Beyond One Million Genomes

D3.3
The B1MG data analysis challenge

Project Title (grant agreement No)	Beyond One Million Genomes (B1MG) Grant Agreement 951724		
Project Acronym	B1MG		
WP No & Title	WP3 - Standards & Quality Guidelines		
WP Leaders	Ivo Gut (CNAG), Jeroen Belien (VUmc)		
Deliverable Lead Beneficiary	3 - CRG		
Deliverable	D3.3 - The B1MG data analysis challenge		
Contractual delivery date	31/05/2023	**Actual delivery date**	31/10/2023
Delayed	Yes		
Authors	Ivo Gut (CNAG), Gabriela Aguileta (CNAG)		
Contributors	Edwin Cuppen (HMF), Valtteri Wirta (KI), Eivind Hovig (UiO), Gert Matthijs (KU Leuven)		
Acknowledgements (not grant participants)			
Deliverable type	Report		
Dissemination level	Public		

Beyond One Million Genomes
B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Document History

Date	Mvm	Who	Description
23/10/2023	0v1	Ivo Gut (CRG)	Initial draft written and circulated to WP participants for feedback
23/10/2023	0v2	Nikki Coutts (ELIXIR Hub)	Circulated to OG, Stakeholders and GB for review
31/10/2023	0v3	Ivo Gut (CRG) & Gabriela Aguileta Estrada (CNAG)	Final comments closed
31/10/2023	1v0	Nikki Coutts (ELIXIR Hub)	Version uploaded to the EC Portal

Table of Contents

1. Executive Summary

Germline and tumor whole genome sequencing (WGS) have now become a standard procedure, integral to both research and clinical practices. However, the diversity in analytical approaches across laboratories remains pronounced. This diversity calls for the establishment of cohesive standards, a need that has yet to be sufficiently addressed. Presently, there exists a scarcity of comprehensive schemes designed to authenticate or set benchmarks for the effectiveness of germline and tumor WGS pipelines.

Addressing this gap, the European H2020 initiative 1+MG has emerged with a specific mission: to bridge the connection between genomic and health data analyses. Achieving this mission mandates a meticulous exploration of existing voids and optimal methodologies within germline and tumor WGS. This is not only crucial for enhancing the quality of outcomes but also for fostering reproducibility and engendering trust among stakeholders.

To achieve these objectives, the collaborative efforts of the 1+MG and B1MG projects have been mobilised. The central focus lies in the orchestration of a somatic WGS benchmarking initiative, encompassing three distinct challenges:

Wet Lab Challenge: This segment scrutinises the library preparation and sequencing stages, with an emphasis on evaluating the precision and robustness of these processes.

Full Pipeline Challenge: Encompassing library preparation, sequencing, and data analysis, this challenge offers a comprehensive evaluation of the end-to-end workflow. The goal is to assess the integrity of the entire pipeline in generating reliable results.

Dry Lab Challenge: The data analysis pipeline takes centre stage in this challenge, as it seeks to appraise the computational methodologies employed in deciphering and interpreting the genomic data.

By structuring these challenges, the 1+MG and B1MG projects have contributed significantly towards harmonising WGS practices, fostering a unified understanding of best practices, and nurturing confidence among stakeholders. This progressive approach not only ensures high-quality outcomes but also supports the critical drive for reproducibility and reliability within the realm of genomic and health data analysis.

To this date, the 1+MG WG4 has organised a comprehensive quality comparison for all the stages of the somatic whole genome variant calling process. As described above, we have divided the workflow into three main tasks: the wetlab, the full pipeline, and the dry lab challenges. For each of these stages, we have collected results from all the participating labs and obtained the relevant quality metrics. The comparison of results across all labs has provided the baseline for the construction of a curated dataset of somatic variants with the highest reliability. This goldset establishes the standard of quality against which individual laboratory observations are measured.

The 1+MG WG4 has provided best practices for whole genome somatic variant calling through a comprehensive benchmark of quality metrics for all stages of the process. This work has also contributed to the generation of a goldset of somatic variant calls, for both small and large variants. In a larger framework, the 1+MG WG4 sets the quality requirements of genomic data for cross-border access and for personalised medicine practice.

2. Contribution towards project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

	Key Result No and description	Contributed
Objective 1 Engage local, regional, national and European stakeholders to define the requirements for cross-border access to genomics and personalised medicine data	**1.** B1MG assembles key local, national, European and global actors in the field of Personalised Medicine within a B1MG Stakeholder Coordination Group (WP1) by M6.	Yes
	2. B1MG drives broad engagement around European access to personalised medicine data via the B1MG Stakeholder Coordination Portal (WP1) following the B1MG Communication Strategy (WP6) by M12.	No
	3. B1MG establishes awareness and dialogue with a broad set of societal actors via a continuously monitored and refined communications strategy (WP1, WP6) by M12, M18, M24 & M30.	No
	4. The open B1MG Summit (M18) engages and ensures that the views of all relevant stakeholders are captured in B1MG requirements and guidelines (WP1, WP6).	No
Objective 2 Translate requirements for data quality, standards, technical infrastructure, and ELSI into technical specifications and implementation guidelines that captures European best practice	**Legal & Ethical Key Results**	
	1. Establish relevant best practice in ethics of cross-border access to genome and phenotypic data (WP2) by M36	Yes
	2. Analysis of legal framework and development of common minimum standard (WP2) by M36.	No
	3. Cross-border Data Access and Use Governance Toolkit Framework (WP2) by M36.	No
	Technical Key Results	
	4. Quality metrics for sequencing (WP3) by M12.	Yes
	5. Best practices for Next Generation Sequencing (WP3) by M24.	Yes
	6. Phenotypic and clinical metadata framework (WP3) by M12, M24 & M36.	No
	7. Best practices in sharing and linking phenotypic and genetic data (WP3) by M12 & M24.	No
	8. Data analysis challenge (WP3) by M36.	Yes
	Infrastructure Key Results	

	9. Secure cross-border data access roadmap (WP4) by M12 & M36.	No
	10. Secure cross-border data access demonstrator (WP4) by M24.	No
Objective 3 Drive adoption and support long-term operation by organisations at local, regional, national and European level by providing guidance on phased development (via the B1MG maturity level model), and a methodology for economic evaluation	**1.** The B1MG maturity level model (WP5) by M24.	No
	2. Roadmap and guidance tools for countries for effective implementation of Personalised Medicine (WP5) by M36.	No
	3. Economic evaluation models for Personalised Medicine and case studies (WP5) by M30.	No
	4. Guidance principles for national mirror groups and cross-border Personalised Medicine governance (WP6) by M30.	Yes
	5. Long-term sustainability design and funding routes for cross-border Personalised Medicine delivery (WP6) by M34.	No

3. Methods

3.1 Somatic Benchmarking Scheme Design

The CNAG organises and takes part in the WGS benchmark for somatic variants. Participant laboratories come from different European countries, and for this initial evaluation there are a total of 9 participant institutes. The overall organisation of the benchmark (Fig. A1) was as follows: 1) The organiser obtained the samples from the Medical University of Graz (MUG), as part of the EASI Genomics initiative. CNAG did the DNA extraction from frozen tissue; 2) A Material Transfer Agreement (MTA) following all the applicable legal and ethical requirements was signed between each participant and the Organiser; 3) CNAG distributed and sent the material (described below) to each participant laboratory; 4) Participants prepared the libraries and sequenced the samples according to their own SOPs; 5) Somatic variant calling was done by participants following their own standard pipelines; 6) Participants sent their results back to the organiser, including a SAV (output file with metrics that is produced by the sequencing instrument), a FASTQ file and the result of variant calling of both small and large variants in the VCF format; 7) CNAG built a gold set using CNAG's and all the participants' results, as well as Nanopore sequencing of the cancer samples to validate structural variants; 8) CNAG uses the gold set to benchmark the results from each participant and produce reports assessing the performance of all participant laboratories. Figure 1 in the Appendix shows the scheme of the somatic benchmarking.

3.2 Test Items

Participants received DNA from 8 Tumour/Normal pairs extracted from frozen tissue: 2 head and neck squamous cell carcinoma samples, 3 lung squamous cell carcinoma samples, and 3 clear cell renal carcinoma samples, as well as their matching normal samples. Table 1 in the Appendix contains the list and description of each of the Tumour/Normal pairs.

3.3 Quality Metrics Evaluation

Together with participants, a set of Quality Metrics was defined to evaluate the performance of each participant laboratory for library preparation, sequencing and data analysis (Deliverable D3.1[1]).The assigned value and acceptable range of results were established, according to the guidelines provided in ISO13528 (2015). According to the definition of the guidelines, an assigned value is an estimate of the value of the measure and that is used for calculating scores. Table 2 in the Appendix shows the QC metrics used to compare across laboratories. The chosen QC metrics provide a measure of the quality of the performance of the laboratory at each stage of the benchmark. The organiser will assess QC metrics of sequencing, sample preparation, post alignment, and variant calling of all the participants in order to compare the performance across laboratories.

3.4 Gold Set Generation

A gold set is a curated set of variants that we are highly confident that are true. To obtain a gold set for the somatic benchmark, the FASTQ data was merged by concatenating the FASTQ files submitted by all participants. Collaborators ran their different SOP pipelines to call small and large somatic variants and returned the results to CNAG. Additionally, nanopore data was used to confirm the large variants.
For the construction of the goldset, CNAG compared the VCF files obtained from the merged FASTQ files by the participants to establish the set of calls that were identically called by all labs. For the rest of calls, CNAG set a rule to establish which calls, although not called by all methods, were still highly reliable to be included in the goldset. In this way, true variants called by most methods but missed by some could be recovered and retained in the goldset. Lastly, for the difficult variants, especially in the case of structural and copy number variants, CNAG developed a variant voter tool to visually curate the most difficult cases.

3.5 Dry Lab Challenge

The goal of the Dry Lab Challenge was to evaluate the impact of the bioinformatics pipeline on the quality of somatic variant calling. To this end, CNAG distributed to all the participating labs the same FASTQ files corresponding to 3 pairs of tumor normal samples: HNO-002, THX-001 and URO-003. Each lab processed the FASTQ files, ran their standard bioinformatics pipelines and submitted the resulting VCF files and TSV files, for the small and large variants, respectively.

[1] https://zenodo.org/record/5018495#.ZFoIuuzMJjc

CNAG compared the results across all lab submissions.

3.6 Benchmark variant voter

The variants called by all participants were used in the construction of the gold set. In the case of discrepant variants, CNAG developed a variant voter tool, which is an R shiny app, available online to all participants through Shinyapps.io, where they were presented with genomic viewer screenshots representing discrepant variants and voters were able to cast their votes more efficiently. The manual curation of difficult variants has helped in the construction of the gold set.

