EXCELERATE Deliverable D8.6

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	Documentation on adequate quality reporting standards for genomics datasets	
WP No.	8	
Lead Beneficiary:	CNAG-CRG	
WP Title	Use Case C: ELIXIR infrastructure for Rare Disease research	
Contractual delivery date:	31 May 2019	
Actual delivery date:	23 May 2019	
WP leader:	Ivo Gut and Marco Roos	8 - CRG; 6 - LUMC (LTP)
Partner(s) contributing to this deliverable:	CNAG-CRG, NBIC (LUMC)	

Authors and Contributors:

Justin P. Whalley, Jean-Rémi Trotta, Steve Laurie, Leslie Matalonga, Marta Gut, Sergi Beltran and Ivo G. Gut

Reviewers:

None

1. Table of contents

Horizon 2020 grant n. 676559

2. Executive Summary

Rare disease (RD) research faces particular challenges because patient populations, clinical expertise, and research communities are small in number and highly fragmented both geographically and in terms of medical specialty. The scarcity of rare disease patients and their corresponding (gen)omic data has made data sharing one of the fundamental pillars to fasten and improve patient diagnostic and to reach IRDiRC 2017-2027 vision to enable all people living with a RD to receive an accurate diagnosis, care, and available therapy within one year of coming to medical attention.

Different project such as NeurOmics[1], EurenOmics[2], RD-Connect[3] and more recently Solve-RD[4] and EJP-RD[5], infrastructures such as BBMRI and ELIXIR and initiatives such as GA4GH have been working towards this objective. Indeed, rare disease platforms such as RD-Connect GPAP[6] enable controlled data sharing of standardised phenotypic and genomic data. HPO[7], OMIM[8] and Orphanet[9] (ORDO) ontologies are used to collect phenotypic data and GATK best practices[10] and GA4GH[11] standards are followed to collect and process genomic data through a standardised pipeline (Laurie et al., 2016).

Collating genomic data from disparate centers across different countries has largely evidentiate to improve our understanding on rare diseases (Lochmüller et al., 2018[12]). However in order to fully benefit from this unprecedented access to genomic data, care must be given to determine the quality of these genomic datasets, especially when sequenced at different centres, under different protocols and using different technologies. In this sense, several metrics such as depth of coverage, base quality and mapping quality are already broadly used for NGS quality evaluation. However, due to the rapid development of the genome sequencing field, comprehensive quality management considerations are still scarce and although some efforts have been made, there is no current standards for genomic data quality comparison (Endrullat et al 2016[13] and Mahamdallie et al 2018[14]).

In this context, one of the specific objectives of EXCELERATE WP8 was to establish a framework for quality assessment of genomic data (Task 8.1.2) to enable rare disease researchers to easily compare genomic datasets, starting with whole exome sequencing data. In this deliverable, we have explored a rating system based on 5 different quality metrics. This rating system could be used as a starting point for continuing work in the

[1] https://www.neuromics.com/
[2] https://www.eurenomics.eu/
[3] https://rd-connect.eu/
[4] http://solve-rd.eu/
[5] http://www.ejprarediseases.org/
[6] https://platform.rd-connect.eu/
[7] https://hpo.jax.org/
[8] https://omim.org/
[9] https://www.orpha.net
[10] https://software.broadinstitute.org/gatk/
[11] https://www.ga4gh.org/
[12] doi: 10.1038/s41431-018-0115-5
[13] doi: 10.1016/j.atg.2016.06.001
[14] doi: 10.12688/wellcomeopenres.14307.1

Horizon 2020 grant n. 676559

context of WGS sequencing data and the 1M genomes declaration as federated systems across endorser countries will require to compare WES / WGS of different origin.

3. Impact

This deliverable provides the RD-Community with the bases for a reliable framework for assessing genomic data quality WES samples. This work can be used as a starting point to define global quality assessment standards for WES and WGS data in the context of the 1M genomes declaration and the MEGA project.

4. Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Demonstrate, in partnership with the Rare Disease community, how aligned ELIXIR resources enable research, avoid fragmentation and support the development of sustainability models for resources created by the community research projects.	X	

5. Delivery and schedule

The delivery is delayed: Yes • No☑

6. Adjustments made

Not applicable

7. Background information

Background information on this WP as originally indicated in the description of action (DoA) is included here for reference.

Horizon 2020 grant n. 676559

Work package number	8	Start date or starting event:	month 1
Work package title	**Use Case C: ELIXIR infrastructure for Rare Disease research**		
Lead	Ivo Gut (ES) and Marco Roos (NL)		

Participant number and person months per participant

4 - UNIMAN 6.00; 6 - NBIC 0.00 LUMC 6.00; 8 - CRG 38.40; 9 - CIPF 2.66; 12 - BSC 10.00; 22 - NTNU 12.00; 26 - CNRS 12.00; 30 - CNR 6.39; 32 - UL 15.00; 38 - DTU 12.00; 47 - FPS 4.54

The International Rare Diseases Research Consortium (http://www.irdirc.org) established the ambitious goal of developing 200 new therapies by 2020. ELIXIR as a whole and in particular this Work Package is aligned with this effort. The overall objective of this Work Package (WP) is to address the needs of the rare diseases community through the instantiation of the ELIXIR resources described in WP1-5. These resources do not constitute a replacement of the current research projects organized around the rare diseases area. Indeed the aim is to empower them and to help in the sustainability of the resources created by these projects in the long term. This WP is organised around the actors that play a major role on the development of these new therapies. These actors are the main users of the ELIXIR infrastructure: data generators and curators (usually personnel working in hospitals, genomics-based companies, and members of large research consortia), researchers (bioinformaticians, geneticists, and clinical doctors), diagnosis companies, CROs (usually SMEs), and the pharmaceutical industry among others.

Objectives

WP8 aims to empower actors involved in the development of new rare diseases therapies through the execution of the following specific objectives:

1. Build the ELIXIR registry of data resources and analysis tools critical for the development of the rare disease research. (Task 8.1)
   - Continuous monitoring of resources and tools in Rare-diseases.
   - Implementation of a system for the generation of datasets adequate for the assessment of methods in the area of rare- diseases.
   - Implementation of the ELIXIR rare-disease portfolio in the ELIXIR registry.
2. Implementation of a technical framework for the comparison and standardization of services useful for the rare-disease communities. (Task 8.2)

Horizon 2020 grant n. 676559

3. Collaboration with the rare-disease communities for the organization of training courses, workshops and jamborees. (Task 8.3)

Work Package Leads: Ivo Gut (ES) and Marco Roos (NL)

Description of work and role of partners

Task 8.1: The ELIXIR portfolio of data resources developed in collaboration with the rare diseases communities (69.4PM)

Subtask 8.1.1 Monitoring of resources and tools. (25.4PM)

There is a wide range of data resources and analysis methods used in the rare-disease area. Many of those resources are provided by ELIXIR Nodes, for example the European Genome-Phenome archive (EGA) currently stores data from major research initiatives in rare diseases like the RD-connect project. In this subtask we will review the current data resources and evaluate their usability and potential impact on the rare disease community. An important aspect of the evaluation will be the security of the data that is a key aspect in rare disease domain given the low frequency of the

associated genomic variants in the population.

One critical aspect of the development of the registry is to engage the different communities in the submission and rating of the tools. In this task we will work together with representatives of the major projects in the field of rare- diseases to create a customized portfolio of ELIXIR tools and services devoted to assist them in the development of these new therapies. As an example we will ask for proposals of tools that serve to interpret the effect of genomics variants on a group of patients that belong to the same family. We strongly believe that this link between the end- users and the tools developers will help ELIXIR to understand better the problems that are actually facing the main actors in the rare diseases research and hence to better solutions. The final outcome of this task will be the ELIXIR data resources and analysis tools useful to the rare disease communities.

Partners: NO, ES, SI, IT, NL

Subtask 8.1.2: Creation of reference datasets adequate for the specific assessment of methods and standards in the area of rare-diseases. (30PM)

While the creation of these tools should stay as a priority for researchers, large scale projects, SMEs and the industry increasingly need access to benchmarked methods on which to build their analysis strategies.

