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

Deliverable 13.2

Completed pathway analysis workflow (data analysis including both types of genetic variant linking and network creation)

Including

AD 50: Case study specific (proof-of-concept) and generic multi-omics analysis pipelines as part of subtask 13.1.9 and in alignment with the deliverable D11.19 and subtask 11.4.2

And

AD 51: Report for training purposes for Pillar 3 based on the workshops, analyses and VP deployment from pilot case studies

Organisation name of lead beneficiary for this deliverable:
36 – UM

Project partners involved: 65 – Radboudumc, 64 – LUMC, 76 – ELIXIR/EMBL-EBI, 1 – INSERM-AMU, 82 – ACURARE

Authors of the deliverable report: Friederike Ehrhart, Cenna Dornboos, Ozan Ozisik, Aishwarya Iyer, Sarah Hunt, Nazli Sila Kara

Due date of deliverable: month 48

Dissemination level: public

Table of Contents

1. Summary

In this deliverable we summarise the progress of multi omics data analysis using molecular pathways for rare diseases. We have identified 4 case studies to develop (multi) omics analysis workflows, including both data-driven and prior knowledge-(pathway) driven methods. The datasets contained at least two, but different combinations of metabolomics, proteomics, peptidomics, transcriptomics and genetic variants (whole exome sequencing) data. Those analysis methods are using combinations of both pathway and network approaches, so we are reporting large parts already ahead for the next deliverable 13.3 "*Completed network analysis workflow (active node detection, lifestyle factor network evaluation and extended network analysis for drugs and toxic compounds*)".

For each of the case studies we have at least one publication in preparation, workflows available on GitHub or WorkflowHub, and at least one pathway. The pathways on WikiPathways are already connected to the virtual platform and the workflows on WorkflowHub are being prepared to connect to the virtual platform.

2. General introduction

This deliverable shows the progress in rare disease (multi)omics data analysis on a group of defined case studies – and beyond. The original idea for this work package was based on the idea that due to the lack of a large number of patients and samples, especially, rare disease data analysis profits from extrinsic data analysis – extrinsic meaning including prior knowledge from known molecular interactions (pathways, see also **D13.1**) and well annotated genetic data.
During the project work progression, the collaborative effort of the people in this work package succeeded to create several workflows that are also able to use intrinsic data analysis – data driven analysis – alone and in combination with prior knowledge to create disease networks and draw conclusions from the case study data. It is therefore not possible – or at least not recommended - anymore to create a strict separation between pathway analysis and network analysis (**D13.3**). This deliverable therefore shows the **current status of both network analysis including and using pathways** on the case study data. Table 1 summarizes the current status of the publications in preparation, the scripts and workflows for the respective data analysis (**AD50**), and the specific pathways involved. We are furthermore including the teaching and training events and available training materials (**AD51**).

Table 1. Overview of publications (or current status of preparation), workflows, and involved pathways of the WP13 case study data analysis. The links to the workflows explicitly cover AD50.

Title	Link to publication, preprint, or draft with estimated publication date	Link to involved workflows (GitHub/ WorkflowHub/ Zenodo) & data (FAIR data points or other repositories)	Link to involved pathways*
Comparative network analysis workflows for Huntington's disease data	Draft available. Estimated submission date: early 2023	In preparation	www.wikipathways.org/instance/WP3853 www.wikipathways.org/instance/WP4222
Overlap of vitamin A and vitamin D target genes with CAKUT-related processes	Link to publication: https://pubmed.ncbi.nlm.nih.gov/35528959/	https://www.doi.org/10.5281/zenodo.4501623	www.wikipathways.org/instance/WP5053 www.wikipathways.org/instance/WP4823 www.wikipathways.org/instance/WP5052 www.wikipathways.org/instance/WP4830
A formalization of one of the main claims of "Overlap of vitamin A and vitamin D target genes with CAKUT-related processes" by Ozisik et al. 2021.	Link to publication: https://content.iospress.com/journals/data-science/5/1	*Not Applicable*	*Not Applicable*
Integrative analysis of CAKUT multiomics data	Draft available. Estimated submission date: early 2022	https://workflowhub.eu/projects/40#workflows https://wp13.fdps.ejprd.semlab-leiden.nl/catalog/4cad6f79-a7e1-46ef-8706-37f942f4aaea	www.wikipathways.org/instance/WP4172
Working title: CAKUT pathways paper	In preparation, estimated submission date: end of 2023	*Not Applicable*	www.wikipathways.org/instance/WP5053 www.wikipathways.org/instance/WP4823 www.wikipathways.org/instance/WP5052 www.wikipathways.org/instance/WP4830
Multi-omics molecular signatures in Inclusion Body Myositis	Draft available. Estimated submission in March 2023	https://github.com/jdwijnbergen/IBM_ASI_workflow	*Not Applicable*

Title	Link to publication, preprint, or draft with estimated publication date	Link to involved workflows (GitHub/ WorkflowHub/ Zenodo) & data (FAIR data points or other repositories)	Link to involved pathways*
Transcriptomics and Genetic Variance Analysis in Human Inclusion Body Myositis, using Pathway and Network Analysis Approaches	Master thesis available on request	https://github.com/fehrhart/Master-Thesis-Inclusion-Body-Myositis	www.wikipathways.org/instance/WP5120
Working title: Transcriptomics based patient stratification of PSVD	In preparation, draft available here: PSVD draft paper		www.wikipathways.org/instance/WP5269
Working title: Integrated transcriptomics and metabolomics data analysis on PSVD	In preparation		www.wikipathways.org/instance/WP5269
orsum: a Python package for filtering and comparing enrichment analyses using a simple principle	https://pubmed.ncbi.nlm.nih.gov/35870894/	https://github.com/ozanozisik/orsum	*Not Applicable*
FAIR Data Point Populator paper	In preparation, submission planned for June 2023	https://github.com/jdwijnbergen/fdp-populator/tree/VP https://github.com/LUMC-BioSemantics/ejprd-wp13-metadata	*Not Applicable*
Exploring Pathway Interactions to Detect Molecular Mechanisms of Disease: 22q11.2 Deletion Syndrome.	https://doi.org/10.21203/rs.3.rs-2093258/v1	https://github.com/woosubs/PathwayInteraction	www.wikipathways.org/instance/WP4657

