www.emif.eu

European Medical Information Framework

Grant Agreement nº115372

D13.5 Data analysis and visualisation tools, including workflows for linkage with omics data v2

WP13 – Analysis, processing & visualisation methods and tools

V2.0
[Final]

Lead beneficiary: EMBL
Date: 13/09/2016
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**	13.5	**Title**	Data analysis and visualisation tools, including workflows for linkage with omics data v2
Work package	**Number**	13	**Title**	Analysis, processing & visualisation methods and tools

Delivery date	**Contractual**	Month 48	**Actual**	19/10/2016
Status	Current version / V2.0		Draft □	Final ☑
Nature	Report ☑ Prototype □ Other □			
Dissemination Level	Public ☑ Restricted □ Confidential □ (See note[1])			

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

[1] The title of the deliverable will be available at the public website (http://www.emif.eu). The availability of the deliverable document will depend on the degree of confidentiality that is assigned:

- **Public (PU):**
  - The executive summary will be available as soon as the document is submitted to the IMI JU.
  - A request button leading to a form will serve to request the full document.
  - After filling in some contact information, the user will be allowed to directly download the document through the website.
- **Restricted (RE):**
  - The executive summary will be available as soon as the document is submitted to the IMI JU.
  - A request button leading to a form will serve to request the full document.
  - Upon request, the PM will forward the request to the main author(s) to decide how to proceed.
  - In the Platform, active approval from the main author(s) and passive consent from partners will be needed to accept a deliverable sharing request.
  - A Non-Disclosure Agreement (NDA) may be placed, if needed.
- **Confidential (CO):** The title and the executive summary will be displayed on the website as soon as the document is submitted to the IMI JU.

DOCUMENT HISTORY

NAME	DATE	VERSION	DESCRIPTION
Natalja Kurbatova	13/09/2016	1.0	First draft
Rudi Verbeeck, Wouter Dhaeze	22/09/2016	1.5	Review and comments
Natalja Kurbatova	24/10/2016	2.0	Final report

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.
- **Ansible**. It's a free-software platform for configuring and managing computers that combines multi-node software deployment, ad hoc task execution, and configuration management.
- **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.
- **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 the deliverable D13.5 remains the same – needs of the EMIF verticals. We are continuing the development of "Multi-omics Research Environment" (MORE) described in the D13.3 and D13.4 by using EMBL-EBI Embassy Cloud that allows us to provide high performance computing. The tasks we performed for the deliverable D13.5 are focused on detailed testing of MORE components with publicly available datasets, gathering EMIF-AD user experiences, transferring MORE components from one cloud platform to the others to ensure flexibility of MORE. Cloud platforms we are working with include: VMWare, OpenStack and Amazon Web Services.

At the current stage, "Multi-omics Research Environment" (v2) consists of the following components:

- iRODS for data sharing
- Docker cluster for pipeline sharing & data analysis
- tranSMART for clinical data
- RCloud for R parallel computing

All components are connected and use a shared file system. Docker cluster and RCloud benefit from cluster computing.

Here we define Ansible scripts that are used for MORE maintenance and installation. We describe publicly available datasets we use for testing. Furthermore, we detail the data analysis we have performed using MORE.

Deliverable D13.5 is a continuation of WP13 previous deliverables D13.1 ("Evaluation of technologies and tools available for data analysis and visualisation"), D13.2 ("Data analysis tools for vertical projects v1"), D13.3 ("Data analysis and visualisation tools, including workflows for linkage with omics data v1") and D13.4 ("Data analysis tools for vertical projects v2").

KEY WORDS (Wordle style)[2]

[2] 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 have developed a “Multi-omics Research Environment” (MORE) as an open-source project. Source code is available on the github repository: https://github.com/olgamelnichuk/ansible-vcloud.

In the previous deliverables we’ve described MORE components and data analysis pipelines in detail. This deliverable focuses on comprehensive testing of MORE components with publicly available datasets, gathering EMIF-AD user experiences and transferring MORE components from one cloud platform to another to ensure flexibility. Cloud platforms we are working with include: VMWare, OpenStack and Amazon Web Services.

