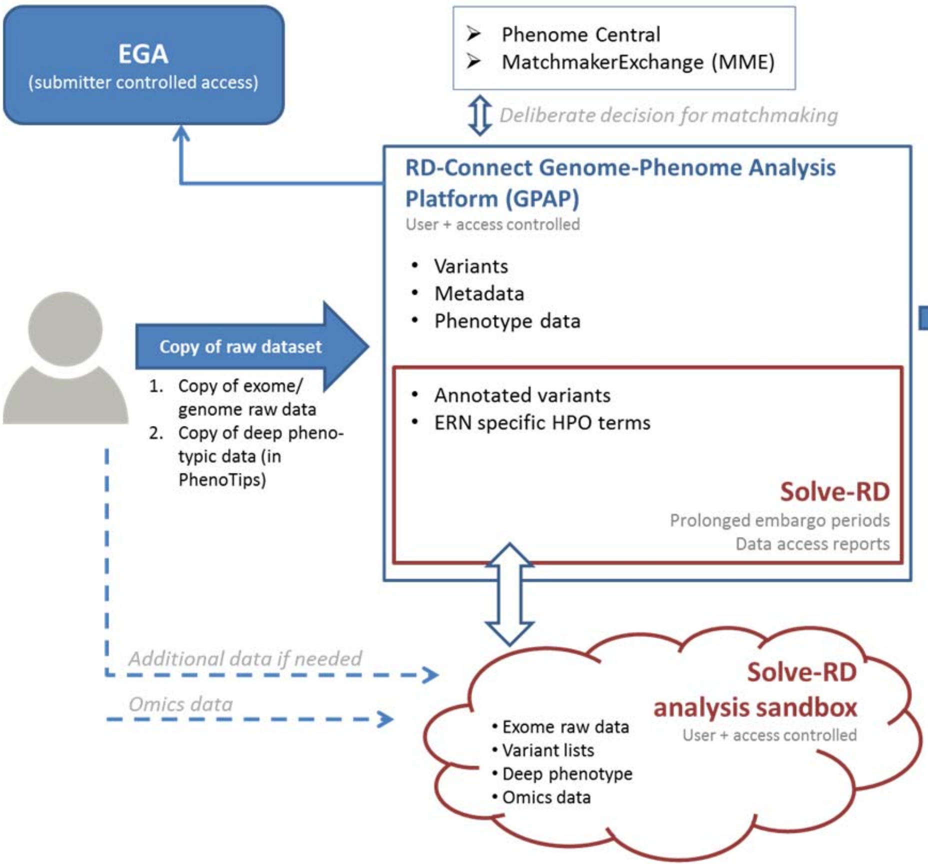
Solve-RD Data Sharing Policy

Version	Final V6
Date	19.03.2019
Authors	Sergi Beltran, Tony Brookes, Holm Graessner, Tina Harmuth, Morris Swertz, Birte Zurek
Approved by SC	08.04.2019

Definitions

- Raw dataset: exome or genome sequencing data in FASTQ or BAM format with corresponding phenotype data in HPO format from a single individual.
- Data collection: multiple raw datasets from a single centre.
- Processed dataset: a raw dataset whose exome or genome sequencing data has been processed through the Solve-RD/RD-Connect variant calling analysis pipeline. It includes the variant calls, some metrics and the corresponding phenotypic information.
- RD-Connect Genome-Phenome Analysis Platform (GPAP): platform that enables collation, processing, analysis, interpretation and sharing of integrated genome and phenome datasets. The RD-Connect GPAP will only retain processed datasets (https://platform.rd-connect.eu/).
- European Genome-phenome Archive (EGA): storage facility that is used by the RD-Connect GPAP for long-term storage of the raw datasets (https://ega-archive.eu).
- Solve-RD analysis sandbox: secure compute environment / private 'cloud' for tailored and *de novo* bioinformatics analysis. Data will be deleted after the Solve-RD analysis have been finalized that is at the latest two years after the end of the project (i.e. after December 31st 2024).
- PhenoTips: user friendly tool integrated in the RD-Connect GPAP to collate and store the phenotypic information from a dataset using ontologies and standards such as HPO, ORDO and OMIM (https://platform.rd-connect.eu/, https://phenotips.org).

