RNA-SEQ ANALYSIS SUMMARY

Analysis for Elisa Randi (PI: C. Szabo)

by Leonore Wigger

December 2020

## SAMPLES

In total, 32 samples were sequenced for this project. The present analysis only included a subset of them.

In this analysis, there are 12 samples from 6 human skin fibroblast lines. They originate from 2 different biobanks. Three cell lines are from children with Down Syndrome (DS) and three from healthy children. Each DS individual is matched to a healthy individual from the same biobank having the same age and gender. For each cell line, there is one sample with treatment (T) and one sample without treatment (0).

Selection of samples: Two of the three matched pairs are siblings: 368 and 369 are sisters (9 and 12 years old), 286 and 285 are male twins (5 years old). The third pair does not consist of siblings but has already been featured in previous publications as a matched pair: Detroit 551 and Detroit 531 (newborns).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SampleID | SampleName | Treatment | Status | Gender | Age | Origin | Matched\_pair |
| 13 | H\_Detroit\_551\_0 | 0 | Healthy | Female | Newborn | ATCC | P2\_Newborn\_F\_AT |
| 14 | H\_Detroit\_551\_T | T | Healthy | Female | Newborn | ATCC | P2\_Newborn\_F\_AT |
| 03 | H\_286\_0 | 0 | Healthy | Male | 5 YR | Le Jeune Institute | P7\_ChildTwin\_M\_JI |
| 04 | H\_286\_T | T | Healthy | Male | 5 YR | Le Jeune Institute | P7\_ChildTwin\_M\_JI |
| 29 | H\_368\_0 | 0 | Healthy | Female | 12 YR | Le Jeune Institute | P8\_ChildSibling\_F\_JI |
| 30 | H\_368\_T | T | Healthy | Female | 12 YR | Le Jeune Institute | P8\_ChildSibling\_F\_JI |
| 19 | D\_Detroit\_539\_0 | 0 | Tris21 | Female | Newborn | ATCC | P2\_Newborn\_F\_AT |
| 20 | D\_Detroit\_539\_T | T | Tris21 | Female | Newborn | ATCC | P2\_Newborn\_F\_AT |
| 09 | D\_285\_0 | 0 | Tris21 | Male | 5 YR | Le Jeune Institute | P7\_ChildTwin\_M\_JI |
| 10 | D\_285\_T | T | Tris21 | Male | 5 YR | Le Jeune Institute | P7\_ChildTwin\_M\_JI |
| 31 | D\_369\_0 | 0 | Tris21 | Female | 9 YR | Le Jeune Institute | P8\_ChildSibling\_F\_JI |
| 32 | D\_369\_T | T | Tris21 | Female | 9 YR | Le Jeune Institute | P8\_ChildSibling\_F\_JI |

## RNA-seq data processing

Data cleaning: Purity-filtered reads were adapters and quality trimmed with Cutadapt (v. 1.8, Martin 2011). Reads matching to ribosomal RNA sequences were removed with fastq\_screen (v. 0.11.1). Remaining reads were further filtered for low complexity with reaper (v. 15-065, Davis et al. 2013).

Alignment to the **genome** using **HTSeq**: Reads were aligned against the Homo\_sapiens.GRCh38.98 genome using STAR (v. 2.5.3a, Dobin et al. 2013). The number of read counts per gene locus was summarized with htseq-count (v. 0.9.1, Anders et al. 2014) using Homo\_sapiens.GRCh38.98 gene annotation. Quality of the RNA-seq data alignment was assessed using RSeQC (v. 2.3.7, Wang et al. 2012). **This data was used for the analysis**.

Alignment to the **transcriptome** using **RSEM**: Reads were also aligned to the Homo\_sapiens.GRCh38.98 transcriptome using STAR (v. 2.5.3a, Dobin et al. 2013) and the estimation of the isoforms abundance was computed using RSEM (v. 1.2.31, Li and Dewey 2011). This data was not used for the analysis.

**Directory:** count\_data **count data file used in analysis**: htseq.genes.counts

### Quality Check after alignment (provided by pipeline)

QC prepared after alignment is good. A global QC report on all samples is available in **globalQC\_report.pdf**. Two samples have a slight bias in RSeQC gene body coverage (samples 11,12 yellow and brown lines) but it is within the range that we consider normal variation.

