Supplementary Material

March 31, 2022

1 Datasets

This section describe the datasets analysed by MOMIC and table **??** summarises type of data used as input and their sizes.

GWES: Microarray: We have used data from the Gene Expression Omnibus (GEO) database [1]. The first dataset, GSE48350 contains microarray data from normal controls (aged 20-99 years) and Alzheimer's disease cases, from 4 brain regions: hippocampus, entorhinal cortex, superior frontal cortex, post-central gyrus. Changes in expression of synaptic and immune related genes were analyzed, investigating age-related changes and AD-related changes, and region-specific patterns of change. These AD cases were processed simultaneously with the control cases (young and aged) included in GSE11882 (GSE11882 dataset contains data exclusively from normal control brains). The second dataset, GSE15222, is also related to genetic control of human brain transcript expression in Alzheimer's disease.

GWES: RNASeq. We have produced a synthetic dataset simulating astrocytes from AD cases and controls, illustrating the diversity of data layers that can be integrated using the tools provided by MOMIC.

GWAS: We have prepared a dataset using data from the HapMap Project [3]. This preparation consists on updating the datasets to rsID, removing SNPs with minor allele frequency and adding fake case/control phenotypes. The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain.

Data from the International Genomics of Alzheimer's Project (IGAP) [2] have been used to illustrate the GWAS Meta-analysis protocol. IGAP is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative – EADI the Alzheimer Disease Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium – GERAD). In stage 2,

Table 1: Sizes of input datasets

Protocol	Dataset	Type	Size
GWAS	1KG	bed	1.2 GB
GWES Micorarray	GSE15222	csv (expression matrix)	17 MB
GWES Micorarray	GSE48350	CEL (raw data)	1.3 GB
GWES RNASeq	Synthetic data	FASTQ	7.2 GB
Proteomics	BLSA	csv (intensity matrix)	11 MB

11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2.

Proteomics: The dataset used for the proteomics testing is the anonymized Baltimore Longitudinal Study of Aging (BLSA) [4]. This study is a clinical research program on human aging that began in 1958. Volunteers of different ages join the study when they are healthy, and have follow-up visits for life. Visits last for multiple days. Participants are evaluated for many physical elements as well as for brain function. Physical tests are given. Information on mood, personality, and social aspects of life is also collected. This program has contributed more than any other research project to our understanding of aging.

2 Protocol diagrams

MOMIC currently compiles protocols for whole genome SNP data (GWAS), mRNA expression (both from arrays and RNAseq experiments) and protein data. Along with enrichment analysis and methods for combining heterogeneous data at different molecular levels. Figures 1 to 4 reveal the collection of notebooks designed for each protocol.

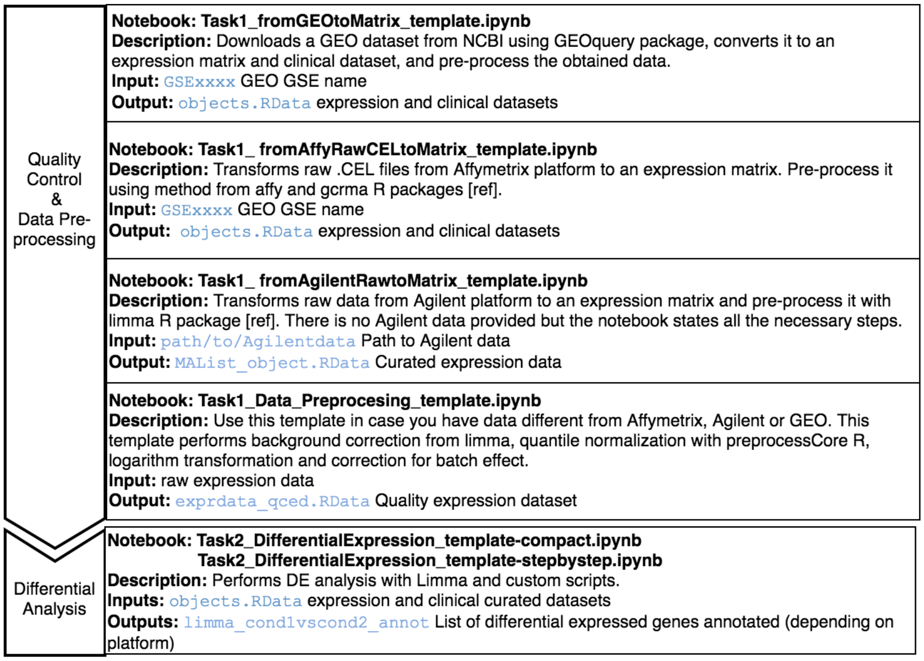
References

[1] E. Clough and T. Barrett. The gene expression omnibus database. In *Statistical genomics*, pages 93–110. Springer, 2016.

