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 11.16
First 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: ELIXIR/EMBL-EBI (EBI & ELIXIR-ES[BSC]); INSERM-AMU; LBG(LBI-RUD); UMCG, Radboudumc; WSI-DECIPHER (not funded, in-kind contribution).

Due date of deliverable: month 12

Dissemination level:
Public

Table of contents

List of figures

1. Introduction

EJP RD Pillar 2 Task 11.4 *"provision of RD analysis and data sharing capabilities through online resources*" has as main goal 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.

During the EJP RD General Assembly in Gdansk (Sept 17-19, 2019), one of the Work Foci kick-launched was "**Resources for experimental data analysis and interpretation**". This Work Focus (WF) encompasses many of the Task 11.4 activities and focuses. During the meeting, a parallel session was dedicated to this WF and it was agreed to split discussions in three sub-groups:

- **User-friendly genomics analysis** (coordinated by Sergi Beltran [CNAG-CRG]) including mainly discussions on RD-Connect GPAP and DECIPHER, as well as prioritization, implementation and testing of new features by the Rare Disease (RD) community.
- **Cloud computing and multi-omics analysis** (coordinated by Morris Swertz [UMCG], Matthias Haimel [LBG (LBI-RUD)] and Salvador Capella-Gutierrez [ELIXIR-BSC]) including mainly discussions on use-cases for collaborative analysis of multi-omics data, as well as development and implementation of supporting pipelines/workflows and Information Technology (IT). This group provides the main link to WP13 (mainly through liaisons with Peter-Bram 't Hoen [Radboudumc] and Chris Evelo [UM])
- **Information and annotation resources** (coordinated by Sarah Hunt, [ELIXIR/EMBL-EBI] and Jennifer Harrow [ELIXIR/EMBL-EBI]) including discussions and prioritization activities for new features on resources relevant for variant annotation such as VEP, neXtProt and UniprotKB, but also new features for tools such as UMD Predictor for genetic variant pathogenicity assessment.

In 2019, Task 11.4 members have participated in 2 surveys which are, and will be, key to map RD community needs and resources for experimental data analysis and interpretation.

The first survey was organised within Pillar 2 (led by Franz Schaefer [UKL-HD] and Mary Wang [FTELE]) to capture input from the ERNs. From the 291 individuals that answered the survey, 48.8% indicated that they generate or use -omics data. Genomics is the type of data mostly generated or used, followed by transcriptomics, epigenomics, metabolomics and proteomics. To mention some relevant data types, 30.9% individuals indicated that they generate or use exome data, 24.1% genome data, 17.2% RNA-Seq and 12% DNA methylation data. Also, 19.6% of individuals indicated that they have different types of 'omics data originating from the same sample, which would be very relevant to identify use cases for cloud computing and multi-omics analysis. The survey also highlighted that many of the resources included within the EJP RD are not that well-known or used by the RD community.

The second survey was organised by the Work Package (WP)11 ("*common virtual platform for discoverable data and resources for RD research*"), coordinated by Giselle Kerry [EMBL-EBI]. It captured input from the 24 resources that answered; 9 of those (37.5%) are RD specific, and 4 (17%) are IRDiRC recognised resources. The survey highlighted that although most resources use some standards and ontologies, there is room for improvement on data mapping and further data standardisation and FAIRification.

Based on the respondents of the first survey, the WF on "Resources for experimental data analysis and interpretation" has identified some members to provide use cases to this WF (see below) and will also invite members to collaborate in prioritising new developments and implementations based on user's needs. Series of webinars targeting the RD community has been programmed for 2020 to accomplish at least 3 goals:

1) disseminate among the RD community the resources available in EJP RDs Pillar 2 to support RD research;
2) provide training on these resources and
3) capture input from the users on those resources, including needs not currently covered.

This deliverable reports mainly on outcomes from Task 11.4 regarding processed genome-phenome datasets, multi-omics use cases analysed, new cloud functionalities and new online analysis functionalities and tools.

2. Processed Genome-Phenome datasets

During 2019, the RD-Connect Genome-Phenome Analysis Platform (GPAP) hosted at the CNAG-CRG has significantly increased the number of processed exomes and genomes integrated with phenotypic information. On December 20th, 2019, a total of 11,526 experiments (genomes, exomes and panels) were released, albeit some under embargo and only available to some specific users for the time being. This triples the number of experiments released in November 26th, 2018, which were 3,684. A lot of the new data has been contributed by 4 European Reference Networks (ERN ITHACA, NMD, RND and GENTURIS) which are beneficiary partners in EU H2020 Solve-RD project. However, data released in 2019 also includes 502 experiments (406 exomes and 96 genomes) from the RD community that have not been submitted to the Solve-RD project. The GPAP currently has over 500 authorised users which have gone through the validation procedure indicated in its Code of Conduct.