This deliverable describes the “Multi-omics Research Environment” at the version v2:

- MORE components (detailed description in D13.3);
- Omics data analysis pipelines (detailed description in D13.4);
- Ansible scripts for MORE maintenance and installation;
- Analysis we’ve done on MORE within EMIF-AD collaboration.

The "Multi-omics Research Environment” components are installed on EMBL-EBI’s Embassy Cloud VMWare with an allocation of CPU, RAM and storage resources. We tested the same components installation and maintenance on other cloud platforms: EMBL-EBI’s Embassy Cloud OpenStack and Amazon Web Services.

The users of the environment are the researchers from the EMIF AD vertical. User accounts are created on an individual basis by request and verification. All components are interconnected to provide solutions for integrative data analysis (see Figure 1).

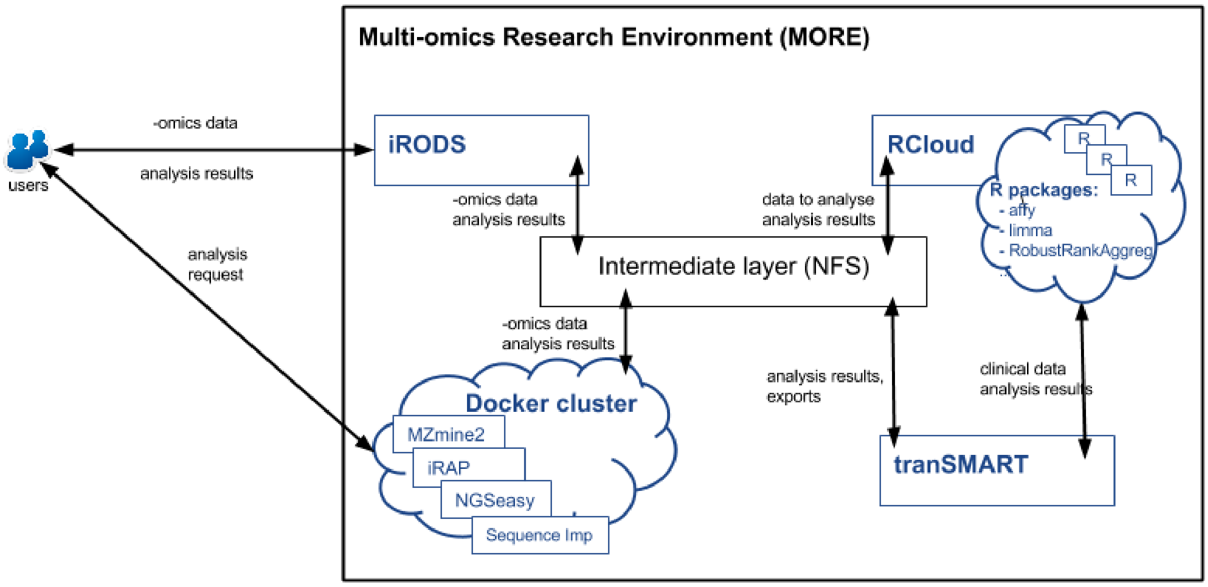

Figure 1: *Components of the "Multi-omics Research Environment"*

More details about MORE components and data security issues are available in the deliverable D13.3.

Docker cluster consists of VMs with installed docker software and adapted for parallel computing and dockerized -omics data analysis pipelines:

- 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;
- LIMIX pipeline for the different types of quantitative trait locus (QTL) analysis.

More details about MORE pipelines are available in the deliverable D13.4. The new LIMIX pipeline is described at page 15 together with the AD datasets that have been used for the comprehensive testing and analysis we've performed within EMIF-AD collaboration, namely genotype imputation and eQTL analysis for the "AddNeuroMed" dataset.

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

Ansible scripts for the environment installation, transfer and rescaling are described on page 13.

2. "MORE" COMPONENTS – OVERVIEW

At the current stage "Multi-omics research environment" (v2) consists of the following components:

- **iRODS** for data sharing;
- **Docker cluster** for pipeline sharing;
- **tranSMART** for clinical data;
- **RCloud** for R parallel computing.

