Investigation Of Novel ORFs In Mouse Cell Line And Human Tissues

A Thesis submitted to
Indian Institute of Science Education and Research (IISER), Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Prashant Uniyal
(20131058)

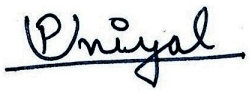
Project Supervisor
Dr. Sudhakaran Prabakaran
Department of Biology, IISER Pune &
Department of Genetics, University of Cambridge

Indian Institute of Science Education and Research (IISER) Pune
Dr. Homi Bhabha Road, Pashan, Pune 411008, INDIA.

March, 2019

Certificate

This is to certify that this dissertation entitled “Investigation Of Novel ORFs In Mouse Cell Line And Human Tissues” towards the partial fulfillment of the BS-MS Dual Degree programme at the Indian Institute of Science Education and Research (IISER), Pune represents original research carried out by Prashant Uniyal at Indian Institute of Science Education and Research (IISER), Pune under the supervision of Dr. Sudhakaran Prabakaran, India DBT - Cambridge Lecturer, Department of Biology, IISER Pune and Department of Genetics, University of Cambridge, during the academic year 2018-2019.



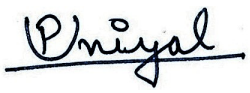
Student
Prashant Uniyal

Supervisor
Dr. Sudhakaran Prabakaran

Date: 20th March, 2019

Declaration

I hereby declare that the matter embodied in the report entitled “Investigation Of Novel ORFs In Mouse Cell Line And Human Tissues” are the results of the work carried out by me at the Department of Biology, Indian Institute of Science Education and Research (IISER) Pune, under the supervision of Dr. Sudhakaran Prabakaran and the same has not been submitted elsewhere for any other degree.



Student
Prashant Uniyal

Supervisor
Dr. Sudhakaran Prabakaran

Date: 20th March, 2019

Abstract

It has become quite apparent that the genomes of many organisms are much more complex than thought before the usage of routine high throughput sequencing of genomes of various organisms. It is quite well known that there are coding regions of the genome that are transcribed and translated to form functional proteins. The transcription and translation is not restricted to these regions but other non coding regions are also transcribed and translated. These noncoding transcriptional events have been claimed to be 'transcriptional noise' but we think otherwise. We show that these noncoding transcriptional events are not noise by studying Nascent RNA sequences from *Mus musculus* and think that they can play an important role in various cell functions. The work done in the project shows that in Nascent RNA sequences from *Mus musculus*, there was no differential expression between knock-out of a histone variant (which would leave the enhancers and chromatin open for non-specific transcriptions to happen and therefore increasing noisy transcription) when compared to wild type. This analysis rules out that transcription of sORFs occurs due to noisy transcriptional events. Having established that sORFs are not biological noise and we want to try and aim to further strengthen this argument by finding expression of sORFs in healthy tissues by analyzing GTEx datasets. The GTEx dataset, being a huge collection of mRNA data from normal human tissues, helped us understand and quantify the expression of sORFs and other Novel ORFs at a large scale. The project then goes on to study how various noncoding regions like sORFs, altORFs, pseudogenes and de novo genes are expressed in 53 healthy human tissue types from the GTEx database and quantifies their expression in these tissues.

Contents

1. Introduction …… 8

2. Methods …… 11

2.1 sORFs transcription in mouse Nascent RNA-seq data …… 11

2.2 Expression of Novel ORFs in various Human tissues …… 19

3. Results and Discussion …… 25

3.1 Analysis of the mouse Nascent RNA-seq data …… 25

3.2 Analysis of expression levels of Novel ORFs in normal human tissues .. 30

4. Conclusion and Future Directions…… 39

5. References …… 41

List of Figures

Figure 1: Graph showing per base sequence quality and adapter content for NEBNext12 ……..……..…….. 12
Figure 2: Workflow of the quantification of Nascent RNA Sequencing data using StringTie ……..…… 13
Figure 3: Workflow of the quantification of Nascent RNA Sequencing data using MAPS ……... 17
Figure 4: Distribution of RNA-Seq samples for the 53 human tissue types …….. 19
Figure 5: sORF transcripts generated from bedtools ……... 23
Figure 6: Cluster dendogram for a DE analysis methods …… 26
Figure 7: Expression values in TPM for the H2afz and H2afv genes ……. 27
Figure 8: Plot for log(FPKM+0.0000001) of NEB06 (-TAM) against NEB02 (+TAM) for the time point 1 and 2 ……. 29
Figure 9: Distribution of lengths of sORFs and sORFs annotations …….. 30
Figure 10: Distribution of the lengths of sORF transcripts overlapped with the …... 31
Figure 11: Log mean TPM expression of novel ORF transcripts in GTEx tissues ... 32
Figure 12: The number of novel ORF transcripts expressed in GTEx tissues ……. 35
Figure 13: The number of uniquely expressed novel ORF transcripts in GTEx tissues ……... 37

Acknowledgements

I would like to extend my sincere and heartfelt gratitude towards my project supervisor Dr. Sudhakaran Prabakaran, for his continuous encouragement and guidance at every step of my project. He has invested a great deal of time and effort in my project, and this thesis would not have been possible without his expert mentorship. I would like to express my gratitude to my thesis advisory committee member, Dr. Krishanpal Karmodiya, for his insightful inputs into my project.

I would also like to thank Dr. Kiran Padmanabhan and his lab at ENS Lyon for providing us the mouse Nascent RNA sequencing data.

I would like to thank all the members of Prabakaran lab especially Dr. Shraddha Puntambekar, Narendra Meena, Matt Neville, Chaitanya Erady and Robin Kohze for helping me at various steps during the project and for providing me with their useful suggestions and valuable discussions.

I would also like to thank Department of Biology, IISER Pune for again providing me with all the necessary facilities for the completion of my work.

I am extremely thankful to my family and friends for all the emotional support and their confidence in me.

1. Introduction

Since the advent of routine high throughput sequencing of genomes of a number of organisms, it has become apparent that these genomes are much more complex than previously thought. Not only there are known coding regions or genes that are transcribed and translated to form proteins, recent evidence has shown that transcription and translation can happen pervasively in the human genome and these events are not just restricted to the coding regions (Prabakaran et al., 2014). The magnitude of these pervasive transcriptional and translational events is generally underestimated by the previously common strategies such as microarrays and some of the existing RNA sequencing pipelines. Systematic identification of various pervasive transcriptional events have been possible due to various advanced techniques such as Gro-seq, NET-seq, metabolic labeling (Bhatt et al., 2012; Herzel and Neugebauer, 2015) and nascent RNA sequencing. Along with these, identification of various translation events from noncoding regions has been made possible due to various Systems Proteogenomics approaches and Ribosome profiling studies (Chew et al., 2013).

It is not known whether the noncoding transcriptional events are biological noise or they are translated into peptide products with some biological functions. Various noncoding transcriptional events have been claimed to be ‘transcriptional noise’ but we think otherwise that these noncoding transcriptional events are not noise and can play an important role in various cell functions.

It is important for cells, tissues and various organisms to have an idea of the time during the day as it is helpful to regulate various functions and controls inputs from the environment as they fluctuate during a day i.e. circadian rhythm that are changes that oscillate during the 24 hour period cycle. A nucleosome is a basic structural unit of chromatin and is very important for DNA packaging. It is a histone octamer formed by 2 copies of each H2A, H2B, H3 and H4 (the core histones) with 147 base pairs wrapped around it. There are variations of the histone that are also present as a barrier to various enzymes for accessing the underlying DNA for replication and transcription processes.

H2A.Z is a variant of H2A, which is encoded by two different genes H2afz and H2afv. It is somewhat similar to the conventional H2A histone. H2afz gene is located on the chromosome 3 and H2Afv gene is located on the chromosome 11 in the *Mus Musculus* genome. The variation in the histone can change the structure of the the nucleosome. Nucleosomes containing both histone variants H3.3 and H2A.Z are even less stable than nucleosomes containing the conventional H3.3 and H2A histones (Jin and Felsenfeld, 2007), which suggests that the DNA wraps around the nucleosome which contains the histone variant H2A.Z and H3.3 is bound loosely. Our collaborators wanted to study the role of the the dynamics and states of these chromatin as these could be affected by these histone variants and to understand their role in circadian rhythm and gave us the Nascent RNA Seq data that was sequenced from

Various noncoding transcriptional events have been claimed to be 'transcriptional noise' but we think these noncoding transcriptional events can play an important role in various cell functions. The project looks at non canonical open reading frame like Small open reading frames (sORFs) which can be defined as open reading frames smaller than 100 amino acids. We are mainly investigating sORFs because the non-coding regions have been shown to be transcribed and translated by previous studies (Bazin et al., 2017; Olexiouk et al., 2018).

It is our claim that a mutant histone protein leaves the enhancers and chromatin open for non-specific transcriptions to happen and hence one should find random ORFs transcribed. But if there is no differential expression it could mean that these events are not 'biological noise'. This project involves analysis of nascent RNA sequences obtained from a mouse embryonic fibroblasts (MEFs) that are bound to chromatin in the nucleus and we use that information to infer the prevalence of various noncoding transcriptional events. We wanted to study if knocking out the H2afz-H2afv genes which code for the H2A.Z histone mutant show any differential expression of genes when compared to the wild type and if there is any expression of non coding transcription events likes sORFs.

In *Mus musculus*, knocking-out of a histone variant would leave the enhancers and chromatin open for non-specific transcriptions to happen and

therefore increasing noisy transcription compared to wild type. If we find no differential expression between the knock-out and the wild type cell lines, it hints towards the fact that transcription of sORFs is not due to noisy transcriptional events.

We claim to establish that sORFs are not biological noise and we want to try and aim to further strengthen this argument by finding the expression of sORFs in various tissues. The GTEx dataset, being a huge collection of mRNA data from human tissues, would help us understand the expression of sORFs and other Novel ORFs and quantify the expression at a large scale. We study the expression of various other novel ORFs like altORFs, pseudogenes and de novo genes in the 53 tissue types in the GTEx dataset. AltORFs stand for alternative open reading frames which are non canonical open reading frames but contain start codons that are different from the canonical open reading frames and therefore altORFs code for an alternative protein (Vanderperre et al., 2012). Pseudogenes are genomic DNA sequences that are related to normal genes. These gene generally have lost at least some functionality, relative to the complete gene, either in terms gene expression in the cell or their ability to code for proteins. De novo genes are those novel coding genes that have evolved from previously noncoding regions, thus generating entirely novel proteins.

2. Methods

2.1 sORFs transcription in mouse Nascent RNA-seq data:

The first part of the project involves analysis mouse Nascent RNA-seq data to find out differential expression. It also involves finding the expression of sORFs in this mouse Nascent RNA-seq data.

2.1.1 Nascent RNA-Seq Samples:

Four nascent, nuclear RNA samples from Mus musculus Embryonic Fibroblasts (MEFs) were sequenced using the Illumina HiSeq 4000 workflow generating a paired end read data. Two replicate samples (NEBNext02, NEBNext04) (denoted by +TAM) were treated with Tamoxifen, a drug which leads to the deletion of H2afz and H2afv genes. The remaining two replicate samples (NEBNext06, NEBNext12) (denoted by -TAM) were wild-type cell samples. The sequence length of the paired end sample reads was 150 base pairs and the total sequences were about 100Mb (Mega base pairs) in each of the four samples. These samples were kindly provided to us by Dr. Kiran Padmanabhan's Lab at ENS Lyon, France.

2.1.2 Quality Check of the Nascent RNA-Seq Samples:

The four Nascent RNA sample reads were analyzed using FASTQC (Andrews, 2010), a tool which is commonly used for checking the quality of the reads. It was observed that the 'per base sequence quality' which determines the overall quality of the sequencing was good but a substantial amount of 'adapter content' was present in all of the reads. Figure 1(a) shows the graph for per base

sequence quality for the sample NEBNext12 and Figure 1(b) shows the graph for adapter content in NEBNext12 as a representative of the rest of the samples.

The adapter content in the Nascent RNA sample reads is due to the adapters which are synthetic sequences and are added to the samples to facilitate RNA sequencing. Removing adapter content is necessary as they can cause errors in the alignment process and an increased number of unaligned reads, since the adapter sequences are synthetic and do not occur in the genomic sequences.

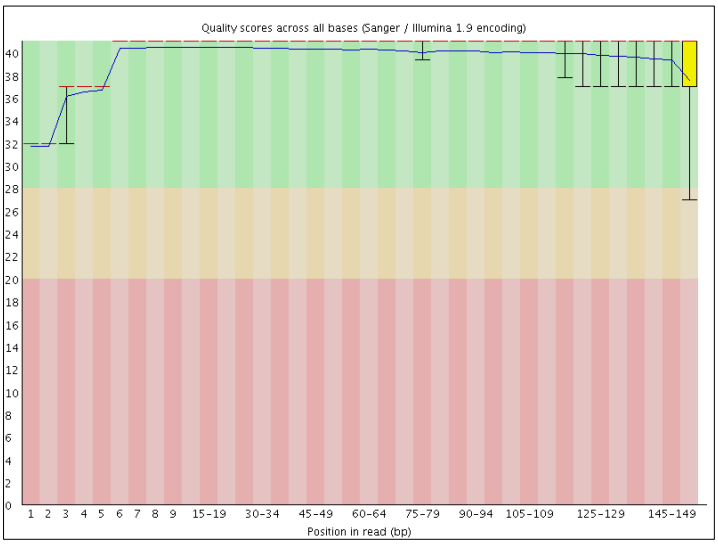


Figure 1(a): Graph showing per base sequence quality for NEBNext12

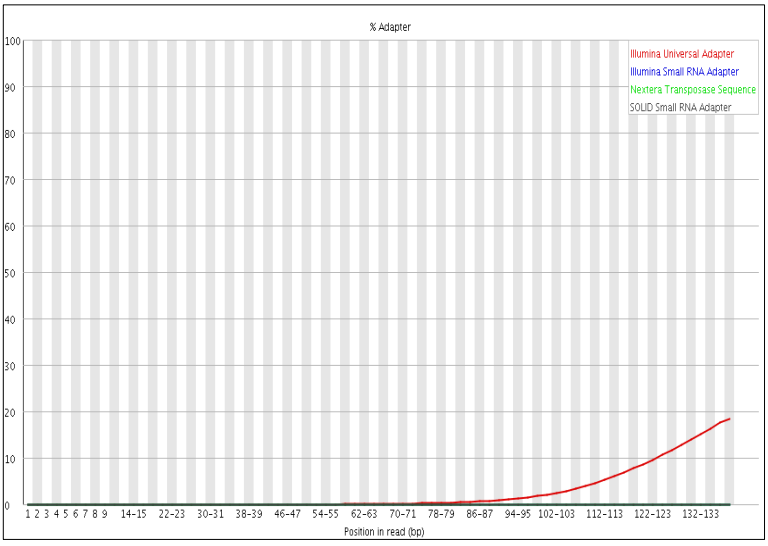


Figure 1(b): Graph showing adapter content for NEBNext12

Due to the significant amount of adapter content in all of the reads, a NGS data trimming tool called Trimmomatic (Bolger et al., 2014) was used to remove these adapters. After the Trimmomatic's adapter removal and quality check in between 72.73% to 76.09% of the forward and reverse input reads were left. This was a significant loss in the amount of data. Trimmomatic was also used with only adapter removal and quality check settings but didn't result in any significant increase in the result. Another adapter removal tool known as Cutadapt (Saeidipour and Bakhshi, 2013) was used to see if this loss could be avoided but using Cutadapt instead resulted in sequences of different lengths. Hence, the output generated by Trimmomatic, having no adapter content, was used for further analysis.

