EJP RD
European Joint Programme on Rare Diseases

H2020-SC1-2018-Single-Stage-RTD
SC1-BHC-04-2018
Rare Disease European Joint Programme Cofund

Grant agreement number 825575

Del 11.18
Third Report on processed genome-phenome datasets and multi-omics use cases analysed, including description of new cloud and online analysis functionalities and tools

Organisation name of lead beneficiary for this deliverable:
Partner 45 – CNAG-CRG

Contributors: EBI 76-ELIXIR/EMBL[EBI, BSC (ELIXIR-ES); SIB (ELIXIR-CH), CSC (ELIXIR-FI), UU (ELIXIR-SE)];1-INSERM[INSERM-AMU,];4-LBG(LBI-RUD);35-UMCG, WSI-DECIPHER (Associated Partner);65-Radboudumc; 64-LUMC& LUMC-Endo-ERN;25-FTELE;36-UM;82-ACURARE;44-ISCIII

Due date of deliverable: month 36

Dissemination level: Public

Table of Contents

1. Introduction

The main goal of EJP RD Pillar Task 11.4 "*provision of Rare Diseases analysis and data sharing capabilities through online resources*" is to improve and scale up genomic, phenomic and multi-omics analysis, integration and sharing capabilities in order to contribute to the achievement of the IRDiRC diagnostics goals. Most of the work from Task 11.4 is addressed within the "Resources for Experimental Data Analysis and Interpretation" Work Focus, which is divided in three sub-groups:

- **User-friendly genomics analysis** (coordinated by Sergi Beltran, CNAG-CRG)
- **Cloud computing and multi-omics analysis** (coordinated by Morris Swertz, Matthias Haimel and Salvador Capella-Gutierrez from UMCG, LBG (LBI-RUD) and ELIXIR-BSC respectively)
- **Information and annotation resources** (coordinated by Sarah Hunt and Jennifer Harrow, ELIXIR/EMBL-**EBI**)

This deliverable reports the progress made during 2021 on each of those sub-groups regarding the genome-phenome datasets processed, the use cases analysed, and the new tools and functionalities developed. Note that "Cloud computing and multi-omics analysis" and "Information and annotation resources" sub-groups are highly aligned with the work done in Work Package 13 (WP13) "*Enabling multidisciplinary, holistic approaches for rare disease diagnostics and therapeutics*".

2. User-friendly genomics analysis

2.1. RD-Connect Genome-Phenome Analysis Platform

The RD-Connect GPAP (https://platform.rd-connect.eu), hosted at CNAG-CRG, is an IRDiRC recognised online platform that facilitates collation, sharing, analysis and interpretation of integrated genome-phenome datasets for Rare Disease diagnosis and gene discovery. Clinicians and researchers from the RD community can apply to register, which will enable them to submit and/or analyse data in the system.

The RD-Connect GPAP currently has 343 groups with a total of 683 authorized users.

2.1.1. New functionalities developed

During 2021 new developments were introduced in the RD-Connect GPAP that have improved the user experience. Below is a brief description of the most relevant features funded by EJP-RD:

- We have continued to improve the efficiency of the RD-Connect GPAP data workflow, particularly with regards to automation in order to improve efficiency and reduce the risk of potential human error,
    - Automation through Jenkins of the reception and variant calling pipeline improves traceability and reduces human effort.

  - We have removed indel recalibration from the variant calling pipeline in lines with updates in GATK Best Practices, which improves analysis speed.
  - We have automated the addition of Runs of Homozygosity to the RD-Connect GPAP
  - We added the option in the annotation pipeline not to save intermediate results. In this way we use less space, and the processing time is lower.
  - The annotation pipeline can now work with gVCFs and a HAIL sparse matrix, while before it accepted only a multisample VCFs.
- We have been developing a new user interface (called NextGPAP), for which we have a version in alpha-testing. We are identifying and resolving final bugs prior to releasing in production, which we expect to do in early 2022
- We have reorganized the internal server infrastructure to cope with the larger amounts of data we are receiving, especially with respect to WGS data. We have increased the platform capacity, and now have five servers available for production.
- Following Decipher's move to using the GRCh38 reference genome by default, the GPAP MME (Matchmaker Exchange) API was improved so that it can take into account the genome reference in the request and can respond accordingly.
- The version of the database that Exomiser uses was updated in 2021 to v12.0, improving the accuracy of the results returned, and improving the probability of reaching a diagnosis for the RD case being analysed.
- The Cohort App module has been fully released within the RD-Connect GPAP (Figure 1). On the server side, we have improved the performance of the filtering functionality by harmonising the intersection of experiment metadata with the participant phenotypic data in our server infrastructure. We have also created a dedicated API to create, save, and export cohorts in CSV format in an interactive way, allowing such cohorts to be used as the input to an analysis within the GPAP analysis module. The User Interface (UI) and the User Experience (UX) have been further adapted to enable filtering by multiple field values and by combinations of ontology terms (OMIM, HPO, ORDO).

