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 13.3

Completed network analysis workflow (active node detection, lifestyle factor network evaluation and extended network analysis for drugs and toxic compounds)

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

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

Authors of the deliverable report: Friederike Ehrhart, Cenna Doornbos, Ozan Ozisik, Daphne Wijnbergen, Aishwarya Iyer, Sarah Hunt, Anaïs Baudot, Peter 't Hoen

Due date of deliverable: month 54

Dissemination level: public

1. Summary

In this deliverable we complete the network analysis workflows, including active node detection, lifestyle factor network evaluation and extended network analysis for drugs and toxic compounds. As indicated there, a large part of the network analysis was already done by the **previous deliverable 13.2** so we are focussing this report on the updates and developments since Dec 2022.
In short, we have continued the work on the four case studies developing (multi) omics analysis workflows. Some of the studies that were in preparation by Dec 22 are now submitted, others are closer to finalising. We have developed three different workflows to investigate rare disease – chemical compounds interaction networks and will continue working on demonstrating use cases in the following year. Additional works in this period include the FAIR data cube and the FAIR Data Point Populator, and the ELIXIR Rare Disease Systems Biology Service Bundle.

2. General introduction

This final deliverable for WP13 summarises the efforts of updating and extending the resources for systems biology research on rare diseases by providing example case studies with fully FAIR and open data and workflows. In this deliverable we show the progress of the last 6 months on the development of the case study analysis, rare disease networks including works on integration of chemical compounds. We have developed several approaches for hypothesis generation on the interaction of chemical compounds with rare disease – for drug repurposing but also for potential toxic interactions. Additional works in this period include tools and methods to improve and support FAIR data handling with e.g. the FAIR data point populator that allows to create a FAIR data point based on a simple Excel spreadsheet containing the relevant information. In collaboration with ELIXIR rare disease community the systems biology service bundle was created, a comprehensive collection including guides and tutorials on tools and methods for systems biology on rare diseases. In the final section we describe the "lessons learned" with the future needs we identified and works we will continue working on. In the following period until August 2024, we plan to have all the scientific papers and workflows submitted. Table 1 summarises the efforts of the past 6 months since the **previous deliverable 13.2**.

Table 1. Overview of new and updated publications (or their current status), workflows, and involved pathways additionally to and since Dec. 2022.

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*
Network-based approaches for the interpretation of transcriptomics data in rare diseases, an application to Huntington's disease	Draft available. Estimated submission date: July 2023	In preparation	www.wikipathways.org/instance/WP3853 www.wikipathways.org/instance/WP4222
Integrative analysis of CAKUT multiomics data	Submitted. Preprint available at bioRxiv.	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: 2024	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 September 2023	https://github.com/jdwijnbergen/IBM_ASI_workflow	Not Applicable
Machine learning in Huntington's disease: exploring the Enroll-HD dataset for prognosis & driving capability prediction	Accepted for publication, preprint: https://www.researchsquare.com/article/rs-2648484/v1	https://github.com/JasperO98/hdml/tree/main and an archived version at 10.5281/zenodo.7620222	
Working title: Transcriptomics based patient stratification of PSVD	In preparation, draft available here: PSVD draft paper		www.wikipathways.org/instance/WP5269

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*
Working title: Integrated transcriptomics and metabolomics data analysis on PSVD	In preparation, estimated submission in 2024.		www.wikipathways.org/instance/WP5269
FAIR Data Point Populator paper	In preparation, submission planned for November 2023	https://github.com/jdwijnbergen/fdp-populator/tree/VP https://github.com/LUMC-BioSemantics/ejprd-wp13-metadata	Not Applicable
ODAMNet: a Python package to identify molecular relationships between chemicals and rare diseases using overlap, active module and random walk approaches	In preparation, estimated submission in July 2023	https://github.com/MOohTus/ODAMNet	All rare disease pathways from WikiPathways

*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. Updates and new developments for the case studies

In this chapter we provide an update on the works on the 4 different case studies as described in the previous deliverable: 1. Huntington's disease, 2. Congenital Anomalies of the Kidney and the Urinary Tract (CAKUT) from ERKNet, 3. sporadic Inclusion Body Myositis (sIBM) from EURO-NMD, and 4. Idiopathic Non-Cirrhotic Intrahepatic Portal Hypertension (INCPH) from ERN RARE-LIVER. The overview of the diseases studied, omics data available, and approaches used can be found in the Table 2 below. The updates on the different case studies are presented 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 different multiomics analysis methods**
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 in preparation

3.1. Huntington's disease

[Involved partners: LUMC]

The Huntington's disease (HD) case study was chosen at the very beginning of this project. As a follow up to the network analysis work, we have extended our analytical work with three more developments that apply machine learning models to clinical data of Huntington's disease and the construction of knowledge graphs for a drug repurposing workflow and for investigating the role of iron in Huntington's disease.

