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 11.19
Fourth 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

Collaborators: 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, DECIPHER (Associated Partner);65-Radboudumc; 64-LUMC& LUMC-Endo-ERN;25-FTELE;36-UM;82-ACURARE;44-ISCIII

Due date of deliverable: month 48

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 coordinated by Sergi Beltran (CNAG-CRG).

This deliverable reports the progress made during 2022 regarding the genome-phenome datasets processed, the use cases analysed, and the new tools and functionalities developed. Note that some actions related to cloud computing, multi-omics analysis and annotation resources 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 the CNAG-CRG, is an 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, share and analyse data in the system.

The RD-Connect GPAP is an IRDiRC recognised resource and at the end of 2022 had 356 groups using it, with a total of 734 authorized users.

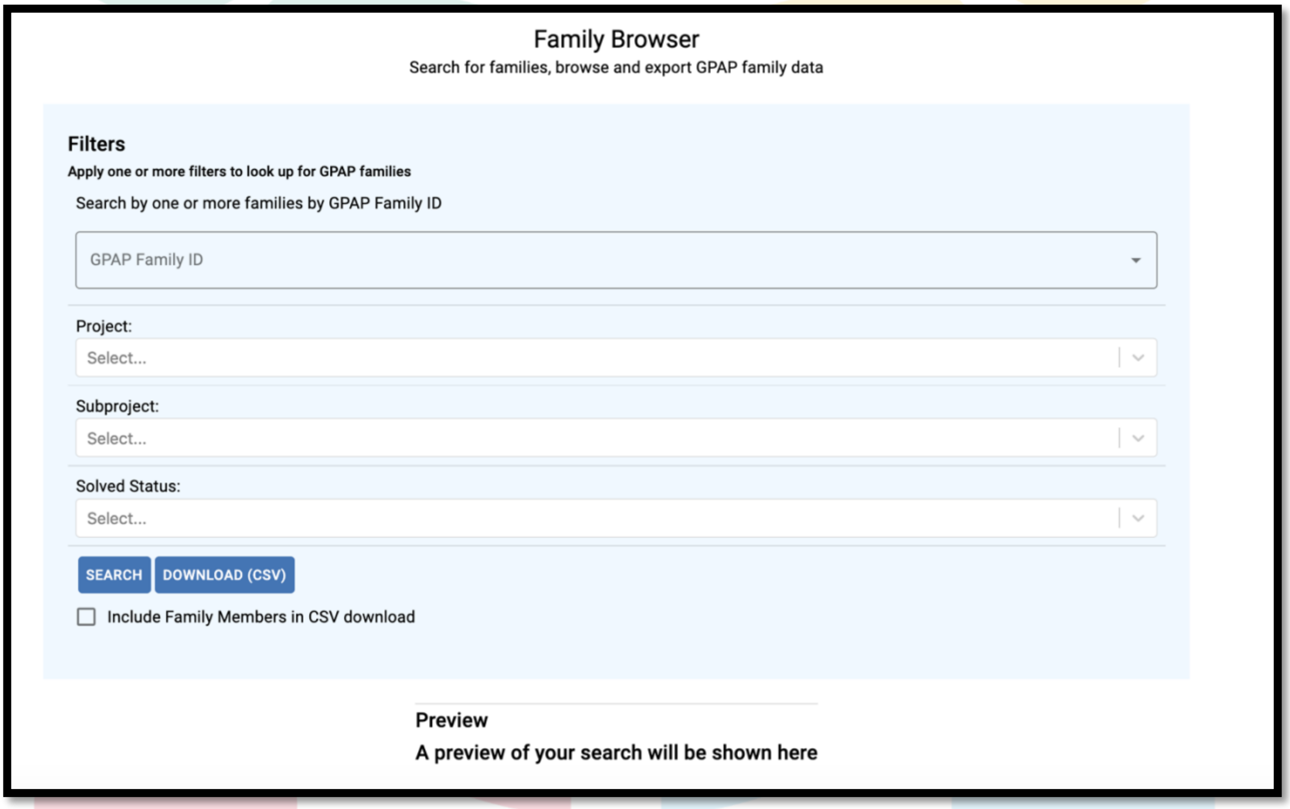
2.1.1. New functionalities developed

During 2022 new developments were introduced to 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,
    - We have been working on the implementation of Grch38 pipeline adaptation in the RD-Connect GPAP. We have updated all the annotations to GRCh38 (VEP, CADD) and we have moved to gnomAD whole genome version. The aim is to process and analyse data both in GRCh37 and GRCh38. It is still pending to test the results and then adapt the frontend to use the GRCh38 related API, and we expect to finish the implementation during the first semester of 2023.

