Journal of
Next Generation Sequencing & Applications

Research Article Open Access

Comparison of Variant Calls from Whole Genome and Whole Exome Sequencing Data Using Matched Samples

Björn N[1,a], Pradhananga S[2,a,*], Sigurgeirsson B[2,3], Lundeberg J[2], Gréen H[1,2,4,b] and Sahlén P[2,b]

[1]Clinical Pharmacology, Division of Drug Research, Department of Medical and Health Sciences, Linköping University, Linköping, Sweden
[2]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, Division of Gene Technology, KTH Royal Institute of Technology, Solna, Sweden
[3]School of Engineering and Natural Sciences, University of Iceland, Reykjavík, Iceland
[4]Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
a N.B. and S.P. share first authorship.
b H.G. and P.S. share last authorship.

Abstract

Whole exome sequencing (WES) has been extensively used in genomic research. As sequencing costs decline it is being replaced by whole genome sequencing (WGS) in large-scale genomic studies, but more comparative information on WES and WGS datasets would be valuable. Thus, we have extensively compared variant calls obtained from WGS and WES of matched germline DNA samples from 96 lung cancer patients. WGS provided more homogeneous coverage with higher genotyping quality, and identified more variants, than WES, regardless of exome coverage depth. It also called more reference variants, reflecting its power to call rare variants, and more heterozygous variants that met applied quality criteria, indicating that WGS is less prone to allelic drop outs. However, increasing WES coverage reduced the discrepancy between the WES and WGS results. We believe that as sequencing costs further decline WGS will become the method of choice even for research confined to the exome.

Keywords: Whole genome sequencing; Whole exome sequencing; Coverage; Depth; Genotyping quality; Discordant; Concordant; Variant calls; Single-nucleotide variants

Abbreviations: WGS: Whole Genome Sequencing; WES: Whole Exome Sequencing; SNV: Single-Nucleotide Variant; INDEL: Insertion or Deletion; REF: Reference Homozygous; HET: Heterozygous; HOM: Variant Homozygous

Introduction

Whole Exome Sequencing (WES) has proven utility for accessing sequences of the human genomes protein-coding regions. These regions account for 1.5% of the genome, so WES requires far fewer sequencing reads than whole genome sequencing (WGS). Thus, it enables cheaper analysis of multiple samples and facilitates large-scale genomic investigations of exomic sequence variants in diverse contexts, including mutations involved in carcinogenesis and other complex diseases. However, preparing samples for WES is labor-intensive and the exome capture probes used can introduce inherent GC biases in PCR amplification, resulting in uneven coverage and increases in frequencies of duplicated fragments [1-4]. In contrast, WGS enables access to a larger portion of the genome, including protein-coding exons, introns, non-coding RNA, regulatory and intergenic regions. Thus, WGS requires large quantities of sequencing reads for adequate coverage, which can limit both numbers of samples used in studies and clinical implementation of the technique [5]. Moreover, as WGS datasets are much larger than corresponding WES datasets, their storage and analysis are more computationally demanding [5-8]. However, WGS library preparation is PCR-free with current protocols [9], and generally requires less sequencing coverage than preparation of WES libraries.

As sequencing costs are falling [7,10] the financial advantage of WES over WGS is declining, while the informational advantages of WGS are maintained. Thus, many researchers are now using WGS rather than WES even for exomic studies. Hence, there is a clear need for robust comparisons of the two approaches in terms of the quality and abundance of data provided on genomic regions that they both cover.

There have been several comparisons of WGS and WES datasets. For example, 10 matched pairs of WES and WGS datasets obtained from analyses of non-tumor samples from The Cancer Genome Atlas have been compared [11]. The findings included indications that WGS and WES required 14X and 39X coverage, respectively, to provide 95% sensitivity in detecting heterozygous variants. A more recent study, based on samples from six individuals, found that WGS data is of higher quality, more uniform and includes fewer false positives than WES data [12]. Another recent comparison, based on sets of five matched samples, concluded that in clinical settings WGS is better than WES for addressing known disease-causing mutations in WES target regions [13]. However, all these studies were limited to a few samples, and broader comparisons would be valuable. Therefore, in the study reported here we prepared matched WES and WGS libraries using germline DNA samples from 96 human lung cancer patients. The WES and WGS libraries were sequenced using Illumina HiSeq 2500 and Illumina HiSeq X Ten platforms, respectively. Reads from all 192 sequencing libraries were aligned to a reference genome and variants were called from the aligned reads in an identical manner. Variants called from the WES and WGS datasets, and quality metrics of the data acquired by the two sequencing approaches, were then compared. The results provide an unprecedentedly detailed comparison of paired WES and WGS datasets, acquired and analyzed using the most recently released hardware and software tools, at the time of the study.

***Corresponding author:** Sailendra Pradhananga, Department of Gene Technology, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Solna, Sweden, Tomtebodavagen 23A, SE–17165 Solna, Sweden, Tel: + (46)720241259; E-mail: sailendra.pradhananga@scilifelab.se

Received May 21, 2018; **Accepted** July 25, 2018; **Published** August 01, 2018

Citation: Björn N, Pradhananga S, Sigurgeirsson B, Lundeberg J, Green H, et al. (2018) Comparison of Variant Calls from Whole Genome and Whole Exome Sequencing Data Using Matched Samples. Next Generat Sequenc & Applic 5: 154. doi:10.4172/2469-9853.1000154

Materials and Methods

Samples and ethical approval

Between 2006 and 2008, patients diagnosed with non-small cell lung cancer at the Karolinska University Hospital, Stockholm, Sweden, were recruited for the study. The regional ethics committee in Stockholm approved the study (DNR-03-413 with amendment 2016/258-32/1) and patients provided written informed consent in accordance with the Helsinki Declaration. For this technical study, peripheral blood samples from 96 of the recruited patients were used.

