Beyond One Million Genomes

D3.1

Quality metrics for sequencing

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 (CRG), Jeroen Belien (VUmc)		
Deliverable Lead Beneficiary	3 - CRG		
Deliverable	D3.1 - Quality metrics for sequencing		
Contractual delivery date	31/05/2021	Actual delivery date	28/05/2021
Delayed	No		
Authors	Ivo Gut, Lucia Estelles		
Contributors	Edwin Cuppen (HMF), Valtteri Wirta (KI), Eivind Hovig (UiO), Pim Volkert (Nictiz), Gert Matthijs (KU Leuven)		
Acknowledgements (not grant participants)			
Deliverable type	Report		
Dissemination level	Public		

Document History

Date	Mvm	Who	Description

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

01/02/2021	0v1	Ivo Gut (CRG)	[Initial draft circulated to WP participants for feedback]
13/05/2021	0v2	Nikki Coutts (ELIXIR Hub)	Version circulated to B1MG-OG, B1MG-GB & Stakeholders for feedback
27/5/2021	0v3	Ivo Gut (CRG) & Lucia Estelles (CRG)	[B1MG-OG/GB/stakeholder comments addressed]
28/5/2021	1v0	Nikki Coutts (ELIXIR Hub)	Final version uploaded to the EC Portal
23/06/2021	1v1	Xenia Perez, Juan Arenas & Nikki Coutts (ELIXIR Hub)	Template issue fixed. Reformatted final version resubmitted to EC and Zenodo

Table of Contents

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

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

1. Executive Summary

Next Generation Sequencing (NGS) is becoming increasingly used in clinical settings for the genomic analysis of germline and cancer samples. Hence, there is a need to establish guidelines that cover the minimum quality requirements for the generation of whole genome sequencing (WGS) and whole exome sequencing (WES) data. NGS pipelines are comprised of several elements, all of which contribute to the end quality of the result, from the reception of the samples to delivery of the outcomes. For this reason, quality control (QC) steps should be incorporated into the workflow to ensure that the data is fit for use, and its usage poses no risk to the patient.

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

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.	No
	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	No
	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.	No
	Infrastructure Key Results	
	9. Secure cross-border data access roadmap (WP4) by M12 & M36.	No

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

	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.	No
	5. Long-term sustainability design and funding routes for cross-border Personalised Medicine delivery (WP6) by M34.	No

3. Methods

Next-generation sequencing (NGS) is starting to be used in clinical diagnostics. Contrary to earlier technologies it interrogates far larger parts of a genome and does not only interrogate well established genome positions for presence or absence of a pathogenic event, but delivers all information related to genomic variation. It relies on producing unprecedented amounts of sequencing data in a single analytical test. It allows sequencing entire human genomes to identify pathogenic variants in the germline or somatic mutations in tumor. Alternatively, key parts of a genome can be selected using capturing technology, typically these are panels of genes. This even allows very cost-effective capture of the entire exonic (protein coding) part of the human genome. As the technology is still young, there are huge differences in how the technology is applied in different laboratories. How the technology is applied can have profound impact on a result and obviously downstream decisions. The objective of this deliverable is to establish a common set of metrics that can be used by laboratories across the 1+Million Genome Initiative so that data can be used without a need to go through the compute intensive initial steps of analysis.

An NGS analysis can be split into five parts, the input material, sample preparation, sequencing, data analysis and data interpretation. Each part has specific quality control procedures associated with it and ISO standards for the preparation of input material exist. However, in particular sample preparation, and data analysis are handled very differently by different laboratories. We are establishing a common set of metrics to improve harmonization across the 1+Million Genomes Initiative. In the beginning, we discussed examples from different large-scale laboratories, how they run their production process and assure quality throughout. This was followed with two rounds of a survey that went into a lot of detail for each possible measure that laboratories might collect. We have worked on comparing what kind of metrics different laboratories across the network apply. This information will be used to define the metrics for the assessment of the data that will come from an inter-laboratory comparison (ILC) that is currently

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

underway in EASI-Genomics and in which many of the B1MG/1+MG participants have engaged to participate.

The objective is to establish quality standards that can be applied widely. A second objective, as with any inter-laboratory comparison, is to help laboratories improve their procedures by exchanging with each other. The ILC will be split into the wet-lab (sample preparation and sequencing) and the dry-lab (bioinformatics analysis) part. From our experience so far, we believe that the guidelines for the wet-lab will be more flexible, while both parts will be guided by the outcome.

4. Description of work accomplished

4.1 Quality evaluation of NGS data

In this work, we have surveyed 22 laboratories across 13 European countries that participate in the 1+MG project (Figure 1). Most of these participants are hospitals and/or research organisations (Figure 2), where NGS is used mainly for cancer and rare genetic diseases (Figure 3). For cancer samples, WES is more used than WGS. Both WGS and WES are used for germline samples. Although most institutes perform NGS for both diagnostics and research purposes, few laboratories reported being accredited or following ISO standards (Figure 4).

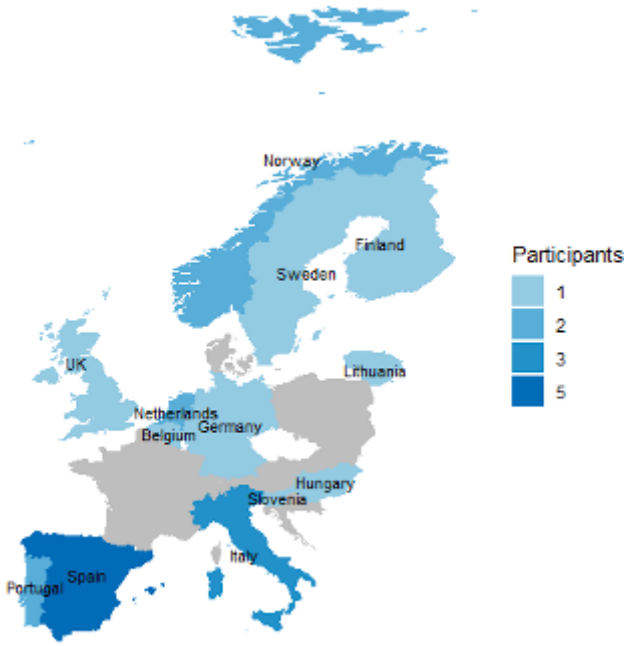


Figure 1. Participants of the survey: 22 laboratories across 13 countries.

We aim to understand how participants carry out clinical NGS protocols in their labs and how they address quality control (QC) of their samples in each step through the pipeline. Typically, NGS pipelines can be broken down into five successive activities, which are pre-analytical processing, library preparation, sequencing, data quality control and data analysis. We will

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

describe how laboratories undergo these steps for cancer and germline samples, and for WGS and WES workflows.

4.1.1 Pre-analytical processing

DNA extraction and purification or concentration steps are carried out in 10 of the surveyed labs for both germline and cancer. In contrast, high molecular weight DNA isolation is used by 6 of the labs for cancer and 5 for germline (Figure 5).

The most frequent measurement methods are based on gel electrophoresis (Bioanalyser), fluorimetry (Pico- and Ribo-Green) and spectrophotometry (Nanodrop) (Figure 6). Other methods include Agilent's DNA integrity number (DIN), Qubit, Agilent's Femtopulse and Dropsense.

Quantity, quality, and integrity are valued as good sample criteria (Figure 7).

4.1.2 Library preparation

Most participants who perform whole exome sequencing use in-solution capture kits rather than array-based methods (Figure 8). Several enrichment kits for WES and library preparation kits for WGS are commercially available and widely used. Among kits for WES for germline and cancer, Twist BioScience core exome and comprehensive WES, Roche Nimblegen SeqCap, IDT xGen, Agilent SureSelect and, Illumina TruSeq and Nextera can be found (Figure 9). For WGS library preparation, Illumina TruSeq DNA PCR-Free kit is the most widely used (Figure 10).

Quality metrics reported at this stage by participants are described in Table 1, where absorbance 260/280 is the most reported metric.

Table 1. Quality metrics for library preparation and recommended ranges, as reported by surveyed laboratories.

Metric	Recommended ranges
Absorbance 260/280	From 1.7 to 2.0
Absorbance 260/230	From 1.8 to 2.0
Insert size	Depending on the application (e.g., from 200 to 300 bp for WES cancer, or from 350 to 400 for WGS cancer)
High molecular weight (HMW)	Above 10 kb
DNA Integrity Number (DIN)	Thresholds from 3 to 7, depending on reporting lab

4.1.3 Sequencing

Illumina is the most used sequencing platform, particularly NovaSeq for both germline and cancer (Figure 11). Out of the participants who replied, Ion Torrent is used by 6 labs for cancer

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

and 3 for germline, PacBio by 1 for cancer and 3 for germline, Nanopore by 2 for cancer and 4 for germline, and 454 by 3 for cancer.

