Deliverable

D4.8 Complete Solve-RD bioinformatics platform operational	
Version | Status	V2 | final
Work package	WP4
Lead beneficiary	CNAG-CRG
Due date	31.12.2021 (M48)
Date of preparation	21.12.2023
Target Dissemination Level	Public
Author(s)	Sergi Beltran (CNAG-CRG), Lennart Johansson (UMCG), Steven Laurie (CNAG-CRG), Luca Zalatnai (CNAG-CRG), Morris Swertz (UMCG), Mallory Freeberg (EMBL-EBI), Davide Piscia (CNAG-CRG), Anthony Brookes (ULEIC)
Reviewed by	Peter Robinson (JAX), Birte Zurek (EKUT)
Approved by	Holm Graessner (EKUT)

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:
Provide complete Solve-RD bioinformatics platform.

Abstract:

As a cornerstone for the success of Solve-RD in terms of solving unsolved rare disease cases, a critical objective is to reuse, enhance and deploy existing solutions for core analytics support, databasing, data discovery, and data sharing. Here we describe the overall Solve-RD data workflow, and how the different components of the Solve-RD bioinformatics platform connect to facilitate the process of solving unsolved cases. The workflow includes data submission, management, analysis and interpretation processes in the RD-Connect Genome-Phenome Analysis Platform (GPAP); raw and processed data archiving at the European Genome-phenome Archive (EGA, Hinxton, UK); data analysis and directory structures; and the functionalities of the Rare Disease Data about Data database (RD3) and Discovery Nexus.

Introduction:

Solve-RD aims to solve unsolved rare disease cases. To this end, data from ~22,000 Whole Exome Sequencing (WES) or Whole Genome Sequencing (WGS) experiments performed in genetic diagnostics labs across Europe, together with corresponding phenotypic data from participants have been collected by the European Reference Networks (ERNs) participating in Solve-RD with the goal of jointly analysing these samples within the Data Analysis Task Force (DATF) and the Data Interpretation Task Forces (DITF). In addition around, 3,500 WGS and over 3,500 other -omics experiments will be performed for selected patients, producing new data (Zurek et al., 2021). These numbers make it possible to create cohorts of patients with the same disease or similar phenotypes. However, collecting the data and enabling large-scale analysis is a challenge, further compounded by the fact that as a result of collecting samples from so many different sources the quality and type of data is very heterogeneous, as each laboratory has differences in their procedures and samples have been processed during different time periods, resulting in procedures (e.g. an updated enrichment kit) changing even within the same laboratory.

To overcome these challenges the Solve-RD project has created an ecosystem to receive data and metadata from external centres, perform standardised initial processing for genomic data, centrally store the data, make data available to the researchers at their respective research environments, allow for data analysis, track samples and metadata and allow for flexible querying to create cohorts of samples based on similarity in phenotypes or genetic variants. Here we describe the components of this ecosystem, their implementation, and how these features contribute to the ambitious goal of solving unsolved rare disease cases.

The work described in this document has also been described in the paper “A unified data infrastructure to support large-scale rare disease research” published on MedRXiV: https://www.medrxiv.org/content/10.1101/2023.12.20.23299950v1 (Johansson et al., 2023).

Report:

Overall Solve-RD data workflow

Focusing on the WES and WGS data, the overall data workflow follows the structure shown in Figure 1 of the “Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases” paper (Zurek et al., 2021).

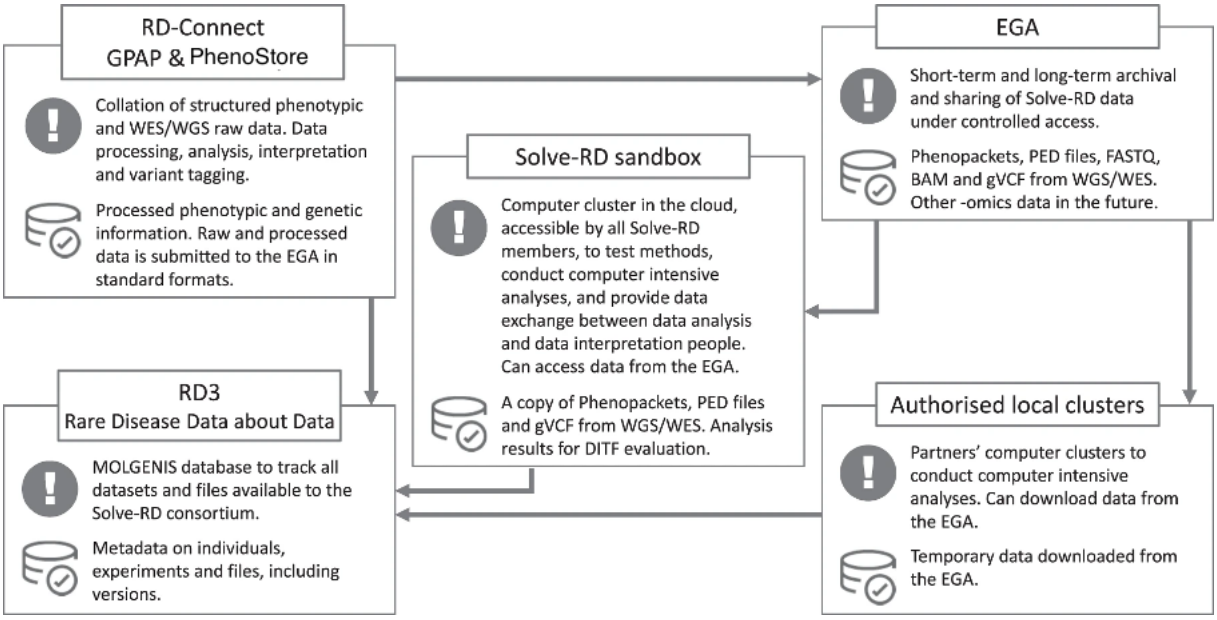

Figure 1: Solve-RD data infrastructure (from Zurek et al., 2021)

