12-11-2015 12:00 AM

Genes and Gene Networks Related to Age-associated Learning Impairments

Raihan Uddin
The University of Western Ontario

Supervisor
Shiva M. Singh
The University of Western Ontario

Graduate Program in Biology
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Genes and Gene Networks Related to Age-associated Learning Impairments

(Thesis format: Integrated Article)

by

Raihan K. Uddin

Graduate Program in Biology

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

The School of Graduate and Postdoctoral Studies

The University of Western Ontario

London, Ontario, Canada

November, 2015

Abstract

The incidence of cognitive impairments, including normal age-associated spatial learning impairment (ASLI), has risen dramatically in past decades due to increasing human longevity. As such trends are expected to continue it has become imperative to better understand the underlying molecular biology and genetics of ASLI. In this study, data from a number of past gene expression microarray studies in rats are integrated and used to perform a meta- and network analysis aimed at identifying key ASLI genes and gene networks.

To ensure the generation of biologically relevant results, I first examine the importance of data selection and data preprocessing. This analysis shows that for effective downstream analysis to take place, both batch effects and outlier samples must be properly removed.

Next, using a set of selected datasets, I perform a meta-analysis and identify a number of significant differentially expressed genes across both age and ASLI in rats. Knowledge based gene network analysis shows that these genes affect many key functions and pathways in aged compared to young rats. These expression dependent functional changes might manifest as various neurodegenerative diseases/ disorders or to syndromic memory impairments at old age. Other aging related molecular changes might result in altered synaptic plasticity, thereby leading to normal, non-syndromic learning or spatial learning impairments such as ASLI.

Lastly, to overcome the limitations of traditional microarray data analysis, I employ a reverse-engineering mathematical modeling approach (called weighted gene co-expression network analysis or WGCNA) to identify key genes and their networks in ASLI. Using this approach I identify several reproducible network modules each highly significant with genes functioning in specific biological functional categories. It identifies a “learning and memory” specific module containing many potential key ASLI hub genes, some of which are also identified (but not prioritized) in the meta-analysis. Many of these candidate hub genes not

only show differential co-expression between young and aged networks, but are also reproducible in independent datasets. Functions of these ASLI hub genes link a different set of mechanisms to learning and memory formation, which meta-analysis was unable to detect.

Modern meta- and network approaches as implemented in this study can be applied to any large-scale dataset to identify potential key molecules and networks and thus generate new hypotheses. Future follow up research can help understand and pinpoint possible molecular mechanisms underlying complex behavioral traits such as cognitive impairments including ASLI.

Keywords

Aging, spatial learning, cognitive impairments, gene expression microarray, data integration, meta-analysis, inter-array correlation, outlier removal, batch effect, effect size, pathway analysis, co-expression networks, WGCNA, RDAVIDWebService, network module, hub gene.

Co-Authorship Statement

Chapter three of this thesis contains material from the manuscript entitled, “Hippocampal Gene Expression Meta-Analysis Identifies Aging and Age-Associated Spatial Learning Impairment (ASLI) Genes and Pathways” by Raihan K. Uddin and Shiva M. Singh, which was published in PLOS One in 2013. Here Raihan K. Uddin performed the experiments, analyzed the data, and wrote the manuscript. Dr. Singh provided manuscript editing, feedback, and supervision.

Acknowledgments

This thesis would not have been possible without the support of many over the years. First, I would like to thank the Department of Biology for giving me the opportunity to undertake this research as a part-time student. My sincere thanks to the department and the University for their financial support for the duration of my study. I also thank my colleagues in the department and current and past members of the Singh lab for their continuous encouragement, support, and well wishes.

I would like to thank my committee advisors Dr. Kathleen Hill, Dr. Graham Thomson, and Dr. Mark Daley for their valuable input and guidance during my Ph.D. I would like to especially thank Dr. Jim Karagiannis for reviewing this thesis and providing helpful comments and valuable input.

I consider myself truly fortunate and blessed to have Dr. Shiva Singh as my mentor who gave me endless support and complete freedom to cherish independent thinking and to achieve what I have today.

My wholehearted thanks go to the friends and families in London for their overwhelming support during many difficult times.

Finally, I dedicate this work to my wife Shahnaz and daughters Nafisha and Ridika. Without their unending love, support, and patience none of these would have been possible.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AKT	protein kinase B
ALL5_COMMON	probe sets common among all five studies
AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ARACNE	algorithm for the reconstruction of accurate cellular networks
ASLI	age-associated spatial learning impairment
AY	aged vs. young
B7	Burger et al. 2007 dataset
B8	Burger et al. 2008 dataset
BL	Blalock et al. 2003 dataset
BH	Benjamini and Hochberg
CA1	cornu ammonis subdivision 1
CA3	cornu ammonis subdivision 3
CaMK	calcium/calmodulin-dependent protein kinase
cAMP	cyclic adenosine monophosphate
CBP	CREB binding protein
cDNA	complementary DNA
CEL	Affymetrix CEL data file format
CNS	central nervous system

CREB	cAMP responsive element binding protein
DAVID	the database for annotation, visualization and integrated discovery
DNA	deoxyribonucleic acid
Eif2	eukaryotic translation initiation factor 2
ERK	mitogen-activated protein kinase 1
GA	general aging
GABA	gamma-aminobutyric acid
GANSI	general aging genes associated with non-syndromic learning impairments
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GASI	general aging genes associated with syndromic learning impairments
GDP	guanosine diphosphate
GTP	guanosine triphosphate
HC	Hierarchical clustering
IAC	inter-array correlation
IPA	ingenuity pathway analysis
IQR	interquartile range
IU	impaired vs. unimpaired
K9	Kadish et al. 2009 dataset
Kv	voltage-gated potassium channels

LTD	long-term depression
LTP	long-term potentiation
MAPK	mitogen-activated protein kinase
MBEI	model based expression index
MM	mismatch
NMDA	N-methyl-D-aspartate
NUSE	normalized unscaled standard errors
PGL	published gene lists
PI3K	phosphatidylinositol 3-kinase
PKA	protein kinase A
PM	perfect match
R7	Rowe et al. 2007 dataset
RAE_EXCLUSIVE	probe sets exclusive to the RAE230A chip type
RGU_EXCLUSIVE	probe sets exclusive to the RGU34A chip type
RLE	relative log expression
RMA	robust multi-array approach
RNA	ribonucleic acid
SMD	standardized mean difference
WGCNA	weighted gene co-expression network analysis

Chapter 1 Introduction

1 Age-associated spatial learning impairment: genes and gene networks discovery from high-throughput large scale gene expression data

1.1 Aging and cognition

Aging has debilitating effects on many human physiological functions (e.g. vision, hearing, hormonal balance, motivation, physical activity, speed of movement, coordination, and cognition) (Glorioso and Sibille, 2011; Young, 1997). However, one of the most significant effects of aging is the decrease in normal brain function, particularly, cognition and memory, which is critical to carry out daily life activities (Sharma et al., 2010). Symptoms related to cognitive decline range from benign senescent forgetfulness (as seen in normal aging) to the memory loss that characterizes various age-related neurodegenerative disorders such as Alzheimer's (Landfield et al., 1992; Mattson and Magnus, 2006; Tanzi and Bertram, 2001), Parkinson's (Everse and Coates, 2009; Poletti et al., 2011), and Huntington's diseases (Dumas et al., 2013). Deficits in memory function may also arise from other psychiatric and neurological disorders such as mental retardation, autism, attention deficit disorder, learning disability, schizophrenia, and depression (Khan et al., 2014). These symptoms and disorders in the aging brain are often characterized by various physical changes including losses of white matter integrity, cortical thickness, grey matter volumes, metabolic activity, and neurotransmitter functions (Li and Rieckmann, 2014). However, normal aging by itself is associated with variable rates of cognitive performance and motor decline, which is generally gradual and progressive and can be severely impairing for the most seriously affected individuals.

Normal aging of the brain is generally described as the aging of the central nervous system in the absence of clinically-diagnosed neurodegenerative or psychiatric diseases, or of a related pathology (Glorioso and Sibille, 2011; Sharma et al., 2010). This normal age related memory decline, termed as "age-associated memory impairments", is generally observed

over age 50 and its prevalence is estimated to be 35 – 98% (Larrabee and Crook, 1994). While molecular changes occurring during normal brain aging substantially overlap with those observed in the context of many age-gated neurodegenerative and psychiatric diseases, it is necessary to investigate this aspect separately from those affected by aging-related disease processes in order to understand the contribution of normal aging to memory deficits. As the general life expectancy of human populations is increasing, understanding brain aging and aging-related cognitive declines has become a key challenge for neuroscience and psychology in the 21st century (Gallagher et al., 2003; Li and Rieckmann, 2014; Sharma et al., 2010).

1.2 Cognitive processes and their impairments through normal aging

Cognition is a broad term that applies to processes such as memory, association, language, attention, concept formation, and problem solving (Sharma et al., 2010). Cognitive processes are mental processes by which knowledge is acquired through perception, intuition, reasoning, judgment, and learning. Decline in these processes is therefore characterized by increasing difficulty with speech, coordination, learning, and the processing of new information quickly. These manifestations are highly heterogeneous and can be individual, family, or population specific. They continue to increase with the current trend in longevity in most populations (Burger et al., 2007; Glorioso et al., 2011; Peleg et al., 2010). As such they are emerging as a major societal challenge.

Normal age-associated declines in neurological functioning have been extensively studied. These works have demonstrated a ~40 to 60% decline in cognitive speed at age 80 compared to age 20 in non-demented adults (Glorioso and Sibille, 2011; Lindenberger, 2014). Interestingly, aging is known to differentially affect aspects of neurological functioning. For example, the so-called ‘‘crystallized abilities’’ related to knowledge or expertise, such as vocabulary, world knowledge, general knowledge, implicit memory, and occupational expertise do not decline or may even show improvement over the life span. In contrast, the ‘‘fluid abilities’’ or those reliant on processing speed, problem solving,

inhibitory function, working memory, long-term memory, and spatial ability decline with age. Studies have clearly shown that like humans other non-human primates also exhibit age-related declines in cognitive abilities such as reasoning, mental speed, memory, and spatial learning (Burgess et al., 2002). These findings are further supported by studies involving aging canines and rodents (Burger et al., 2007; Glorioso et al., 2011; Keller, 2006; Peleg et al., 2010).

1.3 Learning and memory

Learning is the process by which we acquire knowledge about the world and memory is the process by which that knowledge is encoded, stored, and later retrieved (Sharma et al., 2010). Learning is the process that modifies subsequent behavior while memory is the ability to remember past experiences. Memory is one of the earliest cognitive functions to show decline during aging (Sharma et al., 2010).

From mollusks to mammals, memory can be generally categorized into short-term and long-term memory (Khan et al., 2014; Paul et al., 2009; Sharma et al., 2010). Short-term memory has a limited capacity and lasts for a short period of time. Long-term memory is often divided into two main types: declarative (or explicit) memory and procedural (or implicit) memory. Declarative memory answers the question “what”, and includes knowledge of facts and events (that can be conventionally transmitted or expressed) such as places, people, and things, and the meaning of these facts. Declarative memory refers to those memories that can be consciously recalled (or "declared"). It can be further sub-divided into episodic memory and semantic memory. Episodic memory deals with personally experienced events specific to a particular context such as time and place, and conscious recollection of those events. However, semantic memory involves knowledge of these facts, meanings, and concepts taken independent of the context in which they were learned. Procedural memory, acquired through repetition and practice, answers the question “how”. It is the unconscious memory of skills, habits, and how to do things, particularly the use of objects or movements of the body, such as tying a shoelace, playing a guitar, or riding a

bike. Procedural or implicit memory deals with information about motor or perceptual skills that may not be orally transmitted.

Episodic memory, which depends on the ability to remember in a determined temporal and spatial context, is especially vulnerable to normal aging and shows a decline with age. While semantic memory does not show any age-related decline (rather, it improves with age), procedural memory remains relatively unaffected by age (Khan et al., 2014; Sharma et al., 2010).

1.4 Spatial memory

Spatial memory is typically conceptualized as a subtype of episodic memory because it stores information within the spatio-temporal frame (Rolls, 2013; Sharma et al., 2010). Spatial memory answers the question “where”. It can be defined as that brain function responsible for recognizing, codifying, storing, and recovering spatial information about the arrangement of objects or specific routes (Paul et al., 2009). Spatial memory is represented in the brain by at least two different dimensions or reference frames: egocentric (personal/body reference frame) and allocentric (external/environmental reference frame) (Galati et al., 2010; Hartley et al., 2014; Paul et al., 2009). Spatial frameworks tied to a particular body part, object or action, are represented by egocentric referencing (e.g. catching a ball or picking a fruit from a tree). Throughout the brain, individual neurons are often found to have spatially restricted firing fields, which carry egocentric spatial information about the source of sensory information or destination of planned actions (Hartley et al., 2014). For example, a neuron in the primary visual cortex might respond to a stimulus in a particular part of the visual field, while a neuron in the primary somatosensory cortex might respond to a tactile stimulation of a particular body part, and the firing of a motor neuron might help to direct limb movements in a specific direction. In each case, neural activity reflects the spatial relationship between a stimulus or response and a part of the body.

Spatial frameworks that are fixed with respect to the outside world, independent of particular actions, body parts, and objects are represented by allocentric referencing (e.g. navigating long distances over natural terrain or through a new city). Considerable evidence supports that aged humans have trouble navigating or finding their way in a large environment and remembering spatial relationships among landmarks (Sharma et al., 2010). In some studies, humans show a 30-80% drop in performance of spatial memory tasks with advancing age. Similar findings related to ASLI are supported by numerous studies in rats (Blalock et al., 2003; Burger et al., 2007; Rowe et al., 2007), mice (Pawlowski et al., 2009; Peleg et al., 2010; Schimanski and Nguyen, 2004; Verbitsky et al., 2004), and Monkey (Gallagher et al., 2003).

1.5 Spatial memory and hippocampus

Neuroimaging studies in the medial temporal lobes and prefrontal cortex have shown an age-related decrease in functional activity that is subsequently linked to poorer memory performance (Khan et al., 2014). The different types of long-term memory are stored in different regions of the brain and undergo quite different processes. Declarative memories are processed by the hippocampus, entorhinal cortex and perirhinal cortex (all within the medial temporal lobe of the brain), but are consolidated and stored in the temporal cortex and elsewhere. Procedural memories, on the other hand, do not appear to involve the hippocampus at all, and are encoded and stored by the cerebellum, putamen, caudate nucleus and the motor cortex, all of which are involved in motor control.

During the past decades strong evidence has emerged showing that the hippocampus is critical to learning and memory, particularly age-associated allocentric spatial memory (Hartley et al., 2014; Paul et al., 2009; Rolls, 2013; Sharma et al., 2010). This allocentric spatial long-term learning and memory framework is dependent on a specialized system centered on the hippocampus, especially the right hippocampus, a phylogenetically ancient and well-preserved structure, which in humans is found deep in the medial temporal lobes (Burgess et al., 2002; Hartley et al., 2014; Paul et al., 2009). The parts of the hippocampus that are of most interest to spatial memory are the dentate gyrus and the CA1, and CA3

regions of the Cornu Ammonis (CA). The encoding process of allocentric spatial memory in the hippocampal neuron is characterized by a localized activity called “place fields” (Gallagher et al., 2003). Notable differences are observed in the dynamic properties (e.g. being less stable) of place fields in older animals compared to aged. Studies of human subjects with hippocampal damage provide evidence that this brain region plays a critical role in spatial or topographical memory (Burgess et al., 2002). Spatial function is also dependent on those medial temporal and parietal regions through which the hippocampus receives its input (Hartley et al., 2014).

The hippocampus contains certain spatial cells that provide an exquisitely detailed representation of an animal’s current location and heading (Hartley et al., 2014). The major categories of spatial cells include place cells, head direction cells, grid cells, and boundary cells, each of which has a characteristic firing pattern. These cells seem to be able to create a mental/cognitive map of space or spatial information in the hippocampus (Burgess et al., 2002; Paul et al., 2009; Sharma et al., 2010).

In summary, the hippocampus is integral to memory function (including spatial memory) and is greatly affected by aging (Burgess, 2002; Gallagher et al., 2003; Morris et al., 1982). Furthermore, it is among the first regions to be affected during dementia (Mesulam, 1999; Pawlowski et al., 2009; Small et al., 2002; Verbitsky et al., 2004). However, the mechanisms underlying age dependent cognitive impairments, including spatial learning impairments such as ASLI are not as well understood.

1.6 Functional and structural changes in the brain are connected to cellular morphology

Numerous studies including longitudinal and cross-sectional fMRI studies have demonstrated consistent grey and white matter volume loss or changes in specific areas in the brain with age (Glorioso and Sibille, 2011; Khan et al., 2014). Decrease in grey matter volume is observed in specific areas of the frontal cortex and is consistent among studies. In contrast, the hippocampus and amygdala display variable effects or unchanged white

matter volumes between studies. The area-specificity of these changes is consistent with age-related cognitive changes. Thus, substantial evidence indicates that structural and functional changes correlate with cognitive changes. However, studies have also shown that there is little or no neuronal death during normal aging. Instead, grey matter volume losses appear to result from age-related reduction in dendritic spine density and synaptic losses. These findings are also supported by age-associated changes in the glial processes (Glorioso and Sibille, 2011; Khan et al., 2014).

In addition to structural, functional, and cellular morphological changes, aging neurons show vulnerability to metabolic changes, cellular insult, and other environmental factors (Glorioso and Sibille, 2011). Progressive morphological and molecular changes within life-long existing neurons and glia likely underlie age-related cognitive, motor, and mood changes and disease susceptibility. As the neurons age they display evidence for increasing DNA damage, accumulation of reactive oxygen species, calcium dysregulation, mitochondrial dysfunction, and inflammatory processes (Foster and Kumar, 2002; Kelly et al., 2006). Improving cellular metabolic environment through diet and caloric restriction seem to improve memory performance in various studies (Zeier et al., 2011).

1.7 Molecular mechanisms of memory

Human subjects as young as 14 years of age display molecular changes in the brain on a continuum with age-associated changes that extend throughout old age (Erraji-Benchekroun et al., 2005). It is thus likely that molecular aging partially extends from developmental processes (Glorioso and Sibille, 2011). The changes in dendritic spine density, and synaptic losses during normal brain aging contributing to learning and memory impairments, have been attributed to the underlying molecular mechanism known as synaptic plasticity. Synaptic plasticity is the process by which connections between co-active neurons are strengthened or weakened (Kelly et al., 2006; Neves et al., 2008). Functional and structural changes in dendritic spines and synapses are considered to be the basis of learning and memory in the brain. This subject has been well studied in the brain for general cognition, as well as in the hippocampus in relation to spatial learning (Khan et al., 2014;

Neves et al., 2008; Schimanski and Nguyen, 2004). The much studied model of synaptic plasticity known as long-term potentiation (LTP) was first identified in the hippocampus. Memory processing in the hippocampus was found associated with synaptic stability and the conversion of simple synapses into complex synaptic structures such as perforated synapses and multi-synaptic boutons. Indeed, an increase in the number of perforated synapses has been associated with the induction and maintenance of LTP (Khan et al., 2014). Evidence also suggests that LTP can critically influence the expression of spatial learning and memory (Schimanski and Nguyen, 2004). Other forms of activity-dependent synaptic plasticity have been documented, including long-term depression (LTD), EPSP-spike (E-S) potentiation, depotentiation, and de-depression (Neves et al., 2008).

Synaptic plasticity and related molecular mechanisms contributing to learning and memory formation in the brain are thought to be driven by many unique genetic modulators. These modulators include neurotrophins (e.g. BDNF), neurotransmitters such as serotonin, dopamine, glutamate, as well as many other neurological disease-related genes (Glorioso and Sibille, 2011; Khan et al., 2014). BDNF is one of the best characterized modulator of normal brain aging (Tapia-Arancibia et al., 2008). It is an activity dependent secreted growth factor that declines steadily with age in the brain. Serotonin has been hypothesized to play a role in normal brain aging, as its levels and selected receptor functions are age-regulated. Moreover, serotonin shares signaling pathways with other known age-regulatory molecules, such as BDNF and IGF-1 (Mattson et al., 2004a; Mattson et al., 2004b). Glutamate, the brain predominant excitatory neurotransmitter, is also a probable candidate for modulating brain aging, as it facilitates the release of BDNF and is essential for LTP, synaptic plasticity, and neurogenesis (Mattson, 2008). There is also strong correlative and causative evidence that dopamine plays a role in the modulation of brain aging (Backman et al., 2006). For example, it is implicated in several age-gated diseases, including Parkinson's disease, Huntington's disease, schizophrenia, and bipolar disorder. Components of the dopamine system also decline with age, including the dopamine transporter as well as the dopamine 1 (D1) and dopamine 2 (D2) receptors. In addition to declining dopamine signaling having cross-talk with caloric restriction pathways and directly mediating age related cognitive decline,

dopamine pathways also have cross-talk with ROS and other age-related molecular pathways (Glorioso and Sibille, 2011). However, beyond these genetic modulators, there seems to be other mechanisms or factors involved in the synaptic plasticity process.

1.8 Transcriptional regulation of memory formation and consolidation

Aging and age-associated cognitive impairments are complex and multifactorial and involve both genetic as well as environmental determinants (Buckner, 2004; Finch and Tanzi, 1997). Over the past few decades, a significant amount of evidence demonstrates that alterations in gene expression (transcription and translation) and protein degradation in neurons across several brain regions are required for proper memory storage and retrieval (Jarome and Lubin, 2014). Among the mechanisms required for synaptic plasticity, coordinated changes in gene expression are essential for the consolidation and maintenance of most lasting forms of memory (Kandel, 2001; Penney and Tsai, 2014). Indeed, these transcriptional and translational changes result in structural and functional changes to synapses, leading to alterations in synaptic plasticity (Kandel, 2001; Penney and Tsai, 2014).

An additional level of transcriptional regulation occurs in the form of chromatin remodeling, also known as epigenetic modification, whereby the accessibility of specific regions of the genome to the transcription machinery can be modulated by local posttranslational modification of histone proteins (Penney and Tsai, 2014). Epigenetic modifications have emerged as an attractive molecular genetic mechanism involved in transient and persistent gene transcriptional regulation during long-term memory formation and storage (Franklin and Mansuy, 2010; Graff and Mansuy, 2008; Jarome and Lubin, 2014; Levenson and Sweatt, 2005; Liu et al., 2009; Sweatt, 2010). There is strong evidence that epigenetic mechanisms are critical regulators of learning-dependent synaptic plasticity. The most important or relevant molecular mechanisms include DNA methylation and the modification of histone proteins by acetylation, phosphorylation, and methylation, among others. Interestingly, short-term memory, which usually lasts from minutes to few hours, does not require new RNA and protein synthesis. Posttranscriptional modification of existing molecules is

sufficient for short-term memory formation and storage. However, the formation of long-term memory requires several hours and involves new RNA and protein synthesis that sequentially occurs at precise times during the process (e.g. training) (Abel and Lattal, 2001). Chromatin remodeling in the hippocampus is necessary for stabilizing long-term memory, including spatial memory (Peleg et al., 2010; Penney and Tsai, 2014; Sweatt, 2010). Thus gene expression has been identified as the key mechanism by which changes can occur in the cellular state of key molecules such as neurotrophins and neurotransmitters, as well as many epigenetic factors (Barco et al., 2006; Kandel, 2001; Levenson et al., 2006; Peleg et al., 2010). Although, studies have identified some genes involved in the regulation of synaptic plasticity processes and provided some insight into the mechanisms of learning and memory formation, the molecular mechanisms of age-associated learning impairments remains to be fully understood.

1.9 Assessment of hippocampus dependent spatial memory in animals

In order to understand the molecular mechanisms at work in ASLI, animal models have been used extensively in the past. Evaluation of hippocampal-based spatial learning and memory has been assessed by numerous behavioral paradigms in rodents (Paul et al., 2009; Sharma et al., 2010). These paradigms include various forms of mazes such as the T-maze, the radial-arm maze, the Barnes circular maze, the Morris water maze and other experimental devices (Morris et al., 1982; Morris et al., 1986). The Morris water maze is one of the most widely used methods for the evaluation of allocentric spatial learning and memory including ASLI. Learning is faster in this device than in other mazes (radial maze, circular maze) and often considered as the gold standard. The Morris water maze consists of a round pool filled with water made opaque using milk or white paint. Animals learn to locate the platform hidden (2-3 cm) below the water from four different starting points. Since water immersion represents an aversive stimulus, training starts with a period of habituation during which the animals are immersed in the water and allowed to swim for a few minutes without a platform. Later on, a platform is placed in a fixed position in one of the sectors (quadrants) and the animals go through a period of acquisition. During this period they are given a

variable number of daily trials and animals learn the location of the hidden platform based on distal cues. With time the latency to locate the platform decreases. The strength of learning is evaluated afterwards by a probe trial in which the hidden platform is removed and the amount of time spent in the former region of platform is measured. The Morris water maze is most suitable for rats whereas the T-maze, Barnes maze or radial mazes are often used for mice.

1.10 High throughput microarray studies involving ASLI

As explained above memory formation involves transcriptional, translational, and epigenetic changes triggered by the postsynaptic activation of neurotransmitter receptors (Barco et al., 2006; Kandel, 2001; Levenson and Sweatt, 2006; Peleg et al., 2010). Attempts in the last decade to gain insight into aging and age-associated learning impairments have been aided by advances in genome-wide methods and technologies, particularly gene expression studies involving microarrays. Microarray technology, which interrogates thousands of genes in a single experiment, has seen tremendous growth in the last decade and has become increasingly accessible and affordable. As a result, there has been an influx of large amounts of data, much of which has been deposited in various public data repositories. These repositories also contain data from studies involving animal models and microarrays, more specifically, from studies that attempted to understand the gene expression changes related to aging and age-associated memory impairments in the hippocampus in humans (Lu et al., 2004) using post-mortem tissues (Glorioso and Sibille, 2011) and in animal models such as monkeys (unpublished) and in rodents after behavioral training using the Morris water maze (Blalock et al., 2003; Burger et al., 2007; Burger et al., 2008; Haberman et al., 2013; Kadish et al., 2009; Pawlowski et al., 2009; Rowe et al., 2007; Verbitsky et al., 2004) or other learning paradigm (Peleg et al., 2010). Data from these microarray gene expression studies were generally used in differential expression analysis and functional and pathway analysis. Results show that learning induces a complex reprogramming of gene expression involving the coordinated regulation of many genes, which is also affected by aging processes.

However, the results of the individual studies are heterogeneous and often difficult to interpret. They often highlight different gene sets and pathways, have limited conclusions, and do not consider their broader implications that may go beyond individual experiments. It is therefore desirable to integrate results from these studies towards a consensus view of the genes affected and the molecular mechanisms underlying brain aging and age-associated learning impairments. This is now possible because of the availability of considerable amount of original microarray data in the public microarray data repositories, as well as the availability of improved statistical analytical methods. Meta-analysis is one such method that can integrate results from multiple independent studies. Meta-analysis also offers powerful ways that have been used in the past to identify genes that are significantly differentially expressed between two treatment groups (Ch'ng et al., 2015; Huan et al., 2015; Uddin and Singh, 2013).

1.11 Inferring gene (regulatory and co-expression) networks from microarray gene expression profiles

Gene regulation is one of the most important biological processes in organisms. Genes interact in networks, where the expression level of one gene is governed by the combined action of multiple other genes to execute various cellular functions in response to both endogenous (e.g. developmental) and exogenous (e.g. light) stimuli (Aluru et al., 2013) (Ahmad et al., 2012). The elucidation of these complex inter-gene interactions is fundamental to biological discoveries.

The biological effect of one gene can dynamically affect the expression and subsequent action of other genes through a complex network of interactions often referred to as a gene network. A gene network is loosely defined as a set of genes that interact with one another through transcription factors, DNA segments, and other gene and protein products thereby governing the rate at which genes in the network are transcribed into mRNAs (Rau et al., 2010). The regulatory mechanisms/architecture controlling gene expression also controls subsequent cellular behavior such as development, differentiation, homeostasis and response to stimuli. There are experimental methods based on, for example, chromatin

immunoprecepitation, DNaseI hypersensitivity, or protein-binding assays that are capable of determining the nature of gene regulation in a given system, but they are time-consuming, expensive and require antibodies for each transcription factor (Maetschke et al., 2014). As a result, the ability to model complex regulatory interactions such as those in gene networks, understanding the topology and elements of a gene network, and how they behave under different experimental paradigms has been of increasing interest as a key to metazoan systems biology (Bonneau, 2008; Cooke et al., 2009; Long et al., 2008).

Gene network modeling can infer networks where interactions between two genes can refer to an indirect regulation via proteins, metabolites, and ncRNA that have not been measured directly. The networks can also include physical interactions, if the two interacting partners are a transcription factor and its target, or two proteins in the same complex (Bansal et al., 2007). Gene networks provide a systematic understanding of molecular mechanisms underlying biological processes (Allen et al., 2012). A gene network analysis is able to identify regulatory or co-expression relationships from thousands of gene expression profiles generated from microarray experiment and is also able to depict a graphical network representation of the underlying regulatory and signaling processes. Large body of microarray data contains information which may allow reconstruction of regulatory networks (Needham et al., 2009). Significant gene lists from a single- or meta-analysis of microarray studies are traditionally used for functional or pathway analysis in order to understand their biological significance. However, molecular pathway analysis of differentially expressed genes obtained from expression profiles is constrained by the current state of molecular knowledge and does not provide a prioritization of molecules within the affected pathways (Gaiteri et al., 2014).

Although, the traditional expression analysis methods have significant potential to infer candidate genes that may contribute to a certain signaling pathway, it is not often possible to determine which transcription and regulatory factors mediate this regulation (Needham et al., 2009). Moreover, traditional functional and pathway analysis can help us identify only known interactions that are present in currently available knowledge bases. For example,

for many significant genes, information in the knowledge base is not available about the genes' function and pathways, or how the genes co-express or interact with one another in a gene regulatory network. Often, in a gene interaction network that is based on a knowledge base, association is made between two or more genes solely based on their co-citation in the literature databases, when in reality no biological interaction exists between them. Moreover, traditional pathways or regulatory networks constructed from a set of gene lists do not reveal the importance of genes that are key modulators in the pathways. This can be overcome by constructing gene networks using mathematical modeling approaches. In complex multi-factorial traits such as learning and memory formation many genes are involved in a complex regulatory or co-expression relationship. Inference of gene network models can help identify key genes and their networks based on gene expression data alone.

1.12 Methods available for constructing gene networks from microarray data

In recent years, microarray gene expression data have been used extensively for inferring gene networks from a wide variety of sources, for example, from Yeast (Dawy et al., 2011; Friedman et al., 2000; Friedman, 2004; Nachman et al., 2004; Nariai et al., 2004; Wang et al., 2009; Zoppoli et al., 2010), Bacteria (Hodges et al., 2010; Wang et al., 2010), virus (Recchia et al., 2008), Arabidopsis (Locke et al., 2005; Needham et al., 2009; Zeilinger et al., 2006), honey bee (Ko et al., 2009), mouse (Ghazalpour et al., 2006; Ko et al., 2009), human B cells (Basso et al., 2005), breast cancer (Niida et al., 2008; Schafer and Strimmer, 2005), brain transcriptome (Levine et al., 2013; Miller et al., 2010; Voineagu et al., 2011), and transcriptional changes in Alzheimer's disease and normal aging (Miller et al., 2008). However, no such published research is available to date in the context of ASLI.

A large number of gene network inference methods have been developed for steady-state data over the past two to three decades (Allen et al., 2012; Bansal et al., 2007; De Smet and Marchal, 2010; Emmert-Streib et al., 2012; Hache et al., 2009; Maetschke et al., 2014; Margolin et al., 2006; Markowetz and Spang, 2007; Olsen et al., 2009; Penfold and Wild,

2011; Villaverde and Banga, 2014; Werhli and Husmeier, 2007). Gene network models have been constructed from microarray datasets using a variety of reverse engineering machine learning and statistical methods. The earliest proposed models for gene networks from microarray data include co-expression networks (Eisen et al., 1998), weight matrices (Weaver et al., 1999), and discrete Boolean models (Akutsu et al., 1999; D'Haeseleer et al., 2000). These methods suffered from disadvantages and information loss. Subsequent commonly used unsupervised statistical methods that have been proposed for this purpose can be classified into four categories (Allen et al., 2012; Aluru et al., 2013; Emmert-Streib et al., 2012):

1. Probabilistic network – based approaches
2. Correlation-based method
3. Partial-correlation-based methods
4. Information-theory based methods

1.12.1 Probabilistic network – based approaches

These are mainly based on Bayesian probability theory (Neapolitan, 2009) and are referred to as Bayesian networks. A Bayesian network is a probabilistic graphical network model that represents a set of relationship or interactions depicted by edges or arrows between variables (e.g. genes as nodes). In mathematical terms a Bayesian network, defined as (G, P), consists of a Directed Acyclic Graph such as (G) and a joint probability distribution (P) of the variables that together satisfy the Markov condition (Djebbari and Quackenbush, 2008; Neapolitan, 2009; Needham et al., 2007).

Bayesian networks have become popular methods for modeling gene regulatory networks from microarray data, since they are able to represent complex stochastic processes and allow combinatorial and non-linear relationships among variables of complex biological systems (Friedman et al., 2000; Hartemink et al., 2002). The resulting networks provide a high level description of the gene expression system by predicting how genes interact with each other through regulatory actions (Dawy et al., 2011). Networks constructed using Bayesian network theory also accommodate missing data by modeling the effect of hidden variables such as genes, transcription factors and proteins not included on a particular

microarray (Needham et al., 2007). They offer a simple way to visualize the structure of the model and express causal relationships or dependencies between variables (Neapolitan, 2009; Needham et al., 2007). Bayesian networks can also incorporate prior knowledge of gene relationships (Hecker et al., 2009; Ko et al., 2009; Needham et al., 2007).

Bayesian networks have also been used to infer gene networks from microarray gene expression data and from a wide variety of sources in many areas of the biological sciences , for example, to infer cellular networks (Friedman, 2004), transcriptional regulation (Brun et al., 2007; Cooke et al., 2009), genetic networks (Djebbari and Quackenbush, 2008), phylogenetic networks (Strimmer and Moulton, 2000), protein signaling pathways (Sachs et al., 2005), protein-protein interactions (Burger and van Nimwegen, 2008; Woolf et al., 2005), and biological pathways (Hodges et al., 2010; Ko et al., 2009). Bayesian network inference has also been used in systems biology (Troyanskaya et al., 2003) and transcription regulatory module discovery (Huttenhower et al., 2009).

For time-series data, two commonly used methods are dynamic Bayesian networks (Husmeier, 2003; Perrin et al., 2003; Zou and Conzen, 2005) and ordinary differential equations (Cao and Zhao, 2008; Chen et al., 1999). Dynamic Bayesian networks are an extension of Bayesian network, which allows a dynamic process to be modeled. Bayesian network methods seem to show the greatest promise in the analysis of steady state expression data to find causal relationship among the variables (Friedman et al., 2000; Needham et al., 2007).

1.12.2 Correlation-based method

Correlation-based methods (Langfelder and Horvath, 2008; Li et al., 2009; Zhang and Horvath, 2005) are one of the most popular gene network modeling approach. In WGCNA (Zhang and Horvath, 2005), a relatively new statistical approach, an undirected correlation or co-expression network is created by calculating connection strength between each pair of genes. The resulting data provide a network adjacency or connection strength matrix. The connection strength between each pair of genes is the absolute Pearson correlation of

their expression profiles from microarray data raised to a power of beta. Beta is the weight, a soft threshold, and is determined in such a way so that the resulting network follows approximate scale free topology. However, a hard (Carter et al., 2004) threshold is also applied to determine the biological meaningfulness of the connections. These correlation-based methods have been used in several studies and have shown that they are not only useful in interpreting biological results but also in identifying important hub genes and gene modules (Li et al., 2009; Mao et al., 2009; Maschietto et al., 2015; Mason et al., 2009; Ruan et al., 2010; Stuart et al., 2003; Torkamani et al., 2010; Voineagu et al., 2011; Ye and Liu, 2015).

The WGCNA method has been successfully applied in recent studies to identify several novel disease-related genes (Carlson et al., 2006; Ghazalpour et al., 2006; Horvath et al., 2006; Oldham et al., 2006). The WGCNA R package (Langfelder and Horvath, 2008) implements both weighted and unweighted correlation networks and identifies modules or subnetworks using hierarchical clustering approaches. Aside from the many functions available for network construction and module/sub-network identification, the R package also provides functions for calculating topological properties and network visualization. Co-expression networks serve mainly to explore the functionality of genes on a systems level (Zhang and Horvath, 2005) and do not aim to be causal representations of regulatory networks. However, they can include both indirect and direct relationships between pair of genes.

1.12.3 Partial-correlation-based methods

Partial-correlation-based methods are based on graphical Gaussian model theory (Cox and Wermuth, 1996; Dempster, 1972; Koller and Friedman, 2009). Ggraphical Gaussian model is a graphical model which assumes that all variables are distributed according to a multivariate normal distribution with a specific structure of the inverse of the covariance matrix (Emmert-Streib et al., 2012). Partial-correlation-based methods infer the conditional dependency by the non-zero entries in the concentration matrix, $C = \left[c_{i,j}\right] = S^{-1}$ also called the precision matrix, which is the inverse of the covariance matrix. The zero entries

$c_{i,j} = 0$ in the concentration matrix imply conditional independency between the expression levels of gene *i* and *j* given the expression of all other genes; in other words, two genes do not interact directly with each other (Allen et al., 2012).

1.12.4 Information-theory based methods

The most popular information-theory-based methods are the relevance networks (Butte and Kohane, 2000) and mutual information networks such as ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) (Basso et al., 2005; Margolin et al., 2006). The principle idea of relevance networks is to compute all mutual information values for all pairs of genes, for a given dataset, and declare mutual information values as relevant if their corresponding correlation coefficient value is larger than a given threshold. The resulting network is constructed based on this threshold by including an edge between two genes in the respective adjacency matrix of the network; otherwise no edge is included between them. ARACNE is similar to relevance networks and uses mutual information to determine the dependency among the genes. However, ARACNE adds a second step in which it then removes indirect interactions using a process known as data processing inequality (DPI). ARACNE has been successfully applied to construct gene regulatory networks in the context of specific cellular types, and demonstrated good performance (Basso et al., 2005; Margolin et al., 2006).

A number of other mutual information based methods have also been described in the literature, for example, CLR (context likelihood of relatedness) (Faith et al., 2007), C3NET (conservative causal core) (Altay and Emmert-Streib, 2010), SA-CLR (synergy augmented CLR) (Anastassiou, 2007; Watkinson et al., 2009); MRNET (maximum relevance, minimum redundance) (Meyer et al., 2007), and others (reviewed in (Emmert-Streib et al., 2012).

1.13 Exploring genes and gene networks in ASLI

In light of the discussion above it is clear that the molecular mechanisms of learning and memory formation are highly complex and have yet to be fully understood. It is therefore critical to identify important genes and explore how they communicate in molecular

networks in order to understand the molecular mechanisms contributing to learning and memory formation. This is timely, given that the use and availability of microarrays in animal models (e.g. rats, mice) of ASLI have generated a large body of genome-wide gene expression results. In addition, several statistical and mathematical modeling approaches have been developed – such as meta-analysis and gene co-expression network analysis – that utilize microarray gene expression data. However, no such published research is available to date in the area of ASLI that uses meta-analysis or gene co-expression network analysis. Therefore, I propose the following hypothesis and major objectives.

1.14 Hypothesis and objectives

I hypothesize that it is possible to identify novel genes and their co-expression networks critical during aging and learning impairments using meta-analysis and mathematical modeling on microarray gene expression data. I have the following three objectives to test my hypothesis.

Objective 1 (chapter 2): Perform selection, collection, quality control, and preprocessing of ASLI gene expression data, and examine their importance for downstream meta- and network analysis.

Objective 2 (chapter 3): Integrate a set of microarray gene expression data using meta-analysis methods, identify and characterize genes that may be involved in ASLI, and identify and characterize gene networks based on existing knowledge.

Objective 3 (chapter 4): Identify key genes and their networks in ASLI by gene co-expression network modeling using WGCNA.

This research would allow one to identify key genes that may be affected by age, learning impairments, or learning impairments associated with aging for rats. Moreover, it will offer valuable insight into the possible regulatory mechanisms of the genes involved and the specific role they may play in cognitive impairments, specifically ASLI, which can provide valuable information for generating new hypotheses for future experimental research. This

research will have significant implications for studying complex disorders from a broader system's perspective.

1.15 References

Abel, T., Lattal, K.M., 2001. Molecular mechanisms of memory acquisition, consolidation and retrieval. Curr Opin Neurobiol. 11**,** 180-7.

Ahmad, F.K., Deris, S., Othman, N.H., 2012. The inference of breast cancer metastasis through gene regulatory networks. J Biomed Inform. 45**,** 350-62.

Akutsu, T., Miyano, S., Kuhara, S., 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput. 17-28.

Allen, J.D., Xie, Y., Chen, M., Girard, L., et al., 2012. Comparing statistical methods for constructing large scale gene networks. PLoS One. 7**,** e29348.

Altay, G., Emmert-Streib, F., 2010. Inferring the conservative causal core of gene regulatory networks. BMC Syst Biol. 4**,** 132.

Aluru, M., Zola, J., Nettleton, D., Aluru, S., 2013. Reverse engineering and analysis of large genome-scale gene networks. Nucleic Acids Res. 41**,** e24.

Anastassiou, D., 2007. Computational analysis of the synergy among multiple interacting genes. Mol Syst Biol. 3**,** 83.

Backman, L., Nyberg, L., Lindenberger, U., Li, S.C., et al., 2006. The correlative triad among aging, dopamine, and cognition: current status and future prospects. Neurosci Biobehav Rev. 30**,** 791-807.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D., 2007. How to infer gene networks from expression profiles. Mol Syst Biol. 3**,** 78.

Barco, A., Bailey, C.H., Kandel, E.R., 2006. Common molecular mechanisms in explicit and implicit memory. J Neurochem. 97**,** 1520-33.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., et al., 2005. Reverse engineering of regulatory networks in human B cells. Nat Genet. 37**,** 382-90.

Blalock, E.M., Chen, K.C., Sharrow, K., Herman, J.P., et al., 2003. Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment. J Neurosci. 23**,** 3807-19.

Bonneau, R., 2008. Learning biological networks: from modules to dynamics. Nat Chem Biol. 4**,** 658-64.

Brun, M., Kim, S., Choi, W., Dougherty, E.R., 2007. Comparison of gene regulatory networks via steady-state trajectories. EURASIP J Bioinform Syst Biol. 82702.

Buckner, R.L., 2004. Memory and executive function in aging and AD: multiple factors that cause decline and reserve factors that compensate. Neuron. 44**,** 195-208.

Burger, C., Lopez, M.C., Feller, J.A., Baker, H.V., et al., 2007. Changes in transcription within the CA1 field of the hippocampus are associated with age-related spatial learning impairments. Neurobiol Learn Mem. 87**,** 21-41.

Burger, C., Lopez, M.C., Baker, H.V., Mandel, R.J., et al., 2008. Genome-wide analysis of aging and learning-related genes in the hippocampal dentate gyrus. Neurobiol Learn Mem. 89**,** 379-96.

Burger, L., van Nimwegen, E., 2008. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol. 4**,** 165.

Burgess, N., 2002. The hippocampus, space, and viewpoints in episodic memory. Q J Exp Psychol A. 55**,** 1057-80.

Burgess, N., Maguire, E.A., O'Keefe, J., 2002. The human hippocampus and spatial and episodic memory. Neuron. 35**,** 625-41.

Butte, A.J., Kohane, I.S., 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput. 418-29.

Cao, J., Zhao, H., 2008. Estimating dynamic models for gene regulation networks. Bioinformatics. 24**,** 1619-24.

Carlson, M.R., Zhang, B., Fang, Z., Mischel, P.S., et al., 2006. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics. 7**,** 40.

Carter, S.L., Brechbuhler, C.M., Griffin, M., Bond, A.T., 2004. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics. 20**,** 2242-50.

Ch'ng, C., Kwok, W., Rogic, S., Pavlidis, P., 2015. Meta-Analysis of Gene Expression in Autism Spectrum Disorder. Autism Res.

Chen, T., He, H.L., Church, G.M., 1999. Modeling gene expression with differential equations. Pac Symp Biocomput. 29-40.

Cooke, E.J., Savage, R.S., Wild, D.L., 2009. Computational approaches to the integration of gene expression, ChIP-chip and sequence data in the inference of gene regulatory networks. Semin Cell Dev Biol. 20**,** 863-8.

Cox, D.R., Wermuth, N., 1996. Multivariate Dependencies: Models, Analysis and Interpretation., Vol., Chapman and Hall, London.

D'Haeseleer, P., Liang, S., Somogyi, R., 2000. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics. 16**,** 707-26.

Dawy, Z., Yaacoub, E., Nassar, M., Abdallah, R., et al., 2011. A multiorganism based method for Bayesian gene network estimation. Biosystems. 103**,** 425-34.

De Smet, R., Marchal, K., 2010. Advantages and limitations of current network inference methods. Nat Rev Microbiol. 8**,** 717-29.

Dempster, A., 1972. Covariance selection. Biometrics. 28**,** 157-175.

Djebbari, A., Quackenbush, J., 2008. Seeded Bayesian Networks: constructing genetic networks from microarray data. BMC Syst Biol. 2**,** 57.

Dumas, E.M., van den Bogaard, S.J., Middelkoop, H.A., Roos, R.A., 2013. A review of cognition in Huntington's disease. Front Biosci (Schol Ed). 5**,** 1-18.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 95**,** 14863-8.

Emmert-Streib, F., Glazko, G.V., Altay, G., de Matos Simoes, R., 2012. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Front Genet. 3**,** 8.

Erraji-Benchekroun, L., Underwood, M.D., Arango, V., Galfalvy, H., et al., 2005. Molecular aging in human prefrontal cortex is selective and continuous throughout adult life. Biol Psychiatry. 57**,** 549-58.
Everse, J., Coates, P.W., 2009. Neurodegeneration and peroxidases. Neurobiol Aging. 30**,** 1011-25.
Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., et al., 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 5**,** e8.
Finch, C.E., Tanzi, R.E., 1997. Genetics of aging. Science. 278**,** 407-11.
Foster, T.C., Kumar, A., 2002. Calcium dysregulation in the aging brain. Neuroscientist. 8**,** 297-301.
Franklin, T.B., Mansuy, I.M., 2010. The prevalence of epigenetic mechanisms in the regulation of cognitive functions and behaviour. Curr Opin Neurobiol. 20**,** 441-9.
Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. J Comput Biol. 7**,** 601-20.
Friedman, N., 2004. Inferring cellular networks using probabilistic graphical models. Science. 303**,** 799-805.
Gaiteri, C., Ding, Y., French, B., Tseng, G.C., et al., 2014. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav. 13**,** 13-24.
Galati, G., Pelle, G., Berthoz, A., Committeri, G., 2010. Multiple reference frames used by the human brain for spatial perception and memory. Exp Brain Res. 206**,** 109-20.
Gallagher, M., Bizon, J.L., Hoyt, E.C., Helm, K.A., et al., 2003. Effects of aging on the hippocampal formation in a naturally occurring animal model of mild cognitive impairment. Exp Gerontol. 38**,** 71-7.
Ghazalpour, A., Doss, S., Zhang, B., Wang, S., et al., 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet. 2**,** e130.
Glorioso, C., Oh, S., Douillard, G.G., Sibille, E., 2011. Brain molecular aging, promotion of neurological disease and modulation by Sirtuin5 longevity gene polymorphism. Neurobiol Dis. 41**,** 279-90.
Glorioso, C., Sibille, E., 2011. Between destiny and disease: Genetics and molecular pathways of human central nervous system aging. Prog Neurobiol. 93**,** 165-81.
Graff, J., Mansuy, I.M., 2008. Epigenetic codes in cognition and behaviour. Behav Brain Res. 192**,** 70-87.
Haberman, R.P., Colantuoni, C., Koh, M.T., Gallagher, M., 2013. Behaviorally activated mRNA expression profiles produce signatures of learning and enhanced inhibition in aged rats with preserved memory. PLoS One. 8**,** e83674.
Hache, H., Lehrach, H., Herwig, R., 2009. Reverse engineering of gene regulatory networks: a comparative study. EURASIP J Bioinform Syst Biol. 617281.
Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., Young, R.A., 2002. Combining location and expression data for principled discovery of genetic regulatory network models. Pac Symp Biocomput. 437-49.

Hartley, T., Lever, C., Burgess, N., O'Keefe, J., 2014. Space in the brain: how the hippocampal formation supports spatial cognition. Philos Trans R Soc Lond B Biol Sci. 369**,** 20120510.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., et al., 2009. Gene regulatory network inference: data integration in dynamic models-a review. Biosystems. 96**,** 86-103.

Hodges, A.P., Dai, D., Xiang, Z., Woolf, P., et al., 2010. Bayesian network expansion identifies new ROS and biofilm regulators. PLoS One. 5**,** e9513.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., et al., 2006. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci U S A. 103**,** 17402-7.

Huan, T., Esko, T., Peters, M.J., Pilling, L.C., et al., 2015. A Meta-analysis of Gene Expression Signatures of Blood Pressure and Hypertension. PLoS Genet. 11**,** e1005035.

Husmeier, D., 2003. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics. 19**,** 2271-82.

Huttenhower, C., Mutungu, K.T., Indik, N., Yang, W., et al., 2009. Detailing regulatory networks through large scale data integration. Bioinformatics. 25**,** 3267-74.

Jarome, T.J., Lubin, F.D., 2014. Epigenetic mechanisms of memory formation and reconsolidation. Neurobiol Learn Mem. 115**,** 116-27.

Kadish, I., Thibault, O., Blalock, E.M., Chen, K.C., et al., 2009. Hippocampal and cognitive aging across the lifespan: a bioenergetic shift precedes and increased cholesterol trafficking parallels memory impairment. J Neurosci. 29**,** 1805-16.

Kandel, E.R., 2001. The molecular biology of memory storage: a dialogue between genes and synapses. Science. 294**,** 1030-8.

Keller, J.N., 2006. Age-related neuropathology, cognitive decline, and Alzheimer's disease. Ageing Res Rev. 5**,** 1-13.

Kelly, K.M., Nadon, N.L., Morrison, J.H., Thibault, O., et al., 2006. The neurobiology of aging. Epilepsy Res. 68 Suppl 1**,** S5-20.

Khan, Z.U., Martin-Montanez, E., Navarro-Lobato, I., Muly, E.C., 2014. Memory deficits in aging and neurological diseases. Prog Mol Biol Transl Sci. 122**,** 1-29.

Ko, Y., Zhai, C., Rodriguez-Zas, S., 2009. Inference of gene pathways using mixture Bayesian networks. BMC Syst Biol. 3**,** 54.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques., Vol., MIT Press, Cambridge, MA.

Landfield, P.W., Thibault, O., Mazzanti, M.L., Porter, N.M., et al., 1992. Mechanisms of neuronal death in brain aging and Alzheimer's disease: role of endocrine-mediated calcium dyshomeostasis. J Neurobiol. 23**,** 1247-60.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9**,** 559.

Larrabee, G.J., Crook, T.H., 3rd, 1994. Estimated prevalence of age-associated memory impairment derived from standardized tests of memory function. Int Psychogeriatr. 6**,** 95-104.

Levenson, J.M., Sweatt, J.D., 2005. Epigenetic mechanisms in memory formation. Nat Rev Neurosci. 6**,** 108-18.

Levenson, J.M., Roth, T.L., Lubin, F.D., Miller, C.A., et al., 2006. Evidence that DNA (cytosine-5) methyltransferase regulates synaptic plasticity in the hippocampus. J Biol Chem. 281**,** 15763-73.

Levenson, J.M., Sweatt, J.D., 2006. Epigenetic mechanisms: a common theme in vertebrate and invertebrate memory formation. Cell Mol Life Sci. 63**,** 1009-16.

Levine, A.J., Miller, J.A., Shapshak, P., Gelman, B., et al., 2013. Systems analysis of human brain gene expression: mechanisms for HIV-associated neurocognitive impairment and common pathways with Alzheimer's disease. BMC Med Genomics. 6**,** 4.

Li, H., Sun, Y., Zhan, M., 2009. Exploring pathways from gene co-expression to network dynamics. Methods Mol Biol. 541**,** 249-67.

Li, S.C., Rieckmann, A., 2014. Neuromodulation and aging: implications of aging neuronal gain control on cognition. Curr Opin Neurobiol. 29**,** 148-58.

Lindenberger, U., 2014. Human cognitive aging: corriger la fortune? Science. 346**,** 572-8.

Liu, L., van Groen, T., Kadish, I., Tollefsbol, T.O., 2009. DNA methylation impacts on learning and memory in aging. Neurobiol Aging. 30**,** 549-60.

Locke, J.C., Southern, M.M., Kozma-Bognar, L., Hibberd, V., et al., 2005. Extension of a genetic network model by iterative experimentation and mathematical analysis. Mol Syst Biol. 1**,** 2005 0013.

Long, T.A., Brady, S.M., Benfey, P.N., 2008. Systems approaches to identifying gene regulatory networks in plants. Annu Rev Cell Dev Biol. 24**,** 81-103.

Lu, T., Pan, Y., Kao, S.Y., Li, C., et al., 2004. Gene regulation and DNA damage in the ageing human brain. Nature. 429**,** 883-91.

Maetschke, S.R., Madhamshettiwar, P.B., Davis, M.J., Ragan, M.A., 2014. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. Brief Bioinform. 15**,** 195-211.

Mao, L., Van Hemert, J.L., Dash, S., Dickerson, J.A., 2009. Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics. 10**,** 346.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., et al., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 7 Suppl 1**,** S7.

Markowetz, F., Spang, R., 2007. Inferring cellular networks--a review. BMC Bioinformatics. 8 Suppl 6**,** S5.

Maschietto, M., Tahira, A.C., Puga, R., Lima, L., et al., 2015. Co-expression network of neural-differentiation genes shows specific pattern in schizophrenia. BMC Med Genomics. 8**,** 23.

Mason, M.J., Fan, G., Plath, K., Zhou, Q., et al., 2009. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics. 10**,** 327.

Mattson, M.P., Maudsley, S., Martin, B., 2004a. A neural signaling triumvirate that influences ageing and age-related disease: insulin/IGF-1, BDNF and serotonin. Ageing Res Rev. 3**,** 445-64.

Mattson, M.P., Maudsley, S., Martin, B., 2004b. BDNF and 5-HT: a dynamic duo in age-related neuronal plasticity and neurodegenerative disorders. Trends Neurosci. 27**,** 589-94.

Mattson, M.P., Magnus, T., 2006. Ageing and neuronal vulnerability. Nat Rev Neurosci. 7**,** 278-94.
Mattson, M.P., 2008. Glutamate and neurotrophic factors in neuronal plasticity and disease. Ann N Y Acad Sci. 1144**,** 97-112.
Mesulam, M.M., 1999. Neuroplasticity failure in Alzheimer's disease: bridging the gap between plaques and tangles. Neuron. 24**,** 521-9.
Meyer, P.E., Kontos, K., Lafitte, F., Bontempi, G., 2007. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol. 79879.
Miller, J.A., Oldham, M.C., Geschwind, D.H., 2008. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. J Neurosci. 28**,** 1410-20.
Miller, J.A., Horvath, S., Geschwind, D.H., 2010. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc Natl Acad Sci U S A. 107**,** 12698-703.
Morris, R.G., Garrud, P., Rawlins, J.N., O'Keefe, J., 1982. Place navigation impaired in rats with hippocampal lesions. Nature. 297**,** 681-3.
Morris, R.G., Hagan, J.J., Rawlins, J.N., 1986. Allocentric spatial learning by hippocampectomised rats: a further test of the "spatial mapping" and "working memory" theories of hippocampal function. Q J Exp Psychol B. 38**,** 365-95.
Nachman, I., Regev, A., Friedman, N., 2004. Inferring quantitative models of regulatory networks from expression data. Bioinformatics. 20 Suppl 1**,** i248-56.
Nariai, N., Kim, S., Imoto, S., Miyano, S., 2004. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. Pac Symp Biocomput. 336-47.
Neapolitan, R.E., 2009. Probabilistic Methods for Bioinformatics, Vol., Elsevier, Morgan Kaufmann.
Needham, C.J., Bradford, J.R., Bulpitt, A.J., Westhead, D.R., 2007. A primer on learning in Bayesian networks for computational biology. PLoS Comput Biol. 3**,** e129.
Needham, C.J., Manfield, I.W., Bulpitt, A.J., Gilmartin, P.M., et al., 2009. From gene expression to gene regulatory networks in Arabidopsis thaliana. BMC Syst Biol. 3**,** 85.
Neves, G., Cooke, S.F., Bliss, T.V., 2008. Synaptic plasticity, memory and the hippocampus: a neural network approach to causality. Nat Rev Neurosci. 9**,** 65-75.
Niida, A., Smith, A.D., Imoto, S., Tsutsumi, S., et al., 2008. Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells. BMC Bioinformatics. 9**,** 404.
Oldham, M.C., Horvath, S., Geschwind, D.H., 2006. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci U S A. 103**,** 17973-8.
Olsen, C., Meyer, P.E., Bontempi, G., 2009. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. EURASIP J Bioinform Syst Biol. 308959.
Paul, C.M., Magda, G., Abel, S., 2009. Spatial memory: Theoretical basis and comparative review on experimental methods in rodents. Behav Brain Res. 203**,** 151-64.
Pawlowski, T.L., Bellush, L.L., Wright, A.W., Walker, J.P., et al., 2009. Hippocampal gene expression changes during age-related cognitive decline. Brain Res. 1256**,** 101-10.

Peleg, S., Sananbenesi, F., Zovoilis, A., Burkhardt, S., et al., 2010. Altered histone acetylation is associated with age-dependent memory impairment in mice. Science. 328**,** 753-6.
Penfold, C.A., Wild, D.L., 2011. How to infer gene networks from expression profiles, revisited. Interface Focus. 1**,** 857-70.
Penney, J., Tsai, L.H., 2014. Histone deacetylases in memory and cognition. Sci Signal. 7**,** re12.
Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., et al., 2003. Gene networks inference using dynamic Bayesian networks. Bioinformatics. 19 Suppl 2**,** ii138-48.
Poletti, M., Emre, M., Bonuccelli, U., 2011. Mild cognitive impairment and cognitive reserve in Parkinson's disease. Parkinsonism Relat Disord. 17**,** 579-86.
Rau, A., Jaffrezic, F., Foulley, J.L., Doerge, R.W., 2010. An empirical Bayesian method for estimating biological networks from temporal microarray data. Stat Appl Genet Mol Biol. 9**,** Article 9.
Recchia, A., Wit, E., Vinciotti, V., Kellam, P., 2008. Computational inference of replication and transcription activator regulator activity in herpesvirus from gene expression data. IET Syst Biol. 2**,** 385-96.
Rolls, E.T., 2013. The mechanisms for pattern completion and pattern separation in the hippocampus. Front Syst Neurosci. 7**,** 74.
Rowe, W.B., Blalock, E.M., Chen, K.C., Kadish, I., et al., 2007. Hippocampal expression analyses reveal selective association of immediate-early, neuroenergetic, and myelinogenic pathways with cognitive impairment in aged rats. J Neurosci. 27**,** 3098-110.
Ruan, J., Dean, A.K., Zhang, W., 2010. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol. 4**,** 8.
Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., et al., 2005. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 308**,** 523-9.
Schafer, J., Strimmer, K., 2005. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics. 21**,** 754-64.
Schimanski, L.A., Nguyen, P.V., 2004. Multidisciplinary approaches for investigating the mechanisms of hippocampus-dependent memory: a focus on inbred mouse strains. Neurosci Biobehav Rev. 28**,** 463-83.
Sharma, S., Rakoczy, S., Brown-Borg, H., 2010. Assessment of spatial memory in mice. Life Sci. 87**,** 521-36.
Small, S.A., Tsai, W.Y., DeLaPaz, R., Mayeux, R., et al., 2002. Imaging hippocampal function across the human life span: is memory decline normal or not? Ann Neurol. 51**,** 290-5.
Strimmer, K., Moulton, V., 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol Biol Evol. 17**,** 875-81.
Stuart, J.M., Segal, E., Koller, D., Kim, S.K., 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science. 302**,** 249-55.
Sweatt, J.D., 2010. Neuroscience. Epigenetics and cognitive aging. Science. 328**,** 701-2.
Tanzi, R.E., Bertram, L., 2001. New frontiers in Alzheimer's disease genetics. Neuron. 32**,** 181-4.
Tapia-Arancibia, L., Aliaga, E., Silhol, M., Arancibia, S., 2008. New insights into brain BDNF function in normal aging and Alzheimer disease. Brain Res Rev. 59**,** 201-20.

Torkamani, A., Dean, B., Schork, N.J., Thomas, E.A., 2010. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. Genome Res. 20**,** 403-12.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., et al., 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A. 100**,** 8348-53.

Uddin, R.K., Singh, S.M., 2013. Hippocampal gene expression meta-analysis identifies aging and age-associated spatial learning impairment (ASLI) genes and pathways. PLoS One. 8**,** e69768.

Verbitsky, M., Yonan, A.L., Malleret, G., Kandel, E.R., et al., 2004. Altered hippocampal transcript profile accompanies an age-related spatial memory deficit in mice. Learn Mem. 11**,** 253-60.

Villaverde, A.F., Banga, J.R., 2014. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. J R Soc Interface. 11**,** 20130505.

Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., et al., 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 474**,** 380-4.

Wang, M., Augusto Benedito, V., Xuechun Zhao, P., Udvardi, M., 2010. Inferring large-scale gene regulatory networks using a low-order constraint-based algorithm. Mol Biosyst. 6**,** 988-98.

Wang, Y., Zhang, X.S., Xia, Y., 2009. Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. Nucleic Acids Res. 37**,** 5943-58.

Watkinson, J., Liang, K.C., Wang, X., Zheng, T., et al., 2009. Inference of regulatory gene interactions from expression data using three-way mutual information. Ann N Y Acad Sci. 1158**,** 302-13.

Weaver, D.C., Workman, C.T., Stormo, G.D., 1999. Modeling regulatory networks with weight matrices. Pac Symp Biocomput. 112-23.

Werhli, A.V., Husmeier, D., 2007. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. Stat Appl Genet Mol Biol. 6**,** Article15.

Woolf, P.J., Prudhomme, W., Daheron, L., Daley, G.Q., et al., 2005. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. Bioinformatics. 21**,** 741-53.

Ye, H., Liu, W., 2015. Transcriptional networks implicated in human nonalcoholic fatty liver disease. Mol Genet Genomics.

Young, A., 1997. Ageing and physiological functions. Philos Trans R Soc Lond B Biol Sci. 352**,** 1837-43.

Zeier, Z., Madorsky, I., Xu, Y., Ogle, W.O., et al., 2011. Gene expression in the hippocampus: regionally specific effects of aging and caloric restriction. Mech Ageing Dev. 132**,** 8-19.

Zeilinger, M.N., Farre, E.M., Taylor, S.R., Kay, S.A., et al., 2006. A novel computational model of the circadian clock in Arabidopsis that incorporates PRR7 and PRR9. Mol Syst Biol. 2**,** 58.

Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 4**,** Article17.

Zoppoli, P., Morganella, S., Ceccarelli, M., 2010. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics. 11**,** 154.

Zou, M., Conzen, S.D., 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics. 21**,** 71-9.

Chapter 2 Data Collection and Preprocessing

2 Proper data collection and preprocessing of microarray gene expression data are critical for effective downstream analysis

2.1 Introduction

Microarray technology allows simultaneous measurement of the expression level of tens of thousands of genes in a single experiment (Lockhart et al., 1996; Schena et al., 1995). Gene expression analysis using microarrays allowed new insights into the cells and revolutionized research in many areas of biological science. Further, public microarray data repositories have allowed submission of data by independent scientists for thousands of gene expression studies. These data are available to the general public. This has created unique opportunities to undertake appropriate meta-analysis studies (Rung and Brazma, 2013; Tseng et al., 2012). Gene expression data from multiple relevant studies can be combined to obtain a more precise estimate of gene expression differentials and pathway signatures in the context of a well-designed biological problem. Meta-analysis of microarray gene expression data increases the statistical power to accurately and reliably characterize gene expression patterns (Ramasamy et al., 2008; Rodriguez-Zas et al., 2008). It has also proved to be highly useful in the field of gene network analysis. For example, Oldham et.al. (2006) investigated the functional organization of the transcriptome in distinct regions of the human brain. They assembled four independent microarray datasets generated from 160 human brain control samples. In a comparative meta-analysis, in order to better understand the neurodegenerative disease pathways, Miller et.al. (2010) merged data from over 1000 microarray samples from 18 human and 20 mouse datasets, representing various diseases, brain regions, study designs, and Affymetrix platforms.

One major challenge in combining microarray data across study and platform in meta-analysis is the issue of heterogeneity. Some of the major sources of heterogeneity

include: a) differences in the technology used in the study; b) differences in experimental design (subject or sample, and goal of the study); c) multiple different probes for the same gene; d) variability in probes used by different platforms; e) choice of preprocessing algorithms by the original study-authors; and f) differences in quantification of gene expression (Goldstein et al., 2010; Moreau et al., 2003; Ramasamy et al., 2008). These can make any form of processed data or result from such studies (e.g. published gene list, p-values, ranks, or processed expression matrix) unsuitable for use in a meta-analysis. (Ramasamy et al., 2008). For example, published gene lists (PGLs) generally represent only a subset of the genes actually studied, and information from many genes will be completely absent. Both the gene expression data matrix and PGL depend heavily on the choice of the preprocessing algorithm used. Furthermore, PGL is also affected by the choice of analysis method, the significance threshold, and the annotation builds used in the original study, all of which usually differ between studies. Therefore, it is recommended that feature-level extraction output or original raw expression data files (rather than preprocessed files) from different studies be used for any meta-analysis (Goldstein et al., 2010; Ramasamy et al., 2008). In case of Affymetrix microarrays the original raw expression data files are called CEL files.

Gene expression changes that are detected in a microarray could reflect selective, biologically relevant alterations in transcription level commonly referred to as biological variation or they could reflect variations caused by many kinds of experimental artifacts known as technical variation (Bolstad et al., 2003; Durinck, 2008; Pevsner, 2009; Talloen and Gohlmann, 2009). The sources for experimental (technical) variations include within-study variations such as the biological experiment (e.g. RNA isolation, RNA purity or quantity, tissue heterogeneity, inter-individual variation), microarray experiment (e.g. reverse transcription of mRNA, different labeling efficiency of fluorescently labeled nucleotides, print-tip effects, fluorescent scanner settings, signal measurement), manufacturing of the microarrays and all necessary reagents (e.g. batch-to-batch variation), and same experiment conducted in different labs (between-study variation). These variations may result in differences in brightness among slides. The purpose of

preprocessing is to deal with such unwanted technical artifacts. Preprocessing is performed on raw microarray data to remove the systematic bias in the data as much as possible while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription (Schuchhardt et al., 2000). Following array quality control, three key steps in preprocessing are background correction, normalization, and summarization.

Quality control can be performed at different stages of array preprocessing starting from the beginning (Ramasamy et al., 2008; Talloen and Gohlmann, 2009). Quality control involves both image quality and data quality assessment. Major experimental errors can be and should be detected early in the preprocessing by simply viewing microarray images using, for example, MAS 5.0 (Affymetrix, 2001), affyPLM (Bolstad, 2006), Simpleaffy (Wilson and Miller, 2005) or dChip software (Li and Wong, 2001a). However, quality control using affyPLM is the most recommended and adopted approach in the meta-analysis and microarray data user community (Goldstein et al., 2010).

Background correction is performed to remove background noise from the measured signal. Typical examples of non-biological background signal are nonspecific signals such as unspecific binding of transcripts, background signal from incomplete washing of the array, or background patterns across arrays, which have a drastic influence on the weak signal compared to the larger signal in terms of $\log_2$ fold changes (Talloen and Gohlmann, 2009). Normalization is a process that allows the comparison of gene expression levels among multiple microarrays in a single experiment or across multiple experiments, platforms, and studies. Summarization refers to the process in which information about multiple probes is integrated to yield a single measurement for the expression level of one transcript.

Various methods have been proposed for each of the preprocessing steps and thus offer a great number of possible combinations of choices (Lim et al., 2007; Talloen and

Gohlmann, 2009). This presents a difficult challenge to the microarray user community with regards to deciding on which method would perform the best. Some of the commonly used preprocessing procedures for Affymetrix microarrays are summarized in Table 2.1. Common background correction and normalization methods include linear scaling (implemented in Affymetrix MAS 5.0 software) (Affymetrix, 2001), quantile normalization (Bolstad et al., 2003), GC-RMA (Wu et al., 2004), variance stabilization normalization (VSN) (Huber et al., 2002), and rank-invariant normalization (Li and Wong, 2001a; Schadt et al., 2001; Stuart et al., 2001; Tseng et al., 2001). For Affymetrix arrays, there have been several approaches to summarizing probe-level data. Three of the most popular methods are Tukey biweight weighted average (implemented in Affymetrix MAS 5.0 software) (Affymetrix, 2001), model based expression index (MBEI) (Li and Wong, 2001b), and robust multi-array approach (RMA) (Irizarry et al., 2003a). Though each method has its own advantages and disadvantages, they have now become industry standard and have been implemented in the R BioConductor package (http://www.bioconductor.org/packages/release/bioc/html/affy.html).

Table 2.1 A summary of four commonly used preprocessing procedures, MAS5, MBEI, RMA, and GCRMA.

Procedure	Background correction	Normalization	Summarization	Reference
MAS5	Ideal (full or partial) MM subtraction	Linear scaling	Tukey biweight	(Hubbell et al., 2002)
RMA	Signal (exponential) and noise (normal) close-form transformation	Quantile	Median polish	(Irizarry et al., 2003b)
GCRMA	Optical noise, probe affinity and MM adjustment	Quantile	Median polish	(Wu et al., 2004)
MBEI	None	Invariant set	Multiplicative model fitting	(Li and Wong, 2001a)

Note. Table adopted from (Lim et al., 2007).

Several groups compared and evaluated the performance of different preprocessing (normalization and summarization) methods, particularly, MAS 5.0, MBEI, and RMA in the context of traditional differential expression analysis (Bolstad et al., 2003; Irizarry et al., 2003b; Irizarry et al., 2006) or gene network analysis (Harr and Schlotterer, 2006; Lim et al., 2007). Irizarry et al. (2003b) show that, the RMA method has a slight edge over MAS 5.0 and MBEI in terms of superior sensitivity and specificity (i.e. the true and false detection rate). Harr and Schlotterer (2006) concluded that for the detection of differentially expressed genes the RMA/GCRMA normalization methods are superior, and for network analysis of co-expressed genes within a single array the MBEI summary method performs significantly better. However, Lim et al. (2007) benchmarked four commonly used normalization procedures (MAS5, RMA, GCRMA and MBEI) (Table 2.1) in the context of reverse engineering of protein–protein and protein–DNA interactions and suggest that MAS5 provides the most faithful cellular network reconstruction. So, it appears that there is no 'golden standard' and no method is best under every circumstance (Cope et al., 2004; Goldstein and Guerra, 2010; Irizarry et al., 2006). All the methods in Table 2.1 generally perform equally well. The choice often depends on personal preference and the type of downstream application of the preprocessed data. Previous research suggests that correlations among co-regulated genes are sensitive to different processes during the normalization procedure (Harr and Schlotterer, 2006). Therefore, it is necessary to evaluate these methods before proceeding into any kind of meta- or gene network analysis.

However, often some of the technical/systematic variations described above, for example, reagents used from different lots or arrays handled by different technicians at different days cannot be corrected by the above normalization processes. The term "batch" generally refers to microarrays processed at one site over a short period of time using the same platform. The cumulative error introduced by these time and place-dependent experimental variations is referred to as "batch effects" (Chen et al., 2011; Leek et al., 2010). Batch effect is a particularly challenging problem when combining microarray data across study and platform. In gene expression studies, the greatest

source of differential expression is nearly always across batches rather than across biological groups, which can lead to confusing or incorrect biological conclusions owing to the influence of technical artefacts (Leek et al., 2010).

A number of batch effect identification and removal methods have been described in the literature for microarray data. These include, distance-weighted discrimination (DWD) (Benito et al., 2004), mean-centering (PAMR) (Sims et al., 2008), surrogate variable analysis (SVA) (Leek and Storey, 2007), geometric ratio-based method (Ratio_G) (Luo et al., 2010), an Empirical Bayes method called ComBat (Johnson et al., 2007), singular value decomposition (SVD) (Alter et al., 2000), standardization (Location/Scale adjustment model) (Li and Wong, 2001b), and a ratio-based method with arithmetic mean (Ratio_A) (Luo et al., 2010). The ComBat method has been found to perform equally well or better than most other approaches (Chen et al., 2011; Leek et al., 2010; Luo et al., 2010). Some methods (e.g. DWD or SVD) require large numbers of samples (e.g. more than 25) and can process only two batches at a time (Johnson et al., 2007). However, ComBat does not have those limitations. ComBat also does not affect correlation in cross-platform normalization (Sirbu et al., 2010). This method has been used to successfully remove batch effects in some recent microarray publications (Konstantinopoulos et al., 2011; Larsen et al., 2014; Stein et al., 2015). The ComBat method has also been identified as the preferred method for between-study, cross-study, or cross-platform normalization (Chen et al., 2011; Leek et al., 2010; Luo et al., 2010). Thus, ComBat becomes a better choice to apply in this study to remove batch effects as part of the preprocessing steps.

Therefore, variation in the way microarray studies are conducted makes it critical to carefully consider the sources of heterogeneity when selecting data for large-scale meta-analysis. Selecting datasets with available CEL files would allow all data to be subjected to the same rigorous quality control check. Further, the availability of several preprocessing methods requires assessing these methods and identifying a method that will perform satisfactorily. Using CEL files will help to remove any systematic differences

and all good quality arrays can then be preprocessed consistently using the same procedure for all studies. Therefore, this chapter examines objective one, specifically, the goals are: 1) to perform selection, collection, and quality control of ASLI gene expression data; and 2) to assess several preprocessing methods, identify a method, and to apply the method on the data.

2.2 Methods

2.2.1 Data collection and selection

I primarily used the GEO (http://www.ncbi.nlm.nih.gov/geo/) and the ArrayExpress (http://www.ebi.ac.uk/array express/) microarray data repositories to search for microarray gene expression datasets using the keyword "memory and brain". I also used the PubMed literature database to search for relevant studies (Figure 2.1). Careful review of the published articles referencing these data revealed that the goals of these studies were varied, and included many different learning paradigms, test conditions, and different tissue types. This observation necessitated the establishment of some data selection criteria for any downstream analysis in order to minimize heterogeneity among datasets and to obtain biologically meaningful results. Therefore, in this research I followed a conservative data selection process (Table 2.2). I focused on datasets generated from carefully designed behavioral studies involving hippocampus dependent ASLI in Fischer 344 strain of male rats (*Rattus norvegicus*) using Affymetrix® expression arrays. These selected studies investigated the spatial learning tasks in young, adult, and/or old animals using only the Morris water maze as the training and assessment protocol. Affymetrix raw data (CEL files) for the selected studies were either directly downloaded from the GEO website or obtained through personal communication with the original authors.

2.2.2 Quality control

All arrays were first assessed for image quality using dChip software (Li and Wong, 2001a) (http://biosun1.harvard.edu/complab/dchip/). Minor contaminations present in

a few of the arrays were corrected using the built in image gradient correction algorithm in dChip by adjusting the background brightness of the contaminated area to a level similar to the background of the surrounding clean region.

All subsequent data preparation, preprocessing, and statistical analyses were performed in R (http://cran.r-project.org/, a freely available programming language), using appropriate software packages. The data quality was assessed using RNA degradation ratios, relative log expression (RLE), and normalized unscaled standard errors (NUSE) plots using the *simpleaffy* and *affyPLM*, packages in Bioconductor (http://www.bioconductor.org/) and the RMAExpress software in R following standard procedures (Bolstad et al., 2005).

Table 2.2 Data selection criteria.

Selection Category	Criteria		
Learning paradigm	Spatial learning		
Training and Diagnostic protocol	Morris water maze		
Species/strain	Rat (Rattus norvegicus) – male Fischer 344 strain		
Age category	Young	Adult	Old
Age	3 – 6 months	9 – 14 months	24 – 26 months
Tissue/RNA	Hippocampus total RNA		
Microarray platform	Affymetrix®		
Microarray experiment and data standard	MIAME (Minimum Information About a Microarray Experiment, http://fged.org/projects/miame/)		

2.2.3 Data preprocessing for meta-analysis

I performed an initial evaluation of five different normalization methods, which were MAS5, RMA, MBEI PM only, MBEI PM – MM, and a recently developed single channel microarray normalization method called SCAN (Piccolo et al., 2012). The question was which normalization method would remove batch effects most effectively. For this purpose, each dataset was normalized with the above methods and then subjected to ComBat batch correction. RMA methods removed batch effects comparatively better

than all other methods consistently in all five datasets (result not shown), and was therefore chosen to perform all preprocessing in this research.

The overall data preprocessing steps are shown diagrammatically in Figure 2.1. Within-study normalization and expression measurement were performed using the RMA methods (Bolstad et al., 2003) with default options in the *affy* package in R (Gautier et al., 2004). Within-study batch correction was performed using the ComBat method (Johnson et al., 2007). Array hybridization dates were retrieved from CEL files and used as processing batches to perform batch correction. Age and spatial learning impairment were used as covariates. It was made sure that each group is well represented in each study during batch correction, even after removal of bad or outlier arrays.

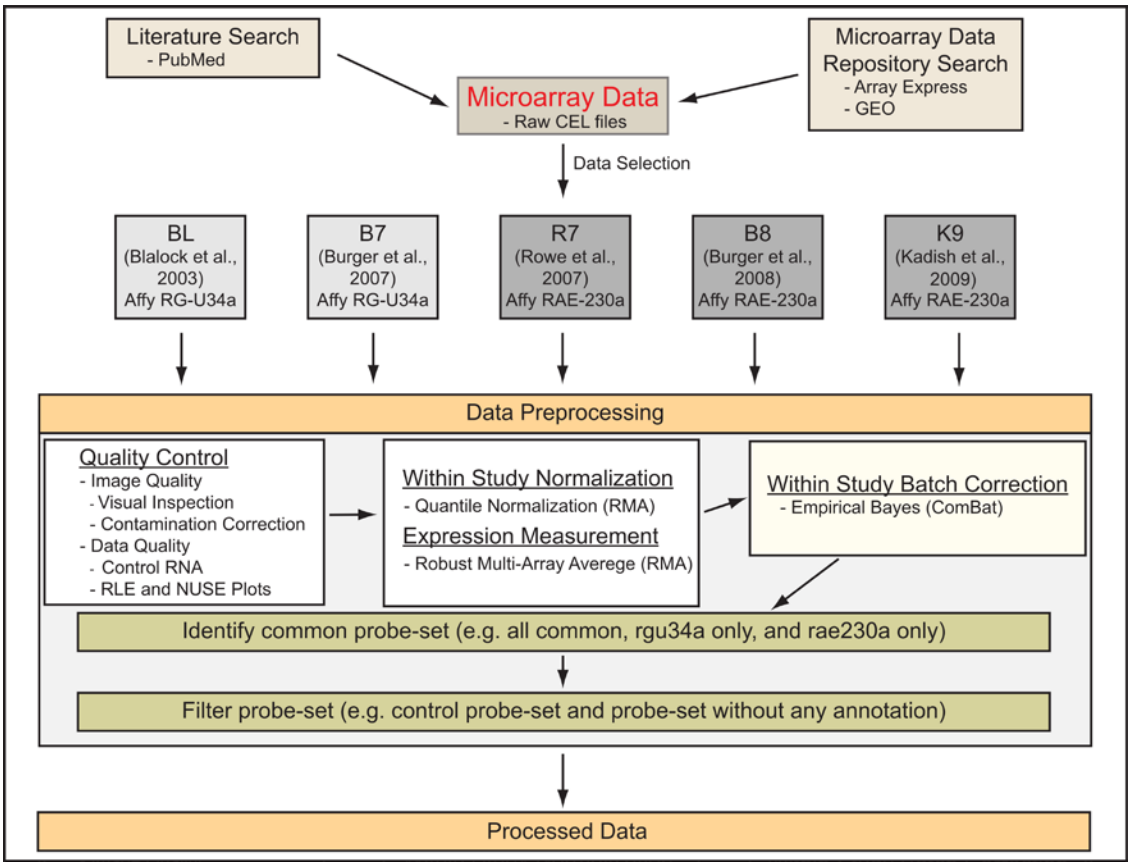


Figure 2.1 Data preprocessing workflow for meta-analysis.

2.2.4 Combining data for meta-analysis: common probe set approach

A gene can have multiple probe sets or often the same probe set can be associated with different gene symbols due to changes or updates in the databases. As a result, gene names or symbols do not serve as a good ID to combine data across microarray platforms. Therefore, in preparation to combine data across two different platforms (i.e. RAE230A and RGU34A) I decided to combine data at the probe set level rather than at the gene level.

A common probe set file that contains best matching pairs of probe sets representing the same gene in the two chip types (i.e. RGU34A and RAE230A) was downloaded from the Affymetrix website (www.affymetrix.com). Applying the common file and the *genefilter* package in R, probe sets from all studies belonging to the two different chip types were merged into three categories as follows: i) rgu_exclusive, probe sets exclusive to the RGU34A chip type, ii) all5_common, probe sets common among all five studies, and iii) rae_exclusive, probe sets exclusive to the RAE230A chip type. Control probe sets and probe sets without any annotation were filtered out.

2.2.5 Data preprocessing for network analysis

Network analysis and module detection can be severely biased by the presence of outlying microarray samples (Miller et al., 2010; Oldham et al., 2008). So, it is important to identify and remove such samples in each dataset during the pre-processing steps prior to network construction. Moreover, it is often meaningful to reduce the number of genes (to most connected genes) for network analysis; otherwise it may become computationally very intensive. Therefore, data selected for network analysis underwent additional preprocessing steps. All datasets were processed identically for consistency and the overall process is described as follows.