Title	Link to publication, preprint, or draft with estimated publication date	Link to involved workflows (GitHub/ WorkflowHub/ Zenodo) & data (FAIR data points or other repositories)	Link to involved pathways*
Converging pathways found in copy number variation syndromes with high schizophrenia risk	https://biorxiv.org/cgi/content/short/2022.02.07.479370v2	*Not Applicable*	www.wikipathways.org/instance/WP4905 www.wikipathways.org/instance/WP4906 www.wikipathways.org/instance/WP4657 www.wikipathways.org/instance/WP3998 www.wikipathways.org/instance/WP4932 www.wikipathways.org/instance/WP4940 www.wikipathways.org/instance/WP4942 www.wikipathways.org/instance/WP4950 www.wikipathways.org/instance/WP4949 www.wikipathways.org/instance/WP5221 www.wikipathways.org/instance/WP5222 www.wikipathways.org/instance/WP5223 www.wikipathways.org/instance/WP5224
EJP RD Ndex network repository	https://www.ndexbio.org/#/networkset/d4048ad7-1281-11ea-bb65-0ac135e8bacf	*Not Applicable*	All pathways from WikiPathways are regularly uploaded to Ndex: https://www.ndexbio.org/#/user/363f49e0-4cf0-11e9-9f06-0ac135e8bacf
SUsPECT: A pipeline for variant effect prediction based on custom long-read transcriptomes for improved clinical variant annotation	https://doi.org/10.1101/2022.10.23.513417	https://github.com/cmbi/SUsPECT	*Not Applicable*

Title	Link to publication, preprint, or draft with estimated publication date	Link to involved workflows (GitHub/ WorkflowHub/ Zenodo) & data (FAIR data points or other repositories)	Link to involved pathways*
A Systems Biology Workflow to Support the Diagnosis of Inherited Metabolic Disorders: a study on Pyrimidine and Urea Cycle disorders	Preprint: https://doi.org/10.1101/2022.01.31.21265847	https://github.com/BiGCAT-UM/IMD-PUPY	www.wikipathways.org/instance/WP4224 www.wikipathways.org/instance/WP4225 www.wikipathways.org/instance/WP4571 www.wikipathways.org/instance/WP4584 www.wikipathways.org/instance/WP4583

*Please note that the pathways listed here involve both the pathways created newly for this case study and the pathways that already existed in the WikiPathways database and are relevant to the case study.

3. The case studies

As part of our work in WP13 we recruited participants from projects funded previously by E-Rare Cofund or EJP RD to provide and collaborate on their (multi)omics data case studies. Within these case studies we developed and applied our data analysis methods from each participating WP13 partner. Our first use case was Huntington's disease which served as a pilot study to apply and test several methods from each group. Subsequently, we sent out a call to European Reference Networks (ERNs) for collaboration on their research questions and multiomics datasets, which would be the use cases. In the end of the process, three use cases were selected: Congenital Anomalies of the Kidney and the Urinary Tract (CAKUT) from ERKNet, sporadic Inclusion Body Myositis (sIBM) from EURO-NMD, and Idiopathic Non-Cirrhotic Intrahepatic Portal Hypertension (INCPH) from ERN RARE-LIVER. An overview of the diseases studied, omics data available, and approaches used can be found in the table 2 below. The different use cases are discussed in the subsequent sections.

Table 2. Case study description

CASE STUDY	OMICS DATA	METHODS AND RESULTS
Huntington's disease	mRNA	6 different network analysis methods** and a comparison method* for those
CAKUT – Congenital anomalies of the kidney and urinary tract	Proteomics, Peptidomics, miRNA	3 (4) different analysis methods**, modelling overlap with nutritional metabolites (Vitamin A/D#)
IBM – Inclusion body myositis	mRNA, WES	2 different approaches ongoing**, inclusion of genetic variant analysis and interpretation in the network
INCPH – Idiopathic non-cirrhotic portal vein hypertension	mRNA, metabolomics	2 approaches ongoing**, inclusion of toxicology

*Paper published https://pubmed.ncbi.nlm.nih.gov/35870894/

#Paper published https://pubmed.ncbi.nlm.nih.gov/35528959/

** Paper in preparation

3.1. Huntington's disease

[Involved partners: LUMC, INSERM-AMU, UM, RUMC, ACURARE]

This case study was, in contrast to the other studies, not selected from a previously EJP RD funded collaboration, but chosen at the very beginning of the project from a repository of publicly available rare disease datasets. This Huntington's disease dataset was the case study for the collaborative network analysis group to test our methods and our approach on a publicly available rare disease dataset, before applying it to a dataset from ERNs. For this analysis we used the publicly available dataset provided by Labadorf et al. under the accession number GSE64810[1]. We applied six different network-based approaches, namely WGCNA, wTO-CoDiNA,

[1] https://pubmed.ncbi.nlm.nih.gov/26636579/

PathVisio/Cytoscape/CyTargetLinker, EnrichNet, pathfindR, and MOGAMUN (**Figure 1** below). These methods include both, extrinsic (prior (pathway) knowledge driven) and intrinsic methods of data analysis.