In short, experimental metadata is first submitted to the RD-Connect GPAP, and corresponding phenotypic data to PhenoStore, where patient, phenotypic and family information are stored. Sequence data is submitted as FASTQ, BAM, or CRAM files to CNAG-CRG via a RedIris Aspera server and processed with a standard alignment and variant calling pipeline (Laurie et al, 2016) with the aim of homogenising results and facilitating systematic analysis, interpretation, and comparisons. Processed data, in the form of BAM and gVCFs per experiments, is transferred to the European Genome-phenome Archive (EGA) where it is archived (note that for novel -omics samples the direction is reversed with FASTQs being submitted to the EGA by the data providers, and then distributed to the relevant centre for analysis). At the EGA a unique file identifier is added to each individual file and data made available for download to partners' local clusters and the two Solve-RD sandboxes for analysis. In parallel the Solve-RD Rare Disease Data about Data (RD3) database collects metadata on subjects, samples, experiments and files from these sources. The data is structured into freezes and patches. The freezes consist of data from subjects and experiments that have been submitted prior to one of three deadlines, meaning that each freeze consists of a fixed number of experiments and participants. Changes in data or metadata for these subjects are captured in patches, leaving the original dataset on which analyses have been performed intact, making reanalysis possible.

RD-Connect Genome-Phenome Analysis Platform (GPAP) and PhenoStore

Architecture of the RD-Connect GPAP

The RD-Connect GPAP variant storage and processing system has been built on fully scalable technologies. The core functionalities of the platform are provided by three distinct modules which allows the different data to be integrated: *Data Management*, *PhenoStore*, and *Genomic Analysis*, each of which consists of at least one server and one client, all being connected to the same Identity Provider (Keycloak, https://www.keycloak.org/). The *Data Management* module facilitates the uploading of raw sequencing data, together with associated metadata required for processing. *PhenoStore* facilitates the submission, management and sharing of phenotypic descriptions of all participants. The *Genomic Analysis* module enables researchers/clinicians to analyse their own cases, and can discover and access data from cases shared by other researchers, in order to identify disease-causing variants.

In addition to these core modules, other microservice modules have been developed, such as those providing MME and Beacon V1 (Fiume et al., 2019) services.

Data management

A requirement for collaboration in the RD-Connect GPAP is that all submitted genomic data be accompanied by a corresponding pseudonymised phenotypic description of the individual from whom the sample was taken, which is captured within the PhenoStore module, developed primarily within the EJP-RD project. A deep phenotypic record for affected individuals can be generated using HPO (Human Phenotype Ontology) terms. Where the RD diagnosis is known at the time of submission, or upon case resolution, an ORDO (Orphanet Rare Disease Ontology) and/or OMIM (Online Mendelian Inheritance in Man) identifier can be added to the record, together with a description of the causative variant(s).

The *Data Management* module facilitates the upload of data and metadata and allows users to keep track of all the data they have submitted. With this module they can easily check for how many participants they have entered data into the system, which types of experiments are linked to them, and which associated files were uploaded.

Data submission is undertaken in three steps. First, the user creates a phenotypic record in the bespoke application *PhenoStore*, describing the case and any family members for which genomic data is available. In the second step the submitter must provide a small amount of essential metadata to allow data processing to be undertaken correctly. This metadata describes the type of experiment performed *e.g.* genomic library preparation and sequencing strategy, and to which participant the data belongs. These first two steps can either be performed directly within a user-friendly graphical user interface (GUI) or via bulk upload using MS Excel templates. In the final step, high-speed transfer of the raw sequencing data, in standard formats (FASTQ, BAM or CRAM) is facilitated via use of an Aspera server, provided by the Spanish academic and research network RedIris (https://www.rediris.es/).

Interoperability: data and communication standards

To enable interoperability and overcome language barriers, the RD-Connect GPAP has been designed to use widely adopted and machine readable international and community standards and ontologies whenever possible. Within the PhenoStore module, patient descriptions are recorded using HPO, ORDO and OMIM (Amberger et al., 2015) terminology, and phenotypic records can be exported to the GA4GH approved phenopackets file format (https://github.com/phenopackets/phenopacket-schema), and family trees in PLINK PED format (Purcell et al., 2007). Genomic alignments and variants are stored and transferred (*e.g.* to the EGA) in GA4GH approved BAM, CRAM, and VCF formats, respectively (https://www.ga4gh.org/genomic-data-toolkit/). Biological annotations, available in the Data Analysis module are provided by Ensembl VEP (McLaren et al., 2016), and supplemented with data from other genomics community resources such as ClinVar (Landrum et al., 2018), gnomAD (Karczewski et al., 2020), and PanelApp (Martin et al., 2019). Data discovery and sharing is achieved through the implementation of GA4GH Beacon V1 and MME APIs (Buske et al., 2015).

