

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 11.17

Second 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, INSERM-RaDiCo]; 4-LBG (LBI-RUD); 35-UMCG, WSI-DECIPHER (Associated Partner); 65-Radboudumc; 64-LUMC & LUMC-Endo-ERN; 25-FTELE; 73-BBMRI-ERIC; 26-ISS; 36-UM; 82-ACURARE; 44-ISCI; 47-CIBER; 18-UKL-FR.

Due date of deliverable: month 24

Dissemination level: Public

Table of Contents

1. Introduction	2
2. User-friendly genomics analysis	3
2.1. RD-Connect Genome-Phenome Analysis Platform	3
2.1.1. New functionalities developed.....	3
2.1.2. New datasets analysed	5
2.2. DECIPHER	6
3. Cloud computing and multi-omics analysis	7
3.1. New cloud functionalities	7
3.1.1. Cloud infrastructure	7
3.1.2. Functionalities and pipelines	7
3.2. Multi-omics use cases analysed	10
4. Information and annotation resources	12
4.1. Tools to predict pathogenicity of genetic variants	12
4.2. The Ensembl Variant Effect Predictor	15

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 2020 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.

Overall, the RD-Connect GPAP currently has 209 groups with a total of 432 authorized users.

2.1.1. New functionalities developed

During 2020 new developments were introduced in the RD-Connect GPAP that have positively affected the user experience. Below is a brief description of the most relevant ones funded by EJP-RD:

- Work on increasing the efficiency of the RD-Connect GPAP data workflow and on minimising potential human errors by automating many processing steps, and by making relevant changes to reduce computing time was performed.
 - On the data submission and processing side, the system was made much more efficient by:
 - automating the whole processing pipeline, which is now able to detect new sequencing data files in the RedIris (Spanish academic and research network that provides advanced communication services to the scientific community and national universities; www.rediris.es) Aspera server;
 - checking if the uploaded files are complete according to the provided metadata by the user, and downloading them to the computing cluster;
 - transforming them back to FASTQ (if starting from a BAM or CRAM) and evaluating their quality;
 - trimming the sequencing adapters, mapping the reads to the reference genome and doing the variant calling to generate a gVCF file. If any of the steps fail, a warning is issued.
 - The other major changes are related to the improvement of the ElasticSearch query, the database structure and the ETL (Extract, Transform, Load) pipeline, which includes variant annotation. These changes have been made to enable incremental uploads, substantially reducing the time between data processing and data release in the platform, because it is now not necessary to consolidate the gVCFs from all the samples, a process which is very computer intensive. The introduction of modifications in the WES/WGS processing pipeline will be maintained to further automatize it, with the aim to complete the work in 2021.
- The “Search across all” functionality (figure 1) allows users to query for variants in the whole RD-Connect GPAP database. However, as the number of datasets

increased, the time-response of the queries also increased, affecting the performance of the tool. Therefore, some internal modifications were introduced to importantly reduce the time-response of the “Search across all” feature. First, the structure in which data was stored was changed. Instead of having one main Index table, multiple smaller Index tables were created, which allow the system to perform multiple searches at the same time. Additionally, the search does not need any more to query the whole database, but only those Index tables that have the requested information. Finally, changes in the filtering order have also been implemented. All these modifications have importantly reduced the time-response of the “Search across all” functionality, which has provided a much better user-experience.

- The “Search across cohort” functionality was introduced (Figure 1), in which users are able to create cohorts based on ERNs or Phenotypic information and perform a query through the GPAP-Analysis module only to the participants in this cohort.
- A new RD-Connect GPAP API (Application Programming Interface) for programmatic access was developed (Figure 2). This new API allows to do direct queries to the platform through a Python client. This provides a more flexible search compared to the filters available through the standard User Interface. In addition, this API has a login access, so the data accessible by the user is restricted to the same data (s)he can access through the RD-Connect GPAP GUI (Graphical User Interface). The API is being tested internally and is not yet accessible to the all the users.
- Exomizer has been updated to v12.0, which better annotates, filters and prioritizes likely causative variants.
- ClinVar has been updated to the version released in June 2020, which performs a more accurate filtering of variants compared to the previous version of January 2020. The frequency of ClinVar updates is planned to be increased in the future.
- A new RD-Connect GPAP Playground (<https://playground.rd-connect.eu/>) has been released. This Playground contains the full RD-Connect GPAP Analysis and RD-Connect GPAP-Phenostore modules with mock-up data. This development, freely accessible, allows potential users to try the different tools and resources available inside RD-Connect GPAP, which will help them understand the potential it has for their research projects before registering to the GPAP. The Playground is used often in training sessions and workshops; in 2020 this included 3 events organised by EJP RD:
 - Training on strategies to foster solutions of undiagnosed rare disease cases (online meeting - April 2020)
 - EJP RD General Assembly (online meeting – September 2020)
 - 2nd Training Course on “Quality assurance, variant interpretation and data management in the NGS diagnostics era (online meeting – October 2020)

Note: RD-Connect GPAP developments and features related to data submission, management and deposition from the user perspective are reported in the Deliverable 11.12 “Second version Additional facilities integrated to resources

regarding data deposition and access, including user guidelines and documentation”.

