Exosome mRNA localisation

Directing mRNA localisation to exosomes

Harry Biggs

A thesis submitted for the degree of Master of Science in Genetics, at the University of Otago, Dunedin, New Zealand

January 2019
Abstract

Exosomes are small vesicles containing proteins, miRNAs and mRNAs. Exosomes are secreted and taken up by neighbouring cells and the contents move into the recipient cell, where they are biologically active (e.g. the mRNA can be translated). Exosome contents are not present at the same proportions of a cell, meaning there must be active sorting of contents into exosomes. Currently the signals that determine which mRNAs are localised into exosomes are of great interest, as being able to manually add a localisation signal would allow for movement of contents that otherwise could not be spread between cells (such as the mRNA of a transmembrane protein). One such signal has been identified in glioblastoma mRNA 3’ UTRs and the results replicated in HEK293 cells, however, is yet to be tested in other cell types.

Astrocytes are a cell type that perform regulatory functions in the brain, by providing nutrients and aiding the development of neurons. They able to actively divide, unlike neurons meaning they are much easier to grow in a cell culture environment for study and research.

This project’s aim was to study mRNA isolated from astrocyte-derived exosomes, develop bioinformatic methods to determine any loading signals present in exosome mRNAs using external datasets, identify these signals present in astrocyte exosomes and then validate these signals \textit{in vitro}. Protocols for enriching a mixed neural cell culture for astrocytes were successfully established and exosomes were isolated. Unfortunately the RNA extractions from exosomes were unsuccessful. The \textit{in silico} analysis involved the processing of 8 individual datasets from independent labs, re-analysing all raw data and generating new datasets containing mRNA 3’ UTRs to suit this projects aims. RNA sequence motifs, miRNA binding sites, and RNA secondary structures were all carefully analysed, revealing large differences in putative 3’ UTR elements present in exosome mRNAs between different cell types. These data suggest that mRNA localisation to exosomes may be both species and cell type specific.
Acknowledgements

No thesis is the work of one person and this one is no exception. I would first and foremost like to thank Alison Clare, Hannah Best, Isaiah Cheong and Hollie Wicky for teaching me the skills necessary for the lab work portion of this thesis, and for your endless patience and wisdom. I genuinely could not have finished without you. To everyone else in my lab, Jade, Tobi, Sophie, Jaz, Maddie, Ben, Steph, Luci and Indranil thank you for your optimism, chat and motivation.

Thank you to my supervisor, Steph Hughes for giving me guidance through this project. You were always there when I needed you.

Thank you to my parents, who constantly checked in to offer help and advice. You were both crucial to this project.

Darren and DJ, thank you for always being there to answer my endless questions about bash, how servers worked and programs I had no idea about, your patience, expertise and kidness never made me feel unwelcome.

To people in the Gardner lab, Steph, Beth and Paul thank you for explaining math, listening and giving advice, it was never wrong or unappreciated. I am in awe of all of your drive and skill.

To my theatre troupe Sacrilege, without that outlet I would have gone more insane than I did.

To the rest of my friends, Malcolm, Shereen, John, Kieran, Josh, Sarah, Shannon, Astra, Gabe, Riku, Tom and so many more, thank you for always making me laugh and smile.

And finally to my wonderful partner, you kept the wind in my sails and my boat moving forward. Every day is a joy when it is spent with you.
# Contents

Abstract i
Acknowledgements ii
Contents iii
List of Figures vi
List of Tables viii
List of abbreviations ix

## 1 Introduction
1.1 Project Overview ......................................................... 1
1.2 Exosomes ................................................................. 2
  1.2.1 Overview .......................................................... 2
  1.2.2 Biogenesis and contents ........................................... 2
  1.2.3 Exosome functions .................................................. 5
  1.2.4 Molecular signals for exosome loading .......................... 5
  1.2.5 Zip code identification in exosomal mRNA ................. 6
1.3 Astrocytes ............................................................... 8
  1.3.1 Overview .......................................................... 8
  1.3.2 Regular function ................................................... 8
  1.3.3 Astrocytes and exosomes ........................................... 9
  1.3.4 Micro-RNAs .......................................................... 10
1.4 RNA localisation signals .............................................. 10
1.5 Bioinformatic predictions and models ............................... 11
  1.5.1 Motif searches ..................................................... 11
  1.5.2 RNA secondary structure .......................................... 13
  1.5.3 Aims ................................................................. 14

## 2 Methods
2.1 Animal protocols ....................................................... 16
### 2.1.1 Euthanasia

#### 2.2 Cell culture

- **2.2.1 Culture preparation**
- **2.2.2 Culturing mouse brain cells**
- **2.2.3 Cell maintenance**
- **2.2.4 Passaging cells**
- **2.2.5 Removal of microglia**
- **2.2.6 Mixed neural cultures**
- **2.2.7 Fixing cells**
- **2.2.8 Immunocytochemistry**
- **2.2.9 Image quantification**
- **2.2.10 Preparing exosome-free media**
- **2.2.11 Exosome isolation**
- **2.2.12 Electron Microscopy**
- **2.2.13 Cell RNA extraction**
- **2.2.14 Exosome RNA extraction**
- **2.2.15 Qubit quality control**

### 2.2.13 Cell RNA extraction

- **2.3 Bioinformatics overview**

- **2.3.1 Dataset generation**
- **2.3.2 MEME**
- **2.3.3 TOMTOM**
- **2.3.4 Cisfinder**
- **2.3.5 Glam2**
- **2.3.6 Targetscan**
- **2.3.7 CMfinder and Infernal**

### 3 Results

#### 3.1 Cell culture results

- **3.1.1 Characterisation of primary astrocyte cultures**
- **3.1.2 Characterisation of exosome isolates from cell culture media**
- **3.1.3 RNA extraction from isolated vesicles**

#### 3.2 Bioinformatic analysis of exosome mRNA

- **3.2.1 Dataset overview**
- **3.2.2 Very little intersection of differentially expressed genes in exosomes from different cell types is observed**
- **3.2.3 MEME analysis identifies enriched motifs**
- **3.2.4 A motif comparison using TOMTOM identifies consistently predicted RBPs within species**
- **3.2.5 Cisfinder used for further motif identification**
4 Discussion

4.1 Isolation and content characterisation of astrocyte exosomes

4.1.1 Astrocyte culture characterisation

4.1.2 Exosome isolation

4.2 In silico prediction of mRNA localisation signals for exosomes

4.2.1 Different cell types have distinct exosome gene populations

4.2.2 Motifs are enriched in exosomes but are not common across cell types

4.2.3 Previously reported localisation signals are not enriched in all cell types

4.2.4 The puzzle of exosome localisation is still missing some pieces

4.3 Future directions

4.4 Overall conclusion

5 Appendix

5.1 Solutions and Buffers

5.2 Antibody dilutions

5.3 Quantification of CD68 levels

5.4 MEME command line arguments

5.5 TOMTOM command line arguments

5.6 Cisfinder search parameters

5.7 Bolukbasi PFM

5.8 GLAM2 command line arguments

5.9 CMfinder and Infernal command line arguments

5.10 Description of online appendices
# List of Figures

1.1 Exosome biogenesis ................................................. 3  
1.2 Core Bolukbasi sequence ........................................... 6  
1.3 Bolukbasi stem loop .................................................. 7  
1.4 Astrocyte functions .................................................. 9  
1.5 Cisfinder algorithm ................................................... 12  
1.6 Potential loading mechanisms ....................................... 15  

2.1 Bioinformatic workflow ............................................... 22  
2.2 Dataset generation workflow ......................................... 24  
2.3 Dataset generation .................................................... 25  

3.1 Immuno timecourse ................................................... 30  
3.2 Neuronal cultures .................................................... 31  
3.3 Shaken cells comparison .............................................. 31  
3.4 Exosome TEM .......................................................... 32  
3.5 RNA rin scores ......................................................... 33  
3.6 Gene overlap .......................................................... 35  
3.7 MEME motifs .......................................................... 36  
3.8 Position bar graphs .................................................... 37  
3.9 Motif position means ................................................ 38  
3.10 MEME kmers .......................................................... 39  
3.11 RNA binding proteins ............................................... 41  
3.12 RBP controls .......................................................... 42  
3.13 Cisfinder results ...................................................... 43  
3.14 GLAM2 motif .......................................................... 44  
3.15 GLAM2 motif frequency .............................................. 45  
3.16 Bolukbasi motif frequency ........................................... 46  
3.17 Targetscan miRNAs ................................................... 47  
3.18 CMfinder models ...................................................... 49  

4.1 Exosome isolate comparison ......................................... 51  
4.2 Final conclusions visualised ......................................... 54
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Cisfinder parameters</td>
<td>65</td>
</tr>
<tr>
<td>5.2 Bolukbasi PFM</td>
<td>65</td>
</tr>
</tbody>
</table>
List of Tables

2.1 GEO experiments downloaded ........................................ 25
3.1 The names, species and cell origins of the datasets generated .................. 34
3.2 The number of genes present in each dataset .................................. 34
3.3 The size in bases of each dataset generated .................................... 34
3.4 Consistently found potential binding proteins discovered ...................... 42
3.5 The functions of the overlapping miRNAs identified by TargetScan ........... 47
5.1 Antibodies used for immunocytochemistry ...................................... 63
5.2 Cell counts used for microglia quantification (expressed as averages) ........ 63
**List of abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD68</td>
<td>Cluster of differentiation 68</td>
</tr>
<tr>
<td>CM</td>
<td>Covariance model</td>
</tr>
<tr>
<td>CNS</td>
<td>Central nervous system</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation maximisation</td>
</tr>
<tr>
<td>ESCRT</td>
<td>Endosomal sorting complexes required for transport</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal bovine serum</td>
</tr>
<tr>
<td>GFAP</td>
<td>Glial fibrillary acidic protein</td>
</tr>
<tr>
<td>GLAM2</td>
<td>Gapped local alignment of motifs 2</td>
</tr>
<tr>
<td>HEK293</td>
<td>Human embryonic kidney cell line</td>
</tr>
<tr>
<td>MAP2</td>
<td>Microtubule associated protein 2</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MEME</td>
<td>Multiple EM for motif elicitation</td>
</tr>
<tr>
<td>miRNA</td>
<td>MicroRNA</td>
</tr>
<tr>
<td>MVB</td>
<td>Multivesicular body</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate buffer saline</td>
</tr>
<tr>
<td>RBP</td>
<td>RNA binding proteins</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

DISCLAIMER: Due to a problem highlighted with this analysis, some data (but not methodology) presented in this thesis are likely to be incorrect. Data used in the bioinformatic analyses was generated using data from series matrices. No Log2 transformation was applied to these matrices before proceeding to downstream processing, as it was thought that this transformation had already been applied. Additionally, the cell type HMEC-1 should be discluded from all results as the format was not a series matrix. The outcome of this error early on in the process means some results presented in section 3.2 and it’s associated Figures would likely be different. As the results will be affected, discussion sections 4.2, 4.3 and 4.4 will be affected as well. However none of the methodology presented in section 2.3 aside from subsection 2.3.1 will be affected, merely the results stemming from these methods.

1.1 Project Overview

The initial concept of this project was to establish a protocol for growing astrocytes, to then extract their exosomes. RNA would then be extracted from these exosomes and then sequenced using high fidelity sequencing techniques. Bioinformatic analysis would then be performed on this data to find putative repeating signals or ‘motifs’ that may be involved in exosome localisation, which would then be confirmed in the lab. However, with large delays and setbacks in the cell culture process, more bioinformatic analysis was undertaken using previously published datasets to find more putative motifs, whose presence could then be check for in the astrocyte RNA sequencing data. Exosomes were harvested after an astrocyte growing protocol was established, and RNA extracted from these exosomes, but due to an issue with RNA quality no sequencing was performed. Instead a thorough bioinformatic analysis across multiple cell types was performed.
1.2 Exosomes

1.2.1 Overview

Exosomes are small extracellular vesicles with a diameter of 40-100 nm that are present in all eukaryotes (Qin and Xu 2014). They are released from several cell types and have been reported to have a variety of functions including roles in immune responses and neurodegenerative disorders (Zhang et al. 2015). Exosomes are formed as a part of endocytosis, when many small vesicles bud inwardly into the endosome to create several small vesicles, these small vesicles then fuse with the cell membrane to become exosomes (Zhang et al. 2015). While originally thought to be a simple mechanism of cellular waste disposal (Qin and Xu 2014), the discovery that the contents of exosomes can be taken up by cells revealed that these are also a method of intercellular communication (Zhang et al. 2015). This uptake by other cells is also implicated in metastasis and tumour development, as non-cancerous cells can be made to express genes associated with the spread of cancer via the delivery of those genes mRNAs or proteins (Azmi et al. 2013).