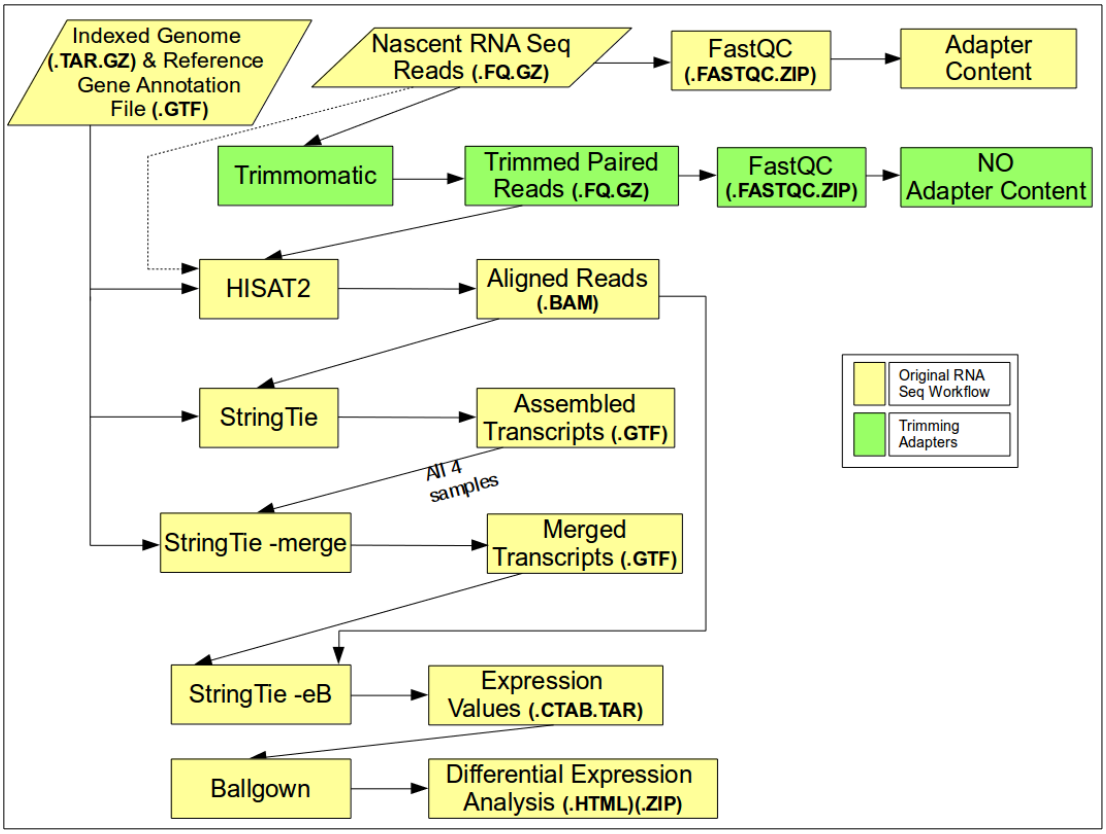


Figure 2: The image shows the workflow of the various steps involved in the quantification of Nascent RNA Sequencing data.

2.1.3 Sequence Alignment:

A number of steps are to be followed to get differential expression between the two conditions. Figure 2 shows the workflow and the various steps involved in the quantification of Nascent RNA Sequencing data. This workflow was the same that was suggested by a Nature protocol paper (Pertea et al., 2016). For alignment of the reads to a indexed reference genome, an updated version of commonly used HISAT (Kim et al., 2015), was used. HISAT which stands for hierarchical indexing for spliced alignment of transcripts, is a tool for mapping RNA-seq reads. HISAT2 with '-dta' parameter (for downstream transcript assembly) was used for the alignment of the indexed reference mouse genome to the four Nascent-RNA Seq samples which generates BAM output files. The Downstream transcriptome assembly (-dta) parameter allows HISAT2 to report alignments for transcript assemblers like

StringTie. HISAT2 then requires longer anchor lengths for de novo discovery of splice sites which leads to fewer alignments with short-anchors and increases the spped and efficiency of transcript assemblers like String Tie. The indexed reference mouse genome build GRCm38 for HISAT2 splice aware aligner was downloaded from ftp://ftp.ccb.jhu.edu/pub/infphilo/hisat2/data/grcm38_tran.tar.gz. The file had HGFM index for reference plus transcripts and was aligned with both the forward and reverse stands for each of the four samples separately. An overall alignment rate in between 95.69% to 96.90% was achieved. This high alignment rate could be because of the fact that the reads had already been trimmed before the alignment and selected for high quality.

2.1.4 Transcript Assembly:

A transcript assembly tool known as StringTie (Pertea et al., 2015) was used for assembling the aligned reads into transcripts with the help of a reference gene annotation file downloaded from Ensembl (Zerbino et al., 2018). String is a commonly used assembler which assembles RNA-Seq alignments into potential transcripts. It uses a novel network flow algorithm and is fast and efficient and hence is widely used for assembling transcripts. The reference annotation file was downloaded from ftp://ftp.ensembl.org/pub/release-91/gtf/mus_musculus/Mus_musculus.GRCm38.91.gtf.gz. This file along with the BAM files generated by HISAT2 is given as an input to StringTie which assembles the potential transcripts.

StringTie with the ‘-merge’ option is then used on the assembled transcripts along with the reference genome are then merged into a single merged transcript file containing a list of non-redundant transcripts using. This merged transcript GTF file along with BAM files containing aligned reads were again given to StringTie with parameters ‘-eB’ to calculate transcript TPM or FPKM values for each sample for the differential expression analysis in .ctab files.

2.1.5 Differential Expression Analysis:

To find any differential expressions between the two conditions (+TAM/KO and -TAM/wt), the .ctab files containing the transcript TPM were given as an input to Ballgown (Frazee et al., 2015). The total number of transcripts and unique genes on merging the assembled transcripts of all the four samples were 135932 and 49563 respectively. A ‘minimum abundance variance across samples’ parameter with a default variance of 1 is used in Ballgown to remove transcripts with low variance (less than 1) which often occur in RNA-Seq data sets. The number of transcripts and unique genes left after filtering were 12718 and 7990 respectively.

	Transcripts	Genes
Number	12718	7990
q-val < 0.05	0	0
q-val < 0.1	0	0
q-val < 0.2	0	0

Table 1: The table shows that no transcript or gene had q-values less than 0.05 or 0.1 or 0.2 indicating that no differentially expressed transcripts and genes are observed at the transcript or at the gene level. q-values are the adjusted p-values accounting for the false discovery rate (FDR).

2.1.6 RNA-Seq Quantification workflow on CGC:

A RNA-Seq Quantification (HISAT2, StringTie) workflow for differential expression analysis based a on Nature protocol paper (Pertea et al., 2016) was also used for finding differential expression between the two conditions. This complete workflow has been implemented on the Cancer Genomics Cloud by Seven Bridges Genomics (http://cgc.sbgenomics.com). The workflow takes an indexed reference genome along with a reference gene annotation file and generates the output for

Ballgown for expression analysis at one go and is very fast and convenient than performing the above steps individually. The total number of transcripts and unique genes on merging the assembled transcripts of all the four samples were 146513 and 53229 respectively. The number of transcripts and unique genes left after the variance filtering were 20751 and 10530 respectively.

	Transcripts	Genes
Number	20751	10530
q-val < 0.05	0	0
q-val < 0.1	0	0
q-val < 0.2	1	0

Table 2: The table shows that only transcript had q-values (q-values are the adjusted p-values accounting for the false discovery rate) less than 0.2. The transcript (with a q-val of 0.1851414149) was ENSMUST00000021296 which comes from a gene Tmem101, a transmembrane protein.

The RNA-Seq Quantification (HISAT2, StringTie) workflow was run again but this time on the raw reads in which Trimmomatic was not run. The reason was to find out whether the DE genes were lost by trimming almost 25% of the data. Ballgown gave that the total number of transcripts and unique genes on merging the assembled transcripts of all the four samples were 146653 and 53387 respectively. The number of transcripts and unique genes left after the variance filtering were 20899 and 10634 respectively.

	Transcripts	Genes
Number	12718	7990
q-val < 0.05	0	0
q-val < 0.1	0	0
q-val < 0.2	0	0

Table 3: The table shows that no transcript or gene had q-values (adjusted p-values accounting for the false discovery rate) less than 0.05 or 0.1 or 0.2 indicating that no differentially expressed transcripts and genes are observed.

2.1.7 Transcript Assembly using MAPS:

To detect rare open reading frames which StringTie may have missed, a new transcript assembly tool known as **MAPS** (M-rna Assembly for ProteogenomicS) (Ma et al., 2018) was also used to assemble the transcripts. MAPS optimizes for both read support and transcriptome diversity, which allows it assemble transcripts containing rare open reading frames. Figure 3 shows the various steps involved in the quantification of Nascent RNA Sequencing data by using StringTie (colored in yellow) as a transcript assembler and MAPS as a transcript assembler (colored in cyan).

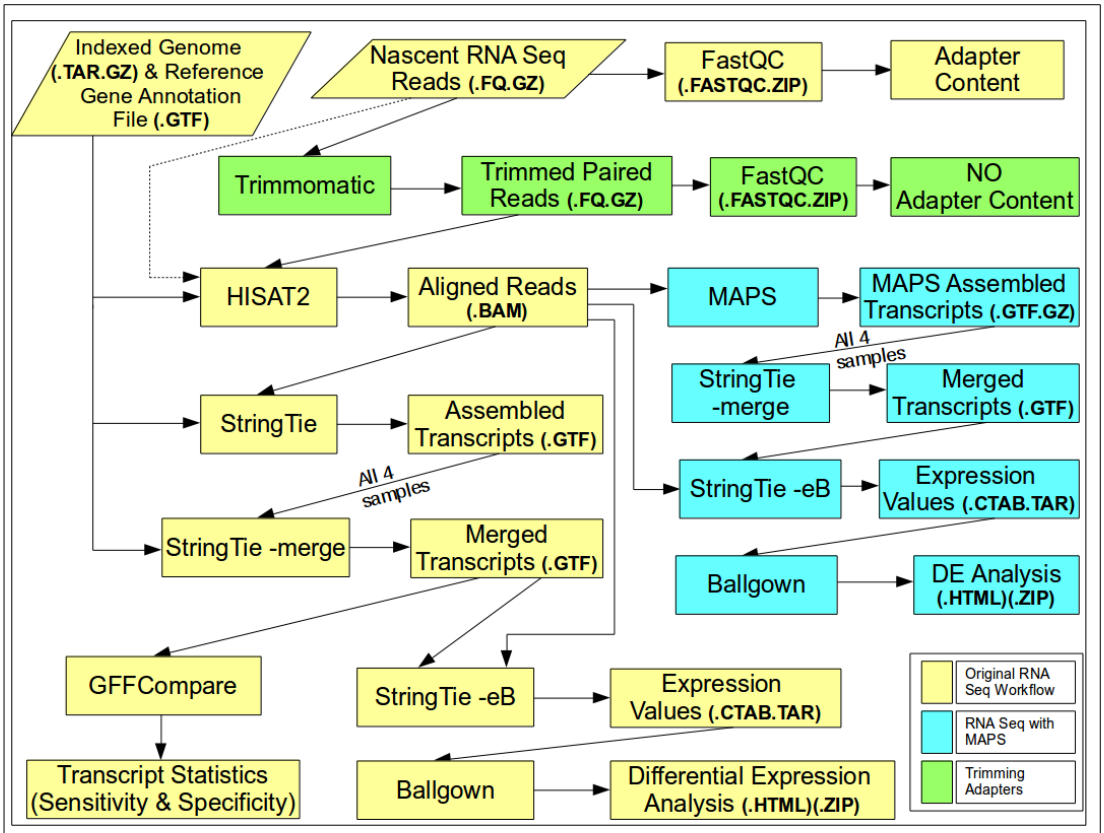


Figure 3: The image shows the various steps involved in the quantification of Nascent RNA Sequencing data by using StringTie (colored in yellow) as a transcript assembler and MAPS as a transcript assembler (colored in cyan).

Ballgown showed that the total number of transcripts and unique genes on merging the assembled transcripts of all the four samples were 163981 and 64352 respectively. This is much more than the number of transcripts or genes assembled

by StringTie. But the number of transcripts and unique genes left after the variance filtering were 8306 and 5931 respectively.

	Transcripts	Genes
Number	8306	5931
q-val < 0.05	0	0
q-val < 0.1	0	0
q-val < 0.2	0	0

Table 4: The table shows that no transcript or gene had q-values (adjusted p-values accounting for the false discovery rate) less than 0.05 or 0.1 or 0.2 indicating that no differentially expressed transcripts and genes are observed even when assembled with MAPS.