All the datasets have been processed with the RD-Connect GPAP standardised SNV and InDel variant calling pipeline, which is mostly based on GATK best practices (*Laurie et al. 2016*, PMID: 27604516). Downstream processing includes annotation using Ensembl's Variant Effect Predictor (from EBI), together with additional sources such as gnomAD, ClinVar and internal allele frequency. During 2019 the identification of Runs of Homozygosity (RoH) was automated and included as part of the standard pipeline with PLINK software. Identified RoH are made available in the GPAP to support with filtering variants. Furthermore, an initial pipeline for Copy Number Variant (CNV) was set up which, for the time being, uses only ExomeDepth software. It is worth noting that while for SNVs and InDels, datasets can be processed individually to generate a gVCF, CNV processing needs to be done in batches by grouping those exomes captured with the same kit. During 2019 over 400 datasets from 2 different capture kits were processed including mostly data from Solve-RD; it is planned to expand CNV detection to other datasets during 2020.

3. Multi-omics use cases analysed

During the first year of the EJP RD project, EJP RD partners Radboudumc (Peter-Bram 't Hoen, lead), UM, INSERM-AMU, UMCG, LUMC, ACU/ACURARE, have focused on establishing collaborations with ERNs and identified some initial use cases to work on during the second year. It is aimed with the use cases to:

- show the added value of multi-omics data for diagnosis, disease modifiers (including environmental exposure), identification of drug targets and druggable pathways, toxic side effects (WP13);
- show the added value of data FAIRification for data integration (in connection with tasks 11.3 "*data deposition and access to data infrastructures for RD research*", 11.4 "*provision of RD analysis and data sharing capabilities through online resources*" and WP12 " *enabling sustainable FAIRness and Federation at the record level for RD data, patients and samples*");
- demonstrate effective use of prior knowledge on pathways (WP13 "*enabling multidisciplinary, holistic approaches for rare disease diagnostics and therapeutics*");
- benchmark different network and pathway-based data integration methods (WP13).

The use cases are being used to set up analysis infrastructure, workflows and pipelines. To develop the EJP RD infrastructure, the workflows and pipelines will be made generic, reusable for others and deployed in the virtual platform.

The use cases were selected from initial respondents to the ERN survey, and the respondents who claimed to have multi-omics data available were approached with an additional short survey to make an inventory of available datasets. The datasets were then selected according to the following criteria:

- Multi-omics data available from >20 human samples per condition
- Preferably paired (from same individuals and same cells/ tissue)
- Data quality had been investigated
- Rich metadata were available (to be made FAIR)
- Data could be shared and published (possibly restricted access)
- ERN members were committed to spend at least a few hours per month on feedback and discussion of results

Three use-cases met all these criteria and were therefore selected:

- Pagan et al. (Barcelona) (ERN RareLiver): Differential diagnosis between idiopathic non-cirrhotic intrahepatic portal hypertension and cirrhotic portal hypertension (Metabolomics and transcriptomics data available)
- Schanstra et al. (INSERM - Toulouse) (ERN ERKNet): Identification of disease modifiers/pathways for CAKUT (congenital anomalies of the kidney and urinary tract) during pregnancy (miRNA, proteomics, peptidomics, metabolomics data available)
- Udd et al. Helsinki (Helsinki) (ERN EURO-NMD): Modifiers of disease severity in Inclusion Body Myositis (Genomics, transcriptomics data available)

A generic analysis strategy was defined and initial efforts on data FAIRification were made in the WP12/13 Maastricht workshop (November 26-29, 2019). After the completion of the data transfer agreement and the set-up of a collaborative research environment with restricted access, data analysis will start.

Furthermore, several datasets have been processed, mainly to gain experience and to setup the systems for production. LBG (LBI-RUD) processed 309 ATAC-Seq samples using standardised pipelines and 30 WES samples using cloud-based solutions. These were in-house samples used for pipeline / technology development and to gain experience with different setups. The cloud solutions were run on AZURE (cromwell, WDL docker) and the Vienna scientific cluster (using singularity instead of docker).