1. Machine learning application on the Enroll-HD data.

The Enroll-HD is an integrated clinical research platform operating in the field of neurology that encompasses more than 20.000 participants in different clinical sites worldwide, including baseline and follow up visits for each patient, in both manifest and premanifest stages. Enroll-HD is collecting longitudinal observational data regarding the clinical aspects of HD, covering motor, cognitive, and behavioural domains as well as information regarding the medication and nutritional supplements that patients are receiving.

We have applied machine learning models to investigate whether the application of ML can lead to better predictions regarding disease progression and the quality of life of patients. We developed models to improve the prediction of the age at onset (AAO) and compared it to the well-established Langbehn formula. In addition, we used recurrent neural networks (RNNs) to demonstrate the utility of ML methods for complementing the clinician's experience-based decision making with a data-driven assessment on driving capability. Our approach can be of paramount importance for making predictions for assessments that clinicians have to perform that concern the quality of life of patients, such as the driving capability or the ability to go to work. The preprint is available here: https://www.researchsquare.com/article/rs-2648484/v1 and the paper has been accepted for publication by the Orphanet Journal for Rare Diseases.

2. Drug repurposing workflow for rare diseases (in collaboration with WP11)

We created an automated workflow which mines and integrates data from different online data sources into a knowledge graph and uses graph-based analysis methods together with machine learning to produce a ranked list of candidate compounds. As a case study we used Huntington's disease, but our method can be easily generalised to be applied to other rare diseases as well. The integration of diverse data sources into a knowledge graph allows us to capture and represent complex relationships between drugs and targets that might not be apparent through individual data sets alone, providing a comprehensive view of the underlying druggable rare disease biology. Through sophisticated graph algorithms, we can traverse the knowledge graph to identify patterns in the data and predict protentional new relationships between different entities, which supports hypothesis generation. Our automated workflow leverages the BioKnowledge reviewer library [https://doi.org/10.1093/database/baaa015]to extract and integrate data from online sources into a knowledge graph. Starting with a list of seed nodes, a Monarch network is created by including the first layer of neighbours and relations from Monarch for each seed node, along with their ortholog-phenotype nodes.

3. Knowlegde graph for investigating the iron in Huntington's Disease in collaboration with WP11

In this project, we present the application of a Structured Review (SR) to explore the relationship of iron with Huntington's Disease. A SR organizes and semantically represents the current knowledge around a research hypothesis in a structured manner, enabling semantic querying and data mining. Abnormal accumulation of iron in the brain has been associated with several other neurodegenerative diseases. Therefore, current therapies often include iron chelators to combat iron build up. Our SR is a knowledge graph that includes information surrounding the iron hypothesis in HD. We constructed a HD knowledge graph that integrates genes, anatomy, genotypes, variants, physiology and disorders as concepts and the relationships between these concepts. We apply basic machine learning link prediction to produce hypotheses by exploring the relationships between HD and iron.

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[1]. 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.

3.3. 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.

3.3.2. Meta-analysis of Inclusion Body Myositis transcriptomics datasets

As a follow-up to the Multiomics analysis in Inclusion Body Myositis, and as a use case for a FAIR based workflow, we performed a meta-analysis on Inclusion Body Myositis datasets. In five datasets that met the inclusion criteria, we performed a differential gene expression analysis. We then aggregated the resulting gene ranking lists using Robust Rank Aggregation (RRA). Finally, we enriched the results using Gene Ontology, protein-protein interactions, Metabolite data and tissue specific expression data. For future work on a FAIR based workflow, we are also identifying the metadata that is needed to perform this analysis in an automatic manner.

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

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. Pathway analysis using transcriptomics data to understand molecular processes affected in 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 original plan to stratify the patients using RNA sequencing data of PSVD patients had to be re-directed to another use case as the number of samples which had sequencing information for PSVD were only 8. Thereby, the power of the analyses would not be enough for the a.

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.

