Funded by the European Union

UNDINE	
Project number:	101057100
Project name:	The human genetic and immunological determinants of the clinical manifestations of SARS-CoV-2 infection: Towards personalised medicine
Topic:	HORIZON-HLTH-2021-DISEASE-04-07
Type of action:	HORIZON Research and Innovation Actions (RIA)
Starting date of action:	1 June 2022
Project duration:	48 months
Project end date:	31 May 2026
Deliverable number:	D8.5
Deliverable title:	Updated Data Management Plan
Document version:	Ver1.0
WP number:	WP8
Lead beneficiary:	AU
Main author(s):	Trine Mogensen (AU), André Walter (AU)
Internal reviewers:	Lluis Quintana-Murci (IP), Laurent Abel (INSERM)
Nature of deliverable:	R
Dissemination level:	PU
Delivery date from Annex 1:	M24
Actual delivery date:	M24

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HADEA. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

The purpose of this DMP is to describe the data management life cycle for all data that will be generated and processed by the UNDINE project. It defines how the entirety of the collected research data will be handled during the project time and beyond. Deliverable 8.3 is designed to be a dynamic document, continuously updated throughout the project time, at least every six months. All the data providing beneficiaries and associated partners will be active in the collection, storage and archiving of data. This will be done in full alignment with the FAIR principles, the Horizon Europe open access strategy and the EU Public Health Emergency strategy for data sharing. Project administration related documents like deliverables, reports or meeting minutes are out of the scope of this document and will be described separately.

Abbreviations

auto-Abs	Autoantibodies
CC-BY-NC	Creative Commons by NonCommercial (license)
COVID-19	Coronavirus disease
D	Deliverable
DDI	Data documentation initiative
EC	European Commission
IAV	Influenza A virus
IEI	Inborn errors of immunity
IFN	Interferon
iPSCs	Induced pluripotent stem cells
MIS-C/A	Multisystem inflammatory syndrome (in *C*hildren or *A*dults)
PBMCs	Peripheral blood mononuclear cells
POC	Point of care
(r)eQTLs	Expression quantitative trait loci or response quantitative trait loci
SARS-CoV-2	Severe acute respiratory syndrome coronavirus
scRNA-seq	Single Cell RNA sequencing
WES	Whole exome sequencing
WGS	Whole genome sequencing
WP	Work package
WT	Work task

Contents

List of Figures

1 Data summary

1.1 Objectives of UNDINE and the data required

Following the ambitions and scope of the Horizon Europe call “HORIZON-HLTH-2021-DISEASE-04-07: Personalised medicine and infectious diseases: understanding the individual host response to viruses (e.g., SARS-CoV-2)”, the overall objective of the UNDINE project is to gain knowledge on human genetics and immune responses underlying the different disease manifestations of SARS-CoV-2 infection, aiming at a swift translation into the. New ground-breaking discoveries are intended to be made, expanding the genetic knowledge and deepening the immunological insight to both innate and adaptive immunity. UNDINE will further explore new options for diagnostic tests and personalized medicine approaches to account for the effects of different SARS-CoV-2 variants. The project is ambitious in a sense as it incorporates novel concepts and approaches, for example the inclusion of inborn errors of immunity (IEI) or their autoimmune phenocopies. Yet, the proposed research plan is based on a solid foundation of preliminary data.

The key hypothesis of UNDINE is that most patients with one or another clinical manifestation of SARS-CoV-2 infection have pre-existing and causal genetic and/or immunological anomalies. Hence, we aim to collect data that will uncover those anomalies. The general approach follows a classical ‘bed side to bench’ and ‘bench to bed side’ approach, generating data to satisfy the following sub-objectives:

a) To define the human genetic and immunological basis of the diverse SARS-CoV-2 disease manifestations,
b) To study the role of auto-Abs to type I IFNs in the pathogenesis of COVID-19 pneumonia and other SARS- CoV-2 disease manifestations,
c) To characterize the impact of novel SARS-CoV-2 variants on the various clinical presentations, in patients with inborn errors of, or auto-Abs to type I IFN immunity,
d) To develop ready-to-use and Point-of care (POC) diagnostic tests for large-scale and accurate detection of auto-Abs to type I IFNs,
e) To improve our knowledge on COVID-19 pathogenesis (human genetics, immune phenotyping, autoantibodies).

1.2 The multidisciplinary approach

To achieve these goals, the UNDINE project coordinates a multidisciplinary and translational research effort relying on a strong and synergic combination of assets: a) unique cohorts of patients and healthy individuals in 11 EU countries and 2 associated countries, UK and Switzerland (cf. D1.1), b) leading expertise in the identification of genetic IEI and auto-Ab-driven phenocopies underlying severe COVID-19, c) a prominent international position of our industrial partner bioMérieux in the development of immunological tests relevant to SARS-CoV-2 infection.