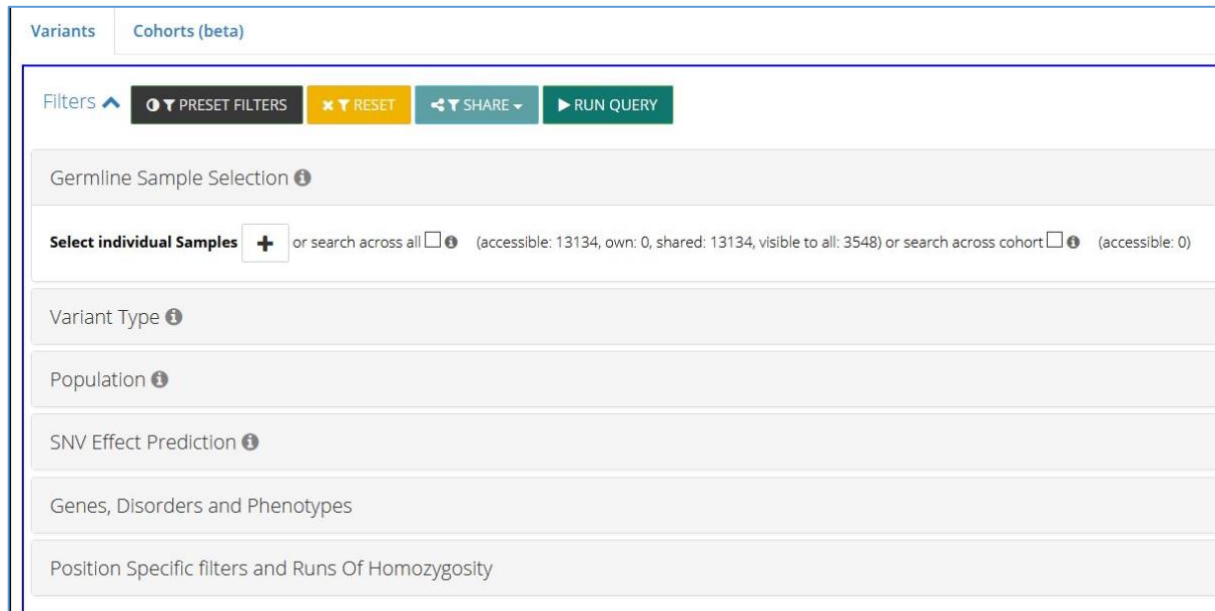


Figure 1. Screenshot of GPAP Analysis module showing the “Search across all” and “Search across cohort” features.

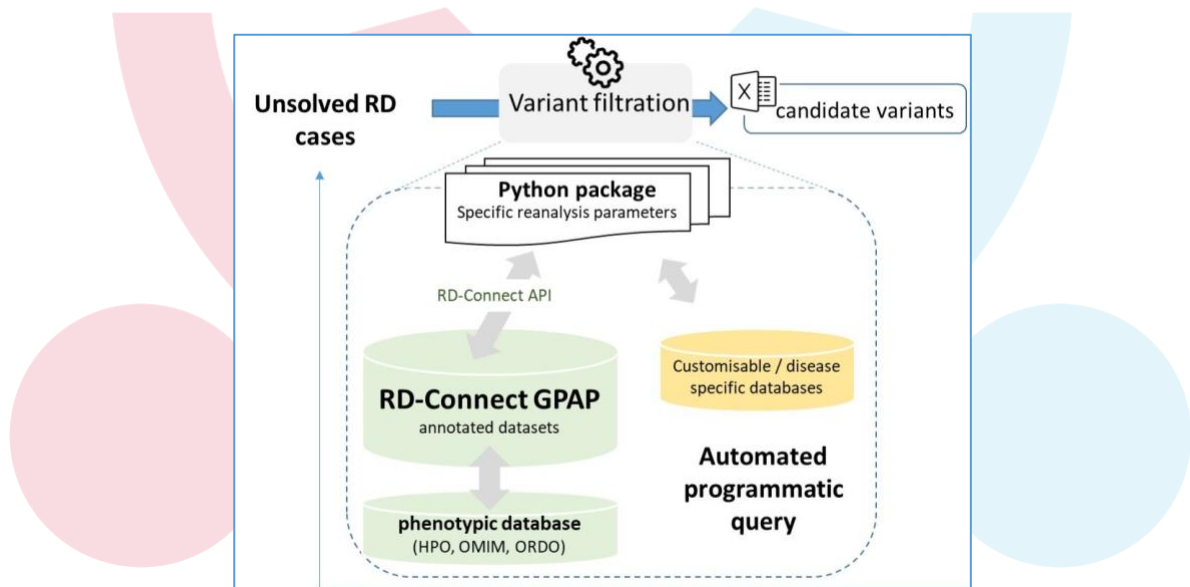


Figure 2. Summary of the GPAP API for programmatic access workflow.

