

Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing

Ahmed Alfares, MD, FACMGG¹, Taghrid Aloraini, MS², Lamia Al subaie, MS³, Abdulelah Alissa, MS², Ahmed Al Qudsi, MS², Ahmed Alahmad, MS², Fuad Al Mutairi, MD^{3,4}, Abdulrahman Alswaid, MD^{3,4}, Ali Alothaim, MD^{2,4}, Wafaa Eyaid, MD^{3,4}, Mohammed Albalwi, PhD^{2,4}, Saeed Alturki, MD, PhD^{2,4} and Majid Alfadhel, MD^{3,4}

Purpose: Whole-exome sequencing (WES) and whole-genome sequencing (WGS) are used to diagnose genetic and inherited disorders. However, few studies comparing the detection rates of WES and WGS in clinical settings have been performed.

Methods: Variant call format files were generated and raw data analysis was performed in cases in which the final molecular results showed discrepancies. We classified the possible explanations for the discrepancies into three categories: the time interval between the two tests, the technical limitations of WES, and the impact of the sequencing system type.

Results: This cohort comprised 108 patients with negative array comparative genomic hybridization and negative or inconclusive WES results before WGS was performed. Ten (9%) patients had positive WGS results. However, after reanalysis the WGS

hit rate decreased to 7% (7 cases). In four cases the variants were identified by WES but missed for different reasons. Only 3 cases (3%) were positive by WGS but completely unidentified by WES.

Conclusion: In this study, we showed that 30% of the positive cases identified by WGS could be identified by reanalyzing the WES raw data, and WGS achieved an only 7% higher detection rate. Therefore, until the cost of WGS approximates that of WES, reanalyzing WES raw data is recommended before performing WGS.

Genet Med advance online publication 22 March 2018

Key Words: consanguinity; detection rate; reanalysis of raw data; whole-exome sequencing; whole-genome sequencing

INTRODUCTION

Whole-exome sequencing (WES) and whole-genome sequencing (WGS) methods are increasingly being applied in clinical medicine for the diagnosis of genetic and inherited disorders. WGS covers up to 98% of the whole human genome, and WES covers nearly 95% of the coding regions but only 1–2% of the genome. Compared with WGS, WES has a lower cost per sample, a greater depth of coverage in target regions, and fewer storage requirements, and the data analysis is easier to perform.^{1,2} Depending on the sequencing platform and implemented pipeline, WES generally requires a minimum coverage of 20–40×, whereas WGS only requires a mean of 14 reads to achieve a 95% on-target single-nucleotide polymorphism detection sensitivity.^{3,4} The reported diagnostic yield of WES generally ranges from 25 to 35% (refs. 5–7), and in consanguineous populations, this diagnostic yield can reach 49% (refs. 8,9). However, WGS is considered a more powerful tool than WES for detecting potential disease-causing variations and could be used as a single test to capture nearly all known genetic variations, including

single-nucleotide variants, small insertions/deletions (indels), and copy-number variants,¹⁰ even within the regions of the human genome covered by WES.^{11,12} Previous studies reported the diagnostic yield of WGS is 21–34% in individuals with a broad spectrum of disorders or congenital malformations and neurodevelopmental disorders.^{13,14} A more recent study,¹⁵ published mid-2017, reported diagnostic yield of WGS up to 41%, which is nearly 26% higher than that of WES, and that the yield is up to 57–73% in critically ill infants and in patients with severe intellectual disability.^{16,17} However, few studies comparing WES and WGS in clinical settings using a retrospective data analysis have been performed. In the present study, we performed a comparative retrospective analysis of 108 genetics patients who underwent both clinical WES and clinical WGS.

