Next Generation Sequencing in Disease-Relevant Tissues



Harry Clifford

St. Edmund Hall

University of Oxford

Thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2016

Abstract

Studies of RNA and the transcriptome are of great importance in providing functional information, and unravelling the genetic mechanisms that underlie complex disorders and diseases. With the vast majority of complex disease-associated variants falling outside protein-coding regions of the genome, it is likely that variations in gene expression regulation will be essential to understanding disease aetiology. Information on RNA quantity and splicing isoforms is therefore likely to be crucial for understanding complex pathologies of deleterious genetic variation. The advent of next generation sequencing has allowed the development of an assortment of technologies for interrogating aspects of the genome, one of which is high-throughput RNA sequencing (RNA-Seq). This technology allows rapid, relatively cheap, and accurate quantification of transcripts at a genome-wide scale. By providing a greater number of advantages and fewer caveats than alternative methods of transcriptome quantification, RNA-Seq is often described as a disruptive technology that is likely to supersede most others. Throughout this thesis, I have sought to demonstrate how these advantages assist in revealing significant and novel developmental, noncoding, coding, and alternative isoform information of relevance to disorders and diseases. I take advantage of methods that utilize the truly genome-wide coverage of RNA-Seq, that quantify large numbers of transcripts, and that interrogate novel splicing

events. More specifically, I present (i) the identification of novel biomarkers of the various placode-derived vertebrate cranial nerves, (ii) differential gene networks which highlight the genetics of autism intellectual disability co-morbidity, and (iii) differential gene expression underlying a form of severe influenza susceptibility. In addition to these studies, this thesis presents an R package for RNA-Seq time-series experiments, including functionality for efficient model-based clustering, and the integration of gene ontology information for cluster number selection and for subsequent profiling. Overall, this thesis demonstrates that RNA-Seq is a powerful tool for understanding disease aetiology.

Acknowledgements

Thanks to my supervisors Chris Ponting and Wilfried Haerty for their continuous assistance, advice, and ideas, for which I am forever grateful. I am very thankful for having had the opportunity to work with them and for all the time and effort they put into guiding and supporting me.

Thanks to my collaborators for their efforts in each of these projects, for generating the data for analysis, and for validation and/or interpretation of results. None of these projects would have been possible without their help. For this, I thank Cedric Patthey, Sebastian Shimeld, Jo Begbie, Anna Dulneva, Esther Becker, Lili Wang, Shokouh Makvandi-Nejad and Tao Dong. I also thank Andreas Heger, Grant Belgard, and everyone in the Ponting research group for their technical support and useful discussions.

Thanks to the Medical Research Council for primary funding and to St. Edmund Hall for additional funding and a great college environment.

Finally, thanks to Elizabeth Emsley, Gary Clifford, Matthew Clifford, Oliver Clifford, and all of my family and friends (both in and out of University) that have provided me with such great moral support throughout.

Table of Contents

Chapter 1 - Introduction 5

RNA and the Transcriptome 5

Quantifying the Transcriptome 13

RNA-Seq in Diseases and Disorders 29

Project Aims and Thesis Structure 34

Chapter 2 - Materials and Methods..... 36

Quality Control Procedures 36

Read Alignment and Quantification 41

Pipeline Formation 45

Early Analyses 51

Chapter 3 – Identifying Molecular Signatures of the Placode-Derived Cranial Sensory Ganglia in the Developing Chick..... 55

Abstract **55**

3.1 Introduction **57**

A Brief Overview of Germ Layer Formation 57

Neurogenic Placodes and Cranial Nerves 58

The Sensory Placodes and Associated Disorders 62

Determining Biomarkers for the Placode-Derived Cranial Sensory Ganglia 66

3.2 Methods **71**

Read Alignment and Quantification of Expression Levels 71

Principal Component Analysis (PCA) 72

Differential Expression Analysis 72

High-Confidence Marker Selection 73

Functional Annotation 74

3.3 Results **75**

Read Alignment and Quantification of Expression Levels 75
Principal Component Analysis (PCA) 78
Differential Expression Analysis and Functional Annotation 88
High-Confidence Marker Selection and Functional Annotation 92
3.4 Discussion **101**

Chapter 4 – Expression of Autism-Associated Genes in Developing Purkinje Cells Predicts Intellectual Disability Co-Morbidity Status 110
Abstract **110**
4.1 Introduction **112**
The Purkinje Cells of the Cerebellum 112
Autism Spectrum Disorder and the Cerebellum 116
Investigating Purkinje Cells and ASD 119
4.2 Methods **121**
Read Processing and Alignment 121
Weighted Gene Co-expression Network Analysis (WGCNA) 122
Fuzzy Clustering 123
Differential Expression Analysis 125
GO, KEGG, and MGI Enrichment 125
Disease Enrichment Analyses 126
Comparative Analysis across Brain Regions 128
Permutation Testing 128
4.3 Results **130**
Read Processing and Alignment 130
Weighted Gene Co-expression Network Analysis (WGCNA) 145
Fuzzy Clustering 164
Differential Expression Analysis 176
Comparative Analysis across Brain Regions 191
Permutation Testing 200
4.4 Discussion **207**

Chapter 5 – RNA-Seq Ontology-assisted Clustering and K Selection: the ROCKS Package 212

Abstract **212**

5.1 Introduction **214**

5.2 Methods **222**

Implementation 222

Methods used in Additional External Packages 226

Testing 228

5.3 Vignette **230**

Input Data 230

Socket Cluster for Parallel Processing 233

Model-Based Clustering 235

Akaike Information Criterion (AIC) 237

Enrichment Testing 238

Determining K 244

Session Information 251

5.4 Results **252**

Time Testing 252

Determining K 256

5.7 Discussion **269**

Chapter 6 – Differential Expression and Splicing Analysis Provides Novel Insight into the Transcriptomic Effects of *IFITM3* SNP rs12252 in Influenza 272

Abstract **272**

6.1 Introduction **274**

Influenza as a Global Issue 274

The Types and Subtypes of the Influenza Virus 278

Gaining Entry into the Host Cell 279

Investigating *IFITM3* and the rs12252 SNP 281

6.2 Methods **286**

Read Processing and Alignment 286

Differential Expression Analysis 288

Differential Splicing Analysis 289
Differential Exon Analysis 292
GO and KEGG Enrichment 292
6.3 Results **293**
Read Processing and Alignment 293
Investigating the Effects of Infection 301
Differential Expression Analysis 314
Differential Splicing Analysis 332
Differential Exon Analysis 336
6.4 Discussion **339**

Chapter 7 – Conclusions and Future Perspectives 346
Future Directions 346
Future Directions of Next Generation Sequencing 349
Future Directions of RNA-Seq 351

References 356

Appendix A – Supplemental Meterials for Chapter 4 .. 383

Appendix B – Supplemental Materials for Chapter 5 .. 396
B1 - Main Function Help Pages 396
B2 – Raw Function Codes 430
B3 – Internal Function Help Pages 447

Appendix C – Supplemental Materials for Chapter 6 .. 456

Chapter 1 – Introduction

RNA and the Transcriptome

The central dogma of molecular biology describes the information transfer from nucleic acids through to proteins. Francis Crick originally coined the central dogma in 1958 (although he later acknowledged dogma to be a misnomer), formalizing it subsequently in 1970. This stated that information cannot be transferred back from protein to either protein or nucleic acid (Crick, 1970). This flow between nucleic acids for the passing of information from DNA to RNA, known as transcription, is the process of compiling a strand of messenger RNA that is complementary to the DNA sequence being transcribed. The DNA then acts as a stable storage or reference of genetic material in the cell nuclei, while conveying information through RNA. Nevertheless, in opposition to the central dogma, reverse transcription is observed in RNA viruses, endowing the ability to transcribe RNA into DNA that can be integrated into the host genome. Information is also passed from DNA to DNA during replication, and RNA to RNA in replication of RNA viruses (Termin & Mizutani, 1970).

The one-way flow from nucleic acids to proteins described in the central dogma is known as translation. Translation is observed following the relocation of messenger RNA (mRNA) from the nucleus to a ribosome, which is formed of ribosomal RNA (rRNA) and proteins, where it is used as a template for aligning and assembling a chain of amino acids. Transfer RNAs (tRNAs) then act as the

physical link between amino acids and codons (nucleotide triplets) allowing the formation of an amino acid chain that directly corresponds to the mRNA sequence (Schweet & Heintz, 1966).

The process of translation is the transmission of information originally embodied by Francis Crick's description, and although this hypothesis of a sole direction of information flow is still widely upheld in the scientific community, it is frequently questioned. One major example of this is epigenetic modifications such as DNA methylation. DNA methylation is the biochemical process of adding a methyl group to a DNA nucleotide (usually cytosine), which is associated with gene expression variation through either recruitment of histone modification proteins (resulting in chromatin remodeling) or by physically impeding the binding of transcription factors (Razin & Cedar, 1991). Methylation status is usually acquired through the action of DNA methyltransferases, which are proteins responsible for catalyzing the addition of a methyl group to a DNA nucleotide (Jurkowska *et al.* 2011). This has also been used to argue that, although the sequence of DNA remains unchanged, the effects of these proteins on methylation, and the changes in gene expression, contradict the central dogma (Peedicayil, 2005).

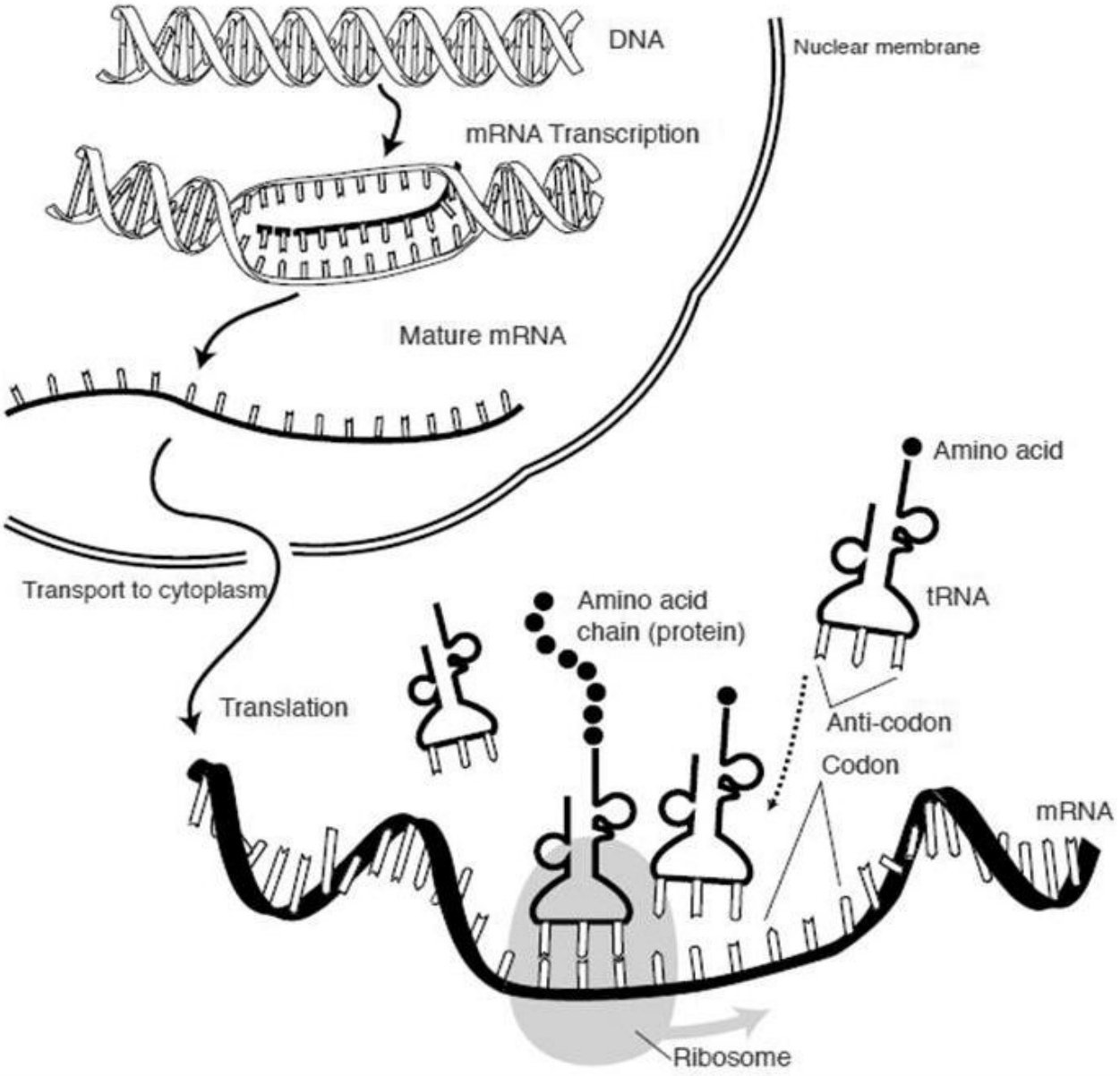


Figure 1.1. A schematic representation of the central dogma of molecular biology, depicting the main flow of information from DNA to RNA to protein. The diagram displays the process of transcription from DNA to mRNA, subsequent mRNA maturation and exportation from the nucleus, and mRNA translation into protein. The ribosome (partially formed of rRNA) is also shown assembling protein via codons/anti-codons attached to amino acids by tRNAs. This Figure has been reproduced from Figure 4 of Vensko & Ferguson, 2010.

Although the central dogma is frequently challenged, it is undeniable that RNA plays a key role in a large number of genetic processes, and our understanding of these many roles continues to increase. The three classes of RNA that play vital roles in protein synthesis are the aforementioned mRNA, tRNA, and rRNA, whose respective primary roles are: transporting copies of genetic information from DNA, matching the relevant amino acid to a codon, and forming a ribosome that catalyzes assembly of an amino acid chain (Lodish *et al.* 2000).

The entire set of RNA molecules transcribed in a cell (or set of cells) at any one time can be defined as its transcriptome. The transcriptome differs from the genome (which is generally unchanging for any given cell lineage, with the exception of somatic mutations) in that it is highly variable and temporally regulated. One example of this temporal variability is in the transcriptomic changes observed throughout the progression of the eukaryotic cell cycle, which is dependent upon hundreds of genes undergoing expression changes throughout its various stages (Archer *et al.* 2011). This has also been widely observed and studied in yeast, with at least nine transcription factors identified as playing central roles in cell-cycle regulation, and a large number of downstream targets including a range of cyclins, DNA replication proteins, and structural proteins (e.g. histones) (Simon *et al.* 2001). Many other cycles are also known to coincide with transcriptomic alteration, including those underlying the ultradian (<24 hour), circadian (~24 hour), and infradian (>24 hour) rhythms in complex organisms. Ultradian rhythms include examples such as heart rate, urination, and appetite; circadian rhythms relate to the day-night biological “clock”; infradian rhythms include examples such as

menstruation or breeding cycles (Wijnen *et al.* 2006). There are also many more examples of transcriptomic variability other than temporal effects, such as environmental factors. Examples of these include the effects of surrounding cells, such as cell-cell communication in embryogenesis (Piras *et al.* 2014); physiological insult, for example under heat-shock (Sakatani et al. 2013); or pathological disruption, for example in innate immune responses (Glennon *et al.* 2015). Overall, the fluid nature of the transcriptome, and its ability to change and adapt, make it a challenging aspect of biology to study.

The diversity of the transcriptome is also much greater than it appears, for example due to alternative splicing, which has been demonstrated to occur in ~95% of human intron-containing genes (Pan *et al.* 2008). Splicing is a widespread mechanism for intron removal and exon ligation through the RNA-protein complex known as the spliceosome, which consists of ~180 proteins with different functions, and five small nuclear RNAs (snRNAs) (Wahl *et al.* 2009). Alternative splicing allows differential inclusion of exons in the mature mRNA, thereby producing multiple mRNA isoforms from just a single gene. Examples of many different types of these can be seen in Figure 1.2, with a specific biological example displayed in Figure 1.3. The results of alternative splicing can range from different protein products, to a loss of the transcript altogether through nonsense-mediated decay (Matlin *et al.* 2005). The effects of this phenomenon on transcriptome diversity can be vast, for example the *Drosophila melanogaster* gene *Dscam* encodes a cell surface protein utilized in the developing brain for axon guidance, which exists as products of up to 38,016 alternatively spliced isoforms (Schmucker & Flanagan, 2004).

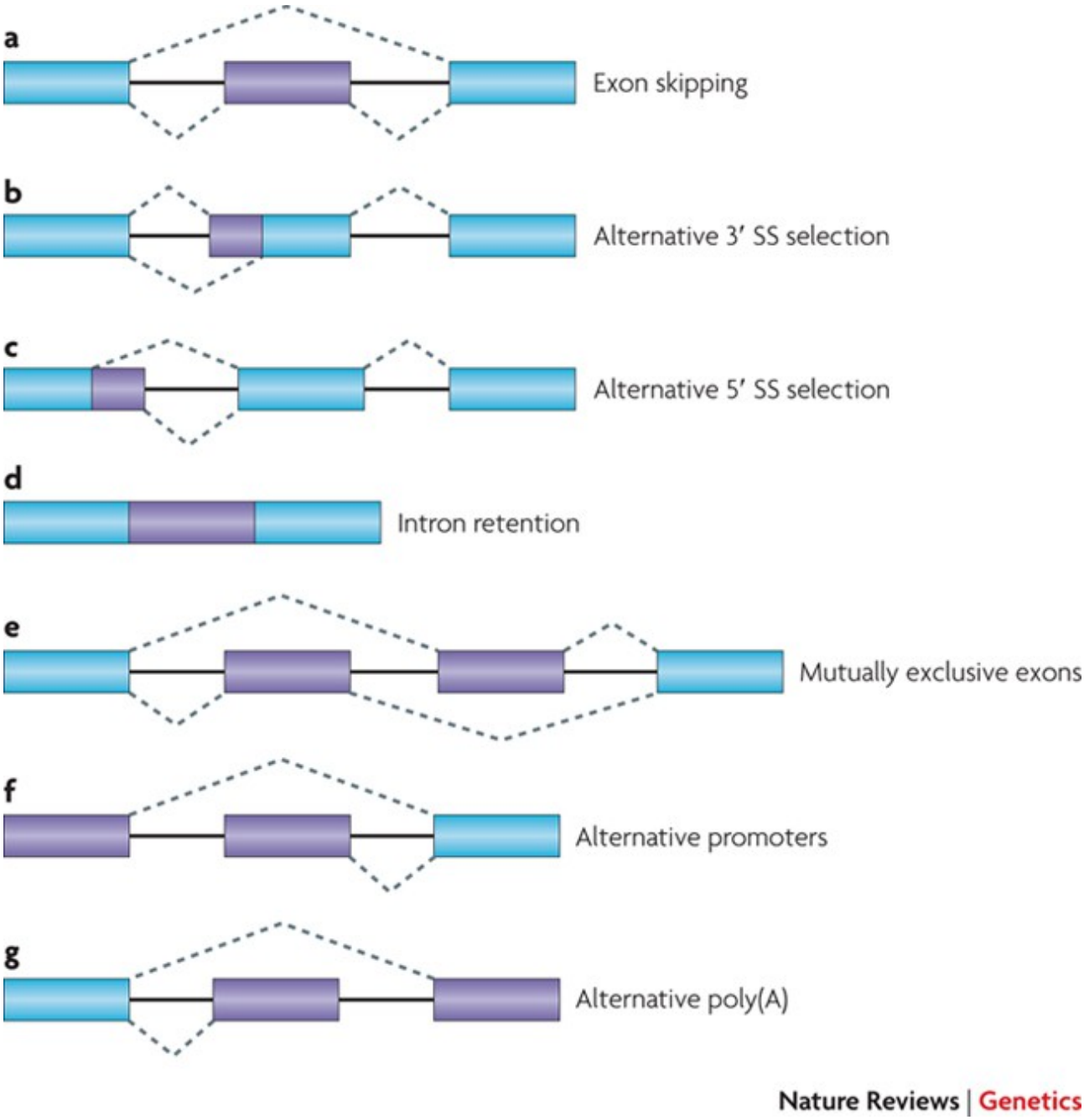


Figure 1.2. A schematic of several different types of alternative splicing (AS) events. Constitutive exons are displayed in blue and alternatively spliced regions in purple, with introns represented by solid lines, and splicing options by dashed lines. (A) displays exon skipping, in which an exon is completely excluded from the resultant transcript. This is the most common form of AS in higher eukaryotes, accounting

for nearly 40% of all AS events (Alekseyenko *et al.* 2007). (B-C) show AS events involving the inclusion of a Splice Site (SS) at the end of an exon. Of all AS events in higher eukaryotes, 18.4% are of this form at 3' SSs (B) and 7.9% are of this form at 5' SSs (Keren *et al.* 2010). (D) displays the rarest type of AS event in vertebrates and invertebrates (less than 5%; Alekseyenko *et al.* 2007; Kim *et al.* 2008), which involves inclusion of an intron in the final transcript. In contrast to vertebrates and invertebrates, this is the most common event in plants, fungi and protozoa (Kim *et al.* 2008). (E-G) display alternative transcript variants (whereby one of two options is selected), including mutually exclusive exons (E), alternative promoter usage (F), and alternative polyadenylation (G). This figure has been reproduced from Box 1 of Keren *et al.* 2010.

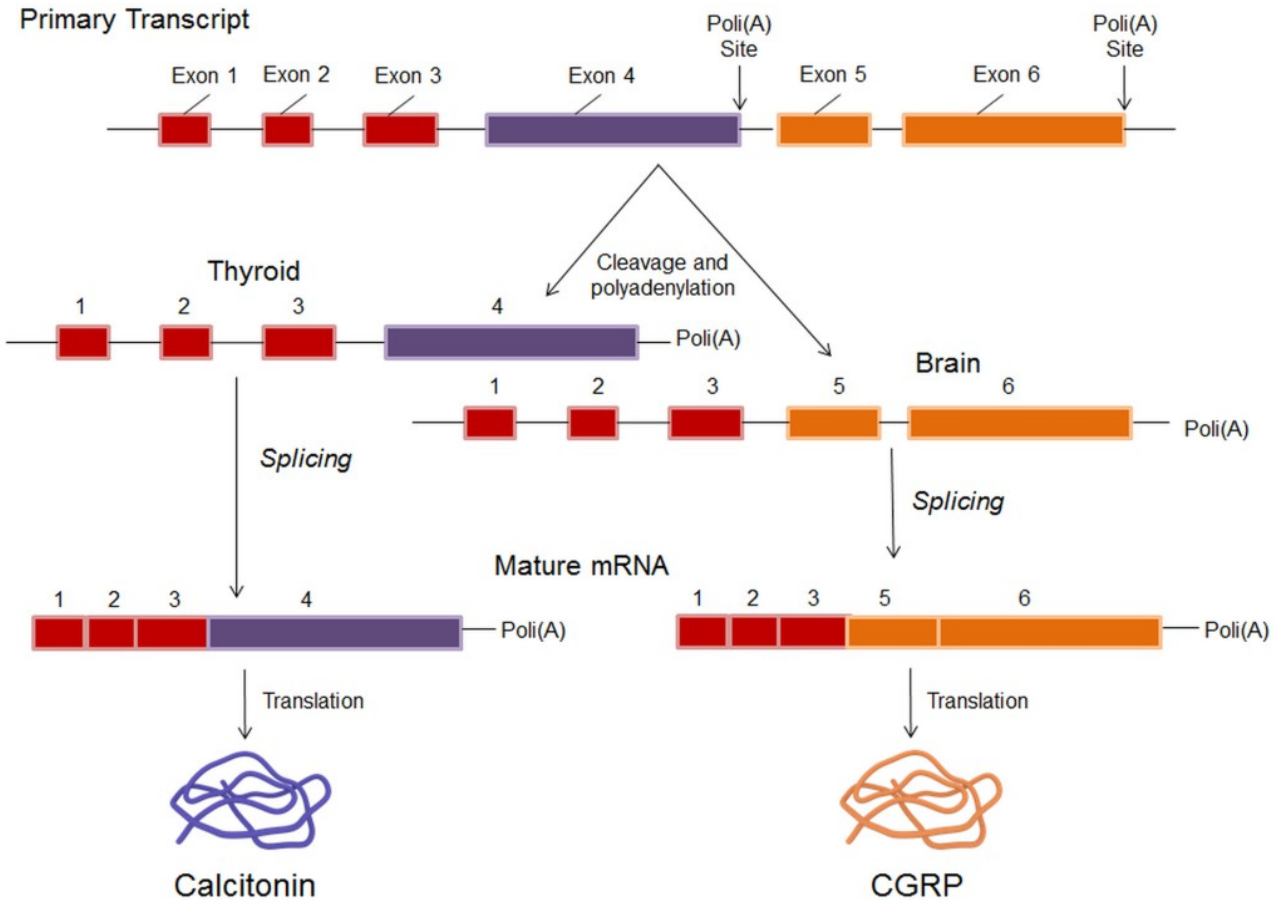


Figure 1.3. A schematic of a specific alternative splicing example observed in the *CALCA* gene (Camacho *et al.* 2014). The *CALCA* gene, as seen here, contains a stop signal/poly(A) site after exon 4, which allows the formation of the calcitonin protein in the thyroid. Through the skipping of this exon, the formation of a different protein (CGRP) occurs in the brain. This figure has been reproduced from Figure S1 of Camacho *et al.* 2014.

Quantifying the Transcriptome

The first technology widely available for simultaneous quantification of transcriptome-wide gene expression levels was the cDNA microarray. This hybridization/probe-based approach evolved from Southern blotting, a process by which fragmented DNA with an attached substrate is probed with a known DNA sequence (Southern, 1975). By applying this principle of complementary sequence hybridization in combination with radioactive or fluorescent detection strategies, and synthesizing these probes at a high density (on a porous membrane "chip"), gene expression values can be determined across thousands of loci dependent on the resultant signal strengths (Schena *et al.* 1995). Figure 1.4 provides details of a typical microarray experiment. Microarrays show a number of advantages in providing a low-cost solution to genome-wide expression quantification, and are still frequently used as such.

However sequencing-based transcriptome quantification methods are fast becoming a disruptive technology, by providing great improvements on many aspects of microarrays. For example, microarrays include a higher level of background signal (due to ambiguous hybridization; Okoniewski & Miller, 2006), a smaller dynamic range (including a ceiling on quantification due to probe number limits, and poor sensitivity at low expression; Draghici *et al.* 2006), and inconsistencies in technical replication (for example, ozone levels are shown to alter the efficacy of some cyanine dyes; Fare *et al.* 2003). As sequencing-based methods determine RNA sequence as opposed to determining their abundance via their hybridization to a known probe, RNA-

Seq analysis is unaffected by many of the issues observed in microarrays. For example, probe ambiguity is not a problem, and the number of reads can always be increased to overcome the issue of limited dynamic range. However, in practice, these issues are not always avoided. In RNA-Seq there remain errors in sequencing and the mis-mapping of reads which can lead to inaccuracies similar to those arising from probe ambiguity. Additionally, any reads that align to multiple places in the genome are often discarded due to an inability to determine the original origin of the read (see Chapter 2 – Read Alignment and Quantification). The expense incurred by deeper sequencing also often provides a somewhat similar upper limit to the dynamic range of RNA-Seq. Similarly, the poor quality of RNA sample can have a negative effect on the complexity of the final RNA-Seq library, resulting in similarly low sensitivity at low abundance.

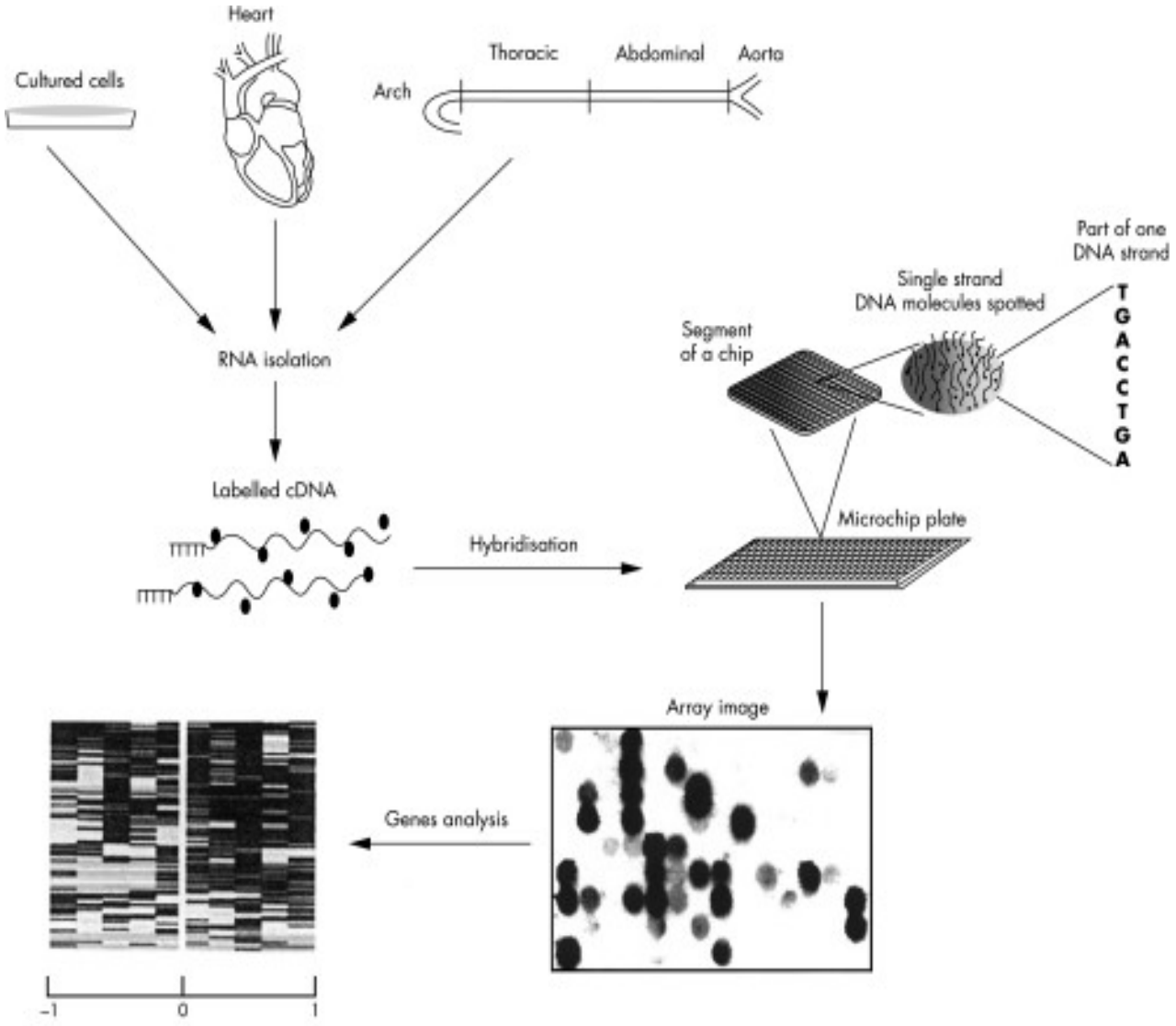


Figure 1.4. A schematic of a typical microarray experiment, in this case in cardiovascular research. This includes RNA isolation, probe preparation and labelling, hybridization to a microchip of single strands of DNA, visualization of signal intensity and/or colour, and transformation into comparable values. This figure has been reproduced from Figure 1 of Napoli *et al.* 2003.

Sequencing-based approaches allow a more quantitatively accurate result through counting the number of transcripts, compared to spot intensities in microarrays, which fall in non-discrete gradients. One of the first examples of the sequencing-based approach was Serial Analysis of Gene Expression (SAGE), in which a restriction enzyme is used to cut single short segments from transcripts, which are then ligated to each other before being sequenced (Velculescu *et al.* 1995). SAGE also has many variants, such as Cap Analysis Gene Expression (CAGE), which is used to specifically sequence the start sites of transcripts and allows investigation of promoters (Shiraki *et al.* 2003). Although these, and many more methods, have been proposed as sequencing-based approaches, overall these have been seen as expensive and labour intensive. However, the introduction of Next Generation Sequencing (NGS) is revolutionizing this, and (although the number of array data sets is currently still outgrowing RNA-Seq sets) sequencing-based approaches are rapidly becoming more appealing compared to hybridization-based approaches.

NGS is a catch-all term used to describe a number of different modern sequencing technologies, which perform sequencing processes in parallel on a library of DNA (or cDNA) fragments so as to produce millions of read sequences simultaneously. These NGS technologies provide a higher-throughput alternative to traditional sequencing methods, and have developed into a fast and relatively cheap method for sequencing genomes. For example, the human genome project was undertaken with Sanger-based sequencing methods, taking 13 years and with an overall cost of approximately $3 billion (International Human Genome Sequencing Consortium, 2004; Wadman,

2008). By comparison, in 2014 Illumina announced the release of the NextSeq 500 and the HiSeq X10 NGS sequencing platforms capable of sequencing a human genome in under 30 hours and at a cost of less than $1,000 respectively (Hayden, 2014). Presently, there are multiple next generation sequencing platforms both available currently and in development, each with their own distinct advantages and disadvantages (see Chapter 7). The most commonly used range of platforms at present are those developed by Illumina, which perform sequencing by synthesis. This begins with a library of denatured single strands grafted to a flowcell, followed by the addition of the four nucleotides containing different cleavable fluorescent dyes and a removable blocking group, allowing the visualization of the incorporation of each complementary base one at a time.

The development of these NGS sequencing platforms has enabled a vast assortment of methods for analysing RNA. Illumina themselves promote over 75 library preparations for which NGS technology can be applied to interrogate different RNA processes, including methods to analyse RNA transcription, RNA-protein interactions, low-level RNAs, and RNA modifications/structure (Illumina Inc. 2015). Just a few interesting examples of these methods include TRAP-Seq, Ribo-Seq, and SHAPE-Seq. Translating Ribosome Affinity Purification (TRAP) sequencing is a method devised for specifically isolating (and subsequently sequencing) what is known as the translatome, the subset of the transcriptome which proceeds to be translated (Jiao & Meyerowitz, 2010). As the quantification of mRNA levels as a whole is an imperfect proxy for analysing protein synthesis, this method enables direct monitoring of

translation. The TRAP process is implemented using an animal model (for example, the mouse line RiboTag) which expresses an epitope-tagged version of the 80s ribosomal protein RPL22 (Sanz *et al.* 2009). The addition of a hemagglutinin (HA) epitope, which is consequently integrated into actively translating polyribosomes, allows for the targeted affinity purification of these using antibodies conjugated to agarose beads. This isolates (along with the polyribosomes) any associated polysomal mRNAs that are being translated. Reverse transcription of these mRNAs can generate a cDNA library for use with NGS systems to enable an accurate quantification of the translatome. Although TRAP has the distinct disadvantage of requiring animal models rather than humans, TRAP-Seq has proved useful in identifying subtle differences in cell types, for example in defining neuronal subpopulations from hypothalamic/brainstem LepRb cells (Allison *et al.* 2015). The aims of Ribo-Seq lie much in the same vein, with the goal of uncovering information on the translatome, but doing so by determining the positions of translating ribosomes (Ingolia *et al.* 2009). The key principle of Ribo-Seq involves the degradation of mRNA using ribonucleases, which leave only ribosomes and fragments of ribosome-protected mRNAs. Following isolation of mRNA-ribosome complexes with centrifugation, protein removal and rRNA depletion, and finally reverse transcription of mRNAs, the cDNA library is once again produced. One major disadvantage of Ribo-Seq is that typical reads are approximately just 30 nucleotides in length (Michel *et al.* 2013), which can make some downstream analyses difficult (for example analyses requiring alignment across splice junctions). However, the advantages of increased resolution (compared to TRAP-Seq), that it can be utilized in any organism,

and that it ensures the ribosomes are actively translating the mRNA (unlike, for example, in cases where ribosomes are bound to the 5' untranslated region) make Ribo-Seq a useful method in quantifying and analysing active translation sites. Finally, Selective 2'-Hydroxyl Acylation analysed by Primer Extension (SHAPE) sequencing provides a method for analysing RNA structure (Mortimer *et al.* 2012). This covers a contrasting area of the transcriptome, by revealing the secondary and tertiary structural changes in RNA with the aim of identifying functional regions, for example in noncoding RNAs with regulatory roles. The idea of SHAPE-Seq is centred on a reagent known as 1M7, which undergoes hydrolysis and forms adducts preferentially on unpaired RNA nucleotides (for example, in the head of a hairpin loop). This reagent also blocks reverse transcription, meaning when the 1M7-treated RNA is reverse transcribed only cDNA from sequence upstream of a 1M7 adduct is produced. The sequencing of cDNA from this process therefore reports information on structure at an individual nucleotide resolution. Additionally, with RNA barcoding, it is possible to employ this in extracting information from multiple RNAs. However, the major caveat of SHAPE-Seq is that there is no guarantee that folding of the RNAs *in vitro* will reflect that which occurs *in vivo*. These are just a few examples of how NGS platforms can be utilized in investigating RNA biology (a more comprehensive list is available at https://liorpachter.wordpress.com/seq/ [accessed 5/6/16]).

RNA-sequencing (RNA-Seq) is a technology that harnesses the capabilities of NGS for transcriptome sequencing, allowing a genome-wide sequencing-based method of transcriptomic profiling superior to previous array-based methods

(Wang *et al.* 2009). Figure 1.5 displays a schematic of a typical RNA-Seq experiment, including the processing of RNAs for application of NGS, and some upstream computational procedures.

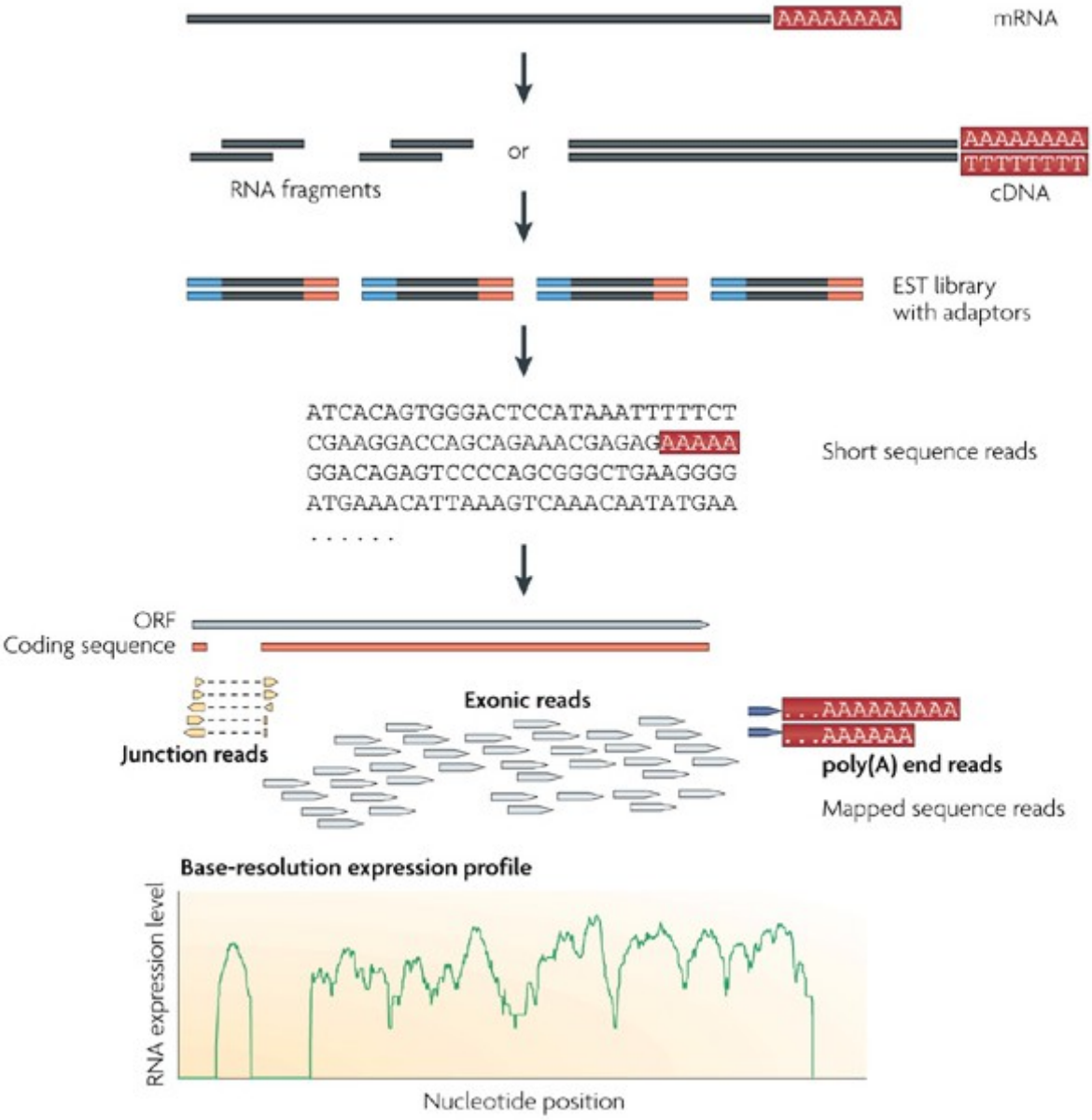


Figure 1.5. A schematic of a typical RNA-Seq experiment. Firstly, RNAs are fragmented and converted into a library of cDNA fragments. Sequencing adaptors (seen here in blue) are then added to the fragments. This allows next generation sequencing to be applied to each cDNA, resulting in a short sequence from each fragment. Certain library preparations also allow for "stranded" RNA-Seq, in which different adapters are ligated to the 5' and 3' ends of the reads. This

enables information on whether the strand is sense or antisense to be retained throughout the sequencing process. In single-end sequencing, one end of the fragment is sequenced, whereas in paired-end both are sequenced (allowing greater understanding of the coverage of the fragment and more junction reads, which in turn can provide greater power for alternative splicing analyses). The final part of this schematic shows results from some early computational analysis, including the alignment of these reads to a reference genome or transcriptome, and the subsequent expression profile based on the number of reads aligned at each nucleotide position (for an in depth representation of computational workflow, see Figure 1.6). This figure has been reproduced from Figure 1 of Wang *et al.* 2009.

One issue that has been identified in standard RNA-Sequencing library preparation is the inability to correctly infer from which strand the sequenced RNA has been transcribed. Although it would be possible to infer the strand based on which the original RNA sequence is complement to, the preparation of RNA fragments for sequencing results in a double stranded cDNA library (as can be observed in Figure 1.5), and as either of these cDNA strands can be subsequently sequenced, it is not possible to know whether the final sequence is the original or complement of the RNA. This poses a problem in cases where genomic loci contain overlapping transcription units on both strands. These are known as natural antisense transcripts (NATs), and around 15% of human protein-coding genes are thought to have an associated NAT (Lapidot & Pilpel, 2006). In these cases, without information on which strand RNA has been transcribed from it is impossible to infer from which transcriptional site the RNA arose (with the exception of reads that encompass GT-AG splice sites, for which strandedness can be inferred). Similarly, non-overlapping loci of identical sequence (and therefore identical complementarity in antisense transcription) (trans-NATs) also present this issue. Stranded RNA-Seq has been developed as a solution to this, with the methods underlying these protocols being split into two classes: strand-specific adapter ligation or chemical strand modification. Illumina have developed protocols in both of these classes. Illumina developed a strand-specific adapter ligation method which involves the process of linking different adapters to the 5’ and 3’ of single-stranded DNA through a T4 DNA ligase (Kuhn & Frank-Kamenetskii, 2005). This enables the determination of directionality directly in the single stranded cDNA, which both simplifies the process of RNA-seq library preparation in general, and circumvents the loss of

strand information when forming double stranded cDNA. However, the chemical strand modification method developed by Illumina is actually more commonly used (in the TruSeq Stranded Total RNA Sample Prep Kit) and is identified as the leading protocol (Mills *et al.* 2013). This involves the incorporation of deoxyuridine triphosphates (dUTPs) into the reverse transcription reaction that produces the second strand of the cDNA, resulting in a strand containing deoxyuridine residues instead of deoxythymidine (as in the first strand). Following ligation of "Y-shaped adapters" (non-complementary double-stranded sequences), Illumina then uses a modified polymerase in amplification which is not functional with deoxyuridine, and therefore forms a library from only the first strand. The Y-shaped adapters mean that this first strand is resultantly spanned by different adapter sequences, once again retaining information on directionality. The use of these techniques for NAT analysis can vastly increase the effectiveness of studies, considering these transcripts are relatively commonplace in the mammalian transcriptome, are likely to play important regulatory roles (Katayama *et al.* 2005), and have been previously associated with human disorders (Faghihi & Wahlestedt, 2009).

There are some drawbacks to RNA-Seq, the most renowned of which include GC-biases and the issues associated with short read sequencing. GC-bias is a phenomenon observed in most NGS systems as a natural result of sequencing by synthesis, the result of which (in RNA-Seq) is an underrepresentation of both GC-rich and GC-poor regions (Risso *et al.* 2011). One major disadvantage of short read sequencing includes the inability to cover the entire transcript isoform (Figure 1.5). This means that analyses on elements such as splicing

events relies on junction-spanning reads and probabilistic models, which would not be as heavily relied upon if the entire RNA or fragment could be sequenced. Although these drawbacks are not insignificant, the many advantages of RNA-Seq have made it a widely used method. Additionally, there are several new RNA-Seq technologies either under development or newly released, which show promise in resolving many of these issues (see Chapter 7).

An overview of a full standard computational work-flow of an RNA-Seq experiment can be seen in Figure 1.6A. For more details on software and how these work, see Chapter 2. The raw data received from sequencing is commonly reported in the form of a FASTQ file, a text-based format containing strings of biological sequence and subsequent quality values corresponding to each sequence letter (Cock *et al.* 2010). Each quality value conveys the probability that the corresponding base call is correct. This is based on the historical Phred score, a Sanger sequencing metric algorithmically determined from electropherogram peak resolution and shape, the Illumina equivalent of which is largely the same, based on parameters relevant to sequencing chemistry (Illumina Inc. 2011). The quality values can be used for applying quality control procedures and ensuring that the data are sufficiently robust. Alignment of these reads to the complementary sequence of a reference genome (or de novo assembly without a reference) then allows for the counting of numbers of reads overlapping genomic features. The genome-wide digital quantification of expression is what provides RNA-Seq with its many advantages over the spot intensity quantification used by hybridization-based

methods. A few examples of the resultant analyses available which highlight the advantages of RNA-Seq can be seen in Figure 1.6B. RNA-Sequencing allows full genome-wide quantification analysis, not limited to those features chosen to be included in a microarray, and not limited to quantification ceiling by probe number limits. RNA-Seq also allows for quantification of novel transcripts, whose sequences are unlikely to be included in microarrays, such as non-coding RNAs, and for which (although lncRNA and miRNA microarrays exist) detecting low expression levels can exhibit poor sensitivity in microarrays. Finally, RNA-Seq allows for the more accurate estimation of splicing variants, with reads spanning multiple exons. There is a microarray equivalent for investigating splicing, which is the exon junction microarray. This uses probes specific to potential splice sites. However, neither Illumina RNA-Seq nor exon junction microarrays provide full length transcript information (although RNA-Seq methods for this are emerging; see Chapter 7), and so all isoform quantifications are based on probabilistic models. The advantages of Illumina RNA-Seq are its abilities to interrogate exons <25bp in length, to identify novel splicing events, and to more effectively investigate complex splicing events with numerous variants (Kolle, 2007).

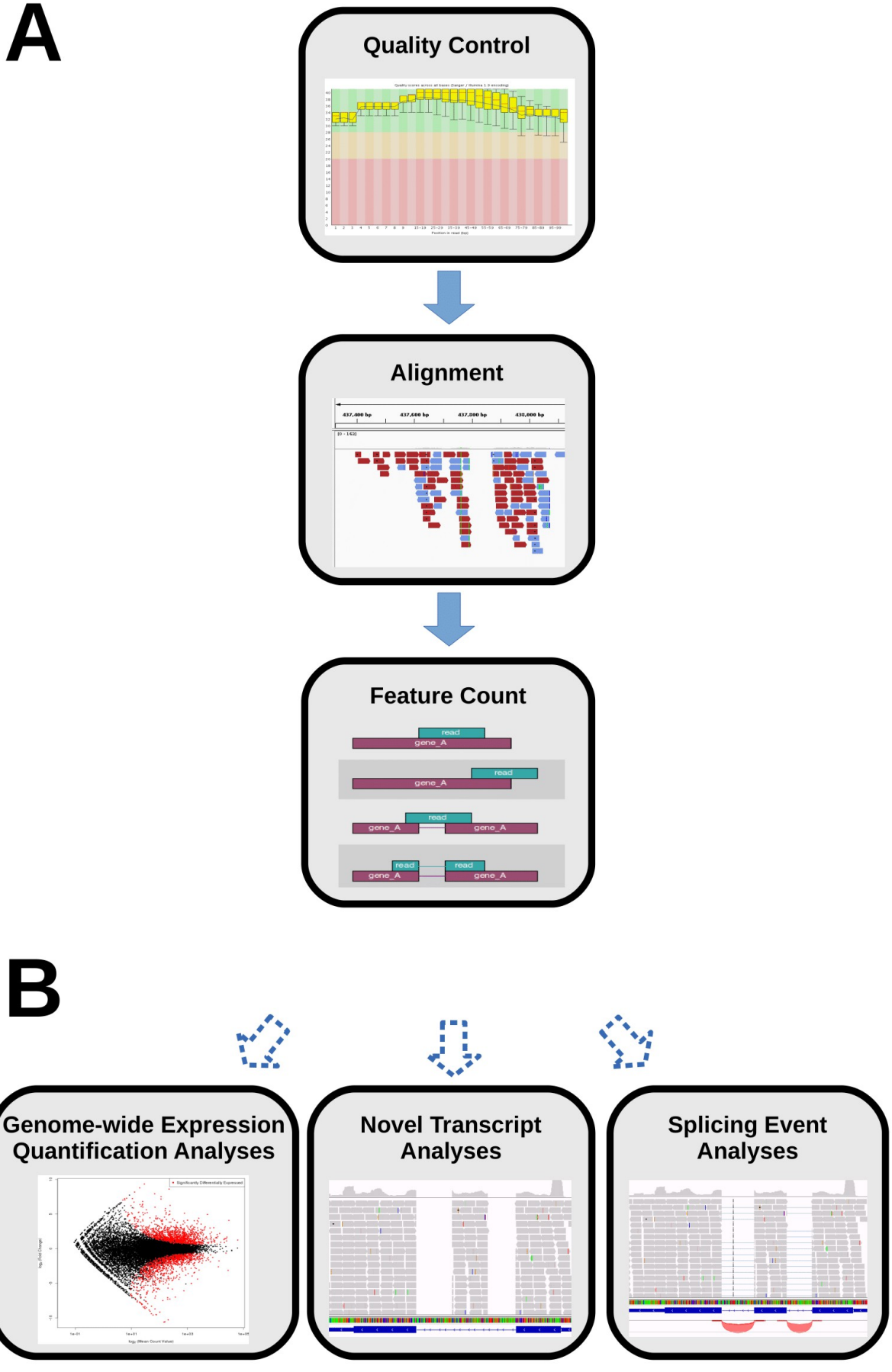

A
Quality Control
Alignment
Feature Count
read
gene_A
B
Genome-wide Expression Quantification Analyses
Novel Transcript Analyses
Splicing Event Analyses

Figure 1.6. The computational work-flow and potential analyses of a typical RNA-Seq experiment. Panel (A) displays the early computational processes required to produce quantifications for genomic features across the transcriptome. Panel (B) represents examples of potential further analyses that are facilitated by the advantageous aspects of RNA-Seq over alternative methods. This can include basic quantification comparisons (e.g. differential expression analysis), identification of novel transcripts (e.g. lowly expressed transcripts or non-coding RNAs), or characterizing of splicing.

RNA-Seq in Diseases and Disorders

It has been estimated that just 4.9% of disease-associated Single-Nucleotide Polymorphisms (SNPs) lie within protein-coding regions of the genome (Figure 1.7; Maurano *et al.* 2012). This suggests that much of the etiology of these complex diseases occurs through changes in expression as opposed to changes in the protein itself. Although studies frequently provide a deep genetic knowledge for numerous diseases, many of these do not supply real insight into the functional contribution and role of the identified variants in the onset and development of diseases (Costa *et al.* 2013). Due to this, RNA-Seq is acquiring an increasingly important role in unraveling the genetic mechanisms and pathogenesis underlying many complex diseases and disorders.

The role of RNA-Seq in determining expression Quantitative Trait Loci (eQTLs) is also important. Previously, eQTLs have been defined as a term describing a statistically significant genotype-gene expression level correlation (Majewski & Pastinen, 2011). The ability of RNA-Seq to readily provide data on allele-specific expression (not available from gene expression microarrays), and on RNA isoform expression, provides a distinct advantage in investigating these eQTLs (Pickrell *et al.* 2010).

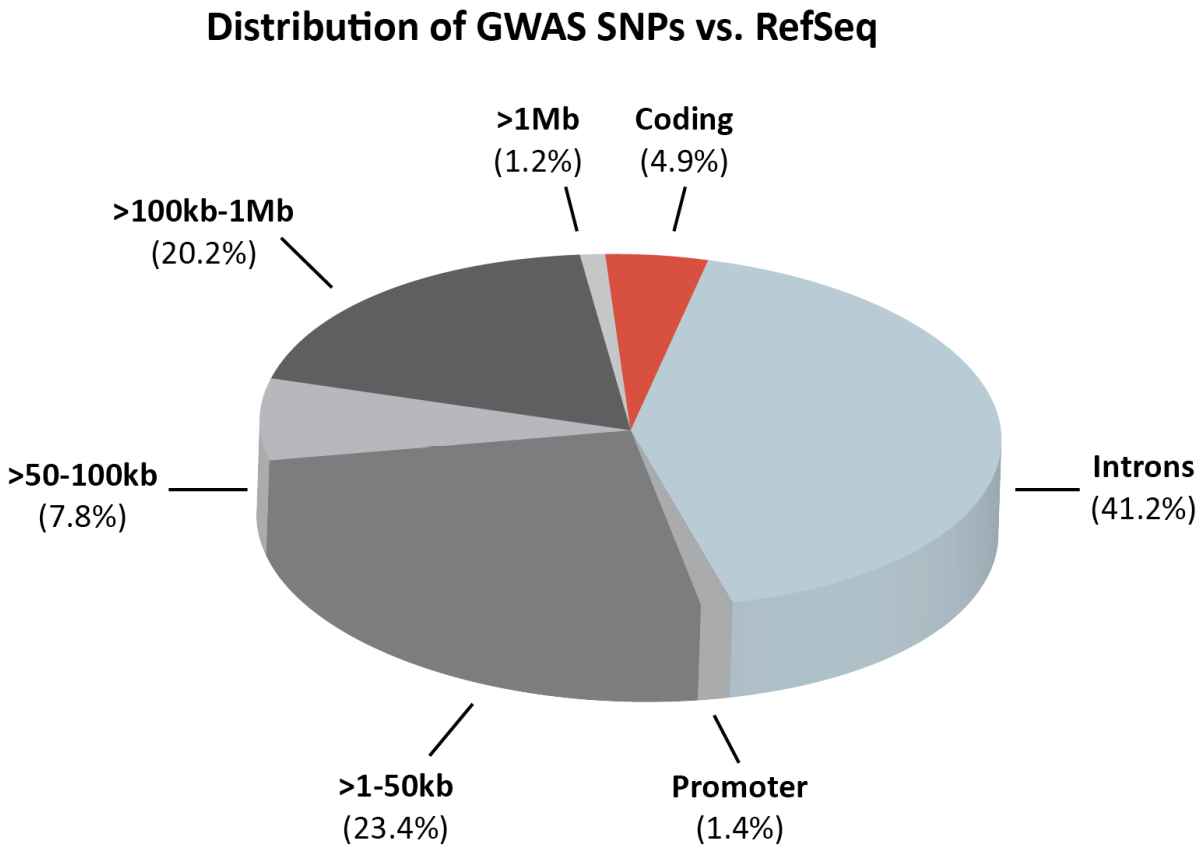


Figure 1.7. A pie chart displaying the distribution of GWAS SNPs by location relative to genic features. This figure has been reproduced from Figure S1B of Maurano *et al.* 2012. The GWAS SNPs in this study were derived from the NHGRI GWAS catalog (Welter *et al.* 2014; now the NHGRI-EBI GWAS catalog). This catalog is a quality controlled, manually curated, literature-derived collection of all published genome-wide association studies assaying at least 100,000 SNPs and all SNP-trait associations with p-values < 1.0 x 10^{-5}. This is formed from both array and sequencing methods, ensuring the list does not just reflect SNP array designs.

There are many examples of RNA-Seq having been used successfully to interrogate a disease, of which cancer is a good instance. One example is in the investigation of variants underlying Granulosa-cell tumours (GCTs), which account for less than 5% of all ovarian cancers and display greater genomic stability in cytogenetic studies than the more common epithelial ovarian cancers. Due to this stability, it was hypothesized that a common variant could be causative, and RNA-sequencing was used to uncover a GCT-specific FOXL2 variant (Shah *et al.* 2009). A second example is in the investigation of alternative splicing in breast cancer transcriptomes, for which sub-type specific isoforms were discovered that have not been recognized in normal human samples. The study uncovered multiple examples of exon skipping and intron retention as predominant splice events in breast cancer, and was the first comprehensive portrait of splicing signatures in breast cancer sub-types (Eswaran *et al.* 2013).

RNA-Seq has also been successful outside of cancer studies. For example, one study utilized RNA-Seq to identify the splicing differences in Alzheimer's-associated brain regions, and suggested alternative splicing and promoter usage of the *APOE* gene could be involved in the progression of neurodegeneration (Twine *et al.* 2011). Another study used RNA-Seq to identify significant transcriptomic differences in schizophrenic patients. This uncovered more than 1000 genes with differential splicing and 2000 genes with alternative promoter usage, in addition to functional map construction to determine three functional clusters highly relevant to schizophrenia (Wu *et al.* 2012).

Overall, RNA-Seq is a powerful tool for interrogating the transcriptomic signatures of many diseases and disorders, and this can successively provide a much greater understanding of the underlying mechanisms. As mentioned previously, there are also many advantages provided by RNA-Seq compared to other technologies. These advantages can be useful in providing novel insights into diseases and disorders. For example, the complete genome-wide quantification of expression provided by RNA-Seq can be used for novel cell-type marker discovery, which in turn could be used for disease models; this, in addition to novel transcript identification, can also be used for the building of more accurate gene regulatory networks, and capturing the relationships between genes for understanding disease etiology; and the ability of RNA-Seq to accurately identify novel splicing events can also provide insight into the changing transcriptome, for example in immunology and disease response (Figure 1.8). In this thesis, I cover all three of these major examples of novel uses for RNA-Seq in relation to diseases and disorders.

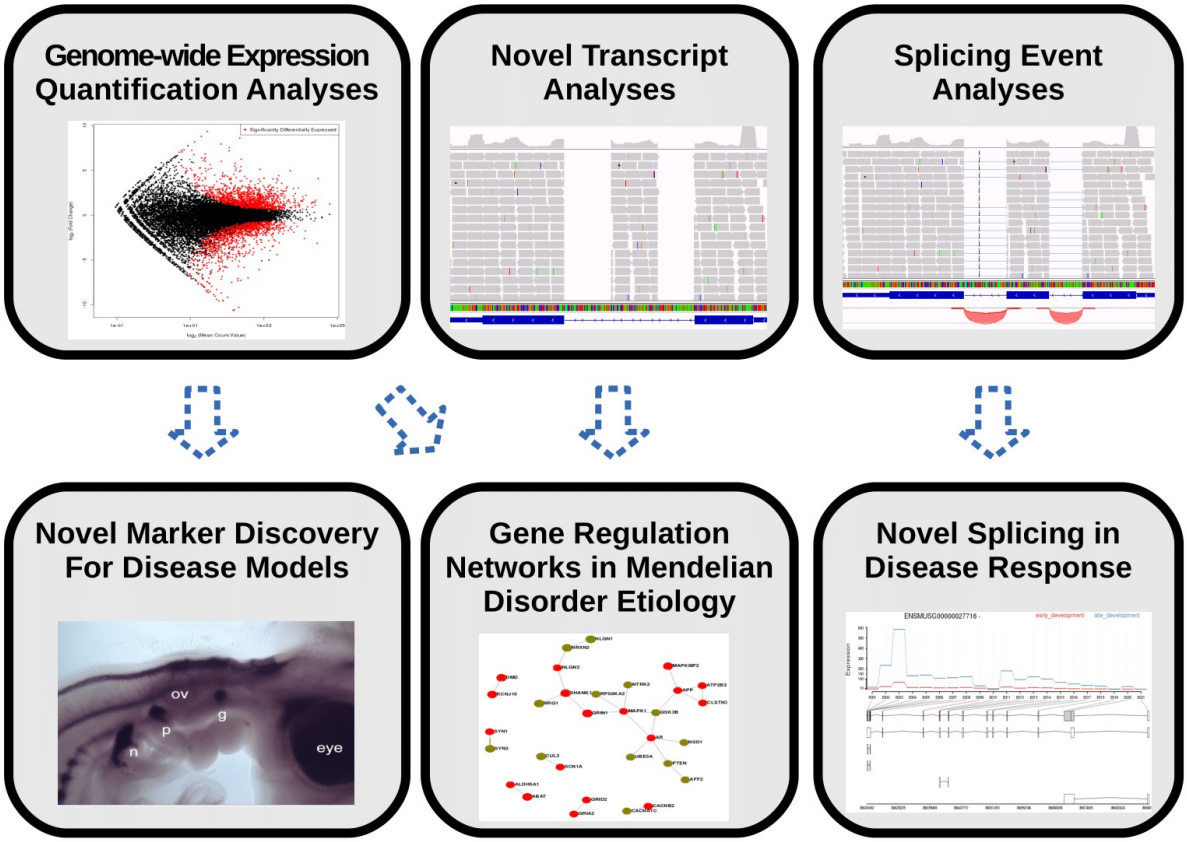


Figure 1.8: A continuation of Figure 1.6B, displaying how the aforementioned RNA-Seq analyses could be used to investigate and gain knowledge on various diseases and disorders. These three approaches are those employed in this thesis.

Project Aims and Thesis Structure

In this thesis, I present three RNA-Seq projects for various disease-relevant tissues. For these, I perform a series of bioinformatic analyses that utilize the advantages and novel qualities of RNA-Seq, in addition to producing an R package that further improves these analyses, to explore and understand these tissues and cells, and their underlying associations with diseases and disorders.

In Chapter 2, I review the materials and methods used throughout this thesis across all RNA-Seq projects, including the pipelines developed for quality control procedures and sequence alignment/quantification.

In Chapter 3, I investigate the transcriptomic signatures of developing chicken placodes, disruptions of which have known associations with disorders such as deafness and migraine, to identify practical genetic markers that could be used in future for distinguishing sensory neuronal cell-type models.

In Chapter 4, I investigate the role of developing mouse cerebellar Purkinje cells in autism spectrum disorder, exploring the compositions of gene expression clusters and autism-associations, and drawing comparisons to the neocortex to examine co-morbidity with intellectual disability.

In Chapter 5, I present an R package for clustering and functional annotation, which provides faster analysis by combining the most robust methods for RNA-

Seq into a single package that allows parallel processing, and which introduces cluster number selection appropriate for RNA-Seq using this functional annotation.

In Chapter 6, I investigate a genotypic variant in the *IFITM3* gene as part of a pilot study into the known associations with severe influenza response in humans, by analysing the differing immune response and varying genetic pathways of human dendritic cells in response to H1N1 influenza viral infection.

In Chapter 7, I conclude with a general discussion of the findings of this thesis, and discuss the future directions of transcriptomics, including developments in methods of sequencing, and single-cell RNA-Sequencing, along with the potential of these in investigating diseases and disorders.

Chapter 2 – Materials and Methods

Quality Control Procedures

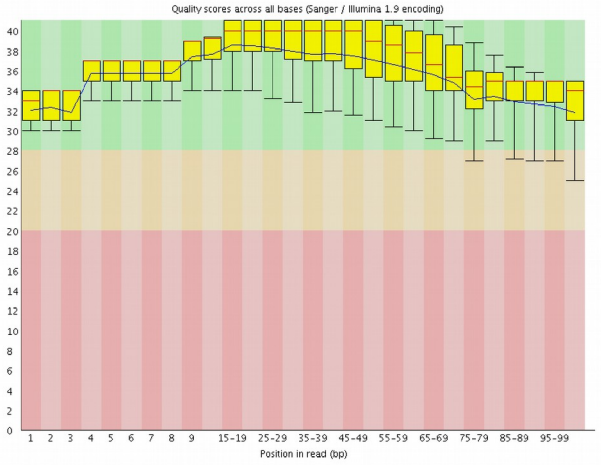
The Quality Control (QC) procedures executed throughout this thesis serve the purpose of assessing the quality of the data, and also correcting the raw sequence reads for any problems or biases observed. Although most sequencers generate QC reports themselves, these are usually focused on problems from the sequencing itself. In this thesis, a further initial assessment of the data was carried out using FastQC (Andrews, 2010), which is a program that provides a modular set of analyses to give an impression of the overall quality of the raw sequence reads through multiple visualizations. The modules returned by FastQC provide information relevant to the sequencing itself, but also relating to problems that may occur in generating the starting library material (Figure 2.1).

In incidences of poor base quality (phred < 20), the FastX-toolkit (Hannon Lab, 2009) was used, including the FASTQ Quality Trimmer for cutting reads down if their quality scores drop below a certain threshold towards the end of the read, and the FASTQ Quality Filter for filtering reads of low quality throughout. In incidences of adapter sequence contamination, Cutadapt (Martin, 2011) was used for cutting reads down based on the presentation of adapter sequences.

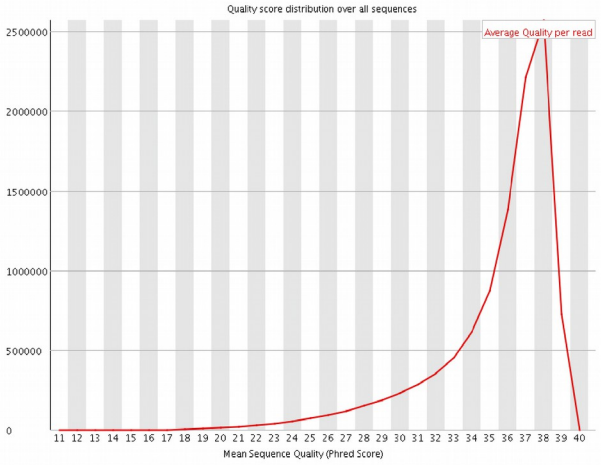
It is important to note here that none of the programs used in this pipeline are

specific for RNA-Seq. This allows for the pipeline to be used generically with different NGS outputs. There are more specific QC tools available that could be useful in further interrogating RNA-Seq outputs, such as RNA-SeQC (Deluca *et al.* 2012) or RSeQC (Wang *et al.* 2012). These tools include additional metrics such as aligned reads distribution, read coverage, strand specificity, and RNA integrity.

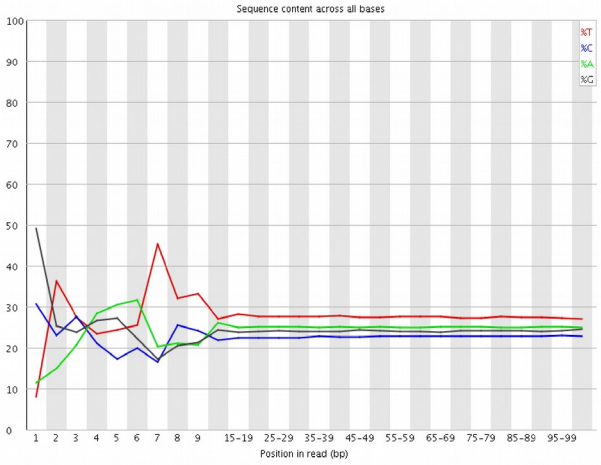
A


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

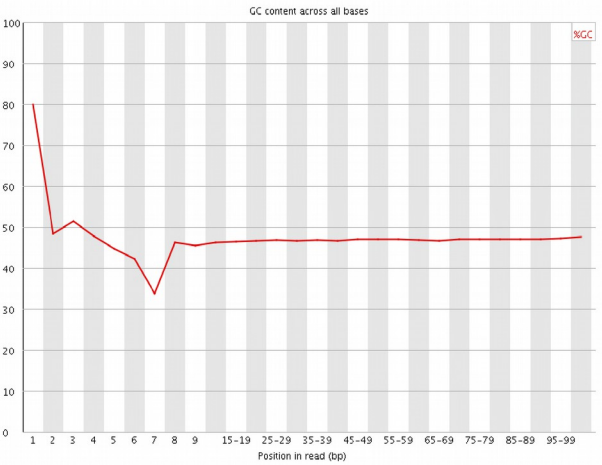
B


Quality score distribution over all sequences
Average Quality per read
Mean Sequence Quality (Phred Score)

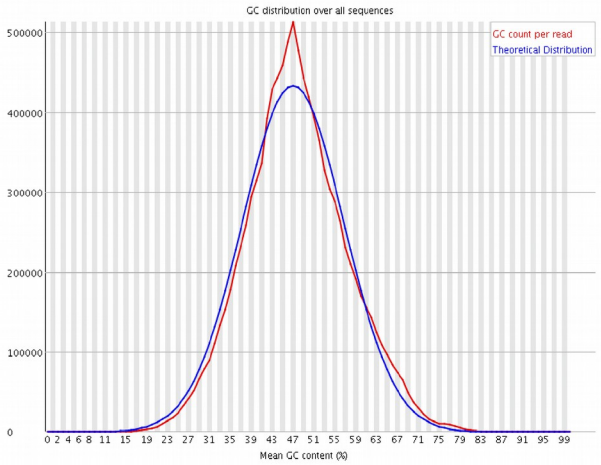
C


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)

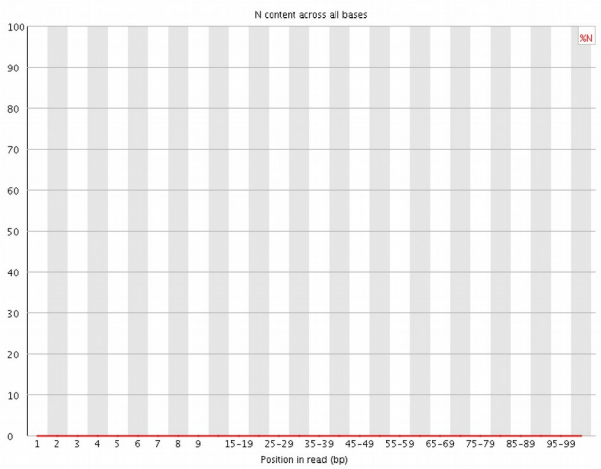
D


GC content across all bases
%GC
Position in read (bp)

E


GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)

F


N content across all bases
%N
Position in read (bp)

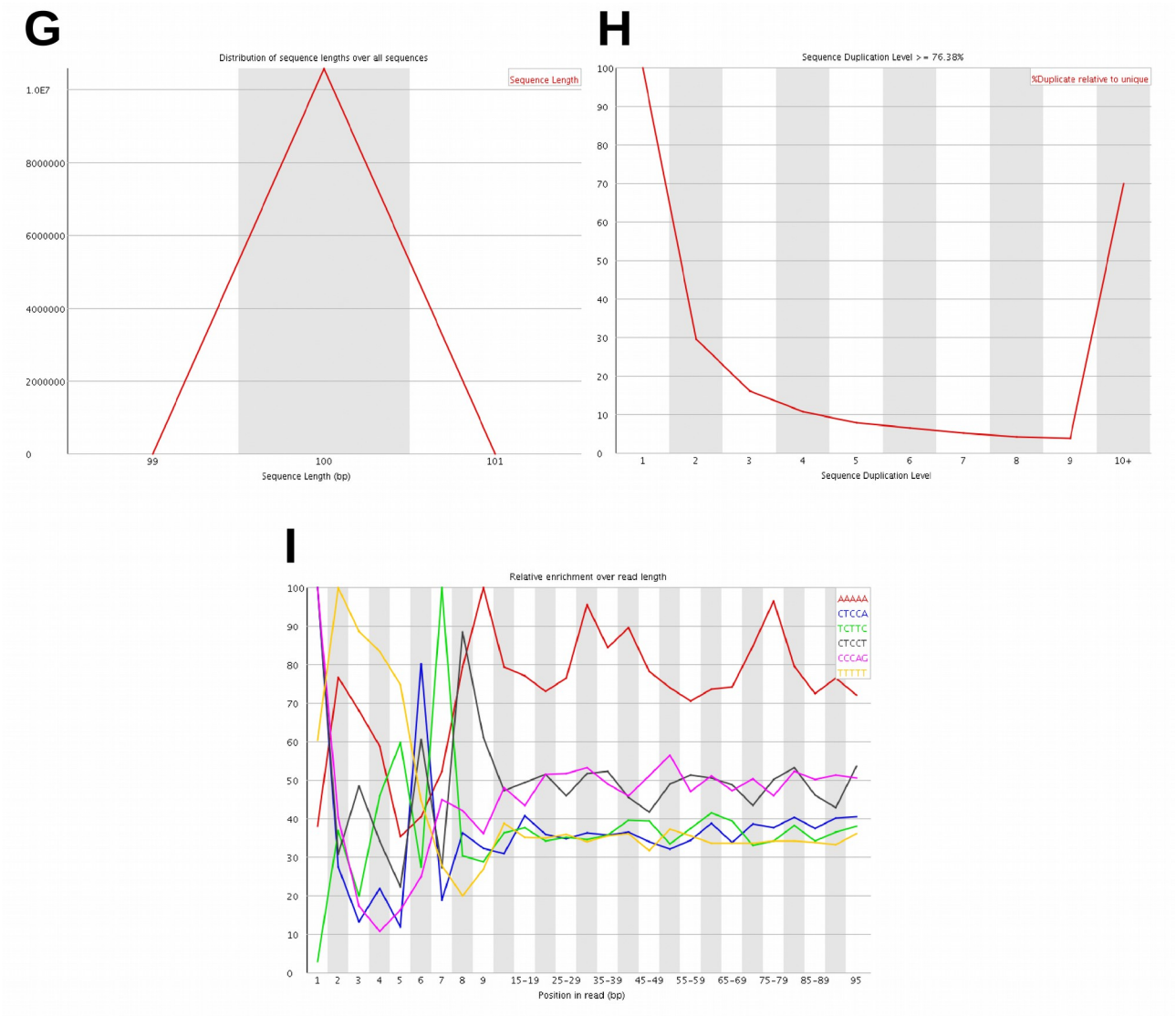


Figure 2.1. Examples of each of the visuals provided by FastQC (Andrews, 2010). The output includes plots to assess quality scores, including a box and whisker plot of quality score distributions across bp position within the reads (A), as well as a plot displaying the number of reads across mean quality scores (B). There are also plots for assessing sequence content, including percentage occurrence of each base across the read (C), GC content both across the reads (D) and throughout the library in general (E), and occurrence of Ns (no base called) across the reads (F). Finally, there is assessment of the

distribution of read lengths (G), read duplication (H), and kmer overrepresentation across the reads (I). The FastQC output also includes a summary of basic file statistics, such as the total number of sequences and %GC overall, and also includes details on any over-represented sequences, such as adapters, which it identifies. The particular examples in this figure are from a sample described in Chapter 6 (CC genotype, H7 infected, replicate 1, lane 175580). In this example, three of these fail the FastQC measures, including per base sequence content (C), per base GC content (D), and sequence duplication levels (H). FastQC is a program for use with all types of NGS outputs, and these particular failings are expected with RNA-Seq data sets. For example, the issues in per base sequence and GC content occur within the first 10 bps, which is due to "random" hexamer priming (ensuring sequencing of many different fragments) during library preparation for RNA-Seq, in which the hexamers are not truly random. The issue in sequence duplication levels can also be expected due to highly expressed genes increasing the number of reads with the same sequences.

Read Alignment and Quantification

Following quality control procedures, the reads are aligned to the relevant reference genome. As of 2012, there are at least 60 aligners available for use in high-throughput sequencing alignment (Fonseca *et al.* 2012), and so selection of a suitable aligner can be difficult. For RNA-Seq it is important to use an aligner which is splice-aware and allows for long distance alignment of reads across splice junctions. Splice-aware aligners accommodate junction-spanning reads through splitting the reads into smaller segments. This is executed through one of two methods depending on the aligner, either "exon-first" alignment or "seed-and-extend" alignment. Exon-first alignment begins by aligning the entire read, and then searching for spliced alignment for the remaining reads, whereas seed-and-extend alignment uses short subsequences, or seeds, determined by breaking down the sequences of the reads and aligning these to the genome. The seed-and-extend aligners are slower but offer the advantage of higher sensitivity (Benjamin *et al.* 2014).

Although many popular aligners use exon-first alignment (e.g. TopHat2; Kim *et al.* 2013), a seed-and-extend aligner is used throughout this thesis due to the increased sensitivity associated with these. Namely, the Genomic Mapping and Alignment Program (GMAP) (Wu & Watanabe, 2005) and its Genomic Short-read Nucleotide Alignment Program (GSNAP) are used. GSNAP is a seed-and-extent short-read aligner that can be used for both single- and paired-end reads as short as 14nt. Most importantly, it is capable of SNP-tolerant alignment (meaning resultant read count values are not biased by SNPs), and

through probabilistic models, long-distance splicing alignment (including interchromosomal splicing).

Following read alignment, the data can be quantified through counting the number of reads that overlap features of the genome. For this, HTSeq (Anders *et al.* 2015) was used for counting reads either contained within gene boundaries, or within specific transcript boundaries when implementing splicing analyses and determining differential exons/isoforms (Figure 2.2). Gene/transcript boundary information was acquired in the form of Gene Transfer Format (GTF) files from the Ensembl genome database (Flicek *et al.* 2014), which provide gene structure information conveying to the relevant reference sequence.

Although HTSeq (intersection_strict) was considered a suitable method for counting reads, it is a conservative method and has several drawbacks. Firstly, this method does not count reads for which any part of the read is aligned outside of an exon. The advantage of this is that it prevents the counting of reads obtained through genomic DNA contamination or from immature unspliced pre-mRNA. However, some of these cases could instead be due to, for example, correct reads with a short adapter (or other) sequence which aligns outside of the exon (which could be allowed with an error/SNP tolerant aligner). Secondly, HTSeq automatically discards reads which align to multiple regions of the genome or align to regions of overlapping gene annotations. The advantage of this is that it prevents ambiguous counting of reads towards an annotation when it is uncertain as to which the read should truly be aligned.

However, this makes certain regions of the genome impossible to quantify, for example in gene families of similar sequence, in tandem repeats, or in overlapping genes. It has been estimated that approximately 3% of genes are completely undetectable despite the use of multiple aligners/counters due to issues such as these (Robert & Watson, 2015). Overall, the counting method used throughout this thesis is conservative, and thus has multiple drawbacks, but whilst it does reduce the number of true positives counted, it also ensures a minimal number of false positives.

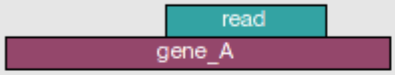
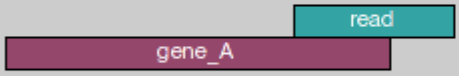
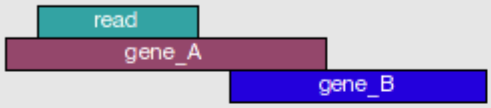
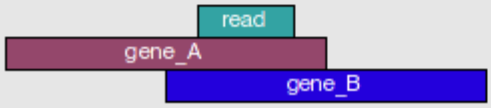
	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

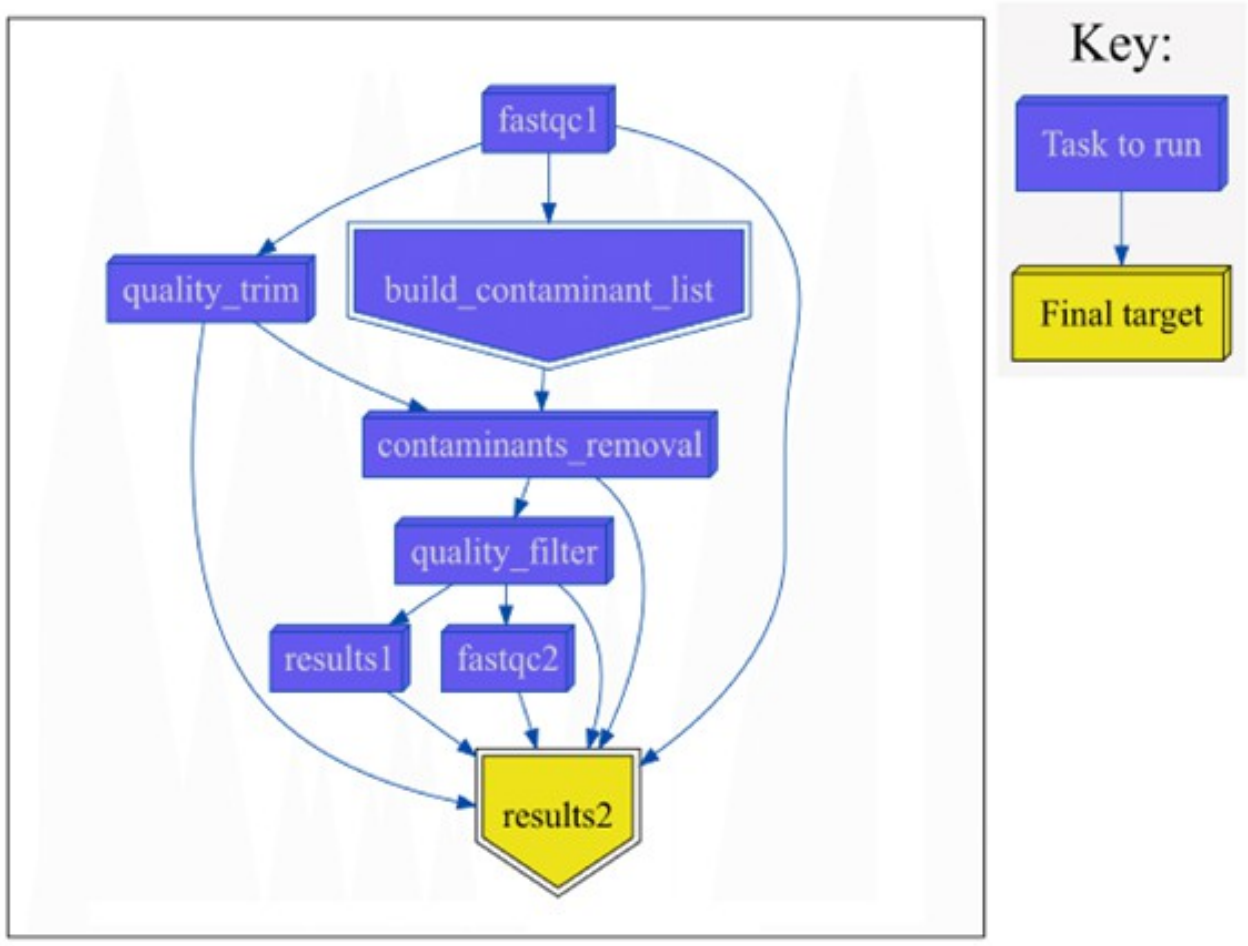
Figure 2.2. A schematic diagram demonstrating the various settings available for HTSeq counting and under which circumstances a read is counted towards a feature. The intersection_strict setting is used in the analyses throughout this thesis. This is the most conservative setting, only counting a read toward a feature when it is completely contained within that feature. This figure is reproduced from *http://www-huber.embl.de/users/anders/HTSeq/doc/count.html.*

Pipeline Formation

The methods presented thus far are a combination of techniques commonly used for RNA-Seq data in forming utilizable data for further analyses, and the software listed are used throughout this thesis. For this, I combined these programs into multiple reusable pipelines, to allow the analyses to be run conveniently and in parallel on multiple files. These were built using a Python computation pipeline library called Ruffus (Goodstadt, 2010). Ruffus provides a series of decorators which track the progression of file names before and after the application of a function, without altering the underlying function itself. This provides multiple advantages, such as allowing the application of the function to many files in parallel, and in case of any errors in the function it also allows for corrections and rerunning of the pipeline from the point of error onwards (rather than starting the whole process from the beginning).

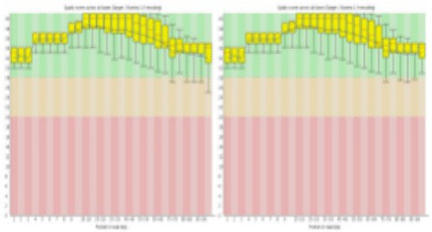
An example of one of these pipelines can be seen in Figure 2.3A, displaying a schematic of the first of these pipelines built for this thesis, which is centered around quality control procedures. An example of a summary file output from this pipeline can also be seen in Figure 2.3B. A second pipeline was also assembled for the read alignment and feature counting, which includes multiple additional options for potential file manipulation and early differential expression analyses (Figure 2.4).

A


fastqc1
quality_trim
build_contaminant_list
contaminants_removal
quality_filter
results1
fastqc2
results2
Key:
Task to run
Final target

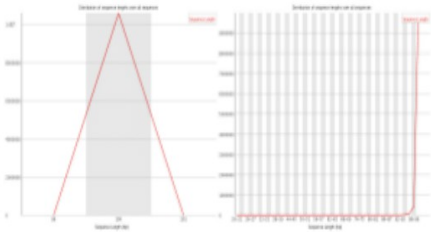
Filename	lane175580_CC--rep1_H7infected_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10561793
Sequence length	100
%GC	47

Discarded 0 (0%) too-short reads.
Before and after pipeline:

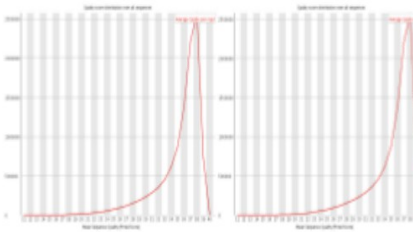


Overrepresented sequences:
0 before pipeline, 0 after pipeline.

