Data Curator:

The instrument that was used for analysis was the Illumina HiSeq2000 system; these systems are relatively common; however, they have recently become obsolete and replaced with newer models.[2] System and application version 1.9 was used for encoding and proper analysis. The Illumina TruSeq RNA Sample Preparation kit and SBS kit v3 was used in tangent with Illumina manufactures protocol to prepare the samples for sequencing.[1] The data was generated with the Rattus norvegicus Rnor_6.0 genome.[1-3] There were a total of 15 samples in the tox group we have selected. Tox group 3 provided us with six controls and nine samples. The nine samples were categorized into three different modes of action (MOA), and each sample according to their MOA had the same chemical, vehicle, and route (Table #).

toxgroup_3_rna_info

Run	mode_of_action	chemical	vehicle	route
SRR1178008	AhR	LEFLUNOMIDE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178009	AhR	LEFLUNOMIDE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178010	AhR	LEFLUNOMIDE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178014	CAR/PXR	FLUCONAZOLE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178021	CAR/PXR	FLUCONAZOLE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178047	CAR/PXR	FLUCONAZOLE	CORN_OIL_100_%	ORAL_GAVAGE
SRR1177981	DNA_Damage	IFOSFAMIDE	SALINE_100_%	ORAL_GAVAGE
SRR1177982	DNA_Damage	IFOSFAMIDE	SALINE_100_%	ORAL_GAVAGE
SRR1177983	DNA_Damage	IFOSFAMIDE	SALINE_100_%	ORAL_GAVAGE
SRR1178050	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178061	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178063	Control	Vehicle	CORN_OIL_100_%	ORAL_GAVAGE
SRR1178004	Control	Vehicle	SALINE_100_%	INTRAPERITONEAL
SRR1178006	Control	Vehicle	SALINE_100_%	INTRAPERITONEAL
SRR1178013	Control	Vehicle	SALINE_100_%	INTRAPERITONEAL

Table #.
The mode_of_action (MOA) represents the mediator or receptor process for each sample. The three different MOA's that are present in tox group 3 is aryl hydrocarbon receptor (AhR), orphan nuclear hormone receptors (CAR/PXR), and on receptor-mediated—DNA damage (DNA_Damage). Chemical represents the chemicals that the mouse was dosed with during the experiment. For tox group 3, the only chemicals that were used were LEFLUNOMIDE, FLUCONAZOLE, and IFOSFAMIDE. The vehicle represents the substance used to house the chemical for injection into the mouse. The two vehicles used were 100% corn oil (CORN_OIL_100_%) and 100% saline (SALINE_100_%). The route represents how the chemical along with the vehicle substance was administered into the mouse, ORAL_GAVAGE represents that it was introduced into the mouse orally by force while INTRAPERITONEAL means it was introduced by needle injection around the abdomen of the mouse.

For comparison to microarray data, the microarray that was used was an Affymetrix microarray. The average library size for the RNAseq analysis was about 20 million base pairs (bp_, while the reads per sequence are about 100 bp long. The reads are paired-end the output of two fastq files by the same sample confirms this.[5] The sample files were first run through fastqc and then through the STAR aligner against the Rattus norvegicus Rnor_6.0 genome. Multiqc was used to pool all of the samples together in a comprehensive report for streamline analysis (Fig. #).

Sample Name	% Aligned	M Aligned	% Dups	% GC	Length	M Seqs
SRR1178063_1_control	88.3%	39.3				
SRR1178061_1_control	89.2%	56.5				
SRR1178050_1_control	87.6%	14.1	54.6%	48%	100 bp	16.1
SRR1178047_1	84.2%	14.4	48.5%	49%	100 bp	17.1
SRR1178021_1	83.6%	14.6	48.7%	49%	100 bp	17.5
SRR1178014_1	81.6%	14.3	53.9%	49%	50 bp	17.5
SRR1178013_1_control	89.5%	14.4	58.4%	49%	101 bp	16.1
SRR1178010_1	90.7%	16.9	62.5%	49%	101 bp	18.6
SRR1178009_1	90.2%	16.3	62.2%	49%	101 bp	18.1
SRR1178008_1	88.1%	13.3	56.2%	49%	101 bp	15.2
SRR1178006_1_control	88.8%	19.1	59.2%	49%	101 bp	21.5
SRR1178004_1_control	87.7%	17.2	61.1%	48%	101 bp	19.6
SRR1177983_1	85.0%	13.7	57.6%	48%	101 bp	16.2
SRR1177982_1	88.6%	15.2	58.5%	48%	101 bp	17.2
SRR1177981_1	85.5%	12.2	56.1%	48%	101 bp	14.2
otherSTARoutputfiles	85.5%	12.2				

