## Steps

#### Alignment

TopHat is deprecated, so I wasn’t sure what RNA aligner would be best to use. I did some research and decided to use STAR. The dataset isn’t that large, so I wasn’t worried about speed, and STAR seems to be very popular now for this sort of task.

#### Processing

The full process can be run by calling bash ./Question1.sh. This script will do the following:

1. collect the necessary files.
2. install the required dependencies.
3. index the Danio genome using STAR genomeGenerate.
4. align each of the RNA reads to the reference using STAR.
5. produce count tables using HTSeq.
6. call an Rscript file that processes the HTSeq count table using DESeq2.
7. call a python file that condenses and cleans up the results.

The resulting DESeq2\_results\_gene\_name.csv file contains all differently expressed genes using an FDR-adjusted p-value cutoff of 0.05, sorted by FDR-adjusted p-value. The GeneID column indicates the row from the HTSeq count table, and the gene\_name connects that column back to the gene name from the Danio reference gtf file.

## Results

There are 484 genes that are differently expressed with a p-value below 0.05 using this method. When sorted by p-value, the 5 most differently expressed are:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| GeneID | baseMean | log2  FoldChange | lfcSE | stat | pvalue | padj | gene\_name |
| ENSDARG00000091150 | 12524.0926 | 13.6124185 | 1.09241196 | 12.4608838 | 1.22E-35 | 7.09E-33 | **MKI67** |
| ENSDARG00000057484 | 10991.341 | 13.4240798 | 1.12924438 | 11.8876658 | 1.37E-32 | 2.66E-30 | **srsf2** |
| ENSDARG00000074849 | 7542.32393 | 12.8807934 | 1.08302696 | 11.8933266 | 1.28E-32 | 2.66E-30 | **rac1** |
| ENSDARG00000077375 | 5147.88343 | 12.3297637 | 1.08559274 | 11.3576328 | 6.80E-30 | 9.87E-28 | **KIF22** |
| ENSDARG00000023920 | 4542.24251 | 12.149189 | 1.07999046 | 11.2493485 | 2.33E-29 | 2.71E-27 | **llgl2** |

I searched these genes in the Danio genome in the NCBI gene database and summarized the descriptions:

|  |  |
| --- | --- |
| Gene Name | Function (summarized from NCBI gene database) |
| MKI67 | Involved in regulation of chromosome segregation and regulation of mitotic nuclear division. |
| srsf2 | serine/arginine (SR)-rich family of pre-mRNA splicing factors |
| rac1 | RAS superfamily of small GTP-binding proteins. Members of this superfamily appear to regulate a diverse array of cellular events, including the control of cell growth, cytoskeletal reorganization, and the activation of protein kinases |
| KIF22 | Predicted to enable ATP hydrolysis activity; microtubule binding activity; and microtubule motor activity. Acts upstream of or within several processes, including central nervous system development; somitogenesis; and thigmotaxis. |
| llgl2 | Predicted to enable GTPase activator activity and myosin II binding activity. Acts upstream of or within with a positive effect on cell projection assembly. |

MK167 is involved in regulating cell division, which is obviously an important step of the early embryo development.

KIF22 is important in brain development, which is expected to happen between the 2-cell and 6-hour mark.

LLGL2 and RAC1 are involved in the physical growth of the cell itself, LLGL2 about the direction the cell extends (eg. A neuron making a connection to another neuron), and RAC1 about the reorganization of the cytoskeleton that is necessary as the cell changes phases from growth to division.

SRSF2 is involved in differential splicing of mRNA, which I think it the more interesting gene from these 5. It makes sense that there are many splice variants that would be different between the 2-cell and 6-hour time points, but seeing the expression of this gene change doesn’t tell us what mRNA is being spliced differently, just that there must be splice variants between the two time points.