The intended integration of the transcriptomic and the metabolomics data for PSVD had to be redirected to another use case since the metabolomics data had 500 metabolites measured with less than 200 metabolites annotated with a unique persistent identifier. Also, given that the raw data was not available to re-annotate the metabolic data we could not go forward with workflow development.

4. Environment/drugs/AOP network integration

4.1. ODAMNet

Environmental factors are external conditions that can affect the health of living organisms. For a number of rare genetic diseases, an interplay between genetic and environmental factors is known or suspected. However, the studies are limited by the scarcity of patients and the difficulties in gathering reliable exposure information.
In order to aid in fostering research between environmental factors and rare diseases, we proposed ODAMNet, a Python package to investigate the possible relationships between chemicals, which are a subset of environmental factors, and rare diseases.
In ODAMNet, targets of the chemicals are retrieved from the Comparative Toxicogenomics Database (CTD) [Davis et al., 2023], and rare disease pathways are retrieved from WikiPathways [Martens et al., 2021]. We used three different and complementary bioinformatics approaches to integrate the data: overlap analysis, active module identification and random walk with restart. Of note, for the network-based approaches, i.e., active module identification and random walk with restart, the biological interaction networks are downloaded from NDEx [Pratt et al., 2015].
ODAMNet allows systematic analysis of chemical - rare disease relationships and generation of hypotheses for further investigation of effect mechanisms.

4.2. RareGenoScope

The RareGenoScope is an RShiny application that allows the creation of molecular networks for drug repurposing. The app was developed as bachelor thesis works by Mehaa Prabakhar (supervisor: Friederike Ehrhart) and identifies gene-disease associations from DisGeNET database, protein-protein interactions from STRING and drug-target information from DrugBank and creates a network from this information that identifies potential repurposable drugs (Figure 1). The network can also serve as a basis for further network extensions or other investigations. The code is available here [https://github.com/fehrhart/RareGenoScope-Application].

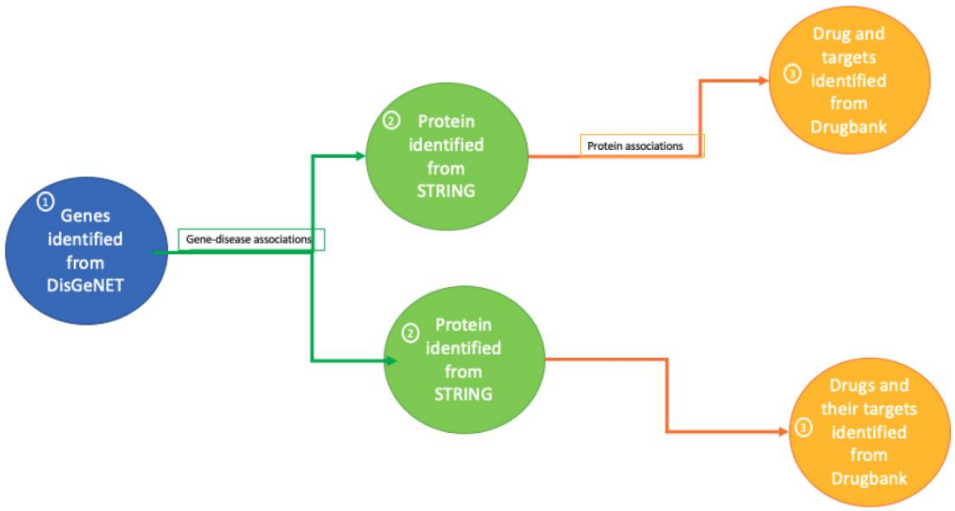

Figure 1: RareGenoScope application scheme. Starting from a disease it queries the associated genes from DisGeNET, adds protein-protein-interactions for those from STRING database and finally identifies potential drug targets among the proteins and adds the associated drugs from DrugBank.

4.3. Molecular AOPs – first results and outlook

Adverse outcome pathways (AOPs) describe how a molecular initiating event, such as a chemical compound binding to a receptor and triggering a signalling cascade, leads to adverse outcomes, such as liver failure, in patients (more information at https://aopwiki.org/). *Molecular* AOPs (https://www.wikipathways.org/communities/aop.html) capture the molecular parts of this pathway and are due to their universal annotations interoperable with any other type of pathways – including rare disease pathways. Currently, there are more rare disease pathways than molecular AOPs available but there is a group of liver toxicity AOPs we could start working with. In a first brainstorming and hacking session we tried several possibilities how to approach investigating the overlap between AOPs and rare disease pathways on liver related diseases (PSVD), respectively, liver toxicity AOPs. The first results indicated that there are genes/proteins that overlap in some of the pathways, but we need a clear filter and hypothesis to clear out general hub genes like MAPK1 or MAPK3. Starting in September 23 a student of Maastricht's Master biomedical sciences program will do his master thesis internship on overlapping network models for molecular AOPs and rare diseases.

