Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease

Christian R. Marshall[1]*, Shimul Chowdhury[2], Ryan J. Taft[3], Mathew S. Lebo[4,5], Jillian G. Buchan[6,#], Steven M. Harrison[4,5], Ross Rowsey[7], Eric W. Klee[7,8], Pengfei Liu[9], Elizabeth A. Worthey[10,%], Vaidehi Jobanputra[11,13], David Dimmock[2], Hutton M. Kearney[7], David Bick[10], Shashikant Kulkarni,[9,12] Stacie L. Taylor[3], John W. Belmont[3], Dimitri J. Stavropoulos[1], Niall J. Lennon[5], on behalf of the Medical Genome Initiative.

[1]Genome Diagnostics, The Hospital for Sick Children, Toronto, ON, Canada

[2]Rady Children's Institute for Genomic Medicine, San Diego, CA, USA.

[3]Illumina Inc., San Diego, CA, USA

[4]Laboratory for Molecular Medicine, Partners HealthCare Personalized Medicine, Cambridge, MA, USA

[5]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[6]Stanford Medicine Clinical Genomics Program, Stanford Health Care, Stanford, CA, USA

[7]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA

[8]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

[9]Baylor Genetics Laboratories, Houston, TX, USA.

[10]HudsonAlpha Institute for Biotechnology, Huntsville, AB, USA

[11]New York Genome Center, New York, NY, USA

[12]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

[13]Department of Pathology and Cell Biology, Columbia University Irving Medical Center (CUIMC), New York, NY, USA

[#]current address of JGB is Department of Laboratory Medicine, University of Washington, Seattle, WA, USA

[%]current address is Center for Genomic Data Sciences, University of Alabama at Birmingham, AB, USA

Running head: *Analytical validity of clinical genome sequencing*

*Corresponding author:

Christian Marshall

Genome Diagnostics,

Department of Paediatric Laboratory Medicine

The Hospital for Sick Children

555 University Ave, Room 3203B

Toronto, ON, Canada M5G 1X8

ABSTRACT

Whole-genome sequencing (WGS) has shown promise in becoming a first-tier diagnostic test for patients with rare genetic disorders, however, standards addressing the definition and deployment practice of a best-in-class test are lacking. To address these gaps, the Medical Genome Initiative, a consortium of leading health care and research organizations in the US and Canada, was formed to expand access to high quality clinical WGS by publishing best practices. Here, we present consensus recommendations on clinical WGS analytical validation with a focus on test development, upfront considerations for test design, test validation practices, and metrics to monitor test performance. This work also provides insight into the current state of WGS testing at each member institution, including the utilization of reference and other standards across sites. Importantly, members of this Initiative strongly believe that clinical WGS is an appropriate first-tier test for patients with rare genetic disorders and at minimum is ready to replace chromosomal microarray analysis and whole-exome sequencing. The recommendations presented here should reduce the burden on laboratories introducing WGS into clinical practice and support safe and effective WGS testing for diagnosis of germline disease.

INTRODUCTION

Advances in next-generation sequencing (NGS) over the last decade have transformed genetic testing, increasing diagnostic yield and decreasing the time to reach a diagnosis.[1-5] Targeted NGS multigene panels have come into widespread use and whole-exome sequencing (WES) is a powerful aid in the diagnosis of patients with non-specific phenotypic features[6-10] and critically ill neonates,[11] where the differential diagnosis often includes multiple rare genetic disorders.[12] These approaches, however, have both workflow and test content limitations that may constrain their overall efficacy.

Whole-genome sequencing (WGS) can address many of the technical limitations of other NGS approaches including lack of completeness[13,14] and the detection of structural and complex variants.[15] WGS also enables the identification of non-coding variants, such as pathogenic variations disrupting regulatory regions, non-coding RNAs, and mRNA splicing.[16-18] Emerging uses of WGS include HLA genotyping,[19] pharmacogenetic testing,[20] and generation of polygenic risk scores.[21] Several studies have demonstrated the advantages of WGS for the identification of clinically relevant variants in a wide range of cohorts[22-26] and have shown the diagnostic superiority of WGS compared with conventional testing in pediatric patients[27-29] and critically ill infants.[30,31] As an efficient and cost-effective test, WGS is poised to replace targeted/exome sequencing and chromosomal microarray as a first-line laboratory approach in the evaluation of children and adults with a suspected genetic disorder.[28,32,33] WGS also has the benefit of periodic re-analysis across multiple variant types, which will increase diagnostic efficacy through updated annotation and analysis techniques.[34]

Although the stage is set for widespread adoption of clinical WGS, technical challenges remain and standards that address both the definition and the deployment practices of a best-in-class clinical WGS test have not been fully defined. Professional bodies have made progress in providing guidance for clinical WGS test validation[35,36] and best practices for benchmarking with reference standards and recommended accuracy measures are beginning to emerge.[37-39] However, these recommendations do not address specific challenges related to the setup of clinical WGS. The work presented here identifies the technical challenges and gaps in existing frameworks relevant to the analytical validation of a clinical WGS test and provides a real-world snapshot of the progress of laboratories within the Medical Genome Initiative for offering a clinical WGS test. We surveyed members of the initiative on all aspects of clinical WGS analytical validation and where possible present practical recommendations based on the consensus of the group. These recommendations may reduce the burden on laboratories who wish to introduce WGS into clinical practice and, more importantly, support safe and effective WGS testing for diagnosis of germline disease.

OVERVIEW OF CLINICAL WHOLE-GENOME SEQUENCING

All clinical diagnostic testing, including WGS, encompasses the entire process from obtaining a patient specimen to the delivery of a clinical report. The technical and analytical elements of clinical WGS can be separated into three stages: sample preparation including extraction and library preparation followed by sequence generation (Primary); read alignment and variant detection (Secondary); and the annotation, filtering, classification, prioritization, and interpretation of variants (Tertiary)[40]

(**Figure 1**). These components are common to all high-throughput sequencing tests and informatics pipelines but differences in components (e.g. informatics algorithms) will result in differences in data quality and accuracy. The focus of this manuscript is the Primary and Secondary analysis as these steps directly relate to the evaluation of test performance for the analytical validation of clinical WGS. Elements critical to establishing the analytical validity are described below in three sections: [1] test development and optimization, [2] test validation, and [3] quality management. Major steps and activities in the analytical validation are shown in **Figure 2** with key definitions in **Box 1.** Consensus recommendations spanning these sections are summarized in **Box 2**.

TEST DEVELOPMENT AND OPTIMIZATION

Test Definition Considerations

Analytical validation requirements will vary based on test definition and can include both technical considerations and the patient population under study. Although clinical WGS may be used for multiple indications (e.g. inherited disorders, cancer, healthy individuals), this document focuses on using clinical WGS for individuals with a suspected germline disorder as the primary use case. The principles of analytical validity described here, however, are applicable to all uses of clinical WGS.

Clinical WGS tests are predicated on a specific test definition that delineates both the variant types to be reported and the regions of the genome that will be interrogated (including any limitations), which may vary depending on the variant type. Importantly, test definitions should consider specimen source because this can influence the variant

types that can be reported. For example, if using a specimen source expected to yield limited DNA quantity, PCR for library preparation may be required and reporting of copy number variants and repeat expansions will be affected. Clear test definitions and identification of factors affecting reportable variant types will provide clarity to ordering physicians.

Classes of clinically relevant genetic variation detectable by clinical WGS include single nucleotide variants (SNVs), small deletions, duplications, insertions (indels), structural variation (SV) including copy number (CNV) and balanced rearrangements, mitochondrial (MT) variants, and repeat expansions (REs).[15] **A clinical whole-genome sequencing test should aim, wherever possible, to analyze and report on all possible detectable variant types. We recommend SNVs, indels, and copy number variants (CNVs) as a viable minimally appropriate set of variants for a WGS test. Laboratories should further aim to offer reporting of mitochondrial variants, repeat expansions, some structural variants and selected clinically relevant pseudogenes (Box 2, Supplementary Figure 1; Supplementary Table 1**). We also note that laboratories may not be able to validate all classes of variation prior to initial launch of clinical WGS, and that a phased approach to validation and subsequent test offering may be necessary. Regardless of the variant types a laboratory may choose to report, a thorough performance comparison between the WGS test and any current testing methodology is warranted. **Clinical WGS test performance should aim to meet or exceed that of any tests that it is replacing. If clinical WGS is deployed with any established gaps in performance compared to current gold standard tests, it should be noted on the test report (see Box 2).** The most immediate and

obvious use of clinical WGS is replacement of genome wide tests such as WES and CMA as there is evidence that WGS is analytically superior to these methods.[27,32] However, it should be noted that the robust detection of low-level mosaicism represents an important limitation of clinical WGS (at 40X mean depth) compared to WES or targeted panels where loss of performance may be a significant issue for some indications (e.g. epileptic encephalopathy).[41] Moreover, although other more complex variant classes like those mentioned above (e.g. mitochondrial variants with varying levels of heteroplasmy, REs, etc) can be identified using WGS, we recognize that in some cases the detection accuracy of these variant types may not yet be equivalent to currently accepted assays. There is still inherent value in including these variant classes to the test definition of clinical WGS to ensure as complete a test as possible. Laboratories planning to return these variant types should describe the limitations of their tests appropriately in the report and consider a confirmatory testing strategy before reporting. **Supplementary Table 1** offers examples of specific loci that could be offered as part of a clinical WGS test and others that are emerging. As with any genetic assay, the test definition should clearly state that a negative report in these instances does not preclude a diagnosis.

Upfront Considerations for Test Design

Upfront considerations for WGS test design such as sample and library preparation, sequencing methodology, sequence analysis, and annotation generally follow current guidelines[35,36,42] and are discussed in the **Supplementary Text**. More complex test design considerations that are most specific to clinical WGS, such as evaluation of

metrics to determine suitable WGS test coverage and the number and type of samples necessary for validation are discussed below.

Evaluation of Genome Coverage, Completeness and Callability

Defining and evaluating high quality genome coverage is one of the most important considerations in clinical WGS test development since it directly relates to the amount of data required to accurately identify variants of interest. **Metrics that measure genome completeness should be used to define the performance of clinical WGS and include sequencing uniformity and depth of coverage. These measures should be monitored with respect to callable regions of the genome and the variant calling accuracy for each variant class and should be subsequently compared to orthogonally investigated truth sets (Box 2).** While universal cutoffs are not yet established, a combination of depth of coverage, base quality, and mapping quality is recommended to assess callability.[43] The majority of laboratories in this initiative calculate both raw and usable coverage, the latter metric relating to reads used in variant detection and excluding poorly mapped reads, low quality base pairs, and overlapping paired reads. All sites have evaluated the performance of clinical WGS using varying mean depth of coverage and assessed the completeness and accuracy of variant calling in specific target files such as a reference standard, or comparison to the method clinical WGS is replacing (e.g. WES) (**Supplementary Figures 2 and 3)**. Variability in assessment methodology can result in differences in metrics and cutoffs (**Table 1**); however, reference genomes performed similarly among the groups in this

consortium (**Supplementary Table 2**). If the laboratory is providing WGS from different DNA sources, these evaluations should be completed for each specimen type.