1.2.2 Biogenesis and contents

Exosomes, like other extracellular vesicles (EVs) are originally formed in a cell by the inward invagination of an internalised endosome, becoming a multivesicular body (MVB) which then fuses with the cell membrane and releases the vesicles (Kowal et al. 2014). Multiple molecules are involved in the release of these vesicles, including endosomal sorting complexes required for transport (ESCRT), lipids and tetraspanins (Kowal et al. 2014). The molecules involved in exosome release also exhibit cell type specificity (Hessvik and Llorente, 2018).
Figure 1.1: The formation and release of exosomes. This diagram shows the mechanism of exosome release, from the inward budding of an endosome which is then trafficked by ESCRT (endosomal sorting complexes required for transport) to the cytosol where the formation of multivesicular bodies (MVB) occurs. Trafficking to the cell membrane is done by RAB proteins, and the release of the MVB and its exosomes is mediated by SNARES (Soluble NSF attachment protein receptors). Also shown are the associated protein chaperons that help transport the vesicles to a target location. Hessvik and Llorente, 2018.

Generally exosome isolations are performed using ultracentrifugation or bead filtration from cultured cell media. As these techniques separate vesicles purely on size alone and exosomes are generally isolated from media containing other organisms serum, vesicles from the serum in the media can be co-extracted and contaminate results (Tosar et al. 2017). This issue can be overcome using serum depleted media so when examining exosome literature it is crucial to check if serum depleted media was used. These vesicles typically contain a range of molecules, including proteins, RNAs (mRNA and miRNA) and lipids, including cholesterol and fatty-acyl chains (Yu et al. 2014). These contents are not random cellular debris but rather have been shown to represent specific parts of the cell, mostly the endosome and the cytosol (Thery et al. 2001) and this composition varies by cell type, however similarities exist (Rashed et al. 2017). The lipid composition of exosomes is markedly different to a regular whole cell membrane as the exosome lipids are enriched with cholesterol, sphingomyelin, ceramide and phosphatidylserine (Kowal et al. 2014). Proteins have always been shown to be present in exosomes (Rashed et al. 2017) and as previously stated are not random but demonstrate a bias in origin specificity (Rashed et al. 2017). Transmembrane proteins are, for example, loaded into the exosome during the MVB budding step of exosome generation (Thery et al. 2002), however also present are targeting and fusion proteins and chaperons such as heat shock proteins (Rashed et al. 2017). The microRNAs (miRNAs) present in exosomes are quite often studied in the
contexts of diseases and cancers (Lopez and Granados-Lopez 2017) and are secreted from both primary tumours and tumours arising from metastasis (Dhondt et al. 2016). These RNAs are also shown to have functions in diseases such as ischemic heart disease (Chen et al. 2017), renal pathology (Pomatto et al. 2017) and also in postnatal development via mechanisms such as breast milk (Golan-Gerstl et al. 2017). Exosomal messenger RNA (mRNA) is less studied than either miRNA or proteins however again are known to have roles in cancer (Guo et al. 2017). Exosomal mRNA is also examined in diseases such as Rheumatic disease (Cosenza et al. 2017) and are useful biomarkers for everything from cancer prognosis to transplant rejection prediction (Guo et al. 2017, Zhang et al. 2017). While exosomes and their contents have several critical roles in diseases, due to the lack of comprehensive exosome characterisation the exact role of exosomes remains to be discovered.

Fusion of MVBs with the cell membrane to facilitate exosome release from the origin cell appears to not be ubiquitous between cell types as several different mechanisms and protein complexes have been identified (Kowal et al. 2014) however some mechanisms appear to be conserved. Such identified complexes are generally in either the RAB family or the Soluble NSF attachment protein receptor (SNARE) family which are associated with endosome trafficking and recycling (Kowal et al. 2014). In terms of exosome secretion, RAB27a appears to be the controlling factor as when inhibited it inhibits exosome secretion as well as other soluble factors (Bobrie et al. 2012). This mechanism was further confirmed by several other studies in other cell lines (Hoshino et al. 2013, Peinado et al. 2012).

Once exosome secretion has occurred, as they are a method of cell—cell communication they require targeting signals to correctly identify the recipient cell. This is accomplished by transmembrane proteins called tetraspanins located on the surface of the exosome membrane (Perez-Hernandez et al. 2013). Exosome delivery is regulated by these transmembrane proteins as tetraspanins also have receptors on the cell surface membrane that they are able to bind to (Jiang et al. 2015). When these tetraspanins are removed via genetic knockout, the concentration of exosomes in neighbouring cells are depleted (Perez-Hernandez et al. 2013).

After exosomes have bound to their receptors on the receiving cell, internalisation via endocytosis of the exosome occurs (Tian et al. 2013, Escudero et al. 2014). After internalisation occurs, the exosomes then usually diffuse into the cytoplasm around the internalisation site, or can undergo rapid transport via the actin cytoskeleton (Tian et al. 2013). For those vesicles undergoing actin transport, the end destination is generally a lysosome (Tian et al. 2013). Once the exosome has been delivered to the lysosome, the time spent in the lysosome impacts the efficacy of the signalling molecules within the exosome, as degradation of these molecules can occur (Escudero et al. 2014). The lysosome then breaks apart, releasing its contents into the recipient cell, including molecules that are able to effect a change in cellular function. This allows cells to regulate, at a distance, the properties of receptor cells (Baixauli et al. 2014).
1.2.3 Exosome functions

Exosomes have been linked to many functions and roles throughout the body, from regulating immune responses to lipid transport (Record et al. 2014). Fully defining the role of exosomes in a cell is currently not possible, as the role differs from cell to cell. The role of exosomes is also constantly in flux as they respond to environmental changes, as differences in exosome content have been observed as reactions to stimuli such as immune reactions (Greening et al. 2015) or oxidative stress (Saeed-Zidane et al. 2017). These changes effect vast responses in the recipient cell, including potentially remodeling the epigenome (Steinbichler et al. 2017). Accurately defining exosome functions will likely require years of focused study on specific cell types, with the view that one answer will not fit all questions.

1.2.4 Molecular signals for exosome loading

As exosomal composition is not reflective of the origin cell (Villarroya-Beltri et al. 2014) it is likely mechanisms of exosome loading are in place to target specific molecules. The mechanisms of exosomal content loading control are still being researched. It is theorised that protein complexes and conserved nucleotide sequences that these bind to complexes are the mechanisms by which the cell controls which molecules are present in exosomes (Thoms et al. 2015, Batagov et al. 2011). Currently identified protein loading mechanisms are adapter proteins that recruit Mtr4, an exosome associated helicase, to RNAs with specific nucleotide motifs (Thoms et al. 2015). Nop53 then targets Mtr4 to pre–ribosomal particles for exosome pre–processing and Utp18 separately recruits Mtr4 to cleaved rRNA fragments destined for loading into the exosomes (Thoms et al. 2015). Nop53 and Utp18 both have a conserved amino acid motif, which in both proteins bind the ‘arch’ domain of Mtr4, to target it to substrates (Thoms et al. 2015). The conservation of these amino acid motifs suggests a generalised protein mechanism for loading of RNAs into exosomes (Thoms et al. 2015). While this demonstrates one such pathway, a great variety of loading pathways exist (Villarroya-Beltri et al. 2014).

Another identified mechanism used for molecule loading involves ESCRT machinery, however as ESCRT is mainly associated with lysosome degradation control the role of ESCRT in extracellular vesicles is more controversial (Villarroya-Beltri et al. 2014). Syndecan heparan sulphate proteoglycans and their cytoplasmic adaptor syntenin control the formation of exosomes by acting with ALIX, which interacts directly with ESCRT proteins to traffic required proteins to the exosomes (Baietti et al. 2012). The ESCRT pathway sorts proteins to be loaded into exosomes based on ubiquitination (Villarroya et al. 2013). These ubiquitinated proteins are then sorted to endosome membranes through an Hrs domain (Villarroya-Beltri et al. 2013). The exact signals derived from the ubiquitination of these proteins is still undefined, as ubiquitin is capable of forming many different modifications due to the complexity of this post translational modification (Swatek and Komander et al. 2016).
1.2.5 Zip code identification in exosomal mRNA

How the cell regulates exosomal content loading has in recent years been the subject of a lot of research. In 2011, three 8 base motifs were identified that were enriched in the mRNAs of secreted exosomes (Batagov et al. 2011). These three motifs were reported to be associated with stem loop structures that RNA can form (Harada, 2015), setting the theme for ongoing research in this particular field. Further work on stem loops was continued and a 25 base ‘consensus’ sequence was identified in exosome 3’ UTRs in 2012 (Bolukbasi et al. 2012) and again was reported to form stem loop structures (the core sequence is highlighted in red in Figure 1.2 and shown in Figure 1.3).

Figure 1.2: The 5 base core sequence identified by Bolukbasi et al. 2012. Sequences are are 3’ UTRs identified by a preliminary alignment to have some consensus. Shown top are the base positions for each sequence. On the left are the microarray probe names for each sequence. Highlighted in red is the core sequence with a mismatch allowance of one base and the red square demonstrates the core consensus sequence. Figure 2 from Bolukbasi et al. 2012.
Figure 1.3: The stemloop secondary structure predicted by Bolukbasi et al. 2012. In red is the core sequence identified to enrich mRNA loading into exosomes. In green is the site that was mutagenised to prevent stemloop formation and confirm that the stemloop was required for exosome localisation. Figure S1 from Bolukbasi et al. 2012.

However even though the authors showed that this sequence enriched green fluorescence protein (GFP) expression in exosomes, as they used a global alignment tool (CLUSTAL), which returns sequences based on genomic position and does not account for conserved sequences that are in different positions, this sequence is in doubt. In addition to this, in Figure 1.2, not every sequence in their Figure has the core ‘consensus’ sequence, and some that are highlighted only have 4/5 bases matching the consensus. It should not be ruled out however that the motif could be a low fidelity motif, like some well characterised RNA binding proteins (RBPs) (Blower 2013). In addition, using global alignment tools to identify RNA motifs needs to be complemented with a downstream clustering analysis (Sabarinathan et al. 2018). To combat this, a program such as MEME (Multiple Em for Motif Elicitation) can be used for a more accurate motif search. MEME allows for a more accurate motif search because instead of simply aligning a list of given sequences and then returning a consensus sequence, it instead takes that list of sequences and uses a ‘position-dependent letter-probability matrix’ (Bailey et al. 2009) to describe the common bases at the position in the motif. MEME finds motifs by “fitting a mixture model to the data (a set of unaligned sequences), which then returns a model of each motif and threshold which together can be used as a Bayes-optimal classifier for searching for occurrences of the motif in other databases” (Bailey and Elkan, 1994). This is more reliable than a simple consensus sequence as it takes into account variability in the motif position and the probability of several common recurrences of two or more differing bases in the motif. Further work must be done to characterise these potentially conserved exosomal motifs, as they could open up avenues in many disciplines of science, including gene therapy and medical science as utilising them could increase cell to cell signaling which in many cases is impacted by disease.
1.3 Astrocytes

1.3.1 Overview

Astrocytes are typically star shaped cells with extended processes that allow them to interact with neurons to maintain cellular homeostasis, regulate blood flow and assist in neuronal metabolism (Schweinhuber et al. 2015). While astrocytes are the most abundant cell in the nervous system (Malik et al. 2014), there are many variations in cell number and structure that are exhibited by astrocytes in the central nervous system (CNS) that are not well understood (Khakh and Sofroniew 2015). Astrocytes typically come in two forms: fibrous and protoplasmic, associated with white matter and grey matter of the brain respectively (Oberheim et al. 2012). These two forms show strikingly different physiology, with fibrous astrocytes being less complex and tending to grow in bundles and protoplasmic astrocytes containing more extended processes and distributed in a more uniform fashion (Oberheim et al. 2012). Full characterisation of astrocyte function and distribution is inhibited by a lack of reliable molecular markers that are specifically expressed by astrocytes (Molofsky and Deneen 2015). Thus far glial fibrillary acidic protein (GFAP) is the most commonly used marker however estimates suggest this only captures 15% of astrocyte density and is also expressed in other neural cells (Khakh and Sofroniew 2015). An alternative, pan-astrocyte marker is S100\(\beta\), however this marker also stains other glial cells. This lack of reliable, cell specific markers is a hindrance in functional studies where characterisation of astrocyte changes in disease is crucial.

