

Module 3.2: Learn - Coding

3.2 Coding

In this module, all students in the CURE will analyze the expression of XIST in the CCLE. In addition to learning about how XIST expression relates to various features of the patient and tumor from which the cell line was derived, we can use XIST expression to predict the sex chromosome complement of the cell lines. Cell lines expressing high levels of XIST would have at least two X chromosomes, while cells expressing very low levels would only have one X chromosome. In later modules, we will be analyzing Y chromosome gene expression, so cell lines that have one X and one Y chromosome (eg, XY genotype) should have very low expression of XIST and higher expression of chromosome Y genes. Gene by gene, we are collecting evidence for a predicted sex chromosome complement and setting the stage for which sex chromosome genes can be involved in altered cell biology in tumor cells.

Assignment: Analyze XIST expression in CCLE

1. Download the CCLE data using the instructions below
2. Run the provided template code line by line to perform basic analysis of XIST expression in the CCLE
 - Learn how to read data
 - Learn how to subset the data for what particular questions
 - Learn different ways to plot the data so we can learn from the data
 - Learn how to write out an output file to share with others

Download CCLE data set and template code for analysis

If you are working with the RStudio Server on Sol supercomputer, we have downloaded the data and put it in a sl to copy to your scratch:

1. Fire up an RStudio Interactive session (see Module 1.2 Coding if you have forgotten how)
2. Click on the Terminal tab down near the Console tab
 - Here you are using a shell on Sol because that's where the RStudio Server is being hosted from
3. Navigate to your scratch drive

```
cd /scratch/username
```

4. Make a directory for your work

```
mkdir CCLE_project
```

5. Copy over expression data file

```
cp /data/compres/CCLE_CURE_2023/CCLE_RNAseq_genes_counts_20180929.csv CCLE_proje
```

6. Copy over cell line annotation file

```
cp /data/compres/CCLE_CURE_2023/Cell_lines_annotations_20181226.txt CCLE_project
```

7. Copy over template code

```
cp /data/compres/CCLE_CURE_2023/CCLE_gene_expression.Rmd CCLE_project
```

If you are working with RStudio Desktop on your computer, you can use the Sol Dashboard to download the files instructions, but refer to Module 1 Coding for more information.

1. Connect to the ASU VPN
2. Access the Sol Dashboard by opening a browser and going to <https://login.sol.rc.asu.edu/> (<https://login.sol.rc.asu.edu/>)
3. Perform authentication if prompted
4. Select the Files menu on the top left
5. Select Home Directory

6. Once your home directory contents are visible, select the 'Change directory' button by your directory path
7. Copy and paste the path to the data in the 'Change Directory' window under 'Path' : `/data/compres/CCLE_CL`
8. When the contents of this directory load up, you will see buttons to the right of the file names that have an option to download the three files (expression data file, annotation file, and template code Rmd) to your local drives

Now you should be able to navigate to your directory containing the code and data in your Files tab in RStudio and use the variables in your code.

Template code for gene expression analysis in CCLE



In order to help you get started with analysis in this class, the instructors have created an R Markdown to analyze the CCLE cell lines.


This file is meant to get you started by showing you how to read data in from files, pull out information for the genes you are interested in (XIST for this module, another sex chromosome gene for the next module), plot the data in sensible ways, identify trends that will allow us to make assertions about the data, and write out tables to output files that can be shared.

The idea is for you to go line by line and study the steps being taken and the functions used to take those steps. The code uses functions in R that you learned about in the tutorial you completed in Modules 1 and 2. The code also has comments to help you understand the idea behind each step. It is the instructors hope that this provides you a starting point and then you can modify parts of the code so that you can ask more questions as we study the data.


Using ggplot2 package in R for data visualization

In the template code, we use a very powerful package in R called ggplot2 to visualize the data (make plots). This package allows you to create a “grammar of graphics” or a base representation of the input data so that you can easily switch out between different data and aesthetic options. The figures made with ggplot2 can be exported in a variety of formats, which can help a lot when creating figures in publications and poster presentations.

