TransQuin Report (Sample)

Ted Wong

1 December 2015

Introduction

An accurate quantification of variability in RNA sequencing analysis is critical for bioinformatics analysis. To control for these sources of variabiltiy, Garvan Institute of Medical Research has developed a set of spiked-in synthetic sequins known as TransQuin. The sequins emulate a diverse range of genomic features and they are combined together across a range of concentration to formulate a mixture to emulate quantitative features.

Anaquin is a statistical software toolkit built for TransQuin. It is designed to analyze outputs generated by the NGS tools, and accommodates the analysis with detailed statistics on the sequins. Please refer to www.sequin.xyz for more details.

Data

Homo spaiens melanoma cell lines (Hs): These human data correspond to a comparison between a melanoma cell line expressing the Microphtalmia Transcription Factor (MiTF) and a melanoma cell line in which small interfering RNAs (siRNAs) are used against MiTF in order to lower its expression.

Mus musculus muscle stem cells (Mm): These data are related to a transcriptome study where the expression of miRNAs was measured in three different cellular stages of the skeletal muscel lineage in adult mouse.

Mixture

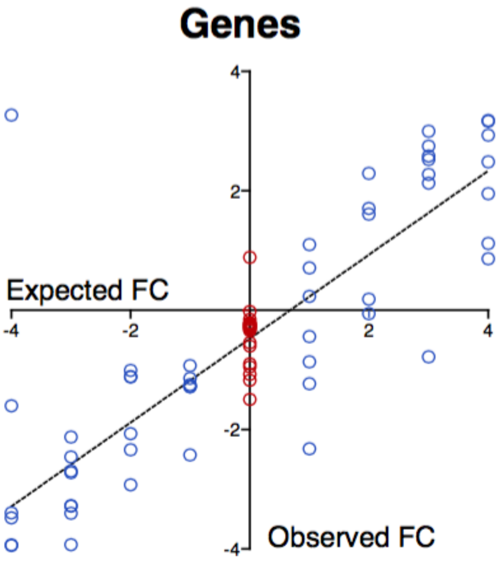
The range of RNA abundance in TransQuin Mix 1 and Mix 2 is used to assess the dynamic range of an experiment. The experiment has a dynamic range and the reference sample RNA-Seq experiment dynamic range spans the design dynamic range. This difference is because of increased sequencing depth in the reference sample experiment. Note that the observed TransQuin control signal-abundance relationship is intended for qualitative assessment of dynamic range. The ERCC controls, as used in these differential gene expression experiments, are not recommended for chemical calibration. The mRNA-enchrichment process, in particular ploy-A selection, can bias the expected signal-abundance relationship of ERCC controls, which have ploy-A tails ranging from ~20 to 26nt, significantly shorter than endogenous transcript ploy-A tails.

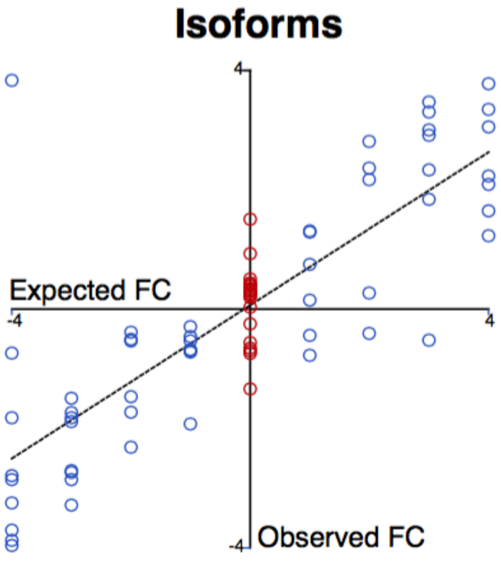
Methods

The datasets included in this study were obtained from two different Illumina sequencing machines, differing in their read length and overall throughput but sharing the same sequencing technology that takes place on a glass slide called a 'flow cell'. A flow cell is made up of eight independent sequencing areas, or 'lanes'. Libraries are deposited on these lanes in order to be sequenced. A library contains cDNAs representative of the RNA molecules that are extracted form a given culture or tissue and are pre-processed in order to be adapted to the sequencing procedire. Simularly to microarrays, the library composition reflects the RNA repertoire expressed in the corresponding culture or tissue. The 'library size' refers to the number of mapped short reads obtained from the sequencing process of the library. In this study, a single library was sequenced in each lane.