4.1.4 Data quality control

A range of tools is used for quality control, where FastQC[1] stands out (Figure 12). Other QC tools are VerifyBamID[2], Picard[3], MultiQC[4], and Fastp[5]. Frequent quality metrics for both germline and cancer include quality scores across all bases, sequence length distribution, average quality per read, percent of duplicated sequences, and percent of each base in the sequence (Figure 13).

4.1.5 Data analysis

More participants reported using a pipeline for germline than for cancer, with 11 for germline WES and 10 for germline WGS, compared to 7 for cancer WES and 4 for cancer WGS (Figure 14). In most of these cases, they run their own in-house pipelines, with some following the GATK best practices and a few that use or plan to use DRAGEN.

Findings are mostly validated using Sanger sequencing. Surveyed laboratories use the sequence QC, other NGS approach, large-scale comparisons versus internal datasets or trio analysis. In terms of metrics, laboratories rely on mean coverage, evenness of coverage, percent of paired reads mapping to different chromosomes, the ratio of edits between paired reads and somatic mutation calling coverage in the case of cancer (Figure 15). Only between 3 and 5 of the surveyed laboratories for cancer WES, germline WES, cancer WGS and germline WGS provide interpretation (Figure 16).

4.2 Evaluation of NGS legacy data

Working with data sequenced over several years can prove challenging. A good understanding of its source and quality is necessary beforehand to establish whether it is viable for the intended usage. A preliminary evaluation may be performed by using a set of defined metrics for legacy data. Whalley et al.[6] developed a star-based rating system that utilizes five quality control metrics for cancer WGS:

- Mean coverage, defined as the number of reads covering each position excluding N-ref bases, low quality reads and duplicate reads.
- Evenness of coverage may be calculated as the ratio of the median coverage over the mean coverage or as the variation of normalised coverage in 10 kb genomic windows.

[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[2] Zhang F, Flickinger M, Gagliano Taliun SA, et al. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res*. 2020;30(2):185-194. doi:10.1101/gr.246934.118
[3] Broad Institute. Picard Toolkit. 2019. http://broadinstitute.github.io/picard/.
[4] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016. doi:10.1093/bioinformatics/btw354
[5] Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. In: *Bioinformatics*. ; 2018. doi:10.1093/bioinformatics/bty560
[6] Whalley JP, Buchhalter I, Rheinbay E, et al. Framework for quality assessment of whole genome cancer sequences. *Nat Commun*. 2020;11(1):1-8. doi:10.1038/s41467-020-18688-y

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

The latter method requires correction for GC-dependent coverage bias using the ACEseq algorithm[7] for somatic copy number variants.

- Somatic mutation calling coverage, which measures how much of the cancer genome is covered to call a somatic mutation. MuTecT[8] calculates if enough coverage is present in both the tumor and normal, that is 14x per tumor and 8x in the matched normal.
- Paired reads mapping to different chromosomes, given that an excess of these reads may be due to technical artefacts.
- Ratio of difference in edits between paired reads, calculated as the ratio of mismatches in read 1 and read 2 across the whole dataset.

In addition to these five metrics, an additional measurement to consider when evaluating quality for cancer data is tumor purity. The tool ABSOLUTE[9] estimates tumor purity and ploidy by using copy number data, statistical models of recurrent cancer karyotypes, and allelic fractions if available.

[7] Kortine Kleinheinz; Isabell Bludau; Daniel Hübschmann; Michael Heinold; Philip Kensche; Zuguang Gu CLMHWKPMIVRWIM-S project; BBRSREMS. ACEseq – allele specific copy number estimation from whole genome sequencing. *bioRxiv*. 2017. doi:https://doi.org/10.1101/210807
[8] Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219. doi:10.1038/nbt.2514
[9] Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol*. 2012;30(5):413-421. doi:10.1038/nbt.2203

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

5. Results

Summary of metrics for NGS QC evaluation

Table 2 summarises metrics used for NGS QC evaluation, with consideration of the measurements recommended by the Medical Genome Initiative consortium[10]. These metrics may include filters for non-N reference positions, base quality (BQ ≥ 20), coverage or reads mapping quality (MAPQ ≥ 10).

Table 2. Metrics for NGS QC evaluation.

	QC metric
Pre-analytical processing	DNA quantity as measured by gel electrophoretic methods, fluorometric methods, and spectrophotometric methods
Library preparation	Library quantification by OD measurement Insert size determination
Sequencing	Yield related metrics • Total bases sequenced • Bases sequenced with Q ≥ 30 • Percent of Passing Filter (PF) bases
	Read related metrics • Mean or median sequence length (or other metrics derived from sequence length) • Percent of each base in the sequence • Percent of N bases in the sequence
	Base quality metrics • Percent bases with Q ≥ 30 • Mean or median base quality per read • Statistics on Q-Scores across all bases
Data quality control and data analysis	Coverage related metrics • Mean or median autosome coverage • Coverage at a specific depth and/or MAPQ • Coverage at specific genomic regions • Coverage in GC regions • Evenness of coverage • Mean coverage per chromosome • Ratio between mean coverage on chromosomes over autosome mean coverage • Percentage of ontarget bases sequenced among all bases

[10] Marshall CR, Chowdhury S, Taft RJ, et al. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. *npj Genomic Med*. 2020;5(1):47. doi:10.1038/s41525-020-00154-9

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

	Alignment metrics • Percent of PF reads aligned, with or without MAPQ filters • Percent of paired reads mapping to different chromosomes, with or without MAPQ filters • Percent of duplicate reads • Percent of adapter dimer reads • Percent of mismatches in read 1, read 2, or both • Ratio of edits (mismatches between read 1 and read 2) • Mean or median insert size, or fragment length (or other metrics derived from insert size distribution)
	Variant calling metrics • Percent callability • Percent callability in specific regions • Number of variants • Number of multiallelic variants • Number of variants per chromosome • Number of SNPs • Number of INDELs • Number of SVs • Number of CNVs • Percent SNV changes (i.e.: %C > T) • Transitions/transversions ratio • Heterozygous/homozygous ratio • DP or GQ distribution derived metrics • INDELs length distribution derived metrics • Variant impact derived metrics
	Cross-individual contamination
	PhiX metrics • PhiX error rate, overall • PhiX error rate by lane
	Somatic metrics • Somatic mutation calling coverage • Tumor purity • Tumor ploidy • Number of novel variants • Alternative allele distribution derived metrics
Legacy data	Mean coverage
	Evenness of coverage
	Somatic mutation calling coverage (cancer only)
	Percent of paired reads mapping to different chromosomes
	Ratio of edits (mismatches between read 1 and read 2).
	Tumor purity (cancer only)

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

We developed a second questionnaire based on Table 2, to understand which QC metrics are more broadly used among participants and if they compute other additional metrics to evaluate NGS data. Six institutes provided detailed feedback (Table 3). We will describe the findings of the survey below. Figures can be found on Appendix II.

Table 3. Institutes that provided feedback about QC metrics and type of NGS data that their laboratories process.

Organisation	WES cancer	WES germline	WGS cancer	WGS germline
ARC-Net Applied Research on Cancer centre at the University of Verona, Italy			x	x
Clinical Institute of Genomic Medicine at the University Medical Centre of Ljubljana, Slovenia		x		x
Instituto Nacional de Saúde Doutor Ricardo Jorge, Portugal		x		
National Institute of Oncology, Hungary		x		
SciLifeLab, Sweden	x	x	x	x
Università Cattolica del Sacro Cuore, Italy		x		
Total	**1**	**5**	**2**	**3**

5.1 Library preparation

Insert size determination and library quantification by OD measurement are performed by most surveyed laboratories, where the former was more widely reported (Figure 17). One participant reported measuring libraries before capture with fluorimetry (Qubit) and electrophoresis (Fragment Analyzer), and quantifying with Qubit only after capture. Another participant performs OD measurement before capture and OD measurement with sizing after capture.

5.2 Yield metrics

Total bases sequenced and total bases sequenced with Q ≥ 30 are used by all the surveyed institutes. The percent of passing filter (PF) bases is similarly used (Figure 18). Other metrics evaluated are the percent of Q30 per read, including index reads, and deduplicated reads with acceptable mapping quality.

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

5.3 Read metrics

The percent of each base in the sequence is broadly evaluated, while the percent of N bases in the sequence is less common. Metrics derived from sequence length, like the mean and median, are employed as well (Figure 19). The percent of phasing and pre-phasing per read was reported in addition to the metrics in the questionnaire.

5.4 Base quality metrics

Percent of bases with Q ≥ 30 is the most broadly used metric in this category. Average base quality per read is also used; the median value is less common though (Figure 20). Additionally, a participating lab reported evaluating Phred scores per position in the read.

5.5 Alignment metrics

Percent of PF reads aligned (commonly without MAPQ filters), percent of duplicate reads, and mean or median insert size (or fragment length) are the most assessed alignment metrics (Figure 21).

