



Transcriptional Regulatory Networks in the Mouse Hippocampus

NAME : Cameron Ross MacPherson
STUDENT NUMBER : 2651322
SUPERVISOR : Professor Vladimir B Bajic

A thesis submitted in partial fulfillment of the requirements for the degree of Magister Scientiae in the South African National Bioinformatics Institute, University of the Western Cape.

2007



Stanford - South Africa

Biomedical Informatics Program



Keywords

Brain / Neuro-anatomy

Hippocampus

Ammon's horn

Neurodegenerative disease

Alzheimer's disease

Gene expression

Transcription regulation

Regulatory potential

Transcription factor

Promoter

Transcription factor binding site

Transcription regulatory network

Allen Brain Atlas

Abstract

Neurological diseases are socially disabling and often mortal. To efficiently combat these diseases, a deep understanding of involved cellular processes, gene functions and anatomy is required. However, differential regulation of genes across anatomy is not sufficiently well understood. This study utilized large-scale gene expression data to define the regulatory networks of genes expressing in the hippocampus to which multiple disease pathologies may be associated. Specific aims were: identify key regulatory transcription factors (TFs) responsible for observed gene expression patterns, reconstruct transcription regulatory networks, and prioritize likely TFs responsible for anatomically restricted gene expression. Most of the analysis was restricted to the CA3 sub-region of Ammon's horn within the hippocampus. We identified 155 core genes expressing throughout the CA3 sub-region and predicted corresponding TF binding site (TFBS) distributions. Our analysis shows plausible transcription regulatory networks for twelve clusters of co-expressed genes. We demonstrate the validity of the predictions by re-clustering genes based on TFBS distributions and found that genes tend to be correctly assigned to groups of previously identified co-expressing genes with sensitivity of 67.74% and positive predictive value of 100%. Taken together, this study represents one of the first to merge anatomical architecture, expression profiles and transcription regulatory potential on such a large scale in hippocampal sub-anatomy.

Thesis structure

This report presents results that have originated from a collaborative project between the South African National Bioinformatics Institute (SANBI) and the Allen Institute for Brain Science (AIBS, Seattle, USA) over a course of two years between 2006 and 2007.

The collaborative project has addressed multiple goals:

1. Discern the correlation between classically defined neuro-anatomy and gene expression patterns in the adult hippocampus (AIBS).
2. Determine if gene expression patterns are able to delineate between high-resolution sub-anatomies (AIBS).
3. Identify gene expression patterns that explain physiological differentiation across hippocampal neuro-anatomy (AIBS).
4. Identify TFs that may play a role in maintaining adult hippocampal anatomically restricted gene expression patterns (SANBI).
5. Determine regulatory potential by means of TFs associated with the promoters of all genes expressing in the considered regions of hippocampus (SANBI).
6. Reconstruct hypothetical transcriptional regulatory networks in the mouse hippocampus (SANBI).
7. Prioritize candidate TFs computationally determined to most likely regulate hippocampal gene expression (SANBI).

This report is divided into 6 chapters as follows:

Chapter 1 introduces the role of TFs in controlling regulatory and signal transduction pathways that act to pattern the mouse brain during development and maintain gene expression in the adult mouse. The chapter provides a brief introduction of different methods used to analyze gene expression levels and transcription regulation in the context of the adult mouse brain.

Chapter 2 describes the basis of the hippocampal gene expression that was analyzed for evidence of transcriptional regulation. The majority of this chapter is derived from results discussed in the manuscript Thompson *et al.* (2007). This chapter discusses the biology of the hippocampal formation in detail and correlates the well-known physiology with gene expression data derived from the Allen Brain Atlas.

Chapter 3 introduces the approach used to identify possible transcriptional regulatory elements controlling the expression observed and discussed in Chapter 2. This chapter describes results I have obtained and have been previously presented in the internal SANBI document, Bajic *et al.* (2006A). A brief discussion of the data is made highlighting the impact of the study in terms of identifying candidate TFs responsible for the normal expression of genes in adult mouse hippocampus.

Chapter 4 discusses an analysis of the data produced by the methods described in Chapter 3 and presents visualization of the reconstructed transcription regulatory networks specific to clusters of gene expression data described in Chapter 2. This chapter describes

results I have obtained and have been previously presented in the internal SANBI document, Bajic *et al.* (2006B).

Chapter 5 describes a software tool for maintaining and presenting projects that contain data that may be represented as a network. The tool is a compilation of methods used to generate the regulatory networks in Chapter 4 and it has been coded in HTML and Python based CGI.

Chapter 6 discusses the entire study drawing information from all chapters in order to highlight the main findings of this thesis.



Declaration

I declare that “*Transcriptional Regulatory Networks in the Mouse Hippocampus*” is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Name: Cameron Ross MacPherson
2007



Date: 9th November

Signed: _____

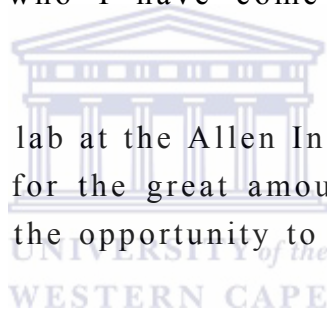
Acknowledgements

Here, I would like to acknowledge that whilst this thesis represents the work of one person it is in fact the result of many. To those that have guided, supported, and freely given their time, it is with the greatest of thanks that these words come:

To my supervisor, Professor Vladimir B Bajic: You have changed the way I think and have been a constant support during my studies. Without you and SANBI I would never have dreamt to experience what I have so soon in my career.

To SANBI, its staff who sacrifice every day to make life just a little easier, and its students who I have come to respect not only as colleagues but friends.

To Doctor Susan Sunkin's lab at the Allen Institute for Brain Science, my thanks go out to you for the great amount of work you have all done and for allowing me the opportunity to work with some amazing data.



To the Stanford South African Biomedical Informatics program who has funded my training and research, my thanks for allowing me to participate in such a prestigious training program.

To those that have been there for me from the 'get-go':

Father, you are steadfast and the roots of our family. Thank you for all that you have done! Mother, you are hope in a bottle and I could never fail in your eyes. Thank you for your courage. Grandmother, grandfather, I wish you could see this. So much of me was made from memories of our past. Vovo and John, you are nothing if not pragmatic, a stolid presence and always willing to guide. In so many ways you are a mirror image to mum, thank you. Sharon, you are forever by my side, thank you for your faith.

"You can know the name of a bird in all the languages of the world, but when you're finished, you'll know absolutely nothing whatever about the bird..."

Richard Feynmann (1988).

Table of Contents

Cover Page	1
Keywords	2
Abstract	3
Thesis structure	
4	
Declaration	7
Acknowledgements	8
Contents	9
List of figures	
12	
List of tables	14
List of abbreviations	14



Chapter 1: Introduction

Transcriptional regulation is a key to understanding the regulation of cellular processes 16

From the developing to the developed brain: The importance of TFs in patterning the adult mouse brain 20

Diseases associated with the hippocampus as a motivation for the study of transcription regulation potential 25

Chapter 2: A discrete molecular architecture underlies the cellular patterning and function of the hippocampus

Computational analysis of hippocampal gene expression 29

Differential gene expression along axes in hippocampal sub-anatomy 31

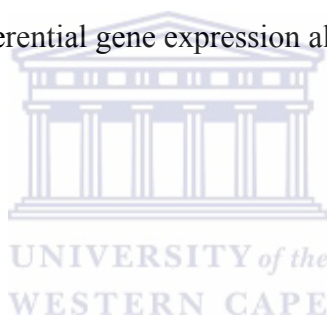
Functions of CA3 genes correlate with gene expression along CA3 axes	38
--	----

Chapter 3: Predicted transcription factor binding sites in promoters of differentially expressed genes of CA3 identify groups of co-expressing genes

Introduction	43
Methods	47
Results	54
Conclusion	64

Chapter 4: Networks of Gene-to-TranscriptionFactor edges elucidate key nodes in the regulatory potential behind differential gene expression along the septo-temporal axis of CA3 anatomy

Introduction	66
Methods	69
Results and Discussion	72



Chapter 5: Graph EZ software tool

Graph EZ as a tool for creating, maintaining, and analyzing network-based projects using an HTML interface	84
Creating a Project	85
The Edit Project Tab	86
The View Project Tab	87
The Analyze Project Tab	88
Discussion	89

Chapter 6: Discussion and concluding remarks

90

References

99

Appendix A: Supplementary Figures

118

Appendix B: Supplementary Tables

119

Appendix C: Supplementary Data Files

120



List of Figures

1. Breakdown of the adult mouse brain highlighting the anatomy and location of Ammon's horn in the hippocampus (adapted from ABA software, www.brain-map.org)
2. Distribution of gene expression across nine CA3 sub-regions in Ammon's horn (adapted from (Thompson et al.))
3. ISH figures of genes expressing in adult CA3 tissues highlighting nine sub-region boundaries (Thompson et al.)
4. 3D Representation of regionalized expression of genes across CA3 (Thompson et al.)
5. Twelve of the most representative gene expression profile clusters (Thompson et al.)
6. Distributions of gene expression across the septo-temporal axis of CA3 for functional groups of genes: (A) Cell Adhesion; (B) Ion-Channel; (C) Transcription Factors (Thompson et al.)
7. Association of gene expression patterns of each of 12 clusters with 9 physical segments of the CA3 hippocampus region in mouse.
8. Explanation of columns in Supplementary Data Files 01 through 12.
9. Distribution of TFBSs across 12 CA3 clusters. ORI value was not less than 3. The background set was mouse promoter set.
10. Distribution of TFBS with ORI not less than 4 and when promoter background is used.
11. Distribution of TFBS that are unique to the clusters.
12. Distribution of significant TFBSs obtained by contrasting cluster target promoter data with the non-promoter (random) genomic background. ORI not less than 2 is used.
13. Heat-map distribution of gene clusters obtained by support tree clustering of annotated promoters using Euclidean distance.

14. (Top left) Transcriptional regulatory network for cluster 8. (Bottom left) Transcriptional regulatory network for cluster 10. (Right) Transcriptional regulatory network for combined clusters 8 and 10. TFBS identified as significant for both clusters are highlighted in yellow, TFBS unique to the individual clusters are depicted in blue.
15. Distributions of TFBS across CA3 gene expression clusters
16. Distribution of TFBS across genes and clusters
17. A) Union of TFBS networks for CA3 (blue) and CA2 (green) regions in adult mouse brain. B) Basic hippocampal anatomy
18. Cluster specific TFBSs filtered by the CA2-CA3 comparison and the percentage of clustered genes they were found in
19. Screen Dump of the create project page
20. Screen Dump of the view tab



List of Tables

1. Summary of known brain diseases associated with the hippocampus
2. Summary of Differential Gene Expression Along Axes in the Hippocampus
3. Summary of gene expression defining CA3 fields
4. Distribution of TSSs relative to clusters
5. List(subset of Supplementary Table 01) of significant TFBSs and their distribution across 12 clusters
6. Link of TFBS and TFs that are known o bind such motifs

List of Abbreviations

3D - Three dimensional

ABA - Allen Brain Atlas

AD - Alzheimer's disease

AIBS - Allen Institute for Brain Science

APP - Amyloid precursor protein

C1,C2,...,C12 - Gene expression cluster 1 through 12

CA - Ammon's horn

CAGE - Cap analysis of gene expression

CGI - Common gateway interface

DG - Dentate gyrus

DLB - Dementia, lewy body

DNA - Deoxyribonucleic acid

EST - Expression sequence tag

FTD - Frontotemporal dementia

GC - Gene cluster



GO - Gene Ontology

GUI - Graphical user interface

HD - Huntington's disease

HIP - Hippocampal formation

HPF - Hippocampal formation

ISH - In situ hybridization

LTD - Long term depression

LTP - Long term potentiation

MEB - Muscle, Eye-Brain Disease

MRC - Medical Research Council

NMF - Non-negative matrix factorisation

ORI - Over representation index

PPV - Positive predictive value

RHP - Retrohippocampal region

RNA - Ribonucleic acid

SAGE - Serial analysis of gene expression

SANBI - South African National Bioinformatics Institute

TC - Tag cluster

TF - Transcription factor

TFBS - Transcription factor binding site

TRH - Thyrotropin releasing hormone

TSS - Transcription start site

TU - Transcriptional unit

VHC - Hierarchical clustering of the voxels



Chapter 1: Introduction

Transcriptional regulation is a key to understanding the regulation of cellular processes

Multicellular organisms develop through a complex process from a single-celled zygote (Cuenca et al., 2003) to their adult forms (Shingleton et al., 2005). Through development, an organism's cells are layered into tissues and further folded into organs with distinctive functionality (Cruzet et al., 2006). This patterning of cells and tissues is reflected in gene expression, since genes express differently across different anatomies, tissues and cell types (Sood et al., 2005). The process of regulating gene transcription must be tightly controlled throughout development in order to ensure the correct functioning and cellular patterning of tissues within an organism (Berger et al., 2007).

Transcription regulation is a directed process integral to the control of gene expression (Reymann and Borlak, 2006). Genes are categorized as being expressed when the transcription process results in gene transcripts. Different forms of transcripts exist, the most exploited kind is messenger RNA (mRNA), which is generated by further processing of primary transcripts (Sheth and Parker, 2003).

Gene expression can be assessed using the concentrations of mRNA transcripts (Wang et al., 2006).

Multiple genes may be associated with a particular function. To determine the relationship that enables functionality between these genes is a hard problem. Gene expression analysis is a powerful approach that may be employed to elucidate on this issue (Wang et al., 2006). Generally, a high-throughput gene expression experiment aims to capture the expression state(s) of a set of genes within cellular or tissue samples from localized anatomy or along a time course. The more samples we analyze the more the expression analysis becomes informative for three reasons: 1/ we analyze the same sample multiple times for confidence in our results; 2/ we analyze the samples over time and gain an understanding of how genes express through time; and 3/ we analyze samples taken at the same time over different tissues and gain an understanding of how genes express spatially. It is through these kinds of studies that the gene expression profile may be described for particular anatomy and associated to lower and higher order biological functions.

Determining the gene expression profile for any particular anatomy certainly answers what genes are being expressed and when. However, it does not answer why. An interesting and important function of one class of gene is to code for proteins that influence the transcription of other genes (Yi et al., 2007). These proteins, known as transcription factors (TFs) and co-factors, provide a powerful means that allows the

cell to preferentially activate transcription of targeted genes in specific cells, under specific conditions and in specific timing. TFs are nuclear proteins that bind to short specific DNA motifs, called TF binding sites (TFBSs), in the regulatory regions of genes (Lin et al., 2007). Not all promoters contain the same TFBSs. This characteristic allows TFs and their complexes (that may include co-factors) to target specific promoters. Complexes of TFs and co-factors function to activate (and at times inhibit) transcription of the genes to whose promoters they are bound (Motohashi et al., 2006).

Here, we present results of an extensive study into the association of transcription regulation potential and gene expression profiles related to the anatomy of the adult mouse brain. The study focuses on the patterning of cells and tissues within the CA3 region of the hippocampus. Using in situ hybridization data, the gene expression profile for the CA3 region under study, as well as its neighboring CA2 region, was obtained using the Allen Brain Atlas, ABA (Lein et al. 2007), from collaborators at the Allen Institute for Brain Science, Seattle, USA. Additionally, TFBSs were predicted and mapped to genes contributing to the gene expression profile. Networks of associations have been made between genes, TFs, and sub-regions of CA3. The result is an in-depth description of the transcription regulation potential of the CA3 region of Ammon's horn in a normal adult mouse brain. The study has revealed transcription regulation programs that are likely to control anatomically restricted gene expression in the studied regions. It has also implicated the combinatorial effect of

several TFs as likely contributors to the specialized gene expression. These results may serve as a model study for the analysis of the regulatory potential of neurodegenerative diseases originating in the adult mouse hippocampus.



From the developing to the developed brain: The importance of TFs in patterning the adult mouse brain

The adult mouse brain is comprised of three parts, the cerebrum, brain stem, and cerebellum. This thesis focuses on a particular part of the cerebrum, Ammon's horn (CA), which is situated within the hippocampal region (HIP). The HIP and retrohippocampal region (RHP) make up the hippocampal formation (HPF, Figure 1). The HPF is part of the limbic system. The limbic system (derived from the latin word *limbus*, meaning edge), is located within the brain and forms within the early stages of the embryo. The limbic system is considered to be involved in emotion and memory related functions. Additionally, it is closely linked with the endocrine and autonomic nervous systems such as the fight-or-flight response. The functions of the HPF have been largely associated with long-term memory formation and more recently anxiety related behaviors (Bannerman et al. 2004). It is for this reason that neurodegenerative diseases originating in the hippocampus, such as Alzheimer's disease, show initial symptoms of memory loss (Farlow et al. 1994; Karlinsky et al. 1992; Rossi et al. 2004).

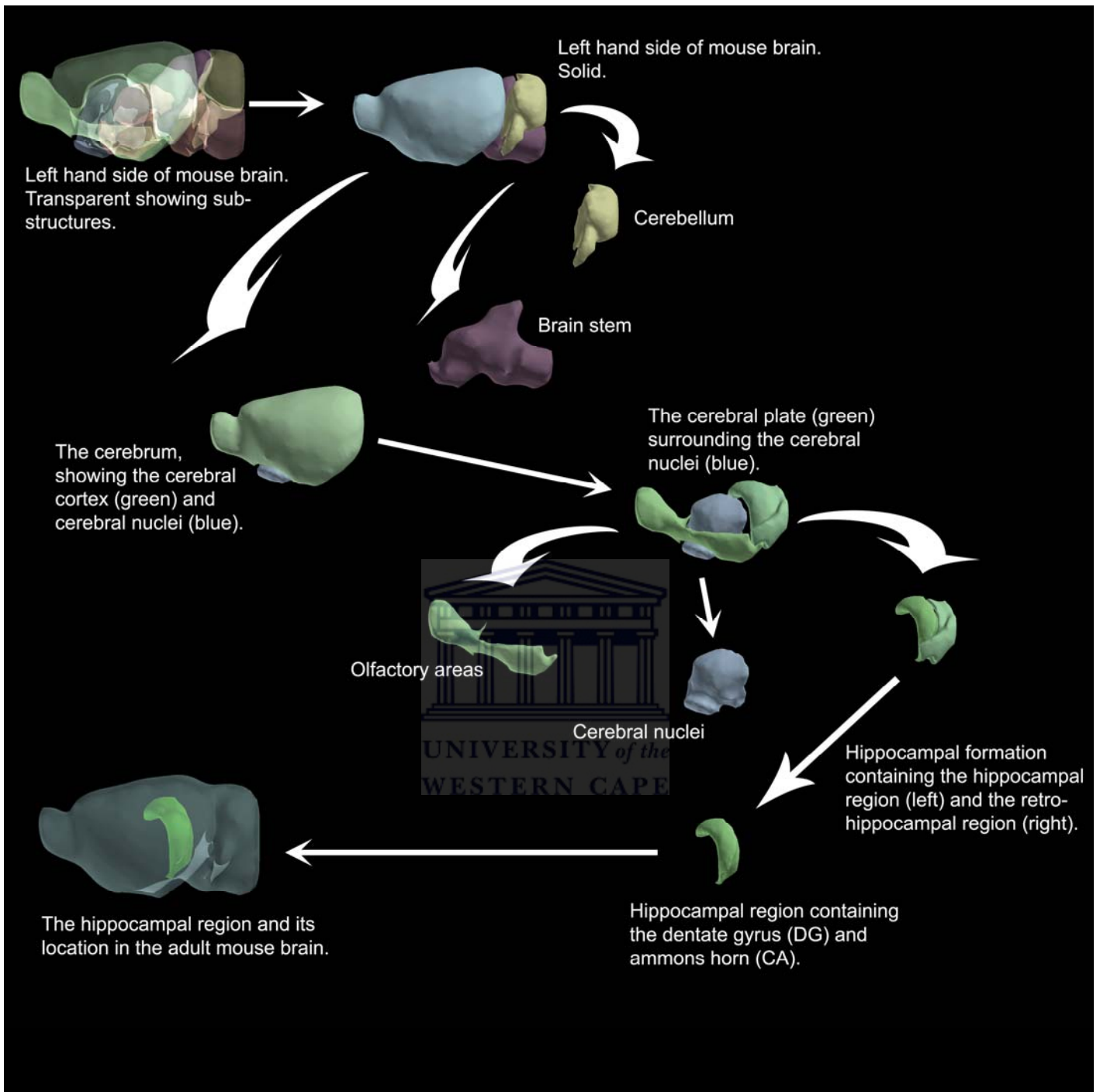


Figure 1: Breakdown of the adult mouse brain highlighting the anatomy and location of Ammon's horn in the hippocampus (part of the figure taken from ABA software, www.brain-map.org)

An in depth understanding of brain anatomy, cellular patterning and anatomic and molecular functions is required to combat disorders such as Alzheimer's disease. The brain is a complex set of cellular and tissue structures. The cells that form these structures during development have to migrate, proliferate, and become polarized. Throughout neural development, organisms use complex signaling pathways to direct cell fate and polarity. For example, the wingless type MMTV integration site (Wnt) signaling pathway is tightly associated with correct development and cellular organization in the mid- and hindbrain (Galceran et al. 2000). The Wnt signaling pathway consists of TFs, such as those from the lymphoid enhancer binding factor 1 (LEF1)/T-cell specific transcription factor (TCF) family, that mediate transcription of Wnt signaling proteins. Wnt signaling proteins are extra-cellular and they target cells both over short and long distances from regions they were translated. Cell signaling enables the Wnt signaling pathway to mediate neuronal morphogenesis, the cellular patterning of brain anatomy, during development (Galceran et al. 2000). Lef1 mutant mice show abnormal development of cell populations in the dentate gyrus of the HIP. Furthermore, Lef1_lacZ fusion genes prevent binding of the LEF1 TF to DNA and additionally inhibit transcription activation by other TFs in the LEF1/TCF family. Lef1_lacZ fusion gene mice display a complete lack of the HIP (Galceran et al. 2000). Similar developmental deficiencies have been reported from mutations in the Smad-interacting protein-1 (Sip1) TF (Miquelajauregui et al. 2007).

