DMAC Training Modules

Kyle Roell, Lauren Koval, Julia Rager

2021-06-03

Contents

1 Introduction **2**

2 Setting Up Your R Environment **2**

2.1 R and RStudio . . . 3

2.2 Scripting Basics . . . 3

3 The Basics for Data Organization **3**

3.1 Basic Data Manipulation . . . 3

4 Finding and Visualizing Data Trends **3**

4.1 Heat maps . . . 3

4.2 Clustering . . . 3

4.3 Data reduction (PCA) . . . 4

4.4 Basic Statistical Tests and Visualizations of Data . . . 4

5 Multi-Omics Analyses for Environmental Health **4**

5.1 Exposomics . . . 4

5.2 Transcriptomics . . . 4

5.3 Genome-wide MicroRNA . . . 4

5.4 Genome-wide DNA Methylation . . . 5

5.5 Proteomics . . . 5

6 Mixtures Analyses for Environmental Health **5**

6.1 Sufficient Similarity . . . 5

6.2 Mixtures Modeling through qgcomp . . . 5

7 Environmental Health Databases **5**

7.1 Comparative Toxicogenomics Database (CTD) . . . 5

7.2 Gene Expression Omnibus (GEO) . . . 5

7.3 NHANES . . . 5

1 Introduction

The UNC-Superfund Research Program (SRP) seeks to develop new solutions for reducing exposure to inorganic arsenic and prevent arsenic-induced diabetes through mechanistic and translational research.

The Data Analysis and Management Core (DMAC) provide the UNC-Superfund Research Program with critical expertise in bioinformatics, statistics, data management and data integration. Our goal is to support the data management, integration, and analysis needs of the researchers to reveal multi-factorial determinants of inorganic arsenic-induced metabolic dysfunction/diabetes.

2 Setting Up Your R Environment

R is a free, open source programming language

- Anyone can download and use
  - Doesn't require a license
  - Good for reproducible analyses
- Large, diverse collection of packages
- Comprehensive documentation

2.1 R and RStudio

2.1.1 Downloading R and RStudio

2.1.2 Installing R and RStudio

2.1.3 Installing and Loading Packages

2.2 Scripting Basics

2.2.1 Setting Working Directory

2.2.2 Importing and Exporting Files

2.2.3 Viewing Data

3 The Basics for Data Organization

3.1 Basic Data Manipulation

3.1.1 Merging

3.1.2 Merging processed data with metadata file?

3.1.3 Cast

3.1.4 Melt

3.1.5 Filtering & subsetting

3.1.6 Tidyverse stuff (pivots)

4 Finding and Visualizing Data Trends

4.1 Heat maps

4.1.1 pheatmap

4.1.2 heatmap2

4.1.3 superheat

4.2 Clustering

Examples with genomics: Rager et al. 2014

4.2.1 Hierarchical

4.2.2 K-means

4.3 Data reduction (PCA)

4.3.1 Visualize PCA Plot

4.3.2 Identify % of variance captured

4.4 Basic Statistical Tests and Visualizations of Data

Need an example dataset – maybe ELGAN shuffled/deidentified, with made-up environmental exposure column?

4.4.1 Normality

Kruskal wallis? The other one that I always forget? Shapiro wilks? - histogram

4.4.2 T-tests – column charts

4.4.3 Regression: linear regression and logistic regression

4.4.4 ANOVA + delicate commentary

4.4.5 Chi-squared test – box plots

4.4.6 Fisher's exact test

5 Multi-Omics Analyses for Environmental Health

5.1 Exposomics

5.1.1 Placenta Exposome

about to be submitted to EI Dust NTA data

5.2 Transcriptomics

5.2.1 DESeq2 / RNAseq

Wildfire dataset, available through GEO

5.3 Genome-wide MicroRNA

Rager et al. 2014 miRNAs

5.4 Genome-wide DNA Methylation

5.4.1 Illumina array data

```
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58499
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28368
```

5.5 Proteomics

```
Bailey et al arsenic dataset maybe?
```

6 Mixtures Analyses for Environmental Health

6.1 Sufficient Similarity

Botanicals example with chemistry and tox profiling – Julia has dataset

6.2 Mixtures Modeling through qgcomp

Could use published wildfire analysis here (Rager et al. 2021, STOTEN), or online example provide through Alex Keil's studies

7 Environmental Health Databases

7.1 Comparative Toxicogenomics Database (CTD)

7.2 Gene Expression Omnibus (GEO)

7.3 NHANES