Sample preparation and sequencing

QIAamp DNA mini-kits (VWR International, Stockholm, Sweden) were used according to the manufacturer's recommended protocol to extract DNA from the 96 collected peripheral blood samples. From each of the DNA samples two sequencing libraries were prepared: one for WES and one for WGS, respectively using a Nextera Rapid Capture Exome kit (FC-140-1003) and a TruSeq DNA PCR-Free Library Preparation kit (FC-121-3001) (both from Illumina, San Diego, CA, USA). The WES and WGS libraries were subsequently sequenced at the ISO-certified sequencing center of the Science for Life Laboratory, Stockholm, Sweden, using Illumina HiSeq 2500 and Illumina HiSeq X Ten platforms, respectively. A previous internal pilot study revealed that these two platforms yield comparable and unbiased results, see Supplementary Document S1.

Alignment and variant calling

Raw fastq files were mapped to the human reference genome, GRCh37/hg19, using the BWA aligner [14] (version 0.7.8) and variants were called using the Genome Analysis Toolkit (GATK) [15] (version 3.3.0) applying the developers' best practices [16]. Variant calling was confined to the exome target region as defined by the Nextera Rapid Capture Exome Targeted Regions Manifest (version 1.2) using the HaplotypeCaller module of GATK [15].

Subgrouping and joint genotyping

The WES and WGS coverage of the paired samples (Figure 1) shows uneven exome coverage and uniform genome coverage. To assess the impact of coverage on sequencing we only included samples with either very high or very low exome coverage. To do this, the samples were split into the subgroups SC_high and SC_low representing the 24 samples with the highest and the 24 samples with the lowest mean WES coverage, respectively. This approach provided a robust design for comparisons of high and low WES coverage with WGS. In further analyses we assessed the variant calls from the WGS samples and corresponding variant calls from WES runs with different coverage in SC_high and SC_low separately. The samples with intermediate WES coverage were not investigated further, except when looking into coverage over differing GC content. We used the GenotypeGVCFs module of GATK [15] for joint genotype calling to obtain information on variants evident in each of these subgroups.

Filtering of quality variants

High-quality SNVs and INDELs in SC_low and SC_high samples were filtered using the GATK [15] VariantFiltration module with the filtering criteria listed in Table 1, extracted from GATK's documentation on VariantFiltration and VariantRecalibrator. This "hard" filtration option was applied instead of the filtering offered by the VariantRecalibrator and ApplyRecalibration modules because of their requirements for higher sample numbers.

Comparison of WES and WGS variants

Only bi-allelic loci detected by WES and WGS were considered in the comparison. All genetic variants were categorized into a specific type and group. The variant types were categorized as: reference homozygous (REF), for the genomic loci which are identical to the reference genome; heterozygous (HET), for loci which differ from the reference genome in a single base; or variant homozygous (HOM), for loci that differ from the reference genome on both bases (alleles). The variant groups were categorized as: called in both (CB), for variants called using both WES and WGS with the same zygosity; discordant, for variants called using both WES and WGS, but with differing zygosity; exome only, for variants only called using WES; or genome only, for variants only called using WGS. The number, coverage and quality of the variant types and groups categorized in this manner in the subgroups SC_low and SC_high were then compared between WES and WGS. All presented plots and statistical comparisons were prepared using the statistical software R [17] (version 3.4.2).

Coverage and GC content

The relationship between mean coverage and GC content was compared in WES and WGS datasets from all 96 samples to validate, with a bigger sample cohort, previously reported findings that WGS coverage is more robust than WES coverage across the entire GC spectrum [13,18].

Results

To compare the results obtained using the WGS and WES technologies, we determined fundamental characteristics including numbers of variants called, coverage, and genotyping quality metrics. We also analyzed discordant variants, i.e., genomic calls for which WGS and WES analyses indicated differing zygosity.

Sequencing and mapping in SC_low and SC_high subgroups

The WES and WGS platforms on average generated 80.32 and 720.63 million raw read pairs per sample, respectively. For the subgroups SC_low and SC_high the WGS generated 704.67 ± 66.23 and 736.61 ± 99.01 million read pairs, respectively. Further, when mapping the WGS reads to the predefined Exome Targeted Region SC_low and SC_high respectively yielded 18.04 ± 1.94 and 20.13 ± 4.56 million aligned reads (p=0.04, t-test). Although slightly different, it has to be considered relatively homogeneous. For WES on the other hand, we observed 47.34 ± 6.76 and 109.40 ± 9.45 (p=2.20e^{-16}, t-test) million aligned reads for SC_low and SC_high, respectively. However, this difference was expected as the two subgroups were selected based on their exome sequencing coverage. An overview of the sequencing output for the samples in SC_low and SC_high is given in the Supplementary Tables S1, S2 and S3.

Number and coverage of called variants in SC_low and SC_high subgroups

We then examined average numbers of SNVs in both SC_low and SC_high subgroups and observed that WES coverage of SNVs was positively correlated with concordance rates between WES and WGS calls (p<2.22e^{-16}, Fisher's exact test; Figures 2A, 2B and Table 2). Additionally we, as expected, observed that increases in WES coverage increased the number of heterozygous (HET) variant calls (p=3.85e^{-9}, Fisher's exact test; Figure 2C and 2D).

