Module 2.1: Learn - Biology

2.1 Biology/Stats

For this research project, we will be downloading processed gene expression data from the CCLE. However, it is very important with big data science to understand how the data was generated so that you know what kinds of analysis can be used given the data you have and how to correctly interpret results that you attain from data analysis.

If you go to the more recent paper on CCLE, this is how the paper describes how the gene expression data was generated:

“RNA-seq reads were aligned to the GRCh37 build of the human genome reference using STAR 2.4.2a59. The GENCODE v19 annotation was used for the STAR alignment and all other quantifications. Gene level RPKM and read count values were calculated using RNA-SeQC v1.1.860. Exon–exon junction read counts were obtained from STAR. Isoform-level expression in TPM (transcripts per million) was quantified using RSEM v.1.2.22. All methods were run as part of the pipeline developed for the GTEx Consortium (**https://gtexportal.org** (https://gtexportal.org))61.”

https://www.nature.com/articles/s41586-019-1186-3#Sec7 (https://www.nature.com/articles/s41586-019-1186-3#Sec7) “

[From **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6697103/** (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6697103/)]

Let’s break this down step by step.

RNA Sequencing (RNA-seq)

In this course we will be using RNA sequencing data which can be used to determine gene expression. A gene is considered to be expressed when it is actively transcribed; some genes are constantly expressed while others are conditionally expressed and vary with tissue, age, sex, and response to environmental stimuli. If you need more information on the basics of gene expression (transcription) and its relevance to bioinformatics, refer to your Biostar Handbook: Biology for Bioinformaticians. RNA sequencing can be used to measure the expression of all genes at the same time, which is what makes it an ‘omics’ technology.

This video gives a visual explanation of how RNA sequencing works step by step. In the example they describe, they are looking for genes that are expressed at different levels in normal versus mutated cells, but in our case we will be looking at gene expression differences between different cell lines in the CCLE.

Video. Statquest: A gentle intro to RNAseq (https://drive.google.com/file/d/1nEuBZAVyx1yFkoHqmGUBOsIeJsvD1oY0/view?usp=sharing) [18:26] | **transcript (https://drive.google.com/file/d/1x_ZZ-I2hajDUjiEmgDYsMHKTKywEotcU/view?usp=sharing)** |

Preparing a sequencing library

The first step to RNA sequencing is to prepare a sequencing library, which means to generate a set of short sequences that can be read by the DNA sequencer. This is almost always done using a kit that can be purchased from companies that specialize in this like Illumina.

This figure shows what this process involves:

Figure. Preparing an RNAseq library. [1.1_C_PrepRNAseq.png (https://drive.google.com/open?id=1gH7wu63lqFrTyv9vJCjHDx4V1wPDghJv)]. The steps in preparing an RNA-seq library are to isolate the RNA, break it into small fragments, convert it into cDNA (complementary DNA), add sequencing adaptors, PCR amplify, and then perform QC to verify library concentration and fragment length.

1. Isolate the RNA from cells that were treated to prevent RNA degradation (stored in cold temperatures with RNAase inhibitors)
2. Break the RNA into fragments because the sequencing machine performs best with 200-300 bp fragments and mRNA transcripts can be thousands of bases long.
3. Convert RNA into double stranded cDNA because DNA is more stable than RNA and can be amplified using a PCR reaction
4. Add sequencing primers which will allow the sequencer to read the fragments and identify the nucleotide sequence
5. PCR amplify the library of fragments from the adapter to the end of the fragment
6. Perform quality control to verify library concentration and library fragment lengths to ensure the sequencing reactions were accurate and precise (quality control to make sure there weren't any unexpected technical problems).

Sequence of library with next-generation sequencer

After the library is prepared with fragmented cDNA that has been PCR amplified, the fragments are laid out in a grid called a flo cell. The machine attaches fluorescent probes to the bottom base of each sequence with a different color for each base (A, G, C, T). The machine takes an aerial photo from above where it can detect each base by color; after it registers the lit base in each sequence it washes off the color from the probe. Now the machine can attach the next row of probes and repeat the process until it determines the sequence of each fragment.

