MGI DNBSeq-G400 vs Illumina Novaseq RNASeq recall test

Stéphane Plaisance [VIB - Nucleomics Core, nucleomics@vib.be]

March 10th, 2020 - version 2.0

Contents

last edits: Wed Mar 11, 2020

Introduction

Evaluate two RNA library samples and their in-vitro mix after sequencing on the MGI DNBSeqG400 and Illumina Novaseq devices and RNASeq mapping to the human genome hg38 build using STAR *(v2.7.3a)*

REM: In this work, we do not assess differential expression of genes and only want to see if in-vitro count mixtures from the two RNA samples leads to a similar assessment of gene expression on both platforms.

We chose to use TPM counts rather than RPKM or RNASeq normalized counts used in the regular Nucleomics Core analysis pipeline.

TPM are equalized for all samples to report a sum of all counts of 1 within each sample and are therefore ideal to produce in-silico mixtures as done below.

For more information, please view the following video: **RPKM, FPKM and TPM, clearly explained** (Starmer, n.d.)

Method

Libraries and sequencing were performed by Kizi Coeck (Nucleomics Core) during the MGI training week (Feb-2020) and later on the Novaseq are not discussed here.

The DNBSEQG400 Lane2 (L02) was loaded with a mixture of 16 final barcoded libraries in three groups (4bc, 4bc, 8bc sets).

The Illumina Novaseq lane2 (L02) was loaded with a mixture of 4 single read libraries prepared from the same RNA. Since this was sequence as SE75 on Novaseq, we also create pseudo samples from the MGI data buy trimming the first 75 bases of read_1 files. The additional fastq fiels were processed as for the original files to evaluate the effect of (size + pair-end sequencing) on the final gene counts.

MGI DNBSEQG400 samples

barcode	RNA sample	comment
1	HumanRef	100% HumanRef
2	HumanRef	100% HumanRef
3	HumanRef	100% HumanRef
4	HumanRef	100% HumanRef
13	HumanBrain	100% HumanBrain
14	HumanBrain	100% HumanBrain
15	HumanBrain	100% HumanBrain
16	HumanBrain	100% HumanBrain
97	R75B25	75% HumanRef + 25% HumanBrain
98	R75B25	75% HumanRef + 25% HumanBrain
99	R75B25	75% HumanRef + 25% HumanBrain
100	R75B25	75% HumanRef + 25% HumanBrain
101	R25B75	25% HumanRef + 75% HumanBrain
102	R25B75	25% HumanRef + 75% HumanBrain
103	R25B75	25% HumanRef + 75% HumanBrain
104	R25B75	25% HumanRef + 75% HumanBrain

Illumina Novaseq samples

sample	RNA sample	comment
1	100UnivHumanRef_S1	100% HumanRef
2	100HuBrain_S2	100% HumanBrain
3	75UnivHuRef_25HuBrain_S3	75% HumanRef + 25% HumanBrain
4	25UnivHuRef_75HuBrain_S3	25% HumanRef + 75% HumanBrain

STAR analysis

The classical STAR alignment method was applied based on info from the Nucleomics Core pipeline

```
    --outFilterMismatchNmax 10 \
    --outFilterMismatchNoverLmax 0.3 \
    --alignSJDBoverhangMin 3 \
    --alignSJoverhangMin 5 \
    --alignIntronMin 21 \
    --alignIntronMax 500000 \
    --outFilterMultimapNmax 10 \
    --outSJfilterOverhangMin 12 30 30 30 \
    --outWigType None \
    --outSAMprimaryFlag OneBestScore
```

And some additions from the STAR tutorial pages.

- sorted BAM output for PASS-2
- STAR counts per gene for PASS-2

In short, the reads were first mapped against the human reference genome hg38 to identify splice sites and the collection of obtained splice events were stored.

A second STAR alignment pass was then performed that takes into account the merge splice event database from pass-1 to optimize sequence alignments. Besides the usual NC settings, a few additional settings were added to directly produce gene counts (not using third party software as this is done for NC RNASeq analysis).

As extra pseudo-samples, we trimmed MGI reads_1 data to 75b to create Novaseq-like single-end data. This left about 18 million reads per sample.

R analysis of STAR counts

The raw STAR counts from the 16 PE100 + 16 SE75 +4 SE75 = 36 concurrent STAR analyses were merged to a single large table (60617 rows x 36 columns)

Merging of the STAR raw count files

We converted the STAR raw counts to “transcripts per million” (TPM) for each of the 36 samples, then compared counts between platforms.

The full table of results and filtered version were used to plot PCA and hierarchical clustering figures shown below.