We next investigated whether increases in WES read depth improved variant call coverage, relative to WGS values. We found

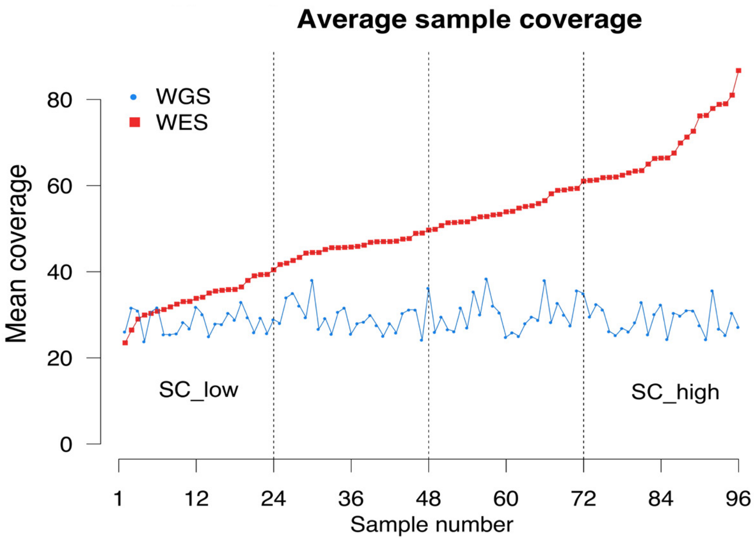

Figure 1: The mean coverage per sample for the matched WES (whole exome sequencing) and WGS (whole genome sequencing) data sorted on increasing exome coverage. From this the samples were divided in to two subgroups: SC_low (sub comparison of samples with low exome coverage) including the 24 samples with the lowest mean exome coverage and SC_high (sub comparison of samples with high exome coverage) including the 24 samples with the highest mean exome coverage.

