MGI DNBSeq-G400 vs Illumina Novaseq analysis of RNASeq recall and effet of read length

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 trimmed 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 PE100 reads

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

MGI DNBSEQG400 SE75 in-vitro trimmed reads

SE-samples	RNA sample	comment
Ref_SE75_1	HumanRef	100% HumanRef
Ref_SE75_2	HumanRef	100% HumanRef
Ref_SE75_3	HumanRef	100% HumanRef
Ref_SE75_4	HumanRef	100% HumanRef
Brain_SE75_13	HumanBrain	100% HumanBrain
Brain_SE75_14	HumanBrain	100% HumanBrain
Brain_SE75_15	HumanBrain	100% HumanBrain
Brain_SE75_16	HumanBrain	100% HumanBrain

SE-samples	RNA sample	comment
R75B25_SE75_97	R75B25	75% HumanRef + 25% HumanBrain
R75B25_SE75_98	R75B25	75% HumanRef + 25% HumanBrain
R75B25_SE75_99	R75B25	75% HumanRef + 25% HumanBrain
R75B25_SE75_100	R75B25	75% HumanRef + 25% HumanBrain
R25B75_SE75_101	R25B75	25% HumanRef + 75% HumanBrain
R25B75_SE75_102	R25B75	25% HumanRef + 75% HumanBrain
R25B75_SE75_103	R25B75	25% HumanRef + 75% HumanBrain
R25B75_SE75_104	R25B75	25% HumanRef + 75% HumanBrain

Illumina Novaseq SE75 reads

NS-sample	RNA sample	comment
NS_SE75_1	100UnivHumanRef_S1	100% HumanRef
NS_SE75_2	100HuBrain_S2	100% HumanBrain
NS_SE75_3	75UnivHuRef_25HuBrain_S3	75% HumanRef + 25% HumanBrain
NS_SE75_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 in-silico samples, we trimmed MGI reads_1 data to 75b to create Novaseq-like single-end data.

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.

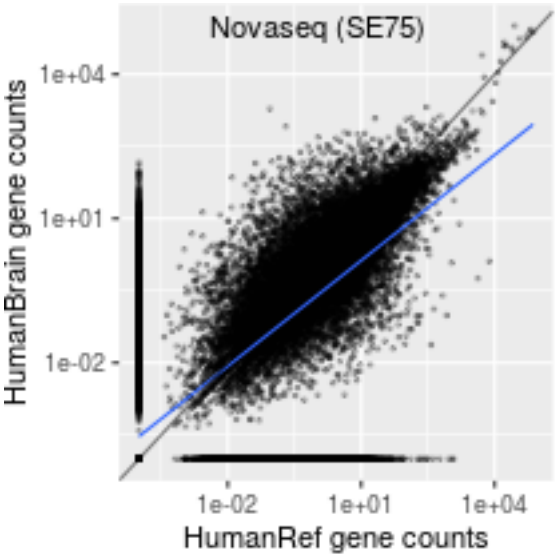
Differential expression within datasets

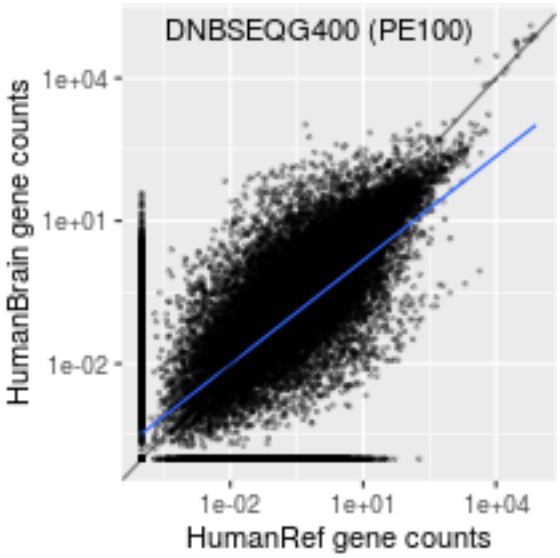
The Human brain vs HumanRef scatter plots from scalled expression counts (TPM) of all genes in both Illumina Novaseq and MGI are shown below.

