Investigation of transcription and translation from sORFs and altORFs in naive B and T cells in *M. musculus*

BS-MS Thesis

Chaitanya Erady

20131080

Under the guidance of

Dr. Sudhakaran Prabakaran

Department of Genetics,

University of Cambridge

Certificate

This is to certify that this dissertation entitled "*Investigation of transcription and translation from sORFs and altORFs in naive B and T cells in M. musculus"* towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Ms. Chaitanya Erady at IISER-Pune under the supervision of Dr. Sudhakaran Prabakaran, Department of Genetics, University of Cambridge, during the academic year 2017-2018

Student:
Chaitanya Erady

Supervisor:
Dr. Sudhakaran Prabakaran

Declaration

I hereby declare that the matter embodied in the report entitled entitled "*Investigation of transcription and translation from sORFs and altORFs in naive B and T cells in M. musculus"* are the results of the work carried out by me at the Department of Biology, IISER-Pune, under the supervision of Dr. Sudhakaran Prabakaran and the same has not been submitted elsewhere for any other degree.

Student:
Chaitanya Erady

Supervisor:
Dr. Sudhakaran Prabakaran

Abstract

Identification of transcription and translation from noncoding regions augmented the complexity of the genome, a problem further compounded by novel open reading frames (ORFs) present within noncoding as well as genic regions with as of yet unclear functions. In this study, we investigated two novel ORFs: short ORFs (sORFs) and alternative ORFs (altORFs), within mouse naive B and T cells. We established evidence of transcription for 2721 sORFs and 4251 altORFs and found 3604 sORFs and 2104 altORFs to be translated. We also identified 289 sORFs and 980 altORFs as differentially expressed (DE) between B and T cells. Furthermore, PCA analysis indicated that transcript expression levels of these novel ORFs are significant and sufficient to distinguish between the two cell types. Additionally, differential methylation (DM) analysis of these differentially expressed novel ORFs and protein-coding transcripts allowed us to identify 117, 139 and 1398 DMRs upstream, downstream and within the body of DE sORFs, 199, 257 and 28497 near DE altORFs and 1712,1679 and 24910 near protein-coding transcripts. Moreover, 46 sORFs containing DMRs were identified in the upstream and downstream regions of protein-coding transcripts indicating that expression of DE protein-coding transcripts might be affected by sORFs. Also, we found no evidence of LINE/SINE repeat elements regulating expression of DE sORFs. Here, we present a framework for a systematic investigation of transcription and translation from novel ORFs that could be utilised to ascertain their functions or identify potential diseases variants present within them.

Contents

List of Figures

List of Tables

Acknowledgement

I am extremely grateful to Dr. Sudhakaran Prabakaran for his guidance throughout this project. I would also like to thank Dr. Ruchi Chauhan for carrying out the extraction of mouse B and T cells from mouse spleens followed by FACS sorting and for doing the proteomic analysis. Additionally, I am grateful to Dr. Marco Chiappello for helping us with the proteomic analysis, Mr. Adam Andreani for the initial analysis of this study and Dr. Matt Wayland for helping with the set-up of cloud analysis pipelines. I would like to thank Prof. Anne Ferguson-Smith's lab for providing mouse spleen samples, Dr. Cristina Pina in whose lab B and T cell sorting was done and Prof. Kathryn Lilley in whose lab proteomics was done.

I am also grateful to the members of Prabakaran lab, especially Felix Jackson, David Chong, Jean Nel and Narendra Meena for the many productive discussions we had that proved very helpful in completing this project. Finally, I would like to thank Dr. Girish Ratnaparkhi for hosting me in his lab at IISER-Pune for the duration of this project.

Introduction

The information within genes to be translated into a protein is present in one of the six possible reading frames: +1, +2, +3, -1, -2 or -3, where + and – designate opposite chromosome strands. One of these six reading frames called an open reading frame (ORF) or canonical ORF, contains a continuous stretch of codons bound by start and stop sites that get transcribed and translated into a protein. The current annotated ORFs are the most well studied genomic components, but they comprise only about 2-3% of the total genome. Until recently, the remaining 98% of the genome was dubbed as 'junk DNA' because they presented no evidence of apparent protein-coding or biochemical functions (St. Laurent, et al., 2014).

Primarily, attribution of functions to genomic regions was done by modifying the genotype of an organism and studying the resultant changes in the phenotype. This method failed to ascribe functions to noncoding regions as the forward genetic screens used were focused on protein-coding regions. Moreover, modifications introduced in noncoding regions, which were at best considered regulatory elements, conferred phenotypic changes that were too small to be detected, a problem further intensified by the redundancy in noncoding sequences (Kapranov and St. Laurent, 2012). An unbiased study to identify RNAs transcribed in the human genome using tiling arrays with randomly selected RNA fragments and various other sequencing methods initiated by ENCODE, provided the first evidence of pervasive transcription of the human genome (Birney et al., 2007). These transcripts which were stable enough to be detected but whose functions could not be ascertained were annotated as 'dark matter' RNAs.

Further research in this area led to the identification and classification of several such dark matter transcripts including but not limited to long noncoding RNAs (lncRNAs), microRNAs (miRNAs) and small nucleolar RNAs (snoRNA). Now, databases dedicated to cataloguing noncoding RNAs, for example, NONCODE database, exist for several organisms (Zhao et al., 2015). Interest in these regions burgeoned with the finding that ~90% of genome-wide association study (GWAS) hits or disease variants are present within noncoding regions (St. Laurent, et al.,

2014). Moreover, the involvement of noncoding transcripts in regulating diseases like cancer, neurodegenerative disorders and autoimmune diseases garnered further interest (Hrdlickova et al., 2014). For example, multiple lncRNAs have been implicated for their role in cancer wherein their upregulation or downregulation can increase or slow down cancer development by interfering with cellular processes like DNA repair (Hrdlickova et al., 2014). Thus, the focus is shifting from protein-coding genes to include noncoding elements in the hope of identifying crucial disease variants or single nucleotide polymorphisms (SNPs) that can be targeted for disease diagnosis and developing therapeutic strategies (St. Laurent, et al., 2014).

Now, there is also a growing understanding that these noncoding regions do harbour the potential to code for proteins and protein-like products (Prabakaran et al., 2014). In a previous study of mouse neurons at our lab, translations from several of these noncoding regions including 5'UTRs, 3'UTRs, introns, pseudogenes etc. were identified (Prabakaran et al., 2014). Thus, for regions which were labelled as 'junk DNA' or noncoding regions, evidence for transcription, translation and their significance in terms of hosting several disease-associated variants highlights the importance of studying these regions. Moreover, this also calls for a change in or revision of current genome annotations.

Additionally, transcription and translation are not limited to the information encoded by one ORF of a protein-coding gene as previously believed. Using techniques like ribosome footprint profiling, proteogenomic analysis and computational studies, several novel ORFs within these protein-coding genes as well as within noncoding regions have been identified in humans and other organisms. Furthermore, proteins or protein-like products from these novel ORFs seem to contribute to proteomic diversity (Raj et al., 2016; Gong et al., 2014). For example, in lower organisms like bacteria and viruses with smaller genomes, coding capacity is optimised by utilising leaky ribosome scanning to initiate translation from multiple AUGs allowing for translation from several novel ORFs (Gong et al., 2014). Therefore, the idea that one CDS (coding sequence) codes for only one protein no longer holds true (Mouilleron, Delcourt and Roucou, 2016). Although novel ORFs have been identified, our knowledge on the subject is still limited and so in this project we focus on the study of two novel ORFs: sORFs and altORFs, using proteogenomic analysis.

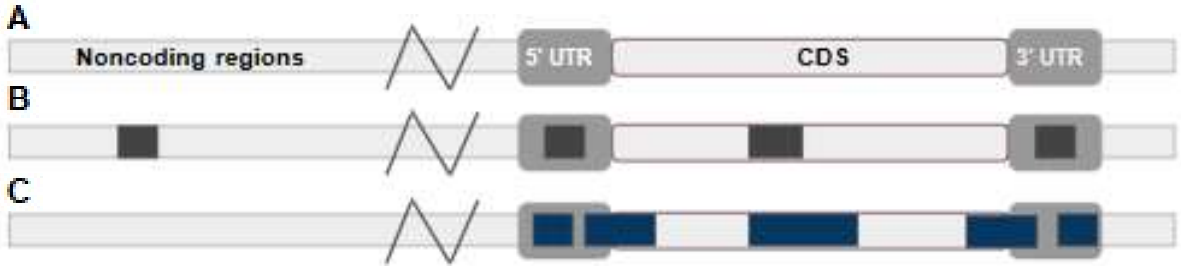


Fig. 1: Genomic regions containing novel ORFs (A) Reference for figures in (B) and (C) depicting different genomic regions including the coding sequence of a gene (CDS), 5' and 3' UTRs as well as a kink to denote the location of noncoding genes far and near from the gene. (B) sORFs (black rectangles) identified within 5' and 3' UTRs, CDS of a gene and noncoding regions. (C) altORFs (blue rectangles) are present in the 5' and 3' UTRs, in CDS of a gene or are found overlapping the CDS and the 5' or 3' UTRs.

sORFs and canonical ORFs both have start and stop codons enclosing nucleotide sequences with the potential to be translated (Hellens et al., 2016). The difference is in their lengths wherein sORFs are short with a length of < 100 aa or 300 nucleotides (Hellens et al., 2016; Basrai, M.A. et al., 1997). Furthermore, for sORFs, the start codon is not necessarily an AUG and about nine other alternative start codons have been identified (Olexiouk V, et al., 2016). The location of sORFs within the genome can be within the 5' or 3' untranslated regions (UTRs) of a gene, within the CDS of a gene albeit in an alternative frame relative to the canonical ORF or within noncoding regions (Hellens et al., 2016) (Fig. 1B).

Although significant numbers of sORFs are present within the genome, they remained undetected because of their small size or expression levels (Basrai, M.A. et al., 1997). Even though computational predictions to identify ORFs within a genome exist, they employ thresholds of 100 aa as the minimum length resulting in the inability of such tools to detect sORFs (Kastenmayer, 2006; Hanada et al., 2010). Creation of programs explicitly designed to identify sORFs like sORFfinder (Hanada et al., 2010), as well as identification of their translated products using ribosome profiling has aided in the discovery of sORFs (Olexiouk V, et al., 2016). Involvement of sORFs and their short peptides (sPEPs) in key cellular processes like chromosome segregation, genome stability, transport etc. in *S. cerevisiae*, regulation

of growth and development in plants and their ability to activate transcription factors to regulate Drosophila development has raised interest in sORFs (Kastenmayer, 2006; Kondo, T. et al., 2010; Hanada, K. et al., 2013; Hellens et al., 2016).

Another novel ORF of interest is alternative open reading frame (altORF). They, like canonical open reading frames, contain codons demarcated from the remaining genomic sequence by start and stop codons but the start codons used are different and therefore altORFs code for an alternative protein (Vanderperre, Lucier and Roucou, 2012). Another difference between altORFs and canonical ORFs is in their length wherein altORFs have a median length of 57 aa in humans, but the median length for proteins from canonical ORFs is 344 aa (Vanderperre et al., 2013). Furthermore, altORFs can be present entirely within the CDS but in an alternative reading frame relative to the canonical ORF, within the 3' or 5' UTRs or overlapping the UTRs and the CDS (Vanderperre et al., 2013) (Fig. 1C). In organisms with small genomes like viruses, altORFs are common and are employed as an alternative protein generating mechanism to increase the diversity of the virus proteome (Vanderperre, Lucier and Roucou, 2012). Studies on altORFs have additionally revealed that they are significantly conserved and therefore possibly play an important role in an organism (Vanderperre, Lucier and Roucou, 2012). There is also evidence for translated products of altORFs cooperating with proteins encoded by canonical ORFs to regulate the latter's function (Samandi et al., 2017). Moreover, altORF proteins have also been identified as potential biomarkers and therapeutic targets for diseases like cancer (Vanderperre, Lucier and Roucou, 2012). Finally, an estimated average of 3.88 altORFs is present for one mRNA (Vanderperre et al., 2013). Thus, altORFs are important regions of the genome that require further investigation.

Although sORFs and altORFs have been identified and research work to discover their functions is being carried out, we focus on investigating their presence and possible functions within mouse naive B and T cells using proteogenomic analysis. Proteogenomic analysis involves discerning the identity of a protein by mapping MS/MS spectra of the isolated proteome to a custom reference protein database, generated using information from transcripts identified for that particular organism (Nesvizhskii, 2014). This approach is much better than the current proteomic analysis, which relies on mapping peptide spectra to known protein databases like

UniPort, as it allows for the identification of sample-specific proteins or protein-like products. Furthermore, proteogenomic analysis is different from ribo-seq technique in that the latter uses information of ribosomes attached to RNAs to identify actively translated RNAs (Ingolia, 2016).

