



# Quality control Sarek workflow supporting documentation

## Summary

In the context of IMPACT-Data, a working group of genomics experts worked towards the generation of a quality control (QC) workflow specific for germline exome samples. The final product consists of a modified version of the well-known Nextflow workflow Sarek from the nf-core community. This pipeline has been modified adding new QC metrics and tools to obtain a broader view of the quality details of the data and the pipeline. All the new implemented features were selected by this group of experts as part of the genomics work package (WP3) within the IMPACT-Data project.



*El proyecto IMPACT-Data (Exp. IMP/00019) ha sido financiado por el Instituto de Salud Carlos III, co-financiado por el Fondo Europeo de Desarrollo Regional (FEDER, “Una manera de hacer Europa”).*

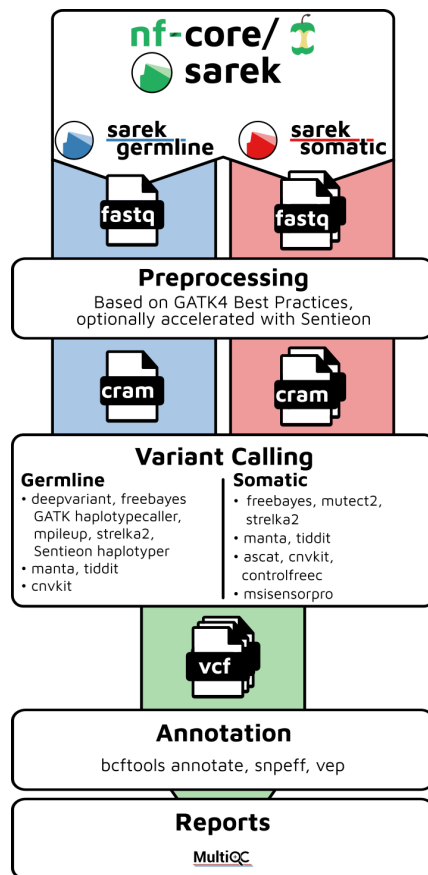
## Main Contributors

Name	Institute	Position
Jordi Rambla	CRG	EGA Team Head (WP3 lead)
Amy Curwin	CRG	EGA Project Manager (WP3 coordination)
Arnau Soler	CRG	EGA Bioinformatician
Sheila Zúñiga	INCLIVA	Head of Bioinformatics Unit
Igor Ruiz de los Mozos	NASERTIC	Senior Bioinformatician
Isabel Cuesta de la Plaza	ISCI	Head of Bioinformatics Unit
Sarai Varona Fernandez	ISCI	Bioinformatician
Jorge Amigo	FPGMX	Head of Bioinformatics Unit
Pablo Minguez	IIS-FJD	Head of Bioinformatics Unit
David Cordero	ICO-IDIBELL	Head of Bioinformatics Unit
Aitor Zarandona	Biobizkaia	Bioinformatician
Naiara Garcia	Biobizkaia	Head of Bioinformatics Unit

## Background

The main goal of this work, as part of IMPaCT-Data [1], was to create an easy to use pipeline able to detect genomic variants complete with rich and deep QC. To this end, among the group of genomic experts, it was decided to start with germline exome samples and to modify the existing nf-core/Sarek workflow [2].

Sarek [3,4], part of the nf-core [5] community, is a comprehensive workflow crafted for identifying variants in both whole genome and targeted sequencing data (Figure 1). While initially tailored for analysis in humans and mice, its adaptability extends to any species with a reference genome. Moreover, Sarek efficiently manages tumor/normal pairs and has the potential to incorporate additional relapses.



**Figure 1:** Schematic representation of the official Sarek workflow.

In order to be aligned with the FAIR principles [6], the pipeline was constructed using Nextflow [7], a versatile workflow tool that enables seamless task execution across diverse computing infrastructures. By leveraging Docker [8] and Singularity [9] containers and Bioconda [10, 11], installation becomes straightforward, and the generated results boast high reproducibility. Notably, the implementation of this pipeline in Nextflow DSL2 [12] employs a

container-per-process approach, significantly simplifying the maintenance and updates of software dependencies.

Implementing the new QC pipeline within this existing Sarek workflow and using Nextflow merges all the excellent characteristics from both parts to create an easy to use product able to help all scientists.

## Methodology

A set of 40 metrics related to human genomic data and the toolset or software to extract these metrics, were decided upon by the genomics experts within IMPACT-Data. The goal of this current work was to implement the different QC metrics in the different data formats (ie. FASTQ [13], BAM [14] and VCF [15, 16]) for germline whole exome sequencing data.

Within this subworkflow there are 11 tools used to obtain the metrics.

- Fastp [17]
- FastQC [18]
- Mosdepth [19]
- GATK-Picard CollectHsMetrics [20, 21]
- Samtools flagstat [14, 22]
- GATK-Picard MarkDuplicates [20, 21]
- GATK-Picard CollectInsertSizeMetrics [20, 21]
- Sex.DetERRmine [23]
- Somalier [24]
- Fastq-Screen [25]
- Bcftools [22, 26]
- Vcftools [15]
- MultiQC [27]

In this documentation, you can find all the detailed information of the metrics implemented in Sarek and how, why or what is taken into account in each of them.

Here you can find the GitHub repository version of the modified Sarek workflow with information on how to use the original and the new features of the pipeline.

<https://github.com/EGA-archive/sarek-IMPACT-data-QC>

# Metrics Index

	Metric	Tool
1	<a href="#">Percentage passed reads</a>	fastp
2	<a href="#">Percentage filtered reads</a>	fastp
3	<a href="#">Total bases sequenced before filtering</a>	fastp
4	<a href="#">Total bases sequenced after filtering</a>	fastp
5	<a href="#">Bases sequenced with Q <math>\geq</math> 30</a>	fastp
6	<a href="#">Mean or median sequence length</a>	FastQC
7	<a href="#">Percent of each base in the sequence</a>	FastQC
8	<a href="#">Percent of N bases in the sequence</a>	FastQC
9	<a href="#">Presence of adaptors</a>	FastQC
10	<a href="#">Mean or median base quality per read</a>	FastQC
11	<a href="#">Statistics on Q-Scores across all bases</a>	FastQC
12	<a href="#">Sequence length distribution</a>	FastQC
13	<a href="#">Overrepresented level</a>	FastQC
14	<a href="#">Error rate</a>	FastQC
15	<a href="#">K-mer analysis</a>	FastQC
16	<a href="#">N fragment</a>	FastQC
17	<a href="#">Raw mean or median coverage</a>	Mosdepth
18	<a href="#">Coverage at specific depth</a>	Mosdepth
19	<a href="#">Percentage of ontarget bases sequenced among all mapped bases</a>	CollectHsMetrics
20	<a href="#">Percentage ontarget bases sequenced</a>	CollectHsMetrics
21	<a href="#">Percent of paired reads mapping to different chromosomes, with MAPQ<math>\geq</math>5</a>	samtools flagstat
22	<a href="#">Percent of duplicate reads</a>	MarkDuplicates
23	<a href="#">Median insert size</a>	CollectInsertSizeMetrics