4. New cloud functionalities

With the advent of next-generation sequencing, large scale analysis services are needed for rare disease research and diagnostics. That is a huge challenge for most rare disease laboratories, given the complexity and scale of the necessary analysis pipelines while considering Ethical, Legal, and Social Implications (ELSI) aspects such as GDPR.

In close collaboration between EJP RD WP11 – Task 11.4 and Solve-RD WP4; and building on many previous projects such as RD-connect, EXCELERATE, GEN2PHEN and CORBEL, suitable Cloud and analysis functionalities have been designed, with the mission to deliver generic solutions that can be used within the context of different projects, in particular involving ERNs. In 2019, UMCG invested in the new development of an 'Analytical Sandbox' for new innovative pipelines aiming first at bioinformaticians. This includes a 'metadatabase' called 'RD3' to ensure that patients, samples and files that are input and output in the analysis can be tracked and traced.

EJP RD in its first year has invested the general design and first implementations of the necessary 'virtual' platform. Meanwhile, EJP RD WP11 – Task 11.4 has used Solve-RD as starting customer, to ensure that any new developments can immediately tested in real research settings. Within the EJP RD project, the sandbox environment is generalized so that it can be easily implemented within a wide range of projects. A relationship with The ELIXIR Rare Disease Service Bundle was established for further dissemination and to provide training on working in a High-Performance Computing (HPC) environment.

In 2019, the Solve-RD project experience was used for EJP RD to create a procedure to set up a virtual machine ('sandbox') making use of secure data storage provided by the ELIXIR EBI EMBASSY. This sandbox is available through a secure online environment and will enable users to deposit, share and analyse phenotypic, genomic and multi-omics data. To support FAIRification of data the RD3 database has been designed of which an instance is set up within the Solve-RD project.

In the first year of the EJP RD project several basic sandbox environments that can be accessed through a private-public keypair have been set up. These environments use a Linux OS and have been set up using CentOS7 with Spacewalk for package distribution and management. Job scheduling is performed using Slurm Workload Manager 17.11.9 on 12 available nodes, each with 14 cores.

In the different sandboxes, an LMOD module system provides a toolchain of commonly used available tools that are deployed using Easybuild and can be easily loaded through a module system. For pilot tests, tools can be installed in user defined directories and upon request, to be used in large-scale analyses; they can be added to the toolchain upon request. Within the Solve-RD project such requests have been made (e.g. Exomiser).

Specifically, for the EJP RD project, a virtual machine is currently being set up to provide an analysis cloud for the use-cases of the three different ERNs (ERKNet, ERN EURO-NMD and Rare-Liver), each with 1 Tb of storage space and shared resources.

In this context, ELIXIR-ES (BSC) have also looked into workflows reproducibility in collaboration with other projects like EOSC Life. Specifically, the aim is the adoption of 1) CWL (Common Workflow Language) as a mechanism to specify any workflow; 2) RO (Research Objects) Crate as a mechanism to work towards reproducibility and repeatability; and 3) Galaxy as a friendly front-end to implement workflows. Reproducibility is important to facilitate the deployment, using software containers, of analytical workflows anywhere including those places that are not directly connected to the Internet for security reasons.

In the future closer connection to WP13 will be sought to provide infrastructure for other use-cases. Within the sandbox environment, a collaboration on implementation of existing pipelines for multi-omics analysis will be performed.

5. New online analysis functionalities and tools

EJP-RD Task 11.4 is, among other, committed to providing user-friendly analysis capabilities to the RD community. Main developments have been done on the RD-Connect Genome-Phenome Analysis Platform (GPAP), DECIPHER and other tools.

