# Test Method Evaluation (TME) Summary for MDL

**Test Name: Variant Calling for Next-Generation Sequencing Assays**

**Section 1:**  **Purpose, Method & Utility** (determined prior to validation through test development)

**1.1 Purpose**

**1.11 Changes to Existing Test Only**

**Intended Change**

Outline the intended change to the test and its impact on test performance.

**Existing SOPs related to test** (for existing tests only)

Reference Strategy and MGAP

**1.12 New Tests Only**

**Intended Use or Application**

Next-Generation Sequencing (NGS) alignments/mapping are used as the main input to determine possible DNA variations [single or multiple (~25bp) nucleotides] in all regions covered by the different test assays. The custom pipeline uses several third-party developed software for calling variants for all samples in one NGS run, which are them compared, merged, saved and displayed in a table for easy assessment.

**Locus/Gene/Marker**

Refer to the specific gene panels for the genes, regions, exons (introns) and locations covered

**Reference Sequence**

Refer to the SOPs for the gene panels for the reference sequence transcript ID(s)

**1.2 Method**

**1.21 Methodology Outline**

NGS reads are aligned/mapped against a reference genome, currently GRCh37, resulting in binary files (BAM) with indexed nucleotides. Each location can have more or less coverage (number of reads under the nucleotide/region/exon) with varying basecalling quality for every single nucleotide in every single read outputted by the sequencer. Variant callers scan BAM files using BED coordinates that define regions for analysis, and by using mathematical and statistical approaches determine if the nucleotide composition differences in a column under a nucleotide and the reference genome in determined location(s) is an actual variant or not.

Each variant caller uses a different approach for the determination of DNA variants (SNPs, indels, etc) but mostly basecalling quality of the location in question and neighbouring nucleotides, frequency of changes in the sequenced reads, excess coverage among other factors.

Published and online references to each variant caller used in the pipeline are in the folder included with this file.

The pipeline provides an automated processing where many required steps are performed, manipulating input and output files to adjust aspects of the data to maximize the statistical power of variant callers. At the same time, the code generates supporting data for quality control and/or assurance of the data generated in each step.

**1.22 Existing SOPs related to methodology**

Inherited Cancerplus Rapid Capture Enrichment Assay (Disorder SOP)

Nextera Rapid Capture Enrichment

**1.23 References**

List any relevant documentation (internal and external) indicating the derivation of the test.

PDFs and other documents included with this file.

**1.3 Utility**

**1.3.1 Applicability (aka Utility)/Limitations (**Factors determined through test development)

This method is designed to detect genomic variants in any short-read next-generation sequencing experiment, such as single nucleotide polymorphisms and short deletions and insertions (indels). Low frequency variants of any type might be missed as the variant callers employed have a detection limited to 20-25% frequency for heterozygote changes. They are mainly used for hereditary variants and not for somatic-base samples.

Regions of low coverage depth can also influence the results, as GC content and other DNA structures repeats/SINE/LINE, etc).

**1.3.2 Scope**

Samples are analysed independently from batch data obtained (FASTQ) from next-generation sequencers and converted and manipulated into formats (BAM, SAM, etc) that allow analysis by variant callers and other software.

Different extraction methods and individual sample issues should not interfere on results for other samples in the same batch, except for the Copy Number Variation methodology (covered in another TME and SOP).

Performance of the algorithms can be related to DNA quality, extraction methods, basecalling of sequence data, indexing amount, but QA/QC data can provide clues on what and where the issue occurred.

**1.3.3 Selectivity** (how well it can distinguish target signal from other components)

Variant determination (SNPs, indels) can suffer with poor nucleotide coverage and base quality, and these can be mitigated with changes in the caller parameters to improve sensitivity of the algorithm. Other aspects that can improve overall variant calling is to provide well established limits of detection to allow the technologist to determine if a variant is properly called or not based on statistical knowledge obtained with some other procedures, such as downsampling or artificially simulated variants with different levels of quality, frequency, location, etc.

**1.3.4 Interferences** (presence of a substance that would cause the reaction to fail)

Regular issues related to DNA quality, NGS library preparation, NGS indexing, among others can interfere with variant calling.