- Removal of outlier array
- Data normalization and batch correction
- Filtering of unwanted probe sets

2.2.5.1 Removal of outlier array

For each dataset, original microarray CEL files were read into R, background corrected using the RMA method in the *affy* package (http://www.bioconductor.org/packages/release/bioc/html/affy.html) and initial un-normalized expression matrices were created. Outlier samples were removed using the inter-array correlation (IAC) approach as described previously (Miller et al., 2010; Oldham et al., 2008). Briefly, IAC was defined as the Pearson correlation coefficient of the expression levels for a given pair of microarrays (using all probe sets). The distribution of IACs within a dataset was visualized as a histogram (frequency plot), while the relationships between arrays were visualized as a dendrogram using average linkage hierarchical clustering with 1-IAC as a distance metric. Samples with low mean IACs (i.e. arrays with mean IAC more than two to three standard deviations below average) and/or samples that exhibited divergent clustering were excluded. This process was repeated until no outlier arrays remained.

2.2.5.2 Data normalization and batch correction

Following outlier removal, absence and presence call information for all probe sets were extracted directly from CEL files using the mas5calls() function in the *affy* package in R. Probe sets that were called "absent" in more than 90% of the samples were filtered out. Next, RMA quantile normalization was performed on each dataset as described before. Batch effect was removed from each dataset using the ComBat batch correction method as described for meta-analysis.

2.2.5.3 Filtering of unwanted probe sets

Unwanted probe sets include control probe sets and those not associated with known genes and were removed. Next, the genefilter package (http://www.bioconductor.org/packages/release/bioc/html/genefilter.html) was used to keep only the probe sets that were associated with some genes (i.e. probe sets for which annotation was available). Many genes contain more than one probe set. To

allow comparison across Affymetrix platforms, only a single probe set for each gene was kept by using a function (CollapseGenesRai(…) in Appendix 6.2.1) modified from (Miller et al., 2010). For this purpose, if a gene contained two or more probe sets, the probe set with the highest connectivity across samples was kept. The remaining probe sets were used for gene network construction using WGCNA.

2.3 Results

2.3.1 Data collection and selection

A search in the ArrayExpress and the GEO public microarray data repositories reveals that there is a large body of microarray data available involving cognitive impairments (Figure 2.2). This search identified 38 unique studies with over 800 assays for rats involving cognitive impairments in the brain. Review of these data and their associated published articles revealed that these studies investigated spatial and associative learning impairments in the brain using the Morris water maze or fear conditioning assessment protocols, respectively, with or without the effect of aging. They also include different tissue types, drug responses, candidate genes, effects of aging alone, effects of different spatial learning tests, or the effects of specific neurodegenerative diseases (e.g. Alzheimer's disease). After careful examination of these datasets and using my data selection criteria (Table 2.2), I have identified five individual studies that investigated only hippocampus dependent ASLI as assessed by the Morris water maze test. Data from these five studies consist of a total of 287 arrays (one animal per assay), which used two different Affymetrix chip types, RG_U34a and RAE230A (Table 2.3).

The data represented young and aged rats that were learning unimpaired and aged rats that were learning impaired from a set of results published during 2003 to 2009. The selected datasets will be referred in this study as BL (Blalock et al., 2003), B7 (Burger et al., 2007), R7 (Rowe et al., 2007), B8 (Burger et al., 2008), and K9 (Kadish et al., 2009). These data would allow one to assess combined gene expression changes related to aging, as well as ASLI in rats across multiple studies. These studies investigated spatial

learning tasks in young (3 – 6 months old), adults (9 – 14 months old), and aged (24 – 26 months old) animals using the Morris water maze as the training and assessment protocol. However, the adult animals were not included in this analysis. The BL and K9 studies were similar in design where only the unimpaired young and impaired aged animals were considered for comparison. The B7, R7, and B8 studies were similar in design where both young and aged groups had impaired and unimpaired animals as well as additional controls (e.g. cage controls, stress controls, and controls for visual impairment). A total of 265 arrays were finally selected following a quality assessment.

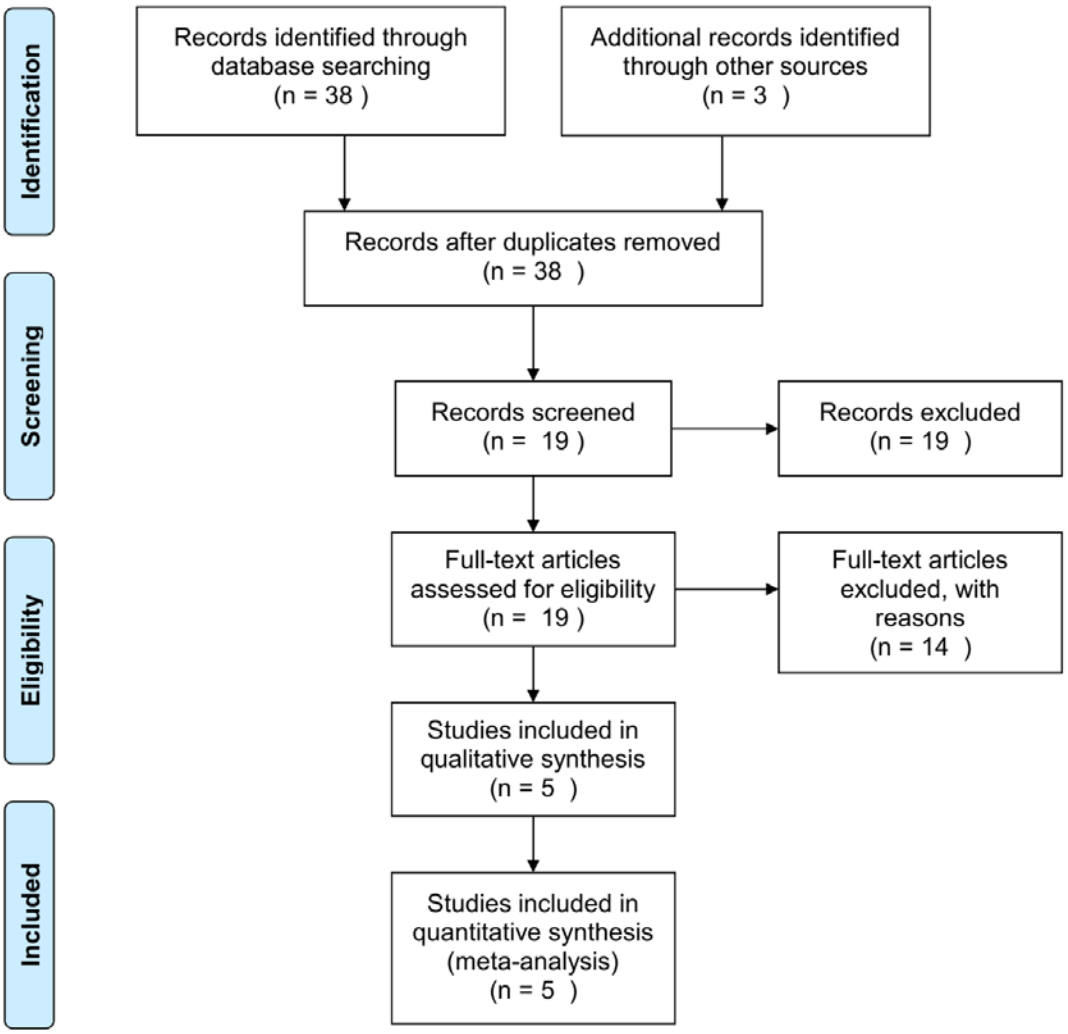


Figure 2.2 Data selection process. Search in the public microarray data repositories identified 38 microarray datasets involving cognitive impairments. I excluded 19 datasets that were either not relevant to this study or were not associated with any publication. I excluded 14 more studies as they involved different learning paradigms, test conditions, and outcomes in mice. I finally selected five studies that dealt with hippocampus dependent age-associated spatial learning in rats.

Table 2.3 Age-associated spatial learning impairment (ASLI) datasets for rats.

Dataset ID	Reference	Affymetrix Array Type	Number of Assays (one animal/array)
BL	Blalock et al. 2003 (Blalock et al., 2003)	RG_U34	29
B7	Burger et. al. 2007 (Burger et al., 2007)	RG_U34	79
R7	Rowe et. al. 2007 (Rowe et al., 2007)	RAE230A	50
B8	Burger et. al. 2008 (Burger et al., 2008)	RAE230A	80
K9	Kadish et al. 2009 (Kadish et al., 2009)	RAE230A	49

2.3.2 Quality control

Image analysis in dChip software identified several arrays with minor contamination or spots as seen in Figure 2.3. If not corrected properly, such contamination can affect any downstream normalization and expression level comparison. Any array with contaminated spots was corrected using dChip (Li and Wong, 2001b) resulting in the creation of a new CEL file that was used in all subsequent analysis. Arrays displaying major hybridization problems (e.g. variable background brightness, uneven hybridization) were discarded (Figure 2.4). The remaining arrays were checked for their data quality based on the β-actin 3'/5' and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) 3'/5' ratios (Figure 2.5 and Figure 2.6) using the *simpleaffy* package in R (http://www.bioconductor.org/packages/release/bioc/html/simpleaffy.html), as well as RLE-NUSE T2 plots (Figure 2.7 and Figure 2.8) using the RMAExpress (http://rmaexpress.bmbolstad.com/). The 3'/5' ratios for β-actin and GAPDH for BL, R7, and K9 arrays mostly fell within -1 to +2 (Figure 2.5). For B7 and B8, these ratios ranged from -2 to over +3 for some arrays (Figure 2.6). Figure 2.7 and Figure 2.8 show RLE-NUSE plots for B8 and B7 datasets which had few bad quality arrays. A number of arrays in these two datasets did not meet the quality requirement and were removed. For

example, B8 arrays 76, 34, and 24 were above 99% cutoff (solid red line) and were removed (Figure 2.7). Arrays 56, 93, and 80 were between the 95% (dotted red line) and 99% cutoff. Arrays 93 and 80 also contained very high probe outliers (data not shown) and were removed. However, array 56 was not removed because it had only ~4% probe outliers (data not shown). For B7 (Figure 2.8), arrays 8 and 83 were above 99% cutoff (solid red line) and arrays 97, 39, and 30 were between the 95% cutoff (dotted red line) and 99% cutoff. These five arrays were also removed from further consideration.

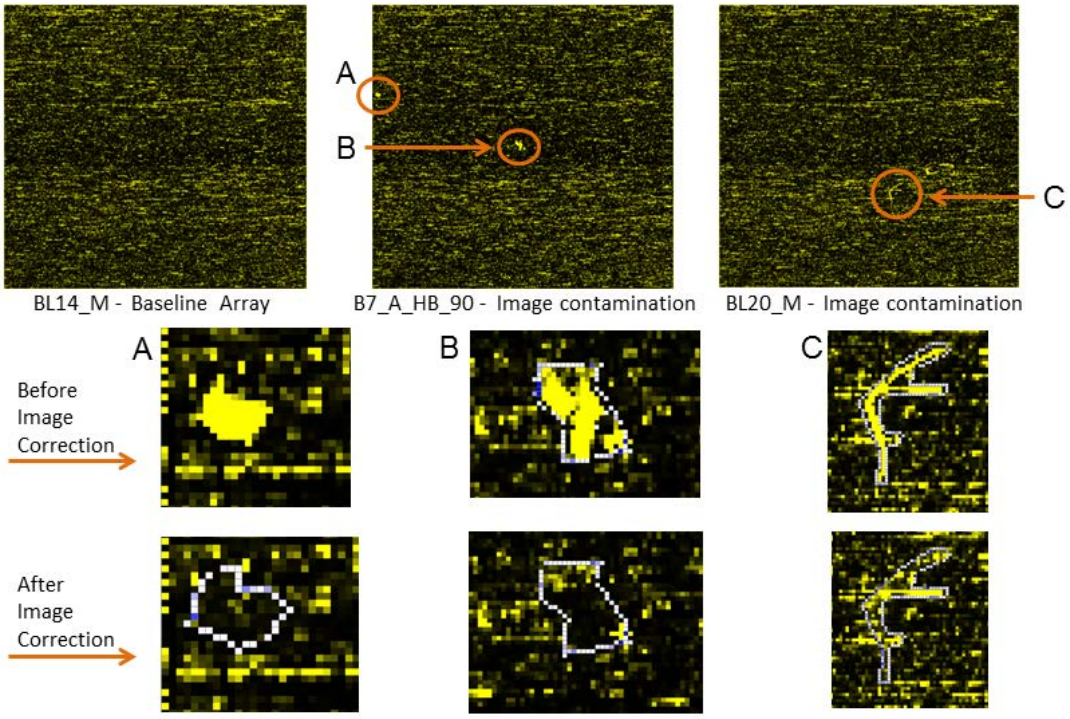
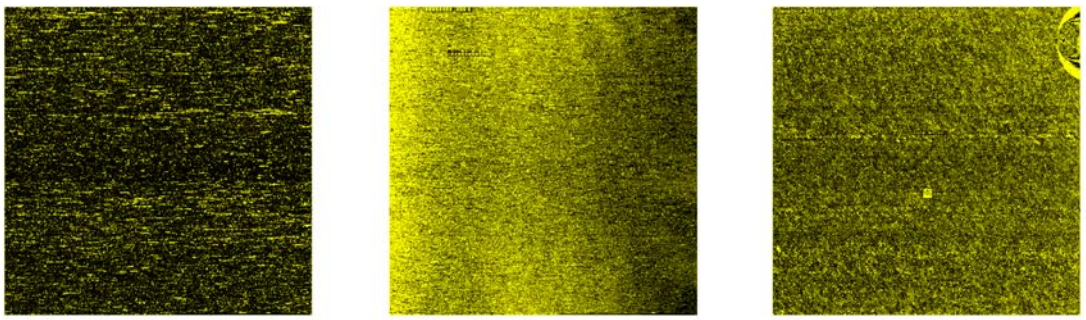


Figure 2.3 Image contamination corrections using the image gradient correction algorithm in dChip. Three contaminated areas (A, B, and C) from two representative arrays are shown in the top panel. The bottom two panels show enlarged views of these areas before and after image contamination correction. In this process, each area was outlined in dChip (middle panel) and the background brightness of the contaminated area was adjusted to a level similar to the background of the surrounding clean region (bottom panel).

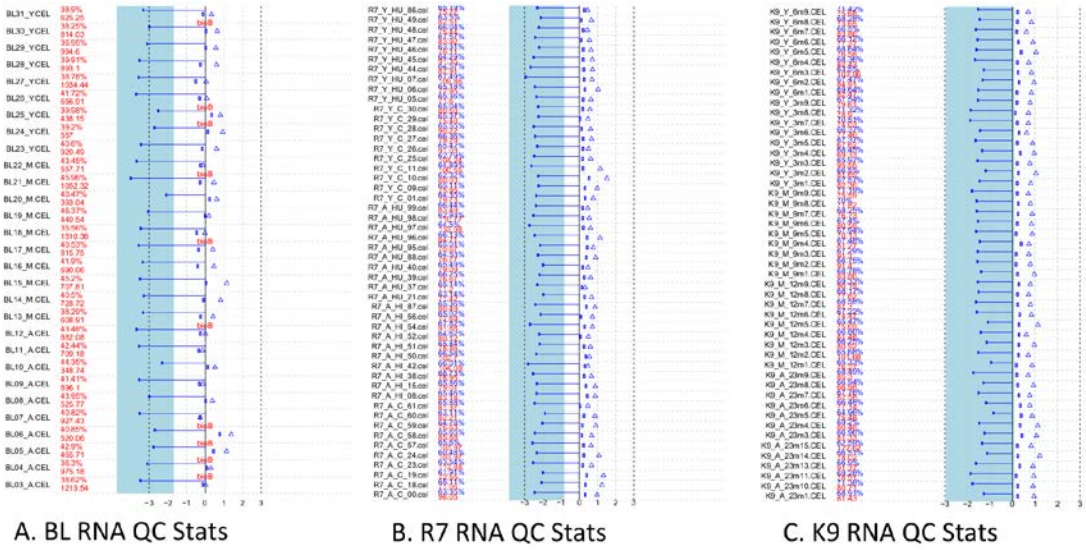


A. A base line array (BL_14_M)

B. Image too bright (B7_A_HI_83)

C. Image too bright and messy (B8_Y_H_80)

Figure 2.4 Example of bad quality arrays. : A) a base line good quality array, B) an array image that is too bright, and C) an array with image defects that is also too bright.



A. BL RNA QC Stats

B. R7 RNA QC Stats

C. K9 RNA QC Stats

Figure 2.5 RNA quality assessments of BL, R7, and K9 datasets. For each array, the corresponding triangle represents β-actin 3'/5' ratio and the circle represents GAPDH 3'/5' ratio. For each dataset, the outer (at +3) and inner (at -3) vertical dotted lines represent the recommended ratio boundaries.

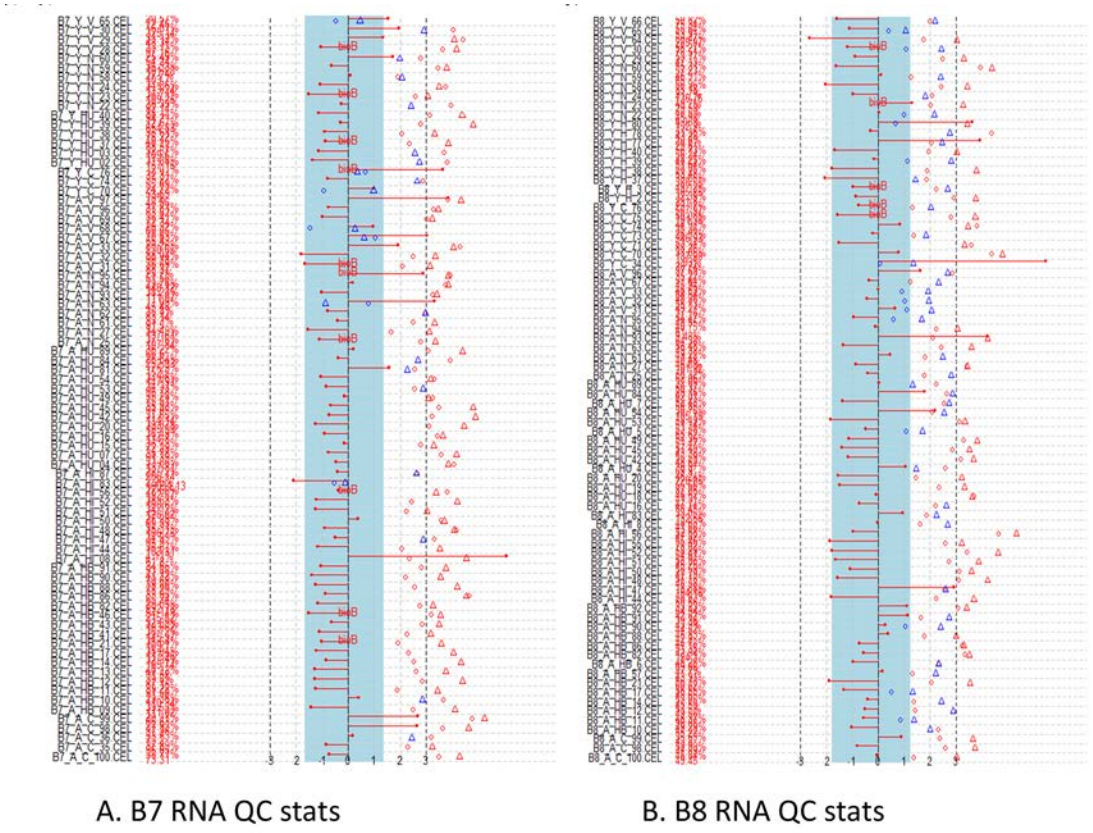


Figure 2.6 RNA quality assessments of B7 and B8 datasets. For each array, the corresponding triangle represents β-actin 3’/5’ ratio and the circle represents GAPDH 3’/5’ ratio. For each dataset, the outer (at +3) and inner (at -3) vertical dotted lines represent the recommended ratio boundaries.

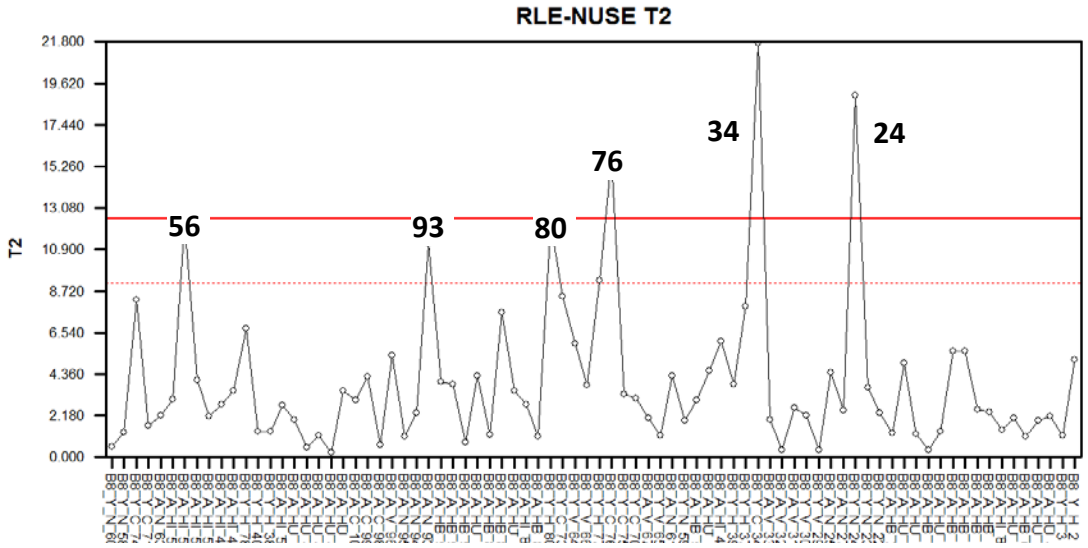


Figure 2.7 B8 RLE-NUSE plot using RMAExpress. In this figure, arrays 76, 34, and 24 are above 99% cutoff (solid red line). Arrays 56, 93, and 80 are just below the 99% cutoff but above 95% cutoff (dotted red line).

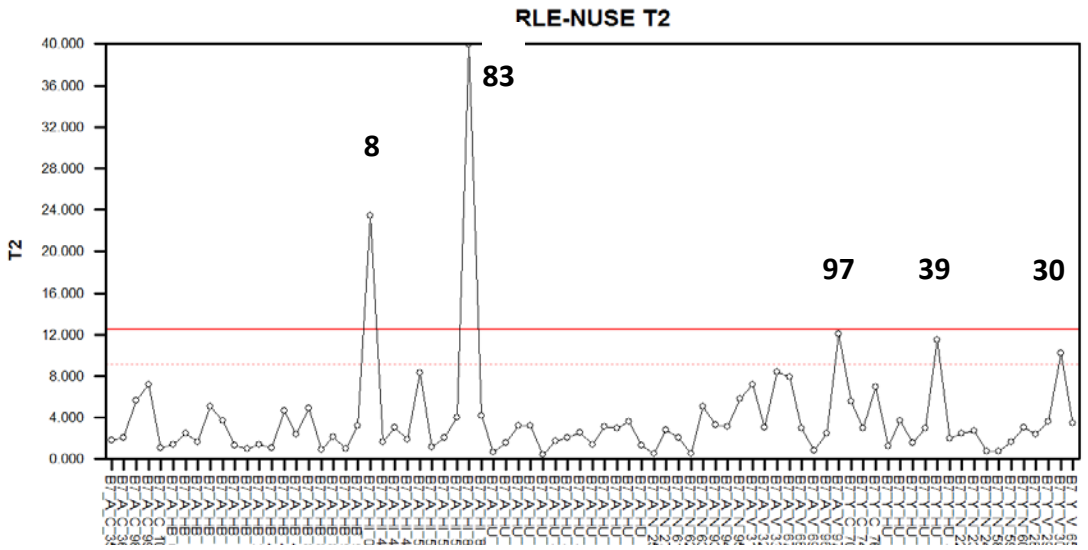


Figure 2.8 B7 RLE-NUSE plot using RMAExpress. In this figure, arrays 8 and 83 are above 99% cutoff (solid red line). Arrays 97, 39, and 30 are just below the 99% cutoff but above 95% cutoff (dotted red line).

2.3.3 Data preprocessing for meta-analysis

Within-study normalization was performed on five selected datasets (Table 2.3). Figure 2.9 to Figure 2.13 show the boxplots of the arrays in each dataset before and after RMA normalization. The results show that the RMA method was able to properly normalize the datasets with reference to the baseline array. However, hierarchical clustering analysis performed on the normalized data shows that batch effects are clearly evident in all studies even after normalization, though at variable degrees. Arrays that were hybridized on the same date as a batch (represented by the same color) are clustered together in the dendrograms (Figure 2.14 to Figure 2.18). I used ComBat to remove batch effects. Batch effects were completely removed from the BL, B7, and K9 data and significantly removed from the B7 and B8 data, as they clustered together more based on their phenotypes such as aged or young (Figure 2.14 to Figure 2.18).

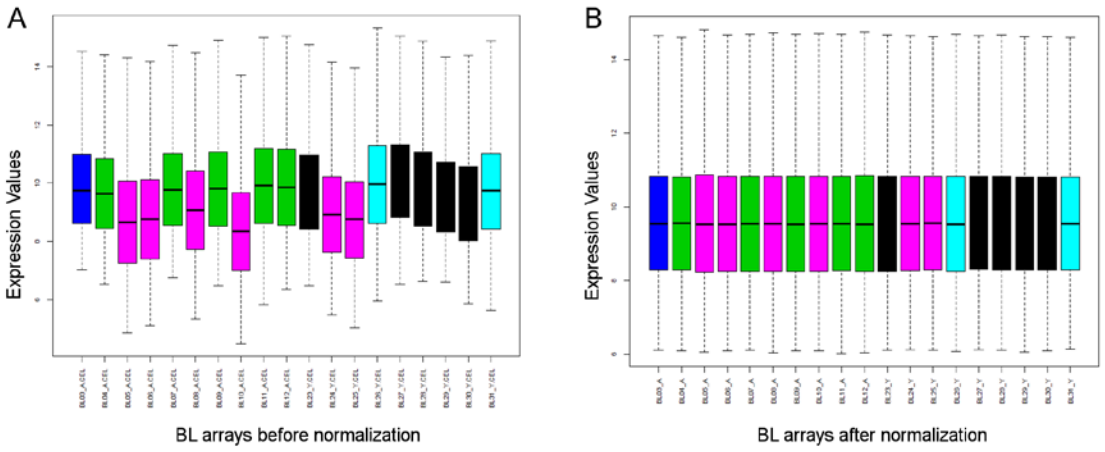


Figure 2.9 Boxplots of BL dataset before (A) and after (B) RMA normalization. Each color represents a batch of arrays that were hybridized and processed at the same time.

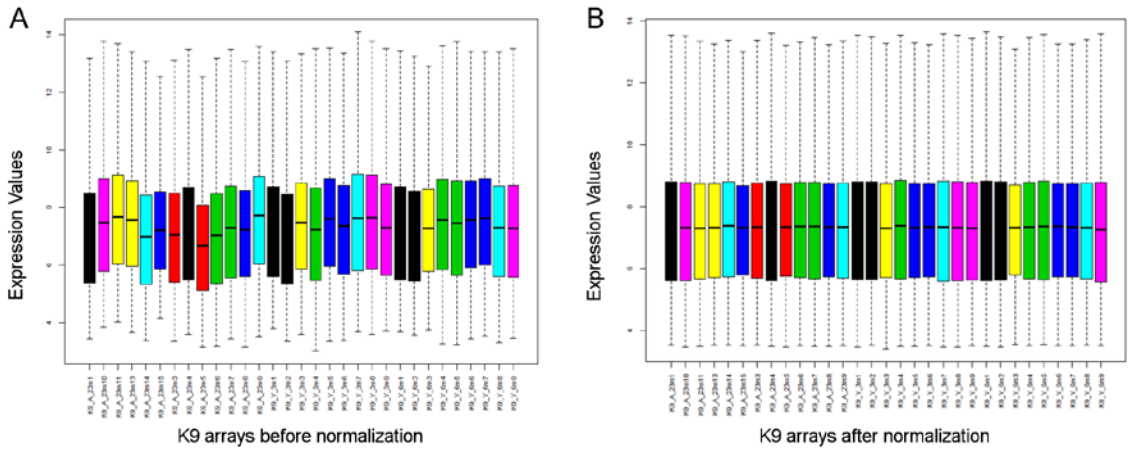


Figure 2.10 Boxplots of K9 dataset before (A) and after (B) RMA normalization. Each color represents a batch of arrays that were hybridized and processed at the same time.

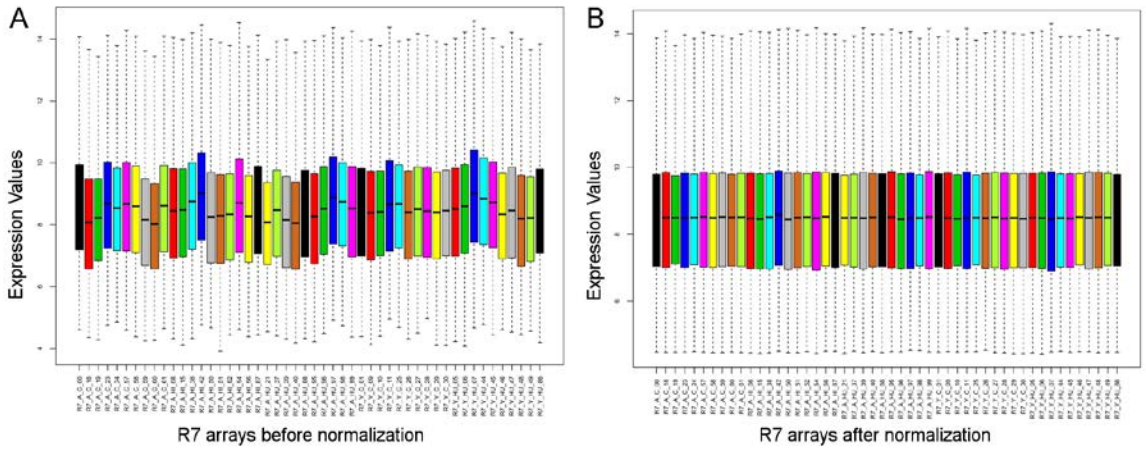


Figure 2.11 Boxplots of R7 dataset before (A) and after (B) RMA normalization. Each color represents a batch of arrays that were hybridized and processed at the same time.

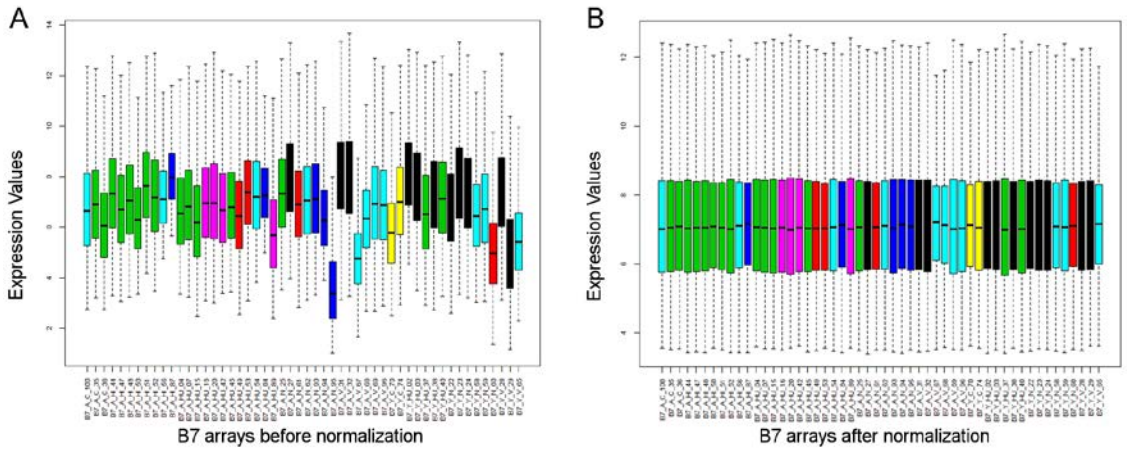


Figure 2.12 Boxplots of B7 dataset before (A) and after (B) RMA normalization. Each color represents a batch of arrays that were hybridized and processed at the same time.

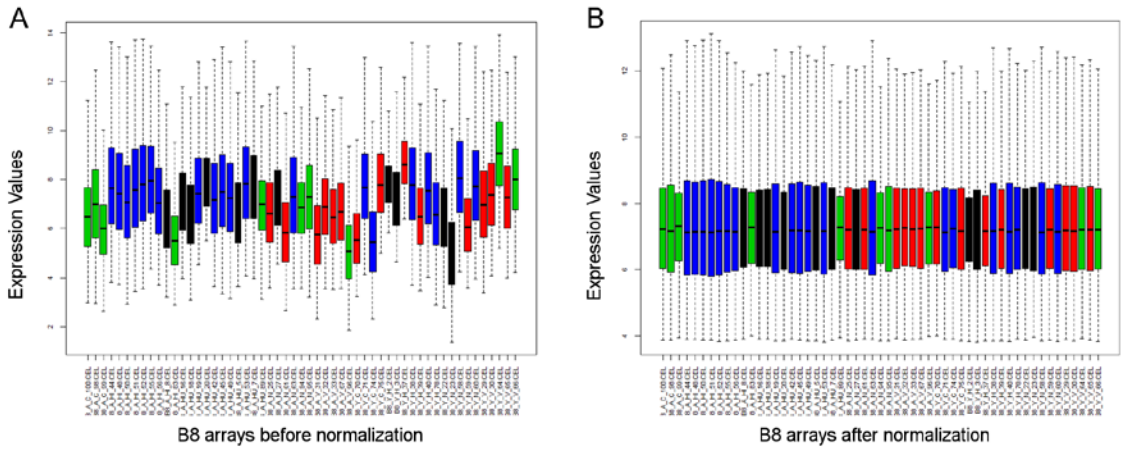


Figure 2.13 Boxplots of B8 dataset before (A) and after (B) RMA normalization. Each color represents a batch of arrays that were hybridized and processed at the same time.

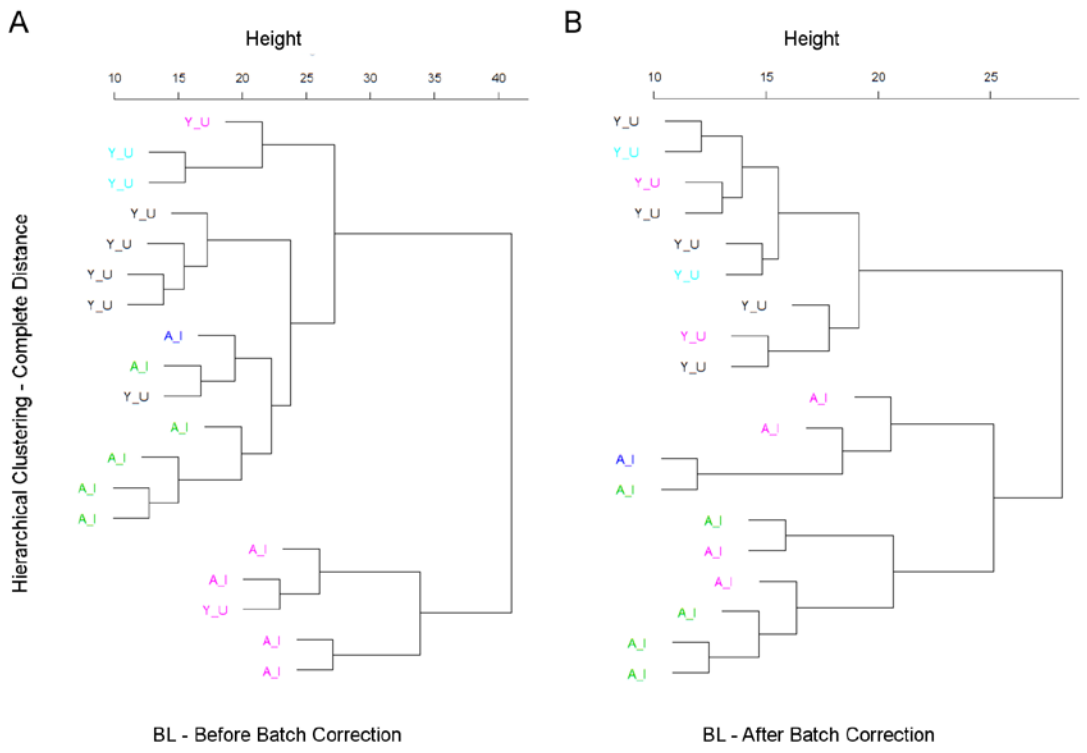


Figure 2.14 Hierarchical clustering of RMA normalized BL data. Each color represents a batch of arrays, which were hybridized and processed at the same time. Batch effects are evident even after normalization and before batch adjustment (A) as arrays are mostly clustered in batches (same color). However, following Empirical Bayes adjustment, arrays are clustered based on aged and young phenotypes irrespective of batches (B). Leaf labels: A, aged; Y, young; I, impaired; U, unimpaired.

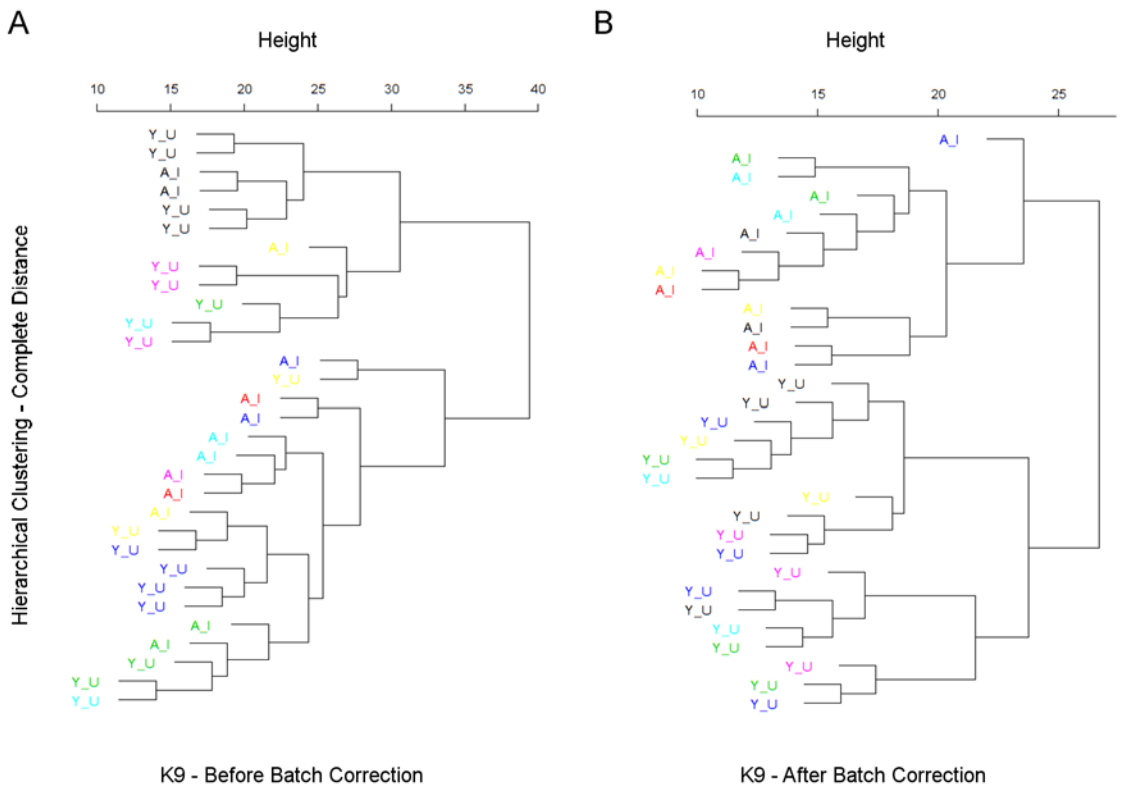


Figure 2.15 Hierarchical clustering of RMA normalized K9 data. Each color represents a batch of arrays, which were hybridized and processed at the same time. Batch effects are evident even after normalization and before batch adjustment (A) as arrays are mostly clustered in batches (same color). However, following Empirical Bayes adjustment, arrays are clustered based on aged and young phenotypes irrespective of batches (B). Leaf labels: A, aged; Y, young; I, impaired; U, unimpaired.

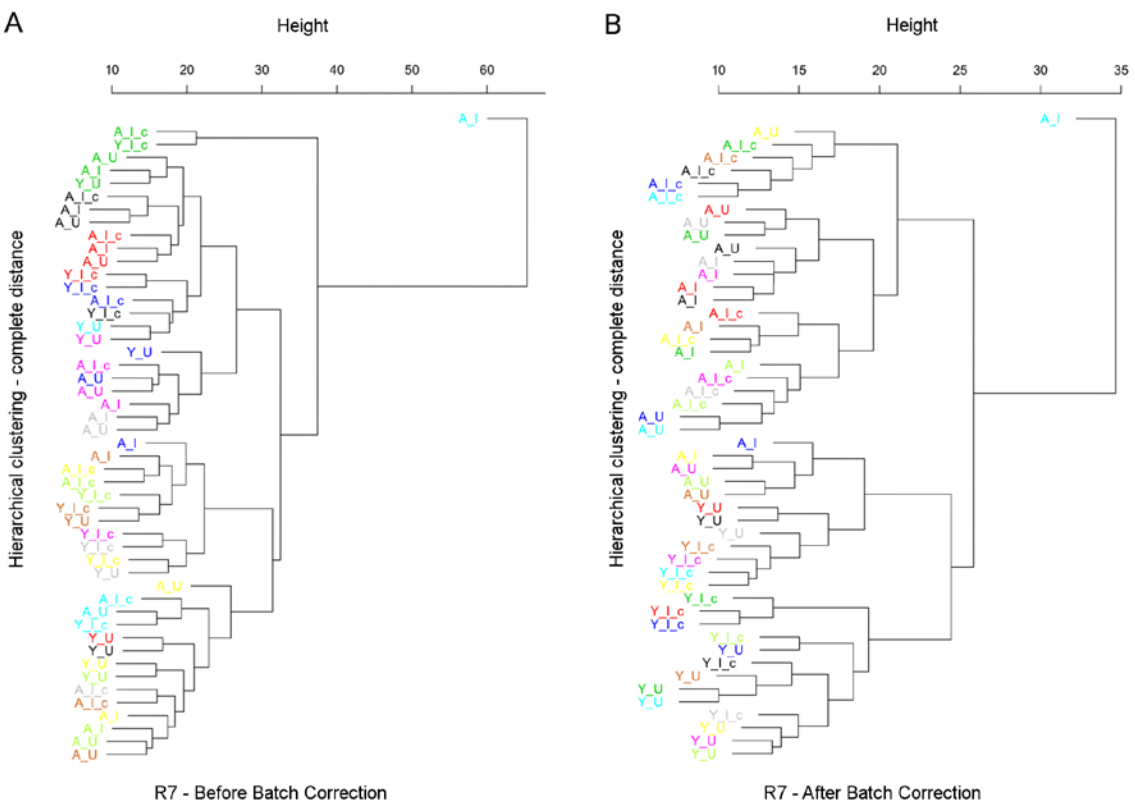


Figure 2.16 Hierarchical clustering of RMA normalized R7 data. Each color represents a batch of arrays, which were hybridized and processed at the same time. Batch effects are evident even after normalization and before batch adjustment (A) as arrays are mostly clustered in batches (same color). However, following Empirical Bayes adjustment, arrays are clustered based on aged and young phenotypes irrespective of batches (B). Leaf labels: A, aged; Y, young; I, impaired; U, unimpaired; c, control.

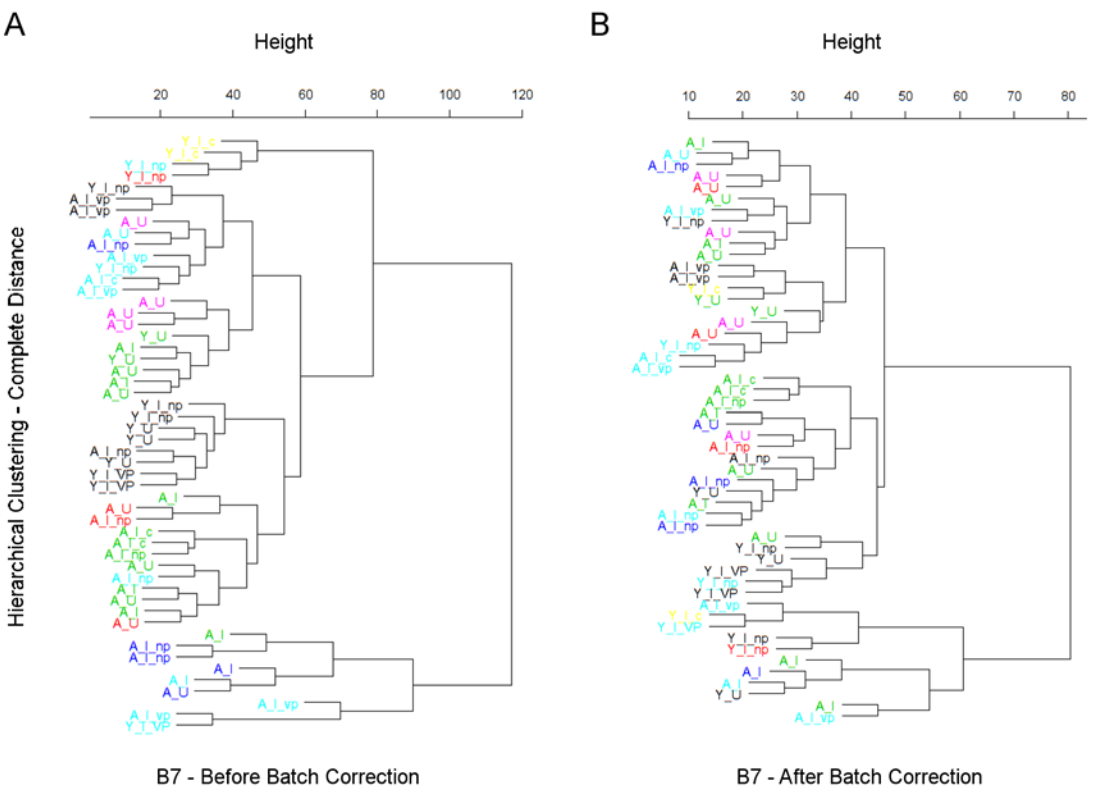


Figure 2.17 Hierarchical clustering of RMA normalized B7 data. Each color represents a batch of arrays, which were hybridized and processed at the same time. Batch effects are evident even after normalization and before batch adjustment (A) as arrays are mostly clustered in batches (same color). However, following Empirical Bayes adjustment, batch effects have improved (B). Leaf labels: A, aged; Y, young; I, impaired; U, unimpaired; c, control.

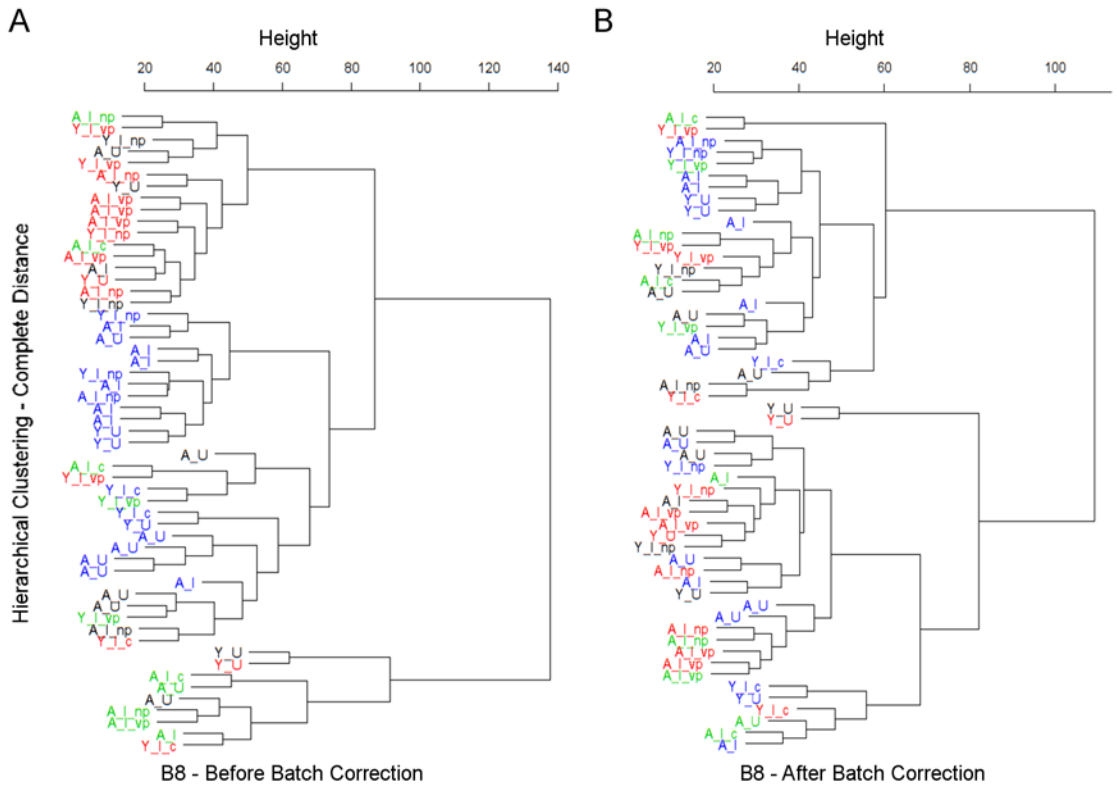


Figure 2.18 Hierarchical clustering of RMA normalized B8 data. Each color represents a batch of arrays, which were hybridized and processed at the same time. Batch effects are evident even after normalization and before batch adjustment (A) as arrays are mostly clustered in batches (same color). However, following Empirical Bayes adjustment, batch effects have improved (B). Leaf labels: A, aged; Y, young; I, impaired; U, unimpaired; c, control.

2.3.4 Data preprocessing for network analysis

2.3.4.1 Removal of outlier array

Outlier samples were removed from the unnormalized expression data based on the IAC method. As an example, typical results obtained from this procedure are described below for R7 dataset. This dataset contained a total of 50 arrays. The IAC histogram (Figure 2.19-A) showed that the mean IAC for these 50 arrays was 0.994, which was very good. However, the distribution of arrays was not bell shaped, which indicated the presence of outlier samples.

One way to view the outlier samples is by performing average linkage hierarchical clustering using 1 – IAC as a distance metric. Another way to visualize outliers is to calculate the mean IAC for each array and examine this distribution in a scatterplot. Hierarchical clustering (HC) (Figure 2.20 A) showed that the sample R7_A_HI_38 indicated by the first branch in the HC was an obvious outlier. In the scatterplot (Figure 2.20 B), the same outlier was visible seven standard deviations below the mean IAC.

After removing the sample R7_A_HI_38 a new IAC matrix was calculated with the remaining 49 arrays. The resulting histogram of IAC (Figure 2.19 B) showed that the mean IAC improved slightly to 0.995. However, the IAC HC dendrogram and mean IAC scatterplot revealed the presence of three more outliers, which were R7_A_C_19, R7_A_HU_21, and R7_Y_C_10 (Figure 2.21 A). These outliers were 2 to 3.5 standard deviations below the mean IAC (Figure 2.21 B). Next, these three outlier samples were removed and a new IAC matrix was calculated. The result (Figure 2.22) did not show the presence of any new outliers.

Once outlier arrays were removed from R7 dataset 46 arrays remained. The outlier removal process was repeated for other datasets such as B8 (Figure 2.23), K9 (Figure 2.24), B7 (Figure 2.25), and BL (Figure 2.26). The final number of arrays that remained for each dataset is shown in Table 2.4.

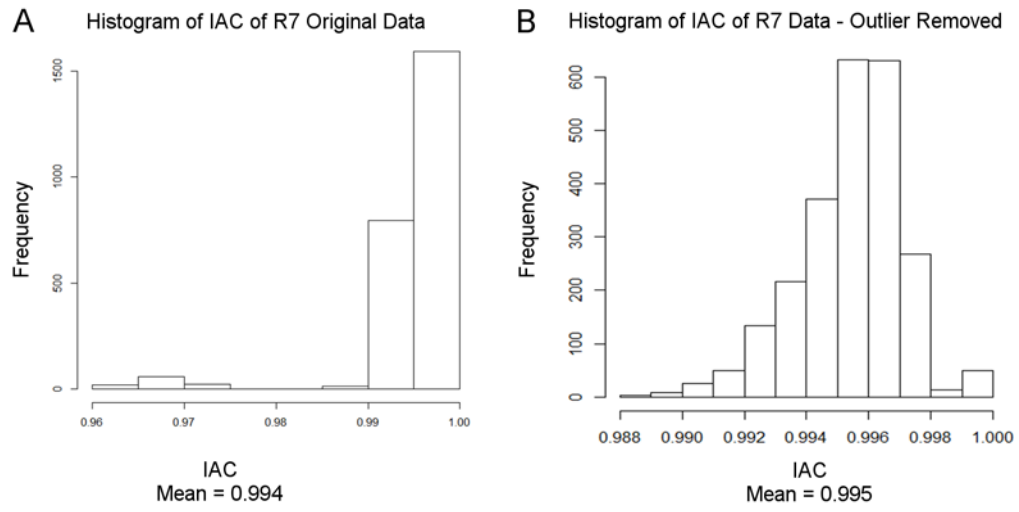


Figure 2.19 Checking outlier arrays in R7 data. A. Histogram of IAC of the unnormalized R7 dataset with no outlier samples removed. The mean IAC is 0.994. The distribution is skewed due to the presence of outlier samples. B. Histogram of IAC of the same dataset after removing one outlier sample. The mean IAC is 0.995 but the distribution is still skewed due to the presence of more outlier samples.

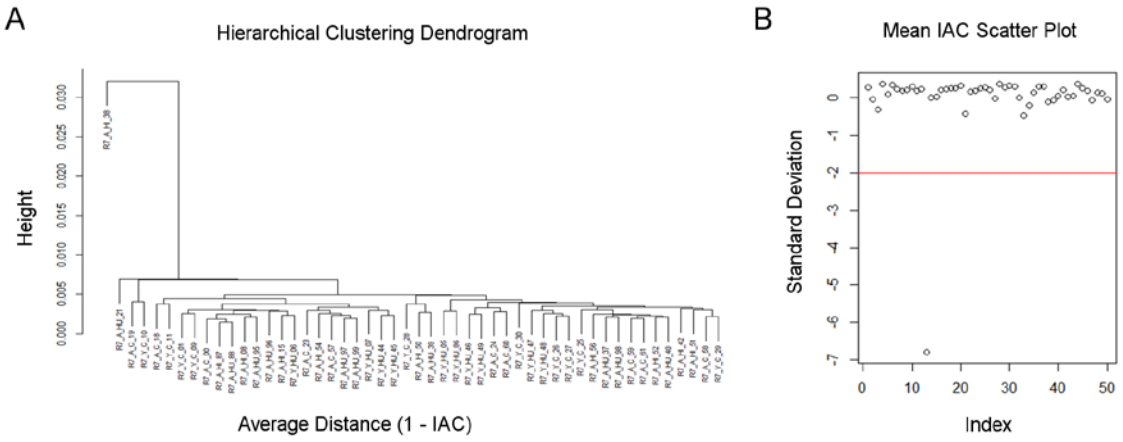


Figure 2.20 Removing outlier arrays in R7 data. A) hierarchical clustering of R7 dataset using 1-IAC as a distance metric. Sample R7_A_HI_38 indicated by the first branch is an obvious outlier. B) Scatter plot of the mean IAC of the same samples.

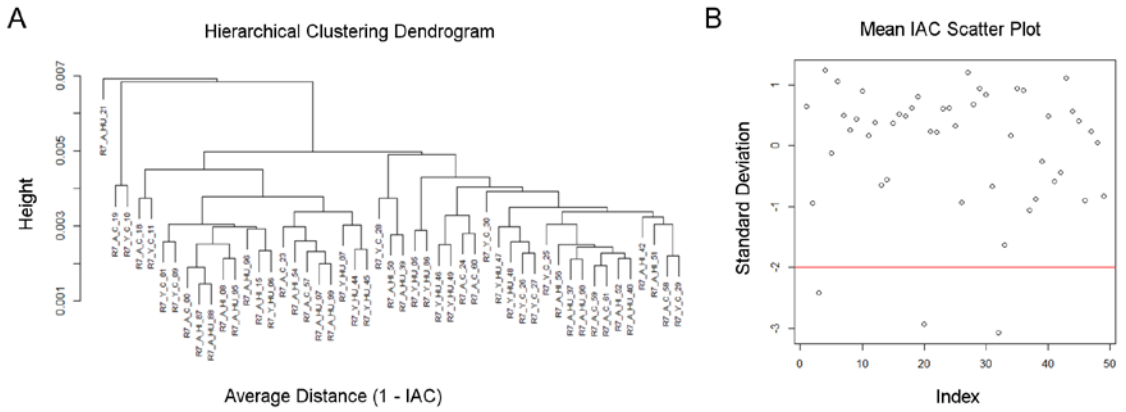


Figure 2.21 Removing outlier arrays in R7 data continued. IAC hierarchical clustering (A) and mean IAC scatter plot of the 49 samples of R7 dataset (B) showing the presence of three more outlier arrays.

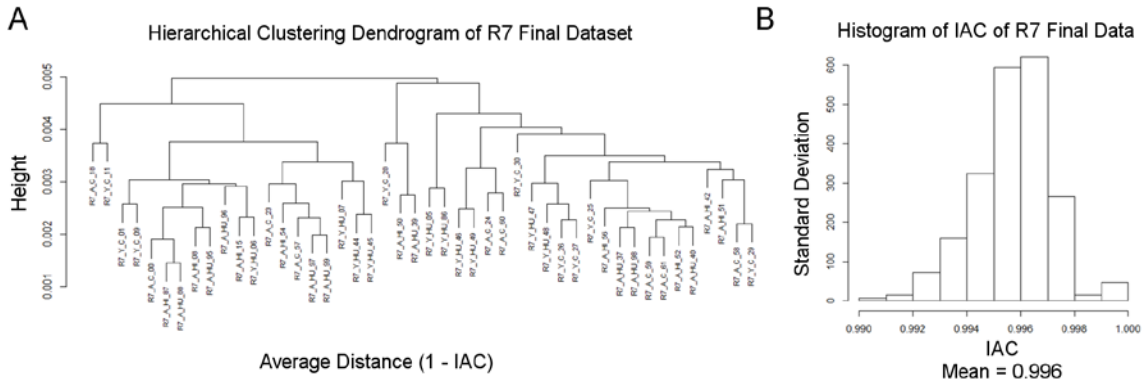


Figure 2.22 Final R7 data quality after removing outliers. IAC hierarchical clustering (A) and histogram of IAC of the final 46 samples of R7 dataset (B) showing no more obvious outliers. The mean IAC has improved to 0.996.

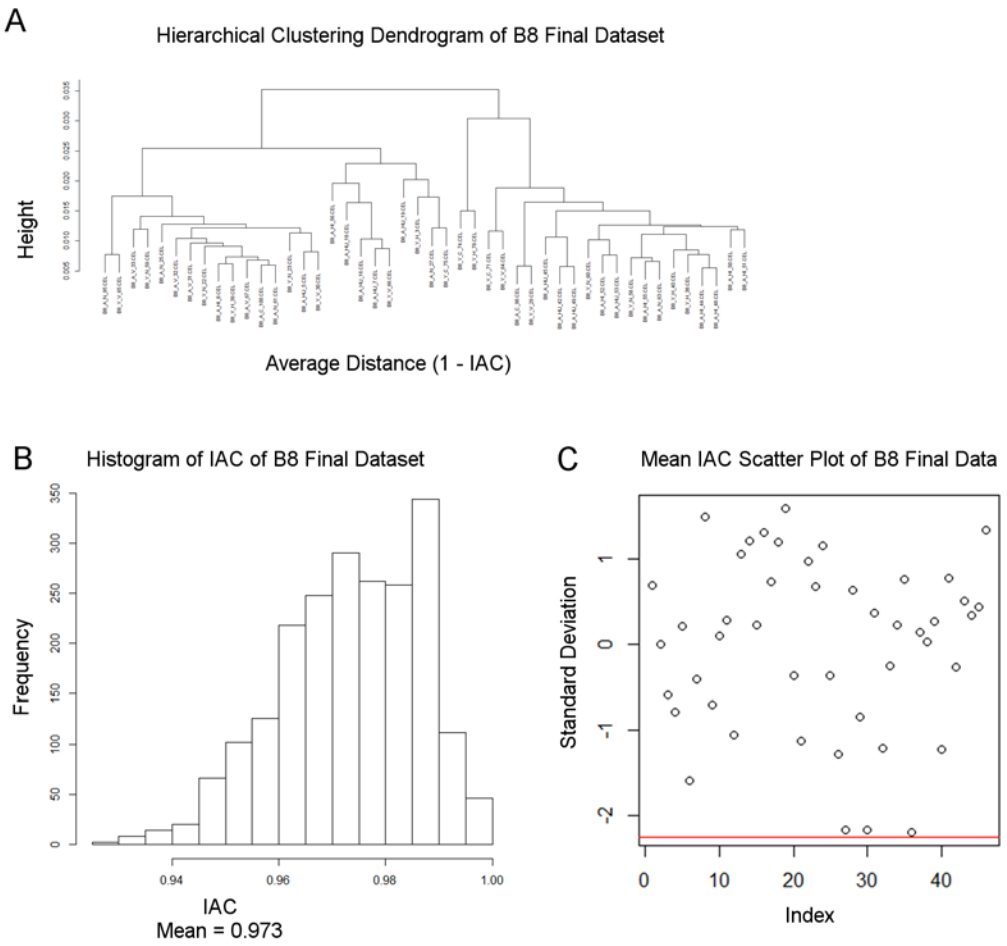


Figure 2.23 Final B8 data quality after removing outliers. IAC hierarchical clustering (A), histogram of IAC (B), and mean IAC scatter plot of the final 46 samples of B8 dataset (C) showing no more obvious outliers. The mean IAC is 0.973.

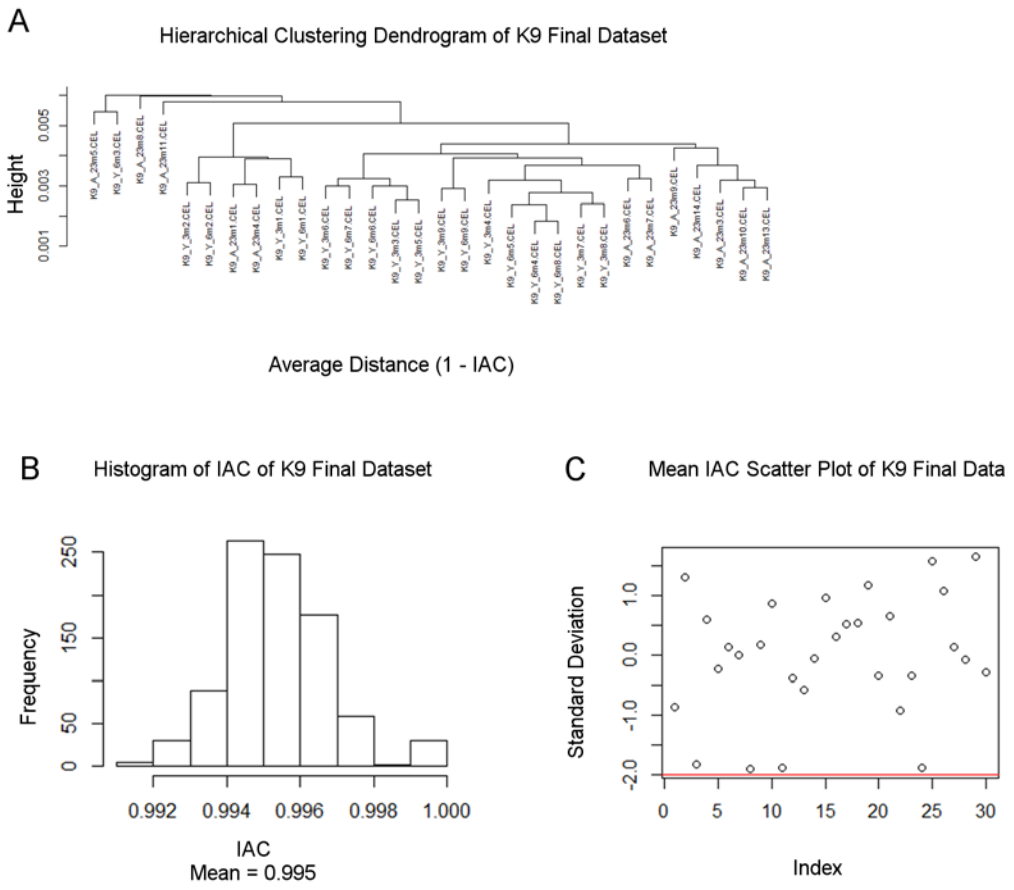


Figure 2.24 Final K9 data quality after removing outliers. IAC hierarchical clustering (A), histogram of IAC (B), and mean IAC scatter plot of the final 30 samples of K9 dataset (C) showing no more obvious outliers. The mean IAC is 0.995.

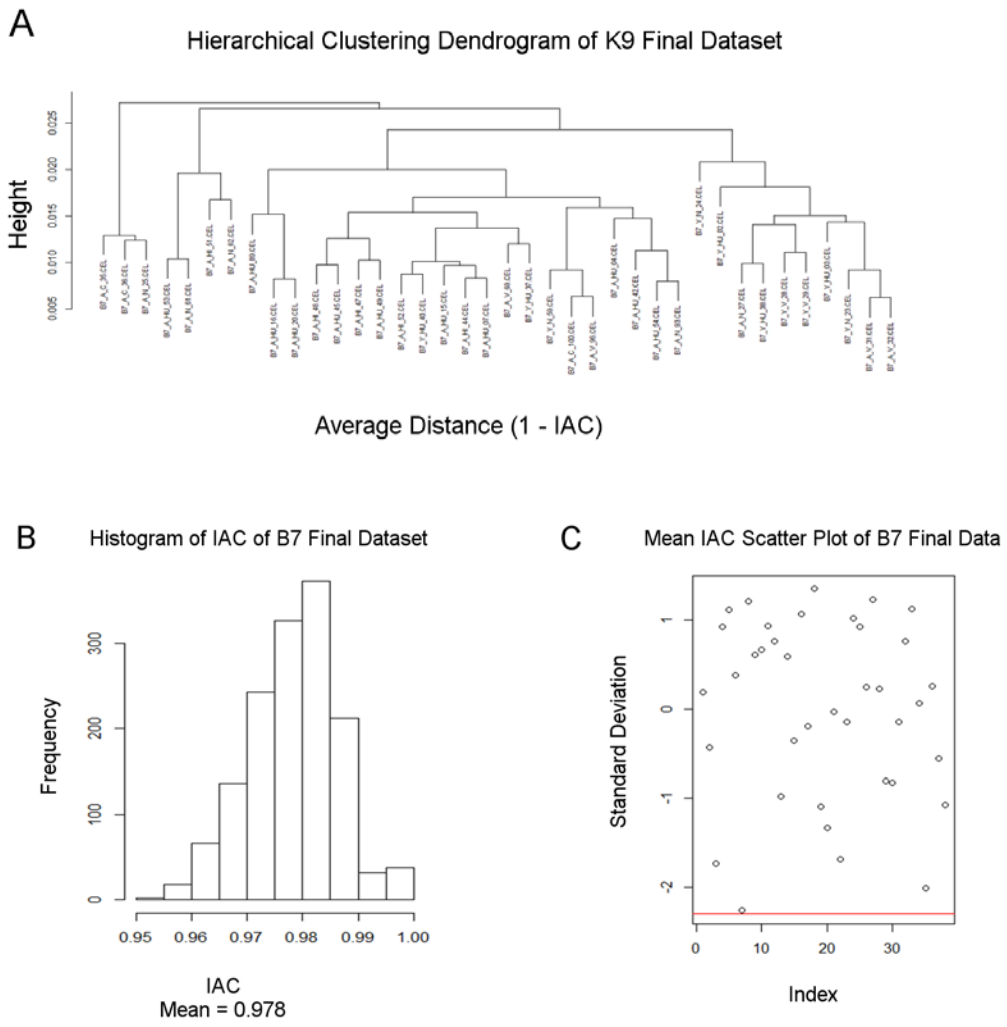


Figure 2.25 Final B7 data quality after removing outliers. IAC hierarchical clustering (A), histogram of IAC (B), and mean IAC scatter plot of the final 38 samples of B7 dataset (C) showing no more obvious outliers. The mean IAC is 0.978.

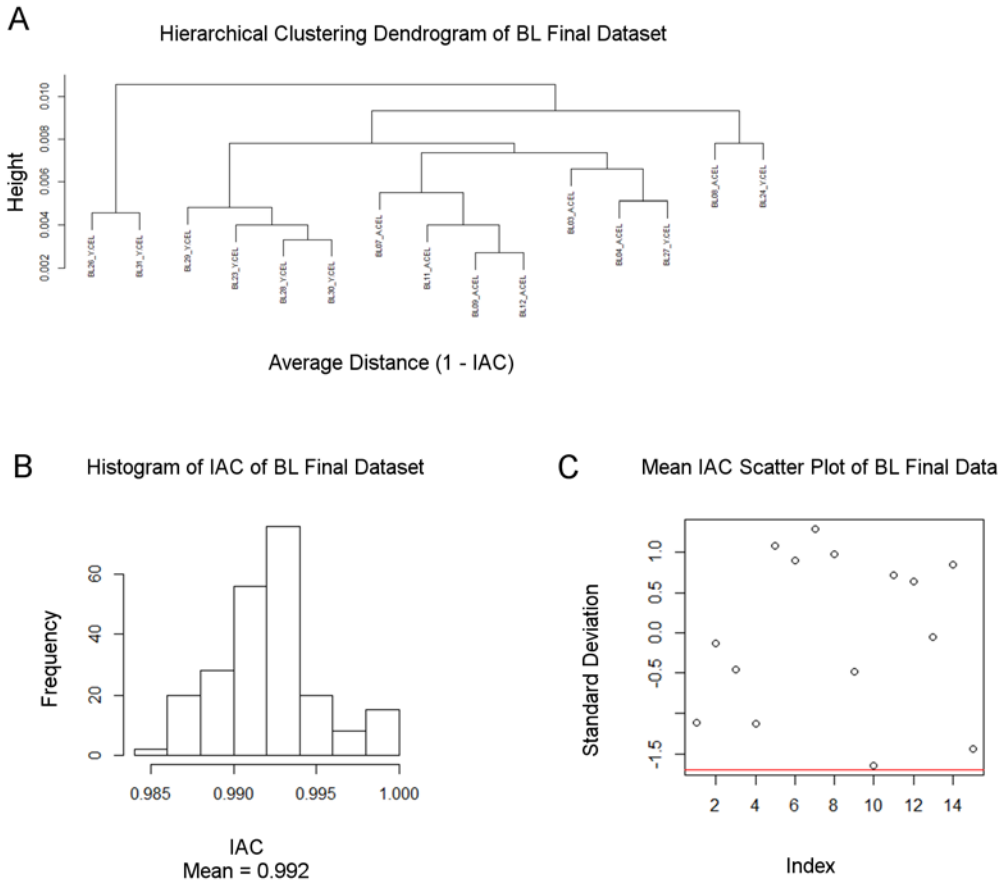


Figure 2.26 Final BL data quality after removing outliers. IAC hierarchical clustering (A), histogram of IAC (B), and mean IAC scatter plot of the final 16 samples of BL dataset (C) showing no more obvious outliers. The mean IAC is 0.992.

Table 2.4 Number of arrays selected from each datasets after preprocessing.

Study name	Original number of arrays	After quality control but before removing outliers			Final number of arrays after removing outlier		
		Total	Young	Aged	Total	Young (Y)	Aged (A)
B7	79	74	18	56	38	10	28
R7	50	50	21	29	46	19	27
B8	80	75	23	52	46	18	28
K9	49	49	18	13	30	18	12
BL	29	29	9	10	15	8	7

2.3.4.2 Data normalization and batch correction

The R7 dataset used the RAE230A chip type which contained a total of 15923 probe sets. After AP call filtering (i.e. excluding probe sets that were called “absent” in more than 90% of the samples) there were 11591 probe sets left. The quantile normalization and batch correction were performed on the AP filtered dataset using the *affy()* and ComBat packages, respectively, in R as described in the meta-analysis section. The box plot results before and after normalization and the hierarchical clustering results before and after batch correction were similar or slightly better to those described for meta-analysis (results not shown). The quantile normalization and ComBat batch correction was performed on other datasets in a similar manner, however, no AP call filtering was performed on B7 and BL because doing so led to a missing gene problem later in the meta- or network analysis.

2.3.4.3 Filtering of unwanted probe sets

Annotation filtering using the *genefilter* package in R in the R7 dataset showed that out of 11591 probe sets selected after AP call filtering, only 9435 were associated with some genes. Many genes contained duplicate or multiple probe sets, in which case probe sets with the highest connectivity were kept. After removing duplicate or multiple probe sets for a gene, a total of 8053 probe sets/genes in 46 arrays were finally selected for network analysis. Repeating this filtering process resulted in 4829 probe sets for both B7 and BL, and 7157 and 8250 probe sets for B8 and K9 datasets, respectively (Table 2.5).

2.3.5 Separate Aged and Young

At this point the aged and young samples were separated and checked for group-wise IAC based quality to make sure the distribution of all samples fell within 2 to 3 standard deviations below the mean IAC (Appendix 6.1.1 to Appendix 6.1.6). For R7 aged and young, the data quality was improved to a mean IAC of 0.998. For B8 datasets the mean IAC was 0.952 for young and 0.957 for aged, and all samples were distributed within 2 to 3 standard deviations below the mean IAC. For K9 young the mean IAC was 0.997 and all

samples were distributed within 2 standard deviations below the mean IAC. For B7 aged the mean IAC was 0.983 with all samples distributed within 3 standard deviations below the mean IAC.

Table 2.5 Number of probe sets selected from each datasets after preprocessing.

Dataset	Total genes in the array	After AP call filtering	After annotation filtering	After multiple and duplicate probe set filtering
B7	8799	Not done**	7246	4829
R7	15923	11591	9435	8053
B8	15923	10293	8075	7157
K9	15923	12279	9698	8250
BL	8799	Not done**	7246	4829

Note. **Not done because doing so led to a missing gene problem later in the meta- or network analysis.

2.4 Discussion

The goals in this chapter were to perform selection, collection, quality control, and preprocessing of ASLI gene expression data and examine their importance for effective downstream meta- and network analysis. These goals were accomplished by defining a data selection process to select studies that are homogeneous, collecting suitable gene expression datasets in ASLI, and finally, by assessing and applying suitable quality control and preprocessing measures on the selected datasets.

Although meta-analysis often includes a large number of unrelated studies, I followed a more conservative data selection approach in this study in order to concentrate on microarray gene expression datasets that focused on the hippocampus dependent ASLI as assessed by the Morris water maze test. The goal was to reduce sources of heterogeneity as much as possible. A major difficulty in combining results from independent studies is the occurrence of study heterogeneity. Studies that are superficially similar may in fact differ in many ways, some of which can be quite subtle

(Goldstein and Guerra, 2010). In general, studies carried out by different investigators may vary in scientific research goals, population of interest, handling of subjects, study design, quality of implementation, treatment dosage and timing, outcome definition or measures, and statistical methods of analysis (Goldstein and Guerra, 2010). Indeed, the data collection result showed that among the 38 studies only five matched my selection criteria in terms of major study goal, selection of animal model, and the assessment of learning impairment. Therefore, the choice of my data selection approach was appropriate. Moreover, the choice of starting the data preprocessing with original raw expression data (CEL files) was the right one, which gave me the opportunity to perform consistent quality assessment, preprocessing, and filtering of imperfect arrays and outlier values. The image and data quality assessment results (Figure 2.3 to Figure 2.8) show that even carefully performed experiments can have imperfect arrays and require close inspection. This observation allowed me to exclude these arrays from my analysis, which was not done in the original publications. For example, the image quality assessment allowed me to correct image contamination in a few of the arrays (Figure 2.3).

The signal intensity ratio of the 3' probe set over the 5' probe set of the housekeeping genes gives an indication of the integrity of starting RNA and efficiency of first strand cDNA synthesis. The signal of each probe set reflects the sequence of the probes and their hybridization properties. The 3'/5' ratios for β-actin and GAPDH in the BL, R7, and K9 arrays mostly fell within the generally recommended range of -3 to +3 (Figure 2.5), while the ratios of some of the arrays in B7 and B8 were outside of these ranges (Figure 2.6). However, the RNA quality results for all arrays were considered along with the results from their image quality and RLE-NUSE plots following recommendations in the literature (http://www.affymetrix.com/support/help/faqs/ge_assays/faq_17.jsp). For arrays to be of good quality their RLE-NUSE values should fall below the 95% cutoff line (dotted red line in Figure 2.7 and Figure 2.8). In addition, arrays that have interquartile range (IQR) such as RLE-IQR > 0.75 and NUSE-IQR > 0.075 should be removed (Aluru et al., 2013). In general, all arrays in BL, R7, and K9 were of very good quality compared to

B7 and B8. However, considering results from all quality assessment steps together, arrays that were outside of the good quality range and which also had greater than 15% array outlier values were excluded from further consideration.

The RMA method was chosen to perform the preprocessing of microarrays which included background correction, normalization, and summarization. RMA was chosen for two reasons: 1) it performed slightly better than MAS in removing batch effect, and 2) when MAS background corrected data were used to create networks, the recommended soft powers for R7, B8, and K9 were far off from each other (this was particularly true for B8 and K9), and a lower power close to that of R7 would not produce approximate scale free topology for B8 and K9.

Since batch effects generally lead to increased variability and decreased power to detect a real biological signal (Leek and Storey, 2007), batch effects were carefully removed from each dataset. During the removal process it was made sure that the samples were not confounded. If batch effects are confounded with an outcome of interest it can result in misleading biological or clinical conclusions (Leek et al., 2010). An example of confounding is when all of the cases are processed on one day and all of the controls are processed on another. The ComBat method allowed correction of batch effects and the removal of any unexplained technical variations from all datasets. The results (Figure 2.14 to Figure 2.18) confirmed the findings of recent studies (Johnson et al., 2007; Leek et al., 2010) and demonstrated the necessity of removing batch effects from microarray data before integrating them in any analysis.

Data selected for network analysis went through a more rigorous preprocessing with the addition of an IAC based outlier removal process. In a network analysis, the expression pattern of a gene is compared to that of other genes in the dataset, often using probabilistic, mutual information, or correlation based network inference methods. Interconnectedness of genes is assessed and evaluated to understand their role in a network. Therefore, the presence of outlier microarray samples can severely bias

network analysis. The IAC based outlier identification and removal process undertaken in this research ensured that the outliers were removed from all datasets in a consistent and unbiased manner (following recommended protocols) (Miller et al., 2010; Oldham et al., 2006; Oldham et al., 2008). Questionable samples were removed from each dataset while maintaining a fine balance between quality and number of samples required for network analysis. For example, the mean IACs for all datasets were 0.996 (R7, Figure 2.22), 0.973 (B8, Figure 2.23), 0.995 (K9, Figure 2.24), 0.978 (B7, Figure 2.25), and 0.992 (BL, Figure 2.26). The values for B8 and B7 were slightly lower compared to others. The histogram results show that a few more arrays could have been removed, however, that would have resulted in a loss of one or more of the sample types from a dataset, and made the data unfit for batch normalization. However, the IAC values indicate that the overall consistency of gene expression among samples in each dataset used for network construction was very comparable.

In summary, this chapter dealt with collection, selection, and preparation of ASLI microarray gene expression datasets for this study. Even though the initially selected 38 microarray studies in cognitive impairment apparently looked similar, they actually varied in terms of major study goal, selection of animal model, and the assessment of learning impairment. This made my choice of a more conservative data selection approach logical, which resulted into a selection of five ASLI datasets. A detailed inspection of data quality revealed the presence of imperfections in some arrays as well as the presence of outlier arrays and batch effects. Working directly from raw expression CEL data files and applying proper quality control and preprocessing on the data resulted in improved data quality. The ComBat method enabled the correction of batch effects and removal of unexplained technical variations from all datasets. Further, the IAC based outlier identification and removal process undertaken in this research ensured that the outliers were removed from all datasets in a consistent and unbiased manner. The data were prepared to combine across individual studies at the probe set level, which is expected to produce the best outcome. The results at each stage of

quality control, preprocessing, filtering, and data integration indicate satisfactory outcomes and make the data ready for downstream meta- and network analysis.

2.5 References

Affymetrix, 2001. Microarray suite user guide version 5.0. Vol., ed.^eds. Affymetrix Inc, Santa Clara CA.

Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A. 97**,** 10101-6.

Aluru, M., Zola, J., Nettleton, D., Aluru, S., 2013. Reverse engineering and analysis of large genome-scale gene networks. Nucleic Acids Res. 41**,** e24.

Benito, M., Parker, J., Du, Q., Wu, J., et al., 2004. Adjustment of systematic microarray data biases. Bioinformatics. 20**,** 105-14.

Blalock, E.M., Chen, K.C., Sharrow, K., Herman, J.P., et al., 2003. Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment. J Neurosci. 23**,** 3807-19.

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 19**,** 185-93.

Bolstad, B.M., Collin, F., Brettschneider, J., Simpson, K., et al., 2005. Quality assessment of affymetrix GeneChip data. In: Bioinformatics and computational biology solutions using R and Bioconductor. Statistics for biology and health, Vol., R. Gentleman, R.A. Irizarry, V.J. Carey, S. Dudoit, et al., ed.^eds. Springer, New York, pp. 33 - 48.

Bolstad, B.M., 2006. affyPLM: Methods for fitting probe-level models. R package version 1.10.0. http://bmbolstad.com/. Vol., ed.^eds.

Burger, C., Lopez, M.C., Feller, J.A., Baker, H.V., et al., 2007. Changes in transcription within the CA1 field of the hippocampus are associated with age-related spatial learning impairments. Neurobiol Learn Mem. 87**,** 21-41.