1.3.2 Regular function

Astrocytes are the predominant cell type in the brain and as such perform several roles including creating and maintaining the blood brain barrier, genesis of synapses, regulating brain metabolism, and neurotransmission (Molofsky and Deneen 2015), however the full scope of these duties remain undefined (Oberheim et al. 2012). Once astrocytogenesis has occurred, astrocytes migrate to their final destination for colonisation, where they remain stationary and begin to produce an extracellular matrix for the support of the surrounding neurons (Mofolsky and Deneen 2015, Biran et al. 1999). During early postnatal development, astrocyte functions are mostly centered around assembling presynaptic and postsynaptic synapses (Clarke and Barres 2013). This is accomplished using secreted molecules such as thrombospondins and glypicans (Clarke and Barres 2013). During this phase, astrocytes are also extending their processes to allow interaction with multiple synapses, which for a single astrocyte can be as many as 100,000 (Halassa et al. 2007). Once synapse maturation has occurred, astrocytes exhibit more of a ‘caretaker’ role where synaptic function is regulated via cytokine and growth factor release (Guillamon-Vivancos et al. 2014). Maintaining homeostasis of the synaptic fluid surrounding the astrocyte and its associated neurons is another way that mature astrocytes regulate synapse function (Simard and Nedergaard 2004). Astrocytes regulate the pH, ions and...
Figure 1.4: Astrocyte roles and activities. This shows astrocytes performing multiple functions, such as regulating blood flow to neurons via direct interaction with blood vessels and synapses, having its own action potentials with $Ca^{2+}$ ions similar to neurons, and regulation of synapse homeostasis by helping regulate neurotransmitters. Wolf and Kirchoff 2008.

neurotransmitter levels of the interstitial fluid and in particular control the osmolarity of neurons and astrocytes via the vast amounts of water channel proteins (aquaporins) present in astrocytes (Simard and Nedergaard 2004, Obara et al. 2008). Maintenance functions of astrocytes also include regulating the flow of blood to neurons via the neurovascular junction (Wolf and Kirchoff 2008). Astrocytes also use this neurovascular junction to assist in metabolic regulation via glucose delivery to neurons and by holding large stores of glucose in granules near dense synaptic regions that allow for reserves of energy for neurons to call upon, as seen in Figure 1.4 (Phelps 1972). Cells in the brain do not take cholesterol from the blood stream but rather utilise cholesterol synthesised by glial cells, mainly by astrocytes (Guillamon-Vivancos et al. 2014). These examples highlight the diverse and important roles astrocytes play in the CNS.

1.3.3 Astrocytes and exosomes

Many of the cells in the CNS, including astrocytes, secrete exosomes (Guescini et al. 2010). Recently, these secreted vesicles have been studied in the context of cancers (Ipas et al. 2015, Zhang et al. 2015) due to their ability to secrete carcinogenic mRNAs that can be expressed by neighbouring cells (Azmi et al. 2013). In contrast, in the brain they are most commonly examined for their potential as therapeutic targets for CNS diseases such as Alzheimers disease or Neural ceroid lipofuscinoses (NCLs) (Kawikova and Askenase 2015). While no one is yet to fully identify the complement of proteins, mRNAs and other molecules that are secreted in astrocyte exosomes, this method of intercellular communication should be characterised in the context of both healthy and disease states of the cell. This knowledge would allow the identification of potential gene therapy targets if molecular mechanisms were identified.
1.3.4 Micro-RNAs

In relevance to exosomes, miRNAs can form stem loops as secondary structure (discussed later) and can bind to mRNAs. This was described as having a role in exosome localisation in 2012 by Bolukbasi et al., where the upregulation of miR-1284 lead to an increase in exosome localisation. Micro-RNAs (miRNAs) have several roles in developmental biology, and were discovered over 30 years ago (Horvitz and Sulston 1980). These RNAs do not encode for genes but instead exhibit regulatory roles as single stranded oligonucleotides, mostly by base pairing with mRNAs to cause silencing of that gene (Fire et al. 1998). Most miRNAs undergo the same process to convert the primary miRNA transcript into the mature miRNA. Either during or directly after transcription, the enzyme Drosha cleaves the stem loop precursor away from the total transcript sequence (Lee et al. 2003). The stem loop precursor is then transported into the cytosol, where Dicer (an endonuclease) cleaves the loop region of precursor, releasing the mature miRNA transcript (Hutvagner et al. 2001). The mature transcript is loaded into the RNA induced silencing complex (RISC) which is then directed to the highly specific target mRNA, leading to translational repression and eventually degradation of the target mRNA (Haas et al. 2016). Several exceptions exist to this pathway, for example Dicer independent biogenesis where after transcription and Drosha cleavage the precursor binds to Ago2 instead of Dicer. Ago2 then cleaves the target RNA as normal and the remaining stem loop sequence is trimmed to yield the pre-loaded mature miRNA (Cheloufi et al. 2010). Prediction of miRNA binding sites is mostly accomplished through bioinformatic approaches, as miRNAs exhibit Watson-Crick base pairing methods to predict miRNA targets are fairly simplistic, however the binding sequence (known as the seed sequence) is only 6 bases long (Lewis et al. 2003). This leads to many false negatives when one single method is used. To improve accuracy several factors are incorporated, such as sequence conservation between species, flanking sequences, and limiting searches to certain regions of the gene (as the best defined targets are in 3’ UTRs) (Ekimler and Sahin 2014). Finding and defining any potential consistently present miRNA binding sites present in exosome 3’ UTRs would imply a potential role in exosome loading for mRNAs.

1.4 RNA localisation signals

RNA localisation signals are involved in redirecting various RNAs to specific locations. Signals are not currently well defined, however have some consistency between them. The mRNA must have a cis acting sequence, and an RNA binding protein able to bind to that sequence (Hamilton and Davis, 2011). These cis acting sequences can be either a specific sequence or a stem-loop structure and are normally located in the 3’ UTR (Di Liegro et al, 2017). When the two bind, this complex is then transported to the final destination in the cell via the cytoskeleton. When only one acting sequence is required for localisation it is known as a ‘zipcode’ (Kislauskis and Singer 1992). As this method involves multiple elements all requiring specificity, a variety of
tools are required to identify both each element individually and the sum of the elements. Additionally, localisation can also be affected by several low specificity elements acting in concert (Blower, 2013). This means that multiple elements present in the same mRNA can increase the localisation activity, making it important to check for multiple occurrences of the element per mRNA as well as single occurrences. Finding RNA localisation elements presents a significant challenge as they are unique, multifaceted, and can have low specificity. However accurately defining a localisation signal for mRNA in exosomes would allow several applications to be developed for this unique method of cellular communication.

1.5 Bioinformatic predictions and models

1.5.1 Motif searches

A motif is a ‘dominant or recurring idea’, and in genetics motifs often refer to short sequences that occur several times in a given set of genes or larger sequences. Motifs can be found using a variety of algorithms, however most still suffer from large amount of false positives, especially when considering de novo motif discovery (Tompa et al. 2005). Previous work to increase accuracy of predictions mostly has involved the incorporation of closely related species sequences (Hawkins et al. 2009), however in cases where either those sequences are not available or there are no closely related sequences this approach is not applicable. Other strategies involve using multiple algorithms to find motifs and combining the results (Osada et al. 2004). The two algorithms used in this thesis are expectation maximisation and position frequency matrix estimation.

Expectation maximisation allows the finding of repeated elements present in a single FASTA file, which is useful when searching for motifs in a single file, however is less applicable when searching across multiple files (Bailey et al. 2015). The program MEME uses the expectation maximisation algorithm. First the frequency of bases is counted in each position and placed in a matrix. Each entry of the matrix is then divided by the total number of bases under investigation. To obtain the probability of each base, the total of each base is also divided by the total number of nucleotides. Then each position is divided by the overall probability of the specific base present at that position in the matrix. The value of 0 in $\log_{10}$ is then set to -10 and the matrix is multiplied by $\log_{10}$. For a specific motif, the matrix entries for the required bases can then be added together and divided by the number of entries to give a numerical score of that motif.

Position frequency matrix estimation uses a matrix estimated directly from word counts, meaning it is able to compare different results more accurately as word counts are much more transferable across different files (Sharov and Ko, 2009). This comparison across files allows for identification of possible consistently present localisation signals. The algorithm has several steps, and is demonstrated graphically in Figure 1.5 (Sharov and Ko, 2009).
A. Nucleotide substitution matrix for word ‘ATGCAAAT’

\[
\begin{array}{c|cccc}
\text{Position} & i = 1 (A) & i = 2 (C) & i = 3 (G) & i = 4 (T) \\
\hline
p = 1 & ATGCAAAT & GTGCAAAT & ATGCAAAT & TTGCAAAT \\
p = 2 & AAGCAAAT & AGGCAAAT & AGGCAAAT & ATGCAAAT \\
p = 3 & ATACAAAT & ATGCAAAT & ATGCAAAT & ATGCAAAT \\
p = 4 & ATGAAAAT & ATGAAAAT & ATGAAAAT & ATGAAAAT \\
p = 5 & ATGCAAAT & ATGCAAAT & ATGCAAAT & ATGCAAAT \\
p = 6 & ATGCAAAT & ATGCAAAT & ATGCAAAT & ATGCAAAT \\
p = 7 & ATGCAAAT & ATGCAAAT & ATGCAAAT & ATGCAAAT \\
p = 8 & ATGCAAAT & ATGCAAAT & ATGCAAAT & ATGCAAAT \\
\end{array}
\]

\[W_{pl} = \]

B. Frequency substitution matrices for the test and control sequences

\[
\begin{array}{c|cccc}
\text{Position} & i = 1 & i = 2 & i = 3 & i = 4 \\
\hline
p = 1 & 200 & 46 & 43 & 120 \\
p = 2 & 42 & 52 & 44 & 200 \\
p = 3 & 45 & 40 & 200 & 37 \\
p = 4 & 38 & 200 & 57 & 55 \\
p = 5 & 200 & 48 & 43 & 145 \\
p = 6 & 200 & 52 & 48 & 100 \\
p = 7 & 200 & 59 & 42 & 30 \\
p = 8 & 80 & 47 & 65 & 200 \\
\end{array}
\]

\[T_{pl} = \]  

\[
\begin{array}{c|cccc}
\text{Position} & i = 1 & i = 2 & i = 3 & i = 4 \\
\hline
p = 1 & 150 & 13 & -5 & 77 \\
p = 2 & -15 & -6 & 1 & 150 \\
p = 3 & -1 & 12 & -15 & -16 \\
p = 4 & -4 & 150 & 6 & 9 \\
p = 5 & 150 & -4 & -1 & 93 \\
p = 6 & 150 & 14 & 5 & 53 \\
p = 7 & 150 & 3 & 1 & -1 \\
p = 8 & 32 & 10 & 19 & 150 \\
\end{array}
\]

\[C_{pl} = \]

C. Subtraction of matrices

\[
\begin{array}{c|cccc}
\text{Position} & i = 1 & i = 2 & i = 3 & i = 4 \\
\hline
p = 1 & 150 & 13 & 150 & 0 \\
p = 2 & -15 & 150 & 0 & 0 \\
p = 3 & -150 & 150 & 150 & 0 \\
p = 4 & -4 & 150 & 0 & 0 \\
p = 5 & 0 & 150 & 0 & 0 \\
p = 6 & 150 & 14 & 0 & 0 \\
p = 7 & 150 & 3 & 0 & 0 \\
p = 8 & 32 & 10 & 0 & 0 \\
\end{array}
\]

D. Negative values are replaced by zero

\[
\begin{array}{c|cccc}
\text{Position} & i = 1 & i = 2 & i = 3 & i = 4 \\
\hline
p = 1 & 63 & 5 & 0 & 32 \\
p = 2 & 0 & 0 & 0 & 100 \\
p = 3 & 0 & 0 & 100 & 0 \\
p = 4 & 0 & 0 & 0 & 5 \\
p = 5 & 62 & 0 & 0 & 38 \\
p = 6 & 68 & 6 & 2 & 24 \\
p = 7 & 97 & 2 & 1 & 0 \\
p = 8 & 15 & 5 & 9 & 71 \\
\end{array}
\]

E. Normalized position frequency matrix (PFM)

\[
\begin{array}{c|cccc}
\text{Position} & i = 1 & i = 2 & i = 3 & i = 4 \\
\hline
p = 1 & 1 & 1 & 1 & 1 \\
p = 2 & 1 & 1 & 1 & 1 \\
p = 3 & 1 & 1 & 1 & 1 \\
p = 4 & 1 & 1 & 1 & 1 \\
p = 5 & 1 & 1 & 1 & 1 \\
p = 6 & 1 & 1 & 1 & 1 \\
p = 7 & 1 & 1 & 1 & 1 \\
p = 8 & 1 & 1 & 1 & 1 \\
\end{array}
\]