The UNDINE consortium takes advantage of those assets and the well-developed preliminary data. For example, Members of the consortium have already identified candidate gene variants underlying new IEI underlying critical COVID pneumonia, MIS-C/A and their cellular phenotypes. First data are also available for auto-Abs to type I IFNs, their possible genetic origin and the nature of their disease-causing potential in various SARS-CoV-2 disease manifestations (Fig 1.).

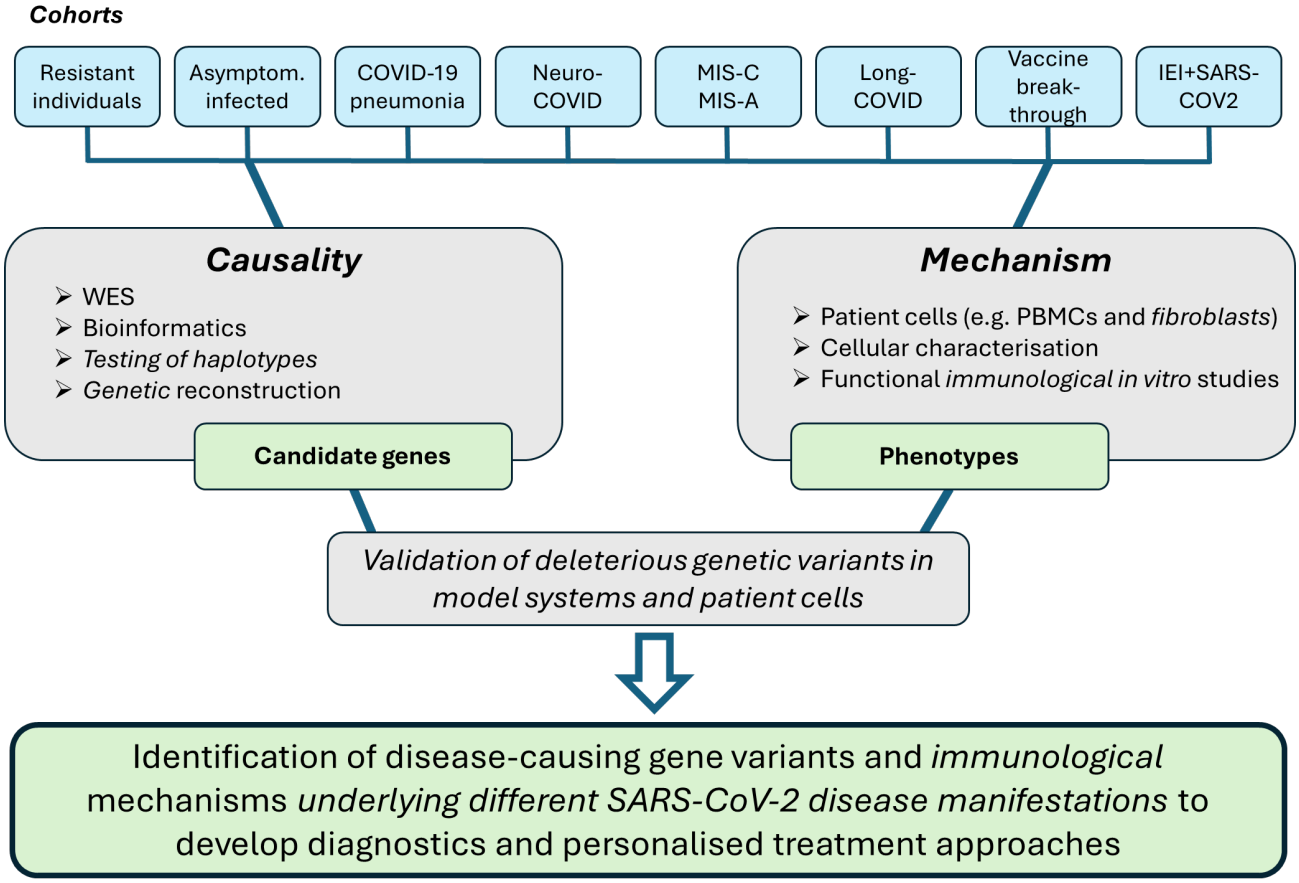

Figure 1. Flowchart illustrating the use of samples taken from the various cohorts for the generation of data in UNDINE.

1.3 Summary of data that will be generated

In particular, the following data will be generated during this project:

a) Clinical information on medical history, disease course, blood tests, imaging results, family history of patients with COVID disease manifestations: including I) critical COVID pneumonia, II) SARS-CoV2 resistance, III) MIS-C/A, IV) COVID toes, V) Long-COVID, VI) neuroCOVID and VII) “breakthrough cases” despite vaccination (cf. D1.1),

b) Whole exome sequencing (WES) and whole genome sequencing (WGS) data and the results of their bioinformatic analysis,

c) Data from biochemical, immunological and cellular functional analyses of patient PBMCs, fibroblasts or other modelling systems (cell lines expressing gene variants found in patients or patient-derived iPSCs) stimulated with relevant ligands or infected by different strains of SARS-CoV2,

d) Serological data on auto-antibodies and cytokines from patients in all categories,

e) Innovation data acquired through the development of patents or analyses for commercial use in a hospital or an outpatient clinic setting.

