Chapter

Clinical Validation of a Whole Exome Sequencing Pipeline

Debra O. Prosser, Indu Raja, Kelly Kolkiewicz, Antonio Milano and Donald Roy Love

Abstract

Establishing whole exome sequencing (WES) in an accredited clinical diagnostic space is challenging. The validation (as opposed to verification) of an approach that will lead to clinical reports requires adhering to international guidelines and recommendations and developing a robust analytical pipeline that can scale due to the increasing clinical demand for comprehensive gene screening. This chapter will present a step-wise approach to WES validation that any laboratory can follow. The focus will be on highlighting the pivotal technical issues that must be addressed in validating WES and the analytical tools and QC metrics that must be considered before implementing WES in a clinical environment.

Keywords: whole exome sequencing, next-generation sequencing, validation, bioinformatics, diagnostics

1. Introduction

The decision as to which type of genetic test should be implemented by a clinical laboratory is largely driven by the type of referrals received by the laboratory and the complexity of patients' clinical phenotypes. In the main, testing has advanced from single-gene to multi-gene panels in which next-generation sequencing (NGS) has offered the technical means of undertaking this approach at low cost and high throughput. However, with the increasing awareness of genetic heterogeneity combined with gene discovery, whole exome sequencing (WES) offers laboratories a more streamlined approach. By implementing a single wet-work pipeline of exome capture coupled with the ability to analyze a virtual gene panel or report on the whole exome, laboratories can perform NGS in a more efficient manner.

Since the inception of NGS over a decade ago, multiple recommendations and guidelines have been published for NGS [1–3]. Using these guidelines, the College of American Pathologists (CAP) and Association for Molecular Pathology (AMP) published their Practical Framework for Designing and Implementing NGS Tests for Inherited Disorders in 2019 [4], and this is available through the CAP website (https://www.cap.org/member-resources/precision-medicine/next-generation-sequencing-ngs-worksheets).

We adopted this framework to establish a diagnostic NGS service using whole exome sequencing as our capture procedure and analyzing virtual gene panels or WES for reporting purposes.

The framework provides guidance and editable worksheets for the five steps involved in test establishment and validation.

1. Test design: setup

2. Assay design and optimization

3. Test validation

4. Quality management

5. Bioinformatics and IT

Throughout the validation process, it is essential that the NGS workflow is informed by the real-world local environment in which clinical testing will be performed.

2. Test design: setup

In view of the diverse range of referrals made to the authors' genetics laboratory (serving the needs of a 400-bed women and children's hospital in the Middle East), a whole exome capture solution was chosen for library preparation. The principal motivation behind this determination was to achieve an efficient workflow that would allow appropriate batching coupled with a time-limited turnaround time (TAT) for all referrals.

The limited number of staff in the authors' laboratory demanded a WES workflow that could be easily automated, twinned with a data analysis package that would allow secure remote access with a strong databasing function. The whole exome solution capture by SOPHiA™ Genetics was chosen for library preparation. This platform allows for the analysis of WES, clinical exome sequencing (CES) and clinical gene panels, together with the identification of single-nucleotide variants (SNVs) and copy number variants (CNVs) using SOPHiA™ DDM software.

3. Assay design and optimization

The validation pipeline needs to be grounded from the beginning in terms of the requirements of the test, which must take into account the sample types the laboratory will receive and the parameters that need to be satisfied (see **Table 1**).

Routinely, whole blood samples collected in EDTA are received by the authors' laboratory for testing. Therefore, our validation focused only on genomic DNA extracted from whole blood using our standard methods. The baseline validation of the WES data required the inclusion of two HapMap gDNA samples: the NIST control (NA12878) and the commercial control (SG063) supplied by SOPHiA™ Genetics.

The WES capture by SOPHiA™ Genetics was used for library preparation following all the steps as set out by the automated WES 32 reaction protocol. For instrumentation, our validation was restricted to automated library preparation using the PE Sciclone® G3 NGS workstation and sequencing using the Illumina® HiSeq4000 platform.

