Computational Analysis of Long Non-Coding RNA Sequences in Mouse Trophoblast Stem Cells

Zhiyue Zhang

BIOL395

Calabrese Lab

Spring 2021

Introduction

Genomes store the instructions to produce and maintain organisms. To understand the execution and regulation of such information, DNA transcription and translation have been extensively studied, with much attention given to proteins and coding RNAs. However, only 2% of human genes are protein-coding; the rest codes for non-coding RNAs (ncRNAs) that are translated into little to no proteins. Despite evidence of their biological importance, numerous ncRNAs lack complete identification and functional characterization (Diamantopoulos et al., 2018).

A major class of ncRNAs are long non-coding RNAs (lncRNA) longer than 200 base pairs. Statistics from GENCODE annotations record about 18,000 lncRNA loci in the human genome, but NONCODE - another database dedicated to ncRNA - has an estimate of nearly 100,000 (Uszczynska-Ratajczak et al., 2018) . Research on lncRNAs in the past decade have shown their key role in gene regulation: through interaction with chromatins, RNAs, and proteins, they modulate multiple levels of gene expression, such as transcription, splicing, translation, RNA stability, as well as the localization and function of proteins. The malfunction of lncRNAs and the consequent misregulation of genes are associated with a wide range of human diseases: multiple types of cancer, as compiled in databases including Lnc2Cancer and the Cancer LncRNA Census; developmental disorders, such as the Rett syndrome and autism; disorders in the brain, heart, immune system, and so on (Petazzi et al., 2013; Statello et al., 2021; Tang et al., 2017) . The pressing need to further study lncRNAs is evident.

One prominent example of biologically important lncRNAs is *Xist*, the lncRNA responsible for X-chromosome inactivation (XCI) - the process of inactivating one of the two X chromosomes in female mammal embryonic cells. By wrapping around the X chromatin, *Xist* recruits protein modifiers to remodel that chromatin and repress its transcription, thus achieving dosage compensation - equal amount of gene expression between males and females (Loda & Heard, 2019). *Airn* and *Kcnq1ot1* are two other transcription-repressive lncRNAs that work in similar fashions, but on smaller regions of the genome (Andergassen et al., 2019; Pandey et al., 2008).

The Calabrese Lab studies how lncRNAs repress transcription, using *Xist*, *Airn* and *Kcnq1ot1* as models (abbreviated as *X/A/K* below). Computationally, the lab has designed a sequence comparison method called SEEKR that relates RNA sequence to its underlying biological function. While traditional alignment algorithms including BLAST fail to detect sequence similarities between functionally similar lncRNAs (such as *X/A/K*), SEEKR is able to do so. Instead of depending on evolutionary relatedness like BLAST does, SEEKR sorts lncRNAs based on *k*-mer contents, which are short motifs - wide-spread sequence patterns that link to similar biological functions (Kirk et al., 2018)

In my project, we hypothesized that *Xist*'s function of gene silencing represents a much larger class of lncRNAs, which operate on smaller scales of the genome. To test this hypothesis, I used SEEKR to identify lncRNAs expressed in mouse trophoblast cells (TSCs) whose sequences resemble that of *Xist*. First, to extract only expressed lncRNAs, I aligned the RNA-seq data of TSCs to the genome, quantified gene expression, and used the expression levels of the whole genome to determine which lncRNAs qualify as "expressed". Next, I ran SEEKR to compare the sequences of *X/A/K* with those of expressed lncRNAs, classified the expressed lncRNAs based on sequence similarity, and analyzed multiple properties of each group. Out of the 137,034 RNAs sequenced in TSCs, I identified 26,835 lncRNAs and 2,803 expressed lncRNAs. Most importantly, my data indicates that in the search for novel lncRNAs with *Xist*-like functions, promising directions to investigate are lncRNAs that are unspliced, have lower GC-content, and are longer in RNA length.

Methods

The methods used in this project are computation-intensive, including UNIX commands and Python programs. To enhance the accessibility of viewing and reusing code, my entire computation log with annotation is uploaded on my GitHub repository.

