Practical guidelines for quality control of WGS results in population-scale initiatives

## A project proposal for GHIF

*Initiators: Mar Gonzalez-Porta, Justin Zook, Nate Olson*

* *Sponsor: Oliver Hofmann to*

Robust quality control (QC) of results is key for the successful delivery of population-scale sequencing efforts, even more so when the scope of such efforts comprises clinical samples.

Multiple studies have discussed the need to evaluate the performance of variant calling pipelines prior to introducing them in production [12-16], and consortia like GIAB and SEQC-II have provided practical guidelines on how to achieve that using well-characterised cell lines (i.e. reference materials or RMs) [17,20]. However, cell-line based RMs do not fully represent the heterogeneity and diversity of real samples nor give metrics for every sample run, and it is thus important to continue to monitor the quality of results beyond initial methods validation.

For production samples, QC of WGS results can be achieved through a range of tools that compute metrics from FASTQ, BAM, and/or VCF files (see e.g. [1-11]). Recommendations on which metrics to include in routine QC have been discussed in published guidelines [12-16] and continue to be actively developed [18-19]. Nonetheless, contrary to the situation encountered when discussing best practices for benchmarking [17], standardised definitions and implementations of recommended QC metrics have yet to be addressed. For example, guidelines will often refer to the need to track genome coverage, and multiple tools exist to directly report this metric or to produce intermediate outputs that can be used to calculate it (i.e. picard, samtools, sambamba, indexcov, mosdepth to name a few). Yet the interpretation and comparison of results require an accurate understanding of how the metric was calculated, e.g. does it include all chromosomes or autosomes only, are N bases masked, does the metric exclude duplicates and soft-clipped bases, does it include any base quality or mapping quality filters, what is the window size used, etc. Even though such details can have a substantial effect on the results, they are often not reflected in the documentation of the tool, and it is necessary to inspect the source code to retrieve them.

Given the increase in the number of population-scale studies, we believe that creating a common framework for the QC of WGS results is needed to ensure that data generation adheres to published guidelines, in-turn establishing confidence in the data quality and facilitating the exchange of results across initiatives at a later stage. We propose to engage with GHIF/GA4GH participants and relevant tool developers to work on a reference implementation that would provide practical recommendations on this matter. In particular, such work would complement existing guidelines by providing (i) standardised definitions for key QC metrics, (ii) tools for calculating them, and (iii) benchmarking resources that would aid in the interpretation and monitoring of results.

## Delivering the proposal

During the Fall 2020 GHIF meeting, we made a call to the community to create standardised implementations for QC of WGS results [21]. In this document, we intend to kick off the effort by (i) encouraging interested initiatives to comment on their needs when it comes to WGS QC (user stories), (ii) define the scope of the proposal, (iii) participate in a proof-of-concept exercise to explore the degree of overlap across QC workflows, and (iv) identify next steps.

### User stories

**Data production**

When sequencing thousands of genomes, having a robust QC pipeline is key to ensure that data pass pre-defined QC requirements and that those are consistently met along the duration of the project. Given the large number of tools available to calculate QC metrics, initiatives will often be faced with a first curation step to decide which metrics to carry forward for ongoing monitoring. Sometimes, multiple tools will report metrics that share a common name but, when compared, will have different values given the same input data. This difference in implementations is not always obvious, and adds complexity to the curation. In addition, if an external sequencing provider is engaged, initiatives will need to make sure that they adopt exactly the same metrics so that contractual requirements can be adequately evaluated. Altogether this can be a time-consuming exercise.

**Data sharing**

When accessing results from an external study, initiatives will often seek to select samples that meet certain QC requirements. For example, an initiative may be interested in sourcing external samples that have been sequenced at an equivalent depth to their own. If this information is not available when querying the external data, or if it’s not been defined with a common language, they will need to spend time generating it, and hence analysing data that eventually may not be relevant.

Sample Quality Metrics

“Not all coverage is equal.” “More coverage is not always better.” “Coverage coming from quality samples is better than coverage coming from low quality samples.” Sample QC metrics are not standardized! Labs have certain thresholds on how samples perform other than coverage information. It reflects the overall amount of DNA available. I have seen where samples look perfectly good in terms of coverage metrics such as average coverage depth but when analyzed for variants, lots of false positives come up. I found out that low DNA concentration seems to correlate with a higher number of false positives. Higher than usual de novo variants, higher than usual CNV calls.

**Others**

*<Please expand as needed>*

### Scope

For the initial round of this project, we propose to focus on QC of human WGS datasets (germline), generated with short-read technologies (Illumina) for research or clinical use.

Whilst the actual shape of the project is open for discussion and will be guided by the participating initiatives, the proposal has been based on previous work from the GA4GH Benchmarking Team in defining best practices for benchmarking germline small variant calls. Their work has been compiled in a publication [17] and GitHub repository [22], and was structured in 3 areas: definitions, reference tool implementations, and benchmarking resources. We intend to adopt an equivalent structure here.

