Deliverable

D2.3 Guidelines for Quality Control metrics	
Version | Status	V1 | final
Work package	WP2
Lead beneficiary	CNAG-CRG
Due date	31.12.2019 (M24)
Date of preparation	18.11.2019
Target Dissemination Level	Public
Author(s)	CNAG-CRG: Leslie Matalonga, Jean-Rémi Trotta, Steven Laurie, Carles Garcia, Marcos Fernández, Sergi Beltran RUMC: Christian Gilissen
Reviewed by	Karolis Sablauskas (RUMC), Stephan Ossowski (EKUT)
Approved by	Alexander Hoischen (RUMC)

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257.

Explanation according to GA Annex I:

Guidelines for Quality Control metrics provided by the Data Analyses Task Force.

Abstract:

Solve-RD has collated over 8,400 standardised phenotypic and genomic datasets from partners across different ERNs and countries. To ensure best practices and standardisation of the process, HPO[1], OMIM[2] and Orphanet[3] (ORDO) ontologies are used to collect phenotypic data, and GATK best practices[4] and GA4GH[5] standards are followed in the collection and processing of genomic data through a standardised pipeline (Laurie et al., 2016).

As data submitted to Solve-RD for reanalysis has been sequenced at a variety of different centres, under different protocols and using different technologies it is fundamental to ensure a minimum quality of geno-pheno datasets to guarantee proper downstream analyses and results. Therefore, several quality control metrics have been established, and are now automatically performed for each of the samples entering the Solve-RD project.

Here we report on the established framework for quality assessment (processing checkpoints, genome coverage metrics, phenotypic data and sample relatedness) for RD-Connect and Solve-RD data and provide guidelines to enable Solve-RD data submitters and Data Analysis Task Force (DATF) members to easily assess the quality of the provided data and compare genomic datasets before undertaking further downstream analyses.

Introduction:

Over 8,400 genomic datasets (~8000 exomes and ~400 genomes) together with their corresponding phenotypic information have been collected and processed by Solve-RD (deliverable D2.6).

To ensure best practices and standardisation of the process, HPO[6], OMIM[7] and Orphanet[8] (ORDO) ontologies are used to describe phenotypic data, and GA4GH[9] standards are followed in the collection and processing of genomic data through a standardised pipeline based on GATK best practices[10] (Laurie et al., 2016). Data, mainly in FASTQ, BAM, gVCF and Phenopackets formats, is made available to all Solve-RD partners either through the RD-Connect Genome-Phenome Analysis Platform (GPAP)[11], the European Genome-Phenome Archive (EGA)[12] and/or the Solve-RD Sandbox environment.

To date, datasets coming from 32 different partners who are part of one of the four core ERNs (GENTURIS, ITHACA, NMD and RND) or a Solve-RD associated UDN program have been processed. This data has been sequenced at different centres, under different proto-

[1] https://hpo.jax.org/
[2] https://omim.org/
[3] https://www.orpha.net
[4] https://software.broadinstitute.org/gatk/
[5] https://www.ga4gh.org/
[6] https://hpo.jax.org/
[7] https://omim.org/
[8] https://www.orpha.net
[9] https://www.ga4gh.org/
[10] https://software.broadinstitute.org/gatk/
[11] https://platform.rd-connect.eu/
[12] https://www.ebi.ac.uk/ega/home

cols and using different technologies and therefore it is fundamental to ensure it meets a minimum quality threshold to guarantee proper downstream analyses and results. A specific framework for quality assessment has been established in the context of data submitted to the RD-Connect GPAP by Solve-RD. Several quality control steps and metrics have been set up and are automatically computed for each of the samples entering the study:

- **Quality control metrics for genomic data**

Several checkpoints during genomic data processing ensure that the different steps (sequencing, mapping and variant calling) have been successful. A dedicated coverage pipeline has been setup and is run on all the data submitted by Solve-RD to the RD-Connect GPAP. Corresponding quality control reports will be made available to all Solve-RD partners through the Sandbox or sent on demand.

Solve-RD also has a dedicated Working Group on "Relatedness and runs of homozygosity" led by Stephan Ossowski (EKUT) to ensure experimentally measured kinship is in accordance with that declared by collaborators in the corresponding phenotypic record, prior to performing any pedigree based analyses. In addition, relatedness between all submitted samples is calculated to identify double entries, multiple samples of a single patient diagnosed at multiple sites or unknown relationships between patients.