TFs play an important role in both the development of the brain as well as regulating gene expression in the adult brain. Yet, the usage of TFs and signaling pathways differ between the developing and adult mouse brain. In developmental brain many genes in pathways, like those within the Wnt signaling pathway, function to promote cellular proliferation (Wodarz and Nusse 1998) and dendrite growth (Keeble et al. 2006). Conversely, these functions are not thought to be significant within the adult brain, and loss of adult neurons is generally irreversible (Eriksson et al. 1998), though there are exceptions. For example, the dentate gyrus contains a pool of neuronal precursors (Kuhn et al. 1996). However, barring these exceptions, gene products that function to inhibit cellular growth and repopulation are up-regulated in the adult brain (Shin et al. 2002; Cayuso et al. 2006). Thus, the gene expression profile within the adult mouse brain is different to that of the developing embryonic brain. Yet, even though different molecular pathway usage is observed, gene expression profiles in the adult brain are still able to specifically delineate between gross anatomy originally defined by developmental gene expression patterns. Recent evidence suggests that this gene expression based delineation has a high resolution within the hippocampus and is able to distinguish between the CA subfields, CA1, CA2, and CA3 (Datson et al. 2004; Lein et al. 2004). This suggests that since gene expression profiles differentiate between adult brain anatomic regions, that they may also be associated with different anatomic cognitive functions. Furthermore, these functions may differentiate along axes within specific anatomy, as is observed in

gene expression studies of the CA region of the hippocampus (Lein et al. 2005). Functional differentiation within the hippocampus has been observed commonly in species, for example rat (Bannerman et al. 2002; Moser and Moser 1998), monkey (Colombo et al. 1998), and human (Small et al. 2001).

The CA region, or Ammon's horn, displays functional differentiation across multiple axes. Many previous studies into hippocampal physiology show preferential targeting of in-coming, afferent (Petrovich et al. 2001), and out-going, efferent (Risold and Swanson 1997; Verwer et al. 1997), neurons of the CA sub-anatomy. Cellular specialization is also observed in the CA anatomy in an axis dependant manner (Jung et al. 1994). Similarly, functional association studies have identified disease models that display pathological symptoms differentially across the CA region (Racine et al. 1977; Bragdon et al. 1986; Ashton et al. 1989). In this context, the study aims to characterize the regulatory potential of TFs responsible for the anatomically defined gene expression data as provided by (Lein et al. 2004).

Diseases associated with the hippocampus as a motivation for the study of transcription regulation potential

Neurological diseases affect multiple cellular types through multifactorial processes, both molecular and environmental (Gatz et al. 2006). To combat these debilitating diseases, an understanding of the complex molecular processes of the brain is required. We investigated large-scale gene expression data at anatomic sub-region level localized within the hippocampus. The hippocampus was chosen because of its particular susceptibility to disease, displaying pathology for many neurological disorders (Table 1). Furthermore, the hippocampus offers a comparatively easy model through which to study neurophysiology because of its structured cell layers and highly ordered synaptic interactions with adjacent brain structures. In what follows, we briefly describe several dementias, incidence rates and genetic etiology with specific focus on Alzheimer's disease (Table 1).

Amongst the many neurodegenerative diseases, Alzheimer's disease (AD) is most commonly associated with the hippocampus and has an incidence rate greater than any other form of progressive dementia. AD, initially manifesting in the hippocampus, is characterizable by extra-cellular β -amyloid plaques and intra-cellular neurofibrillary tangles. This pathology has led to the proposition of many genetic hypotheses as to the etiology of AD, including the mutation of amyloid precursor protein (APP) (Goate et al. 1991), duplication of APP

(Rovelet-Lecrux et al. 2006), and mutation of presenilin-1 and presenilin-2 (Tomita et al. 1997).

Whilst the onset of AD has been reported before the age of 65, this is more rare than the senile variant. Additionally, due to the longevity that first world populations enjoy, AD is at times reflected on as a 'first world disease'. Taken together with the poor ability to diagnose the disease prior-mortem both in first world and especially in developing countries, it is likely that global incidence rates for AD are far below the true statistics. In 2005 a comprehensive study done by Ferri et al. revealed global estimates for AD at 24.3 million with an increase of 4 to 6 million annually. Rates of incidence in developing countries were predicted to increase 100-300% between the years 2001 and 2040 (Ferri et al. 2005). Specifically, a report compiled by the Medical Research Council (MRC) in South Africa indicated that AD was within the top twenty causes of death in Cape Town (Groenewald et al. 2003). Since the entire African continent comprises of developing countries, the above-mentioned estimates pose worrying questions for mental health in Africa. Additionally, treatment of dementia is exceedingly expensive and worldwide costs have been estimated to exceed the 315 billion US dollar mark (Wimo et al. 2007). To date, there is no known cure for Alzheimer's disease.

Several additional dementias exist for which pathologies have been observed and documented in the hippocampus. These dementias, including HD, PD4, FTD, DLB, and MEB (see Table 1 for full names, summary and references) are not tied to the hippocampus as closely as

AD. However, studying their genetic etiology might give clues to dementia in general and help determine why the hippocampus is found to be part of so many neuropathologies. This study describes a model for the normal state of gene expression and its regulatory mechanisms within the context of the adult mouse hippocampus. The model has potential to contribute to research of senile dementia and enable identification of candidate genes and TFs as potential drug targets for these diseases. In particular, the results of this study may prove beneficial in identifying candidate genes for knockout mice to facilitate research on the mouse model.



Table 1: Summary of known brain diseases associated with the hippocampus

Disease	Symbol	Onset	Cause	Inheritable	Chromosome	First Described	OMIM	References
Alzheimer's Disease	SD	Late and early onset. Pre-senile SD occurs at ages below 65.	Unknown.	Yes (evidence for many forms)	20p, 19p13.2, 17q23.1, 17q23, 17q11.2, 12p11.23-q13.12, 12p13.3-p12.3, 11q23.2-q24.2, 10q24, 10q24, 7q36, 7q36, 4p14	Dr. Alois Alzheimer (1907)	104300	Mosconi et al. 2003; Hampel et al. 2005; Li et al. 2003
Huntington's Disease	HD	Juvenile forms exist. Symptoms of dementia usually appear between 30 and 40 years of life.	Multiple trinucleotide CAG repeats.	Autosomal dominant.	4p16.3	George Huntington (1872)	143100	Clarke et al. 2000; Jiang et al. 2003; Reddy et al. 1998
Lewy Body Parkinson's Disease	PD4*	45 years of age.	Triplication of alpha-synuclein gene	Autosomal dominant.	4q21	Waters and Miller (1994)	605543	Muenter et al. 1998
Frontotemporal Dementia	FTD	From 32-58. Penetrance equals 15% at 40 and 80% at 45.	Mutation in the gene, microtubule-associated protein tau.	Autosomal dominant on the female line.	17q21.1, 14q24.3		600274	Neumann et al. 2006; Willhelmsen et al. 2004
Dementia, Lewy Body	DLB	62 years of age.	Mutation of alpha- and beta-synuclein genes.	Autosomal dominant.	5q35, 4q21	Frederick Lewy (1912), characterised the lewy body.	127750	Wakabayashi et al. 1998; Ohtake et al. 2004; Khachaturian 1985
Muscle-Eye-Brain Disease	MEB	Congenital.	Mutations in POMGNT1 and FKRP genes.		19q13.3, 1p34-p33	Raitta et al. 1978	253280	Michele et al. 2002
CREUTZFELDT-JAKOB DISEASE	CJD	On transmission.	Mutation of PRNP gene. Transmissible.	Non-genetic inheritance.	20pter-p12, 6p21.3	Jakob et al. 1950	123400	Meissner et al. 2005

Chapter 2: A discrete molecular architecture underlies the cellular patterning and function of the hippocampus

Computational analysis of hippocampal gene expression

The Allen Brain Atlas (ABA) (www.brain-map.org), is comprised of 21,500 genes and their expression profiles mapped onto a three-dimensional adult mouse brain. The process to map just one gene three-dimensionally required the sectioning of the P56 male C57BL/6J mouse brain into 25 μ m thick slices. The sections were then probed with fluorescent single stranded RNA molecules that emitted colored light when exposed to certain wavelengths. The probes targeted and bound particular mRNA transcripts within cells in each section. This technique, termed in situ hybridization (ISH) described in Lein et al. (2007), was used and each gene active within a brain section was spatially mapped back to a structural reference atlas (Dong 2007) and the concentration of transcripts quantified. This was performed for all 21,500 genes across over 6,000 mouse brains, the completion of ABA required 1 million sections with over 600 terabytes of data required to digitally store ABA (Lein et al. 2007).

The data structures that form the backbone to the ABA make it possible to easily query the atlas for genes that express only in

certain brain regions with some user defined expression threshold. This data structure, termed a voxel, is a matrix of 3D co-ordinates and gene expression values.

Two independent and unbiased computational methods were used to analyze voxels comprised of 2,686 genes that were highly expressed within the hippocampus. The primary concern was whether or not the voxels contained enough pertinent expression data to identify and discriminate between classically defined neuro-anatomy within the hippocampus. Using non-negative matrix factorization (NMF) (Lee and Seung 1999), a method similar to principal component analysis, several groups of voxels that subdivided the hippocampus correctly into the DG, CA3 and CA1 regions were identified (Thompson et al. 2007). Additionally, the method identified an expression domain spanning all three regions of the HIP sub-anatomy on the temporal pole; it is evident that the temporal pole of the HIP displays a far different expression profile relative to the septal portion. However, NMF failed to accurately distinguish between high-resolution subfields (Thompson et al. 2007); an alternative approach of hierarchical clustering of the voxels (VHC) was used instead. The VHC metric was taken as the Pearson's correlation of expression data between the voxels. This VHC approach resulted in two large clusters representative of the hippocampus. Subdivided, one of these clusters represented voxels spanning major excitatory cell layers of CA and the DG with an additional cluster of voxels containing expression data similar to the region found by the NMF analysis spanning the temporal HIP. Further sub-division of

these clusters revealed high-resolution sub-domains within the CA1, CA3 and DG fields. Keeping the voxel data structure intact throughout the analysis allowed for the clustered data to be easily mapped back to the ABA.

Both NMF and VHC methods, unbiased and independent of each another, were able to identify similar domains across major subfields of the hippocampus. Since voxel data is based exclusively on gene expression data, these results indicate that genes express together in ways that are able to identify and discriminate between functionally and classically defined neuroanatomy. Taking into account that expression of single genes is unable to delineate between neuroanatomy, this suggests that such expression data may be used to identify cohorts of active genes whose products might be working together in an anatomically restricted manner in the adult mouse brain. Furthermore, the data contained in the voxels, coupled with the high-resolution clusters identified by the VHC method, helped to identify genes in novel regions within the HIP sub-anatomy.

Differential gene expression along axes in hippocampal sub-anatomy

Bordering the CA3 region in the hippocampus are the CA2, CA1, and DG. The small CA2 region separates CA3 and CA1. In these regions (CA1, CA2 and CA3) as well as in the DG region, gene expression patterns are observed to be distinguishing. These expression patterns

display differential gene expression across the sub-anatomy in an axes dependant manner. Our study focuses on the CA3 sub-anatomy and a summary of the CA1 and DG axis-dependant gene expression patterns can be found in Table 2.

Table 2: Summary of Differential Gene Expression Along Axes in the Hippocampus. Lct (lactase); Cyp7v1 (cyclophilin); Ptpro (protein tyrosine phosphatase); Igfbp6 (insulin-like growth factor binding protein); Trhr (thyrotropin releasing hormone receptor); Cpne7 (copine E); Dio3 (deiodinase, iodothyronine type III)

Anatomy	Symbol	Axis		Genes	
		Pole A	Pole B	Pole A	Pole B
Dentate Gyrus	DG	Dorsal Half	Ventral Half	Lct	Trhr
		Dorsal Two Thirds	Ventral Third	Cyp7v1	Cpne7
Ammon's Horn (CA1)	CA1	Septal-Distal	Temporal	Lct	Dio3
		Distal	Proximal	Ptpro	
		All	Distal Pole	Igfbp6	

As previously described by Lein et al. (2007), the CA3 region is divided into two axes, the septo-temporal axis and the proximal/distal, along which differential gene expression can be observed. The expression profile of CA3 is heterogeneous and many genes (see Figure 2) show anatomically restricted CA3 expression. Notably, the expression of genes within CA3 shows discrete and diffuse anatomical boundaries suggesting a specialized architecture and role within CA3.

Observing the boundaries evidenced in the CA3 gene expression patterns reveals nine discrete sub-fields within the CA3 anatomy. These sub-fields have been labeled 1 through 9 and divide the CA3 septo-temporal axis in groups of three: fields (1,2,3) at septal pole; fields (4,5,6) at mid-septo-temporal third; and fields (7,8,9) at temporal pole. Each field within a group further divides that portion of the CA3 region along the proximal/distal axis. Examples of genes were found that define the borders to each of the fields and can be viewed in Table 2. Furthermore, using double fluorescent ISH, genes shown to discriminate between fields 1 through 9 were further analyzed at the borders of these fields to determine how discrete or diffuse the gene expression is at the borders. Results, displayed in Figure 3, show highly discrete boundaries with no co-labeling of cells. Although, some boundaries have diffuse cell-border layers with differently labeled cells co-mingled.

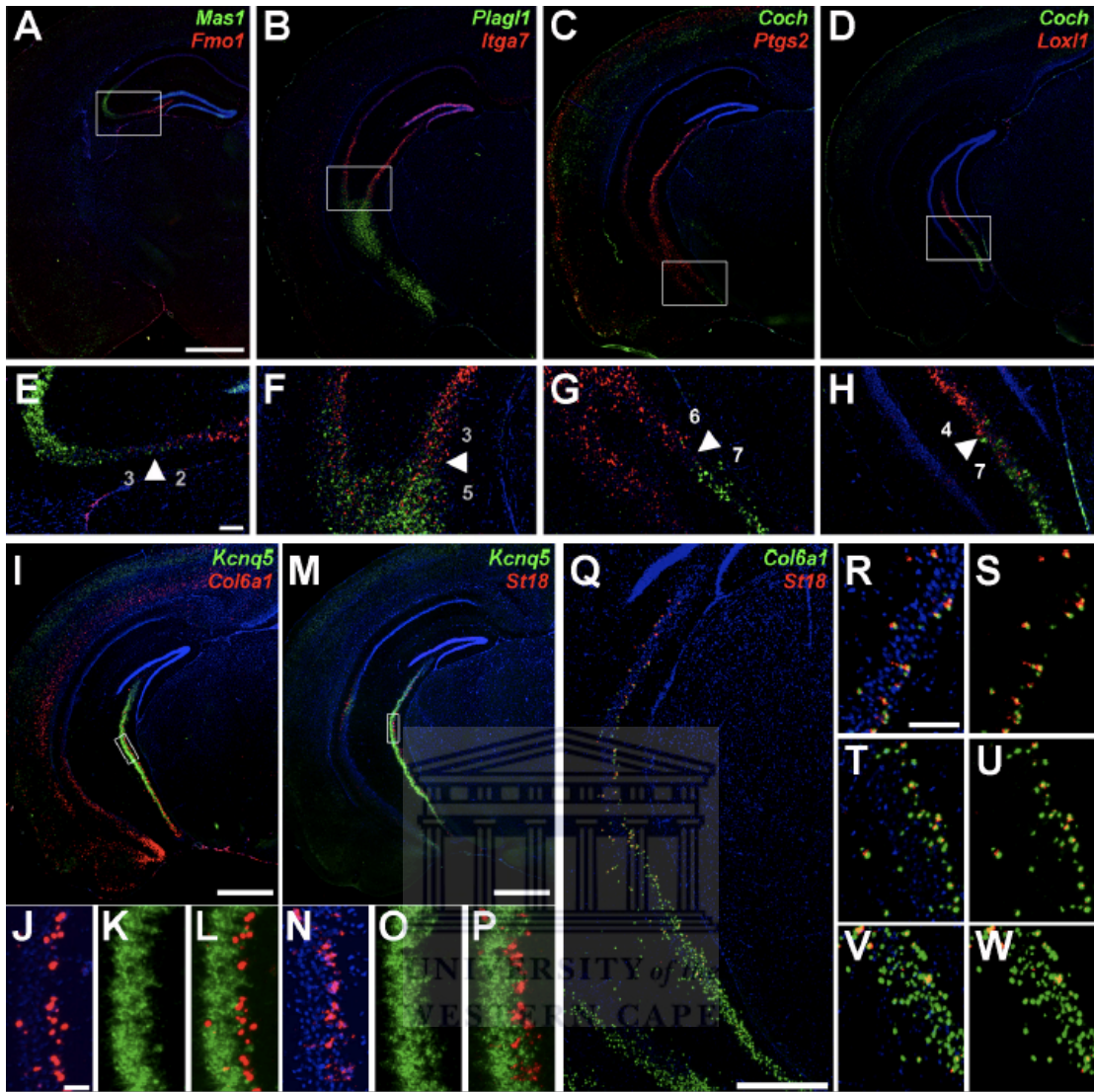


Figure 3: In situ hybridization (ISH) figures of genes expressing in CA3 tissues highlighting 9 sub-region boundaries (Thompson et al. 2007). **A-H:** Double fluorescent ISH for pairs of genes defining reciprocal boundaries in CA3 at low magnification (**A-D**) or high magnification (**E-H**). Sections are counterstained with DAPI (blue). **A, E,** *Mas1* (green) and *Fmo1* (red); **B, F,** *Plagl1* (green) and *Itga7* (red); **C, G,** *Coch* (green) and *Ptgs2* (red); **D, H,** *Coch* (green) and *Loxl1* (red). **I-W:** Double fluorescent ISH for pairs of genes differentiating inner from outer (adjacent to stratum oriens) pyramidal cells in CA3 at low (**I,M,Q**) and high (**J-P,R-W**) magnification. *Kcnq5* (green) with *Col6a1* (red; **I-L**) or with *St18* (red; **M-P**), with single gene labeling (**J,K** and **N,O**) or co-labeling (**I,L** and **M,P**). **Q-W:** *St18* labels a subpopulation of *Col6a*-expressing cells with differential co-labeling in septal (**R,S**), mid-septotemporal (**T,U**), or temporal CA3 (**V,W**). Methodological details are provided in (Thompson et al. 2007).

Table 3: Summary of gene expression defining CA3 fields. Each row indicates a border between two of the 9 anatomical fields/regions of CA3 as defined by gene expression data, Examples of genes expressing in a region specific manner are also given.

