BF528 Individual Project
Jason Yeung

Project 3: Concordance of microarray and RNA-Seq differential gene expression
Role: Data Curator

Introduction

The goal of Wang et al. was to analyze concordance between RNA expression datasets generated using microarray and RNA-seq techniques. Here, the microarray dataset has been processed for downstream analyses already, so the focus is on processing the RNA-seq dataset. The initial data curation steps are important in all RNA-seq analysis workflows, in order to perform initial data quality control and identify any potential outlier samples that can influence further analyses.

Methods

These quality control steps were done by first identifying and running FastQC on the nine treatment samples that are part of toxgroup 2. STAR aligner was then used to align each sample against the rat genome, as described in the paper. Alignment statistics from running STAR are reported in Table 1 below. MultiQC was also run to compile FastQC and STAR statistics into a single report. Relevant plots from the MultiQC report can also be referenced in Figures 1 and 2 in the results section below.

Results

Sample	Uniquely Aligned Reads	Multi-mapped Reads	Unmapped Reads
SRR1177966	15,983,143 (84.2%)	807,481 (4.3%)	2,134,238 (11.2%)
SRR1177969	16,494,573 (85.4%)	695,194 (3.6%)	2,087,700 (10.8%)
SRR1177970	16,262,255 (85.1%)	670,339 (3.5%)	2,148,781 (11.2%)
SRR1177993	17,094,859 (86.2%)	858,349 (4.3%)	179,631 (9.1%)
SRR1177994	18,514,298 (88.0%)	859,510 (4.1%)	1,612,036 (7.7%)
SRR1177995	17,661,373 (87.6%)	794,497 (3.9%)	1,633,442 (8.1%)
SRR1177998	20,972,287 (88.8%)	936,865 (4.0%)	1,578,961 (6.7%)
SRR1178001	22,219,635 (89.1%)	974,756 (3.9%)	1,642,973 (6.6%)
SRR1178003	22,041,739 (89.2%)	926,439 (3.7%)	1,670,458 (6.8%)

Table 1. Table of STAR read and alignment statistics.

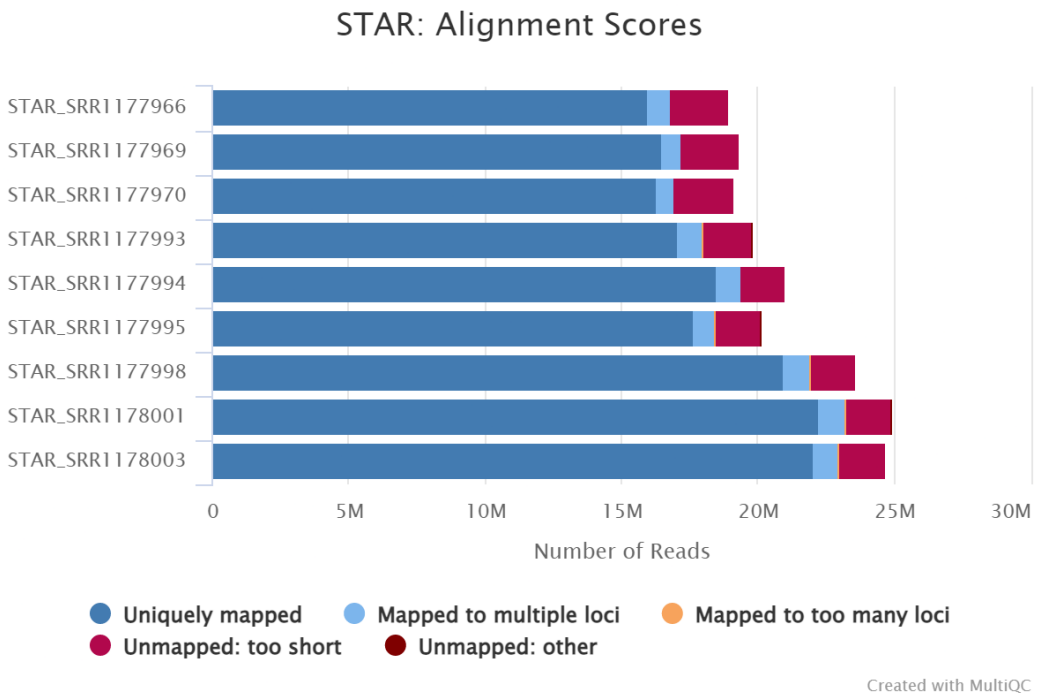

Figure 1. STAR alignment scores showing breakdown of reads mapped to rat genome for each sample.