Burger, C., Lopez, M.C., Baker, H.V., Mandel, R.J., et al., 2008. Genome-wide analysis of aging and learning-related genes in the hippocampal dentate gyrus. Neurobiol Learn Mem. 89**,** 379-96.

Chen, C., Grennan, K., Badner, J., Zhang, D., et al., 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One. 6**,** e17238.

Cope, L.M., Irizarry, R.A., Jaffee, H.A., Wu, Z., et al., 2004. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics. 20**,** 323-31.

Durinck, S., 2008. Pre-processing of microarray data and analysis of differential expression. Methods Mol Biol. 452**,** 89-110.

Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A., 2004. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 20**,** 307-15.

Goldstein, D.R., Delorenzi, M., Luthi-Carter, R., Sengstag, T., 2010. Comparison of meta-analysis to combined analysis of a replicated microarray study. In: Meta-analysis and combining information in genetics and genomics. Chapman & Hall/CRC Mathematical and Computational Biology Series, Vol., D.R. Goldstein, R. Guerra, ed.^eds. CRC Press, pp. 135-156.

Goldstein, D.R., Guerra, R., 2010. A brief introduction to meta-analysis, genetics and genomics. In: Meta-analysis and combining information in genetics and genomics. Chapman & Hall/CRC Mathematical and Computational Biology Series, Vol., D.R. Goldstein, R. Guerra, ed.^eds. CRC Press, pp. 3-20.

Harr, B., Schlotterer, C., 2006. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. Nucleic Acids Res. 34**,** e8.

Hubbell, E., Liu, W.M., Mei, R., 2002. Robust estimators for expression analysis. Bioinformatics. 18**,** 1585-92.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., et al., 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 18 Suppl 1**,** S96-104.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., et al., 2003a. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31**,** e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., et al., 2003b. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 4**,** 249-64.

Irizarry, R.A., Wu, Z., Jaffee, H.A., 2006. Comparison of Affymetrix GeneChip expression measures. Bioinformatics. 22**,** 789-94.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 8**,** 118-27.

Kadish, I., Thibault, O., Blalock, E.M., Chen, K.C., et al., 2009. Hippocampal and cognitive aging across the lifespan: a bioenergetic shift precedes and increased cholesterol trafficking parallels memory impairment. J Neurosci. 29**,** 1805-16.

Konstantinopoulos, P.A., Cannistra, S.A., Fountzilas, H., Culhane, A., et al., 2011. Integrated analysis of multiple microarray datasets identifies a reproducible survival predictor in ovarian cancer. PLoS One. 6**,** e18202.

Larsen, M.J., Thomassen, M., Tan, Q., Sorensen, K.P., et al., 2014. Microarray-based RNA profiling of breast cancer: batch effect removal improves cross-platform consistency. Biomed Res Int. 2014**,** 651751.

Leek, J.T., Storey, J.D., 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3**,** 1724-35.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., et al., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 11**,** 733-9.

Li, C., Wong, W., 2001a. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol. 2**,** RESEARCH0032.

Li, C., Wong, W.H., 2001b. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A. 98**,** 31-6.
Lim, W.K., Wang, K., Lefebvre, C., Califano, A., 2007. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics. 23**,** i282-8.
Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., et al., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol. 14**,** 1675-80.
Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., et al., 2010. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 10**,** 278-91.
Miller, J.A., Horvath, S., Geschwind, D.H., 2010. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc Natl Acad Sci U S A. 107**,** 12698-703.
Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., et al., 2003. Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet. 19**,** 570-7.
Oldham, M.C., Horvath, S., Geschwind, D.H., 2006. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci U S A. 103**,** 17973-8.
Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., et al., 2008. Functional organization of the transcriptome in human brain. Nat Neurosci. 11**,** 1271-82.
Pevsner, J., 2009. Gene Expression: Microarray Data Analysis. In: Bioinformatics and Functional Genomics. Vol., ed.^eds. Wiley.
Piccolo, S.R., Sun, Y., Campbell, J.D., Lenburg, M.E., et al., 2012. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics. 100**,** 337-44.
Ramasamy, A., Mondry, A., Holmes, C.C., Altman, D.G., 2008. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med. 5**,** e184.
Rodriguez-Zas, S.L., Ko, Y., Adams, H.A., Southey, B.R., 2008. Advancing the understanding of the embryo transcriptome co-regulation using meta-, functional, and gene network analysis tools. Reproduction. 135**,** 213-24.
Rowe, W.B., Blalock, E.M., Chen, K.C., Kadish, I., et al., 2007. Hippocampal expression analyses reveal selective association of immediate-early, neuroenergetic, and myelinogenic pathways with cognitive impairment in aged rats. J Neurosci. 27**,** 3098-110.
Rung, J., Brazma, A., 2013. Reuse of public genome-wide gene expression data. Nat Rev Genet. 14**,** 89-99.
Schadt, E.E., Li, C., Ellis, B., Wong, W.H., 2001. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem Suppl. Suppl 37**,** 120-5.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 270**,** 467-70.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., et al., 2000. Normalization strategies for cDNA microarrays. Nucleic Acids Res. 28**,** E47.

Sims, A.H., Smethurst, G.J., Hey, Y., Okoniewski, M.J., et al., 2008. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. BMC Med Genomics. 1**,** 42.

Sirbu, A., Ruskin, H.J., Crane, M., 2010. Cross-platform microarray data normalisation for regulatory network inference. PLoS One. 5**,** e13822.

Stein, C.K., Qu, P., Epstein, J., Buros, A., et al., 2015. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. BMC Bioinformatics. 16**,** 63.

Stuart, R.O., Bush, K.T., Nigam, S.K., 2001. Changes in global gene expression patterns during development and maturation of the rat kidney. Proc Natl Acad Sci U S A. 98**,** 5649-54.

Talloen, W., Gohlmann, H., 2009. Data analysis preparation. In: Gene Expression Studies Using Affymetrix Microarrays. Vol., ed.^eds. Chapman and Hall/CRC.

Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C., et al., 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. 29**,** 2549-57.

Tseng, G.C., Ghosh, D., Feingold, E., 2012. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 40**,** 3785-99.

Wilson, C.L., Miller, C.J., 2005. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics. 21**,** 3683-5.

Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., et al., 2004. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association. 99**,** 909-917.

Chapter 3 Meta-Analysis

3 Hippocampal gene expression meta-analysis identifies aging and age-associated spatial learning impairment (ASLI) genes and pathways

3.1 Introduction

In addition to the more commonly used differential expression analysis of microarray data, which identifies a list of genes that are differentially expressed within the dataset of an individual experiment, various meta-analysis approaches have also been described in the past. Meta-analysis, which combines the results of independent but related experiments in a relatively inexpensive way, has the ability to increase the statistical power to obtain a more precise estimate of gene expression differences. This approach for uncovering of a significant effect from a combined analysis, where individual studies have not yielded any positive or reliable findings, has emerged as an essential tool for modern genetics and genomic analysis (Goldstein and Guerra, 2010). Meta-analysis, where each study dataset is analyzed independently and then the results from all studies are combined, is more advantageous than mega-analysis. In mega-analysis information across studies is pooled into a single dataset for analysis, often after minor correction. As a result, mega-analysis suffers from many drawbacks. Because even after correction and adjustment, study observations may well remain too heterogeneous for pooling (Goldstein and Guerra, 2010).

There are many ways to combine the results across microarray studies and platforms (Goldstein et al., 2010; Moreau et al., 2003; Ramasamy et al., 2008; Sirbu et al., 2010). These generally fall into four generic approaches such as vote counting, combining ranks, combining p-values, and combining effect sizes. As discussed in chapter two, in order to eliminate bias due to specific algorithms that were used in the original studies, and to allow consistent handling of all datasets, one should use the feature-level extraction output or original raw data such as CEL files, and convert those to gene expression data matrix in a consistent manner. However, in practice, vote counting, combining p-values, and many combining ranks methods are not designed to work with CEL files. Moreover, techniques

under these three categories suffer from many limitations. For example, vote counting and some combining ranks techniques only consider the genes declared significant based on some arbitrary threshold in the original studies, and do not consider information from all available genes. Many of these techniques do not treat frequently studied and rarely studied genes present in newer microarrays equally, and do not produce highly accurate results when the number of studies is small.

Combining effect sizes using an inverse-variance method (Cochran, 1937; Fleiss, 1993) can overcome these limitations. It is considered to be the most comprehensive approach for meta-analysis of gene expression microarrays (Ramasamy et al., 2008). In addition, this method offers several other decisive advantages. For example, it yields a biologically interpretable discrimination measure, which is the pooled effect size of differential expression and its standard error. Combining effect sizes is the only technique that weights the contribution of each study by its precision, which is related to the study sample size. Further, the use of effect size, a unit-less measure not dependent on sample size, facilitates the combining of signals from different technology platforms. Thus, combining effect sizes method presents itself as a promising technique to follow in this research.

Combining effect sizes using the inverse-variance techniques has been used frequently by many researchers in meta-analysis of microarrays (Goldstein et al., 2010; Stevens and Doerge, 2005). In this method (Borenstein et al., 2009), standardized mean difference (SMD) can be used as a study-specific effect size when dealing with typical microarray studies involving two treatment groups. SMD can be calculated as the Cohen's d, which is the difference in two group means standardized by its pooled standard deviation. By pooling the two estimates of the standard deviation, a more accurate estimate of their common value is obtained. The SMD thus serves as an index that would be comparable across studies. So, the first step is to calculate the SMD effect size and the variance associated with the effect size for every gene in every study. SMDs from every study are then combined using either the fixed or the random effect model.

Under the fixed effect model it is assumed that there is one true effect size μ, which is shared by all the studies included. It follows that the combined effect is an estimate of this common effect size. Thus, the observed effects will be distributed about μ, with a variance σ^2 that depends primarily on the sample size for each study. So, the observed effect T_i is determined by the common effect μ plus the within-study error ε_i as $\mathrm{T}_i = \mu + \varepsilon_i$. By contrast, under the random effects model the true effect could vary from study to study. Rather than assuming that there is one true effect this allows that there is a distribution of true effect sizes. The combined effect therefore cannot represent the one common effect, but instead represents the mean of the population of true effects. The observed effect T_i is sampled from a distribution with true effect θ_i, and (within-study) variance σ^2. The true effect θ_i, in turn, is sampled from a distribution with mean μ and (between-study) variance τ^2. So, the observed effect T_i is determined by the true effect θ^i plus the within-study error ε_i. In turn, θ_i is determined by the mean of all true effects μ and the between-study error ζ^i as $\mathrm{T}_i = \theta_i + \varepsilon_i = \mu + \zeta^i + \varepsilon_i$ (Borenstein et al., 2009).

The study-specific effect sizes for every gene are then combined across studies into a weighted average, with more weight given to studies with larger sample sizes, which again is thought to be more precise compared to studies with smaller sample sizes. The study weights are inversely proportional to the variance of the study specific estimates. Thus, this meta-analysis method using SMD effect size seems to be a statistically sound method to combine microarray data across microarray studies and platforms. Therefore, the goal of this chapter is to accomplish objective two, which is to integrate the ASLI microarray gene expression datasets selected in chapter two (Table 2.3 and Table 2.4) using meta-analysis methods, and thereby identify and characterize genes that may be involved in ASLI, as well as to identify and characterize gene networks based on existing biological knowledge.

3.2 Methods

All statistical analyses were performed in R using appropriate software packages.

3.2.1 Data integration

Microarray data across two different Affymetrix platforms in all five studies were integrated at the common probe set level. Towards that goal, probe sets common to both the RGU34A and RAE230A chip types or only either of the chip types were identified and grouped into three categories: rgu_exclusive, all5_common, and rae_exclusive (see Section 2.2.4 for details). Each probe set specific data and their analysis outcome from all studies were combined in two ways (Figure 3.1): 1) effect size integration (which combined the estimated effect size results), and 2) direct data integration (which combined the preprocessed data first before any analysis).

3.2.1.1 Effect size integration

I estimated effect sizes on the within-study batch-corrected data using the random effect size model as follows. First, for each probe set, study-specific sample sizes, mean expression measures, and standard deviations were computed for each comparison. In order to understand the effect of age and spatial learning impairment, data were analyzed in two ways, e.g. by comparing samples across age (aged vs. young, AY) and across learning impairment (aged-impaired vs. aged-unimpaired, IU), respectively. Next, the *meta* package in R (http://cran.r-project.org/ web/packages/meta/meta. pdf) was used to calculate each study-specific SMD (Cohen's d) for each probe set, and later, probe set SMDs for all studies in each category (e.g. rgu_exclusive, all5_common, and rae_exclusive) were pooled utilizing Hedges' adjusted g (Borenstein et al., 2009) to obtain the final random effect size for each probe set. Effect size values for all probe sets from all three categories were then combined together, annotated, and summarized. Duplicate probe sets and multiple probe sets annotated to the same gene were summarized by keeping the probe set with the lowest p-value (of the z-value) for the gene (Rhodes et al., 2002). Uninformative probe sets were filtered out by removing probe sets whose expression values had a coefficient of variation of zero across all arrays and probe sets with a p-value (of the effect size z-value) greater than 0.1. The p-values of the treatment effect for all probe sets were adjusted with the

Benjamini and Hochberg (BH) multiple testing correction (Benjamini and Hochberg, 1995) in R.

3.2.1.2 Direct data integration

This was done by a cross-study and cross-platform normalization process by first combining data separately for each category (e.g. rgu_exclusive, all5_common, and rae_exclusive) and then adjusting data across all studies. For each category, data were adjusted similarly as within-study batch correction, however, considering individual studies as separate batches. Next, differential expression analysis was performed by comparing the data in two ways as above e.g. AY and IU using the *limma* R software package (Smyth, 2004). Significant differentially expressed genes from all three categories were combined together, annotated, and summarized as described above. Duplicate and multiple probe sets issues and multiple testing corrections were also handled similar to the effect size analysis.

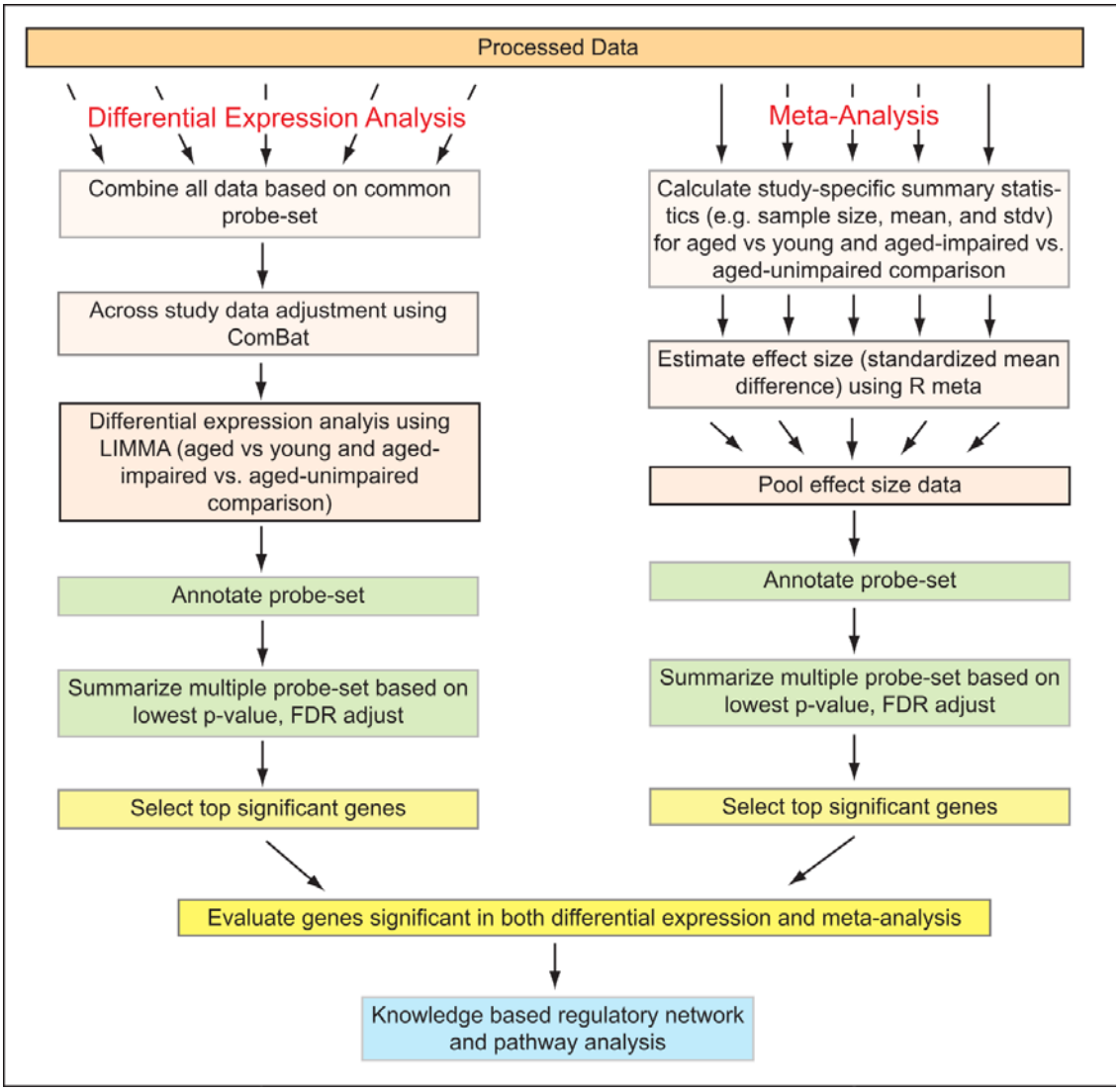


Figure 3.1 A summary of the meta-analysis workflow. Five individual studies (BL, B7, R7, B8, and K9) were selected for this meta-analysis. The studies involved two different array platforms, Affymetrix RG-U34a and RAE-230a. Following preprocessing, data were integrated across studies and across array platforms and analyzed in two ways: meta-analysis using random effect size model and differential expression analysis using the *limma* software. Top significant genes were used to identify enriched functions and pathways and to construct knowledge based gene regulatory networks using the Ingenuity Pathway Analysis (IPA) software (http://www.ingenuity.com).

3.2.2 Functional and Pathway Analysis

Functional and pathway analysis was performed mainly using the IPA software. Datasets containing identifiers of significant (p-value ≤ 0.05) differentially expressed genes from the AY or IU comparisons along with their corresponding effect size estimates (as fold-change values) and p-values were used as input. Identifiers that were successfully mapped to their corresponding objects in the IPA knowledge base were considered for functional, network, and canonical pathway analysis.

For functional analysis the mapped identifiers that were associated with biological functions and/or diseases in the IPA knowledge base were considered. Right-tailed Fisher's exact test was used to calculate a p-value determining the probability that each biological function and/or disease assigned to the dataset is due to chance alone. The expression levels (up- or down-regulation) for all of the input genes in each function annotation category were compared with the information stored for those genes in the IPA knowledge base to predict whether the expression patterns correspond to the activation state (decreased or increased) for that category.

For network analysis the mapped identifiers were overlaid onto a global molecular network developed from information contained in the IPA knowledge base. Networks of network eligible molecules were then algorithmically generated based on their connectivity. Next, the functional analysis of a network identified the biological functions and/or diseases that were most significant to the molecules in the network based on the association of the network molecules with the biological functions and/or diseases in the IPA knowledge base. Right-tailed Fisher's exact test was used to calculate a p-value determining the probability that each biological function and/or disease assigned to that network is due to chance alone.

Canonical pathway analysis identified the pathways from the IPA library of canonical pathways that were most significant to the gene lists. All the mapped identifiers from the dataset that were associated with a canonical pathway in the IPA knowledge base were

considered for the analysis. The significance of the association between the dataset and the canonical pathway was measured in two ways: a) using a ratio of the number of molecules from the dataset that map to the pathway divided by the total number of molecules that map to the canonical pathway, and b) using the Fisher's exact test by calculating a p-value determining the probability that the association between the genes in the dataset and the canonical pathway is explained by chance alone.

3.3 Results

3.3.1 Data Integration

Data were integrated between the RGU34A chip which had a total of 8799 probe sets and the RAE230A chip that had a total of 15923 probe sets. After data integration, the rgu_exclusive category contained 2356 probe sets exclusive to the RGU34A array only. The all5_common category included 6384 RGU34A unique probe sets mapping to 5435 RAE230A unique probe sets that are common among all five studies. Finally, the rae_exclusive category contained 10,431 probe sets exclusive to the RAE230A array type.

3.3.2 Gene identification and functional analysis

3.3.2.1 Aged vs. young (AY)

In order to assess the effect of aging, a comparison was made between aged vs. young animals. After combining probe sets from all three categories and after summarization I had effect size estimates for 10,619 unique annotated genes. After filtering, there were 3235 genes left, of which 2245 genes were found significant with a p-value ≤ 0.05 (Table S1 in (Uddin and Singh, 2013)) and 1753 genes were found significant after BH multiple testing correction (p-value ≤ 0.05). Among the 1753 genes, 874 genes had an I^2 (ratio of true heterogeneity to total variation) value of 0% while 1347 genes had an I^2 value under 40%. Differential expression analysis was also performed on the datasets in parallel to the effect size analysis. Using the 3235 genes from the effect size analysis, their AY differential expression levels (log fold-changes and corresponding p-values) were calculated and BH adjusted similarly to that of the effect size data. This resulted in a total of 1946 genes (p-

value < 0.05) and 1569 genes after BH adjustment (p-value < 0.05). Table 3.1 shows some of the effect size and differential expression analyses results for the top 10 most up- and down-regulated genes in the aged animals (compared to the young animals). The forest plots of two representative genes *C3* (complement component) (up-regulated) and *Tubb2b* (tubulin, beta 2B class IIb) (down-regulated) are presented in Figure 3.2.

Functional and Pathway analysis were performed using the IPA software. For this analysis, I considered the significant genes based on unadjusted p-value (p-value ≤ 0.05) of the random effect size, which resulted in a total of 2245 genes. These genes were used as input in the IPA of which 2240 were mapped to their corresponding objects in the IPA knowledge base. The functional analysis identified the biological functions and/or diseases that were most significant to the mapped gene list (activation z-score value-cutoff of 1.980). The IPA functional analysis predicts that comparatively more functions are decreased than increased in the aged animals. Table 3.2 shows a summary of the most significant functions, increased or decreased, as predicted by the IPA algorithm based on the expression levels of the genes in the dataset. The results show that the functions that are specifically decreased include cell viability of central nervous system cells, formation of cells, quantity and synthesis of inositol phosphate, and axonogenesis. Thus they affect the cell death and survival, cellular growth and proliferation, carbohydrate metabolism, molecular transport, small molecule biochemistry, cell morphology, and nervous system development and function in the aged animals. Major functions categories that see an increase are cellular movement, cellular development, and connective tissue development and function. The specific functions of the genes in this category include the migration of cells and differentiation of chondrocytes. I generated biological knowledge based gene interaction networks for the AY significant genes. A representative network graph is presented in Figure 3.4, which shows the network interactions of some of the aging and learning genes. Additional networks are presented in Appendix 6.3.1 to Appendix 6.3.5. A summary of the functions for the top five most significant networks is given in Table 3.3. The most critical canonical pathways that are affected in the aged animals include Eif2 (eukaryotic translation initiation factor 2) signaling, antigen presentation, and Ox40 (tumor necrosis factor) signaling pathways (Table 3.4).

Table 3.1 Top ten most up- and down-regulated genes (based on effect size) in the AY comparison.

Up-regulated genes		Effect size results			Differential expression results		
Probe ID	Symbol	Effect size (ES)	ES z-value	p-value of z-value	pBH of z-value	LogFC of DE	pBH of DE
1398892_at	Npc2	3.988	3.16	0.002	0.009	0.474	0
X52477_at	C3*	3.812	3.716	0	0.002	0.730	0
X13044_g_at	Cd74*	3.389	3.148	0.002	0.009	0.916	0
M15562_g_at	HLA-DRA*	3.236	3.284	0.001	0.007	1.011	0
1368187_at	Gpnmb*	3.189	2.827	0.005	0.017	0.610	0
L03201_at	Ctss*	3.110	3.362	0.001	0.006	0.368	0
1373575_at	Fcer1g*	2.846	2.821	0.005	0.018	0.473	0
1370885_at	Ctsz	2.606	3.201	0.001	0.008	0.432	0
J03752_at	Mgst1*	2.544	3.229	0.001	0.024	0.362	0
1376652_at	C1qa*	2.519	3.709	0	0.007	0.488	0
Down-regulated genes		Effect size results			Differential expression results		
Probe ID	Symbol	Effect size (ES)	ES z-value	p-value of z-value	pBH of z-value	LogFC of DE	pBH of DE
1376319_at	Sema3c*	-3.674	-3.867	0	0.001	-0.588	0
X57281_at	Glra2	-2.589	-4.599	0	0	-0.528	0
1388821_at	Trib2	-2.029	-2.578	0.01	0.028	-0.253	0
1388750_at	Tfrc*	-1.853	-2.576	0.01	0.028	-0.203	0
L03294_at	Lpl*	-1.803	-5.42	0	0	-0.395	0
1374966_at	Dcx*	-1.783	-3.42	0	0.005	-0.262	0
1389533_at	Fbln2	-1.756	-2.332	0.02	0.04	-0.192	0
D45412_s_at	Ptpro*	-1.721	-2.499	0.013	0.032	-0.319	0
M58369_at	Pnlip*	-1.618	-3.26	0.001	0.007	-0.200	0
X03369_s_at	Tubb2b*	-1.607	-2.907	0.004	0.015	-0.174	0

Top genes identified by IPA are indicated by an asterisk (*). Legends: ES, effect size; pBH, p-value with Benjamini and Hochberg correction; FC, fold change; DE, differentially expressed.

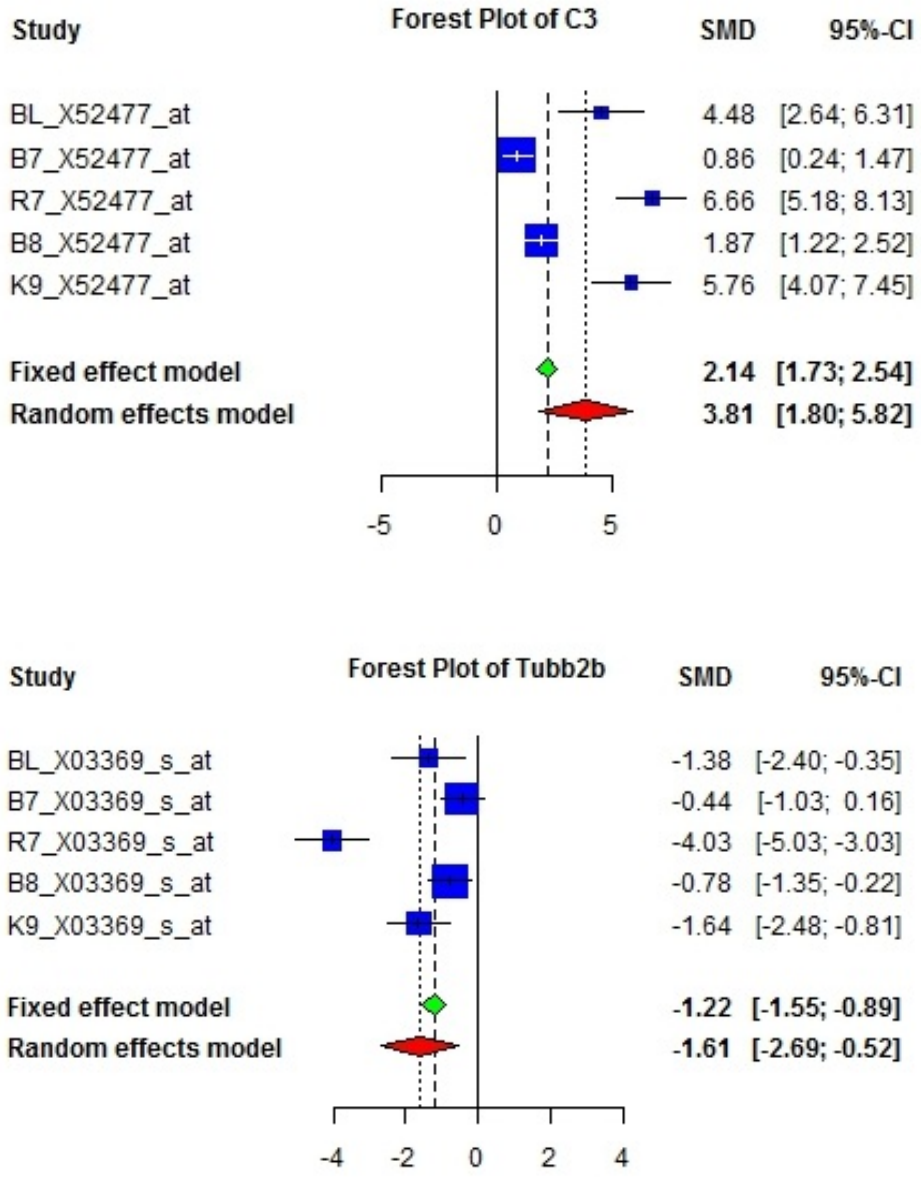


Figure 3.2 Forest plots of two representative significant genes in the aged rats. *C3* is up-regulated (top) and *Tubb2b* is down-regulated (bottom) in the aged rats. For the selected probe set for each gene the individual study specific SMD and their 95% confidence intervals (CI) are plotted and shown on each row. The effect size results are shown at the bottom of each plot.

Table 3.2 Significantly increased or decreased functions and associated genes in the AY comparison.

Functions Annotation	p-value	Predicted activation state	Activation z-score	High-level functions category	Genes
Cell viability of central nervous system cells	0.00 to 0.02	Decreased	-2.757 to -2.000	Cell death and survival	*ApoE*[A,L,SL], *Atf3*, *Bdnf*[L,SL], *Cdk5r1*[L,SL], *Cycs*, *Hspb1*, *Ide*, *Igf2*, *Ntf3*[L], *Plagl1*, *Prkcg*[L,SL], *Rela*[L], *Serpini1*, *Sh3kbp1*, *Slc11a2*[L], *Vegfa*[L], *Vip*
Formation of cells	0.01	Decreased	-2.376	Cellular growth and proliferation	*Bdnf*[L,SL], *Egr1*[L], *Fgf18*, *Icam1*, *Igf2*, *Nppa*, *Pf4*, *S100b*[L], *Sdc2*, *Wt1*
Quantity and synthesis of inositol phosphate	0.02	Decreased	-2.186	Carbohydrate metabolism, molecular transport, et.	*Agtr1*, *Avp*[L], *Cckbr*, *Gal*[L], *Gnaq*, *Grp*[L], *Icam1*, *Mas1*, *Pthlh*, *Rgs2*, *Rgs3*, *S1pr1*, *Trhr*
Axonogenesis	0.01	Decreased	-1.980	Cell morphology, assembly and organization, nervous system development and function	*Actb*, *Actr3*, *Agrn*, *Bdnf*[L,SL], *Cck*, *Cntn2*[L], *Igf1r*, *L1cam*[L,SL], *Mbp*, *Picalm*, *Ppp2ca*, *Snap91*, *Stk11*
Migration of cells	0.00 to 0.01	Increased	2.158	Cellular movement	*Abcc1*, *Actr3*, *Agt*, *Aif1*, *Anxa2*, *Bcar1*, *Bdnf*[L,SL], *C3*, *Cck*,, *Ccl3l1/Ccl3l3*[L,SL], *Ccl5*, *Cd44*, *Cd82*, *Cdc42*, *Dnm2*, *Drd5*[L,SL], *Gucy1a3*, *Gucy1b3*, *Icam1*, *Nfkbia*[L], *Ntf3*[L], *Pten*[L], *Reln*, *Stat3*, *Scpep1*, *Tac1*[L], *Tgfa*, *Tgfa*, *Tgfb1*, *Tgfb2*, *Tpm1*, *Tubb2b*, *Vegfa*[L], and etc.
Differentiation of chondrocytes	0.01	Increased	2.183	Cellular development, connective tissue development and function	*Grn*, *Por*, *Rela*[L], *Tgfb1*, *Thrb*[L]

Note. Genes in red were up-regulated and in green were down-regulated in the aged rats. Genes annotated as aging, learning, and spatial learning in the IPA knowledge base are indicated by "[A]", "[L]", and "[SL]", respectively.

Table 3.3 Major functions associated with the top five networks in the AY comparison.

Network ID	Top functions associated with the networks	IPA score	Total focus genes
1	Molecular transport, cell-to-cell signaling and interaction, nervous system development and function	25	35
2	Endocrine system disorders, gastrointestinal disease, metabolic disease	21	33
3	Cellular assembly and organization, tissue development, cell morphology	17	30
4	Cell-to-cell signaling and interaction, cell signaling, molecular transport	14	28
5	Drug metabolism, protein synthesis, cancer	14	28

Table 3.4 Top canonical pathways in the AY comparison.

Name	p-value	Ratio
EIF2 signaling pathway	2.36E-07	58/170 (0.341)
Antigen presentation pathway	6.01E-05	14/27 (0.519)
OX40 signaling pathway	1.91E-04	19/60 (0.317)
Chondroitin sulfate degradation pathway	4.96E-03	6/14 (0.429)
IL-17A signaling in gastric cells pathway	5.17E-03	10/24 (0.417)
Complement system pathway	1.69E-02	10/32 (0.312)

3.3.2.2 Aged-impaired vs. aged-unimpaired

In order to assess the effect of ASLI, a comparison was made between the aged-impaired vs. aged-unimpaired (IU) rats where I included three sets of controls (e.g. cage controls, visual controls, and stress controls (no platform during memory test in the water maze)) in the aged-impaired group as was done in the B7 and B8 studies (Burger et al., 2007; Burger et al., 2008). After combining probe sets from all three categories and after summarization there were 10,412 unique annotated genes with effect size estimates. After filtering out uninformative genes there were 1310 genes left, of which 787 were found significant with a p-value ≤ 0.05 (Table S2 in (Uddin and Singh, 2013)). Among the 787 genes, 59 were

significant with BH adjusted p-value ≤ 0.05 and 55 of these genes have an I^2 value of 0%. Differential expression analysis for the 1310 IU genes identified 460 significant genes (p-value ≤ 0.05), of which 92 were significant with p-value ≤ 0.05 after correction. However, among the 92 genes significant in the differential expression analysis, 44 were also present in the effect size meta-analysis (p ≤ 0.05) category. Table 3.5 shows some of these effect size and differential expression analyses results for the top 10 most up- and down-regulated genes in the aged-impaired (as compared to the aged-unimpaired) animals. Figure 3.3 shows the forest plots of two representative genes *Arc* (activity-regulated cytoskeleton-associated protein) (down-regulated) and *Marcks* (myristoylated alanine-rich protein kinase C substrate) (up-regulated).

A total of 738 IU genes with significant effect sizes (p-value ≤ 0.05) were used as input for the functional analysis in IPA. Though cell viability of hippocampal neurons and CNS (central nervous system) cells, cell-to-cell signaling, and molecular transport were the top functions in the results, none were statistically significant. However, when I reanalyzed with an IU effect size dataset that was generated without any controls, four functions (e.g. molecular transport, cellular development, cellular growth and proliferation, and connective tissue development and function) were significantly decreased (results not shown). The specific functions of these genes in these categories include transport of molecules and proliferation of fibroblast cell lines. In addition, growth of neuritis was also decreased among others. Similar to AY, I generated biological knowledge based gene interaction networks for the IU related genes (Appendix 6.4.1 to Appendix 6.4.4). A summary of the functions for the top five most significant networks is given in Table 3.6. The canonical pathways that are most affected in the aged-impaired compared to the aged-unimpaired animals include Nurr77 (nuclear receptor subfamily) signaling in lymphocytes, nNOS (nitric oxide) signaling in neurons, and glutamate receptor signaling (Table 3.7).

Table 3.5 Top ten most up- and down-regulated genes (based on effect size) in the IU comparison.

Up-regulated genes		Effect size results			Differential expression results			
Probe ID	Symbol	Effect size (ES)	ES z-value	p-value of z	pBH of z-value	LogFC of DE	p-value of DE	pBH of DE
1369775_at	Nucks1	1.187	4.105	0	0	0.129	0.008	0.074
S74393_s_at	Pax6	0.881	3.944	0	0.016	0.073	0.014	0.086
M27905_at	Rpl21	0.884	3.965	0	0.016	0.093	0.020	0.1
1388783_at	Hmgb1*	1.124	3.921	0	0.016	0.095	0.055	0.155
U93692_at	Nup88	0.814	3.665	0	0.026	0.083	0.004	0.056
J01436cds_s_at	Cytb	0.827	3.726	0	0.026	0.052	0.119	0.232
1373952_at	Prkag2	1.023	3.610	0	0.033	0.089	0.013	0.084
U78090_s_at	Alg10	0.780	3.522	0	0.033	0.061	0.041	0.135
AB002111_at	Pex12	0.780	3.534	0	0.033	0.100	0.001	0.034
1389373_at	Smad1*	0.949	3.375	0	0.04	0.099	0.045	0.144
Up-regulated genes		Effect size results			Differential expression results			
Probe ID	Symbol	Effect size (ES)	ES z-value	p-value of z	pBH of z-value	LogFC of DE	p-value of DE	pBH of DE
1390518_at	Emid1	-1.259	-4.314	0	0	-0.063	0.049	0.148
rc_AA891838_at	Mrto4	-0.951	-4.224	0	0	-0.095	0.000	0.013
1389264_at	Ankrd54	-1.149	-3.983	0	0.016	-0.088	0.004	0.056
1369203_at	Wif1*	-0.980	-3.478	0	0.034	-0.056	0.023	0.107
U19866_at	Arc	-0.764	-3.46	0	0.034	-0.215	0.000	0.008
1376569_at	Klf2*	-0.960	-3.405	0	0.04	-0.182	0.000	0.013
rc_AA800613_at	Zfp36	-0.750	-3.396	0	0.04	-0.089	0.008	0.073
1398380_at	Vwa1	-0.94	-3.350	0	0.04	-0.098	0.002	0.038
1368451_at	Hrh3*	-0.940	-3.349	0	0.04	-0.094	0.005	0.063
S49760_g_at	Dgka	-0.711	-3.226	0.0013	0.051	-0.081	0.006	0.063

Top genes identified by IPA are indicated by an asterisk (*). Legends: ES, effect size; pBH, p-value with Benjamini and Hochberg correction; FC, fold change; DE, differentially expressed.

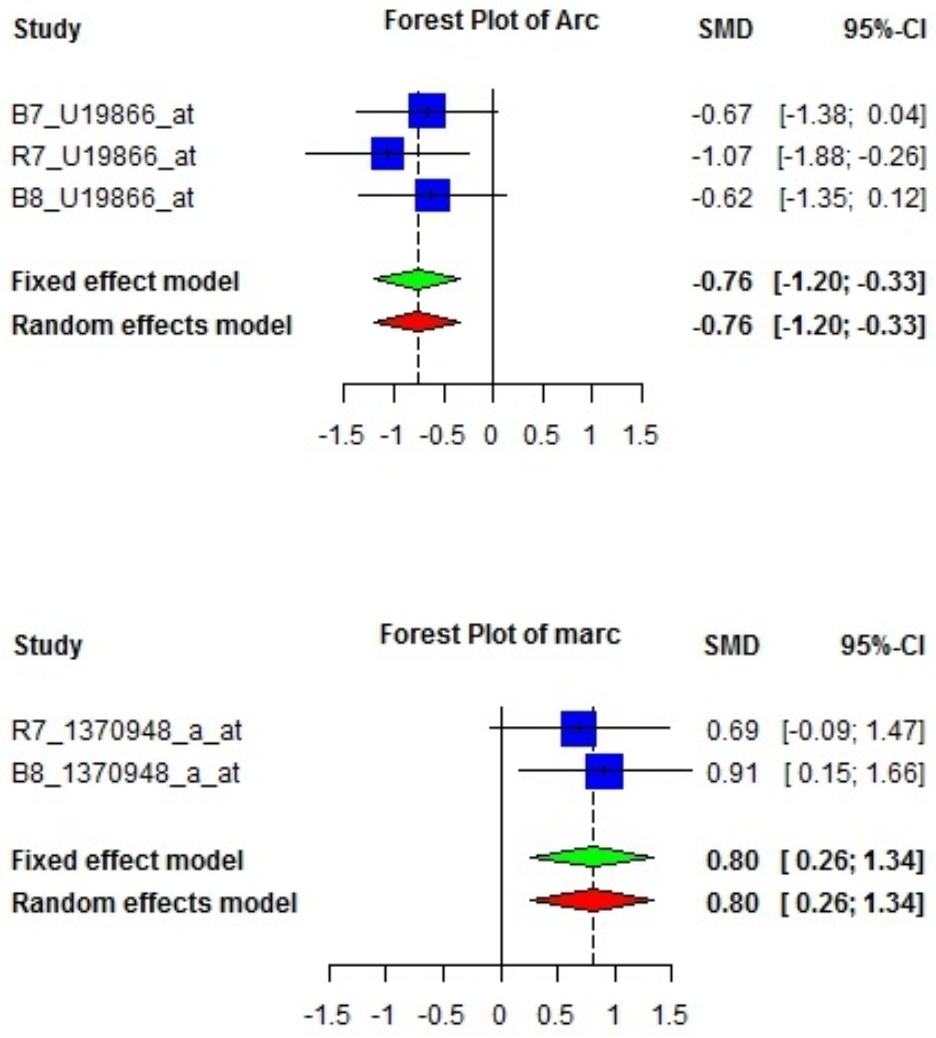


Figure 3.3 Forest plots of two representative significant genes in the aged-impaired rats. *Arc* is down-regulated (top) and *Marcks* is up-regulated (bottom) in the aged-impaired rats. For the selected probe set for each gene the individual study specific SMD and their 95% confidence intervals (CI) are plotted and shown on each row. The effect size results are shown at the bottom of each plot.

Table 3.6 Major functions associated with the top five networks in the IU comparison.

Network ID	Top functions associated with the networks	IPA score	Total focus genes
1	Neurological disease, tissue morphology	29	27
2	Cellular growth and proliferation, cancer, cell death and survival	16	19
3	Cell-to-cell signaling and interaction, nervous system development and function, carbohydrate metabolism	14	18
4	Cell death and survival, cellular development, hematological system development and function	10	15
5	Cell death and survival, metabolic disease, cellular function and maintenance	8	13

Table 3.7 Top canonical pathways in the IU comparison.

Name	p-value	Ratio
Nur77 signaling in T lymphocytes	6.13E-04	13/51 (0.255)
nNOS signaling in neurons	5.13E-03	12/46 (0.261)
Glutamate receptor signaling	5.68E-03	13/60 (0.217)
Calcium-induced T lymphocyte apoptosis	1.07E-02	12/57 (0.211)
Glutamate dependent acid resistance	1.48E-02	2/2 (1)

3.3.3 Aging and learning related genes

I searched the IPA knowledge base for genes that are annotated as aging related and genes that are annotated as learning related, particularly spatial learning. IPA recorded a total of 93 genes related to general aging in its database, of which five, *Adraid* (all-trans retinoic acid-induced differentiation factor), *Aldoc* (aldolase C, fructose-bisphosphate), *Clu* (clusterin), *ApoE* (apolipoprotein E), and *Mapt* (microtubule-associated protein tau) (Figure 3.4) were present in my AY significant gene list (p-value ≤ 0.05). Further, there were 401 genes annotated as learning genes in the IPA knowledge base, of which 177 were categorized under spatial learning. Among these learning genes 86 (30 of which were spatial learning related) were present in my AY comparison (Table S3 in (Uddin and Singh, 2013)) and 48 (15 of which were spatial learning related) were present in the IU comparison (Table S4 in (Uddin and Singh, 2013)) with p-value ≤ 0.05. Among the 86 genes for AY and 48 genes for IU, 15 were found common. These genes were considered as the ASLI associated genes.

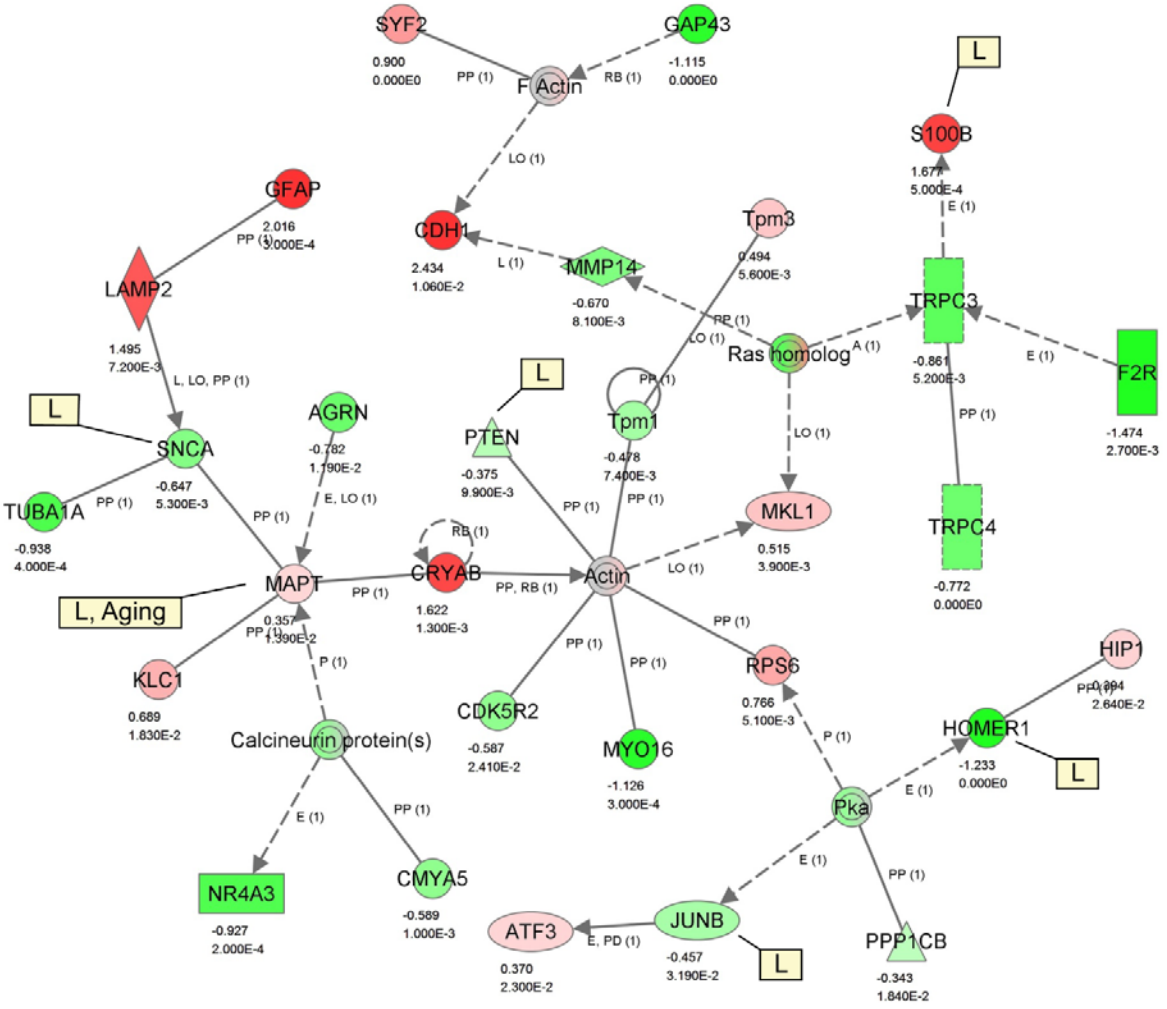


Figure 3.4 Network AY-3 from the AY comparison. Major functions of this network are cellular assembly and organization, tissue development, and cell morphology. Each biological relationship (an edge) between two genes (nodes) is supported by at least one reference from the literature or curated information stored in the IPA knowledge base. The intensity of the node color indicates the degree of up- (red) or down- (green) regulation represented by the effect size as observed in the AY comparison (see Section 3.3.2.1). The effect size and p-value for each gene is shown below the gene symbol. Edges are displayed with various labels that describe the nature of relationship between the genes (e.g. P for phosphorylation, PP for protein-protein binding, PD for protein-DNA binding, A for activation, E for expression, L for proteolysis, LO for localization, RB for regulation of

binding). Any specific findings for a gene whether it is associated with aging (A), learning (L), and/or spatial learning (SL) is presented inside a rectangle beside that gene.

3.4 Discussion

3.4.1 Effective meta-analysis necessitates proper data integration

Meta-analysis has emerged as an essential tool in modern genetic and genomic analysis (Goldstein and Guerra, 2010). It can uncover a significant effect from a combined analysis as integration of a broader and/or richer collection of data has the potential to generate results that have greater confidence, and place less reliance on a single dataset (Goldstein and Guerra, 2010; Ramasamy et al., 2008). I followed a set of standard generalized steps critical for effective meta-analysis that had been recommended in the literature (Chang et al., 2013; Goldstein and Guerra, 2010; Moreau et al., 2003; Nazri and Lio, 2012; Ramasamy et al., 2008; Stevens and Doerge, 2005). I formulated a set of specific objectives and explicitly defined the outcome to be extracted from each study (Section 1.14). I identified relevant primary studies, established inclusion/exclusion criteria for these studies as well as detailed data collection and selection processes, and executed careful data quality control and preprocessing on the selected data for meta-analysis (Chapter 2). In this chapter, I decided on the meta-analysis methods including ways to handle between-study heterogeneity. I performed a random effect size meta-analysis by keeping the individual studies separate and then only combining the probe set specific effects. I also performed the traditional differential expression analysis in parallel to the effect size analysis after merging all probe set data into a single pool through the process of cross-study and cross-platform data normalization (Figure 3.1). Even though the differential expression analysis was able to detect significant differential expression level, the difference was smaller compared to the effect size. Overall, the effect size analysis seems to be a better approach than differential expression analysis, particularly when combining data from different studies and platforms. Nonetheless, the differential expression results helped me verify the effect size outcomes and better screen the aging and ASLI associated genes.

It is important to point out that during the data integration process I worked at the probe set level rather than at the gene level. This is essential when combining data from independent microarray results from different platforms. Therefore, I integrated all data first before doing any filtering, annotation, and summarization. In the final filtering process some genes with higher effect sizes and p-values (of the effect size z-values) > 0.1 were removed. This was based on the observation that, a gene may have a higher effect size but not necessarily a lower p-value (Table S1 and S2 in (Uddin and Singh, 2013)). This is due to either the heterogeneity among studies or the fact that some datasets are lacking the expression information for that particular probe set. Also the genes whose treatment effect sizes are either zero or close to zero have higher p-values. These genes were therefore filtered out. The data integration method adopted has prevented any loss of information and generated a number of differentially expressed genes even after multiple testing corrections (Table S1 and S2 in (Uddin and Singh, 2013)) particularly for the AY comparison. It is also important to mention that effect size estimates of some of these genes e.g. *C3* and *Tubb2b* (Figure 3.2) present some degree of heterogeneity. It is not unexpected in a meta-analysis as the heterogeneity may arise, as in this case, from differences in the details of the Morris water maze training, memory test and sample collection procedure, and other experimental variables pertaining to the individual studies. However, during the selection of the aging and ASLI related genes that had high heterogeneity, I made sure that the estimates of the effect size are in the same direction.

3.4.2 Knowledge based gene networks provide useful insight with some limitations

Functional and pathway analysis was performed using IPA. Beside IPA, a number of software other programs are available, which can streamline the data analysis process, including pathway analysis (e.g. Partek (www.partek.com), GENIES (Kotera et al., 2012), and TM4 (Saeed et al., 2003)). Some also integrate various knowledge bases such as protein-protein interaction data, GO (gene ontology), and pathway information. These platforms provide a relatively quick analysis of microarray gene expression data for general biologists. However, in order to use these platforms, raw microarray data must be preprocessed

beforehand following standard pre-processing methods, proper quality control must be maintained, and at each step the output results must be verified for accuracy and consistency in the biological context being investigated. Although some of them offer functions to perform some data preprocessing or meta-analysis from microarray data, but they are limited. IPA was the best choice to perform functional and pathway analyses using the significant genes generated from my effect size meta-analysis, mainly, to take advantage of its rich manually curated biological knowledge base as well as its built in network construction methods

In order to include more genes in IPA, I considered the unadjusted p-value (≤ 0.05) of the random effect size for gene selection (and used 2245 genes from the AY comparison). Also, I analyzed data in IPA with lower number of genes following more stringent criteria such as using p-value ≤ 0.005 (e.g. 888 genes) or BH corrected p-value ≤ 0.05 (e.g. 1753 genes) for the AY comparison. It was satisfactory to note that the IPA analysis returned similar results. Also the expression levels (up- or down-regulation) identified in this meta-analysis for all or most of the genes in each function annotation category in the AY comparison did correspond to the predicted activation state (decreased or increased) for that category as supported by the literature in the IPA knowledge base. Further, I was able to verify the results by literature review using PubMed. The gene networks created in IPA (Figure 3.4 and Appendix 6.3.1 to Appendix 6.4.4) show how the significant genes may interact with other genes in networks. The results indicate that majority of the learning genes reside in the periphery on these networks with only a few (e.g. one or two) interactions. For example, in the AY networks, three out of four learning genes (e.g. *Camk4* (calcium/calmodulin-dependent protein kinase IV), *Synj1*, and *L1cam*) in network AY-1, all four learning genes in (e.g. *Nfkb1*, *Crem*, *C3*, and *Slc11a2*) in network AY-2, four out of six learning genes (e.g. *Pten*, *Homer1*, *JunB*, and *S100b*) in network AY-3 (Figure 3.4), and four out of five learning genes (e.g. *Tnc*, *Drd5*, *Arc*, and *Prkar2b*) in network AY-5 share only one or two interactions with other genes in the network. Similar observation can be made in the other AY or IU networks. In addition, the results show that some of the significant genes may function as hub genes. A hub gene is usually a gene with many connections or interactions with other

genes. For example, the gene *Cacna1b* (calcium channel, voltage-dependent, N type, alpha 1B subunit) interacts with a large number of other genes in network AY-1 (Appendix 6.3.1), which include the learning genes *Camk4*, *Synj1* and *Grm4*. It also includes *Tubb2b*, *Mapre2*, and other potassium and calcium channel genes (e.g. *Kcna1* (potassium channel, voltage gated shaker related subfamily A, member 1), *Kcnma1* (potassium channel, calcium activated large conductance subfamily M alpha, member 1), *Kcna4* (potassium channel, voltage gated shaker related subfamily A, member 4), and *Cacnb2* (calcium channel, voltage-dependent, beta 2 subunit)).

The protein encoded by *Cacna1b* (effect size -0.421 and p-value 0.0039) is a pore-forming subunit of an N-type voltage-dependent calcium channel, which controls neurotransmitter release from neurons (Currie, 2010). The activity and kinetics of several types of calcium channels are regulated by *Cacnb2* (effect size -0.443 and p-value 0.0349) and this gene has recently been found as a risk locus for five major psychiatric disorders including autism spectrum disease (Breitenkamp et al., 2014). Neuronal Ca^{2+} plays a critical role as an intracellular second messenger, linking neuronal excitability with many kinds of cellular biological events including synaptic plasticity (Berridge et al., 1998; Bito, 1998; Bliss and Collingridge, 1993). Ca^{2+} ions bind to calmodulin (CaM) and form a complex, which mediates a significant part of signaling downstream. An important target for the Ca^{2+}/CaM complex is the Ca^{2+}/calmodulin-dependent protein kinases (CaMKs) (Bito and Takemoto-Kimura, 2003; Takemoto-Kimura et al., 2003). CaMKs such as *Camk4* can then activate a number of other targets such as CREB (cAMP responsive element binding protein) and play a significant role in learning and memory formation through the activation of CREB signaling (Baudry et al., 2014; Bito and Takemoto-Kimura, 2003; Miyamoto, 2006; Sweatt, 2001; Thomas and Huganir, 2004).

Voltage-gated potassium (Kv) channels like *Kcna1* (Kv1.1) (effect size 0.651 and p-value 0.0001) and *Kcna4* (Kv1.4) (effect size -0.399 and p-value 0.017) represent the most complex class of voltage-gated ion channels (Lai and Jan, 2006; McKeown et al., 2008). They serve a diverse function in the cell include regulating neurotransmitter release. Specific

potassium channels, gated by intracellular calcium elevation, have been associated with synaptic plasticity (Kurotani et al., 2013; Voglis and Tavernarakis, 2006). The Kv1 potassium channels are generally activated by the binding of the beta subunits (e.g. *Kcnab1* or *Kcnab2*) and play important role in learning and memory in the hippocampal pyramidal neurons. For example, deletion of *Kcnab1* in mice results in increasing neuronal excitability facilitating LTP induction and improving learning and memory in aged mice (Murphy et al., 2004). *Kcnma1* (effect size -0.507 and p-value 0.005) channels can be formed by 2 subunits: the pore-forming alpha subunit, which is the product of this gene, and the modulatory beta subunit. Intracellular calcium regulates the physical association between the alpha and beta subunits. *Kcnma1* has been implicated in cognitive impairments (Higgins et al., 2008).

Other noticeable hub genes include *Nfkb* complex (effect size 0.619 and p-value 0.0042) and PKC(s) in network AY-2 (Appendix 6.3.2), *Mapk1* (p38) (effect size -0.55 and p-value 0.14) in network AY-4 (Appendix 6.3.3), NMDA receptor in network AY-5 (Appendix 6.3.4), and the kinases (e.g. *Akt*, *ERKs*, and *PI3K*) in network AY-6 (Appendix 6.3.5). Some of these are also present as hub genes in the IU networks, such as *Nfkb* complex, *Akt* (protein kinase B or PKB), and ERKs (mitogen-activated protein kinases) in network IU-1, and PI3K (phosphatidylinositol 3-kinase) in network IU-4. The broader implications of some of these genes in ASLI are discussed in more details in Chapter 5.

It is important to note that Fischer 344 strain of rats have a median life-span of 23-31 months in captivity (Coleman et al., 1977; Sass et al., 1975). Their normal age-related incidence of neoplasms and degenerative diseases is high, particularly, once the rats pass 24 months of age (Coleman et al., 1977; Sass et al., 1975). Also, the effect of aging and ASLI on brain gene expression is evident in the aged (24-26 months old) in comparison to the young (3-6 months old) rats. Indeed, it is expected that studies on animals beyond 26 weeks of age may show involvement of additional genes in this phenomenon and the effects observed could be more pronounced at later stages of the rat's life-span.

In summary, I followed a set of standard generalized steps critical for effective meta-analysis that had been recommended in the literature. I performed a random effect size

meta-analysis by keeping the individual studies separate and then only combining the probe set specific effects. The probe set level data integration method adopted here has prevented any loss of information and generated a larger number of differentially expressed genes even after multiple testing corrections. GO and pathway analysis results relating to these genes support the fact that the genes and pathways identified in this analysis follow biological expectations. The genes identified (Table 3.2) are known to partake in aging and in learning impairments. This conclusion is also supported by follow up analysis including regulatory interaction networks based on known functions and interaction. However, the pathway analysis reveals three important shortcomings of such traditional analysis using microarray gene expression data: 1) the regulatory interaction relationships among genes are based on curated information from published literature stored in biological knowledge data base only; 2) the genes known as aging or learning based on the current biological knowledge are all scattered in different networks; and 3) the hub genes express at a comparatively lower level. Therefore, results from the analysis in this chapter are not able to provide a complete picture as to how the candidate learning genes co-express in the context of aging as well as learning or how their combined action may contribute to the ASLI phenotype in rats. Particularly, this type of pathway analysis is limited for genes for which no interaction or regulatory information is available in the literature or biological knowledge base.

3.5 References

Baudry, M., Zhu, G., Liu, Y., Wang, Y., et al., 2014. Multiple cellular cascades participate in long-term potentiation and in hippocampus-dependent learning. Brain Res.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 57**,** 299-314.

Berridge, M.J., Bootman, M.D., Lipp, P., 1998. Calcium--a life and death signal. Nature. 395**,** 645-8.

Bito, H., 1998. The role of calcium in activity-dependent neuronal gene regulation. Cell Calcium. 23**,** 143-50.

Bito, H., Takemoto-Kimura, S., 2003. Ca(2+)/CREB/CBP-dependent gene regulation: a shared mechanism critical in long-term synaptic plasticity and neuronal survival. Cell Calcium. 34**,** 425-30.

Bliss, T.V., Collingridge, G.L., 1993. A synaptic model of memory: long-term potentiation in the hippocampus. Nature. 361**,** 31-9.
Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. Introduction to Meta-Analysis, Vol., Wiley.
Breitenkamp, A.F., Matthes, J., Nass, R.D., Sinzig, J., et al., 2014. Rare mutations of CACNB2 found in autism spectrum disease-affected families alter calcium channel function. PLoS One. 9**,** e95579.
Burger, C., Lopez, M.C., Feller, J.A., Baker, H.V., et al., 2007. Changes in transcription within the CA1 field of the hippocampus are associated with age-related spatial learning impairments. Neurobiol Learn Mem. 87**,** 21-41.
Burger, C., Lopez, M.C., Baker, H.V., Mandel, R.J., et al., 2008. Genome-wide analysis of aging and learning-related genes in the hippocampal dentate gyrus. Neurobiol Learn Mem. 89**,** 379-96.
Chang, L.C., Lin, H.M., Sibille, E., Tseng, G.C., 2013. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics. 14**,** 368.
Cochran, W., 1937. Problems arising in the analysis of a series of similar experiments. J R Stat Soc. 102-118.
Coleman, G.L., Barthold, W., Osbaldiston, G.W., Foster, S.J., et al., 1977. Pathological changes during aging in barrier-reared Fischer 344 male rats. J Gerontol. 32**,** 258-78.
Currie, K.P., 2010. G protein modulation of CaV2 voltage-gated calcium channels. Channels (Austin). 4**,** 497-509.
Fleiss, J.L., 1993. The statistical basis of meta-analysis. Stat methods Med Res. 121-145.
Goldstein, D.R., Delorenzi, M., Luthi-Carter, R., Sengstag, T., 2010. Comparison of meta-analysis to combined analysis of a replicated microarray study. In: Meta-analysis and combining information in genetics and genomics. Chapman & Hall/CRC Mathematical and Computational Biology Series, Vol., D.R. Goldstein, R. Guerra, ed.^eds. CRC Press, pp. 135-156.
Goldstein, D.R., Guerra, R., 2010. A brief introduction to meta-analysis, genetics and genomics. In: Meta-analysis and combining information in genetics and genomics. Chapman & Hall/CRC Mathematical and Computational Biology Series, Vol., D.R. Goldstein, R. Guerra, ed.^eds. CRC Press, pp. 3-20.
Higgins, J.J., Hao, J., Kosofsky, B.E., Rajadhyaksha, A.M., 2008. Dysregulation of large-conductance Ca2+-activated K+ channel expression in nonsyndromal mental retardation due to a cereblon p.R419X mutation. Neurogenetics. 9**,** 219-23.
Kotera, M., Yamanishi, Y., Moriya, Y., Kanehisa, M., et al., 2012. GENIES: gene network inference engine based on supervised analysis. Nucleic Acids Res. 40**,** W162-7.
Kurotani, T., Miyashita, T., Wintzer, M., Konishi, T., et al., 2013. Pyramidal neurons in the superficial layers of rat retrosplenial cortex exhibit a late-spiking firing property. Brain Struct Funct. 218**,** 239-54.
Lai, H.C., Jan, L.Y., 2006. The distribution and targeting of neuronal voltage-gated ion channels. Nat Rev Neurosci. 7**,** 548-62.

McKeown, L., Swanton, L., Robinson, P., Jones, O.T., 2008. Surface expression and distribution of voltage-gated potassium channels in neurons (Review). Mol Membr Biol. 25**,** 332-43.

Miyamoto, E., 2006. Molecular mechanism of neuronal plasticity: induction and maintenance of long-term potentiation in the hippocampus. J Pharmacol Sci. 100**,** 433-42.

Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., et al., 2003. Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet. 19**,** 570-7.

Murphy, G.G., Fedorov, N.B., Giese, K.P., Ohno, M., et al., 2004. Increased neuronal excitability, synaptic plasticity, and learning in aged Kvbeta1.1 knockout mice. Curr Biol. 14**,** 1907-15.

Nazri, A., Lio, P., 2012. Investigating meta-approaches for reconstructing gene networks in a mammalian cellular context. PLoS One. 7**,** e28713.

Ramasamy, A., Mondry, A., Holmes, C.C., Altman, D.G., 2008. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med. 5**,** e184.

Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D., et al., 2002. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res. 62**,** 4427-33.

Saeed, A.I., Sharov, V., White, J., Li, J., et al., 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 34**,** 374-8.

Sass, B., Rabstein, L.S., Madison, R., Nims, R.M., et al., 1975. Incidence of spontaneous neoplasms in F344 rats throughout the natural life-span. J Natl Cancer Inst. 54**,** 1449-56.

Sirbu, A., Ruskin, H.J., Crane, M., 2010. Cross-platform microarray data normalisation for regulatory network inference. PLoS One. 5**,** e13822.

Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 3**,** Article3.

Stevens, J.R., Doerge, R.W., 2005. Combining Affymetrix microarray results. BMC Bioinformatics. 6**,** 57.

Sweatt, J.D., 2001. The neuronal MAP kinase cascade: a biochemical signal integration system subserving synaptic plasticity and memory. J Neurochem. 76**,** 1-10.

Takemoto-Kimura, S., Terai, H., Takamoto, M., Ohmae, S., et al., 2003. Molecular cloning and characterization of CLICK-III/CaMKIgamma, a novel membrane-anchored neuronal Ca2+/calmodulin-dependent protein kinase (CaMK). J Biol Chem. 278**,** 18597-605.

Thomas, G.M., Huganir, R.L., 2004. MAPK cascade signalling and synaptic plasticity. Nat Rev Neurosci. 5**,** 173-83.

Uddin, R.K., Singh, S.M., 2013. Hippocampal gene expression meta-analysis identifies aging and age-associated spatial learning impairment (ASLI) genes and pathways. PLoS One. 8**,** e69768.

Voglis, G., Tavernarakis, N., 2006. The role of synaptic ion channels in synaptic plasticity. EMBO Rep. 7**,** 1104-10.

Chapter 4 Gene Network Analysis

4 Gene network construction using the WGCNA approach identifies a key ASLI network module and several candidate hub genes

4.1 Introduction

Mathematical modeling of gene networks from gene-expression data begins with the understanding that the information captured by microarray experiments is far richer than that which is obtained by a list of differentially expressed genes. The availability of large scale genome-wide gene expression microarray data has inspired the development of a large number of gene network inference algorithms as reviewed in Chapter 1. Literature review in PubMed suggests that mathematical modeling approaches utilizing steady-state gene expression data to model consensus gene networks have been in frequent use (e.g. probabilistic Bayesian networks (Friedman et al., 2000), partial-correlation based approaches (GeneNet) (Opgen-Rhein and Strimmer, 2007), information-theoretic approaches (ARACNE) (Margolin et al., 2006), and correlation based methods (WGCNA) (Langfelder and Horvath, 2008)). However, many of these methods suffer from major limitations. Though Bayesian network approaches appear to be highly promising, there is rarely any publication where the authors demonstrate wet-bench/experimental validation of the implementation of their algorithm. A major practical challenge in using Bayesian networks to infer gene network is that the structure learning of the network is NP-hard (non-deterministic polynomial-time hard) for score-based approaches (Chickering, 1996). Also the network learning process is computationally complex, as the number of possible graphs increases super-exponentially with the number of genes, and an exhaustive search is untraceable. Hence, Bayesian networks or dynamic Bayesian networks can be applied only to relatively small networks (Ahmad et al., 2012; Emmert-Streib et al., 2012).

ARACNE, another very popular mutual information based approach, also shows a fate similar to Bayesian networks (i.e. a lack of any experimentally verified publications following the original work)(Basso et al., 2005). Several studies have compared a number of these popular algorithms to model gene networks, and provided comprehensive evaluations and suggestions to help choose proper statistical methods for constructing large scale gene networks (Allen et al., 2012; Maetschke et al., 2014; Nazri and Lio, 2012; Song et al., 2012; Villaverde and Banga, 2014). They compared and evaluated the methods in terms of sensitivity and specificity in identifying the true connections and the correct hub genes, the ease of use, and computational speed. For example, Allen et al. (2012) compared eight different methods such as GeneNet (Opgen-Rhein and Strimmer, 2007), SPACE (Sparse PArtial Correlation Estimation) (Peng et al., 2009), WGCNA, ARACNE, and four Bayesian Networks methods such as BNArray (Chen et al., 2006), B-course (Myllymaki et al., 2002), BNT (murphy, 2001), and Werhli's implementation of Bayesian network (Werhli et al., 2006). They concluded that each method has its own advantages, however, GeneNet, WGCNA, and ARACNE performed well in constructing the global network structure with simulated data; GeneNet and SPACE performed well in identifying a few connections with high specificity. With real *E. coli* data, their results indicated that WGCNA and ARACNE performed best and were relatively more robust. Moreover, WGCNA methods were suitable for detecting network modules or sub-networks, identifying hub genes (which are likely to be the disease driver genes), and selecting candidate genes as biomarkers.

Maetschke et al. (2014) recently compared the prediction accuracy of 17 different unsupervised methods that included popular methods such as Pearson's correlation (used in WGCNA), ARACNE, MRNET (Meyer et al., 2007), CLR (Faith et al., 2007), relevance networks (Butte et al., 2000), and GENIE (Kotera et al., 2012). They also compared these methods against successful supervised and semi-supervised methods. Their conclusion is that simple correlation methods such as Pearson correlation are as accurate as much more complex methods, yet much faster and parameter-less. Song et al. (2012) compared correlation and mutual information based approaches, and

confirmed close relationships between these methods. Moreover, they suggest that robust measure of correlation leads to modules that are superior to mutual information based modules in terms of gene ontology enrichment.

In summary, among the methods that are used to model causal interactions and networks, there is a lack of agreement in the scientific community as to which method performs best. Many methods, or their variations, have been claimed to be performing better than others. However, only mathematical (and seldom experimental) evidence is provided to support these claims. For many methods, there are too many variations (particularly for Bayesian networks) (Ahmad et al., 2012; Beal et al., 2005). Many, understandable only to the mathematician, statistician, or computational biologist, are not in an easy-to-use format and require a steep learning curve. For general biologists with limited computational knowledge, it poses a challenge to select a suitable gene network algorithm to generate biologically meaningful networks. Therefore, correlation-based methods such as WGCNA are gaining popularity in the biological scientist community.

Correlation networks are widely used to explore, analyze, and visualize high-dimensional data, for example in finance (Mangegna and Stanley, 2000), gene expression analysis (Butte et al., 2000; Mason et al., 2009; Miller et al., 2008; Oldham et al., 2006; Oldham et al., 2008; Plaisier et al., 2009; Rickabaugh et al., 2015; Ye and Liu, 2015), or metabolomics (Steuer, 2006). Their popularity is owed to a large extent to the ease with which a correlation network can be constructed, as this requires only two simple steps: i) the computation of all pairwise correlations for the investigated variables, and ii) a thresholding or filtering procedure to identify significant correlations, and hence edges of the network (Opgen-Rhein and Strimmer, 2007). Correlation-based methods are the most straightforward way to explore the gene co-expression network.

When the mRNA expression of two or more genes are correlated across multiple samples, these genes are said to be ‘coexpressed’ (Gaiteri et al., 2014). Correlation network analysis using WGCNA can infer these co-expression links from microarray

expression profiles. WGCNA has been successfully used in recent years in a number of biological and cellular contexts (over 30 publications in 2014 alone) (Fontenot and Konopka, 2014; Fuller et al., 2007; Levine et al., 2013; Mason et al., 2009; Miller et al., 2010; Oldham et al., 2008; Plaisier et al., 2009; Rickabaugh et al., 2015; Ye and Liu, 2015). It enables a more systematic and global interpretation of gene expression data. WGCNA takes an unbiased approach for ascertaining the relationships among all genes queried across all samples in a dataset (Langfelder and Horvath, 2008; Zhang and Horvath, 2005), and identifies biologically meaningful 'modules' that are often comprised of functionally related genes. Gene relationships within a given module can then be assessed using a number of visualization tools such as Visant (Hu et al., 2004) or Cytoscape (Shannon et al., 2003). The graphical representation of a module aids in rapidly identifying hub genes and other biologically meaningful patterns of co-expression within a given module. Overall, WGCNA provides an approach for prioritizing specific genes from large expression datasets, particularly those with biologically salient relationships that might otherwise be missed using differential expression approaches (Fontenot and Konopka, 2014).

As discussed in Chapter 3, differential expression analysis (followed by functional and pathway analysis using IPA) to identify ASLI gene networks was limited to the current IPA knowledge base. IPA pathway analysis could only model gene networks based on information that was available in the literature. Therefore, those analyses were unable to fully utilize the gene transcript expression information captured by the microarray data. WGCNA may overcome such limitations. Numerous studies have applied gene co-expression network analysis using WGCNA to associate co-expression modules with brain and psychiatric diseases (de Jong et al., 2010; Miller et al., 2008; Torkamani et al., 2010; Voineagu et al., 2011). Oldham et al. (2006) investigated the conservation and evolution of gene co-expression networks in human and chimpanzee brains, and shed light on the molecular bases of primate brain organization. Miller et al. (2008) employed WGCNA to explore commonalities and differences between normal aging and pathological aging in Alzheimer's disease, resulting in the identification of biologically

relevant modules conserved between Alzheimer's disease and aging. However, no study investigating gene network modeling in ASLI appeared in the literature. It was thus necessary to initiate such a study to explore and identify functional modules and gene hubs in the context of ASLI. In this chapter, I perform a co-expression network analysis (using WGCNA) to fulfill objective three as outlined in Chapter 1. The specific goals of this aspect of the study are: 1) to separate aged and young samples and create gene network models from the exploratory datasets, 2) to perform a differential network analysis between aged and young networks, and 3) to evaluate results (significant functional modules and hub genes) by comparing them against the validation datasets.

4.2 Methods

All data preparation steps including WGCNA, GO, and other statistical analyses were performed in R using appropriate software packages.

4.2.1 Data selection for network analysis

The five datasets, R7, B8, K9, B7, and BL (Table 2.3) were assessed and used in WGCNA. These datasets were already quality checked and normalized, and had outliers removed and batch effect adjusted (Table 2.4). For each dataset, aged and young samples were separated and assessed further for the presence of array outliers (Appendix 6.1.1 to Appendix 6.1.6). Since the WGCNA network construction method is correlation based, before proceeding with network analysis I wanted to make sure that the correlations between genes in each dataset were reasonable as suggested in the literature (Miller et al., 2010). This was done by calculating Pearson’s correlations between the expression levels of each pair of genes in the aged or young preprocessed datasets and by plotting the correlation values in histogram plots in R.

4.2.2 Co-expression network analysis using the WGCNA approach

Co-expression analysis using WGCNA generally begins with calculating pairwise correlations between all gene expression profiles in a dataset and creating an adjacency

matrix (i.e. a correlation matrix). Often the analysis is restricted to some fraction of the original gene set (typically several thousand of the most highly correlated genes). Once the correlation matrix is built, it is raised to a power to approximate scale free topology. Genes with highly correlated expression profiles are then grouped together in clusters called modules, and networks are constructed. Each module may correspond to biological pathways or similarly functionally associated groups of genes.

Using the preprocessed transformed data (genes in columns and samples in rows), gene networks were constructed for aged and young using the WGCNA R package (Langfelder and Horvath, 2008; Zhang and Horvath, 2005) following the approaches described in (Miller et al., 2008; Miller et al., 2010; Oldham et al., 2006; Oldham et al., 2008). The overall network analysis process for a single dataset is described below, which involves the following main steps.

1. Determining the weights or soft power beta
2. Creating an adjacency (connection strength) matrix
3. Filtering out genes with very low connectivity
4. Creating and visualizing a whole network
5. Creating and visualizing network modules
6. Exploring the functional significance of modules
7. Validating network modules
8. Differential network analysis of young vs. aged
9. Identifying and validating hub genes

4.2.3 Determining the weights or soft power beta

Once it was confirmed that the correlations between genes in each dataset were reasonable, soft threshold power beta (Appendix 6.15.1) was determined for each dataset by using the function pickSoftThreshold(. . .) in the WGCNA R package.

4.2.4 Creating an adjacency (connection strength) matrix

A weighted correlation between two genes represents connection strength between the genes in a network. For each dataset, a network adjacency or connection strength matrix (network data) (Appendix 6.15.1) was created by taking the signed correlations of the gene expression values between each pair of genes raised to a power of beta. Beta is the weight, a soft threshold, and was determined in such a way so that the resulting network follows approximate scale free topology. The values in the diagonal (self-correlation) were converted to zero.

4.2.5 Filtering out genes with very low connectivity

To save computational time, genes were filtered out from a network adjacency matrix based on their connectivity (i.e. only genes with reasonably high connectivity were kept for network analysis). The overall connectivity for each gene (denoted by *k*) is the sum of connection strengths (weighted correlation) between that gene and all other genes in the network. It is scaled to lie between 0 and 1 and represents how strongly a gene is connected to all other genes in the network.

4.2.6 Creating and visualizing a whole network

A co-expression network can be created using all the genes in an adjacency matrix. In a co-expression network, an edge between two genes (nodes) represents a co-expression relationship. For each dataset a network interaction file was created from its adjacency matrix (see Appendix 6.8.1 for details), and used in Cytoscape for visualization and analysis.

4.2.7 Creating and visualizing network modules

Following filtering (Section 4.2.5), an adjacency matrix contained genes with reasonably high network connectivity. This adjacency matrix was used to determine a network topological overlap, construct a hierarchical clustering dendrogram of 1 – topological overlap, determine network modules using a hybrid tree-cutting algorithm, and to visualize network modules.

Network analysis often results in a large number of modules. It is sometimes useful to reduce the number of modules by merging those whose expression profiles are very similar. This was accomplished by the WGCNA function mergeCloseModules(…), which merged modules whose member genes were highly co-expressed. To calculate the co-expression similarity of entire modules, their module eigengenes were calculated. The module eigengene is defined as the first principal component of a given module. It can be considered as a representative of the gene expression profiles in a module (see Appendix 6.15.1 for details). The module eigengenes were clustered on their consensus correlation, which was the minimum correlation across the two sets.

4.2.8 Exploring the functional significance of modules

A list of genes belonging to each network module was exported to tab delimited text files along with all necessary information. For each module there were two files, the first file contained a list of genes with their gene symbols, mean expression, module names, and intramodular connectivity. This file was used for GO analysis using DAVID (The Database for Annotation, Visualization and Integrated Discovery) (**http://david.abcc.ncifcrf.gov/**) (Huang da et al., 2009a; Huang da et al., 2009b; Huang et al., 2007). The second file contained co-expression interaction information between each pair of genes in a module along with the topological overlap and correlation information. This interaction file was used for network visualization and analysis.

Functional Annotation Clustering analysis was performed In DAVID using the gene list for each young network module. Since, the network analysis generated a large number of modules in young and aged groups in multiple datasets, it was not efficient to perform online analysis one module at a time. In this research, DAVID web-services were accessed programmatically to perform GO analysis by using an R package called RDAVIDWebService (Fresno and Fernandez, 2013). It is a versatile R interface to all DAVID web service functionalities.

In summary, a DAVIDWebService object is created inside R; desired annotation categories e.g. GO biological process, molecular function, and cellular component are set; and a gene list (belonging to a module) along with the species name and background gene list is passed to the DAVID database online. The R object would in turn retrieve all requested information from the DAVID database. Since gene symbols can be confusing and often fail to produce a perfect match, the corresponding affymetrix IDs were used to query the DAVID database. GO functional annotation information was obtained for all modules in the young and the aged categories.

The functional Annotation Clustering analysis function in DAVID uses a novel algorithm to measure relationships among the annotation terms based on the degrees of their co-association genes, and organizes functionally similar gene groups into functional annotation clusters (http://david.abcc.ncifcrf.gov/) (Huang da et al., 2009a; Huang da et al., 2009b; Huang et al., 2007). The clustering algorithm is based on the hypothesis that similar annotations should have similar gene members. The algorithm adopts kappa statistics to quantitatively measure the degree of the agreement as to how genes share the similar annotation terms. It uses fuzzy heuristic clustering to classify the groups of similar annotations according to kappa values. The Kappa Statistic is a chance corrected measure of agreement between two sets of categorized data. In this sense, the more genes share common annotations, the higher the chance they will be grouped together. In DAVID, for each functional cluster an enrichment score is calculated. This enrichment score is the geometric mean (in -log scale) of the p-values of all member annotation terms and is used to rank their biological significance (Huang da et al., 2009b). Thus, the top ranked annotation clusters will most likely have consistently lower p-values for their annotation members.

The significance of a gene-enrichment p-values for each annotation term is first calculated based on a modified Fisher exact test method known as the EASE score (Hosack et al., 2003), which is more conservative than the Fisher exact p-value (Huang da et al., 2009b). For example, in human genome backgrounds (30,000 gene total), 40

genes are involved in the p53 signaling pathway. A given gene list has found that three out of 300 genes belong to the p53 signaling pathway. Then the question is if 3/300 is more than by random chance compared to the human background of 40/30000. Usually a p-value has to be equal or smaller than 0.05 for it to be considered strongly enriched in the annotation categories. The default threshold of the EASE score was set at 0.1.

4.2.9 Validating network modules

Network modules for young and aged were compared across studies and platforms for their repeatability. This was done in two ways: a) module preservation and b) module overlap.

4.2.9.1 Module preservation

Module preservation statistics (Langfelder et al., 2011; Miller et al., 2010; Zhang and Horvath, 2005) can qualitatively and quantitatively measure network preservation at the module level. This statistics is implemented in the WGCNA R package (Langfelder and Horvath, 2008). The module preservation analysis in this research was performed following the recommendation in Miller et al. (2010). As a qualitative assessment, the gene module assignment from one network is mapped on the same genes in the second network. The results are then plotted in a dendrogram, which offers a visual mean to qualitatively assess preservation.

Quantitative measure of network preservation at the module level takes advantage of the modulePreservation(…) function built into the WGCNA R library. This function assesses how well a module in one study is preserved in another study using a number of statistics. Module preservation was estimated quantitatively between the young and the aged networks in different datasets. In all comparisons, the R7 top most connected genes, their transcription profiles, and their module assignments were used as a reference.