Trimmed reads: 0 (0.0%)
Too short reads: 0 (0.0%)



Discarded 0 (0%) low-quality reads.
Before and after pipeline:

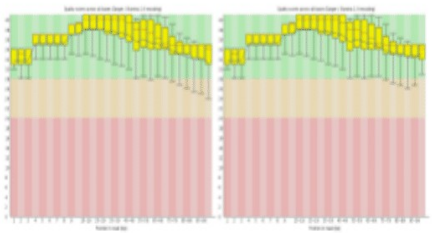


Summary (before processing)

Filename	lane175580_CC--rep1_H7infected_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	10561793
Sequence length	100
%GC	47

Quality Trim

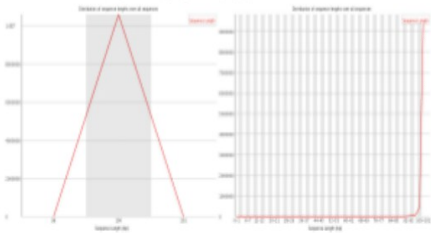
Discarded 160 (0%) too-short reads.
Before and after pipeline:



Contaminants Removal

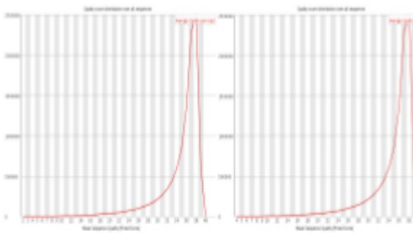
Overrepresented sequences:
0 before pipeline, 0 after pipeline.

Trimmed reads: 0 (0.0%)
Too short reads: 0 (0.0%)



Quality Filter

Discarded 0 (0%) low-quality reads.
Before and after pipeline:

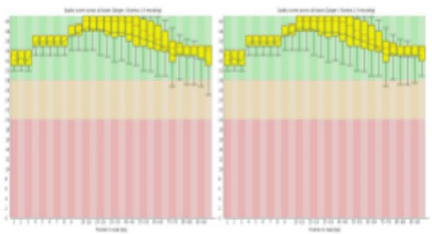


Summary (before processing)

Filename	lane175580_CC--rep1_uninfected_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	9648411
Sequence length	100
%GC	48

Quality Trim

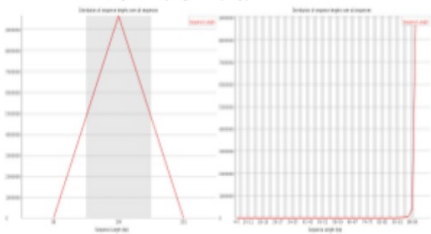
Discarded 0 (0%) too-short reads.
Before and after pipeline:



Contaminants Removal

Overrepresented sequences:
0 before pipeline, 0 after pipeline.

Trimmed reads: 0 (0.0%)
Too short reads: 0 (0.0%)



Quality Filter

Discarded 0 (0%) low-quality reads.
Before and after pipeline:

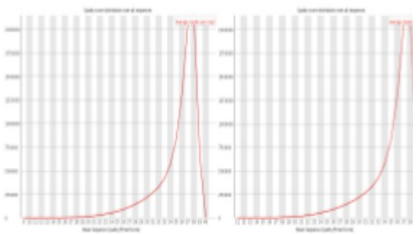


Figure 2.3. A schematic of the quality control pipeline built and used throughout this thesis. The software used includes FastQC visualization, FastQ quality trimming, Cutadapt contaminant removal, and FastQ quality filtering (A). This also includes an option for compiling a custom list of contaminant sequences (as opposed to the one provided with FastQC), and the pipeline ends by performing FastQC a second time and assembling a summary file of before and after visualizations. This summary file includes the sections of the FastQC output relevant to the processes implemented throughout the pipeline (quality trimming, contaminant removal, and quality filtering), displaying how the pipeline has changed these (B).

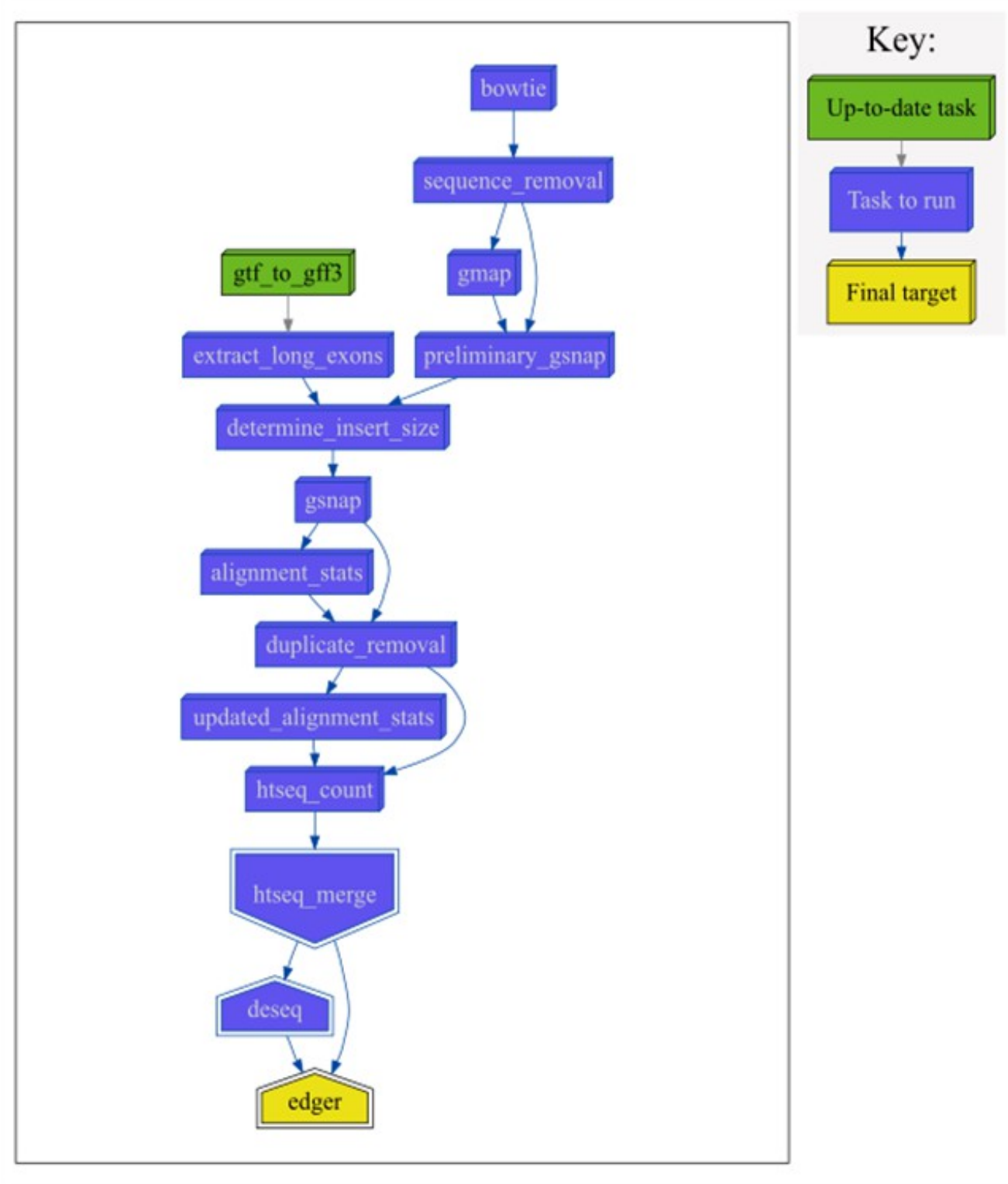


Figure 2.4. A schematic of the read-alignment/read-count/differential-expression pipeline built and used throughout this thesis. The core software used consists of the aforementioned Gsnap alignment and HTSeq feature quantification. Additional and optional software added to this pipeline includes Bowtie2 (Langmead & Salzberg, 2012)

removal of transcriptomic sequences (using Bowtie over Gsnap as transcriptomic sequences do not require spliced alignment), MISO (Katz *et al.* 2010) determination of insert size (including a python script to convert gtf to gff3 files if required), Samtools (Li *et al.* 2009) alignment statistics (and other file manipulations throughout the pipeline, e.g. SAM file to BAM file compression), and Picard tools (Wysoker *et al*. 2013) for duplicate read removal. Some basic early stage analyses have also been added to the end of this pipeline, to be used when appropriate, including DESeq (Anders & Huber, 2010) and EdgeR (Robinson *et al.* 2010) differential expression analyses (see Early Analyses section), as well as principal component analysis and heatmap visualization. These additional stages can be easily switched off in the pipeline if the user does not wish to perform these.

Early Analyses

Some methods commonly used throughout this thesis and subsequent to the application of the pipelines include differential expression analysis, and enrichment analyses.

Differential expression analysis is that which seeks to identify genes that are expressed at different levels in different conditions. This testing can be used in the end stages of the standardized pipeline I developed when appropriate, with the process automated to perform the analysis across all paired combinations of conditions. Two different R packages, DESeq (Anders & Huber, 2010) and EdgeR (Robinson *et al.* 2010) (both differential expression packages), are implemented so as to ensure results across packages are robust and reproducible. DESeq and EdgeR are mostly the same in terms of methodology. Specifically, they both aim to test whether the measurements for the expression of a given gene in one condition comes from a significantly different distribution to another. The null hypothesis of this is that there is no difference, and that in the given gene the expression values in one condition come from the same distribution as another condition. This involves some analysis prior to the statistical test, which approximates a suitable distribution of the expression data. As the expression data is based on count data, a Poisson distribution would normally be most suitable; however (as mentioned previously) there is an overdispersion problem in RNA-Seq data, which is the presence of a greater than expected variance at larger mean counts, and is more suited to the negative binomial distribution. Both DESeq and EdgeR use this model in

addressing the underlying hypothesis with a customised statistical test (a variant of the Fisher's exact test, but suited to this distribution). Overall, the methodology of DESeq and EdgeR are largely the same. However, they differ in two key respects: their normalization of the data prior to analysis, and their estimation of the dispersion for the negative binomial distribution (Anders *et al.* 2013).

Normalization of the data is key in ensuring differences between samples are due to true biological variation as opposed to a simple difference in library size. Classically, this could be done through dividing all values by the total read count (thereby giving resultant values totaling the same in each sample); however DESeq and EdgeR provide more complex and better fitting solutions (Dillies *et al.* 2013). DESeq normalizes the count data with a customised method aimed at maintaining the assumption that the majority of genes are not differentially expressed. DESeq calculates a scaling factor for a given sample as the ratio between the median readcount value across all genes within that sample and the geometric mean across all samples. This is based on the idea that genes which are not differentially expressed should have similar readcounts across all samples, and a resultant ratio of one. EdgeR normalizes the data with Trimmed Mean of M-values (TMM) normalization. TMM also assumes the majority of genes are not differentially expressed, and is calculated as the weighted mean of log ratios between a given sample and all other samples. In lack of differential expression this should be close to one, and if it is not, it provides a correction factor for the total sample size.

The other key area that DESeq and EdgeR differ in is their estimation of the dispersion for the negative binomial distribution to be used in statistical testing. EdgeR "squeezes" feature-level dispersion estimates towards a trended mean, whereas DESeq uses the maximum individual dispersions and the dispersion-mean trend (Anders *et al.* 2013). The overall result of this is a more conservative result from DESeq.

One final method commonly used throughout this thesis is enrichment analyses, such as Gene Ontology (GO) enrichment. There are many functional enrichment analysis tools available and widely used for these analyses (including 68 mentioned in one review alone: Table 1, Huang *et al.* 2009). These rely on various functional annotation databases (such as ontologies or pathways) and their associations with each gene across the genome. A given list of genes can be profiled with these methods by testing for enrichments of functional annotations. The statistical method used by these tools also varies. However, the most predominant of these is the Fisher's exact test or variants thereof (e.g. the hypergeometric test) (Huang *et al.* 2009). The Fisher's exact test (Fisher, 1922) is used to evaluate a contingency table which displays the relationship of two variables and the co-occurrence between the two. In the case of GO enrichment analysis, for any given ontology, this would display the co-occurrence between the genes in the provided list and the genes associated with the ontology. For example, this may look something like the following:

	Genes associated with oxidative stress.	Genes not associated with oxidative stress.
Genes in the provided gene list.	28	146
Genes not in the provided gene list.	257	11,422

A Fisher's exact test can then be used to determine if the number of oxidative stress genes are significantly higher (or lower) in the gene list than in the general genetic population. The test works using an exact method, meaning all possible permutations (all possible tables resulting in the same total row and column sizes) of the observed data are enumerated, and the probability of each calculated. Contingency tables with fairly equal values across the table are more likely, and hence will return the highest probability. The final p-value of the test is then calculated as the sum of the probabilities of the observed table and all those of a lower probability (thereby providing the overall probability of observing a table this rare or rarer).

Many enrichment tools will also require a "background" gene list to be specified. These are the genes selected to be included in this test, and are key in minimizing bias in the result. If a bias has been introduced in obtaining the gene list (for example, if these genes were identified in mouse and are now being tested for enrichment in humans, the list would be biased to those with orthologs only), then the background should consist of only genes that could be identified with the bias in place (in this example, only genes with human-mouse orthologs).

Chapter 3 – Identifying Molecular Signatures of the Placode-Derived Cranial Sensory Ganglia in the Developing Chick.

Abstract

The cranial sensory ganglia embody neuronal populations with distinct functions, or sensory modalities. The distinct ganglia produced by the neurogenic placodes in vertebrate development provide a suitable model for the investigation of individual sensory modality acquisition. Presently, there are limited genetic markers available for the examination of the molecular pathways that underlie the developmental process. This study uses RNA-sequencing to explore the transcriptome-wide profiles of the developing cranial sensory ganglia populations in the embryonic chick. Investigation into relationships of the neurogenic placodes confirmed the division into somato- and viscero-sensory neurons, and provided new evidence into the subdivision of the somatic class into general and special somatosensory. Through differential expression analysis across the ganglia, and subsequent stringent selection criteria, this study also identifies robust genetic markers for the placode-derived cranial sensory ganglia. This constitutes 134 high-confidence unambiguous markers for the trigeminal, vestibulo-acoustic and epibranchial neuronal populations, of which 113 are novel. 21 markers were also confirmed by *in situ* hybridisation (of 26 further considered). These markers can be used in investigations into the acquisition of specific cranial sensory modalities.

Publication and Contributions

This chapter is based on the work described in my joint lead-author paper published in *Neural Development* (Patthey, Clifford *et al.* 2016). Cedric Patthey, Jo Begbie and Sebastian M. Shimeld led procedures prior to RNA-Seq, with Cedric Patthey performing all dissections, fluorescence-activated cell sorting, and RNA preparations. Sequencing was completed by the Wellcome Trust Centre for Human Genetics. All subsequent computational genomics analysis and interpretations of the results was performed by me, with Wilfried Haerty and Chris P. Ponting advising. Final wholemount *in situ* hybridizations for confirmation were performed by Jo Begbie.

3.1 Introduction

A Brief Overview of Germ Layer Formation

Throughout embryogenesis in vertebrates, fate maps of structural formations in different species all exhibit slight variations of the same basic plan (Wolpert *et al.* 2015). Following the formation of a zygote at fertilization, mitotic cell divisions result in the formation of what is known as a blastula: a hollow sphere of cells (blastomeres), surrounding an inner cavity of fluid (the blastocoele). This single-layered blastula then eventually undergoes a phase called gastrulation, during which it develops into a multi-layered structure. At this point, invertebrate organisms with radial symmetry will produce two germ layers, the endoderm and ectoderm, whereas vertebrates (of bilateral symmetry) form three germ layers, the endoderm, mesoderm and ectoderm (Stern, 2004).

The endoderm is formed as cells migrate inward to form a characteristic layer. These cells proceed to form the epithelial lining of the gastrointestinal tract, the glands for the pancreas and liver that open into the gastrointestinal tract, as well as the lining of the lower respiratory tract including the trachea, bronchi and alveoli. The endoderm also contributes to the linings of the endocrine glands, auditory system and urinary system. The mesoderm is formed on top of the endoderm as a secondary layer, and proceeds to form somites: blocks of mesoderm forming rostrocaudally and leading to vertebrate segmentation in the form of vertebrae of the spine, rib cage, etc. Somites then form the

myotome, sclerotome, and dermatome, which all in turn form the muscle tissue, cartilage/bone, and skin tissue (respectively) of the organism. Finally, the ectoderm is the outermost germ layer which proceeds to form multiple structures including the central nervous system, peripheral nervous system, the epidermis and subcutaneous glands. The ectoderm splits into three well studied key structures: the surface ectoderm, the neural crest, and neural tube (Stern, 2004; Wolpert *et al.* 2015).

Neurogenic placodes and Cranial Nerves

Neurogenic placodes are transient embryonic structures of thickened columnar epithelium, which can be identified as focal thickenings in the embryonic ectoderm, documented as structures fundamental to the formation of the cranial sensory systems (Ariëns-Kappers *et al.* 1960). The neurogenic placodes proceed to form, or contribute to, the paired sense organs of the head, as well as five of the twelve cranial nerves. Those which contribute to the paired sense organs include the olfactory placode, forming the olfactory neurons and supporting cells of the nasal epithelia, the lens placode, forming the lenses of the eyes, and the otic placode, forming the hair cells of the inner ear. Lateral line placodes also exist in anamniote vertebrates and give rise to a lateral line system of mechanoreceptors, sensing disturbances in surrounding water, and electroreceptors, sensing changes in electrical fields. The placodes which contribute to the cranial nerves include the trigeminal placode, the otic placode (contributing to both the paired sense organ and the cranial nerve), and the three epibranchial placodes: geniculate, petrosal and nodose (D'amico-Martel &

Noden, 1983; Schlosser, 2006; Figure 3.1). The remaining cranial nerves form from non-placodal origins in the neural crest, and have traditionally been easier to study, and hence gained more attention, due to their compact receptor cell arrays and innervation (Barlow, 2002). This study will focus on the five cranial nerves with placodal-origins, and the molecular signatures of these sensory placodes in contributing to cranial nerve formation.

The positions where the placodes of the cranial nerves form in the embryo are very stereotypical of their roles (Begbie *et al.* 1999; D'amico-Martel & Noden, 1983; Figure 3.1). The trigeminal placode forms two lobes together in close proximity to the Central Nervous System (CNS) around the midbrain-hindbrain junction. The otic placode forms directly caudally of the trigeminal. The epibranchial placodes all form above the branchial clefts, dorsal and caudal to the apposition sites between the ectoderm and the endodermal pharyngeal pouches. The geniculate placode forms over the first pouch, petrosal over the second, and nodose over the caudal pouches (Graham & Shimeld, 2013).

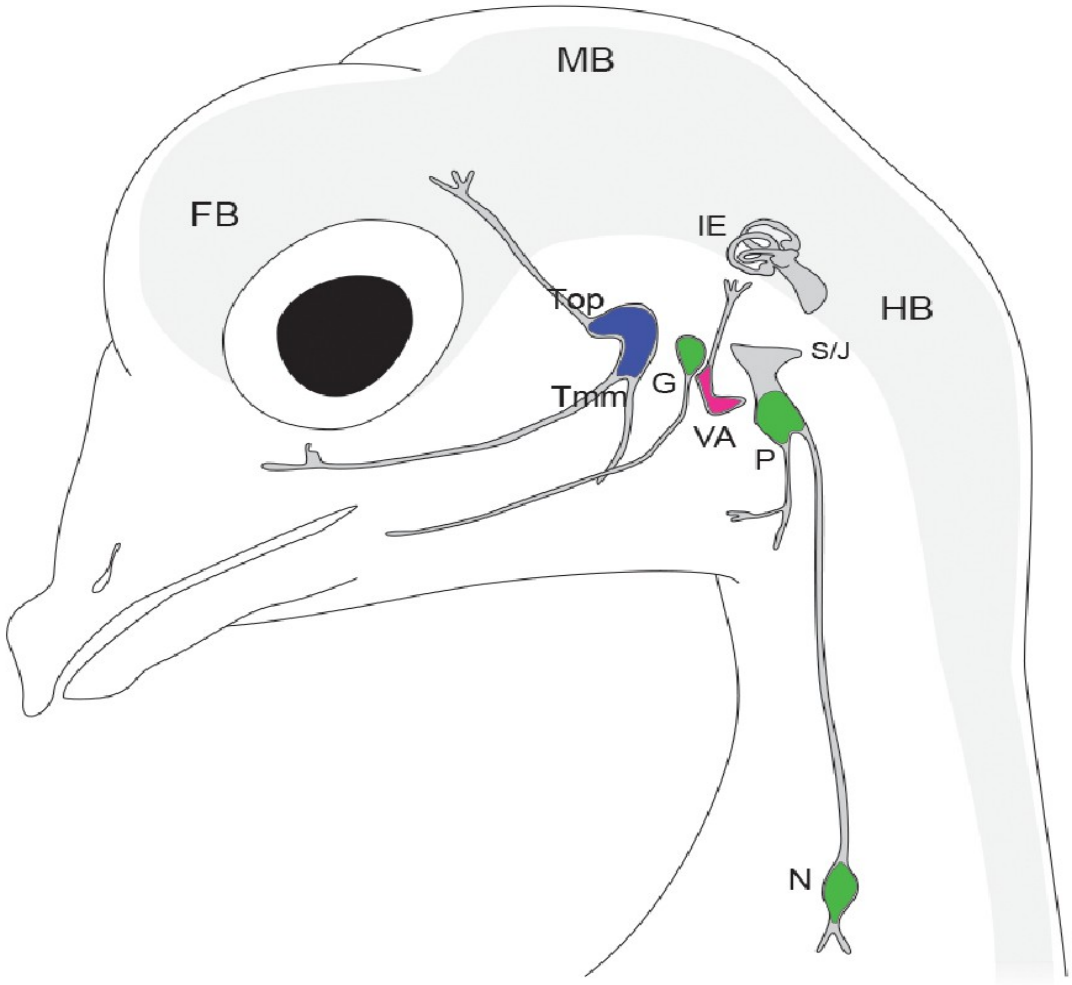


Figure 3.1. A schematic of the cranial sensory ganglia at HH38 (embryonic day 12) in the developing chick. The two lobes of the trigeminal, the ophthalmic and maxillo-mandibular (Top and Tmm), are indicated in blue as the general somatosensory ganglia; the vestibulo-cochlear (VA) are indicated in pink as special somatosensory; the epibranchial placodes, geniculate, petrosal and nodose (G, P and N), are indicated in green as the viscerosensory. Additional labels have been added for reference, including the forebrain (FB), midbrain (MB) and hindbrain (HB), as well as the position of the inner ear (IE), and superior-jugular ganglionic complex (S/J) (a well-marked ganglionic enlargement of the vagus nerve / P). Figure adapted by Jo Begbie, from D'amico-Martel and Noden (1983).

The trigeminal placode forms the trigeminal nerve (cranial nerve V), which is named as such due to its separation into three main branches later in development. As the trigeminal placode forms, it does so firstly into two molecularly distinct lobes, named the ophthalmic (or the profundal) and the maxillo-mandibular (or the trigeminal). The ophthalmic lobe forms one main nerve branch, which provides the transmission of somatosensory information for the lower forehead, upper eyelids, and nasal mucosa. The maxillo-mandibular lobe forms two main nerve branches, the maxillary and mandibular nerves, both of which also provide somatosensory transmission. The maxillary nerve provides this for the lower eyelid, nose, rostral cheek, upper jaw and teeth. The mandibular branch provides this for the caudal cheek, lower jaw and teeth, and the chin (Walker, 2013).

The otic placode (otherwise known as the vestibulo-acoustic placode) forms the vestibulocochlear nerve (cranial nerve VIII). As previously mentioned, this contributes to the paired sense organ as well as the cranial nerve, forming the hair cells of the inner ear with the former contribution, and providing sensory transmission in the inner ear with the latter. This sensory transmission is provided by two main branches of this nerve, which includes the vestibular nerve, carrying information on equilibrium and balance, and the cochlear nerve, carrying auditory information (Bell *et al.* 2008).

Of the three epibranchial placodes, geniculate forms the facial nerve (cranial nerve VII), petrosal forms the glossopharyngeal nerve (cranial nerve IX), and nodose forms the vagus nerve (cranial nerve X). These are all sensory

neurons, of which the majority of roles relate to afferent innervation (i.e. information from sense organs to the CNS). The facial nerve is implicated in taste in two thirds of the tongue and in facial expression muscle control. The glossopharyngeal nerve is implicated in taste in the final third of the tongue, as well as in motor control of the pharynx and saliva secretion. Finally, the vagus nerve transmits information from the heart, lungs, and other abdominal structures, including information on heart rate and blood pressure (Walker, 2013).

The ganglia of the cranial sensory placodes can also be characterized as having distinct sensory modalities according to the function of their associated cranial nerve. The trigeminal ganglion is mostly involved in touch, pain, and temperature sensation; the otic ganglion is involved in balance and hearing; the epibranchial ganglia are involved in chemical sensing including taste, blood gas levels, and pressure changes in blood vessels (Walker, 2013). Overall, these modalities correspond to general somatosensation, special somatosensation, and viscerosensation respectively (Nieuwenhuys, *et al.* 1998).

The Sensory Placodes and Associated Disorders

The placode-derived cranial nerves also have associations with many disorders unique to each of them, which highlight the importance of understanding the underlying differences between them, and investigating the differential development of these nerves from placodes.

The trigeminal nerve of the trigeminal placode is well-known in medical literature due its association with multiple disorders including trigeminal neuralgia and headaches. Although trigeminal neuralgia is usually caused by complications with surrounding structures (the most common cause thought to be compression by the nearby superior cerebellar artery) (Jannetta, 1967), and therefore the molecular signatures of the trigeminal nerve may not be of use to this disorder, the cause for cluster headaches is unknown and in some occurrences have been found to be heritable. Cluster headaches belong to a group of headaches all associated with the trigeminal nerve, called Trigeminal Autonomic Cephalalgias (TACs), and is named due to the symptoms of recurring and intense unlateral headaches, diagnosed from a minimum of five attacks lasting from 15 to 180 minutes each (Leroux & Ducros, 2008). Molecular genetic causes of cluster headaches remain elusive, however epidemiological surveys have found that first-degree relatives of sufferers have a 5-18 times higher risk of suffering cluster headaches themselves compared to the general population, while second degree relatives have a 1-3 times higher risk (Pinessi *et al.* 2005; Leone *et al.* 2001; El Amrani *et al.* 2002).

The vestibulocochlear nerve also has many associated disorders, including neurofibromatosis type II, and vestibular neuritis. Neurofibromatosis type II is otherwise known as MISME syndrome, standing for Multiple Inherited Schwannomas, Meningiomas and Ependymomas, with symptoms of non-malignant tumours of the nervous system, predominantly in the vestibulocochlear nerve or surrounding tissues, and caused by a mutation of

the tumour suppressor gene Neurofibromin 2, leading to an autosomal dominant disorder (Schroeder *et al.* 2014). Vestibular neuritis, otherwise known as labyrinthitis, is a disorder of the inner ear caused by changes to the vestibulocochlear nerve, resulting in vertigo, nausea due to issues with balance and sometimes hearing loss. There are a wide range of known causes, the main of which is viral, with frequent evidence of a recent viral upper respiratory tract infection that is thought to be causative in many acute cases, and with chronic occurrences theorized by some to be caused through infection by a virus of the herpes simplex family (Cooper, 1993; Arbusow *et al.* 2000). Diseases that cause permanent damage to these nerves also highlight the importance of studying these nerves developmentally, if it is hoped that medical intervention can be successful, for example by stem cell treatments.

The epibranchial placodes also lead to many medical disorders. Neuralgia of the facial, glossopharyngeal and vagus nerves are all potential disorders, leading to pain in the areas corresponding to their respective functions (Jannetta, 1980). Problems with the vagus nerve are also known to cause a vasovagal response, which can lead to malaise or fainting due to a drop in heart rate and/or blood pressure. These problems can include a range of triggers depending on the individual, with the result of one or both of activation of the parasympathetic nervous system and inactivation of the sympathetic, causing the problems in heart rate or blood pressure (Yakushin *et al.* 2014). Finally, problems with the facial nerve are also well documented, such as in Bell's palsy, which manifests as a partial or complete paralysis of one side of the face. The major causes of this are also thought to be due to a

virus of the herpes family, with suspected pathogens including the varicella-zoster virus and the Epstein-Barr virus (Zandian *et al.* 2014).

To further investigate these specific disorders, it is important to question and understand the processes underpinning the specific sensory modalities of the sensory placodes. Understanding their different developmental processes and what make these ganglia phenotypically distinct could not only reveal key processes underlying their various associated disorders, but also provide much needed biomarkers for their distinctive phenotypes. These biomarkers could then be used for the identification of the cell population, cell sorting techniques, and therefore in measuring disorder progression. In addition, these biomarkers could be used to unequivocally determine the fate adopted by cells, for example in aiming to derive sensory neurons from stem cells.

A broad range of markers distinguishing the phenotypes will be required for further experimentation. Even ganglia of the same modality can differ depending on their origin. For example, in mice, investigation of the *Scn10a* gene promoter has shown endogenous expression of the product *Nav1.8* in neural-crest derived, but not placode-derived, cranial sensory neurons (Lu *et al.* 2015). Another example is in nociceptive C-fibre sensory neurons innervating the lung, which are phenotypically distinct depending on their neural-crest or placode-derived origin (Nassenstein, *et al.* 2010). These examples reinforce the importance of biomarkers for placode-derived cranial sensory neurons, specific to both their phenotype and their origin.

Determining Biomarkers for the Placode-Derived Cranial Sensory Ganglia

Recent publications have investigated the transcriptomes of somatosensory neurons (Chiu *et al.* 2014). Nevertheless, these have not encompassed the many other sensory modalities of the cranial sensory neurons. To this end, this study is aimed at developing a data resource to address the scarcity of markers distinguishing the cranial sensory ganglia through RNA-Sequencing of five distinct ganglia: the two lobes of the trigeminal: ophthalmic (Top) and maxillo-mandibular (Tmm); the otic/vestibulo-acoustic (VA); two of the epibranchial placodes: petrosal (P) and nodose (N). Unfortunately, it was not possible to capture the geniculate epibranchial placode (G).

The organization and functions of the placode-derived cranial sensory ganglia are well conserved across vertebrates, and the timing and localization of the distinct ganglia have been carefully documented in the chicken throughout their development (Graham & Shimeld, 2013; Hamburger & Hamilton, 1992). This study takes advantage of this knowledge by analyzing RNA-Sequencing applied to the ganglia in chicken, to characterize these at early stages of differentiation. The detailed understanding of developmental timing available in the chicken made it possible for our collaborators to dissect the ganglia separately at the relevant stage of formation based on their individual timescales. Top, Tmm and VA were harvested at Hamburger-Hamilton (HH) stage 18 (a staging system based on morphological landmarks in the chick; Hamburger & Hamilton, 1992), while P and N were harvested at HH23. These timings ensure a consistent level of differentiation between placodes (Figure

3.2), and therefore can be used to ensure the capture of exclusively placode-derived neurons (for example, in the trigeminal ganglion, where neural-crest derived neurons begin to contribute to the ganglion at later stages). Cells were then subjected to immunostaining allowing isolation of cells positive for Neurofilament-M (an antibody to detect filaments found exclusively in the neuronal cytoskeleton) through fluorescence-activated cell sorting (Figure 3.3; Table 3.1), before RNA extraction and RNA-Sequencing multiplexed in triplicate with the Illumina Hi-Seq 2000 for 100bp paired-end reads. This data was then used for all computational analyses in this study.

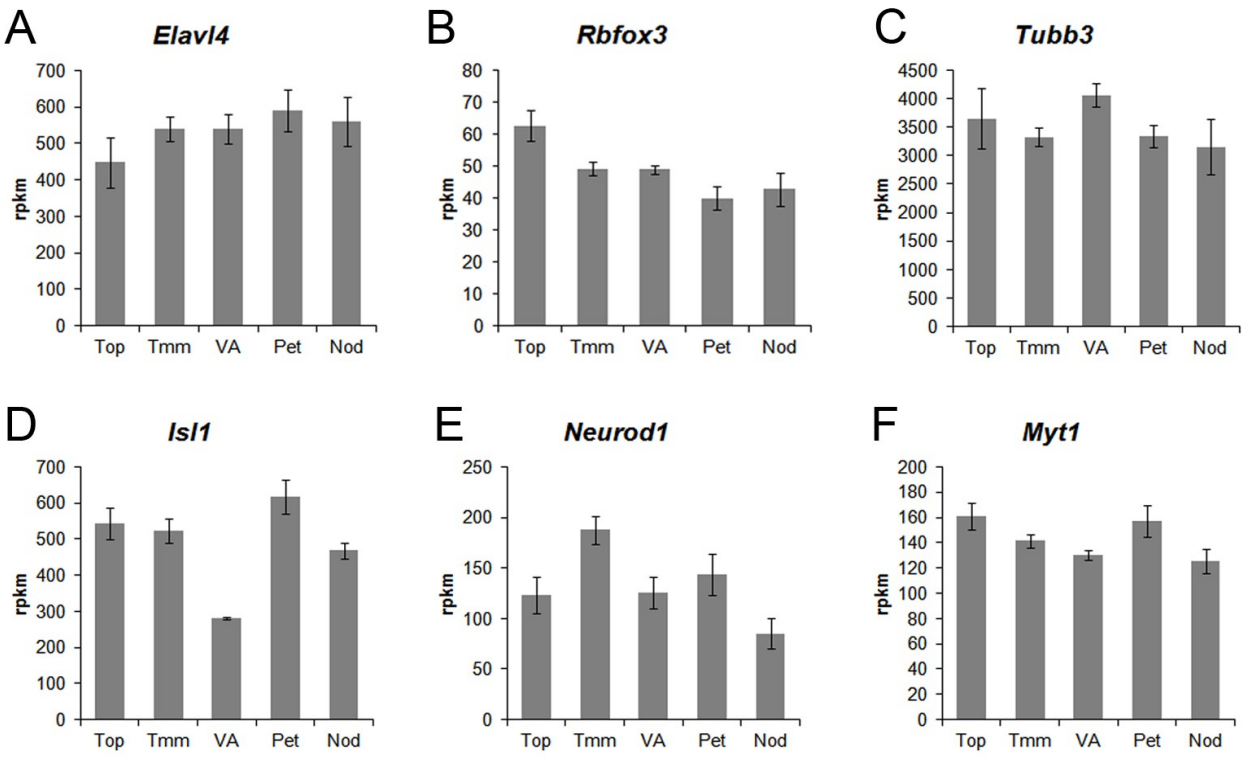


Figure 3.2. Expression levels of neuronal differentiation markers, ensuring samples were taken during the same stage of differentiation for each placode (figure produced by collaborator Cedric Patthey).

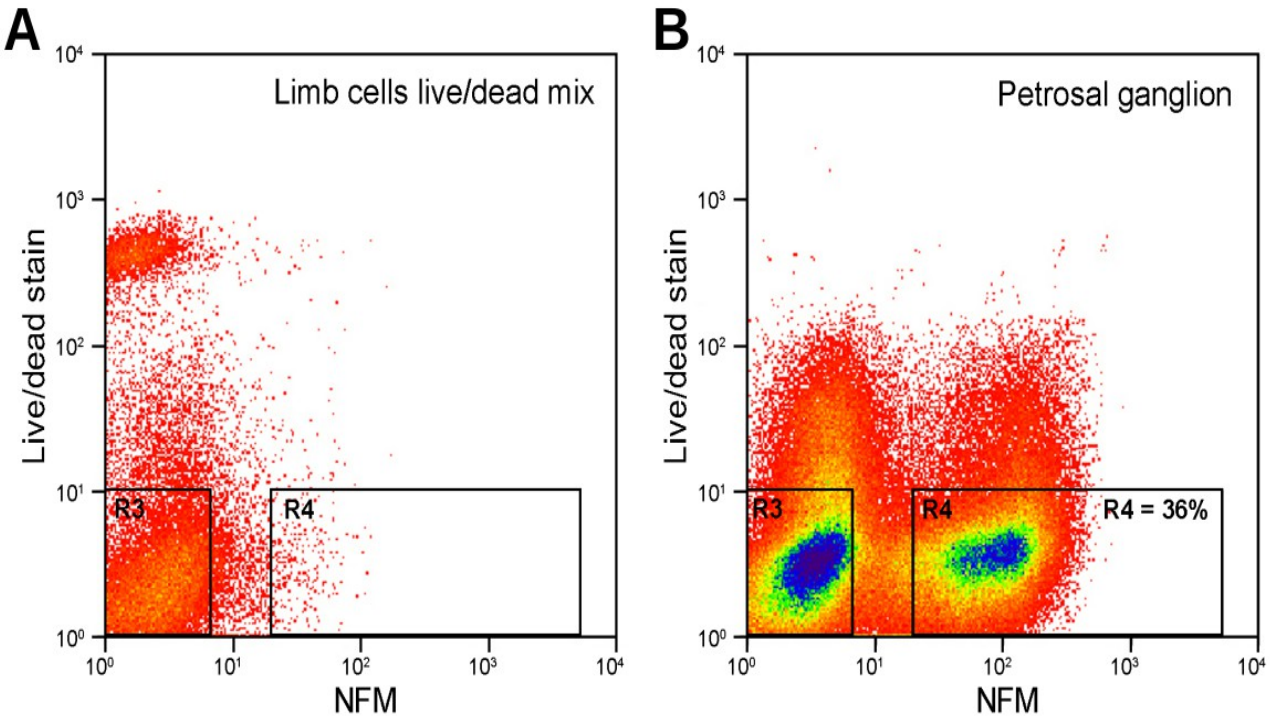


Figure 3.3. Representative FACS plots of cells stained for live-dead stain and the neuronal marker NFM, as performed by Cedric Patthey for RNA-Sequencing, with figure produced by Cedric Patthey. (A) Control cell population: limb bud cells devoid of neurons, containing 50% dead cells. (B) Petrosal ganglion cell population with 36% of cells NFM positive and dead cell marker negative.

Table 3.1. A summary for each of the three replicates of each of the five placodes studied, providing insight into the quantities of biological material retrieved prior to RNA-Sequencing (provided by collaborator Cedric Patthey).

	# Embryos	# Ganglia	Cells Sorted (%)	Cells Sorted (thousands)
Top Rep1	42	83	13.7	50
Top Rep2	48	95	12.6	41
Top Rep3	42	84	14.9	41
Tmm Rep1	42	83	22.2	100
Tmm Rep2	38	75	24.5	120
Tmm Rep3	42	84	21.9	130
VA Rep1	40	80	34.8	84
VA Rep2	38	75	33.8	69
VA Rep3	42	84	29.5	89
P Rep1	55	110	38.1	86
P Rep2	46	92	25.8	20
P Rep3	50	100	28.3	50
N Rep1	55	110	38.2	168
N Rep2	50	100	26.0	110
N Rep3	61	122	41.3	290

3.2 Methods

Read Alignment and Quantification of Expression Levels

Initial sequencing and read quality were assessed using FastQC (Andrews, 2010). Despite poly-A selection prior to RNA-Sequencing, rRNA reads were identified as a potential contaminant. The raw reads were aligned to chicken rRNA sequences and removed from the data-set, followed by the alignment of the reads to the galGal4 chicken genome assembly (Ensembl Release 71) (Flicek *et al.* 2014) with the splice-aware short read aligner Gsnap (Wu & Watanabe, 2005). This was initially performed with no restriction on insert size, from which the expected insert size mean and standard deviation were then determined with MISO (Katz *et al.* 2011). These insert size estimations were then used for guidance in performing a secondary alignment. When relevant (see high confidence marker selection) duplicated reads were identified with Picard Tools' Mark Duplicates function (Wysoker *et al.* 2013), with removal of these reads conservatively ensuring no PCR artifacts. Following this, HTSeq-count (Anders *et al.* 2015) was used (on intersection strict settings) to count uniquely aligned reads using gene models from Ensembl galGal4 reference genome 71. The read-alignment/read-count/differential-expression pipeline (see Chapter 2.3 Pipeline Formation) was used to apply these tools.

Principal Component Analysis (PCA)

PCA was implemented for visualization of sample relationships to one another. Heatmap visualization and hierarchical clustering were also performed to corroborate PCA findings. Hierarchical clustering was performed with the complete (maximum) linkage method and using euclidean distance measures. All of these visualizations were implemented with all count data, with stabilized variance through DESeq. DESeq advises variance-stabilization for all methods of clustering or ordination in multidimensional data where the data is not homoskedastic (i.e. where the variance depends on the mean), which is true in RNA-Seq, for which the variance grows with the mean (Anders & Huber, 2010).

Differential Expression Analysis

Differential expression analysis was implemented through both DESeq (Anders & Huber, 2010) and EdgeR (Robinson *et al.* 2010) to provide a higher confidence in downstream analyses. Both of these provide their own method of normalization, including one based on median log deviation (through DESeq) and one based on Trimmed Mean of Ms (TMM) normalization (through EdgeR). Following normalization, both DESeq and EdgeR were run in a pairwise manner, identifying significantly differentially expressed genes with higher levels of expression in each individual placode (in turn) in comparison to all other placodes as a group, therefore identifying the genes that make each placode unique. In addition to this, we report significantly differentially expressed genes for pairwise comparisons between N/P and Top/Tmm/VA, and between

N/P/VA and Top/Tmm, to identify contributory genes in the sensory modalities. The differential expression analyses were implemented with a q-value (p-value adjusted for False Discovery Rate [FDR]) threshold of 0.05 to control the expected false positive rate under multiple comparisons.

High-Confidence Marker Selection

Markers were selected from differentially expressed genes using strict and conservative criteria to extract a list of robust and high-confidence markers. These criteria were applied to all significantly differentially expressed genes for all pairwise possibilities (not just those representing modalities), allowing us to report high-confidence markers for many ganglia and combinations of ganglia. Firstly, only genes determined to be highly significantly differentially expressed (q-value < 0.01) by both methods of differential expression analysis (DESeq and EdgeR) were considered. Secondly, to fully ensure no PCR artifacts had contributed to selection, both methods of differential expression were re-run with duplicate reads removed post-alignment, and only genes determined to be highly significantly differentially expressed (q-value < 0.01) both with and without duplicate removal (in both methods) were considered. Thirdly, the expression, estimated as Reads Per Kilobase per Million reads (RPKM), was calculated for each gene, and only those with an RPKM of 5.0 or higher in the group of interest (i.e. in the "upregulated" group of the pairwise comparison) were considered. RPKM was used to ensure gene length was not a confounding factor in gene selection. Fourthly, the mean expression level in the group of interest was set at a minimum of 3.0 fold change higher than the mean

expression level in its counterpart (i.e. the group not of interest in the pairwise comparison). This was to ensure the group of interest would be obviously discernible from its counterpart in situ. Fifthly, the expression level for any single sample in the group of interest was set to a minimum of 1.5 fold change higher than any single sample in the opposing (pairwise) counterpart, to ensure a consistent difference between all biological replicates. Finally, when comparing groups of placodes (i.e. N/P vs. Top/Tmm/VA), the mean expression level of the placodes in the group of interest was set at lower than a 2.0 fold change from one another, to ensure consistency across the placodes within the group of interest.

Functional Annotation

Gene ontology enrichment analysis was implemented through DAVID (Huang *et al.* 2009). This analysis was applied to the differentially expressed genes identified, with those identified as upregulated for the group of interest as the foreground, and with a background of all expressed genes. A conservative measure was employed when defining a gene as expressed, taking only those with a minimum of 80% coverage in at least one transcript of at least one sample. This conservative background was then used for GO enrichment analysis with DAVID, with a threshold for significance set at 0.05 for the Benjamini and Hochberg (1995) q-value. Redundant GO terms were also then removed, with the REduce VIsualize Gene Ontology online tool (REVIGO) (Supek et al. 2011), with strict settings on allowed level of GO similarity, allowing only those with a similarity score of 0.4 or below to be retained.

3.3 Results

Read Alignment and Quantification of Expression Levels

RNA-Sequencing produced a mean of ~77 million reads per replicate, ranging from ~30-116 million. Although this is a larger than expected range, there was no evidence of lane bias in the multiplex sequencing. RNA Integrity Number (RIN) values were all ≥ 8.0, ensuring the RNA sequenced was of good quality. FastQC returned no evidence of low quality reads, and so read trimming and filtering were not deemed necessary. Initial alignment determined an average expected insert size (the inner distance between the paired ends) of 116bp with a 72bp standard deviation, values which were then used for re-aligning the reads. This resulted in a mean alignment of ~83% (range: ~66-89%). This percentage appears lower than expected in some samples, and although this could be due to the low quality of the chicken reference genome, we thought it prudent to investigate the effectiveness of the poly-A selection. By aligning the reads to UCSC galGal4 rRNA transcripts, we found ~7-17% of reads aligning to these sequences. Due to the poly-A selection used in library preparation, rRNA contamination should not be present in this dataset. Nevertheless, it is possible that such contamination could be caused by paraformaldehyde fixation of the tissue samples. This could have potentially resulted in a crosslinking event between ribosomes and RNA which were subsequently poly-A selected, leading to sample contamination. Removal of these reads improved the total alignment to the chicken reference genome to a mean of ~88%, ranging from ~69-93%. Following this, counts were calculated over all 17,108 Ensembl gene models,

with an average of 11,800 genes (~69%) with an RPKM > 0.3 (a reasonable threshold for background expression [Ramsköld *et al*. 2009]), ranging from 10,897 (~64%) to 12,190 (~71%). A full summary of results for these metrics is provided in Table 3.2.

Table 3.2. A summary of results for each of the three replicates of each of the five placodes studied, providing insight into the quality of the samples and data. This includes quantities and integrity values (QRNA and RIN) for the RNA used for sequencing, as well as reads returned from sequencing, subsequent mapping percentages, and numbers of genes with base-level expression (RPKM > 0.3).

	QRNA (ng)	RIN	Reads (million)	Mapping (%)	Genes Expressed (RPKM > 0.3)
Top Rep1	70	8.5	35.1	84.4	10,897
Top Rep2	86	8.6	71.5	89.0	11,841
Top Rep3	53	8.0	47.1	69.1	11,580
Tmm Rep1	106	8.2	79.9	88.7	11,892
Tmm Rep2	140	9.4	116.2	92.8	11,867
Tmm Rep3	106	9.5	88.6	91.9	11,891
VA Rep1	137	8.0	81.3	88.4	12,039
VA Rep2	133	8.7	84.8	91.7	11,863
VA Rep3	104	9.0	69.5	90.6	11,765
P Rep1	236	8.8	109.4	89.9	12,097
P Rep2	70	8.3	29.9	86.0	11,313
P Rep3	124	8.8	57.7	89.6	11,821
N Rep1	250	8.7	84.3	90.3	12,190
N Rep2	254	9.3	95.8	90.8	11,864
N Rep3	225	9.3	103.7	91.7	12,082

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was applied to the expression count data for all genes across all samples. The first Principal Component (PC) explain a much greater proportion of the variance (36%) than subsequent PCs (Figure 3.4), and reveals two distinct clusters of cranial sensory ganglia, separating Top, Tmm and VA, from N and P (Figure 3.5A). This reflects the known segregation of somatosensory (Top/Tmm/VA) and viscerosensory (N/P) neurons. Although this segregation also reflects the differing embryonic stages of dissection, six known markers of neuronal differentiation (*ELAL4, ISL1, MYT1, NEUROD1, RBFO,* and *TUBB3*) display similar levels of expression across the two sets of samples, supporting the hypothesis that the clusters reflect sensory phenotype rather than maturation level.

This principal component also displays a moderate separation of the somatosensory ganglia into general (Top/Tmm) and special (VA), which is further corroborated by analysis into further principal components. The fourth principal component shows more distinct separation of VA from the other somatosensory ganglia, explaining 9% of the variance (Figure 3.5B). Overall, these principal components provide evidence for the subdivision of these ganglia, not only on the anatomical level, but also on the molecular level, and that this transciptome-wide data can be used to identify subdivision of the somatosensory modalities. All remaining principal components are shown in Figure 3.6.)

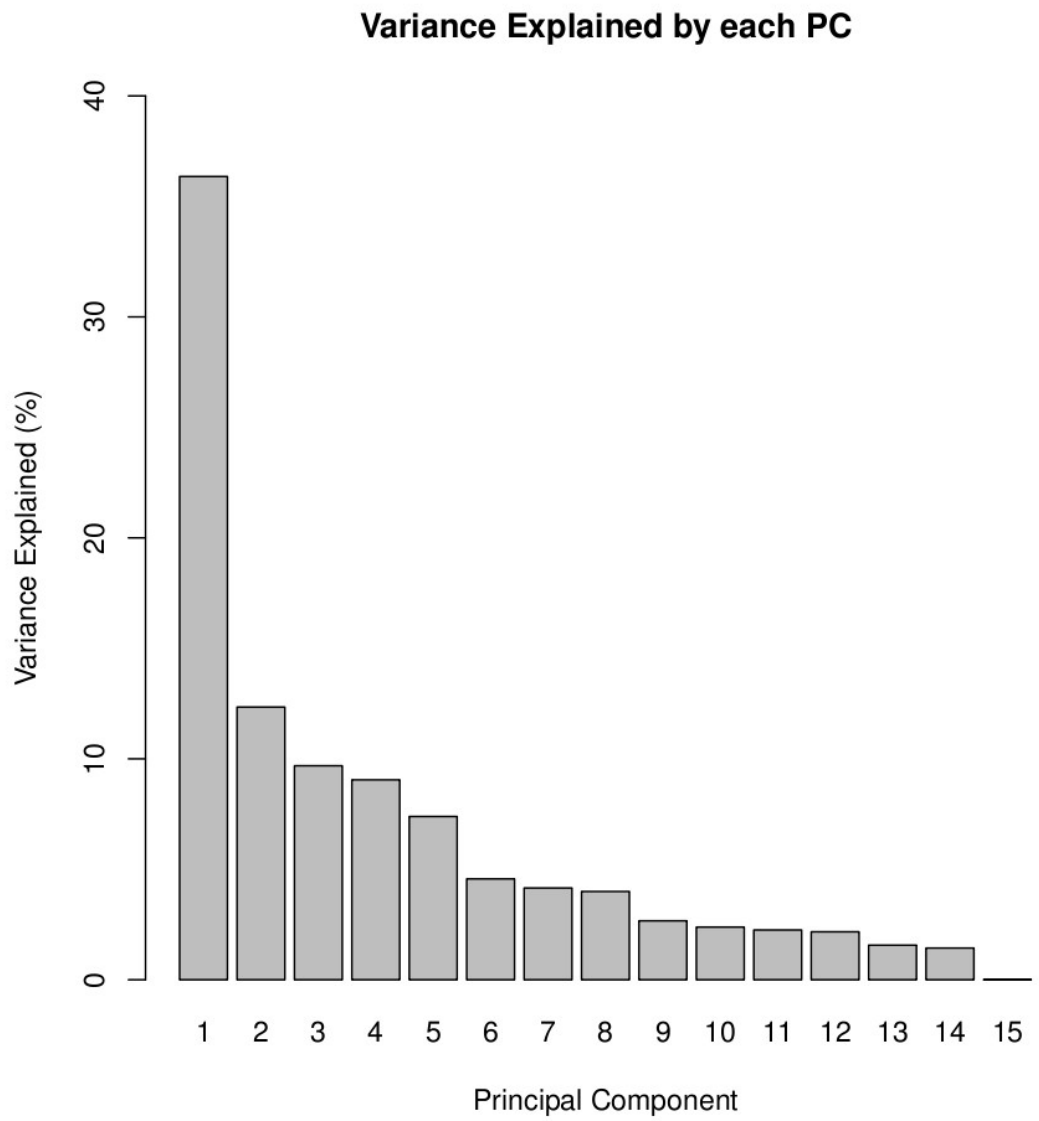


Figure 3.4. A bar plot of the variance explained by each Principal Component for all of the data, showing that PC1 explains a considerably larger amount of the variance than subsequent PCs.

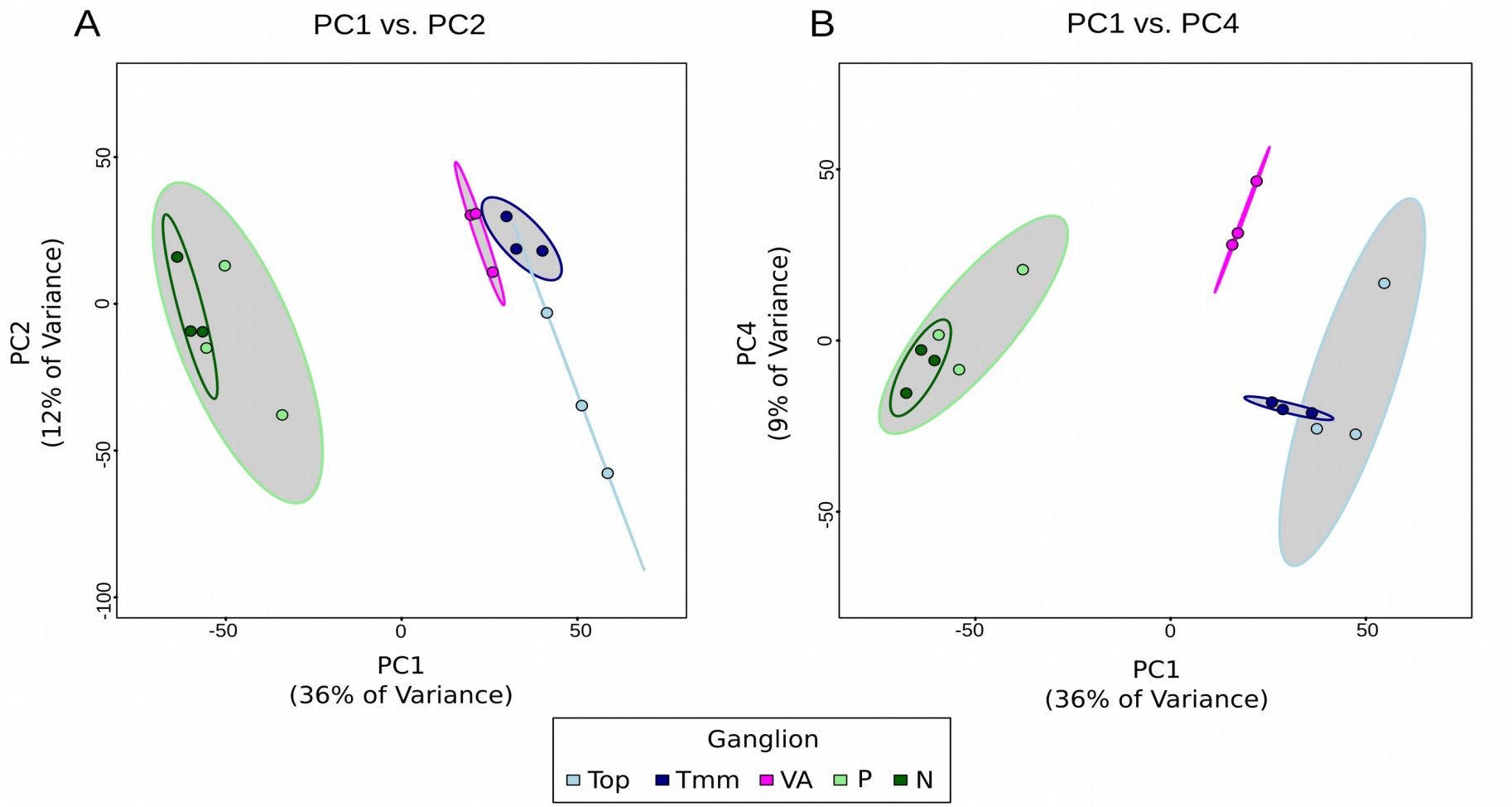


Figure 3.5. Visualizations of some of the major PCs of the data, displaying ganglion-specific clustering. (A) A plot of the first and second PC. (B) A plot of the first and fourth PC. Points and ellipses are coloured by ganglion, with ellipses representing the 90% confidence interval for the groupings.

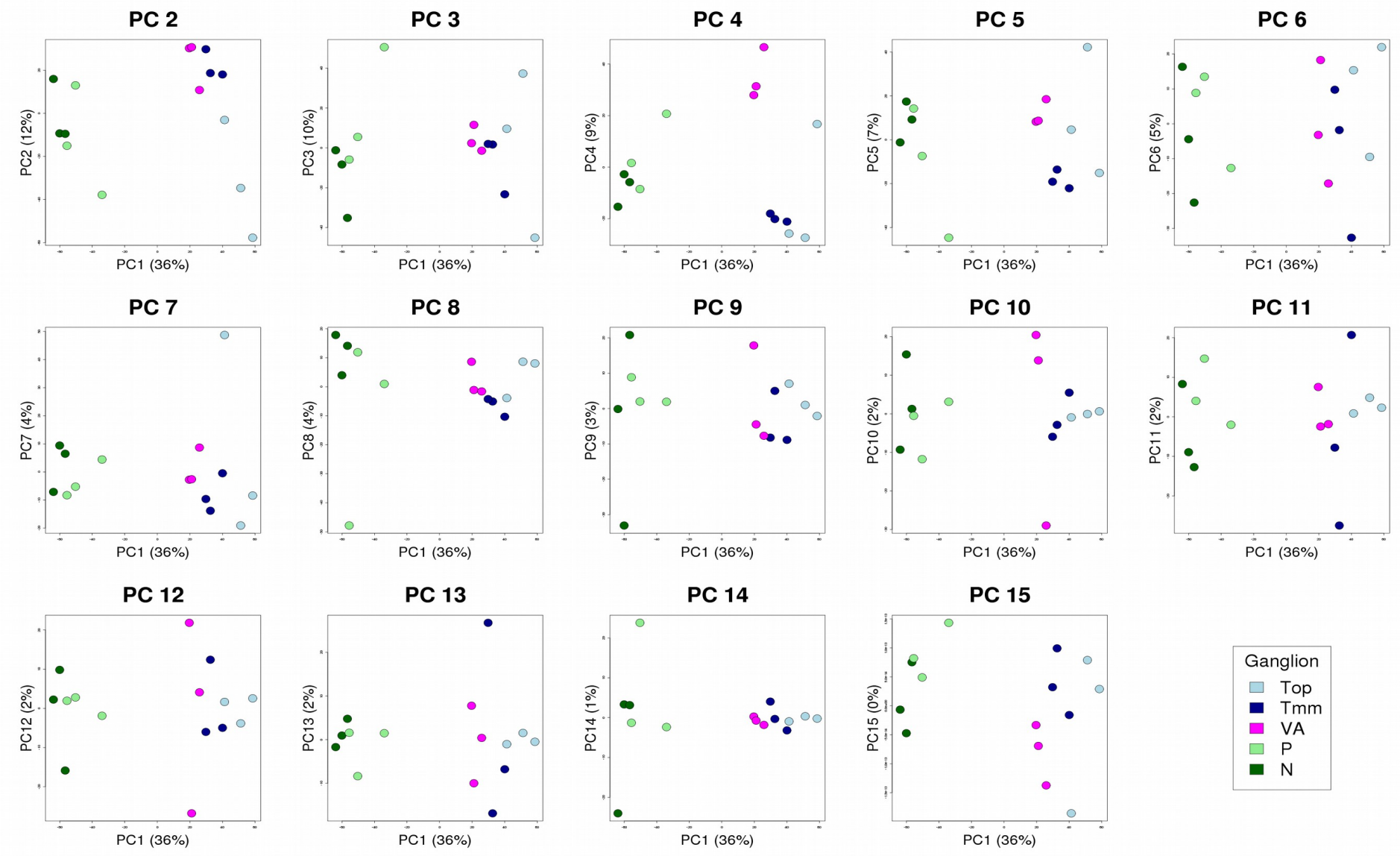
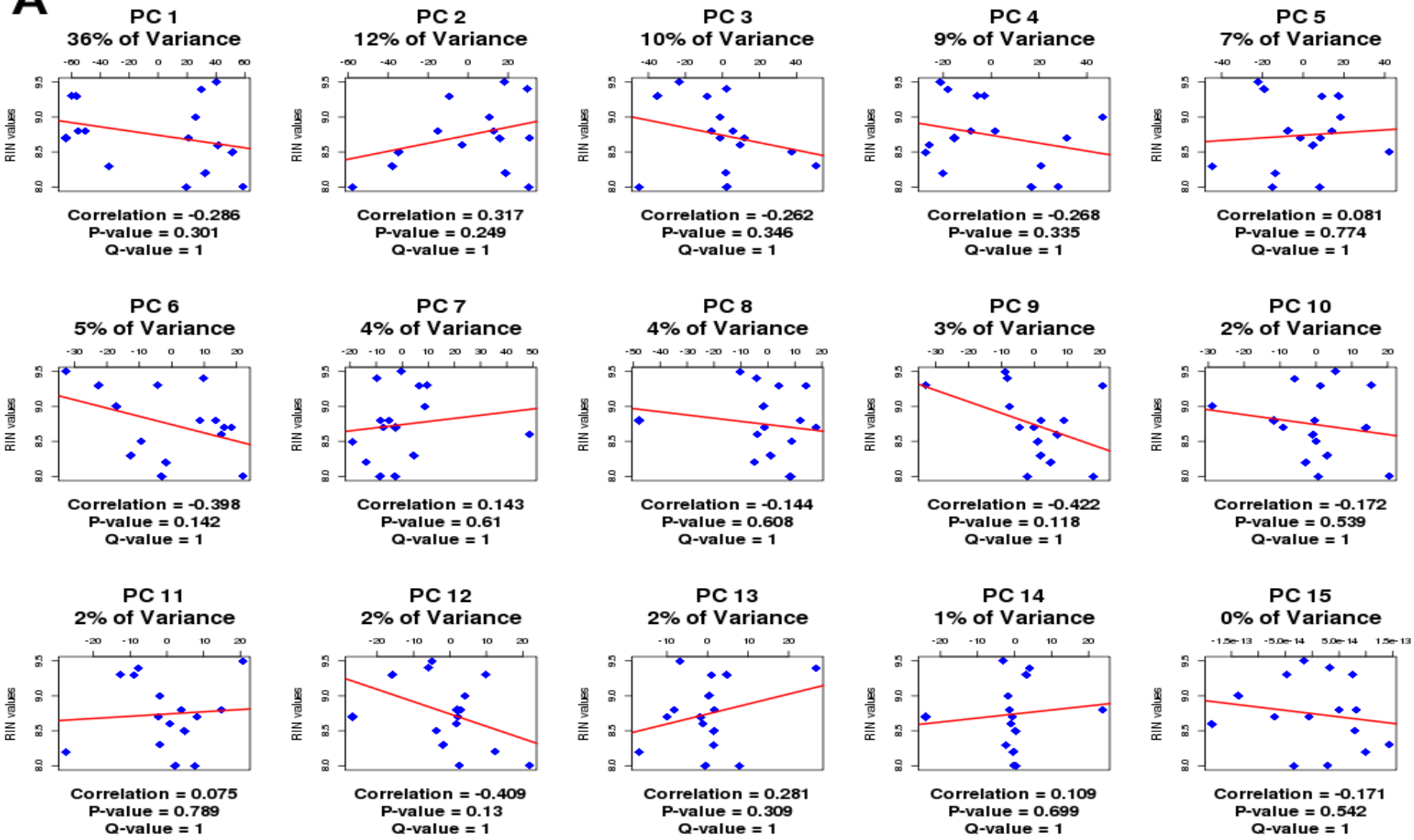


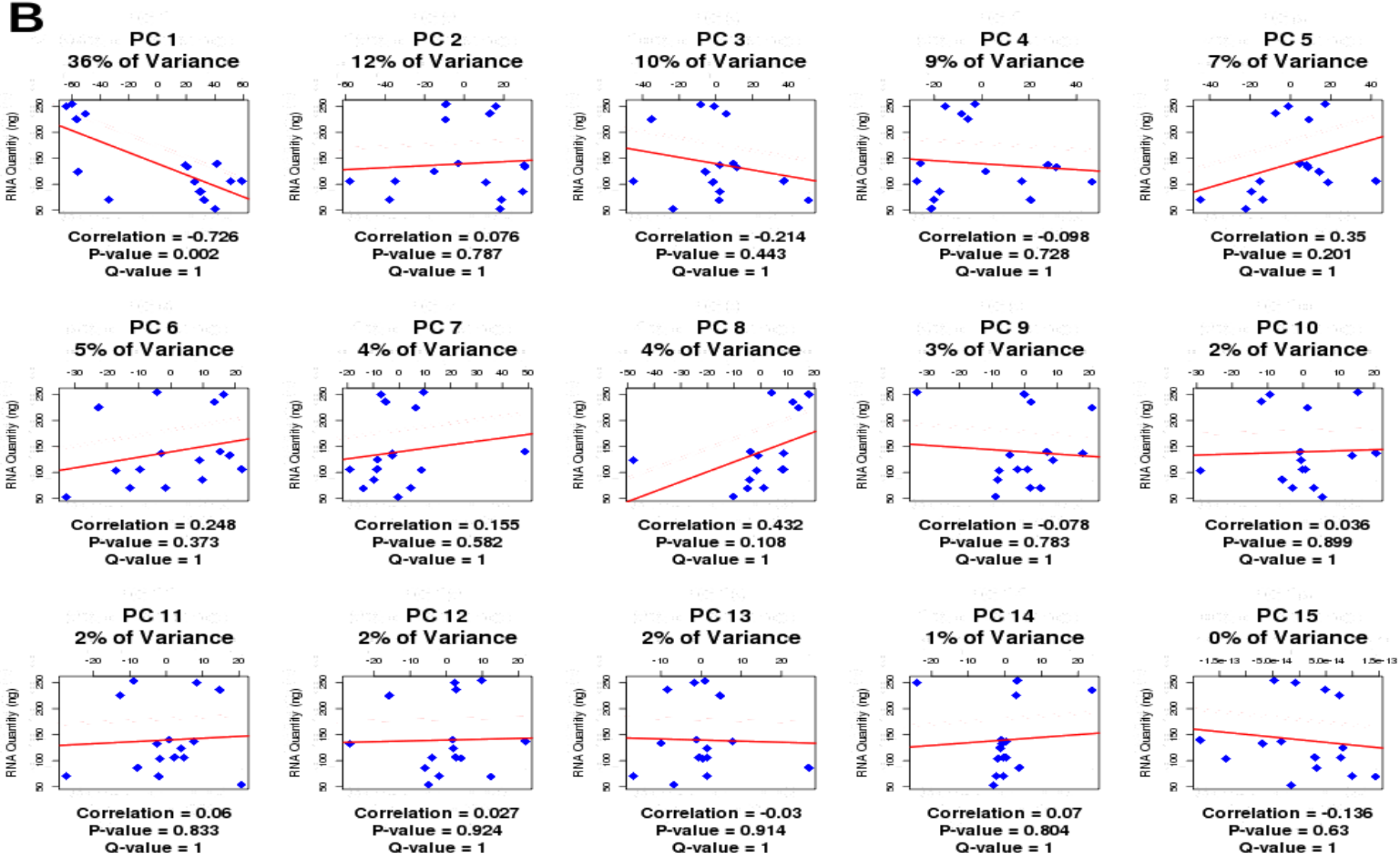
Figure 3.6. Visualizations of PCs 2-15 of the data, displayed against PC1, with points coloured by ganglion.

Finally, all principal components were assessed for correlation with RIN, RNA yield, and sequencing depth, to ensure that there was no statistically significant correlation that could potentially be a confounding factor (Figure 3.7). Of all correlation tests performed, one showed signs of significance, with PC1 showing strong correlation with RNA quantity (Figure 3.7B, PC1). However, this was not statistically significant when correcting for multiple testing, yielding a non-significant q-value. In addition, this is not reflected in correlation tests between PC1 and read depth (Figure 3.7C, PC1), suggesting that this has not had a significant effect on the RNA-Sequencing (or the findings shown in Figure 3.5).

A


PC 1
36% of Variance
Correlation = -0.286
P-value = 0.301
Q-value = 1
PC 2
12% of Variance
Correlation = 0.317
P-value = 0.249
Q-value = 1
PC 3
10% of Variance
Correlation = -0.262
P-value = 0.346
Q-value = 1
PC 4
9% of Variance
Correlation = -0.268
P-value = 0.335
Q-value = 1
PC 5
7% of Variance
Correlation = 0.081
P-value = 0.774
Q-value = 1
PC 6
5% of Variance
Correlation = -0.398
P-value = 0.142
Q-value = 1
PC 7
4% of Variance
Correlation = 0.143
P-value = 0.61
Q-value = 1
PC 8
4% of Variance
Correlation = -0.144
P-value = 0.608
Q-value = 1
PC 9
3% of Variance
Correlation = -0.422
P-value = 0.118
Q-value = 1
PC 10
2% of Variance
Correlation = -0.172
P-value = 0.539
Q-value = 1
PC 11
2% of Variance
Correlation = 0.075
P-value = 0.789
Q-value = 1
PC 12
2% of Variance
Correlation = -0.409
P-value = 0.13
Q-value = 1
PC 13
2% of Variance
Correlation = 0.281
P-value = 0.309
Q-value = 1
PC 14
1% of Variance
Correlation = 0.109
P-value = 0.699
Q-value = 1
PC 15
0% of Variance
Correlation = -0.171
P-value = 0.542
Q-value = 1
RIN values

B


PC 1
36% of Variance
Correlation = -0.726
P-value = 0.002
Q-value = 1
PC 2
12% of Variance
Correlation = 0.076
P-value = 0.787
Q-value = 1
PC 3
10% of Variance
Correlation = -0.214
P-value = 0.443
Q-value = 1
PC 4
9% of Variance
Correlation = -0.098
P-value = 0.728
Q-value = 1
PC 5
7% of Variance
Correlation = 0.35
P-value = 0.201
Q-value = 1
PC 6
5% of Variance
Correlation = 0.248
P-value = 0.373
Q-value = 1
PC 7
4% of Variance
Correlation = 0.155
P-value = 0.582
Q-value = 1
PC 8
4% of Variance
Correlation = 0.432
P-value = 0.108
Q-value = 1
PC 9
3% of Variance
Correlation = -0.078
P-value = 0.783
Q-value = 1
PC 10
2% of Variance
Correlation = 0.036
P-value = 0.899
Q-value = 1
PC 11
2% of Variance
Correlation = 0.06
P-value = 0.833
Q-value = 1
PC 12
2% of Variance
Correlation = 0.027
P-value = 0.924
Q-value = 1
PC 13
2% of Variance
Correlation = -0.03
P-value = 0.914
Q-value = 1
PC 14
1% of Variance
Correlation = 0.07
P-value = 0.804
Q-value = 1
PC 15
0% of Variance
Correlation = -0.136
P-value = 0.63
Q-value = 1
RNA Quantity (ng)

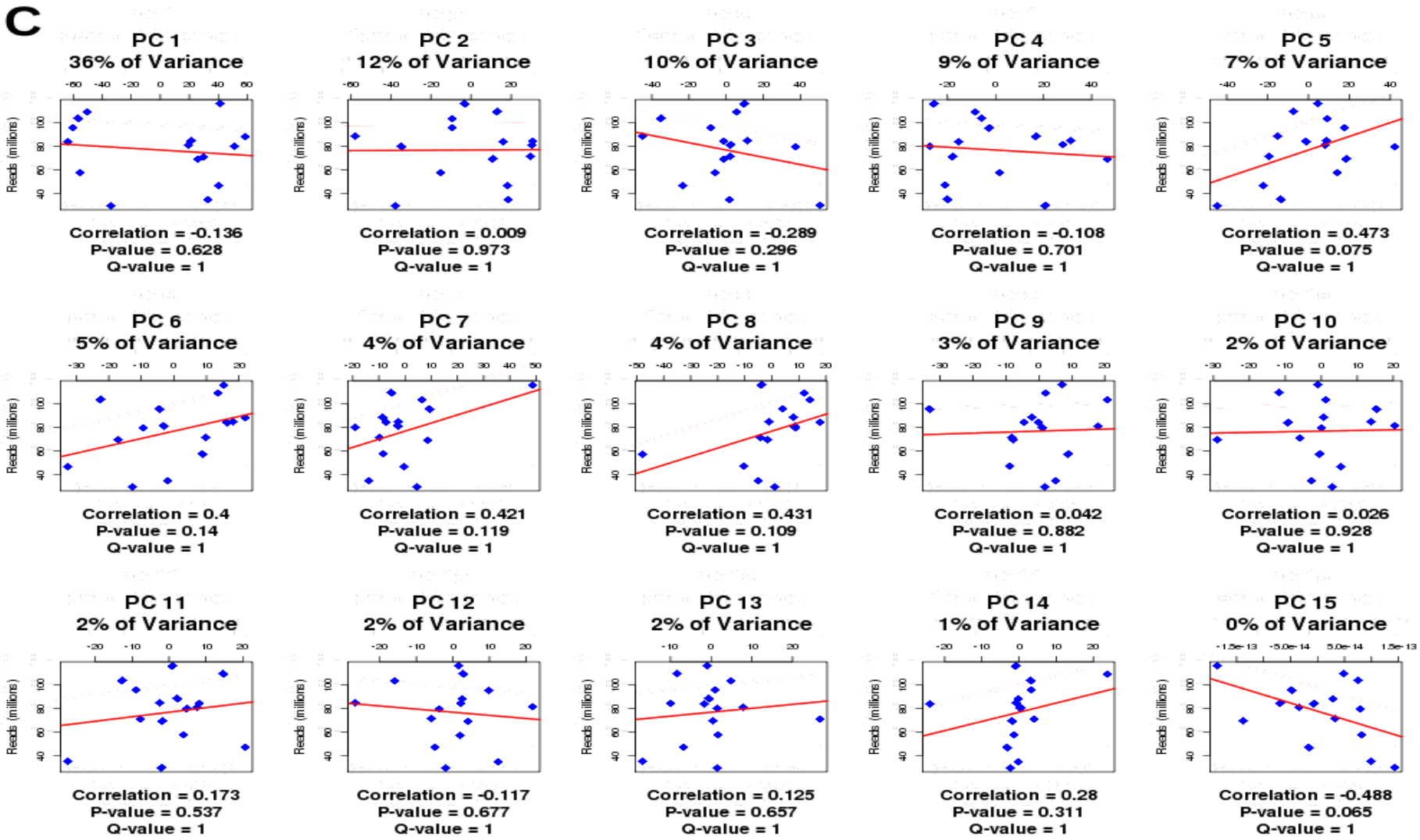


Figure 3.7. Scatter plots investigating each PCs correlation with RIN (A), RNA quantity (B), and number of reads (C). With a linear model fitted (in red), and providing correlation value (r), p-value and q-value.

The relationships seen in the PC analysis are further confirmed by visualization through heatmaps and by hierarchical clustering of the data (Figure 3.8). The first two major clusters are formed of P/N, and Tmm/Top/VA. The P/N cluster shows no specific distinction between the two placodes. The Tmm/Top/VA cluster shows a separation of two of the Top replicates (although the other is integrated with the Tmm replicates), and then Tmm and VA form two distinctive clusters. With the exception of the second replicate of Top, this shows an expected separation of replicates, firstly by structure (which placode they belong too), and secondly by modality.

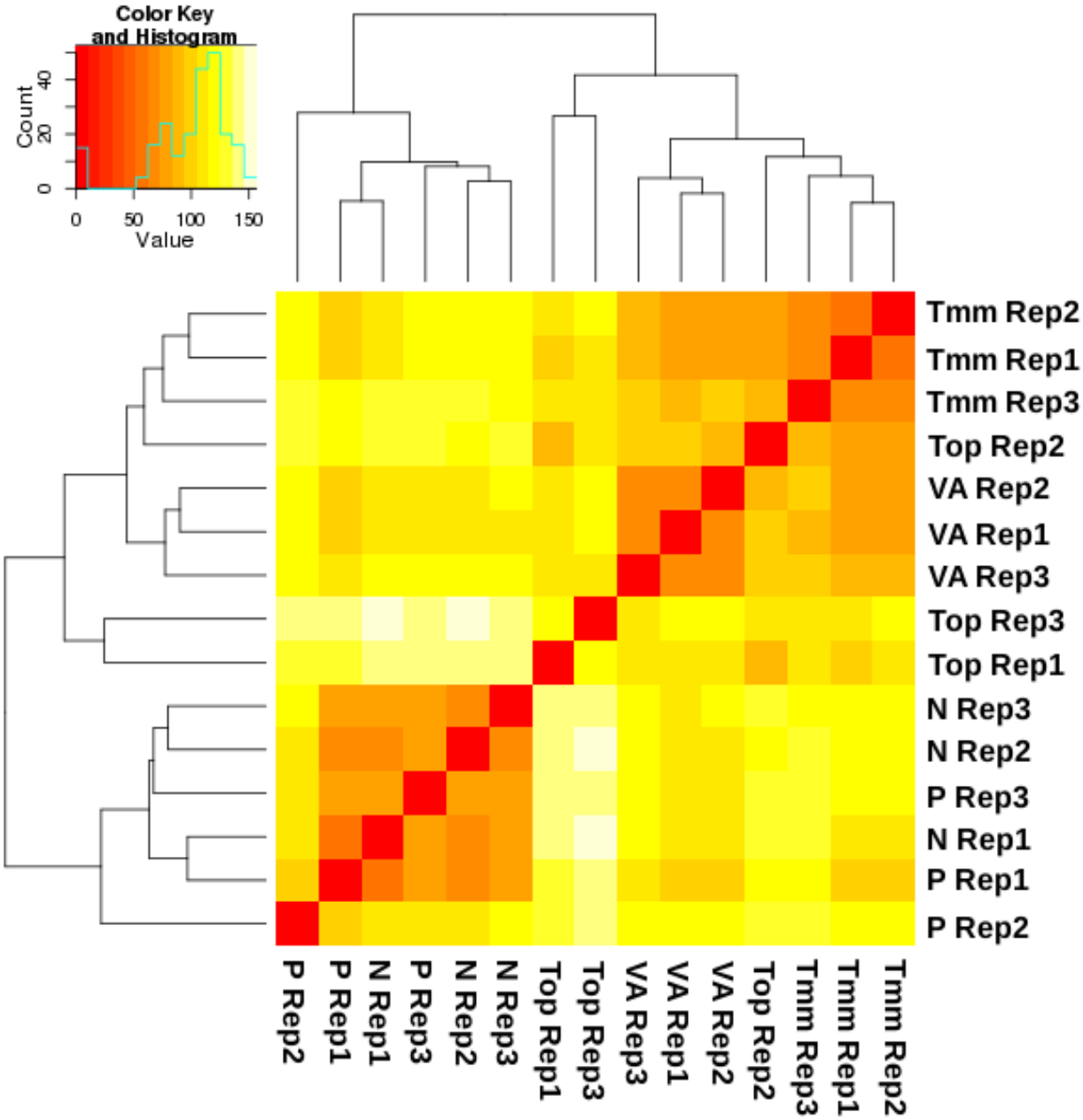
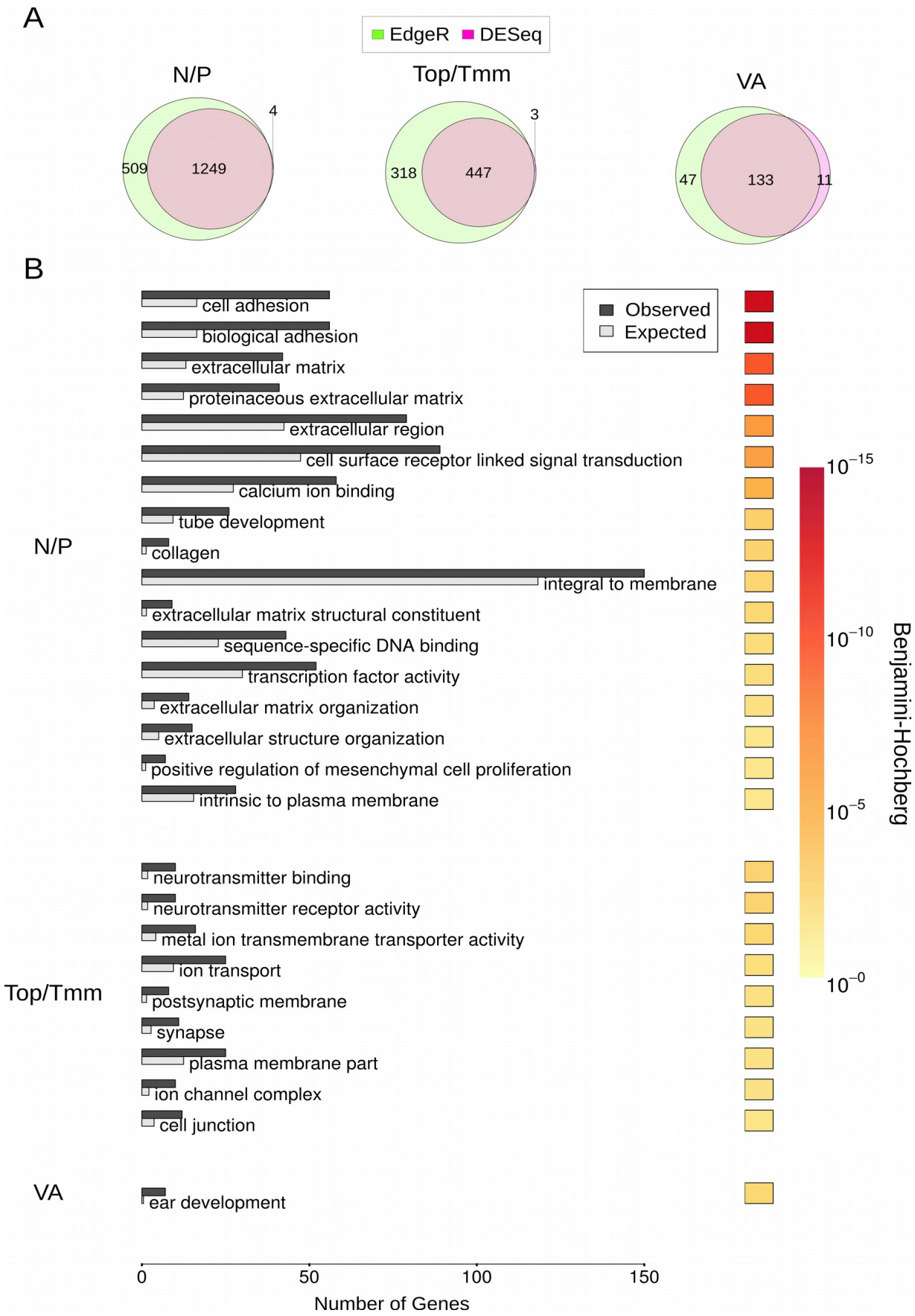


Figure 3.8. A heatmap with conjoining hierarchical clustering dendrogram for all of the count data of all of the replicates of the data set. A histogram key in the upper left corner demonstrates the variance-stabilized count value used for heatmap. The hierarchical clustering was implemented with euclidean distance measure and complete (maximum) linkage method.

Differential Expression Analysis and Functional Annotation

The intersect of differential expression analyses (EdgeR and DESeq) returned a total of 1829 genes separating the modalities, with 1249 genes up-regulated in the viscerosensory (N/P), 447 up-regulated in the general somatosensory (Top/Tmm), and 133 up-regulated in the special somatosensory (VA) ganglia (Figure 3.9A-B). A further 708 differentially expressed genes were identified in tests for all somatosensory ganglia (Top/Tmm/VA), and with additional testing for up-regulation in each unique ganglia, 169 were identified for N, 7 for P, 60 for Tmm, and 107 for Top (Figure 3.9C-D).

Gene ontology functional annotation enrichment analysis was implemented for each of these. In investigating the modalities, the N/P grouping displays a diverse spread of categories, with adhesion and membrane proteins being particularly enriched. The Top/Tmm grouping displayed the most neuronal terms, with enrichments in neurotransmitter, synapse, plasma membrane, and ion transport/channel activity. The VA grouping returned significant enrichment for a single term, however (satisfyingly) this was for ear development (Figure 3.9B). The additional grouping test of Top/Tmm/VA returned functional enrichments similar to that of Top/Tmm, including strong enrichment of genes associated with plasma membrane and ion transport/channel activity terms, as well as transmission of nerve impulse. The functional annotations for each unique ganglion returned further significant enrichments for N and Top, including plasma membrane, neuroactive ligand and cell adhesion related terms for N, and channel/ion transport activity for Top (Figure 3.9D).


