Project 5: Individual Project

Group: Dreadlocks

Michael Peters

Report 1: Data Curator - Concordance of Microarray and RNA-Seq Differential Gene Expression
Report 2: Programmer - Concordance of Microarray and RNA-Seq Differential Gene Expression

Data availability: Scripts used in both analyses are aggregated as a single Nextflow script in the following GitHub repository: https://github.com/mpeters15/BF_528_project5

Introduction

Wang et al. present a study comparing gene expression data from RNA-Seq and microarray technologies. Their study uses a set of toxicological treatments with known mechanisms of action (MOA) measured in rat liver. In an attempt to reproduce the RNA-Seq sample statistics and alignment generated by Wang et al, this report performs the tasks of gathering sample files for a chosen subset of Wang et al's data (i.e. a toxgroup), aligning each sample from this subset's data against the rat genome, and gathering quality control data on these data. The results from this report are important for the overall study as it ensures both that the sequencing performed on these samples by Wang et al. was done correctly and that the data used in downstream analyses is of high-enough quality to make reliable conclusions on.

Methods

In Wang et al's analysis, Illumina RNA-seq and Affymetrix microarray data were generated from rat liver samples. These data sets are made available to us from accessions SRP039021, GSE55347, and GSE47875, and are categorized in toxgroups. A subset of samples from toxgroup 6 are chosen for this report. Toxgroup 6 contains nine experimental and six control samples in total. The three different chemical agents and their known associated MOAs in toxgroup 6 are as follows: 3-Methylcholanthrene and aryl hydrocarbon receptor (AhR), Fluconazole and orphan nuclear hormone receptors (CAR/PXR) and Pirinixic acid and peroxisome proliferator-activated receptor alpha (PPARA). Six appropriate control samples are also analyzed.

Raw, paired-end RNA-seq counts data from the nine experimental samples from toxgroup 6 are compiled and analyzed using *FastQC* (v0.11.7) to determine the quality of the raw sequence data. The nine samples are then aligned to the provided reference rat genome using the *STAR* aligner (v2.6.0c). The alignment statistics generated from this step, as well as the *FastQC* metrics, are then run through *MultiQC* (v1.6) to generate a summary report containing quality control metrics and alignment statistics for all samples.

Results

FastQC, *STAR*, and *MultiQC* are used to process each treatment sample's raw RNA-seq data and generate quality control metrics for it. Table 1 displays select alignment statistics generated by *STAR*. Table 2 depicts additional quality statistics generated by *FASTQC* and *STAR*, and aggregated by *MultiQC*. All samples show a median 85.1 ± 2.6% uniquely aligned reads, indicating that the samples are successfully aligned, with low variance. All samples additionally have a low percentage of unmapped reads.

Figure 1A displays the quality scores for each sample's base calls. Higher scores indicate better base calls. The quality of the calls lessens as the run progresses, which is a common occurrence. All of the samples have a low percentage of base calls at each position for which an N was called, meaning that, during sequencing, the sequencer was able to make base calls with sufficient confidence at least 98.5% of the time (Figure 1B). One potentially concerning attribute

of most of the samples is a high relative level of duplication found for every sequence. Figure 1C indicates the occurrence of either specific enrichments of subsets or the presence of low complexity contaminants, as seen by the spikes towards the right of the plot. Figure 1D shows mostly acceptable mean quality values across each base position in each sample.

Figure 2 displays the number of mapped and unmapped genes in *STAR* alignment. Again, it is observed that most of the genes are uniquely mapped within each sample. Most samples have similar amounts of successfully mapped reads.