Fig #. MultiQC table
Table of the group analysis produced by MulitQC. % aligned represents the percentage of sequences that were successfully uniquely aligned to the reference genome. M aligned is the number of uniquely aligned sequences by the millions. % dups is the percentage of duplications per sample. % GC is the percentage of GC base pair (bp) per sample. Length is the length of bp per read, M seq is the total sequence number by million base pairs.

No samples appeared to have a quality score that was alone too low; each sample seemed to follow the same trend in terms of the higher the bp position count, the more the quality declined (Fig #). As opposed to excluding any samples, a better approach would be to trim the samples from the beginning bp to about the 70th bp position. The quality score graph does, however, raise concern considering that the controls are also included and appear to have the same trend line. I do not believe that all of the samples having the same trend line to be 100% accurate due to the assumption of the control to be of higher quality throughout the whole sample as opposed to decreasing along with the samples.

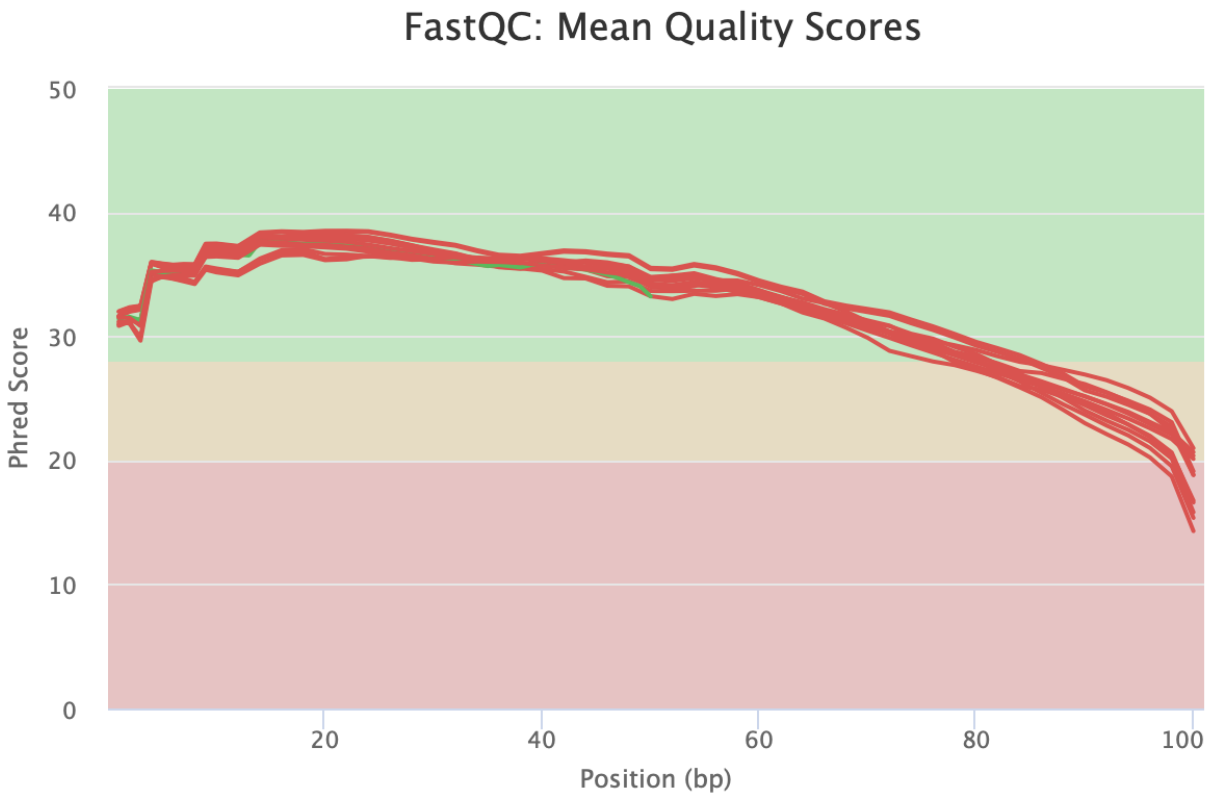

Fig #.
Created by MultiQC, a chart containing all the samples processed with an x-axis representing the base pair count of 1 sequence. The max on the x-axis is 100bp, which is the max average base pair per sequence. The Phred score on the y-axis is a quality meter with starting at 20, and going up represents preferred good reads.

Two main issues with that data that were observed are overrepresented sequences and duplication. Overrepresented sequence refers to a sequence that seemed to reappear for alignment more often than expected (Fig #). The potential source of error that was provided by the FastQC analysis program seemed to be due to TruSeq adapter index 4.[4,5] The adapter is what binds to the flow cell provided in the TruSeq RNA Sample Preparation kit used for sample prep, which implies that the issue lies within the adapter and most likely occurred during the preparation step. [2]

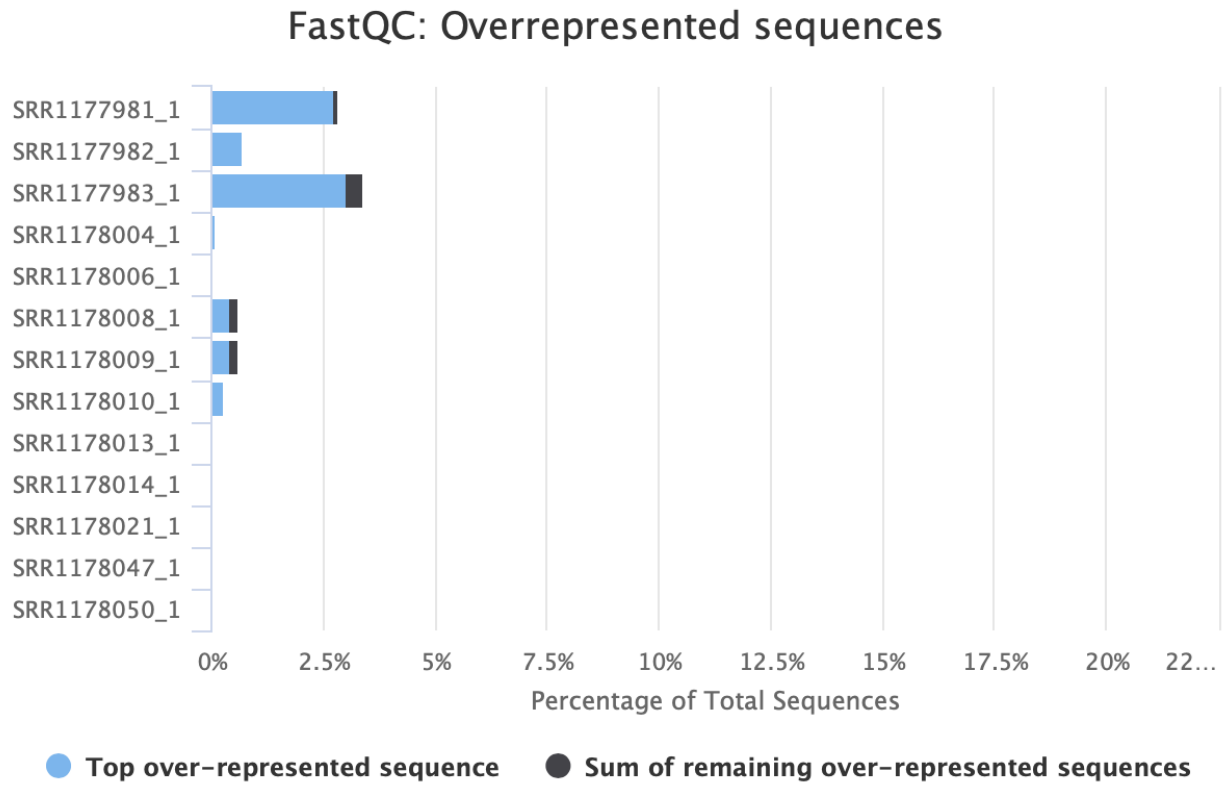

Fig #.
Produced by FastQC showing the percentage of total sequences overrepresented on the x-axis and the sample names on the y-axis.