3.7 Final reports

ISO 17043 defines the requirements of the participants' reports. We have decided to make two reports: a general report and a participants' report. We deliver the same general report to all participants. We deliver a personalised report to each participant separately.

4. Description of work accomplished

4.1 1+MG WG4 Somatic Benchmark

4.1.1 Sequencing and alignment QC metric comparisons

CNAG obtained the QC metrics for each stage of the variant calling process (Table 2).

For the alignment stage, the QC metrics estimated and compared across labs were:

- % duplicate reads
- Median insert size
- Mean coverage
- Evenness of coverage
- % chimeras

To evaluate the performance of each lab relative to all others, we used z-scores, which measure how much an observation deviates from the mean of observations in terms of standard deviation units. Based on absolute z-scores (|z-score|), there are three possible outcomes for each observation: acceptable (below 2 |z-scores|), questionable (between 2-3 |z-scores|), and unacceptable (above 3 |z-scores|). Depending on the QC metric, the limits that define the outcomes are established by the thresholds of each measure, as they can be upper, lower or both. Additionally, CNAG used the inter-quartile range (IQR) to determine which observations are outliers. Given a distribution of observations, the interquartile range can be estimated. An observation is an outlier if it has a value 1.5 times greater than the IQR, or 1.5 times less than the IQR. The plots corresponding to these comparisons are shown in the Results section below.

4.1.2 Variant calling and somatic QC metric comparisons

For the somatic variant calling stage, the QC metrics estimated and compared across labs were:

- Ti/Tv ratio
- Number of somatic SNVs (single nucleotide variants)
- Number of somatic INDELs (insertions and deletions)
- Number of somatic SVs (structural variants)
- Number of somatic CNVs (copy number variants)

The classification of observations by outcome and outlier criteria was done as described above for the alignment QC metrics. The plots corresponding to these comparisons are shown in the Results section below.

4.2 Construction of goldsets

4.2.1 From merged FASTQ to multiple VCF comparisons

Participant labs produced a merged FASTQ file per sample, by merging the FASTQ files generated by each participant lab. The total depth of the calls obtained from the merged files was around 600X. This depth was desirable to gain reliability in the calling.

CNAG received a VCF file at 600X, per sample, from the participants and did the comparisons across all datasets to extract the reliable variants that went into building the goldset.

The variants identically called by all laboratories using different calling pipelines, were kept for the goldset. For the rest of calls, CNAG set a rule to established which calls, although not called by all methods, were still highly reliable to be included in the goldset. In this way, true variants called by most methods but missed by some could be recovered and retained in the goldset. Lastly, for the difficult variants, especially in the case of structural and copy number variants, CNAG developed a variant voter tool to visually curate the most challenging cases.

4.2.2 Benchmarking of somatic variant calls

Comparisons of the variants produced by the different participating labs, against the goldset, allowed CNAG to benchmark the calls from the different labs.

We obtained the F1, Precision and Recall metrics for all labs with respect to the goldset. To do so, CNAG generated a nextflow pipeline for the benchmark that used different tools: hap.py, prep.py, and rep.py, for the small variants (SNVs and indels), and truvari for the large variants (structural and copy number variants).

4.2.3 Dry Lab challenge comparison of somatic variant calls

CNAG compared the dry lab challenge results using the z-scores approach to determine the outcome of each observation, as described above (4.1.1).

5. Results

5.1.1 Alignment QC metric comparisons

In the Deliverable B1MG D3.2 Best practices for Next Generation Sequencing, we have presented the results of the comparison of the sequencing QC metrics across participating labs. Here, we present the comparison of the alignment QC metrics across labs.

% duplicate reads

This metric refers to the fraction of mapped sequence that is marked as duplicate. The indicative value reported in the literature (Marshall et al. 2020) is < 10%. Results above this value are deemed either questionable or unacceptable.

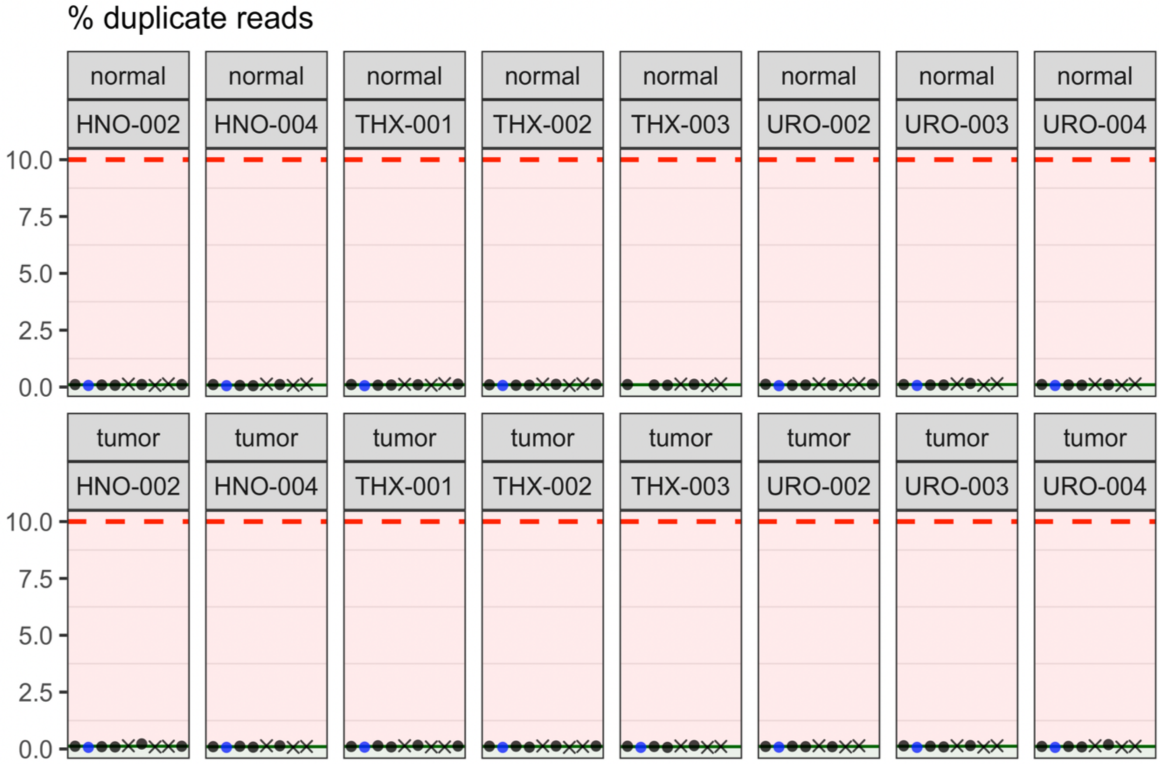

Figure 1: Percentage of duplicate reads. QC metrics for the normal samples are in the top panel and for the tumor samples in the bottom panel. The indicative value (dashed line) is 10% and the assigned value (solid line) corresponds to the overall mean for all observations of normal samples. The blue dots correspond to the CNAG values for this QC metric. Dots indicate accredited labs and crosses those that are not accredited. All the observations are at zero or near zero.

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Median insert size

This metric refers to the median insert size of all paired end reads where both ends mapped to the same chromosome. The indicative value reported in the literature (Marshall et al. 2020) is > 300. Results below this value are deemed either questionable or unacceptable.

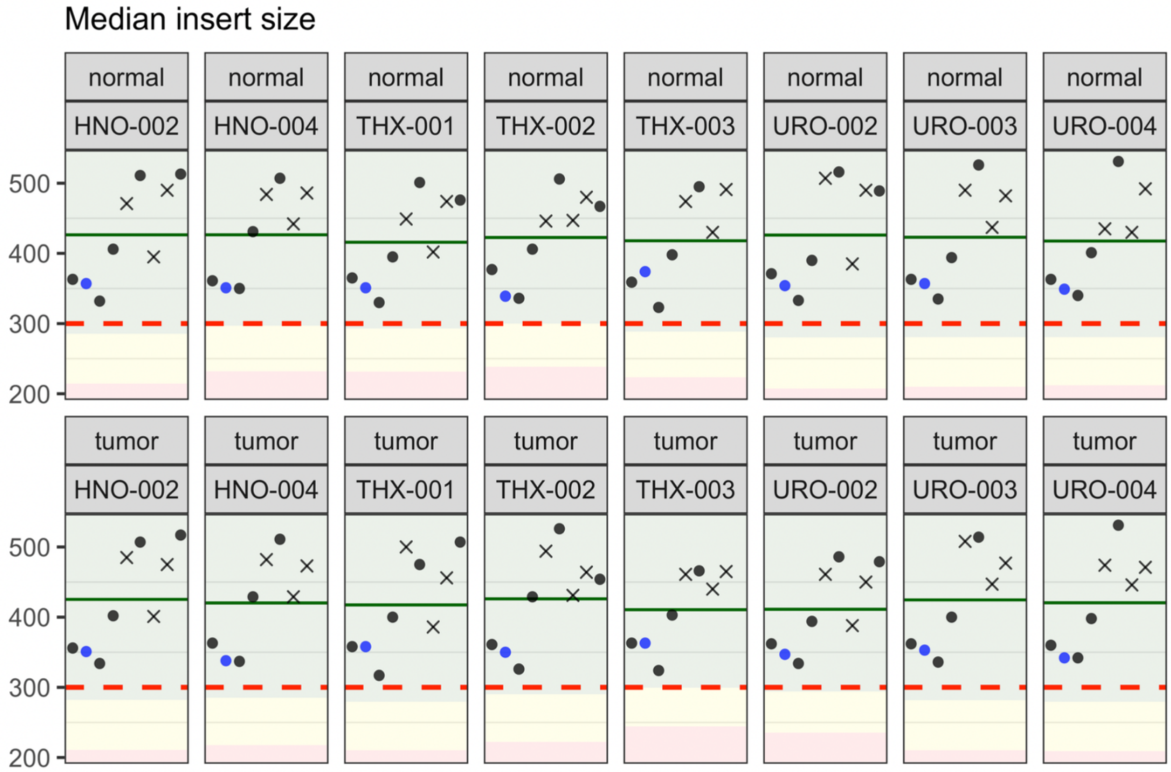

Figure 2: Median insert size. QC metrics for the normal samples are in the top panel and for the tumor samples in the bottom panel. The indicative value (dashed line) is 300 base pairs (bp) and the assigned value (solid line) corresponds to the overall mean for all observations of normal samples. The blue dots correspond to the CNAG values for this QC metric. Dots indicate accredited labs and crosses those that are not accredited. All the observations are acceptable, as they are above 300 bp, the indicative value.

Mean coverage

This QC metric is the average number of times a position in the genome is sequenced, considering only uniquely aligned reads (MAPQ≥20). The indicative value reported in the literature (Marshall et al. 2020) is ≥ 30. Results below this value are deemed either questionable or unacceptable.