MATERIALS AND METHODS

This study was approved by the Institutional Research Board of King Abdullah International Medical Research Center. The data collection and analysis were conducted retrospectively at

¹Department of Pediatrics, College of Medicine, Qassim University, Qassim, Saudi Arabia; ²Department of Pathology and Laboratory Medicine, King Abdulaziz Medical City, Riyadh, Saudi Arabia; ³Division of Genetics, Department of Pediatrics, King Abdulaziz Medical City, Riyadh, Saudi Arabia; ⁴King Abdullah International Medical Research Centre, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia. Correspondence: Ahmed Alfares (fars@qu.edu.sa)

Submitted 28 September 2017; accepted 5 January 2018; advance online publication 22 March 2018. doi:10.1038/gim.2018.41

King Abdulaziz Medical City, Riyadh, Saudi Arabia, by reviewing the files of all patients with genetic diseases who had been followed up at the genetics clinic. All cases that underwent both WES and WGS between 2013 and 2017 were enrolled irrespective of their phenotype. DNA sequencing

using both WES and WGS was performed at commercial College of American Pathologists/CLIA-accredited laboratories. WES was performed using one of two different systems, i.e., the Illumina NextSeq, Illumina HiSeq (Illumina Inc, CA, USA) or Ion Proton system (Thermo Fisher Scientific, MA, USA), and