24	<a href="#">MAD insert size</a>	CollectInsertSizeMetrics
25	<a href="#">Capture efficiency</a>	CollectHsMetrics
26	<a href="#">Sex determination</a>	Sex.DetERRmine / Somalier / Mosdepth
27	<a href="#">Expected species</a>	Fastq-Screen
28	<a href="#">Expected kinship</a>	Somalier
29	<a href="#">Sample duplication</a>	Somalier
30	<a href="#">DP distribution</a>	Bcftools
31	<a href="#">GQ distribution</a>	Vcftools
32	<a href="#">Strand bias</a>	Vcftools
33	<a href="#">Allelic read percentages</a>	Custom script
34	<a href="#">Number of SNPs</a>	Bcftools
35	<a href="#">Number of indels</a>	Bcftools
36	<a href="#">Number of variants</a>	Bcftools
37	<a href="#">Number of multiallelic variants</a>	Bcftools
38	<a href="#">Heterozygous/homozygous ratio for SNVs</a>	Custom script
39	<a href="#">Transitions/transversions ratio for SNVs</a>	Bcftools
40	<a href="#">Percent SNV changes</a>	Bcftools

## Details per metric

**DISCLAIMER:** For all metrics, bear in mind there is not a specific threshold, only indicative values. The final criteria is dependent on the type of study, analysis and/or technology used and left to the user perspective. More information is provided per metric in the below tables. Remember, these apply to germline whole exome sequencing data.

## Results Directory Structure

The default results directory structure is as follows:

```
...
{outdir}
├── csv
├── multiqc
├── pipeline_info
├── preprocessing
│   ├── markduplicates
│   │   └── <sample>
│   ├── recal_table
│   │   └── <sample>
│   ├── recalibrated
│   │   └── <sample>
├── reference
├── reports
│   ├── <tool1>
│   ├── <tool2>
│   │   ├── <sample1>
│   │   └── <sample2>
├── <impact_qc>
│   ├── <tool1>
│   ├── <tool2>
│   │   ├── <sample1>
│   │   └── <sample2>
work/
.nextflow.log
...
```

## 1. Percentage passed reads

<b>Name</b>	Percentage passed reads
<b>Definition</b>	Percentage of number of reads that passed the filters calculated from fastp metrics $[(\text{Number of reads after filtering} / \text{Number of reads before filtering}) * 100]$ .
<b>Input</b>	FASTQ
<b>Tool</b>	fastp
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'PCT_PASSED_READS' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'total reads before and after filtering' - 'reports/fastp/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Percentage of passed reads" metric from fastp is an informative value for understanding the quality and usability of your sequencing data before it undergoes further analysis. This metric represents the proportion of reads that are kept from the initial dataset during the quality control (QC) steps, which include reads with good quality and the desired criteria among others.</li> <li>This metric could be very different across different analyses since it depends on a lot of things. A bad value could be below ~80%, but this is not critical if other metrics support it. For example, it depends on coverage, the number of sequenced bases expected, the quality of them, the filters used and the study design or technology.</li> </ul>

## 2. Percentage filtered reads

<b>Name</b>	Percentage filtered reads
<b>Definition</b>	Percentage of number of reads removed calculated from fastp metrics $[100 - ((\text{Number of reads after filtering} / \text{Number of reads before filtering}) * 100)]$ .
<b>Input</b>	FASTQ
<b>Tool</b>	fastp
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'PCT_FAILED_READS' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'total reads before and after filtering' - 'reports/fastp/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Percentage of filtered reads" metric from fastp is an informative value for understanding the quality and usability of your sequencing data before it undergoes further analysis. This metric represents the proportion of reads that were removed from the initial dataset during the quality</li> </ul>



	<p>control (QC) steps, which may include the removal of reads with low quality reads and ends, adapters, very short reads, among others.</p> <ul style="list-style-type: none"> <li>This metric could be very different across different analyses since it depends on a lot of things. A bad value could be above ~20%, but this is not critical if other metrics support it. For example, it depends on coverage, the number of sequenced bases expected, the quality of them, the filters used and the study design or technology.</li> </ul>
--	---

### 3. Total bases sequenced before filtering

<b>Name</b>	Total bases sequenced before filtering
<b>Definition</b>	Number of bases sequenced before the filtering step.
<b>Input</b>	FASTQ
<b>Tool</b>	fastp
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>'N_SEQ_BASES_BEFORE' - General Statistics Table</li> </ul> <p>Workflow output ('{outdir}/')</p> <ul style="list-style-type: none"> <li>'total bases before filtering' - 'reports/fastp/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Total bases sequenced before filtering" metric represents the cumulative count of nucleotide bases generated before filtering, crucial for assessing sequencing depth and coverage. High sequencing depth is vital for reliable variant detection, especially for variants with low allele frequencies.</li> <li>This metric guides in evaluating the efficiency of the sequencing run, aids in calculating the necessary coverage for accurate variant calling, and helps in planning computational resources for subsequent analyses.</li> <li>In workflows like nf-core/Sarek, it serves as a foundational quality control measure, ensuring the data's adequacy for comprehensive genomic analyses and variant detection. This value should be close to the value defined at the time of sequencing.</li> </ul>

### 4. Total bases sequenced after filtering

<b>Name</b>	Total bases sequenced after filtering
<b>Definition</b>	Number of bases sequenced after the filtering step.
<b>Input</b>	FASTQ
<b>Tool</b>	fastp

<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>'N_SEQ_BASES_AFTER' - General Statistics Table</li> </ul> <p>Workflow output ('{outdir}/')</p> <ul style="list-style-type: none"> <li>'total bases after filtering' - 'reports/fastp/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Total bases sequenced after filtering" metric represents the cumulative count of nucleotide bases after the filtering step, important to know how many bases the sample has left after the filtering processes have been applied. As the number of total bases sequenced before filtering (metric 1), this metric helps in understanding if the sample could assess the wanted sequencing depth and coverage.</li> <li>This value must be proportional to the percentage of filtered reads and should not fall below 80% of the total. Remember that it is not a definitive value, it must always take into account other factors such as other metrics or the design of the study.</li> </ul>

## 5. Bases sequenced with $Q \geq 30$

<b>Name</b>	Bases sequenced with quality $\geq 30$
<b>Definition</b>	Total number of bases sequenced with at least a quality of 30.
<b>Input</b>	FASTQ
<b>Tool</b>	fastp
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>'N_SEQ_BASES_Q30_BEFORE' - General Statistics Table</li> <li>'N_SEQ_BASES_Q30_AFTER' - General Statistics Table</li> </ul> <p>Workflow output ('{outdir}/')</p> <ul style="list-style-type: none"> <li>'total bases with Q30 before and after filtering' - 'reports/fastp/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Bases sequenced with <math>Q \geq 30</math>" metric, reported by tools like fastp, indicates the number of bases with a quality score of 30 or higher, signifying a base call accuracy of 99.9%. This measure is critical for evaluating the reliability of sequencing data, as it quantifies the proportion of the dataset that is of high quality and likely to be accurate. High-quality bases are essential for confident variant calling and downstream analyses, ensuring that errors in base calling minimally impact the results.</li> <li>This metric serves as a key indicator of data quality in the Sarek nf-core workflow, aiding in the optimization of sequencing strategies and the assessment of sequencing performance. Depending on every study, different quality scores could be enough for the subsequent analyses, a good value should represent that more than 75% of the bases are above Q30.</li> </ul>

