Contract no. 101057548

EPIVINF

Epigenetic regulation of host factors in viral infections

D1.12 Data management plan - update

ACTION: Research & Innovation Action (RIA)

CALL: HORIZON-HLTH-2021-DISEASE-04

TOPIC: HORIZON-HLTH-2021-DISEASE-04-07

Due Date of Deliverable	**31 Aug 2024**	**Completion Date of Deliverable**	**05 Aug 2024**
Deliverable leading partner	IRSICAIXA	**Author**	IRSICAIXA
WP Nº	WP1	**WP Title**	Harmonization of clinical data and cohort studies

Project starting date	01/09/2022	**Project Duration**	60 months

Dissemination Level	**PU**	**Public**	✓
	SEN	**Sensitive**	

This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement Nº 101057548

TABLE OF CONTENTS

1. INTRODUCTION / EXECUTIVE SUMMARY

This report is the second deliverable of Task 1.12 “Data Management” and describes the updated Data Management Plan (DMP) for the EPIVINF project, funded by the EU’s Horizon Europe Programme under Grant Agreement number 101057548. The purpose of the DMP is to provide an overview of all datasets collected and generated by the project, and to define the EPIVINF consortium’s data management policy regarding these datasets.

The EPIVINF DMP follows the structure of the Horizon Europe DMP template. It reflects the status of the data that is collected, processed, or generated, as well as the methodologies and standards followed. It specifies whether and how data will be shared and/or made open, and how it will be preserved.

This DMP outlines the general policy and approach to data management in EPIVINF. This includes for example topics like data and metadata collection, publication and deposition of open data.

2. DATA SUMMARY

In the EPIVINF project, we collect and generate a broad set of clinical and molecular data. Both types of data are to be kept separately from each other and only combined for data analysis. For the general aim of the project, the understanding of how acute viral infections impact host factors involved immune control and neurological health, the most important data types from a molecular perspective are (1) genome wide methylation sequencing data and (2) single cell and bulk transcriptomes. Methylation and transcriptomic data are both generated from different peripheral mononuclear blood cells and from samples collected across a large number of different clinical cohorts, including individuals with and without HIV or SARS-CoV-2 infection and with variable disease courses related to these infections. It includes individuals followed longitudinally and participants in vaccine trials, forth both, HIV and SARS-CoV-2. For the genome wide methylation analysis, we are sequencing 500 million reads (paired end) per sample to allow for sufficient coverage of the CpG islands. For the single cell transcriptomes, we likewise sequence around 500 million reads per sample. This deep coverage is required to detect ~5000 genes in ~5,000 cells per samples. All genomic raw data (we assume 50 TB raw data over the entire project) is stored at the Institut Germans Trias i Pujol (IGTP) and in an external hard drive. Then genomic and clinical data will be integrated at the Centre for Bioinformatics in Saarbrücken, Germany where backups of the data will be stored. The raw data in fastq format are also uploaded to the European Genome-phenome Archive (EGA). Typically, at the time of publishing, including at the pre-print stage, we will also make the EGA data accessible for the public.

Next Generation Sequencing (NGS) antibody repertoire data generated in the project consist of Illumina MiSeq raw sequence read files. Two files are produced per sample/library sequenced, a Read 1 and a Read 2 file, forward and reverse direction reads of each library that are further processed to produce a full-length library encompassing the full VDJ reads. The Read1 and Read2 files are in a compressed fastq format, the size of which is relative to the library size; we generally aim for a size of approximately 1 million sequences per library. A library of 1 million sequences will have compressed Read1 and Read2 fastq files of approximately 250 Mb in size. These data are stored at SciLifeLab on the high security Bianca server at Uppmax, https://www.uppmax.uu.se/ to adhere to GDPR.

T- and B-cell repertoire sequencing data are generated using OS-T and OS-B Omniscope proprietary sequencing technology where the format of raw data is FASTQ and contigs. All data are stored at AWS in encrypted S3 buckets with restricted access.

The clinical data are managed with REDCap. Clinical database storage is centralise at the Institut Germans Trias i Pujol (IGTP), where both, the coordinating Institution (IrsiCaixa) and one of the main clinical sites (Infectious disease unit from Hospital Germans Trias i Pujol) of the project are located. However, we are evaluating all the legal requirements to transfer pseudonymised data from external centres to IGTP. For data privacy reasons access to that database is strictly regulated, even though only pseudonymized data are stored there. For analyses of the molecular data, computational biologist will get access to the clinical data as export from REDCap in CSV formatted files or through secure identified application programming interface access to just read the data without possibilities to modify them.