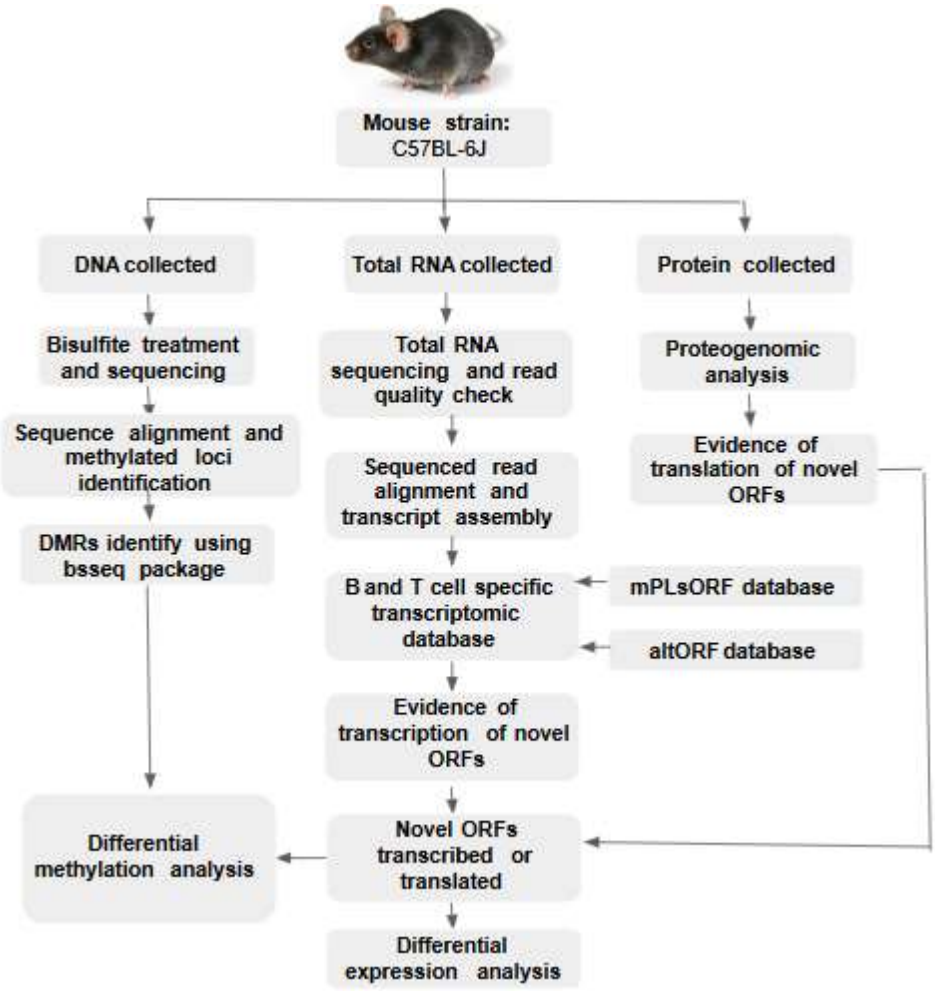


Fig. 2: Workflow depicting major steps involved in this project. (mPLsORF database stands for mouse Prabakaran Lab sORF database).

The objective of this project is to evaluate the presence of transcription and translation of novel ORFs like sORFs and altORFs in mouse resting B and naive T cells and investigate whether they are regulated by differential methylation. This work is the first such systematic analysis of sORFs and altORFs in mouse B and T cells. Even though our interest is primarily in the exploration of noncoding regions, expression of known transcripts and proteins have been evaluated for comparison.

Throughout this study, our definition of noncoding regions include lncRNAs, introns, intergenic regions, pseudogenes, UTR's as well as alternative frames in exonic regions relative to canonical ORFs (i.e. +2, +3 frame if canonical ORF is assumed as +1 frame).

In this study, we created our own sORF and altORF databases with information curated from online databases. Additionally, we isolated transcripts and proteins from our mouse samples and used computational tools and proteogenomic analysis to evaluate evidence for transcription and translation of sORFs and altORFs. Next, we asked the question whether sORF and altORF transcript expression levels are differentially expressed between B and T cells. If they are differentially expressed, it could hint towards the fact that these novel ORFs are involved in cell-specific functions. Moreover, we performed a PCA analysis to evaluate if the transcript expression levels of these novel ORFs is sufficient to distinguish between the two cell types in question thereby verifying whether expression levels from noncoding regions are significant enough to differentiate between cell types. Finally, we identify differentially methylated regions (DMRs) in B and T cells and evaluate whether sORFs and altORFs are regulated by methylation. Similarly, we evaluate whether differentially expressed sORFs and altORFs are regulated by repeat elements like LINEs and SINEs and whether protein-coding genes are regulated by sORFs. The focus is not on what type of methylation and therefore the exact nature of regulation but rather we try to identify the possibility of such a regulation taking place.

This work (Fig. 2) highlights the emerging importance of noncoding regions as well as novel ORFs in the genome. Initially considered to have no function, we also show from their expression levels as well as differential methylation studies that they could be important components of the cell. This project lays a foundation for several other works performed in our lab including predicting sORF structures, identifying cancer mutations that map to sORFs and evaluating sORFs to identify potential pathogenic variants (mutations that increase predisposition to a disease) thereby aiding in the development of diagnostic markers for diseases.

Materials and Methods

1. Sample information

Cells extracted from the spleen of 3 Male and 3 Female C57BL-6J mice were FACS sorted to isolate resting B and naive CD4+ T cells. Total RNA was extracted from each of the 12 samples (3 B-male, 3 B-female, 3 T-male and 3 T-female). Proteins were collected from each of the 12 samples, but those from the same sub-group (B-male, B-female, T-male or T-female) were pooled together to gather sufficient protein for further analysis. DNA was extracted from resting B and naive CD4+ T cells isolated from 2 Male and 2 Female C57BL-6J mice. This work was done in Dr. Fergusson-Smith's lab at the University of Cambridge. (GEO accession: GSE94671; Ferguson-Smith et al., 2017).

2. Data analysis

We utilised a cloud-based platform Cancer Genomics Cloud (www.cancergenomicscloud.org) for data analysis. We used the CGC platform for read quality checks, read alignment, SAM to BAM file conversion and coordinate to nucleotide sequence conversion using FastQC, HISAT2, Picard SortSam and Bedtools GetFasta respectively which are tools already available in CGC. For transcript assembly, we uploaded and used StringTie in CGC. For differential expression analysis, Ballgown a Bioconductor package was used. Further analysis involving processing data and creating plots were done using a combination of R and UNIX shell bash commands.

3. Creation of transcriptomic database and identifying differential expressed transcripts

Library preparation of total RNA collected from the 12 mouse samples was performed using Illumina stranded total RNA library prep kit. Paired-end sequencing of reads was performed using the Illumina HiSeq 2500 platform (GEO accession: GSE94671; Ferguson-Smith et al., 2017). This work was done in Dr. Ferguson's lab. Quality of sequenced reads was determined using FastQC which was run with

default settings. Primary assembly sequence and comprehensive gene annotation files for C57BL-6J, release version M12, was used as the reference genome for our analysis (GENCODE, 2016).

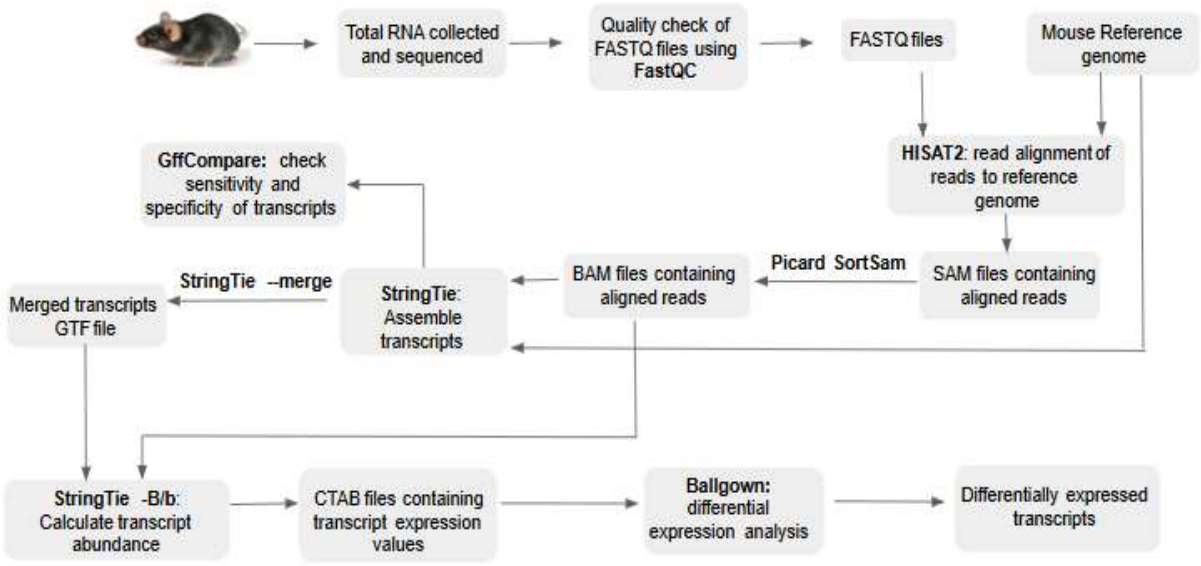


Fig. 3: Workflow describing the various tools and steps involved starting from sample collection to identifying differentially expressed transcripts. FASTQ, SAM, BAM, GTF and CTAB are different file formats utilised in our analysis.

Using HISAT2-build, exon and splice-site coordinates were extracted from the reference annotation and the genome index file thus created was used to align paired-end sequenced reads to the reference genome using HISAT2 (Kim D, et al., 2015). HISAT2 was run with default settings and the –dta option to ensure that stranded information is retained after alignment (Kim D, et al., 2015) (Fig. 3). SAM files that were generated were converted to BAM files using Picard SortSam that further sorts aligned reads based on their genomic coordinates (Broad Institute, Picard SortSam).

Aligned reads in the BAM files and the reference genome were used to assemble sample-specific transcripts using StringTie run with default settings and the –fr option which assumes that reads were generated from a stranded library (Pertea et al., 2016) (Fig. 3). GFFcompare (Pertea et al., 2016) was run on the 12 output GTF files to assess the sensitivity and specificity of the assembled transcripts (Fig. 3). As

recommended by the HISAT-StringTie-Ballgown pipeline, StringTie –merge was run on all the 12 sample GTF files and the reference genome to generate a final merged transcript file containing a list of non-redundant transcripts (Pertea et al., 2016). Although StringTie can assemble novel transcripts, the –merge step can cause a loss of some of the identified novel transcripts. This step is still recommended to ensure that a transcript identified as novel is not an incompletely assembled transcript by verifying for its presence across all samples.

This merged transcript GTF file, our transcriptomic database, along with the 12 BAM files containing aligned reads were used for a second StringTie run with parameters '-Be' to calculate transcript FPKM values for each sample (Pertea et al., 2016).The 12 CTAB output files were used by Ballgown, an R/Bioconductor package, to carry out differential expression analysis (Frazee AC et al., 2017). Ballgown's 'stattest' function performs a $\log_2$ transformation on the library-normalised FPKM values, fits the data to standard linear models and calculates p and q values for the transcripts (Frazee AC, 2017). Transcripts with q values (number of significant results that are false positives) < 0.01 were called differentially expressed. The sequencing analysis pipeline was set up in CGC by Dr. Matt Wayland, and Ballgown analysis was done by Mr. David Chong. My task involved verifying read assembly by HISAT2 and performing transcript assembly using StringTie.

4. Creation of B and T cell-specific transcriptomic databases

We used the StringTie –merge results to create our cell-specific transcriptomic databases. If a transcript's gene is identified from the reference genome, the corresponding gene id, transcript id and gene name are generated by StingTie otherwise a random gene and transcript id with the prefix 'MSTRG' is listed out for each transcript (Pertea et al., 2016). We define novel transcripts as those without a corresponding gene name and annotated/known transcripts as those which have been assigned a reference gene by StringTie. Furthermore, we merged the information in the 12 CTAB files containing transcript FPKM values with the merged transcript file to create the final transcriptomic database.

This transcriptomic database was then filtered to remove duplicates based on chromosome number, transcript start and end coordinates and strand information.

Additionally, the database was filtered to remove transcripts with '0' FPKM values for all the 12 samples. The remaining transcripts were categorised into four sub-groups: B-male, B-female, T-male or T-female, based on whether at least one out of three samples corresponding to a sub-group had a non-zero FPKM value. Finally, the transcripts were categorised into B or T cell based on whether the transcript was present in at least one of the two sub-groups corresponding to a particular cell-type. All the transcripts that were present in B cells were extracted to create a B cell-specific transcriptomic database, and similarly, all transcripts present in T cells were used to create T cell-specific transcriptomic database.

The transcript coordinates in B and T cell-specific transcriptomic databases were used to extract the corresponding nucleotide sequence from the reference genome using Bedtools Getfasta available in CGC (Quinlan and Hall, 2010). Bedtools Getfasta was run with default settings and with the name parameter ="True", which ensures that the name column of the input BED file is used as the header for the output FASTA file (Quinlan and Hall, 2010). The output FASTA files generated for B and T cells are our B and T cell-specific nucleotide databases. Transcripts in our B and T cell-specific nucleotide databases were analysed computationally in 6 frames to determine all possible translated products and create a custom reference protein database. The resulting peptide sequences were catalogued as B and T cell-specific proteogenomic databases.