A
EdgeR
DESeq
N/P
509
1249
4
Top/Tmm
318
447
3
VA
47
133
11
B
Observed
Expected
N/P
cell adhesion
biological adhesion
extracellular matrix
proteinaceous extracellular matrix
extracellular region
cell surface receptor linked signal transduction
calcium ion binding
tube development
collagen
integral to membrane
extracellular matrix structural constituent
sequence-specific DNA binding
transcription factor activity
extracellular matrix organization
extracellular structure organization
positive regulation of mesenchymal cell proliferation
intrinsic to plasma membrane
Top/Tmm
neurotransmitter binding
neurotransmitter receptor activity
metal ion transmembrane transporter activity
ion transport
postsynaptic membrane
synapse
plasma membrane part
ion channel complex
cell junction
VA
ear development
0
50
100
150
Number of Genes
10^{-15}
10^{-10}
10^{-5}
10^{-0}
Benjamini-Hochberg

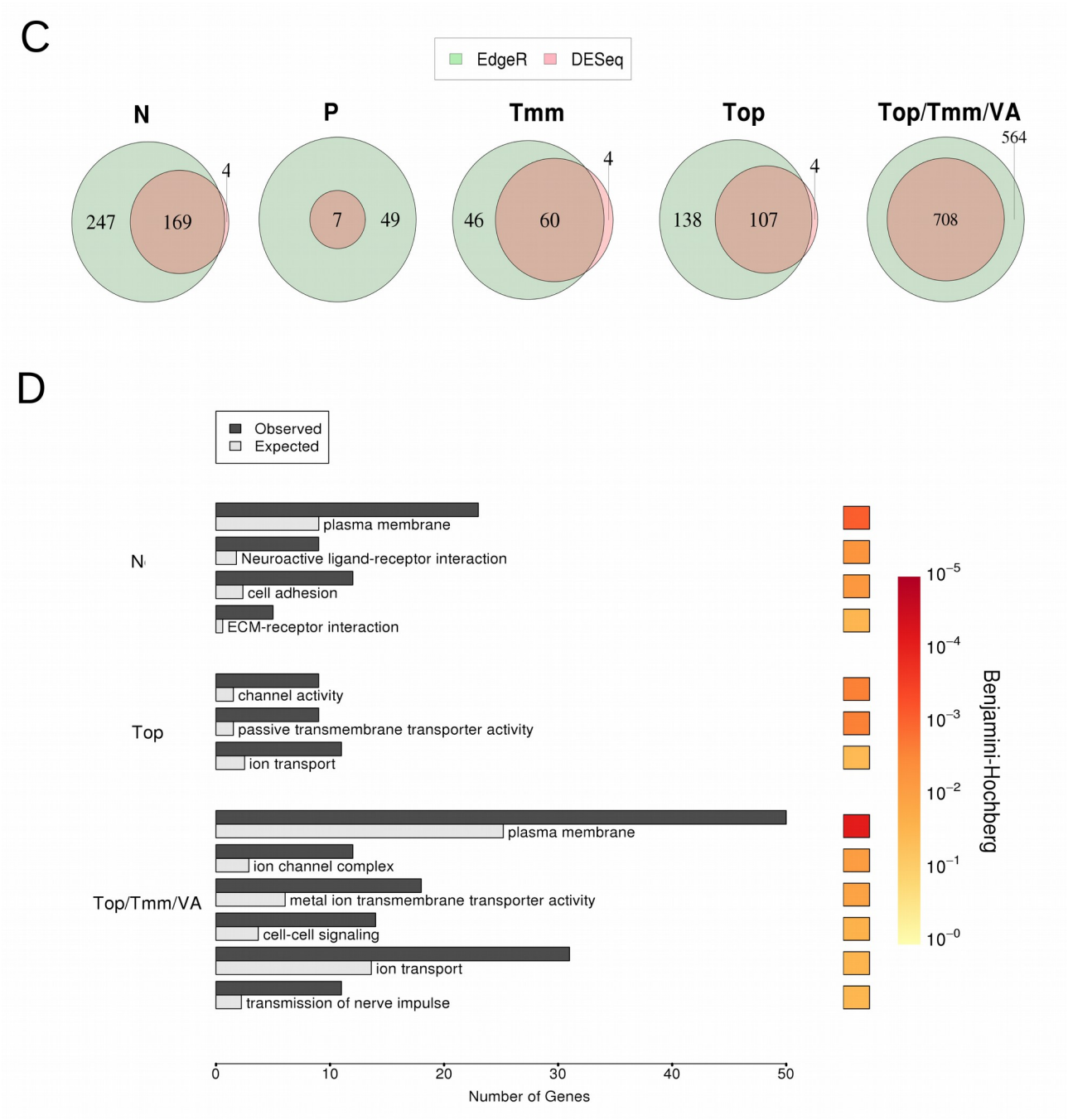


Figure 3.9. The intersection of genes returned by the two differential expression tools, with accompanying significantly enriched functional annotations. Venn diagrams display the intersect for comparisons between modalities (A), including those genes significantly upregulated in N and P, Top and Tmm, and VA. Additional Venn

diagrams also display the intersect for tests of significantly upregulated genes in individual placodes, including N, P, Tmm and Top, in addition to a test for upregulation in the grouping Top/Tmm/VA (C). Below these are statistically significantly enriched gene ontology terms (q-value < 0.05) for the corresponding Venn diagrams above (B and D). This displays the enrichment through the observed number of genes compared to the expected, and with corresponding q-values displayed in an attached Hinton plot. Partially redundant terms have been removed through REVIGO.

High-Confidence Marker Selection and Functional Annotation

In accordance with the methods, all data preparation steps (quality control, mapping, and counting) were re-run with duplicate reads removed as an additional step (Figure 3.10), and the initial selection of markers were taken from the intersect of those genes significantly differentially expressed across DESeq/EdgeR both with and without duplicate removal runs, the results of which can be seen in Figure 3.11.

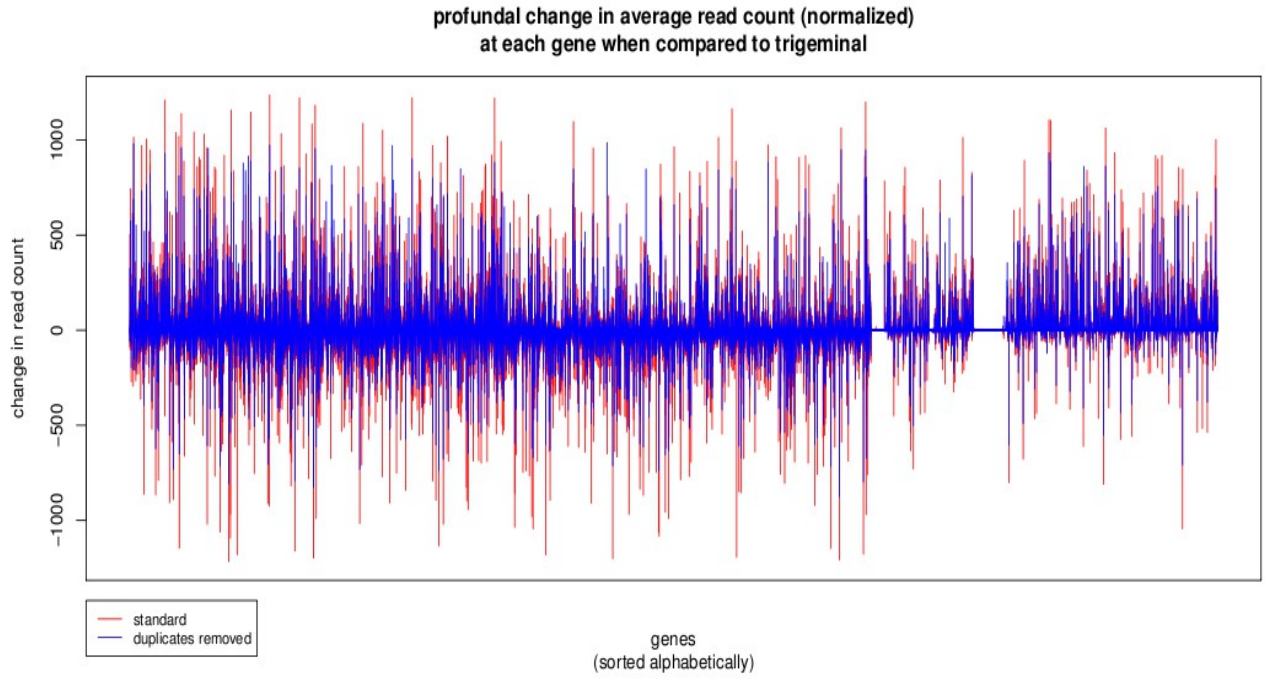


Figure 3.10. A representative bar plot of the effects of duplicate removal on fold change between samples. For this example, these are the mean fold changes for all gene expression values between the profundal and trigeminal ganglia, with original fold changes in red, and fold changes after duplicate removal in blue. In this case, a mean of 45.17% of reads were removed.

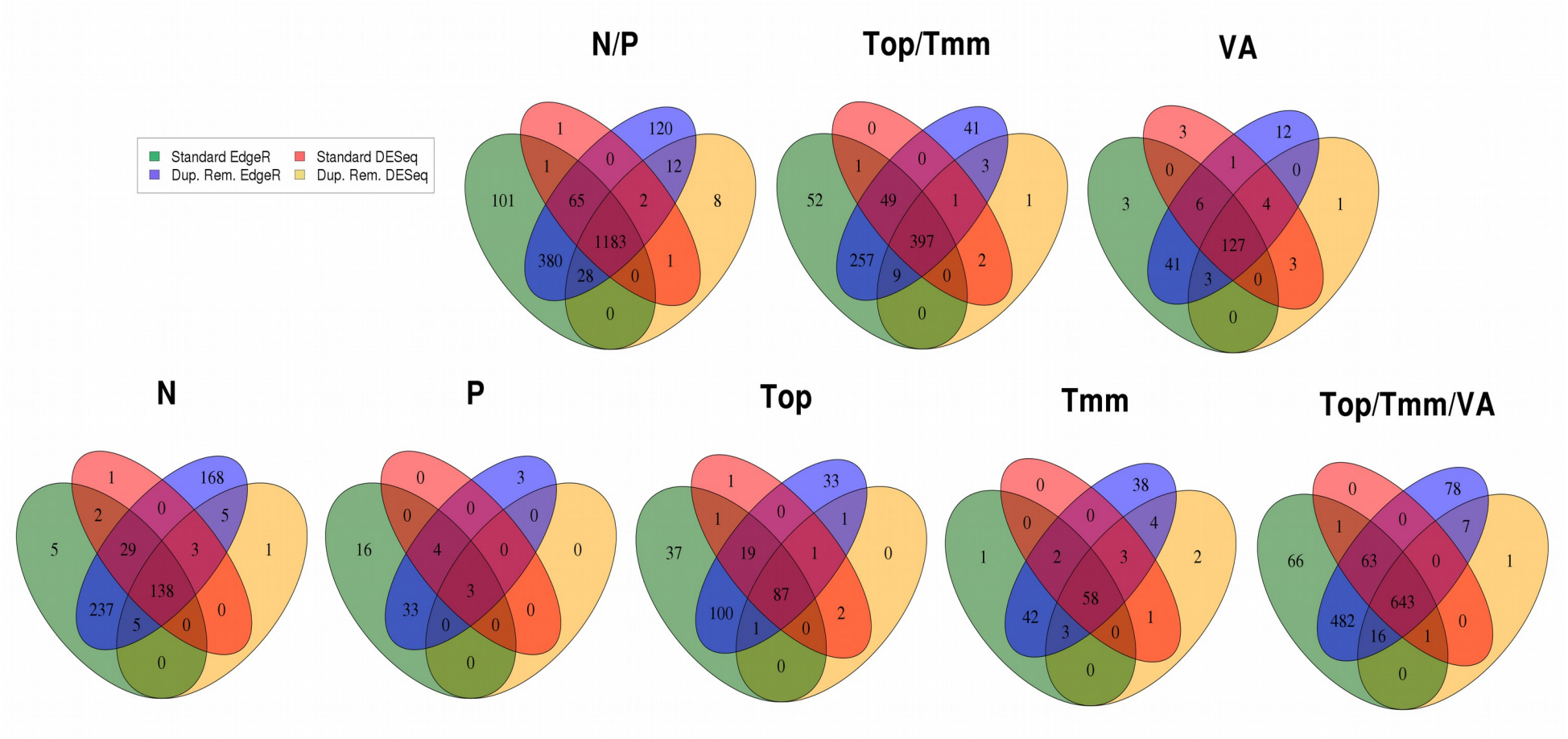


Figure 3.11. Four-way Venn diagrams representing the intersect of the two differential expression methods applied (EdgeR and DESeq), and their intersects with results from data with duplicate reads removed.

The subsequent high confidence marker selection criteria were then applied (see Methods) to return a final total of 134 markers across all pairwise comparisons in differential expression analysis, of which 113 are not previously reported. For the individual placodes, markers were returned for Top, Tmm, VA and N (P returned zero), with this selection process returning a total of 2, 3, 15 and 7 markers respectively. For groups of placodes, markers were returned for Top/Tmm, Top/VA, Tmm/VA, Top/Tmm/VA, Top/P/N, VA/P/N and P/N, with this selection process returning 9, 1, 1, 20, 1, 3 and 72 markers, respectively (Table 3.3).

Table 3.3. All 134 marker genes for all placode groups, including DESeq-normalized count means and the fold change observed.

Upregulated Ganglia	Ensembl ID	Common Name	Mean Deseq-Normalized Read Counts			Q-Value
			Upregulated Ganglia Mean	Other Ganglia Mean	Fold Change	
Top	ENSGALG00000005246	*Pax3*	4828.08	131.84	36.62	4.51E-063
	ENSGALG00000023608	-	168.58	23.33	7.23	4.80E-006
Top + Tmm	ENSGALG00000003005	*Chrnb4*	743.28	29.17	25.48	4.74E-024
	ENSGALG00000003014	*Chrna3*	497.30	41.25	12.06	8.86E-009
	ENSGALG00000003106	*Tox2*	648.73	128.78	5.04	6.47E-021
	ENSGALG00000003895	*Prdm12*	1958.35	200.17	9.78	8.81E-050
	ENSGALG00000007588	*Gad2*	849.69	48.01	17.70	5.06E-064
	ENSGALG00000009012	*Znf385b*	378.47	49.87	7.59	2.39E-020
	ENSGALG00000015597	*Cpz*	609.19	70.15	8.68	1.46E-017
	ENSGALG00000026017	*Drg11*	2238.05	273.67	8.18	1.22E-032
	ENSGALG00000027720	*Otx2*	680.96	22.40	30.41	7.01E-080
Tmm	ENSGALG00000007710	-	101.33	25.74	3.94	5.37E-007
	ENSGALG00000020292	*Gabre*	128.00	31.64	4.05	1.32E-005
	ENSGALG00000021455	*Asip*	76.32	16.18	4.72	7.81E-005
Top + VA	ENSGALG00000016908	*Pcdh9*	348.17	58.92	5.91	2.41E-013
Tmm + VA	ENSGALG00000006245	*Klf15*	66.82	9.70	6.89	2.45E-012
Top + Tmm + VA	ENSGALG00000000420	*Klhdc7a*	689.21	89.59	7.69	6.30E-008
	ENSGALG00000000426	*Igsf21*	1221.65	172.19	7.09	4.85E-023
	ENSGALG00000000912	*Kcnab2*	762.94	153.08	4.98	1.28E-018
	ENSGALG00000001150	-	353.54	59.71	5.92	5.66E-006
	ENSGALG00000002015	*Sncg1*	714.57	91.16	7.84	1.43E-016
	ENSGALG00000004322	-	905.61	44.14	20.52	3.72E-048
	ENSGALG00000005651	*Klhl36*	1264.07	263.65	4.79	3.43E-021
	ENSGALG00000007781	*Gal3st1*	223.23	42.53	5.25	9.69E-017
	ENSGALG00000009514	*Nrcam*	737.79	84.41	8.74	6.22E-013
	ENSGALG00000011236	*Begain*	488.56	65.87	7.42	9.77E-011
	ENSGALG00000012357	*Cxcr4*	590.26	92.29	6.40	9.99E-018
	ENSGALG00000016035	*Gfra4*	3453.71	249.92	13.82	1.50E-015
	ENSGALG00000016826	*Adprhl1*	227.09	18.75	12.11	1.29E-039
	ENSGALG00000020199	*Gpm6a*	1654.07	198.95	8.31	8.59E-022
	ENSGALG00000023638	-	152.39	25.04	6.09	2.62E-008
	ENSGALG00000027241	*Tbxa2r*	695.53	52.19	13.33	8.70E-022
	ENSGALG00000027265	*Ajap1*	130.60	17.81	7.33	9.16E-018
	ENSGALG00000028048	*Ptn*	527.35	17.46	30.20	4.10E-033
	ENSGALG00000028055	-	65.83	10.48	6.28	1.30E-005
	ENSGALG00000028196	-	312.79	5.31	58.88	1.34E-019
VA	ENSGALG00000001181	*Kcnj5*	65.53	9.41	6.96	5.61E-008
	ENSGALG00000001346	-	33.46	6.10	5.49	3.19E-004
	ENSGALG00000007141	*Lrrk1*	821.74	105.46	7.79	7.66E-020
	ENSGALG00000009645	*Esrrg*	1941.14	260.91	7.44	6.64E-020
	ENSGALG00000011009	*Scn2a*	282.42	44.72	6.32	1.32E-014
	ENSGALG00000011040	*Scn3a*	899.24	31.13	28.88	1.76E-063
	ENSGALG00000011166	-	268.23	12.04	22.27	1.88E-034
	ENSGALG00000012303	*Pou6f2*	88.26	4.01	22.03	5.74E-023
	ENSGALG00000012894	*Chrm2*	78.86	11.65	6.77	1.44E-008
	ENSGALG00000013193	*Irx2*	907.99	155.58	5.84	1.84E-014
	ENSGALG00000014872	*Fgf10*	102.51	10.93	9.38	1.44E-016
	ENSGALG00000016022	*Runx1*	581.59	19.17	30.34	1.54E-040
	ENSGALG00000023415	*Hmx3*	119.71	11.49	10.42	1.50E-016
	ENSGALG00000025748	*Fgf9*	677.22	55.59	12.18	2.16E-038
	ENSGALG00000027830	*Cacng2*	1331.21	261.81	5.08	2.97E-007
Top + P + N	ENSGALG00000013022	*Cacna1c*	686.77	141.54	4.85	1.78E-005
VA + N + P	ENSGALG00000001143	*Ets1*	330.11	31.41	10.51	2.86E-020
	ENSGALG00000008716	*Flrt3*	320.56	45.55	7.04	2.80E-008
	ENSGALG00000028821	*Glipr2*	195.17	57.72	3.38	2.46E-008

Upregulated Ganglia	Ensembl ID	Common Name	Mean Deseq-Normalized Read Counts			Q-Value
			Upregulated Ganglia Mean	Other Ganglia Mean	Fold Change	
P + N	ENSGALG00000000327	*Fbn3*	1354.30	143.17	9.46	2.53E-010
	ENSGALG00000000337	*Ctxn1*	272.64	6.43	42.38	1.56E-025
	ENSGALG00000000461	*Mrc2*	560.80	109.27	5.13	2.79E-015
	ENSGALG00000001063	*Prdm16*	693.56	120.99	5.73	4.25E-004
	ENSGALG00000001071	*Eln*	409.99	33.32	12.30	1.84E-011
	ENSGALG00000001434	*Stom*	446.20	52.29	8.53	8.72E-027
	ENSGALG00000001735	*Lhx4*	203.85	12.64	16.13	2.44E-030
	ENSGALG00000001896	*Ntng1*	430.48	65.27	6.60	3.91E-026
	ENSGALG00000001916	*Slit3*	493.99	69.41	7.12	4.59E-034
	ENSGALG00000002086	*Rasl10b*	65.77	4.95	13.28	5.48E-013
	ENSGALG00000002203	*Fgf18*	139.51	18.31	7.62	1.52E-011
	ENSGALG00000002298	*Col8a2*	188.99	28.39	6.66	1.15E-017
	ENSGALG00000003324	*Prrx1*	626.92	15.74	39.82	5.59E-041
	ENSGALG00000003470	*Zmat4*	124.48	16.59	7.50	8.39E-025
	ENSGALG00000003521	*Tpm1*	646.15	131.24	4.92	1.54E-015
	ENSGALG00000003578	*Fn1*	1804.49	153.34	11.77	3.67E-015
	ENSGALG00000003580	*Mmp2*	723.85	100.54	7.20	1.90E-020
	ENSGALG00000003644	-	260.13	49.44	5.26	3.43E-008
	ENSGALG00000004183	*Fam20a*	396.55	73.94	5.36	1.74E-024
	ENSGALG00000004290	*Angptl1*	159.61	5.62	28.42	7.73E-014
	ENSGALG00000004445	*Slc13a3*	131.37	35.23	3.73	1.24E-012
	ENSGALG00000005321	*Lama5*	867.90	94.73	9.16	2.33E-015
	ENSGALG00000005410	*Wnt5a*	307.37	56.39	5.45	5.37E-009
	ENSGALG00000006259	*Kirrel*	1357.09	304.86	4.45	5.67E-008
	ENSGALG00000006584	-	105.25	18.49	5.69	3.80E-004
	ENSGALG00000006700	*Proser2*	160.24	18.42	8.70	1.46E-027
	ENSGALG00000007047	*Gal*	188.18	5.63	33.43	1.79E-038
	ENSGALG00000007392	*Adamts9*	1331.43	83.94	15.86	1.08E-005
	ENSGALG00000007706	*Fgf8*	76.89	8.75	8.78	9.65E-010
	ENSGALG00000007866	*Bcl11a*	824.14	176.42	4.67	1.03E-009
	ENSGALG00000008018	*Snai1*	235.40	12.58	18.71	5.69E-012
	ENSGALG00000008352	*Ckmt1a*	103.69	19.99	5.19	4.67E-015
	ENSGALG00000008661	*Cyr61*	130.19	9.50	13.70	7.96E-005
	ENSGALG00000009437	*Mecom*	1398.70	34.05	41.08	1.04E-006
	ENSGALG00000009530	*Pdzrn4*	2140.92	31.78	67.36	2.26E-106
	ENSGALG00000009612	*Tgfb2*	988.97	158.84	6.23	9.58E-034
	ENSGALG00000009739	*Amigo2*	616.18	27.88	22.10	2.25E-013
	ENSGALG00000010071	*Mab21l2*	1366.45	135.42	10.09	3.80E-053
	ENSGALG00000010294	*Rps6kl1*	947.51	251.44	3.77	1.95E-018
	ENSGALG00000011207	*Mkp3*	451.41	127.73	3.53	3.40E-013
	ENSGALG00000011295	*Socs2*	120.85	4.12	29.34	6.00E-026
	ENSGALG00000011309	*Plxnc1*	776.16	91.66	8.47	9.35E-011
	ENSGALG00000011592	*Mras*	242.38	43.25	5.60	1.34E-019
	ENSGALG00000012956	*Tmtc1*	544.27	80.36	6.77	3.87E-022
	ENSGALG00000012964	*Ankh*	204.67	61.11	3.35	7.22E-011
	ENSGALG00000012998	*Wnt5b*	89.47	8.37	10.69	1.75E-018
	ENSGALG00000013071	*Cald1*	830.48	146.17	5.68	2.50E-012
	ENSGALG00000013124	*Fhod3*	482.85	67.98	7.10	5.40E-018
	ENSGALG00000013454	*Hoxb1*	216.64	4.79	45.25	1.12E-035
	ENSGALG00000014189	*Sult4a1*	1188.21	292.14	4.07	1.60E-013
	ENSGALG00000014250	*Phox2b*	809.86	18.35	44.13	5.60E-104
	ENSGALG00000014875	*Hcn1*	410.59	20.49	20.04	3.52E-037
	ENSGALG00000015476	*Clvs1*	199.57	18.76	10.64	3.87E-022
	ENSGALG00000015729	*Lpar1*	185.25	19.00	9.75	2.36E-013
	ENSGALG00000016358	*Ptchd1*	727.47	128.96	5.64	1.26E-005
	ENSGALG00000016698	*Shox*	829.62	23.83	34.82	2.33E-084
	ENSGALG00000017059	*Mab21l1*	509.74	55.45	9.19	6.30E-014
	ENSGALG00000021589	*Rtn4rl1*	218.36	22.49	9.71	1.40E-016
	ENSGALG00000023113	*Prox2*	2271.39	26.34	86.22	3.25E-034
	ENSGALG00000023847	-	174.96	18.34	9.54	2.25E-013
	ENSGALG00000023879	-	70.82	12.04	5.88	7.26E-004
	ENSGALG00000025774	-	395.25	11.60	34.07	1.01E-031
	ENSGALG00000026059	-	76.48	1.83	41.76	1.72E-014
	ENSGALG00000026557	*Pgf*	213.42	17.15	12.44	1.10E-013
	ENSGALG00000026763	-	275.35	24.08	11.43	7.05E-010
	ENSGALG00000026821	*Runx3*	541.77	26.15	20.72	2.31E-014
	ENSGALG00000027980	*Nanos1*	270.65	54.39	4.98	3.97E-014
	ENSGALG00000028236	*Prrx2*	759.97	37.46	20.29	5.79E-016
	ENSGALG00000028301	*Trim47*	134.71	25.31	5.32	1.22E-017
	ENSGALG00000028402	-	382.89	14.75	25.96	1.12E-035
	ENSGALG00000028506	*Nhsl2*	262.89	72.79	3.61	2.17E-014
	ENSGALG00000029175	*Efnb2*	748.11	106.76	7.01	1.55E-015
N	ENSGALG00000000284	*Hoxb4*	461.89	29.66	15.57	1.02E-018
	ENSGALG00000007819	*Pdzrn3*	832.82	83.37	9.99	4.23E-007
	ENSGALG00000015309	*Scml4*	837.81	70.58	11.87	5.28E-005
	ENSGALG00000023419	*Hoxd4*	286.66	14.45	19.84	2.87E-035
	ENSGALG00000023420	-	813.16	54.98	14.79	5.28E-043
	ENSGALG00000027903	*Hoxb5*	1497.79	61.17	24.48	4.27E-017
	ENSGALG00000028726	*Hoxb6*	124.62	2.72	45.85	3.17E-014

Hierarchical clustering of the markers reflects the divisions observed in the PCA (Figure 3.12A), although it shows slightly greater distinction of VA, P and N, with bootstrap values >98% for these nodes. It also confirms that there is very little separation between Top and Tmm, the two lobes of the trigeminal placode (and so not an unexpected result), with bootstrap values of 63% and 42%.

Gene ontology functional annotation analysis was also implemented for all of these markers as a group, to demonstrate the underlying functionality of the genes differentiating the placodes (Figure 3.12B). The most striking of these results is significant enrichment of genes associated with transcription factor activity, and sequence-specific DNA binding, both indicative of the underlying gene networks expected to give rise to these differences. There is also significant enrichment of extracellular compartment terms, which may be a reflection of signaling processes.

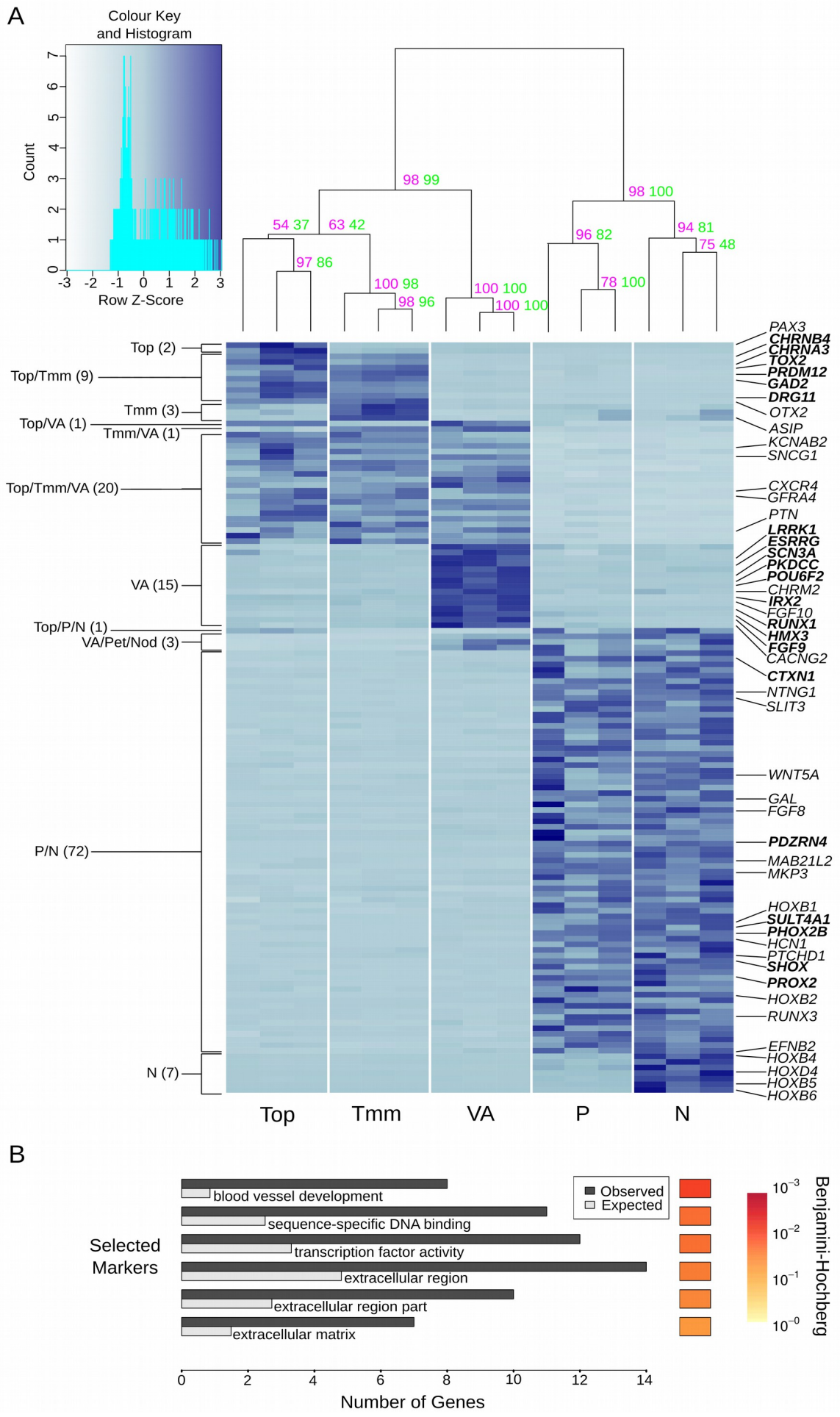

A
Colour Key and Histogram
Count
Row Z-Score
98 99
54 37
63 42
97 86
100 98
98 96
100 100
100 100
98 100
96 82
94 81
75 48
78 100
Top (2)
Top/Tmm (9)
Tmm (3)
Top/VA (1)
Tmm/VA (1)
Top/Tmm/VA (20)
VA (15)
Top/P/N (1)
VA/Pet/Nod (3)
P/N (72)
N (7)
PAX3
CHRNB4
CHRNA3
TOX2
PRDM12
GAD2
DRG11
OTX2
ASIP
KCNAB2
SNCG1
CXCR4
GFRA4
PTN
LRRK1
ESRRG
SCN3A
PKDCC
POU6F2
CHRM2
IRX2
FGF10
RUNX1
HMX3
FGF9
CACNG2
CTXN1
NTNG1
SLIT3
WNT5A
GAL
FGF8
PDZRN4
MAB21L2
MKP3
HOXB1
SULT4A1
PHOX2B
HCN1
PTCHD1
SHOX
PROX2
HOXB2
RUNX3
EFNB2
HOXB4
HOXD4
HOXB5
HOXB6
Top
Tmm
VA
P
N
B
Selected Markers
blood vessel development
sequence-specific DNA binding
transcription factor activity
extracellular region
extracellular region part
extracellular matrix
Observed
Expected
Benjamini-Hochberg
Number of Genes

Figure 3.12. Expression patterns and functional annotation of the 134 marker genes. (A) Variance-stabilized gene expression of all markers (as displayed in Table 3.3). The numbers of the markers for each ganglion or group of ganglia are displayed on the left, with genes of interest on the right, including those selected for validation by *in-situ* hybridization (in bold). The dendrogram displays the hierarchical clustering of these markers (with euclidean distance measure and complete/maximum linkage method), with the approximately unbiased probability and bootstrap probability values (10 million iterations) for each node displayed in pink and green respectively. (B) Significantly enriched GO terms amongst the 134 markers, with redundant GO terms removed using REVIGO and a Hinton plot displaying the FDR (Benjamini and Hochberg) corrected p-value.

3.4 Discussion

The cranial sensory ganglia are important in performing distinct sensory functions. However, the underlying mechanisms for their development and their specific sensory modalities are not fully understood, and have been restricted by a lack of markers. This study pursued the identification of the underlying molecular signatures for these ganglia and their modalities, reporting the genome-wide expression profiles of these through RNA-Sequencing in the developing chick embryo. PCA revealed distinct clusters and illustrated the segregation of the ganglia by modality, into the epibranchial/viscerosensory, trigeminal/general somatosensory, and otic/special somatosensory. We report a library of differentially expressed genes underlying their differences, in addition to 134 marker genes which, following stringent selection criteria, provide important insight and information on the mechanisms underlying the development of the cranial sensory ganglia.

The high confidence markers reported in this study provide an important tool for future investigation into these ganglia, and the confidence of these is therefore of utmost importance. Many of these markers have now been confirmed by collaborators, Cedric Patthey and Jo Begbie, through in situ hybridization, the results of which (and the selection rationale for which markers to validate) can be found in Figures 14-16.

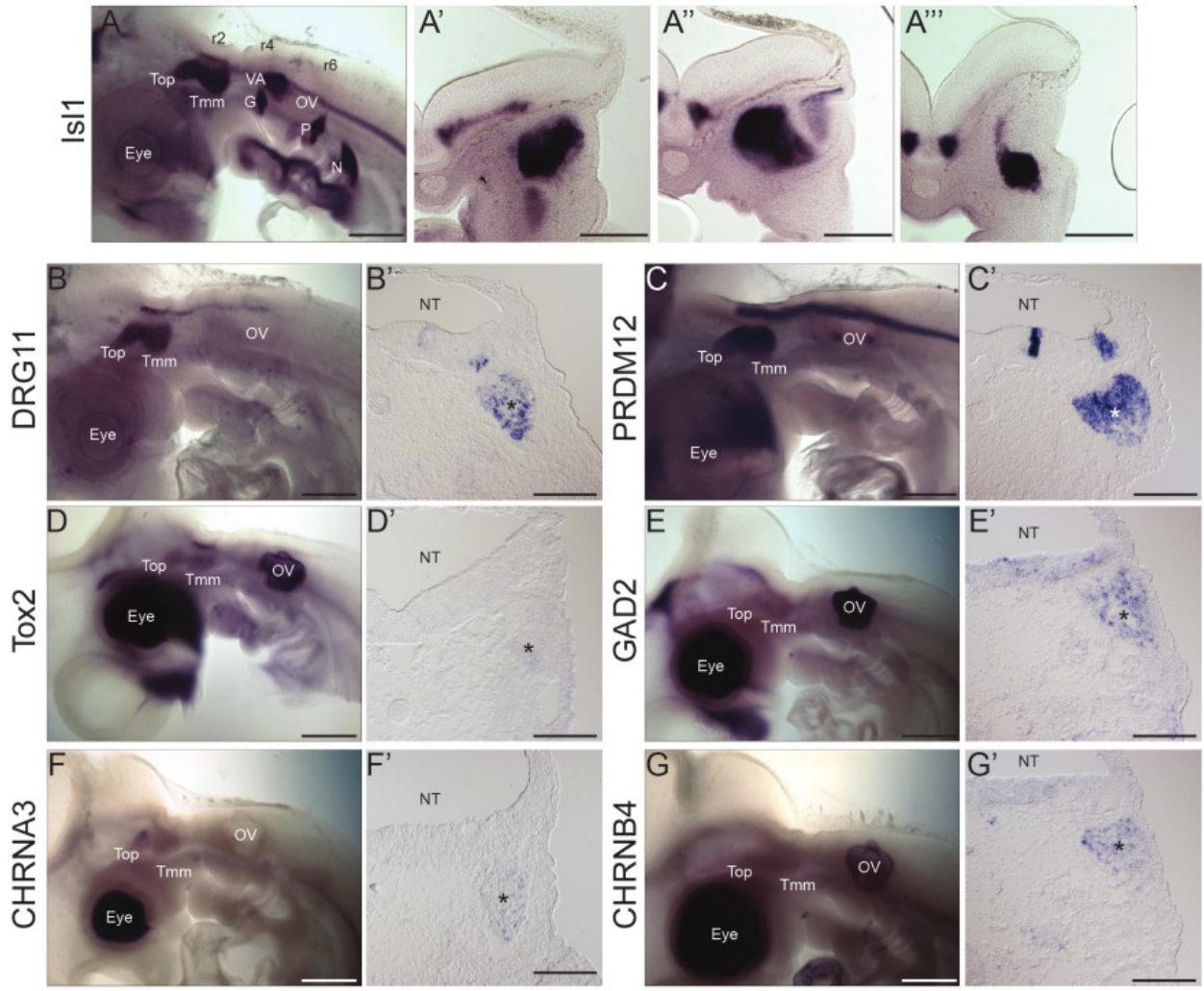


Figure 3.14. *In-situ* validation of expression patterns of trigeminal ganglion-specific markers (performed by collaborators Jo Begbie and Cedric Patthey, with figure produced by Jo Begbie). (A) Wholemount in situ hybridization of *ISL1* at *HH21* to provide anatomical localisation of all cranial sensory ganglia. (A'-A''') Transverse sections stained with *ISL1* at *HH21* to provide comparative sections for other markers at the level of Tmm ganglion (A'); VA ganglion (A''); and P ganglion (A'''). (B-G) Wholemount in situ hybridization of trigeminal ganglion specific markers at HH21. (B'-G') Transverse sections through Tmm at HH21 showing ganglion specific expression of the indicated gene.

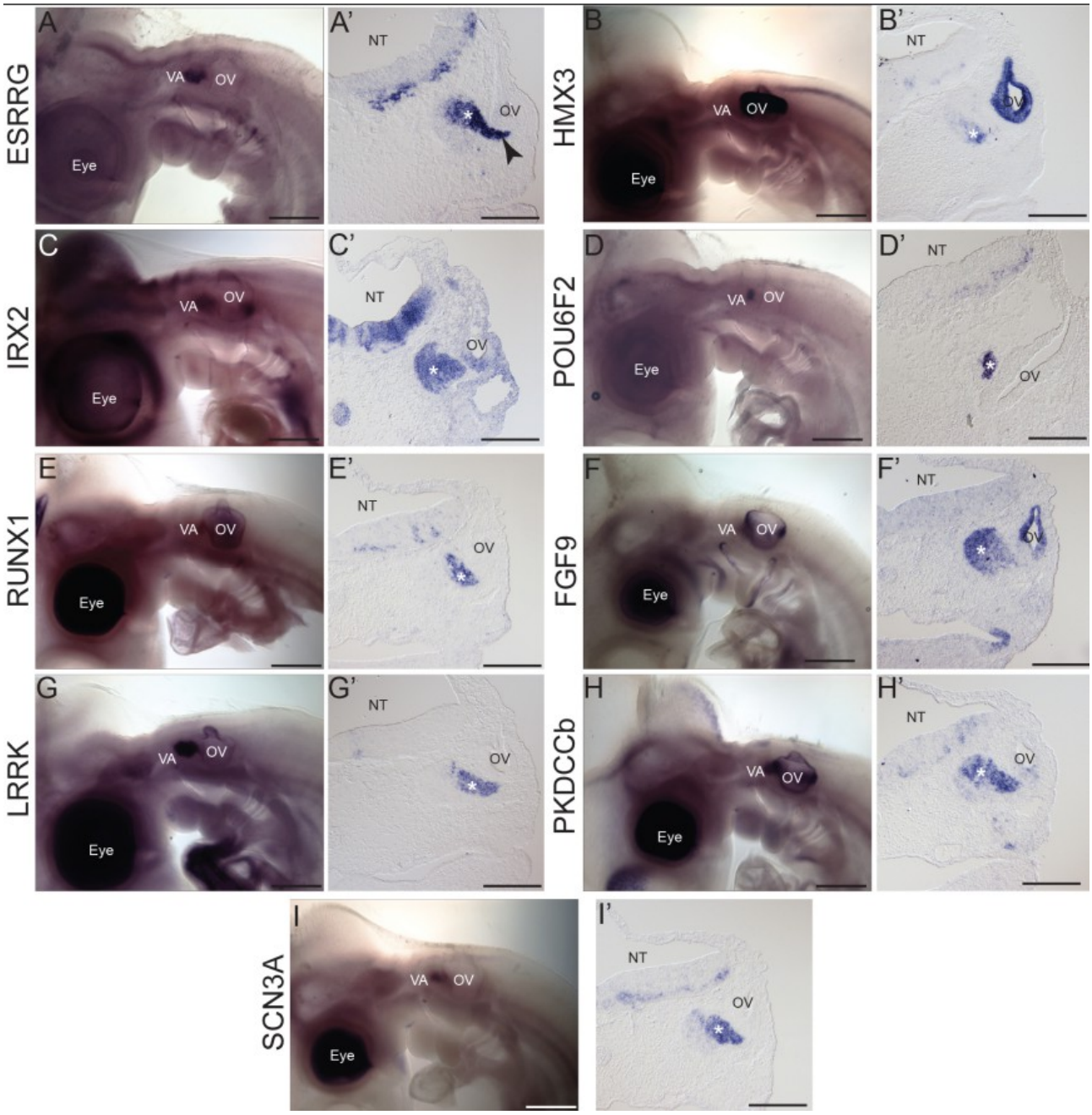


Figure 3.15. *In-situ* validation of expression patterns of vestibulo-acoustic ganglion-specific markers (performed by collaborators Jo Begbie and Cedric Patthey, with figure produced by Jo Begbie). (A-H) Wholemount in situ hybridization of vestibulo-acoustic ganglion-specific markers at HH21. (A'-H') Transverse sections through VA at HH21 showing ganglion-specific expression of the indicated gene.

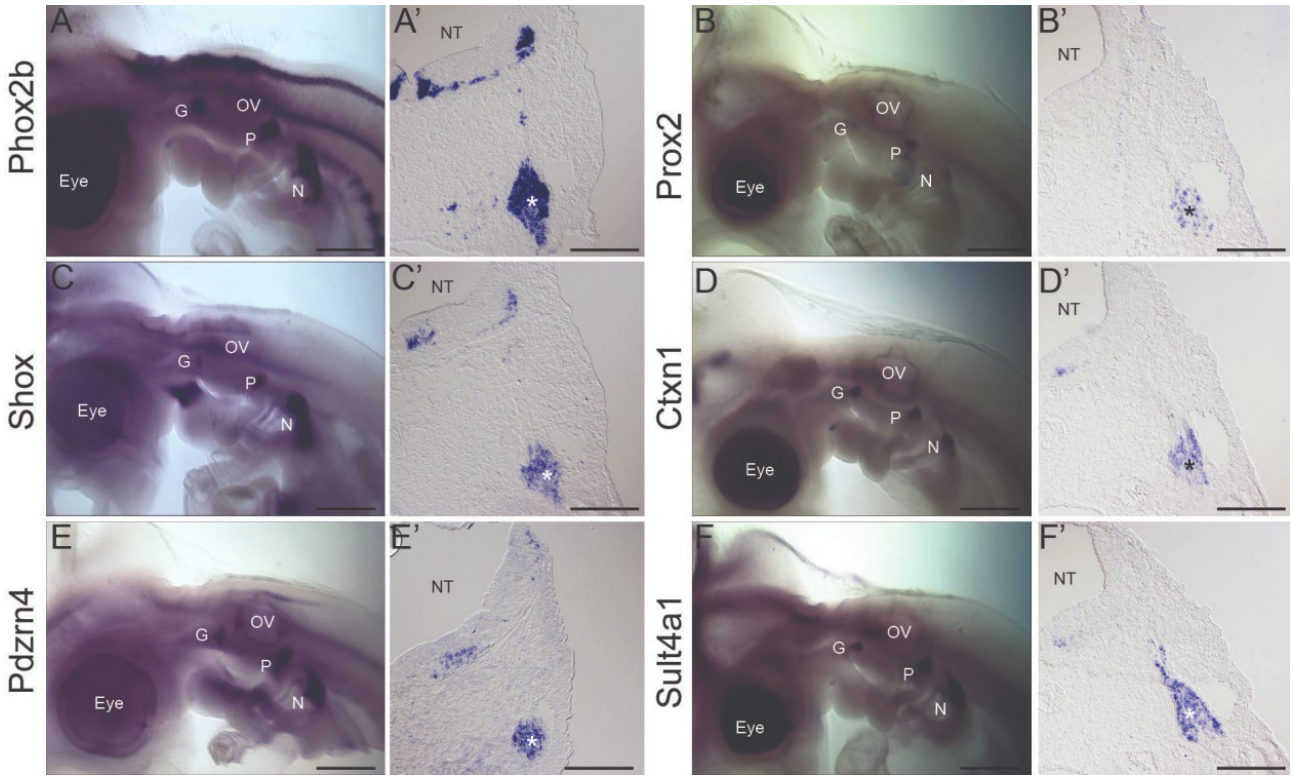


Figure 3.16. *In-situ* validation of expression patterns of epibranchial ganglion-specific markers (performed by collaborators Jo Begbie and Cedric Patthey, with figure produced by Jo Begbie). (A-F) Wholemount in situ hybridization of epibranchial ganglion specific markers at HH21. (A'-F') Transverse sections through P at HH21 showing ganglion-specific expression of the indicated gene.

The validity of the markers is further bolstered by the presence of the few genes already widely known to be expressed specifically in the specific cranial sensory ganglia: for example, *PAX3* as a marker for Top (Dude *et al.* 2010), *DRG11* as a marker for Top/Tmm (Rebelo *et al.* 2007), and *PHOX2B* as a marker for P/N (Brunet and Pattyn, 2002). In addition to this, those markers differentiating N constitute a large number of *HOX* genes, including B4-6 and D4, with the most anterior *HOX* genes, B1-2, included in the N/P markers. This reflects the well-known distribution of *HOX* genes along the rostro-caudal axis (Nodel, 1993). On the other hand, many of the markers reported in this study were not previously known or expected, and are entirely novel. To our knowledge, this is the first report of a molecular distinction between "general" and "special" somatosensory neurons.

The novel high confidence markers displayed in Figure 3.14 that are specific to the trigeminal ganglion (when compared to the other cranial sensory ganglia studied), include *PRDM12, TOX2, GAD2, CHRNA3*, and *CHRNB4.* The PRDM family consists of a series of zinc finger transcription factors which are broadly implicated in transcriptional regulation through chromatin modification, either through methyltransferase activity or recruitment of chromatin remodelling complexes (Di Zazzo *et al.* 2013). More specifically, *PRDM12* is associated with a well characterized pathology in humans, with *PRDM12* mutations underlying an autosomal recessive disorder known as hereditary sensory and autonomic neuropathy (HSAN) type VIII (Chen et al. 2015). HSAN8 presents as

congenital insensitivity to pain, with disrupted pain and temperature sensation. This in turn results in a plethora of symptoms observed in babies, including lip/tongue lesions, facial scratching and finger biting. TOX2 is also a transcription factor and, although less is known about its role specifically, the TOX1-4 subfamily are a group of proteins which share an HMG-box DNA binding domain, and that have known roles in chromatin modification for transcription, replication, and DNA repair (Stros *et al.* 2007). *GAD2* is one of two genes which encode different forms of the Glutamic Acid Decarboxylase (GAD) enzyme. The GAD enzyme is responsible for catalysing the decarboxylation of glutamate to GABA, which is then used as a neurotransmitter. GAD65 specifically (the *GAD2* form), also has associations with neuropathic pain, identified due to the reduction of GAD65 terminals in sciatic nerve trauma in rats (Lorenzo *et al.* 2014). *GAD2* expression has also been described specifically in the adult trigeminal ganglion in rats (Hayasaki *et al.* 2006). Finally, *CHRNA3* and *CHRNB4*, both members of the Cholinergic Receptor Nicotinic subunit (CHRN) family, encode subunits of cholinergic receptors that respond to the neurotransmitter acetylcholine. Neuronal nicotinic acetylcholine receptors are well associated with pain perception and specifically associated with the adult trigeminal ganglion in rats (Liu *et al.* 1998).

The high confidence markers specific to the vestibulo-acoustic ganglion displayed in Figure 3.15 include *IRX2*, *POU6F2*, *FGF9*, *LRRK1*, *PKDCCB* and *SCN3A*. *ESRRG*, *HMX3* and *RUNX1* are also displayed as controls, which (although no molecular distinction between special and general

somatosensory has been described previously) are known markers specifically for inner ear development and for somatosensory systems (Hermans-Borgmeyer *et al.* 2000; Feng & Xu, 2010; D'Autréaux *et al.* 2011). *IRX2* and *POU6F2* are both homeobox genes, and have previously been implicated in developmental patterning, but are yet to be reported as being expressed in the vestibulo-acoustic ganglion. *FGF9* has also been implicated previously in development, with the Fibroblast Growth Factor (FGF) gene family well-known in embryonic development, cell growth, morphogenesis and signalling pathways, of which *FGF9* has been previously implicated in later inner ear hair cell development (Pirvola *et al.* 2004). *LRRK1* and *PKDCCB* both encode kinases, a large family of proteins that govern many cellular processes, such as growth, differentiation and proliferation (Lahiry *et al.* 2010). Due to these roles, kinases have been targeted as part of many neuronal disorder therapies (Chico *et al.* 2009), and although *LRRK1* and *PKDCCB* have not previously been associated with the vestibulo-acoustic ganglion specifically, variants in *LRRK1* have been associated with neurological disorders such as Parkinson's disease (Schulte *et al.* 2014). Finally, *SCN3A* encodes a voltage-gated sodium channel, and is responsible for the generation and propagation of action potentials in neurons. Interestingly, variants in *SCN2A* and *SCN3A* are strongly associated with autosomal dominant non-syndromic hearing loss known as DFNA16, which is characterized by progressive hearing impairment (Kasai *et al.* 2001).

Finally, the high confidence markers displayed in Figure 3.16, specific to

the epibranchial placodes, include *PROX2*, *SHOX*, *CTXN1*, *PDZRN4* and *SULT4A1*. The first of these, *PROX2* and *SHOX,* are both homeobox transcription factors, and so (as before) are known in developmental patterning in various structures. The remaining markers we report are not well studied. *CTXN1* is named cortexin to reflect its enriched expression in the rodent cortex. The small CTXN1 protein contains a putative single membrane-spanning domain in the centre, which means it likely functions as an integral membrane protein that mediates extra- or intra-cellular signalling of cortical neurons during forebrain development (Watson *et al.* 1994). *PDZRN4* is also largely uncharacterized, having been identified *in silico* and only known for its protein being homologous to *PDZRN1,* which is implicated in cell fate determination through a Notch signalling pathway (Katoh, 2004). Finally, *SULT4A1* is a member of the sulfotransferase family, which encode a series of cytosolic enzymes with proposed roles in modulation of neurotransmitters, with SULT4A1 being specifically localized to the cerebral cortex, cerebellum, pituitary, and brainstem, in humans and rats (Liyou *et al.* 2003).

Regarding the use of these newly identified markers, and immediate possibilities for further studies of the cranial sensory systems, these markers provide a resource for tracking adoption of cell fate and further investigating the phenotype of the systems. Although we mostly report transcriptional regulators and components of signaling pathways, which do not necessarily inform us immediately about modality, these are key to understanding the process of acquiring modalities and distinct

phenotypes. For example, gene knockdowns of these markers could be extremely revealing, and would be likely to uncover phenotypic aspects of the different cell type identities. This set of novel markers can also be used to unequivocally determine the fate adopted by cells, for example (as mentioned previously) when deriving sensory neurons from stem cells. At present, the molecular markers used to recognize somatosensory, as opposed to viscerosensory, neurons are very limited and typically consist of: *BRN3A+, DRG11+, RUNX1+, PHOX2A-, PHOX2B-* and *RET-*. Using just these markers, one study found *PHOX2B* to be the main regulatory switch between somatosensory and viscerosensory ganglia (D'Autréaux *et al.* 2011). The markers we provide could be used in experiments such as this to more robustly confirm this switch in phenotype, as well as to reveal novel information on whether the resultant cells are more similar to trigeminal or otic cells.

Overall, this study provides an early, but essential, insight into the molecular signatures underlying the placodes of the cranial sensory systems. These markers can form the basis of future studies, important in investigations into the development of these cranial nerves, as well as in deriving these specific neurons in a clinical setting.

Chapter 4 - Expression of Autism -Associated Genes in Developing Purkinje Cells Predicts Intellectual Disability Co-Morbidity Status

Abstract

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disease, for which the underpinning molecular mechanisms remain incompletely understood. Through time-series analysis of the transcriptome dynamics underlying developing Purkinje Cells (PCs) in the mouse, this study reports a single co-expressed cluster of genes whose expression is positively correlated with development and is enriched for ASD-associated genes. This gene cluster was found to be specific to developing PCs rather than to the developing mouse neocortex, in which a second temporally regulated ASD gene module was identified. The composition of ASD risk genes within these two distinct clusters is significantly different in their association with intellectual disability (ID), consistent with the existence of two genetically and spatiotemporally distinct ASD endophenotypes. This study defines a specific cluster of ASD genes, enriched in developing PCs, which predict ID co-morbidity status.

Publication and Contributions

This chapter is based on the work described in my lead-author paper submitted to *Biological Psychiatry*. Anna Dulneva and Esther B. E. Becker led procedures prior to RNA-Seq, with Anna Dulneva performing dissections, laser capture microdissection, and RNA preparations. Sequencing was completed by the Wellcome Trust Centre for Human Genetics. All subsequent computational genomics analysis and interpretation was performed by me, with Wilfried Haerty and Chris P. Ponting advising. I'd also like to thank Andreas Heger for technical support and T. Grant Belgard for critical comments on the publication manuscript.

4.1 Introduction

The Purkinje Cells of the Cerebellum

The cerebellum is a key structure of the brain known for its major role in motor co-ordination and learning. Known as the "little brain", the human cerebellum contains over 50% of the total neurons of the brain, despite only accounting for approximately 10% of the total brain volume (Johnson & Byrne, 1998). The majority of the cerebellum's outputs are involved in the motor system and consequently damage to the cerebellum is associated with deficiencies in motor control and posture. Specifically, its key roles are in refining the commands of the motor system to ensure high levels of accuracy and adaptation. These roles include motor learning, maintenance of balance (and posture), voluntary movement co-ordination, and cognitive function. The motor learning of the cerebellum includes a chief role in fine-tuning accurate movements through trial-and-error, important in mastering motor processes (Thach, 1998). The maintenance of balance is executed through interaction with outputs from the vestibular receptors of the inner ear, from which the cerebellum transmits commands to motor neurons for changes in posture in order to compensate for shifts in body position, thus maintaining balance (Denoth *et al.* 1979). The voluntary movement co-ordination of the cerebellum is related to ensuring co-ordination of multiple muscle groups (in terms of power and timing) to ensure fluid body movements (Johnson & Byrne, 1998). Finally, the cognitive processes executed include the lesser known

and less understood functions of the cerebellum. Despite the cerebellum's historic association with motor control, an increasing implication in higher cognitive functions has included aspects such as attention, language and executive functions (Koziol *et al.* 2014). For example, patients with cerebellar lesions have been shown to have reduced performance in specific attention tasks (Gottwald *et al.* 2003). The implications of the cerebellum's role in language and executive function are largely due to associations with working memory, including verbal working memory. Verbal working memory is the ability to hold and process verbal information, for which the pre-frontal cortex has been most commonly implicated. Nevertheless, there is now increasing evidence of involvement in these processes by the cerebellum, including specific roles in error-driven adjustment, and internal timing (Ben-Yehudah *et al.* 2007). Due to increasing information on the plethora of functions in which the cerebellum is involved, cerebellar dysfunction is becoming further associated with an increasing number of human disorders, including both motor diseases (ataxia, dystonia) (Klockgether, 2011) and non-motor diseases (autism spectrum disorder, dyslexia, fetal alcohol syndrome, etc.) (Reeber *et al.* 2013; Becker & Stoodley, 2013).

The cerebellum hosts many cell types, including non-neuronal cells, such as glia, and a range of neuronal cells, such as cerebellar granule cells, Golgi cells, Basket/Stellate cells, and Purkinje Cells (PCs) (Purves *et al.* 2001). Glia are observed throughout the nervous system, maintaining homeostasis (for example, by supplying oxygen), providing protection (for

example, by providing insulation, as well as destroying pathogens), in addition to actually physically supporting the neurons in both the central and peripheral nervous systems (Greener, 2015). Granule cells of the cerebellum are small, densely packed, and the most abundant class of neurons in the human brain, constituting over half of all neurons of the CNS (Williams and Herrup, 1988). Mossy fibers (a form of axon and the major input to the cerebellum, transmitting from a variety of sources) are the sole provider of signals to the cerebellar granule cells. The abundance of these granule cells allows the information from the mossy fibers to be recoded in an expansive way, before relaying along parallel fibers to the PCs for processing. The Golgi cells also receive input from mossy fibres, and connect to both granule cells and the parallel fibers of the granule cells, allowing for an inhibitory feedback network (Purves *et al.* 2001). Basket and Stellate cells are interneurons that also interact with PCs, inhibiting their response, and are differentiated from one another based on their morphology and location (rather than role) in being located in the inner and outer regions of the cerebellar molecular layer respectively. The PCs themselves are the most histologically striking cell type and among the most distinctive neurons in the brain, with a profusely branched but severely flattened dendritic tree (Purves *et al.* 2001) (Figure 4.1).

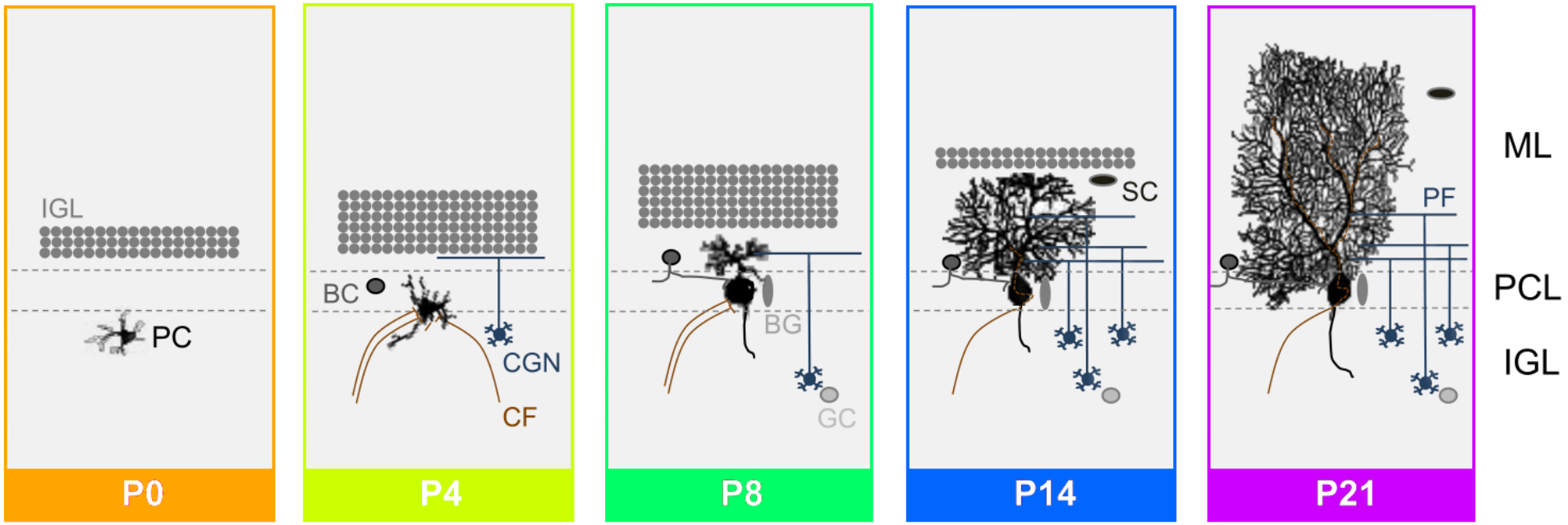


Figure 4.1. A schematic diagram of the development of a PC in the mouse, from P0 to P21. Extensive morphological changes can be seen throughout the development, including the formation of synaptic connections with climbing fibers (CF), Basket cells (BC) and parallel fibers (PF) of cerebellar granule neurons (CGN). BG: Bergmann glia; EGL: external granular layer; GC: Golgi cells; IGL: internal granular layer; ML: molecular layer; PCL: Purkinje cell layer; SC: Stellate cells.

The PCs provide the sole output of the cerebellum, and therefore play a vital role in the function of the cerebellum. Additionally, they act as organizers of the cerebellar cortex, including neurogenesis, and arrangement of the cortical layering (Anderson *et al.* 1996). This makes PCs key to the cerebellum's development and ongoing processes, and thus important to study.

Autism Spectrum Disorder and the Cerebellum

Autism spectrum disorder (ASD) is clinically and etiologically diverse, constituting a heterogeneous group of neurodevelopmental disorders. It is generally defined by a triad of symptoms including deficits in social cognition, deficits in communication, and restricted interests / repetitive behaviors. Social deficits include difficulty empathizing, and difficulty in understanding the interactions of neurotypicals; communication deficits include delays in learning to speak, and a lack of communicative behavior such as eye contact; restricted and repetitive behaviors include difficulty in coping with changes (for example, moving furniture), difficulty engaging in imaginative play (and so often play in solitary), and repetitively stacking or lining up objects (Rapin & Tuchman, 2008). There are also a large range of co-morbid conditions associated with ASD, just a few examples of which include Attention Deficit Hyperactivity Disorder (ADHD), Obsessive Compulsive Disorder (OCD), and anxiety disorders, as well as reports of a low intelligence quotient (IQ, measured nonverbally), gastrointestinal symptoms, sleep problems, and motor deficits (Maski *et*

al. 2011; Mannion *et al.* 2013}. One study estimating co-morbidity in 89 children and adolescents with ASD, reported 46% prevalence of a co-morbid disorder, rising to 79% when including intellectual disability. These included 18% with ADHD, 10% with epilepsy, and 16% with an anxiety disorder. They additionally report 79% of patients with at least one gastrointestinal problem and 81% with a sleep problem (Mannion *et al.* 2013). Deficits in motor skills are noted in high frequency amongst ASD patients, with studies reporting as high as 79% (Green *et al.* 2009) or 100% (Pan *et al.* 2009) of patients with a motor deficit (depending on the method of assessment).

Although the contributions of genetic and environmental factors still remain largely undefined, its high heritability is undeniable. Twin studies suggest heritability of 0.7-0.9 (Geschwind, 2009), and siblings of those with autism are approximately 25 times more likely to be autistic (Jorde *et al.* 1991). Genome-wide studies have also exposed hundreds of ASD-associated contributory genetic variants (Willsey & State, 2015; Devlin & Scherer, 2012; Berg & Geschwind, 2012). There are also Mendelian disorders which include ASD as a symptom, for example in Fragile X syndrome, a trinucleotide repeat expansion disorder resulting in failure to express the FMR-1 protein required for normal neural development, and subsequently symptoms of intellectual disability (among others) (Verkerk *et al.* 1991). Fragile X is the most common known single gene cause of ASD, co-occurring in approximately 5% of all ASD cases (Budimirovic & Kaufmann, 2011), with around 50% of all Fragile X patients also exhibiting

ASD (Abbeduto *et al.* 2014). The large genetic heterogeneity of ASD provides a challenge to those attempting to understand the underlying causes. Attempts to tackle this have included the curation of databases containing information on the many genes studied and their respective associations with ASD. One database currently at the forefront of this is the Simons Foundation Autism Research Initiative (SFARI) Gene database (Basu *et al.* 2009), which contains records of hundreds of genes and their ASD-associations, with critical evaluation of each (including gene scores based on published literature), and information on hundreds of mouse model lines for ASD.

Overall, the genetic heterogeneity of ASD alongside the broad clinical phenotype it presents provides a major challenge for understanding the underlying pathophysiology. In investigations into the regions of the brain affected in autism, multiple regions have been implicated. The neocortex is commonly implicated (Stoner *et al.* 2014), including in functional MRI (Redcay & Courchesne, 2008), gene expression (Chow *et al.* 2012), and neuron-count studies (Courchesne *et al.* 2011). However, the cerebellum has also been highlighted as a key region (Becker, 2014; Wang *et al.* 2014). A key sign of its implicated role is the finding that developmental injury to the human cerebellum is the greatest single non-heritable risk for ASD (Wang *et al.* 2014). In addition to this, statistically significant reduction in grey matter in the cerebellum of patients with ASD has been identified through imaging meta-analysis (Stoodley, 2014), the extent of which can be used to successfully predict autism severity (D'Mello *et al.*

2015). It has also been noted that PCs are specifically reduced in number and density in ASD (Palmen, 2004). The role of PCs is corroborated further by PC-specific conditional mouse models which lack the autism-associated genes *Tsc1* and *Tsc2* (Tsai *et al.* 2012; Reith *et al.* 2013). Overall, there is strong evidence for PC dysfunction in development contributing to ASD, PC dysfunction at the genetic and molecular level, however, remains poorly understood.

Investigating Purkinje Cells and ASD

This study builds on what is already known about the associations of PCs with ASD, and aims to confirm these findings by identifying and profiling clusters of genes over PC development. This includes further investigation into complex aspects of ASD, such as Intellectual Disability (ID) co-morbidity, and comparative analysis with the neocortex. This was undertaken with the aim of both gaining a greater understanding of gene expression patterns through PC development, and further investigating the potential associations with autism at a genomic level. To perform this analysis, data was acquired by collaborators Esther Becker and Anna Dulneva, who performed laser capture microdissection of PCs at five key developmental time-points: P0, P4, P8, P14, and P21 (Figure 4.1). This was implemented for 1000 PCs collected randomly across the cerebellum from 2-3 tissue sections per mouse (C57BL6, male). For P0, more tissue sections were required, with PCs isolated from 6 tissue sections. RNA was then extracted from these samples and sequenced in triplicate with the

Illumina Hi-Seq 2000 for 100bp paired-end reads. This data was then provided for all computational analyses in this study.

4.2 Methods

Read Processing and Alignment

Initial sequencing and read quality was assessed using FastQC (Andrews, 2010). SMARTer PCR cDNA Synthesis Kit had been used to increase the quantity of materials for sequencing, and this was evident with SMARTer adapter contamination in the raw sequences (see Results). Trimming of these adapter sequences was implemented with Cutadapt (Martin, 2011), allowing a maximum trimming of any given read down to 30bp before disposal of the read. This resulted in substantial disposal of reads (particularly of the 3' reads of each pair) (see Results). To circumvent this issue, the 5' and 3' reads were trimmed with differing variables, with 5' reads permitted a maximum trimming down to 50bp before disposal, and 3' reads permitted a maximum trimming down to 17bp (the minimum allowed by Gsnap – see below) before disposal. This allowed the retention of 5' reads with a higher confidence in alignment, and therefore permitted a greater allowance for 3' read trimming to a shorter length, because the alignment of the 3' read can then be guided (with expectation of a nearby mapped locus) by the 5' pair, using expected insert size settings. Alignment was implemented with the splice-aware short read aligner Gsnap (Wu & Watanabe, 2005), to the mouse reference genome GRCm38/mm10. To provide extra confidence in the alignment, only those reads that were uniquely aligned to the genome were retained, and then count tables were produced with HTSeq (on intersection strict settings)

(Anders *et al.* 2015) to count reads mapped to Ensembl gene models (Ensembl 72 GRCm38.p1).

Weighted Gene Co-expression Network Analysis (WGCNA)

Weighted Gene Co-expression Network Analysis (WGCNA) is an R package that facilitates the formation of clusters (or modules) from gene expression data based on co-expression (Langfelder & Horvarth, 2008). The basic theory behind the analysis is firstly to measure concordance of gene expressions with Pearson correlation, and secondly to transform the resultant Pearson correlation matrix with a power adjacency function to form a weighted network. In the WGCNA package the power adjacency function used is known as the scale free topology criterion. Scale free topology ensures the presence of "hub nodes" that are connected to a large number of other nodes (as observed in biological scenarios, such as metabolic networks; Ravasz *et al.* 2002), and refers to a specific frequency distribution of the connectivity (a skewed distribution, allowing for many genes to have few connections, and few genes to have many; a log transformation of this distribution would be roughly linear). The parameters of this model are then selected so as to result in the highest mean number of connections. WGCNA then forms a Topological Overlap Matrix (TOM) from a generalized topological overlap dissimilarity measure: this determines distances between pairs of nodes in a network through information on the number of shared neighbors (measured at varying orders, e.g. an n-th step neighbour). WGCNA then uses this data to

determine the clusters, through use of average linkage hierarchical clustering with a dynamic tree cutting algorithm.

Here, WGCNA is used to identify modules of co-expressed genes underlying the PC development. Genes with zero expression across all samples or zero variance were removed from the analysis. For subsequent analyses, the modules were reduced to only genes with a module membership >0.8, therefore only considering the core of the module.

Fuzzy Clustering

Fuzzy clustering is a form of K-means clustering algorithm, which achieves its namesake by allowing any given object (in this case a gene) to be assigned to multiple clusters, with varying levels of membership applied in R using the Mfuzz package (Kumar & Futschik, 2007). K-means clustering itself is an iterative method of clustering, which works by positioning centroids in the data (randomly in the first instance), and then subsequently assigning all of the objects of the data to their nearest centroid. This is then repeated, re-positioning the centroids to the centre of their assigned objects, and then re-assigning the objects to their nearest centroid, until the centroids no longer move, or until a set maximum number of iterations is reached. With fuzzy clustering, assignment is continuous and based on distance. In this study, fuzzy clustering is used as a secondary method of analysis. It is recommended that WGCNA is run with as many replicates as possible, and so to ensure

confidence in any positive results it was thought prudent to also utilize an additional independent approach to ensure robust findings.

The number of centroids to use in the data, and subsequently the number of clusters formed (K), can be decided in a variety of ways (for which there is no gold standard). The most common methods applied to K-means algorithms are internal evaluation methods, evaluating the clusters resulting from that value of K, to determine if these are more suitable than any other values of K. In this study, three well-established methods were used to ensure the most suitable value of K was chosen: the silhouette-width method (Rousseeuw, 1987; Clifford *et al.* 2011), the Calinski-Harabasz method (Calinski & Harabasz, 1974), and the elbow method (Thorndike, 1953). The silhouette-width method determines the average dissimilarity of each datum with all other data of the same cluster, calculating its "silhouette" based on how well matched it is to its current cluster compared to its neighbouring cluster. The average silhouette of each cluster can then be taken, and the average of these averages can provide a metric for how well the clusters are formed in general. The Calinski-Harabasz method works in the same vein, determining an index from the clustering of the data, from which the K with the highest Calinski-Harabasz index is deemed the most appropriate. This index is based on the between-cluster sums of squares, known as the "separation" of the clusters, divided by the within-cluster sums of squares, known as the "cohesion" of the clusters. It therefore looks to maximize the distance between clusters, while minimizing the distance within. Finally, the elbow

method is implemented based on the percentage of variance explained by the clusters. The variance explained is plotted against each value of K, and the optimum value is taken from the "elbow point" at which K is minimized and the variance explained is maximized.

Differential Expression Analysis

Following WGCNA and fuzzy clustering analysis, methods for acquiring statistically significant clusters positively or negatively correlated with developmental time were utilized. This was implemented through the time-series differential expression functionality of DESeq2 (Love *et al.* 2014). DESeq2 utilizes much of the same methodology of DESeq, but also hosts tools for applying a likelihood ratio test, suitable for time-series analysis with the aim of comparing models with and without time factor. This testing was used with FDR<0.05 as a threshold for significant results. Results were then subsequently split by fold change into positive and negative groups, resulting in the formation of two respective clusters positively or negatively correlated with developmental timing.

GO, KEGG, and MGI Enrichment

To successfully profile the developing PCs, various enrichment tests were performed at various stages of the analysis, including for associated GO terms, KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa & Goto, 2000), and MGI phenotypes (Mouse Genome

Informatics) (Blake *et al.* 2014). All enrichment testing was performed with two-tailed Fisher's exact testing, which was provided by the GOSeq package correcting for gene-length bias in differential expression selection (Young *et al.* 2010). The background used for testing consisted of all genes with non-zero expression in any samples or non-zero variance across all sample. Gene Ontology was performed with GOSeq-provided GO terms for mm10 Ensembl gene IDs. Resultant significantly enriched GO terms were then filtered through REVIGO (REduce and VIsualize Gene Ontology) (Supek *et al.* 2011), removing redundant GO terms based on semantic similarity. This was used on a moderately strict setting, allowing a similarity value of < 0.5.

Disease Enrichment Analyses

Enrichment testing was implemented with multiple customized neurological disorder-associated gene lists. As with previous enrichment testing, two-tailed Fisher's exact testing was used through the GOSeq package correcting for gene-length bias. The background used for this testing once again consisted of all genes with non-zero expression in any sample or non-zero variance across all samples. In addition to this, the background was further reduced to protein coding genes only, as a conservative measure to match the constitution of the customized foreground gene lists. When performing testing against any human foreground gene lists, the background was reduced even further to only those genes with 1:1 orthologs between human and mice. Although this

additional reduction of the background makes statistically significant enrichments less likely for the human results than for the mouse, it is necessary to ensure fair and robust enrichment testing. The neurological disorders investigated include autism (our primary disorder of interest), ataxia (an expected enrichment due to associations with the cerebellum), and schizophrenia (no known association with the cerebellum). The autism-associated gene lists include both a human candidate list (292 genes) and mouse model list (150 genes), obtained from the SFARI database (accessed 2nd Dec 2014) (Basu *et al.* 2009). The ataxia-associated gene lists were curated by Esther Becker, with candidate lists obtained through literature searches (Table A1 [Appendix A]), and once again consisting of a human candidate list (57 genes) and a mouse model list (21 genes). The schizophrenia-associated genes were obtained from a human candidate gene list (293 genes) produced by the Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). In addition to these curated gene lists, a list of the genes of 842 transcripts that bind Fragile-X Mental Retardation Protein (FMRP-bound transcripts), as produced by Darnell *et al.* (2011), was also tested, providing an independent list which is also known to be associated with autism (Steinberg & Webber, 2013). Similarly to the background data, these lists have been reduced to protein coding genes, and (where appropriate) genes with 1:1 mouse-human orthologs only.

Comparative Analysis across Brain Regions

To investigate the specificity of these results in the PC data, and to put these findings into context, comparative analyses were performed on equivalent transcriptomic data in the developing mouse neocortex, produced by a study from Fertuzinhos *et al.* (2014). Data was taken for the fourth layer of the neocortex (the middle layer and the only isolated layer available), at developmental timings P4, P6, P8, P10 and P14, with the aim of maintaining as much consistency with the PC data as possible. Although batch effects would likely pose a large problem, it was ensured that none of our analyses involved direct comparisons of expression values across tissues. Identical data processing procedures (including QC, alignment, and read counting) were applied to this data as with the PC data. Subsequently, WGCNA was applied with cluster patterns and enrichments compared across the two tissues.

Permutation Testing

The application of clustering analyses across two brain regions allowed the investigation of whether genes within these specific modules might contribute to distinct ASD endophenotypes. As Intelligence Quotient (IQ) is proposed to define clinical subgroups of ASD (Waterhouse, 2013), investigation of differences in the Intellectual Disability (ID) subgroup (IQ < 70) across the ASD-enriched gene clusters was deemed an appropriate next step.

Permutation testing was performed to statistically quantify differing trends in odds ratios so as to retrieve an empirical p-value. Odds ratios were calculated by dividing each of the two odds observable in a contingency table, and were therefore useful in quantifying the presence of one property against the presence of another. Permutation testing was then used to compare PC and neocortex enrichments, with stratified autism human-candidate gene lists separated into those genes with ID association, and those with no known ID association. Odds ratios were obtained from the contingency tables for 4 enrichment tests: autism ID-associated genes enriched in the main PC cluster; autism non-ID-associated genes enriched in the main PC cluster; autism ID-associated genes enriched in the main neocortex cluster; and autism non-ID-associated genes enriched in the main neocortex cluster. Ratio of Odds Ratios (RORs) were then obtained across the two tests for each tissue type (PCs and neocortex), and these were compared directly. This same comparison was made with data permuted 10^8 times, in order to obtain the empirical p-value for the difference in ROR we observed in our data set. The simulations of contingency tables for each permutation were implemented using Patefield's algorithm (Patefield, 1981), and those tables containing zeroes were adjusted using Haldane's correction (Haldane, 1956). Overall, this permutation testing was run twice, once with the syndromic and once with the non-syndromic human autism-associated genes, with significant results determined as those with a q-value < 0.05.

4.3 Results

Read Processing and Alignment

RNA-Sequencing produced a mean of ~61 million reads per replicate, ranging from ~38-100 million. Although this was a larger range than expected, there was little evidence of this having substantial effects on the data. Correlation testing between principal components (PCs) of the data and the read counts showed one significant correlation, with PC12 at a q-value of 0.02. However, PC12 explains only 3% of the variance of the data and effects of this are thus likely to be minimal (Figure 4.2). The initial application of adapter trimming, allowing trimming of a read to as low as 30bp before disposal, caused substantial read loss, especially amongst the 3' reads. Of the 5' reads, a mean of 6.03% of reads were disposed of (ranging from 1.97-17.93%), whereas of the 3' reads, a mean of 37.61% of reads were disposed of (ranging from 30.95-43.78%). As mentioned in the methods, this was deemed to be excessive, and adapter trimming was re-applied with variable thresholds for disposal. Subsequently, of the 5' reads, a mean of 6.14% of reads were discarded when they reached a length smaller than 50bp (ranging from 2.03-18.22%), while of the 3' reads, a mean of 15.34% of reads were discarded when their lengths became smaller than 17bp (ranging from 11.01-24.87%) (Figure 4.3).

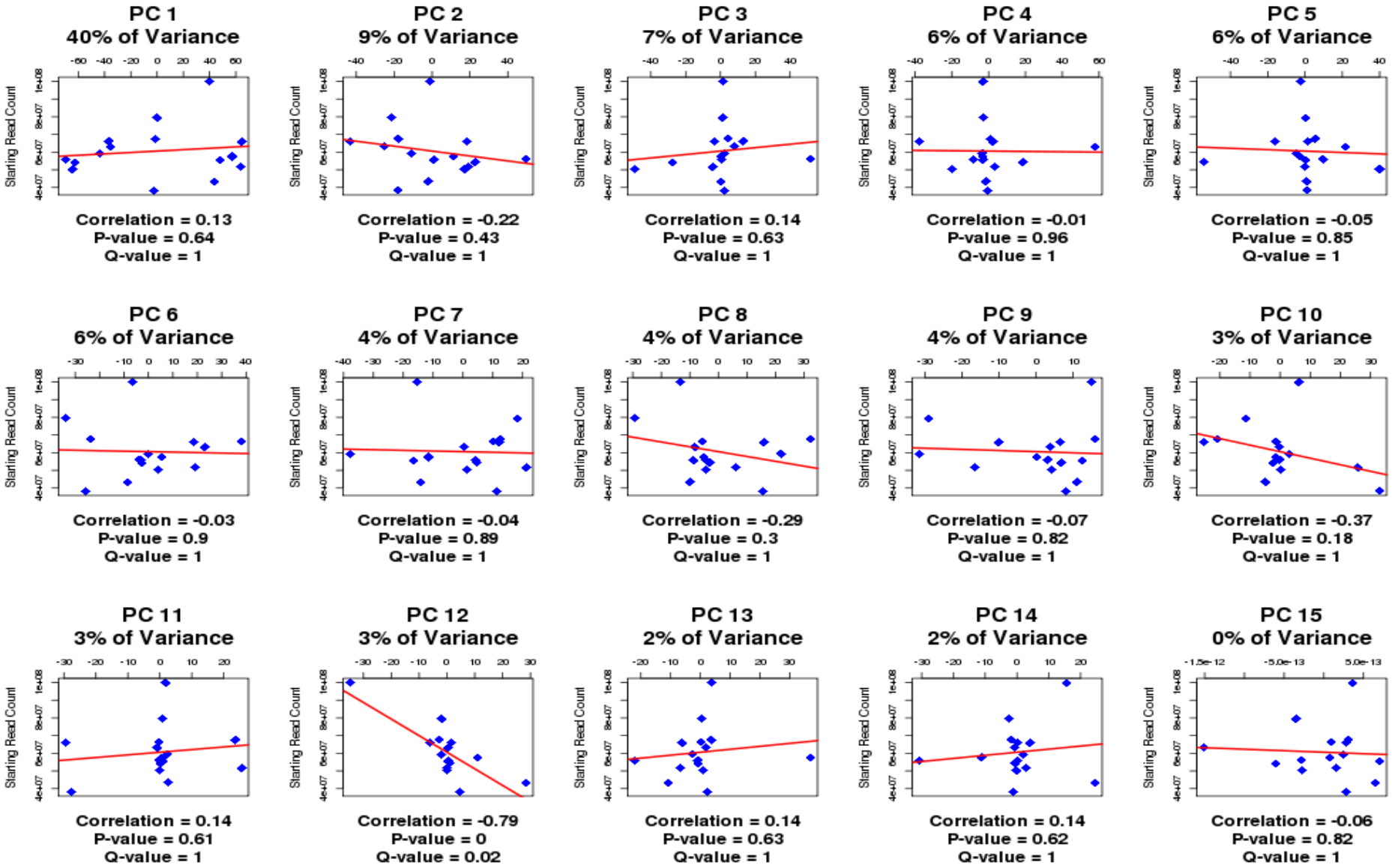
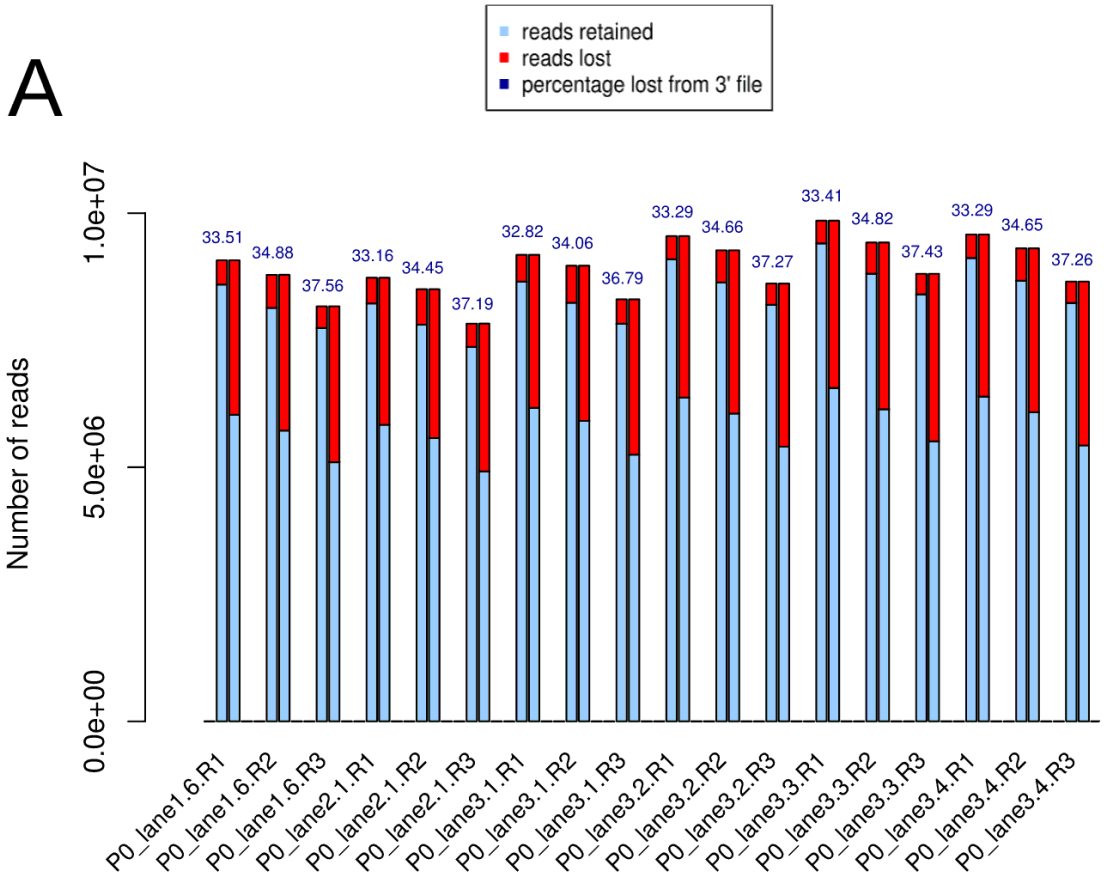

PC 1
40% of Variance
Correlation = 0.13
P-value = 0.64
Q-value = 1
PC 2
9% of Variance
Correlation = -0.22
P-value = 0.43
Q-value = 1
PC 3
7% of Variance
Correlation = 0.14
P-value = 0.63
Q-value = 1
PC 4
6% of Variance
Correlation = -0.01
P-value = 0.96
Q-value = 1
PC 5
6% of Variance
Correlation = -0.05
P-value = 0.85
Q-value = 1
PC 6
6% of Variance
Correlation = -0.03
P-value = 0.9
Q-value = 1
PC 7
4% of Variance
Correlation = -0.04
P-value = 0.89
Q-value = 1
PC 8
4% of Variance
Correlation = -0.29
P-value = 0.3
Q-value = 1
PC 9
4% of Variance
Correlation = -0.07
P-value = 0.82
Q-value = 1
PC 10
3% of Variance
Correlation = -0.37
P-value = 0.18
Q-value = 1
PC 11
3% of Variance
Correlation = 0.14
P-value = 0.61
Q-value = 1
PC 12
3% of Variance
Correlation = -0.79
P-value = 0
Q-value = 0.02
PC 13
2% of Variance
Correlation = 0.14
P-value = 0.63
Q-value = 1
PC 14
2% of Variance
Correlation = 0.14
P-value = 0.62
Q-value = 1
PC 15
0% of Variance
Correlation = -0.06
P-value = 0.82
Q-value = 1
Starling Read Count

Figure 4.2. Correlation plots between principal components of the samples, and the initial read counts. Each data point (in blue) represents a single replicate of a sample; linear models have been fitted (in red) and correlation tests applied and FDR corrected. Only one principal component shows statistically significant correlation (PC12), yet this explains only 3% of the variance.