4.2.9.2 Module overlap

Comparing networks by calculating module overlap allows one to determine whether a module that was found in one dataset can also be found in another dataset (Horvath, 2011; Miller et al., 2010). For example, to validate the existence of a module, it is desirable to show that the module is reproducible in a second independent dataset. Module overlap is a cross-tabulation-based statistics implemented in the WGCNA function overlapTable(. . .). This function determines whether clusters or modules in an exploratory or reference dataset are found in a validation or test dataset. These statistics do not assume a network or do not require transcription profiles. Instead, module assignments in both the reference and the test data are needed. Fisher's exact test is used to calculate a p-value of significance of pair-wise module overlap.

In this research, module overlaps were calculated along with their significance of overlaps between the young modules and between the aged modules in different datasets following the approach described in Oldham et al. (2008). In brief, top most connectivity genes common between a network from R7 (exploratory set) aged (or young) and another aged (or young) network from a validation set were selected. Next, the module labels between the two networks were matched using the matchLabels(…) function in WGCNA. The purpose was to see which modules in one network contain a significant number of overlapping genes with modules in the second network. This function reassigns module labels in the second network such that corresponding modules are assigned the same color label. For example, the brown module in network N1 has a significant number of genes overlapping with the black module in network N2. The matchLabels(…) function will re-label the black module in network N2 as brown. After matching labels between the modules in exploratory and validation networks, the overlapTable(…) function was used to find percentage overlaps and significance p-values.

4.2.10 Differential network analysis of young vs. aged

Differential network analysis allows one to compare two different networks side by side, for example, between a control and a disease network. Networks for several interesting modules identified in this research were visualized side by side between the young and aged groups using Cytoscape and compared for their differential co-expression.

4.2.11 Identifying and validating hub genes

Top hub genes were identified using the method topGenesKME(…) (Miller et al., 2010). This method determines which genes have extremely high module eigengene-based connectivity or k_{ME} values in both networks. Module eigengene-based connectivity k_{ME}, also known as module membership, is calculated for each gene. It is defined by correlating each gene's expression profile with the module eigengene of a given module (Langfelder and Horvath, 2008; Zhang and Horvath, 2005). Hub genes were validated by assessing their repeatability in networks constructed from independent datasets and by investigating their functions in relevant pathways.

4.2.11.1 Repeatability

Repeatability of the candidate hub genes were assessed as follows. For each module, hub genes present in the exploratory (R7) networks were checked for their presence as hub genes in the validation networks (e.g. B8, K9, or B7) with high k_{ME} values as well as with t-test p-values ≤ 0.05 (between two networks). In cases where a module from an exploratory network matched to multiple modules in a validation network, genes from multiple significant modules in the validation network were combined together and then compared to the hub genes in the exploratory network module.

4.2.11.2 Literature search

Literature searches were performed using PubMed to explore characteristics and functions of selected ASLI candidate hub genes and their relationship to learning and memory formation.

4.3 Results

In order to model, explore, and identify ASLI genes and their networks, I next describe the application of WGCNA to the current analysis. This analysis followed a detailed and through investigation that included the identification of GO enriched significant functional modules and hub genes, as well as validation of results using independent datasets. The results are described below.

4.3.1 Data selection for network analysis

Based on the quality of data and number of samples (Section 2.3.4 and Table 2.4), R7 aged (R7-A) and young (R7-Y) datasets were chosen as the exploratory datasets; B8 young (B8-Y), K9 young (K9-Y), B7 aged (B7-A), and B8 aged (B8-A) datasets were chosen as the validation datasets (Table 4.1). These six datasets were used for the construction of WGCNA networks. Since, after preprocessing, the B7 young, K9 aged, and both the BL young and aged groups did not have sufficient number of samples for WGCNA, they were excluded from this network analysis. The networks were constructed for each of the aged and young datasets separately (i.e. B7-A, B8-Y, B8-A, K9-Y, R7-Y, and R7-A). However, GO based functional analysis and visualization was done only for the networks from R7 young and aged exploratory datasets, and the results were validated independently in networks constructed from the validation datasets.

Table 4.1 Datasets selected for WGCNA network analysis.

Dataset	Number of young sample	Exploratory / Validation (Young)	Number of aged sample	Exploratory / Validation (Aged)
B7			28	Validation
R7	19	Exploratory	27	Exploratory
B8	18	Validation	28	Validation
K9	18	Validation		

4.3.2 Determining the weights or soft power beta

Since the WGCNA network construction approach is based on measuring pairwise gene-gene correlations, I wanted to make sure that the overall correlation quality would be acceptable before investing in extensive network analysis. The histogram plots of the correlations between genes in the aged and young preprocessed datasets show that the correlations are reasonable as they are mostly centered at zero (Figure 4.1).

The next step in network construction is to determine the soft threshold power. To choose a power beta for computing the connection strengths, the WGCNA function pickSoftThreshold(…) makes use of the scale-free topology criterion (Langfelder and Horvath, 2008; Zhang and Horvath, 2005). It focuses on the linear regression model fitting index (denoted as R^2 or scale.law.R.2) that quantifies how well a network satisfies a scale-free topology. The function calculates connectivity *k*, which for each gene is the sum of connection strengths with the other network genes. Connectivity was calculated for each gene in a dataset using a set of powers from 1 to 20. For each power the R^2 was then calculated and returned along with other information on connectivity. The function pickSoftThreshold(…) estimated an appropriate soft-thresholding power from the set. For each dataset, it also returned a data frame containing the fit indices for scale free topology.

Table 4.2 shows such a data frame from R7-Y as an example of a typical result. The columns contain the soft-thresholding power, adjusted R^2 for the linear fit, the linear coefficient, mean connectivity, median connectivity and maximum connectivity. Given this table the power six presents as the best overall fit for R7-Y with $R^2 = 0.88$ and mean.k = 39.35 (the target values are $R^2 > 0.8$ and mean.k > 30).

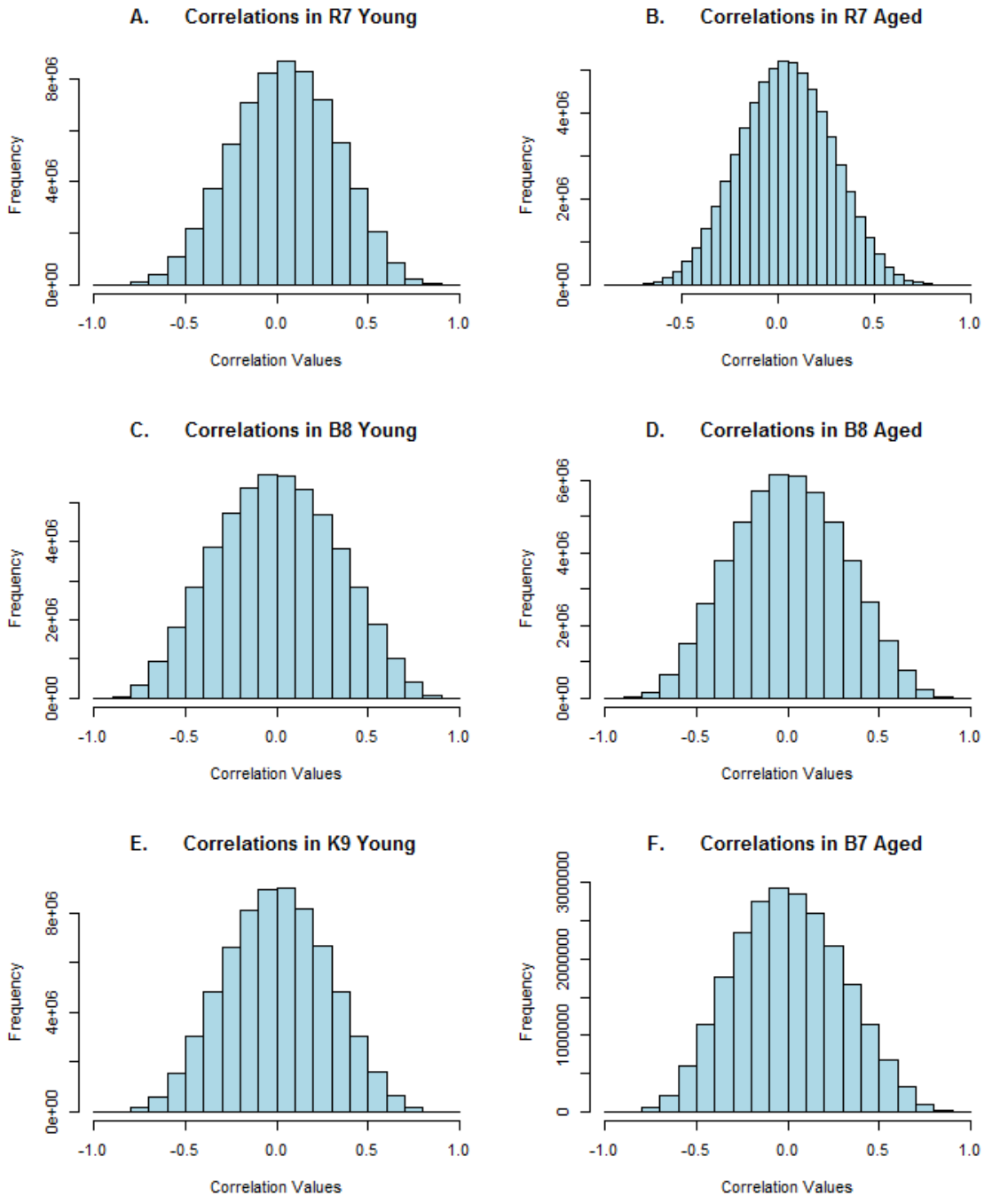


Figure 4.1 Histogram of correlations between genes in each dataset selected for WGCNA network analysis. Correlations are centered at zero for R7-Y, B8-Y, B8-A, K9-Y, and B7-A, and close to zero for R7-A.

Table 4.2 R7 young data power table. The first column lists the soft threshold Power, the second column reports the resulting scale free topology fitting index R^2 (scale.law.R.2), the third column reports the slope of the fitting line, the fourth column reports the fitting index for the truncated exponential scale free model, the remaining columns list the mean, median and maximum connectivity. The slope of the regression corresponds to the value gamma for the scale free distribution.

Power	R^2	slope	Truncated R^2	mean.k.	median.k.	max.k.
1	0.00	0.22	0.96	1850.26	1832.41	2620.20
2	0.13	-1.42	0.95	640.17	619.17	1214.80
3	0.36	-1.87	0.96	271.45	254.39	658.61
4	0.60	-1.93	0.98	130.99	118.04	394.23
5	0.78	-2.13	0.98	69.31	59.78	265.08
6	0.88	-2.27	0.98	39.35	32.37	191.35
7	0.93	-2.35	0.98	23.63	18.41	146.66
8	0.96	-2.33	0.98	14.87	10.91	116.58
9	0.98	-2.29	0.98	9.72	6.70	95.32
10	0.98	-2.21	0.98	6.58	4.26	79.68
12	0.98	-2.04	0.98	3.28	1.84	58.45
14	0.99	-1.88	0.99	1.80	0.87	44.89
16	0.99	-1.75	0.99	1.06	0.44	35.57
18	0.99	-1.65	0.99	0.67	0.23	28.82
20	0.99	-1.58	0.99	0.44	0.13	23.86

As examples of further analyses of network topology for various soft-thresholding powers, results from R7-Y and R7-A are shown in Figure 4.2 and Figure 4.3. These figures present some of the information from a power table (such as Table 4.2) in a graphical format. Figure 4.2.A and Figure 4.3.A show the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis) for the young and aged samples, respectively. These plots help to visualize how the scale-free fit depends on the power parameter beta. The smallest power of beta is chosen where the R^2 curve seems to saturate. The horizontal red line corresponds to $R^2 = 0.80$ as a general cutoff. Based on this scale free topology model fit analysis the soft-threshold power for both R7-Y and R7-A was determined to be 6. This power also results in an approximate straight line relationship in the scale-free topology plots in Figure 4.2.B and Figure 4.3.B. According to these plots, the black linear regression line leads to a fitting index of $R^2 = 0.86$ for R7-Y and $R^2 = 0.81$ for R7-A. The red line represents the better fit provided by an exponentially truncated power law, which leads to a fitting index of $R^2 = 0.98$ for R7-Y and $R^2 = 1.0$ for R7-A.

Performing similar analyses, the soft powers for B8-Y and B8-A were determined to be 10 and 8, respectively. For the B7-A dataset, the soft power was 9 and for K9-Y it was 10.

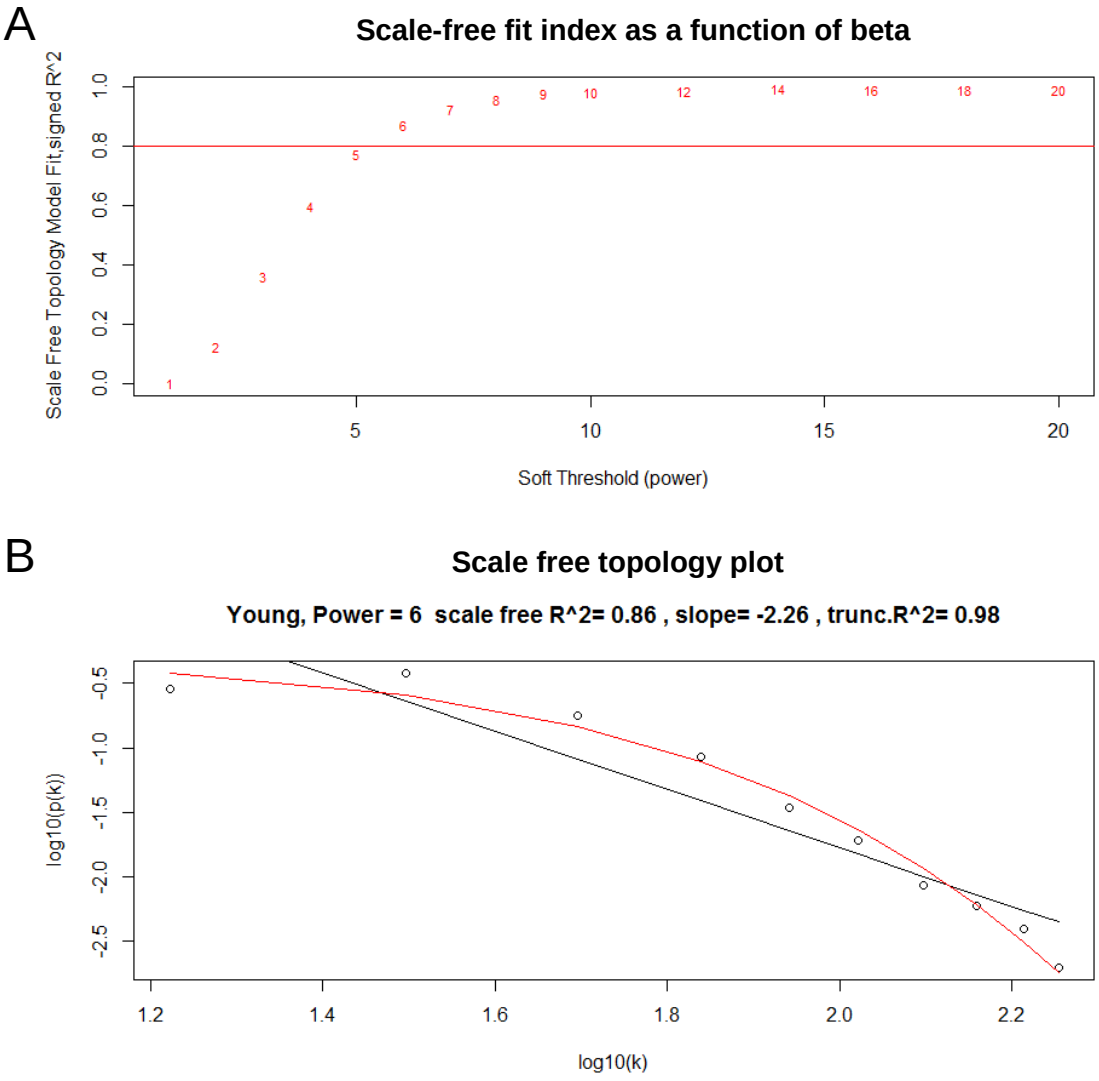


Figure 4.2 Analysis of network topology for various soft-thresholding powers for the R7 young dataset. A) Scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). B) Scale free topology plot shows the log-log plot between frequency of connectivity p(k) and connectivity k for determining whether the network exhibits scale-free topology.

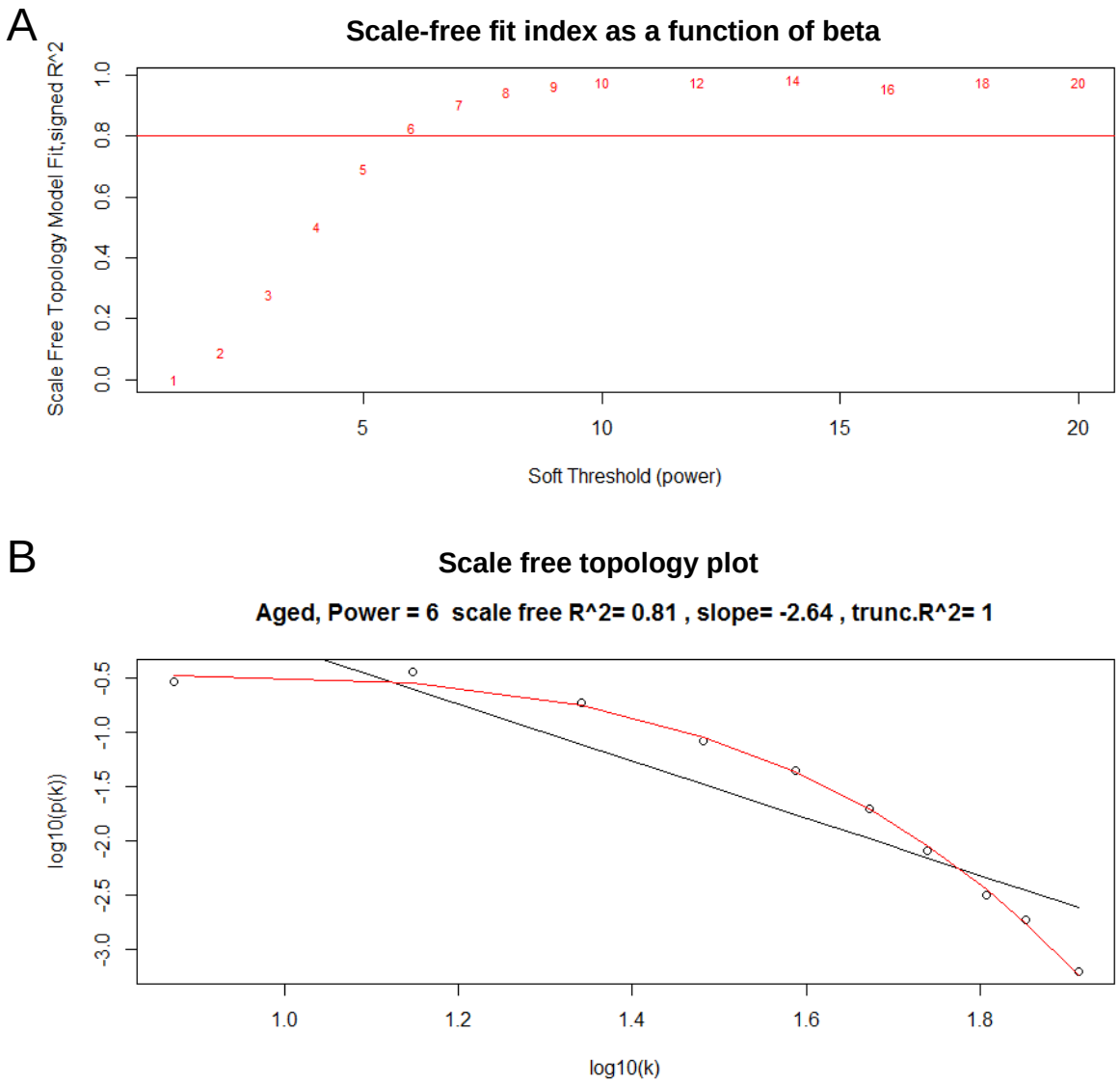


Figure 4.3 Analysis of network topology for various soft-thresholding powers for the R7 aged dataset. A) Scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). B) Scale free topology plot shows the log-log plot between frequency of connectivity p(k) and connectivity k for determining whether the network exhibits scale-free topology.

4.3.3 Creating adjacency (connection strength) matrices

The genes that remained after preprocessing and filtering (Table 2.5 and Table 4.5) were used to calculate the signed Pearson correlation coefficients for all pairwise comparisons of gene-expression values across all young and aged samples. The correlation matrix for each group was then transformed into a matrix of connection strengths (i.e. an "adjacency" matrix) using a soft power beta as determined above. This resulted in a network adjacency matrix for each dataset, for example, for R7 it generated an 8053x8053 matrix. Figure 4.4 shows a portion (6x8) of such data matrices for the R7-Y and R7-A samples.

```
> adj.y.top [1:6, 1:8]
        A1cf    A2m   Aaas   Aacs  Aadat   Aamp   Aars Aarsd1
A1cf  0.0000 0.0114 0.0069 0.0022 0.0094 0.0047 0.0256  0.012
A2m   0.0114 0.0000 0.0430 0.0229 0.0094 0.0125 0.0762  0.093
Aaas  0.0069 0.0430 0.0000 0.0040 0.0174 0.3529 0.0596  0.028
Aacs  0.0022 0.0229 0.0040 0.0000 0.0850 0.0074 0.0616  0.125
Aadat 0.0094 0.0094 0.0174 0.0850 0.0000 0.0037 0.0058  0.045
Aamp  0.0047 0.0125 0.3529 0.0074 0.0037 0.0000 0.0409  0.024
>

> adj.a.top [1:6, 1:8]
        A1cf    A2m  Aaas   Aacs  Aadat   Aamp   Aars Aarsd1
A1cf  0.0000 0.0084 0.014 0.0438 0.0054 0.0104 0.0529 0.0334
A2m   0.0084 0.0000 0.040 0.0241 0.1663 0.0361 0.0323 0.0159
Aaas  0.0144 0.0403 0.000 0.0177 0.1660 0.0927 0.0542 0.0072
Aacs  0.0438 0.0241 0.018 0.0000 0.0154 0.0099 0.0585 0.0761
Aadat 0.0054 0.1663 0.166 0.0154 0.0000 0.0250 0.0083 0.0023
Aamp  0.0104 0.0361 0.093 0.0099 0.0250 0.0000 0.1603 0.0026
>
```

Figure 4.4 A portion (6x8) of the R7 network adjacency data matrix.

4.3.4 Filtering out genes with very low connectivity

First, connectivity value for each gene was calculated from the adjacency matrix. Next for each dataset, the average median connectivity k_{med} was used as a cutoff value to filter out genes with very low connectivity. For R7-Y k_{med} was 0.46 and for R7-A k_{med} was 0.54. I selected the average k_{med} = 0.5 as the minimum connectivity cutoff, which removed 2379 genes, leaving 5674 high connectivity genes for the R7 network analysis (Table 4.5). For B8 and K9 the median k_{med} was 0.4 and 0.35, which resulted in 5202 and 4796 high connectivity genes, respectively. The number of B7 genes was already low

and close to the numbers of other filtered datasets. So, in order to prevent information loss no filtering was done on these B7 genes.

4.3.5 Creating and visualizing a whole network

After connectivity based filtering, all filtered genes were used to determine network topological overlaps and gene co-expression interactions. The topological overlap is a measure of node similarity and for two separate nodes it reflects their relative interconnectedness (i.e. how close the neighbors of gene 1 are to the neighbors of gene 2). Using the 5674 high connectivity genes in R7 a co-expression network was created for visualization (Appendix 6.8.1). Figure 4.5 below shows the co-expression network of 5000 highly connected genes in R7-Y.

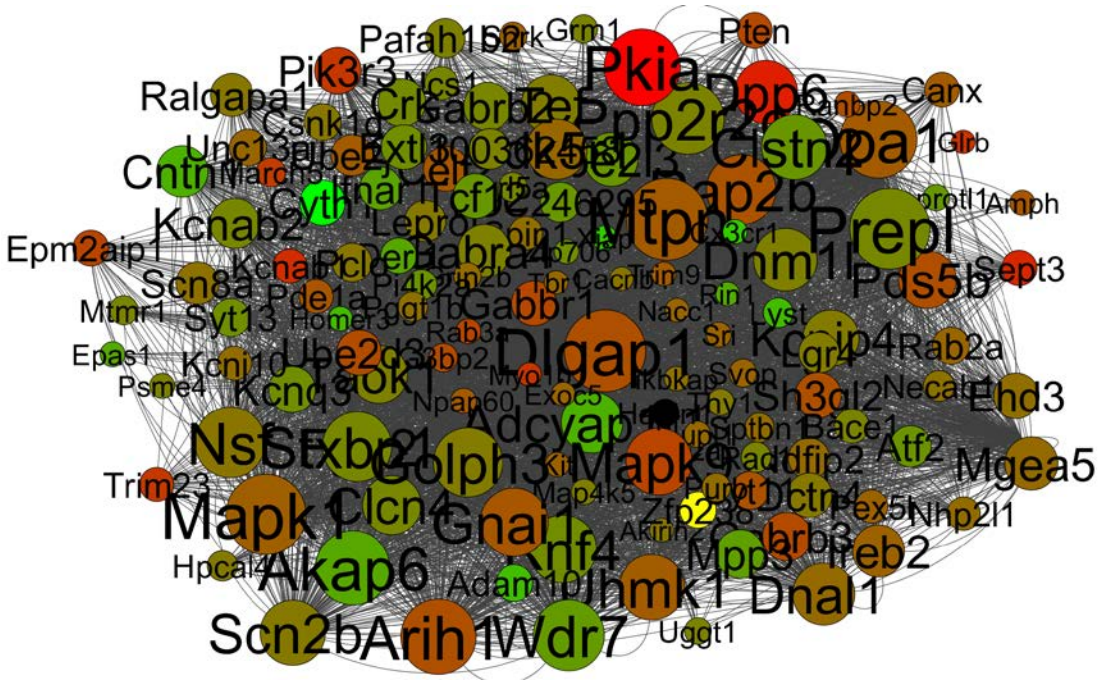


Figure 4.5 A co-expression network using 5000 most highly connected genes from the R7 young dataset. Each node represents a gene whose color represents its differential expression values (log fold change) between the young and aged samples. Red color means the gene shows higher expression in R7-Y compared to R7-A, and green represents the opposite in the young. Node sizes are proportional to the number of co-expression connections with other genes in the network.

4.3.6 Creating and visualizing network modules

A major goal of gene correlation network analysis is to identify groups of highly interconnected genes (Oldham et al., 2006; Zhang and Horvath, 2005) termed as modules. The expression profiles of genes in a module are highly correlated across the samples. In a co-expression network, modules are identified by searching for genes with similar patterns of connection strengths to other genes, or genes with high topological overlap. The topological overlap values are calculated using the adjacency and connectivity values, which determine which genes will be in which module and form a network. The values range between 1 and 0 representing maximum and minimum interconnectedness. The module identification method in WGCNA is based on using a node dissimilarity measure in conjunction with a clustering method. Since the topological overlap matrix is non-negative and symmetric, it is turned into a dissimilarity measure by subtracting from one. Genes are hierarchically clustered using the average linkage method, taking 1-topological overlap as the distance measure and modules are determined by choosing a height cutoff for the resulting dendrogram. In the dendrogram, discrete branches of the tree correspond to modules of co-expressed genes. Following these steps, gene network modules for the young and aged samples were identified separately for each dataset using the filtered weighted correlation matrices as prepared above.

Figure 4.6 and Figure 4.7 show the hierarchical dendrograms of topological overlaps for the 5674 genes in R7-Y and R7-A, respectively. There are several height cut-off algorithms implemented in the WGCNA R package. In this research the cut-tree hybrid method was chosen to pick a height cut-off and to identify modules, which are shown in the panel below the dendrograms. The default lowest cut-off resulted in six modules in the young network and 15 modules in the aged network. Each module is labeled with a unique color (except grey) for easy visualization and understanding. The color grey is preserved for genes that do not belong to any module.

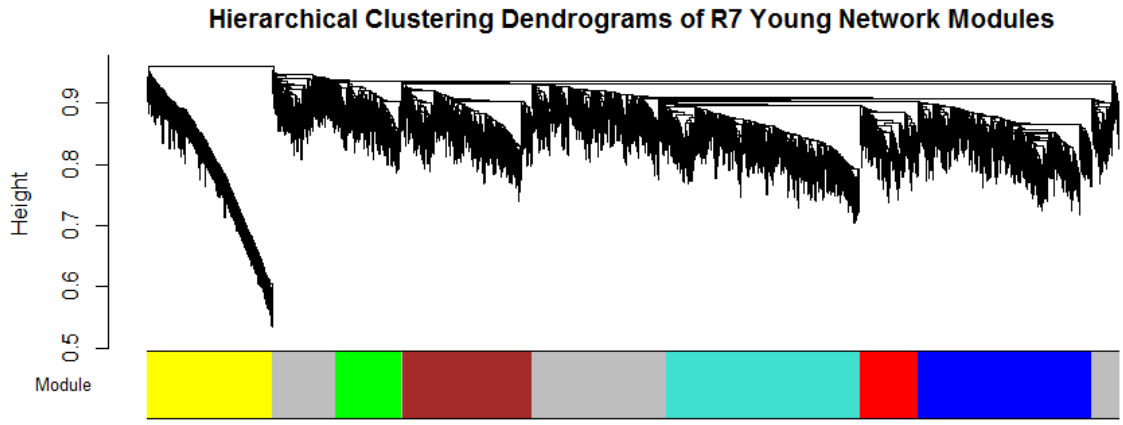


Figure 4.6 Hierarchical clustering dendrogram of topological overlaps of R7-Y genes. The cut-tree hybrid method was used to pick a height cut-off and to identify modules, which are shown in the panel below the dendrogram. Each module is labeled with a unique color for easy visualization and understanding.

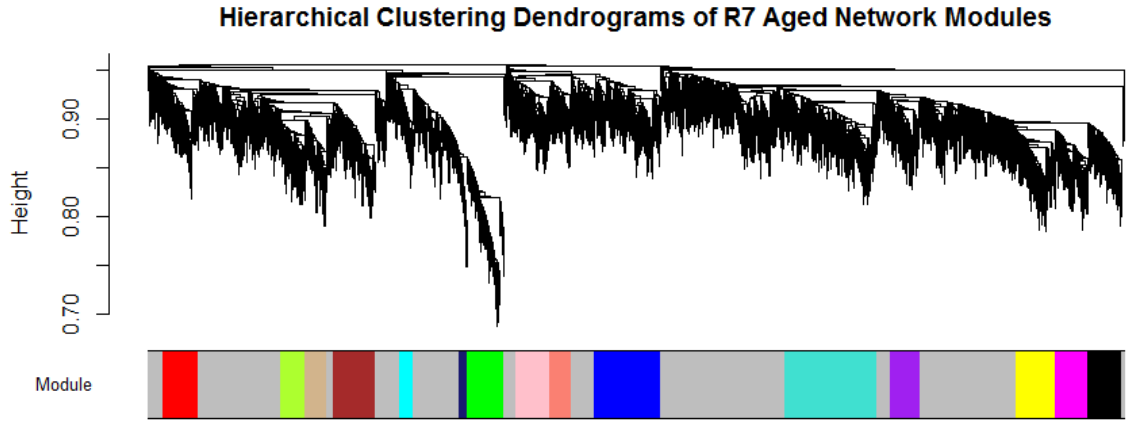


Figure 4.7 Hierarchical clustering dendrogram of topological overlaps of R7-A genes. The cut-tree hybrid method was used to pick a height cut-off and to identify modules, which are shown in the panel below the dendrogram. Each module is labeled with a unique color for easy visualization and understanding.

The aged network resulted in many modules, most with small numbers of genes, for example, 13 of the modules had fewer than 300 genes each and 9 of them had less than 200 genes each (result not shown). For better comparison, the number of modules in the aged network was brought closer to that of the young network. This was accomplished by merging the modules using the WGCNA function mergeCloseModules(...). Figure 4.8 presents an average linkage hierarchical plot of the module eigengenes of the 16 aged modules (including the grey module). It shows that some modules (e.g. the black and turquoise, cyan and green, blue and salmon, etc.) are clustered very close together. The three red lines represent the tree cut line at different heights and the numbers on the right (corresponding to the lines) represent the expected number of resulting merged modules that each cut will produce. In order to keep the module numbers similar to that of the young network, a cut height of 0.4 was chosen that generated seven modules in the aged network (including the grey module) (Figure 4.9).

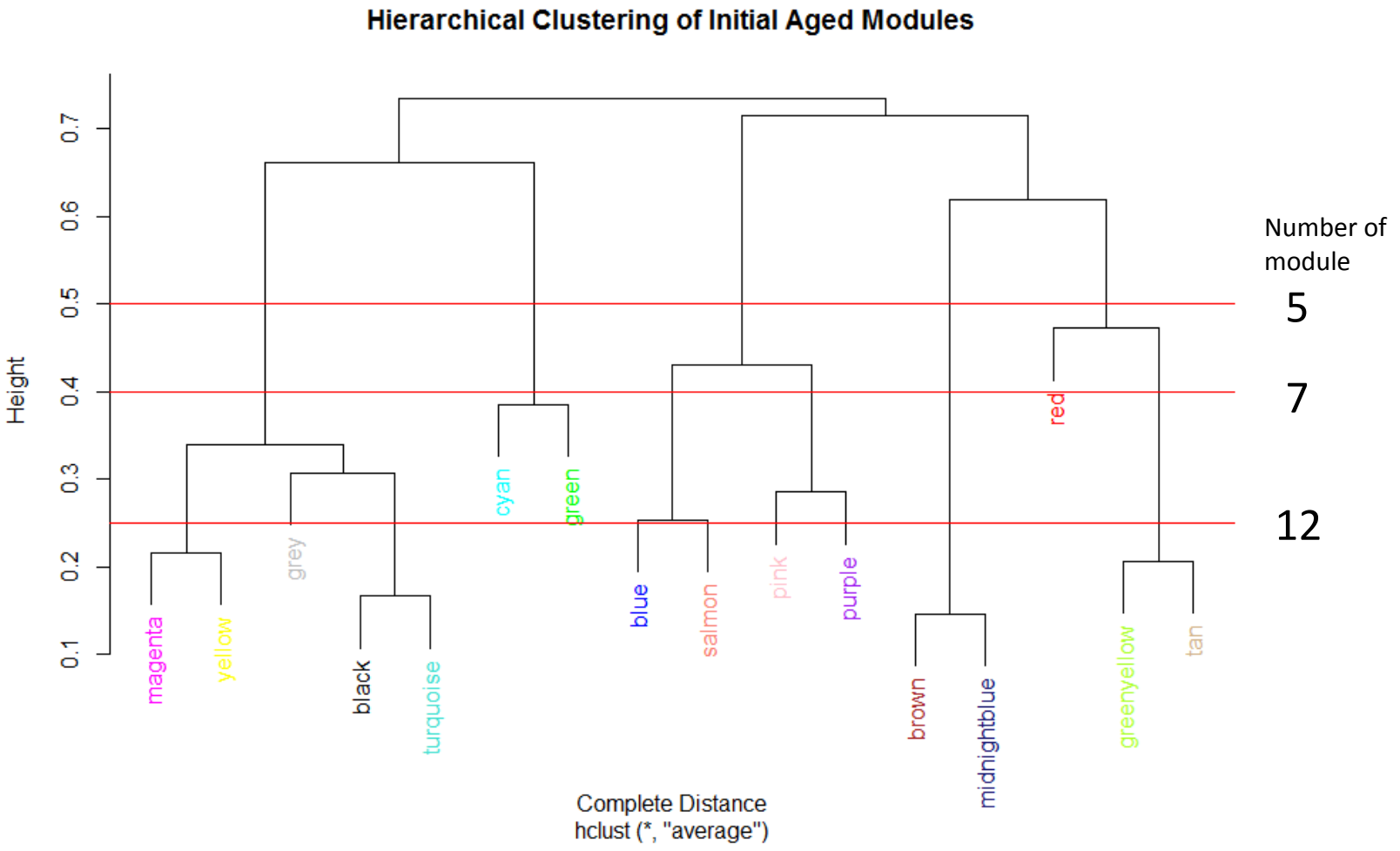


Figure 4.8 Hierarchical clustering of the initial 16 aged modules.

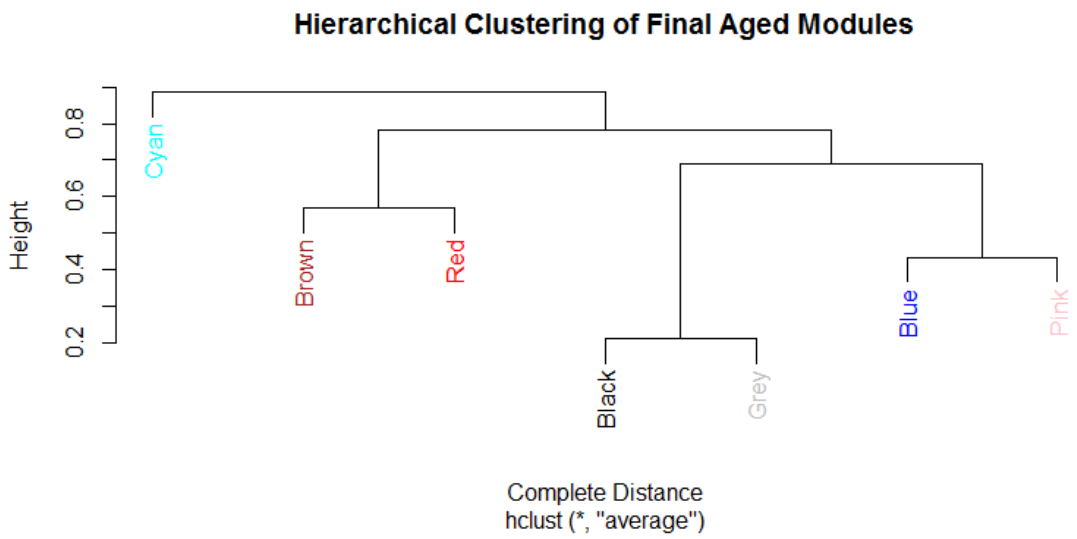


Figure 4.9 Hierarchical clustering of the final aged modules.

Figure 4.9 shows that the old modules (Figure 4.8) are merged into new modules, for example, the magenta, yellow, black, and turquoise modules are merged into a new module named black; cyan and green are merged into cyan; blue and salmon are merged into blue; pink and purple are merged into pink; brown and midnightblue are merged into brown; and red, greenyellow, and tan modules are merged into red. The grey module contained genes that did not belong to any module and remained separate. Since module names/labels in a network were randomly generated, the seven aged modules were matched to the seven young modules to check for similarity and module overlap of gene members (Appendix 6.9.1 and Appendix 6.9.2; see Section 4.3.8.2 for details). Once a significant match was found, modules in the aged network were renamed after the matched young network module names. The Table 4.3 shows the final modules in the young and aged networks along with the number of genes belonging to each module. In addition, Table 4.3 shows which aged modules are matched to which young modules. The black module from the aged network had genes matching significantly to both the blue and brown modules in the young network. The aged brown, red, and cyan modules matched to the green, red, and yellow young modules, respectively, while the blue and pink aged modules matched a single turquoise

young module. This module matching process is helpful when comparing similar modules between networks, for example, aged vs. young.

Table 4.3 Modules in the R7 young and aged networks. There were seven modules in each group including the grey module. Aged modules were matched to the young modules to find modules containing the maximum number of matching genes. Once identified, the aged module names were changed to match the respective young module names for easy comparison.

Samples	Module names						
Young	Blue	Brown	Green	Grey	Red	Turquoise	Yellow
# of genes	1015	759	380	1319	341	1129	731
Aged (original labels**)	Blue (Black)	Brown (Black)	Green (Brown)	Grey (Grey)	Red (Red)	Turquoise (Blue & Pink)	Yellow (Cyan)
# of genes	1151	1151	554	2600	206	508 & 366	289

** (original labels/names of the aged modules before matching to the young modules are in bracket)

The co-expression network interaction files for the genes belonging to each of the R7 young modules were created using the method described in Appendix 6.8.1. For clarity, only the top 500 to 600 most connected genes and their co-expression interactions were used to create each module network. Co-expression information from all modules were combined and imported into the Cytoscape for visualization. Figure 4.10 shows all six modules in the R7 young networks where the modules are represented by the color of their respective names (e.g. the blue module is represented by the color blue).

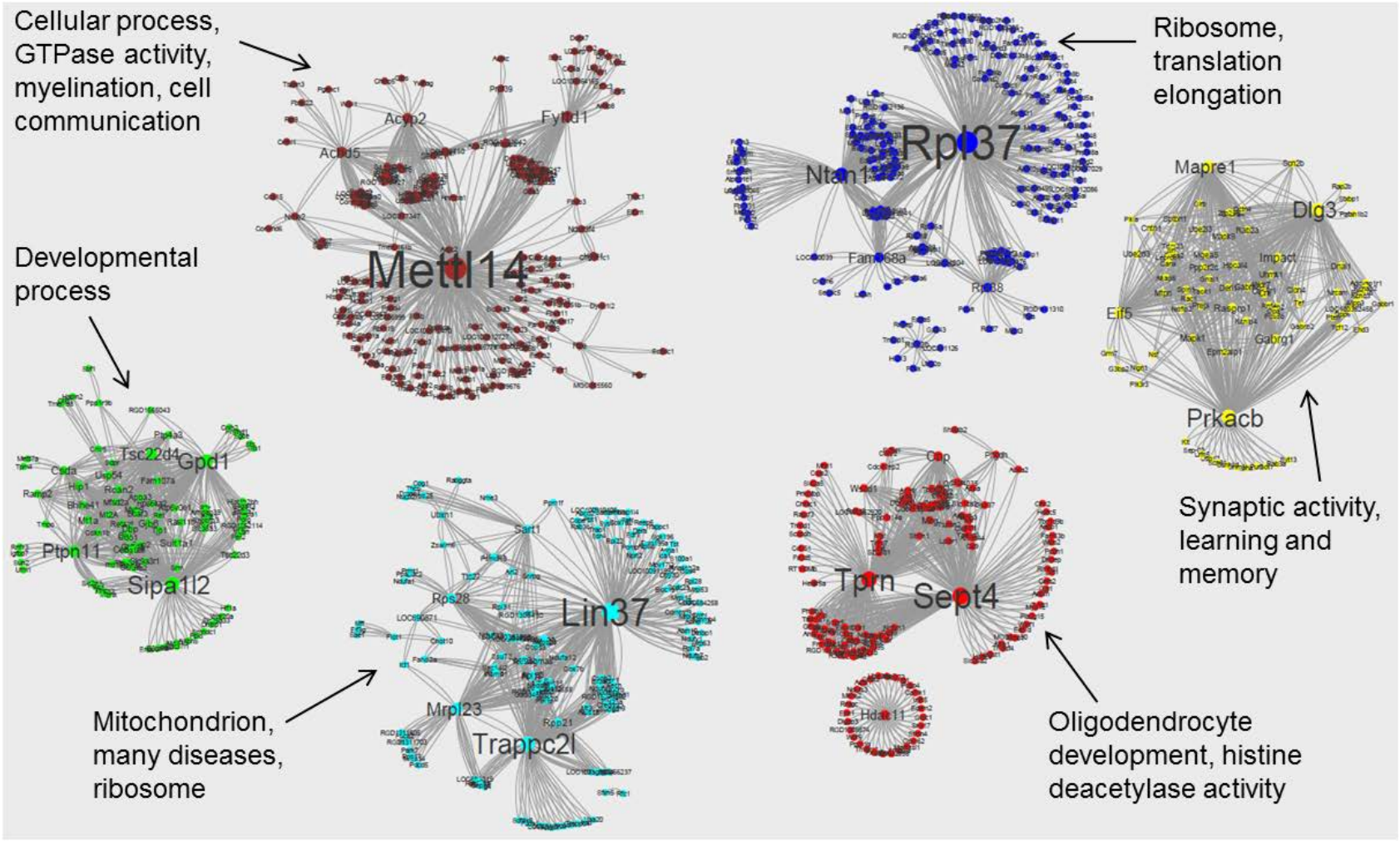

Cellular process, GTPase activity, myelination, cell communication
Mettl14
Acyp2
Fyttd1
Acbd5
Ribosome, translation elongation
Rpl37
Ntan1
Fam98a
Rps8
Developmental process
Tsc22d4
Gpd1
Ptpn11
Sipa1l2
Lin37
Sart1
Rps28
Mrpl23
Rpp21
Trappc2l
Mitochondrion, many diseases, ribosome
Tprn
Sept4
Cnp
Hdac11
Oligodendrocyte development, histine deacetylase activity
Mapre1
Dlg3
Impact
Eif5
Gabrg1
Prkacb
Synaptic activity, learning and memory

Figure 4.10 All six modules in the R7 young networks. The modules are represented by the color of their respective names, for example, the blue module is represented by the color blue. The most significant GO functional categories represented by the genes belonging to each module are also shown.

4.3.7 Exploring the functional significance of modules

Biological significance analysis of the network modules was performed using the functional annotation clustering analysis in DAVID that utilizes the GO and other biological pathway information databases. DAVID is a web-based high-throughput functional annotation bioinformatics resource. It provides a comprehensive set of functional annotation tools to understand biological meaning behind large lists of genes. For any given gene list, DAVID tools are able to identify enriched biological themes, particularly GO terms and discover enriched functionally-related gene groups.

DAVID functional annotation clustering analysis was used through the RDAVIDWebService tool in R. DAVID also allows one to identify the most relevant (overrepresented) biological terms associated with a given gene list. The DAVID database offers extended annotation coverage with over 40 annotation categories, including GO terms, protein-protein interactions, protein functional domains, disease associations, bio-pathways, sequence features, homology, and many more (Huang da et al., 2009b). However, for reasons of simplicity and to better understand the biological significance of the network modules identified above, only the biological processes (BP), molecular functions (MF), and cellular components (CC) GO terms and all KEGG Pathway terms were included in the functional annotation clustering analysis.

Affymetrix probe set identifiers of all the genes belonging to a network module (Table 4.3) were used as the input gene list. The total number of genes from the RAE230A array for the R7 dataset (after preprocessing and filtering) was 5674, and was used as a background population. *Rattus norvegicus* was used as species.

The function getClusterReportFile(. . .) in RDAVIDWebService was used with default parameters to retrieve all relevant information (Appendix 6.6.1). Next getClusterReport(…) function was used to extract the functional annotation chart file, which was saved as a text file and later analyzed. An enrichment score cutoff of 1.0 was used to minimize the number of clusters that were returned. Table 4.4 shows the

summary result of GO analysis for the young modules. The most significant GO functional categories represented by the genes belonging to each module are also shown in Figure 4.10. A module-wise detail report can be found in Appendix 6.7.1 to Appendix 6.7.6.

The results show that, in general, each module is highly enriched with genes functioning in broad but distinct GO functional categories or biological pathways with highly significant enrichment scores. A brief description of the results for each module is given in the following sections.

Table 4.4 GO functional analysis summary for the R7 young modules.

Module	Major GO Categories	p-value
Blue	ribosome, translation elongation	9.85E-08 to 2.02E-09
Brown	cellular process, GTPase activity, myelination, cell communication	0.02 to 0.006
Green	developmental process	9.36E-04
Red	oligodendrocyte development, histine deacetylase activity	0.01 to 0.005
Turquoise	mitochondrion, many diseases, ribosome	1.20E-04 to 3.12E-06
Yellow	synaptic activity, synaptic transmission , learning and memory	2.94E-04 to 4.77E-15

4.3.7.1 Blue Module

The blue module contained a total of 1015 genes, of which 999 corresponding Affymetrix IDs were found in the DAVID database. Results show that this module is highly enriched with genes functioning in two distinct functional categories with very high enrichment scores (Appendix 6.7.1). The first cluster of genes is localized in the ribosome and contributes to the translational pathway (enrichment score 5.08, p-values 0.001 to 1.68E-06 after Benjamini multiple testing correction). The second cluster of genes is localized in the mitochondria and contributes to cellular metabolic and biosynthetic process pathways (enrichment score 2.40, p-value 0.05 to 7.59e-06 after multiple testing correction).

4.3.7.2 Brown Module

This module contained 759 genes, of which 744 were found in the database. This module is highly enriched with clusters of genes that are part of the intracellular organelles or macromolecular complex and may function in a variety of cellular processes including binding, transport, hydrolase or GTPase/ATPase activity, myelination, and cellular homeostasis (enrichment score 1.1 to 1.6 and p-values ranging from 0.05 to 0.001) (Appendix 6.7.2).

4.3.7.3 Green Module

In this module, 375 Affymetrix IDs corresponding to 380 genes were identified in the DAVID database. Genes in this module are significant with two main GO term clusters (enrichment score 0.99 to 1.56 and p-values ranging from 0.05 to 5.91e-04) (Appendix 6.7.3). The first cluster is enriched with genes in developmental process pathways and the second cluster is enriched with genes contributing to GTPase regulator activity functions. Some of these genes are known to reside in the extracellular space.

4.3.7.4 Red Module

The red module contained a total of 341 genes, 335 of which were mapped in the DAVID database. The red module is enriched with three main clusters (enrichment score 1.0 to 1.87 and p-values ranging from 0.03 to 0.005) (Appendix 6.7.4). The first cluster of genes functions in the histone deacetylase pathway and the second cluster of genes contribute to glial cell differentiation or oligodendrocyte differentiation functions. The third cluster of genes (a majority of which reside in the cell membrane or integral to membrane) respond to the hormone stimulus pathway.

4.3.7.5 Turquoise Module

This module contained 1129 genes. Affymetrix IDs for 1118 of these genes were found in the database. There are four highly significant gene clusters that belong to the turquoise module (enrichment score 1.63 to 1.93, p-values 0.02 to 0.001 after Benjamini multiple testing correction) (Appendix 6.7.5). DAVID analysis shows that a large number

of genes in this module are located in or associated with, ribosome or mitochondrion. In fact, many of these genes are mitochondrial ribosome genes, for example, about 19 of the 45 ”ribosome” genes are also part of the 175 “mitochondrion” genes. A number of these genes show repeated appearance in different clusters that are highly enriched with Huntington’s disease, Parkinson’s disease, Alzheimer’s disease, and oxidative phosphorylation KEGG pathway terms.

4.3.7.6 Yellow Module

Affymetrix probe set IDs corresponding to 723 of the 731 yellow module genes were mapped in the DAVID database. Functional clustering analysis using the mapped genes shows a very large number of significant hits even after Benjamini multiple testing corrections, and with very high enrichment scores (Appendix 6.7.6).

Cluster one has the highest enrichment score of 5.92 with a p-value range of 0.03 to 1.37e-5 (after Benjamini multiple testing correction). Genes in this cluster are enriched in the regulation of cellular localization, secretion, transport, cell communication, and synaptic transmission GO biological process terms.

Cluster two contains a large number of genes that are part of the plasma membrane or are integral to the membrane (enrichment score of 5.91 with a p-value range of 0.03 to 1.55e-10 after Benjamini multiple testing correction). A closer look at the genes from cluster one (e.g. 37 genes in regulation of cellular localization or 53 genes of regulation of transport) and cluster two genes (e.g. 182 plasma membrane and 156 transport genes) show that many of these genes are localized in the plasma membrane or integral to membrane and contribute to cellular localization and transport.

Cluster three contains genes that are enriched with GO molecular functions such as transmembrane receptor activity and molecular signal transducer activity (enrichment score of 5.07 with a p-value range of 0.05 to 3.40e-05 after Benjamini multiple testing correction).

Cluster four shows enrichment of genes in two GO categories, cellular components and biological processes (enrichment score of 4.85 with a p-value range of 0.02 to 2.10e-12 after Benjamini multiple testing corrections). The GO cellular components results indicate that some genes are part of the synapse, plasma membrane or postsynaptic membrane. The GO biological processes enrichment indicates that these genes contribute to cell-cell signaling and synaptic transmission.

Cluster five echoes the result of cluster four GO cellular component enrichment. It shows that the genes are significantly enriched with GO cellular component terms, which indicates that these genes are part of synapse, axon, neuron projection, or postsynaptic density (enrichment score of 4.74 with a p-value range of 0.02 to 2.10e-12 after Benjamini multiple testing correction).

Cluster six is very interesting as it shows enrichment of genes mostly in GO molecular functions such as gated or voltage-gated ion channel activity and ion binding and transport (enrichment score of 3.16 with a p-value range of 0.02 to 4.88e-05 after Benjamini multiple testing correction).

Cluster seven is enriched with genes contributing to various neurological system processes such as cognition, behavior, learning, and memory (enrichment score of 3.05 with a p-value range of 0.03 to 3.97e-05). However, after Benjamini multiple testing correction, only the neurological system processes GO biological process term was found significant (p-value = 0.011).

All these clusters have many genes in common in closely related cellular component, function or pathway categories (based on my manual comparison).

4.3.8 Validating network modules

The gene expression data were compared as follows: R7 young vs. B8 and K9 young; R7 aged vs. B8 and B7 aged (Figure 4.11).

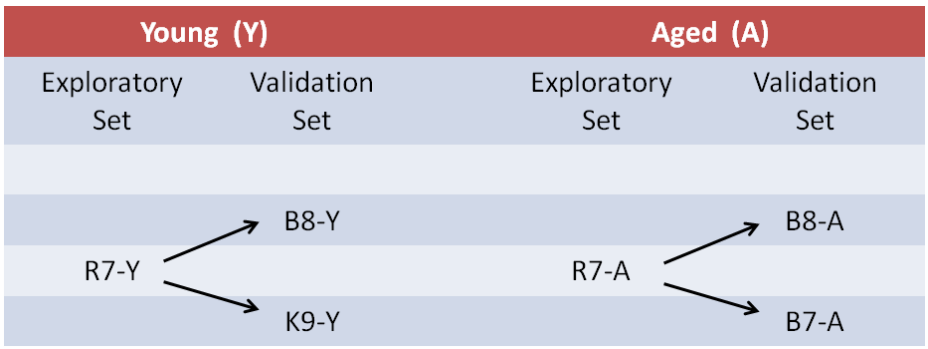


Figure 4.11 Network validation analyses strategies across multiple independent datasets. This figure shows how the networks of the young and aged samples were compared among independent datasets. The gene expression data were compared as follows: R7 young vs. aged; R7 young vs. B8 and K9 young; R7 aged vs. B8 and B7 aged. However, in each comparison the R7 young network module definition was used as a reference and networks were created from gene expression data accordingly for comparison.

4.3.8.1 Module preservation

Module preservation was assessed quantitatively where the R7 5674 top most connected genes from the young networks were compared to the same genes in other datasets to see how well the module assignment of these R7 genes and their module-wise functions are preserved in other datasets. However, in each comparison the R7-Y network module definition was used as a reference. For example, in the comparison between R7-Y vs. B8-Y, the same R7 top most connected 5674 genes were selected from B8-Y. Next, the same R7-Y gene module definition was mapped to the B8-Y genes. There was an exception for the R7-A vs. B7-A comparison where only 2140 genes were used because only these genes were common between the two different chip types used in the two independent studies.

Network preservations were estimated using the modulePreservation(…) function built into the WGCNA library by keeping the maximum module size at 700 and using 30 permutations. The results are summarized in the bar plot in Figure 4.12. It presents the

preservation of R7 young and aged modules in each comparison as $Z_{summary}$ statistics along the x-axis. All the R7 young modules (e.g. brown, yellow, turquoise, blue, green, and red) along with their major significant functional categories are represented in the y-axis. Except the green module, all other modules generally show moderate to high preservation across independent studies. The brown module shows the highest preservation among all the modules while the green module shows the lowest preservation. All modules in general in the R7 aged vs B7 aged comparison shows comparatively lower preservation than in the other comparisons.

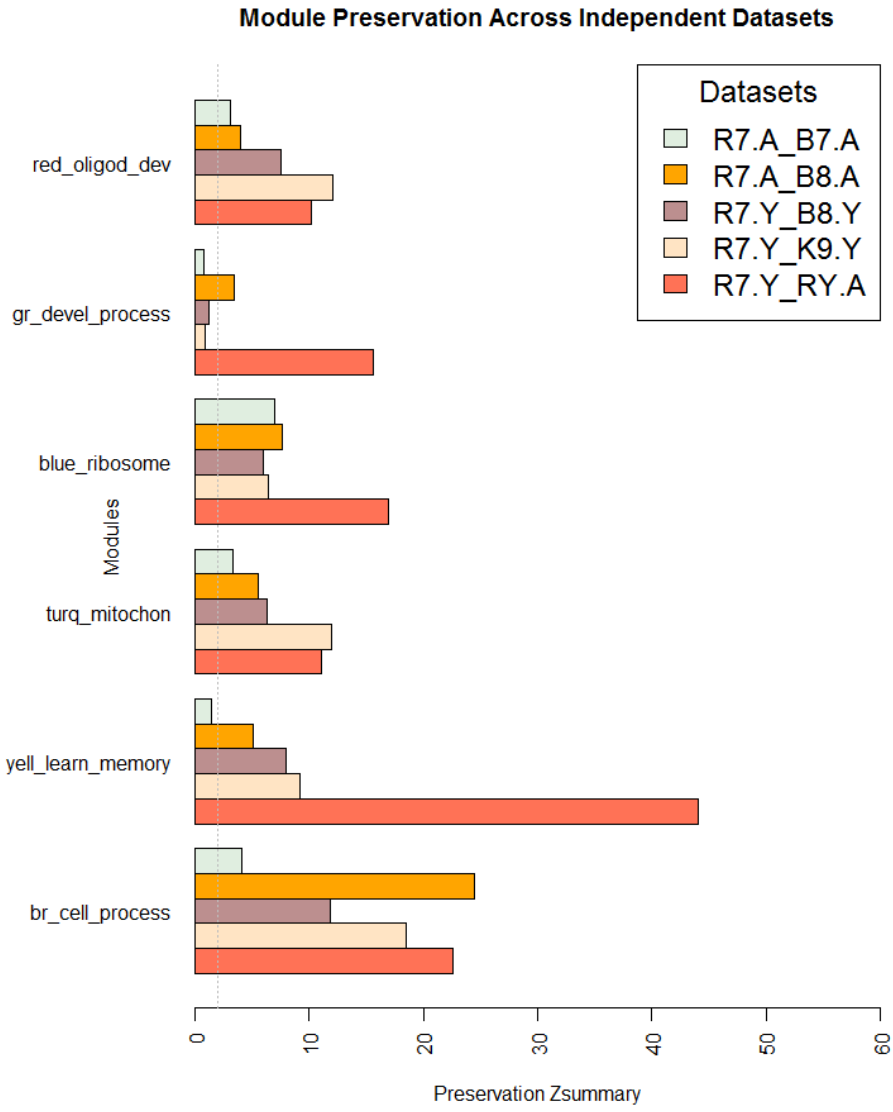


Figure 4.12 Preservation of R7 young network modules across studies, age, and platform. The x-axis presents the preservation $Z_{summary}$ statistics and the y-axis represents the R7-Y modules such as brown, yellow, turquoise, blue, green, and red along with their major significant functional categories. In each comparison R7 module assignment was used as a reference. The preservation of modules in R7-Y vs. R7-A is shown as a guide. The vertical dotted line at $Z_{summary}$ score 2.0 indicates the borderline between no preservation and very weak preservation. Generally, $5<Z<10$ indicates moderate preservation and $Z>10$ indicates high preservation. Legends: gr, green; turq, turquoise; yell, yellow; br, brown.

4.3.8.2 Module overlap between networks

Comparing networks by calculating module overlaps between networks provides another way to validate network modules using independent datasets. I performed a pair-wise comparison for all datasets as explained in Section 4.2.9.2 and in Figure 4.11. After merging datasets by matching genes, there were 3626 top most connectivity genes common between R7 and B8, 3138 between R7 and K9, and 2140 between R7 and B7 networks (Table 4.5).

Table 4.5 Gene selection for network comparison. This table shows the number of genes that remained for network analysis and comparison after low-connectivity gene filtering and after matching the R7 young module labels to the B8, K9, and B7 data. For B7, no low-connectivity filtering was done because of the already low number of remaining genes.

	Number of genes remained for network analysis				
Dataset	After preprocessing	Median Connectivity (kmed)	After low-connectivity gene filtering	Module overlap comparison	After matching R7 young genes
R7	8053	0.5	5674		--
B8	7157	0.4	5202	R7-Y_B8-Y R7-A_B8-A	3626
K9	8250	0.35	4796	R7-Y_K9-Y	3138
B7	4829	- - -	4829	R7-A_B7-A	2140

Once two datasets had the same matching genes selected, next, for each comparison (e.g. between R7-Y and B8-Y) all modules were compared between the two datasets (i.e. the module assignment of the genes in R7 were matched to the same genes in B8). For each comparison, the results generated an overlap table and a p-value table showing the number of genes that matched between each pair of modules and their associated p-value significance, respectively (Appendix 6.9.3 to Appendix 6.9.6). From these results, percentage overlap for each module was calculated by dividing the total genes matched

to a module (e.g. number of genes from an R7 module matching to the genes from a module in the second dataset) with the total matched to all modules (e.g. number of genes from an R7 module matching to the genes in all modules (max shared) in the second dataset). In cases where an R7 module was matched to multiple modules in the second network, overlap with the lowest p-value was considered. For example, the R7-Y yellow module genes (731) matched to only 85 genes in the B8-Y red module with the lowest p-value (highest match), while they matched to 385 genes in the B8 young network shared by all the modules. Therefore, the percentage overlap is 85/385 = 22.08% with a p-value of 8.50e-09. The final results for all four comparisons (column five in Table 4.5) are summarized in the bar plots in Figure 4.13 for young and in Figure 4.14 for aged networks.

For the young, all modules in R7-Y were compared for their significant overlap in B8-Y and K9-Y (Figure 4.13). The results show that except the blue module in the R7-Y vs. B8-Y comparison, all modules show a significant repeatability with a p-value < 0.05. The red module showed the maximum overlap trailed by brown, turquoise, yellow, green, and blue.

For the aged, all modules in R7-A (using the R7 young module definition) were compared for their significant overlap in B8-A and B7-A (Figure 4.14). The results show that all modules demonstrate a significant repeatability with a p-value < 0.05 across independent datasets. The blue module showed the maximum overlap trailed by turquoise, brown, yellow, red, and green.

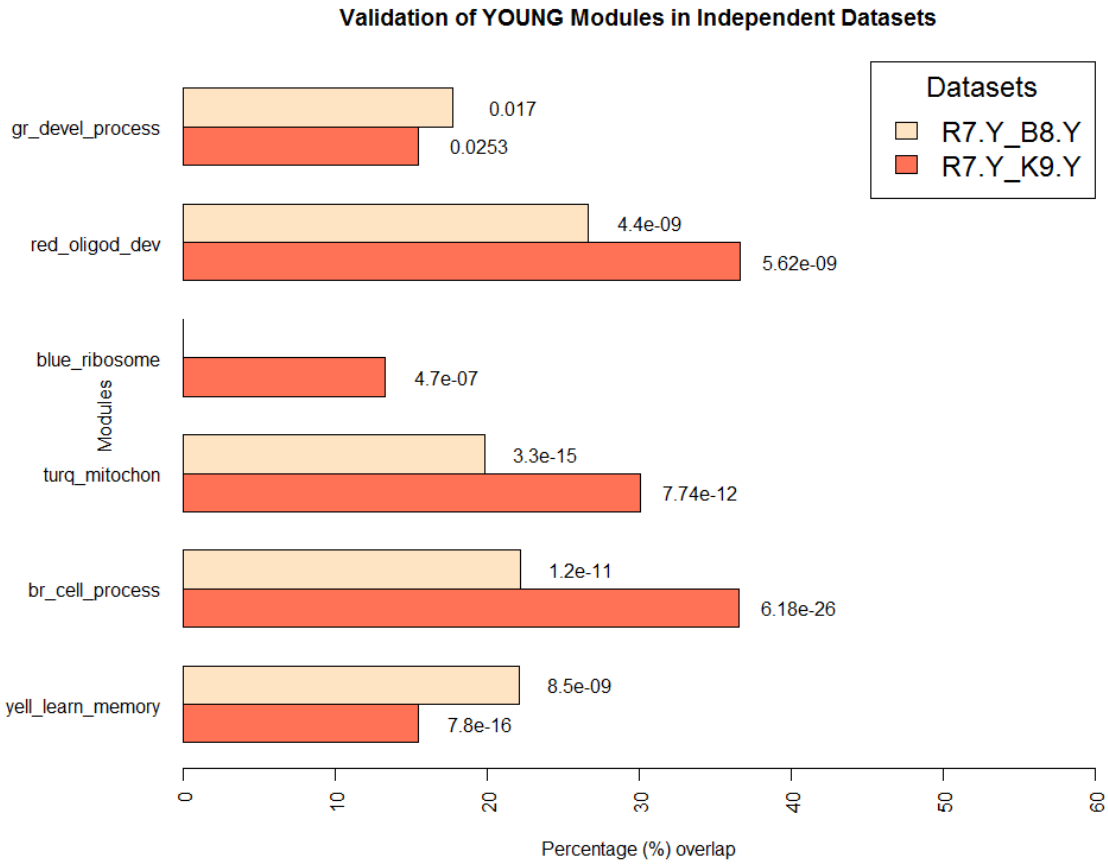


Figure 4.13 Validation of young modules in independent datasets. All modules in R7-Y were compared for their significant overlaps in B8-Y and K9-Y. The percentage overlap is shown on the x-axis and the modules, along with their broad significant GO categories, are shown on the y-axis. Legends: gr, green; turq, turquoise; yell, yellow; br, brown.

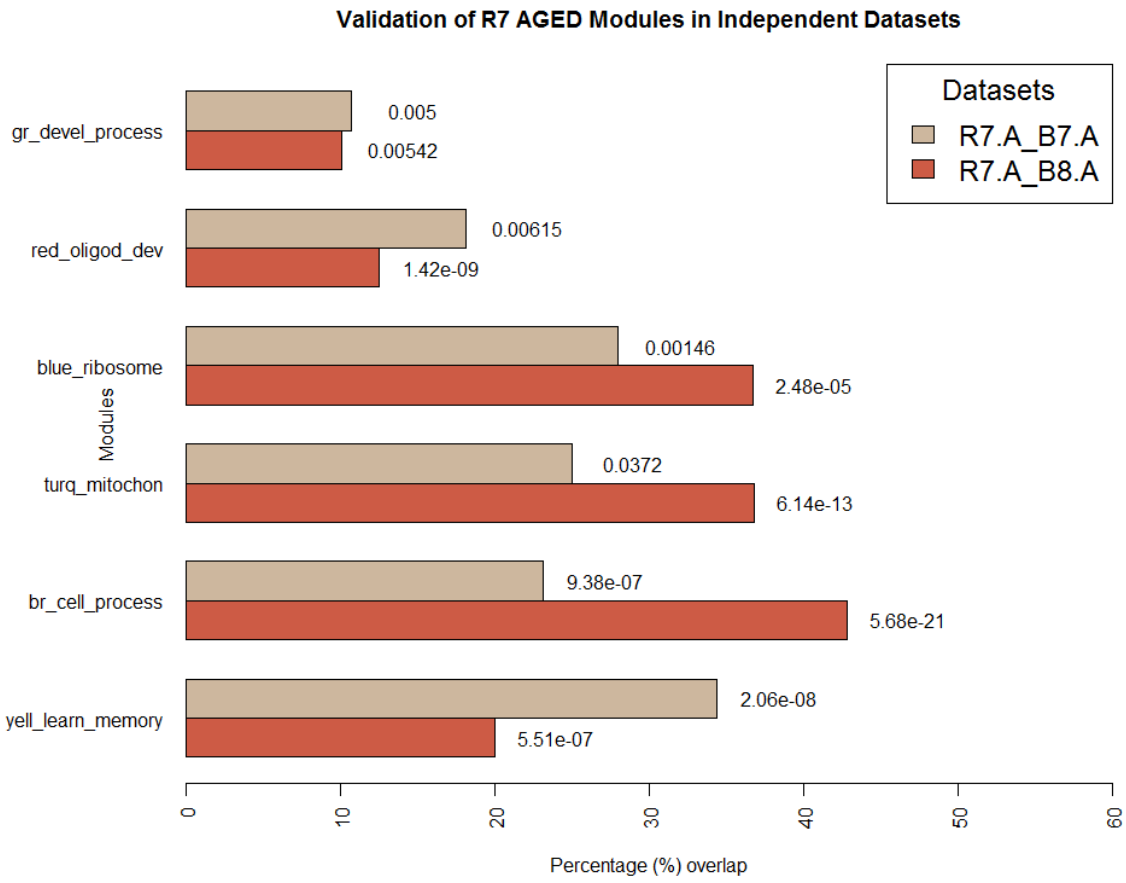


Figure 4.14 Validation of aged modules in independent datasets. All modules in R7-A were compared for their significant overlaps in B8-A and B7-A. The percentage overlap is shown on the x-axis and the modules, along with their broad significant GO categories, are shown on the y-axis. Legends: gr, green; turq, turquoise; yell, yellow; br, brown.

4.3.9 Differential network analysis of young vs. aged

In order to assess the changes in co-expression patterns of the young as they age and how the aging would affect learning impairments, I compared several interesting network modules between young and aged networks generated from the R7 data. This comparative investigation involved visualizing them side by side, comparing expression patterns between networks, and searching for key genes. In addition, it involved identifying the key genes' functions and pathways that can help explain the learning differences as well as the aging effect that had been observed between the young and aged animals. Differential expression levels for the top 5674 genes in the R7 data were calculated by using the *limma* package in Bioconductor. The log fold changes of expression differences between young and aged for all genes were saved as a tab delimited text file, and later loaded as node attributes in Cytoscape for each module.

Figure 4.15 presents the differential co-expression networks of the yellow module between young and aged rats, which demonstrates a clear difference in expression patterns between the young and the aged genes. The majority of the genes in the aged yellow network show lower expression compared to the young. In addition, the comparative analysis demonstrates differential co-expression for many genes between the two networks (i.e. some genes display more co-expression interaction than others and this varies between the young and the aged networks). The results allow one to identify a number of key genes for further investigation (see Section 4.3.10).

Differential co-expression networks for the brown ("cellular processes, GTPase activity"), green ("developmental process"), and red ("oligodendrocyte development, histine deacetylase activity") modules are presented in Figure 4.16 to Figure 4.18. Like the yellow module, the majority of the genes in the aged brown network show lower expression compared to the young. However, this type of expression differences is not so dramatic, rather mixed, in the green and red modules. Like the yellow, all these modules display differential co-expression.

R7 Yellow Module

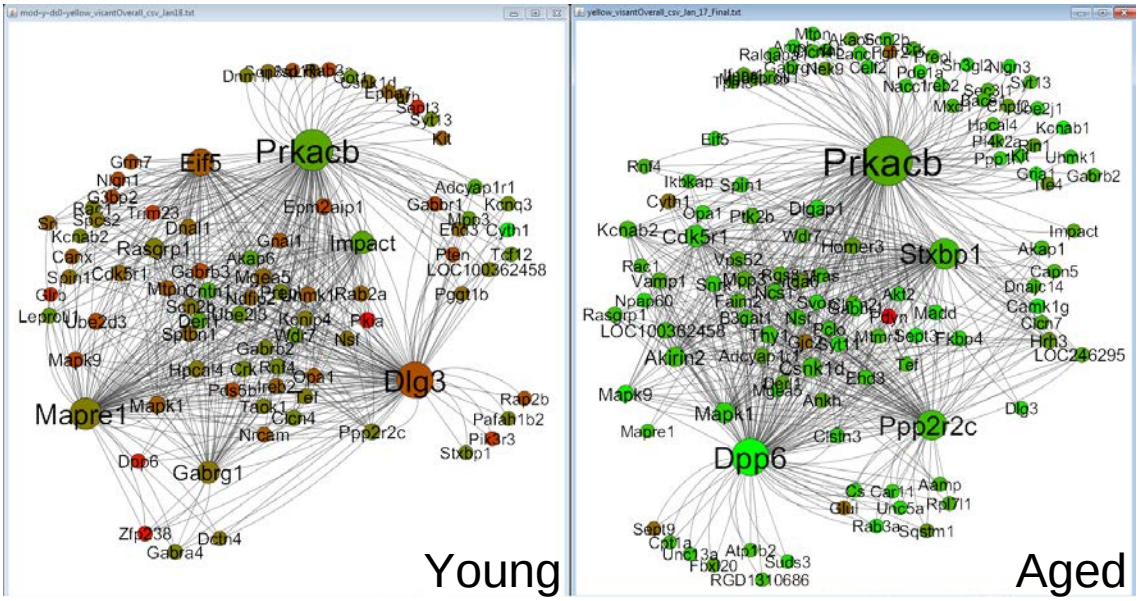


Figure 4.15 Differential co-expression network analysis for the yellow "learning" module in the young and aged in R7. The color of each node displays differential expression level (log fold change value) between young and aged samples. Each node size is proportional to the number of co-expression interaction the node has. Legends: red is upregulation; green is downregulation.

R7 Brown Module

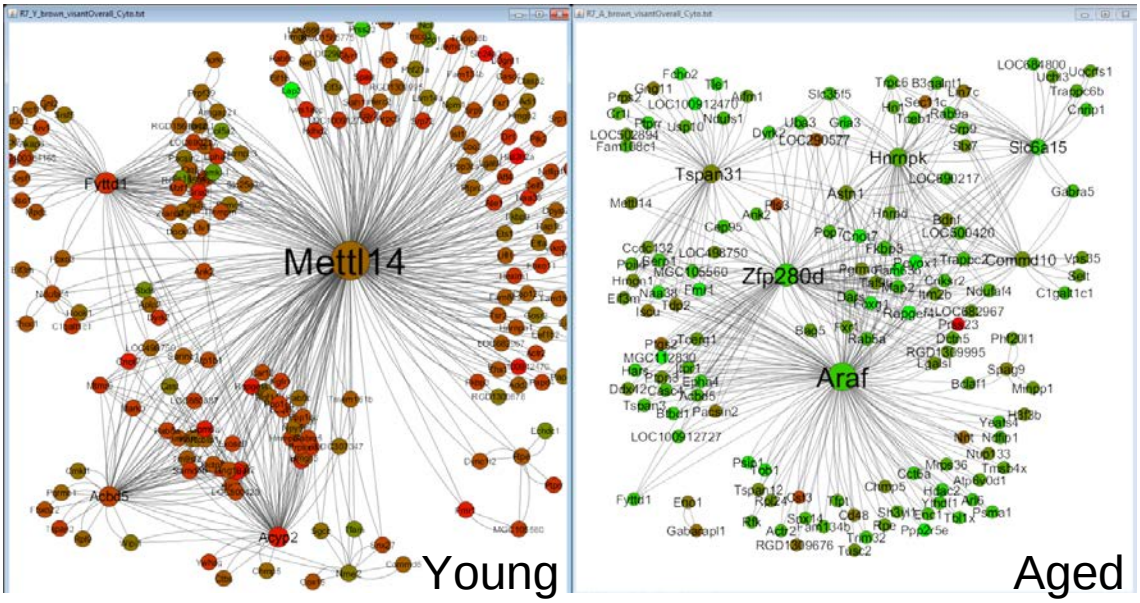


Figure 4.16 Differential co-expression network analysis for the brown module in the young and aged in R7. The color of each node displays differential expression level (log fold change value) between young and aged. Each node size is proportional to the number of co-expression interaction the node has. Legends: red is upregulation; green is downregulation.

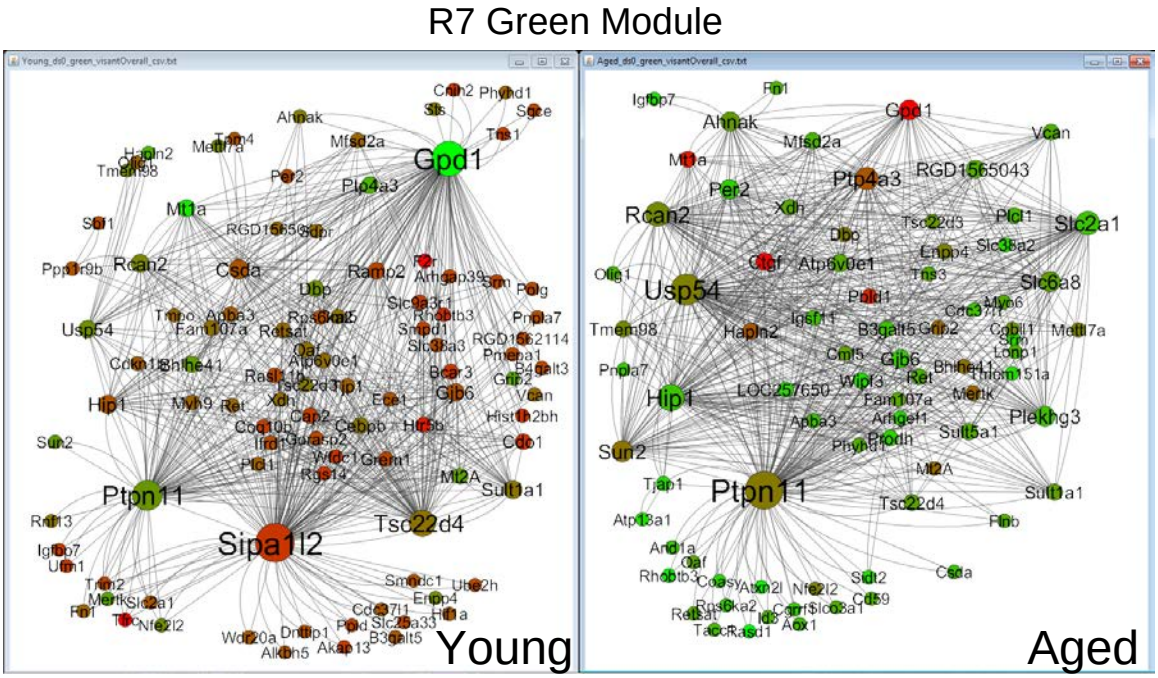


Figure 4.17 Differential co-expression network analysis for the green module in the young and aged in R7. The color of each node displays differential expression level (log fold change value) between young and aged. Each node size is proportional to the number of co-expression interaction the node has. Legends: red is upregulation; green is downregulation.

R7 Red Module

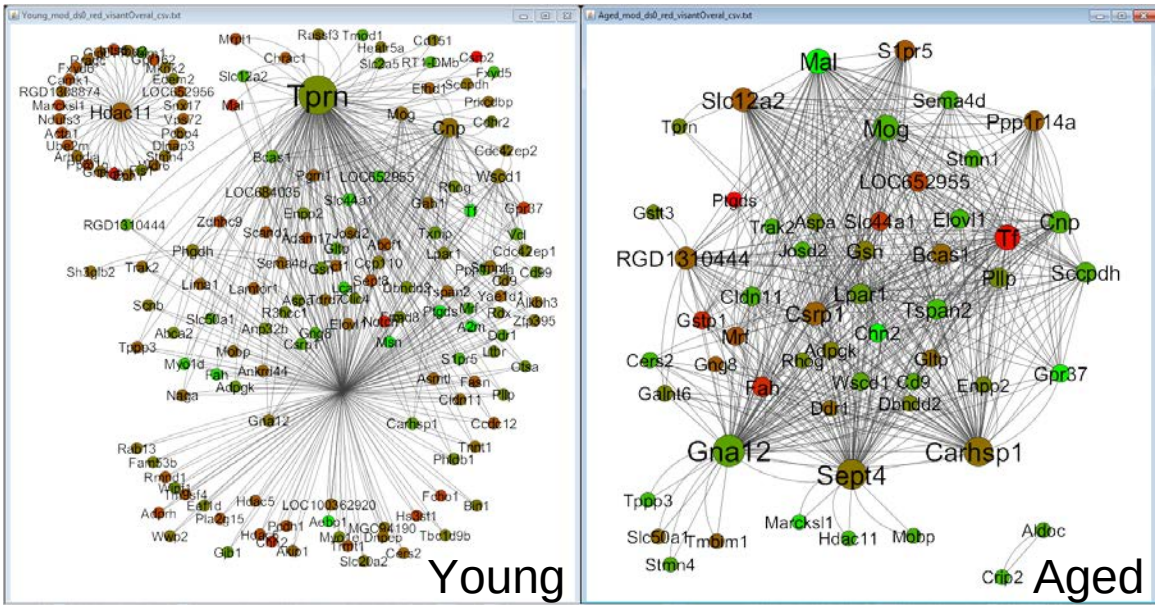


Figure 4.18 Differential co-expression network analysis for the red module in the young and aged in R7. The color of each node displays differential expression level (log fold change value) between young and aged. Each node size is proportional to the number of co-expression interaction the node has. Legends: red is upregulation; green is downregulation.

Differential co-expression for a subnetwork of a subset of genes from the original network can also be investigated. For example, Figure 4.19 shows a yellow module subnetwork involving *Cdk5r1*, a significant learning gene identified in the previous chapter through meta-analysis. This figure shows that *Cdk5r1* is co-expressing with only five other genes in the young subnetwork. While in the aged subnetwork, this gene is co-expressing with a much larger number of genes. Thus there is a difference in the number of co-expression connections (that *Cdk5r1* has with other genes) between the young and aged subnetworks. For example, the number of co-expression interactions for *Prkacb* and *Cdk5r1* is much higher in the aged network than in the young yellow network. Interestingly, *Mapre1*, *Dlg3*, *Impact*, and *Gabrg1* is only present in the young subnetwork, whereas, *Dpp6*, *Stxbp1*, *Kcnab2*, *Mapk1*, *Mapk9*, *Ppp2r2c* and others are only present in the aged subnetwork with large number of co-expression connections.