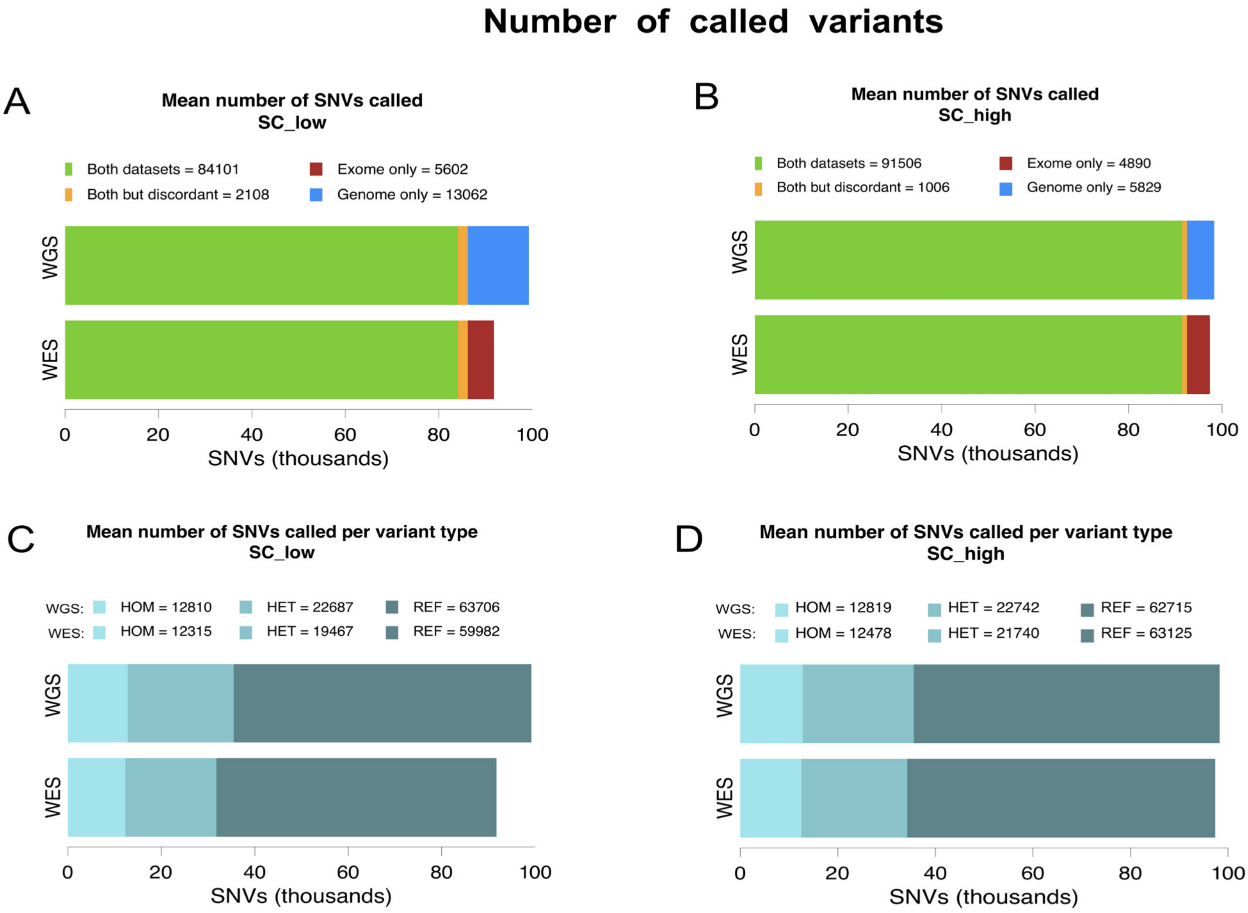

Figure 2: Depicts the average number of variants called in WGS (whole genome sequencing) and WES (whole exome sequencing) for SC_low (sub comparison of samples with low exome coverage) and SC_high (sub comparison of samples with high exome coverage): A) for SC_low and B) for SC_high show the mean number of variants found in both WGS and WES, both but discordant (different variant type in WGS and WES), exome only and genome only. C) for SC_low and D) for SC_high show the mean number of variants for the different variant types (reference homozygous (REF), heterozygous (HET) and variant homozygous (HOM)).

Parameters	SNVs	INDELs
Qual By Depth (QD)	<2.0	<2.0
RMS Mapping Quality (MQ)	<40.0	n. a.
Fisher Strand (FS)	>60.0	>200.0
Strand Odds Ratio (SOR)	>3.0	>10.0
Mapping Quality Rank Sum Test (MQ Rank Sum)	<-12.5	n. a.
Read Pos Rank Sum Test (Read Pos Rank Sum)	<-8.0	<-20.0
Inbreeding Coefficient (Inbreeding Co-eff)	n. a.	<-0.8

Table 1: Parameters used for hard filtering scheme to filter SNVs and INDELs using the GATK module VariantFiltration applied to WES and WGS data of SC_low and SC_high separately.

Number of variants	WES	WGS
SC_low	91811 ± 1182	99271 ± 1802
SC_high	97401 ± 432	98340 ± 2435
Coverage of variants		
SC_low	37.25 ± 5.66	29.60 ± 3.02
SC_high	74.56 ± 11.91	30.25 ± 3.58
Genotyping quality of variants		
SC_low	75.83 ± 5.19	80.84 ± 5.27
SC_high	92.54 ± 1.87	81.43 ± 5.42
Discordant variants	**WGS/WES**	
SC_low	2108 ± 523	
SC_high	1007 ± 429	

Table 2: Summary statistics of different parameters in WGS and WES comparisons displaying average numbers ± the standard deviation.

WES variants had higher coverage depth than WGS variants (p = 9.78e^{-7} and 2.86e^{-16}, t-test; for SC_low in Figure 3A and SC_high in Figure 3B, respectively). Additionally, as expected, the average depth of WES coverage in SC_high exceeded the average depth of WES coverage in SC_low (p=3.03e^{-15}, t-test). Moreover, WES coverage of discordant variants was lower for the SC_low subgroup than for the SC_high subgroup (p=1.28e^{-13}, t-test; Figures 3A and 3B). In addition, the coverage of variants called only using WES were significantly lower for homozygous (HOM) variant calls than for heterozygous (HET) variant calls (p<2.22e^{-16}, t-test; for both SC_low and SC_high), regardless of the coverage in each subgroup (Figures 3C and 3D). A slight reduction in coverage for the same variant types is also seen for WGS. Low coverage can be due to difficulty in alignment of sequences, hence HOM variant calls in areas with low mappability or alignability should be interpreted cautiously, whatever sequencing method is used.

Genotyping quality in SC_low and SC_high subgroups

We then examined genotyping quality metrics for variant calls in both subgroups. As shown in Figures 4A, 4B, and Table 2, WES-based genotyping quality increased with increases in coverage (difference in this respect between SC_low and SC_high subgroups: p=8.35e^{-15}, t-test). In addition, at both low and high WES coverage, WES-based calls of discordant variants had lower quality than corresponding WGS-based calls (p=2.26e^{-16} and 4.11e^{-8}, t-test; for SC_low and SC_high, respectively). These results confirm that increasing exome coverage improves the genotyping quality and accuracy of variant calls, thereby diminishing discordant calls. Further evaluation of the homozygous (HOM) variant calls only called using each sequencing technology had poorer genotyping quality than other types of variants (Figures 4C and 4D).

Discordant variant calls in SC_low and SC_high subgroups

Discordant variants are the genomic loci where variant calls were of different zygosity when comparing between the two sequencing technologies. Although there were few discordant variants, 2108 and 1007 respectively in the SC_low and SC_high subgroups (Figures 5A, 5B and Table 2) they had interesting characteristics in terms of coverage and genotyping quality. Frequencies of all categories of these variant call discordances significantly differed between the SC_low and SC_high subgroups (p<0.01, Fisher's exact test) except for variants discordantly called as homozygous (HOM) by WES and reference (REF) by WGS, and vice-versa, however, these types of discordances were almost never occurring and of low quality in both subgroups and using both sequencing platforms. In addition, discordant variants called as heterozygous (HET) in WGS datasets but either reference (REF) or homozygous (HOM) in WES datasets had superior WGS genotyping quality (Figures 5C and 5D) and coverage (Figures 5E and 5F) in both the SC_low and SC_high subgroups. This implies that many of the discordant variants called after WES are potentially false positive calls. Of particular interest are the discordant variants that were called as heterozygous (HET) in WGS datasets but as reference (REF) in WES datasets (in which these variants were poorly genotyped, even at higher coverage). This shows that WGS is a powerful method for genotyping variants even at its, compared to WES, moderate average coverage.