F. Patterns of 8-mer words with and without gaps

G. Filling the gaps and extending PFMs

H. Clustering and combining PFMs

\[
\begin{array}{c|cccc}
\text{Position} & i = 1 & i = 2 & i = 3 & i = 4 \\
\hline
p = 1 & 1 & 1 & 1 & 1 \\
p = 2 & 1 & 1 & 1 & 1 \\
p = 3 & 1 & 1 & 1 & 1 \\
p = 4 & 1 & 1 & 1 & 1 \\
p = 5 & 1 & 1 & 1 & 1 \\
p = 6 & 1 & 1 & 1 & 1 \\
p = 7 & 1 & 1 & 1 & 1 \\
p = 8 & 1 & 1 & 1 & 1 \\
\end{array}
\]

\[
\begin{array}{c|cccc}
\text{Position} & i = 1 & i = 2 & i = 3 & i = 4 \\
\hline
p = 1 & 1 & 1 & 1 & 1 \\
p = 2 & 1 & 1 & 1 & 1 \\
p = 3 & 1 & 1 & 1 & 1 \\
p = 4 & 1 & 1 & 1 & 1 \\
p = 5 & 1 & 1 & 1 & 1 \\
p = 6 & 1 & 1 & 1 & 1 \\
p = 7 & 1 & 1 & 1 & 1 \\
p = 8 & 1 & 1 & 1 & 1 \\
\end{array}
\]

Figure 1.5: The Cisfinder algorithm demonstrated visually. The algorithm involves 8 steps. First nucleotide substitution is performed, the frequency substitution for both the test and control sequences. The substraction of the two matrices occurs followed by the replacement of negative values by 0. The matrix is then normalised to obtain a motif PFM that can be transferred between sequences. Patterns of words with and without gaps are analysed, gaps are extended and then PFMs are combined and clustered to return the motifs enriched in one sequence compared to another. Sharov and Ko, 2009.
Another level of motif comparison involves breaking motifs down into the possible combinations of different bases or letters present in several motifs, known as kmers (Beckmann et al. 2014). For example, a motif with the bases GGTAAT has kmers GGT, GTA, TAA, AAT. If a specific kmer is enriched across multiple motifs it may imply a small binding site or other element that is present, and 3 base motifs are much harder to detect in a large amount of sequences. Combined these approaches allow an in silico approach to discovering short sequences that have biological meaning in a high throughput manner, and at a volume that experimental determination cannot compete with.

1.5.2 RNA secondary structure

RNA molecules are capable of base pairing within themselves, allowing it to twist and conform to shapes, such as loops and knots, this shape is referred to as a secondary structure. Secondary structures in RNAs can be determined experimentally by a variety of methods, however in general, these methods are time consuming, expensive and difficult (Yonemoto et al. 2015). Finding secondary structures present in biologically relevant RNAs in silico is therefore the preferred method however this approach still has some severe drawbacks, with the benefit of massively increasing the throughput. Most RNA secondary structure finding algorithms involve assessing the minimum free energy of the structure (Zuker and Stiegler, 1981). Minimum free energy predictions are based on the thermodynamic stability of the predicted structure, where the lower the score the more stable the structure is. This method is widely used, and picking the most likely structure is based on assessing each potential structures energy and aggregating the most likely structures (Yonemoto et al. 2015). Minimum free energy is the most stable structure, but can fall down with the inclusion of so called wobble base pairs where non canonical base pairs (Mostly G with U) occur (Ananth et al. 2013). The other issue (as with any in silico prediction) is that the predicted structures don’t correspond necessarily to the biologically present RNA secondary structure. This method used in isolation can predict the most likely structures present in either in a single RNA or an alignment with a high level of sequence conservation (Will and Jabbari, 2016).

In instances where groups of sequences have no sequence conservation but instead a conserved structure, a covariance model (CM) can be generated. A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus, so in many cases, it is more capable of identifying RNA homologs that conserve their secondary structure more than their primary sequence. Covariance models are generated from an initial sequence alignment and can then be used to search for other occurrences of the structure in non aligned sequences, one tool for doing this is CMfinder (Eddy and Durbin, 1994). CMfinder begins by constructing a heuristic initial alignment to identify the approximate position and potential structure of motifs present. Candidate selection for potentially stable secondary structures are then performed using Vienna, an RNA secondary structure tool (Hofacker et al. 1994).
To refine these candidate results, a consensus alignment is generated by comparing their predicted secondary structures again using Vienna. To further improve these alignments, an EM-algorithm (Similar to MEME described earlier) is used to estimate both a secondary structure and motif instance. Probability refinement is then performed. A second phase of EM then occurs, and the CM is then used to scan through each sequence to treat the top sequence hits as candidates, further refining the candidate alignments. More diverse RNA candidates are then discovered by using the EM algorithm with the candidate motif as a seed, then motifs are combined using a hierarchical algorithm to return the final suggested covariance models of conserved RNA secondary structures. This approach has several advantages in that sequences with low conservation can be searched, with low error rates and high sensitivity (Eddy and Durbin, 1994). RNA secondary structures are highly biologically relevant and are involved in localisation signals.

1.5.3 Aims

Overall this project aims to identify molecular loading signals for mRNA into an exosome. We specifically aim to address the question of what RNA is present in an astrocyte exosome and, using a bioinformatic analysis, whether these mRNA contain any signals that are present in other cell types.
Figure 1.6: The overall question that this thesis aims to address, what factors cause an mRNA to be localised to an exosome? As Exosomes bud inward to a multivesicular body, localisation of the contents occurs from within the cytoplasm. The signals for this are as yet not well defined. This thesis specifically focuses on the localisation signals for mRNA.
Chapter 2

Methods

For a complete list of solutions and buffers see Appendix 5.1
All centrifuge steps were performed using a Thermo Scientific Heraeus Multifuge3S (Thermo Fisher scientific, NZ).

2.1 Animal protocols

2.1.1 Euthanasia

All animals used were listed under the ET 25/14 and ET 34/17. C57BLK/6 Mice pups were euthanised at either two or three days old (P2 or P3) using 30 mg mL$^{-1}$ Sodium Pentabarbitone via peritoneal injection and checked for complete anesthesia using a toe pinch reflex test.

2.2 Cell culture

2.2.1 Culture preparation

Two days prior to culturing, flasks were treated with a 1.5 µg L$^{-1}$ poly-lysine dilution and left to incubate in tinfoil overnight at room temperature. The next day the poly-lysine was removed and 10% Fetal Bovine Serum (Moregate Biotech, NZ) (FBS) in DMEM (GE life sciences) was sterilised with a 0.2 µm filter and placed in the wells or flasks overnight in a Thermo scientific Steri-Cycle HEPA class 100 incubator at 37 °C and 5% CO$_2$. Also at this time, 3 fire polished pasteur pipettes of decreasing pore size were moulded and autoclaved.

2.2.2 Culturing mouse brain cells

After pups were euthanised, the head was removed and skin and skull peeled away. The brain was extracted and finely diced with scalpel blades then placed in 1 mL of cold L15 complete. Diced tissue was transferred to a 2 mL microfuge tube and was allowed to settle. Supernatant was removed and 1 mL of digestion media per brain was added. Samples were then placed on
the MACS mix rotating incubator (Miltenyi Biotec, Germany) and allowed to rotate in incubator for 15 min. Samples were then taken off the rotator and tissue allowed to settle. Digestion solution was then replaced with 1 mL Blocking solution. Samples were then rotated for a further 10 min. Blocking solution was replaced with 1.5 mL fresh blocking solution in a 15 mL Falcon tube and trittrated to a cloudy solution with each pasteur pippette prepared earlier. Samples were passed through 100 µm pore cell strainers and washed with the rest of the trituration solution then centrifuged at 100 xg for five min to pellet cells. Cells were then plated at 200,000 cells per well for a 24 well plate or 5,000,000 for a Corning 75 cm² flask.

2.2.3 Cell maintenance

Every 3 days a half media change was performed on cell cultures. Media was warmed for thirty min in a 37.5 °C water bath. Cultures then had all of their media removed and placed into a 15mL Falcon tube. Cultures had half volume fresh warmed media placed on them. Tubes were then centrifuged for five min at 1000 xg to remove cell debris. Cultures then had half volume centrifuged media placed on them and returned to the incubator.

2.2.4 Passaging cells

When cells reached 95% - 100% confluence, media was warmed for thirty min in a 37.5 °C water bath. Cultures had all media removed and discarded. Cultures were washed with dPBS then were placed in 1/3 volume of total media TrypLE™ Express enzyme (Invitrogen, NZ) for 10 min in the incubator. Physical force (tapping flasks against hand and table) was used to remove cells from the flask when necessary and cell separation was observed under a light microscope. When 95% of the cells had been lifted from the flask, half the volume of TrypLE was removed and placed into a new flask, and both new cultures had fresh media placed on them.

2.2.5 Removal of microglia

To remove the microglia present in the astrocyte cultures, the cultures were placed in a — shaker and placed in an incubator for 6 hours. The cultures were then retrieved and all media was discarded. Fresh media was then added to the cultures, 3 mL for a T25 and 9 milliL for a T75.

2.2.6 Mixed neural cultures

As neurons do not actively divide, different methods were undertaken to grow a mixture of primary cells to ensure the previous methods were obtaining an enriched culture of astrocytes. Culturing method and maintenance was kept the same, with a media containing neural growth
factors (full media in the appendix) being used. Cells were plated into a 24 well plate and not shaken or passaged. All downstream methods were the same.

2.2.7 Fixing cells

To perform immunocytochemistry, cells were plated and fixed into 24-well plates. Cells were first split as above and counted and plated at a concentration of 200,000 cells per well and left overnight to recover. Cells were fixed by removing all growth media and washing the well with 100 μL of dPBS. Cells were then placed in 4% paraformaldehyde (PFA) for 15 min. The PFA was removed and another 100 μL dPBS wash was performed. Wells then had 1 mL fresh dPBS placed in them for storage.

2.2.8 Immunocytochemistry

To preblock the cells, 3% PBS Normal goat serum (NGS) in PBS was incubated in the wells for an hour at 500 μL per well. Primary antibodies were then diluted as per appendix section 2. Antibody solutions were then added at 500 μL per well once the pre-block was removed. Plates were then sealed with parafilm and refrigerated overnight at 4 °C. The next day the primary antibody was removed and a 500 μL wash of PBS-T was added to each well and the plate was placed on the shaker for 10 min. This was repeated three times. Secondary antibodies were diluted as stated in section 2 of the appendix and 500 μL of the solution was then added per well. Plates were wrapped in foil and left on the shaker for 1 hour 30 min. The mix was then removed and another three PBST washes were performed. The final wash was then removed and 300 μL PBS was added to each well. DAPI nuclear staining was then performed by making a 0.5 μg mL⁻¹ solution of DAPI (4,6-diamidino-2-phenylindole) and 500 μL was added to each well without removing the PBS. Plates were then incubated for five min at room temperature. All liquid was then removed from well and 500 μL fresh PBS was added. Images were captured using a Nikon eclipse ti2 inverted microscope for quantification.

2.2.9 Image quantification

To quantify the percent of CD68 positive contaminating microglia present in cultures, images taken were imported into the image processing program ImageJ (Abramoff et al, 2004). Images were either converted into color via the ‘merge channels’ option for representative images, or converted into 8-bit images for further processing. To quantify the number of cells, the ‘analyse particles’ option was used on the image containing the cell type of interest. These counts were saved into an excel spreadsheet and read into R for processing.
2.2.10 Preparing exosome-free media

Fetal Bovine Serum (FBS) contains contaminating bovine exosomes which must be removed prior to use. To prepare exosome free-media for exosome harvest, half the volume of astrocyte media was first prepared, with the total volume of FBS required, and ultracentrifuged at 100,000 xg for 17 hours. Ultracentrifuge tubes were balanced using DMEM. The rest of the required media was then filter sterilised into the ultracentrifuged portion and stored at 4 °C. The two portions were combined the next morning post centrifuge.

2.2.11 Exosome isolation

To harvest exosomes from astrocytes, cell cultures were incubated for 48 h in exosome free media. All growth media was then collected into 50 mL falcon tubes and centrifuged at 450 xg for 5 min. Supernatant was then transferred to a new 50 mL Falcon tube and centrifuged at 2000 xg for 20 min to remove cells and large debris. Supernatant was then transferred to a 32 mL ultracentrifuge tube and balanced with dPBS. Tubes were ultracentrifuged for 25,000x g for 30 min at 4°C and then supernatant was transferred to a new ultracentrifuge tube and ultracentrifuged for 100,000 xg for 70 min at 4°C. The supernatant was then discarded and pellets were washed in 1 mL dPBS and then filled with dPBS. Tubes were then ultracentrifuged at 100,000 xg for 70 min at 4°C. Supernatant was then discarded and pellet was either resuspended in 30 µL dPBS for storage at -80 °C or was immediately used for downstream RNA extraction.