Sample	Total Reads	Avg. Read Length	Uniquely Mapped Reads	Multi-Mapped Reads	Unmapped Reads
SRR1177963	17897455	202	84.83%	3.70%	11.28%
SRR1177964	19342910	202	85.33%	3.67%	10.78%
SRR1177965	16849678	202	85.10%	3.85%	10.82%
SRR1177997	19746775	202	89.17%	3.88%	6.77%
SRR1177999	21838440	202	88.72%	3.92%	7.10%
SRR1178002	18844950	202	89.13%	3.92%	6.73%
SRR1178014	17524782	100	83.52%	6.57%	9.41%
SRR1178021	17497925	200	81.95%	5.77%	11.98%
SRR1178047	17093302	200	83.98%	5.85%	9.73%

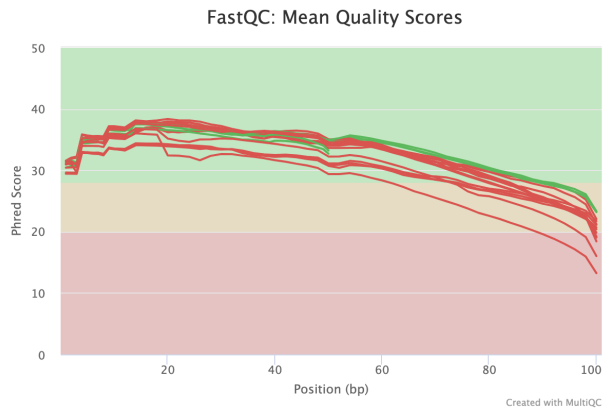
Table 1. Select *STAR* Alignment Statistics, as generated by the aligner.

Sample Name	% Aligned	M Aligned	% Dups	% GC	Length	M Seqs
SRR1177963	84.80%	15.2				
SRR1177963_1			54.80%	48%	101 bp	17.9
SRR1177963_2			52.30%	48%	101 bp	17.9
SRR1177964	85.30%	16.5				
SRR1177964_1			57.30%	48%	101 bp	19.3
SRR1177964_2			54.80%	48%	101 bp	19.3
SRR1177965	85.10%	14.3				
SRR1177965_1			54.50%	48%	101 bp	16.8
SRR1177965_2			51.80%	48%	101 bp	16.8
SRR1177997	89.20%	17.6				
SRR1177997_1			59.60%	49%	101 bp	19.7
SRR1177997_2			58.60%	49%	101 bp	19.7
SRR1177999	88.70%	19.4				

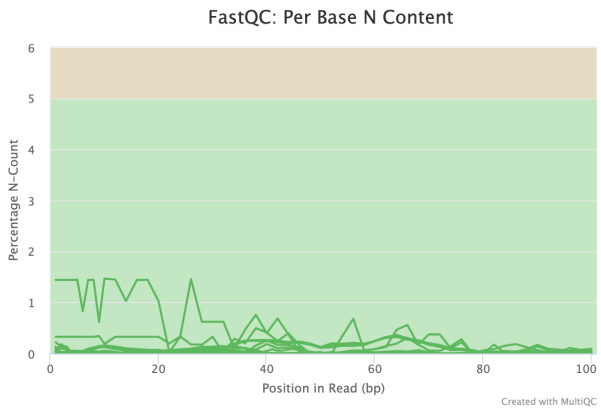
SRR1177999_1			60.20%	49%	101 bp	21.8
SRR1177999_2			58.90%	49%	101 bp	21.8
SRR1178002	89.10%	16.8				
SRR1178002_1			58.50%	49%	101 bp	18.8
SRR1178002_2			57.60%	49%	101 bp	18.8
SRR1178014	83.50%	14.6				
SRR1178014_1			53.90%	49%	50 bp	17.5
SRR1178014_2			51.90%	49%	50 bp	17.5
SRR1178021	82.00%	14.3				
SRR1178021_1			48.70%	49%	100 bp	17.5
SRR1178021_2			46.40%	49%	100 bp	17.5
SRR1178047	84.00%	14.4				
SRR1178047_1			48.50%	49%	100 bp	17.1
SRR1178047_2			47.30%	49%	100 bp	17.1

Table 2. Metrics created by *FASTQC* and *STAR*, and aggregated by *MultiQC*. *FASTQC* metrics are located in the non-shaded rows, while the shaded rows indicate *MultiQC* metrics. The metrics that *FastQC* generates includes the percentage of duplicate sequences, the average percentage of GC content, the read length, and the total number of sequences (in millions). *STAR* produces metrics regarding the percent of reads that are uniquely aligned against the reference genome and the number of uniquely mapped reads (in millions).