5. Creation of sORF database

Mouse Prabakaran Lab sORF database (mPLsORF) was created using information curated from two sources: sORFs.org and SmProt (Olexiouk V, et al., 2016; Hao Y, et al., 2017). sORFs.org is a repository that contains a list of sORFs which have been computationally predicted and experimentally verified using ribosome profiling (Olexiouk V, et al., 2016). We exported mouse sORFs from sORFs.org with default filters except for FLOSS classification which was set to 'GOOD' and 'EXTREME'. SmProt contains a list of experimentally validated small peptides identified in several species including mouse (Hao Y, et al., 2017). We extracted mouse sORFs from SmProt with filter parameters set to 'ALL'. The downloaded information from SmProt did not provide chromosome information for sORFs. A macros code was, therefore,

run on the SmProt website to specifically extract this chromosome information which was available on the webpage but not in the downloaded TXT file. Furthermore, each sORF obtained from SmProt has a designated id with the format 'SPROMUSXXXX', but two sORFs had a wrong id beginning with 'PROMUS' which was manually corrected. Also, two sORF entries were not from mouse but from human and rat and were subsequently removed from further analysis.

Both databases had several duplicate entries which were removed by filtering them based on chromosome location and amino acid sequence. Finally, we assigned a unique sORF id with the format 'mPLsORFXXXXXXXXXX', where X denotes a number, to each sORF entry and created our sORF database with the following columns: "Organism_name", "Source_database", "Chromosome_number", "Start_coordinate", "End_coordinate", "Strand", "Amino_acid_sequence". There are still a few sORFs in our database with the same chromosome coordinates, but these duplicates were not removed because their corresponding amino acid sequences were different. A length distribution for sORFs in mPLsORF database, as well as a pie chart depicting the proportion of different sORF annotations, was also created. Using the sORF information in our mPLsORF database, we constructed FASTA headers which combined with their respective amino acid sequences generated our sORF fasta database.

6. Creation of altORF database

Information for altORFs identified in mouse was downloaded from Roucou's lab, Université de Sherbrooke, Canada (Vanderperre et al., 2013). Of the altORFs identified, a few had multiple chromosome numbers assigned to it. These were removed from our dataset to generate a file compatible with tools used for downstream analysis. There are a few altORFs in our database with the same chromosome coordinates which were retained owing to a difference in their amino acid sequences. Additionally, a plot displaying length distribution of altORFs in our database was also generated. As done for sORFs, information available in our altORF database was used to create FASTA headers and the corresponding amino acid sequence was used as the FASTA sequence to generate our altORF fasta database.

7. Identifying evidence for transcription of novel ORFs in B and T cells:

GTF files for sORFs, altORFs and B and T cell transcripts were created and sorted first according to their chromosome number and then according to their start coordinates in ascending order. Bedtools intersect was used to identify coordinate overlap between sORFs or altORFs and B or T cell transcripts. The following parameters were used for this run: parameter -'f' = 0.99, signifying only sORFs/altORFs that overlap 99% of the transcript coordinates will be called; parameter -'wo' to generate information on the sORF or altORF, the transcript it matches to and the total number of nucleotide overlap between the two; parameter –'s' to only map sORFs to transcripts if they are from the same strand (Quinlan and Hall, 2010). altORFs were run without the –'s' parameter because the altORF database contains no strand information. Thus, there is a possibility that more altORFs with evidence of transcription are being identified than if the -s parameter was used. altORF and sORF ids were extracted from the output TXT files generated by bedtools getfasta and filtered to create a unique list of sORFs or altORFs with evidence of transcription.

8. Differential Expression Analysis:

Differential expression analysis was performed using Ballgown. The identified differentially expressed (DE) transcripts, called if the transcript q value was less than 0.01, were categorised into four: protein-coding transcripts identified using transcript annotation information from Ensembl BioMart (Ensembl Biomart, Ensembl genes 91), sORF transcripts identified as DE transcripts that mapped to sORFs, altORF transcripts which are DE transcripts that mapped to altORFs and finally transcripts that belonged to none of these three categories. DE sORFs and altORFs were categorised into B-male, B-female, T-male, T-female, B cells or T cells depending on their FPKM values.

We performed PCA to determine if sORF/altORF transcript expression levels could distinguish between cell types and different genders of the same cell type. For this, we used FPKM values of transcripts that map to sORFs and altORFs for each of the 12 samples, $\log_2$ normalised the data using a pseudo-count of 1 ($\log_2$(number+1)) and centred the data around its median by subtracting the median from each data

point. The final normalised dataset was then run using the R function prcomp (R core team (2015), version 3.2.3, function prcomp) with scale and centre parameters set to “FALSE”. PCA plots for B VS T cells and Male VS Female cells were plotted along two principal component axes: PC1 and PC2. Similarly, PCA plots were generated for known transcripts.

9. Proteogenomic Analysis: Identifying evidence for translation of novel ORFs

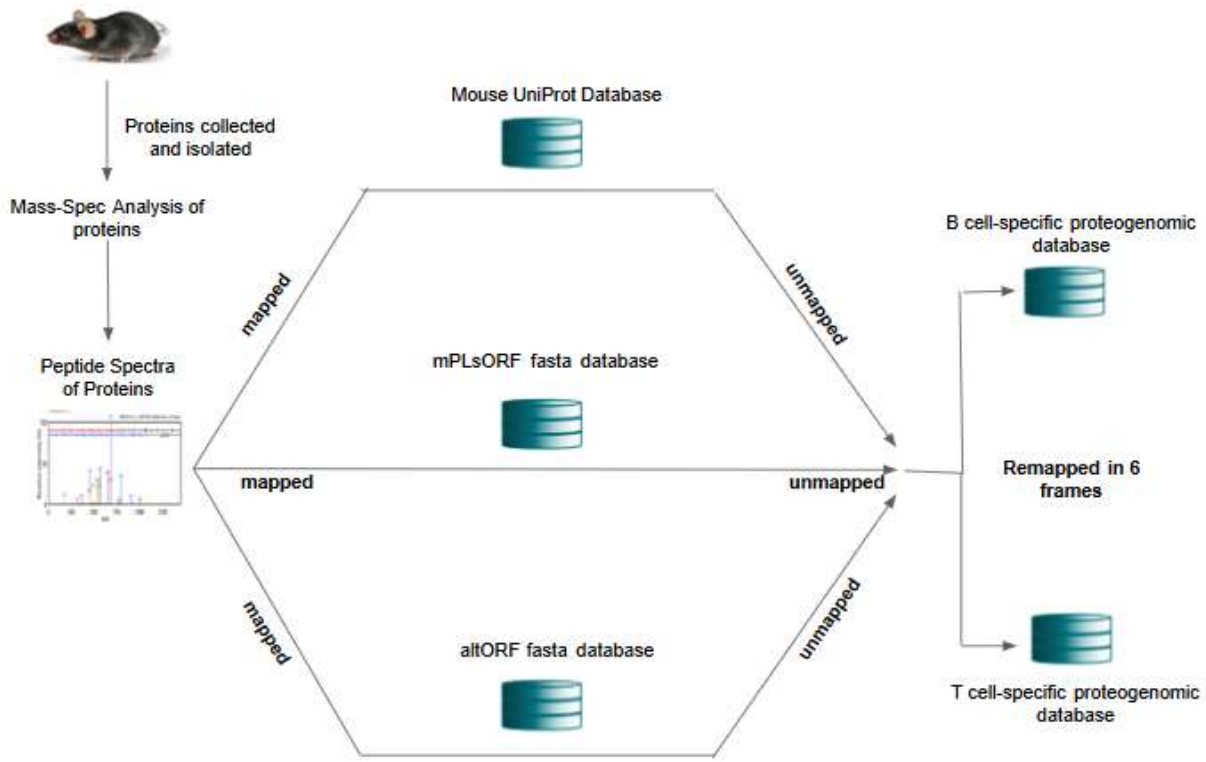


Fig. 4: Proteogenomic workflow to identify novel proteins in our sample. Mass-spec analysis generated peptide spectra of proteins identified in our sample which is then mapped to one of the three databases shown in blue. Proteins identified by mapping to these three databases and that were retained after applying a 1% FDR cutoff were labelled as known proteins, sORF proteins or altORF proteins. Remaining peptide spectra which did not map to any of the three databases were searched against B and T cell-specific proteogenomic database to identify novel proteins in our sample.

Proteins obtained from the four sub-groups (B-male, B-female, T-male or T-female) were run on a gel, protein bands were cut out following which proteins were extracted. Mass-spectrometry analysis of the isolated proteins generated peptide spectra for proteins identified in our sample. This experimental work was carried out by Ruchi, a visiting scientist in Prabakaran lab at the University of Cambridge. The peptide spectra were then mapped to proteins in UniProt database (UniProt, 2017), sORF fasta database, as well as the altORF fasta database and only those within top 1% FDR cutoff, were retained (Fig. 4). This workflow allowed us to identify known proteins, sORF protein-like products and altORF proteins present in our sample.

The output file with details about sORF, altORF or known proteins identified in our sample along with their protein abundance and FDR values were filtered to remove 'cRAP' which are proteins that are either contaminants or were introduced accidentally. Next, only those proteins with 'Medium/High' FDR values were retained. Finally, entries with no abundance values for all the four sub-groups were removed. After filtering, sORFs, altORFs and known proteins with evidence of translation in our mouse samples were generated. The unmapped spectra form each of these analyses (peptide spectra mapping to UniProt, sORF database and altORF database) was searched against the B cell and T cell proteogenomic database (Fig. 4). We also compared sORF and altORF ids identified with evidence of transcription or translation to determine the intersection which denotes sORFs or altORFs with evidence of transcription and translation in mouse B and T cells.

10. Differential methylation analysis

DNA isolated from our samples were pooled together before library preparation and subjected to oxidation and bisulfite treatment. After library preparation, whole genome oxidative bisulfite sequencing was performed using Illumina HiSeq 2500 platform (GEO accession: GSE94674, Ferguson-Smith et al., 2017). The sequenced reads are then aligned using Bismark which can output a BEDGRAPH file with information on the methylated loci or CpGs. (Krueger F., 2011). This work was done in Dr. Ferguson's lab.

To identify DMRs between B and T cells we used the bsseq package (Hansen KD, 2012). The BEDGRAPH file containing methylated loci is first converted to a BS-

object, a file format used by bsseq functions, using an R script (Hansen KD, 2012). An additional smoothening step is applied to this data using the function BSmooth which calculates methylation estimates for a genomic region utilising a minimum of 200 smoothening windows wherein each window contains at least 20 methylation loci (Hansen KD, 2012). After smoothening, we select for CpGs with a minimum coverage of two in at least three out of four samples corresponding to a cell type. Here, the coverage is defined as the sum of methylated and unmethylated reads corresponding to a CpG locus. The function BSmooth.tstat with parameter 'estimate.var' set to "group2" or B cell uses t-statistics to compare CpGs between the two groups (B cell and Tcell). The resulting BSseqTstat object is fed as an input to the function dmrFinder and using an alpha value of 0.05, differentially methylated regions (DMRs) between B cell and T cell were determined. This output was further filtered to retain only those DMRs containing at least 3 CpGs spaced within 300bp of each other (Hansen KD, 2012).

To evaluate whether differential methylation regulates the expression of protein-coding transcripts, sORFs and altORFs, DMRs identified between B and T cells were mapped to upstream, downstream regions (3000 bp window) and the body (within start and stop coordinates of transcripts) of DE protein-coding transcripts, DE sORFs and DE altORFs using bedtools intersect. To explore the possibility of regulation of sORFs and altORFs by repeat elements like LINE or SINE, we mapped mouse LINE/SINE (downloaded from repeatmasker.org for mm10), to upstream and downstream regions (3000 bp window) of DE sORFs, altORFs and protein-coding transcripts and then mapped DMRs to any LINE/SINE identified. Furthermore, to evaluate if sORFs regulate expression of protein-coding transcripts, we mapped 6248 sORFs to upstream and downstream regions of protein-coding transcripts and investigated if DMRs are present within any identified sORFs. The sample collection, bisulfite sequencing, Bismark alignment were done in Dr. Fergusson-Smith's lab at the University of Cambridge. Mapping of LINE/SINE to sORF/altORF and mapping of sORF to protein-coding transcripts was performed by Mr. Narendra Meena. My task involved identifying DMRs between B-male and T-male, creating files with a list of protein-coding transcripts, sORFs and altORFs and identifying DMRs near protein-coding transcripts DE sORFs and DE altORFs.