Reference Standard Materials and Sample Types for Clinical WGS Validation

High quality reference standard materials and positive controls with associated truth data sets are a necessary resource for laboratories offering clinical WGS. **The analytical validation of clinical WGS should include publicly available reference standards in addition to commercially available and laboratory held positive controls for each variant and type. For variant types commonly addressed by the field, including SNVs and indels, a low minimal number of controls can be utilized if these include well-accepted reference standards. For variant types where standards are still evolving (e.g. REs), a larger number of samples should be employed (Box 2)**. The National Institute of Standards and Technology (NIST) NA12878 genome and Platinum Genomes are routinely utilized by NGS laboratories seeking to establish WGS analytical validity.[42] These genomes have the benefit of thousands of variants that have been curated and confirmed across many technologies.[44,45] Within this initiative, all groups have used NA12878 for validation and most groups also utilize the Ashkenazi Jewish and Chinese ancestry trios from the Personal Genome Project that are available as NIST Reference Materials (**Supplementary Table 3**).

The ability to sub-categorize analytical performance by variant type is another benefit of using well-characterized reference materials. Genome-wide estimates of sensitivity often mask poor performance in certain sequence contexts or across different

variant attributes. Understanding performance in difficult regions of the genome is important for accurately representing the limitations of the assay and setting benchmarks against which new analytical tools and methods can be developed. The Global Alliance for Genomics and Health (GA4GH) Benchmarking Team recently developed tools (https://github.com/ga4gh/benchmarking-tools) to evaluate performance in this way. Currently, all members of this initiative have incorporated or intend to use the results of such an analysis in their analytical validation study.

Reference standard materials alone are not sufficient to establish validity of a test, however. For example, both the specimen and disease context must also be taken into consideration when sourcing samples for a validation study. For clinical WGS laboratories in this consortium, specimen context has included determination of the acceptable sample types (e.g. blood, saliva, tissue) with associated representative positive controls. Some pathogenic variants, including short tandem repeats, low copy repeats, SVs with breakpoints within non-unique sequences, paralogs and pseudogenes, occur in regions of the genome that are difficult to sequence, align, and map. If analysis and reporting of these loci is planned, the laboratory should perform validation assessments on samples with these specific variant types to determine robustness. Since performance expectations may not be well established for these variants, a large number of positive controls should be used (see below and **Supplementary Table 3**).

TEST VALIDATION

Clinical WGS requires a multi-faceted approach to analytical validation due to the large number of rare genetic disease loci, the number and different classes of variation that can be detected, and the genomic-context driven variability in variant calling accuracy. Traditional summary statistics defining performance metrics across the entire assay are necessary, but not sufficient. **The analytical validation framework should include metrics that account for genome complexity, with special attention to sequence content and variant type (Box 2).** For example, SNV and CNV variants have different calling constraints that can be affected differently by low-complexity sequence**.** Specific test validation recommendations that address these and other clinical WGS specific validation requirements are discussed in detail below. Other considerations that are not unique to clinical WGS include sequencing bias, repeatability, limits of detection, interference, homology, software validation, and specific disease variant validation; these are discussed in the **Supplementary Text**.

Performance Metrics, Variant Type, and Genomic Context

Analytical validation is the first step in ensuring diagnostic accuracy and is classically measured in terms of sensitivity (recall) and specificity. However, this initiative agrees with current recommendations from the GA4GH to use precision as a more useful metric than specificity owing to the large number of true negatives expected by clinical WGS.[38] The FDA suggests similar, albeit slightly different metrics, for validation of NGS assays including positive percent agreement (sensitivity), negative percent agreement (specificity) and technical positive predictive value (equivalent to precision above) as

well as reporting the lower bound of the 95% confidence interval (CI).[39] Relevant definitions are provided in **Box 1**.

This initiative recommends following published guidelines as described above, as the performance metrics are generally applicable to clinical WGS. In addition to the global metrics of accuracy (sensitivity, precision), repeatability (technical replicates performed under identical conditions), reproducibility (comparison of results across instruments), and limits of detection assessment (e.g. mosaic SNVs) should also be measured. For SNV and indels, gold standard reference data are available, as described above and can be used to calculate performance metrics.[42] Other variant types may not have standard truth sets available, so comparative metrics should be confined to positive and negative percent agreement (PPA and NPA, respectively) against laboratory or commercially acquired samples assessed using a precedent technology. Laboratories may also consider creating virtual datasets and analytically mixed specimens for validation of the variant types that may not have standard truth sets available.

Performance thresholds should be predetermined and matched to clinical requirements for low diagnostic error rates. Flexibility in performance thresholds at the stage of variant calling may be acceptable as long as these deviations are documented and laboratory procedures include additional confirmatory assessments. These can include additional bioinformatics analyses, manual inspection by analysts, and orthogonal laboratory testing. The amount of data being examined in a clinical WGS test requires that confirmatory methods be restricted to small subsets of the data with potentially high clinical impact. No calls and invalid calls should not be used in

calculations of sensitivity, precision, or TPPV in the validation of variant calling. Instead, these should be documented separately as part of the accuracy of the test and, where possible, genomic intervals that routinely have low map quality and coverage should be flagged in the clinical WGS test definition.

Identification of different variant types require unique calling algorithms, resulting in differences in analytical performance. Further stratification by size is warranted for some common variant types to provide greater insight into overall test performance. For example, GA4GH recommends binning insertions, deletions, and duplications into size bins of <50bp, 50-200bp, and >200bp[46], although it is important to note that most laboratories in this Initiative assess additional smaller bins (**Supplementary Figure 4**). Similarly, CNV size bins and minimum cutoffs are frequently tied to maximum resolution of current clinical microarrays, which can vary from 20 kb to 100 kb, depending on the platform used. Laboratories in this initiative that currently offer CNVs as part of the test report events down to 1-2 kb using a depth-based CNV caller, whereas smaller CNV events require split or anomalous read pair information partnered with a depth assessment.

Variant calling performance can be affected by the sequence context of the region itself, or, in the case of large variants, the surrounding bases. Currently, there are no best practices for the identification of systematically problematic regions or comprehensive population-level truth data sets, but all members of this Initiative have developed internal methods to identify such regions. These include regions where clinical WGS may perform poorly, including paralogous genes, which are excluded from the test definition in order to guide appropriate clinical ordering. The Initiative also

recommends that regions identified as systematically problematic, or that negatively affect variant calling tied to particular variant types, are documented as part of the test validation, with clear mention of the limitations of the test on the reports and made available to ordering clinicians as appropriate.

Sample Number and Type for Validation

Determining the number of samples and specimen types required for clinical WGS validation is challenging due to the fact that the test interrogates thousands of disease genes with different variant types predominating (SNV vs CNV vs RE) in a variety of sequence contexts. It is the Initiative's position that it is not technically or practically feasible to validate all possible pathogenic variants genome-wide, and that the repeatable and accurate assessment of genome reference standards is sufficient to establish global accuracy. This assessment, however, should include all reportable disease-associated regions. Variant or disease-specific controls should be employed for any variant beyond SNVs and indels. The samples chosen for these validations should include the most commonly affected genes and variants detected for the most commonly referred clinical indications. The number of specific variants that should be assessed may vary according to variant type, genomic context, and the availability of appropriate reference samples. Where possible, however adhering to a statistically rigorous approach similar to that outlined by Jennings et al,[47] which incorporates a confidence level of detection and required probability of detection, is recommended. When applying this method and requiring a 95% reliability with 95% confidence interval, at least 59 variants should be used in the performance assessment as has been

previously published.[36,47] One can then apply this calculation to those variant classes and genomic contexts that are both germane and feasible to establish the overall analytical validation plan and sample/variant counts. Among members of this Initiative, the number of non-reference standard positive controls used for test validation was dependent on the variant type and ranged between 4 to 175 cases with more cases typically used to validate CNV and RE calling and reporting.

QUALITY MANAGEMENT

As with any laboratory test, groups performing clinical WGS should have a robust quality management program in place for quality control and assurance. Test run quality metrics and performance thresholds for clinical WGS should be assessed at the sample level as part of quality control. A quality assurance program should periodically monitor quality metrics over time and identify trends in test performance related to reagent quality, equipment performance, and technical staff. Clinical WGS sample level quality metrics describe whether the biological specimen and end-to-end test are technically adequate [i.e. whether the test provides the expected analytical sensitivity and technical positive predictive value for all variant types (SNVs, indels, CNVs, SVs)] within the reportable range of the genome established during test validation. In the following sections we briefly describe relevant test performance and quality metrics unique to clinical WGS. Important metrics, definitions and suggested cutoffs for pass/fail and monitoring of clinical WGS are shown in **Table 1 and Supplementary Table 4,** respectively. Additional details for quality management are available in the **Supplementary Text.** As part of the quality management plan, laboratories should have

a strategy in place for minor and major updates of the tools and algorithms used in the various steps of the analytical pipeline including careful re-evaluation of previously run datasets or re-calculation of the performance specifications (**Supplementary Text**)

Sample, Extraction and Library Preparation

The yield and quality (e.g. fluorometry and size range) of the DNA should have defined criteria for acceptance that allows a DNA sample to be passed to library preparation and sequencing. For clinical WGS, sample pooling and molecular barcoding is utilized in the majority of laboratories. Some platforms benefit from a dual-barcoding strategy (i.e. a barcode on each end of the library molecule) to reduce the possibility of barcode hopping on the flowcell.[48] Performance metrics (e.g. library concentration) with acceptance thresholds must be established and the results from each sample must be documented. For sample and library preparation, procedures are needed to detect and interpret systematic drops in quality and/or the percentage of samples meeting minimum quality requirements. A control for library preparation may be used to monitor quality and troubleshoot preparation versus sample issues.