Note: RD-Connect GPAP developments and features related to Data Submission, and Data Management from the user perspective are reported in Deliverable 11.13 "Third version - Additional facilities integrated to resources regarding data deposition and access, including user guidelines and documentation".

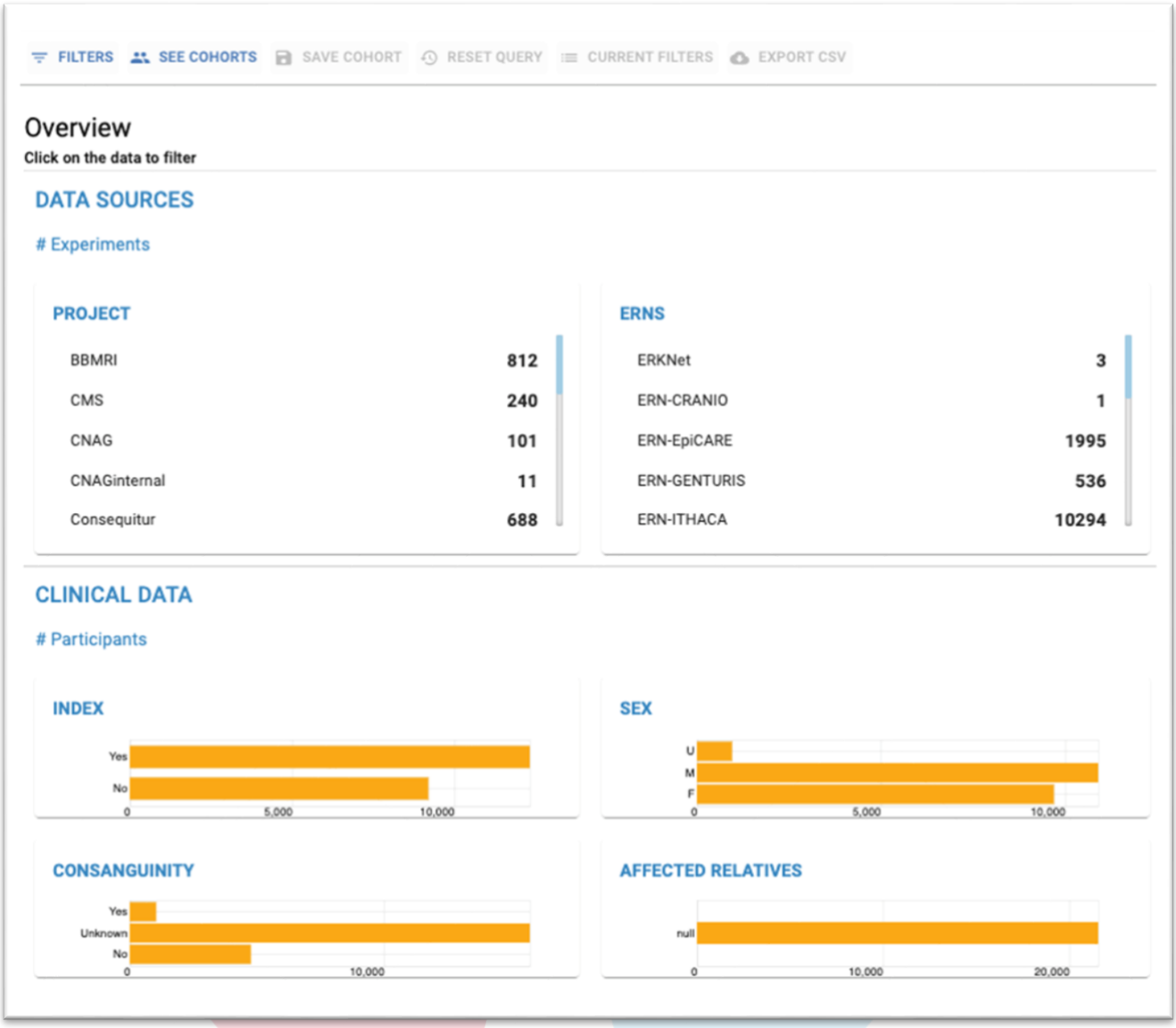

Figure 1. Screenshot of the GPAP Cohort App module

2.1.2. New datasets processed and analysed

The number of processed datasets in the RD-Connect GPAP has increased substantially in 2021. In 2020 the GPAP had 13.134 processed datasets (exomes, genomes and panels). In December 2021 this amount has increased by 7,605 processed datasets, currently containing 20,739 (18,732 exomes, 1,890 genomes and 117 panels) with their corresponding phenotypic information.

Of the 7,605 new datasets processed in RD-Connect GPAP during 2021, 738 have been submitted by a variety of RD-Connect GPAP users from across Europe, while the GPAP has also collated and processed 6,867 datasets from European registered users who have submitted these datasets for analysis with the H2020 Solve-RD project.

2.2. DECIPHER

DECIPHER (https://www.decipher.sanger.ac.uk) is a web platform that helps clinical and research teams to assess the pathogenicity of variants and to share rare disease patient records. DECIPHER is an EJP RD associated partner (not funded by EJP RD) and supports the EJP RD project.

DECIPHER provides a plethora of variant interpretation interfaces including a genome browser, protein browser, matching patient/variant interface, ACMG pathogenicity interface and patient assessment module.

This year, a major new version of DECIPHER has been released. DECIPHER now visualises genomic data in GRCh38, allowing the most up-to-date gene build information and transcripts, to enable accurate variant interpretation.

In addition, additional gene-disease annotations have been incorporated, including ClinGen gene-disease validity, ClinGen dosage sensitivity and Gene Curation Coalition (GenCC) annotations. New predictive gene scores (LOEUF and sHet) have also been added, as well as the variant predictive score REVEL, which predicts the pathogenicity of missense variants.