The clinical data (under pseudonymized codification) that are used for the analysis will be attached to the molecular data in the European Genome-phenome Archive (EGA), to allow for analysis and interpretation of the molecular data by third parties.

In the project we are reusing data for two different reasons. First, we have accessed publicly available and in-house molecular data sets to support the analysis of the newly generated data, but also based on existing sequencing data and on shallow sequencing of the newly collected samples, we can compute duplication rates and other metrics. Based on these metrics we can compute the optimal dilution of samples to sequencing lanes, making the coverage of samples more uniformly distributed. These readily available annotated data support the training of statistical models to improve the annotation of the newly sequenced data. The second source of available data that are incorporated in the project is the peer reviewed literature and annotated databases. Using artificial intelligence, we have implemented models that support the identification of causal cascades from the secondary analysis of the newly generated data. Altogether, we plan to make use of 200 TB readily existing data.

Of note we think that the data generated in this project are of use for a broad research community and therefore deserves access at multiple scales.

1. As we re-use fastq raw data from other projects, we are convinced that computational experts will benefit from access to the raw data generated in EPIVINF as well. These will be made available via EGA as described.

2. With the fastq data in EGA, we always make available the count matrices, i.e. the data after primary analysis and quality control. Typically, these data are already more than one order of magnitude below the raw data and typically accessed by computational experts and advanced life scientist with some background in Phyton or R programming.

3. The key findings are made available as web services or databases. This allows medical researchers and life science researchers to have a look at the data with minimal effort, e.g. by typing in a gene name on a straightforward web site. The respective repositories are published with the original manuscripts and in case they find recognition in the community, as separate web services.

In addition to the newly generated molecular and clinical data, we also consider software that we generate as one type of data. We will make source code accessible via GitHub repositories and might publish models as web services.

Additionally to the data mentioned above, there will be other types of data generated in validation studies which are listed following:

Proteomics and ELISA assays: Plasma and CSF samples are used in high-throughput proteomics panels to analyse up to 380 analytes in an exploratory manner to identify potential epigenetically regulated factors. Raw data are received and stored in Excel files (.xslx, .csv, < 500Kb). ELISA and SIMOA assays

are conducted to further validate potential signatures (.csv, < 100Kb/file). Data are collected as either relative or absolute values of the levels of specific marker in the respective sample fluid.

RT-PCRs: RT-PCR is conducted in isolated bulk populations and total PBMCs and cellular fraction of CSF to validate the gene expression of specific candidates and virus quantification. The data resulted from real time PCR instruments are stored in Excel files (.xlsx, .csv, <100Kb/file) and the amplification plots as jpg files.

Flow cytometry: Flow cytometry data are generated to describe specific phenotypes and functional characteristics of different cell populations in samples drawn from different compartments (blood, CSF). These include percentages of cells expressing specific (functional) markers and intensity of the flow cytometric signal. Analyses are generally run with individual samples of 0.2 to 1-x10^6 cells in blood specimen and fewer (approx. 10^4-10^5 cells) in the CSF fractions. Raw data files in flow cytometry instrument are saved as (.fcs) format. Around 300MB per patient is being stored. The intermedium analyses, considering gating strategy, are processed in FlowJo software and stored as .wsp. Finally, the definitive output with the corresponding gating strategy is saved as well as .xslx (< 100Kb/file).

Elispot assays: Elispot assays are used to enumerate the number of antigen specific cells in a sample. In general, 10^5 cells are being exposed to a specific antigen and the number of responding cells is extrapolated to their frequency in 1x10^6 PBMC. Data are retrieved from the ELISpot counter in .png files and data are transferred to Excel files for integration in overall analyses (.xlsx, < 100Kb/file) and .jpg for plate image capturing.

Humoral antiviral activity / ADCC: Humoral immunity is measured by Elisa-based assays or by functional tests using flow cytometry (see above for both platform). Specific titter of antibodies are being collected from neutralization or binding assays, with the data being directly converted from the Elisa reader raw data file into excel and/or csv files (< 100Kb/file).

Data generated are associated with a README file in order to collect standardised metadata for all kinds of data.