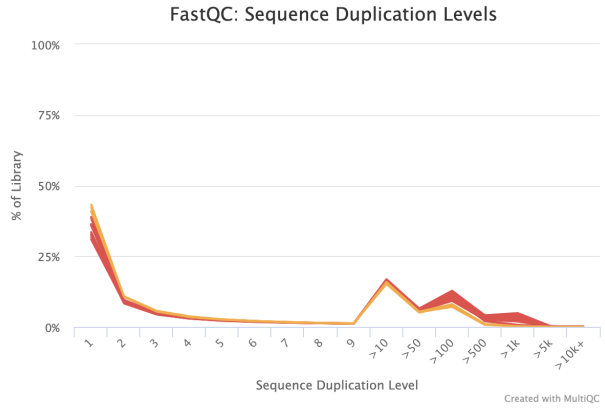
A.



B.



C.



D.

Figure 1. *MultiQC* plots displaying various metrics generated by *FASTQC*. **A.** The mean quality score across each base position in the read, across each sample. **B.** The percentage of base calls at each read position for which an N was called. **C.** The relative level of duplication found for every sequence. **D.** The mean quality value across each base position in the read.

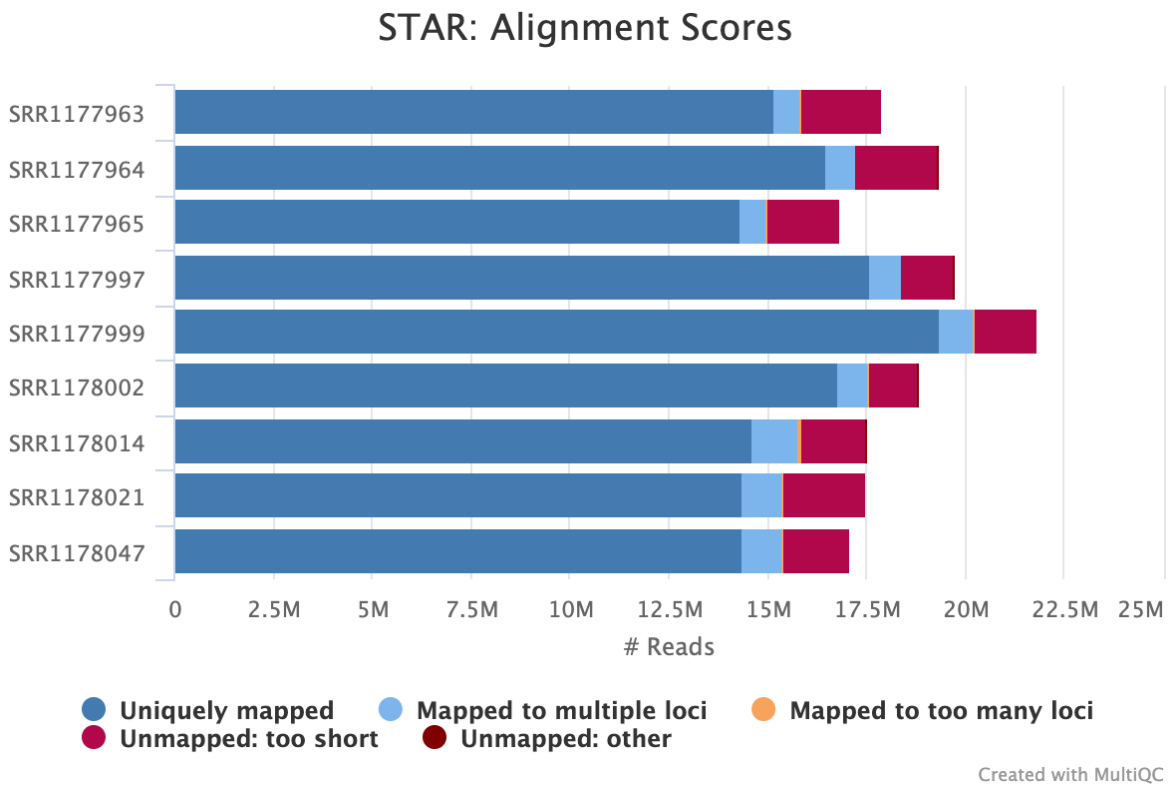


Figure 2. *MultiQC* metrics generated by *STAR*, showing the number of mapped and unmapped genes.