5.1. RD-Connect Genome-Phenome Analysis Platform

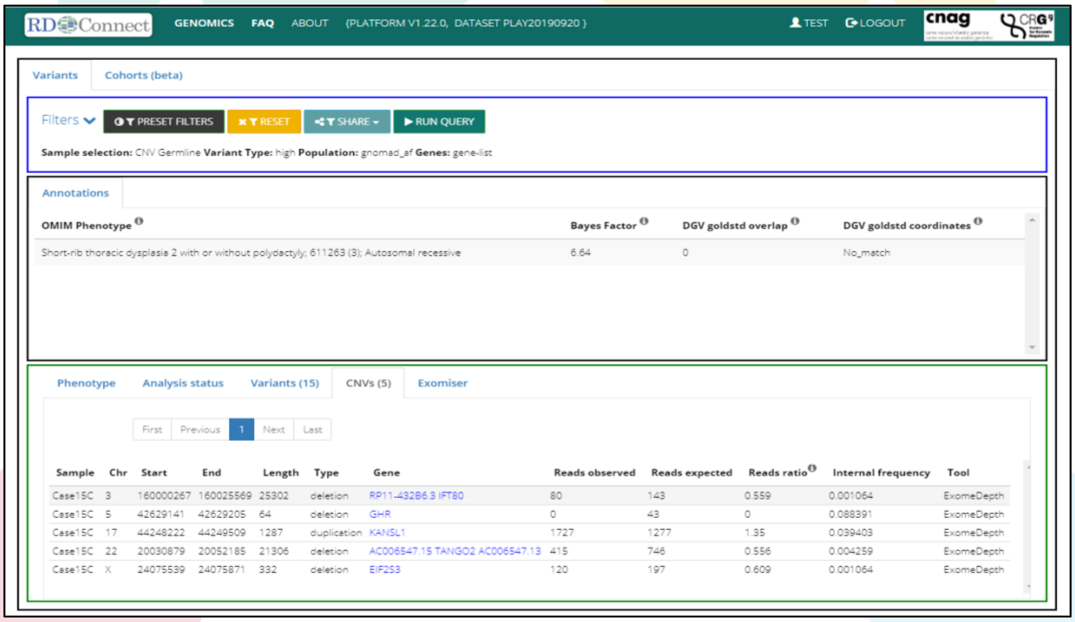
The RD-Connect GPAP (https://platform.rd-connect.eu) is an IRDiRC recognised online platform that facilitates collation, sharing, analysis and interpretation of integrated genome-phenome datasets for Rare Disease diagnosis and gene discovery. During 2019, CNAG-CRG has continued to develop the system based on input from the users. Some of the new functionalities implemented thanks to EJP RD funding include:

- implementation of semi-automated Copy Number Variant (CNV) analysis pipeline (only a few datasets have been processed so far; information is provided for each)
- inclusion of CNV results for exomes and genomes submitted to the GPAP (figure 1)
- integration of Mendelian.co to generate on-the-fly in-silico gene panels (figure 2)
- integration of PanelApp to generate on-the-fly in-silico gene panels (figure 2)

- feature to connect RD-Connect users between them regarding a specific experiment (an email is sent to the user initiating a request on a specific experiment and to its submitter); in collaboration with Solve-RD
- connection with DECIPHER through MatchMaker Exchange API (figure 3)
- implementation of MatchMaker Exchange metrics API
- new use cases have been created to provide training on the GPAP. A total of 15 use cases are now available in the GPAP's playground, with a redesigned homepage (https://playground.rd-connect.eu/) (figure 4)
- EJP RD logos have been included in the RD-Connect GPAP website and playground (figure 4)

Furthermore, the RD-Connect GPAP is participating in the GA4GH Discovery Workstream and partners in the development of the new Beacon API implementation, mainly to realize novel matchmaking functionalities.

Besides the user oriented features, backend changes have been made to improve the code, architecture and scalability (e.g. improved client version control, change to ElasticSearch 6, scripts to upload BAMs and VCFs to Aspera, updates on data tracking, transition to hail for some features, update of annotation sources, prototype for the re-design of the Graphical User Interface, prototype for incremental uploading of datasets).