## 6. Mean or median sequence length

<b>Name</b>	Mean or median sequence length
<b>Definition</b>	Mean or median length of the sequenced reads
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Median Read Length' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'Sequence length' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Mean or median sequence length" metric from FastQC provides an average or midpoint value of the sequence lengths after quality control processes, such as trimming for quality and adapter removal. This metric is essential for understanding the uniformity and integrity of the sequencing reads, which impacts the accuracy of alignment, variant calling, and overall analysis quality.</li> <li>In workflows like Sarek nf-core, evaluating this metric helps ensure that the sequencing data meets the necessary criteria for reliable downstream analyses, allowing for adjustments in data processing or experimental design if needed. It needs to be the expected length in the study design or technology.</li> </ul>

## 7. Percent of each base in the sequence

<b>Name</b>	Percent of each base in the sequence
<b>Definition</b>	The percentage of each of the different nucleotides (A,T,C,G) in all the sequenced bases
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Per Base Sequence Content' - FastQC section</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'Per Base Sequence Content' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Percent of each base in the sequence" metric from FastQC quantifies the distribution of nucleotide bases (A, T, G, C) across all sequences, offering insights into the composition bias of the sequencing data. This information is crucial for identifying any abnormal base composition that might suggest sequencing biases, contamination, or</li> </ul>

	<p>errors in library preparation.</p> <ul style="list-style-type: none"> <li>• Within the Sarek nf-core workflow, understanding base composition is vital for ensuring the quality and accuracy of the sequencing data ahead of complex analyses like variant calling, as biases can affect sequence alignment and the detection of genomic variants.</li> </ul>
--	--

## 8. Percent of N bases in the sequence

<b>Name</b>	Percent of N bases in the sequence
<b>Definition</b>	The percentage of the bases that the sequencer was not able to make a basecall for this base (N).
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>• 'Per Base N Content' - FastQC section</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>• 'Per Base N Content' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Percent of N bases in the sequence" metric from FastQC measures the proportion of undetermined nucleotides (N) in your sequencing data, which can arise from low-quality base calls or sequencing errors. A high percentage of Ns can significantly impact the quality of downstream analyses, such as alignment and variant calling, by reducing the usable data.</li> <li>• Within the context of the Sarek nf-core workflow, maintaining a low percentage of N bases is crucial for the accuracy and reliability of genomic analyses, as it ensures that the majority of the sequencing data represents confidently identified nucleotides. In order to correct this, reads with Ns could be discarded or at least trimmed till the Ns are reached.</li> </ul>

## 9. Presence of adaptors

<b>Name</b>	Presence of adaptors
<b>Definition</b>	If the sequence still has presence of adaptors (short pieces of DNA, which attach to the DNA fragments used in the sequencing step).
<b>Input</b>	FASTQ

<b>Tool</b>	FastQC*
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Adapter Content' - FastQC section</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'Adapter Content' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Presence of adaptors" metric from FastQC identifies sequences that contain sequencing adaptor remnants, indicating that not all adaptor sequences were successfully removed during the preprocessing phase. This is critical because adaptor sequences can interfere with read alignment and variant calling by falsely aligning to the reference genome or creating artifacts.</li> <li>Within the Sarek nf-core workflow, identifying and trimming adaptor sequences is essential for ensuring data integrity, as it prevents the introduction of sequencing artifacts into the analysis, thereby enhancing the accuracy of the genomic analyses conducted. No adaptor presence should be found in the data.</li> </ul>

## 10. Mean or median base quality per read

<b>Name</b>	Mean or median base quality per read
<b>Definition</b>	Mean or median quality score of the reads (Phred score).
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Per Sequence Quality Scores' - FastQC section</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'Per Sequence Quality Scores' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Mean or median base quality per read" metric from FastQC provides an aggregate measure of the sequencing quality for each read, typically calculated using Phred scores. This metric is fundamental for assessing the overall quality of sequencing data, as higher base quality scores indicate more accurate base calls.</li> <li>Within the Sarek nf-core workflow, analyzing this metric helps to ensure that only high-quality reads are used in downstream analyses like variant calling, improving the reliability of the results and minimizing the chances of errors or false positives in the final variant dataset. As in the metric 3, this metric could have different thresholds and the user needs to take into account different metrics or variables together.</li> </ul>

## 11. Statistics on Q-Scores across all bases

<b>Name</b>	Statistics on Q-Score across all bases
<b>Definition</b>	The mean quality value across each base position in the read.
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Sequence Quality Histograms' - FastQC section</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'Sequence Quality Histograms' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Statistics on Q-Scores across all bases" from FastQC offer a detailed distribution of quality scores for all nucleotides in the sequencing data, providing insight into the overall data quality. This encompasses average scores, the distribution of high versus low scores, and variations across the sequencing run. Such statistics are crucial for evaluating the consistency and reliability of sequencing outputs, guiding quality control measures.</li> <li>In the context of the Sarek nf-core workflow, this comprehensive view of Q-Scores helps in fine-tuning the analysis pipeline, ensuring that downstream processes like variant calling are based on data of uniformly high quality, thereby enhancing the accuracy and confidence in the identified genomic variants. Values above 30 are normally said to be of good quality (Phred score).</li> </ul>

## 12. Sequence length distribution

<b>Name</b>	Sequence length distribution
<b>Definition</b>	The distribution of the lengths of all the reads.
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Sequence Length Distribution' - FastQC section</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'Sequence Length Distribution' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Sequence length distribution" metric from FastQC illustrates the range and frequency of read lengths within your sequencing data after quality control steps like trimming. This distribution is key for assessing the consistency of read lengths and identifying any deviations from</li> </ul>

	<p>expected lengths, which could indicate issues with sequencing or sample preparation.</p> <ul style="list-style-type: none"> <li>In the Sarek nf-core workflow, understanding sequence length distribution aids in optimizing alignment and variant calling processes, as certain analyses may require reads of specific lengths for optimal performance. This metric ensures that the data used in downstream analyses meets the necessary criteria for accurate and reliable genomic analysis.</li> </ul>
--	---

### 13. Overrepresented level

<b>Name</b>	Overrepresented sequenced level
<b>Definition</b>	The sequences that are overrepresented.
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>'Top overrepresented sequences' - FastQC section</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>'Top overrepresented sequences' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Overrepresented sequence level" metric from FastQC identifies sequences that appear more frequently in the dataset than expected by chance. This could indicate biological significance, such as a highly expressed gene, or technical issues like contamination, PCR duplication, or adapter dimer formation.</li> <li>In the context of the Sarek nf-core workflow, analyzing overrepresented sequences helps in assessing the quality of the sequencing run and the integrity of the sample preparation process. It's crucial for ensuring that the data used for downstream analyses, such as variant calling, is free from biases or artifacts that could compromise the accuracy of the results.</li> </ul>

### 14. Error rate

<b>Name</b>	Error rate
<b>Definition</b>	Errors in the number of bases called when a single nucleotide occurs more than once in consecutive order in a sequence
<b>Input</b>	FASTQ

<b>Tool</b>	FastQC*
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>• 'Per Sequence Quality Scores (shows it when warning)' - FastQC section</li> </ul> <p>Workflow output ('{outdir}/')</p> <ul style="list-style-type: none"> <li>• 'Per Sequence Quality Scores (shows it when warning)' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Error rate" metric from FastQC estimates the frequency of incorrect basecalls across the sequencing data, providing a direct measure of the sequencing accuracy. This rate is crucial for gauging the reliability of the data: lower error rates indicate higher quality sequencing, which is essential for accurate alignment and variant calling.</li> <li>• Within the Sarek nf-core workflow, maintaining a low error rate is vital to ensure the integrity and trustworthiness of the genomic analysis, as even small errors can significantly impact the identification and interpretation of genetic variants.</li> </ul>