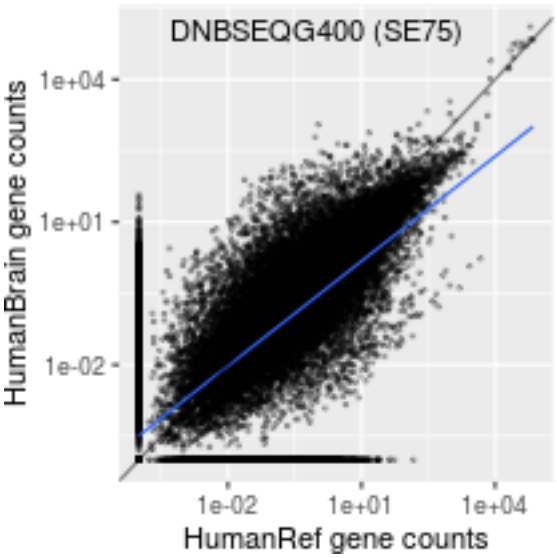
The black line shows the theoretical diagonal while the blur line represents the data average (loess).

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.





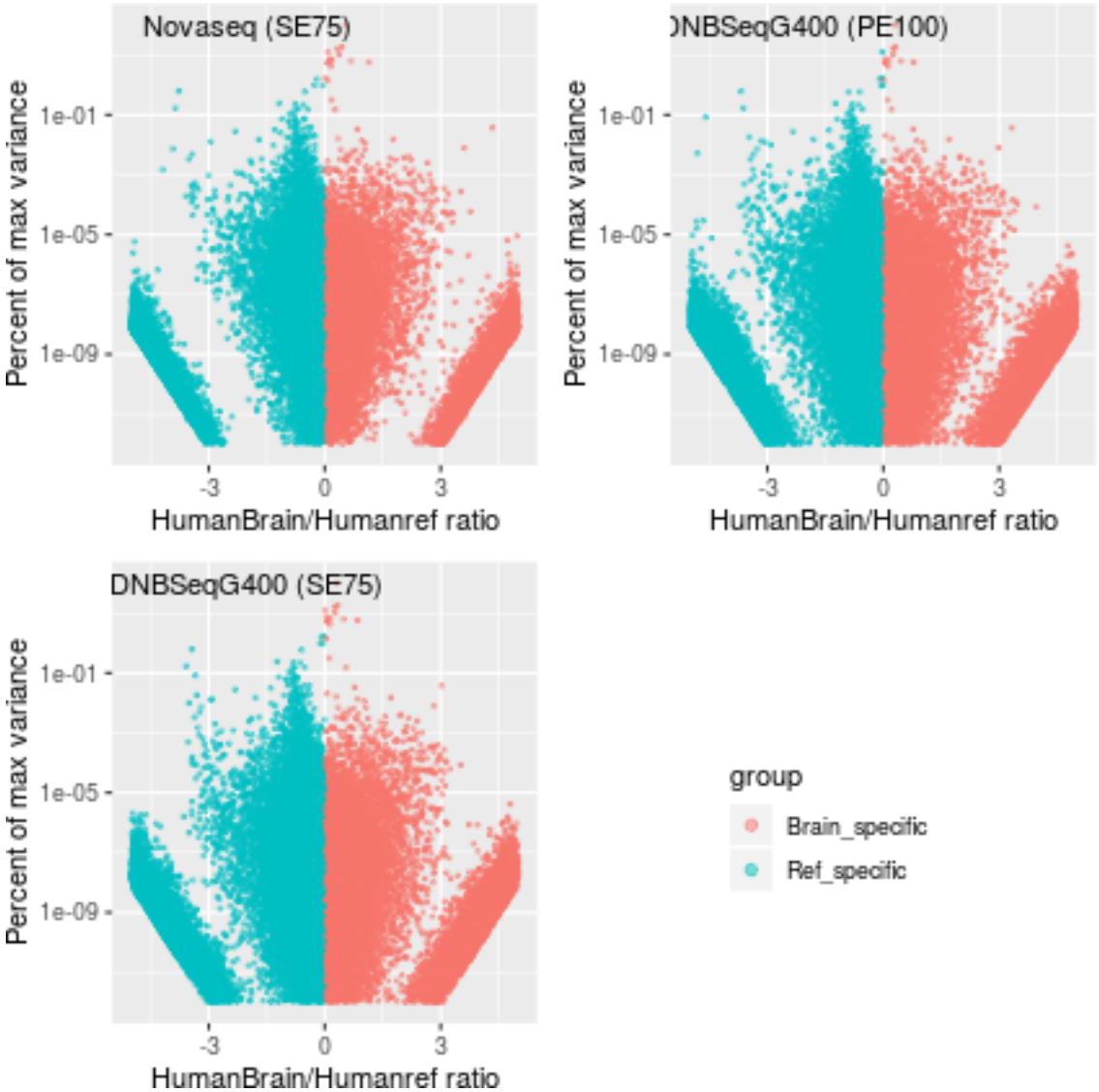


Both plots are globally very similar and probably show the same outliers. The read length does not seem to affect the data distribution as seen by the third plot nbased on pseudo-reads SE75 derived from the MGI data.