Results and Discussion

Quantification of expressed lncRNAs in TSCs

In a previous study, other members of the Calabrese Lab generated the total RNA-seq data of mouse TSCs (Schertzer et al., 2019). I aligned the sequence data to the whole mouse genome with RSEM and bowtie. Out of the 137,034 RNA sequenced, I extracted 26,835 lncRNAs.

To determine what expression level qualifies as the benchmark for "expressed" lncRNAs, I graphed the TPM values of all RNAs and lncRNAs in TSCs (Figure 1).

The graphs match our expectation that lncRNAs are generally less transcribed than coding RNAs. The inflection point of the expression levels of all TSC RNAs is roughly TPM = 0.25. Thus, we determined any lncRNA with TPM >= 0.25 to be expressed lncRNAs. Out of 26,835 lncRNAs, we found 2,803 expressed lncRNAs, which are more likely to be functionally active in mouse TSCs.

General trends observed in expressed TSC lncRNAs in relation to k-mer similarity with Xist, Airn, and Kcnq1ot1

With SEEKR, I compared the *k*-mer profiles of the 2803 expressed lncRNAs to *X/A/K*. The resulting file contains 3 lists of Pearson's R-values, quantifying each expressed lncRNA's sequence similarity to *X/A/K*.

All 3 lists of R-values are approximately normally distributed (Figure 2). For each comparison to *X/A/K*, I partitioned the 2803 lncRNAs into 4 groups - the 99.9th percentile (with z >= 3), and 3 approximately equal-numbered tersiles for the rest (with z < 3). The number of lncRNAs in each group is 31/924/924/924 for *Xist*, 30/925/924/924 for *Airn*, and 30/925/924/924 for *Kcnq1ot1*. Here, lncRNAs in the 99.9th percentile have the highest sequence similarity to *X/A/K*, and thus are more likely to resemble their biological functions.

For each list, I analyzed the number of unspliced RNAs, GC-content, RNA length, and gene expression per group and observed trends (Table 1). Most prominently, as evidenced by binomial tests, the number of unspliced lncRNAs in each 99.9th percentile is significantly higher than the whole sample at the 0.01 significance level (Table 2). Namely, lncRNAs with higher similarity to *X/A/K* are significantly more likely to be unspliced.

Table 1

	Number of lncRNA	**Number of unspliced lncRNA**	**Mean GC-content**	**Mean length (base pair)**	**Mean log2(TPM)**
X	31/924/924/924	30/574/370/387	0.3942/0.4218/ 0.4769/0.5408	73391/11281/ 3587/2761	0.6/0.06/ 0.34/0.18
A	30/925/924/924	29/666/334/332	0.4340/0.4466/ 0.4672/0.5245	93164/13075/ 2291/1686	-0.38/-0.32/ 0.28/0.65
K	30/925/924/924	29/578/365/389	0.3980/0.4187/ 0.4762/0.5445	85456/11019/ 2895/3390	-0.37/-0.03/ 0.39/0.25

Summary of the properties analyzed for each group. The order shown here of the 4 groups is based on decreasing R-values: 99.9th percentile/first tersile/second tersile/third tersile. X/A/K each refers to the list of 2803 expressed lncRNAs when compared versus *Xist, Airn,* or *Kcnq1ot1*.

Consistent with such a pattern, and the fact that the spliced-out introns are generally AT-rich (Amit et al., 2012) , lncRNAs with higher similarity to *X/A/K* display significantly lower GC-content. Tukey's HSD tests confirm that the

GC-content in groups with higher similarity is significantly lower than groups with lower similarity at the 0.01 significance level, except for two comparisons: *Airn*'s 99.9th percentile versus first tersile, which has little difference in mean (difference in mean = 0.0125); *Kcnq1ot1*'s 99.9th percentile versus first tersile, which has a *p*-value just above 0.01 ($p = 0.0124$) (Table 3).

Similarly, lncRNAs with higher sequence similarity to *X/A/K* generally have significantly longer RNA length, which is again consistent with the fact that they are mostly unspliced. Such a pattern in length is confirmed by Tukey's HSD tests at the 0.01 significance level, except for the second versus third tersiles in each comparison to *X/A/K*, which display little difference in length ($p > 0.7$) (Table 4).