[2] J.-C. L. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. 45:1452–1458, 2013.

[3] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, et al. The international hapmap project. 2003.

[4] N. W. Shock. Normal human aging: The baltimore longitudinal study of aging. 1984.



Notebook: Task1_fromGEOtoMatrix_template.ipynb
Description: Downloads a GEO dataset from NCBI using GEOquery package, converts it to an expression matrix and clinical dataset, and pre-process the obtained data.
Input: `GSExxxx` GEO GSE name
Output: `objects.RData` expression and clinical datasets

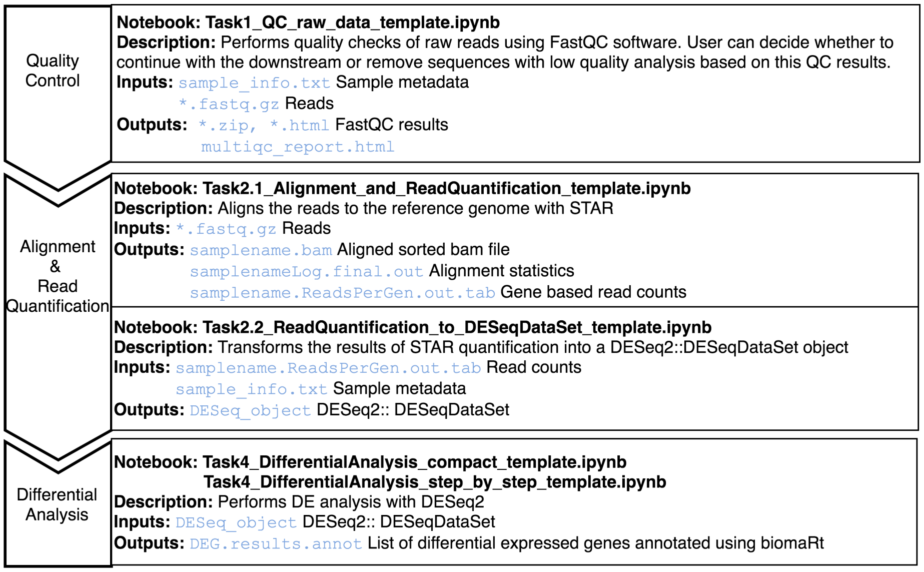
Notebook: Task1_ fromAffyRawCELtoMatrix_template.ipynb
Description: Transforms raw .CEL files from Affymetrix platform to an expression matrix. Pre-process it using method from affy and gcrma R packages [ref].
Input: `GSExxxx` GEO GSE name
Output: `objects.RData` expression and clinical datasets

Notebook: Task1_ fromAgilentRawtoMatrix_template.ipynb
Description: Transforms raw data from Agilent platform to an expression matrix and pre-process it with limma R package [ref]. There is no Agilent data provided but the notebook states all the necessary steps.
Input: `path/to/Agilentdata` Path to Agilent data
Output: `MAList_object.RData` Curated expression data

Notebook: Task1_Data_Preprocesing_template.ipynb
Description: Use this template in case you have data different from Affymetrix, Agilent or GEO. This template performs background correction from limma, quantile normalization with preprocessCore R, logarithm transformation and correction for batch effect.
Input: raw expression data
Output: `exprdata_qced.RData` Quality expression dataset

Notebook: Task2_DifferentialExpression_template-compact.ipynb
Task2_DifferentialExpression_template-stepbystep.ipynb
Description: Performs DE analysis with Limma and custom scripts.
Inputs: `objects.RData` expression and clinical curated datasets
Outputs: `limma_cond1vscond2_annot` List of differential expressed genes annotated (depending on platform)