Variance plot as a function of the expression ratio within datasets

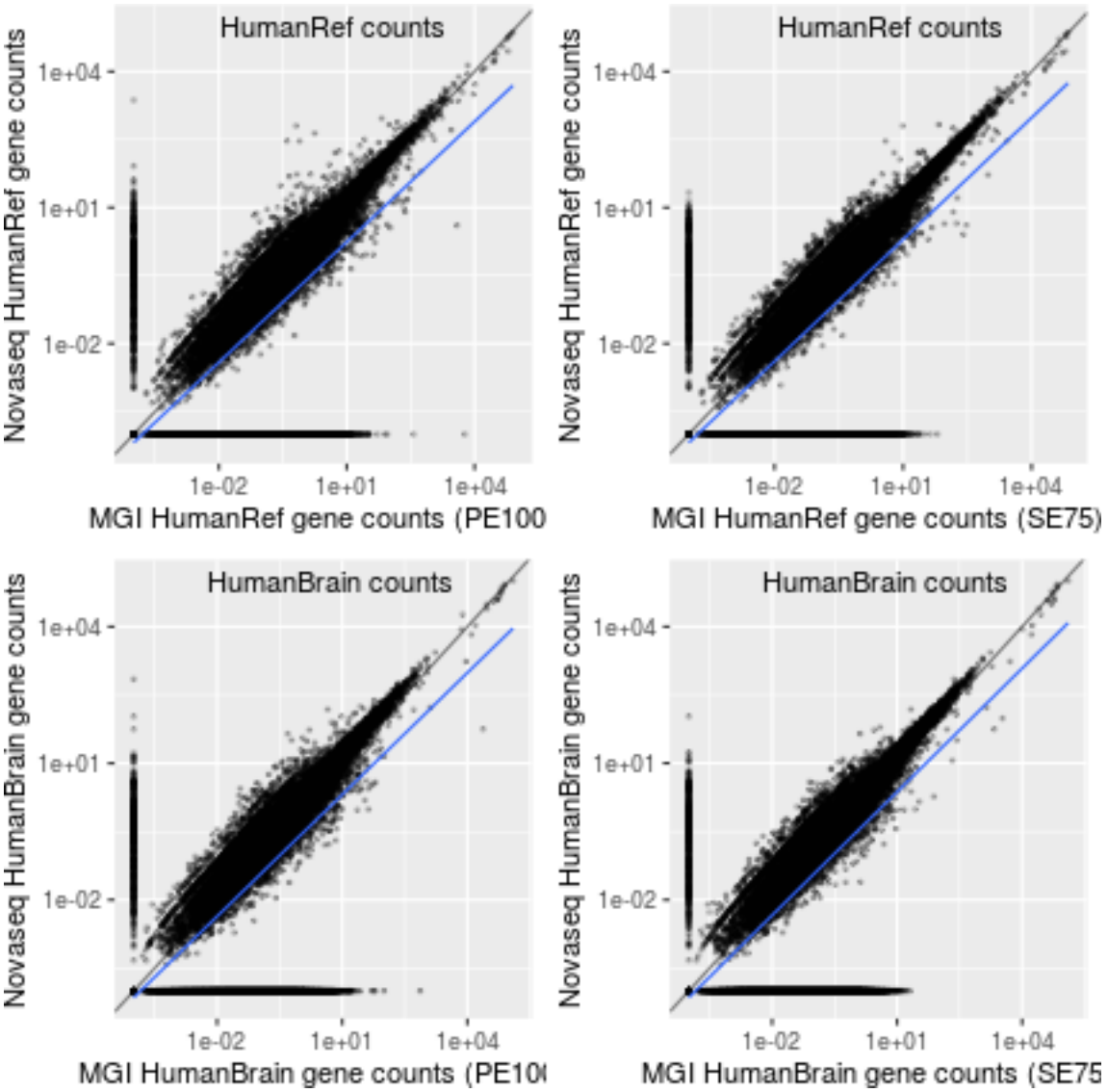
The following plots reports for each gene, the fraction of the max-variance across all genes (y-axis) as a function of the ratio of [the mean gene count in the HumanBrain] / [the mean count in the HumanRef] (x-axis; both axis are log-scaled).