Lastly, I examined the expression levels (TPM) per group (Figure 3). As evidenced by Tukey's HSD tests, in terms of gene expression, each of the 99.9th percentile is not significantly different from the other tersiles at the 0.01 significance level (Table 5).

In conclusion, within TSCs, those expressed lncRNAs displaying the highest (99.9th percentile) sequence similarity to *X/A/K* in terms of *k*-mer profiles are significantly more likely to be unspliced, have lower GC-content, and be longer in length. These patterns signify promising directions to look when searching for novel lncRNAs with *Xist*-like functions (e.g., cis-repressive). However, there is not such a significant correlation detected between these lncRNAs' gene expression levels and the similarity of their *k*-mer profiles with *X/A/K*.

One limitation of this study is that the functional implication of whole *k*-mer profiles as calculated by SEEKR is yet unclear; for instance, the *k*-mer average of *Xist* may not be most relevant to *Xist*'s biological function. To further investigate the relation between the sequences and functions of lncRNAs, meaningful future projects could include computation of regional variation of *k*-mer contents, biochemical experiments of lncRNAs' functional properties (e.g., localization in cells, interaction with RNA binding proteins) that test SEEKR's effectiveness linking sequence to function, and similar investigation of RNA-seq data from other cell communities (e.g., mouse embryonic stem cells, human cells).

References

1. Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., Pupko, T., & Ast, G. (2012). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Reports*, *1*(5). https://doi.org/10.1016/j.celrep.2012.03.013
2. Andergassen, D., Muckenhuber, M., Bammer, P. C., Kulinski, T. M., Theussl, H. C., Shimizu, T., Penninger, J. M., Pauler, F. M., & Hudson, Q. J. (2019). The Airn lncRNA does not require any DNA elements within its locus to silence distant imprinted genes. *PLoS Genetics*, *15*(7). https://doi.org/10.1371/journal.pgen.1008268
3. Diamantopoulos, M. A., Tsiakanikas, P., & Scorilas, A. (2018). Non-coding RNAs: the riddle of the transcriptome and their perspectives in cancer. *Annals of*

Translational Medicine, *6*(12). https://doi.org/10.21037/atm.2018.06.10

4. Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., Wooten, J. S., Baker, A. R., Sprague, D., Collins, D. W., Horning, C. R., Wang, S., Chen, Q., Weeks, K. M., Mucha, P. J., & Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, *50*(10). https://doi.org/10.1038/s41588-018-0207-8
5. Loda, A., & Heard, E. (2019). Xist RNA in action: Past, present, and future. In *PLoS Genetics* (Vol. 15, Issue 9). https://doi.org/10.1371/journal.pgen.1008333
6. Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-DiNardo, D., & Kanduri, C. (2008). Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation. *Molecular Cell*, *32*(2). https://doi.org/10.1016/j.molcel.2008.08.022
7. Petazzi, P., Sandoval, J., Szczesna, K., Jorge, O. C., Roa, L., Sayols, S., Gomez, A., Huertas, D., & Esteller, M. (2013). Dysregulation of the long non-coding RNA transcriptome in a Rett syndrome mouse model. *RNA Biology*, *10*(7). https://doi.org/10.4161/rna.24286
8. Schertzer, M. D., Braceros, K. C. A., Starmer, J., Cherney, R. E., Lee, D. M., Salazar, G., Justice, M., Bischoff, S. R., Cowley, D. O., Ariel, P., Zylka, M. J., Dowen, J. M., Magnuson, T., & Calabrese, J. M. (2019). lncRNA-Induced Spread of Polycomb Controlled by Genome Architecture, RNA Abundance, and CpG Island DNA. *Molecular Cell*, *75*(3). https://doi.org/10.1016/j.molcel.2019.05.028

9. Statello, L., Guo, C. J., Chen, L. L., & Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. In *Nature Reviews Molecular Cell Biology* (Vol. 22, Issue 2). https://doi.org/10.1038/s41580-020-00315-9

