## Introduction

Every organism is made up of cell(s) that contain genes which accounts for the differences among different species. Genes are specific regions in the organism's genome which are transcribed into RNA transcripts during transcription. A diversity of mature RNA transcripts can be synthesized from a same gene via the transcription and alternative splicing process. During alternative splicing, Introns, which are non-coding regions of an RNA transcript, are "removed" while the remaining coding regions of the same RNA transcript, known as the Exons, are eventually joined. After which, the RNA transcripts will be translated into proteins that will be responsible for the differentiated functions of different cells.

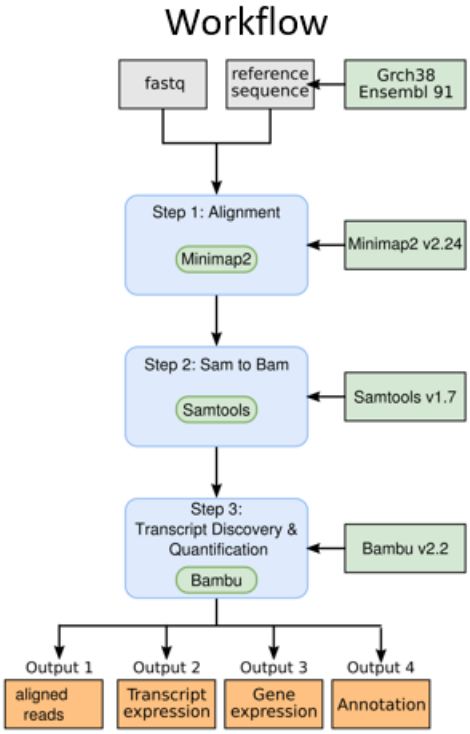
Therefore, if we would like to investigate the differential expressions of genes, we would need to analyse the transcriptomes (gene isoforms) synthesized. RNA-sequencing (RNA-Seq) is a technique used to profile transcriptomes, allowing for transcript discovery and transcript quantification.

In this project, I will be running a workflow to demonstrate a simple RNA-sequencing pipeline.

## Methods

The inputs for the workflow are downloaded from the S3 bucket using the AWS CLI. Then, Nextflow is used as the workflow manager tool to orchestrate and execute the workflow in a Unix terminal.

The parameters needed to start the pipeline are "reads", "refFa", "refGtf" and "outdir". All outputs from step 2 and 3 will be flushed into the "outdir" directory specified by the user.




**Step 1: Alignment with Minimap 2**

**Input:** reference sequence (refFa), RNA sample (reads)
**Output:** aligned reads (.sam file)

**Step 2: Sam to Bam Conversion with Samtools**

**Input:** output of Step 1
**Output:** aligned reads ($outdir/aligned_reads.bam)

**Step 3: Transcript Discovery & Quantification with Bambu**

**Input:** output of step 2, reference annotations (refGtf), reference sequence (refFa)
**Output:** transcript expression ($outdir/counts_transcript.txt), gene expression ($outdir/counts_gene.txt), extended annotations ($outdir/extended_annotations.gtf)

Using the .gtf file, we prepare the gene annotations object using the *prepareAnnotations* function. The *bambu* function then takes in the annotations, the reference sequence, and align reads file from previous steps to perform the transcript discovery and quantification. The outputs from *bambu* are then flushed into an output directory specified.

## Datasets

- **Full-sized dataset used:**

1. Reference sequence (.fa file): hg38_sequins_SIRV_ERCCs_longSIRVs.fa
2. Reference annotations (.gtf file): hg38_sequins_SIRV_ERCCs_longSIRVs_v5_reformatted.gtf
3. Reads (Fastq file): SGNex_Hct116_directRNA_replicate3_run1.fastq.gz
4. Aligned reads (.bam file): SGNex_Hct116_directRNA_replicate3_run1.bam

The reference genome used for this project is the human genome version 38. The RNA reads that is used for transcript discovery and quantification here comes from the HCT116 cell line which is a human colorectal carcinoma cell line originating from an adult male.

- **Scaled-down dataset used:**

1. Reference sequence (.fa file): hg38_chr22.fa
2. Reference annotations (.gtf file): hg38_chr22.gtf
3. Reads (Fastq file): A549_directRNA_sample2.fasq.gz

The reference sequence for the scaled-down dataset is only of chromosome 22 of the human genome version 38 while the RNA reads used comes from the A549 cell line which consists of lung carcinoma epithelial cells.

