Concordance of microarray and RNA-Seq differential gene expression

BF528 Individual Project
Zhiyu Zhang (Data curator, Programmer)
group dreadlocks

Introduction

Data Curator

The goal of the data curator role in this project is to align short reads from the RNA sequencing experiment conducted by Wang et al. [1] to the rat genome, and to perform quality control on both the original sequencing data and the alignment result. This step is crucial for the overall study, because the validity and accurate interpretation of all downstream analysis is ensured by inputting reliable raw data at the start of the pipeline.

Programmer

The goal of the programmer role in this project is to quantify mRNA abundance as a measure of gene expression, and to perform differential expression analysis on read counts. This step provides analytical results from the RNA-Seq platform in order to compare differentially expressed genes (DEGs) and enriched pathways found by RNA-Seq and Affymetrix microarray in following concordance analyses.

Method

Data Curator

Nine treatment (two runs each) and six control samples of paired-end RNA-Seq samples were obtained from the group project directory on SCC. Treatment samples consisted of three replicates for each of three toxicological treatments from toxgroup 6 (3-methylcholanthrene, fluconazole, pirinixic acid). Quality control was performed on fastq files of the nine treatment samples using *FastQC* version 0.11.9[2], and reads were aligned to rat genome using the *STAR* aligner version 2.6.0c[3] and a genome index provided for this project. *MultiQC* version 0.9.1a0[4] was used to summarize quality control and alignment results.

Programmer

Read summarization was achieved using the *Subread featureCounts* package (1.6.2)[5], and *MultiQC* was used to report mapping statistics. Multi-mapped reads were not counted. Genes with zero count across all samples were removed from the count matrix, and differential expression analysis was performed with *DESeq2* (1.30.1)[6] in R version 4.0.3 for the three chemical treatment samples with their corresponding control samples. Effect size shrinkage was applied to DE results using the *apeglm* package[7] in order to obtain better gene rankings. Histograms, volcano plots and top 10 DEGs were generated using *apeglm* shrunken estimates, while significantly differentially expressed genes at P value < 0.05 results were generated without shrinkage.

Results

<u>Data Curator</u>

According to FastQC report, 13 out of 18 samples failed per base sequence quality check. Mean quality scores appeared to drop towards end positions, illustrated in **Figure 1**. 14 samples failed the sequence duplication levels check, with the rest marked as warnings. All samples failed the per base sequence content check and passed the rest of status checks.

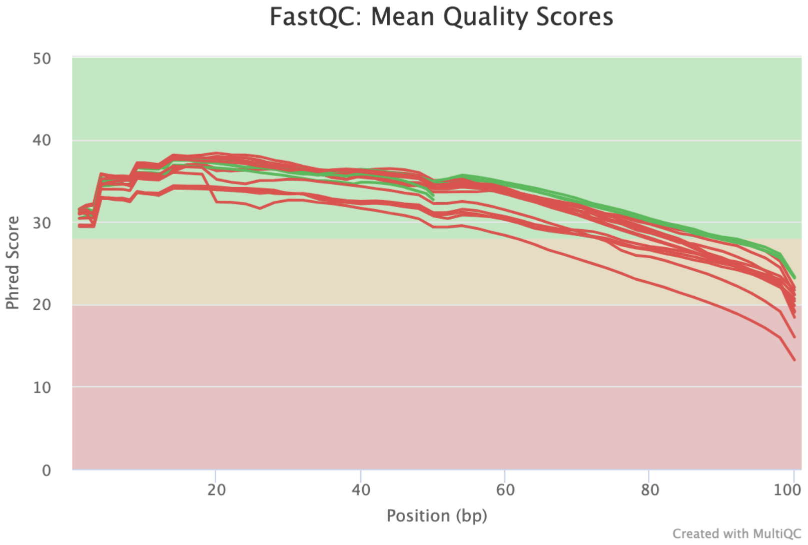

Figure 1. Quality drop at end positions

Read and alignment statistics are summarized in **Table 1**.

