www.emif.eu

European Medical Information Framework

Grant Agreement nº115372

D13.4 Data analysis tools for vertical projects v2

WP13 – Analysis, processing & visualisation methods and tools

V2.0
[Final]

Lead beneficiary: EMBL
Date: 26/01/2016
Nature: P
Dissemination level: PU

TABLE OF CONTENTS

DOCUMENT INFORMATION

Grant Agreement Number	115372	**Acronym**	EMIF
Full title	European Medical Information Framework		
Project URL	http://www.emif.eu		
IMI Project officer	Ann Martin (Ann.Martin@imi.europa.eu)		

Deliverable	**Number**	13.4	**Title**	Data analysis tools for vertical projects v2
Work package	**Number**	13	**Title**	Analysis, processing & visualization methods and tools

Delivery date	**Contractual**	Month	**Actual**	26/01/2015
Status	Current version / V2.0		Draft ☐ Final ☑	
Nature	Report ☑ Prototype ☐ Other ☐			
Dissemination Level	Public ☑ Restricted ☐ Confidential ☐			

Authors (Partner)	Natalja Kurbatova (EMBL-EBI)			
Responsible Author	Natalja Kurbatova		**Email**	natalja@ebi.ac.uk
	Partner	EMBL	**Phone**	+44 (0) 1223 492 597

DOCUMENT HISTORY

NAME	DATE	VERSION	DESCRIPTION
Natalja Kurbatova	30/12/15	1.0	First draft
Alvis Brazma	08/01/16	1.1	Review and changes
Rudi Verbeeck	12/01/16	1.2	Review and changes
Natalja Kurbatova	18/01/16	1.3	Review and changes
Natalja Kurbatova	21/01/16	1.4	Review and changes
Natalja Kurbayova	26/01/16	2.0	Final version

DEFINITIONS

- **Analysis (Broad definition).** Any “analytical method” used to get insights from data, based on descriptive or predictive statistics, modelling, simulation, graphs and other visualisation methods.
- **Cluster**. A computer cluster consists of a set of loosely or tightly connected computers that work together so that, in many respects, they can be viewed as a single system.
- **Cloud computing.** Cloud computing is defined as a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications.
- **Interface (In computing).** It is a device or program enabling a user to communicate with a computer.
- **LC-MS.** Liquid chromatography-mass spectrometry is an analytical chemistry technique that combines the physical separation capabilities of liquid chromatography with the mass analysis capabilities of mass spectrometry.
- **NGS.** Next-generation sequencing, also known as high-throughput sequencing is the term used to described a number of different modern sequencing technologies, like Illumina sequencing, Roche 454 sequencing, Ion torrent sequencing, SOLiD sequencing.
- **Pipeline (In the scope of this document)**. Pipeline is a chain of data analysis components. Terms “pipeline” and “workflow” are interchangeable.
- **Platform (In the scope of this document).** The meaning of the term “platform” is very similar to the term “framework” – any base of technologies on which other technologies or processes are built. Platform in most of the cases has tools for developers and may provide computational power.
- **URL**. It stands for uniform resource locator is a reference to a resource that specifies the location of the resource on a computer network and a mechanism for retrieving it. A URL is a specific type of uniform resource identifier (URI).
- **VM (In computing).** In computing, VM stands for a virtual machine. Virtual machine is an emulation of a particular computer system. Virtual machines operate based on the computer architecture and functions of a real or hypothetical computer, and their implementations may involve specialized hardware, software, or a combination of both.
- **Workflow (In the scope of this document).** A series of computational steps usually programmed to run at once. Terms “pipeline” and “workflow” are interchangeable.

EXECUTIVE SUMMARY

The main driver of deliverables D13.3 and D13.4 is the needs of the EMIF verticals. During our face to face meetings and conference calls the EMIF AD vertical requested a multi-omics data sharing solution that allows metadata attachment to the data, a pipeline sharing solution and high performance computing. In EMIF vertical projects are using tranSMART for clinical data analysis, but the –omics data analysis problems are solved with the help of R or by using specific pipelines.

Taking into account all those considerations data analysis tools for vertical projects v2 includes tranSMART, R Cloud and a number of pipelines and data analysis tools adapted for cloud computing. The adaptation and optimization process depends on the individual pipeline, for example some of the bioinformatics pipelines requires docker image creation and/or adding of the parallel computing support for the efficient work on a cluster. In some special cases when user interface support is need, e.g for MZmine2 tool, the adaption means creation of the GUI forwarding mechanism.

Cloud computing provides users with a number of benefits: reduction of computational costs, universal access, up to date software, choice of applications, flexibility. However, there are a lot of different cloud platforms that vertical projects potentially can choose: Amazon Web Services, OpenStack, VMWare, Google Cloud, etc.