Explanation figure of data sharing process


EGA
(submitter controlled access)
➢ Phenome Central
➢ MatchmakerExchange (MME)
Deliberate decision for matchmaking
RD-Connect Genome-Phenome Analysis Platform (GPAP)
User + access controlled
• Variants
• Metadata
• Phenotype data
• Annotated variants
• ERN specific HPO terms
Solve-RD
Prolonged embargo periods
Data access reports
Copy of raw dataset
1. Copy of exome/ genome raw data
2. Copy of deep pheno-typic data (in PhenoTips)
RD Community
Metadata for DISCOVERY (enable finding) of datasets
Additional data if needed
Omics data
Solve-RD analysis sandbox
User + access controlled
• Exome raw data
• Variant lists
• Deep phenotype
• Omics data

Most important points of Solve-RD data sharing policy

- All exome/genome and phenotypic data that will be collated in task 1.2 of Solve-RD will be submitted to the RD-Connect Genome-Phenome Analysis Platform (GPAP)
- All IT systems used in Solve-RD comply with the General Data Protection Regulation, GDPR (Regulation (EU) 2016/679)
- Options for uploading datasets include single dataset upload and bulk upload
- Specific data access stipulations for Solve-RD in GPAP:
  - All data submitters will be able to see which other users have accessed their submitted datasets and when
  - If justified Solve-RD data submitters can define longer embargo periods before data become accessible to other users

Upload of exome/genome and phenotypic data to the RD-Connect GPAP

All exome/genome and phenotypic data that will be collated in task 1.2 of Solve-RD will be submitted to the RD-Connect Genome-Phenome Analysis Platform, an IRDiRC recognised resource. This empowers the clinical end-users and their research teams to analyse and interpret their own data and actively participate in solving cases rather than only handing it over to the project and waiting for results.

Where unsolved processed datasets from previous projects are already available in the platform, there will be an option to assign these to the Solve-RD project, while also keeping the assignment to the original project (if this is the case). An option to assign a new or existing dataset to a specific European Reference Network (ERN) will also be available.

Options for uploading

One raw dataset is considered to include both **phenotypic data** in the form of HPO terms and **exome/genome data** of a patient, preferentially in FASTQ format, although BAM files are also accepted.

There are two options for uploading datasets.

i. **Data collections from centres that will upload fewer than 100 raw datasets**: raw datasets will be uploaded using the standard RD-Connect GPAP upload interface, which includes user friendly PhenoTips templates to enter the phenotypic information and user-friendly tables to upload the genomic data and metadata. An option to bulk upload the genomic data and metadata is also available.

ii. **Data collection from centres that will upload 100 or more raw datasets**: in cooperation with CNAG-CRG in Barcelona there will be the opportunity to discuss a customised bulk upload option to facilitate uploading large data collections.

All sequencing data are submitted as raw data in FASTQ (or BAM) format and are processed through the same Solve-RD pipeline. The processed datasets will be stored in the platform indefinitely, or as long as the funding allows, to allow future reanalysis, reinterpretation or increase statistical power (see below for long-term storage of raw data at the EGA). Clinical interpretation of the data (final or temporary) can be entered in the RD-Connect platform but this is not a requirement at upload.

The raw data will be stored in the European Genome-phenome Archive (see below) for long-term archiving.

Data security and accessibility

RD-Connect pays strict attention to data quality and security and the data in the Platform meet high quality and safety standards. The RD-Connect registration process includes user validation as defined in the RD-Connect GPAP Code of Conduct (https://rd-connect.eu/gpap-code-conduct), which all users must confirm they accept. Additionally, the PI/group leads must sign the Adherence Agreement.