Field Border			Genes expressing on field borders		
1	/	2	LOC433436	/	Car12
2	/	3	Fmo1	/	Mas1
3	/	5	Itga7	/	Plagl1
5	/	6	D330017J20Rik	/	Rprm
6	/	7	Ptgs2	/	Coch
4	/	7	Loxl1	/	Coch
8	/	9	Grp	/	
8	/	9	Coch	/	

Viewing the CA3 sub-fields identified by gene expression data in 3D shows that the fields covering the temporal third portion of CA3 do so in a segregated manner. The remaining fields are organized into bands orientated diagonally from the septal/distal pole to the temporal/proximal pole (Figure 4). Such organization is remarkably similar to the recurrent associational projections found in the CA3 region (Ishizuka et al. 1990).

Whilst there are certain genes that express in patterns that define the CA3 field borders, the majority of genes express across various combinations of these fields. These gene expression patterns cluster into twelve groups describing the most common forms of expression patterns observed in the CA3 region (Figure 5).

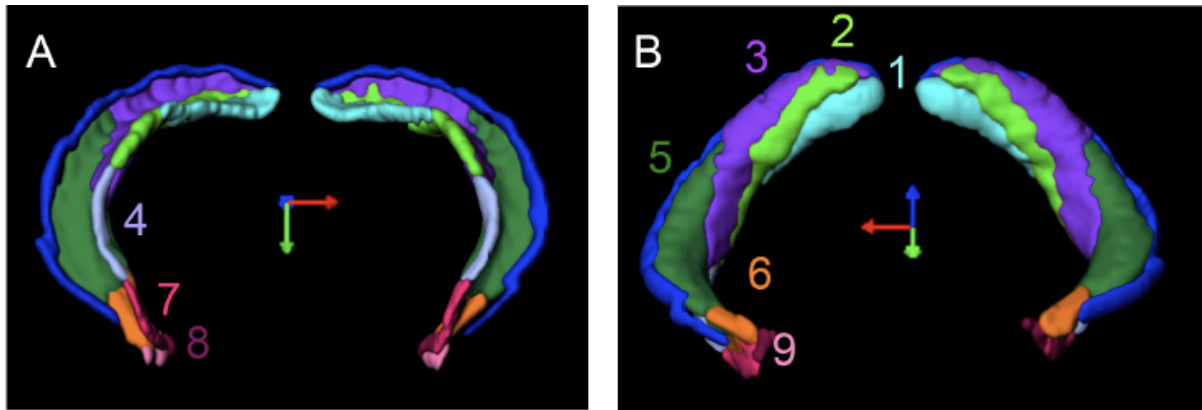


Figure 4: 3D Representation of regionalized expression of genes across CA3 (Thompson et al. 2007). **A,B:** 3D models of CA3 subdivisions delineated by gene expression boundaries. CA2 is included in this model (dark blue) for reference. Two different views are shown to illustrate major boundaries, including differentiation along the portion of CA3 proximal to the DG (**A**) and diagonal banding in the septal portion of CA3 (**B**). 3D orientation bars in (**A,B**): green (ventral); blue (pointing into page in **A**, rostral); red (lateral).

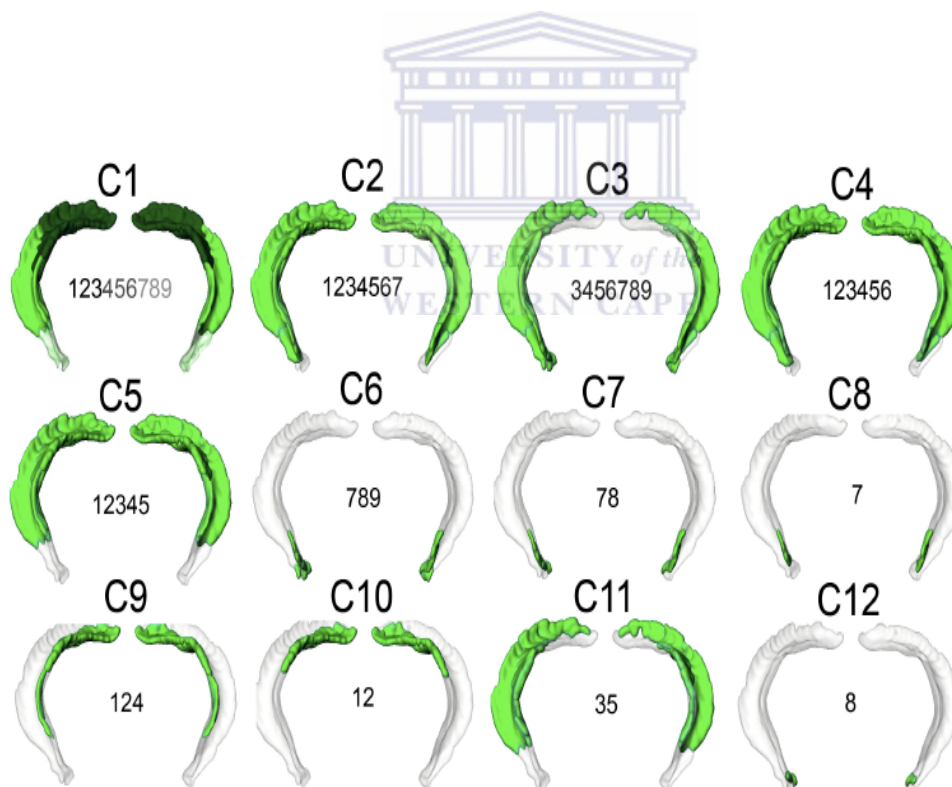


Figure 5: Twelve of the most representative gene expression profile clusters (Thompson et al. 2007). **C1-C12:** 12 most common expression patterns observed for individual genes in CA3. C1 represents a septal-high to temporal-low step gradient of expression. Numbers in C1-C12 indicate the divisions in Figure 4 spanned by each cluster.

Functions of CA3 genes correlate with gene expression along CA3 axes

Whilst genes express differently across the CA3 region in combinations of discrete fields of banded expression patterns, there are still questions about the cause and effect of this patterning. Functional analysis using public ontologies was made (Thompson et al. 2007) and identified several functional categories of genes showing differential expression in CA3. Predominantly, three large functional groups were observed: Cellular communication, including cell-adhesion factors and neuro-peptides; Physiology, including ion channels; and Genetic regulation, including TFs.

Those genes involved with cellular communications conformed to the general observations of polarized expression patterns along the septo-temporal axis. For example, Figure 6A, cell adhesion factors localize to the septal pole (Sema5a, Ephb1 and Epha7); temporal pole (Efna5, Mylk and Pcdh7); and across all CA3 fields (Pvr13 and Cdh2). This seems to suggest that neuronal cell populations display differential connectivity (afferent/efferent) in CA3 and that the connectivity is similar to the banding pattern observed in CA3 fields. Indeed, there is strong evidence that afferent and efferent neuronal projections exhibit discrete borders spanning the hippocampus (Dolorfo and Amaral 1998; Ishizuka et al. 1990; Risold and Swanson 1997; van Groen et al. 2003; Verwer et al. 1997). Notably, CA3 pyrimidal cells display comparable patterns of neural projections (Ishizuka et al. 1990) to the molecular banding patterns

observed for the two thirds of septal CA3. In the remaining temporal third of CA3, Petrovich et al. (2001) describes afferents from the amygdala akin to the patterns observed in the CA3 defined fields 7,8 and 9 (see Figure 4). However, it should be noted that cell adhesion is generally thought of as a developmental function required for initial neural tissue patterning in association with axon guidance molecules. This ambiguity might best be explained by adult DG, which maintains precursor cell populations in and actively persists in neurogenesis (Zhao et al. 2006). Additionally, recent research has implicated cell adhesion molecules in the maintenance of synaptic networks and their plasticity (Pinkstaff et al. 1999; de Wit and Verhaagen 2003). Taken together, it would seem that a large majority of expressed genes within the adult hippocampus maintain the patterning set out by the developmental process. One possible explanation to maintaining these signals could be that they would prompt the correct axonal path-finding of granule cells from the DG to the CA3 region (Zhao et al. 2006).

Research into physiological disorders like epilepsy has produced a wealth of data indicating differential usage of cellular processes along different axes in the hippocampus. The CA1 region of Ammon's horn displays differential induction of long term -potentiation (LTP) and -depression (LTD) dorso-ventrally (Izaki et al. 2000; Papatheodoropoulos and Kostopoulos 2000). Also the CA3 region is ventrally susceptible to epileptiform bursting (Bragdon et al. 1986). Additionally, the differential CA3 afferents (described previously) along the septo-temporal axis transmit neurotransmitter types differentially (dopamine, serotonin,

Npy, TRH) (Verney et al. 1985; Gage and Thompson; Köhler et al. 1987; Pazos et al. 1985). This patterning is reciprocated by the proper presence of receptor (Drd2, Htr4, Npy1r, Trhr) transcripts that show matching expression patterns to neurotransmitter connectivity. Of consequence, the physiological properties governing neurological systems are well associated with ion channels (Zuberi and Hanna, 2001). There are approximately 23 genes (see Figure 6B) associated with ion channels whose expression varies along the septo-temporal axis across CA3.

It appears as though the CA3 region of the hippocampus is segregated into 9 fields (based on gene expression) that define different biological and neural functions across the anatomy. The regulation that causes such gene expression should, presumably, be mirrored by suitable expression of TFs. Indeed, sets of TFs have been identified that show both polarized and concordant expression patterns. For example, the temporal pole transcribes TFs such as Fos12, Nab1 and Etv1; the septal pole transcribes Dscr1; while Foxo1 and Kcnip3 show expression in central regions (see Figure 6C). It is considered that TFs act in cohorts to regulate transcription correctly (Mata et al. 2007). Thus, it is not surprising that the TF expression patterns may share some similarities. However, the surprising fact is that they show high similarity of expression profiles that correlate with the nine discrete CA3 fields. Due to the tightly regulated and highly correlated expression patterns observed by TFs spanning the CA3, it is likely that the TFs responsible for transcriptional regulation of the genes expressing within the CA3 anatomy are part of the observed genes

expressed in CA3 (Figure 6,C). Yet, from expression data we do not know which TFs regulate which genes. It is likely that many of the co-expressed genes (genes in a cluster) are controlled by similar sets of TFs. We used this hypothesis and applied it to 12 observed clusters of similarly expressed genes (Figure 5) to find out TFs that potentially control expression of genes in these clusters. This has been reported on in chapter 3.



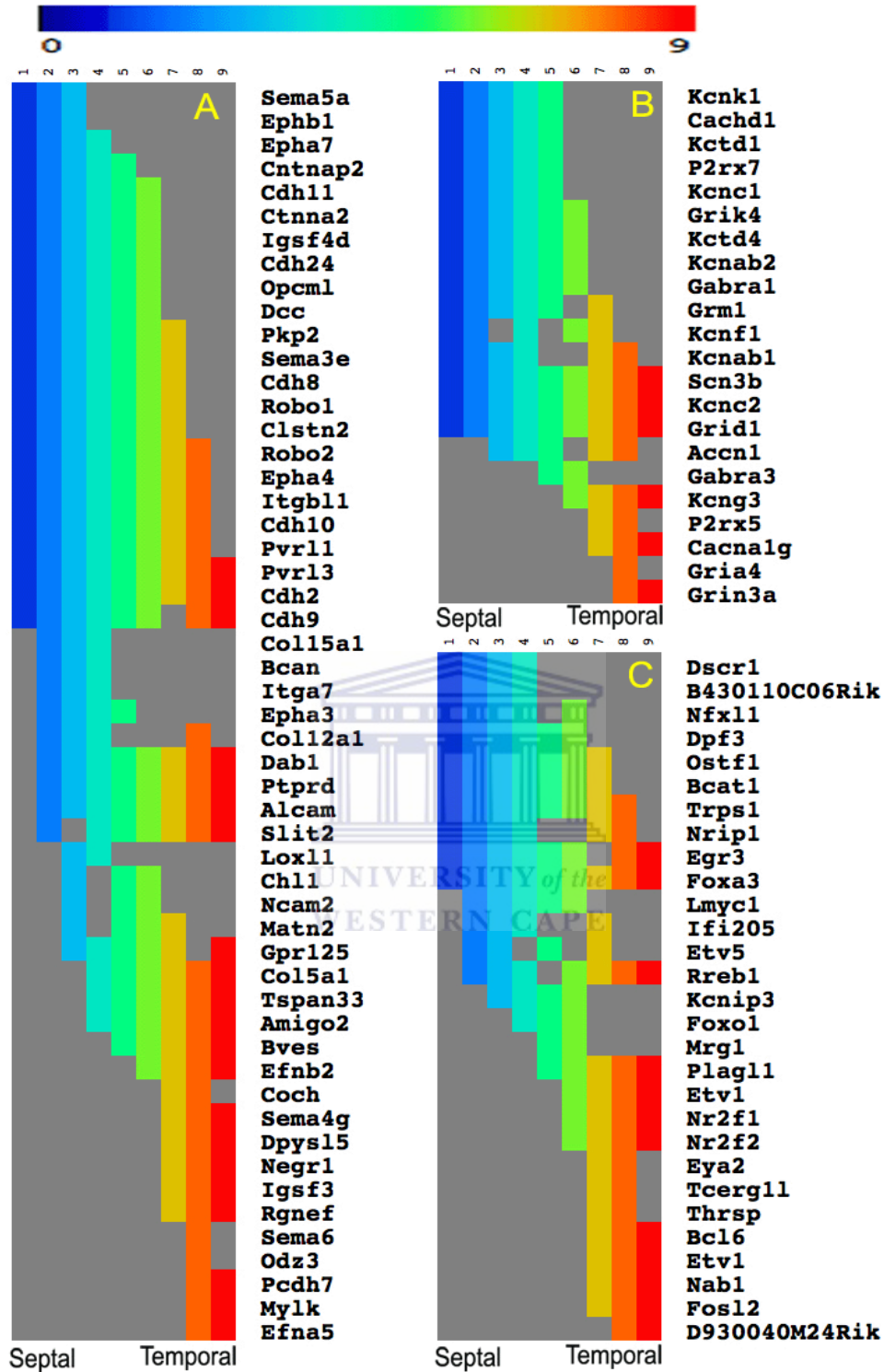


Figure 6: Distributions of gene expression across the septo-temporal axis of CA3 for functional groups of genes: (A) Cell Adhesion; (B) Ion-Channel; (C) Transcription Factors (Thompson et al. 2007).

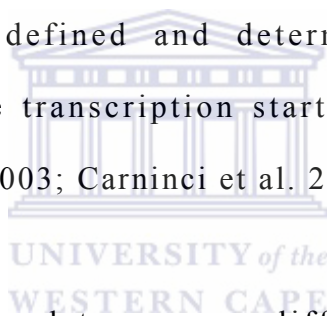
Chapter 3: Predicted transcription factor binding sites in promoters of differentially expressed genes of CA3 identify groups of co-expressing genes

Introduction

Regulation of gene transcription and the control of factors affecting that regulation amongst different cell populations is an important problem that still requires a significant amount of research to solve (Bannerwarth et al. 2006; Bernat et al. 2006). Interaction between TFs and their cognate TFBSs are currently thought of as having the greatest potential to affect transcription regulation (Kadonaga 2004). Thus, since promoters contain combinations of TFBSs they are targets of relevant TFs; and specific complexes of TFs, TFBSs and various co-factors participate cooperatively in the initiation of transcription (Kaiser and Meisterernst 1996).

Identifying those factors that play pivotal roles in the mechanics regulating transcription initiation is far from simple and current methods remain inadequate. The study primarily aimed to address this issue and identify associated TFs and TFBSs likely to regulate transcription initiation in a set of genes that express in well-defined patterns within the hippocampus. A secondary aim was conceived to realise a network of TFs and target gene promoters based on observed

gene expression across sub-anatomy in the hippocampus. To the best of my knowledge, a detailed computational analysis of the transcription regulation potential for anatomically restrictive gene expression in the hippocampus sub-regions has not been done to date. To construct such a spatial regulatory network we first required to identify the TF-gene associations that would serve as the building blocks for the network. To identify such TF-gene associations we used methodology first described in (Bajic et al. 2004) and later applied in (Bajic et al. 2006; Chong et al. 2007). The methodology required an accurate set of promoters for the genes under study in order to predict possible TFBSs. Promoters were defined and determined as the immediate surrounding region of the transcription start site (TSS) for each gene under study (Shiraki et al. 2003; Carninci et al. 2005; Carninci et al. 2006).



Promoters of genes observed to express differentially in portions of the CA3 region of the hippocampus were analysed for potential TFBS content. Since differential expression of genes was observed in nine discrete CA3 sub-regions, we attempted the identification of sets of TFs that may be responsible for such spatially restricted expression. To do this, twelve gene clusters determined from the expression of 155 individual genes across the nine CA3 sub-regions were made and their promoters analysed for TFBS content. In total we associated 610 TSSs/promoters with these 155 genes.

The methodology applied in this study identifies the TFBSs most dominant in the promoters of individual genes in each cluster as

compared to background data sets. The study used two background data sets, one consisting of 39156 mouse promoters and the other of 41000 random mouse genomic sequences. The resultant TFBSs were then used to annotate promoters of all 155 genes expressing in the CA3 region. We note that if TFBSs determined in this way (see methods for details) annotate only promoters found in one gene cluster then one could claim that these TFBSs are unique for that cluster and potentially regulate expression of genes within that cluster. Yet, the methodology does not guarantee such uniqueness. We examined this uniqueness of TFBSs to gene clusters by testing the potential of associated TFBSs to be representative of different gene clusters. Our methodology re-clustered genes using the distribution of TFBSs in annotated promoters. This contrasted the original gene clusters that were based on gene expression data only. Such re-clustering regrouped 67.74% of genes equivalent to those groups found in the original gene clusters. Promoters were annotated with a total of 178 predicted TFBS of which the top five most frequently found were '+1 Octamer' (32 promoters), '-1 Octamer' (25 promoters), '-1 myogenin/NF-1' (24 promoters), '+1 PBX' (22 promoters) and '+1 Pax-6' (19 promoters). In contrast there are 21 TFBSs associated with only one transcript and when considering cluster specificity, 58.99% are associated with only one anatomically restricted gene cluster, 24.16% in 2 clusters, 11.24% in 3 clusters, 4.49% in 4 clusters, and only 1.12% of TFBS are shared with 5 gene clusters. We hypothesise the following four TFBS to be likely to regulate gene expression across CA3 sub-anatomy: '+1 PPAR direct

repeat 1' and '+1 Tst-1' (9 promoters each) and '+1 MAF' and '-1 Cdc5' (8 promoters each).

In summary we have collected a set of predicted TFBSs for genes whos expression has been well-characterised (Thompson et al. 2007) in the CA3 region of the hippocampus. We used these TFBS collections to infer and describe the potential transcription regulation for different gene clusters and the role of TFs in controlling regional gene expression in CA3. Additionally we used this data to reconstruct parts of the spatially restricted transcription regulatory networks that potentially explain the observed gene expression across CA3 sub-anatomy.



Results from this study have produced the first detailed computational analysis of the mouse CA3 region and sub-regions in terms of transcription regulation (Thompson et al. 2007). Additionally, this study reports an initial survey into the regulatory potential of genes in the CA3 region by identifying likely TFs and cognate TFBSs. We show that each of the CA3 gene expression clusters is associated with a unique set of TFBSs. Hence we hypothesize that the observed expression patterns in each cluster may be at least partially explained by the identified TFBS sets. Thus, computationally derived regulatory potential of genes expressing in the hippocampus CA3 region reveals specificities in promoter content that could provide more insights into anatomically restricted gene expression in CA3 sub-regions.

Methods

Genes

A list of 234 genes (Thompson et al. 2007) shown to express in the CA3 region of the mouse hippocampus were provided by collaborators at AIBS. However, only 155 genes were unambiguously clustered into twelve different groups (see Figure 7). This study focused only on these 155 genes in 12 clusters.

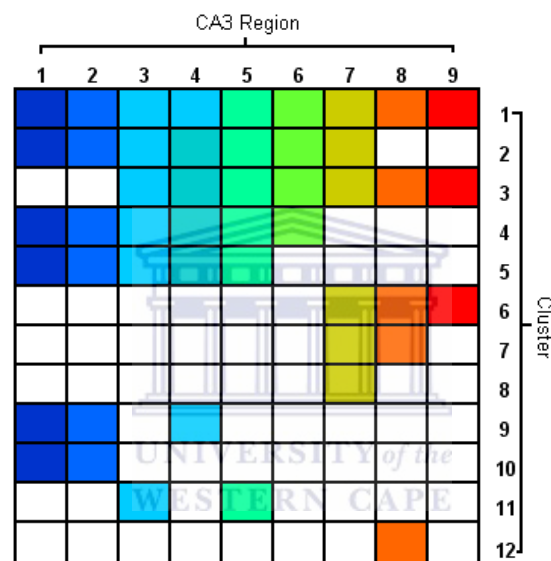


Figure 7: Association of gene expression patterns of each of 12 gene clusters with 9 physical segments of the CA3 hippocampus region in mouse. Each colour represents a different sub-region of the CA3 anatomy.