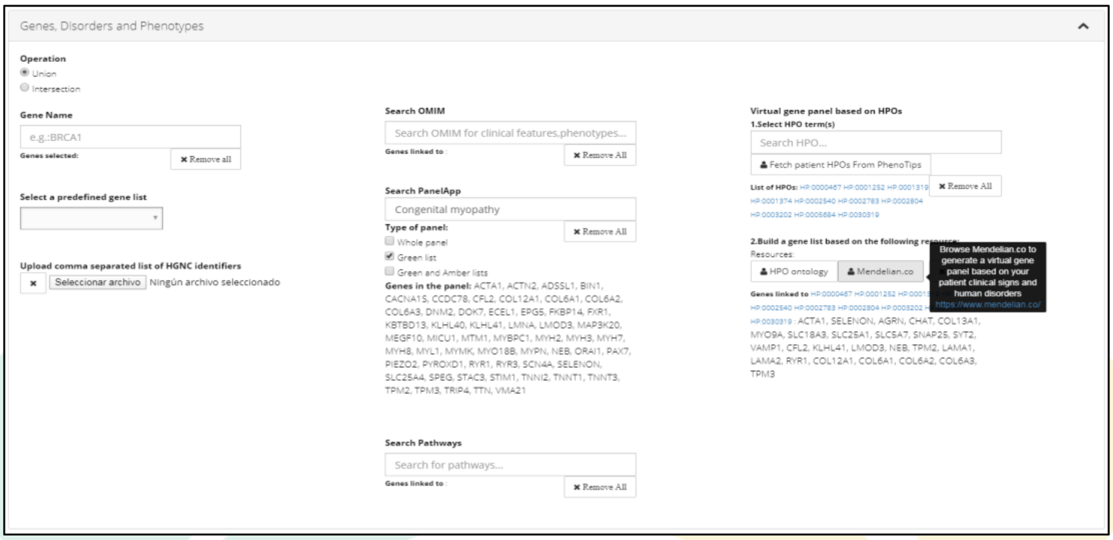
{PLATFORM V1.22.0, DATASET PLAY20190920 }

Sample selection: CNV Germline **Variant Type:** high **Population:** gnomad_af **Genes:** gene-list

OMIM Phenotype	Bayes Factor	DGV goldstd overlap	DGV goldstd coordinates
Short-rib thoracic dysplasia 2 with or without polydactyly; 611263 (3); Autosomal recessive	6.64	0	No_match

Sample	Chr	Start	End	Length	Type	Gene	Reads observed	Reads expected	Reads ratio	Internal frequency	Tool
Case15C	3	160000267	160025569	25302	deletion	RP11-432B6.3 IFT80	80	143	0.559	0.001064	ExomeDepth
Case15C	5	42629141	42629205	64	deletion	GHR	0	43	0	0.088391	ExomeDepth
Case15C	17	44248222	44249509	1287	duplication	KANSL1	1727	1277	1.35	0.039403	ExomeDepth
Case15C	22	20030879	20052185	21306	deletion	AC006547.15 TANGO2 AC006547.13	415	746	0.556	0.004259	ExomeDepth
Case15C	X	24075539	24075871	332	deletion	EIF2S3	120	197	0.609	0.001064	ExomeDepth

Figure 1. CNV results displayed in the RD-Connect GPAP Graphical User Interface.



Genes, Disorders and Phenotypes

Operation
● Union
○ Intersection

Gene Name
e.g.:BRCA1
Genes selected:

Select a predefined gene list

Upload comma separated list of HGNC identifiers

Ningún archivo seleccionado

Search OMIM
Search OMIM for clinical features,phenotypes...
Genes linked to :

Search PanelApp
Congenital myopathy
Type of panel:
□ Whole panel
☑ Green list
□ Green and Amber lists
Genes in the panel: ACTA1, ACTN2, ADSSL1, BIN1, CACNA1S, CCDC78, CFL2, COL12A1, COL6A1, COL6A2, COL6A3, DNM2, DOK7, ECEL1, EPG5, FKBP14, FXR1, KBTBD13, KLHL40, KLHL41, LMNA, LMOD3, MAP3K20, MEGF10, MICU1, MTM1, MYBPC1, MYH2, MYH3, MYH7, MYH8, MYL1, MYMK, MYO18B, MYPN, NEB, ORAI1, PAX7, PIEZO2, PYROXD1, RYR1, RYR3, SCN4A, SELENON, SLC25A4, SPEG, STAC3, STIM1, TNNI2, TNNT1, TNNT3, TPM2, TPM3, TRIP4, TTN, VMA21

Search Pathways
Search for pathways...
Genes linked to :

Virtual gene panel based on HPOs
1.Select HPO term(s)
Search HPO...

List of HPOs: HP:0000467 HP:0001252 HP:0001319 HP:0001374 HP:0002540 HP:0002783 HP:0002804 HP:0003202 HP:0005684 HP:0030319