DECIPHER has continued to incorporate information from the ClinGen Sequence Variant Interpretation Working Group in regard to recommendations for interpreting the ACMG/AMP pathogenicity evidence criteria, including a functional evidence PS3/BS3 criterion interactive decision tree. ClinGen expert panel specifications for ACMG/AMP variant interpretation recommendations are also now displayed in the sequence variant pathogenicity interface.

Interpretation interfaces for copy-number variants have been further developed. Additional structural variants tracks are now available on the genome browser displaying ClinVar structural variants, DGV gold variants, gnomAD structural variants and ClinGen dosage sensitivity regions. An interface for recording the evidence used to classify copy-number variants (CNVs) according to the ACMG/ClinGen Technical standards for the interpretation of CNVs has also been developed.

3. Cloud computing and multi-omics analysis

3.1. New cloud functionalities

3.1.1. Cloud infrastructure

The cloud infrastructure of EJP consists of two components: a virtual research environment (VRE) for high performance bioinformatics computing and as a side-cart a meta-database called RD3 to keep track of data files and their context (patients, samples, projects).

Currently there are four VRE instances. Within one (Gearshift) UMCG has setup an EJP-RD specific environment. Other cloud environments can be set up on request

to benefit the wider European research community. To increase sustainability, the UMCG is currently implementing a fifth site of the virtual cluster environment (VRE) in a test environment (Marvin). The hardware on which this VRE is run is located at Dutch National University HPC cloud provider SURFsara. Using the network of the X-Omics initiative the goal of this VRE is to enable researchers to perform multi-omics analyses (not EJP-RD funded). This new VRE makes use of the integrated rule-oriented data system (iRODS) for data management. The Openstack is deployed using Terraform. By making this transition we also proof portability of the virtual environment created. Most other efforts to enhance the computing environment has been focussing on functionalities and pipelines described below.