## 15. K-mer analysis

<b>Name</b>	K-mer analysis
<b>Definition</b>	K-mer refers to all possible subsequences (of length k) from a nucleic acid sequence.
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>• 'Status checks' - FastQC section</li> </ul> <p>Workflow output ('{outdir}/')</p> <ul style="list-style-type: none"> <li>• 'Status checks' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "K-mer analysis" from FastQC examines the frequency of all possible subsequences of length k (where k is a positive integer) in the sequencing data. This analysis helps identify overrepresented or unusual patterns in the sequence data that could indicate contamination, repeats, or biases in the sequencing library preparation.</li> <li>• Within the Sarek nf-core workflow, k-mer analysis contributes to quality control by highlighting potential issues that could affect downstream analyses like alignment and variant calling. By detecting these anomalies early, researchers can take corrective actions to ensure that the data used in subsequent steps is of high quality and reliability, leading to more accurate genomic interpretation.</li> </ul>



## 16. N fragment

<b>Name</b>	N fragment
<b>Definition</b>	Number and/or percentage of ambiguous calls
<b>Input</b>	FASTQ
<b>Tool</b>	FastQC*
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Per Base N Content' - FastQC section</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'Per Base N Content' - 'reports/fastqc/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "N fragment" metric from FastQC identifies and quantifies segments within sequencing reads that consist solely of 'N' bases, where 'N' represents an undetermined nucleotide. In sequencing data, 'N' bases occur when the sequencing instrument is unable to confidently call a base. A high number of N fragments can indicate low-quality sequencing runs or problems with the sequencing chemistry.</li> <li>Within the Sarek nf-core workflow, monitoring and minimizing N fragments is important for maintaining data quality, as excessive undetermined bases can impair the accuracy of read alignment and variant calling, affecting the overall reliability of the genomic analysis outcomes.</li> </ul>

\* *FastQC (metrics 6-16) generates a highly informative color scheme, highlighting correct values in green, warning areas in yellow, and problematic values in red across all metrics. Take these color schemes into account to know if your data is considered correct or problematic.*

## 17. Raw mean or median coverage

<b>Name</b>	Raw mean or median coverage
<b>Definition</b>	The average number of times a position in the genome is sequenced, considering only uniquely aligned reads.
<b>Input</b>	BAM
<b>Tool</b>	Mosdepth
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Median' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'reports/mosdepth/.'</li> </ul>

<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Raw mean or median coverage" metric from Mosdepth represents the average or midpoint number of reads aligning to a reference sequence across the entire dataset or a specified region. This measurement is fundamental in assessing the sequencing depth, providing insights into whether the genomic or targeted regions of interest have been sequenced sufficiently to detect variants reliably.</li> <li>In the context of the Sarek nf-core workflow, understanding raw coverage is crucial for evaluating the quality of the sequencing and the potential for accurate variant detection. Adequate coverage ensures that variants, especially those with low allele frequencies, can be identified with high confidence, thereby enhancing the robustness of genomic analyses. In this context, the median coverage should be near 40X or above, to ensure good results.</li> </ul>
-------------	---

## 18. Coverage at specific depth

<b>Name</b>	Coverage at specific depth (C1, C8, C10, C20, C30, C100)
<b>Definition</b>	Number of reads for the targeted regions that are equal or greater to a particular depth of coverage.
<b>Input</b>	BAM
<b>Tool</b>	Mosdepth
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>'≥ 1X, ≥ 8X, ≥ 10X, ≥ 20X, ≥ 30X, ≥ 100X' - General Statistics Table</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>'reports/mosdepth/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Coverage at specific depth" metrics (i.e., C1, C8, C10, C20, C30, C100) from Mosdepth quantify the proportion of the genome or targeted regions that are covered by at least a specific number of reads, such as 1, 10, or any other threshold. These metrics are crucial for understanding the depth of sequencing across the genome, indicating how uniformly the target regions are sequenced. For instance, C1 indicates the percentage of bases covered by at least one read, while C10 indicates the percentage covered by ten or more reads.</li> <li>In the context of the Sarek nf-core workflow, these metrics help in assessing the adequacy of coverage for reliable variant calling. High coverage across targeted regions ensures that variants, including those that are rare or occur in difficult-to-sequence areas, are detected with greater accuracy and confidence. One has to make sure that the coverage does not change unexpectedly and does not decrease or rise abruptly in the different target zones.</li> </ul>

## 19. Percentage of on-target bases sequenced among all mapped bases

<b>Name</b>	Percentage of on-target bases sequenced among all mapped bases
<b>Definition</b>	The fraction of PF_BASES_ALIGNED located on or near a baited region (ON_BAIT_BASES + NEAR_BAIT_BASES)/PF_BASES_ALIGNED.
<b>Input</b>	BAM
<b>Tool</b>	Picard CollectHsMetrics
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Selected bases' - HSMetrics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'impact_qc/gatk4/collecthsmetrics/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Percentage of on-target bases sequenced among all mapped bases" from Picard's CollectHsMetrics quantifies the proportion of sequenced bases that align to the intended target regions as compared to the total number of bases that align anywhere in the genome. This metric is pivotal for evaluating the specificity and efficiency of targeted sequencing efforts, such as those employed in exome sequencing or targeted panel sequencing projects. A high percentage indicates that a large portion of the sequencing effort was focused on the regions of interest, which is desirable for targeted sequencing projects.</li> <li>Within the Sarek nf-core workflow, assessing this metric is crucial for ensuring that the sequencing data is highly relevant to the study's objectives, maximizing the utility and cost-effectiveness of the sequencing run by focusing analysis on regions with the highest likelihood of containing clinically or functionally relevant variants.</li> </ul>

## 20. Percentage on-target bases sequenced

<b>Name</b>	Percentage on-target bases sequenced
<b>Definition</b>	The number of aligned, de-duped, on-target bases out of all of the PF bases available.
<b>Input</b>	BAM
<b>Tool</b>	Picard CollectHsMetrics
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Usable bases on-target' - HSMetrics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'impact_qc/gatk4/collecthsmetrics/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Percentage of on-target bases" metric from CollectHsMetrics, a part</li> </ul>

	<p>of the Picard toolkit, measures the proportion of all sequenced bases that align to target regions designated for a hybrid selection or targeted sequencing project. This metric is essential for assessing the efficiency and focus of the sequencing effort, indicating how much of the sequencing output is directly relevant to the areas of interest. A high percentage suggests effective targeting and capture, implying that the sequencing resources are being utilized efficiently to generate data that is most pertinent to the study's goals.</p> <ul style="list-style-type: none"> <li>In the Sarek nf-core workflow, this metric helps in evaluating the performance of the capture and enrichment process, ensuring that the data analysis can proceed with a high degree of confidence in the coverage and specificity of the target regions, which is crucial for accurate variant detection and characterization.</li> </ul>
--	---