10. Tang, J., Yu, Y., & Yang, W. (2017). Long noncoding RNA and its contribution to autism spectrum disorders. In *CNS Neuroscience and Therapeutics* (Vol. 23, Issue 8). https://doi.org/10.1111/cns.12710

11. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., & Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. In *Nature Reviews Genetics* (Vol. 19, Issue 9). https://doi.org/10.1038/s41576-018-0017-y

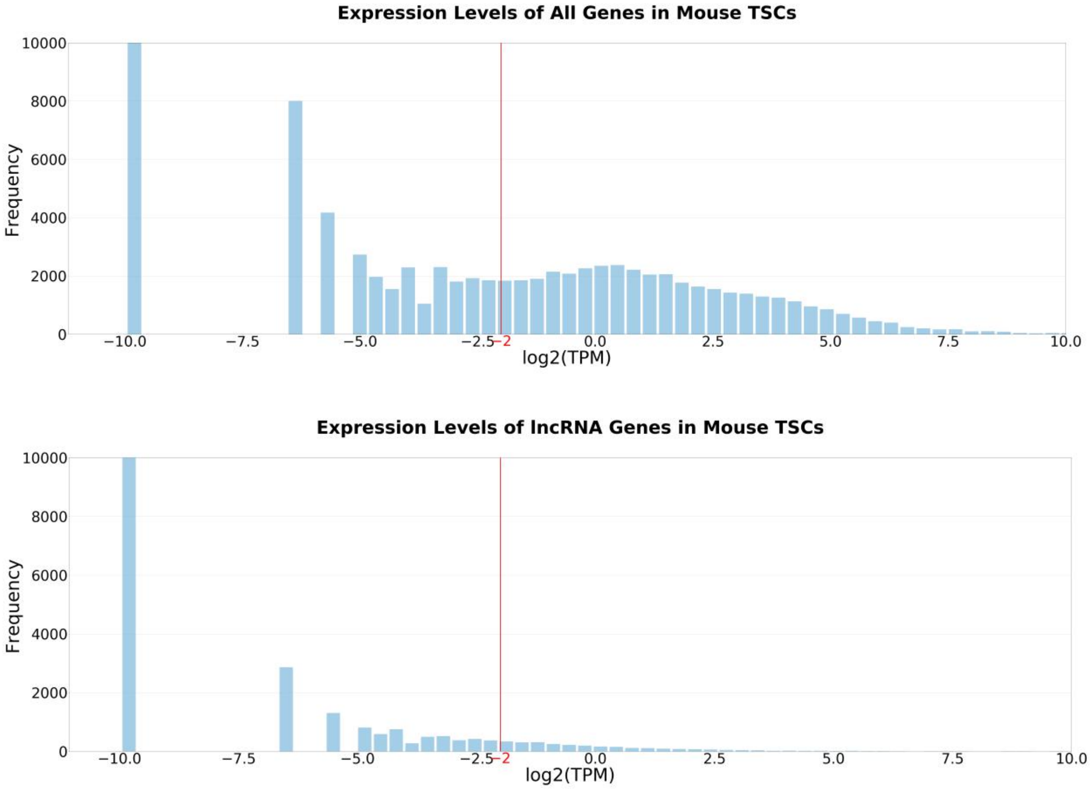
