**Integrated Assignment, Part 1**

**Exploratory Data Analysis of Biological Data using R  
May 21-22, 2015, Toronto, ON**

**Teaching Assistants**: Lauren Erdman and Ben Brew

The Integrated Assignment will provide you with an opportunity for practice with concepts that have been introduced during the lecture.

## Overview

The Cancer Cell Line Encyclopedia project (CCLE; [http://www.broadinstitute.org/ccle/home](http://www.broadinstitute.org/ccle/homed)) gathered multiple types of data on a large panel of human cancer cell lines, in order to predict drug response.

Results from the first phase of the study were published in:

Barretina et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603-607 (2012).

Advances in genotyping, transcriptional, and pharmacological profiling technologies enable researchers to characterize the interaction between molecular profiles and drug responses of cancer cells. Analyses of these interactions will facilitate the development of targeted therapies.

The data were retrieved from the Cancer Cell Line Encyclopedia. In order to reduce computing time, only a subset of genes is included in the expression and copy-number profiles.

We will use the following datasets from the Encyclopedia:

1. Sample information of the assayed cell lines (pheno)
2. Expression profiles of the cell lines (expr)
3. Copy-number profiles of the cell lines (cn)

To simplify the assignment, all the necessary pre-processing steps have been done, and the data has been incorporated into an R package. The pre-processing scripts are available for your reference.

## Descriptions of the datasets

A panel of cancer cell lines was characterized using various genomic technologies. For a total of 476 cell lines, the DNA copy-number and RNA expression are included in the dataset.

In this section, we will be working mainly with the DNA copy-number and the RNA expression data, visualized below as heatmaps for a subset of 16 cell lines and 14 genes. Note that there are many more cell lines and genes in the dataset. In each case, the cell lines appear along the rows, and the gene names along the columns.

|  |  |
| --- | --- |
| C:\Users\DavidS\Documents\Academia\courses\ta\workshop\assign\2014\ccle_cn.png  **DNA copy-number changes of select genes**. *Blue*, loss. *White*, balanced. *Orange*, gain. Basic concepts | C:\Users\DavidS\Documents\Academia\courses\ta\workshop\assign\2014\ccle_expr.png  **Relative RNA expressions of select genes**. *Green*, relatively low expression. *White*, medium expression. *Red*, relatively high expression. |

**DNA copy-number**

* Humans have a diploid genome and inherit one copy of each gene from each parent.
* Cancer cells often lose one or both copies of tumour suppressor genes and gain multiple copies of proto-oncogenes.
* The copy-number values are log2 ratios compared to a diploid reference as determined on the Affymetrix SNP 6.0 platform. Values near zero indicate no change, positive values indicate gain, and negative values indicate loss.

**RNA expression**

* RNA expression depends on cell state, responds to extracellular signals, and can reflect activities of biological processes or pathways.
* The values are continuous log2 measurements on the Affymetrix U133plus2 platform.

## Instructions

### Section 1 Retrieve data and set working directory

Open the *Stats2015\_IntegratedAssignment\_Questions.R* file. (Make sure you put it in a file w/all the downloads)

Next, set your working directory in R using the setwd(“PATH TO YOUR WORKING DIRECTORY”) command.

Note: once you’re cursor reaches the inside of the quotes in setwd(“”)

you can press ‘tab’ and you’re able to use your arrow keys to choose the folders you want sequentially.

### Section 2 Import the data into R

Load the CCLE data using load(“ccleCgc.rda”).

### Section 3 Examine the environment and objects

We will examine basic properties (e.g. variable class and dimensions) of the R objects imported into the environment. These properties are also displayed in RStudio.

Follow instructions in *Stats2015\_IntegratedAssignment \_Questions.R* and replace each ???? with R code to answer the questions.

### Section 4 Convert and rearrange the data:

Datasets typically need to be restructured into a common format that facilitates downstream analyses. This step can be laborious and time-consuming, so it has already been done. If you wish to see what had been done to the raw data, extract *CCLE\_preprocess.zip* and read the *README.txt* file therein.

Follow the instructions in *Stats2015\_IntegratedAssignment\_ Questions.R*.

### Section 5 Examine the data distributions

We will examine how the expression matrix (expr) and copy-number matrix (cn) are distributed, by plotting histograms and density plots.

We can plot histograms with hist() and overlay density plots using lines() and density().

Follow the instructions in *Stats2015\_IntegratedAssignment \_Questions.R.*

(*Optional*: examine how log() changes data distributions.