Coverage over GC content

Previous reports have implicated that GC-rich sequences are particularly prone to sequencing errors and bias in the sequencing platforms. Therefore, as shown in Figure 6, we compared the mean coverage of WES and WGS over the entire GC spectrum using all of our sequenced samples. The WES coverage increased sharply from 20% to 30% GC content, then declined sharply when the GC content exceeded

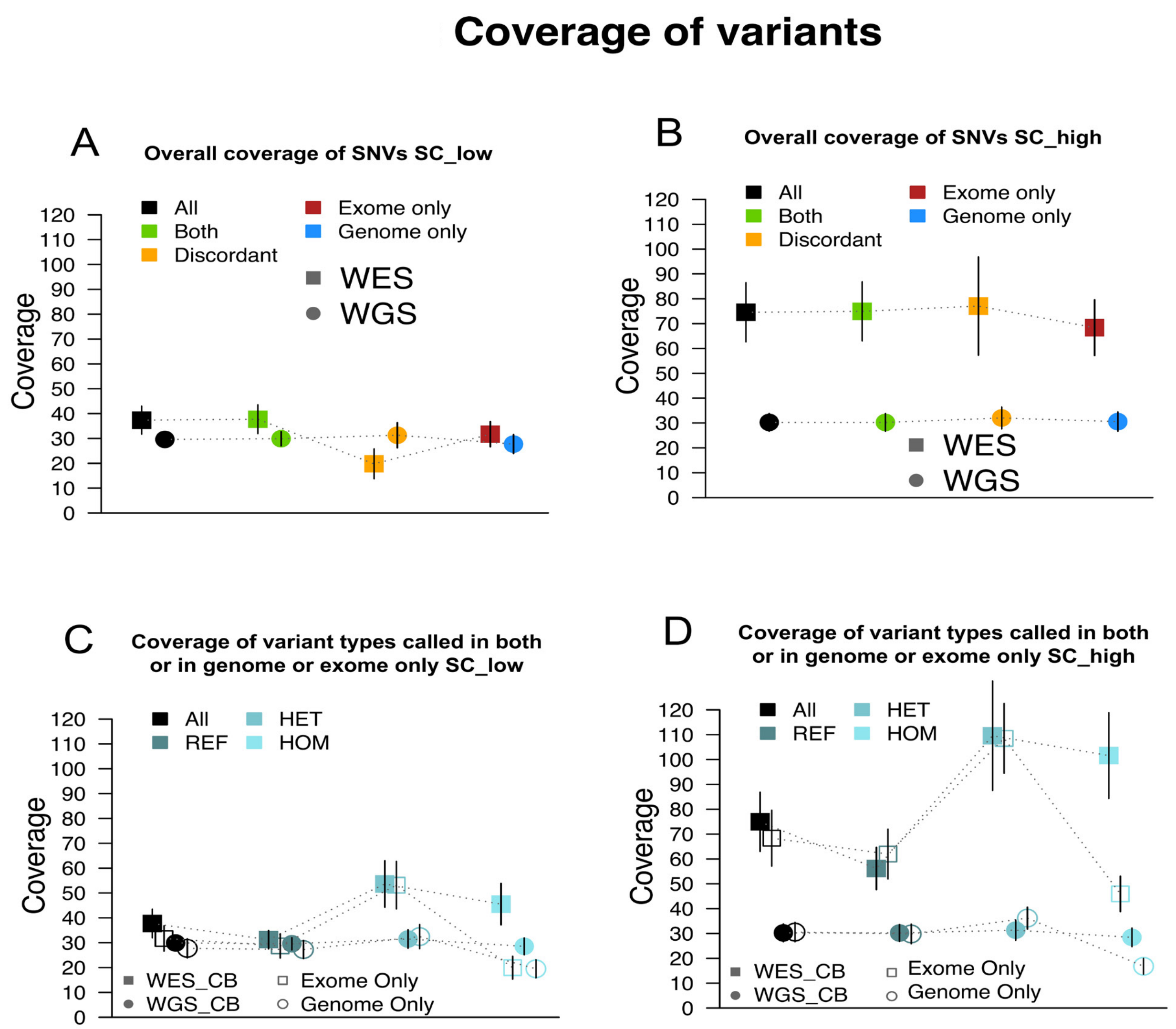

Figure 3: Average coverage of called variants in WGS (whole genome sequencing) and WES (whole exome sequencing) for SC_low (sub comparison of samples with low exome coverage) and SC_high (sub comparison of samples with high exome coverage): A) for SC_low and B) for SC_high show the average coverage of all variants, variants called in both datasets, discordant variants and variants found only in exome or genome. C) for SC_low and D) for SC_high show the average coverage of different variant types (reference homozygous (REF), heterozygous (HET) and variant homozygous (HOM)) separately for all variants, variants called in both (CB) datasets, WES_CB and WGS_CB, and variants called in exome only or genome only. All error bars denote standard deviation.

60%. However, WGS provided more homogenous and stable coverage than WES across the entire GC spectrum, validating previous findings that WGS provides superior coverage to WES, especially for GC-rich regions [13,18].

Discussion

In this study, we compared features of variants called in WES- and WGS-based analyses of matched DNA samples at various coverage depths to evaluate differences in the technologies' variant calling quality. In order to conduct a robust comparison, we leveraged this large cohort of matched patient samples (compared to previous comparison studies) to generate WGS and WES at high and low coverage. Furthermore, by using matched biological samples for WGS and WES library generation, sequencing and data analysis the introduction of biological variation otherwise affecting the comparison of WES and WGS data should be reduced.