Gene functional properties

The DAVID package (<http://david.abcc.ncifcrf.gov/>) was used to search the Gene Ontology for any over-represented categories of genes amongst the 12 gene clusters.

Target promoters and background sequences

For each of the 155 genes we determined the location of the TSS using available CAGE tags (Shiraki et al. 2003; Carninci et al. 2005; Carninci et al. 2006), which are each linked to the transcriptional unit (TU) of the gene (Carninci et al. 2005; Carninci et al. 2006). Hence, linking each gene to a promoter through a TSS. However, since TSS locations are determined based on CAGE tags, it is possible that one gene could have multiple promoters and that each promoter could have multiple TSSs. Table 4 presents the distribution of TSSs per gene cluster with a total of 610 TSSs.

Table 4: Distribution of TSSs relative to clusters (Bajic et al. 2006A).

Number of TSS locations	cluster
51	1
79	2
20	3
176	4
48	5
64	6
40	7
24	8
21	9
42	10
30	11
15	12

Whilst the above use of CAGE tags provided a set of promoters for our target data, we required a comprehensive set of all mouse promoters and random non-promoter mouse DNA. These sets of

sequences served as background sets to determine whether or not the target promoter sets (determined by TSS analysis for each of the clusters) contained overrepresented TFBS.

A background mouse reference promoter set of 39156 promoters was prepared as follows (Bajic et al. 2006): Promoters were defined as the region -1000 to +200 surrounding the position of the TSS. A TSS was immediately accepted should the first 5' nucleotide of a CAGE tag or 5' ditag agree with the first 5' nucleotide of a corresponding full-length cDNA (flcDNA). However, this did not always occur and in these cases we chose a TSS location supported by a tag cluster (TC). Each TC contained at least 10 CAGE tags. The chosen TSS had to be supported by at least six tags within a cluster and pass the added restriction of having support from transcriptional evidence associated with the TC such as EST, flcDNA, and/or long SAGE.

A random mouse DNA set of 41000 background sequences was selected as follows: Random DNA sequences of 1200bp in length were selected from all mouse chromosomes. The number of sequences selected from each chromosome was proportional to the length of each chromosome.

Mapping of promoter elements

We used only the mammalian subset of available TFBS position weight matrix models found in TRANSFAC Professional (Ver. 9.4) database (Matys et al. 2006). These matrices were used to find likely TFBSs on either of the positive and/or negative DNA strands in each extracted promoter (target and background). A threshold (designated

as *minFP*) was chosen for the matrix models, which optimises the core and matrix scores (Kel et al. 2003) such that the predicted set of TFBSs contain a minimum number of false positives.

Determination of the most dominant TFBSs

To resolve the set of dominant TFBSs in the target promoters we used a method that calculates a likelihood of observing TFBSs in target against background promoters, as described in (Bajic et al. 2004). Two background sets were at our disposal: 1/ Whole mouse promoter set (39156 promoters); 2/ Random set (41000 random genomic sequences). TFBSs found within target promoters were ranked highest to lowest based on their over-representation index (ORI), which is a ratio of concentration in target promoters over concentration in background sets and is further described in (Bajic et al. 2004). Thus, if the concentration of a particular TFBS in the target promoter set equals that of the background (i.e. no enrichment) then the ORI value would equal 1; the greater the presence of a TFBS in the target set is, relative to the background, the higher the ORI value. Using a right-side Fisher's exact test (based on a hyper-geometric distribution) we consider the null-hypothesis that the set of TFBSs found in the target set are proportionally the same as the background set. P-values can be viewed in the provided reports (Supplementary Data File 01 through 12), where corrected (Bonferroni) p-values not greater than 0.05 are annotated with a '+' sign after the ORI value. Supplementary data files 01 through 12 contain summaries on top ranked TFBSs. Each TFBS was selected according to the following criteria: Each TFBS had to

score an ORI value of at least 3 and occur within at least 20% of target promoters from one of the 12 gene clusters. A cartoon description of the format used in Supplementary data files 01 through 12 is displayed in Figure 8.

TFBS pattern	ORI	% TAR	% BCG	Prob TARGET	Prob BACKGR	# TAR	# BCG	TOT TAR	TOT BCG	p value
+1 FOXD3	9.547	23.53	9.78	5.23E-04	1.32E-04	12	3831	51	39156	1.00E+00
-1 PEA3	6.362	49.02	19.36	4.58E-04	1.82E-04	25	7590	51	39156	2.01E-03

Figure 8: Explanation of columns in Supplementary Data Files 01 through 12 (Bajic et al. 2006A).

Annotation of promoters by TFBSs

We used TFBSs with ORI value greater than 3 and present in at least 20% of target promoters to annotate promoters of all genes expressing in CA3. The TFBSs are DNA strand specific and are either found on the positive or negative DNA strands. Thus the TFBSs are written with the prefix '+1' or '-1' to indicate strand.

Clustering and generation of heat-map graphics

TIGR-MeV (Saeed et al. 2003) software was used to cluster genes according to their TFBS annotations. TIGR-MeV was also used to graphically present results, whilst the CytoScape suite (Shannon et al. 2003) was used to generate figures of network graphs.

Generation of data for network representation

- **Gene lists: Discontinuities/Irregularities**

We excluded from the analysis genes that appeared in the original list of twelve clusters provided by AIBS, but which could not be matched to FANTOM3 data for the purposes of identifying associated TSSs.

- **Promoters of the target sets: Summary for TSSs**

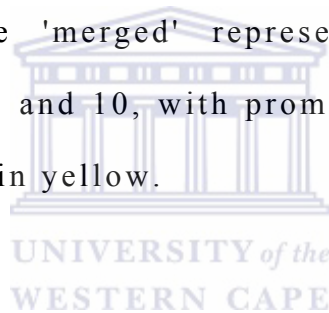
We provided information for each gene name, the genomic location of the TSS and the number of CAGE tags in each tag cluster. The conditions under which the background promoters were determined had initially been too restrictive for the identification of promoters in the target sets and were subsequently relaxed. Hence tag clusters were required to contain at least five CAGE tags and at least three of those tags supporting the TSS.

- **Gene Clusters: Visual representation of the association between identified TFBSs and genes, irrespective of the individual promoters**

Heatmaps were generated by mapping promoters to strand specific TFBSs using the log of the number of TFBSs found in each promoter (log value indicated by colour).

- **Network visualization: A visualization of the gene cluster information using Cytoscape**

The networks consist of TFBS→gene edges. These were created by first associating promoters to genes and then promoters to TFBSs. The networks are depicted in Figures 14. TFBSs that were mapped to both the forward and reverse strand of gene promoters have been merged into single TFBS nodes for this representation and are displayed as blue rectangles. Genes are represented as orange circles and linked to their binding sites by black edges. Red edges reflect manually curated information on protein-protein interactions between the TFs. Figure 16 depicts enlarged the 'merged' representation of a union of information for clusters 8 and 10, with promoter elements common to both clusters highlighted in yellow.



Results

Spatial regulatory networks: The most dominant TFBSs for individual clusters

Information within the Supplementary Data files 01 through 12 describes each of the 12 clusters by what dominant TFBSs are present in each cluster (see Methods and Figure 8 for discription of files).

From Supplementary data files 01 through 12 we used the TFBSs to annotate each clustered genes' promoter. We note that some TFBSs appear to be found in several clusters, yet 83.15% are found associated with two clusters at most, suggesting that TFBSs are utalised in a directed manner for regulation of transcription initiation.

Mapping of TFBSs to our target DNA sequences produced a large set containing many false predictions due to imperfect models. Whilst expected, it remains unfortunate that there is no computational method that can resolve this issue. To increase quality of the mapped set and reduce the frequency of false predictions we contrasted the target set with background promoter sets (Bajic et al. 2004) and discarded all TFBS mappings that did not show enrichment in the target set.

We visualised the distribution of identified dominant TFBSs in each of the 12 clusters in Figure 9, which displays a tendency of TFBSs to be cluster specific. Figure 9 displays a set of TFBSs that were determined using the background data set of 39156 mouse promoters

and an ORI of not less than 3 (see Methods). The high ORI criteria filtered out TFBSs that are less enriched in the target promoter sets than the background. Again, note that these TFBSs cluster generally along the anatomically restricted gene clusters.

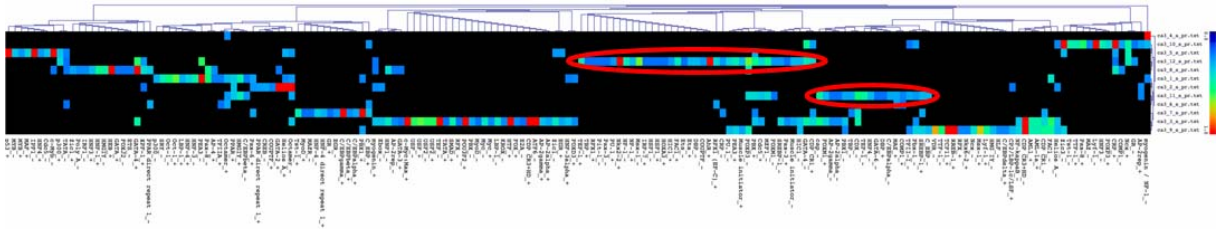


Figure 9: Distribution of TFBSs (horizontal) across 12 CA3 clusters (vertical). ORI value was not less than 3. The background set was mouse promoter set

The ORI criteria becomes increasingly more restrictive the higher it is set, as can be seen when ORI is not less than 4 in Figure 10. The distribution of TFBSs that are uniquely associated with promoters from only one cluster are displayed in Figure 11.

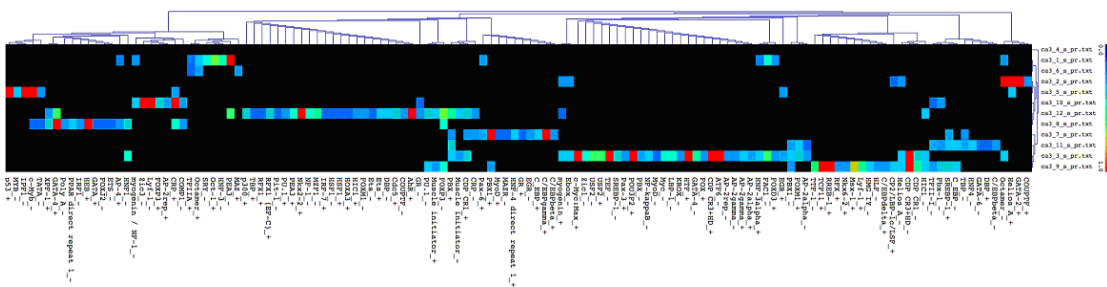


Figure 10: Distribution of TFBSs (horizontal) across 12 CA3 clusters (vertical). ORI not less than 4 and when promoter background is used

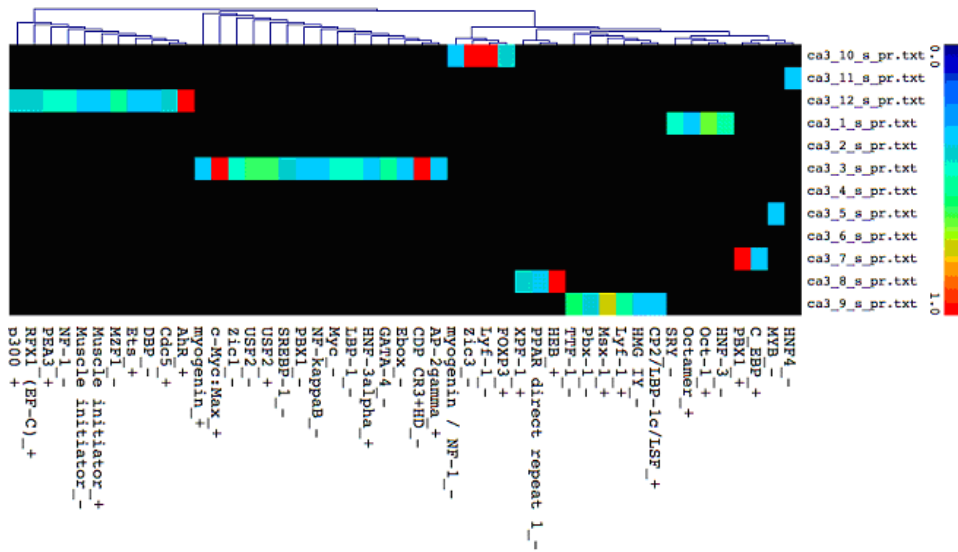


Figure 11: Distribution of TFBSs (horizontal) across 12 CA3 clusters (vertical). Here, TFBSs unique to one of the 12 clusters are displayed.

The distribution of TFBSs across gene promoters within the twelve clusters was checked to determine if it remained qualitatively similar when using a background of random genomic sequences as apposed to the mouse promoter sets. Indeed, the qualitative similarity is high in that specialization of the collections of TFBSs associated with the various clusters could be observed.

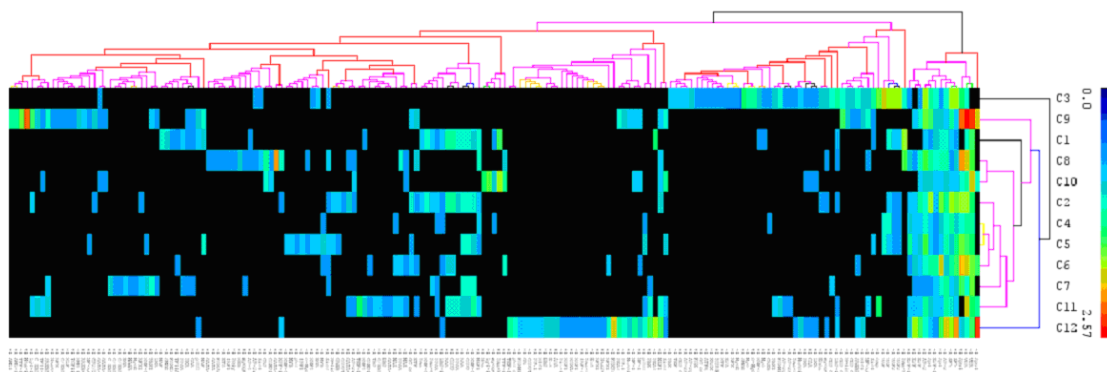
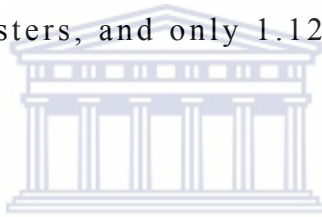


Figure 12: Distribution of significant TFBSs obtained by contrasting cluster target promoter data with the non-promoter (random) genomic background. ORI not less than 2 is used

A subset of Supplementary Table 01 in Table 5 shows dominant TFBSs and the frequency that they were found in each of the 12 clusters. The study identified 178 unique TFBSs that were used to annotate all extracted promoters for the set of clustered genes. In summary, the top five most frequently found TFBSs: '+1 Octamer' (32 promoters), '-1 Octamer' (25 promoters), '-1 myogenin/NF-1' (24 promoters), '+1 PBX' (22 promoters) and '+1 Pax-6' (19 promoters). In contrast, when we consider how unique TFBSs are, we find: 21 TFBSs that appear in only one transcript; 58.99% TFBSs appear exclusively in genes of one anatomically restricted gene cluster, 24.16% in 2 clusters, 11.24% in 3 clusters, 4.49% in 4 clusters, and only 1.12% are shared with 5 gene clusters.



It is likely that '-1 Octamer' and '+1 PBX', which are found in 5 different clusters each, share the least chance to regulate anatomically restricted gene expression since they represent the most promiscuous TFBSs within the set.

Table 5: List of significant TFBSs and their distribution across 12 clusters (C1,C2,...,C11,C12). The list is a subset of Supplementary Table 01.

TF	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
+1 AP-4	2	0	0	0	0	0	0	0	4	0	0	0
+1 FOXD3	3	0	1	0	0	0	0	0	0	0	0	0
+1 HMGY1	5	0	0	0	0	0	0	0	0	0	0	6
+1 HNF-3 α	2	0	3	0	0	0	0	0	0	0	0	0
+1 HNF-4	4	0	0	0	0	0	0	0	0	0	0	0
+1 Helios	4	7	0	0	6	0	0	0	0	0	0	0
+1 LBP-1	4	0	0	0	0	0	0	0	0	0	0	0
+1 Oct-1	6	0	0	0	0	0	0	0	0	0	0	0
+1 Octamer	6	0	0	15	0	5	0	0	0	0	0	6
+1 PBX1	2	0	0	0	0	0	6	0	2	0	0	0
+1 PPAR	4	9	0	0	0	0	0	0	0	0	0	5
+1 Pax-6	5	7	0	0	0	0	7	0	0	0	0	0
+1 TFIIA	2	0	0	0	0	7	0	0	0	0	0	0
-1 C/EBP β	4	0	0	0	0	0	0	0	0	0	0	4
-1 FAC1	5	0	3	0	0	0	0	0	4	0	0	0
-1 HNF-3	6	0	0	0	0	0	0	0	0	0	0	0
-1 Oct-1	6	0	0	0	0	0	0	0	0	0	0	0
-1 Octamer	4	9	2	0	5	0	0	0	0	0	0	5
-1 PEA3	5	0	0	0	0	0	0	0	0	0	0	4
-1 Pax-8	5	0	0	0	0	0	0	5	0	0	0	0
-1 SRY	6	0	0	0	0	0	0	0	0	0	0	0
-1 myogenin	5	0	0	14	0	0	0	0	0	0	5	0

Intuitively, those TFBSs that influence anatomically restricted gene expression most, should also be those that are exclusively found to be associated with one gene expression cluster. Furthermore, these TFBSs should be found in the majority of promoters within that cluster. Using this idea as criteria we ranked cluster specific TFBSs and find the top 4 as: ‘+1 PPAR direct repeat 1’, (9 promoters), ‘+1 Tst-1’ (9 promoters), ‘+1 MAF’ (8 promoters) and ‘-1 Cdc5’ (8 promoters). A cursory inspection of peer-reviewed literature reveals that PPAR (Teboul et al. 2001) and MAF (Li-Weber et al. 1997) are known controllers of tissue- and cell-specific transcription, whilst Tst-1 is thought to influence the control of tissue-specific expression (Josephson et al. 1998). Nothing similar could be found for Cdc5.

Re-clustering data based on annotation by significant TFBSs

In order to evaluate how well the computationally determined significant TFBSs relate to the twelve anatomically defined gene clusters we re-clustered data using only significant TFBSs for promoter annotation. This clustering we termed feature-based clustering. Thus, we expected to find that there would be a minimal difference between feature-based clusters and those clusters derived anatomically by gene expression. At first glance there might be a perceived error in logic to this evaluation as circularity of arguments: The TFBSs are derived in a stepwise manner, cluster-by-cluster, and thus, one might expect a minimal difference in feature-based clusters and those by expression based clustering which would render this test meaningless. However, TFBSs are derived based on sequence data and a static background promoter sets. The significant TFBSs found need not be cluster-specific although they could be derived for a specific cluster. The reason is that no condition is used to restrict TFBSs to be specific for only one individual cluster. Indeed, some of the TFBSs found to be dominant in one cluster, are also found to be dominant in other clusters. Hence, if the promoters with annotated TFBSs cluster in a highly correlated fashion with the anatomically restricted gene expression clusters, then this method has extracted the part of the information that specifies promoter properties in a potential role of regulating gene expression.