Other computer systems used in Solve-RD will adhere to similar standards.

(i) Who can access the processed dataset/s in the RD-Connect Genome-Phenome Analysis Platform?

- Data submitters can request an embargo period for each of their datasets. During the embargo period the data is accessible only to the members from the submitter group and the Solve-RD user group. However, the members of the original submitter group can share it specifically with other group/s of users (from another PI/group lead).
- After the embargo period set by the submitter, datasets uploaded to the RD-Connect GPAP will become accessible to all authorised scientists and clinicians who have gone through the strict registration and verification process.
- The embargo period is considered to start at the moment a specific processed dataset (genomic plus phenotypic data) is made accessible to the dataset submitter.

(ii) Who can access Solve-RD processed datasets within the RD-Connect GPAP and under which conditions?

- Account registration: Every PI/group lead enrolled in Solve-RD or in a participating ERN will undergo the full RD-Connect registration process and will then enrol members of their team. All the users under the responsibility of one PI/group lead are assigned to the same user group and have the same user permissions but have different usernames and passwords to enable user-specific logs. In addition, every PI/group that will upload samples for Solve-RD and that is not a Solve-RD beneficiary will have to sign an association agreement with Solve-RD containing the Solve-RD Data Sharing Policy and Publication Policy.

- Solve-RD tagging of datasets: New datasets uploaded specifically for the Solve-RD project will be assigned to the Solve-RD project to allow project-wide sharing and monitoring. Pre-existing unsolved processed datasets can be added to the Solve-RD project (while also keeping the original project tag). Datasets must also be tagged with the name of the submitting European Reference Network (ERN) so that it is possible to follow numbers submitted per ERN.

- No embargo period within Solve-RD: Data become accessible to all other authorised Solve-RD users immediately after submission. Solve-RD users can analyse and query their own datasets as well as datasets submitted by other Solve-RD users.

- Access of Solve-RD datasets by other users of the RD-Connect Genome-Phenome Analysis Platform that are not part of Solve-RD:

  i. Embargo periods of up to twelve months: As part of the online data submission process, Solve-RD users can easily define an embargo period of up to twelve months before data become accessible to other users of the RD-Connect Genome-Phenome Analysis Platform. Solve-RD users selecting and embargo period of up to six months will not be asked for any additional information, but Solve-RD data submitters requesting an embargo period between six and twelve months will be asked to type a short justification. RD-Connect has agreed to delegate approval of embargo periods between six and twelve months to the Solve-RD Steering Committee, which will only contact Solve-RD data submitters if the embargo period has been denied or more details are needed.

  ii. Embargo periods longer than twelve months: depending on the origin and nature of data (for example if they are diagnostic data) Solve-RD users can request (in written form at the time of submission) an embargo period longer than twelve months before data become accessible to other users of the RD-Connect Genome-Phenome Analysis Platform. Requests longer than twelve months will require acceptance by the Solve-RD Steering Committee and by the RD-Connect GPAP Access Committee since they are considered non-compliant with IRDiRC principles for rapid data release. Retrospective requests for prolonged embargo periods are not possible.

- Access reports: within the RD-Connect GPAP data management portal, a dataset submitter is able to see which other users have accessed at which date a given dataset through a specific query on that dataset or through a general query.

(iii) What can users do with the data?

- During embargo: Datasets are only accessible to the submitter and the Solve-RD project members. Members from the submitter group can share specific datasets with other RD-Connect GPAP users. Users with access to the datasets will be able to discover, query, analyse, interpret and tag them. If the submitter group has opened the dataset to matchmaking through MatchMaker Exchange, internal RD-Connect users and external users across the globe performing a matchmaking query may be informed that there is a dataset containing a potential match, but the user cannot see the relevant dataset and must contact the submitter to find out more details or request sharing.