Data analysis and interpretation functionalities

The *Genomic Analysis* module is the analytical core of the RD-Connect GPAP, where researchers and clinicians analyse their own RD cases in isolation or in combination with other data shared with them. The primary use case for the analysis is to discover disease-causing variants in a submitted index case or family structure, and the system supports variant analysis in both a "diagnostic" paradigm (known variants or variants in known genes) and a "discovery" paradigm (potentially damaging variants in genes not previously associated with disease).

To begin an analysis, the user selects the samples they want to analyse and defines the mode of inheritance they wish to investigate. Following this, a range of filters can be applied. The user can add or remove filtering terms at will, and rerun their query, which will return updated results in seconds. Summary results are displayed with colour-coding to highlight variants likely to be of particular interest, *e.g.* variants recorded to be Likely Pathogenic or Pathogenic in ClinVar, and/or those predicted to be damaging by tools such as SIFT (Kumar et al, 2009) and PolyPhen2 (Adzhubei et al, 2013). Furthermore, the RD-Connect GPAP provides fully detailed annotations for each variant in an extended results section.

In addition to the more standard features, the RD-Connect GPAP uniquely brings together information from a wide variety of sources which facilitate variant filtration, prioritisation and interpretation, beyond that provided by simple biological annotation. These include the ability to generate candidate gene lists on-the-fly via API connections to resources such as Genomics England PanelApp, HPO, and DisGeNET. Tracking of the status of ongoing analysis of specific cases is also enabled, including time-stamping, username recording, and the outcome of the analysis (*e.g.* solved case, negative case, variant under segregation analysis *etc.*).

Query results can be saved through the generation of unique URLs, thus allowing users to return to their analyses at any point in the future, and share results easily with other authorised users in the system.

Variants of interest can be tagged and user interpretation applied, which is then visible to other users. Furthermore, when a user confirms a variant as disease causing within the platform, relevant information about it can be filled in by the user; the information requested is aligned with what ClinVar requires.

Data discovery

Users in the RD-Connect GPAP can discover data internally by searching across all GPAP participants, by searching within the GPAP cohorts, and by internal matchmaking.

The search-across-all feature allows users who have identified a variant of interest in a certain gene to search across all genomic datasets available to them within the RD-Connect GPAP to attempt to discover participants with similar or identical variants. If they identify such a variant, a contact button allows them to contact the submitter of the relevant participants to discuss the candidate variants further.

A combinatorial query on the participants' information (ORDO, HPO, assigned ERN) allows *in-silico* cohorts of experiments to be defined to which the search-across-all feature can be applied to focus the query.

There is also an opportunity for external data discovery through the Matchmaker Exchange (MME) API and through Beacon v1 API (Fiume et al., 2019). The MME API is implemented for internal matchmaking (between datasets available in the GPAP) and also externally, with the MME platform (Philippakis et al., 2015), via the connection to other nodes. Furthermore, the RD-Connect GPAP is included in the Beacon Network (https://beacon-network.org/) via the implementation of the GA4GH Beacon API v1, which enables queries to determine if a particular variant is found in at least one of the experiments in the RD-Connect GPAP. Likewise, any interesting variant found in the RD-Connect GPAP can be immediately queried in other Beacons through the Beacon Network. More information about matchmaking strategies are reported in more detail in Solve-RD deliverables D2.5 and D2.20.

European Genome-phenome Archive (EGA): raw data archiving

To ensure availability of the data during the project and beyond, processed data and metadata is archived at the European Genome-phenome Archive (EGA; https://ega-archive.org/), which is a service for permanent archiving and sharing of identifiable genetic and phenotypic data (Freeberg et al., 2021). The data is uploaded to a submission box and, after curation, archived

and released in a dataset. Metadata and data files are assigned EGA accessions, functioning as a unique identifier (UID). File metadata is submitted using a manifest file which is processed by EGA to directly link files to the correct datasets and ensure files are not corrupted during processing. Novel -omics raw data and file metadata are uploaded directly from the data provider to the EGA using the same protocols, and subsequently archived and released.

After the data are successfully archived and released, the EGA provides access to the data only upon approval by the associated Data Access Committee (DAC) for specified individuals. The EGA currently provides Solve-RD data access to Solve-RD project partners to support data analysis and interpretation goals, and will enable access in the future to any approved researchers.

Solve-RD Sandbox: shared HPC compute environment

Sandbox instances and deployment

In addition to local downloads of the data, to enable researchers to jointly analyse the available data and metadata, shared High Performance Computing (HPC) virtual research environments (VRE), called 'sandboxes' have been created. The project made use of two VRE, the first using EMBL-EBI's private and secure Embassy Cloud, based in Hinxton, UK (http://www.embassycloud.org/) and the second at the UMCG in Groningen, NL (http://docs.gcc.rug.nl/gearshift/). The EMBASSY VRE is only accessible by members of the Solve-RD project, while the UMCG VRE is shared with other projects. Here, a dedicated Solve-RD group is present with access restricted to Solve-RD members. Generalisation of the VRE is being performed within the EJP-RD project.