The evaluation of the methods requires the adequate selection of the datasets and benchmarking strategies. The systems for the selection of the datasets for the benchmarking have to be fast and effective to enable the continuous evaluation of the methods, as described in WP2. We will collaborate with the ELIXIR benchmarking strategy (WP2) to build the appropriate strategies for the selection of the datasets (subtask 8.1.1 above) and with the rare- disease communities to implement the adequate quality reporting standards. Moreover we will integrate these pipelines in the ELIXIR

benchmarking framework (WP2) to continuously monitor the selected methods with the newly generated datasets.

Partners: ES, DK, IT, FR, SI, UK

Subtask 8.1.3 Implementation of the ELIXIR rare-disease portfolio in the ELIXIR registry. (14PM)

The ELIXIR registry will be a reference for the research community (WP1), as it will reflect the quality and the real-time status of the services included on it. This registry will act as a one-stop shop for services provided by ELIXIR. The goal is to allow users from the different countries, communities and projects to discover which are the tools available at a given time, with the associated information about the community based rating (see WP2), instructions for correct use and associated examples We will encourage tools developers to adopt the EDAM standard to describe their tools and to share several metrics about the performance and usage of these of the tools (see description in WP1)

Those services promoted as relevant by the end-users will be listed in a special section in the ELIXIR registry.

Partners: DK, ES, FR.

Task 8.2: Standardisation of rare disease services in collaboration with the RD communities. (36PM)

The ecosystem of RD services will inevitably be a combination of distributed and centralized resources, because of the sheer number of rare diseases and rare disease organisations, as well as legal and ethical constraints between countries and communities. At the same time, because of the low frequency in the population, combining data across patient registries, biobanks, and -omics databases is the single most important way of getting new insights towards new treatments.

One of the most recurrent issues when attempting to perform research across resources is the lack of standards or the poor adoption of existing standards by RD stakeholders. Rare disease standards concern different types of data including genomic and phenotypic characteristics, causative genetic variation status, quality criteria, analysis protocols, supporting evidence and follow-up indicators. These problems will be analysed in workshops including experts in semantic web, linked data technologies and rare-disease experts (see previous experiences and proposal in “Bring Your Own Data (BYOD) bootcamps”, in WP5). The initial experience with this methodology (see 61) is that a critical bottleneck is the identification of the most appropriate terms and identifiers to annotate data for cross- resource questions. Based on this experience we aim to address two major 'white spots' in the available infrastructure for Rare-diseases: (i) the current infrastructure of the rare disease platform: RD-Connect, does not contain backbone services for functional interlinking, (ii) a majority of RD sources are not equipped to provide data, metadata, and data updates using appropriate standard procedures. To address these needs we will work together with WP5, the rare-disease communities and the RD-Connect project to (i) deploy and test the services and guidelines for standardization 'at the source', (ii) provide standardized interfaces that

Rare-disease communities can work with from a central location, (iii) build capacity in the RD community by enabling them to work with these services themselves.

Partners: FR, ES, DK, NL.

Task 8.3: Training workshops targeting different user communities. (32PM)

In this task training workshops and courses will be delivered, in partnership with WP11 "EXCELERATE Training Programme". The training will be approached from two sides. First, in collaboration with the Train the Researcher task in WP11 we will train rare diseases' researchers in the use of relevant tools, standards and infrastructure produced by ELIXIR. Second, we will run "feedback workshops" in which those who are developing the methods will be exposed directly to problems faced by the rare disease community. These userthons will help to shape the ELIXIR portfolio.

The direct collaboration with WP11 Train the Researcher will ensure that researchers are trained to a high standard in state-of-the-art analysis techniques for rare disease data and that innovative training approaches developed in this task are applied elsewhere in ELIXIR.

Partners: UK, SI, NL.

