Concordance of microarray and RNA-Seq differential gene expression

Chris Lin, Yuehting Wang, and Cody Webb

Introduction

In Wang et al.[1] , the authors investigated the concordance in genome-wide expression profiling between RNA-sequencing and microarrays. This was done by generating both Illumina RNA-seq and Affymetrix microarray data from the same set of liver samples of rats treated with 27 chemicals with known modes of action. The authors found that generally, concordance between RNA-seq and microarrays was determined by treatment effect size, gene-expression abundance, and the biological complexity of the drug modes of action. In this project, we sought to reproduce their findings using similar methods as in Wang et al. First, the publically available RNA-seq data from the authors was acquired. As the microarray data was already processed for us, quality control and alignment to the rat genome was only performed on the RNA-seq data. Differential expression analysis was then performed on both the RNA-seq and pre-normalized microarray expression data. Lastly, concordance of differentially expressed genes (DEGs) between platforms was calculated in an effort to determine if our findings were similar to those of Wang et al.

Data

Male Sprague-Dawley rats were administered one of 27 chemicals orally (three rats per chemical with matching number of controls). RNA was then isolated from liver tissue, and analyzed using both Illumina RNA-seq and Affymetrix microarrays.

RNA-sequencing was performed with the Illumina HiScanSQ or HiSeq2000 systems according to manufacturer's protocol using the Illumina TruSeq RNA Sample Preparation Kit and SBS Kit v3 (San Diego, CA). mRNA of 118 samples were sequenced at depths of around 23–25 million paired-end 100 bp reads. RNA-seq data was deposited in the Sequence Read Archive (NCBI) under ascension number SRP024314.

Fragmented cDNA from liver RNA was hybridized to the Affymetrix whole genome GeneChip® Rat Genome 230 2.0 Array according to protocols outlined in the Affymetrix GeneChip® Expression Analysis Technical Manual (subsection: Eukaryotic Target Hybridization) using the GeneChip® Hybridization, Wash, and Stain Kit (P/N 900720). Scanning of arrays was performed on the GeneChip® Scanner 3000 7G, and normalized with GeneSpring GX 11.5.1 (Agilent Technologies

Santa Clara, CA). All normalized data can be found in the Gene Expression Omnibus[2] (NCBI) under the accession number GSE47792.

For this project, we focused on the analysis of a single subset of samples, termed "Toxgroup 3", consisting of three drug treatments over three replicates each and six controls. All reads consisted of paired-reads with a length of 100 or 101 base pairs, with the exception of SRR1178014, which had a length of 50 base pairs (Table 1).

Methods

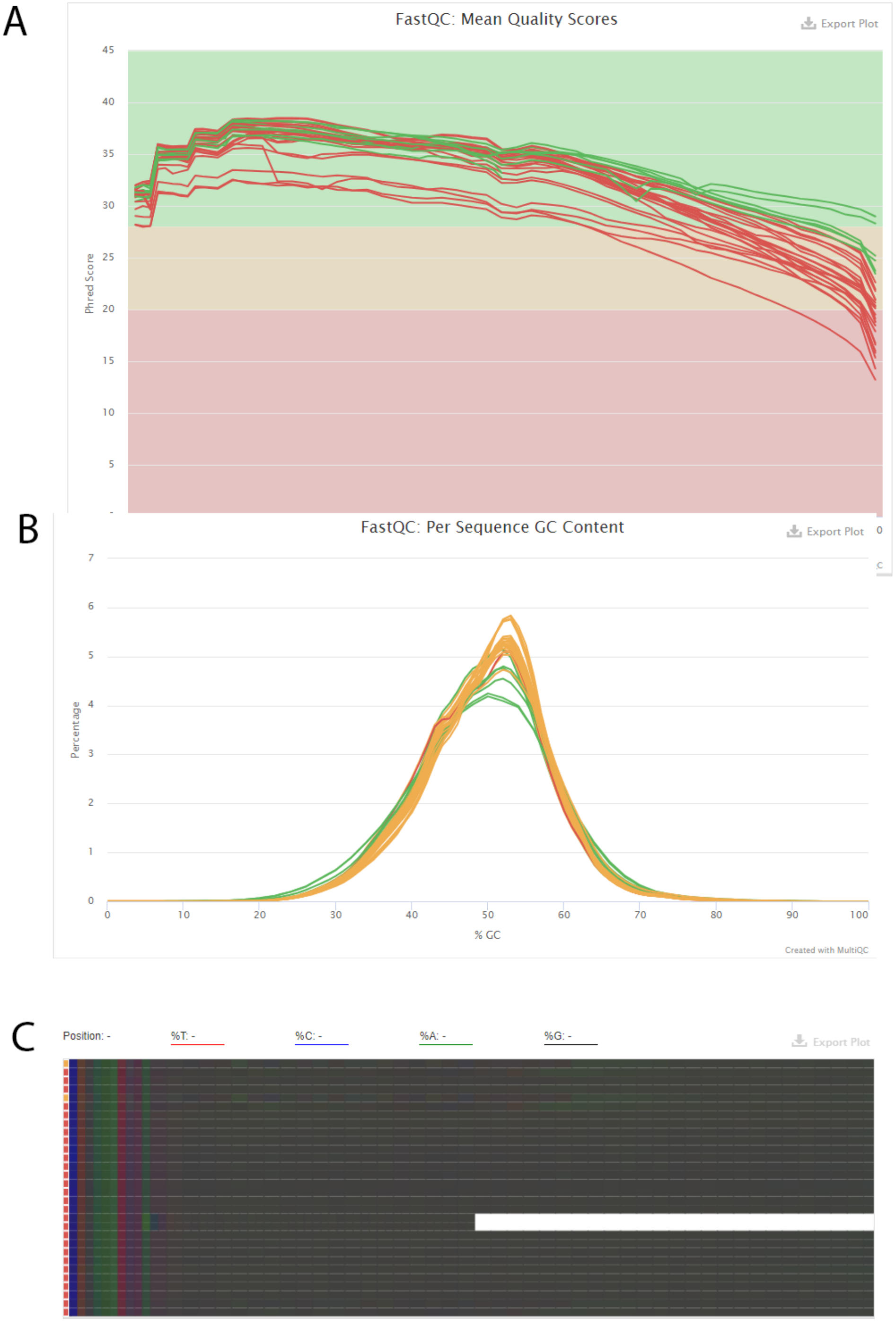
In order to compare differential gene expression of RNA-seq to microarrays, the RNA-seq data first needs to be aligned to the rat genome. This allows for quantification of which genes are being differentially expressed in each method. Corresponding paired-end samples were aligned to the rat genome using STAR[3] (Table 2). After alignment, MultiQC[4] was used on previously generated FastQC results[5] and the reports generated from STAR to determine the quality of the reads.