The EMBASSY VRE has a storage of 30 Tb and 12 compute nodes with 14 cores/node and 56,072 Mb RAM/node. The UMCG VRE has a shared storage with other projects, of which 200 Tb is in use by the Solver-RD project. 10 compute nodes are available with 22 cores/node and 205,490 Mb RAM/node. Both systems are deployed using the same playbook (https://github.com/rug-cit-hpc/league-of-robots), based upon Linux CentOS7 (https://www.centos.org/) with Spacewalk (https://spacewalkproject.github.io/) for package distribution and management. Openstack is used for virtualization. In addition, tools and pipelines are deployed using a second repository (https://github.com/molgenis/ansible-pipelines) making use of the LMOD module system (https://github.com/TACC/Lmod) and Easybuild (https://github.com/easybuilders/easybuild).

Both clusters are in internal networks that are not directly accessible from the internet. Access is possible via dedicated jumphosts, security hardened machines not involved in any data storage or processing. Using asymmetric cryptography via a private-public key pair (http://docs.gcc.rug.nl/fender/accounts/), users can login to the jumphost to be directly redirected to the main HPC cluster.

Data analysis and directory structure

Within the Solver-RD project data analysis is performed within the DATF, while interpretation of variants is done by the DITFs. Within the DATF, analyses are divided over several Working Groups (WGs): six focusing on the reanalysis of short-read WES data and novel -omics short-read WGS data: SNV-indel, CNV, regions of homozygosity and relatedness, *de novo* variant analysis, meta-analysis and short-tandem repeat analysis with additional working groups focusing on long-read genomics and other -omics data such as transcriptomics and proteomics.

Directory structures on both VRE are identical, with a division in the main Solver-RD group folder into a temporary (*tmp*) and a permanent (*prm*) storage folder. Data stored in the *tmp* folder can be used for analysis. Folder structures within tmp are managed by the users and could be either a user-specific folder or a DATF working group (WG) folder, allowing for pilot

analyses or analyses of patient cohorts. Scripts are managed using the SLURM workload manager (https://slurm.schedmd.com/).

The *prm* directory contains a directory for each DATF WG. Each WG has appointed a data manager who is allowed to copy, move and remove data to and from the *prm* folder on the EMBASSY VRE. Within each DATF WG folder, a subfolder was created for each of the ERNs as well as an *all_erns* folder and a *sandbox_only* folder. Results to be shared with a specific ERN are copied into the respective directory and data of interest for all ERNs (e.g. general coverage statistics) are copied to the *all_erns* folder. The *sandbox_only* folder is meant for large (intermediate) files that should be archived, for instance CNV tool count statistics that are costly to calculate. An extra folder called *sftp* was created containing two folders: DATF and DITF, with symlinks to all DATF WG folders and subfolders, except the *sandbox_only* folder. The first we structured per DATF WG containing subfolders for each DITF, the second was structured the other way around, one main folder per DITF, each with subfolders of the DATF WGs. A scheme of the structure described above is shown in Figure 2.