Since the average run can be around 400,000,000 reads, probe colors can be too light to detect or have "low diversity" which masks a probe of one color in a region saturated with probes of another color, making it difficult to identify with confidence which base was in that position. Either of these would be considered low quality bases and, if many are in the same sequence, can be filtered out when we trim.

Figure. The process of sequencing using a flo cell and fluorescent probes. [1.1_D_FloCell.png (https://drive.google.com/open?id=1Xtfa-dJmIbvOBZQPH0ZfTDIA2JOTIhm2)]. The probes attach to the first base in each sequence, are photographed, and then washed away; this process is repeated until the full sequence of each fragment is determined.

Human reference genome

The human reference genome is a sequence of DNA nucleotides accepted to be the field’s best knowledge of the sequence of the human genome. This sequence is assembled from high-quality sequencing reads and is meant to be a consensus sequence capturing variation across multiple populations. Thus when looking at the genome sequence of a specific human cell or human tissue, we can compare the sequence to the reference genome and note sites at which the observed genome is different from the reference genome (variants). In our case, we are going to use the human reference genome sequence to map our RNA sequencing to the genome and use annotations about where genes are to figure out which genes are being expressed. Additional information about how the human reference genome is generated can be found in the Additional Resources section below.

One important thing to know about the human reference genome is that there have been several recent releases and it’s important to know which one was used to generate the data you are working with. There is an international consortium called the Genome Reference Consortium that has used massive parallel DNA sequencing to produce sequencing reads that span the entire human genome and assemble them into a full sequence. They released a version of the human reference genome sequence most recently called GRCh37 in 2009 and GRCh38 in 2013. In 2022, another consortium called the Telomere-to-Telomere (T2T) released a full human reference genome said to not have any gaps (regions that could not be confidently assembled). Each of these can be reference genome sequences can be re-released with more annotation information in resources like GENCODE (https://www.gencodegenes.org/human/) and UCSC genomics (**https://genome.ucsc.edu/** (https://genome.ucsc.edu/)); each release of the reference genome is given a version number so that it is clear which release people were using when they did their analysis. Many studies and tools are currently operating with GRCh38 including the CCLE, but the field is slowly transitioning to the T2T reference genome. Further, there are tools called LiftOver that allow to map results from one reference genome assembly to another, but it is always preferred that you keep the reference genome the same for all analyses to ensure that the alignment step is as accurate as possible.

If you look at the download site for the GRCh38 human reference genome, you can see that it contains a lot of information for scientists to use including the total length and information about gaps and how the reads were assembled:

https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/ (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/)

If you to ‘Download Assembly’ in this link on the NCBI site, you can see that you can download the sequence as DNA, RNA, and protein along with GFF/GTF files which are annotations of where the genes, transcripts, coding sequences, etc are located by chromosome and positions on the chromosomes. When you sequence your own samples, you can align the samples to the reference genome sequence and then use the annotation to figure out if each aligned read maps to what chromosome and what known gene or known feature if any.

Sequence alignment to reference genome

Sequencing reads can be aligned or mapped to it to determine how many reads match to each gene using the coordinates of each gene given in the genome annotation. The figure below shows how sequencing reads are mapped to genes during the alignment step. Alignment can be done to the genome sequence (which is what was done for the CCLE) or to a transcriptome sequence generated from the genome sequence if you are looking for reads that cover splice sites (reads that span the junction between two exons).

Figure: Schematic of alignment of RNA sequencing reads after adapter trimming (**Source (https://nbisweden.github.io/workshop-ngsintro/2001/slide_rnaseq.html#18)**)

There are many algorithms available to align RNA sequencing reads to the human reference genome. The CCLE used a very popular one for RNA seq data called STAR, but another popular one is Hisat2. Both produce an alignment file (SAM or BAM format) which contains the positions in a chromosome where each sequencing read best lines up on the human reference genome sequence.