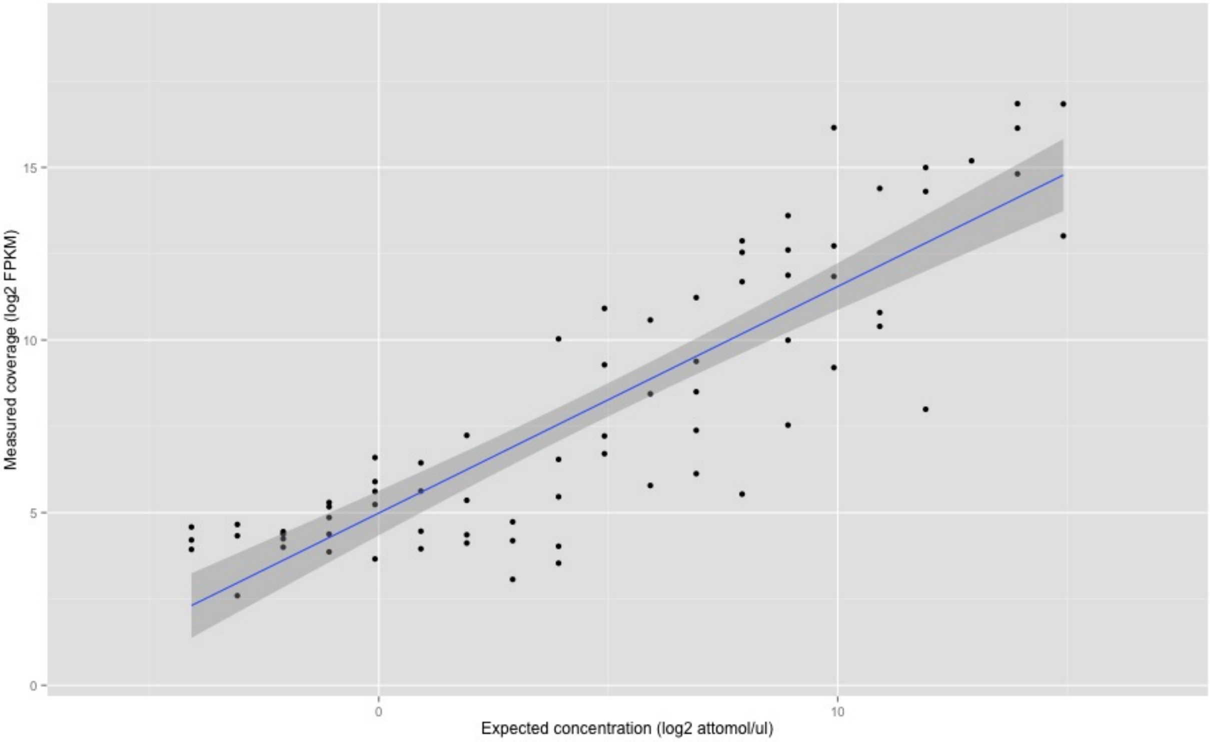
Expression Analysis

Out of 160 sequins, 150 of them have been detected.





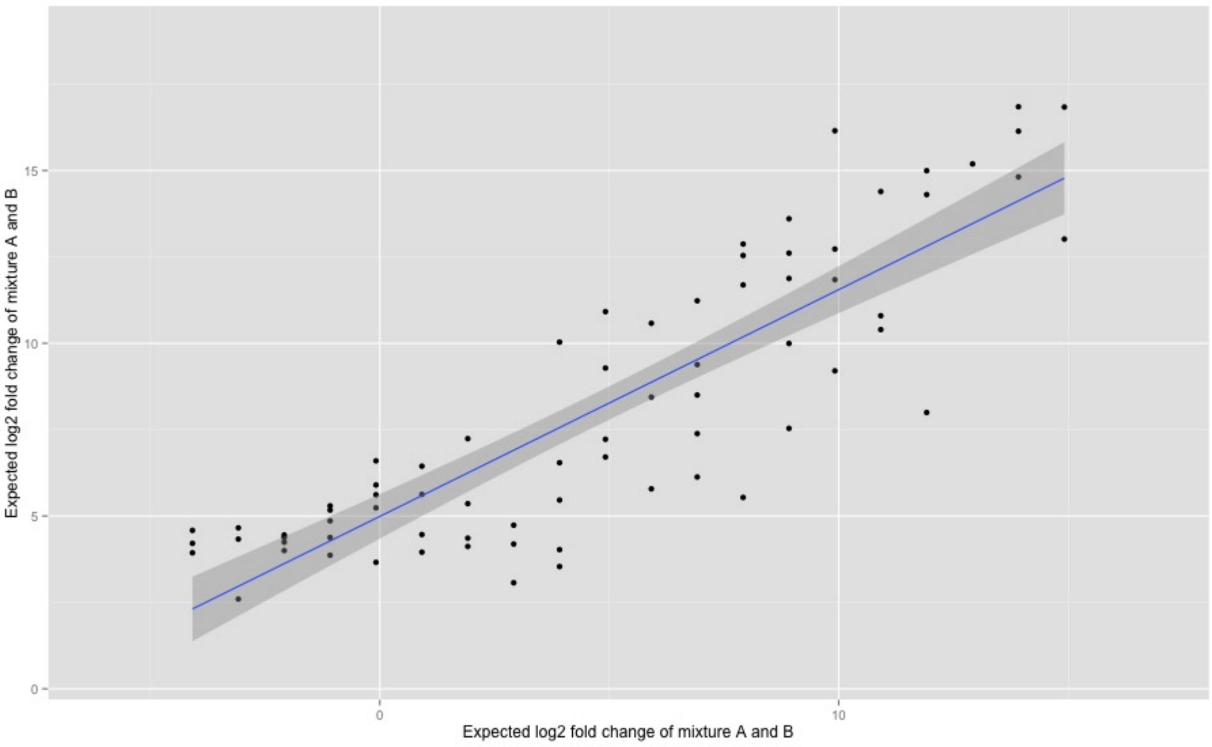
The observed TransQuin known-measured relationship is intended for qualitative assessment of dynamic range. A linear regression and it's 95% confidence interval is also plotted.