## 21. Percent of paired reads mapping to different chromosomes, with MAPQ>=5

<b>Name</b>	Percent of paired reads mapping to different chromosomes, with MAPQ>=5
<b>Definition</b>	The percent of paired reads mapping to different chromosomes with MAPQ>=5 is a measure of the proportion of paired-end sequencing reads where one read maps to one chromosome and its mate maps to a different chromosome with a high confidence mapping quality score of 5 or greater.
<b>Input</b>	BAM
<b>Tool</b>	samtools flagstat
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'PCT_DIFF_CHR_MAPQ5' - General Statistics Table</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Percent of paired reads mapping to different chromosomes, with MAPQ&gt;=5" metric from samtools flagstat highlights the proportion of read pairs in a dataset where each read maps to a different chromosome and both reads have a mapping quality score (MAPQ) of 5 or higher. This MAPQ threshold suggests a relatively high confidence in the mapping location. Such inter-chromosomal read pairs can be indicative of structural variations, such as translocations, or they might result from sequencing or alignment errors.</li> <li>In the context of the Sarek nf-core workflow, this metric is important for several reasons: <ul style="list-style-type: none"> <li>Structural Variant Detection: It provides a clue about the presence of potential structural variations, which are crucial for understanding genomic rearrangements associated with various</li> </ul> </li> </ul>

	<p>genetic conditions and cancers.</p> <ul style="list-style-type: none"> <li>○ Quality Control: A high percentage of such read pairs might signal issues with the sample, such as contamination or chromosomal abnormalities, or it might indicate problems with the sequencing or alignment processes.</li> <li>● Assessing this metric helps in filtering for high-confidence mapping results and aids in the interpretation of structural genomic variations, enhancing the overall quality and reliability of genomic analyses derived from the workflow.</li> <li>● This metric can vary depending on the specific sequencing technology, experimental design, and data processing pipeline used. However, in typical high-quality sequencing data, you might expect this percentage to be relatively low, often in the range of 0.1% to 2%.</li> <li>● It's important to note that the MAPQ score reflects the confidence in the mapping of a read to its assigned location in the reference genome. A higher MAPQ score indicates higher confidence in the mapping. Therefore, when considering paired reads mapping to different chromosomes with a MAPQ score of 5 or greater, you would generally expect these mappings to be reliable.</li> </ul>
--	---

## 22. Percent of duplicate reads

<b>Name</b>	Percent of duplicate reads
<b>Definition</b>	Percentage of reads that are duplicates, where duplicate reads are defined as originating from a single fragment of DNA.
<b>Input</b>	BAM
<b>Tool</b>	Picard MarkDuplicates
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>● 'Duplication' - General Statistics Table</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>● 'reports/markduplicates/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>● The "Percent of duplicate reads" metric from Picard's MarkDuplicates indicates the proportion of reads identified as duplicates, reflecting PCR amplification or library preparation redundancy. High duplicate rates may suggest over-amplification or low library complexity, affecting data quality and variant calling accuracy by biasing read counts.</li> <li>● In the Sarek nf-core workflow, managing this metric is crucial for ensuring that sequencing resources yield efficient, high-quality data, enabling more</li> </ul>

	accurate genomic analyses by minimizing artificial inflation in read depth at variant sites. Near the 10% could be considered a normal percentage.
--	--

## 23. Median insert size

<b>Name</b>	Mean or median insert size, or fragment length (or other metric derived from insert size distribution)
<b>Definition</b>	Median insert size of the reads.
<b>Input</b>	BAM
<b>Tool</b>	Picard CollectInsertSizeMetrics
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Insert Size' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'impact_qc/gatk4/collectinsertsizemetrics/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Median insert size" metric from Picard's CollectInsertSizeMetrics quantifies the average or midpoint length of the DNA fragments between paired-end reads, offering insight into the library preparation quality. Variations in insert size can indicate issues with library preparation or potential genomic rearrangements.</li> <li>Within the Sarek nf-core workflow, this metric is essential for assessing sequencing library quality and ensuring consistency in sequencing data, which impacts alignment accuracy and the reliability of structural variant detection, contributing to the overall integrity of genomic analyses. This metric should be the same as the insert size calculated in the study design.</li> </ul>

## 24. MAD insert size

<b>Name</b>	MAD insert size
<b>Definition</b>	The median absolute deviation of insert sizes over the "core" of the distribution. If the distribution is essentially normal then the standard deviation can be estimated as $\sim 1.4826 * MAD$ .
<b>Input</b>	BAM
<b>Tool</b>	Picard CollectInsertSizeMetrics
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'MAD_INSERT_SIZE' - General Statistics Table</li> </ul>

	Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'impact_qc/gatk4/collectinsertsizemetrics/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "MAD (Median Absolute Deviation) insert size" from Picard's CollectInsertSizeMetrics provides a measure of variability in the lengths of DNA fragments between paired-end reads, focusing on the dispersion around the median insert size. This statistic is less sensitive to outliers than the standard deviation, offering a robust view of insert size consistency.</li> <li>In the context of the Sarek nf-core workflow, analyzing MAD insert size helps in evaluating the quality of library preparation and sequencing consistency. A low MAD value indicates uniform fragment sizes, which is critical for accurate read mapping and reliable detection of structural variants, ensuring the overall quality of genomic data analysis.</li> </ul>

## 25. Capture efficiency

<b>Name</b>	Capture efficiency (fold enrichment)
<b>Definition</b>	The fold by which the baited region has been amplified above genomic background.
<b>Input</b>	BAM
<b>Tool</b>	Picard CollectHsMetrics
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Fold enrichment' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'impact_qc/gatk4/collecthsmetrics/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Capture Efficiency" metric from CollectHsMetrics, part of the Picard toolkit, assesses the effectiveness of the target enrichment process in a hybrid capture-based sequencing workflow. It calculates the proportion of on-target reads out of all reads that pass the quality filters, providing insight into how well the capture system performs in isolating the regions of interest from the entire genomic sample. High capture efficiency indicates a successful enrichment process, where a significant majority of the sequenced reads are from targeted regions, enhancing the depth of coverage and the sensitivity of variant detection in these areas.</li> <li>In the Sarek nf-core workflow, monitoring capture efficiency is crucial for ensuring the high quality and cost-effectiveness of sequencing efforts, particularly in applications like exome sequencing or targeted gene panel analyses, where the focus is on analyzing specific genomic regions.</li> </ul>

## 26. Sex determination

<b>Name</b>	Sex determination
<b>Definition</b>	Determination of sample's sex.
<b>Input</b>	BAM
<b>Tool</b>	Sex.DetERRmine, Somalier, Mosdepth (XY coverage)
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>• 'Relative Coverage' - Sex.DetERRmine</li> <li>• 'Sex' - Statistics Somalier</li> <li>• 'XY coverage' - Mosdepth</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>• 'impact_qc/sexdetermine/.'</li> <li>• 'impact_qc/somalier/.'</li> <li>• 'reports/mosdepth/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• Sex determination from Sex.DetERRmine and Mosdepth, are methods used to infer an individual's biological sex based on the genomic data derived from sequencing. This approach typically examines the coverage or read depth across sex chromosomes (X and Y) in comparison to autosomes. A higher relative coverage of the X chromosome and minimal to no coverage of the Y chromosome suggests a female (XX) genotype, while significant coverage of both X and Y chromosomes indicates a male (XY) genotype. In Somalier, another method is used, it extracts informative sites to evaluate the counts of heterozygous and homozygous sites in the X chromosome.</li> <li>• This analysis is crucial for verifying sample identity and integrity, as mismatches between reported and determined sex can indicate sample contamination, mislabeling, or other issues. In all types of analyses, accurate sex determination is important for downstream analyses and ensures the quality of the metadata. But in addition, in analyses such as variant determination, it can play a more important role as sex-specific considerations can affect the interpretation of genomic variants.</li> </ul>