Overall the consortium aims at an open access and open data sharing strategy in accordance with EU visions and recommendations. Moreover, all data are intended to follow the FAIR principles (see 2. below) to the widest possible extent, within national and local legal directives and rules (GDPR and specific rules concerning whole exome data) and with respect to protect sensitive personal information of patients. To achieve this goal the following general measures will be taken (more details under 2.-7. below):

a) All publications will be open access, and when possible data and preprint manuscripts will be shared publicly. All data in publications will be deposited at data repositories for gene sequencing and HLA phenotyping.
b) All beneficiaries and associated partners will have access to the data within the consortium and in accordance with the Consortium Agreement. This ensure sharing gene sequences and potentially disease causing variants immediately after sequencing, in order to accelerate further data generation and within-consortium collaboration.
c) Unpublished data will be presented and shared at scientific and medical conferences in order to promote fast dissemination of results to the research community as well as to the clinical medical community. It is the explicit aim to promote the UNDINE project, and to ultimately facilitate the transition from research to clinical application and patient benefit with other colleagues in the European Union and worldwide.

Both, submitted Deliverables as well as preprint data and results judged as important and reproducible will be shared through as member-restricted section on the UNDINE homepage. Accordingly, intermediate results of the various WPs will be presented and discussed locally and virtually at regular progress meetings of the entire consortium (every three months). Apart from the UNDINE website, the internal knowledge and data-sharing as well as the public visibility of the project are facilitated through the use of multi-media channels, such as Twitter (#undine-project) and Slack (for quick communication within the consortium).

1.3.1 Data re-used in UNDINE

The composition of cohorts (WP1) used in the UNDINE project are based on the Covid Human Genetic Effort international consortium (see: www.covidhge.com). The main reason for reusing those data is the need for large datasets to decipher genetic and immunologic features that define the various complications with COVID19. Such large sets cannot be compiled during the duration of the UNIDNE project alone.

Work package 2 (WT 2.1 to 2.3, IP) takes advantage of three existing cohorts (EvoImmunoPop, SeroPrevHK-CoV-19, and Milieu Intérieur) for which DNA genotyping and/or sequencing data have already been generated. These genetic data will be reused for both a) the genetic demultiplexing of cells within each single cell RNA-sequencing (scRNA-seq) library and b) the mapping of eQTL and reQTLs. In addition, for both cohorts, the generated data will be integrated with pre-existing individual

metadata on demographic variables (age and sex), and previous exposure to viral infections measured by VirScan3, which detects antibodies against >1,200 viral strains.

In particular, the following data will be re-used:

b) Genotyping/Whole genome sequencing

- For all three cohorts (EvoImmunoPop, SeroPrevHK-CoV-19 & Milieu Intérieur), final genotype calls are stored in gzipped vcf format. This format is standard for sequencing/ genotyping and can be easily reused by the community with open software such as vcftoofs/bcftools.

c) VIRscan data (reused)

- For EvoImmunoPop and Milieu Intérieur cohorts, VIRscan results are stored in .tsv format and include: 1 epitope file per sample in .tsv format, indicating the epitope enrichments observed after immunoprecipitation and a count matrix in .tsv format, showing the number of enriched epitopes per virus for each sample. This format is highly interoperable and can be easily reused and opened without the need of a specialized software.

It shall be noted here that the re-use of genetic data from EvoImmunoPop and SeroPrevHK-CoV-19 cohorts is restricted to academic research on the variability of the human immune response. The use of genetic, demographic and VirScan3 data from the Milieu Intérieur cohort is restricted to research (academic or private) on the variability of the human immune response.

The human sequencing data used for WP3 have in part been generated already in the past by the Laboratory of Human Genetics of Infectious Diseases (JL Casanova lab, Partner 2). Existing data will be re-used. The existing data comprise: 1) whole exome sequencing data, in various states of processing; 2) pseudonymized phenotypic data; 3) pseudonym keys. WES data will be reused and generated as follows:

a) *Data capture tools:* GATK, BWA, SAMtools, etc;
b) *File types:* Raw data, alignments, variant calls, variant annotations;
c) *File formats:* .fastq, .bam, .vcf, .csv. In principle, all original research data generated during the project will remain re-usable for reanalysis.

