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 11.20
Fifth 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 65

Dissemination level: Public

Table of Contents

1. Introduction

The main goal of EJP RD Pillar 2 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 "Experimental Data, Biomaterial and other Resources" Work Focus.

This deliverable reports on the progress made between January 2023 and August 2024 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 (RD) 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 in August 2024 had 367 registered groups with a total of 801 authorized users.

2.1.1. New functionalities developed

In this reporting period 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:

- The implementation of a Nextflow pipeline on GRCh38 has been tested both in the CNAG cluster and in Amazon Web Services. It includes calling and annotation of SNVs, InDels, CNVs and SVs. It generates VCFs and CRAMs to enable visualisation. From October 2024 all new submitted datasets to the RD-Connect GPAP will be processed with the GRCh38 pipeline. Old datasets previously processed with GRCh37 will be slowly reprocessed to GRCh38.
- The new user interface (called NextGPAP) was released and further improved with new features and resolution of bugs, including adaptations to GRCh38 reference. The old interface is planned to be discontinued from mid September 2024.

- Improvements have been made to the implementation of the IGV genome browser, which will allow soon to visualise also VCFs from different variant types. These developments have been tested internally and will be released for the RD-Connect GPAP with the release of the GRCh38 version.
- A system to update ClinVar annotation easily has been implemented.

2.1.2. New datasets processed and analysed

The number of processed datasets in the RD-Connect GPAP has increased by 2,156 between January 2023 and August 2024. The platform currently includes 30,011 datasets (26,255 exomes, 3,622 genomes, 134 panels) with their corresponding phenotypic information.

Of the 2,156 new datasets processed in RD-Connect GPAP, 736 have been submitted by a variety of RD-Connect GPAP users from across Europe, while the GPAP has also collated and processed 1,420 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 extend the annotations DECIPHER provides to support variant interpretation. We now display AlphaMissense (PMID:37733863) scores, which categorise missense variants as likely pathogenic or benign using knowledge of conservation and the variant location in predicted 3 dimensional protein structure, UTRannotator (PMID:32926138) results which predict when variants in 5'untranslated regions of transcripts will create or disrupt upstream open reading frames and PhyloP (PMID:19858363) measures of conservation, which can be helpful when applying ACMG/AMP classifications. Results of Multiplexed Assays of Variant Effect (MAVEs) which interrogate cellular phenotype (from MaveDB, PMID: 31679514) are displayed to help identify variants impacting function. We have implemented an interactive decision tree tool for the evaluation of functional data following current recommendations (PMID:31892348). To enable simple access to other relevant resources, we now provide links from DECIPHER gene-disease association pages to the OpenTargets platform, which collates information to support

therapeutic target identification and prioritisation, ClinGen Variant Curation Expert Panel recommendations and ACMG secondary finding information.

3. Cloud computing and multi-omics analysis

3.1. New cloud functionalities

3.1.1. Cloud infrastructure

The cloud infrastructure of EJP RD 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 v7.9.0 as latest release. VIP now supports both short- and longread data starting from raw data to intermediate data e.g. FastQ, CRAM or VCF for all different variant types. To warrant FAIRness VIP uses Apptainer used as container.

VIP is now based on GRCh38, support for GRCh37 and T2T support is available via the lift over tool. In addition to the current decision tree we have created a custom decision tree to assess non-coding variants making use of new annotations.

Proof of concept pipelines incorporating RNA sequencing data and methylation information from short- and longread sequencing data respectively have been developed.

The BSC has released the latest version of the Workflow Execution Service backend (WfExS-backend), which at the time of writing is 1.0.0b1. Since the past reporting period, the workflow orchestrator has matured enough to be able to consume both prospective and retrospective provenances in the form of RO-Crates following the Workflow Run RO-Crate profile (link) from workflow instantiations. This eases both reproducible (same workflow, container images and inputs) and replicable analyses (some of the original inputs and parameters are replaced by other ones). The core of the export steps has gained several plugins, so it is currently possible to export to Zenodo, B2SHARE and Dataverse, among others.