2.1.2. New datasets analysed

RD-Connect GPAP has increased the number of processed datasets during 2020. In 2019 the GPAP included 11.526 processed datasets (exomes, genomes and panels).

In December 2020 this amount has increased by 1.608 processed datasets, currently containing 13.134 (12.377 exomes, 640 genomes and 117 panels) with their corresponding phenotypic information.

From the 1.608 new datasets processed in RD-Connect GPAP during 2020, 185 have been submitted by RD-Connect GPAP users from across Europe, which were processed and made available. The GPAP has also collated and processed 1.423 datasets from European registered users which 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.

A major new version of DECIPHER was released in June 2020. Previously DECIPHER supported the deposition, interpretation and sharing of sequence variants and copy-number variants. The new version supports aneuploidy/segmental aneuploidy, uniparental disomy, short tandem repeats, inversions and insertions (including mobile element and retrogene insertions). Interpretation interfaces are provided for each type of variant. The new version also enables the grouping of variants, allowing the representation, for example, of compound heterozygous variants or rare pathogenic haplotypes.

Tools to assist in the interpretation of splice site variants have also been improved, SpliceAI scores are now displayed for deposited variants and DECIPHER/ClinVar splice region/site variants are displayed on the protein browser.

In addition, DECIPHER has incorporated further information from the ClinGen Sequence Variant Interpretation Working Group regarding recommendations for interpreting the ACMG pathogenicity evidence criteria, including an interactive flow chart to determine the strength of the loss of function PVS1 criterion.

3. Cloud computing and multi-omics analysis

3.1. New cloud functionalities

3.1.1. Cloud infrastructure

The UMCG has set up four different cloud instances [<http://docs.gcc.rug.nl/>], including two environments (Talos and Hyperchicken) specifically meant for infrastructure development. The 'Gearshift' cloud, located at the UMCG is a cloud environment where compute resources are shared between projects, allowing for more efficient use of compute nodes that are available for all projects on the cloud and less idle time, benefiting all users. A separate EJP RD environment has been added to the Gearshift cloud to provide functionalities to EJP RD associated projects. Other cloud environments can be set up on request to benefit the wider European research community. Current login procedures are based upon an SSH public-private keypair and make use of a jump-host to ensure safe access. For the 'Fender' cloud hosted at the EMBL-EBI Embassy and funded by Solve-RD through EBI, it is used as a use-case to pilot large-scale collaboration between researchers around Europe with many samples. Storage is funded by the use-case projects.

3.1.2. Functionalities and pipelines

The UMCG cloud environment consists of four layers, the data layer, tool layer, meta layer and view layer. Since the deployment of the cloud environments is designed to be reproducibly deployed on different sets of equivalent hardware, all functionalities and pipelines can be easily added to all instances of the cloud if not already available. The data, tool and view layer are currently available to the EJP RD community. The meta layer is currently piloted in the Solve-RD use-case. Instances for other groups datasets can be set up on request.

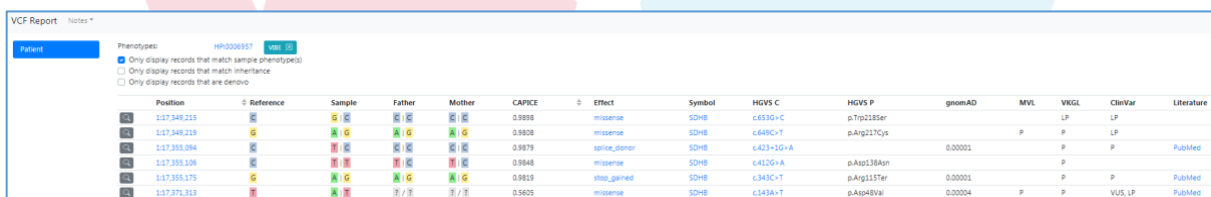
The data and view layer are supported by the High-Performance Computing (HPC) Cluster environment that was expanded to further support omics analysis. Eight tools for CNV and SNV-indel calling and interpretation (ngs-bits, ExomeDepth, Conifer, ClinCNV, Capice, Scramble, Lirical and AnnotSV) have been added to the Easybuild toolchain [<https://github.com/molgenis/easybuild-easyconfigs>]. Moreover, a new GATK4-based NGS DNA analysis pipeline [https://github.com/molgenis/NGS_DNA] and an RNA analysis pipeline [https://github.com/molgenis/NGS_RNA] have been updated and implemented using the Molgenis Compute workflow language. All tools and workflows can be reproducibly added to any instance of the cloud environment (including future instances). Because the UMCG cloud environment is set up for research purposes new tool versions will not automatically be updated but can be added upon request.