For the work packages 4-6, basic clinical information has been for more than 5000 patients: age, sex, medical history, serology results, viral SARS-CoV-2 tests results, auto-abs neutralizing type I IFNs and other biological data. Those data were collected from a retrospective cohort and that has been included on the clinical program C10-13 promoted by INSERM and coordinated by the Human Genetics of Infectious Diseases laboratory and will be reused to generate new results. The information is stocked in local servers.

1.3.2 Data generation in relation to UNDINE's work packages

The format, size and collection of data on the cohorts that are used in UNDNE are already specified in Deliverable 1.1. Those data are the sole focus of WP1. The format of the data in WP1 will be:

a) databases, spreadsheets, texts

b) clinical data (pseudonymized),
c) physical data (DNA, RNA, PBMCs, serum),
d) molecular data (BAM files, VCF files, xls files)
e) most used formats for data sharing: xls, doc, pzf (prism), flow jo (flow data), pdf (images, adobe imager).

The data collection in WP2 will comprise single-cell transcriptional profiles (scRNA-seq) on PBMCs from 300 healthy individuals of West European, Central African and East Asian origin stimulated by SARS-CoV-2, IAV or mock controls. This allows for the characterization of the variation in the immune response to SARS-CoV-2 in individuals from different ancestries worldwide (cf. WT 2.1 in the GA). To analyse the effects of sex and age on immune response variation, we will further acquire single-cell transcriptional profiles (scRNA-seq) on PBMCs from 512 healthy individuals of West European ancestry, stratified by age (from 20 to 70 years old) and sex, stimulated by SARS-CoV-2, IAV or mock controls (cf. WT 2.2 in the GA). We will integrate these data with previously generated genetic data (genotyping or WGS) and VIRscan3 data from the same individuals to assess the contribution of genetics and previous pathogenic exposure to variation in immune response to SARS-CoV-2 (cf. WT 2.3 in the GA). For WP 2 (lead partner: IP) the following data types will be generated:

- scRNAseq sequence data in .fastq format (gzipped, text files), 3 files per library: cell barcodes and UMI (unique molecular identifiers), 3' RNA insert and sample index,
- scRNAseq aligned sequences in .bam format, 1 file per library (textual),
- Count matrices in .mtx format (numeric) indicating the number of UMIs associated with each gene and cell, 2 files per library (full & filtered),
- Demultiplexing files in text format indicating the samples to which the cells are associated, 2 files per library, 1 for each program used (demuxlet & freemuxlet),
- Data formats correspond to standard formats for sequencing data and scRNA-seq count data that can be reused by the community (as generated by widely used tools such as CellRanger and STARsolo).

This satisfies the objective 1 from the UNDINE project: “To define the human genetic and immunological basis of the diverse SARS-CoV-2 disease manifestations”. scRNA-seq and associated genotyping/sequencing, VIRscan and demographic data may be useful to any researcher interested in replicating UNDINE’s results, or wishing to address specific questions on the variability of human leukocyte transcriptional responses to SARS-CoV-2 and its genetic and non-genetic basis.

To identify individuals resistant to SARS-CoV-2 infection (resistors) and to perform genetics analysis of resistors and identify candidate variants in WP3 (objectives 1 & 2, cf. Tab 3.1a in the GA), WES data are needed. Original research data will be created to assess the candidate variants and their immunological and functional consequences, and to identify biological markers of resistance to infection with SARS-CoV-2 (objectives 3 & 4). Once accessible, those data will be useful for Human geneticists, immunologists, virologists and epidemiologists.

For WPs 4-6, we generate data that include:

a) Immunological data for the detection of auto-antibodies against type I IFN,
b) Next generation sequencing (NGS) data: WES/WGS from the DNA of patients with auto-Abs,
c) Confirmatory Sanger sequencing on DNA samples of those patients with rare variants identified by WES,
d) Biochemical and molecular biology data on cells (including RT-qPCR, Western-blots, ELISA, immunolabelling and flow cytometry).

Different types of data will be generated depending on the specific technique considered. WGS/WES and Sanger data will consist of annotated assemblies and raw sequence reads (FASTQ and BAM formats and/or other genetic data arising from new generation sequencing of the individuals on the study). Raw data from neutralization assays for autoantibodies generated .csv and .xls files. Raw Western blots, and immuno-labelling data generate TIFF images. All manipulations are made on copies or the original files; quantifications of biochemical and molecular biology data were provided as spreadsheets preferably in .csv format; clinical data were provided as spreadsheets based on medical records and questionnaires, preferably in .csv format. The genomic and transcriptomic data generated in the context of these work packages will be of long-term value for the scientific community. We will make sure they are made accessible for further use through controlled-access (e.g. dbGaP, as described below). The detailed clinical data obtained on study participants will only be kept for the duration of the project, then archived for 10 years.

