European Genomic Data Infrastructure

Deliverable D6.6

Report outlining the recommendations on data curation and ELSI compliance

Project Title Grant agreement no	**Genomic Data Infrastructure** Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP6: Data Management		
WP Leaders	Rob Hooft (21. HRI)		
Deliverable Lead Beneficiary	8. VIB		
Contractual delivery date	31/01/2024	**Actual delivery date**	28/02/2024
Delayed	Yes		
Partner(s) contributing to deliverable	VIB		
Authors	Dilza Campos (VIB)		
Contributors	N/A		
Acknowledgements	N/A		
Reviewers	Teresa D'Altri (CRG), Marco Morelli (UniSR)		

Log of changes

Date	Mvm	Who	Description
19/01/2024	0v1	Dilza Campos (VIB)	First draft of complete document to be reviewed by WP6 members

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

26/01/2024	0v2	Dilza Campos (VIB)	First draft modified by addressing WP6 members suggestions
06/02/2024	0v3	Teresa D'Altri (CRG), Marco Morelli (UniSR)	Revision of the second draft
07/02/2024	0v4	Dilza Campos (VIB)	Second draft modified by addressing reviewers suggestions
20/02/2024	0v5	Mercedes Rothschild Steiner (ELIXIR Hub)	Version ready for submission following Management Board review
28/02/2024	1v0	Mercedes Rothschild Steiner (ELIXIR Hub)	Final version to EC Portal

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

European Genomic Data Infrastructure

Table of contents

Contents

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

European Genomic Data Infrastructure

1. Executive Summary

It is a challenging task to work with genomic data sequenced over several years (legacy and future data). The aim of this deliverable is to gather recommendations and best practices for the process of curating data for ingestion into GDI nodes. Data curation is a broad term that includes managing data throughout its lifecycle, overlapping with other tasks and aspects of the GDI project. Since one important aspect of data curation is to ensure its quality—the availability and fitness for use and reuse-, this report will focus on data quality aspects and ELSI compliance during the submission process to the data repository. It also discusses some possibilities on how to assure data quality and elaborates on the concepts of data quality and data curation, with a strong focus on genomic data. This report then provides a list of resources for genomic data quality assessment.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

2. Contribution towards project outcomes

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

[Select 'Yes' (at least one) if the deliverable contributed to the key result, otherwise select 'No'. For more details of project outcomes, see here]

	Contributed
Outcome 1 Secure federated infrastructure and data governance needed to enable sustainable and secure cross border linkage of genomic data sets in compliance with the relevant and agreed legal, ethical, quality and interoperability requirements and standards based on the progress achieved by the 1+MG initiative.	Yes
Outcome 2 Platform performing distributed analysis of genetic/genomic data and any linked clinical/phenotypic information; it should be based on the principle of federated access to data sources, include a federated/multi party authorisation and authentication system, and enable application of appropriate secure multi-party and/or high-end computing, AI and simulation techniques and resources.	Yes
Outcome 3 Clear description of the roles and responsibilities related to personal data and privacy protection, for humans and computers, applicable during project lifetime and after its finalisation.	No
Outcome 4 Business model including an uptake strategy explaining the motivation, patient incentives and conditions for all stakeholders at the different levels (national, European, global) to support the GDI towards its sustainability, including data controllers, patients, citizens, data users, service providers	No

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

(e.g., IT and biotech companies), healthcare systems and public authorities at large.	
Outcome 5 Sustained coordination mechanism for the GDI and for the GoE multi-country project launched in the context of the 1+MG initiative.	No
Outcome 6 Communication strategy – to be designed and implemented at the European and national levels.	No
Outcome 7 Capacity building measures necessary to ensure the establishment, sustainable operation, and successful uptake of the infrastructure.	No
Outcome 8 Financial support to the relevant stakeholders to enable extension, upgrade, creation and/or physical connection of further data sources beyond the project consortium or to implement the communication strategy and for capacity-building.	No

