Concordance of Microarray and RNA-seq Gene Expression

David Lenci, Nikita Tomar, and Daniel Gealow

Introduction

RNA sequencing (RNA-seq) has grown as a widely used technology in research that allows for unbiased genome wide expression analysis. Prior to the establishment of RNA-seq, the principle technology for performing transcription analysis was microarrays, which allowed for a large amount of biological discovery and research [2]. Due to the growth of RNA-seq as a popular research method many have begun directly comparing it to microarrays by performing the same sequencing analyses using both technologies. These endeavors have yielded varying results with some establishing that RNA-seq exhibits lower precision for weekly expressed genes, and others determining that RNA-seq is in fact more sensitive for gene detection [3][4]. The researchers in Wang et al believe this is due to the fact that these studies utilized few treatment methods, and therefore did not cover a wide enough range of biological complexity [5].

In order to obtain a proper comparison between the two technologies with the needed coverage of biological systems, Wang et al, under the umbrella of the third phase of the Sequencing Quality Control (SEQC) project, conducted a complete comparison of RNA-seq and microarray technology. Specifically, their project's aim was to compare the two technologies through the analysis of gene expression from rat liver tissue exposed to various chemicals [5]. This project aims to replicate some of the analysis results from Wang et al.

Data

Three male Sprague-Dawley rats were exposed to one of 27 chemicals, RNA was isolated from these mice and analyzed with microarray chips (Affymetrix) and RNA-seq. Our analysis will only look at three of these chemicals for a total of 9 mice. The microarray data was from the Affymetrix GeneChip® Rat Genome 230 2.0 array[1]. The RNA reads are paired ends and input reads range from 15559784 to 19627402 with an average of 17293468 reads. Each single read is around 100 bp except for one sample which has 50 bp for each read in the pair. Microarray and RNA-seq data are from NCBI services and databases from accession SRP024314. For our analysis, samples from toxgroup 3 were chosen, representing Leflunomide, Fluconazole, and the relevant controls. Sample IDs for the RNA-seq data were extracted from the toxgroup 3 rna data table, representing the experimental and control samples. FastQC (version 0.11.7) was run on files with these ids. Next, an alignment to the provided genome was performed using STAR (version 2.6.0c) to generate BAM files for the nine experimental samples. MultiQC (version 1.10.1) was performed on fastq files and STAR alignments.

Methods

Quality Control and Sample Alignment

FastQC results showed an average of 16 million reads per sample out of which roughly 50% are duplicates. All samples had total reads in comparable ranges with no samples having too many or too less read counts. All sequences maintained a Phred score of approximately 30-35 at least through 70 bp. FastQC reports red flagged 15 of these sequences for poor mean quality scores, but because the low Phred score began late in the read and STAR aligner takes into account the Phred score, reads were not trimmed before alignment. FastQC red flagged SRR1177981 and SRR1177983 as overrepresented and red flagged 13 samples for high sequence

duplication levels but these flags were ignored because RNA-Seq gives non-homogenous sequencing results which are tied to differences in expression and hence the differences in RNA template is observed. STAR alignment QC showed that more than 80% of all reads were mapped to the genome for each sample with read counts ranging from 11-16 million across the samples. These statistical values imply that our samples are of high quality RNA-Seq reads. These samples were then used for further analysis.

Figure 1. MultiQC Report. Samples failed Mean Quality Scores (a), and Sequence Duplication Levels (b) but passed Per Sequence Quality Scores (c), Per Base N Content (d),Adapter Content from MultiQC Report (e),And Per sequence GC content (Other metrics from MultiQC report had varied amounts of failed, warning, and passed samples).