For work package 7, we will reuse some data previously generated from patients with acute COVID-19 (Rodriguez et al. 2020, *Cell Reports Medicine*), MIS-C (Consiglio et al. 2020, *Cell*), because they will act as references for all newly created datasets from patients with Long-COVID (cf. WP7-description in Tab 3.1a of the GA). Those datasets will eventually comprise immune cell data (Mass cytometry), plasma protein data, Bulk mRNA-seq. data, sc-RNA-seq data, WGS and functional immune cell response data. The purpose of the data collection in WP7, in particular, is to gain a basic understanding of the immunology of Long-COVID and a possible genetic personal predisposition (cf. D7.1). These findings will be of great value to researchers interested in immune dysregulation, function in health and disease.

1.3.3 Data provenance and size

All partners within UNDINE are involved in WP1 and recruit patients. The basis of all studies is a set of defined and unique cohorts of patients and healthy individuals providing genetic material. Deliverable 1.1 provide clear definitions and sizes for all the used clinical phenotypes. The spectrum covers a broad range of rare or common, mild or severe COVID-19 phenotypes. The data used to define the cohorts originate from sequencing, flow cytometry, western blot, demographic information,

clinical information and analysed and filed using prism, adobe, R, Microsoft office, flowjo. The size of the according data set is expected to reach approx. 3TB.

WP2 mainly uses ScRNA-seq data, which are obtained from frozen PBMCs collected from adult, healthy volunteers in either Ghent (Belgium, EvoImmunoPop cohort), Hong Kong (SAR China, SeroPrevHK-19 cohort) or Rennes (France, Milieu Intérieur cohort). Genotyping/sequencing data are obtained from DNA that is extracted from blood samples from the same donors. In addition, VIRscan data used in WP2 were previously generated from plasma samples obtained from the same research participants. Those reused data were generated at Institut Pasteur (Quintana-Murci Lab, partner 3). And, all newly generated data for this WP will be generated at the same place. ScRNA-seq from will be aligned/counted using STAR solo. Genotype calls are obtained from GenomeStudio software (genotyping data), or following GATK best practices pipeline (WGS data). VIRscan data are analysed with custom scripts. We expect the generated data to consist in 406 scRNA-seq libraries, with an average of ~42 Gb of fastq files, ~36 Gb of bam files, and 0.5 Gb of count and demultiplexing data per library, for a total of 32 Tb of data. The genotyping/sequencing data required for the according study consists of ~7 Gb (EvoImmunoPop and SeroPrevHK-CoV-19) and ~215 Gb (Milieu Intérieur) of .vcf files.

As mentioned above, the human sequencing data used for WP3 have in part been generated in the past, and existing data will be re-used. The existing data comprise: WES data, in various states of processing; pseudonymized phenotypic data; and pseudonym keys. Data capture tools will be GATK, BWA, SAMtools, etc; File types: Raw data, alignments, variant calls, variant annotations; and File formats: .fastq, .bam, .vcf, .csv.

The size of the dataset generated in WP4-6 will reach around 20TB, while WP7 is expected to generate data of a size of approx. 10TB.

2 FAIR data

2.1 Applying FAIR to UNDINE's data

All data metadata generated in the work packages of UNDINE will be aligned to the FAIR principles, to be findable, accessible, interoperable and reusable. It is in the joint interest of the consortium and the EC to ensure the highest possible value of the generated data even beyond the duration time of the UNDINE project. Since 'accessible' in FAIR does not mean Open without constraint, the UNDINE consortium will, however, reserves the right to protect some of the most sensitive data, at least for a defined and limited amount of time (see below). In particular, the handling of patient data requires the consent of the patient that data produced by the use of their samples can be shared. While we will ensure that this will be included in the questionnaires given to the patients, their consent is not guaranteed. Moreover, as our industry partner is working on a new diagnostic tool based on

UNDINE's results, confidential commercial information may accrue, which cannot be shared with the public. For any confidential information, we will try to use anonymization techniques to maximise the number of data sharing options.

Apart from those restrictions, we ensure that all data, which are shared publicly and with the scientific community will be provided in a format that allows the swift re-use in new and follow-up projects. Key to this is the utilisation of common software and vocabularies, analysis-platforms, repositories and distribution channels, as described below.

2.2 Making data *findable*, including provisions for metadata

All cohort data created and monitored in WP1 use unique identifiers for genes, such as UniProt ID/ GenBaank. Metadata will include but are not limited to: filtering strategies for analysis of whole exome sequencing, analysis strategy for transcriptomics, methodology for validation experiments. The connection between metadata and data will be made clear by unique identifiers.