- After embargo: Datasets are accessible to the other authorised users within the RD-Connect GPAP, who will also be able to discover, query, analyse, interpret and tag them. The datasets also need to be specifically opened by the submitter group for matchmaking.
- Download of data: Direct download of full datasets is not possible at any time. Download of search results will be restricted to the user group that submitted the dataset and users with whom it may have been specifically shared by the submitter group.

(iv) Data security

Data is stored in a computer cluster with a restricted access policy, limited internet access and daily backups. Databases are using distributed filesystems, limiting the risk of physical attacks. All communications are encrypted. Security of the platform was audited in October 2017 with no major risks being identified. Platform requests and user actions are safely logged for audit purposes. Documentation and procedures are currently being adapted for the new General Data Protection Regulation, GDPR (Regulation (EU) 2016/679).

Matchmaking and discovery

The RD-Connect GPAP is integrated in the Beacon Network (https://beacon-network.org), a project by the Global Alliance for Genomics and Health (GA4GH).

RD-Connect participates in MatchMaker Exchange (MME, http://www.matchmakerexchange.org) and MME is functional for internal GPAP queries and bi-lateral queries to PhenomeCentral. Bi-lateral queries to DECIPHER are being implemented. Additionally, patient profiles can be pushed to PhenomeCentral (https://www.phenomecentral.org) from the RD-Connect GPAP PhenoTips instance by the data submitters. The dataset submitter must specifically enable matchmaking at the time of submission or in the data management portal. This permission can also be enabled or disabled at any later stage. Increasing powerful and precise forms of matchmaking will be developed by Solve-RD in conjunction with GA4GH and others, to enable ever more sophisticated dataset discovery and matchmaking with more options for data protection. Dataset submitters will have to specifically enable the datasets for new forms of discovery and/or matchmaking outside the RD-Connect GPAP and/or Solve-RD if they are less restrictive than the current MME v1.

Patient security and confidentiality

To protect patient privacy, explicitly identifiable patient information is never submitted to the Platform. The submitting clinician stores the data in the manner appropriate for their own centre and links it at a secure local level to the unique RD-Connect IDs. Patient identities are therefore not stored on the Platform and cannot be accessed by Platform users. Only the researcher who submitted the data has the key to identify the IDs corresponding to his/her data.

(v) Return of results

Solve-RD is a research project and as such cannot ensure the quality standards required for genetic diagnostics. It is thus the responsibility of the clinician/researcher who submitted the data and who is the only one who has access to the patient and family to validate any novel genes and to return the results to the patient and her/his family (depending on the consent given).

It is possible that other GPAP users identify genetic variants in sequencing data submitted by Solve-RD partners. These variants may explain the cause of the patients' disease but they may also be completely unrelated – so called 'incidental findings'. Some clinicians don't want to be informed about such incidental findings. We have thus put a system in place where the data submitter can indicate for every patient if they DO NOT WANT to be notified about (forced) incidental findings: "This patient has not consented to be notified about incidental findings and I don't want to be contacted regarding incidental findings on this patient". The system to contact other users, which would basically send an email to both data submitter and the "data analyst" to put them in contact on a certain experiment, includes this message so that the "data analyst" would know that he/she should NOT contact the submitter for any findings which are not related to the patients' disease.

(vi) Storage of raw data in the European Genome-phenome Archive (EGA)

The raw data will be stored indefinitely for long-term access at the European Genome-phenome Archive (EGA), a secure, controlled-access repository. The EGA serves as an archive for publications as well as data on several levels, including the raw data (to allow future reanalysis using other algorithms) and the genotype calls (information about pathogenic genetic variants) provided by the data submitters.

The EGA provides the necessary security required to control access to the data and maintain patient confidentiality. Data can be accessed only by authorised researchers and clinicians. In all cases, data access decisions are made not by the EGA but by an appropriate Data Access Committee, which can be the person or group submitting the data.

