CIDC 2.0 Bioinformatic Pipelines
WES Validation

Essex Management

National Cancer Institute (NCI)

06/14/2024

Version 1.0

Draft

Submitted To:
Daoud Meerzaman
Computational Genomics and Bioinformatics Branch (CGBB)
National Cancer Institute
Center for Biomedical Informatics & Information Technology
Rockville, MD 20850

Submitted by:
Nick Renzette, Jennifer Harvey
Essex Management, LLC
11140 Rockville Pike | Suite 332
Rockville, MD 20852-3149
DUNS: 829872345
Cage Code: 5CYC9

1. Introduction 3

2. WES Pipeline – Validation Dataset 6

3. WES Pipeline – Validation Method 11

4. WES Pipeline – Validation Results 12

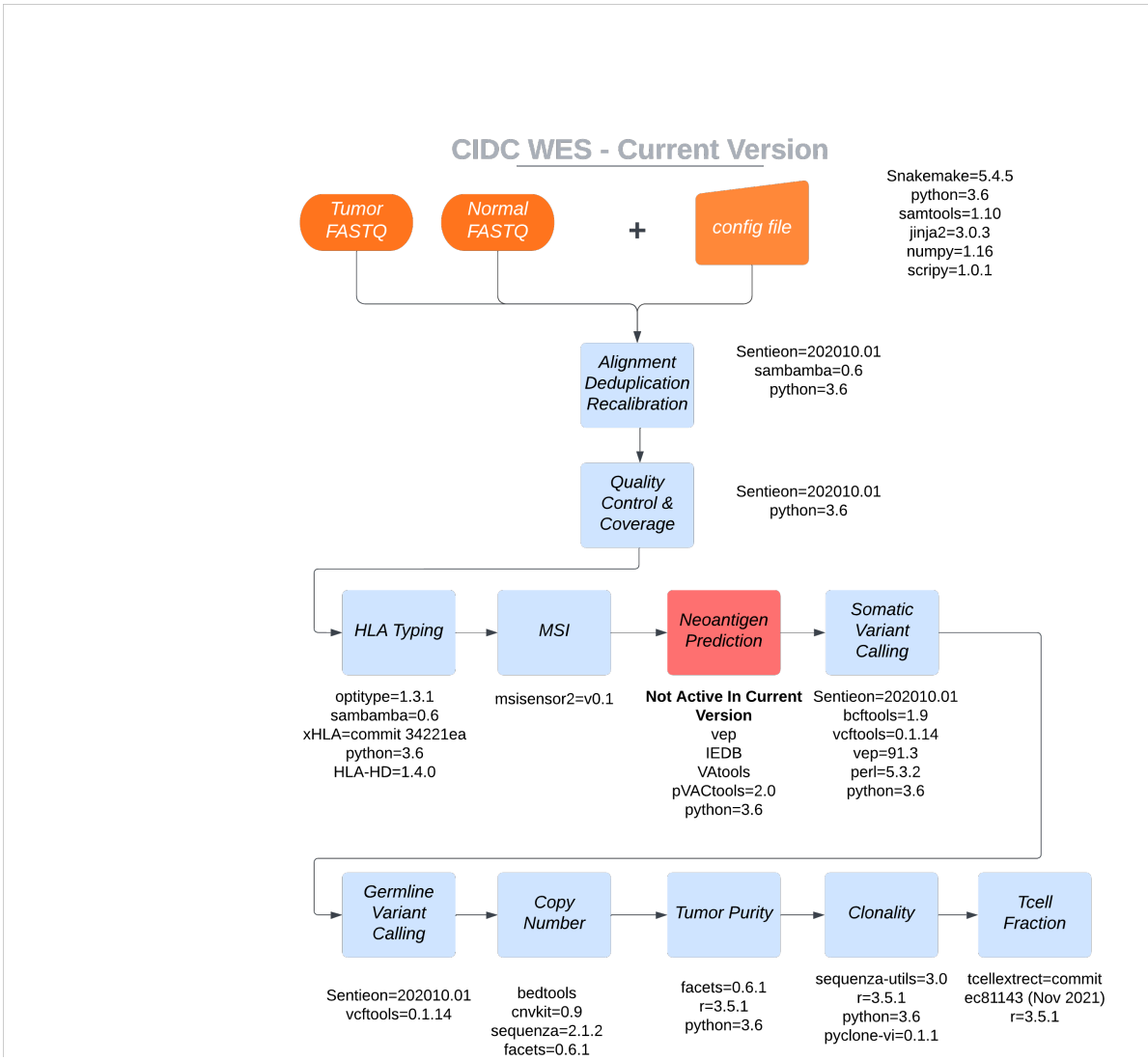
5. WES Pipeline – Appendix I – Tabular Results 17

1. INTRODUCTION

As part of the planned CIDC enhancements after the migration to the National Cancer Institute (NCI), the bioinformatic pipelines were reviewed by members the National Cancer Institute Computational Genomics and Bioinformatics Branch (NCI-CGBB) and Essex Management (EM). These reviews were carried out to determine which, if any, changes could be made to the pipeline to satisfy the following goals and objectives:

Goals	Objectives
Update and clean up code so that it is easy to read and understand for everyone working in the same code base, thus, making it easier to maintain, debug, and update.	Provide code review to determine areas for clean up and refactoring
Maintain industry standard software to optimize the best combination of biochemistry, mathematics, computer science, data science, and modern data analytics tools.	Review the performance of current packages to determine if current functionality is meeting customer needs (CIMAC Input)
Maintain current software versioning to optimize vendor support and application performance.	Provide gap analysis on current software versions to determine what tools to upgrade
Provide enhancements to pipeline to improve current functionality/performance and better support the analysis of DNA, while maintaining backward compatibility with previous versions.	Add new functionality and features as desired by the stakeholder community

The reviews created schematics of the extant pipelines stood up after the migration from Dana Farber Cancer Institute (DFCI) to NIH STRIDES. The **Whole Exome Sequencing (WES)** pipeline was divided into 11 distinct modules:

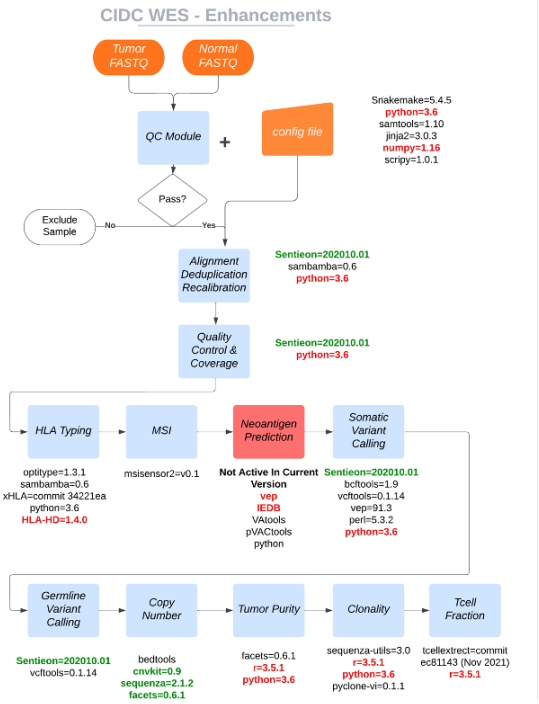


Note that in the original pipeline, the neoantigen prediction module was present but not run.