Discussion

The analyses performed in this report are done so with the goal of ensuring that this data is of high-enough quality to be able to be used in downstream analysis. Every sample's sequencing data appears to be of good quality. Around 85% of reads are uniquely aligned to the reference genome, and during sequencing, the sequencer was able to make base calls with sufficient confidence almost all of the time. One area for concern, however, may be the elevated duplication rate observed, indicating the possible presence of contaminants in the sequenced sample. Of importance, the data analyzed in this report appear usable for further analysis. This helps support the original results of Wang et al.

References

1. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.
3. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr
4. Martin Morgan (2021). BiocManager: Access the Bioconductor Project Package Repository. R package version 1.30.16. https://CRAN.R-project.org/package=BiocManager
5. Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, 1 October 2016, Pages 3047–3048, https://doi.org/10.1093/bioinformatics/btw354
6. Wang, Charles et al. "The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance." Nature biotechnology vol. 32,9 (2014): 926-32. doi:10.1038/nbt.3001

Introduction

Wang et al. present a study comparing gene expression data from RNA-Seq and microarray technologies. Their study uses a set of toxicological treatments with known mechanisms of action (MOA) measured in rat liver. In an attempt to reproduce a portion of read counting analysis and RNA-Seq differential expression analysis done by Wang et al, this report performs the tasks of using *STAR* alignments generated during prior analysis of Wang et al's paper, counting reads mapped to genomic features, such as genes, exons, promoters and genomic bins, and running differential expression analysis with *DESeq2*. The results from the gene count analysis and the differential gene expression analysis are important for the overall study because it generates the data that will later be compared to similarly-generated microarray data to identify concordance of microarray and RNA-Seq differential gene expression. It provides quantification of transcript abundance measurements for comparison with the microarray data. Additionally, it provides the basis for a gene-set enrichment analysis similar to that which was performed in the original study.

Methods

In Wang et al's analysis, Illumina RNA-seq and Affymetrix microarray data were generated from rat liver samples. These data sets are made available to us from accessions SRP039021, GSE55347, and GSE47875, and are categorized in toxgroups. A subset of samples from toxgroup 6 are chosen for this report. Toxgroup 6 contains nine experimental and six control samples in total. The three different chemical agents and their known associated MOAs in toxgroup 6 are as follows: 3-Methylcholanthrene and aryl hydrocarbon receptor (AhR), Fluconazole and orphan nuclear hormone receptors (CAR/PXR) and Pirinixic acid and peroxisome proliferator-activated receptor alpha (PPARA). Six appropriate control samples are also analyzed.

For each of the nine treatment samples of toxgroup 6, alignments generated by the STAR aligner in a prior analysis are used to count reads against a gene annotation using the program *featureCounts* (v1.6.2). *featureCounts* is a read summarization program that counts reads mapped to genomic features, such as genes, exons, promoters and genomic bins. This generates nine files containing counts data for each of the bam files fed to the program. *MultiQC* (v1.6) then is run on all of the resulting counts files. The counts files from each sample are joined into a single comma-delimited text file via R. The first column of this file contains one gene name per row, and the remaining columns hold the counts taken from each sample's counts file. A box plot to depict the distribution of each sample's counts is then generated.