For the meta-layer, in addition to the HPC environment the meta-database (RD3) has been expanded so that it can handle patches of datasets. For instance, if new phenotypes are seen in patient, or found to be absent, this information can be handled in a patch, leaving the metadata that was available for earlier analyses intact. Furthermore, support for other omics data is being implemented. Experiment information is different for the various omics and experiment types. Next to WES and

Whole Genome Sequencing (WGS), firstly, support for RNA sequencing data is being implemented. For all implementations RD3 is made compatible with common data elements (CDE) and (upcoming) standards are connected, such as the Global Alliance for Genomics and Health (GA4GH) approved Phenopackets. For the future the plan is to make the model compatible with the EJP RD record level (CDE) model by implementing a semantic endpoint. As an extension, the possibility to connect RD3 to the query builder Work Focus is being explored.

ELIXIR AAI (Authentication Authorization Infrastructure) login has been added to RD3; the addition of this way of access to the HPC cluster is being performed.

A view layer is added to the cloud to be run on command line or through a web service. A Variant Interpretation Pipeline (VIP) has been developed, currently focused on genetic variants, of which the latest release is version 1.4.1 [<https://github.com/molgenis/vip/releases/>]. This modular pipeline enables the addition of analysis specific building blocks. Currently such blocks are being added, based upon existing resources, such as the Ensembl Variant Effect Predictor (see section 4.2). The pipeline consists of three steps, annotation, filtration/prioritization and reporting. The input of the pipeline is an unannotated VCF file. Using phenotypic terms coded in an ontology (e.g., HPO) patient specific variant prioritization can be performed. Different outputs showing prioritized or filtered variants can be created, including an interactive html file that can be viewed in a web browser. Future developments may include other omics data types, such as RNA sequencing data to give information on splicing or expression effects and further aid prioritization. VIP is available within the UCMG cloud instances, but is readily available for the wider community, because releases can also be downloaded and used locally (Figure 3). Releases will remain available on GitHub, but new functionalities, such as copy number variant prioritization, will be added to future releases.



Position	Reference	Sample	Father	Mother	CAPICE	Effect	Symbol	HDV5 C	HDV5 P	gnomAD	MVL	VKGL	ClinVar	Literature
117,349,215	C				0.8893	missense	SDHB	<G53G>C	p.Trp218Ser			LP	LP	
117,349,219	G				0.8808	missense	SDHB	<G49C>T	p.Arg217Cys		P	P	LP	
117,351,084	C				0.9879	splice_donor	SDHB	<A21+1G>A		0.00001		P	P	PubMed
117,351,106	C				0.9848	missense	SDHB	<A12G>A	p.Arg158Asn			P	P	PubMed
117,351,175	C				0.9819	miss_gained	SDHB	<348C>T	p.Arg117Ter	0.00001		P	P	PubMed
117,371,313	C				0.9505	missense	SDHB	<143A>T	p.Arg49Ile	0.00004	P	P	VUS, LP	PubMed

Figure 3. View of mock-up data in VIP interactive html output

Cloud functionalities are developed in connection to other EU funded projects. With funding from the CINECA project a direct connection with the EGA archive has been established through a Filesystem in User-Space (FUSE) client. New developments on cloud functionalities are piloted with several use-cases, including the Solve-RD project in which currently over 8,000 Whole Exome Sequencing (WES) samples are being reanalyzed.

New data processing and analysis tools and pipelines and their implementation are funded through EJP RD. This includes the exploration of implementing pipelines using different workflow specification languages, (e.g., Molgenis compute and Common Workflow Language (CWL), and workflow managers (e.g., Nextflow), which might

have their own pseudo-language to specify the workflow. As analysis reproducibility and provenance is important to build trust among stakeholders, the focus is also on using software containers (e.g., docker and singularity), and provenance mechanisms (e.g., RO-Crate). Indeed, current plans include the exploration of mechanisms to access available workflows in the EOSC-Life WorkflowHub infrastructure (<https://workflowhub.eu/>), using the GA4GH TRS (Tool Registry Service) and materialize the workflow (e.g., associated software containers and any necessary data) in the jumping/login node, prior to the proper execution in the cloud facilities. Execution will be performed using software compatible with the GA4GH WES (Workflow Execution Service) and TES (Task Execution Service) depending whether the workflow is run locally to a specific cloud instance (the expected scenario) or is distributed across different computer systems, which is unexpected considering the existing limitations imposed by EU GDPR on the management of sensitive data. This enable using workflows in computational infrastructures that are not directly connected to the Internet. This effort will ensure using state-of-the-art analytical workflows maintained by the Life Science community and its rapid uptake by EJP RD. In addition, secure login and data access are funded through EJP RD as well as data FAIRification including RD3 development prioritizing standards from GA4GH and ELIXIR.