Sequencing

Quality metrics are calculated for every run of the instrument. Test development optimization and validation processes establish which metrics are reviewed for every sample and the specifications for each. Important sequencing metrics include the amount of purity filtered data produced (PF bases), the alignment rate of these bases (PF reads aligned %), the predicted raw and usable coverage of the genome (mean

coverage), proportion of reads that are duplicates (% duplication), and any evidence of sample contamination (% contamination). Mean coverage and completeness of coverage are commonly used metrics but as discussed previously these may be calculated differently across groups (see previous section on coverage evaluation). Examples of performance metrics used by members of this consortium to evaluate WGS for Pass/Fail and monitoring are listed in **Table 1** and **Supplementary Table 4**, respectively. However, it is important to note that at the time of publication the Initiative was unable reach consensus as to which metrics should be used and the corresponding thresholds that need to be met to qualify as a passing clinical WGS test. There was general agreement on the types of measures that are important but often these were calculated in different ways and coming to a consensus was difficult. This is likely a reflection of the evolving technology and the way in which each group validated testing in the absence of accepted guidelines.

One of the biggest challenges for laboratories offering clinical WGS is the application of controls to comply with regulatory guidelines. Guidelines recommend the use of positive, negative, and sensitivity controls (e.g. CAP Molecular Pathology Checklist, August 2018 – MOL.34229 Controls Qualitative Assays) to ensure that all steps of the assay are successfully executed without contamination. **Ongoing quality control of a clinical whole-genome test should include identification of a comprehensive set of performance metrics, continual monitoring of these metrics across samples over time, and the use of positive controls on a periodic basis dependent on overall sample volume (Box 2).** Although the inclusion of a control reference standard in every sequencing run is ideal, it is not practical or financially

viable for a laboratory performing clinical WGS. Moreover, the use of positive and negative controls may be informative for the overall performance of a sequencing run, but will not be reflective of sample specific differences and may incorrectly indicate adequate test performance.

There are additional positive and negative control strategies that some laboratories may choose to employ. Some of the groups in the Initiative use PhiX standing for the empirical measure of sequencing error rate. For variant positive controls, one approach is the use of low-level spike-ins of well-characterized positive control samples that include a spectrum of variants in each sequencing run. Similarly, some groups in the Initiative are exploring the use of synthetic spike-in constructs, including Sequins,[49] which can be added to a run at a low level (<1% of reads) and enable a performance assessment that can serve as a process control for at least some variant types. Within this Initiative, most groups run a reference standard at periodic intervals and check for deviations from expected calling accuracy and concordance with previously run samples.

Secondary and Tertiary Analysis

After sequencing and demultiplexing, secondary analysis consisting of alignment, mapping and variant calling is performed. A detailed description of secondary analysis monitoring is further discussed in the **Supplementary Text,** with a minimum set of performance passing metrics shown in **Table 1**. For clinical WGS, it is particularly important to monitor global mapping metrics and assess clinically significant loci for completeness (e.g. OMIM genes, ClinVar pathogenic variants). Following variant

calling, it is necessary to express variants in a standardized nomenclature using guidelines from the Human Genome Variation Society (HGVS) and link externally-available information to each variant, such as population allele frequencies or occurrences in variant databases. This process is collectively referred to as variant annotation and is the first step in tertiary analysis. Proper annotation of variants is necessary for tertiary analysis or the downstream variant filtering, prioritization, and classification for reporting. Because of the breadth and complexity of genomic variation identified through clinical WGS, a mix of HGVS and the International Standing Committee on Human Cytogenetic Nomenclature (ISCN) are currently employed by all members of the Initiative (**see Supplementary Text**).

Summary

Clinical WGS is poised to become a first-tier test for the diagnosis of those individuals with suspected genetic disease. Although some guidelines are beginning to emerge that offer recommendations for the analytical validation of genome testing, specific challenges related to the set up and deployment of clinical WGS are not addressed. In this document we outlined consensus recommendations for the analytical validation of clinical WGS, based on the experiences of members of the Medical Genome Initiative. We focused on providing practical advice for test development optimization, validation practices and ongoing quality management for the deployment of clinical WGS. Even amongst members within the initiative it was often challenging to come to a consensus on specific recommendations, since there are often different but equally valid approaches to the analytical validation of WGS. However, members of this Initiative

agreed upon the endorsement of clinical WGS as a viable first-tier test for individuals with rare disorders and that it should replace CMA and WES. The recommendations provided here are meant to represent a snapshot of the current state of the field and we expect best practices to continue to evolve. Nonetheless, the practical advice in this document and future Medical Genome Initiative publications should aid laboratories in introducing WGS into clinical practice. To this end, our group is committed to providing best practices on clinical WGS topics both upstream and downstream from analytical validity including genome interpretation, data infrastructure, and clinical utility measures.

ACKNOWLEDGEMENTS

The authors thank Michael Eberle, Mar Gonzàlez-Porta, Cinthya Zepeda Mendoza, Cherisse Marcou, Jaime Lopez, R. Tanner Hagelstrom, and Kirsten Curnow for critical comments and review of the manuscript.

COMPETING INTERESTS

SLT, RJT, and JWB are current employees and shareholders of Illumina Inc.

AUTHOR CONTRIBUTIONS

CRM prepared the manuscript. CRM, JWB, NJL, SC, RJT conceived the idea and wrote the paper. SC, CRM, DJS, NJL, MSL, collected and analyzed survey data. SC, NJL, MSL, CRM, VJ, PL contributed original data and analysis tools. JGB, SMH, RR, EWK, DJS, SC, NJL, MSL contributed writing of manuscript sections. HMK, DB, DD, SK, EAW provided design advice and critical review of the manuscript.

DISPLAY ITEMS

Box 1: Definitions of Key Terms

Term	Definition
Analytical Validity	A measure of the accuracy with which a test predicts a genetic change.
Callable Region (Callability)	Regions of the genome where accurate genotype calls can be reliably derived. Typically expressed as a percentage of non-N reference calls with a passing genotype across a target (Whole genome, OMIM genes)
Completeness	Proportion of the genome, or a select region of interest (e.g. exons), that have sufficient, high-quality sequencing reads to enable identification of variants.
Negative Percent Agreement (NPA)	Equivalent to Specificity. The proportion of correct calls in the absence of a variant, reflecting the frequency of false positives.
No-call or Invalid call	A position within the testing interval where no variant call can be been made.
Orthogonal Confirmation	Verification of a specific variant call using a different testing modality.
Positive Percent Agreement (PPA)	Equivalent to Recall/Sensitivity. Ability of the test to correctly identify variants that are present in a sample, reflecting the frequency of false negatives
Precision	Equivalent to TPPV. The fraction of variant calls that match the expected, reflecting the number of false positives per test.
Predicted zygosity	In diploid organisms, one allele is inherited from the male parent and one from the female parent. Zygosity is a description of whether those two alleles have identical or different DNA sequences.
Read depth	A measure of the number of sequence reads that are aligned to a specific base or locus.
Repeatability	The percent agreement between the results of successive tests carried out under the same conditions of measurement.
Reproducibility	The percent agreement between the results of tests under a variety of (e.g. different operators, machines, time frames).
Sensitivity or Recall	Equivalent to PPA. Ability of the test to correctly identify variants that are present in a sample, reflecting the frequency of false negatives.
Specificity	Equivalent to NPA. The proportion of correct calls in the absence of a variant, reflecting the frequency of false positives.
Technical Positive Predictive Value (TPPV)	Equivalent to Precision. The fraction of variant calls that match the expected, reflecting the number of false positives per test.
Uniformity of Coverage	A measure of the evenness of sequencing read distribution along a genomic interval. Note that the interval may encompass the majority of the genome.

Box 2: Key Summary Statements for Analytical Validation of WGS

1. A clinical whole-genome sequencing test should aim, wherever possible, to analyze and report on all possible detectable variant types. We recommend SNVs, indels, and copy number variants (CNVs) as a viable minimally appropriate set of variants for reporting in a WGS test. Laboratories should further aim to offer reporting of mitochondrial variants, repeat expansions, some structural variants and selected clinically relevant pseudogenes (Box 2, Supplementary Figure 1; Supplementary Table 1).
2. Clinical WGS test performance should aim to meet or exceed that of any tests that it is replacing. If clinical WGS is deployed with any established gaps in performance compared to current gold standard tests, it should be noted on the test report.
3. Metrics that measure genome completeness should be used to define the performance of clinical WGS and include sequencing uniformity and depth of coverage. These measures should be monitored with respect to callable regions of the genome and the variant calling accuracy for each variant class compared to orthogonally investigated truth sets.
4. The analytical validation of clinical WGS should include publicly available reference standards, in addition to commercially available and laboratory held positive controls for each variant and type. For variant types commonly addressed by the field, including SNVs and indels, a low minimal number of controls can be utilized if these include well-accepted reference standards. For variant types where standards are still evolving (e.g. REs), a larger number of samples should be employed.
5. The analytical validation framework should include metrics that account for genome complexity, with special attention to sequence content and variant type.
6. Ongoing quality control of a clinical whole-genome test should include identification of a comprehensive set of performance metrics, continual monitoring of these metrics across samples over time, and the use of positive controls on a periodic basis dependent on overall sample volume.

REFERENCES

1 Boycott, K. *et al.* The clinical application of genome-wide sequencing for monogenic diseases in Canada: Position Statement of the Canadian College of Medical Geneticists. *J Med Genet* **52**, 431-437, doi:10.1136/jmedgenet-2015-103144 (2015).

2 ACMG. Points to consider in the clinical application of genomic sequencing. *Genet Med* **14**, 759-761, doi:10.1038/gim.2012.74 (2012).

3 Gullapalli, R. R. *et al.* Clinical integration of next-generation sequencing technology. *Clin Lab Med* **32**, 585-599, doi:10.1016/j.cll.2012.07.005 (2012).

4 Matthijs, G. *et al.* Guidelines for diagnostic next-generation sequencing. *Eur J Hum Genet* **24**, 2-5, doi:10.1038/ejhg.2015.226 (2016).

5 Vrijenhoek, T. *et al.* Next-generation sequencing-based genome diagnostics across clinical genetics centers: implementation choices and their effects. *Eur J Hum Genet* **23**, 1142-1150, doi:10.1038/ejhg.2014.279 (2015).

6 Farwell, K. D. *et al.* Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. *Genet Med* **17**, 578-586, doi:10.1038/gim.2014.154 (2015).

7 Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *Jama* **312**, 1880-1887, doi:10.1001/jama.2014.14604 (2014).

8 Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *Jama* **312**, 1870-1879, doi:10.1001/jama.2014.14601 (2014).

9 Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305-1314, doi:10.1016/s0140-6736(14)61705-0 (2015).

10 Stark, Z. *et al.* A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet Med* **18**, 1090-1096, doi:10.1038/gim.2016.1 (2016).