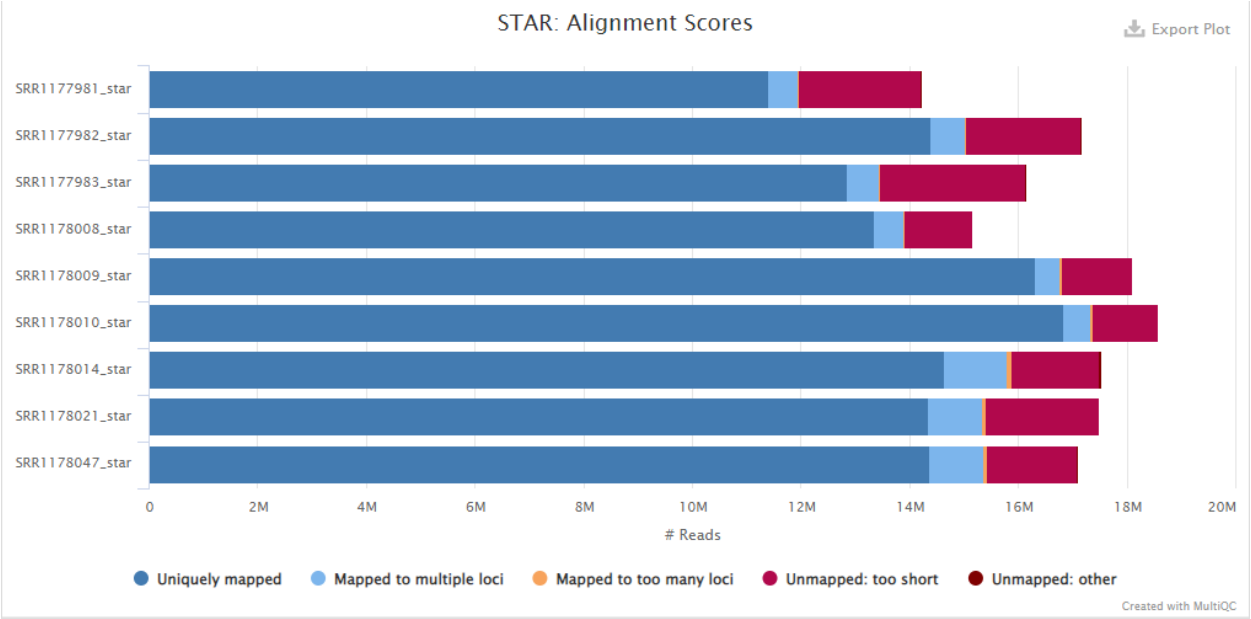


Figure 2: STAR alignment scores for each of nine sample alignments by number of reads

FeatureCounts

With the generated bam files for the RNA-seq data, we then utilized the featureCounts package in order to generate count files to be analyzed later on. FeatureCounts was run with options “-T 16”, or 16 threads, and with an RNA annotation file in gtf format provided to us. This generated a counts file for each bam file as well as a summary output file. MutliQC was then run on the output counts and summary files in order to assess the quality of our data. From the MultiQC we found that all nine samples hovered around 60% alignment, and the distribution of counts for each sample can be seen in supplemental figure 1. Additionally, there appeared to be no significant outliers in our nine samples.