5. 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:

5.1. FAIR Data Cube

[In collaboration with WP11, involved partners: (RUMC)]

The FAIR Data Cube (FDCube) is developed together with the Dutch X-omics project in order to provide a data infrastructure for federated data analysis (Figure 2). It is a

tool that brings algorithms to the data to facilitate data analysis. The data can be analysed within one or acros multiple FAIR data points. The metadata from the FAIR data points is used to tailor the algorithms for analysis and acquire the data access point, data access for federated analysis is managed by Vantage6 while the data itselfs remains within its protected environment. Furthermore, he FDCube adopts the FAIR principles and provides guidelines and examples for ISA-Tab-based metadata formats to create FAIR data, promoting data reusability.

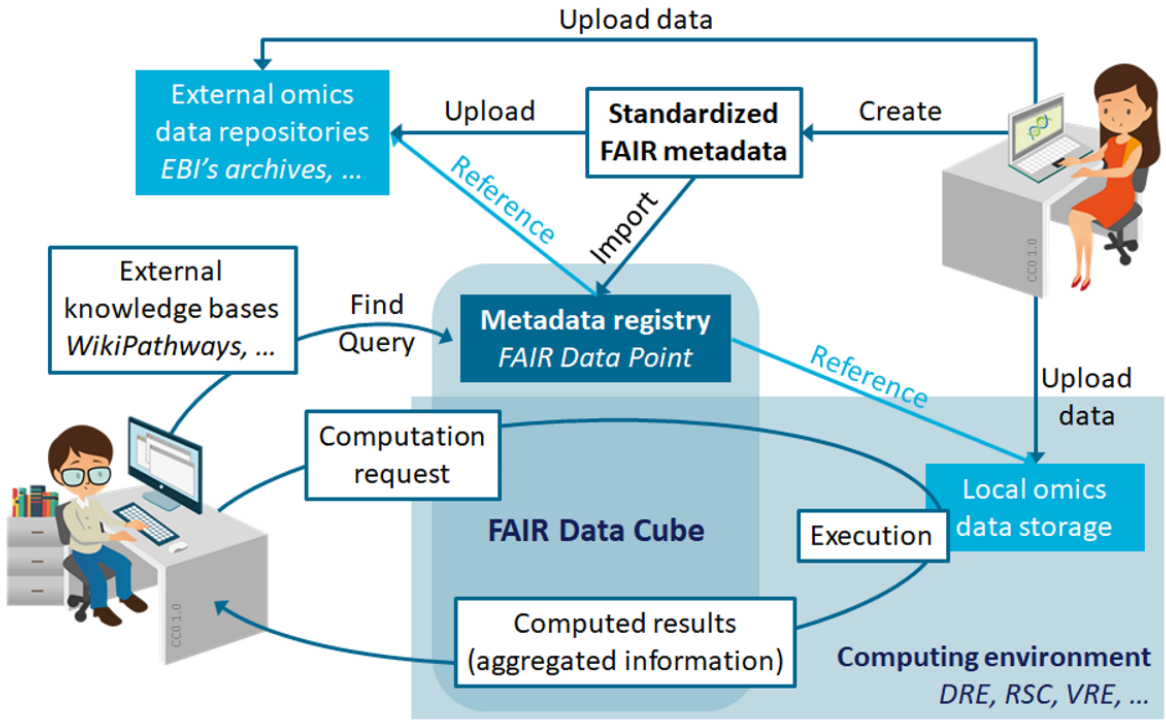

Figure 2: FAIR Data Cube

In order to facilitate the integration of Vantage6 into the FDCube, we organized a hackaton together with the Vantage6 development team, EJPRD WP11, and EJPRD WP13 to promote the analysis of federated analysis for the FDCube, but also other federated analysis tools that are generated within EJPRD. Furthermore, a workshop was organized in order to aid FDCube users through multiple use cases and tutorials. The preprint is available at https://doi.org/10.1101/2023.04.23.23289000.