Data visualization

In this section we compare 100%-Ref and 100%-Brain counts between platforms and with each other within platform. The other two samples are not included here.

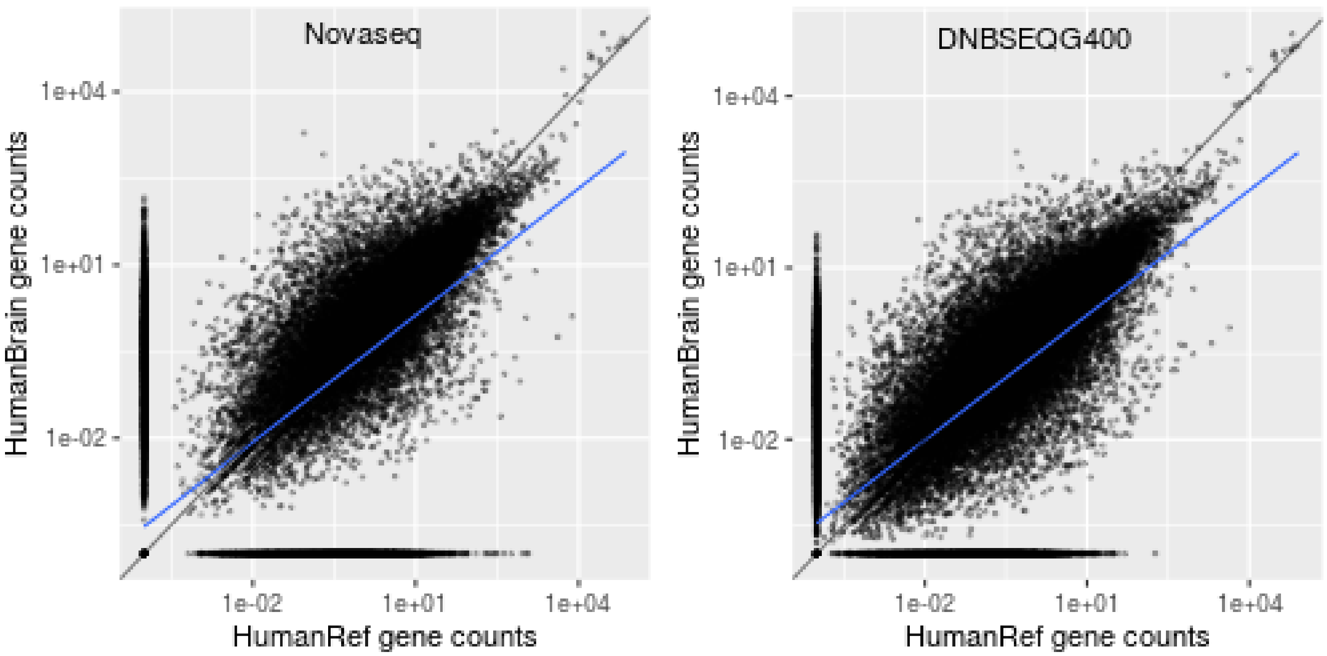
Differential expression within datasets

The Human brain vs HumanRef scatter plots from scalled expression counts (TPM) of all genes in both Illumina Novaseq and MGI are shown below.

The black line shows the theoretical diagonal while the blur line represents the data average (loess).

The dots on the left an dbottom are not expressed in one sample.

The data was not normalized as for a regular RNASeq analysis and should therefore be taken as-is.