In addition, UMCG has been further generalizing RD3 (first used in Solve-RD, and now further developed in EJP-RD) to be useful for the wider community, with or without Sandbox. In particular for EJP-RD,

- An automated workflow has been created by UMCG to automatically process new data in selected European Genome-Phenome Archive (EGA) datasets.
- The RD3 database workflow that is used to manage the underlying structure was rewritten to improve the efficiency and reproducibility of processing of new data. This makes the process of modifying the RD3 more manageable for data managers and allows for fine-grained customizations for new projects. All code and documentation relating to the RD3 database workflow was made open source so that RD3 could be used by collaborators and the wider scientific community.
- To proof the reusability of RD3 by the wider community, LBI-RUD has created an instance of RD3 for the dataset described in section 3.2, i.e., to make that data queryable. Regular meetings between UMCG and LBI-RUD have taken place so that an exchange of knowledge could occur, and improvements could be made to RD3.
- In collaboration with the FAIR data stewards, the UMCG has started to map RD3 to the FAIR genomes semantic model (http://fairgenome.github.io).
    - As a use-case LBI-RUD and UMCG improved FAIRification of the functional LBI-RUD RD3 node by establishing an installation of the RD3 node at the LBI-RUD premise. The RD3 installation was performed in close exchange with UMCG to improve the process and the expanded the data model for the local needs. The established LBI-RUD RD3 supports the management of local samples with whole exome sequence data, the performed analyses and resulting files.
- Currently attributes are being collected to create a table to capture metadata of RNA-sequencing experiments.

3.1.2. Functionalities and pipelines

The UMCG has implemented the DROP pipeline (https://github.com/gagneurlab/drop) into the VRE to integrate RNA and DNA sequence data analysis.

The UMCG and BSC have implemented WfExS (https://github.com/inab/WfExS-backend) in the VRE. To demonstrate the functionality and to open up workflows present in WorkflowHub to the EJP-RD VRE, we have set up the workflow implementation procedure and successfully implemented two workflows, present on WorkflowHub, one in Nextflow (https://workflowhub.eu/workflows/106) and one in CWL (https://workflowhub.eu/workflows/107)). Currently work is being performed to import easily import well-described workflows into an HPC compute environment.

LBI-RUD and WP13 are setting up well-described workflows in workflow-hub for the workflows created in WP13. Resources required for a complete registration include i) published docker container e.g., BioContainers), ii) a public GitHub repository and iii) a workflow file (compatible with workflow-hub) in GitHub. Docker images created in the CAKUT case are being improved to enable reproducible implementation in a VRE.

3.2. Multi-omics use cases analysed

CAKUT multi-omics analysis has been performed in the anDREa Digital Research Environment (DRE). Scripts and workflows are being adapted to allow for analysis in the VRE. A Snakemake file is being worked on to allow the automated import into wokflow-hub. The created Docker containers should be used by the pipeline. Internal source code, data and the Snakemake file are being worked on to be made publicly available as part of an EJP-RD GitHub repository.
Inclusion Body Myositis (IBM), the second case study, is also being analysed at the same DRE and relevant Docker containers as well as pipelines will be provided to the UMCG team once analyses are finished by all teams. The same applies to the data and workflows from the third case study, Idiopathic Portal Hypertension (INCPH).

Besides the use cases data, several samples have been analysed as part of pipeline and infrastructure development, implementation, and testing. The UMCG has analysed 357 WES samples plus 115 WES controls (70 of which were trios add-in.)

4. Information and annotation resources

The following tools are useful for variant filtering and prioritisation and are further developed partly thanks to EJP-RD. They can be used independently, in analysis workflows (e.g. in the Sandbox) or connected to other resources. For example, the RD-Connect GPAP uses VEP to annotate the genetic variants before they are uploaded to the GPAP.