After thorough review, the following enhancements were planned for the WES Pipeline:

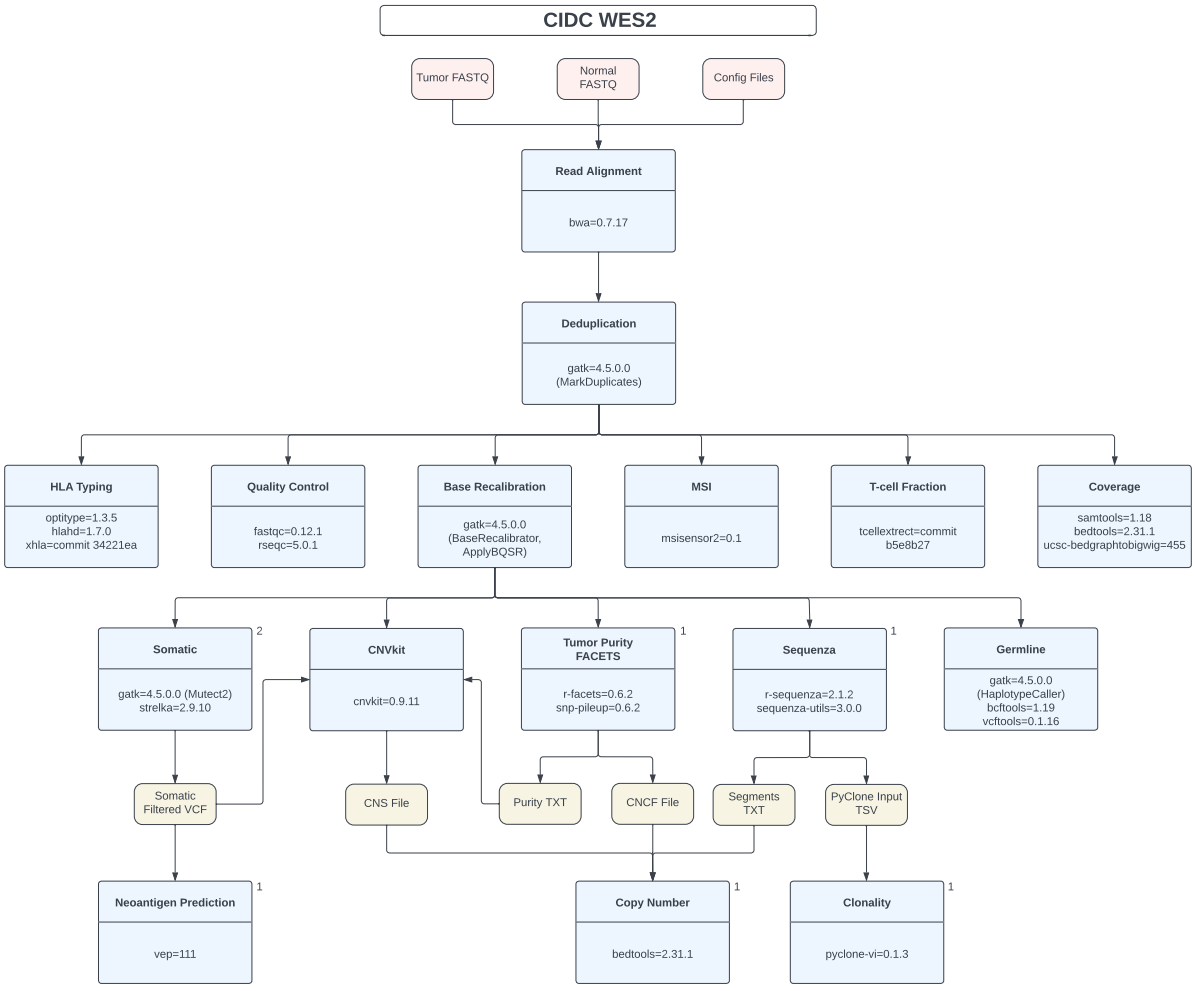
CIDC WES – Recommended Enhancements

WES Pipeline Recommendations	Reason
Add Neoantigen Functionality	Increased utility
Upgrade numpy beyond 1.19	CBIIT identified security vulnerabilities (High - null pointer; critical - Numpy Deserialization of Untrusted Data)
Upgrade scipy to 1.10.0	CBIIT identified security vulnerabilities (Moderate - memory leak)
Upgrade Variant Effect Predictor (VEP) to 109.3 (May 2023)	Improved annotation based on the most recent Gencode gene models (Gencode v35).
Upgrade HLA-HD to 1.7.0 released Feb 2023	Reduced Memory Usage: Version 1.7.0 consumes approximately 75% less memory compared to version 1.6.1. We are on 1.4.0
Upgrade r	Current version unsupported; update past v3.6
Upgrade python	Current version unsupported; update past v3.7
Add file validation (QC Module)	To identify corrupt or incorrectly formatted files
Evaluate use of Sentieon	8 releases have updated speed and accuracy. **CAVEAT: Difference in SNV/Indel Output**
Evaluate CNV Callers	Known source of variability



The specific changes associated with the pipeline are enumerated in the figures above. For each module, code refactoring was also performed to increase readability and ease future maintenance and/or upgrade efforts.

Final Design



[1] Tumor-only samples are not run in the module.
[2] Tumor-only samples are run with Mutect2 but not with Strelka in the Somatic Module.

2. WES Pipeline – Validation Dataset

Validation Dataset:

For validation, it would be possible to compare the output of the enhanced pipeline to the original pipeline. While this method would be suitable to determine result continuity associated with the transition, it would not be suitable for determining result accuracy. The aligner and variant caller (Sentieon 202010.01) will be updated to open source components in the enhanced pipeline, which will likely lead to vastly differing results (https://support.sentieon.com/versions/202308.01/manual/appendix/releasenotes/ and https://github.com/lh3/bwa/releases). Thus, to verify the accuracy of the enhanced pipeline, we will use truth sets obtained from external sources.

The source fastq files for small variant calling will be collected from NIST's Genome in a Bottle (GIAB) project, a consortium dedicated to authoritative characterization of the human genome. As such, they distribute high-quality sequencing data and benchmarking datasets for 7 samples (1 HapMap genome [NA12878], and two son/father/mother trios of Ashkenazi Jewish and Han Chinese ancestry).
For CNV benchmarking, data from the triple negative breast cancer line HCC1395 that has been extensively characterized will be used. These data have previously been used in a comprehensive CNV caller benchmarking study[1]. The total set is as follows:

Sample	Description	Result Evaluated	Source – Sequencing Reads	Source – Benchmarking Data
NA12878_HG001	HapMap sample	Alignment / Small Variants	https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/Nebraska_NA12878_HG001_TruSeq_Exome/NIST-hg001-7001-ready.bam	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/NA12878_HG001/latest/GRCh38/
HG002_NA24385_son	Ashkenazi Jewish trio – son	Alignment / Small Variants	https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG002-EEogPU_v02-KIT-Av5_AGATGTAC_L008.posiSrt.markDup.bam	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/latest/
HG003_NA24149_father	Ashkenazi Jewish trio – father	Alignment / Small Variants	https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG003-EEogPU_v02-KIT-Av5_TCTTCACA_L008.posiSrt.markDup.bam	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG003_NA24149_father/latest/
HG004_NA24143_mother	Ashkenazi Jewish	Alignment / Small Variants	https://ftp.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSample

[1] Masood et al. (2024) Under Review

	trio – mother – replicate 1		HG004_NA24143_mother/OsloUniversityHospital_Exome/151002_7001448_0359_AC7F6GANXX_Sample_HG004-EEogPU_v02-KIT-Av5_CCGAAGTA_L008.posiSrt.markDup.bam	s/giab/release/AshkenazimTrio/HG004_NA24143_mother/latest/
WES_EA_T1	HCC1395 Replicate 1 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_EA_T_1.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_EA_N_1.bwa.dedup.bam	https://zenodo.org/records/10081574
WES_FD_T1	HCC1395 Replicate 2 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_FD_T_1.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_FD_N_1.bwa.dedup.bam	https://zenodo.org/records/10081574
WES_FD_T2	HCC1395 Replicate 3 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_FD_T_2.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/So	https://zenodo.org/records/10081574

			matic_Mutation_WG/data/WES/WES_FD_N_2.bwa.dedup.bam	
WES_FD_T3	HCC1395 Replicate 4 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_FD_T_3.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_FD_N_3.bwa.dedup.bam	https://zenodo.org/records/10081574
WES_IL_T1	HCC1395 Replicate 5 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_IL_T_1.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_IL_N_1.bwa.dedup.bam	https://zenodo.org/records/10081574
WES_IL_T2	HCC1395 Replicate 6 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_IL_T_2.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_IL_N_2.bwa.dedup.bam	https://zenodo.org/records/10081574
WES_IL_T3	HCC1395	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/R	https://zenodo.org/records/10081574

	Replicate 7 – WES dataset		eferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_IL_T_3.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_IL_N_3.bwa.dedup.bam	
WES_LL_T1	HCC1395 Replicate 8 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_LL_T_1.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_LL_N_1.bwa.dedup.bam	https://zenodo.org/records/10081574
WES_NC_T1	HCC1395 Replicate 9 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_NC_T_1.bwa.dedup.bam https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_NC_N_1.bwa.dedup.bam	https://zenodo.org/records/10081574
WES_NV_T1	HCC1395 Replicate 10 – WES dataset	CNV Calling	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_NV_T_1.bwa.dedup.bam	https://zenodo.org/records/10081574

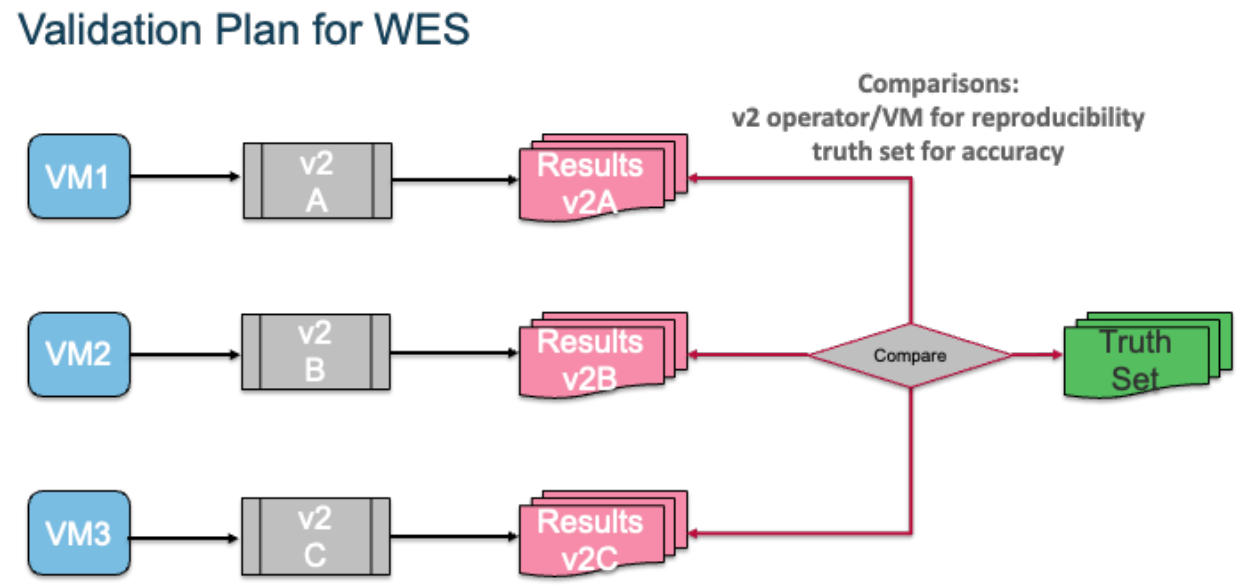
			https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG/data/WES/WES_NV_N_1.bwa.dedup.bam	

Additional details about the GIAB data and analyses:
Source of hg38 reference file used by GIAB:
https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/references/

GIAB recommended analysis tools for small variant benchmarking:
https://github.com/Illumina/hap.py
https://github.com/ga4gh/benchmarking-tools/blob/master/resources/high-confidence-sets/giab.md

3. WES Pipeline – Validation Method

Validation Design: The validation design is shown below. It will be used to compare the results of the enhanced pipeline (v2) to industry-accepted truth sets as well as assess inter-operator and inter-VM variability.



Acceptance Criteria:

Small Variants

Recall: > 0.90

Precision: > 0.90

CNV

Sensitivity: > 0.8

Specificity: > 0.8

Note: The GIAB BED file defines the high-confidence regions for scoring small variants. Other regions will be excluded from the analysis.

Investigation of Discrepancies:

If the comparison metrics do not meet the acceptance criteria, an investigation will be carried out by members of NCI-CGBB and EM. The members to perform the investigation will be designated by Daoud Meerzaman, based on expertise and availability. The investigation should last no longer than 10 business days, at which time a report that outlines the problem and suggests solutions will be presented to Daoud Meerzaman.