  - We have been working on the implementation of new annotations (Intervar, REVEL Score, SpliceAI, MitoMap Polymorphism). In this way the users could have more information on variants, especially non-coding variants, facilitating interpretation.

- We have been developing a new user interface (called NextGPAP), for which we have a version in beta-testing available in the RD-Connect GPAP Playground.
- We have added two new views to the RD-Connect GPAP Cohort App module. The Family View enables users to look for specific Family IDs or search family by solved status and/or projects; and it is also possible now to export the data from the module. With the Tagged Variant View users can see all tagged variants in the RD-Connect GPAP.



Family Browser

Search for families, browse and export GPAP family data

Filters

Apply one or more filters to look up for GPAP families

Search by one or more families by GPAP Family ID

GPAP Family ID

Project:

Select...

Subproject:

Select...

Solved Status:

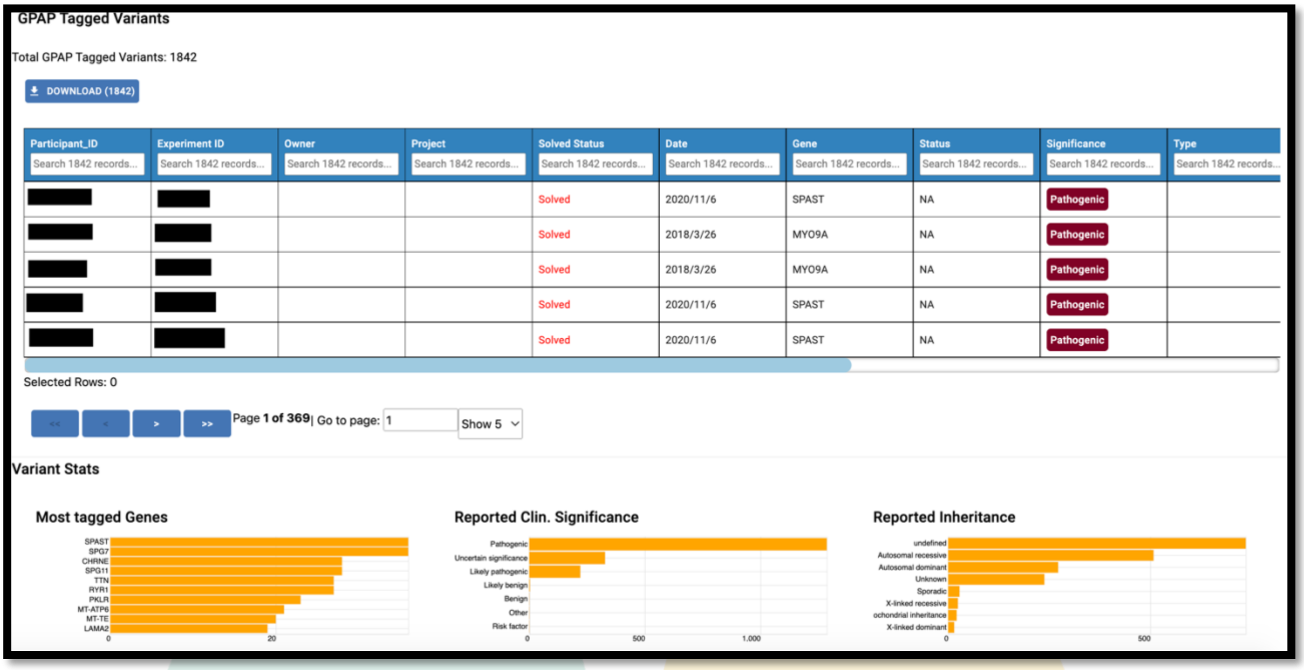
Select...