**Metric definitions**

* Compile a list of commonly used metrics with detailed definitions.
* Define which metadata needs to be included when sharing metrics results to make it easier to understand what’s being assessed, e.g. has the metric been calculated on the entire genome / a subset, does it include duplicates, does it have any baseQ/mapQ filters… This would make it easier for the data holder to decide if two metrics can be directly compared. It will also guide the extent of the definitions compiled in the previous item.
* Explore defining a new file format/schema to make it easier to report QC metric outputs; e.g. could use a combination of bitwise flags and optional info fields to capture how a given metric has been calculated.

**Reference tool implementations**

* Compile a list of tools/pipelines that can be used to obtain the set of metrics discussed above.
* Engage with tool developers to close any implementation gaps.
* Explore making our tools/pipelines available within GA4GH’s TRS.

**Benchmarking resources**

* Run our reference tools on publicly available data from GIAB samples. We could then discuss which outputs are “directly comparable” vs. which remain “functionally equivalent”.
* Compile “common practices” for QC from participating initiatives, e.g. setting thresholds, identifying outliers, inspecting shifts in trends... This would provide a chance to explain how the proposed QC fits into the population study workflow and how the metrics are used. It could be a general discussion or we could identify a relevant dataset to showcase results. We wouldn’t expect to find a single solution here, instead we would aim to share what the different initiatives are doing.

### Proof of concept

To get a first sense of the current QC setup across initiatives, we propose a “proof of concept” exercise, whereby we encourage interested parties to run their existing sample QC pipeline on publicly available data, and submit its outputs and metric definitions to a [public GitHub repository](https://github.com/c-BIG/wgs-sample-qc/tree/main/proof-of-concept). This will allow us to assess how much overlap there is across workflows and kick off discussions on which outputs are “directly comparable” vs. which remain “functionally equivalent”. We kick off this knowledge-sharing exercise by providing details from the Singapore National Precision Medicine programme (Singapore NPM), and encourage others to add to it.

#### Singapore NPM

**Submission**

<https://github.com/c-BIG/wgs-sample-qc/tree/main/proof-of-concept/sg-npm>

**Comments**

* We are establishing our QC workflow for NPM phase 2 (100K genomes) and would benefit from a set of GHIF-recommended metrics to ensure that our processes remain aligned with the community.
* Singapore NPM comprises multiple partners who will access not only variant calls but also raw data (CRAM). We have a need to share QC metrics alongside sequencing results, so that collaborators involved in downstream analysis can better define relevant cohorts of samples for their studies. We would benefit from community-backed ways of sharing the QC results.
* Our QC pipeline has been designed to run independently of our alignment and variant calling workflow. This was an intentional decision to ensure that the same QC checks can be run regardless of choice of upstream analysis, and remains a requirement for our QC process. We are aware that this comes at a cost in speed, but this is not our focus at the moment.
* See here for our current QC pipeline: <https://github.com/c-BIG/NPM-sample-qc>. It combines several commonly used tools to report >80 QC metrics, and supports auto-generation of metric definitions from code comments. We are open to updating it to align with the current proposal.

#### *<Include your initiative here>*

## 

## References

1. samtools: <http://www.htslib.org/doc/samtools.html>
2. Bcftools: <http://samtools.github.io/bcftools/>
3. Picard: <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. *Bioinform Oxf Engl* **32**, 3047–8 (2016).
5. Zhang, F. *et al.* Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res* (2020) doi:10.1101/gr.246934.118.
6. FastQC: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
7. Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor–normal pairs. *Bioinformatics* **32**, 3196–3198 (2016).
8. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinform Oxf Engl* **32**, 292–4 (2015).
9. Yan, M. Y., Ferguson, B. & Bimber, B. N. VariantQC: a visual quality control report for variant evaluation. *Bioinform Oxf Engl* (2019) doi:10.1093/bioinformatics/btz560.
10. Pedersen, B. S., Collins, R. L., Talkowski, M. E. & Quinlan, A. R. Indexcov: fast coverage quality control for whole-genome sequencing. *Gigascience* **6**, 1–6 (2017).
11. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2017).
12. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* **15**, 733–747 (2013).
13. Roy, S. et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines. J Mol Diagnostics 20, 4–27 (2018).
14. U.S. Food and Drug Administration (FDA). Considerations for Design, Development, and Analytical Validation of Next Generation Sequencing-Based In Vitro Diagnostics Intended to Aim in the Diagnosis of Suspected Germline Diseases. Updated 13 April 2018.
15. Aziz, N. *et al.* College of American Pathologists’ Laboratory Standards for Next-Generation Sequencing Clinical Tests. *Arch Pathol Lab Med* **139**, 481–493 (2015).
16. Gargis, A. S. *et al.* Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol* **30**, 1033–1036 (2012).
17. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**, 555–560 (2019).
18. Initiative, M. G. *et al.* Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. *Npj Genom Medicine* **5**, 47 (2020).
19. TC215/SC1 - ISO standard on “Quality Control Metrics for DNA Sequencing”
20. Fang, L. T. et al. Establishing reference samples for detection of somatic mutations and germline variants with NGS technologies. Biorxiv 625624 (2019) doi:10.1101/625624.
21. <https://broadinstitute.swoogo.com/ghif2020/agenda>
22. <https://github.com/ga4gh/benchmarking-tools>

## 