Author: Anna Nia

Date: 01/13/2020

Scripts that I have written for the manuscript titled "Efficient Identification of Multiple Pathways: RNA-Seq Analysis of Livers from 56Fe Ion Irradiated Mice"

Tutorial: Calculating Adjacency Matrix

This script takes the user through a practical example of generating an adjacency matrix from expression data using WCGNA, which can then be used for further calculations using the modularity maximization algorithm.

Dependencies:

- gdata
- dplyr
- Biobase
- convert
- IRanges
- edgeR
- GenomicRanges
- DESeq2
- limma
- rJava
- xlsx
- biomaRt
- enrichR
- devtools
- WGCNA

Input file requirements:

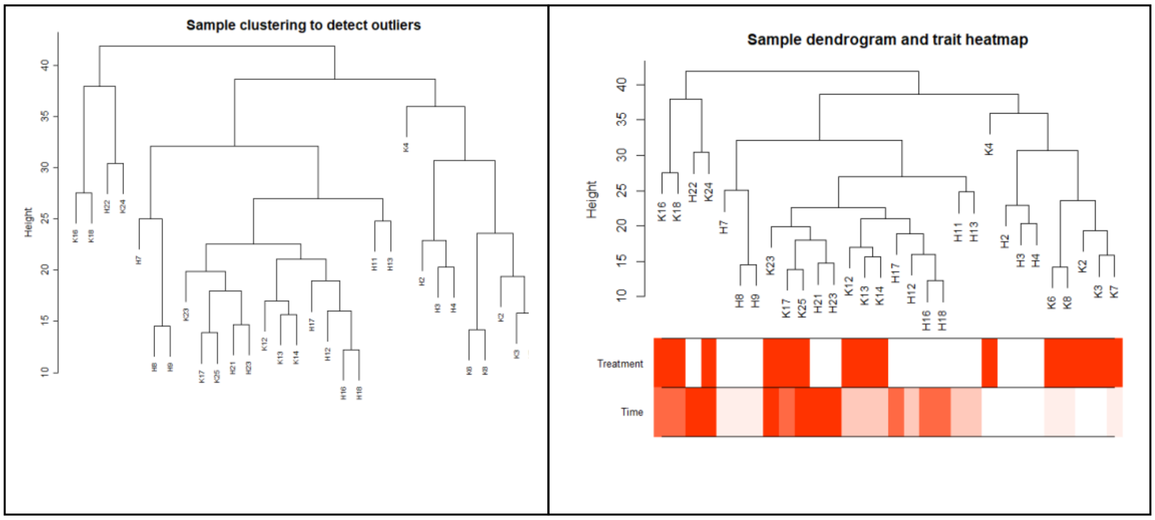
1. **Expression data**
    a. table of integer read counts, with rows corresponding to genes and columns to independent libraries. The counts represent the total number of reads aligning to each gene (or other genomic locus).
2. **Phenotype data**
    a. Phenotypic data summarizes information about the samples (e.g., sex, age, and treatment status; referred to as ‘covariates’). The information describing the samples can be represented as a table with S rows and V columns, where V is the number of covariates, and S the number of samples.

In this example we will analyze the gene expression of $_{56}$Fe irradiated C57 mice liver tissue samples. Reads were aligned to the mouse GRCm38 reference genome using the STAR alignment program, version 2.5.3a, with the recommended ENCODE options. The -quantMode GeneCounts option was used to obtain read counts per gene, based on the Gencode release M14 annotation file.

