Bioinformatics Workflow: Transcript Quantification of MSC populations.

Barry Digby
ADB10-19, Áras De Brún, National University of Ireland, Galway
b.digby237@gmail.com

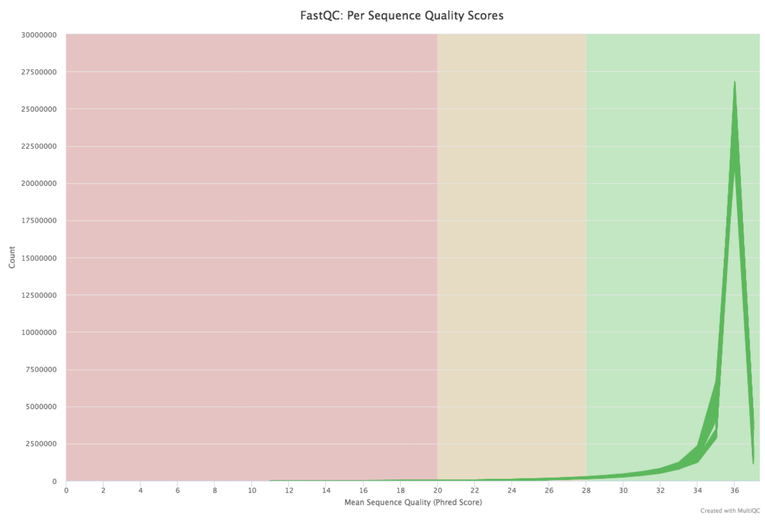
OrbsenTherapeutics
Redefining Cell Therapy

1. Experiment Design, Hypothesis

The experimental design included 32 total samples of RNA, originating from 4 donors. Each donor provided 8 samples, which can be further subdivided into the treatment group and the control group. Within each group, the cell types are classified as plastic adherent (PA), CD362 treated Aria, CD362 TYTO and untreated Aria. The goal of the researcher was to perform differential gene expression on each of the cell types. Statistically, this can be carried out by testing the null hypothesis that there is no difference between cell types. Results will be reported in p-values, and adjusted p-values, capturing genes that reject the null hypothesis and thus highlight the gene-level differences in cell type populations.

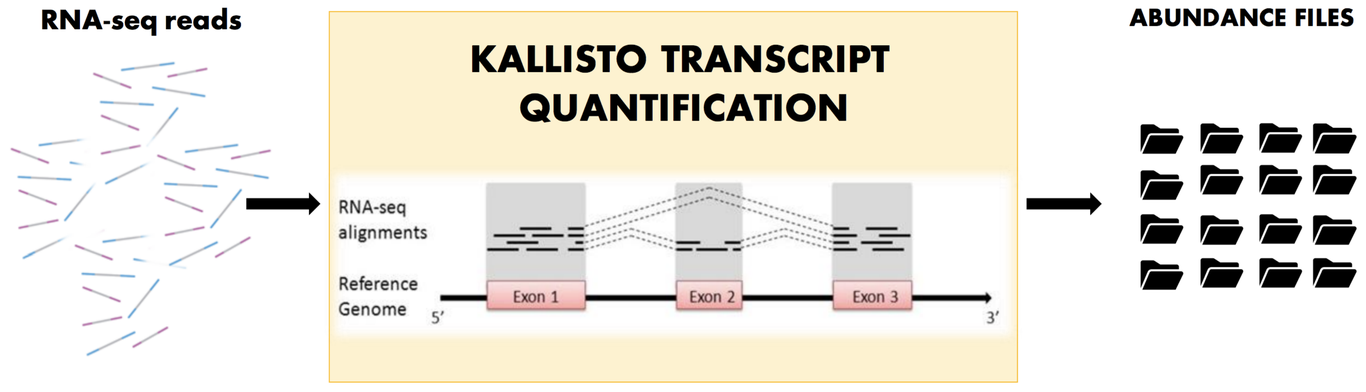
2. QC/Trimming Data

The raw data must be trimmed to ensure maximum read quality for downstream alignment. This includes removing adapter sequences and trimming poor quality bases from the reads. The figure below shows the majority of reads have a base calling accuracy of 99.97488114%.



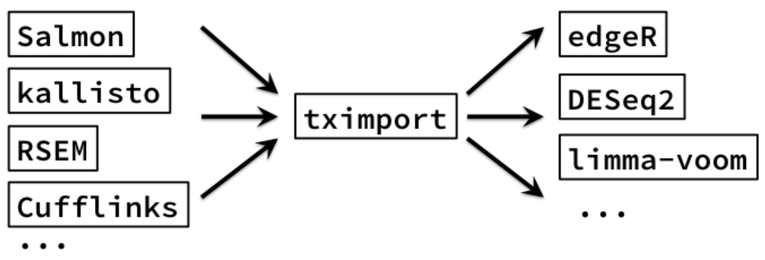
3. Transcriptome Profiling

To carry out transcriptome profiling using the trimmed high quality reads, Kallisto was used. Kallisto works by building a reference index for the human genome, and mapping the RNA-seq reads to the indexed genome using pseudoalignments. Pseudoalignment of reads preserves the key information needed for quantification, and is executed at a much faster rate than other available quantification tools. The output files from Kallisto contain the transcript abundances.