The MultiQC report of samples from Toxgroup 3 showed warnings or failures in the categories of *Mean Quality Scores*, *Per Sequence GC Content*, *Per Base Sequence Content*, *Sequence Duplication Levels*, and *Overrepresented sequences*. Phred scores of 22 of the 30 samples dropped below 25. This only occurred in the final 10 base pairs of each sample (Fig 1A), which is expected due to phasing errors from Illumina sequencing. 23 samples showed a slightly shifted distribution that did not match the expected normal distribution in regards to *Per Sequence GC Content* (Fig 1B). Twenty-eight sequences failed the *Per Base Sequence Content*, while two sequences flagged a warning (Fig 1C). However, the majority of non-uniform nucleotide distribution only occurred within the first 10-15 bases, which is not unusual given the type of library kit that was used. Lastly, while the majority of our sequences were unique (Fig 1D, 1E), some sequence duplication was seen in the levels higher than nine. However, this is to be expected in RNA-seq data for genes that are highly expressed. As a result, no samples from Toxgroup 3 were omitted.

Table 1. STAR alignment scores for each chemical treatment and controls in Toxgroup 3.

Sample Name	Chemical Treatment	Uniquely mapped	Mapped to multiple loci	Mapped to too many loci	Unmapped : too short	Unmapped : other	% Aligned
SRR1177981	Ifosafmide	11,486,014	546,933	7,726	2,175,870	12,808	91.3%
SRR1177982	Ifosafmide	14,502,948	644,332	12,668	1,993,281	15,452	90.7%
SRR1177983	Ifosafmide	12,964,436	581,265	10,529	2,581,512	14,539	89.8%
SRR1178008	Leflunomide	13,418,052	546,382	24,900	1,159,162	7,576	88.5%
SRR1178009	Leflunomide	16,419,670	468,098	14,466	1,201,035	7,235	88.4%
SRR1178010	Leflunomide	16,956,547	515,772	14,275	1,078,500	7,425	87.9%

SRR1178014	Fluconazole	14,645,914	1,194,171	41,502	1,606,368	36,827	84.5%
SRR1178021	Fluconazole	14,467,021	1,042,096	30,913	1,947,397	10,498	84.5%
SRR1178047	Fluconazole	14,447,020	1,039,926	42,556	1,553,546	10,254	83.6%
SRR1178050	Control	13,785,744	625,248	16,210	1,641,729	17,688	82.7%
SRR1178061	Control	56,357,222	2,585,915	64,976	4,189,626	69,616	80.7%
SRR1178063	Control	39,137,119	1,586,584	32,401	3,750,736	26,727	80.3%
SRR1178004	Control	17,339,919	853,610	28,834	1,352,638	47,048	89.3%
SRR1178006	Control	19,191,606	850,597	24,198	1,392,277	30,080	89.1%
SRR1178013	Control	14,459,638	578,672	14,462	1,025,762	16,078	85.7%



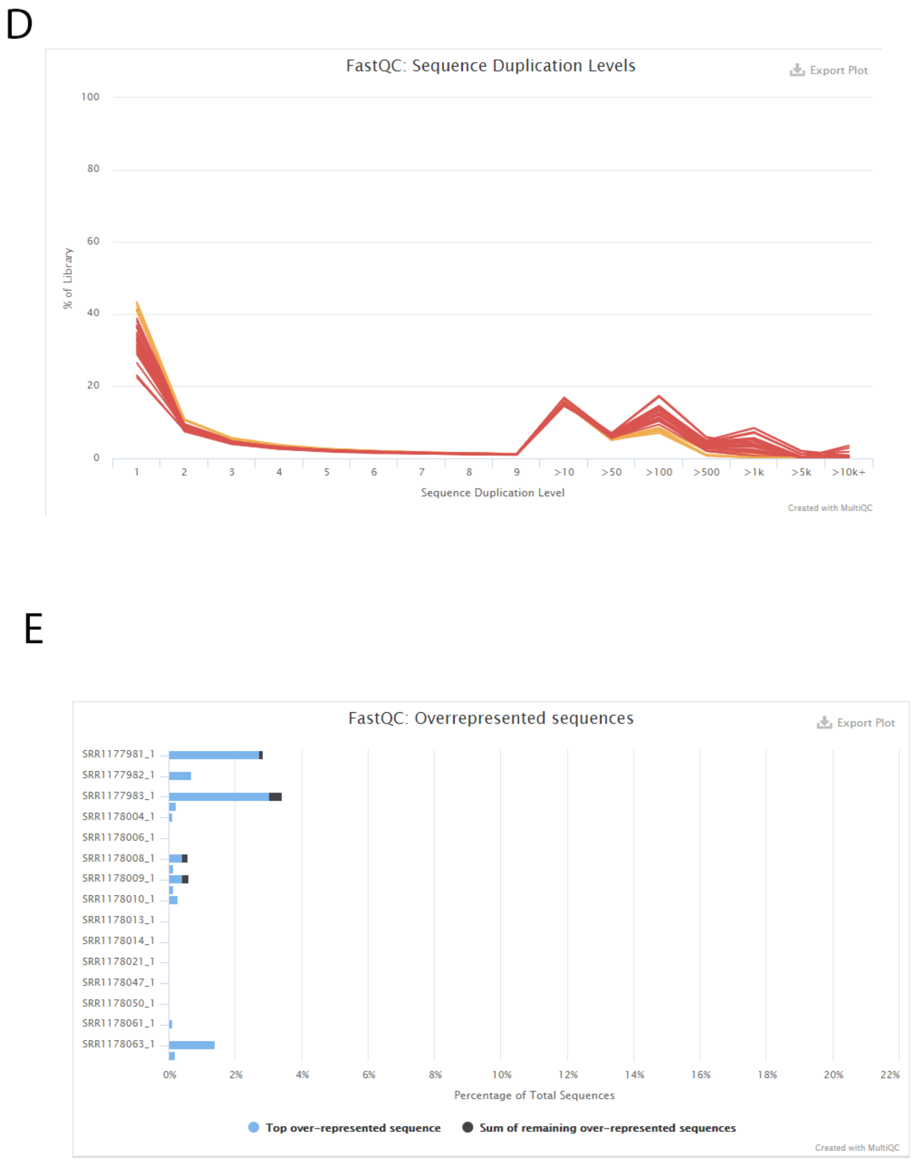


Figure 1. MultiQC output categories from Toxgroup 3 that included warnings or failures.