5.6 Coverage metrics

Mean and/or median autosome coverage are common metrics to assess coverage. A subset of the surveyed participants computes coverage by chromosome and/or mitochondria. Coverage at specific regions is computed for ACMG genes, specific disease sets, regions based on tumor type, and relevant targets. Some laboratories report coverage at a specific depth and/or MAPQ, although most of them did not provide details. Coverage in GC regions is also used; a laboratory achieves this by plotting coverage vs percent GC. Evenness of coverage is used by a few surveyed labs (Figure 22).

The percent of on-target bases sequenced among all bases is used by all participants who perform exome sequencing of germline variants (WES germline).

5.7 Variant calling metrics

All surveyed laboratories consider the total number of variants as a QC metric, and a subset of them additionally evaluate the counts of SNVs, INDELs, SVs and CNVs separately. The heterozygous/homozygous and transitions/transversions ratio are also widely considered (Figure 23).

One surveyed laboratory reported using specific genomic regions for callability based on the tumor type.

5.8 Cross-individual contamination

Four institutes reported measuring cross-individual contamination (Figure 24). One of them performed this by estimating the profile of heterozygosity.

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

5.9 PhiX metrics

Some surveyed laboratories report using PhiX, and some do not use PhiX. Laboratories that use PhiX usually evaluate the error rate (Figure 25).

5.10 Somatic metrics

Only one participating laboratory responded to this question. They evaluate tumor purity, ploidy, the somatic mutation calling coverage, and the alternative allele distribution (Figure 26).

6. Discussion

We have established a picture of quality metrics that are used across many laboratories of the 1+Million Genomes Initiative. Metrics can cover all five stages of an NGS analytical procedure, input DNA, sample preparation, sequencing, data analysis and data interpretation. For this deliverable the focus is on sample preparation and sequencing together and data analysis. For the needs of the 1+Million Genomes Initiative two angles are important, whether the quality metrics are used to retrospectively vet a dataset, or whether metrics are used to establish the best possible standard operating procedure for a laboratory. For retrospective evaluation of whole genome sequencing applied to cancer (tumor-normal comparison) for the identification of somatic mutations we could build on work done by some of the participants of the 1+MG WG4 in the PanCancer study of the International Cancer Genome Consortium. In that instance a set of 5 salient metrics were established that allow identifying whether a dataset is of sufficient quality to be included in a study or not (without having access to laboratory metadata). With this any legacy dataset can be evaluated. For data that is newly generated, the same five metrics can be applied, though other parameters can be added. In cancer studies the cellularity of cancer cells in a sample are important, because this information can be used to determine thresholds of somatic mutations detection.

We have also established that when the entire process is under the control of the data user, substantially more metrics can be collected. Through the survey of the participating laboratories we have collected all values that are collected. For the inter-laboratory comparison, that is currently being launched, in the EU-funded project EASI-Genomics and in which several of the B1MG WP3 laboratories will be participating a subset of these metrics will be collected in order to establish thresholds as a guideline.

7. Conclusions

Establishing metrics is of great importance. 5 metrics to assess a tumor-normal whole-genome dataset retrospectively now exist and these can be taken for post hoc vetting of a dataset. At the same time they can be applied for newly generated data. Through the work of the 1+MG WG4 we have collected information on all of the metrics that are collected at different stages of the sequencing process. In a benchmark currently carried out in the EU-funded project EASI-Genomics we will collect many of the metrics that have been put forward. This will allow a classification of the measures that are determining for the quality of the outcome.

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

8. Next steps

Our next step is collecting the data and metadata from the EASI-Genomics WGS cancer benchmark. This benchmark has been slowed down by the lockdowns, as collecting tumors from cancer surgeries is more difficult. We are also in the preparation of the germline benchmark. This consists of writing a specification that can be used in the application to the local ethics committee that will underwrite the donor sample collection.

9. Impact

Evidently, salient guidelines for the quality of NGS analyses are needed by all laboratories and in particular for ambitious initiatives like the 1+Million Genomes they are essential. Only data that has been generated exceeding common and high standards will be useful for this initiative. Without the outcomes will be fraught with many false positive observations. The value of the entire system depends critically on the quality of the baseline information.

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

10. Appendix I: Additional survey figures

Table 4. Participants of the survey, 13 countries and 22 centers.

Country	Organisation
Belgium	KU Leuven
Finland	FIMM Sequencing Unit (Institute for Molecular Medicine Finland, University of Helsinki)
Germany	German Cancer Research Center
Hungary	National Institute of Oncology
Italy	ARC-Net
Italy	San Camillo-Forlanini hospitals
Italy	University of Verona
Lithuania	Vilnius University Hospital Santaros Klinikos
Netherlands	Hartwig Medical Foundation
Netherlands	Radboudumc
Norway	Oslo University Hospital
Norway	Telemark Hospital Trust
Portugal	Instituto Nacional de Saude Dr Ricardo Jorge
Portugal	University of Aveiro
Slovenia	Clinical institute of Genomic Medicine, University medical centre Ljubljana
Spain	Carlos III Institute of Health
Spain	CNIC
Spain	Fundacion Publica Galega de Medicina Xenomica
Spain	Hospital Universitario Ramon y Cajal
Spain	ISCIII
Sweden	Clinical Genomics, SciLifeLab
United Kingdom	Genomics England

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

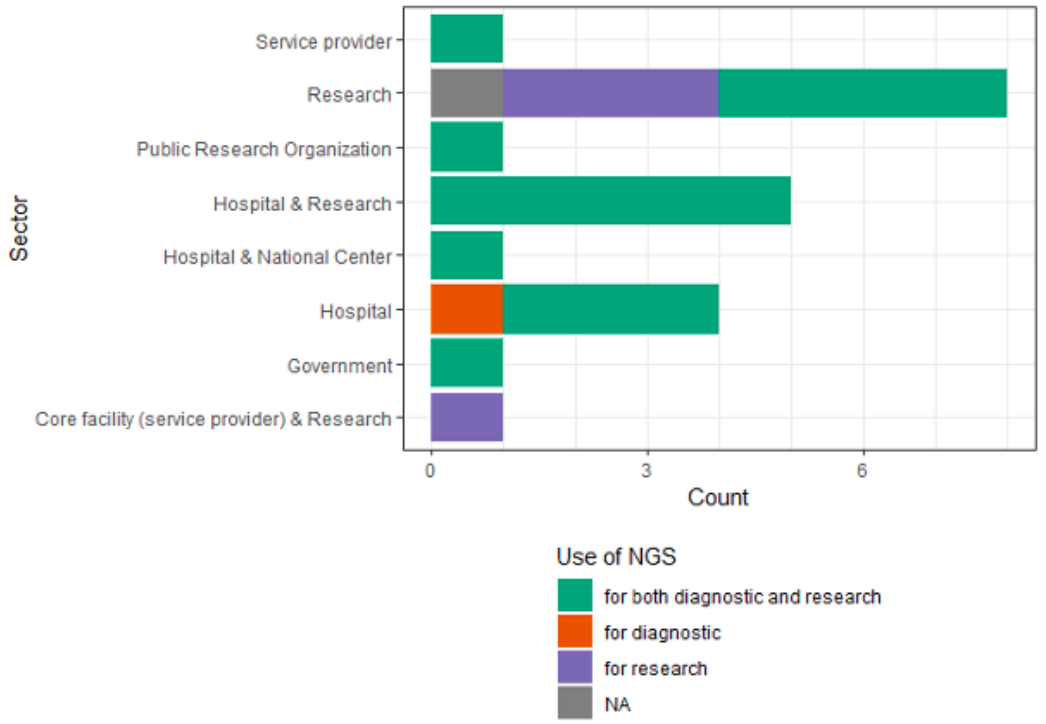
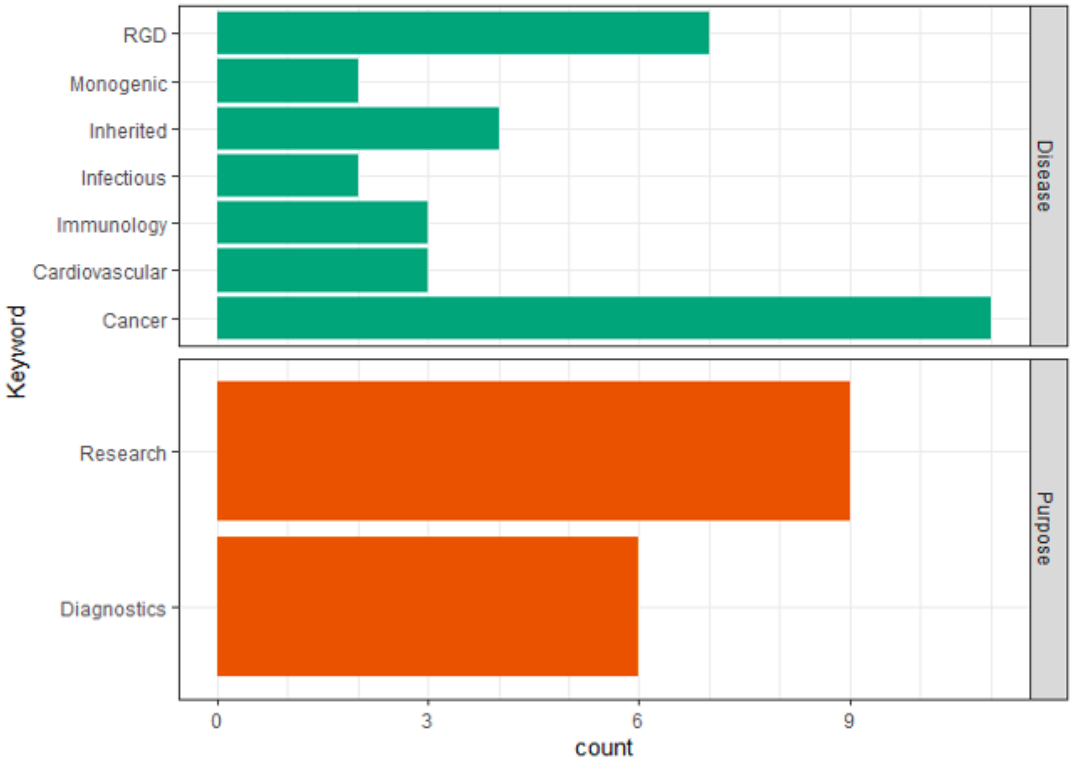


