Integrated Assignment, Part 1

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

Teaching Assistants: David Shih and Catalina Anghel

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) 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:

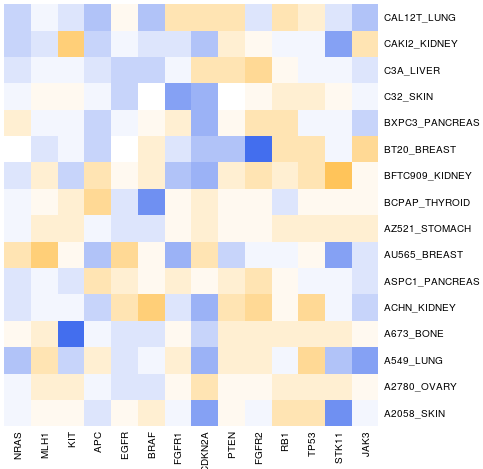
1. Sample information of the assayed cell lines (`pheno`)
2. Expression profiles of the cell lines (`expr`)
3. Copy-number profiles of the cell lines (`cn`)

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.

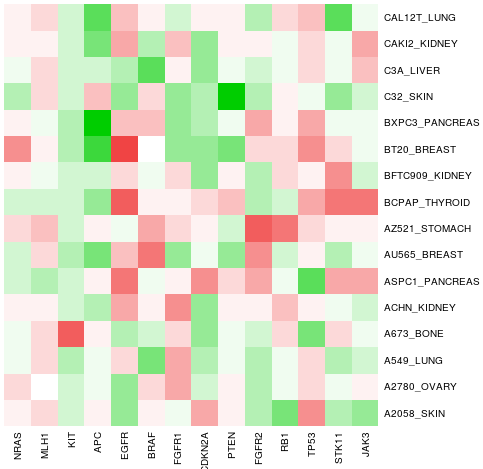
Descriptions of the datasets

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.



DNA copy-number changes of select genes. *Blue*, loss. *White*, balanced. *Orange*, gain.



Relative RNA expressions of select genes. *Green*, relatively low expression. *White*, medium expression. *Red*, relatively high expression.

Basic concepts

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 $\log_2$ 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 $\log_2$ measurements on the Affymetrix U133plus2 platform.

Instructions

Section 1 Retrieve and import data

Go to workshop wiki and download the R package *CCLE_0.1.1.tar.gz* (RStudio users on any platform) or *CCLE_0.1.1.zip* (non-RStudio users on Windows).
Mac: Right-click on the link and select *Save link as...* to prevent automatic extraction.

Install the R package from the local file.

Section 2 Import the data into R

Open the *Stats2015_IntegratedAssignment_Part1_Questions.R* file.

Load the CCLE library using `library(CCLE)`, and import the data using `data(ccleCgc)`.

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_Part1_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_Part1_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_Part1_Questions.R.*

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

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_Part1_Questions.R.*

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_Part1_Questions.R*.