A
reads retained
reads lost
percentage lost from 3' file
Number of reads
0.0e+00
5.0e+06
1.0e+07
33.51
34.88
37.56
33.16
34.45
37.19
32.82
34.06
36.79
33.29
34.66
37.27
33.41
34.82
37.43
33.29
34.65
37.26
P0_lane1.6.R1
P0_lane1.6.R2
P0_lane1.6.R3
P0_lane2.1.R1
P0_lane2.1.R2
P0_lane2.1.R3
P0_lane3.1.R1
P0_lane3.1.R2
P0_lane3.1.R3
P0_lane3.2.R1
P0_lane3.2.R2
P0_lane3.2.R3
P0_lane3.3.R1
P0_lane3.3.R2
P0_lane3.3.R3
P0_lane3.4.R1
P0_lane3.4.R2
P0_lane3.4.R3

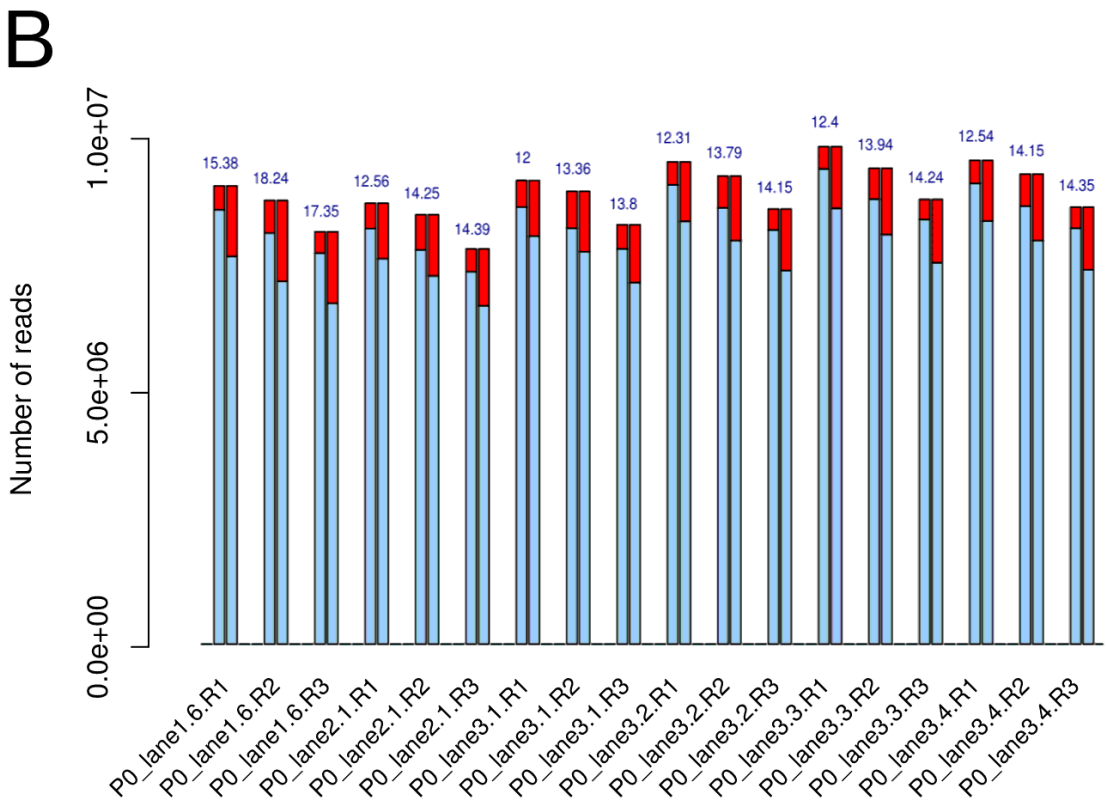

B
Number of reads
0.0e+00
5.0e+06
1.0e+07
15.38
18.24
17.35
12.56
14.25
14.39
12
13.36
13.8
12.31
13.79
14.15
12.4
13.94
14.24
12.54
14.15
14.35
P0_lane1.6.R1
P0_lane1.6.R2
P0_lane1.6.R3
P0_lane2.1.R1
P0_lane2.1.R2
P0_lane2.1.R3
P0_lane3.1.R1
P0_lane3.1.R2
P0_lane3.1.R3
P0_lane3.2.R1
P0_lane3.2.R2
P0_lane3.2.R3
P0_lane3.3.R1
P0_lane3.3.R2
P0_lane3.3.R3
P0_lane3.4.R1
P0_lane3.4.R2
P0_lane3.4.R3

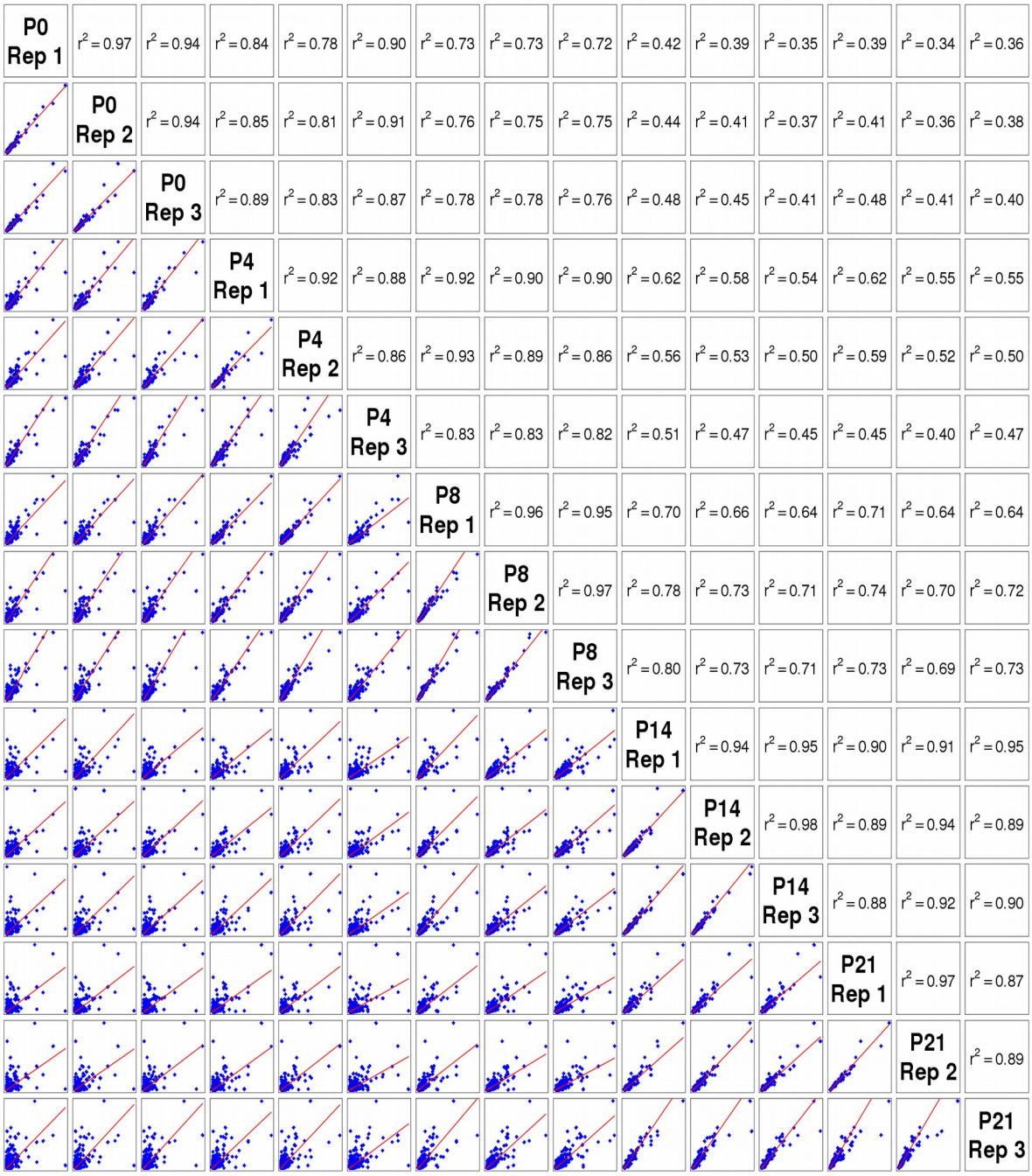
Figure 4.3. Bar plots displaying an example of the effects of adapter trimming (and subsequent read disposal) on read counts at one of the time-points (P0). This is shown for the initial adapter trimming (A), where all reads trimmed to below 30bp were removed, and for the differential adapter trimming (B), where all 5' reads trimmed to below 50bp, and all 3' reads trimmed to below 17bp, were removed. This quality control step was implemented prior to consolidating the data from across different RNA-Seq lanes (hence the replicates are split into multiple lanes). Each pair of bars are for the 5' and 3' reads, from left to right, respectively. The light blue of the bar indicates retained reads, and the red of the bar shows the reads lost from trimming. The dark blue numbers above the bars indicate the percentage of reads lost from the 3' bar.

The reads showed a very high percentage mappability, with a mean of 99.01% of reads aligning to the mouse reference genome (ranging from 98.96-99.07%). Of those that aligned, a mean of 74.36% aligned to a unique position in the genome (ranging from 72.17-76.91%), and these were the only reads retained for subsequent analyses (Table A2). This maintained a good level of coverage, with 19,343 (50.16%) of genes maintaining an RPKM > 0.3 (a threshold that has been previously suggested as suitable for a background expression level; Ramsköld *et al.* 2009).

Several preliminary analyses of these gene expression levels confirmed the quality of the data, the success of the experimental design, and showed the consistency of expression as expected. Replicates showed high levels of correlation where expected, with mean correlation coefficients of gene expression values being considerably higher between replicates within a time-point, with an r^2 of 0.93 (ranging from 0.86 to 0.98), compared with across timepoints, with an r^2 of 0.65 (ranging from 0.34 to 0.95) (Figure 4.4). Principal Component Analysis (PCA) was carried out as a further quality control measure, as well as to visualize the extent of the effect of developmental timing on the data. PCA was used in correlation analyses to quantify the quality of the samples, including correlation with RIN (Figure 4.5A), and with RNA concentration (Figure 4.5B). Although slight correlations were evident, none were significant following multiple testing corrections, providing confidence that these metrics had little effect on the resultant findings. PCA also displayed distinct separation by

developmental timepoint using the first two components (Figure 4.6), which together explain 49% of the variance. This demonstrates that the PCs were effectively captured at these stages. Finally, visualization of common gene marker expression for specific cell type was implemented to confirm that PCs had been successfully isolated and contributed the majority of the captured RNA. For this, transcript abundance of markers was compared in Reads Per Kilobase per Million reads (RPKM), to account for sample size and gene length, for markers of various cell types observed in the cerebellum. These included cerebellar granule neurons, Basket and Stellate cells, Golgi cells, and glia (Figure 4.7). These data indicate a clear predominance for PCs, with PC-specific markers *Calb1* and *Itpr1* displaying the strongest expression, and increasing with postnatal developmental time. Markers for the other cell types demonstrated little-to-no expression, with the exception of the glial cell marker *Gdf10*, which suggests minor contamination by glial cells. This is likely to reflect the close physical association of Bergmann glia with PCs (Yamada & Watanabe, 2002).

A


P0 Rep 1
r² = 0.97
r² = 0.94
r² = 0.84
r² = 0.78
r² = 0.90
r² = 0.73
r² = 0.73
r² = 0.72
r² = 0.42
r² = 0.39
r² = 0.35
r² = 0.39
r² = 0.34
r² = 0.36
P0 Rep 2
r² = 0.94
r² = 0.85
r² = 0.81
r² = 0.91
r² = 0.76
r² = 0.75
r² = 0.75
r² = 0.44
r² = 0.41
r² = 0.37
r² = 0.41
r² = 0.36
r² = 0.38
P0 Rep 3
r² = 0.89
r² = 0.83
r² = 0.87
r² = 0.78
r² = 0.78
r² = 0.76
r² = 0.48
r² = 0.45
r² = 0.41
r² = 0.48
r² = 0.41
r² = 0.40
P4 Rep 1
r² = 0.92
r² = 0.88
r² = 0.92
r² = 0.90
r² = 0.90
r² = 0.62
r² = 0.58
r² = 0.54
r² = 0.62
r² = 0.55
r² = 0.55
P4 Rep 2
r² = 0.86
r² = 0.93
r² = 0.89
r² = 0.86
r² = 0.56
r² = 0.53
r² = 0.50
r² = 0.59
r² = 0.52
r² = 0.50
P4 Rep 3
r² = 0.83
r² = 0.83
r² = 0.82
r² = 0.51
r² = 0.47
r² = 0.45
r² = 0.45
r² = 0.40
r² = 0.47
P8 Rep 1
r² = 0.96
r² = 0.95
r² = 0.70
r² = 0.66
r² = 0.64
r² = 0.71
r² = 0.64
r² = 0.64
P8 Rep 2
r² = 0.97
r² = 0.78
r² = 0.73
r² = 0.71
r² = 0.74
r² = 0.70
r² = 0.72
P8 Rep 3
r² = 0.80
r² = 0.73
r² = 0.71
r² = 0.73
r² = 0.69
r² = 0.73
P14 Rep 1
r² = 0.94
r² = 0.95
r² = 0.90
r² = 0.91
r² = 0.95
P14 Rep 2
r² = 0.98
r² = 0.89
r² = 0.94
r² = 0.89
P14 Rep 3
r² = 0.88
r² = 0.92
r² = 0.90
P21 Rep 1
r² = 0.97
r² = 0.87
P21 Rep 2
r² = 0.89
P21 Rep 3

B

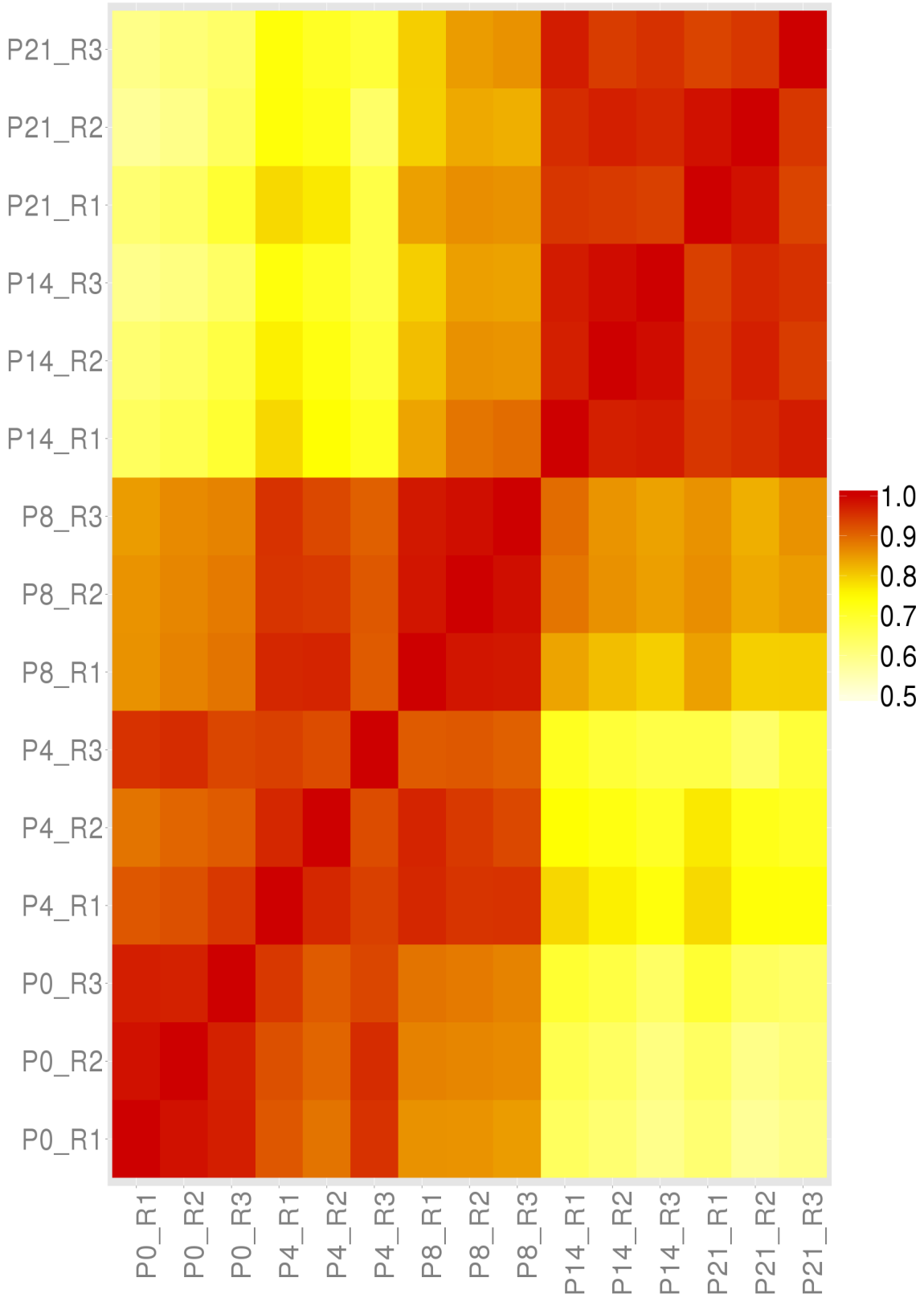
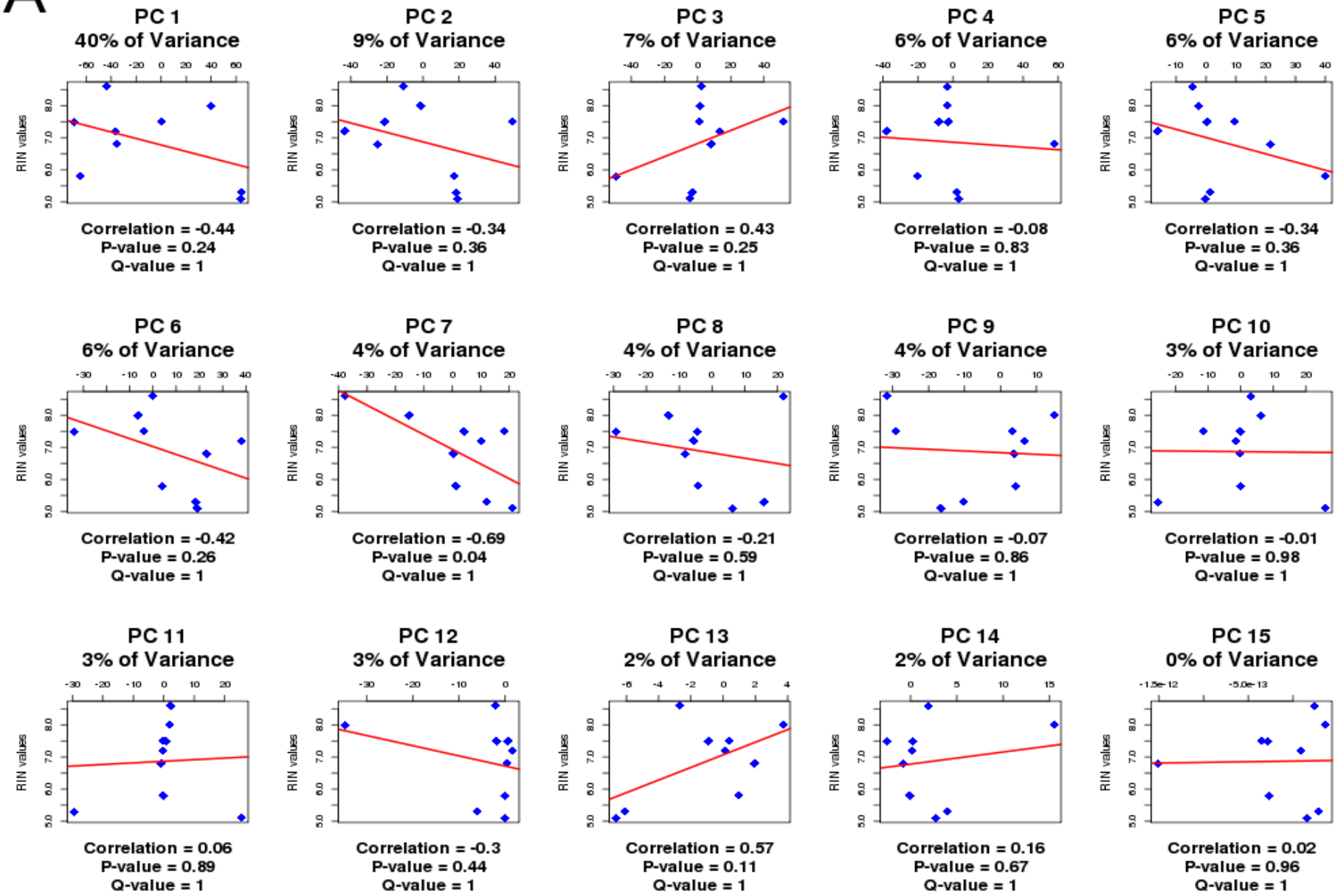

P21_R3
P21_R2
P21_R1
P14_R3
P14_R2
P14_R1
P8_R3
P8_R2
P8_R1
P4_R3
P4_R2
P4_R1
P0_R3
P0_R2
P0_R1
P0_R1
P0_R2
P0_R3
P4_R1
P4_R2
P4_R3
P8_R1
P8_R2
P8_R3
P14_R1
P14_R2
P14_R3
P21_R1
P21_R2
P21_R3
1.0
0.9
0.8
0.7
0.6
0.5

Figure 4.4. Pairwise correlations (r^2) for gene expression data between samples. (A) displays r^2 values between replicates in the upper right quadrant, and correlation plots in the bottom left. For these plots, each data point (in blue) represents a single gene, with expression of one replicate on the x-axis, and expression of the other on the y-axis. Linear models have been fitted (in red). Correlation values are shown to be significantly higher between replicates of the same developmental time-point, or replicates of nearby time-points, compared to those further away in time. (B) displays these same r^2 values in a heatmap for easier visualization.

A


PC 1
40% of Variance
Correlation = -0.44
P-value = 0.24
Q-value = 1
PC 2
9% of Variance
Correlation = -0.34
P-value = 0.36
Q-value = 1
PC 3
7% of Variance
Correlation = 0.43
P-value = 0.25
Q-value = 1
PC 4
6% of Variance
Correlation = -0.08
P-value = 0.83
Q-value = 1
PC 5
6% of Variance
Correlation = -0.34
P-value = 0.36
Q-value = 1
PC 6
6% of Variance
Correlation = -0.42
P-value = 0.26
Q-value = 1
PC 7
4% of Variance
Correlation = -0.69
P-value = 0.04
Q-value = 1
PC 8
4% of Variance
Correlation = -0.21
P-value = 0.59
Q-value = 1
PC 9
4% of Variance
Correlation = -0.07
P-value = 0.86
Q-value = 1
PC 10
3% of Variance
Correlation = -0.01
P-value = 0.98
Q-value = 1
PC 11
3% of Variance
Correlation = 0.06
P-value = 0.89
Q-value = 1
PC 12
3% of Variance
Correlation = -0.3
P-value = 0.44
Q-value = 1
PC 13
2% of Variance
Correlation = 0.57
P-value = 0.11
Q-value = 1
PC 14
2% of Variance
Correlation = 0.16
P-value = 0.67
Q-value = 1
PC 15
0% of Variance
Correlation = 0.02
P-value = 0.96
Q-value = 1
RIN values

B

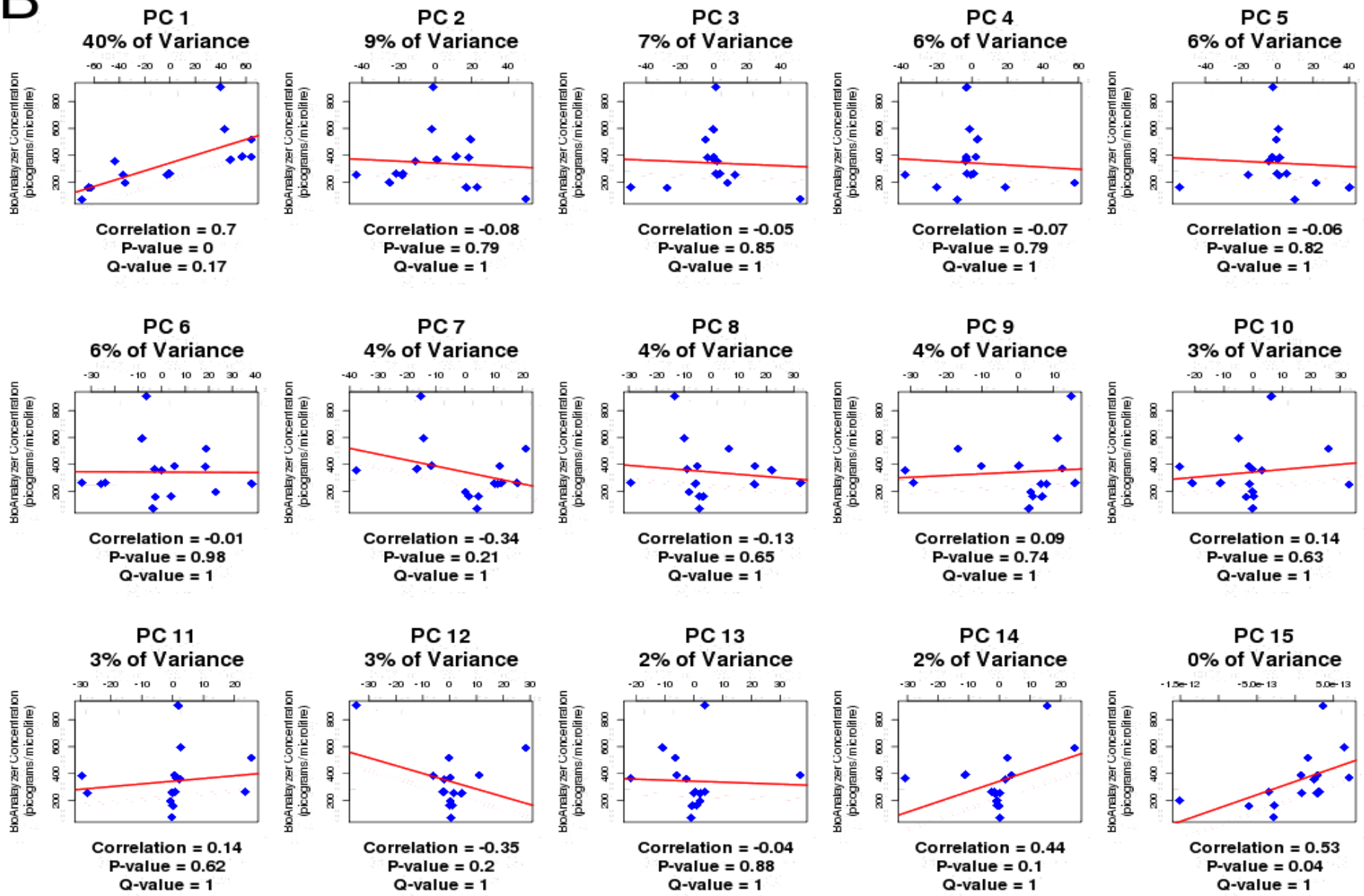

PC 1
40% of Variance
Correlation = 0.7
P-value = 0
Q-value = 0.17
PC 2
9% of Variance
Correlation = -0.08
P-value = 0.79
Q-value = 1
PC 3
7% of Variance
Correlation = -0.05
P-value = 0.85
Q-value = 1
PC 4
6% of Variance
Correlation = -0.07
P-value = 0.79
Q-value = 1
PC 5
6% of Variance
Correlation = -0.06
P-value = 0.82
Q-value = 1
PC 6
6% of Variance
Correlation = -0.01
P-value = 0.98
Q-value = 1
PC 7
4% of Variance
Correlation = -0.34
P-value = 0.21
Q-value = 1
PC 8
4% of Variance
Correlation = -0.13
P-value = 0.65
Q-value = 1
PC 9
4% of Variance
Correlation = 0.09
P-value = 0.74
Q-value = 1
PC 10
3% of Variance
Correlation = 0.14
P-value = 0.63
Q-value = 1
PC 11
3% of Variance
Correlation = 0.14
P-value = 0.62
Q-value = 1
PC 12
3% of Variance
Correlation = -0.35
P-value = 0.2
Q-value = 1
PC 13
2% of Variance
Correlation = -0.04
P-value = 0.88
Q-value = 1
PC 14
2% of Variance
Correlation = 0.44
P-value = 0.1
Q-value = 1
PC 15
0% of Variance
Correlation = 0.53
P-value = 0.04
Q-value = 1
BioAnalyzer Concentration (picograms/microlitre)

Figure 4.5. Correlation plots with principal components of the samples, including correlation with RNA Integrity Number (RIN) (A) and RNA concentration (BioAnalyzer Concentration) (B). Each data point (in blue) represents a single replicate of a sample; linear models have been fitted (in red) and correlation tests applied and FDR corrected. No principal component for either comparison showed significant correlation following FDR corrections.

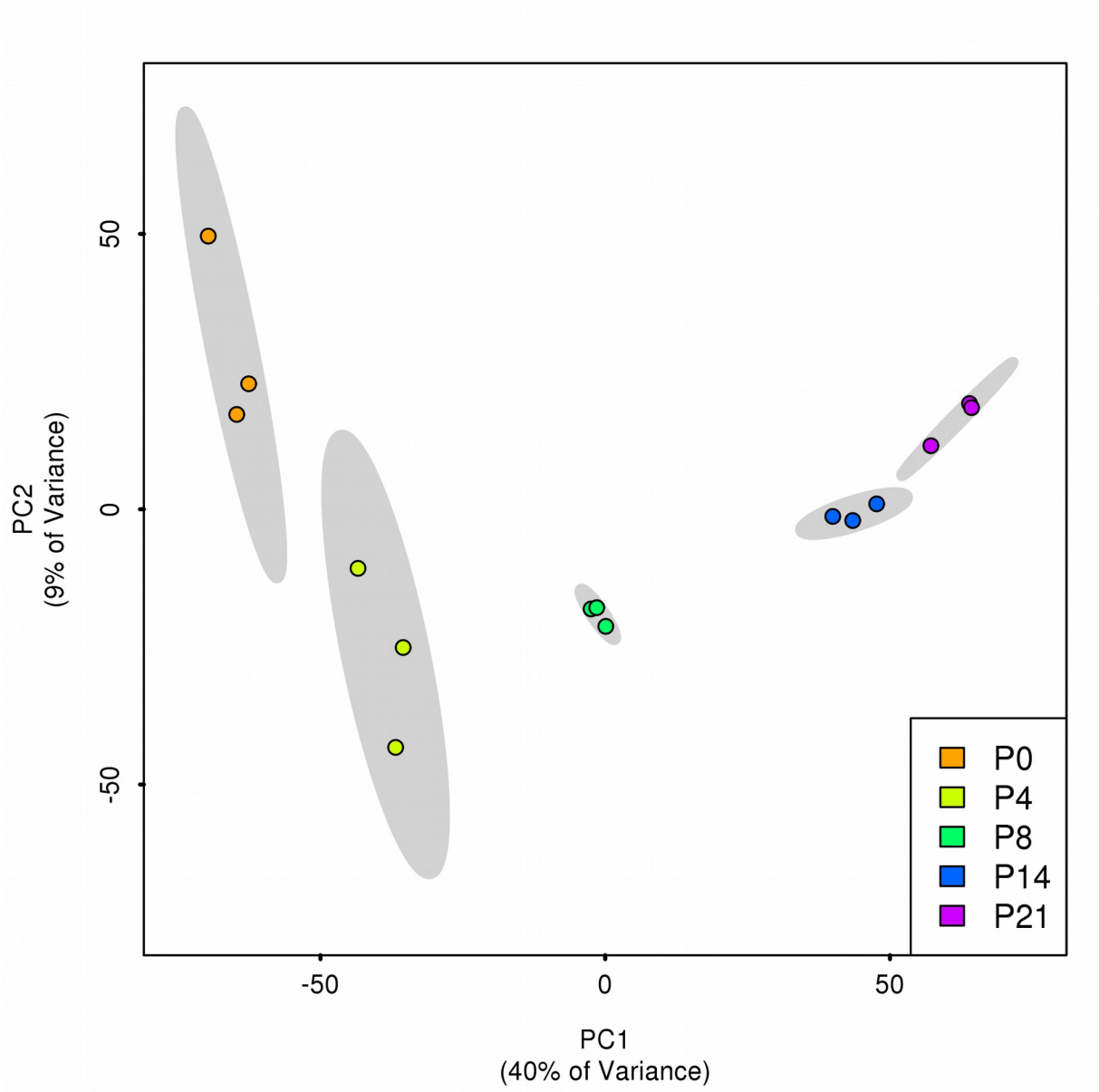


Figure 4.6. A scatter plot of the first two principal components of the expression data, indicating that the PCs were successfully captured at distinct and independent developmental timings. Samples are coloured by time-point (in correspondence with Figure 4.1), and grey ellipses have been included representing 95% confidence intervals. Therefore, the first two principal components, explaining nearly half of the variance, display full separation and are significantly different across developmental timepoints.

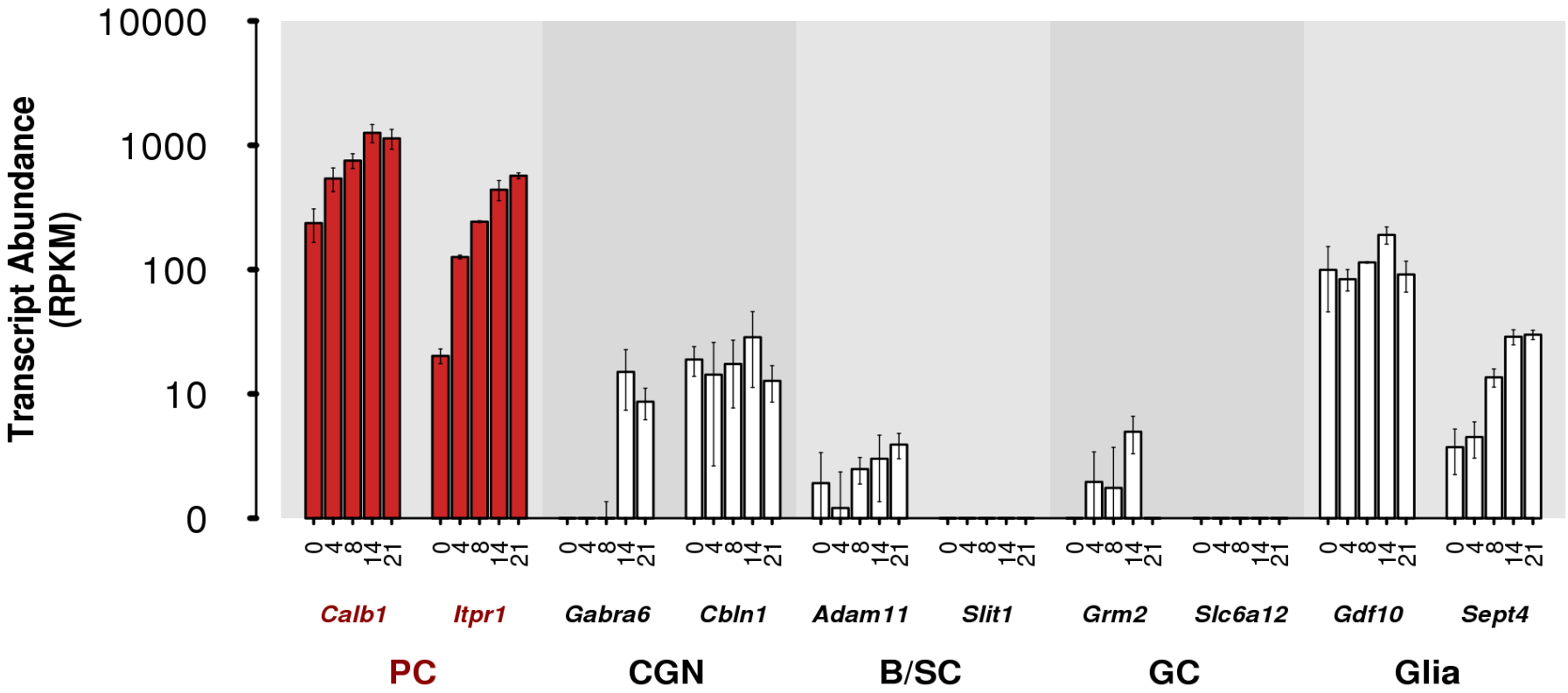


Figure 4.7. Bar plots displaying the expression of known cell-type specific markers across the developmental time-points, showing predominant capture of PCs. Gene expression data has been normalized for sample size and gene length (RPKM) and displayed on a logarithmic scale. The PC data is highlighted in red, with whiskers indicating standard deviation. CGN: cerebellar granule neurons; B/SC: basket/stellate cells; GC: Golgi cells; BG: Bergmann glia.

Weighted Gene Co-expression Network Analysis (WGCNA)

As mentioned in the methods, this analysis was implemented on genes with non-zero expression in at least one sample, and with non-zero variance across all samples, which were found to constitute 54.27% of the total genes in the reference genome. Subsequent WGCNA algorithms were applied, separating the data into 46 clusters initially. Modules were then merged based on eigengene similarity, using a dendrogram height of 0.5 (merging all modules on the same branch above this height), leaving a total of 17 gene clusters from the data (Figure 4.8). To allow for greater visualization of the clusters, the standard colour palette used by WGCNA was reduced to only colours within a specific RGB range to ensure visibility (removing colours such as ivorywhite). This was followed by plotting of all of the clusters (Figure 4.9).

Of these clusters, only two yielded significantly enriched functional annotations (of GO, KEGG, and MGI enrichment analyses), and these were the two clusters significantly positively and negatively correlated with developmental time (Figure 4.10). The top 5 significant GO enrichment for each category, and for the positively and negatively correlated temporal clusters, can be seen in Figure 4.11. For the positively correlated cluster, many of these are focused around membrane, cytoplasm, and transporter activity. Significant enrichment of KEGG pathways is also displayed, with some more informative results. The most striking of these is the KEGG ataxia pathway (Figure A1 [see Appendix]), which further reinforces our

understanding of PC development, because abnormal development and dysfunction of PCs are known to underlie many cerebellar ataxias (Becker, 2014; Chopra & Shakkottai, 2014). In addition to this, there are enriched processes or phenotypes such as tremors, and nervous system electrophysiology. These also reiterate what is already known about PC development, and suggest that this cluster is suitable for exploitation in exploring genetic correlates of disorders in developing PCs. Significant GO enrichment for the negatively correlated temporal cluster can also be seen, including multiple terms associated with ribosomes and RNA binding. Significantly enriched MGI phenotypes confirmed the primary enrichment for ribosomal terms, in addition to cell cycle and DNA replication, which itself can be associated with proliferation-related terms in the KEGG analysis.

A
Height
0.0
0.2
0.4
0.6
0.8
1.0
MEmagenta
MElightslateblue
MEmediumpurple
MEmediumvioletred
MEdarkviolet
MEmaroon
MEblue
MEfirebrick
MElightskyblue
MEdarkmagenta
MEblueviolet
MEdeepskyblue
MElightcoral
MEgreenyellow
MEcoral
MEgold
MElimegreen
MEmediumorchid
MEdarkgoldenrod
MElightsalmon
MElightgoldenrod
MEgreen
MElightseagreen
MEmediumseagreen
MEhotpink
MEcyan
MElawngreen
MEchocolate
MEforestgreen
MEchartreuse
MEgoldenrod
MEdarkorchid
MEdarksalmon
MEdodgerblue
MEaquamarine
MEdarkturquoise
MEindianred
MEbrown
MEdarkred
MEcornflowerblue
MEmediumturquoise
MEdarkorange
MEmediumspringgreen
MEnavy
MEdeeppink
MEmediumslateblue
Clustering of module eigengenes
B
Height
0.5
0.6
0.7
0.8
0.9
1.0
1.1
MEdarkturquoise
MEnavy
MEaquamarine
MElightslateblue
MEdarksalmon
MEdarkorchid
MEdodgerblue
MEmediumseagreen
MEchartreuse
MEblue
MEindianred
MEgreenyellow
MEdarkorange
MElimegreen
MEcornflowerblue
MEchocolate
MElightgoldenrod
Merged clustering of module eigengenes

C

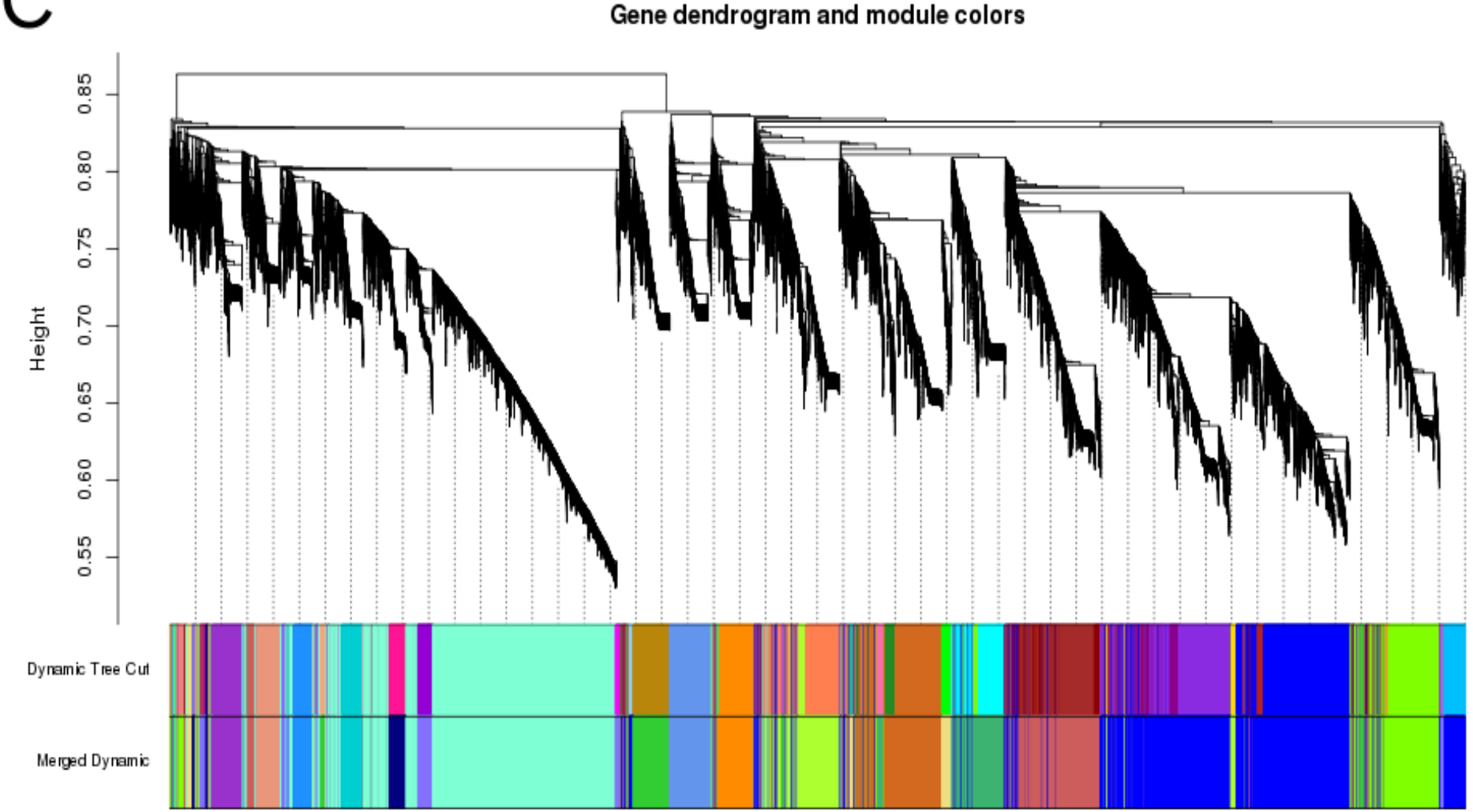

Gene dendrogram and module colors
Height
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Dynamic Tree Cut
Merged Dynamic

Figure 4.8. Dendrograms displaying the clusters formed from WGCNA. Hierarchical clustering of the eigengenes of the original 46 clusters can be seen in (A), followed by the same clustering of eigengenes after merging of similar clusters, leaving 17, in (B). (C) displays hierarchical clustering of all of the genes in the data-set (post-QC), with assignment to the clusters (as seen in (A) and (B)) displayed in the coloured bars below. The top coloured bar is representative of the 46 clusters (as shown in (A)) following dynamic tree cutting, whereas the lower coloured bar shows this same information following merging of similar clusters (as shown in (B)). All clustering was performed using a euclidean distance measure and the average linkage method.

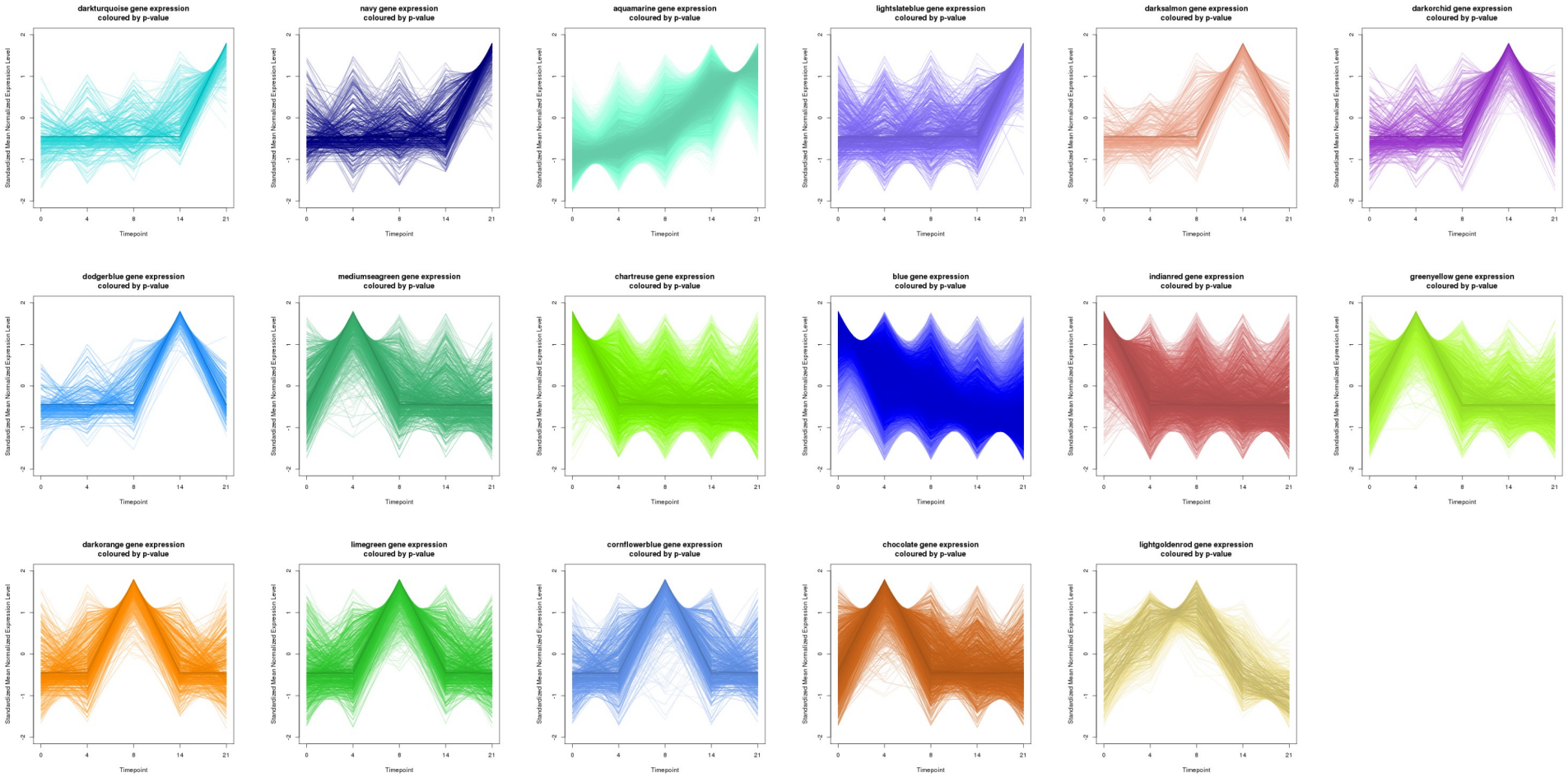
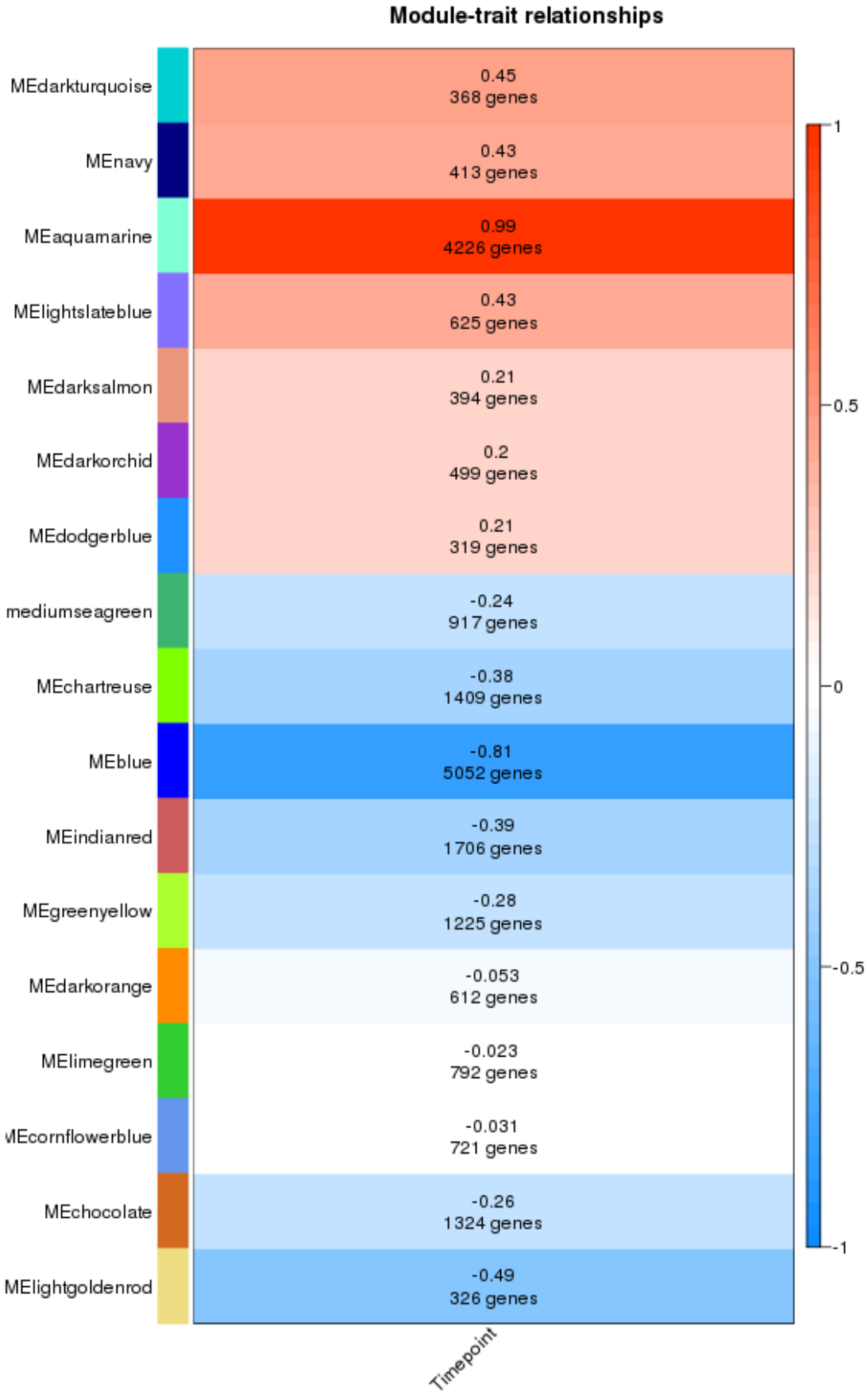


Figure 4.9. Line graphs displaying standardized gene expression across each of the 17 clusters identified through WGCNA. Each plot represents a single cluster, and each line represents the expression of a single gene over the course of developmental timing. Expression data was standardized to a mean of zero, and a standard deviation of one.

A


Module-trait relationships
MEdarkturquoise
0.45
368 genes
MEnavy
0.43
413 genes
MEaquamarine
0.99
4226 genes
MElightslateblue
0.43
625 genes
MEdarksalmon
0.21
394 genes
MEdarkorchid
0.2
499 genes
MEdodgerblue
0.21
319 genes
mediumseagreen
-0.24
917 genes
MEchartreuse
-0.38
1409 genes
MEblue
-0.81
5052 genes
MEindianred
-0.39
1706 genes
MEgreenyellow
-0.28
1225 genes
MEdarkorange
-0.053
612 genes
MElimegreen
-0.023
792 genes
MEcornflowerblue
-0.031
721 genes
MEchocolate
-0.26
1324 genes
MElightgoldenrod
-0.49
326 genes
Timepoint
1
0.5
0
-0.5
-1

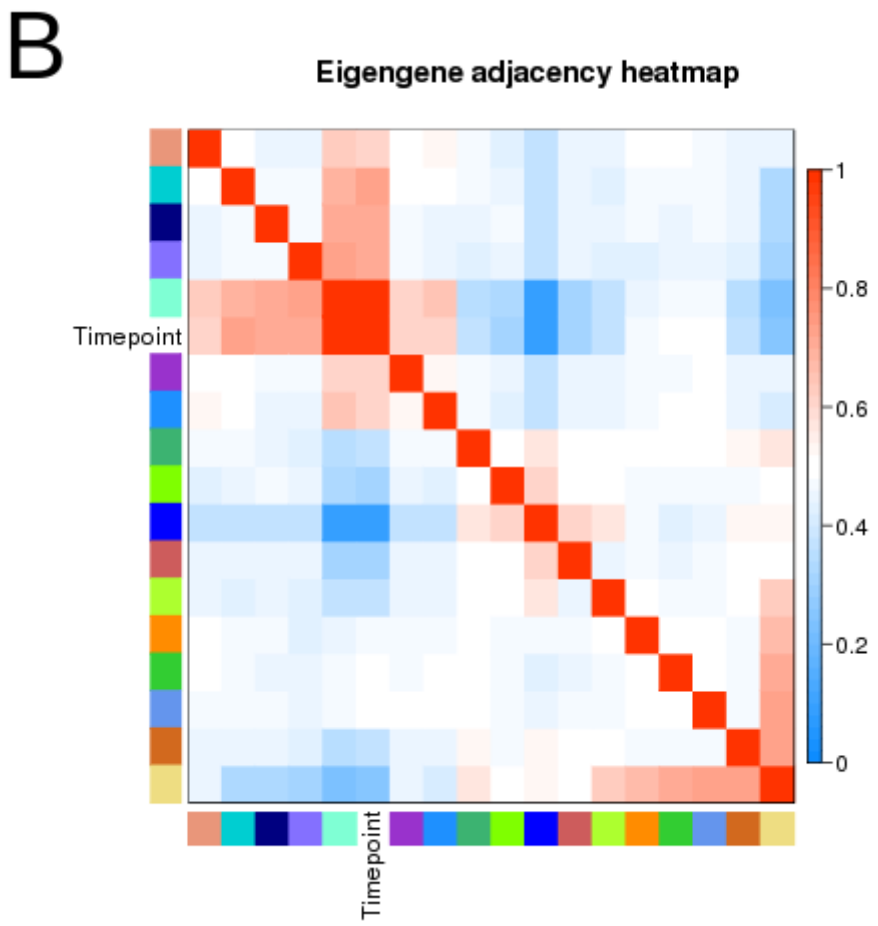
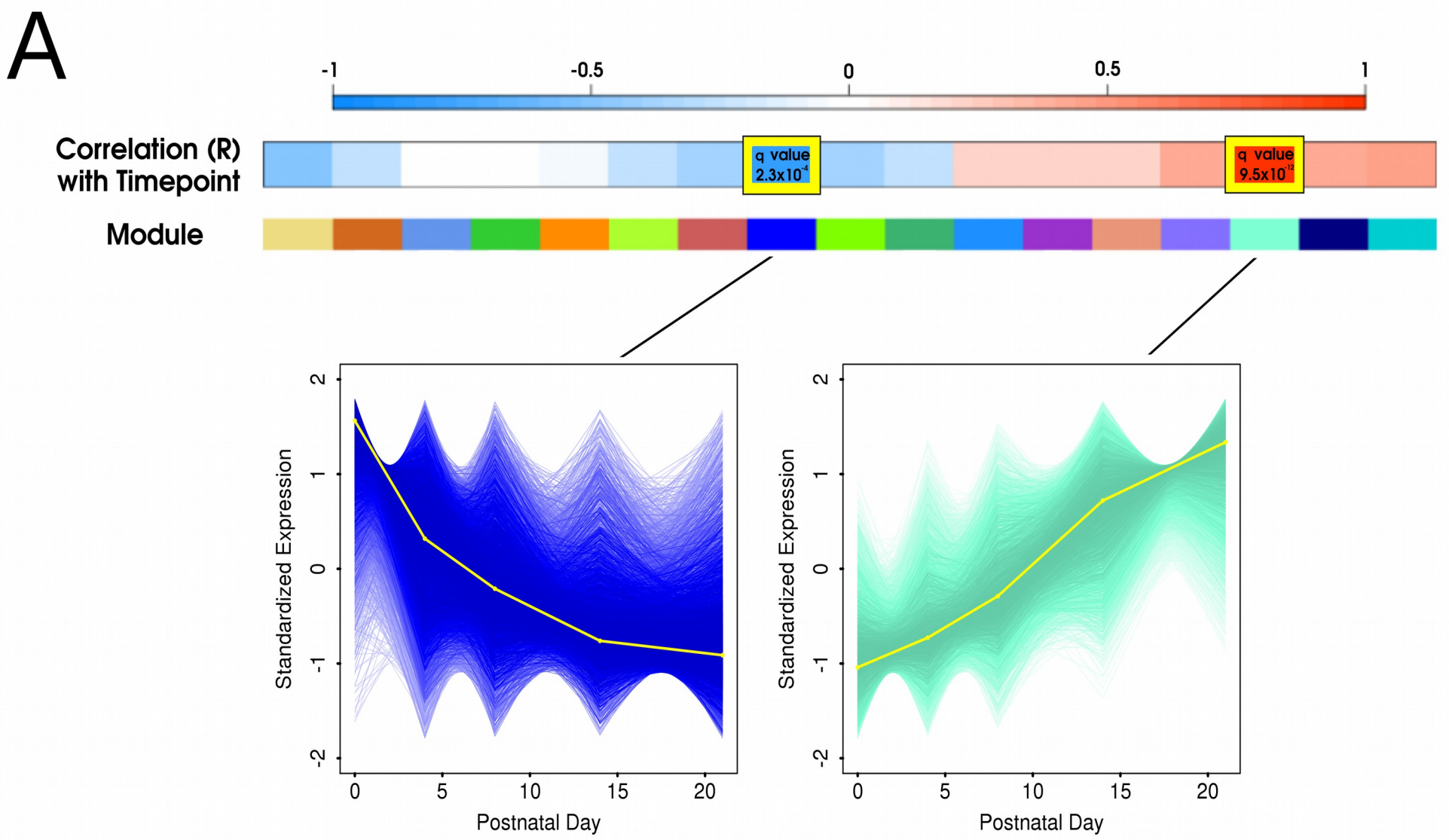


Figure 4.10. Heat plots displaying correlations between the WGCNA cluster eigengenes and developmental timepoint. Panel (A) displays the correlation (r) between each of the clusters and timepoint, coloured by correlation and including information on the number of genes in each cluster. Panel (B) displays similar correlation information, but in a full heatmap comparing each cluster with one another. Timepoint has also been included as a factor in (B), showing strong correlation with the aquamarine cluster.


A
-1
-0.5
0
0.5
1
Correlation (R)
with Timepoint
q value
2.3x10-4
q value
9.5x10-12
Module
Standardized Expression
Postnatal Day
Standardized Expression
Postnatal Day

B

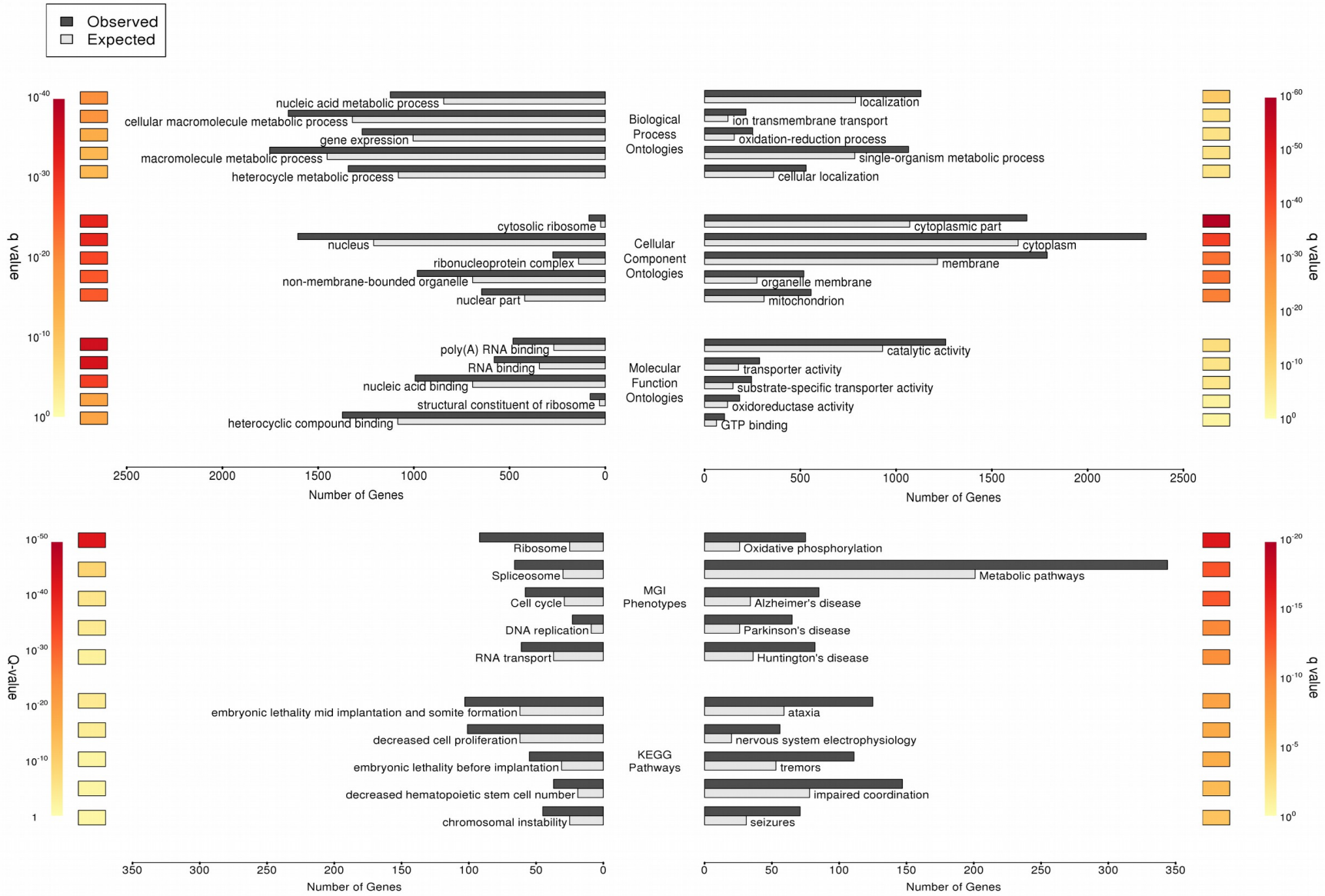

Observed
Expected
q value
Biological Process Ontologies
nucleic acid metabolic process
cellular macromolecule metabolic process
gene expression
macromolecule metabolic process
heterocycle metabolic process
localization
ion transmembrane transport
oxidation-reduction process
single-organism metabolic process
cellular localization
Cellular Component Ontologies
cytosolic ribosome
nucleus
ribonucleoprotein complex
non-membrane-bounded organelle
nuclear part
cytoplasmic part
cytoplasm
membrane
organelle membrane
mitochondrion
Molecular Function Ontologies
poly(A) RNA binding
RNA binding
nucleic acid binding
structural constituent of ribosome
heterocyclic compound binding
catalytic activity
transporter activity
substrate-specific transporter activity
oxidoreductase activity
GTP binding
Number of Genes
Q-value
MGI Phenotypes
Ribosome
Spliceosome
Cell cycle
DNA replication
RNA transport
Oxidative phosphorylation
Metabolic pathways
Alzheimer's disease
Parkinson's disease
Huntington's disease
KEGG Pathways
embryonic lethality mid implantation and somite formation
decreased cell proliferation
embryonic lethality before implantation
decreased hematopoietic stem cell number
chromosomal instability
ataxia
nervous system electrophysiology
tremors
impaired coordination
seizures
Number of Genes

Figure 4.11. The two WGCNA clusters returning significant functional annotations, both of which are significantly correlated (negatively and positively) with developmental timing. (A) displays the correlation data for the WGCNA clusters (as in Figure 4.10A), with correlation test q-values displayed for the two significantly correlated (blue and aquamarine) modules, and with attached plots of the expression values over development for the two modules. For these, each line represents the expression of a single gene over the course of developmental timing. An additional line in yellow has been added to indicate the eigengene of the cluster. Expression data was standardized to a mean of zero, and standard deviation of one. (B) consists of bar plots of functional annotations (for the blue module on the left, and aquamarine module on the right). This displays the five most statistically significant enrichments for phenotype (MGI), pathway (KEGG) and various ontologies (biological process, cellular component, and molecular function). Overall, functional annotation of the module negatively correlated with time suggests strong association with cell formation and proliferation functions. In contrast, the module positively correlated with developmental time is enriched for Synapse GO terms, observed in both biological process (synaptic transmission) and cellular component (synapse), suggesting strong association with synapse formation. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

Next, enrichment of disorder-related genes was investigated across the 17 WGCNA clusters (Table 4.1). In investigating enrichment of the ataxia gene lists, only the cluster positively correlated with development returned significant results (Figure 4.12), both for the human ataxia gene list, with a 4.3-fold increase on the expected gene occurrence and q-value of $1.94x10^{-5}$, and for the mouse ataxia gene list, with a 8.8-fold increase on the expected gene occurrence and q-value of $1.02x10^{-7}$. Similar enrichment testing was implemented with the SFARI lists, and remarkably, significant enrichment was seen in the same cluster for the mouse-model SFARI list. The fold-enrichment observed was not as large as seen in the ataxia gene lists, which is perhaps to be expected due to the cerebellum's strong association with ataxia; however, this enrichment was still statistically significant. This was observed as a 2.4-fold enrichment for mouse ASD-associated genes, with a q-value of $9.4x10^{-3}$. This significant enrichment was not observed for the human ASD-associated genes (despite a p-value bordering statistical significance, of 0.07, this was not sufficiently robust enough for the large false discovery rate corrections required for the high number of WGCNA modules). Finally, the clusters were tested for enrichment of the human schizophrenia gene list and, as expected, there were no noteworthy enrichments (with testing on the cluster positively correlated with development returning a p-value of 0.59 and a q-value of 1).

Table 4.1. Enrichment testing for disorder-related genes amongst the PC clusters defined by WGCNA. For each disorder in each cluster this includes: the enrichment of the disorder gene list in the cluster, the odds ratio of the gene list over the cluster, the p-value (length corrected), and the q-value (FDR-corrected p-value). Statistically significant results are highlighted in yellow.

Gene-list	Cluster	Enrichment	Odds Ratio	P-value (length corrected)	Q-value (n=102)
Ataxia Human List	aquamarine	4.26	5.72	3.81E-007	1.94E-005
	blue	0.95	0.95	6.70E-001	1.00E+000
	chartreuse	0.00	NA	1.00E+000	1.00E+000
	chocolate	0.86	0.86	6.99E-001	1.00E+000
	cornflowerblue	0.00	NA	1.00E+000	1.00E+000
	darkorange	0.00	NA	1.00E+000	1.00E+000
	darkorchid	0.00	NA	1.00E+000	1.00E+000
	darksalmon	0.00	NA	1.00E+000	1.00E+000
	darkturquoise	0.00	NA	1.00E+000	1.00E+000
	dodgerblue	0.00	NA	1.00E+000	1.00E+000
	greenyellow	0.00	NA	1.00E+000	1.00E+000
	indianred	0.00	NA	1.00E+000	1.00E+000
	lightgoldenrod	0.00	NA	1.00E+000	1.00E+000
	lightslateblue	0.00	NA	1.00E+000	1.00E+000
	limegreen	0.00	NA	1.00E+000	1.00E+000
	mediumseagreen	1.36	1.37	5.32E-001	1.00E+000
	navy	0.00	NA	1.00E+000	1.00E+000
Ataxia Mouse List	aquamarine	8.79	19.38	1.00E-009	1.02E-007
	blue	0.00	NA	1.00E+000	1.00E+000
	chartreuse	0.00	NA	1.00E+000	1.00E+000
	chocolate	2.14	2.20	3.76E-001	1.00E+000
	cornflowerblue	0.00	NA	1.00E+000	1.00E+000
	darkorange	0.00	NA	1.00E+000	1.00E+000
	darkorchid	0.00	NA	1.00E+000	1.00E+000
	darksalmon	0.00	NA	1.00E+000	1.00E+000
	darkturquoise	0.00	NA	1.00E+000	1.00E+000
	dodgerblue	0.00	NA	1.00E+000	1.00E+000
	greenyellow	0.00	NA	1.00E+000	1.00E+000
	indianred	2.03	2.09	3.91E-001	1.00E+000
	lightgoldenrod	0.00	NA	1.00E+000	1.00E+000
	lightslateblue	0.00	NA	1.00E+000	1.00E+000
	limegreen	0.00	NA	1.00E+000	1.00E+000
	mediumseagreen	0.00	NA	1.00E+000	1.00E+000
	navy	0.00	NA	1.00E+000	1.00E+000
Autism Human List (S+1-6)	aquamarine	1.41	1.47	6.98E-002	1.00E+000
	blue	1.39	1.42	2.13E-001	1.00E+000
	chartreuse	0.65	0.64	8.78E-001	1.00E+000
	chocolate	0.50	0.49	9.34E-001	1.00E+000
	cornflowerblue	0.74	0.74	7.01E-001	1.00E+000
	darkorange	0.96	0.96	5.73E-001	1.00E+000
	darkorchid	0.37	0.36	9.11E-001	1.00E+000
	darksalmon	1.44	1.46	2.57E-001	1.00E+000
	darkturquoise	0.51	0.51	8.14E-001	1.00E+000
	dodgerblue	2.86	3.02	8.79E-003	2.24E-001
	greenyellow	0.49	0.48	8.83E-001	1.00E+000
	indianred	1.42	1.45	1.98E-001	1.00E+000
	lightgoldenrod	3.53	3.77	5.16E-002	1.00E+000
	lightslateblue	0.97	0.97	5.86E-001	1.00E+000
	limegreen	1.03	1.03	5.16E-001	1.00E+000
	mediumseagreen	1.60	1.63	1.47E-001	1.00E+000
	navy	0.46	0.45	8.49E-001	1.00E+000
Autism Mouse List	aquamarine	2.36	2.64	2.77E-004	9.42E-003
	blue	1.43	1.46	2.33E-001	1.00E+000
	chartreuse	0.60	0.59	8.72E-001	1.00E+000
	chocolate	0.90	0.90	6.42E-001	1.00E+000
	cornflowerblue	0.59	0.58	7.96E-001	1.00E+000
	darkorange	1.14	1.14	4.93E-001	1.00E+000
	darkorchid	0.00	NA	1.00E+000	1.00E+000
	darksalmon	0.81	0.80	6.70E-001	1.00E+000
	darkturquoise	0.00	NA	1.00E+000	1.00E+000
	dodgerblue	1.57	1.58	3.29E-001	1.00E+000
	greenyellow	0.42	0.42	8.92E-001	1.00E+000
	indianred	0.57	0.56	8.64E-001	1.00E+000
	lightgoldenrod	0.00	NA	1.00E+000	1.00E+000
	lightslateblue	1.68	1.70	3.28E-001	1.00E+000
	limegreen	1.68	1.70	2.71E-001	1.00E+000
	mediumseagreen	0.44	0.43	8.93E-001	1.00E+000
	navy	0.00	NA	1.00E+000	1.00E+000
Schizophrenia Human List	aquamarine	0.97	0.97	5.94E-001	1.00E+000
	blue	1.38	1.42	1.37E-001	1.00E+000
	chartreuse	1.29	1.31	2.82E-001	1.00E+000
	chocolate	0.84	0.83	7.18E-001	1.00E+000
	cornflowerblue	0.37	0.36	9.36E-001	1.00E+000
	darkorange	0.63	0.62	8.25E-001	1.00E+000
	darkorchid	0.73	0.72	7.60E-001	1.00E+000
	darksalmon	0.95	0.95	6.17E-001	1.00E+000
	darkturquoise	0.00	NA	1.00E+000	1.00E+000
	dodgerblue	1.43	1.45	3.44E-001	1.00E+000
	greenyellow	0.72	0.72	7.80E-001	1.00E+000
	indianred	0.31	0.30	9.89E-001	1.00E+000
	lightgoldenrod	1.19	1.20	5.76E-001	1.00E+000
	lightslateblue	1.40	1.42	3.61E-001	1.00E+000
	limegreen	0.00	NA	1.00E+000	1.00E+000
	mediumseagreen	0.26	0.26	9.79E-001	1.00E+000
	navy	1.36	1.37	3.72E-001	1.00E+000

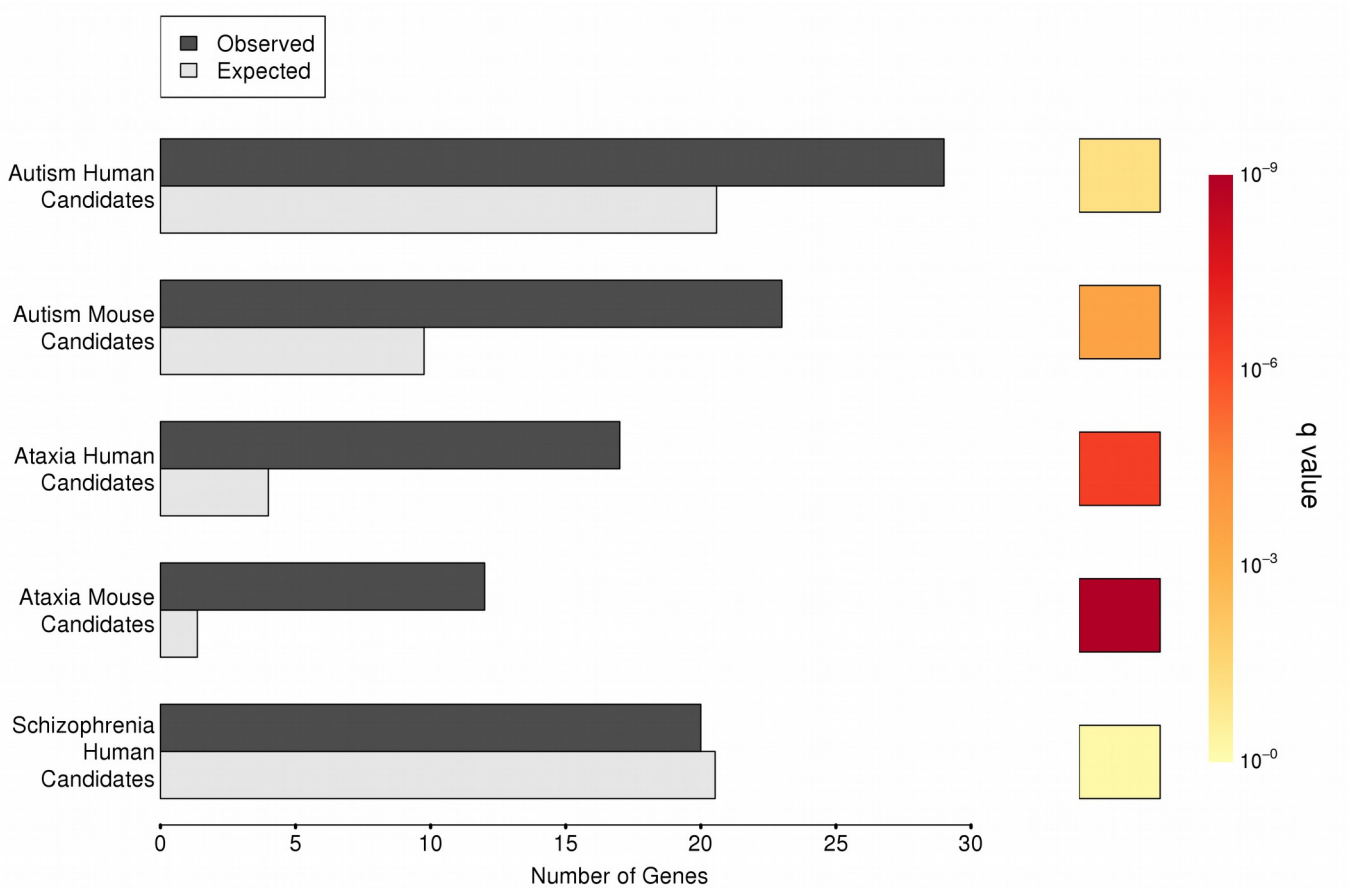


Figure 4.12. A bar plot of disease enrichments observed in the WGCNA cluster positively correlated with developmental timing (aquamarine), including ataxia, autism and schizophrenia associated gene lists (as displayed in Table 4.1). The autism mouse, ataxia human, and ataxia mouse lists all returned significant q-values. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

Finally, enrichment testing for the independent FMRP-bound transcript gene list was also implemented, with the hope of further corroborating autism-related findings (Table 4.2). This showed strong enrichment in this same positively correlated cluster (Figure 4.13), with a 2.7-fold enrichment and q-value of 6.30×10^{-21}. To ensure this FMRP-bound transcript gene list was distinct from the SFARI gene lists, the intersect was visualized between the two gene lists, showing a relatively small level of overlap (Figure 4.14).

Table 4.2. Enrichment testing for genes of FMRP-bound transcripts amongst the PC clusters defined by WGCNA. For each cluster this includes: the enrichment of the FMRP-bound transcript genes in the cluster, the odds ratio of the FMRP-bound transcript gene list over the cluster, the p-value (length corrected), and the q-value. Statistically significant results are highlighted in yellow.

Cluster	Enrichment	Odds Ratio	P-value (length corrected)	Q-value (n=102)
aquamarine	2.70	3.41	6.17E-023	6.30E-021
blue	1.65	1.75	2.13E-002	1.00E+000
chartreuse	0.17	0.15	1.00E+000	1.00E+000
chocolate	0.33	0.31	9.99E-001	1.00E+000
cornflowerblue	0.00	NA	1.00E+000	1.00E+000
darkorange	0.11	0.10	1.00E+000	1.00E+000
darkorchid	0.33	0.31	9.65E-001	1.00E+000
darksalmon	0.00	NA	1.00E+000	1.00E+000
darkturquoise	0.00	NA	1.00E+000	1.00E+000
dodgerblue	0.15	0.14	9.89E-001	1.00E+000
greenyellow	0.08	0.08	1.00E+000	1.00E+000
indianred	0.27	0.25	1.00E+000	1.00E+000
lightgoldenrod	2.09	2.24	7.73E-002	1.00E+000
lightslateblue	0.32	0.30	9.76E-001	1.00E+000
limegreen	0.00	NA	1.00E+000	1.00E+000
mediumseagreen	0.08	0.08	1.00E+000	1.00E+000
navy	0.15	0.14	9.88E-001	1.00E+000

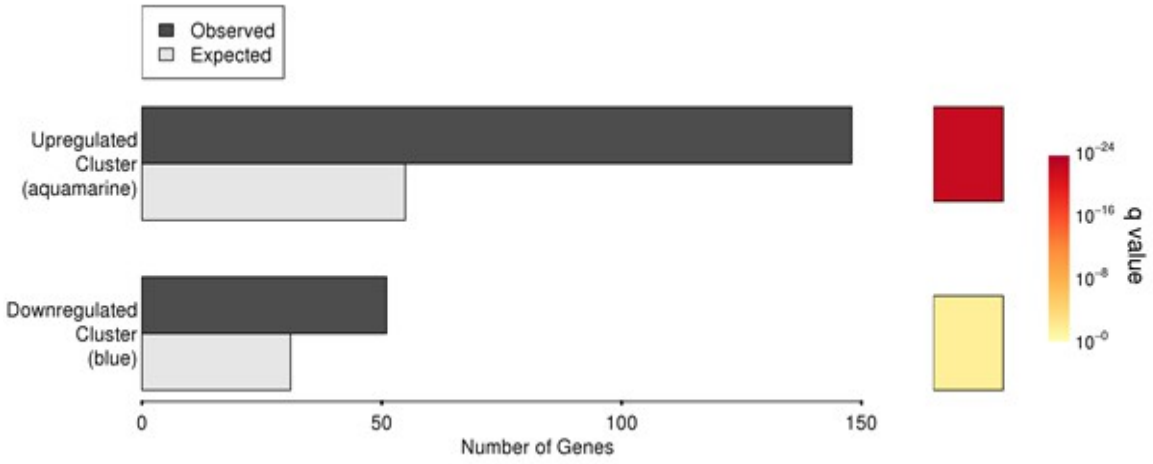


Figure 4.13. A bar plot displaying the enrichment of genes of FMRP-bound transcripts, observed in the WGCNA clusters positively (aquamarine) and negatively (blue) correlated with developmental timing, respectfully (as displayed in Table 4.2). Genes of FMRP-bound transcripts are strongly significantly enriched in the positively correlated cluster. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

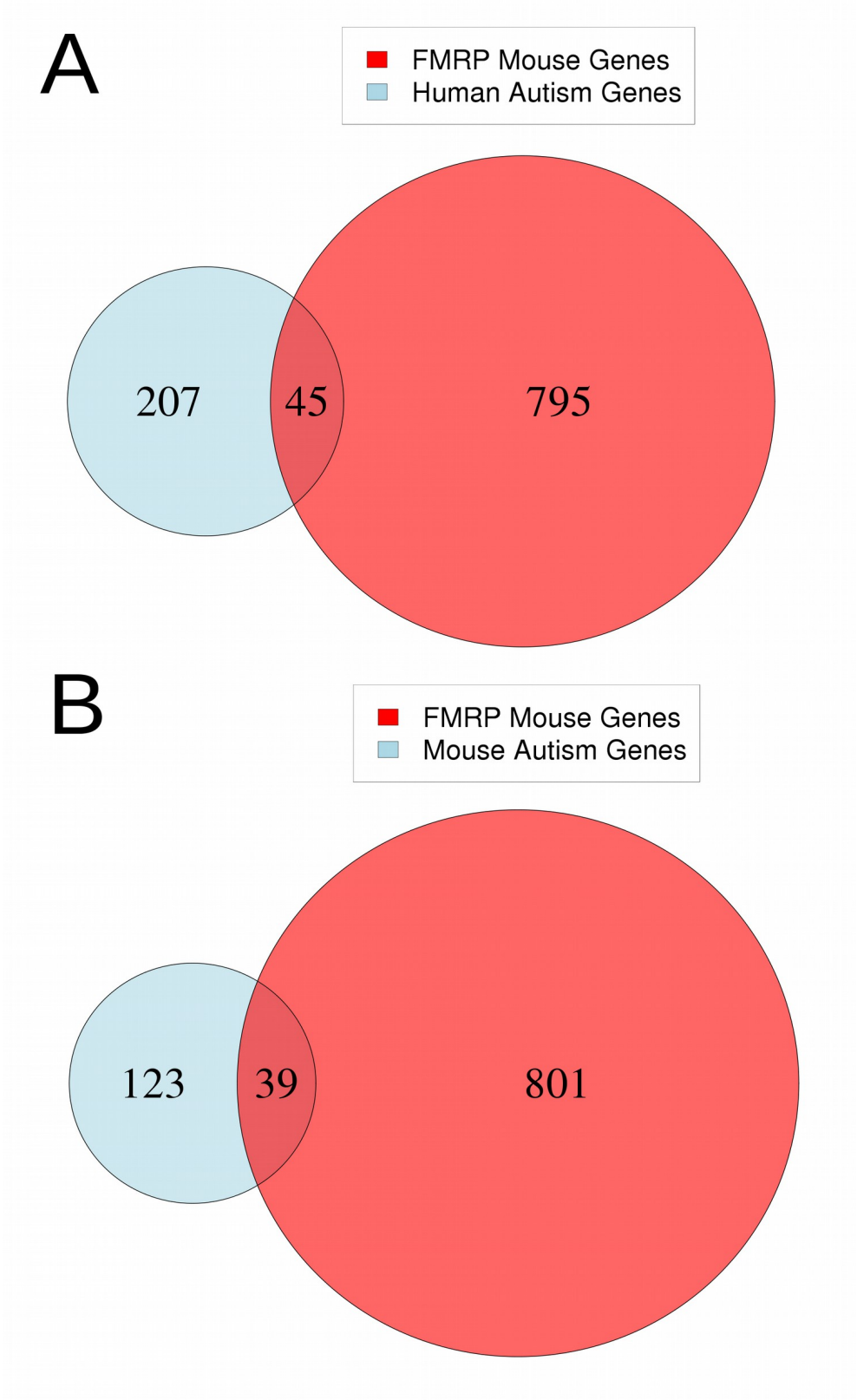


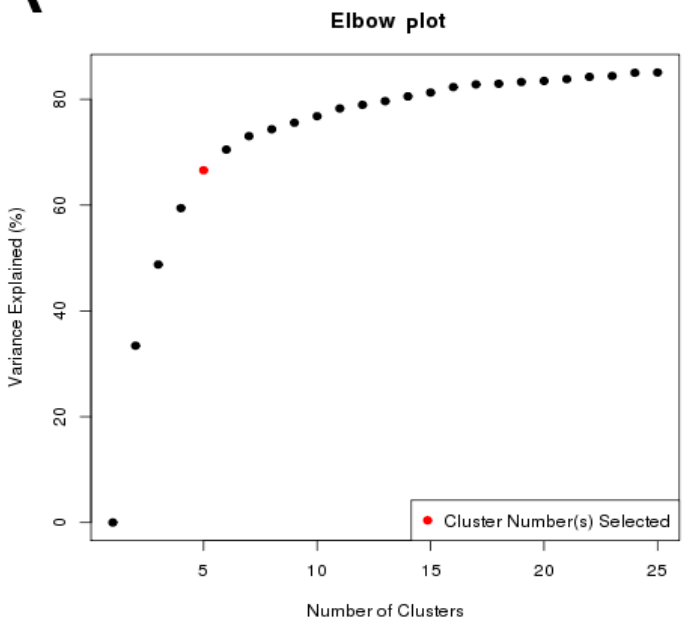
Figure 4.14. Venn diagrams displaying the overlap between autism-associated gene lists and the FMRP-bound transcript gene list, including the SFARI human candidate gene list (A) and the SFARI mouse model gene list (B).