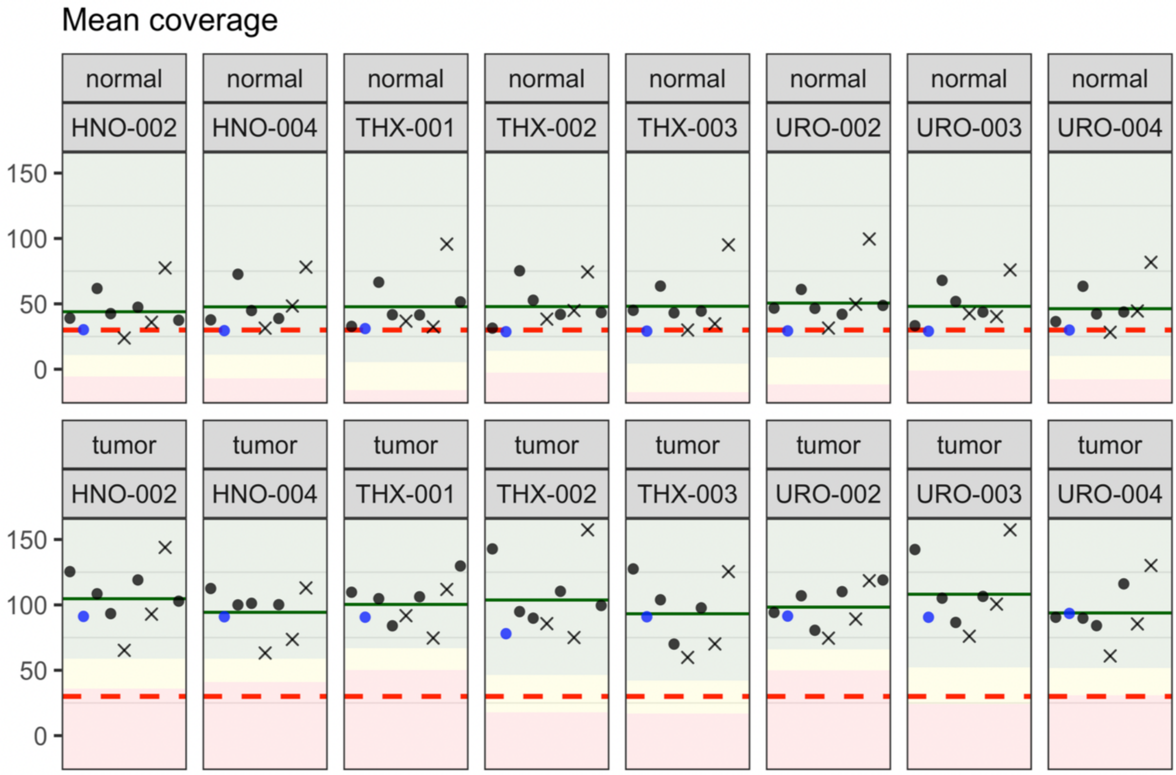

Figure 3: Mean coverage. QC metrics for the normal samples are in the top panel and for the tumor samples in the bottom panel. The indicative value (dashed line) is ≥ 30 and the assigned value (solid line) corresponds to the overall mean for all observations of normal samples. The blue dots correspond to the CNAG values for this QC metric. Dots indicate accredited labs and crosses those that are not accredited. All the observations are acceptable, as they are above 30, except for one observation in the normal sample HNO-002 that is slightly below 30.

Evenness of coverage

This QC metric is the median divided by the mean coverage. The optimal is to get an even coverage. If that is the case, values of the mean and median are close. The indicative value is therefore ≈ 1. Results well above or below this value are deemed either questionable or unacceptable.

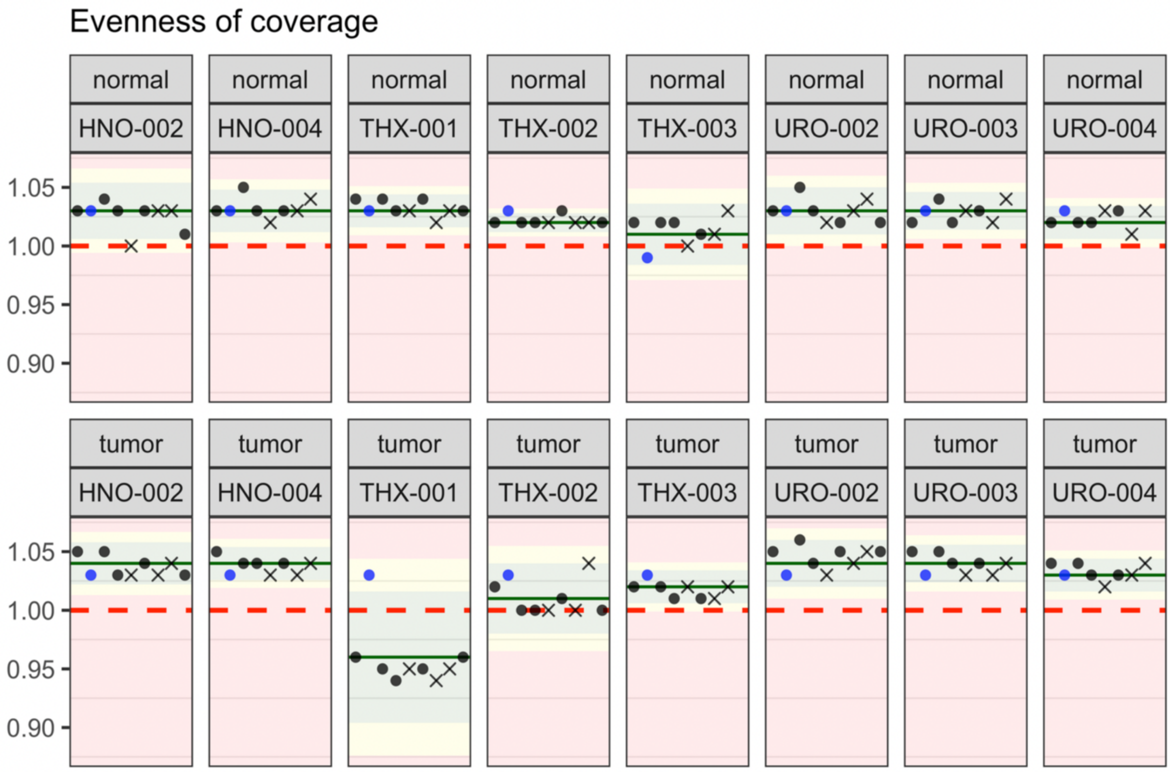

Figure 4: Evenness of coverage. QC metrics for the normal samples are in the top panel and for the tumor samples in the bottom panel. The indicative value (dashed line) is ≈ 1, and the assigned value (solid line) corresponds to the overall mean for all observations of normal samples. The blue dots correspond to the CNAG values for this QC metric. Dots indicate accredited labs and crosses those that are not accredited. Almost all the observations are acceptable, as they are ≈ 1, the optimal value. There are some questionable observations (yellow area) in normal samples HNO-002, HNO-004, THX-002, and URO-002, as well as in tumor samples THX-001, THX-002, and URO-002. Sample THX-001 has the mean farthest from 1 of all samples.

% chimeras

This QC metric is the fraction of reads that map outside of a maximum insert size or have the two ends mapping to different chromosomes. The indicative value reported in the literature (Marshall et al. 2020) is < 0.5. Results above this value are deemed either questionable or unacceptable.

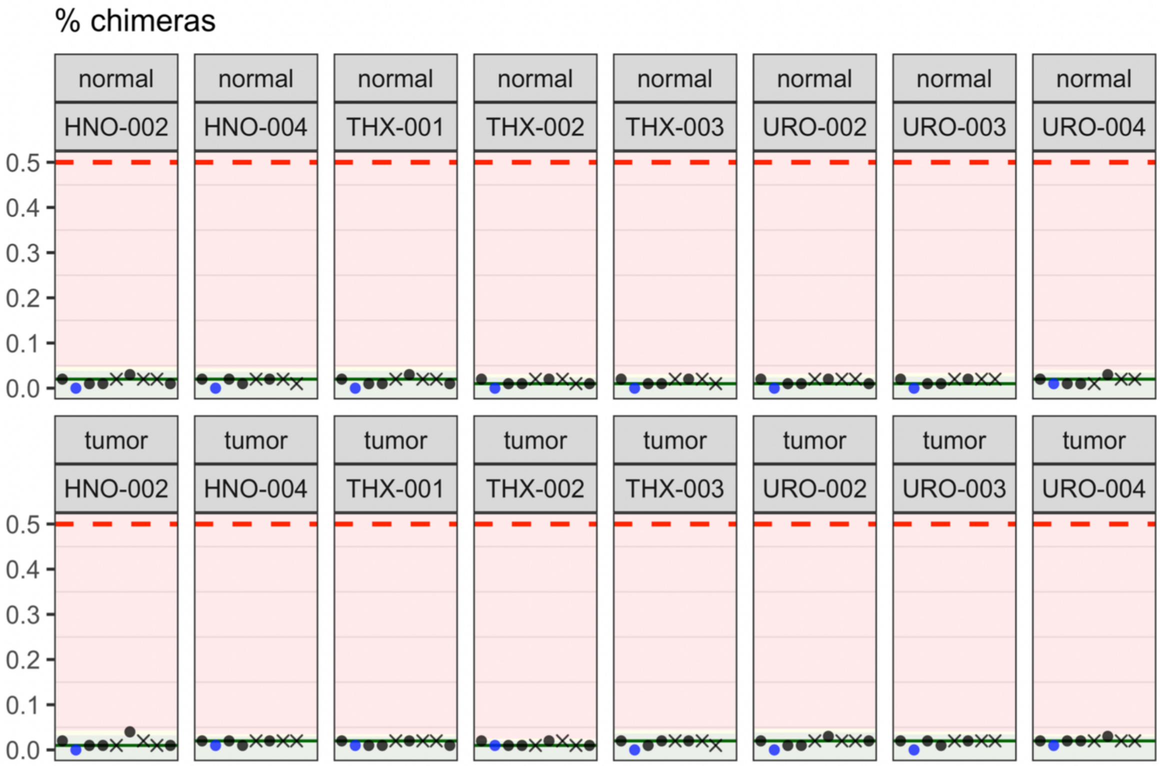

Figure 5: Percentage of chimeras. QC metrics for the normal samples are in the top panel and for the tumor samples in the bottom panel. The indicative value (dashed line) is < 0.5, and the assigned value (solid line) corresponds to the overall mean for all observations of normal samples. The blue dots correspond to the CNAG values for this QC metric. Dots indicate accredited labs and crosses those that are not accredited. All observations are well below the indicative value, and all are in the acceptable region, near zero.