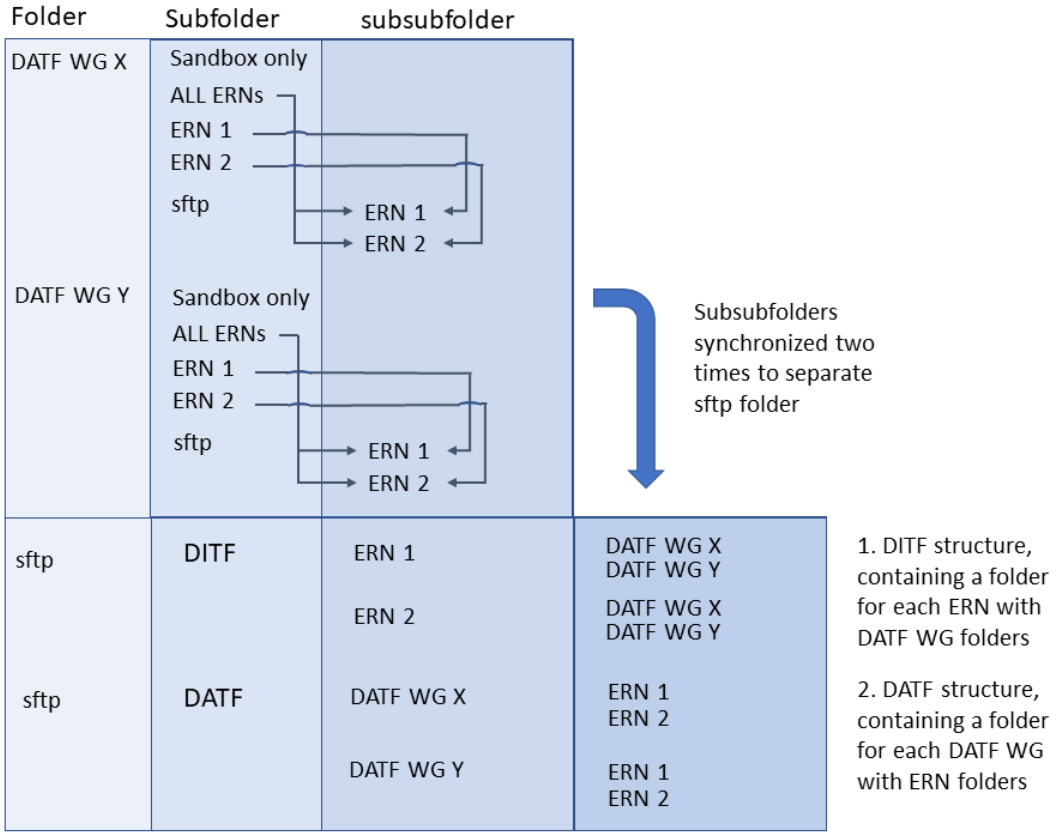

Figure 2: Sandbox folder structure. Data is organised by the DATF working groups (DATF WG) either in folders per ERN or in a common folder (for data intended for all ERNs). Additionally, large files that should be kept but not shared are stored in a 'sandbox only' folder. All data to be shared with the ERNs is linked to in an sftp folder with a subfolder per ERN. The thin arrows indicate which folder is linked to where. These folders are further synchronised to two folders: DATF and DITF, each with the same information (indicated by the thick arrow). The DATF folder has the same structure as the initial sftp folder (a folder for each DATF WG with subfolders per ERN) and the DITF folder has the turned around structure (a folder for each DITF ERN with subfolders per working group). This structure makes it easy for both DATF and DITF to browse the data (e.g. all CNV data or all data from ERN ITHACA).

In addition to these folders an *ega-fuse-client* (https://github.com/EGA-archive/ega-fuse-client) folder is present in the prm folder, giving direct access to the Solve-RD datasets archived at the EGA. This connection to the EGA was set up within the CINECA project, benefiting also the Solve-RD project.

All data in the EMBASSY VRE *prm* folder are synchronised to the UMCG VRE prm folder, resulting in a backup of the data as well as the possibility to access the data from both sites.

Transfer server

To allow non-bioinformaticians easy access to the DATF results. The data in VRE *prm*/*sftp* folders is copied to a download server that can be accessed using a private-public keypair, but without the extra security of a jumphost. The sftp server can be accessed using freely available software such as WinSCP (https://winscp.net), MobaXTerm (https://mobaxterm.mobatek.net) or Cyberduck (https://cyberduck.io).

A general overview of the connections between the EGA, the two VRE instances and the transfer server can be seen in Figure 3. This figure also shows the most important workspaces and file locations.

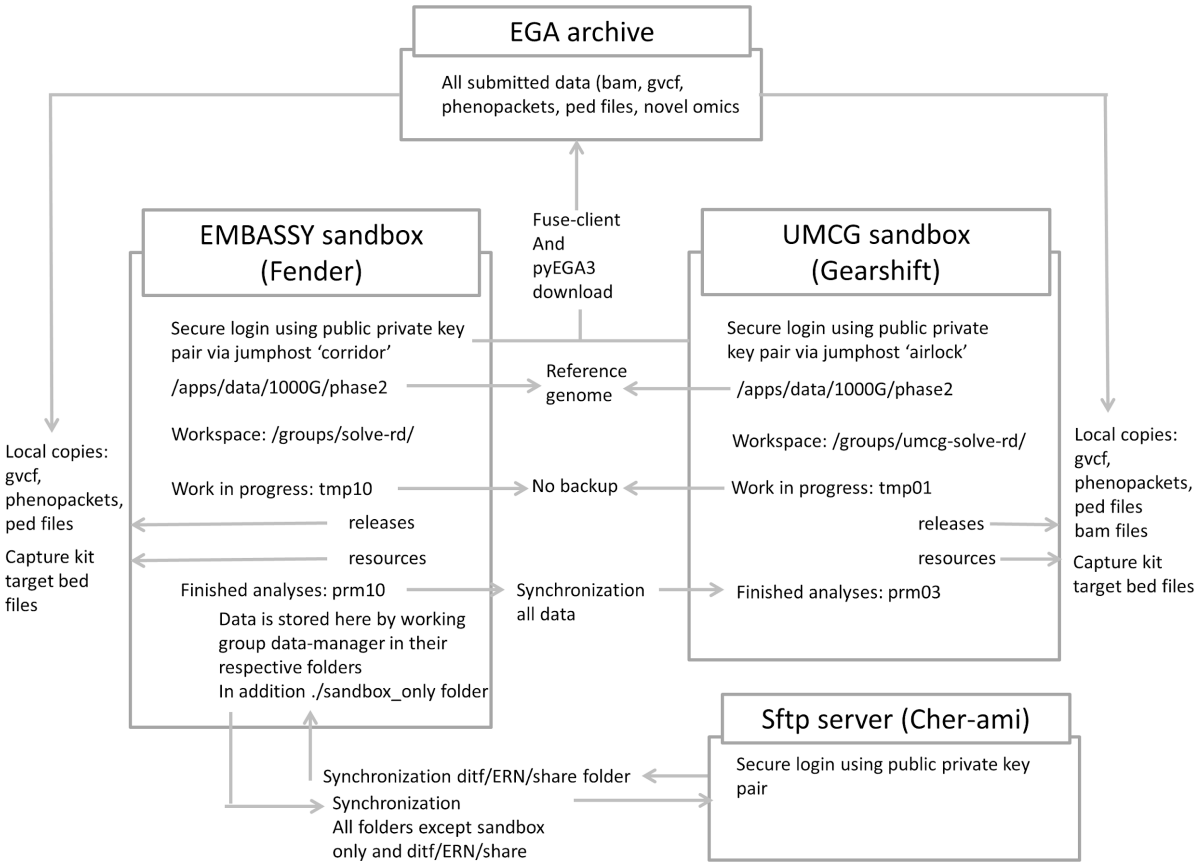

Figure 3: EGA and sandbox synchronisation structure, file locations and workspaces. The EMBASSY and UMCG sandbox have identical structures, split in temporary and permanent storage and, for all releases, local copies of the same files are present (initially also bam files were present on the UMCG sandbox). In addition, all data from the EGA is accessible on both sandboxes via Fuse-client and pyEGA3 download. Both sandboxes have key files available, such as bed files from the enrichment kits and the reference genome used for alignment in the Solve-RD project. Temporary storages are not backed up. Storing data on the permanent storage is only possible in the EMBASSY sandbox, which is synchronised to the UMCG sandbox to serve as a backup. Selected folders as shown in figure 2 are further synchronised to a transfer sftp server. A selected folder 'share' allows to store and share data directly on the transfer sftp server and is synchronised to the EMBASSY sandbox.