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.



Comparing MGI-PE100 to NS-SE75 gives correlation coefficients of **0.97** (HumanRefB) and **0.98** (HumanBrain), indicating that both platforms have a similar image of the RNA composition.

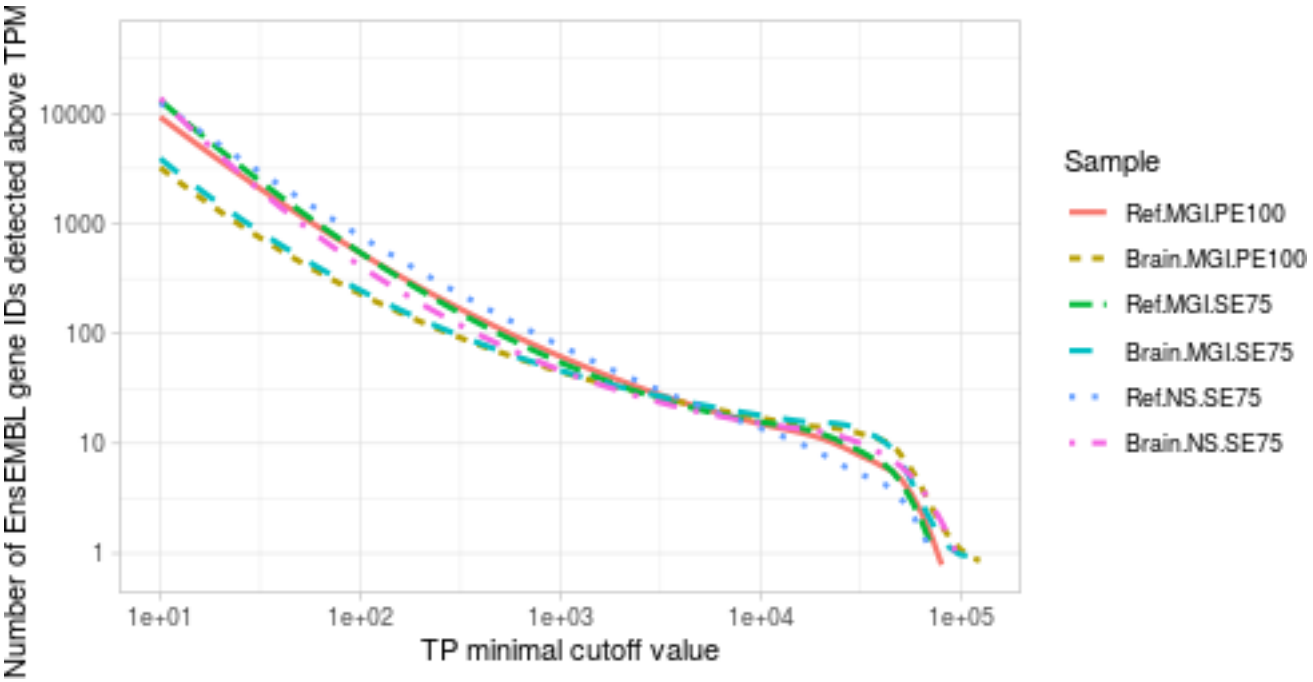
Comparing MGI-SE75 (trimmed-reads) to NS-SE75 gives correlation coefficients of **0.95** (HumanRefB) and **0.98** (HumanBrain), similar to above and indicating that the read nature and length has little effect on the results.

These 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 unlikely due to the difference in read length as demonstrated by the right plots.

Gene level detection limits in both platforms

Plot the number of genes above minimal TPM counts in both platforms and datasets. This plots aims at detecting gene-specific differences between platforms (ie. one patform detecting more genes at given low- or high- expression levels). Such differences should not exist unless some transcripts are not as efficiently sequenced as others depending on the technology.

REM: The data being normalized per million, we do not plot here absolute counts but relative fraction of all counts.



As expected, no clear difference is seen between platforms for the same RNA sample.

Comparison between platforms and effect of read lengths

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.