Table 1. Read and alignment statistics

Sample ID	Total Sequences	Read length	Uniquely mapped	Mapped to multiple loci	Mapped to too many loci	Unmapped
SRR1177963	17897455	202	85.5%	3.8%	0.1%	10.6%
SRR1177964	19342910	202	86.0%	3.8%	0.1%	10.2%
SRR1177965	16849678	202	85.7%	4.0%	0.1%	10.2%
SRR1177997	19746775	202	89.9%	4.0%	0.1%	6.1%
SRR1177999	21838440	202	89.3%	4.0%	0.1%	6.4%
SRR1178002	18844950	202	89.7%	4.0%	0.1%	6.1%
SRR1178014	17524782	100	83.6%	6.8%	0.2%	9.4%

SRR1178021	17497925	200	82.7%	6.0%	0.2%	11.2%
SRR1178047	17093302	200	84.5%	6.1%	0.2%	9.2%

Programmer

The majority of reads across the nine treatment samples were assigned to genomic features (54%-62%). Around 12% of reads were multimapped for 3-methylcholanthrene- and fluconazole-treated samples, while pirinixic acid-treated samples had a relatively higher multimapping rate (~19%).

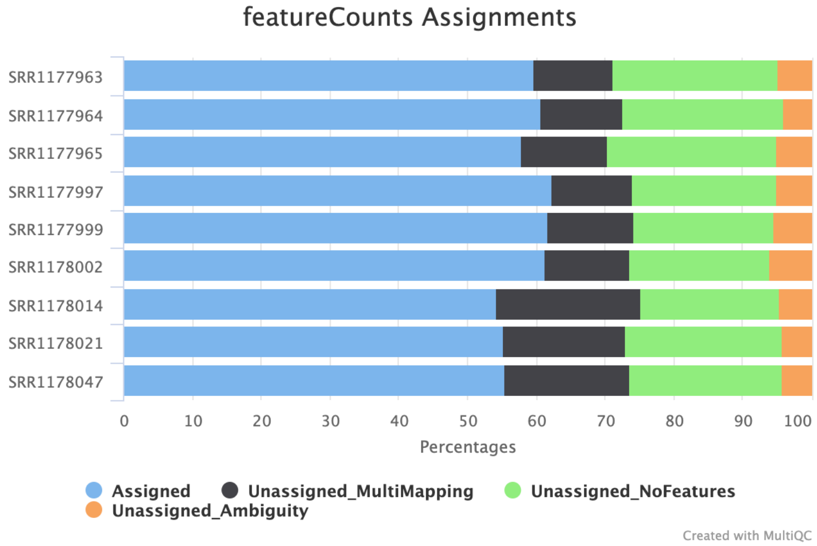

Figure 2. Summary of read counts mapped for genomic features

Count distributions are illustrated in **Figure 3**. Read counts were relatively uniformly distributed across treatment samples (**Figure 3**). After removing zero-count genes, there were 11001 genes in the count matrix. Number of differentially expressed genes with adjusted p-value < 0.05 for each chemical treatment are reported in **Table 2**. Top 10 DEGs are reported in supplementary **Table S1**. Histograms of log 2 fold change values (**Figure 4**) and volcano plots of log2 fold change vs unadjusted -log10 P-values (**Figure 5**) were used to visually inspect and compare DE results.