Next, a counts matrix for the control samples appropriate to toxgroup 6 is created using a provided sample metadata file. These counts data are merged with the counts data for the treatment samples. The merged data frame is then subset into three new data frames. Each of these three data frames contains the counts data for one of the three modes of action in toxgroup 6 (AhR, CAR/PXR, and PPARA), as well as the counts data for the control samples delivered with the same vehicle as the treatment with that mechanism of action.
After counting reads, these counts are analyzed via the program *DESeq2* (v1.32.0) to determine differential expression between the treated samples and appropriate controls. *DESeq2* is run on

each of the three data frames. The *lfcShrink()* function is then applied to the results object of each *DESeq2* analysis. The differential expression results are saved after sorting by adjusted p-value, and the number of genes significant at p-adjust < 0.05 are reported. A histogram of fold change values from the significant DE genes for each analysis, as well as a scatter plot of fold change vs nominal p-value, is created.

Results

Figure 1 shows a summary of the statistics generated by *featureCounts*, displaying the numbers of reads that map to various features for each sample. Table 1 shows these statistics in more detail. The proportions of assigned vs unassigned reads are similar across samples, and no significant issues are evident in these results. Figure 2 shows the distribution of gene counts across the samples from toxgroup 6. Again, the distribution of counts appears to be similar across samples, suggesting that there are no striking issues with the counts data used in this report. The similar distribution across samples is important in ensuring that no single sample is overrepresented in the differential gene expression analysis.

DESeq2 analysis is performed on the counts data produced by *featureCounts*. Analysis with *DESeq2* provides information on the differential expression of genes between the treated samples and the appropriate controls mentioned in this report. Under the mechanisms of action of AhR, CAR/PXR, and PPARA, 335 genes, 3735 genes, and 2779 genes are differentially expressed relative to the appropriate controls, respectively, at a threshold of adjusted p-value < 0.05. The top ten most differentially expressed genes for each MOA are shown in Table 2(A-C). In addition, a histogram that represents the relative frequencies of different log2FC values, split by mechanism of action, is produced. Similarly, for each mechanism of action, a scatter plot is generated. This scatter plot represents the relationship between log2FC values and nominal p-values. Both plots are shown in Figure 3(A-B). All mechanisms of action show more positive regulation than negative, as interpreted by the higher proportion of positive log2FC features vs the negative log2FC features seen in Figure 3A. In Figure 3B, samples with the AhR mechanism of action and treated with 3-Methylcholanthrene show significantly fewer counts of significantly differentially expressed genes when compared to the other mechanisms of action.

