Deliverable

D4.3 Central RD-Connect database serving Solve-RD, including user authentication and authorization	
Version | Status	V1 | final
Work package	WP4
Lead beneficiary	CNAG-CRG (Sergi Beltran)
Due date	30.06.2018 (M6)
Date of preparation	04.09.2018
Target Dissemination Level	Public
Author(s)	Ricky Joshi (CNAG-CRG), Sergi Beltran (CNAG-CRG)
Reviewed and approved by	Anthony Brookes (ULEIC)

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

Explanation according to GA Annex I:

Database and data sharing aspects of the project.

Abstract:

As a cornerstone for the success of Solve-RD, a critical objective is to maintain the strict dyad of guaranteeing data protection and privacy while facilitating data sharing. Solve-RD will employ the main RD-Connect database infrastructure (also known as RD-Connect Genome-Phenome Analysis Platform (GPAP)) to pool and enable controlled access to a large number of harmonised and integrated datasets from unsolved rare disease cases. RD-Connect participates in the main GA4GH and IRDiRC activities focused on facilitating sensitive data sharing to minimize siloed data limitations while ensuring privacy. Here we will describe the actions taken to ensure RD-Connect is serving Solve-RD, including the stringent steps for user authentication and authorization and how researchers can share their data.

Introduction:

The RD-Connect Genome-Phenome Analysis Platform (GPAP) is one of the flagship tools developed by FP7 RD-Connect project and a recognized resource by IRDiRC (International Rare Diseases Research Consortium). RD-Connect GPAP aims to cover the following needs: 1) Enable standardised collection, integration, storage and reuse of linked genomic and phenotypic data and metadata using widely used languages and ontologies. 2) Provide an infrastructure to enable researchers to securely store, share and re-use linked raw and processed genomic and phenotypic data from individuals with rare diseases (RD) and relatives. 3) Standardise variant calling and annotation to facilitate integration and comparability. 4) Provide a powerful and user-friendly analysis interface that enables researchers to analyse and interpret the full genomic datasets they submit for both diagnosis and gene discovery on an individual patient basis. 5) Enable researchers to find “matching” patients in other databases worldwide. 6) Enhance discovery of new genes, pathways and therapeutic targets through integrated bioinformatics tools. 7) Link genomic and phenotypic datasets to RD biobanks and registries though RD-Connect catalogues.

The user-friendly online system allows users to analyse and query their own data as well as data submitted by others that has been made accessible to authorised users. The platform has supported the discovery of several new RD genes and phenotypes.

The RD-Connect genome-phenome analysis platform (https://platform.rd-connect.eu) is unique in its combination of scope and functionality. While some commercial platforms allow analysis of NGS data either for a fee or attached to the cost of the sequencing itself, they lack the crucial association to standardised clinical information, data discoverability and sharing and reuse features of RD-Connect. Publicly funded systems such as the European Genome-phenome Archive provide long-term storage and sharing of raw NGS data but act primarily as file-stores that lack the associated deep phenotype information, functionality analysis and data integration provided by RD-Connect. In RD-Connect, an instance of PhenoTips allows user-friendly collation and accession to individual profiles coded with the Human Phenotype Ontology, OMIM and ORDO. The corresponding genomic data (genomes, exomes or panels) are analysed through a standard analysis pipeline and made accessible in the system. The genetic nomenclature follows HGVS recommendations. HGVS, HPO, OMIM and ORDO are IRDiRC recognised resources.

The unique strategy begins with submission of the raw .bam or .fastq files, which is essential in order to allow data from multiple sequencing providers to be processed through a standard pipeline to ensure comparability. The raw data and metadata are stored for long-term access at the European Genome-phenome Archive (EGA), a secure, controlled-access repository, while the processed data and metadata are made accessible online for real-time analysis in the RD-Connect genome-phenome analysis interface. The standard analysis pipeline for aligning, variant calling and annotating raw exome and genome data is applied to all incoming data and the fully integrated results are made available to authorised users through the online interface.

Report:

Solve-RD will leverage the fully operational RD-connect platform for integrating genomic information and standardised phenotypic information from large patient cohorts for gene discovery through its advanced bioinformatics analysis and annotation pipeline. Safe, secure management of Solve-RD data sharing policy is paramount. Solve-RD 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 acceptance. Additionally, the PI/group leads must sign the Adherence Agreement (https://rdconnect.eu/gpap-adherence). The RD-Connect Code of Conduct and procedures were modified in May 2018 to comply with the EU General Data Protection Regulation.