In this sense, LBI-RUD worked on analyses to be compatible with the workflow definition language (WDL) and the Task Execution Service (TES) specification supported by GA4GH. These specifications decouple the executable workflow from the physical server infrastructure and provides the flexibility to move defined analyses between cloud-based systems or compute clusters. For development and testing purposes, a TES compatible environment was created on the local compute infrastructure. The previous established WES processing pipeline for the genome reference 37 (GRCh37) was made fully compatible and tested with the WDL specification. This pipeline was further extended to enable the detection of copy number variation (CNV), also implemented in WDL format and compatible with TES. In addition, the publicly available WES for the genome reference 38 (GRCh38) provided by GATK (<https://github.com/gatk-workflows/gatk4-exome-analysis-pipeline>) was selected to test the TES compatibility of the internal infrastructure.

Besides the cloud IT described above, EJP RD WP13 is also working with another cloud environment to analyse the initial use cases (see section 3.2). This is the anDREa (AZURE) cloud computing environment. The reason for this was to avoid any delays on using the actual EJP RD use case data while the cloud IT described above and implemented specifically for EJP RD was evaluated regarding EU GDPR.

BSC is developing a high-level workflow execution service (WfExS) backend as part of their involvement in EJP RD. First implementation iteration of this development is supporting both CWL; Nextflow workflows support is ongoing. Workflow to be executed has to be available either in a git repository or findable at a GA4GH TRSv2 service which supports describing the workflows through RO-Crate (currently WorkflowHub instances).

Backend receives a list of parameters, which can be either inline parameters or inputs, and a list of outputs, which will be either proposed filenames for those outputs or wildcard patterns to match local filenames. Input parameters are described through URIs, currently standard URLs, and in a future iteration some CURIE namespaces from identifiers.org / n2t.net which are going to be supported (for instance, EGA files / datasets). These inputs are downloaded and cached (whenever it is allowed), in order to avoid downloading the very same copy of the workflows or reference genomes, for instance. Supported workflow execution engines are also downloaded and installed, using the more adequate versions, as there are some workflows which report the minimal version of the engine to be used (for instance, Nextflow ones). All the workflow steps have to be based on public docker or singularity containers. The kind of environment where the workflows from EJP RD and other projects are going to be run was considered; the use of singularity runtime was decided (docker container images can be re-assembled as singularity ones when pulled), this keeps open the options to run the workflows both in HPC and cloud environments. For the very same reason, preconditions materialization phase, which is where containers are fetched, is detached from the execution one, as this last can have restricted internet access.

Some features of the minimum viable product milestone being pursued include trusted, secure and reproducible workflow execution using EGA files and datasets. In order to provide all the needed hints to have reproducible executions, output metadata from the WfExS backend is an RO-Crate with all the execution provenance: concrete repository checkout hashes, concrete engine used, complete list of inputs (even implicit ones). But, at the same time, a secure execution is going to be achieved using FUSE encfs encrypted directories for intermediate results, and final results are going to be encrypted using crypt4gh GA4GH standard and the public keys of the researchers, so the results can be safely moved outside the execution environment through unsecure networks and storages.

3.2. Multi-omics use cases analysed

Three selected multi-omics use cases were selected by WP13:

- "INCPH" Pagan et al. (Barcelona) (ERN RareLiver): differential diagnosis between Idiopathic Non-Cirrhotic intrahepatic Portal Hypertension (INCPH) and cirrhotic portal hypertension. Metabolomics of serum and proteomics/transcriptomics of liver biopsies, which include liver tissue from 21 INCPH and 22 patients with cirrhosis (CH) as well as from 15 Histologically Normal Livers (HNL) as controls for transcriptomics. Serum from 30 INCPH patients, 34 patients with cirrhosis and 33 healthy individuals were used for metabolomics. Data Transfer Agreement (DTA) is finalized by the end of 2020 and presumably data will be deposited in the beginning of 2021.
- "CAKUT" Schanstra et al. (INSERM - Toulouse) (ERN ERKNet): identification of disease modifiers/pathways for congenital anomalies of the kidney and urinary tract (CAKUT) during pregnancy (miRNA, proteomics, peptidomics, metabolomics data available). Data generator provided miRNA, proteome and peptidome data for a total of 162 samples; however not all are matching

samples and not all 'omics data sets have data for all 162 samples. Data provider has already analysed peptidome data.