Data structure: freezes and patches

Solver-RD project releases the data collected for re-analysis in freezes, each containing data submitted before one of three deadlines. At the time of writing, two freezes have been released, the first released in early 2020, including data from 8,393 participants, and the second released in mid 2021, including data from 3,242 participants. To allow for changing data, such as the addition of new phenotypic information on a participant, or the correction of an error, patches may be applied to the freezes. Patched files are released with a date inserted between filename and extension (FILENAME.YYYY-MM-DD.extension). Patch data is archived at the EGA in a separate dataset. Within the VRE a freeze and patch structure is set up with a single master folder containing original files as well as patched files. A directory is created per freeze and patch with symlinks to the files included in the specific patch release, typically a mix with the majority of files included in the previous patch and some new changed files.

Authorised local clusters

Some of the pipelines needed for omics data analysis are very computer intensive, and it was not reasonable to expand the Sandbox up to the specifications that would be required for a short period of time. Therefore, Solve-RD relies on local clusters from partners for conducting certain tasks, such as mapping of sequencing reads to the reference genome and calling genetic variants. Up to date, the consortium has approved three clusters onto which Solve-RD data can be downloaded and analysed. These are the clusters from the following partners: EKUT (Tübingen, Germany), RUMC (Nijmegen, The Netherlands) and CNAG-CRG (Barcelona, Spain). Data can also be downloaded to the UMCG (Gröningen, The Netherlands) cluster from within the Sandbox.

RD3: Rare Disease Data about Data database

The Solve-RD Rare Disease Data about Data (RD3) database is the *spider in the web,* containing information on subjects, samples, experiments and files. The database is created in MOLGENIS (Swertz et al., 2010) accessible via a web-interface (https://Solve-RD.gcc.rug.nl/) and via local login or FusionAuth single-sign on login (https://fusionauth.io/), enabling direct usage of the *Discovery Nexus* search capabilities described in the next section. Generalisation and further FAIRification of RD3 is being done within the EJP-RD project. Within RD3 all relevant metadata that is needed for research is collected. The database is divided in separate sections. Firstly, the WES/WGS reanalysis data, split in a section per freeze and patch. Each of these sections have the same format.

The **subjects** table contains information on the participants as collected by the RD-Connect GPAP PhenoStore. It is imported via phenopackets and PED files archived at the EGA. The subjects are identified based on their P-ID. For each subject the P-IDs of the parents are given if included in the project as well as the family number to identify all subjects part of the same family. Furthermore, the sex as provided by the submitter and the disease and/or the phenotypes known to be present or absent in the subject are listed. For each subject it is also recorded if they are considered to be affected by a condition or not (e.g. a child is affected and both parents are unaffected by a condition). Information on who has submitted the case, if they are allowed to be recontacted in case of incidental findings or if the case is retracted is also stored. Finally, the subjects table shows if the case is solved. Because this information is actually updated by the clinicians or clinical researchers in the RD-Connect GPAP PhenoStore, a connection between the two applications allows the solved status to be updated daily.

From each subject, zero or more samples may be derived. Sample metadata is collected in the **samples** table. Each sample is given a sample-ID (S-ID) for unique identification. For every S-ID, the P-ID of the subject from which it is derived is shown as well as the tissue type (e.g. whole blood) from which the sample was taken.

On each sample, zero or more experiments can be performed (e.g. WES or WGS on DNA isolated from the sample). Information on these experiments is collected in the **experiments** table. Each type of experiment has its own specific lay-out. For WES, the enrichment kit used is captured as well as the sample preparation method. In addition the metrics "% of the target covered >20x" and "average target coverage" are collected. Data tabs for other -omics data types are still under development.

For each family, subject and experiment files are archived at the EGA. RD3 captures this information in the **files** table. Here, for each file the path in the VRE *ega-fuse-client* within the dataset is given with its checksum information enabling a sanity check on copies of this file. In addition, information is given on the filetype, the experiment it belongs to and the corresponding EGA accession number.

Somewhat hidden in RD3 are the portal tables. These tables enable import of data from outside RD3, such as the solved status. Multiple portal tables are present to enable import of information on novel -omics, for instance using a manifest file.

Discovery Nexus

Whereas all metadata is stored in RD3 and simple searches can be performed, more complex queries to build meaningful cohorts cannot be performed. Therefore, the Discovery Nexus has been developed, powered by Cafe Variome 2 (Lancaster et al:, 2015, and https://github.com/Cafe-Variome/CafeVariome). From a technical point of view Discovery Nexus is a self-contained software package and runs on a server separate to RD3. However, from a user perspective it seamlessly functions as part of RD3, including the single sign-on described earlier. Discovery Nexus can be accessed directly from the RD3 main menu where an interface appears (see Figure 4), consisting of five sections that together enable a user to create a query of the Solve-RD data.