A critical additional consideration was the need for copy number variant calls to be made. This required a minimum batch number of eight patients and high coverage requirements, which involved restricting the number of samples per Illumina® HiSeq4000 lane to one pool of eight patients.

Test requirements	Must have	Nice to have
WES		Y
CES	Y	
Clinical panels	Y	
CNV detection	Y	
Necessary sample throughput per month	16	32
How deeply does each position need to be covered for accurate variant calling (if known—otherwise address during test optimization)	>20x	>50x
DNA from whole blood collected in EDTA	Y	
DNA from external/commercial sources (limitations)	Y	
Required/expected TAT	3 months	2 months
Combine different tests (existing or planned) within a sequencing run	Y	

WES, whole exome sequencing; CES, clinical exome sequencing; CNV, copy number variant; TAT, turnaround time.

Table 1.
Test requirements and limitations.

Importantly, the naming of the sequence files (.bam,. FASTQ, etc.) should be considered during the early phase of test design and validation. File conventions that are used for the bioinformatic process may be limited in terms of the type of special characters and/or character length. Following recommendations in the CAP/AMP-Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines [5], the identity of the sample must be preserved throughout all steps of the bioinformatic pipeline. These authors recommend the following four unique identifiers that should be applied to the sample file name:

i. Unique sample identifier

ii. Unique patient identifier

iii. Unique run identifier

iv. Laboratory location identifier

It is essential that the file naming convention that is decided upon for validation adheres to the above recommendations and can be universally implemented for all subsequent testing.

4. Test validation

Test validation mandates a need for accuracy, precision and stability. These assessments must be made in the context of expected clinical workloads and performance. For the authors' laboratory, the sample batch size was set at 16 samples per validation batch and a total of three validation runs performed over differing days with differing technologists.

Analytical performance was characterized by the assessment of precision, sensitivity and concordance of variant calls against previously validated data.

Inter-run and intra-run data were achieved by replicate analysis of two HapMap gDNAs, the NIST sample, NA12878, and the commercial control supplied by SOPHiA™ Genetics, SG063, as well as four well-characterized clinical samples previously reported by accredited laboratories. The remaining samples included a representative group of the clinical samples received by the authors' laboratory (see **Table 2**).

The complete NGS workflow should be included in the validation, from library preparation to bioinformatic analysis to report generation, which is highlighted below.

- Sample collection and DNA extraction. Genomic DNA is extracted and purified from blood samples using either the Gentra® PureGene® DNA Blood Mini Kit or the QIAsymphony® DSP DNA Midi kit (QIAGEN, Hilden, Germany). DNA quality is initially assessed by NanoDrop™ spectrophotometry.

  Genomic DNA preparation. The initial preparation of gDNA used in NGS library preparation is the most critical step in the NGS workflow, and the care and time taken here are key to successful library amplification and sequencing.

  High-quality gDNA can be by quantified using a Qubit™ fluorometer followed by sequential dilution with further quantification to the desired input concentration. It is essential to minimize pipetting gDNA volumes of less than 5 μl for dilution. In our study, gDNA is prepared to a working concentration of 40 ng/μl. After Qubit™ quantification, the integrity of the gDNA can be analyzed using an Agilent TapeStation 4200. Samples with a DNA integrity number (DIN) of greater than 7.5 can proceed to WES capture.

- Library preparation, targeted capture and sequencing. Whole exome sequencing was performed according to the SOPHiA™ Whole Exome Solution 32 Samples User Guide, in combination with the SOPHiA™ Library Preparation and Capture User Guide—automation with PerkinElmer Sciclone® G3 NGS workstation. Each validation run consists of 16 samples that are divided into 2 pools of 8 samples each, as shown in the validation grid in **Table 3**.

  The SOPHiA™ WES protocol for library construction subjects genomic DNA (200 ng) to enzymatic fragmentation, end repair and A-tailing. All these steps occur using a Sciclone® G3 NGS workstation. The adapter-ligated DNA is then amplified in a limited way via an eight-cycle PCR protocol.