Data security

Data is stored in a computer cluster with a restricted access policy, limited internet access and daily backups. Databases are using distributed file systems, 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 adapted for the new General Data Protection Regulation, GDPR (Regulation (EU) 2016/679).

Online registration, user authorization and authentication

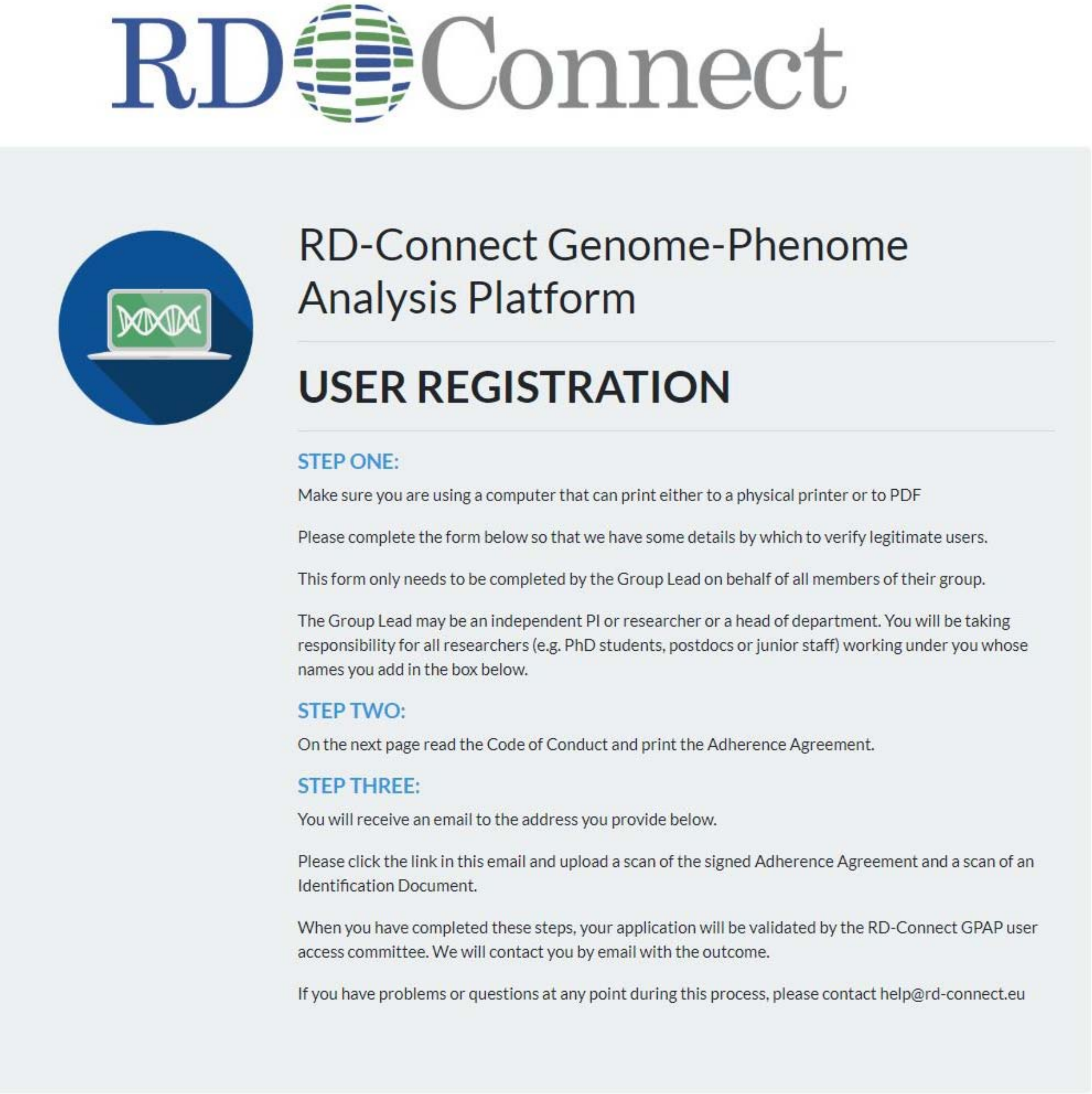
All collaborators of Solve-RD must fill out an online registration form (see Figure 1). The online registration consists of three steps and must be filled out by the group lead on behalf of all members of their group.

Step 1: Complete online registration form indicating group information and participation in Solve-RD (see Figure 2). The group lead will take responsibility for all researchers assigned to their group.

Step 2: Read and accept the Code of Conduct and print the Adherence Agreement.