3. FAIR DATA

3.1 Making data interoperable

Molecular data (transcriptomics, epigenomics and single cell sequencing) will be mainly stored in the European Genome-phenome Archive (EGA), together with the metadata (age, sex, diagnosis and similar clinical variables). EGA has become the most used repository for sensitive human genomic data associated in Europe. The stored datasets will be easily identifiable thanks to the use of Persistent Identifiers such as Digital Object Identifiers (DOI) and by adding a significant set of key words to the data sets. Nevertheless, for some specific data such as the NGS antibody repertoire data, these will be published to Bianca server at Uppmax where a digital object identifier (DOI) will be also assigned.

With REDCap, we are using a metadata-driven software toolset to collect, store and monitor the clinical parameters. The export of REDCap will be attached to all submissions to public repositories and likewise a rich set of keywords, including controlled vocabularies like MeSH, will be added to ensure data are findable.

For other data types like Flow Cytometry, other discipline and trusted repositories will be considered (e.g. FlowRepository, ImmPort). Also, generalist repositories like CORA.RDR (Dataverse) or Zenodo, compliant with FAIR principles might be used. Such information will be completed in future versions of this DMP.

3.2 Making data accessible

Most of the molecular data along with metadata will be deposited in EGA. The access to the data will be controlled by the Data Access Committee, composed by the researchers involved in the collection and analysis of the data, who will be responsible for approving access to the datasets. Access to each dataset will be covered by a Data Access Agreement, which defines the terms and conditions to use it. The NGS antibody repertoire data will be published to Bianca server at Uppmax in which a Data Access Agreement will be also needed

The other datasets generated, will also be made publicly available in other repositories, as long as there are no conflicts with intellectual property rights or sensitive data protection. All selected repositories will provide unique and persistent identifiers for each of the submitted datasets and provide a mechanism to track if the data are modified (e.g., the day of last modification of the data or a versioned identifier).

T- and B-cell repertoire sequencing data generated using OS-T and OS-B Omniscope proprietary sequencing technology will be made available to project partners via Omniscope web-based Immuneportal and can be downloaded either over the web (HTTPS) or over SFTP. Due to intellectual property rights, these data will be not made openly available, though contigs could be shared with the scientific community through a data sharing agreement.

We will make most of the data accessible at the earliest convenience, at the very latest when we make a description of the data available as part of any preprint in which the data are used. We will ensure that filing of intellectual property will not delay access to the data by more than 12 weeks.

We plan to make our work accessible at least under Attribution-NonCommercial-ShareAlike. Depending on the data set and study, and also the place where the data are published, even less restrictive licenses might apply. All software and models that we develop will be either published as web services or made available via GitHub, unless restrictions by third party software that is included (e.g. third-party libraries) prohibit it.

With the submissions and entries in the online repositories, we also will ensure that links to web services and special data sets are accessible. Site accesses and downloads of the data sets are for us a key performance indicator that is monitored monthly. In case we observe low usage rates we further try to promote the data sets actively in social media posts, on conferences, and publications and review our access modalities to eliminate any potential bottlenecks for broader access.

3.3 Making data interoperable

By uploading data in repositories, we ensure technical and semantic interoperability as well as the use of open and community accepted formats. For molecular data we will upload it to EGA, the fastq format is the de-facto standard of the raw data. It contains the reads from the sequencer along with sequence quality information. Because of the size of data sets we do not plan to make intermediate files directly accessible (i.e. BAM files). However, the results of the primary analysis after QC will be uploaded to the respective repositories (e.g. count matrices) and will be made available in standard formats (e.g. as Seurat objects). For the other datasets, associated with validation experiments, we will use as well open and community accepted formats, such as .fcs files for flow cytometry experiments. In future versions of the DMP, we will further refine the plans for these datasets.

Finally, each of the published datasets will be linked to the associated publications and related datasets. When possible, the use of controlled vocabularies like MeSH will be used.

3.4 Increase data re-use

With each data set uploaded to a repository, we provide very detailed descriptions of the methodology used to generate the data. As for the source code that we will deposit in GitHub, we commonly follow the GitLab documentation style guide (https://docs.gitlab.com/ee/development/documentation/styleguide/).

By uploading all data to persistent repositories (EGA), we ensure that all data remain accessible after the end of the project. Additionally, all the code will be made available on GitHub to ensure the reproducibility and transparency of the analyses/computational analysis and facilitate the reproducibility of the analyses. To enable highest quality of data, we implemented very stringent quality filters for all data types. These quality filters have been defined from the state-of-the art literature in the field, own data sets and publicly available data sets. For example, genome wide methylation sequencing data have to be covered by at least 500 million reads, and only CpG sites with at least 5 unique starting point reads (i.e. reads after removal of duplicates) covering the site will be included in the analysis. For the single cell data, at least 500 million reads per sample are required, only cells with at least 75,000 reads covering the cell, and at least 2,000 genes covered within the cell are included. Moreover, MT contamination filters apply. However, we also provide all raw data with the quality information as fastq files, allowing expert users that want to re-use the data to apply own quality control filters.

