Performance characterization of PCR-free whole genome sequencing for clinical diagnosis

Zhou, Guiju; Zhou, Meizhen; Zeng, Fanwei; Zhang, Ningzhi; Sun, Yan; Qiao, Zhihong; Guo, Xueqin; Zhou, Shihao; Yun, Guojun; Xie, Jiansheng

Total number of authors:
20

Published in:
Medicine (United States)

Link to article, DOI:
10.1097/MD.0000000000028972

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Observational Study

Medicine®

OPEN

Performance characterization of PCR-free whole genome sequencing for clinical diagnosis

Guiju Zhou, MD[a], Meizhen Zhou, MS[b], Fanwei Zeng, MS[b,c], Ningzhi Zhang, BS[d], Yan Sun, PhD[b], Zhihong Qiao, MS[e,f], Xueqin Guo, MS[g], Shihao Zhou, BS[h], Guojun Yun, MS[i], Jiansheng Xie, MD[j], Xiaodan Wang, MS[e,f], Fengxia Liu, MS[e,f], Chunna Fan, MS[e,f], Yaoshen Wang, BS[e,f], Zhonghai Fang, MS[e,f], Zhongming Tian, BS[e], Wentao Dai, PhD[e,f], Jun Sun, PhD[e,f], Zhiyu Peng, PhD[b], Lijie Song, MS[e,f,k,*]

Abstract

To evaluate the performance of polymerase chain reaction (PCR)-free whole genome sequencing (WGS) for clinical diagnosis, and thereby revealing how experimental parameters affect variant detection.

Five NA12878 samples were sequenced using MGISEQ-2000. NA12878 samples underwent WGS with differing deoxyribonucleic acid (DNA) input and library preparation protocol (PCR-based vs PCR-free protocols for library preparation). The depth of coverage and genotype quality of each sample were compared. The performance of each sample was measured for sensitivity, coverage of depth and breadth of coverage of disease-related genes, and copy number variants. We also developed a systematic WGS pipeline (PCR-free) for the analysis of 11 clinical cases.

In general, NA12878-2 (PCR-free WGS) showed better depth of coverage and genotype quality distribution than NA12878-1 (PCR-based WGS). With a mean depth of ~40×, the sensitivity of homozygous and heterozygous single nucleotide polymorphisms (SNPs) of NA12878-2 showed higher sensitivity (>99.77% and >99.82%) than NA12878-1, and positive predictive value exceeded 99.98% and 99.07%. The sensitivity and positive predictive value of homozygous and heterozygous indels for NA12878-2 (PCR-free WGS) showed great improvement than NA128878-1. The breadths of coverage for disease-related genes and copy number variants are slightly better for samples with PCR-free library preparation protocol than the sample with PCR-based library preparation protocol. DNA input also influences the performance of variant detection in samples with PCR-free WGS. All the 19 previously confirmed variants in 11 clinical cases were successfully detected by our WGS pipeline (PCR free).

Different experimental parameters may affect variant detection for clinical WGS. Clinical scientists should know the range of sensitivity of variants for different methods of WGS, which would be useful when interpreting and delivering clinical reports.

Editor: Ivana Kavecan.

GZ, MZ, FZ, and NZ contributed equally to this work.

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-IRB19019) and was performed in accordance with the Declaration of Helsinki.

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 CNP0001068. The datasets of the GIAB samples used and analyzed during the current study are available from the corresponding author on reasonable request. The data of the 11 clinical cases generated and analyzed during this study is not publicly available as they are patient samples and sharing them could compromise research participant privacy.

The authors declare that they have no competing interests.

This work was partly supported by the Key research and development Program of Anhui Province (202104j07020022). This work was also supported by the Special Foundation for High-level Talents of Guangdong (Grant 2016TX03R171). These are non-profit research projects funded by the Chinese government, which played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Posted history: This manuscript was previously posted to bioRxiv: doi: https://www.biorxiv.org/content/10.1101/2020.06.19.160739v1.

The authors have no conflicts of interest to disclose.

Supplemental Digital Content is available for this article.

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. The datasets generated during and/or analyzed during the current study are not publicly available, but are available from the corresponding author on reasonable request.

[a] *Department Obstetrics and Gynecology, The Second Affiliated Hospital, Anhui Medical University, Hefei, China,* [b] *BGI Genomics, BGI-Shenzhen, Shenzhen, China,* [c] *Department of Biology, Faculty of Science, University of Copenhagen, Copenhagen, Denmark,* [d] *Fuyang People's Hospital, 63 Luci Street, Fuyang, Anhui Province, China,* [e] *Tianjin Medical Laboratory, BGI-Tianjin, BGI-Shenzhen, Tianjin, China,* [f] *Binhai Genomics Institute, BGI-Tianjin, BGI-Shenzhen, Tianjin, China,* [g] *BGI-Wuhan Clinical Laboratories, BGI-Shenzhen, Wuhan, China,* [h] *Department of Genetics and Eugenics, Changsha Hospital for Maternal & Child Health Care Affiliated to Hunan Normal University, Changsha, Hunan Province, China,* [i] *Rehabilitation Ward, Shenzhen Children's Hospital, 7019 Yitian Road, Futian District, Shenzhen, Guangdong Province, China,* [j] *Department of Prenatal Diagnosis, The University of Hongkong Shenzhen Hospital, 1 Haiyuan one Road, Shenzhen, Guangdong Province, China,* [k] *Bacterial Interactions and Evolution Group, Bioengineering, Technical University of Denmark, Kongens Lyngby, Denmark.*

[*] *Correspondence: Lijie Song, Tianjin Medical Laboratory, BGI-Tianjin, BGI-Shenzhen, Tianjin 300308, China (e-mail: songlijie@bgi.com).*

How to cite this article: Zhou G, Zhou M, Zeng F, Zhang N, Sun Y, Qiao Z, Guo X, Zhou S, Yun G, Xie J, Wang X, Liu F, Fan C, Wang Y, Fang Z, Tian Z, Dai W, Sun J, Peng Z, Song L. Performance characterization of PCR-free whole genome sequencing for clinical diagnosis. Medicine 2022;101:10(e28972).