11 Meng, L. *et al.* Use of Exome Sequencing for Infants in Intensive Care Units: Ascertainment of Severe Single-Gene Disorders and Effect on Medical Management. *JAMA Pediatr* **171**, e173438, doi:10.1001/jamapediatrics.2017.3438 (2017).

12 Delaney, S. K. *et al.* Toward clinical genomics in everyday medicine: perspectives and recommendations. *Expert Rev Mol Diagn* **16**, 521-532, doi:10.1586/14737159.2016.1146593 (2016).

13 Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* **112**, 5473-5478, doi:10.1073/pnas.1418631112 (2015).

14 Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A. & Gilissen, C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum Mutat* **36**, 815-822, doi:10.1002/humu.22813 (2015).

15 Bick, D., Jones, M., Taylor, S. L., Taft, R. J. & Belmont, J. Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. *J Med Genet*, doi:10.1136/jmedgenet-2019-106111 (2019).
16 Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* **46**, 61-64, doi:10.1038/ng.2826 (2014).
17 Merico, D. *et al.* Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. *Nat Commun* **6**, 8718, doi:10.1038/ncomms9718 (2015).
18 Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e524, doi:10.1016/j.cell.2018.12.015 (2019).
19 Hayashi, S. *et al.* ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data. *BMC Genomics* **19**, 790, doi:10.1186/s12864-018-5169-9 (2018).
20 Cohn, I. *et al.* Genome sequencing as a platform for pharmacogenetic genotyping: a pediatric cohort study. *NPJ Genom Med* **2**, 19, doi:10.1038/s41525-017-0021-8 (2017).
21 Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224, doi:10.1038/s41588-018-0183-z (2018).
22 Carss, K. J. *et al.* Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *Am J Hum Genet* **100**, 75-90, doi:10.1016/j.ajhg.2016.12.003 (2017).
23 Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344-347, doi:10.1038/nature13394 (2014).
24 Yuen, R. K. *et al.* Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat Med* **21**, 185-191, doi:10.1038/nm.3792 (2015).
25 Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet* **47**, 717-726, doi:10.1038/ng.3304 (2015).
26 Scocchia, A. *et al.* Clinical whole genome sequencing as a first-tier test at a resource-limited dysmorphology clinic in Mexico. *NPJ Genom Med* **4**, 5, doi:10.1038/s41525-018-0076-1 (2019).
27 Stavropoulos, D. J. *et al.* Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. *NPJ Genom Med* **1**, doi:10.1038/npjgenmed.2015.12 (2016).
28 Clark, M. M. *et al.* Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom Med* **3**, 16, doi:10.1038/s41525-018-0053-8 (2018).
29 Soden, S. E. *et al.* Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Sci Transl Med* **6**, 265ra168, doi:10.1126/scitranslmed.3010076 (2014).
30 Farnaes, L. *et al.* Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom Med* **3**, 10, doi:10.1038/s41525-018-0049-4 (2018).

31 Saunders, C. J. *et al.* Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med* **4**, 154ra135, doi:10.1126/scitranslmed.3004041 (2012).
32 Lionel, A. C. *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*, doi:10.1038/gim.2017.119 (2017).
33 Gross, A. M. *et al.* Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genet Med*, doi:10.1038/s41436-018-0295-y (2018).
34 Costain, G. *et al.* Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. *Eur J Hum Genet* **26**, 740-744, doi:10.1038/s41431-018-0114-6 (2018).
35 Aziz, N. *et al.* College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med* **139**, 481-493, doi:10.5858/arpa.2014-0250-CP (2015).
36 Roy, S. *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. *J Mol Diagn* **20**, 4-27, doi:10.1016/j.jmoldx.2017.11.003 (2018).
37 Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* **37**, 561-566, doi:10.1038/s41587-019-0074-6 (2019).
38 Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**, 555-560, doi:10.1038/s41587-019-0054-x (2019).
39 FDA. (ed US Food and Drug Administration) (2018).
40 Richards, 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* **17**, 405-424, doi:10.1038/gim.2015.30 (2015).
41 D'Gama, A. M. & Walsh, C. A. Somatic mosaicism and neurodevelopmental disease. *Nat Neurosci* **21**, 1504-1514, doi:10.1038/s41593-018-0257-3 (2018).
42 Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* **15**, 733-747, doi:10.1038/gim.2013.92 (2013).
43 Goldfeder, R. L. & Ashley, E. A. A precision metric for clinical genome sequencing. *bioRxiv*, 051490, doi:10.1101/051490 (2016).
44 Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**, 246-251, doi:10.1038/nbt.2835 (2014).
45 Zook, J. *et al.* Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. *bioRxiv*, 281006, doi:10.1101/281006 (2018).
46 GA4GH. *Benchmarking Performance Stratification for SNVs and Small Indels*, <https://github.com/ga4gh/benchmarking-tools/blob/master/doc/standards/GA4GHBenchmarkingPerformanceStratification.md> (2017).

47 Jennings, L. J. *et al.* Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn* **19**, 341-365, doi:10.1016/j.jmoldx.2017.01.011 (2017).
48 Costello, M. *et al.* Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332, doi:10.1186/s12864-018-4703-0 (2018).
49 Hardwick, S. A., Deveson, I. W. & Mercer, T. R. Reference standards for next-generation sequencing. *Nat Rev Genet* **18**, 473-484, doi:10.1038/nrg.2017.44 (2017).

FIGURE LEGENDS

Figure 1. **Clinical Whole-genome Sequencing Workflow**

The workflow for clinical WGS involves three major analysis steps spanning wet laboratory and informatics processes: Primary (blue) analysis refers to the technical production of DNA sequence data from biological samples; Secondary (green) analysis includes the identification of DNA variants; and Tertiary (yellow) analysis refers to the annotation of variants and the subsequent filtering, triaging, classification, and interpretation. Health record information and phenotype can be mined and converted to Human Phenotype Ontology (HPO) terms to aide variant interpretation. Primary analysis involves the sample and library preparation and sequencing with base calling followed by extensive quality control (QC). During this stage, genotyping with an orthogonal method (SNP-array or targeted assay) is performed for QC purposes. Secondary analysis involves mapping, read alignment, and variant calling. Different classes of variation (SNVs, SV, CNVs, mitochondrial, repeat expansions) will use different algorithms that can be run in parallel. Aside from QC of alignment and variant calling, the orthogonal genotyping can be used to ensure no sample mix up has occurred throughout the workflow. Tertiary analysis begins with the annotation of variants followed by the filtering, stratification, and variant prioritization depending on the phenotype and clinical indication for testing. Classification of variants according to ACMG guidelines and final case interpretation will ultimately be driven by the case phenotype, and variants can be reported based on primary, secondary, or pharmacogenetics following any necessary confirmation method. Confirmation may be performed with an orthogonal wet lab method or in silico examination of the data based

on how the test was validated. Throughout the process, collection of aggregate data will be necessary to generate internal allele frequencies and for sharing of interpreted data with repositories.

Figure 2. Key steps in the analytical validation of a clinical WGS test

Key steps in the analytical validation of clinical WGS include test development optimization, test validation and quality management. Each step involves activities that lead to defined outcomes.