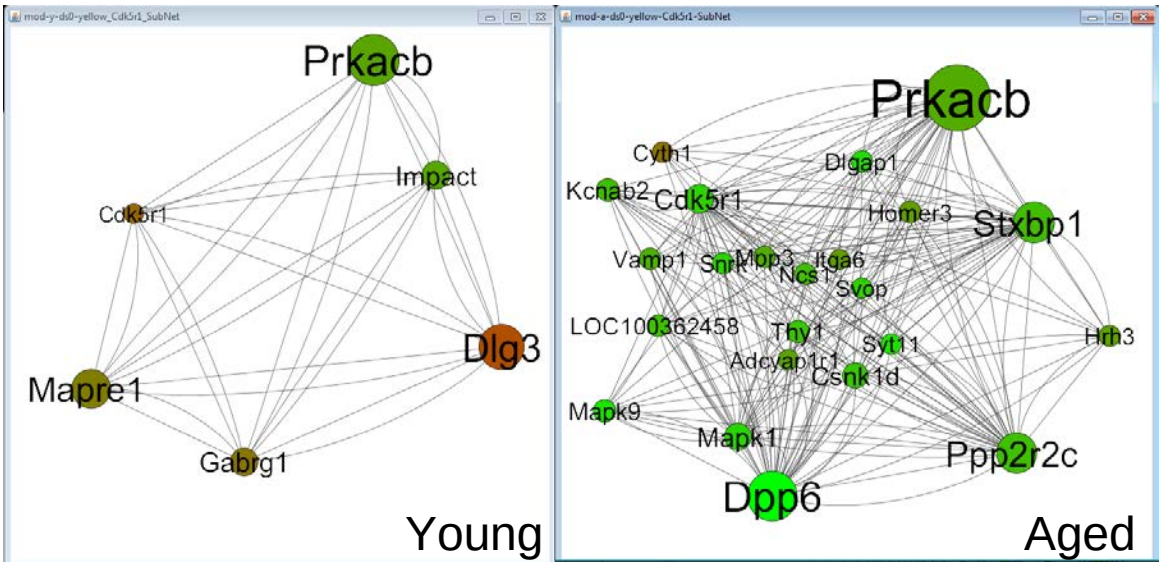


Figure 4.19 First neighbors of Cdk5r1 in the R7 yellow module.

4.3.10 Identifying and validating ASLI candidate hub genes

In a co-expression network, genes that are highly connected with many other genes are called hub genes. These genes show significant correlation with the module eigengenes and have high within-module connectivity. After closely studying the networks in young and aged, I have identified a set of key hub genes in each module. Some of the hub genes in the yellow module in R7 are shown in Table 4.8, which include *Camk1g*, *Cdk5r1*, *Cntn1*, *Dlg3*, *Dlgap1*, *Dpp6*, *Eif5*, *Gabrg1*, *Impact*, *Kcnab2*, *Mapk1*, *Mapre1*, *Ndfip2*, *Ppp2r2c*, *Prkacb*, *Pten*, *Rasgrp1*, *Scn2b*, and *Stxbp1*. These hub genes show differences in the number of co-expression interactions they have between the young and aged networks, and form a set of candidate hub genes for ASLI. Literature searches show that many of these genes function as kinases (e.g. *Camk1g*, *Dlg3*, *Mapk1*, and *Prkacb*) or phosphatases (*Ppp2r2c*) or are involved in the function of ion channels (*Dpp6*, *Gabrg1*, *Kcnab2*) (Table 4.7). Some of them are already known as learning genes and were identified in my meta-analysis. Table 4.6 shows the number of significant AY meta-analysis genes that are also members of different modules in the R7-Y network. Particularly, it shows that there are 165 AY significant meta-analysis genes in the yellow module. A list of these genes is presented in Appendix 6.5.1 which also includes few candidate hub genes.

Table 4.6 Significant AY meta-analysis genes common in R7-Y modules.

R7 Modules	Number of genes	Number of AY meta-analysis genes matching to each module
Blue	1015	275
Brown	759	195
Green	380	130
Grey	1319	334
Red	341	133
Turquoise	1129	275
Yellow	731	165

Effect size estimates from my meta-analysis results for the above ASLI candidate hub genes are summarized in Appendix 6.10.1. In addition, I have created individual forest plots for some of these hub genes, which are presented in Appendix 6.10.2 to Appendix 6.10.16. The combined meta-analysis results for these hub genes show that they were expressed at a very low level in the brain with comparatively lower standardized mean differences between young and aged, and thus failed to appear towards the top in the differentially expressed aging or learning gene list (Tables S1 and S2 in (Uddin and Singh, 2013)).

Table 4.7 Top candidate ASLI hub genes in the yellow module of the R7 dataset. Genes with an '*' were also identified as learning genes in my meta-analysis.

Hub Gene	Function Description	Reference
Camk1g	Encodes a protein similar to calcium/calmodulin-dependent protein kinase (CaMK), but its exact function is not known. CaMKs activated by the neuronal Ca2+ influx phosphorylate cAMP (cyclic adenosine monophosphate) responsive element binding protein (CREB), which has been implicated in spatial learning and memory formation.	(Thomas and Huganir, 2004; Voglis and Tavernarakis, 2006)
*Cdk5r1**	Involved in the pathology of Alzheimer's disease through the deregulated activity of Cdk5 (cyclin-dependent kinase 5), and also involved in synaptic plasticity, and learning and memory.	(Angelo et al., 2006; Shukla et al., 2012)
Cntn1	Contributes to the formation and function of neuronal connections, axon-glia communication, and necessary for myelin sheath formation by oligodendrocytes.	(Colakoglu et al., 2014; Ranscht, 1988)
*Dlg3**	Encodes a member of the membrane-associated guanylate kinase protein family; may play a role in clustering of N-methyl-D-aspartate (NMDA) receptors at excitatory synapses. It is highly enriched in the postsynaptic density (PSD), and plays essential roles in synaptic organization and plasticity.	(Elias and Nicoll, 2007; Elias et al., 2008; Wei et al., 2015)
Dpp6	Encodes an auxiliary subunit of voltage-gated potassium-4 channels and regulates the A-type K+ current gradient, which regulates dendritic excitability.	(Nadal et al., 2003; Wolf et al., 2014)
Eif5	Make 80S ribosomal initiation complex functional for translation.	(Si et al., 1996)
Gabrg1	Belongs to the ligand-gated ionic channel family. It is an integral membrane protein and plays an important role in inhibiting neurotransmission.	(Pirker et al., 2000; Ye and Carew, 2010)
*Kcnab2**	Encodes one of the beta subunits of the shaker-related Kv channels (Kv1.1 to Kv1.8) and found as a component of almost all potassium channel complexes containing Kv1 α subunits. It is a learning gene that is known to contribute to certain types of learning	(McKeown et al., 2008; Voglis and Tavernarakis, 2006)
*Mapk1**	Encodes a member of the MAP kinase family and is known as a learning gene. Hippocampal expression of Mapk1 is essential for synaptic plasticity and spatial learning.	(Selcher et al., 2001; Sweatt, 2001; Thomas and Huganir, 2004)
Mapre1	It is involved in the regulation of microtubule structures	(Kim et al., 2013;

	and chromosome stability.	Tirnauer et al., 2002)
Ndfip2	Affects receptor tyrosine kinase signaling by ubiquitinating several key components of the signaling pathways through binding to E3 ubiquitin ligases.	(Cristillo et al., 2003; Mund and Pelham, 2010)
Ppp2r2c	Ppp2r2c gene encodes one of the four B regulatory subunits of the PP2A (protein phosphatase 2A) enzyme complex. Inhibition of PP2A by inhibitor I1PP2A results in deficits in spatial reference memory and memory consolidation in adult rats.	(Backx et al., 2010; Xu et al., 2006)
Prkacb	Encodes the catalytic beta subunit of protein kinase A (PKA). PKA activates CREB and contributes to learning induced gene expression. Prkacb expression is required for LTP in the Hippocampus.	(Howe et al., 2002; Nguyen and Woo, 2003; Qi et al., 1996)
*Pten**	It modulates activation of the phosphatidylinositol 3-kinase (PI3K)/ protein kinase B (Akt) pathway. PTEN independently controls the structural and functional properties of hippocampal synapses and plays a direct role in activity-dependent hippocampal synaptic plasticity such as LTP and LTD.	(Blair and Harvey, 2012; Maehama and Dixon, 1998; Sperow et al., 2012)
Rasgrp1	It is a guanine nucleotide-exchange factor. When it is activated by Ca2+/calmodulin and diacylglycerol (DAG), it facilitates exchange of GDP to GTP and activates Ras.	(Stone, 2006)
Scn2b	Scn2b is a complex glycoprotein comprised of an alpha subunit and often one to several beta subunits. It was reported to have a role in epilepsy.	(Baum et al., 2014; XiYang et al., 2015)
Stxbp1	Plays a role in release of neurotransmitters via regulation of syntaxin, a transmembrane attachment protein receptor.	(Kurps and de Wit, 2012)

Legend: *Camk1g*, Calcium/calmodulin-dependent protein kinase I gamma; *Cdk5r1*, Cyclin-dependent kinase 5, regul. subunit 1 (p35); *Cntn1*, Contactin 1; *Dlg3*, Discs, large homolog 3; *Dlgap1*, Discs, large homolog-associated protein 1; *Dpp6*, Dipeptidyl-peptidase 6; *Eif5*, Eukaryotic translation initiation factor 5; *Gabrg1*, Gamma-aminobutyric acid (GABA) A receptor, gamma 1; *Impact*, Impact RWD domain protein (RWDD5); *Kcnab2*, Potassium channel, voltage gated shaker related subfamily A regulatory beta subunit 2; *Mapk1*, Mitogen-activated protein kinase 1 (ERK); *Mapre1*, Microtubule-associated protein, RP/EB family, member 1; *Ndfip2*, Nedd4 family interacting protein 2; *Ppp2r2c*, Protein phosphatase 2, regulatory subunit B, gamma; *Prkacb*, Protein kinase, cAMP-dependent, catalytic, beta; *Pten*, Phosphatase and tensin homolog; *Rasgrp1*, RAS guanyl releasing protein 1 (calcium and DAG-regulated); *Scn2b*, Sodium channel, voltage-gated, type II, beta; *Stxbp1*, Syntaxin binding protein 1.

The candidate ASLI hub genes were checked for their repeatability in networks constructed independently from B8, K9, and B7. The results are summarized in Table 4.8. Details of the hub gene validation data are available in Appendix 6.11.1 to Appendix 6.11.6. The results show that a number of hub genes from the yellow module are repeated in one or more independent datasets in B8, K9, or B7 with a p-value ≤ 0.05. From the R7 yellow module *Prkacb*, *Scn2b*, *Cntn1*, *Pten*, and *Ndfip2* were found present as hub genes in the K9 network; *Dlgap1* was found in the B7 and B8 networks; and *Camk1g* was found repeated in the B7 network. Notably, many of these hub genes were in the list of top 20 mean KME values in other networks, but their p-values were not significant, for example, *Dlg3*, *Mapre1*, *Dpp6*, *Stxbp1*, *Impact*, and *Mapk1*.

For the brown module, a set of candidate hub genes were identified and their repeatability as hub genes in networks constructed independently from datasets B8, K9, and B7 were checked. The results are summarized in Table 4.9 and the detail validation data are available in Appendix 6.13.1 to Appendix 6.13.6.

Table 4.8 Significant ASLI candidate hub genes from the yellow "learning" module and their repeatability in independent datasets.

Gene symbol	Number of co-expression in R7 network		Hub gene repeated in study	t-test p-value	Known learning gene
	Young	Aged			
Camk1g	0	4	B7-A	0.0003	No
Cdk5r1	5	22			Yes
Cntn1	6	0	K9-Y	0.0186	No
Dlg3	63	1			Yes
Dlgap1	0	7	B7-A, B8-A	0.0332	No
Dpp6	2	68			No
Eif5	36	1			No
Gabrg1	23	1			No
Impact	24	1			No
Kcnab2	2	10			Yes
Mapk1	9	19			Yes
Mapre1	49	1			No
Ndfip2	4	0	K9-Y	0.0217	No
Ppp2r2c	6	47			No
Prkacb	76	103	K9-Y	0.0523	No
Pten	2	0	K9-Y	0.0308	Yes
Rasgrp1	15	5			No
Scn2b	5	1	K9-Y	0.0028	No
Stxbp1	1	49			No

Table 4.9 Significant hub genes in the brown "cell process" module and their repeatability in independent datasets.

Hub genes in R7	Number of co-expression interaction in R7		Hub gene repeated in study	t-test p-value
	Young	Aged		
Acbd5	33	2		
Acyp2	36	0		
Araf	0	89	B8	0.0772*
Astn1	0	13		
Commd10	2	16		
Eif3m	1	2	B7, B8	0.0038
Fyttd1	45	2		
Hnrnpk	1	31		
Mettl14	178	2		
Mtmr6	5	0	K9	0.0234
Rpe	5	1	B8, K9	0.0001
Slc6a15	0	20	B7, B8	0.0097
Tspan31	0	30		
Zfp280d	0	57		

4.4 Discussion

WGCNA provides a simple methodology with which to construct gene co-expression network models from microarray gene expression data. I employed WGCNA for the first time in the analysis of ASLI microarray gene expression data. A key step in the network construction process was to determine the soft power beta. To choose a cutoff value to select the soft power, I made use of the scale-free topology criterion (Zhang and Horvath, 2005). The function pickSoftThreshold(…) estimated appropriate soft-thresholding powers for each dataset. It is recommended that a soft power greater than the power corresponding to R^2 value > 0.80 and a slope of the regression line between -1 to -2 produces approximate scale-free topology (Carlson et al., 2006; Horvath et al., 2006; Oldham et al., 2006). In order to meet the scale-free criterion and to have the same soft power for aged and young, I chose a soft power of 6 for R7, which was above the 0.80 threshold for a R^2 cutoff (Figure 4.2 and Figure 4.3). The power tables (e.g. Table 4.2) show that the resulting slope (minus the gamma parameter of the scale-free plot) looks reasonable. The slopes corresponding to the soft power 6 were -2.26 for young and -2.64 for aged. However, a slope of up to -3.4 was used in some instances (Miller et al., 2010). In case of R7, above soft power 6, the scale free topology fit did not improve much and showed saturation. There is a natural trade-off between maximizing the scale-free topology model fit (R^2) and maintaining a high mean (mean k > 30) number of connections (Zhang and Horvath, 2005). A signed $R^2 > 0.80$ can lead to a network satisfying scale-free topology at least approximately, while an R^2 value close to 1 may lead to networks with very few connections. In addition, the mean connectivity should be high enough so that the network contains sufficient information (e.g. for module detection). Thus selecting a soft power 6 was reasonable.

As a pre-processing step towards module detection, I restricted the network construction to genes with reasonably high connectivity. This helped eliminate genes that did not change their expression much between young and aged. These genes do not contribute to the correlation matrix. Further, this filtering process does not lead to a

big loss of information since module genes tend to have high connectivity (Zhang and Horvath, 2005; Oldham et al., 2006). Toward this end, average median connectivity in the aged and young groups was considered as a cutoff. After filtering out genes with very low connectivity, close to 5000 genes were selected from each dataset for network analysis, which was reasonable. Often three to four thousand genes are used in such analyses (Carlson et al., 2006; Zhang and Horvath, 2005). However, I wanted to include more genes because after matching genes among datasets in subsequent steps, the final number of genes can decrease significantly. Notably, genes with the most variable expression patterns across conditions (variable genes) can also be used, instead of the most connected genes, to create networks (Miller et al., 2008).

Next, the selected most connected genes were used to create co-expression networks. A number of free software applications are available to visualize network graphs of all or any single network module (e.g. igraph (http://igraph.org/), Gephi (http://gephi.github.io/), VisANT (http://visant.bu.edu/), Tulip (http://tulip.labri.fr/TulipDrupal/), CGV (http://www.informatik.uni-rostock.de/~ct/software/CGV/CGV.html), and Cytoscape (http://www.cytoscape.org/)). In this research I evaluated some of them and decided to use Cytoscape for all network visualization because of its strength and versatility. For R7 data the WGCNA started out with 5674 genes. Co-expression analysis of 5000 highly connected genes in R7-Y resulted in a dense mass of highly interconnected network (Figure 4.5). As expected, this type of network is not very helpful, which necessitates breaking it down to meaningful clusters or modules. Modules were identified from each dataset using the topological overlap. The topological overlap is considered a highly robust measure of network interconnectedness that combines the adjacency of two genes and the connection strengths these two genes share with other genes (Mason et al., 2009). To calculate the topological overlap for a pair of genes, their connection strengths with all other genes in the network were compared. The topological overlap values were used as input for average linkage hierarchical clustering. Modules were defined as branches of the resulting cluster tree (Langfelder et al., 2008). This module detection procedure has

been used in many applications (Carlson et al., 2006; Fuller et al., 2007; Ghazalpour et al., 2006; Horvath et al., 2006; Maschietto et al., 2015; Oldham et al., 2006; Oldham et al., 2008; Yang et al., 2014; Ye and Liu, 2015). Subsequent analysis identified six modules in the young network and 15 modules in the aged network in R7. Often it is desirable to have small number of large modules for comparing networks (Miller et al., 2010). Therefore, it was reasonable in this study to reduce the total number of unique (excluding the grey) modules in the aged network to six by merging the related modules. This allowed comparing networks between young and aged or comparing the aged networks between two different datasets.

Since gene network modules often correspond to biological pathways, focusing the analysis on modules (and their highly connected intramodular hub genes) amounts to a biologically meaningful data reduction scheme (Levine et al., 2013). One popular approach to understand the biological significance of coexpressed modules is to perform GO functional annotation enrichment analysis. Grouping genes based on functional similarity can systematically enhance the biological interpretation of large lists of genes derived from high throughput studies. A number of gene functional enrichment analysis tools are available (e.g. DAVID, GenMAPP (http://www.genmapp.org/), GOstats (https://bioconductor.org/packages/release/bioc/html/GOstats.html), EASE (https://david.ncifcrf.gov/content.jsp?file=/ease/ease1.htm&type=1), AmiGO (http://amigo.geneontology.org/amigo), and gProfiler (http://biit.cs.ut.ee/gprofiler/)). DAVID was used in this analysis because its functional annotation clustering report groups/displays similar annotations together which makes the biology clearer. If genes share a similar set of those terms, they are most likely involved in similar biological mechanisms. Interestingly, this analysis resulted in some very exciting outcomes in R7 young networks (i.e. genes in each module did correspond to a broad but distinct GO functional category) (Figure 4.10). Of particular interest to this research was the observation that, among all the modules, the yellow module was highly enriched with genes functioning in learning and memory related functions and pathways. This is

attributed to the fact that the datasets used in this research were indeed experimentally enriched (through the Morris water maze learning training) to identify genes involved in learning and memory impairment. Therefore, based on its GO functional analysis results, the yellow module is termed here as the “learning and memory” module.

The identified modules were validated in networks across independent datasets by comparing their preservation and repeatability (Figure 4.11). The rational was that, if a co-expression module is enriched with genes altogether serving a distinct function or phenotype critical for survival, then the module or its co-expression property should be preserved and repeatable. The results show that the preservation of all modules between young and aged within the same R7 study was higher compared to their preservation across studies (Figure 4.12). In general (except the green module), all other modules show moderate to high preservation across independent studies. The brown module shows the highest preservation among all the modules while the green module shows the lowest preservation. The yellow module shows the highest preservation between young and aged in the R7 dataset. However, it was not 100%, which could be attributed to the difference in the co-expression observed between R7 young and aged (Figure 4.15). This could be related to the decreased learning and memory in the aged animals observed in all studies. The modules identified in R7 young networks also showed repeatability in independent datasets with significant p-values (Figure 4.13 and Figure 4.14).

Cross-tabulation-based module validation measures that are employed in WGCNA provide powerful statistics which can be used to quantitate the extent to which disease related modules are present in other datasets (Levine et al., 2013). These module preservation and module overlap methods have been successfully used in the past in studying the preservation of clusters in human and chimpanzee brain networks (Oldham et al., 2006), in comparing human and mouse brain modules (Miller et al., 2010), and in comparing modules in different brain regions in human (Oldham et al., 2008). However, variations across studies, quality of data, and missing genes can all affect the outcome

of the network comparison and validation across studies. For example, all modules in general in the R7-A vs. B7-A comparison showed comparatively lower preservation than in the other comparisons. The reason for this poor performance is that the comparison was made across two independent studies and across two different chip types. The RGU34A chip type used in the B7 study had only 8799 genes. After various filtering B7 had only 4829 genes left to compare with the 5674 top most connected genes in R7-A. Taking the genes common between the two sets resulted in only 2140 genes and module preservation was estimated using these genes. Therefore, a lot of genes were missing in the comparison, for example, the RGU34A exclusive genes and 10431 RAE230A exclusive genes were not part of the comparison. This resulted in a lower $Z_{summary}$ score compared to others. However, four of the six modules still show moderate module preservation.

WGCNA is particularly very useful in identifying functional gene network modules and hub genes in a network, and also in comparing different networks. However, for shedding light on the causal processes underlying the observed data, correlation networks have some limitations. This is due to the fact that correlations confound direct and indirect associations (and thus between cause and effect) (Opgen-Rhein and Strimmer, 2007). However, despite the limitations in distinguishing direct causal interaction from indirect, WGCNA has become the most used gene network modeling approach due to its advantages in dissecting gene functional relationships in the form of hubs and modules.

Typical analysis of gene co-expression seeks to associate co-expression modules with disease or other phenotypic traits recorded in the same dataset. For instance, if the average expression of a particular module is higher in patients with more severe pathology, then the activity of genes in that module can be potentially linked to that pathological trait (Ghazalpour et al., 2006). While it would be desirable to identify causal molecular systems behind pathology, the trait-module association may be a downstream effect of the pathology (Gaiteri et al., 2014). However, this was not

possible in this research because trait related information was not available for any of the studies in the public database.

In summary, the major objective in this chapter was to identify and use a mathematical modeling approach that could better utilize the information captured in microarray data that traditional analysis was not able to do (Chapter 3). The major goal was to overcome some of the limitations observed in the traditional meta- and pathway analysis (Chapter 3) and identify novel ASLI related genes and their networks that are not limited to biological knowledge base alone. Based on the literature analysis WGCNA offered the best choice. I set R7-Y and R7-A as the exploratory datasets and used WGCNA to create gene network models from them. This analysis has identified as set of network modules from R7-Y, each of which is highly enriched with genes functioning in broad but distinct GO functional categories or biological pathways. Interestingly, the analysis pointed to a single (yellow) module that was highly enriched with genes functioning in learning and memory related functions and pathways. Subsequent, differential network analysis (Figure 4.15) and literature analysis (Table 4.7) of this yellow "learning and memory" module in R7-Y and R7-A allowed me to identify a set of novel ASLI candidate hub genes, some of which show significant repeatability in networks from independent validation datasets. These hub genes are highly co-expressed with other genes in the yellow network, which not only show differential expression but also differential co-expression and differential connectivity. The known function of these hub genes (Table 4.7) indicate that they play key roles in critical pathways, including kinase and phosphatase signaling, in functions related to various ion channels, and in maintaining neuronal integrity relating to synaptic plasticity and memory formation. Future study of these hub genes may help identify the molecular mechanisms responsible for age associated learning impairment, including spatial learning.

4.5 References

Ahmad, F.K., Deris, S., Othman, N.H., 2012. The inference of breast cancer metastasis through gene regulatory networks. J Biomed Inform. 45**,** 350-62.

Allen, J.D., Xie, Y., Chen, M., Girard, L., et al., 2012. Comparing statistical methods for constructing large scale gene networks. PLoS One. 7**,** e29348.
Angelo, M., Plattner, F., Giese, K.P., 2006. Cyclin-dependent kinase 5 in synaptic plasticity, learning and memory. J Neurochem. 99**,** 353-70.
Backx, L., Vermeesch, J., Pijkels, E., de Ravel, T., et al., 2010. PPP2R2C, a gene disrupted in autosomal dominant intellectual disability. Eur J Med Genet. 53**,** 239-43.
Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., et al., 2005. Reverse engineering of regulatory networks in human B cells. Nat Genet. 37**,** 382-90.
Baum, L., Haerian, B.S., Ng, H.K., Wong, V.C., et al., 2014. Case-control association study of polymorphisms in the voltage-gated sodium channel genes SCN1A, SCN2A, SCN3A, SCN1B, and SCN2B and epilepsy. Hum Genet. 133**,** 651-9.
Beal, M.J., Falciani, F., Ghahramani, Z., Rangel, C., et al., 2005. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics. 21**,** 349-56.
Blair, P.J., Harvey, J., 2012. PTEN: a new player controlling structural and functional synaptic plasticity. J Physiol. 590**,** 1017.
Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R., et al., 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A. 97**,** 12182-6.
Carlson, M.R., Zhang, B., Fang, Z., Mischel, P.S., et al., 2006. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics. 7**,** 40.
Chen, X., Chen, M., Ning, K., 2006. BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. Bioinformatics. 22**,** 2952-4.
Chickering, D.M., 1996. Learning Bayesian networks is NP-complete. Networks. 121-130.
Colakoglu, G., Bergstrom-Tyrberg, U., Berglund, E.O., Ranscht, B., 2014. Contactin-1 regulates myelination and nodal/paranodal domain organization in the central nervous system. Proc Natl Acad Sci U S A. 111**,** E394-403.
Cristillo, A.D., Nie, L., Macri, M.J., Bierer, B.E., 2003. Cloning and characterization of N4WBP5A, an inducible, cyclosporine-sensitive, Nedd4-binding protein in human T lymphocytes. J Biol Chem. 278**,** 34587-97.
de Jong, S., Fuller, T.F., Janson, E., Strengman, E., et al., 2010. Gene expression profiling in C57BL/6J and A/J mouse inbred strains reveals gene networks specific for brain regions independent of genetic background. BMC Genomics. 11**,** 20.
Elias, G.M., Nicoll, R.A., 2007. Synaptic trafficking of glutamate receptors by MAGUK scaffolding proteins. Trends Cell Biol. 17**,** 343-52.
Elias, G.M., Elias, L.A., Apostolides, P.F., Kriegstein, A.R., et al., 2008. Differential trafficking of AMPA and NMDA receptors by SAP102 and PSD-95 underlies synapse development. Proc Natl Acad Sci U S A. 105**,** 20953-8.
Emmert-Streib, F., Glazko, G.V., Altay, G., de Matos Simoes, R., 2012. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Front Genet. 3**,** 8.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., et al., 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 5**,** e8.

Fontenot, M., Konopka, G., 2014. Molecular networks and the evolution of human cognitive specializations. Curr Opin Genet Dev. 29**,** 52-9.

Fresno, C., Fernandez, E.A., 2013. RDAVIDWebService: a versatile R interface to DAVID. Bioinformatics. 29**,** 2810-1.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. J Comput Biol. 7**,** 601-20.

Fuller, T.F., Ghazalpour, A., Aten, J.E., Drake, T.A., et al., 2007. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome. 18**,** 463-72.

Gaiteri, C., Ding, Y., French, B., Tseng, G.C., et al., 2014. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav. 13**,** 13-24.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., et al., 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet. 2**,** e130.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., et al., 2006. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci U S A. 103**,** 17402-7.

Horvath, S., 2011. Weighted Network Analysis: Applications in Genomics and Systems Biology, Vol., Springer-Verlag New York.

Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., et al., 2003. Identifying biological themes within lists of genes with EASE. Genome Biol. 4**,** R70.

Howe, D.G., Wiley, J.C., McKnight, G.S., 2002. Molecular and behavioral effects of a null mutation in all PKA C beta isoforms. Mol Cell Neurosci. 20**,** 515-24.

Hu, Z., Mellor, J., Wu, J., DeLisi, C., 2004. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics. 5**,** 17.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37**,** 1-13.

Huang da, W., Sherman, B.T., Lempicki, R.A., 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4**,** 44-57.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., et al., 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35**,** W169-75.

Kim, M.J., Yun, H.S., Hong, E.H., Lee, S.J., et al., 2013. Depletion of end-binding protein 1 (EB1) promotes apoptosis of human non-small-cell lung cancer cells via reactive oxygen species and Bax-mediated mitochondrial dysfunction. Cancer Lett. 339**,** 15-24.

Kotera, M., Yamanishi, Y., Moriya, Y., Kanehisa, M., et al., 2012. GENIES: gene network inference engine based on supervised analysis. Nucleic Acids Res. 40**,** W162-7.

Kurps, J., de Wit, H., 2012. The role of Munc18-1 and its orthologs in modulation of cortical F-actin in chromaffin cells. J Mol Neurosci. 48**,** 339-46.
Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9**,** 559.
Langfelder, P., Zhang, B., Horvath, S., 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 24**,** 719-20.
Langfelder, P., Luo, R., Oldham, M.C., Horvath, S., 2011. Is my network module preserved and reproducible? PLoS Comput Biol. 7**,** e1001057.
Levine, A.J., Miller, J.A., Shapshak, P., Gelman, B., et al., 2013. Systems analysis of human brain gene expression: mechanisms for HIV-associated neurocognitive impairment and common pathways with Alzheimer's disease. BMC Med Genomics. 6**,** 4.
Maehama, T., Dixon, J.E., 1998. The tumor suppressor, PTEN/MMAC1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. J Biol Chem. 273**,** 13375-8.
Maetschke, S.R., Madhamshettiwar, P.B., Davis, M.J., Ragan, M.A., 2014. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. Brief Bioinform. 15**,** 195-211.
Mangegna, R.N., Stanley, H.E., 2000. An Introduction to Econophysics: Correlations and Complexity in Finance. , Vol., Cambridge University Press, Cambridge, UK.
Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., et al., 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 7 Suppl 1**,** S7.
Maschietto, M., Tahira, A.C., Puga, R., Lima, L., et al., 2015. Co-expression network of neural-differentiation genes shows specific pattern in schizophrenia. BMC Med Genomics. 8**,** 23.
Mason, M.J., Fan, G., Plath, K., Zhou, Q., et al., 2009. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics. 10**,** 327.
McKeown, L., Swanton, L., Robinson, P., Jones, O.T., 2008. Surface expression and distribution of voltage-gated potassium channels in neurons (Review). Mol Membr Biol. 25**,** 332-43.
Meyer, P.E., Kontos, K., Lafitte, F., Bontempi, G., 2007. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol. 79879.
Miller, J.A., Oldham, M.C., Geschwind, D.H., 2008. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. J Neurosci. 28**,** 1410-20.
Miller, J.A., Horvath, S., Geschwind, D.H., 2010. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc Natl Acad Sci U S A. 107**,** 12698-703.
Mund, T., Pelham, H.R., 2010. Regulation of PTEN/Akt and MAP kinase signaling pathways by the ubiquitin ligase activators Ndfip1 and Ndfip2. Proc Natl Acad Sci U S A. 107**,** 11429-34.

murphy, K., 2001. The bayes net toolbox for matlab. . Computing science and statistics 1024–1034. .

Myllymaki, P., Silander, T., Tirri, H., Uronen, P., 2002. B-course: A web-based tool for bayesian and causal data analysis. . International Journal on Artificial Intelligence Tools 369-388.

Nadal, M.S., Ozaita, A., Amarillo, Y., Vega-Saenz de Miera, E., et al., 2003. The CD26-related dipeptidyl aminopeptidase-like protein DPPX is a critical component of neuronal A-type K+ channels. Neuron. 37**,** 449-61.

Nazri, A., Lio, P., 2012. Investigating meta-approaches for reconstructing gene networks in a mammalian cellular context. PLoS One. 7**,** e28713.

Nguyen, P.V., Woo, N.H., 2003. Regulation of hippocampal synaptic plasticity by cyclic AMP-dependent protein kinases. Prog Neurobiol. 71**,** 401-37.

Oldham, M.C., Horvath, S., Geschwind, D.H., 2006. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci U S A. 103**,** 17973-8.

Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., et al., 2008. Functional organization of the transcriptome in human brain. Nat Neurosci. 11**,** 1271-82.

Opgen-Rhein, R., Strimmer, K., 2007. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol. 1**,** 37.

Peng, J., Wang, P., Zhou, N., Zhu, J., 2009. Partial Correlation Estimation by Joint Sparse Regression Models. J Am Stat Assoc. 104**,** 735-746.

Pirker, S., Schwarzer, C., Wieselthaler, A., Sieghart, W., et al., 2000. GABA(A) receptors: immunocytochemical distribution of 13 subunits in the adult rat brain. Neuroscience. 101**,** 815-50.

Plaisier, C.L., Horvath, S., Huertas-Vazquez, A., Cruz-Bautista, I., et al., 2009. A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. PLoS Genet. 5**,** e1000642.

Qi, M., Zhuo, M., Skalhegg, B.S., Brandon, E.P., et al., 1996. Impaired hippocampal plasticity in mice lacking the Cbeta1 catalytic subunit of cAMP-dependent protein kinase. Proc Natl Acad Sci U S A. 93**,** 1571-6.

Ranscht, B., 1988. Sequence of contactin, a 130-kD glycoprotein concentrated in areas of interneuronal contact, defines a new member of the immunoglobulin supergene family in the nervous system. J Cell Biol. 107**,** 1561-73.

Rickabaugh, T.M., Baxter, R.M., Sehl, M., Sinsheimer, J.S., et al., 2015. Acceleration of age-associated methylation patterns in HIV-1-infected adults. PLoS One. 10**,** e0119201.

Selcher, J.C., Nekrasova, T., Paylor, R., Landreth, G.E., et al., 2001. Mice lacking the ERK1 isoform of MAP kinase are unimpaired in emotional learning. Learn Mem. 8**,** 11-9.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13**,** 2498-504.

Shukla, V., Skuntz, S., Pant, H.C., 2012. Deregulated Cdk5 activity is involved in inducing Alzheimer's disease. Arch Med Res. 43**,** 655-62.
Si, K., Das, K., Maitra, U., 1996. Characterization of multiple mRNAs that encode mammalian translation initiation factor 5 (eIF-5). J Biol Chem. 271**,** 16934-8.
Song, L., Langfelder, P., Horvath, S., 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics. 13**,** 328.
Sperow, M., Berry, R.B., Bayazitov, I.T., Zhu, G., et al., 2012. Phosphatase and tensin homologue (PTEN) regulates synaptic plasticity independently of its effect on neuronal morphology and migration. J Physiol. 590**,** 777-92.
Steuer, R., 2006. Review: on the analysis and interpretation of correlations in metabolomic data. Brief Bioinform. 7**,** 151-8.
Stone, J.C., 2006. Regulation of Ras in lymphocytes: get a GRP. Biochem Soc Trans. 34**,** 858-61.
Sweatt, J.D., 2001. The neuronal MAP kinase cascade: a biochemical signal integration system subserving synaptic plasticity and memory. J Neurochem. 76**,** 1-10.
Thomas, G.M., Huganir, R.L., 2004. MAPK cascade signalling and synaptic plasticity. Nat Rev Neurosci. 5**,** 173-83.
Tirnauer, J.S., Grego, S., Salmon, E.D., Mitchison, T.J., 2002. EB1-microtubule interactions in Xenopus egg extracts: role of EB1 in microtubule stabilization and mechanisms of targeting to microtubules. Mol Biol Cell. 13**,** 3614-26.
Torkamani, A., Dean, B., Schork, N.J., Thomas, E.A., 2010. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. Genome Res. 20**,** 403-12.
Uddin, R.K., Singh, S.M., 2013. Hippocampal gene expression meta-analysis identifies aging and age-associated spatial learning impairment (ASLI) genes and pathways. PLoS One. 8**,** e69768.
Villaverde, A.F., Banga, J.R., 2014. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. J R Soc Interface. 11**,** 20130505.
Voglis, G., Tavernarakis, N., 2006. The role of synaptic ion channels in synaptic plasticity. EMBO Rep. 7**,** 1104-10.
Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., et al., 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 474**,** 380-4.
Wei, Z., Behrman, B., Wu, W.H., Chen, B.S., 2015. Subunit-specific regulation of N-methyl-D-aspartate (NMDA) receptor trafficking by SAP102 protein splice variants. J Biol Chem. 290**,** 5105-16.
Werhli, A.V., Grzegorczyk, M., Husmeier, D., 2006. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics. 22**,** 2523-31.
Wolf, E.J., Rasmusson, A.M., Mitchell, K.S., Logue, M.W., et al., 2014. A genome-wide association study of clinical symptoms of dissociation in a trauma-exposed sample. Depress Anxiety. 31**,** 352-60.

XiYang, Y.B., Wang, Y.C., Zhao, Y., Ru, J., et al., 2015. Sodium Channel Voltage-Gated Beta 2 Plays a Vital Role in Brain Aging Associated with Synaptic Plasticity and Expression of COX5A and FGF-2. Mol Neurobiol.

Xu, Y., Xing, Y., Chen, Y., Chao, Y., et al., 2006. Structure of the protein phosphatase 2A holoenzyme. Cell. 127**,** 1239-51.

Yang, Y., Han, L., Yuan, Y., Li, J., et al., 2014. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun. 5**,** 3231.

Ye, H., Liu, W., 2015. Transcriptional networks implicated in human nonalcoholic fatty liver disease. Mol Genet Genomics.

Ye, X., Carew, T.J., 2010. Small G protein signaling in neuronal plasticity and memory formation: the specific role of ras family proteins. Neuron. 68**,** 340-61.

Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 4**,** Article17.

Chapter 5 Discussion

5 Discussion

This research attempted to integrate microarray gene expression data generated from multiple independent studies in the context of ASLI in rats. The goal was to investigate genes that may be involved in ASLI and also the way that these genes may interact in networks contributing to or affecting various signaling pathways, which ultimately modulate ASLI phenotype. Chapter 2 examined the hypothesis that proper microarray data quality control and preprocessing are essential for any downstream analysis, whether it is for large scale data integration through meta-analysis, or for gene network analysis. My research reconfirms the notion proposed in the recent literature that when integrating data from multiple independent studies, data quality control should be considered as one of the most important steps in preprocessing. This was accomplished by removing outlier arrays and probe sets, following appropriate normalization methods, and removing batch effects. In Chapter 3, I integrated probe set level data from five independent studies involving ASLI. I used standardized mean difference effect size based meta-analysis followed by GO and pathway analysis. Interestingly, a large number of genes were identified that were differentially expressed between young and aged rats. I attribute this to the proper preprocessing, data integration, and meta-analysis methods that were applied to the gene expression data. Pathway analysis revealed that as the rats age many major pathways are affected. This finding is attributed to the aberrant gene expression patterns observed in old rats. However, our understanding of the candidate genes' functions and pathway interactions is limited to the current knowledge base. In addition, there is no prioritization of molecules within the knowledge-based network models of affected pathways. Moreover, the traditional methods are unable to fully utilize all the information that is contained within the microarray data. Therefore, Chapter 4 dealt with the suggestion that recent mathematical modeling approaches have the potential to utilize the gene interaction information present in microarray data and to help identify useful new candidate genes

and their networks. In this respect, I explored the use of co-expression networks using WGCNA, which allowed me to identify a set of network modules and ASLI candidate hub genes. These modules and candidate hub genes are repeatable across independent datasets. The implications of some of my major findings are discussed in detail under the following themes:

1. The effect of differential gene expression on aging and learning
2. Co-expression to cofunctionality – from the perspective of modules
3. Gene co-expression to co-functionality – from the perspective of hub genes
4. Differential expression vs. differential co-expression vs. differential connectivity
5. New insight into the molecular mechanisms of learning and memory formation
6. Study strength and limitations
7. Future directions
8. Conclusions

5.1 The effect of differential gene expression on aging and learning

Traditional differential expression analysis and the effect size meta-analysis, using the probe sets integrated from five independent studies involving ASLI, generated a larger number of differentially expressed genes (Chapter 3) between young vs. aged and aged-unimpaired vs. aged-impaired. I have performed a comprehensive functional and pathway analysis of these genes using the IPA knowledge base. This analysis has revealed major functions and pathways that are affected in the aged as well as aged-impaired animals. The results show that aging is affected by the genes functioning in cell viability, axonogenesis, and inositol phosphate metabolism. Further, these genes contribute to the imbalance in many major function categories including molecular transport, cell to cell signaling and interaction, and nervous system function. Considering the effect of the most significant differentially expressed genes on cellular biology, these genes could be classified into three distinct but non-exclusive categories: general aging (GA) genes that are associated with aging related disorders and not associated with any learning impairments; general aging genes associated with

syndromic learning impairments (GASI); and general aging genes associated with non-syndromic learning impairments (GANSI). Given the confounding effect of aging on learning impairments one may expect an overlap in the three groups of genes. Below I will summarize some key findings about some of the genes from each of the above three categories (GA, GASI, and GANSI). These genes presented significant up- or down regulation in the AY and IU comparisons (Table 3.1 and Table 3.5), and some of them were also identified as contributing to significantly increased or decreased function in the aged animals (Table 3.2).

5.1.1 GA or general aging genes

A majority of the genes that fall into this category were up-regulated in the aged rats in comparison to the young rats, and many have been implicated in disease vulnerability at old age in humans and animals. These GA genes affect a number of pathways including the Eif2 signaling, antigen presentation, complement system, and Ox40 signaling pathways (Table 3.4). The EIF2 signaling is activated (through the phosphorylation of eIF2α) in response to a wide array of cellular stresses to protects cells by reducing the general rate of protein synthesis while facilitating programs of stress-induced gene expression (Donnelly et al., 2013). OX40 is a member of the tumor necrosis factor (TNF) receptor family and plays a key role in the survival and homeostasis of effector and memory T cells and T-cell-mediated inflammatory diseases (Ishii et al., 2010).

The GA genes that are of special interest to this discussion are *C3*, *Cd74* (CD74 molecule, major histocompatibility complex, class II invariant chain), *Ctss* (cathepsin S), *Ctsz* (cathepsin Z), *Agt* (angiotensinogen), *Mbp* (myelin basic protein), and *Cck* (Cholecystokinin). Specifically, *C3*, *Cd74*, and *Agt* expression level was increased (Table 3.2, migration of cells function) and they affect the endocrine system disorders, gastrointestinal disease, and metabolic disease functions. *C3* (Table 3.1 and Figure 3.2) plays a central role in the activation of the complement system and is needed to restore tissue injury. However, inappropriate or excessive activation of the complement system can lead to cell death and tissue destruction, thus contributing to further injury and

impaired wound healing (Cazander et al., 2012). These consequences are clinically manifested in various disorders (Maier et al., 2008). *Cd74* (Table 3.1) participates in several key processes of the immune system including antigen presentation, B-cell differentiation, and inflammatory signalling. Overexpression of *Cd74* has been reported in some inflammatory diseases and several forms of cancer (Borghese and Clanchy, 2011), and is also known as an indicator of disease in some conditions. The longer form of CD74 also interacts with CTSS by direct binding (Mihelic et al., 2008), and both *Ctss* and *Ctsz* are also highly up-regulated in the aged rats (Table 3.1). Further, there is strong evidence implicating different AGT molecular variants as the cause of human essential hypertension and organ damage during aging (Arnold et al., 2013).

Expression of *Mbp* is known to decrease and *Cck* is known to affect the axonogenesis function in the aged animals (Table 3.2). My analysis has revealed an increased expression of *Mbp* and a decreased expression of *Cck* in the aged rats. MBP is a major constituent of the myelin sheath of oligodendrocytes and has an important role in the pathophysiology of multiple sclerosis (Moscarello et al., 2007), which is a chronic inflammatory and neurodegenerative disease of the CNS of unknown cause. *Cck* is extensively expressed in the brain and a number of diverse changes to hippocampal *Cck* expression profiles have been documented in various models of epilepsy (Wyeth et al., 2012). *Cck* is also known to have a role in modulating the neuronal network of anxiety and panic disorders that involve other parts of the brain (e.g. amygdala and hypothalamus) (Zwanzger et al., 2012). Such results argue that the GA genes in general are associated with reduction in physiological and immunological efficiency leading to deterioration (senescence) with advancing age in the aged rats.

5.1.2 GASI or general aging genes associated with syndromic learning impairments

Deterioration of mental and physical state is very common with advancing age and manifests itself in various syndromes. It is apparent that many syndromes associated with aging are also involved in memory loss and learning impairments. One such

syndrome is Alzheimer's disease, which has been studied extensively. Among the GASI genes identified in this analysis that have been implicated in Alzheimer's disease or late-onset Alzheimer's disease include *ApoE* (Bu, 2009; Corder et al., 1993), *Mapt* (Maeda et al., 2006), *Igf1r* (insulin-like growth factor 1 receptor) (O'Neill et al., 2012), *Clu* (Chen et al., 2012; Ferrari et al., 2012), *Picalm* (phosphatidylinositol binding clathrin assembly protein) (Chen et al., 2012; Ferrari et al., 2012), *Cdk5r1* (cyclin-dependent kinase 5, regulatory subunit 1, p35) (Shukla et al., 2012), and *Ide* (Insulin degrading enzyme) (Miller et al., 2003). *ApoE*, *Mapt*, *Igf1r*, *Clu*, and *Picalm* were up-regulated, and *Cdk5r1* and *Ide* were down-regulated in the aged animals compared to the young. These GASI genes may also lead to syndromic learning impairments by affecting various key neuronal functions. For example, *ApoE*, *Cdk5r1* and *Ide* are known to decrease cell viability and *Picalm* and *Igf1r* are known to affect axonogenesis (Table 3.2).

Specifically, *ApoE* and *Mapt* have been annotated as aging and learning genes in the IPA knowledge base (Table S3 in (Uddin and Singh, 2013)). *ApoE* gene is known as the strongest risk factor for age-related cognitive decline during normal ageing (Alzheimer's, 2012). APOE isoforms differentially regulate Aβ (amyloid β-peptide) aggregation and clearance in the brain, and have distinct functions in regulating brain lipid transport, glucose metabolism, neuronal signaling, neuroinflammation, and mitochondrial function (Liu et al., 2013). Toxicity of Aβ also depends on *Mapt* (Figure 3.4). Increase in MAPT levels may represent a very early sign of NFT (neurofibrillary tangle) formation and Alzheimer's disease in humans (Maeda et al., 2006). Down-regulated *Igf1r* activity has been implicated with prolonged human lifespan (O'Neill et al., 2012). When considering age-related neurodegeneration in Alzheimer's disease, signaling through the *Igf1r* is disturbed in the Alzheimer's disease patients' brain. An increased level of *Igf1r* has been reported in the degenerating synapses of the cerebral cortex within and surrounding Aβ plaques in people with Alzheimer's disease compared to people of the same age without the disease (O'Neill et al., 2012). Through the deregulated activity of *Cdk5*, *Cdk5r1* is involved in the pathology of Alzheimer's disease (Shukla et al., 2012), synaptic plasticity, learning, and memory (Angelo et al., 2006). IDE is involved in the degradation of Aβ and

other bioactive peptides (e.g. insulin and IGF-1 and IGF-2 in vitro) (Nalivaeva et al., 2012). PICALM plays a critical role in iron homeostasis and cell proliferation (Scotland et al., 2012). PICALM knockdown can result in reduced APP (amyloid precursor protein) internalization and Aβ generation, while overexpression can cause increased APP internalization and amyloid plaque load (Xiao et al., 2012). Irregularities in the Aβ clearance pathway are thought to initiate Aβ and tau protein accumulation in specific brain regions and consequent toxic events that lead to synaptic dysfunction and neurodegeneration in Alzheimer's disease. This is associated with the progressive destruction of synaptic circuits controlling memory and higher mental function.

Besides the above genes associated with Alzheimer's disease, there is a number of GASI genes associated with other age-related disease syndromes and related memory impairment. For example, *Cntn2* (Contactin-2), a learning gene, is up-regulated in the aged (Table S3 in (Uddin and Singh, 2013)), *Hmgb1* (high mobility group box-1) is up-regulated in the aged-impaired (Table 3.5), and *Tubb2b* is down-regulated in the aged rats (Table 3.1). *Cntn2* plays a role in the formation of axon connections (Lin et al., 2012) and autoimmune responses to *Cntn2* have been implicated in multiple sclerosis (Derfuss et al., 2009). Studies show that cellular stress, trauma, and inflammatory condition can also result in the up-regulation of *Hmgb1* in the hippocampus in aged rats, which results in reduced cognitive function in a reversal learning version of the Morris water maze test (He et al., 2012; Klune et al., 2008). Further, *Tubb2b* is a major component of microtubules cytoskeletal structures essential for cell motility and function and one of the top ten most down-regulated genes in the AY comparison (Table 3.1). A spectrum of neurological disorders (e.g. Polymicrogyria) characterized by abnormal neuronal migration, differentiation, organization, axon guidance, and maintenance has recently been associated with various mutations in *Tubb2b* (Cushion et al., 2013; Romaniello et al., 2012). In summary, a number of genes identified in the aged and aged-impaired animals are associated with a number of syndromes and fall in the category of GASI genes, which may contribute to the memory loss and learning impairments observed in the aged-impaired animals.

5.1.3 GANSI or general aging related genes associated with non-syndromic learning impairments

It is apparent that a majority of the differentially expressed genes in the aged or aged-impaired animals are known to facilitate learning and memory formation and are not implicated in any syndromes. They have been annotated as learning or spatial learning genes in the IPA knowledge base (Table S3 and S4 in (Uddin and Singh, 2013)). The canonical pathways that are most relevant to the GANSI genes functioning in the brain include nNos signaling pathway and glutamate receptor signaling pathway, which were identified most significant in the IU comparison (Table 3.7). nNos (Bartus et al., 2013; Shen et al., 2012) and glutamate receptors (Menard and Quirion, 2012) play an important role in neurotransmission and are critical to LTP, memory formation and synaptic plasticity.

The genes that deserve particular attention in the GANSI category are the 59 genes identified in the IU comparison following BH correction (effect sizes $p \leq 0.05$). These genes were differentially expressed in the aged rats with spatial learning impairment as compared to those without spatial learning impairment. *Arc*, a learning gene, is one of the most interesting of these 59 genes and is among the top ten most down-regulated genes in the aged-impaired animals (Table 3.5). The immediate-early gene *Arc* (aka *Arg3*) (Figure 3.3) expression is found to be vital for spatial memory consolidation and long-term synaptic plasticity in a variety of hippocampal-dependent and hippocampal-independent tasks, including spatial learning in the Morris water maze (Bramham et al., 2010; Shepherd and Bear, 2011). *Arc* is known for its tight experience-dependent regulation, dendritic mRNA transport, and local protein expression in activated synapses. For example, blocking *Arc* expression either using *Arc* knockout mice (Plath et al., 2006) or intra-hippocampal injections of *Arc* antisense oligonucleotides (Guzowski et al., 2000) is known to impair or prevent LTP without affecting short-term memory performance.

When I considered the larger list of 787 differentially expressed genes in the IU comparison (BH uncorrected, $p \leq 0.05$), I also found 48 genes annotated as learning or spatial learning genes in the IPA knowledge base (Table S4 in (Uddin and Singh, 2013)). Some of the interesting learning genes among theses 48 genes include *Camk2a* (calcium/calmodulin-dependent protein kinase II alpha), *Creb1*, *Crem* (cAMP responsive element modulator), *Egr1* (early growth response 1), *Homer 1* (homer homolog 1) (Figure 3.4), *Junb* (jun B proto-oncogene) (Figure 3.4), *Psen2* (presenilin 2), *Slc11a2* (solute carrier family 11), and *Marcks*. Particularly, *Marcks* ($p = 0.004$) (Figure 5D) is highly up-regulated in the aged-impaired animals. Timofeeva and colleagues (2010) recently reported that local infusions of MARCKS long peptide into the rat hippocampus resulted in a dramatic impairment of both working and reference memory in a dose-dependent manner with robust impairment at higher doses (Timofeeva et al., 2010), most likely through the inhibition of alpha7 nicotinic acetylcholine receptors (Gay et al., 2008). Thus, our analysis has identified these two genes, *Arc* and *Marcks*, as prime candidates for further investigation for their role in ASLI.

Additional GANSI genes include *Bdnf* (brain-derived neurotrophic factor), *Ntf3* (neurotrophin 3), *Igf2*, *Serpini1* (neuroserpin), *Gucy1a3* (guanylate cyclase 1, soluble, alpha 3), *Gucy1b3* (guanylate cyclase 1, soluble, beta 3), *Avp* (arginine vasopressin), *Gnaq* (guanine nucleotide binding protein), *Grp* (gastrin releasing peptide), *Pthlh* (parathyroid hormone-like hormone), *Trhr* (thyrotropin-releasing hormone receptor), *Agrn* (agrin), *L1cam* (Cell adhesion molecule L1), and *Ppp2ca* (protein phosphatase 2, catalytic subunit, alpha isozyme). These differentially expressed genes in the AY comparison play a critical role in the increase or decrease of several significant functions (Table 3.2) in the aged animals. Since a majority (73%) of the aged animals in the AY comparison were also impaired in the spatial learning task, it is not surprising that some of the aging genes may also contribute to the ASLI in these animals. Below I highlight some major functions of these GANSI genes.

For example, the genes *Bdnf* , *Ntf3*, *Igf2*, and *Serpini1* were down-regulated in the aged animals and are known to decrease cell viability of CNS cells (Table 3.2). The expression of neurotrophins such as *Bdnf* and *Ntf3* is strongly associated with synaptic function and plasticity. Specifically, *Bdnf* is known as a strong mediator for LTP (long term potentiation) in the hippocampus and play an essential role in memory formation in the adult brain (Park and Poo, 2013). *Igf2* is a late response gene regulated by the CREB-C/EBP pathway and plays a critical role in memory consolidation and enhancement (Chen et al., 2011). Furthermore, injections of recombinant IGF-II into the hippocampus after either training or memory retrieval significantly enhance memory retention and prevent forgetting. Neuroserpin e.g. *Serpini1* expression is involved in regulating the proteolytic balance associated with axonogenesis and synaptogenesis during development and synaptic plasticity in the adult (Lee et al., 2008; Osterwalder et al., 1996).

Further, *Gucy1a3* and *Gucy1b3* are involved in the increase of cellular movement function (Table 3.2). They are soluble guanylate cyclase (sGC) and are part of the nitric oxide (NO)/sGC/cGMP dependent protein kinase (PKG) signaling pathway that plays a key role in memory processing (Bartus et al., 2013; Shen et al., 2012). Inhibition of sGC, of PKG or of cGMP-degrading phosphodiesterase has been found to impair LTP (Monfort et al., 2004). Both GUCY1A3 and GUCY1B3 were found down-regulated in the aged animals, which may explain the ASLI in these animals.

The products of the genes *Avp* (Poulin and Pittman, 1993), *Gnaq* (Montmayeur et al., 2011), *Grp* (Roesler and Schwartsmann, 2012), *Pthlh* (Smogorzewski and Islam, 1995), and *Trhr* (Ramsdell and Tashjian, 1986) maintain the quantity and synthesis of IP_3 (inositol 1,4,5-triphosphate) level in the cell (Table 3.2) and were down-regulated in the aged rats. These genes facilitate IP_3 production in the brain and some through the activation of phospolipase C (PLC) (Montmayeur et al., 2011; Poulin and Pittman, 1993; Ramsdell and Tashjian, 1986; Roesler and Schwartsmann, 2012; Smogorzewski and Islam, 1995). Some of them (e.g. AVP (Ebstein et al., 2012) and GRPs (Roesler and

Schwartsmann, 2012)) are specifically involved in regulating cognition and memory. IP_3 is an important second messenger in the neuron produced from phosphatidylinsitol biphosphate (PIP_2) and cleaved by PLC. IP_3 binds to IP_3 receptors, which are gated Ca^{2+} channels that release calcium from the endoplasmic reticulum in to the cytosol (Finch and Augustine, 1998). Ca^{2+} in turn controls many different signaling events within neurons, including neurotransmitter release and gene expression in the cell nucleus. At least two Ca^{2+}-activated protein kinases (e.g. Ca^{2+}/calmodulin-dependent protein kinase (CaMKII) and protein kinase C (PKC)) have been implicated in LTP induction. LTP is the underlying cellular molecular mechanism that correlates with learning and memory formation (Foster, 2012). Thus, down regulation of the genes *Avp*, *Gnaq*, *Grp*, *Pthlh*, and *Trhr* can have a negative effect on the Inositol phospholipid-calcium-CamK-protein kinase C transduction pathway through decreased quantity and synthesis of IP_3 in the aged animals and directly or indirectly contribute to age-associated non-syndromic learning impairments such as ASLI.

A number of genes e.g. *Agrn*, *L1cam*, *Ppp2ca* that were down-regulated in the aged animals that demonstrated spatial learning impairment. Lower expression of these genes is known to decrease axonogenesis (Table 3.2). They play a critical role in neurite outgrowth, synaptogenesis, and synaptic plasticity. For example, high level of *Agrn* (Figure 3.4) expression was found in regions of the adult brain that show extensive synaptic plasticity. Recent studies demonstrated a substantial loss of excitatory synapses in the adult transgenic mice brain that lacked *Agrn* expression. Furthermore, they demonstrated inhibition of synaptogenesis by *Agrn* antisense oligonucleotides or *Agrn* siRNA in neuronal cell culture (Daniels, 2012). *L1cam* promotes the outgrowth of neurites and thereby contributes to formation of neuronal connections, learning, and memory (Kenwrick et al., 2000; Maness and Schachner, 2007) via activation of the mitogen-activated protein kinase (MAPK) pathway (Poplawski et al., 2012). *Ppp2ca* (aka Pp2a) is involved in Ca2+-dependent dephosphorylation of SNAP-25 (Iida et al., 2013) and SNAP-25 phosphorylation plays an important role in neural plasticity and long-term potentiation in the hippocampus (Genoud et al., 1999).

In conclusion, aged animals display a significant decrease in cell viability, axonogenesis, and inositol phosphate metabolism. They also show a significant increase in the migration of cells and differentiation of cells functions due to altered gene expression. The regulatory interactions of the differentially expressed genes seems to affect molecular transport, cell to cell signaling and interaction, nervous system development and function, and cell death and survival. The genes that are known to be involved in the above functional changes and/or those that present the most significant expression changes in the aged or aged-impaired animals could be broadly classified into three major categories such as GA, GASI, and GANSI. The GA genes are mostly involved in inflicting various aging related senescence (e.g. stress, disorders, and inflammation conditions) and generally are not associated with any learning impairment. The GASI genes, on the other hand, are associated with age-related neurological disease syndromes e.g. Alzheimer's disease, which generally affect normal cognitive functioning and may result into syndromic memory impairments. The most important group of genes perhaps is the GANSI genes, most of which show down-regulation in the aged or aged-impaired rats and by themselves usually are not associated with any syndromes. These genes affect various signal transduction pathways and functions in the brain contributing to the disruption of proper learning and memory formation. I propose that the GANSI genes should form the foundation of future studies in understanding age-associated memory impairments such as ASLI. These GASI and GANSI genes form a set of interesting candidates for future investigations as to how they interact with each other, how they are regulated, and what target genes they may affect in order to elucidate the mechanisms behind aging and age-associated spatial learning impairment.

5.2 Co-expression to co-functionality – from the perspective of modules

One useful property of a co-expression network is module. In a module the expression patterns of the genes are mutually correlated (Langfelder and Horvath, 2008). The focus on co-expression modules, each consisting of possibly hundreds of genes with common co-expression across samples, allows for a biologically motivated reduction of data while

also alleviating the problem of multiple comparisons (Levine et al., 2013). Further, just as correlated genes tend to have similar biological functions, on a larger scale, modules tend to contain genes with similar biological functions (Lee et al., 2004).The results obtained using WGCNA in this research and the follow up network analysis support these hypotheses. For example, the use of WGCNA reduced R7 data into a few biologically meaningful co-expression modules. The follow up GO analysis and literature search results were persuasive enough to indicate that each module gene set likely serve a distinct major biological function, thus, pointing to the above widely held notion of "co-expression to co-functionality". It is important to note that the networks and modules constructed from R7 microarray data were based on the gene expression patterns alone (i.e. there was no prior knowledge of the genes' function at the time of network construction). Once the networks were divided into modules and their module-wise GO functional analysis was performed, it was indeed observed that each module pointed to a broad but distinct category of biological function, and genes in each module shared similar subcategories of functions all converging to the broad functional category of the module (Table 4.4, Figure 4.10, and Appendix 6.7.1 to Appendix 6.7.6). Particularly, the genes in the yellow module showed significant enrichment in GO functions and pathways related to learning and memory formation in the brain. Although, the other modules are enriched with functions not directly related to learning and memory, they are critical for normal neuronal processes such as communication, growth, development, and maintenance. For example, genes in the brown module are significantly enriched in functions contributing to the various cellular processes and communication (Table 4.9), the green module genes in developmental processes (Appendix 6.7.3), and the red module genes in oligodendrocyte development (Appendix 6.7.4).

Thus, alteration of these modules' normal module-wise functions at old age through altered gene expression, as observed in the datasets, has the potential to affect normal functioning of learning and memory formation process. Preservation of these modules were not only validated across networks created from independent datasets, but also

the gene members of these modules demonstrated significant module membership (module overlap) across the independent networks (Figure 4.12 to Figure 4.14).

Gene co-expression analysis studies in multiples species, tissues, and platforms have shown that co-expressed genes tend to be functionally related (Obayashi et al., 2008; Oldham et al., 2008; Williams and Bowles, 2004). In order to investigate, whether observed clusters or modules of co-expressed genes are of functional significance, Lee and Sonnhammer (2003) observed that genes involved in the same biochemical pathways tend to be clustered together in a number of eukaryotic genomes. By a heuristic generalization known as “guilt by association”, it has been computationally established that functionally related genes are organized into co-expression networks, in practice assisting functional annotation of uncharacterized genes (Michalak, 2008). For example, physically interacting proteins in yeast were found to be encoded by co-expressed genes (Ge et al., 2001; Wuchty et al., 2006). These observations likely have inspired the development of co-expression network analysis methods. Gene network modeling using co-expression approaches provide insight into cellular activity as genes that are co-expressed often share common functions (Piro et al., 2011). Such networks have been widely used to study many diseases and phenotypes because of their ease of use and their ability to provide more biologically meaningful results (Chen et al., 2008; Gargalovic et al., 2006; Holtman et al., 2015; Maschietto et al., 2015; Min et al., 2012; Rickabaugh et al., 2015; Spiers et al., 2015; Ye and Liu, 2015; Zhou et al., 2014).

Microarray data captures functional relationship among genes that can provide biologically relevant information. In traditional microarray data analysis, however, these relationships remain essentially unexplored. Thus, a modular approach to gene function through WGCNA co-expression analysis provides a sensible way to extract such functional information from large microarray datasets in a biologically meaningful way. Particularly, my analyses have shown that specific learning associated functional gene modules can be identified through co-expression network modeling where genes in the

module show significant enrichment in learning and synaptic plasticity related GO functions.

5.3 Gene co-expression to co-functionality – from the perspective of hub genes

Hub genes play a central role in the structure of co-expression networks as they are often relevant to the function of regulatory networks. The ability to efficiently transit cellular signals within and between co-expressed clusters is facilitated by “hubs”, which are connected to a large number of nodes (Gaiteri et al., 2014). Analysis of the yeast protein-protein interaction network revealed that highly connected nodes are more likely to be essential for survival (Carter et al., 2004; Han et al., 2004; Jeong et al., 2000). Literature analysis indicate that the combined effect of the functions of the hub genes that are co-expressing together in individual modules may in fact contribute to the co-functionality of the whole module as discussed below.

5.3.1 Hub genes in the brown “cellular processes” module

A number of hub genes in this module contribute to various cellular communication and processes (Figure 4.16). For example, *Araf* (A-Raf proto-oncogene) is a proto-oncogene that belongs to the RAF subfamily of the serine/threonine protein kinase family, and is involved in cell growth and development (Mooz et al., 2014). *Fyttd1* (forty-two-three domain containing 1 aka UIF) was named UIF because it interacts with UAP56 (ATP-dependent RNA helicase). *Fyttd1* is an mRNA export adaptor, recruited to mRNA by other factors, binds to mRNA, and efficiently exports nuclear mRNA to the cytoplasm in vertebrates and other animals (Hautbergue et al., 2009). *Mettl14* (methyltransferase like 14) encodes a protein that catalyzes m(6)A RNA methylation. Together with METTL3, the only previously known m(6)A methyltransferase, these two proteins form a stable heterodimer core complex of METTL3-METTL14 that functions in cellular m(6)A deposition on mammalian nuclear RNAs (Liu et al., 2014). *Mtmr6* (myotubularin related protein 6) is a negative regulator of the Ca2+-activated K+ channel KCa3.1 (Srivastava et al., 2005) and plays a role in apoptosis (Zou et al., 2009). *Rpe* (ribulose-5-phosphate-3-

epimerase) catalyzes the reversible conversion of D-ribulose 5-phosphate to D-xylulose 5-phosphate and is an important enzyme for the cellular response against oxidative stress. *Rpe* functions in the pentose phosphate pathway (PPP). PPP confers protection against oxidative stress by supplying NADPH necessary for the regeneration of glutathione, which detoxifies H_2O_2 into H_2O and O_2 (Liang et al., 2011). *Slc6a15* (solute carrier family 6, member 15) encodes a member of the solute carrier family 6 protein family. The encoded protein is a Na+-dependent neutral amino acid transporter, thought to play a role in neuronal amino acid transport (Broer et al., 2006), and may be associated with major depression (Kohli et al., 2011). The *Tspan31* (tetraspanin 31, aka SAS) gene encodes a cell-surface protein that is a member of the transmembrane 4 superfamily, also known as the tetraspanin family (Jankowski et al., 1994). This protein mediates signal transduction events that play a role in the regulation of cell development, activation, growth and motility (Wright and Tomlinson, 1994). *Tspan31* is associated with tumorigenesis and osteosarcoma (Ragazzini et al., 1999).

5.3.2 Candidate ASLI hub genes in the yellow "learning and memory" module

The co-expression networks of the yellow "learning and memory" module (Figure 4.15) display a tight interrelationship of a large number of nodes with some hub genes. What is most interesting is that the co-expression of these hubs and nodes, as demonstrated by the WGCNA analysis, is not a random aggregation of some genes. Literature review suggests that the correlated expression pattern of the hub genes in the yellow networks may in fact be highly coordinated, and inside the young rats' hippocampus they may be serving a common purpose. This purpose could be to maintain the functional integrity of the normal process of learning and memory formation mechanisms, which are disrupted in the aging brain. I have short listed 19 genes as candidate ASLI hub genes from both the young and aged networks based on their co-expression connection to other genes. These genes include *Camk1g*, *Cdk5r1*, *Cntn1*, *Dlg3*, *Dlgap1*, *Dpp6*, *Eif5*, *Gabrg1*, *Impact*, *Kcnab2*, *Mapk1*, *Mapre1*, *Ndfip2*, *Ppp2r2c*, *Prkacb*, *Pten*, *Rasgrp1*, *Scn2b*, and *Stxbp1*. Below I will discuss literature findings of some of these candidate ASLI hub genes in

combination with the results from the meta-analysis. This will show that some of these hub genes are already known as key learning and memory genes and have well established roles in memory functions. While for others, information is emerging indicating their direct or indirect role in learning and memory.

5.3.2.1 Camk1g (calcium/calmodulin-dependent protein kinase IG)

This gene encodes a protein similar to calcium/calmodulin dependent protein kinase. Calcium ions binding to calmodulin can regulate protein phosphorylation/ dephosphorylation. Neuronal Ca^{2+} is known to play a critical role as an intracellular second messenger, linking neuronal excitability with many kinds of cellular biological events including synaptic plasticity, neuronal cell survival, and apoptosis (Berridge et al., 1998; Bito, 1998; Bliss and Collingridge, 1993). Ca^{2+} ions bind to calmodulin, a ubiquitous and evolutionary well conserved intracellular Ca^{2+} receptor, and form a complex, which mediates a significant part of signaling downstream. Although, a large number of molecules have been shown to be targeted and activated by the Ca^{2+}/ calmodulin complex, one subgroup of multifunctional kinases, Ca^{2+}/calmodulin-dependent protein kinases (CaMKs), has been ascribed a prominent role (Bito and Takemoto-Kimura, 2003; Takemoto-Kimura et al., 2003). CaMKs such as CaMKII can then activate a number of other targets such as CREB, which is important in synaptic plasticity and learning. CaMKs play a significant role in learning and memory formation through the activation of CREB signaling (Baudry et al., 2014; Bito and Takemoto-Kimura, 2003; Sweatt, 2001; Thomas and Huganir, 2004). It is very likely that *Camk1g*, which has not been reported before in relation to memory impairment, may function in a similar manner. It is likely that down-regulation of *Camk1g* in the aged rats (Appendix 6.10.1 and Appendix 6.10.2) may in fact contribute to ASLI in those animals.

5.3.2.2 Cdk5r1 (cyclin-dependent kinase 5, regulatory subunit 1)

Cdk5r1 (aka p35, p23; p25; CDK5R; NCK5A; CDK5P35; p35nck5a) is one of the genes identified as a general aging genes associated with syndromic learning impairments (Chapter 3) and is implicated in Alzheimer's disease or late-onset Alzheimer's disease

(Shukla et al., 2012). *Cdk5r1* was down-regulated in the aged animals (effect size = -0.66, p-value = 0.03) (Appendix 6.10.1 and Appendix 6.10.3) compared to the young, and is known to decrease cell viability (Table 3.2). *Cdk5r1* is involved in the pathology of Alzheimer's disease (Shukla et al., 2012), synaptic plasticity, learning, and memory through the deregulated activity of *Cdk5* (Angelo et al., 2006). The protein CDK5R1 is a neuron-specific activator of CDK5; the activation of CDK5 is required for proper development of the central nervous system.

A literature review revealed contrasting roles of *Cdk5* in learning and memory formation. These findings suggest that *Cdk5* not only promotes LTP and LTD, but also counteracts changes induced by LTP and LTD to maintain neuronal network stability, thereby functioning as a homeostatic regulator of synaptic plasticity (Shah and Lahiri, 2014). For example, a positive role of *Cdk5* in promoting synaptic plasticity was identified in *Cdk5r1*$^{-/-}$ (p35$^{-/-}$) mice, which display depotentiation (Ohshima et al., 2005). Likewise, Cdk5 is transiently upregulated in mice that are exposed to stress and facilitates context-dependent fear conditioning (Fischer et al., 2002). In contrast, Hawasli et al. (2007) demonstrated that the initial loss of Cdk5 in Cdk5 conditional knockout mice results in enhanced LTP, NMDA-receptor-mediated synaptic plasticity, and improved performance in hippocampal behavioral learning tasks, which highlights a negative role of Cdk5 in learning and memory formation. The p25 form of *Cdk5r1* is the principal activator of *Cdk5*. It is generated as a calcium-dependent degradation product of p35 form of *Cdk5r1* (Seo et al., 2014). The p25 form generation is associated with normal memory formation in the mouse hippocampus. In addition, p25 overexpression enhances synaptogenesis. Therefore, it is possible that p25 generation may act as a molecular memory mechanism that is impaired in early Alzheimer's disease (Giese, 2014; Seo et al., 2014). Interestingly, results from recent studies indicate that Cdk5/p35 is required for motor learning and involved in long-term synaptic plasticity (He et al., 2014). However, further studies are required to understand the specific roles *Cdk5r1* play in ASLI, and the mechanisms involved.

5.3.2.3 Cntn1 (Contactin 1)

Cntn1 is a membrane glycoprotein that provides critical signals for axon–glia communication in CNS myelin. It is expressed in a variety of neurons and contributes to the formation and function of neuronal connections (Ranscht, 1988). *Cntn1* has been studied as a prime candidate for multiple sclerosis (Colakoglu et al., 2014). Gene ablation in mice shows that *Cntn1* is necessary for myelin sheath formation by oligodendrocytes as well as for the establishment of paranodal axoglial junctions, which regulate the domain organization and enable rapid nerve impulse conduction of myelinated nerves (Colakoglu et al., 2014). *Cntn1*deficiency also resulted in mislocalized potassium Kv1.2 channels, abnormal myelin terminal loops, and reduced numbers and impaired maturation of sodium channel clusters along with significant hypomyelination (up to 60% myelin loss). Interestingly, *Cntn2* (Table 3.2), identified as a significant gene in the meta-analysis, also plays a role in the formation of axon connections (Lin et al., 2012). Autoimmune responses to *Cntn2* have been implicated in multiple sclerosis (Derfuss et al., 2009). However, *Cntn1* was up regulated in the aged rats (effect size = +0.36, pvalue = 0.04) in my combined meta-analysis with the K9 study showing a down regulation (Appendix 6.10.1 and Appendix 6.10.4). So, like *Cntn2*, *Cntn1* may simply play a role as a GASI gene that affects learning through the process of normal aging.

5.3.2.4 Dlg3 (Discs, large homolog3)

Dlg3, also known as synapse-associated protein 102 (SAP102), is a scaffolding protein highly enriched in the postsynaptic density (PSD), and plays an essential role in synaptic organization and plasticity (Elias and Nicoll, 2007). *Dlg3* interacts directly or indirectly with major types of glutamate receptors. It binds directly to *N*-methyl-d-aspartate receptors (NMDARs), anchors receptors at synapses, and facilitates transduction of NMDAR signals (Wei et al., 2015). *Dlg3* null mice that survive into adulthood show impairments in synaptic plasticity and spatial learning (Cuthbert et al., 2007). Accordingly, recent studies have demonstrated that *Dlg3* plays an important role in excitatory synapse formation through the PAK (p21-activated kinases) signaling pathway

(Murata and Constantine-Paton, 2013). These findings are consistent with other electrophysiological studies showing that *Dlg3* regulates glutamate receptor trafficking during synaptogenesis (Elias et al., 2008). These studies demonstrate that expression of *Dlg3* is critical in proper cognitive development and functioning including spatial learning. In my meta-analysis, *Dlg3* showed lower expression (effect size = -0.69, p-value 0.06) in the aged rats compared to the young in three of the five studies (Appendix 6.10.1 and Appendix 6.10.5), and did not make it to the significant gene list because of its p-value. However, it is known as a learning gene (Section 3.3.3) (IPA). In the WGCNA networks this gene showed strong differential co-expression. Taken together, *Dlg3* presents itself as a highly promising candidate ASLI gene.

5.3.2.5 Dpp6 (Dipeptidyl-peptidase 6)

Dpp6 has been studied for its association with autism (Marshall et al., 2008), and for its relation to multiple (Brambilla et al., 2012) and lateral sclerosis (Blauw et al., 2010). However, not much is known about the direct role of *Dpp6* in learning and memory. The DPP6 protein is an auxiliary subunit of voltage-gated potassium-4 channels (Kv4). DPP6 influences neuronal excitability and communication of excitability to distal dendrites, very likely by regulating the A-type K+ current gradient (Nadal et al., 2003). Dendritic excitability has been found to be critical for synaptic integration and excitation (Wolf et al., 2014). Hippocampal recordings from *Dpp6* knock-out mice demonstrate a decrease in this gradient and increased dendritic excitability (Sun et al., 2011). Given the role of DPP6 in synaptic integration, it is possible that this protein also plays a role in dissociation. Dissociation is a state that is defined by poor integration of incoming sensory experiences and problems with region-specific cognitive processes that are ordinarily organized dynamically across time. Intriguingly, in addition to autism spectrum disorder and multiple sclerosis, *Dpp6* has also been implicated as a potential susceptibility gene in Schizophrenia (Tanaka et al., 2013). Dendritic excitability may turn out to be a common function affected by these neurological diseases. Hippocampal neurons lacking DPP6 show a sparser dendritic branching pattern along with fewer

spines throughout development and into adulthood. Thus, *Dpp6* plays an important role in cell adhesion and motility, impacting the hippocampal synaptic development and function (Lin et al., 2013). *Dpp6* is another hub gene that showed lower expression in the aged compared to young in the meta-analysis in 4 of the 5 studies (effect size = -0.42, p-value = 0.38), with only B7 showing higher expression in the aged (Appendix 6.10.1 and Appendix 6.10.6). Interestingly, Kcnab2, a voltage-gated potassium channel co-expressed with Dpp6 in the aged yellow "learning and memory" module, also showed decreased expression in the aged compared to the young. Future studies should investigate specific role of *Dpp6* in ASLI.

5.3.2.6 Eif5 **(**Eukaryotic translation initiation factor-5)

Eif5 interacts with the 40S initiation complex to promote hydrolysis of bound GTP with concomitant joining of the 60S ribosomal subunit to the 40S initiation complex. The resulting functional 80S ribosomal initiation complex is then active in peptidyl transfer and chain elongations (Si et al., 1996). *Eif5* is up-regulated in the aged rats (effect size = 0.42, p-value 0.04) (Appendix 6.10.1 and Appendix 6.10.7). Not much is known about Eif5's involvement in learning and memory impairment. Interestingly, EIF2 signaling pathway was one of the top canonical pathways that was affected by the GA genes in the meta-analysis.

5.3.2.7 Gabrg1 (gamma-aminobutyric acid (GABA) A receptor, gamma 1)

GABA can inhibit action potential firing in mammalian neurons. GABAA receptor (GABAAR) channels mediate the majority of inhibitory neurotransmissions in the mammalian brain. These receptors are pentamers assembled from a large family of subunits, of which 19 members have so far been identified. Receptors targeted to the synaptic compartment are composed of two α, two β, and a single γ subunit (Pirker et al., 2000). The ionotropic GABA receptors are usually inhibitory because their associated channels are permeable to Cl^-; the flow of the negatively charged chloride ions inhibits

postsynaptic cells since the reversal potential for Cl^- is more negative than the threshold for neuronal firing.

Neurofibromin (NF1), a RasGAP, restricts GABA release from inhibitory neurons and is important for memory formation (Costa et al., 2002; Cui et al., 2008). Cui et al. (2008) demonstrate that the learning deficits in a mouse model of neurofibromatosis type I are caused by increased hippocampal GABA release, which dampens hippocampal synaptic plasticity and consequently leads to hippocampal-dependent learning deficits.

Gabrg1 was upregulated in 3 of the 5 studies in the meta-analysis in this research (Appendix 6.10.1 and Appendix 6.10.8). Interestingly few other GABA receptors were also co-expressed along with *Gabrg1*, for example, *Gabbr1*, *Gabrb2*, *Gabrb3*, and *Gabra4*, which showed mixed expression in the aged (some up and some down).

5.3.2.8 Kcnab2 (potassium channel, voltage gated shaker related subfamily A regulatory beta subunit 2)

Kcnab2 (aka AKR6A5; KCNA2B; HKvbeta2; KV-BETA-2; HKvbeta2.1; HKvbeta2.2) is known as a learning gene (Section 3.3.3) (IPA). Voltage-gated potassium (Kv) channels represent the most complex class of voltage-gated ion channels from both functional and structural standpoints (Lai and Jan, 2006; McKeown et al., 2008). Their diverse functions include regulating neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume. Four sequence-related potassium channel genes - shaker, shaw, shab, and shal - have been identified in Drosophila, and each has been shown to have human homolog(s). *Kcnab2* gene encodes one of the beta subunits of the shaker-related Kv channels (Kv1.1 to Kv1.8) and this subunit is found as a component of almost all potassium channel complexes containing Kv1 α subunits (McKeown et al., 2008). This association of beta subunits with Kv1 channels not only increases the potential for diversity, it also indicates that the functional properties of individual channels are governed by the specific combination of alpha and beta subunits present in the channel

complex (Rhodes et al., 1996). For example, *Kcnab2* can alter functional properties of the *Kcna4* gene product (Kv1.4).

Specific potassium channels, gated by intracellular calcium elevation, have been associated with synaptic plasticity (Voglis and Tavernarakis, 2006). Specific, non-synaptic voltage-gated potassium (Kv) channels are also important for controlling neuron membrane electrical excitability and are localized to axons, somata and dendrites. Deletion of *Kcnab2* in mice leads to deficits in associative learning and memory and loss of this gene function likely contributes to the cognitive and neurological impairments in humans (Perkowski and Murphy, 2011).

In my meta-analysis, *Kcnab2* showed an effect size of 0.04 with a p-value of 0.77 (Appendix 6.10.1). This is due to the fact that the SMD was slightly down in the aged in R7 and K9, but up in BL, B7 and B8 (Appendix 6.10.9). Given the diverse and delicate nature of these ion channels, which are constantly changing in quantity and locations, none of the studies was successful in recording the exact expression of this gene. Nonetheless, given the involvement of the Kv1 channels in synaptic plasticity, the exact function of this hub gene in ASLI requires future study.

5.3.2.9 Mapk1 (mitogen-activated protein kinase 1)

Mapk1 (aka ERK; p38; p40; p41; ERK2; ERT1; ERK-2; MAPK2; PRKM1; PRKM2; P42MAPK; p41mapk; p42-MAPK) encodes a member of the MAP kinase family. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), are serine/threonine kinases, which act as an integration point for multiple biochemical signals. They are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation, and development (Cuadrado and Nebreda, 2010). The activation of ERKs requires their phosphorylation by upstream kinases. Upon activation, these kinases translocate to the nucleus of the stimulated cells, where they phosphorylate nuclear targets. The targets of ERKs include transcription factors, cytoskeletal proteins, regulatory enzymes and, importantly, other kinases (Thomas and Huganir, 2004). The

role of *Mapk1* in learning and memory is well known (Selcher et al., 2001; Sweatt, 2001). MAP kinases are activated in neurons in response to excitatory glutamatergic signaling, which controls many forms of synaptic plasticity that are thought to underlie higher brain process such as learning and memory. The forms of long-term memory in mammals in which the involvement of ERK have been best- characterized are spatial learning and fear conditioning. Specifically, the necessity for hippocampal ERK activation for spatial memory formation was clearly demonstrated in mice using the Morris water maze (Blum et al., 1999; Selcher et al., 1999), and for fear conditioning using context dependent audible cue and subsequent heat-shock (Atkins et al., 1998; Schafe et al., 2000). In all these studies, the MEK (an early activator of MAPK) inhibitors greatly impaired memory retention in spatial learning or far less frequent freeze in fear conditioning experiments. *Mapk1* is known as a learning gene (Section 3.3.3) (IPA) and was down regulated in the aged (effect size = -0.55. p-value = 0.14) (Appendix 6.10.1 and Appendix 6.10.10), which is consistent with the finding reported in the literature.