5.1.2 Variant calling and somatic QC metric comparisons

Ti/Tv ratio

This metric is the count of transitions (from purine to purine or from pyrimidine to pyrimidine) over the count of transversions (from purine to pyrimidine or from pyrimidine to purine). The indicative value ≈ 2 is reported in the literature (Wang, Peng, and Leal 2014). Results well above or below this value are deemed either questionable or unacceptable.

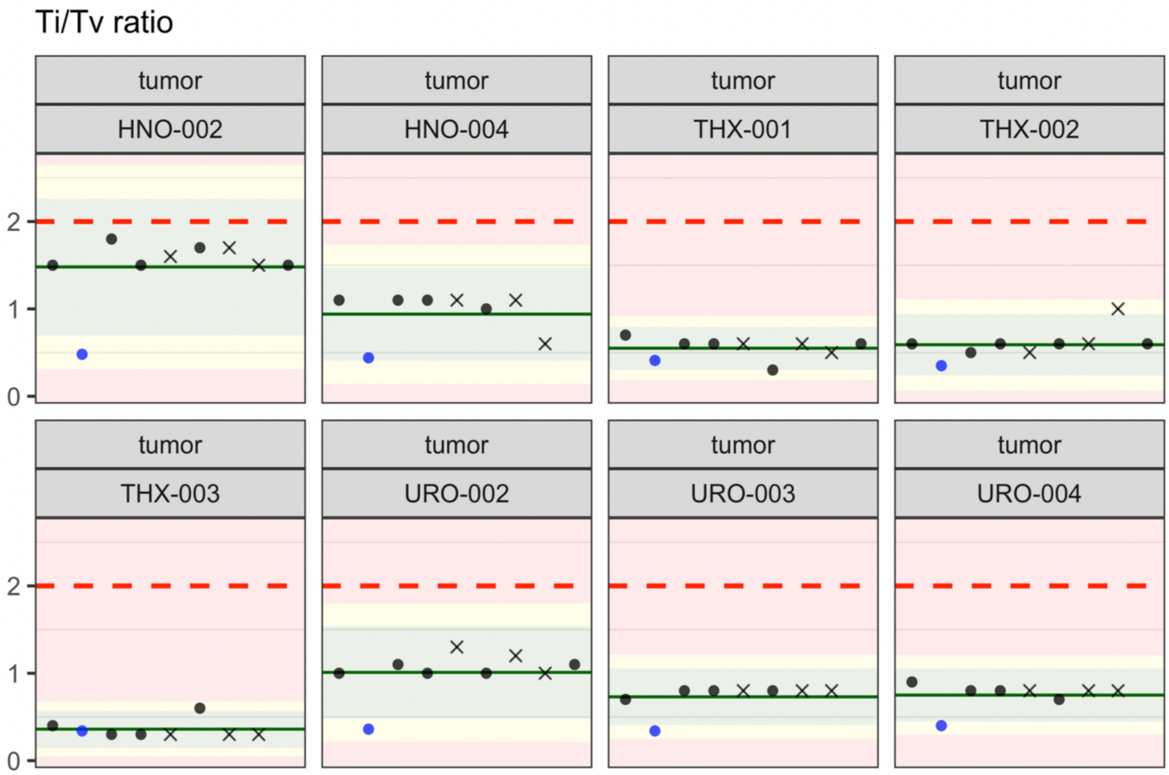

Figure 6: Ti/Tv ratio. The indicative value (dashed line) is ≈ 2, and the assigned value (solid line) corresponds to the overall mean for all observations of normal samples. The blue dots correspond to the CNAG values for this QC metric. Dots indicate accredited labs and crosses those that are not accredited. All observations are below 2, the indicative value. Nevertheless, most values fall within the acceptable region. The questionable results were observed in samples HNO-002, THX-002, THX-003, URO-002, URO-003, and URO-004.

Number of somatic SNVs (single nucleotide variants)

This metric corresponds to the number of passing bi-allelic single nucleotide variant calls (i.e., non-reference genotypes). An average 3.3 million SNVs per human genome were reported in the literature (Shen et al. 2013). This figure is illustrative only and varies from individual to individual. There is no fixed indicative value.

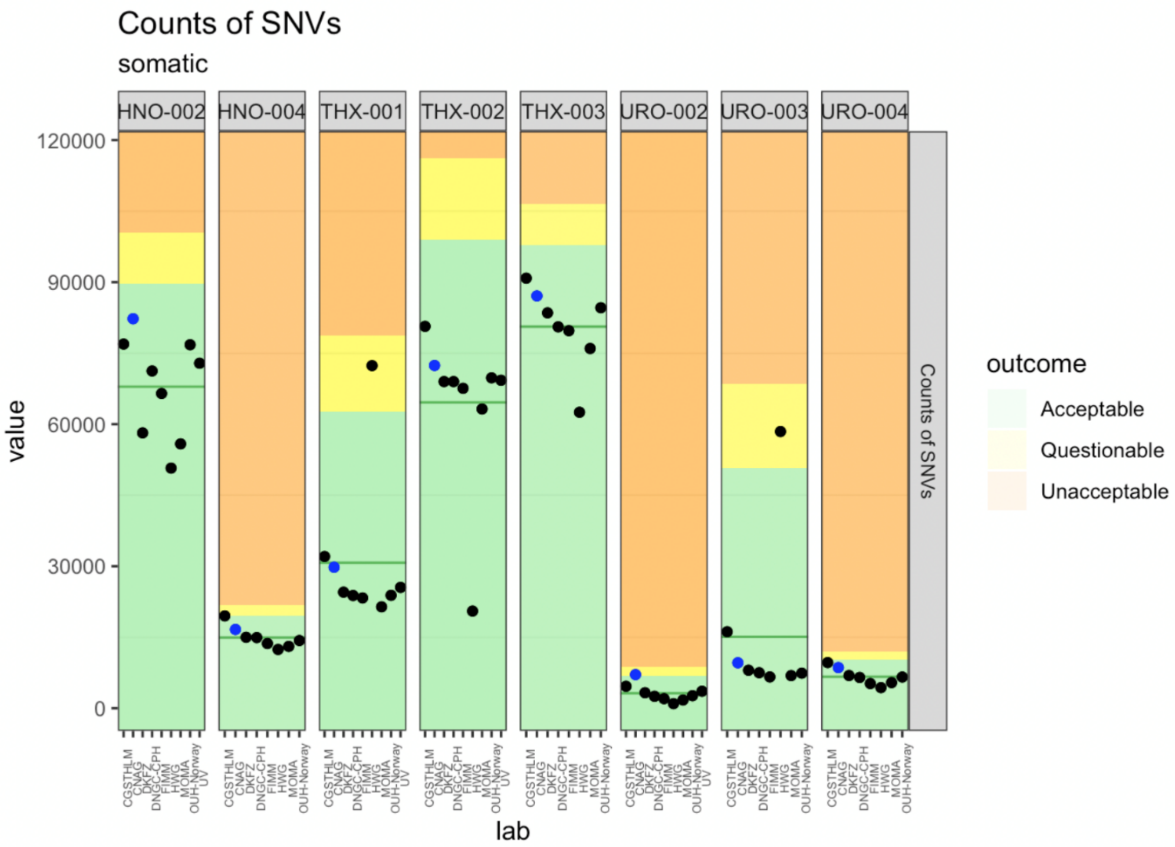

Figure 7: Counts of Single nucleotide variants (SNVs). The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis.

We observed a higher degree of dispersion around the mean than in the alignment QC metrics. At this stage of the whole genome sequencing process, the impact of the different bioinformatics pipelines used by the participating laboratories can be seen. It is also interesting to note that results among the samples obtained from tissues affected vary by the different cancer types. The HNO (head and neck) and THX (lung) samples show higher dispersion of results than those of the URO (renal) cancer type.

Number of somatic INDELs (small insertions and deletions)

This metric refers to the number of passing indel calls. An average 492,000 INDELs per human genome were reported in the literature (Shen et al. 2013). This figure is illustrative only and varies from individual to individual. There is no fixed indicative value.

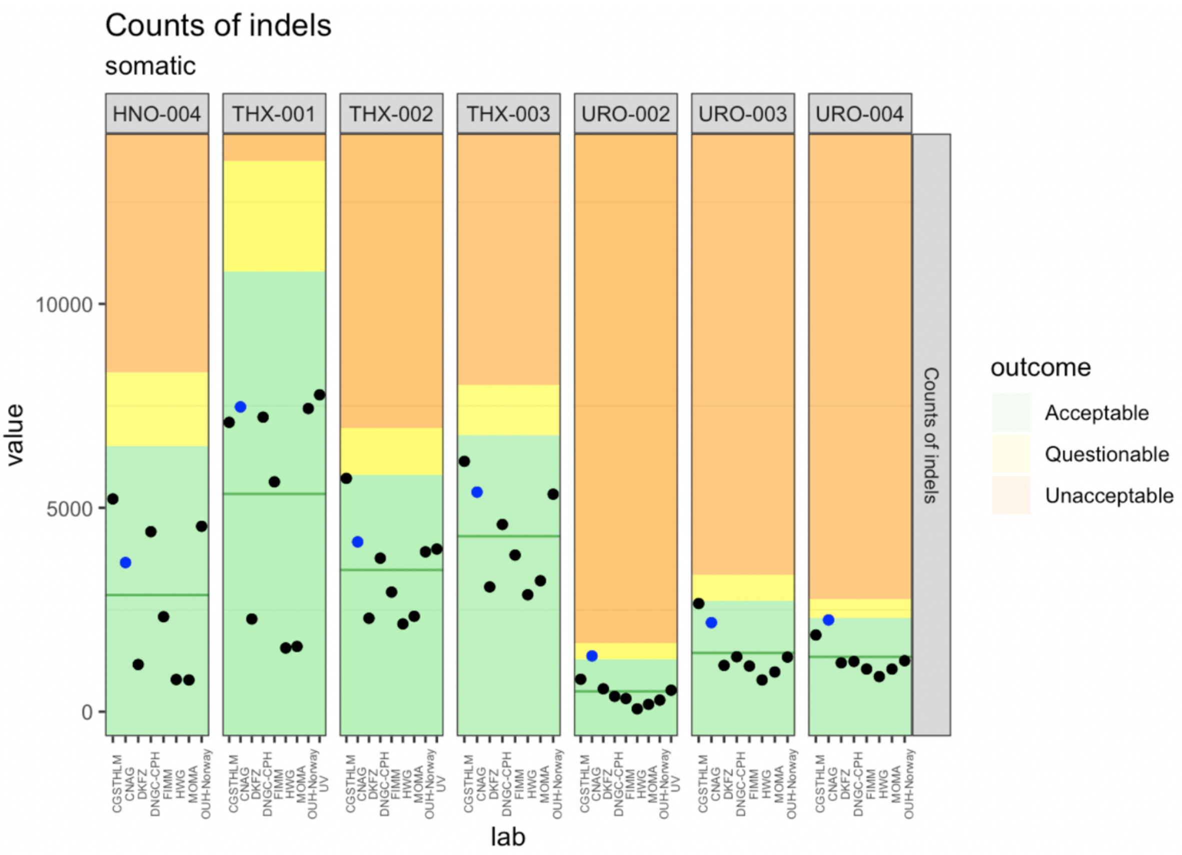

Figure 8: Counts of single nucleotide deletions and insertions (indels). The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis.