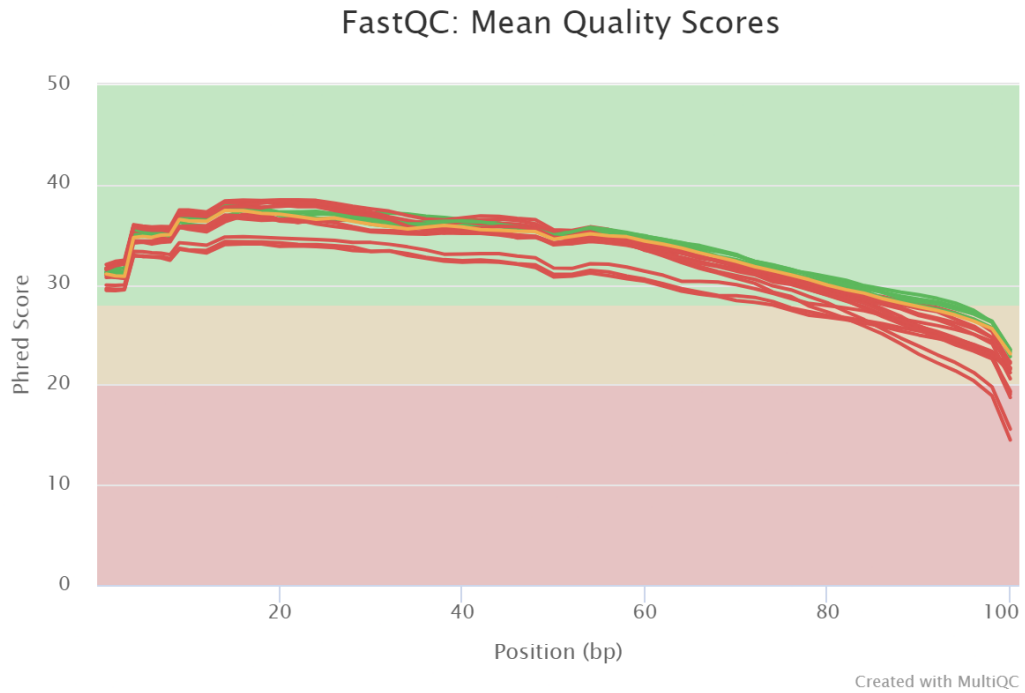

Figure 2. FastQC results showing Phred quality score per sample across all basepair positions.

The resulting alignment statistics from mapping the reads to the rat genome using STAR are shown in Table 1 and Figure 1 above. Across all samples, the percentage of uniquely aligned reads was relatively high, suggesting that the samples and sequencing results are viable. Figure 2 shows the Phred quality score for each sample across all basepair positions. The Phred score is relatively high for all reads, with quality dropping off towards the end of the reads, which is often observed in Illumina sequencing due to inefficient sequencing by synthesis when reads are longer in length. However, this is not a significant problem because the majority of nucleobases in these reads are high quality.

Discussion

The goal of these initial preprocessing steps was to perform quality control on and align the RNA-seq reads to the rat genome. If any samples had a low percentage of reads mapped to the genome or low quality reads, it may be an indication of problems during experimental design or problems during the sequencing process. In this case, using toxgroup 2 as a subset of the overall dataset, it seems that all nine treatment samples showed high read quality and mapped well to the respective genome. They are therefore informative samples that can be used for analyses further downstream.

References

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Łabaj, P. P., Kreil, D. P., Megherbi, D., Gaj, S., Caiment, F., van Delft, J., … Tong, W. (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nature biotechnology, 32(9), 926–932. https://doi.org/10.1038/nbt.3001

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England), 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Ewels, P., Magnusson, M., Lundin, S., Käller M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics (2016). https://doi.org/10.1093/bioinformatics/btw354