Received: 1 December 2021 / Received in final form: 28 January 2022 / Accepted: 10 February 2022

http://dx.doi.org/10.1097/MD.0000000000028972

Abbreviations: ACMG = The American College of Medical Genetics and Genomics, CNVs = copy number variants, DNA = deoxyribonuclei acid, DP = depth of coverage, GQ = genotype quality, Indels = small insertions and deletions, MPS = massively parallel sequencing, PCR = polymerase chain reaction, PPV = positive predictive value, SNPs = single nucleotide polymorphisms, SNVs = single-nucleotide variants, SVs = structure variations, WGS = whole-genome sequencing.

Keywords: clinical diagnosis, deoxyribonucleic acid input, polymerase chain reaction-free, sequencing depth and coverage, whole genome sequencing

1. Introduction

Massively parallel sequencing (MPS) technology is more and more widely used in genomic research and real clinical setting, which has revolutionized clinical genetic diagnosis. Recently, whole genome sequencing (WGS) has been gradually implemented in the diagnosis of rare and undiagnosed clinical cases,[1–3] making it possible to be a routine in clinical care. Focusing on whole genome scale, WGS can not only be used to detect single-nucleotide variants (SNVs) and small insertions/deletions (Indels), but it can also be used to identify structure variations (SVs).[4–6] What is more, WGS can reduce the cost derived by the need of other tests,[7] and provide higher diagnostic yields than targeted panels.[8]

The process of WGS mainly includes 3 steps: template preparation (isolation of nucleic acid), library preparation (end repairing, adapter addition, optional PCR amplification), and sequencing (sequencing preparation, instrument operation). The results of clinical WGS may be influenced by factors related to the 3 steps, such as quality of genomic DNA,[9,10] methods used for library preparation,[11–13] and differing sequencing platforms.[12,14] After sequencing, a bioinformatics pipeline (sequencing data quality control, alignment, variant calling, and interpretation) will be implemented. The comparability of WGS can be improved by implementing a standardized bioinformatics analysis pipeline. The first 3 steps (template preparation, library construction, and sequencing) could largely influence the quality of WGS data. As for the bioinformatics analysis of WGS, there is already a well-accepted pipeline for the analysis of SNV and indel for WGS data with short read, including alignment with Burrows-Wheeler Aligner,[15] and variant calling with Genome Analysis Toolkit.[16] There are also well-established algorithms for SV detection from short read sequencing data. Great tools and algorithms may improve the sensitivity for variant detection, however, without high quality sequencing data, it is hard to generate good results. As long as WGS data is generated, sometimes there's little things we can do to improve the quality of WGS data. So, the performance of different experimental methods for clinical WGS and how these experimental parameters affect variant detection becomes a relevant research topic.

As for template preparation, library construction, and sequencing, various methods have been provided by different sequencing platforms. Broadly, library preparation of WGS can be classified into 2 groups, PCR-based library preparation protocol versus PCR-free library preparation protocol. Each method has both common and specific variables related to the required DNA input, read length, and cost-effectiveness. These variables could influence the overall quality of WGS data, thus impact the sensitivity of variant detection. Specifically, in addition to evaluating the sensitivity of and breadth of coverage of WGS, we investigated the effects of library preparation method (PCR-based vs PCR-free protocols) and DNA input using MGISEQ-2000 platform in this study.

In the present study, we systematically compared 5 WGS data generated from NA12878 samples. We compared the sensitivity of WGS using samples by differing library preparation protocols (PCR-based vs PCR-free protocols) and DNA inputs (1 μg, 500, 300, and 200 ng). We also compared the yield and quality of sequencing data, depth of coverage, genotype quality, sensitivity for variant detection, and breadth of coverage for each sample. The performance of each method was systematically analyzed and compared, thereby revealing how these experimental parameters affect variant detection. Generally, samples using PCR-free library preparation protocol and DNA input of 1 μg showed the highest performance in depth of coverage (DP) and genotype quality (GQ) distribution, variant detection, and breadth of coverage for disease-related genes and copy number variants (CNVs). We used the WGS pipeline (PCR-free) for the analysis of 11 clinical cases.

2. Methods

2.1. Samples and overall study design

This study and all the protocols followed herein were approved by the Institutional Review Board of BGI (NO. BGI-IRB19019). To investigate the performance of PCR-free WGS for clinical diagnosis, Genome in a Bottle NA12878 was collected and sequenced 5 times with differing library preparation protocols and total DNA inputs (Table 1). All the samples were sequencing

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	PE150	90.72
NA12878-2	WGS	MGISEQ-2000	1 μg	PCR free	PE150	83.01
NA12878-3	WGS	MGISEQ-2000	500 ng	PCR-free	PE150	82.23
NA12878-4	WGS	MGISEQ-2000	300 ng	PCR-free	PE150	79.82
NA12878-5	WGS	MGISEQ-2000	200 ng	PCR-free	PE150	88.39

WGS = whole-genome sequencing.