We demonstrated that application of several network-based analysis methods can help with rare disease research by increasing the soundness of common results unveiled by different methods and by providing new valuable insights discovered by the different methods. Our results provide new perspectives and disease understanding that can serve as hypotheses for future wet or dry lab experiments. The manuscript is in preparation. Furthermore, during the analysis process of this first case study a need to be able to compare the results of different extrinsic and intrinsic data analysis methods with each other was born. To address this, we developed a tool that can summarize different annotations with the use of ontologies. More details about the tool: orsum, Ozisik et al. 2022[2], see also chapter 4 "Additional tools, software and applications" below.

[2] https://pubmed.ncbi.nlm.nih.gov/35870894/

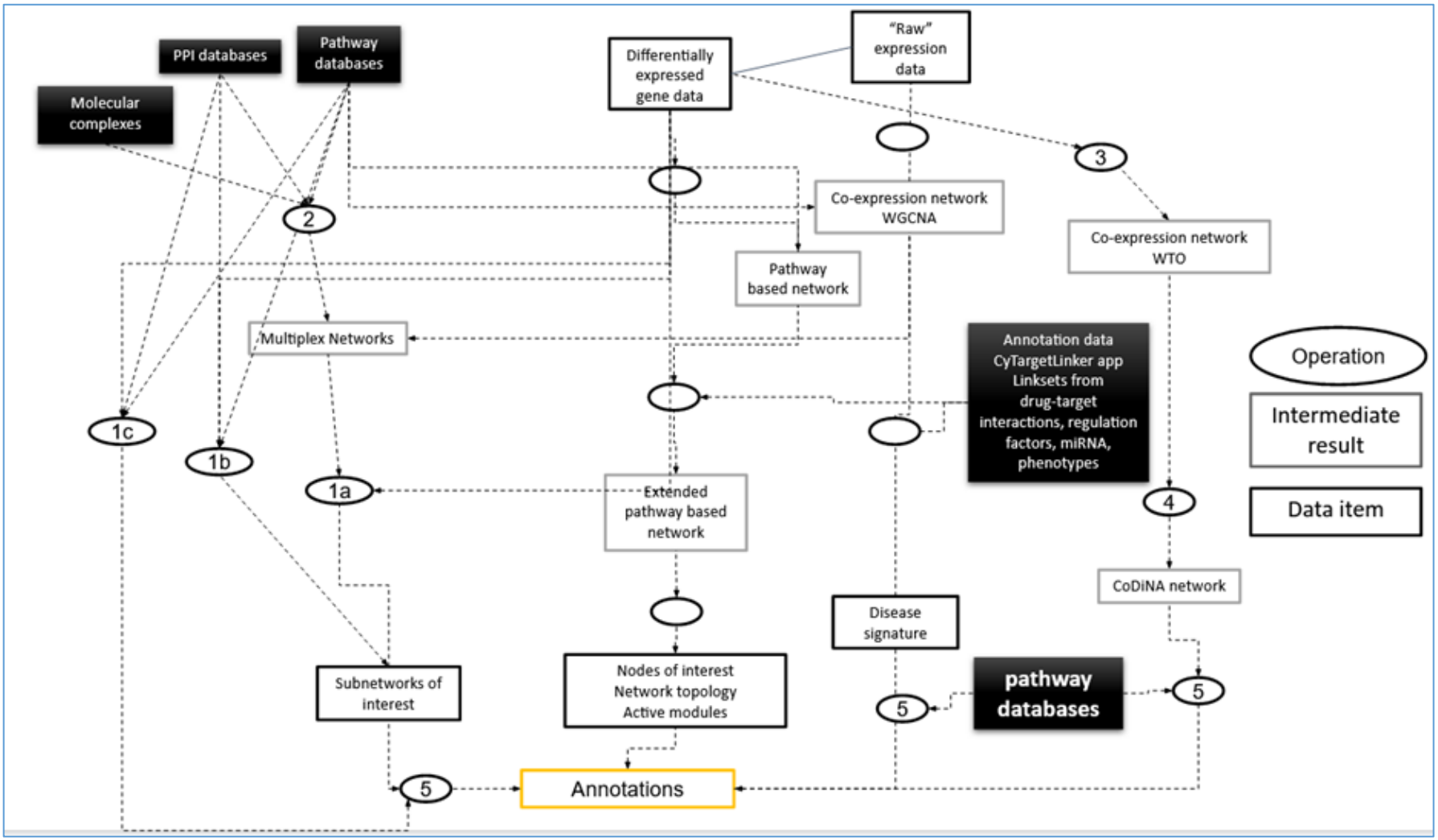

Figure 1. Workflow of data, analysis methods and extrinsic knowledge application to analyse the Huntington's disease transcriptomics dataset. **1a**: MOGAMUN, **1b**: PathfindR, **1c**: Enrichnet, **2**: MOGAMUN network construction, **3**: Co-expression networks construction, wTO, **4**: Differential co-expression network analysis, CoDiNA, **5**: pathway analysis/getting pathway annotations, 6: WCGNA co-expression network construction, 7: Network extension with CyTargetLinker, 8: Extracting information.

3.2. Congenital anomalies of the kidney and urinary tract

[Involved partners: RUMC, UM, INSERM-AMU, LUMC, ACURARE]

Congenital Anomalies of the Kidney and Urinary Tract (CAKUT) are a group of abnormalities affecting the kidneys and their outflow tracts. In the European Union, the overall prevalence of CAKUT (in live plus stillbirths) between 2013 and 2019 was approximately 35:10,000[3]. CAKUT presents high clinical variability in terms of observed anomaly and the severity. Approximately 40 different genes are known to be associated with monogenic causes of CAKUT in humans, but they explain only 5% to 20% of the cases. From the CAKUT dataset, being the first dataset that became available to WP13, we performed multiple studies as follows.