(A) Mean quality scores from MultiQC output of treatment and control samples. Scores are generally above 25 until the last 10 base pairs.

(B) Per sequence GC content from MultiQC output of treatment and control samples. Twenty-three samples did not match the expected normal distribution.

(C) Per base sequence content from MultiQC output of treatment and control samples. The majority of non-uniform nucleotide distribution occurs within the first 10 to 15 base pairs.

(D) Sequence duplication levels from MultiQC output of treatment and control samples. Some sequence duplication is observed in levels greater than nine.

(E) Overrepresented sequences from MultiQC output of treatment and control samples. Overrepresented sequences make up a small portion of our samples.

After we performed the alignment using STAR, we used the featureCounts utility from the subread[6] tool to count reads per reference rat gene[9]. We then used MultiQC[4] to examine the output count files. This (Fig. 2) showed that around 60% of reads are assigned to exons. The sample SRR1178009 and the sample SRR1778010 with the highest assignments had slightly higher read counts as well.

The Fluconazole group had lower assignment, possibly due to its shorter read lengths, limiting the mappability. Finally we combined the counts files from each sample into a single file and created a box plot (Fig. 3) containing count distributions for each of the samples. There is a single outlier above the counts in the SRR1178009 sample. This may be a result of strong differential expression in the gene. As a result, we did not drop the outlier.

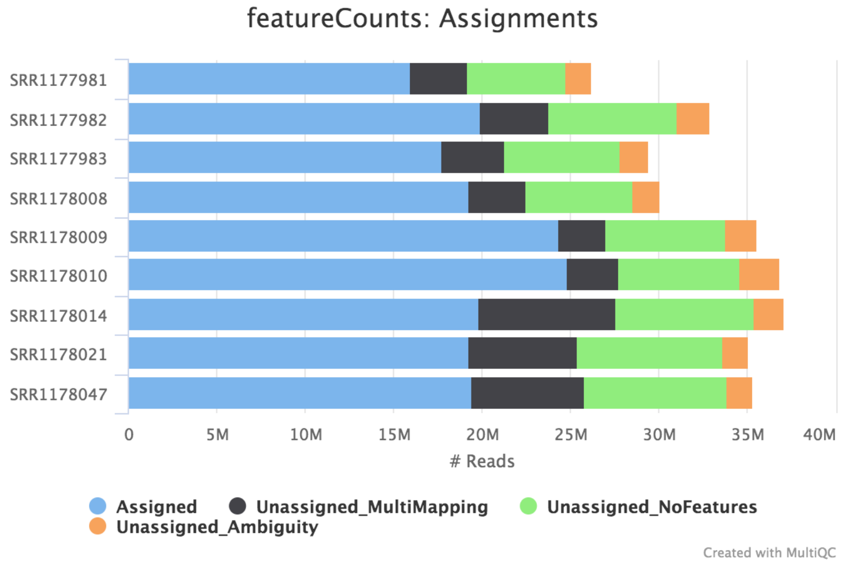


Figure 2. MultiQC results from the featureCounts files. The SRR1178014, SRR1178021 and SRR1178047 which belong to the Fluconazole group have more unassigned or multi mapping regions than other samples.

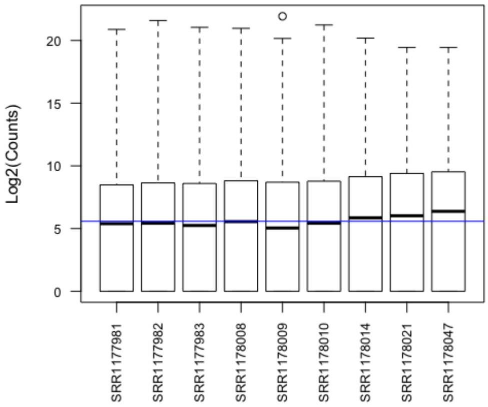
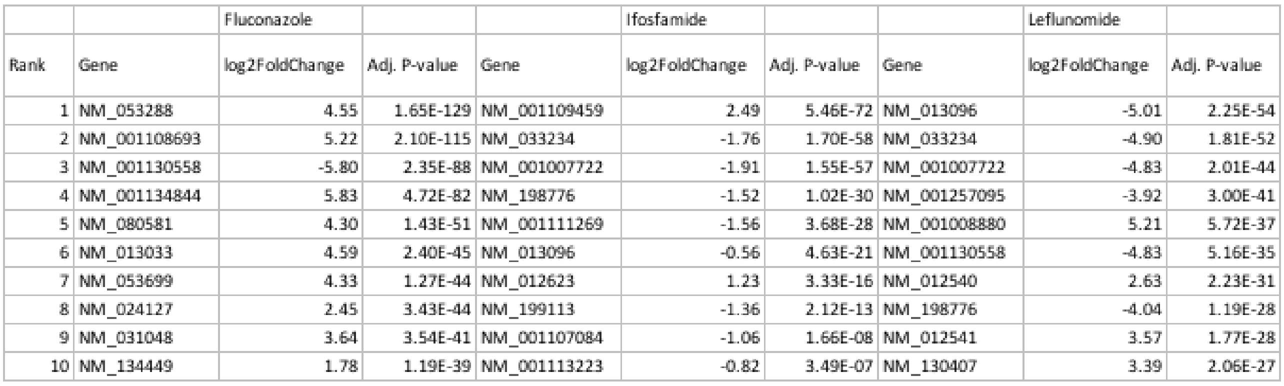


Figure 3. From the boxplots we see that overall the density distributions of raw log-intensities are not identical but still not very different.