WGS provides information on larger proportions of genomes than WES, so the analysis was limited to the regions targeted by the exome capture probes. In these regions, WGS identified more variants than WES at both low and high exome coverage (SC_low and SC_high, respectively). This may be because WGS can evenly cover entire target regions, while WES covers parts of the target regions poorly (or not at all), due to inefficiencies and/or biases of the target capture probes [2,19]. Importantly, overall more reference (REF) variants were identified in the WGS datasets, through use of GATK module HaplotypeCaller, which compiles information on all positions with variations in cohorts of samples, even those present in just one sample, in genomic variant call format (gVCF) files. The higher frequency of detectable REF variants in WGS datasets reflects the greater power of WGS to detect rare variants. However, the discrepancy in numbers of variants called in WGS and WES datasets could be diminished by

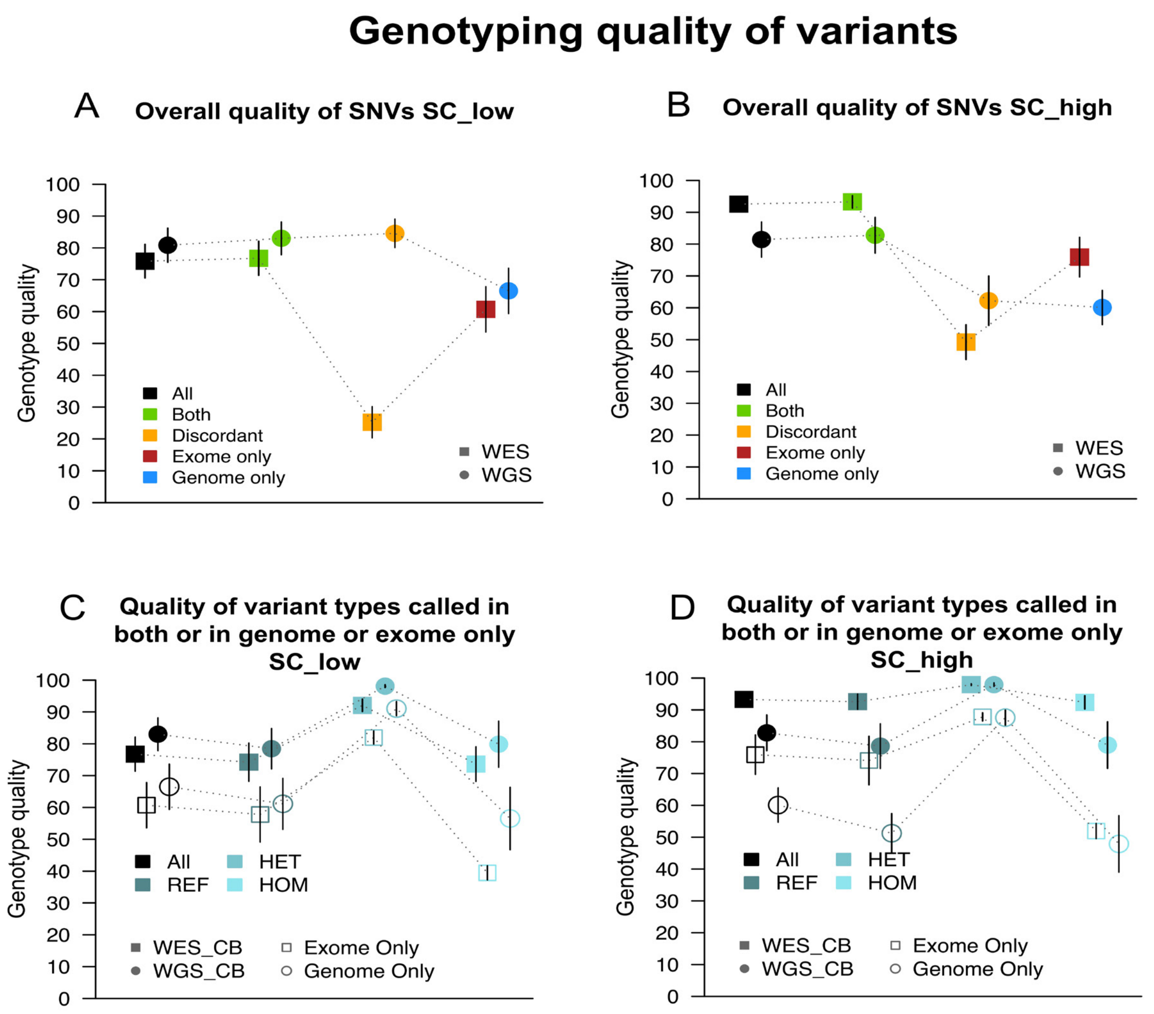

Figure 4: Average genotype quality of called variants in WGS (whole genome sequencing) and WES (whole exome sequencing) for SC_low (sub comparison of samples with low exome coverage) and SC_high (sub comparison of samples with high exome coverage): A) for SC_low and B) for SC_high show the average quality of all variants, variants called in both datasets, discordant variants and variants found only in exome or genome. C) for SC_low and D) for SC_high show the average quality of different variant types (reference homozygous (REF), heterozygous (HET) and variant homozygous (HOM)) separately for all variants, variants called in both (CB) datasets, WES_CB and WGS_CB, and variants called in exome only or genome only. All error bars denote standard deviation.

increasing the coverage depth of the exome sequencing.