3.2.1. Integrative analysis of CAKUT multiomics data.

The CAKUT dataset consisted of three different datasets: miRNome, peptidome and proteome and was provided by a group from Toulouse University (part of ERKNet). These different omics datasets were integrated and analysed using three different methods that were developed within the EJP RD WP13 and these workflows are available at WorkflowHub database: mixOmics, momix, and pathway analysis. They allowed for the identification of affected molecular groups and networks in CAKUT patients. In addition, FAIRification of the datasets was performed and coupled to the EJP RD FAIR data point, to make the data available for future systematic omics integration and analyses. Despite the complementing features of the three different workflows, they all pointed towards an important role for collagen in CAKUT disease development. Furthermore, high overlap was seen with the PI3K-Akt signalling pathway on WikiPathways (WP4172), indicating both upregulation and downregulation of proteins involved in PI3K-Akt signalling (**Figure 2**). The study is in preparation.

[3] https://eu-rd-platform.jrc.ec.europa.eu/eurocat/eurocat-data/prevalence_en

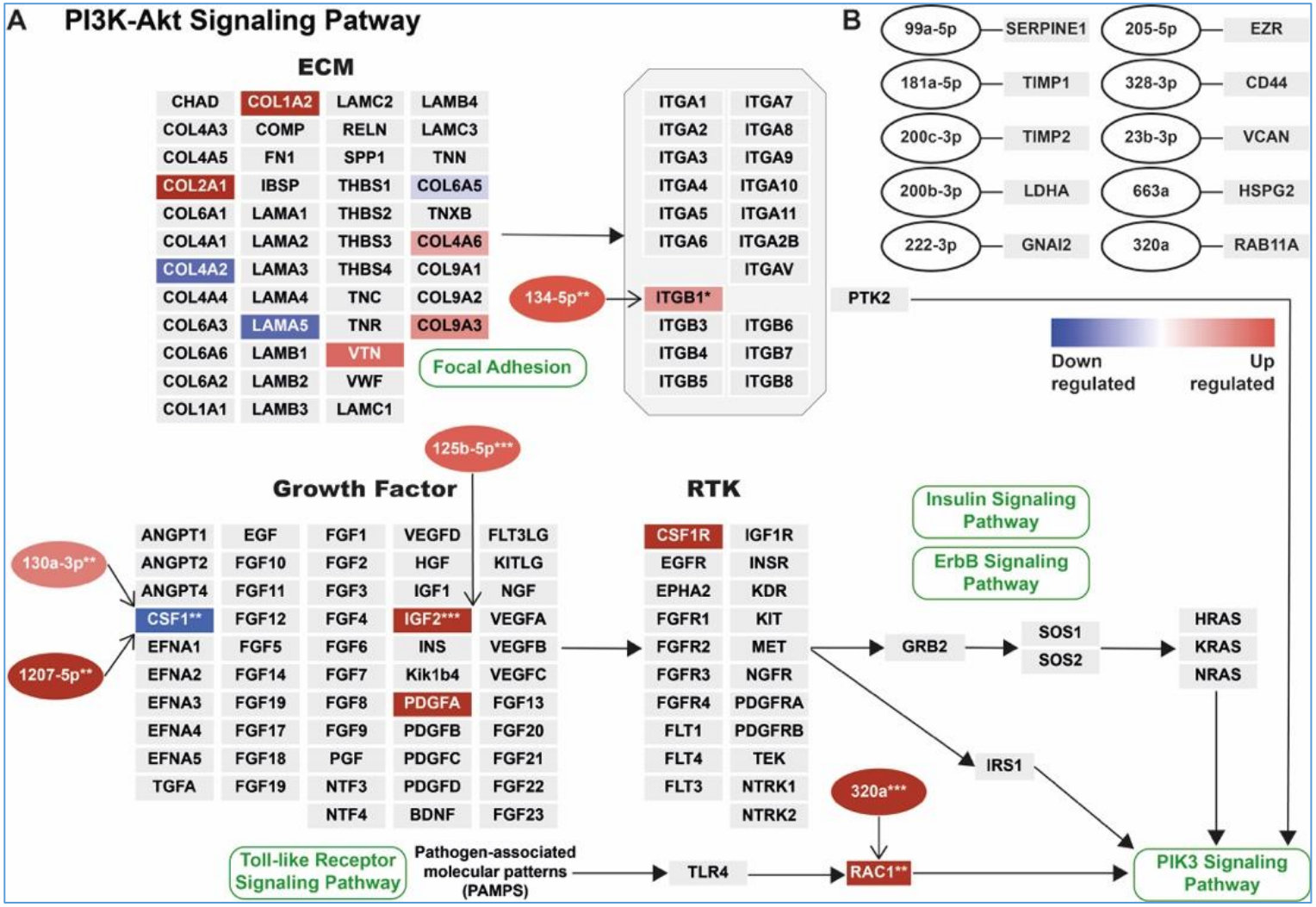

Figure 2. Affected part of the PI3K-Akt signalling pathway including results from miRNA and proteomics data

3.2.2. Molecular pathways of kidney development

There are currently about 40 genes known to cause CAKUT but the molecular pathways finally leading to CAKUT are not well understood. We recruited interested bio/medical researchers within ERKNet to participate in several pathway curation workshops (see also **D13.1**). The result of these workshops is a group of molecular pathways containing biomarkers of kidney development. This is the first extensive, literature annotated, human and machine-readable pathway on kidney development. There is a publication planned to be submitted in 2023.

3.2.3. Overlap of vitamin A and vitamin D target genes with CAKUT-related processes

The clinical variability and the complex aetiology of CAKUT cases suggest a multifactorial origin with complex interactions of both genetic and environmental factors contributing to the disease. Vitamins are among the environmental factors that can contribute to disease aetiology. There have been studies that present association of Vitamin A and Vitamin D with renal development, so we decided to investigate the overlap of these disease and vitamin related genes. This study is already reaching towards **D13.3**.