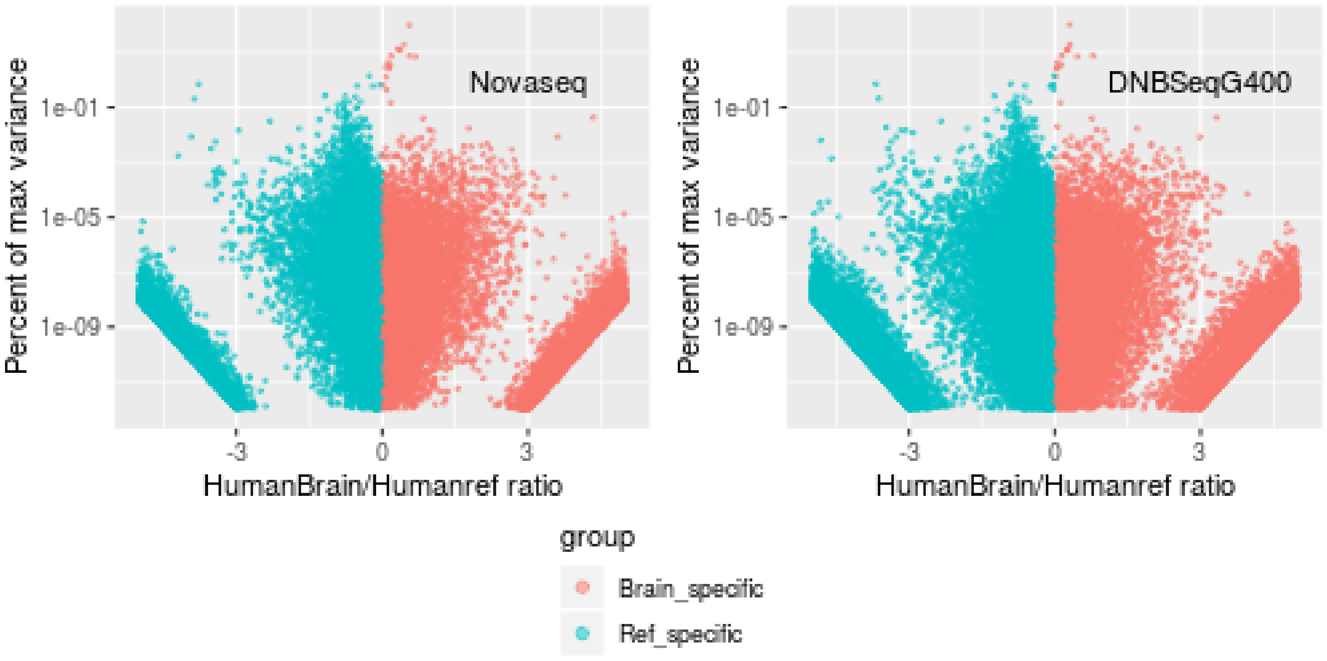


Both plots are globally very similar and probably show the same outliers.

Variance plot as a function of the expression ratio within datasets

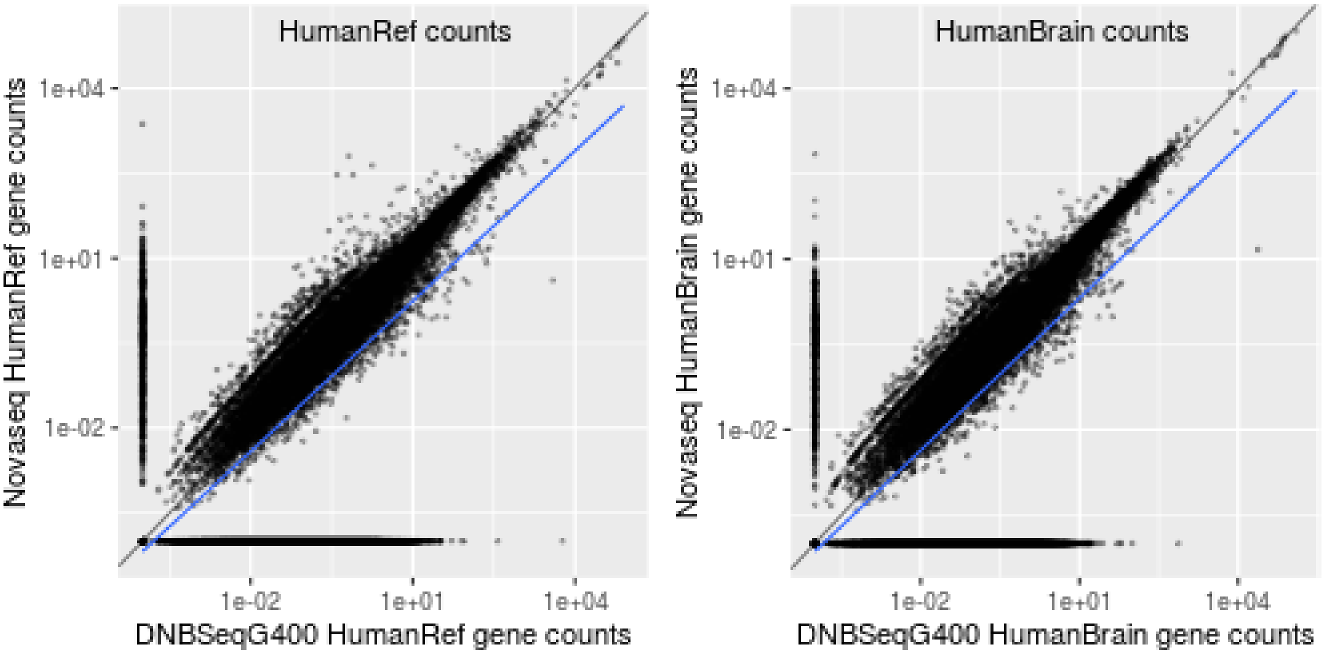
The following plots reports for each gene, the fraction of the max-variance across all genes (y-axis) as a function of the ratio of [the mean gene count in the HumanBrain] / [the mean count in the HumanRef] (x-axis; both axis are log-scaled).

