STAT5003 Group Project
Semester 1, 2019

Background

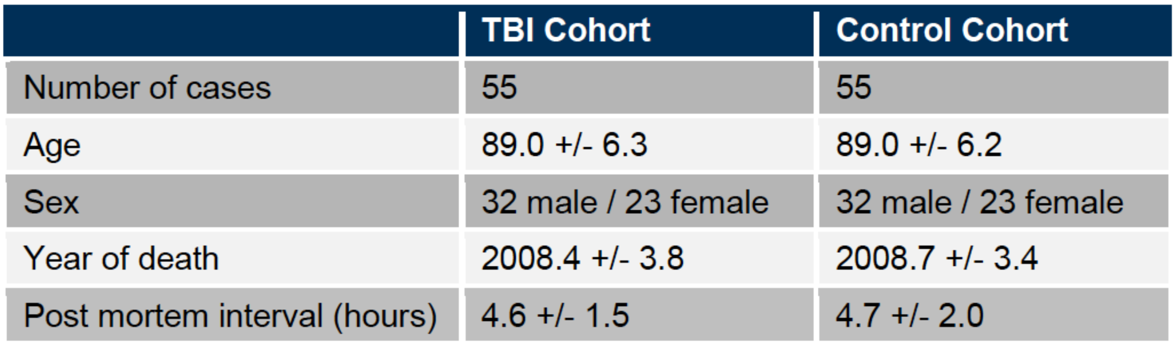
Dementia is a group of conditions characterized by gradual loss of cognitive functions such as memory and reasoning. In Australia, it is estimated that 1 in 10 people over the age of 65 had dementia in 2016; 11,000 people die of dementia each year, making it the second leading cause of death in Australia.

Alzheimer's disease is one specific form of dementia. It is a common form of dementia causing as many as 50 to 70% of all dementia cases. Alzheimer's disease is characterized by the build-up of abnormal proteins that form plagues and tangles in the brain. It is degenerative and currently there is no cure.

Data set details

The data set we will be using for this project is called the "Aging, Dementia and Traumatic Brain Injury (TBI) Study" provided by the Allen Institute for Brain Science. The project website is here: http://aging.brain-map.org/overview/home
The original aim of this study is to characterize the brains of people who had exposure to TBI versus controls. As such, the study design includes a cohort of 55 TBI cases with 55 age matched controls. Details of the cohort is in Table 1.

Table 1. TBI and control cohort characteristics.



	TBI Cohort	Control Cohort
Number of cases	55	55
Age	89.0 +/- 6.3	89.0 +/- 6.2
Sex	32 male / 23 female	32 male / 23 female
Year of death	2008.4 +/- 3.8	2008.7 +/- 3.4
Post mortem interval (hours)	4.6 +/- 1.5	4.7 +/- 2.0

The data sets produced by this study include the following:

1. **Histology and immunohistochemistry (IHC):** Image data and quantitative image metrics to assess β-amyloid, tau, and α-synuclein pathologies as well as the overall local pathological state of tissue samples from each donor.
2. **In situ hybridization (ISH):** High-resolution ISH image data of six canonical marker genes for astrocytes, oligodendrocytes, and neuronal subtypes in parietal cortex, temporal cortex, and hippocampus.
3. **RNA-Seq:** RNA sequencing data for temporal cortex, parietal cortex, cortical white matter, and hippocampus isolated by macrodissection.

4. **Protein quantification by Luminex:** Luminex assays to assess protein levels of neuropathological and immune system targets, complementing measurements from traditional antibody-based (IHC) methods.
5. **Isoprostane quantification:** Gas chromatography mass spectrometry (GC/MS) quantitation of isoprostanes to measure oxidative stress and to assess free radical injury.
6. **Specimen metadata:** De-identified clinical data (including Alzheimer's disease, dementia, and TBI diagnoses) for each case.

All data can be downloaded here: http://aging.brain-map.org/download/index
Data are in three main files under the following headings:

- De-identified clinical information (including Alzheimer's disease, dementia, and TBI diagnoses) for all donors included in the study. **(Data type 6)**
- Values for Luminex protein quantification, isoprostane quantification and immunohistochemistry pathology metrics for tissue specimens. **(Data type 1, 2, 4, 5)**
- Both normalized (as displayed in the RNA-Seq page heatmap) and unnormalized gene-level FPKM values for all samples. (zip) **(Data type 3)** This file is over 100Mb in size.