5.3.2.10 Mapre1 (Microtubule-associated protein, RP / EB family, member 1; aka EB1)

Although the end-binding protein *Mapre1* is well known for its role in regulating microtubule dynamics (Tirnauer et al., 2002), its role in learning and memory is not as well understood. In a recent study, Oz and colleagues (Oz et al., 2014) showed that *EB1* (*Mapre1*) binds to ADNP (activity-dependent neuroprotective protein) at the NAP motif (8-amino acid peptide) and regulates dendritic spine growth, ultimately leading to prevention of neuronal death and protection against cognitive deficiencies in mice. ADNP is essential for brain formation and is shown to contribute to aging. Down-regulation of *EB1* promotes non-small-cell lung cancer cell death by inducing ROS-mediated, NF-κB-dependent Bax signaling cascades (Kim et al., 2013). Recent studies suggest that *EB1* is associated with a variety of microtubule-mediated cellular activities in various systems, including migration, cell division, and morphogenesis (Kim et al., 2013). Thus, *EB1* plays a crucial role in ADNP function along with other molecules including EB3 and PSD-95 (*Dlg4*) (Oz et al., 2014). Thus *EB1* possibly contributes to

neurite outgrowth, the growth cone, axonal transport, and synaptic plasticity. Down-regulation of *Mapre1* (*EB1*) in the aged rats (effect size = -0.41, p-value = 0.02) (Appendix 6.10.1 and Appendix 6.10.11) and the associated deficiency in learning of these aged rats is in line with the findings in the literature.

5.3.2.11 Ndfip2 (Nedd4 family interacting protein 2)

Ndfip1 and *Ndfip2* are related endosomal membrane proteins that bind to and activate members of the Nedd4 family of E3 ubiquitin ligases (Cristillo et al., 2003). These ligases in turn affect receptor tyrosine kinase signaling by ubiquitinating several key components of the signaling pathways. They associate with the EGF (epidermal growth factor) receptor and PTEN (another learning gene, Section 5.3.1.14), and control the ubiquitination and abundance of PTEN, c-Cbl, and Src family kinases. *Ndfip2* also binds to and is phosphorylated by Src and Lyn, and can act as a scaffold for Src phosphorylation of *Ndfip1* and potentially other substrates. Depletion of *Ndfip1* inhibits Akt activation in EGF-stimulated HeLa cells, stimulates activation of Jnk, and enhances cell multiplication. *Ndfip1* and *Ndfip2* are physically and functionally associated with multiple components of the EGF signaling cascade, and their levels modulate the relative output of different signaling pathways (Mund and Pelham, 2010). It is possible that *Ndfip2*, which was down-regulated in the aged (effect size = -0.38, p-value = 0.22) (Appendix 6.10.1) compared to the young, might be working in the same fashion as NGF (nerve growth factor) to influence TrkA pathway. In the brain, NGF can phosphorylate tyrosine kinase receptor TrkA in the plasma membrane which later activates a number of downstream pathways (Purves et al., 2004). In fact, EGF and NGF, use the same pathway such as Raf → MEK → ERK to promote distinct outcomes in PC12 cell line, which include neuritogenesis, gene induction, and proliferation (Vaudry et al., 2002). However, EGF and NGF likely work differently and on different receptor tyrosine kinases (Lee et al., 2002). For example, K252a, a well-established inhibitor of Trk tyrosine kinases inhibits NGF activation of Trk receptors and the subsequent biological effects of neurotrophins, without affecting other receptor tyrosine kinases, such as the EGF and

FGF receptors (Berg et al., 1992). In addition, the duration of signaling through ERKs may produce different outcomes of EGF and NGF stimulation. EGF induces rapid and transient Ras- and Rap1-dependent ERK phosphorylation, whereas NGF stimulation of ERK is both rapid and sustained, with sustained activation dependent on signaling to ERK through Rap1 (Qui and Green, 1992; York et al., 2000).

5.3.2.12 Ppp2r2c (Protein phosphatase 2, regulatory subunit B, gamma)

Ppp2r2c gene encodes one of the four B regulatory subunits of the PP2A (protein phosphatase 2A) enzyme complex. The enzyme PP2A is a Serine/Threonine phosphatase that plays an important role in cell-cycle regulation, control of cell-growth, regulation of multiple signal transduction pathways, cytoskeleton dynamics, and mobility (Xu et al., 2006). *Ppp2r2c* was down-regulated in the aged rats (effect size -0.43, p-value 0.22) in my meta-analysis (Appendix 6.10.1 and Appendix 6.10.12).

The exact function of *Ppp2r2c* is not yet known. However, Backx et al. (2010) reported a unique expression pattern for *Ppp2r2c* with a very high expression in the hippocampus in *in situ* experiments in normal adult mice. In addition, they found PPP2R2C is disrupted in autosomal dominant intellectual disability. Combined with its unique expression pattern in mouse brain, this suggests a role for *Ppp2r2c* in synaptic plasticity and hence learning and memory (Backx et al., 2010).

PP2A (*Ppp2r2c*) deficiency is a cause of the abnormal hyperphosphorylation of tau, which composes neurofibrillary tangles (NFTs) in the Alzheimer's disease brain. Studies have shown that PP2A activity was 30% lower in the brains of Alzheimer's disease patients as compared to controls (Sontag et al., 1996; Vogelsberg-Ragaglia et al., 2001). Inhibition of PP2A by inhibitor I of PP2A (I_1^{PP2A}) results in deficits in exploratory activity, spatial reference memory, and memory consolidation in adult rats, which leads to hyperphosphorylation of tau, neurodegeneration, and cognitive impairment in rats (Wang et al., 2015).

5.3.2.13 Prkacb (Protein kinase, cAMP-dependent, catalytic, beta, also known as cbeta)

Prkacb (*Cbeta*) is a catalytic beta subunit of cAMP-dependent PKA. PKA mediates the required transcriptional events by phosphorylating transcription factors such as CREB. PKA plays a major role in long-term changes in synaptic strength in the brain (Nguyen and Woo, 2003) and has been well known for its critical role in learning and memory formation (Waltereit and Weller, 2003). In mouse, PKA *Cbeta* subunit gene *Prkacb* gives rise to several splice variants that are specifically expressed in discrete regions of the brain. A mutation in mouse *Cbeta* specifically targeting the *Cbeta1*-subunit isoform was studied (Qi et al., 1996). Homozygous mutants showed normal viability and no obvious pathological defects, despite a complete lack of *Cbeta1*. However, these mutant mice demonstrated impaired synaptic transmission in the Schaffer collateral-CA1 pathway of the hippocampus. The authors provided direct genetic evidence that the *Cbeta1* isoform is required for long-term depression and depotentiation, as well as the late phase of long-term potentiation in the Schaffer collateral-CA1 pathway. Others also reported similar role for *Cbeta* in memory formation in mice (Howe et al., 2002). These findings are in-line with the results of the current analysis. The *Cbeta* gene was down-regulated in the aged rats compared to the young with an effect size of -0.1214 and p-value of 0.59 (Appendix 6.10.1 and Appendix 6.10.13). Findings from this WGCNA analysis indicate an involvement of *Prkacb* gene in age associated spatial learning and memory impairment. *Prkacb* becomes a new spatial learning candidate gene that requires further investigations.

5.3.2.14 Pten (Phosphatase and tensin homolog)

PTEN modulates activation of the PI3K/Akt pathway. Specifically, PTEN inhibits downstream activation of the PI3K pathway (Maehama and Dixon, 1998). Therefore, deletion of the *Pten* gene results in hyperactivation of the PI3K signaling pathway, which then leads to increased activation of the downstream effectors such as Akt. *Pten* gene

was down-regulated in the aged rats compared to the young with an effect size of -0.37 and p-value of 0.01 (Appendix 6.10.1).

PTEN is highly expressed in neurons and several lines of evidence support a role for PTEN in regulating important neuronal functions (Blair and Harvey, 2012). For example, familial mutations that result in PTEN inactivation have been linked to neurological disorders such as ataxia, mental retardation and seizures (Backman et al., 2001). Moreover, loss of PTEN function at early stages of development results in widespread deficits in neuronal growth, synaptogenesis, and synaptic plasticity suggesting additional roles for PTEN in these processes. A recent study demonstrated that the structural and functional properties of hippocampal synapses are independently controlled by PTEN, and PTEN plays a direct role in activity-dependent hippocampal synaptic plasticity, namely LTP and LTD (Sperow et al., 2012). In this study, deficits in both hippocampal LTP and LTD were observed in PTEN knockout (PTEN−/−) mice. Postnatal deletion of PTEN also resulted in hippocampal-specific memory deficits in these mice as significant impairments in spatial memory tasks performed in the Morris water maze were observed. Deletion of PTEN can also result in deficits in contextual learning and trace fear conditioning (Lugo et al., 2013).

5.3.2.15 Rasgrp1 (RAS guanyl releasing protein 1 (calcium and DAG-regulated))

Rasgrp1 is a member of a family of four GEFs (guanine nucleotide-exchange factors) (Ras GEFs). *RasGRP1* possesses a catalytic region consisting of a REM (Ras exchange motif) and a CDC25 (cell division cycle 25) domain. *RasGRP1* also possesses a DAG – binding C1 domain and a pair of EF hands that bind calcium. Ras proteins cycle between GDP-bound ‘off’ and GTP-bound ‘on’ states and serve to link membrane receptor signals to internal effector pathways. *Rasgrp1* is activated by Ca2+/calmodulin and DAG, and promotes the dissociation of GDP from Ras family proteins, which facilitates the exchange of GDP for GTP, and thus enhances the activity of Ras family proteins (Stone, 2006).

Rasgrp1 was down-regulated in the aged with an effect size of -0.50 and p-value of 0.23 in 4 of the 5 studies assessed in this meta-analysis (Appendix 6.10.1 and Appendix 6.10.14). In a recent expression array analysis, *Rasgrp1* had exhibited significant expression changes in the CA3 region of the hippocampus. Particularly, this gene was up-regulated in the partial learning-activated group compared to controls (Haberman et al., 2008). Ras family proteins play important roles in mediating cell proliferation, differentiation, and survival during development. Recently, a growing body of evidence suggests that they are also critically engaged in memory formation and can modify neuronal function and structure, leading to changes in synaptic strength and neuronal firing rates (Ye and Carew, 2010). The best-characterized downstream signaling cascade of Ras family proteins is the mitogen-activated protein kinase (MAPK) cascade, mainly extracellular signal-regulated kinase 1/2 (ERK) of the MAPK family has been implicated in the formation of enduring memory, but is not required for short-term memory (Adams and Sweatt, 2002; Sharma et al., 2003). Interestingly, Mapk1 (ERK) is also another candidate hub gene, which was down-regulated in the aged rats in the meta-analysis. Indeed, *Rasgrp1* may be a novel link between molecules activated in behavioral paradigms such as phospholipase C and the well-known Ras–MAPK pathway (Buckley and Caldwell, 2004).

5.3.2.16 Scn2b (Sodium channel, voltage-gated, type II, beta)

Sodium channels are complex glycoproteins comprised of an alpha subunit and often one to several beta subunits (Johnson et al., 2007). *Scn2b* encodes the beta 2 subunit of the type II voltage-gated sodium channel. Though this gene was reported to have a role in epilepsy (Baum et al., 2014), its role in brain aging and memory impairment is largely unknown. In a recent study XiYang and colleagues (2015) observed that the mRNA and protein expressions of *Scn2b* were up-regulated in the prefrontal cortex in SAMP8 (senescence-accelerated mice prone 8) mice at 8 months of age. At this stage these mice also generally show impaired learning and memory functions in the Morris water maze test. These authors also observed that in SCN2B knockdown mice a down-

regulation of SCN2B level by about 60% resulted in improvement in the hippocampus-dependent spatial recognition memory and LTP. In addition, SCN2B down-regulation was associated with up-regulation of COX5A and BDNF as well as downregulation of FGF-2. They suggested that SCN2B could play an important role in the aging-related cognitive deterioration. In my meta-analysis, *Scn2*b was found to be slightly down-regulated in the aged (effect size = -0.275, p-value = 0.07) in all 5 studies compared to the young rats (Appendix 6.10.1 and Appendix 6.10.15), which is somewhat opposite to the findings by XiYian group. Interestingly, Lu et al. (2004) reported a down-regulation of SCN2B in the aging human prefrontal cortex. The difference in reported functions for this gene could most likely be related to species and/or the strain of animal used. However, the role of this gene in learning and memory impairment needs further investigation, particularly, as it is known to influence several other genes in modulating synaptic plasticity.

5.3.2.17 Stxbp1 (syntaxin binding protein 1)

Stxbp1, also known as *Munc18-1*, is down-regulated in the aged rats (effect size = -0.32, p-value = 0.07) in my meta-analysis, that demonstrated spatial learning impairment compared to the young (Appendix 6.10.1 and Appendix 6.10.16). *Stxbp1* plays a role in release of neurotransmitters via regulation of syntaxin, a transmembrane attachment protein receptor (Kurps and de Wit, 2012).

The role of *Stxbp1* in learning and memory impairment is not known decisively. Cao et.al. (2012) reported that increased hippocampal SNAP-25 and *Munc18-1* positively correlated with spatial learning decline and might be involved in the age-related impairment of spatial learning and memory in Kunming mice. However, Dachtler et al. (2014) reported that *Stxbp1* was significantly decreased in the hippocampus of *Nrxn2α* (neurexin 2 alpha) KO mice, which exhibit deficits in sociability and social memory in relation to autism. This decreased expression of *Stxbp1* is suggestive of deficiencies in presynaptic vesicular release, which may potentially contribute to the altered behavioral state of *Nrxn2α* KO mice. Loss of *Nrxn2α* has been argued to have a causal role in the

genesis of autism-related behaviors in mice. However, they noted normal cognitive performance in these mice in hippocampus-dependent step-through passive avoidance tests (Dachtler et al., 2014). They also found significantly decreased expression of several other synaptic proteins including *Dlg4* (PSD-95) in *Nrxn2α* KO mice. *Munc18-1* has been shown to interact presynaptically with neurexins to facilitate presynaptic vesicular release and may be critical for important neurotransmitter release (Rizo and Sudhof, 2002). In addition, A 21% decrease in the abundance of *Munc18-1* in the brain has previously been found in *Nlgn1* (neuroligin 1) KO mice that display impaired spatial memory and increased repetitive behavior (Blundell et al., 2010).

5.4 Differential expression vs. differential co-expression vs. differential connectivity

Differential co-expression refers to changes in gene-gene correlations between two sets of phenotypically distinct samples (de la Fuente, 2010). Changes in gene-gene correlation may occur in the absence of differential expression, meaning that a gene may undergo changes in regulatory pattern that would be undetected by traditional differential expression analysis (Gaiteri et al., 2014). The fact that the altered regulatory patterns observed within tissues across phenotypic states in manners that are reflected in altered co-expression networks has been shown in aging mice (Southworth et al., 2009), across corticolimbic regions in major depression (Gaiteri et al., 2010) and between miRNA's in Alzheimer's disease (Bhattacharyya and Bandyopadhyay, 2013).

In light of the discussions in the previous sections, what becomes apparent is that the differential expression and differential co-expression analysis resulting from this research may be pointing to distinct cellular mechanisms involved in ASLI, which are working at different levels in the cell. For example, differential expression meta-analysis has identified a large number of genes showing significantly altered expression in the aged rats compared to young rats (Tables S1 and S2 in (Uddin and Singh, 2013)). These genes include many immediate early (e.g. *Arc*) or late phase genes (during gene expression) as well as other genes contributing to aging and ASLI as GA, GASI, and GANSI

genes. Major functions disrupted by these genes include cell viability, axonogenesis, quantity and synthesis of IP3, and formation of cells.

On the other hand differential co-expression analysis has identified a set of modules each with distinct functions. In addition, it has identified a set of candidate ASLI hub genes in one of those modules. From the known function of these hub genes (Section 5.3.1) it is evident that many of these genes function as kinases and phosphatases in the neuronal information flow process, starting from the synaptic junctions/synapses to the nucleus to activate various transcription factors. Though scattered in different networks, meta-analysis has also identified few hub genes functioning as kinases or in ion channels. Thus the hub genes may be triggering one or more mechanisms that activate other key factors in a number of pathways, which set the stage for the expression of several immediate early or late phase genes, which again most likely activate the expression of majority of the differentially expressed genes. Learning in the young animals most likely induces such mechanisms that synchronously regulate transcription of multiple genes, and may potentially generate co-expression relationships.

Another important observation to note is that all the learning related genes identified in the differential expression and IPA analyses (and genes they generally interact with) are scattered in different networks and pathways (Appendix 6.3.1 to Appendix 6.4.4). In contrast, differential co-expression analysis identified many known learning genes (or genes that appear to be contributing to learning and memory functioning) that are highly concentrated and co-expressed in the yellow “learning and memory” module.

Interestingly, the candidate ASLI hub genes are expressed at a comparatively lower level, with small differences in expression (e.g. effect size) between young and aged samples. For example, the *Prkacb* hub gene was not known to be a learning gene (IPA) (Tables S3 and S4 in (Uddin and Singh, 2013)). Like *Prkacb*, most of the hub genes failed to show significant effect size or differential expression values and remained undetected in the meta-analysis (Tables S1 and S2 in (Uddin and Singh, 2013)). This fact highlights the importance of alternate analysis like WGCNA to identify genes that are not detected

using the traditional methods. Similar observations have been demonstrated in past studies (Rhinn et al., 2012). For example, the alpha synuclein gene variant “aSynL”, containing a long 3’sUTR, was identified as the most differentially coexpressed gene in several Parkinson’s disease datasets. However, aSynL was not highly differentially expressed and thus would have likely been overlooked by traditional microarray analysis (Gaiteri et al., 2014). Thus, through the identification of modules, hubs, and differential co-expression analysis, WGCNA can be used to prioritize specific phenotype-related important molecules.

Another very interesting property of co-expression networks is the network connectivity. My findings (Appendix 6.14.1) support the newly emerging hypothesis (Miller et al., 2008; Oldham et al., 2006) that differential connectivity is different from differential expression. During the WGCNA network construction process, I selected genes with high connectivity and filtered out all low connectivity genes (Table 4.5). The observation is that the resulting network modules represent a set of highly connected genes as hubs that were virtually absent in the differentially expressed top gene list and vice versa. In fact, it has been reported that gene-gene correlations in disease can occur with or without changes in expression (Hudson et al., 2009). In addition, differentially expressed genes in some complex psychiatric diseases can have low connectivity, which reside on the periphery of co-expression networks for neuropsychiatric disorders such as depression, schizophrenia, and bipolar disorder (Gaiteri and Sibille, 2011; Gaiteri et al., 2014).

5.5 New insight into the molecular mechanisms of learning and memory formation

Here I will discuss what is already known from the literature about the molecular mechanisms of learning and memory formation and how the candidate ASLI hub genes from this research fit into that scenario.

Several major signaling pathways seem to modulate synaptic plasticity mechanisms in the brain and have been implicated in learning and memory formation processes (Baudry et al., 2014; Nguyen and Woo, 2003; Sweatt, 2001; Ye and Carew, 2010). Some of the major pathways relevant to this study include the PKA, CaMKs, MAPK, and PI3K/Akt pathways that have been implicated in LTP formation. LTP is a synaptic plasticity mechanism and a cellular correlates thought to underlie learning and memory. Following external stimulation, a set of crucial upstream events are necessary for their activation, which include NMDA receptors and the resulting calcium influx.

Calcium-dependent phosphorylation of CREB is primarily caused by PKA, CaMK and MAP kinase, which leads to prolonged CREB phosphorylation. CREB in turn contributes to the transcription of a set of immediate early genes implicated in learning and memory formation. CREB is thought to mediate long-lasting changes in brain function. For example, CREB has been implicated in spatial learning, behavioral sensitization, long-term memory of odorant-conditioned behavior, and long-term synaptic plasticity (Alberini, 2009; Chen et al., 2010; Sweatt, 2010; Thomas and Huganir, 2004). The ASLI candidate hub genes that are important in the CREB related pathways include *Camk1g*, *Dlg3*, *Dlgap1*, *Dpp6*, *Kcnab2*, *Mapk1*, and *Stxbp1*. For example, *Stxbp1* plays a role in releasing of neurotransmitters via regulation of syntaxin (Section 5.3.1.17) and may serve to transfer of signal through the synapse. *Dlg3* binds directly to NMDA receptors, anchors receptors at synapses, and facilitates transduction of NMDAR signals (Section 5.3.1.4). CaMKs, particularly CaMKII has been shown to be directly activated by calcium influx through the NMDA receptor. *Camk1g* may function in this CaMK pathway (as discussed in Section 5.3.1.1) to modulate CREB phosphorylation. *Camk1g* co-expression with other learning genes such as *Mapk1* (Section 5.3.1.9), *Kcnab2* (Section 5.3.1.8), and *Dpp6* (Section *5.3.1.5*), functioning in the MAPK pathway or in various ion channels indicate a potential co-functioning of these genes towards learning and memory formation. Some may involve a feed-back loop type activation/mechanism. For example, during the early phase of LTP at postsynaptic terminals of CA1 hippocampal neurons, calcium entering through AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic

acid) and NMDA receptors activates CaMKII, which phosphorylates Kv channels and increases neuronal excitability (Sweatt, 2001). Similarly, *Mapk1*, stimulated by elevated levels of cAMP as a result of calcium entry and subsequent activation of adenylyl cyclase-1, phosphorylates the A-type potassium channel (Kv1.4 and Kv4.2) resulting in increased depolarization, allowing influx of Ca^{2+} through the NMDA and voltage-gated Ca^{2+} channels, which results in increased cAMP levels in the hippocampus in mice. The increase in Ca^{2+} and cAMP induces the MAPK pathway. Thus, the induced pathway activates additional pools of MAPK1, some of which can further increase phosphorylation of Kv1.4 and Kv4.2, whereas others may phosphorylate nuclear targets. Voltage-gated potassium (Kv) channels play important roles in regulating the excitability of neurons and other excitable cells. Subthreshold activating, rapidly inactivating, A-type K^+ currents are non-uniformly expressed in the primary apical dendrites of rat hippocampal CA1 pyramidal neurons, with density increasing with distance from the soma (Hoffman et al., 1997). These changes correlate with impaired spatial memory and context discrimination (Morozov et al., 2003). Note that the ASLI candidate genes *Kcnab2* gene encodes one of the beta subunits of the Kv channels (Kv1.1 to Kv1.8) (Section 5.3.1.8). And the role of *Mapk1* through MAPK (ERK) signaling is not only documented in LTP, but also in spatial learning (Section 5.3.1.9). DPP6 may take part by regulating the A-type K+ current gradient, ultimately contributing to synaptic integration and dendritic excitability (Section 5.3.1.5). The action potential firing and dendritic excitability must be balanced by inhibition in hippocampal neuron. This is likely achieved by *Gabrg1* and a number of other GABA receptors that demonstrated co-expression in the yellow module (Section 5.3.1.7).

Dendritic integration of synaptic inputs is fundamental to information processing in neurons of diverse function, serving as a link between synaptic molecular pathways and higher-order network function (Sun et al., 2011). Dendritic ion channels play a critical role in regulating such information processing and are targets for modulation during synaptic plasticity (Shah et al., 2010). Normal experience-dependent changes in the excitability of dendrites (dendritic plasticity), involving the down-regulation of A-type K^+

currents by down-regulation of *Dpp6* (observed here), may represent a mechanism by which neurons store recent experience in individual dendritic branches (Makara et al., 2009). Down-regulation of *Kcnab2* may contribute to the reduction of A-type potassium channel currents through reduced availability of Kv1.4. Future studies are required to investigate the effect of *Dpp6* and *Kcnab2* in synaptic development and spatial memory formation.

Prkacb, a new ASLI candidate in the PKA pathway (Section 5.3.1.13), once activated by a variety of upstream signals, including calcium, can phosphorylate and regulate a variety of downstream signaling cascades linked to regulation of transcription and translation (Baudry et al., 2014). It can phosphorylate AMPA and NMDA receptors and regulate their functions.

Another pathway that is making itself relevant in this big picture is the PI3K/Akt pathway. A set of genes involved here include the ASLI candidate genes *Ndfip2*, *Pten*, and *Rasgrp1*. In the brain, tyrosine kinase receptor TrkA is phosphorylated on the plasma membrane by the binding of another growth factor NGF, which later activates three major signaling pathways: the PI 3 kinase pathway leading to activation of Akt kinase, the ras pathway leading to MAP kinases, and the PLC pathway leading to release of intracellular Ca2+ and activation of PKC (Purves et al., 2004). *Ndfip2* affect tyrosine kinase signaling pathway through Nedd4 ligases, which associate with EGF receptor and *Pten* (Section 5.3.1.11). Based on literature information it can be hypothesized that *Ndfip2* may modulate the EGF signaling cascade; it is possible that *Ndfip2* might be working in the same fashion as NGF in the brain to influence not only Akt kinase pathway through Akt, but also other pathways such ras, MAPK, and PLC. In fact, EGF and NGF share the same Raf → MEK → MAPK pathway to promote distinct outcomes (Vaudry et al., 2002). Therefore, the role of *Ndfip2* in learning and memory can be investigated in a future experiment.

MAPKs are normally inactive in neurons but become activated when they are phosphorylated by other kinases. In fact, MAPKs are part of a kinase cascade in which

one protein kinase phosphorylates and activates the next protein kinase in the cascade (Purves et al., 2004). The extracellular signals that trigger these kinase cascades are often extracellular growth factors that bind to receptor tyrosine kinases that, in turn, activate monomeric G-proteins such as Ras. *Rasgrp1*, once activated by Ca2+/calmodulin and DAG, facilitates the exchange of GDP for GTP and may trigger downstream *Mapk1* signaling (Section 5.3.1.15). Once activated, MAPKs can phosphorylate transcription factors, proteins that regulate gene expression.

Although, *Pten* is known to play a direct role in regulating hippocampal synaptic plasticity (Section 5.3.1.14), the precise mechanisms underlying *Pten* modulation of synaptic plasticity such as LTP and LTD are not fully known. Recent studies suggest its involvement in postsynaptic mechanism as PTEN inhibition promotes AMPA receptor trafficking to synapses leading to a persistent increase in excitatory synaptic strength in adult hippocampal slices (Moult et al., 2010). On the other hand, enhanced PTEN lipid phosphatase activity has been reported to depress excitatory synaptic transmission, which in turn is required for NMDA receptor-dependent LTD (Jurado et al., 2010). In light of this research, *Pten* is an excellent candidate to study further for it potential involvement in ASLI and the mechanisms in play.

Co-expression of genes like *Cntn1*, *Mapre1*, etc., which have known functions in neuronal structure, indicates that these genes play an essential role in learning and memory along with other genes discussed above. For example, *Mapre1* is well known to regulate microtubule dynamics (Section 5.3.1.10) and *Cntn1* is necessary for myelin sheath formation by oligodendrocytes and provides critical signal in axon-glia communication. *Ppp2r2c*, another new ASLI candidate gene forms a part of PP2A, which catalyzes a broad range of substrates (Section 5.3.1.12).

Taken together, this research has identified a set of candidate hub genes that all co-express together in a single gene network module. These genes are known to participate in multiple different cellular signaling pathways such as PKA, MapK, and CamK as discussed above. Overall, reversible phosphorylation of proteins by kinase and

phosphatase enzymes constitutes some major forms of signaling (Backx et al., 2010). These different signaling cascades converge on a common set of mechanisms: 1) post-translational protein modifications, 2) translational regulation, and 3) regulation of gene expression (Baudry et al., 2014; Purves et al., 2004; Sweatt, 2010). Ultimately, these mechanisms are linked to a few of the common events responsible for LTP such as increased number of postsynaptic receptors, and increased dendritic spines. In fact, these mechanisms are not isolated; rather, multiple cross-talk between the signaling pathways exist, which suggests that depending on the conditions, various form of LTP or LTD can be triggered with different features (Middei et al., 2014)(Baudry etal 2014). Thus, the signaling pathways are involved in the mechanism of synaptic plasticity, which in turn is the molecular mechanism for learning and memory. (Barco et al., 2006; Chen et al., 2010; Sweatt, 2001). Thus, co-expression of the hub genes along with other genes in the yellow module seems to be leading to a common function in the hippocampus in the brain, which in this case is ASLI. Results from the meta-analysis for these genes strengthen this conclusion. Down-regulation of the majority of the hub genes in the aged rats (Figure 4.15) may play a critical role in the spatial learning impairment in the Morris water maze protocol. Interestingly, many of the hub genes' individual expression patterns follow what is reported in the literature in respect to their potential role in aging associated learning and memory impairment, for example *Camk1g*, *Dlg3*, *Dpp6*, *Mapk1*, *Mapre1*, *Ndfip2*, *Ppp2r2c*, *Pten*, *Prkacb*, and *Rasgrp1*. Some other hub genes such as *Cdk5r1*, *Cntn1*, *Impact*, *Kcnab2*, *Scn2b*, and *Stxbp1* may have more indirect role. The main function of this second category of genes may involve contributing to the regulation of normal neuronal structure and functions, dysregulation of which become vulnerable at old age, and thus may indirectly contribute to the overall instability of the memory formation mechanism.

In this research, the findings of a specific "learning and memory" module and the associated key hub genes with their known role in learning and memory formation offer a promising insight and a plausible logical expansion to our existing knowledge about

the molecular correlates of the mechanisms underlying memory formation and synaptic plasticity.

5.6 Study strength and limitations

The strength of this research lies in the choice of scientifically sound techniques and approaches that were widely used by the research community. Since impurities in data generally carry over and bias any subsequent analysis, I started with raw data and placed high priority in the quality control, preprocessing, and selection of data for meta- and network analysis. Additionally, I have performed step-wise outlier removal and strict batch correction to make sure no spurious clusters were generated from a single dataset, which might lead to spurious modules. It is often common for meta-analysis to include large number of studies. Several studies using WGCNA (Oldham et al., 2008) integrate data from a variety of tissue type, experiments, and even species, which helped find genes implicating in broad categories of phenotypic differences. However, I followed a set of conservative data selection criteria as I sought to identify genes and networks in a very specific phenotype (e.g. ASLI). I have tried to maintain data sample size as large as possible for network analysis. This was challenging, particularly, during the batch effect correction for B7 and B8 because of the presence of few poor quality arrays in those dataset. The most important fact to consider during batch effect correction is to make sure that every phenotypic group (e.g. impaired, unimpaired, various control groups) is represented in every batch (Johnson et al., 2007; Leek et al., 2010). After removing unsuitable/ poor quality arrays, and after preprocessing, B7 young, K9 aged, and both BL young and aged groups were remained with 12 samples or less, these were excluded. In order to achieve the best quality results in gene network / WGCNA analysis all young or aged groups contained a minimum of 18 samples. This is in line with previous correlation network studies using WGCNA (Miller et al., 2008; Oldham et al., 2006) where 18 to 20 samples were successfully used between control and test. Moreover, constructing networks from samples from mice of common genetic background allows co-expression networks to be constructed with fewer samples

(Gaiteri et al., 2014), which was also true in my research as all rats used were from Fischer 344 strain. Overall, the methods adopted provided step-wise processes to prepare data for downstream analyses. Undertaking of these processes resulted in a logical outcome in the form of verifiable results both in meta- and network analysis.

One major limitation with the data was that extra, array preparation information (e.g. RNA isolation, reagents, array hybridization, etc.) was not available, so they could not be used to further correct additional batch effect that may be have remained for B7 and B8 data. Similarly, detailed experimental/phenotypic information was also not available. So, they could not be used to associate co-expression modules with disease or other phenotypic traits. Often, in a co-expression network analysis, module membership can be compared between cases and controls, among different tissues, species, or other phenotypes or clinical traits (de Jong et al., 2010; Fuller et al., 2007; Ghazalpour et al., 2006; Plaisier et al., 2009).

One major challenge when combining data across microarray studies and platforms is to handle missing probe sets. The same probe set may not be present in different arrays, may get filtered out during the preprocessing steps, or may be excluded in the later part of an analysis due to lack of annotation or low connectivity. Therefore, to minimize the effect of absence of a probe set in a dataset, I worked at the probe set level, and recorded the number of studies each probe set was present. Random effect size meta-analysis model with inverse variance technique was the perfect choice in this situation because this model considers sample sizes, number of studies, within-study and between-study variability. Compared to average microarray results, this overall approach resulted in a large number of significant differentially expressed genes between young and aged samples. In addition, forest plots provide an efficient way to view data size, within study variations, data strength, and heterogeneity.

For gene network modeling, the use of WGCNA produced satisfactory and promising results. The strength of WGCN lies in its simplicity and ability to model gene co-expression in the form of modules and hubs, which showed biologically meaningful

functions and reproducibility in independent datasets. This has made the WGCNA approach very popular in recent years among general biologists with little computer background (e.g. there are now over 50 applied research article using WGCNA); while many other methods (such as Bayesian networks and ARACNE) are mostly confined to the computational scientist community and are in active development. Unfortunately, the major limitation of WGCNA is that it can't distinguish direct regulatory relationship from indirect based on gene expression data alone. However, a literature review suggests that in the yellow "learning" module, hub genes *Dpp6* and *Kcnab2* may have a more direct regulatory relationship, while *Ndfip2* and *Pten* may have an indirect regulatory relationship. This indicates that additional knowledge bases can aid in characterizing close regulatory relationship among gene members in a module. Despite this disadvantage, use of WGCNA in this research was successful and identified one specific module and a set of hub genes that showed differential co-expression between the young vs. aged networks, which may play a key role in learning impairments in the aged, compared to the young rats.

5.7 Future directions

The candidate ASLI genes (including hub genes) and gene networks identified in this research through meta- and network analysis become excellent candidates for further investigations. Particularly, the hub genes can provide a different perspective on gene regulation as they can serve as excellent targets to examine the biological significance of a network module. They could be targeted to see not only a perturbation effect of altered regulation on network module structure and function, but for therapeutic use as well. Co-expression modules are not in fact completely modular as there are often correlations among the members of different modules (Gaiteri et al., 2014). Therefore, any perturbation effect will likely extend outside of a module and will need to be studied. Since, differential co-expression is likely related to altered gene regulation, experiments involving ChIP, or ChIP-seq of potential transcription factors, can be designed to capture related gene regulatory mechanisms after any perturbation.

Epigenetic mechanism are also intimately involved during the gene expression process in learning and memory formation (Franklin and Mansuy, 2010; Graff and Mansuy, 2008; Levenson and Sweatt, 2005; Levenson and Sweatt, 2006; Sweatt, 2010). So, changes in chromatin structure, methylation and acetylation pattern, as well as miRNA population changes should also be investigated.

For the purpose of future investigation, the candidate ASLI hub genes could be grouped into three categories: 1) Hub genes whose role in learning (including spatial learning) is more transparent than others (i.e. gene with well-established roles in memory, for example, *Camk1g*, *Dlg3*, *Mapk1*, *Ppp2r2c*, and *Prkacb*), 2) Hub genes (e.g. *Cdk5r1*, *Cntn1*, *Scn2b*, *Stxbp1*, *Eif5*, and *Gabrg1*) where there is not enough information in the literature to support which direction their expression pattern contributes to the ASLI phenotype, and 3) Hub genes where information is emerging indicating their direct or indirect role in learning and memory (e.g. *Pten*, *Kcnab2*, *Mapre1*, *Ndfip1*, *Rasgrp1*, and *Dpp6*).

One way to learn the specific effects of hub genes is through knockout experiments. This is because the hub genes are likely to act as drivers of the disease status due to their key positions in the gene networks (Allen et al., 2012). It is known that transmission of signal through scale-free cellular networks is unlikely to be affected by random node deletion; rather it is especially vulnerable to targeted hub attack (Albert et al., 2000). This observation is supported by examples from multiple molecular and brain networks in which hub targeting leads to crucial functional impairment (Stam et al., 2007). Practically, hub genes have been the specific focus for investigations into many disease-correlated modules (Maschietto et al., 2015; Miller et al., 2008; Ray et al., 2008; Torkamani et al., 2010; Voineagu et al., 2011; Ye and Liu, 2015). Analysis of hub genes has been shown to be a promising approach in identifying key genes in many other phenotypic conditions (Holtman et al., 2015; Kendall et al., 2005; Mani et al., 2008; Nibbe et al., 2010; Rickabaugh et al., 2015; Slavov and Dawson, 2009; Spiers et al., 2015; Zhou et al., 2014). Such genes are often of biological interest because of their critical involvement in regulatory pathways or sub-networks and these genes often incur a

substantial effect on the pathways as a whole. The candidate ASLI hub genes identified in this research may very likely present a snapshot of what is going on inside brain cells during the memory formation process.

5.8 Conclusions

Despite significant research in the past, ASLI genes and networks remain largely unclear and were the main focus of this dissertation. The major goal of this research was to combine gene expression data from multiple independent but related studies and identify genes and gene networks in ASLI in rats. During the data collection and selection process, I learned that even though there were many microarray studies related to cognitive impairments, they actually varied in terms of major study goal, selection of animal model, and the assessment of learning impairment. By following a more conservative data selection approach I was able to select five ASLI related datasets. A detailed inspection of data quality revealed the presence of imperfections in some arrays as well as the presence of outlier arrays and batch effects. By applying appropriate methods, I satisfactorily removed unsuitable arrays and corrected batch effects, and prepared all five datasets to combine at the probe set level. My research supports previous findings and emphasizes that proper data quality control and preprocessing are essential when combining data from several studies in a meta- or network analysis.

In order to integrate the selected ASLI datasets, I adopted the random effect size meta-analysis method in this research. The goal was to identify and characterize genes that may be involved in ASLI, as well as to identify and characterize gene networks based on existing biological knowledge. I implemented a probe set level data integration method, which prevented loss of information from data. The results show that a large number of genes are differentially expressed across age and across spatial learning impairment between young and aged rats. I attribute this to the proper preprocessing, data integration, and meta-analysis methods that were applied to the gene expression data. This meta-analysis allowed the identification of pertinent lists of aging and ASLI related

genes. GO and pathway analysis results relating to these genes support the fact that the genes and pathways identified in this analysis follow biological expectations. Further, the follow up analysis offered a novel insight into the underlying molecular pathways associated with aging and age-related non-syndromic memory impairments such as ASLI. The results indicate that the aged animals display a significant decrease in cell viability, axonogenesis, and inositol phosphate metabolism. Based on the known function of the significant genes, they logically fall into three major categories such as GA, GASI, and GANSI. The GA genes are mostly involved in aging related processes and generally are not associated with any learning impairment. The GASI genes, on the other hand, are associated with age-related neurological disease syndromes that generally affect normal cognitive functioning and hence may result into syndromic memory impairments. The most interesting group of genes are the GANSI genes, most of which show down-regulation in the aged or aged-impaired rats and by themselves usually are not associated with any syndromes. I report that altered expression of the GANSI category of genes affects major pathways and functions at old age, and may play a significant role in ASLI in rats. These genes affect various signal transduction pathways and functions in the brain such as molecular transport, cell to cell signaling and interaction, and nervous system development and function ultimately contributing to the disruption of proper learning and memory formation processes. I identified a set of these GANSI genes, which include some genes that express at a low level and appear as potential hub genes in the knowledge based AY or IU gene networks (Appendix 6.3.1 to Appendix 6.4.4). I propose that the selected GANSI genes should form the foundation of future studies in understanding age-associated memory impairments such as ASLI.

One of the limitations in the traditional meta- and network analysis is that gene networks and regulatory interactions among the genes in these networks are modeled based on current biological knowledge only. Another limitation is that they are not able to identify a single network that could be solely associated with ASLI (as I found that the candidate ASLI genes were all scattered in different networks). There is no prioritization of molecules within the knowledge-based network models of affected pathways.

Moreover, the traditional methods are unable to better utilize all the information that is contained within the microarray data. In order to overcome these limitations in traditional meta- and pathway analysis, I explored the option of using a mathematical modeling approach that could better utilize the information captured in microarray data. I chose to use WGCNA, applied it on a set of R7 exploratory datasets, and identified a set of gene network modules. To my satisfaction, WGCNA offered a way of prioritizing the molecules solely based on data and without any knowledge of their functions (i.e. by grouping genes into co-expressing network modules). This finding was confirmed by the follow up GO analysis which showed that each module is highly enriched with genes functioning in some broad but distinct GO functional categories or biological pathways. Further, these modules show significant repeatability in independent young and aged validation datasets. Interestingly, this analysis identified a single learning and memory related module and within the module a set of unique hub genes related to ASLI. Though some of the significant genes identified through meta-analysis are replicated in the "learning and memory" module, but majority of the candidate ASLI hub genes from this module remained undetected by the meta- and differential expression analysis. Some of these hub genes also show significant repeatability in networks generated from independent validation datasets. These hub genes are highly co-expressed with other genes in the "learning and memory" module. In network comparison between young and aged, these genes not only show differential expression but also differential co-expression and differential connectivity. The known function of these hub genes indicate that they play key roles in critical pathways relating to synaptic plasticity and memory formation. Collectively, they provide a deeper understanding of the mechanisms that may be involved. These candidate ASLI hub genes seem highly promising to investigate further to understand the regulatory networks in ASLI.

Co-expression network analysis as applied in this research shows how to transform large-scale gene expression microarray data involving spatial learning impairment in rats into several testable hypotheses related to ASLI. This type of analysis can complement

traditional analysis of microarray data and can help better understand how genes interact with each other, how they are regulated, and what target genes they may affect in order to elucidate the mechanisms behind complex phenotype such as aging and age-associated spatial learning impairment. In closing, it is possible to extract interesting and useful information about genes and their networks in a specific biological context from large scale data using meta- and mathematical modeling approaches.

5.9 References

Adams, J.P., Sweatt, J.D., 2002. Molecular psychology: roles for the ERK MAP kinase cascade in memory. Annu Rev Pharmacol Toxicol. 42**,** 135-63.

Alberini, C.M., 2009. Transcription factors in long-term memory and synaptic plasticity. Physiol Rev. 89**,** 121-45.

Albert, R., Jeong, H., Barabasi, A.L., 2000. Error and attack tolerance of complex networks. Nature. 406**,** 378-82.

Allen, J.D., Xie, Y., Chen, M., Girard, L., et al., 2012. Comparing statistical methods for constructing large scale gene networks. PLoS One. 7**,** e29348.

Alzheimer's, A., 2012. 2012 Alzheimer's disease facts and figures. 8, 131–168 (2012). Alzheimers Dement. 8**,** 131-168.

Angelo, M., Plattner, F., Giese, K.P., 2006. Cyclin-dependent kinase 5 in synaptic plasticity, learning and memory. J Neurochem. 99**,** 353-70.

Arnold, A.C., Gallagher, P.E., Diz, D.I., 2013. Brain renin-angiotensin system in the nexus of hypertension and aging. Hypertens Res. 36**,** 5-13.

Atkins, C.M., Selcher, J.C., Petraitis, J.J., Trzaskos, J.M., et al., 1998. The MAPK cascade is required for mammalian associative learning. Nat Neurosci. 1**,** 602-9.

Backman, S.A., Stambolic, V., Suzuki, A., Haight, J., et al., 2001. Deletion of Pten in mouse brain causes seizures, ataxia and defects in soma size resembling Lhermitte-Duclos disease. Nat Genet. 29**,** 396-403.

Backx, L., Vermeesch, J., Pijkels, E., de Ravel, T., et al., 2010. PPP2R2C, a gene disrupted in autosomal dominant intellectual disability. Eur J Med Genet. 53**,** 239-43.

Barco, A., Bailey, C.H., Kandel, E.R., 2006. Common molecular mechanisms in explicit and implicit memory. J Neurochem. 97**,** 1520-33.

Bartus, K., Pigott, B., Garthwaite, J., 2013. Cellular targets of nitric oxide in the hippocampus. PLoS One. 8**,** e57292.

Baudry, M., Zhu, G., Liu, Y., Wang, Y., et al., 2014. Multiple cellular cascades participate in long-term potentiation and in hippocampus-dependent learning. Brain Res.

Baum, L., Haerian, B.S., Ng, H.K., Wong, V.C., et al., 2014. Case-control association study of polymorphisms in the voltage-gated sodium channel genes SCN1A, SCN2A, SCN3A, SCN1B, and SCN2B and epilepsy. Hum Genet. 133**,** 651-9.

Berg, M.M., Sternberg, D.W., Parada, L.F., Chao, M.V., 1992. K-252a inhibits nerve growth factor-induced trk proto-oncogene tyrosine phosphorylation and kinase activity. J Biol Chem. 267**,** 13-6.

Berridge, M.J., Bootman, M.D., Lipp, P., 1998. Calcium--a life and death signal. Nature. 395**,** 645-8.

Bhattacharyya, M., Bandyopadhyay, S., 2013. Studying the differential co-expression of microRNAs reveals significant role of white matter in early Alzheimer's progression. Mol Biosyst. 9**,** 457-66.

Bito, H., 1998. The role of calcium in activity-dependent neuronal gene regulation. Cell Calcium. 23**,** 143-50.

Bito, H., Takemoto-Kimura, S., 2003. Ca(2+)/CREB/CBP-dependent gene regulation: a shared mechanism critical in long-term synaptic plasticity and neuronal survival. Cell Calcium. 34**,** 425-30.

Blair, P.J., Harvey, J., 2012. PTEN: a new player controlling structural and functional synaptic plasticity. J Physiol. 590**,** 1017.

Blauw, H.M., Al-Chalabi, A., Andersen, P.M., van Vught, P.W., et al., 2010. A large genome scan for rare CNVs in amyotrophic lateral sclerosis. Hum Mol Genet. 19**,** 4091-9.

Bliss, T.V., Collingridge, G.L., 1993. A synaptic model of memory: long-term potentiation in the hippocampus. Nature. 361**,** 31-9.

Blum, S., Moore, A.N., Adams, F., Dash, P.K., 1999. A mitogen-activated protein kinase cascade in the CA1/CA2 subfield of the dorsal hippocampus is essential for long-term spatial memory. J Neurosci. 19**,** 3535-44.

Blundell, J., Blaiss, C.A., Etherton, M.R., Espinosa, F., et al., 2010. Neuroligin-1 deletion results in impaired spatial memory and increased repetitive behavior. J Neurosci. 30**,** 2115-29.

Borghese, F., Clanchy, F.I., 2011. CD74: an emerging opportunity as a therapeutic target in cancer and autoimmune disease. Expert Opin Ther Targets. 15**,** 237-51.

Brambilla, P., Esposito, F., Lindstrom, E., Sorosina, M., et al., 2012. Association between DPP6 polymorphism and the risk of progressive multiple sclerosis in Northern and Southern Europeans. Neurosci Lett. 530**,** 155-60.

Bramham, C.R., Alme, M.N., Bittins, M., Kuipers, S.D., et al., 2010. The Arc of synaptic memory. Exp Brain Res. 200**,** 125-40.

Broer, A., Tietze, N., Kowalczuk, S., Chubb, S., et al., 2006. The orphan transporter v7-3 (slc6a15) is a Na+-dependent neutral amino acid transporter (B0AT2). Biochem J. 393**,** 421-30.

Bu, G., 2009. Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. Nat Rev Neurosci. 10**,** 333-44.

Buckley, C.T., Caldwell, K.K., 2004. Fear conditioning is associated with altered integration of PLC and ERK signaling in the hippocampus. Pharmacol Biochem Behav. 79**,** 633-40.

Cao, L., Wang, F., Yang, Q.G., Jiang, W., et al., 2012. Reduced thyroid hormones with increased hippocampal SNAP-25 and Munc18-1 might involve cognitive impairment during aging. Behav Brain Res. 229**,** 131-7.

Carter, S.L., Brechbuhler, C.M., Griffin, M., Bond, A.T., 2004. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics. 20**,** 2242-50.
Cazander, G., Jukema, G.N., Nibbering, P.H., 2012. Complement activation and inhibition in wound healing. Clin Dev Immunol. 2012**,** 534291.
Chen, D.Y., Stern, S.A., Garcia-Osta, A., Saunier-Rebori, B., et al., 2011. A critical role for IGF-II in memory consolidation and enhancement. Nature. 469**,** 491-7.
Chen, G., Zou, X., Watanabe, H., van Deursen, J.M., et al., 2010. CREB binding protein is required for both short-term and long-term memory formation. J Neurosci. 30**,** 13066-77.
Chen, L.H., Kao, P.Y., Fan, Y.H., Ho, D.T., et al., 2012. Polymorphisms of CR1, CLU and PICALM confer susceptibility of Alzheimer's disease in a southern Chinese population. Neurobiol Aging. 33**,** 210 e1-7.
Chen, Y., Zhu, J., Lum, P.Y., Yang, X., et al., 2008. Variations in DNA elucidate molecular networks that cause disease. Nature. 452**,** 429-35.
Colakoglu, G., Bergstrom-Tyrberg, U., Berglund, E.O., Ranscht, B., 2014. Contactin-1 regulates myelination and nodal/paranodal domain organization in the central nervous system. Proc Natl Acad Sci U S A. 111**,** E394-403.
Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., et al., 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 261**,** 921-3.
Costa, R.M., Federov, N.B., Kogan, J.H., Murphy, G.G., et al., 2002. Mechanism for the learning deficits in a mouse model of neurofibromatosis type 1. Nature. 415**,** 526-30.
Cristillo, A.D., Nie, L., Macri, M.J., Bierer, B.E., 2003. Cloning and characterization of N4WBP5A, an inducible, cyclosporine-sensitive, Nedd4-binding protein in human T lymphocytes. J Biol Chem. 278**,** 34587-97.
Cuadrado, A., Nebreda, A.R., 2010. Mechanisms and functions of p38 MAPK signalling. Biochem J. 429**,** 403-17.
Cui, Y., Costa, R.M., Murphy, G.G., Elgersma, Y., et al., 2008. Neurofibromin regulation of ERK signaling modulates GABA release and learning. Cell. 135**,** 549-60.
Cushion, T.D., Dobyns, W.B., Mullins, J.G., Stoodley, N., et al., 2013. Overlapping cortical malformations and mutations in TUBB2B and TUBA1A. Brain. 136**,** 536-48.
Cuthbert, P.C., Stanford, L.E., Coba, M.P., Ainge, J.A., et al., 2007. Synapse-associated protein 102/dlgh3 couples the NMDA receptor to specific plasticity pathways and learning strategies. J Neurosci. 27**,** 2673-82.
Dachtler, J., Glasper, J., Cohen, R.N., Ivorra, J.L., et al., 2014. Deletion of alpha-neurexin II results in autism-related behaviors in mice. Transl Psychiatry. 4**,** e484.
Daniels, M.P., 2012. The role of agrin in synaptic development, plasticity and signaling in the central nervous system. Neurochem Int. 61**,** 848-53.
de Jong, S., Fuller, T.F., Janson, E., Strengman, E., et al., 2010. Gene expression profiling in C57BL/6J and A/J mouse inbred strains reveals gene networks specific for brain regions independent of genetic background. BMC Genomics. 11**,** 20.

de la Fuente, A., 2010. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet. 26**,** 326-33.

Derfuss, T., Parikh, K., Velhin, S., Braun, M., et al., 2009. Contactin-2/TAG-1-directed autoimmunity is identified in multiple sclerosis patients and mediates gray matter pathology in animals. Proc Natl Acad Sci U S A. 106**,** 8302-7.

Donnelly, N., Gorman, A.M., Gupta, S., Samali, A., 2013. The eIF2alpha kinases: their structures and functions. Cell Mol Life Sci.

Ebstein, R.P., Knafo, A., Mankuta, D., Chew, S.H., et al., 2012. The contributions of oxytocin and vasopressin pathway genes to human behavior. Horm Behav. 61**,** 359-79.

Elias, G.M., Nicoll, R.A., 2007. Synaptic trafficking of glutamate receptors by MAGUK scaffolding proteins. Trends Cell Biol. 17**,** 343-52.

Elias, G.M., Elias, L.A., Apostolides, P.F., Kriegstein, A.R., et al., 2008. Differential trafficking of AMPA and NMDA receptors by SAP102 and PSD-95 underlies synapse development. Proc Natl Acad Sci U S A. 105**,** 20953-8.

Ferrari, R., Moreno, J.H., Minhajuddin, A.T., O'Bryant, S.E., et al., 2012. Implication of common and disease specific variants in CLU, CR1, and PICALM. Neurobiol Aging. 33**,** 1846 e7-18.

Finch, E.A., Augustine, G.J., 1998. Local calcium signalling by inositol-1,4,5-trisphosphate in Purkinje cell dendrites. Nature. 396**,** 753-6.

Fischer, A., Sananbenesi, F., Schrick, C., Spiess, J., et al., 2002. Cyclin-dependent kinase 5 is required for associative learning. J Neurosci. 22**,** 3700-7.

Foster, T.C., 2012. Dissecting the age-related decline on spatial learning and memory tasks in rodent models: N-methyl-D-aspartate receptors and voltage-dependent Ca2+ channels in senescent synaptic plasticity. Prog Neurobiol. 96**,** 283-303.

Franklin, T.B., Mansuy, I.M., 2010. The prevalence of epigenetic mechanisms in the regulation of cognitive functions and behaviour. Curr Opin Neurobiol. 20**,** 441-9.

Fuller, T.F., Ghazalpour, A., Aten, J.E., Drake, T.A., et al., 2007. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome. 18**,** 463-72.

Gaiteri, C., Guilloux, J.P., Lewis, D.A., Sibille, E., 2010. Altered gene synchrony suggests a combined hormone-mediated dysregulated state in major depression. PLoS One. 5**,** e9970.

Gaiteri, C., Sibille, E., 2011. Differentially expressed genes in major depression reside on the periphery of resilient gene coexpression networks. Front Neurosci. 5**,** 95.

Gaiteri, C., Ding, Y., French, B., Tseng, G.C., et al., 2014. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav. 13**,** 13-24.

Gargalovic, P.S., Imura, M., Zhang, B., Gharavi, N.M., et al., 2006. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proc Natl Acad Sci U S A. 103**,** 12741-6.

Gay, E.A., Klein, R.C., Melton, M.A., Blackshear, P.J., et al., 2008. Inhibition of native and recombinant nicotinic acetylcholine receptors by the myristoylated alanine-rich C kinase substrate peptide. J Pharmacol Exp Ther. 327**,** 884-90.

Ge, H., Liu, Z., Church, G.M., Vidal, M., 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet. 29**,** 482-6.

Genoud, S., Pralong, W., Riederer, B.M., Eder, L., et al., 1999. Activity-dependent phosphorylation of SNAP-25 in hippocampal organotypic cultures. J Neurochem. 72**,** 1699-706.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., et al., 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet. 2**,** e130.

Giese, K.P., 2014. Generation of the Cdk5 activator p25 is a memory mechanism that is affected in early Alzheimer's disease. Front Mol Neurosci. 7**,** 36.

Graff, J., Mansuy, I.M., 2008. Epigenetic codes in cognition and behaviour. Behav Brain Res. 192**,** 70-87.

Guzowski, J.F., Lyford, G.L., Stevenson, G.D., Houston, F.P., et al., 2000. Inhibition of activity-dependent arc protein expression in the rat hippocampus impairs the maintenance of long-term potentiation and the consolidation of long-term memory. J Neurosci. 20**,** 3993-4001.

Haberman, R.P., Lee, H.J., Colantuoni, C., Koh, M.T., et al., 2008. Rapid encoding of new information alters the profile of plasticity-related mRNA transcripts in the hippocampal CA3 region. Proc Natl Acad Sci U S A. 105**,** 10601-6.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., et al., 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 430**,** 88-93.

Hautbergue, G.M., Hung, M.L., Walsh, M.J., Snijders, A.P., et al., 2009. UIF, a New mRNA export adaptor that works together with REF/ALY, requires FACT for recruitment to mRNA. Curr Biol. 19**,** 1918-24.

Hawasli, A.H., Benavides, D.R., Nguyen, C., Kansy, J.W., et al., 2007. Cyclin-dependent kinase 5 governs learning and synaptic plasticity via control of NMDAR degradation. Nat Neurosci. 10**,** 880-6.

He, H.J., Wang, Y., Le, Y., Duan, K.M., et al., 2012. Surgery upregulates high mobility group box-1 and disrupts the blood-brain barrier causing cognitive dysfunction in aged rats. CNS Neurosci Ther. 18**,** 994-1002.

He, X., Ishizeki, M., Mita, N., Wada, S., et al., 2014. Cdk5/p35 is required for motor coordination and cerebellar plasticity. J Neurochem. 131**,** 53-64.

Hoffman, D.A., Magee, J.C., Colbert, C.M., Johnston, D., 1997. K+ channel regulation of signal propagation in dendrites of hippocampal pyramidal neurons. Nature. 387**,** 869-75.

Holtman, I.R., Raj, D.D., Miller, J.A., Schaafsma, W., et al., 2015. Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis. Acta Neuropathol Commun. 3**,** 31.

Howe, D.G., Wiley, J.C., McKnight, G.S., 2002. Molecular and behavioral effects of a null mutation in all PKA C beta isoforms. Mol Cell Neurosci. 20**,** 515-24.

Hudson, N.J., Reverter, A., Dalrymple, B.P., 2009. A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. PLoS Comput Biol. 5**,** e1000382.

Iida, Y., Yamamori, S., Itakura, M., Miyaoka, H., et al., 2013. Protein phosphatase 2A dephosphorylates SNAP-25 through two distinct mechanisms in mouse brain synaptosomes. Neurosci Res.

IPA, 2013. Ingenuity Pathway Analysis (IPA) knowledge base, http://www.ingenuity.com/products/ipa. Vol., ed.^eds.

Ishii, N., Takahashi, T., Soroosh, P., Sugamura, K., 2010. OX40-OX40 ligand interaction in T-cell-mediated immunity and immunopathology. Adv Immunol. 105**,** 63-98.

Jankowski, S.A., Mitchell, D.S., Smith, S.H., Trent, J.M., et al., 1994. SAS, a gene amplified in human sarcomas, encodes a new member of the transmembrane 4 superfamily of proteins. Oncogene. 9**,** 1205-11.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., et al., 2000. The large-scale organization of metabolic networks. Nature. 407**,** 651-4.

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 8**,** 118-27.

Jurado, S., Benoist, M., Lario, A., Knafo, S., et al., 2010. PTEN is recruited to the postsynaptic terminal for NMDA receptor-dependent long-term depression. EMBO J. 29**,** 2827-40.

Kendall, S.D., Linardic, C.M., Adam, S.J., Counter, C.M., 2005. A network of genetic events sufficient to convert normal human cells to a tumorigenic state. Cancer Res. 65**,** 9824-8.

Kenwrick, S., Watkins, A., De Angelis, E., 2000. Neural cell recognition molecule L1: relating biological complexity to human disease mutations. Hum Mol Genet. 9**,** 879-86.

Kim, M.J., Yun, H.S., Hong, E.H., Lee, S.J., et al., 2013. Depletion of end-binding protein 1 (EB1) promotes apoptosis of human non-small-cell lung cancer cells via reactive oxygen species and Bax-mediated mitochondrial dysfunction. Cancer Lett. 339**,** 15-24.

Klune, J.R., Dhupar, R., Cardinal, J., Billiar, T.R., et al., 2008. HMGB1: endogenous danger signaling. Mol Med. 14**,** 476-84.

Kohli, M.A., Lucae, S., Saemann, P.G., Schmidt, M.V., et al., 2011. The neuronal transporter gene SLC6A15 confers risk to major depression. Neuron. 70**,** 252-65.

Kurps, J., de Wit, H., 2012. The role of Munc18-1 and its orthologs in modulation of cortical F-actin in chromaffin cells. J Mol Neurosci. 48**,** 339-46.

Lai, H.C., Jan, L.Y., 2006. The distribution and targeting of neuronal voltage-gated ion channels. Nat Rev Neurosci. 7**,** 548-62.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9**,** 559.

Lee, F.S., Rajagopal, R., Chao, M.V., 2002. Distinctive features of Trk neurotrophin receptor transactivation by G protein-coupled receptors. Cytokine Growth Factor Rev. 13**,** 11-7.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., et al., 2004. Coexpression analysis of human genes across many microarray data sets. Genome Res. 14**,** 1085-94.
Lee, J.M., Sonnhammer, E.L., 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13**,** 875-82.
Lee, T.W., Coates, L.C., Birch, N.P., 2008. Neuroserpin regulates N-cadherin-mediated cell adhesion independently of its activity as an inhibitor of tissue plasminogen activator. J Neurosci Res. 86**,** 1243-53.
Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., et al., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 11**,** 733-9.
Levenson, J.M., Sweatt, J.D., 2005. Epigenetic mechanisms in memory formation. Nat Rev Neurosci. 6**,** 108-18.
Levenson, J.M., Sweatt, J.D., 2006. Epigenetic mechanisms: a common theme in vertebrate and invertebrate memory formation. Cell Mol Life Sci. 63**,** 1009-16.
Levine, A.J., Miller, J.A., Shapshak, P., Gelman, B., et al., 2013. Systems analysis of human brain gene expression: mechanisms for HIV-associated neurocognitive impairment and common pathways with Alzheimer's disease. BMC Med Genomics. 6**,** 4.
Liang, W., Ouyang, S., Shaw, N., Joachimiak, A., et al., 2011. Conversion of D-ribulose 5-phosphate to D-xylulose 5-phosphate: new insights from structural and biochemical studies on human RPE. FASEB J. 25**,** 497-504.
Lin, J.F., Pan, H.C., Ma, L.P., Shen, Y.Q., et al., 2012. The cell neural adhesion molecule contactin-2 (TAG-1) is beneficial for functional recovery after spinal cord injury in adult zebrafish. PLoS One. 7**,** e52376.
Lin, L., Sun, W., Throesch, B., Kung, F., et al., 2013. DPP6 regulation of dendritic morphogenesis impacts hippocampal synaptic development. Nat Commun. 4**,** 2270.
Liu, C.C., Kanekiyo, T., Xu, H., Bu, G., 2013. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol.
Liu, J., Yue, Y., Han, D., Wang, X., et al., 2014. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nat Chem Biol. 10**,** 93-5.
Lu, T., Pan, Y., Kao, S.Y., Li, C., et al., 2004. Gene regulation and DNA damage in the ageing human brain. Nature. 429**,** 883-91.
Lugo, J.N., Smith, G.D., Morrison, J.B., White, J., 2013. Deletion of PTEN produces deficits in conditioned fear and increases fragile X mental retardation protein. Learn Mem. 20**,** 670-3.
Maeda, S., Sahara, N., Saito, Y., Murayama, S., et al., 2006. Increased levels of granular tau oligomers: an early sign of brain aging and Alzheimer's disease. Neurosci Res. 54**,** 197-201.
Maehama, T., Dixon, J.E., 1998. The tumor suppressor, PTEN/MMAC1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. J Biol Chem. 273**,** 13375-8.
Maier, M., Peng, Y., Jiang, L., Seabrook, T.J., et al., 2008. Complement C3 deficiency leads to accelerated amyloid beta plaque deposition and neurodegeneration and

modulation of the microglia/macrophage phenotype in amyloid precursor protein transgenic mice. J Neurosci. 28**,** 6333-41.
Makara, J.K., Losonczy, A., Wen, Q., Magee, J.C., 2009. Experience-dependent compartmentalized dendritic plasticity in rat hippocampal CA1 pyramidal neurons. Nat Neurosci. 12**,** 1485-7.
Maness, P.F., Schachner, M., 2007. Neural recognition molecules of the immunoglobulin superfamily: signaling transducers of axon guidance and neuronal migration. Nat Neurosci. 10**,** 19-26.
Mani, K.M., Lefebvre, C., Wang, K., Lim, W.K., et al., 2008. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol Syst Biol. 4**,** 169.
Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., et al., 2008. Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet. 82**,** 477-88.
Maschietto, M., Tahira, A.C., Puga, R., Lima, L., et al., 2015. Co-expression network of neural-differentiation genes shows specific pattern in schizophrenia. BMC Med Genomics. 8**,** 23.
McKeown, L., Swanton, L., Robinson, P., Jones, O.T., 2008. Surface expression and distribution of voltage-gated potassium channels in neurons (Review). Mol Membr Biol. 25**,** 332-43.
Menard, C., Quirion, R., 2012. Group 1 metabotropic glutamate receptor function and its regulation of learning and memory in the aging brain. Front Pharmacol. 3**,** 182.
Michalak, P., 2008. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 91**,** 243-8.
Middei, S., Ammassari-Teule, M., Marie, H., 2014. Synaptic plasticity under learning challenge. Neurobiol Learn Mem. 115**,** 108-15.
Mihelic, M., Dobersek, A., Guncar, G., Turk, D., 2008. Inhibitory fragment from the p41 form of invariant chain can regulate activity of cysteine cathepsins in antigen presentation. J Biol Chem. 283**,** 14453-60.
Miller, B.C., Eckman, E.A., Sambamurti, K., Dobbs, N., et al., 2003. Amyloid-beta peptide levels in brain are inversely correlated with insulysin activity levels in vivo. Proc Natl Acad Sci U S A. 100**,** 6221-6.
Miller, J.A., Oldham, M.C., Geschwind, D.H., 2008. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. J Neurosci. 28**,** 1410-20.
Min, J.L., Nicholson, G., Halgrimsdottir, I., Almstrup, K., et al., 2012. Coexpression network analysis in abdominal and gluteal adipose tissue reveals regulatory genetic loci for metabolic syndrome and related phenotypes. PLoS Genet. 8**,** e1002505.
Monfort, P., Munoz, M.D., Kosenko, E., Llansola, M., et al., 2004. Sequential activation of soluble guanylate cyclase, protein kinase G and cGMP-degrading phosphodiesterase is necessary for proper induction of long-term potentiation in CA1 of hippocampus. Alterations in hyperammonemia. Neurochem Int. 45**,** 895-901.

Montmayeur, J.P., Barr, T.P., Kam, S.A., Packer, S.J., et al., 2011. ET-1 induced Elevation of intracellular calcium in clonal neuronal and embryonic kidney cells involves endogenous endothelin-A receptors linked to phospholipase C through Galpha(q/11). Pharmacol Res. 64**,** 258-67.

Mooz, J., Oberoi-Khanuja, T.K., Harms, G.S., Wang, W., et al., 2014. Dimerization of the kinase ARAF promotes MAPK pathway activation and cell migration. Sci Signal. 7**,** ra73.

Morozov, A., Muzzio, I.A., Bourtchouladze, R., Van-Strien, N., et al., 2003. Rap1 couples cAMP signaling to a distinct pool of p42/44MAPK regulating excitability, synaptic plasticity, learning, and memory. Neuron. 39**,** 309-25.

Moscarello, M.A., Mastronardi, F.G., Wood, D.D., 2007. The role of citrullinated proteins suggests a novel mechanism in the pathogenesis of multiple sclerosis. Neurochem Res. 32**,** 251-6.

Moult, P.R., Cross, A., Santos, S.D., Carvalho, A.L., et al., 2010. Leptin regulates AMPA receptor trafficking via PTEN inhibition. J Neurosci. 30**,** 4088-101.

Mund, T., Pelham, H.R., 2010. Regulation of PTEN/Akt and MAP kinase signaling pathways by the ubiquitin ligase activators Ndfip1 and Ndfip2. Proc Natl Acad Sci U S A. 107**,** 11429-34.

Murata, Y., Constantine-Paton, M., 2013. Postsynaptic density scaffold SAP102 regulates cortical synapse development through EphB and PAK signaling pathway. J Neurosci. 33**,** 5040-52.

Nadal, M.S., Ozaita, A., Amarillo, Y., Vega-Saenz de Miera, E., et al., 2003. The CD26-related dipeptidyl aminopeptidase-like protein DPPX is a critical component of neuronal A-type K+ channels. Neuron. 37**,** 449-61.

Nalivaeva, N.N., Beckett, C., Belyaev, N.D., Turner, A.J., 2012. Are amyloid-degrading enzymes viable therapeutic targets in Alzheimer's disease? J Neurochem. 120 Suppl 1**,** 167-85.

Nguyen, P.V., Woo, N.H., 2003. Regulation of hippocampal synaptic plasticity by cyclic AMP-dependent protein kinases. Prog Neurobiol. 71**,** 401-37.

Nibbe, R.K., Koyuturk, M., Chance, M.R., 2010. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol. 6**,** e1000639.

O'Neill, C., Kiely, A.P., Coakley, M.F., Manning, S., et al., 2012. Insulin and IGF-1 signalling: longevity, protein homoeostasis and Alzheimer's disease. Biochem Soc Trans. 40**,** 721-7.

Obayashi, T., Hayashi, S., Shibaoka, M., Saeki, M., et al., 2008. COXPRESdb: a database of coexpressed gene networks in mammals. Nucleic Acids Res. 36**,** D77-82.

Ohshima, T., Ogura, H., Tomizawa, K., Hayashi, K., et al., 2005. Impairment of hippocampal long-term depression and defective spatial learning and memory in p35 mice. J Neurochem. 94**,** 917-25.

Oldham, M.C., Horvath, S., Geschwind, D.H., 2006. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci U S A. 103**,** 17973-8.

Oldham, M.C., Konopka, G., Iwamoto, K., Langfelder, P., et al., 2008. Functional organization of the transcriptome in human brain. Nat Neurosci. 11**,** 1271-82.
Osterwalder, T., Contartese, J., Stoeckli, E.T., Kuhn, T.B., et al., 1996. Neuroserpin, an axonally secreted serine protease inhibitor. EMBO J. 15**,** 2944-53.
Oz, S., Kapitansky, O., Ivashco-Pachima, Y., Malishkevich, A., et al., 2014. The NAP motif of activity-dependent neuroprotective protein (ADNP) regulates dendritic spines through microtubule end binding proteins. Mol Psychiatry. 19**,** 1115-24.
Park, H., Poo, M.M., 2013. Neurotrophin regulation of neural circuit development and function. Nat Rev Neurosci. 14**,** 7-23.
Perkowski, J.J., Murphy, G.G., 2011. Deletion of the mouse homolog of KCNAB2, a gene linked to monosomy 1p36, results in associative memory impairments and amygdala hyperexcitability. J Neurosci. 31**,** 46-54.
Pirker, S., Schwarzer, C., Wieselthaler, A., Sieghart, W., et al., 2000. GABA(A) receptors: immunocytochemical distribution of 13 subunits in the adult rat brain. Neuroscience. 101**,** 815-50.
Piro, R.M., Ala, U., Molineris, I., Grassi, E., et al., 2011. An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction. Eur J Hum Genet. 19**,** 1173-80.
Plaisier, C.L., Horvath, S., Huertas-Vazquez, A., Cruz-Bautista, I., et al., 2009. A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. PLoS Genet. 5**,** e1000642.
Plath, N., Ohana, O., Dammermann, B., Errington, M.L., et al., 2006. Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. Neuron. 52**,** 437-44.
Poplawski, G.H., Tranziska, A.K., Leshchyns'ka, I., Meier, I.D., et al., 2012. L1CAM increases MAP2 expression via the MAPK pathway to promote neurite outgrowth. Mol Cell Neurosci. 50**,** 169-78.
Poulin, P., Pittman, Q.J., 1993. Arginine vasopressin-induced sensitization in brain: facilitated inositol phosphate production without changes in receptor number. J Neuroendocrinol. 5**,** 23-31.
Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W.C., et al., 2004. Neuroscience, Vol., Sinauer Associates, Inc., Sunderland, MA, USA.
Qi, M., Zhuo, M., Skalhegg, B.S., Brandon, E.P., et al., 1996. Impaired hippocampal plasticity in mice lacking the Cbeta1 catalytic subunit of cAMP-dependent protein kinase. Proc Natl Acad Sci U S A. 93**,** 1571-6.
Qui, M.S., Green, S.H., 1992. PC12 cell neuronal differentiation is associated with prolonged p21ras activity and consequent prolonged ERK activity. Neuron. 9**,** 705-17.
Ragazzini, P., Gamberi, G., Benassi, M.S., Orlando, C., et al., 1999. Analysis of SAS gene and CDK4 and MDM2 proteins in low-grade osteosarcoma. Cancer Detect Prev. 23**,** 129-36.
Ramsdell, J.S., Tashjian, A.H., Jr., 1986. Thyrotropin-releasing hormone (TRH) elevation of inositol trisphosphate and cytosolic free calcium is dependent on receptor

number. Evidence for multiple rapid interactions between TRH and its receptor. J Biol Chem. 261**,** 5301-6.
Ranscht, B., 1988. Sequence of contactin, a 130-kD glycoprotein concentrated in areas of interneuronal contact, defines a new member of the immunoglobulin supergene family in the nervous system. J Cell Biol. 107**,** 1561-73.
Ray, M., Ruan, J., Zhang, W., 2008. Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases. Genome Biol. 9**,** R148.
Rhinn, H., Qiang, L., Yamashita, T., Rhee, D., et al., 2012. Alternative alpha-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology. Nat Commun. 3**,** 1084.
Rhodes, K.J., Monaghan, M.M., Barrezueta, N.X., Nawoschik, S., et al., 1996. Voltage-gated K+ channel beta subunits: expression and distribution of Kv beta 1 and Kv beta 2 in adult rat brain. J Neurosci. 16**,** 4846-60.
Rickabaugh, T.M., Baxter, R.M., Sehl, M., Sinsheimer, J.S., et al., 2015. Acceleration of age-associated methylation patterns in HIV-1-infected adults. PLoS One. 10**,** e0119201.
Rizo, J., Sudhof, T.C., 2002. Snares and Munc18 in synaptic vesicle fusion. Nat Rev Neurosci. 3**,** 641-53.
Roesler, R., Schwartsmann, G., 2012. Gastrin-releasing peptide receptors in the central nervous system: role in brain function and as a drug target. Front Endocrinol (Lausanne). 3**,** 159.
Romaniello, R., Tonelli, A., Arrigoni, F., Baschirotto, C., et al., 2012. A novel mutation in the beta-tubulin gene TUBB2B associated with complex malformation of cortical development and deficits in axonal guidance. Dev Med Child Neurol. 54**,** 765-9.
Schafe, G.E., Atkins, C.M., Swank, M.W., Bauer, E.P., et al., 2000. Activation of ERK/MAP kinase in the amygdala is required for memory consolidation of pavlovian fear conditioning. J Neurosci. 20**,** 8177-87.
Scotland, P.B., Heath, J.L., Conway, A.E., Porter, N.B., et al., 2012. The PICALM protein plays a key role in iron homeostasis and cell proliferation. PLoS One. 7**,** e44252.
Selcher, J.C., Atkins, C.M., Trzaskos, J.M., Paylor, R., et al., 1999. A necessity for MAP kinase activation in mammalian spatial learning. Learn Mem. 6**,** 478-90.
Selcher, J.C., Nekrasova, T., Paylor, R., Landreth, G.E., et al., 2001. Mice lacking the ERK1 isoform of MAP kinase are unimpaired in emotional learning. Learn Mem. 8**,** 11-9.
Seo, J., Giusti-Rodriguez, P., Zhou, Y., Rudenko, A., et al., 2014. Activity-dependent p25 generation regulates synaptic plasticity and Abeta-induced cognitive impairment. Cell. 157**,** 486-98.
Shah, K., Lahiri, D.K., 2014. Cdk5 activity in the brain - multiple paths of regulation. J Cell Sci. 127**,** 2391-400.
Shah, M.M., Hammond, R.S., Hoffman, D.A., 2010. Dendritic ion channel trafficking and plasticity. Trends Neurosci. 33**,** 307-16.

Sharma, S.K., Sherff, C.M., Shobe, J., Bagnall, M.W., et al., 2003. Differential role of mitogen-activated protein kinase in three distinct phases of memory for sensitization in Aplysia. J Neurosci. 23**,** 3899-907.
Shen, F., Li, Y.J., Shou, X.J., Cui, C.L., 2012. Role of the NO/sGC/PKG signaling pathway of hippocampal CA1 in morphine-induced reward memory. Neurobiol Learn Mem. 98**,** 130-8.
Shepherd, J.D., Bear, M.F., 2011. New views of Arc, a master regulator of synaptic plasticity. Nat Neurosci. 14**,** 279-84.
Shukla, V., Skuntz, S., Pant, H.C., 2012. Deregulated Cdk5 activity is involved in inducing Alzheimer's disease. Arch Med Res. 43**,** 655-62.
Si, K., Das, K., Maitra, U., 1996. Characterization of multiple mRNAs that encode mammalian translation initiation factor 5 (eIF-5). J Biol Chem. 271**,** 16934-8.
Slavov, N., Dawson, K.A., 2009. Correlation signature of the macroscopic states of the gene regulatory network in cancer. Proc Natl Acad Sci U S A. 106**,** 4079-84.
Smogorzewski, M., Islam, A., 1995. Parathyroid hormone stimulates the generation of inositol 1,4,5-triphosphate in brain synaptosomes. Am J Kidney Dis. 26**,** 814-7.
Sontag, E., Nunbhakdi-Craig, V., Lee, G., Bloom, G.S., et al., 1996. Regulation of the phosphorylation state and microtubule-binding activity of Tau by protein phosphatase 2A. Neuron. 17**,** 1201-7.
Southworth, L.K., Owen, A.B., Kim, S.K., 2009. Aging mice show a decreasing correlation of gene expression within genetic modules. PLoS Genet. 5**,** e1000776.
Sperow, M., Berry, R.B., Bayazitov, I.T., Zhu, G., et al., 2012. Phosphatase and tensin homologue (PTEN) regulates synaptic plasticity independently of its effect on neuronal morphology and migration. J Physiol. 590**,** 777-92.
Spiers, H., Hannon, E., Schalkwyk, L.C., Smith, R., et al., 2015. Methylomic trajectories across human fetal brain development. Genome Res. 25**,** 338-52.
Srivastava, S., Li, Z., Lin, L., Liu, G., et al., 2005. The phosphatidylinositol 3-phosphate phosphatase myotubularin- related protein 6 (MTMR6) is a negative regulator of the Ca2+-activated K+ channel KCa3.1. Mol Cell Biol. 25**,** 3630-8.
Stam, C.J., Jones, B.F., Nolte, G., Breakspear, M., et al., 2007. Small-world networks and functional connectivity in Alzheimer's disease. Cereb Cortex. 17**,** 92-9.
Stone, J.C., 2006. Regulation of Ras in lymphocytes: get a GRP. Biochem Soc Trans. 34**,** 858-61.
Sun, W., Maffie, J.K., Lin, L., Petralia, R.S., et al., 2011. DPP6 establishes the A-type K(+) current gradient critical for the regulation of dendritic excitability in CA1 hippocampal neurons. Neuron. 71**,** 1102-15.
Sweatt, J.D., 2001. The neuronal MAP kinase cascade: a biochemical signal integration system subserving synaptic plasticity and memory. J Neurochem. 76**,** 1-10.
Sweatt, J.D., 2010. Neuroscience. Epigenetics and cognitive aging. Science. 328**,** 701-2.
Takemoto-Kimura, S., Terai, H., Takamoto, M., Ohmae, S., et al., 2003. Molecular cloning and characterization of CLICK-III/CaMKIgamma, a novel membrane-anchored neuronal Ca2+/calmodulin-dependent protein kinase (CaMK). J Biol Chem. 278**,** 18597-605.

Tanaka, S., Syu, A., Ishiguro, H., Inada, T., et al., 2013. DPP6 as a candidate gene for neuroleptic-induced tardive dyskinesia. Pharmacogenomics J. 13**,** 27-34.
Thomas, G.M., Huganir, R.L., 2004. MAPK cascade signalling and synaptic plasticity. Nat Rev Neurosci. 5**,** 173-83.
Timofeeva, O.A., Eddins, D., Yakel, J.L., Blackshear, P.J., et al., 2010. Hippocampal infusions of MARCKS peptides impair memory of rats on the radial-arm maze. Brain Res. 1308**,** 147-52.
Tirnauer, J.S., Grego, S., Salmon, E.D., Mitchison, T.J., 2002. EB1-microtubule interactions in Xenopus egg extracts: role of EB1 in microtubule stabilization and mechanisms of targeting to microtubules. Mol Biol Cell. 13**,** 3614-26.
Torkamani, A., Dean, B., Schork, N.J., Thomas, E.A., 2010. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. Genome Res. 20**,** 403-12.
Uddin, R.K., Singh, S.M., 2013. Hippocampal gene expression meta-analysis identifies aging and age-associated spatial learning impairment (ASLI) genes and pathways. PLoS One. 8**,** e69768.
Vaudry, D., Stork, P.J., Lazarovici, P., Eiden, L.E., 2002. Signaling pathways for PC12 cell differentiation: making the right connections. Science. 296**,** 1648-9.
Vogelsberg-Ragaglia, V., Schuck, T., Trojanowski, J.Q., Lee, V.M., 2001. PP2A mRNA expression is quantitatively decreased in Alzheimer's disease hippocampus. Exp Neurol. 168**,** 402-12.
Voglis, G., Tavernarakis, N., 2006. The role of synaptic ion channels in synaptic plasticity. EMBO Rep. 7**,** 1104-10.
Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., et al., 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 474**,** 380-4.
Waltereit, R., Weller, M., 2003. Signaling from cAMP/PKA to MAPK and synaptic plasticity. Mol Neurobiol. 27**,** 99-106.
Wang, X., Blanchard, J., Tung, Y.C., Grundke-Iqbal, I., et al., 2015. Inhibition of Protein Phosphatase-2A (PP2A) by I1PP2A Leads to Hyperphosphorylation of Tau, Neurodegeneration, and Cognitive Impairment in Rats. J Alzheimers Dis. 45**,** 423-35.
Wei, Z., Behrman, B., Wu, W.H., Chen, B.S., 2015. Subunit-specific regulation of N-methyl-D-aspartate (NMDA) receptor trafficking by SAP102 protein splice variants. J Biol Chem. 290**,** 5105-16.
Williams, E.J., Bowles, D.J., 2004. Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res. 14**,** 1060-7.
Wolf, E.J., Rasmusson, A.M., Mitchell, K.S., Logue, M.W., et al., 2014. A genome-wide association study of clinical symptoms of dissociation in a trauma-exposed sample. Depress Anxiety. 31**,** 352-60.
Wright, M.D., Tomlinson, M.G., 1994. The ins and outs of the transmembrane 4 superfamily. Immunol Today. 15**,** 588-94.
Wuchty, S., Barabasi, A.L., Ferdig, M.T., 2006. Stable evolutionary signal in a yeast protein interaction network. BMC Evol Biol. 6**,** 8.

Wyeth, M.S., Zhang, N., Houser, C.R., 2012. Increased cholecystokinin labeling in the hippocampus of a mouse model of epilepsy maps to spines and glutamatergic terminals. Neuroscience. 202**,** 371-83.

Xiao, Q., Gil, S.C., Yan, P., Wang, Y., et al., 2012. Role of phosphatidylinositol clathrin assembly lymphoid-myeloid leukemia (PICALM) in intracellular amyloid precursor protein (APP) processing and amyloid plaque pathogenesis. J Biol Chem. 287**,** 21279-89.