  Post-amplification cleanup of the libraries is carried out using the Sciclone® G3 NGS workstation, and libraries are prepared for quantitation with a dilution factor of 4.

  Amplified libraries are analyzed using Qubit™ fluorometer and Agilent TapeStation 4200 to assess the quantity and quality of each individual library. Library DNA fragments should have a size distribution between 300 and 700 bp. Genomic DNA that has been fragmented, end repaired, A-tailed and adapter-ligated can then be considered library DNA, which is ready for pooling and then hybridization and capture. In the case of the SOPHiA™ WES protocol, eight samples are pooled (200 ng of each library) per capture.

  Prepared pools are hybridized for 4 h followed by post-capture amplification and cleanup on the Sciclone® G3 NGS workstation.

  Final library quantification is performed for each captured library pool using a Qubit™ fluorometer and Agilent TapeStation 4200. Subsequent pools are

Sample ID	Description	Purpose	Purpose (detail)	Specific variant/s of interest	Variant type	Measured metric
VAL-1	NA12878	Baseline validation	N/A	N/A	N/A	Intra-run variability Inter-run variability
VAL-2	SG063	Baseline validation	N/A	N/A	N/A	Intra-run variability Inter-run variability
VAL-3	Anonymized patient specimen	Baseline validation	Variant type	Ciliopathy gene panel CCDC39:c.2017G > T p.(Glu673*) CCDC39: Deletion of exons 14 to 20	SNV CNV	Inter-run variability Sensitivity
VAL-4	Anonymized patient specimen	Baseline validation	Variant type prevalent in gene	Single-gene analysis CFTR:c.1521_1523delCTT p.(Phe508del)	DEL	Inter-run variability Sensitivity
VAL-5	Anonymized patient specimen	Baseline validation	Variant type	Craniosynostosis gene panel CACNA1H:c.4318_4319delinsGC p.(Phe1440Ala)	DELINS	Inter-run variability Sensitivity
VAL-6	Anonymized patient specimen	Baseline validation	Variant type prevalent in gene	Tuberous sclerosis gene panel TSC2: Deletion of exons 2 to 16	CNV	Inter-run variability Sensitivity
VAL-7	Anonymized patient specimen	Gene-specific validation	Variant type	Arrhythmia cardiomyopathy gene panel SCN5A:c.4867C > T p.(Arg1623*)	SNV (stop)	Sensitivity
VAL-8	Anonymized patient specimen	Gene-specific validation	Variant type	Custom panel of 196 genes 200 genomic co-ordinates	SNV DEL/ DUP	Sensitivity
VAL-9	Anonymized patient specimen	Gene-specific validation	Variant type	Paroxysmal Dystonia gene panel Del 16p11.2 chr16:29,656,684-30,190,568	CNV	Sensitivity
VAL-10	Anonymized patient specimen	Gene-specific validation	Variant type	Leukodystrophy gene panel MLC1:c.908_918delinsGCA p.(Val303Glyfs*96)	DELINS	Sensitivity
VAL-11	Anonymized patient specimen	Gene-specific validation	Variant type	Epilepsy gene panel WWOX: Deletion of exons 1–5	CNV	Sensitivity
VAL-12	Anonymized patient specimen	Gene-specific validation	Variant range	Epilepsy gene panel	SNV DEL/ DUP	Sensitivity
VAL-13	Anonymized patient specimen	Gene-specific validation	Variant type	Single-gene analysis CFTR: deletion of exons 4–8	CNV	Sensitivity