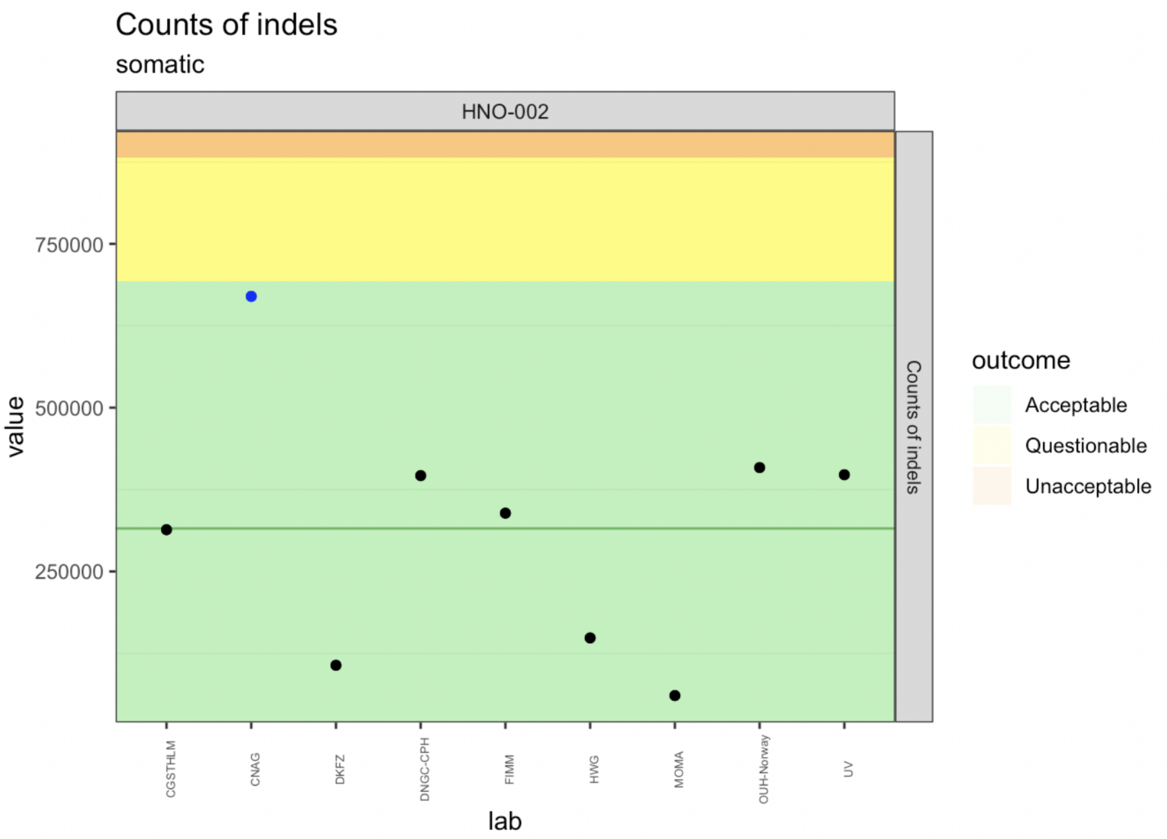

Figure 9: Counts of single nucleotide deletions and insertions (indels) for the HNO-002 sample. The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis.

In the counts of indels there is even higher dispersion of results than in the counts of SNVs. Figure 8 shows the counts of indels for all samples, except HNO-002. We had to represent these results in two separate figures because the scale of the counts is ten times higher in HNO-002, shown in Figure 9. Similarly to SNV counts, for indel counts results differ by lab but also by cancer type, with HNO (head and neck), and THX (lung) samples exhibiting higher variability than URO (renal) samples.

Number of somatic SVs (structural variants)

This QC metric is the number of passing structural variant calls. We count as structural variants break ends (BNDs) that map to two sites in the genome, thereby avoiding problems of BND ambiguous assignment and loose definitions of complex rearrangement events identified by different bioinformatics tools. For this metric there is no fixed indicative value.

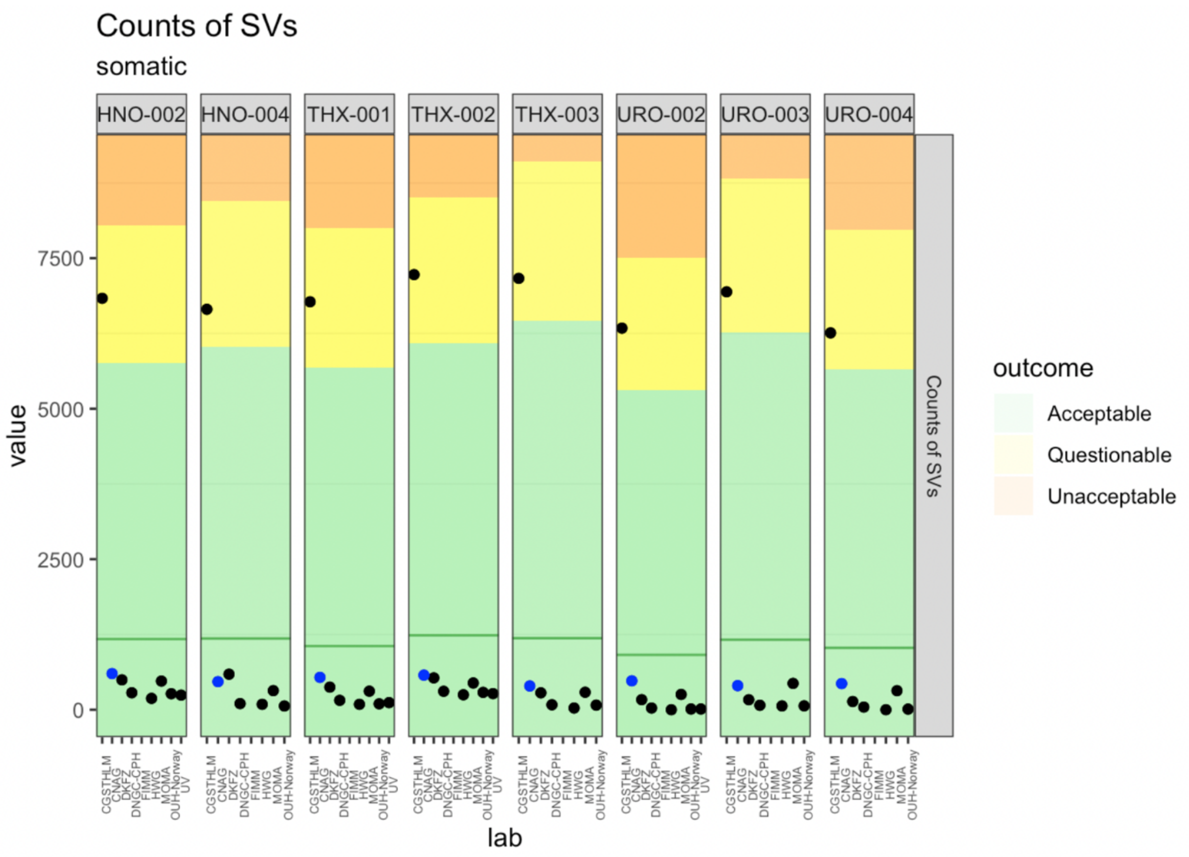

Figure 10: Counts of structural variants (BNDs). The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis.

The counts of structural variants (BNDs), show less variability than the counts of either SNVs or Indels. Nevertheless, it is important to mention that here we have only counted BNDs, and excluded more complex structural variants, such as caused by chromothripsis, or chromosome rearrangements. It is also necessary to understand how different tools define each type of event, to count only comparable types of structural variants. In these results, the same lab (CGSTHLM) appears as an outlier and that is because they used different tools to identify structural variants and reported the union of the variants called. It has become clear through these comparisons, that the bioinformatics pipeline differences have a large impact on the identification of structural variants.

Number of somatic CNVs (copy number variants)

This QC metric reports the number of passing copy number variant calls. Here we count as CNVs, duplications (DUP) and deletions (DEL, as reported by different bioinformatics tools. There is no fixed indicative value.

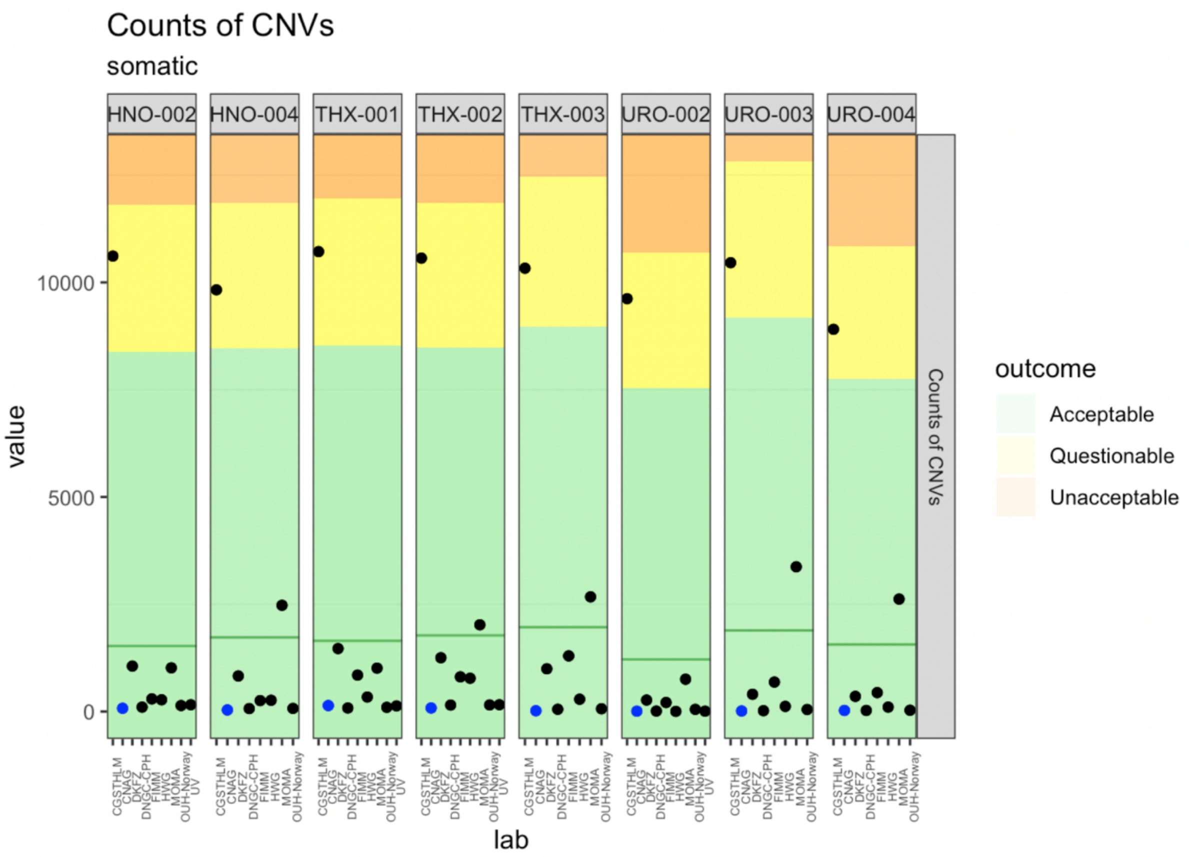

Figure 11: Counts of copy number variants: deletions (DEL) and duplications (DUP). The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis.