XiYang, Y.B., Wang, Y.C., Zhao, Y., Ru, J., et al., 2015. Sodium Channel Voltage-Gated Beta 2 Plays a Vital Role in Brain Aging Associated with Synaptic Plasticity and Expression of COX5A and FGF-2. Mol Neurobiol.

Xu, Y., Xing, Y., Chen, Y., Chao, Y., et al., 2006. Structure of the protein phosphatase 2A holoenzyme. Cell. 127**,** 1239-51.

Ye, H., Liu, W., 2015. Transcriptional networks implicated in human nonalcoholic fatty liver disease. Mol Genet Genomics.

Ye, X., Carew, T.J., 2010. Small G protein signaling in neuronal plasticity and memory formation: the specific role of ras family proteins. Neuron. 68**,** 340-61.

York, R.D., Molliver, D.C., Grewal, S.S., Stenberg, P.E., et al., 2000. Role of phosphoinositide 3-kinase and endocytosis in nerve growth factor-induced extracellular signal-regulated kinase activation via Ras and Rap1. Mol Cell Biol. 20**,** 8069-83.

Zhou, Y., Xu, J., Liu, Y., Li, J., et al., 2014. Rat hepatocytes weighted gene co-expression network analysis identifies specific modules and hub genes related to liver regeneration after partial hepatectomy. PLoS One. 9**,** e94868.

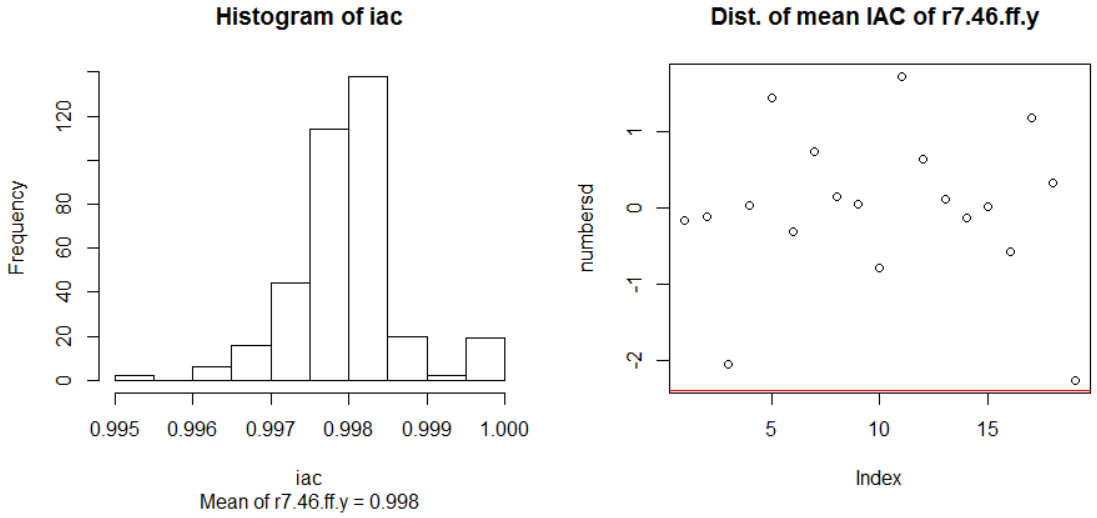
Zou, J., Chang, S.C., Marjanovic, J., Majerus, P.W., 2009. MTMR9 increases MTMR6 enzyme activity, stability, and role in apoptosis. J Biol Chem. 284**,** 2064-71.

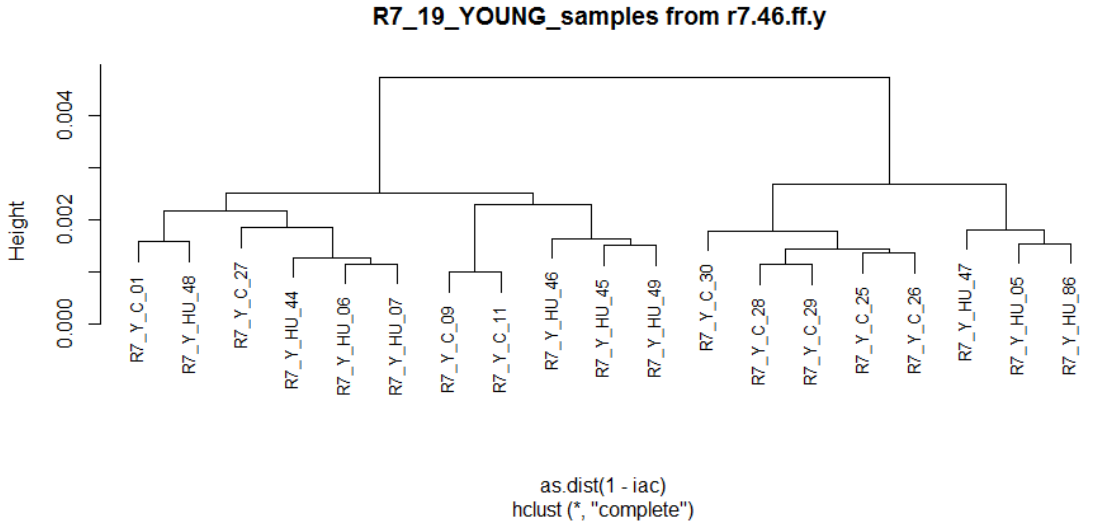
Zwanzger, P., Domschke, K., Bradwejn, J., 2012. Neuronal network of panic disorder: the role of the neuropeptide cholecystokinin. Depress Anxiety. 29**,** 762-74.

6 Appendices

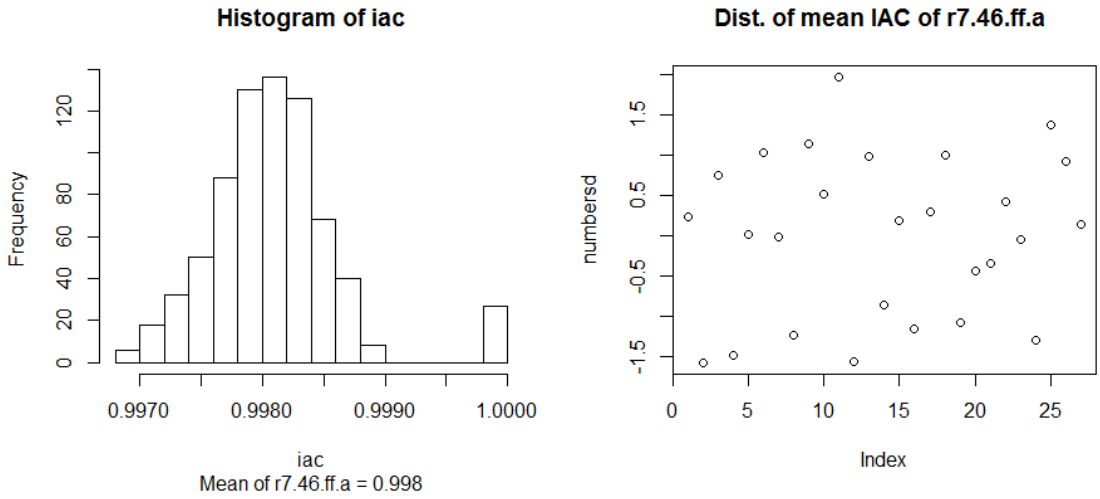
6.1 IAC based quality status of individual young and aged sample groups for the final datasets selected for network analysis.

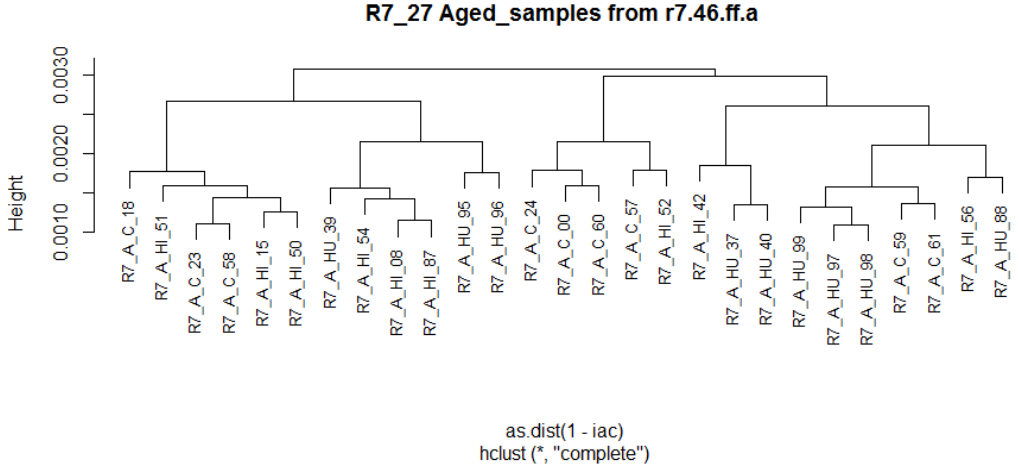
Appendix 6.1.1 IAC based quality check for R7 young dataset. The mean IAC for the 19 young samples were 0.998 (A) and all arrays were 2.5 standard deviations below the mean (B). No outlier is evident in the hierarchical clustering dendrogram (C).



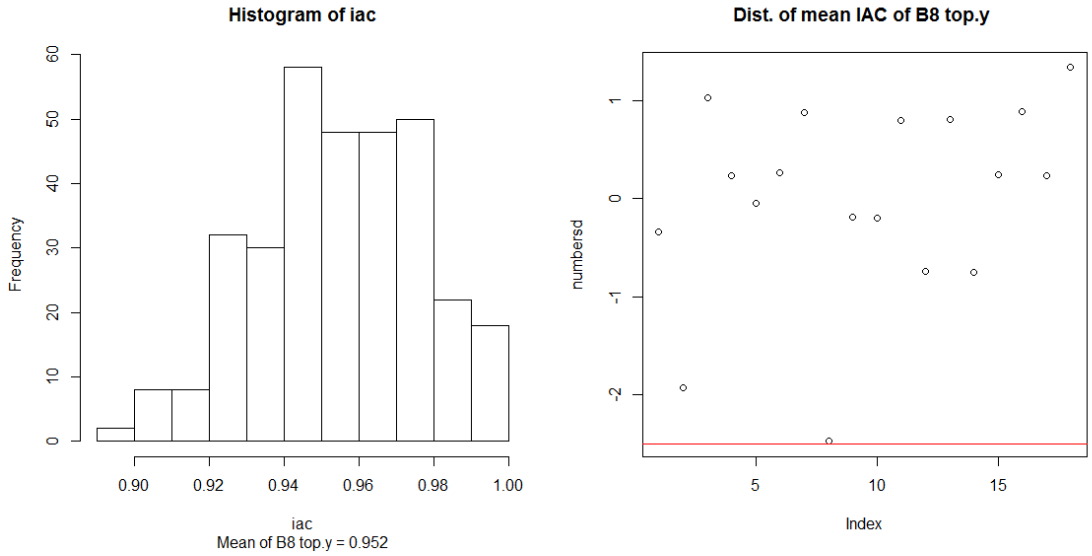


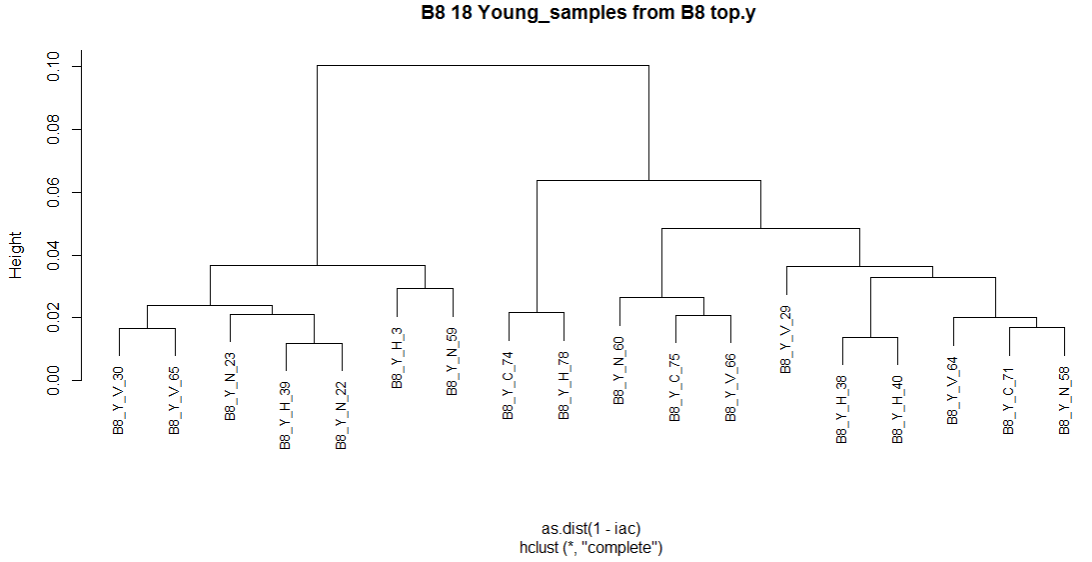
Appendix 6.1.2 IAC based quality check for R7 aged dataset. The mean IAC for the 27 aged samples were 0.998 (A) and all arrays were 2 standard deviations below the mean (B). No outlier is evident in the hierarchical clustering dendrogram (C).



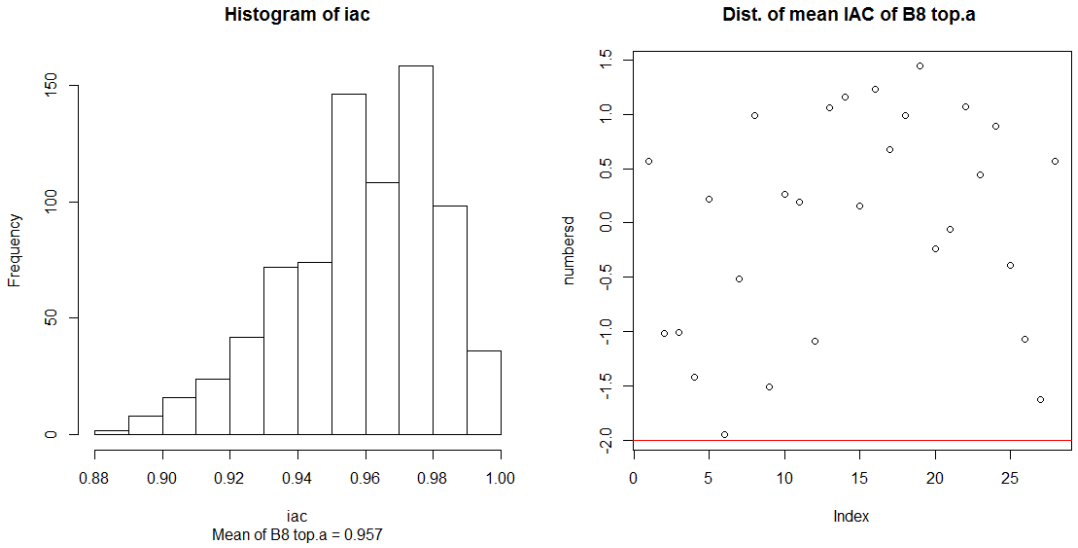


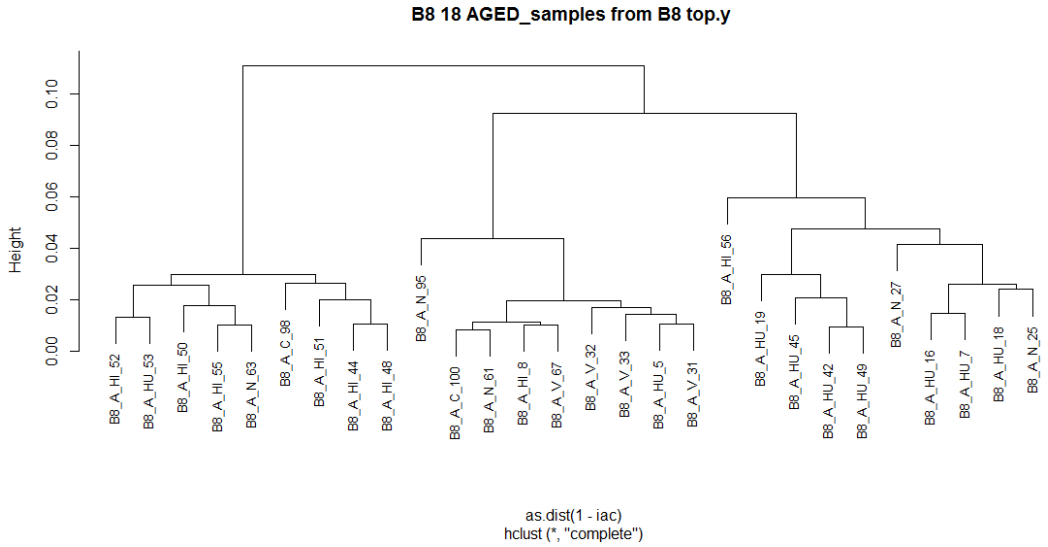
Appendix 6.1.3 IAC based quality check for B8 young dataset. The mean IAC for the 18 young samples were 0.952 (A) and all arrays were within 3 standard deviations below the mean (B). No outlier is evident in the hierarchical clustering dendrogram (C).



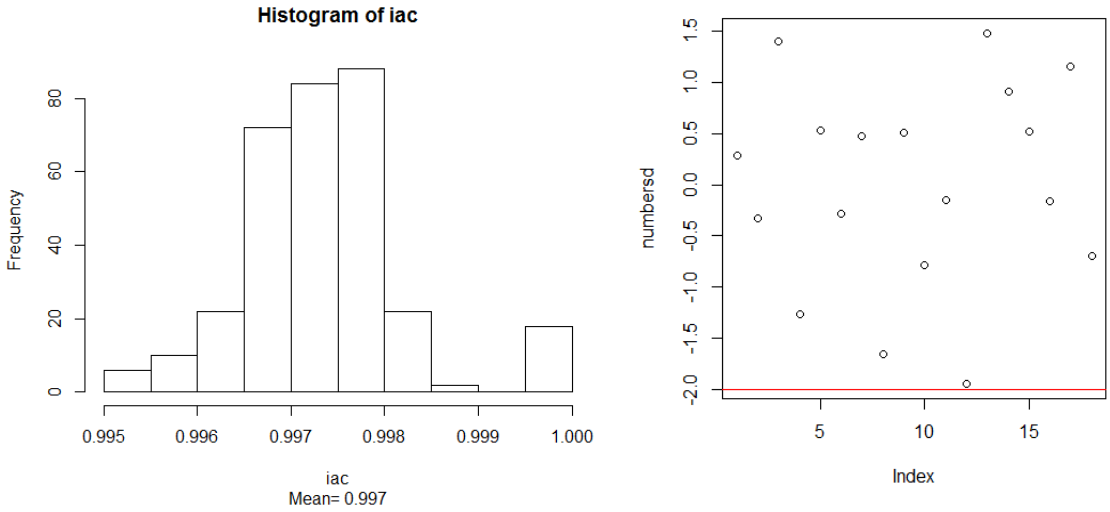


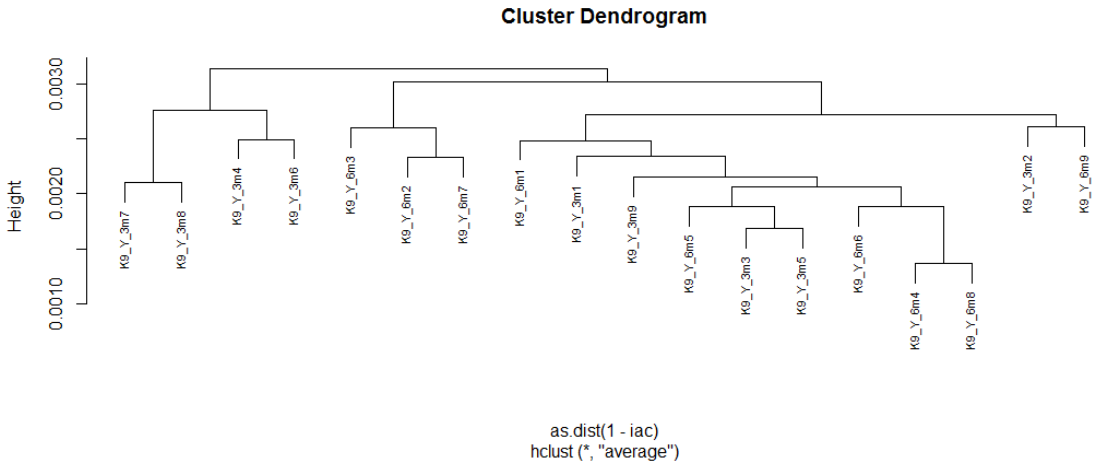
Appendix 6.1.4 IAC based quality check for B8 aged dataset. The mean IAC for the 28 aged samples were 0.957 (A) and all arrays were 2 standard deviations below the mean (B). No outlier is evident in the hierarchical clustering dendrogram (C).



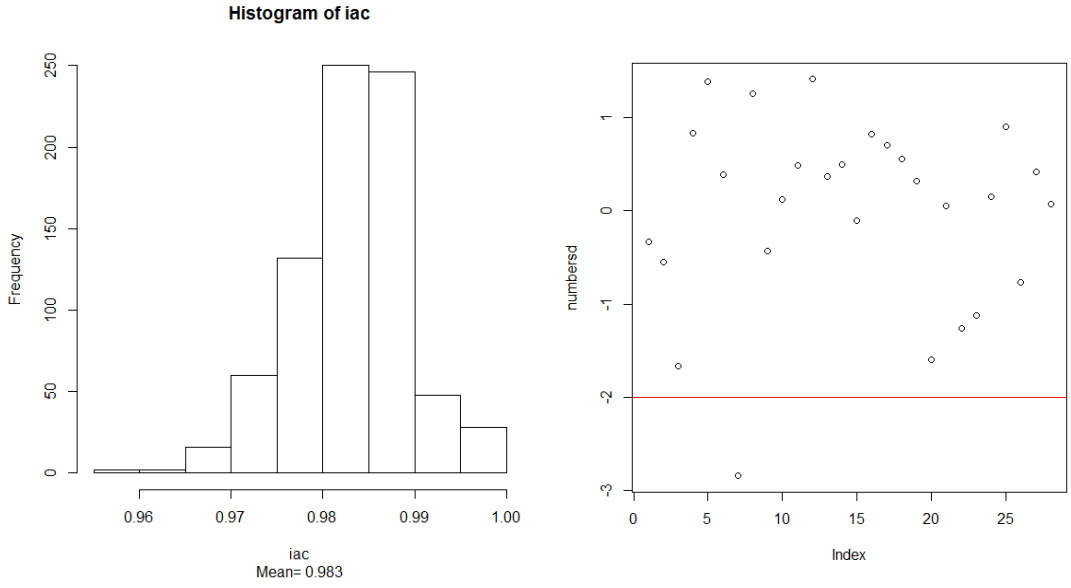


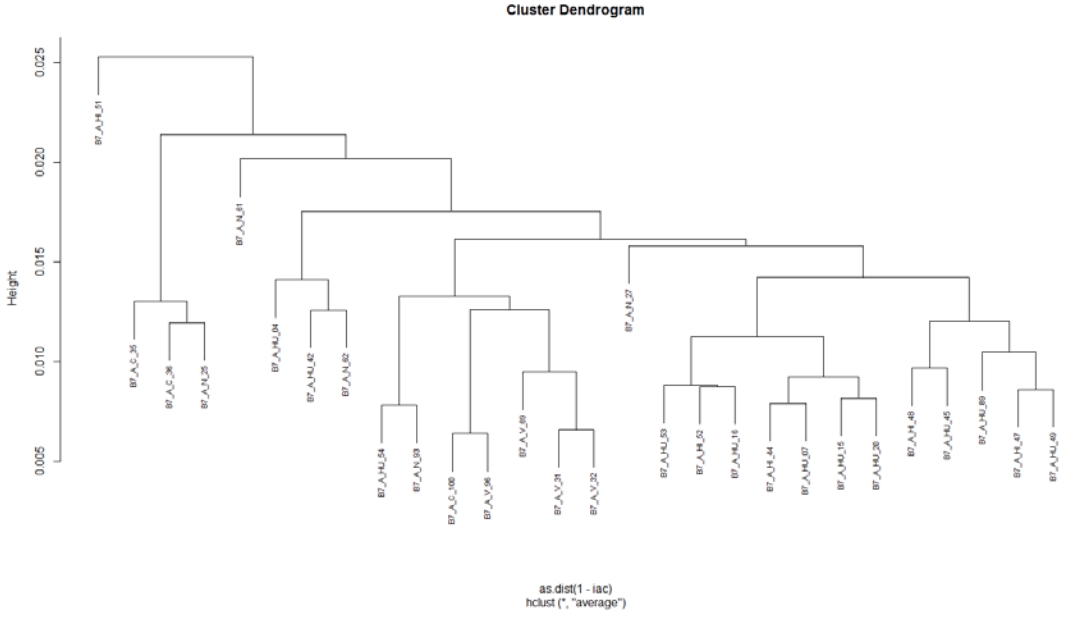
Appendix 6.1.5 IAC based quality check for K9 young dataset. The mean IAC for the 18 young samples were 0.997 (A) and all arrays were 2 standard deviations below the mean (B). No outlier is evident in the hierarchical clustering dendrogram (C).





Appendix 6.1.6 IAC based quality check for B7 aged dataset. The mean IAC for the 28 young samples were 0.983 (A) and all arrays were 3 standard deviations below the mean (B). No outlier is evident in the hierarchical clustering dendrogram (C).





6.2 R Function CollapseGenesRai

Appendix 6.2.1 R function collapseGenesRai(…)

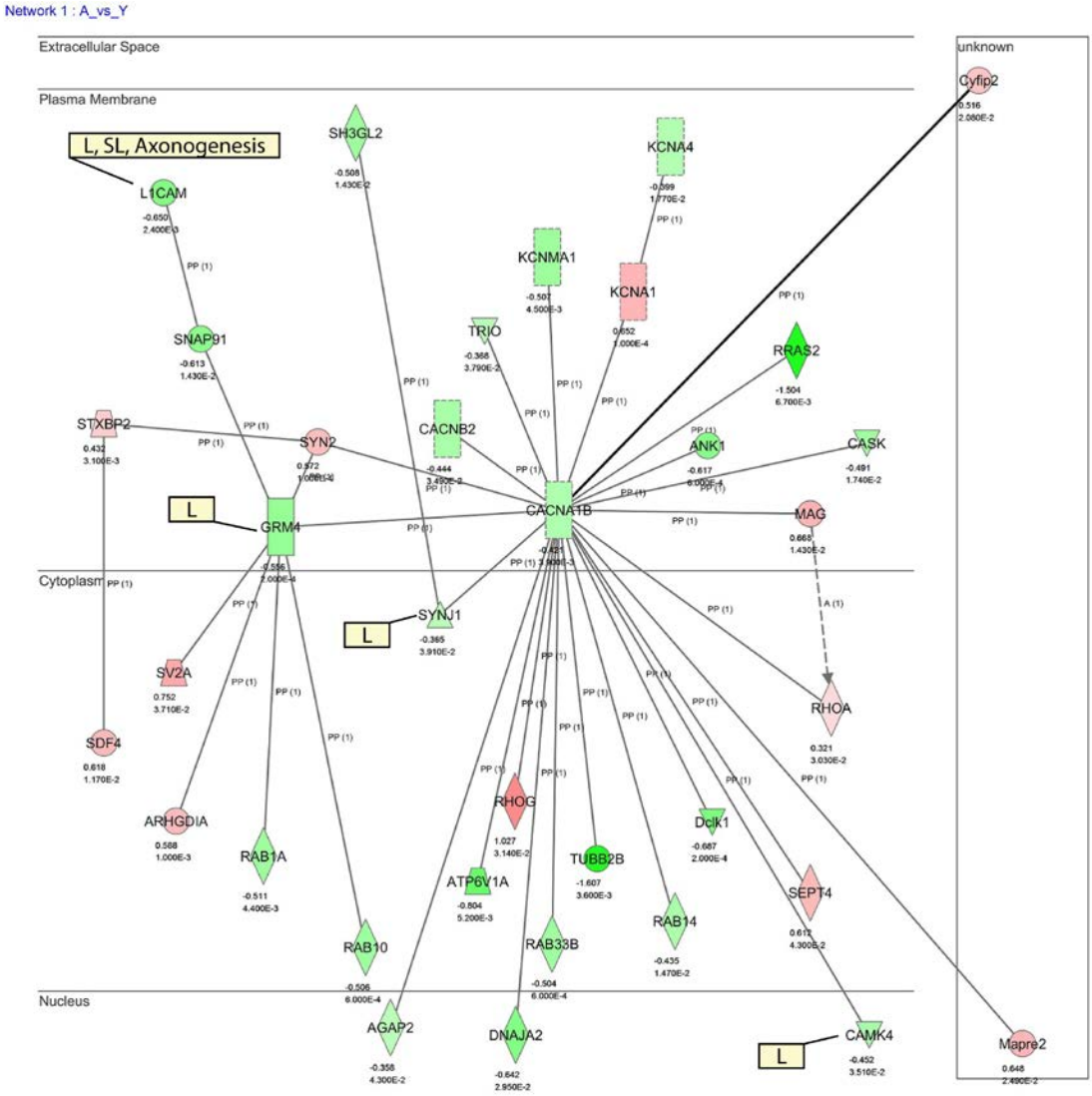
Many genes contain duplicate or multiple probe sets, in which case this function can select a single probe sets with the highest connectivity.

```
collapseGenesRai <- function(dat, allGenes, allProbes, abs=FALSE)
{
   ## Collapse genes with multiple probe sets together using the following algorthim:
   # 1) if there is one probe set/gene = keep
   # 2) if there is two or more take the probe set with max connectivity
   # dat is an expression matrix with rows=genes and cols=samples
   # function will return a list object of dat matrix and gene/probe set matrix

   names(allGenes) = allProbes
   probes = rownames(dat)
   genes  = allGenes[probes]
   tGenes = table(genes)
   datOut=matrix(0,nrow=length(tGenes),ncol=length(colnames(dat)))
   colnames(datOut) = colnames(dat)
   rownames(datOut) = sort(names(tGenes))
   ones = sort(names(tGenes)[tGenes==1])
   more = sort(names(tGenes)[tGenes >= 2])
   gp = matrix(0, nrow=length(tGenes), ncol = 2) ## matrix to hold gene , pset
   rownames(gp) = sort(names(tGenes))
   colnames(gp) = c("genes", "probes")

  for (g in ones){
     datOut[g,] = as.numeric(dat[probes[genes==g],])## just copy the expr data for
     #these ones, no need to do anything and fill out the datOut matrix
     #for the respective genes, genes with two or more pset are not
     #filled in the matrix datOUt – they remain zeo, see below
     gp[g, ] = as.character(c(g, probes[genes==g]))
   }
   for (g in more){
     datTmp = dat[probes[genes==g],]
     adj = cor(t(datTmp))^2  # choose power = 2 for connectivity
     datOut[g,] = as.numeric(datTmp[which.max(rowSums(adj)),])
     datTmp.pset = as.character(rownames(datTmp))
     gp[g, ] = as.character(c(g, datTmp.pset[which.max(rowSums(adj))] ))
   }
   return(list(datOut, gp))
}
```

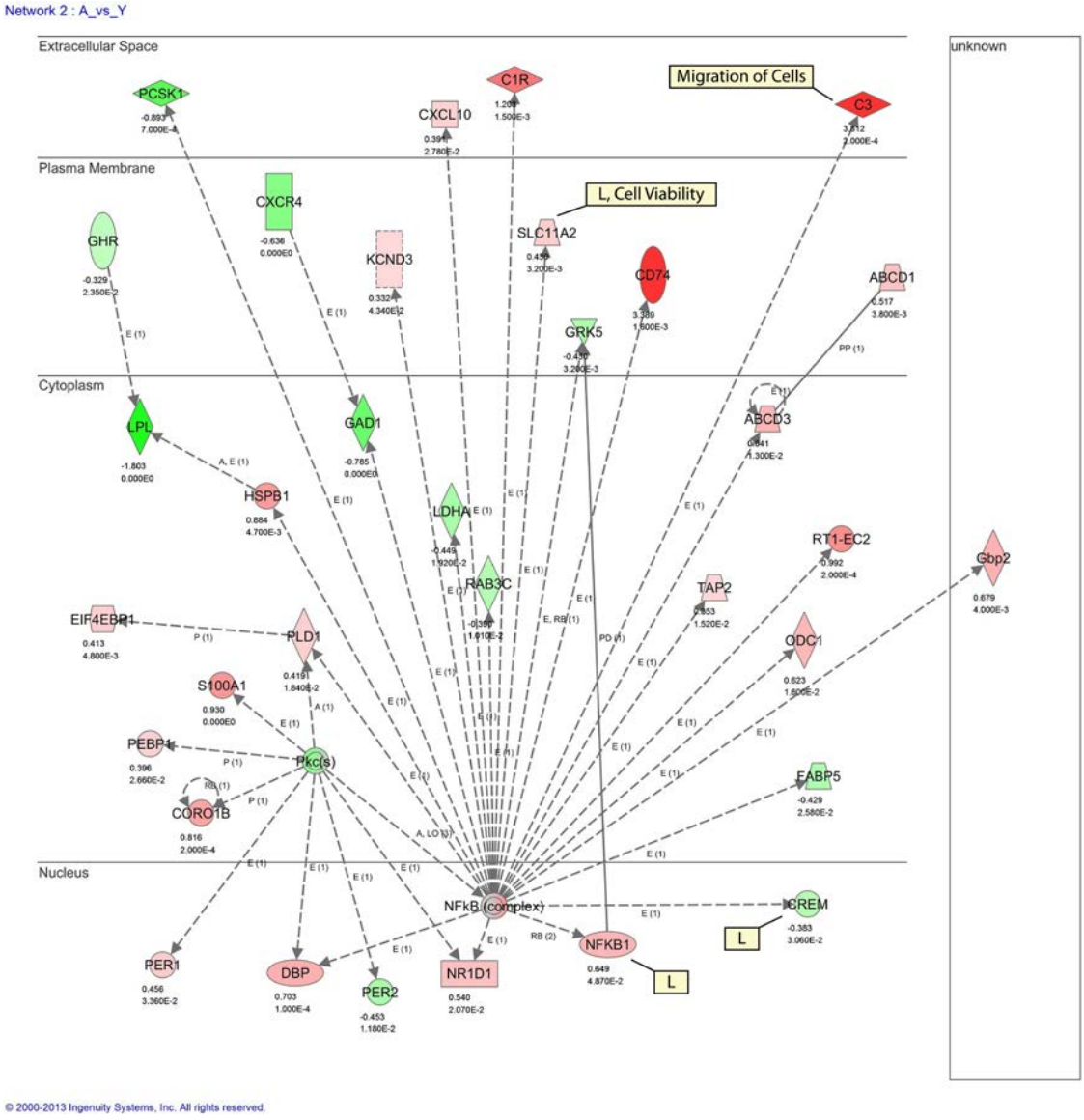
6.3 Knowledge based networks from the AY comparison

In each network, each biological relationship (an edge) between two genes (nodes) is supported by at least one reference from the literature or curated information stored in the IPA knowledge base. The intensity of the node color indicates the degree of up- (red) or down- (green) regulation represented by the effect size as observed in the AY comparison (see Section 3.3.2.1). The effect size and p-value for each gene is shown below the gene symbol. Edges are displayed with various labels that describe the nature of relationship between the genes (e.g. P for phosphorylation, PP for protein-protein binding, PD for protein-DNA binding, A for activation, E for expression, L for proteolysis, LO for localization, RB for regulation of binding). Any specific findings for a gene whether it is associated with aging (A), learning (L), and/or spatial learning (SL) is presented inside a rectangle beside that gene.

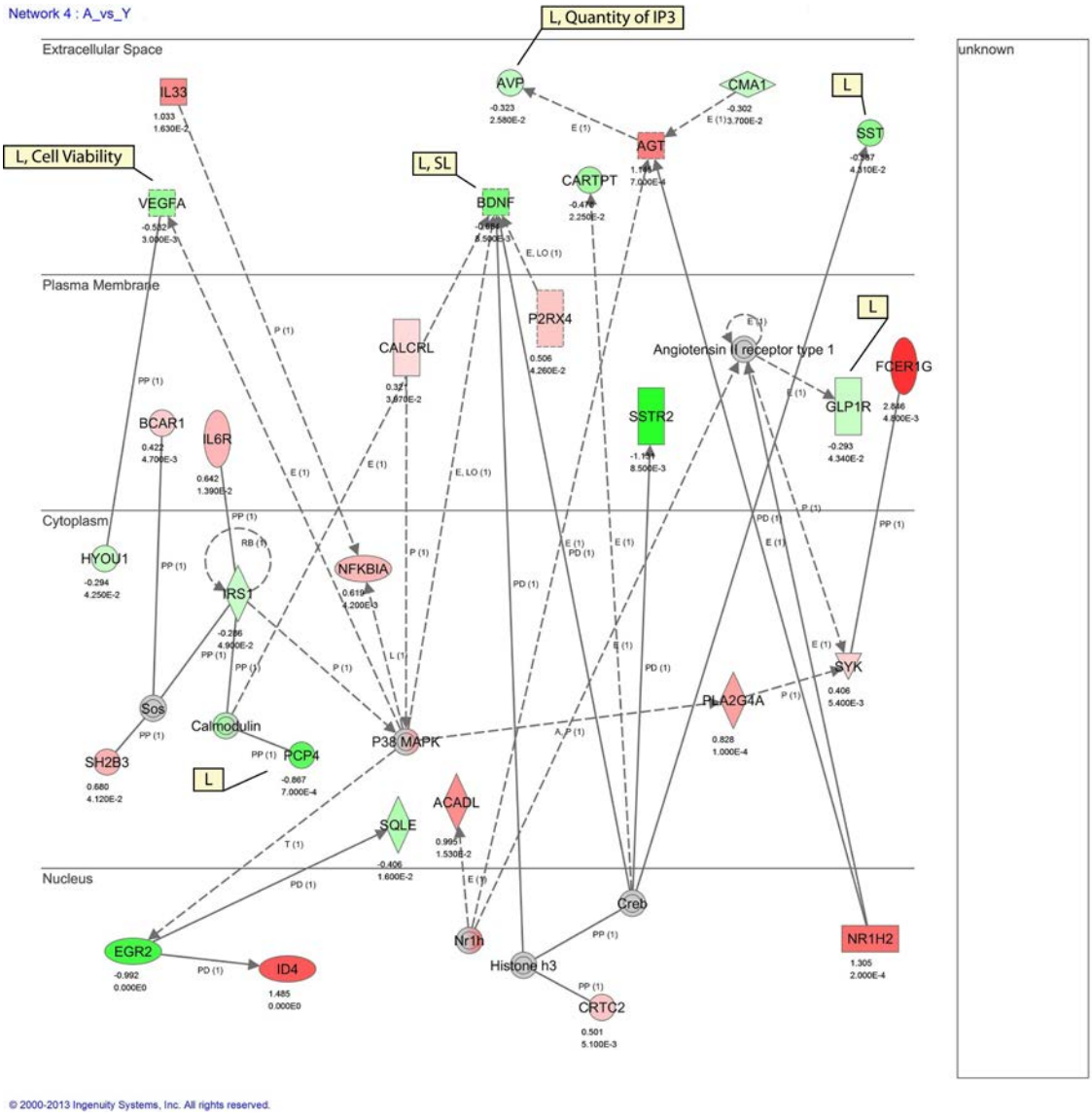
Appendix 6.3.1 Network AY-1: Molecular transport, cell-to-cell signaling and interaction, nervous system development and function.



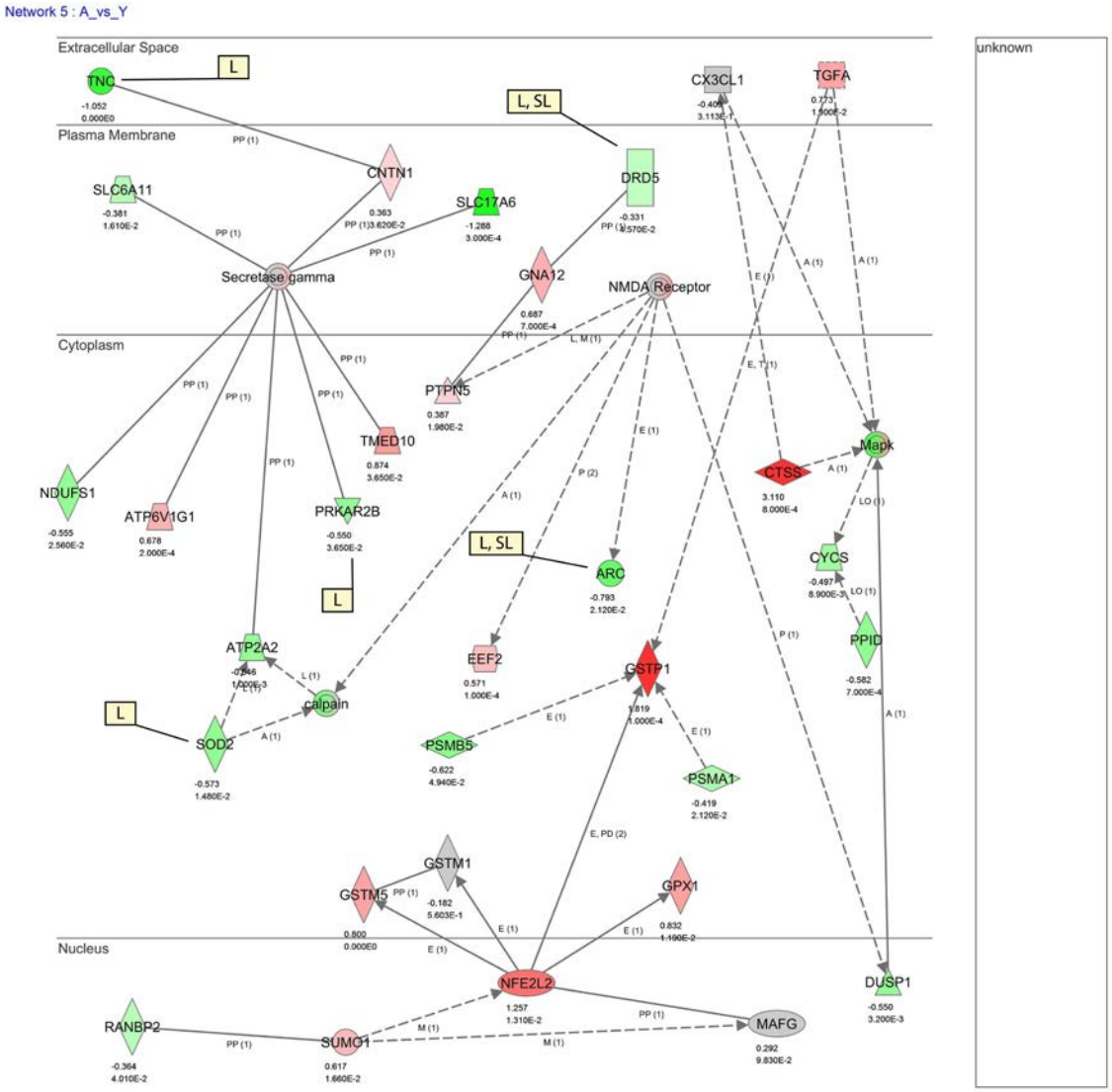
Appendix 6.3.2 Network AY-2: Endocrine system disorders, gastrointestinal disease, metabolic disease.



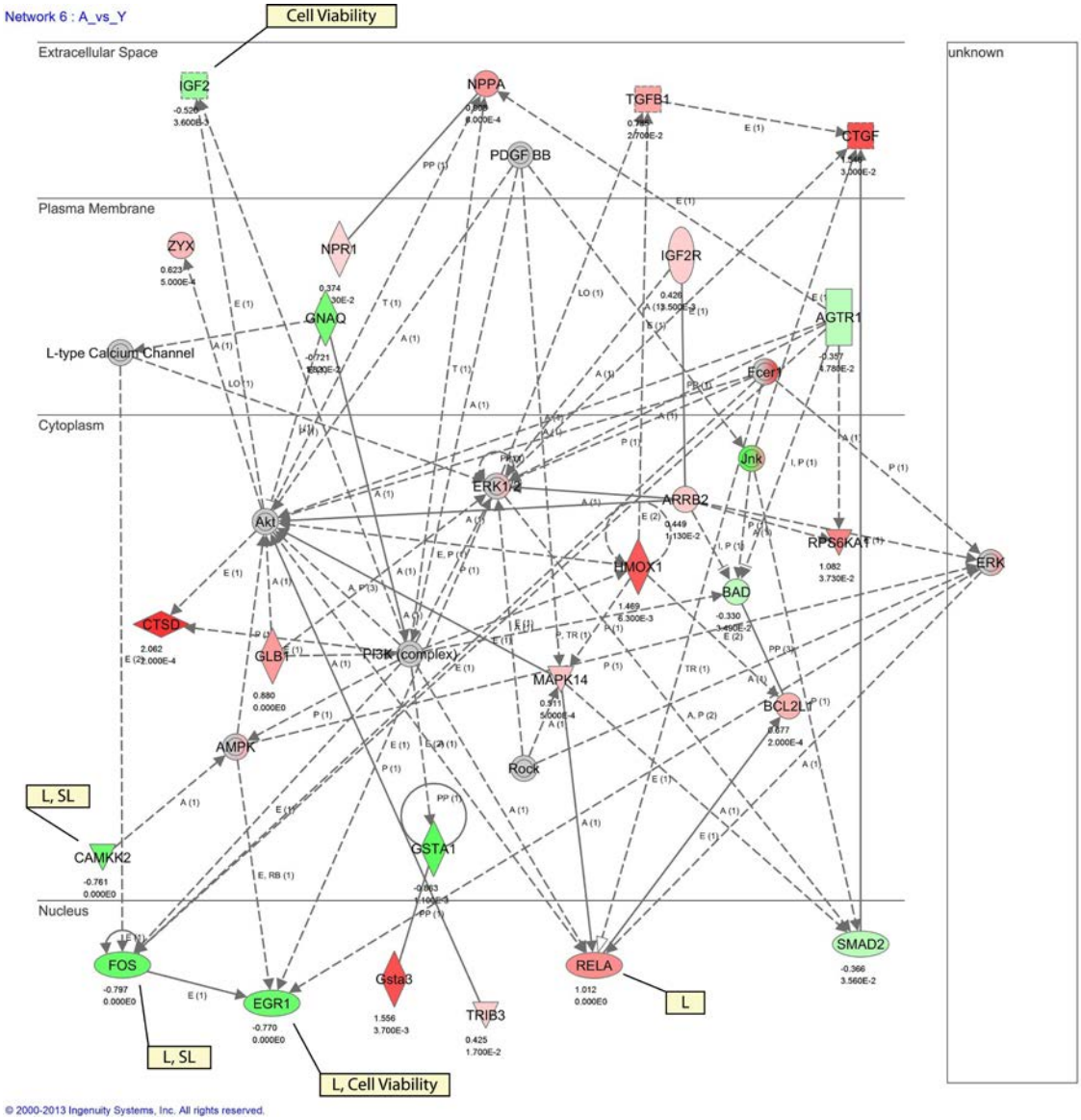
Appendix 6.3.3 Network AY-4: Cell-to-cell signaling and interaction, cell signaling, molecular transport.



Appendix 6.3.4 Network AY-5: Drug metabolism, protein synthesis, cancer.



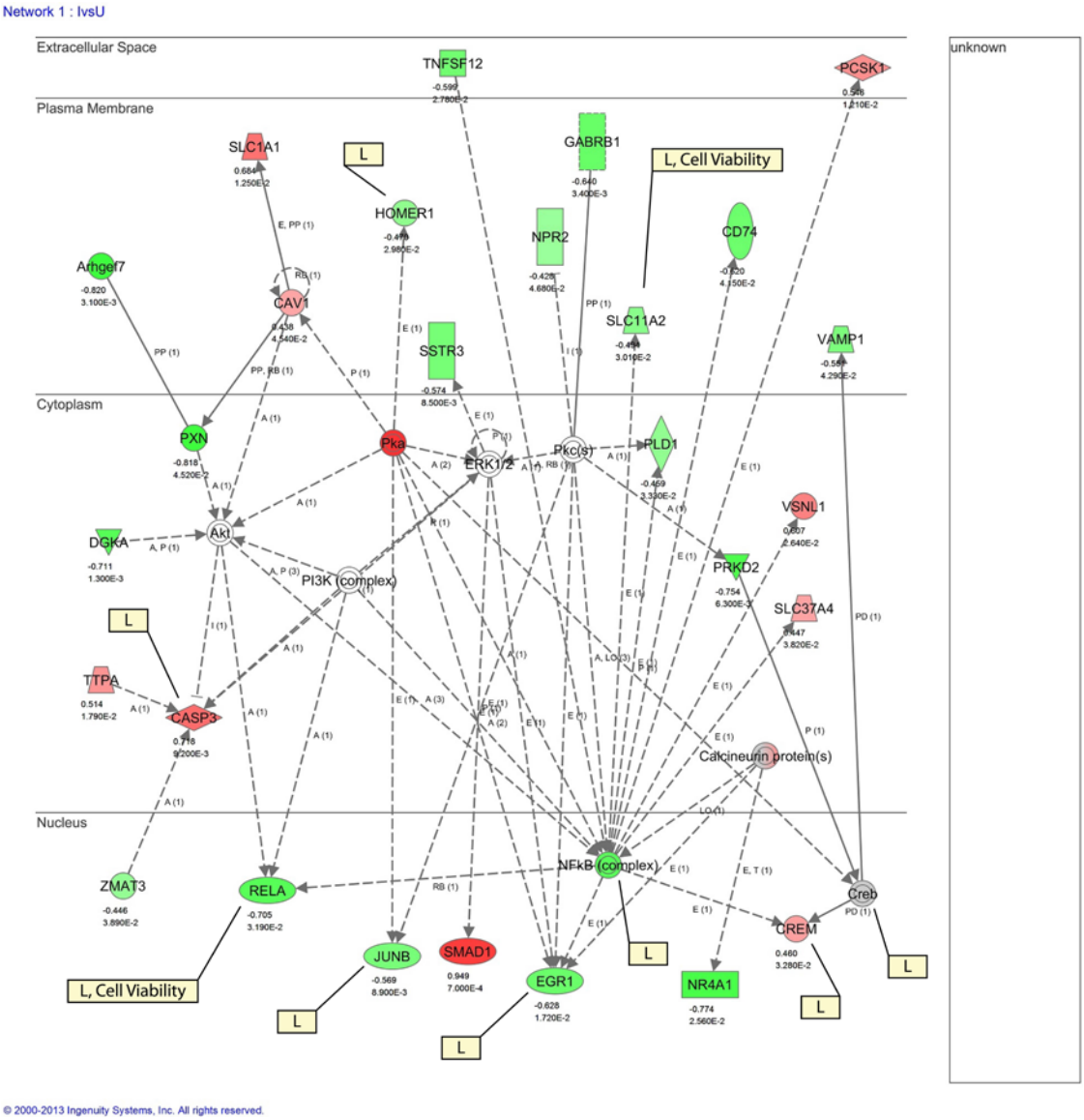
Appendix 6.3.5 Network AY-6: Cell death and survival, renal necrosis/cell death, lipid metabolism.



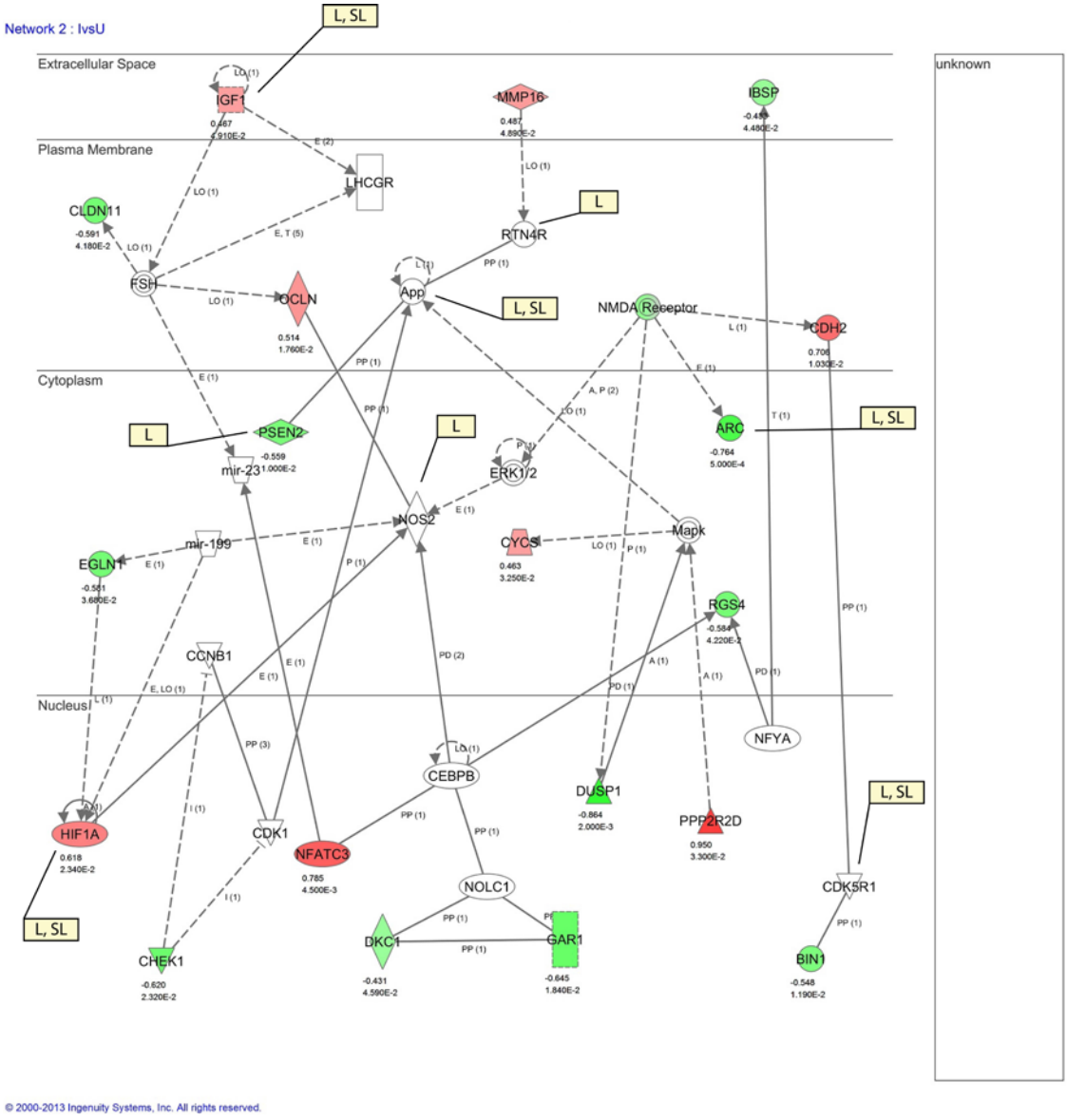
6.4 Knowledge based networks from the IU comparison

Each biological relationship (an edge) between two genes (nodes) is supported by at least one reference from the literature or curated information stored in the IPA knowledge base. The intensity of the node color indicates the degree of up- (red) or down- (green) regulation represented by the effect size as observed in the IU comparison (see Section 3.3.2.2). The effect size and p-value for each gene is shown below the gene symbol. Edges are displayed with various labels that describe the nature of relationship between the genes (e.g. P for phosphorylation, PP for protein-protein binding, PD for protein-DNA binding, A for activation, E for expression, L for proteolysis, LO for localization, RB for regulation of binding). Any specific findings for a gene whether it is associated with aging (A), learning (L), and/or spatial learning (SL) is presented inside a rectangle beside that gene.

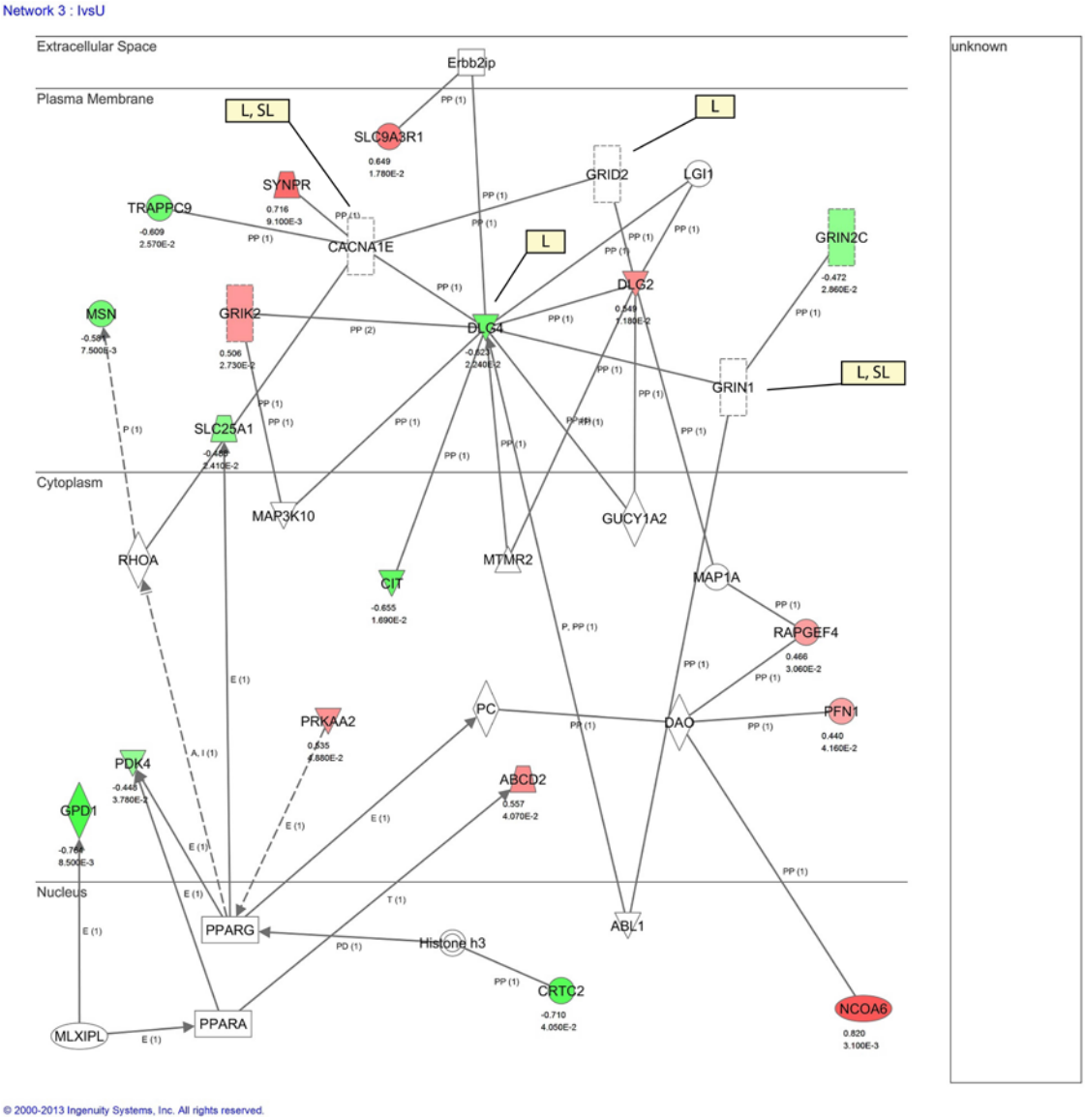
Appendix 6.4.1 Network IU-1: Neurological disease, tissue morphology.



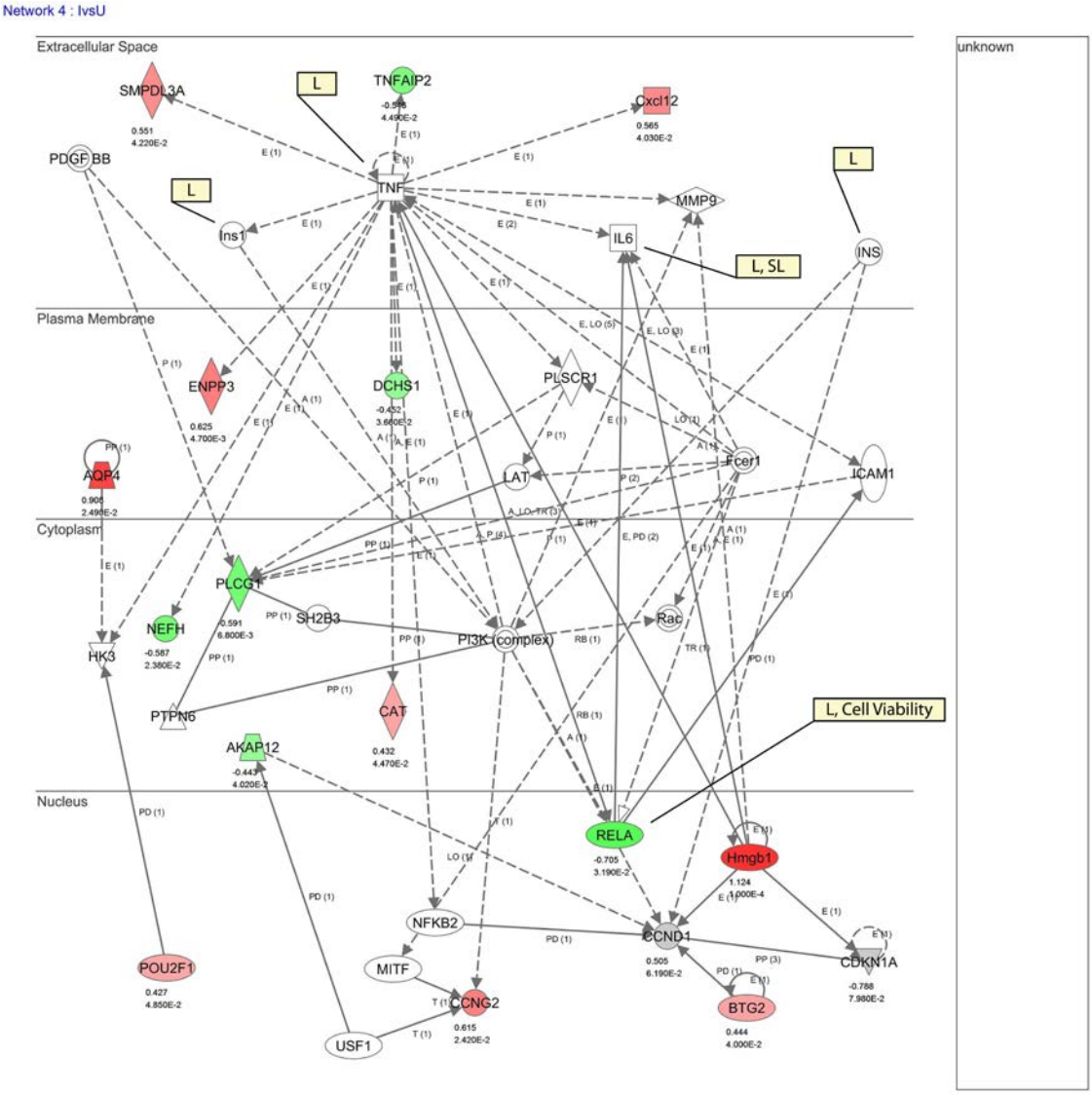
Appendix 6.4.2 Network IU-2: Cellular growth and proliferation, cancer, cell death and survival.



Appendix 6.4.3 Network IU-3: Cell-to-cell signaling and interaction, nervous system development and function, carbohydrate metabolism.



Appendix 6.4.4 Network IU-4: Cell death and survival, cellular development, hematological system development and function.



6.5 Significant meta-analysis genes in the yellow module

Appendix 6.5.1: The image (R screenshot) below shows the 165 significant meta-analysis genes which are also present in the yellow module. The candidate ASLI hub genes are highlighted in yellow.

```
> tmp = meta.in.5674[meta.in.5674$module == "yellow" ,5]

> tmp
  [1] Abhd13       Acsl1        Actb         Adora1       Aftph        Agap2
  [7] Amdhd2       Ank1         Ankrd28      Ap3b1        Apc          Arih1
 [13] Armcx3       Arrb2        Atp1a1       Atp6v0a2     B3gat2       Bcl2l1
 [19] Camk1g       Cask         Cdk5r1       Chchd4       Cml3         Cntn1
 [25] Cntn2        Col4a1       Cox18        Crem         Cryab        Cx3cr1
 [31] Dcaf6        Dctn4        Dlgap2       Dnm1l        Dusp3        Eftud2
 [37] Eif2ak4      Eif3j        Eif5         Elf1         Epm2aip1     Fam115a
 [43] Fbxl20       Fgd4         Fgfr2        Ftsj2        G3bp2        Gadd45a
 [49] Ghr          Git2         Glul         Gne          Gnl3l        Grm1
 [55] Gstm4        Hapln1       Hsf2         Ide          Impact       Ing1
 [61] Inha         Ireb2        Jag2         Kcnc2        Kcnv1        Klhl7
 [67] Kpna1        Laptm5       Lepr         Lfng         LOC100362458 LOC100363863
 [73] LOC100363987 LOC100909788 LOC100912981 LOC246295    LOC302022    LOC684996
 [79] Mapk9        Mapre1       Mas1         Mdm2         Mmp24        Mro
 [85] Mycn         Naa35        Nagk         Nap1l5       Ncaph2       Ncor1
 [91] Nlgn1        Nlgn3        Nploc4       Nr1h2        Nrcam        Ntrk2
 [97] Nupl1        Odc1         P2ry12       Papola       Pds5b        Pias2
[103] Pkia         Plcb4        Plekha6      Ppp4r1       Prdx6        Prr3
[109] Psme3        Pten         Pygm         Qprt         Ralgapa1     Ranbp2
[115] Rapgef2      Rbm39        RGD1306820   RGD1311578   Rnf4         Robo1
[121] S100b        Scn8a        Sh3bgrl2     Sh3gl2       Sidt1        Slc31a1
[127] Slc35a1      Slc9a3       Slc9a8       Slco2b1      Snrk         Sod2
[133] Sorcs3       Spock1       Sptbn1       Sqstm1       Srsf10       St6galnac3
[139] Surf2        Syngr1       Synj2bp      Tank         Tceb3        Tm2d1
[145] Tpm1         Tpm3         Trio         Trpc3        Trpm1        Txnl1
[151] Uba5         Ube2a        Ube2d3       Ube2l3       Usp12        Vamp5
[157] Vegfa        Vmp1         Yipf4        Yme1l1       Zbtb17       Zcchc6
[163] Zdhhc2       Zfp292       Zfp692
```

6.6 RDAVIDWebService

Appendix 6.6.1 GetGeneCategoriesReport – Default Parameters

Arguments	Description
object	DAVIDWebService class object
fileName	Character with the name of the file to store the Report
threshold	Numeric with the EASE score (at most equal) that must be present in the category to be included in the report. Default value is 0.1.
count	Integer with the number of genes (greater equal) that must be present in the category to be included in the report. Default value is 2.
type	Character with the type of cluster to obtain Term/Genes. Default value "Term".
overlap	Integer with the minimum number of annotation terms overlapped between two genes in order to be qualified for kappa calculation. This parameter is to maintain necessary statistical power to make kappa value more meaningful. The higher value, the more meaningful the result is. Default value is 4L. The 'L' suffix is used to qualify any number with the intent of making it an explicit integer.
initialSeed, finalSeed	Integer with the number of genes in the initial (seeding) and final (filtering) cluster criteria. Default value is 4L for both.
linkage	Numeric with the percentage of genes that two clusters share in order to become one.
kappa	Integer (kappa * 100), with the minimum kappa value to be considered biological significant. The higher setting, the more genes will be put into unclustered group, which lead to higher quality of functional classification result with a fewer groups and a fewer gene members. Kappa value 0.3 starts giving meaningful biology based on our genome-wide distribution study. Anything below 0.3 have great chance to be noise.

Appendix 6.6.2 Terms used in the getFunctionalAnnotationChartFile output columns under the getGeneCategoriesReport file.

Terms	Description
Category	Factor with the main categories under used in the present analysis
Term	Character with the name of the term in format id~name (if available)
Count	Integer with the number of ids of the gene list that belong to this term
X	After converting user input gene IDs to corresponding DAVID gene ID, it refers to the percentage of DAVID genes in the list associated with a particular annotation term. Since DAVID gene ID is unique per gene, it is more accurate to use DAVID ID percentage to present the gene-annotation association by removing any redundancy in user gene list, i.e. two user IDs represent same gene.
PValue	Numeric with the EASE Score of the term (see DAVID Help page)
Genes	Character in comma separated style with the genes present in the term
List.Total, Pop.Hits, Pop.Total:	Integers (in addition to Count) to build the 2x2 contingency table in order to compute the EASE Score (see DAVID Help page).
Fold.Enrichment	Numeric with the ratio of the two proportions. For example, if 40/400 (i.e. 10%) of your input genes involved in "kinase activity" and the background information is 300/30000 genes (i.e. 1%) associating with "kinase activity", roughly 10% / 1% = 10 fold enrichment.
Bonferroni, Benjamini, FDR	Numerics with p-value adjust different criteria (see p.adjust)

6.7 Gene Ontology Analysis

Appendix 6.7.1 Gene Ontology functional analysis output for the R7 young blue module.

Blue Module			
Annotation Cluster 1	Enrichment Score: 5.076594912569949		
Category	Term	PValue	Benjamini
GOTERM_MF_ALL	GO:0003735~structural constituent of ribosome	2.02E-09	1.68E-06
KEGG_PATHWAY	rno03010:Ribosome	6.02E-08	9.63E-06
GOTERM_BP_ALL	GO:0006412~translation	9.85E-08	2.70E-04
GOTERM_CC_ALL	GO:0005840~ribosome	1.45E-07	3.71E-05
GOTERM_BP_ALL	GO:0006414~translational elongation	1.16E-06	1.59E-03
Annotation Cluster 2	Enrichment Score: 2.3961195219961517		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0044237~cellular metabolic process	2.49E-05	2.26E-02
GOTERM_BP_ALL	GO:0008152~metabolic process	6.78E-05	4.55E-02
GOTERM_BP_ALL	GO:0009058~biosynthetic process	9.27E-05	4.96E-02
Annotation Cluster 3	Enrichment Score: 1.8912996680281784		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0005739~mitochondrion	1.49E-08	7.59E-06

Benjamini: Benjamini multiple testing corrected p-value

Appendix 6.7.2 Gene Ontology functional analysis output for the R7 young brown module.

Brown Module			
Annotation Cluster 1	Enrichment Score: 1.6083199429749104		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0044446~intracellular organelle part	6.02E-03	5.25E-01
GOTERM_BP_ALL	GO:0009987~cellular process	1.68E-02	9.20E-01
GOTERM_MF_ALL	GO:0005488~binding	4.93E-02	9.58E-01
Annotation Cluster 2	Enrichment Score: 1.4028984566788694		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0007017~microtubule-based process	1.18E-03	9.58E-01
GOTERM_CC_ALL	GO:0005874~microtubule	9.26E-03	3.99E-01
Annotation Cluster 3	Enrichment Score: 1.261249164925707		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0032991~macromolecular complex	1.79E-03	5.87E-01
Annotation Cluster 4	Enrichment Score: 1.1955856826429254		
Category	Term	PValue	Benjamini
GOTERM_MF_ALL	GO:0016818~hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	1.61E-02	9.64E-01
GOTERM_MF_ALL	GO:0003924~GTPase activity	3.78E-02	9.57E-01
GOTERM_MF_ALL	GO:0042626~ATPase activity, coupled to transmembrane movement of substances	5.01E-02	9.50E-01
GOTERM_MF_ALL	GO:0019001~guanyl nucleotide binding	5.54E-02	9.45E-01
Annotation Cluster 6	Enrichment Score: 1.1398291604515942		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0051641~cellular localization	2.72E-02	9.54E-01
GOTERM_BP_ALL	GO:0046907~intracellular transport	4.22E-02	9.60E-01
Annotation Cluster 7	Enrichment Score: 1.1021127923779181		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0006873~cellular ion homeostasis	6.51E-03	9.46E-01
GOTERM_BP_ALL	GO:0042552~myelination	7.42E-03	9.17E-01
GOTERM_BP_ALL	GO:0007272~ensheathment of neurons	9.86E-03	9.10E-01
GOTERM_BP_ALL	GO:0007154~cell communication	2.06E-02	9.38E-01

Benjamini: Benjamini multiple testing corrected p-value

Appendix 6.7.3 Gene Ontology functional analysis output for the R7 young green module.

Green			
Annotation Cluster 1	Enrichment Score: 1.5613644814308199		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0032502~developmental process	9.36E-04	8.39E-01
GOTERM_BP_ALL	GO:0007275~multicellular organismal development	3.54E-03	7.49E-01
GOTERM_BP_ALL	GO:0048731~system development	4.76E-03	7.35E-01
Annotation Cluster 2	Enrichment Score: 1.2600508748712447		
Category	Term	PValue	Benjamini
GOTERM_MF_ALL	GO:0030695~GTPase regulator activity	5.58E-03	6.47E-01
GOTERM_MF_ALL	GO:0060589~nucleoside-triphosphatase regulator activity	7.93E-03	5.89E-01
GOTERM_BP_ALL	GO:0035023~regulation of Rho protein signal transduction	3.52E-02	9.64E-01
GOTERM_MF_ALL	GO:0008047~enzyme activator activity	4.50E-02	9.23E-01
Annotation Cluster 3	Enrichment Score: 0.992893123718538		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0016265~death	6.48E-02	9.76E-01
GOTERM_BP_ALL	GO:0012501~programmed cell death	8.99E-02	9.78E-01
GOTERM_BP_ALL	GO:0008219~cell death	9.80E-02	9.74E-01
Annotation Cluster 4	Enrichment Score: 0.9869303313882487		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0044421~extracellular region part	5.91E-04	1.79E-01
GOTERM_CC_ALL	GO:0005615~extracellular space	1.80E-03	2.59E-01

Benjamini: Benjamini multiple testing corrected p-value

Appendix 6.7.4 Gene Ontology functional analysis output for the R7 young red module.

Red Module			
Annotation Cluster 1	Enrichment Score: 1.8696327308817076		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0006476~protein amino acid deacetylation	7.73E-03	9.99E-01
GOTERM_MF_ALL	GO:0004407~histone deacetylase activity	1.15E-02	9.99E-01
GOTERM_MF_ALL	GO:0033558~protein deacetylase activity	1.15E-02	9.99E-01
Annotation Cluster 2	Enrichment Score: 1.5025568983239734		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0048709~oligodendrocyte differentiation	5.64E-03	1.00E+00
GOTERM_BP_ALL	GO:0010001~glial cell differentiation	7.84E-03	9.93E-01
GOTERM_BP_ALL	GO:0042063~gliogenesis	8.49E-03	9.82E-01
GOTERM_BP_ALL	GO:0021782~glial cell development	1.69E-02	9.95E-01
GOTERM_BP_ALL	GO:0014003~oligodendrocyte development	1.69E-02	9.90E-01
Annotation Cluster 3	Enrichment Score: 1.2016709930074059		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0009725~response to hormone stimulus	2.09E-02	9.82E-01
GOTERM_BP_ALL	GO:0009749~response to glucose stimulus	3.63E-02	9.93E-01
Annotation Cluster 4	Enrichment Score: 1.0221510133086018		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0016020~membrane	3.24E-02	9.74E-01
GOTERM_CC_ALL	GO:0016021~integral to membrane	3.62E-02	9.54E-01

Benjamini: Benjamini multiple testing corrected p-value

Appendix 6.7.5 Gene Ontology functional analysis output for the R7 young turquoise module.

Turquoise Module			
Annotation Cluster 1	Enrichment Score: 1.9334188254920037		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0005739~mitochondrion	3.12E-06	1.56E-03
KEGG_PATHWAY	rno05016:Huntington's disease	1.20E-04	1.95E-02
KEGG_PATHWAY	rno05012:Parkinson's disease	1.64E-04	1.35E-02
KEGG_PATHWAY	rno00190:Oxidative phosphorylation	2.78E-04	1.52E-02
KEGG_PATHWAY	rno05010:Alzheimer's disease	5.20E-04	2.12E-02
GOTERM_MF_ALL	GO:0050136~NADH dehydrogenase (quinone) activity	1.35E-03	3.43E-01
GOTERM_MF_ALL	GO:0008137~NADH dehydrogenase (ubiquinone) activity	1.35E-03	3.43E-01
GOTERM_CC_ALL	GO:0031090~organelle membrane	1.97E-03	2.18E-01
GOTERM_CC_ALL	GO:0070469~respiratory chain	2.08E-03	1.87E-01
Annotation Cluster 2	Enrichment Score: 1.7247276117582555		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0005840~ribosome	7.14E-06	1.78E-03
GOTERM_MF_ALL	GO:0003735~structural constituent of ribosome	8.52E-04	5.49E-01
GOTERM_CC_ALL	GO:0030529~ribonucleoprotein complex	6.94E-03	2.35E-01
GOTERM_BP_ALL	GO:0006412~translation	9.84E-03	1.00E+00
Annotation Cluster 3	Enrichment Score: 1.6999225473101733		
Category	Term	PValue	Benjamini
KEGG_PATHWAY	rno05016:Huntington's disease	1.20E-04	1.95E-02
KEGG_PATHWAY	rno05012:Parkinson's disease	1.64E-04	1.35E-02
KEGG_PATHWAY	rno00190:Oxidative phosphorylation	2.78E-04	1.52E-02
KEGG_PATHWAY	rno05010:Alzheimer's disease	5.20E-04	2.12E-02
GOTERM_CC_ALL	GO:0070469~respiratory chain	2.08E-03	1.87E-01
Annotation Cluster 4	Enrichment Score: 1.6261616792501006		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0005761~mitochondrial ribosome	7.53E-03	2.22E-01
GOTERM_CC_ALL	GO:0000313~organellar ribosome	7.53E-03	2.22E-01
Annotation Cluster 5	Enrichment Score: 1.35843773527133		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0043227~membrane-bounded organelle	6.16E-03	2.65E-01

Benjamini: Benjamini multiple testing corrected p-value

Appendix 6.7.6 Gene Ontology functional analysis output for the R7 young yellow module (truncated).