## STATISTICS

### Normalization and Data Transformation

Statistical analysis was performed for genes independently in R (R version 4.0.3). Genes with low counts were filtered out according to the rule of 1 count(s) per million (cpm) in at least 1 sample. Library sizes were scaled using TMM normalization. Subsequently, the normalized counts were transformed to cpm values and a log2 transformation was applied by means of the function cpm with the parameter setting prior.counts = 1 (EdgeR v 3.30.3; Robinson et al. 2010).

**Number of genes measured in Hyp, by biotype category (HTSEQ)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | total\_gtf | sum\_greater\_0 | in\_analysis |
| IG\_or\_TR\_gene | 408 | 19 | 2 |
| IG\_or\_TR\_pseudogene | 237 | 4 | 0 |
| IG\_pseudogene | 1 | 0 | 0 |
| lncRNA | 16837 | 7321 | 863 |
| miRNA | 1879 | 109 | 2 |
| misc\_RNA | 2211 | 116 | 1 |
| Mt\_rRNA | 2 | 2 | 1 |
| Mt\_tRNA | 22 | 12 | 4 |
| polymorphic\_pseudogene | 42 | 11 | 1 |
| processed\_pseudogene | 10164 | 1889 | 105 |
| protein\_coding | 19947 | 16239 | 12352 |
| pseudogene | 18 | 3 | 0 |
| ribozyme | 8 | 0 | 0 |
| rRNA | 52 | 0 | 0 |
| rRNA\_pseudogene | 499 | 4 | 0 |
| scaRNA | 49 | 7 | 0 |
| scRNA | 1 | 1 | 0 |
| snoRNA | 942 | 97 | 0 |
| snRNA | 1901 | 75 | 1 |
| sRNA | 5 | 0 | 0 |
| TEC | 1061 | 375 | 32 |
| transcribed\_processed\_pseudogene | 495 | 203 | 37 |
| transcribed\_unitary\_pseudogene | 130 | 74 | 13 |
| transcribed\_unprocessed\_pseudogene | 921 | 412 | 110 |
| translated\_processed\_pseudogene | 2 | 1 | 0 |
| translated\_unprocessed\_pseudogene | 2 | 2 | 0 |
| unitary\_pseudogene | 98 | 16 | 1 |
| unprocessed\_pseudogene | 2629 | 299 | 21 |
| vaultRNA | 1 | 0 | 0 |
| TOTALS | 60564 | 27291 | 13546 |

**Column legend**:

* *total\_gtf*: All genes to which the read counts were mapped.
* *sum\_greater\_0*: Genes that have at least 1 read count in at least 1 sample.
* *in\_analysis*: Genes that have at least 1 count per million (cpm) in at least 1 sample.

**Data directory (normalized counts):** normdata\_v2\_final\_3pairs

### Quality control: Clustering and PCA

After data normalization, a quality control analysis was performed through hierarchical clustering and sample PCA (plots of first two principal components). We colored the plots in different ways to see how samples cluster and which variables might have influence on gene expression. Each pdf file contains four clustering and PCA plots prepared with decreasing numbers of genes: (all available genes, 5000 most variable, 1000 most variable, 500 most variable).

**Directory:** QC

#### Observations in clustering and PCA

We see the same patterns in this subset of data that we see when all cell lines are included

**colored\_by\_condition.SampleCluster\_PCA.pdf**  
Treatment: The untreated and treated sample from each cell line cluster tightly together. This indicates that there is very little treatment effect. The two samples from the same cell line behave like replicates of one another.

Down Syndrome: There isn’t a clear separation of DS and healthy samples. This indicates that there are other effects that are stronger than DS and that need to be adjusted for in order to see the DS effect.

**colored\_by\_biobank.SampleCluster\_PCA.pdf**  
Samples are from two biobanks, but there are not enough samples included in this analysis to clearly see biobank effect. The two samples from cell line Detroit\_551 cluster apart from the rest of the samples, including far away from its matched cell line, Detroit\_539.

**colored\_by\_gender.SampleCluster\_PCA.pdf**  
While samples are not completely separated by gender, there is a demarcation line, indicating a gender effect on gene expression.

**colored\_by\_matched\_pair.SampleCluster\_PCA.pdf**  
Two of the matched pairs cluster together (285 and 285; 368 and 369). However, Detroit\_551 and Detroit\_539 are far apart despite being a matched pair.