- **Quality control metrics for phenotypic data**

As specified in the provided guidelines[13] (deliverable D1.3[14]), several phenotypic information fields are mandatory when submitting phenotypic information. Compliance with these criteria is checked before generation and transfer of the corresponding phenopackets to the EGA / Sandbox.

Here we report on the established framework for quality assessment for RD-Connect and Solve-RD data and provide guidelines to enable Solve-RD data submitters and Data Analysis Task Force (DATF) members to easily assess the quality of the data and compare genomic datasets before undertaking further downstream analyses.

Report:

In this report we describe the different quality metrics that are being performed on all submitted Solve-RD data and will be made available to Solve-RD partners. They are divided in four sections:

1) Processing quality control
2) Coverage metrics
3) Phenotypic data
4) Relatedness information

1. Processing quality control

Processing here refers to the conversion of raw sequencing read data, submitted as either BAMs or pairs of FastQ files, through realignment and variant calling to generate gVCF files per chromosome for each submitted experiment. Experiments may fail during the standard

[13] https://rd-connect.eu/phenotips-guide/

[14] http://solve-rd.eu/wp-content/uploads/2019/02/D1.3-Training-modules-guidance-document-and-online-help-module-for-collection-of-phenotypes.pdf

processing for a variety of reasons. The most commonly encountered reasons are the following:

1) Submitted files are truncated or corrupted, preventing processing
2) FastQs are incorrectly paired, or read names are not unique
3) Read groups in read of BAM file do not match those declared in the header

Whenever such issues are encountered they are communicated back to the submitting group, and they are offered the opportunity to resubmit the correct version of the files for the experiment. When this is not possible, or when resubmitted files also fail during processing, the submitting group is informed again, and the corresponding sequencing data, metadata and phenotypic records are purged from the GPAP.

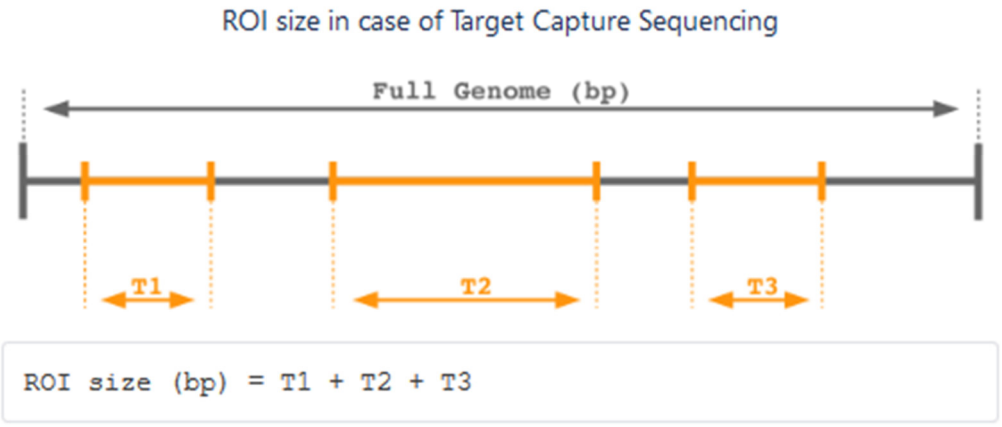
2. Coverage metrics

TSV files with the following quality metrics will be made available to Solve-RD partners for each of the submitted genomic datasets:

a) RoI Size

Region of Interest (RoI) size depends on the type of experiment performed (exome or genome sequencing in the context of Solve-RD). For genome sequencing it represents the size of the genome. For exome sequencing it represents the size of the BED file passed as targeted regions.

RoI size is calculated as follows:

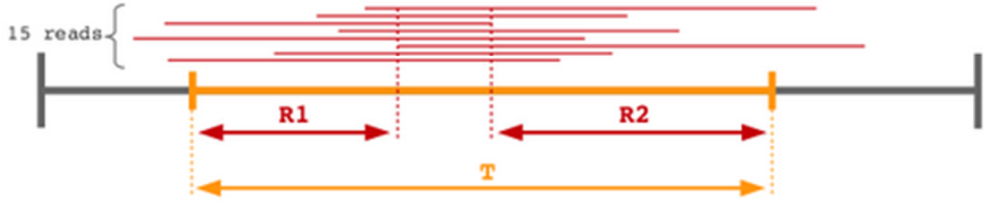


b) Mean and Median Coverage

Mean and median coverage is reported for the total RoI i.e. for the sum of the targets in the respective bed file.