Figure 2. Sector and usage of NGS of surveyed participants.



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

Figure 3. Disease analysed by survey participants and purpose, which were reported as free text.

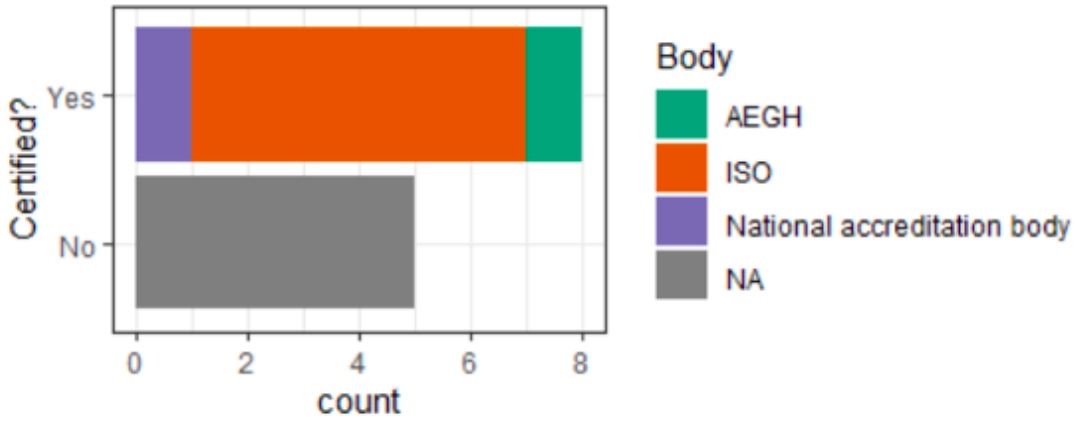


Figure 4. Certification or accreditation status of participants.

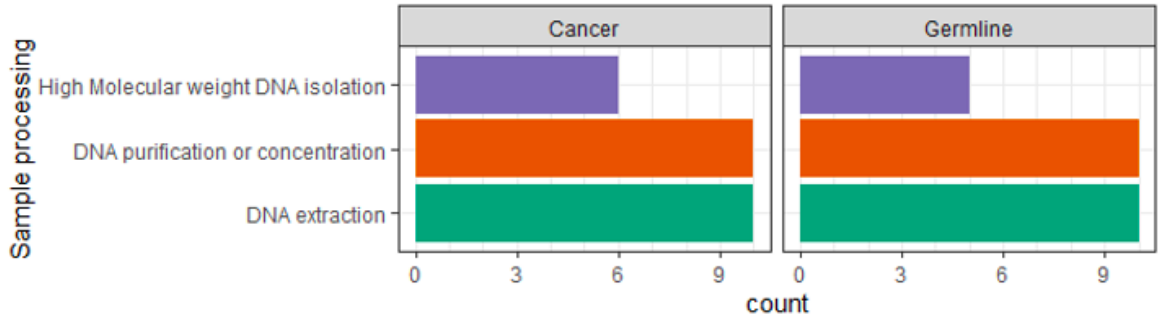


Figure 5. Pre-analytical sample processing performed by participants.

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

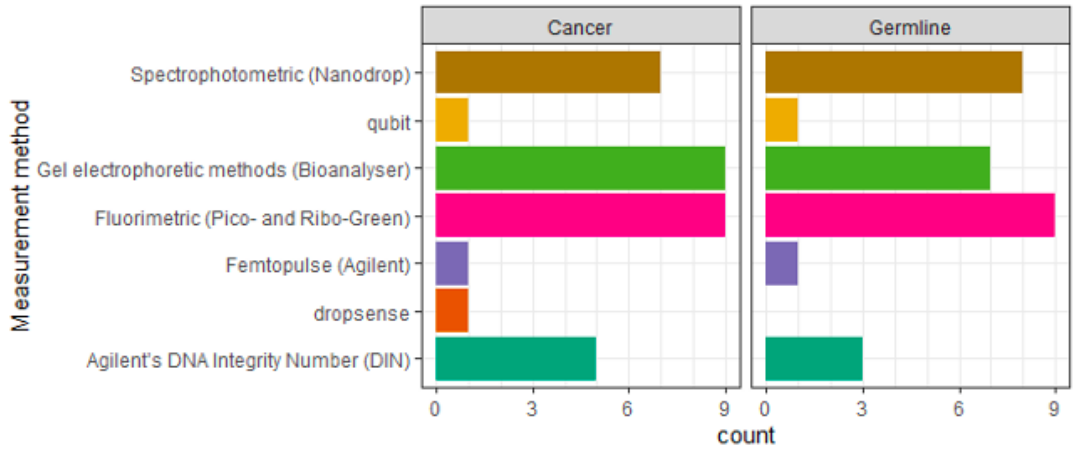


Figure 6. Measurement method for pre-analytical sample processing performed by participants.

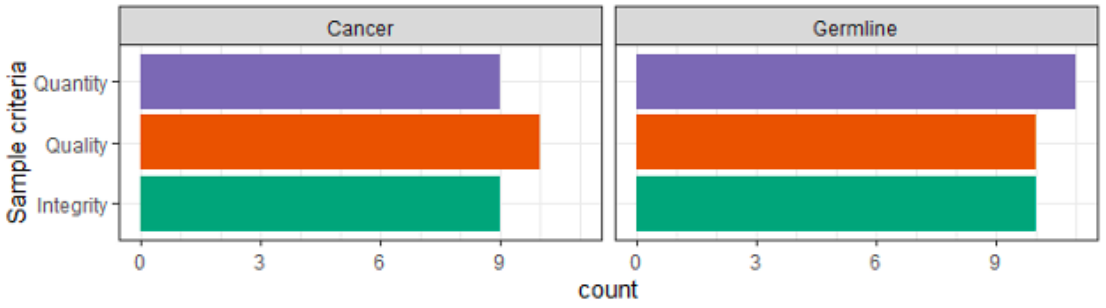


Figure 7. Sample criteria used by participants

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

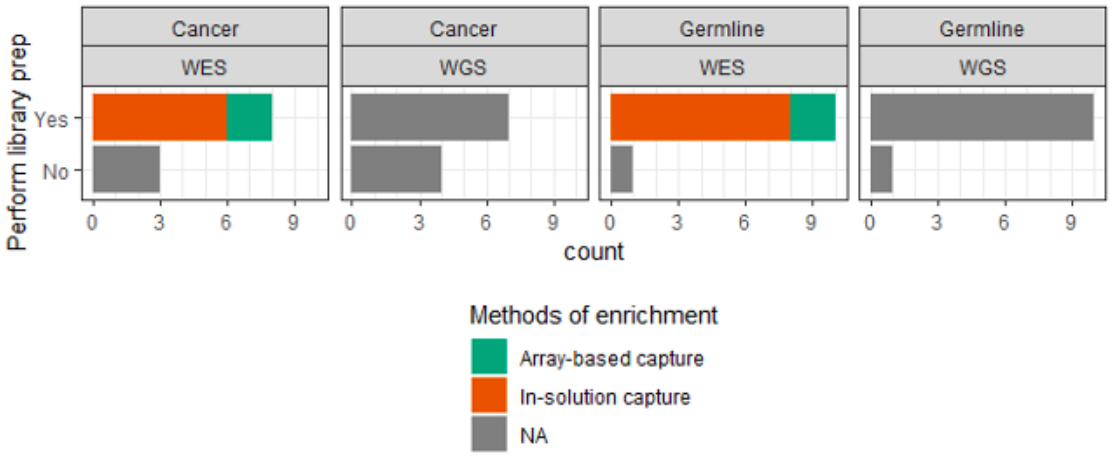


Figure 8. Participants that perform library prep and methods of enrichment used for exome sequencing.