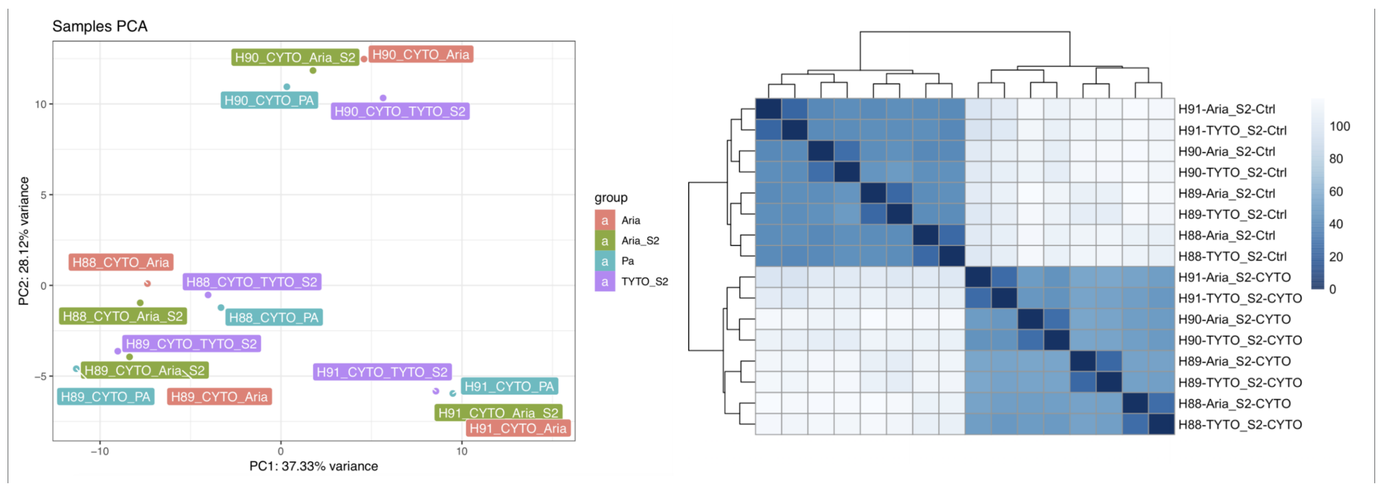
4. Transcripts to Gene counts

Kallisto outputs transcript abundance estimates, however this quantification method can not be used downstream for gene-level analysis. In order to correct this the **tximport** package must be used to import and summarize transcript-level abundance estimates and match them to gene ID's for gene-level summarization.



5.DESeq2

DESeq2 uses the summarized gene counts as input, and can produce useful plots for data quality assessment and quality control. Below are examples of PCA (left) and sample to sample distance heatmaps (right), showing how the samples tend to cluster together. In this experiment the samples tended to cluster together by donor, indicating the inherent genetic variation between samples outweights the variation of celltype effect.



6. Differential Expression

To conduct differential expression analysis between the cell types, the sample donor variation must be accounted for. This can be modeled in DESeq2 using:

$$\text{design} = \sim \text{Donor} + \text{celltype}$$

This design controls for the sample donor variation, while testing for the effect of cell types. Up regulated and Down regulated genes were reported in output files, subject to the following statistical filtering: Independent hypothesis weighting, adjusted P-value < 0.05.

To further elucidate the functions of the differentially expressed genes, the genes were mapped to their Biological Processes. The output files contain an ordered list, according to P-value, containing the Biological Process and the differentially expressed genes involved in the process.

7. Results

Results have been uploaded to Github for convenient access, found at the following link: https://github.com/BarryD237/Galway_Genomics_2019. The page hosting the results contains a number of directories:

1. **QC Report**: A data quality assessment of the 32 samples, identifying the main sources of variation in the dataset.
2. **Differential Expression**:
   - This DE results are divided by condition: Control samples and CYTO treated samples.
   - Within the sample condition folders there are pairwise comparisons of each cell type, 12 in total. Each folder contains the following files:
     (a) Up regulated: up regulated genes in the cell type.
     (b) Down regulated: down regulated genes in the cell type.
     (c) GOresults_up: Biological Processes the up regulated genes are involved in.
     (d) GOresults_down: Biological Processes the down regulated genes are involved in.
   - It is important to note that a folder called "vs_Aria" means the reference cell type is Aria. In this folder, TYTO_S2 vs Aria means that differentially expressed genes are in the TYTO_S2 cell type when compared to Aria.

Naming convention: CD362 Aria = Aria_S2. CD362 TYTO = TYTO_S2.
Plastic Adherent = PA. CD362 negative Aria = Aria.