Results and Discussion

1. Creation of cell-type specific transcriptomic database and identifying differential expressed transcripts

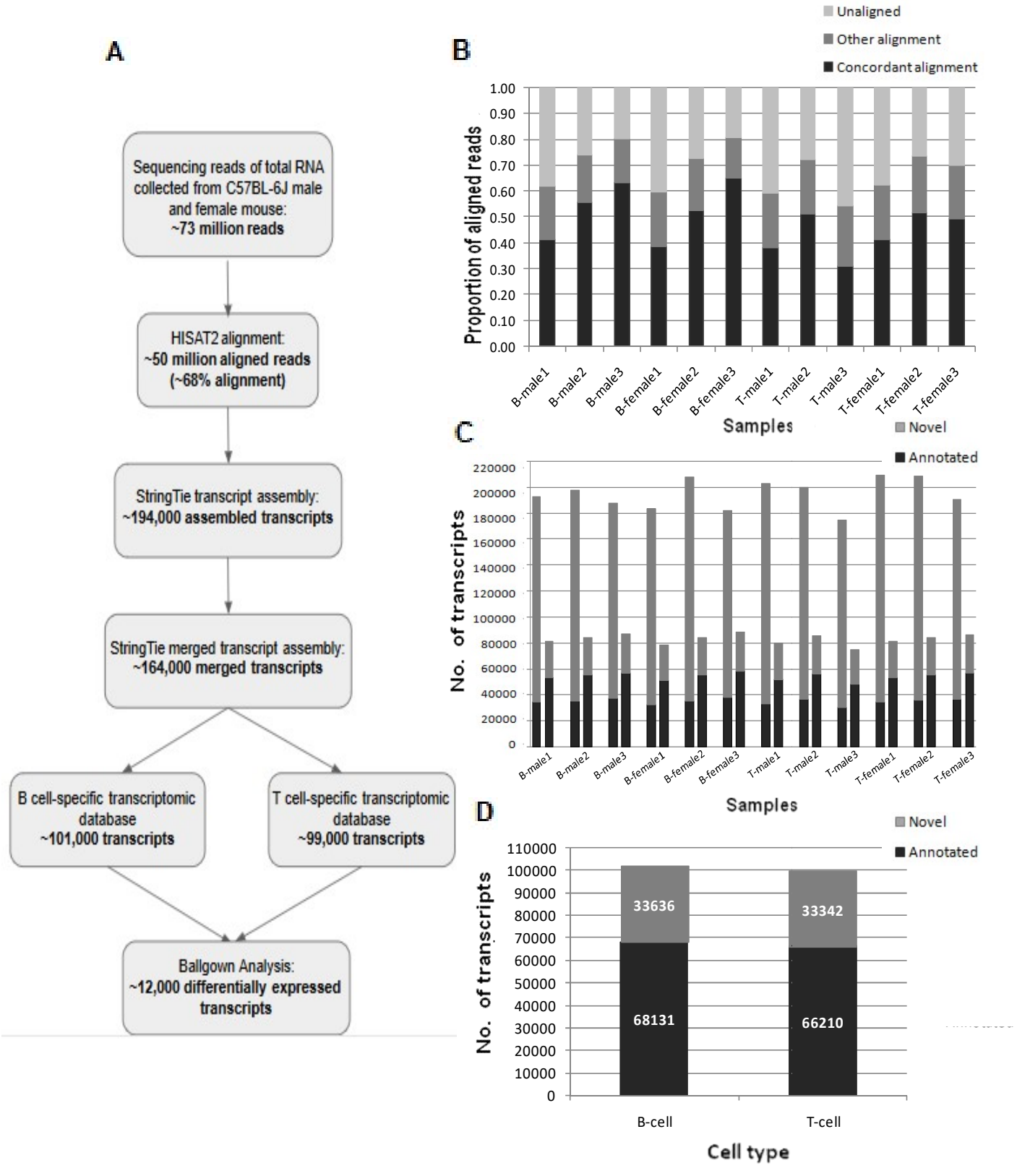


Fig. 5: Identifying transcripts in our sample. **(A)** The workflow from sequencing reads to identification of differentially expressed transcripts along with the number of reads or transcripts identified at each stage. **(B)** Proportion of different types of aligned reads (y-axis) for each sample (x-axis). Concordant alignment (black) refers to reads that aligned to the reference genome with a specified orientation (forward-reverse) and within a specific distance with respect to each other. Other alignment (dark grey) contains reads that mapped concordantly >1 times, discordant aligments as well as single read of a mate pair that aligned atleast 1 time. Unaligned (light grey) refers to reads that aligned to the genome 0 times. **(C)** Number of transcripts (y-axis) in each sample (x-axis) identified after two StringTie runs. Transcripts identified after first StringTie run are shown first amongst the two bars corresponding to one sample. Transcripts identified after StringTie merge is denoted by the second and smaller bar for each sample. Transcripts were colourcoded dark grey if they are novel and black if they are annotated. **(D)** Number of transcripts (y-axis) in each cell –specific transcriptomic database (x-axis) Transcripts were colourcoded dark grey if they are novel and black if they are annotated

No. of transcripts after StringTie merge	No. of transcripts after removing unlocalised contigs	No. of transcripts without FPKM = '0' for all 12 samples	No. of transcripts after duplicate removal	No. of transcripts in cell-specific transcriptomic database
164491	164274	111417	109441	B cell: 101767
				T cell: 99552

Table 1: No. of transcripts identified at different filtering stages of B and T cell-specific transcriptomic database creation. Unlocalised contigs refers to transcripts assembled from reads which map to regions whose chromosome number is known but the exact order and orientation of these regions within the chromosome is unknown.

DE transcripts after Ballgown analysis	DE transcripts after removing unlocalised contigs	DE transcripts after removing duplicates
12138	12079	12009

Table 2: No. of DE transcripts identified after different stages of filtering.

Our first goal is to systematically investigate evidence for transcription and translation of novel ORFs from mouse B and T cells. This mandates that we isolate and identify transcripts and proteins from the same mouse samples thereby also requiring that we create our own transcriptomic database for B and T cells. Furthermore, since our interest lies in investigating transcripts from noncoding regions, we collected total RNA from our samples, i.e. without a poly-A selection as some noncoding transcripts do not have a poly(A) tail.(Zhang et al., 2014). Our samples, therefore, contain rRNAs which we did not deplete post-sequencing as our initial analysis showed evidence of sORFs within rRNAs (data not shown). There are several programs and pipelines available to assemble transcripts from sequenced reads, but we chose the HISAT-StringTie-Ballgown pipeline as it requires less memory and is therefore much faster than prevalent tools utilised for the same analysis (Pertea, M. 2016).

Read alignment: ~73 million sequenced reads of total RNA from our samples were aligned to the reference genome using HISAT2 resulting in ~50 million aligned reads (Fig. 5A). The alignment rate for our samples ranges from ~50-80% averaging at ~68% (Fig. 5B).

Transcript Assembly: An average of ~194000 transcripts was assembled for each of our samples by StringTie. Merging the transcripts across the 12 samples resulted in a total of ~164,000 transcripts which after filtering amassed to ~109,000 transcripts (Fig. 5A, Table 1). Comparing the number of transcripts before and after running StringTie –merge as seen in Fig. 5C, there is a loss of a significant amount of transcripts, mostly novel whereas an increase in the number of annotated transcripts is observed across all samples.

Creating cell-type specific transcriptomic database: Based on the transcript FPKM values ~101,000 transcripts in B and ~99,000 transcripts in T cell-specific transcriptomic database were identified (Fig. 5A). The number of annotated and novel transcripts is similar between B and T cells, but this is not representative of the number of transcripts common to both these cells (Fig. 5D). Of the 109441 transcripts, 91878 transcripts are common to B and T, 9889 are unique to B and 7674 are unique to T cells.

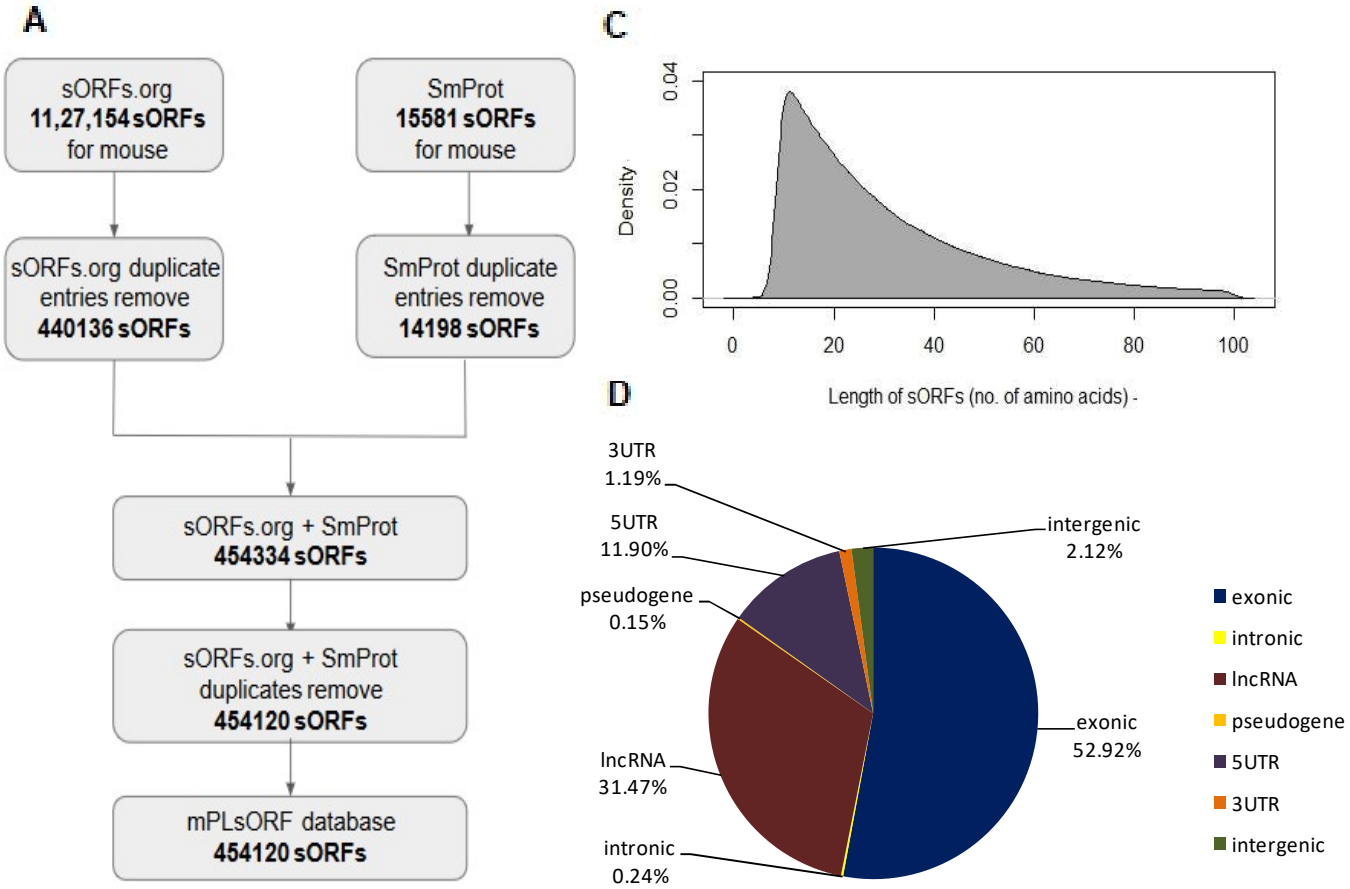
Identifying differentially expressed transcripts: ~12,000 transcripts were identified as differentially expressed between B and T cells called using a q value of less than 0.01, using Ballgown (Fig. 5A, Table 2).

An observation that immediately stands out from Fig. 5C is the difference in the number of transcripts identified before and after the StringTie –merge run. Here, annotated transcripts refer to transcripts for which a reference gene has been determined by StringTie. StringTie –merge functions to create a non-redundant set of transcripts by comparing assembled transcripts across all samples (Pertea, M. 2016). So, the significant decrease in the number of novel transcripts can be because these transcripts are present in only a few samples subsequently leading to their elimination after the merge step. Also, some of the novel transcripts could be incompletely assembled transcripts which after the merge step, wherein required read information for complete assembly is gathered from other samples, leads to a decrease in the number of novel transcripts. This can lead to an increase in the number of annotated transcripts if the completely assembled transcript is ascribed a reference gene by StringTie. Thus, although the merge step results in reduction of number of potential novel transcripts, it is recommended as it ensures that the generated merged transcripts, both novel and annotated, are genuine as determined by comparing read information across multiple samples.

Approximately 12000 DE transcripts were identified between B and T cells (Fig. 5A; Table 2). Although one might expect a number greater than the sum of 9889 transcripts unique to B cells and 7674 transcripts unique to T cells, a total of ~17000 transcripts to be differentially expressed, we identified only 12000. This is probably because of the parameter we set (qval <0.01), due to which some of the FPKM values for the ~17000 weren't statistically significant to be considered differentially expressed. From literature we know how B and T, especially the activated versions of these cells have different genomes due to somatic hyper-mutations, DNA rearrangements and recombinations (Smith M. 2016) yet an interesting observation is the similarity between the transcriptome of resting B and naive T cells, ~80% as determined by our analysis. From our data, we cannot comment on the similarity between the transcriptome of activated forms of B and T cells, but it suggests the possibility that most differences between B and T cell transcripts arise after their activation. Through this work, we successfully created our cell-specific transcriptomic

database and identified differentially expressed transcripts between the two cell types under consideration.