8. Appendix 1: Documentation on adequate quality reporting standards for genomics datasets

8.1 Background

Rare disease (RD) research faces particular challenges because patient populations, clinical expertise, and research communities are small in number and highly fragmented both geographically and in terms of medical specialty. The scarcity of rare disease patients and their corresponding (gen)omic data has made data sharing one of the fundamental pillars to fasten and improve patient diagnostic and to reach IRDiRC 2017-2027 vision to enable all people living with a RD to receive an accurate diagnosis, care, and available therapy within one year of coming to medical attention.

Different project such as NeurOmics[15], EurenOmics[16], RD-Connect[17] and more recently Solve-RD[18] and EJP-RD[19], infrastructures such as BBMRI and ELIXIR and initiatives such

[15] https://www.neuromics.com/
[16] https://www.eurenomics.eu/
[17] https://rd-connect.eu/
[18] http://solve-rd.eu/
[19] http://www.ejprarediseases.org/

as GA4GH have been working towards this objective. Indeed, rare disease platforms such as RD-Connect GPAP[20] enable controlled data sharing of standardised phenotypic and genomic data. HPO[21], OMIM[22] and Orphanet[23] (ORDO) ontologies are used to collect phenotypic data and GATK best practices[24] and GA4GH[25] standards are followed to collect and process genomic data through a standardised pipeline (Laurie et al., 2016).

Collating genomic data from disparate centers across different countries has largely evidentiate to improve our understanding on rare diseases (Lochmüller et al., 2018[26]). However in order to fully benefit from this unprecedented access to genomic data, care must be given to determine the quality of these genomic datasets, especially when sequenced at different centres, under different protocols and using different technologies. In this sense, several metrics such as depth of coverage, base quality and mapping quality are already broadly used for NGS quality evaluation. However, due to the rapid development of the genome sequencing field, comprehensive quality management considerations are still scarce and although some efforts have been made, there is no current standards for genomic data quality comparison (Endrullat et al 2016[27] and Mahamdallie et al 2018[28]).

In this context, one of the specific objectives of EXCELERATE WP8 was to establish a framework for quality assessment of genomic data (Task 8.1.2) to enable rare disease researchers to easily compare genomic datasets, starting with whole exome sequencing data. In this deliverable, we have explored a rating system based on 5 different quality metrics. This rating system could be used as a starting point for continuing work in the context of WGS sequencing data and the 1M genomes declaration as federated systems across endorser countries will require to compare WES / WGS of different origin.

8.2 Report

8.2.1 Literature review

In order to define which Quality Control (QC) measures best reflect the quality of the sequencing on WES data, we first performed a literature review of whole exome sequencing (WES) in rare diseases. Understandably quality control is not the main focus of these papers, but it is notable for those that mention it that average read depth (Steven et al 2013 [29]; Balint et al 2015[30]), percentage of the target region covered by at least 20X

[20] https://platform.rd-connect.eu/
[21] https://hpo.jax.org/
[22] https://omim.org/
[23] https://www.orpha.net
[24] https://software.broadinstitute.org/gatk/
[25] https://www.ga4gh.org/
[26] doi: 10.1038/s41431-018-0115-5
[27] doi: 10.1016/j.atg.2016.06.001
[28] doi: 10.12688/wellcomeopenres.14307.1
[29] doi: 10.1016/j.ajhg.2013.01.016
[30] doi: 10.1002/mds.26355

reads (Antony et al., 2013[31]; Baynam et al 2015[32]; Onoufriadis et al 2013[33]) or measurements based on the CCDS (consensus coding sequence) exons (Srour et al., 2012[34]) are relevant. The evenness of coverage across the exons of the genes under investigation was also important for some studies (Futema et al., 2014[35]). Several studies used Picard to analyse and report on their sequencing (Cirak et al 2013[36]; Daoud et al 2015[37]). It is also noteworthy that the Deciphering Developmental Disorders study[38], concentrated the QC measures computation on regions covered by exome capture kit probes (Wright et al., 2015[39]).