c) Cn

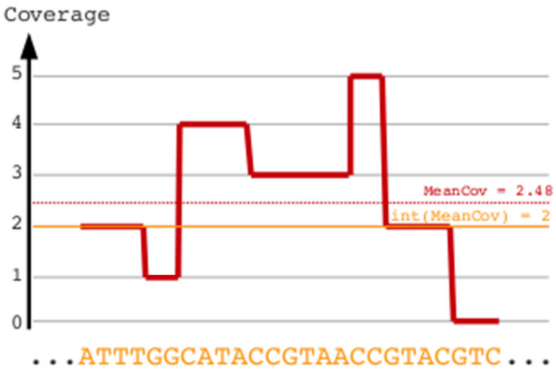
Percentage of bases sequenced in the genome or RoI with a depth of coverage greater than or equal to n. Metrics are provided for C1,C10,C15,C20,C30,C50,C100.




```
C15 = ( 1 - ( (R1+R2) / T ) ) *100
```

d) Evenness and Median Evenness

The more even the coverage across all RoIs, the better the quality of the data for calling SNVs and CNVs. Evenness represents the proportion of the RoI that is covered by at least the mean coverage across the entire RoI.




```
cov = round(MeanCov)
P(n) = Sum(i=1..n; #bases[base_coverage>=i])

Evenness = (P(cov) * 100) / (cov * #bases)
```

e) Sex Ratio

Sex Ratio is calculated as the ratio between mean coverage of ChrY compared to mean coverage of all autosomes. This allows us to compare the experimentally predicted sex to that submitted in the phenotypic record. Values around 0 are expected to be from females while values around 0.5 are expected to be from males.

f) Gene and Transcript Level Coverage

Mean and median coverage and Cn, as described in sections b and c above, are also provided individually for all Ensembl version75 and RefSeq coding genes and coding transcripts.

3. Phenotypic data

Phenotypic data is collated using HPO[15], OMIM[16] and Orphanet[17] (ORDO) ontologies. As described in deliverable D1.1, data can be submitted either using dedicated phenotypic forms[18] or by bulk upload using a specific excel file[19] provided by RD-Connect GPAP.

[15] https://hpo.jax.org/
[16] https://omim.org/
[17] https://www.orpha.net
[18] https://rd-connect.eu/phenotips-guide/
[19] https://drive.google.com/file/d/1R9DL6y7Cq-vAqW8DLQObw0L5EDqTdPa4/view

a) Minimum Fields Required

In the context of the project and to ensure proper phenotypic description for further analysis and interpretation of the (gen)omic data, several fields have been made mandatory: local ID, sex, year of birth, pedigree, affected status, 5 positive HPO terms (for index cases only) and clinical diagnosis (using ORDO ontology).

b) Phenopackets

Before creation and submission of the corresponding phenopackets[20] generated for phenotypic data export across the project, an internal quality control is performed to check that all mandatory information has been included in the record. Missing information is requested to the corresponding partner, although it is not always possible to obtain it all.

Further phenotypic quality checks are planned to be included in the phenopacket schema as part of the GA4GH Clinical Phenotype Data Capture Workstream. Once released, we plan to also include those in our data quality workflow.

4. Relatedness Information

Solve-RD has a dedicated Working Group (WG) on relatedness to provide users with relatedness scores between all the individuals analysed, accurate to 3rd degree. This WG aims to provide Data Interpretation Task Forces (DITFs) and ERNs with accurate and efficient relatedness scores for all individuals analysed in Solve-RD to ensure experimentally measured kinship is in accordance with that declared by collaborators in the corresponding phenotypic record, prior to performing any pedigree based analyses. Furthermore, we computationally identify consanguineous cases and compare the results with the phenotypic record. Finally, we identify unknown relationships between all submitted samples and identify double entries and highly similar genomes (indicating that samples of the same person were taken multiple times potentially in different hospitals). For these purposes we used the tools SampleSimilarity and RohHunter (both ngs-bits package), KING and Plink. Information will be made available in the Sandbox environment.

Conclusion:

We have established a framework for quality assessment (processing checkpoints, genome coverage metrics, phenotypic data and sample relatedness) for all WES and WGS Solve-RD datasets. Once data is processed, the corresponding quality metrics reports are provided to Solve-RD partners to ensure minimum quality requirements before undertaking further downstream analyses within the project.

[20] http://phenopackets.org/