The first section focuses on the **subject** and provides the option to filter by gender or family type (e.g. singletons, trios). Optionally, only subjects considered to be affected by a condition of interest can be selected. Note that the filters stack onto each other and that if you select *Trio* and *Affected*, not the entire trio is returned, but only the affected family members (e.g. only the child).

The second section focuses on the subject **phenotypes**. Here, HPO terms of interest can be selected. By default, the setting is applied strictly: all selected HPO terms have to match exactly to the subject's phenotype. By default the setting is applied strictly: all selected HPO terms themselves, or their children, have to match exactly to the subject's phenotype. However, the setting can be set more leniently in two different ways. Firstly, it can be determined that not only the exact match to the HPO term is returned, but also similar terms as determined using established similarity measurements such as Resnick, as summarised by Deng et al., 2015, and in particular the Rel measure (Schlicker et al., 2006). The underlying method used for this utilises pre-calculated similarity measurements represented on a phenotype/subject graph allowing near instantaneous searching of a subject based on similar term or set of terms. In the interface the user can intuitively control this setting by setting a slider somewhere between 'minimum' and 'exact'. Secondly the HPO search can be set to have only a subset of the selected terms match the subject. In this way a broad search can be performed in which many phenotypes are allowed, but not all have to be present. Here also a slider allows the user to have 'any' to 'all' HPO terms be required to return a sample.

The third section uses the subject's **disease** information, coded in ORDO and OMIM. As before, by default the setting is applied strictly: only subjects with the exact disease (or child disease term in the ontology) will be matched. This can be applied more leniently by utilising the HPO-ORDO Ontological Module from Orphanet (HOOM http://www.orphadata.org/). One slider ('ORDO Match Scale') allows subjects to be matched based on overlap of HOOM mapped HPO terms between diseases. The 'HPO Term Pairwise Similarity' slider expands this overlap matching to include similar HPO terms as described in the phenotypes section above.

In addition, both the phenotypes and disease sections can use HOOM mappings to include subjects that are either connected to ORDO or HPO codes respectively in each section.

The fourth section allows for creation of a **cohort** with certain **types of variants in selected genes** or **pathways**. If in an experiment a variant of a certain type has been found in a specifically selected gene, or in a gene part of a certain pathway, the subject on whose sample this experiment has been performed will be returned. Optionally, a maximum allele frequency can be filled in, meaning that only variants with a frequency below the set value will be taken into account.

In the fifth section, to prevent incidental findings, the **ERNs** that the user wants to query must be specifically selected. For instance, if a search is performed to include patients with a loss-of-stop variant in a cancer related gene, you may only want to return subjects submitted by ERN-GENTURIS.

Once the appropriate sections of interest are filled in, a query can be executed, firstly returning the count of subjects in the different selected ERNs matching the query. Secondly, for each ERN the list of entries appears. From this list the subjects of interest can be selected and full data can be obtained through linking back to RD3 and automatically setting a data item filter for these subjects.

From a data management perspective, the data in RD3 lies behind the Discovery Nexus Query for all fields except the one to query variants. For this field the annotated variants produced by the SNV-Indel working group are used as input.