Sample ID	Description	Purpose	Purpose (detail)	Specific variant/s of interest	Variant type	Measured metric
VAL-14	Anonymized patient specimen	Gene-specific validation	Variant range	Neuropathy gene panel	SNV DEL/ DUP	Sensitivity
VAL-15	Anonymized patient specimen	Gene-specific validation	Variant range	Cholestasis gene panel	SNV DEL/ DUP	Sensitivity
VAL-16	Anonymized patient specimen	Gene-specific validation	Variant type	Tuberous sclerosis gene panel (2 genes) TSC2:c.5238_5255del p.(His1746_Arg1751del)	DEL	Sensitivity
VAL-17	Anonymized patient specimen	Chromosomal CNV validation	Variant type	Molecular karyotype referral Dup 22q11.21 chr22:18,661,724-21,809,099	CNV	Sensitivity
VAL-18	Anonymized patient specimen	Gene-specific validation	Variant range	Primary ciliary dyskinesia gene panel DNAH5: Gain of exons 1 to 50 DNAH5:c.5503C > T p.(Gln1835*)	SNV CNV	Sensitivity
VAL-19	Anonymized patient specimen	Gene-specific validation (pseudogene)	Variant range	Inherited cancer gene panel CDKN2A:c.9_32dup p.(Ala4_Pro11dup)	SNV DEL	Sensitivity
VAL-20	Anonymized patient specimen	Gene-specific validation	Variant range	Custom panel of 196 genes 200 genomic coordinates	SNV DEL/ DUP	Blind analysis
VAL-21	Anonymized patient specimen	Chromosomal CNV validation	Variant type	Molecular karyotype referral Duplication at 16p13.11, deletion at 12p31 and duplication at Xp21.1	CNV	Sensitivity
VAL-22	Anonymized patient specimen	Gene-specific validation	Variant type prevalent in gene	Single-gene analysis DMD: duplication exons 45–62	CNV	Sensitivity
VAL-23	Anonymized patient specimen	Gene-specific validation	Variant type prevalent in gene	Dystrophinopathy gene panel DMD: deletion of exons 8–34	CNV	Sensitivity
VAL-24	Anonymized patient specimen	Gene-specific validation	Variant range	Custom panel of 196 genes 200 genomic co-ordinates	SNV DEL/ DUP	Sensitivity
VAL-25	Anonymized patient specimen	Gene-specific validation (pseudogene)	Pseudogene	Custom panel of nine genes	SNV DEL/ DUP	Sensitivity

Sample ID	Description	Purpose	Purpose (detail)	Specific variant/s of interest	Variant type	Measured metric
VAL-26	Anonymized patient specimen	Gene-specific validation	Variant type	Primary Immunodeficiency gene panel TBX1:c.1383_1421del p.(Ala464_Ala476del)	DEL	Sensitivity
VAL-27	Anonymized patient specimen	Gene-specific validation	Variant type	Dilated cardiomyopathy gene panel TTN:c.75984_75985insTACCA p.(Ala25329Tyrfs*32)	INS	Sensitivity
VAL-28	Anonymized patient specimen	Gene-specific validation	Variant type	Pediatric cancer gene panel SMARCB1:c.159_160delinsTATCTGGAGGCG (p.Leu54Ilefs*20)	DELINS	Sensitivity

DEL, deletion; INS, insertion; DUP, duplication; SNV, single-nucleotide variant; CNV, copy number variant.

Table 2.
Sample list.

		Run 001			Run 002			Run 003	
Pool A	**A**	VAL-1 NA12878	VAL-4	**A**	VAL-5	**VAL-13**	**A**	**VAL-21**	VAL-5
	B	**VAL-3**	VAL-10	**B**	VAL-2 SG-063	VAL-15	**B**	VAL-28	VAL-1 NA12878
	C	**VAL-11**	VAL-1 NA12878	**C**	**VAL-17**	VAL-2 SG-063	**C**	VAL-1 NA12878	VAL-24
	D	VAL-2 SG-063	VAL-12	**D**	VAL-16	VAL-19	**D**	**VAL-3**	VAL-25
Pool B	**E**	VAL-1 NA12878	**VAL-6**	**E**	VAL-2 SG-063	VAL-20	**E**	**VAL-22**	**VAL-6**
	F	VAL-7	VAL-8	**F**	**VAL-6**	**VAL-3**	**F**	**VAL-23**	VAL-2 SG-063
	G	VAL-2 SG-063	**VAL-9**	**G**	VAL-14	VAL-1 NA12878	**G**	VAL-4	VAL-27
	H	VAL-5	VAL-2 SG-063	**H**	**VAL-18**	VAL-4	**H**	VAL-1 NA12878	VAL-26

Copy number variant (CNV) samples are indicated in bold.