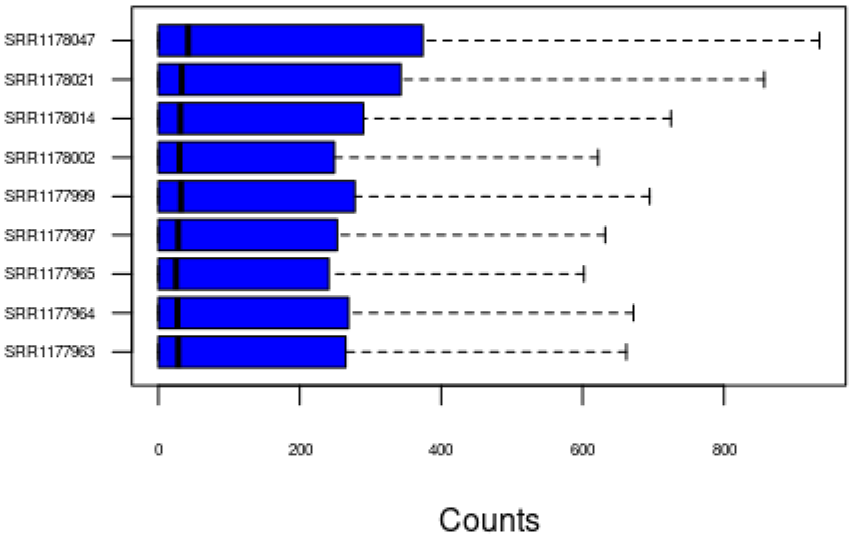

Figure 3. Boxplot of read counts distributions for treatment samples

Table 2. Number of genes significant at p-adjust < 0.05

MOA	Chemical treatment	# DEGs
AhR	3ME	296
CAR/PXR	FLU	3697
PPARA	PIR	2624

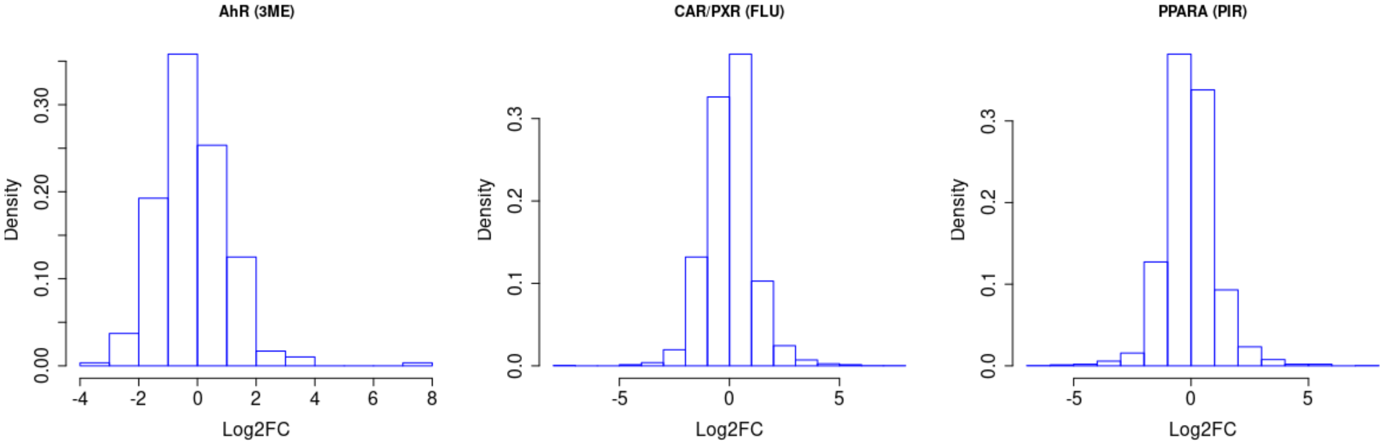

Figure 4. Histograms of log2 Fold Change values

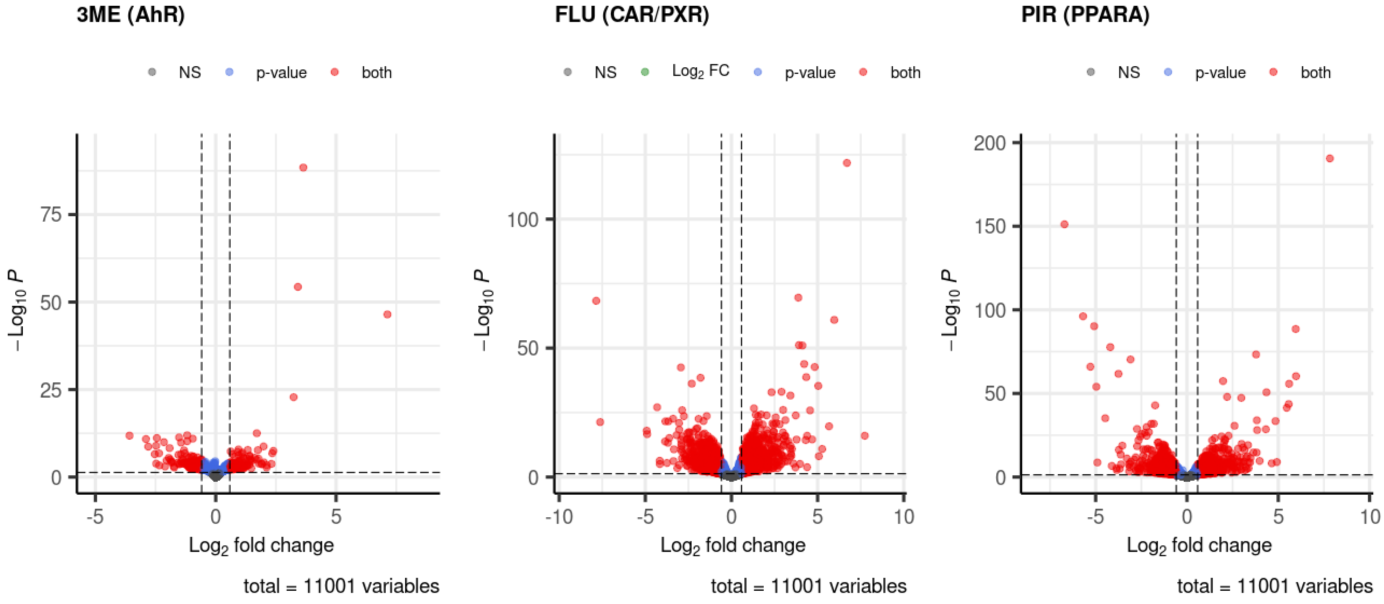

Figure 5. Significantly differentially expressed genes by fold change and P values. Blue points correspond to DEGs with P<0.05, red points correspond to DEGs with P<0.05 and absolute FC > 1.5.

Discussion

Data Curator

Quality control of raw sequencing data and alignment results in RNA-Seq experiments is important for accurate interpretation of downstream analysis. The metrics that failed or received warnings according to *FastQC* are likely a result of signal decay or phasing, which are inherent characteristics of reads generated by Illumina sequencing or RNA-Seq experiments in general[8]. Therefore, the quality of both the raw sequencing reads and the alignment results are considered acceptable. On average, 86% of reads were uniquely mapped and 5% were mapped to multiple loci to the rat genome.

Programmer

Read summarization was performed using *featureCounts* in order to quantify gene expression for different samples. On average less than 60% of reads across the nine treatments were assigned to genomic features, and a significant portion of reads were unassigned due to multi-mapping or not overlapping with any feature. The identity of these reads remained unclear. Differential expression analysis found 296, 3697 and 2624 DEGs for 3-methylcholanthrene-, fluconazole-, and pirinixic acid-treated samples respectively compared to untreated control samples.

Supplementary Materials

Table S1. Top 10 DEGs ranked by adjusted P value