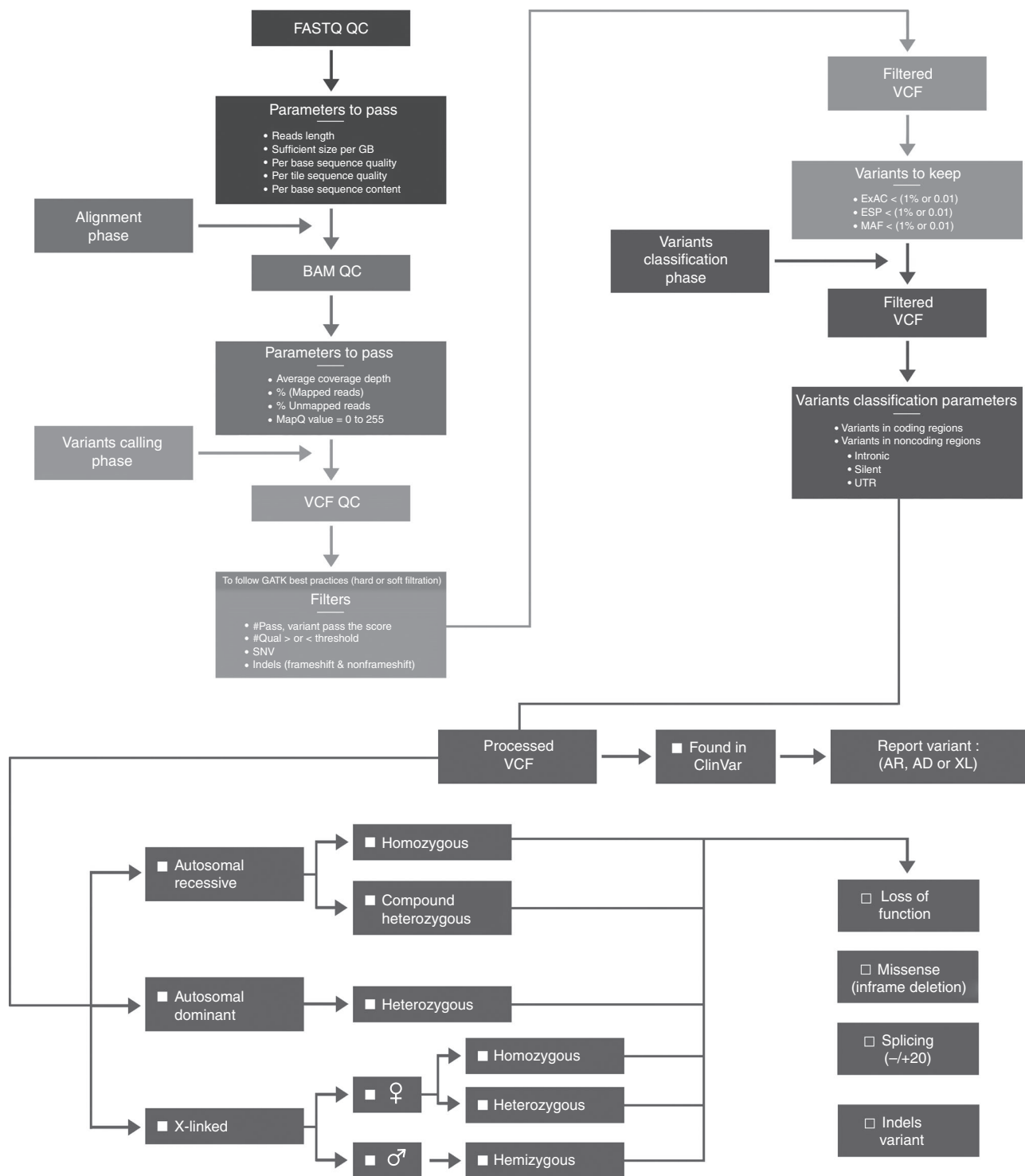


Figure 1 Illustration of the automated bioinformatics pipeline workflow to generate processed vcf files followed by manual filtration. AD, autosomal dominant; AR autosomal recessive; ESP, Exome Sequencing Project; ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; QC, quality control; SNV, single-nucleotide variant; UTR, untranslated region; VCF, variant call format; XL, X-linked.

WGS was performed using the HiSeq 4000. The average coverage depth for WES cases was $\sim 95\times$ with minimum coverage of $20\times$ for any considered variant, and average coverage depth for WGS cases was $\sim 30\times$. All patient information was collected from the electronic system, and we excluded patients with only WGS or WES, patients with limited or no clinical information, and patients with limited or no raw data. Variant call format (vcf) files were generated and the raw data analysis was performed in cases in which the final molecular results showed discrepancies, with positive results in WGS and either negative or inconclusive results in WES; a complete illustration of the pipeline and filtration processes is shown in **Figure 1**. Customization of each step is based on the sequencing systems, bioinformatics pipeline, and the type of the capturing kits. We classified the possible explanations for the discrepancies into three categories. First, during the time interval between the two tests, as part of further studies of the underlying disorder, additional clinical information or novel genes/variants could have been discovered. Second, deep intronic or large deletion variants may not have been detected due to the technical limitations of WES. Third, the type of sequencing system may have had an impact on the final results.

We examined the vcf files from both the WES and the WGS studies that were generated for the same case at the time of collection of the final molecular results. The identified phenotype-related variants were crosschecked using the same genome map and genomic coordinates. In addition to performing several laboratory checkpoints and assessing quality-control measures that were already implemented using different identifiers, such as medical records number, sample ID, laboratory accession number, and date of birth, multiple single-nucleotide polymorphisms were crosschecked between the WES and the WGS vcf files to confirm the identity of the individuals and to verify that both vcf files belonged to the same patient. For each variant, we confirmed whether it was identified in the vcf file and verified the read depth and coverage of the reference and alternate alleles in both the WES and the WGS studies (**Table 1**). Finally, we further reanalyzed each variant identified by WGS for clinical significance and classified the variants as pathogenic/likely pathogenic, a variant of uncertain significance, or benign according to the American College of Medical Genetics and Genomics criteria.¹⁸ Detailed clinical information in Human Phenotype Ontology format and variant classification of all positive cases are provided in **Supplementary Material 2 online**, and detailed family pedigrees for all positive cases with segregation analysis are provided in the **Supplementary Material** as well. All identified disease-causing variants were confirmed by either Sanger sequencing or another related method (i.e., fragment analysis, quantitative polymerase chain reaction, and fluorescence in situ hybridization studies). Several tools were used for the raw data analysis, including Alamut Visual (<http://www.interactive-biosoftware.com/alamut-visual/>), VarSeq from GoldenHelix (<http://www.goldenhelix.com/>), the University of California–Santa Cruz Genome

Browser (<https://www.genome.ucsc.edu/>), the Integrative Genomics Viewer (<http://software.broadinstitute.org/software/igv/>), Genome Analysis Toolkit Best Practices (<https://software.broadinstitute.org/gatk/>), SAMtools (<http://www.samtools.sourceforge.net/>), and Freebayes (<https://github.com/ekg/freebayes>).