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

3. Overall scope and introduction

Genomic data is considered by General Data Protection Regulation (GDPR)[1] as one of the 'special categories' of personal data that 'merit higher protection', meaning that it is subject to strict national regulatory frameworks. Also, there are differences amongst European Union member states in the concrete implementation of the GDPR into national regulations, complicating the cross-border sharing of sensitive data. In many countries, most of the genomic data is being generated in healthcare settings[2] and in medical centers lacking a research focus, these datasets are usually stored in siloed, inaccessible locations. The World Economic Forum estimates that 97% of all hospital data is never used beyond primary attention[3]. Other main sources of genomic data are large research projects and direct-to-consumer DNA tests. However, to be able to use this data and to generate knowledge and products to improve human health, the sharing of this siloed data is essential.

To facilitate cross-border sharing of genomic data in Europe the signatory countries of the 1+ Million Genomes (1+MG) initiative committed with the aim of ensuring that appropriate technical infrastructure is available across the EU, allowing for secure, federated access to genomic data. The 1+MG infrastructure plans to allow sharing of genomic data with harmonised data governance. The Beyond 1 Million Genomes (B1MG) project provided coordination and support to the 1+MG initiative in aspects such as ELSI and infrastructure requirements, as well as data quality and standards for interoperability among countries.

The scope of this deliverable is to gather the recommendations for the process of curating data for ingestion into GDI nodes. To properly carry out the curation work, data curators/stewards need to have appropriate expertise, skills and consolidated curation policies should be put in place. It is important that standard operating procedures are set to guide data curators/stewards on how to proceed in case data does not comply with the quality standards set by the node. These recommendations align with GDI WP4 activities related to standard operating procedures and builds on the previous work done by the B1MG project and the 1+MG Framework.

4. Methods

This deliverable is the result of a literature review on genomic data curation activities and data quality parameters for human genomic data. A review of the content in the 1+MG Framework was performed to identify available guidance. Discussions about data curation were carried out in work package 6 meetings to achieve a common understanding of the scope of the deliverable.

[1] https://gdpr-info.eu/
[2] https://doi.org/10.1038/s41431-021-00976-w
[3] https://www.weforum.org/agenda/2019/12/four-ways-data-is-improving-healthcare/

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

5.1 Data curation definition

Data curation is defined by the American Network of the National Library of Medicine[4] (NNLM) as "the ongoing processing and maintenance of data throughout its lifecycle to ensure long term accessibility, sharing, and preservation". Data curation in this context is composed of data management and digital preservation and involves processes such as adding metadata to make data more findable and understandable, ingesting data into a repository, and validating it. NNLM definition is broader than the definition of data curation used by Huan et al[5]: "Data curation is the process of managing data, including ensuring its quality—the availability and fitness for use and reuse". This report will focus on the process of ensuring data quality and ELSI compliance during the submission process to the data repository (e.g., a GDI node). A list of resources for guidance on genomic data quality and ELSI compliance is provided, as well as possibilities on how to assure data quality.

5.2 Data quality dimensions

There is no agreement upon one definition for data quality[6] as it is regarded as a multidimensional and hierarchical concept instead of a one-dimensional assessment. Data quality dimensions are measurable characteristics or attributes of data that are used to assess the overall reliability of the data. Although there are more than 30 data quality dimensions[7], the more frequent ones are data completeness, accuracy, consistency, timeliness, validity and uniqueness. The main objective of quality assessment is to ensure that high-quality data is data that represents its underlying real-world phenomena correctly.