Figure 1: Clinical WGS Workflow

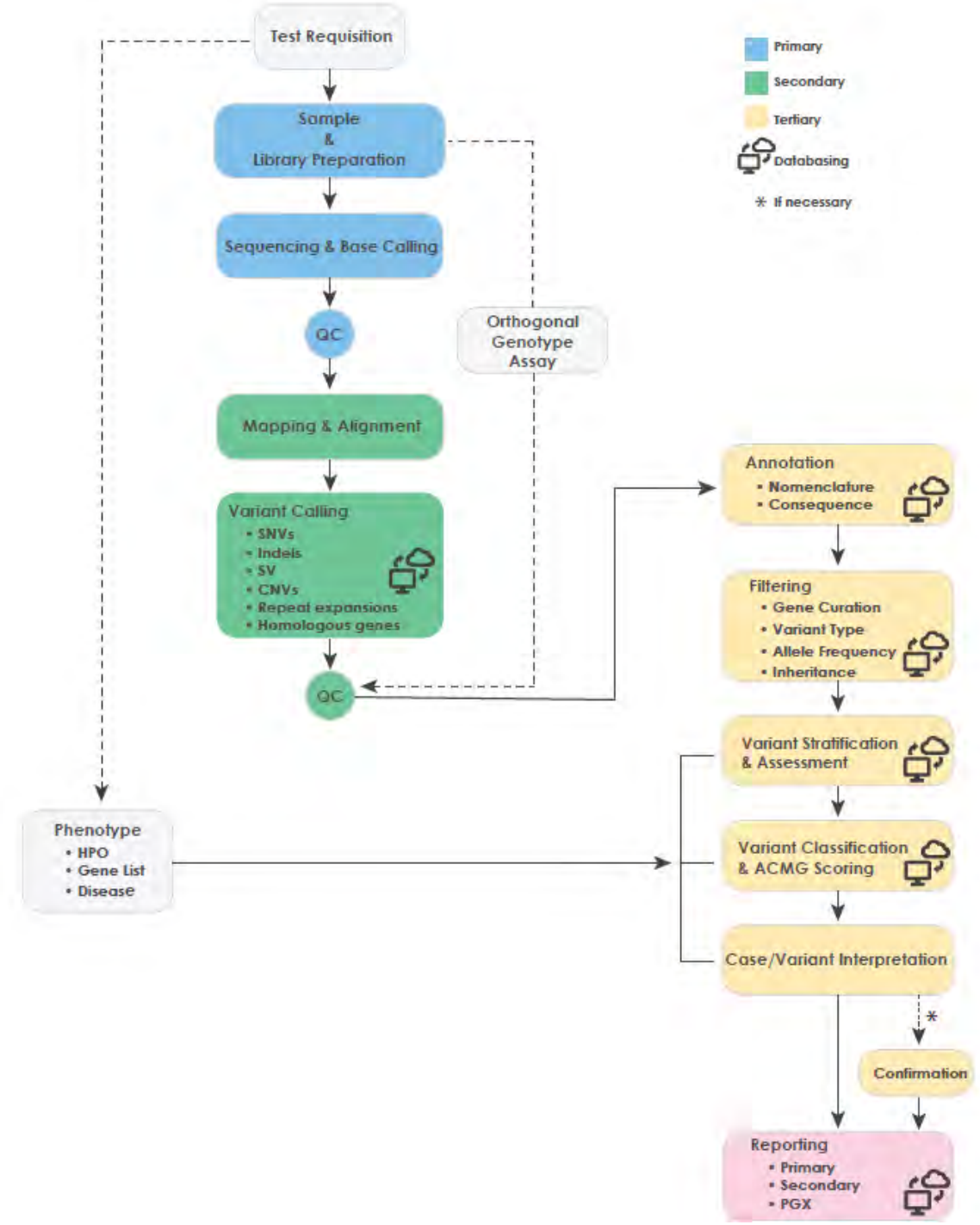


Figure 2. Key steps in the analytical validation of a clinical WGS test

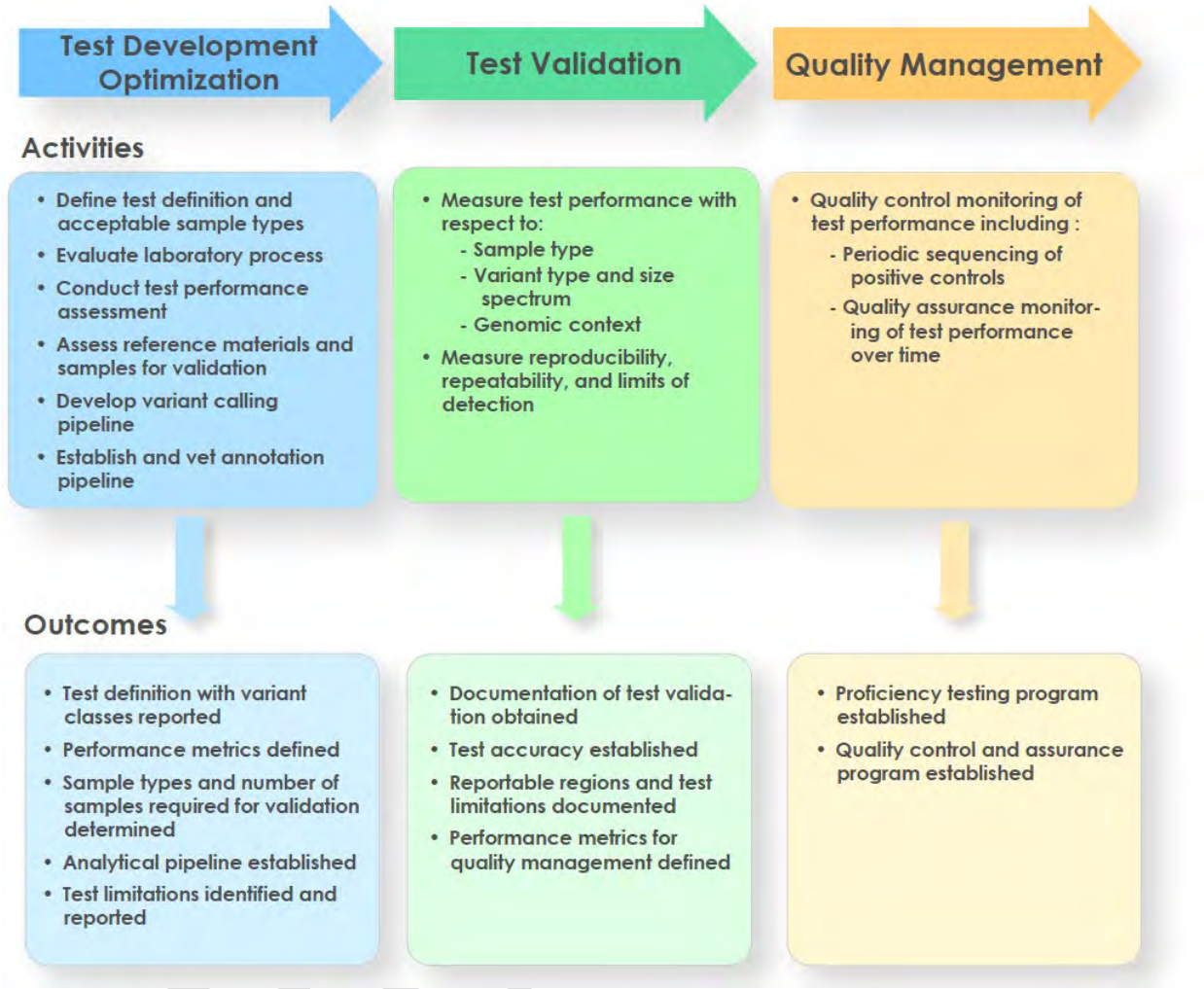


Table 1. Examples of Pass/Fail performance metrics for clinical WGS[a]

Metric	Description	Type (Threshold)
Sample Identity	Concordance with genotype (orthogonal measurement)	PASS/FAIL (match)
Contamination	The estimated level of sample cross-individual contamination based on a genotype-free estimation.	PASS/FAIL (≤2%)
Gb >=Q30	Total Gb of data with base quality score >Q30	PASS/FAIL (<85Gb)
Autosome Mean Coverage[b]	The mean coverage across human autosomes, after all filters are applied.	PASS/FAIL (≥30-40X)
% Callability[c]	Percent of non-N reference positions in autosomal chromosomes with a passing genotype call	PASS/FAIL (>95%)

[a]See Supplementary Table 4 for additional recommended metrics for monitoring clinical WGS test performance
[b]Members of the Initiative use either 30 or 40X mean coverage as a cutoff.
[c]Callability, or the fraction of the genome where accurate calls can be made can be calculated in different ways. The description in the table represents one way to calculate callability but there are others including using the percentage of base pairs that reach a read depth (RD) of 20 with base quality (BQ) and mapping quality (MQ) of 20.

Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease

Christian R. Marshall[1]*, Shimul Chowdhury[2], Ryan J. Taft[3], Mathew S. Lebo[4,5], Jillian G. Buchan[6,#], Steven M. Harrison[4,5], Ross Rowsey[7], Eric W. Klee[7,8], Pengfei Liu[9], Elizabeth A. Worthey[10,%], Vaidehi Jobanputra[11,13], David Dimmock[2], Hutton M. Kearney[7], David Bick[10], Shashikant Kulkarni,[9,12] Stacie L. Taylor[3], John W. Belmont[3], Dimitri J. Stavropoulos[1], Niall J. Lennon[5], on behalf of the Medical Genome Initiative.

Contents

I. SUPPLEMENTARY TEXT

II. Test Development and Optimization

There are several components of WGS test design that should be taken into consideration as part of test development and optimization, including sample and library preparation, sequencing methodology, sequence analysis, and annotation. These components generally follow current guidelines and are summarized below.

Sample preparation. Sample preparation for clinical WGS largely follow standard laboratory recommendations wherein proper labeling and sample tracking is critical to ensure the integrity of the final result. In the majority of cases, the laboratory may receive purified genomic DNA, whose quality meets or exceeds pre-defined parameters, or biological samples which include, but are not limited to, whole blood, blood spot on filter paper, saliva, buccal brush/swab, urinary sediment, tissue, or cultured cells, ensuring that there is sufficient sample for DNA preparation. Clinical WGS using short read technology is generally robust to DNA source but some specimen types, like saliva, have varying degrees of non-human DNA contamination that decrease overall genome coverage. In addition, the small amount of DNA needed (1ug) for PCR-free library preparation and followed by short read technology sequenced generally allows most specimen types of DNA sources to be used; however, it should be noted that some specimen types like blood spots may not yield enough DNA for PCR-free libraries and may limit downstream analysis (i.e. copy number or repeat expansion analysis). Moreover, it should be noted that long-read sequencing technologies will have different DNA requirements and the laboratory may want to determine if the extraction technique is compatible. All members of this Initiative have validated clinical WGS with whole

blood as a sample type, but only a few laboratories have validated alternate tissue sources like buccal, saliva, fibroblasts, and blood spots.

Given the multi-step processes needed to generate a final result, a sample tracking procedure is recommended during tube and instrument transfers and to confirm the integrity of the final results. There is no standard method employed by Initiative members, but examples of sample tracking include comparison of WGS data to SNPs genotyped with a multiplex assay or custom microarray, STR marker analysis or spike-in methodology. Similar methods should be used when case-parent trios are sequenced or when other family members are included in the clinical testing strategy. Formal checks for Mendelian errors to establish parentage and to assess relatedness among other family members should be computed using standard methods.

Library preparation. PCR-free library preparation is preferred for clinical WGS because it improves the evenness of coverage and other bioinformatics analyses such as calling copy number variants calling[1] and repeat-disease testing expansions.[2] However, when using DNA isolated from saliva samples, contamination of non-human (bacterial) sequence may be significant when using a PCR-free library. Members of this initiative have observed bacterial contamination of between 8-30%, which affects overall coverage (unpublished data). When DNA quantity is limited, standardized amplification methods can be used for clinical WGS but limitations should be noted including increased PCR duplicates. Current sample preparation methods utilize sample indexing in which a sample-identifying oligonucleotide is introduced during the library preparation to allow for sample pooling. Even in cases where there is a 1:1 match between sequence yield from a single lane of a flowcell and the desired coverage of a single

WGS sample, there are benefits to sample pooling. Pooling several samples and spreading the library across several lanes or instruments can offset the effects of any lane to lane or instrument to instrument variability. Similarly, if more sequence data is needed to reach coverage targets, pooling many samples and sequencing can be a cost-effective option to top up coverage. Incorrect index assignment may be an issue with PCR-free libraries and compounded by choice of downstream sequencing technology; thus, using uniquely dual indexed libraries is recommended to mitigate index-hopping error. Pooling of libraries is common practice to achieve desired overall coverage but is a practical issue for laboratories to ensure that samples are balanced as there is no well-accepted solution for how best to quantify and normalize before loading onto the sequencers. Library quantification using quantitative PCR is standard across laboratories with some using low pass sequencing with a lower throughput sequencer (e.g. MiSeq™ or iSeq™ 100) to ensure final pooling is balanced.

Sequencers, Sequencing Methodology and Sequence Generation. There are several factors in the sequencing process that contribute directly to the analytical validity of the data being produced including the inherent accuracy and error rates of the sequencing chemistry and instrument detection method, inter- and intra- instrument variability, quantity and quality of sequence data produced per sample, mean coverage and evenness of coverage of the genome, alignment and mapping accuracy, and systematic bias or noise in the data. All of these factors contribute to the performance of downstream variant calling.

There are a handful of sequencing platforms on the market today, each with distinguishing features such as read length, error rates, and cost per base.[3,4] However,

once throughput and cost are taken into consideration there are limited sequencing technology options for routine high throughput clinical WGS. The members of this initiative are using both Illumina HiSeq™ and NovaSeq™ platforms for clinical WGS and all sites have either validated or are intending to validate the test on the NovaSeq™ 6000.