All components are connected and use a shared file system. Docker cluster and RCloud benefit from cluster computing usage.

iRODS is the integrated Rule-Oriented Data-management System, a community-driven, open source, data grid software solution. Fundamentally, iRODS helps researchers, archivists and others manage (organize, share, protect and preserve) large sets of computer files.

In the current version (v2) of "Multi-omics Research Environment" iRODS consists of three parts: an iCAT Postgres database to store meta-data, an iCAT enabled resource server to store actual files and one more resource server for the data replication. All three components ideally should be located on different physical volumes to make sure that uploaded data are securely stored and will not be lost if one of the hard disks is down.

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 cluster of docker machines gives the possibility to run dockerized analysis pipelines in parallel.

We use a NFS server for data exchange: to make -omics data from iRODS available for the docker nodes to analyse them and to store analysis results. We have implemented the solution with OpenLava for workflow management amongst docker nodes. The dockerized pipelines available in the MORE v2 are described in the next section.

The open source **tranSMART** platform provides researchers with a single self-service web portal with access to phenotypic, 'omics, and unstructured text-based data from multiple sources, combined with search and analysis capabilities. A more detailed description of tranSMART can be found in deliverables D13.2 ("Data analysis tools for vertical projects v1"), D14.2 ("A data management solution for vertical projects, version 1") and D14.5 ("A data management solution for vertical projects, version 2").

tranSMART instance version 1.2 is installed on one of the virtual machines of Embassy Cloud. The Postgres database is located on another VM.

RCloud is an R processing framework that is scalable and distributed. It allows applications to perform R analysis on any biological data using numerous packages from Bioconductor and CRAN repositories. RCloud facilitates creation and management of parallel computing clusters and allows applications to perform computational tasks in parallel on the cluster. A more detailed description of R Cloud can be found in deliverable D13.1.

We are using the R Cloud component of “Multi-omics Research Environment” for R development to combine biostatisticians’ and bioinformaticians’ efforts in analysis of clinical and multi-omics data. R Cloud allows us to run heavy R jobs in parallel using a cluster of virtual machines under orchestration of OpenLava. Integration of tranSMART and RCloud allows us to accelerate R computing runs by tranSMART immensely.

3. “MORE” PIPELINES – OVERVIEW

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.

Docker cluster consists of VMs with installed docker software and adapted for parallel computing and dockerized -omics data analysis pipelines:

- iRAP pipeline to analyse transcriptomics sequencing data (https://github.com/nunofonseca/irap);
- NGSeasy pipeline to analyse genomics sequencing data (https://github.com/KHP-Informatics/ngseasy);
- MZmine2 to analyse proteomics and metabolomics LC-MS data (http://mzmine.github.io);
- Sequence Imp pipeline to analyse microRNA sequencing data (ftp://ftp.ebi.ac.uk/pub/contrib/enrightlab/kraken/SequenceImp/src/seqimp-13-095/doc/imp.html);
- LIMIX pipeline for the different types of quantitative trait locus (QTL) analysis (https://github.com/PMBio/limix).