Our provided solution “Multi-omics Research Environment” is transferrable between different cloud platforms and has specialised components for clinical and –omics data analysis. In addition, “Multi-omics Research Environment” implies a flexible architecture that allows new tools and pipelines to be easily added upon request of the vertical projects. “Multi-omics Research Environment” is described in details in the deliverable D13.3. The tranSMART application is described in deliverable D13.2.

The following pipelines and tools are available in the “Multi-omics Research Environment” at the current stage (v2):

- tranSMART for clinical data analysis;
- R Cloud for R parallel computing and R specific analysis;
- iRAP pipeline adapted for the Docker cluster to analyse transcriptomics sequencing data;
- NGSeasy pipeline adapted for the Docker cluster to analyse genomics sequencing data;
- MZmine2 adapted for the Docker cluster to analyse proteomics and metabolomics LC-MS data;
- Sequence Imp pipeline adapted for the Docker cluster to analyse microRNA sequencing data.

The tranSMART instance, Docker cluster and R Cloud are connected and use a shared file system. The Docker cluster and R Cloud benefit from the scalability of cluster computing

usage – multiple VMs, job queues and task scheduler. New resources are added when needed.

Deliverable D13.4 is a continuation of WP13 previous deliverables D13.1, D13.2 and D13.3.

KEY WORDS (Wordle style)[1]

[1] http://www.wordle.net/

1. INTRODUCTION

EMIF work-package WP13 is about developing analysis, processing and visualization methods and tools, in particular to aid EMIF verticals.

We've developed "Multi-omics Research Environment" as the open-source project available on the github repository: https://github.com/olgamelnichuk/ansible-vcloud. An institute with the access to the cloud platform can take the whole project or its separate parts and create an instance of the "Multi-omics Research Environment". Since the cloud platform solution provides scalability feature the instance of the environment can be aimed either for a particular single cohort analysis or for the large collaborative vertical project depending on the availability of computational resources. One instance of "Multi-omics Research Environment" is hosted at the EMBL-EBI on Embassy Cloud platform and is accessible by EMIF AD vertical researchers. At the moment this instance is aimed for "1000 samples AD cohort" data analysis.

In order to meet the needs of the vertical projects we've provided the following tools and pipelines as a part of a "Multi-omics Research Environment":

- tranSMART for clinical data analysis;
- R Cloud for R parallel computing and R specific analysis;
- iRAP pipeline adapted for the Docker cluster to analyse transcriptomics sequencing data;
- NGSeasy pipeline adapted for the Docker cluster to analyse genomics sequencing data;
- MZmine2 adapted for the Docker cluster to analyse proteomics and metabolomics LC-MS data;
- Sequence Imp pipeline adapted for the Docker cluster to analyse microRNA sequencing data.

tranSMART, R Cloud and Docker cluster, where pipelines and tools for –omics data are running, are interconnected to provide solutions for the integrative data analysis (see Figure 1).

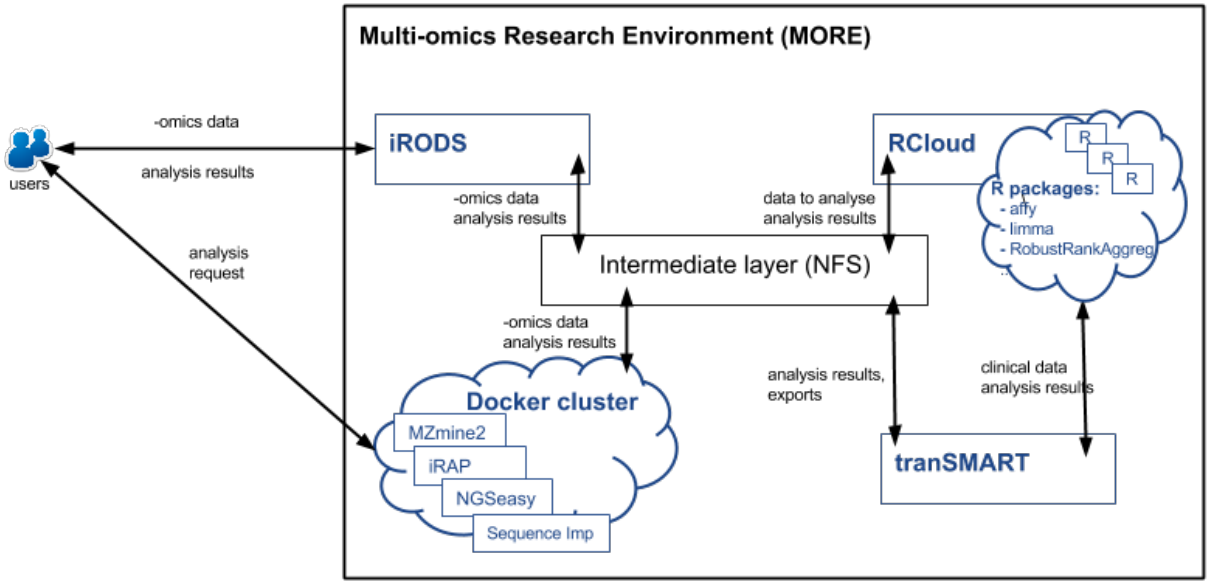

Figure 1 Components of the "Multi-omics Research Environment"