Links

1. Github page for the computational methods used in this project

   https://github.com/zhiyuezhang7/tsc_lncRNA_project

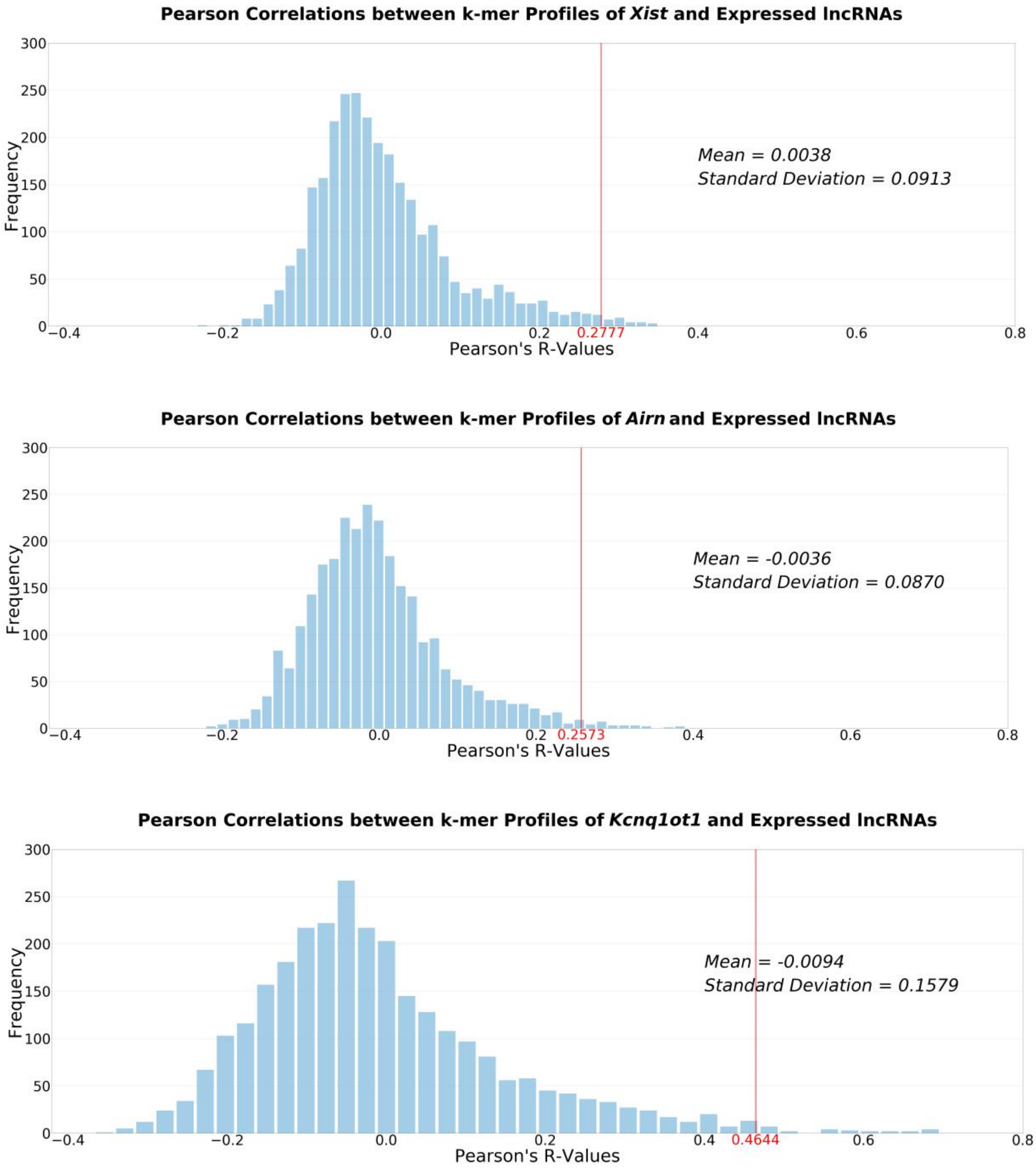
Appendix

Figure 1



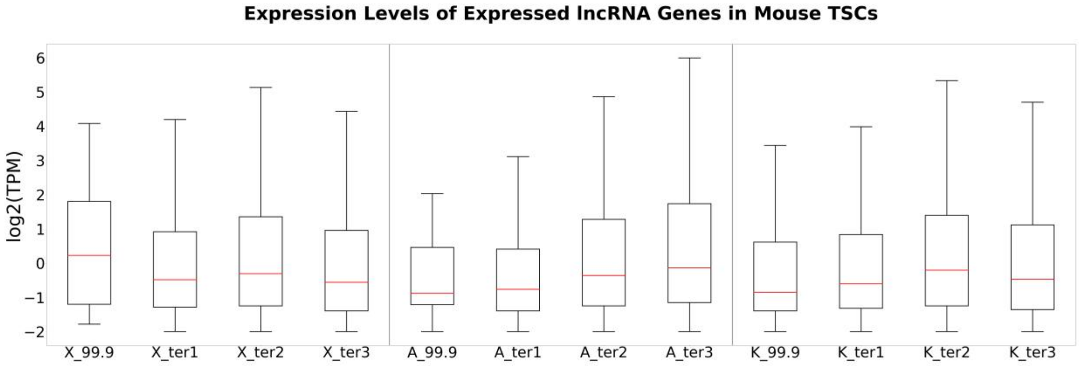
Expression levels (log2(TPM)) of all RNAs and only lncRNAs in mouse TSCs. The red vertical lines mark the rough inflection point of the TPM of all TSC RNAs, which determines the benchmark for an RNA to be "expressed".

