material

The data that support the findings of this study have been deposited in the CNSA (<https://db.cngb.org/cnsa/>) of CNGBdb with accession code CNP0000813. The data of the 8 clinical cases generated and analyzed during the current study is not publicly available as they are patient samples and sharing them could compromise research participant privacy. The datasets of the GIAB samples used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

ZYP is an Editorial Board Member of BMC Medical Genomics. The authors declare that they have no competing interests.

Funding

This work was supported by the Special Foundation for High-level Talents of Guangdong (Grant 2016TX03R171). This work was also supported by Beijing Municipal Science & Technology Commission (NO. Z181100001918013). These projects are non-profit research projects by government, and had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

ZYP, JS, and YS designed the research. YS wrote the first draft of the article. CNF, LJS, XDW, ZYY, ZPX, CNS, HYZ, WD, YL, YQL and JS designed and performed the experiments. FXL, YSW, ZHF, RH, ZHW, JGP and ZYP performed data analysis. HH provided patient specimens and conducted histopathological examinations. ZYP, JS, YS, FXL, CNF, YSW and LJS contributed to drafting and revising the manuscript. All authors reviewed the manuscript.

Acknowledgments

We thank all the blood donors for their invaluable contribution to this study.

References

1. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491-8.
2. Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics.* 2014;15:247.
3. Carson AR, Smith EN, Matsui H, Braekkan SK, Jepsen K, Hansen JB, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics.* 2014;15:125.
4. Kong SW, Lee IH, Liu X, Hirschhorn JN, Mandl KD. Measuring coverage and accuracy of whole-exome sequencing in clinical context. *Genet Med.* 2018;20(12):1617-26.
5. Pang AW, Macdonald JR, Yuen RK, Hayes VM, Scherer SW. Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. *G3 (Bethesda).* 2014;4(1):63 – 5.
6. Fang H, Wu Y, Narzisi G, O'Rawe JA, Barron LT, Rosenbaum J, et al. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med.* 2014;6(10):89.
7. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? *Hum Genet.* 2016;135(3):359 – 62.

8. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A*. 2015;112(17):5473-8.
9. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet*. 2007;80(4):588–604.
10. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum Mutat*. 2015;36(8):815 – 22.
11. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*. 2010;362(13):1181-91.
12. Herdewyn S, Zhao H, Moisse M, Race V, Matthijs G, Reumers J, et al. Whole-genome sequencing reveals a coding non-pathogenic variant tagging a non-coding pathogenic hexanucleotide repeat expansion in C9orf72 as cause of amyotrophic lateral sclerosis. *Hum Mol Genet*. 2012;21(11):2412-9.
13. Bae BI, Tietjen I, Atabay KD, Evrony GD, Johnson MB, Asare E, et al. Evolutionarily dynamic alternative splicing of GPR56 regulates regional cerebral cortical patterning. *Science*. 2014;343(6172):764-8.
14. Weedon MN, Cebola I, Patch AM, Flanagan SE, De Franco E, Caswell R, et al. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet*. 2014;46(1):61 – 4.
15. Zhu Q, Hu Q, Shepherd L, Wang J, Wei L, Morrison CD, et al. The impact of DNA input amount and DNA source on the performance of whole-exome sequencing in cancer epidemiology. *Cancer Epidemiol Biomarkers Prev*. 2015;24(8):1207-13.
16. Londin ER, Keller MA, D'Andrea MR, Delgrosso K, Ertel A, Surrey S, et al. Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics*. 2011;12:464.
17. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. 2011;12(2):R18.
18. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol*. 2011;29(10):908 – 14.
19. Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, Hannula S, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol*. 2011;12(9):R94.
20. Meienberg J, Zerjavic K, Keller I, Okoniewski M, Patrignani A, Ludin K, et al. New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res*. 2015;43(11):e76.
21. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep*. 2015;5:17875.
22. Normand EA, Alaimo JT, Van den Veyver IB. Exome and genome sequencing in reproductive medicine. *Fertil Steril*. 2018;109(2):213 – 20.

23. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, et al. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med*. 2013;15(9):733 – 47.
24. Thiffault I, Farrow E, Zellmer L, Berrios C, Miller N, Gibson M, et al. Clinical genome sequencing in an unbiased pediatric cohort. *Genet Med*. 2019;21(2):303 – 10.
25. Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, et al. Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. *NPJ Genom Med*. 2016;1.
26. de Vries BB, Pfundt R, Leisink M, Koolen DA, Vissers LE, Janssen IM, et al. Diagnostic genome profiling in mental retardation. *Am J Hum Genet*. 2005;77(4):606 – 16.
27. Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, et al. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet*. 2006;38(9):1038-42.
28. Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. *Annu Rev Genet*. 2011;45:203 – 26.
29. Zhang Y, Haraksingh R, Grubert F, Abyzov A, Gerstein M, Weissman S, et al. Child development and structural variation in the human genome. *Child Dev*. 2013;84(1):34–48.
30. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, et al. The diploid genome sequence of an Asian individual. *Nature*. 2008;456(7218):60 – 5.
31. Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, et al. Comprehensive variation discovery in single human genomes. *Nat Genet*. 2014;46(12):1350-5.
32. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011;21(6):974 – 84.
33. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009;6(9):677 – 81.
34. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15(6):R84.
35. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol*. 2014;32(3):246 – 51.
36. Jamal SM, Yu JH, Chong JX, Dent KM, Conta JH, Tabor HK, et al. Practices and policies of clinical exome sequencing providers: analysis and implications. *Am J Med Genet A*. 2013;161A(5):935 – 50.
37. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297 – 303.
38. Haraksingh RR, Abyzov A, Urban AE. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics*.

- 2017;18(1):321.
39. Parikh H, Mohiyuddin M, Lam HY, Iyer H, Chen D, Pratt M, et al. svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*. 2016;17:64.
 40. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods*. 2015;12(8):780-6.
 41. Zhou B, Ho SS, Zhang X, Pattni R, Haraksingh RR, Urban AE. Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J Med Genet*. 2018;55(11):735 – 43.
 42. Robinson PN, Kohler S, Oellrich A, Sanger Mouse Genetics P, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*. 2014;24(2):340-8.
 43. Kohler S, Schulz MH, Krawitz P, Bauer S, Dolken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*. 2009;85(4):457 – 64
 44. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-60.
 45. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i90.
 46. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med*. 2017;19(2):249 – 55.
 47. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2016;44(D1):D862-8.
 48. Mitchell JA, McCray AT. The Genetics Home Reference: a new NLM consumer health resource. *AMIA Annu Symp Proc*. 2003:936.
 49. Cooper DN, Krawczak M. Human Gene Mutation Database. *Hum Genet*. 1996;98(5):629.
 50. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(Database issue):D514-7.
 51. Pavan S, Rommel K, Mateo Marquina ME, Hohn S, Lanneau V, Rath A. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLoS One*. 2017;12(1):e0170365.
 52. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2.

Figures

Figure 1

Study design

a Sensitivity and PPV of SNP detection in down-sampling samples of NA12878-1

b Sensitivity and PPV of indel detection in down-sampling samples of NA12878-1

Figure 2

Sensitivity and PPV of variant calls from 12 down-sampling samples of NA12878-1

Figure 3

CNV detection in the 12 down-sampling samples of NA12878-1 using 3 CNV call sets

Figure 4

Depth and breadth of coverage for disease-associated genes and CNVs

Variant call set	Variant type	Sensitivity	PPV
High-confidence call set	SNP	99.39%	99.33%
	Indel	92.15%	89.68%
Gold standard set	SNP	99.48%	99.86%
	Indel	96.36%	97.93%

Figure 5

Variants analysis in the Chinese trios

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTables.xlsx](#)
- [Supplementarymaterial.pdf](#)