tranSMART is a knowledge management application that enables scientists to develop and refine research hypotheses by investigating correlations between genetic and phenotypic data. Within the "Multi-omics Research Environment" tranSMART is a core component for the clinical data analysis. A detailed description of tranSMART analysis options for the vertical projects is provided in deliverable D13.2.

R Cloud gives the opportunity to use any of the existing R packages for the data analysis. In addition, R Cloud supports parallel R computing. For example, a researcher from a vertical project can install and use R package "RobustRankAggreg" developed by University of Tartu (Kolde et. al), get the list of genes/transcripts and store them back into tranSMART with the scores calculated by the Robust Rank Aggregation algorithm.

Docker is an open-source project that automates the deployment of applications inside software containers, by providing ad additional layer of abstraction and automation of operating-system-level virtualization. From the end-user perspective Docker encapsulates software dependencies, for example those dependencies are especially complex in case of next generation sequencing pipelines, and provides ready to go solutions that can be used without time-consuming pipeline installation process. A Docker cluster is a cloud platform solution – a cluster of VMs under orchestration of OpenLava (LSF) with Docker installed on each VM. Docker cluster gives the possibility to run dockerized analysis pipelines in parallel.

Both R Cloud and tranSMART's analytical part are mostly supposed to be used for the down-stream analysis. Non-massive data like microarray results can also be analysed in the R Cloud; however, vast amounts of data from next generation sequencing at the pre-processing stage need a special approach and a lot of computational power. For such data we have provided a Docker cluster solution with a number of pipelines adapted for cloud computing.

At the moment we have pipelines to deal with genomics, transcriptomics, small RNAs sequencing, proteomics and metabolomics data.

All the mentioned components make up a very flexible system that can satisfy all possible needs of the vertical projects. The new dockerized pipelines can be added to the Docker cluster by vertical project request.

2. tranSMART

tranSMART is a knowledge management application that enables scientists to develop and refine research hypotheses by investigating correlations between genetic and phenotypic data.

Within the “Multi-omics Research Environment” tranSMART is a core component for the clinical data analysis. In addition, tranSMART instance is integrated with R Cloud that allows parallel R computation.

tranSMART analytical part includes the following analysis and visualisation types:

- Summary statistics;
- ANOVA and following box plot generation;
- Linear Regression analysis and following scatter plot generation;
- Logistic Regression;
- Correlation Analysis;
- Fisher Exact Test;
- Standard heatmaps;
- Survival Analysis;
- Principal Component Analysis;
- Hierarchical Clustering;
- K-Means Clustering;
- Marker Selection;
- Cross trial query: requires carful harmonization of the contributing datasets. See deliverables D11.3 and 14.5 for details on data harmonization in tranSMART. The main effort has been spent on loading cohort data for cross trial analysis, rather than expanding tranSMART functionality.

The more detailed description of tranSMART functionality can be found in deliverable D13.1. For the details and examples of tranSMART clinical data analysis possibilities see deliverable D13.2.

3. DOCKER CLUSTER PIPELINES AND TOOLS

Docker is an open-source project that automates the deployment of applications inside software containers, by providing an additional layer of abstraction and automation of operating-system-level virtualization. A Docker cluster is a cloud platform solution – a cluster of VMs under orchestration of OpenLava (LSF) with Docker installed on each VM (see Figure 2). More details about the Docker cluster can be found in deliverable D13.3.