Because different tools define duplications and deletions differently, we observe some dispersion in the results, by lab. Here also, CGSTHLM appears as an outlier because they report the union of the calls produced by different bioinformatics programs.

5.2.1 Dry Lab Challenge somatic calling evaluation

To investigate the impact of the different bioinformatics pipelines on the quality of the somatic variant calls, independently of the sequencing and alignment stages, CNAG compared the results of the different pipelines ran by the participants. Each pipeline used a different somatic caller. We asked participants to submit a VCF for SNVs and indels, and a TSV for the structural variants (SVs) and copy number variants (CNVs).

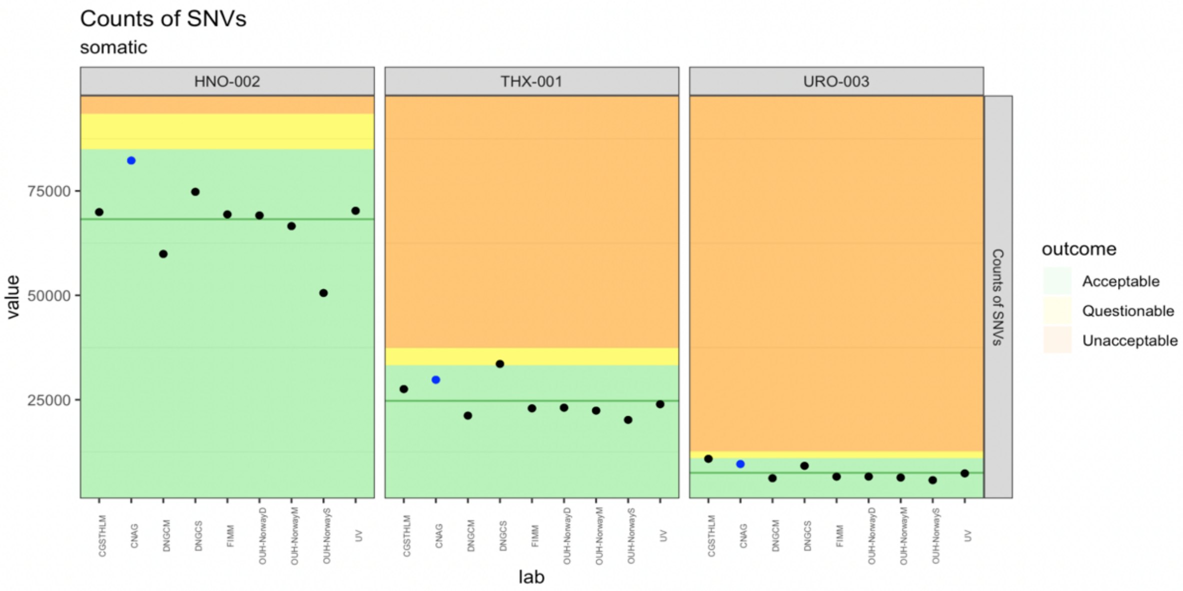

Figure 12: Counts of SNVs. The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis. The HNO-002 sample was the most variable in number of SNVs, THX-001 was an intermediate and URO-003 showed little variability.

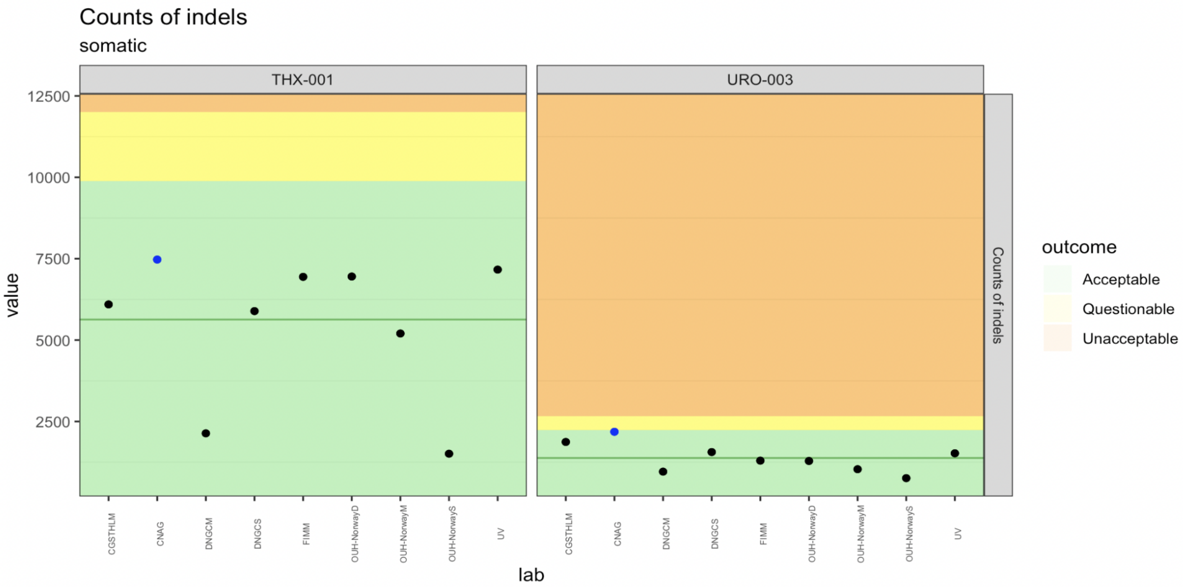

Figure 13: Counts of Indels for THX-002 and URO-003. The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis. THX-001 shows more variability in terms of indel count than URO-003, which showed little variability. Notably, the 3 labs running Dragen got comparable results (FIMM, OUH-NorwayD, and UV). Mutect and Strelka give similar results in 2 out of 3 pipelines that implement them (DNGCM, and OUH-NorwayS). DNGCM and DNGCS, ran by the same lab, differ by over 2500 indels.

Beyond One Million Genomes
B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

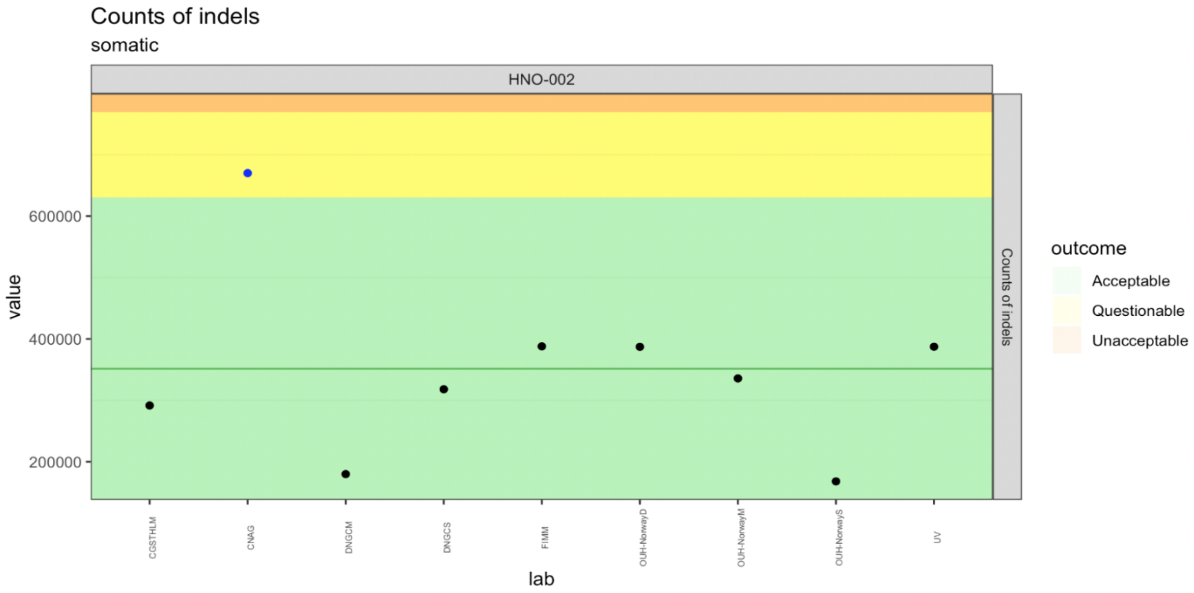

Figure 14: Counts of Indels for HNO-002. The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis. We had to make a different plot because HNO-002 has roughly 9 more times the amount of indels in comparison with THX-001 and URO-003. Notably, the 3 labs running Dragen got comparable results (FIMM, OUH-NorwayD, and UV). CNAG ran Mutect and Strelka jointly and it appears as an outlier at the top, in the questionable (yellow) area. Mutect and Strelka give similar results in 2 out of 3 pipelines that implement them (DNGCM, and OUH-NorwayS). DNGCM and DNGCS, ran by the same lab, differ by roughly 100,000 indels.