Step 3: Upload a scan of the signed Adherence Agreement and a scan of an Identification Document.

Finally, the registration will be validated by the RD-Connect GPAP user access committee to ensure authenticity.



RD-Connect Genome-Phenome Analysis Platform

USER REGISTRATION

STEP ONE:

Make sure you are using a computer that can print either to a physical printer or to PDF

Please complete the form below so that we have some details by which to verify legitimate users.

This form only needs to be completed by the Group Lead on behalf of all members of their group.

The Group Lead may be an independent PI or researcher or a head of department. You will be taking responsibility for all researchers (e.g. PhD students, postdocs or junior staff) working under you whose names you add in the box below.

STEP TWO:

On the next page read the Code of Conduct and print the Adherence Agreement.

STEP THREE:

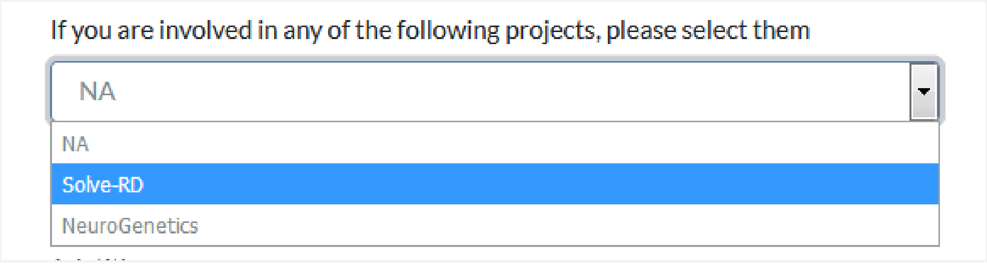
You will receive an email to the address you provide below.

Please click the link in this email and upload a scan of the signed Adherence Agreement and a scan of an Identification Document.

When you have completed these steps, your application will be validated by the RD-Connect GPAP user access committee. We will contact you by email with the outcome.

If you have problems or questions at any point during this process, please contact help@rd-connect.eu

Figure 1: A screen shot of the RD-Connect Genome-Phenome Analysis Platform (GPAP) user registration interface.



If you are involved in any of the following projects, please select them

NA

Figure 2: Users can specify they will submit data to the Solve-RD project. EKUT will confirm the group is a Solve-RD beneficiary or has signed the proper collaboration agreement with Solve-RD.

Account registration: Every PI/group lead enrolled in Solve-RD will undergo the full Solve-RD 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.

Authentication and Authorization: Every PI/group lead and its enrolled members will be registered in the RD-Connect Central Authentication system (CAS) and their account will be activated for data access once they read, understand and accept online the RD-Connect Code of Conduct.

Central RD-Connect GPAP for data management

Data upload: All exome/genome and phenotypic data will be submitted to the RD-Connect Genome-Phenome Analysis Platform, an IRDiRC recognised resource. 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. An option to assign a new or existing dataset to a specific European Reference Network (ERN) has been made available to keep track of submitted datasets by ERN. 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. When uploading fewer than 100 raw datasets, the standard RD-Connect GPAP upload interface can be used, which includes user friendly PhenoTips templates to enter the phenotypic information and user-friendly tables to upload the genomic data and metadata. For more than 100 raw datasets, a customised bulk upload option is available in conjunction with CNAG-CRG in Barcelona, which has developed an Excel template in collaboration with EKUT.

Solve-RD tagging of datasets: A system to tag datasets to multiple projects (including Solve-RD) has been implemented for the project. 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 as dual-tagged (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.

Data embargo: Data is 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. A data embargo may be enforced for other users of RD-Connect GPAP that are not part of Solve-RD. For an embargo period of up to six months, no additional information will be required but for an embargo period between six and twelve months, Solve-RD users will be asked to type a short justification that will be reviewed by Solve-RD Steering Committee. Embargo periods longer than twelve months can also be requested (in written form at the time of submission) but will require acceptance by the Solve-RD Steering Committee and by the RD-Connect GPAP Data Access Committee since they are considered non-compliant with IRDiRC principles for rapid data release.

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.

Activity logs

To increase transparency, the dataset submitter is able to see within the RD-Connect GPAP data management portal which other users have accessed their data and at which date, through a specific query on that dataset or through a general query. 882 experiments are already available as part of Solve-RD in the RD-Connect GPAP.

Conclusion:

The RD-Connect GPAP has been adapted to serve the needs of the Solve-RD project and has already registered users for the project, which have uploaded new phenotypic and genomic data or dual-tagged already existing data.