Fuzzy Clustering

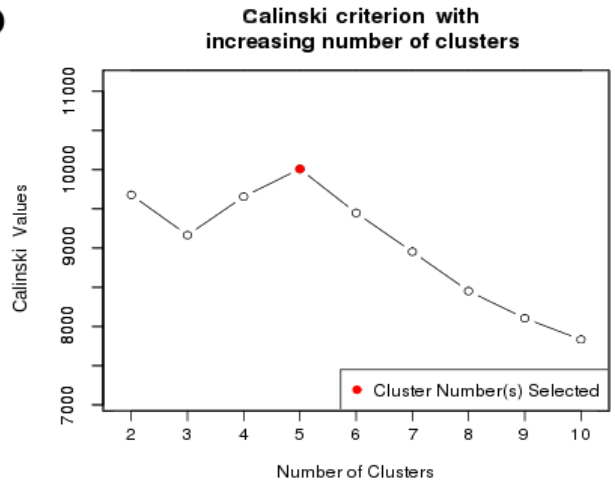
Fuzzy clustering algorithms were applied as a secondary method of analysis to reinforce WGCNA results. The number of clusters (K) was decided through application of three individual methods: the silhouette-width method, the Calinski-Harabasz method, and the elbow method (Figure 4.15). All of these methods suggest a K of five would form optimal clusters, and so this value was applied in subsequent analyses (Figure 4.16). The same disease enrichment analyses were implemented on the five clusters as with the WGCNA modules (Table 4.3). The third cluster, of similar temporal patterning to the WGCNA cluster positively correlated with time, was also the only cluster to return significant enrichments for disease gene lists. These results were also in the same vein, with significant enrichment for the human candidate ataxia gene list, the mouse model ataxia gene list, and the mouse model autism gene list, providing confidence in our WGCNA results (Figure 4.17). The ataxia lists both had high fold-enrichment values, with the human list at 3.7-fold and a q-value of 6.71×10^{-3}, and with the mouse list at 5.7-fold and a q-value of 6.71×10^{-3}. The mouse ASD gene list had a significant enrichment of 2.4-fold, resulting in a q-value of 1.15×10^{-2}. As with the WGCNA, this fuzzy cluster showed an increased enrichment for the human ASD list (1.2-fold), however this was not significant. The same test for human schizophrenia-associated genes was also applied, and also displayed no noteworthy enrichment (0.9-fold). Finally, FMRP-bound transcript enrichment testing was employed for the fuzzy clusters (Table 4.4). Once

again, this corroborated our findings with WGCNA, with the third cluster enriched for genes of FMRP-bound transcripts, exhibiting a 2.6-fold enrichment and q-value of 6.50×10^{-13} (Figure 4.18), with no noteworthy enrichment in other clusters.

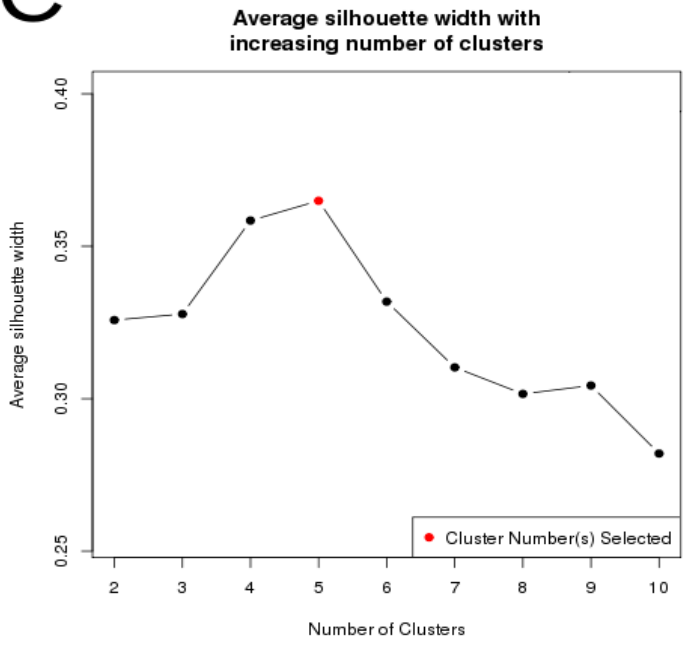
A


Elbow plot
Variance Explained (%)
0
20
40
60
80
5
10
15
20
25
Number of Clusters
Cluster Number(s) Selected

B


Calinski criterion with
increasing number of clusters
Calinski Values
7000
8000
9000
10000
11000
2
3
4
5
6
7
8
9
10
Number of Clusters
Cluster Number(s) Selected

C


Average silhouette width with
increasing number of clusters
Average silhouette width
0.25
0.30
0.35
0.40
2
3
4
5
6
7
8
9
10
Number of Clusters
Cluster Number(s) Selected

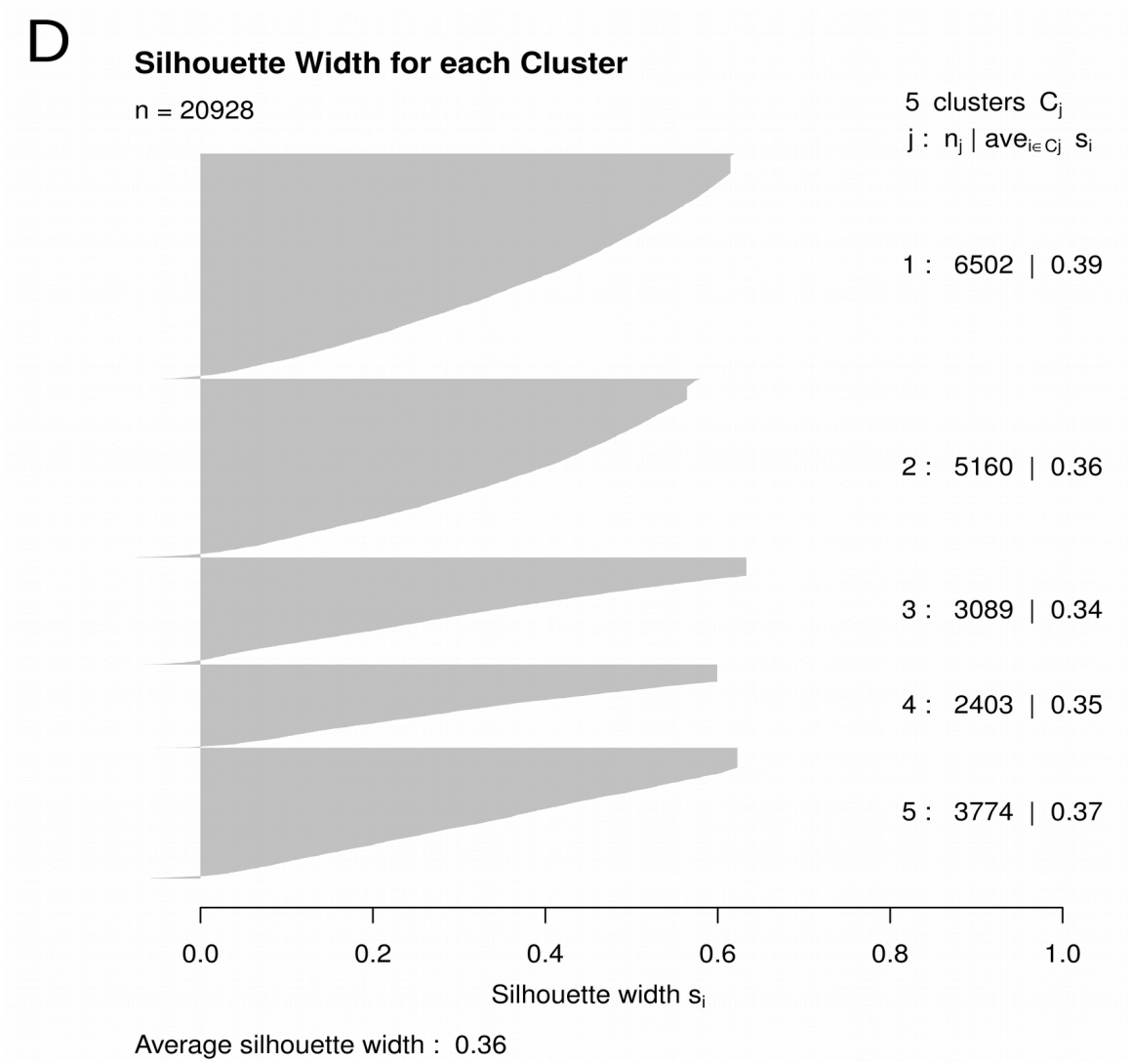
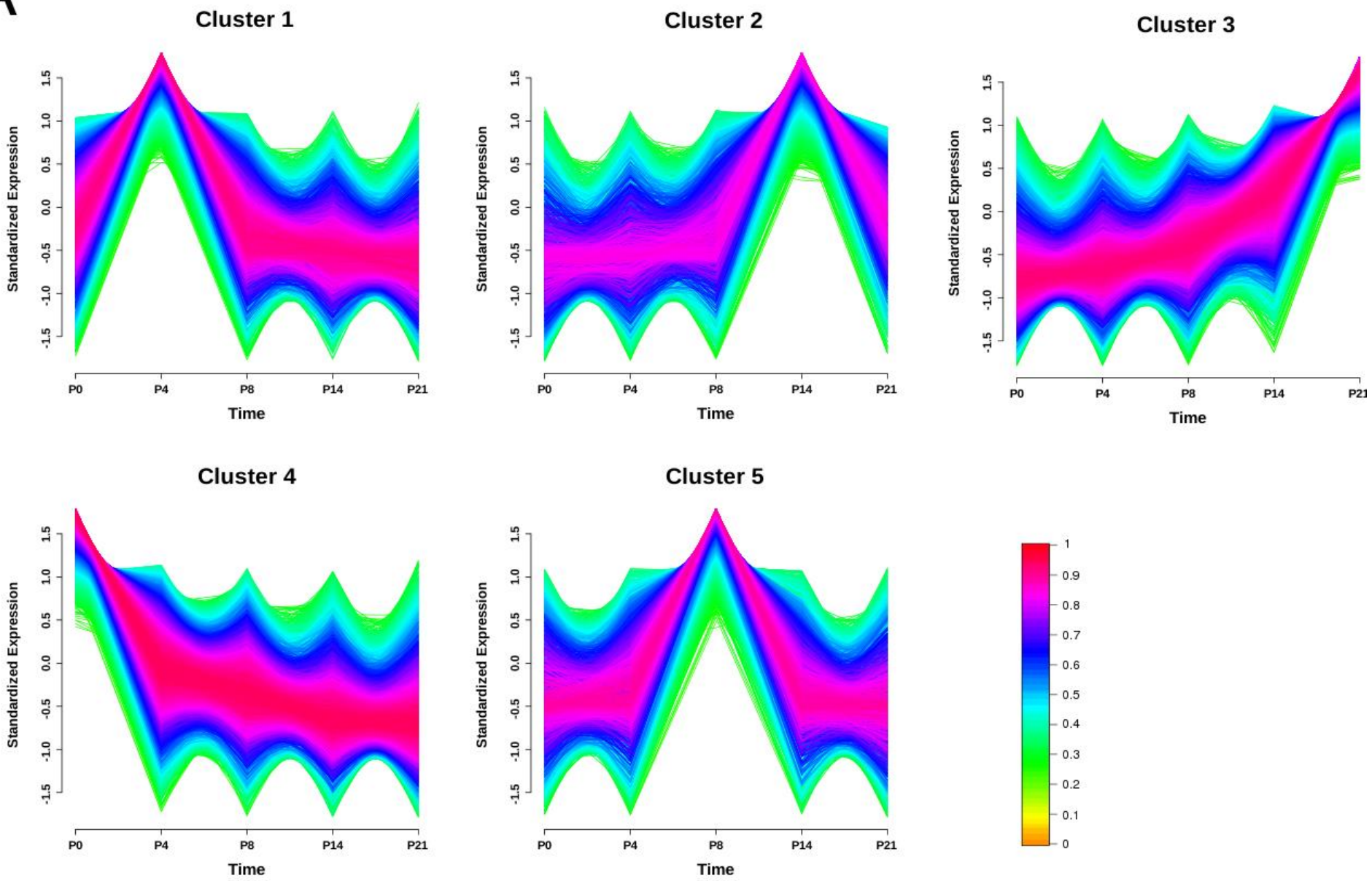


Figure 4.15. Graphical representation of the three methods used to select the number of clusters (K) for fuzzy clustering of the PC expression data. This includes selecting an elbow point (A) between the number of clusters and variance explained, maximizing the Calinski-Harabasz criterion (B), and maximizing the average silhouette width (C). A bar plot of the silhouette width is also displayed (D) for the selected clustering (K=5) for all genes, ordered by cluster and then by silhouette width, and including average silhouette width per cluster and overall average silhouette width.

A


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Standardized Expression
Time
P0
P4
P8
P14
P21
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

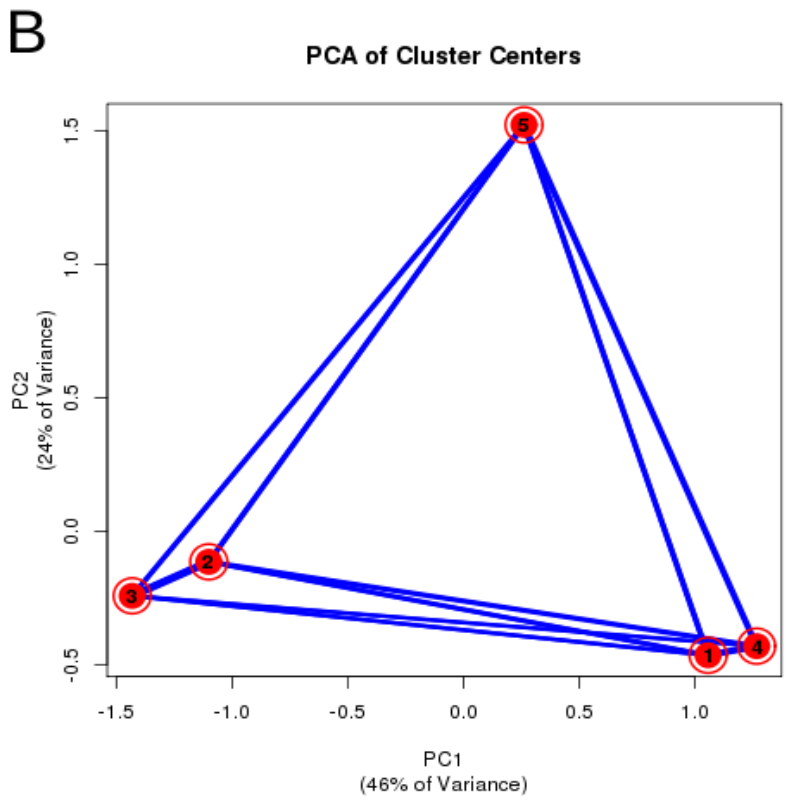


Figure 4.16. Plots displaying information on the fuzzy clustering approach. Panel (A) consists of line graphs displaying standardized gene expression across each of the five clusters identified. Each plot represents a single cluster, and each line represents the expression of a single gene over the course of developmental timing. Expression data was standardized to a mean of zero, and standard deviation of one. Lines are coloured by membership value (as displayed by the included key). Only genes with a membership value > 0.8 were used for subsequent analyses (pink and red lines), to ensure only the genes confidently assigned to the clusters were analysed. Panel (B) is a scatter plot displaying the first two principal components of the cluster centres, providing information on the cluster relationships to one another. The overlap of genes between clusters is represented by connecting lines, where the width represents the strength of overlap.

Table 4.3. Enrichment testing for disorder-related genes amongst the PC clusters defined by fuzzy clustering. For each disorder in each cluster this includes: the enrichment of the disorder gene list in the cluster, the odds ratio of the gene list over the cluster, the p-value (length corrected), and the q-value. Statistically significant results are highlighted in yellow.

Gene-list	Cluster	Enrichment	Odds Ratio	P-value (length corrected)	Q-value (n=30)
Ataxia Human List	cluster 1	0.93	0.92	6.17E-001	1.00E+000
	cluster 2	0.93	1.78	2.75E-001	1.00E+000
	cluster 3	3.67	4.36	2.38E-004	6.71E-003
	cluster 4	0.42	0.40	9.60E-001	1.00E+000
	cluster 5	0.00	NA	1.00E+000	1.00E+000
Ataxia Mouse List	cluster 1	0.00	NA	1.00E+000	1.00E+000
	cluster 2	0.00	NA	1.00E+000	1.00E+000
	cluster 3	5.70	7.63	4.47E-004	6.71E-003
	cluster 4	0.56	0.54	8.43E-001	1.00E+000
	cluster 5	1.67	1.70	4.55E-001	1.00E+000
Autism Human List (S+1-6)	cluster 1	0.72	0.71	8.03E-001	1.00E+000
	cluster 2	0.69	0.68	7.18E-001	1.00E+000
	cluster 3	1.24	1.26	2.94E-001	1.00E+000
	cluster 4	0.83	0.81	8.96E-001	1.00E+000
	cluster 5	1.00	1.00	4.62E-001	1.00E+000
Autism Mouse List	cluster 1	0.78	0.77	7.30E-001	1.00E+000
	cluster 2	0.28	0.27	9.60E-001	1.00E+000
	cluster 3	2.39	2.62	1.15E-003	1.15E-002
	cluster 4	0.63	0.61	9.55E-001	1.00E+000
	cluster 5	0.93	0.93	5.53E-001	1.00E+000
Schizophrenia Human List	cluster 1	0.63	0.62	9.28E-001	1.00E+000
	cluster 2	0.85	0.84	6.93E-001	1.00E+000
	cluster 3	0.91	0.90	6.91E-001	1.00E+000
	cluster 4	1.19	1.22	1.89E-001	1.00E+000
	cluster 5	0.42	0.41	9.74E-001	1.00E+000

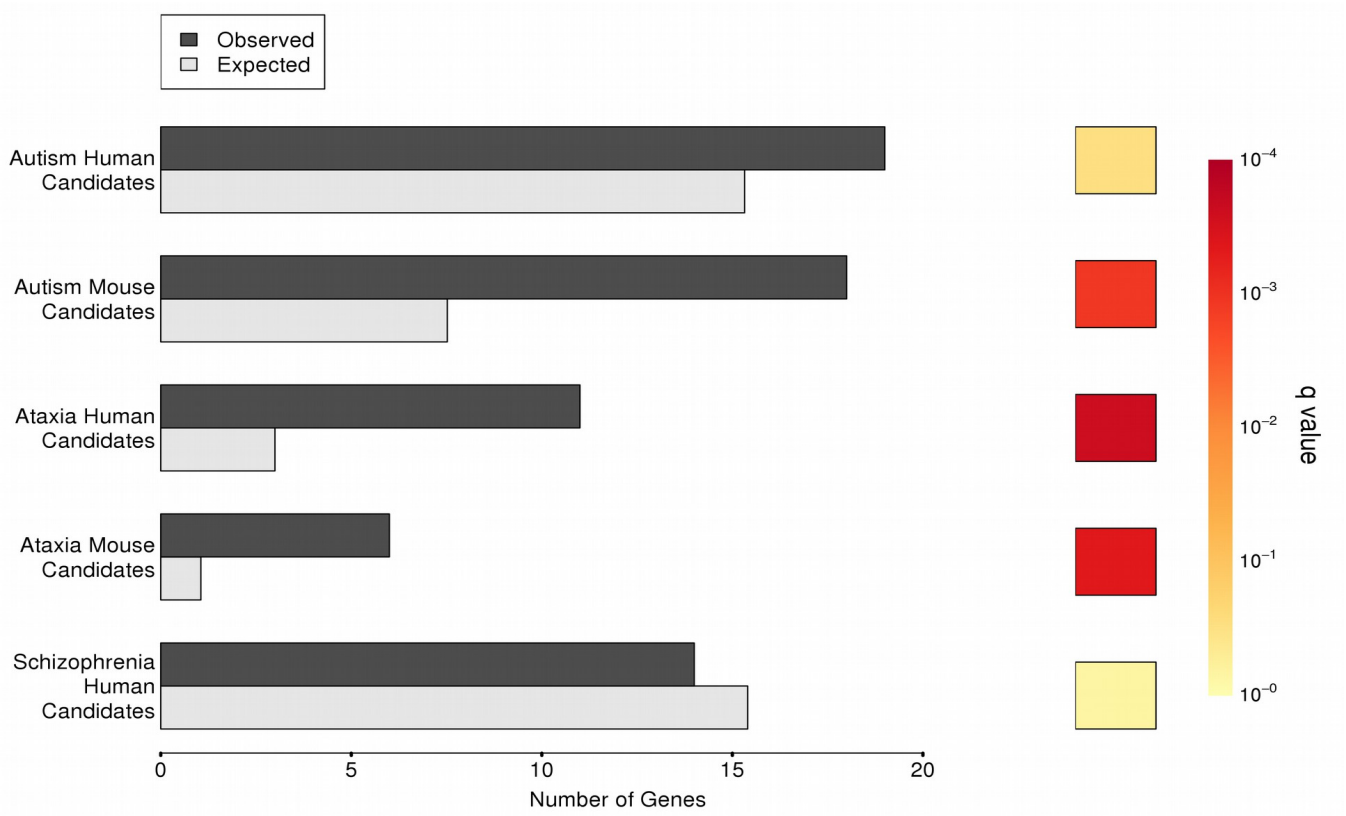


Figure 4.17. A bar plot of disease enrichments observed in the fuzzy cluster positively correlated with developmental timing (cluster 3), including ataxia, autism and schizophrenia associated gene lists (as displayed in Table 4.3). The autism mouse, ataxia human, and ataxia mouse lists all returned significant q-values, reflecting the results observed through WGCNA. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

Table 4.4. Enrichment testing for genes of FMRP-bound transcripts amongst the PC clusters defined by fuzzy clustering. For each cluster this includes: the enrichment of the FMRP-bound transcript genes in the cluster, the odds ratio of the FMRP-bound transcript gene list over the cluster, the p-value (length corrected), and the q-value. Statistically significant results are highlighted in yellow.

Cluster	Enrichment	Odds Ratio	P-value (length corrected)	Q-value (n=30)
cluster 1	0.29	0.27	1.00E+000	1.00E+000
cluster 2	0.10	0.10	1.00E+000	1.00E+000
cluster 3	2.57	3.09	2.17E-014	6.50E-013
cluster 4	0.75	0.72	9.99E-001	1.00E+000
cluster 5	0.09	0.08	1.00E+000	1.00E+000

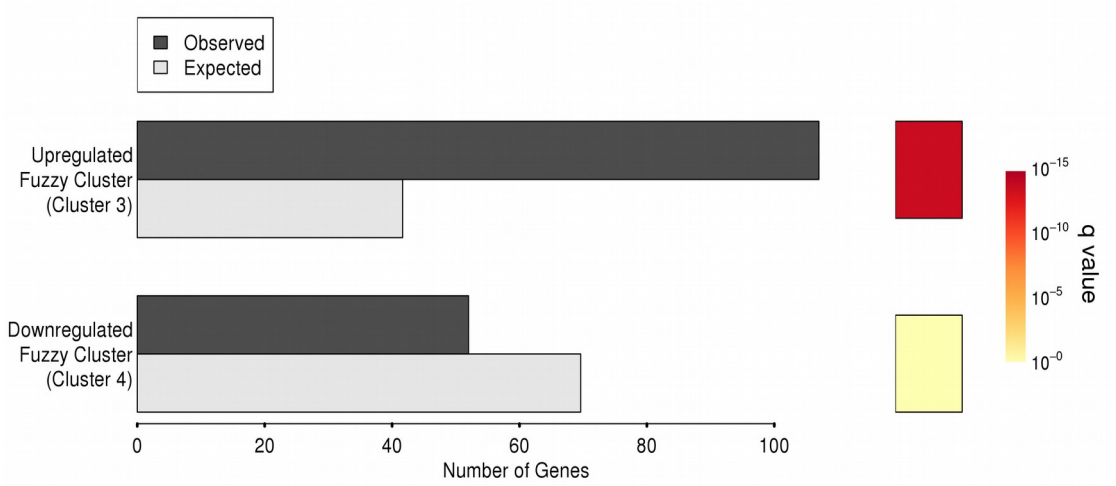


Figure 4.18. A bar plot displaying the enrichment of genes of FMRP-bound transcripts, observed in the fuzzy clusters positively (cluster 3) and negatively (cluster 4) correlated with developmental timing, respectfully (as displayed in Table 4.4). Genes of FMRP-bound transcripts are strongly significantly enriched in the positively correlated cluster, reflecting the results observed through WGCNA. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

To put these findings into perspective, it was important to investigate how similar the results from WGCNA and fuzzy clustering methods were. To address this, the two clusters positively and negatively correlated with developmental time (respectively) were directly compared between methods to determine the extent of their intersect (Figure 4.19). This confirms that the two methods are neither wholly different, nor the same, with approximately half of the smaller of the clusters also occurring in the larger (i.e. half of fuzzy cluster 3 occurs in WGCNA aquamarine cluster, and half of WGCNA blue cluster occurs in fuzzy cluster 4). In spite of the differences between these methods, enrichment results were well replicated, providing confidence in the finding of ataxia and ASD gene enrichment in those cluster positively correlated with developmental time.

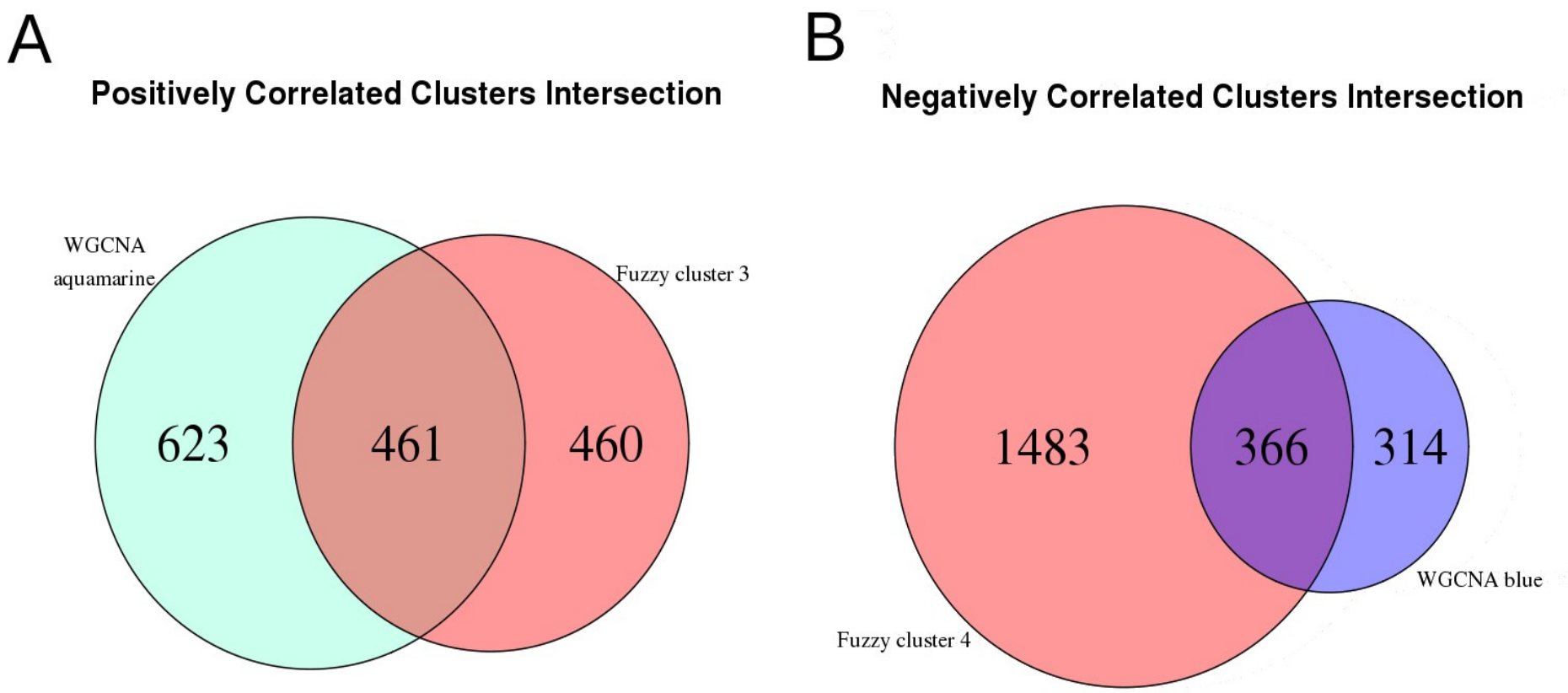


Figure 4.19. Venn diagrams displaying the overlap between the respective clusters correlated with developmental timing and their counterparts in alternative clustering methods, including WGCNA and fuzzy clustering. These are displayed for the positively correlated (A) or negatively correlated (B) clusters.

Differential Expression Analysis

Following positive results for ASD-enrichment in WGCNA or fuzzy-clustering derived modules that are positively correlated with developmental timing, differential expression analysis was implemented to optimize a cluster of this pattern. This was used to obtain a positively correlated cluster (and a negatively correlated cluster) of genes statistically significantly differentially expressed over time (see Methods). This resulted in two large clusters: a positively correlated cluster of 2,004 genes, and a negatively correlated cluster of 2,511 genes (Figure 4.20).

The intersection between these clusters, and clusters of the same pattern but defined by previous methods, were subsequently visualized (Figure 4.21). This showed that either of these differentially expressed clusters encompassed nearly all of the genes in their respective WGCNA counterparts. 1,052 of the 1,084 positively correlated WGCNA cluster genes, and 561 of the 680 negatively correlated WGCNA cluster genes, also occurred in their respective equivalent differentially expressed cluster. In addition, the positively correlated differentially expressed cluster encompassed over two-thirds of its fuzzy cluster counterpart, containing 659 of the 921 fuzzy clustering genes. These results validate the successful production of optimized, and statistically significant, clusters correlated with developmental timing.

A final GO enrichment analysis was implemented once more for these optimized clusters. For the positively correlated cluster, the top significantly enriched GO terms (Figure 4.22A) included similar terms to that of the equivalent WGCNA cluster, including cytoplasm and membrane related. However, they also include further terms related to neurons, such as neuron part, and SNARE/syntaxin binding, which are known to be localized to neuron plasma membranes to play a key role in synaptic vesicle fusion (Rizo & Rosenmund, 2008). Significant MGI phenotypes and KEGG pathways are also displayed (Figure 4.22B), confirming the inclusion of neuronal genes with the long-term potentiation MGI phenotype and the nervous system electrophysiology KEGG pathway. In addition, the KEGG pathways reaffirm what was seen in the WGCNA module, with the ataxia KEGG pathway. Satisfyingly, PC morphology is also seen amongst the KEGG pathways. All of these enrichment terms suggest that this cluster may have included many more genes related to the PC's neuronal development than the previous WGCNA cluster. For the negatively correlated cluster, many terms that were seen before have also been reaffirmed (Figure 4.23), once again displaying many terms related to the ribosome and cell cycle.

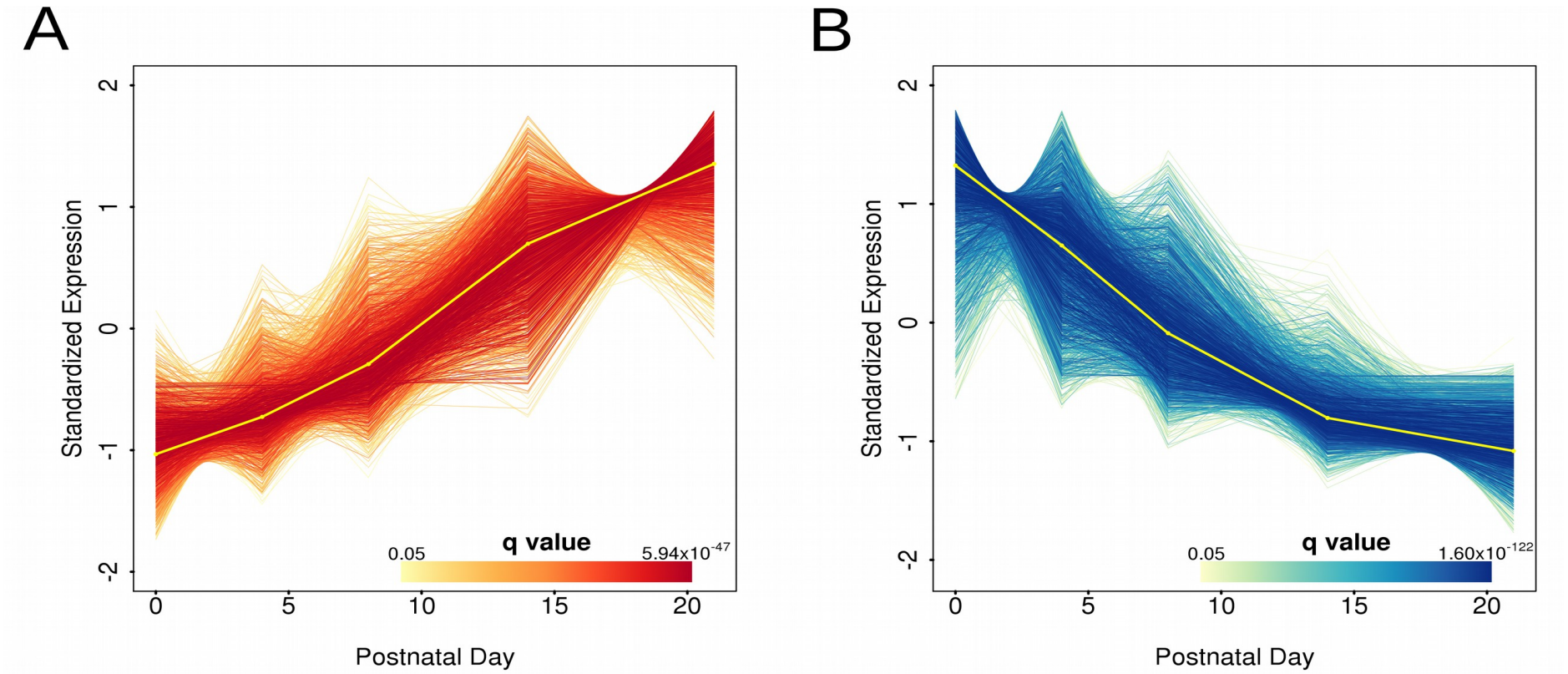


Figure 4.20. Line graphs displaying the differential expression approach to clustering, to form two clusters of significantly differential genes: one positively correlated (A), and one negatively correlated (B), with developmental timing. The threshold for significance was a p-value of 0.05, corrected for FDR. Each line on the plot represents the expression of a single gene over the course of developmental timing. An additional line in yellow has been added to indicate the eigengene of the cluster. Expression data was standardized to a mean of zero, and standard deviation of one.

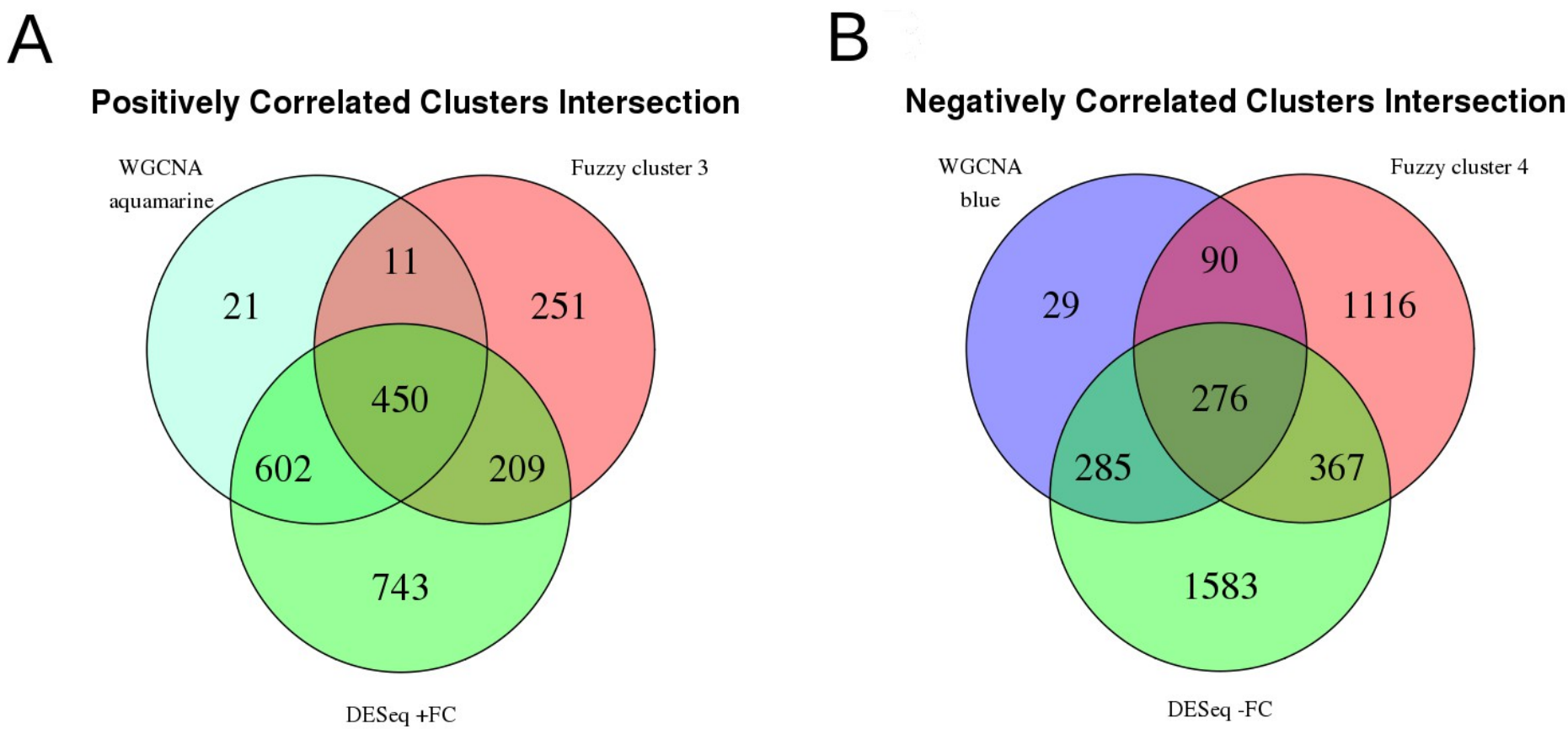


Figure 4.21. Venn diagrams displaying the overlap between the respective clusters correlated with developmental timing and their counterparts in alternative clustering methods, including WGCNA, fuzzy clustering, and differential expression (DESeq). These are displayed for the positively correlated (A) and negatively correlated (B) clusters.

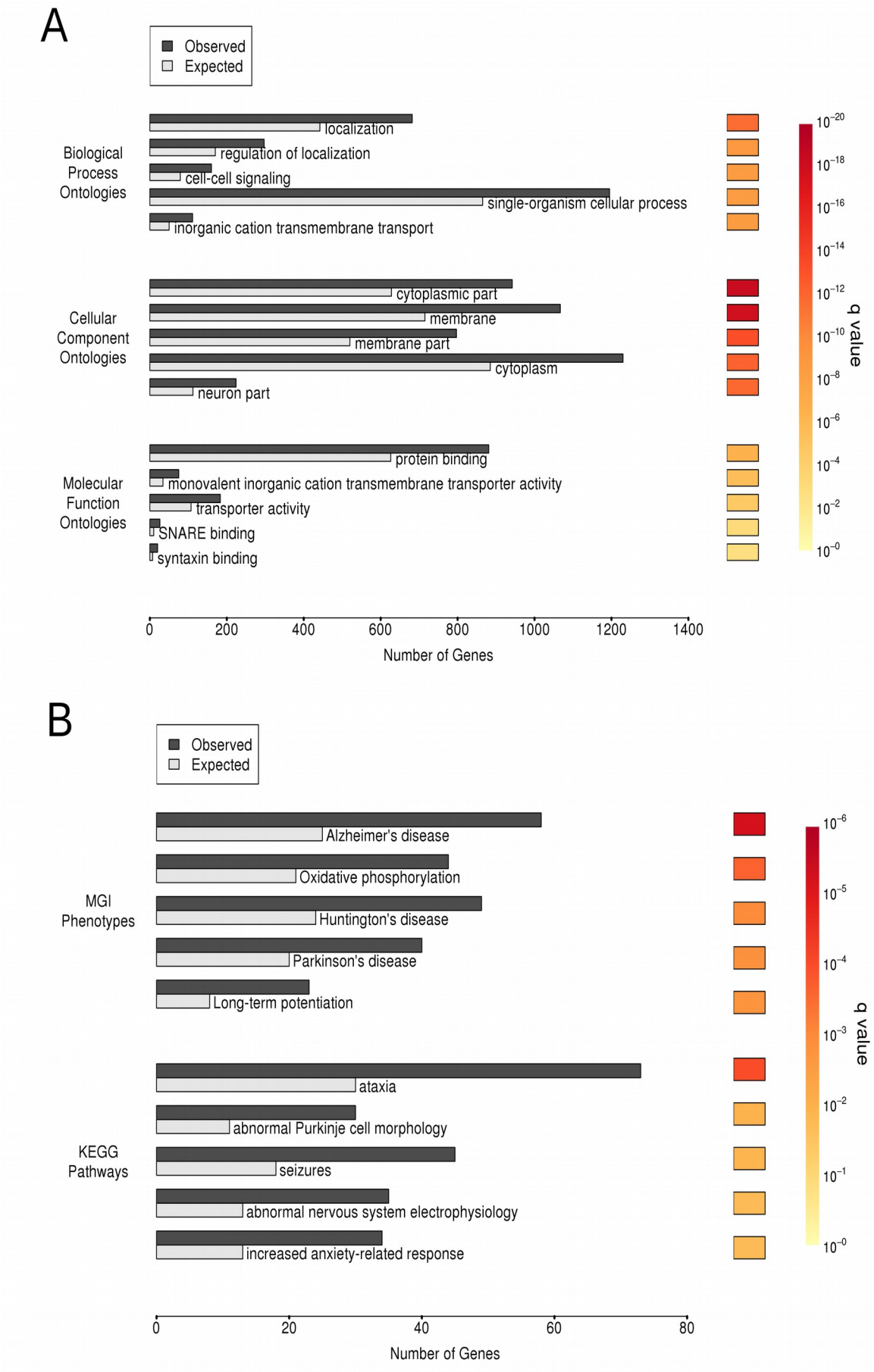

A
Observed
Expected
Biological Process Ontologies
localization
regulation of localization
cell-cell signaling
single-organism cellular process
inorganic cation transmembrane transport
Cellular Component Ontologies
cytoplasmic part
membrane
membrane part
cytoplasm
neuron part
Molecular Function Ontologies
protein binding
monovalent inorganic cation transmembrane transporter activity
transporter activity
SNARE binding
syntaxin binding
0
200
400
600
800
1000
1200
1400
Number of Genes
q value
10^{-20}
10^{-18}
10^{-16}
10^{-14}
10^{-12}
10^{-10}
10^{-8}
10^{-6}
10^{-4}
10^{-2}
10^{-0}
B
Observed
Expected
MGI Phenotypes
Alzheimer's disease
Oxidative phosphorylation
Huntington's disease
Parkinson's disease
Long-term potentiation
KEGG Pathways
ataxia
abnormal Purkinje cell morphology
seizures
abnormal nervous system electrophysiology
increased anxiety-related response
0
20
40
60
80
Number of Genes
q value
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{-0}

Figure 4.22. Bar plots of significant functional annotations for the differentially expressed cluster positively correlated with developmental timing, displaying the five most statistically significant enrichments for each category. This includes the various GO term categories (A) (biological process, cellular component, and molecular function) and phenotype (MGI) and pathway (KEGG) categories (B). Synapse-related GO terms were significantly enriched in all three GO term categories, and phenotype and pathway categories, suggesting that this module is largely associated with synapse formation (specifically in the PCs). Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

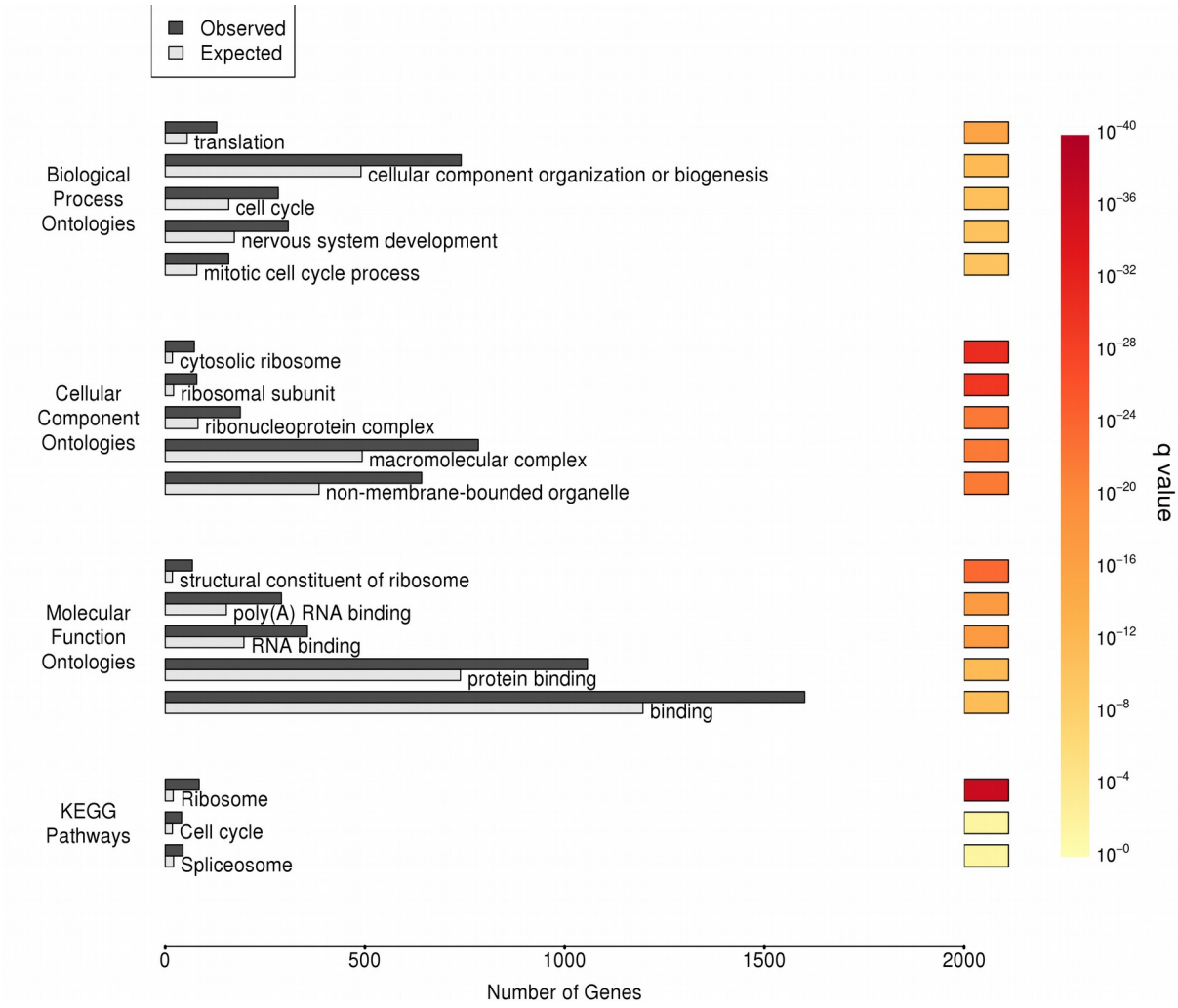


Figure 4.23. Bar plots of significant functional annotations for the differentially expressed cluster negatively correlated with developmental timing, displaying the five most statistically significant enrichments for each category. This includes the various GO term categories (A) (biological process, cellular component, and molecular function) and pathway (KEGG) categories (B). Notably, both ribosome-related and cell cycle-related annotations were returned in both GO terms and KEGG pathways. Only three pathways were significantly enriched, and zero phenotypes. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

As with the previous clustering methods, disease enrichment analyses were implemented on both these differentially expressed clusters (Table 4.5). As expected, the cluster positively correlated with developmental timing returned significant enrichments for many of the disease gene lists (Figure 4.24). As before, the ataxia lists both had high fold-enrichment values, with the human list at 2.7-fold and a q-value of $9.70x10^{-5}$, and with the mouse list at 5.6-fold and a q-value of $6.98x10^{-8}$ (a p-value considerably lower than observed in previous equivalent clusters of different methods). Also as before, the mouse ASD gene list had a significant enrichment of 1.9-fold, resulting in a q-value of $2.24x10^{-3}$. In addition to this, the human ASD gene list now returned a statistically significant result, with a q-value of $6.44x10^{-3}$. The same test for human schizophrenia-associated genes was also applied and, as before, displayed no noteworthy enrichment (1.0-fold). The negatively correlated cluster also returned no significant enrichment for any of the disease gene lists, as expected.

Finally, supplementary enrichment testing of the FMRP-bound transcript gene list was employed once more (Table 4.6). Unlike before, for this method of clustering the FMRP-bound transcript gene list was significantly enriched in both the positively and negatively correlated clusters (Figure 4.25). This was likely due to the increased inclusion of neuronal development genes in both clusters, and included a very strong enrichment for the positively correlated cluster, with a 2.3-fold enrichment and q-value of $4.76x10^{-30}$, and a weaker enrichment in the negatively

correlated cluster, with a 1.4-fold enrichment and a q-value of 1.75×10^{-6}.

To further investigate the enrichment of the human SFARI gene list, the gene list was stratified based on confidence. The SFARI resource annotates genes by degrees of confidence in their associations with ASD, including categories from 1-6 (decreasing in confidence) and information on whether they are associated with syndromic forms of ASD. By removing many of the redundant genes, it was discovered that the enrichment observed could be strengthened (Figure 4.26). Firstly, with syndromic genes removed, the gene list is 1.5-fold enriched. Further reduction of the gene list, removing each of the lower confidence categories in turn, strengthens the significance. This peaks at categories 1-4, with a 1.7-fold enrichment. Very few genes are considered to be of high confidence, and further removal of categories caused a large reduction in the size of the gene list. This may be the reason for loss of significance for just the highest confidence genes.

Table 4.5. Enrichment testing for disorder-related genes amongst the PC clusters defined by differential expression. For each disorder in each cluster this includes: the enrichment of the disorder gene list in the cluster, the odds ratio of the gene list over the cluster, the p-value (length corrected), and the q-value. Statistically significant results are highlighted in yellow.

Gene-list	Cluster	Enrichment	Odds Ratio	P-value (length corrected)	Q-value (n=12)
Ataxia Human List	-ve fold change	0.37	0.33	9.94E-001	1.00E+000
	+ve fold change	2.74	3.71	1.62E-005	9.70E-005
Ataxia Mouse List	-ve fold change	0.68	0.64	8.18E-001	1.00E+000
	+ve fold change	5.60	14.91	5.82E-009	6.98E-008
Autism Human List (S + 1-6)	-ve fold change	1.14	1.17	1.34E-001	3.06E-001
	+ve fold change	1.42	1.53	6.44E-003	1.93E-002
Autism Mouse List	-ve fold change	1.09	1.11	4.13E-001	6.42E-001
	+ve fold change	1.85	2.11	5.59E-004	2.24E-003
Schizophrenia Human List	-ve fold change	1.16	1.20	1.53E-001	3.06E-001
	+ve fold change	1.04	1.05	4.28E-001	6.42E-001

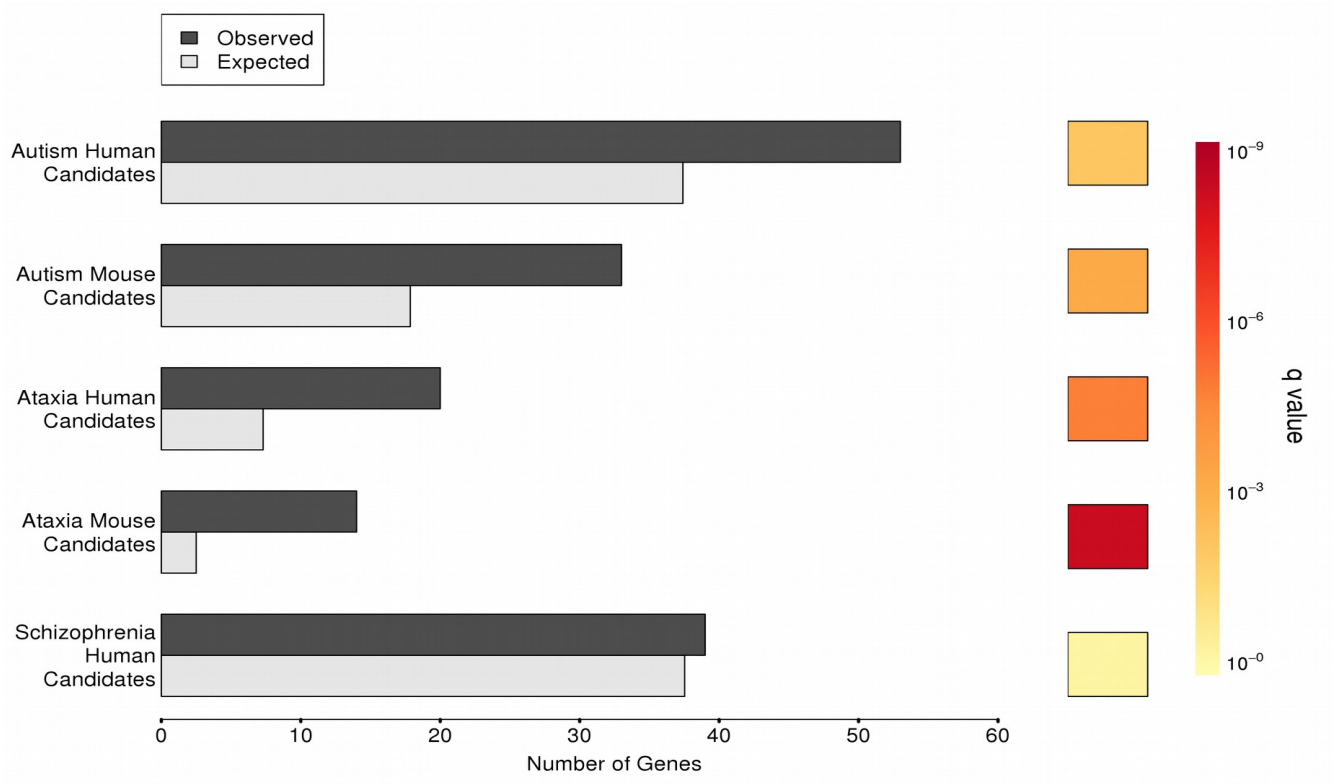


Figure 4.24. A bar plot of disease enrichments observed in the differentially expressed cluster positively correlated with developmental timing, including ataxia, autism and schizophrenia associated gene lists (as in Table 4.5). The autism human, autism mouse, ataxia human, and ataxia mouse lists all returned significant q-values, reflecting most of the results observed through WGCNA and fuzzy clustering (with the additional significant enrichment of the autism human list). Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

Table 4.6. Enrichment testing for genes of FMRP-bound transcripts amongst the PC clusters defined by differential expression (DE-Seq2). For each cluster this includes: the enrichment of the FMRP-bound transcript genes in the cluster, the odds ratio of the FMRP-bound transcript gene list over the cluster, the p-value (length corrected), and the q-value. Statistically significant results are highlighted in yellow.

Cluster	Enrichment	Odds Ratio	P-value (length corrected)	Q-value (n=12)
-ve fold change	1.36	1.49	2.91E-007	1.75E-006
+ve fold change	2.30	3.05	3.97E-031	4.76E-030

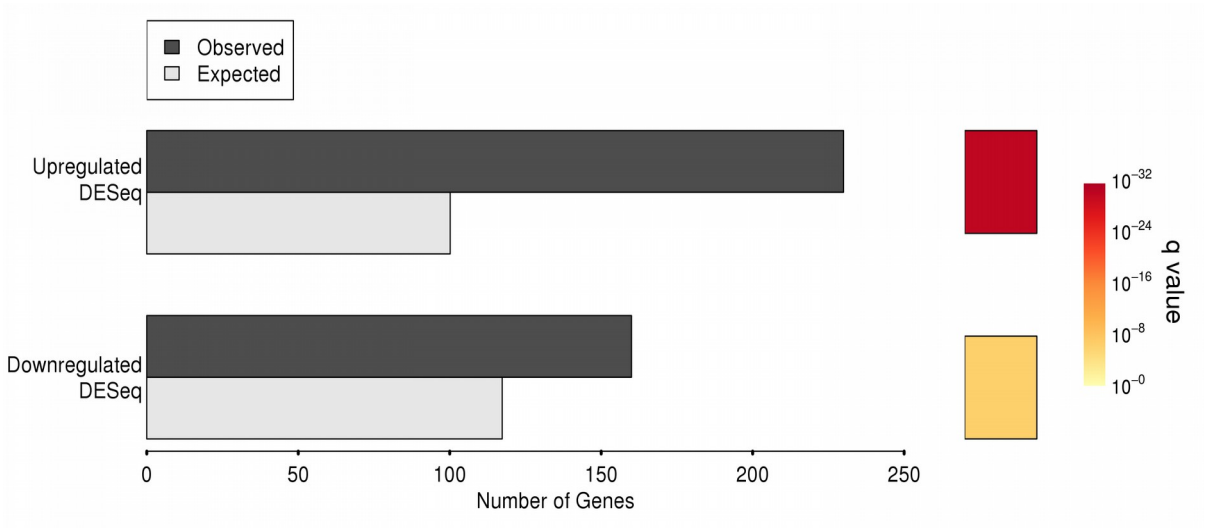


Figure 4.25. A bar plot displaying the enrichment of genes of FMRP-bound transcripts, observed in the differentially expressed clusters (as displayed in Table 4.6). Unlike with the previous clustering approaches demonstrated, the FMRP-bound transcript gene list is strongly significantly enriched in both of these clusters. Nevertheless, the trend of increased enrichment in the cluster positively correlated with developmental timing (compared to the negatively correlated) remains. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

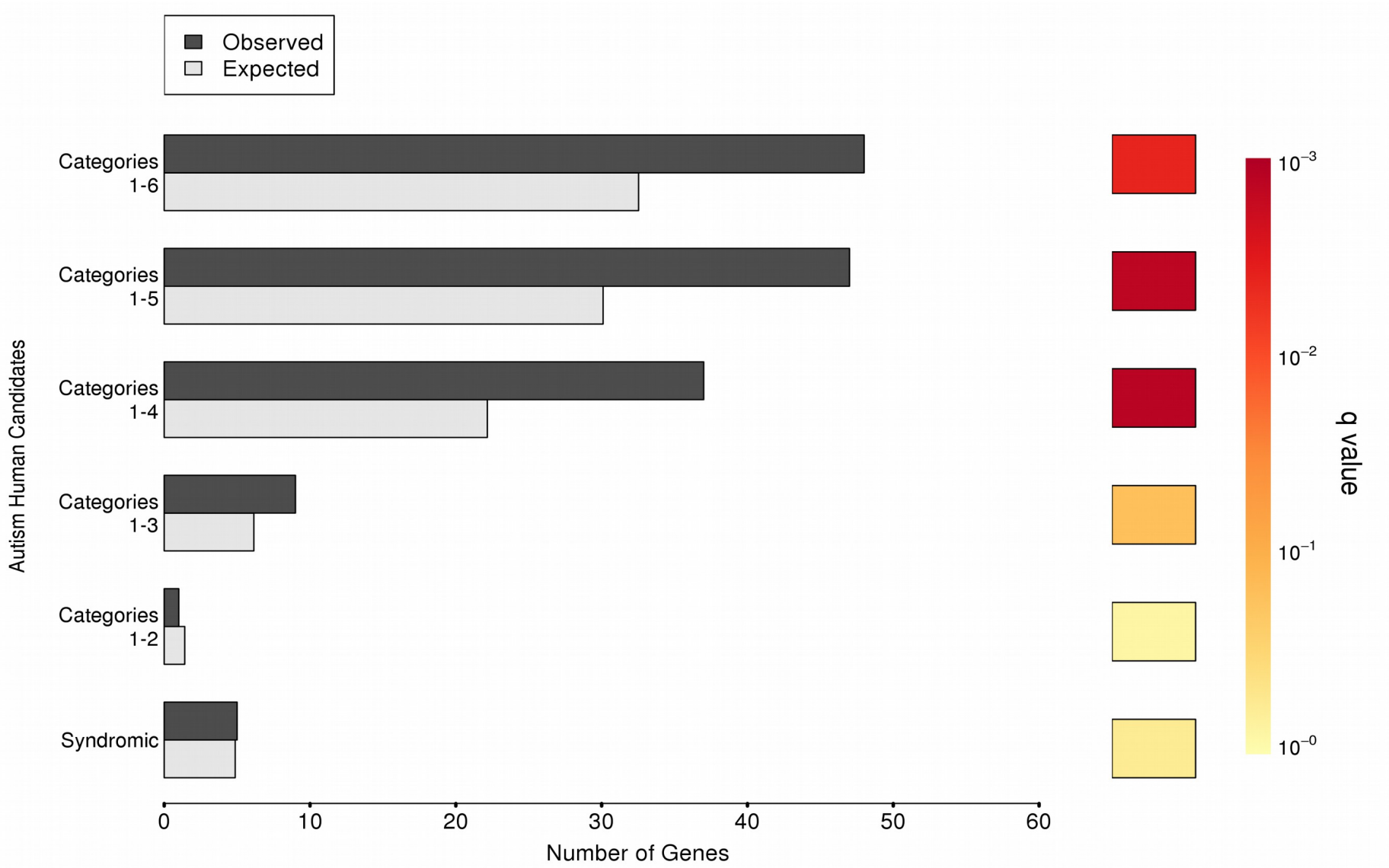

Observed
Expected
Categories 1-6
Categories 1-5
Categories 1-4
Categories 1-3
Categories 1-2
Syndromic
Autism Human Candidates
0
10
20
30
40
50
60
Number of Genes
10^{-3}
10^{-2}
10^{-1}
10^{-0}
q value

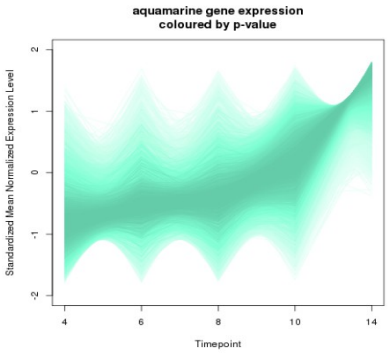
Figure 4.26. A bar plot based on the enrichment of the SFARI human autism-associated gene list in the differentially expressed cluster positively correlated with developmental timing, displaying varying results based on stratification of the gene list. From top to bottom, this displays enrichment testing with SFARI categories of least confidence removed. SFARI categories 5 & 6 include respective hypothesized but untested genes and genes with evidence against a role in autism; categories 3 & 4 include genes with suggestive and minimal evidence respectively; categories 1 & 2 include high confidence and strong candidate genes, respectively. A category of genes predisposing autism in the context of syndromic disorders is also included. Genes from categories 1-4 exhibit the greatest significance (with a q-value of 4.1×10^{-3}). Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

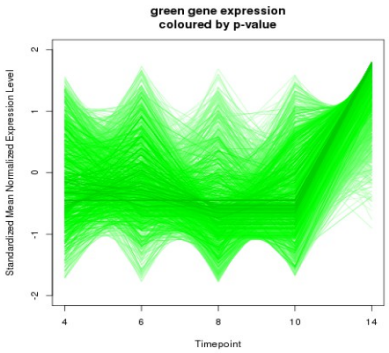
Comparative Analysis across Brain Regions

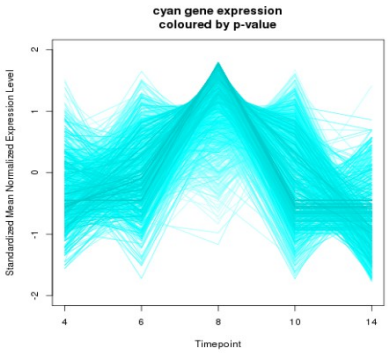
As was initially implemented for the PC data analysis, WGCNA was applied to the neocortex data, forming 12 gene clusters (Figures 27, S2 and S3). Only one of these clusters was significantly enriched for the ASD-associated gene lists and, in contrast to the gene cluster identified in the developing PCs, this ASD-enriched neocortex cluster does not show positive differential expression over development, but is patterned in an arc (Figure 4.28A). This was a significant enrichment observed in both the human candidate and mouse model ASD gene lists (Table 4.7 and Figure 4.29), with a 2.9-fold enrichment for human SFARI ASD-associated genes (categories 1-4 and syndromic) and a q-value of 3.09×10^{-4}. Similarly, the significant enrichment for mouse SFARI ASD-associated genes was due to a 3.0-fold enrichment, with a q-value of 0.01. Of the ataxia and schizophrenia lists, there was one significant enrichment observed, which was for the mouse model ataxia list, but enriched in the cluster positively correlated with developmental timepoint (the same patterning as that in the PC data), with a 3.2-fold enrichment and q-value of 4.93×10^{-3}. Once again, enrichment for genes of FMRP-bound transcripts was tested and, as with the PC data, there was significant enrichment in the ASD-enriched cluster (Table 4.8). This was observed with a fold-enrichment of 3.7 and q-value of 2.67×10^{-8}. However, for the neocortex, there was a second significant enrichment identified, in the cluster positively correlated with developmental timing, with a 1.9-fold enrichment and 3.39×10^{-19} q-value. There was a third cluster showing signs of enrichment, which is a cluster of very

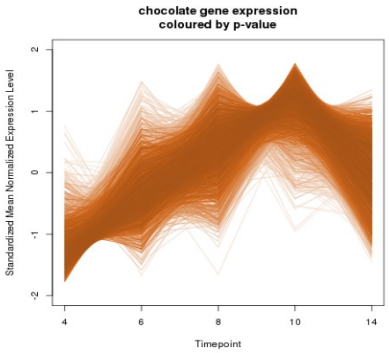
similar temporal expression pattern to the ASD-enriched cluster of the neocortex (the darkviolet cluster; arcing over development), with a 3.9-fold change and p-value of 0.04, however this was not robust enough for significance following multiple testing correction (q-value of 0.91).

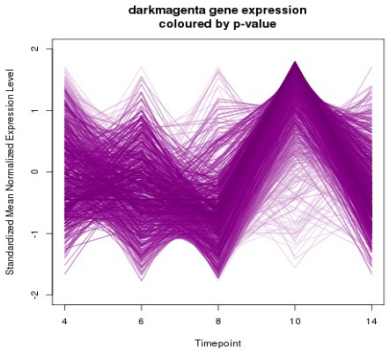
The ASD-enriched modules in both brain regions were then cross-examined to explore further contrasting features. Firstly, the genes in each cluster showed low coherency when investigating their expression in the opposite brain region, i.e. the genes in the PC cluster did not form a consistent cluster in the neocortex, and vice versa (Figure 4.28B) (although this is not necessarily unexpected; Figure A4). In this same vein, the degree of overlap between the ASD-associated genes in the two clusters was also quantified with an intersection analysis (Figure 4.28C). Of the 182 genes in the human and mouse SFARI gene lists, 59 were present in the PC ASD-enriched cluster, while 39 were present in the neocortex equivalent. Random sampling determined an expected overlap of eight genes across the two; however, only a single gene (*Maoa*) was present in both clusters. This is a statistically significant low degree of overlap, as determined by Fisher's exact testing, with a p-value of 5.8×10^{-4}. These observations across the two clusters suggest they are not maintained across the neuronal subpopulations, with distinctive ASD-associated gene enrichments.

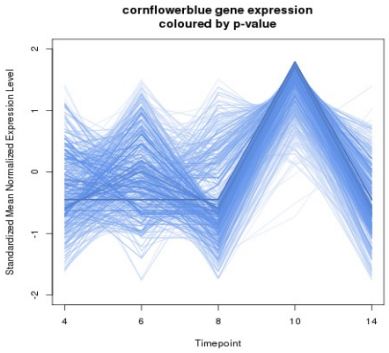

aquamarine gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint

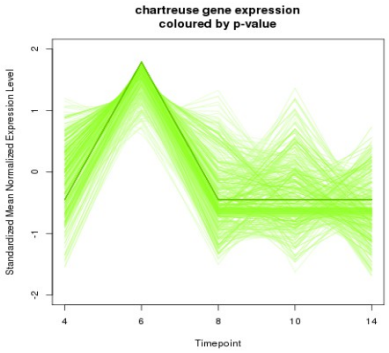

green gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint

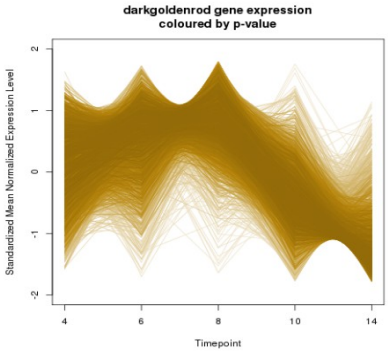

cyan gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint

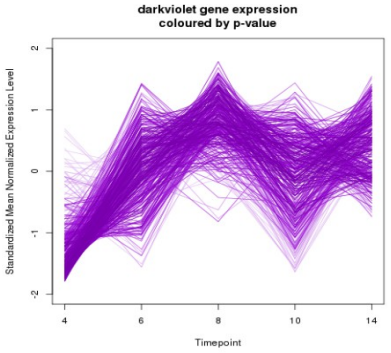

chocolate gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint

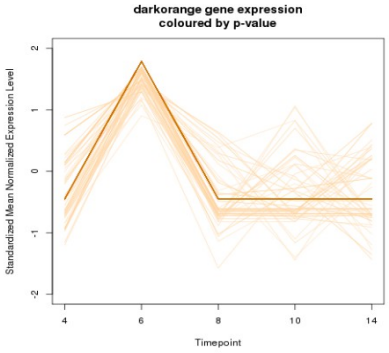

darkmagenta gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint

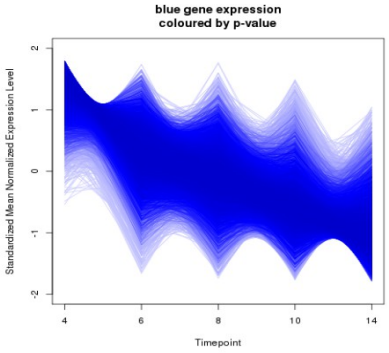

cornflowerblue gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint


chartreuse gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint


darkgoldenrod gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint


darkviolet gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint


darkorange gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint


blue gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint

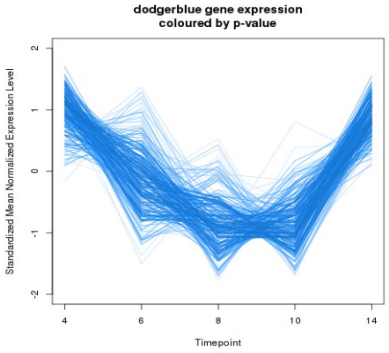

dodgerblue gene expression
coloured by p-value
Standardized Mean Normalized Expression Level
Timepoint

Figure 4.27. Line graphs displaying standardized gene expression across each of the 12 clusters identified in the neocortex data through WGCNA. Each plot represents a single cluster, and each line represents the expression of a single gene over the course of developmental timing. Expression data was standardized to a mean of zero, and standard deviation of one.

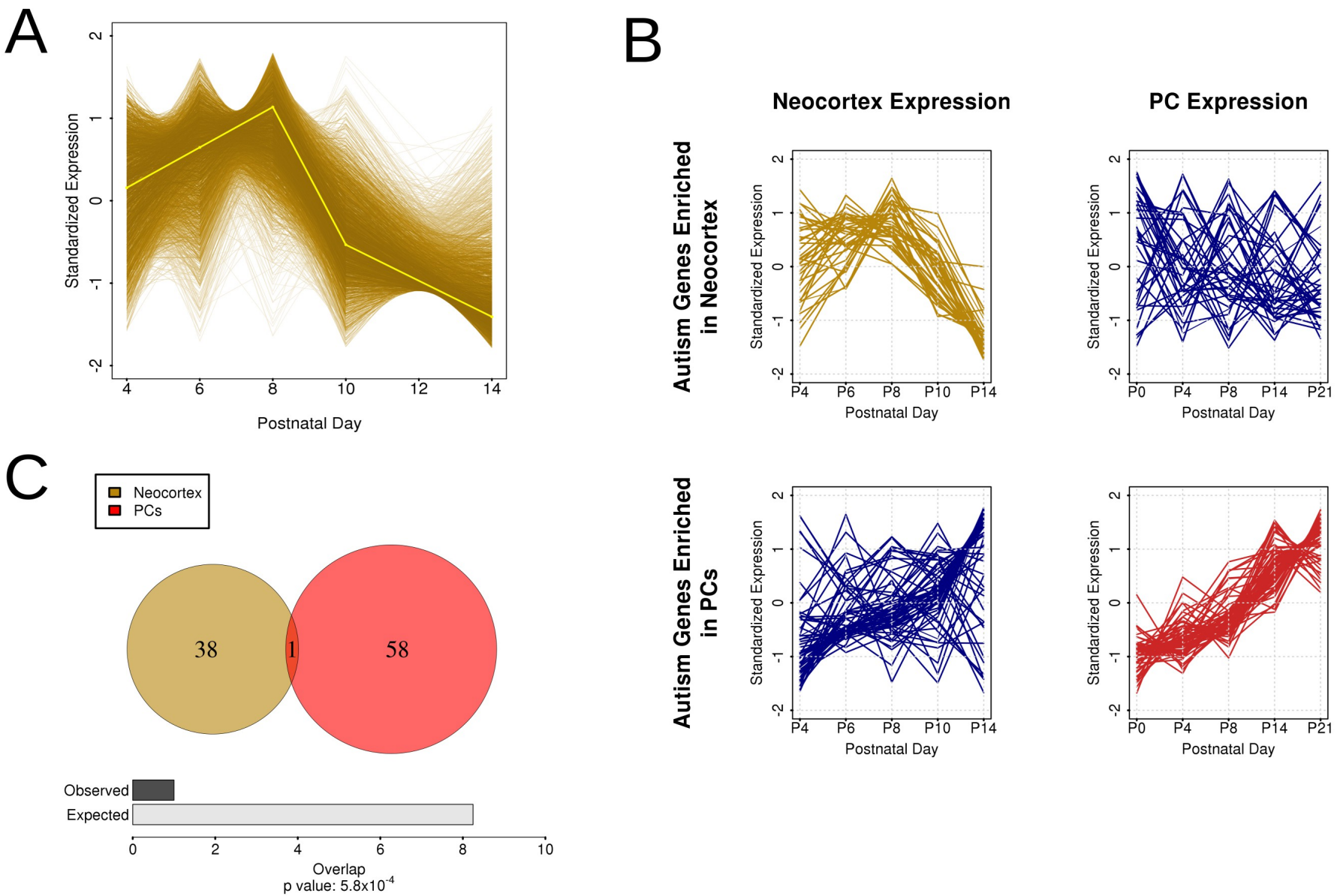

A
Standardized Expression
Postnatal Day
B
Neocortex Expression
PC Expression
Autism Genes Enriched in Neocortex
Autism Genes Enriched in PCs
Standardized Expression
Postnatal Day
C
Neocortex
PCs
38
1
58
Observed
Expected
Overlap
p value: 5.8x10-4

Figure 4.28. Plots displaying differential clustering between the PC and neocortex data. Panel (A) displays a line graph of the pattern of the ASD-enriched cluster in the neocortex (as identified by WGCNA). Each line on the plot represents the expression of a single gene over the course of developmental timing. An additional line in yellow has been added to indicate the eigengene of the cluster. Expression data was standardized to a mean of zero, and standard deviation of one. Panel (B) displays the expression patterns of both ASD-enriched clusters for the PC and neocortex data, and the expression patterns for these same genes in the opposing tissue, showing a lack of coherency across tissues. Each line on the plot represents the expression of a single gene over the course of developmental timing. Expression data was standardized to a mean of zero, and standard deviation of one. Panel (C) displays a Venn diagram of the significant low degree of overlap between the 182 autism-associated genes (human and mouse) within the PC and neocortex clusters. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

Table 4.7. Enrichment testing for disorder-related genes amongst the neocortex clusters defined by WGCNA. For each disorder in each cluster this includes: the enrichment of the disorder gene list in the cluster, the odds ratio of the gene list over the cluster, the p-value (length corrected), and the q-value. Statistically significant results are highlighted in yellow.

Gene-list	Cluster	Enrichment	Odds Ratio	P-value (length corrected)	Q-value (n=72)
Ataxia Human List	aquamarine	1.34	1.45	1.49E-001	1.00E+000
	blue	0.62	0.58	9.12E-001	1.00E+000
	chartreuse	0.00	NA	1.00E+000	1.00E+000
	chocolate	0.51	0.50	8.80E-001	1.00E+000
	cornflowerblue	0.00	NA	1.00E+000	1.00E+000
	cyan	0.00	NA	1.00E+000	1.00E+000
	darkgoldenrod	2.00	2.13	2.14E-001	1.00E+000
	darkmagenta	0.00	NA	1.00E+000	1.00E+000
	darkorange	0.00	NA	1.00E+000	1.00E+000
	darkviolet	1.52	1.54	6.11E-001	1.00E+000
	dodgerblue	0.00	NA	1.00E+000	1.00E+000
	green	1.15	1.15	5.40E-001	1.00E+000
Ataxia Mouse List	aquamarine	3.24	5.73	1.37E-004	4.93E-003
	blue	0.37	0.34	9.43E-001	1.00E+000
	chartreuse	0.00	NA	1.00E+000	1.00E+000
	chocolate	0.00	NA	1.00E+000	1.00E+000
	cornflowerblue	0.00	NA	1.00E+000	1.00E+000
	cyan	0.00	NA	1.00E+000	1.00E+000
	darkgoldenrod	1.03	1.03	6.53E-001	1.00E+000
	darkmagenta	0.00	NA	1.00E+000	1.00E+000
	darkorange	0.00	NA	1.00E+000	1.00E+000
	darkviolet	4.81	5.02	1.96E-001	1.00E+000
	dodgerblue	0.00	NA	1.00E+000	1.00E+000
	green	0.00	NA	1.00E+000	1.00E+000
Autism Human List (S+1-6)	aquamarine	1.09	1.12	2.98E-001	1.00E+000
	blue	0.72	0.68	9.58E-001	1.00E+000
	chartreuse	0.46	0.45	8.67E-001	1.00E+000
	chocolate	0.54	0.53	9.54E-001	1.00E+000
	cornflowerblue	0.68	0.67	7.50E-001	1.00E+000
	cyan	0.00	NA	1.00E+000	1.00E+000
	darkgoldenrod	2.91	3.35	4.30E-006	3.09E-004
	darkmagenta	0.51	0.50	8.32E-001	1.00E+000
	darkorange	0.00	NA	1.00E+000	1.00E+000
	darkviolet	2.02	2.07	2.23E-001	1.00E+000
	dodgerblue	0.00	NA	1.00E+000	1.00E+000
	green	0.61	0.60	8.06E-001	1.00E+000
Autism Mouse List	aquamarine	1.32	1.41	6.96E-002	8.36E-001
	blue	0.69	0.65	9.56E-001	1.00E+000
	chartreuse	0.00	NA	1.00E+000	1.00E+000
	chocolate	0.61	0.60	9.02E-001	1.00E+000
	cornflowerblue	0.00	NA	1.00E+000	1.00E+000
	cyan	0.00	NA	1.00E+000	1.00E+000
	darkgoldenrod	2.98	3.36	4.27E-004	1.03E-002
	darkmagenta	0.00	NA	1.00E+000	1.00E+000
	darkorange	0.00	NA	1.00E+000	1.00E+000
	darkviolet	2.52	2.60	1.84E-001	1.00E+000
	dodgerblue	0.00	NA	1.00E+000	1.00E+000
	green	0.33	0.32	9.30E-001	1.00E+000
Schizophrenia Human List	aquamarine	1.04	1.05	3.92E-001	1.00E+000
	blue	1.10	1.11	2.75E-001	1.00E+000
	chartreuse	0.00	NA	1.00E+000	1.00E+000
	chocolate	0.90	0.90	6.78E-001	1.00E+000
	cornflowerblue	1.52	1.54	3.10E-001	1.00E+000
	cyan	0.34	0.34	9.46E-001	1.00E+000
	darkgoldenrod	1.32	1.35	1.80E-001	1.00E+000
	darkmagenta	1.14	1.14	4.78E-001	1.00E+000
	darkorange	0.00	NA	1.00E+000	1.00E+000
	darkviolet	2.11	2.18	5.75E-002	8.28E-001
	dodgerblue	1.91	1.95	2.77E-001	1.00E+000
	green	0.45	0.44	9.33E-001	1.00E+000

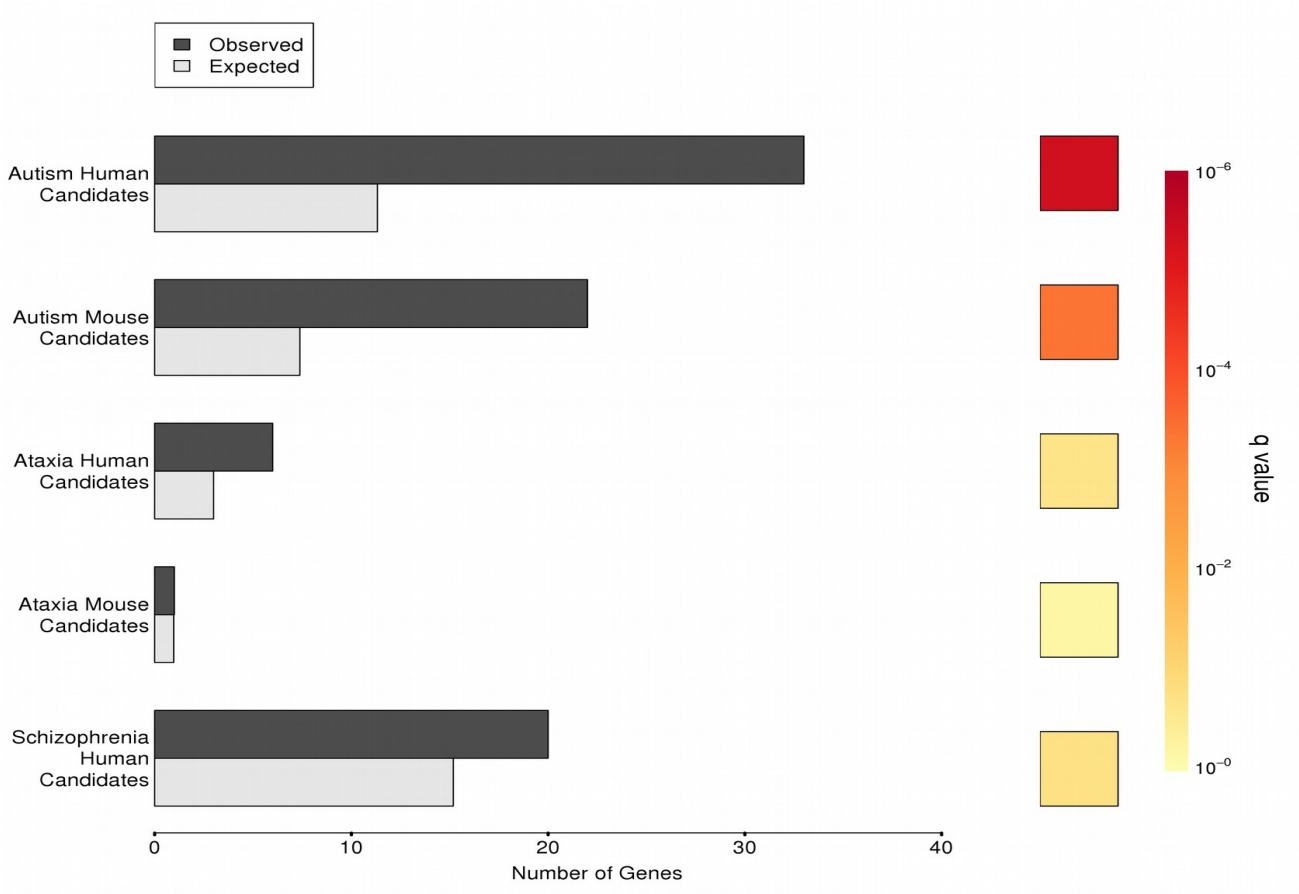


Figure 4.29. A bar plot of disease enrichments observed in the ASD-enriched WGCNA cluster of the neocortex, including ataxia, autism and schizophrenia associated gene lists (as displayed in Table 4.7). The autism human and autism mouse lists both returned significant q-values. Bars represent the number of genes identified, allowing comparison of observed and expected numbers. The q-values are displayed via Hinton plots.