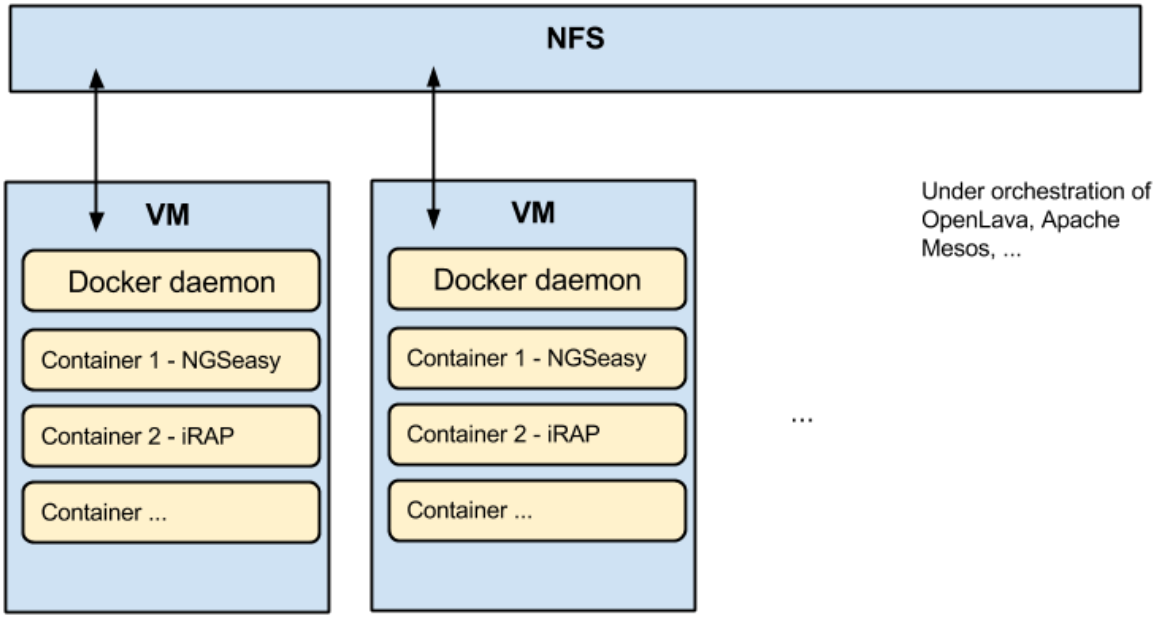

Figure 2 Docker cluster and analysis pipelines in the "Multi-omics Research Environment"

Vast amounts of data from next generation sequencing at the pre-processing stage need a special approach and a lot of computational power.

For such data we have adapted a number of pipelines for the usage of Docker cluster features:

- iRAP pipeline to analyse transcriptomics sequencing data;
- NGSeasy pipeline to analyse genomics sequencing data;
- MZmine2 to analyse proteomics and metabolomics LC-MS data;
- Sequence Imp pipeline to analyse microRNA sequencing data.

3.1. iRAP PIPELINE

iRAP is an integrated RNA-seq analysis pipeline that allows the user to select and apply their preferred combination of existing tools for mapping reads, quantifying expression and testing for differential expression. iRAP also includes multiple tools for gene set enrichment analysis and generates web browsable reports of the results obtained in the different stages of the

pipeline. Depending upon the application, iRAP can be used to quantify expression at the gene, exon or transcript level.

iRAP is implemented on our docker cluster in the “Multi-omics Research Environment”. It can be run in parallel using all benefits of a multi-CPUs cloud platform.

Availability – iRAP is available under General Public License 3 (GPLv3) and although it should be portable to any POSIX-compliant operating system, several third party programs only run on Linux. iRAP can be obtained from (https://github.com/nunofonseca/irap).

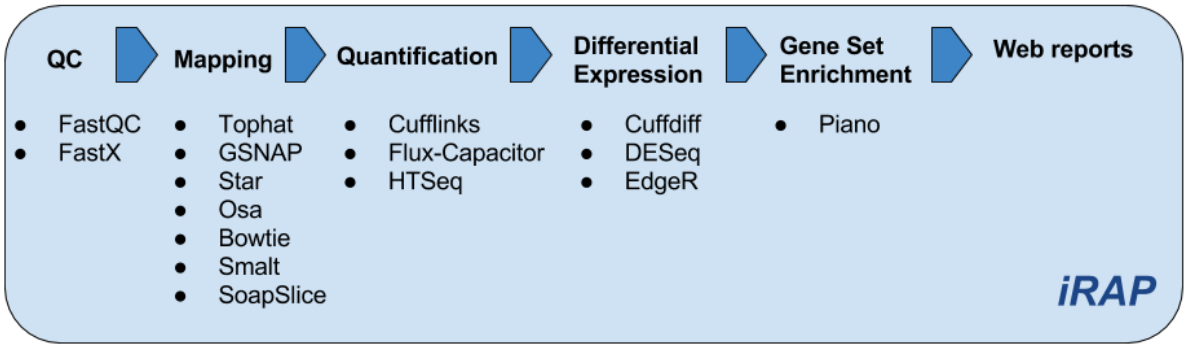

Figure 3 iRAP - an integrated RNA-seq analysis pipeline that allows a user to select a combination of preferred tools