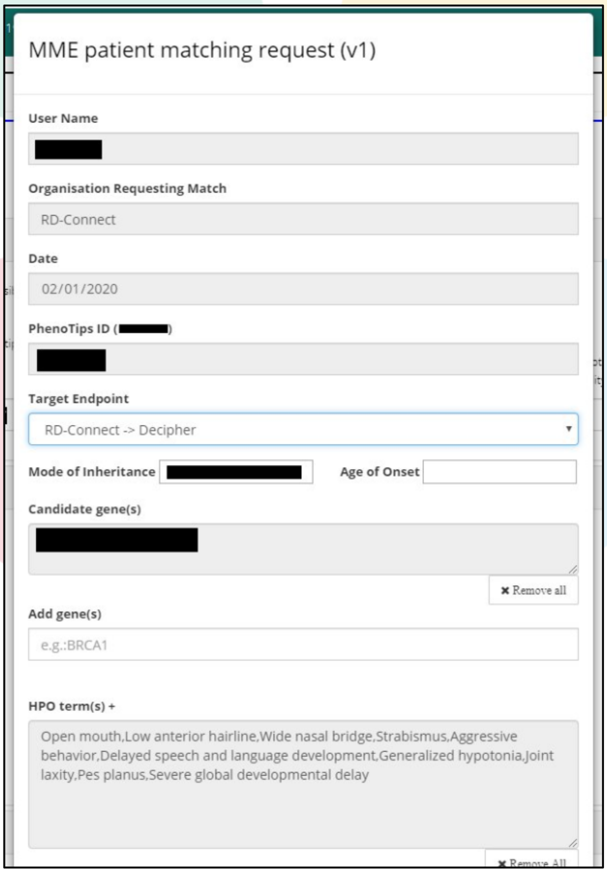
2.Build a gene list based on the following re
Resources:

Browse Mendelian.co to generate a virtual gene panel based on your patient clinical signs and human disorders
https://www.mendelian.co/

Genes linked to HP:0000467 HP:0001252 HP:0001319 HP:0002540 HP:0002783 HP:0002804 HP:0003202 HP:0030319 : ACTA1, SELENON, AGRN, CHAT, COL13A1, MYO9A, SLC18A3, SLC25A1, SLC5A7, SNAP25, SYT2, VAMP1, CFL2, KLHL41, LMOD3, NEB, TPM2, LAMA1, LAMA2, RYR1, COL12A1, COL6A1, COL6A2, COL6A3, TPM3

Figure 2. Screenshot of the RD-Connect GPAP "Genes, Disorders and Phenotypes" filtering panel.

Mendelian.co and PanelApp have been included to generate on-the-fly in-silico gene panels.



MME patient matching request (v1)

User Name
[redacted]

Organisation Requesting Match
RD-Connect

Date
02/01/2020

PhenoTips ID ([redacted]**)**
[redacted]

Target Endpoint
RD-Connect -> Decipher

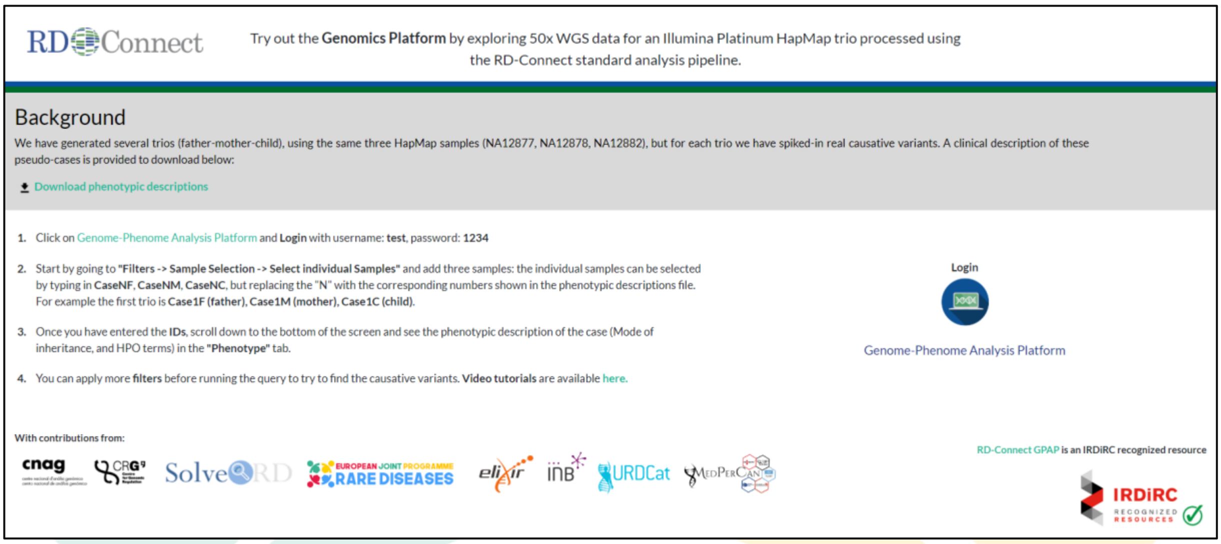
Mode of Inheritance [redacted] **Age of Onset**