After sequencing, manufacturer-supplied software is used for demultiplexing and base calling and data is stored in fastq file format, which includes base quality scores. Base calling quality (expressed as a phred score) is an important measure of confidence of an individual call and is used in downstream variant calling.[5,6] Base quality scores can be influenced by systematic technical errors or biases introduced by the sequencing process itself. Some groups apply base quality score recalibration (BQSR) to apply machine-learning techniques to model and correct for these errors (https://software.broadinstitute.org/gatk/).

Alignment and Variant Calling. Alignment of reads to a reference genome is the most computationally intensive part of clinical WGS analysis, and there are various open source or commercial aligners available that differ in processing speeds. Although relatively standard, the laboratory should examine the effect of aligner not only on the accuracy of SNV and indel calls, but also other classes of genomic variation like SVs and CNVs. Choice of genome build (hg19 versus GRCh38) and alignment method (e.g. use of decoy sequences and alternative loci) can have large effects on the accuracy of more complex variant calls. Similarly, there are many variant callers that have been developed for the detection of specific types of genomic variation including SNVs, multinucleotide variants (MNVs; variants that involve two or more adjacent nucleotide

substitutions), indels, structural variants (SV), copy number variants (CNVs), repeat nucleotide expansions, segments with high homology, and variations in the mitochondrial genome. Evaluation and deployment of several different algorithms may be necessary to achieve the appropriate calling accuracy. Additionally, the reliability and accuracy of variant callers for more complex variants (i.e. SVs) are not yet well established. The majority of participating laboratories in this Initiative are using either BWA/GATK or Illumina DRAGEN™ for genome alignment and small variant calling (SNVs and Indels).

Annotation. Both curated and inferred information is attached to each variant. Different annotations may be used for the major types of allelic variation – SNVs, MNVs, indels, CNVs, repeat expansions, and mitochondrial variants. Information about the quality of the variant call (e.g. the count of the reads in which the variant appears) is carried forward from the variant calling program. A major goal of annotation is to harmonize the variant nomenclature so that variants observed in different individuals can be confidently compared and tabulated. This is relatively simple for small variant calls but becomes more of a challenge when trying to integrate complex genomic variants. The quality of each variant substitution in MNVs is tabulated independently, but the consequence of the adjacent substitutions must be considered together since the prediction for each by itself could be quite different than that inferred from the full sequence.

III. Test Validation

Clinical WGS requires a multi-faceted approach to analytical validation with some aspects that are unique to WGS setup and others that are relevant to analytical validation of laboratory tests in general. Considerations that are not unique to clinical WGS including sequencing bias, repeatability, limits of detection, interference, homology, with points relevant to clinical WGS tests highlighted. In addition, sections on disease specific variant classes validation, and software validation are also discussed below.

Bias. Estimates of analytical performance can be subject to bias and it is important to identify potential sources of error in clinical WGS. Genome reference standards have become more robust and biases associated with their use have become less likely as both deep sequencing on multiple technical platforms with independent chemistries and inheritance information (e.g. through trios) are used to establish truth sets.[7,8] Bias in reference standards can be addressed during validation by judicious use of orthogonal testing methods. Sanger sequencing and CMA are generally used to confirm WGS variant and copy number results, respectively. It is important to note that the error rate for Sanger sequencing may be higher than NGS, and resolution of CMA may be lower than WGS. Sequencing or data biases encountered by the authors included GC and AT rich sequence drop relative to sequencer performance, and systematic biases in variant calling associated with specific aligners and variant callers.

Repeatability and Reproducibility. The validation process should include any known potential technical sources of variation. This could include laboratory staff, flow cell technical replicates, instruments, consumables, and computational pipelines. As part of

the validation process, it is recommended to have different technical scientists prepare libraries and examine within (repeatability) and between run consistency (reproducibility). The laboratory should specify concordance thresholds for repeatability and reproducibility required for diagnostic use prior to validation. The accepted thresholds will vary depending on variant type as one can expect to have much higher concordance for SNVs compared to CNVs where there is greater variability in calling accuracy for breakpoints. For clinical WGS, it is expected that small variant calling should exceed 99% concordance with 90% of the discordance resulting from no-calls and no more than 10% from discordant genotype calls. These studies do not require gold standard reference materials; technical replicates can be used to calculate pair-wise positive agreement or pair-wise negative agreement.

Limit of detection. Germline mosaic and somatic variants relevant to non-neoplastic diseases can be detected with clinical WGS. Given the decreased depth of coverage that is feasible for WGS compared to exome and panel testing, members of this Initiative have noted limited sensitivity for small variants whereas larger mosaic CNVs are easier to detect. Laboratories reporting mosaicism should define the limits of detection for each variant type during validation and this will be tied to coverage targets. While mixing two pure samples at variable percentages is an accepted method to establish mosaicism sensitivity, some patient samples must be used. The limit of detection for clinical WGS should be validated in a defined specimen type such as a blood sample or tissue biopsy.

Interference. The laboratory must identify and document any interfering substances that might occur in routine clinical samples thereby reducing the quality of the overall

sequence. One example is saliva specimens, which contain varying amounts of bacterial genome contamination. High levels of contamination will reduce the number of reads that map to the human reference sequence and will thus reduce usable coverage and as well as the quality and number of variant calls.[9] There are no laboratories in this initiative that currently use saliva as a tissue source for clinical WGS due to this contamination issue. Several are planning to establish performance criteria that will allow for saliva as a valid biospecimen type.

Homology Regions. Clinical WGS produces sequence reads from all parts of the genome. The ability to unambiguously align and map reads to the reference genome is reduced in regions of sequence similarity such as low copy repeats, paralogous genes, pseudogenes and simple repeats. These can be stratified during validation to assess calling accuracy. Regions of homology reduce the sensitivity of clinical WGS by making it difficult or impossible to reliably call variants in some genes. Pseudogenes are particularly problematic as they typically often harbor variants that would be deleterious to function if they occurred in the active copy of a gene. As part of a validation process, a laboratory should identify all clinically relevant areas with these technical issues that are excluded from the WGS test definition and make this list available for upon request.

Disease-Specific Variant Validation: Medically relevant diseases in regions of homology can be assessed and reported. It is now feasible to use clinical WGS data combined with specialized algorithms to provide a definitive genotype for these conditions or signal that there is an abnormality that could be resolved by orthologous testing. Laboratories validating clinical WGS should obtain appropriate samples representing patients with these disorders and demonstrate the ability to detect the specific

abnormality. For laboratories that choose to implement a customized variant caller to address a particular region [e.g. SMN1 and SMN2 associated with Spinal Muscular Atrophy (SMA)], this group recommends a dedicated validation effort employing >50 disease positive samples representing the spectrum of disease-relevant alleles for development and validation. Depending upon the performance of the caller, orthogonal confirmation of findings before reporting may be necessary in some cases.

Software Validation: Validation studies should document all the computational components needed to compile, install, and run the specialized clinical WGS bioinformatics pipelines.[10] Validation is performed on a defined version of the software and data analysis should be repeated each time software components and reference data files are updated. Laboratories should have a clear definition of major versus minor updates that should determine the level of validation. When software components are changed, the testing mechanism must revalidate local test performance and the effect of those changes on the clinical WGS variant calling and annotation process. This is sometimes called “deep testing” to indicate that an effort is made to identify errors or altered test performance elsewhere in the process that may arise when an isolated process is improved. Ideally, software validation should include processes for continuous integration of upgrades and improvements. This may require that software testing is to some degree automated when incremental changes are introduced. It is crucial that new failures are detected efficiently, and the underlying problems identified and addressed.

IV. Quality Management

Quality control and quality assurance are important components of a quality management program. Components relevant to clinical WGS are presented below along with discussion of updating and modification of validation tests.

Alignment and Variant Calling Metrics: After de-multiplexing, it is expected that only a small fraction of reads will fail to be assigned to an expected sample. In alignment, mapping quality scores are assigned and recorded in the output files, which can have a variety of coding and compression formats including SAM, BAM, and CRAM. In variant calling, metrics are associated with each variant as well as global metrics for all variants identified. For example, single base variant call metrics typically include variant call quality score, total depth of coverage at the variant position, variant allele fraction (number and percentage of reads with the variant reported), predicted zygosity, and strand bias for paired-end sequencing. Global variant metrics often captured include percent of different variant types [e.g., heterozygous calls vs homozygous calls, indels vs SNVs, or variant types (i.e. nonsense variants, silent variants, missense variants)], portion and ratios of base substitutions [transition/transversion (Ti/Tv)], percent of novel variants as compared to a standard reference (dbSNP, gnomAD, etc), and concordance rates with reference variant/sequence, as applicable.

Additional quality control metrics also make use of the alignment and variant calling data, in order to detect contamination and to establish sample identity. Genome sequencing can identify instances of contamination from exogenous species (often bacterial or viral species) and patient samples. Contamination thresholds should be set

(typically <1% cutoff) and should be substantially lower than reported minimum detection level used for somatic variant and mosaic variant detection. Because of the numerous steps required in WGS, there is the opportunity for sample or data misidentification between accessioning and reporting. This is especially relevant for the laboratories that do not require orthogonal confirmation of reportable variants. In this instance, the laboratory can identify and utilize markers present in the genome data and compare them to an orthogonal assay to confirm that the report corresponds to the correct individual.

Variant quality scores are generated for each position including those that are apparently homozygous for the reference allele. In general, a laboratory produces a vcf file containing only those positions that differ from reference. A much larger file with scores for all positions called a gvcf can also be generated. Variant quality metrics are central to the process because errors in variant calling will negatively affect annotation, filtering, classification, interpretation and reporting steps. Variant calling metrics are used to estimate the accuracy of a particular genotype call. Through robust validation, the laboratory must establish the acceptable variant quality thresholds, that provide confidence that the variant is a true positive (e.g. variant allele fraction). In addition, the genomic context should be considered since variants within difficult to sequence regions or those with low mapping quality (e.g. repetitive DNA, SINES, segmental duplications) are prone to false positive calls.[11] It is common practice to confirm all variants using an orthogonal method prior to reporting, however as laboratories gain more experience with clinical WGS and collect data regarding variant confirmation rates and test

accuracy, orthogonal confirmation of high confidence variants passing established quality thresholds will no longer be required.[12]

Annotation: Variant annotation names each variant using standardized nomenclature and links it with information from various databases and the medical literature. The Human Genome Variation Society (HGVS) and International Standing Committee on Human Cytogenetic Nomenclature (ISCN) nomenclature guidelines are widely adopted by laboratories, databases, and medical literature, and address the variety of variant types detected by WGS, though guidelines may change over time. Proper annotation of variants is necessary for downstream variant filtering, prioritization, and classification and interpretation. Given the significance of variant annotation, careful evaluation of software and other tools should be carried out to eliminate potential sources of error, such as incorrect application of HGVS/ISCN nomenclature guidelines or incorrect matching of database values to called variants. Currently, there are no consensus standards for variant annotation and therefore no agreed upon acceptance criteria for validation and implementation of variation annotation. Validation of variant annotation tools and annotation sources should focus on evaluating performance such that a uniform and consistent set of annotations are routinely produced by the software and that any differences between expected and observed annotations are systematic, explainable and documented. For clinical WGS, a thorough evaluation of software performance is recommended, as some variant types that may be evaluated and reported as part of a clinical WGS assay, including noncoding variants, CNVs, and other SVs, are not routinely identified by most NGS-based tests. Therefore, these variant types may be at greatest risk for errors with regards to correct application of

HGVS/ISCN nomenclature and ability to correlate variant identity with appropriate databases.