2. Creation of mPLsORF database



B

#sORFs_id	Organism	Source_database	Chromosome	Start_position	End_position	Strand	Amino_acid_sequence
mPLsORF0000000001	Mouse	sORFs.org	chr1	3216022	3216081	-	MQYKDDALIQERLEYETTL
mPLsORF0000000002	Mouse	sORFs.org	chr1	3216022	3216129	-	MRKAILAFECSPSPPRLQYKDDALIQERLEYETTL
mPLsORF0000000003	Mouse	sORFs.org	chr1	3216022	3216144	-	MLDRSLRKAILAFECSPSPPRLQYKDDALIQERLEYETTL
mPLsORF0000000004	Mouse	sORFs.org	chr1	3216022	3216180	-	MFRNRYPAWERHVLDRSLRKAILAFECSPSPPRLQYKDDALIQERLEYETTI
mPLsORF0000000005	Mouse	sORFs.org	chr1	3216022	3216252	-	MRPTAPPTPSSRPPRIEESVIKIDLFRNRYPAWERHVLDRSLRKAILAFECS
mPLsORF0000000006	Mouse	sORFs.org	chr1	3216022	3216261	-	MFQVRPTAPPTPSSRPPRIEESVIKIDLFRNRYPAWERHVLDRSLRKAILAF
mPLsORF0000000007	Mouse	sORFs.org	chr1	3216065	3216103	-	MFPISSKAAVQG
mPLsORF0000000008	Mouse	sORFs.org	chr1	3216104	3216157	-	MGETCVRSKPEKGHFSL
mPLsORF0000000009	Mouse	sORFs.org	chr1	3216138	3216206	-	MKNQSLKLTCSGIDIQHGRDMC
mPLsORF0000000010	Mouse	sORFs.org	chr1	3216138	3216242	-	MHHPPHHLDHHGLKNQSLKLTCSGIDIQHGRDMC
mPLsORF0000000011	Mouse	sORFs.org	chr1	3216203	3216271	-	MCTCVSSETNCTTHPIISTTTD

Fig. 6: Creation of mPLsORF database. **(A)** Steps for creation of mPLsORF database starting from curating sORF information from two source databases: sORFs.org and SmProt. No. of sORFs at each step is also mentioned. **(B)** A snippet of the mPLsORF database. **(C)** Length distribution plot of sORF encoded peptides. **(D)** Proportion of sORFs (402733/454120) with different annotations.

Information on sORFs that have been computationally and experimentally verified is available online. Even though there is evidence for translation of these sORFs, a systematic study of their expression in mouse B and T cells has not been performed. Furthermore, to conduct such an investigation, curating information from several sources was not sufficient due to different cataloguing methods and a large number of duplicates. Thus, we resorted to the creation of our own sORF database.

Information for our database was curated from two online sources: sORFs.org and SmProt containing 11,27,154 and 15,581 mouse sORFs, respectively. (Olexiouk V, et al., 2016; Hao Y, et al., 2017). We pre-processed these two databases extensively because of a large number of duplicates present within and between the two source databases and assigned a unique sORF id to each entry in the final filtered list. Our final sORF database contains a total of 454,120 sORFs with 440,136 entries from sORFs.org and 13,984 entries from SmProt (Fig. 6A). A snippet of the mPLsORF database is shown in Fig. 6B which in addition to the sORF ids and the genomic coordinates of the sORFs contains the amino acid sequences of their peptides curated from the two source databases.

Fig. 6C shows a plot of the length distribution of sORFs within the mPLsORF database. As expected, sORF length, in terms of the number of amino acids present in sORF peptides, varies from 2-100 aa. Also, most sORFs in our database have a length ranging from 10-20 aa. Using sORF annotations available from the two source databases, we plotted a pie chart to highlight the proportion of different types of sORF annotations present in the mPLsORF database (Fig. 6D). Only 402733/454120 sORFs were plotted owing to the lack of annotation information for the remaining sORFs. The three most abundant annotations are exonic sORFs, meaning sORFs located within the exonic part of a gene (Olexiouk V, et al., 2016) comprising ~52% of the total sORFs followed by ~30% sORFs located on lncRNAs and ~11% located in 5'UTRs of a gene. The remaining annotations comprise < 3% of total sORFs.

Although we extensively filtered out sORFs before compiling them into the mPLsORF database, there are still a few sORFs with the same chromosome location which had to be retained because of different corresponding amino acid sequences potentially representing alternatively spliced versions of each other.

Furthermore, a significant drop from 1,142,735 to 454,120 sORFs after extensive processing highlights the requirement for the creation of our own sORF database. From the pie chart depicting different annotation types of sORFs (Fig. 6D), one can infer that sORFs are present in both genic and noncoding regions. A similar distribution of sORFs that are transcribed and translated in our samples was also obtained (Fig. 10) indicating transcription and translation occurs from both genic and noncoding regions of the genome. Although most sORFs are localised in exonic regions, they are not necessarily in-frame with the exons that are translated into known proteins. Thus, exonic sORFs could be from +2 or +3 reading frame of an exon (thus noncoding by our definition), assuming +1 reading frame corresponds to translation of known proteins. Therefore a significant amount of sORFs are present within noncoding regions.

Several small peptides with lengths <20 aa are involved in signalling in Arabidopsis (Murphy et al., 2012). Thus given most sORF peptides are small, it is interesting to speculate their involvement in signalling albeit the exact mechanisms or the nature of signalling is not known. In our lab, to investigate possible sORF functions, we evaluated their amino acid sequences to identify potential binding sites for known protein domains using an online tool called Prosite (Sigrist et al., 2009). Interestingly, the preliminary results hint towards the possible involvement of sORFs in signalling, given their binding partner are mostly kinases or phosphatases (work not shown).

3. Creation of altORF database

In addition to sORFs, we investigate transcription and translation of altORFs in our samples. A list of altORFs identified in mouse, downloaded from Roucou's lab (Vanderperre et al., 2013) was used for our analysis. Some of these altORFs had unclear chromosome location wherein two or more chromosome numbers were attributed to the same altORF. To facilitate optimal downstream analysis, we filtered the downloaded list of altORFs as described in Fig. 7A and identified 2,15,320 altORFs for further use. A length distribution plot for altORFs after the filtering process is shown in Fig. 7B. altORF length, in terms of the number of amino acids in altORF peptides, varies from 29 to 1538 amino acid with a median length of 59 aa and about 90% of altORFs have a length <150 aa.

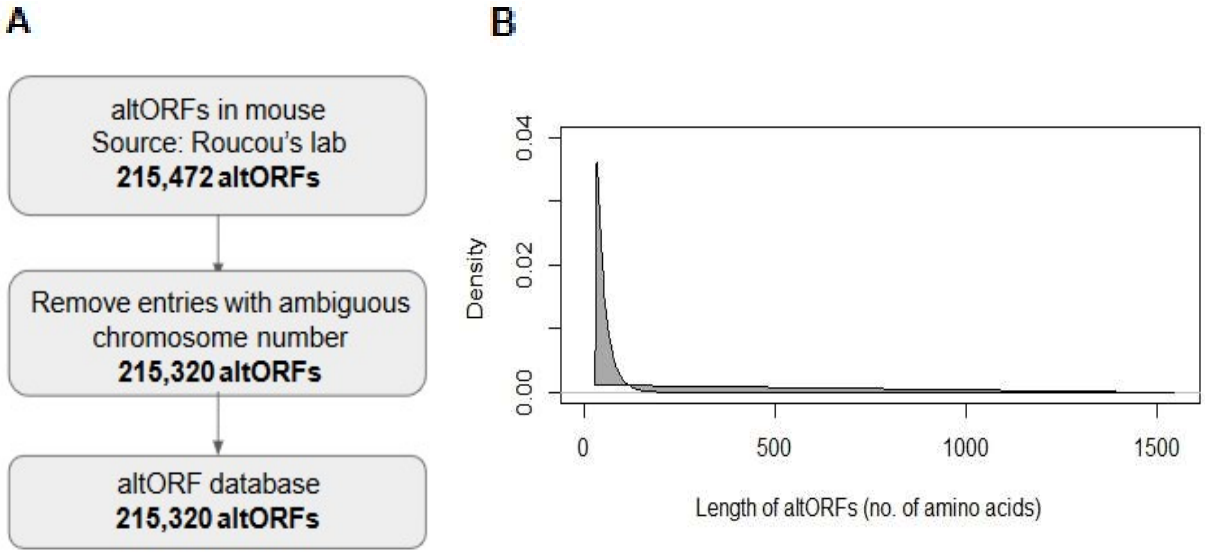


Fig. 7: Creation of altORF database. **(A)** Steps involved in creation of altORF database along with the number of altORFs at each step. **(B)** Length distribution plot for altORF encoded peptides.

In the paper published by Roucou's lab (Vanderperre et al., 2013), the median length of human altORFs is cited as 57 aa as opposed to 344 aa for known proteins. This paper (Vanderperre et al., 2013) also claims that altORFs are highly conserved sequences explaining the small difference in median length seen between human altORFs (57 aa) and mouse altORFs (59 aa). Additionally, even after the filtering step, there were some altORFs with the same chromosome coordinates which were retained because of difference in corresponding amino acid sequence, indicating the possibility of alternative splicing of altORF transcripts. In comparison to the mPLsORF database, the altORF database does not have any associated strand information as it was not available in the downloaded source. This work allowed us to establish a list of altORFs along with their chromosome location and amino acid sequence, for further use.

4. Proteogenomic Analysis

To identify proteins isolated from our mouse samples, their peptide spectra were mapped to three databases: known protein database from UniProt (UniProt, 2017), sORF fasta database and altORF fasta database. After the mapping, proteins were filtered using a 1% FDR cutoff. We, therefore, identified 1850 known proteins, 3604

sORF proteins and 2104 altORF proteins in our mouse samples which were used for further analysis (Fig. 8). Proteins which did not match any of the three databases and thus whose identity is unknown are further evaluated by mapping their spectra to amino acid sequences in our B and T cell-specific proteogenomic database. Although peptide spectra are usually mapped to known protein databases in proteomic analysis and to proteogenomic databases in proteogenomic analysis, we introduced an additional step wherein the spectra are mapped to sORF and altORF fasta databases. This workflow thus allows for the identification of novel proteins that are not sORF, altORF or known proteins, by mapping the unmatched peptide spectra to B and T cell proteogenomic databases. This work is currently being performed in our lab.

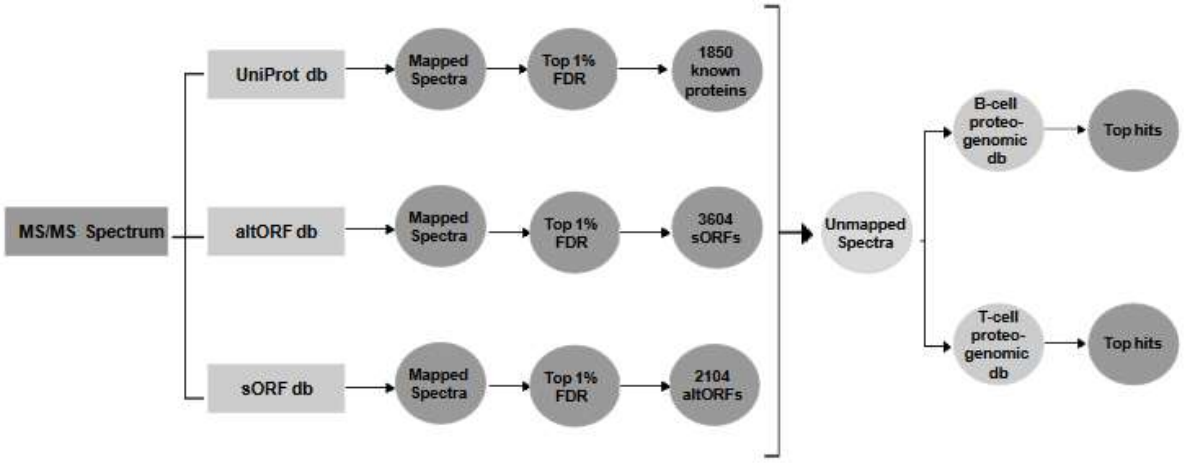


Fig. 8: Proteogenomic workflow used to identify proteins in mouse samples by mapping peptide spectra generated by mass-spectrometry analysis to each of the three databases (UniProt, altORF and sORF databases) individually.

5. Evidence for transcription and translation of sORFs and altORFs in mouse B and T cells

After the successful creation of the sORF database, altORF database and B and T cell-specific transcriptomic database, we evaluated if there is evidence for transcription and translation of these novel ORFs in mouse B and T cells. We utilised bedtools intersect for coordinate mapping such that a positive overlap between novel ORFs and B or T cell transcripts indicates that a transcript is encoded by a genomic

region designated as sORF or altORF and thus we assume that the transcript is a sORF or altORF transcript.