## 27. Expected species

<b>Name</b>	Expected species
<b>Definition</b>	Determine the species: human, mouse, etc.
<b>Input</b>	BAM
<b>Tool</b>	FastQ-Screen



<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>• 'Mapped Reads' - FastQ Screen</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>• 'impact_qc/fastqscreen/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Expected species" metric inferred by FastQ-Screen allows users to assess the presence and proportion of reads from the species expected to be in the sequencing data. By screening reads against a panel of reference genomes, FastQ-Screen quantifies how much of the data aligns to the genome of the expected species versus potential contaminants or other organisms. This is crucial for confirming the purity of the sample and the absence of significant cross-contamination, which is vital for accurate genomic analysis.</li> <li>• In the context of the Sarek nf-core workflow, ensuring that the majority of reads originate from the expected species underpins the validity of downstream analyses, such as variant calling, by reducing the risk of misinterpretation due to foreign DNA sequences.</li> </ul>

## 28. Expected kinship

<b>Name</b>	Expected kinship
<b>Definition</b>	Determine the expected kinship and/or relatedness.
<b>Input</b>	BAM
<b>Tool</b>	Somalier
<b>Where?</b>	<p>MultiQC report ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>• 'Heterozygosity' - Somalier</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>• 'impact_qc/somalier/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Expected kinship" metric inferred by Somalier estimates the genetic relatedness between samples based on shared variants, allowing for the detection of biological relationships (e.g., parent-child, siblings) or identifying sample swaps, contaminations and checking if the metadata are correct. This is particularly important in studies involving multiple samples or individuals, where maintaining sample integrity is crucial.</li> <li>• In the context of the Sarek nf-core workflow, assessing expected kinship helps ensure the accuracy of sample labels and the validity of genetic analyses by confirming that the samples originate from the anticipated individuals or groups, facilitating reliable interpretation of genetic data and variant associations.</li> </ul>

## 29. Sample duplication

<b>Name</b>	Sample duplication
<b>Definition</b>	Verify if two or more samples are identical.
<b>Input</b>	BAM
<b>Tool</b>	Somalier
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Heterozygosity' - Somalier</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'impact_qc/somalier/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "Sample duplication" detection metric checked by Somalier identifies instances where multiple samples in a dataset may actually be from the same individual or represent technical replicates. This is achieved by analyzing the genetic overlap between samples, looking for higher-than-expected similarity in allele frequencies and variant calls that would suggest duplication. Identifying duplicate samples is crucial for accurate data interpretation, as duplicates can skew allele frequency calculations, affect association studies, and lead to erroneous conclusions.</li> <li>In the context of the Sarek nf-core workflow, ensuring sample uniqueness helps maintain the integrity of the genomic analysis, allowing for reliable variant calling and genetic analysis by accurately representing the diversity and true size of the sample cohort.</li> </ul>

## 30. DP distribution

<b>Name</b>	Read depth or depth of coverage distribution
<b>Definition</b>	Read depth or depth of coverage distribution provides information about the distribution of read depth across the variants in a dataset. Where DP is the filtered depth, at the sample level. This gives you the number of passed reads that support each of the reported alleles.
<b>Input</b>	VCF
<b>Tool</b>	Bcftools
<b>Where?</b>	MultiQC report (--tools plotdp) ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Dp Distribution' - Custom</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>'impact_qc/vcftoolscustom/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>The "DP (read depth) distribution" from Bcftools indicates the number of</li> </ul>

	<p>reads covering each variant, offering insights into sequencing depth and coverage uniformity. Adequate sequencing depth is crucial for accurate variant calling, as it increases confidence in variant detection. This distribution is essential for quality control and variant filtering in genomic analyses.</p> <ul style="list-style-type: none"> <li>• Within the Sarek nf-core workflow, examining DP distribution aids in evaluating data quality, optimizing variant filtering parameters, and ensuring the reliability of downstream analyses, such as association studies or clinical variant interpretation. In germline analysis of WES, a good threshold for DP is ~30. Thanks to the DP distribution, one can see if the DP is consistent across all variants.</li> </ul>
--	---

### 31. GQ distribution

<b>Name</b>	Genotype quality (GQ) score distribution
<b>Definition</b>	Distribution of the genotype quality scores across the variants of the dataset. Where GQ represents the Phred-scaled confidence that the genotype assignment (GT) is correct, derived from the genotype PLs ("Normalized" Phred-scaled likelihoods of the possible genotypes).
<b>Input</b>	VCF
<b>Tool</b>	Vcftools
<b>Where?</b>	<p>MultiQC report (--tools plotgq) ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>• 'Gq Distribution' - Custom</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>• 'impact_qc/vcftoolscustom/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "GQ (Genotype Quality) score distribution" from Vcftools provides information about the genotype quality scores across variants in a dataset.</li> <li>• The GQ score represents the confidence in the assigned genotype for each variant. Higher GQ scores indicate more reliable genotype calls. Analyzing the GQ distribution helps assess the overall quality of genotype calls and identify variants with uncertain genotypes.</li> <li>• The GQ distribution is essential for quality control and variant filtering in genomic analyses. Within the Sarek nf-core workflow, examining GQ distributions aids in evaluating data quality, optimizing variant filtering parameters, and ensuring the reliability of downstream analyses, such as association studies or clinical variant interpretation.</li> <li>• It is recommended to filter all variants with a GQ lower than 20 (1 percent</li> </ul>

	<p>chance that the call is incorrect). Use the GQ=20 as a threshold and the variants/samples below could be marked as “wrong” and above that number could be marked as “good” variants. Please bear in mind that the GQ is associated with each sample. For example, a variant called in a trio analysis will have three different GQs, one per each sample. The variant might have a GQ below the threshold in one of the samples while having a GQ above of it in the other samples. In that case, the variant will be marked as "Failed/Not genotyped" in the sample where it had a low GQ and PASS in the others.</p> <ul style="list-style-type: none"> <li>Take into account that the GQ also depends on the AD and DP. One can have really good GQ but very poor coverage, hence that the variant in that case could be incorrect.</li> </ul>
--	--

## 32. Strand bias

<b>Name</b>	Strand bias
<b>Definition</b>	Strand bias is a type of sequencing bias in which one DNA strand is favored over the other, which can result in incorrect evaluation of the amount of evidence observed for one allele vs. the other.
<b>Input</b>	VCF
<b>Tool</b>	Vcftools
<b>Where?</b>	<p>MultiQC report (--tools plotsb) ('{outdir}/multiqc/multiqc_report.html')</p> <ul style="list-style-type: none"> <li>'Sb/Fs/Sap-Srp-Epp/Sp Distribution' - Custom</li> </ul> <p>Workflow output ('{outdir}/.')</p> <ul style="list-style-type: none"> <li>'impact_qc/vcftoolscustom/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>“Strand bias” extracted by Vcftools refers to an assessment of the imbalance in the distribution of sequencing reads aligned to the forward and reverse strands at variant sites. It evaluates whether there is a significant difference in the number of reads supporting a variant allele on one strand compared to the other. Strand bias can occur due to various factors, including PCR amplification biases, sequencing errors, or true biological phenomena. In the context of variant calling, significant strand bias may indicate potential artifacts or errors in variant detection, impacting the reliability of identified variants. Therefore, evaluating and accounting for strand bias is crucial for accurate variant calling and downstream analyses, such as association studies or functional genomics investigations.</li> <li>Within the Sarek nf-core workflow, assessing strand bias helps ensure the quality and integrity of variant calls, improving the accuracy of genomic analyses and enhancing the reliability of research findings.</li> </ul>