Work on manipulating axis labels and a plot title)

### Section 6 Explore the gene expression data

We will start by examining the expression of the tumour suppressor gene *TP53*.

Follow the instructions in *Stats2015\_IntegratedAssignment \_Questions.R.*

### Section 6.5 A taste of ggplot2

Generate a simple ggplot graph and navigate publicly available materials to change the color and size of your points.

You can use: <https://cran.r-project.org/web/packages/ggthemes/vignettes/ggthemes.html> to check out the themes available ☺

### Section 7 Principal component analysis

Principal component analysis (PCA) helps us visualize data by condensing the dimensionality of the data. We will use prcomp()and visualize the data along major PCA axes to discover patterns.

Follow the instructions in *Stats2015\_IntegratedAssignment \_Questions.R.*

**Integrated Assignment, Part 2 (Take-home)**

**Exploratory Data Analysis of Biological Data using R  
May 21-22, 2015, Toronto, ON**

This assignment is an optional continuation of the Integrated Assignment. Some of the topics may cover material from Day 2 of the workshop or may be slightly more advanced.

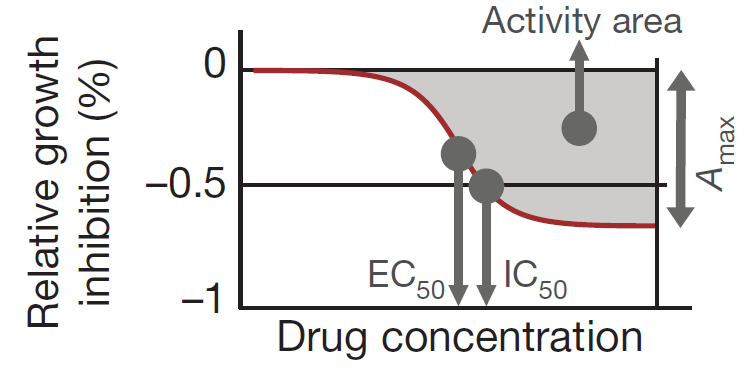
In addition to the datasets described in Part 1, the Encyclopedia also includes:

1. Mutation statuses of the cell lines for a gene panel (mut)
2. Pharmacological profiles of the cell lines in response to a panel of anti-cancer drugs (pharm)

These datasets are summarized in heatmaps below. Only a portion of the data is shown below.

|  |  |
| --- | --- |
| C:\Users\DavidS\Documents\Academia\courses\ta\workshop\assign\2014\ccle_mut.png  **Mutation status of select genes**. *Purple*, missense. *Orange*, nonsense. | C:\Users\DavidS\Documents\Academia\courses\ta\workshop\assign\2014\ccle_ec50.png  **Sensitivity to anti-cancer drugs**. *White*, insensitive. *Red*, sensitive. |

The pharmacological profiles were pre-analyzed in the CCLE project, as described in Barretina, 2012. Briefly, drugs were administered to cell cultures at 8 different concentrations and cell viability was measured and normalized to negative and positive controls (where 0% indicates viability is comparable to negative controls, and -100% indicate viability is comparable to positive controls). Sigmoid curves were fitted to the dose-response data. Activity area, maximum activity (Amax), effective concentration (EC50), and inhibitory concentration (IC50) were determined as illustrated in the below schematic. These four measures of dose-response assess the efficacy of a drug against cancer cell lines. A drug that is effective at inhibiting the growth of a cell line would have low IC50 and EC50, as well as high activity area and maximum activity in the dose-response curve against the given cell line.



## Instructions

(Continued from Part 1)

### Section 8 Explore the gene expression again

RUNX1 is a transcription factor that regulates the differentiation of haematopoeitic stem cells into mature blood cells. It is involved in many different cancers. Further, it is known to physically interact with JUN. Given that physically interacting proteins are likely to be co-expressed, are RUNX1 and JUN expression correlated?

Follow the instructions in *questions.R.*

### Section 9 Explore the pharmacological profiles

Four response variables were measured in the pharmacological study (ic50, ec50, act.area, and act.max). To analyze them visually, we can create scatter plots with two variables at a time.

Such scatter plots show the *bivariate joint distribution* of two variables, while histograms show how a single variable is distributed (i.e. *univariate distribution*).

Follow the instructions in *questions.R.*

### Section 10 Create publication-quality plots

The ggplot2 library has become popular among R users.

Follow the instructions in *questions.R* to create more beautiful plots with ggplot2.