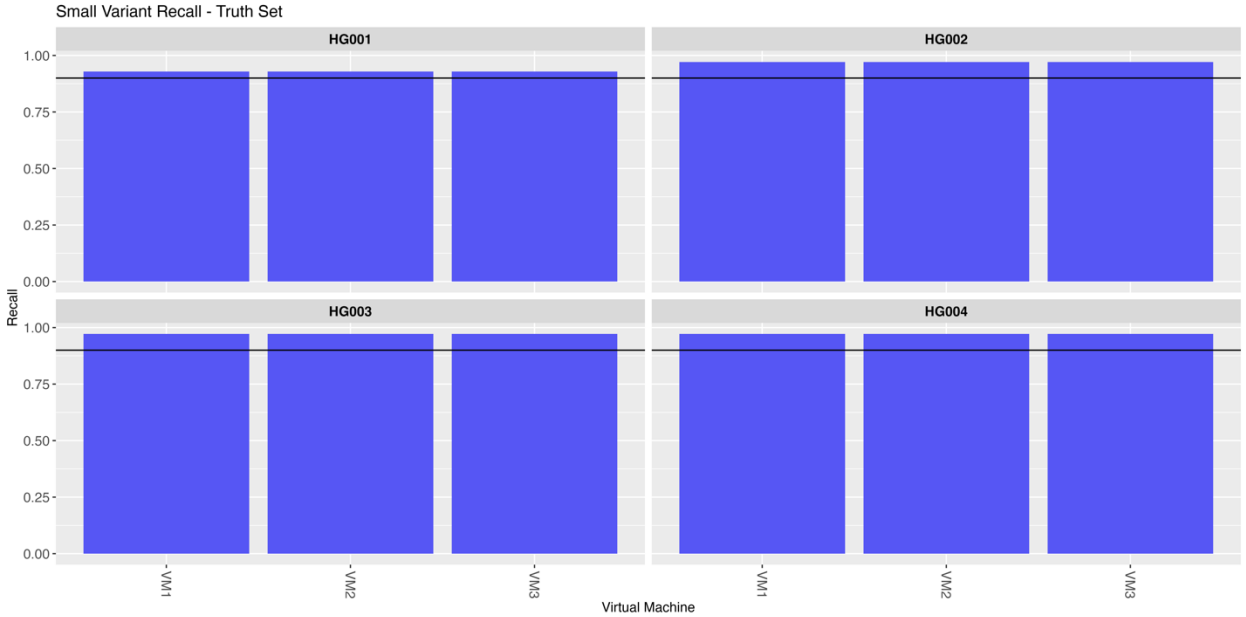
4. WES Pipeline – Validation Results

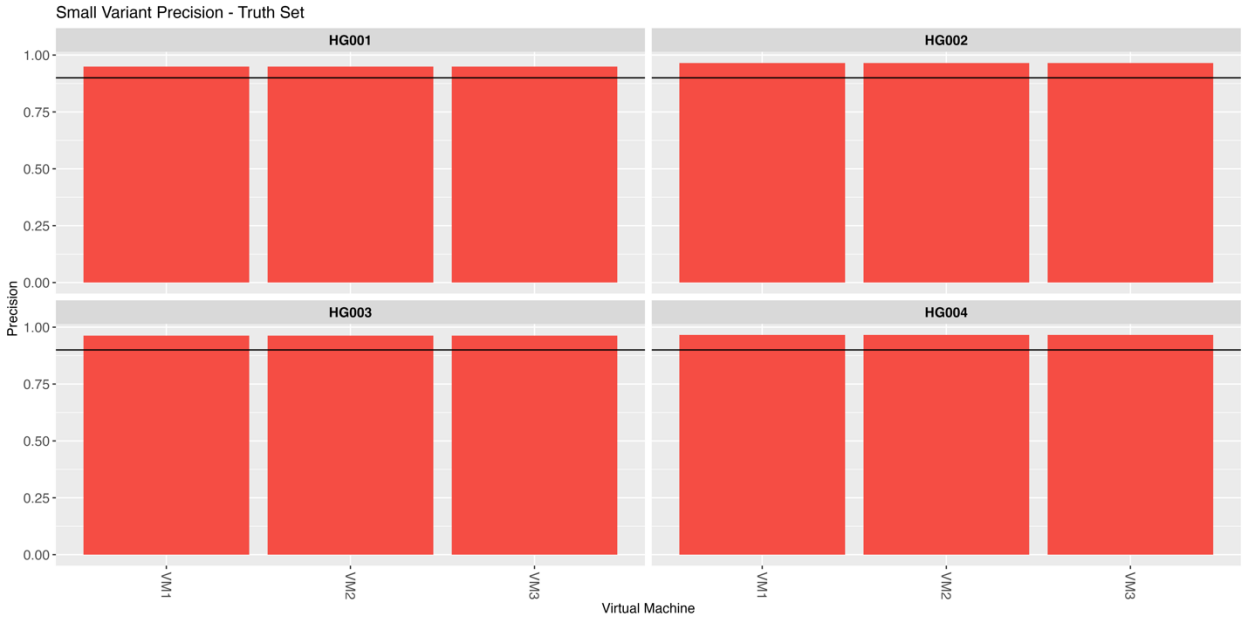
Pipeline Performance Specifications

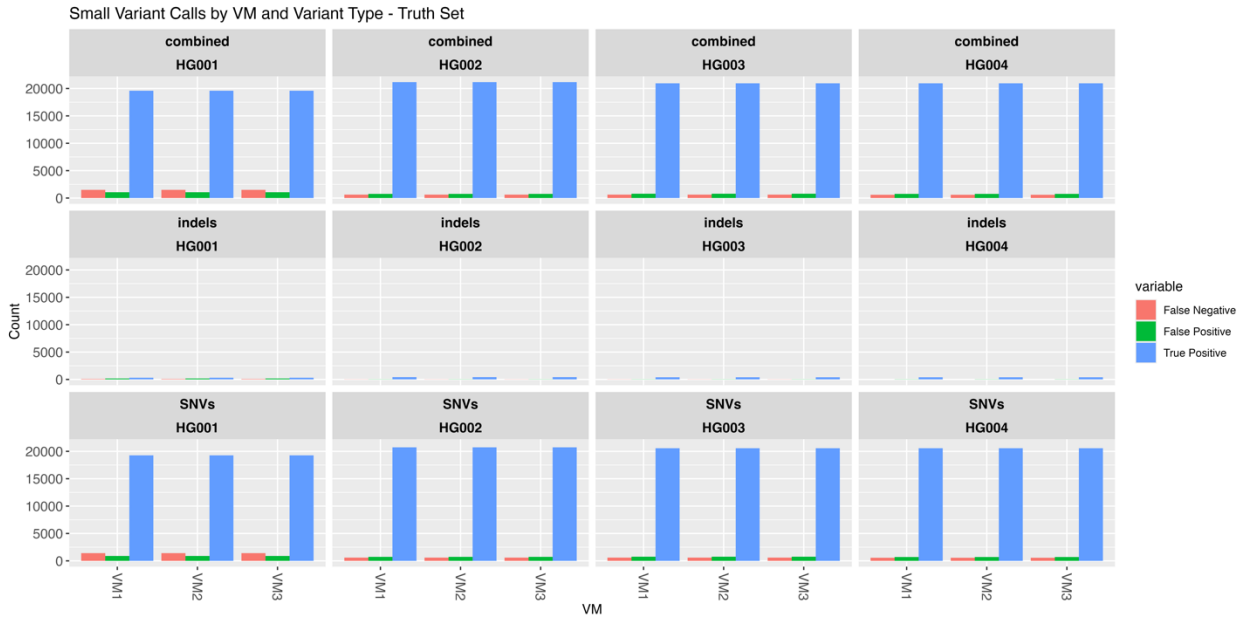
Small Variants		
	Truth Set	**Pipeline Reproducibility**
Recall	0.9610	1.000
Precision	0.9608	1.000

CNV		
	Truth Set	**Pipeline Reproducibility**
Sensitivity	0.8201	1.000
Specificity	0.9406	1.000
Accuracy	0.9239	1.000