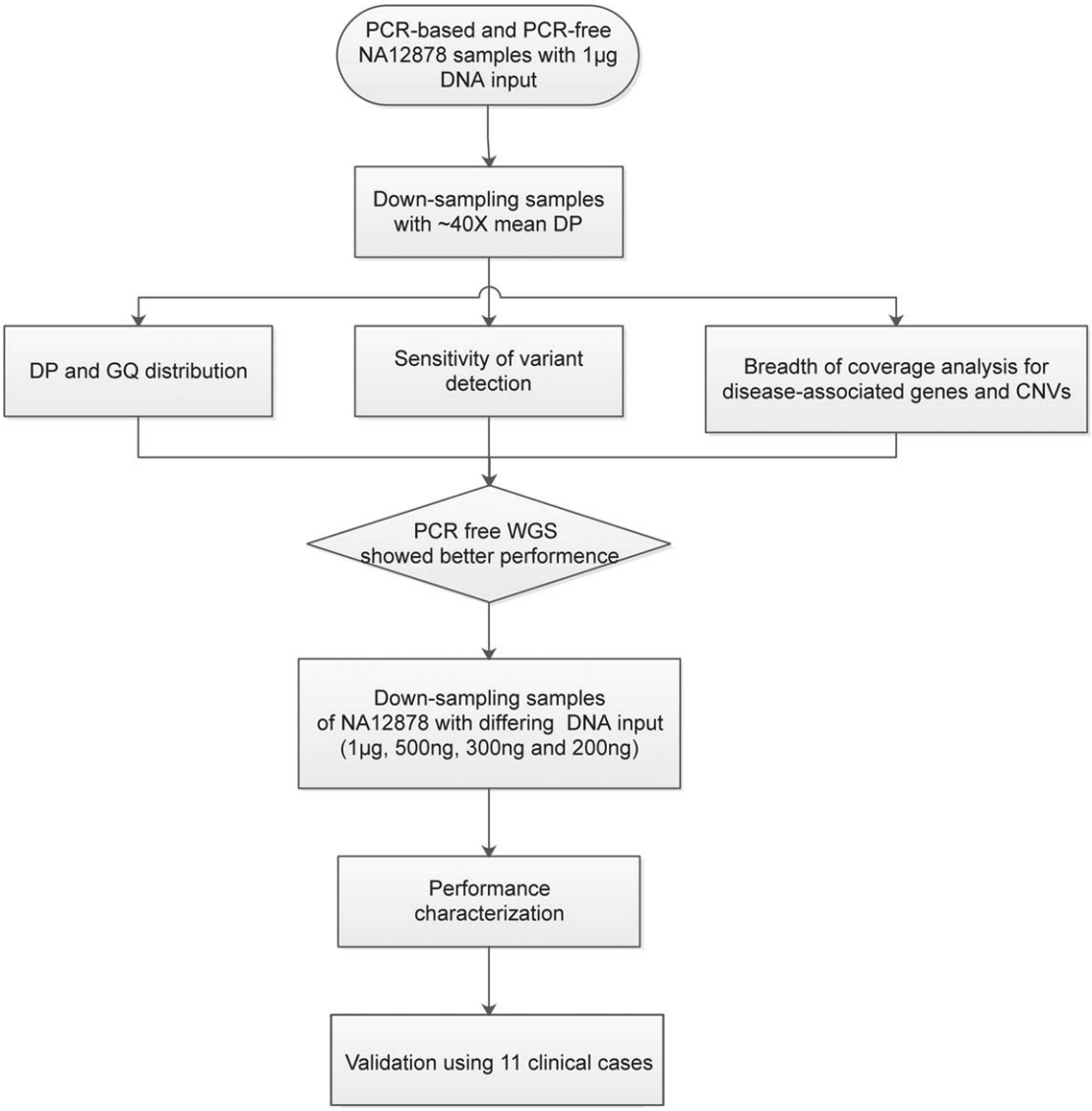

Figure 1. Study design.

on MGISEQ-2000 platform. DNA samples of NA12878 were procured from Coriell (Camden, NJ).

The overall study design is shown in Fig. 1. First, to compare the performance of WGS with PCR-based and PCR-free protocols, analysis of the sensitivity of high-confidence SNPs/indels, breadth of coverage and depth of coverage were performed using PCR-based (NA12878-1) and PCR-free (NA12878-2) down-sampling samples (~40×) of NA12878. Then, using down-sampling samples of NA12878, we also compared the performance of samples (PCR-free WGS) with differing DNA input (1 μg, 500, 300, and 200 ng) (Fig. 1). Finally, 11 clinical cases (with 19 variants previously validated) were collected to test the performance of the WGS pipeline (PCR-free).

2.2. Library preparation, genome sequencing, and bioinformatics analysis

In this study, 1 μg of DNA input were used for library preparation using either PCR-based protocol (MGIEasy FS DNA Library Prep Set, containing PCR-amplification steps after second bead purification) or PCR-free protocol (MGIEasy FS PCR-Free DNA Library Prep Set, omitting the PCR steps). The detailed library preparation workflows can refer to a paper.[17] Libraries with various amount of DNA input (1 μg, 500, 300, and 200 ng) were constructed using PCR-free library preparation protocol (Table 1). After quantification by BMG Labtech (Ortenberg, Germany) FLUOstar Omega and Agilent 2100 Bioanalyzer, (California, USA) all the libraries were then sequenced on the MGISEQ-2000 platform.

A standard bioinformatics analysis pipeline was implemented for all the samples. In short, after sequencing, fastq data was filtered to generate clean reads. The clean reads of each sample were then aligned to hg19 (the human reference genome) by Burrows-Wheeler Aligner.[15] To remove duplicate reads, MarkDuplicates was then used for analysis.[16] Genome Analysis Toolkit package was then used to perform realignment around indels and quality scores re-calibration, and to generate VCF files for each sample for further analysis. Depth and coverage analysis

were performed by BEDTools[18] and bamdst (https://github.com/shiquan/bamdst).