□ Include Family Members in CSV download

Preview

A preview of your search will be shown here

Figure 1. Screenshot of the Family View in the RD-Connect GPAP Cohort App



GPAP Tagged Variants

Total GPAP Tagged Variants: 1842

Participant_ID	Experiment ID	Owner	Project	Solved Status	Date	Gene	Status	Significance	Type
Search 1842 records...	Search 1842 records...	Search 1842 records...	Search 1842 records...	Search 1842 records...	Search 1842 records...	Search 1842 records...	Search 1842 records...	Search 1842 records...	Search 1842 records...
[redacted]	[redacted]			Solved	2020/11/6	SPAST	NA	Pathogenic	
[redacted]	[redacted]			Solved	2018/3/26	MYO9A	NA	Pathogenic	
[redacted]	[redacted]			Solved	2018/3/26	MYO9A	NA	Pathogenic	
[redacted]	[redacted]			Solved	2020/11/6	SPAST	NA	Pathogenic	
[redacted]	[redacted]			Solved	2020/11/6	SPAST	NA	Pathogenic	

Selected Rows: 0

Page 1 of 369| Go to page: 1 Show 5

Variant Stats

Figure 2. Screenshot of the Tagged Variant View in the RD-Connect GPAP Cohort App

- We have implemented Phenopackets version 2.0 in the RD-Connect GPAP PhenoStore module. It is possible now for the users to transfer more information, thanks to the extended data model that is provided by version 2.0 compared to version 1.0.
- We have developed a stronger integration between the Data Management and the PhenoStore modules of the RD-Connect GPAP. Now only the PhenoStore ID is stored in the Data Management, and from PhenoStore user interface, you can now see which experiments are linked to the participants.
- Also, in the RD-Connect PhenoStore module users now can submit an Excel with the participants relation and the pedigree will be drawn automatically. This means that the user does not have to draw the pedigree manually.
- We have implemented an API in the PhenoStore module of the RD-Connect GPAP, which provides autocomplete and search functionality for HPO ontology. In this way we can control better the resource in terms of versioning and monitoring. Those functionalities are then used by the PhenoStore client and batch submission module API.
- We are currently working on PhenoPackets import functionality, so that a user could submit data in Phenopacket format (right now RD-Connect PhenoStore can only export data as PhenoPackets).
- Improvements in PhenoStore - API, it is now easier to collect information about the families, this end point makes easier the integration of family into the GPAP analysis module

- Users can link to the Varsome website and are able to see the ACMG classification there.
- We are testing a new way of updating Clinvar information on already existing datasets without the needs of reannotating all the datasets from scratch. In this way the users will have the latest Clinvar annotation on already available datasets in the RD-Connect GPAP.

2.1.2. New datasets processed and analysed

The number of processed datasets in the RD-Connect GPAP has substantially increased during 2022. In 2021 the GPAP had 20,739 processed datasets (exomes, genomes and panels). In December 2022 this amount has increased by 7,288 processed datasets, currently containing 28,027 (25,187 exomes, 2,706 genomes, 134 panels) with their corresponding phenotypic information.

Of the 7,288 new datasets processed in RD-Connect GPAP during 2022, 829 have been submitted by a variety of RD-Connect GPAP users from across Europe, while the GPAP has also collated and processed 6,459 datasets from European registered users who have submitted these datasets for analysis as part of the H2020 Solve-RD project.

2.2. DECIPHER

DECIPHER (https://www.deciphergenomics.org/) 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.

We have continued to integrate new data sources to support the interpretation of variants including probability of haploinsufficiency (pHaplo) and probability of triplosensitivity (pTriplo) (Collins *et al.*, 2022; https://europepmc.org/article/MED/35917817). We now report ClinGen Clinical Actionability information for genes to provide evidence-based assessments for available clinical interventions and their ability to prevent or mitigate health outcomes. We have also created an extensive range of new pop-up information windows which make it easier to access version information, basic descriptions of and links to further details about the data sources we import.

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 metadatabase called RD3 (Rare Disease Data about Data) to keep track of data files and their context (patients, samples, projects).