	log2 FC	P_adj
Up-regulated AhR		
NM_012541	3.639131922	3.99E-85
NM_130407	3.414055245	2.49E-51
NM_012540	7.132694012	1.23E-43
NM_001109459	3.237287915	4.06E-20
NM_022521	1.707888228	6.51E-10
Down-regulated AhR		
NM_053883	-1.18616178	2.11E-09
NM_001109022	-3.582593466	2.32E-09
NM_134329	-1.522410139	6.24E-09
NM_022866	-2.446048063	8.62E-09
NM_022297	-0.9542582547	1.17E-08
Up-regulated CAR/PXR		
NM_053699	6.701122808	1.76E-118
NM_031605	3.880931861	1.55E-66
NM_013033	5.965896879	3.19E-58
NM_001005384	3.916228008	1.54E-48
NM_144755	4.117686805	1.71E-48
NM_031048	4.218587158	2.35E-41
NM_013105	4.828270012	2.94E-40
NM_053288	4.339586237	2.00E-36
Down-regulated CAR/PXR		
NM_001130558	-7.850731182	1.65E-65
NM_001014166	-2.935967598	4.01E-40
Up-regulated PPARA		
NM_024162	7.828272934	2.96E-187
NM_019157	5.955131078	6.03E-86
NM_012600	3.786615421	7.64E-71
Down-regulated PPARA		

NM_012737	-6.710651137	3.69E-148
NM_017158	-5.689819267	2.69E-93
NM_131903	-5.084782368	1.63E-87
NM_001014063	-4.199845703	3.72E-75
NM_053883	-3.089161447	6.34E-68
NM_001013098	-5.291089415	1.28E-63
NM_001013975	-3.747693324	1.98E-59

References

1. 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
2. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.
3. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15-21.
4. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047-3048.
5. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England), 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656
6. Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.
7. Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895.
8. Mary Piper, R. K. (2018, September 5). Quality control: Assessing FASTQC results. Retrieved May 9, 2021, from Github.io website: https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html