2.3. Sensitivity and positive predictive value (PPV) of variant detection

To evaluate the performance of different experimental methods in identifying true genotypes, NA12878 high-confidence calls (v3.3.2) were recognized as true-positive calls for evaluation. We further restricted the high-confidence calls to the high confidence region to calculate the sensitivity and PPV of different experimental methods. The percentage of high-confidence calls detected by our method in all the high-confidence calls in NA12878 was considered as the sensitivity for variation detection. The percentage of high-confidence calls detected by our method in all the variants detected by our method was considered as the PPV for variation detection. To filter out erroneous variant calls, genotype quality and depth of coverage were used.

2.4. Breadth of coverage for disease-related genes and CNVs

Genes in a single gene set may be incomplete to investigate the breadth of coverage for disease-related genes. In order to include all putative disease-related genes for evaluation, we generalize a new gene list (8394 genes) using the following 5 databases: ClinVar (accessed on February 19, 2019), Genetic Home Reference (accessed on July 2, 2019), Human Gene Mutation Database (HGMD) (professional March 2018), Online Mendelian Inheritance in Man (accessed on April 4, 2018) and Orphanet (accessed on July 2, 2019). NCBI annotation release 104 was used for the annotation of all the gene regions. Transcripts used in the HGMD database got priority for annotation. For genes without a definite transcript in the HGMD database, a combination of the regions of all transcripts was used for annotation. Coverage analysis of each sample for the 8394 genes were performed for evaluation.

For the analysis of the coverage of CNVs, we performed coverage analysis of the NA12878 samples using CNVs from DECIPHER database (version GRCh37_v9.29).

2.5. Analysis of 11 clinical cases and validation

To test the performance of the WGS pipeline (PCR-free) in real clinical setting, a total of 11 clinical cases (19 variants) were recruited. All the 19 variants were confirmed previously by methods other than WGS. One microgram of DNA input were used for library preparation for all the 11 clinical cases. Written informed consent was obtained from all the participants before sample collection.

3. Results

3.1. Overall performance of the 5 NA12878 samples

The libraries of all the 5 NA12878 samples were loaded and sequenced on 2 lanes of MGISEQ-2000. On average, there were 256.80 Gb clean data generated per sample (2 lanes). In this study, an average sequencing depth of 84.83-fold was achieved for each sample (Table 1).

In order to compare various experimental parameters (library preparation protocols and DNA inputs) at constant read depth, clean reads of each sample were randomly down-sampled from each sample using seqtk (https://github.com/lh3/seqtk). Finally, each sample was down-sampled to a sequencing depth of ~40× for further analysis.

As a result, NA12878 samples using PCR-based library preparation protocol (NA12878-1) and PCR-free library preparation protocol (NA12878-2) showed similar mean percentage of >98.88% and 98.62% for regions with ≥ 10× coverage. Samples using PCR-free WGS (NA12878-2, 3, 4, and 5) all showed a mean duplication rate of 2.5%, which was slightly lower than NA12878-1 (duplication rate of 3%).

3.2. Distribution of DP and GQ in PCR-based and PCR-free samples of NA12878

The DP and GQ parameters are widely used for assessment of variation quality in MPS technology. In this study, we investigated the distribution of the 2 main quality parameters (DP and GQ) for variation detection in PCR-based (NA12878-1) and PCR-free (NA12878-2) samples of NA12878 at 1 μg DNA input. Here, down-sampling samples of NA12878 (mean DP of ~40×) was used for comparison. In general, the quality of sample with PCR-free library construction method (NA12878-2) is shown to be better than sample with PCR-based library construction method (NA12878-1) (Fig. 2).

The distribution of DP was normal-like for the 2 samples (Fig. 2). The distribution of DP for all the variants showed more uniform quality for sample with PCR-free library preparation protocol (NA12878-2) than that for sample with PCR-based protocol (NA12878-1), especially for indel detection (Fig. 2). The proportion of variants for NA12878-2 (93.34%) with a DP of >20× was higher than NA12878-1 (89.19%), indicating a better DP distribution for PCR-free library preparation method of WGS.

The vast majority of variants called for all the samples had a GQ close to 100 (Fig. 2). Variants detected by sample with PCR-free WGS (NA12878-2) showed higher quality than those detected by PCR-based WGS (NA12878-1). The proportion of variants called for NA12878-1 (17.42%) with a GQ of <20 was 1.67% more than that for NA12878-2 (15.75%). For indel detection, more proportion of variants were detected when the GQ is <65 in NA12878-1 (Fig. 2). These results showed that the variation quality of PCR-free WGS is better than PCR-based WGS.

3.3. Sensitivity and PPV of variant detection in PCR-based and PCR-free samples of NA12878

In order to investigate the impact of different library preparation methods in identifying true genotypes, NA12878 high-confidence calls (v3.3.2) were recognized as true-positive calls for evaluation. To calculate the sensitivity of each sample, variants located in the high confidence region (v3.3.2) were further recognized as "gold standard" calls. To filter out erroneous variant calls, GQ (≥20) and DP (≥10) were used.