Figure 1: GWES Microarray protocol



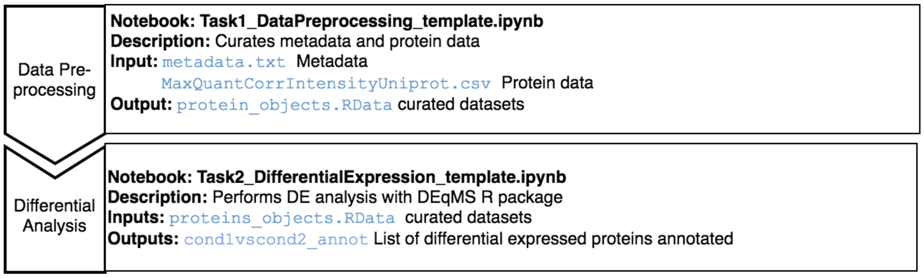
Notebook: Task1_QC_raw_data_template.ipynb
Description: Performs quality checks of raw reads using FastQC software. User can decide whether to continue with the downstream or remove sequences with low quality analysis based on this QC results.
Inputs: `sample_info.txt` Sample metadata
`*.fastq.gz` Reads
Outputs: `*.zip, *.html` FastQC results
`multiqc_report.html`

Notebook: Task2.1_Alignment_and_ReadQuantification_template.ipynb
Description: Aligns the reads to the reference genome with STAR
Inputs: `*.fastq.gz` Reads
Outputs: `samplename.bam` Aligned sorted bam file
`samplenameLog.final.out` Alignment statistics
`samplename.ReadsPerGen.out.tab` Gene based read counts

Notebook: Task2.2_ReadQuantification_to_DESeqDataSet_template.ipynb
Description: Transforms the results of STAR quantification into a DESeq2::DESeqDataSet object
Inputs: `samplename.ReadsPerGen.out.tab` Read counts
`sample_info.txt` Sample metadata
Outputs: `DESeq_object` DESeq2:: DESeqDataSet

Notebook: Task4_DifferentialAnalysis_compact_template.ipynb
Task4_DifferentialAnalysis_step_by_step_template.ipynb
Description: Performs DE analysis with DESeq2
Inputs: `DESeq_object` DESeq2:: DESeqDataSet
Outputs: `DEG.results.annot` List of differential expressed genes annotated using biomaRt