The other datasets will be uploaded to the repositories in open formats or community accepted formats with all the necessary disciplinary metadata to understand how they were generated and facilitate its reuse (e.g. using readme files with the units of measurements, methods, variable definition, etc.). We will cover this part in more detail in the next versions of the DMP.

4. OTHER RESEARCH OUTPUTS

In the EPIVINF project, we will generate software either as stand-alone or as web-services. The source code of on-premise software will be made available via GitHub. Further, we use workflow management systems, such as NextFlow or SnakeMake. In the same manner as we provide source code for on-premise software we will also distribute workflows generated in the project via GitHub – unless licenses of third party software embedded prohibit this. We apply the same FAIR principles to software that also applies to research data. Software will be licensed with permissive (e.g. MIT) or copyleft licenses (e.g. GNU GPL).

5. ALLOCATION OF RESOURCES

We currently plan with direct storage cost of 30.000 Euro on RAID6 storage systems at the Center for Bioinformatics in Saarbrücken over the grant period to hold all data accessible for the consortium. One year after the end of the grant period we will move the data to a less cost intensive network attached storage device to comply with a 10-year minimum storage obligation of the raw data. All data will however be also uploaded to other persistent repositories as described above. The cost will be covered from the overhead of the funded project.

6. DATA SECURITY

The data are stored on RAID6 server storages that are regularly backed up. We will ensure that at each time at least two copies of the data are stored at two different sites. The data storage server room has a very restricted physical access policy. In case of transfer of data via hard drives or over the internet, data will be encrypted.

Data generated by Karolinska Institute (NGS antibody repertoire) are stored securely on in-house firewall-protected servers, and on Bianca server at Uppmax (Uppsala University, https://www.uu.se/en/centre/uppmax/resources/clusters/bianca) to adhere to GDPR. The internal server is backed-up nightly to the central KI storage system.

Data generated by Omniscope (T and B-cell repertoire sequencing) are stored in the cloud using AWS. Raw and processed data are stored in encrypted S3 buckets with highly restricted access for an indefinite period. After 6 months, data is put in long-term storage, which requires 48 hours' notice for retrieval. During processing data is temporarily copied on ephemeral encrypted drives whose contents

are overwritten multiple times with random data on completion to make recovery impossible by a 3rd party.

Furthermore, the different partners have in-house security measures to ensure data privacy as well as regular back-ups and a controlled access to the data they receive or generate. By controlling who can access to a machine, and what permissions each user has on each folder with the data. Physical access policy is controlled, and people are trained in security practices while working in the computer.

Long-time storage is secured on network attached storage devices for at least 10 years and in external repositories, most importantly EGA.

7. ETHICS

All the data processing will be compliant with the GDPR (Regulation (EU) 2016/679) and different safeguards will be applied. On the one hand, clinical data are managed using REDCap (Research Electronic Data Capture), which is hosted in institutional servers with the appropriate security measures to be compliant with GDPR and national laws of data protection. This will guarantee all the data are pseudonymised and that a technical separation is in place (personnel involved in the data analysis have no means to reverse the process). Also, according to the data minimisation principle, only the data necessary for the project are collected. Each of the institutions participating in research activities are responsible to guarantee personal data protection. IRB approval has been obtained to conduct the studies mentioned in this project. Individuals participating in this project come from various pre-existing cohorts for which they signed informed consents, allowing their samples to be used in future related studies. They have also been informed of the possibility to withdraw their consent at any time and were informed that all the data collected prior to leaving the study can be used in research projects, but no new data will be collected.

All external services (e.g. sequencing services) will act as data processors and the controller of the dataset will be responsible to make sure to revise the material and data transfer agreements to be compliant with the appropriate national and international legislation.

8. OTHER ISSUES

The guidelines from einaDMP have been used in the creation of this DMP (https://dmp.csuc.cat/). The data management service and the legal unit of IrsiCaixa (the coordinating institution in EPIVINF) will also participate in the revision of the document and in future updates.

Omniscope will retain the ownership of FASTQ and contigs and will keep them for internal research and commercial purposes.