We used k-means clustering with 12 seeds (Saeed et al. 2003). The actual association of the resulting feature-based clusters is obtained using dominant TFBSs from all 12 gene expression clusters. The correlation of these clusters with the anatomically restricted gene expression clusters can be inspected in Supplementary Figure 01. Each of the sub-figures in Supplementary Figure 01 shows a separate cluster of genes. We highlighted genes by a different colour based on the anatomically restricted gene cluster, so as to be able to easily observe the effects of re-clustering. Thus, if the genes in the feature-based clusters originally came from the same anatomically restricted cluster, then the colour strip on the top of the graph will indicate that colour. Thus, at a glance, a high correlation between original and new clusters can be observed by the preservation of the colour strip. Only 120 of the 155 genes were annotated since our criteria required that each TFBSs score an ORI of at least 3 and were found in at least 20% of the promoters. Thus, only these 120 could be clustered, and whilst we observe 12 clusters we note that 15 genes were unclassifiable based on available annotations. Therefore our sensitivity can be given as approximately 67.74% (105/155) and since no gene that is classified is wrongly classified, our positive predictive value (PPV) is 100%.

Considering the many methods available to cluster data we applied two additional clustering methods (hierarchical and support tree clustering) in order to evaluate whether the results remain similar and are not the artefact of the clustering method used. Depicted in

Supplementary Figure 02, both clustering methods show gene groupings based on TFBS annotations that largely coincide with one another as well as to the original expression based clusters. Figure 13 displays the resulting heatmap of the support tree clustering method. The genes along the horizontal axis are highlighted in one of twelve colours in the same manner used in the k-means clustering. Again it is evident that the distributions of predicted TFBSs across gene promoters contain enough information to partition genes into groups of similar gene expression patterns (although feature-based clustering is used).

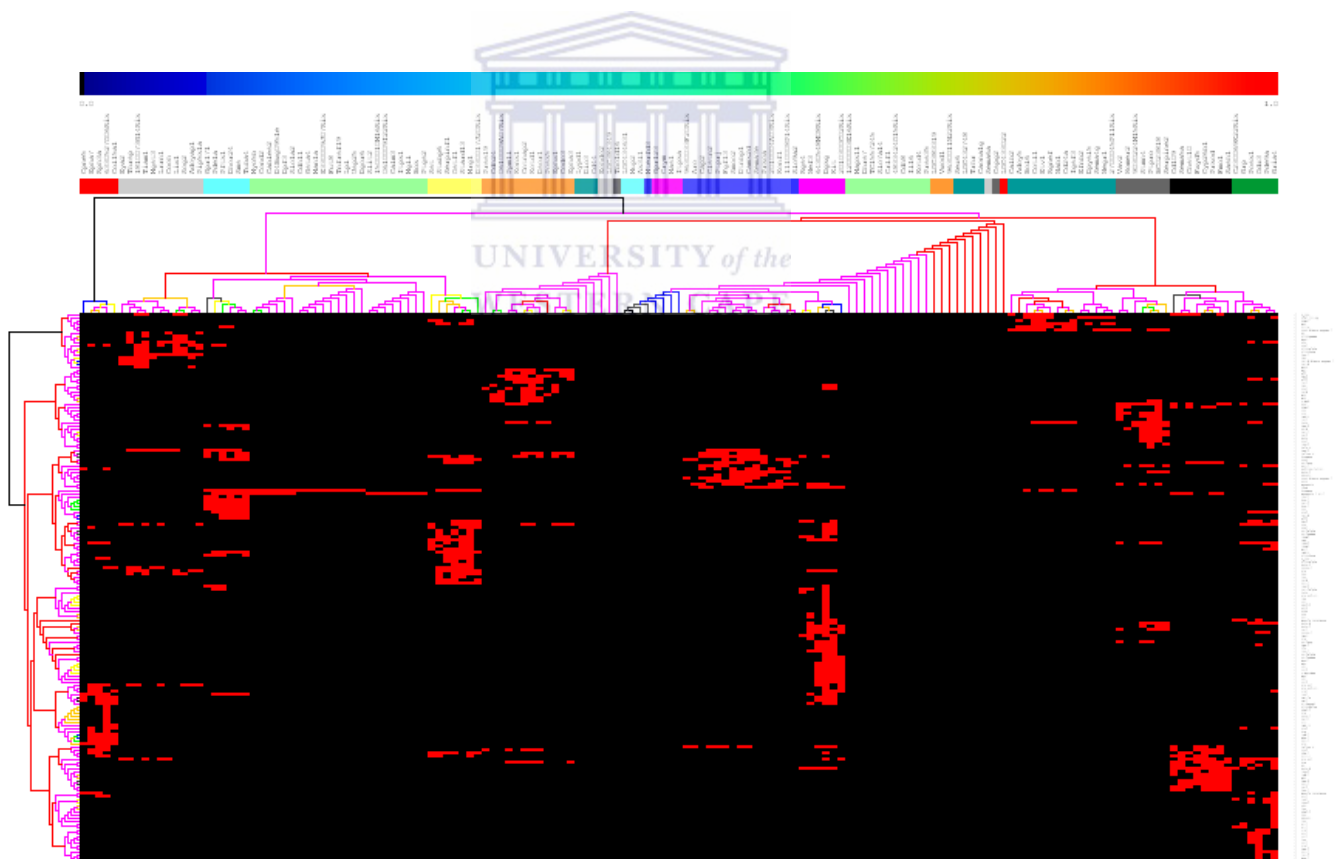


Figure 13: Heat-map distribution of genes (horizontal) obtained by support tree clustering of annotated promoters (using TFBS, vertical) using Euclidean distance

Functional specificity of gene clusters

Whilst there is no reason to expect that members of genes in the same gene cluster would share any functional similarity, it was an intriguing possibility. We queried the DAVID system (Dennis et al. 2003) for Gene Ontology (GO) (Ashburner et al. 2000) categories and analysed if there are any significant functional specializations for genes in each cluster. Unfortunately, the DAVID system could not detect any significant functional GO categories amongst any of the gene expression based clusters. Yet, over 60% of genes in cluster 8 were found to be associated with cell signalling and 50% of genes in cluster 10 were found to be associated with cell growth and maintenance. Hence, we investigated the networks of promoters and TFBSs for clusters 8 and 10, since the remaining 10 gene clusters displayed a large degree of functional diversity.

Transcription regulation networks

As previously mentioned, clusters 8 and 10 contain the most genes with uniform function (similar GO categories). Here we illustrate several networks potentially regulating the transcription of genes found in clusters 8 and 10 (Figure 14). The interesting issue is that few TFs appear to be common to both networks, while the networks are largely controlled by disparate sets of TFs. This can be explained by two arguments: 1/ clusters 8 and 10 contain genes expressing in

two distinct and different anatomic regions; 2/ genes are also associated with different functional GO categories.

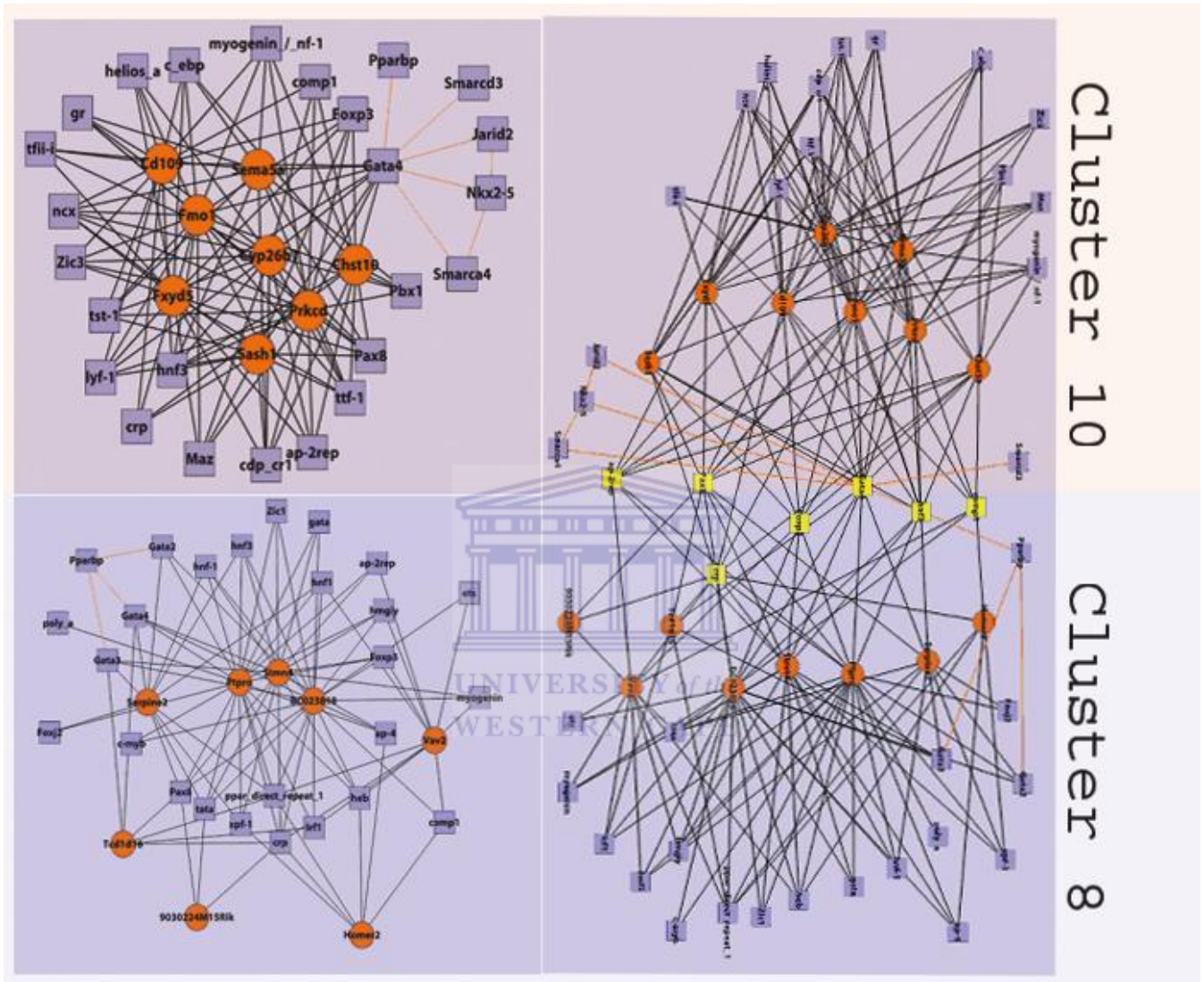


Figure 14. Transcriptional regulatory network for cluster 10 (Top left). Transcriptional regulatory network for cluster 10 (Bottom left). Transcriptional regulatory network for combined clusters 8 and 10 (Right). TFBS identified as significant for both clusters are highlighted in yellow, TFBS unique to the individual clusters are depicted in blue.

Conclusion

Of the whole set of mammalian TFBS matrix models contained within the Transfac Professional database, we have selected a set of 178 dominant TFBSs found in 120 gene promoters across the 12 gene expression clusters from the CA3 region of the hippocampus in the adult mouse. The distributions of the set of 178 TFBSs amongst gene promoters are both largely heterogeneous between clusters but similar within the promoters of each cluster and are able to distinguish between promoters of each gene cluster. Since each gene cluster represents a set of similar and anatomically restricted expressing genes, the sets of TFBSs identified as dominant for any particular cluster contains possible key controllers of targeted transcription for the genes in that cluster. Additionally, the distribution of TFBSs is sufficient to predict groups of similarly expressed genes with some degree of accuracy due to the TFBSs being significantly cluster-specific. However, those TFs that are found significant in several clusters might play dominant roles in defining expression boundaries within CA3 hippocampus as they are shared by different sub-regions. Whilst such boundaries would have an increased significance within the context of the developing mouse brain for the purposes of tissue patterning and defining the cellular layers common to brain tissues, it should not be forgotten that the dentate gyrus, neighbouring the CA3 region, maintains active neuronal precursor cell populations that would require CA3 borders to be maintained for correct innervation via proper axon guidance cues. Ultimately, it seems that targeted

cellular expression in the CA3 region is the result of combinations of TFs. Results generated from this study provide a means for targeted experiments to resolve the function and influence of the most significant TFBSs within each gene clusters.



Chapter 4: Networks of Gene-to-TranscriptionFactor edges elucidate key nodes in the regulatory potential behind differential gene expression along the septo-temporal axis of CA3 anatomy

Introduction

Our primary task has been in locating TFs that putatively control expression of genes in 12 gene clusters determined by anatomically restricted gene expression. These clusters, 'C1' through 'C12', are obtained from genes displaying differential expression across 9 regions along the septo-temporal axis of CA3. Noting that our method revolves around the annotation promoter sequences by TFBSs, our method selects only those TFBSs that are dominant in the promoters of the target gene groups. The premise being that we expect that TFBSs, being targeted DNA motifs, must be present in promoters of similarly expressing genes more often than one would expect in random or non-promoter DNA. In such cases, we consider the possibility that the over-represented TFBSs confer some degree of transcriptional regulatory control to a group of co-expressing genes.

The methodology is limited in that it cannot single out the actual TF that binds to the identified TFBSs. This is a corollary of the following:

a/ TFBSs are computationally determined based on the matrix models of mammalian TFBSs from Transfac Professional database ver 9.4. Matrix models of TFBSs are not perfect and using them to predict TFBSs inevitably introduces false positives to the motif set. Whilst each model predicts sets of motifs on the target DNA that are similar to the motifs from which the model is derived, there is no certainty that the predictions will bind only the TFs whose binding sites were used to create the model. Moreover, it is well known that TFBSs can bind TFs from various remote families and whilst we may for example report that a predicted motif binds Tst-1, other TFs are also known to bind the motif (see Table 6).

b/ There is no common method of naming TFs and assigning them with matrix models from the Transfac Professional database. This issue offers obscure interpretations of the results.

To resolve the above-mentioned issues we processed the Transfac database along with publicly available databases for synonyms of TFs to reduce the multiplicity and redundancy of TF-matrix associations found in the Transfac Professional database. Thus, for each significantly predicted motif we show a detailed list of TFs known to bind such motifs. Additionally, using a network representation and meta-nodes, key genes have been identified that contain unique

distributions of TFBSs. These TFBSs seem to support regulatory control in gene expression clusters. Finally, the relative control that predicted TFBSs may exert on gene expression is dependant on what resolution of anatomy is analyzed. In this case, the most proximally located region to CA3 is CA2. The same method used to predict TFBSs across genes highly expressed in CA3 in Chapter 3 is used for CA2 gene sets. CA2 TFBS distributions were compared to CA3 distribution and those TFBSs specific to CA3 anatomy resolved. Furthermore, data displayed in the form of a network identifies discrete sub-networks of TFBSs particular to gene expression clusters. This demonstrates the potential regulatory mechanisms behind differential septo-temporal expression along the CA3 anatomy of adult mouse hippocampus.



Methods

Construction of Table 6

Cluster specific TFBSs have been extracted in Supplementary Table 02, where TFBSs are ranked according to their frequency. For example, C7 “+1 Tst-1” represents a TFBS that is specific for cluster C7 and is found in the promoters of 9 genes.

Please note that Table 6 contains only data that relates to TFBSs found to be unique for a cluster. This means, every TFBS in Table 1 has been found to be in the promoters of genes from only one of the 12 clusters. If a TFBS appeared in promoters of genes of more than one cluster, it has been excluded from this table. The reason for this is the aim to find the most likely TF candidates that affect anatomically restricted gene expression in CA3 region.

Construction of Figure 15

Data from Supplementary Table 01 was used to map gene clusters to their associated TFBSs. Each edge between the nodes was then weighted according to the frequency of the TFBSs within the cluster. The weightings are viewable within Figure 15 as a thickening of the edge line; the network is coded in XGMML file format, an example of which can be viewed in Supplementary Data File 13.

TFBSs are represented as either green diamonds or triangular nodes. The diamond shape is used to denote TFBSs that appear on both

positive and negative DNA strands, while triangular nodes are used for TFBSs that are found only on one of the DNA strands. These TFBSs are listed in the file Supplementary Table 03. TFBS nodes are linked to the cluster nodes to which they are associated. Each set of cluster-specific TFBSs have been listed in Supplementary Table 04. The cluster-non-specific TFBSs have also been listed at the end of the Supplementary Table 04. Cluster specific TFBSs have been extracted in Supplementary Table 02, where TFBSs are ranked according to their frequency. For example, C7 “+1 Tst-1” represents a TFBS that is specific for cluster C7 and is found in promoters of 9 genes. See Supplementary Figure 03 for the full network of genes and TFBSs.

Construction of figure 16

Figure 16 represents the synthesis of data from Figure 15 as well as data from heat-map Figure 13. The network required introduction of additional gene nodes such that the edges within the network are displayed as TFBS→GeneCluster for those TFBSs that have been associated to multiple gene clusters. Also, edges of the form TFBS→Gene→GeneCluster are used to form cluster-specific elements within each of the 12 clusters. Node are colored and shaped the same as in Figure 15. Gene nodes are green circles. See Supplementary Figure 03 for the full network of genes and TFBSs.

Construction of Figure 17

Distributions of TFBSs for CA2 region were obtained using the same methods as described for Figure 15 except that no gene expression

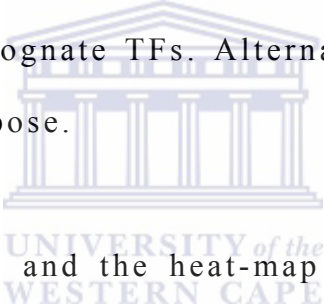
clusters were determined for CA2. Both CA2 TFBSs, as well as mapped CA3 TFBSs for each CA3 cluster, were plotted in the form of a network. Figure 17 depicts a union of TFBSs for CA2 and CA3. The format of Figure 17 is similar to Figure 15 except with the addition of the CA2 meta-node (a node used to classify other nodes into more informative groups) and the coloration of edges. All edges belonging to one of 12 CA3 clusters are colored blue, and green if belonging to CA2. Note that all TFBSs that are common to CA2 and CA3 have both green and blue edges associated with them. Also, all TFBSs common to CA2 and CA3 that were uniquely associated to one of twelve CA3 clusters in Figure 15 have been pulled away from the inner wheel in Figure 17.



Results and Discussion

Exploring the distribution of TFBSs across gene clusters of CA3

It is possible to prioritize TFs of interest that exhibit potential control of any of the 12 considered gene clusters. In Table 6, TFs are ranked according to their anticipated likelihood of relevance to the control of gene expression in a cluster. This relevance is determined based on the number of genes in whose promoters TFBS are found. We consider a TFBS more relevant the more genes' promoters it has been found to be associated with. Furthermore, as TFBSs bind multiple TFs, the same holds for cognate TFs. Alternatively, one could use p-values for the similar purpose.



Corresponding to Table 6 and the heat-map presented in Figure 13, Figure 15 depicts a network of clusters and TFBSs. Initially, clusters of genes restricted to particular sub-regions of CA3 anatomy were analyzed for the presence of dominant putative TFs. This data presented in Figure 15, was part of an effort to understand the common control mechanisms among transcriptional regulation in these clusters. Accordingly, the association between TFBS to gene cluster can be weighted by the number of genes in a cluster the TFBS is associated with and drawn as a set of edges, gene clusters, and TFBSs (Figure 15).

Figure 15 is a basic network demonstrating which TFBSs may regulate which cluster of genes in terms of transcription initiation. However, its simplicity, allows for in depth exploration into possible regulatory relationships applicable within the sub-anatomy of CA3. Figure 15 is constructed from TFBSs and gene cluster (GC) nodes. The link (edge) between the TFBS and GC nodes is derived from the frequency of TFBS associations to genes within a GC.