The bar-shaped group of dots on the left side are HumanRef-specific genes while the bar-shaped group on the right side are HumanBrain-specific genes (these weird shapes come from the log-transformation where one of the ratio value is very small). Genes shown in the center part of the plot are expressed in both samples but show more expression in the HumanRef (left half) or in the HumanBrain (rigt half).



Between dataset plots

The next plots report the gene counts within the same RNA sample between datasets.



The global correlation coefficient of **0.97** (HumanRefB) and **0.98** (HumanBrain) between both platforms is quite good, indicating that both platforms have a similar image of the RNA composition.

TheThese two plots also show that a number of genes are not (or very few) expressed with one platform (left and bottom lines). This may indicates a difference in sensitivity betwen Novaseq and DNBSeqG400 possibly due to the different reads produced for both platforms (single 75b on Novaseq and paired 100b on MGI).

PCA analysis

The TPM counts were further filtered to remove low-variance rows (variance<0.1 - arbitrary choice) where all samples have similar expression values as these rows do not (or much less) contribute to the PCA analysis or clustering in the first place.

PCA plots for the actual sequencing samples

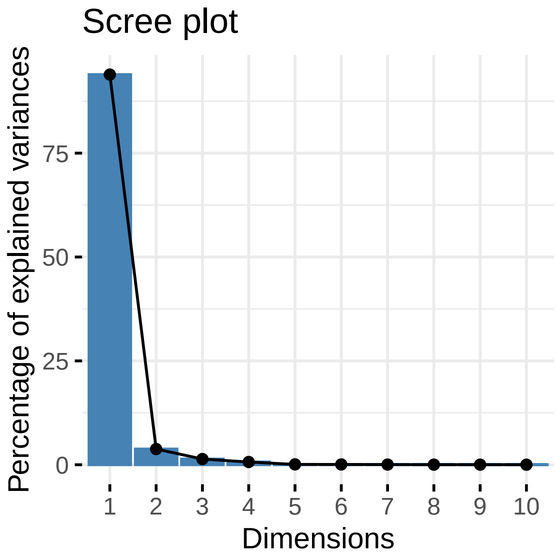
The variance filtering retained 24248 rows out of 60617 EnsEMBL GeneID rows.

The filtered TPM data is used to compute PCA as seen next.