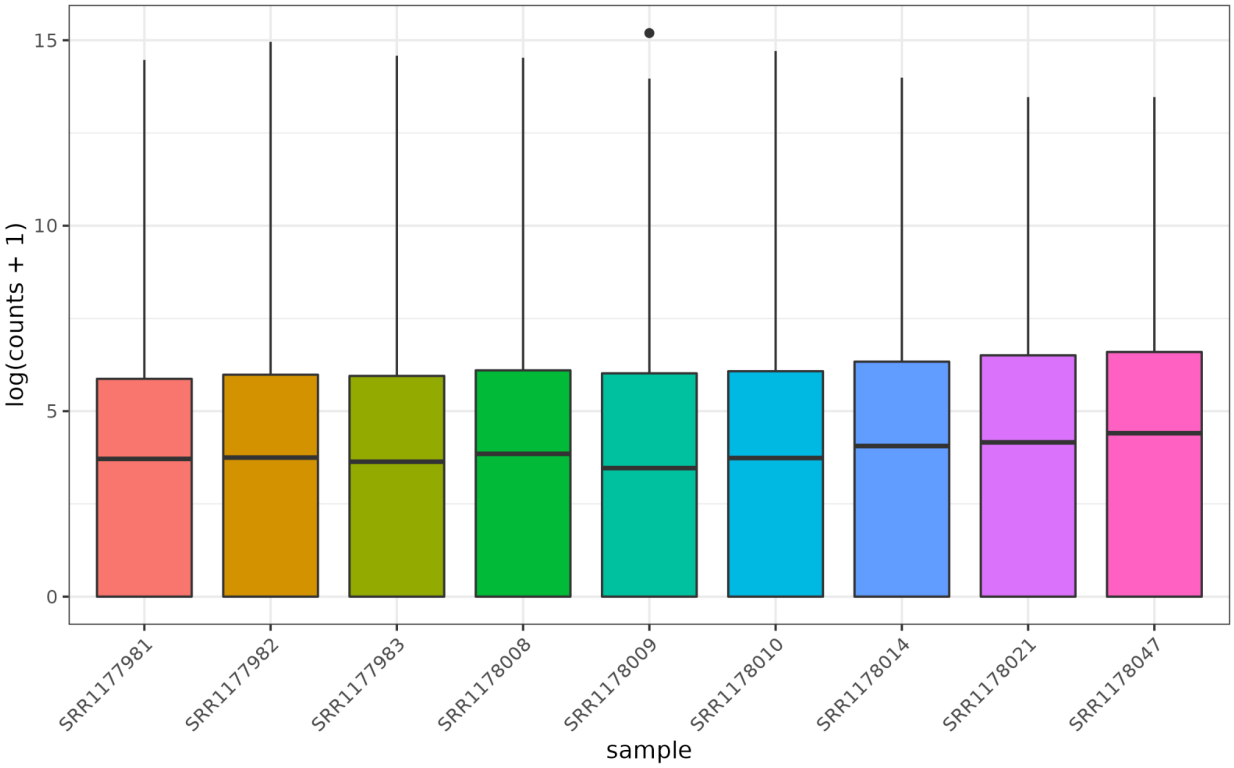


Figure 3 Distribution of Counts: This figure shows the distribution of log(counts +1) for each sample generated by featureCounts.

The count files were then combined into a single csv file, to allow for easy comparison of the different samples. Additionally, figure 3 was generated to demonstrate the distribution of log(counts) across each of the samples, and we can see that the distribution of counts is consistent across all samples.

Results

RNA Sequencing DE Results

Top 10 Differentially Expressed Genes from DESeq2

Ranked by Significance (P-Value)

AhR vs Control[1]			CAR/PXR vs Control[2]			DNA Damage vs Control[3]		
Genes	P-Value	Log2(FC)	Genes	P-Value	Log2(FC)	Genes	P-Value	Log2(FC)
NM_013096	6.941e-61	-9.92	NM_053288	1.449e-136	4.8	NM_033234	3.234e-63	-7.01
NM_033234	1.082e-58	-10.1	NM_001108693	5.238e-118	5.65	NM_001007722	6.742e-62	-6.88
NM_001007722	3.146e-48	-9.22	NM_001130558	2.706e-90	-6.64	NM_198776	1.048e-35	-0.00684
NM_001257095	2.243e-44	-4.56	NM_001134844	9.464e-87	6.93	NM_001111269	2.288e-31	-7.14

Table 1: Top 10 DE Genes from DESeq2: This table shows the top 10 differentially expressed genes generated from DESeq2 ranked by nominal p-value for each condition in the data analyzed.

With the RNA-seq data fully organized into a single csv file we then moved to differential expression (DE) analysis utilizing the DESeq2 R package. First, following the DESeq2 common practices we removed all rows with less than 10 total counts across all samples. Then, with our filtered counts we performed DE analysis for each group in our data. Since we are only looking at a subset of the total data analyzed in Wang et al, we only have four distinct groups: Control, AhR, CAR/PXR, and DNA damaged. Additionally, in our control and sample groups of data there were two subgroups depending on the “vehicle” utilized to generate the data; corn oil and saline. In order to perform DE analysis we compared each sample group to a control group where the “vehicle” used was the same for the sample and control.

Doing this, we generated three different DESeq2 result tables, and then selected for significantly differentially expressed genes with a p-adjusted of < 0.5. From this we found that the AhR group had 1477 genes, CAR/PXR had 3516 genes, and DNA Damage had 119 genes that were significantly differentially expressed. We then pulled out the top 10 genes ranked by p-value as seen in table 1. Additionally we generated histogram and scatter plots to demonstrate the distribution of DE results for each sample group, which can be seen in Figure 5 below.