Table 4.8. Enrichment testing for genes of FMRP-bound transcripts amongst the neocortex clusters defined by WGCNA. For each cluster this includes: the enrichment of the FMRP-bound transcript genes in the cluster, the odds ratio of the FMRP-bound transcript gene list over the cluster, the p-value (length corrected), and the q-value. Statistically significant results are highlighted in yellow.

Cluster	Enrichment	Odds Ratio	P-value (length corrected)	Q-value (n=72)
aquamarine	1.85	2.35	4.71E-021	3.39E-019
blue	0.79	0.76	9.99E-001	1.00E+000
chartreuse	0.10	0.09	9.93E-001	1.00E+000
chocolate	0.85	0.84	9.84E-001	1.00E+000
cornflowerblue	0.14	0.13	9.83E-001	1.00E+000
cyan	0.00	NA	1.00E+000	1.00E+000
darkgoldenrod	3.66	4.91	7.42E-010	2.67E-008
darkmagenta	0.11	0.11	9.65E-001	1.00E+000
darkorange	0.12	0.11	9.64E-001	1.00E+000
darkviolet	3.92	4.75	3.81E-002	9.14E-001
dodgerblue	0.00	NA	1.00E+000	1.00E+000
green	0.20	0.19	9.94E-001	1.00E+000

Permutation Testing

A sanity check was performed to ensure correct algorithms and methods were being applied to the permutation testing. This was executed on a single example contingency table as follows:

	A1	**A2**
B1	4	1
B2	2	3

The sums of the values in each row (5 and 5) and in each column (6 and 4) were then applied to the algorithms for simulation of similar contingency tables (as described in the Methods), and an empirical p-value was then calculated for the above configuration compared to 10^8 simulations. The empirical p-value produced was 0.5239844. This was approximated well the Fisher's exact p-value of 0.5238095.

The same methods were then implemented using the four contingency tables for the biological data of this study (protein-coding 1:1-orthologous expressed genes only), which were as follows:

	ASD-associated genes (ID only)	Not ASD-associated genes (ID only)
Neocortex cluster genes	20	744
Not neocortex cluster genes	72	14,057

	ASD-associated genes (Non-ID only)	Not ASD-associated genes (Non-ID only)
Neocortex cluster genes	13	751
Not neocortex cluster genes	144	13,985

	ASD-associated genes (ID only)	Not ASD-associated genes (ID only)
PC cluster genes	11	1,770
Not PC cluster genes	81	11,901

	ASD-associated genes (Non-ID only)	Not ASD-associated genes (Non-ID only)
PC cluster genes	30	1,751
Not PC cluster genes	127	11,855

Analysis of the odds ratios of these contingency tables revealed a pattern suggesting that the PC gene cluster contained a significantly greater proportion of ID-free ASD-associated genes when compared with the neocortex gene cluster, for which this pattern is inverted, containing a greater proportion of ID-associated ASD genes. The resultant p-value for this comparison was 4.2×10^{-2} (Figure 4.30). By repeating this analysis on SFARI category 1-4 (non-syndromic) genes, and the SFARI syndromic genes separately, this inversion of patterns was found to be significantly pronounced in the syndromic genes (Figure 4.31). The q-value for permutation testing of the syndromic genes only was 6.8×10^{-4} (Figure 4.32), with the testing for categories 1-4 only returning an insignificant q-value (0.24), although the trend was still maintained.

This testing of the four contingency tables could also be implemented in a single test, if we reduce the key information down to one contingency table (although this causes a loss of information on genes outside of the clusters and ASD gene lists, and so ignores differences in numbers of genes expressed in each tissue, and differences in cluster size). For this, a standard Fisher's exact test returns a p-value of 0.0046 (including categories 1-4 and syndromic genes):

	ID ASD-associated	**Non-ID ASD-associated**
Neocortex cluster	20	13
PC cluster	11	30

Overall, all these results indicate a significant difference in the composition of ASD-associated genes underlying the clusters of each of the brain regions in relation to ID. These genetic differences are consistent with the existence of two genetically and spatiotemporally distinct endophenotypes of ASD.

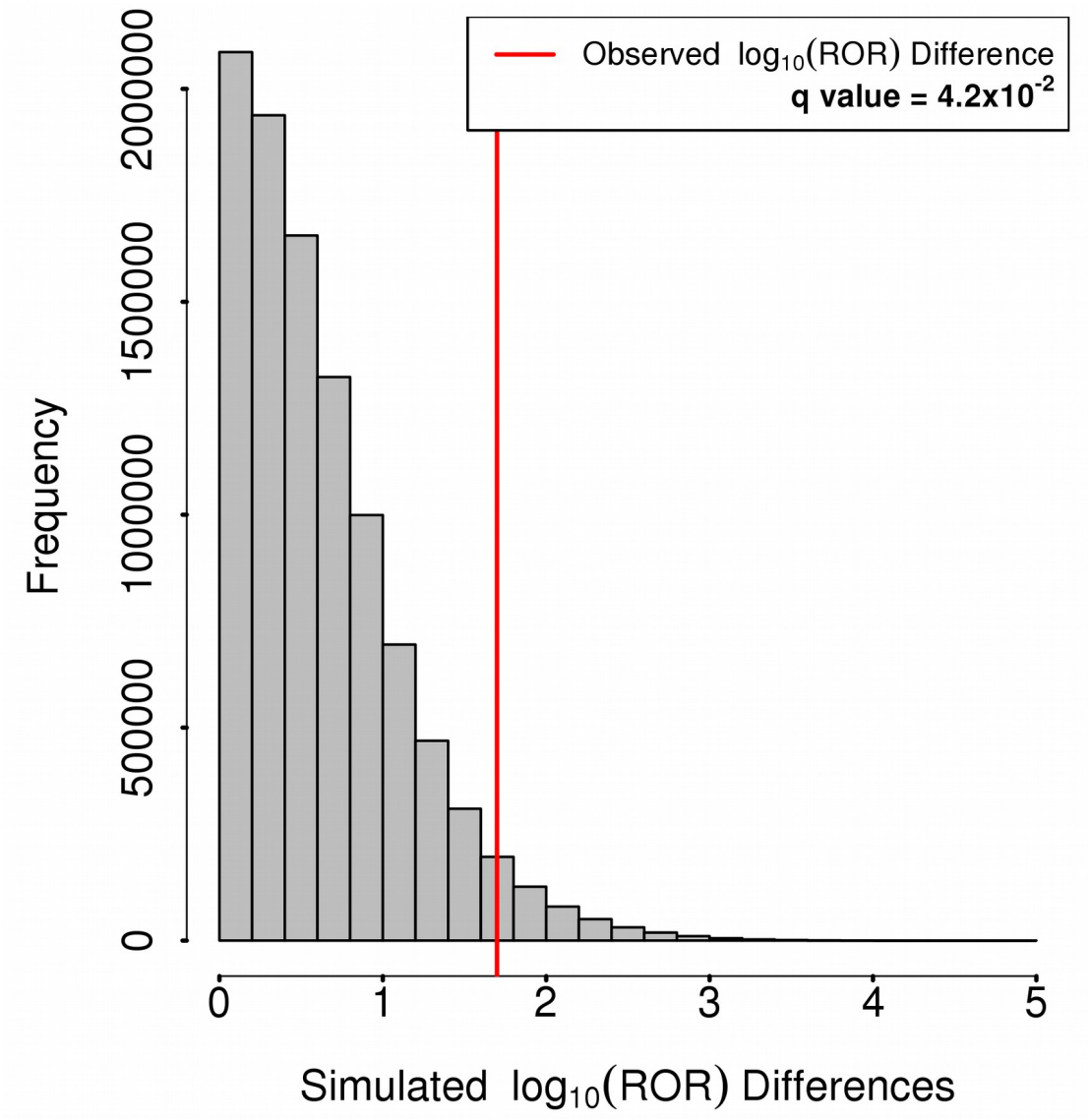


Figure 4.30. A histogram of the log_{10} ratio of odds ratios (RORs) for the simulated contingency tables in the permutation testing. The actual log_{10}ROR observed between the ID/non-ID lists of the neocortex/PC contingency tables is marked with a solid red line, and the subsequent q-value displayed in the legend.

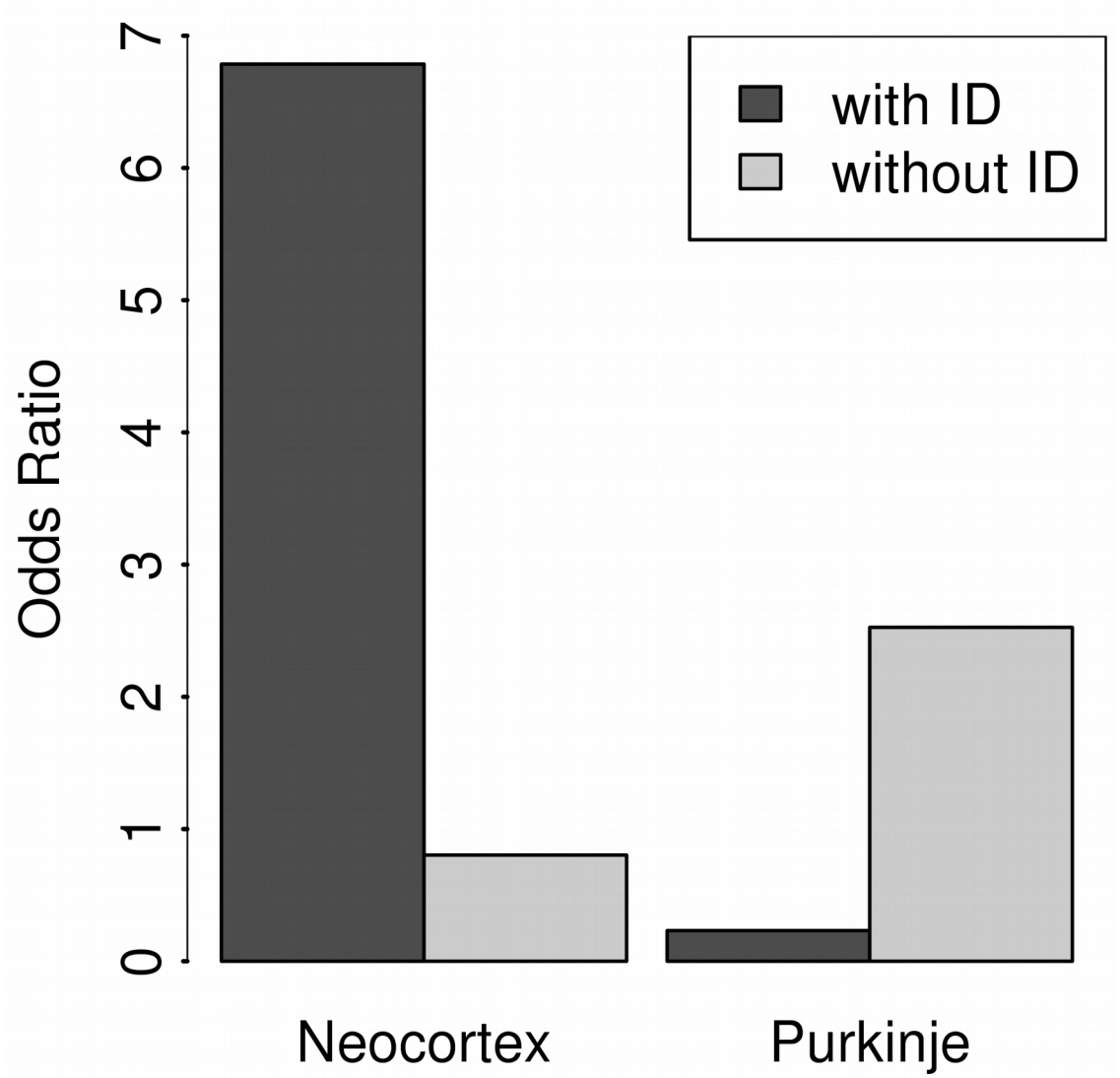


Figure 4.31. A bar plot of the odds ratios observed for the enrichment of SFARI syndromic genes only, stratified into those associated with ID and without ID, and shown for both the PC and neocortex data. The neocortex data displays a large enrichment for ASD-genes associated with ID, and the PC data displays both an enrichment for ASD-genes not associated with ID, as well as a depletion of ASD-genes associated with ASD.

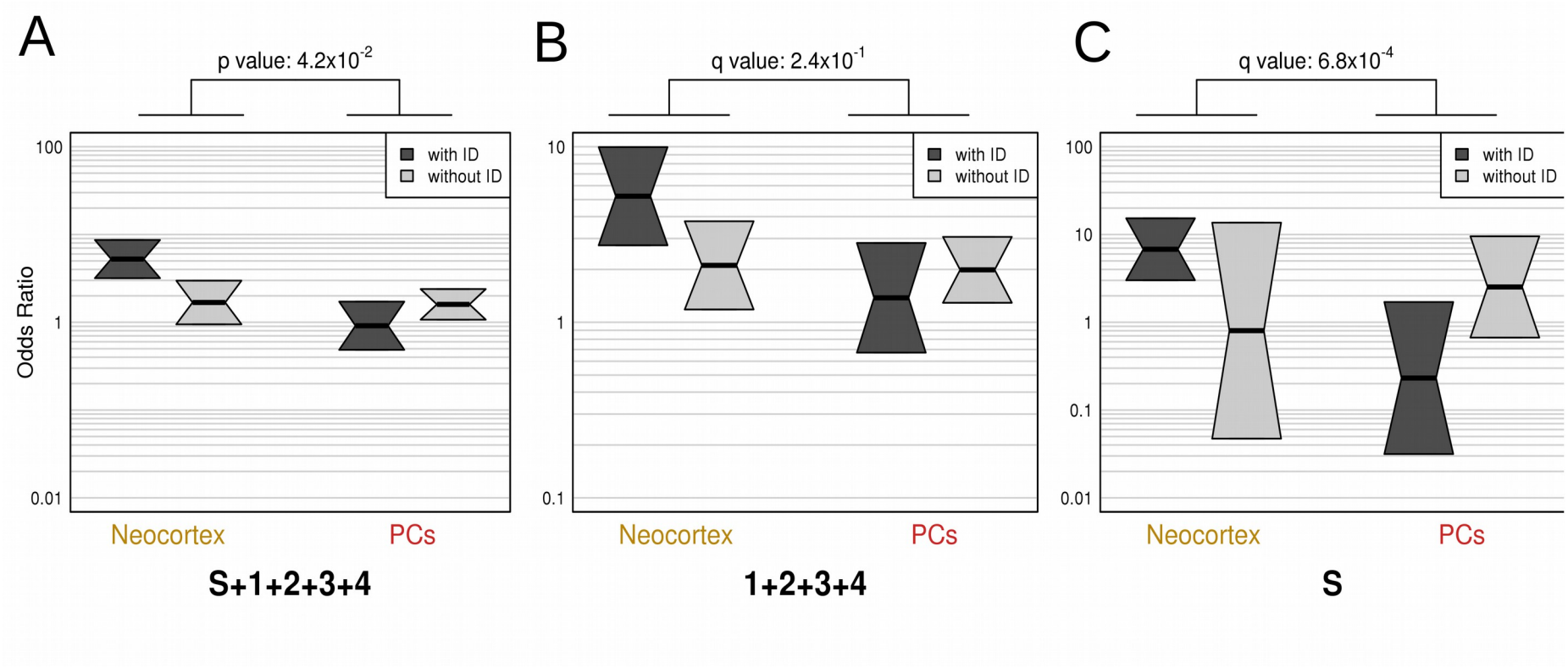


Figure 4.32. Box plots displaying odds ratios for the enrichment of SFARI human autism-associated genes, stratified into those associated with ID and without ID, and shown for both the PC and neocortex data. This displays these plots for categories 1-4 and syndromic genes (A), in addition to category 1-4 genes only (B), and syndromic genes only (C) (as in Figure 4.31). The midline of the boxes signifies the odds ratios, with notches signifying 95% confidence. The odds ratios are displayed on a $\log_{10}$ scale. The p and q values indicate the estimated probability of the difference between the ratios of odds ratios.

4.4 Discussion

In this study, a cluster of genes with significant increase in expression over developmental timing in the PCs has been identified, and shown to have a significant enrichment of ASD-associated genes (both human candidates and mouse models). Importantly, these ASD-associated genes have been found to be strongly associated with a distinct endophenotype: the absence of ID co-morbidity. This has been shown in contrast to an independent gene cluster in the neocortex of the developing mouse, which has also been shown to be enriched for ASD-associated genes, but with ID association. These findings expose a relationship between the spatiotemporal expression pattern of ASD-associated genes and genetically distinct ASD endophenotypes. Due to the heterogeneity of the ASD phenotype, and the many different genes associated with these, it is unsurprising that multiple brain regions, cell types, and developmental time periods could contribute to the disorder. The ASD-gene enriched expression cluster specific to developing PCs identified in this study lends support to the emerging concept of the neuronal populations of the cerebellum being highly vulnerable to ASD (Becker *et al.* 2009; Wang *et al.* 2014).

Multiple clustering methods were used in this study, with varying levels of success. WGCNA successfully returned two useful clusters (positively and negatively correlated over time) with significant GO, KEGG and MGI enrichments, with one of these clusters containing an enrichment for autism-associated genes. However, the remaining clusters were relatively small by

comparison (Figure 4.10A) and returned no convincing functional enrichments. Overall, WGCNA was not successful in fully profiling the PC data. It is possible that this issue could have been resolved with a higher number of replicates. The creators of WGCNA recommend a minimum of 15 samples, with a network based on data from fewer samples than this likely to be too noisy to be biologically meaningful (Langfelder & Horvath, 2014). The PC data contains exactly 15 samples (3 replicates at 5 time points), and therefore was possibly not sufficiently robust for the WGCNA algorithm. Comparatively, the method of silhouette width and fuzzy clustering algorithms returned fewer but larger more robustly sized clusters. This returned five clusters, peaking at each of the five time points, suggesting a more successful clustering.

One other noteworthy study is that of Parikshak *et al.* (2013), which similarly attempted to discern the causative genes in ASD ID comorbidity. This also identified a tightly coalescing module of ASD-associated genes, with implicated functions in early transcriptional regulations and synaptic development. Furthermore, the involvement of FMRP was implicated in the transcription factor networks connecting these functions, and the ID-risk genes were identified as distinct. This was also implemented with WGCNA and SFARI genes, but on transcriptomic data of the neocortex only. In contrast to this, the study presented here demonstrates ASD gene modules within the PCs, before drawing comparisons with the neocortex, providing evidence for cell-specific transcriptome dynamics relating to ASD. Other transcriptomic studies have also identified cerebellar gene networks which have been shown to be enriched in ASD-associated genes, and show partial overlap with our data. For example,

one analysis investigated images of the cerebellar cortex from the Allen Mouse Brain Atlas, defining a co-expression matrix and identifying a higher degree of co-expression connectivity in ASD-associated genes (Menashe *et al.* 2013). This included genes also present in the PC cluster presented here, such as *Aldh5a1*, *Astn2*, *Auts2*, *Dpp10* and *Sez6l2.* A similar analysis built a co-expression network from human BrainSpan data, and identified an ASD enrichment including *ANK2* in the thalamus and cerebellum throughout postnatal development (Willsey *et al.* 2013). All of these studies support the findings presented here, as well as highlighting the need to investigate many neuronal subpopulations across multiple regions of the brain.

These key genes found in our study and in the other aforementioned studies of the cerebellum are of great interest. For example, variants among the *Aldh5a1* gene have been associated with a syndromic from of ASD which should be further investigated. *Aldh5a1* encodes the protein Succinic Semialdehyde Dehydrogenase (SSADH), which is involved in the degradation of the neurotransmitter GABA. GABA transaminase metabolizes GABA into succinic semialdehyde, which in turn is either converted into succinic acid by SSADH or into GHB (gamma-hydroxybutyric acid) by succinic semialdehyde reductase. SSADH Deficiency (SSADHD) therefore results in large increases of GHB (approximately 30-fold) and GABA (approximately 3-fold) (Buzzi *et al.* 2006). The resultant symptoms of this include developmental delay, mental retardation, seizures, non-progressive ataxia, and autistic behaviours (Pearl *et al.* 2003). SSADHD is a very rare autosomal recessive disorder (reported in approximately 350 families; Pearl *et al.* 2003), and therefore further studies in

humans are difficult to implement. However, the associations of this gene in cerebellar development (having been identified in this study and those previously discussed and including symptoms such as ataxia) make this gene worthy of further study in relation to its function in the cerebellum and how this could relate to ASD. Similarly, the gene *Sez6l2* also has interesting associations with both autism and the cerebellum, and may merit further investigation. *Sez6l2* encodes a protein localized to the surface of the cell, and is a paralog of the gene Seizure-related 6 (*SEZ6*). *SEZ6* plays a role in cell-cell signalling and neuronal membrane signalling, with an important role in dendrite elongation and branching. Although the role of *SEZ6L2* is less well characterized, it is thought to contribute to endoplasmic reticulum functions in neurons (UniProt Consortium, 2015). The association of *SEZ6L2* with ASD is also indirect. SEZ6L2 falls within a 500-700kb region of the genome at 16p11.2, along with 23 other genes, which is microdeleted or duplicated in many ASD cases, constituting the second most frequent chromosomal disorder observed in ASD. Further study of the genes in this region found that *SEZ6L2* was a likely candidate due to the association of ASD with a coding variant (Kumar *et al.* 2009). Further study of this gene in relation to its role in the development of PCs could shed more light on its association with ASD.

The novelty of this study lies in both the specific analysis of PCs and in the association of distinct spatiotemporal gene clusters with ID comorbidity. In addition to supporting studies into the associations between ASD and the cerebellum, our gene cluster confirms that many such genes are definitively co-expressed in PCs. This study provides a more specific investigation into

ASD-associated genes and which of these may play a specific role in PC development. Additionally, the comparison of this gene cluster with a second ASD-associated gene cluster identified in the neocortex sheds light on the genetic distinction between many genes associated with ID-comorbidity and others without. The stratification of these ASD-associated genes through the formation of these specific gene clusters in separate tissue/cell types and with these specific temporal patterns is entirely novel and suggestive of genetically distinct endophenotypes.

Overall, the findings of this study will be useful in informing future functional studies of the specific ASD-associated genes identified in the developing PCs, which should allow deeper understanding of the underlying pathophysiological mechanisms of ASD. Using a highly enriched single neuron population at multiple timepoints across a sensitive window of development has allowed for the identification of ASD-specific transcriptomic signatures that suggest genetically distinct ASD endophenotypes that could have otherwise been masked in a traditional whole-tissue approach. These results will hopefully contribute to future studies that may eventually lead to differential diagnosis and patient stratification, including ASD phenotype, severity, and possibly targets for therapeutic interventions.

Chapter 5 – RNA-Seq Ontology-assisted Clustering and K Selection: the ROCKS Package

Abstract

Model-based clustering can provide a method of classification that is statistically suited to RNA-Seq data. However, the time consuming process of applying a statistically complex procedure to the increasingly large RNA-Seq data sets being produced is a distinct disadvantage. Additionally, selecting the number of clusters should be based on more than simply expression pattern similarity. This chapter suggests methods for helping to address these issues, through parallelization of model-based clustering and cluster profiling procedures, and the use of gene ontology semantic similarity analysis to ensure clusters of more ontologically similar genes. For the application of these processes, I present ROCKS: an R package for RNA-Seq Ontology-assisted Clustering and K Selection.

Publication and contributions

This chapter is awaiting manuscript preparation and publication. Package design and all subsequent development and testing were performed by me, with Wilfried Haerty and Chris P. Ponting advising.

5.1 Introduction

Clustering is a statistical classification technique in which the main objective is to form groups or clusters from a set of objects, and for each cluster to contain objects showing a greater similarity with one another than with objects in other clusters. Clustering is an approach to data mining which is utilized in many fields, one of which is for elucidating patterns in large biological data sets such as those produced from next-generation sequencing. For example, in RNA-Sequencing clustering is frequently used to extract commonly occurring gene expression patterns observed over time-series data, where each cluster contains a group of genes with similar expression patterns (Baumgart *et al.* 2014; Cho *et al.* 2014; Klepikova *et al.* 2015; Marthandan *et al.* 2015; Oberstaller *et al.* 2013).

No one specific clustering algorithm is advised for use on all occasions. Regarding RNA-Sequencing, one algorithm commonly used is K-means clustering (Cho et al. 2014; Klepikova et al. 2015; Simm et al. 2015). K-means clustering is a quick and simple method that can be applied to large data sets such as those from RNA-Seq with relatively low compute-intensive processing. The aim of this algorithm is to partition the data into K clusters (a number of clusters selected by the user) in which each observation (in the case of RNA-Seq, each gene) belongs to the cluster with the nearest mean (or centroid). The algorithm achieves this firstly by allocating K centroids at random, and then through an iterative method, in which it assigns each gene to its nearest centroid (the assignment step), followed by the reallocation of the centroid to

the centre of its assigned genes (the update step). The two steps are repeated until the centroids remain constant (or until a specified maximum number of iterations have been reached) (Figure 5.1). This results in each centroid finding a cluster which is fully separated from the other centroids (Lloyd, 1982). The specific derivative of the K-means algorithm that is commonly used for RNA-Seq is the fuzzy clustering (otherwise known as soft clustering) algorithm (Baumgart *et al.* 2014; Marthandan *et al.* 2015; Oberstaller *et al.* 2013). Fuzzy clustering follows the same algorithm as K-means clustering, but allows a gene to be assigned to multiple clusters, with its assignment to each cluster measured with a membership value (calculated from the distance to the centroid).

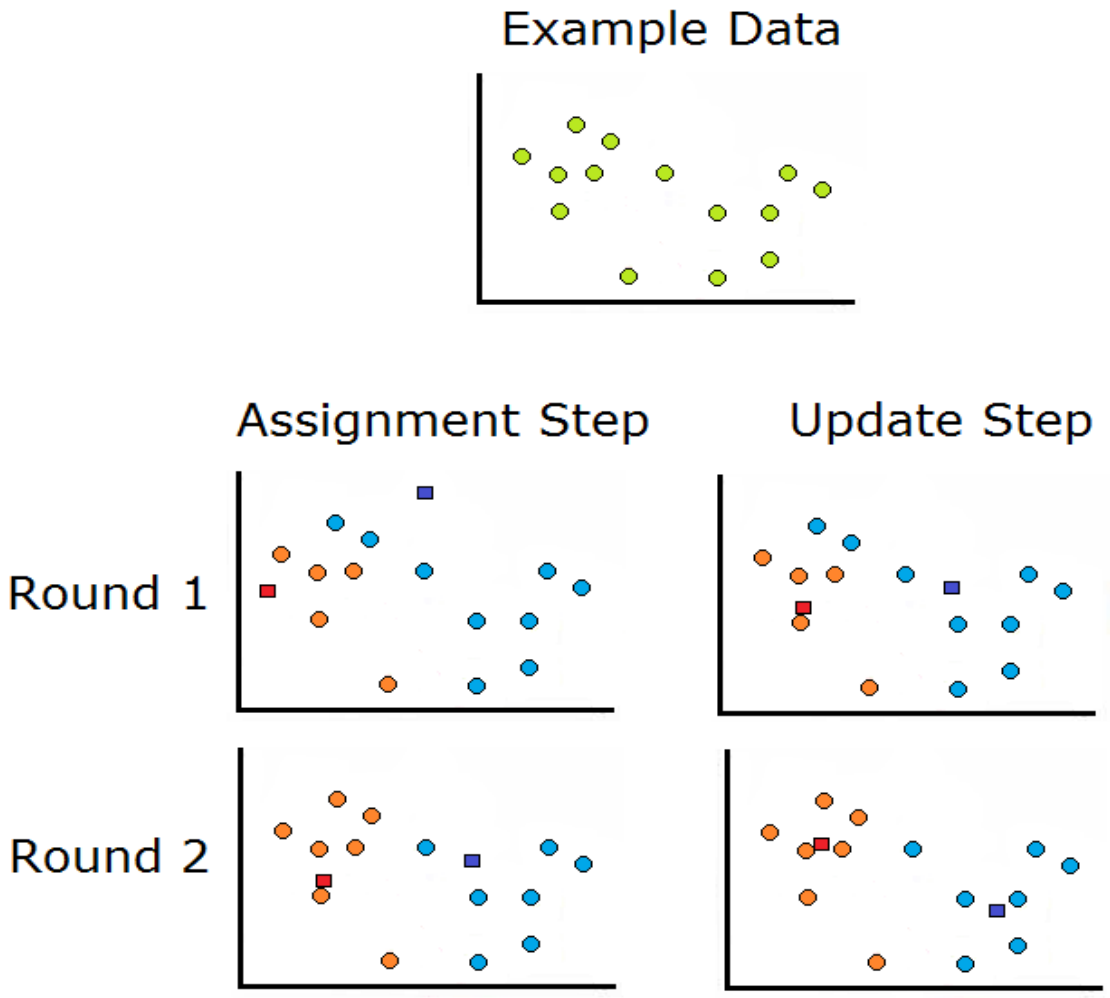


Figure 5.1. A schematic diagram showing two iterations of the K-means clustering method to produce two clusters. Two-dimensional example data is displayed (in green). To initialize the K-means algorithm, centroids (represented here by square points) are placed randomly amongst the data and the two clusters (represented here in orange and blue) are formed of the points nearest to each respective centroid (assignment step, round 1). The centroids are then updated by moving these to the centre of each cluster (update step, round 1). This process is repeated, with data points being re-assigned to the new centroids (assignment step, round 2), and centroids being updated (update step, round 3), until the centroid no longer moves in the update step.

Although the fuzzy clustering algorithm can be used successfully in RNA-Seq, alternative algorithms have been suggested to be better suited, for example model-based clustering (Si *et al.* 2014). The K-means or fuzzy clustering approach to clustering can be described as a heuristic approach, and it may not be entirely suitable for the distribution of the data obtained from RNA-Seq (Figure 5.2). The data analyzed from RNA-Seq is count data, with a count assigned to each gene or transcript, and count data typically follows a Poisson distribution. However, in RNA-Seq data, a greater than expected (in the Poisson distribution) variance at high expression levels is often observed due to natural biological replicate variation, which is referred to as "the over-dispersion problem". Hence, RNA-Seq data is ultimately better fitted to a negative-binomial distribution (Anders & Huber, 2010; Whitaker, 1914). For this reason, a clustering algorithm suggested for RNA-Seq is model-based clustering with a negative-binomial model, for example using the package MBCluster.Seq (Si *et al.* 2014).

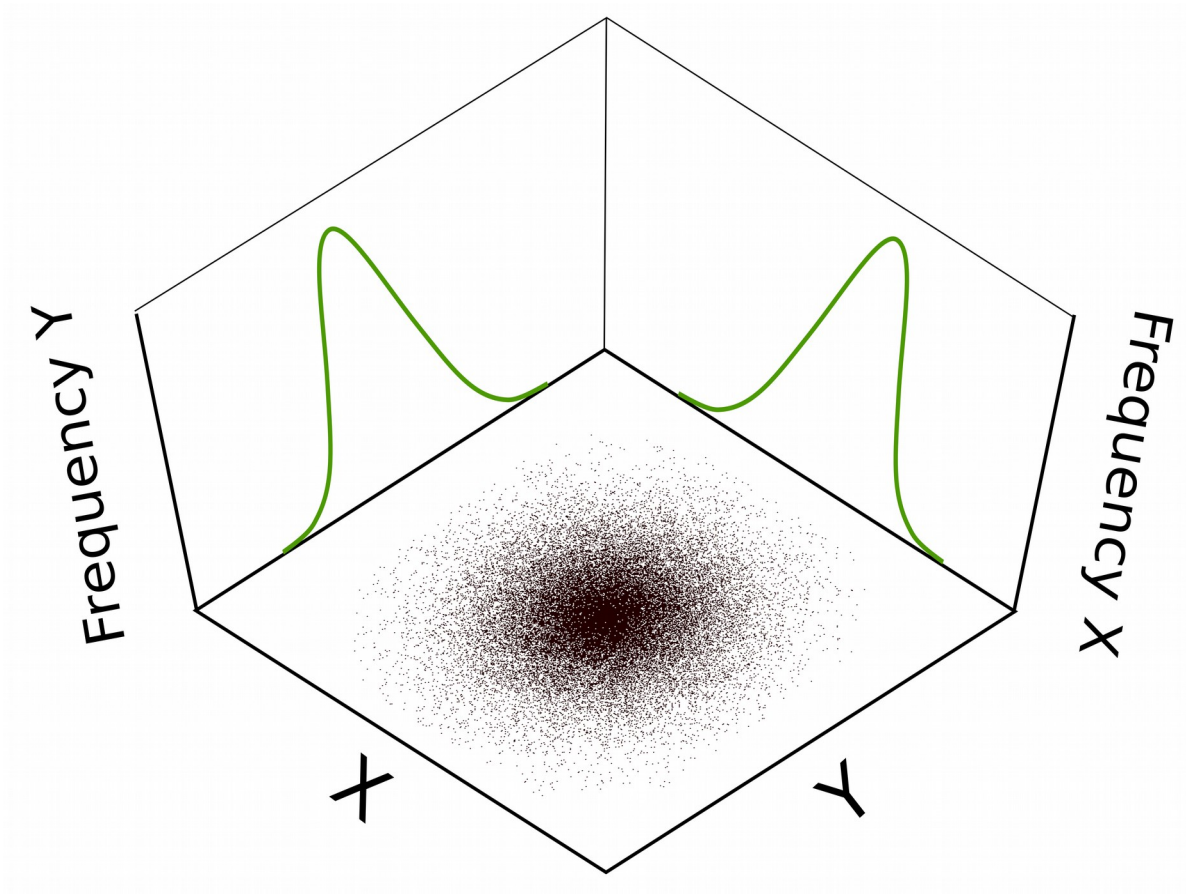


Figure 5.2. A schematic demonstrating a cluster of two-dimensional data points and their distributions. Most heuristic approaches to clustering, such as K-means, have no prior expectation for the data's distribution. A classically envisaged cluster that forms a circle, like the example shown here, follows a Gaussian distribution. RNA-Seq count data is thought to be best modelled by a negative binomial distribution.

Model-based clustering is a form of clustering based on distribution models (such as the negative-binomial) which forms a mixture model (a model which represents multiple subpopulations) and therefore allows clusters to be formed of those objects most likely to belong to the same distribution. This is achieved using the expectation-maximization (EM) algorithm, which is applied in an iterative method similar to K-means. Each iteration for the EM algorithm is formed of the expectation step (the equivalent of the K-means assignment step), which computes the probability of each observation belonging to each model using the current parameter estimates, followed by the maximization step (the equivalent of the K-means update step), which uses the observations to re-compute model parameters (Dempster *et al.* 1977). Although model-based clustering is thought to provide greater accuracy in clustering, it also involves compute-intensive modeling stages.

One difficulty in both of these methods is selecting a value for K (the number of clusters to be formed from the data). With traditional clustering methods such as K-means, a variety of solutions have been applied as useful internal evaluations (i.e. evaluating the resultant clusters formed) for biological data, examples being the Calinski-Harabasz Criterion (based on maximizing the between-cluster variance and minimizing the within-cluster variance) (Calinski & Harabasz, 1974; Richards *et al.* 2008) and silhouette width (a measure for determining the coherence of clusters) (Rousseeuw, 1987; Clifford *et al.* 2011). For model-based clustering an information criterion can be used, which seeks to maximize the likelihood function of the model (the goodness of fit) whilst penalizing the model based on the number of parameters (penalizing

complexity or over-fitting). The log-likelihood value is calculated using maximum likelihood estimates for when a model is applied to a data set, calculating this from the probability of the data points occurring assuming the model is correct. This likelihood will increase as the model becomes more complex, and with each iterative calculation the model better fits the data. To avoid overfitting the data, the information criterion is a log-likelihood penalized by the number of parameters. For Akaike's Information Criterion (AIC), this is -2(log-likelihood – number of parameters) (Akaike, 1974). Therefore, when plotting AIC value vs. K, this will form a characteristic curve (see Results), for which a good estimate of K can be taken as the "elbow point" of the curve, i.e. the point which minimizes AIC and minimizes K, thereby providing the least complex clustering for the best fit. The AIC is the criterion suggested by the authors of MBCluster.Seq, although the functions for calculating this value are not included in the package itself (Si *et al.* 2014).

Although AIC and model-based clustering are methods well suited to clustering RNA-Seq data, there are two main issues that arise in their application. The first is the time required due to the compute-intensive model-based clustering of such large data sets. With some of the larger data sets it can take many hours for a single algorithm to be run. Additionally, determining AIC (or any K-selection statistic) required running the algorithm many times before using AIC to determine the best K value. Secondly, clustering based solely on statistical significance could potentially ignore smaller clusters of biological significance. A likely example would be from a clustering algorithm returning a gene cluster characterized by increasing expression over time, for which this cluster may

actually contain multiple similarly-expressed genes of no relation in terms of biology. Therefore, in such cases, purely statistical approaches to clustering (although producing well-formed clusters) are unable to predict when further separation reflecting biological significance may be required.

The ROCKS package (see below) aims to address both of these significant issues. Firstly, by introducing parallel processing to these procedures, which allows for the multiple runs of model-based clustering and many more runs of downstream analyses for each of the clusters within these, such analyses should be performed within a more reasonable time. Secondly, by incorporating prior knowledge in the form of gene ontological data to complement AIC for cluster number selection. In addition to allowing the user to account for gene ontology similarity within clusters in their selection of K, the package allows the user to ensure that their selection of K maximizes the number of enriched gene ontologies, therefore extracting as much interpretable information as possible from the data.

5.2 Methods

Implementation

This package contains three main components. The first performs a run-through of the negative binomial model-based clustering procedure implemented by MBCluster.Seq. To support parallel processes in this package, these have been written under a wrapper function, cluster.wrapper(), which provides a framework for applying the main functions with varying values of K in parallel. This has been implemented with Parallel Socket Cluster (PSOCK) functions of the doParallel package (*Revolution Analytics & Weston, 2014)*, and hence is utilizable on Windows operating systems. The three main stages of MBCluster.Seq are denoted by three functions which are all contained within this wrapper: RNASeq.Data, KmeansPlus.RNASeq, and Cluster.RNASeq, which serve the purposes of data preparation (including log-quantile data normalization), initial K-means centroid determination, and application of the EM algorithm, respectively. These have been amended to increase functionality, firstly by the addition of AIC calculation, through extraction of the log-likelihood and calculation of the number of parameters from the final negative binomial model produced. Secondly, with the addition of forking (using functions of the R package: parallel). This is an optional extra that allows further parallel processes to be spawned from the PSOCK environments. This enables the user not only to run with a range of K-values in parallel, but to introduce further parallel processes for use within the model-based clustering, allowing an even greater increase in speed for users with more advanced

computer systems. Finally, a visualization functionality has been added which includes a function that allows the user to plot the AIC against the cluster number (aiding in determination of the "elbow point" and the selection of a cluster number), as well as a function allowing the user to plot expression values (with standardized mean and variance) for all of the genes in a specified cluster.

The second component of this package enables the user to perform enrichment analyses on the clusters produced. This functionality can be used for both cluster number selection and for subsequent analysis of the specific gene lists and therefore, in addition to GO enrichment analysis, KEGG pathway enrichment analysis has also been included as an option. The GO enrichment analysis has been implemented with GOSeq (Young *et al.* 2010), an R package designed for RNA-Seq data that includes compensation for gene selection bias in GO enrichment analyses (see *Methods for Additional External Packages*). Once again, PSOCK parallel processing capabilities are provided, allowing the user to perform GO enrichment analysis on all clusters derived from all of the varying K-values, in parallel. Finally, I include functions for applying semantic similarity measures to the resultant gene ontology information. This is implemented through the GOSemSim package (Yu *et al.* 2010), which can be used to determining the semantic similarity between any two gene ontology terms, and is used in this package to assign an average semantic similarity to a given cluster (see *Methods for Additional External Packages*). Once again, this has been given a wrapper function to enable these processes to be applied in parallel (PSOCK) to all of the GO enrichment analyses for all of the clusters.

The third and final component of this package enables the user to visualize this information. This includes functionality for extracting AIC and extracting information on the number of GO terms enriched, as well as the mean semantic similarity, at each value of K and in a format suitable for plotting to identify patterns. In addition to this, there is a function for extracting gene expression information for any cluster selected by the user, and a subsequent function for plotting these expression patterns, allowing the user to directly observe the cluster pattern.

Information on all functions forming the ROCKS package and the use of these can be observed in the package vignette. Figure 5.3 also gives a schematic overview of the ROCKS functions. Finally, the code of each function and function help pages describing what each function does, where it lies within the package, and what other packages it utilizes or is edited from, can be found in Appendix B.

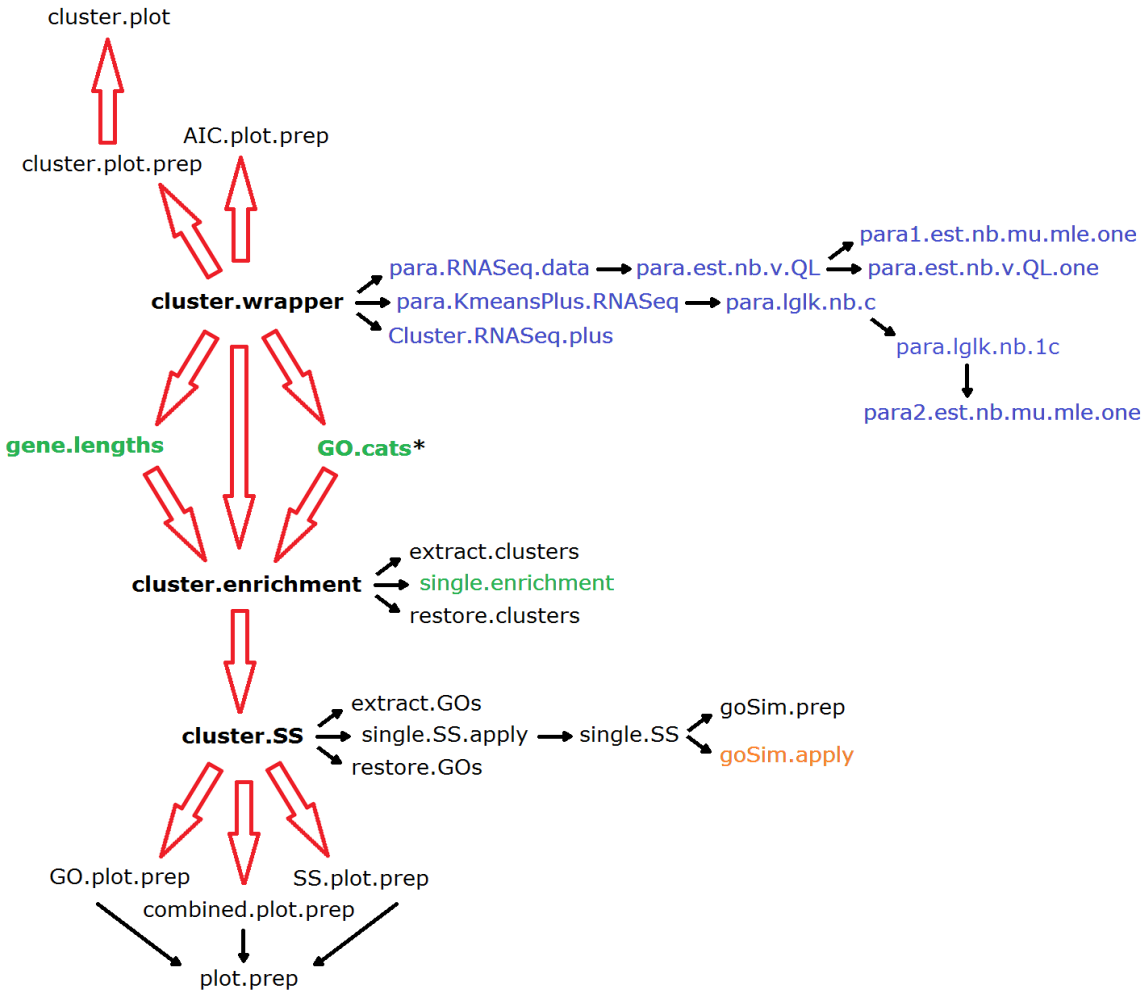


Figure 5.3. A schematic naming the functions written for the ROCKS package and displaying their associations to one another. Red arrows indicate the flow of functions developed for the user (as demonstrated in the package vignette), including core functions for data processing (bold) and additional functions for plotting. Black arrows point to functions applied internally, as part of other functions. Coloured function names indicate utilization of an exterior package, including those that are edited versions of MBCluster.Seq functions (blue), those that contain GOSeq functions (green), and those that contain GOSemSim functions (orange). Additionally, as an alternative to the GO.cats function (*), the KEGG.cats function can also be applied to investigate KEGG pathway enrichments.

Methods used in Additional External Packages

As mentioned above, in addition to the methods of MBCluster.Seq, ROCKS utilizes established packages for GO enrichment and GO semantic similarity calculations. There are many gene ontology enrichment analysis tools available, for example DAVID, topGO and EasyGO (Alexa *et al.* 2006; Huang *et al.* 2009; Huang *et al.* 2009; Zhou & Su, 2007). The GO enrichment tool utilized in ROCKS is GOSeq, an R package developed specifically for RNA-Seq. GOSeq performs GO enrichment analysis whilst compensating for gene length bias, a phenomenon initially observed in RNA-Seq by Oshlack and Wakefield (2009), who concluded that longer transcripts have a greater likelihood of high expression, and therefore often are assigned a greater number of reads. This in turn enables a greater power for statistical testing. GOSeq aims to account for this bias by compensating its enrichment testing in those categories without a preponderance of long genes and therefore a lower likelihood for enrichment. The standard method for testing enrichment uses the Fisher's exact test. However, this assumes each gene has an equal probability of selection, and therefore that the number of genes selected that overlap with those associated with an ontology follows a hypergeometric distribution. To compensate for gene length, GOSeq instead quantifies the likelihood of selection by fitting a monotonic function to this selection likelihood, versus gene length (with genes binned by length to allow a likelihood to be estimated), with this function then incorporated into statistical testing. The statistical testing used is then based on the Wallenius non-central hypergeometric distribution (Chesson, 1976), which is a generalization of the

hypergeometric distribution where items are samples with a particular bias.

In ROCKS, I have supplied functions for applying the full GOSeq process to individual clusters of individual cluster runs, or to all clusters of all cluster runs. For the latter, as mentioned previously, there is functionality for PSOCK parallel processing. The ROCKS package flattens the lists of cluster information to allow GOSeq to be applied to all efficiently; therefore there is no need for additional forking at this stage. In applying GOSeq, the methodology and core principles of the package remain unchanged; however the organization of the package has been adapted for optimized performance in multiple processes. For example, the obtaining of gene length information and GO category information has been moved outside of the core goseq() function to avoid repetitive use and therefore to allow quicker looping when applying the GO enrichment analyses to multiple different datasets.

Semantic similarity calculations are implemented in the ROCKS package through the GOSemSim tool (Yu *et al.* 2010). GOSemSim provides implementation of multiple methods for calculating semantic similarity, all of which provide their own advantages, and none of which is considered the single gold-standard approach. These include both Information Content (IC) and graph-based method. IC-based measures depend on the frequencies of GO terms and their closest common ancestor term, obtaining information for this from annotation databases such as the UniProt knowledgebase. Graph-based measures use the topology of GO graphs and their structure to compute semantic similarity, determining the specificity of a GO term based on its

location in the GO graph (Yu *et al.* 2010). The singular graph-based method provided by GOSemSim was proposed by Wang *et al.* (2007), and involved aggregating the semantic contributions of a GO term's ancestor terms and assigning a value based on the locations of these terms in the GO graph. Wang *et al.* showed that their algorithm in this method performs well on co-expression data, in clustering based on resultant semantic similarity scores.

The ROCKS package applies the Wang method through the GOSemSim package, and applies this calculation to all possible pairings of genes within a cluster, thereby establishing this cluster's mean semantic similarity. As before, there is functionality for applying this process to an individual cluster or to all cluster runs, and in the case of the latter there is functionality for PSOCK parallel processing on flattened cluster lists to maximize efficiency (and negate the need for forking).

Testing

To ensure that ROCKS is robust, it was tested on multiple diverse time-series RNA-Seq data sets. Firstly, on the Purkinje cell data set used previously in this thesis (Chapter 4). Secondly, on an older and smaller dataset generated to demonstrate Cufflinks, applying RNA-Sequencing to a well-studied mouse model of skeletal muscle development (the C2C12 myoblast cell line) (Trapnell *et al.* 2010). Lastly, on a much larger modENCODE data set which investigates the developmental transcriptome of *Drosophila melanogaster* (Graveley *et al.* 2011). The Purkinje data set will be included in the package test data; the

Trapnell and Graveley data sets were obtained from ReCount, a resource providing analysis-ready R ExpressionSet data for RNA-Seq gene count datasets (Frazee *et al.* 2011).

5.3 Vignette

Input Data

ROCKS requires time-series data from an RNA-Seq experiment. This should include samples of multiple replicates over multiple time-points. This can be used by *ROCKS* to perform model-based clustering, which will extract clusters of genes with similar patterns over time, in addition to performing gene annotation enrichment analyses on these clusters.

For fully automated use of *ROCKS*, the experiment should have been executed on a suitable model organism, and reads aligned to a suitable genome version. The organism must be available for applying semantic similarity calculations with the *GOSemSim* package (Yu *et al.* 2010), and the genome version must be available for applying gene ontology calculations with the *goseq* package (Young *et al.* 2010). Suitability of the former can be checked by viewing organism variable options on the help page of the goSim function; suitability of the latter can be checked by loading *geneLenDataBase* of the *goseq* package and running the supportedGenomes function.

```
> library(GOSemSim)
> ?goSim


> library(geneLenDataBase)
> supportedGenomes()
```

Once the data has been determined as suitable, it may also be helpful to ensure that the gene IDs available are supported for automated gene length information download later in this vignette. This can be checked using the supportedGeneIDs function. If there are no suitable supported gene IDs, the IDs will either need to be converted to a format that is supported, or a custom numeric gene length vector can be manually created.

```
> library(geneLenDataBase)
> supportedGeneIDs()
```

One of the initial inputs required by *ROCKS* is a raw count data table of integers. These counts do not need to be normalized. The columns of the count table should correspond to the independent biological replicates of all of the samples of the experiment, with the header formed of unique names. The replicate data does not need to be combined at this stage; the treatment vector to be provided as a secondary input will be used to combine replicates at a later stage. The rows of the count table should correspond to genomic features (e.g. genes), with the row names formed of gene IDs.

For automated downstream analysis, such as gene ontology enrichment, the gene IDs should be in a format recognised by the *goseq* package. This can be tested by loading *geneLenDataBase* of the *goseq* package and running the supportedGeneIDs function.

```
> library(geneLenDataBase)

> supportedGeneIDs()
```

In this vignette, we will be working with a tab-delimited count table which is included with the *ROCKS* package. This count table is from an RNA-Seq experiment for profiling cerebellar Purkinje cell gene expression over five time-points. The function system.file can be used to locate the package and the file inside, which can then be read with the read.delim function.

```
> filename <- system.file("data/counts.txt", package="ROCKS")

> counts <- read.delim(filename, header=T,
row.names=1,stringsAsFactors=F)

> head(counts, n=3)
                 P0_R1 P0_R2 P0_R3 P4_R1 P4_R2 P4_R3 P8_R1 P8_R2 P8_R3
ENSMUSG000000000   483   363   242   267   296   565   238    87   319

               01
ENSMUSG000000000     0     0     0     0     0     0     0     0     0

               03
ENSMUSG000000000    14     6    50     0     0    53     1     7     0

               28
                   P14_R P14_R P14_R P21_R P21_R P21_R
                       1     2     3     1     2     3
ENSMUSG000000000   204   186   171    14   134    87

               01
ENSMUSG000000000     0     0     0     0     0     0

               03
ENSMUSG000000000     7    70    14     3    93    38

               28
```

The final piece of information to be provided to begin is the treatment vector. This is a vector of length equal to the number of columns in the count table, i.e. one entry per replicate, and is used to determine which of these are replicates of one another. For example, a treatment vector of c(1,1,2,2,3,3,4,4) implies that there are eight columns in the count table, the first two being replicates of the first time point, the second two being replicates of the second time-point, and so on.

For the example data we are working on, we have five stages with 3 three replicates for each.

```
> treatment <- rep(1:5, each=3)
> treatment
 [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5
```

Socket Cluster for Parallel Processing

One of the advantages for using *ROCKS* to perform model-based clustering is the introduction of parallel processing to the calculations. This allows the model-based clustering algorithm to be applied multiple times with multiple values of K (the number of clusters to be formed from the data) in a much reduced amount of time.

Many of the functions in *ROCKS* have the ability to work with socket clusters, meaning that multiple sessions of R are run at once over multiple cores. Socket

clusters can be built in any operating system and if they are to be used, should be done so before running the main functions of *ROCKS*. This should be done by loading the *doParallel* package (Revolution Analytics & Weston, 2014), and firstly running the makeCluster function with the desired number of cores specified. The object that this has been stored in should then be passed to the registerDoParallel.

```
> library(doParallel)
> mymaxcores <- 4
> mysockcluster <- makeCluster(mymaxcores)
> registerDoParallel(mysockcluster)
```

This socket cluster can be closed at any point by ending the session or using the stopCluster function on the socket cluster object.

```
> stopCluster(mysockcluster)
```

Once this cluster has been created, simply running one of the functions cluster.wrapper, cluster.enrichment, or cluster.SS will allow these processes to be run in parallel over the number of cores specified.

If the user does not wish to use socket clusters, these functions will also run appropriately on a single core without the creation of a socket cluster.

(Note: when running one of these functions, occasionally not all of the cores in

the socket cluster created are used, or immediately used. This can be due to the socket cluster having more cores than it requires, or simply due to it taking time for the large amount of data to be loaded before it is distributed across all cores.)

Model-Based Clustering

The cluster.wrapper function provides a wrapper for the use of the *MBCluster.Seq* package (Si, 2010) to apply negative binomial distribution-based clustering to the RNA-Seq data. In addition to this, it adds functionality for calculating AIC (see *Akaike Information Criterion* section) (Akaike, 1974), as well as for parallel processing to allow the user to apply different values of K in parallel.

In addition to the count table and treatment vector, the cluster.wrapper function requires a vector containing the K values with which to run the model-based clustering algorithm (nK.vector). For a full run-through of the package, I suggest including values from two to a minimum of thirty (nK.vector=2:30), or testing the function on just a couple of values.

Finally, the parallel processing of this function is enabled to work with a socket cluster, and also with forking. The forking option in this function allows each of the socket cluster sessions created to also create a number of sessions for running on a larger number of cores. This is useful if the number of model-based clustering runs to be completed (the length of nk.vector) is smaller than

the number of cores available, allowing an increase by additional forks for each model-based clustering run. This number of forks per socket cluster session is specified with the additional.forks option. For example, if a socket cluster of four sessions has been created and additional.forks=4, this will allow four forks per socket cluster session, using a total of 16 cores.

If the user does not wish for any additional forks to be created, the additional.forks option should be left as 1 (Note: forking is not possible on the Windows operating system and should be left as 1).

Here we see the use of the cluster.wrapper function on the example data (from *Input data* section).

```
> library(doParallel)
> mymaxcores <- 4
> mysockcluster <- makeCluster(mymaxcores)
> registerDoParallel(mysockcluster)

> clustering.data <- cluster.wrapper(counts, treatment,
+                         nK.vector=2:30, additional.forks=1)

> stopCluster(mysockcluster)
```

This will produce a list of lists, specifically a list of several runs of the model-

based clustering algorithm, which each includes a list. The structure of the first model-based clustering run can be viewed as follows:

```
     > str(clustering.data[[1]])
List of 6
 $ probability: num [1:20928, 1:2] 9.98e-09 6.52e-01 7.82e-02
1.41e-01 1.00 ...
 $ centers    : logFC [1:2, 1:5] -0.4186 0.2502 -0.1904 0.2253
0.0113 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:5] "1" "2" "3" "4" ...
 $ cluster    : Named int [1:20928] 2 1 2 2 1 1 2 2 2 1 ...
  ..- attr(*, "names")= chr [1:20928] "ENSMUSG00000000001"
"ENSMUSG00000000028" "ENSMUSG00000000031" "ENSMUSG00000000037" ...
 $ lglk       : num -62.3
 $ nparams    : num 31401
 $ aic        : num 62927
```

Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a measure of relative quality of a statistical model, which can be used to estimate the ideal value of K by seeking a K-value that minimizes the AIC and hence maximizes the quality of the model (Akaike, 1974).

AIC is a value based on the log-likelihood of the model formed by the model-based clustering algorithm (the goodness of fit of the model) and penalizes this by the number of parameters used to form the model (the complexity). As the K-value increases, the complexity of the model increases, and the goodness of fit of the model increases to a point, which causes a leveling off of the AIC value. When plotted against the K-value this causes a characteristic curve, from which the best K-value can be selected as the value at the "elbow" of the curve.

This AIC is calculated as part of the cluster.wrapper function, and the AIC.plot.prep function extracts this information and formats it correctly for plotting.

Here we can prepare data for visualizing the AIC with the AIC.plot.prep function (see Results for subsequent plots), on the example clustering data (produced in the *Model-based clustering* section).

```
> AIC.matrix <- AIC.plot.prep(clustering.data)
> plot(AIC.matrix)
```

Enrichment Testing

Gene length information is important to consider as a factor of potential bias in gene selection (in our case, selection for a specific cluster), because genes of a

greater length (and hence usually a more robust gene count) are likely to be more robust in statistical testing. This is taken into account at the gene ontology stage by the *goseq* package, which can be used to visualize the relationship between gene length and gene selection (by plotting binned gene length against the proportion of genes in that bin that have been selected). The *goseq* package then determines a probability weighting based on any trend seen from fitting a spline using a penalized constrained least squares method.

To obtain gene length information for this process, *ROCKS* provides a wrapper function called gene.lengths, for utilizing *goseq*'s getlength function. To determine suitable genomes and geneIDs for use with this function see *Input data*.

Here we see the use of the gene.lengths function on the example data.

```
    > bias.data <- gene.lengths(clustering.data, 'mm9',
'ensGene')
    > head(bias.data)
ENSMUSG00000000001 ENSMUSG00000000028 ENSMUSG00000000031
            3213.0             1839.5             2262.0
ENSMUSG00000000037 ENSMUSG00000000049 ENSMUSG00000000056
            4053.0             1180.0             4405.0
```

If there are no supported genomes/gene IDs for the data, it may not be possible to use the gene.lengths function to automatically obtain gene length information. If this is the case, a numeric gene length vector will need to be

created manually in the same format as that produced above.

Gene ontology information can also be downloaded, for which *ROCKS* provides a wrapper function called GO.cats for utilizing *goseq*'s getgo function. This will require the same genome/gene ID information as provided in the *Enrichment testing* section.

Here we see the use of the GO.cats function on the example data.

```
     > gene2cat <- GO.cats(clustering.data, 'mm9', 'ensGene')
     > head(gene2cat, 3)
$ENSMUSG00000000001
     [1]   "GO:0000166"   "GO:0000323"   "GO:0001664"   "GO:0001882"
"GO:0001883"
  [6] "GO:0003674" "GO:0003824" "GO:0003924" "GO:0004871" ...

$ENSMUSG00000000028
   [1]   "GO:0005575"   "GO:0005622"   "GO:0005623"   "GO:0005634"
"GO:0005813"
 [6] "GO:0005815" "GO:0005856" "GO:0006139" "GO:0006259"  ...

$ENSMUSG00000000049
     [1]   "GO:0001525"   "GO:0001568"   "GO:0001667"   "GO:0001935"
"GO:0001936"
  [6] "GO:0001937" "GO:0001944" "GO:0001948" "GO:0003674"  ...
```

As before, if there are no supported genomes/gene IDs for the data, it may not be possible to use the GO.cats function to automatically obtain gene ontology information. If this is the case, a list of genes with relevant gene ontology codes will need to be created manually in the same format as that produced above.

ROCKS also provides functionality for downloading KEGG pathway information using the KEGG.cats function. This information can be used for KEGG enrichment testing of the clusters (as shown in the *Enrichment testing* section). However, it cannot be used for any further downstream analyses (e.g. semantic similarity analysis).

Here we see the use of the KEGG.cats function on the example data.

```
    > KEGGgene2cat <- KEGG.cats(clustering.data, 'mm9',
'ensGene')
    > head(KEGGgene2cat, 3)
$ENSMUSG00000000001
 [1] "04062" "04360" "04530" "04540" "04670" "04730" "04914"
"04916" "04971"

$ENSMUSG00000000028
[1] "04110"
```

```
$ENSMUSG00000000058
[1] "04144" "04510" "05100"
```

ROCKS can now use the information gathered to apply enrichment testing across the clusters. For this, the cluster.enrichment function provides a wrapper for the use of *goseq* enrichment testing.

This function can once again utilize parallel processing through working with a socket cluster (see *Socket clusters for parallel processing* section). This function also flattens the lists of clustering information provided to allow efficient parallel processing across all clusters at once, meaning that there is no need for additional forking capabilities.

Here we see the use of the cluster.enrichment function on the example data.

```
> library(doParallel)
> mymaxcores <- 4
> mysockcluster <- makeCluster(mymaxcores)
> registerDoParallel(mysockcluster)

> GO.output <- cluster.enrichment(clustering.data, bias.data, gene2cat)

> stopCluster(mysockcluster)
```

This will produce a list of lists; specifically, a list of several runs of the model-based clustering algorithm, which each includes a list of gene ontology tables for each cluster. The structure of the first model-based clustering run can be viewed as follows:

```
     > str(GO.output[[1]])
List of 2
 $ cluster1:'data.frame':     18368 obs. of  7 variables:
  ..$ category               : chr [1:18368] "GO:0044425"
"GO:0044444" ...
  ..$ over_represented_pvalue: num [1:18368] 1.08e-32 4.46e-30
1.97e-23 ...
  ..$ over_represented_FDR   : num [1:18368] 1.99e-28 4.09e-26
1.21e-19 ...
  ..$ numDEInCat             : int [1:18368] 2135 2509 1782 1737
782 2904 ...
  ..$ numInCat               : int [1:18368] 4574 5545 3853 3748
1553 6637 ...
  ..$ term                   : chr [1:18368] "membrane part"
"cytoplasmic part" "intrinsic component of membrane" "integral
component of membrane" ...
  ..$ ontology               : chr [1:18368] "CC" "CC" "CC"
"CC" ...
 $ cluster2:'data.frame':     18368 obs. of  7 variables:
  ..$ category               : chr [1:18368] "GO:0003676"
```

```
"GO:0090304" ...
  ..$ over_represented_pvalue: num [1:18368] 1.17e-33 8.45e-32
1.71e-25  ...
  ..$ over_represented_FDR   : num [1:18368] 2.15e-29 7.76e-28
1.04e-21 ...
  ..$ numDEInCat             : int [1:18368] 2033 2438 2114 3471
1256 2495 ...
  ..$ numInCat               : int [1:18368] 2892 3541 3084 5252
1782 3738 ...
  ..$ term                   : chr [1:18368] "nucleic acid
binding" "nucleic acid metabolic process" "RNA metabolic process"
"nucleus" ...
  ..$ ontology               : chr [1:18368] "MF" "BP" "BP"
"CC" ...
```

Determining K

In addition to the AIC calculation, it may be useful to use the ontology information produced to determine the most suitable value for K. One method to do this is to seek converging ontologies as the K value increases and the clusters become more specific, which can be achieved through calculating the mean semantic similarities of the clusters.

Individual semantic similarity calculations can be implemented using the Wang method, a graph-based method which uses the topology of GO graph structure

to compute semantic similarity. This works by determining a numeric semantic value, through aggregating the semantic contributions of a GO term's ancestor terms, and assigning a value based on the locations of these terms in the GO graph.

The cluster.SS function calculates semantic similarity for a cluster using the goSim function of the *GOSemSim* package (which, by default, uses the Wang method). It implements this function for all possible pairwise combinations (only comparing those of the same category: biological processes, molecular functions, and cellular components) for all significantly enriched (q-value < 0.05) terms in the given cluster, and then returns the mean of these values. This is implemented across all gene ontology runs (all clusters of all model-based clustering runs) in the object provided (an output from the cluster.enrichment function) in parallel.

In addition to the gene ontology cluster information, the cluster.SS function requires an organism to be specified, in accordance with the *GOSemSim* package suitability (see *Input data* section).

As previously mentioned, this function can once again utilize parallel processing, and does so through working with a socket cluster. As with the cluster.enrichment function, this function also flattens the lists of gene ontology information provided to allow efficient parallel processing across all results at once, meaning that there is no need for additional forking capabilities.

Here we see the use of the cluster.SS function on the example data.

```
> library(doParallel)
> mymaxcores <- 4
> mysockcluster <- makeCluster(mymaxcores)
> registerDoParallel(mysockcluster)

> SS.output <- cluster.SS(GO.output, organism='mouse',
FDR.lim=0.05)

> stopCluster(mysockcluster)
```

Once again, this will return a list of lists, but with each containing a numeric vector of the total number of gene ontology terms and mean semantic similarity for that cluster. Here is the first list for this data:

```
> SS.output[[1]]
$cluster1
       nGOs          SS
167.0000000   0.2800474

$cluster2
       nGOs          SS
155.0000000   0.2865912
```

As mentioned previously, in addition to the AIC calculation it may be useful to visualize patterns in ontology as the K value changes, to ensure that the most information is extracted from the data. For this, *ROCKS* provides three functions for preparing the data to be plotted.

The first is the GO.plot.prep function, which extracts information on the total number of gene ontologies from all clusters, and does so for all values of K. This is then manipulated to a format suitable for plotting, for which I advise using the scatter.smooth function (although the basic plot function can also be used), which adds a smooth spline. See Results for GO.plot.prep function resultant output plots.

```
> prepdata <- GO.plot.prep(SS.output)
> scatter.smooth(prepdata)
```

The trend in this plot may vary depending on the diversity of the data provided. Ideally, this plot will produce a curved trend in which the sum of the GO terms increases as the K-value increases, before leveling off. From this the user can select an elbow point of K, much like with the AIC values, which seeks to maximize the number of gene ontology terms produced. However, sometimes data can produce a flat linear trend, in which no elbow point can be selected, for which any value of K would be equally appropriate in maximizing gene ontology terms. In this situation, I advise considering semantic similarity

alone (as provided by the next function below) to aid K-selection.

The second function provided is the SS.plot.prep function. This performs the same task as the GO.plot.prep function, except that it compiles the mean semantic similarity values rather than the total gene ontology values. Once again, I advise following this with the scatter.smooth function.

```
> prepdata <- SS.plot.prep(SS.output)
> scatter.smooth(prepdata)
```

This plot should produce a curved trend as the clusters become more specific, and (as before) an elbow point may be informative as one that maximizes semantic similarity.

The third and final of these functions is the combined.plot.prep function. If informative trends have been identified in both of the previous two plots, this function can be useful in aiming to define a single elbow point from a combination of both the total number gene ontology terms and the mean semantic similarity. For each value of K, this takes the product of the sum of gene ontology terms and the normalized semantic similarity (normalized to a range of zero to one). Once again, I advise plotting this with the scatter.smooth function.

```
> prepdata <- combined.plot.prep(SS.output)
> scatter.smooth(prepdata)
```

Finally, *ROCKS* provides functionality for visualising the gene expression patterns of any cluster.

The initial function for this is cluster.plot.prep, which standardizes each gene of the count data to the same scale (to a mean of zero and standard deviation of one) for plotting in a single space, and extracts cluster assignment information for a given K-value, as well as p-values for cluster assignment, from a cluster.wrapper output.

This is then followed by the secondary function, cluster.plot, which plots a graph of standardized expression, with one line per gene, for a given cluster number. Although cluster.plot produces a single plot, when used in a loop this can be used to plot all clusters for a single K-value (as seen in Results).

Here we see the use of the cluster.plot.prep and cluster.plot functions on the example data with a selected K of 18. Firstly plotting the first cluster by itself, and secondly plotting all 18.

(Note: due to the large volume of lines to be drawn, I advise plotting these straight to a file to make the plotting considerably quicker. For example, using the png function.)

```
> png(filename="single-cluster-plot.png")
> cluster.plot(prepdata, K=1)
```

```
> dev.off()


>                                   png(filename="multi-cluster-
plot.png",height=500,width=1000)
> par(mfrow=c(3,6))
> for(i in 1:18){
>     cluster.plot(prepdata, K=i)
> }
> dev.off()
```

Session Information

```
     > sessionInfo()
R version 3.1.1 (2014-07-10)
Platform: x86_64-unknown-linux-gnu (64-bit)
locale:
[1] C
attached base packages:
[1] parallel  stats4    stats     graphics  grDevices utils
datasets
[8] methods   base
other attached packages:
[1] AnnotationDbi_1.28.1 GenomeInfoDb_1.2.3   IRanges_2.0.0
[4] S4Vectors_0.4.0      Biobase_2.26.0       BiocGenerics_0.12.1
[7] RSQLite_1.0.0        DBI_0.3.1
loaded via a namespace (and not attached):
[1] GO.db_3.0.0     GOSemSim_1.24.1 Rcpp_0.11.3
```

5.4 Results

Time Testing

Parallelization becomes a very important tool when inferring appropriate cluster number by internal evaluation techniques on model-based clustering, because the modelling process must be performed multiple times (K) in a single data clustering run, with multiple runs also being performed (K between 2 and n [as selected by the user]). This is of even greater importance for large data sets such as for RNA-Seq, where (for example in our test data sets) the suitable K value could be a double digit number. This clustering process can take hours on a single processor, even when not including any complex K-selection or profiling analyses that the user may wish to follow this with.

Regarding the initial clustering, Tables 1 and 2 show the duration for early stages of analysis. These are reported in clock time (in minutes, as performed on the Oxford University Medical Research Council Functional Genomics Unit computer cluster) rather than process time (as process time is not actually reduced, due to the same processes being run, just simply in parallel). Stages include data preparation (equivalent to MBCluster.Seq's RNASeq.Data function), centre initialization (equivalent to MBCluster.Seq's KmeansPlus.RNASeq function), and EM iterations (equivalent to MBCluster.Seq's Cluster.RNASeq function). These Tables show runs from K of 2 to 20, both with and without the additional forking functionality provided by ROCKS. Forking has been utilized with 20 forks, allowing for some inner

functions of the data preparation and centre initialization stages to be run in parallel. (The EM iteration stage could not be parallelized due to its iterative nature). As the Tables show, the data preparation stage, for which the size of the input data is the main determining factor of duration (independent of K and hence relatively constant within the respective Tables), shows a mean 79.54% reduction with forking (from a mean of 17.64 mins [range: 17.40-18.02 mins] to a mean of 3.61 mins [range: 3.59-3.65 mins]). The centre initialization stage (which does vary with K), shows a constant reduction of approximately 90% (mean: 89.72%, range: 89.46-89.98%, with the longest of these [K=20] reduced from 37.46 to 3.86 minutes). In total, the forking reduces the entire process of clustering of K between 2 and 20 by 73.78% (from 14.23 hours [853.59 minutes] to 3.73 hours [223.83 minutes]). For users with sufficient computational power and cores (for example with a computer cluster) forking for each value of K can be combined with parallel processing of each value of K (with PSOCK). For this same example set, running each K in parallel (20 parallel processes) with 20 forks in each process (totaling 400 cores), reduced the total runtime by 97.75% (from 14.23 hours [853.59 minutes] to 19.20 minutes).

Table 5.1. Non-parallel running of the main stages of model-based clustering and time taken in minutes (as recorded on the Functional Genomics Unit cluster - timings will differ on other systems).

Number of Clusters (K)	The Time Taken for Non-Parallel Runs (Minutes)			
	Data Prep Stage	Centre Initialization Stage	EM Iterations	Total
2	17.52	3.86	1.30	**22.68**
3	17.50	5.63	1.88	**25.01**
4	17.65	7.45	2.47	**27.58**
5	17.51	9.39	2.84	**29.74**
6	17.53	11.38	3.95	**32.85**
7	17.72	13.34	4.31	**35.37**
8	17.46	15.29	4.73	**37.48**
9	17.98	17.08	5.61	**40.68**
10	17.56	18.69	5.14	**41.39**
11	17.40	20.94	6.58	**44.91**
12	17.56	22.88	6.68	**47.12**
13	17.44	24.76	7.49	**49.69**
14	17.92	26.88	8.12	**52.92**
15	17.77	29.00	9.01	**55.78**
16	17.75	30.65	9.52	**57.91**
17	18.02	31.80	9.68	**59.50**
18	17.58	34.00	10.61	**62.18**
19	17.58	35.79	10.81	**64.18**
20	17.70	37.46	11.46	**66.62**
Total	**335.13**	**396.27**	**122.18**	**853.59**

Table 5.2. Parallel running of the main stages of model-based clustering with 20 forks and time taken in minutes (as recorded on the Functional Genomics Unit cluster).

Number of Clusters (K)	The Time Taken for 20 Fork Runs (Minutes)			
	Data Prep Stage	Centre Initialization Stage	EM Iterations	Total
2	3.60	0.40	1.06	**5.06**
3	3.62	0.59	1.22	**5.43**
4	3.62	0.79	2.35	**6.76**
5	3.62	0.98	2.72	**7.32**
6	3.63	1.18	3.38	**8.19**
7	3.64	1.37	3.88	**8.90**
8	3.64	1.56	4.42	**9.63**
9	3.62	1.75	4.18	**9.55**
10	3.65	1.93	4.57	**10.15**
11	3.58	2.16	6.08	**11.82**
12	3.59	2.34	6.91	**12.84**
13	3.61	2.52	7.76	**13.90**
14	3.61	2.70	7.75	**14.05**
15	3.59	2.91	8.23	**14.72**
16	3.59	3.10	9.32	**16.01**
17	3.60	3.29	9.11	**16.00**
18	3.61	3.50	9.79	**16.90**
19	3.62	3.65	10.13	**17.39**
20	3.59	3.86	11.75	**19.20**
Total	**68.64**	**40.59**	**114.60**	**223.83**

In the ROCKS package, the calculation of AIC is integrated into the clustering function. Originally, with MBCluster.Seq, an external function (lglk.cluster) was provided for calculating the log-likelihood of the final model from which the AIC can be calculated. With this being an external function, it involves remodeling the final model and calculating the log-likelihood. An example of this is the application of this function to the Purkinje cell data set post-clustering at just one of the values of K (K=20), which took 18.88 minutes. By adding the AIC calculation to the clustering itself, there is no need to perform remodeling to recalculate log-likelihood, and AIC calculation adds just 8.48 seconds to the clustering function (with the same Purkinje cell data set and K of 20).

Finally, in the case of ROCKS the clustering is followed with more complex GO enrichment and semantic similarity analysis applied to each, which is also going to benefit hugely from parallel processing. For example, in this Purkinje cell data set, testing between a K of 2 and 20 means GO enrichment needs to be run on a total of 209 clusters. Even when the obtaining of gene length information and GO category information is moved outside of the core goseq() function to avoid repetitive use (as mentioned in Methods of Additional External Packages), implementing this without parallel processes took a total of 97.01 minutes. This was reduced to 1.32 minutes when parallelized.

Determining K

For each of our example data sets (see *Testing* in Methods) AIC was calculated with K between 2 and 30 clusters (Figure 5.4).

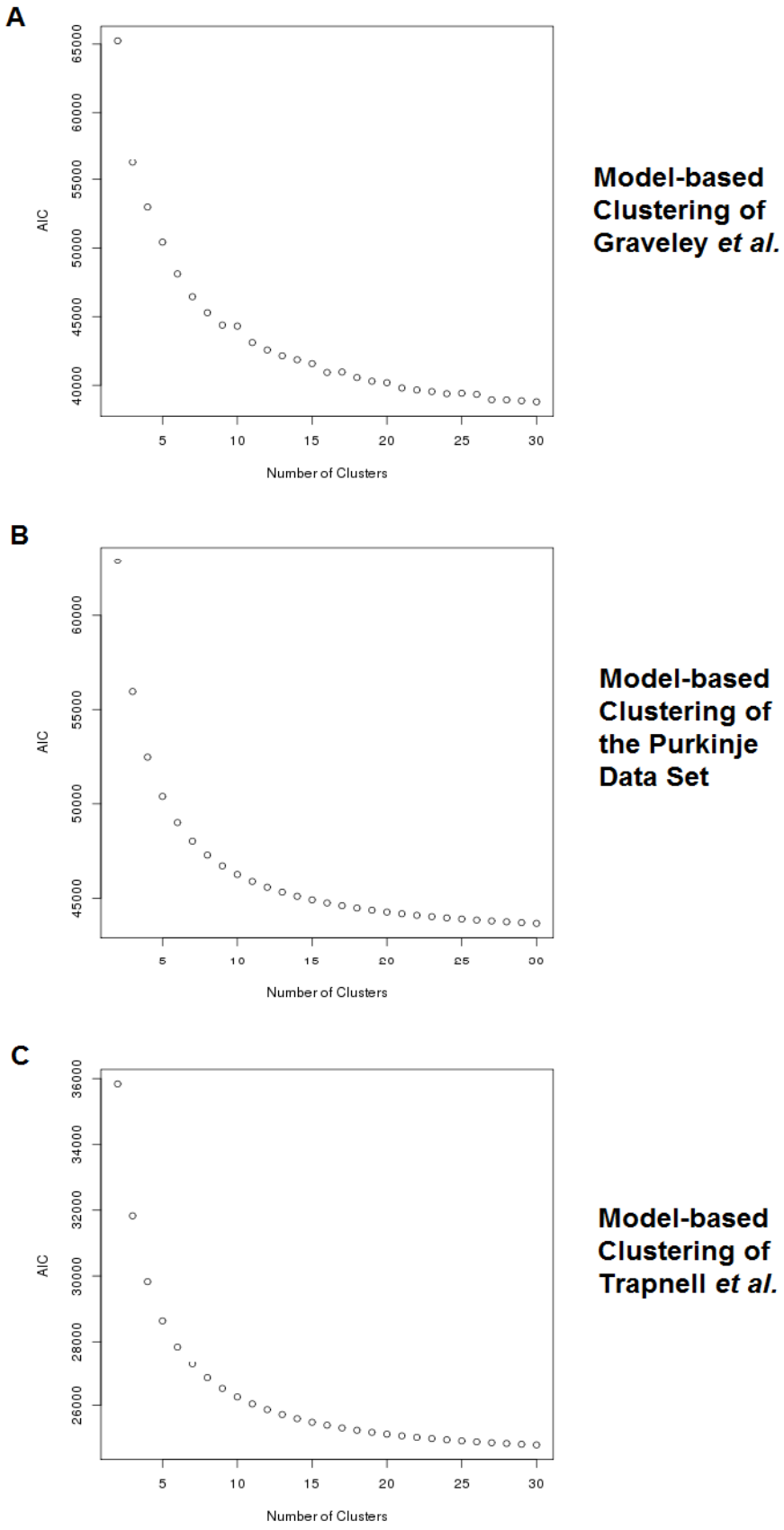


Figure 5.4. AIC calculations from K values of between 2 and 30 for all test data-sets. The elbow of the curve is considered to be the statistically best clustering.

As previously mentioned, AIC can be used to statistically evaluate the clustering model, and to determine the most suitable value of K at the elbow point of the curve. As can be seen in Figure 5.4, determining the elbow point which minimizes the number of clusters and the AIC is not always clear-cut. This is where visualization of the mean semantic similarity of the clusters can assist. The functions provided in this package extract semantic similarity data from the analysis for plotting, and can be seen for each of the example data sets in Figure 5.5. As observed, the semantic similarity follows an intuitive pattern, for which the increase in K leads to increasingly specific clusters, and hence an increasing semantic similarity. Further increases in K then cause this pattern to reach a ceiling, as the clusters become as specific as possible, causing a characteristic curve. Due to the imperfect nature of GO analysis and semantic similarity measures, this is not a perfect curve, but a spline fit to the data captures this pattern. The elbow observed for the curve of the spline can then guide the user as to the smallest value of K for which the semantic similarity is high.

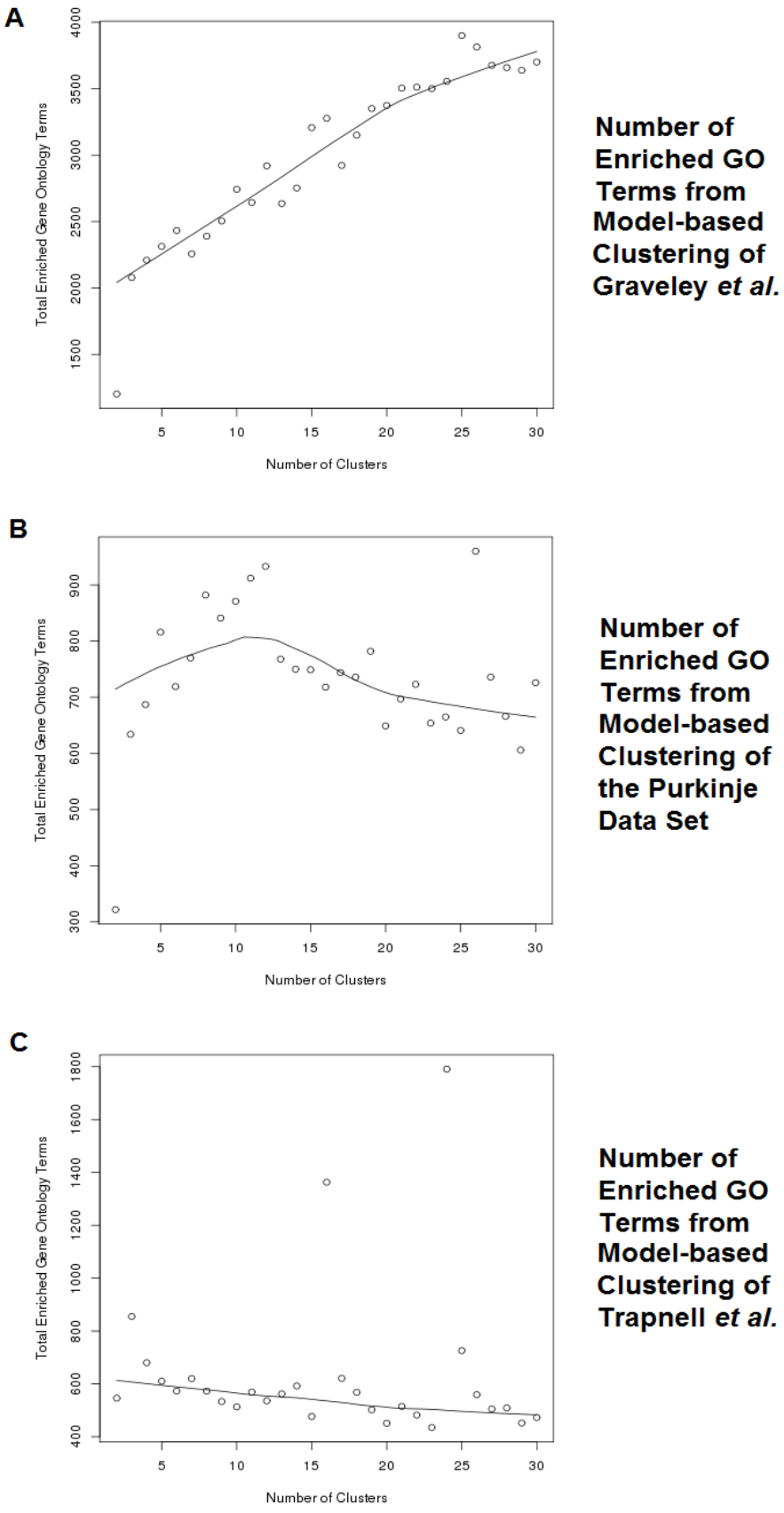


Figure 5.5. The number of statistically significant GO terms reported for K values between 2 and 30 for all test data-sets, with fitted spline.

There are likely to be many applications for users of ROCKS. For those seeking to obtain true biological clusters for specific downstream analysis, this semantic similarity information is likely to be of the greatest importance (used alongside AIC). However, for others the goal may be more to profile the data available in an exploratory manner. For the latter, because GO enrichment analysis is far from perfected, it may be worth adapting the selection of K to maximize the number of enriched GO terms produced and extract as much information as possible from the data. The functions provided by the package extract the number of significantly enriched GOs from the analysis for plotting, and can be seen for each of the example data sets in Figure 5.6. The trend observed for this is less predictable than semantic similarity: as mentioned previously, GO information is imperfect, and in addition to this, the trend identified is likely to be highly influenced by the level of diversity in the specific data set (for example, data which shows changes of expression in very few genes would provide small clusters and a low number of significantly enriched GO terms when increasing cluster size) (see Discussion).

