A quality control portal for sequencing data deposited at the European genome–phenome archive

Dietmar Fernández-Orth, Manuel Rueda, Babita Singh, Mauricio Moldes, Aina Jene, Marta Ferri, Claudia Vasallo, Lauren A. Fromont, Arcadi Navarro and Jordi Rambla

Abstract

Since its launch in 2008, the European Genome–Phenome Archive (EGA) has been leading the archiving and distribution of human identifiable genomic data. In this regard, one of the community concerns is the potential usability of the stored data, as of now, data submitters are not mandated to perform any quality control (QC) before uploading their data and associated metadata information. Here, we present a new File QC Portal developed at EGA, along with QC reports performed and created for 1 694 442 files [Fastq, sequence alignment map (SAM)/binary alignment map (BAM)/CRAM and variant call format (VCF)] submitted at EGA. QC reports allow anonymous EGA users to view summary-level information regarding the files within a specific dataset, such as quality of reads, alignment quality, number and type of variants and other features. Researchers benefit from being able to assess the quality of data prior to the data access decision and thereby, increasing the reusability of data (https://ega-archive.org/blog/data-upcycling-powered-by-ega/).

Keywords: Fastq, quality control, variant call format (VCF), binary alignment map (BAM), European Genome-Phenome Archive (EGA)

Introduction

Next-generation sequencing (NGS) has become the leading method for deoxyribonucleic Acid (DNA) sequencing due to its capacity to process millions of DNA molecules in a single experiment. Despite its wide use, NGS still suffers from several concerns. At the experimental (wet-lab) level, NGS has been shown to display random errors and systematic biases, including: polymerase chain reaction amplification problems, GC-content (or guanine-cytosine content) shift and contamination [1]. Some of these aspects can be quality controlled a posteriori using bioinformatics tools [2]. Historically, quality control (QC) has been based on checking ‘reference’ values for chosen parameters, yet recently algorithms based in machine learning methods have been introduced [3, 4]. Not only the experimental part of NGS is prone to errors, the downstream analysis of the raw data (dry lab) in the form of ‘pipelines’ (including genome mapping, variant calling, etc.) can create biases as well. Therefore, a robust bioinformatics-based QC is a critical step to the correct interpretation of sequencing results.

The European Genome–Phenome Archive (EGA) is a public repository of human genomic and phenotypic data, tasked with the secure maintaining and distribut-
ing of data. EGA is maintained in collaboration between the European Bioinformatics Institute and the Centre for Genomic Regulation [5, 6]. Since its launch in 2008, the EGA has been leading the archiving and distribution of human identifiable genomics data that require controlled access. To date, over 4890 studies, containing up to 7690 datasets, have been deposited and are available to authorized researchers.

In an effort to facilitate researchers the selection of the fittest data for their analysis, the EGA has assembled a set of QC procedures for the file formats where data are stored and distributed (i.e. Fastq, Sequence Alignment Map (SAM), Binary Alignment Map (BAM), Compressed version of BAM (CRAM) and Variant Call Format (VCF) files). Reports resulting from these procedures are embedded in HTML and available through the main EGA website. EGA users can visualize the main attributes of the deposited files and obtain an overall idea about its quality and potential reusability before starting the access requesting process.

Methods

Stored files at EGA

One particular challenge faced by the EGA is how to deal with the great diversity of file formats submitted to it. For nonphenotypic data, the EGA accepts a variety of formats that span all the way from raw data (microarray, Illumina Fastq and Complete Genomics) to processed ones like BAM, CRAM or VCF. In that sense, EGA is a driver project for the Global Alliance for Genomics Health and verifies the deposited data following the specifications for BAM/VCF data.

Performing QC for these different formats requires the use of different tools. We focused our analyses on the file formats that are popular in bioinformatics and that have been deposited at the EGA in significant numbers. At the time of planning for the file QC procedures, Fastq, BAM/CRAM and VCF account for more than 60% of all files deposited at EGA.

File formats analysed and selected tools

Fastq is the de facto standard format for storing the output of NGS instruments. It combines the sequence as well as an associated per base quality score (PHRED score) of a base call. For Fastq files, QC parameters commonly checked include (i) filtering low-quality reads (according to thresholds in PHRED scores) in order to avoid subsequent false positives, and (ii) keeping contamination and other features under control [2].

Several options exist to perform QC on Fastq files [2, 7, 8]. The EGA has selected FastQC because it encompasses most of the desired QC options (e.g. quality score/base-call distribution, detection of contaminants such as adaptors and detection of duplicates, etc.) and it is recognized as the gold standard tool by the community.