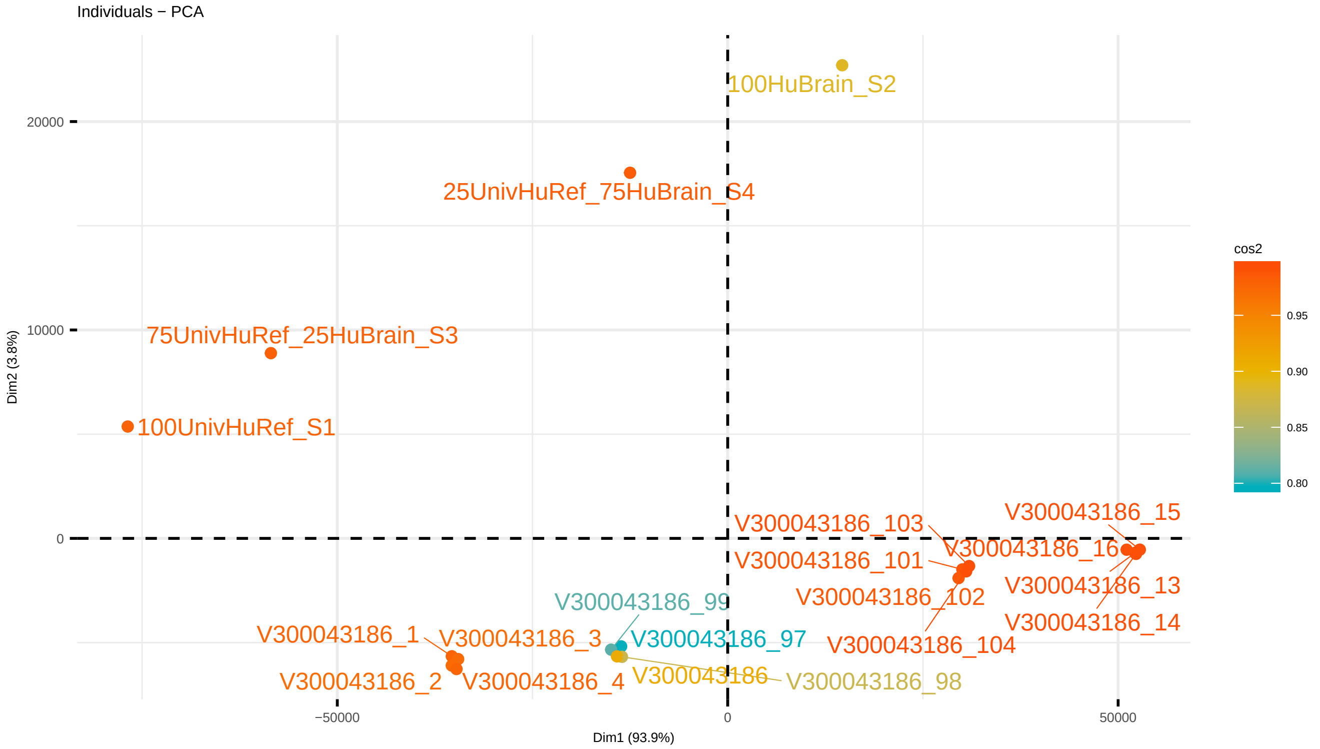
In the first PCA, filtered STAR counts are used to show the variance between all 16+4 samples.

```
Importance of components:
                             PC1       PC2       PC3       PC4       PC5
Standard deviation     3.967e+04 7.962e+03 4.790e+03 3327.0897 1.169e+03
Proportion of Variance 9.391e-01 3.782e-02 1.369e-02    0.0066 8.200e-04
Cumulative Proportion  9.391e-01 9.769e-01 9.906e-01    0.9972 9.980e-01
                             PC6       PC7       PC8       PC9      PC10
Standard deviation     1.076e+03 845.38577 620.63044 567.06816 503.30032
Proportion of Variance 6.900e-04   0.00043   0.00023   0.00019   0.00015
Cumulative Proportion  9.987e-01   0.99916   0.99939   0.99959   0.99974
                           PC11      PC12      PC13      PC14      PC15
Standard deviation     400.2010 329.98204 257.82768 208.24942 160.06742
```

```
Proportion of Variance   0.0001   0.00006   0.00004   0.00003   0.00002
Cumulative Proportion    0.9998   0.99990   0.99994   0.99996   0.99998
                            PC16       PC17  PC18  PC19       PC20
Standard deviation     126.13880 103.89852 78.01 68.38 2.343e-11
Proportion of Variance   0.00001   0.00001  0.00  0.00 0.000e+00
Cumulative Proportion    0.99999   0.99999  1.00  1.00 1.000e+00
```



About 94% of the variance is explained by the first principal component.



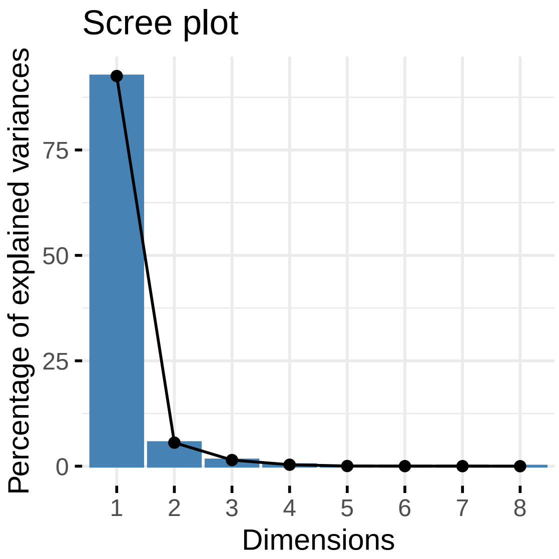
The Illumina samples seem to have less spread than MGI samples but appear in the same order from left to right.