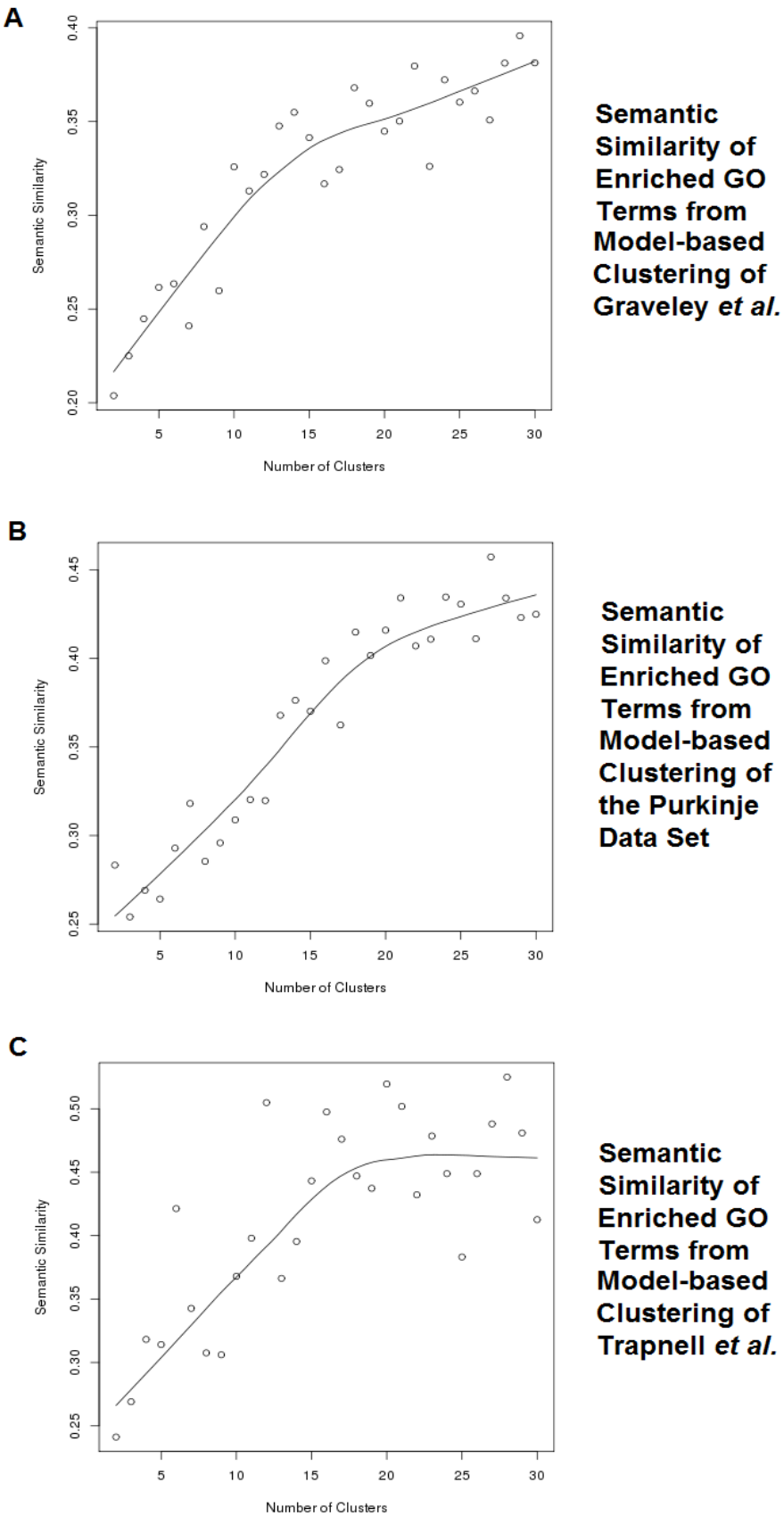


Figure 5.6. The mean semantic similarity of all clusters for K values of between 2 and 30 for all test data-sets, with a fitted spline.

The ROCKS package also includes functionality for combining the two aspects. This multiplies the number of significant GO terms by a standardized semantic similarity. This standardization brings the range of the semantic similarities to between zero and one which, when multiplied by the number of GO terms, provides a metric of semantic similarity adjusted GO count value. This visualization technique can be seen for each of the example data sets in Figure 5.7.

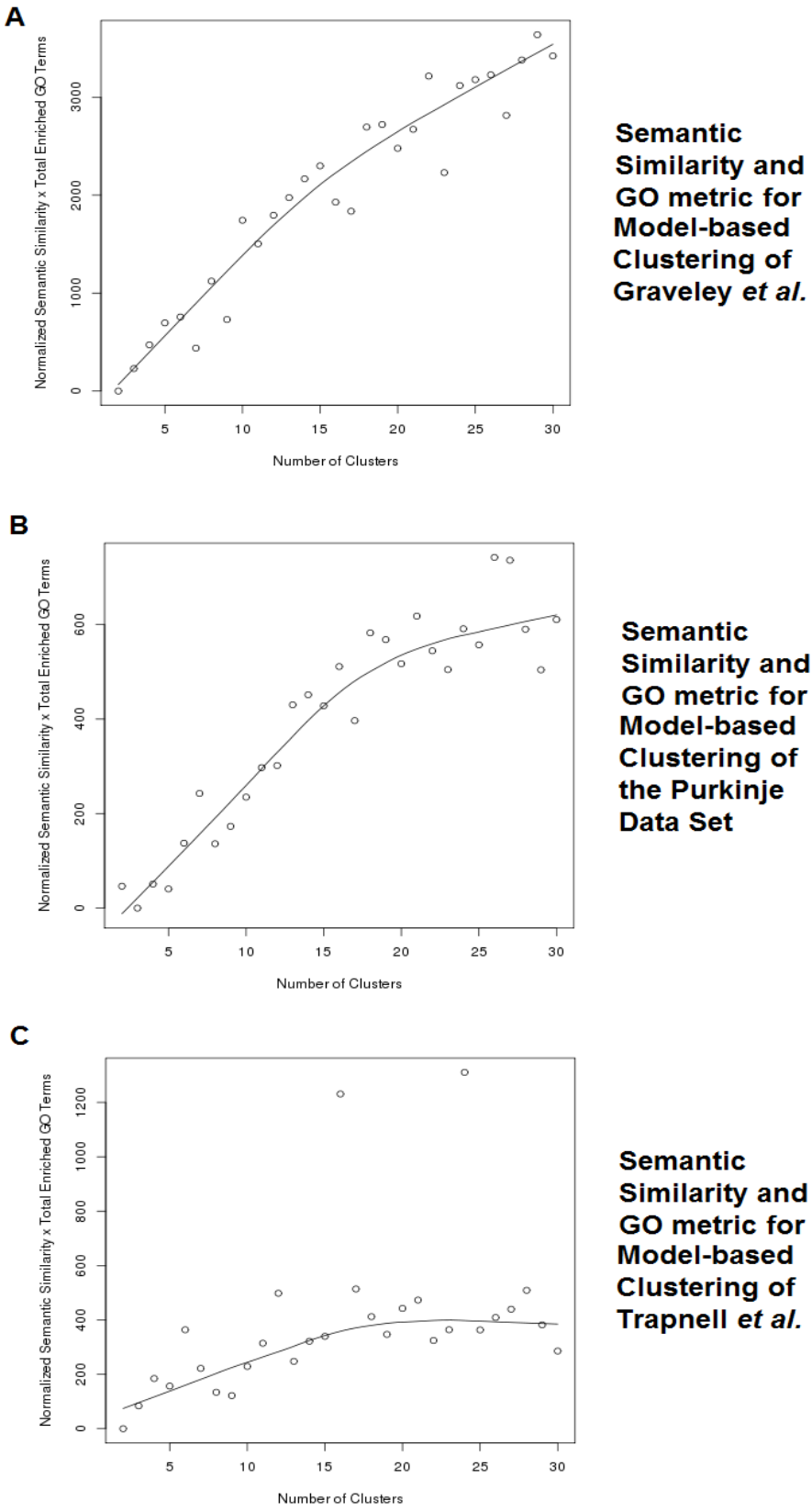


Figure 5.7. A metric combining the number of significant GO terms with the mean semantic similarity of all clusters, reported for K values of between 2 and 30 for all test data-sets, with a fitted spline.

The final set of functions provided by the ROCKS package enables visualization of specified clusters. An example of the use of the plotting function to visualize a single cluster is shown in Figure 5.8. Application of the plotting function to all clusters of a single value of K, for each of the example data sets, is shown in Figure 5.9.

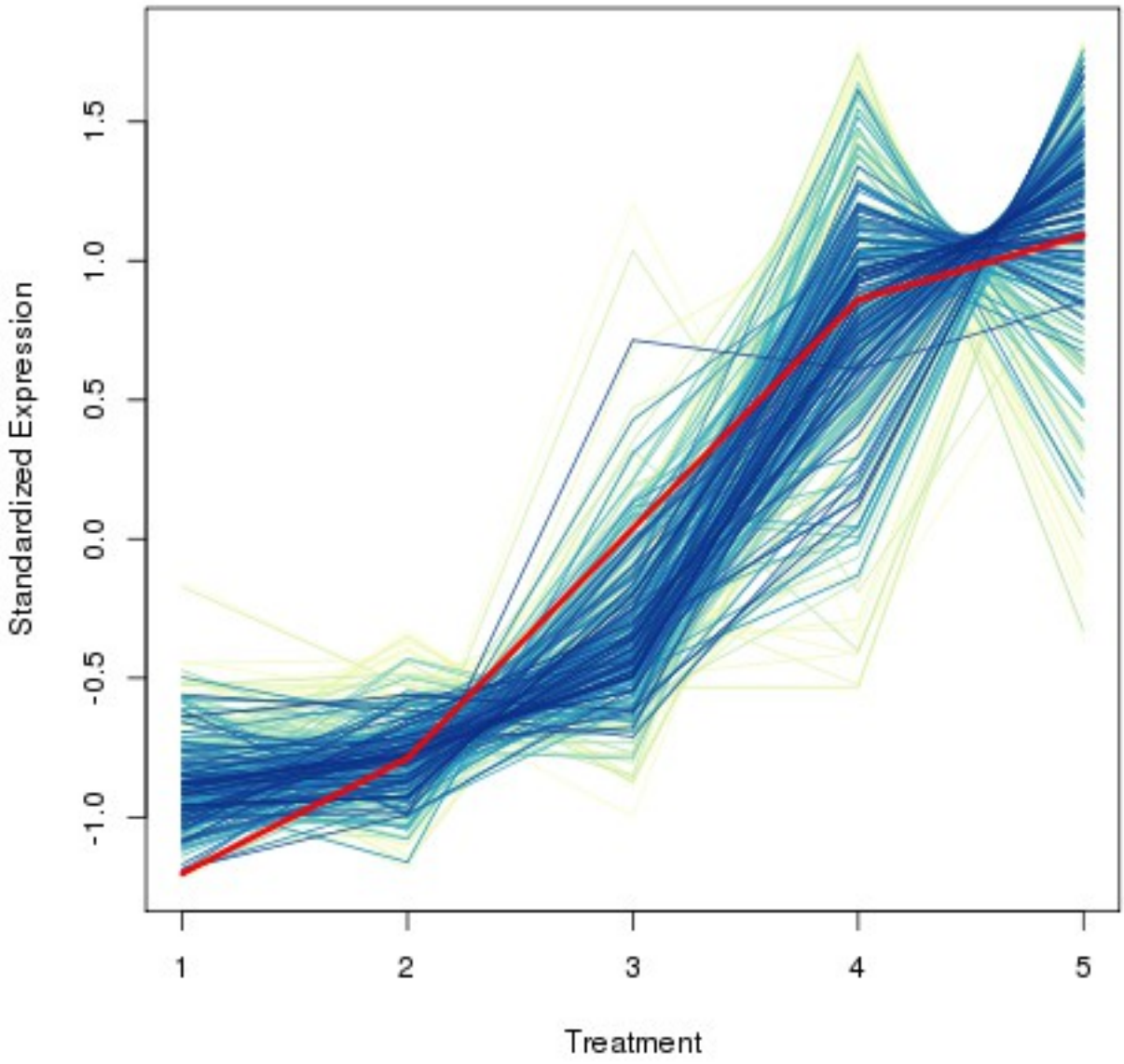
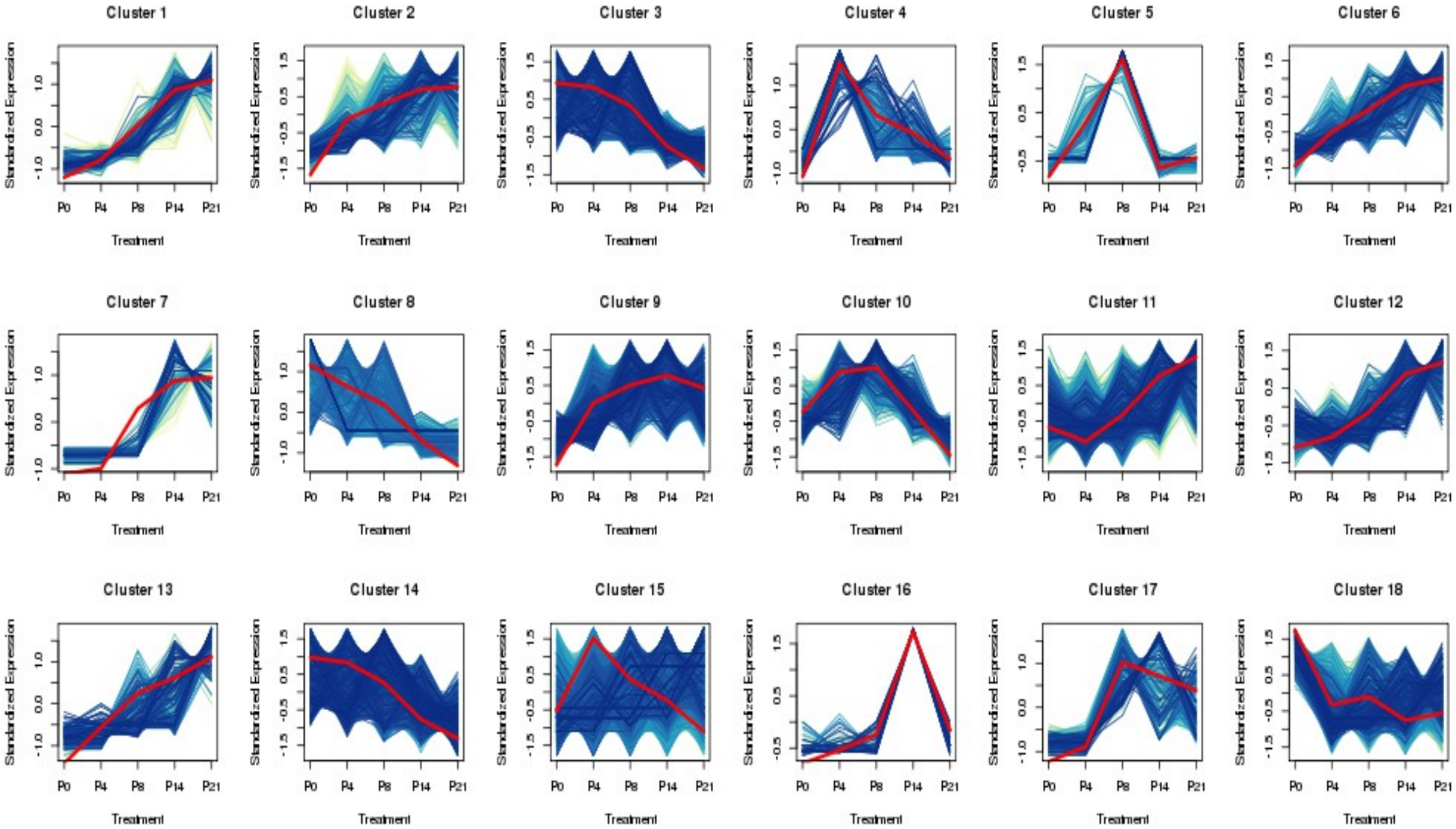
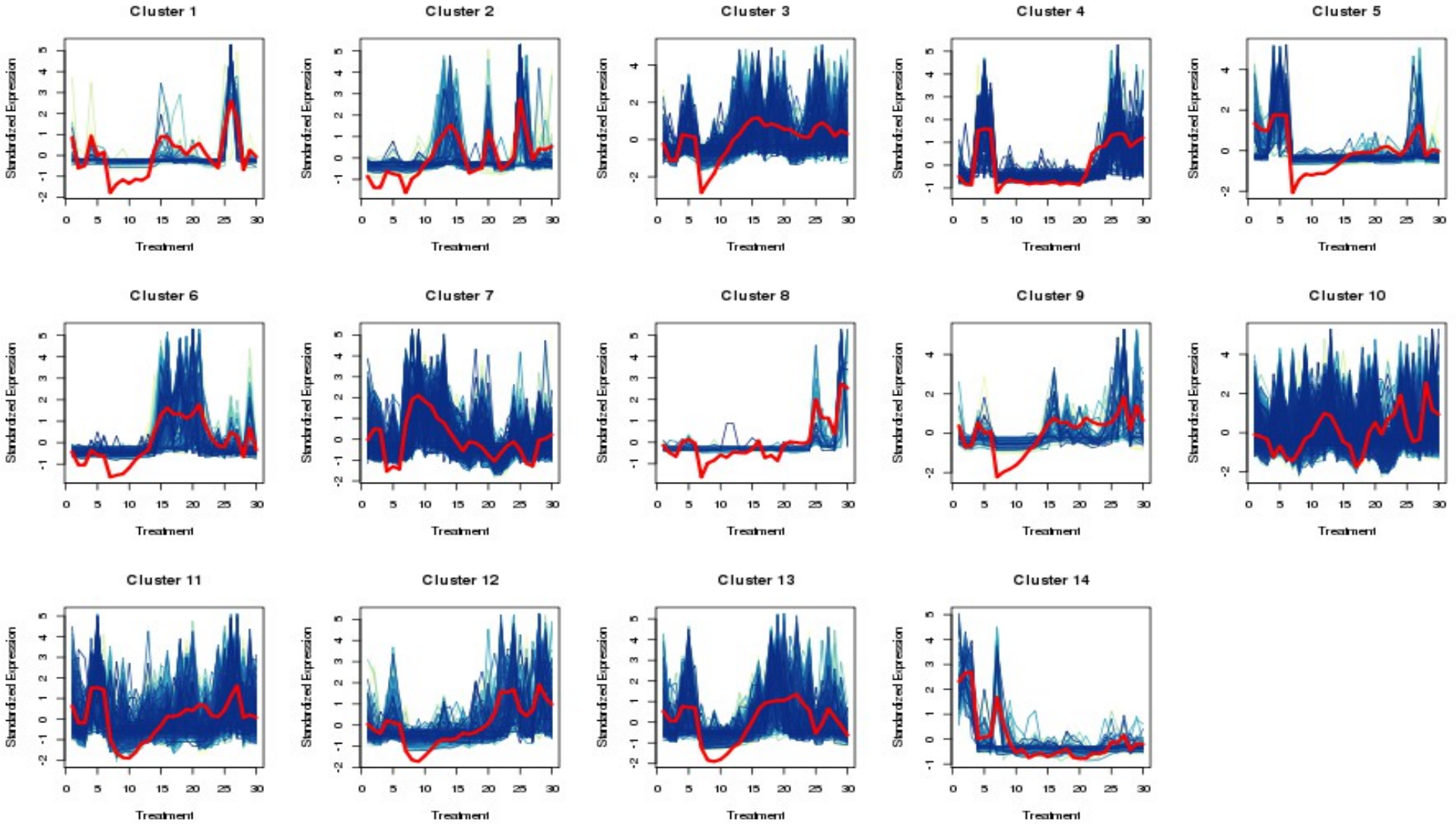


Figure 5.8. Gene expression data from the first cluster of the Purkinje test data set with K=18, plotted with the cluster.plot() plotting function. Each gene within the module is plotted as a line of increasing colour intensity with decreasing p-value (probability of the model explaining the data). Expression values are standardized to a mean of zero and standard deviation of one. The centre of the cluster is displayed in red.


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 14
Cluster 15
Cluster 16
Cluster 17
Cluster 18
Standardized Expression
Treatment
P0
P4
P8
P14
P21

B


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 14
Standardized Expression
Treatment

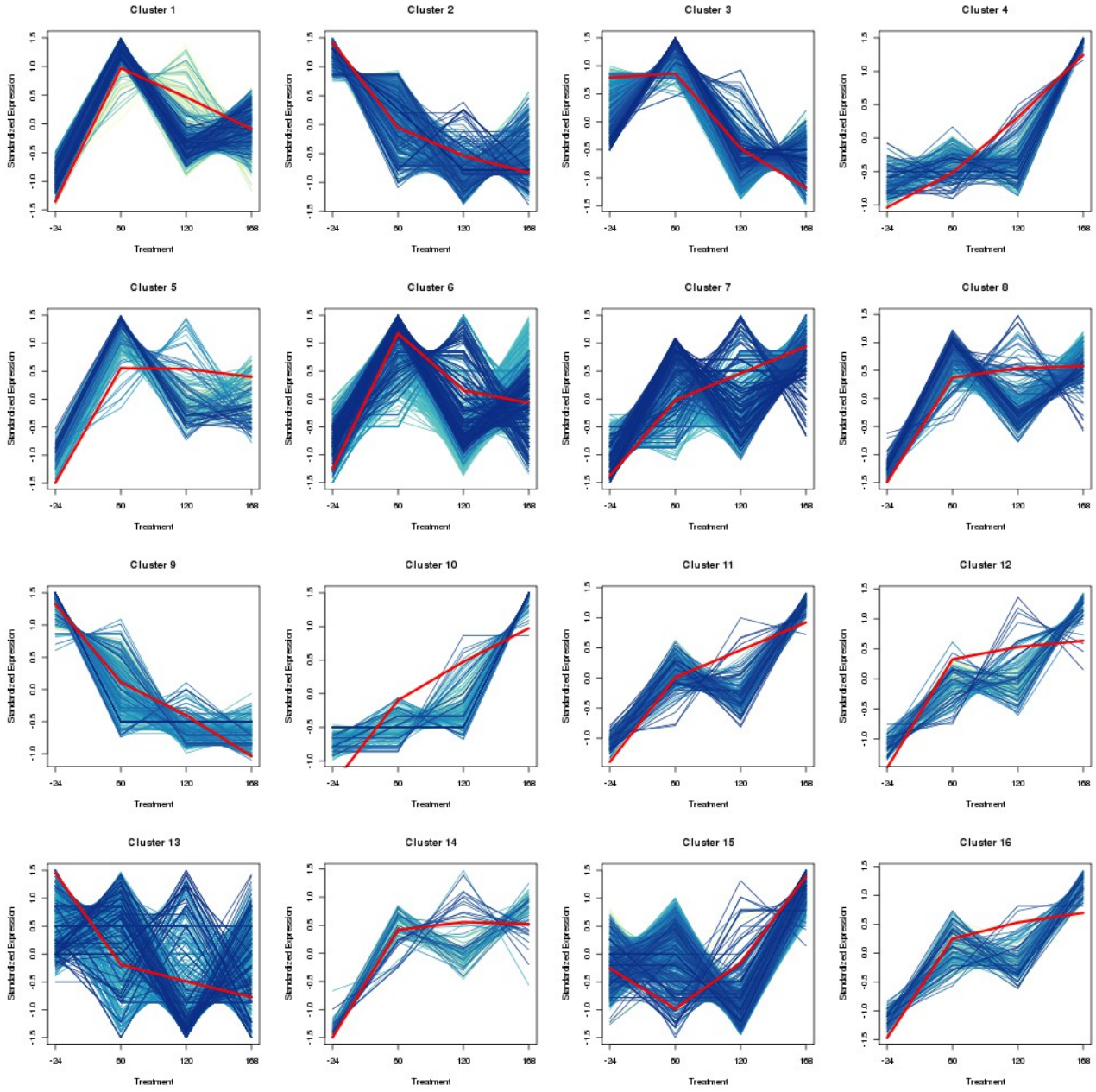


Figure 5.9. Plotting of all clusters, as in Figure 5.8, for a selected value of K for each of the test data sets. Including the Purkinje cell data set with K=18 (A), the Gaveley *et al.* data set with K=14 (B), and the Trapnell *et al.* data set with K=16 (C).

5.5 Discussion

This package provides both a pipeline for easy application of clustering and enrichment analyses (GO and KEGG), with simplified wrapper functions and parallel processing, as well as providing functionality for K-selection methods, firstly with classic AIC, and secondly through integration of GO enrichment and semantic similarity information.

Model-based clustering and AIC are already well established methods for RNA-Seq cluster analysis, and GO enrichment is also a prominent approach for profiling, and providing an overview of the functions in a gene-set. However, using GO enrichment and semantic similarity as a proxy for statistical K-selection terms, such as AIC, could be problematic due to associated issues of reliability. Semantic similarity is dependent on the completeness and quality of ontological information, and much of GO information itself is irregular in nature, including biases in biological research interests and currently available experimental techniques (Pesquita *et al.* 2009). In addition, evidence codes for assigning GO terms to specific genes show further indication of issues, in that the most prevalent GO evidence code is IEA (Inferred from Electric Annotation). This code indicates no direct experimentation has been implemented, nor has manual curation occurred (based on reasonable inference from other assigned GO terms), but the GO term has been assigned based on sequence similarity or model organism annotation databases. Overall, IEA evidence codes have been shown to account for over 97% of all

annotations in the GOA-UniProt database, and less than 2% of GO terms having non-IEA annotations (Camon *et al.* 2004). Conversely, IEA annotations show an increasingly high accuracy and quality, having been reported with precisions of between 91 and 100% (Camon *et al.* 2005). This accuracy is also expected to improve further with increasing experimental annotation of model organisms, and IEA terms are expected, in future, to be widely accepted as reliable (Pesquita *et al.* 2009).

This package aims to maximize the profiling of gene clusters and can enable a substantial increase in novel biological insight compared to some other methods. For example, the use of WGCNA (used to extract a relatively large number of clusters for profiling) in the Purkinje cell dataset yielded very little in terms of significant GO/KEGG/MGI enrichments in these clusters, with 15 of 17 clusters returning no functional annotations (see Chapter 4). In comparison, aiming for a large number of GO terms in this dataset using the ROCKS package, for example with a K of 12 (Figure 5.5), we find that only 1 of 12 clusters returns no enriched GO terms, with the remaining 11 returning a mean of ~85 enriched GO terms, ranging from 31 to 150 terms. This allowed the identification of clusters with dominant roles of cell division, cytoplasmic components, transporter activity, signalling receptor activity, ribosomal subunits, and mitochondrion. Observing the gene expression patterns of each of these clusters alongside this GO enrichment information enables a much greater understanding of the data set and a much more in depth profiling.

Overall, although GO information has its shortcomings, the GO and semantic similarity information that ROCKS retrieves for K-selection can be seen as allowing further guidance within the region of the AIC "elbow", in addition to allowing the user to select a value for K which will return the most (semantically similar) GO terms and therefore more easily assign an overarching profile to a cluster. As the quality of GO terms increases over time, which (as aforementioned) is expected, this method of K-selection will become increasingly precise, potentially allowing K-selection to be based purely on biological over statistical measures.

To conclude, ROCKS capitalizes on clustering and GO methods specific to RNA-Seq to form a single package to perform clustering analysis and profiling suitable for time-series data. This includes parallelization techniques and suitable formation of processes (for example, avoiding unnecessary repetition of certain processes) to increase speed and efficiency in applying these analyses to the increasingly large RNA-Seq data-sets being analysed. In addition to this, ROCKS presents a novel biological method for K-selection to be used alongside the statistical and classical AIC method, which uses the profiling information itself to aid in this selection.

Chapter 6 – Differential Expression and Splicing Analysis Provides Novel Insight into the Transcriptomic Effects of *IFITM3* SNP rs12252 in Influenza

Abstract

Influenza still poses a large problem on a worldwide scale, including hundreds of thousands of deaths annually. Following the poor efficacy of the 2014-15 flu vaccine, studies of the human immune response and underlying susceptibility variants are highly relevant and important. Investigations into the *IFITM3* gene, encoding a transmembrane protein that impedes viral host cell entry, have uncovered a single nucleotide polymorphism (rs12252) with links to severe influenza infection risk. This chapter investigates this phenomenon further, and examines the overall changes in immune response under this variant. Through RNA-Seq of dendritic cells with differing genotypes and infections status, I explore transcriptome-wide changes in expression and splicing. This chapter reports recurring differences in MHC class II cell surface receptor proteins, in addition to observed differences in expression of each rs12252 variant (in heterozygous individuals) when infected. This provides a key pilot study, highlighting important areas for further in-depth investigation.

Publication and contributions

This chapter is based on work carried out as part of a pilot project. Lili Wang, Shokouh Makvandi-Nejad and Tao Dong led procedures prior to RNA-Seq, with Lili Wang performing blood sample extraction, patient genotyping, dendritic cell maturation, influenza infection (when relevant) and fluorescence-activated cell sorting. Sequencing was completed by the Wellcome Trust Centre for Human Genetics. All subsequent computational genomics analysis and interpretation was performed by me, with Wilfried Haerty and Chris P. Ponting advising.

6.1 Introduction

Influenza as a global issue

Influenza, or "the flu", is an infectious disease caused by the influenza virus, which primarily targets cells of the respiratory epithelium. Typical symptoms following infection include cough, fever, and/or general aches, with the infection itself lasting three to seven days. Nevertheless, symptoms following respiratory damage, including the cough, and small-airways disease, may persist for weeks. In addition to this, complications can often occur, with those most commonly leading to hospitalization including laryngotracheobronchitis or pneumonia in children, and chronic pulmonary disease, bacterial pneumonia, and congestive heart failure in adults (Cate, 1987).

The effects of influenza viral infection and the resultant symptoms are by no means trivial, and influenza has been responsible for a large number of deaths worldwide. Early recordings and resultant reports by the World Health Organisation (WHO) put the rate of influenza-related deaths in the region of 250,000 to 500,000 annually in non-pandemic years alone, with between three and five million cases of severe illness in standard annual epidemics. These findings also reported that annual influenza epidemics in the US results in 5-15% of the population being affected by upper respiratory tract infections, and results in a large burden on hospital health care, amounting to costs of between 71 and 167 billion dollars (World Health Organization, 2009). In the developing world, and particularly in the tropics (where it is commonplace for

transmission to occur year-round), the spread of influenza, and the resultant fatality rate, is even greater. This has been found to be especially critical in poorly nourished populations and those with limited access to adequate healthcare, as reported by the WHO during the 2002 Madagascar influenza outbreak. This was also reported to include over 27,000 cases in just three months, with more than 800 deaths despite prompt intervention (World Health Organization, 2009).

Presently, influenza still poses a large problem, including in vulnerable individuals (children and the elderly) in countries where vaccinations are widely available. Statistics on vaccine effectiveness compiled by the US Centers for Disease Control and Prevention from 2004-present show that, although the vaccine seems to have improved over the years, its effectiveness still lies somewhere in the region of half (Table 6.1; Centers for Disease Control and Prevention, 2015).

Table 6.1. A table of compiled studies demonstrating the effectiveness of the influenza vaccine on seasonal influenza in the US, from 2004-2015 (with the exception of season 08-09). As compiled by the Centers for Disease Control and Prevention (2015).

Influenza Season	Reference	Study Site(s)	No. of Patients Studied	Adjusted Overall Vaccine Effectiveness (%)	95% Confidence Interval
2004-05	Belonga *et al.* 2009	WI	762	10	-36, 40
2005-06	Belonga *et al.* 2009	WI	346	21	-52, 59
2006-07	Belonga *et al.* 2009	WI	871	52	22, 70
2007-08	Belonga *et al.* 2011	WI	1914	37	22, 49
2009-10	Griffin *et al.* 2011	WI, MI, NY, TN	6757	56	23, 75
2010-11	Treanor *et al.* 2011	WI, MI, NY, TN	4757	60	53, 66
2011-12	Ohmit *et al.* 2014	WI, MI, PA, TX, WA	4771	47	36, 56
2012-13	McLean *et al.* 2015	WI, MI, PA, TX, WA	6452	49	43, 55
2013-14	Unpublished	WI, MI, PA, TX, WA	5990	51	43, 58
2014-15	Flannery *et al.* 2015	WI, MI, PA, TX, WA	9329	23	14, 31

In addition, the most recent influenza season statistics (2014-15) suggest unprecedentedly low effectiveness of the vaccine. This was developed to target the predominant circulating viruses, influenza A (H3N2) and influenza B (Yamagata lineage), but (as shown in Table 6.1) yielded just 23% effectiveness (with confidence intervals of 14-31%). These data were an early estimate (Flannery et al. 2015), and were further corroborated by a follow-up study of data collected between November 2014 and January 2015, which indicated effectiveness of just 19% (with confidence intervals of 7-29%). The low effectiveness of the vaccine is largely attributed to antigenic drift of one of the targeted strains (influenza A, H3N2) causing a poor match between vaccine and strain. The vaccine was found to be 45% effective (with confidence intervals of 14-65%) against the influenza B strain, but just 18% effective (with confidence intervals of 6-29%) against the influenza A strain (Centers for Disease Control and Prevention, 2015).

In the wake of this recent influenza vaccine mismatch, many global flu experts have called for a delayed selection of strains to vaccinate against, allowing more time to collect data on circulating strains (Marill, 2015). Although this may reduce the likelihood of a mismatching vaccine, viral drift is still unpredictable and the resultant varying vaccine effectiveness highlights the importance of understanding the immune response to influenza. The optimizing of the human immune system may be one of the more successful strategies in minimizing the intensity of the annual influenza season and reducing the rates of hospitalizations and mortality.

The Types and Subtypes of the Influenza Virus

The influenza virus is traditionally classified into types A, B and C, based on their antigenic similarities. Typically, types A and B are the most structurally similar, characterized by a segmented genome of 8 RNAs, whereas type C consists of just 7 (Gao *et al.* 2008). Type C influenza is seen as a relatively mild virus, leading to only minor illness in humans, and not leading to epidemics or pandemics. Types A and B are more severe, with both known to cause epidemics, but only A known to have caused pandemics. The reason for this lies in a widely known characteristic that differentiates type A from B, which is their viable hosts. The natural hosts of type A are birds, but the virus can also infect people, pigs, horses, among other animals, while type B has only been known to infect humans. Type A is also differentiated into subtypes, or strains, based on two glycoproteins on the surface of the virus: hemagglutinin (HA) and neuroaminidase (NA) (e.g. strain H7N2 expresses HA7 and NA2). These subtypes are differentiated based on major antigenic changes in the glycoproteins, known as antigenic shifts, which are thought to occur from reassortment with animal influenza A viruses. This also explains the lack of subtypes seen in influenza B, which is restricted to humans. A lack of major shifts in antigens is also thought to explain the lack of pandemics of influenza B. However both types A and B do have minor antigenic changes (within type A subtypes) due to the accumulation of point mutations, which can lead to antigenic divergence, a process known as antigenic drift (Gupte, 2011). This is a common factor contributing to difficulties in developing an effective and up-to-date vaccine for season influenza.

Gaining Entry into the Host Cell

The entry of a virus into an animal cell is the key initial stage in the virus life cycle, and important in initiating an infection in the host. This is broadly known to occur in three phases: attachment, endocytosis, and uncoating. Regarding the initial attachment of the virus to the host cell, the specificity of viral attachment is mediated by HA, which presents as a homotrimer forming spikes on the viral lipid membrane (with NA largely responsible for the later stages of the viral replication cycle, including the ultimate release of new virions). This initiation of virus infection requires multiple HAs to bind to sialic acids on carbohydrate side chains of cell-surface glycoproteins and glycolipids of the host cell (Sauter et al. 1989; Skehel & Wiley, 2000).

Interestingly, the specificity of this initiation is important in allowing influenza A to perform cross-species infection. Two major linkages are found between the sialic acids and carbohydrate side chains, known as α(2,3) and α(2,6). Human influenza viruses recognize the α(2,6) linkage, avian and equine influenza viruses recognize the α(2,3), and swine influenza viruses recognize both (Skehel & Wiley, 2000), making swine a "mixing vessel" for producing dangerous pathogenic viruses from avian and human influenza (Samji, 2009).

Following initiation and attachment to the host sialic acid residues, endocytosis is induced and the virus enters the cell. The HA precursor consists of two subunits linked by a disulphide bond: HA1, containing the receptor binding

domain, and HA2, containing the fusion peptide (Huang et al. 2003). The endosome in which the virus has entered the cell has a low pH of 5-6, instigating a HA precursor conformational change that exposes the fusion peptide of HA2. This in turn allows fusion of the viral and endosomal membranes (Samji, 2009). Finally, the endosome's low pH also instigates the opening of the M2 ion channel, leading to acidification of the viral core, which in turn leads to release of the viral ribonucleoprotein particles (vRNPs) for entry into the host cell (Pinto & Lamb, 2006). These vRNPs are the entities that proceed to enter the nucleus of the host cell.

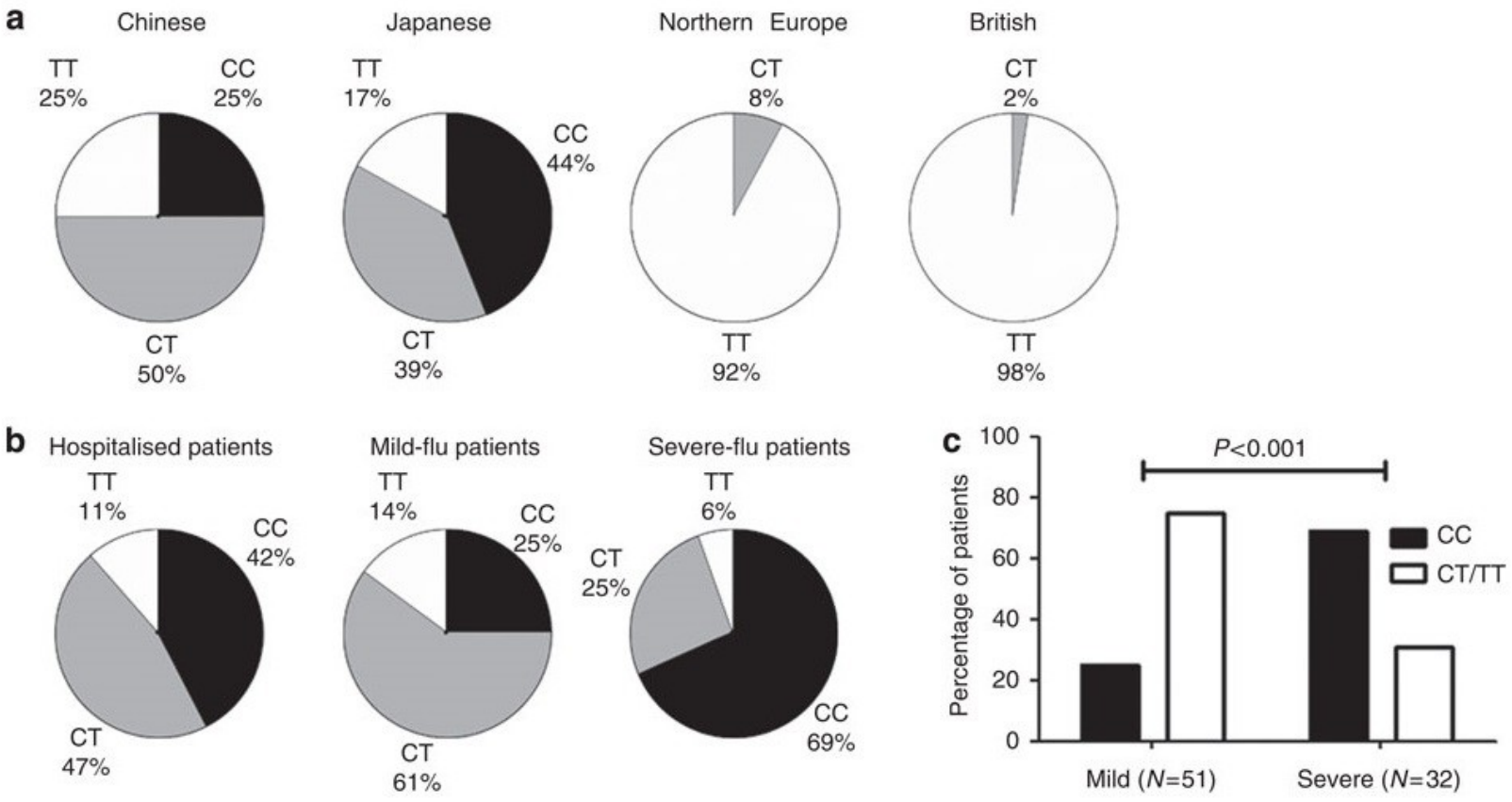
Many methods for prevention of influenza infection are aimed at impeding one of these processes in viral host cell entry. The influenza vaccine is used to stimulate an adaptive immune response that includes the release of neutralizing antibodies specific to the HA and/or NA of the virus. HA-specific antibodies prevent viral entry by binding to two regions of the HA, the head and stalk, with those bound to the head region preventing binding to the sialic acid residues of host cells, and those bound to the stalk region preventing the conformational change required for fusion with the endosome (Kaur *et al.* 2011). Innate immune responses can also play a key role in preventing influenza infection (amongst other viral infections), including anti-viral restriction factors expressed by the host cell to defend against infection. An example of this is the family of InterFeron-Induced TransMembrane (IFITM) proteins, which were initially found to inhibit a range of viruses at a range of virus life cycle stages, including influenza A H1N1, West Nile virus, and dengue virus (Brass *et al.* 2009). Specifically, the IFITM3 protein has been further

characterized and found to play a key role in impeding viral host cell entry, by interfering with the fusion stage (Desai et al. 2014).

Investigating IFITM3 and the rs12252 SNP

As previously mentioned, IFITM3 plays a key role in impeding viral host cell entry. It was also recently shown that a Single Nucleotide Polymorphism (SNP), rs12252, of *IFITM3* occurs in increased frequency in influenza patients with severe infection (Everitt *et al.* 2012). This is thought to be due to the ancestral (C) allele encoding a C-terminally truncated IFITM3 protein. It has been shown that homozygosity of this variant (CC) among patients of European ancestry (which is rare, at ~0.6% population frequency) confers a high risk of severe symptoms. This variant has now also been found to be much more common in those of Chinese ancestry (25%), with a very high occurrence in patients suffering from severe influenza infection (69%) (Figure 6.1; Zhang *et al.* 2013).

A


a
Chinese
TT 25%
CC 25%
CT 50%
Japanese
TT 17%
CC 44%
CT 39%
Northern Europe
CT 8%
TT 92%
British
CT 2%
TT 98%
b
Hospitalised patients
TT 11%
CC 42%
CT 47%
Mild-flu patients
TT 14%
CC 25%
CT 61%
Severe-flu patients
TT 6%
CT 25%
CC 69%
c
P<0.001
Percentage of patients
100
80
60
40
20
0
CC
CT/TT
Mild (N=51)
Severe (N=32)

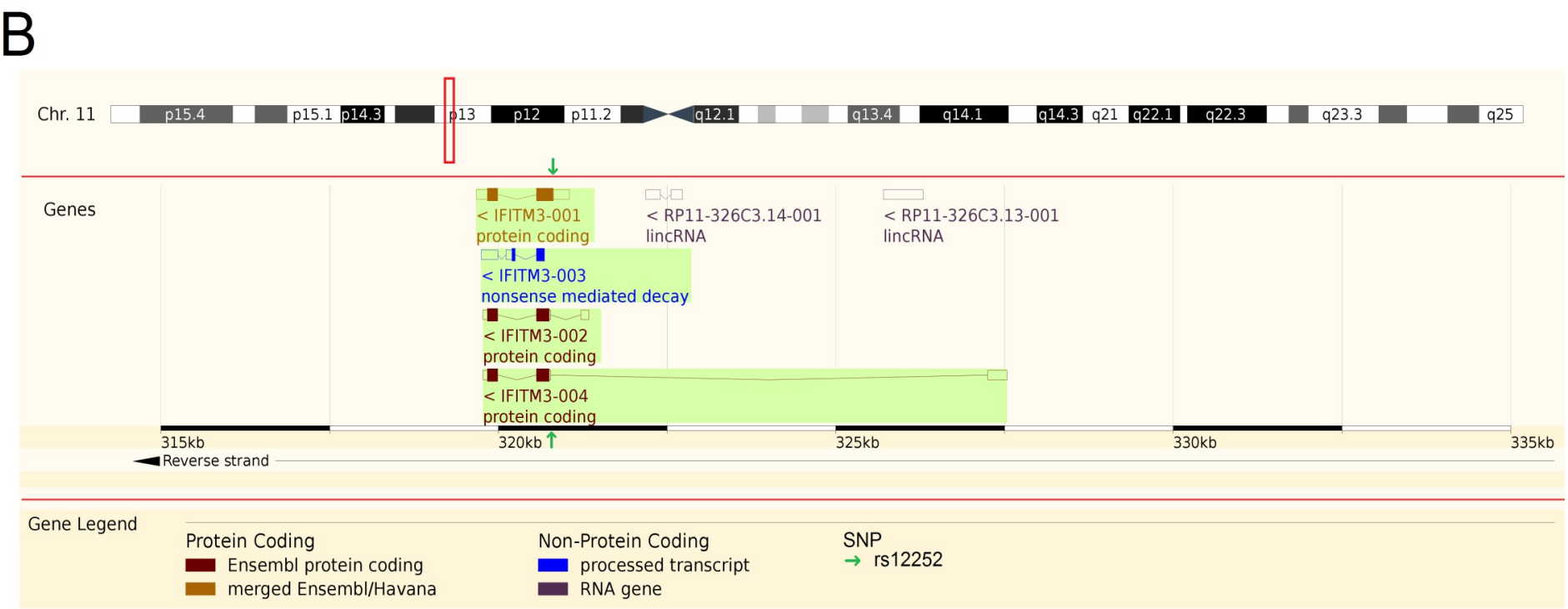


Figure 6.1. Allele frequency, influenza infection severity, and position of *IFITM3* and the rs12252 variant. Panel (A) is reproduced from the rs12252 study by Zhang *et al.* (2013). (Aa) displays the variation of allele frequencies among populations with data from the 1,000 genomes project. (Ab) displays allele frequencies from patients in the Beijing You'an Hospital with confirmed influenza A H1N1 infection. (Ac) displays the proportion of those mild vs. severe patients with each genotype. Panel (B) displays the IFITM3 gene (and isoforms) and rs12252 locus, modified from the Ensembl genome browser (Flicek *et al.* 2014).

It is possible that the higher frequency of the CC variant in *IFITM3* among South East Asia may be facilitating the introduction of new viral subtypes into the population. Therefore it is important to further elucidate the role of IFITM3 and to uncover any additional roles this may have in co-ordination of the immune response, and the effects of the rs12252 variant on these roles. In the wake of disappointing efficacy of the most recent influenza vaccine, and with influenza still remaining a serious worldwide issue, investigating the innate immune response and its ability to prevent influenza host-cell entry could help prompt the development of better therapeutic strategies.

Consequently, this study focuses on differential immunological and host genetic mechanisms due to the rs12252 variant, with the aim of exploring any mechanisms associated with *IFITM3* that may be lost, and any mechanisms that may compensate for the CC genotype. To address this, the experimental design was centered on monocyte samples from individuals of differing rs12252 allele (homozygote C, homozygote T, and heterozygote), and comparing samples infected with influenza against uninfected controls. This was performed as a pilot study to investigate the hypothesis that individuals with the flu-susceptible rs12252-C allele would mount a different immune response to those of rs12252-T, possibly due to an increased rate of cell infection causing the need for a greater response, or due to reliance on other immunological methods to prevent infection in the first place (relying less on the IFITM3 protein).

Analyses throughout this study were performed on data acquired by collaborators Lili Wang, Shokouh Makvandi-Nejad, and Tao Dong, who located 9 donors of varying genotype (3 CC, 3 CT, and 3 TT) and isolated monocytes from blood samples (50ml per sample), which were subsequently differentiated into dendritic cells. These dendritic cells had been used for both infected and uninfected cohorts, through splitting of the samples into two populations and infecting one with the Influenza A H7N9 virus for two hours (GFP was also used as a marker for virus infection, from which GFP positive cells (virus infected cells) were then FAC sorted for RNA extraction). This provided RNA (extracted from all 18 samples - 3 biological replicates of 3 genotypes, infected and uninfected) for poly-A selected 100bp paired-end sequencing, and provided the data to be analyzed throughout this project.

6.2 Methods

Read Processing and Alignment

Quality control was implemented with a series of tools designed to ensure high read quality and lack of adapter sequences, including FastQC (Andrews, 2010), the FASTQ Quality Trimmer and Quality Filter of the FastX-toolkit (Hannon Lab, 2009), and Cutadapt (Martin,2011). These were used to assess the data quality, remove sequences and reads of low quality (below a phred score of 20), and remove adapter associated sequences, respectfully. In removing sequence/trimming of reads, those with a resulting length shorter than 30bp were discarded.

Alignment was implemented with the splice-aware short read aligner Gsnap (Wu & Watanabe, 2005). Insert size and deviation were determined from statistical reports returned by the sequencing machine. Reads were aligned to the human reference genome GRCh38/hg38. Viral sequences were also added to the reference genome (Bao *et al.* 2008), to allow assessment of quantities of viral RNAs. To provide extra confidence in the alignments, only those reads that were uniquely aligned (i.e. aligning to just one position in the genome) were retained. Count tables were then produced with HTSeq (on intersection strict settings) (Anders *et al.* 2015) to count reads mapped to Ensembl gene models. When investigating the reads aligned to the *IFITM3* variant loci, including determining the quantity of reads of each genotype, duplicates were also removed, using Picard Tools MarkDuplicates (Wysoker *et al.* 2013), so as

to ensure any direct comparison of quantities was not biased by PCR artifacts.

The mappability of reads to the *IFITM3* gene is important in ensuring that quantification of this gene's allele-specific expression is accurate. Overall, *IFITM3* is a mappable gene. One study by Robert and Watson (2015), which evaluated the alignment and counting of 1000 simulated reads for each gene of the human genome, found that 958 of 1000 IFITM3 simulated reads could be aligned and counted with STAR aligner (Dobin *et al*. 2013) (a seed and extend aligner similar to GSNAP; Wu & Watanabe, 2005) and HTSeq-count (intersection strict settings) (Anders *et al.* 2015). This compares to a genome-wide mean of ~864 per gene with the same methods.

Preliminary analyses also returned some irregular results, in which some genes associated with the immune system were found to be differentially expressed in both directions, and thus appearing to be significantly upregulated both in a subset of the samples, and outside of the subset (see Differential Expression Analysis). This was found to be due to alternative sequence representations in the MHC region on chromosome 6 (providing multiple versions/copies of the same genes). This is a feature of the Ensembl human reference genome assembly, which includes 261 alternative loci scaffolds, mainly in the MHC region on chromosome 6, and the LRC/KIR complex on chromosome 19. Consequently, count tables were reproduced with those reads that aligned uniquely to these alternative sequences added to those aligned to the initial sequences.

Differential Expression Analysis

Differential expression was analysed with DESeq2 (Love *et al.* 2014). DESeq2 provides functionality for a paired comparison (identifying differentially expressed genes between two groups) in the form of a Wald test, and for multiple comparisons (classically used for time-series data) in the form of a likelihood ratio test.

The Wald testing was used to identify statistically differentially expressed genes across infection status (comparing infected and uninfected samples), and across *IFITM3* genotypes (comparing CC and TT). CT genotypes were excluded from this testing due to the limits of the paired testing, and with CC vs. TT likely to detect the most differences. This also circumvents the assumption of an additive or dominant effect when testing across all genotypes. Finally, Wald testing was also used to assess the interaction between the genotypes and infection status, testing for a difference of differences in expression (i.e. the difference of expression between the genotypes being significantly different depending on the infection status).

The likelihood ratio testing was used in an attempt to provide greater power to the testing of the interaction term. In time-series data, likelihood ratio testing is used to determine whether there is a significant difference in an expression pattern over time, using a constant expression as the null model. In using this for genotype data, the *IFITM3* genotypes CC, CT and TT, can be treated as time-points 1, 2 and 3. Although this assumes an additive model for

expression, it can increase the power of the testing due to inclusion of CT replicates. This was used once more to test for the interaction, looking for a significant difference in expression over genotype in the infected samples, with the expression over genotype in the uninfected samples as the null model.

Differential Splicing Analysis

Differential splicing analysis was analysed with BitSeq (Glaus *et al.* 2012). BitSeq is a tool that uses a probabilistic approach to infer alternative splicing, and does so through the application of a Bayesian approach, to estimate expression levels of full transcripts. BitSeq was found to be one of the most accurate methods for these probabilistic methods or estimation (SEQC/MAQC-III Consortium, 2014). BitSeq's methods include two algorithmic stages: estimation of transcript expression, and differential expression assessment. The transcript expression estimation stage is computed on reads aligned to a transcriptome, and consists of calculating the probability of each read originating from a given transcript fragment. This is used to form a probability distribution of each transcript fragment given the reads, from which the Markov Chain Monte Carlo (MCMC) method is used to generate samples (and subsequent estimates of expression) for each transcript (Figure 6.2). The differential expression assessment stage is then implemented, comparing the distributions represented by MCMC samples (the posterior expression distributions, as displayed in Figure 6.2), through a one sided Bayesian test to determine the Probability of Positive Log Ratio (PPLR). Although this is a good measure for comparing the distributions, it does not provide a standardized p-

value (or a value suitable for multiple testing corrections). Therefore, a stringent threshold for PPLR was used, with MCMC samples deemed significantly differentially spliced when returning a PPLR of exactly 0 or 1 (the very limits of the PPLR metric).

As with previous methods, this allows for paired differential expression testing, and was applied to investigate differences due to infection status (comparing infected and uninfected samples), and due to *IFITM3* genotypes (comparing CC and TT).

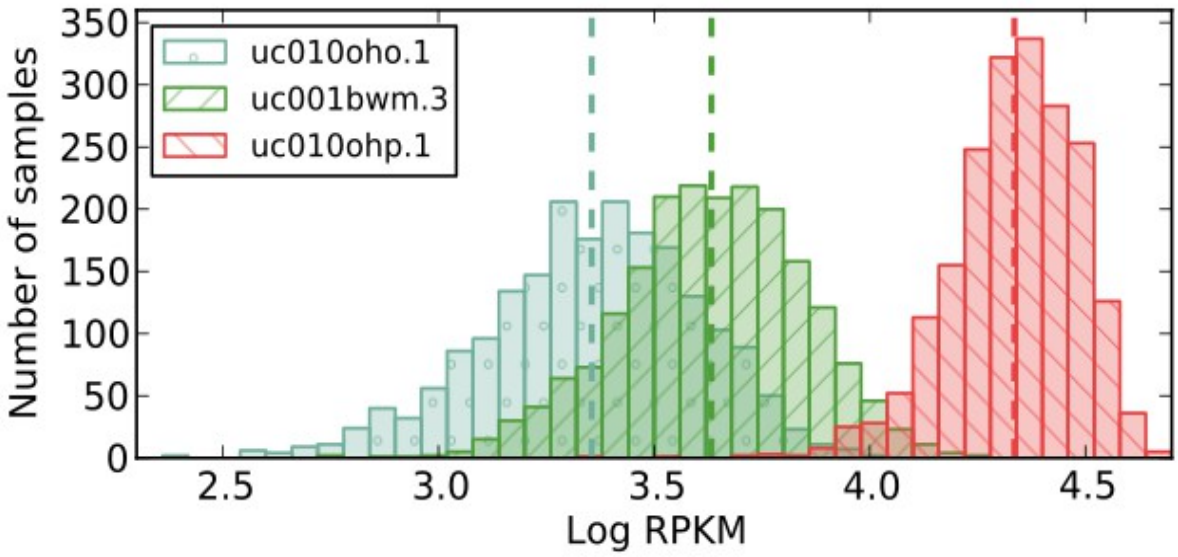


Figure 6.2. Example histograms of MCMC samples, illustrating the marginal posterior distributions for expression levels of three transcripts of one gene. Modified from Figure 4C of the documents presenting BitSeq by Glaus *et al.* (2012).

Differential Exon Analysis

Differential exon analysis was also applied, with the aim of complementing differential splicing results, and possibly pinpointing specific exons that may also indicate changes in splicing. The differential expression of exons was analysed with DEXSeq (Anders *et al.* 2012), which provides functionality similar to that of DESeq/DESeq2: however, it does so for read count data of individual exons, and hence allows differential expression analysis at an exonic level. In turn, this then allows for identification of alternative splicing events. DEXSeq provides functionality for paired differential expression testing (again, similar to that of DESeq), and was applied across *IFITM3* genotypes (comparing CC and TT), to both complement BitSeq data, and visualize splicing events (for example *IFITM3*).

GO and KEGG Enrichment

Functional annotation was implemented on differential expression analyses through the GOSeq package (Young et al. 2010), including enrichment testing for GO terms (GOSeq terms and hg38 Ensembl IDs) and for KEGG pathways (Kanehisa & Goto, 2000). The background used for testing consisted of all genes with non-zero expression in any sample. Resultant significantly enriched GO terms were then filtered through REVIGO (Supek et al. 2011), removing redundant GO terms, with a moderately strict threshold for similarity value of 0.5.

6.3 Results

Read Processing and Alignment

RNA-Sequencing was applied to high quality RNA, with a mean RIN value of 9.8 across samples, ranging from 9.4 to 10. Sequencing was done at relatively high depth, producing a mean of ~131 million reads per replicate, ranging from ~90-159 million. This was to allow for more complex downstream analyses such as investigations of alternative splicing. The data proved to be of extremely high quality, with adapter and quality trimming/filtering leading to disposal of very few reads. Throughout all quality control procedures, a mean of 99.98% of all reads were retained (with a range of just 99.9795% to 99.9804%). Histograms representing the proportion of reads removed due to quality control procedures, and the frequency of samples for these proportions, can be observed in Figure 6.3. Alignment of these reads to the reference genome, including the human genome and viral transcripts, also indicated the data to be of good quality, with a mean of 92.77% of the reads successfully aligning, ranging from 92.35% to 93.23%. Of those aligning, a mean of 78.82% of reads aligned to one location only, ranging from 74.03% to 83.28%. Only these uniquely aligning reads were retained for downstream analysis, to ensure robust expression estimations. The final proportions of reads retained at each of these stages are represented in Figure 6.4, including a break-down by *IFITM3* genotype and infection status. The results for each sample for all of these metrics relating to data quality and alignment procedures can also be found in Table 6.2.

As part of the quality control procedure, the alignment of reads over the *IFITM3* gene was visualized, including the sequences of reads aligned across the rs12252 variant. Unfortunately, although most of the samples were concordant with the expected genotype (with nearly all reads in CC samples containing a C at rs12252, nearly all reads in TT samples containing a T, and reads in CT samples containing a mixture of the two), one of the CT samples (CT replicate 2), proved to have nearly all reads with rs12252-C, suggesting an incorrect genotyping of the original donor (Figure C1). This sample was re-genotyped and shown to be the homozygote CC genotype. Therefore, for all downstream analyses, this sample was changed to a CC *IFITM3* genotype replicate (replicate 4).

The alignment of reads to the viral transcriptome was also assessed in detail, investigating alignment to viral transcripts and to GFP, which was used to select cells with successful infection. As expected, a large proportion (16.42% of the uniquely aligned reads) of reads aligned to these transcripts in the samples which were infected, with next to none (0.01% of the uniquely aligned reads) aligning in the uninfected samples (Figure 6.5A). The abundance of reads aligned to each of the viral transcripts individually was also investigated, as shown in Figure 6.5B. This displays the alignment of reads across the viral transcriptome, and shows an especially high alignment to the gene encoding the NS1 influenza protein. The NS1 protein is responsible for preventing polyadenylation of cellular mRNAs, and does so to interrupt antiviral responses from the host (e.g. of interferon mRNAs) (Nemeroff *et al.* 1998; Wang *et al.*

2000). Interestingly, this expression is also observed at its highest level in the CC *IFITM3* genotype (influenza-susceptible phenotype), followed by the CT, and finally the TT (although these differences are not statistically significant). Multiple viral transcripts also showed zero read alignment: this was due to multi-mapping reads which were discarded. The alignment of non-unique reads (not retained in downstream analyses) across the viral transcriptome, can be observed in Figure 6.5C, showing alignment to those at zero reads-per-million in Figure 6.5B.

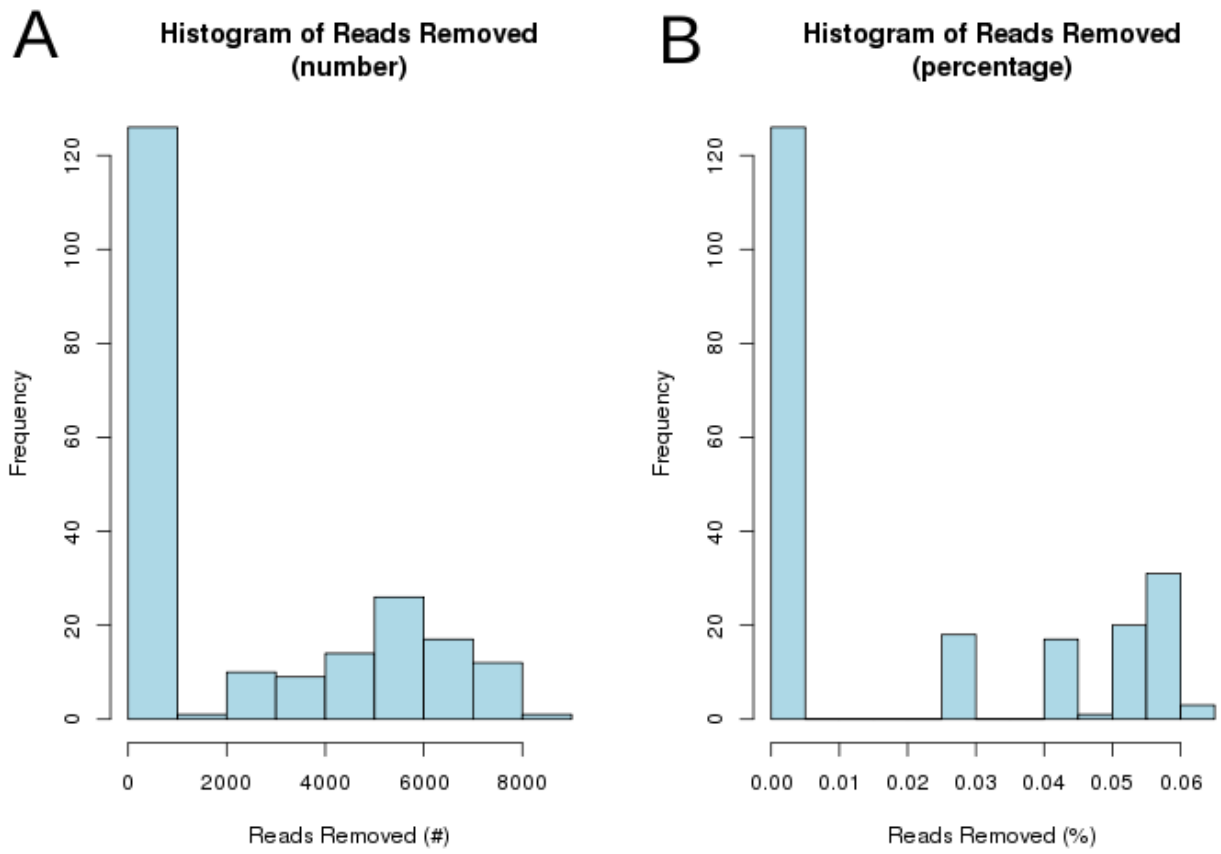


Figure 6.3. Histograms representing the frequency of FastQ files and the number of reads removed (A), also represented as the percentage of reads removed (B), due to quality control procedures. This was performed before merging of FastQ files across lanes, and so represents a total of 216 files.

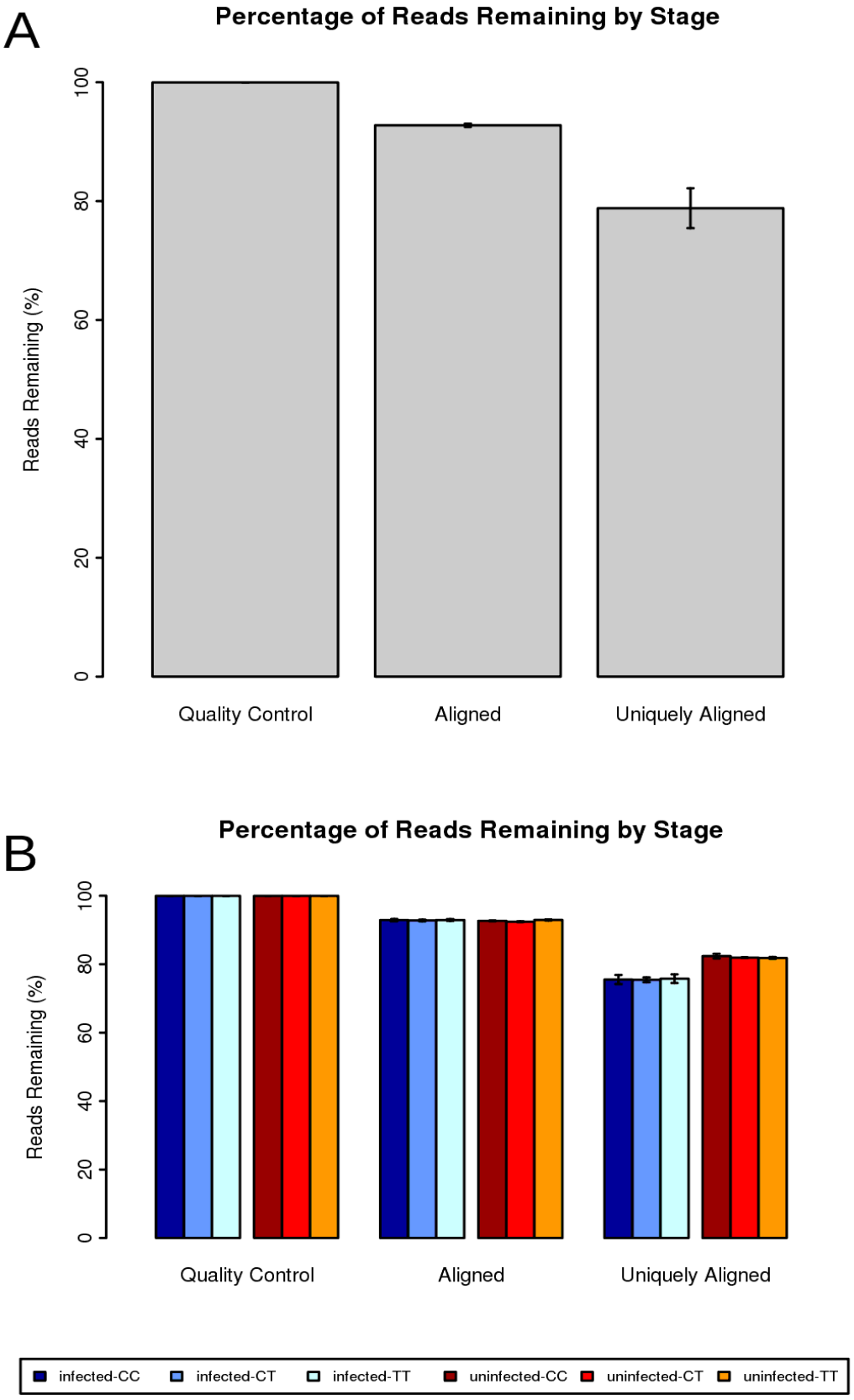
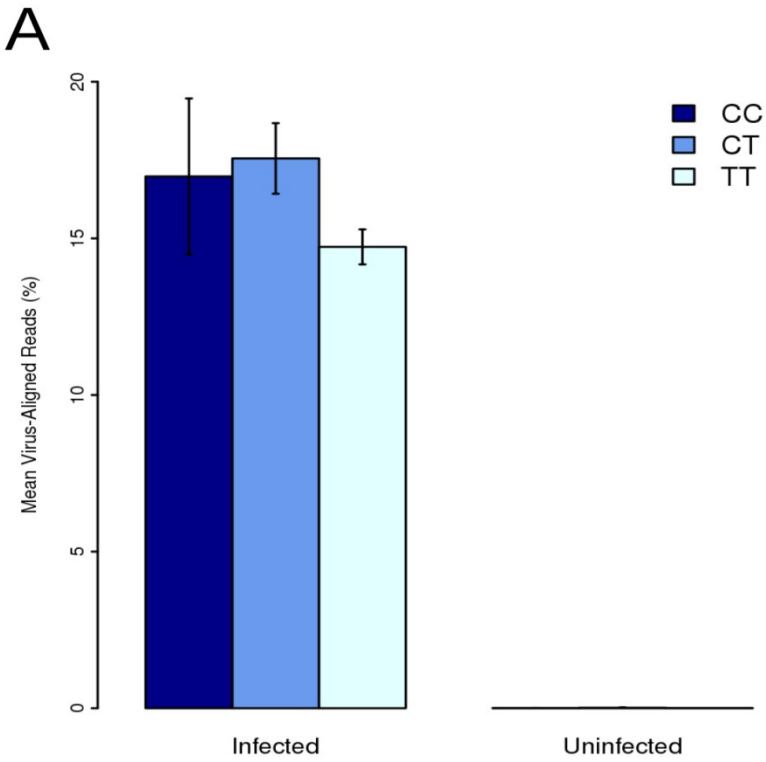
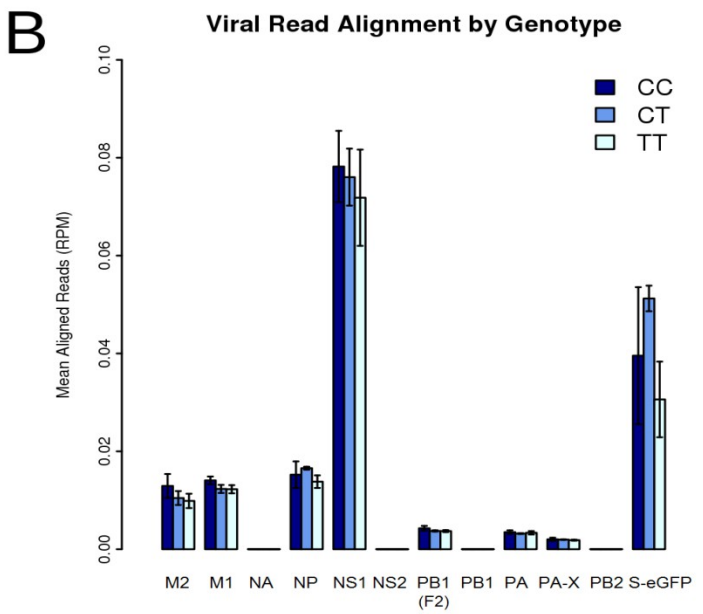


Figure 6.4. Bar charts of proportions of reads remaining at each stage of data preparation. This is shown as a total mean (A), and split by genotype/infection status (B). Whiskers indicate standard deviation.

Table 6.2. Information relating to sample quality (Q RNA and RIN), as well as the information visualized in Figures 3 and 4 for each sample and replicate. (CT replicate 2 is subsequently CC replicate 4)

Genotype	Biological Replicate	Infection Status	Q RNA [ug]	RIN	Raw Reads (millions)	Pass QC (%)	Of QC, Reads Aligned (%)	Of Aligned, Uniquely Aligned (%)
CC	1	infected	1.22	9.7	135685990	99.98	92.50	75.38
		uninfected	3.25	10.0	106544000	99.98	92.56	81.78
	2	infected	6.36	9.5	122700032	99.98	92.98	77.19
		uninfected	8.65	10.0	89696940	99.98	92.70	83.28
	3	infected	7.96	9.8	143393426	99.98	93.23	74.03
		uninfected	6.70	9.7	125647056	99.98	92.74	82.06
CT	1	infected	3.61	9.9	125142154	99.98	93.01	75.38
		uninfected	5.22	10.0	141194854	99.98	92.35	81.83
	2	infected	1.98	9.4	131397550	99.98	92.46	76.31
		uninfected	7.64	10.0	132074322	99.98	92.59	82.03
	3	infected	5.99	9.7	159148822	99.98	92.92	74.66
		uninfected	2.73	10.0	118312806	99.98	92.35	81.92
TT	1	infected	13.01	9.8	156466072	99.98	92.98	76.20
		uninfected	7.44	9.5	127436468	99.98	92.84	81.50
	2	infected	1.34	10.0	152091702	99.98	92.57	74.11
		uninfected	3.34	10.0	117313940	99.98	92.80	81.90
	3	infected	1.69	10.0	120879270	99.98	93.20	77.04
		uninfected	4.74	9.9	154351742	99.98	93.14	82.12


A
Mean Virus-Aligned Reads (%)
0
5
10
15
20
CC
CT
TT
Infected
Uninfected


B
Viral Read Alignment by Genotype
Mean Aligned Reads (RPM)
0.00
0.02
0.04
0.06
0.08
0.10
CC
CT
TT
M2
M1
NA
NP
NS1
NS2
PB1 (F2)
PB1
PA
PA-X
PB2
S-eGFP

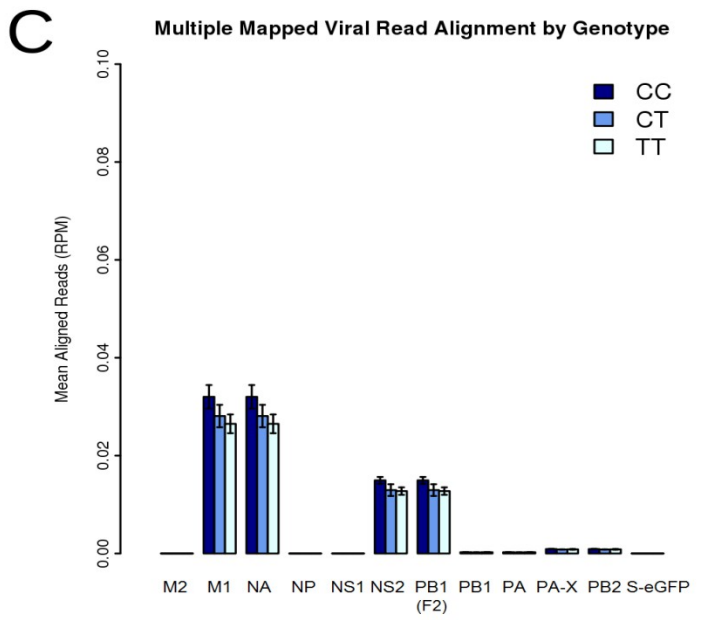

C
Multiple Mapped Viral Read Alignment by Genotype
Mean Aligned Reads (RPM)
0.00
0.02
0.04
0.06
0.08
0.10
CC
CT
TT
M2
M1
NA
NP
NS1
NS2
PB1 (F2)
PB1
PA
PA-X
PB2
S-eGFP

Figure 6.5. Bar plots summarizing the alignment of reads to the viral transcriptome, including the percentage of reads aligned from each sample (A), and to which of the viral RNAs these reads aligned, displaying those uniquely aligned reads (B; retained for all downstream analyses) and multi-mapped reads (C). The multi-mapped reads were visualized to check whether regions with zero uniquely aligning reads was actually due to no expression or due to duplicate sequences. This figure determines that the latter is most likely the case. This visualization was generated following the renaming of sample CT replicate 2 to CC replicate 4 (and so only includes 2 CT samples).

Investigating the Effects of Infection

Principal Component Analysis (PCA) was implemented to investigate the relationships of the samples to one another. The first principal component explains a considerably larger percentage of the variance when compared to subsequent principal components, with the first explaining 66%, and all others explaining 5% or less. This first principal component largely explains infection status, displaying a distinct separation between infected and uninfected samples (Figure 6.6A). This suggests infection had a very large effect on the transcriptome. None of the principal components seems to be explained by the variant in the *IFITM3* gene, suggesting that this has only had a relatively small effect on the transcriptome overall (Figure 6.6B). In addition to this, none of the principal components could also be explained by experimental bias due to read count, RNA integrity, or quantity of RNA used (Figure 6.7).

The expression of key immunological signaling genes was also investigated. Specifically, the expression patterns of type 1 interferons were visualized (Figure 6.8), the proteins of which are (in general) produced upon recognition of a viral infection (Isaacs & Lindenmann, 1957). Many of these interferon type 1 genes show a consistent pattern, with significantly higher expression in the infected samples (see Differential Expression Analysis), as expected. These also show higher expression in samples with the CC *IFITM3* genotype (influenza-susceptible phenotype). This is displayed in a distinct pattern, by which expression in CC is the highest, followed by CT, and then by TT. This also reflects the pattern observed in the expression of the NS1 viral transcript (as

displayed in Figure 6.5B). Together, these findings could suggest that the viral infection has been more potent in the samples of the CC *IFITM3* genotype, as has been observed in the general population (Zhang *et al.* 2013), resulting in the need to mount a larger innate immune response.

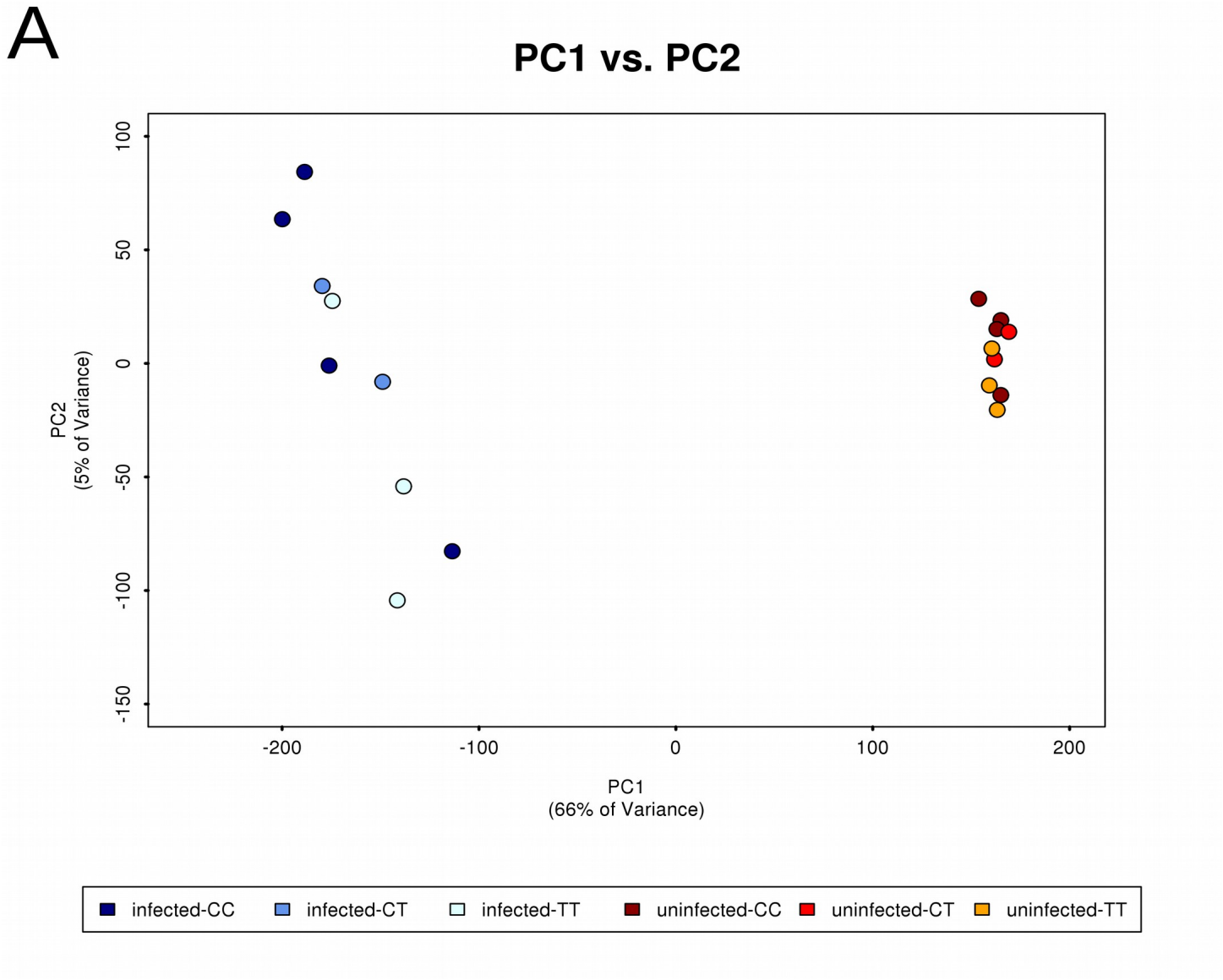
The effect of infection was also investigated in relation to *IFITM3* and the rs12252 variant. The overall expression of the *IFITM3* gene was found to be significantly increased in the infected samples compared to the uninfected, and there was also found to be no significant difference in gene expression in relation to variant genotype (Figure 6.9A; see Differential Expression Analysis). This is as expected, because expression of the IFITM3 protein is generally preventative of infection, and the variant is hypothesized to cause a truncation of this protein, hence being responsible for differential splicing of the gene but not altering overall gene expression.

Further investigation into the reads aligning with the rs12252 variant was also implemented, including quantifying the read coverage of each genotype in each sample (Table 6.3). This uncovered some noteworthy results, specifically in the heterozygote (CT) samples. The two remaining CT samples exhibit a reduced proportion of reads displaying the rs12252-C variant in infected samples, compared to uninfected (Figure 6.9B; Figure C2). In uninfected samples, an average of 69.43% of reads aligned to this loci exhibit a rs12252-C (59.50% and 79.35% in each replicate) which is reduced to an average of 52.22% in infected samples (48.75% and 55.70% in each replicate). The contingency tables for reads in each replicate (as observed in Table 6.3) returned odds

ratios of 1.54 (replicate 1) and 2.29 (replicate 3). The Cochran-Mantel-Haenszel test for count data, which has been described as a Fisher's Exact test of proportions for multiple replicates (Topa *et al.* 2015), was applied in a two-tailed fashion and with exact conditional testing (using strata margins rather than chi-squared distribution). This returned a common OR for the two replicates of 2.16 (CI=1.66-2.84) and significant p-value of 1.80×10^{-9}.

With the rs12252-T variant associated with decreased influenza-susceptibility, this allelic imbalance of increased T allele expression upon infection could be a sign of preferential expression. This finding warrants further investigations, and greater scrutiny is needed to determine whether this hypothesis is true and how this finding could be used to direct further study (see Discussion).