2.2.12 Electron Microscopy

Suspended preparations of exosomes were imaged via the Otago Centre of Electron Microscopy (OCEM). A negative stain was used and images were taken on a Philips CM100 transmission electron microscope. All preparation and imaging was performed by Richard Easingwood (OCEM).

2.2.13 Cell RNA extraction

Unless stated otherwise, all microcentrifuge steps were performed using a Thermo scientific Heraeus Pico 17 centrifuge.

After media was collected for vesicle isolation, fresh media was placed on cultures until a 70 min ultracentrifuge step was being performed. At this point all media again was removed from cultures to proceed with an Invitrogen RNAeasy RNA isolation kit. First, 0.6 mL lysis buffer was added and the flask was scraped with a pipette to aid cell lysis. Cells were then vortexed for 5 s and then passed 5-10 times through a 21-gauge needle attached to a 1 mL syringe. One volume of 70% ethanol was added to each homogenate and was vortexed thoroughly to remove any precipitate that may have formed. Up to 700 µL of the sample was then transferred to
the spin cartridge and microcentrifuged for 12000x g for 15 s at room temperature. This was repeated until the whole sample had been processed. Then, 700 µL of wash buffer I was added to the spin cartridge and microcentrifuged at 12,000 xg for 15 s at room temperature. The flow through and collection tube was discarded and the spin cartridge was placed in a new collection tube. 500 µL of wash buffer II was added to the spin cartridge and microcentrifuged at 12,000 xg for 15 s at room temperature and the flow through was discarded. This was repeated once. To dry the membrane, the spin cartridge was microcentrifuged at 12000 xg for 1 min, the collection tube and flow through was then discarded and the spin cartridge was placed in a recovery tube. To elute the RNA, 30 µL RNase-free water to the center of the spin cartridge and samples were incubated at room temperature for one min. The cartridge was then microcentrifuged for 2 min at 12000 xg at room temperature. To check for the presence of RNA, 1 µL was pipetted onto a NanoDrop (NanoDrop technologies, USA) and measured under the RNA setting. If samples registered positive for RNA at a quantity over 300 ng µL⁻¹, they were split into 5 µL aliquots, and stored at -80 °C.

2.2.14 Exosome RNA extraction

Unless stated otherwise, all microcentrifuge steps were performed using a Thermo scientific Heraeus Pico 17 centrifuge.

Once extracellular vesicle isolation was complete, RNA isolation was performed using the Novus Biologicals exosome RNA kit. First, 700 µL lysis buffer was added to the sample and was incubated for five min at room temperature. 140 µL of chloroform was added and was shaken manually for 30 s. Samples were then incubated for 10 min at room temperature and for one min on ice. Samples were then microcentrifuged at 12000 xg at 4 °C for 10 min using an Eppendorf centrifuge 5417R. The top phase of the resulting precipitant was then transferred to an RNase free tube and 2 volumes of 96% ethanol was added. Half of the volume was added to a spin column and microcentrifuged at 14000 xg for 30 s and was repeated for the remaining half of the sample volume. 400 µL of RNA wash buffer was added and microcentrifuged at 14000 xg for 30 s. This step was repeated twice more. Samples were then microcentrifuged at 14000 xg for 5 min and 15 µL elution buffer was added. Samples were then incubated for 5 min at room temperature. Samples were then microcentrifuged for 2 min at 200 xg and at 1 min for 14000 xg. To check for the presence of RNA, 1 µL was pipetted onto a NanoDrop (NanoDrop technologies, USA) and measured under the RNA setting. If samples registered positive for RNA at a quantity over 5 ng µL⁻¹, they were split into 3 µL aliquots, and stored at -80 °C.

2.2.15 Qubit quality control

To accurately measure the quantity of RNA present in isolated samples, the Invitrogen RNA High Sensitivity Qubit kit was used. To calibrate the qubit instrument, two standards containing
190 µL buffer and 10 µL of RNA standard were made up. Once the instrument was calibrated, the sample mixtures containing 199 µL buffer and 1 µL of sample RNA were prepared, vortexed, and then incubated for 2 mins. These mixtures were then measured under the high sensitivity RNA option.
2.3 Bioinformatics overview

All R scripts available in online supplementary materials
All command line arguments available in Appendix section named after program

Figure 2.1 is a diagram demonstrating the overall workflow of the bioinformatic methods undertaken. These methods were used to identify differences between exosome and cell 3’ UTRs, as previous literature suggested the localisation elements were located in these.

Figure 2.1: The bioinformatic work flow performed, diamonds denote decisions, blue blocks denote types of analyses and red blocks denote searches performed
2.3.1 Dataset generation

All code for dataset generation can be found in the online appendices (link in appendix 5.11) in the script Dataset generation.R under R scripts

Microarray data was downloaded from NCBI's microarray data repository GEO (Gene Expression Omnibus) (https://www.ncbi.nlm.nih.gov/geo/) in the series matrix (series matrices available in online supplementary materials) format, which has already been pre-normalised to analyse differences between exosome and cell 3’ UTRs (Wilhite and Barrett, 2012). Chosen experiments included both cell and exosome data in the same matrix and exosomes in which RNA had been isolated from cells/exosomes cultured in serum free media to prevent contamination of serum derived exosomes. In addition, only data from untreated cells and their exosomes were used. After filtering, 8 out of 52 experiments were found to be suitable for processing. These experiments can be found in the Table 2.1. To reduce computational time, it was decided to only use genes significantly upregulated instead of all the genes present in exosomes, as it was theorised that if an mRNA is significantly enriched in an exosome in comparison to it’s parent cell, it had a higher probability of containing a strong localisation element. Matrices edited to only contain selected samples were processed using limma (Ritchie et al. 2015) in R using Rstudio (Rstudio team, 2015). A Limma toptable, containing logfold changes and adjusted p values was used to identify the significantly enriched mRNAs present in the exosomes. To reduce computational time, the toptable was restricted to the top 1000 mRNAs but this had no effect on results, as the highest amount of significantly upregulated mRNAs in a single dataset was 846. A Bonferroni-Holm correction (Holm, 1979) was applied for the adjusted P values and the threshold for significance was set at 0.05. MRNAs significantly upregulated in exosomes compared to cells were separated into a list for a test dataset (Exosome genes) and mRNAs significantly upregulated in cells compared to exosomes were separated into a list for a control dataset (Cell diff express genes). A FASTA file containing these mRNAs 3’ UTRs was generated using the biomaRt R package, with specific platform and probe IDs specified (Durinck et al. 2009). As a secondary control, a random subset of mRNAs from the total pool of cell expressed mRNAs was generated (Cell random subset), and the 3 UTRs were also appended into a FASTA file using biomaRt and this secondary control file matched the sample size of the exosome dataset. Figure 2.2 is a simple visualisation of where the mRNAs were selected for the test and both control datasets and Figure 2.2 is a flowchart of the overall workflow to generate the datasets. All 3 UTRs were retrieved using Ensembl gene names and only the lowest transcript number (E.g. For Transcript1, Transcript 2, Transcript 3 only Transcript1 was used) was used to account for different transcript variants being present in the datasets, as the first labelled transcript was found first and is best characterised. Exosome datasets were checked for overlapping significantly upregulated mRNAs, to see if any mRNAs were repeatedly being localised to exosomes, using the Upset package in R (Conway et al. 2017). Datasets were treated as separate, and not aggregated into one file for a meta analysis.
Figure 2.2: How each of the cell types datasets were generated. Samples containing untreated cells were extracted manually. The significantly upregulated mRNAs were generated using limma, and the randomly chosen cell mRNAs were taken using the sample function in R. All datasets were kept separate, for all 8 cell types, for a total of 24 datasets.
Figure 2.3: A diagram showing how the genes for each dataset were selected for each cell type. The black vertical lines represent the p value cut off for differential expression, differently oriented lines represent different datasets and where in the pools they were drawn from.

Table 2.1: GEO experiments downloaded

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Cell Type</th>
<th>GEO</th>
<th>Samples used</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEMa-LP</td>
<td>Human epithelium</td>
<td>GSE35388</td>
<td>4-8</td>
<td>Xiao et al. 2012</td>
</tr>
<tr>
<td>HMEC-1</td>
<td>Human endothelium</td>
<td>NA</td>
<td>All</td>
<td>Van Balkom et al. 2015</td>
</tr>
<tr>
<td>H5V</td>
<td>Mouse epithelial</td>
<td>GSE72351</td>
<td>All</td>
<td>Willms et al. 2016</td>
</tr>
<tr>
<td>Tcells</td>
<td>Primary human</td>
<td>GSE50971</td>
<td>2,3,7,8</td>
<td>Villaroya-Beltri et al. 2013</td>
</tr>
<tr>
<td>HMC-1</td>
<td>Human mast cells</td>
<td>GSE25320</td>
<td>1-8</td>
<td>Ekstrom et al. 2012</td>
</tr>
<tr>
<td>MC-9</td>
<td>Mouse liver</td>
<td>GSE24886</td>
<td>1-4,9-12</td>
<td>Eldh et al. 2010</td>
</tr>
<tr>
<td>Macrophages</td>
<td>Primary mouse</td>
<td>GSE54419</td>
<td>4-6,10-12</td>
<td>Singh et al. 2015</td>
</tr>
<tr>
<td>Glioblastoma</td>
<td>Human brain</td>
<td>GSE35444</td>
<td>1,2,5,7,9,10</td>
<td>Bolukbasi et al 2012</td>
</tr>
</tbody>
</table>
2.3.2 MEME

To determine if exosomes from different cell types contain similar motifs, the tool MEME (Multiple EM for Motif elicitation) (Bailey and Elkan, 1995) was used. MEME returns a list of motifs and their E value thresholds based on a single set of unaligned sequences. MEME 4.1 was installed locally and run across all generated FASTA files. The command specified a motif search between 3 and 10 bp and to allow any number of motif sites per 3’ UTR.

Inspection of potential conserved kmers was performed by the ngram package (Schmidt, 2016). Raw motifs (IUPAC uncertainty base notation) were split and 3mers, 4mers and 5mers were generated for both cell and exosome motifs, and the proportions of each kmer was plotted.

To determine if a positional bias was present in the exosome datasets, predicted motif start positions were divided by 3’ UTR length to represent each position as a percentage of the UTR. The means of these motifs were then compared using a non parametric paired Wilcoxon test and were then plotted using ggplot2 in R (Wickham, 2009).

2.3.3 TOMTOM

To analyse what RNA binding proteins the identified exosomes motifs may be binding to, TOMTOM (Gupta et al, 2007), a MEME suite tool was used. TOMTOM compares motifs and returns a P value and E value to indicate the likelihood of the motifs being the same, and motifs were compared to the RNA binding protein database at http://cisbp-rna.ccbr.utoronto.ca. All Human and mouse RNA binding protein motifs were downloaded and TOMTOM was run with a P value threshold of 0.05. These results were the processed in R, and results were examined for overlaps between cell types.

2.3.4 Cisfinder

Cisfinder is similar to MEME in that it identifies motifs. However this tool has the added benefit of being able to compare motifs that are enhanced in one file compared to a control. Cisfinder was run on the webtool available at https://lgsun.irp.nia.nih.gov/CisFinder/. Cisfinder was initially used to identify exosome motifs that were enriched compared to the control files for each cell type (using the arguments as shown in the appendix (Figure 5.1)).

Cisfinder can also search for occurrences of a motif expressed in a position frequency matrix (PFM). This was used to identify occurrences of the motif reported by Bolukbasi et al. in 2012 in datasets to determine if the reported sequence was present at a statistically significant level, using a False Discovery Rate correction (Benjamini and Hochberg, 1995) (FDR <0.05) in other cell types. The exact PFM used (as based on the core motif and core motif base frequency reported with a base on either side as well) can be found in the appendix (Figure 5.2). A motif search was performed to search for the PFM in each of the datasets using the “Search motifs” option. Results were downloaded and processed in R. Exosomes and cells were compared
with a Fishers exact test, and an FDR correction for multiple testing was applied, statistical significance was thresholded at FDR < 0.05.

2.3.5 Glam2

GLAM2 (Gapped Local Alignment of Motifs) (Frith et al. 2008) is another part of the MEME suite that specifically deals with gapped motifs. To investigate a set of repeated elements from Cisfinder, GLAM2 was run locally on the default settings. A putative conserved gapped motif was found, to interrogate this GLAM2 scan was used to search for occurrences of the motif within the FASTA files. All settings were kept default apart from number of alignments reported, which was changed to 200 to ensure every possible occurrence was captured. These results were then read into R and normalised to total size of the FASTA file in bp. To test for enrichment in exosomes, a chi-squared test was used and the frequency of the motif in each dataset was plotted.