Table 3.
Validation grid.

diluted to 20 nM (in a total volume of 20 µl) and subjected to sequencing using an Illumina® HiSeq4000 Sequencing platform.

- Sequence analysis: performance metrics. Baseline performance metrics for the WES validation study must involve the analysis of well-characterized reference samples: the NIST sample (NA12878) and the SOPHiA™ Genetics control SG063. The sequence metrics for each sample in the run must be recorded and averages established using the reference samples. Samples must meet the sequencing metrics shown in **Table 4** in order to reach the threshold for clinical reporting.

  Analytical sensitivity and specificity must be calculated separately for each variant type (SNV, indel, CNV, etc.). Additional runs may be required to meet acceptable confidence intervals for less frequent variant types of insertions and deletions. For 95% confidence and 95% reliability, 59 variants of each type (and insertion/deletion range) should be analyzed [5]. The variant types that do not have strong confidence intervals must be listed in the test limitations of the clinical report until such time that the desired confidence levels have been achieved.

Selected sequencing metrics	Must have	Nice to have
Q30 score	>80	>85
Total number of reads per sample	>70 M	80–100 M
Percentage of mapped reads	>80%	>85%
Total percentage on-target reads	>90%	>95%
Coverage 10% quantile (at this depth, 90% target covered)	20x	50x

Table 4.
Sequencing metrics.

5. Quality management

The worksheets described by Santani et al. [4] set out very clear guidance for all quality aspects that need to be taken into consideration for the test to meet CAP requirements [4]. Through a validation study, the majority of a test's limitations will be discovered and can be recorded against the QC parameters. **Table 5** summarizes quality metrics that need to be addressed.

Section	Category	Criteria	Specific requirement Note that these may vary between tests and laboratories
Pre-analytical QC (per sample)	Specimen quality	Wrong specimen type	Whole blood
		Wrong type of tube	Purple top EDTA tube
		Insufficient quantity	≥0.5 ml
		Clotting (blood only)	No visible clots
		Insufficient labelling	Labelling contains name, DOB, barcode, date of collection
		Expired specimen	≤7 days since collection
		Expired collection tube	Collection tube not expired
	DNA quality and quantity	OD 260/280 ratio	>1.7
		Electrophoretic analysis	Shows intact high molecular weight DNA band
		Quantification	≥500 ng
		DNA integrity number (DIN)	>7.5
Analytical QC (per instrument run)	Instrument run QC	Cluster density	Not taken into account
		Base quality	Q30 ≥ 80
	Pipeline QC	Total reads passing filter	>280 M per lane
		% reads not assigned to any sample	<5%
	Control samples	Positive control	Expected variants found
Analytical QC (per sample)	Library preparation	Fragment size and distribution	>80% of fragments between 300 and 700 bp
		Pooled library concentration	>20 nM
	Sample de-multiplexing	% reads assigned to sample	8–12%
	Read alignment	% Reads aligned to target	>90%
		Distribution of coverage	>95% within 25–200×
		Coverage 10% quantile (at this depth 90% target covered at x)	>40×
		PCR duplicates	<20%
	Specimen identity	Accurate specimen identity, file names with 4 points of identification	All worksheets and transfers during bench work are witness checked for accurate specimen identification
	Data transfer Integrity	Data transfer to secure analysis platform	

Table 5.
Quality management.