### 33. Allelic read percentages

<b>Name</b>	Alternate allelic read percentages
<b>Definition</b>	Allelic read percentages for reference and alternate alleles. Number of reference/alternate allele reads divided to the number of total reads multiplied per 100 for each variant.
<b>Input</b>	VCF
<b>Tool</b>	Bcftools + custom script
<b>Where?</b>	MultiQC report (--tools plotallelicreadpct) ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'Alt Allelic Read Percentages' - Custom</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>"Alternate allelic read percentages" provide the proportion of reads supporting the alternate allele at a given variant locus. This metric is crucial for determining the allelic balance at variant sites and assessing potential biases in sequencing or variant calling processes.</li> <li>Analyzing allelic read percentages helps identify scenarios such as allelic imbalance, where one allele is overrepresented compared to the other, which can result from various factors including sequencing errors, PCR amplification biases, or true biological phenomena such as copy number variations.</li> <li>In the context of genomic analysis workflows like Sarek nf-core, examining allelic read percentages aids in quality control and ensures the accuracy of variant calling. Identifying and correcting biases or artifacts improves the reliability of downstream analyses, such as variant interpretation or association studies.</li> </ul>

### 34. Number of SNPs

<b>Name</b>	Number of SNPs
<b>Definition</b>	Number of SNPs. Where a SNP is a DNA sequence variation that occurs when a single nucleotide (adenine, thymine, cytosine, or guanine) is different from the reference sequence.
<b>Input</b>	VCF
<b>Tool</b>	Bcftools
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>'SNP' - General Statistics Table</li> </ul>

	Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>• 'reports/bcftools/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Number of SNPs" metric from Bcftools tallies the single nucleotide polymorphisms (SNPs) identified in a dataset, representing variations where a single nucleotide differs from the reference genome. SNPs are the most common type of genetic variation among people and are crucial for studies in genetics, including association studies for disease linkage, population genetics, and evolutionary biology.</li> <li>• Within the Sarek nf-core workflow, counting SNPs is fundamental for understanding the genetic diversity and structure of the samples under analysis. This metric helps in identifying potential genetic markers associated with diseases or traits and in exploring the genetic basis of biological diversity, providing a foundation for further genomic and genetic research.</li> </ul>

### 35. Number of indels

<b>Name</b>	Number of indels
<b>Definition</b>	Number of indels. Where an indel, short for "insertion-deletion," is a type of genetic variation that involves the insertion or deletion of nucleotides (the building blocks of DNA) within a genome.
<b>Input</b>	VCF
<b>Tool</b>	Bcftools
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>• 'Indel' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>• 'reports/bcftools/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Number of indels" metric from Bcftools provides a count of insertions and deletions (indels) identified in a dataset, reflecting changes in the genome where bases are either inserted into or deleted from the DNA sequence. Indels can significantly impact gene function and are important for understanding genetic variability, contributing to phenotypic diversity, disease mechanisms, and evolutionary processes.</li> <li>• Within the Sarek nf-core workflow, accurately quantifying indels is crucial for comprehensive variant profiling, aiding in the detection of variants that may have functional implications or be pathogenic. This metric enhances the understanding of the structural aspects of the genome being studied, supporting a wide range of genetic analyses from functional genomics to clinical diagnostics.</li> </ul>

## 36. Number of variants

<b>Name</b>	Number of variants
<b>Definition</b>	Number of variants. Where a variant refers to any difference or alteration in the DNA sequence compared to a reference or standard sequence.
<b>Input</b>	VCF
<b>Tool</b>	Bcftools
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>• 'Vars' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>• 'reports/bcftools/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Number of variants" metric from Bcftools provides a count of all types of genetic variants identified in a dataset, including single nucleotide variants (SNVs), insertions and deletions (indels), and larger structural variations, if analyzed. This comprehensive count reflects the genetic diversity within the sample or population studied and is crucial for a wide range of genomic analyses, from understanding evolutionary processes to identifying genetic factors associated with diseases.</li> <li>• In the context of the Sarek nf-core workflow, this metric aids in quantifying the mutational burden and genomic complexity of the samples, essential for variant discovery, comparative genomics, and the elucidation of genetic underpinnings of phenotypic traits or pathologies. Accurate variant counting is foundational for subsequent analyses, including association studies and functional genomics investigations, guiding interpretations and decisions in both research and clinical settings.</li> </ul>

## 37. Number of multiallelic variants

<b>Name</b>	Number of multiallelic variants
<b>Definition</b>	Number of multiallelic variants. A multiallelic site is a specific locus in a genome that contains three or more observed alleles, counting the reference as one, and therefore allowing for two or more variant alleles.
<b>Input</b>	VCF
<b>Tool</b>	Bcftools
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>• 'N_MULTIALLELIC_VARIANTS' - General Statistics Table</li> </ul> Workflow output ('{outdir}/.')

	<ul style="list-style-type: none"> <li>• 'reports/bcftools/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Number of multiallelic variants" metric from Bcftools quantifies variants in the dataset that have more than one alternate allele at the same locus, indicating a site in the genome where multiple nucleotide variations occur. These multiallelic sites can be indicative of high genetic diversity and complexity within the sample or population, and they are particularly relevant in studies of population genetics, evolutionary biology, and in the context of complex diseases or cancer genomics.</li> <li>• In the Sarek nf-core workflow, identifying and quantifying multiallelic variants are important for understanding the full spectrum of genetic variation, as these variants may affect gene function or regulation in a more complex manner than biallelic variants. Accurate assessment of multiallelic variants supports comprehensive genomic analyses, enabling detailed investigations into genetic structure, variation, and the potential impact of these variants on phenotypes or disease risk.</li> </ul>

### 38. Heterozygous/homozygous ratio for SNVs

<b>Name</b>	Heterozygous/homozygous ratio for SNVs
<b>Definition</b>	The count of heterozygous positions over the count of homozygous non-ref positions.
<b>Input</b>	VCF
<b>Tool</b>	Bcftools + custom script
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>• 'RATIO_HET-HOM' - General Statistics Table</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Heterozygous/homozygous ratio for SNVs" from Bcftools calculates the proportion of single nucleotide variants (SNVs) that are heterozygous (one altered and one reference allele present) to those that are homozygous (both alleles altered from the reference). This ratio is a key indicator of the genetic diversity within a sample and can provide insights into the population history, inbreeding levels, and the potential presence of allelic imbalances that may be relevant in disease contexts, such as cancer or inherited genetic disorders.</li> <li>• In the Sarek nf-core workflow, analyzing this ratio helps in characterizing the mutational landscape of the samples, aiding in the interpretation of genomic data for research and clinical applications. A balanced or expected heterozygous/homozygous ratio supports the accuracy of variant calling processes, while deviations might indicate technical artifacts or biologically significant phenomena.</li> </ul>