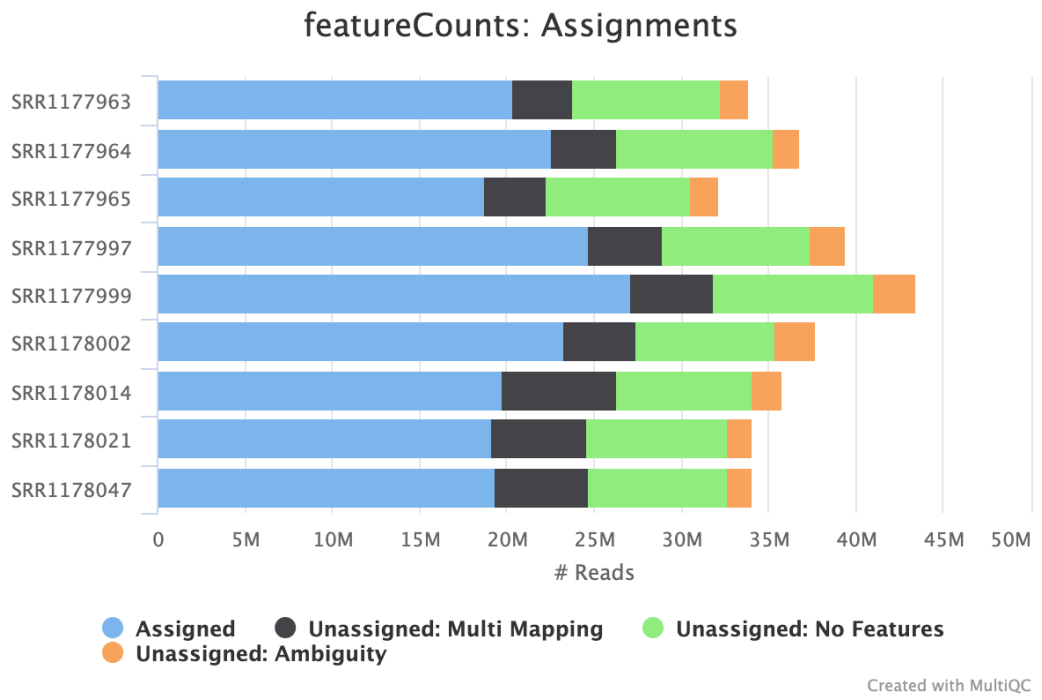


Figure 1. *featureCounts* summary of read assignments aggregated by *MultiQC*.

Sample	Percent Assigned	Percent Multi-Mapped	Percent No Features	Percent Ambiguous
SRR1177963	59.99%	10.36%	24.82%	4.83%
SRR1177964	61.32%	10.25%	24.34%	4.09%
SRR1177965	58.46%	10.81%	25.66%	5.06%
SRR1177997	62.63%	10.68%	21.51%	5.18%
SRR1177999	62.36%	10.87%	21.13%	5.64%
SRR1178002	61.85%	10.83%	21.15%	6.17%
SRR1178014	55.35%	18.14%	21.68%	4.83%
SRR1178021	56.07%	15.91%	23.72%	4.30%
SRR1178047	56.64%	15.79%	23.29%	4.29%

Table 1. Summary statistics generated by *featureCounts* and aggregated by *MultiQC*. For each sample, the percentage of assigned reads vs unassigned reads are shown.

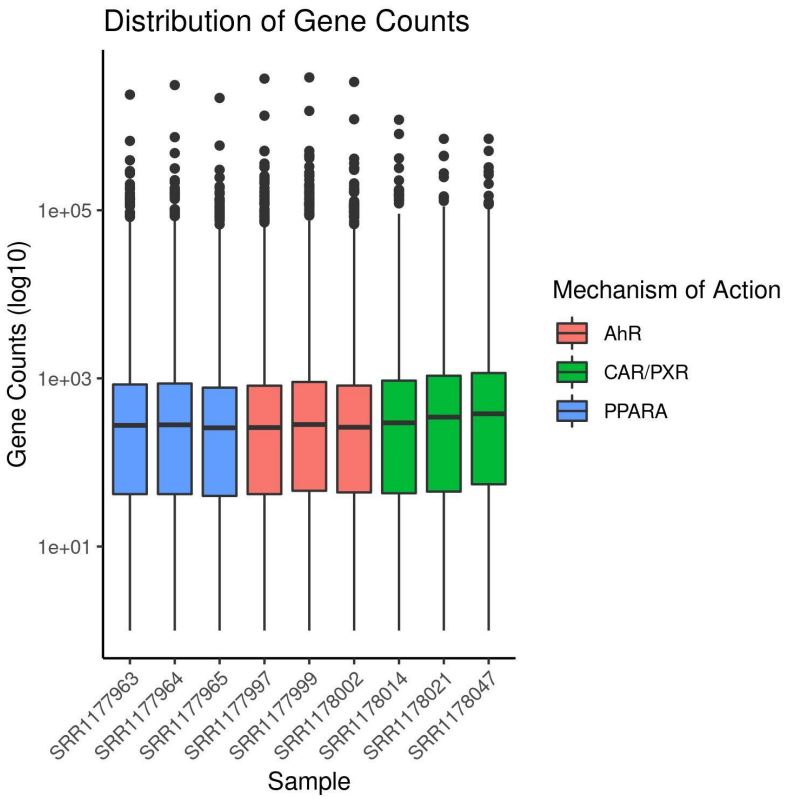


Figure 2. Box plot showing distribution of gene counts across the toxgroup 6 samples. Each color represents the mechanism of action a given sample is associated with. This plot visualizes the median, the 25th and 75th percentiles, two whiskers, and and all outlier points.