### Differential expression

Differential expression was computed with the R Bioconductor package **limma** (Ritchie et al. 2015) by fitting data to a linear model. The approach limma-trend was used. Fold changes were computed and a moderated t-test was applied. P-values were adjusted using the Benjamini-Hochberg (BH) method, which controls for the false discovery rate (FDR). Some tests were run with the full data set and others only with a subset of the data, e.g., with healthy or DS samples, as indicated below.

**Directory:** stat\_tests\_\_v2\_final\_3pairs

#### Comparing DS vs Healthy

**Final approach: Use all 12 samples together**

Two versions, data supplied for both:

**A) allsamples\_Tris21-Healthy\_corrTreatmentMatchedpair.xlsx**

Linear model Compare DS vs healthy, adjust for “Treatment” and for “Matched pair”.

This results in **789** significant genes after multiple testing correction by the Benjamini Hochberg method. **Caveat**: This may underestimate variability and therefore, p-values, i.e. give us too many small p-values. In a regular linear model we cannot account for the fact that the treated and untreated sample of each cell line are highly correlated.

Number of differentially expressed genes:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ALL | UP | DOWN | ALL, FC>2 | UP, FC>2 | DOWN, FC>2 |
| allsamples\_Tris21-Healthy\_corrTreatmentMatchedpair | 789 | 457 | 332 | 335 | 200 | 135 |

**B) allsamplesDupcor\_Tris21-Healthy\_corrTreatmentMatchedpair.xlsx**

Linear model with a factor estimating random effect for cell line (limma function dupcor)  
Compare DS vs healthy, adjust for “Treatment” and for “Matched pair” and add a variable for the random effect.

This results in **77** significant genes after multiple testing correction by the Benjamini Hochberg method.  
This model accounts for the fact that the two samples from the same cell line (treated and untreated) are more correlated to one another than to other samples.

Number of differentially expressed genes:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ALL | UP | DOWN | ALL, FC>2 | UP, FC>2 | DOWN, FC>2 |
| allsamplesDupcor\_Tris21-Healthy\_corrTreatmentMatchedpair | 77 | 48 | 29 | 66 | 40 | 26 |

GTF Recommendation:  
- Use the longer list from A) only if the significant genes that are further away from the top of the list include known ones and seem to make biological sense.  
- Use the smaller list from B) otherwise, for smaller risk of false positives.

#### Comparing Treated vs Untreated

Linear model, paired samples (treated-untreated from each cell line).

This analysis was performed on all samples together, on DS samples only, and on healthy samples only. **No genes were significantly differentially expressed** between treated and untreated conditions in any of these analyses.

TreatmentT-Treatment0\_in\_allsamples.xlsx  
TreatmentT-Treatment0\_in\_DS.xlsx  
TreatmentT-Treatment0\_in\_Healthy.xlsx

### Previously studied gene: CBS

Expectation: In previous studies, this genes had higher expression levels in DS than in healthy. Specifically this was observed when comparing the cell line pair Detroit 539 and Detroit 551

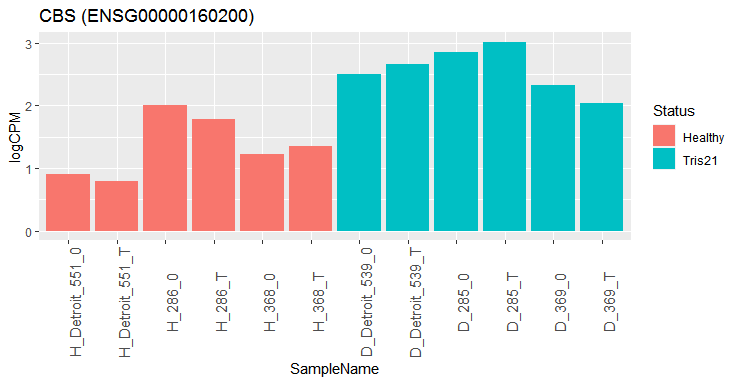
**Visualization in a barplot (Barplots\_CBS.pdf)**:

* In the three matched pairs (six cell lines) included in this analysis, this pattern holds: CBS is lower higher in DS than in healthy.

**Statistical results**:

Overall, the gene is **very lowly expressed** (average log CPM value: 1.95) and **quite variable** (see barplot). It has a **good fold change** and it comes up as significant in the simpler model, but does not come up as significant in the more complex model (which aims to avoid understating p-values due to high correlation of the sample pairs within each cell line).