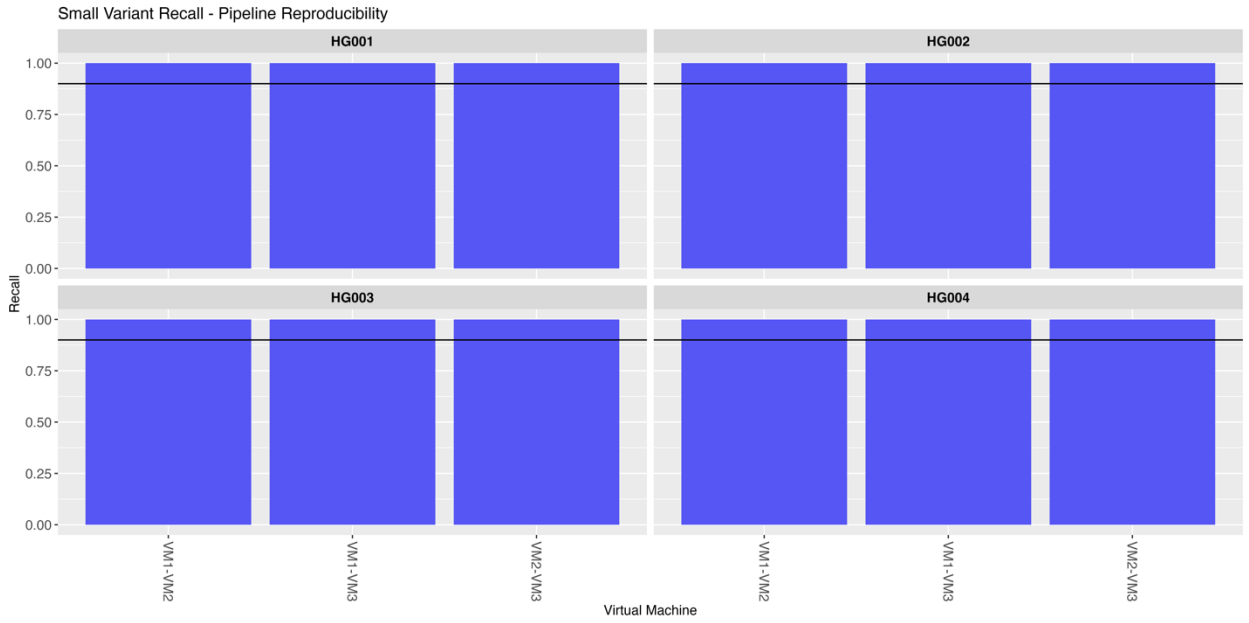
Small Variants - Comparison to Truth Set

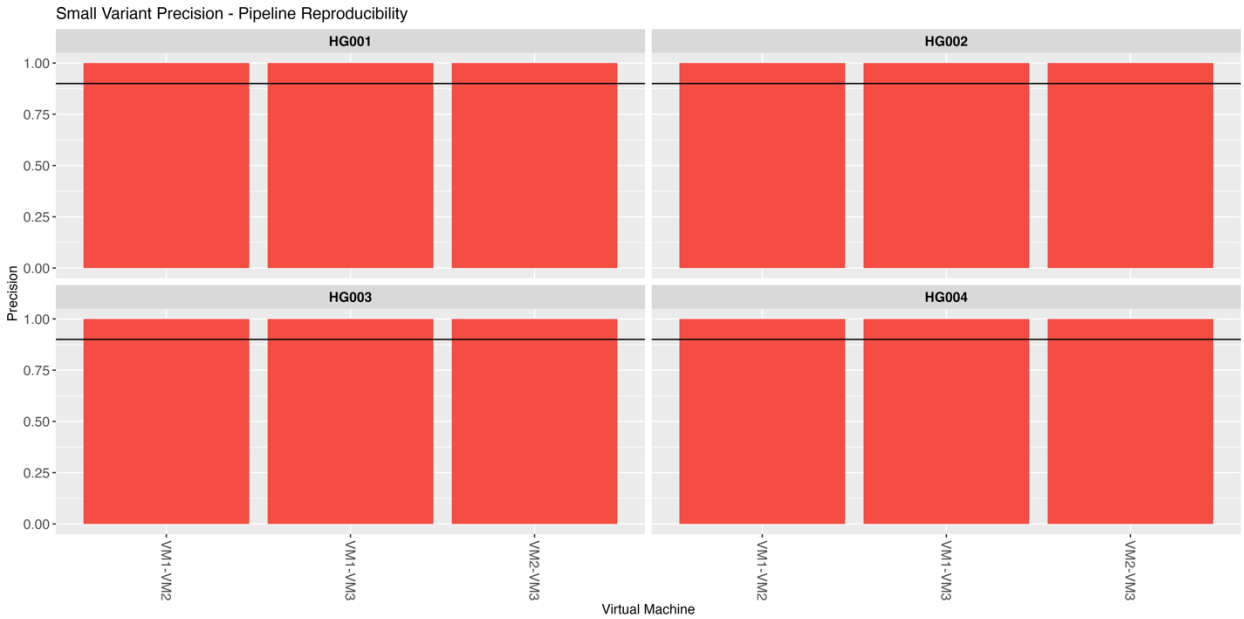

Small Variant Recall - Truth Set
HG001
HG002
HG003
HG004
Recall
1.00
0.75
0.50
0.25
0.00
VM1
VM2
VM3
Virtual Machine

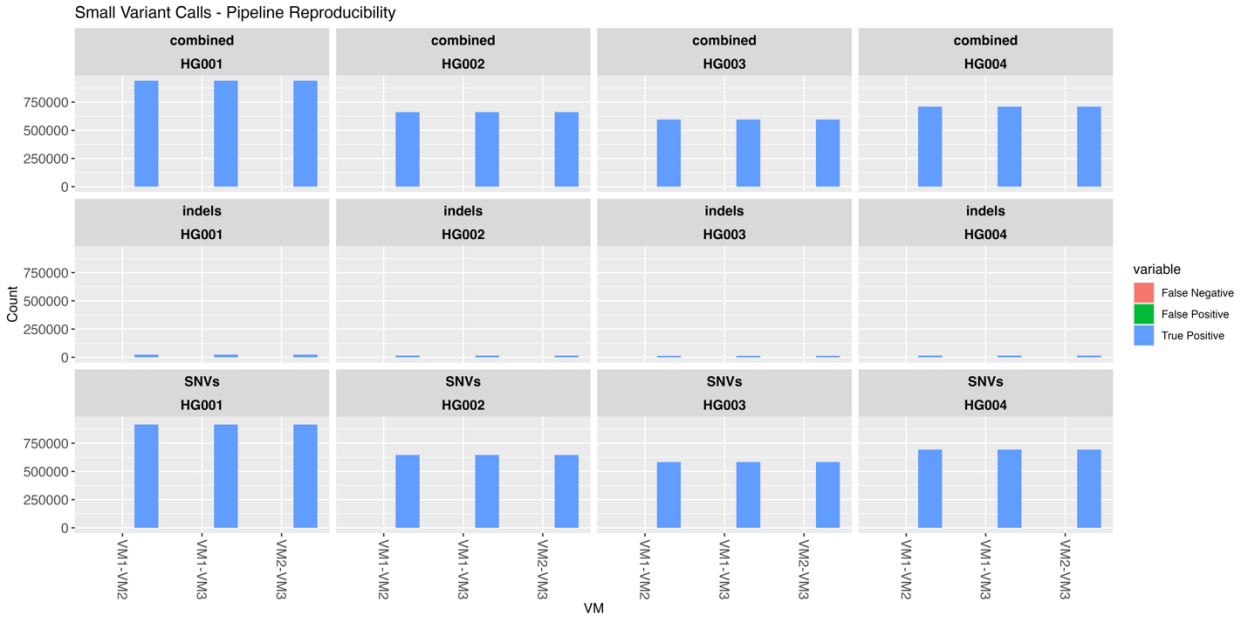

Small Variant Precision - Truth Set
HG001
HG002
HG003
HG004
Precision
1.00
0.75
0.50
0.25
0.00
VM1
VM2
VM3
Virtual Machine