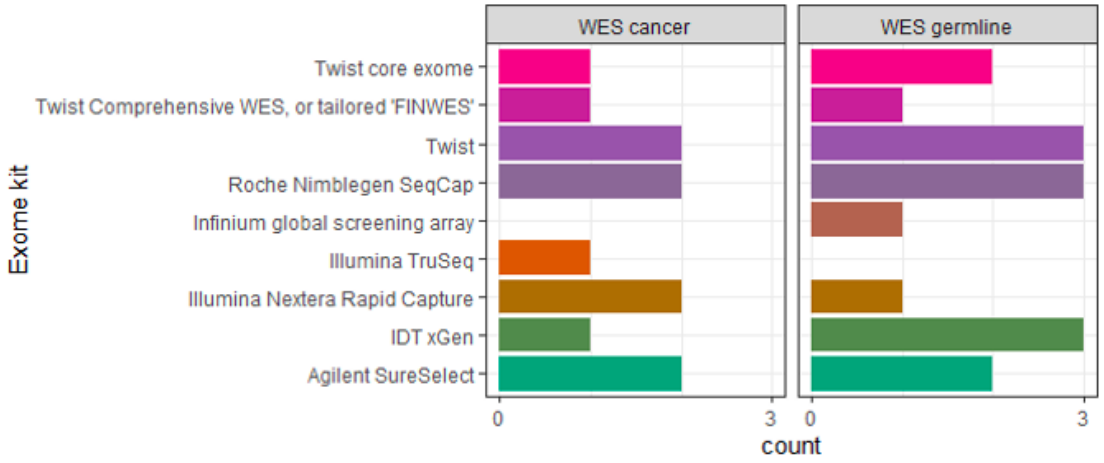


Figure 9. Kits used for exome sequencing.

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

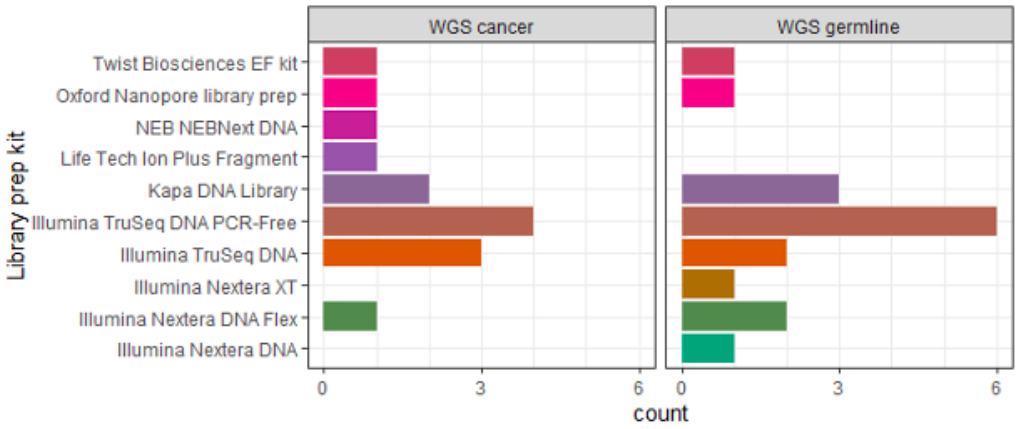


Figure 10. Kits used for WGS library preparation.

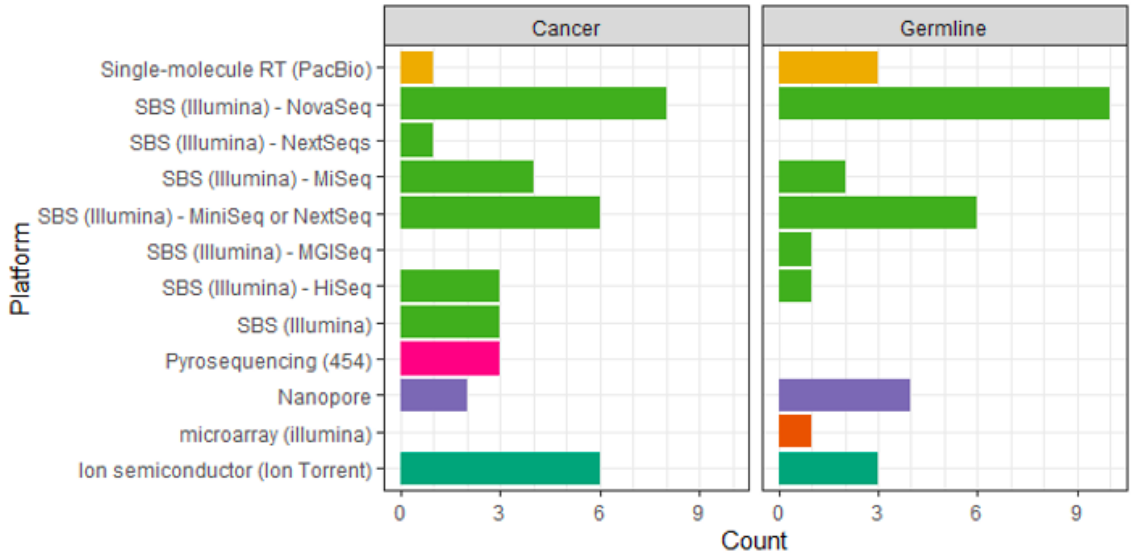


Figure 11. Sequencing platforms.

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

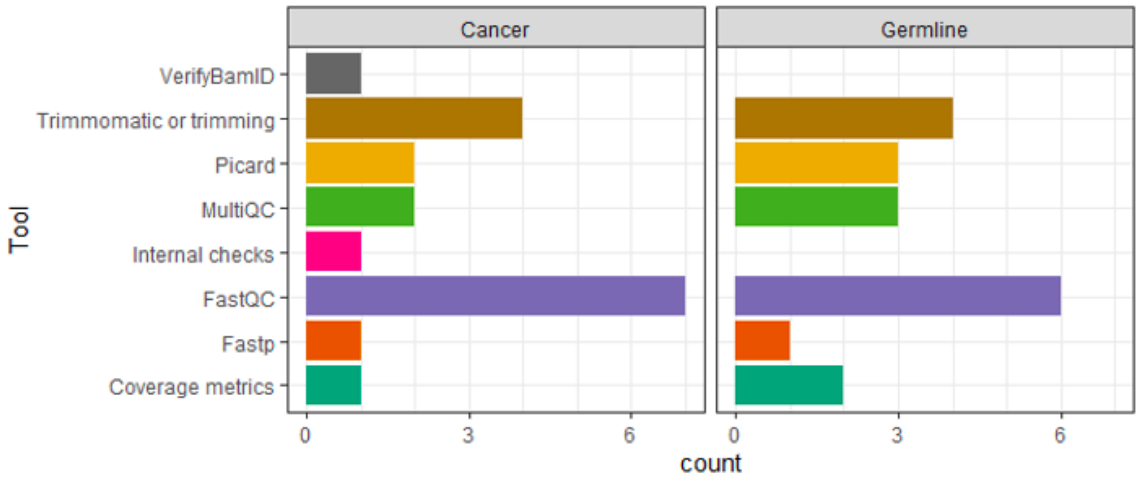


Figure 12. Tools for NGS QC.