In this study we examined the overlap of vitamin A and vitamin D target genes with CAKUT-related gene sets (known disease genes, pathways, GO terms), and we observed significant overlap for vitamin A target genes. We obtained vitamin target genes list from The Comparative Toxicogenomics Database and the gene sets from WikiPathways, Reactome, GO databases and three publications. This study demonstrates a first hallmark to predict possible nutrition – disease interactions specific for patients with rare diseases[4] and invites for extension towards drugs and environmental toxicology. The main statement of this study was furthermore published in an experimental concept paper containing "nanopublications"[5].

3.2.4. CAKUT disease subclassification and patient stratification

There is a knowledge gap in the subclasses of CAKUT, and the complex nature of CAKUT requires a high-level molecular analysis in order to define disease subclasses. An approach with multi-omics data integration provides sufficient means for a comprehensive analysis of the disease. To observe disease subclasses and stratify the patients, we used proteomics, peptidomics, and transcriptomics data and utilized a network-based method, called Similarity network fusion[6], for multi-omics data integration. The main aim of the study is patient stratification and CAKUT disease subclassification to provide a better explanation of the complex disease mechanisms and aid for further possible treatment strategies. The publication is planned to be submitted in 2023.

[4] https://pubmed.ncbi.nlm.nih.gov/35528959/
[5] https://content.iospress.com/articles/data-science/ds210041
[6] https://www.nature.com/articles/nmeth.2810

3.3. Inclusion Body Myositis

[Involved partners: LUMC, UM, RUMC, EMBL-EBI]

Inclusion Body Myositis (IBM) is a rare, acquired muscle disease, occurring in roughly 24.8 to 45.6 individuals per 1 million people. The symptoms start with progressive asymmetric weakness which mainly affects the finger flexors and quadriceps muscles. This gradually leads to an impaired mobility for the patients. The exact mechanisms that lead to the IBM pathology are not known, however there are two main hypotheses. Firstly, the disease is primarily driven by autoimmunity and secondly, the disease is driven by muscle degeneration.

3.3.1. Multiomics molecular signatures in Inclusion Body Myositis

In order to investigate the mechanisms implicated in IBM, we performed a multi omics data integration study. Therefore, we created a large multi-omics disease network using interaction and target data from the STRING and miRTarBase databases. We combined this publicly available data together with the omics data from IBM and healthy patients. This included gene expression, microRNA expression, and variant burden data. We then applied an active subnetwork identification algorithm in order to find dysregulated subnetworks where the multiple omics are integrated. This is also reaching towards **D13.3 - subnetwork identification**. Our analysis revealed five interesting subnetworks that we hypothesize to be involved in IBM. Of interest is one particular subnetwork that implicates both autoimmunity and muscle degeneration. Our results will serve as hypotheses for future wet and dry lab experiments. The paper is planned to be submitted in March 2023.