PCA plots for the group means of the MGI samples and Novaseq samples

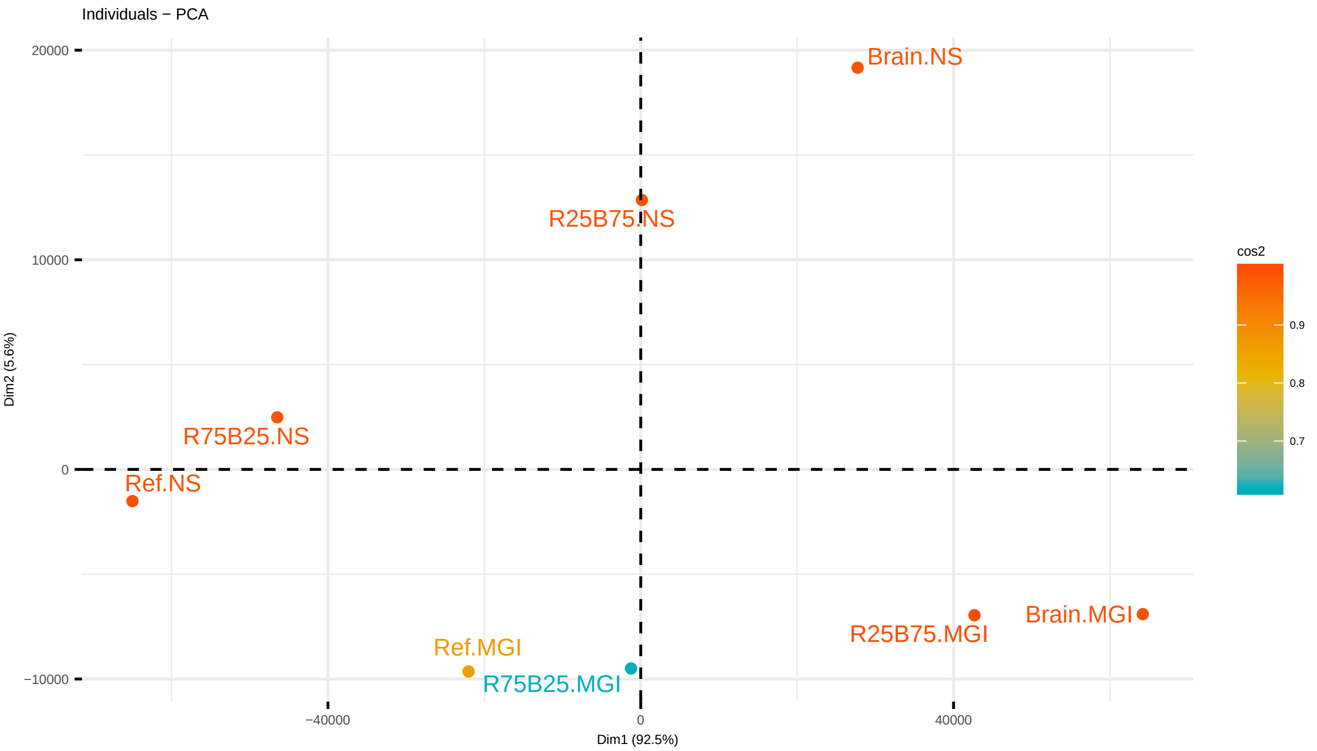
This plot shows the distribution of the 4 MGI group count means together with the 4 corresponding Illumina Novaseq count sets.

The variance filtering retained 23244 rows out of 60617 EnsEMBL GeneID rows.

```
Importance of components:
                             PC1       PC2       PC3       PC4       PC5
Standard deviation     4.405e+04 1.082e+04 5.536e+03 2.700e+03 959.76679
Proportion of Variance 9.253e-01 5.587e-02 1.461e-02 3.480e-03   0.00044
Cumulative Proportion  9.253e-01 9.812e-01 9.958e-01 9.993e-01   0.99972
                             PC6       PC7       PC8
Standard deviation     578.31539 505.75336 9.718e-12
Proportion of Variance   0.00016   0.00012 0.000e+00
Cumulative Proportion    0.99988   1.00000 1.000e+00
```



More than 92% of the variance is explained by the first principal component.