Modification and Updating Clinical WGS tests: Periodic updating or modifying a clinical WGS test is imperative due to the continual development and advancement of all methodological components available for WGS. After initial clinical WGS test validation, the laboratory may need to make changes to the wet bench process and bioinformatics pipeline in order to increase efficiency, improve accuracy, expand reporting of variant classes and ensure that highest quality results are obtained. The procedure for validating and implementing modifications must be described in the laboratories quality management plan. Modifications to the validated clinical WGS test may be minor or a major change. Minor changes such as instrument software updates, updating a version of the tools/algorithm(s) used in the pipeline require running previously run cases/data set, to ensure there are no bugs or inadvertent issues. Major changes such as validating a new instrument (sequencer), additional sample type, library preparation (PCR vs. PCR free), human genome build etc. require recalculating the sensitivity, precision and types of variants detected.

V. SUPPLEMENTARY TABLES

Supplementary Table 1. Variant types detectable and reportable from clinical WGS

Variant Type	Gene(s) (if applicable)	Disorder(s)	References (if applicable)
SNVs and small insertions and deletions[a] (1-50 base pairs)	N/A	Heritable disease	
Copy number variation[a] (deletions and duplications)	N/A	Unbalanced changes including known microdeletion/ duplication syndromes	Gross et al.[13] Stavropoulos et al.[14]
Mitochondrial variation[b] (SNVs, deletions, duplications, heteroplasmy of at least 5%)	N/A	Known mitochondrial disorders	Duan et al.[15]
Structural variants[b]	N/A	Translocations, other genomic rearrangements	
Repeat expansions[c]	*FMR1*	Fragile X and related disorders	Dolzhenko et al.[2]
	HTT	Huntington disease	Dolzhenko et al.[2]
	SCA1	Spinocerebellar ataxia 1	Dolzhenko et al.[2]
	DMPK	Myotonic Dystrophy 1	Dolzhenko et al.[2]
	C9orf72	Amyotrophic lateral sclerosis	
Selected pseudogenes[c]	*SMN1 and SMN2*	Spinal Muscular Atrophy	
	CYP21A2	21-hydroxylase deficiency	
	CYP2D6	Codeine sensitivity	
	HBA1 and HBA2	Alpha thalassemia	
	PMS2	Colorectal cancer	
	PKD1	Polycystic kidney disease 1	

[a]Recommended minimum variant types for clinical validation of WGS
[b]Some initiative groups have clinically validated
[c]Examples of targeted loci that could be validated and reported as part of a clinical WGS test.

Supplementary Table 2: Assessment of Genome Completeness Across Laboratories

Medical Genome Initiative Laboratory	Total Reads[a]	Mean Coverage	Granular 3rd quartile	Granular median	Granular 1st quartile	% bases above 15
1	1,746,955,926	49.76	57	51	45	97.7
2	1,748,355,695	49.8	57	51	45	97.7
3	1,720,118,385	49	56	50	44	98.1

[a]Target is all coding exons +/-2bp as the region of interest with coverage set a variant calling cutoffs (MQ>17, BQ>10) (available by request)

Supplemental Table 3: Reference Standards and Positive Control Resource

		Variant Type								
		SNVs & Indels	CNVs (>10Kb)	SVs	Mitochondrial variants	Pseudogenes	REs	Somatic/ mosaic changes	Literature/Data	Source
Reference Samples	NA12878	100%[a]	40%	0	0	0	0	0	Zook et al[16] FTP Directory	NIST Reference Materials Link
	Other NIST standard (e.g. AJ/Asian trios)	71%	40%	50%	0	0	0	0	Zook et al[16] FTP Directory	NIST Reference Materials Link
	Platinum Genomes	29%	0	0	0	0	0	0	Eberle et al[8]	Platinum Genomes Link
	Venter/HuRef	14%	40%	0	0	0	0	0	Trost et al[1]	HuRef Link
Positive Controls	Disease specific positive controls[b]	86%	80%	50%	100%	100%	100%	50%	GeT-RM Link	GeT-RM Link
	Synthetic controls	0	0	0	33%	0	0	50%	Deveson et al[17]	Sequins Standards Link
	In silico data	0	20%	0	0	0	0	0	Escalona et al;[18] Duncavage et al[19]	
	No. positive control samples	10—85	7—40	>10	4—20	4—40	18—175	N/A		

[a]Percentage of Initiative sites using the specific reference or positive control sample per variant type
[b]Includes clinical sequencing/CMA controls and Coriell lines
SNV: single nucleotide variant; CNV: copy number variant; SV: structural variant; RE: repeat expansion

Supplementary Table 4: Examples of metrics for monitoring clinical WGS test performance

Metric	Description	Typical Value
%Q30 bases total	The percentage of bases that meet Q30 scores	~90%
15X%[a]	The fraction of autosome bases that attained at least 15X sequence coverage in post-filtering bases	96-99%
PF Reads Aligned %	The percentage of PF reads (Passing Filter) that align to the reference sequence.	>98%
PF Aligned Q20 Bases[b]	The number of bases aligned to the reference sequence in reads that were mapped at high quality and where the base call quality was Q20 or higher	>1.0E+11
Adapter %	The fraction of PF reads that are unaligned and match to a known adapter sequence right from the start of the read.	~0.2%
Chimera %	The percentage of reads that map outside of a maximum insert size (usually 100kb) or that have the two ends mapping to different chromosomes.	~1%
Duplication %	The percentage of mapped sequence that is marked as duplicate.	~5%
Median Insert Size[c]	The median insert size of all paired end reads where both ends mapped to the same chromosome	>300bp
Excluded Total %	The percentage of aligned bases excluded due to all filters.	~15%

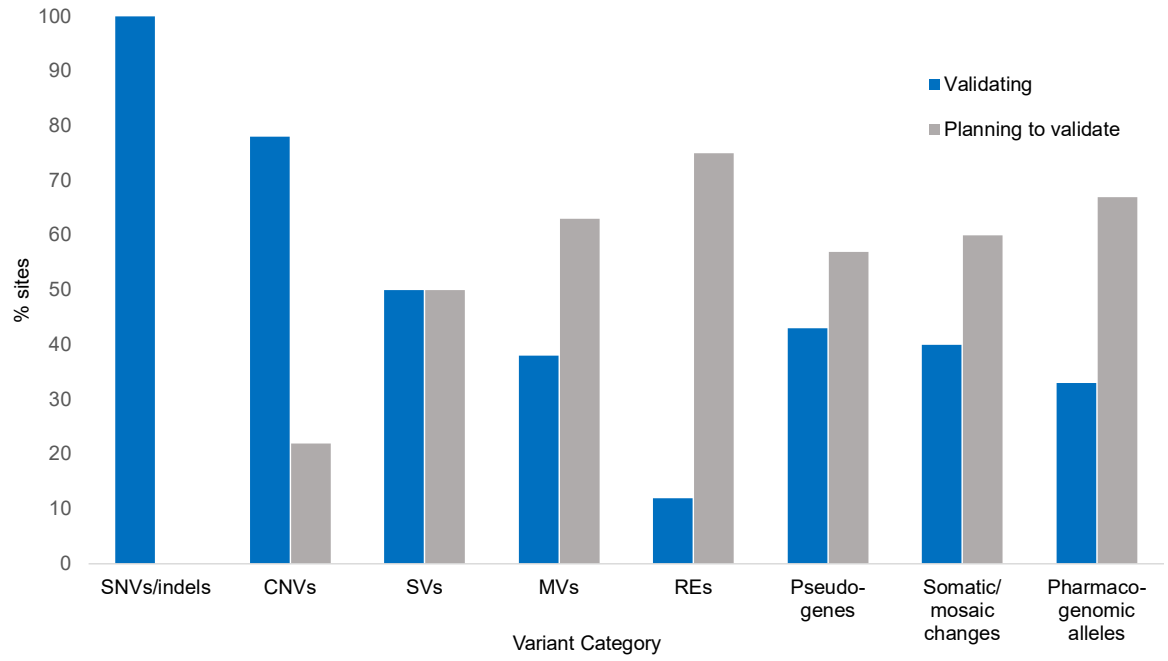
[a]Measure of completeness. Depth of coverage at 20X also used by some groups along with a mapping quality cutoff (>10). Targets will also vary; one can measure across genome, exome, OMIM morbid map genes, positions or exons with known pathogenic regions.

[b]Some groups use Q10 Bases

[c]Also could use mean insert size. The mean insert size of the "core" of the distribution. Artefactual outliers in the distribution often cause calculation of nonsensical mean and stdev values. To avoid this the distribution is first trimmed to a "core" distribution of +/- N median absolute deviations around the median insert size. By default N=10, but this is configurable.