RESULTS

In total, 154 patients recruited for the study had negative array comparative genomic hybridization and negative or inconclusive WES results. However, 36 patients were excluded because their WGS results were incomplete or because further testing, such as a segregation analysis or clinical examination, was required. The remaining 118 patients had complete clinical and final results available; of these patients, 20 (17%) had positive WGS results, 5 (4%) had inconclusive results, and the 93 (79%) remaining patients had negative results. However, we further excluded 10 positive cases, of which 4 cases had loss-of-function variants and 2 cases had missense variants, because the WES was performed many years ago and thus the raw WES data (vcf or BAM) were not available for retrieval and comparison. In the other 4 cases, the variants were found in both the WES and the WGS raw data but were apparently missed and thus not reported in the WES results following the final interpretation and filtration processes. Therefore, in this study, we only considered the 108 patients with complete clinical information and raw data.

Demographic data

The gender distribution was equal between males and females. The cohort was enriched in pediatric (< 14 years) patients (98 (91%) patients) and included only 10 (9%) adults. Additionally, our cohort was enriched in cases from consanguineous unions (76 (70%) cases compared with 32 (30%) nonconsanguineous cases); of the consanguineous cases, 71% were reported as first-cousin unions (**Table 2**).

WGS hit rate before reanalysis of results

Of the 108 patients, 10 (9%) had positive WGS results and previously negative WES results, 5 (5%) had inconclusive results, and 93 (86%) had negative results (**Figure 2**). Of the 5 cases with inconclusive results, 4 with the same variants were found by both WES and WGS, and in case 5, a deep intronic variant was identified only by WGS.

Effect of time interval between WES and WGS

The average time interval between the WES and the WGS for the 10 positive cases was 5 months (1–16 months, SD 4.9). We examined the indication for the testing during the interval between the two tests and whether any new clinical data resulted in significant changes in the classification, and we found that none of the 10 positive cases had new clinically significant data, although certain cases underwent additional investigations, such as brain imaging or tissue analysis. Nevertheless, for cases 1, 2, and 3 (**Supplemental Cases 1–3 online**), the variants were identified using both WES and

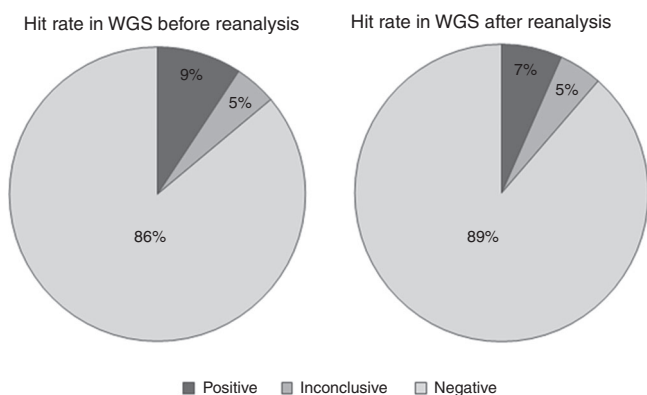
Table 1 List of 10 positive cases detected by WGS but not WES and possible explanations