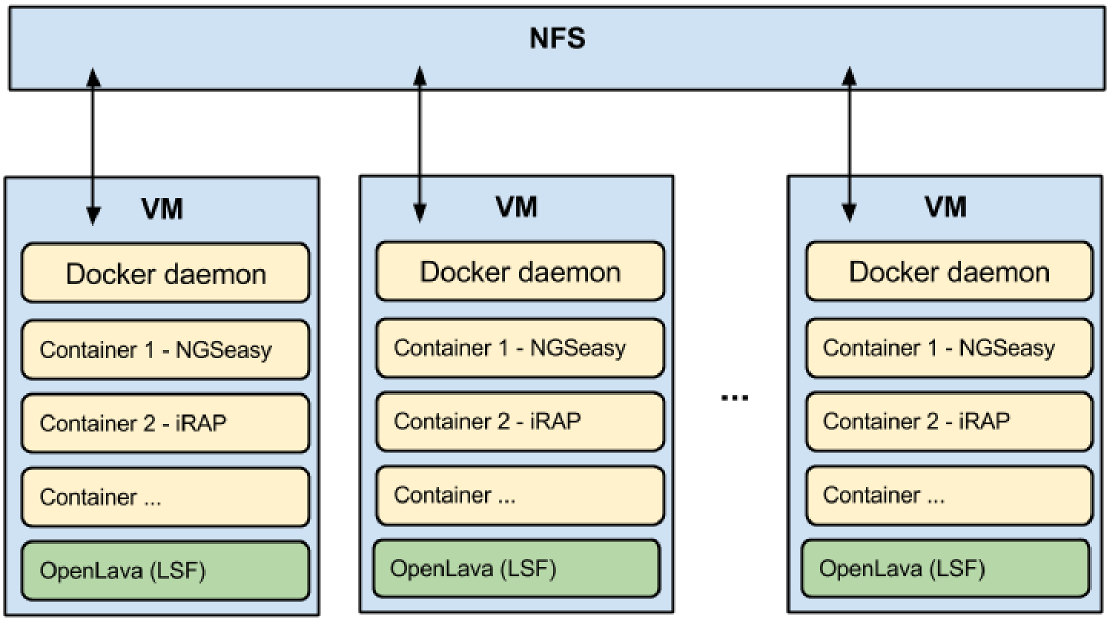

Figure 2: *Docker cluster and analysis pipelines in the "Multi-omics Research Environment"*

The dockerized pipelines can be added to the MORE upon users’ request.

4. ANSIBLE SCRIPTS

Ansible is a simple IT automation engine that automates cloud provisioning, configuration, management, application deployment, intra-service orchestration, and many other IT needs (https://www.ansible.com/how-ansible-works).