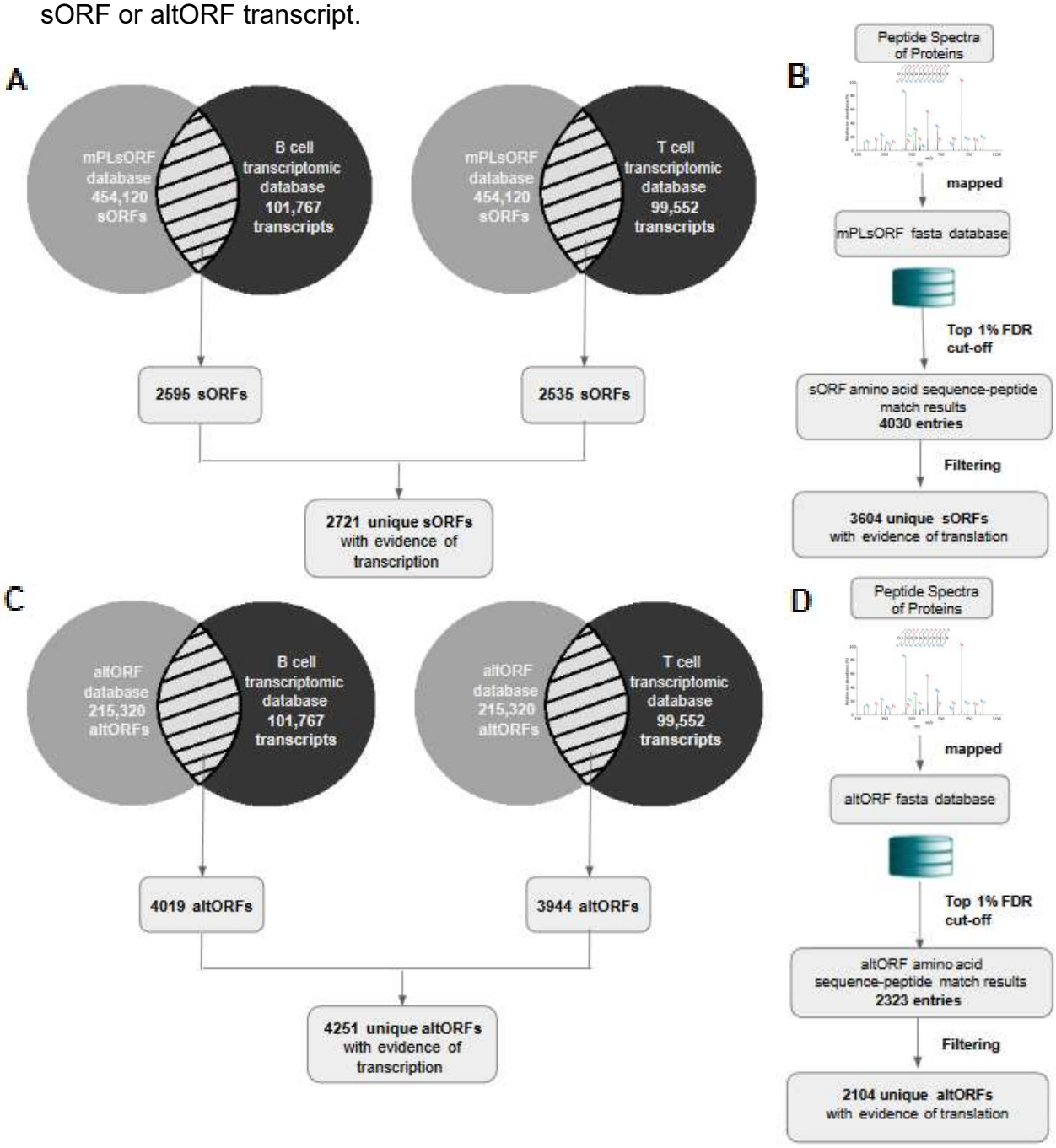


Fig. 9: Evidence for transcription and translation of sORFs and altORFs **(A)** Identifying sORFs with evidence for transcription by mapping sORF coordinates in mPLsORF database to B/T cell transcriptomic database, along with the number of sORFs/transcripts at each step. **(B)** Identifying evidence for translation of sORFs by mapping peptide spectra of proteins to sORF fasta database along with the number of peptides at each step. **(C)** Identifying evidence for transcription of altORFs **(D)** Identifying evidence for translation of altORFs.

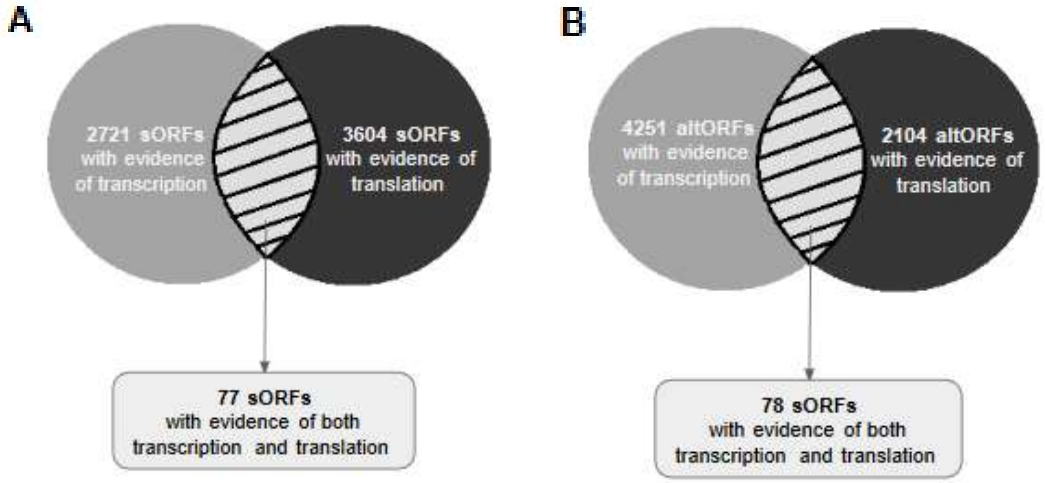


Fig. 10: (A) Identifying sORFs with evidence of both transcription and translation **(B)** Identifying altORFs with evidence of both transcription and translation

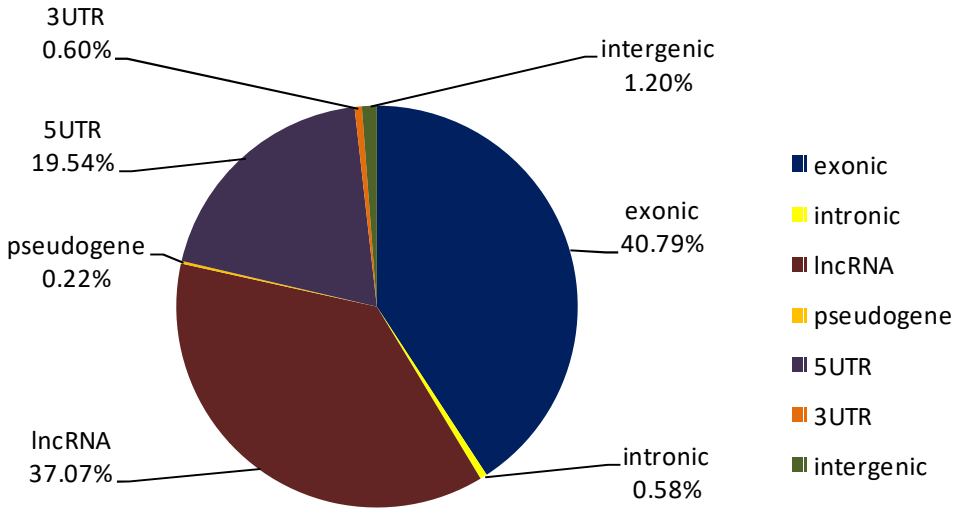


Fig. 11: Proportion of sORFs (4494/6248) with different annotations. These are sORFs with evidence of transcription or translation in mouse B and T cells.

Total no. of entries identified after mapping sORFs to peptide spectra	No. of sORFs after Filter 1: remove 'cRAP'	No. of sORFs after Filter 2: Select sORFs with 'High' FDR	No. of sORFs after Filter 3: Remove sORFs with no abundance values
4030	3988	3988	3604

Table 3: Different filtering steps employed as well as the number of sORF peptides identified at each step of identification of sORFs with evidence of translation.

Total no. of entries identified after mapping altORFs to peptide spectra	No. of altORFs after Filter 1: remove 'cRAP'	No. of altORFs after Filter 2: Remove altORFs with no abundance values
2323	2289	2104

Table 4: Different filtering steps employed as well as the number of altORF peptides identified at each step of identification of altORFs with evidence of translation.

Evidence in mouse B and T cells	Total No. of unique sORFs
sORFs with evidence of transcription	**2721**
sORFs with evidence of translation	**3604**
sORFs with evidence of transcription & translation	**77**
sORFs with evidence of transcription or translation	**6248**

Table 5: Total number of sORFs identified with evidence of transcription or translation in mouse B and T cells

Evidence in mouse B and T cells	Total No. of unique altORFs
altORFs with evidence of transcription	**4251**
altORFs with evidence of translation	**2104**
altORFs with evidence of transcription & translation	**78**
altORFs with evidence of transcription or translation	**6433**

Table 6: Total number of altORFs identified with evidence of transcription or translation in mouse B and T cells

In the overlap between 454,120 sORFs and 101,767 B cell or 99,552 T cell transcript coordinates, we identified 2595 sORFs in B cells and 2535 sORFs in T cells with evidence of transcription which upon comparison revealed, 2721 unique sORFs in mouse B and T cells with evidence of transcription (Fig. 9A). Similarly, in the overlap between 215,320 altORFs and 101,767 B cell or 99,552 T cell transcript coordinates, we identified 4019 altORFs in B cells and 3944 altORFs in T cells with evidence of transcription that further led to the identification of 4251 unique altORFs in mouse B and T cells with evidence of transcription (Fig. 9C). Moreover, to identify evidence of translation of novel ORFs we used peptide spectra mapping results from our proteogenomic analysis and filtered them first to remove contaminant 'cRAP' proteins then filtered the entries to select for only proteins with high FDR values and finally removed protein entries with zero abundance across all samples (Table 3,

Table 4). This allowed for the identification of 3604 sORFs and 2104 altORFs with evidence of translation in mouse B and T cells (Fig. 9B, Fig. 9D).

Comparing sORFs identified with evidence of transcription and translation we found 77 sORFs with evidence of both transcription and translation in our samples and a total of 6248 sORFs with evidence of transcription or translation (Fig. 10A; Table 5). A similar comparison for altORFs revealed 78 altORFs with evidence of both transcription and translation and 6433 unique altORFs with evidence of either transcription or translation (Fig. 10B; Table 6). Furthermore, annotations for 4494 out of 6248 sORFs with evidence of transcription or translation showed that the most abundant sORF annotations are exonic (~40%) followed by lncRNAs (~37%) and 5'UTRs (~19%) while the remaining annotations amount to only a mere ~3% (Fig. 11).

From our data, we observe that more altORF transcripts were identified compared to sORF transcripts. Although this could mean that more altORF transcripts are indeed present in the cell, there is another factor that needs to be considered. The only difference in parameters used for sORF coordinate vs altORF coordinate mapping to transcripts using bedtools intersect was the use of '–s' parameter that takes into consideration the strand information for elements being mapped to each other (Quinlan and Hall, 2010). For altORFs without any strand information, we had to forgo this parameter thereby introducing the possibility that more altORFs are being identified with evidence of transcription than if the '–s' parameter were imposed. Thus, there is a possibility of overrepresentation of altORFs with evidence of transcription, but with the current dataset, we cannot resolve this issue. Even then, a subset of these altORFs is transcribed within B and T cells.

Although for sORFs in the mPLsORF database there is evidence for translation, the cells in which said translation is observed varies from mouse embryonic stem cells to fibroblasts, and very few in comparison are from B cells (Olexiouk V, et al., 2016). This prompted us to investigate translation of sORFs in B and T cells. So, although our sORF database has 4,54,120 sORFs since most of them are from different cell lines, we obtained only 2721 sORFs with evidence of transcription. This small number is also dependent on the transcript assembly done by StringTie which as mentioned before can lead to the loss of novel transcripts after the merge step. All

we can claim with this data in hand is that we identified a subset of sORFs in B and T cells that have evidence of transcription.

Another observation that immediately stands out is the discrepancy between the numbers of transcribed vs translated novel ORFs. Mainly, why are there fewer sORF transcripts compared to sORF proteins and why is there double the number of altORF transcripts compared to altORF proteins? Possible reasons for this, although not exhaustive, are discussed. The possibility that one transcript codes for multiple proteins after undergoing alternative splicing can lead to the identification of a larger number of proteins than transcripts. Also, some sORF transcripts possibly have very low abundances or a short half-life reducing chances of their identification. Finally, loss of novel transcripts during the read alignment and transcript assembly process due to inherent limitations of the transcript assembly tool used (StringTie merge in creating a non-redundant transcript set can cause loss of novel transcript unique to a particular sample) (Pertea et al., 2016) or we lost some possible transcripts at the alignment stage where we aligned the genome not to a cell-specific reference genome, which is unfortunately not available, but rather to a strain-specific reference genome. Our analysis is unable to distinguish between these speculated reasons, and therefore the exact cause of the discrepancy between identified transcripts vs proteins may be one or a combination of reasons mentioned above.

Also since only 77 sORFs and 78 altORFs with both evidence of transcription and translation have been identified, it implies that we have identified a few novel ORF transcripts without their corresponding proteins and a few novel ORF proteins without their corresponding transcripts. One reason could be that although the peptides are present, their small size (as small as 2 aa) makes it difficult to isolate and study them and further they may not pass the FDR filter imposed. We also don't know if all sORFs and altORFs are translated, so another possibility that needs to be considered is that not all novel ORF transcripts have corresponding proteins and instead these transcripts possibly exert their functions at the RNA level. For example, miRNAs bind to complementary RNAs and silence the expression of the latter (Wahid et al., 2010). As for peptides without corresponding transcripts reasons mentioned in the previous paragraph, are potential causes. Finally, through this work, we were able to ascertain the presence of transcribed and translated sORFs and altORFs in mouse B and T cells.