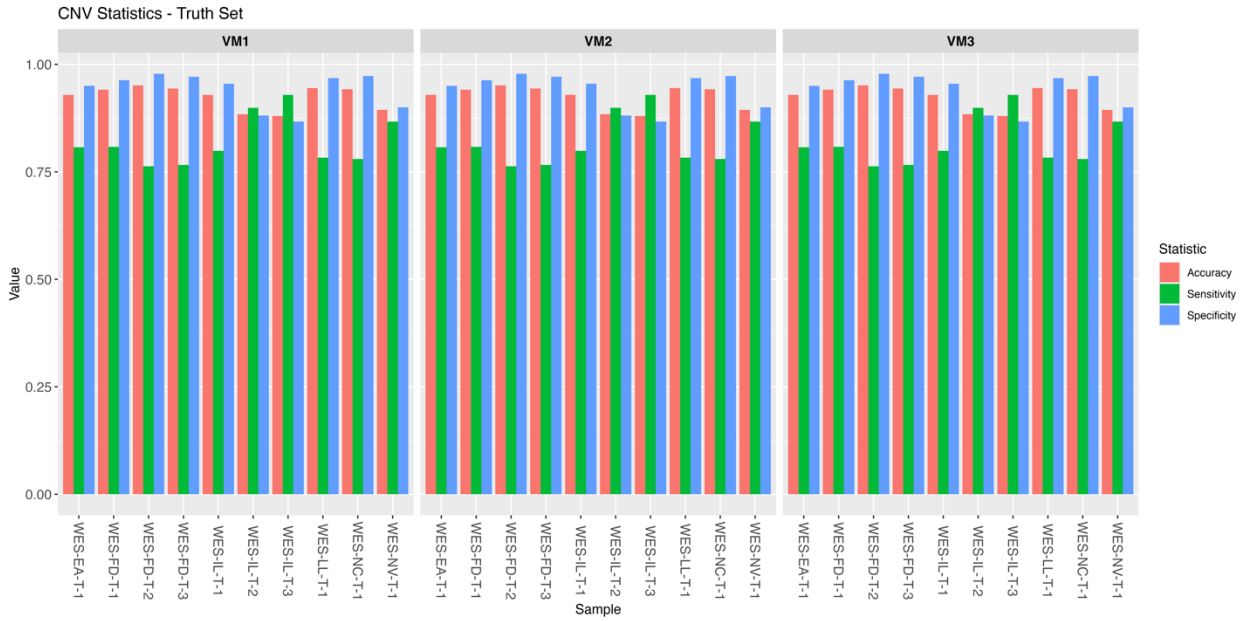
Small Variants – Pipeline Reproducibility