2.1.8 Expression of sORFs in mouse:

To find out whether any sORFs are even expressed in mouse, the Mouse sORF coordinates with annotations downloaded from sORFs.org (Olexiouk et al., 2018) database and SmProt (Hao et al., 2017) with annotations were looked at. This database had been created previously in the lab. BEDTools intersect (Quinlan and Hall, 2010) was then used on this database and the merged transcript from StringTie with a a minimum overlap of 99%. BEDTools intersect finds out overlapping regions between two sets of genomes. It was found that the sORF database overlaps with the merged transcript comprising of all the four samples. The total was 5232 sORFs regions that overlapped with the transcripts, with exonic sORFs – 1362, 5UTR – 1316, uORF – 11, 3UTR – 1, lncrna – 837 and unannotated sORFs being 1705 in number.

2.2 Expression of Novel ORFs in various Human tissues:

The next part of the project involved studying the expression of various novel Open Reading Frames like sORFs, altORFs, pseudogenes and de novo genes in multiple human tissue types.

2.2.1 GTEx dataset:

Genotype-Tissue Expression (GTEx) (Lonsdale et al., 2013) is an ongoing project by the GTEx Consortium which aim to build a resource to study tissue wise expression and regulation of various genes. RNA extraction and sequencing on a number of samples was done using Illumina TruSeq polyA selection protocol, hence the RNA sequenced was primarily mRNA for all GTEx tissues. These samples that have been collected from 53 healthy non-diseased tissue sites from across 714 postmortem donors with a total of 11,688 RNA-Seq samples at its current V7 release (Ardlie et al., 2015). The figure 4 shows the distribution of RNA-Seq samples for the 53 human tissue types.

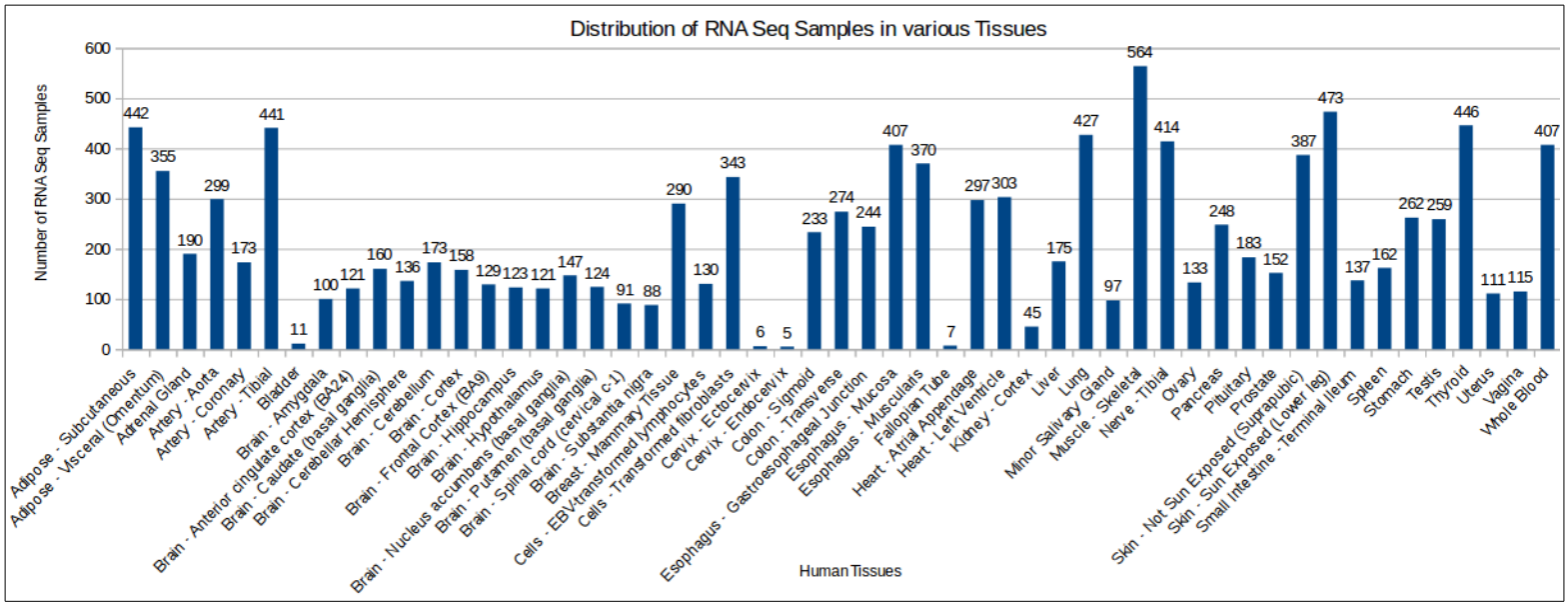


Figure 4: The graph shows the distribution of RNA-Seq samples for the 53 human tissue types.

To find the expression of various novel ORFs in these human tissues, three large files containing various different information about the data were download from the GTEx dataset website. The first file was gencode.v19.transcripts.patched_contigs.gtf which contained the evidence-based annotation of the human genome (GRCh37), version 19 (Ensembl 74) from GENCODE. Since this file had 2,619,449 annotations of CDS, exon, gene, Selenocysteine, start_codon, stop_codon, transcript and UTR, only the transcripts which were 196,520 in number were extracted by sub-selecting those rows in which third column was labeled as ‘transcript’. A ‘chr’ word was added before each entry in the transcript file to make it in alignment with a GTF file for downstream usage. The next file was GTEx_v7_Annotations_SampleAttributesDS.txt which contained the information about which sample ID belong to which tissue types. Each of the 11,688 GTEx samples has a unique ID which doesn’t contain information about its tissue type directly and needs to be mapped to the corresponding human tissue using this file. The third file is GTEx_Analysis_2016-01-15_v7_RSEMv1.2.22_transcript_tpm.txt which contains the expression values in TPM for 196,520 transcripts as row entries for all the 11,688 samples as columns containing a total of 2,296,925,760 data points.

2.2.2 Using LiftOver on the GTEx transcripts:

As mentioned above, about 196,520 transcripts were sub-selected from the bigger gencode.v19.transcripts.patched_contigs.gtf file. The original file which contained the evidence-based annotation of the human genome was (GRCh37) version 19 but all the novel ORF data was (GRCh38) version 39 assembly. Hence, a conversion of these transcripts from hg19 to hg38 assembly was necessary. A commonly used batch coordinate conversion tool called UCSC Liftover (UCSC liftOver) was used to convert the genome annotation file and the genome coordinates from hg19 to hg38 human assembly. LiftOver executible was downloaded from http://hgdownload.cse.ucsc.edu/admin/exe/liftOver.gz along with a hg19ToHg38.over.chain file which has the required liftOver data needed to

convert hg19 (Human Build 37) coordinates to hg38 (Human Build 38). About 195,923 GTEx transcripts were converted from hg19 to hg38.

2.2.3 sORFs and Novel ORFs database:

sORFs stand for small open reading frames which are non canonical open reading frames that are of lengths less than 100 amino acids or 300 nucleotides. The human sORFs coordinates with annotations were downloaded from sORFs.org (Olexiouk et al., 2018). sORFs.org is a database that contains a sORFs which have been computationally predicted along with some sORFs that have been experimentally verified using ribosome profiling. Human sORFs from sORFs.org were downloaded with default filters along with 'GOOD' and 'EXTREME' FLOSS classification. Initially, I downloaded the database from sORFs.org and those entries in which all parameters other than the 'Sorf ID' were exactly the same were removed. It did remove duplicate entries to a certain level but there were still a lot of duplicate entries remaining. It had about 519,698 sORF entries with Sorf ID set in a sequential manner when sorted based on the genomic start site.

Later a proper Novel ORF database was developed by various members of the lab and was uploaded on GitHub. The Main Attributes that were downloaded from Biomart.biobix.be were 'Sorf ID', 'Chromosome', 'Sorf Start', 'Sorf End', 'Strand', 'Spliced Start Parts', 'Spliced Stop Parts', 'Start Codon', 'Sorf Length', 'AA sequence', 'Transcript sequence', 'Biotype', 'Annotation', 'Ensembl Transcript ID' The Genomic coordinates and splicing information were converted into bed12 format for easy processing and then the duplicate entries were removed based on matching genomics coordinates. This database had about **502,056** sORF entries in a GTF file format.

AltORFs stand for alternative open reading frames which are non canonical open reading frames but contain start codons that are different from the canonical open reading frames and therefore altORFs code for an alternative protein (Vanderperre et al., 2012). These altORFs were downloaded from OpenProt (Brunet et al., 2019). OpenProt has all the possible ORFs which are longer than 30 codons,

and has protein conservation, translation and expression as supporting evidence. OpenProt annotates all the proteins that are known and are termed as ‘RefProts’, novel predicted isoforms termed as ‘Isoforms’ and novel predicted proteins from alternative ORFs termed as ‘AltProts’. The following parameter set was used to download the altORFs from (http://www.openprot.org/p/download) with ‘Release’ set to ‘1.3’, ‘Species’ set to ‘Homo Sapiens’, ‘Assembly’ set to ‘GRCg38.p5’, ‘Protein Type’ set to ‘AltProts, Isoforms and Refprots’ and ‘Annotation’ set to ‘Ensembl (GRCh38.83) + RefSeq (GRCh38.p7) + UniProt (2017-09-27)’. An R script was written to select only those entries with Mass Spectrometry evidence with False Discovery Rate of 0.001% and entries with Ribo-Seq evidence with False Discovery Rate of 1%. The final number of entries in the in the curated database was **34,036**.

Pseudogenes are genomic DNA sequences that are related to normal genes. These gene generally have lost at least some functionality, relative to the complete gene, either in terms gene expression in the cell or their ability to code for proteins. De novo genes are those novel coding genes that have evolved from previously noncoding regions, thus generating entirely novel proteins. A similar database for human pseudogenes and de novo genes was created by selecting gene-type as ‘pseudogene’ and ‘de_novo-gene’ respectively. Human pseudogenes database had **15,177** entries whereas the Human de novo genes database had just **41** entries. The creation of the Novel ORF database was a collaborative project and the work was done by various members (Matt Neville, Narendra Meena, Chaitanya Erady and Robin Kohze) of the Prabakaran lab. The Novel ORF database is available on the Prabakaran Lab GitHub page (https://github.com/PrabakaranGroup/nORF-data-prep).

2.2.4 Finding sORFs and Novel ORFs transcripts using BEDTools intersect:

After the GTEx database and the sORF and other Novel ORFs were ready, the next step was to find the transcripts for these sORFs and Novel ORFs. This was done using BEDTools intersect. BEDTools intersect is a tool which finds overlaps

between two sets of genomic features. To find the sORF and other Novel ORF transcripts, BEDTools was used to find overlap between the hg38 GTEx transcript co-ordinates and the human Novel ORFs. We selected those GTEx transcripts that were expressed in sORFs by overlapping the genomic coordinates of the sORFs and GTEx transcripts. The command used was: “bedtools intersect -a gencode.v19.transcripts.patched_contigs_transcript_chr_fixed.gtf -b sorfs.gtf -f 0.99 -s -wo” , where file ‘a’ is the GTEx transcript file and file ‘b’ is the sorfs file in GTF file format. The ‘-f 0.99’ indicates that minimum overlap of 0.99 as a fraction of a is required. The ‘-s’ option is to force ‘strandedness’, i.e. only those hits in the ‘b’ file that overlap with the ‘a’ file on the same strand are reported. The ‘-wo’ option writes the original entries of both ‘a’ and ‘b’ files along with the number of base pairs that overlap between the two files. Only features of A with overlap are reported. Figure 5 shows the gencode.v19.transcripts.patched_contigs_transcript_chr_fixed.gtf as file ‘a’ and the sORFs file as file ‘b’ and the resulting file gives us the sORF transcripts i.e. those genomic coordinates in the GTEx transcripts that lie within the sORF genomic coordinates.

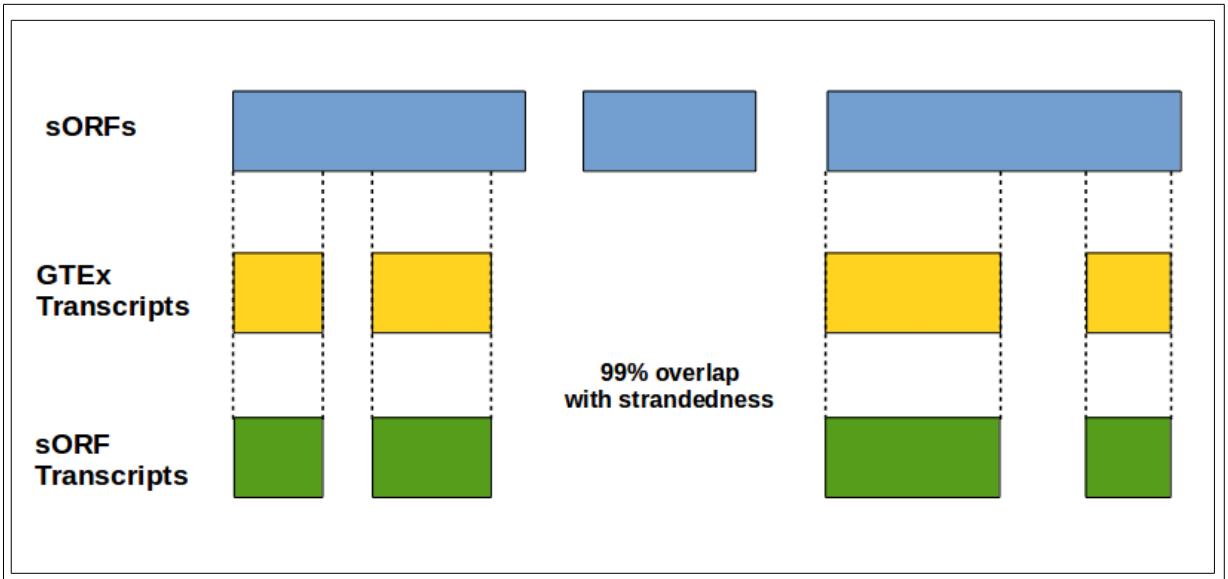


Figure 5: The diagram shows the GTEx transcripts as file ‘a’ and the sORFs file as file ‘b’ and the resulting file gives us the sORF transcripts i.e. those genomic coordinates in the GTEx transcripts that lie within the sORF genomic coordinates. The ‘-f 0.99’ indicates that minimum overlap of 0.99 as a fraction of a is required. The ‘-s’ option is to force ‘strandedness’, i.e. only those hits in the ‘b’ file that overlap with the ‘a’ file on the same strand are reported. The ‘-wo’ option writes the original entries of both ‘a’ and ‘b’ files along with the number of base pairs that overlap between the two files. Only features of A with overlap are reported.