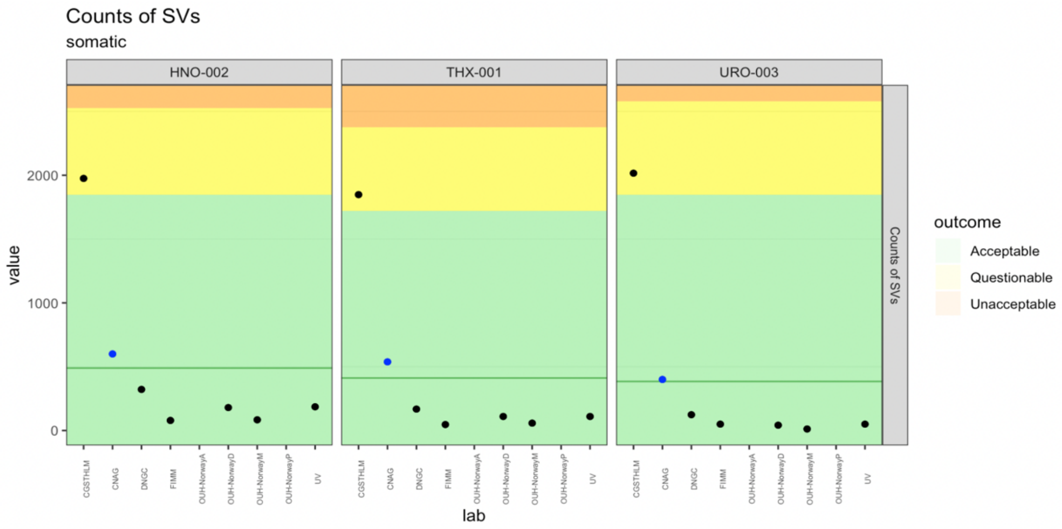

Figure 15: Counts of SVs. The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis. There is remarkably a large agreement in the count of SVs (BND), except for results by CGSTHLM that reported the union of different callers.

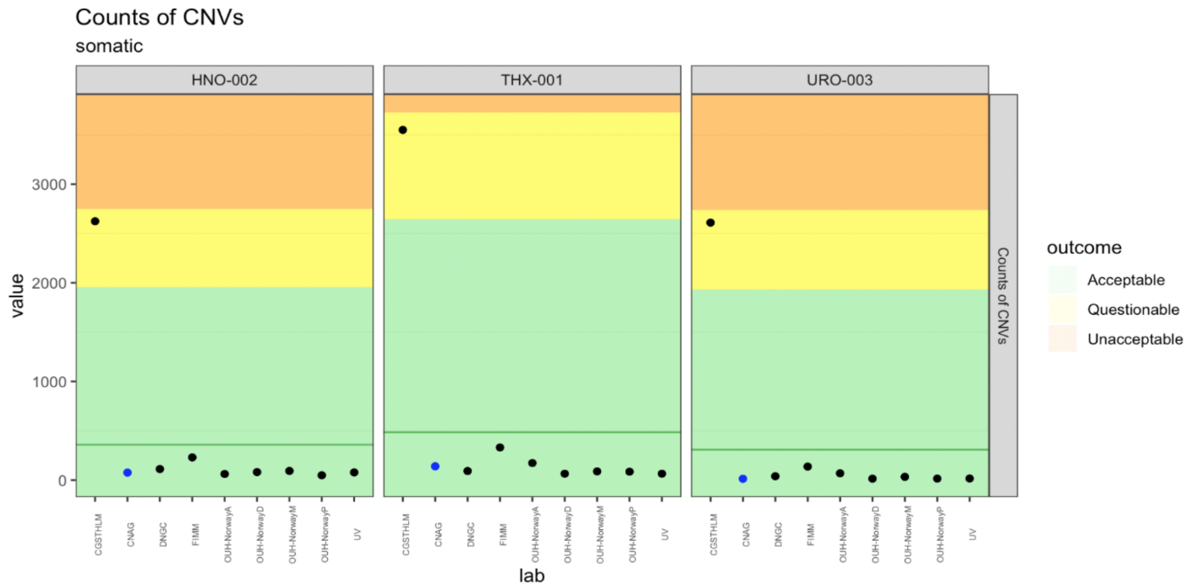

Figure 16: Counts of CNVs. The assigned value (green solid line) corresponds to the mean for all observations of somatic variants. The blue dots correspond to the CNAG values for this QC metric. The names of the participating laboratories are indicated along the x axis. There is also a large agreement in the count of CNVs (DEL, DUP), except for results by CGSTHLM that reported the union of different callers.

5.2.2 From merged FASTQ to multiple VCF comparisons

For the construction of goldsets, we asked participants of the 1+MG WG4 benchmark to merge all the FASTQs produced by all labs, per sample. Using the merged FASTQ participants ran different variant calling pipelines and returned to CNAG the resulting VCF and TSV files, for small and large variants, respectively. The merged datasets had an approximate coverage of 600X, potentially allowing for the detection of variants at low frequencies that are difficult to observe at commonly used depths of 80X. The effort of running bioinformatics pipelines with data at 600X of depth, was significant. Three labs were able to use the merge datasets and run somatic variant calling using Mutect, Dragen and Strelka. CNAG compared the variants predicted by the three pipelines to establish the set of variants predicted by: a) all three, b) by two, and c) exclusively by each method. From these comparisons, a reliable set of variants called by all 3 or by the intersection of 2 methods were extracted. Additionally, we manually inspected the challenging calls, hoping to identify true variants that can only be found with deep sequencing. For structural and copy number variants, the calls will be further validated by comparison with the ONT calls CNAG obtained for the 8 tumor/normal pairs.

The following plots show the comparisons of the calls using the merged dataset generated by CNAG (at ~600X) with the calls of each participant lab (at ~80X).

Comparison of Purity: Merged vs non merged datasets

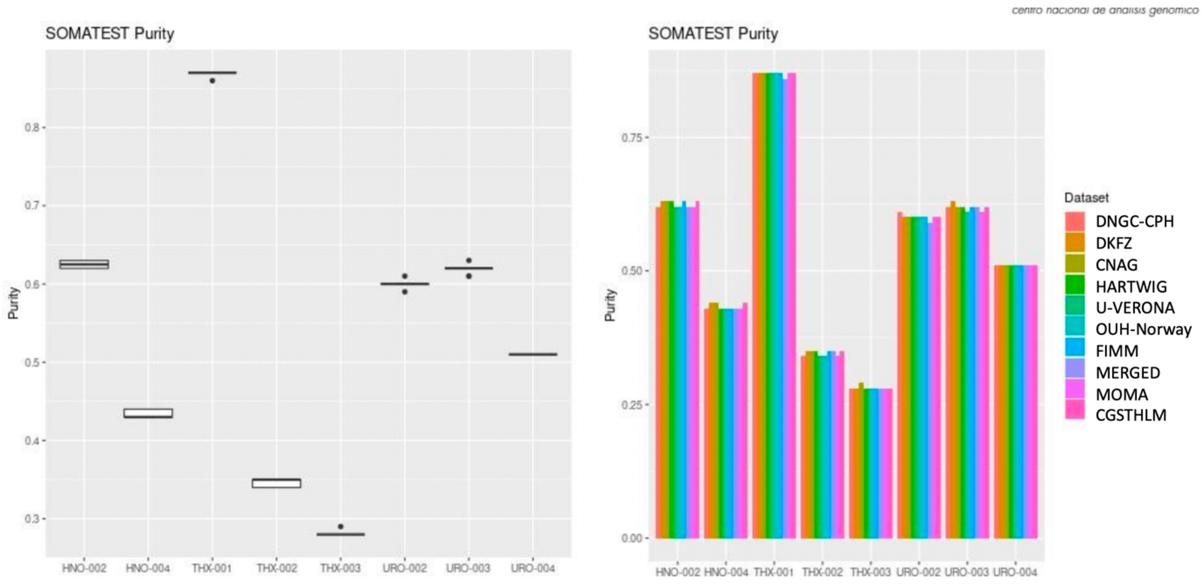

Figure 17: Purity of the tumor samples, for the merged dataset (purple), and individual lab datasets. On the left panel, the boxplots with the dispersion of the observations, and on the right panel, the bar plots with the purity estimates. The names of the participating laboratories are indicated along the x axis. The lowest purity level was found for sample THX-003, and the highest for THX-001. For all samples, the merged and non-merged datasets produced similar levels of purity.

Comparison of Ploidy: Merged vs non merged datasets

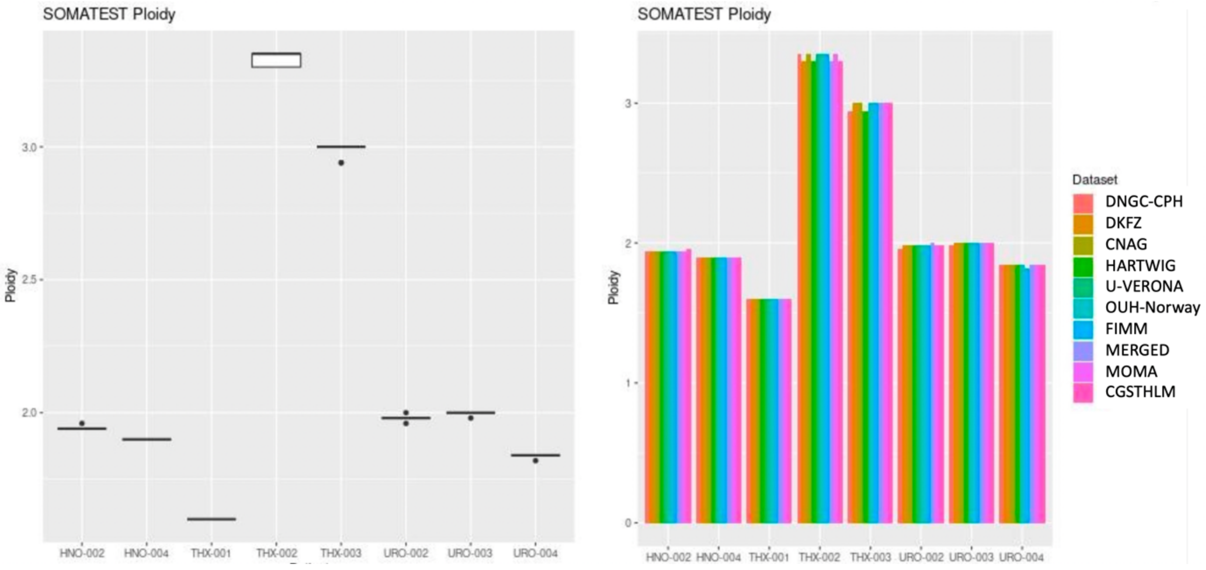

Figure 18: Ploidy of the tumor samples, for the merged dataset (purple), and individual lab datasets. On the left panel, the boxplots with the dispersion of the observations, and on the right panel, the bar plots with the ploidy estimates. The names of the participating laboratories are indicated along the x axis. Merged and non-merged datasets produced similar levels of ploidy. Samples THX-002 and THX-003 have ploidy > 2, and THX-001 has ploidy < 2.