In order to get the RNA-Seq differential expression, first we needed to combine the data matrix above using the control sample matrix. A script was then written in R to run DESeq2[7], comparing each group of the samples to the controls in the combined data matrix. Differentially expressed (DE) genes were extracted using DESeq2[7] from each of the comparisons at an adjusted p-value threshold of 0.05. We obtained 3,506 DE genes in the Fluconazole group, 127 in the Ifosfamide group and 1,379 in the Leflunomide group. The 10 genes that were most differentially expressed for each of our chemicals are listed and sorted by adjusted p-value in the table below (Table 2). The results (Figs. 4 and 5) show that there are more up-regulated genes in the Fluconazole group, more down-regulated genes in the Ifosfamide group and the greatest number of differentially expressed genes in the Leflunomide group.

Table 2. List of Top 10 most differentially expressed genes for RNA-Seq differential expression.



		Fluconazole			Ifosfamide			Leflunomide	
Rank	Gene	log2FoldChange	Adj. P-value	Gene	log2FoldChange	Adj. P-value	Gene	log2FoldChange	Adj. P-value
1	NM_053288	4.55	1.65E-129	NM_001109459	2.49	5.46E-72	NM_013096	-5.01	2.25E-54
2	NM_001108693	5.22	2.10E-115	NM_033234	-1.76	1.70E-58	NM_033234	-4.90	1.81E-52
3	NM_001130558	-5.80	2.35E-88	NM_001007722	-1.91	1.55E-57	NM_001007722	-4.83	2.01E-44
4	NM_001134844	5.83	4.72E-82	NM_198776	-1.52	1.02E-30	NM_001257095	-3.92	3.00E-41
5	NM_080581	4.30	1.43E-51	NM_001111269	-1.56	3.68E-28	NM_001008880	5.21	5.72E-37
6	NM_013033	4.59	2.40E-45	NM_013096	-0.56	4.63E-21	NM_001130558	-4.83	5.16E-35
7	NM_053699	4.33	1.27E-44	NM_012623	1.23	3.33E-16	NM_012540	2.63	2.23E-31
8	NM_024127	2.45	3.43E-44	NM_199113	-1.36	2.12E-13	NM_198776	-4.04	1.19E-28
9	NM_031048	3.64	3.54E-41	NM_001107084	-1.06	1.66E-08	NM_012541	3.57	1.77E-28
10	NM_134449	1.78	1.19E-39	NM_001113223	-0.82	3.49E-07	NM_130407	3.39	2.06E-27

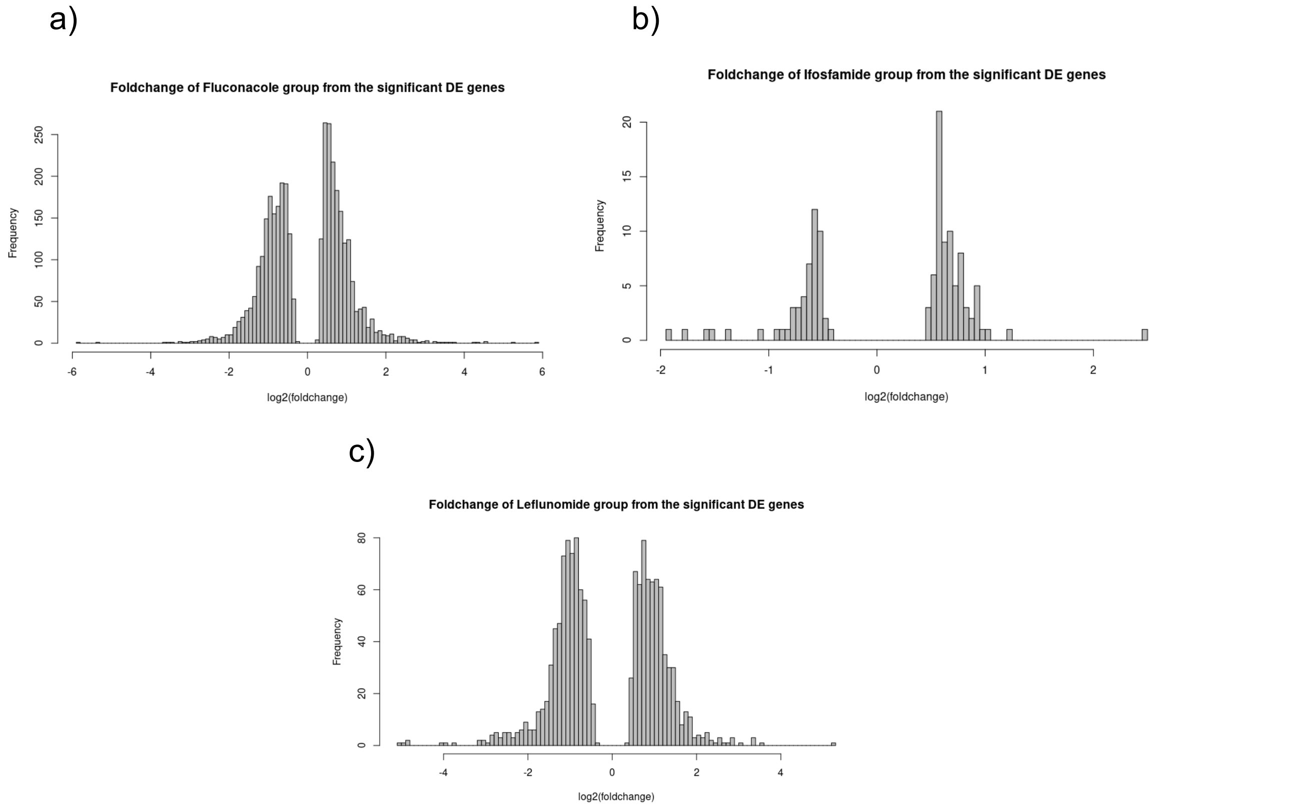


Figure 4. Log Fold Changes of the Differentially Expressed genes of data from RNA-Seq Fluconazole (a), Ifosfamide (b), and Leflunomide (c). Ifosfamide had the lowest number of differentially expressed genes.