NA12878-2 (PCR-free WGS) showed higher sensitivity and PPV for both SNP and indel detection (Fig. 3). For homozygous and heterozygous SNPs detection, the sensitivity of NA12878-2 is slightly better than NA12878-1. PCR-free WGS showed great improvement for homozygous and heterozygous indels detection (Fig. 2). The sensitivity for homozygous and heterozygous indels detection in NA12878-2 is >99.22% and 91.28% respectively, while sensitivity of NA12878-1 for homozygous and heterozygous indels detection is only 88.05% and 88.76%. For SNP and indel detection, the PPV (high confidence region) of NA12878-2

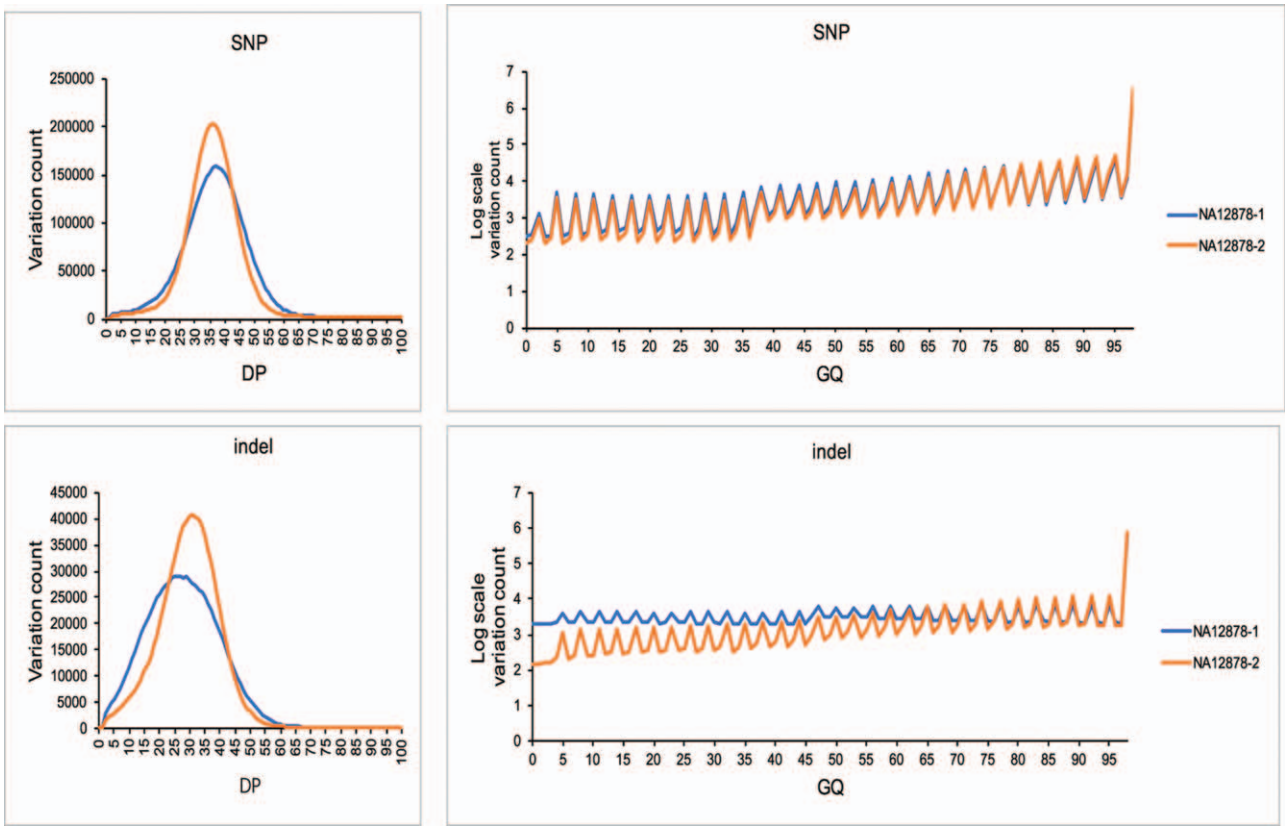

Figure 2. Distribution of (depth of coverage) DP and (genotype quality) GQ in PCR-based and PCR-free samples of NA12878. DP = depth of coverage, GQ = genotype quality

(PCR-free WGS) is also better than NA12878-1 (PCR-based WGS) (Fig. 3). Heterozygous indels of NA12878-1 (PCR-based WGS) showed the lowest PPV of 82.87%. In general, the sensitivity and PPV of variant detection for samples with PCR-free library preparation protocol (NA12878-2) is better than samples with PCR-based library preparation method (NA12878-1) (Fig. 3).

3.4. Depth and breadth of coverage for disease-related genes and CNVs in PCR-based and PCR-free samples of NA12878

In this part, the breadth of coverage for the 8394 genes was first evaluated in PCR-based (NA12878-1) and PCR-free (NA12878-2) samples of NA12878 at 1 μg DNA input. The 8394 disease-