To get the transcripts for the altORFs, pseudogenes and de novo genes databases, the previously mentioned bedtools intersect command (“bedtools intersect -a gencode.v19.transcripts.patched_contigs_transcript_chr_fixed.gtf -b xxxx.gtf -f 0.99 -s -wo”) was executed for altORFs, pseudogenes and de novo genes by replacing the file ‘b’ xxxx.gtf in the bedtools intersect. The sORF transcripts database had **7,361** entries with 4,255 unique sORFs mapping to 2,246 unique GTEx transcripts. The altORF transcripts database had **2,304** entries with altORFs mapping to 1,858 unique GTEx transcripts. The pseudogenes transcripts database had **17,668** entries with pseudogenes mapping to 17,521 unique GTEx transcripts. The de novo genes transcripts database had **193** entries with pseudogenes mapping to all unique GTEx transcripts.

2.2.5 Mapping expression of sORFs and Novel ORFs transcripts in the GTEx dataset:

After the genomic location and and the transcript IDs of the sORFs and other Novel ORFs transcripts were obtained, they were mapped to their TPM expression level using multiple Python codes. A list of all samples was segregated according to their tissues types from the file GTEx_v7_Annotations_SampleAttributesDS.txt. The first Python code selects the TPM expression value of only the sORFs and other Novel ORFs transcripts from the large dataset of all TPM expression values for all transcripts. The second Python code find out the column number of the GTEx samples that map for a particular tissue and write the output into a separate file each for the 53 tissue types. The third Python code finally takes the input of the selected transcript expression from the first code and the column number of the particular samples and give the expression level for each of the 53 human tissues. For each of the tissues, only the mean and standard deviation over all the samples is reported for every sORFs and other Novel ORFs transcript.

3. Results and Discussion

3.1 Analysis of the mouse Nascent RNA-seq data:

3.1.1 HISAT2 overall alignment rate:

When the trimmed reads were aligned to a indexed reference genome using HISAT2, a very high overall alignment rate (96%) was achieved when compared to generally reported alignment rate in the field (70% - 80%). This can be explained in the Table 5 which shows the percentage of reads remaining, the HISAT2 overall alignment rates and the effective alignment rate after various pre-processing steps using Trimmomatic at different parameters and settings. The effective alignment rate is defined as the multiplication of the percentage of the remaining preprocessing reads with the HISAT2 overall alignment rate. The effective alignment rate is very similar to what is generally reported alignment rate in the RNA-Seq field.

	Trimmomatic Settings		
	Default parameters (Adapter removal and Quality Check)	**Default parameters (Adapter removal)**	**No preprocessing**
% of reads remaining after Preprocessing	76.09%	77.92%	100.00%
HISAT2 Overall alignment rate	96.68%	95.56%	80.28%
Effective Alignment Rate	73.56%	74.46%	80.28%

Table 5: This table shows the percentage of reads remaining after various pre-processing steps using Trimmomatic at different settings, the HISAT2 overall alignment rates and the effective alignment rate.

3.1.2 Clustering the two conditions:

Although, it is difficult to perform clustering based on just four samples, the Cluster dendogram of the DE analysis gave a unique observation. The Cluster dendogram tells how ‘close’ or ‘distant’ two samples. All the samples are clustered using the Euclidean distance between log(FPKM+1) values on the gene level (FPKM is Fragments Per Kilobase Million). Only those genes that pass the filtering step are included in the diagram. The input grouping for Ballgown was on the samples being replicates i.e. NEBNext02 and NEBNext04 (+TAM samples) were in the group 1 and the NEBNext06 and NEBNext12 (-TAM samples) were in the group 0. Figure 6 shows the Cluster dendogram for one of the DE analysis. It shows that the replicates are not clustered together based on their expression values. It indicates that there is more variance between the replicates than in between the groups. The similar Cluster dendogram is observed across all the various differential expression analysis methods that were out.

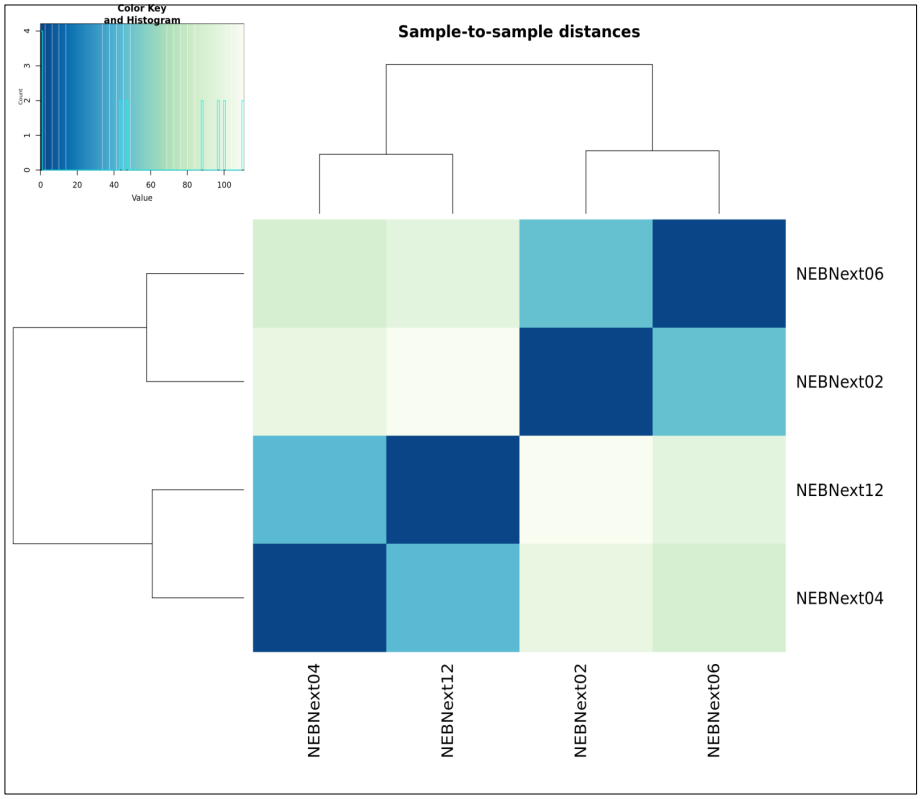


Figure 6: The figure shows the Cluster dendogram for one of the DE analysis methods. All the samples in the Cluster dendogram are clustered using the Euclidean distance between log(FPKM+1) values on the gene level. The diagram shows that the replicates are not clustered together based on their expression values. It indicates that there is more variance between the replicates than in between the groups. This Cluster dendogram is observed across all the various differential expression analysis methods that were out.

3.1.2 Expression values of H2afz and H2afv genes:

When the expression values in the data were searched for H2afz and H2afv genes, it was observed that there was difference in the expression in between the two conditions. It appears that the Knock-out had worked but not completely. As observed earlier, the variance between the replicates was also significantly high. Figure 7 shows the expression values in TPM (Transcripts Per Kilobase Million) for the H2afz and H2afv genes. The -TAM samples which are the wild type have a higher expression values when compared to the +TAM samples which are knock-outs. The knock-outs still have a large amount of expression levels for H2afz gene.

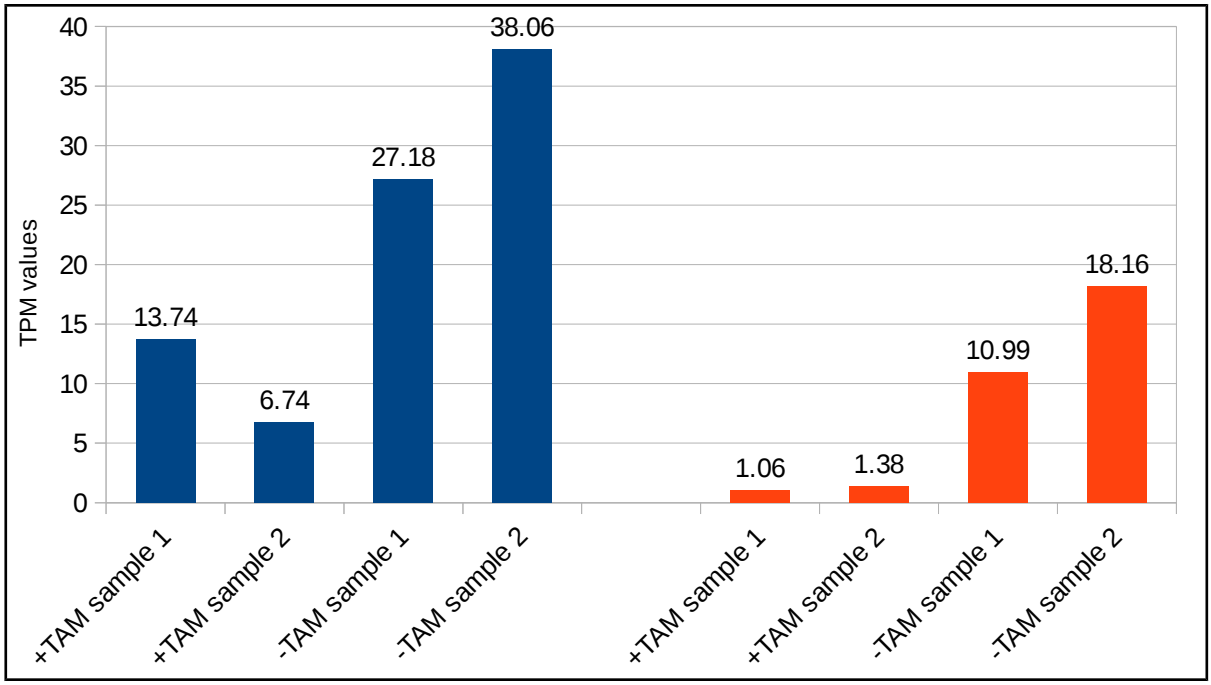
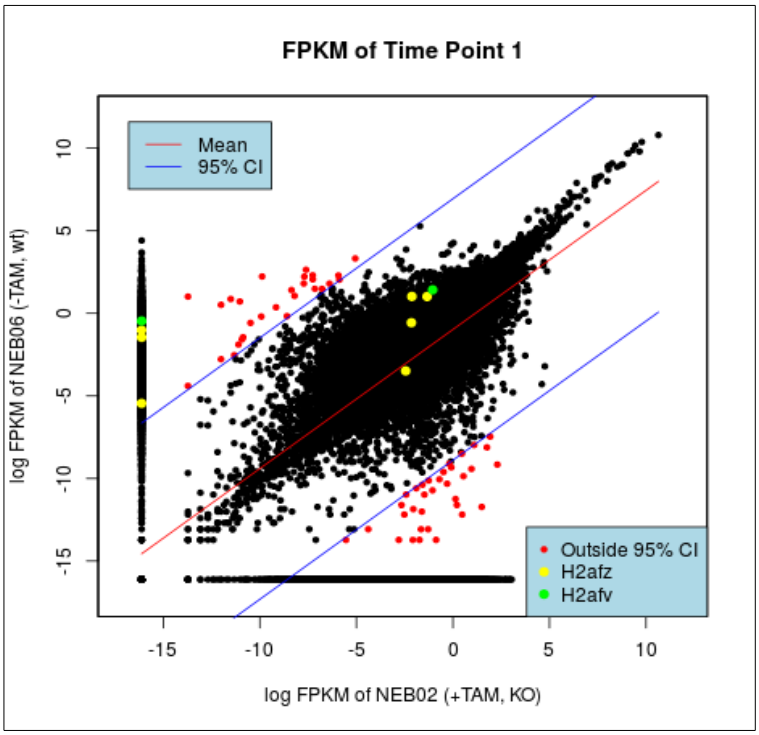


Figure 7: The figure shows the expression values in TPM (Transcripts Per Kilobase Million) for the H2afz and H2afv genes. The -TAM samples which are the wild type which have a higher expression values when compared to the +TAM samples which are knock-outs. The knock-outs still have a large amount of expression levels for H2afz gene.

3.1.3 Analysis of transcript FPKM expression values:

Further, the transcript FPKM expression values for the four samples were analyzed using R. Figure 8(a) shows the plot for log(FPKM+0.0000001) of NEB06 (-TAM) against NEB02 (+TAM) for the first replicate or time point 1 and Figure 8(b) shows the plot for log(FPKM+0.0000001) of NEB12 (-TAM) against NEB04 (+TAM) for the second replicate or time point 2. The black data points are the expression values of 146513 transcripts which were generated by the RNA-Seq Quantification (HISAT2, StringTie) workflow for DE analysis. The yellow and green data points are the expression values of H2afz and H2afv respectively. The red line is the mean regression line and the blue lines are the 95% confidence interval lines. The red data points are the points that lie outside the 95% confidence interval. The points on the extreme left & bottom are those points, whose FPKM = 0 and since this is a log plot and 0.0000001 has been added to the FPKM and ln(0.0000001) equals −16.118095651. Hence the points are on the extremes. The region outside the 95% confidence interval (blue lines) represents the transcripts that are more likely to be differentially expressed between the two conditions as they will have larger variance. As seen, all the transcripts for the H2afz and H2afv genes lie within the 95% confidence interval region and hence are less likely to be differentially expressed. There are only 10 transcripts corresponding to 9 genes which lie outside the 95% interval and are common in the two graphs which could be potentially differentially expressed.

The significant overlapping of the sORF database with the merged transcript comprising of all the four sample done by BEDTools intersect is an evidence of transcription from 'noncoding regions' but these regions are not differentially expressed. Previous studies may not have found these 'novel' transcriptional products because most of the RNA-Seq studies are based on polyA enriched RNAs and this study is based on total RNA.