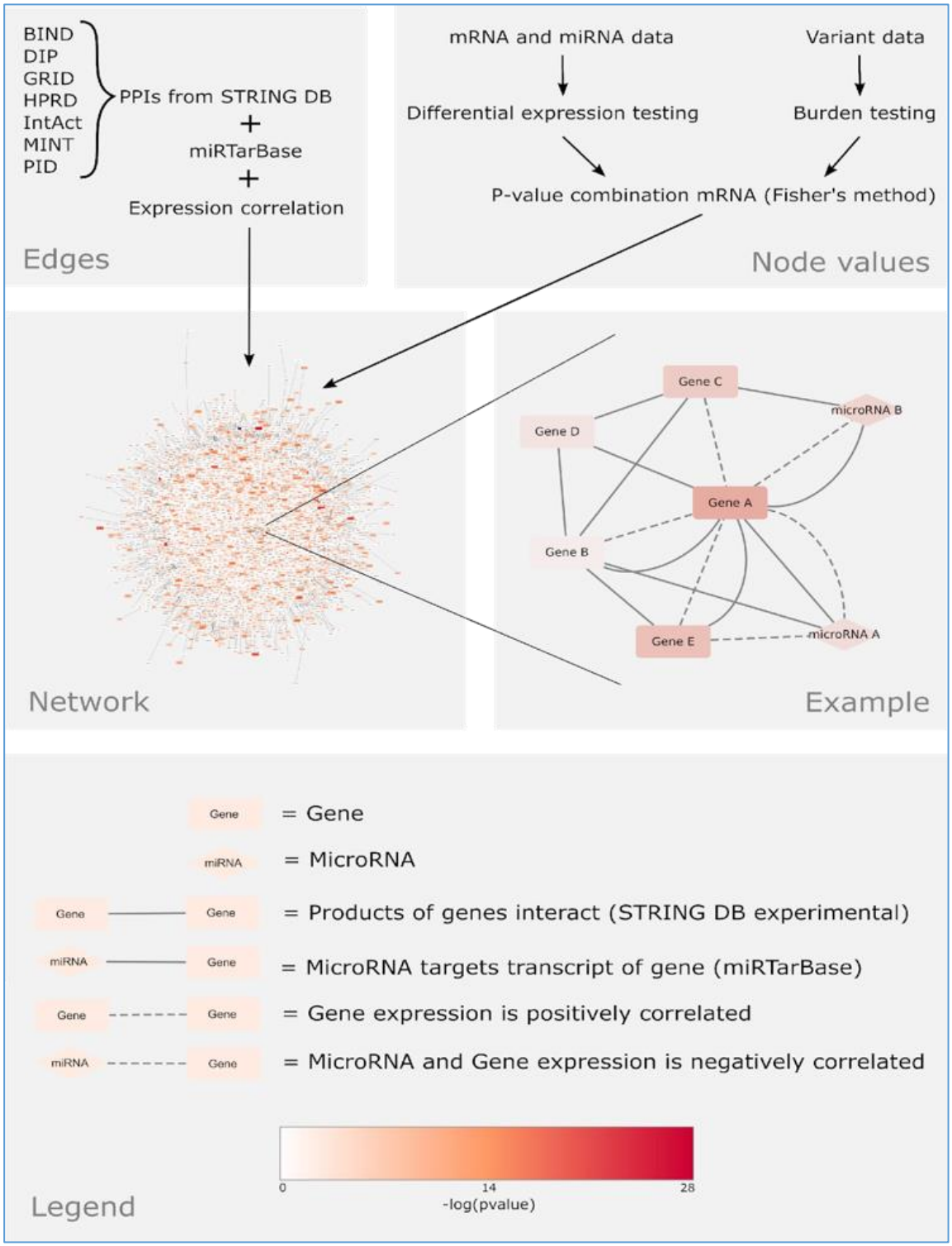

Figure 3. Multi-omics network construction. Data from publicly online databases (STRING and mirTarBase) were used as edges of this network to describe the interactions between gene products and genes with miRNAs. The values of the nodes were obtained from our experimental data. The differential expression values (p-values) were used for each node. In the case of genes where a variant burden p-value was available, we merged the two p-values using the Fisher's method

3.3.2. Transcriptomics and Genetic Variance Analysis in Human Inclusion Myositis, using Pathway and Network Analysis Approaches

[Master thesis by Theodoros Zarothiakis]

In this study, we explored several bioinformatics approaches, by applying a transcriptomics-based analysis workflow. Next, we applied network-based analyses, and integration with genetic, miRNA, and drug target information. In addition, we looked for co-expressed genes by applying weighted gene co-expression analysis. The scripts for the data analysis are available on GitHub[7]. Our results showed major disturbances in the immune response-related biological processes, while many pathways were altered. Candidate drugs and miRNAs that target genes have been identified as candidate treatments options and prognostic biomarkers. They were visualized as networks, combining information from several omics data layers. Here we presented a workflow that tried to elucidate the pathogenesis of this rare disease, to better understand the underlying biology, find potential new and personalized treatments, and provide a novel systems biology approach that could be implemented in the study of other rare diseases.

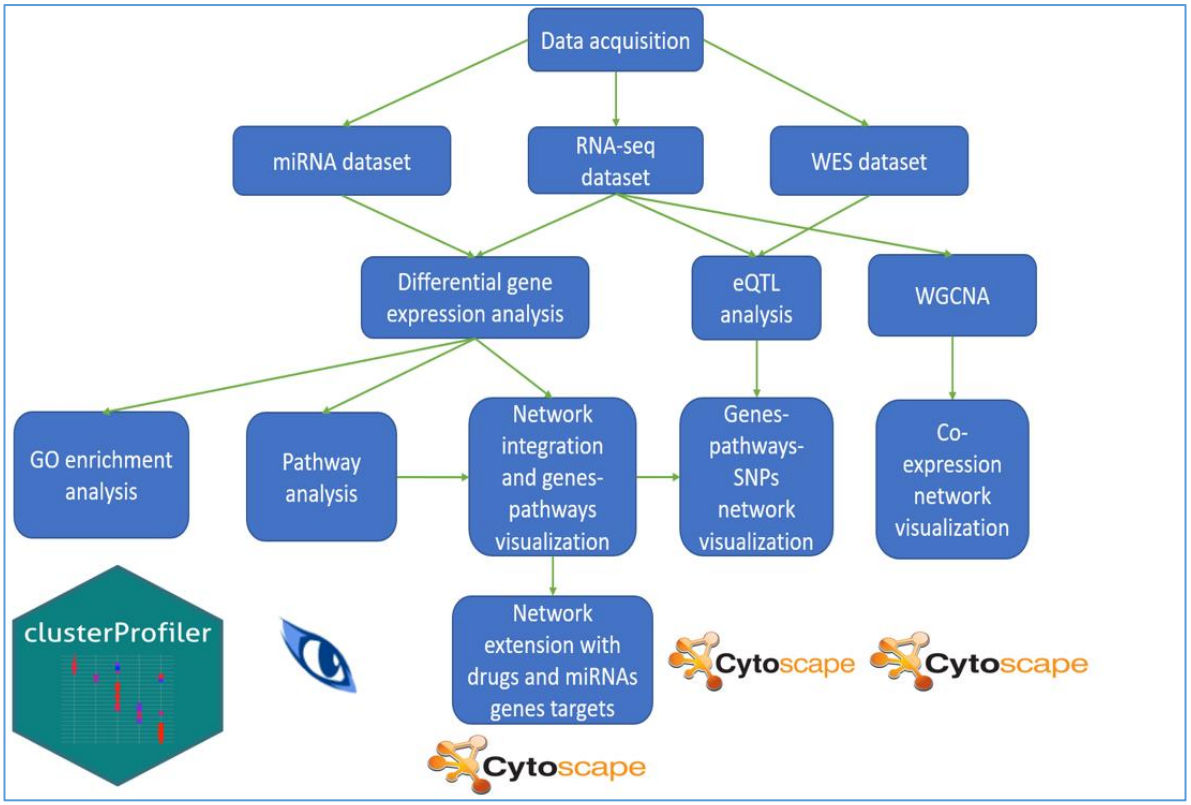

Figure 4. Analysis workflow. Step 1: Acquisition of the three datasets. Step 2: Differential gene expression analysis of the miRNA and mRNA datasets, WGCNA of the normalized counts. Step 3: GO enrichment and pathway overrepresentation analysis of the differentially expressed genes. Step 4: Importing the pathways in Cytoscape, as well as, annotating and merging them with pathway-gene information, for the network analysis. The WGCNA network visualization was done in parallel. Step 5: Extension of the pathway-genes network with drugs and miRNAs targets. Step 6: eQTL analysis and visualization of the identified SNPs, in the pathway-genes network from step 5.