However, to understand the context in which the concept of quality is used, it is essential to take into account the objective driving the utilisation of data. Wang and Strong[8] defined quality as "fitness for use," indicating the importance of determining data quality in the context of its use and highlighting it is a contextual and multidimensional concept. Huang, Jorgensen and Stivilia[9] showed that the role of the individual in the data curation process also influences the aspects considered important from the possible data quality dimensions. They grouped individuals as data collectors, end-users (researchers) or dual-users (individuals with data curation activities who also perform research on the curated data). In their division, data collectors and data custodians are knowledgeable about data collection, storage and maintenance processes. However, since data consumers may know more about the utility of the data, they might be more knowledgeable about data relevancy. In their survey, data curators gave more importance to the accessibility and accuracy

[4] https://www.nnlm.gov/guides/data-glossary/data-curation
[5] https://doi.org/10.1002/asi.21652
[6] https://doi.org/10.4301/S1807-1775202017003
[7] https://dama-nl.org/dimensions-of-data-quality-en/
[8] https://doi.org/10.1016/j.lisr.2014.08.003
[9] https://doi.org/10.1016/j.lisr.2014.08.003

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

dimensions of the data while dual-users and end-users considered the usefulness of data more important than accessibility.

5.3 Metadata and ELSI

Metadata helps users understand, manage, and use the actual data it describes. This additional layer of information enhances the effectiveness and efficiency of working with data. Metadata is the first input to the process of measuring and assessing data quality and includes the essential information necessary to comprehend assumptions about data, providing a starting point for defining expectations related to data quality[10]. It provides the necessary context and information to evaluate the reliability, usability, and trustworthiness of the underlying data. Metadata can describe many aspects of data and is important from the ELSI perspective. Metadata on data access properties (also known as rights metadata) is a form of administrative metadata that describes the type of data access allowed and what are the use conditions of a dataset. There are ontologies developed for data use rights, such as the Global Alliance for Genomics and Health (GA4GH) Data Use Ontology (DUO)[11].

DUO is a standardised, machine-readable vocabulary of data use terms that enables direct matching between data use conditions and intended research use. It is the accepted GA4GH standard for data use terms, with over 200,000 datasets annotated with it. Guaranteeing that the data use terms for a dataset match the research purposes of different researchers or stakeholders using the data is of paramount importance to comply with ELSI requirements for secondary use of genomic data.

DUO does not capture other important ELSI-related data use conditions, such as re-contacting policies or the legal basis on which the data was included. Metadata on changes in datasets derived from the exercise of data subjects rights: deletion, rectification, changed use conditions should also be present. The possibility of the data subject being part of vulnerable groups, such as minors, is also an important field of the ELSI metadata.

Another important ELSI aspect to observe during data curation procedure is related to incidental finding (IFs). Incidental Findings are findings with health relevance for 1+MG data subjects (or their families), revealed through the analysis of genomic and health related data made available through 1+MG infrastructure. 1+MG describes three policy options[12] for incidental findings reporting to data holders and it is important to have metadata available on the IF policy that applies to the dataset being shared for secondary use.

[10] https://doi.org/10.1016/C2011-0-07321-0
[11] https://doi.org/10.1016/j.xgen.2021.100028
[12] https://zenodo.org/records/8279737

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Data provenance[13] tells researchers the origin, changes to, and details supporting the confidence and validity of data. The concept of provenance guarantees that data creators are transparent about their work, and provides a chain of information where data can be tracked during the reuse process. Data provenance is a relevant ELSI aspect and there should be an agreement about what provenance metadata must be collected and stored. Provenance metadata allows tracking the origin, transformation, and manipulation of data throughout the entire analytical pipeline. It is thus important to register metadata information about sample collection, experimental procedures, and data preprocessing steps, such as genome assembly, software tools, parameters, and versions for genomic mapping and variant calling. Provenance metadata provides transparency and aids in the identification of ethical concerns or biases in the data or analysis. It is important to use controlled vocabularies and ontologies when those are available.