Reports generated by FastQC include a section with basic statistics like a number of sequences, % GC, etc., followed by some plots showing diverse features.

SAM is a format for storing read alignments against reference sequences. BAM and CRAM format files are compressed versions of SAM [9, 10]. Popular QC parameters checked include coverage, length of reads and percent of mapped reads. Among stats for QC in SAM/BAM/CRAM, those usually checked are: nonreference allele frequency, depth distribution, stats by quality and per-sample counts and singleton abundance. Taking into account the diverse options available to perform QC on aligned files [11, 12], the EGA has chosen SAM tools as a QC tool for alignment formats, as it allows extracting these features in an efficient way while being a gold standard in bioinformatics analysis [13].

VCF format is a format for storing genomic variations such as single nucleotide polymorphisms (SNPs), insertions, deletions and structural variants together with annotations [14]. Interesting tags for QC that may be taken into account for VCF files are allele frequency, depth distribution, stats by quality and per-sample counts and singleton stats [11, 15]. The tools selected by EGA are vcf tools (options: —TsTv-summary, —SNPdensity 1000, —site-quality, —freq) and bcf tools (options: stats).

File QC implementation

The EGA File QC portal has two components: (i) the QC pipeline, that uses the set of tools described above to generate a report for each file. (ii) The frontend pages, which presents the report results in user-friendly graphics, also per each file. The front end generates an HTML5 webpage using the Django Python framework, and the D3 Javascript library. The front-end webpage is inspired by the existing http://iobio.io/ website for BAM and VCF files adding some custom features [16].

Results

QC reports can be accessed by anonymous EGA users from ega-archive website, browse-dataset-files page, where we present summary-level information about each dataset file and reference to its corresponding QC report. Researchers can also access each QC report, directly from a reference found in a paper or by browsing the EGA catalog.

As a typical user journey, a user could start querying for key tags in the search box on the main page of the EGA website (we searched for H3AFRICA). Among the results, there is a list of datasets matching the query (we selected H3AFRICA TRYPANOSU2). Every dataset details page shows a ‘Browse Files’ button that forwards the user to the list of files available for that dataset (Figure 1). Once there, a table provides information related to the files (Figure 1A) including a link to the QC report when available (Figure 1B). By clicking on the link, the File Portal pops up for the selected EGA file (VCF example: https://filesportal.ega-archive.org/EGAF00002052188).
H3AFRICA TRYPANOCEN2

Dataset ID | Technology | Samples
---|---|---
EGAD00001004220 | Illumina HiSeq 2500 | 41

### Dataset Description

41 samples from Zambia generated for the H3Africa Chip Design Study. The dataset includes BAM, FASTQ and decompressed gVCF files.

| 1237 Files | 10.26 TB |

**Figure 1.** EGA website. Primary file information and how to access the QC report. **(A)** List with the EGA ID files composing the dataset. **(B)** Link to the File QC for each specific ID (https://ega-archive.org/datasets/EGAD00001004220/files).

Also, in ega-archive.org navigation header, About Section, we have added the ‘Quality Control Reports’ (https://ega-archive.org/about/quality-control-reports). Here, we detail on how to access and use QC reports, with a step-by-step guide.

### BAM/VCF QC report description

For BAM and VCF files, the report is divided into two sections (**Figure 2**). **‘File Information’** (**Figure 2A**) shows some sample attributes, as well as information about the origin of the dataset and study. Furthermore, information about publications in which that file has been included is enclosed. The header of the BAM/VCF files is also viewable (**Figure 2B**), allowing the user to get an idea about which workflow and tools were used to generate that specific file. For BAMs, the link to stats plot displays the results of the plot bamstats script according to the SAM tools reference manual (**Figure 2B**).

The second section shows different plots depending on the type of file (**Figure 3**). BAM files comprise a series of plots giving information about base coverage distribution/quality, number of mapped reads, singletons and duplicates, among others (**Figure 3A**). For VCF files, included plots are: variant types, Ts/Tv ratio and variants quality. Details about each plot are available by clicking the information icon included in each plot. Such explanations help to understand/interpret them (**Figure 3B**).

On the upper right side of each plot shown, there is an information tag which opens a detailed description of every plot. This may help users to understand it.

### Discussion

Currently, there are close to 4 million files, summing 14 PB of data stored at the EGA. Up to November 2021, nearly 70% of them came from NGS sequencing results.

As shown in **Figure 4**, Fastq, BAM/CRAM and VCF constitute 99% of the total number of files, and approximately 91% of them have been analysed by the File QC portal workflow successfully. Those failing the QC are usually due to being corrupted or showing unexpected results. These files are then flagged and potential issues are resolved after requesting information from the submitters.