Analysis reproducibility starting from a previously generated Workflow Run RO-Crate can be achieved at different levels. As the embedded RO-Crate metadata records the permanent identifiers of the used inputs and the

workflow, the basic reproducibility level just uses these details to try reproducing the analysis. The intermediate level of reproducibility fixes the specific versions of used container images, using the gathered metadata about the specific container layers used in the previous execution. When the previously generated Workflow Run RO-Crate also holds snapshots of container images, inputs or the workflow, then the highest reproducibility level implemented in WfExS-backend can be achieved, where these payloads are used instead of refetching them. On the engines side, Nextflow workflows are better supported, as both embedded scripts and needed plugins are recorded, improving the chances of reproducibility. As some nf-core workflows limit the execution time of some of their steps, and there is no standardized way among the different workflows to change it, a feature to override these hardcoded time limits has been added to WfExS-backend. should now be better supported, like the VIP workflow and the ones from nf-core community. Last, but not the least important, working directories are more relocatable now, in preparation to allow execution scenarios where WfExS itself is in a docker instance.

The UMCG and BSC have been 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. WfExS configuration files have been adapted to accommodate for the Nextflow setup used by VIP.

3.2. Multi-omics use cases analysed

No additional samples were provided during year 6, therefore there was no integrated analysis of the multi-omics data during year 6.

4. Information and annotation resources

4.1. Tools to predict pathogenicity of genetic variants

The UMCG has created a new CAPICE model to improve variant pathogenicity prediction, thus including more variants and gnomAD homozygosity counts based upon a new XG boost model version.

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.

Multiplexed assays of variant effect (MAVEs) have the potential to help interpret the large numbers of variants of unknown significance currently being identified in clinical sequencing. We have collaborated with the Atlas of Variant Effect to integrate such experimental results from MaveDB (https://doi.org/10.1186/s13059-019-1845-6) into Ensembl VEP to ensure they are easily accessible via simple REST and web interfaces and can be easily integrated into analysis pipelines using Ensembl VEP as a base. To help identify variants linked to human phenotype, we now report when a variant is in the Geno2MP (http://geno2mp.gs.washington.edu) database of rare variants found in people with Mendelian disease. A further extension enables the integration of phenotype associated variant data from the AVADA (https://doi.org/10.1038/s41436-019-0643-6) database for command line Ensembl VEP annotation.

To provide further evidence of possible disease association, Ensembl VEP can now report information from gene paralogs and orthologs. The PhenotypeOrthologous plugin reports any phenotypes associated with mouse and rat genes orthologous to the human gene a variant falls within. The Paralogs plugin annotates variants with any phenotype assertions attached to variants at the same location in a paralogous gene.

We have further enhanced Ensembl VEP's support for structural variants (SVs) by adding support for breakend events and enabling the use of pre-calculated CADD scores of predicted SV deleteriousness (https://doi.org/10.1101/gr.275995.121) via a plugin. To help interpret the effect of gene duplications or deletions, we have developed the DosageSensitivty plugin which integrates pre-calculated scores of haploinsufficiency and triplosensitivity (https://doi.org/10.1016/j.cell.2022.06.036).

We have created multiple new Ensembl VEP extensions to integrate predictions of variant deleteriousness. Support for AlphaMissense (https://doi.org/10.1126/science.adg7492) results, which classifies missense variants as likely pathogenicity or benign using conservation and protein structure information, was made available on the day the scores were released to facilitate their use as quickly as possible. Additional scores of missense variant pathogenicity are available using the new VARITY (https://doi.org/10.1016/j.ajhg.2021.08.012) plugin. To provide predictions of variant impact on gene expression, we have developed a plugin to integrate scores from DeepMind's Enformer (https://doi.org/10.1038/s41592-021-01252-x) method. Additional information on possible variant effect on splicing are

available through the integration of SpliceVault (https://doi.org/10.1038/s41588-022-01293-8), which used GTEx data to predict when a variant may cause cryptic donor or acceptor sites or exon skipping.

New functionality is added to the web tool and REST service, where open data redistribution is permitted. To enable easier use of Ensembl VEP locally for the analysis or restricted access data or for integration of licensed resources, we have made improvements to the Ensembl VEP Docker image, including the addition of Ensembl VEP plugins and their dependencies and support for environmental variables and additional platforms. We have also made a number of other user-requested improvements including more efficient handling of the annotation of individual genotypes in files with multiple samples, an extension to the identify de novo genotypes when trios are analysed and the reporting of summary statistics when using custom datasets.

We have created an updated Beacon endpoint using the latest specification which returns Ensembl VEP results, given a known or novel variant as input to enable integration into the EJP RD virtual platform.