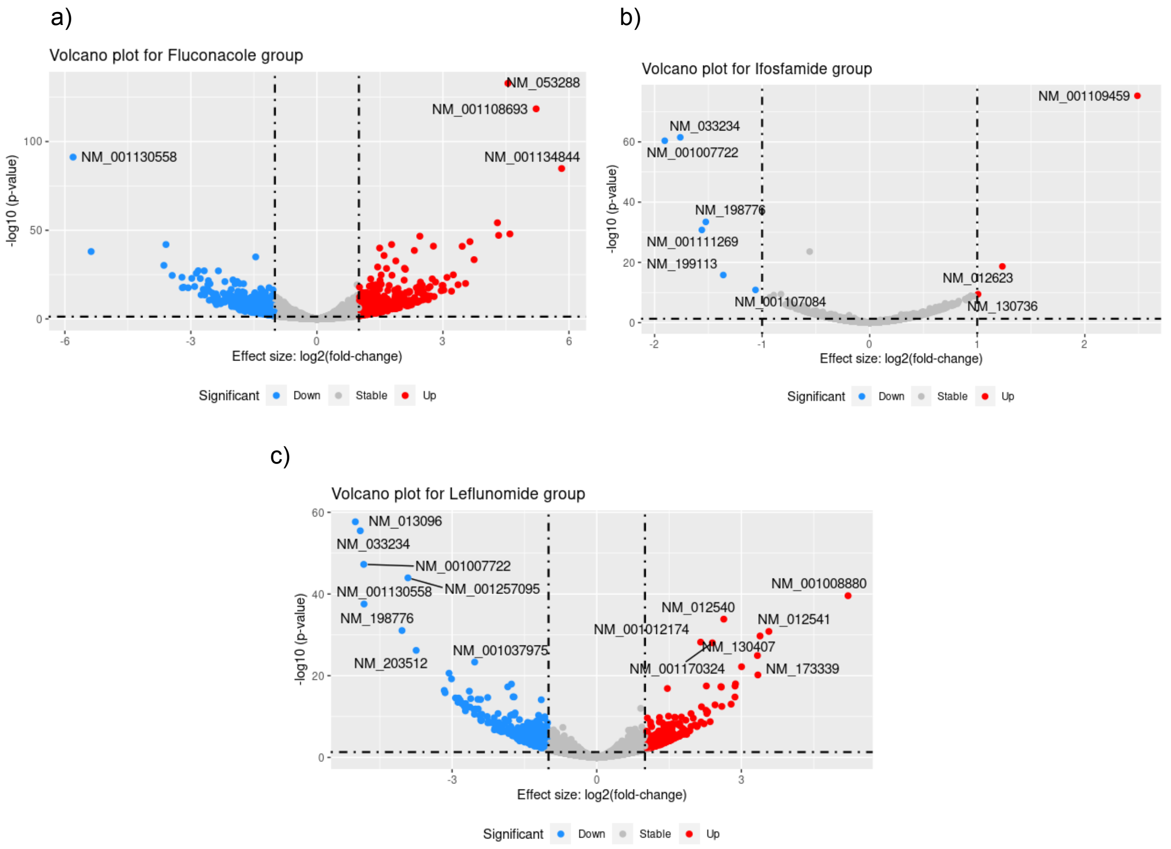


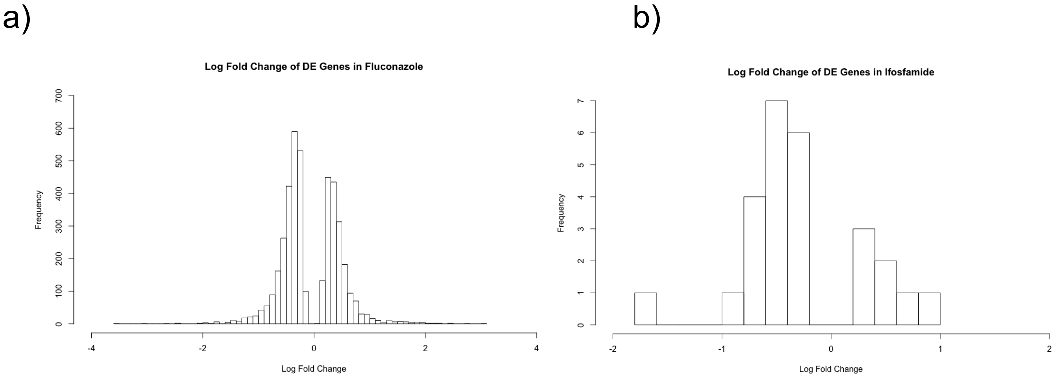
Figure 5. Plots of the effect size and significance (p-value) for differentially expressed genes in Fluconazole (a), Ifosfamide (b), and Leflunomide (c). In (a) Fluconazole group and (c) Leflunomide group, there are much more DE genes with much higher fold change. In contrast, (b) the Ifosfamide group does not seem to differ much.

Next, analysis of differential expression was performed on the microarray data. Using the limma package[8] in R, we compared the gene expression results for each of our three chemicals to our controls. We then took the probes that had an adjusted p-value < 0.05 and used those as our differentially expressed genes. The 10 probes that were most differentially expressed for each of our chemicals are listed below (Table 3).

Table 3. List of Top 10 most differentially expressed genes. Leflunomide has the most highly expressed DE genes, followed by Fluconazole and Ifosfamide.