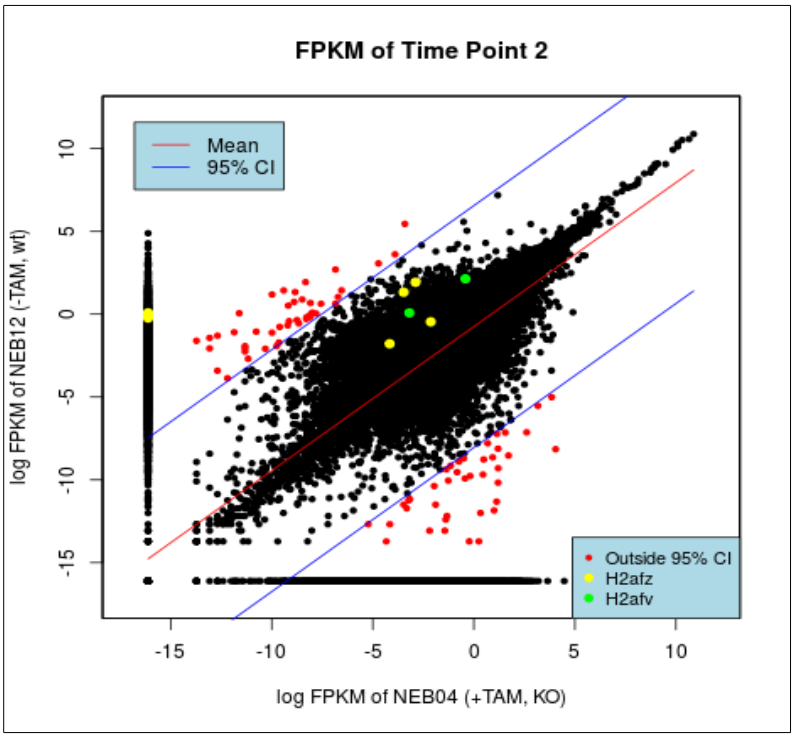


Figure 8(a): shows the plot for log(FPKM+0.0000001) of NEB06 (-TAM) against NEB02 (+TAM) for the first replicate or time point 1.

Figure 8(b) shows the plot for log(FPKM+0.0000001) of NEB12 (-TAM) against NEB04 (+TAM) for the second replicate or time point 2.

The black data points are the expression values of 146513 transcripts which were generated by the RNA-Seq Quantification (HISAT2, StringTie) workflow for DE analysis. The yellow and green data points are the expression values of H2afz and H2afv respectively. The red line is the mean regression line and the blue lines are the 95% confidence interval lines. The red data points are the points that lie outside the 95% confidence interval. The points on the extreme left & bottom are those points, whose FPKM = 0 and since this is a log plot and 0.0000001 has been added to the FPKM and ln(0.0000001) equals −16.118095651. Hence the points are on the extremes.

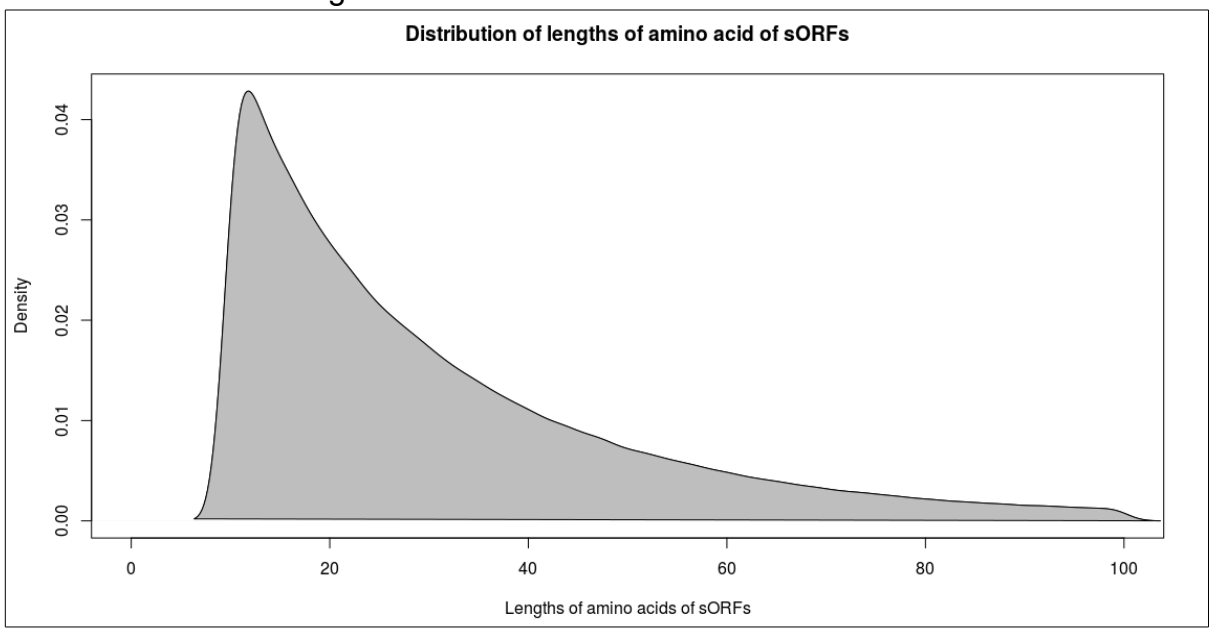
As seen, all the transcripts for the H2afz and H2afv genes lie within the 95% confidence interval which mean they don't have a high deviation from the mean.

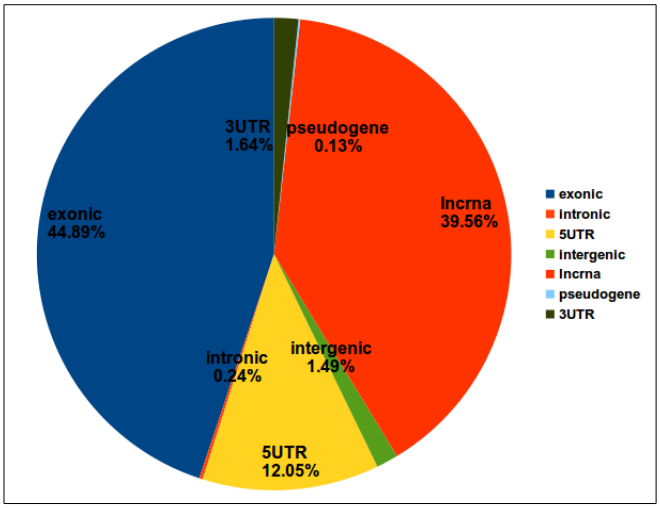
Various noncoding transcriptional events have been claimed to be 'transcriptional noise'. It is our claim that a mutant histone protein leaves the enhancers and chromatin open for non-specific transcriptions to happen and hence one should find random ORFs transcribed which we do not find in this case. We have found no evidence of any differential expression, which supports our claim that these events are not 'biological noise'.

3.2 Analysis of expression levels of Novel ORFs in normal human tissues:

3.2.1 Analysis of the sORF database:

The unfiltered sORF database when download from sorfs.org had about 2182379 entries. A lot of those entries were duplicates and when they were removed, about 502,056 sORF were left in the final database that we used for downstream analysis. Figure 9(a) shows a plot of the length distribution of the sORFs in the sORF database. The length of sORFs in the database various from 10 to 100 amino acids, as expected. Figure 9(b) shows the distribution of the various sORF annotations. The three most abundant annotations are exonic sORFs, i.e. those sORFs which are located in the exonic part of a gene comprising about 45% of the total sORFs, followed by about 40% sORFs located in lncRNAs and about 12% located in 5'UTRs of a gene.





(clockwise from top) Figure 9(a) shows a plot for distribution of the length of amino acids of the sORFs in the curated sORF database. The length of sORFs in the database various from 10 to 100 amino acids. Figure 9(b) shows the distribution of the various sORF annotations showing exonic sORFs (45%), lncRNA (40%) and 5'UTR (12%).

3.2.2 Analysis of the sORF data and GTEx transcripts:

Figure 10 shows the distribution of the lengths of sORF transcripts that were overlapped with the GTEx transcripts. This distribution is not similar to the distribution as shown in Figure 9(a). This indicates that although there are a higher number of sORFs with amino acid length less than 30, there are significant number of sORFs that are even larger in length i.e. greater than 80 amino acid as well.

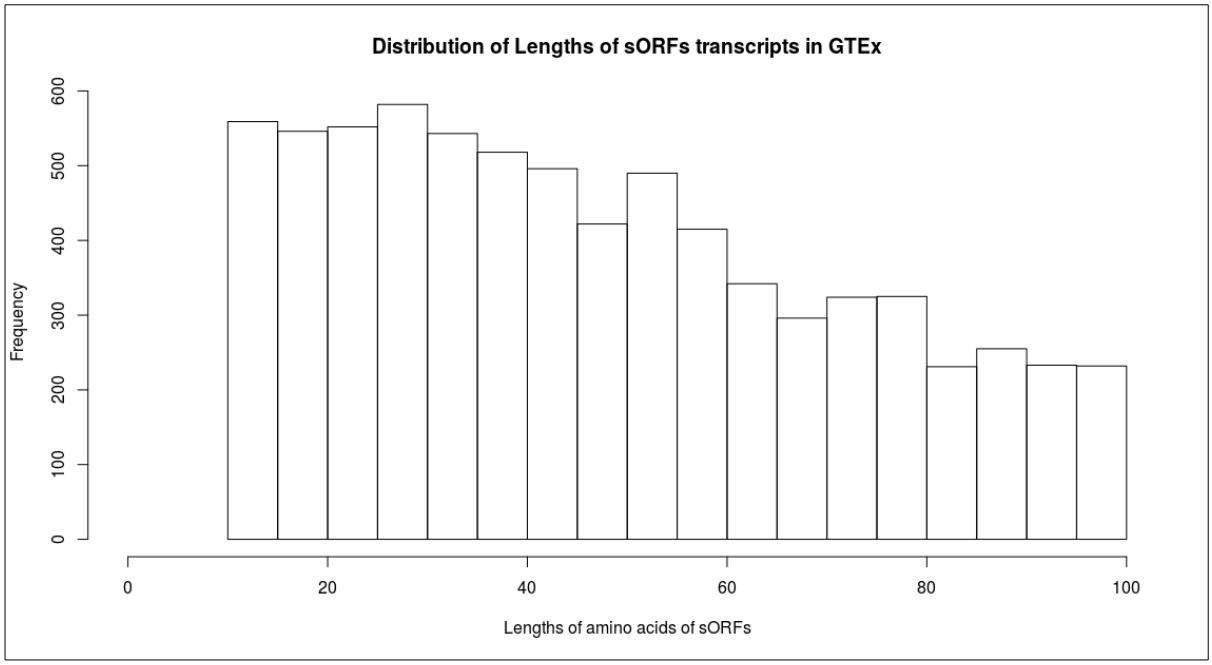


Figure 10: The plot shows the distribution of the lengths of sORF transcripts that were overlapped with the GTEx transcripts. This indicates that although there are a higher number of sORFs with amino acid length less than 30, there are significant number of sORFs that are even larger in length i.e. greater than 80 amino acid as well.

3.2.3 Analysis of expression levels of the sORF and Novel ORFs data and GTEx transcripts for human tissues:

The expression levels for each of the 53 human tissues were generated by using multiple Python scripts as mentioned previously. For each of the tissues, only

the mean and standard deviation over all the samples was reported for every sORFs and other Novel ORFs transcripts. The mean value of these sORFs and other Novel ORFs transcripts was taken and added a small value of 0.001 to this mean. Then the natural log of this value was taken and was plotted as a box plot. Next, we wanted to compare the expression of these Novel ORFs to the expression levels of rest of the protein coding transcripts. For this, all the protein coding transcripts were sub selected from the the GTEx database for all the tissue and their expression levels were found following the same steps as to get the expression levels of the novel ORFs using the Python codes. Then, for each tissue, mean expression values were taken for every transcripts of these novel ORFs and after adding a small value of 0.001, natural log was taken of these mean values. Then the first quartile, second quartile (median) and the third quartile of these log values of the protein coding transcripts were overlaid with the expression levels of the sORFs, altORFs, pseudogenes and the de novo genes transcripts. Figure 11(a) shows the natural log of mean over every sample for each of the tissue expression levels of sORFs transcripts and the natural log of mean over every sample for each of the tissue expression levels of all protein coding transcripts for each of the 53 tissues. Figure 11(b) shows the same plot for altORFs, Figure 11(c) for pseudogenes and Figure 11(d) for de novo genes.