5.2. FAIR Data Point Populator

The FAIR Data Populator is a tool to enables people to make their data FAIR without being an expert in FAIRification. It consists of an Excel template that helps users through documentation, tooltips, validation, and collaborative features. The FAIR Data Point Populator then transforms this data into an RDF format, and publishes this as an entry on a FAIR Data Point.

In addition to the FAIR Data Point schema, the FAIR Data Point Populator has also been updated to work with the EJP-RD metadata schema, to allow resources such as biobanks, patient registries and guidelines to be FAIRified using the FAIR Data Point Populator.

5.3. CWL tutorial and workshop

To aid people in the FAIRification of their workflows, we created a tutorial + workshop for Common Workflow Language (CWL). CWL is a language that defines an analytical workflow based on its input and output data. It allows for analytical workflows to be executed exactly as designed.

The tutorial and workshop consist of a presentation and a number of hands-on examples. The basics of CWL are explained through the presentation, and participants are able to apply this in practice. The materials are available at [https://github.com/jdwijnbergen/CWL_workshop].

5.4. ELIXIR Rare Disease Systems Biology Service Bundle

The European bioinformatics initiative ELIXIR [https://elixir-europe.org/] promotes the development of service bundles which contain information and links towards methods, tools and workflows to solve certain computational problems. The rare diseases community supported an implementation study to develop a systems biology focused service bundle. The systems biology service bundle contains tools, tutorials and information on the following case studies which are closely linked to the WP13 activities in EJP RD:

- Rare disease pathways – identifying and integrating knowledge on rare disease molecular pathways
- Find and map information about a specific variant from (different) resources
- Creation and analysis of rare disease networks
- Analysis of (multi) omics data
- Integration of toxicology, lifestyle and nutrition data
- Import and Export from a FAIR data point

The deliverable draft is available here [link to draft] and will be published at the website of the rare disease service bundles https://elixir-europe.org/services/service-bundles/rare-diseases when reviewed.

6. WP13 tool inclusion to the VP

The WP13 uses and has developed a number of tools/softwares that are generally useful for systems biology or specifically useful for working with rare disease data. In collaboration with WP11 the following tools/softwares and their documentation and tutorials were added to the VP (Table X).

Table 3: Tools included to VP (Level 1)

Tool name	Website	bio.tools	Tutorial/ documentation	Tagged with "rare disease" on bio.tools
PathVisio	https://pathvisio.org/	https://bio.tools/pathvisio	https://pathvisio.org/tutorials	YES
BridgeDb	https://www.bridgedb.org/	https://bio.tools/bridgedb	https://www.bridgedb.org/pages/docs.html	YES
BridgeDbAPI	https://www.bridgedb.org/pages/webservice.html	https://bio.tools/bridgedb_api	https://www.bridgedb.org/pages/tutorials_workflows.html	YES
BridgeDbR	http://bioconductor.org/packages/release/bioc/html/BridgeDbR.html	https://bio.tools/bridgedbr	http://bioconductor.org/packages/release/bioc/manuals/BridgeDbR/man/BridgeDbR.pdf	YES
orsum	https://anaconda.org/bioconda/orsum	https://bio.tools/orsum	https://github.com/ozanozisik/orsum/	NO
ODAMnet	https://pypi.org/project/ODAMNet/	https://bio.tools/odamnet	https://odamnet.readthedocs.io/en/latest/	YES
VEP	http://www.ensembl.org/info/docs/tools/vep/index.html	https://bio.tools/vep	https://www.youtube.com/watch?v=rSIG_OVzyLU	NO
Cytargetlinker	https://cytargetlinker.github.io/	https://bio.tools/CyTargetLinker	https://cytargetlinker.github.io/pages/tutorials	NO
Linkset creator	https://github.com/CyTargetLinker/linksetCreator		https://github.com/CyTargetLinker/linksetCreator	NO
MOGAMUN	https://www.bioconductor.org/packages/release/bioc/html/MOGAMUN.html	https://bio.tools/mogamun	https://www.bioconductor.org/packages/release/bioc/vignettes/MOGAMUN/inst/doc/MOGAMUN_Vignette.html	NO
momix	https://workflowhub.eu/workflows/126		https://workflowhub.eu/workflows/126	NO
mixomics	https://workflowhub.eu/workflows/330		https://workflowhub.eu/workflows/330	NO
WGCNA	https://workflowhub.eu/workflows/331		https://workflowhub.eu/workflows/331	NO
Fair Data Cube	https://github.com/Xomics/FAIRDataCube		https://github.com/Xomics/FAIRDataCube/wiki	NO