The ScRNA-seq data generated for WP2 will be stored on OWEY, the data lake developed by Institut Pasteur (beneficiary 3), and will be assigned a unique DOI, that will serve as persistent identifier. Essential metadata (incl. library design, donor anonymized ID, age, sex, population of origin, sampling site, and sample QC data) will be made publicly available in .tsv format from the OWEY landing page, and provided along with the scRNA-seq datasets. In addition, metadata on the dataset itself will be generated and will follow the metadata format of the European Genome-phenome Archive (EGA): title, author, date, contributor, description, keywords, format, type of resource, etc. At least 3 keywords will be associated with the dataset to facilitate indexing. Summary statistics from eQTL/reQTL mapping will be also be provided and made publicly available to the community through the eQTL catalouge (http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/) and its associated web browser (https://fivex.sph.umich.edu/). The sequencing data of WP3 will receive a DOI-code as their persistent identifier. Unalyzed data will not have a DOI. However, explicit names will be given to the files, which, along with the README files, will ensure their traceability and reuse (also applicable for sequencing data generated in WPs 4-6).

In general, all UNDINE data not yet published will be available to the members of the project in a direct and unlimited way. The same will apply to the published data. High-level data (e.g. summary statistics from genetic analyses) will be made publicly available for unrestricted access. In parallel, additional materials necessary to interpret these data will also be shared, including protocols. The metadata will be filled following the DDI standards. They correspond to biological material, wet-lab protocols (kits, experimental conditions, composition of reagents), data analysis protocols (tools, versions, workflows, etc.).

Finally, all Long-COVID data generated in WP7 will receive persistent identifiers such as GEO accession numbers, FlowRepository IDs and Immport IDs.

2.3 Making data accessible

Non-identifiable metadata and summary statistics will be made openly available to the community. Genetic and transcriptomic data, however, are in most cases considered as sensitive. Accordingly, identifiable data and will therefore be made available under restricted access only, in order to comply with National and European legislations. An embargo will be applied to the data until its publication (within two years of data generation) to preserve intellectual property. After this time period, the dataset is made accessible, perhaps with restricted access. To agree on a consortium-wide data handling policy, all UNDINE beneficiaries will meet on Feb 23rd 2023 in a virtual progress meeting. Here, we will discuss the creation of a 'data access committee' that will have the task to decide on unified access restrictions and access granting of future project results. So far, all beneficiaries have local policies in place, which, however, are not yet synchronised.

The cohort data collected through WP1 are the crucial basis for all experiments of the entire UNDINE project. Genetically validated data will thus first be shared in the academic network only, but readily once available. Cohort numbers will be publicly available, and all data can be used by hospitals and the public once they have been published.

All data generated in WP2 will be made accessible from web interface of the OWEY data lake, maintained by the Institut Pasteur IT department. Arrangements have been made to obtain the required disk space and set up a specific web page to access the dataset (landing page) before the end of the year 2022. The landing page will be associated to a DOI by the Institut Pasteur data librarian (CeRiS – Resource center for Scientific information). All data deposited in OWEY is backed up on magnetic tapes stored at a distinct location within the Institute, and used for long-term storage. The conditions for access are specified on the (publicly accessible) landing page associated with the dataset on OWEY. The person who requests access to the dataset must first fill a form, which provides details regarding the data requester, the data requested, the requester's research project and their affiliations. The data requester will also sign a formal letter, including a verifiable affiliation (in the case of EvoImmunoPop and SeroPrevHK-CoV-19) or a data access agreement (in the case of Milieu Intérieur), where they certify that they will act in compliance with the donor's informed consent, and take the necessary steps to preserve the security and confidentiality of the data provided. They also need to explain how their research project complies with the informed consent signed by research participants. On the basis of these documents, the Legal department of Institut Pasteur and the Data Access Committee will decide whether to grant or deny access. The Data Access Committee will be composed of a jurist, an ethicist and a data protection officer, as well as 3 members from the Human Evolutionary Genetics lab (Lluis Quintana-Murci, Etienne Patin, and Maxime Rotival). Metadata will be made openly available on OWEY and licenced under CC-BY-NC. Data will remain findable via the same platform for an indefinite amount of time.

For WP3, all Partners will use a center-specific dedicated and secured research folder structure for all research data generated. For each analysis, the raw data (in an acceptable open format e.g.

.fsc/.csv/.tiff), and all the files that belong to the experiments and the analyses thereof (including pdf of the lab-journal) will be stored in the raw data directory for each experiment. Also, the metadata of that experiment, including the exact methods/layout of the experiment will be stored in the same folder. This folder will have a specific number and a title. Modifications or updates to these analysis files are tracked using an appended version number at the end of the filename. Combined data analysis files for final publication figures are maintained in separate directories, as they are linked to data from several different experiments performed on different dates. Within the final data analysis directories, there will be a text file containing metadata regarding exactly which source (raw) data is contained within the analysis file, and the directory in which the raw data can be found. The results of WP3 will be published in open access scientific research journals and/or preprint servers. All data and metadata (other than DNA sequence data) underlying published findings will be deposited in an appropriate public repository (such as the generalist repository Dryad (www.datadryad.org) or flow repository (https://flowrepository.org/) to ensure FAIR data and enhancing scientific reproducibility. However, apart from experimental data, WP3 as well as WP7 will generate human genetic data. For reasons of privacy protection, these human genetic data cannot be made available. Hence, all additional/unpublished data and documents in the data package will only be shared upon request and under restrictions. Data for WP7 will be stored in secure repositories, like NCBI GEO, FlowRepository.org, Immport.org accordingly.