A.

MOA	Gene ID	baseMean	log2FoldChange	lfcSE	pvalue	padj
AhR	NM_012541	80326.80848	3.691320551	0.1797076536	1.07E-94	1.14E-90
AhR	NM_130407	546.3575796	3.35242723	0.2190012273	4.57E-54	2.44E-50
AhR	NM_022521	13003.77943	1.747629364	0.2442327568	3.60E-14	1.28E-10
AhR	NM_053883	1310.725154	-1.161961466	0.169852505	4.01E-13	1.07E-09
AhR	NM_012608	64.24548234	-2.86141687	0.4366668911	2.69E-12	5.75E-09
AhR	NM_023094	51.63337157	6.337793376	0.915213741	3.69E-12	6.57E-09
AhR	NM_017061	127.0197977	-2.182772949	0.3381614607	5.13E-12	7.83E-09
AhR	NM_134329	285.7950497	-1.500466159	0.2338976069	6.71E-12	8.97E-09
AhR	NM_022866	1108.166598	-2.407518406	0.3781375691	8.40E-12	9.98E-09
AhR	NM_022297	2888.890563	-0.9221880655	0.1443869649	9.35E-12	1.00E-08

B.

MOA	Gene ID	baseMean	log2FoldChange	lfcSE	pvalue	padj
CAR/PXR	NM_001108693	573.5890483	5.176186653	0.2171788066	4.81E-127	5.94E-123
CAR/PXR	NM_053699	426.5724459	6.60537783	0.2869285783	5.80E-118	3.58E-114
CAR/PXR	NM_001130558	2377.22854	-7.888374094	0.3897122106	2.22E-92	9.12E-89
CAR/PXR	NM_031605	1076.962744	3.820995323	0.2215502399	9.28E-68	2.87E-64
CAR/PXR	NM_013033	1000.389495	6.003437381	0.358289541	4.19E-64	1.04E-60
CAR/PXR	NM_144755	2060.777948	4.102081408	0.272828509	2.59E-52	5.33E-49
CAR/PXR	NM_001005384	1522.844313	3.913407807	0.2688015772	3.13E-49	5.53E-46
CAR/PXR	NM_031048	6984.513451	4.221267446	0.3020639241	1.38E-45	2.13E-42
CAR/PXR	NM_013105	300199.8119	4.856779816	0.3509263396	8.77E-45	1.20E-41
CAR/PXR	NM_001014166	1614.185995	-2.908112407	0.2188036986	1.46E-41	1.80E-38

C.

MOA	Gene ID	baseMean	log2FoldChange	lfcSE	pvalue	padj
PPARA	NM_024162	2298.61223	7.836003425	0.2612374124	1.36E-197	1.56E-193
PPARA	NM_012737	2516.803823	-6.722575355	0.2403061485	3.50E-173	2.00E-169
PPARA	NM_131903	9516.988733	-5.133523985	0.2484947444	4.15E-96	1.58E-92
PPARA	NM_017158	2723.415934	-6.163623093	0.3028134785	2.72E-93	7.79E-90
PPARA	NM_019157	439.4460982	5.989476059	0.2947176505	3.28E-92	7.50E-89
PPARA	NM_053883	994.2520835	-3.059788426	0.1591780358	1.47E-83	2.80E-80
PPARA	NM_001014063	2844.494836	-4.182102957	0.2241666453	6.23E-79	1.02E-75
PPARA	NM_012600	13167.95193	3.83363074	0.2053802867	1.01E-78	1.45E-75
PPARA	NM_001013098	2748.692917	-5.393707723	0.2981973443	1.97E-74	2.51E-71
PPARA	NM_001013975	954.7396869	-3.83570396	0.2172305012	5.32E-71	6.09E-68

Table 2. The top 10 most differentially expressed genes in the **(A.)** AhR **(B.)** CAR/PXR, and **(C.)** PPARA mechanism of action.