6. Differential expression analysis of sORFs and altORFs

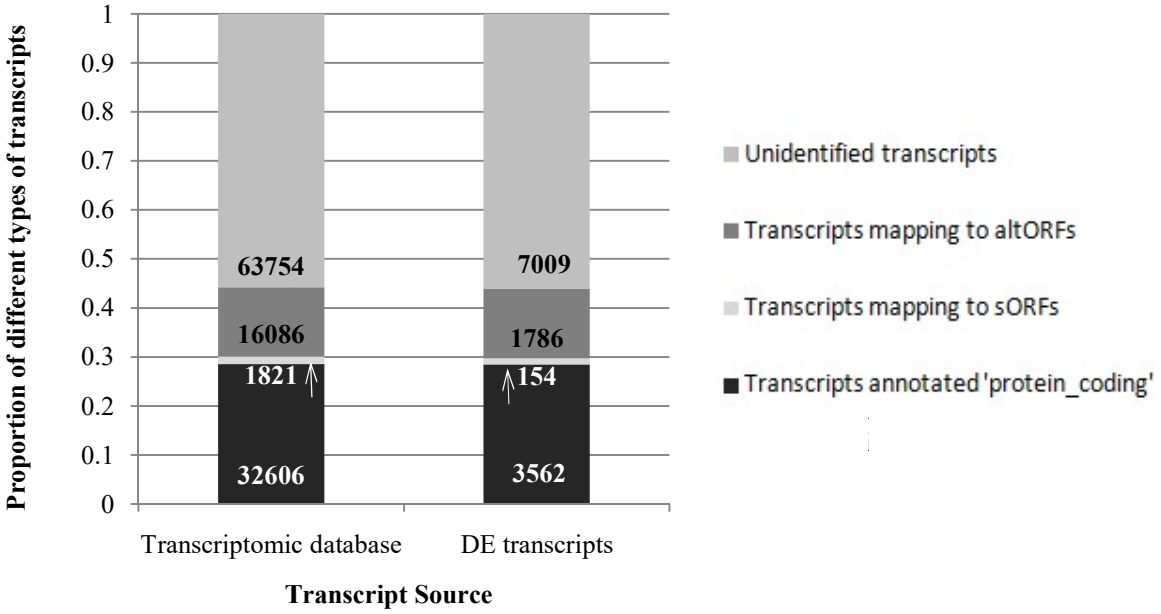


Fig. 12: Proportion of different types of transcripts (y-axis) identified from different transcript datasets (x-axis). Transcriptomic database indicates transcripts identified after StringTie merge. DE transcripts are differentially expressed transcripts identified using Ballgown analysis. Transcripts annotated protein coding are marked in black; sORF transcripts are represented by white bars; altORF transcripts are shown as dark grey and transcripts not belonging to these 3 categories and labelled 'unidentified transcripts' are coloured light grey.

Transcript Source ->	Transcriptomic database	DE transcripts
Transcripts annotated 'protein_coding'	32606	3562
Transcripts mapping to sORFs	1821	154
Transcripts mapping to altORFs	16086	1786
Unidentified transcripts	63754	7009

Table 7: Different types of transcripts in our transcriptomic database and differentially expressed transcripts list along with the number of transcripts of each type is shown.

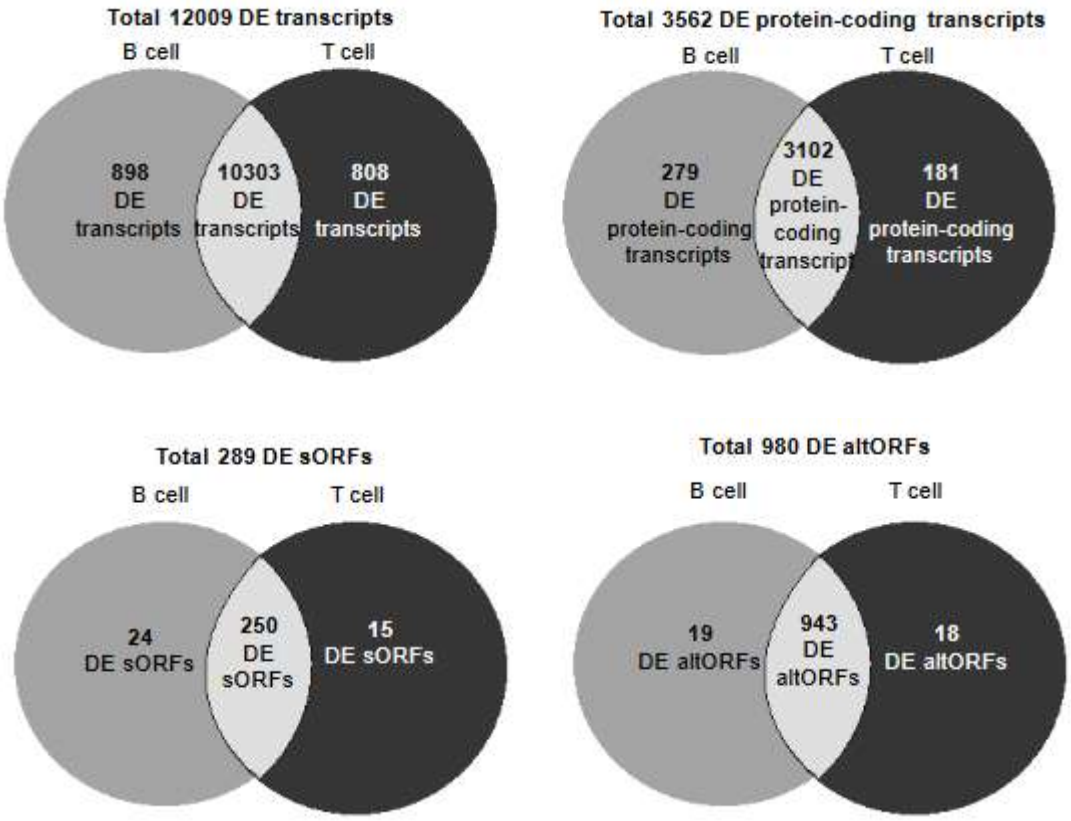


Fig. 13: Classification of differentially expressed transcripts, DE protein coding transcript, DE sORF transcripts and DE altORF transcripts into B and T cell based on transcript FPKM values

We identified evidence for transcription and translation of two novel ORFs namely sORFs and altORFs in mouse B and T cells. Next, we explore the possibility of differences in their transcript expression levels between B and T cell and whether this information can be used to distinguish between the two cell types. Here, we work with two transcript datasets, one our transcriptomic database generated by StringTie merge and the other a list of differentially expressed (DE) transcripts produced after Ballgown analysis. Transcripts were categorised as protein coding by extracting biotype information for their ensembl transcript ids from ensembl biomart (Ensembl Biomart, Ensembl genes 91).

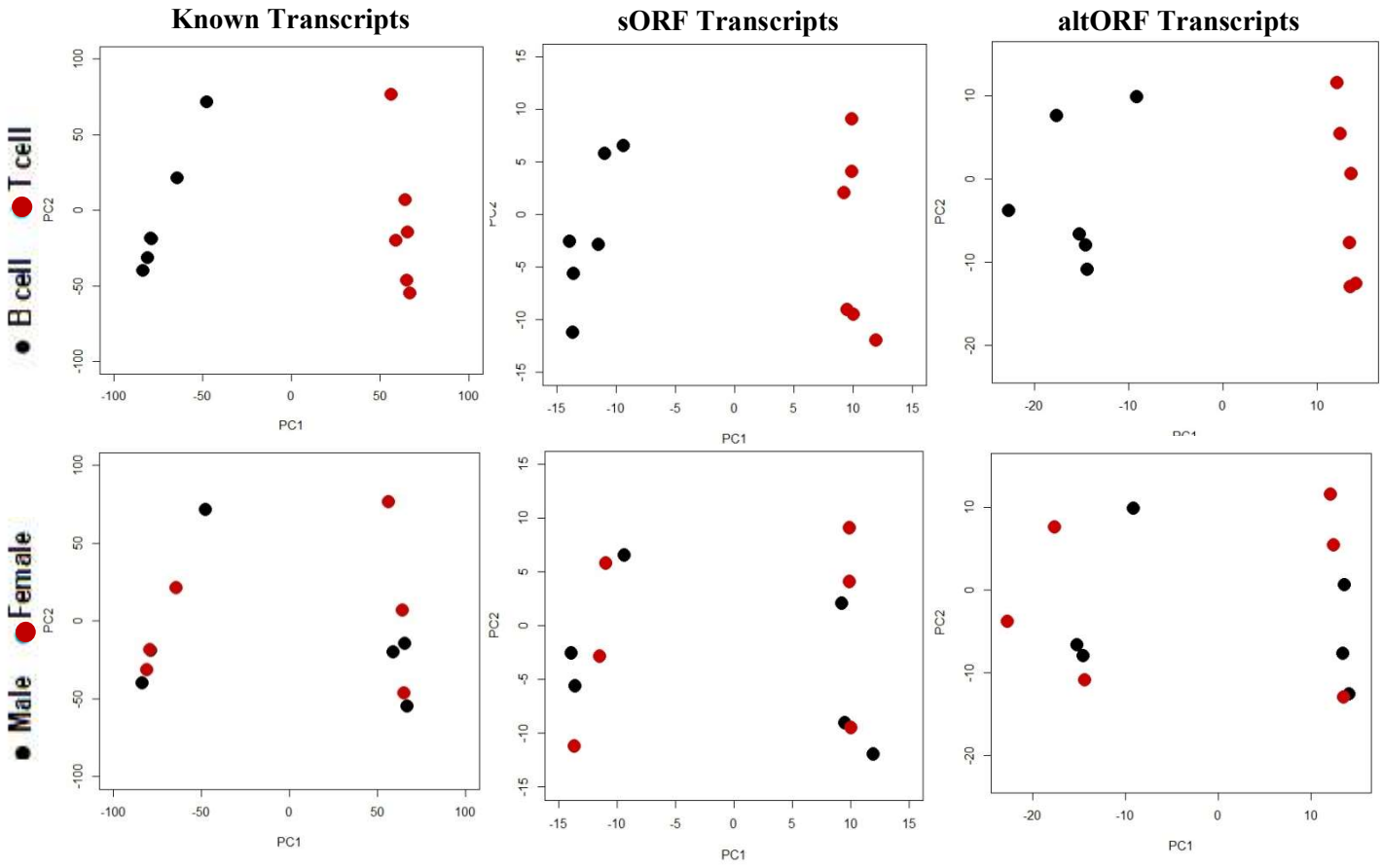


Fig. 14: PCA analysis of transcript expression levels of known, sORF and altORF transcripts. Top panel contains PCA analysis distinguishing between B cells (black) and T cells (red). Bottom panel indicates PCA analysis is unable to distinguish between male (black) and female (red) of a particular cell-type based on the transcript expression levels

We identified 32606 and 3562 protein-coding transcripts in the transcriptomic and DE transcript lists respectively. 1821 and 16086 transcripts from the transcriptomic database mapped to sORFs and altORFs in comparison to 154 and 1786 DE transcripts that mapped to sORFs and altORFs respectively. The identity of 63754 transcripts from the transcriptomic database and 7009 DE transcripts could not be ascertained (Table 7). From Fig. 12 we observe that the proportion of different transcript annotation is almost the same for the transcripts in the transcriptomic database and those which are differentially expressed. Also, most transcripts are unidentified transcripts with the next common annotation being protein-coding followed by altORF and sORF transcripts.

To evaluate whether there are any DE transcripts uniquely expressed in B or T cells, we classified the transcripts depending on their FPKM values. If FPKM values corresponding to a transcript were zero for all samples of a particular cell type, the transcript was deemed absent in that specific cell. 898 DE transcripts unique to B cells and 808 DE transcripts unique to T cells were identified with 10303 common to both cell types. Of the 3562 protein-coding DE transcripts 279 were unique to B and 181 unique to T with 3102 common to both. Similarly, 24 DE transcripts in B, 16 in T and 250 in both were identified for sORF transcripts whereas 19 altORF DE transcripts in B, 18 in T and 943 in both B and T cells were identified (Fig. 13). Furthermore, we did a PCA analysis of transcript expression levels of known transcripts, sORF transcripts and altORF transcripts. Although the transcript expression levels could distinguish between B cell and T cell, it could not distinguish between the male and female of the same cell type (Fig. 14).

Some of the transcripts annotated as unidentified (Fig. 12) contain long intergenic noncoding RNAs (lincRNA), small nucleolar RNAs (snoRNA), mitochondrial RNA (mt-RNA) etc. The nature of the remaining unidentified transcripts without ensembl transcript ids could not be determined. Thus the unidentified category includes transcripts from noncoding regions and other transcripts whose identity could not be ascertained. Also, as seen from Fig. 13, almost 88% of DE transcripts are common to both cell types and only a very small fraction is unique to either cell type. It is known that there are differentially expressed genes between B and T cells (Painter et al., 2011) and therefore the identification of differentially expressed protein-coding transcripts is expected but what is interesting is that sORFs and altORFs are differentially expressed as well. Moreover, our PCA analysis showed that sORFs and altORFs are differentially expressed between B and T cells and their transcript expression levels are sufficient to distinguish between these two cell types. Thus, novel ORFs, as well as transcripts from noncoding regions, can distinguish between B and T cells. We traced back the original cell line for a few of the sORFs from sORFs.org which were exclusively expressed in B or T cells. Interestingly, these sORFs were identified in mouse fibroblasts, mouse brain cells and even mouse stem cells (Olexiouk V, et al., 2016).

It is interesting to speculate the functions of these sORFs identified in B and T cells as well as different cell lines like mouse fibroblasts, mouse brain cells and mouse

stem cells. Are some sORFs involved in regulation of common cellular processes explaining why they are prevalent in multiple cell types or do sORFs regulate cell-specific processes and thus are uniquely expressed in only some cell types? Maybe it is a combination of both. Also, proteins encoded by altORFs add to the diversity of the proteome of a cell (Vanderperre et al., 2013), thus possibly allowing B and T cell in addition to known mechanisms of increasing diversity at the genomic level (Smith,M., 2016), to increase diversity at the proteome level too. These are all speculations which need to be further worked on but our data clearly reveals that there is differential expression of sORFs and altORFs between mouse B and T cells and that sORF and altORF transcripts can be used to distinguish between these two cell types.