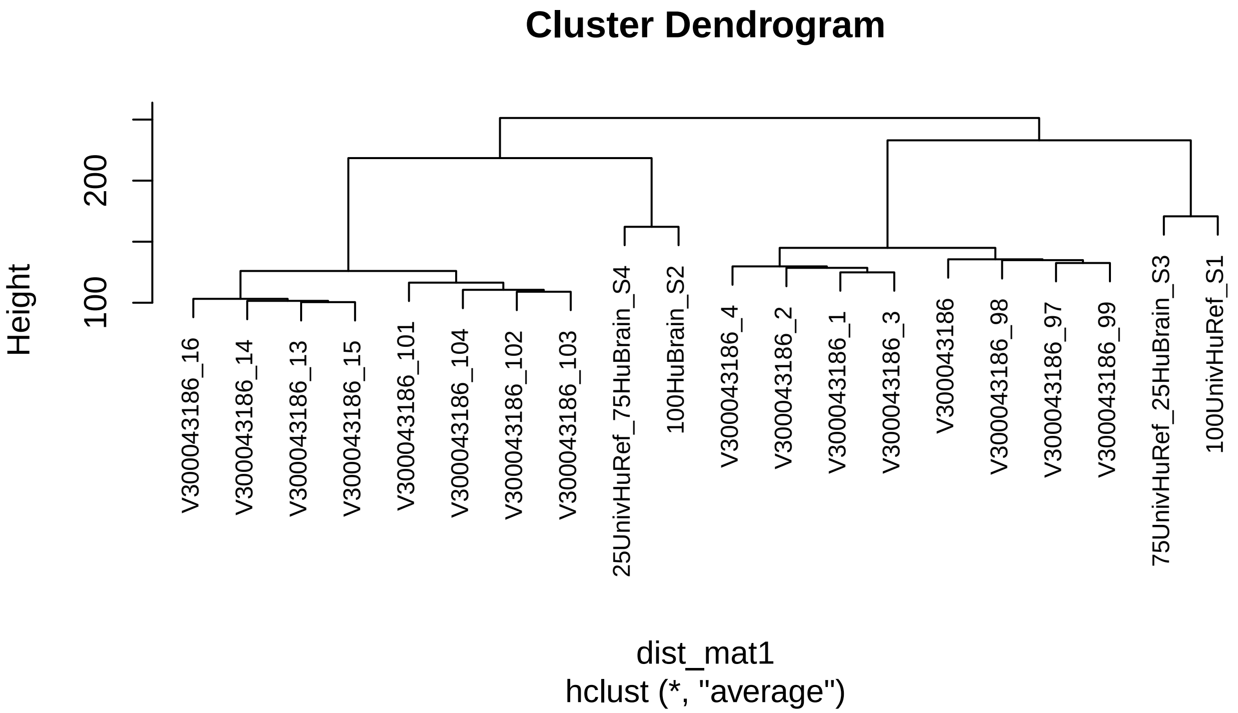
Again, the Novaseq samples have less spread but appear in the same order as the MGI samples.

Hierarchical Clustering analysis

The same data can also be used to plot hierarchical clustering results for the 16+4 sequenced samples and build a tree.

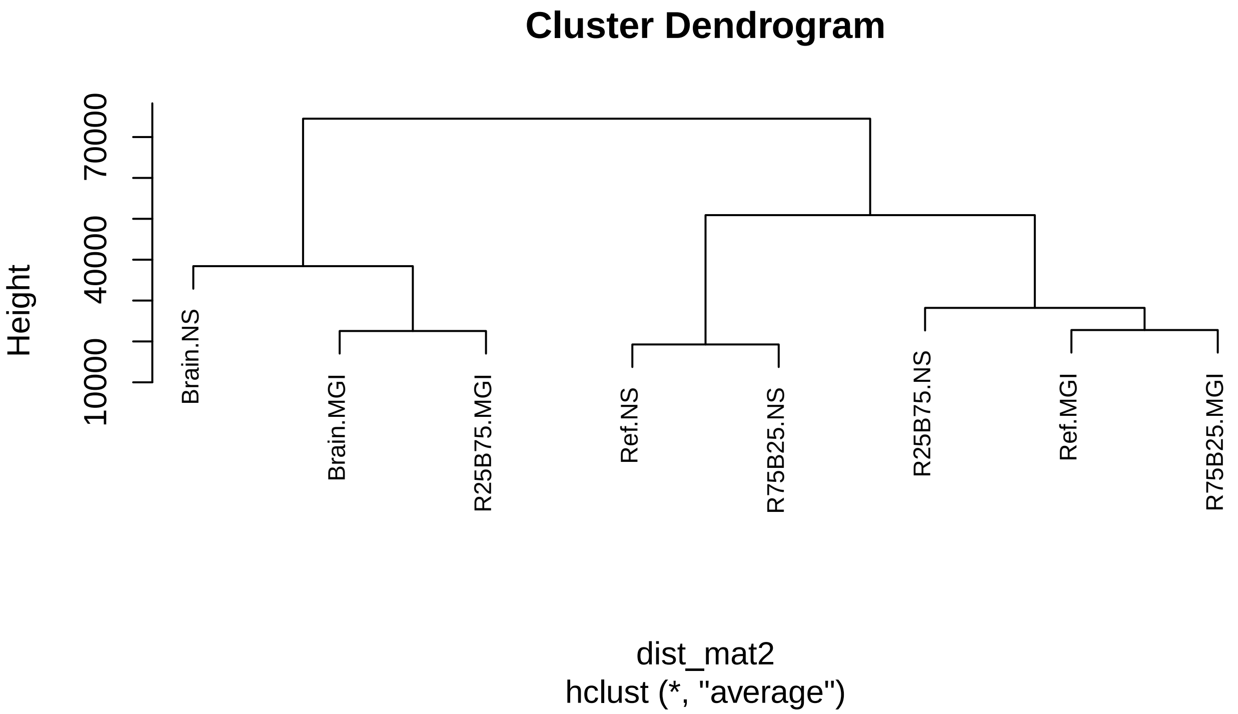
The four groups appear in the order:

- 100% HumanBrain *MGI [bc13..bc16]*
- 75% HUmanBrain 25% HUmanRef *[MGI bc101..bc104]*
- 100% HumanRef *[MGI bc1..bc4]*
- 75% HUmanRef 25% HUmanBrain *[MGI bc97..bc100]*



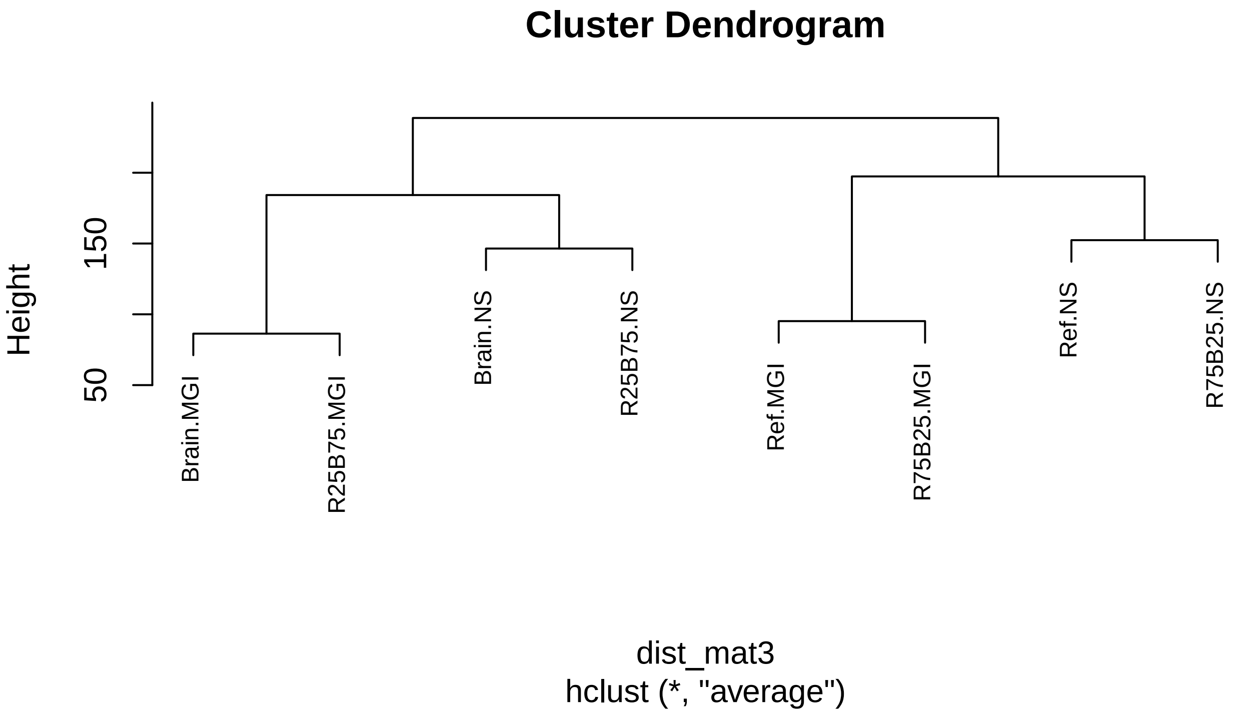
The four MGI samples of each groups nicely cluster together. the mixed-samples with 75% Human Ref are close to the HumanRef samples (idem for the two Brain groups).

the result of comparing the four sample means to the Novaseq samples is shown next



Pure and 75% samples cluster together. MGI and Novaseq are close to each other but distinct.

Below are similar plot after scaling and centering of the data (in case experts would prefer this approach).



The scaling of the data does not change dramatically the aspect of the tree and the same samples are found at the same relative positions.

Conclusion

We show here that sequencing data obtained by mixing two defined samples reflects the abundance of each sample. This expected results comforts us the fact that both platform are similarly robust to detect library complexities of similar nature.

The Illumina Novaseq data is similar but not identical to the MGI data. This could originate from the different sequencing libraries (paired vs single reads of different lengths) or reveals a true difference in the sensitivity of both platforms.

last edits: Wed Mar 11, 2020

VIB NUCLEOMICS CORE more at **http://www.nucleomics.be**

References

Starmer, Josh. n.d. *RPKM, Fpkm and Tpm, Clearly Explained*. https://statquest.org/2015/07/09/rpkm-fpkm-and-tpm-clearly-explained/.