The file QC portal allows for checking the quality of the files, prior to downloading them. On par with the community, we run FastQC, SAM tools and BCF tools on deposited files, as these tools allow us to obtain robust statistics about the quality of the files. For each plot displayed, general feedback on the data is given in the information section, which enables the research community to instantly decide if the data are of suitable quality for their research purpose(s).

In order to decide if a file fulfills the quality criteria, it is recommended to check the ‘File Information’ section (**Figure 2**) as it allows detecting its main characteristics.
at a glance. Each plot is designed to check for different file parameters. A short text explaining how to interpret each plot is provided, helping the researcher to decide the suitability of the file (Figure 3B).

Files are not classified or tagged as having ‘good’ or ‘bad’ quality as that criteria depends on the purpose of the analysis to be done and should be applied by the potential requester.

We welcome users to contribute or suggest additional features to be evaluated and added to our File QC procedure and report. Suggestions can be made by contacting the EGA’s Helpdesk staff. Future implementations of QC on other file formats such as genome-wide association studies in plink format are under development. Regarding Fastq files, integration of MultiQC results for englobing all files within the same dataset is under study [17]. The File QC reports are free, publicly available and an open-source licensed resource.

**Key Points**

- We present a new File QC portal, ready to be used for any desired deposited dataset by the research community to check if Fastq, BAM/CRAM and VCF files within the EGA dataset fulfill applicants’ quality requirements to be used in their own analysis.
A quality control portal developed at EGA

- We analysed quality parameters for more than 1,500,000 files stored at the EGA comprising Fastq, BAM/CRAM and VCF files.
- We used FastQC, SAM tools and BCF tools/VCF tools as gold standard open source tools for checking the quality and generated user-friendly plots to allow users an easy interpretation.

Acknowledgements

We want to thank the team of developers of https://iobio.io/ in the Marth Lab at the Center for Genetic Discovery at the University of Utah as they have inspired us to the creation of our File QC portal. Additionally, acknowledge the support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme/Generalitat de Catalunya.

Funding

The File QC feature project has received funding from Horizon 2020 ELIXIR-CONVERGE project (grant agreement No 871075), the ELIXIR-FHD-IS and LaCaixa Foundation (LCF/PR/CE20/50740008).

References

1. Ari Ş, Arikan M. Next-Generation Sequencing: Advantages, Disadvantages, and Future. In: Hakeem KR, Tombuloglu H, Tombuloglu G (eds), Plant Omics: Trends and Applications. Springer International, 2016, 109–35. https://doi.org/10.1007/978-3-319-31703-8_5.
2. Paszkiewicz KH, Farbos A, O’Neill P, et al. Quality control on the frontier. Front Genet 2014;5:1.
3. Li J, Jew B, Zhan I, et al. Forest QC: quality control on genetic variants from next-generation sequencing data using random forest. PLoS Comput Biol 2019;15 e1007556.
4. Albrecht S, Sprang M, Andrade-Navarro MA, et al. seqQscorer: automated quality control of next-generation sequencing data using machine learning. Genome Biol 2021;22:75.
5. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European genome-phenome archive of human data consensed for biomedical research. Nat Genet 2015;47:7 2015.
6. Freeberg MA, Fromont LA, D’Altri T, et al. The European genome-phenome archive in 2021. Nucleic Acids Res 2021;D980–7.
7. Chanumolu SK, Albahran M, Otu HH. FQStat: a parallel architecture for very high-speed assessment of sequencing quality metrics. BMC Bioinformatics 2019;20:424.
8. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 2011;27:863–4.
9. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.
10. Fritz MHY, Leinonen R, Cochrane G, et al. Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res 2011;21:734–40.
11. Ivanov M, Ivanov M, Kasianov A, et al. Novel bioinformatics quality control metric for next-generation sequencing experiments in the clinical context. Nucleic Acids Res 2019;47:e135.
12. Rausch T, Hsi-Yang Fritz M, Korbel JO, et al. Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. Bioinformatics 2019;35:2489–91.
13. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience 2021;10:giab008.
14. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156–8.
15. Yan MY, Ferguson B, Bimber BN. Variant QC: a visual quality control report for variant evaluation. Bioinformatics 2019;35:5370–1.
16. Miller CA, Qiao Y, DiSera T, et al. bam.iobio: a web-based, real-time, sequence alignment file inspector. Nat Methods 2014;11:1189–9.
17. Ewels P, Magnusson M, Lundin S, et al. Multi QC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32:3047–8.