2.3.6 Targetscan

Targetscan (Angarwal et al. 2015) is a miRNA binding site prediction database. To determine if there were any common or consistently conserved miRNA binding sites present in RNA enriched in exosomes compared to their respective cells, the Targetscan databases for *Homo sapiens* and *Mus musculus* were downloaded and processed in R (http://www.targetscan.org/cgi-bin/targetscan/data_download.vert72.cgi and http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=mmu_72, both “Predicted Conserved Targets” databases). Binding site frequencies for each miRNA were divided by the size of the dataset and compared between exosomes and cells using a chi-squared test. The statistically significant (P < 0.05, Bonferroni-Holm correction) binding sites were then compared between cell types to check for any overlap of miRNA binding sites between cell types.

2.3.7 CMfinder and Infernal

To test the possibility of conserved 3 dimensional RNA secondary structures present in the 3’ UTRs of these genes, CMfinder (Yao et al. 2006) and Infernal (Nawrocki and Eddy 2013) were used. Sequences with over 90% similarity were removed using CD-HIT-EST (Li and Godzik 2006) because covariance models are built on alignments, and overly homologous sequences would bias the structure prediction. CMfinder creates alignments as part of its algorithm and secondary structures were predicted. Once this was done, Infernal was used to generate Covariance Models (CMs). CMbuild was used to build Covariance models off the predicted models generated by CMfinder, followed by CMcalibrate to fit tailing E values. All exosome and cell FASTA files were combined together and CMsearch was used to search
for occurrences for each model in this FASTA file. The tabular output option was used and processed in R. Any model with more occurrences in exosomes and an E value of under 0.05 was included. To test for statistical enrichment a Fishers exact test was used. These tables were then retrieved and processed in R using a Fishers exact test to determine which models were statistically enriched in exosomes (P < 0.05, FDR correction).
Chapter 3

Results

3.1 Cell culture results

3.1.1 Characterisation of primary astrocyte cultures

To establish protocols for gaining an enriched culture of astrocytes from primary neural cultures, immunocytochemistry was performed to characterise the proportion of astrocytes to contaminating microglia in the cell cultures. Glial fibrillary acid protein (GFAP) and CD68 were used to stain astrocytes and microglia respectively. Astrocyte cultures with less than 5% microglia contamination are considered pure (Saura, 2007), so to confirm if the method of shaking cultures and keeping a microglial inhibitor in astrocyte media was effective, a timecourse immunohistochemistry was performed to track the presence of microglia and growth of astrocytes in cultures (Figure 3.1). Cultures were established and as they grew were split into 24 well plates for immunocytochemistry. Cultures were shaken at 12 days for 6 hours at 115 rpm, and the presence of microglia was observed to be under 5% total cells after shaking. To try and reduce the presence of microglia even further samples were shaken again at 21 days for another 6 h however the a large amount of GFAP downregulation was observed. Consistently high downregulation of GFAP was also observed post 3 weeks and as such no cultures were used post 3 weeks of age.
Figure 3.1: The full timecourse immunocytochemistry experiment (representative images shown), with GFAP, CD68 and DAPI staining shown. In brackets is the magnification each image is taken at. For images labelled with (a), the scale bar represents 500 µm and for images labelled with (b) the scale bar represents 100 µm. GFAP downregulation is observed in both treatment conditions as time passes. After observing remnants of microglia in the shaken cultures, cultures were shaken again for 21 hours (compared to 6 hours previously), with this negatively impacting GFAP expression in these cells.

Neurons are not present in primary astrocyte cultures

Microtubule associated protein 2 (MAP2), a neuronal cell marker, was used to determine if neuronal contamination was present in the primary astrocytes cultures. MAP2 was not detected in primary astrocyte cultures, and was detected in mixed neuronal cultures (figure 3.2).
Figure 3.2: Representative images of mixed neural cell cultures (a) and primary astrocyte cultures (b). Blue represents DAPI, green represents GFAP and magenta shows the staining of neurons, absent in astrocyte cultures.

**Qualitative analysis reveals lower levels of CD68**

Across multiple cultures lower levels of CD68 staining was observed, indicating the shaking treatment was removing microglia (Green, arrows pointing to representative staining, Figure 3.3).

Figure 3.3: Representative images used for qualitative analysis. Shown in green is CD68 staining, a universal microglia marker, blue is DAPI nuclei staining. (a) non shaken culture demonstrating higher levels of CD68 staining and (b) is a shaken culture showing lower levels of CD68 staining.

This was then quantified, and the percentage of contaminating microglia was established as being under 5%.
3.1.2 Characterisation of exosome isolates from cell culture media

After a method of enriching a primary culture of neural cells for astrocytes was successfully established, the ultracentrifugation methods were checked for capability of isolating exosomes. These methods were checked using transmission electron microscopy (TEM) (Figure 3.4). As described in previous literature a negative stain TEM revealed vesicles present. In this sample there appears to be low amounts of vesicles, and different vesicles due to the size disparity. This suggests several other vesicle subtypes, as well as exosomes, are being extracted via this method.

![TEM images of exosomes](image1.png)

Figure 3.4: Transmission electron microscopy images (TEM) of a negative stained vesicle sample. Arrows point to the vesicles. Different sizes are also apparent, suggesting different subtypes of extracellular vesicles.

3.1.3 RNA extraction from isolated vesicles

For 9 months, cells were enriched for astrocytes and exosomes were isolated from serum free media. RNA was then extracted from these exosomes and nano-dropped for quantity. A pilot study of 3 samples was generated however a freezer failure thawed and destroyed these. When 5 samples with RNA > 5 ng µL$^{-1}$ had been isolated, a qubit analysis was performed to meet the quality control requirements of RNA seq. Unfortunately, only one sample had a detectable level of RNA present in the qubit analysis. When this was sent to Otago Genomics for further
quality control, their bioanalyser detected only fragmented RNA (Figure 3.5). When this was discovered, taking into consideration the large amount of time it had taken to generate the samples and the work also done in the bioinformatics section, it was decided to not continue with the RNA seq portion of the project.

![Figure 3.5: Rin scores of the exosome (left) and cell (right) RNA submitted to Otago Genomics. Left shows nothing but fragmented RNA with a Rin score of 0 and right shows a regular bioanalyser score with a RIN of 9.5](image)

### 3.2 Bioinformatic analysis of exosome mRNA

#### 3.2.1 Dataset overview

*The raw series matrices used to generate datasets can be found in the online appendix under Series matrices*

To investigate possible localisation signals present in the 3’ UTRs of mRNAs loaded into exosomes (as previous literature states that’s where the localisation signals are located), 8 datasets of various cell types were generated from microarray data available on GEO. Below are summaries of the datasets generated, table 3.1 shows the cell type the datasets were derived from, table 3.2 shows the number of 3’ UTRs in each dataset and table 3.3 shows the size in bases. The random datasets were created with the same number of genes between exo and cell, and the cell differentially expressed dataset contains all the present differentially expressed genes from the master dataset.
Table 3.1: The names, species and cell origins of the datasets generated

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Species</th>
<th>Cell origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glio</td>
<td>Human</td>
<td>Glioblastoma</td>
</tr>
<tr>
<td>H5V</td>
<td>Mouse</td>
<td>Heart cells</td>
</tr>
<tr>
<td>HEMA_LP</td>
<td>Human</td>
<td>Epidermal melanocytes</td>
</tr>
<tr>
<td>HMC</td>
<td>Human</td>
<td>Mast cells</td>
</tr>
<tr>
<td>HMEC</td>
<td>Human</td>
<td>Dermal endothelium</td>
</tr>
<tr>
<td>Macro</td>
<td>Mouse</td>
<td>Primary macrophages</td>
</tr>
<tr>
<td>MC9</td>
<td>Mouse</td>
<td>Liver</td>
</tr>
<tr>
<td>Tcells</td>
<td>Human</td>
<td>Primary Jurkat T-cells</td>
</tr>
</tbody>
</table>

Table 3.2: The number of genes present in each dataset

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Exo, and Cell random</th>
<th>Cell differentially expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glio</td>
<td>202</td>
<td>846</td>
</tr>
<tr>
<td>H5V</td>
<td>61</td>
<td>172</td>
</tr>
<tr>
<td>HEMA_LP</td>
<td>47</td>
<td>758</td>
</tr>
<tr>
<td>HMC</td>
<td>24</td>
<td>758</td>
</tr>
<tr>
<td>HMEC</td>
<td>115</td>
<td>9</td>
</tr>
<tr>
<td>Macro</td>
<td>54</td>
<td>30</td>
</tr>
<tr>
<td>MC9</td>
<td>16</td>
<td>831</td>
</tr>
<tr>
<td>Tcells</td>
<td>291</td>
<td>310</td>
</tr>
</tbody>
</table>

Table 3.3: The size in bases of each dataset generated

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Exo</th>
<th>Cell random subset</th>
<th>Cell differentially expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glio</td>
<td>134236</td>
<td>305115</td>
<td>1247156</td>
</tr>
<tr>
<td>H5V</td>
<td>71424</td>
<td>76502</td>
<td>212437</td>
</tr>
<tr>
<td>HEMA_LP</td>
<td>98230</td>
<td>305115</td>
<td>1409998</td>
</tr>
<tr>
<td>HMC</td>
<td>37252</td>
<td>35585</td>
<td>1385964</td>
</tr>
<tr>
<td>HMEC</td>
<td>327253</td>
<td>291666</td>
<td>9285</td>
</tr>
<tr>
<td>Macro</td>
<td>26269</td>
<td>70004</td>
<td>44707</td>
</tr>
<tr>
<td>MC9</td>
<td>9337</td>
<td>24048</td>
<td>926482</td>
</tr>
<tr>
<td>Tcells</td>
<td>289535</td>
<td>497146</td>
<td>805485</td>
</tr>
</tbody>
</table>
3.2.2 Very little intersection of differentially expressed genes in exosomes from different cell types is observed

Once datasets were generated, a comparison was performed to determine if there were any genes that were present in all exosomes regardless of cell type in the datasets. A comparative analysis (Figure 3.6) revealed very little intersection between genes differentially expressed in exosomes in different cell types, with the largest intersect being 15 genes between the Glio and Tcells datasets (shown in blue).

Figure 3.6: An UpsetR graph showing the overlap of differentially expressed genes in exosomes between different cell types, the number of genes in each set with no intersection (single dots), and the intersections between different sets (dots with lines drawn between them). The largest intersect is between Glio and Tcells, with 15 genes. Sets with no overlap are not shown. Numbers shown are indicative of the number in each set, i.e. $274 + 15 + 1 + 1 = \text{The total number of Tcells differentially expressed genes}$.

3.2.3 MEME analysis identifies enriched motifs

Motifs identified

>All MEME motif results can be found in the online appendices under MEME/10bp.width
MEME Identified motifs present in 3 UTRs significantly upregulated in exosomes compared to cells that were not present in 3 UTRs significantly upregulated in cells compared to exosomes. Each cell type displayed differences in motifs however. Shown below are a subset of motifs that were present in exosomes not in cells (Figure 3.7).

**Figure 3.7:** A motif result from MEME from each exosome dataset from each cell type, that was not present in the cell control files. Each letter is a graphical representation of the probability of that base occurring at that position in the motif. These motifs are also all different from each other, as there were no motifs that were similar between cell types. A) Glio B) HMEC C) H5V D) Macro E) HEMA LP F) MC9 G) HMC H) Tcells

**Motif positions are not distributed differently in exosomes**

All motif position files and bar graphs can be found in the online appendices under MEME/10bp width positions

Position information of each motif was also extracted for each motif. To determine if the position of the motif instead of the sequence itself was different compared to cells, the starting position of each motif was divided by the total length of the 3' UTR, and these were graphed to compare between cells and exosomes. Figure 3.8 shows the graphs generated for the cell type HMC, a graph was generated for each cell type, with the control file used being Cell random subset. All cell types indicated no difference between cells and exosomes for motif positions.
Figure 3.8: Bar graphs showing the frequency of starting positions for each motif expressed as a proportion of the UTR the motif is present in for the cell type HMC. The graphs do not demonstrate any noticeable bias towards any position in a UTR. Datasets used: Cell random subset and Exo

To condense this information, the top five most significant motifs for each dataset were extracted, the mean starting position of each motif was then determined, the means were then compared using a non-parametric paired Wilcoxon test, and no means were significantly different at \( p < 0.05 \). These five means were then plotted (Figure 3.9). This analysis revealed no significant difference in the means of the starting positions of the motifs, suggesting that there was not a positional element involved in the exosome localisation process.
Mean starting positions of the most significant motifs

![Boxplot](image)

**Figure 3.9:** A boxplot demonstrating the mean motif positions for each cell type and their respective exosomes, with no significant difference (Wilcoxon paired test, \( p < 0.05 \)).