3.1.2.Functionalities and pipelines

Since previous reporting the UMCG has had regular releases of the MOLGENIS Variant Interpretation Pipeline (VIP), with v4.12.1 as latest release. To reduce dependency upon specific infrastructure VIP now makes use of singularity containers and has been implemented in Nextflow. Settings are configurable. Both compressed and uncompressed vcf and bcf formats are now supported as input. Several VEP plugins have been incorporated, including SpliceAI, Grantham, PhyloP and UTRannotator. Custom gnomAD annotation files have replaced default VEP Plugins. Improvements have been made to the decision tree, based upon the UMCG diagnostic procedure. Variant pathogenicity prediction for SNV and Indel variants has been improved by creating and incorporating a new CAPICE model. Structural Variant annotation and has been improved by incorporating AnnotSV. In addition, support for reduced penetrance genes for inheritance was added as well as GRCh38 support.

The VIP report has been updated to be more input agnostic. The user-friendly report now shows information from any annotation present in the vcf file. Alignment data can be viewed around the variants in the report.
A first investigation on non-coding annotation tools has started to select tools to incorporate in VIP to improve non-coding variant prioritization, of which UTRannotator has been implemented. Currently, we are working on incorporating GREEN-VARAN functionality.

Due to changes in licencing of DisGeNet and keeping up to date with latest information, VIBE has been removed from new versions of VIP. However in order to improve genotype to phenotype matching, we are currently working on LIRICAL

The BSC has released the latest version of the Workflow Execution Service backend (WfExS-backend), which at the time of writing is 0.9.0. Since the past reporting period, the workflow orchestrator has matured enough to be able to describe both prospective and retrospective provenances in the form of RO-Crates following the Workflow Run RO-Crate profile (link) from workflow instantiations. The core of the export steps has been implemented, so any of the elements involved in a workflow execution (inputs, outputs, provenance, ….) can be exported to a supported facility which is able to provide permanent identifiers. Both internal WfExS-backend cache and Nextcloud are currently supported, and future developments will add support for Zenodo, WorkflowHub and B2SHARE, among others.

Workflow stage declarations now can contain placeholders, which ease the declaration of complex workflows where some keywords shared by several parameters and URLs must match. On the engines side, Nextflow DSL2 workflows should now be better supported, like the VIP workflow and the ones from nf-core community. This is due all the included workflow files are now tracked in order to learn the needed containers. Last, but not the least important, many subcommands which help both in staged workdir and cache management have been added, and some internal optimizations about when to hard link and when to copy contents have been added.

The UMCG and BSC are currently testing usability of WfExS-backend for implementation of workflows from sources other than WorkflowHub (i.e., git repositories like GitHub). VIP 4.9.0 is used as the test model, as well as other ones from communities outside EJP-RD. VIP has been installed on the UMCG VRE using WfExS-backend. This required some manual changes to the configuration file, improvements are made to make the use of WfExS-backend more applicable.

3.2. Multi-omics use cases analysed

Together with the RadboudUMC the UMCG has created the docker images for the CAKUT use case. The docker images have been tested by running the Mixomics data set from WP13, which is one of the three selected use cases. We are working on further validation. No EJP RD partners have asked for analyses of samples. Building on the ongoing work the developments will be made publicly available and registered to be findable in the UMCG Virtual Cluster Environment. The created workflows by WP13 will be deposited in a public git repository and registered in WorkflowHub with the help of the RadboudUMC.

ACURARE analysed the CAKUT use case using the Similarity Network Fusion (https://doi.org/10.1038/nmeth.2810) as a different approach for patient stratification. ACURARE is working on an additional manuscript to publish the findings.

4. Information and annotation resources

4.1. Tools to predict pathogenicity of genetic variants

Since the previous report the UMCG has developed GRCh38 support for the CAPICE variant pathogenicity estimation tool. The usage has been documented on GitHub.

ACURARE has developed a new network-based mutation impact prediction method called predatoR and has submitted a manuscript to bioRxiv (https://doi.org/10.1101/2022.11.29.518310).

4.2. The Ensembl Variant Effect Predictor

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. We have continued to extend its functionalities for the identification of variants potentially involved in rare disease.

We have worked with collaborators to optimise an Ensembl VEP extension which predicts when a variant in the upstream untranslated region of a transcript is likely to create or disrupt the normal open reading frame and thus disrupt normal gene function. This will be available in the VEP REST service and webtool in our December release.

In response to community requests, we have started a series of improvements to Ensembl VEP's annotation of structural variants. The first results of this project, more efficient matching of SVs to reference datasets (such as pathogenic variants in ClinVar and gnomAD variants with population frequency information) and more accurate consequence prediction for CNVs will be available in our December release.