- "IBM" Udd et al. Helsinki (Helsinki) (ERN EURO-NMD): modifiers of disease severity in Inclusion Body Myositis (IBM) (Genomics, transcriptomics, miRNA, lncRNA and other ncRNA data sets are available). For all RNAseq data, except miRNA, data for a total of 33 samples (24 patients vs 9 controls) were provided. There are 18 patient and 9 control samples that have miRNA data, and whole exome data set consist of 23 patients with no controls. Data provider has already analysed mRNA and miRNA data sets.

Currently these use cases are being analysed within WP13 using the anDREa (AZURE) cloud computing environment. Data from two use cases are already deposited at the cloud environment; however, for INCPH case study only the DTA has been signed by the end of 2020. Since data transfer agreements and user code of conducts have been signed and processed for the anDREa environment, analysis will continue to be performed on this location. The relevant workflows and analyses performed on this cloud will be implemented in the UMCG cloud environment described in section 3.1 and run with publicly available datasets known within WP13 to demonstrate that the functionality needed for the use-cases is present. However, for the use cases included in the next phase of WP13 all DTA agreements, infrastructure, data access and required research environments will be managed by UMCG cloud administrators.

Besides the use cases data, several samples have been analysed as part of the pipeline and infrastructure development, implementation and testing. The UMCG has analysed 191 targeted gene panel samples from patients with familial cancer, plus 45 positive controls with likely pathogenic or pathogenic variants matching the patient or family's phenotype, using the Variant Interpretation Pipeline, described in section 3, as well as 77 WES samples from patients with Hypoplastic Left Heart Syndrome (HLHS). These were in-house samples used for pipeline / technology development. The current developments are focused on automatic variant classification and prioritization in the context of a patient's phenotype, ideally resulting in only a few possible causal variants to be manually reviewed by a clinician. The two datasets constitute two use-cases: the first with a focused approach, likely resulting in a single hit, and the second with a wider approach in which a vaguer result is expected and as such is more challenging. The LBI-RUD has analysed 68 samples with the GRCh38 WES pipeline to test the internal TES development compute infrastructure. Furthermore, 72 WES samples were analysed with the updated WES GRCh37 pipeline including the copy number variation (CNV) in order to ensure GA4GH compatibility. All processed samples were in-house samples and analysed for pipeline development purposes.

4. Information and annotation resources

4.1. Tools to predict pathogenicity of genetic variants

During the collection of ACMG-AMP evidence ([Richards et al. 2015](#)) to classify a mutation as benign (class 1), probably benign (class 2), of unknown significance (class 3), probably pathogenic (class 4) or pathogenic (class 5), some of the 28 criteria are more frequently used than others because of their availability (Figure 4).

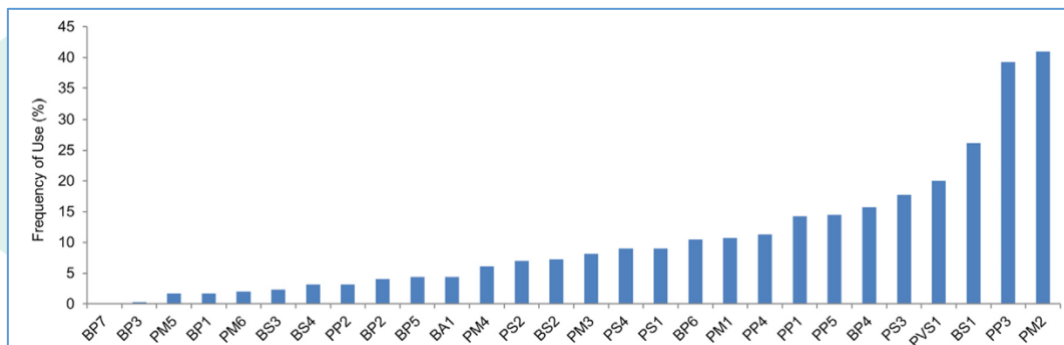


Figure 4. Frequency of Use for Each ACMG-AMP Line of Evidence (from Amendola et al.) (Amendola et al. 2016)

The AMU team has developed the UMD-Predictor ([Salgado et al. 2016](#)) and the Human Splicing Finder ([Desmet et al. 2009](#)) systems that are considered as international references for splice site motif identification (HSF) and pathogenicity prediction (HSF and UMD-Predictor).

Despite their accuracy: 85% for UMD-Predictor (Figure 5), and from 85% for ESE/ESS to about 100% for branch points and splice sites for HSF, these systems can still be improved.