This video explains the idea behind ggplot2 at a conceptual level so that you can better understand the way the r written: [ggplot2 Explained in 5 minutes!](https://www.youtube.com/watch?v=FdVy57oGJuc)  (<https://www.youtube.com/watch?v=FdVy57oGJuc>) [ggplot2 Explain transcript](https://docs.google.com/document/d/1jcG7l01ggy4oHr47TwkQYJ63gO-p1ZtEvO963GrCfYY/edit?usp=sharing)  ([https://docs.google.com/document/d/1jcG7l01ggy4oHr47TwkQYJ63gO-p1ZtEvO963GrCfYY/edit?usp=](https://docs.google.com/document/d/1jcG7l01ggy4oHr47TwkQYJ63gO-p1ZtEvO963GrCfYY/edit?usp=sharing)

This guide discusses the same concepts as the video but has example code with it so you can see the details on functions are used in R: https://uc-r.github.io/ggplot_intro  (https://uc-r.github.io/ggplot_intro)

Once you understand the way ggplot2 works, this is an awesome guide that shows examples of all the different k make and how to modify them to highlight various aspects of the data: <http://www.sthda.com/english/wiki/ggp> (<http://www.sthda.com/english/wiki/ggplot2-essentials>)

Finally, here is a cheat sheet that you can use to remind you how to do things when you are writing new code wit <https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf>  (<https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf>)

Interpreting XIST expression in CCLE

In the previous sections above, we learned that cells containing two X chromosomes will undergo X chromosome them using the expression of XIST, so cell lines that are XX genotype should have high expression of XIST. In X be low levels of XIST because it is activated by having two X chromosomes and only one is present. So we shou from female patients having high XIST expression levels and all cell lines from male patients having low expressi after executing the code in 'CCLE_gene_expression.Rmd' we see a bimodal distribution of XIST expression in ce patients and some cell lines from males having high expression of XIST.

Let's take a minute to think about why that can be the case.

For the cell lines from female patients, if we assume that they had an XX genotype in all cells before some cells t there are at least two possible reasons why cell lines from their tumors could lose expression of XIST. The first is inactivation was turned off by some kind of mutation in the tumor from which the cell line originated, allowing gen X chromosomes (two active X instead of one active X and one inactive X). The second is that a mutation was ac grew and divided in tissue culture in the lab. In order for cells to be able to grow in the tissue culture environmen petri dish placed in an incubator), many mutations have to take place, one of which could turn off X chromosome

literature to support both of these ideas— X chromosome reactivation has been observed in aggressive strains of cancers and gains of X chromosomes have been seen in many other cancers (more on this later modules), while reactivation has been observed in cell line models used to study cellular reprogramming which is happening in cells that are capable of growing indefinitely in lab are selected for. To see a review paper on X reactivation, look resources for section 3.1 Biology/Statistics. What do you think is happening? What do you observe as you explore expression data that makes you think this?


For the cells from male patients, if we assume that they had an XY genotype in all cells before some cells developed these cells that have high XIST expression have gained at least one X chromosome, again either in the original cell line originated or during the process of growing in culture. Gain of chromosome X has been reported in male there have been reports of X inactivation in several types of male cancers, but again this could be something that process of becoming a cell line.

If you wanted to test this theory, hypothetically, how would you do it? If we see certain trends in cell lines, how might whether that trend is found in human tumors?

Additional Resources

If you are curious to know exactly how the instructors got the data files from the CCLE website and were interested we made to make the data easier to work with in R, this section details it for you. Since we have the files you need you should not need to do this, but it's always good to know as much about how the data files you are working with possible so you can use that information during troubleshooting if needed.

To generate the csv file that contains the gene expression matrix, the instructors did the following before the course

1. Opened a browser and went to <https://sites.broadinstitute.org/ccle/>  (<https://sites.broadinstitute.org/ccle/>)
2. Selected 'Datasets' menu
3. Selected 'CCLE data' link
4. For 'Select a dataset to view:', pick 'CCLE 2019'
5. Scrolled down to 'CCLE 2019 All Files'
6. Found 'Cell_lines_annotations_20181226.txt'

7. Used the links to download it
8. Found 'CCLE_RNAseq_genes_counts_20180929.gct.gz'
9. Used the links to download it
10. Used software such as 7zip to extract the gct file (unzip)
11. Used Excel or other spreadsheet program to open the file
12. Removed first two rows which are a header specific to gct format
13. Saved the file as a comma-separated values file with the extension '.csv' (CCLE_RNAseq_genes_cou
14. Saved the annotation txt file and the data csv file in a folder so that you can easily set the path to this f