Figure 2



Pearson correlations between *k*-mer profiles of expressed lncRNAs and *X/A/K*. For each comparison, I calculated the mean, standard deviation, and z-score=3 value in Excel. The red vertical lines mark the values when z-score=3, which correspond to the benchmark for the 99.9th percentile in each list.

Figure 3



Expression levels (log2(TPM)) of the expressed lncRNAs in TSCs. X_99.9 refers to the 99.9th percentile within these lncRNAs, whose sequences are most similar to *Xist*; X_ter1 refers to the first tersile in the rest of these lncRNAs when compared with *Xist*; other labels follow similar patterns, including all the tables below.

Table 2

	***p*-value**
X_99.9	0.0000000062
A_99.9	0.0000000123
K_99.9	0.0000000123

Number of unspliced. The results of binomial tests for the number of unspliced lncRNAs in each of the 99.9th percentile.

Table 3

GC-content. The results of Tukey's HSD tests of the GC-content of lncRNAs per group. The significance level used is all 0.01.

Table 3.1

group1	**group2**	**meandiff**	***p*-adj**	**lower**	**upper**	**reject**
X_99.9	**X_ter1**	0.0277	0.001	0.0048	0.0505	TRUE
X_99.9	**X_ter2**	0.0828	0.001	0.06	0.1056	TRUE
X_99.9	**X_ter3**	0.1467	0.001	0.1239	0.1695	TRUE
X_ter1	**X_ter2**	0.0551	0.001	0.0493	0.0609	TRUE
X_ter1	**X_ter3**	0.119	0.001	0.1132	0.1248	TRUE
X_ter2	**X_ter3**	0.0639	0.001	0.0581	0.0697	TRUE

Table 3.2

group1	**group2**	**meandiff**	***p*-adj**	**lower**	**upper**	**reject**
A_99.9	**A_ter1**	0.0125	0.5854	-0.0188	0.0438	FALSE
A_99.9	**A_ter2**	0.0332	0.0054	0.0019	0.0645	TRUE
A_99.9	**A_ter3**	0.0905	0.001	0.0592	0.1218	TRUE
A_ter1	**A_ter2**	0.0206	0.001	0.0128	0.0285	TRUE
A_ter1	**A_ter3**	0.0779	0.001	0.0701	0.0858	TRUE
A_ter2	**A_ter3**	0.0573	0.001	0.0494	0.0651	TRUE

Table 3.3

group1	**group2**	**meandiff**	***p*-adj**	**lower**	**upper**	**reject**
K_99.9	**K_ter1**	0.0207	0.0124	-0.0004	0.0418	FALSE
K_99.9	**K_ter2**	0.0782	0.001	0.057	0.0993	TRUE
K_99.9	**K_ter3**	0.1465	0.001	0.1254	0.1676	TRUE
K_ter1	**K_ter2**	0.0575	0.001	0.0522	0.0628	TRUE
K_ter1	**K_ter3**	0.1258	0.001	0.1205	0.1311	TRUE
K_ter2	**K_ter3**	0.0683	0.001	0.063	0.0736	TRUE

Table 4

RNA length. The results of Tukey's HSD tests of the length of lncRNAs per group. The significance level used is all 0.01.