Only six samples appeared to have overrepresented sequences with an average percentage of about 2%, which does not lead to recommend a complete do-over of the experiment starting at the peeping stage, but it is something to be noted.

The second issue appears to be duplication, with a majority of the samples achieving above 40% duplication numbers (Fig#). A majority of reasons can cause duplication, with a major one being cluster miscalling. Another possible cause for duplication is an error in the PCR process. [5]

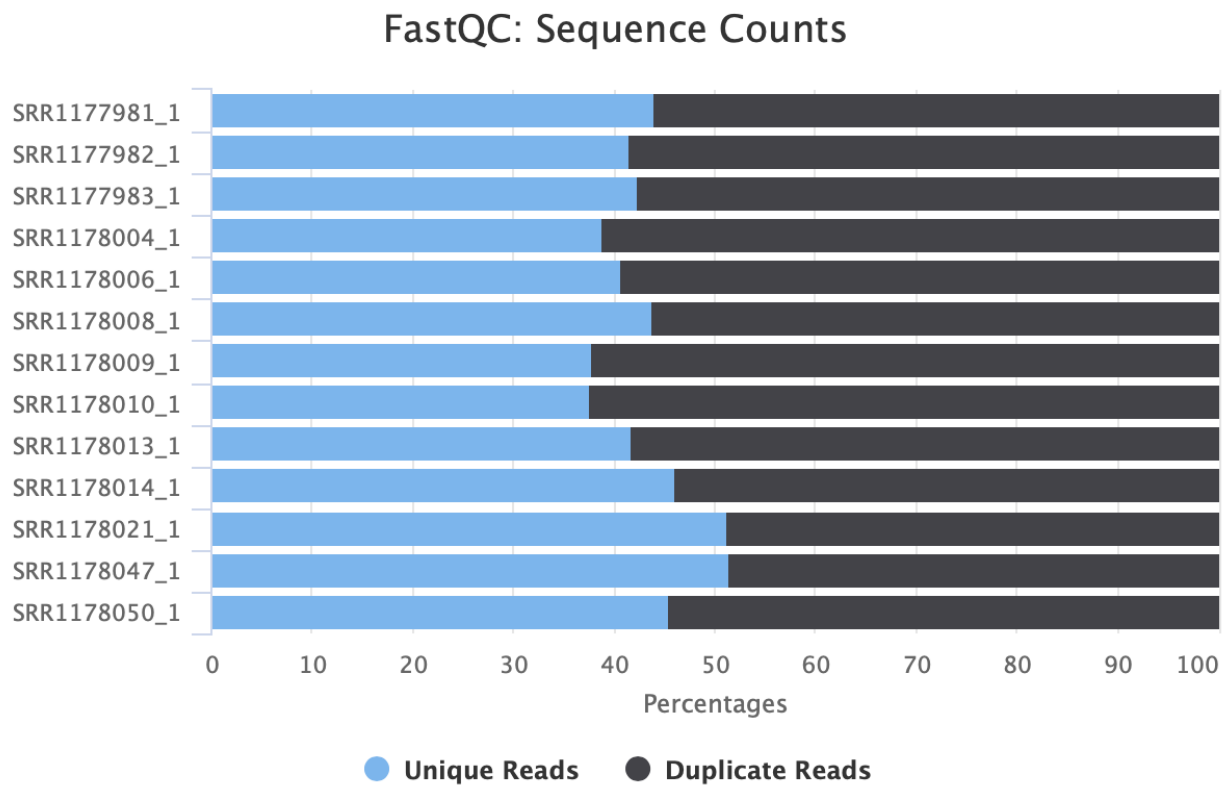

Fig #.
Chart produced by FastQC representing sequence counts per sample with the y-axis showing the percentage of unique reads vs. duplicate reads and the x-axis representing sample names.