		Fluconazole				Ifosfamide				Leflunomide	
Rank	Probe Name	Log Fold Change	Adj. P-value		Probe Name	Log Fold Change	Adj. P-value		Probe Name	Log Fold Change	Adj. P-value
1	1368731_at	1.33	1.62E-10		1376481_at	-0.64	5.45E-03		1370269_at	7.44	4.75E-14
2	1371076_at	2.47	4.11E-10		1387725_at	-0.57	1.34E-02		1387243_at	1.38	4.46E-12
3	1387316_at	2.79	1.64E-08		1383248_at	0.58	1.34E-02		1392946_at	1.60	1.18E-09
4	1395403_at	-3.58	1.68E-08		1377092_at	0.89	1.34E-02		1388611_at	-0.69	2.95E-08
5	1367957_at	2.02	4.68E-08		1375775_at	-0.71	1.34E-02		1376827_at	0.86	3.86E-08
6	1390165_at	-0.94	4.68E-08		1390271_at	0.34	1.34E-02		1372600_at	1.16	1.41E-07
7	1370698_at	1.38	6.13E-08		1370080_at	0.65	1.34E-02		1370244_at	0.55	2.47E-07
8	1370153_at	2.21	1.09E-07		1390019_at	-0.34	1.34E-02		1395403_at	-3.06	4.50E-07
9	1391570_at	1.54	1.37E-07		1387125_at	-0.76	1.50E-02		1370245_at	0.87	4.76E-07
10	1378027_at	-0.78	3.73E-07		1372996_at	0.28	1.63E-02		1387759_s_at	0.90	4.79E-07

The range and frequency of log fold changes between each of our chemicals were what we expected for fluconazole and leflunomide (Fig 6a, b), showing similar trends to the RNA-seq data. The distribution of the log fold change corresponding to the ifosfamide group, does not seem to follow the same pattern as the others. This may be due to the lower number of reads for the ifosfamide group. However, due to the lack of values close to zero, we believe that the microarray is valid.



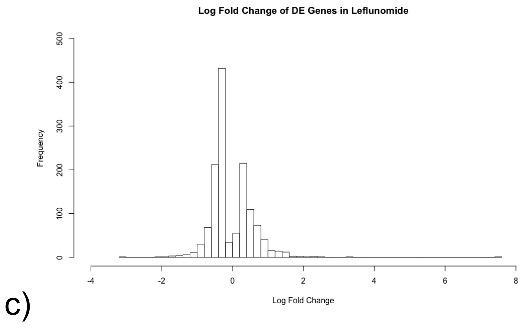


Figure 6. Log Fold Changes of the Differentially Expressed genes in Fluconazole (a), Ifosfamide (b), and Leflunomide (c). Fluconazole had 4,174 differentially expressed genes , Ifosfamide had 26, and Leflunomide had 1,348. Ifosfamide's low differentially expressed gene count is a likely explanation for why its graph is unlike the other two samples.

We also found that similar trends to the RNA-seq data occurred when we plotted the log fold change against the nominal p-value (Fig 7), which indicates that the p-values that are being assigned to our individual probes do seem to be following the patterns that we would expect.

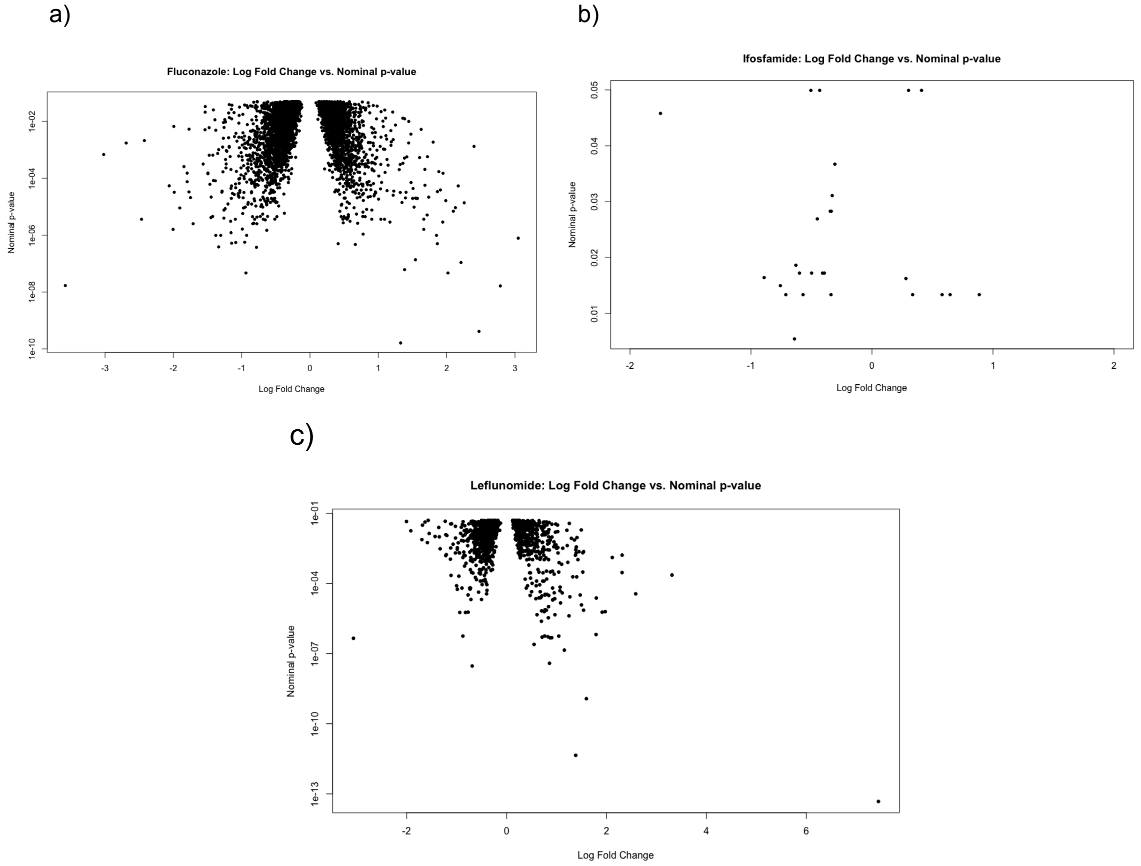


Figure 7. Plots of Log Fold change against nominal p-value for differentially expressed genes in Fluconazole (a), Ifosfamide (b), and Leflunomide (c). As with Figure 7, the number of genes expressed in Ifosfamide is too small to definitively know if this is due to unwanted effects.