TABLE 6. Link of TFBSs and TFs that are known to bind such motifs. **Frequency:** gives the number of genes in the cluster (indicated in second column) whose promoter contain TFBS (column 4) that can bind indicated TF. **Cluster:** indicates the anatomic cluster. **Strand:** indicates the strand where TFBS has been mapped. **TFBS:** this is the mapped binding site. **Transfac ID:** gives Transfac database identifier of the TF that can bind given TFBS. **Transfac gene ID:** gives Transfac ID of the gene that produces TF that can bind given TFBS

Frequency	Cluster	Strand	TFBS	Transfac	Transfac	Swissprot	Entrez
9	C7	1	Tst-1	T00655	G009112	Q03052	5453
9	C7	1	Tst-1	T00656	G009192	P21952	18991
9	C7	1	Tst-1	T00969	G014470	P20267	192110
9	C2	1	PPAR direct repeat 1	T00694	G004881	P23204	19013
9	C2	1	PPAR direct repeat 1	T00991	G014219	P37230	25747
9	C2	1	PPAR direct repeat 1	T01352		P37232	
9	C2	1	PPAR direct repeat 1	T01353		P37233	
9	C2	1	PPAR direct repeat 1	T01354		P37234	
9	C2	1	PPAR direct repeat 1	T02529	G004880	P37238	19016
9	C2	1	PPAR direct repeat 1	T02726	G002888	Q07869	5465
9	C2	1	PPAR direct repeat 1	T02736	G004022		5468
9	C2	1	PPAR direct repeat 1	T03731		Q15832	
9	C2	1	PPAR direct repeat 1	T04780	G002769	O18971	281993
9	C2	1	PPAR direct repeat 1	T04794			
9	C2	1	PPAR direct repeat 1	T05221			
9	C2	1	PPAR direct repeat 1	T05235			
9	C2	1	PPAR direct repeat 1	T05236			
9	C2	1	PPAR direct repeat 1	T05243			
9	C2	1	PPAR direct repeat 1	T05246			
9	C2	1	PPAR direct repeat 1	T05256			

The network representation from Figure 15 is easier to use for analyzing possible regulatory relationships amongst the genes,

remembering that each node that corresponds to a TFBS appears in the network as a consequence of already being found as sufficiently over-represented by our computational analysis.

The Figure 15 network consists of only TF to gene cluster associations and is represented in a circular pattern similar to a wheel, where nodes with more than one association are directly on the wheel. This means that TFBS that appear in only one cluster are depicted as a group outside of the wheel and are associated with that cluster node (coloured red). These unique outer groups of TFBSs represent the greatest potential of the predicted TFBS set to determine anatomically restricted gene expression in the associated gene cluster. However, that being said, TFBSs associated with multiple gene clusters are significantly over-represented in those clusters and cannot be ignored since they certainly have potential for regulating cluster specific gene expression. Indeed, it is most likely that combinations of promiscuous and cluster specific TFBSs cooperate in the control of gene expression across the CA3 anatomy. Yet, cluster specific TFBSs are more suitable for experimental evaluation as results would be simpler to interpret.

Figure 15: Distributions of TFBS across CA3 gene expression clusters



Identification of candidate genes, per cluster (based on the presence of TFBS), for anatomically restricted expression in CA3

The list of TFs in Table 6 is critical for the determination of TF candidates for biological experiments, however, it does not associate TFs to the genes being controlled in the clusters and a relationship between the two cannot be determined. We answer this issue by examining the link between genes and TFBSs and create a network depicting associations between clusters, genes, and TFBSs (Figure 16).

In Figure 16, genes were presented as green circle nodes, while TFBSs are represented as brown diamonds and triangles as described for Figure 15. The inner wheel of Figure 16 is the same as that of Figure 15, the difference being that in Figure 16 each red gene cluster node has been expanded to include genes. Here, expanded clusters are represented by magenta nodes, which display the regulatory potential of TFBS amongst clustered genes. Topology of these magenta cluster nodes is such that each cluster node is associated with genes that are then associated to TFBSs. Due to complexity of the network, as in Figure 16, edges involving TFBS that appear in multiple clusters are not shown.

Two classes of genes and two classes of TFBSs are displayed within Figure 16:

a/ Genes are grouped into two classes, those that are annotated by TFBSs (promiscuous genes), and those that are not annotated with TFBSs (singleton genes). This happens because in the process of selecting dominant TFBSs, many TFBS predictions that were originally mapped to promoters of the genes now appear as singletons are eliminated.

b/ TFBSs are grouped into two classes, those that are mapping to only one promoter (singleton TFBSs) and those that are mapping to more than one promoter within a cluster (promiscuous TFBSs). Note that TFBSs that are not cluster-specific (on the inner wheel of Figure 16) are not included in this class. This simplifies the graphs considerably and still retains those TFBSs that are dominant to specific clusters.



Figure 16: Distribution of TFBS across genes and clusters



Since co-regulated genes are thought to share similar promoter content, it is reasonable to assume that the most promiscuous TFs within the cluster are also the most relevant regarding transcription regulation for that cluster. This assumption relies on the fact that each cluster contains sets of co-expressing genes and that TFBSs common/promiscuous to most of the genes within a cluster are likely to control their gene expression. This data is presented in Supplementary Data File 14. Supplementary data file contains data on each cluster and breaks down those genes and TFBSs classified as ‘promiscuous’ or ‘singleton’. Additionally included at the end of the file, is a list of all genes and TFBSs applicable to the study.

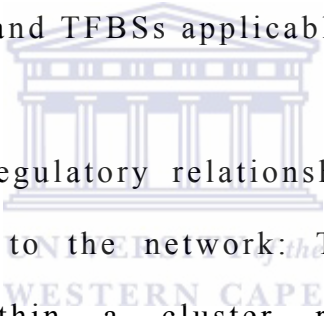


Figure 16 displays the regulatory relationships that have prompted two hypotheses relating to the network: TFs that bind the most promiscuous TFBSs within a cluster represent the core of transcriptional control machinery for that cluster; distributions of TFBS can be used to predict which genes should ‘belong’ to which cluster. The second hypothesis is based on the premise that genes belonging to the cluster will contain similar sets of TFBSs to confer similar expression behavior. This behavior of TFBSs has already been observed in Figure 13, Chapter 3 and Supplementary Figure 1, which shows that TFBSs are able to cluster genes similarly to expression data.

We inspected the ability of TFBS annotation to predict gene placement among the clusters by comparing all genes we classed as

‘singleton’ (Supplementary Data File 14) to those clustered genes that did not fit clearly in any one of the 12 anatomically restricted gene expression clusters. These ambiguously clustered genes are those that display similar expression profiles, but not sufficiently so, to the other genes in a cluster. Thus these genes have shown the presence or absence of gene expression in CA3 regions that are not exactly mirrored by any other CA3 cluster. The comparison found that all partially clustered genes (see Supplementary Table 5) are classed as ‘singleton’ genes (Figure 16) and have not been annotated by dominant TFBSs. Such a result is intriguing because it suggests that the method used to predict TFBSs could help to improve the accuracy of gene expression clustering methods.

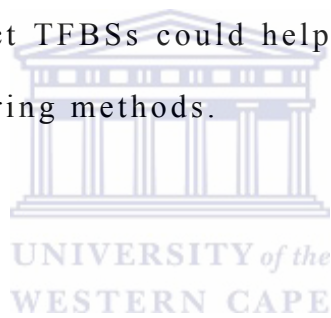


Figure 17:A) Union of TFBS networks for CA3 (blue) and CA2 (green) regions in adult mouse brain. B) Basic hippocampal anatomy



Thus far, all analyses have involved genes expressing differentially across the CA3 region of the hippocampus. Whilst such an analysis has implicated likely key regulatory networks within CA3 sub-anatomy, it cannot be said that these TFBSs represent the CA3 region as a whole. Thus, the nearest neighboring region populated by neural cells, CA2, was analyzed using the same methods as reported for the CA3 analysis in Chapter 3 and compared to CA3 results. The comparison may be viewed in Figure 17. This analysis was important for two reasons:

a/ First, the comparison identified a set of TFBSs common to promoters of genes expressed in both CA2 and CA3 anatomy. A recurring hypothesis here is that the dentate gyrus maintains a population of neural precursor cells and thus it is not unreasonable to propose that common TFBSs between CA2 and CA3 are involved in maintaining anatomical boundaries between the two regions in the adult mouse.

b/ Second, those TFBSs that remain unique to one of the twelve clusters of CA3 after the union between CA2 and CA3 show greater specificity towards CA3. Thus, this comparison identified regulatory networks that are, in the context of Ammon's horn, specific for regionalized expression within CA3. Interestingly, those TFBSs filtered out by this comparison are also those TFBSs that have low ranking in Table 6, suggesting that high frequency of TFBSs in clusters is a good measure for identifying key, cluster specific, TFBSs. Figure 18 shows a low frequency of TFBS to gene associations for those TFBSs filtered by the CA2-CA3 union.

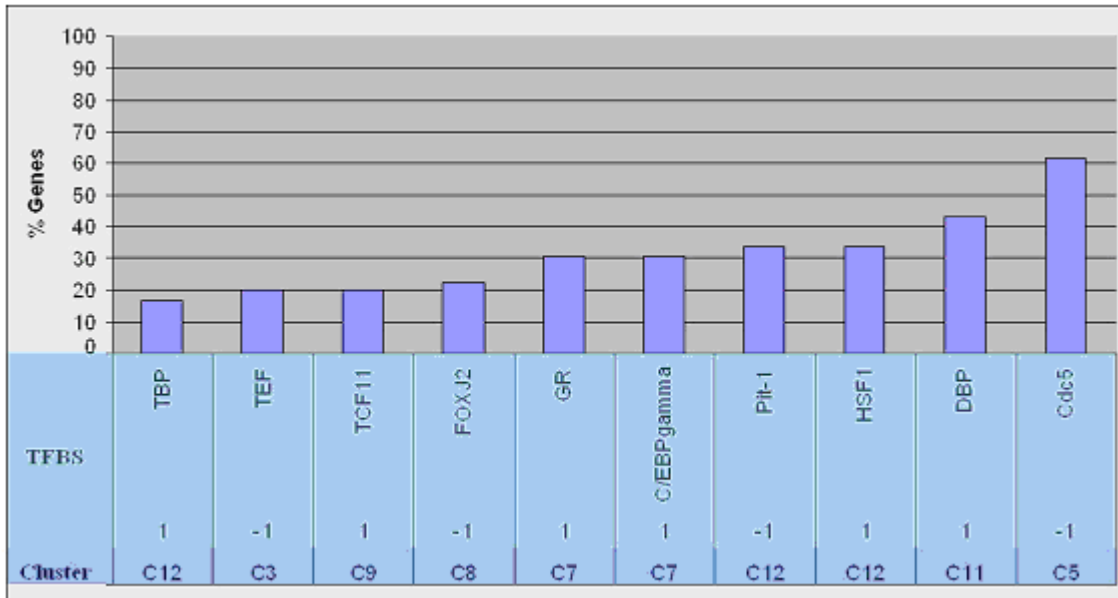


Figure 18: Cluster specific TFBSs filtered by the CA2-CA3 comparison and the percentage of clustered genes they were found in



Chapter 5: Graph EZ software tool

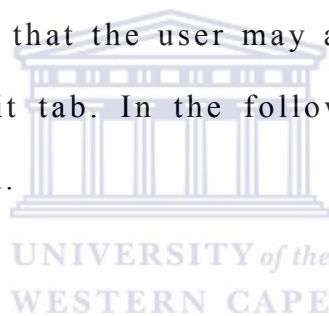
Graph EZ as a tool for creating, maintaining, and analyzing network-based projects using an HTML interface

Modern bioinformatics has seen the development of many data visualization tools. With the inclusion of meta-nodes (nodes used to group other nodes into sub-networks), networks are becoming both an intuitive and analytical method for representing relationships in data. Prior to the use of meta-nodes, networks have been cumbersome to work with for visualization purposes and have been mostly used for identifying key nodes, shortest paths, and various other informative statistics on data.

‘Graph EZ’ has culminated from the combination of Python (www.python.org) scripts and tools required to create the network visualizations presented in this thesis. It is a tool that provides a platform to host data relationships in the form of a network along with applicable annotations, analytical results, and the algorithms for performing many common network analyses such as shortest path algorithms. A single network and all associated data are stored in the form of a project acting as both a summary of prior analyses and as a platform for any new analyses. Graph EZ makes use of a ‘best of both’ philosophy providing functionality to organize networks into

user specified clusters using meta-nodes, whilst remembering the original network topology for analytical purposes. Thus, a user can view a more ‘human readable’ network whilst still being able to analyze it with well-established algorithms.

Graph EZ is a web-based tool, operating on Python and HTML (CSS) code through a Python common gateway interface (CGI). Data is stored in a MySQL database on the host server but may also be accessed through a human readable flat-file format (for example see Supplementary Data File 15). Every project created on the server contains three major tabs that the user may access: the View tab, the Analyze tab, and the Edit tab. In the following sections each tab’s function will be discussed.



Creating a Project

When accessing the home page on the server, a user is presented with two options, to select an available project or to create a new project.

The creation page (Figure 19) requires the input of a title, project description, and a list of edges as the primary index for relationships within the network. Once created, more advanced features for the project are accessible through the Edit tab. Note that once a project has been created the user is redirected to the home page where the project may be accessed.

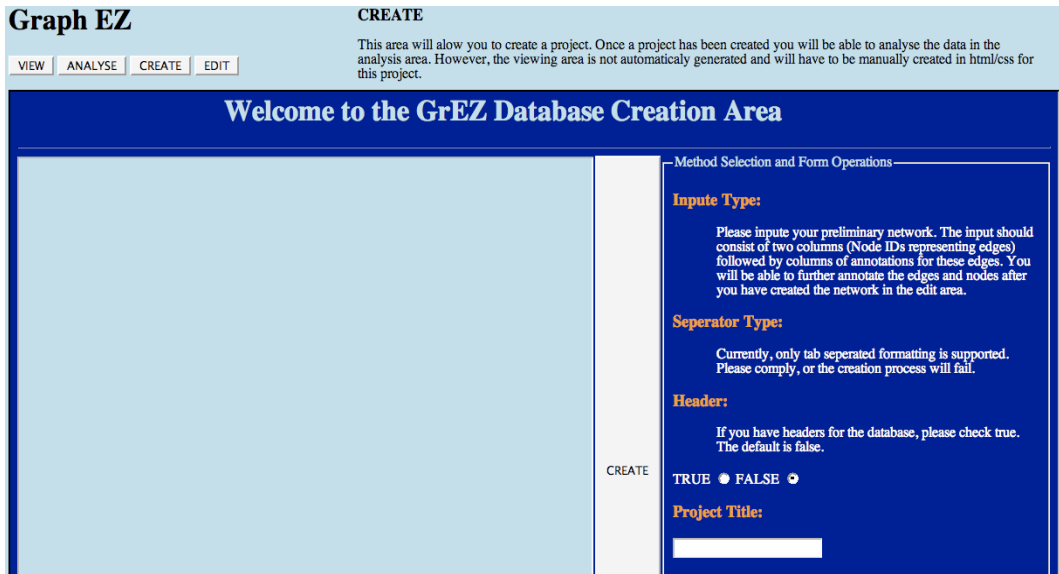


Figure 19: Screenshot of the create project page

The Edit Project Tab

The Edit project tab is accessible once a user has selected a project after clicking on the “select projects” option from the home page. The Edit tab acts as an account for the project and is separated into two sections. The first section allows the user to edit the core network relationships by deleting, adding, or annotating edges. The second section allows the user to edit what will be seen from the View tab (discussed later under View Project).

Editing the core relationships of the network can only be done by deleting, adding, or annotating an edge, whilst nodes may only be annotated. The reason for this is that the presence or absence of a node is dependant on the edges. When annotating nodes, two types of annotations exist: informative and categorical annotations. Informative annotations are descriptive and include alternative Identifiers for nodes. Categorical annotations are used to classify

nodes into different groups or clusters. The categorical annotations are used by native Graph EZ algorithms (rather than 3rd-party plug-in algorithms) that re-organize data into more human readable networks by making use of meta-nodes, for example the network can be arranged to visualize networks in terms of gene expression clusters.

The View Project Tab

The view project tab is the first tab a user will see after selecting the project and will be blank until the user has uploaded content to it through the Edit Project tab. The View Project tab serves to summarize and explore the data. A user is able to upload graphics (usually network figures in 'gif' format) to the project via the Edit tab and make the image inter-actable by selecting regions on the image to annotate with text and more image data. On the view page, these annotations will appear to the user as popup boxes once the mouse has rolled over the required region on the figure (see Figure 20).

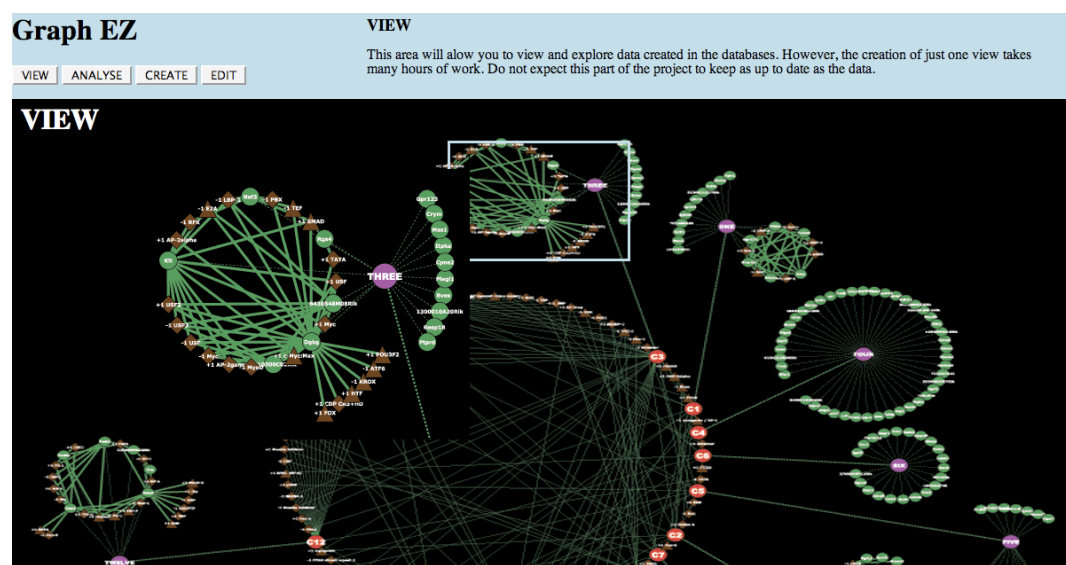
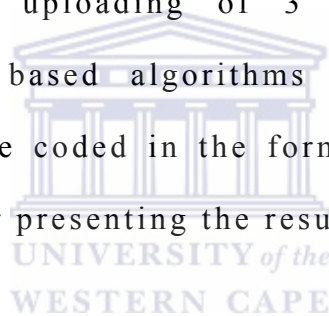


Figure 20: Screenshot of the View tab

The Analyze Project Tab

When analyzing the project, it is important to remember that a user can only analyze the core (without meta-nodes) network. From this tab the re-organized networks with meta-nodes may be generated and downloaded into formats applicable to many network visualization tools such as Cytoscape (Shannon et al. 2003).

This tab hosts algorithms implemented within Graph EZ, but also, provides a section for uploading of 3rd-party plug-ins by the community. Community based algorithms may only act on core network data and must be coded in the form of a CGI script, along with the required code for presenting the results back to the user.



Discussion

Graph EZ is a tool designed to exhibit data that can be described in the form of a network. Whilst it contains some functionality for analytical purposes, its primary function is to effectively describe projects of data both logically and intuitively. It is a tool that has been designed around the goal to facilitate understanding of results between research collaborators working on the same projects or third parties whom are merely interested in the project. Graph EZ was originally developed using Python implementing Python Card (<http://pythoncard.sourceforge.net/>) for the graphical user interface (GUI). This version of Graph EZ supported only the algorithms necessary to

create and analyze networks. Since this version, Graph EZ has begun to be implemented in HTML and Python CGI to make it more accessible. This web-based version, as described in detail previously, is currently still under development at the time this report was written.



Chapter 6: Discussion and concluding remarks

The wealth of data generated by the ABA project is enormous and required 600 terabytes of data to map high-resolution gene expression data down to individual brain sections. This study has made use of the expression data in this resource to dissect, from the viewpoint of transcription regulation, the CA3 anatomy of the adult mouse hippocampus and to investigate possible regulatory networks responsible for correct gene expression.

To discern those genes whose expression most likely contributes to the correct functioning of the hippocampus, 2686 genes were clustered using two distinct and unbiased computational methodologies. Non-negative matrix factorization yielded broad anatomic clusters that corresponded well with classically defined neuro-anatomy distinguishing between the dentate gyrus and Ammon's horn. Yet, this methodology was unable to determine high-resolution sub-anatomy presumably due to background interference from genes expressing in non-neuronal cell populations. Using a hierarchical clustering approach the CA3 portion of Ammon's horn was successfully partitioned into 9 sub-regions generally along the septo-temporal axis. As a result, 155 genes were selected to characterize the CA3 region.