The raw data of NGS type, created in WPs 4-6 will be deposited and used for downstream processing and analysis via the individual laboratories' servers. The final processed files will be made available to all members of the project on a shared and secured server. The first level of data organization will be linked to their origin (RNA-seq, WGS, WES), then within each sub-folder, the data separated according to the coding of each individual that has been generated on the psedonymisation process. All data will be encrypted following a standard nomenclature: kind of experiment-date-patient ID-experimental variable, for example: WB GADPH-20092022-MB0002-PBMCs NS.

2.4 Making data interoperable

The cohort data of WP1 are provided in highly interoperable light formats (e.g. vcfs and xls), which are widely used. All sequencing data are of the other work packages are given out as .fastq and .bam formats, and the count matrices and demultiplexing data are in easily manipulated text formats (e.g. .mtx). All these formats are open and interoperable formats. Additional design files required for data analysis will be made publicly available in the according repositories along with the uploaded files.

To prevent that data have been acquired in a format that depends on a specific software that might not be usable anymore, we aim to save all useful raw data in a format that is open access

proof: .csv for all standard data with values (e.g. ELISA's); .csv or .fsc for flow cytometry data; .tiff for imaging data; and, as mentioned above, .fastq for sequencing data.

Most data from genetic analyses are produced and analysed using DRAGEN software (see emea.illumina.com). For biochemical and molecular biology data we use image acquisition and data analysis software: Image Studio, Workflow, Workout and Prisme. All data and stockage information for each individual will be managed by the LIMS system ModulBio.

Internally, all analyses and experiments are performed based on detailed protocols, which are maintained in Electronic Lab Notebooks of person(s) performing the experiments. The protocols will be prepared in word and saved as PDF. This will always include all the required information to repeat the experiment and a layout of the experiment to track individual samples of the raw data. The information contained within notebook entries corresponds with collected data on the same date, therefore the data can be reproduced by following the protocol therein.

2.5 Increasing data re-use

In principle, all experimental analyses and data will remain re-usable for reanalysis. To prevent that data have been acquired in a format that depends on a specific software that might not be usable anymore, we aim to save all useful raw data in a format that is open access proof:

-.csv for all standard data with values (e.g. ELISA's),
-.csv or .fsc for flow cytometry data,
-.tiff for imaging data,
-.fastq for sequencing data.

A readme file will be provided for each dataset with a summary of the methodology used for analyses and quality control. All code associated with the project will be deposited on a dedicated git-hub page referenced in the associated publication and in the repository.

The scRNA-seq dataset generated as part of WP2 as well as the human genetic data generated in WP3 will not be made freely available as they contain sensitive, identifiable data, therefore no license is associated with it. Access to the dataset will be provided upon request (see 2.3). Once published, the data from this dataset will be available for reuse by the entire scientific community, under a restricted access policy, with the restriction that the data should not be used for commercial purposes.

Long-term data storage: Raw and processed/analysed data will be stored for 15 years after the completion of the project. The data will be stored in widely used file formats to avoid compatibility and accessibility problems in the future.

3 Other research outputs

Within the work packages of UNDINE, a few results will be produced that are not categorised using a “classical” data caption. Those outputs comprise scripts, protocols and software. For example, all scripts and software generated as part of the analysis of WP2 data (cf. WT 2.1 to 2.3 in the GA) will be shared publicly on GitHub. Also, summaries of statistical results based on scRNA-seq data from WP2 (e.g. mean expression profiles of cell types identified a basal state and after SARS-CoV-2 stimulated, eQTL/reQTL summary statistics, logFC in expression between population, etc.) will be made publicly available on the OWEY landing page of IP, as supplementary tables in open access publications or via the EMBL eQTL catalog.

All alternative research outputs generated will be made openly available under a CC-BY-NC license in interoperable file formats (gzipped tsv files for table, and text files for scripts), deposited on well-established public repositories with long-term storage capacity (git-hub, eQTL catalog, scientific journals, etc.).

4 Allocation of resources

All costs associated with the data collection, data storage and repository maintenance are covered by the UNDINE budget. A special case represents work package 1, where a the project has coverage for the salary of post-doc monitoring the cohort composition, which represents the crucial basis for the experiments of all other research work packages of UNDINE. If any additional costs accrue, the individual beneficiaries and associated partners have agreed to compensate from additional funds. For example, the IT department of IP will not only provide technical but also financial support to allow the sharing of WP2 data. All this is always done under the FAIR requirements.