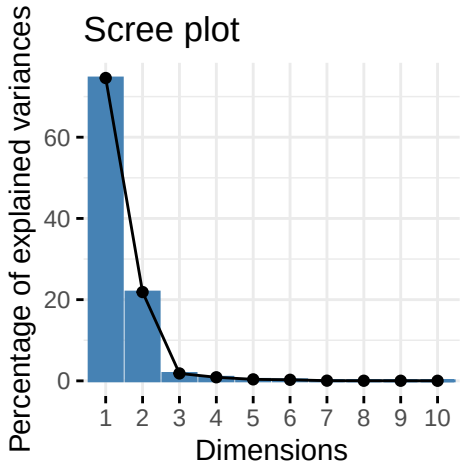
PCA plots for the actual sequencing samples

The variance filtering retained 23951 rows out of 60617 EnsEMBL GeneID rows.

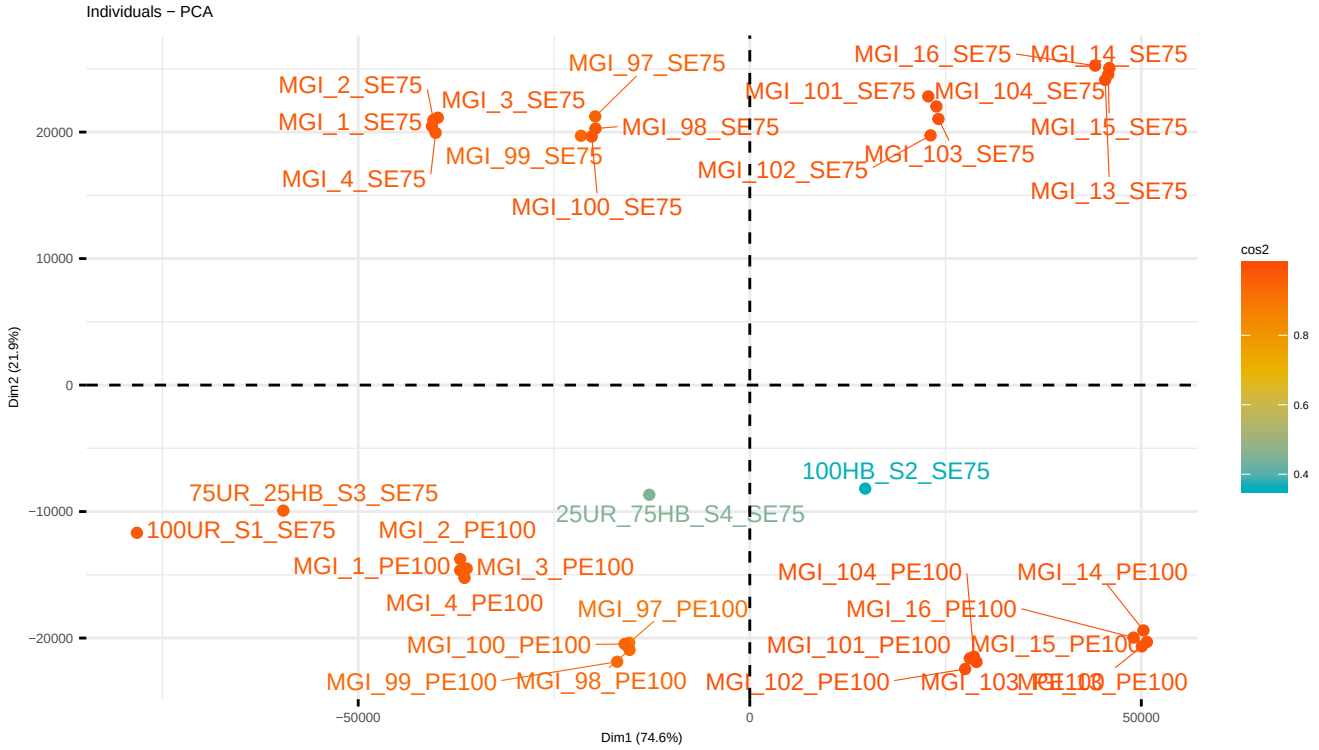
The filtered TPM data is used to compute PCA as seen next.