5.4 Data quality in genomics

Next-Generation Sequencing (NGS) may be performed on any living organism and any specimen that yields DNA and/or RNA (e.g., peripheral blood, saliva, fresh or frozen tissues, cultured cells, formalin-fixed paraffin-embedded tissues, prenatal specimens). Given the focus of the GDI project, we'll focus this chapter on human genomics, which relates to the complete set of DNA in an organism[14]. Sequencing performance may vary by sample type (e.g., saliva-derived DNA performs more poorly on whole genome sequencing compared with capture-based NGS due to contaminating bacterial DNA; Formalin-Fixed Paraffin-Embedded (FFPE) samples perform poorly on long-read sequencing methods because of DNA breaks during sample processing). Besides a variety of specimens, there are several types of sequencing techniques that can be used, depending on the objectives of the sequencing process. Clinical laboratories perform mostly short-read germline DNA sequencing. In research settings, NGS is used in many other contexts, depending on the research question in place. Some of the broader types of NGS use are (non-exhaustive list):

1. Whole genome sequencing
2. Targeted and exome sequencing
3. Transcriptomics
4. Chromatin sequencing
5. Metagenomics
6. Cell-free DNA sequencing

Apart from specimen type and the aim of the experiment, there are also technological aspects that influence data quality (short-read versus long-read sequencing, arrays, DNA capture strategies). In oncology, samples can be derived from normal or tumoral tissues and this will also reflect on parameters like variant allele fraction because of the distribution of tumoral versus non-tumoral

[13] https://www.nnlm.gov/guides/data-glossary/data-provenance#:~:text=The%20term%20%E2%80%9Cdata%20provenance%E2%80%9D%2C,to%20where%20it%20is%20presently .

[14] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/genome

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

tissue sampled. Up to now there is no quality threshold set for all the possible uses of genomic data and the accurate annotation of all this experimental setup in the metadata model is important to data quality dimensions.

There are ontologies aimed to describe the experimental setup in which the genomic data was acquired, including sequencing machines, study set-up design, biological sample types[15] [16]. It is of high relevance to align which ontologies or controlled vocabularies are used as much as possible to facilitate genomic data exchange and interoperability. This approach ensures that the defined concepts are unambiguous and that data users have the same interpretation of data as the data providers intended.

There is a study group in GA4GH working on developing a standardised way of organising the properties and ontologies for genomic experiments and to derive schemas and specifications for genomic experimental metadata.

When it comes to NGS for clinical applications, there are many resources available on how to perform data quality assessment, since the aim of the test is to provide a clinical report or diagnosis to a patient affected by a certain disease. The vast majority of the literature available deals with short-read sequencing of germline DNA derived from blood or saliva. Since most of the genomic data is generated in healthcare settings, it is expected that the majority of data that comes from clinical laboratories comply with at least some of the guidelines described below.

The American College of Medical Genetics published the first guidance for laboratories performing next-generation sequencing (NGS) testing in 2013[17] and the last iteration of its recommendations regarding short-reads NGS data quality for germline variants detection was available in 2021[18]. That set of recommendations does not rely on hard thresholds for data quality, but focuses on the processes to streamline the clinical validation of assays (best practices regarding how to assess assay sensitivity and specificity) using NGS technology. The same holds true for the College of American Pathologists from the United States in its framework for clinical laboratories performing NGS.[19] It provides detailed guidance for laboratories on how to establish the types of specimens and the minimum amount and quality of input DNA required for NGS assays. But the laboratory is responsible for determining the acceptable parameters of sample and data quality.

The European Society of Human Genetics[20] indicate what parameters on NGS data quality should be evaluated and states that thresholds should be in place for data filtering (e.g., average depth of coverage, evenness of coverage, percent genome above minimum mapping quality, and/or callability), but does not set what those thresholds should be.