The adherence to the open access and FAIR principles will be monitored by the project manager of UNDINE, as well as the co-ordination of the data storage and dissemination efforts. Locally, each beneficiary and associated partner will utilise their administrative staff and IT sections to take care of the data management, including the long-term storage and backups of data (for example, using magnetic tapes for backups at IP theoretically allows data storage for an indefinite amount of time).

5 Data security

Samples from each individual are encrypted (anonymised/pseudonymised) once they arrive at the labs. Upon arrival, each individual is given with an internal code (for example generated by LIMS system ModuBio). The coding key remains within the laboratory and it is computer secured. Technical and organizational measures will be taken to guarantee confidentiality, integrity,

availability and resilience of the systems with regard to the processing of data. As general measures in UNDINE we will:

a) deny unauthorized persons access to facilities and data processing systems,
b) prevent unauthorized persons from reading, copying, altering or deleting data in/from data processing systems,
c) ensure that unauthorized persons are not able to read, copy, modify or remove data upon the electronic transfer of data as well as during the transport of data carriers or saving of data thereon,
d) ensure that it is possible to examine and verify if, when and by whom data was entered into the data processing system or if, when and by whom data was modified or removed,
e) ensure that data is protected from accidental destruction or loss,
f) guarantee that the efficacy of technical and organizational measures is regularly reviewed and assessed,
g) implement corrective measures and automatic reporting in case of any suspected data security breach.

In summary, all raw data deriving from any research work package are a priori personal data, which must remain confidential and only accessible to other beneficiaries and associated partners as specified in the consortium agreement. All partners within UNDINE have committed themselves to further take appropriate local measures to ensure data security at their research institute. Accordingly, during the project, all data are stored on a shared storage space, on servers protected internal network firewalls. Access to this space is reserved to the local project participants and other partners within UNDINE. The access will be controlled by a password of at least 12 characters. We recommend that all data are also stored on Dual hard drives that provide an automatic back-up system with saves being performed at regular intervals.

In addition to the UNDINE-wide recommendations, individual partners are welcome to use additional measures to ensure data security. For example, and as mentioned above, IP uses a long-term storage of data on magnetic, located in a secure room with controlled access. WP1 data are stored in a secured server of KUL, with regular backup. The entry is protected by a two-step authentication. All scRNA-seq generated data and the genetic and VirScan3 re-used data as part of the WP2 will be pseudonymized and encrypted with an encryption software provided by IP IT department. Specifically, the scRNA-seq data generated and analyzed are deposited in the secured Institut Pasteur data repository, OWEY, which can be accessed at: https://doi.org/10.48802/owey.e4qn-9190. The genetic data used are also deposited in OWEY and can be accessed at https://doi.org/10.48802/owey.pyk2-5w22. In accordance with the General Data Protection Regulation (GDPR) in force in the European Union, the aforementioned data can only be accessed from the institutional data repository after authorization by the relevant Data Access Committee (DAC). The DAC ensures that data access and use is authorized for academic research relating to the variability of the human immune response, as defined in the informed consent signed by research

participants. For all research data generated for WP3, the lead beneficiary has installed a dedicated folder, which is located on a secured server of the UMCU. Only researchers working on the project and the data managers have access to this folder. A back-up of all research folders is made daily. The adherence to UNDINE's data security policy will be a main focus of the reports of the ethics advisors, who will monitor the compliance with all ethical standards (see previous ethics reports: D9.2 and D9.3 and cf. future reports: D9.4, M36 and D9.5, M48).

6 Ethics

UNDINE handles a large amount of genetic data from research using patient material. Sharing of those very sensitive data can only be performed under a strict restricted access policy, in compliance with National and European data protection laws and the informed consent obtained from study participants. We will make sure that consent for long-term preservation and/or sharing of personal data will be a part of questionnaires and consent forms for all participating patients and healthy donors of biological samples. We further offer that all personal information can be converted into pseudonymized data in a secured and independent file that is saved in a folder with restricted access. A pseudonymized identifier will then be used to link phenotypic data from genetic and experimental analyses.

In particular, ethical issues potentially raised by this project include the collection use of personal data and samples from otherwise healthy young patients with severe COVID-19 forms at the national and international level (and their family members when available) and from healthy or asymptomatic infected individuals. For that, all required ethical approvals have already been obtained from the relevant ethics committees. For every individual included in the study, we obtain informed consent for the preservation of clinical data at the moment of the inclusion. All participants are informed by the investigator physician and they receive a written information note. They have the right to withdraw from the study and request the destruction of samples and data at any time.

Given the nature of the data generated in the UNDINE project, all partners will work closely with their technology transfer offices to ensure that any resources and findings generated that have the potential to be commercialized are treated as sensitive information until intellectual property protection and potential patent filing is in place. These matters of rights are covered and detailed in the consortium agreement.

7 Other issues

None.