The following files will be used as inputs

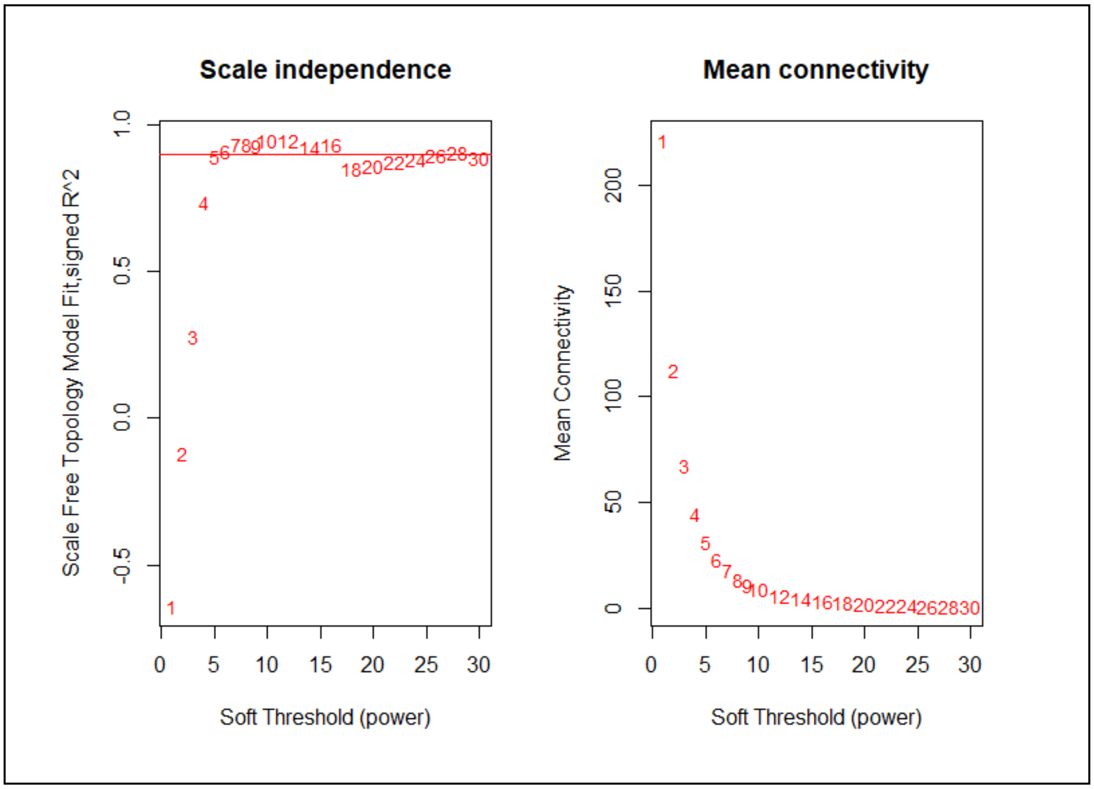
Expression Data File	Phenotype Data File	Description
raw_exprsData_Control_C57_1mo.txt	pData_Fe_Control_C57_1mo.txt	Month 1
raw_exprsData_Control_C57_2mo.txt	pData_Fe_Control_C57_2mo.txt	Month 2
raw_exprsData_Control_C57_4mo.txt	pData_Fe_Control_C57_4mo.txt	Month 4
raw_exprsData_Control_C57_9mo.txt	pData_Fe_Control_C57_9mo.txt	Month 9
raw_exprsData_Control_C57_12mo.txt	pData_Fe_Control_C57_12mo.txt	Month 12
raw_exprsData_Control_C57.txt	pData_Fe_Control_C57.txt	Combined Months 1-12
	pData3.txt	Experimental conditions

1. On line 22, change the working directory path to the location where all of the example files are located.
2. For every time point differential gene expression analysis is conducted using R software package edgeR. First, normalization factors are calculated to scale the raw library sizes. In addition, dispersion parameters based on generalized linear models (GLM) are estimated; in particular, common dispersion for negative binomial GLMs, trended dispersion for negative binomial GLMs using power method, and empirical bayes tagwise dispersions for negative binomial GLMs. Pairwise statistical tests are then conducted between $_{56}$Fe irradiated and non-irradiated control samples (at every time point) using a quasi-likelihood negative binomial generalized log-linear model applied to count data. The Benjamini-Hochberg correction is applied and genes with FDR≤0.05 & fold change ≥1.5 are extracted.
3. The expression values of genes identified as differentially expressed by edgeR are extracted from the “Combined Months 1-12” files. List of differentially expressed genes are merged across time points and deduplicated, leaving the lowest FDR value for each gene. The final list of differentially expressed genes can be re-filtered to a lower FDR, as desired. The raw counts for the selected genes are re-normalized using DESeq. (see edgeR, DESeq)
4. DESeq normalized expression values of differentially expressed genes are loaded into WGCNA for adjacency matrix calculation.
5. By inspection, remove any obvious outliers using a semi-automatic code that only requires a choice of a height cut. (see Tutorial for the WGCNA package for R)



6. We now read in the trait data or experimental conditions and match the samples for which they were measured to the samples in the expression data. We choose the power 16, which is the lowest power for which the scale-free topology fit index curve flattens out upon reaching a high value (in this case, roughly 0.90).

Choosing the soft-thresholding power: analysis of network topology



7. Adjacency matrix at the selected soft-thresholding power threshold is written out as "adjancency_matrix.txt"