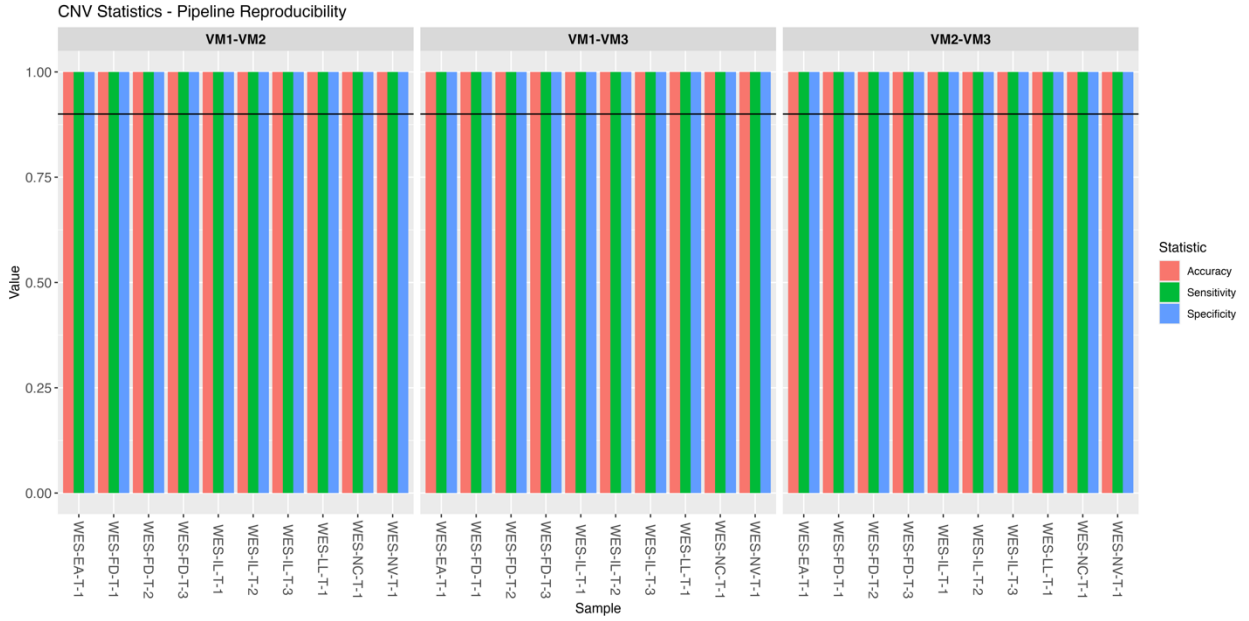




CNV - Comparison to Truth Set



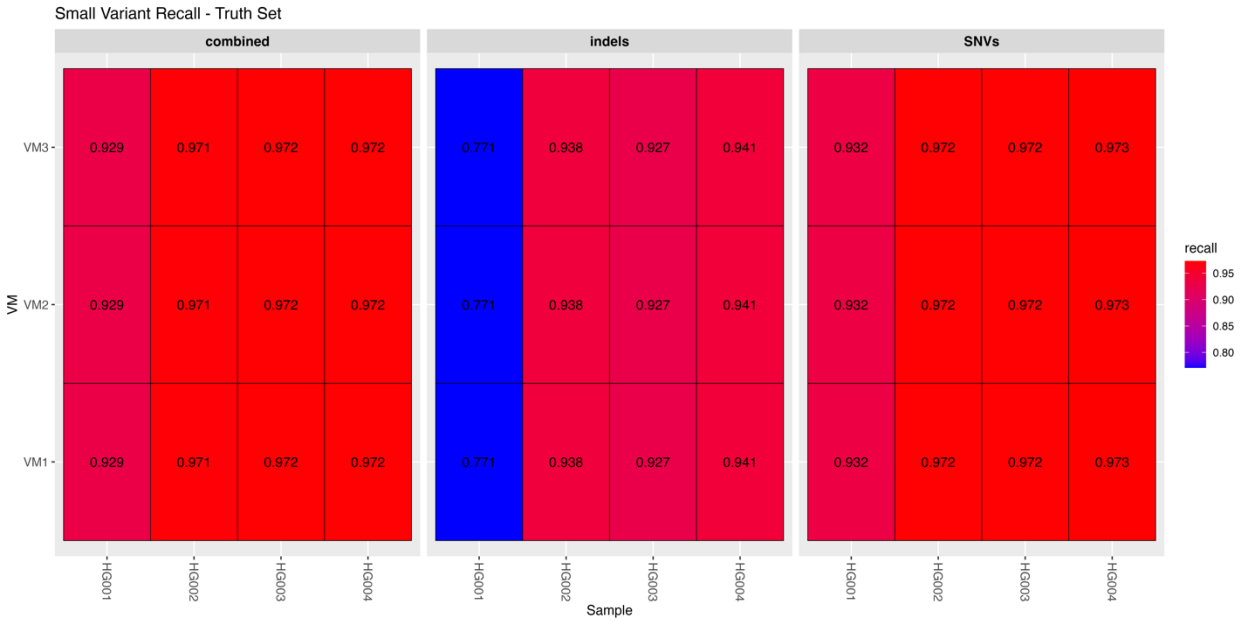
CNV – Pipeline Reproducibility

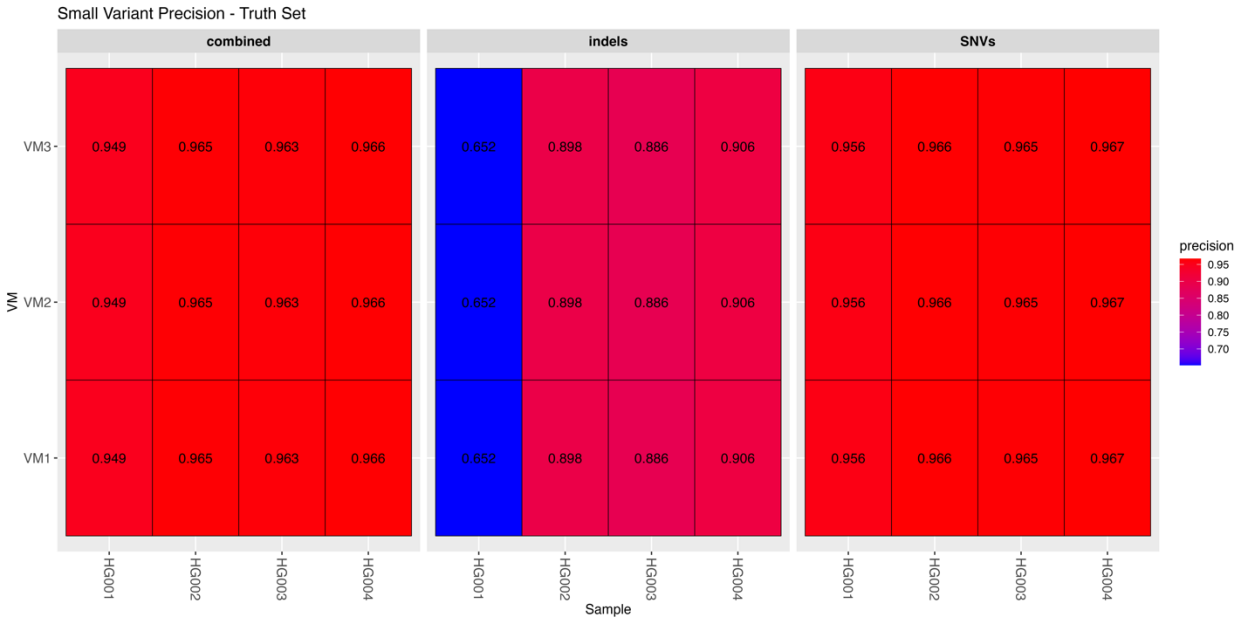


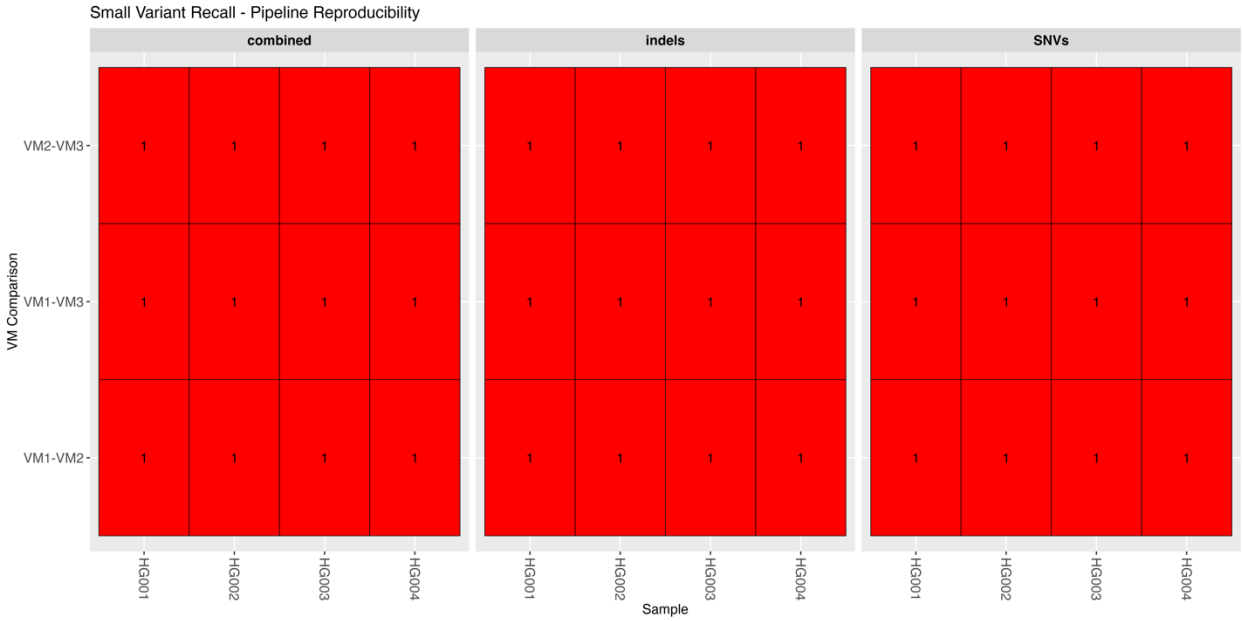
Conclusions: Based on these results, the pipeline meets the acceptance criteria for analytic performance as described prior to commencement of the validation activities. The pipeline generates highly reproducible results, as exemplified by the perfectly matched results created by