[15] https://www.ebi.ac.uk/ols4/ontologies/efo
[16] https://obi-ontology.org/
[17] https://doi.org/10.1038/gim.2013.92
[18] https://www.nature.com/articles/s41436-021-01139-4
[19] https://doi.org/10.1016/j.jmoldx.2018.11.004
[20] https://doi.org/10.1038/s41431-022-01113-x

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

The WP3 from B1MG project surveyed several European laboratories performing NGS on cancer and germline specimens with the aim of establishing a common set of metrics and deriving quality standards that could be applied for quality control[21]. They have recommended thresholds ranges for DNA library preparation and have identified commonly used quality metrics for both germline and cancer NGS (quality scores across all bases, sequence length distribution, average quality per read, percent of duplicated sequences, and percent of each base in the sequence). The EU-funded project EASI-Genomics[22], finished in 2023, aimed to benchmark the metrics identified to establish quality thresholds to provide guidelines for data quality. Their results involve a near exhaustive elaboration of germline WGS data quality thresholds that is about to be finished and later in 2024 the project aims to release the tumour/normal pair benchmarking analysis.

Using data from 2959 normal-tumour genome pairs from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project, researchers[23] were able to define quality metrics covering five important features to assess the quality of cancer genome sequences: mean coverage, evenness of coverage, somatic mutation calling coverage, paired reads mapping to different chromosomes and the ratio of difference in edits between paired reads. These metrics allowed them to create a five-star rating system for cancer genomic data quality. They have made their rating system available as a docker container[24] and argue that their method could be adapted for similar projects that look to use whole-genome sequences from a variety of sources.

An important dimension of data quality is completeness. It measures the extent to which data is comprehensive and lacks gaps or missing elements. To enable measurements of how complete a dataset is for genomic use, one has to compare it against a predefined set of attributes from an agreed upon data model. The 1+MG use cases and B1MG WP3 worked together to define minimal dataset models for rare diseases, cancer, infectious diseases and population genomics (focusing on the Genomes of Europe project). The minimal dataset model of the use cases will be compared to arrive at a set of common core features considered to be the minimal common elements across fields.

5.4.1 How large genomic repositories assess NGS data quality

There are successful established initiatives that enable access to large genomic databases and having a look on how they perform data quality assessment can provide insights on the best approaches to deal with this issue during data curation procedures.

The European Genome-Phenome Archive (EGA) performs quality assessment[25] on deposited data and makes them available to the scientists before the decision of requesting access to the data.

[21] https://doi.org/10.5281/zenodo.4889391
[22] https://www.easi-genomics.eu/home
[23] https://doi.org/10.1038/s41467-020-18688-y
[24] https://dockstore.org/containers/quay.io/jwerner_dkfz/pancanqc:1.2.2?tab=info
[25] https://web2.ega-archive.org/about/quality-control-reports

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

EGA stands that quality assessment and reporting can increase the reusability of data. Raw sequencing files (FASTQ) have parameters like per base sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, sequence duplication levels reported. Alignment files, like SAM/BAM/CRAM have base coverage distribution, base quality, % of mapped reads, % of both mates mapped, singletons, duplicates, among other metrics reported. VCF files have site frequency distribution, Ts/Tv, base changes, indel distribution reported. EGA does not filter or remove data based on these quality parameters.

The UKBiobank is the biggest whole genome dataset in the world and contains whole genome sequencing data on 500,000 participants[26] with clinical data and secondary health data. It provides controlled access to CRAM files and to unfiltered vcf files. There are quality measures available for all samples, such as coverage, yield, proportion of mapped read pairs, and additional QC data.

Genomics England Research Environment contains data from the British 100,000 Genomes Project that sequenced the genomes of 85,000 cancer and rare disease participants along with their continuously updated medical histories. The samples that fail the quality thresholds established by the project are removed from the cohorts and are not available on the datasets.

All of US is a genomic program from the United States launched in 2018 with the aim of sequencing 1 million individuals. The latest release has over 245,000 short-read sequencing WGS. Data quality parameters evaluated are coverage and biological sample concordance parameters (data sanity checks), such as fingerprint concordance, sex concordance, cross-individual contamination rate. Data that do not pass the quality/sanity checks is not included in data releases[27].