**Kmer combinations suggest individual motifs are unique**

To determine if conserved subsequences were present within the identified motifs, Kmer combinations were inspected. This analysis revealed little, showing only a string of adenines as being enriched in exosome motif files across all cell types (Figure 3.10).
Figure 3.10: Kmer combinations of motifs in exosomes versus cells, any kmer with a higher frequency in exosome datasets will be above the line (slope of 1), the kmer most enriched in each case is labelled. Kmers created using the ngram package in R.
3.2.4 A motif comparison using TOMTOM identifies consistently predicted RBPs within species

All TOMTOM databases and results can be found in the online appendices under TOMTOM

Investigation of possible RNA binding protein (RBP) binding site motif similarities was done via a motif comparison tool TOMTOM. The analysis compared proteins predicted to bind identified MEME motifs across different cell types (Figure 3.11).
Figure 3.11: Intersections of RBP binding sites motifs predicted in different cell types split by species, demonstrating consistently predicted RBPs.

After a set of intersecting proteins was identified, a TOMTOM search was run on both control datasets for each cell type, and proteins present in these searches were removed, and the overlaps were examined again (Figure 3.12). For the Cell_random file, no proteins were predicted to bind across cell types from the same species, whereas for the Cell_differentially_expressed control file showed a similar overlap pattern to the overlap without any controls (Figure 3.11) with 3
out of 4 proteins being the same for humans and 1 protein for mice.

Figure 3.12: Protein intersection after eliminating protein binding sites that were present in the control datasets, Demonstrating different proteins remaining in each dataset. Cell diff_exprs maintains some intersection between the datasets of the same species.

As multiple proteins were seen to intersect between specific species, the functions of the proteins with defined binding sites (as some were inferred) in the database were investigated (summarised in table 3.4).

Table 3.4: Consistently found potential binding proteins discovered

<table>
<thead>
<tr>
<th>Species</th>
<th>Protein</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>NOVA2</td>
<td>Alternative splicing factor</td>
</tr>
<tr>
<td>Human</td>
<td>FXR2</td>
<td>Associates with polyribosomes in 60s large subunits</td>
</tr>
<tr>
<td>Human</td>
<td>ZFP36</td>
<td>Binds to 3' UTRs and promotes degradation</td>
</tr>
<tr>
<td>Human</td>
<td>LIN28A</td>
<td>Localises mRNA to ribosomes, P-bodies and stress granules</td>
</tr>
<tr>
<td>Mouse</td>
<td>Ybx2</td>
<td>Cytosolic storage and localisation</td>
</tr>
</tbody>
</table>

3.2.5 Cisfinder used for further motif identification

Cisfinder was used to both identify motifs enriched in the exosome datasets compared to the cells and to analyse previously reported localisation signals.

Motifs identified

All Cisfinder motifs identified can be found in the online appendices under Cisfinder.
Using the ‘identify motifs’ option in Cisfinder, motifs enriched in one FASTA file compared to another can be identified. These motifs can be found in the appendix. An example of a Cisfinder output from the dataset Glio, with Cell Random being used as a comparison is shown in Figure 3.13). Shown in Figure 3.13 are short 3 letter subsequences that appear to repeat throughout the motifs, such as CCT and CTG.

![Figure 3.13: An example Cisfinder enriched motifs output for the cell type Glio, showing left to right: Motif name, Motif logo, Motif pattern, frequency, frequency ratio compared to control FASTA file, information score, bitscore and p value with FDR correction.](image)

### 3.2.6 Gapped motif identification with GLAM2

All GLAM2 original searches can be found in the online appendices under GLAM2/GLAM_searches, GLAM2scan results are under Glam2scan_putative_motif.

After observing possible short repeating elements present in previous motif results, a motif with gaps present was theorised to be present. GLAM2 (Gapped Local Alignment of Motifs) was used to identify any present gapped motifs. Full GLAM2 files can be found in the online appendix.
Figure 3.14: Results of the GLAM2 gapped motif search, with the repeating elements of the motif outlined in red. Notable repeating elements include a core C, flanked by CTG and CCT. Cell types labelled with 'none found' did not have any motifs present with the gapped motif present.

Figure 3.14 Shows a putative gapped motif that was observed across 6 out of 8 datasets. This motif suggested a core C, with flanks consisting of CCT and CTG, with variable space (0-6 bp) around the core C. GLAM2SCAN, a tool used for identifying frequencies of gapped motifs was used to determine if this motif was in fact enriched statistically in exosomes. The Glio variant of the motif was used to search for enrichment and results were downloaded and processed in R (Figure 3.15). Only the Glio and Tcells datasets showed a statistically significant enrichment for the gapped motif.
Previously reported localisation signals are not present in other cell types

Motif enrichment files can be found in the online appendices under Cisfinder/7 bp

Bolukbasi et al. 2012 reported a 25 base localisation signal that enhanced mRNA loading into exosomes. After observing a multitude of different motifs reported across cell types, it was decided to investigate the prevalence of their reported core motif across cell types. The reported 5 base core sequence and a base on either side was converted in a position frequency matrix (PFM) based on Figure 1.2) in the introduction.

The PFM was then uploaded into Cisfinder and the frequency of the PFM was analysed. Results were downloaded and processed in R, shown in Figure 3.16.
The cell type Glio is the dataset that Bolukbasi et al. built the original localisation element from. As this dataset shows statistical enrichment it was taken as an indication that the method was robust. Only the cell type Macro otherwise indicated a significant statistical enrichment of the localisation element in the exosomes.

### 3.2.7 TargetsScan reveals consistently predicted miRNA targets

*TargetsScan databases can be found in the online appendices under TargetsScan databases*

To identify any miRNAs that consistently bound to 3’ UTRs present in exosomes, TargetsScan, a miRNA database was used. Frequency of miRNA binding sites were examined and miRNAs with significantly enriched binding sites in exosome 3’ UTRs (chi squared test, FDR correction, $P < 0.05$) were plotted to examine overlap. In Figure 3.17, of the five datasets that demonstrated significant enrichment of binding sites, there was little overlap, and no consistently enriched miRNAs with binding sites.
Figure 3.17: An UpsetR graph showing the overlap of statistically enriched miRNA binding sites in exosomes, showing common miRNA binding sites, but none consistently predicted across every cell type.

The functions of these miRNA were then checked (summarised in table 3.5).

Table 3.5: The functions of the overlapping miRNAs identified by Targetscan

<table>
<thead>
<tr>
<th>Cell types overlapping</th>
<th>miRNA</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC9 &amp; Macro</td>
<td>miR-125/670</td>
<td>Cellular differentiation, immune regulation</td>
</tr>
<tr>
<td>MC9 &amp; Macro</td>
<td>miR-184</td>
<td>Differentiation of epidermal cells</td>
</tr>
<tr>
<td>MC9 &amp; Macro</td>
<td>miR-191</td>
<td>Reported in over 20 cancers, regulates cellular processes</td>
</tr>
<tr>
<td>MC9 &amp; HEMA_LP</td>
<td>miR-155</td>
<td>Cell proliferation, regulatory roles in macrophages</td>
</tr>
<tr>
<td>MC9 &amp; HEMA_LP</td>
<td>miR-33</td>
<td>Promotes osteoblast differentiation</td>
</tr>
<tr>
<td>Macro &amp; H5V</td>
<td>let-7-5p</td>
<td>Downregulates Casp3, shown to be in exosomes</td>
</tr>
<tr>
<td>MC9 &amp; H5V</td>
<td>miR-33</td>
<td>Downregulates Casp3, shown to be in exosomes</td>
</tr>
</tbody>
</table>
3.2.8 CMfinder and Infernal identify potential secondary mRNA structures

*All CMfinder results can be found in the online appendices under CMfinder*

To identify potentially conserved secondary structures, covariance models were built and analysed using a fishers test in R, out of 38 models that CMfinder identified, 5 were statistically significantly enriched in exosomes after FDR correction ($P < 0.05$). These models can be seen in Figure 3.18. Of the five models, three were generated from one cell type. There are two similar types of structure, with two of them having 3 stems and 3 of them having long stems. Also not present in the structures is the suggested core sequence required for exosome mRNA localisation (Bolukbasi et al. 2012).
Figure 3.18: Structures identified via CMfinder and infernal, statistically enriched in exosomes ($P < 0.05$, FDR correction). Models are labelled after the dataset they were generated in, followed by the CMfinder numbering system.
Chapter 4

Discussion

4.1 Isolation and content characterisation of astrocyte exosomes

4.1.1 Astrocyte culture characterisation

The immunocytochemistry revealed that GFAP was consistently downregulated after 3 weeks of age. This finding is consistent with previous literature (Schildge et al, 2013). Astrocyte purity was determined by comparing the percentage of CD68-labelled microglia to DAPI positive nuclei (taken as a measure of total cells present). This experiment showed the total percentage of microglia present was low (5%) Likely due to the presence of L-LME, a microglia inhibitor that functions by altering lysosome function and making the cell less likely to survive. Shaking the cultures lowered the microglia presence to 3%, to well within recommended purity parameters (Saura, 2007). Shaking the cultures reduced the amount of microglia present as the microglia are less adherent than astrocytes (Wang et al, 2012). The lack of neuronal cells present, as indicated by comparing mixed neural cultures with enriched astrocyte cultures, is consistent with neuron biology as neurons do not actively divide (Rakic, 1985). Since neurons do not actively divide, once cells were removed from the flask during passaging the neurons were unable to recover, resulting in a neuron free culture. This characterisation provided a strong baseline for the subsequent biological experiments.

4.1.2 Exosome isolation

Based on the electron microscopy results presented in results section 3.1.2, the differences in isolated vesicles is apparent. Based on the size of the vesicles extracted it appears both exosomes (20-100 nm) and microvesicles (50-1000 nm) are being extracted (Samanta et al, 2018). This is likely due to the extraction method isolating vesicles based on size and not applying any molecular targeting such as antibodies to harvest exosomes specifically, which exhibit different proteins on their membrane surface compared to microvesicles (Bobrie et al,
The lack of RNA extracted via these methods may relate to either the amount of starting material or the kit used to extract RNA. Astrocytes also may not produce as many exosomes as other cell types, based on a comparative study performed by Goetzl et al, 2016. In this study, the astrocyte exosome count was $88.5 \times 10^9$ /mL compared to mixed neural cultures which produced $307 \times 10^9$ /mL. This lower amount of exosome release was confirmed by comparison with an exosome isolate from mixed primary neural cultures performed by Alison Clare as demonstrated in Figure 4.1. One possible reason for this may be that astrocytes have other methods of communicating with cells, such as physical contact by extended processes (Halassa et al. 2017). Also apparent in the mixed neural culture vesicle isolates are various vesicle subtypes - similar to the different subtypes present in Figure 3.4 but to a greater extent, whether this is due to a greater proportion of different vesicle subtypes to exosomes or a symptom of a greater sample size is unclear.

![Exosomes isolated from primary astrocytes](image1.png) ![Exosomes isolated from primary mixed neural cultures](image2.png)

Figure 4.1: A comparison of exosomes isolated from different cell cultures, demonstrating a reduced amount of exosomes in the astrocyte cultures. Mixed neural exosomes isolated by Alison Clare.

A lower amount of exosomes produced may limit the amount of starting material available and as seen in results 3.1.2, there is a low amount of exosomes present. This limited amount of starting material will impact the RNA available to extract, regardless of RNA kit extraction efficiency. While the end result was not desirable, the proof of principal of extracellular vesicle isolation was confirmed, though further optimisation to selectively isolate exosomes and high quality RNA is still required.
4.2  *In silico* prediction of mRNA localisation signals for exosomes

4.2.1 Different cell types have distinct exosome gene populations

Exosomes are a unique method of cell-cell communication (Valadi et al, 2007) and exhibit active loading during biogenesis as the contents of an exosome are not reflective of the contents of the originating cell. As different cell types express different genes to reflect their function, the exosomes originating from these cells also must contain specific contents to fulfill their roles in communication. The lack of overlap of differentially expressed genes in results reaffirms this as each cell type is directing different contents in their exosomes, to perform functions extracellularly.

4.2.2 Motifs are enriched in exosomes but are not common across cell types

Much like the coding sequence for the gene, each 3’ UTR is unique. This diversity of 3’ UTRs present in exosomes likely contributes to both the distinct differences in motifs present in the 3’ UTRs of mRNAs differentially expressed in exosomes between cell types and the differences present in the motifs discovered within each cell type. The differences in motif discovery between tools used reflects the algorithms used in each program. MEME uses a finite mixture model with an expectation maximisation and Cisfinder uses a position frequency matrix estimated directly from word counts. However despite the disparity of motifs in different cell types, there is evidence that can suggest the presence of similar proteins that bind these motifs across cell types. The motifs discovered by a combination of methods implies the presence of specific cell type loading mechanisms, and possibly a species specific loading mechanism. If specific RNA binding proteins are involved, it must be noted that binding motifs for RBP’s are much less complex and specific than DNA binding proteins (Helder et al. 2016). This may explain the variation in motifs present in these data. In light of the lack of specificity of RNA binding proteins and how disparate the motifs in exosomes across each cell type are, it may not be ruled out that a protein or family of proteins with a low specificity RNA binding domain may have a large role in localising mRNA to exosomes. However further experimental validation will be required to confirm the presence of the predicted proteins in exosomes, with sequencing also possibly being needed for motif confirmation.