[7] https://github.com/fehrhart/Master-Thesis-Inclusion-Body-Myositis

3.4. Porto-Sinusoidal Vascular Disease/ Idiopathic Non-Cirrhotic Portal vein Hypertension

[Involved partners: LUMC, UM, RUMC]

Porto-sinusoidal vascular disease (PSVD) is a complex rare disease affecting the liver and resulting in portal hypertension. The disease is poorly studied leading to challenges in accurate diagnosis, prognosis, and treatment options for these patients. The prevalence of the disease varies widely over the world due to the socioeconomic disparities. In France, the occurrence of PSVD is 4% among 3600 liver biopsies, while in Spain the incidence rate of PSVD with HIV is lower (0.5%). For this use case, we received microarray data, RNA-sequencing data, and metabolomics data with corresponding clinical information of these patients.

PSVD Hackathon: We have conducted a hackathon to discuss the data, analysis, preliminary results, and the pathway constructed by Friederike Ehrhart (UM) with the clinicians. The link to the presentation contents is linked here "PSVD hackathon presentation".

3.4.1. Transcriptomics based patient stratification of PSVD (working title)

Knowledge-based analysis: we created a pathway visualisation of potential genetic causes linked to PSVD (Pathway link) on WikiPathways.

Microarray data analysis: the transcriptomics data was curated, and the sample IDs were matched in collaboration with clinicians and data providers in Barcelona. Data pre-processing and differential gene expression (DEG) analysis has been performed, and the resulting gene ontology terms and pathways were enriched. A publication draft has been written to focus on the pathways and molecular processes affected in PSVD patients compared to healthy controls.
The clinical data is currently being curated to account for the missing data and mismatched samples in collaboration with clinicians/data providers (Genis and Juan Carlos Garcia) in Barcelona.

RNA sequencing data: the RNA sequence reads for the samples are currently being analysed to identify allele specific expression of a gene. This analysis would help provide us patient-specific regulatory networks. Thereby, allowing us to study the heterogeneity between the control and disease and also within the diseased group itself. It could also help us potentially identify SNPs specific to PVSD patients which could be used as biomarkers to predict the risk of developing PSVD.

3.4.2. Integrated transcriptomics and metabolomics data analysis on PSVD (working title)

This work is complemented by integration of the transcriptomics analysis with metabolomics data. Integration will be done on pathway level, since the samples are obtained from two groups of PSVD patients with little overlap. This omics integration and analysis approach will be done to determine whether there are metabolic changes that occur with PSVD progression, which could be used as prognostic biomarkers that can be measured from blood instead of requiring an invasive liver biopsy.

4. Additional tools, software and applications

As part of the work on the use cases, we created a series of reusable tools and workflows and application examples to other rare disorders as listed below:

4.1. orsum

[INSERM-AMU]

Enrichment analyses are widely applied to investigate lists of genes of interest. However, such analyses often result in long lists of annotation terms with high redundancy, making the interpretation and reporting difficult. Long annotation lists and redundancy also complicate the comparison of results obtained from different enrichment analyses.

During the study of Huntington's disease, while analysing the enrichment analysis results obtained by six different methods, these were the problems we encountered. As the available tools were not answering our needs, we developed orsum, a Python package for filtering and comparing enrichment analyses using a simple principle. This study is published[8].

4.2. FAIR Data Point Populator tool

[In collaboration with WP11 (LUMC)]

In order to facilitate making data FAIR, we developed a tool that can allow users with less technical expertise to add entries of their own datasets in a FAIR data point (FDP). The tool receives metadata from an Excel template that is shared with the user. The user needs to fill in the appropriate fields provided in the excel sheet (e.g., title, publisher, language) and share the file with the admin user who, after checking the excel sheet for inconsistencies, uploads the file to the GitHub repository. The tool then converts this file into a machine-readable format (RDF) and connects to the API of an FDP in order to push the metadata to this FDP. The metadata is then publicly available in the same machine-readable format and indexed by the FDP Index. We have successfully applied this method on the metadata of the IBM and CAKUT data. There is no draft available yet but the tool, Excel template, and WP13 metadata are available online. The paper is planned to be submitted to SWAT4HCLS as poster and full paper in June 2023.

4.3. Molecular pathway analysis for rare copy number variation syndromes

[UM]

4.3.1. Exploring Pathway Interactions to Detect Molecular Mechanisms of Disease: 22q11.2 Deletion Syndrome

In this study, an analysis method was developed that finds molecular paths between causative genes for a rare disease and their measurable outcome of differentially expressed genes. We used a transcriptomics dataset from peripheral blood of 22q11.2 deletion syndrome patients and healthy controls to investigate possible paths – also across pathways. These were identified using the WikiPathways database and the

[8] https://pubmed.ncbi.nlm.nih.gov/35870894/

22q11.2 copy number variation syndrome pathway[9]. The outcome suggests an involvement of natural killer cells in the development of psychiatric symptoms in 22q11.2 deletion syndrome patients.

4.3.2. Converging pathways found in copy number variation syndromes (CNVs) with high schizophrenia risk

Copy number variation syndromes (CNVs) are a special case of rare genetic disorders in which there is not one gene the cause of disease but a larger or smaller group of genes which are located on the same locus. The symptoms are usually highly variable, but a group of apparently unconnected CNVs show a high risk of developing schizophrenia. In this study, we created molecular pathways for these specific high-risk CNVs, allowing to investigate the causes and effects on the patient's phenotype on the molecular basis. The most important findings were that all these high risk CNVs share a connection to the BDNF pathway, while not containing more known schizophrenia risk genes, as expected by random chance[10].