**1.3.5 Cross Reactivity** (contamination)

At the pipeline level, DNA barcodes are generated from the NGS read alignments for each sample, and are compared with other test results to determine the uniqueness of the sample to prevent analysis of mixed samples. One caveat is that these values are available after the pipeline is done and sometimes comparison to other tests are delayed due to their schedule.

**1.3.6 Other Considerations**

Factors that may affect the performance of this test include:

* Poor quality DNA. As most of our DNA is extracted from blood, the DNA quality from patients currently undergoing radiation treatment or chemotherapy may be affected.
* DNA extracted by other laboratories with a different non-MDL validated extraction type may contain inhibitors or substances that alter the rate of tagmentation or affect probe capture, resulting in lower reads and coverage than usual.
* Patients with leukemia will likely have many detectable CNVs if tested on DNA from blood. It may be advisable to run both the sample extracted from blood, and a sample from another tissue type (eg. buccal) to do a comparison to see if the CNVs are somatic in origin.
  + Poor equalization of library DNA prior to pooling. Precise quantification of library DNA prior to pooling is very important, as a sample with significantly less reads than others may have especially low read depth in known problem regions, and samples with significantly more reads than others will use up the space on the flow cell, resulting in lower than expected reads for the rest of the samples.
  + Low cluster density due to inaccurate DNA pool quantification will result in global reduction of reads. Low read depth regions will be affected first and will show reduced accuracy of SNP detection.

**1.3.7 Test Development**

The pipeline is agnostic to what type of NGS data being analysed, and it was initially developed and tested with Platinum genome samples ran in the laboratory. Code development follows most steps of the GATK Best practices with modifications for local usage (<https://gatk.broadinstitute.org/hc/en-us/articles/360035894711-About-the-GATK-Best-Practices>).

Each module of the pipeline was tested individually in order to ascertain that the output generated was compatible with the next step and the output from a previous step was suitable for use. New modules added after initial development were tested in the same fashion.

All third-party software used in the pipeline was checked for compatibility in our systems, at the same time that the best set of parameters for each file were optimized and benchmarked manually before being implemented in the pipeline.

Platinum genomes have a set of expertly curated small variants (SNPs, indels) available at the Genome in a Bottle webpage and FTP site (<https://www.nist.gov/programs-projects/genome-bottle>) and the resulting VCFs were used in the initial testing phase of the pipeline. But our samples cover a small subset of the whole genome and whole exome covered by the GiaB consortium and even though the results were promising, there was need for a more thorough analysis of samples that could provide a better assessment of the pipeline outputs, mainly VCF files which are one of the most important outputs, apart from CNV data (related TME/SOP).

In order to properly test, benchmark and validate the results, we used a dataset of 142 publicly available whole exome NGS samples, with curated variants, initially obtained with short-read sequencing and then verified using specific Sanger sequencing techniques (<https://ega-archive.org/studies/EGAS00001001332>). This dataset also provides a list of result files for each sample and a software package that is capable of comparing locally obtained results to the their true set, while calculating sensitivity and specificity values for each software evaluated.

**Section 2:** **Validation &** **Criteria**

Choose one of the following:

**■ Overall aims of validation or verification study:**

**Table

Description automatically generated**

**□ Overall aim of verification study:**

* To determine the overall accuracy of the test

Based on test type, indicate the critical parameters to be met during validation (refer to Tables 1 and 2))

**■** Sensitivity **□** Trueness **□** Reproducibility **□** Limit of quantification – set 50ng/uL

**■** Specificity **□** Repeatability **□** Robustness **□** Linearity

**■** Accuracy **□** Intermediate precision **□** Limit of detection **□** Measurement uncertainty

Define the levels of performance that must be obtained for each parameter checked:

**Accuracy/Sensitivity/Specificity**:

> 95% for each software independently for single nucleotide polymorphisms.

> 90% for each software independently for small insertions and deletions

> 95% overall for each software independently

When combined the 4 or 5 callers should have an accuracy of over 98%, when using a rule of thumb of at least two concordant callers

**Section 3:** **Work Plan/ Study Design and Data Collection**

**3.1 Samples:** (known positives, negatives, DNA from various matrices, reference samples)

As most of the software that provide variant detection used in the laboratory are third-party and open source, the validation can be performed with any type of sample, either extracted and analyzed in house or obtained elsewhere, as these software should be capable of determining variants in any situation. Also, samples and data already clinically validated and available in international data repositories can be an asset due to the sample size, known detected variants and in some cases validation with other sequencing technologies.