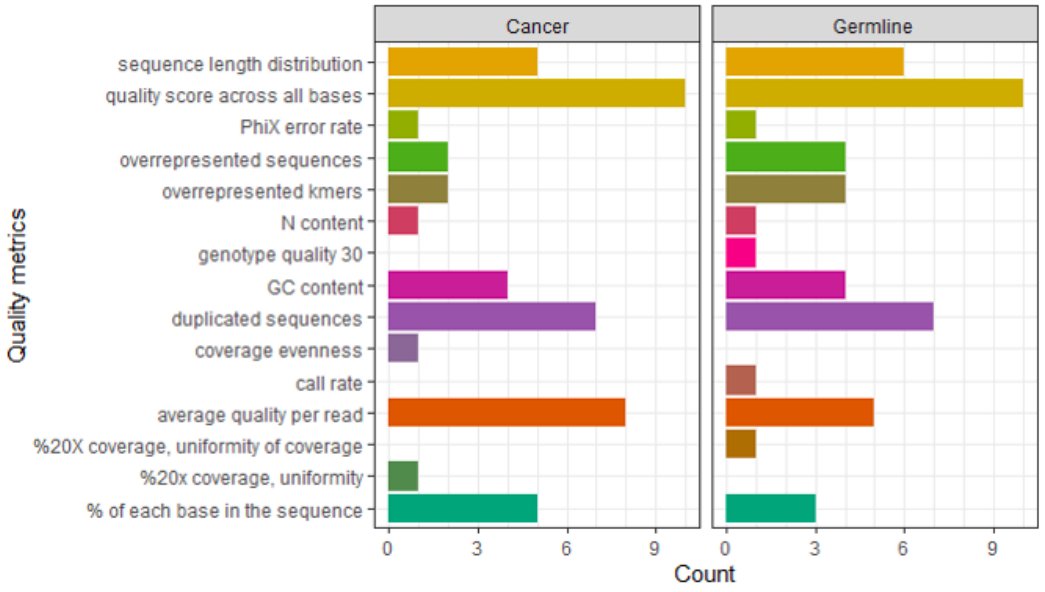


Figure 13. Quality sequencing evaluation metrics.

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

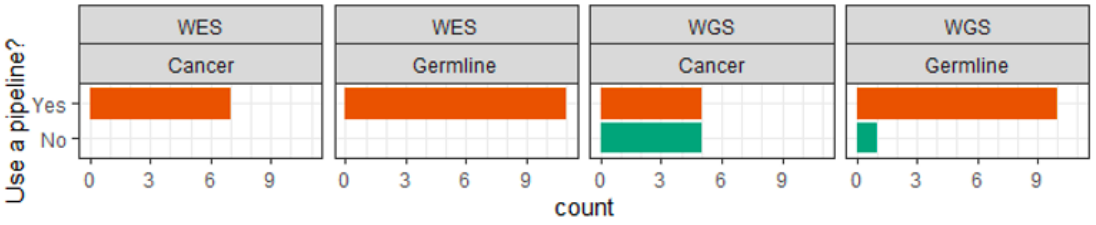


Figure 14. Participants who use an analysis pipeline.

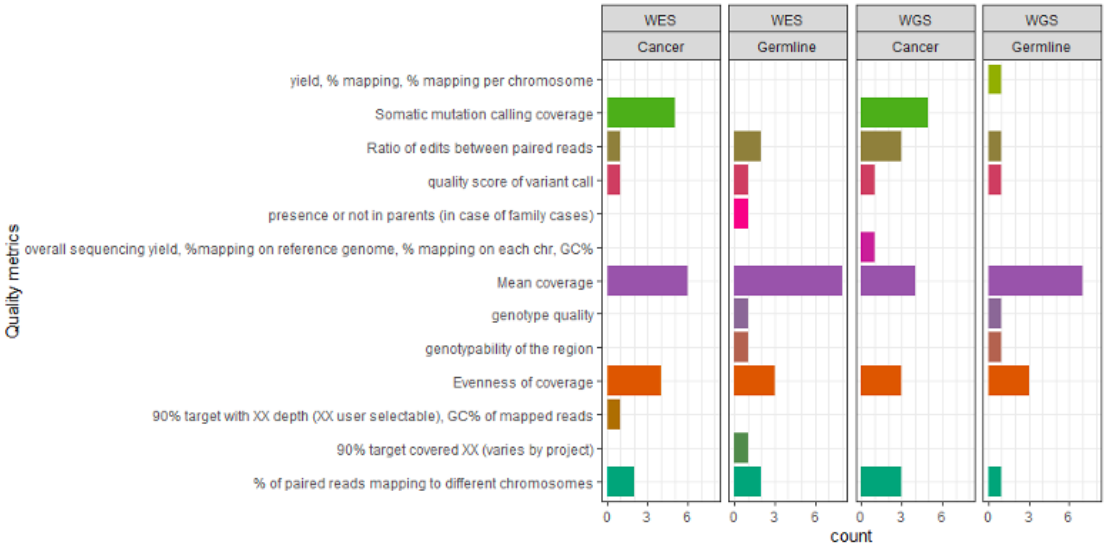


Figure 15. Quality criteria used to confirm findings.

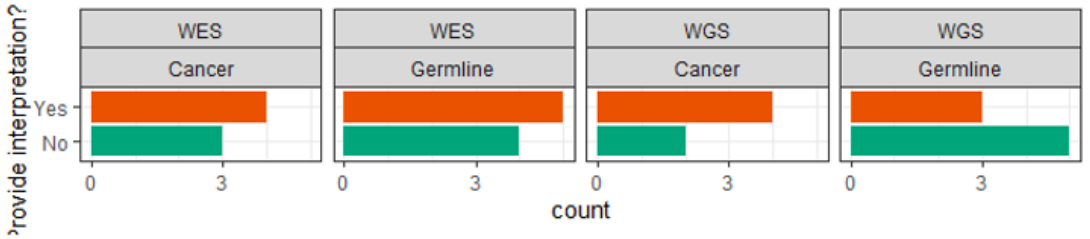


Figure 16. Participants that provide interpretation.

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

11. Appendix II: QC metrics survey figures

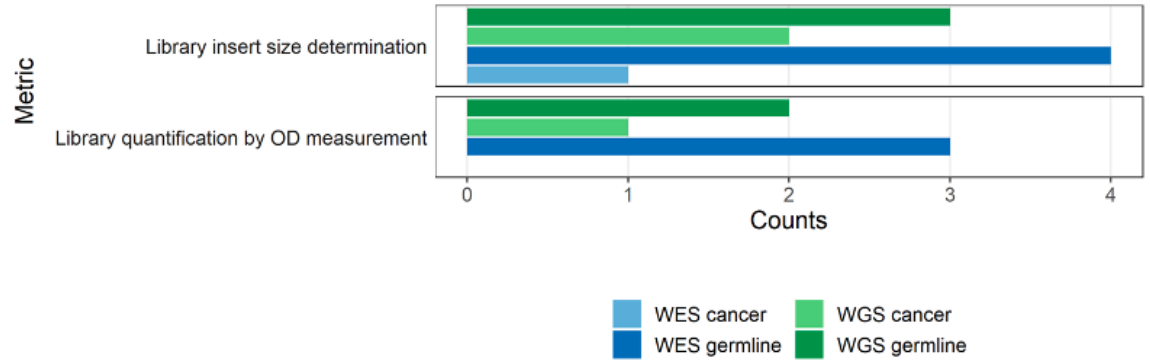
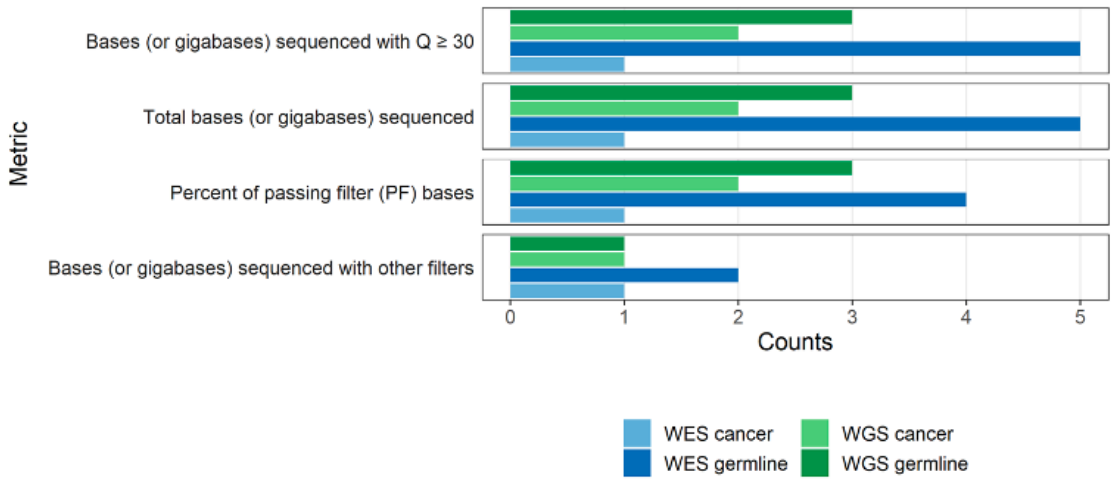


Figure 17. Library preparation metrics measured by 6 surveyed laboratories, where all of them answered this question



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

Figure 18. Yield metrics measured by 6 surveyed laboratories, where all of them answered this question.