This comparison of called variants clearly indicates that WGS datasets are more stable, with less variability between samples, and have higher quality than corresponding WES datasets. Increasing WES coverage increased both the number and quality of called variants, but the higher variability between samples persisted, as shown by the relatively high standard deviation (± 11.97) of exome coverage in the high coverage subgroup SC_high. Some of the discordances with relatively low WES quality may be introduced by PCR during the exome library preparation [1]. Furthermore, the higher frequencies and quality of HET calls provided by WGS indicate that it is less prone to allelic dropouts compared to WES.

Previous authors have recommended application of additional filtering when analyzing WES data [2], with at least 8X coverage and a threshold genotype quality value of 20 for calling WES variants to reduce frequencies of false positives. However, the quality cutoff of 20 might be too lenient for high coverage exome sequencing, as we observed discordant variant calls with higher average genotyping quality (Figure 5D). Therefore, we suggest using a quality cutoff of 40 in new high coverage exome sequencing projects.

The high coverage needed to call HOM and HET variants in WES data robustly, and the generally high genotyping quality of HET variants, indicates bias in the variant calling software. We suggest that different settings of GATK, and other variant callers, should be applied when analyzing WES and WGS data.

The usage of next-generation sequencing technologies, such as WES and WGS, in clinics for improved patient diagnosis and care is emerging; however, it is a future with opportunities as well as challenges. WES-based genetic testing for diagnosis of rare Mendelian disorders has been shown to identify genetic defects in 25% of patients, which is better than

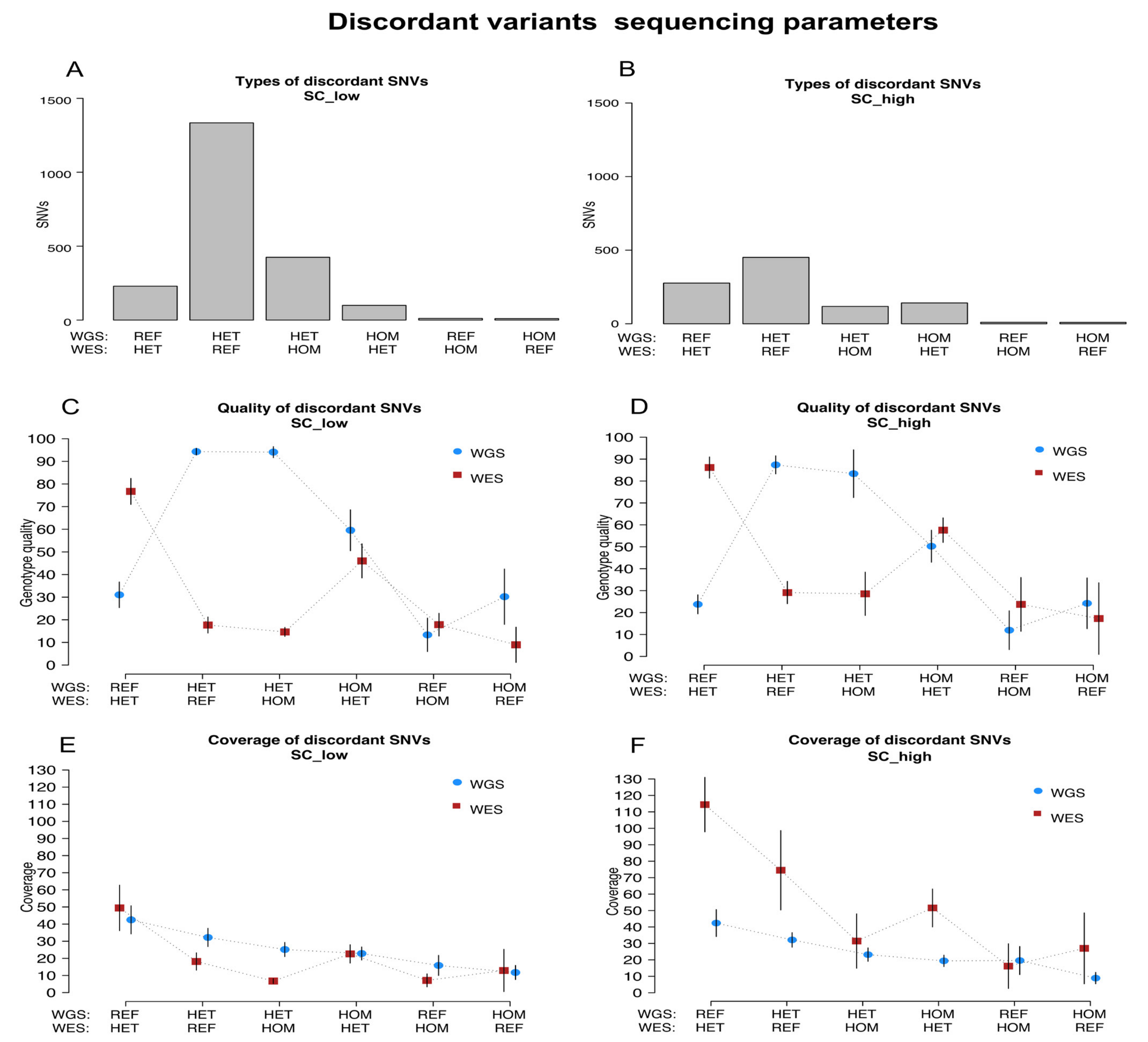

Figure 5: Discordant variant calls by WGS (whole genome sequencing) and WES (whole exome sequencing) for SC_low (sub comparison of samples with low exome coverage) and SC_high (sub comparison of samples with high exome coverage). The x-axis of all plots explains the different types of discordances: A) for SC_low and B) for SC_high shows the average number of the different types of discordances between WGS and WES. C) for SC_low and D) for SC_high shows the average quality for the different types of discordances for WGS and WGS. E) for SC_low and F) for SC_high shows the average coverage for the different types of discordances for WGS and WGS. All error bars denote standard deviation.