Data must be submitted to the EGA at the latest by the end of Solve-RD. At the time of uploading a dataset to the RD-Connect GPAP, the user can indicate if the dataset is already available at the EGA and provide the corresponding reference number. For datasets not yet available at the EGA, the CNAG-CRG will broker the submission to the EGA of the data and metadata uploaded to the RD-Connect GPAP. The original data submitter will be responsible for making decisions regarding the future access to their datasets.

(vii) Solve-RD publications - notification and authorship policy with regard to shared data

All Solve-RD publications are acknowledged to be based on the fundamental principles of open scientific collaboration, reciprocity, attribution and benefit sharing. For any publication resulting from work carried out using data shared or generated through Solve-RD (e.g. for identifying a novel gene), including where data has been accessed through the RD-Connect Genome-Phenome Analysis Platform, the Analysis Sandbox or future Solve-RD systems, the authors should in all cases acknowledge and give appropriate authorship positions to all relevant parties in line with best practice for acknowledgement of scientific contribution including submission of the primary data (also see the Solve-RD Publication Policy).

Examples and further principles are described below.

1. **A publication arising from research in which the party leading the publication ("the PI team") is primarily analysing their own submitted data (example: novel gene discovery by a submitter analysing their own patient cohorts in the RD-Connect GPAP):**

   i. Where a publication only includes data and hypotheses from the PI's own research group, key authorship positions may be held by this group, but the software, tools and resources made use of for the research should be duly acknowledged and referenced in line with the policies for those resources (e.g. see RD-Connect GPAP policy below). Where justified, individuals supporting the bioinformatics analysis or platforms may be approached for co-authorship based on individual scientific contribution.

   ii. Where a publication has involved the use or analysis of data from additional submitters, these submitters should be contacted as soon as possible ahead of publication and invited to provide input as co-authors. The PI team is strongly encouraged to share key authorship positions with other teams that have brought in similar intellectual input and/or fundamental data (e.g. "a second family"). Acknowledgement of bioinformatics support should also be considered as in (i) above.

2. **A publication arising from the analysis of data where the party generating the hypothesis and carrying out the analysis is not themselves the data submitter (example: reanalysis of data by a Solve-RD bioinformatics group that did not submit the data or see the patients):**

   i. Submitters of the data used for the analysis should be contacted as soon as possible ahead of publication and invited to provide input as co-authors. The PI team is strongly encouraged to share key authorship positions with the submitting teams based on the value and amount of data contributed to the publication. If the primary data is the key to discovery, a key authorship position should be discussed with the owner of the primary data.

   ii. Where a publication makes use of data from a large number of submitters or transversal analysis of the Solve-RD cohort, a group authorship for Solve-RD (see Solve-RD Publication Policy) should be considered in order to acknowledge the role of all data submitters equally.

All data access through the RD-Connect Genome-Phenome Analysis Platform is monitored automatically by the system and all other data access for other Solve-RD activities is only to named individuals within the Solve-RD consortium, therefore any breach of the publication policy will be monitored and flagged up to the Solve-RD Steering Committee.

Solve-RD funding acknowledgement

Any publications arising from Solve-RD project funding should acknowledge it in the following way:

> "This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 779257 (Solve-RD)."

RD-Connect GPAP acknowledgement

In addition to authorship positions as described above, any publications that arise from the use of the RD-Connect Genome Phenome Analysis Platform should acknowledge it in the following way:

> "This study makes use of data shared/provided through RD-Connect, which received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 305444."

In addition, the following paper should be cited:

> Lochmüller H & Badowska D, Thompson R, Knoers N, Aartsma-Rus A, Gut I, Wood L, Harmuth T, Durudas A, Graessner H, Schaefer F & Rieß O. RD-Connect, NeurOmics and EURenOmics: Collaborative European Initiative for Rare Diseases. European Journal of Human Genetics. 2018.

ERN acknowledgement

Any publications with contributions from ERNs should acknowledge involved ERNs in the following way:

> "This study was supported by the European Reference Network(s) [add ERN names] (https://ec.europa.eu/health/ern/networks_en)."