* Fold change: 2.34
* adjusted p-value in analysis A): 0.0033
* adjusted p-value in analysis B): 0.11



### RESULT TABLES COLUMNS

The result files contain one row per gene. Columns are defined as follows:

|  |  |
| --- | --- |
| Column | Description |
| logFC | Log2 fold change |
| Fold\_Change or FC\_\* | Antilog of the logFC |
| mean\_all or AveExp | Log2 average expression or CPM across all samples in the model |
| mean\_[\*] | Log2 average expression or CPM of each condition tested |
| t | t statistic (from moderated t-test) |
| F | F statistic (from moderated F-test) |
| P.Value | P-value of the test applied |
| adj.P.Val or FDR | Adjusted p-value computed by the Benjamini-Hochberg method, controlling for false discovery rate (FDR) |
| Next few columns | Gene annotation information |
| Last few columns | Expression level per sample: Normalized counts per million reads (CPM), log2-transformed |

## REFERENCES

* Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17(1), 10-12. doi: dx.doi.org/10.14806/ej.17.1.200.
* Babraham Bioinformatics. <http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/>
* Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N & Enright AJ. (2013). Kraken: A set of tools for quality control and analysis of high-throughput sequence data. Methods 63(1), 41-49. doi: 10.1016/j.ymeth.2013.06.027. PMID: 23816787
* Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15-21. doi: 10.1093/bioinformatics/bts635. PMID: 23104886.
* Anders S, Pyl PT, Huber W. (2014). HTSeq a Python framework to work with high-throughput sequencing data. Bioinformatics. Advance online publication. <doi:10.1093/bioinformatics/btu638>. PMID: 25260700.
* Wang L, Wang S, Li W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics 28(16):2184-5. doi: 10.1093/bioinformatics/bts356. PMID: 22743226.
* Li B, Dewey CN. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323. doi: 10.1186/1471-2105-12-323. PMID: 21816040.
* Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43(7):e47. PMID: 25605792
* Robinson, MD, McCarthy, DJ, Smyth, GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.

### R session info

## R version 4.0.3 (2020-10-10)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 8.1 x64 (build 9600)  
##   
## Matrix products: default  
##   
## locale:  
## [1] LC\_COLLATE=English\_United States.1252   
## [2] LC\_CTYPE=English\_United States.1252   
## [3] LC\_MONETARY=English\_United States.1252  
## [4] LC\_NUMERIC=C   
## [5] LC\_TIME=English\_United States.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] ggpubr\_0.4.0 ggplot2\_3.3.2 readxl\_1.3.1 RColorBrewer\_1.1-2  
## [5] maptools\_1.0-2 sp\_1.4-4 edgeR\_3.30.3 MASS\_7.3-53   
## [9] limma\_3.44.3 knitr\_1.30   
##   
## loaded via a namespace (and not attached):  
## [1] statmod\_1.4.35 locfit\_1.5-9.4 tidyselect\_1.1.0 xfun\_0.15   
## [5] purrr\_0.3.4 splines\_4.0.3 haven\_2.3.1 lattice\_0.20-41   
## [9] carData\_3.0-4 colorspace\_1.4-1 vctrs\_0.3.2 generics\_0.1.0   
## [13] htmltools\_0.5.0 yaml\_2.2.1 rlang\_0.4.7 pillar\_1.4.6   
## [17] foreign\_0.8-80 glue\_1.4.1 withr\_2.3.0 lifecycle\_0.2.0   
## [21] stringr\_1.4.0 munsell\_0.5.0 ggsignif\_0.6.0 gtable\_0.3.0   
## [25] cellranger\_1.1.0 zip\_2.1.1 evaluate\_0.14 labeling\_0.4.2   
## [29] rio\_0.5.16 forcats\_0.5.0 curl\_4.3 highr\_0.8   
## [33] broom\_0.7.2 Rcpp\_1.0.5 scales\_1.1.1 backports\_1.1.10   
## [37] abind\_1.4-5 farver\_2.0.3 hms\_0.5.3 digest\_0.6.25   
## [41] stringi\_1.4.6 openxlsx\_4.2.2 rstatix\_0.6.0 dplyr\_1.0.0   
## [45] cowplot\_1.1.0 grid\_4.0.3 tools\_4.0.3 magrittr\_1.5   
## [49] tibble\_3.0.3 crayon\_1.3.4 tidyr\_1.1.2 car\_3.0-10   
## [53] pkgconfig\_2.0.3 ellipsis\_0.3.1 data.table\_1.13.0 rmarkdown\_2.5   
## [57] R6\_2.5.0 compiler\_4.0.3