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     3.710e+04 2.009e+04 5811.5280 4.038e+03 2.622e+03
Proportion of Variance 7.459e-01 2.187e-01    0.0183 8.830e-03 3.720e-03
Cumulative Proportion  7.459e-01 9.646e-01    0.9829 9.917e-01 9.954e-01
                            PC6       PC7       PC8       PC9      PC10
Standard deviation     2.223e+03 950.88211 777.53849 718.84053 639.30475
Proportion of Variance 2.680e-03   0.00049   0.00033   0.00028   0.00022
Cumulative Proportion  9.981e-01   0.99857   0.99890   0.99918   0.99940
                           PC11      PC12      PC13      PC14      PC15
Standard deviation     547.15098 422.4861 409.97965 351.70548 326.81563
Proportion of Variance   0.00016   0.0001   0.00009   0.00007   0.00006
Cumulative Proportion    0.99957   0.9997   0.99975   0.99982   0.99988
                           PC16      PC17      PC18      PC19      PC20  PC21
Standard deviation     275.92343 249.90797 155.44059 128.32857 107.45911 90.15
Proportion of Variance   0.00004   0.00003   0.00001   0.00001   0.00001  0.00
Cumulative Proportion    0.99992   0.99995   0.99997   0.99998   0.99998  1.00
                          PC22  PC23  PC24  PC25  PC26  PC27  PC28  PC29  PC30
Standard deviation     80.12 74.07 59.48 41.79 39.98 36.49 35.13 30.86 28.35
Proportion of Variance  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
Cumulative Proportion   1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00
                          PC31  PC32  PC33  PC34  PC35      PC36
Standard deviation     27.13 26.29 21.27 18.93 16.87 2.493e-11
Proportion of Variance  0.00  0.00  0.00  0.00  0.00 0.000e+00
Cumulative Proportion   1.00  1.00  1.00  1.00  1.00 1.000e+00
```



About **75%** 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.

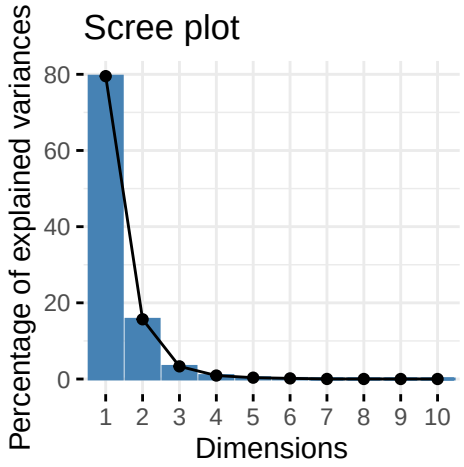
The counts from trimmed MGI-read are distinct from the counts from the full PE100 data, highlighting an yet unexplained effect of read length on the STAR count data.

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

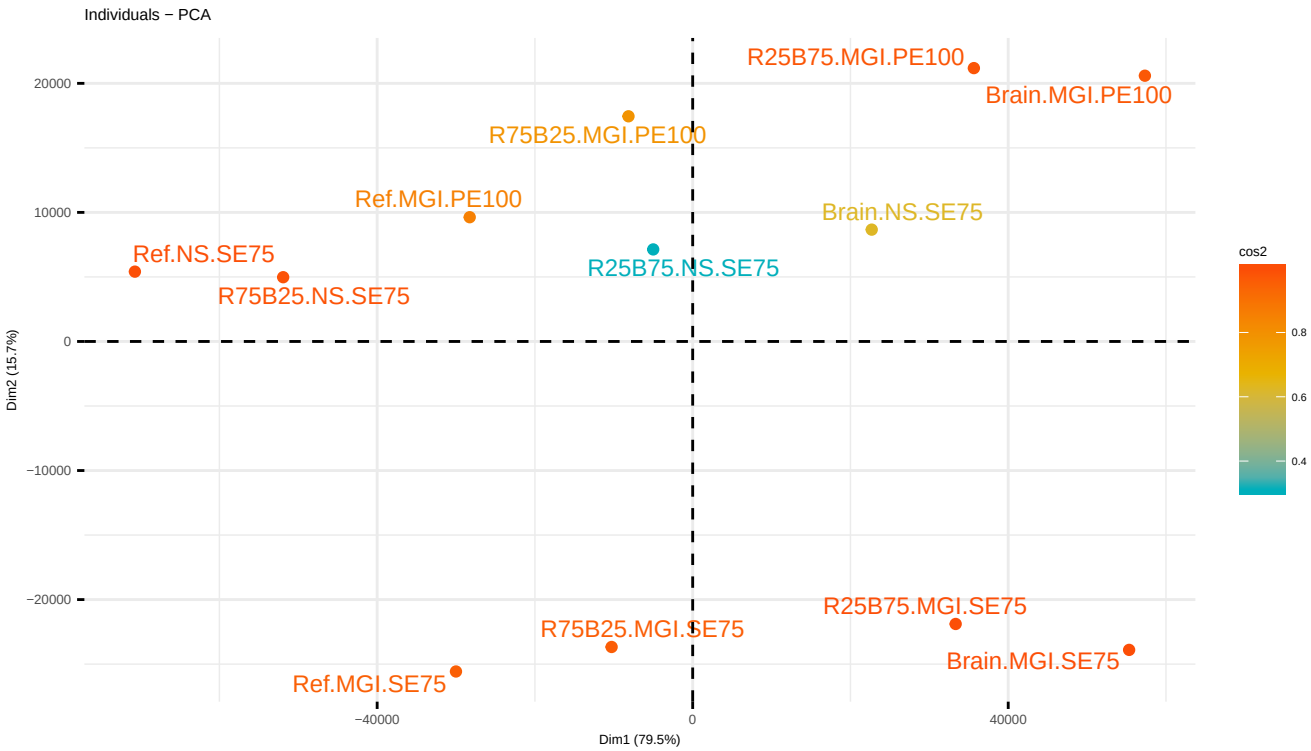
This plot shows the distribution of the 4 MGI group count means together with the 4 corresponding Illumina Novaseq count sets.

The variance filtering retained 22701 rows out of 60617 EnsEMBL GeneID rows.