Quantifying gene expression (counts)

Once alignment is complete, reads mapped to each gene can be counted to give an estimate of the expression of that gene. In the methods, it is stated that the RSEM tool was used to calculate expression, but again many tools are available for this purpose. If you search online for information about this tool (Google for “RSEM gene expression” for example), you will find the paper that describes this method (see additional resources for link). The paper describes that RSEM actually provides a full workflow that includes tools to go from sequence reads, through alignment, and then to quantification, plus some other tools to visualize the results. But, based on the methods in the CCLE paper, we know they did not use this entire workflow, but rather only the last steps using the rsem-calculate-expression function which takes alignment files (CCLE paper says it used the STAR algorithm) and uses interim functions to finally produce an gene abundance. The paper describes how this method estimates gene expression: “The first is an estimate of the number of fragments that are derived from a given isoform or gene. We can only estimate this quantity because reads often do not map uniquely to a single transcript. This count is generally a non-integer value and is the expectation of the number of alignable and unfiltered fragments that are derived from an isoform or gene given the ML abundances. These (possibly rounded) counts may be used by a differential expression method such as edgeR [9] or DESeq [8]. The second measure of abundance is the estimated fraction of transcripts made up by a given isoform or gene. This measure can be used directly as a value between zero and one or can be multiplied by 10^6 to obtain a measure in terms of transcripts per million (TPM). The transcript fraction measure is preferred over the popular RPKM [18] and FPKM [6] measures because it is independent of the mean expressed transcript length and is thus more comparable across samples and species [7].”

Figure. Workflow schematic for the entire RSEM package. We will be just using data generated using the paths after rsem-calculate-expression because we know from the CCLE paper that they used other tools to do data preprocessing and alignment.

Reported sex versus sex chromosome complement

The CCLE has released annotation information about all of the cell lines it contains. For many of the cell lines, there is an assignment of “male” and “female” in a column marked “Gender”. This assignment is very likely to be a self-reported sex from the patient whose tumor was the original source of the cell line.

The basis of this project is that a reported sex of male or female does not truly indicate what sex chromosomes will be found in the cell line. There are several reasons for this:

1. Gain and loss of sex chromosomes have been observed in several cancer types (more on this in later modules)
    - If the tumor cell that was able to grow in lab tissue culture conditions to make the cell line originally had lost sex chromosomes, the cell line would be likely missing sex chromosomes too
    - If the originating tumor cells had not lost sex chromosomes, tumor cells typically exhibit genome instability and sex chromosomes could be lost as the cell lines get cultured in the lab

2. Sex chromosome loss has been observed in males and females as they age
   - There is a large range of values listed in the “Age” column of the CCLE cell line annotations
   - Cell lines from older cancer patients would have been more likely to have lost their sex chromosomes even without having developed cancer
3. A person describes themselves as male or female based on the external genitalia they see, and while people with a vagina, vulva, clitoris, etc are typically XX and people with a scrotum and penis are typically XY, that is not always the case
   - XXY individuals (Kleinfelter’s Syndrome) have two X chromosomes and 1 Y chromosome present with male reproductive organs and typically don’t know that they are not XY
   - X0 individuals (Turner’s Syndrome) have only one X chromosome present with female reproductive organs and often don’t realize they are not XX until they do not develop menstrual periods

Given all of these factors, it is better to determine the sex chromosome complement in each cell line using molecular evidence rather than self-reported sex of the patient. There are many genes that show sex differences in gene expression; we are focusing on genes expressed off of sex chromosomes themselves as a molecular assay to predict which sex chromosomes are present in each cell line and how the levels of those sex chromosomes compare across the data set in the context of other information listed for the CCLE cell lines.