## 39. Transitions/transversions ratio for SNVs

<b>Name</b>	Transitions/transversions ratio for SNVs
<b>Definition</b>	The count of transitions (variant from purine to purine or pyrimidine to pyrimidine) over the count of transversions transitions (variant from purine to pyrimidine or pyrimidine to purine).
<b>Input</b>	VCF
<b>Tool</b>	Bcftools
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html') <ul style="list-style-type: none"> <li>• 'Ts/Tv' - General Statistics Table/Bcftools/Vcftools</li> </ul> Workflow output ('{outdir}/.') <ul style="list-style-type: none"> <li>• 'reports/bcftools/.'</li> <li>• 'reports/vcftools/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Transitions/transversions ratio for SNVs" from Bcftools measures the ratio of transition mutations (purine to purine or pyrimidine to pyrimidine changes) to transversion mutations (purine to pyrimidine or vice versa) among the identified single nucleotide variants (SNVs).</li> <li>• This ratio is a valuable marker for understanding the mutational mechanisms and the nature of genetic changes within a genome. In evolutionary biology, a higher transitions/transversions ratio is expected due to the molecular mechanisms of mutation. In clinical genomics, deviations from expected ratios can indicate exposure to mutagens or the presence of specific mutational processes in diseases like cancer.</li> <li>• Within the Sarek nf-core workflow, analyzing this ratio contributes to the interpretation of the mutational landscape, aiding in the assessment of mutation types that are prevalent in the sample set, which can inform on the underlying biology or potential therapeutic targets.</li> </ul>

## 40. Percent SNV changes

<b>Name</b>	Percent SNV changes
<b>Definition</b>	It refers to the proportion of each genetic variation that involves the substitution of a single nucleotide (A, T, C, or G) for another single nucleotide within a DNA sequence for each of the four nucleotides.
<b>Input</b>	VCF
<b>Tool</b>	Bcftools
<b>Where?</b>	MultiQC report ('{outdir}/multiqc/multiqc_report.html')

	<ul style="list-style-type: none"> <li>• 'Variant Substitution Types' - Bcftools Workflow output ('{outdir}/.')</li> <li>• 'reports/bcftools/.'</li> </ul>
<b>Why?</b>	<ul style="list-style-type: none"> <li>• The "Percent SNV changes" metric from Bcftools calculates the proportion of single nucleotide variant (SNV) changes of a specific type compared to the total number of SNVs identified in the dataset. For example, "% C &gt; T" represents the percentage of SNVs where a cytosine (C) nucleotide in the reference genome is changed to a thymine (T) nucleotide. These specific SNV changes, known as substitution patterns or types, can provide insights into underlying mutational processes, such as deamination or exposure to mutagens, and are relevant in various genomic studies, including cancer research and population genetics.</li> <li>• In the Sarek nf-core workflow, analyzing the percent SNV changes contributes to understanding the mutational landscape of the samples, aiding in the identification of mutational signatures, potential carcinogens, or genomic regions under selective pressure. This metric facilitates the interpretation of genomic data and provides valuable information for investigating biological and evolutionary processes.</li> </ul>

# References

1. IMPACT-Data. (n.d.). Retrieved May 22, 2024, from <https://impact-data.bsc.es/>
2. nf-core. GitHub - nf-core/sarek: Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing. Retrieved from <https://github.com/nf-core/sarek>
3. Garcia M, Juhos S, Larsson M et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants [version 2; peer review: 2 approved] F1000Research 2020, 9:63 doi: 10.12688/f1000research.16665.2.
4. Friederike Hanssen, Maxime U Garcia, Lasse Folkersen, Anders Sune Pedersen, Francesco Lescai, Susanne Jodoin, Edmund Miller, Oskar Wacker, Nicholas Smith, nf-core community, Gisela Gabernet, Sven Nahnsen Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery bioRxiv doi: 10.1101/2023.07.19.549462.
5. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020 Mar;38(3):276-278. doi: 10.1038/s41587-020-0439-x. PMID: 32055031.
6. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
7. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr 11;35(4):316-319. doi: 10.1038/nbt.3820. PubMed PMID: 28398311.
8. Merkel, D. 2014. Docker: lightweight linux containers for consistent development and deployment. Linux Journal, 2014(239), 2. doi: 10.5555/2600239.2600241.
9. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017 May 11;12(5):e0177459. doi: 10.1371/journal.pone.0177459. eCollection 2017. PubMed PMID: 28494014; PubMed Central PMCID: PMC5426675.

10. Anaconda Software Distribution. Computer software. Vers. 2-2.4.0. Anaconda, Nov. 2016. Web.
11. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J; Bioconda Team. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods*. 2018 Jul;15(7):475-476. doi: 10.1038/s41592-018-0046-7. PubMed PMID: 29967506.
12. Nextflow DSL 2 is here! | Nextflow. (n.d.). Retrieved from <https://www.nextflow.io/blog/2020/dsl2-is-here.html>
13. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. 2010 Apr;38(6):1767-71. doi: 10.1093/nar/gkp1137. Epub 2009 Dec 16. PMID: 20015970; PMCID: PMC2847217.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.
15. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*. 2011 Aug 1;27(15):2156-8. doi: 10.1093/bioinformatics/btr330. Epub 2011 Jun 7. PMID: 21653522; PMCID: PMC3137218.
16. Caetano-Anolles, D. (2024, March 21). VCF - variant call format – GATK. GATK. <https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format>
17. Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, *Bioinformatics*, Volume 34, Issue 17, 01 September 2018, Pages i884–i890, doi: 10.1093/bioinformatics/bty560. PubMed PMID: 30423086. PubMed Central PMCID: PMC6129281
18. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

19. Brent S Pedersen, Aaron R Quinlan, Mosdepth: quick coverage calculation for genomes and exomes, *Bioinformatics*, Volume 34, Issue 5, 01 March 2018, Pages 867–868. doi: 10.1093/bioinformatics/btx699. PubMed PMID: 29096012. PubMed Central PMCID: PMC6030888.
20. “Picard Toolkit.” 2019. Broad Institute, GitHub Repository. <https://broadinstitute.github.io/picard/>; Broad Institute
21. McKenna A, Hanna M, Banks E, et al.: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19. PubMed PMID: 20644199; PubMed Central PMCID: PMC2928508.
22. Twelve years of SAMtools and BCFtools Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li *GigaScience*, Volume 10, Issue 2, February 2021, giab008, <https://doi.org/10.1093/gigascience/giab008>
23. TCLamnidis. GitHub - TCLamnidis/Sex.DetERRmine: A python script to calculate the relative coverage of X and Y chromosomes, and their associated error bars, from the depth of coverage at specified SNPs. Retrieved from <https://github.com/TCLamnidis/Sex.DetERRmine>
24. Pedersen BS, Bhetariya PJ, Brown J, Kravitz SN, Marth G, Jensen RL, Bronner MP, Underhill HR, Quinlan AR. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* 2020 Jul 14;12(1):62. doi: 10.1186/s13073-020-00761-2. PMID: 32664994; PMCID: PMC7362544.
25. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res.* 2018 Aug 24;7:1338. doi: 10.12688/f1000research.15931.2. PMID: 30254741; PMCID: PMC6124377.
26. bcftools. (n.d.). Retrieved from <https://www.htslib.org/doc/1.0/bcftools.html>
27. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PubMed PMID: 27312411; PubMed Central PMCID: PMC5039924.