Candidate gene(s)
[redacted]

Add gene(s)
e.g.:BRCA1

HPO term(s) +
Open mouth,Low anterior hairline,Wide nasal bridge,Strabismus,Aggressive behavior,Delayed speech and language development,Generalized hypotonia,Joint laxity,Pes planus,Severe global developmental delay

Figure 3. Screenshot of the RD-Connect GPAP's MatchMaker Exchange menu enabling the new connection with DECIPHER



RD-Connect Try out the **Genomics Platform** by exploring 50x WGS data for an Illumina Platinum HapMap trio processed using the RD-Connect standard analysis pipeline.

Background

We have generated several trios (father-mother-child), using the same three HapMap samples (NA12877, NA12878, NA12882), but for each trio we have spiked-in real causative variants. A clinical description of these pseudo-cases is provided to download below:

Download phenotypic descriptions

1. Click on Genome-Phenome Analysis Platform and **Login** with username: **test**, password: **1234**
2. Start by going to **"Filters -> Sample Selection -> Select individual Samples"** and add three samples: the individual samples can be selected by typing in **CaseNF**, **CaseNM**, **CaseNC**, but replacing the "N" with the corresponding numbers shown in the phenotypic descriptions file. For example the first trio is **Case1F (father), Case1M (mother), Case1C (child).**
3. Once you have entered the **IDs**, scroll down to the bottom of the screen and see the phenotypic description of the case (Mode of inheritance, and HPO terms) in the **"Phenotype"** tab.
4. You can apply more **filters** before running the query to try to find the causative variants. **Video tutorials** are available here.

With contributions from:

RD-Connect GPAP is an IRDiRC recognized resource

Figure 4. Screenshot of re-designed RD-Connect GPAP Playground homepage

5.2. 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 there has been a major restructuring of the DECIPHER database and codebase to ensure that the infrastructure fully facilitates the development of new functionality and features for variants identified by genome-wide analysis. In addition, during 2019, the module to support sequence variant classification according to ACMG pathogenicity evidence has included additional information to support this framework from the ClinGen Sequence Variant Interpretation Group and the Association for Clinical Genomic Science. DECIPHER is a founding member of the Matchmaker Exchange and, during 2019, has connected with RD-Connect.

5.3. Tools to predict pathogenicity of genetic variants

INSERM-AMU has been working to improve the tools to predict the pathogenicity of sequence variations. One of the most efficient systems, UMD-Predictor (*Salgado et al.* 2016), uses a combinatorial approach combining knowledge from various layers of information related to nucleotide variations and proteins' changes. To improve the ability to distinguish benign from pathogenic sequence variations, a new bioinformatics system is being developed to extract information from a new biological layer: the protein functional domains' conservation. In fact, most systems only rely on residue conservation at the protein level itself (is this specific residue conserved in this protein over the evolution?). Nevertheless, proteins often contain functional domains that are repeated and/or shared within multiple proteins. The alignment of these functional domains provides a new layer of information, which is complementary to the protein conservation itself (figure 5). Therefore, a new system is being created to:

- rapidly extract information from any missense mutation including a prediction of its impact at the functional domain level,

- give access to all data through a visual interface displaying functional domains and their critical residues.

The efficiency of the pathogenicity predictions from the system alone or in combination with UMD-Predictor using various reference datasets is currently being evaluated. The analysis of mutations localized in functional domains allowed to define two principal dimensions as illustrated in figure 6. Their combination (yellow area) is predicted to contain pathogenic mutations. The pathogenicity of mutations localized in the white area cannot be predicted by this approach. A preliminary analysis of mutations localized in functional domains from various UMD locus specific databases (http://www.umd.be) or ClinVar (http://clinvar.org) indicates:

- a very high efficiency of this approach to predict the pathogenicity of a subset of mutations localized in functional domains,
- an independence from other existing prediction tools,
- a synergy with the UMD-Predictor system.

In conclusion, the addition of the new "functional domains" granularity in the combinatorial landscape of pathogenicity prediction tools should be an added value to improve the efficiency of existing systems such as UMD-Predictor. In addition, for a subset of missense mutations, the system might provide a new level of predictions that could move those in silico predictions to a higher-ranking category from the ACMG classification. This might allow an easier classification of variants in clinical contexts and ultimately lower the number of unsolved cases.