Sex-chromosome independent effects (hormone levels)

When considering sex differences in cancer, we must also consider sex hormones which can cause sex differences in gene expression independent of genes on the sex chromosomes themselves.   Sex hormones, or gonadal hormones as they are hormones produced by the gonads, are affected by a variety of factors very early in development (see figure below). Many aspects that affect sex hormone levels are determined at fertilization. Sex chromosomes are passed to the offspring and call for specific regulatory mechanisms. XX embryos will have some mechanism of dosage compensation; one of the X chromosomes is inactivated to even out the expression of sex chromosome genes in genetic females compared to genetic males such that they each have only one active X chromosome. Humans have X chromosome inactivation where one of the X chromosomes in XX individuals is activated (more on that in later modules). Other species have imprinting where either the maternal or paternal X chromosome is selectively silenced. Furthermore, there are differences between sexes in epigenetics (chromosome structure) and metabolism (energy processing in the cell and across the whole body) that will affect how tumors form. After gestation and birth, levels of sex hormones are very low until they spike at puberty. After puberty they remain high in males and drop at menopause for females. These differences can have an effect on genes involved in DNA repair, oncogenes (genes that drive cancer development), tumor suppressor genes (genes that stop cancer development), and immunity.

Fig: Factors that affect sex (gonadal) hormones and hormone levels throughout life. Red line is females, blue line is males. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8930612/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8930612/)

Certain cancer types are dependent on sex hormones. Key sex hormones androgen, estrogen, and progesterone are highly involved in the development and progression of breast, ovarian, prostate and endometrial cancers. These cancers, cell lines from which are represented in the CCLE, depend on these hormones for high mitotic rates and increased cell proliferation. We should keep in mind while working with the CCLE data that the conditions cell lines are grown in are very different from conditions in the tumor when it was still in the body.

Journal Club: Method for predicting sex based on gene expression

Now that we have learned about how gene expression is measured and had an introduction to sex chromosomes, let’s read a paper that does something like what we are trying to do with the CCLE but for a totally different purpose. This paper uses RNA-seq to measure gene expression of pig embryos, specifically chromosome Y genes, to predict genetic sex.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6947224/pdf/genes-10-01010.pdf (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6947224/pdf/genes-10-01010.pdf)

As you are reading the paper, take note of the following things:

1. What genes are they using to predict sex? How did they select them?
2. How exactly did they make the prediction of male or female?
    - How does this differ from what we are doing?
3. How do they demonstrate that their method worked?

Please go to the Journal Club assignment for this module to participate in the discussion. Here are those Perusall tutorial videos again if you need them:

- Accessing Perusall through Canvas (https://www.youtube.com/watch?v=bs_Z_3wqib4) (Accessing Perusall From Within Canvas Video Transcript (https://docs.google.com/document/d/1ql6li6Au6ccO-xoTpQRM_ilF5Z6FMeGtbRGGi7BOD4Q/edit?usp=sharing))
- Intro to Perusall (https://www.youtube.com/watch?v=M8bOP7yF_6I) (Perusall Introduction Video Transcript (https://docs.google.com/document/d/1OPT_i7YrembK3518QiKaYcgClgsM-BRbuCCc7Y-BQXU/edit?usp=sharing))

Additional Resources

- Original paper describing the RSEM gene expression quantification method
  - https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323 (https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323)
- Human reference genome
  - https://www.genome.gov/genetics-glossary/Human-Genome-Reference-Sequence#:~:text=A%20human%20genome%20reference%20sequence,sequences%20generated%20in%20their%20studies (https://www.genome.gov/genetics-glossary/Human-Genome-Reference-Sequence#:~:text=A%20human%20genome%20reference%20sequence,sequences%20generated%20in%20their%20studies) .
  - https://en.wikipedia.org/wiki/Reference_genome (https://en.wikipedia.org/wiki/Reference_genome)
- Spectrum of sex differences in cancer
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8930612/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8930612/)
- Sex chromosome loss in aging
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1801353/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1801353/)