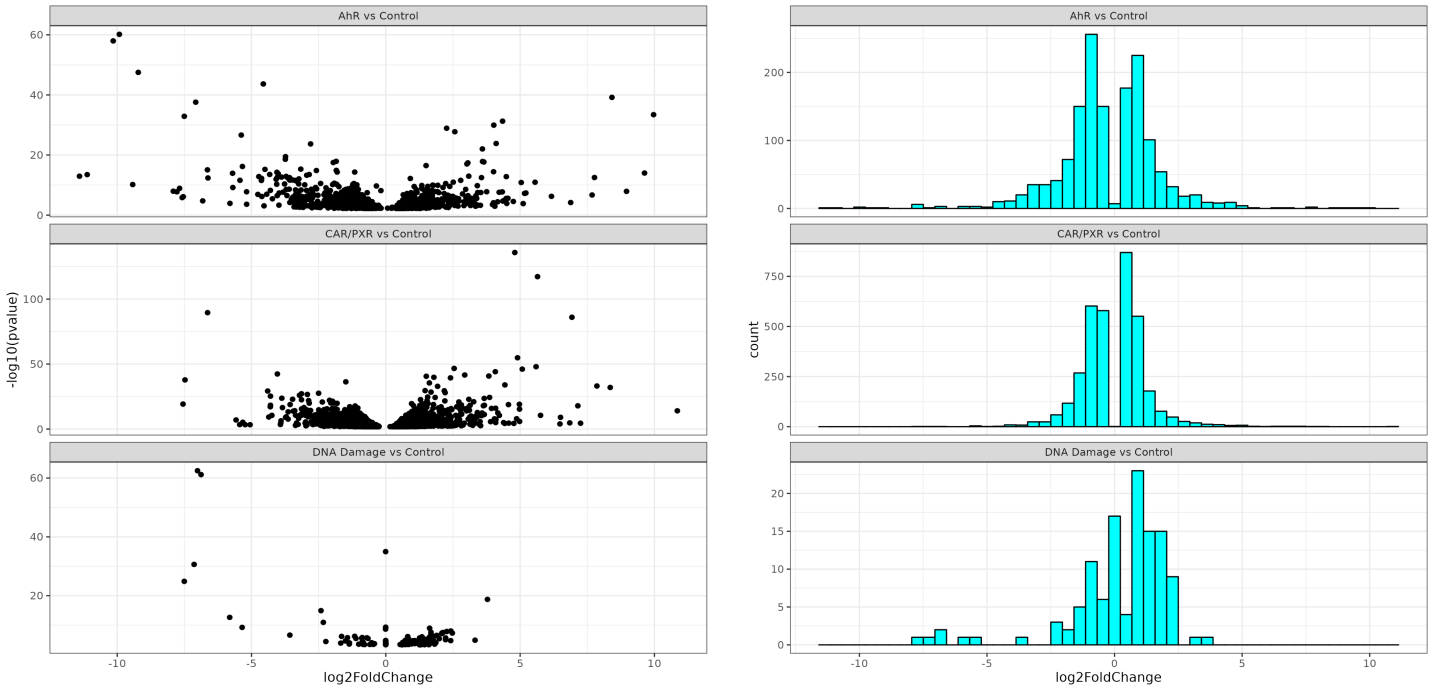


Figure 4: Left figure shows -log10(p-value) vs log2-fold change, and the right figure shows the distribution of counts across our samples. Each row represents one of our outputs from DESeq2, with the top being AhR vs control, middle being CAR/PXR vs control, and bottom being DNA Damaged vs control.

Microarray DE Results

Next, we performed a differential expression analysis of the Affymetrix microarray data, using the limma package in R. We used the annotate and rat2032.db packages from Bioconductor to convert the Affymetrix Rat Genome 230 2.0 Array probe names into gene symbols. For each drug, we created a histogram of log-transformed fold change, and a scatterplot of the fold-change and p-values of significantly differentially expressed genes. The 10 most differentially expressed genes for each drug were as follows:

FLUCONAZOLE (CAR/PXR)

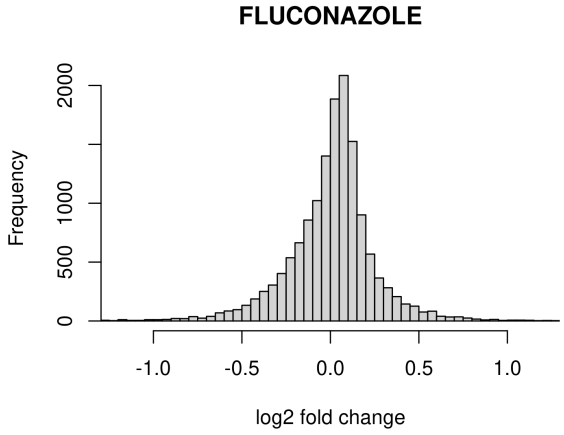
	symbol	probe	logFC	padj
0	Orm1	1368731_at	1.325853	1.618185e-10
1	Cxcl1	1387316_at	2.785198	1.639096e-08
2	Stac3	1395403_at	-3.580996	1.684088e-08
3	Rgs3	1367957_at	2.018340	4.682045e-08
4	C2cd2	1390165_at	-0.937787	4.682045e-08
5	Gdf15	1370153_at	2.209461	1.092788e-07
6	Ablim3	1391570_at	1.542911	1.366638e-07