The next steps will include the fine tuning of parameters and thresholds; the optimization of the user interface; the creation of various API. The corresponding website and tools should be made available to the public by Q4 2020.

Complementing the previously described work, the EBI is currently working to improve support for variants impacting splicing in the Ensembl Variant Effect Predictor (VEP) by integrating results from the Illumina SpliceAI tool. This should be available for command line use in January 2020 and via REST in April 2020. The latest variant pathogenicity prediction algorithms will be investigated and any which compliment or improve upon the currently supported tools will be integrated.

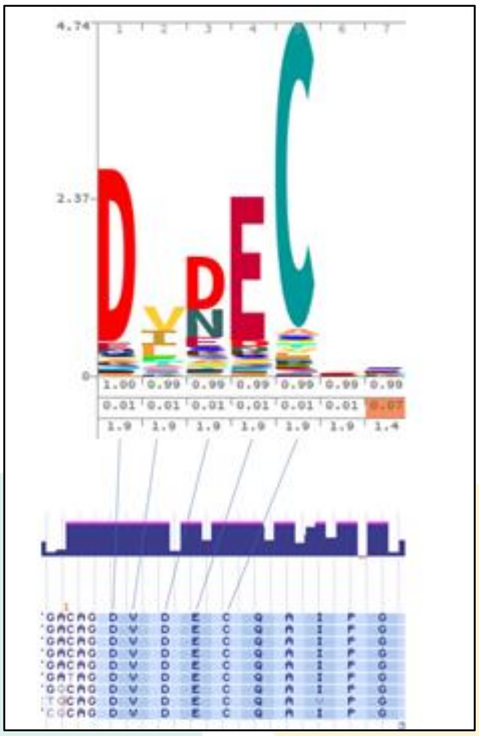


Figure 5. FBN1 calcium binding EGF-like (cb-EGF-like) #1 domain conservation.

Top: at the functional domain level (cb-EGF-like from PFAM). Bottom: at the protein conservation level (UCSC). As illustrated in this figure, while the conservation between species indicate that all residues are highly conserved (almost 100%), their alignment at the functional domain reveals that they are not equivalently important as positions 6 & 7 do not have any functional importance, while position 5 is critical and other positions have various levels of importance

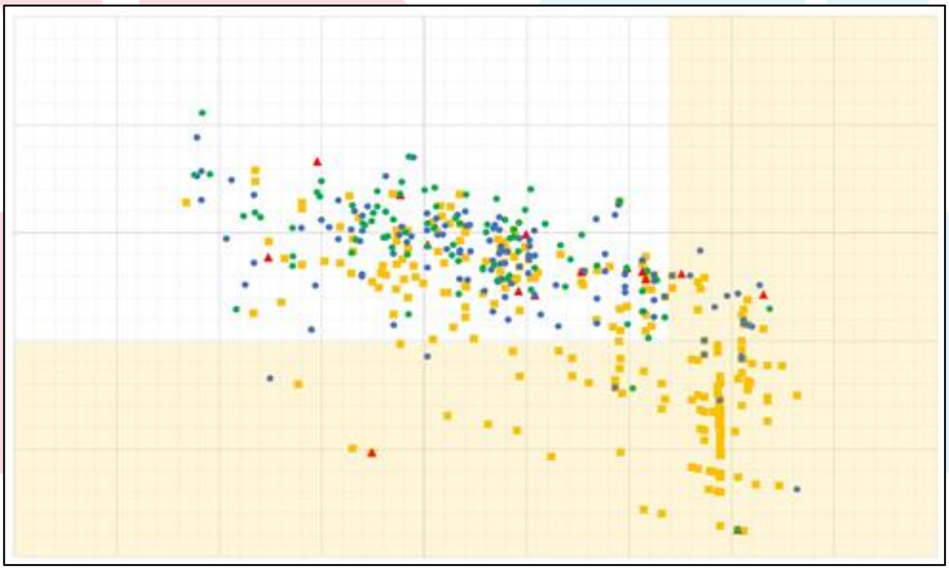


Figure 6. FBN1 mutations localized in functional domains extracted from the UMD international database and GnomAD

Yellow squares and red triangles = pathogenic mutations; green and blue circles = benign mutations.