4.1. DOLPHIN

The classification of variants as pathogenic or benign is the most challenging part of NGS analysis. To facilitate this process, various guidelines as the ACMG/AMP have been produced and are widely used. Nevertheless, despite their apparent simplicity, inconsistent classifications across laboratories have been reported and are mainly linked to the degrees of subjectivity and uncertainty allowed by the ACMG-AMP guidelines. one of the most challenging classification elements is the PM1 "Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation", which has been reported to be used in about 10% of cases. Today, automatic classifiers, as Intervar and Varsome, do not assign the PM1 criterion similarly even if they both use data from same reference databases (e.g., InterPro). Thus, Intervar excludes all protein domains containing variants annotated as benign or common (allele frequency > 5%) and does not consider hotspots. VarSome considers protein domains that contain at least 1 annotated pathogenic mutation if the ratio of pathogenic/non-VUS variants is > 0.5. In addition, these two approaches rely on manual variants' annotation, introducing a bias because only a minority of the observed variants are currently classified. INSERM-AMU therefore built the DOLPHIN system to provide a new layer of information in humans from protein domains. It benefits from the mutation rate that is believed to be different in organisms due to their population number, generation turnover, dissimilar metabolisms, interaction with the environment and reproductive strategies. In addition, it provides a standardized approach in agreement with the objectives of harmonization of the ACMG-AMP recommendations. This system is accessible at https://dolphin.mmg-gbit.eu. Based on variants from the ClinVar database located in protein domains and on a k-means clustering approach, INSERM-AMU demonstrated that DOLPHIN could efficiently assign a PM1 label. This was further extended to all potential mutations of human transcripts leading to 15,378,004 mutations being labelled as PM1 (29.9%). INSERM-AMU thus propose to restrict the PM1 criterion to this subset of mutations. In addition to using DOLPHIN scores to label a PM1 variant, these data can also be used to label a variant as not significantly impacting a protein residue. INSERM-AMU thus propose to create a BP8 criterion "Located in a functional domain without affecting a key residue". This new label could be allocated to 17,197,772 variants (33.4%). Altogether, DOLPHIN will provide a PM1 or BP8 label to 63.2% of mutations localized in protein domains.

4.2. The Ensembl Variant Effect Predictor and neXtProt

The Ensembl Variant Effect Predictor (VEP, https://doi.org/10.1186/s13059-016-0974-4) is a powerful, flexible tool for the annotation and prioritisation of genomic variants. The enhancement of its functionalities for rare disease analysis has continued. In response to requests from users of Ensembl VEP, the autumn Ensembl VEP release will feature more refined prediction of variant consequence around splice sites. The assignment of three additional Sequence Ontology terms (splice_donor_5th_base_variant, splice_donor_region_variant, splice_polypyrimidine_tract_variant) will enable improved prioritisation of variants which may impact splicing. This update applies to our REST, web and command line interfaces. The release also includes an extension to Ensembl VEP to predict whether variants introducing a premature stop codon are likely to result in a truncated mRNA which escapes nonsense-mediated mRNA decay. We have aligned our criteria with the DECIPHER project for consistency across the resources.

The Variant Effect Predictor (VEP) can also be used to predict the potential impact of non-synonymous amino-acid variations on protein structure and function, which can help prioritizing functional validation studies. More specifically, SIFT and PolyPhen-2 scores can be retrieved using VEP for any ENSP single amino-acid change. The mapping of neXtProt isoforms to ENSP identifiers was improved, which resulted in a gain of interoperability between VEP and neXtProt. A new interface for the VEP tool was built in neXtProt , that allows users to call VEP directly from any neXtProt isoform provided it aligns perfectly with an ENSP (https://www.nextprot.org/entry/NX_P52701/gh/calipho-sib/VEP-community-tool) . On this interface, variants can be visualized in the context of other positional annotations such as domains or PTMs. By zooming in on a region of the sequence, one or more variants can be selected by simply clicking on a residue in the sequence and providing the variant amino acid, followed by clicking on the "Get Predictions" button. The "+ Add Variants" button allows one to select multiple variants in one go by entering their position in the sequence and variant amino-acid and obtain their predicted effect by clicking on the "Get Predictions" button. It is also possible to upload a CSV file containing the position in the sequence, the original and variant amino acid for each variant and get their predicted impact. The SIFT and PolyPhen-2 scores are returned in the variant table, which can be exported in CSV format.