##

```
## Call:
## lm(formula = ly ~ lx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8147 -1.5081  0.5811  1.6620  4.6541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.98851    0.32408   15.39   <2e-16 ***
## lx           0.65628    0.04564   14.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.067 on 71 degrees of freedom
## Multiple R-squared:  0.7444, Adjusted R-squared:  0.7408
## F-statistic: 206.8 on 1 and 71 DF,  p-value: < 2.2e-16
```

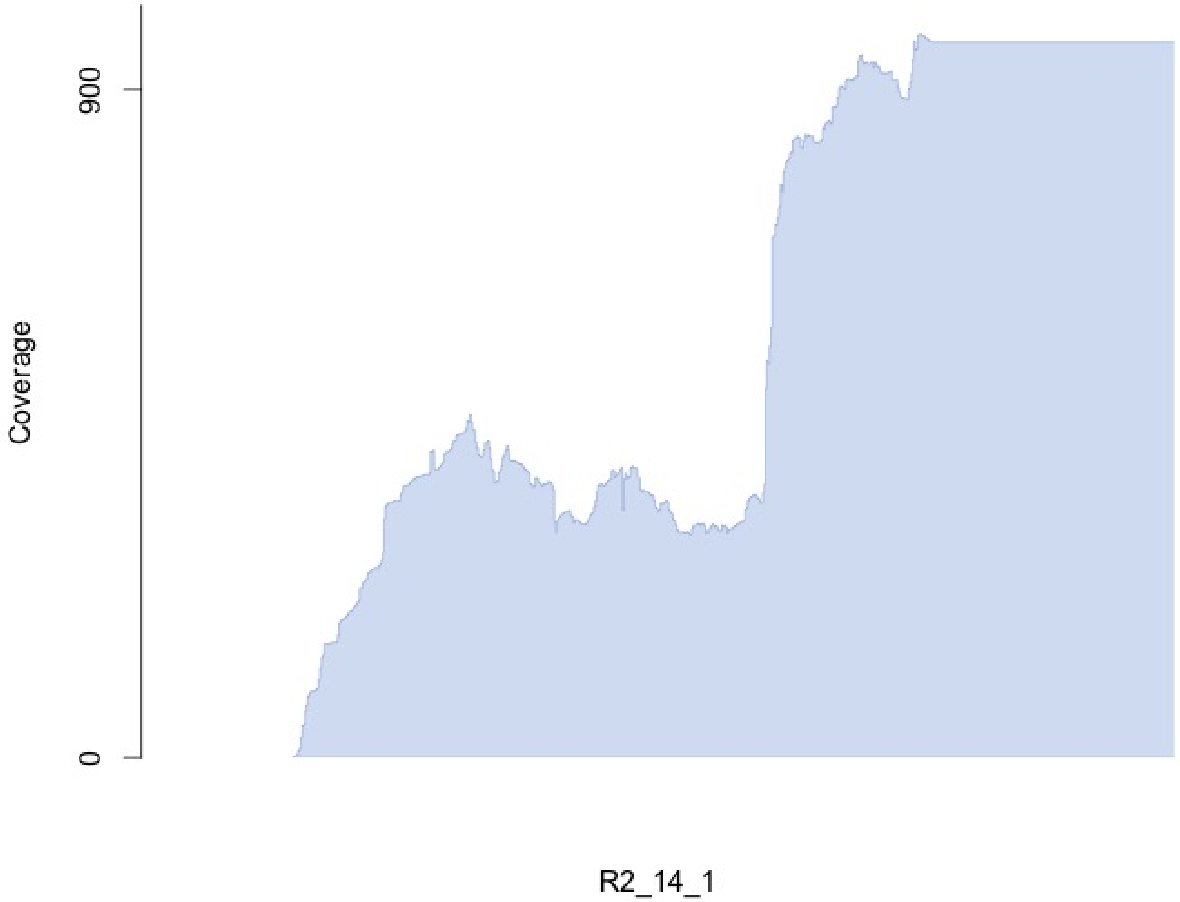
Differential Analysis



```
##
## Call:
## lm(formula = ly ~ lx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8147 -1.5081  0.5811  1.6620  4.6541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.98851    0.32408   15.39   <2e-16 ***
## lx           0.65628    0.04564   14.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.067 on 71 degrees of freedom
## Multiple R-squared:  0.7444, Adjusted R-squared:  0.7408
## F-statistic: 206.8 on 1 and 71 DF,  p-value: < 2.2e-16
```

Apprendix

Density plot for *R2_14_1*


Coverage
R2_14_1