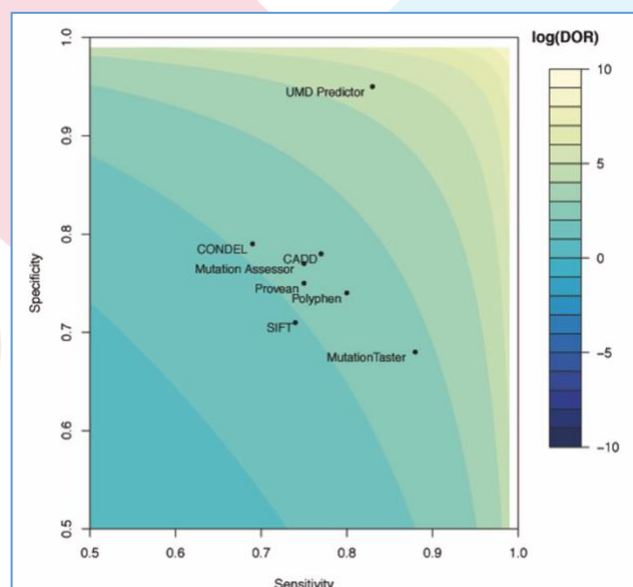


Figure 5. Comparison of the log(DOR) of 7 widely used in-silico pathogenicity prediction systems (CADD, CONDEL, Mutation Assessor, MutationTaster, Polyphen2, Provean, SIFT). Note the log(DOR) 2 logs improvement of UMD-Predictor over the other systems on the Varibench (Nair and Vihinen 2013) + dbSNP (Wheeler et al. 2007) dataset of 17,329 classified variants.

This can be done through the addition of new layers of annotations for UMD-Predictor that uses a combinatorial approach.

It was therefore decided to create the DOLPHIN system to extract additional information from protein domains (PD), this being independent from the protein conservation itself as shown in figure 6.

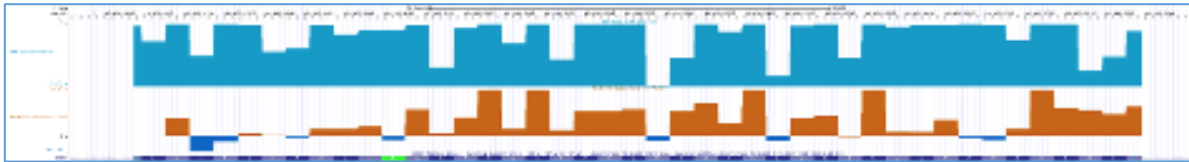


Figure 6. Comparison of protein conservation information provided by the 100-species conservation from UCSC (blue track) to the protein domains conservation score (brown track) for the Calcium Binding Domain #10 of the fibrillin-1. Top = genomic reference sequence (HG38); Bottom = amino acids contained in the domain

To do so, all PD were extracted from eukaryotes and quantitative parameters were defined to characterize each mutation. They can further be used to distinguish pathogenic and non-pathogenic mutations from ClinVar ([Landrum et al. 2020](#)). Currently, a work is being performed on the application of machine learning approaches based on KNN (K-Nearest-Neighbors) to optimize the ability of the system to distinguish benign mutations from pathogenic mutations (Figure 7) in order to generate an innovative pathogenicity prediction score.

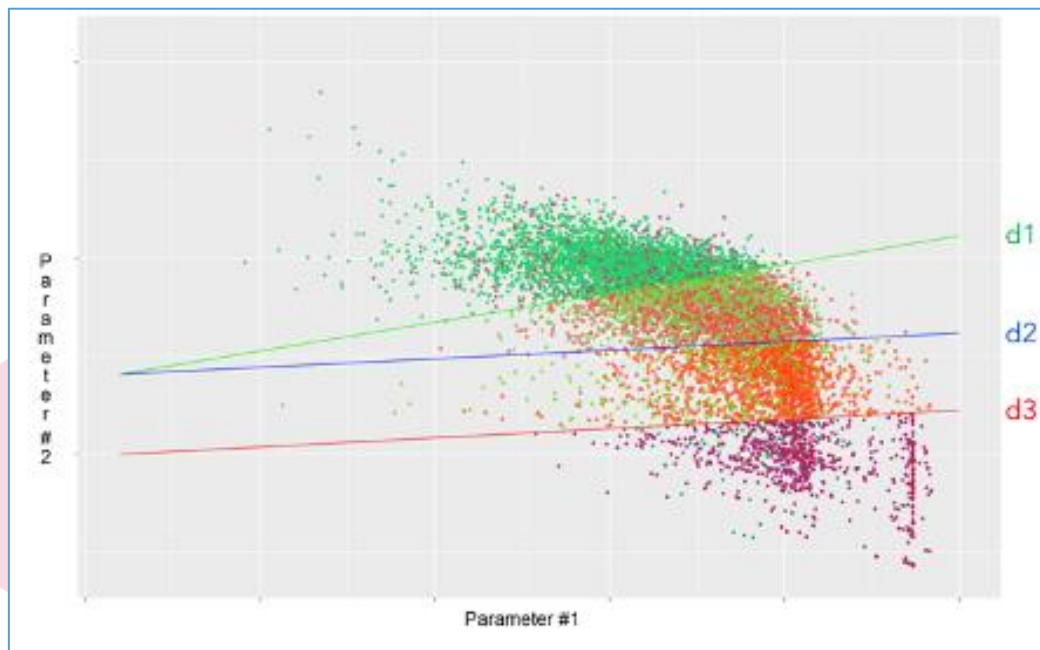


Figure 7. Distribution of pathogenic (red) and benign (green) mutations from ClinVar according to the parameters #1 & #2 extracted from protein domains. The d1, d2 and d3 lines allow to define regions with high confidence of pathogenicity or neutral mutations.