Table 1 - List of major genomic initiatives and its approaches to data quality

Repository	Data Quality Measurements	Data filtering strategy
EGA	Per base sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, sequence duplication levels reported. % of mapped reads, % of both mates mapped, singletons, duplicates, site frequency distribution, Ts/Tv, base changes, indel	Data is not removed if QC fails

[26] https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/genetic-data

[27] https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2022/06/All%20Of%20Us%20Q2%202022%20Release%20Genomic%20Quality%20Report.pdf

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

	distribution reported, among others	
UKBiobank	Coverage, yield, proportion of mapped read pairs, and supplementary files containing QC data	Data is not removed if QC fails
Genomics England	Rare diseases: • 95% of the autosomal genome covered at ≥15x calculated from reads with mapping quality >10 and >85x10^9 bases with Q≥30, after removing duplicate reads and overlapping bases after adaptor and quality trimming. • Germline cross-sample contamination performed using VerifyBamID. Samples with >3% contamination are considered as failing. Cancer: • Tumour Cross-Contamination less than 5% •Germline Cross-Contamination less than 3% • Median Fragment Size greater than 279bp • Excess of Chimeric Reads mean of 0.3% • Percentage of Somatic Mapped Reads mean of 93.4% • Percentage AT Dropout mean of 3.1%	Data is removed if QC fails
AllofUS	Fingerprint concordance, sex	Data is removed if QC fails

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

	concordance, cross-individual contamination rate and coverage	

5.4.2 Quality filtering versus annotation

Large scale genomic projects have chosen to either remove data based on sequencing quality or to allow researchers to access the "poor quality" data, providing many annotations regarding sequencing quality. Having all the data available for secondary use and having an annotation system to describe data quality (such as the star rating system described by Whalley[28] previously) is more aligned with the fit-for-use concept of data quality, since the end-user can decide on the fitness of data for her research purposes. However, the downside of this approach is that even when researchers consider that a certain data's quality is too low to render it useful, the repository has anyway to deal with the costs of maintaining that data.

If the repository filters out data based on hard thresholds, the end-user is ensured that virtually all data available meet the quality requirements set by the repository. However, these thresholds might be difficult to achieve and there probably should be some trade-offs between being severe enough to remove poor quality while not discriminating against samples with not so abundant information on quality standards (e.g., new sequencing technologies). It might also be difficult to set those thresholds for samples from specimens hard to sequence, such as FFPE.

If the repository filters out data based on quality thresholds, the end-user is ensured that all data available meet certain quality requirements. However, these thresholds have to be chosen wisely by technical experts, to avoid leaving useful data outside. Technical reasons for lower quality of certain types of difficult samples have also to be taken into consideration. There are initiatives aiming to help both types of approaches. The EU financed QUANTUM[29] project aims to develop a quality and utility label for datasets that could be adopted in the HealthData@EU, including genomics. The project starts in 2024. Researchers from the Pan-Cancer Analysis of Whole Genomes developed a 5 star rating system for cancer genomic data that could be incorporated into the metadata model. Regarding thresholds for data filtration, the EU financed project EASI-Genomics[30] benchmarked genomic data quality metrics to establish thresholds for data quality; these thresholds could be used for data filtering in repositories.

[28] https://doi.org/10.1038/s41467-020-18688-y

[29] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/org-details/920094665/project/101137057/program/43108390/details

[30] https://www.easi-genomics.eu/home

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

5.4.3 Resources for NGS data quality assessment

There are many resources available to provide guidance on data quality dimensions for data curation procedures. Table 2 is a non-exhaustive list of resources available to guide data quality policies to be implemented by the nodes.