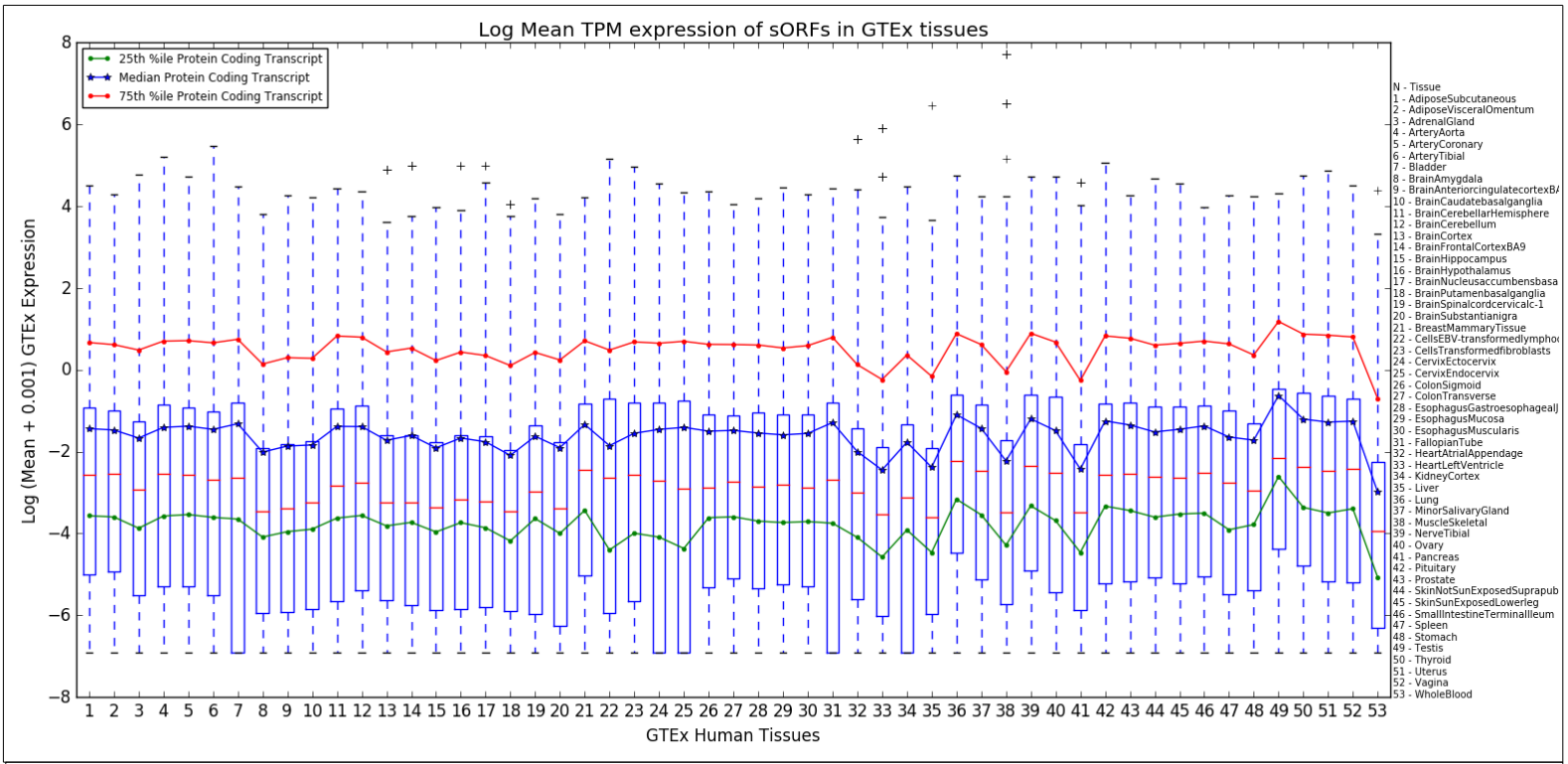


Figure 11(a): The figure shows the natural log of mean over every sample for each of the tissue expression levels of sORF transcripts and the natural log of mean over every sample for each of the tissue expression levels of all protein coding transcripts for each of the 53 tissues.

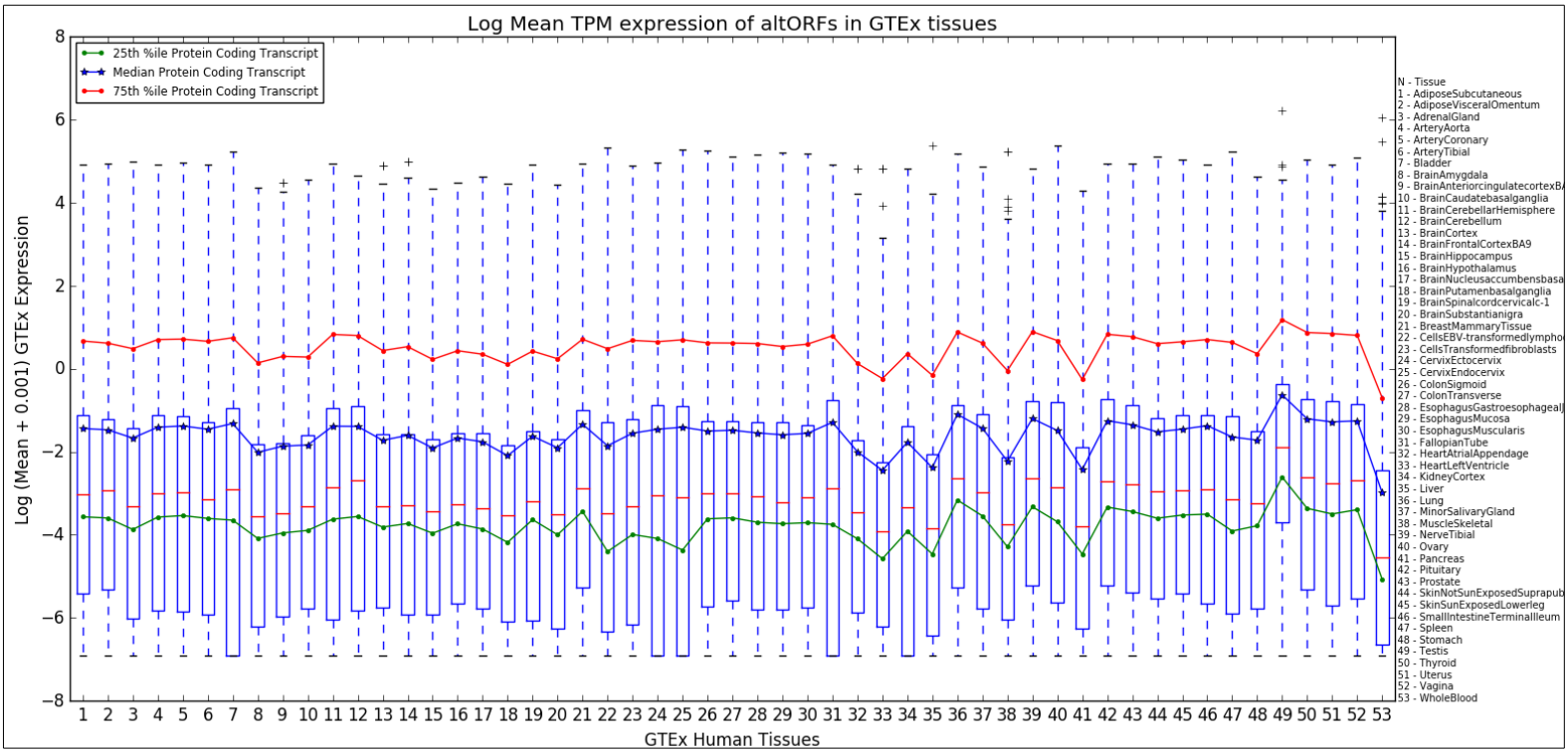


Figure 11(b): The figure shows the natural log of mean over every sample for each of the tissue expression levels of altORF transcripts and the natural log of mean over every sample for each of the tissue expression levels of all protein coding transcripts for each of the 53 tissues.

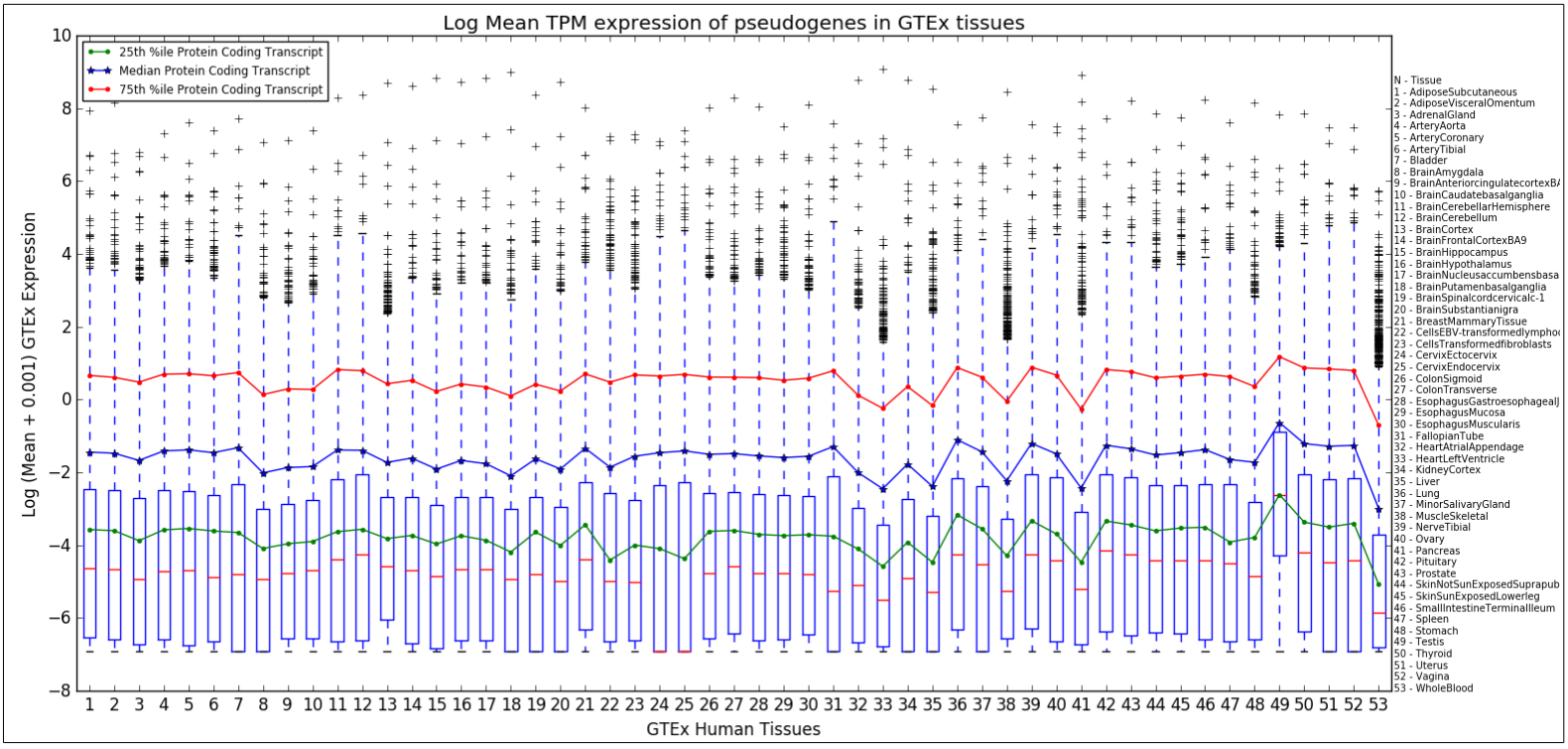


Figure 11(c): The figure shows the natural log of mean over every sample for each of the tissue expression levels of pseudogene transcripts and the natural log of mean over every sample for each of the tissue expression levels of all protein coding transcripts for each of the 53 tissues.

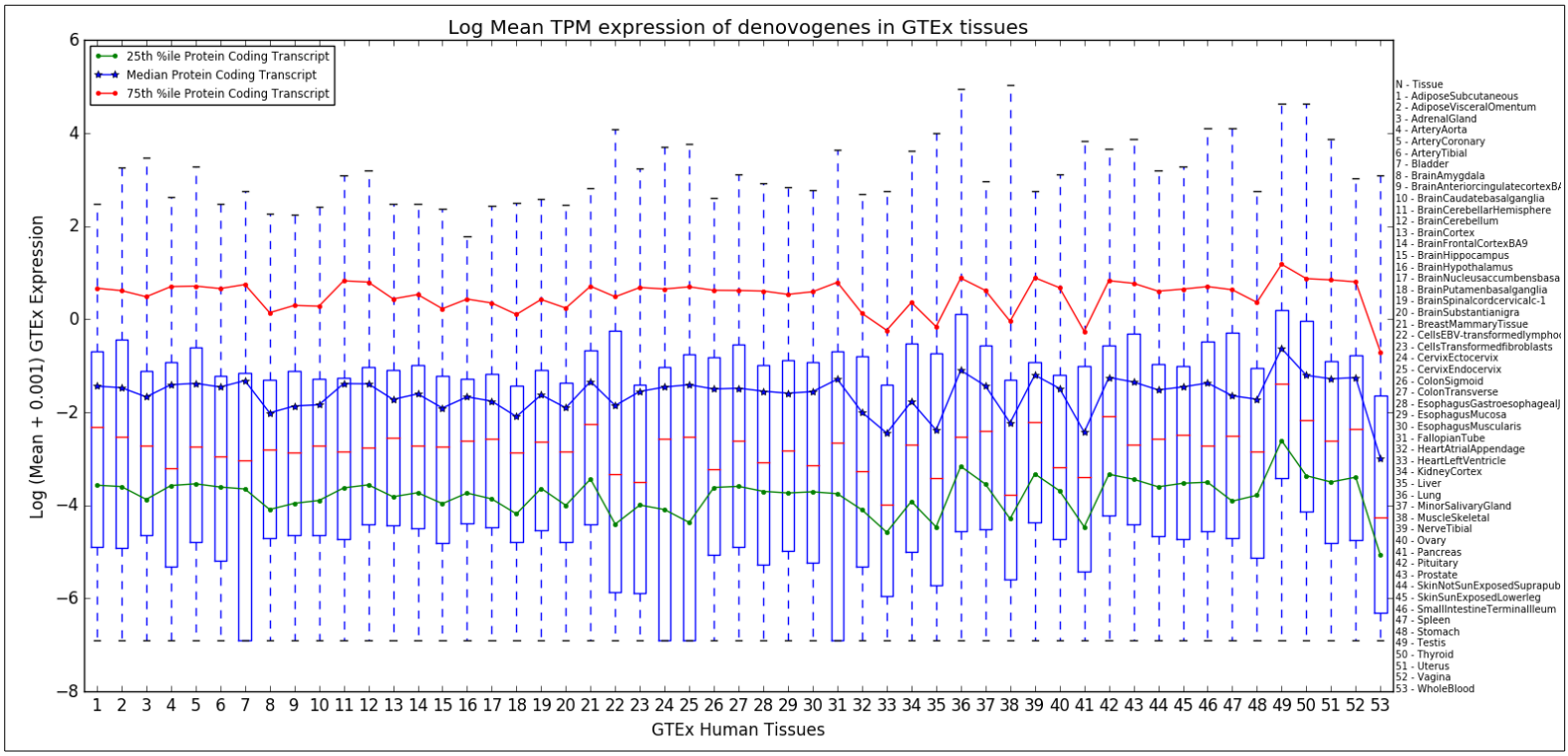


Figure 11(d): The figure shows the natural log of mean over every sample for each of the tissue expression levels of de novo gene transcripts and the natural log of mean over every sample for each of the tissue expression levels of all protein coding transcripts for each of the 53 tissues.