Case	Detected by WGS only	Diagnosis	OMIM entry	Gene	Time interval	CNV	Intronic	g.DNA (GRCh37), transcript ID, c.DNA, amino acid	Coverage in WES (Ref, Alt)	Coverage in WGS (Ref, Alt)
1	No	Spinocerebellar ataxia type 42	616795	CACNA1G	Yes	—	—	Chr17:g.48649284T>C, NM_018896.4, c.632T>C, p.Leu211Pro	43,38	21,17
2	No	Early infantile epileptic encephalopathy type 26	616056	KCNB1	Yes	—	—	Chr20:g.47990875G>A NM_004975.2, c.1222C>T p.Pro408Ser	28,25	16,25
3	No	Neurodegeneration accompanied by iron deposition in the brain	607434	GTPBP2	Yes	—	—	Chr6:g.43589844_43589845del NM_019096.4, c.1527_1528del p.Glu509Aspfs*78	0,57	0,40
4	Yes	Congenital central hypoventilation syndrome	209880	PHOX2B	—	Yes	—	Chr4:g.41748048GCG[25] NM_003924.3c.721GCN(25) repeat expansion [+25]	NA	8,18
5	Yes	Nemaline myopathy 1	609284	TPM3	—	Yes	—	Chr1:g.154142455-154147729del NM_152263.2 large deletion	NA	NA
6	Yes	Tuberous sclerosis type 2	613254	TSC2	—	—	Yes	Chr16:g.2107460C>T NM_000548.3c.848+281C>T	NA	16,20
7	Yes	Mental retardation type 36	615286	ADAT3	Detected in subsequent runs	Detected in subsequent runs	WES	Chr19:g.1912476G>A NM_138422.2c.430G>A, p.Val144Met	Not detected	0,23
8	Yes	Mental retardation type 36	615286	ADAT3	Detected in subsequent runs	Detected in subsequent runs	WES	Chr19:g.1912476G>A NM_138422.2c.430G>A, p.Val144Met	Not detected	0,18
9	Yes	Mental retardation type 36	615286	ADAT3	Detected in subsequent runs	Detected in subsequent runs	WES	Chr19:g.1912476G>A NM_138422.3c.430G>A, p.Val144Met	Not detected	0,19
10	Yes	Congenital disorder of glycosylation type IIIm	300896	SLC35A2	Detected in subsequent runs	Detected in subsequent runs	WES	ChrX:g.48767103C>G NM_001282651.1c.346G>C, p.Ala116Pro	Not detected	12,6

c.DNA, coding DNA; CNV, copy-number variant; g.DNA, genomic DNA; NA, not available; WES, whole-exome sequencing; WGS, whole-genome sequencing.

Table 2 Demographic and diagnostic information for the cohort based on the whole-genome sequencing results

	Positive	Inconclusive	Negative	Total
Gender				
Male (56%)	4 (7%)	4 (7%)	53 (87%)	61
Female (44%)	6 (13%)	1 (2%)	40 (85%)	47
Age				
Adult (9%)	0	0	10 (100%)	10
Pediatrics (91%)	10 (10%)	5 (5%)	83 (85%)	98
Consanguinity				
Yes (70%)	6 (8%)	4 (5%)	66 (87%)	76
No (30%)	4 (13%)	1 (3%)	27 (84%)	32

**Figure 2** Pie chart showing the whole-genome sequencing (WGS) hit rate before and after reanalysis of the clinical information and raw data.

WGS; however, the variants were not reported at the time of WES because the disease-causing genes were only described after the WES was performed (**Table 1** and **Supplementary Material 2 online**).

Effects of structural rearrangements and noncoding variants

Cases 4, 5, and 6 (**Supplemental Cases 4–6 online**) were detected by WGS and not detected by WES. In particular, case 4 presented with idiopathic congenital central hypoventilation syndrome, a disorder caused by polyalanine expansion in *PHOX2B*,¹⁹ with typical presentation. Sanger sequencing of *PHOX2B* was ordered first and failed to identify the expansion, similar to WES; however, WGS showed possible expansion in exon 3, which was confirmed by fragment analysis (data not shown). Meanwhile, case 5 included a large deletion spanning different exons in the *TPM3* gene (exon 3 to exon 9). In case 6, the variant was a deep intronic single-nucleotide variant in the *TSC2* gene (**Table 1**) (**Supplementary Material 1 online, Figure 1**).