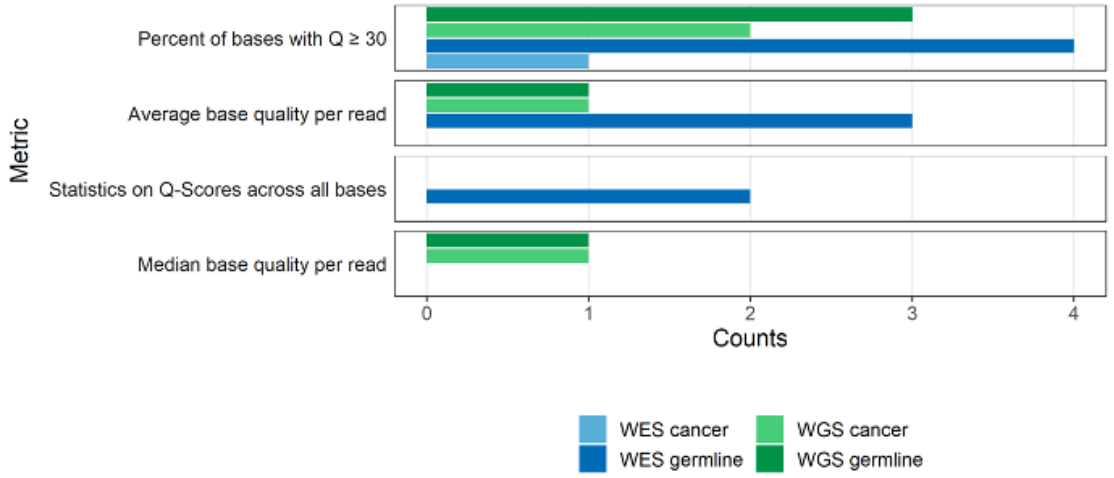


Figure 19. Read based metrics measured by 6 surveyed laboratories, where 5 of them answered this question.

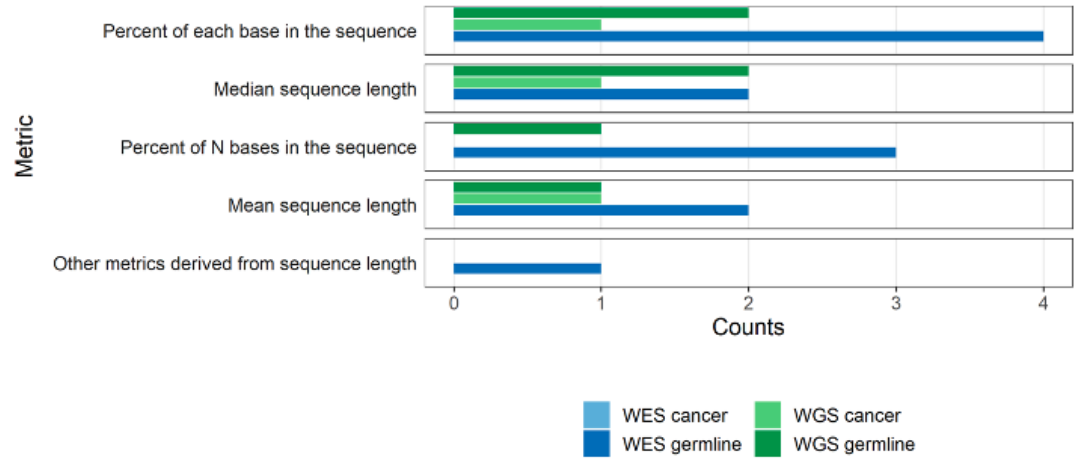


Figure 20. Base quality metrics measured by 6 surveyed laboratories, where all of them answered this question.

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

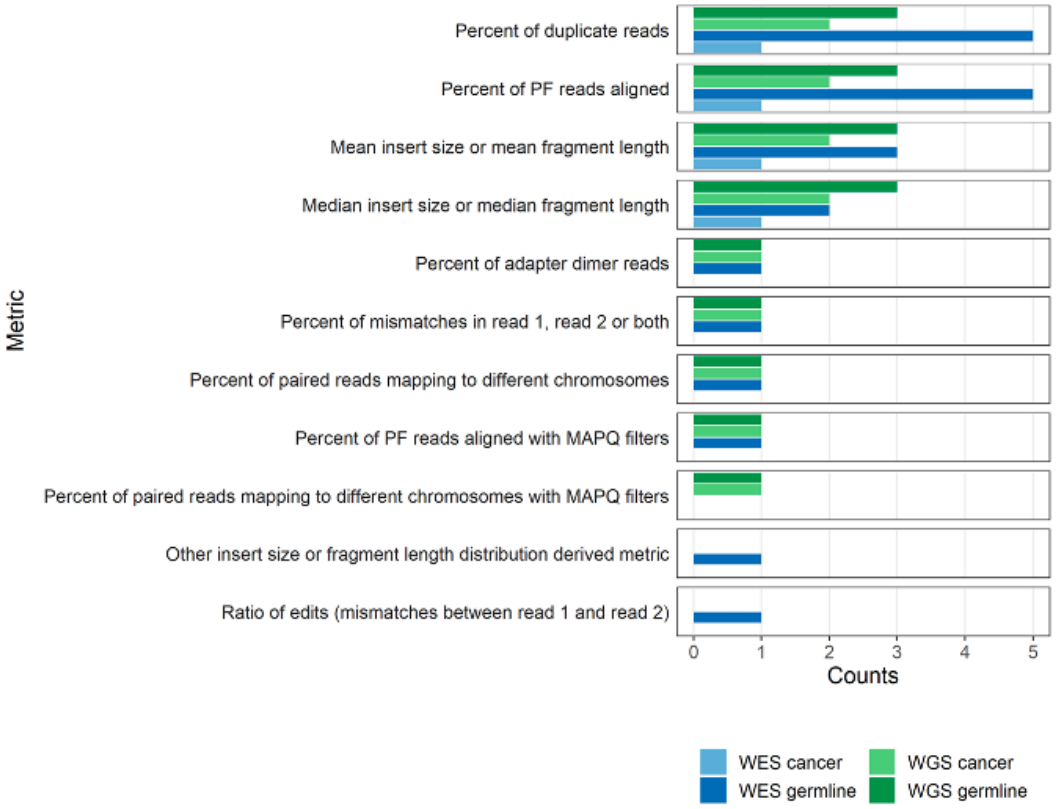


Figure 21. Alignment metrics measured by 6 surveyed laboratories, where all of them answered this question.

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

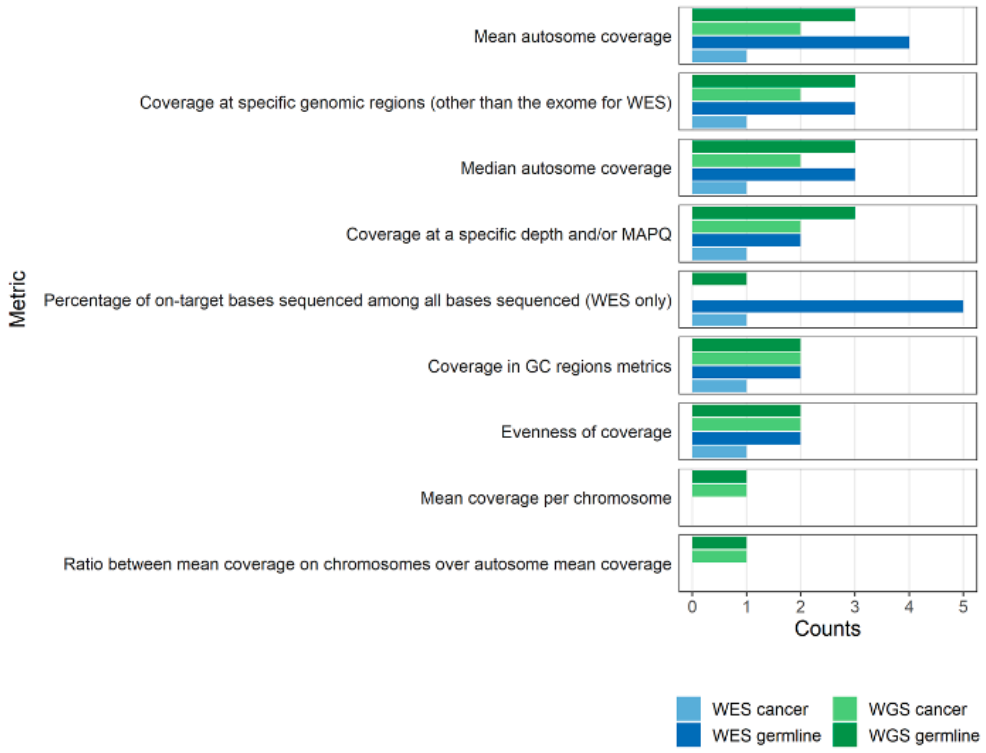


Figure 22. Coverage metrics measured by 6 surveyed laboratories, where all of them answered this question.