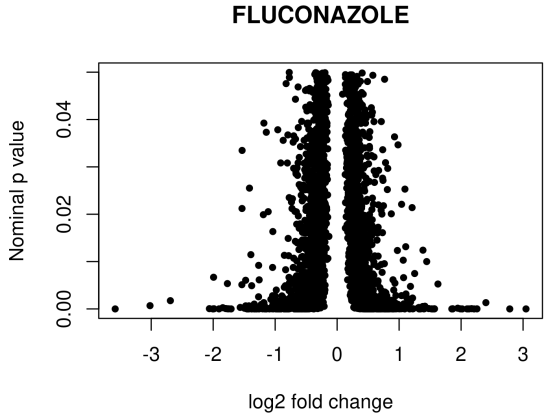
```
7  Nectin3  1378027_at -0.783955  3.729952e-07
8      Id4  1394022_at -1.336862  3.864863e-07
9  Mybbp1a  1387779_at  0.662647  4.706798e-07

IFOSFAMIDE (DNA damage)
      symbol       probe     logFC      padj
0    Adamts9  1376481_at -0.640727  0.005448
1       Gulo  1387725_at -0.569992  0.013365
2       Fmo5  1383248_at  0.579198  0.013365
3      Socs3  1377092_at  0.886421  0.013365
4      Odf3b  1375775_at -0.712282  0.013365
5    Plekha6  1390271_at  0.335791  0.013365
6      Hmox1  1370080_at  0.645470  0.013365
7     S100a9  1387125_at -0.757299  0.014970
8  LOC684270  1372996_at  0.280365  0.016251
9       Pklr  1368651_at -0.891421  0.016415

LEFLUNOMIDE (AhR)
   symbol       probe     logFC          padj
0  Cyp1a1  1370269_at  7.440868  4.745228e-14
1  Cyp1a2  1387243_at  1.383740  4.462421e-12
2   Il1r1  1392946_at  1.596782  1.179146e-09
3   Tcea3  1388611_at -0.688520  2.949004e-08
4    Eml4  1376827_at  0.858660  3.857063e-08
5  Fbxo31  1372600_at  1.156193  1.406275e-07
6    Ctsl  1370244_at  0.550195  2.470171e-07
7   Stac3  1395403_at -3.064409  4.498964e-07
8  R3hdm2  1373814_at -0.873889  5.600699e-07
9    Glrx  1367705_at  1.044114  5.600699e-07
```

Table 2. Most differentially expressed probes and corresponding genes for each treatment.