```
Importance of components:
                            PC1       PC2       PC3       PC4       PC5
Standard deviation     4.138e+04 1.838e+04 8455.8549 4.428e+03 2.770e+03
Proportion of Variance 7.952e-01 1.568e-01    0.0332 9.110e-03 3.560e-03
Cumulative Proportion  7.952e-01 9.520e-01    0.9852 9.943e-01 9.979e-01
                            PC6      PC7       PC8       PC9      PC10
Standard deviation     1.893e+03 652.9769 491.93235 388.08080 353.82686
Proportion of Variance 1.660e-03   0.0002   0.00011   0.00007   0.00006
Cumulative Proportion  9.995e-01   0.9997   0.99985   0.99992   0.99998
                           PC11      PC12
Standard deviation     218.38174 5.496e-12
Proportion of Variance   0.00002 0.000e+00
Cumulative Proportion    1.00000 1.000e+00
```



About **80%** 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.

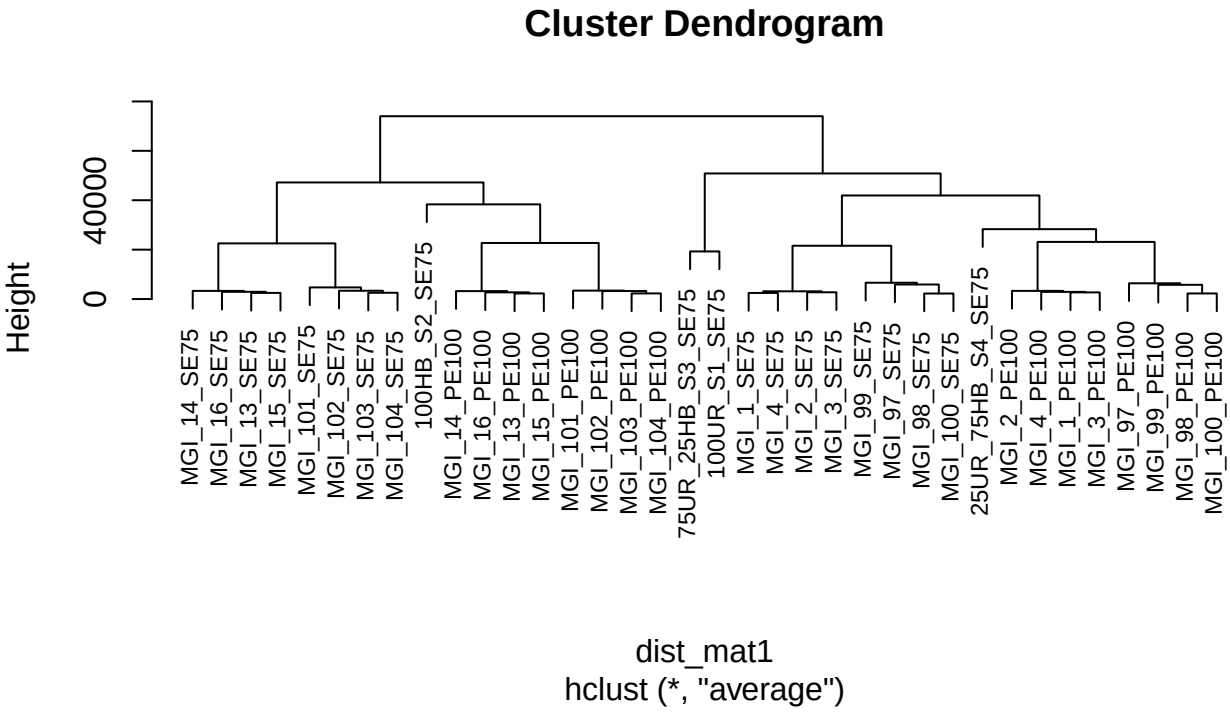
The difference between MGI-PE100 counts and MGI-trimmed SE75 cunts is also seen in this plot, confirming a bias due to read length.

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:

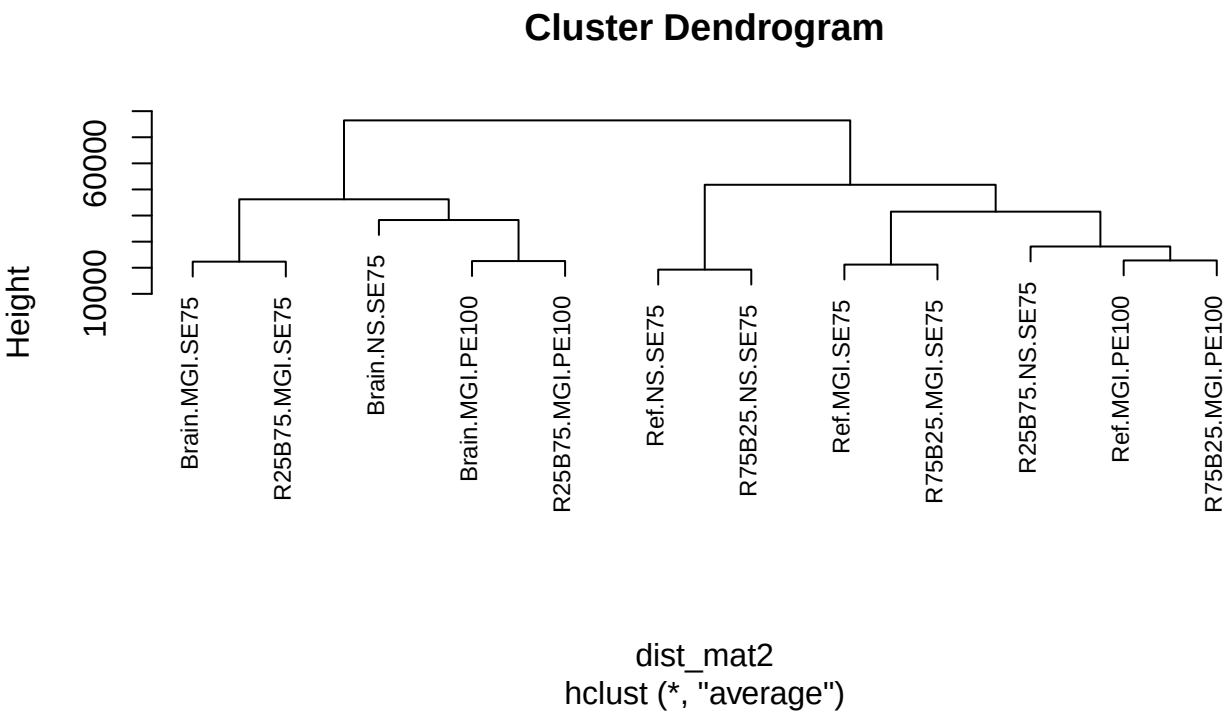
- 100% HumanBrain *MGI [bc13..bc16]*
- 75% HUmanBrain 25% HUmanRef *[MGI bc101..bc104]*
- 100% HumanRef *[MGI bc1..bc4]*
- 75% HUmanRef 25% HUmanBrain *[MGI bc97..bc100]*



The four MGI samples of each groups nicely cluster together, and PE100 and SE75 data are very simiular. the mixed-samples with 75% Human Ref are close to the HumanRef samples (idem for the two Brain groups).

The four Novaseq samples are distinct from the MGI samples but group with the closest related counterparts.

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.

MGI PE100 and SE75 trimmed-read counts cluster separately.

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, even when trimming MGI reads to ressemble SE75 Novaseq reads. This could originate from the different sequencing libraries or reveals a true difference in the sensitivity of both platforms. The difference is identified by PCA analysis but not clearly visible when directly comparing counts from the two platforms.

last edits: Wed Mar 11, 2020

References

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