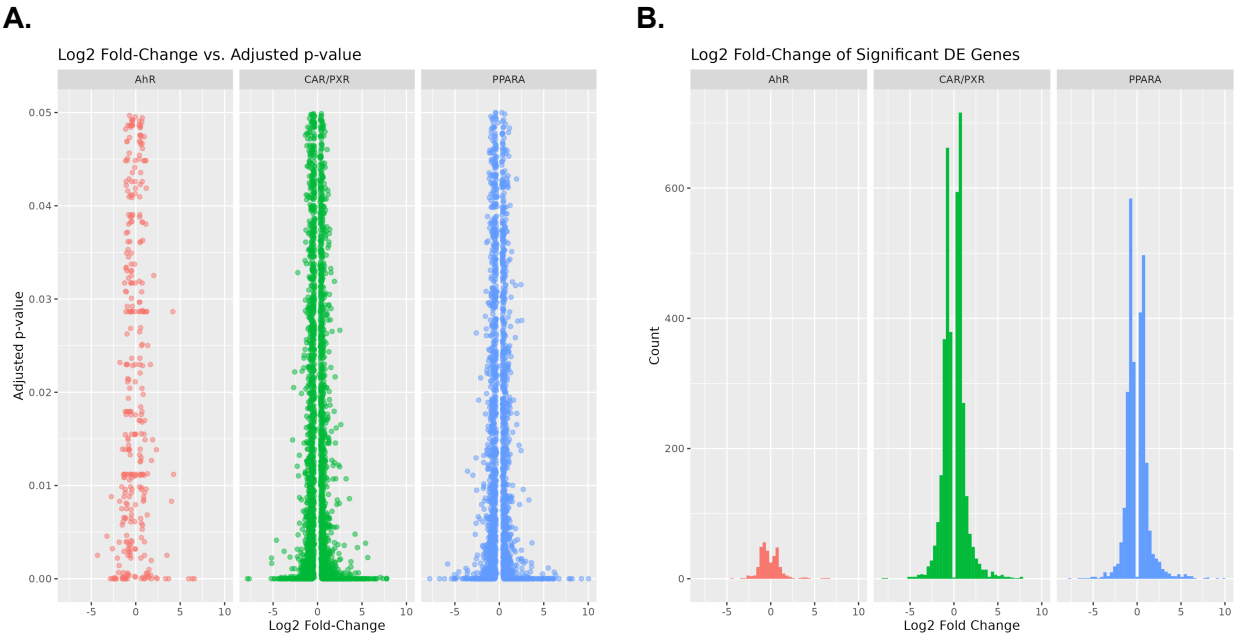


Figure 3. Plots of significant differentially expressed genes as determined by *DESeq2*. **A.** Histogram of fold change values for each sample, split by mechanism of action. **B.** Scatter plot of log2 fold-change vs p-value from the significant differentially expressed genes determined via *DESeq2* analysis, split by mechanism of action.

Discussion

The analyses performed in this report are done so with the goals of preparing data to be run through differential expression analysis, identifying the quality of that data, and performing the differential expression analysis itself. No glaring issues are found during the *featureCounts* analysis, and no one sample appears to be overrepresented in the data set. Under the mechanisms of action of AhR, CAR/PXR, and PPARA, 335 genes, 3735 genes, and 2779 genes are differentially expressed relative to the appropriate controls, respectively. All mechanisms of action show more positive regulation than negative, and samples with the AhR mechanism of action show significantly fewer counts of significantly differentially expressed genes when compared to the other mechanisms of action.

References

1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.
2. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

3. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13. PMID: 24227677.
4. Love MI, Huber W and Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, pp. 550. http://dx.doi.org/10.1186/s13059-014-0550-8
5. Martin Morgan (2021). BiocManager: Access the Bioconductor Project Package Repository. R package version 1.30.16. https://CRAN.R-project.org/package=BiocManager
6. Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, 1 October 2016, Pages 3047–3048, https://doi.org/10.1093/bioinformatics/btw354
7. Wang, Charles et al. “The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance.” Nature biotechnology vol. 32,9 (2014): 926-32. doi:10.1038/nbt.3001