The CA3 region is observed to express 155 genes differentially across the anatomy in an axis dependant manner. Moreover, the expression seems to be tightly regulated and regionalized to CA3 sub-anatomy. Observed with the aid of computer graphics the nine regions, subdividing CA3, appear to be banded diagonally along the septo-temporal axis. The sub-regions of CA3 show discrete borders of gene expression between all regions. Thus, it seems that genes tend to either express or not express in differing combinations of these sub-regions. These differential gene expression patterns have not previously been observed for this many genes in the CA3 region. Although, many previous studies have indeed identified much evidence to suggest that the CA3 anatomy is organized septo-temporally. The CA3 region displays anatomical differentiation, and afferent and efferent innervation occurs differentially along the septo-temporal axis. Furthermore, afferent neurons make differential use of neuro-transmitters in compliance to the septo-temporal theme. Additionally, neurotransmitter chemo-receptors types within the CA3 region parallel the differential afferent neurons according to which neuro-transmitters are used.

Thus, it seems that whilst genes express differentially along the septo-temporal axis, so too is the anatomy of CA3 arranged indicating that CA3 functionality is also spatially organized along the septo-temporal axis. To investigate the potential of functional differences along this axis, gene symbols were queried against resources such as the Gene Ontology. Results indicate three major functional themes of

genes with varying expression along the septo-temporal axis. These themes are cellular communication, regulation, and physiology. All three themes displayed similar profiles of gene expression and all themes contain genes only expressing at either pole as well as genes expressing throughout the CA3 region. Twenty-two genes coding for ion channel related functions demonstrated physiological differentiation whilst 53 genes coding for cellular adhesion products expressed in similar patterns. Initially it seemed odd that cell adhesion genes would be upregulated within the adult. These gene's products along with axon guidance signaling molecules are usually associated with the developing brain anatomy and function in correct axonal path finding. Since much of the adult brain is geared towards inhibiting neurogenesis and instead focusing on neuronal cell maintenance, it seems unnecessary to express genes coding for cell adhesion molecules. An exception is the case of the CA3, where not only do we find the expression of cell adhesion genes, but also, they are tightly regulated in a regionalized manner. This might best be explained by the presence of a persistent and active population of neural progenitors in the dentate gyrus. Newly matured cells might require the regionalized expression patterns observed in cell adhesion genes in the CA3 for correct axonal path finding and CA3 innervation.

Of the three major functional themes identified using ontology's, that of regulation is perhaps the most intriguing. Here, 29 genes were found to code for TFs that had varying expression along the septo-temporal axis. These TFs mirror gene expression patterns found in

both the cell-cell signaling and physiology themes, suggesting that they are at least part of a set of key regulators responsible for observed CA3 gene expression. Yet, at this stage of the study there was no context for how these TFs regulated genes within CA3. The link between the TF regulator and the target gene was not yet made. To create this link, a predictive approach established by (Bajic et al. 2004) was made and related to twelve clusters of unambiguously expressing genes. These clusters displayed twelve expression profiles based on regionalized CA3 expression. Of the 155 genes expressing in CA3 the majority expressed in perfect correlation with one of the 12 clusters, whilst the minority displayed slight deviation of expression patterns. Such clustering shows much promise for biological significance relating to these clusters. Additionally, one might expect that highly overlapping expression patterns in gene clusters are putatively the result of similar transcriptional control elements such as TFs.

To determine the most likely transcriptional control elements, the first step required the extraction of a possible 610 TSSs representative of the 155 genes expressing in CA3. With these TSSs the promoter regions were defined as 1000bp upstream and 200bp downstream of each TSS. Each promoter was analyzed for the presence of motifs that possibly bound TFs. These TFBSs were determined using position weight matrix models from the literature based Transfac Professional database. Strict criteria ensured a minimum false positive predictive rate, yet large lists of predicted

TFBSs were still generated. To truncate these lists further and increase data quality, the predicted TFBSs were compared to the background data sets generated by random mouse DNA as well as background mouse promoter sets. Predictions that did not appear to be with overrepresentation index greater than three fold in the target versus background promoter sets were rejected. The subsequent TFBS list represented high quality predicted data that had been passed through multiple additional filters. Once the list of dominant TFBSs was ascertained, TFs that could bind TFBSs were determined by making use of the Transfac Professional database. Consequently, mapping these TFs back to the genes, yielded a transcriptional regulatory network specific for the expression of genes within the CA3 region. Additionally, reorganizing the network such that genes were grouped into twelve similarly expressing gene clusters identified regulatory sub-networks specific for each cluster. Because each cluster demonstrates a high degree of uniformity in gene expression, specific sub-networks of regulatory elements seem to suggest that those TFs identified by this approach may play key role in the regulation of regionalized CA3 gene expression.

Unfortunately, this predictive approach cannot pinpoint unique associations of the type TF→TFBS. This is because many TFs are known to bind many different TFBSs, while many TFBS can bind various TFs. This many-to-many relationship results in multiple combinations of TF→TFBS associations, many of which may not be applicable to the given study. Finally, TF nomenclature within the

Transfac Professional database is not always clear, thus data must be hand curated because it is not amenable to automatic processing. Most of these problems were resolved by integrating the Transfac database with publicly available databases by linking them over the Entrez gene identifiers.

Of the 178 TFBSs utilized for promoter annotation, 59% (105/178) appear associated with only one anatomically restricted gene cluster, while 83% (148/178) are associated with at most two clusters. This implies a potential specialization of identified TFBSs in control of genes of different anatomically restricted gene clusters. Further inspection of the distribution of TFBSs amongst the CA3 expressing genes resulted in the clustering of genes based on TFBS information. Surprisingly, a high degree of similarly expressing genes was grouped together. Comparisons between these clusters and the original 12 produced by gene expression data indicated that 105 of 155 genes were possible to re-clustered, a sensitivity of 67.47%. All of the genes associated to clusters have been assigned to the original 12 expression based clusters from which they originated, providing thus positive predictive value of 100%. This evidence suggests that the predicted TFBS list contains enough pertinent data to parallel the biology behind CA3 gene expression. Here, it should be acknowledged that the sensitivity might increase if the strict criteria, required to generate the TFBS list, was relaxed.

In contrast, whilst the genes analyzed were specific for CA3 expression, the same cannot be said for the predicted TFBSs. Indeed, genes expressing in the neighboring CA2 region might also contain similar distributions of TFBSs. If this were the case, there would be no grounds to claim that regionalized gene expression within CA3 is due to specific CA3 regulatory networks. Whilst analyzing CA2 gene promoters for TFBS distributions did reveal commonly predicted TFBS, the comparison of the two sets produced an insignificant number of common TFBSs. Moreover, of those common TFBSs, the majority were non-specific for any of the 12 expression clusters and those that were, were only found in a minority of gene's promoters in each cluster suggesting that they were largely irrelevant.

As a whole this study has set out to define (identify, prioritize, and reconstruct a network of TFs) a molecular model of normal gene expression and regulation in the adult mouse hippocampus with specific focus on the CA3 region of Ammon's horn. Because the hippocampus constitutes a highly ordered cellular architecture, it is particularly compliant to the approaches used in this study to identify discrete regions of gene expression and transcriptional control in its sub-anatomy. The hippocampus has shown to differentially express genes across many axes, and the CA3 region displays this most aptly. Several classes of genes express differentially across the regionalized CA3 anatomy demonstrating functional differentiation of ion channel and cell-adhesion related genes. Differential regulation of gene expression across regions of CA3 is putatively due to differential

action of TFs, and indeed, this is what is observed in CA3 anatomy. Reconstructed transcription regulatory networks define the transcription potential for adult CA3 expressing genes and display plausible regulatory sub-networks for 12 clusters of co-expressed genes. Interestingly, though somewhat expected, genes clustered based on their promoter content (distribution of TFBSs), group co-expressed genes together to a large extent.

Neurodegenerative diseases are socially crippling and often mortal. Understanding both the anatomical architecture and molecular processes of a normal brain is crucial to the kinds of studies required to combat brain disease. This study represents one of the first to merge anatomical architecture, expression profiles and transcription regulatory potential on such a large scale in hippocampal sub-anatomy. It would not be feasible without the gene expression atlas project conducted by the Allen Institute for Brain Science. Already this study has identified a set of TFs that may explain gradients of cell adhesion molecules observed across the CA3 area. If indeed these are required for correct axonal path finding by dentate gyrus neural progenitors then they would make good candidates for studies into combating neurodegenerative disease and injuries through neural and stem cell transplants. Additionally, identified TFs appearing to control discrete groups of co-expressing genes are ideally suited for knock-out mice studies.

Finally, the study has made an inventory of TFs that potentially control genes in CA3 hippocampal region and whose combination of specific TFs in promoters of genes could be responsible for anatomically restricted gene expression. The results of this study suggest that it is a combinatorial effect of TFs that are providing specificity of gene expression in the adult CA3 region. Although we are able to find, or prioritize, certain TFs hinting to being unique to certain gene clusters, it is, however, highly unlikely that these TFs act in a singular capacity and more likely that they represent a core part of the transcriptional regulatory machinery responsible for the anatomically restricted gene expression patterns observed in the CA3 anatomy of the adult mouse brain.

In the future we would like to identify the biological properties of the identified TFBS for all clusters by: a/ clearly identifying modules of TFs based on their composition, arrangement and potential protein-protein interaction, b/ a comparison of said modules with the background distributions.

References

Ashburner, M., Ball, C.S., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25(1): 25-9.

Ashton, D., Van Reempts, J., Haseldonckx, M., and Willems, R. (1989). Dorsal-ventral gradient in vulnerability of CA1 hippocampus to ischemia: a combined histological and electrophysiological study. *Brain Research* 487(2): 368-72.

Bajic, V., Hofmann, O., MacPherson, C., Kaur, M., and Hide, W. (2006A). Progress report: Analysis of promoter properties for the clusters in the hippocampus CA3 region in mouse. South African National Bioinformatics Institute, internal document (08 July).

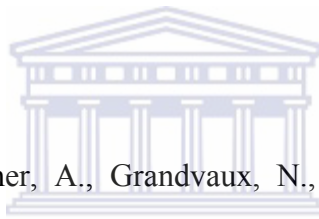
Bajic, V., MacPherson, C., Schmeier, S., Hofmann, O., Kaur, M., and Hide, W. (2006B). Progress report No.2: Analysis of promoter properties for the clusters in the hippocampus CA3 region in mouse. South African National Bioinformatics Institute, internal document (12 November).

Bajic, V., Choudhary, V., and Hock, C. (2004). Content analysis of the core promoter region of human genes. *In Silico Biology* 4:109-125.

Bajic, V., Tan, S., Christoffels, A., Schonback, C., Lipovich, L., and Yang, L.. (2006). Mice and men: their promoter properties. *PLoS Genetics* 2(4): e54.

Bannerman, D.M., Deacon, R.M.J., Offen, S., Friswell, J., Grubb, M., and Rawlins, J.N.P. (2002). Double dissociation of function within the hippocampus: spatial memory and hyponeophagia. *Behavioral Neuroscience* 116(5): 884-901.

Bannerman, D.M., Rawlins, J.N.P., McHugh, S.B., Deacon, R.M.J., Yee, B.K., Bast, T., Zhang, W-N., Pothuizen, H.H.J., and Feldon, J.. (2004). Regional dissociations within the hippocampus--memory and anxiety. *Neuroscience and Biobehavioral Reviews* 28(3): 273-83.



Bannerwarth, S., Laine, S., Daher, A., Grandvaux, N., Clerzius, G., Leblanc, A., and Gagnon, A.. (2006). Cell-specific regulation of TRBP1 promoter by NF-Y transcription factor lymphocytes and astrocytes. *Journal of Molecular Biology* 355(5): 898-910.

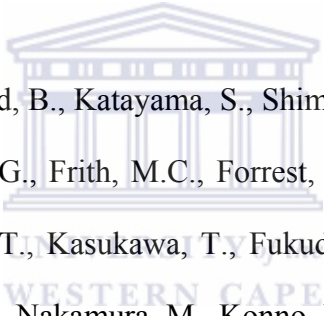
Bergen, A.W., Baccarelli, A., McDaniel, T.K., Kuhn, K., Pfeiffer, J., Bender, P., Jacobs, K., Packer, B., Chanock, S.J. and Yeager, M. (2007). Cis sequence effects on gene expression. *BMC Genomics* 8:296.

Bernat, J., Crawford, G., Ogurtsov, A., Collins, F., Ginsburg, D., Kondrashov, A., Kondrashov, A.. (2006). Distant conserved sequences flanking endothelil-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. *Human Molecular Genetics* 15(13): 2098-2105.

Bragdon, A.C., Taylor, D.M., and Wilson, W.A. (1986). Potassium-induced epileptiform activity in area CA3 varies markedly along the septotemporal axis of the rat hippocampus. *Brain Research* 378(1): 169-73.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple, B.P., de Bono, B., Della, G.G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojobori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasaki, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan, B.M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schonbach, C., Sekiguchi, K., Semple, C.A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoya, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., van Nimwegen, E., Verardo, R., Wei, C.L., Yagi, K.,

Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusica, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., and Hayashizaki, Y.; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). (2005). The transcriptional landscape of the mammalian genome. *Science* 309(5740): 1559-63.



Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., Forrest, A.R., Alkema, W.B., Tan, S.L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S.M., Wells, C.A., Orlando, V., Wahlestedt, C., Liu, E.T., Harbers, M., Kawai, J., Bajic, V.B., Hume, D.A., and Hayashizaki, Y.. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* 38(6): 626-35.

Cayuso, J., Ulloa, F., Cox, B., Briscoe, J., and Martí, E. (2006). The Sonic hedgehog pathway independently controls the patterning, proliferation and survival of neuroepithelial cells by regulating Gli activity. *Development* 133(3): 517-28.

Chong, A., Zhang, Z., Choi, K.P., Choudhary, V., Djamgoz, M.B.A., Zhang, G., Bajic, V.B. (2007). Promoter profiling and coexpression data analysis identifies 24 novel genes that are coregulated with AMPA receptor genes, GRIAs. *Genomics* 89(3): 378-84.

Clarke, G., Collins, R.A., Leavitt, B.R., Andrews, D.F., Hayden, M.R., Lumsden, C.J., McInnes, R.R. (2000). A one-hit model of cell death in inherited neuronal degenerations. *Nature* 406(6792): 195-9.

Colombo, M., Fernandez, T., Nakamura, K., and Gross, C.G.. (1998). Functional differentiation along the anterior-posterior axis of the hippocampus in monkeys. *Journal of Neurophysiology* 80(2): 1002-5.

Creuzet, S.E., Martinez, S. and Le Douarin, N.M. (2006). The cephalic neural crest exerts a critical effect on forebrain and midbrain development. *Proceedings of the National Academy of Sciences of the United States of America* 103(38): 14033-14038.

Cuenca, A.A., Schetter, A., Aceto, D., Kempfues, K. and Seydoux, G. (2003). Polarization of the *C.elegans* zygote proceeds via distinct establishment and maintenance phases. *Development* 130(7): 1255-1265.

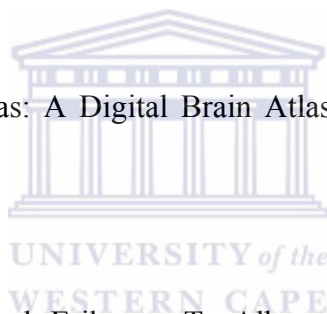
Datson N.A., Meijer, L, Steenbergen, P.J., Morsink, M.C., van der Laan, S., Meijer, O.C., and de Kloet, E.R.. (2004). Expression profiling in laser-microdissected hippocampal subregions in rat brain reveals large subregion-specific differences in expression. *The European journal of neuroscience* 20(10): 2541-54.

de Wit, J., and Verhaagen, J. (2003). Role of semaphorins in the adult nervous system. *Progress in Neurobiology* 71(2-3): 249-67.

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A.. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology* 4(5): P3.

Dolorfo, C.L., and Amaral, D.G.. (1998). Entorhinal cortex of the rat: topographic organization of the cells of origin of the perforant path projection to the dentate gyrus. *The Journal of Comparative Neurology* 398(1): 25-48.

Dong, H. (2007). The Allen Atlas: A Digital Brain Atlas of C57BL/6J Male Mouse. (in press) John Wiley & Sons.



Eriksson, P.S., Perfilieva, E., Bjork-Eriksson, T., Alborn, A-M., Nordborg, C., Peterson, D.A., and Gage, F.H.. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine* 4(11): 1313-1317.

Farlow, M., Murrell, J., Ghetti, B., Unverzagt, F., Zeldenrust, S., and Benson, M. (1994). Clinical characteristics in a kindred with early-onset Alzheimer's disease and their linkage to a G-->T change at position 2149 of the amyloid precursor protein gene. *Neurology* 44(1): 105-11.

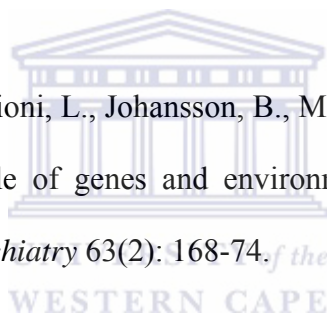
Ferri, G., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., Hasegawa, K., Hendrie, H., Huange, Y., Jorm, A., Mathers, C., Menezes, P., Rimmer, E.,

and Scazufca, M. (2005). Global prevalence of dementia: a Delphi consensus study. *The Lancet* 366(9503): 2112-2117.

Gage, F.H., and Thompson, R.G. (1980). Differential distribution of norepinephrine and serotonin along the dorsal-ventral axis of the hippocampal formation. *Brain Research Bulletin* 5(6): 771-3.

Galceran, J., Miyashita-Lin, E., Devaney, E., Rubenstein, J., and Grosschedl, R. (2000). Hippocampus development and generation of dentate gyrus granule cells is regulated by LEF1. *Development* 127(3): 469-482.

Gatz, M., Reynolds, C.A., Fratiglioni, L., Johansson, B., Mortimer, J.A., Berg, S., Fiska, A., and Pedersen, N.L.. (2006). Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry* 63(2): 168-74.



Goate, A., Chartier-Harlin, M.C., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., and Haynes, A., Irving, N., and James, L.. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* 349(6311): 704-6.

Groenewald, P., Bradshaw, D., Nojilana, B., Bourne, D., Nixon, J., Mahomed, H., Daniels, J. (2003). Cape Town Mortality, 2001, Part III, Cause of death profiles for each sub-district. Medical Research Council, Cape Town, South Africa.

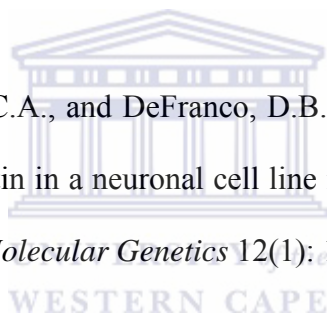
Hampel, H., Bürger, K., Pruessner, J.C., Zinkowski, R., DeBernardis, J., Kerkman, D., Leinsinger, G., Evans, A.C., Davies, P., Moller, H-J., and Teipel, S.J. (2005). Correlation of

cerebrospinal fluid levels of tau protein phosphorylated at threonine 231 with rates of hippocampal atrophy in Alzheimer disease. *Archives of Neurology* 62(5): 770-3.

Ishizuka, N., Weber, J., and Amaral, D.G. 1990. Organization of intrahippocampal projections originating from CA3 pyramidal cells in the rat. *The Journal of Comparative Neurology* 295(4): 580-623.

Izaki, Y., Takita, M., and Nomura, M.. (2000). Comparative induction of long-term depression between dorsal and ventral hippocampal CA1 in the anesthetized rat. *Neuroscience Letters* 294(3): 171-4.

Jiang, H., Nucifora, F.C., Ross, C.A., and DeFranco, D.B.. (2003). Cell death triggered by polyglutamine-expanded huntingtin in a neuronal cell line is associated with degradation of CREB-binding protein. *Human Molecular Genetics* 12(1): 1-12.



Josephson, R., Müller, T., Pickel, J., Okabe, S., Reynolds, K., Turner, P.A., Zimmer, A., and McKay, R.D. (1998). POU transcription factors control expression of CNS stem cell-specific genes. *Development* 125(16): 3087-100.

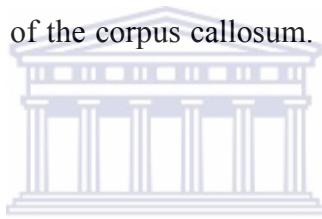
Jung, M.W., Wiener, S.I., and McNaughton, B.L.. (1994). Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. *The Journal of Neuroscience* 14, no. 12, 7347-56.

Kadonaga, J. (2004). Regulation of RNA Polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116:247-257.