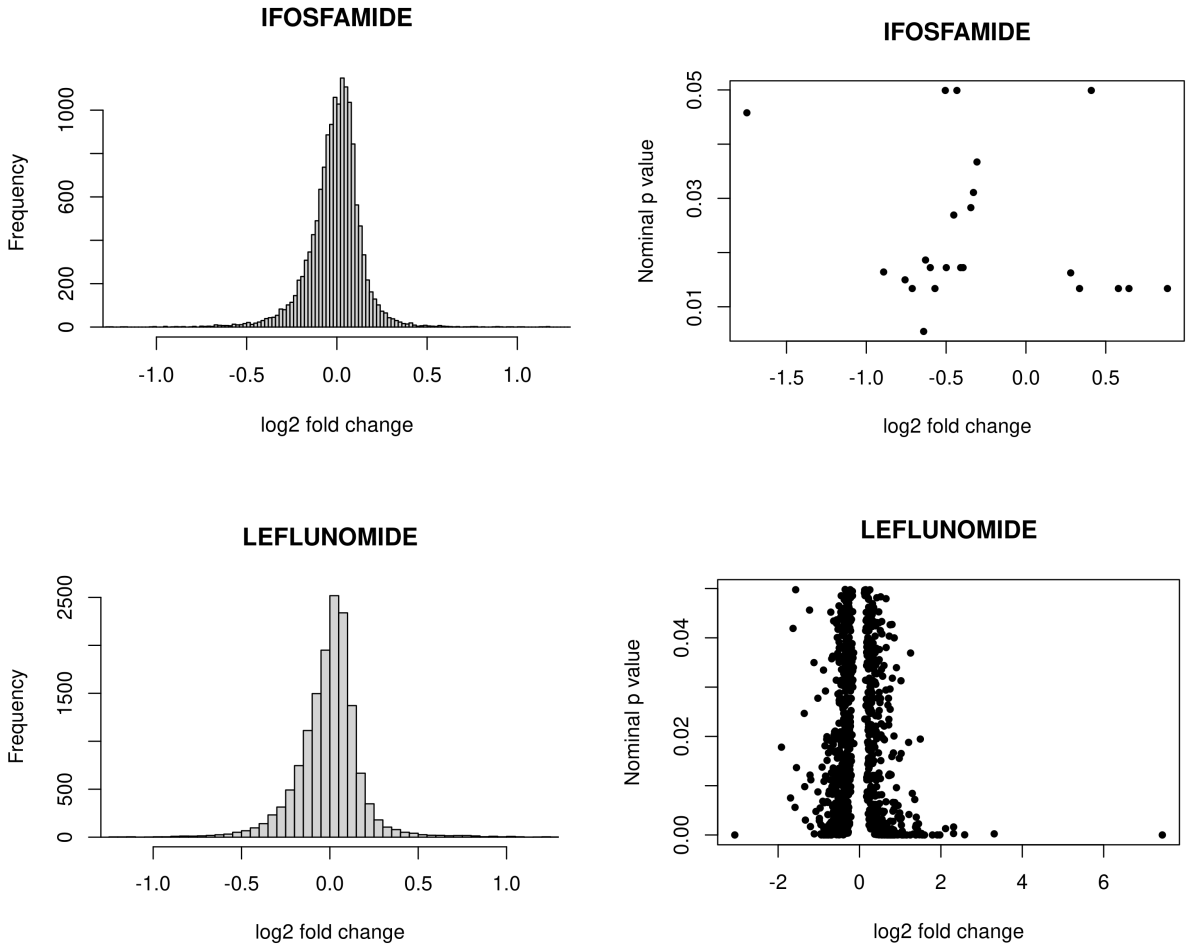


Figure 5. Left: Histograms of log-transformed fold change distributions of all genes. Right: Plots of BH-adjusted p-value vs. logFC. The main body of the histograms are approximately normal, but there are some outlying points that greatly changed in expression visible on the scatterplots.

As with DESeq, histograms and scatter plots were created to display the distribution of expression changes, shown in Figure 5.

Finally, we performed an analysis of concordance between the two methods of measuring gene expression. We converted the reference sequence IDs to gene symbols using the provided table. Since the set of genes detected by RNA-seq and those detected by the microarray were not identical, we restricted our analysis to genes in the intersection of these sets. The number of genes that were differentially expressed according to both DESeq and Limma was counted and

corrected based on the overlap that would be expected by chance, and concordance was calculated as the size this “true” overlap divided by the average of the number of differentially expressed genes detected by each of the two methods. (A concordance of 1.0 indicates complete overlap; a concordance of zero indicates exactly the expected overlap for random sets; a negative concordance indicates a smaller overlap than expected at random.) This analysis was repeated thrice for each drug, with the full set of genes common to the two methods (“all”), the subset of genes with above-median absolute expression (“high”), and the subset with below-median absolute expression (“low”) according to DESeq. The results are given in Table 3 and depicted visually in the discussion section below.

Drug	subset	deseq	limma	concordance
FLUCONAZOLE	all	3019	2340	0.45403
	high	2065	1594	0.563064
	low	954	746	0.369625
IFOSFAMIDE	all	97	20	0.014125
	high	42	14	0.033912
	low	55	6	-0.001025
LEFLUNOMIDE	all	1255	786	0.225527
	high	715	523	0.296804
	low	540	263	0.188674

Table 3. Numerical results of concordance analysis. For each analysis, “deseq” is the number of genes differentially expressed in RNA-seq, “limma” is the number D.E. in the microarray, and concordance is calculated as described above.

Discussion

We were interested in whether there was a correlation between the number of D.E. genes and the concordance between the two methods of measuring it. Figures 6 and 7 plot this relationship across all nine analyses we performed.

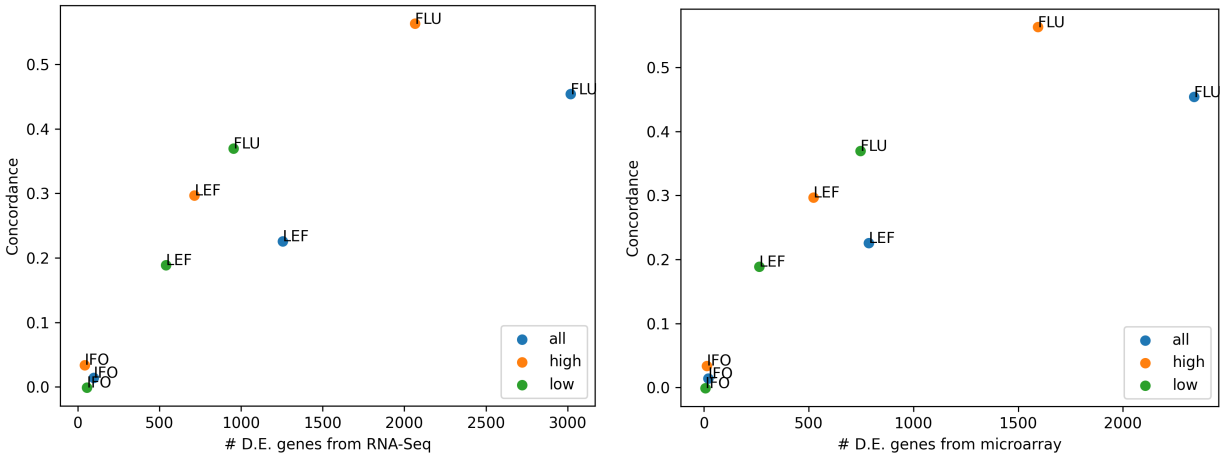


Figure 6. Plots of concordance against number of differentially expressed genes in RNA-seq (left) and microarray (right).