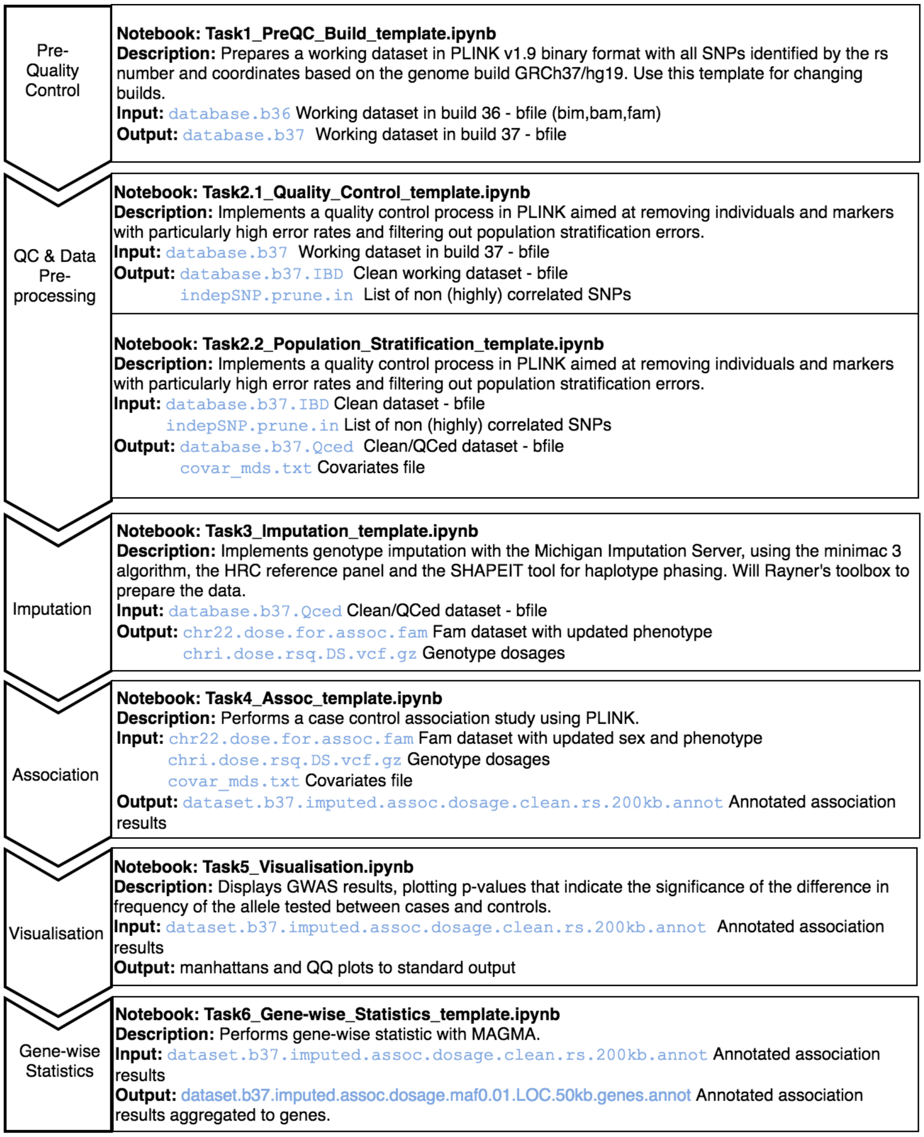
Figure 2: GWES RNASeq protocol



Notebook: Task1_DataPreprocessing_template.ipynb
Description: Curates metadata and protein data
Input: `metadata.txt` Metadata
`MaxQuantCorrIntensityUniprot.csv` Protein data
Output: `protein_objects.RData` curated datasets

Notebook: Task2_DifferentialExpression_template.ipynb
Description: Performs DE analysis with DEqMS R package
Inputs: `proteins_objects.RData` curated datasets
Outputs: `cond1vscond2_annot` List of differential expressed proteins annotated

Figure 3: Proteomics protocol



Notebook: Task1_PreQC_Build_template.ipynb
Description: Prepares a working dataset in PLINK v1.9 binary format with all SNPs identified by the rs number and coordinates based on the genome build GRCh37/hg19. Use this template for changing builds.
Input: `database.b36` Working dataset in build 36 - bfile (bim,bam,fam)
Output: `database.b37` Working dataset in build 37 - bfile

Notebook: Task2.1_Quality_Control_template.ipynb
Description: Implements a quality control process in PLINK aimed at removing individuals and markers with particularly high error rates and filtering out population stratification errors.
Input: `database.b37` Working dataset in build 37 - bfile
Output: `database.b37.IBD` Clean working dataset - bfile
`indepSNP.prune.in` List of non (highly) correlated SNPs

Notebook: Task2.2_Population_Stratification_template.ipynb
Description: Implements a quality control process in PLINK aimed at removing individuals and markers with particularly high error rates and filtering out population stratification errors.
Input: `database.b37.IBD` Clean dataset - bfile
`indepSNP.prune.in` List of non (highly) correlated SNPs
Output: `database.b37.Qced` Clean/QCed dataset - bfile
`covar_mds.txt` Covariates file

Notebook: Task3_Imputation_template.ipynb
Description: Implements genotype imputation with the Michigan Imputation Server, using the minimac 3 algorithm, the HRC reference panel and the SHAPEIT tool for haplotype phasing. Will Rayner's toolbox to prepare the data.
Input: `database.b37.Qced` Clean/QCed dataset - bfile
Output: `chr22.dose.for.assoc.fam` Fam dataset with updated phenotype
`chri.dose.rsq.DS.vcf.gz` Genotype dosages

Notebook: Task4_Assoc_template.ipynb
Description: Performs a case control association study using PLINK.
Input: `chr22.dose.for.assoc.fam` Fam dataset with updated sex and phenotype
`chri.dose.rsq.DS.vcf.gz` Genotype dosages
`covar_mds.txt` Covariates file
Output: `dataset.b37.imputed.assoc.dosage.clean.rs.200kb.annot` Annotated association results

Notebook: Task5_Visualisation.ipynb
Description: Displays GWAS results, plotting p-values that indicate the significance of the difference in frequency of the allele tested between cases and controls.
Input: `dataset.b37.imputed.assoc.dosage.clean.rs.200kb.annot` Annotated association results
Output: manhattans and QQ plots to standard output

Notebook: Task6_Gene-wise_Statistics_template.ipynb
Description: Performs gene-wise statistic with MAGMA.
Input: `dataset.b37.imputed.assoc.dosage.clean.rs.200kb.annot` Annotated association results
Output: dataset.b37.imputed.assoc.dosage.maf0.01.LOC.50kb.genes.annot Annotated association results aggregated to genes.

Figure 4: GWAS protocol