	Considerations	Resources	Links
Gene selection	Clinical association	ClinGen	https://clinicalgenome.org/curation-activities/gene-disease-validity/ https://search.clinicalgenome.org/kb/gene-validity
		GeneReviews	https://www.ncbi.nlm.nih.gov/books/NBK1116/
Gene analysis	Appropriate transcripts	LRG RefSeq	https://www.lrg-sequence.org/ https://www.ncbi.nlm.nih.gov/refseq/rsg/
	Pseudogenes	Pseudogene PanelApp – Genes and Entities	http://pseudogene.org/ https://panelapp.genomicsengland.co.uk/panels/entities/?tag=locus-type-pseudogene
	Evaluated homopolymeric regions	Ivády et al. [6]	DOI: 10.1186/s12864-018-4544-x
	Mutation spectrum—reported deep intronic and/or promoter region variants	PanelApp—Genes and Entities	https://panelapp.genomicsengland.co.uk/panels/entities/?
	CNV analysis	ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/
		Decipher	https://decipher.sanger.ac.uk/
	Establish if critical variants are not covered by assay		
Virtual panel creation	Expert reviewed panels	PanelApp	https://panelapp.genomicsengland.co.uk/
		ClinGen	https://www.clinicalgenome.org/data-sharing/clinvar/
	Phenotype-driven	HPO	https://hpo.jax.org/app/

Table 6.
Considerations for gene selection, analysis and virtual panel creation.

6. Bioinformatics and IT

To assess accuracy, genetic variants must be compared against publicly available reference data obtained from 1000 Genomes Project.

Clinical association, gene validity and mutation spectrum are applied to the creation of virtual gene panels in order to aid variant interpretation and reporting. The considerations associated with constructing virtual gene panels and the analysis of variants are shown in **Table 6**.

7. Conclusions

The decision to implement WES in a clinical diagnostic environment is one that must take into account local context, which encompasses clinical complexity, staff resources, equipment resources and bioinformatic expertise. The decisions described here were made based on the above considerations with a view to establishing opportunity, the most important of which was to have a WES pipeline that could scale over time in terms of patients tested and with the potential to be a regional resource.

It should be stressed, however, that a WES pipeline is sandwiched by two critical elements: first, the need to focus on the quality and accurate quantitation of genomic DNA; which dictates the quality of everything that happens downstream, and second, to understand that the identification of DNA variants is technically demanding but the classification of those variants is not currently a fully automated process. The former can sometimes be overlooked, while the latter can be a daunting exercise. It is perhaps the subject of another book chapter to discuss the approaches to variant classification.

Conflicts of interest

The authors declare no conflicts of interest.

Thanks

The authors wish to thank Mr. Duncan Kay of Custom Science (NZ) for his generous suggestions regarding commercial providers for WES data analysis and Javier Botet of Sophia Genetics for his advice regarding quality management considerations.

Author details

Debra O. Prosser, Indu Raja, Kelly Kolkiewicz, Antonio Milano
and Donald Roy Love*
Division of Pathology Genetics, Department of Pathology, Sidra Medicine, Doha, Qatar

*Address all correspondence to: dlove@sidra.org

IntechOpen

References

[1] Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, et al. ACMG clinical laboratory standards for next-generation sequencing. Genetics in Medicine. 2013;**15**(9):733-747. DOI: 10.1038/gim.2013.92

[2] Hegde M, Santani A, Mao R, Ferreira-Gonzalez A, Weck KE, Voelkerding KV. Development and validation of clinical whole-exome and whole-genome sequencing for detection of germline variants in inherited disease. Archives of Pathology & Laboratory Medicine. 2017;**141**:798-805. DOI: 10.5858/arpa.2016-0622-RA

[3] Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. European Journal of Human Genetics. 2016;**24**:2-5. DOI: 10.1038/ejhg.2015.226

[4] Santani A, Simen BB, Briggs M, Lebo M, Merker JD, Nikiforova M, et al. Designing and implementing NGS tests for inherited disorders a practical framework with step-by-step guidance for clinical laboratories. The Journal of Molecular Diagnostics. 2019;**21**:369-374. DOI: 10.1016/j.jmoldx.2018.11.004

[5] Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. The Journal of Molecular Diagnostics. 2018;**20**(1):4-27. DOI: 10.1016/j.jmoldx.2017.11.003

[6] Ivády G, Madar L, Dzsudzsák E, Koczok K, Kappelmayer J, Krulisova V, et al. Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. BioMed Central Genomics. 2018;**19**:158. DOI: 10.1186/s12864-018-4544-x