Deliverable

D1.5 Guidelines for collection of experimental data	
Version	Status	V1	final
Work package	WP1
Lead beneficiary	CNAG-CRG (Sergi Beltran)
Due date	31.12.2018 (M12)
Date of preparation	18.12.2018
Target Dissemination Level	Public
Author(s)	Ricky Joshi (CNAG-CRG), Sergi Beltran (CNAG-CRG)
Reviewed by	Holm Graessner (EKUT), Alexander Hoischen (RUMC)
Approved by	Ana Rath (INSERM-Orphanet)

Abstract:

This report provides information on the guidelines that have been developed for experimental data collection for Solve-RD. These guidelines aim to aid and facilitate all data upload from all Solve-RD collaborators and ensure high quality standards are met.

Introduction:

Many Solve-RD partners, associated partners and networks will submit unsolved exomes and genomes as well as phenotype and clinical information from rare disease patients and their family members to the RD-Connect Genome-Phenome Analysis Platform (GPAP). We have prepared a number of guidelines with step-by-step instructions to facilitate this. This report deals exclusively with the collation of experimental data. Collation of phenotypic data is addressed in deliverable D1.4, and information on training modules and help provided for the latter is addressed in D1.3.

Report:

Genomics and other -omics data will be collected in a secure manner using RD-Connect GPAP's web-based submission platform. Once all phenotypic data has been submitted to PhenoTips (see Deliverable D1.3), the user may proceed to the experimental data and meta-data upload at https://platform.rd-connect.eu/datasubmission/. Here, the user will find a three-step workflow to submit essential meta-data related to the experimental data they wish to submit (Fig1). The three steps include 1) submission of participant set, 2) submission of experiment set and 3) submission of data files meta-data and data upload itself. The three steps are defined in more detail below.
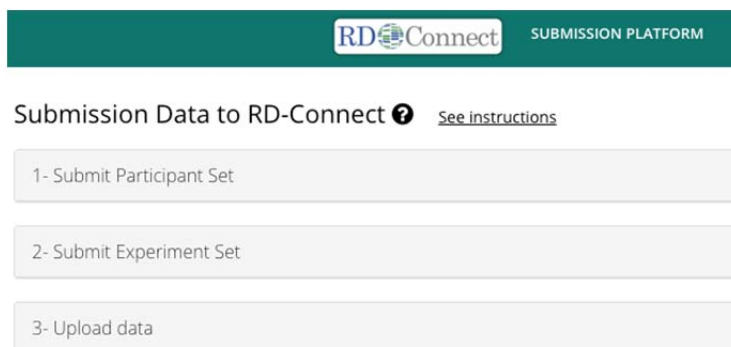

Fig1. RD connect submissions platform for meta-data and data upload and "See instructions" link.

The submission page has a "See instructions" link (Fig1) that leading to a page with the guidelines to submit this information, either through the interface or by uploading pre-filled Excel files. These guidelines are available at https://rd-connect.eu/data-submission-guide/ and a screenshot is provided in Fig2. In addition, we hold regular tele-conferences and webinars to help all users to correctly upload their meta-data and assists with any queries they may have (our last webinar was uploaded to YouTube and can be found here: https://www.youtube.com/watch?v=56omyLC0wQw&t=13s). We also have a help-desk email to provide case-by-case support at: help@rd-connect.eu.

1. Data Submission Platform 🔗

The **Data Submission Platform** can be found at https://platform.rd-connect.eu/datasubmission/

You will need a list of the PhenoTips IDs corresponding to the genomic data you wish to submit.

There are 3 steps to submitting genomic data to the platform:

1. Submit participant data (requires that PhenoTips entry is already complete)
2. Submit corresponding experiment data
3. Transfer genomic data files

Fig2. Screenshot of experimental data submission guidelines available online.

Data quality for both online and Excel submission is enforced at each step as mandatory data fields must be completed. These mandatory fields are highlighted by red asterisks (see Fig2, 3 and 4 for examples). The required meta-data also includes that required for the European Genome-Phenome Archive (EGA) in order to be able to submit this data to the EGA at a later stage as agreed in the Solve-RD Data Sharing Policy.

Step1- Submit Participant Set. At this step, the user will be required to declare the RD-Connect individual IDs for those individuals for which he/she intends to submit data. A summary of the information the user entered in PhenoTips (see deliverable D1.4) will be displayed (e.g. sex, mode of inheritance, consanguinity etc. (Fig2). There is also a checkbox to authorize data discovery from this patient through Matchmaker Exchange (www.matchmakerexchange.org), to which the RD-Connect GPAP is connected. Below, Fig3 shows an example of Step1-Submit Participant set user interface.

Fig3. Step1- Submit Participant set user interface. Mandatory fields are marked with a red asterisk.

Step 2- Submit Experiment Set. Users are required to upload basic meta-data regarding their sequencing experiments, including type of experiment, library information, capture kit used, as well as tissue and Biobank names (see Fig4). Users can indicate which project their data belongs to, and are allowed to select a data embargo of up to 6 months for datasets submitted to any project and up to 12 months for datasets submitted to the Solve-RD project (justification needs to be provided in the adjacent text box as explained in the Solve-RD Data Sharing Policy). In addition, each dataset can be associated to an ERN, which will facilitate both accounting in Solve-RD but also interaction with the ERNs between the Data Analysis Task Force and the Data Interpretation Forces. Below, Fig4 shows an example of Step 2- Submit Experiment Set user interface.

Fig4. Step 2- Submit Experiment Set user interface. Mandatory fields are marked with a red asterisk.

Step 3- Upload data. The third and final step requires to provide meta-data related to the files that will be actually uploaded to the RD-Connect GPAP. The meta-data fields include file type, sequencing strategy, read length, instrument model and sequencing center (see Fig5).

Fig5. Step 3- Upload data user interface. Mandatory fields are marked with a red asterisk.

Finally, once Step3- (Upload data) has been satisfactorily completed and the "submit" button clicked, users are sent, via email, the login credentials for the genomic data transfer, which is done through Aspera (Fig6). Once the transfer has been completed, data processing is initiated and the resulting data subsequently released to the RD-Connect GPAP where it is ready for analysis by Solve-RD partners and the Data Analysis Task Force.

Fig6. Aspera Client login page for genomic data transfer to CNAG for processing.

Conclusion:

The RD-Connect GPAP submission platform has been set up to serve the needs of the Solve-RD project, including assignment of datasets to the projects and association with an ERN. Furthermore, the system collects the required meta-data to submit to the EGA, at a later stage, the raw data collected as part of Solve-RD. Finally, we have developed guidelines and provide support to facilitate data submission to the system. Over 2,300 experimental data sets and the corresponding meta-data have already been submitted to the Solve-RD project.