1000G

A Data Management Plan created using DMPonline

Creators: Sander W. van der Laan, Vinicius Tragante, PLEASE UPDATE YOUR DETAILS, Kristel Kool, Vinicius Tragante, Jessica van Setten, Charlotte Onland

Affiliation: Other

Funder: European Commission

Template: UMC Utrecht DMP

ORCID iD: 0000-0001-6888-1404

ORCID iD: 0000-0002-2360-913X

ORCID iD: 0000-0002-4934-7510

ORCID iD: 0000-0002-8223-8957

ID: 80380

Start date: 01-01-2021

End date: 01-01-3000

Last modified: 06-01-2022

1000G

1. General features

1.1. Please fill in the table below. When not applicable (yet), please fill in N/A.

DMP template version	29 (don't change)
ABR number *(only for human-related research)*	n/a
METC number *(only for human-related research)*	C-01.18
DEC number *(only for animal-related research)*	n/a
Acronym/short study title	1000G
Name Research Folder	
Name Division	Laboratories, Pharmacy, and Biomedical genetics
Name Department	Central Diagnostic Laboratory
Partner Organization	
Start date study	2021-01-01
Planned end date study	3000-01-01
Name of datamanager consulted*	Saskia Haitjema
Check date by datamanager	January 6th 2022

1.2 Select the specifics that are applicable for your research.

- Fundamental / translational study
- Non-WMO

We will use data from the 1000G (https://www.internationalgenome.org) as reference in many studies or as part of a course curriculum in practicals to learn genetic analyses methods.

The 1000 Genomes Project created a catalogue of common human genetic variation, using openly consented samples from people who declared themselves to be healthy. The reference data resources generated by the project remain heavily used by the biomedical science community.
The International Genome Sample Resource (IGSR) maintains and shares the human genetic variation resources built by the 1000 Genomes Project. We also update the resources to the current reference assembly, add new data sets generated from the 1000 Genomes Project samples and add data from projects working with other openly consented samples.

2. Data Collection

2.1 Give a short description of the research data.

Subjects	Volume	Data Source	Data Capture Tool	File Type	Format	Storage space
Human	1	Genotype data	R, SNPTEST, GCTA, etc	PLINK-format, Oxford-format	.vcf, .bed/.bim/.fam, .gen/.sample	±1Tb

2.2 Do you reuse existing data?

- Yes, please specify

Existing data from the 1000G:
- Genotype data
- Some 'clinical' data, i.e. age (when available), sex, relationships (parent-child)

2.3 Describe who will have access to which data during your study.

Please note, that the data has been de-identified for the purpose of public sharing.

Type of data	Who has access
Pseudonymized data	Research team, Datamanager

2.4 Describe how you will take care of good data quality.

#	Question	Yes	No	N/A
1.	Do you use a certified Data Capture Tool or Electronic Lab Notebook?			x
2.	Have you built in skips and validation checks?			x
3.	Do you perform repeated measurements?			x
4.	Are your devices calibrated?			x
5.	Are your data (partially) checked by others (4 eyes principle)?			x
6.	Are your data fully up to date?	x		
7.	Do you lock your raw data (frozen dataset)	x		
8.	Do you keep a logging (audit trail) of all changes?	x		
9.	Do you have a policy for handling missing data?			x
10.	Do you have a policy for handling outliers?	x		

2.5 Specify data management costs and how you plan to cover these costs.

#	**Type of costs**	**Division** ("overhead")	**Funder**	**Other** (specify)
1.	Archiving	x		
2.	Storage	x		
3.	Maintenance Dataset		x	
4.	Datamanager	x		
5.	Data analysis tool	x		

2.6 State how ownership of the data and intellectual property rights (IPR) to the data will be managed, and which agreements will be or are made.

The International Genome Sample Resource (IGSR) and the 1000 Genomes Project

IGSR was set up to ensure the future usability and accessibility of data from the1000 Genomes Project and to extend the data set produced by the 1000 Genomes Project to include new data generated from the 1000 Genomes Project samples and new populations where sampling has been carried out in line with IGSR sampling principles.