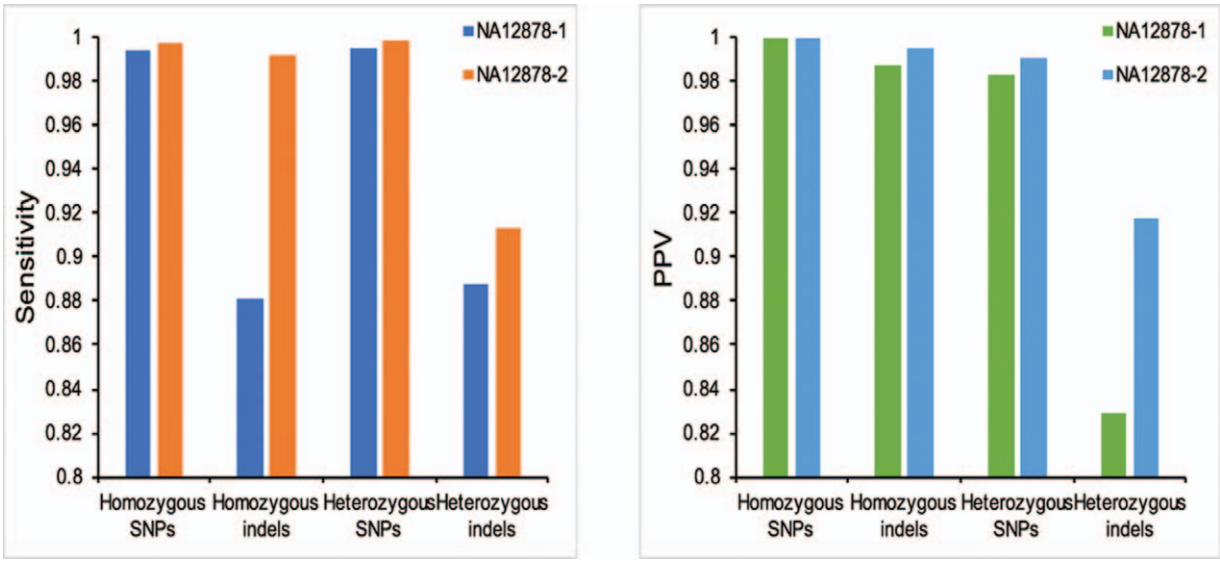

Figure 3. Sensitivity and positive predictive value (PPV) of variant detection in PCR-based and PCR-free samples of NA12878.

related genes were compiled using 5 databases (ClinVar, Genetic Home Reference, HGMD, Online Mendelian Inheritance in Man, and Orphanet). The percent of targeted bases covered at >10× depth has been reported to be related to the sensitivity for heterozygous SNV detection in whole-exome sequencing.[19] Here, we calculated the percent of bases covered at >10× depth for exons of the all the 8394 genes. As a result, none of the samples of NA12878 covered 100% of the coding exons. The results of samples with PCR-free library preparation protocol method (NA12878-2) is slightly better than samples with PCR-based library preparation method (NA12878-1). For NA12878-2, the percent of bases covered at >10× depth for the 8394 putative disease-related genes was >99.84%, while 99.44% of the exon regions was covered in NA12878-1.

We also compared the breadth of coverage performance of the 2 samples for The American College of Medical Genetics and Genomics (ACMG) 59 genes[20] (see Table S1, Supplemental Digital Content, http://links.lww.com/MD/G640). The proportion of the ACMG 59 genes covered 100% (>10×) was 98.30% and 93.22% for NA12878-2 and NA12878-1 respectively. Sites of all genes that are covered >10× was 99.97% and 99.78% for sample with PCR-free library preparation method (NA12878-2) and sample with PCR-based library preparation method (NA12878-1). The breadths of coverage are slightly better for sample with PCR-free protocols. We also examined finished genes of the ACMG 59 gene set at ≥20× coverage that could provide 99% sensitivity for heterozygous SNVs.[19] A percentage of 79.66% and 57.63% genes were covered 100% for sample with PCR-free library preparation method (NA12878-2) and sample with PCR-based library preparation method (NA12878-1) respectively.

In this study, the breadth of coverage of CNVs in the DECIPHER database was also investigated for the 2 NA12878 samples (see Table S2, Supplemental Digital Content, http://links.lww.com/MD/G641). Most CNVs in the DECIPHER database can be well covered (>95%) at >10× depth for the 2 samples. The breadths of coverage are slightly better for sample with PCR-free protocol for CNVs in the DECIPHER database. A percentage of 92.42% CNVs showed better coverage for NA12878-2 (PCR-free WGS) than NA12878-1 (PCR-based WGS).

3.5. Impact of DNA input in PCR-free samples of NA12878

After comparison of the performance of WGS with PCR-based and PCR-free protocols, we also investigated the impact of DNA input (1 μg, 500, 300, and 200 ng) on variant detection in PCR-free samples of NA12878 (NA12878-2, 3, 4, and 5). We compared the DP and GQ, sensitivity for SNV/indels detection, breadth of coverage of disease-related genes, and CNVs in the 4 PCR-free samples of NA12878.

First, we investigated the GQ and DP distribution in PCR-free samples of NA12878 (NA12878-2, 3, 4, and 5). In general, the distribution of DP was normal-like for all the samples. The proportion of variants with ≥10× depth increased with increasing DNA input (Fig. 4). NA12878-2 (1 μg DNA input) showed the highest proportion of 98.62% with ≥10× depth (Fig. 4). The GQ of vast majority of variants called by PCR-free samples of NA12878 (NA12878-2, 3, 4, and 5) is ~100. The proportion fluctuated along with the GQ scores. For both SNP and indel detection, the proportion of variants with ≥20 GQ also increased with increasing DNA input (Fig. 4). NA12878-2 (1 μg DNA input) showed the highest proportion of 99.37% with ≥20 GQ (Fig. 4). These results showed that the performance of samples with higher DNA input is better than samples with lower DNA input for PCR-free WGS. DNA input may influence the variant quality in samples with PCR-free WGS.

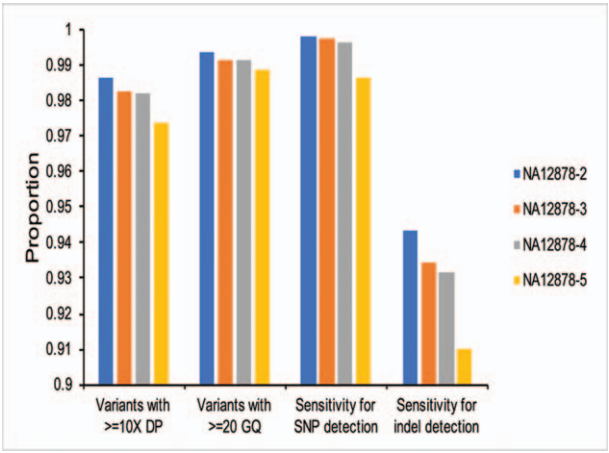

Figure 4. Impact of DNA input on variant detection in PCR-free samples of NA12878.