B1MG has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No 951724

Comparison of somatic SNVs and Indels using merged datasets and 3 different somatic callers (Mutect, Strelka and Dragen)

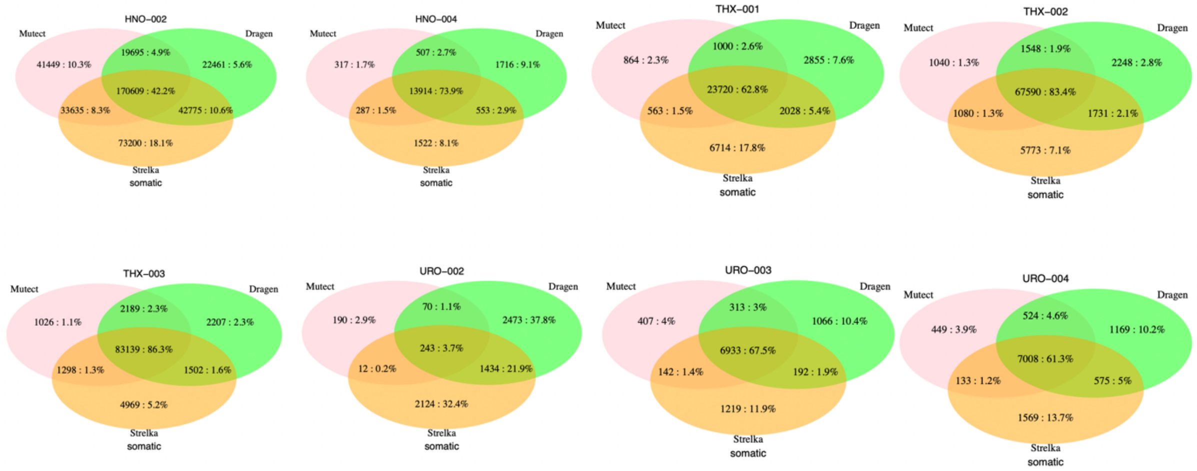

Figure 19: Intersection of the somatic variant calls (SNVs and Indels) predicted by Mutect, Strelka and Dragen using merged datasets (~600X) by different labs.

6. Discussion

6.1 Impact of sequencing protocol on the sequencing QC metrics

As noted in the Deliverable D3.2, despite small differences in sequencing and followed protocols, the results across centres do not deviate much from the mean and conform to expectations given by common practice and the literature. Moreover, the indicative values were good indicators of performance, and all observations fall near expected values. The differences in sequencing performance were not pronounced. At this stage, differences in sequencing machine and protocol do not appear to have a large impact on the quality achieved.

6.2 Impact of alignment protocol on the alignment QC metrics

The alignment QC metrics varied in the degree of conformity to expectations. The percentage of duplicate reads, the median insert size and the percentage of chimeras show a relatively even distribution of observations around the mean, and practically all results are in the acceptable z-score region. However, the mean coverage and the evenness of coverage metrics do show some variability around the mean, as well as some questionable observations. Because the percentage of chimeras and duplicate reads appear unbiased, variability in the mean and evenness of coverage may be better explained by differences in bioinformatic processing. Either by the choice of aligner or by specific parameters built into the tools. In any case, the differences in mean coverage observed among labs are small, and practically all samples achieve the indicative value of ≥ 30. In terms of evenness of coverage, tumor sample THX-001, has a mean of 0.95, the farthest from the indicative value of ≈1. This may indicate specific biases for this sample

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

that need to be further investigated. GC content biases could explain such deviations from a balanced coverage.

6.3 Impact of variant calling on the somatic variant QC metrics

Sequencing QC metrics for the somatic variants showed the most variability, compared to the sequencing or alignment stages. For both small (SNVs and Indels) and large (SVs and CNVs) type of variants, differences by type of cancer were apparent in all comparisons. These differences were more pronounced than differences by lab for a given sample. This is an encouraging indication that the variant calling reflects true biological signals, and not simply noise. It is clear, also, that the specific settings and tools used in the bioinformatic processing have a large impact on the results and likely explain to a large degree the observed differences.

For SNVs and Indels, the HNO (head and neck) and THX (lung) samples show higher dispersion of results than those of the URO (renal) cancer type. Clearly, sample HNO-002 has the highest number of indel variants, compared to all other samples. URO samples, on the other hand, have the lowest number of variants, especially once the PASS filters have been applied.

For SVs and CNVs, we observe less variability of results than for the SNVs and Indels but nevertheless, we note that the counts of these variants vary depending on the bioinformatics tools used. The differences stem from the definition each tool has about what constitutes a deletion and a duplication. This ambiguity is even more pronounced for structural variants. That is why we have limited the count to include only break ends that map in two sites of the genome.

Despite the sources of variability, it is remarkable that we see some agreement in the counts of SVs (BND) reported by different labs that used different variant calling tools. Counts of CNVs vary more and this variability may also be explained by the definitions of each event type built into different callers.

All participating labs achieve the same levels of purity and ploidy of the somatic variants. Samples THX-002 and THX-003 have the highest ploidy and THX-001 the lowest, suggesting possible structural changes relative to the normal. As seen in the counts of SVs, these samples have a slightly higher number than the HNO or URO samples.

6.4 Impact of the variant calling pipeline on the somatic variant QC metrics (Dry Lab Challenge)

We have observed that there is a striking difference in the number of the predicted SNVs and indels both by sample and, to a lesser extent, by pipeline. This is especially the case for the HNO-002 sample, that appears to have a very large number of indels, between 9 and 10 times more than samples THX-001 and URO-003. It appears that Dragen produces the most similar results, even when run by different labs. The most dissimilar results come from merging results by more than one caller, as in the case of CNAG. Mutect and Strelka run by different labs, do not appear to produce comparable results.

On the other hand, by limiting the structural variants we counted to BND events, the variability due to the different tools is not very big. Most callers give comparable results. For CNVs the

B1MG has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No 951724

similarity is even bigger. Manta, Ascat, Purple and Dragen were used to call the large variants. In this case, Dragen ran by FIMM seems to be a little bit off in the comparisons. Overall, we see more variability for structural than for copy number variants.

6.5 Variant call agreement between 3 tools using merged data

Dragen, Strelka and Mutect were used to call the variants of the merged datasets. The calls were compared to determine the level of agreement between the 3 methods. We observe a moderate level of agreement of ~60% on average across samples. The calls made by all three methods for the 600X merged data have a high level of confidence so they can be considered for the goldset. For the calls made by combinations of 2 methods, VAF plots were made to determine which calls are safe to be considered for the goldset and for all the difficult variants, for the rest of callsets we curated variants manually using the variant voter app developed by CNAG.

7. Conclusions

To this date, the 1+MG WG4 has orchestrated an exhaustive quality assessment encompassing every facet of the somatic whole-genome variant calling workflow. At each crucial juncture within this process, we have compiled the outcomes submitted by all participating laboratories and extracted pertinent quality metrics. The comparison of results across all labs has provided the baseline for the construction of a curated dataset of somatic variants with the highest reliability. This goldset establishes the standard of quality against which individual laboratory observations are measured.

The 1+MG WG4 has contributed with best practices for whole genome somatic variant calling through a comprehensive benchmark of quality metrics for all stages of the process. This work has also generated goldsets of somatic variant calls, for both small and large variants. In a larger framework, the 1+MG WG4 sets the quality requirements of genomic data for cross-border access and for personalised medicine practice.

8. Next steps

8.1 Curating challenging variants for the goldset

The variants called by all participants were used in the construction of the gold set. In the case of discrepant variants, CNAG developed an R shiny app variant voter tool, available online to all participants, where they can view screenshots representing discrepant variants and vote to validate or not the calls in question.

8.3 Final individual and general reports

ISO 17043 defines the requirements of the reports the participants will receive at the end of the benchmark. We have decided to make two reports: a general report and an individualised participant report. We will deliver the same general report to all participants. We will deliver a personalised report to each participant separately.

9. Impact

With the current proliferation of sequencing projects, a critical challenge arises: the absence of standardised procedures and essential quality assessment metrics to be followed by these initiatives. Notably, the sequencing of tumor DNA introduces its own unique set of technical and bioinformatics complexities. Additionally, the rapid integration of novel sequencing technologies into analytical pipelines, each tailored to the practices of different laboratories, adds even more heterogeneity. This underscores the pressing need to establish uniform practices for Next Generation Sequencing (NGS) and to mandate regular accreditation and validation of laboratory protocols. A pivotal initial stride in this direction involves launching a benchmarking initiative for somatic variant calling, assessing the efficacy of methods routinely employed across diverse laboratories. The findings from this task serve as a compass pointing toward optimal practices, enhancing reproducibility, and upholding the highest standards of reliability, particularly in the context of tumor sequencing. It is evident that the incorporation of more quality control metrics and a comparative analysis with a gold standard dataset will furnish a more comprehensive roadmap for implementing best practices at every juncture of the sequencing and variant calling process.

10. Appendix (Supplementary Tables and Figures)

Table 1. The 8 Tumour/Normal pairs used in the somatic benchmarking

Sample Name Tumour	Sample Name N	Cancer Type
URO-002-03	URO-002-17	Clear Cell Renal Carcinoma
URO-003-01	URO-003-03	Clear Cell Renal Carcinoma
URO-004-02	URO-004-06	Clear Cell Renal Carcinoma
HNO-002-07	HNO-002-01	Head & Neck Squamous-Cell Carcinoma
THX-001-06	THX-001-02	Lung Squamous-Cell Carcinoma
THX-002-07	THX-002-01	Lung Squamous-Cell Carcinoma
HNO-004-07	HNO-004-01_02	Head & Neck Squamous-Cell Carcinoma
THX-003-09	THX-003-04	Lung Squamous-Cell Carcinoma

Table 2. QC metrics for evaluation

	Metric	Method	Threshold	Indicative value
Sequencing metrics (T+N)	% bases with Q ≥ 30	Extract from Illumina SAV	Lower	R1, R2: ≥ 85%
	% PhiX alignment		Lower	R1, R2: 1%
	PhiX error rate		Upper	R1, R2: 0% (optimal)

	% passing filter (PF) clusters		Lower	100% (optimal)
	% phasing		Upper	R1, R2: 0% (optimal)
	% prephasing		Upper	R1, R2: 0% (optimal)
Alignment metrics (T+N)	% duplicate reads	Picard	Upper	< 10%
	Median insert size		Lower	> 300
	Mean coverage		Lower	Normal: ≥ 30; Tumour: ≥ 70
	Evenness of coverage		Both	1
	% chimeras		Upper	< 1%
Variant calling metrics (T+N)	Ti/Tv ratio	Picard	Both	≈ 2.0
	% callability	Custom script	Lower	> 95%
Somatic metrics	Tumour purity	FACETS	Both	N/A
	Number of somatic SNVs	Picard		

	Number of somatic INDELs			
	Number of somatic SVs	Calculated only for full pipeline challenge using the information from variant callers		
	Number of somatic CNVs			