The 1000 Genomes Project ran between 2008 and 2015, creating the largest public catalogue of human variation and genotype data. As the project ended, the Data Coordination Centre at EMBL-EBI received funding from the Wellcome Trust to create IGSR with the following aims:

1. Ensure the future access to and usability of the 1000 Genomes reference data
2. Incorporate additional published genomic data on the 1000 Genomes samples
3. Expand the data collection to include new populations not represented in the 1000 Genomes Project

Disclaimer

IGSR is part of EMBL - European Bioinformatics Institute (EBI) and the "Terms of Use for EMBL-EBI Services" apply to online services, data and software provided by IGSR. The following information specific to IGSR should be read in addition to the Terms of Use for EMBL-EBI Services.

In relation to the "Data Services" section of the "Terms of Use for EMBL-EBI Services", users should be aware that data made available by IGSR comes from many different owners and that consequently restrictions on different pieces of data within IGSR and rights claimed on pieces of data vary. These variations can occur both between and within subsets of data in IGSR. In addition, restrictions and claimed rights may vary over time.

Where specific restrictions or claimed rights have been made known to IGSR, that information will be provided by IGSR with the data, however, IGSR can not guarantee the information being accurate for any purpose. It remains the responsibility of users to ensure that their exploitation of the data does not infringe any of the rights of third parties, including the data owners.

Data from the 1000 Genomes Project is now available without embargo, following the final publication from the project. Use of the data should be cited in the usual way, with current details available at http://www.internationalgenome.org/faq/how-do-i-cite-1000-genomes-project.

Data from the Human Genome Structural Variation Consortium (HGSVC) continues the philosophy of the 1000 Genomes Project, making data available prior to publication in line with Fort Lauderdale principles, allowing others to use the data but allowing the data producers to make the first presentations and to publish the first paper with global analyses of the data. Users should see the data reuse statement accompanying the data.

For all data collections in IGSR, please check the accompanying data reuse statements and cite any available publications appropriately.

For any enquiries, including the terms of use of data and citation, please contact info@1000genomes.org.

Privacy and Cookies

As noted above, IGSR is part of EMBL - European Bioinformatics Institute (EBI) and EMBL-EBI's policies apply.

We collect information related to our legitimate interests in providing services to you, to help improve our resources and for the purposes of day to day running of the IGSR resources and underlying infrastructure. Information collected may remain stored beyond the life of the service.

When browsing our website we collect information about your IP address, date/time of visit, page visited, browser type, data transferred and success of the request. This collection and processing is done by EMBL-EBI. We also use Google Analytics.

IGSR uses Google Analytics as a third party tracking service, but we don't use it to track you individually or collect personal data. Instead it collects information about website performance and how users navigate through and use our site helping us design better interfaces.

Google Analytics gathers certain simple, non-personally identifying information over time, such as your anonymised IP address, browser type, internet service provider, referring and exit pages, time stamp and similar information. We do not link this information to any of your personal data.

Google provides further information about its own privacy practices and offers a browser add-on to opt out of Google Analytics tracking.

IGSR uses cookies to ensure you are aware of our cookie and personal data policies. By using our website, you agree that we can place these types of cookies on your computer or device. If you disable your browser or device's ability to accept cookies your ability to use our services will suffer. You can view more details about the cookies in use on EMBL-EBI's sites from https://www.ebi.ac.uk/about/cookie-control.

Further details are available in the Privacy Notice for this service.

There is also a Privacy Notice for our FTP site and a Privacy Notice for our helpdesk at info@1000genomes.org, which are specific to those services.

3. Personal data (Data Protection Impact Assessment (DPIA) light)

Will you be using personal data (direct or indirect identifying) from the Electronic Patient Dossier (EPD), DNA, body material, images or any other form of personal data?

- Yes, go to next question

3.1 Describe which personal data you are collecting and why you need them.

Which personal data?	Why?
Genotyping data	To answer the research question.

3.2 What legal right do you have to process personal data?

- Study-specific informed consent

3.3 Describe how you manage your data to comply to the rights of study participants.

Right	Answer
Right of Access	Research data are coded, but can be linked back to personal data, so we can generate a personal record at the moment the person requires that. This needs to be done by an authorized person.
Right of Rectification	The authorized person will give the code for which data have to be rectified.
Right of Objection	We use informed consents.
Right to be Forgotten	In the informed consent we state that the study participant can stop taking part in the research. Removal of collected data from the research database cannot be granted because this would result in a research bias.

3.4 Describe the tools and procedures that you use to ensure that only authorized persons have access to personal data.

We use a secured Research Folder Structure that ensures that only authorized personnel has access to personal data, including the key table that links personal data to the pseudoID.