7. Differential methylation analysis

DMRs near ->	DE sORFs	DE altORFs	DE protein-coding transcripts
Upstream region	117	199	1712
Downstream region	139	257	1679
Body	1398	28497	24910

Table 8: Number of DMRs identified in upstream, body and downstream regions of DE sORFs, DE altORFs and DE protein-coding transcripts

LINE/SINE near ->	DE sORFs	DE altORFs	DE protein-coding transcripts
Upstream region	453	2533	6972
Downstream region	330	2494	6255

Table 9: Number of LINE/SINE repeat elements identified in upstream and downstream regions of DE sORFs, DE altORFs and DE protein-coding transcripts

DMRs in LINE/SINE near ->	DE sORFs	DE altORFs	DE protein-coding transcripts
Upstream region	0	3	21
Downstream region	1	11	18

Table 10: Number of DMRs identified within LINE/SINE repeat elements found in the upstream and downstream regions of DE sORFs, DE altORFs and DE protein-coding transcripts

Regions	sORFs near DE protein-coding transcripts	DMRs in sORFs near DE protein-coding transcripts
Upstream region	24	4
Downstream region	22	1

Table 11: Number of sORFs identified in upstream and downstream regions of DE protein-coding transcripts and the number of DMRs found within these sORFs.

In our attempt to identify regulation of novel ORFs by differential methylation we first identified 140,929 DMRs between B cells and T cells using bsseq package. We then mapped these DMRs to the upstream, body and downstream regions (3000 bp windows upstream and downstream) of DE sORFs, altORFs and protein-coding transcripts (DE elements). We identified 117 DMRs in the upstream, 139 in the downstream region and 1398 in the body of DE sORFs, 199 DMRs in the upstream, 257 DMRs in the downstream region and 28497 in the body of DE altORFs and 1712 DMRs in the upstream, 1679 DMRs in the downstream region and 24910 in the body of DE protein-coding transcripts (Table 8).

To evaluate the possibility of regulation of DE elements by LINE/SINE repeat elements, we mapped LINE/SINE to upstream and downstream regions of these DE elements. 453 and 330 unique LINE/SINE elements were found in the upstream and downstream regions of DE sORFs, 2533 and 2494 unique LINE/SINE elements in the upstream and downstream regions of DE altORFs and 6972 and 6255 unique LINE/SINE elements were found in the upstream and downstream regions of DE protein-coding transcripts (Table 9). Furthermore, mapping DMRs to these LINE/SINE elements identified near DE elements resulted in the identification of 0 and 1 DMRs within LINE/SINE in the upstream and downstream regions of DE sORFs, 3 and 11 DMRs within LINE/SINE in the upstream and downstream regions of DE altORFs and 21 and 18 DMRs within LINE/SINE in the upstream and downstream regions of DE protein-coding transcripts (Table 10). Additionally, to evaluate if differentially methylated sORFs regulate expressions of DE protein-coding transcripts, we first mapped sORFs to upstream and downstream regions of protein-coding transcripts and then mapped DMRs to these sORFs. We thus identified 4 and 1 DMRs in sORFs upstream or downstream to protein-coding transcripts (Table 11).

From this analysis, we identified DMRs in the upstream and downstream regions of DE elements. We also determined the presence of LINE/SINE repeat elements near these DE elements and furthermore identified DMRs within these repeat elements. Finally, we found sORFs in the upstream and downstream regions of protein-coding transcripts and these sORFs contained DMRs within them. Thus, we have only identified DMRs in the regions mentioned above. Since the presence of DMRs does not guarantee regulation of transcript expression further work regarding identifying

significant and functional DMRs is required. For example, we could generate multiple random sets of DMRs and overlap it to DE elements and use this information to evaluate whether the DMRs identified near DE elements are significant. We could also undertake comparative genomics studies to identify any cis-regulatory regions in genomic regions containing DMRs. One observation that clearly stands out from our data is the lack of DMRs in LINE/SINE near sORFs highlighting the likelihood that sORFs are not regulated by repeat elements.

Conclusion

We investigated transcription and translation of two novel ORFs in mouse naive B and T cells. Using transcript coordinate mapping to novel ORFs and proteogenomic analysis, we identified evidence for transcription and translation of sORFs and altORFs in mouse B and T cells. Furthermore, we determined that sORF and altORF transcripts are differentially expressed between B and T cells and their transcript expression levels are significant and sufficient enough to distinguish between the two cell types. Additionally, to understand the regulation of these DE transcripts, we identified differentially methylated regions between B and T cells and found evidence for their localisation within upstream, body and downstream regions of DE sORFs and DE altORFs. Also, LINE/SINE elements were found near these DE novel ORFs, and although a few DMRs within these repeat elements near DE altORFs were identified, we found no significant evidence of DMRs in LINEs/SINEs present near DE sORFs indicating that DE sORFs are not regulated by repeat elements. Finally, we identified sORFs containing DMRs in the upstream and downstream regions of DE protein-coding transcripts. Although we identified DMRs within several upstream and downstream regions of DE elements further work is required to establish whether or not the genomic regions under consideration are involved in regulation of DE elements. This work in addition to providing a framework for a systematic analysis of novel ORFs highlights novel ORFs and their transcribed and translated products as significant components of the genome. Using results obtained from this work, sORF structure predictions and identification of disease variants within sORFs were conducted. Future work would involve identification of novel transcripts in our sample using the proteogenomic workflow described in this study and a rigorous analysis to determine whether DE protein-coding transcripts are regulated by sORFs thus ascribing a potential function to sORFs in mouse B and T cells.

References

1. Basrai, M.A., Hieter, P. and Boeke, J.D., (1997). Small open reading frames: beautiful needles in the haystack. Genome research, 7(8), pp.768-771.
2. Birney, E., Stamatoyannopoulos, J., Dutta, A., Guigó, R., Gingeras, T., Margulies, E., Weng, Z., Snyder, M., Dermitzakis, E., Stamatoyannopoulos, J., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799-816.
3. Broad Institute, Picard SortSam [online: http://broadinstitute.github.io/picard]
4. Cancer Genomics Cloud, Seven Bridges [online: www.cancergenomicscloud.org] Documentation available at: www.docs.cancergenomicscloud.org/docs
5. Ensembl Biomart, Ensembl genes 91 [online: asia.ensembl.org/biomart/martview]
6. Ferguson-Smith et al. (2017). *GEO Accession viewer*. [online: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94671]
7. Ferguson-Smith et al. (2017). *GEO Accession viewer*. [online: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94674]
8. Fu J, Frazee AC, Collado-Torres L, Jaffe AE and Leek JT (2017). *ballgown: Flexible, isoform-level differential expression analysis*. R package version 2.10.0.
9. GENCODE, version M12 [online: gencodegenes.org/mouse_releases/12.html]
10. Gong, Y., Chen, G., Chen, C., Kuo, R. and Shih, S. (2014). Computational Analysis and Mapping of Novel Open Reading Frames in Influenza A Viruses. *PLoS ONE*, 9(12), p.e115016.
11. Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K. and Shiu, S. (2010). sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics*, 26(3), pp.399-400.
12. Hanada, K., Higuchi-Takeuchi, M., Okamoto, M., Yoshizumi, T., Shimizu, M., Nakaminami, K., Nishi, R., Ohashi, C., Iida, K., Tanaka, M. and Horii, Y., (2013). Small open reading frames associated with morphogenesis are hidden in plant genomes. Proceedings of the National Academy of Sciences, 110(6), pp.2395-2400
13. Hansen KD, Langmead B and Irizarry RA (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10), pp. R83.[online: bioconductor.org/packages/release/bioc/html/bsseq.html]

14. Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, Zhang B, Zhang D, Qin Y, Yang F, Chen R. (2017) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief Bioinform. pii: bbx005 [online:bioinfo.ibp.ac.cn/SmProt/browse.php]
15. Hellens, R., Brown, C., Chisnall, M., Waterhouse, P. and Macknight, R. (2016). The Emerging World of Small ORFs. *Trends in Plant Science*, 21(4), pp.317-328.
16. Hrdlickova, B., de Almeida, R., Borek, Z. and Withoff, S. (2014). Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10), pp.1910-1922.
17. Ingolia, N. T. (2016). Ribosome footprint profiling of translation throughout the genome. *Cell*, *165*(1), 22-33.
18. Kapranov, P. and St. Laurent, G. (2012). Dark Matter RNA: Existence, Function, and Controversy. *Frontiers in Genetics*, 3.
19. Kastenmayer, J. (2006). Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Research*, 16(3), pp.365-373.
20. Kim D, Langmead B and Salzberg SL.(2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods,* 357–360. [online:ccb.jhu.edu/software/hisat2/manual.shtml]
21. Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F. and Kageyama, Y., (2010). Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science, 329(5989), pp.336-339.
22. Krueger, F. and Andrews, S. (2011). *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications*. Bioinformatics. 27(11):1571-2. [online: rawgit.com/FelixKrueger/Bismark/master/Docs/Bismark_User_Guide.html]]
23. Mouilleron, H., Delcourt, V. and Roucou, X. (2016). Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Research*, 44(1), pp.14-23.
24. Murphy, E., Smith, S. and De Smet, I. (2012). Small Signaling Peptides in Arabidopsis Development: How Cells Communicate Over a Short Distance. *The Plant Cell*, 24(8), pp.3198-3217.
25. Nesvizhskii, A. (2014). Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11(11), pp.1114-1125.

26. Nesvizhskii, A. (2014). Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11(11), pp.1114-1125.
27. Olexiouk, V., Crappé, J., Verbruggen, S., Verheggen, K., Martens, L. and Menschaert, G. (2016). *sORFs.org : a repository of small ORFs identified by ribosome profiling*. Nucleic Acids Res. 44(D1):D324-9 [online: sorfs.org/BioMart]
28. Painter, M., Davis, S., Hardy, R., Mathis, D. and Benoist, C. (2011). Transcriptomes of the B and T Lineages Compared by Multiplatform Microarray Profiling. *The Journal of Immunology*, 186(5), pp.3047-3057.
29. Pertea, M. (2016). *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown*. *Nature Protocols* volume 11, 1650–1667. GFFCompare [online: ccb.jhu.edu/software/stringtie/gffcompare.shtml]; StringTie [online: ccb.jhu.edu/software/stringtie/index.shtml?t=manual]
30. Prabakaran, S., Hemberg, M., Chauhan, R., Winter, D., Tweedie-Cullen, R., Dittrich, C., Hong, E., Gunawardena, J., Steen, H., Kreiman, G. and Steen, J. (2014). Quantitative profiling of peptides from RNAs classified as noncoding. *Nature Communications*, 5, p.5429.
31. Quinlan, A. and Hall, I. (2010). *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics. 26(6):841-2 [online:bedtools.readthedocs.io/en/latest/content/tools/]
32. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [online: https://www.R-project.org/; function prcomp documentation available at: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html]
33. Raj, A., Wang, S., Shim, H., Harpak, A., Li, Y., Engelmann, B., Stephens, M., Gilad, Y. and Pritchard, J. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife*, 5.
34. Samandi, S., Roy, A., Delcourt, V., Lucier, J., Gagnon, J., Beaudoin, M., Vanderperre, B., Breton, M., Motard, J., Jacques, J., Brunelle, M., Gagnon-Arsenault, I., Fournier, I., Ouangraoua, A., Hunting, D., Cohen, A., Landry, C., Scott, M. and Roucou, X. (2017). Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*, 6.
35. Searle, B. C. (2010), Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics, 10: 1265–1269 [online:proteomesoftware.com/products/scaffold/]

36. Sigrist, C., Cerutti, L., de Castro, E., Langendijk-Genevaux, P., Bulliard, V., Bairoch, A. and Hulo, N. (2009). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, 38(suppl_1), pp.D161-D166.
37. Smith,M. (2016). Unravelling Complexities in Genetics and Genomics: Impact on Diagnosis Counseling and Management. World Scientific Publishing Co. Pte, Ltd.
38. St. Laurent, G., Vyatkin, Y. and Kapranov, P. (2014). Dark matter RNA illuminates the puzzle of genome-wide association studies. *BMC Medicine*, 12(1).
39. UniProt: the universal protein knowledgebase. (2017). *Nucleic Acids Research*. 45: D158-D169 [online: uniprot.org/proteomes/]
40. Vanderperre, B., Lucier, J. and Roucou, X. (2012). HAltORF: a database of predicted out-of-frame alternative open reading frames in human. *Database*, 2012(0), pp.bas025-bas025.
41. Vanderperre, B., Lucier, J., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F. and Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS ONE*, 8(8), p.e70698. [online:roucoulab.com/p/downloads
42. Wahid, F., Shehzad, A., Khan, T. and Kim, Y. (2010). MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1803(11), pp.1231-1243.
43. Zhang, Y., Yang, L. and Chen, L. (2014). Life without A tail: New formats of long noncoding RNAs. *The International Journal of Biochemistry & Cell Biology*, 54, pp.338-349.
44. Zhao, Y., Li, H., Fang, S., Kang, Y., wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M. and Chen, R. (2015). NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Research*, 44(D1), pp.D203-D208. [online: noncode.org/]