Instructions

Exploratory data analysis (**5 marks**)

Before we embark on building any predictive models, first step in any data science project is to perform some exploratory data analysis. One key aspect of this is to produce some insightful visualisations to summarise the data set (e.g. PCA or tSNE plots). You may also want to extract summary statistics and/or perform density estimation of some of the features, and to check the data for outliers and/or missing values.

There are two main parts to this project.

Part One (10 marks)
Design a classification procedure to predict dementia.

For this part, please use **data types 1-5**, and only the following **seven demographic and clinical features**:

- age, sex, apo_e4_allele, education_years, age_at_first_tbi, longest_loc_duration, num_tbi_w_loc

There are multiple columns with clinical diagnosis information. For this part of the project, use the column named "**act_demented**" as the dementia diagnosis.

For the RNAseq data modality, please use the **unnormalized FPKM values**.

This dataset is very high-dimensional so you may want to consider some feature selection procedure. You may also want to think about how to best combine features from different data modalities. Comment on whether some variables are more predictive than others.

Benchmark the performance of your classifier and comment on how well you expect it to generalize.

Part Two (10 marks)
Extract some other insight from the dataset.

Work with your group to define another problem that you can explore using this dataset. The problem should be cast as a classification, regression or unsupervised clustering problem (or some combination of the three). You are allowed to augment your analysis with any other data set(s) you can find.

Some example problems:

- Are gene expression and protein quantification age dependent? (Regression)
- Are gene expression and protein quantification sex dependent? (Classification)
- Do gene expression and protein quantification cluster by brain regions? (Classification/Clustering)

You will need to come up with some suitable metric to assess the performance of your analysis.

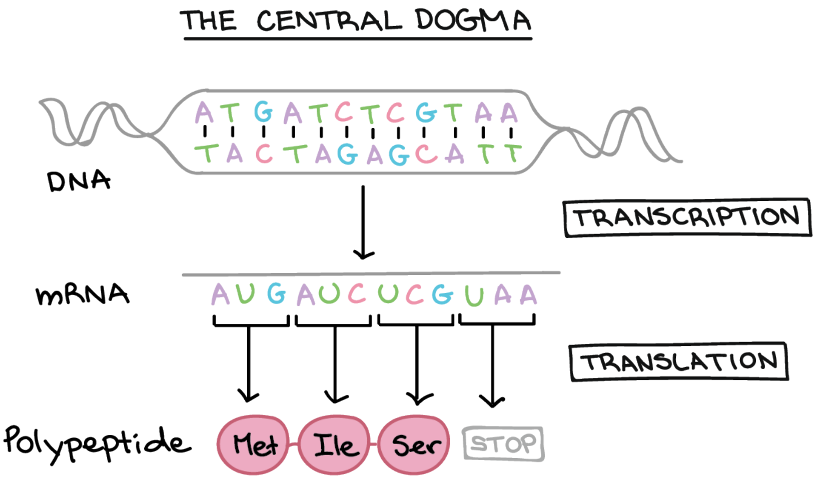
Marking rubric

Marks for this project will be determined by how well the above key points are addressed. Both the R markdown report and the oral presentation will be used in the assessment.

For the oral presentation, each group will have a maximum of **10 mins**. Each group will receive a single mark and all members in a group will have the same mark unless group members are noted to contribute unequally.

Key points	Exceptional	Proficient	Fair	Developing	Inadequate
(1) Exploratory data analysis. Visualization, clustering, density estimation, data missingness etc.	5	4	3	2	1
(2) Classification procedure to predict dementia; feature selection, curation, selection	7	6	4	2	1
(3) Validation and benchmark of dementia prediction results	3	2.5	2	1.5	1
(4) Additional analysis – define the problem and propose a suitable analytical procedure, include some benchmarking of results	10	8	6	4	2
(5) Presentation quality	5	4	3	2	1

Crash course in biology



Source: Khan Academy

Key ideas:

- Every person has one set of DNA that defines who we are. A human genome has around 3 billion DNA letters but only < 3% are located in **genes**.
- DNA are **transcribed** into RNA, usually this is done per gene. One gene can be transcribed many times, resulting in lots of copies of a gene's RNA
- mRNA are **translated** into proteins. More mRNA generally leads to more proteins – but not always the case, many layers of regulation.