DNA input may also influence the detection sensitivity of NA12878 samples. In order to investigate the impact of DNA input in identifying true genotypes, NA12878 high-confidence calls (v3.3.2) were recognized as true-positive calls for evaluation. To calculate the sensitivity of each sample, variants located in the high confidence region (v3.3.2) were further recognized as "gold standard" calls. To filter out erroneous variant calls, GQ (≥20) and DP (≥10) were used. With the same library preparation protocols, the sensitivity visibly increased with increasing DNA input (Fig. 4), indicating that variant detection sensitivity is positively correlated with increasing DNA input in this study (Fig. 4). As a result, NA12878-2 showed the highest sensitivity for both SNP (99.80%) and indel (94.34%) detection. Heterozygous indels of sample NA12878-5 (200 ng DNA input) showed the lowest sensitivity of 87.92%. DNA input may influence the detection sensitivity in samples with PCR-free WGS.

We also investigated the breadth of coverage of PCR-free samples of NA12878 (NA12878-2, 3, 4, and 5) in the 8394 disease-related genes and CNVs from DECIPHER database. To filter out erroneous variant calls, GQ (≥20) and DP (≥10) were used. In general, PCR-free samples with differing DNA input showed similar coverage of both putative disease-related genes and CNVs. NA12878-5 with the lowest DNA input of 200 ng showed more lower coverage regions in this study.

3.6. Analysis of 11 clinical cases

In the present study, 11 clinical cases were collected between October 2018 and January 2021 in Changsha Hospital for Maternal & Child Health Care Affiliated to Hunan Normal University, Shenzhen Children's Hospital and The University of Hongkong Shenzhen Hospital to test the performance of the WGS pipeline (PCR-free). All the 19 variants in the 11 cases were confirmed previously by methods other than WGS, including 7 known variants (5 variants were classified as pathogenic, 1 variant was classified as likely pathogenic, and 1 variant was classified as variant of uncertain significance) and 12 novel

Table 2

Summary of detected variants in 11 clinical cases.

Sample name	Final diagnosis/Inheritance	Variant	Zygosity	ACMG variant classification*	Status†	Result of WGS pipeline (PCR-free)
P1	Mental retardation, autosomal dominant 23/AD	NM_001080517.1(SETD5):c.3167C>T(p.Ala1056Val)	het	VUS	novel	Detected
	Pitt-Hopkins syndrome/AD	NM_001083962.1(TCF4):c.305-20T>C	het	VUS	novel	Detected
	Asparagine synthetase deficiency/AR	NM_133436.3(ASNS):c.-59-9delT	hom	VUS	novel	Detected
	Mental retardation, autosomal dominant 52/AD	NM_018489.2(ASH1L):c.1304C>T(p.Pro435Leu)	het	VUS	novel	Detected
P2	Neurofibromatosis 1/AD	NM_000267.3(NF1):EX1-EX58E Del	het	P	25325900	Detected
P3	Spinocerebellar ataxia, autosomal recessive 8/AR	NM_152393.2(KBTBD5):c.1516A>C(p.Thr506Pro)	hom	P	31360996; 23746549	Detected
P4	Developmental and epileptic encephalopathy 11/AD	NM_021007.2(SCN2A):c.3955C>T(p.Arg1319Trp)	het	P	28379373	Detected
P5	Leukoencephalopathy with ataxia/AR	NM_004366.5(CLCN2):c.773-15C>G	het	VUS	novel	Detected
	Leukoencephalopathy with ataxia/AR	NM_004366.5(CLCN2):c.233G>A(p.Arg78His)	het	VUS	novel	Detected
	Deafness, autosomal dominant 3A/AD	NM_004004.5(GJB2):c.299_300delAT(p.His100Argfs*14)	het	P	20095872; 12111646; 21162657	Detected
P6	Brugada syndrome 3/AD	NM_000719.6(CACNA1C):c.5747A>G(p.Gln1916Arg)	het	LP	28493952; 27871843; 27005929	Detected
	Brugada syndrome 3/AD	NM_000719.6(CACNA1C):c.4393T>C(p.Phe1465Leu)	het	VUS	29306897	Detected
P7	Epidermolysis bullosa with congenital localized absence of skin and deformity of nails/AD	NM_000094.3(COL7A1):c.5990G>A(p.Gly1997Asp)	het	VUS	novel	Detected
P8	Myopathy, centronuclear, 1/AD	NM_022485.4(MTMR14):c.1577G>T(p.Arg526Leu)	het	VUS	novel	Detected
P9	Mental retardation, X-linked, syndromic, Houge type/XL	NM_014927.3(CNKSR2):c.1393+10C>G	hem	VUS	novel	Detected
	Helsmoortel-Van der Aa syndrome/AD	NM_015339.2(ADNP):c.3137A>G(p.Gln1046Arg)	het	VUS	novel	Detected
P10	Sotos syndrome 1/AD	NM_022455.4(NSD1):c.2704G>T(p.Glu902*)	het	LP	novel	Detected
P11	Retinitis pigmentosa 39/AR; Usher syndrome type 2A/AR	NM_206933.2(USH2A):c.3788G>A(p.Trp1263*)	het	P	21686329	Detected
	Retinitis pigmentosa 39/AR; Usher syndrome type 2A/AR	NM_206933.2(USH2A):c.5572+1136G>A	het	VUS	novel	Detected

* LP = likely pathogenic, P = pathogenic, VUS = variant of uncertain significance, WGS = whole-genome sequencing.

† "novel" indicates that the variant has not yet been reported as far as we know. Numbers are the PMIDs of the literature where the variation has been reported.

variants (1 variant was classified as likely pathogenic, and 11 variants were classified as variant of uncertain significance) (Table 2).[21] We applied the WGS pipeline (PCR-free) to all the 11 clinical cases. All the 19 previously confirmed variants were also successfully detected using the WGS pipeline (PCR-free) (Table 2). These results further demonstrated the sensitivity of the method.