Kaiser, K., and Meisterernst, M. (1996). The human general co-factors. *Trends Biochemical Science* 21(9): 342-345.

Karlinsky, H., Vaula, G., Haines, J.L., Ridgley, J., Bergeron, C., Mortilla, M., Tupler, R.G., Percy, M.E., Robitaille, Y., and Noldy, N.E. (1992). Molecular and prospective phenotypic characterization of a pedigree with familial Alzheimer's disease and a missense mutation in codon 717 of the beta-amyloid precursor protein gene. *Neurology* 42(8): 1445-53.

Keeble, T.R., Halford, M.M., Seaman, C., Kee, N., Macheda, M., Anderson, R.B., Stacker, S.A., and Cooper, H.M. (2006). The Wnt receptor Ryk is required for Wnt5a-mediated axon guidance on the contralateral side of the corpus callosum. *The Journal of Neuroscience* 26, no. 21(24): 5840-8.



Kel, A.E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research* 31(13): 3576-9.

Khachaturian, Z.S. (1985). Diagnosis of Alzheimer's disease. *Archives of Neurology* 42(11): 1097-105.

Köhler, C., Schultzberg, M., and Radesäter, A.C. (1987). Distribution of neuropeptide Y receptors in the rat hippocampal region. *Neuroscience Letters* 75(2): 141-6.

Kuhn, H., Dickinson-Anson, H., and Gage, F. (1996). Neurogenesis in the dentate gyrus of the adult rat: age-related decrease of neuronal progenitor proliferation. *Journal of Neuroscience* 16(6): 2027-2033.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755): 788-91.

Lein, E.S., Callaway, E.M., Albright, T.D., and Gage, F.H. (2005). Redefining the boundaries of the hippocampal CA2 subfield in the mouse using gene expression and 3-dimensional reconstruction. *The Journal of Comparative Neurology* 485(1): 1-10.

Lein E.S., Hawrylycz, M.J, Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., Chen, L., Chen, L., Chen, T-M., Chin, M.C., Chong, J., Crook, B.E., Czaplinska, A., Dang, C.N., Datta, S., Dee, N.R., Desaki, A.L., Desta, T., Diep, E., Dolbeare, T.A., Donelan, M.J., Dong, H-W., Dougherty, J.G., Duncan, B.J., Ebbert, A.J., Eichele, G., Estin, L.K., Faber, C., Facer, B.A., Fields, R., Fischer, S.R., Fliss, T.P., Frensley, C., Gates, S.N., Glattfelder, K.J., Halverson, K.R., Hart, M.R., Hohmann, J.G., Howell, M.P., Jeung, D.P., Johnson, R.A., Karr, P.T., Kawal, R., Kidney, J.M., Knapik, R.H., Kuan, C.L., Lake, J.H., Laramée, A.R., Larsen, K.D., Lau, C., Lemon, T.A., Liang, A.J., Liu, Y., Luong, L.T., Michaels, J., Morgan, J.J., Morgan, R.J., Mortrud, M.T., Mosqueda, N.F., Ng, L.L., Ng, R., Orta, G.J., Overly, C.C., Pak, T.H., Parry, S.E., Pathak, S.D., Pearson, O.C., Puchalski, R.B., Riley, Z.L., Rockett, H.R., Rowland, S.A., Royall, J.J., Ruiz, M.J., Sarno, N.R., Schaffnit, K., Shapovalova, N.V., Svisay, T., Slaughterbeck, C.R., Smith, S.C., Smith, K.A., Smith, B.I., Sodt, A.J., Stewart, N.N., Stumpff, K.R., Sunkin, S.M., Sutram, M., Tam, A., Teemer, C.D., Thaller, C., Thompson, C.L., Varnam, L.R., Visel, A., Whitlock, R.M., Wohnoutka, P.E., Wolkey, C.K., Wong, V.Y., Wood, M., Yaylaoglu, M.B., Young, R.C., Youngstrom, B.L., Yuan, X.F., Zhang, B.,

Zwingman, T.A., Jones, A.R. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445(7124): 168-76.

Lein, E.S., Zhao, X., and Gage, F.H. (2004). Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput in situ hybridization. *The Journal of Neuroscience* 24(15): 3879-89.

Li, Y.J., Oliveira, S.A., Xu, P., Martin, E.R., Stenger, J.E., Scherzer, C.R., Hauser, M.A., Scott, W.K., Small, G.W., Nance, M.A., Watts, R.L., Hubble, J.P., Koller, W.C., Pahwa, R., Stern, M.B., Hiner, B.C., Jankovic, J., Goetz, C.G., Mastaglia, F., Middleton, L.T., Roses, A.D., Saunders, A.M., Schmechel, D.E., Gullans, S.R., Haines, J.L., Gilbert, J.R., Vance, J.M., Pericak-Vance, M.A., Hulette, C., Welsh-Bohmer, K.A. (2003). Glutathione S-transferase omega-1 modifies age-at-onset of Alzheimer disease and Parkinson disease. *Human Molecular Genetics* 12(24): 3259-67.

Li-Weber, M., Laur, O., Davydov, I., Hu, C., Salgame, P., and Krammer, P.H.. (1997). What controls tissue-specific expression of the IL-4 gene? *Immunobiology* 198(1-3): 170-8.

Lin, C-Y., Vega V.B., Thomsen J.S., Zhang, T., Kong, S.L., Xie, M., Chiu, K.P., Lipovich, L., Barnett, D.H., Stossi, F., Yeo, A., George, J., Kutznetsov, V.A., Lee, Y.K., Charn, T.H., Palanisamy, N., Miller, L.D., Cheung, E., Katzenellenbogen, B.S., Ruan, Y., Bourque, G., Wei, C-L. and Liu, E.T. (2007). Whole-genome cartography of estrogen receptor α binding sites. *PLoS Genetics* 3(6):e87.

Mata, J., Wilbrey, A. and Bahler, J. (2007). Transcriptional regulatory network for sexual differentiation in fission yeast. *Genome Biology* 8:217.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research* 34(1): D108-10.

Meissner, B., Westner, I.M., Kallenberg, K., Krasnianski, A., Bartl, M., Vargas, D., Bosenberg, C., Kretzschmar, H.A., Knauth, M., Schulz-Schaeffer, W.J., and Zerr, I. (2005). Sporadic Creutzfeldt-Jakob disease: clinical and diagnostic characteristics of the rare VV1 type. *Neurology* 65(10): 1544-50.

Michele, D.E., Barresi, R., Kanagawa, M., Saito, F., Cohn, R.D., Satz, J.S., Dollar, J., Nishino, I., Kelley, R.I., Somer, H., Straub, V., Mathews, K.D., Moore, S.A., Campbell, K.P. (2002). Post-translational disruption of dystroglycan-ligand interactions in congenital muscular dystrophies. *Nature* 418(6896): 417-22.

Miquelajauregui, A., Van de Putte, T., Polyakov, A., Nityanandam, A., Boppana, S., Seuntjens, E., Karabinos, A., Higashi, Y., Huylebroeck, D., and Tarabykin, V. (2007). Smad-interacting protein-1 (Zfhx1b) acts upstream of Wnt signaling in the mouse hippocampus and controls its formation. *Proceedings of the National Academy of Sciences of the United States of America* 104(31): 12919-24.

Mosconi, L., Sorbi, S., Nacmias, B., De Cristofaro, M.T.R., Fayyaz, M., Cellini, E., Bagnoli, S., Bracco, L., Herholz, K., and Pupi, A. (2003). Brain metabolic differences between sporadic and familial Alzheimer's disease. *Neurology* 61(8): 1138-40.

Moser, M.B., and Moser, E.I. (1998). Functional differentiation in the hippocampus. *Hippocampus* 8(6): 608-19.

Motohashi, H., Katsuoka, F., Miyoshi, C., Uchimura, Y., Saitoh, H., Francastel, C., Engel, J.D. and Yamamoto, M. (2006). *Molecular and Cellular Biology* 26(12): 4652-4663.

Muentner, M.D., Forno, L.S., Hornykiewicz, O., Kish, S.J., Maraganore, D.M., Caselli, R.J., Okazaki, H., Howard, F.M., Snow, B.J., and Calne, D.B.. (1998). Hereditary form of parkinsonism--dementia. *Annals of Neurology* 43(6): 768-81.

Neumann, M., Sampathu, D.M., Kwong, L.K., Truax, A.C., Micsenyi, M.C., Chou, T.T., Bruce, J., Schuck, T., Grossman, M., Clark C.M., McCluskey, L.F., Miller, B.L., Masliah, E., Mackenzie, I.R., Feldman, H., Feiden, W., Kretschmar, H.A., Trojanowski, J.Q., and Lee, V.M-Y. (2006). Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* 314(5796): 130-3.

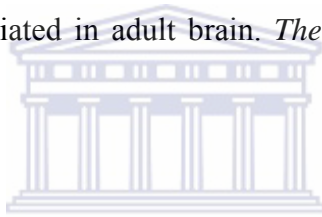
Ohtake, H., Limprasert, P., Fan, Y., Onodera, O., Kakita, A., Takahashi, H., Bonner, L.T., Tsuang, D.W., Murray, I.V.J., Lee, V,M-Y., Trojanowski, J.Q., Ishikawa, A., Idezuka, J., Murata, M., Toda, T., Bird, T.D., Leverenz, J.B., Tsuji, S., and La Spada, A.R. (2004). Beta-synuclein gene alterations in dementia with Lewy bodies. *Neurology* 63(5): 805-11.

Papathodoropoulos, C., and Kostopoulos, G. (2000). Decreased ability of rat temporal hippocampal CA1 region to produce long-term potentiation. *Neuroscience Letters* 279(3): 177-80.

Pazos, A., Cortés, R., and Palacios, J.M. (1985). Thyrotropin-releasing hormone receptor binding sites: autoradiographic distribution in the rat and guinea pig brain. *Journal of Neurochemistry* 45(5): 1448-63.

Petrovich, G.D., Canteras, N.S., and Swanson, L.W.. (2001). Combinatorial amygdalar inputs to hippocampal domains and hypothalamic behavior systems. *Brain Research Reviews* 38(1-2), 247-89.

Pinkstaff, J.K., Detterich, J., Lynch, G., and Gall, C. (1999). Integrin subunit gene expression is regionally differentiated in adult brain. *The Journal of Neuroscience* 19(5): 1541-56.



Racine, R., Rose, P.A., and Burnham, W.M.. (1977). Afterdischarge thresholds and kindling rates in dorsal and ventral hippocampus and dentate gyrus. *The Canadian journal of neurological sciences* 4(4): 273-8.

Reddy, P.H., Williams, M., Charles, V., Garrett, L., Pike-Buchanan, L., Whetsell, W.O., Miller, G., and Tagle, D.A. (1998). Behavioural abnormalities and selective neuronal loss in HD transgenic mice expressing mutated full-length HD cDNA. *Nature Genetics* 20(2): 198-202.

Reymann, S. and Borlak, J. (2006). Transcriptome profiling of human hepatocytes treated with Aroclor 1254 reveals transcription factor regulatory networks and clusters of regulated genes. *BMC Genomics* 7:217.

Risold, P.Y., and Swanson, L.W. (1997). Connections of the rat lateral septal complex. *Brain Research Reviews* 24(2-3): 115-95.

Rossi, G., Giaccone, G., Maletta, R., Morbin, M., Capobianco, R., Mangieri, M., Giovagnoli, A.R., Bizzi, A., Tom,aino, C., Perri, M., Di Natale, M., Tagliavini, F., Bugiani, O., and Bruni, A.C. (2004). A family with Alzheimer disease and strokes associated with A713T mutation of the APP gene. *Neurology* 63(5): 910-2.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerrière, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., Dubas, F., Frebourg, T., and Campion, D. (2006). APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genetics* 38(1): 24-6.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, E., Trush, V., and Quakenbush, J. (2003). TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* 34(2): 374-8.

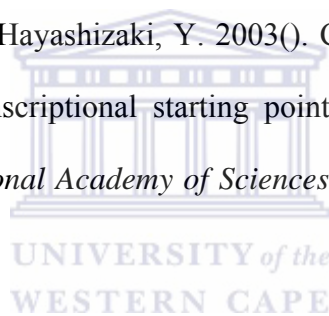
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Shwikowaski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13(11): 2498-504.

Sheth, U. and Parker, R. (2003). Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science* 300(5620): 805-808.

Shin, M.H., Lee, E-G., Lee, S-H., Lee, Y.S., and Son, H.. (2002). Neural cell adhesion molecule (NCAM) promotes the differentiation of hippocampal precursor cells to a neuronal lineage, especially to a glutamatergic neural cell type. *Experimental & Molecular Medicine* 34(6): 401-10.

Shingleton, A.W., Das, J., Vinicius, L. and Stern, D.L. (2005). The temporal requirements for insulin signaling during development in drosophila. *PLoS Biology* 3(9): e289.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., and Hayashizaki, Y. 2003(). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. . *Proceedings of the National Academy of Sciences of the United States of America* 100:15776-15781.



Small, S.A., Nava, A.S., Perera, G.M., DeLaPaz, R., Mayeux, R., and Stern, Y. (2001). Circuit mechanisms underlying memory encoding and retrieval in the long axis of the hippocampal formation. *Nature Neuroscience* 4(4): 442-449.

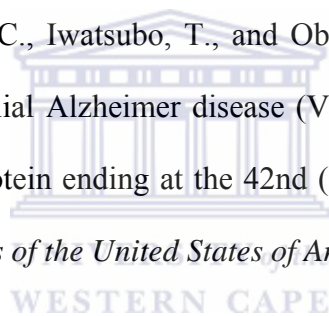
Sood, R., Zehnder, J.L., Druzin, M.L. and Brown, P.O. (2006). Gene expression patterns in human placenta. . *Proceedings of the National Academy of Sciences of the United States of America* 103(14): 5478-5483.

Teboul, L., Febbraio, M., Gaillard, D., Amri, E.Z., Silverstein, R., and Grimaldi, P.A. (2001). Structural and functional characterization of the mouse fatty acid translocase

promoter: activation during adipose differentiation. *The Biochemical Journal* 360(2): 305-12.

Thompson, C., MacPherson, C., Pathak, S., Lydia, N., Allison, C., Schmeier, S., Hofmann, O., Kaur, M., Riley, Z., Hide, W., Sunkin, S., Bernard, A., Puchalski, R., Gage, F., Hawrylycz, M., Jones, A., Bajic, V.B., and Lein, E. Molecular partitioning of the hippocampus into novel subdomains with implications for functional architecture and transcriptional regulation. Manuscript submitted to *Neuron*, 2007.

Tomita, T., Maruyama, K., Saido, T.C., Kume, H., Shinozaki, K., Tokuhira, S., Capell, A., Walter, J., Grunberg, J., Haass, C., Iwatsubo, T., and Obata, K. (1997). The presenilin 2 mutation (N141I) linked to familial Alzheimer disease (Volga German families) increases the secretion of amyloid beta protein ending at the 42nd (or 43rd) residue. *Proceedings of the National Academy of Sciences of the United States of America* 94(5): 2025-30.



van Groen, T., Miettinen, P., and Kadish, I. (2003). The entorhinal cortex of the mouse: organization of the projection to the hippocampal formation. *Hippocampus* 13(1): 133-49.

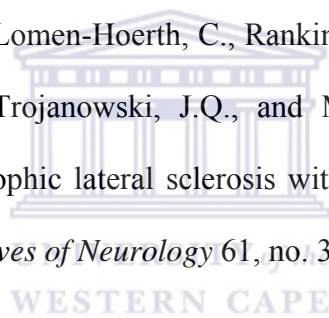
Verney, C., Baulac, M., Berger, B., Alvarez, C., Vigny, A., and Helle, K.B. (1985). Morphological evidence for a dopaminergic terminal field in the hippocampal formation of young and adult rat. *Neuroscience* 14(4): 1039-52.

Verwer, R.W., Meijer, R.J., Van Uum, H.F., and Witter, M.P. (1997). Collateral projections from the rat hippocampal formation to the lateral and medial prefrontal cortex. *Hippocampus* 7(4): 397-402.

Wakabayashi, K., Hayashi, S., Ishikawa, A., Hayashi, T., Okuizumi, K., Tanaka, H., Tsuji, S., and Takahashi, H. (1998). Autosomal dominant diffuse Lewy body disease. *Acta neuropathologica* 96(2): 207-10.

Wang, Y., Barbacioru, C., Hyland, F., Xiao, W., Hunkapiller, K.L., Blake, J., Chan, F., Gonzalez, C., Zhang, L. and Samaha, R.R. (2006). Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. *BMC Genomics* 7:59.

Wilhelmsen, K.C., Forman, M.S., Rosen, H.J., Alving, L.I., Goldman, J., Feiger, J., Lee, J.V., Segall, S.K., Kramer, J.H., Lomen-Hoerth, C., Rankin, K.P., Johnson, J., Feiler, H. S., Weiner, M.W., Lee, V.M-Y., Trojanowski, J.Q., and Miller, B.L. (2004). 17q-linked frontotemporal dementia-amyotrophic lateral sclerosis without tau mutations with tau and alpha-synuclein inclusions. *Archives of Neurology* 61, no. 3, 398-406.



Wimo, A., Winblad, B., and Jonsson, L. (2007). An estimate of total worldwide societal costs of dementia in 2005. *Alzheimer's & Dementia* 81-91.

Wodarz, A., and Nusse, R.. (1998). Mechanisms of Wnt signaling in development. *Annual review of cell and developmental biology* 14:59-88.

Yi, J-H., Park, S-W., Kapadia, R. and Vemuganti, R. (2007). Role of transcription factors mediating post-ischemic cerebral inflammation and brain damage. *Neurochem Int.* 50(7-8): 1014-1027.

Zhao, C., Teng, E.M., Summers, R.G., Ming, G-L., and Gage, F.H. (2006). Distinct morphological stages of dentate granule neuron maturation in the adult mouse hippocampus. *The Journal of Neuroscience*, 26(1): 3-11.

Zuberi, S.M. and Hanna, M.G. (2001). Current topic: Ion channels and neurology. *Arch. Dis. Child.* 84:277-280.



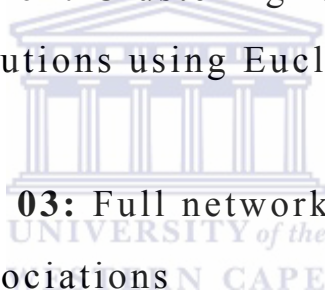
Appendix A: Supplementary Figures

Contents:

Supplementary Figure 01: K-means gene clusters based on TFBS distribution amongst CA3 expressing genes

Supplementary Figure 02: Clustering results of gene|TFBS distributions using Euclidean distance

Supplementary Figure 03: Full network of Cluster-to-Gene-to-TFBS associations



Appendix B: Supplementary Tables

Contents

Supplementary Table 01: List of significant TFBSs and their distribution across 12 CA3 clusters

Supplementary Table 02: Candidate transcription factor binding sites ordered from the most promising to the least

Supplementary Table 03: Description of all predicted TFBSs and their shape within network representations

Supplementary Table 04: Cluster Specific and Non-Specific TFBS

Supplementary Table 05: Overlap of Singleton Genes with those that demonstrate ambiguous clustering in the 12 CA3 clusters

Appendix C: Supplementary Data Files

Content:

- Supplementary Data File 01:** TFBS mappings to CA3 Cluster 1
- Supplementary Data File 02:** TFBS mappings to CA3 Cluster 2
- Supplementary Data File 03:** TFBS mappings to CA3 Cluster 3
- Supplementary Data File 04:** TFBS mappings to CA3 Cluster 4
- Supplementary Data File 05:** TFBS mappings to CA3 Cluster 5
- Supplementary Data File 06:** TFBS mappings to CA3 Cluster 6
- Supplementary Data File 07:** TFBS mappings to CA3 Cluster 7
- Supplementary Data File 08:** TFBS mappings to CA3 Cluster 8
- Supplementary Data File 09:** TFBS mappings to CA3 Cluster 9
- Supplementary Data File 10:** TFBS mappings to CA3 Cluster 10
- Supplementary Data File 11:** TFBS mappings to CA3 Cluster 11
- Supplementary Data File 12:** TFBS mappings to CA3 Cluster 12
- Supplementary Data File 13:** Example 'XGMML' code for cytoscape networks
- Supplementary Data File 14:** Breakdown of cluster specific genes and TFBSs presented in Figure 20
- Supplementary Data File 15:** Example of EZ Graph human readable flat file format