After these quality control steps, we finally set out to compare the microarray and RNA-seq datasets. We found the concordance of the microarray and the RNA-seq data by finding the ratio of the overlap of the two sets. First, the names of the RNA-seq had to be matched to the probes in the microarray data. When doing this, we found that some probes mapped to multiple IDs. When this occurred, we took the first ID in the list and disregarded the others. Since we were not looking for biological significance of the probes, only whether there was a match in each data set, we found this to be a reasonable way to handle the issue. Once the RefSeqIDs had been matched to their corresponding Probe IDs, we looked to find which genes were differentially expressed in the microarray data and the RNA-seq data. We defined an intersection between two genes if the probe and corresponding RefSeqID were found in both differentially expressed datasets and the log fold change was in the same direction for both datasets.

Concordance was calculated to be $2n_0/(n_1+n_2)$, where n_0 is the number of gene probes that overlap for each chemical, n_1 is the size of the set of DE genes in the RNA-seq data, and n_2 is the size of the set of DE genes in the microarray data.

Results

First, we determined the concordance of the RNA-seq and microarray results for each individual chemical. Using only the DE genes, the concordance was 83.12% for Fluconazole, 1.30% for Ifosfamide, and 19.43% in Leflunomide. These results are quite different from the results that were found in the original study. In the original study, the concordance values ranged from 20-60%. We believe this discrepancy is because of our inability to replicate the background-correction technique that was outlined in the original paper. It is unclear if that was due to a lack of information or due to technical issues. It appears that the results with high concordance in the original study had their concordance inflated while the results with low concordance had their results deflated in our analysis.

However, despite the fact that the sheer numbers are different from the results that were found in the original study, the overall patterns that were found were consistent with their results. There was a linear relationship between the number of differentially expressed genes in each set and the concordance between them (Fig. 8). This was true for both the number of genes in RNA-seq and the number of genes in microarray.

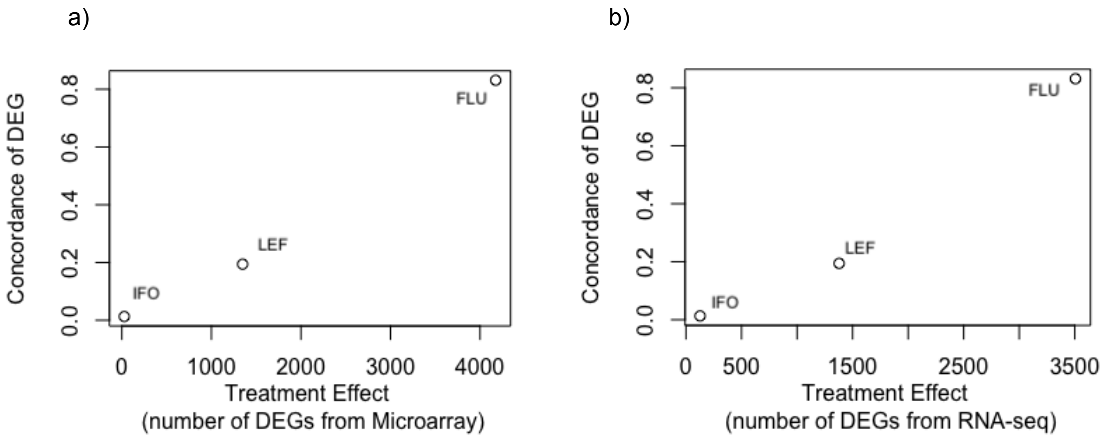


Figure 8. Concordance of the Differentially Expressed Genes (DEGs) plotted against the number of DEGs that were present in the sample. The number of DEGs either comes from the microarray data (a) or the RNA-seq data (b) . The same pattern seems to hold true for both samples of concordance increasing linearly with increasing numbers of DEGs. Our values are slightly different than the values in the original study, but still hold true to the same pattern.

Alongside that, we wanted to see if the genes that were being expressed were being expressed at roughly the same level across the microarray and RNA-seq data. To see if this was happening, we plotted the log fold changes for microarray and the log fold changes for the RNA-seq against each other.

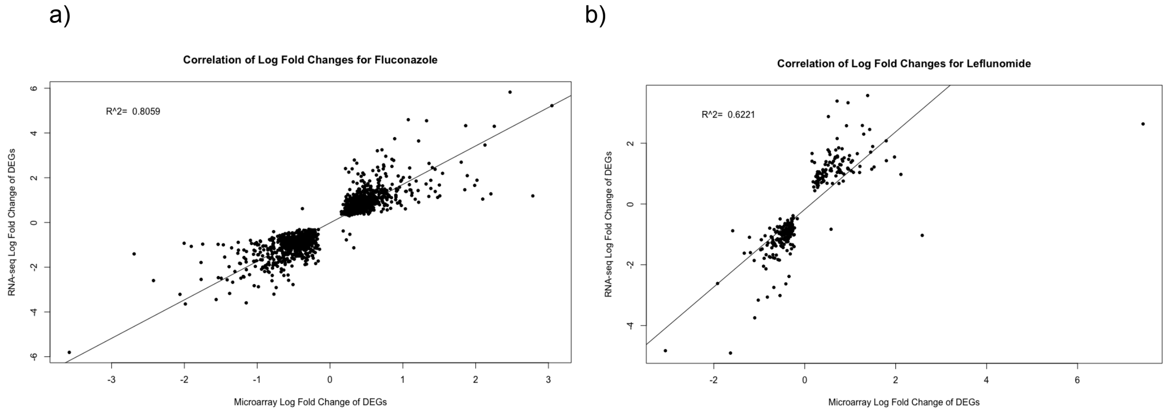


Figure 9. Plotting of the differentially expressed genes for Fluconazole (a) and Leflunomide (b). Ifosfamide was omitted due to insufficient overlapping of the RNA-seq and microarray data. The sample with more differentially expressed genes behaves in a more linear fashion in this case. p<2E-16 for both Fluconazole and Leflunomide.

We also decided to see if the above and below median results of the original study were reflected in our analysis. When we did that, we found that their original results of above median expression DEGs having higher concordance than below median expression DEGs did hold true. Although as with our previous results in Fig. 8, since our values of concordance were calculated differently than in the original study the patterns are the same but the values themselves are different.