Table 2 - Resources available for data curation

Resources for Data Curation	Description
https://framework.onemilliongenomes.eu/	Series of resources based on the output of the 1+MG and B1MG projects that provide guidance on ELSI, data quality, data standards, and technical infrastructure standards and APIs.
https://www.nature.com/articles/s41467-020-18688-y	Quality rating system for cancer genomic data
https://www.ga4gh.org/product/experiments-metadata-standard/	Metadata framework for experimental annotation of NGS Data
https://www.ga4gh.org/	Global organisation developing standards, policy frameworks, and tools for the sharing of genomic and other related health data.
https://fairgenomes.org/	Standardised semantic metadata scheme to describe various aspects (patients and their clinical data, samples, assays, analysis procedures, consent) of the genomics data
https://zenodo.org/records/5706412	MINSEQE describes the Minimum Information about a high-throughput nucleotide SEQuencing Experiment that is needed to enable the unambiguous interpretation and facilitate reproduction of the results of the experiment.

6. Recommendations on data curation and ELSI compliance

Regardless of the approach chosen (either filtering out "low-quality" data or labelling the data quality parameters), it is important that each node aiming to have a data repository develop or follow agreed upon guidelines (possibly derived from decisions of 1+MG) and standard operating

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

procedures on how to deal with data quality dimensions. It is important to have those guidelines and procedures clear and well described for data submitters. Some aspects to be included in the data curation procedures are:

1. Have data curators/stewards and also researchers involved in the processes of developing data standards
2. Define a set of well-established parameters to be evaluated for data quality filtering/labelling
3. Have documented standardised procedures to check for data discrepancies (e.g., biological sex checking versus sex on metadata description) and other data quality parameters
4. Have documented standardised procedures to contact data submitters for addressing data quality issues before data is available for secondary use
5. Have documented standardised procedures to address data quality issues after data has been made available for secondary use (e.g., procedure to re-contact data submitters in case of discrepancies found by researchers re-using the data)
6. Enforce the use of standards and common ontologies for both sequencing and phenotypic and clinical data
7. Enforce the use of the specified metadata model and adherence to recommended minimum information standards

This document describes the current status of data curation and ELSI compliance discussions on the GDI project with a focus on quality aspects. It points to resources from the 1+MG Framework and other tools for guidance.

7. Discussion

Despite many papers and resources available in the literature regarding some features of genomic data quality and curation, there is little empirical evidence regarding how data quality issues are perceived by scientists. Wang and Strong[31] observed that end-users of genomic data tend to give more weight to quality aspects of data that are different from the ones data curators are most concerned with. Involving researchers in the decisions about the data curation guidelines seems to be a reasonable approach to avoid storing data that won't be used. There are no available statistics from data repositories regarding data requests per quality parameters, however this type of analysis would also be influenced by factors such as the impact factor of the original publication, area of research, sequencing technology, and so on.

The fitness-for-purpose definition of data quality is important and captures the principle of quality, however this concept is abstract. Thus, it is a challenge to measure data quality using only this holistic construct or definition. It needs to be broken into measurable characteristics, known as data quality dimensions, for data quality assessment to be actionable.

[31] https://doi.org/10.1016/j.lisr.2014.08.003

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

This report did not elaborate on data quality dimensions for healthcare data, however this subject was explored in depth by the B1MG WP3 in their report entitled "Phenotypical and clinical metadata framework"[32] and the individuals interested in this aspect of data quality can find many resources there.

8. Conclusions, Impact and Next steps

Since working with genomic data sequenced over several years (legacy and future data), is challenging, proper data curation procedures are of central relevance to ensure data present in the database accurately reflect real-world phenomena. Data curation is the process of managing data, including ensuring its quality. This report provides arguments both in favour and against filtering data out based on data quality or labelling data quality in the metadata fields. The reader can also find recommendations on how to establish procedures to perform data curation and there is a list of resources for the evaluation of several quality dimensions. Discussion around data curation procedures will not be restricted to this report, but will continue during the project. Next discussions will involve how to help nodes decide if data quality should be used for filtering data or for labelling data and it is expected that this deliverable can help with the decision-making process.

[32] https://doi.org/10.5281/zenodo.10058688

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.