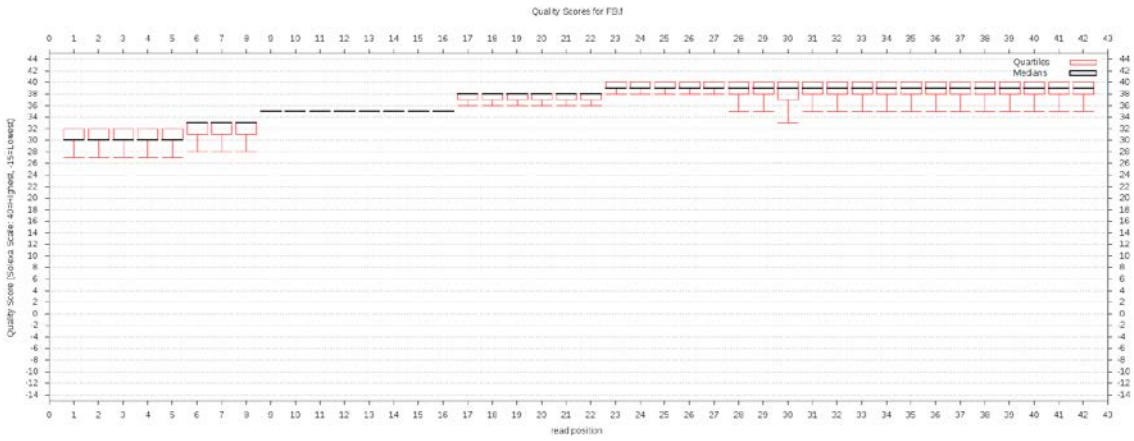

Figure 4 iRAP – example of a generated report “Quality scores”

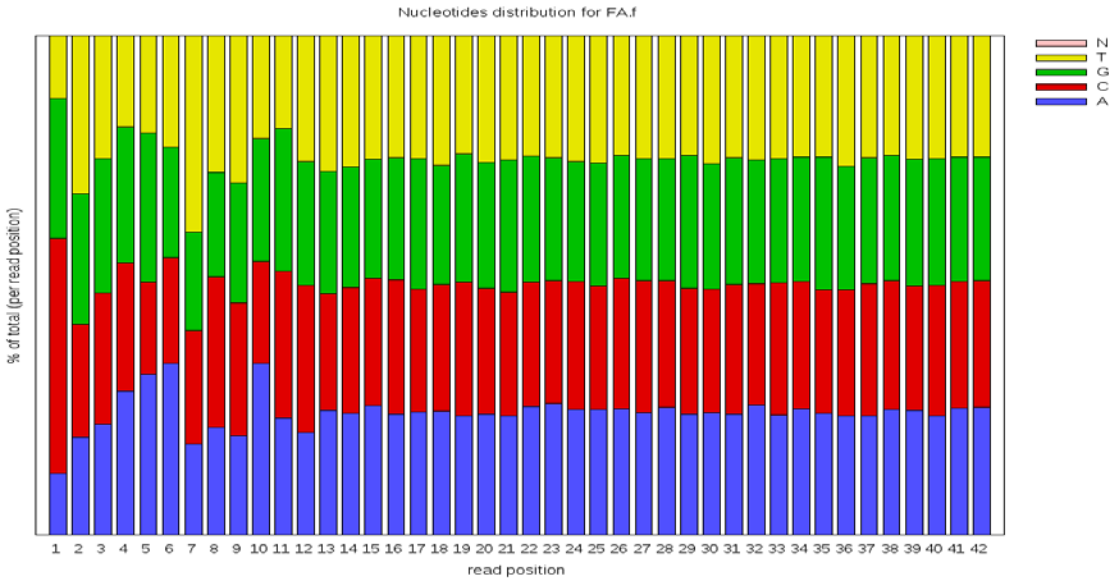

Figure 5 iRAP - example of a generated report "Nucleotides distribution"

3.2. NGSeasy PIPELINE

NGSeasy – dockerized next generation sequencing pipeline for the genomics data. Similar to the iRAP pipeline, NGSeasy consists of multiple sequential steps and the user has an opportunity to choose the preferred tool for each step. The basic pipeline contains all the tools needed for manipulation and quality control of raw fastq files (ILLUMINA focused), SAM/BAM manipulation, alignment, cleaning (based on GATK best practises) and first pass variant discovery. Separate docker containers are provided for in-depth variant annotation, structural variant calling, basic reporting and visualisations.