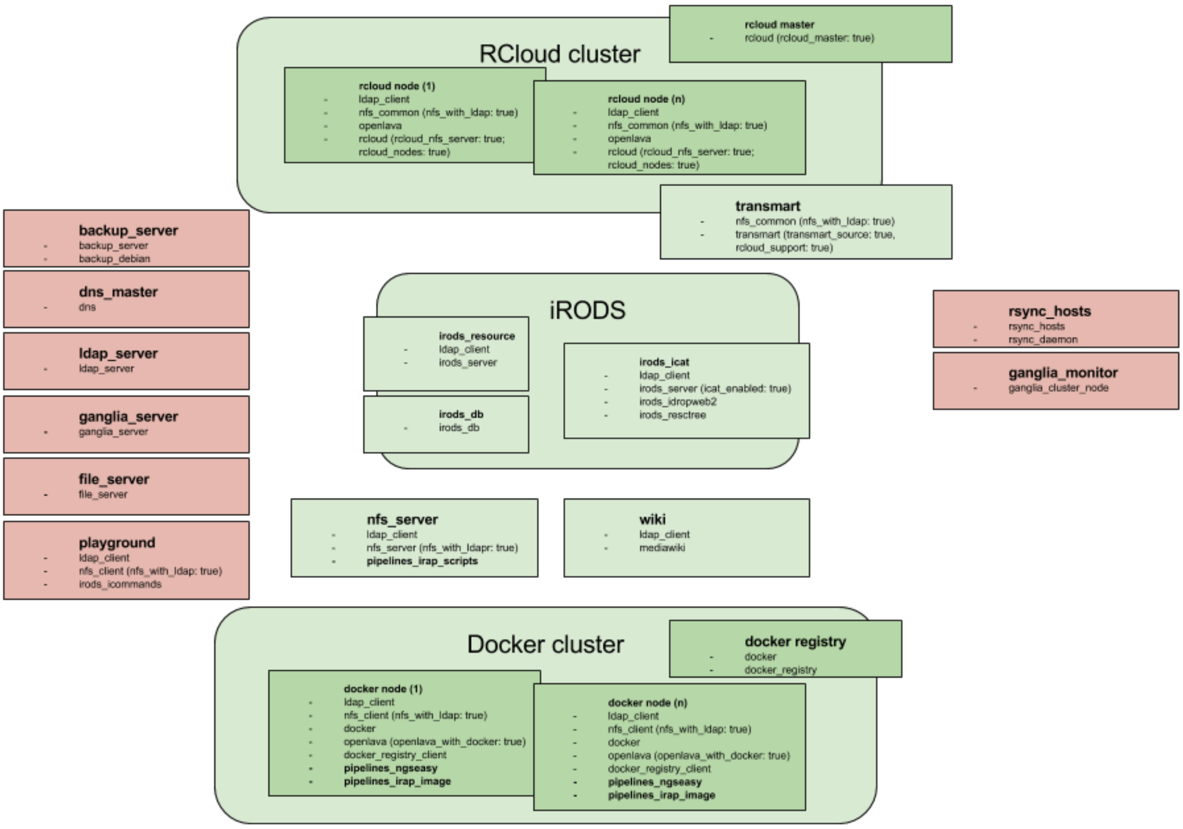

Figure 3 *Ansible scripts for "Multi-omics Research Environment" v2. Ansible scripts for the components accessible to the end user are coloured in green and system support/administration components are coloured in pink.*

From the technical point of view, the MORE can be divided into the following components:

- RCloud computational cluster that is shared between RCloud and tranSMART;
- iRODS with its three parts;
- docker cluster with analysis pipelines and
- a number of system support/administration components like LDAP server, Ganglia server etc.

On **Figure 3** components accessible to the end user are coloured in green and system support/administration components are coloured in pink. For each individual component we've created an Ansible script that, with minimal effort, installs/monitors/changes the component or applications it consists of.

Ansible scripts allow us to create a flexible, easily transferable and scalable environment.

Transferability here means that the same components and as the result the whole environment can be installed/maintained on different cloud platforms.

Scalability here means that additional components, like a docker cluster node, can be added into the system when needed. For example, when the monitoring tools (ganglia server) notify us that the docker cluster is fully used and there is a risk of lack of computational resources we can run a special Ansible script that will add new docker cluster nodes. In reality, system administrator's interaction is still needed but work is mostly automated.

5. DATASETS AND PERFORMED ANALYSIS

5.1. Dataset

For testing and as part of the collaboration with the EMIF-AD vertical we are working with the publicly available AddNeuroMed dataset.

AddNeuroMed is a cross European, public/private consortium developed for AD biomarker discovery (http://www.ncbi.nlm.nih.gov/pubmed/19906259).

The AddNeuroMed dataset consists of clinical information loaded into tranSMART, gene expression data from Illumina microarrays (723 samples), genetic data also done using chip technologies (1063 samples) and available in three batches. Gene expression data are available for two matches done on different Illumina microarrays. The intersection of the probes consists of 4867 genes. Samples from all three batches have been genotyped. However, the number of SNPs differ as well: 503585 for the batch 1, 651187 for the batch 2 and 662073 for the batch 3 after quality control (see page 15) and filtering procedures. After the imputation procedure (see page 16) merged genotyping matrix contains 12973077 genotypes (SNPs and indels).

Since not all the samples have expression levels measured, the intersection of two experiments (genotyping and gene expressions) has only 319 samples. So we ended up with a genotyping matrix (DNA matrix) having dimensions 1063x12973077 and a gene expression matrix (RNA matrix) with dimensions 723x4867.

5.2. Quality Control

Firstly, we’ve done quality control analysis for the genotyping data from batch 3 by excluding individuals with:

- gender mismatches,
- check and filter on MAF and call rate ((missingness < 3% and MAF > 1%) or (missingness < 1% and 1% < MAF < 3%));
- an individual call rate <=98%;
- individuals with autosomal heterozygosity outside ±4 standard deviation (SD) of the mean heterozygosity;
- duplicates and cryptically related by calculating identity by descent (IBD) estimates for all possible pairs of individuals and removing one of each pair with an IBD estimate >=0.1875 (the level expected for second cousins).

And excluding SNPs with:

- missingness >= 3%,
- missingness < 3% and MAF < 0.05,
- missingness < 1% and MAF>=0.05
- SNPs with HWE $p<10^{-4}$ in controls.

The call rate for a given SNP is defined as the proportion of individuals in the study for which the corresponding SNP information is not missing. Call rate of 95%, meaning we retain SNPs for which there is less than 5% missing data (**missingness**). More stringent cut points (e.g., less than 5%) may be employed in smaller sample settings.

MAF stands for minor allele frequency and refers to the frequency at which the second most common allele occurs in a given population.