conventional diagnostic method [20]. Furthermore, the WGS platform also provides better diagnostic yields than conventional methods in a recent study of 103 patients with suggested underlying genetic disorders [21]. WGS and partly also WES can be especially important in heterogeneous cohorts where conventional tests are not inclusive, as these newer methods have a more agnostic approach. This suggests that WGS and WES have immense potential for future clinical applications. However, there are still multiple challenges (some beyond the scope of this paper) inhibiting their widespread clinical implementation. Such as accuracy and meaningful interpretation and integration of vast volumes of sequencing data and the lack of comparative data and standardized guidelines for clinical use [5,8,13,21].

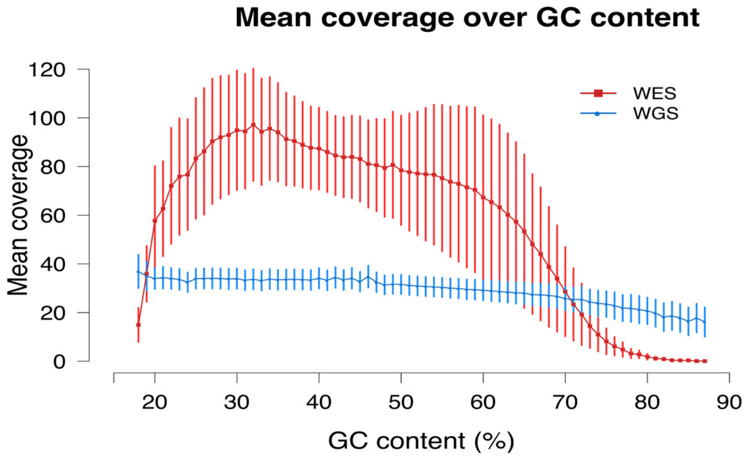

Figure 6: The mean coverage over GC content for WGS (whole genome sequencing) and WES (whole exome sequencing) samples. Error bars denote standard deviation.

Conclusion

To conclude, we found that preparing WGS libraries using the TruSeq DNA PCR-Free Library Preparation kit and sequencing them with an Illumina HiSeq X Ten platform to an average depth of ~30X generally provides higher quality and calls more variants than preparation of WES samples with a Nextera Rapid Capture Exome kit and sequencing them to an average depth of 20-80X using an Illumina HiSeq 2500 platform. However, WES performance can be improved, while maintaining some of its cost advantage, by increasing coverage depth. We believe that as sequencing costs further decline next-generation sequencing technologies clinical implementations will advance and WGS will become the method of choice, even for questions confined to the exome.

Acknowledgements

This work was financially supported by grants from the Swedish Cancer Society, the Swedish Research Council, Linköping University, ALF grants Region Östergötland, the Funds of Radiumhemmet and Marcus Borgströms stiftelse. We gratefully acknowledge the provision of massive parallel sequencing facilities, computational infrastructure, and support by the Science for Life Laboratory (SciLifeLab-Stockholm, Sweden), National Genomics Infrastructure (NGI-Sweden), Uppmax (Sweden) and the National Bioinformatics Infrastructure Sweden (NBIS-Sweden).

Conflicts of Interest

The authors have no interests that could be construed as conflicts of interest.

References

1. Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. Biotechniques 52: 87-94.
2. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C (2015) Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. Hum Mutat 36: 815-822.
3. Chilamakuri CS, Lorenz S, Madoui MA, Vodák D, Sun J, et al. (2014) Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15: 449.
4. Shigemizu D, Momozawa Y, Abe T, Morizono T, Boroevich KA, et al. (2015) Performance comparison of four commercial human whole-exome capture platforms. Sci Rep 5: 127-142.
5. Ormond KE, Wheeler MT, Hudgins L, Klein TE, Butte AJ, et al. (2010) Challenges in the clinical application of whole-genome sequencing. Lancet 375: 1749-1751.
6. Walter K, Min JL, Huang J, Crooks L (2015) The UK10K project identifies rare variants in health and disease. Nature 526: 82-90.
7. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30:418-426.
8. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C (2015) Big data: Astronomical or genomical. PLoS Biol 13: e1002195.
9. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, et al. (2009) Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes. Nature Methods 6: 291-295.
10. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, et al. (2016) The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biol 17: 53.
11. Meynert AM, Ansari M, Fitz Patrick DR, Taylor MS (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. BMC bioinformatics 15: 247.
12. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, et al. (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. PNAS 112: 5473-5478.
13. Meienberg J, Bruggmann R, Oexle K, Matyas G (2016) Clinical sequencing: Is WGS the better WES? Hum Genet 135: 359-362.
14. H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303
16. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491.
17. R-Core Team (2013) A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
18. Meienberg J, Zerjavic K, Keller I, Okoniewski M, Patrignani A, et al. (2015) New insights into the performance of human whole-exome capture platforms. Nucleic acids research 43: 76.
19. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, et al. (2011) Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29: 908.
20. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl 369: 1502-1511.
21. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, et al. (2018) Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med 20: 435-443.