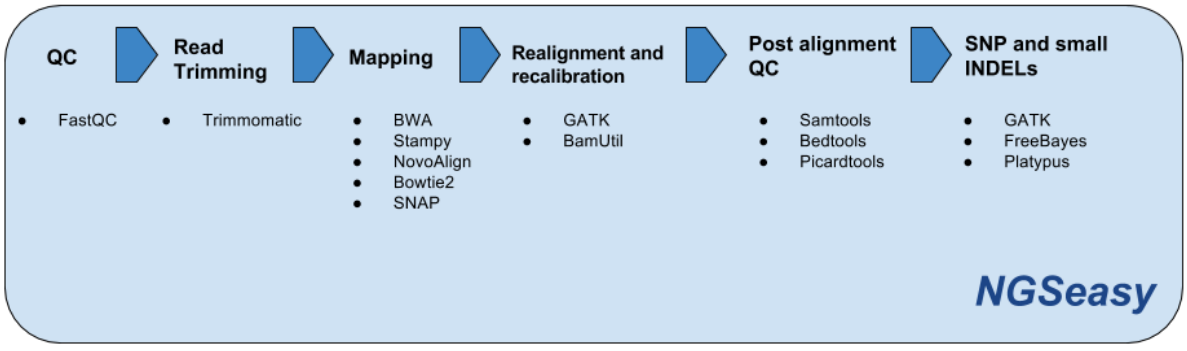

Figure 6 NGSeasy pipeline's components and tools

Availability – NGSeasy is available under General Public License 3 (GPLv3). However, a user has to purchase a GATK license if he/she belongs to a for-profit organisation and wants to use GATK tools. NGSeasy can be obtained from (https://github.com/KHP-Informatics/ngseasy).

The NGSeasy pipeline is implemented on our Docker cluster in the "Multi-omics Research Environment". It can be run in parallel mode using all benefits of a multi-CPUs (VMs) cloud platform.

3.3. MZmine2 TOOL

MZmine2 is open-source software for mass-spectrometry data processing, with the main focus on LC-MS data. MZmine2 can read and process both unit mass resolution and exact mass resolution data in the following formats: mzML, mzXML, mzData, NetCDF, Thermo RAW and Waters RAW. MZmine2 steps include: raw data filtering and smoothing, peak detection (mass detection followed by deconvolution), different peak list methods, statistical analysis (Cluster analysis, Heat Maps, PCA, etc.) and visualisation possibilities (Chromatogram plots, 2D and 3D view of datasets, Histograms, Scatter plots etc.). Some of the visualisation examples are shown on Figures 7,8.