HWE stands for Hardy Weinberg Equilibrium. Violations of HWE can be an indication of the presence of population substructure or the occurrence of a genotyping error. While they are not always distinguishable, it is a common practice to assume a genotyping error and remove SNPs for which HWE is violated.

After all the filtering procedures genotype data files have been remapped to genome reference hg19 build 37 (UCSC) and SNPs have been flipped when needed.

5.3. Genotype Imputation

We have merged all three batches of genotype data and cut files by chromosomes to reduce the requirements for the computational resources. Then we've performed the analysis called genotype imputation.

Imputation in genetics refers to the statistical inference of unobserved genotypes. It is achieved by using known haplotypes in a population, for instance from the 1000 Genomes Project in humans, thereby allowing to test initially-untyped genetic variants for association with a trait of interest. Genotype imputation hence helps tremendously in narrowing-down the location of probably causal variants in genetic association studies like eQTL. Genotype imputation requires a lot of storage and computational resources (www.g3journal.org/content/1/6/457.full).

We have used the computational cluster to run all the imputation jobs in parallel. The first job was phasing: statistical estimation of haplotypes from genotype data. For haplotype phasing we've used tool called SHAPEIT2 (http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004234).

A haplotype is, in the simplest terms, a specific group of genes or alleles that progeny inherited from one parent. Phased data are ordered along one chromosome and so from these data you know the haplotype. Unphased data are simply the genotypes without regard to which one of the pair of chromosomes holds that allele.

After haplotype phasing we've performed the imputation by using IMPUTE2 software (http://www.nature.com/ng/journal/v44/n8/pdf/ng.2354.pdf) and 1000 Genome reference panels. For the imputation we split the whole-chromosome analysis into manageable chunks of 5-Mb regions.

Post imputation jobs involved:

- Merge imputed chunks together to form a file for each chromosome;
- Merge files by chromosome to form a file for the whole genome;

- Clean IMPUTE2 output format;
- Filter by probability threshold 0.95;
- Convert output file into PLINK format (that was the original binary format);
- Quality control (MAF and call rate) followed by filtering out low quality SNPs;
- Create genotype matrix to satisfy LIMIX pipeline input file requirements.

The results of imputation are available for the EMIF-AD collaborators through a private FTP site.

5.4. eQTL Analysis

We've performed eQTL analysis by using "Multi-omics Research Environment" and LIMIX pipeline.

An eQTL (expression quantitative trait locus) is a locus that explains a fraction of the genetic variance of a gene expression phenotype. eQTL analysis trying to find those eQTLs by looking for strong statistical associations between expression levels and variants in genotyping data (www.ncbi.nlm.nih.gov/pmc/articles/PMC3682727/).

LIMIX is a flexible and efficient linear mixed model library with interfaces to Python that has been wrapped up into a docker pipeline. Genomic analyses require flexible models that can be adapted to the needs of the user. LIMIX is smart about how particular models are fit to safe computational cost.

First the analysis configuration file need to be created. It consists of:

- references to the matrices, like DNA matrix, RNA matrix, genotypes positions, covariates, etc.;
- number of parameters needed to be defined for the successful LIMIX run, like cis window which defines the region around gene for the cis (local) eQTL analysis, expression matrix transformation method, correlation method etc. and
- different thresholds like the minimal expression value etc.

We have performed a cis eQTL analysis – local search in the region next to a gene for associated genotypes (SNPs and indels) - and a trans eQTL analysis – distant search for genotypes associated with a gene. There are no associations found in the cis regions. However, there are a lot of trans associations and at the moment we are trying to adjust LIMIX parameters and interpret the results.

6. NEXT STEPS

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** A representative set of the different modalities (metabolomics, proteomics, genomics) in this cohort will be processed on the Multi-omics Research Environment.
- **Further gathering of EMIF AD user experiences** with the “Multi-omics Research Environment” especially for the analysis of 1000 samples AD cohort data.
- **Further modification of “Multi-omics Research Environment”** which includes also the possibility to delete existing components or add new ones based on feedback gathered from EMIF users.

These modifications and final tuning will lead us to the last remaining deliverable: D13.6 (Final suite of modules and tools for data analysis, visualisation and linkage of EHR data with omics data).