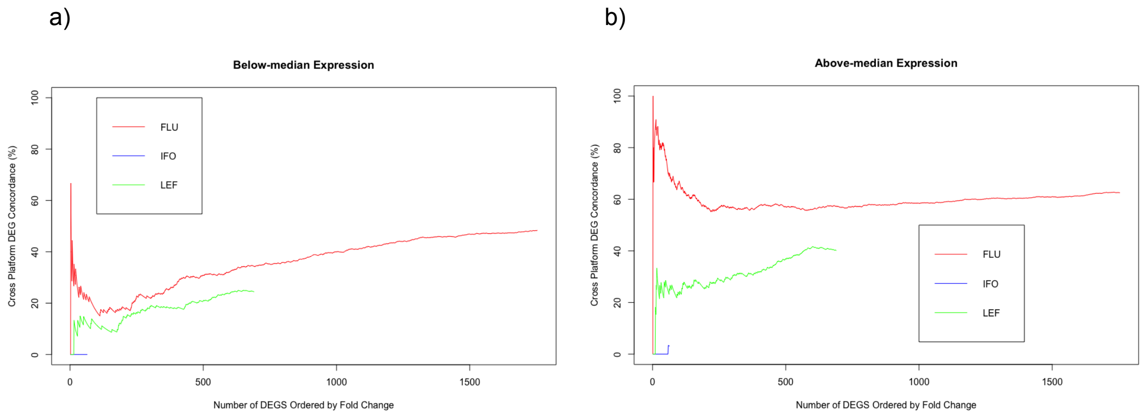


Figure 11. The below- (a) and above-median (b) expression of genes, based on the number of DE genes. The genes are ordered from largest-fold change to smallest low change. Similarly to previous results, we found that Ifosfamide did not give us enough information to be particularly relevant.

When comparing the results of each of the nine concordances that we calculated, we find some relatively surprising results. We find that the above median results are only the largest concordance in the ifosfamide and leflunomide trials, whereas the fluconazole actually had a higher concordance when all of the data was considered despite the fact that the sum of its parts was lower. This might be due to our original estimates for concordance being inflated for high values and deflated for lower values, although it is hard to say for sure.

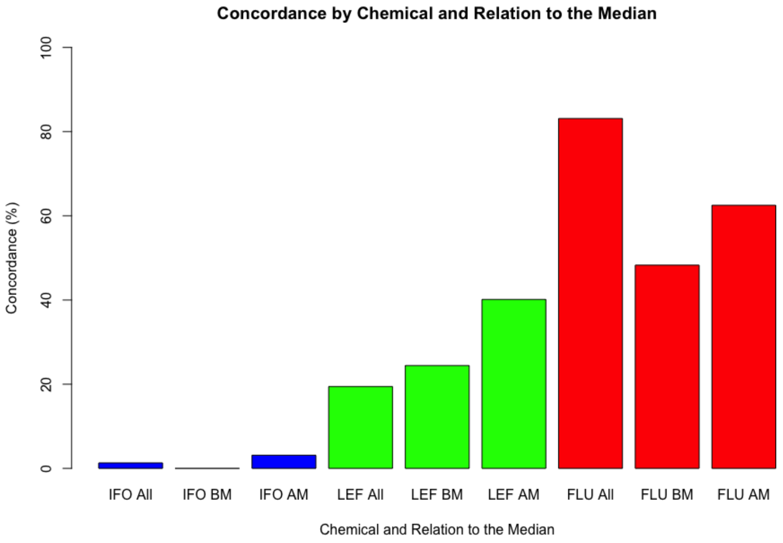


Figure 12. Concordance of all nine conditions that were taken in our experiment, with three separate chemicals (Ifosfamide, leflunomide, and fluconazole) and three separate ways of measuring the data (all DE genes, DE genes below the median, and DE genes above the median).

Discussion

After investigating one of the toxgroups administered to the mice in the original experiment, we found positive albeit mixed results. We found that the information from our analysis tended to inflate results with high concordances while deflating results with low concordances. However, despite this we found that the patterns that were present in our study were still the same as in the original.

Overall, we could conclude that the more genes that were investigated in the microarray and RNA-seq data, the higher agreement there was between the two sets. We found this to be true even though we could not replicate their background correction. This suggests that RNA-seq and microarray have a high degree of accuracy between them as long as a large number of genes are expressed. While we did not replicate everything in the original study as closely as we would have liked, we did enough to verify their main results.

Conclusion

We can successfully conclude that there is significant concordance between RNA-seq and microarray data when there are a large number of genes that are differentially expressed. The less expression that occurs, the more that RNA-seq and microarray results diverge from each other.

The main issue in our analysis is that we were unable to do background corrected concordance analysis. When we followed the method that was laid out in the original paper, we found that our concordance results were not in the range of 0-100%, which does not make any sense at all. It is unclear whether this was due to technical errors on our part or due to a miscommunication in the original paper. We believe that this is likely the cause of our disparity of the values of concordance for each chemical. Another issue was that our Ifosfamide results seemed to have significantly less overlap in our results than in the original paper, resulting in extremely low concordance scores. It appears to have been a relatively low quality sample since in our microarray data, only 26 genes were differentially expressed: more than two orders of magnitude lower than Fluconazole. The Ifosfamide sample also only had a single overlap between the RNA-seq and microarray datasets. A higher quality sample with more expression than this would have gone a long way in assisting in our journey to replicate this study.

However, these hardships and differences in our respective studies did not distort the results significantly enough to cause us to change our overall conclusion. For each part of the study that we set out to replicate, we were able to verify the results, even though some of our methods may not have been as precise as the original study.

References

1. 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
2. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update.
Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.
3. Dobin, Alexander et al. “STAR: ultrafast universal RNA-seq aligner.” *Bioinformatics (Oxford, England)* vol. 29,1 (2013): 15-21. doi:10.1093/bioinformatics/bts635
4. 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
5. Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data
6. Liao Y, Smyth GK, and Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics, 2013. subread(Version1.6.2)
7. Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, **15**, 550.
8. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic Acids Research*, **43**(7), e47. doi: 10.1093/nar/gkv007.
9. Rat RefSeq RNAs (release version 52, March 5, 2012) downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/refseq)