4. Discussion

In this study, we focused on the impact of 2 experimental parameters in the upstream step of WGS analysis (library preparation protocol and DNA input) on variant detection. We comprehensively analyzed and compared the performance of each method using 5 NA12878 samples. After down-sampling to a sequencing depth of ~40×, the performance of GQ and DP for different samples were evaluated first. In addition, we further assessed the variation detection sensitivity with high-confidence calls in the high confidence region from Genome in a Bottle. The breadth of coverage of disease-related genes and CNVs was also compared. As a result, samples with PCR-free protocol showed better performance in DP and GQ distribution, SNV/indel detection, and breadth of coverage of disease-related genes and CNVs, thereby revealing how experimental parameters affect variation detection. In this study, the analysis of samples with different experimental methods provided additional insight and choice for clinical variant detection.

Generally, various sequencers share a basic MPS workflow, including preparation of template, library construction, sequencing, and analysis. Various experimental parameters were provided by different platforms. As for the DNA input, the data generated by extreme low DNA input may not always pass the quality control for different platforms, and that's the reason why we selected 200 ng as the lowest amount of DNA input. The amount of DNA used for library construction can be much lower than 200 ng, such as DNA extracted from plasma and dried-blood spot. The performance characterization of extremely low amount DNA input (<50 ng) is another interesting research topic. Another limitation of this study is that, we did not perform CNV detection comparison in the 5 NA12878 samples, because there's no well-accepted "gold standard" CNV call set for benchmarking, nor "best practices" workflow for the detection of CNVs. Instead, the depth and breadth of coverage for CNVs was evaluated using the 5 NA12878 samples.

The successful applications of WGS in real clinical setting requires comprehensive assessment of experimental parameters. In this study, we have systematically evaluated the performance of

different methods for clinical WGS, and which illustrates how experimental parameters affect variant detection. The results provide additional insight and choice for clinical variant detection.

Acknowledgments

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

Author contributions

Guiju Zhou, Meizhen Zhou, Fanwei Zeng, Ningzhi Zhang, Yan Sun, and Lijie Song designed the research. Yan Sun wrote the first draft of the article. Shihao Zhou, Guojun Yun, Jiansheng Xie collected samples and made clinical diagnosis. Xiaodan Wang, Chunna Fan, and Wentao Dai designed and performed the experiments. Zhihong Qiao, Fengxia Liu, Yaoshen Wang, Zhonghai Fang, Zhongming Tian, Jun Sun, and Zhiyu Peng performed data analysis. Guiju Zhou, Meizhen Zhou, Fanwei Zeng, Ningzhi Zhang, Yan Sun, Lijie Song, Xueqin Guo, and Wentao Dai contributed to drafting and revising the manuscript. All authors reviewed the manuscript.

Conceptualization: Fanwei Zeng.
Data curation: Zhihong Qiao, Xiaodan Wang, Fengxia Liu, Chunna Fan, Yaoshen Wang, Zhonghai Fang, Zhongming Tian, Wentao Dai, Jun Sun, Zhiyu Peng.
Formal analysis: Guiju Zhou, Meizhen Zhou, Ningzhi Zhang, Yan Sun, Lijie Song.
Funding acquisition: Guiju Zhou.
Investigation: Guiju Zhou, Meizhen Zhou, Ningzhi Zhang, Yan Sun, Lijie Song.
Methodology: Guiju Zhou, Meizhen Zhou, Fanwei Zeng, Ningzhi Zhang, Yan Sun, Lijie Song.
Project administration: Guiju Zhou, Meizhen Zhou, Ningzhi Zhang, Yan Sun, Lijie Song.
Resources: Shihao Zhou, Guojun Yun, Jiansheng Xie, Yan Sun, Lijie Song.
Supervision: Guiju Zhou, Meizhen Zhou, Ningzhi Zhang, Yan Sun, Lijie Song.
Validation: Yan Sun.
Writing – original draft: Yan Sun.
Writing – review & editing: Guiju Zhou, Meizhen Zhou, Fanwei Zeng, Ningzhi Zhang, Xueqin Guo, Wentao Dai, Yan Sun, Lijie Song.

References

[1] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17:333–51.
[2] DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8.
[3] Meynert AM, Ansari M, FitzPatrick DR, Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics 2014;15:1–11.
[4] 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;4:63–5.
[5] Fang H, Wu Y, Narzisi G, et al. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med 2014;6:1–17.
[6] Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? Hum Genet 2016;135:359–62.
[7] Soden SE, Saunders CJ, Willig LK, et al. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Sci Transl Med 2014;6:265ra168.
[8] Lionel AC, Costain G, Monfared N, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med 2018;20:435–43.
[9] Zhu Q, Hu Q, Shepherd L, 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;24:1207–13.
[10] Londin ER, Keller MA, D'Andrea MR, et al. Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. BMC Genomics 2011;12:1–9.
[11] Aird D, Ross MG, Chen WS, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 2011;12:1–14.
[12] Clark MJ, Chen R, Lam HY, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 2011;29:908–14.
[13] Sulonen AM, Ellonen P, Almusa H, et al. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol 2011;12:1–18.
[14] Meienberg J, Zerjavic K, Keller I, et al. New insights into the performance of human whole-exome capture platforms. Nucleic Acids Res 2015;43:e76.
[15] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
[16] McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.
[17] Li Q, Zhao X, Zhang W, et al. Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS platform. BMC Genomics 2019;20:1–13.
[18] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.
[19] Kong SW, Lee IH, Liu X, Hirschhorn JN, Mandl KD. Measuring coverage and accuracy of whole-exome sequencing in clinical context. Genet Med 2018;20:1617–26.
[20] Kalia SS, Adelman K, Bale SJ, 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;19:249–55.
[21] Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17:405–24.