4.4. NDEx network repository

[UM]

Already during the first workshop and hackathon "Molecular pathways for rare disease (FAIR) data analysis", in Maastricht 26-29/11/2019, a user account for EJP RD was created on the NDE network database[11]. The first large CAKUT related networks were uploaded there. In parallel, the WikiPathways account regularly publishes all pathways, including the rare disease pathways, as networks on its own NDEx account. The technical developments on WikiPathways itself make this specific task obsolete for the individual pathways since all pathways from WikiPathways are now on NDEx[12].

4.5. Improving variant analysis options for rare disease

[EBI-EMBL, RUMC]

The reference human gene annotation produced by Ensembl/GENCODE and RefSeq contains high-quality, well-supported transcripts but does not include extensive tissue-specific transcript data. Long read transcriptomic sequencing is now making tissue-specific transcript sets more accessible. We have created a NextFlow pipeline [https://github.com/cmbi/SUsPECT] built around Ensembl VEP to facilitate more detailed variant analysis with respect to disease-relevant transcript sets. It takes custom gene annotation in the commonly used GTF format and variants in VCF format as input. It predicts molecular consequence of the variants on the transcript and calculates PolyPhen-2 scores for all missense changes. This provides detailed tissue-specific annotation to enable more specific variant prioritisation. See SUsPECT publication[13] for more details.
See also Deliverable **D11.18** for a summary of new functionality added to Ensembl VEP for rare disease analysis.

[9] Preprint: https://www.researchsquare.com/article/rs-2093258/v1
[10] Preprint: https://biorxiv.org/cgi/content/short/2022.02.07.479370v2
[11] https://www.ndexbio.org/#/networkset/d4048ad7-1281-11ea-bb65-0ac135e8bacf
[12] https://www.ndexbio.org/#/user/363f49e0-4cf0-11e9-9f06-0ac135e8bacf
[13] https://www.biorxiv.org/content/10.1101/2022.10.23.513417v1

4.6. A Systems Biology Workflow to Support the Diagnosis of Inherited Metabolic Disorders: a study on Pyrimidine and Urea Cycle disorders

[UM]

A workflow was developed to visualize clinical data on metabolic pathway models for Pyrimidine and Urea Cycle disorders, which can be difficult to diagnose with the current methodology. The visualizations were used by trained clinical geneticist to derive a diagnosis, resulting in nine correct diagnoses out of 16 patient samples; an additional four visualizations provided a future direction for additional tests. Several issues were discovered when integrating data from different resources, which could be alleviated in the future by adding persistent identifiers to (clinical) biomarker data, allowing automated data downloads from relevant databases, and creating computer-readable pathway models from pathway figures. The presented workflow is adaptable to analyse different types of Inherited Metabolic Disorders, difficult patient cases, and functional assays in the future, which opens up the possibility for usage in the diagnostic workflow.

6. Appendix

6.1. AD 50: Case study specific (proof-of-concept) and generic multi-omics analysis pipelines as part of subtask 13.1.9 and in alignment with the deliverable D11.19 and subtask 11.4.2

This information is mostly covered in **Table 1** – the links to actual scripts and workflows. As mentioned in **D11.19** chapter **3.1.2** the EJP RD VRE analysis platform can run workflows directly from WorkflowHub [https://workflowhub.eu]. Currently, there are three complete workflows from the CAKUT data analysis study available [https://workflowhub.eu/projects/40#workflows], but we intend to have all scripts for all case studies publicly available as Common Workflow Language (CWL) or snakemake workflows, including required docker or singularity file, by the end of the project (see also **D11.19**, chapter **3.2**).

6.2. AD 51: Report for training purposes for Pillar 3, based on the workshops, analyses and VP deployment from pilot case studies

The common efforts on multi omics data analysis within WP13 required from the first year on regular teaching and training efforts within the WP but also for external participants. Table 3 summarises the officially given workshops, hackathons and other training activities.

Additionally, it should also be mentioned that teaching on rare diseases and the analysis methods is part of several WP13 members occupation as academic teachers. As stated already in **D13.1** report, training on creation and curation of molecular pathways for rare diseases contributed to the high number of now available rare disease pathways and about 30 students profited directly from this training during their academic education.

Table 3. Overview of workshops/hackathons other training activities

Title	When	Where
Workshop and Hackathon: Molecular pathways for rare disease (FAIR) data analysis [link]	26-29/11/2019	Maastricht, NL
Huntington's disease hackathon	12/01/2021 & 29/01/2021	EJP RD Microsoft Teams
Inclusion body myositis FAIR hackathon	17/12/2020	EJP RD Microsoft Teams
Inclusion body myositis case study - Pathway curation workshop	15/03/2022	EJP RD Microsoft Teams
Congenital Anomalies of the Kidney and the Urinary Tract pathway curation workshops	04/02/2021 & 25/08/2021	EJP RD Microsoft Teams

Title	When	Where
Porto-Sinusoidal Vascular Disease hackathon	12/04/2022	EJP RD Microsoft Teams
Bring Your Own Omics Data workshop [link - EJP RD consortium restricted]	17-18/11/2022	Nijmegen, NL
WF Variants hackathon (IBM use case)	01/02/2022	EJP RD Microsoft Teams
Omics data analysis & FAIR data stewardship – courses in collaboration with Helis Academy Interreg project [link]	2019	Several courses in Leiden, Eindhoven and Maastricht
Workflow creation hackathon	Feb 2023	TBD

Table 4. Training materials available online

Title	Link to material
Helis Academy – Omics data analysis & FAIR data stewardship	Link to material collection
ELIXIR TeSS	Courses: • WikiPathways • Identifier Mapping Service • network biology workflow • SPARQLing course
WikiPathways Academy	WikiPathways academy