Effect of sequencing system

Cases 7, 8, and 9 (**Supplemental Cases 7–9 online**) involved a frequent variant in the *ADAT3* gene, and case 10

(**Supplemental Case 10**) involved a variant in *SLC35A2*. For these four cases, WES was performed using the Ion Proton system, and the variants were missed and thus undetected (**Table 1**). Unfortunately, no BAM files were available from the Ion Proton system for visual inspection or confirmation of the coverage of these two genes, and the outsourced sequencing laboratory does not keep FASTQ or BAM files for longer than 6 months. However, in the subsequent WES analyses and tests using the Illumina system, we were able to detect the *ADAT3* variant in the heterozygous state in four individuals and to identify one positive case with the same variant in *SLC35A2*. Additionally, in new analyses, we confirmed that there was enough coverage of both genes (~30× for *SLC35A2* and ~90× for *ADAT3*) (**Supplementary Material 1 online, Figure 2**).

Impact of consanguinity

Autosomal recessive (AR) and homozygous variants represented 50% of the positive cases, and autosomal dominant (AD) and X-linked cases represented 50% of the positive cases. Of the seven positive cases detected only by WGS, four had a documented history of consanguinity, three (75%) had an AR disorder with homozygous variants, and one (25%) had an AD disorder with a heterozygous variant. Among the remaining three positive cases with no documented history of consanguinity, one had an AR disorder with a homozygous variant, one had an AD disorder with a heterozygous variant, and one (a male patient) had an X-linked dominant disorder with a hemizygous variant. Of these seven cases, family history was significant in three (43%) and unremarkable in four (57%) (**Supplementary Material online, Tables 1 and 2**).

WGS hit rate after reanalysis of raw data

Of the 10 cases identified as positive by WGS, only 7 (cases 4–10) were detected by WGS and not detected by WES; hence, compared with WES, the WGS hit rate decreased to only 7% (7/105) (**Figure 2**).

DISCUSSION

Whereas previous reports have shown that the WGS hit rate was nearly 26% over WES, in our study, without considering any reanalysis of the raw data or the clinical information, the hit rate was 9%. However, after reanalyzing and comparing the raw data from both the WES and WGS, the hit rate for the positive cases detected only by WGS decreased to 7%, which illustrates the importance of reanalyzing WES results before performing further testing. Furthermore, when only considering the 7 cases (cases 4–10) detected by WGS but not by WES, in three cases, the variant in *ADAT3* was not identified; however, in subsequent analyses, the *ADAT3* and *SLC35A2* genes were well covered, and we were able to identify heterozygous variants in other carriers and one additional positive case of *SLC35A2* with the same variant. If we exclude these four cases, the WGS hit rate decreases even further, to only 3% (3/101). However, because these variants remained undetected after reanalyzing the vcf files, WGS might have

been required. Therefore, we considered these cases to belong to the group of cases detected by WGS.

In this study, there are several different explanations for why variants were not reported in WES final results: the time interval between the WES and the WGS (three variants), the limitations of WES coverage or technology (i.e., for copy-number and deep intronic variants) (three variants), and the limitations of the sequencing systems (four variants). Of the 10 positive cases, 3 could be detected by reanalyzing the raw data and vcf files from previous WES analyses before performing WGS, particularly via reassessment of newly discovered genes or reported variants, which accounted for 30% of positive cases in our study that were not detected by WES but were subsequently detected by WGS. Considering an average time interval of only 5 months between WES and WGS, the utility of WES reanalysis could be even greater in the context of a longer time interval between the two tests. Therefore, until the cost of WGS approximates that of WES, this approach could help to increase the detection of variants by WES.