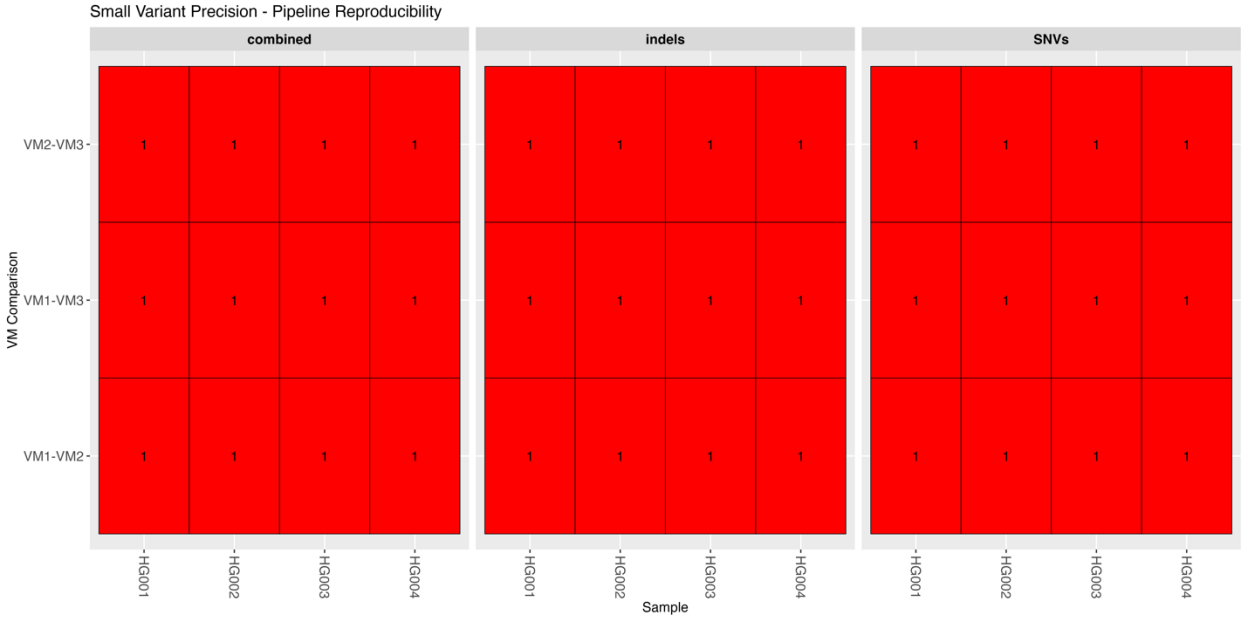
three users on three virtual machines. The pipeline also shows excellent performance metrics for calling small variants. The pipeline does show reduced sensitivity (however, above the acceptance criteria) for small indels and CNVs. This characteristic is not unexpected given the known limitations of the short-read whole exome sequencing in general (for example, see *Masood et al. Evaluation of somatic copy number variation detection by NGS technologies and bioinformatics tools on a hyper-diploid cancer genome (2023)*). The specificity of the pipeline is excellent regardless of the variant type, which is a desirable trait for pipelines used for analysis of clinical data. Given these results, it is suggested that this pipeline be released for analysis of CIMAC-CIDC WES data.

5. WES Pipeline – Appendix I – Tabular Results

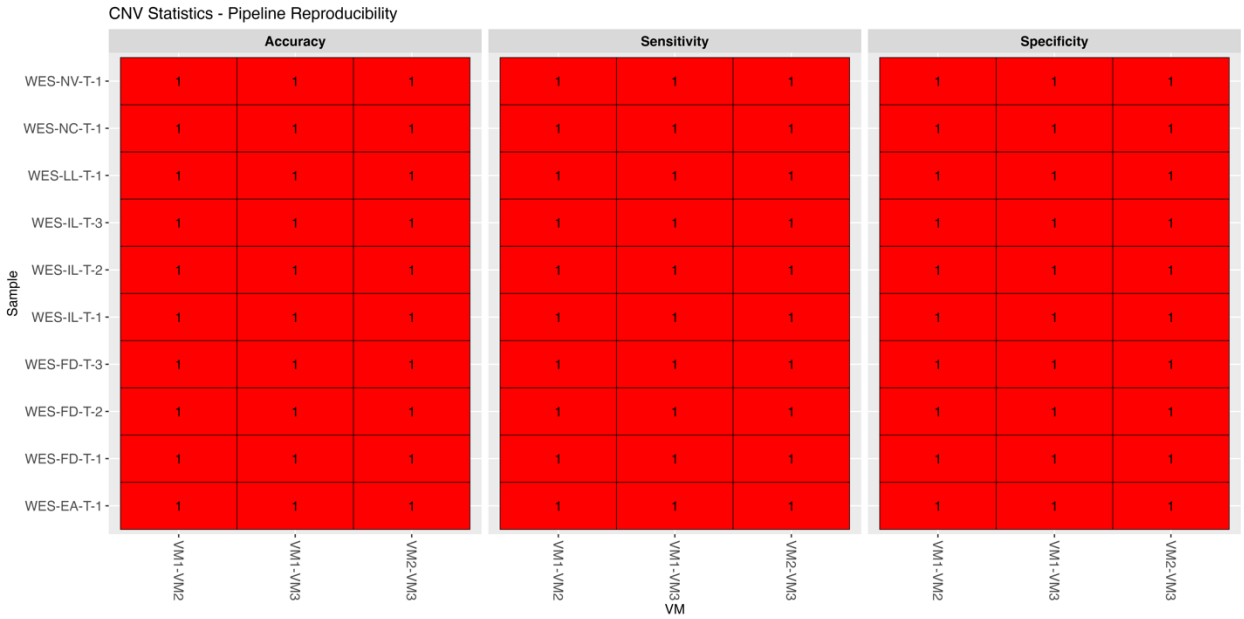



Small Variant Precision - Truth Set
combined
indels
SNVs
VM3
VM2
VM1
VM
0.949
0.965
0.963
0.966
0.652
0.898
0.886
0.906
0.956
0.966
0.965
0.967
precision
0.95
0.90
0.85
0.80
0.75
0.70
HG001
HG002
HG003
HG004
Sample


Small Variant Recall - Pipeline Reproducibility
combined
indels
SNVs
VM2-VM3
VM1-VM3
VM1-VM2
VM Comparison
1
HG001
HG002
HG003
HG004
Sample


Small Variant Precision - Pipeline Reproducibility
combined
indels
SNVs
VM2-VM3
VM1-VM3
VM1-VM2
VM Comparison
1
HG001
HG002
HG003
HG004
Sample


CNV Statistics - Truth Set
Accuracy
Sensitivity
Specificity
WES-NV-T-1
WES-NC-T-1
WES-LL-T-1
WES-IL-T-3
WES-IL-T-2
WES-IL-T-1
WES-FD-T-3
WES-FD-T-2
WES-FD-T-1
WES-EA-T-1
Sample
0.894 0.894 0.894 0.867 0.867 0.867 0.9 0.9 0.9
0.942 0.942 0.942 0.78 0.78 0.78 0.973 0.973 0.973
0.945 0.945 0.945 0.783 0.783 0.783 0.968 0.968 0.968
0.88 0.88 0.88 0.929 0.929 0.929 0.867 0.867 0.867
0.884 0.884 0.884 0.899 0.899 0.899 0.881 0.881 0.881
0.929 0.929 0.929 0.799 0.799 0.799 0.955 0.955 0.955
0.944 0.944 0.944 0.766 0.766 0.766 0.971 0.971 0.971
0.951 0.951 0.951 0.763 0.763 0.763 0.978 0.978 0.978
0.941 0.941 0.941 0.808 0.808 0.808 0.963 0.963 0.963
0.929 0.929 0.929 0.807 0.807 0.807 0.95 0.95 0.95
VM1
VM2
VM3
VM
Value
0.95
0.90
0.85
0.80


CNV Statistics - Pipeline Reproducibility
Accuracy
Sensitivity
Specificity
Sample
WES-NV-T-1
WES-NC-T-1
WES-LL-T-1
WES-IL-T-3
WES-IL-T-2
WES-IL-T-1
WES-FD-T-3
WES-FD-T-2
WES-FD-T-1
WES-EA-T-1
1
VM1-VM2
VM1-VM3
VM2-VM3
VM