All the above figures show the expression levels of various non coding regions like sORFs, altORFs, pseudogenes and de novo genes in all the 53 healthy tissue types. The transcripts of these novel ORFs in the GTEx database for healthy human tissues show a relatively lower mean amount of expression in terms of TPM (transcripts per kilobase million) when compared to transcripts to all the possible 81,575 coding transcripts in the entire GTEx database for all the 53 tissue types.

3.2.4 Analysis of the tissue-wise expression of the Novel ORF transcripts in the GTEx dataset:

Out of the total detected 7,361 sORF transcripts, 2,304 altORF transcripts, 17,668 pseudogene transcripts and 193 de novo gene transcripts, not all are expressed evenly in every tissue. Here in Figure 12(a)-(d), we look at the number of sORF transcripts and other Novel ORF transcripts that are expressed in each of the 53 tissue types present in the GTEx dataset.

The sum of the number of sORF transcripts that had a non-zero mean expression, and hence were expressed in all the GTEx human tissues was 301,943 which have been distributed by the tissue type in Figure 12(a). The sum of the number of altORF transcripts that had a non-zero mean expression was 96,138 which have been distributed by the tissue type in Figure 12(b).

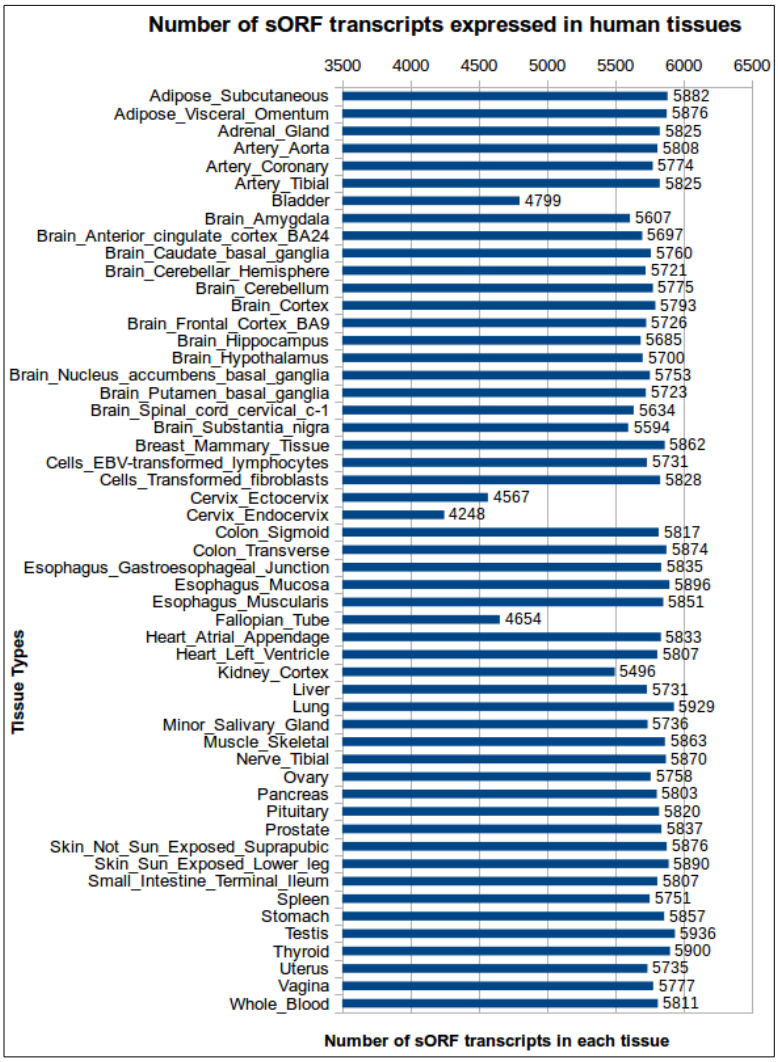


Figure 12(a): The figure shows the number of sORF transcripts expressed for each of the 53 tissues.

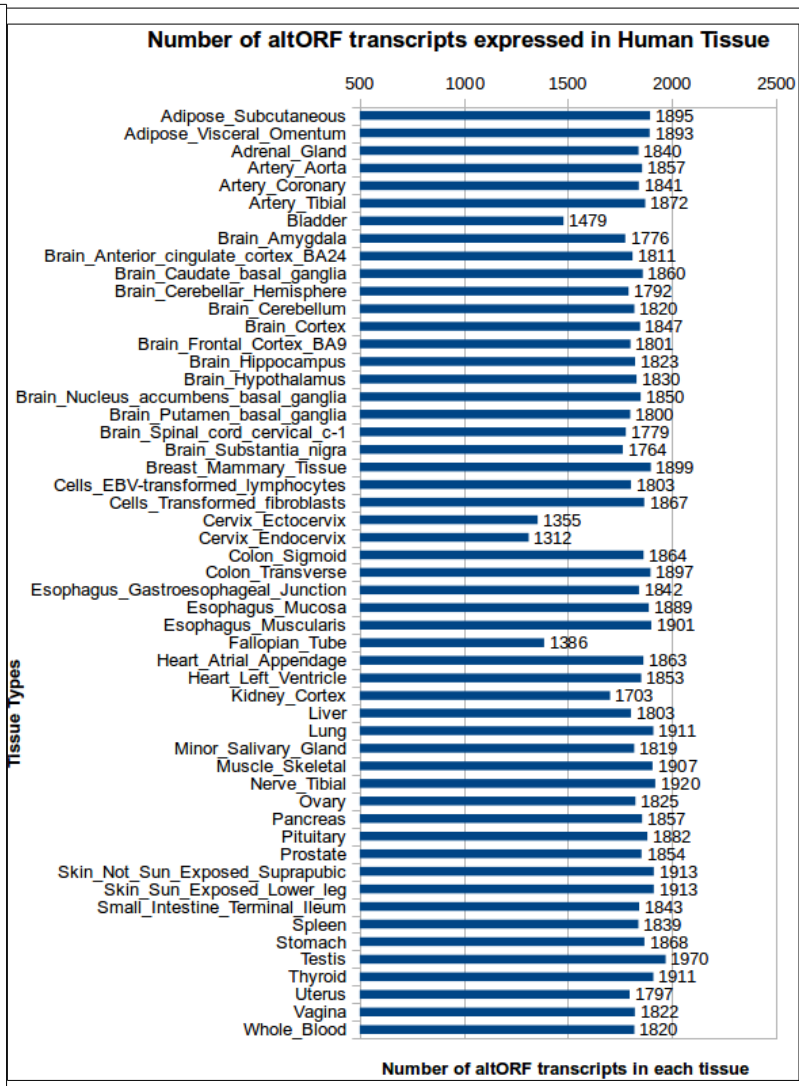


Figure 12(b): The figure shows the number of altORF transcripts expressed for each of the 53 tissues.

The sum of the number of pseudogene transcripts and de novo gene transcripts that had a non-zero mean expression, and hence were expressed in all the GTEx human tissues was 713,464 and 9,161 respectively which have been distributed by the tissue type in Figure 12(c) and Figure 12(d) respectively.

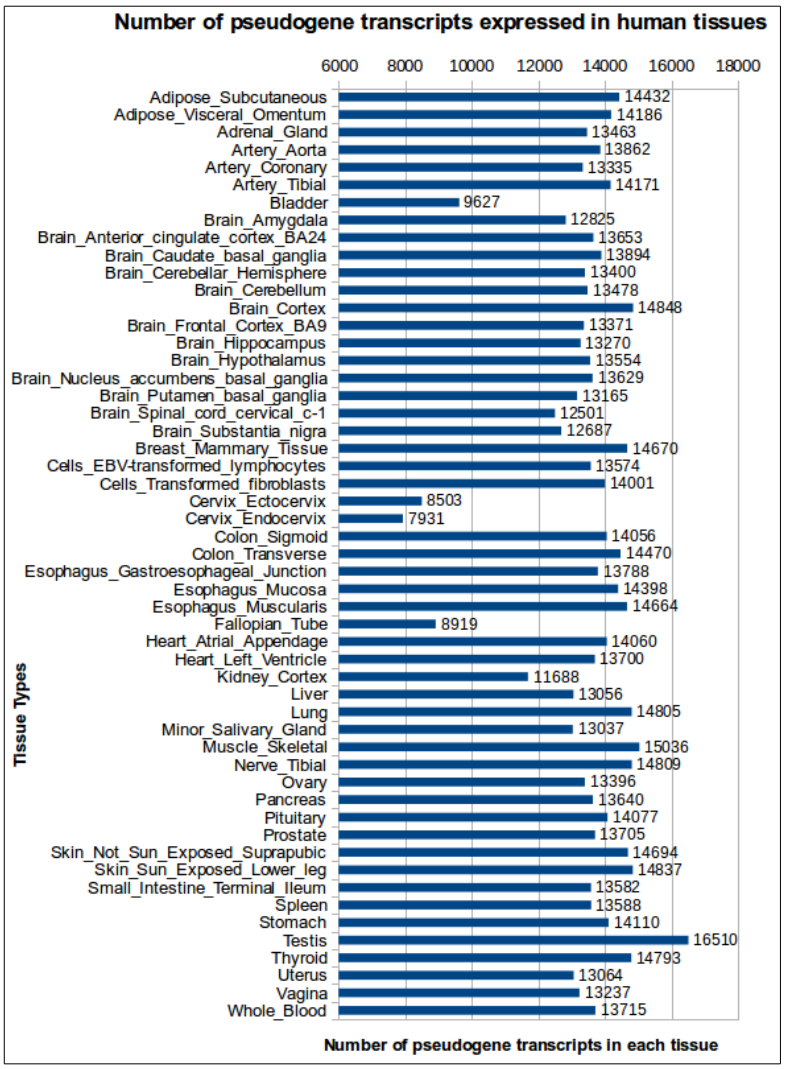


Figure 12(c): The figure shows the number of pseudogene transcripts expressed for each of the 53 tissues.

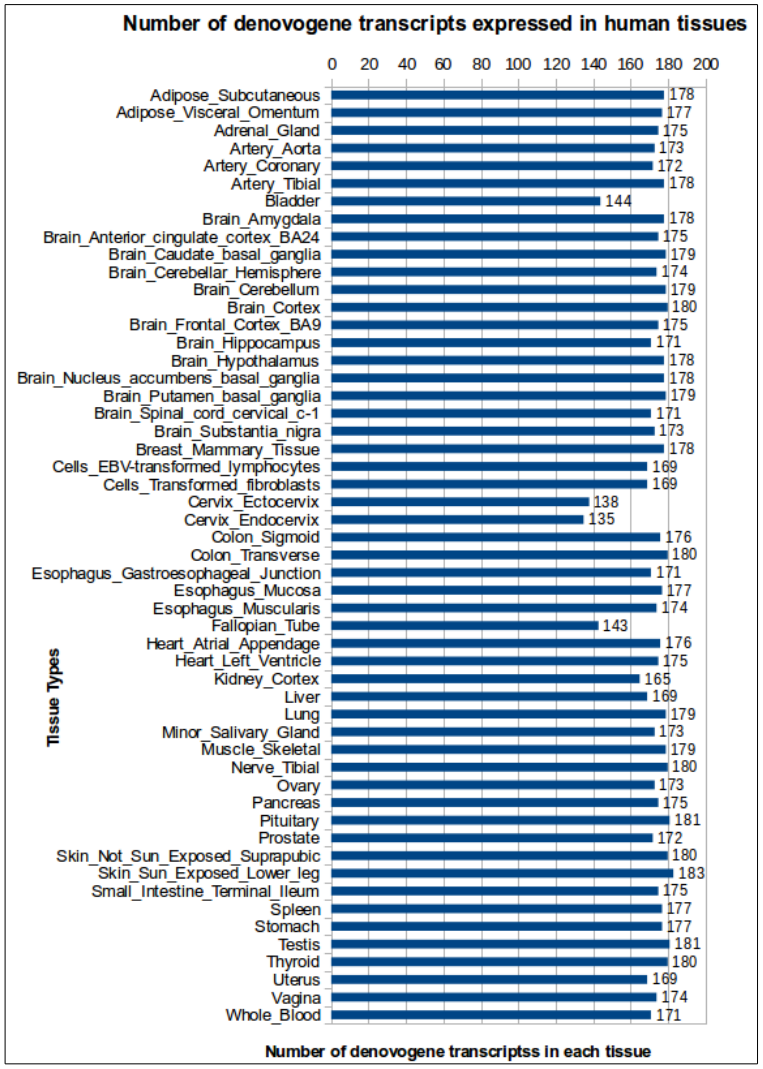


Figure 12(d): The figure shows the number of de novo gene transcripts expressed for each of the 53 tissues.

The above graphs in Figures 12(a) - (d) show the distribution of number of various Novel ORF transcripts like sORF, altORF, pseudogene and de novo gene transcripts in all the 53 tissue types present in the GTEx datasets. We observe that the Novel ORF transcripts are expressed lower in Bladder, Cervix and Fallopian Tube tissue samples. This could be due to the fact that these tissue have less number of samples in the GTEx data itself.

3.2.5 Analysis of the tissue-wise expression of unique Novel ORF transcripts in the GTEx dataset in the human genome:

We also looked at the number of unique sORF transcripts and other Novel ORF transcripts that are expressed in each of the 53 tissue types present in the GTEx dataset. A unique Novel ORF transcript is an ORF transcript that is expressed only in one tissue types and not expressed in any other tissue. There are a various Novel ORF transcripts that are expressed in two or three tissues only but those haven't considered here.