The European Genome-Phenome Archive (EGA) provides a collection of 142 NGS exome samples that were used to evaluate the sensitivity and specificity of variation detection in NGS exome samples, which in turn were validated with traditional Sanger-style sequencing (<https://ega-archive.org/studies/EGAS00001001332>).

From the study publication abstract (https://europepmc.org/article/MED/27158454):

“*To provide a useful community resource for orthogonal assessment of NGS analysis software, we present the ICR142 NGS validation series. The dataset includes high-quality exome sequence data from 142 samples together with Sanger sequence data at 704 sites; 416 sites with variants and 288 sites at which variants were called by an NGS analysis tool, but no variant is present in the corresponding Sanger sequence. The dataset includes 293 indel variants and 247 negative indel sites, and thus the ICR142 validation dataset is of particular utility in evaluating indel calling performance. The FASTQ files and Sanger sequence results can be accessed in the European Genome-phenome Archive under the accession number EGAS00001001332.*”

The same group also provides an analysis package to assess sensitivity and specificity from the results automatically.

**3.2 Study Design**

The pipeline would be run using a reduced workload, where metric files are not generated, to improve time performance, and each VCF file generated by a caller will be used to determine the level of specificity and sensitivity. A report using the package provided by the study’s authors (<https://europepmc.org/article/MED/30483600>) for each caller used will be generated and provided as evidence for validation procedures.

The final reports are in Word format and provide a quick assessment of the values obtained for each software .

**Section 4: Director Approval** Submit TME Folder for approval to proceed.

Investigating Scientist: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Director Approval to proceed: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Test Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Section 5: Results** **and Data Analysis** Attach all records of results obtained.

**5.1 Results Summary**

We downloaded the FASTQs of the 142 whole-exome samples from the EGAS00001001332 study, and ran the pipeline code using four different applications for variant calling

GATK v3.5-0-g36282e4

GATK v4.2.4.0

Octopus v0.7.4

Freebayes v1.2.0-2-g29c4002

A VCF file was generated for each sample and each variant caller, and the results were analyzed by the ICR142\_Benchmarker-1.0.2. All outputs are available in this TME folder.

**5.2 Overall Conclusions**

State explicitly if requirements in Section 2 have been met and give any other relevant conclusions.

The results of this validation meet the requirements as set out in Section 2. Except for Freebayes’ sensitivity, all three variant callers reached of exceeded the values defined in Section 2. This also falls into the at least two callers are required to assign a variant that is used in the pipeline, when all results are combined.

**Table: ICR142 specificity values**

|  |  |  |  |
| --- | --- | --- | --- |
| Caller | Overall | SNPs | Indels |
| GATK v3 | **94%** | **95%** | **94%** |
| GATK V4 | **93%** | **90%** | **93%** |
| Freebayes | **93%** | **93%** | **95%** |
| Octopus | **93%** | **90%** | **94%** |

**Table: ICR142 sensitivity values**

|  |  |  |  |
| --- | --- | --- | --- |
| Caller | Overall | SNPs | Indels |
| GATK v3 | **96%** | **98%** | **95%** |
| GATK V4 | **97%** | **99%** | **97%** |
| Freebayes | **76%** | **71%** | **89%** |
| Octopus | **95%** | **98%** | **94%** |

**5.3 Estimates of Accuracy and Measurement of Uncertainty**

The results are available in this TME folder and were outputted by the ICR142\_Benchmarker software.

**5.4 Sources of uncertainty:**

We used externally generated samples to validate the software, and there is not feasible way to assess uncertainty for in house samples.

Please refer to TME for CNV detection in Autosomal Genes using NGS read analysis for a related assessment.

**Section 6:** **Quality Control & Proficiency Testing**

**6.1 Quality Control** Details of QC Samples and Procedures

.

**6.2 Proficiency Testing Schemes** Details of external quality assurance measures

The lab will participate in EMQN proficiency testing at least once per year.

**Section 7:** **Method Authorization**

State whether the results fulfill the validation criteria. List any specific derived limitations to reproduce the outcome.

**This test method is fit for its intended use in the *ACH Molecular Diagnostic Laboratory* with the following limitations:**

Comment on the potential influence of the uncertainty on the reliability of the result.

Laboratory Director/Geneticist Authorization: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_