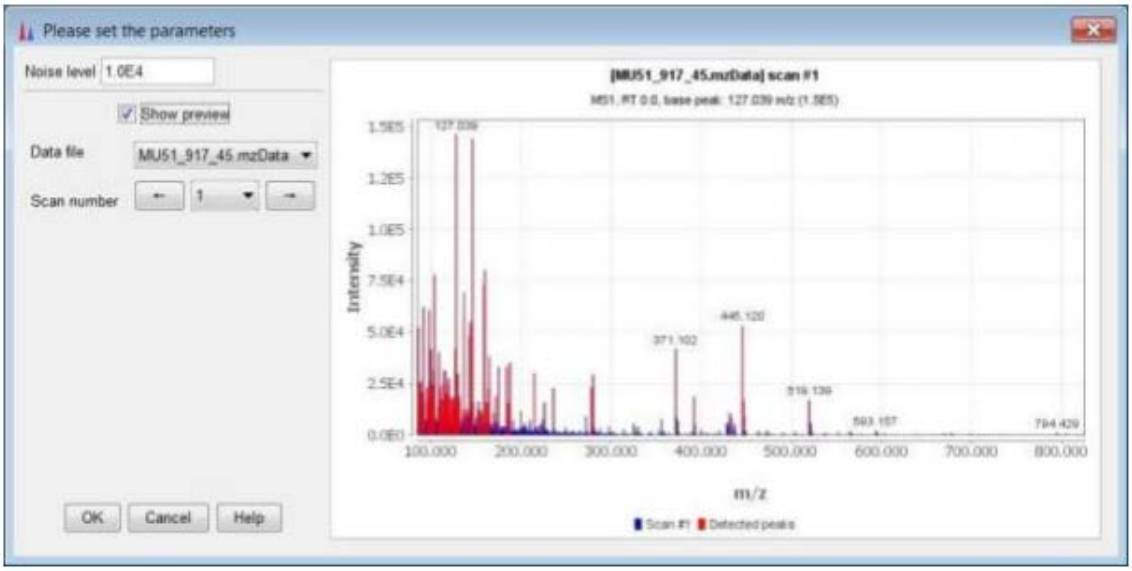

Figure 7 MZmine2 visualisation example – chromatogram plot

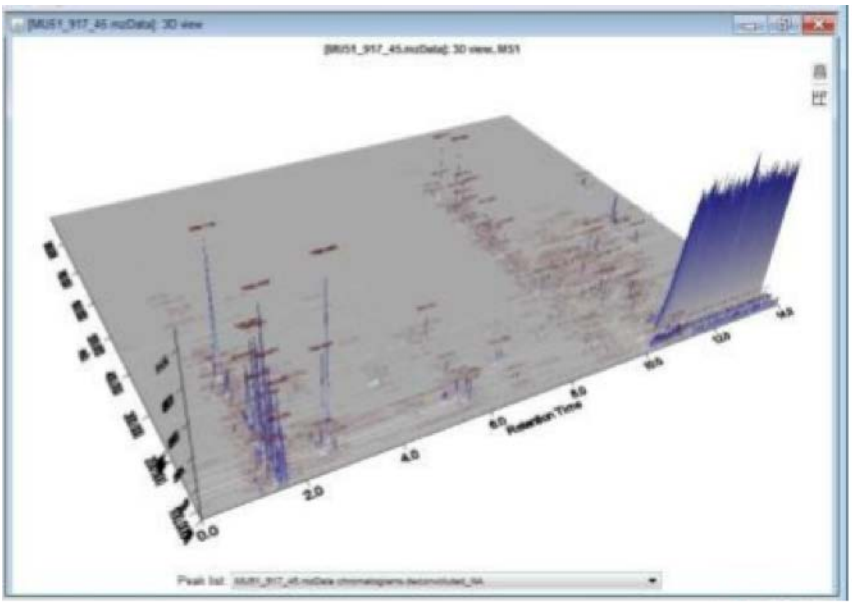

Figure 8 MZmine2 visualisation example – 3D view of the dataset

Availability – MZmine2 is available under General Public License 2 (GPLv2). MZmine2 can be obtained from (http://mzmine.github.io).

MZmine2 is implemented on our Docker cluster in the "Multi-omics Research Environment". Since the MZmine2 graphical user interface is needed for the productive data analysis we use special software (xpra server) to forward MZmine2 GUI and provide encrypted data communication between the docker container and user's local machine. On the local machine Xpra client can be used to display the MZmine2 interface. Xpra is an open-source multi-platform persistent remote display server and client for forwarding applications and desktop screens.

3.4. Sequence Imp PIPELINE

The *SequenceImp* pipeline incorporates bioinformatics tools such as *Bowtie*, *Reaper*, *Tally* and various R Bioconductor packages into a system for analysing high-throughput sequencing studies (microRNA sequencing) that can generate sequences in a FASTQ format. The pipeline is a Unix command line application that allows the simultaneous analysis of multiple FASTQ files through a single command. The pipeline uses pre-built annotation files to compute read/feature overlaps.

The pipeline proceeds via a series of sequential steps:

1. ***organise*** This sets up an analysis directory and configures the user information provided into a structure and formats that the pipeline can navigate.
2. ***reaper*** This uses the *Reaper* software to trim adapter sequences from reads and then uses *Tally* to collapse multiple identical sequences into a single non-redundant copy, while recording read depth as a separate parameter in the read header.

3. ***filter*** This filters reads based on a set of user defined parameters provided as part of the configuration file.
4. ***align*** This aligns reads to the relevant genome using *Bowtie*, providing output in either *Bowtie* or SAMformat. Metrics for assessing the features represented in a sample are plotted.
5. ***features*** This conducts preliminary feature analysis, either cross-referencing genome alignments with miRBase mature miRNA coordinates or comparing the processed reads to canonical repeat sequence data.

Sequence Imp pipeline is fully automated and adapted for the Docker cluster usage. Examples of QC plots that are automatically generated by Sequence Imp during the analysis process are available on Figure 7.

Availability – Sequence Imp pipeline is available under General Public License 3 (GPLv3). Sequence Imp can be obtained from:

ftp://ftp.ebi.ac.uk/pub/contrib/enrightlab/kraken/SequenceImp/src/seqimp-13-095/doc/imp.html

Figure 9 Sequence Imp pipeline's automatically generated QC plots

4. R ENVIRONMENT - R CLOUD

4.1. Data analysis in R Cloud

R Cloud is an R processing framework, scalable and distributed for exposure of R/Bioconductor packages to Java applications. It allows applications to perform R analysis on any biological data using numerous packages from Bioconductor and CRAN repositories. From the user perspective it means that R Cloud allows analysing –omics and clinical data by using preferred R packages. In addition, R Cloud allows running heavy R jobs in parallel. R Cloud has an interface component, which is running on a client machine whilst all R computations are done on a cloud platform.

R Cloud and tranSMART are connected and use a shared file system within the “Multi-omics Research Environment”.

For example, a researcher from the vertical projects can install and use R package “RobustRankAggreg” developed by University of Tartu (Kolde et. al), get the list of genes/transcripts and store them back into tranSMART with the scores calculated by the Robust Rank Aggregation algorithm.

Availability – R Cloud is available under General Public License 3 (GPLv3). R Cloud can be obtained from (https://github.com/andrewtikhonov/RCloud).

The more detailed description of R Cloud can be found in deliverables D13.1 and in D13.3.