3.5 Describe how you ensure secure transport of personal data and what contracts are in place for doing that.

In case we need to transport personal data with colleagues, we use Surffilesender with encryption.
In such events that we collaborate with outside collaborators, we first set up a Research Agreement and/or Data Transfer Agreement regardless of the current consortium agreement.

4. Data Storage and Backup

4.1 Describe where you will store your data and documentation during the research.

The digital files will be stored in a secured Research Folder Structure of the UMC Utrecht. We will need +/- 1 Tb storage space, so the capacity of the network drive will be sufficient.
For purposes of analyses digital files are partly and temporarily stored on the high-performance computer cluster (HPC) facilitated by the institute or a UMC Utrecht owned and managed device.
Data storage is only accessible to authorized personnel.

4.2 Describe your backup strategy or the automated backup strategy of your storage locations.

All (research) data is stored on UMC Utrecht networked drives from which backups are made automatically twice a day by the division IT (dIT).
We will have multiple copies 1) at the HPC, and 2) at the UMC internal network.

5. Metadata and Documentation

5.1 Describe the metadata that you will collect and which standards you use.

We do not collect anything else, but the data we can obtain through a download. This includes relationships (parent-child), age (when available), and sex.

5.2 Describe your version control and file naming standards.

We will use GitHub as version control with a specific GitHub repository for the each individual project.
We will use the release-system native of GitHub and where possible link it to Zenodo (code only!).

6. Data Analysis

6 Describe how you will make the data analysis procedure insightful for peers.

We will write an analysis plan in which we state why we will use which data and which statistical analysis we plan to do in which software. The analysis plan will be stored at GitHub or potentially through a pre-registration server, e.g. OSF. This way this will be findable for our peers.

7. Data Preservation and Archiving

7.1 Describe which data and documents are needed to reproduce your findings.

The data package will contain: the raw data, the study protocol describing the methods and materials, the script to process the data, the scripts leading to tables and figures in the publication, a codebook with explanations on the variable names, and a 'read_me.txt' file with an overview of files included and their content and use.

Where it is relevant this is amended by an Electronic Lab Notebook (ELN) and handwritten (legacy) lab journals.

After finishing the project, documentation for the ELN will be stored at the UMC Utrecht [GIVE FULL PATH] and is under the responsibility of the Principal Investigator of the research group.

** I will update 'XXX' in this answer when available.*

7.2 Describe for how long the data and documents needed for reproducibility will be available.

Data and documentation needed to reproduce findings from this WMO study will be stored for at least 15 years.

7.3 Describe which archive or repository (include the link!) you will use for long-term archiving of your data and whether the repository is certified.

We do not 'own' the data, it is controlled/managed by the IGSR: The International Genome Sample Resource. We will only keep copies for local use, and potentially archive projects through Archivemetica and share codes used publications etc through DataverseNL according to the principles of FAIR. At the same time a copy will remain at the department server in the existing Research Folder Structure and is under the responsibility of the Principal Investigator of the research group.

7.4 Give the Persistent Identifier (PID) that you will use as a permanent link to your published dataset.

When we get DOI-codes we will update this plan to included these.

8. Data Sharing Statement

8.1 Describe what reuse of your research data you intend or foresee, and what audience will be interested in your data.

Specifically the methods and codes developed for the use of this data will be of interest to our peers. Since the data is managed by the IGSR: The International Genome Sample Resource we refrain from stating anything regarding data re-use, other than that in general these data make for an excellent population reference for multiple purposes.

8.2 Are there any reasons to make part of the data NOT publicly available or to restrict access to the data once made publicly available?

- Yes (please specify)

As the data is privacy-sensitive, and managed by the IGSR: The International Genome Sample Resource we will refrain from sharing these data publicly; this should go through IGSR.

8.3 Describe which metadata will be available with the data and what methods or software tools are needed to reuse the data.

Publications will be open access. The study protocol and this Data Management Plan will also be available.
Along with the publication, the codebook of the data and scripts of analyses will be available through GitHub.
Data (raw or processed) will be accessible under conditions set forward by the IGSR: The International Genome Sample Resource.

8.4 Describe when and for how long the (meta)data will be available for reuse

- Other (please specify)

Meta data will be accessible under conditions set forward by the IGSR: The International Genome Sample Resource.

8.5 Describe where you will make your data findable and available to others.

We will publish and archive publication, codes, etc as described above through Archivemetica (local archiving) and DataverseNL (public) with a note that the data will be accessible under conditions set forward by the IGSR: The International Genome Sample Resource.