Table 4.1

group1	**group2**	**meandiff**	***p*-adj**	**lower**	**upper**	**reject**
X_99.9	**X_ter1**	-62109.6	0.001	-72632.525	-51586.7	TRUE
X_99.9	**X_ter2**	-69803.7	0.001	-80326.564	-59280.7	TRUE
X_99.9	**X_ter3**	-70629.8	0.001	-81152.745	-60106.9	TRUE
X_ter1	**X_ter2**	-7694.04	0.001	-10375.244	-5012.83	TRUE
X_ter1	**X_ter3**	-8520.22	0.001	-11201.425	-5839.01	TRUE
X_ter2	**X_ter3**	-826.181	0.7462	-3507.3858	1855.024	FALSE

Table 4.2

group1	**group2**	**meandiff**	***p*-adj**	**lower**	**upper**	**reject**
A_99.9	**A_ter1**	-80089.4	0.001	-90073.003	-70105.7	TRUE
A_99.9	**A_ter2**	-90872.9	0.001	-100856.75	-80889.1	TRUE
A_99.9	**A_ter3**	-91478.7	0.001	-101462.48	-81494.9	TRUE
A_ter1	**A_ter2**	-10783.6	0.001	-13286.687	-8280.46	TRUE
A_ter1	**A_ter3**	-11389.3	0.001	-13892.418	-8886.2	TRUE
A_ter2	**A_ter3**	-605.731	0.862	-3109.5184	1898.057	FALSE

Table 4.3

group1	**group2**	**meandiff**	***p*-adj**	**lower**	**upper**	**reject**
K_99.9	**K_ter1**	-74436.6	0.001	-84866.099	-64007.1	TRUE
K_99.9	**K_ter2**	-82560.8	0.001	-92990.501	-72131.2	TRUE
K_99.9	**K_ter3**	-82066.1	0.001	-92495.735	-71636.4	TRUE
K_ter1	**K_ter2**	-8124.22	0.001	-10739.123	-5509.33	TRUE
K_ter1	**K_ter3**	-7629.46	0.001	-10244.358	-5014.56	TRUE
K_ter2	**K_ter3**	494.7652	0.9	-2120.84	3110.37	FALSE

Table 5

Gene expression. The results of Tukey's HSD tests of the log2(TPM) of lncRNAs per group. The significance level used is all 0.01.

Table 5.1

group1	group2	meandiff	*p*-adj	lower	upper	reject
X_99.9	**X_ter1**	-0.5408	0.4997	-1.7451	0.6635	FALSE
X_99.9	**X_ter2**	-0.2586	0.9	-1.4629	0.9457	FALSE
X_99.9	**X_ter3**	-0.4242	0.6691	-1.6285	0.7801	FALSE
X_ter1	**X_ter2**	0.2822	0.0218	-0.0246	0.5891	FALSE
X_ter1	**X_ter3**	0.1166	0.6204	-0.1902	0.4235	FALSE
X_ter2	**X_ter3**	-0.1656	0.3337	-0.4725	0.1412	FALSE

Table 5.2

group1	group2	meandiff	*p*-adj	lower	upper	reject
A_99.9	**A_ter1**	0.059	0.9	-1.1444	1.2624	FALSE
A_99.9	**A_ter2**	0.6595	0.3198	-0.5439	1.863	FALSE
A_99.9	**A_ter3**	1.0262	0.0396	-0.1773	2.2296	FALSE
A_ter1	**A_ter2**	0.6005	0.001	0.2988	0.9022	TRUE
A_ter1	**A_ter3**	0.9672	0.001	0.6655	1.2689	TRUE
A_ter2	**A_ter3**	0.3667	0.001	0.0649	0.6685	TRUE

Table 5.3

group1	**group2**	**meandiff**	***p*-adj**	**lower**	**upper**	**reject**
K_99.9	**K_ter1**	0.3372	0.8022	-0.8838	1.5581	FALSE
K_99.9	**K_ter2**	0.7577	0.2143	-0.4633	1.9786	FALSE
K_99.9	**K_ter3**	0.6182	0.3927	-0.6028	1.8392	FALSE
K_ter1	**K_ter2**	0.4205	0.001	0.1144	0.7266	TRUE
K_ter1	**K_ter3**	0.281	0.0221	-0.0251	0.5872	FALSE
K_ter2	**K_ter3**	-0.1395	0.4879	-0.4457	0.1667	FALSE