Yellow			
Annotation Cluster 1	Enrichment Score: 5.919466002347552		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0060341~regulation of cellular localization	4.47E-09	1.37E-05
GOTERM_BP_ALL	GO:0032879~regulation of localization	6.73E-09	1.03E-05
GOTERM_BP_ALL	GO:0051046~regulation of secretion	1.85E-08	1.89E-05
GOTERM_BP_ALL	GO:0051049~regulation of transport	2.07E-08	1.59E-05
GOTERM_BP_ALL	GO:0010646~regulation of cell communication	1.04E-07	6.38E-05
GOTERM_BP_ALL	GO:0050804~regulation of synaptic transmission	2.94E-04	3.69E-02
Annotation Cluster 2	Enrichment Score: 5.910461921496712		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0005886~plasma membrane	1.06E-12	1.55E-10
GOTERM_CC_ALL	GO:0016021~integral to membrane	9.47E-08	5.21E-06
GOTERM_BP_ALL	GO:0051179~localization	8.41E-05	1.83E-02
GOTERM_BP_ALL	GO:0006810~transport	1.52E-04	2.56E-02
GOTERM_MF_ALL	GO:0005215~transporter activity	1.34E-03	6.18E-02
Annotation Cluster 3	Enrichment Score: 5.071792828679952		
Category	Term	PValue	Benjamini
GOTERM_MF_ALL	GO:0004888~transmembrane receptor activity	3.96E-08	3.40E-05
GOTERM_MF_ALL	GO:0004871~signal transducer activity	6.69E-07	1.92E-04
GOTERM_BP_ALL	GO:0007166~cell surface receptor linked signal transduction	2.25E-06	9.85E-04
KEGG_PATHWAY	rno04080:Neuroactive ligand-receptor interaction	2.53E-05	3.67E-03
GOTERM_BP_ALL	GO:0007186~G-protein coupled receptor protein signaling pathway	5.88E-04	5.33E-02
GOTERM_MF_ALL	GO:0004930~G-protein coupled receptor activity	4.42E-03	1.23E-01
Annotation Cluster 4	Enrichment Score: 4.853521005791371		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0045202~synapse	4.77E-15	2.10E-12
GOTERM_CC_ALL	GO:0044456~synapse part	2.05E-14	4.52E-12
GOTERM_CC_ALL	GO:0045211~postsynaptic membrane	3.04E-07	1.34E-05
GOTERM_BP_ALL	GO:0007267~cell-cell signaling	1.04E-06	5.32E-04
GOTERM_CC_ALL	GO:0030054~cell junction	3.12E-06	1.06E-04
GOTERM_BP_ALL	GO:0007268~synaptic transmission	8.45E-06	3.24E-03

GOTERM_BP_ALL	GO:0007154~cell communication	1.18E-05	4.01E-03
GOTERM_BP_ALL	GO:0050877~neurological system process	3.97E-05	1.10E-02
GOTERM_BP_ALL	GO:0019226~transmission of nerve impulse	4.67E-05	1.19E-02
GOTERM_BP_ALL	GO:0003008~system process	9.75E-05	1.85E-02
GOTERM_BP_ALL	GO:0044057~regulation of system process	1.01E-03	7.30E-02
GOTERM_BP_ALL	GO:0007610~behavior	1.06E-03	7.29E-02
GOTERM_BP_ALL	GO:0042391~regulation of membrane potential	5.31E-02	6.27E-01
Annotation Cluster 5	Enrichment Score: 4.747184695084769		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0045202~synapse	4.77E-15	2.10E-12
GOTERM_CC_ALL	GO:0030424~axon	1.80E-07	8.80E-06
GOTERM_CC_ALL	GO:0014069~postsynaptic density	8.74E-07	3.49E-05
GOTERM_CC_ALL	GO:0043005~neuron projection	2.57E-06	9.42E-05
GOTERM_CC_ALL	GO:0042995~cell projection	5.38E-06	1.69E-04
GOTERM_CC_ALL	GO:0033267~axon part	1.00E-04	2.75E-03
GOTERM_CC_ALL	GO:0044463~cell projection part	1.59E-04	4.12E-03
GOTERM_CC_ALL	GO:0030425~dendrite	1.10E-03	2.00E-02
GOTERM_CC_ALL	GO:0043679~nerve terminal	4.30E-02	3.62E-01
GOTERM_CC_ALL	GO:0043025~cell soma	5.43E-02	4.01E-01
GOTERM_CC_ALL	GO:0043197~dendritic spine	1.42E-01	6.62E-01
Annotation Cluster 6	Enrichment Score: 3.1609377923530992		
Category	Term	PValue	Benjamini
GOTERM_MF_ALL	GO:0022836~gated channel activity	1.14E-07	4.88E-05
GOTERM_MF_ALL	GO:0005216~ion channel activity	6.95E-07	1.49E-04
GOTERM_MF_ALL	GO:0015267~channel activity	7.59E-07	1.31E-04
GOTERM_MF_ALL	GO:0022832~voltage-gated channel activity	3.86E-06	4.74E-04
GOTERM_MF_ALL	GO:0005244~voltage-gated ion channel activity	3.86E-06	4.74E-04
GOTERM_MF_ALL	GO:0005261~cation channel activity	7.81E-06	8.39E-04
GOTERM_BP_ALL	GO:0006811~ion transport	8.29E-05	1.94E-02
GOTERM_MF_ALL	GO:0030955~potassium ion binding	2.45E-04	1.61E-02
GOTERM_MF_ALL	GO:0005249~voltage-gated potassium channel activity	7.86E-04	4.14E-02
GOTERM_CC_ALL	GO:0034702~ion channel complex	8.27E-04	1.72E-02
GOTERM_MF_ALL	GO:0005267~potassium channel activity	9.31E-04	4.60E-02
GOTERM_CC_ALL	GO:0005887~integral to plasma membrane	9.96E-04	1.89E-02
GOTERM_MF_ALL	GO:0005215~transporter activity	1.34E-03	6.18E-02
GOTERM_MF_ALL	GO:0015075~ion transmembrane transporter activity	2.06E-03	8.10E-02
GOTERM_MF_ALL	GO:0022857~transmembrane transporter	2.36E-03	8.82E-02

	activity		
Annotation Cluster 7	Enrichment Score: 3.05116145469033		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0050877~neurological system process	3.97E-05	1.10E-02
GOTERM_BP_ALL	GO:0003008~system process	9.75E-05	1.85E-02
GOTERM_BP_ALL	GO:0007610~behavior	1.06E-03	7.29E-02
GOTERM_BP_ALL	GO:0007613~memory	1.06E-03	7.17E-02
GOTERM_BP_ALL	GO:0007611~learning or memory	1.72E-03	1.02E-01
GOTERM_BP_ALL	GO:0050890~cognition	2.18E-03	1.19E-01
GOTERM_BP_ALL	GO:0007612~learning	2.67E-02	4.58E-01
Annotation Cluster 8	Enrichment Score: 2.7718503600877216		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0051046~regulation of secretion	1.85E-08	1.89E-05
GOTERM_BP_ALL	GO:0050804~regulation of synaptic transmission	2.94E-04	3.69E-02
GOTERM_BP_ALL	GO:0051588~regulation of neurotransmitter transport	5.02E-04	4.70E-02
GOTERM_BP_ALL	GO:0046928~regulation of neurotransmitter secretion	5.02E-04	4.70E-02
GOTERM_BP_ALL	GO:0044057~regulation of system process	1.01E-03	7.30E-02
GOTERM_BP_ALL	GO:0007612~learning	2.67E-02	4.58E-01
GOTERM_BP_ALL	GO:0051966~regulation of synaptic transmission, glutamatergic	6.14E-02	6.53E-01
GOTERM_BP_ALL	GO:0048168~regulation of neuronal synaptic plasticity	1.21E-01	8.04E-01
GOTERM_BP_ALL	GO:0048167~regulation of synaptic plasticity	1.26E-01	8.13E-01
GOTERM_BP_ALL	GO:0048169~regulation of long-term neuronal synaptic plasticity	1.60E-01	8.66E-01
Annotation Cluster 9	Enrichment Score: 2.7629140992115166		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0007267~cell-cell signaling	1.04E-06	5.32E-04
GOTERM_BP_ALL	GO:0007268~synaptic transmission	8.45E-06	3.24E-03
GOTERM_BP_ALL	GO:0007154~cell communication	1.18E-05	4.01E-03
GOTERM_BP_ALL	GO:0048489~synaptic vesicle transport	1.57E-02	3.54E-01
GOTERM_BP_ALL	GO:0001505~regulation of neurotransmitter levels	1.87E-02	3.88E-01
Annotation Cluster 10	Enrichment Score: 2.5594188277341177		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0051046~regulation of secretion	1.85E-08	1.89E-05
GOTERM_BP_ALL	GO:0046883~regulation of hormone secretion	1.05E-03	7.40E-02
GOTERM_BP_ALL	GO:0051048~negative regulation of	1.48E-03	9.21E-02

	secretion		
GOTERM_BP_ALL	GO:0051047~positive regulation of secretion	3.71E-03	1.59E-01
GOTERM_BP_ALL	GO:0002791~regulation of peptide secretion	6.56E-03	2.28E-01
GOTERM_BP_ALL	GO:0051050~positive regulation of transport	7.20E-03	2.35E-01
Annotation Cluster 11	Enrichment Score: 2.4611965954562742		
Category	Term	PValue	Benjamini
KEGG_PATHWAY	rno04080:Neuroactive ligand-receptor interaction	2.53E-05	3.67E-03
GOTERM_BP_ALL	GO:0030817~regulation of cAMP biosynthetic process	1.16E-04	2.07E-02
GOTERM_BP_ALL	GO:0030814~regulation of cAMP metabolic process	1.16E-04	2.07E-02
GOTERM_BP_ALL	GO:0030808~regulation of nucleotide biosynthetic process	2.54E-04	3.65E-02
GOTERM_BP_ALL	GO:0051051~negative regulation of transport	3.40E-04	3.94E-02
GOTERM_BP_ALL	GO:0030799~regulation of cyclic nucleotide metabolic process	3.65E-04	4.06E-02
GOTERM_BP_ALL	GO:0007186~G-protein coupled receptor protein signaling pathway	5.88E-04	5.33E-02
GOTERM_BP_ALL	GO:0006140~regulation of nucleotide metabolic process	7.12E-04	5.59E-02
GOTERM_BP_ALL	GO:0051048~negative regulation of secretion	1.48E-03	9.21E-02
GOTERM_BP_ALL	GO:0045761~regulation of adenylate cyclase activity	2.17E-03	1.20E-01
GOTERM_BP_ALL	GO:0031279~regulation of cyclase activity	2.17E-03	1.20E-01
GOTERM_BP_ALL	GO:0051339~regulation of lyase activity	3.01E-03	1.43E-01
GOTERM_BP_ALL	GO:0019932~second-messenger-mediated signaling	3.59E-03	1.61E-01
GOTERM_BP_ALL	GO:0019933~cAMP-mediated signaling	4.09E-03	1.67E-01
GOTERM_MF_ALL	GO:0004930~G-protein coupled receptor activity	4.42E-03	1.23E-01
GOTERM_BP_ALL	GO:0019935~cyclic-nucleotide-mediated signaling	7.14E-03	2.36E-01
GOTERM_BP_ALL	GO:0007194~negative regulation of adenylate cyclase activity	2.65E-02	4.58E-01
GOTERM_BP_ALL	GO:0031280~negative regulation of cyclase activity	2.65E-02	4.58E-01
GOTERM_BP_ALL	GO:0007631~feeding behavior	7.05E-02	6.78E-01
GOTERM_BP_ALL	GO:0007188~G-protein signaling, coupled to cAMP nucleotide second messenger	1.27E-01	8.14E-01

GOTERM_BP_ALL	GO:0007187~G-protein signaling, coupled to cyclic nucleotide second messenger	1.51E-01	8.57E-01
GOTERM_BP_ALL	GO:0007193~inhibition of adenylate cyclase activity by G-protein signaling	2.87E-01	9.53E-01
Annotation Cluster 12	Enrichment Score: 2.038710771650711		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0042734~presynaptic membrane	3.85E-04	9.37E-03
GOTERM_MF_ALL	GO:0008066~glutamate receptor activity	1.65E-02	3.01E-01
GOTERM_BP_ALL	GO:0007215~glutamate signaling pathway	2.03E-02	4.03E-01
KEGG_PATHWAY	rno04720:Long-term potentiation	5.43E-02	3.33E-01
Annotation Cluster 13	Enrichment Score: 1.9337071768017942		
Category	Term	PValue	Benjamini
GOTERM_CC_ALL	GO:0033267~axon part	1.00E-04	2.75E-03
GOTERM_CC_ALL	GO:0030673~axolemma	4.69E-03	7.37E-02
GOTERM_CC_ALL	GO:0032589~neuron projection membrane	1.73E-02	2.13E-01
GOTERM_CC_ALL	GO:0031253~cell projection membrane	2.29E-02	2.47E-01
Annotation Cluster 14	Enrichment Score: 1.8096942956298074		
Category	Term	PValue	Benjamini
GOTERM_BP_ALL	GO:0032501~multicellular organismal process	6.07E-04	5.34E-02
GOTERM_BP_ALL	GO:0048468~cell development	1.22E-03	7.99E-02
GOTERM_BP_ALL	GO:0030030~cell projection organization	2.00E-03	1.16E-01
GOTERM_BP_ALL	GO:0048518~positive regulation of biological process	2.27E-03	1.19E-01
GOTERM_BP_ALL	GO:0048666~neuron development	2.36E-03	1.21E-01

Benjamini: Benjamini multiple testing corrected p-value

6.8 Using Cytoscape

Appendix 6.8.1 Creating network interaction files from gene expression data and visualizing in Cytoscape.

Create network interaction file in R as follows.

1. Load a gene expression data matrix (rows representing genes and column representing samples) and subset / select data for a set of probe sets or genes for which network file will be created. For example,

```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
. . .
> load("~/R7-rma-wgcna_Nov_13.RData")# Load project data
> dim(y.top)  # gene expression data matrix

[1] 5674   19
>
> y.top[1:6,1:6] # view a portion of the data to verify
 R7_Y_C_01 R7_Y_C_09 R7_Y_C_11 R7_Y_C_25 R7_Y_C_26 R7_Y_C_27
A1cf   7.136422  7.144941  7.104992  7.227441  7.177077  7.312507
A2m    8.577820  8.654121  8.613125  8.403527  8.453800  8.633708
Aaas   8.203427  8.293036  8.274909  8.215817  8.229674  8.212953
Aacs   8.938654  9.027368  9.073445  8.999271  8.903219  8.983764
Aadat  7.847731  7.805649  7.906973  7.925234  8.082314  8.046692
Aamp  11.038918 11.058953 11.095012 10.979411 11.018721 11.009730
```

2. Create a gene list from the data

```
> allGenes = as.character(row.names(y.top))
>
> allGenes[1:10]
 [1] "A1cf"   "A2m"    "Aaas"   "Aacs"   "Aadat"  "Aamp"   "Aars"   "Aarsd1" "Aass"
[10] "Aatf"
>
> length(allGenes)
[1] 5674
```

3. Prepare a file to store data

```
> fileName = "Network_Int_Data"
>
```

```
> a = rep(fileName,length(allGenes))
>
> length(a)

[1] 5674
```

4. Call a WGCNA function visantPrepOverall(. . .) [REF miller etal 2010] to create network files. In the following example, it is instructed to generate 500 most highly connected interactions using a soft power of 6.

```
> visantPrepOverall(a, fileName, t(y.top), allGenes, 500, 6, signed = TRUE)
Network_Int_Data_connectivityOverall.csv written.
201 0.3975 6
Network_Int_Data_visantOverall.csv written.
>
```

5. The file ..._visantOverall.csv is imported in Cytoscape to visualize resulting network. However, this file needs to be prepared as follows.
6. In order to preserve some gene names, which are often confused as date and modified by Excel to read as date, the file ..._visantOverall.csv is opened first in a note pad program e.g. NotePad++ and saved as a text file, for example ..._visantOverall.txt.

   Note – Data can also be exported from R as tab delimited text file, from which desired columns can be loaded into Cytoscape. For data loaded from .csv file in Cytoscape, gene names are shown in the graphs as comma quoted.
7. Next, the file is imported in Excel using "comma separated values" option and saved as a tab delimited text file.
8. Optionally, a network file can also be created in .sif format from the ..._visantOverall.csv file (by keeping only the two interaction columns separated by a column in between that could be filled with specific interaction type e.g. protein-protein)

Import network interaction file into Cytoscape and visualize:

1. Start Cytoscape
2. Import network from file option in Cytoscape and load the ..._visantOverall.csv or the .sif format file following prompt or as File→ import → network → from file → filename.
3. Analyze network using the option from the tools menu.
4. Use layout "Edge weighted spring embedded"
5. Can use Cytoscape to control network nodes and edges using styles. For example,
   - Nodes sizes can be made proportional to the degree (of connectivity) to make hub genes bigger than others
   - Edge bundling can be used to clearly separate hubs/clusters
   - Opacity can be used to highlight important genes, etc.

- See Cytoscape wiki manuals and tutorials (http://wiki.cytoscape.org/Cytoscape_3/UserManual) for details.

Load differential expression data for network genes into Cytoscape.

1. Save data as tab delimited text file
2. Import data (e.g. differential expression values) file in Cytoscape as File→ import → Table → from File → filename.

   Note: While working with multiple networks in the same Cytoscape workspace, for example, when comparing networks, load data file separately for each network and associate only to that network, otherwise changing expression value based color for one network (e.g. control) will also automatically change for another (e.g. experimental).

Note: In the limma differential expression analysis for the aged ~ young,

- the design was A ~ Y
- so any + value means "Age Upregulated"
- so any - value means "Young Upregulated"
- so map color on networks to expr/logFC values as follows:
    - Young network:
        - - logFC value == Red (expr value was high in Young)
        - + logFC value == Green (expr value was down in Young)
    - Age network:
        - - logFC value == Green (expr value was down in Aged)
        - + logFC value == Red (expr value was high in Aged)

6.9 Module Overlap Tables

Appendix 6.9.1 R7-Y to R7-A overlap table showing the number of genes that matches between each pair of modules.

Young \ Aged	Black (pink)	Brown (black)	Green (brown)	grey	Red (red)	Turquoise (blue)	Yellow (cyan)
blue	54	**350**	39	499	28	40	5
brown	21	**308**	84	314	9	11	12
green	6	72	**104**	162	12	18	6
grey	113	159	113	660	51	218	5
red	11	21	91	142	**63**	13	0
turquoise	**135**	188	54	541	37	**173**	1
yellow	26	53	69	282	6	35	**260**

Appendix 6.9.2 R7-Y to R7-A overlap table showing the p-values of the matches between each pair of modules in the above table.

Young \ Aged	Black (pink)	Brown (black)	Green (brown)	grey	Red (red)	Turquoise (blue)	Yellow (cyan)
blue	9.57E-01	**2.98E-32**	1.00E+00	1.02E-02	9.62E-01	1.00E+00	1.00E+00
brown	1.00E+00	**5.41E-44**	1.10E-01	9.96E-01	1.00E+00	1.00E+00	1.00E+00
green	1.00E+00	7.68E-01	**1.31E-24**	9.11E-01	7.36E-01	1.00E+00	1.00E+00
grey	3.22E-04	1.00E+00	9.59E-01	2.59E-04	3.26E-01	4.20E-25	1.00E+00
red	9.98E-01	1.00E+00	1.18E-20	9.51E-01	**1.93E-29**	1.00E+00	1.00E+00
turquoise	**3.90E-15**	1.00E+00	1.00E+00	6.12E-02	7.86E-01	**2.42E-15**	1.00E+00
yellow	1.00E+00	1.00E+00	6.45E-01	1.00E+00	1.00E+00	1.00E+00	**9.10E-214**

Appendix 6.9.3 R7-Y vs. B8-Y.

Table: B8 young modules after matching their module names to the modules of R7.

black	blue	brown	greenyellow	grey	magenta	pink	purple	red	turquoise
200	492	467	154	937	178	197	154	447	400

Table: R7-Y to B8-Y overlap table showing the number of genes that matches between each pair of modules.

R7-Y \ B8-Y	black	blue	brown	greenyellow	grey	magenta	pink	purple	red	turquoise
blue	36	128	92	7	171	55	54	36	63	78
brown	**62**	90	**125**	7	127	20	34	9	60	29
green	13	41	**42**	9	64	5	13	2	29	19
red	3	13	11	7	68	4	9	6	**56**	**33**
turquoise	25	75	42	38	197	45	30	41	82	**142**
yellow	17	39	**80**	21	90	10	15	13	**85**	15

Table: R7-Y to B8-Y overlap table showing the p-values of the matches between each pair of modules in the above table.

R7-Y \ B8-Y	black	blue	brown	greenyellow	grey	magenta	pink	purple	red	turquoise
blue	7.77E-01	2.08E-04	5.57E-01	1.00E+00	9.31E-01	2.19E-04	5.25E-03	1.55E-01	1.00E+00	5.97E-01
brown	**1.11E-08**	4.15E-02	**1.18E-11**	1.00E+00	9.78E-01	9.62E-01	2.73E-01	1.00E+00	9.18E-01	1.00E+00
green	5.52E-01	5.41E-02	**1.65E-02**	6.87E-01	3.61E-01	9.93E-01	5.29E-01	1.00E+00	5.49E-01	9.54E-01
red	1.00E+00	1.00E+00	1.00E+00	8.00E-01	1.73E-02	9.94E-01	8.18E-01	8.92E-01	**4.35E-09**	**2.06E-02**
turquoise	9.98E-01	9.98E-01	1.00E+00	7.51E-02	1.43E-01	3.91E-02	9.62E-01	2.15E-02	8.08E-01	**3.25E-15**
yellow	8.70E-01	9.87E-01	**2.81E-06**	1.35E-01	8.91E-01	9.94E-01	9.42E-01	8.49E-01	**8.51E-09**	1.00E+00

Appendix 6.9.4 R7-Y vs. K9-Y.

Table: K9 young modules after matching their module names to the modules of R7.

black	blue	brown	green	grey	magenta	pink	purple	red	turquoise	yellow
149	249	543	343	687	48	136	28	155	633	167

Table: R7-Y to K9-Y overlap table showing the number of genes that matches between each pair of modules.

R7-Y \ K9-Y	black	blue	brown	green	grey	magenta	pink	purple	red	turquoise	yellow
blue	20	**77**	128	**94**	105	13	32	7	16	62	25
brown	12	24	**159**	**67**	71	11	11	2	11	44	23
green	12	10	32	**31**	50	2	8	3	7	35	10
red	3	14	12	6	56	0	11	3	**24**	**78**	6
turquoise	26	46	39	52	181	4	44	6	41	**195**	15
yellow	25	17	**97**	27	69	5	4	2	19	57	**59**

Table: R7-Y to K9-Y overlap table showing the p-values of the matches between each pair of modules in the above table.

R7-Y \ K9-Y	black	blue	brown	green	grey	magenta	pink	purple	red	turquoise	yellow
blue	9.62E-01	**4.69E-07**	5.79E-04	**9.75E-06**	9.94E-01	9.00E-02	7.67E-02	2.48E-01	9.98E-01	1.00E+00	9.05E-01
brown	9.91E-01	9.86E-01	**6.18E-26**	**1.26E-03**	9.99E-01	5.94E-02	9.88E-01	9.17E-01	9.98E-01	1.00E+00	5.50E-01
green	2.38E-01	9.65E-01	7.22E-01	**2.53E-02**	1.56E-01	8.21E-01	6.47E-01	2.63E-01	8.77E-01	8.57E-01	6.31E-01
red	9.98E-01	8.13E-01	1.00E+00	1.00E+00	6.61E-02	1.00E+00	3.16E-01	2.95E-01	**8.27E-05**	**5.62E-09**	9.76E-01
turquoise	8.66E-01	8.36E-01	1.00E+00	9.98E-01	2.98E-05	9.94E-01	7.50E-04	5.36E-01	4.58E-02	**7.74E-12**	1.00E+00
yellow	5.42E-02	9.98E-01	**1.13E-05**	9.97E-01	9.77E-01	7.10E-01	1.00E+00	8.71E-01	5.21E-01	9.98E-01	**7.80E-16**

Appendix 6.9.5 R7-Y vs. B8-A.

Table: B8 aged modules after matching their module names to the modules of R7.

black	blue	brown	green	grey	pink	red	turquoise	yellow
347	109	194	348	1534	31	471	302	290

Table: R7-Y to B8-A overlap table showing the number of genes that matches between each pair of modules.

R7-Y \ B8-A	black	blue	brown	green	grey	pink	red	turquoise	yellow
blue	**97**	**40**	24	83	292	0	55	65	64
brown	48	8	**83**	**106**	229	3	28	12	46
green	17	5	11	**35**	103	2	29	15	20
red	8	5	1	5	99	1	**59**	18	14
turquoise	74	31	5	18	268	9	**154**	**111**	47
yellow	21	2	**37**	42	185	5	26	9	**58**

Table: R7-Y to B8-A overlap table showing the p-values of the matches between each pair of modules in the above table.

R7-Y \ B8-A	black	blue	brown	green	grey	pink	red	turquoise	yellow
blue	**8.20E-05**	**2.48E-05**	9.98E-01	3.10E-02	8.65E-01	1.00E+00	1.00E+00	2.45E-01	1.81E-01
brown	8.40E-01	9.97E-01	**5.68E-21**	**8.40E-14**	8.15E-01	8.81E-01	1.00E+00	1.00E+00	4.62E-01
green	9.26E-01	8.51E-01	7.35E-01	**5.42E-03**	3.80E-01	6.11E-01	6.70E-01	9.03E-01	4.35E-01
red	1.00E+00	7.67E-01	1.00E+00	1.00E+00	8.27E-02	8.44E-01	**1.42E-09**	4.86E-01	8.04E-01
turquoise	2.42E-01	1.73E-02	1.00E+00	1.00E+00	9.99E-01	1.42E-01	**6.14E-13**	**4.51E-13**	9.55E-01
yellow	9.99E-01	1.00E+00	**2.12E-04**	2.01E-01	9.35E-03	2.28E-01	1.00E+00	1.00E+00	**5.51E-07**

Appendix 6.9.6 R7-Y vs. B7-A.

Table: B7 aged modules after matching their module names to the modules of R7.

black	blue	brown	green	grey	magenta	pink	purple	yellow
230	100	298	234	714	118	174	115	157

Table: R7-Y to B7-A overlap table showing the number of genes that matches between each pair of modules.

R7-Y \ B7-A	black	blue	brown	green	grey	magenta	pink	purple	yellow
blue	37	**28**	**75**	30	110	19	28	6	12
brown	28	12	**69**	21	106	12	19	13	12
green	17	3	20	**25**	42	7	8	6	8
red	17	4	8	**25**	54	9	10	5	6
turquoise	46	**25**	33	45	135	23	32	15	24
yellow	30	3	43	33	114	12	35	**37**	**54**

Table: R7-Y to B7-A overlap table showing the p-values of the matches between each pair of modules in the above table.

R7-Y \ B7-A	black	blue	brown	green	grey	magenta	pink	purple	yellow
blue	5.37E-01	**1.46E-03**	**9.23E-06**	9.42E-01	7.57E-01	5.43E-01	5.39E-01	1.00E+00	1.00E+00
brown	7.83E-01	7.32E-01	**9.38E-07**	9.92E-01	1.41E-01	9.02E-01	8.89E-01	8.12E-01	9.95E-01
green	2.87E-01	9.60E-01	4.32E-01	**5.04E-03**	7.65E-01	6.34E-01	8.79E-01	7.53E-01	7.97E-01
red	3.09E-01	8.98E-01	9.99E-01	**6.15E-03**	8.31E-02	3.49E-01	7.00E-01	8.77E-01	9.50E-01
turquoise	1.85E-01	**3.71E-02**	1.00E+00	2.79E-01	1.57E-01	3.33E-01	4.30E-01	9.32E-01	8.20E-01
yellow	9.62E-01	1.00E+00	9.04E-01	9.04E-01	8.02E-01	9.87E-01	1.39E-01	**2.55E-05**	**2.06E-08**

Appendix 6.9.7 R7-Y to B8-Y percentage overlap table.

R7 Young Modules (R7-Y)	Total in R7-Y	R7-Y module matched to B8-Y modules	Maximum gene shared with B8-Y all modules	Total genes matched to the best matched module	p-value	Overlap (%)
brown	759	black, brown	563	125	1.20E-11	22.20
yellow	731	brown, red	385	85	8.50E-09	22.08
turquoise	1129	turquoise, greenyellow	717	142	3.30E-15	19.80
blue	1015	black, brown	720	92	5.60E-01	12.78
green	380	brown	237	42	1.70E-02	17.72
red	341	red, turquoise	210	56	4.40E-09	26.67

Appendix 6.9.8 R7-Y to K9-Y percentage overlap table.

R7 Young Modules (R7-Y)	Total in R7-Y	R7-Y module matched to K9-Y modules	Maximum gene shared with K9-Y all modules	Total genes matched to the best matched module	p-value	Overlap (%)
brown	759	brown, green	435	159	6.18E-26	36.55
yellow	731	brown, yellow	381	59	7.80E-16	15.49
turquoise	1129	red, turquoise	649	195	7.74E-12	30.05
blue	1015	blue, green	579	77	4.70E-07	13.30
green	380	green	200	31	2.53E-02	15.50
red	341	Red	213	78	5.62E-09	36.62

Appendix 6.9.9 R7-Y to B8-A percentage overlap table.

Young Modules (R7-Y)	Total in R7-Y	R7-Y module matched to B8-A all modules	Maximum gene shared with B8-A all modules	Total genes matched to the best matched module	p-value	Overlap (%)
brown	759	brown, green	194	83	5.68E-21	42.78
yellow	731	yellow, brown	290	58	5.51E-07	20.00
turquoise	1129	turq, red, blue	302	111	6.14E-13	36.75
blue	1015	blue, green	109	40	2.48E-05	36.70
green	380	green	348	35	5.42E-03	10.06
red	341	red	471	59	1.42E-09	12.53

Appendix 6.9.10 R7-Y to B7-A percentage overlap table.

Young Modules (R7-Y)	Total in R7-Y	R7-Y module matched to B7-A all modules	Maximum gene shared with B7-A all modules	Total genes matched to the best matched module	p-value	Overlap (%)
brown	759	brown	298	69	9.38E-07	23.15
yellow	731	yellow, purple	157	54	2.06E-08	34.39
turquoise	1129	blue	100	25	3.72E-02	25.00
blue	1015	blue	100	28	1.46E-03	28.00
green	380	green	234	25	5.00E-03	10.68
red	341	green	138	25	6.15E-03	18.12

6.10 Meta-Analysis of the ASLI Candidate Hub Genes

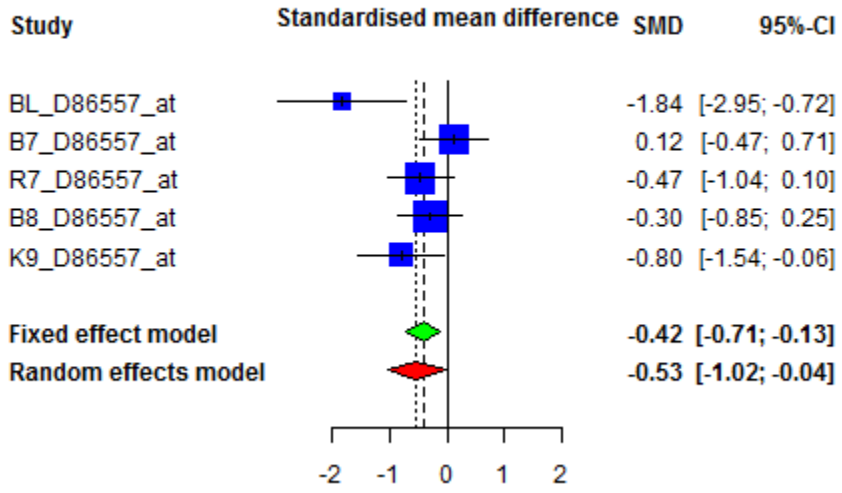
Appendix 6.10.1 Effect size estimates of top candidate ASLI hub genes in R7 (yellow) "learning and memory" module.

Gene Symbol	Probe ID (RGU34A or RAE230A)	Rat Gene ID	Number Of Study	ES	z-value	p-value of z-value	pBH of z-value	Q value	p-value of Q	tau.2	I2
Camk1g	D86557_at	171358	5	-0.53	-2.10	0.04	0.19	10.63	0.03	0.19	62.40
Cdk5r1	rc_AA850669_at	116671	2	-0.66	-2.21	0.03	0.17	1.22	0.27	0.04	18.10
Cntn1	D38492_at	117258	5	0.36	2.09	0.04	0.20	5.46	0.24	0.04	26.70
Dlg3	1388280_a_at	58948	3	-0.69	-1.92	0.06	0.25	7.35	0.03	0.28	72.80
Dlgap1	U67987_s_at	65040	5	-0.12	-0.47	0.64	0.80	12.35	0.01	0.23	67.60
Dpp6	M76426_at	29272	5	-0.42	-0.88	0.38	0.64	36.78	0.00	0.99	89.10
Eif5	rc_AI012604_at	56783	5	0.42	2.05	0.04	0.21	7.41	0.12	0.09	46.00
Gabrg1	X57514_at	140674	5	0.22	1.32	0.19	0.46	5.25	0.26	0.03	23.90
Impact	1375310_at	497198	3	0.41	2.28	0.02	0.15	0.53	0.77	0.00	0.00
Kcnab2	X76724_at	29738	5	0.04	0.29	0.77	0.88	3.61	0.46	0.00	0.00
Mapk1	1398346_at	116590	3	-0.55	-1.47	0.14	0.41	7.98	0.02	0.31	74.90
Mapre1	1375525_at	114764	3	-0.41	-2.28	0.02	0.15	1.04	0.60	0.00	0.00
Ndfip2	1389364_at	361089	3	-0.38	-1.22	0.22	0.50	5.84	0.05	0.19	65.70
Ppp2r2c	D38261_at	117256	5	-0.43	-1.24	0.22	0.50	20.01	0.00	0.46	80.00
Prkacb	D10770_s_at	293508	5	-0.12	-0.53	0.60	0.78	9.36	0.05	0.15	57.30

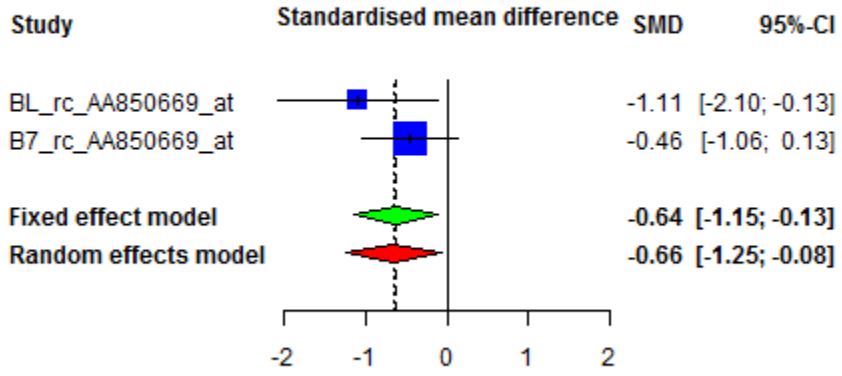
Pten	rc_AA963447_at	50557	5	-0.37	-2.58	0.01	0.09	1.49	0.83	0.00	0.00
Rasgrp1	AF060819_s_at	29434	5	-0.50	-1.20	0.23	0.51	28.63	0.00	0.73	86.00
Scn2b	U37147_at	25349	5	-0.28	-1.84	0.07	0.28	4.21	0.38	0.01	4.90
Stxbp1	1370840_at	25558	3	-0.32	-1.79	0.07	0.29	0.57	0.75	0.00	0.00

Legends: ES, effect size; pBH, p-value with Benjamini and Hochberg correction; FC, fold change; DE, differentially expressed; Q = Cochran's Q test for significant heterogeneity; I^2 = Ratio of true heterogeneity to total variation.

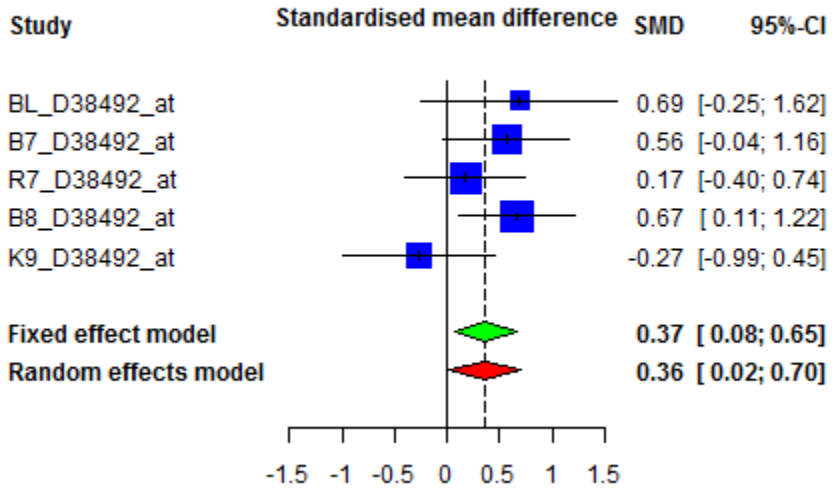
Appendix 6.10.2 Forest plot of *Camk1g*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



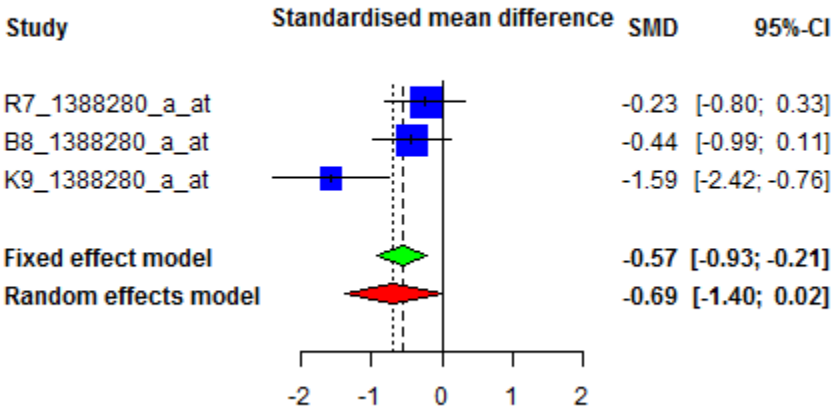
Appendix 6.10.3 Forest plot of *Cdk5r1*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



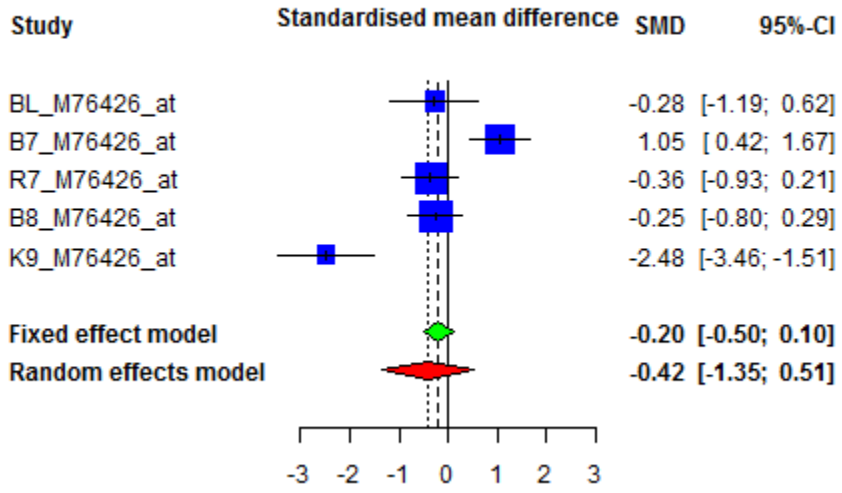
Appendix 6.10.4 Forest plot of *Cntn1*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



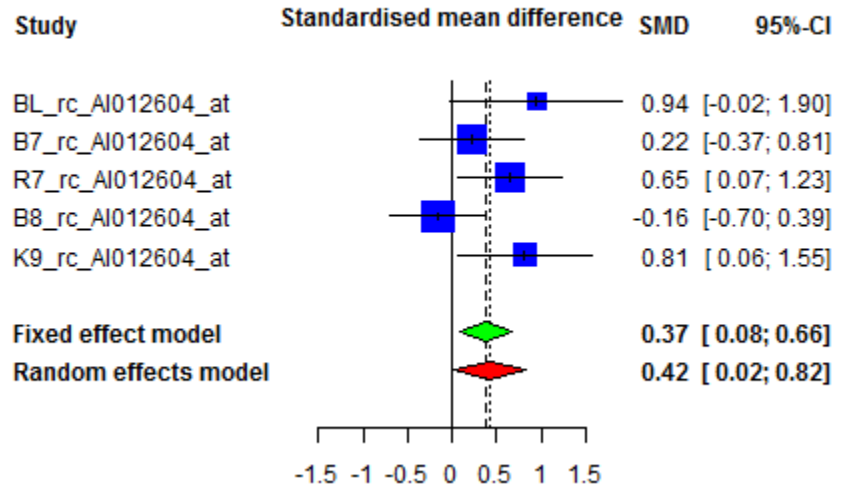
Appendix 6.10.5 Forest plot of *Dlg3*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



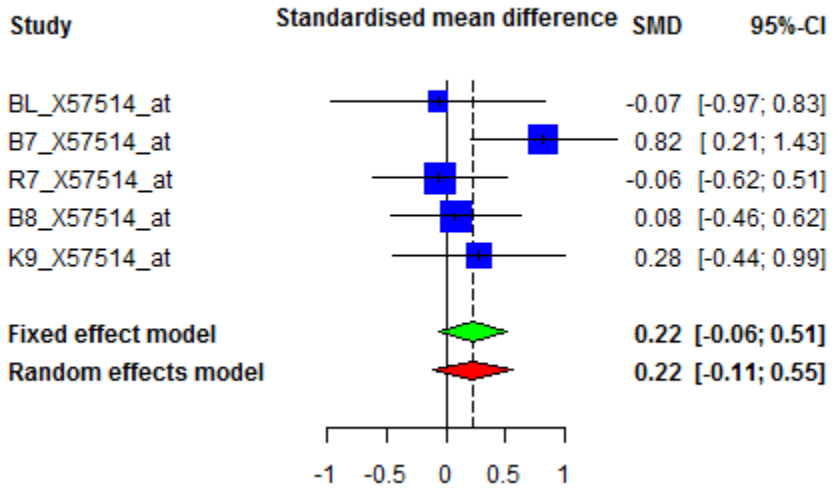
Appendix 6.10.6 Forest plot of *Dpp6*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



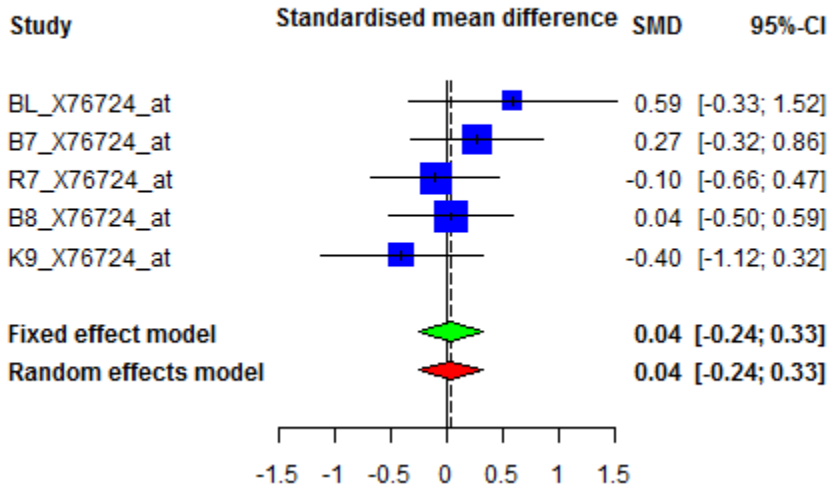
Appendix 6.10.7 Forest plot of *Eif5*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



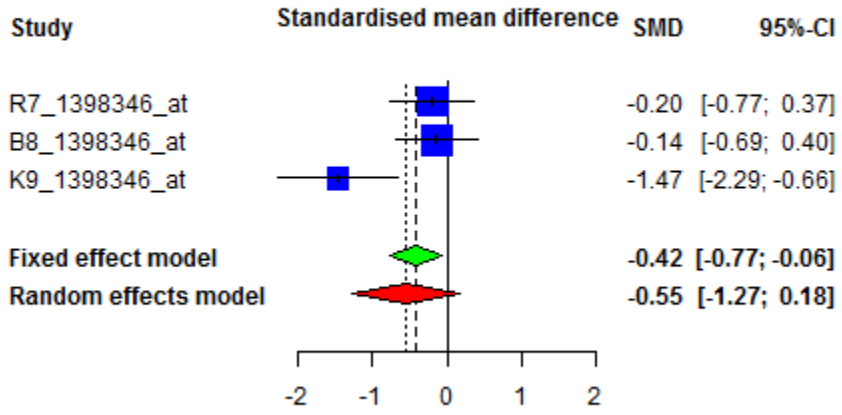
Appendix 6.10.8 Forest plot of *Gabrg1*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



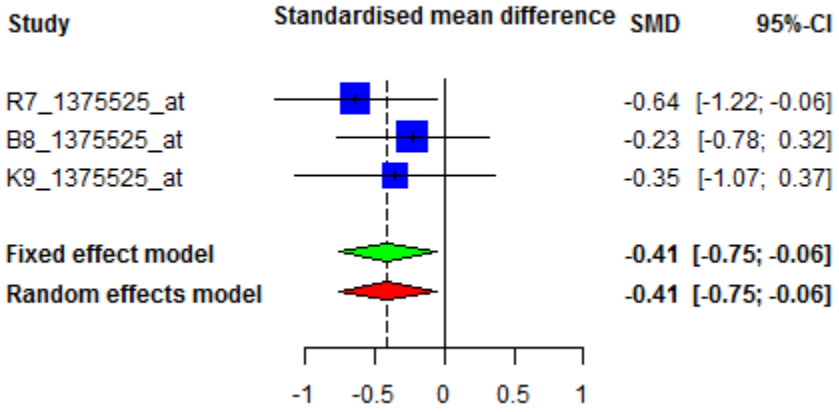
Appendix 6.10.9 Forest plot of *Kcnab2*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



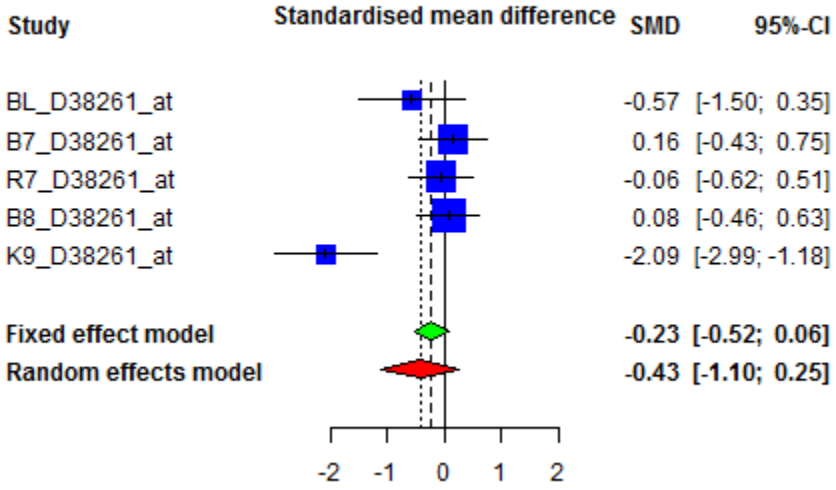
Appendix 6.10.10 Forest plot of *Mapk1*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



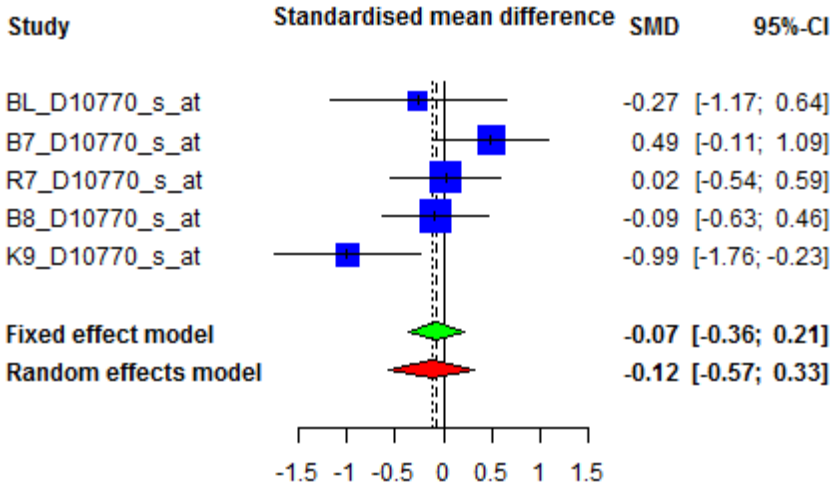
Appendix 6.10.11 Forest plot of *Mapre1*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



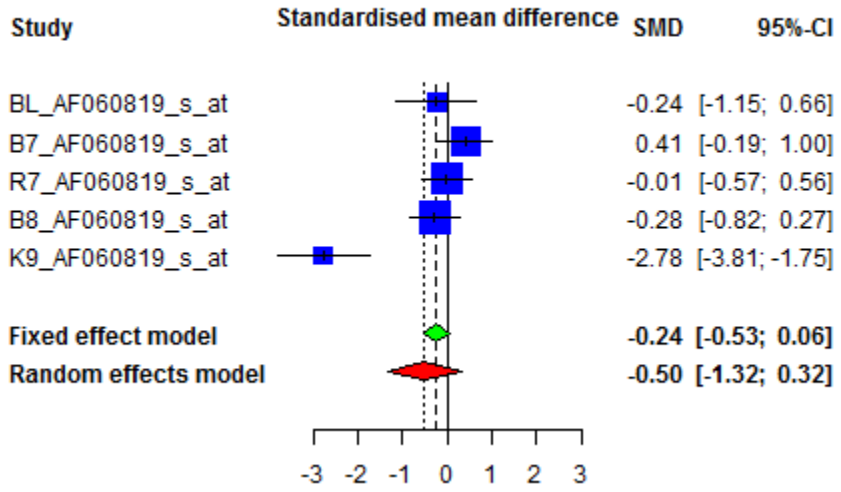
Appendix 6.10.12 Forest plot of *Ppp2r2c*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



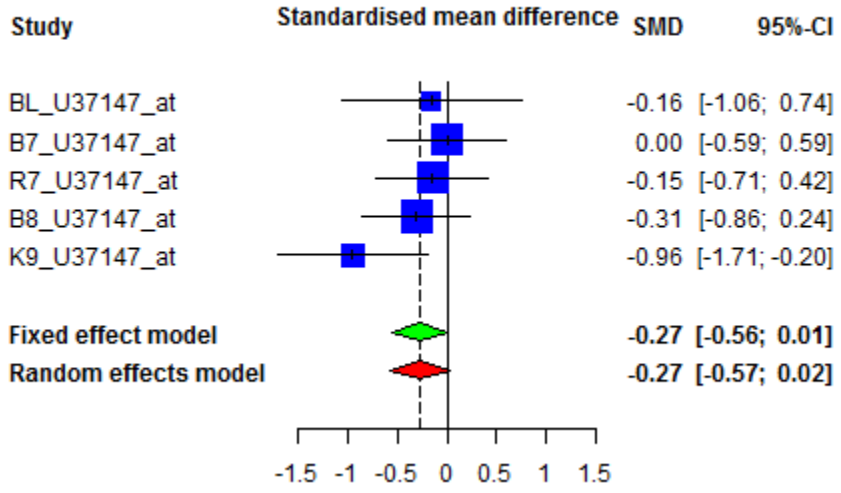
Appendix 6.10.13 Forest plot of *Prkacb*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



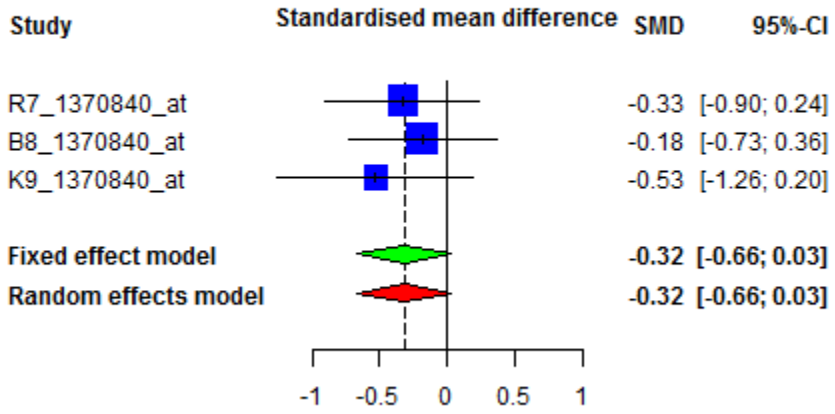
Appendix 6.10.14 Forest plot of *Rasgrp1*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



Appendix 6.10.15 Forest plot of *Scn2b*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



Appendix 6.10.16 Forest plot of *Stxbp1*. For the selected probe set for this gene the individual study specific SMDs and their 95% confidence intervals are plotted and shown on each row. The effect size results are shown at the bottom of the plot.



6.11 Validation of Hub Genes: Yellow Module

Appendix 6.11.1 Repeatability of young R7 yellow module hub genes in B8 young matching (red and brown) modules. Twenty hub genes with the highest mean $\boldsymbol{k_{IM}}$ are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B8 KME	Mean KME	t-test p-value
Dlgap1*	0.98	0.69	0.83	0.11
Fkbp1a	0.88	0.73	0.80	0.06
Rab3a	0.75	0.82	0.79	0.03
Ppp4r2	0.74	0.83	0.79	0.04
Xpr1	0.84	0.71	0.78	0.05
Glul	0.86	0.68	0.77	0.08
Dlg3*	0.99	0.51	0.75	0.20
Stxbp1*	0.96	0.52	0.74	0.18
Sri	0.86	0.61	0.74	0.10
Got1	0.78	0.68	0.73	0.04
Zfp292	0.49	0.96	0.73	0.20
Psme4	0.84	0.55	0.70	0.13
Cacng3	0.71	0.65	0.68	0.03
Mapre1*	0.99	0.35	0.67	0.28
Cnpy2	0.56	0.78	0.67	0.11
Nsf	0.95	0.36	0.65	0.27
Dpp6*	0.93	0.38	0.65	0.26
Arfgap1	0.62	0.67	0.65	0.03
Odc1	0.38	0.91	0.65	0.25
Pafah1b2	0.73	0.55	0.64	0.09

Appendix 6.11.2 Repeatability of young R7 yellow module hub genes in young K9 matching (brown and yellow) modules. Twenty hub genes with the highest mean k_{IM} are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	K9 KME	Mean KME	t-test p-value
Scn2b*	0.93	0.93	0.93	0.00
Prkacb*	1.00	0.85	0.92	0.05
Pclo	0.88	0.92	0.90	0.01
Dctn4	0.85	0.87	0.86	0.01
Cacnb4*	0.78	0.93	0.86	0.06
Ndfip2*	0.87	0.81	0.84	0.02
Mtpn	0.96	0.72	0.84	0.09
Cntn1*	0.86	0.81	0.83	0.02
Impact*	0.97	0.68	0.83	0.11
Dnal1	0.92	0.71	0.82	0.08
Pten*	0.78	0.86	0.82	0.03
G3bp2	0.88	0.75	0.81	0.05
Dnm1l	0.94	0.68	0.81	0.10
Trim23	0.89	0.69	0.79	0.08
Ranbp2	0.84	0.72	0.78	0.05
Akap6	0.95	0.60	0.77	0.14
Tmem30a	0.76	0.77	0.77	0.00
Fam91a1	0.59	0.94	0.77	0.14
Atf2	0.78	0.75	0.77	0.01
Arl1	0.81	0.70	0.76	0.05

Appendix 6.11.3 Repeatability of young R7 yellow module hub genes in young B8 matching (brown and red), and young K9 matching (brown and yellow) modules. Twenty hub genes with the highest mean $\boldsymbol{k_{IM}}$ are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B8 KME	K9 KME	Mean KME	t-test p-value
Ppp4r2	0.74	0.83	0.48	0.68	0.02
Tmf1	0.48	0.80	0.76	0.68	0.02
Klhl7	0.25	0.99	0.76	0.67	0.09
Tmem30a	0.76	0.46	0.77	0.66	0.02
Xpr1	0.84	0.71	0.39	0.65	0.04
Dnm1l	0.94	0.31	0.68	0.64	0.07
Papola	0.58	0.71	0.62	0.63	0.00
Cntn1*	0.86	0.20	0.81	0.62	0.10
Mapk1*	0.96	0.32	0.53	0.60	0.08
Pafah1b2	0.73	0.55	0.49	0.59	0.01
Gnai1	0.94	0.26	0.55	0.58	0.10
Kdm1b	0.31	0.73	0.69	0.57	0.05
Nlgn1	0.59	0.58	0.51	0.56	0.00
Tm2d1	0.32	0.59	0.75	0.56	0.05
Gpm6b	0.64	0.38	0.55	0.52	0.02
Tardbp	0.50	0.75	0.31	0.52	0.05
Fbxo8	0.56	0.38	0.53	0.49	0.01
Prepl	0.96	0.24	0.26	0.49	0.18
Pggt1b	0.84	0.21	0.40	0.48	0.12
Rragd	0.63	0.55	0.27	0.48	0.05

Appendix 6.11.4 Repeatability of aged R7 yellow module hub genes in B8 aged matching (red and brown) modules. Twenty hub genes with the highest mean $\boldsymbol{k_{IM}}$ are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B8 KME	Mean KME	t-test p-value
Stxbp1*	0.95	0.70	0.83	0.10
Dlgap1*	0.75	0.86	0.81	0.04
Psmd8	0.59	0.93	0.76	0.14
Nsf	0.90	0.58	0.74	0.13
Dpp6*	0.94	0.48	0.71	0.20
Glul	0.75	0.67	0.71	0.04
Zfp706	0.61	0.76	0.69	0.07
Tmem30a	0.82	0.54	0.68	0.13
Ptk2b	0.56	0.75	0.66	0.10
Gpm6b	0.34	0.92	0.63	0.27
Zfp238	0.77	0.48	0.63	0.14
Rab3a	0.61	0.65	0.63	0.02
Rac1	0.77	0.49	0.63	0.14
Ube2l3	0.69	0.57	0.63	0.06
Thy1	0.67	0.57	0.62	0.05
Rnf4	0.85	0.37	0.61	0.24
Trim9	0.66	0.56	0.61	0.05
Skp1	0.52	0.68	0.60	0.08
Vps52	0.63	0.56	0.60	0.04
Pkia	0.62	0.56	0.59	0.04

Appendix 6.11.5 Repeatability of aged R7 yellow module hub genes in B7 aged matching (purple and yellow) modules. Twenty hub genes with the highest mean k_{IM} are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B7 KME	Mean KME	t-test p-value
Nek9	0.77	0.82	0.79	0.02
Il1rap	0.79	0.77	0.78	0.01
Lyst	0.83	0.69	0.76	0.06
Prkacb*	1.00	0.45	0.72	0.23
Kit	0.94	0.46	0.70	0.21
Camk1g*	0.70	0.70	0.70	0.00
Gpam	0.71	0.67	0.69	0.02
Lgr4	0.75	0.62	0.68	0.06
B3gat1	0.77	0.58	0.67	0.09
Mapk1*	0.94	0.40	0.67	0.24
Akap1	0.67	0.66	0.67	0.00
Zfp706	0.61	0.72	0.67	0.05
Gabbr1	0.83	0.49	0.66	0.16
Atp1b2	0.75	0.55	0.65	0.09
Grm7	0.78	0.51	0.65	0.13
Kcnq3	0.78	0.51	0.65	0.13
Bmp3	0.55	0.70	0.62	0.08
Bcl2l1	0.71	0.54	0.62	0.09
Nlgn3	0.73	0.51	0.62	0.11
Tef	0.87	0.37	0.62	0.24

Appendix 6.11.6 Repeatability of aged R7 yellow module hub genes in B7 aged matching (purple and yellow) modules and aged B8 matching (brown and yellow) modules. Hub genes with the highest mean k_{IM} are shown. Candidate hub genes are marked by an '*' beside them. There were only five hub genes that were common among the three networks.

Hub Gene	R7 KME	B7 KME	B8 KME	Mean KME	t-test p-value
Zfp706	0.61	0.72	0.76	0.70	0.00
Dlgap1*	0.75	0.43	0.86	0.68	0.03
Sqstm1	0.46	0.70	0.31	0.49	0.05
Grb2	0.23	0.43	0.52	0.39	0.04
Zfp386	0.22	0.23	0.52	0.32	0.08

6.12 Validation of Hub Genes: Brown Module

Appendix 6.12.1 Repeatability of young R7 brown module hub genes in young B8 matching (black and brown) modules. Twenty hub genes with the highest mean $\boldsymbol{k_{IM}}$ are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B8 KME	Mean KME	t-test p-value
Rpe*	0.89	0.90	0.89	0.00
Zhx1	0.76	1.00	0.88	0.09
Ate1	0.71	1.00	0.85	0.11
Myo5b	0.79	0.91	0.85	0.05
Phf20l1	0.76	0.86	0.81	0.04
B3galnt1	0.88	0.67	0.78	0.09
Ctsl1	0.82	0.73	0.77	0.04
Aff4	0.80	0.74	0.77	0.03
Cast	0.87	0.67	0.77	0.08
Ist1	0.73	0.80	0.77	0.03
Slc6a15	0.69	0.81	0.75	0.05
Eif5b	0.69	0.77	0.73	0.03
Tmem161b	0.88	0.56	0.72	0.14
Snx2	0.46	0.97	0.72	0.22
Fpgt	0.66	0.73	0.70	0.03
Sucla2	0.54	0.85	0.70	0.14
Fbxo3	0.79	0.60	0.69	0.09
Tox4	0.66	0.72	0.69	0.03
Chchd3	0.76	0.62	0.69	0.07
Gng10	0.78	0.57	0.68	0.10

Appendix 6.12.2 Repeatability of young R7 brown module hub genes in young K9 matching (brown and green) modules. Twenty hub genes with the highest mean k_{IM} are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	K9 KME	Mean KME	t-test p-value
Rpe*	0.89	0.87	0.88	0.01
Fmr1	0.89	0.82	0.85	0.03
Tcerg1	0.69	1.00	0.85	0.11
Mtmr6	0.86	0.80	0.83	0.02
Aff4	0.80	0.85	0.82	0.02
Btbd1	0.82	0.81	0.81	0.01
Fyttd1	0.92	0.63	0.77	0.12
Ndufaf4	0.88	0.66	0.77	0.09
Psip1	0.65	0.86	0.75	0.09
LOC100912470	0.76	0.75	0.75	0.01
Glyr1	0.77	0.73	0.75	0.02
Taf9b	0.66	0.83	0.75	0.07
Arpc5	0.53	0.94	0.74	0.17
B3galnt1	0.88	0.59	0.74	0.13
Hars	0.67	0.80	0.73	0.06
Tm9sf2	0.87	0.60	0.73	0.12
Wdr13	0.72	0.74	0.73	0.01
Rapgef4	0.79	0.66	0.73	0.06
Fkbp3	0.75	0.70	0.72	0.02
Tmem161b	0.88	0.57	0.72	0.13

Appendix 6.12.3 Repeatability of young R7 brown module hub genes in matching young B8 black and brown, and K9 brown and green modules. Twenty hub genes with the highest mean $\boldsymbol{k_{IM}}$ are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B8 KME	K9 KME	Mean KME	t-test p-value
Rpe*	0.89	0.90	0.87	0.88	0.00
Zhx1	0.76	1.00	0.68	0.81	0.01
Aff4	0.80	0.74	0.85	0.80	0.00
Ate1	0.71	1.00	0.61	0.77	0.02
Eif5b	0.69	0.77	0.73	0.73	0.00
B3galnt1	0.88	0.67	0.59	0.71	0.01
Slc6a15	0.69	0.81	0.63	0.71	0.01
LOC100910334	0.48	0.77	0.86	0.71	0.03
Ctsl1	0.82	0.73	0.54	0.69	0.01
Glyr1	0.77	0.57	0.73	0.69	0.01
Fkbp3	0.75	0.58	0.70	0.68	0.01
Tmem161b	0.88	0.56	0.57	0.67	0.02
Fpgt	0.66	0.73	0.61	0.67	0.00
Phf20l1	0.76	0.86	0.38	0.66	0.05
Naa38	0.55	0.60	0.85	0.66	0.02
Hnrpd	0.45	0.85	0.67	0.66	0.03
Etfa	0.59	0.62	0.70	0.64	0.00
Cnksr2	0.70	0.63	0.57	0.63	0.00
Trappc6b	0.55	0.76	0.58	0.63	0.01
Fyttd1	0.92	0.30	0.63	0.61	0.08

Appendix 6.12.4 Repeatability of aged R7 brown module hub genes in aged B8 matching (brown and green) modules. Twenty hub genes with the highest mean $\boldsymbol{k_{IM}}$ are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B8 KME	Mean KME	t-test p-value
Araf	1.00	0.78	0.89	0.08
Spag9	0.98	0.75	0.87	0.09
Ndufs1	0.75	0.96	0.85	0.08
Slc6a15	0.93	0.73	0.83	0.07
Spast	0.90	0.73	0.81	0.07
Mtmr6	0.69	0.93	0.81	0.10
MGC112830	0.82	0.78	0.80	0.01
Casc4	0.87	0.72	0.80	0.06
Ccdc104	0.78	0.79	0.79	0.00
RGD1309995	0.67	0.91	0.79	0.10
Dars	0.67	0.88	0.77	0.08
Aff4	0.78	0.77	0.77	0.00
Pkn2	0.55	1.00	0.77	0.18
Slc30a5	0.79	0.75	0.77	0.02
Fmr1	0.90	0.63	0.77	0.11
Rapgef4	0.93	0.60	0.76	0.14
Rpl4	0.81	0.71	0.76	0.04
Psip1	0.77	0.74	0.75	0.02
Eif3m	0.83	0.67	0.75	0.07
Foxg1	0.78	0.70	0.74	0.03

Appendix 6.12.5 Repeatability of aged R7 brown module hub genes in aged B7 brown module. Twenty hub genes with the highest mean k_{IM} are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B7 KME	Mean KME	t-test p-value
Cct3	0.80	0.93	0.86	0.05
Slc6a15	0.93	0.67	0.80	0.10
Aifm1	0.74	0.85	0.80	0.04
Eif3m	0.83	0.73	0.78	0.04
Hdac2	0.69	0.81	0.75	0.05
Hdhd2	0.62	0.86	0.74	0.10
Etfa	0.70	0.76	0.73	0.03
Dync1i2	0.63	0.77	0.70	0.06
Calm2	0.58	0.81	0.69	0.10
Ctsl1	0.81	0.56	0.69	0.12
Spast	0.90	0.45	0.68	0.20
Ivns1abp	0.44	0.88	0.66	0.20
Foxg1	0.78	0.53	0.65	0.12
Trim32	0.70	0.60	0.65	0.05
Rap1b	0.61	0.68	0.64	0.03
Naca	0.70	0.56	0.63	0.07
Sec62	0.56	0.69	0.62	0.07
Psma1	0.63	0.61	0.62	0.01
Tspan3	0.76	0.46	0.61	0.15
Zranb2	0.65	0.54	0.59	0.06

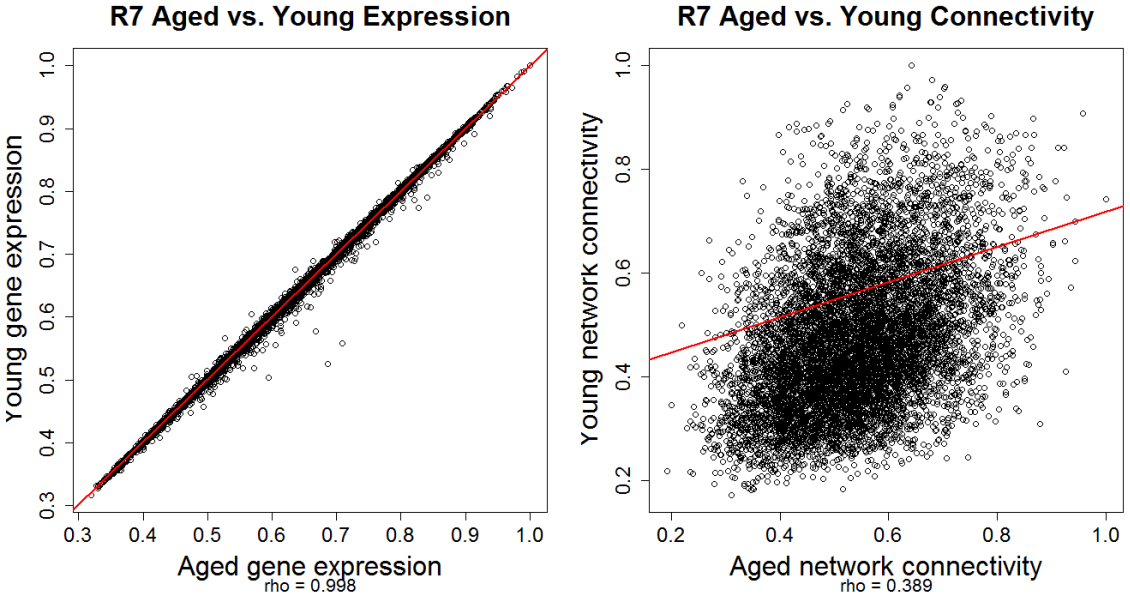
Appendix 6.12.6 Repeatability of aged R7 brown module hub genes in matching aged B7 brown module, and B8 brown and green modules. Twenty hub genes with the highest mean k_{IM} are shown. Candidate hub genes are marked by an '*' beside them.

Hub Gene	R7 KME	B7 KME	B8 KME	Mean KME	t-test p-value
Slc6a15	0.93	0.67	0.73	0.78	0.01
Eif3m	0.83	0.73	0.67	0.74	0.00
Spast	0.90	0.45	0.73	0.69	0.03
Hdac2	0.69	0.81	0.58	0.69	0.01
Foxg1	0.78	0.53	0.70	0.67	0.01
Ctsl1	0.81	0.56	0.62	0.66	0.01
Etfa	0.70	0.76	0.50	0.65	0.01
Trappc6b	0.56	0.58	0.80	0.65	0.01
Ppp1cb	0.62	0.45	0.77	0.62	0.02
Rap1b	0.61	0.68	0.55	0.61	0.00
Cnbp	0.58	0.43	0.81	0.61	0.03
Hook1	0.59	0.36	0.84	0.60	0.05
Psma1	0.63	0.61	0.54	0.59	0.00
Dync1i2	0.63	0.77	0.37	0.59	0.04
Naca	0.70	0.56	0.45	0.57	0.02
Gnl3	0.61	0.56	0.53	0.56	0.00
Ivns1abp	0.44	0.88	0.34	0.56	0.08
Map2k1	0.47	0.56	0.64	0.56	0.01
Zranb2	0.65	0.54	0.47	0.55	0.01
Ate1	0.64	0.46	0.50	0.54	0.01

6.13 Differential expression vs. differential connectivity

Appendix 6.13.1: Comparing gene expression and connectivity between young and aged R7 samples using scatter plots.

Mean gene expressions for each gene across all aged and young arrays were calculated. The gene expression values were then scaled to lie between 0 and 1 by dividing them with the maximum mean expression. The young and aged mean scaled gene expression values were plotted in the scatter plot (on the left image below) along the y and x axis, respectively. Similarly, connectivity for each gene was calculated, scaled to lie between 0 and 1, and plotted on the right scatter plot. Spearman's rank correlation test (rho) was performed between young vs. aged gene expression as well as young vs. aged network connectivity. The results show that the overall gene expressions between the young and aged samples are highly correlated (rho = 0.998), which is not the case for the gene networks connectivity patterns between the same samples as the correlations between them are very weak (rho = 0.389). This observation highlights the fact that differential connectivity is not the same as differential expression.



6.14 Glossary of terms

Appendix 6.14.1 Glossary of terms used in the thesis.

Adjacency matrix: The connection strengths in an undirected network can be represented by an adjacency matrix, a symmetric matrix whose entries lie between 0 and 1. The element a_{ij} is the connection strength between nodes *i* and *j*. As a convention, the diagonal elements are set to 1, $a_{ij} = 1$ (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Affymetrix oligonucleotide microarray: In Affymetrix expression array, oligonucleotides of 25 base pairs in length are used to probe genes. There are two types of probes: reference probes that match a target sequence exactly, called the *perfect match* (PM), and partner probes which differ from the reference probes only by a single base in the center of the sequence, called the *mismatch* (MM) probes. Typically 16–20 of these probe pairs, each interrogating a different part of the sequence for a gene, make up what is known as a probe set. Some more recent arrays, such as the HG-U133 arrays, use as few as 11 probes in a probe set (Lipshutz et al., 1999; Warrington et al., 2000).

Co-expression (correlation) Network: Co-expression networks are undirected gene networks. The nodes of such a network correspond to genes (and their expression profiles), and edges between genes represent connection strengths (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Connection Strength: Connection strength between a pair of genes is determined by the pairwise correlations between their expression profiles (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Connectivity: For each gene, the connectivity is defined as the sum of connection strengths with the other network genes: $k_i = \sum_{u \neq i} a_{ui}$. In co-expression networks, the connectivity measures how correlated a gene is with all other network genes (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Data normalization: The term “normalization” as applied to microarray data does not refer to the normal (Gaussian) distribution, but instead it refers to the process of correcting two or more datasets prior to comparing their gene expression values (Pevsner, 2009).

Gene co-expression network: In gene co-expression networks, the nodes represent genes (or probe sets of a microarray) measured across a given set of microarray samples and the connections represent the strength of co-expression. Various measures of co-expression can be used, for example Pearson or robust correlation (in which case the co-expression network is also a correlation network), information-theoretic methods such as mutual information, and other measures of co-expression similarity (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Hub gene: This loosely defined term is used as an abbreviation of “highly connected gene". By definition, hub genes inside co-expression modules tend to have high connectivity (i.e. genes with many connections with other genes) (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Intramodular connectivity k_{IM}: Intramodular connectivity measures how connected or co-expressed a given gene is with respect to the genes of a particular module. The intramodular connectivity may be interpreted as a measure of module membership (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Mismatch (MM): Same as PM but with a single base change for the middle (13th) base. The purpose is to measure non-specific binding and background noise.

Module eigengene E: The module eigengene *E* is defined as the first principal component of a given module. It can be considered as a representative of the gene expression profiles in a module (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Module membership also known as eigengene-based connectivity k_{ME}: For each gene, a "fuzzy" measure of module membership is defined by correlating its gene expression profile with the module eigengene of a given module. For example, $MM^{blue}(i) = K_{cor,i}^{blue} = cor(X_i, E^{blue})$ measures how correlated gene *i* is to the blue module eigengene. $MM^{blue}(i)$ measures the membership of the *i*-th gene with respect to the blue module. If $MM^{blue}(i)$ is close to 0, the *i*-th gene is not part of the blue module. On the other hand, if $MM^{blue}(i)$ is close to 1 or -1, it is highly connected to the blue module genes. The sign of module membership encodes whether the gene has a positive or a negative relationship with the blue module eigengene. The module membership measure can be defined for all input genes (irrespective of their original module membership). It turns out that the module membership measure is highly related to the intramodular connectivity k_{IM}. Highly connected intramodular hub genes tend to have high module membership values to the respective module (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Module: Module consists of a group of genes whose expression profiles are highly correlated across the samples. Module is a type of sub-network which consists of clusters of highly interconnected genes. In an unsigned co-expression network, modules correspond to clusters of genes with high absolute correlations. In a signed network, modules correspond to positively correlated genes (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Network density: The mean adjacency (connection strength) among all nodes in a network (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

NP-hard problem: A problem is NP-hard if an algorithm for solving it can be translated into one for solving any NP-problem (nondeterministic polynomial time) problem. NP-hard therefore means "at least as hard as any NP-problem," although it might, in fact, be harder (Weisstein, 2015).

Perfect match (PM): A 25-mer probe complementary to a reference sequence of interest (e.g. part of a gene)

Probe: An oligonucleotide of 25 base-pairs ("25-mer") in length

Probe set: Typically 16–20 probe pairs, each interrogating a different part of the sequence for a gene, make up what is known as a probe set (Lipshutz et al., 1999; Warrington et al., 2000).

Scale-free network: Scale-free network has grown a lot of interest in recent years. The term scale-free refers to the distribution principle of how many links there are per node. The defining property of scale-free networks is that the probability that a node is connected with *k* other nodes decays as a power law distribution (Barabasi and Albert, 1999; Barabasi and Bonabeau, 2003; Jeong et al., 2000). For example, the probability distribution function $P(k)$ of the degree *k* of scale-free networks is described by $P(k) \approx k^{-\gamma}$. Many real world networks show the properties of scale-free network, for example, the physical structure of the internet (router level, domain level, and web links), social networks like e-mail networks, the structure of software modules, etc. Interestingly many biological networks such as yeast protein-protein interaction network (Carter et al., 2004; Han et al., 2004; Jeong et al., 2001) also demonstrate scale-free property. Scale-free networks are extremely heterogeneous, their topology being dominated by a few highly connected nodes (hubs) which link the rest of the less connected nodes to the system. (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Soft-threshold: Soft-threshold is a value that is used to raise the power of gene co-expression measures in weighted co-expression networks. It is determined in such a way so that the resulting network follows approximate scale free topology (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Sub-network: A subnetwork of a network can be any collection (subset) of nodes from the network, together with the adjacencies (connection strengths) between the nodes.

Thus, a subnetwork of a network also forms a (smaller) network on its own (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Topological overlap and topological overlap matrix (TOM): A major goal of gene correlation network analysis is to identify groups or "modules" of highly interconnected genes. Modules are groups of genes whose expression profiles are highly correlated across the samples. Modules are identified by searching for genes with similar patterns of connection strengths to other genes, or genes with high topological overlap. The topological overlap is a measure of node similarity. Topological overlap of two nodes reflects their relative interconnectedness (i.e. how close the neighbors of gene 1 are to the neighbors of gene 2). In order to identify network modules, generalized topological overlap matrix (GTOM) is calculated using the adjacency and connectivity values. The topological overlap values determine which genes will be in which module and form a network. The values range between 1 and 0 representing maximum and minimum interconnectedness. Module identification method in WGCNA is based on using node dissimilarity measure in conjunction with a clustering method. Since topological overlap is non-negative and symmetric, it is turned into a dissimilarity measure by subtracting from one. Genes are average linkage hierarchically clustered using 1-topological overlap as the distance measure and modules are determined by choosing a height cutoff for the resulting dendrogram. In the dendrogram, discrete branches of the tree correspond to modules of co-expressed genes (Langfelder and Horvath, 2008; Zhang and Horvath, 2005).

Weighted gene co-expression Network: Weighted gene co-expression network is created by raising the absolute value of the correlation between the expression profiles of a pair of genes to a power $\beta \geq 1$ (soft thresholding). This approach emphasizes high correlations at the expense of low correlations. Specifically, the function $a_{ij} = |cor(x_i, x_j)|^{\beta}$ represents the adjacency of an unsigned network. However, using the absolute value of the correlation may obscure biologically relevant information, since no distinction is made between gene repression and activation. In contrast, in signed

networks the similarity between genes reflects the sign of the correlation of their expression profiles. A simple transformation of the correlation is used and the adjacency is defined by the adjacency function $a_{ij} = [\frac{1}{2}\ (1\ +\ cor(x_i, x_j))]^{\beta}$. As in the unsigned measure, the signed similarity takes on a value between 0 and 1. Note that the unsigned similarity between two oppositely expressed genes $(cor(x_i, x_j) = -1)$ equals 1, while it equals to 0 for the signed similarity. Similarly, while the unsigned co-expression measure of two genes with zero correlation remains zero, the signed similarity equals to 0.5. (Langfelder and Horvath, 2008; Mason et al., 2009; Zhang and Horvath, 2005).

6.15 References

Barabasi, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science. 286**,** 509-12.

Barabasi, A.L., Bonabeau, E., 2003. Scale-free networks. Sci Am. 288**,** 60-9.

Carter, S.L., Brechbuhler, C.M., Griffin, M., Bond, A.T., 2004. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics. 20**,** 2242-50.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., et al., 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 430**,** 88-93.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., et al., 2000. The large-scale organization of metabolic networks. Nature. 407**,** 651-4.

Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature. 411**,** 41-2.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 9**,** 559.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., Lockhart, D.J., 1999. High density synthetic oligonucleotide arrays. Nat Genet. 21**,** 20-4.

Mason, M.J., Fan, G., Plath, K., Zhou, Q., et al., 2009. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics. 10**,** 327.

Pevsner, J., 2009. Gene Expression: Microarray Data Analysis. In: Bioinformatics and Functional Genomics. Vol., ed.^eds. Wiley.

Warrington, J.A., Nair, A., Mahadevappa, M., Tsyganskaya, M., 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. Physiol Genomics. 2**,** 143-7.

Weisstein, E.W., 2015. MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/NP-HardProblem.html. Vol. 2015, ed.^eds.

Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 4**,** Article17.

Raihan K. Uddin

Curriculum Vitae

Education

2010 – present	Ph.D. Candidate (part-time), Department of Biology, University of Western Ontario, London, ON, Canada
2003	Diploma in Computer Programmer Analyst, Fanshawe College, London, ON, Canada
2000	M. Sc. In Zoology (Molecular Biology), University of Western Ontario
1993	M. Sc. in Zoology, Jahangirnagar University, Savar, Bangladesh
1990	B. Sc. In Life Science (Zoology), Jahangirnagar University (Gold Medalist)

Honours and Awards

2002	Fanshawe College London Life Award, London, ON, Canada
1998 – 2000	President Special Graduate Scholarship and Special University Scholarship, University of Western Ontario
1990	B.Sc. University Best, Prime Minister's Gold Medal and Scholarship, Jahangirnagar University
1990	B.Sc. University Best, A.F.M Kamaluddin Gold Medal, Institute of Life Sciences, Jahangirnagar University.

Research Experience

2003 – 2006	Bioinformatics Associate, Department of Biology, University of Western Ontario

Publications

UDDIN R.K., Singh S.M. **(2013)**: Hippocampal gene expression meta-analysis identifies aging and age-associated spatial learning impairment (ASLI) genes and pathways. PLoS One. 2013 Jul 18;8(7).

UDDIN R.K., Singh S.M. **(2010)**: Metabolomics in drug response and addiction. Addictive disorder and substance abuse, edited by B.A. Johnson. Springer.

Singh S.M., Treadwell J., Kleiber M.L., Harrison M., **UDDIN K.R. (2007)** Analysis of behavior using genetical genomics in mice as a model: from alcohol preferences to gene expression differences. Genome, 50(10):877-97. Review.

UDDIN R.K., Singh S.M. **(2007)**: Ethanol-responsive genes: Identification of transcription factors and their role in metabolomics. Pharmacogenomics J., Feb; 7(1):38-47. Epub 2006 May 2.

UDDIN R.K., Zhang Y., Siu V.M., Fan Y-S., O'Reilly R.L., Rao J., Singh S.M. **(2006)**: Breakpoint Associated with a novel 2.3 Mb deletion in the VCFS region of 22q11 and the role of Alu (SINE) in recurring microdeletions. BMC Med Genet., Mar 2;7:18.

UDDIN R.K., Singh S.M. **(2006)**: Cis-regulatory sequences of the genes involved in apoptosis, cell growth and proliferation may provide a target for some of the effects of acute ethanol exposure. Brain Res., May 9;1088(1):31-44. Epub 2006 Apr 21.

Loney K.D., **UDDIN R.K.**, Singh S.M. **(2006):** Analysis of metallothionein brain gene expression in relation to ethanol preference in mice using cosegregation and gene knockouts. Alcoholism: Clin. Exp. Res., 30(1): 15-25.

UDDIN R.K., Treadwell J.A, Singh S.M. **(2005)**: Towards unraveling ethanol-specific neuro-metabolomics based on ethanol responsive genes in vivo. Neurochemical Research, 30(9): 1179-1190.

Loney K.D., **UDDIN R.K.**, Singh S.M. **(2003)**: Strain-specific brain metallothionein II (MT-II) gene expression, its ethanol responsiveness, and association with ethanol preference in mice. Alcoholism: Clin. Exp. Res., 27(3): 388-395.

Murphy B.C., Chiu T., Harrison M., **UDDIN R.K.**, Singh S. M. **(2002)**: Examination of ethanol responsive liver and brain specific gene expression, in the mouse strains with variable ethanol preferences, using cDNA expression arrays. Biochemical Genetics, 40(11/12): 395-410.

UDDIN R.K., Baqui M. A. **(1995)**: Effects of photoperiod on the wing dimorphism of Brown Planthopper (BPH) Nilaparvata Lugens (Stal). Bangladesh Journal of Entomology, 3(1 & 2): 95-98.

UDDIN R.K., Baqui M. A., Muhibullah M. **(1994)**: Effects of crowding and rearing on wing dimorphism of Brown Planthopper (BPH) *Nilaparvata Lugens* (Stal). Jahangirnagar University Journal of Life Sciences, 18: 193-203.

UDDIN R.K., Baqui M. A. **(1993)**: Some aspects of biology of Brown Planthopper (BPH) *Nilaparvata Lugens* (Stal) (Homoptera: Delphacidae). Bangladesh Journal of Life Sciences, 5(1): 59-64.

Teaching Experience

2006	Lecturer, Bio 461f – "From Genes to Genome", Department of Biology, University of Western Ontario
1998 – 2000	Teaching assistant, Bio 1001/1002 (Biology for Science) and Bio 2290 (Scientific Methods in Biology), Department of Biology, University of Western Ontario
1993 – 1997	Lecturer, animal biology and molecular cell biology courses, Jahangirnagar University, Bangladesh