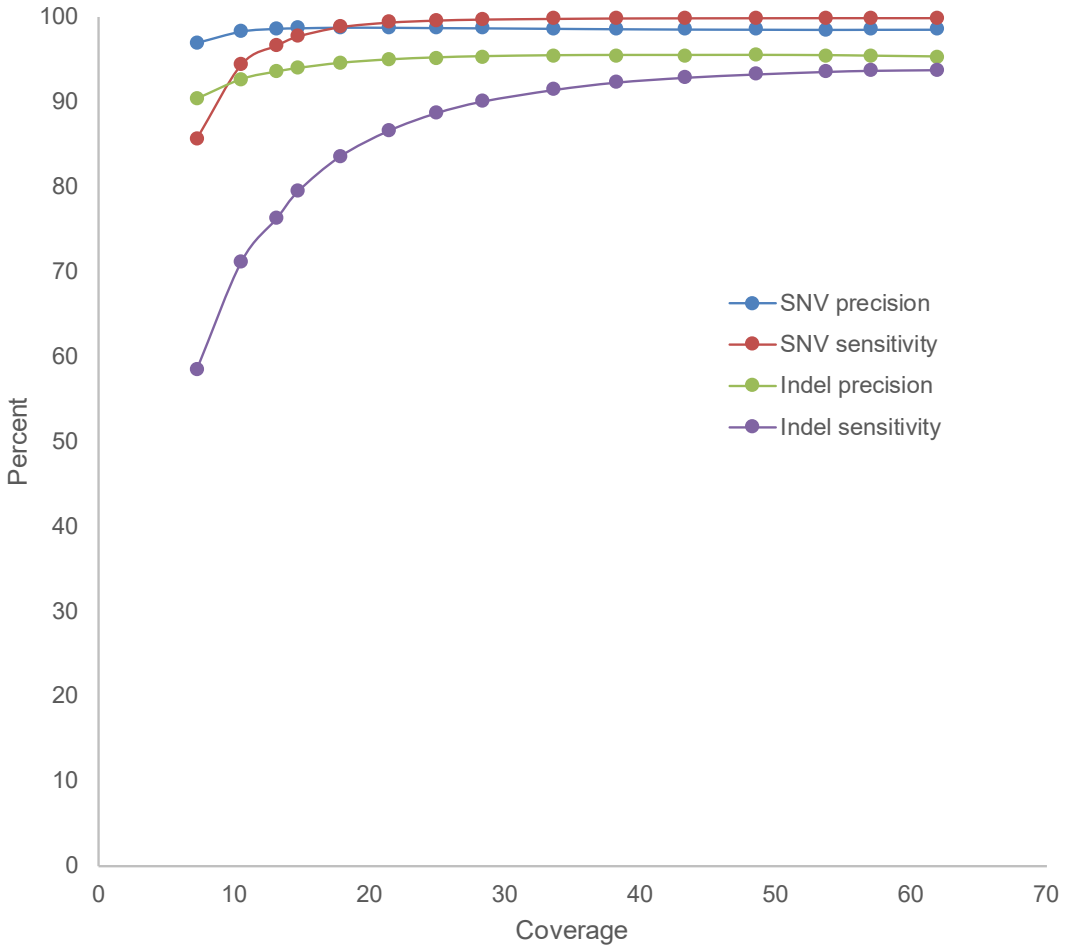
VI. SUPPLEMENTARY FIGURES

Supplementary Figure 1: Analytical Validation of Variant Categories Across Medical Genome Initiative Sites.



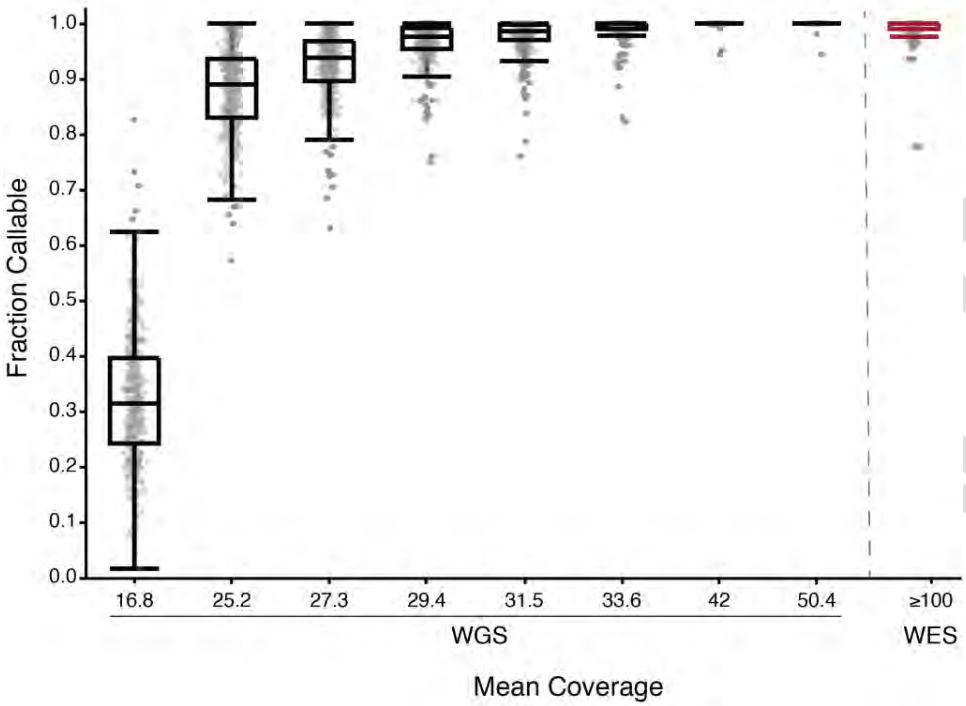
The percentage of Initiative sites that have either validated (blue) or plan to validate (grey) across the spectrum of variant types detectable by clinical whole-genome sequencing. SNV: single nucleotide variant; CNV: copy number variant; SV: structural variant; RE: repeat expansion

Supplementary Figure 2: Coverage titration for SNVs and Indel Accuracy



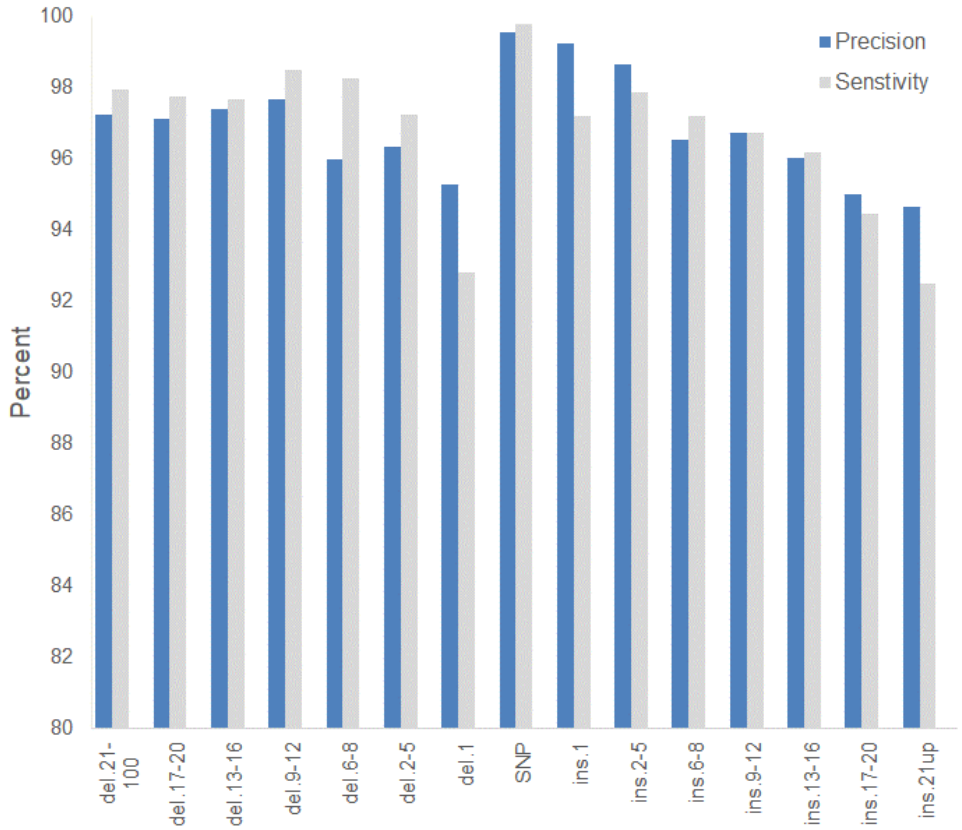
Representative coverage titration measuring SNV and indel precision using the Consensus Coding Sequence (CCDS) and the NA12878 Reference Standard truth set. Note that precision is equivalent to technical positive predictive value (TPPV).

Supplementary Figure 3: Fraction of callable bases in WGS versus WES.



Fraction of callable bases in the ACMG secondary finding 56 genes at various whole-genome sequencing (WGS) mean coverages. A base within in the interval is considered 'callable' if it is covered at ≥20X, has a mapping quality ≥20, and a base quality of ≥20. The fraction of the interval that is callable is equivalent between WGS and whole-exome sequencing (WES) once genome mean coverage reaches >30X.

Supplementary Figure 4: Accuracy of indel calling across size ranges.



Representative data depicting the accuracy of indel calling across size ranges using the NA12878 Reference Standard truth set.

VII. SUPPLEMENTARY REFERENCES

1. Trost B, Walker S, Wang Z, et al. A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. *Am J Hum Genet.* 2018;102(1):142-155.
2. Dolzhenko E, van Vugt J, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* 2017;27(11):1895-1903.
3. Mardis ER. DNA sequencing technologies: 2006-2016. *Nat Protoc.* 2017;12(2):213-218.
4. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* 2016;17(6):333-351.
5. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 1998;8(3):175-185.
6. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 1998;8(3):186-194.
7. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data.* 2016;3:160025.
8. Eberle MA, Fritzilas E, Krusche P, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* 2017;27(1):157-164.
9. Mandelker D, Schmidt RJ, Ankala A, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet Med.* 2016;18(12):1282-1289.
10. Roy S, Coldren C, Karunamurthy A, 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. *J Mol Diagn.* 2018;20(1):4-27.
11. Goldfeder RL, Priest JR, Zook JM, et al. Medical implications of technical accuracy in genome sequencing. *Genome Med.* 2016;8(1):24.
12. Lincoln SE, Truty R, Lin CF, et al. A Rigorous Interlaboratory Examination of the Need to Confirm Next-Generation Sequencing-Detected Variants with an Orthogonal Method in Clinical Genetic Testing. *J Mol Diagn.* 2019;21(2):318-329.
13. Gross AM, Ajay SS, Rajan V, et al. Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genet Med.* 2018.
14. 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.
15. Duan M, Chen L, Ge Q, et al. Evaluating heteroplasmic variations of the mitochondrial genome from whole genome sequencing data. *Gene.* 2019;699:145-154.
16. Zook JM, McDaniel J, Olson ND, et al. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol.* 2019;37(5):561-566.

17. Deveson IW, Chen WY, Wong T, et al. Representing genetic variation with synthetic DNA standards. *Nat Methods.* 2016;13(9):784-791.
18. Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat Rev Genet.* 2016;17(8):459-469.
19. Duncavage EJ, Abel HJ, Merker JD, et al. A Model Study of In Silico Proficiency Testing for Clinical Next-Generation Sequencing. *Arch Pathol Lab Med.* 2016;140(10):1085-1091.