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

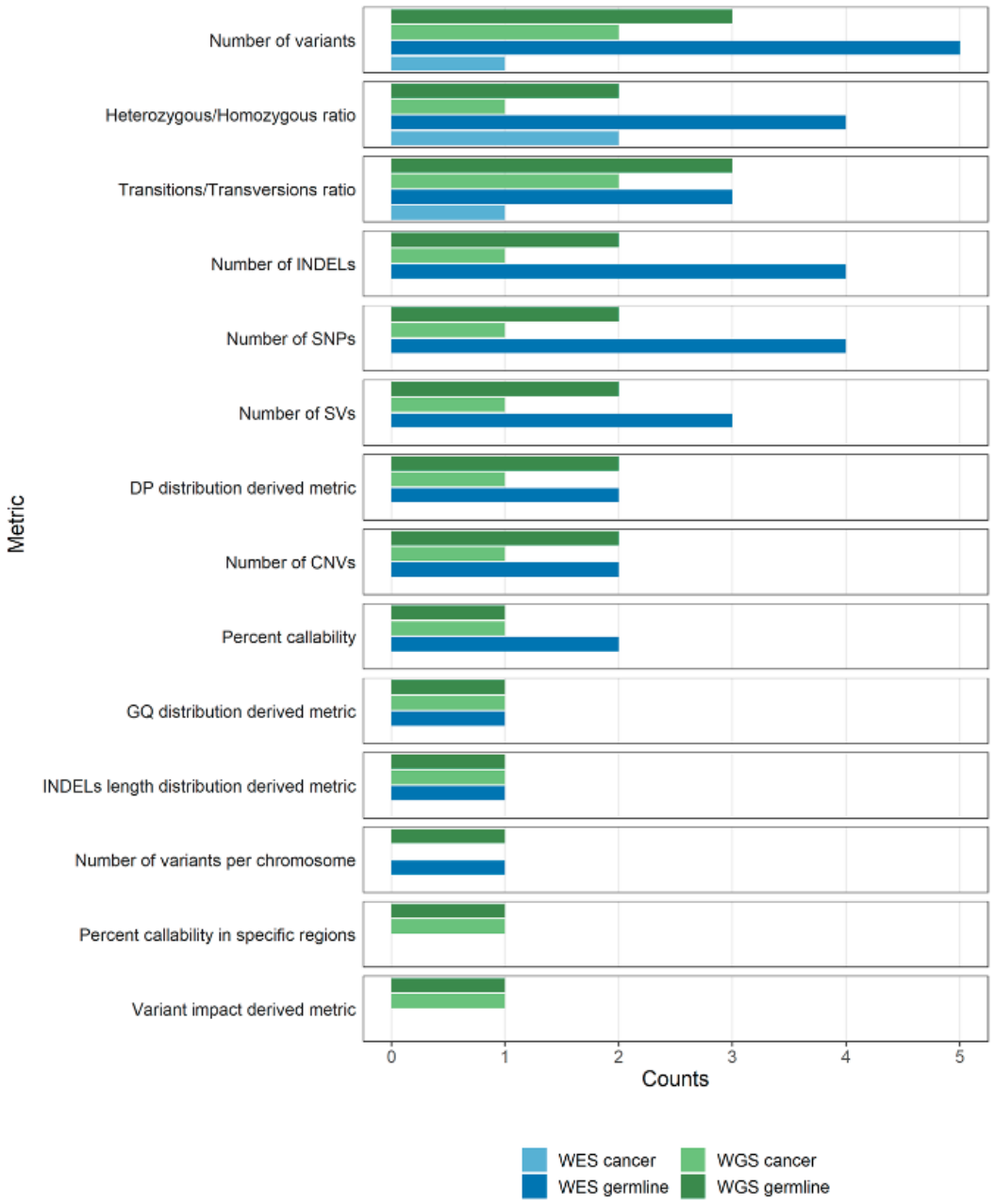


Figure 23. Variant calling metrics measured by 6 surveyed laboratories, where all of them answered this question.

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

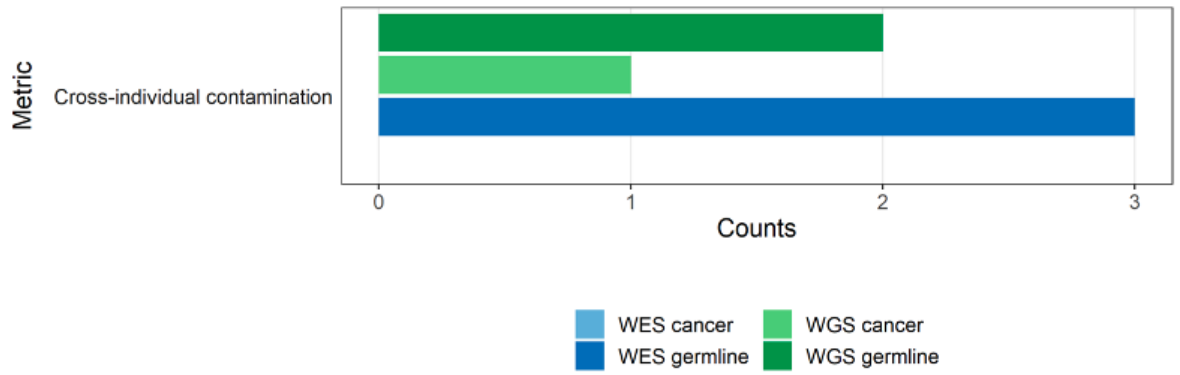


Figure 24. Surveyed laboratories that estimate cross-individual contamination, where 4 of them answered this question.

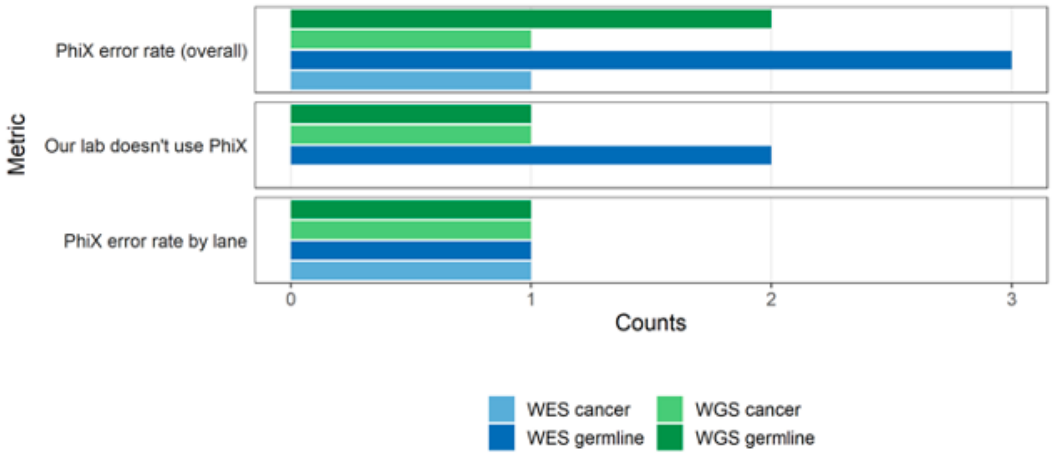


Figure 25. PhiX metrics measured by 6 surveyed laboratories, where all of them answered this question.

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

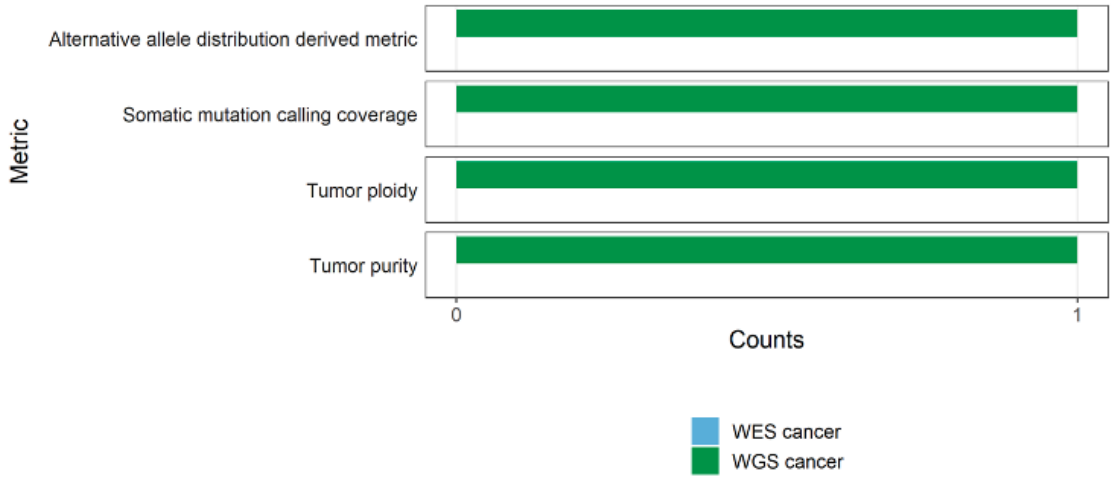


Figure 26. Somatic metrics measured by 6 surveyed laboratories, where one of them answered this question

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