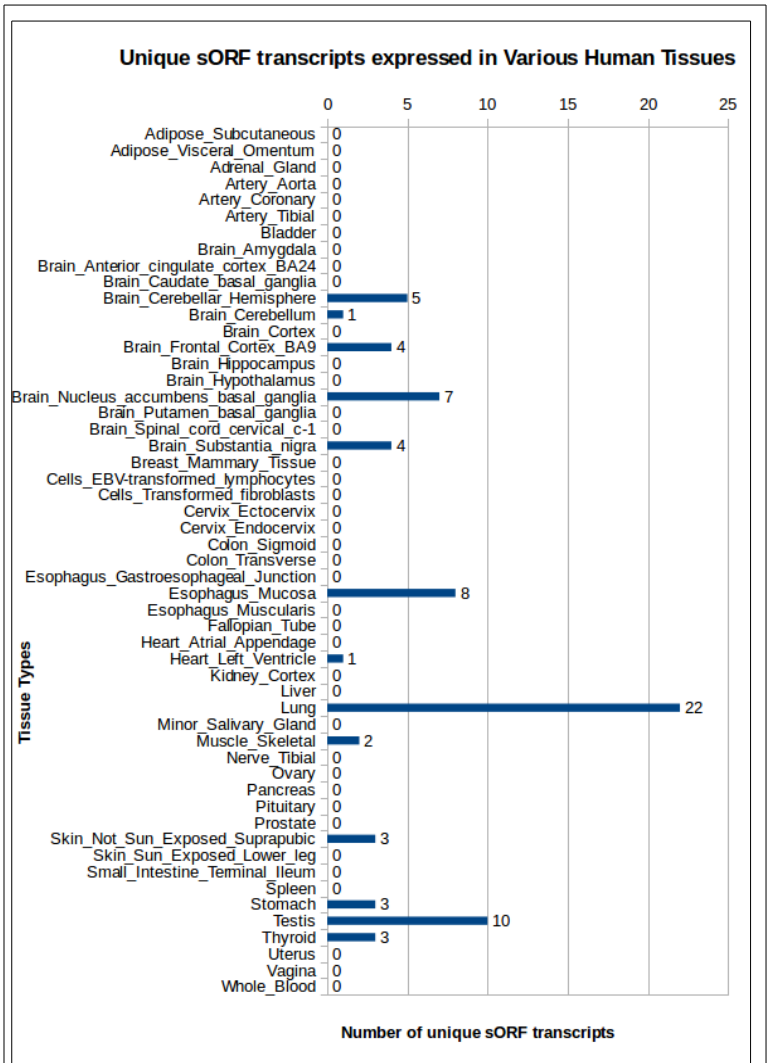


Figure 13(a): The figure shows the number of uniquely expressed sORF transcripts for each of the 53 tissues.

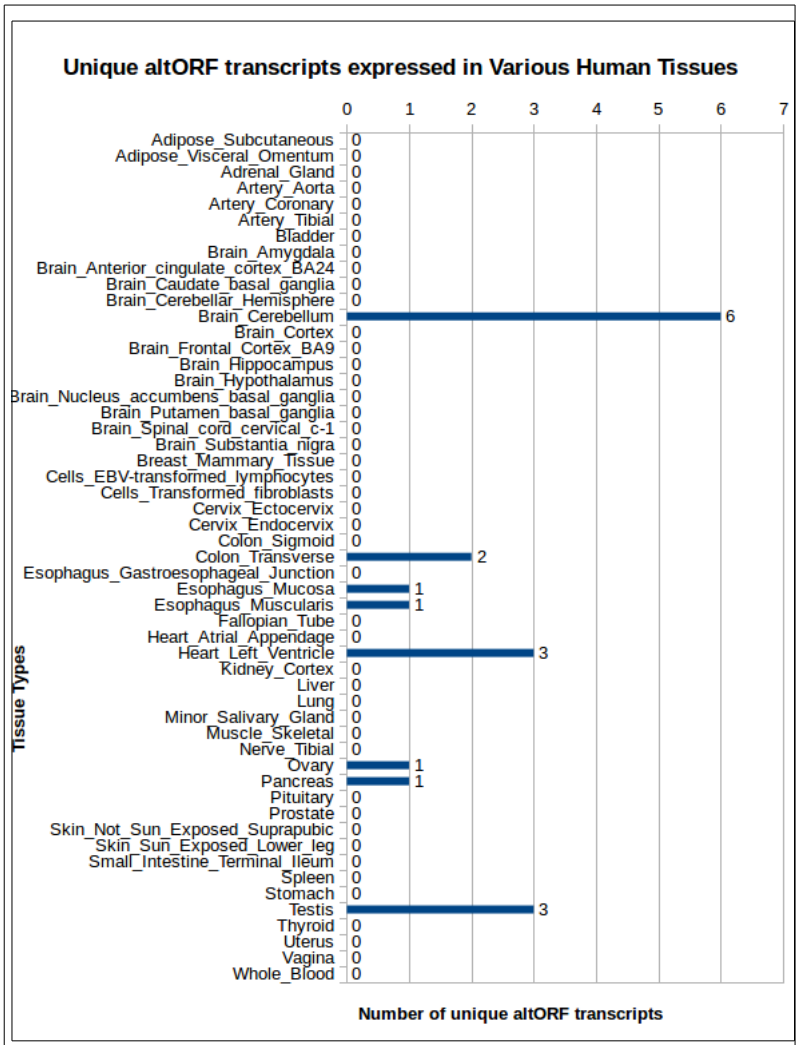


Figure 13(b): The figure shows the number of uniquely expressed altORF transcripts for each of the 53 tissues.

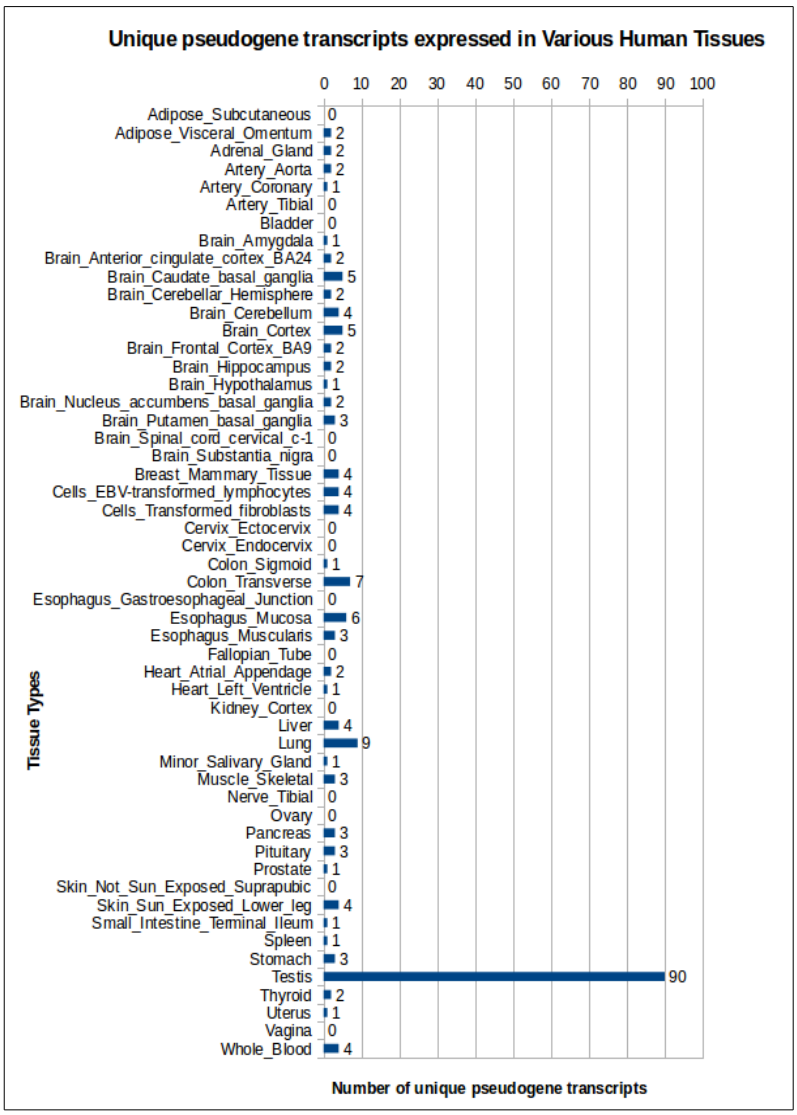


Figure 13(c): The figure shows the number of uniquely expressed pseudogene transcripts for each of the 53 tissues.

The total number of unique sORF transcripts that were expressed in all the tissue types present in the GTEx dataset was **73** which have been distributed by the tissue type in Figure 13(a). The total number of unique altORF transcripts that were expressed in all the tissue types was **18** which have been distributed by the tissue type in Figure 13(b). The total number of unique pseudogene transcripts that were expressed in all the tissue types present in the GTEx dataset was **198** which have been distributed by the tissue type in Figure 13(c).

There were **no** uniquely expressed de novo gene transcripts in any of the tissue types.The tissue type with the highest number of uniquely expressed sORFs was ‘Lung’ followed by ‘Testis’. The tissue type with the highest number of uniquely expressed altORFs was ‘Brain Cerebellum’ followed by ‘Heart Left Ventricle’ and ‘Testis’. The tissue type with the highest number of uniquely expressed pseudogenes was ‘Testis’ followed by ‘Lung’.

4. Conclusion and Future Directions:

Various noncoding transcriptional events have been claimed to be ‘transcriptional noise’ but to show otherwise, we investigate differential expression between mutant histone variant and wild type samples with two replicate each. The mutant samples were NEBNext02, NEBNext04 (denoted by +TAM) which were treated with Tamoxifen, a drug which leads to the deletion of H2afz and H2afv genes and the wild type samples were NEBNext06, NEBNext12 (denoted by -TAM). These were nuclear RNA samples from *Mus musculus* Embryonic Fibroblasts (MEFs) which had sequenced using the Illumina HiSeq 4000. We do not find any significant differential expression levels of genes or transcripts in between the two conditions. It was our claim that a mutant histone protein leaves the enhancers and chromatin open for non-specific transcriptions to happen, therefore increasing noisy transcription, and hence one should find random ORFs transcribed which we do not find in this case. We found no evidence of any differential expression in the known and sORF transcripts between the two conditions, which supports our claim that these events are not ‘biological noise’. We also found that there was no differentially expressed

After showing in *Mus musculus* that transcription of sORFs is not due to noisy transcriptional events, we further strengthen this argument by finding the expression of sORFs and other novel ORFs in various tissues types. The GTEx dataset, being a huge collection of mRNA data from normal human tissues, helped us understand and quantify the expression of sORFs and other Novel ORFs at a large scale. We also look at the expression levels of multiple novel ORFs like sORFs, altORFs, pseudogenes and de novo genes in 53 healthy tissue types. We find transcripts of novel ORFs in the GTEx database and then compared these transcripts to all the possible 81,575 coding transcripts in the entire database for all the 53 tissue types. The transcripts of these novel ORFs in the GTEx database for healthy human tissues show a relatively lower mean amount of expression in terms of TPM (transcripts per kilobase million) when compared to transcripts to all the possible 81,575 coding transcripts in the entire GTEx database for all the 53 tissue types. We

also find the total number of novel ORFs that are expressed in various GTEx tissue types and also find the uniquely expressed novel ORFs in each tissue type.

To find functional significance of sORFs and other Novel ORFs, we are also looking whether there are any mutations that map to the novel ORFs that are expressed in various tissue types. We will use the HGMD (Human Gene Mutation Database) (Stenson et al., 2009) and the COSMIC (Catalogue Of Somatic Mutations In Cancer) (Tate et al., 2019) databases for characterizing mutations in the novel ORFs. We are also looking whether these Novel ORF transcripts have a differential expression in cancer data using the TCGA (The Cancer Genome Atlas) (Weinstein et al., 2013) dataset.

In the future, we also want to look into the Transcription Factor Binding Site and promoter regions in various novel transcripts which are uniquely expressed in the GTEx tissues and try to figure out a reason for their selective expression. We also want to predict structure for the various sORFs and other Novel ORFs that have a selective expression using their amino acid sequences.

5. References

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Babraham Bioinforma.

Ardlie, K.G., DeLuca, D.S., Segrè, A. V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-.).

Bazin, J., Baerenfaller, K., Gosai, S.J., Gregory, B.D., Crespi, M., and Bailey-Serres, J. (2017). Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. Proc. Natl. Acad. Sci. 201708433.

Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279–290.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Brunet, M.A., Brunelle, M., Lucier, J.F., Delcourt, V., Levesque, M., Grenier, F., Samandi, S., Leblanc, S., Aguilar, J.D., Dufour, P., et al. (2019). OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res.

Chew, G.-L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F., and Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development *140*, 2828–2834.

Frazee, A.C., Pertea, G., Jaffe, A.E., Langmead, B., Salzberg, S.L., and Leek, J.T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat. Biotechnol. *33*, 243–246.

Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., Zhang, B., Zhang, D., Qin, Y., Yang, F., et al. (2017). SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief. Bioinform.

Herzel, L., and Neugebauer, K.M. (2015). Quantification of co-transcriptional splicing from RNA-Seq data. Methods *85*, 36–43.

Jin, C., and Felsenfeld, G. (2007). Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes Dev.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: A fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet.

Ma, J., Saghatelian, A., and Shokhirev, M.N. (2018). The influence of transcript assembly on the proteogenomics discovery of microproteins. PLoS One *13*.

Olexiouk, V., Van Criekinge, W., and Menschaert, G. (2018). An update on sORFs.org: A repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. *46*, D497–D502.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. *11*, 1650–1667.

Prabakaran, S., Hemberg, M., Chauhan, R., Winter, D., Tweedie-Cullen, R.Y., Dittrich, C., Hong, E., Gunawardena, J., Steen, H., Kreiman, G., et al. (2014). Quantitative profiling of peptides from RNAs classified as noncoding. Nat. Commun. *5*.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics.

Saeidipour, B., and Bakhshi, S. (2013). The relationship between organizational culture and knowledge management,& their simultaneous effects on customer relation management. Adv. Environ. Biol. *7*, 2803–2809.

Stenson, P.D., Mort, M., Ball, E. V., Howells, K., Phillips, A.D., Cooper, D.N., and Thomas, N.S.T. (2009). The human gene mutation database: 2008 update. Genome Med.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res.

UCSC liftOver LiftOver.

Vanderperre, B., Lucier, J.F., and Roucou, X. (2012). HAltORF: Adatabase of predicted out-of-frame alternative open reading frames in human. Database.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Sander, C., Stuart, J.M., Chang, K., Creighton, C.J., et al. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res.