5. CONCLUSION

Our solution for the data analysis for vertical projects consists of three components of "Multi-omics Research Environment": tranSMART, R Cloud and Docker cluster with a number of adapted –omics data processing pipelines. Docker cluster approach gives the possibility to add particular tools and pipelines when necessary. For example, at the moment we are adapting a LIMIX pipeline for different types of the QTL analysis. We are planning to use a LIMIX pipeline for the AD vertical project integrative data analysis. At the same time we are open for the discussions and are ready to add new pipelines for different types of –omics data analysis.

The "Multi-omics Research Environment" has been created to meet needs of EMIF vertical projects. For example, 1000 samples AD cohort consists of the following data types: clinical harmonized meta-data, metabolomics data, genomics data, etc. All these data types can be analysed in our instance of "Multi-omics Research Environment" available on Embassy Cloud: by using special –omics pipelines available on Docker cluster, by using R Cloud for the parallel R works or for the specific R analysis, by using tranSMART for the clinical data analysis. Figure 10 shows this example and the components where different data types are analysed. The tranSMART instance here is used to store harmonized clinical data, to stratify those data when needed and to select the need information for the –omics data analysis. The results of the downstream analysis can be stored back to the tranSMART to enrich the data.

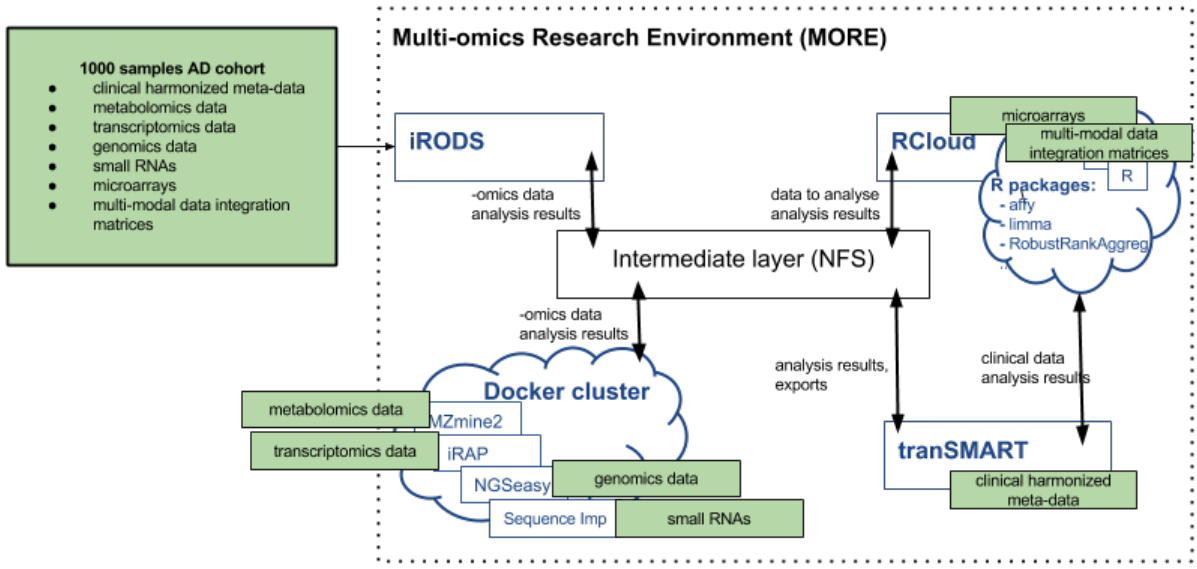

Figure 10 "Multi-omics Research Environment" usage for the 1000 samples AD cohort data analysis

From the user administration perspective usually different data types are uploaded by a number of individual institutions. For example, institute A generated metabolomics data, institute B generated genomics data and researchers from both institutes would like to perform the downstream analysis. All components of "Multi-omics Research Environment"

are using LDAP server for user authentication and authorization. Coming back to the example from Figure 10, first institute A uploads and pre-process metabolomics data, then analysis results are shared either through iRODS or through group permission mechanism on NFS with the institute B.

We will continue our work with "Multi-omics Research Environment". Our next steps include:

- **Collecting and analysing multi-omics data from 1000 samples AD cohort** by using "Multi-omics Research Environment"; in this process the users of the environment will be representatives from AD modalities (metabolomics, proteomics, genomics).
- **Gathering of EMIF AD users experiences** with the “Multi-omics Research Environment”.
- **Modification of “Multi-omics Research Environment”** which includes also possibility to delete exiting components or add new ones based on gathered EMIF users experience.

These modifications and final tuning will lead us to the three remaining deliverables: D13.5 (Data analysis and visualisation tools, including workflows for linkage with omics data v2) and eventually to the D13.6 (Final suite of modules and tools for data analysis, visualisation and linkage of EHR data with omics data).

The “Multi-omics Research Environment” can be easily modified for the needs of the EMIF-Metabolic Research Topic upon request since it has the generic potential.