Previous work done in this field by Kojima et al. 2018 has shown that by selecting an exosome specific protein (CD63) and fusing it with an RNA binding protein with a defined binding motif, inserting the binding motif into a 3’ UTR and including a cytosolic delivery protein can very effectively load a mRNA of choice into an exosome. While this paper did not identify the naturally occurring loading mechanisms, it did present a proof of concept that we can direct
exosome loading manually. Discovering the method through which exosomes are loaded in vivo would remove the need for this protein fusion method, however the two could by combined to direct large amounts of mRNA to an exosome.

4.2.3 Previously reported localisation signals are not enriched in all cell types

In 2012, Bolukbasi et al. used a method of sequential alignment of 3’ UTRs to discover a motif that appeared to be involved in mRNA localisation to exosomes. This method was used to allow for potential motifs that were present at a low frequency in a large sample size of 3’ UTRs. The drawback of this approach was that if the motif had conserved bases in different positions the approach would not have identified these. This discovery hinted that all cell types may localise their mRNAs to exosomes via similar mechanisms. When this hypothesis was interrogated, the lack of statistical significance across several cell types for the occurrence of this motif suggests that this motif is not the main localisation signal in other cell types. However considering the Glio and Tcells datasets were both significantly enriched for the motif and had the largest number of overlapping genes, it cannot be ruled out that this motif has some importance in RNA localisation, and may be one of several factors that are involved in this process. Another example of a localisation factor that may not be involved everywhere but still vital for some cells would be the putative gapped motif that was predicted using GLAM2. While being unable to state that this motif is statistically enriched in all cell types, the mere presence of a similar motif may imply that it is not a single stretch of bases but rather a variable gapped motif that has a binding site for a low specificity family of RNA binding proteins.

4.2.4 The puzzle of exosome localisation is still missing some pieces

Micro RNAs (miRNAs) have been acknowledged as biologically important for over 30 years, since their first discovery in 1993 (Lee et al. 1993). However as these miRNAs can bind to multiple mRNAs (Martinez et al. 2002), defining their specific gene targets can be difficult when viewed at a cellular level. To this end, various miRNA binding site prediction tools have been developed. The most precise of these is the TargetScan algorithm, with the lowest false positive rate of recently benchmarked tools (Riffo-Campos et al. 2016). This is why TargetScan was selected for prediction of possible miRNAs that could bind to 3’ UTRs in exosomes. Two mouse cell types exhibited the highest level of overlap of miRNAs, however one of these cell types had over 100 predicted miRNAs. While this could suggest a conserved set of species specific miRNAs that have a role in exosome loading, it seems more likely that the large amount of available miRNAs in that cell type biased the results. The otherwise low amount of statistically enriched miRNAs may stem from the con of Targetscan having the lowest false positive rate, which leads to a low sensitivity or high false negative rate. This was deemed acceptable as over-prediction can be dangerous in an exploratory analysis, leading researchers
down fruitless paths. Experimental validation such as miRNA seq for exosomes needs to be performed before any conclusions can be drawn.

In a similar vein, correctly predicting secondary structures present in RNAs is also a non-trivial matter (Lorenz et al. 2016). Compounding this difficulty was the decision to find consensus secondary structures using covariance models. First described in 1994 by Eddy and Durbin, covariance models are capable of predicting conserved secondary structures with little sequence conservation. The method selected for this search was CMfinder (Yao et al. 2006), as it has a base accuracy of 79% when predicting secondary structures. However this still allows for a 20% error rate and after E value filtering, most models with an E value of under 0.05 were only present in 9 3’ UTRs in exosomes and not present at all in cells. This highlights the need for caution to be exercised when performing bioinformatic searches, as it can be simple to overestimate the importance of any one result. One of the most accurate structure interpreting tools is RNAalifold (Lorenz et al. 2016) and thus this was used to fold the CMs statistically over-represented. The secondary structures predicted in this thesis may provide an indication of a low frequency of structure conservation across cell types, or it may simply be the aberrant results of throwing over 800 sequences at a prediction program and seeing what sticks. Figure 4.2 shows a suggestion for how exosome mRNA signals may function

![Diagram showing exosome mRNA signals](image)

Figure 4.2: The proposed state of localisation signals, where both cell type specific and re-occurring signals are prevalent. Multiple variants of localisation mechanisms are likely to be present.

### 4.3 Future directions

In future, exosome isolation using molecular capture techniques such as magnetic beads or protein binding columns for exosome specific isolates would be preferable, however this approach will need more starting material to compensate for the loss of other, contaminating material. Quantification of amount of exosome release will also need to be performed for each cell type being investigated. These approaches will allow generation of high quality, highly specific exosome data, including RNA and protein databases. From high quality data, the bioinformatic
approaches can also be improved. However in future, every cell type needs to be treated as unique when identifying exosome loading signals, followed by experimental validation to confirm each signal. Further analyses on these data could involve using all genes in the exosomes instead of the significantly upregulated ones, to capture a wider possible range of localisation methods. Taking multiple randomly sampled datasets from cell data as comparisons would allow an understanding of the variance inherently present in taking a random subset.

4.4 Overall conclusion

This thesis aimed to address the question of how mRNA is localised to exosomes through both \textit{in vivo} and \textit{in silico} methods. After establishing methods for enriching a primary neural culture for astrocytes, exosomes were isolated from these cultures. Extraction of RNA from exosomes was attempted however was unsuccessful. In the bioinformatic analysis, the evidence of cell type specific loading signals was gathered, and more work needs to be done before making any conclusions on the mechanisms of loading for exosomes.
Bibliography


Chapter 5
Appendix

5.1 Solutions and Buffers

Astrocyte maintenance media (1 L)
100 mL heat inactivated FBS (final concentration 10%)
10 mL Penicillin/Streptomycin
10 mL L-Glutamine
880 mL DMEM with 4.5 g L\(^{-1}\) Glucose
100 µL 1 M L-LME stock made to 0.1 mM

L15 complete (10 mL), Filter Sterilised
9.3 mL Gibco L15 + phenol red
0.1 mL Penicillin/Streptomycin
0.6 mL Glucose stock

Digestion media
12U mL Worthington Papain
1 µL DNAse I
In 1 mL of L15 complete

Blocking solution
1 µL DNAse I
20 µL B27
in 1 mL L15 complete

Tituration solution
See Blocking solution

Mixed neural media (20 mL)
19 mL Neurobasal A
0.6 mL L-Glutamine
0.1 mL Glutamax
0.2 mL Pen/Strep
0.4 mL B27

5.2 Antibody dilutions

Table 5.1: Antibodies used for immunocytochemistry

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Supplier</th>
<th>Catalogue #</th>
<th>Dilution factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-glia fibrillary acid protein (GFAP)</td>
<td>Sigma-Aldrich</td>
<td>G3893</td>
<td>1:1000</td>
</tr>
<tr>
<td>Anti-cluster of differentiation (CD68)</td>
<td>AbD Serotec</td>
<td>MCA1957GA</td>
<td>1:250</td>
</tr>
<tr>
<td>Anti-microtubule associated protein 2 (MAP2)</td>
<td>Synaptic systems</td>
<td>188 004</td>
<td>1:1000</td>
</tr>
<tr>
<td>Goat anti-mouse</td>
<td>Invitrogen</td>
<td>62-6520</td>
<td>1:1000</td>
</tr>
<tr>
<td>Goat anti-rat</td>
<td>Invitrogen</td>
<td>A-11006</td>
<td>1:1000</td>
</tr>
<tr>
<td>Goat anti-guinea pig</td>
<td>Invitrogen</td>
<td>A-11076</td>
<td>1:1000</td>
</tr>
</tbody>
</table>

5.3 Quantification of CD68 levels

Table 5.2: Cell counts used for microglia quantification (expressed as averages)

<table>
<thead>
<tr>
<th>Label</th>
<th>Not shaken</th>
<th>Shaken</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAPI</td>
<td>1141.5</td>
<td>626</td>
</tr>
<tr>
<td>CD68</td>
<td>80</td>
<td>24</td>
</tr>
</tbody>
</table>

5.4 MEME command line arguments

Running MEME

```bash
for file in *.txt
do
    mkdir $file\_dir
    nohup meme $file -oc $file\_dir -dna -minw 3 -maxw 10 -mod anr
```

63
Extracting just positions from MEME results
Written by Bethany Jose

!/usr/bin/perl
use strict; use warnings;
my $file=$ARGV[0];

open(FILE,'<',$file);

my $motif;
while(<FILE>){
    if($_=~/Motif\s+(\S+)\s+/){
        $motif=$1;
    }
    elsif($_=~/^ENS\s+\d+\s+\S+\s+\S+\s+\S+/){
        my $line = $_;
        chomp($line);
        print $line, "\t$motif\n";
    }
}

close FILE;

5.5 TOMTOM command line arugments

Altering position matrices to MEME format

for file in *.txt
    do
        name="ID 
        file_name=$(basename $file)
        c="$name$file_name"
        echo $c >>All_motifs.txt
        awk '{gsub("Pos","PO",$0);print}' $file
        >>All_motifs.txt
        echo "//" >>All_motifs.txt
    done
Running TOMTOM

for file in ./MEME_texts/Human/*.txt
do
echo $file
tomtom -oc ./$file\_OUT -thresh 0.05 $file
All_motifs_MEME.txt
done

5.6 Cisfinder search parameters

<table>
<thead>
<tr>
<th>File type</th>
<th>File name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence file #1 (test)</td>
<td>Tcells_exo_3un.fa</td>
</tr>
<tr>
<td>Sequence file #2 (control)</td>
<td>Tcells_cell_3un.fa</td>
</tr>
</tbody>
</table>

Parameters for motif identification (modify if needed)

<table>
<thead>
<tr>
<th>No.</th>
<th>Use repeats for search</th>
<th>Forward</th>
<th>Search in strands</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>FDR</td>
<td>No</td>
<td>Adjust for CG/AT ratio and CpG</td>
</tr>
<tr>
<td>2</td>
<td>Count motif once per sequence</td>
<td>Z-score+info</td>
<td>Score motifs by</td>
</tr>
<tr>
<td>0.9</td>
<td>Match enrichment ratio (test vs. control)</td>
<td>1000</td>
<td>Maximum enrichment in repeats (ratio)</td>
</tr>
<tr>
<td></td>
<td>Clustering threshold</td>
<td>560</td>
<td>Maximum number of motifs to find</td>
</tr>
</tbody>
</table>

Figure 5.1: The parameters used in the enriched motifs cisfinder search

5.7 Bolukbasi PFM

>Enhanced_Zipcode_7bp
30 45 25 25
20 50 20 10
10 10 20 55
5 15 75 5
0 75 0 25
0 60 25 15
0 20 40 40

Figure 5.2: The generated PFM from the reported core motif from Bolukbasi et al 2012.
5.8 GLAM2 command line arguments

glam2 n Glio_cell_3utr_new.txt -o Glam2_Glio_Cell_new

Running GLAM2SCAN

glam2scan n Putative_conserved_motif.txt Tcells_exo_3utr.txt
-O Putative_motif_Tcells_Exo -n 200

Taking everything with a score of 15 or over

grep .+[56789] ./Putative_motif_Tcells_Exo/*.txt
>Tcells_Exo_Glam2scan_score_15.txt

5.9 CMfinder and infernal command line arguments

Running CD-HIT-EST

for file in *exo*.txt
    do
        cd-hit-est -i $file -o $file\_CD_hitted
    done

Running CMfinder

for file in *\_hitted
    do
        cmfinder.pl --def $file
    done

Running CMbuild

for file in *\_motif*
    do
        cmbuild $file\_CM $file
    done

Running CMcalibrate

for file in *\_CM
    do cmcalibrate $file
    done

Running CMsearch
for G in *_CM
    do
        cmsearch --tblout $G\_Searched  $G  Total FASTA.txt
>>Cmfindersearched_models.txt
done

5.10  Description of online appendices

Online appendices are available at:
https://drive.google.com/drive/folders/1XM1WrjuzH9lkd8e36kiWm3ggLLF9_s?usp=sharing
Raw data is available in appropriate folders, all scripts under R scripts