Preliminary data demonstrated that this new layer of information can efficiently predict the pathogenicity of disease-causing mutations localized in PD with a positive

predictive value of 0.85 vs. 0.77 to 0.85 for the most accurate predictors (UMD-Predictor, Provean and REVEL).

As these predictions are only available for mutations localized in PD, the potential benefit of adding this information to other predictors was evaluated. Interestingly, findings suggest that it will mostly benefit UMD-Predictor, which is already one of the most accurate predictors (figure 8).

DOLPHIN accurately predicts pathogenic mutations corresponding to 271 false negative mutations from UMD-Predictor, 68 for Mutation Taster, 51 from REVEL and 7 from CADD. At this stage, the training process of some predictors that may introduce bias in this analysis is not considered. As UMD-Predictor is a unique combinatorial system that does not require training, these results are highly relevant for this system.

The next step will be to evaluate how this new layer of information can be included in UMD-Predictor to improve its accuracy. In theory, if the addition of false negatives, accurately predicted by DOLPHIN, is only considered, a 10% improvement on the Matthews correlation coefficient (MCC) is observed. Nevertheless, this does not take into account positive synergistic effects on common false negatives or negative synergistic effects.

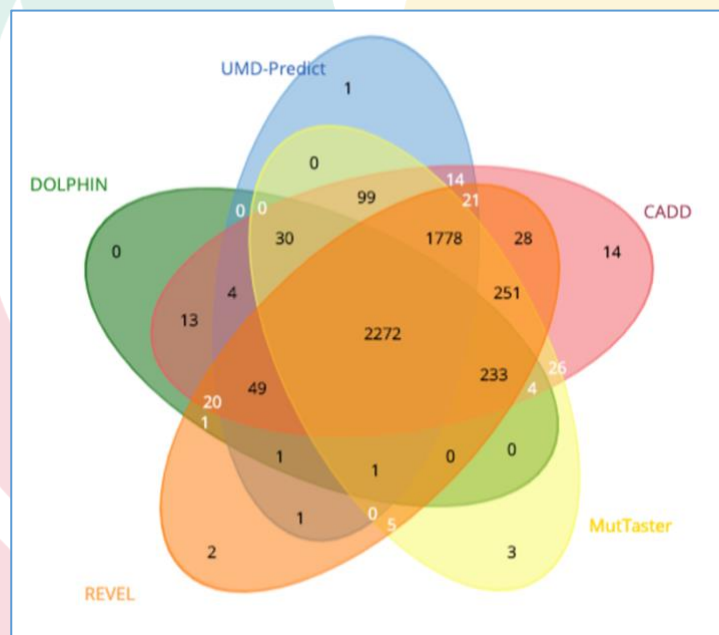


Figure 8. Venn diagram of true pathogenic mutations predicted by REVEL, Mutation Taster (MutTaster), CADD, UMD-Predictor and DOLPHIN.

In conclusion, the new "functional domains" granularity in the global landscape of classification of mutations may strongly improve the ability to use the ACMG/AMP PM1 criteria "Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation". That is today difficult to use because of its highly variable interpretation between laboratories and users. It was demonstrated that the DOLPHIN system allows revisiting this PM1 classification criteria thus allowing an easier classification of amino acid substitutions localized in protein domains that account for almost 40.2% of proteins and where most

pathogenic mutations are located. A manuscript is now under preparation and should be submitted early next year. Concomitantly the DOLPHIN website will be opened to all researchers to access this information.

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. The enhancement of its options for rare disease analysis functionalities has been continued. SpliceAI (<https://doi.org/10.1016/j.cell.2018.12.015>) has been reported to be an effective tool for the accurate identification of genomic variants which disrupt normal pre-mRNA splicing (<https://doi.org/10.1101/781088>). In response to requests from users of Ensembl VEP, a pipeline to calculate SpliceAI predictions for new MANE transcripts as they are created was developed. Results from this new analysis will first be available early 2021. A VEP extension to integrate results from the DisGeNET database of disease associations was also created and results were made available via the command line, REST and web interfaces. These updates were presented at the Genomics of Rare Disease Conference 2020.