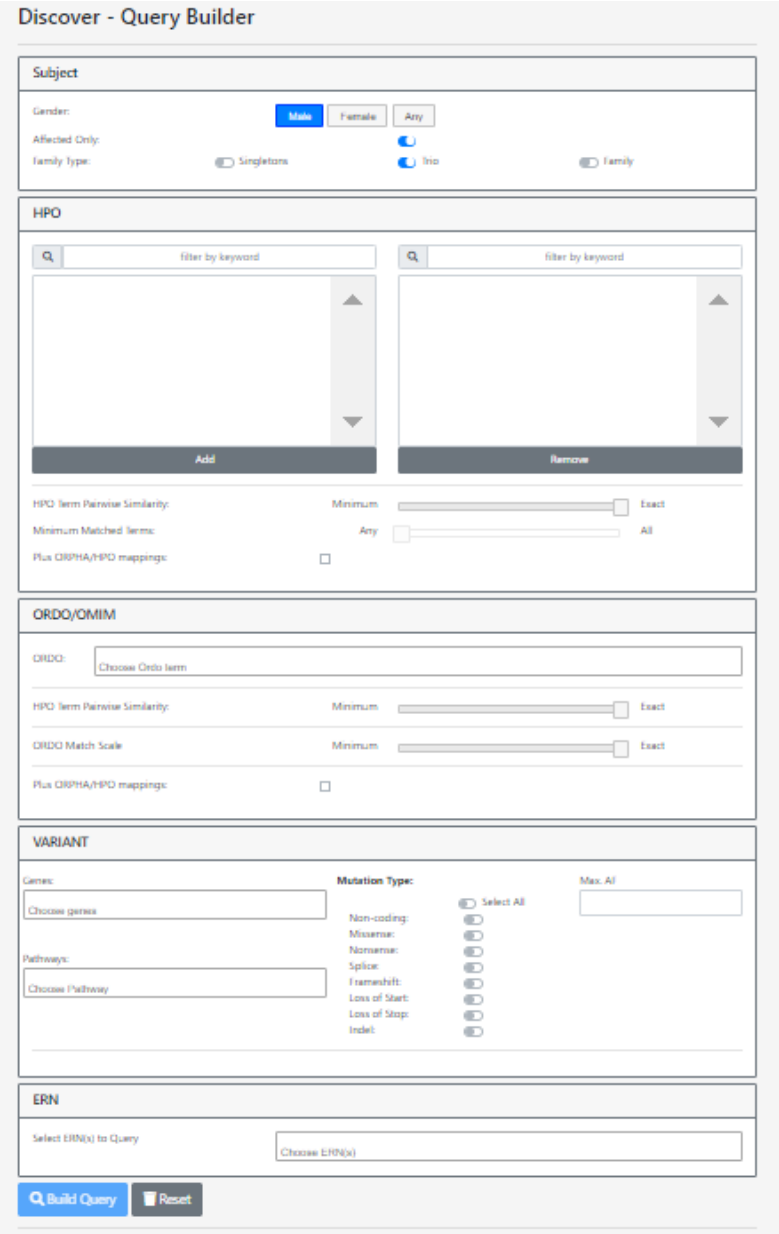

Figure 4. Screenshot of the Discovery Nexus showing the search fields

Conclusion

The Solve-RD bioinformatics platform is fully operational, serving the needs of the consortium even beyond what was initially envisioned. The platform has been built by bringing together, adapting and further developing existing resources (e.g. EGA, RD-Connect GPAP, local clusters) and new ones (e.g. Sandbox, RD3, Discovery Nexus; developed over existing tools such as MOLGENIS and Cafe Variome). By building on previous technologies rather than reinventing the wheel, Solve-RD has been very cost-efficient in deploying an innovative platform that is able to cater for the whole data cycle during the project, but also ensuring that the data is still FAIR (Findable, Accessible, Interoperable, Reusable) after the project is finished. Most of the components from the Solve-RD bioinformatics platform can be reused and adapted for other projects, therefore opening the door to future sustainability. The approach taken by Solve-RD provides a good example of how to implement a powerful bioinformatics infrastructure for a specific project and might be exportable to other projects with the need to collect, manage, archive, share, analyse and interpret thousands of datasets.

References

Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. 2013; Chapter 7:Unit7.20. doi:10.1002/0471142905.hg0720s76

Amberger J.S., Bocchini C.A., Schiettecatte F., Scott A.F. and Hamosh A. OMIM.org. Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, Volume 43, Issue D1, 28 January 2015; Pages D789–D798. https://doi.org/10.1093/nar/gku1205

Buske O. J., Schiettecatte F., Hutton B, Dumitriu S., et al. The Matchmaker Exchange API: Automating Patient Matching Through the Exchange of Structured Phenotypic and Genotypic Profiles. *Human Mutation* 2015; DOI: 10.1002/humu.22850

Deng, Y., Gao, L., Wang, B., and Guo, X. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLoS One 10*, 2015; e0115692.

Fiume, M., Cupak, M., Keenan, S., Rambla S., de la Torre S., et al. Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol 37*, 2019; 220–224. https://doi.org/10.1038/s41587-019-0046-x

Freeberg, M.A., Fromont L.A., D´Altri T., Romero A.F., et al. The European Genome-Phenome Archive in 2021. *Nucleic Acids Res.*, 2021; DOI:10.1093/nar/gkab1059

Johansson, L. F., Laurie, S., Spalding, D., Gibson, S., Ruvolo, D., et al..,. A unified data infrastructure to support large-scale rare disease research. MedRXiV 2023; 2023.12.20.23299950. https://doi.org/10.1101/2023.12.20.23299950.

Karczewski, K.J., Francioli, L.C., Tiao, G. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 2020; 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 2009; 4(7):1073-1081. doi:10.1038/nprot.2009.86

Lancaster, O., Beck, T., Atlan, D., Swertz, M., Thangavelu, D., Veal, C., Dalgleish R. and Brookes A.J. Cafe Variome: General-Purpose Software For Making Genotype-Phenotype Data Discoverable In Restricted Or Open Access Contexts. *Human Mutation* 2015, 36, 957–964. doi: 10.1002/humu.22841

Landrum M.J., Lee J.M., Benson M., Brown G.R., Chao C., et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018 Jan 4.; doi: 10.1093/nar/gkx1153

Laurie S., Fernandez-Callejo M., Marco-Sola S., Trotta J-R., Camps J., et al. From wet-lab to variations: concordance and speed of bioinformatics pipelines for whole genome and whole exome sequencing. *Human Mutation* 2016; 37(12):1263-1271. doi: 10.1002/humu.23114

Martin A.R., Williams E., Foulger R.E., Leigh S., Daughterty L.C., *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet* 51, 2019; 1560–1565. https://doi.org/10.1038/s41588-019-0528-2

McLaren, W., Gil, L., Hunt, S.E. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122, 2016; https://doi.org/10.1186/s13059-016-0974-4

Philippakis AA, Azzariti DR, Beltran S, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat*. 2015;36(10):915-921. doi:10.1002/humu.22858

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007; 81:559–575.

Schlicker A., Domingues F., Rahnenführer J., Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 7: 302, 2006; doi:10.1186/1471-2105-7-302PMID:16776819

Swertz M.A., M. Dijkstra, T. Adamusiak, J.K. van der Velde, A. Kanterakis et al. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics* 2010; doi: 10.1186/1471-2105-11-S12-S12

Zurek, B., Elwanger K. Vissers, L.E.L.M., Schüle, R. Synofzik M., et al. Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. *EJHG*. 2021; 29:1325-1331.