In either method of counting, there is a clear positive relationship between concordance and number of D.E. genes, confirming our hypothesis. This makes sense, because if a drug does not cause significant changes to gene expression, then most genes that appear to be D.E. will just be chance variations. It also appears that the concordance is relatively higher at smaller numbers of genes for either subset than for the full set of genes.

We were also interested to see which drugs had the best concurrence, and whether there was more agreement between methods for strongly or weakly expressed genes. The results from all analyses are displayed in the bar chart in figure 7.

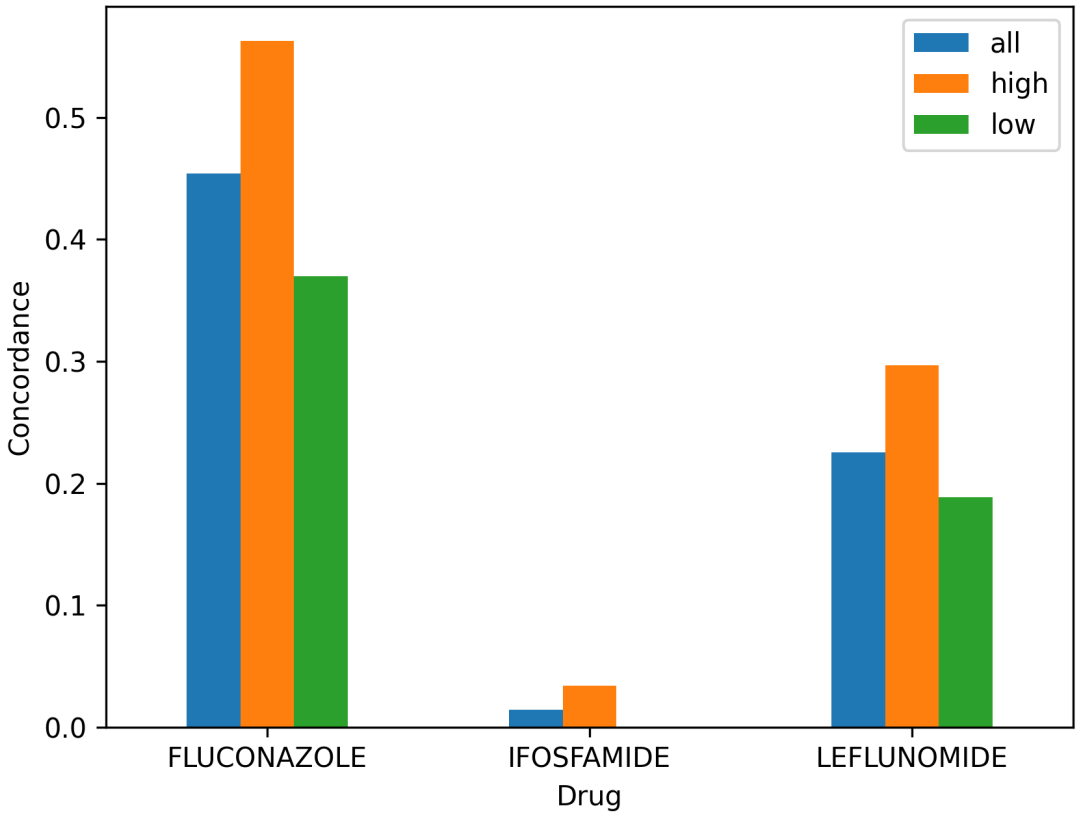


Figure 7. Concordance in full and subset analyses for each of the three drugs tested.

It is clear upon inspection that there was the most concordance between the methods on the subset of highly-expressed genes. Also, Fluconazole had relatively high concordance in all the subsets, while Ifosfamide had very low concordance (even negative for the weakly-expressed subset.)