## Results

- **Screenshot of successful workflow execution (on scaled-down dataset):**

```
✓ 3m [20] ! nextflow run workshop/nextflow/workflow_longReadRNASeq.nf -resume \
           --reads $PWD/workshop/fastq/A549_directRNA_sample2.fastq.gz \
           --refFa $PWD/workshop/reference/hg38_chr22.fa \
           --refGtf $PWD/workshop/reference/hg38_chr22.gtf \
           --outdir $PWD/workshop/results/

     N E X T F L O W  ~  version 22.04.5
     Launching `workshop/nextflow/workflow_longReadRNASeq.nf` [cranky_swartz] DSL2 - revision: cf811cdc2c
     [-        ] process > MINIMAP2_ALIGN -
     [-        ] process > SAM_TO_BAM     -
     [-        ] process > BAMBU          -

     executor >  local (1)
     [6a/b7b525] process > MINIMAP2_ALIGN [  0%] 0 of 1
     [-        ] process > SAM_TO_BAM     -
     [-        ] process > BAMBU          -

     executor >  local (2)
     [6a/b7b525] process > MINIMAP2_ALIGN [100%] 1 of 1 ✓
     [7b/460a69] process > SAM_TO_BAM     [  0%] 0 of 1
     [-        ] process > BAMBU          -

     executor >  local (2)
     [6a/b7b525] process > MINIMAP2_ALIGN [100%] 1 of 1 ✓
     [7b/460a69] process > SAM_TO_BAM     [  0%] 0 of 1
     [-        ] process > BAMBU          -
```

```
executor >  local (3)
[6a/b7b525] process > MINIMAP2_ALIGN [100%] 1 of 1 ✓
[7b/460a69] process > SAM_TO_BAM     [  0%] 0 of 1 ✓
[53/981669] process > BAMBU          [  0%] 0 of 1

executor >  local (3)
[6a/b7b525] process > MINIMAP2_ALIGN [100%] 1 of 1 ✓
[7b/460a69] process > SAM_TO_BAM     [100%] 1 of 1 ✓
[53/981669] process > BAMBU          [  0%] 0 of 1

executor >  local (3)
[6a/b7b525] process > MINIMAP2_ALIGN [100%] 1 of 1 ✓
[7b/460a69] process > SAM_TO_BAM     [100%] 1 of 1 ✓
[53/981669] process > BAMBU          [100%] 1 of 1 ✓
Completed at: 06-Sep-2022 06:31:55
Duration    : 3m 44s
CPU hours   : 0.1
Succeeded   : 3
```

**Transcript discovery & quantification of full dataset with Bambu**

- **Number of novel transcripts discovered:** 972
- **5 most highly expressed genes:**
  "ENSG00000198786", "ENSG00000156508", "ENSG00000140988", "ENSG00000111640", "ENSG00000184009"
- **Minimum number of transcripts per gene:** 1
- **Maximum number of transcripts per gene:** 193
- **Average number of transcripts per gene:** 3.44

**Summary/Discussion**

This project is mainly to demonstrate an RNA sequencing processing pipeline which takes an RNA reads sample to align against a reference genome to discover novel transcripts and genes which will allow scientists to further investigate the differential gene expressions' effects on abnormal cells.

Due to the limitations of the compute power of my laptop, I could only run the workflow on the scaled-down dataset from the week 3 workshop and use the already aligned bam file from S3 to run the transcript analysis on the full dataset with Bambu in R.

Full sized genome datasets are usually huge in size and will require more scalable data processing pipelines. The bottleneck that is causing the pipeline to break when ran on the full dataset is at the nucleotides alignment stage with Minimap2. Perhaps an improvement that can be explored is to use index partitioning when running the Minimap2 stage. This will likely reduce the amount of DRAM usage but might increase the latency. If given a budget, running the pipeline on GPUs on cloud platforms can also be considered. Especially for use cases where we need to process multiple transcriptome samples at once to compare the transcript discoveries, the pipeline will need to be tweaked to accommodate such an increase in memory requirements.

**References**


1. https://www.nature.com/scitable/definition/intron-67/#:~:text=Introns%20are%20noncoding%20sections%20of,for%20proteins%20are%20called%20exons.
2. https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461
3. http://sg-nex-data.s3-website-ap-southeast-1.amazonaws.com/
4. https://github.com/GoekeLab/sg-nex-data
5. https://www.nextflow.io/blog/2021/setup-nextflow-on-windows.html
6. https://bioconductor.org/packages/release/bioc/vignettes/bambu/inst/doc/bambu.html
7. https://www.creative-biogene.com/support/hct-116-cell-line.html
8. https://www.synthego.com/a549-cells#:~:text=A549%20cells%20are%20lung%20carcinoma,to%20work%20with%20these%20cellss