Previously, we showed that most positive cases in clinical whole-exome studies in Saudi Arabia were in consanguineous populations with AR disorders.⁸ Although the sample size in this study was small, i.e., 108 patients and 10 positive cases, no difference in WGS hit rate was observed between the consanguineous and the nonconsanguineous cases. Additionally, the modes of inheritance in the positive cases were equally distributed between AR and non-AR disorders, likely because most of the positive cases were detected by WES, so the utility of WGS might be limited in our population. However, further studies with larger sample sizes are required to confirm our observation.

The major limitation of WGS is cost. In this study, all clinical testing was performed as a part of routine clinical patient care; in this context, each WES costs approximately \$1,200, the calculated WES reanalysis costs approximately \$250 (accessible raw data storage and retrieval, reanalysis time with 5 hours on average for each case), and each WGS costs approximately \$4,200. If we include the 4 positive cases in WGS that were missed by WES during interpretation, then for all 112 cases (108 cases and 4 cases missed during interpretation), approximately \$604,800 was spent to perform both WES and WGS. Half of the positive (7/14) cases detected by WGS could have been identified by reanalyzing the raw WES data from all 112 cases, which would cost around \$28,000, thus saving a substantial amount of money. The likely explanation is that the gene was reported and discovered either during the interval between the two tests or secondary to the interpretation and filtration processes during the final steps before reporting, rather than being related to the technical advantages of WGS. However, we spent an additional \$529,200 from the health-care system budget to achieve a 7% higher detection rate.

In conclusion, although WGS is a more powerful tool than WES, in this study, we showed that WGS has additional, but limited, clinical utility compared with reanalyzing WES

data, and until the cost of WGS approximates that of WES, reanalyzing WES raw data is recommended before performing WGS.

SUPPLEMENTARY MATERIAL

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

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

- Saunders CJ, Miller NA, Soden SE, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med*. 2012;4:154ra135.
- Berg JS, Khoury MJ & Evans JP. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genet Med* 2011;13:499–504.
- 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: 815–22.
- Meynert AM, Ansari M, FitzPatrick DR & Taylor MS. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* 2014;15:247.
- Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N Engl J Med*. 2013;369:1502–11.
- Lee H, Deignan JL, Dorrani N, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 2014;312:1880–7.
- Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 2014;312:1870–9.
- Alfares A, Alfadhel M, Wani T, et al. A multicenter clinical exome study in unselected cohorts from a consanguineous population of Saudi Arabia demonstrated a high diagnostic yield. *Mol Genet Metab*. 2017;121: 91–5.
- Monies D, Abouelhoda M, AlSayed M, et al. The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes. *Hum Genet* 2017;136:921–39.
- Fang H, Wu Y, Narzisi G, et al. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med*. 2014;6:89.
- Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A*. 2015;112:5473–8.
- Meinenberg J, Bruggmann R, Oexle K & Matyas G. Clinical sequencing: is WGS the better WES? *Hum Genet* 2016;135:359–62.
- Stavropoulos DJ, Merico D, Jobling R, et al. Whole genome sequencing expands diagnostic utility and improves clinical management in pediatric medicine. *NPJ Genom Med* 2016;1; 15012; doi: 0.1038/npjgenmed. 2015.12.
- Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet* 2015;47:717–26.
- 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*; advance online publication, 3 August 2017; doi: 10.1038/gjm.2017.119.
- Gilissen C, Hehir-Kwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* 2014;511:344–7.
- 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.
- Li MM, Datto M, Duncavage EJ, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn*. 2017;19:4–23.
- Weese-Mayer DE, Berry-Kravis EM, Zhou L, et al. Idiopathic congenital central hypoventilation syndrome: analysis of genes pertinent to early autonomic nervous system embryologic development and identification of mutations in PHOX2b. *Am J Med Genet A*. 2003;123A:267–78.