Conclusion

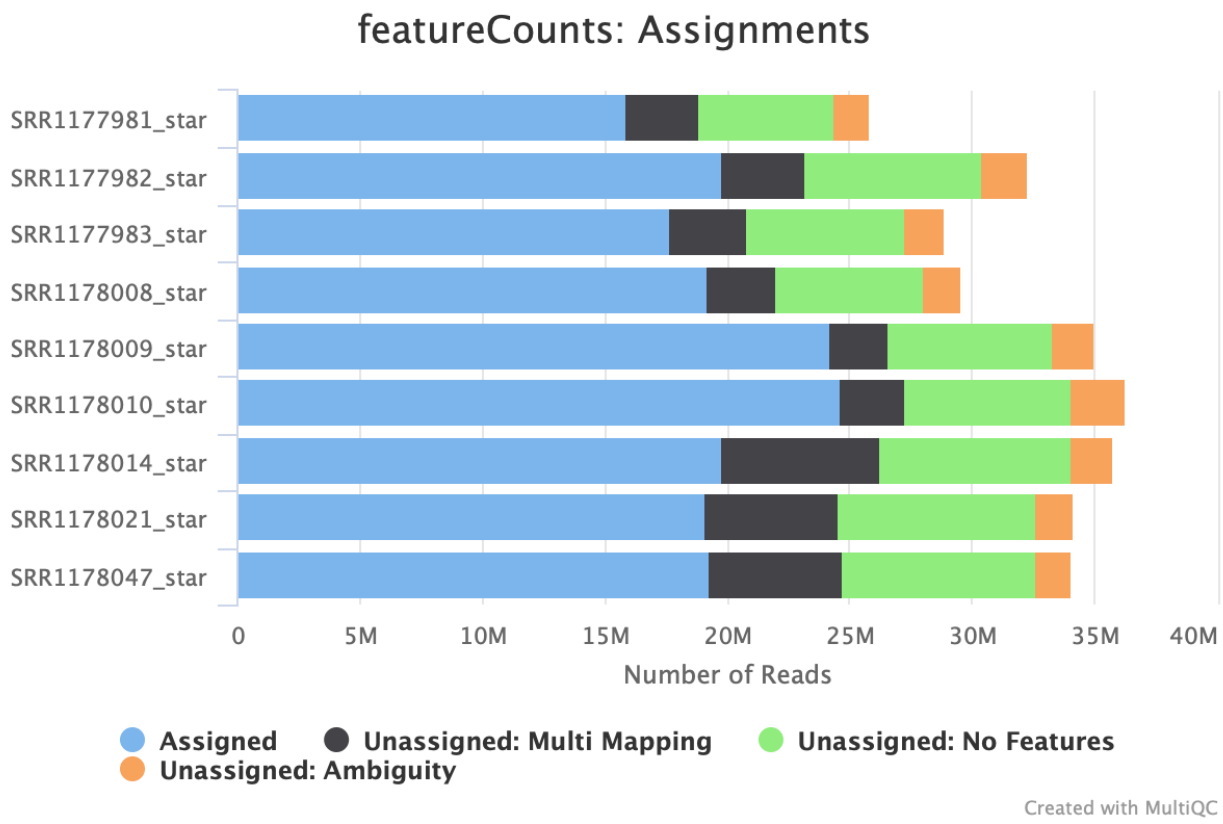
We were able to replicate the paper's findings that concordance was higher when the effect was greater. We were also able to replicate the finding that concordance was higher when considering a subset of genes with higher absolute expression.

A major challenge of this analysis was conversion of probes and reference sequences to standard gene identifiers. Many were ambiguous, or entirely absent from the mapping tables we used. Consequently, the total set of genes detected by each method was not the same–about a third of the genes had to be excluded since they were only measured by one method but not the other. We suspect this may have deflated concordance–if both methods were measuring the same genes, there would be more opportunity for overlap.

References

1. 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/
2. Chen, Minjun, et al. “A Decade of Toxicogenomic Research and Its Contribution to Toxicological Science.” *Toxicological Sciences: An Official Journal of the Society of Toxicology*, vol. 130, no. 2, Dec. 2012, pp. 217–28. *PubMed*, https://doi.org/10.1093/toxsci/kfs223.
3. Łabaj, Paweł P., et al. “Characterization and Improvement of RNA-Seq Precision in Quantitative Transcript Expression Profiling.” *Bioinformatics*, vol. 27, no. 13, July 2011, pp. i383–91. *PubMed Central*, https://doi.org/10.1093/bioinformatics/btr247.
4. Mooney, Marie, et al. “Comparative RNA-Seq and Microarray Analysis of Gene Expression Changes in B-Cell Lymphomas of Canis Familiaris.” *PLoS ONE*, vol. 8, no. 4, Apr. 2013, p. e61088. *PubMed Central*, https://doi.org/10.1371/journal.pone.0061088.
5. Wang, Charles, et al. “A Comprehensive Study Design Reveals Treatment- and Transcript Abundance–Dependent Concordance between RNA-Seq and Microarray Data.” *Nature Biotechnology*, vol. 32, no. 9, Sept. 2014, pp. 926–32. *PubMed Central*, https://doi.org/10.1038/nbt.3001.

Supplemental Figures:



Supp. Figure 1, MultiQC featureCounts: This figure shows us the number of reads for each of our samples, as well as where those reads are distributed across four groups: assigned, unassigned: multi mapping, unassigned: no features, and unassigned: ambiguity. As we can see, for all samples a large majority of our reads were successfully assigned by featureCounts.