In order to establish quality standard measurements, we have made use of the QC measures used in the literature, as well as measures that were found important in our sequencing experience at the CNAG-CRG[40]. To do so we have built on previous work on whole genome sampling in cancer (Whalley et al., 2017[41]) where defining a threshold and assigning a star for each QC measure a sample passed provided a very useful summary of the sequences quality.

8.2.2 Assembly of a limited set of non-redundant measures

Five QC measures, linked to different aspects of the genomic sequence quality have been selected:

1. Median coverage
2. Evenness of coverage
3. Median / Mean coverage
4. Percentage of chimeras
5. Mismatch rate

Median coverage, evenness of coverage and the median over mean coverage ratio, all reflect how well the exome has been sequenced. While making use of the paired read ends to calculate the percentage of chimeras and the mismatch rate in edits on read one and read two, gives some insight into the quality of the starting material and the library preparation for sequencing.

Unlike whole genome sequencing, where the QC measures can be based on the whole genome, the regions of the genome to calculate the QC measures need to be defined as multiple exome capture kits are available. Therefore, to keep a level of consistency, we

[31] doi: 10.1002/humu.22261
[32] doi: 10.1002/ajmg.a.37070
[33] doi:10.1016/j.ajhg.2012.11.002
[34] doi: 10.1016/j.ajhg.2012.02.011
[35] doi: 10.1136/jmedgenet-2014-102405
[36] doi: 10.1093/brain/aws312
[37] doi: 10.1016/j.ejmg.2015.08.001
[38] https://www.ddduk.org/
[39] doi:10.1016/S0140-6736(14)61705-0
[40] https://www.cnag.crg.eu/
[41] doi: 10.1101/140921

propose to calculate QC measures on the aligned sequences of the CCDS (consensus coding sequence) exons in the regions covered by the probes of the exome capture kit.

8.2.2.1 Median Coverage

Higher median coverage results in an increase of reads, which in turn increases the number and the sensitivity of single nucleotide variants called (Clark et al 2011[42]). The median is calculated on the set of the number of reads covering each position in the CCDS exons covered by the exome capture kit probes (with low quality and duplicate reads excluded so we do not inflate the number of reads).

8.2.2.2 Evenness of Coverage

As well as the depth of coverage, we need the exome to be evenly covered to call SNVs and other variants across the targeted region. To calculate the evenness of coverage, the number of bases which are covered by the median coverage or higher are added together divided by the number of bases multiplied by the median coverage.

8.2.2.3 Median Coverage over Mean Coverage ratio

The skewness of the coverage can be measured by calculating the ratio of the median coverage over the mean coverage. This value identifies sequencing quality independently of the median coverage and the evenness of coverage. An ideally sequenced sample would have a ratio of one, with the mean value the same as the median value, not skewed by very low or high coverage in certain regions.

8.2.2.4 Percentage of chimeric reads

Chimeric reads are paired reads that map outside of a maximum insert size or that have the two ends mapping to different chromosomes or the first and second read map to the same strand. The percentage of chimeric reads is a good indicator for the quality of the library constructed while preparing the sample for sequencing. Chimeric reads are likely to emerge as a result of an unspecific ligation process during library preparation.

8.2.2.5 Mismatch Rate

In paired end sequencing, bases which are different to the reference (edits) should roughly be shared between read one and read two. If there is an imbalance in the number of edits between the paired reads, this suggests damage in sequencing runs, possibly due to DNA degradation in collection and preparation. To quantify this, we calculate the ratio of the maximum of the number the edits in read 1 or read 2, divided by the number the edits in read 1 or read 2, which ever is the minimum.

8.2.3 Rating system

The measures described above can be used to construct a rating system. Each QC measure could grant a point if the sample is within a defined threshold, giving a final score from 0 to 5. This score could enable an easy comparison between different WES samples.

[42] doi: 10.1038/nbt.1975

8.3 Conclusion

The quality metrics explored for WES data assessment together with the establishment of a rating system to compare samples quality would provide a more reliable framework to the rare disease community for genomic data analysis of shared data. This approach could be used as a starting point to define global quality assessment standards for WES and WGS data in the context of the 1M genomes declaration and the MEGA project.