A


PC1 vs. PC2
PC2
(5% of Variance)
100
50
0
-50
-100
-150
-200
-100
0
100
200
PC1
(66% of Variance)
infected-CC
infected-CT
infected-TT
uninfected-CC
uninfected-CT
uninfected-TT

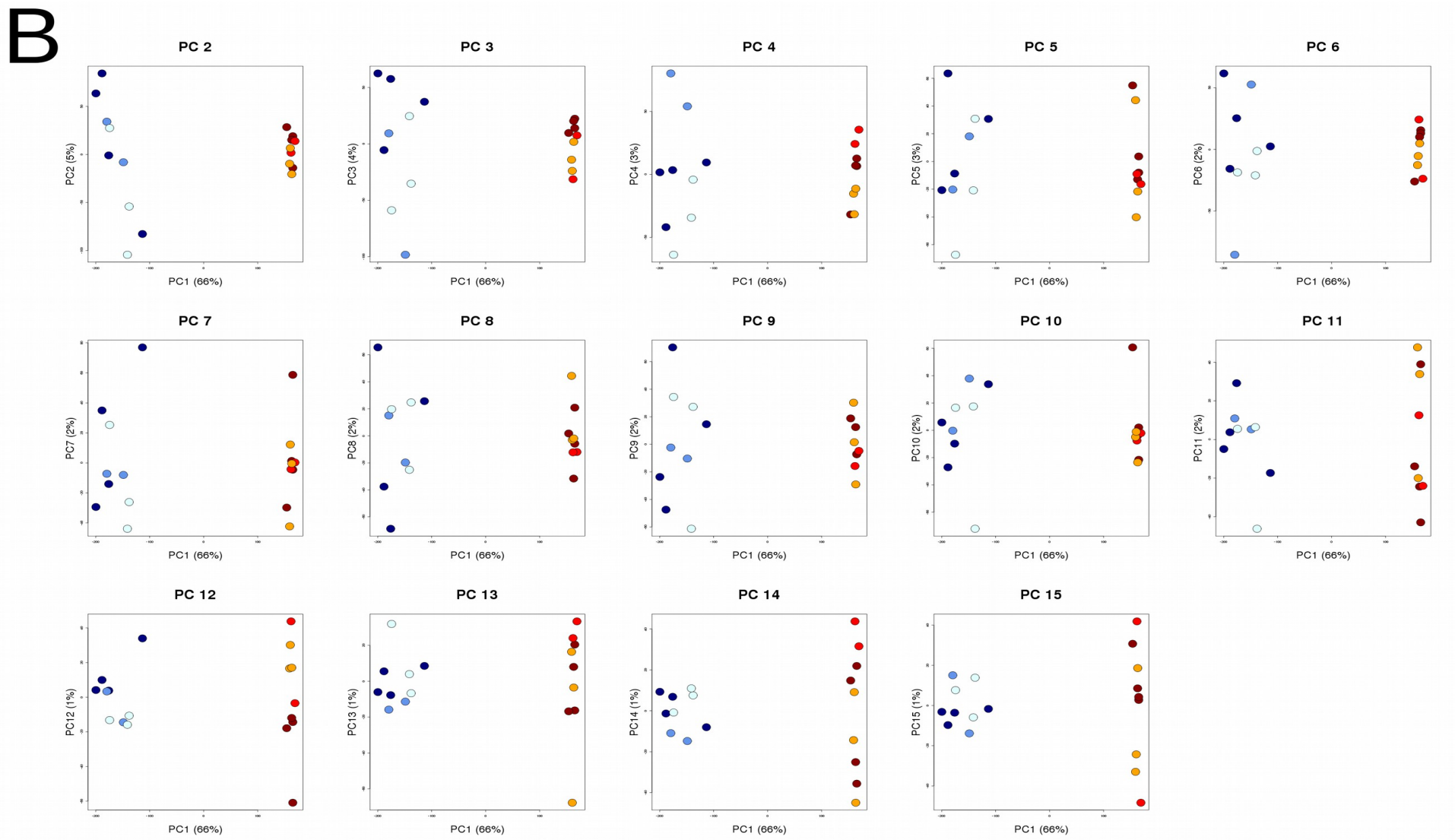
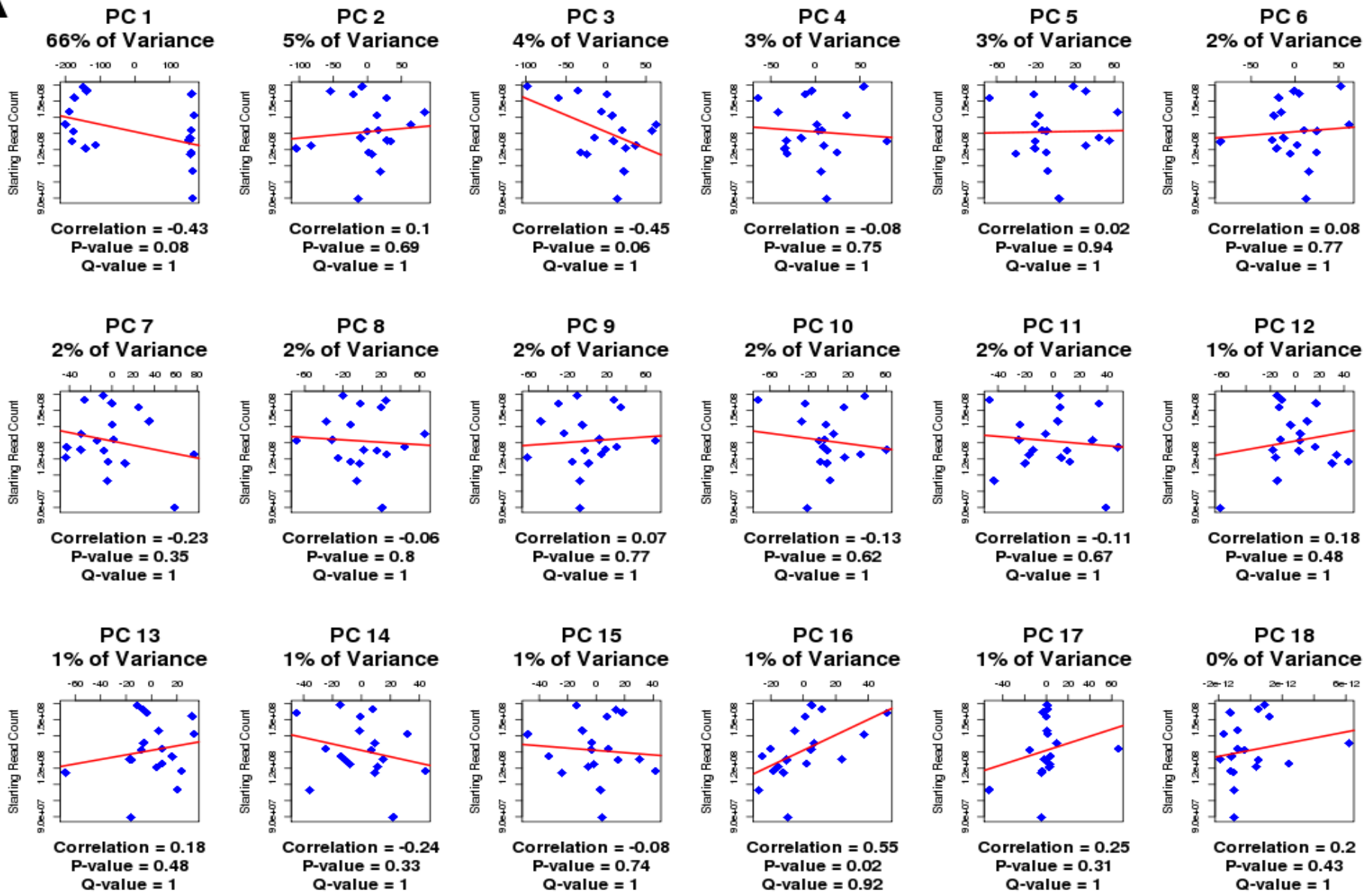
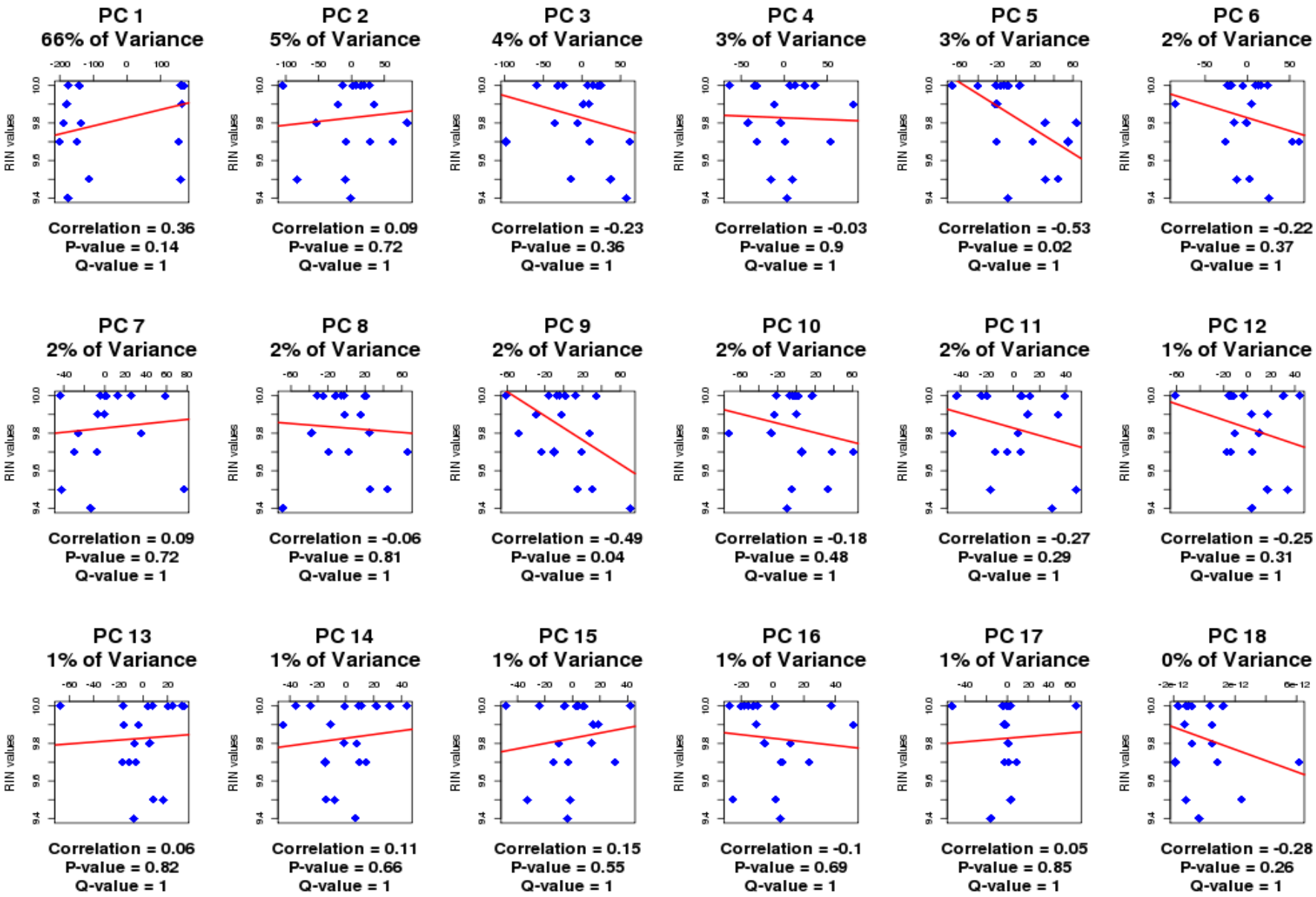


Figure 6.6. Visualizations of the principal components (PCs) of the data, displaying clustering specific to infection status on PC1 (A). None of the top 15 PCs displays evidence of separation based on genotype (B).

A


PC 1
66% of Variance
Correlation = -0.43
P-value = 0.08
Q-value = 1
PC 2
5% of Variance
Correlation = 0.1
P-value = 0.69
Q-value = 1
PC 3
4% of Variance
Correlation = -0.45
P-value = 0.06
Q-value = 1
PC 4
3% of Variance
Correlation = -0.08
P-value = 0.75
Q-value = 1
PC 5
3% of Variance
Correlation = 0.02
P-value = 0.94
Q-value = 1
PC 6
2% of Variance
Correlation = 0.08
P-value = 0.77
Q-value = 1
PC 7
2% of Variance
Correlation = -0.23
P-value = 0.35
Q-value = 1
PC 8
2% of Variance
Correlation = -0.06
P-value = 0.8
Q-value = 1
PC 9
2% of Variance
Correlation = 0.07
P-value = 0.77
Q-value = 1
PC 10
2% of Variance
Correlation = -0.13
P-value = 0.62
Q-value = 1
PC 11
2% of Variance
Correlation = -0.11
P-value = 0.67
Q-value = 1
PC 12
1% of Variance
Correlation = 0.18
P-value = 0.48
Q-value = 1
PC 13
1% of Variance
Correlation = 0.18
P-value = 0.48
Q-value = 1
PC 14
1% of Variance
Correlation = -0.24
P-value = 0.33
Q-value = 1
PC 15
1% of Variance
Correlation = -0.08
P-value = 0.74
Q-value = 1
PC 16
1% of Variance
Correlation = 0.55
P-value = 0.02
Q-value = 0.92
PC 17
1% of Variance
Correlation = 0.25
P-value = 0.31
Q-value = 1
PC 18
0% of Variance
Correlation = 0.2
P-value = 0.43
Q-value = 1
Starting Read Count

B


PC 1
66% of Variance
Correlation = 0.36
P-value = 0.14
Q-value = 1
PC 2
5% of Variance
Correlation = 0.09
P-value = 0.72
Q-value = 1
PC 3
4% of Variance
Correlation = -0.23
P-value = 0.36
Q-value = 1
PC 4
3% of Variance
Correlation = -0.03
P-value = 0.9
Q-value = 1
PC 5
3% of Variance
Correlation = -0.53
P-value = 0.02
Q-value = 1
PC 6
2% of Variance
Correlation = -0.22
P-value = 0.37
Q-value = 1
PC 7
2% of Variance
Correlation = 0.09
P-value = 0.72
Q-value = 1
PC 8
2% of Variance
Correlation = -0.06
P-value = 0.81
Q-value = 1
PC 9
2% of Variance
Correlation = -0.49
P-value = 0.04
Q-value = 1
PC 10
2% of Variance
Correlation = -0.18
P-value = 0.48
Q-value = 1
PC 11
2% of Variance
Correlation = -0.27
P-value = 0.29
Q-value = 1
PC 12
1% of Variance
Correlation = -0.25
P-value = 0.31
Q-value = 1
PC 13
1% of Variance
Correlation = 0.06
P-value = 0.82
Q-value = 1
PC 14
1% of Variance
Correlation = 0.11
P-value = 0.66
Q-value = 1
PC 15
1% of Variance
Correlation = 0.15
P-value = 0.55
Q-value = 1
PC 16
1% of Variance
Correlation = -0.1
P-value = 0.69
Q-value = 1
PC 17
1% of Variance
Correlation = 0.05
P-value = 0.85
Q-value = 1
PC 18
0% of Variance
Correlation = -0.28
P-value = 0.26
Q-value = 1
RIN values

C

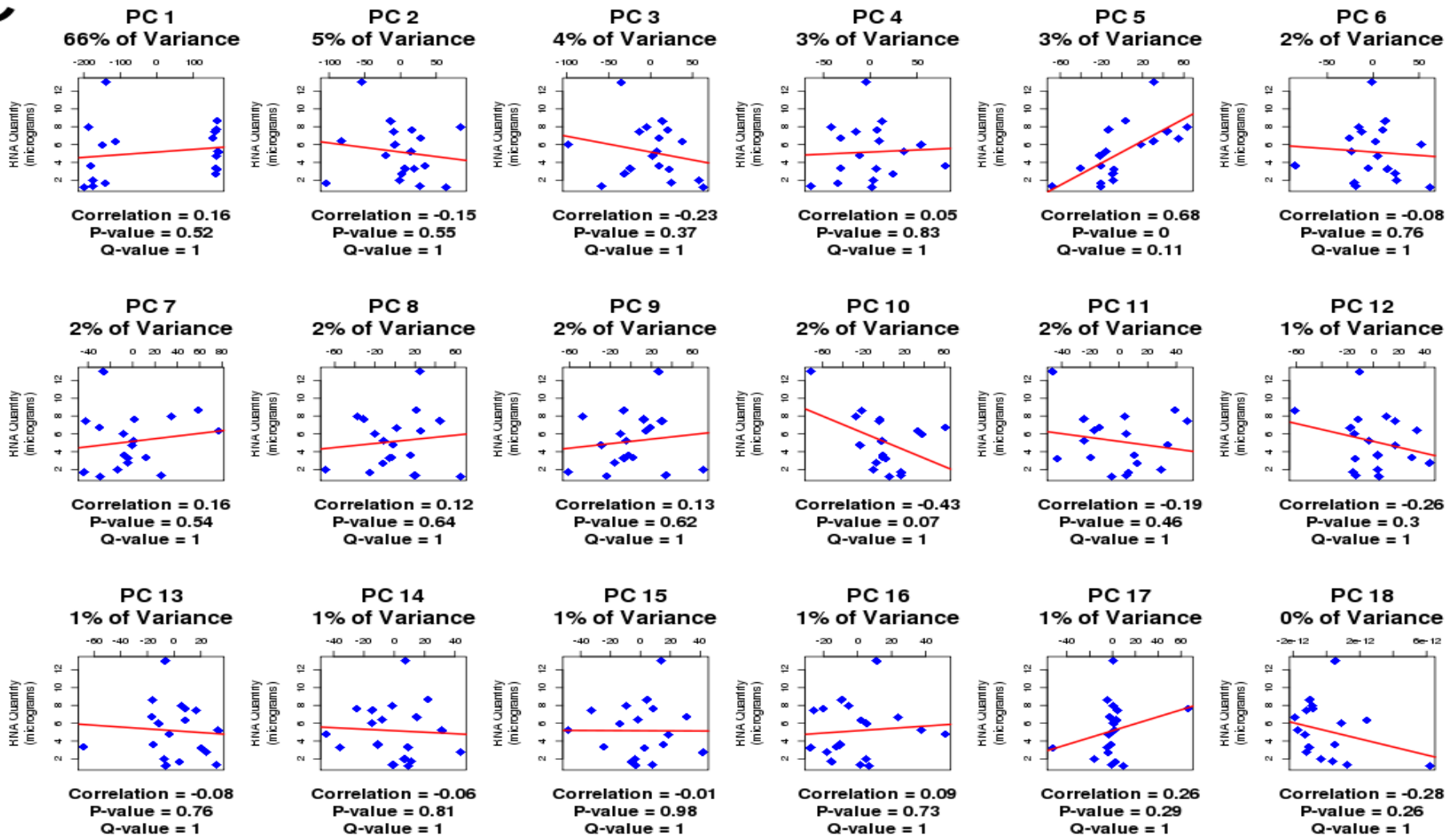


Figure 6.7. Scatter plots investigating each PC's correlation with RIN (A), RNA quantity (B), and number of reads (C). With a linear model fitted (in red), and providing correlation value (r), p-value and q-value.

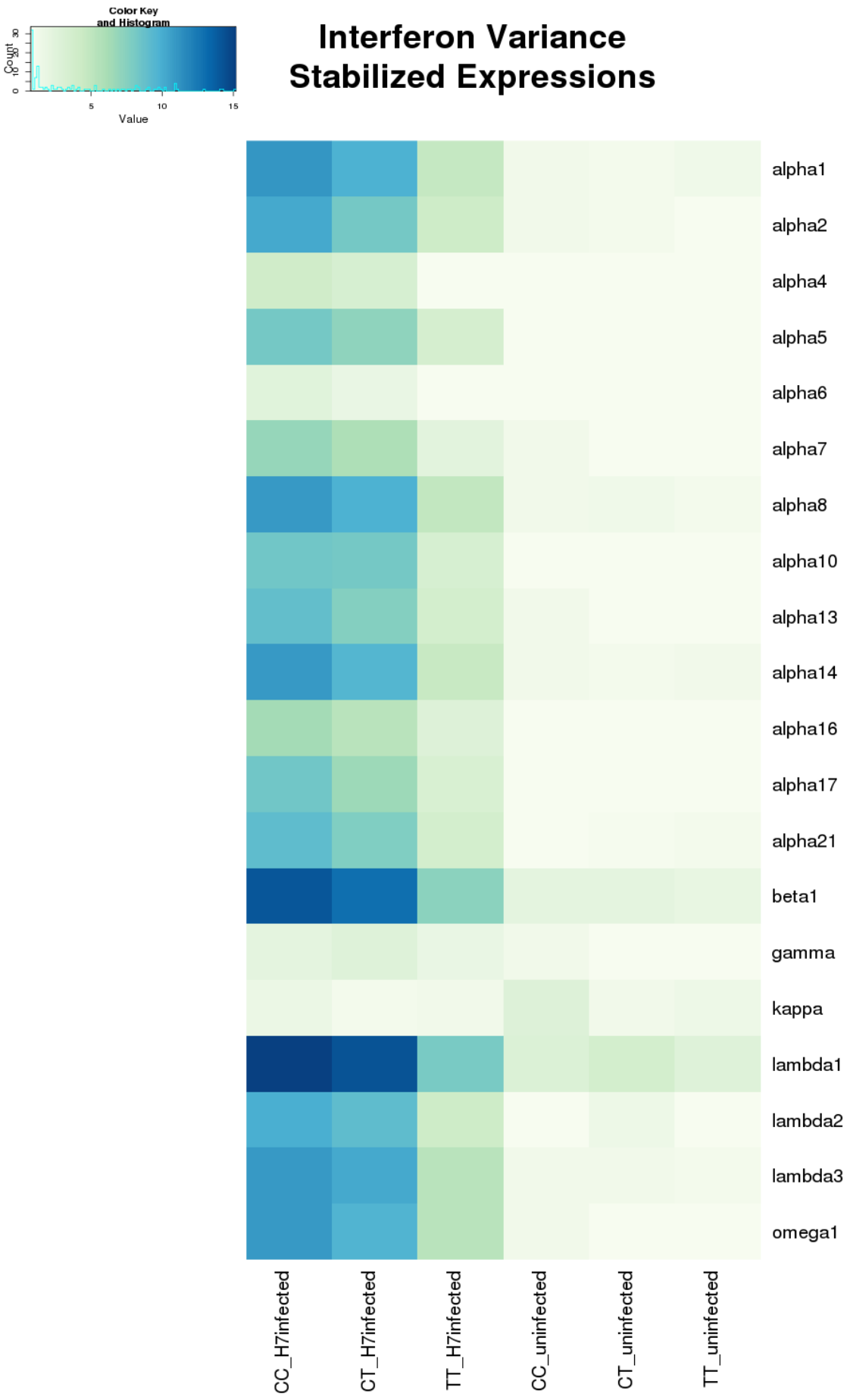
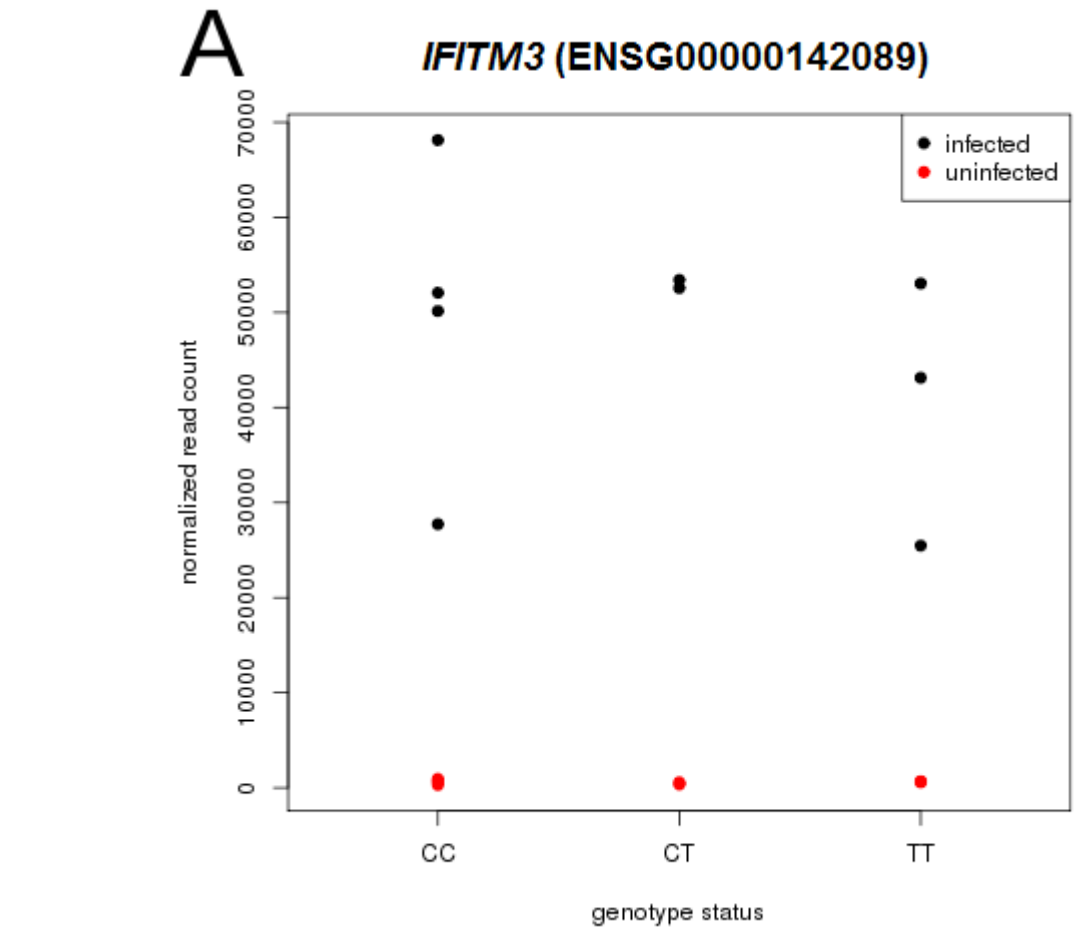

Color Key
and Histogram
Count
Value
Interferon Variance
Stabilized Expressions
alpha1
alpha2
alpha4
alpha5
alpha6
alpha7
alpha8
alpha10
alpha13
alpha14
alpha16
alpha17
alpha21
beta1
gamma
kappa
lambda1
lambda2
lambda3
omega1
CC_H7infected
CT_H7infected
TT_H7infected
CC_uninfected
CT_uninfected
TT_uninfected
Samples

Figure 6.8. A heatmap of expression data for the Type I Interferon genes. The higher expression in CC-genotype infected samples may suggest a compensatory mechanism, or may be due to heavier levels of successful infection. This figure corroborates that of Figure 6.5B, in which the RNA for the viral protein NS1 is seen to be more highly expressed in the CC-genotype (due to the associations of NS1 with interferon). Read counts for this data were variance stabilized with DESeq2.


A
IFITM3 (ENSG00000142089)
infected
uninfected
normalized read count
0
10000
20000
30000
40000
50000
60000
70000
CC
CT
TT
genotype status

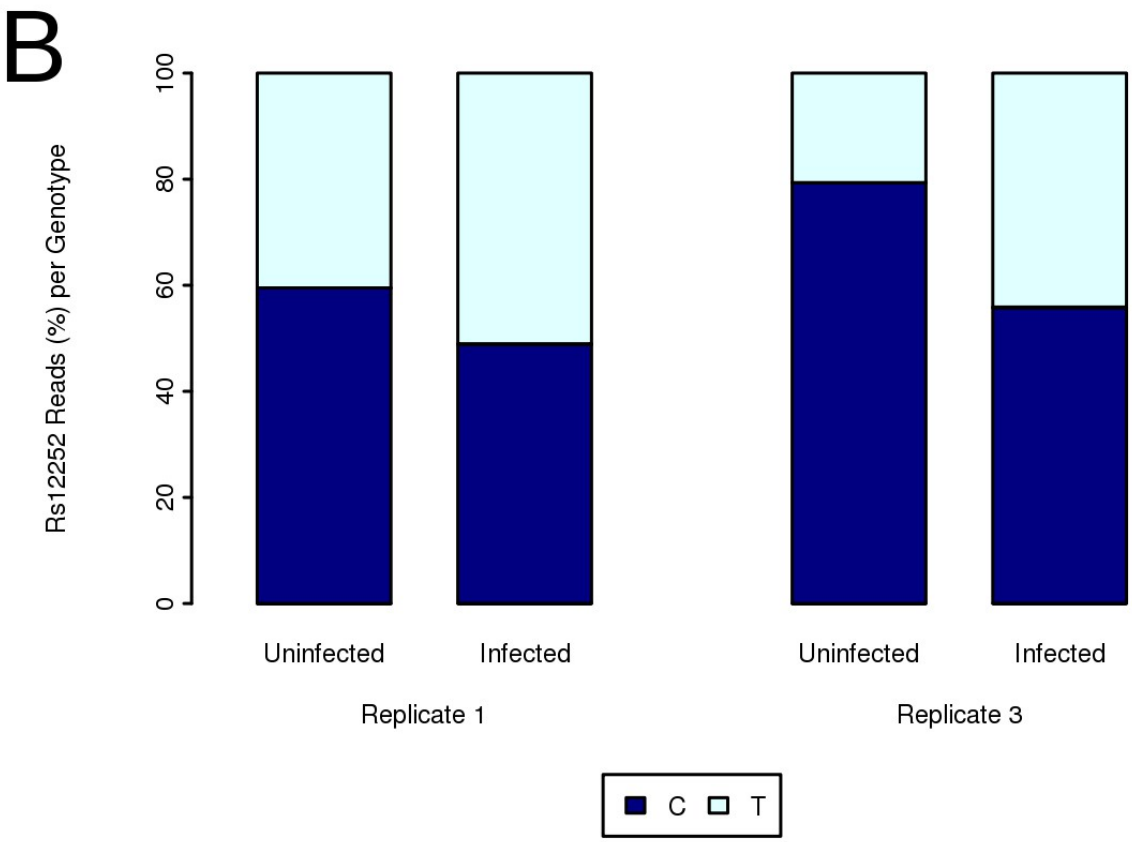

B
Rs12252 Reads (%) per Genotype
0
20
40
60
80
100
Uninfected
Infected
Uninfected
Infected
Replicate 1
Replicate 3
C
T

Figure 6.9. Quantification of reads aligned to the *IFITM3* gene. Panel (A) displays the (DESeq2) normalized read count for all of the replicates of the data, and shows increased but constant expression in infection across the genotypes. Panel (B) displays the quantities of reads exhibiting each form of the variant in both of the heterozygote replicates, and shows a reduction in rs12252-C reads in the infected samples. Counting of reads of each variant was implemented following duplicate removal, so as to avoid the possibility of PCR duplicates affecting the proportions.

Table 6.3. The raw count data corresponding to Figure 6.9B and all other samples, including the number of reads exhibiting each variant of the *IFITM3* rs12252 SNP, in both raw and normalized (reads per million) form. (Due to the gene being on the negative strand, A corresponds to the T variant, and G corresponds to the C variant)

Genotype	Infection status	Replicate	SNP Rs12252 reads				SNP Rs12252 reads per million			
			A	C	G	T	A	C	G	T
CC	infected	1	15	3	11516	3	1.47E-007	2.93E-008	1.13E-004	2.93E-008
		2	6	4	6231	1	5.87E-008	3.91E-008	6.09E-005	9.78E-009
		3	15	8	14939	11	1.47E-007	7.82E-008	1.46E-004	1.08E-007
		4	14	5	11115	5	1.37E-007	4.89E-008	1.09E-004	4.89E-008
	uninfected	1	1	1	273	0	9.78E-009	9.78E-009	2.67E-006	0.00E+000
		2	4	0	78	0	3.91E-008	0.00E+000	7.63E-007	0.00E+000
		3	1	0	161	0	9.78E-009	0.00E+000	1.57E-006	0.00E+000
		4	0	0	219	0	0.00E+000	0.00E+000	2.14E-006	0.00E+000
CT	infected	1	6005	22	5738	5	5.87E-005	2.15E-007	5.61E-005	4.89E-008
		3	5398	19	6816	5	5.28E-005	1.86E-007	6.66E-005	4.89E-008
	uninfected	1	49	0	72	0	4.79E-007	0.00E+000	7.04E-007	0.00E+000
		3	32	0	123	0	3.13E-007	0.00E+000	1.20E-006	0.00E+000
TT	infected	1	10627	21	33	7	1.04E-004	2.05E-007	3.23E-007	6.84E-008
		2	12884	31	9	15	1.26E-004	3.03E-007	8.80E-008	1.47E-007
		3	6306	10	15	7	6.17E-005	9.78E-008	1.47E-007	6.84E-008
	uninfected	1	217	0	6	0	2.12E-006	0.00E+000	5.87E-008	0.00E+000
		2	226	1	1	1	2.21E-006	9.78E-009	9.78E-009	9.78E-009
		3	148	0	1	0	1.45E-006	0.00E+000	9.78E-009	0.00E+000

Differential Expression Analysis

Differential expression analysis was used to determine genes of statistically significantly differential expression across infected and uninfected samples, across samples of *IFITM3* CC and TT genotype, and for the interaction of these two. Finally, this interaction was also separately tested with all three genotypes (CC, CT, and TT), assuming an additive pattern of expression, through likelihood ratio testing.

These tests returned a large number of genes differing in expression due to infection status (Figure 6.10A). Of the 64,253 human coding, non-coding and pseudogene models utilized in read-counting, 32,702 exhibited non-zero read counts in these samples. Of these, 19,387 genes were differentially expressed due to infection (with an FDR corrected p-value < 0.05). This was composed of 11,532 genes significantly higher in infected samples, and 7,855 significantly lower. The finding that over half of the expressed genes (59.28%) were significantly differentially expressed exposes the extremity of the effect of a severe viral infection. As expected, the differentially expressed genes with the very lowest p-values (Figure 6.10B) included many immune-system-related genes. This included many interferon-related genes (e.g. *IFIT3*, *IFIT2*, *IFI44*, *IFI35*, etc.), among which was the gene *IFITM1* of the IFITM family, all of which have strong anti-viral properties.

Due to the over-large number of genes deemed to be significantly differentially expressed, a threshold for $\log_2$ fold change was applied of 3 (and -3) before

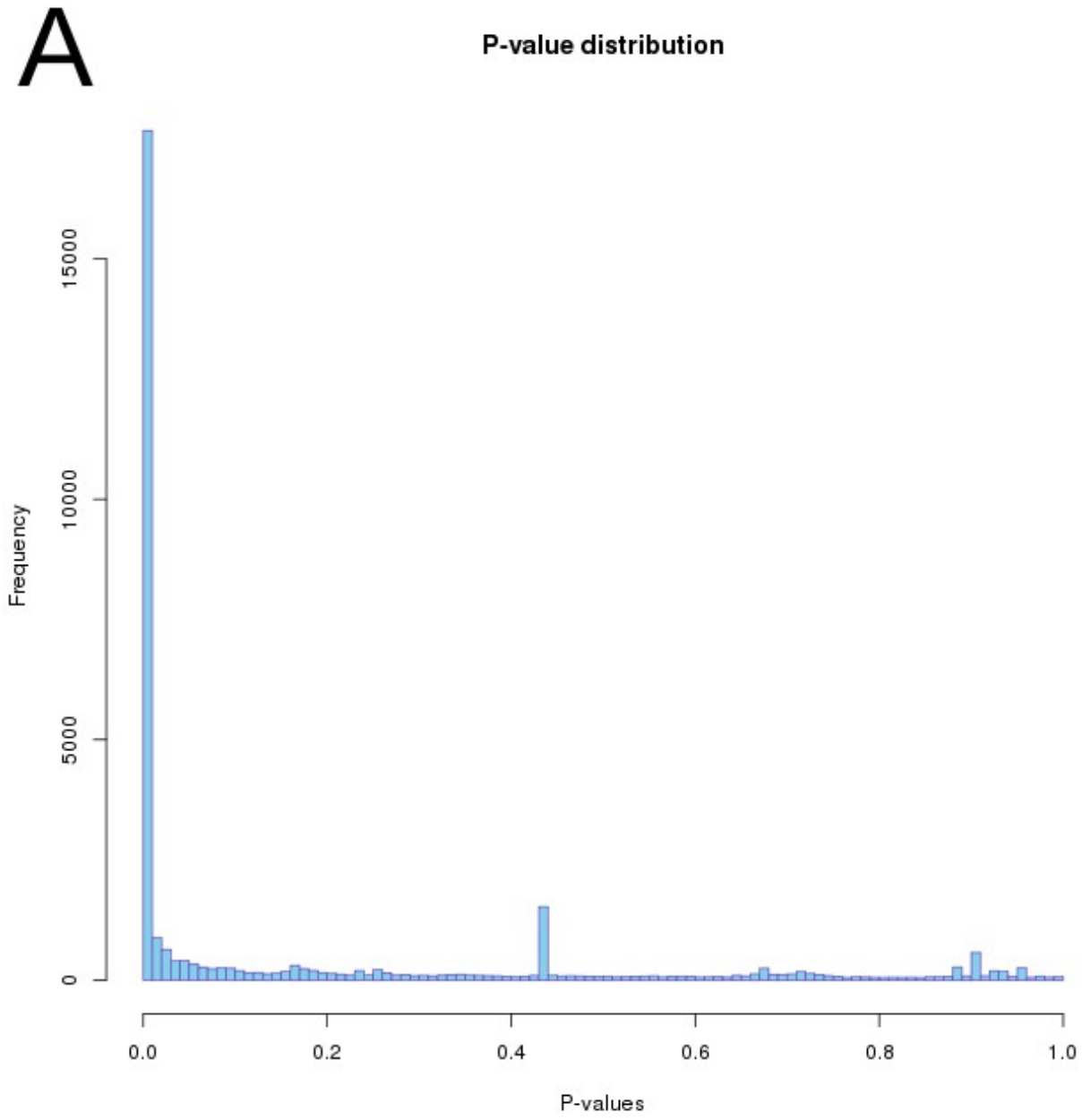
applying GO and KEGG enrichment analyses. This reduced the number of genes to 5,155 expressed more in infected samples, and 1,397 expressed more in uninfected samples (Figure 6.10C). GO and KEGG enrichment analyses confirmed much of what we would expect in the infected cells. Analyses of those genes more highly expressed with infection returned a total of 120 significantly enriched GO terms (after passing through REVIGO) and 20 significantly enriched KEGG pathways (FDR < 0.05). The top 5 most significant GO terms of each ontology category, and the top 5 most significant KEGG pathways, are displayed in Figure 6.11. Extracellular space appears to be the most highly enriched GO term, referring to gene products with functions outside of the cell but still within the organism. This is supported by the cell periphery and extracellular region GO terms. These terms are seen alongside the defence response and response to other organism / biotic stimulus GO terms, all implicating the immune response. Cytokine activity is also seen, which play an important role in immunity, such as response to infection and inflammation. Enriched KEGG pathways also implicate immune response, including for example cytokine-cytokine receptor interaction; the Jak-STAT signalling pathway, which is responsible for transmitting information from extracellular chemical signals to the nucleus for gene expression changes (including in immunity); and the RIG-I-like receptor signalling pathway, which includes pattern recognition receptors for sensing viruses. All of these enrichments provide confidence in the successful infection of the cells in this experiment. In contrast to these findings, the only functional annotation of genes more highly expressed in uninfected samples included one significantly enriched GO term, which was cilium organization. Although cilia are known to

be involved in bacterial infection resistance, there are no implications for their involvement in viruses such as influenza (Harford & Hamlin, 1952).

Unlike the large numbers of differentially expressed genes identified with infection status, analysis across samples with differing *IFITM3* genotypes returned only 29 genes that were determined to be significantly differentially expressed (FDR < 0.05) (Figure 6.12A). This included 10 genes of higher expression in the CC samples compared to TT, and 19 genes of lower expression (Figure 6.12B). Two of the most striking genes on this list include Human Leukocyte Antigen (HLA) genes, *HLA-DRB1* and *HLA-DPB1*. These are both genes of the human class II Major Histocompatibility Complex (MHC), and encode antigen-presenting proteins. These antigen-presenting proteins are anchored to membranes, and present extracellular peptides. Class II proteins, such as HLA-DRB1 and HLA-DPB1, perform this role in antigen-presenting cells, to stimulate T-helper cells, and subsequently B-cells, resulting in the production of antibodies specific to the antigen (Janeway *et al.* 2001). The differences between HLA-DRB1 and HLA-DPB1 are not widely known, however this should be further investigated in the context of *IFITM3* genotype (see Discussion). The 29 genes identified in this differential expression analysis were too few in number to return any significant GO or KEGG enrichments.

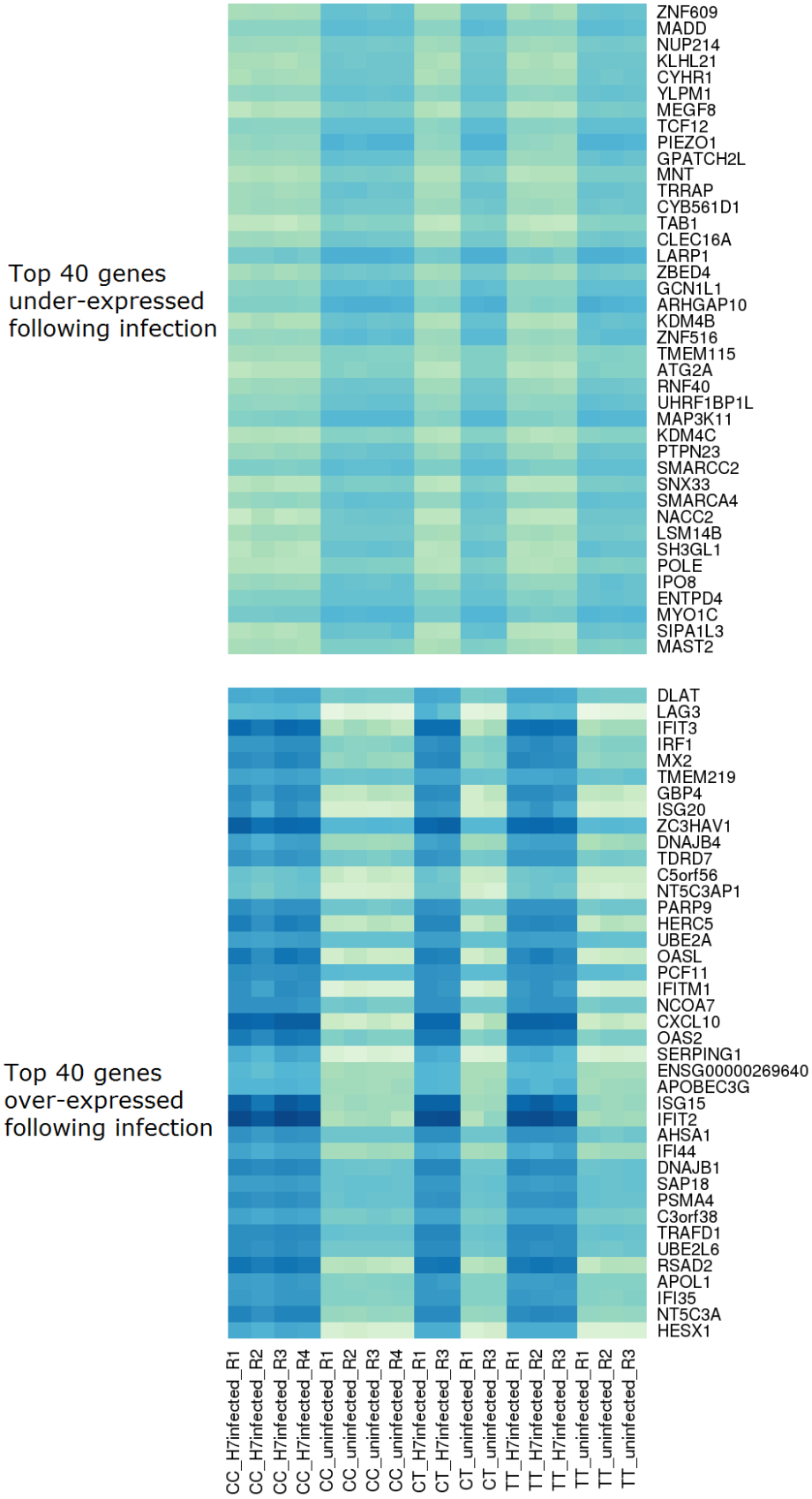
Differential expression analysis was also applied to the interaction of *IFITM3* genotype and infection status, with the aim of reporting genes for which the change in expression between genotypes is significantly different dependent on infection status. As with the testing across genotypes, this returned no results

(Figure 6.13A). With a significance threshold of FDR < 0.05, there were zero genes found to be significantly differentially expressed. Even when using a less stringent significance threshold of FDR < 0.1, this only returned a single gene: *FABP5L7* (/*FABP5P7*; Ensembl ID: ENSG00000234964). For this gene, the overall expression is increased in infected samples compared to uninfected, with the CC genotype exhibiting a much lower expression in those uninfected samples than the TT genotype (Figure 6.13B). Although *FABP5L7* is a pseudogene of the fatty acid binding protein 5 gene, and it has not been studied in depth, *FABP5* (along with *FABP5L7*) was named by Buschow *et al.* (2010) among a list of genes which contribute proteins to MHC class II carrying exosomes (a secretion of antigen presenting cells) (Supplementary Table 1A of Buschow *et al.* 2010), and so may be worth studying further.


A
P-value distribution
Frequency
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
P-values

B

Top 80 Differentially Expressed Genes



C

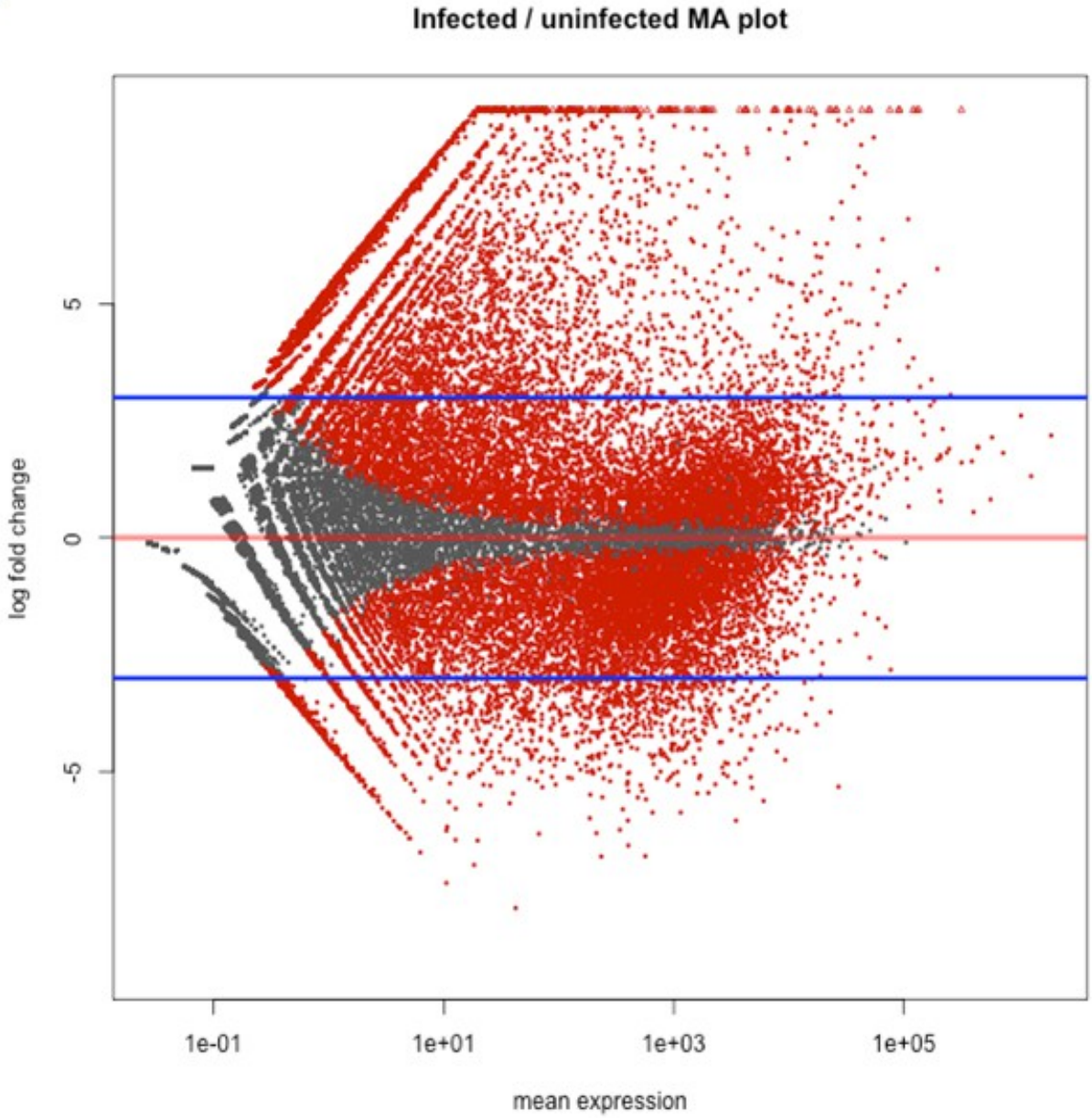
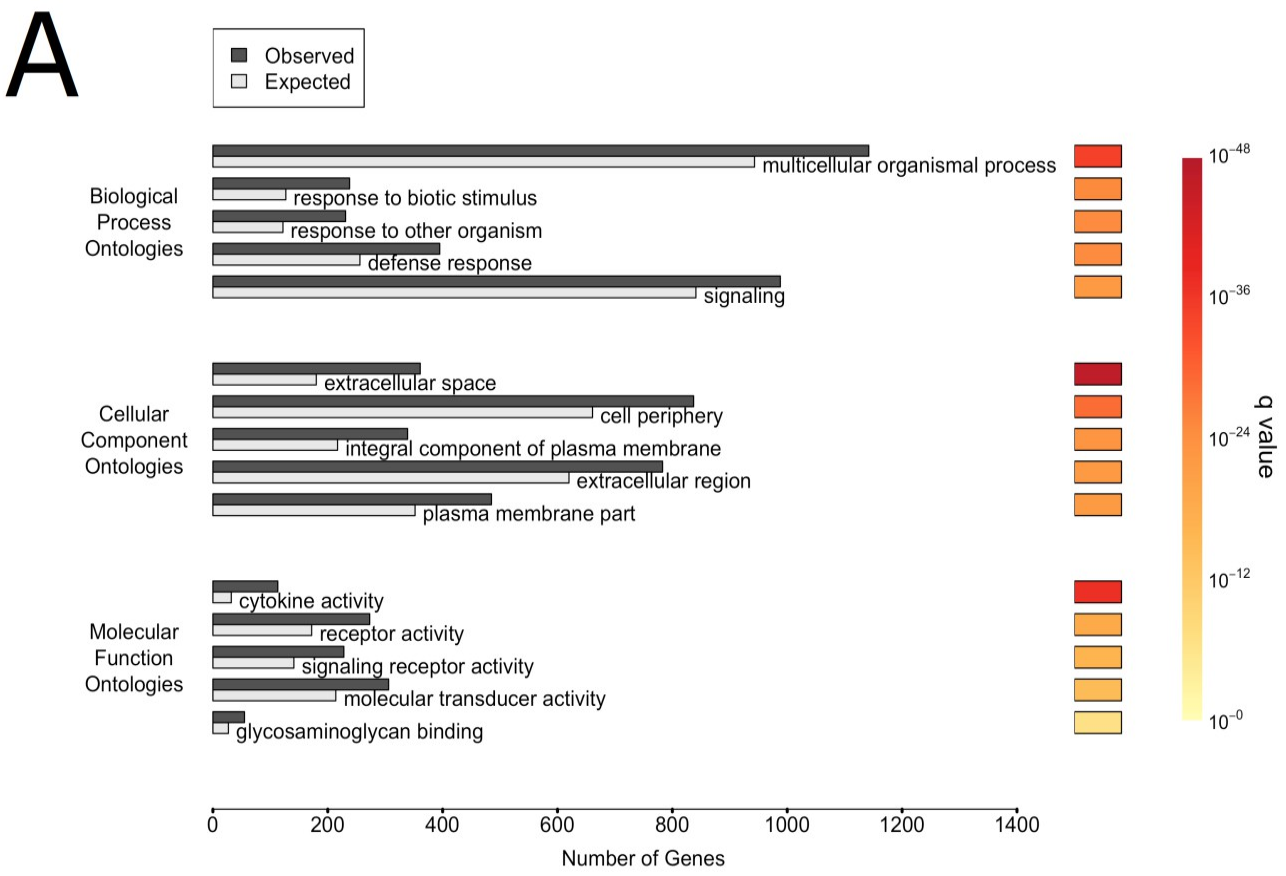

Infected / uninfected MA plot
log fold change
5
0
-5
1e-01
1e+01
1e+03
1e+05
mean expression

Figure 6.10. Visualization of some of the results from differential expression analysis across infection status of the samples. Panel (A) displays a histogram of p-values for each of the genes across the genome, demonstrating a large frequency of low p-values. Panel (B) displays the top 40 differentially expressed genes (by q-value) for upregulation of both uninfected and infected samples respectively. Panel (C) displays an MA plot of gene expression $\log_2$ fold change between infected and uninfected samples, over mean expression. Genes initially determined as significantly differentially expressed are displayed in red. The threshold for additional filtering by fold change is displayed with blue horizontal lines, at $\log_2$ fold change of 3 and -3.


A
Observed
Expected
Biological Process Ontologies
multicellular organismal process
response to biotic stimulus
response to other organism
defense response
signaling
Cellular Component Ontologies
extracellular space
cell periphery
integral component of plasma membrane
extracellular region
plasma membrane part
Molecular Function Ontologies
cytokine activity
receptor activity
signaling receptor activity
molecular transducer activity
glycosaminoglycan binding
0
200
400
600
800
1000
1200
1400
Number of Genes
10^{-48}
10^{-36}
10^{-24}
10^{-12}
10^{-0}
q value

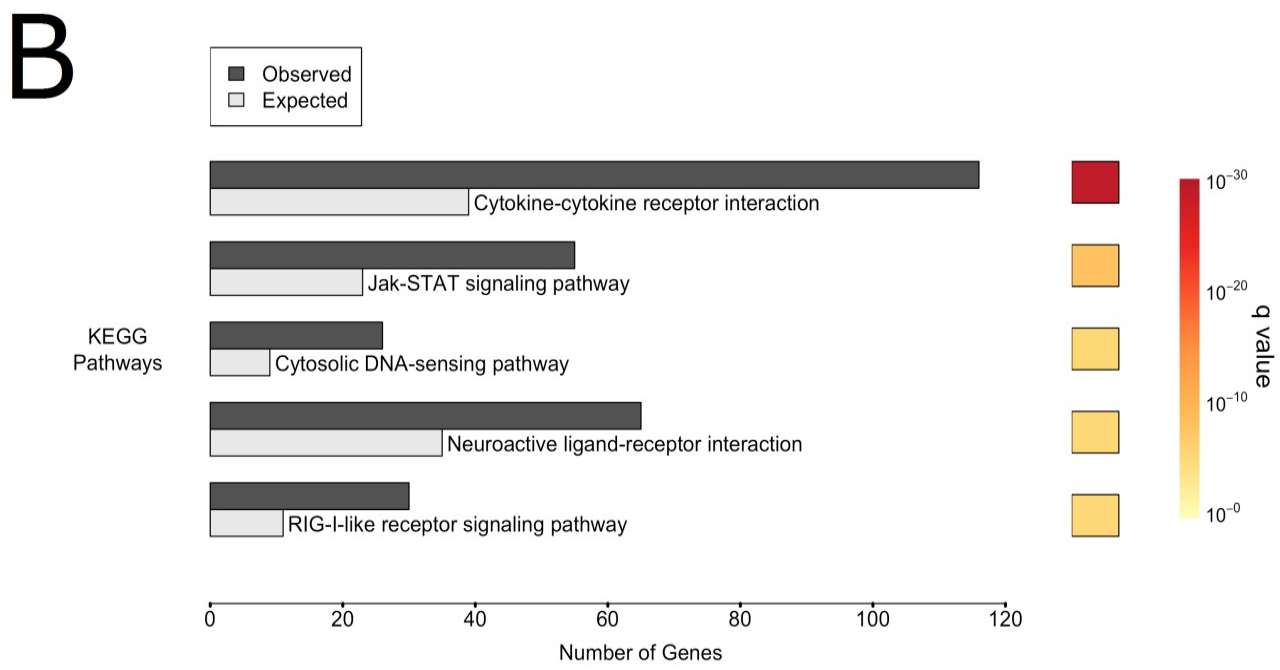
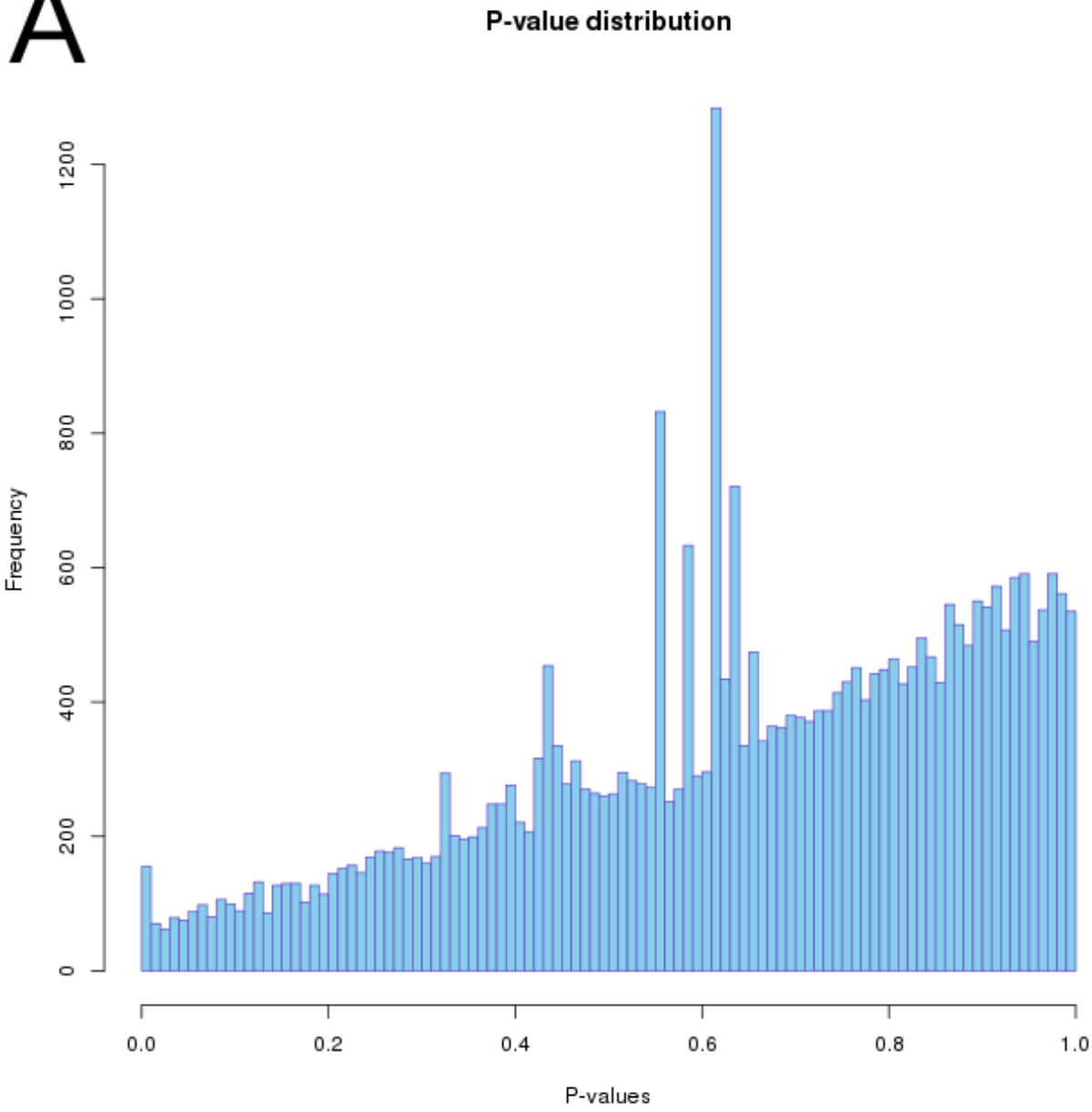

B
Observed
Expected
KEGG Pathways
Cytokine-cytokine receptor interaction
Jak-STAT signaling pathway
Cytosolic DNA-sensing pathway
Neuroactive ligand-receptor interaction
RIG-I-like receptor signaling pathway
0
20
40
60
80
100
120
Number of Genes
10^{-30}
10^{-20}
10^{-10}
10^{-0}
q value

Figure 6.11. Bar plots of the top 5 significant functional annotations (by q-value) for GO terms (A) and KEGG pathway (B) enrichments for genes upregulated following infection. Bars represent the number of genes expected and observed for each term. Q-values are displayed via the attached Hinton plots.

A


P-value distribution
Frequency
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
P-values

B

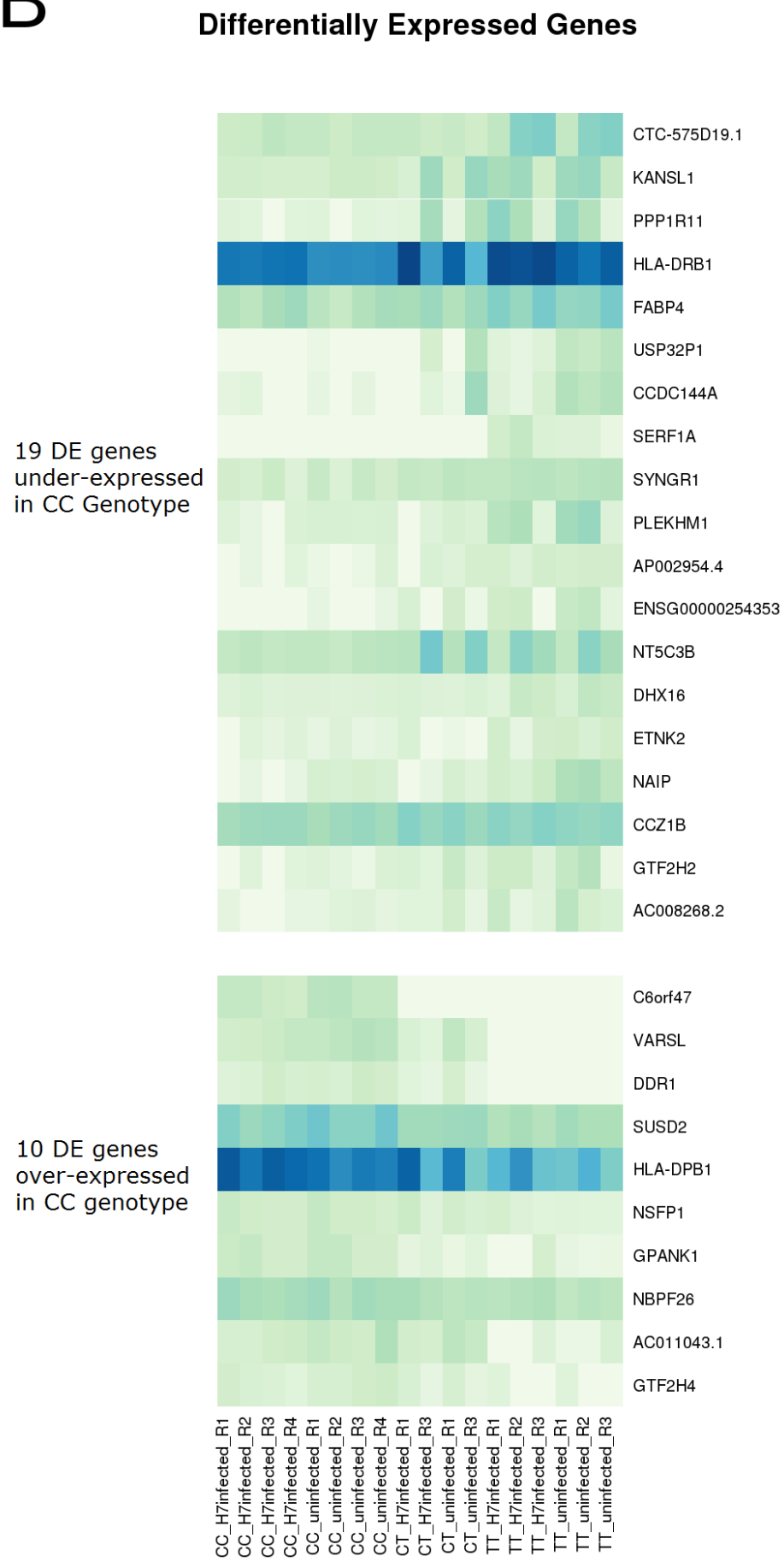
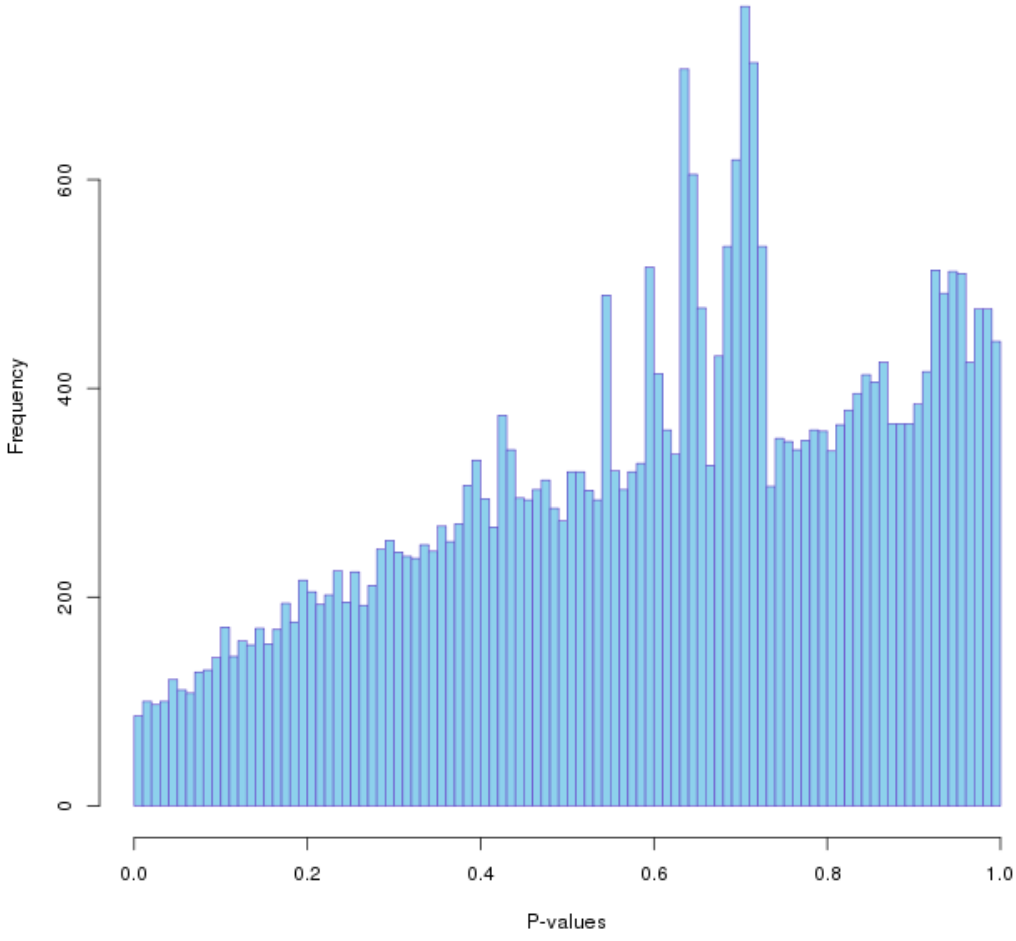

Differentially Expressed Genes
19 DE genes under-expressed in CC Genotype
CTC-575D19.1
KANSL1
PPP1R11
HLA-DRB1
FABP4
USP32P1
CCDC144A
SERF1A
SYNGR1
PLEKHM1
AP002954.4
ENSG00000254353
NT5C3B
DHX16
ETNK2
NAIP
CCZ1B
GTF2H2
AC008268.2
10 DE genes over-expressed in CC genotype
C6orf47
VARSL
DDR1
SUSD2
HLA-DPB1
NSFP1
GPANK1
NBPF26
AC011043.1
GTF2H4
CC_H7infected_R1
CC_H7infected_R2
CC_H7infected_R3
CC_H7infected_R4
CC_uninfected_R1
CC_uninfected_R2
CC_uninfected_R3
CC_uninfected_R4
CT_H7infected_R1
CT_H7infected_R3
CT_uninfected_R1
CT_uninfected_R3
TT_H7infected_R1
TT_H7infected_R2
TT_H7infected_R3
TT_uninfected_R1
TT_uninfected_R2
TT_uninfected_R3
Samples

Figure 6.12. Visualization of some of the results from differential expression analysis across *IFITM3* rs12252 genotype of the samples. Panel (A) displays a histogram of p-values for each of the genes across the genome and, in contrast to that seen in Figure 6.10A, demonstrates a low frequency of low p-values (and hence very few results). Panel (B) displays all 29 differentially expressed genes for upregulation with both TT and CC genotypes respectively.

A


P-value distribution
Frequency
0
200
400
600
0.0
0.2
0.4
0.6
0.8
1.0
P-values

B

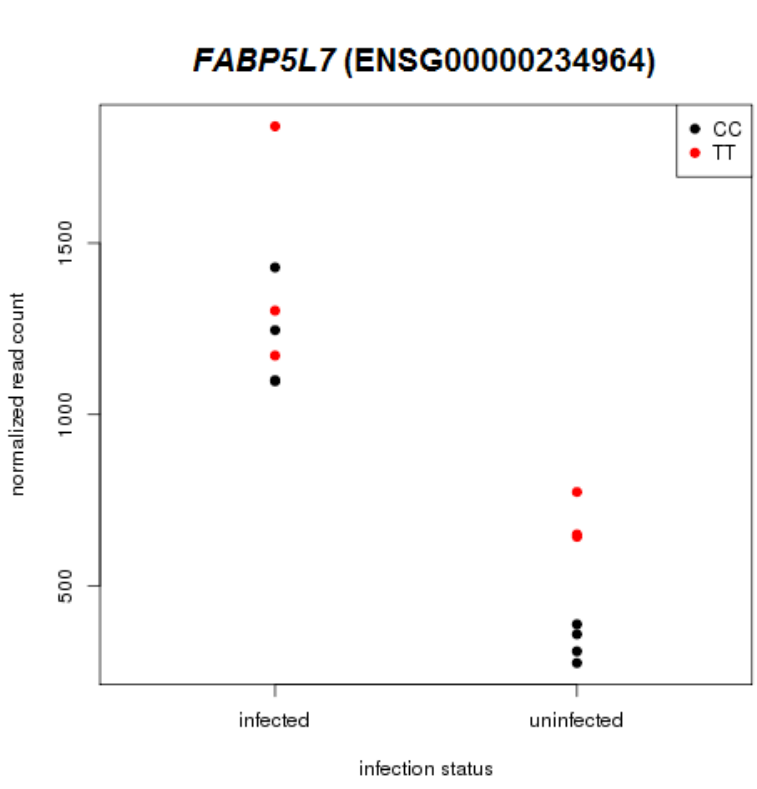
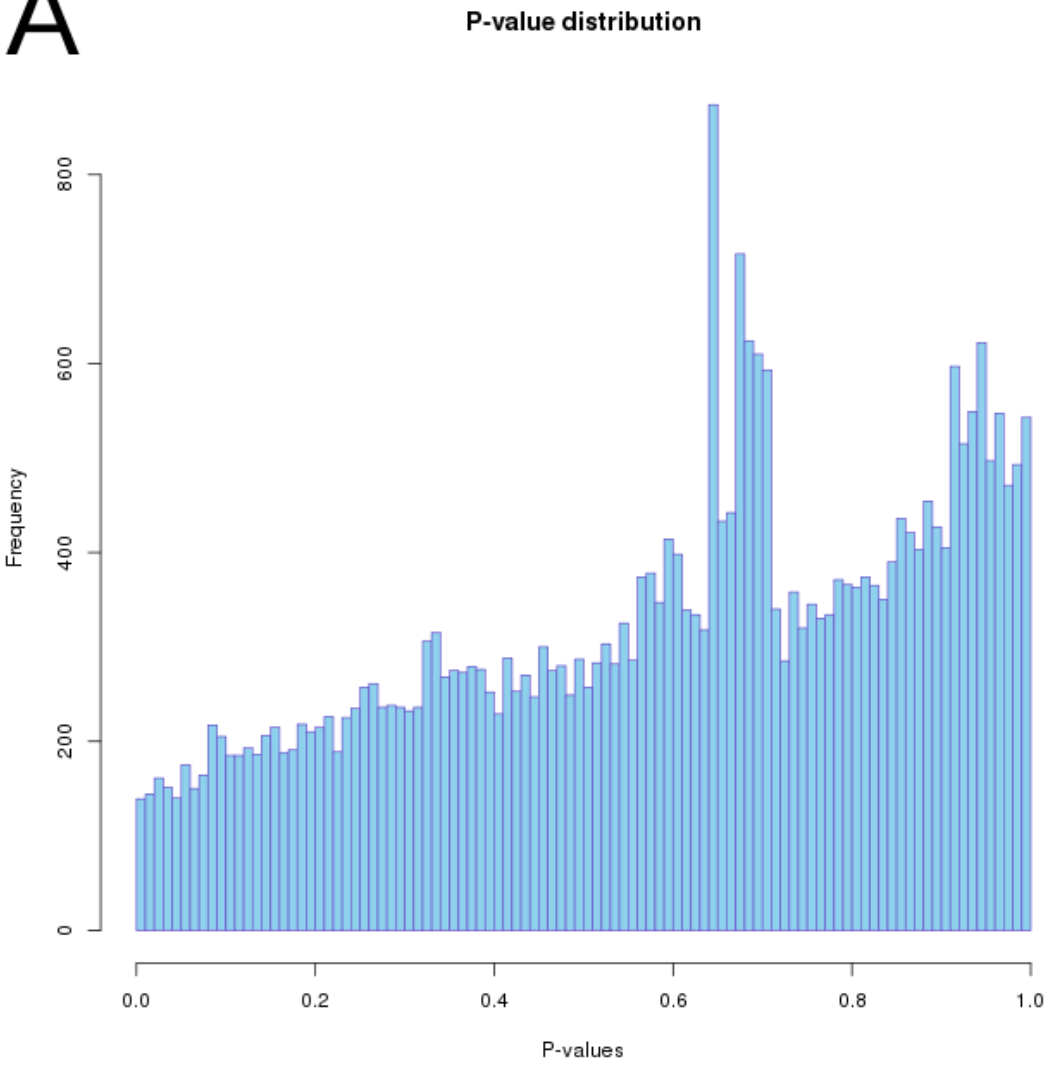

FABP5L7 (ENSG00000234964)
CC
TT
normalized read count
500
1000
1500
infected
uninfected
infection status

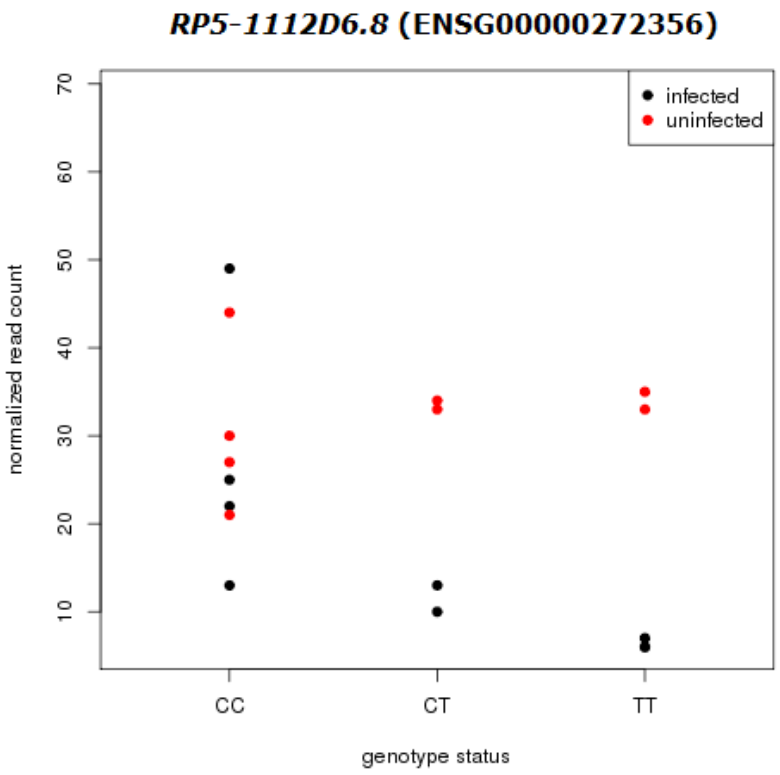
Figure 6.13. Visualization of some of the results from differential expression analysis across the interaction of infection status and *IFITM3* rs12252 genotype. Panel (A) displays a histogram of p-values for all genes across the genome, of which only one gene was found to have an FDR-corrected value below 0.1. Panel (B) displays the expression values for this gene (*FABP5L7*) in each of the replicates of each of the samples.

Finally, differential expression analysis was performed with a likelihood ratio test, which was also aimed at identifying genes of the *IFITM3* genotype and infection status interaction, but with the advantage of allowing for testing across all three genotypes (CC, CT, and TT). This provides greater power for identifying differentially expressed genes, but does assume an additive pattern of expression. However, as with the paired interaction testing, this also produced uninformative results (Figure 6.14A). No gene was found to be significantly differentially expressed with an FDR threshold < 0.05, and only two at a less stringent threshold of 0.1. These two genes are *RP5-1112D6.8* (Ensembl ID: ENSG00000272356) and *RPL32P1* (Ensembl ID: ENSG00000224098), both encoding ribosomal proteins. Little is known of the gene RP5-1112D6.8, but the expression pattern shows a consistently high expression across genotypes when uninfected, with a lowering of expression in TT genotypes upon infection (Figure 6.14B). Interestingly, the *RPL32P1* gene is a known processed pseudogene, but is contained within an intron of the *HLA-DPB1* gene (same strand), and thought to have been produced by integration of a reverse transcribed mRNA into the genome (Young & Trowsdale, 1985). With a consistently low expression across genotypes when infected, and then an increased expression in the CC genotype when uninfected (Figure 6.14C), this could corroborate the identification of the *HLA-DPB1* gene in the paired genotype differential expression analysis, in which it was also upregulated in the CC genotype (Figure 6.12B). Although these two genes are transcribed from different promoters, transcription of these two genes may be coupled.

A


P-value distribution
Frequency
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
P-values

B


RP5-1112D6.8 (ENSG00000272356)
infected
uninfected
normalized read count
10
20
30
40
50
60
70
CC
CT
TT
genotype status

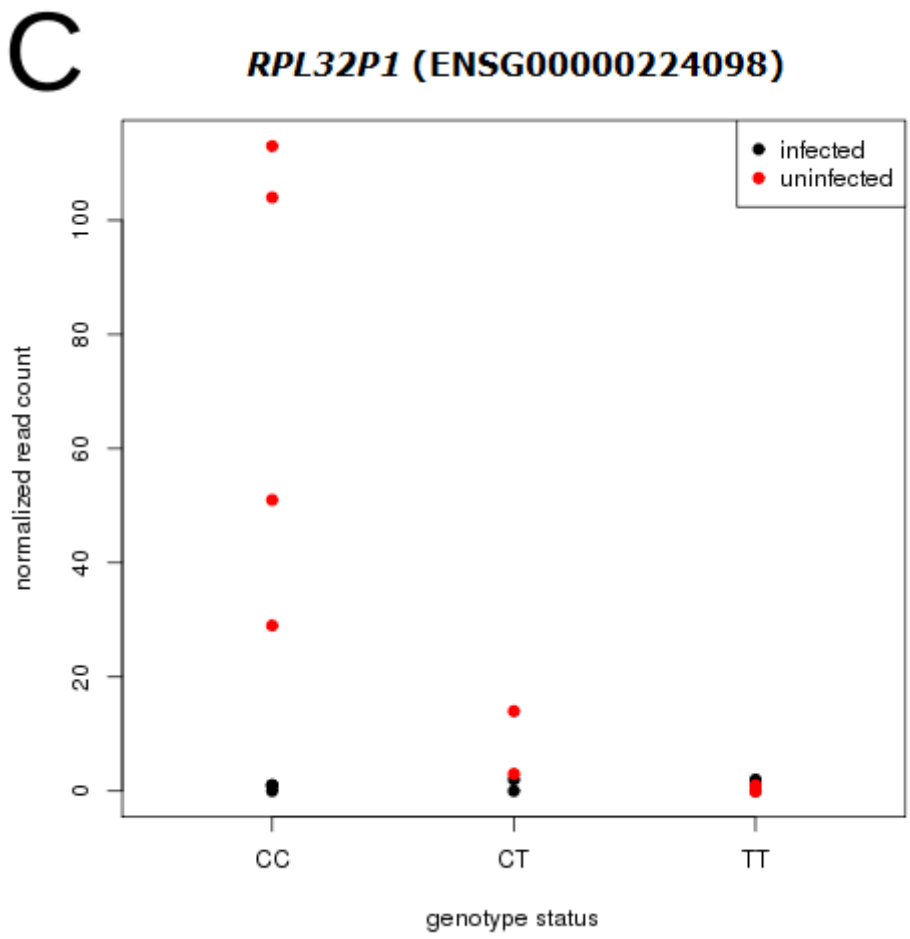


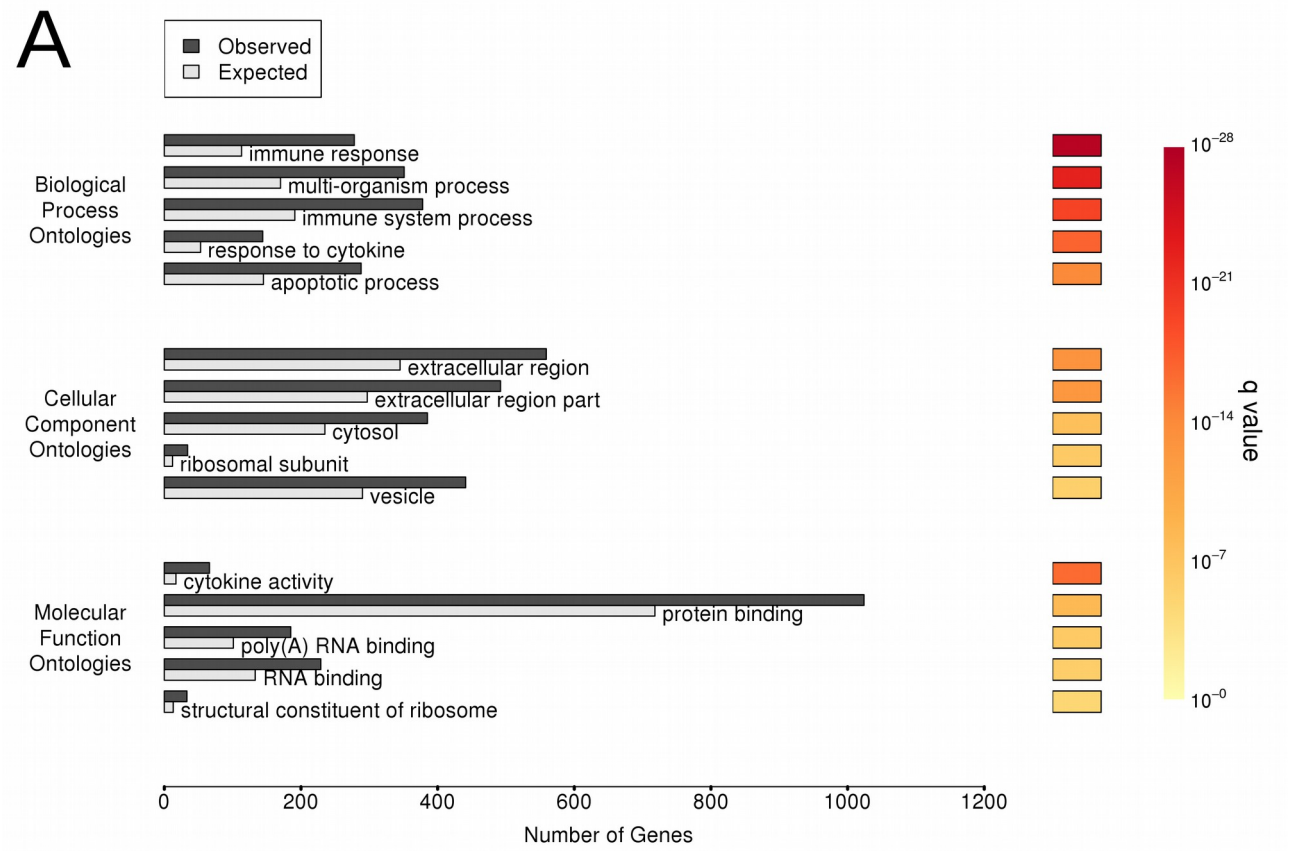
Figure 6.14. Visualization of some of the results from differential expression analysis across all *IFITM3* rs12252 genotypes (including heterozygotes), through likelihood ratio testing and assumed additive allelic expression. Panel (A) displays a histogram of p-values for each of the genes across the genome, for which only two genes were found to have an FDR-corrected value below 0.1. Panels (B) and (C) display the expression values for these genes (RP5-1112D6.8 and RPL32P1) in each of the replicates of each of the samples. The expression patterns identified in the interferon type I genes (Figure 6.8) appear to fit a pattern that should be identified by this test. However there may have been insufficient power to determine these as statistically significant.

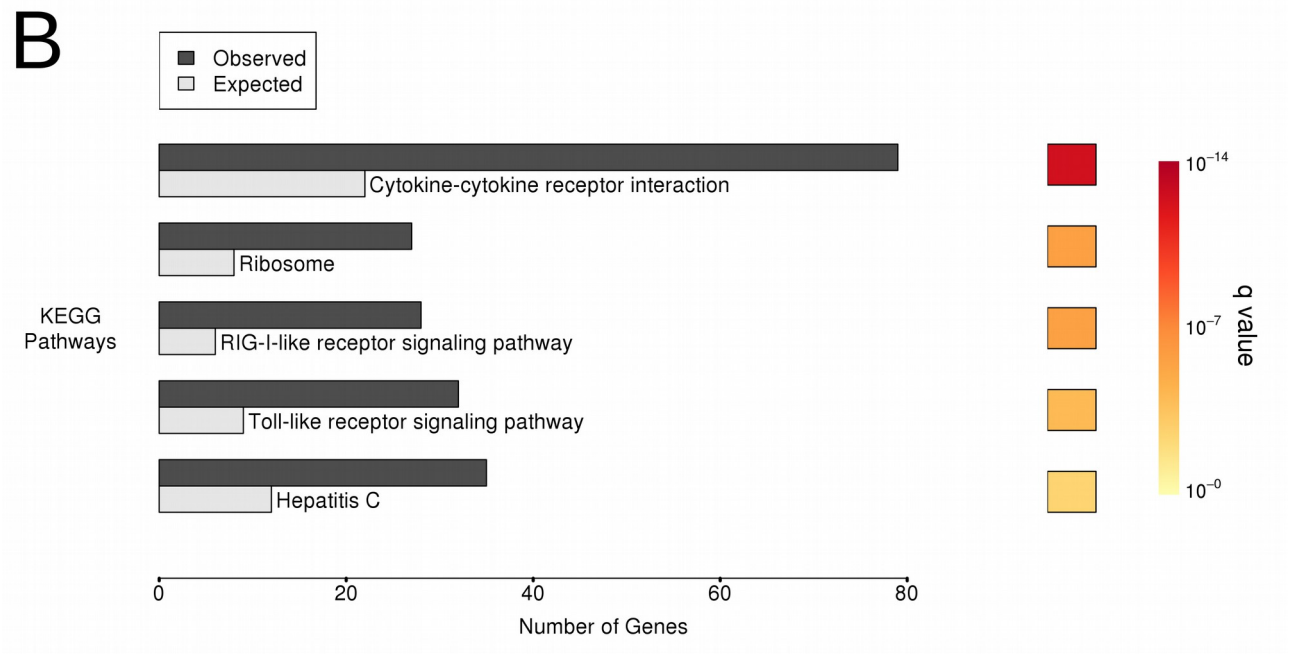
Differential Splicing Analysis

Differential splicing was implemented through probabilistic models to determine the expected expression of isoforms across the genome. BitSeq allows for these models of expected expression to be compared in a paired manner, and so were used to determine differential isoforms across *IFITM3* genotype and infection status. In terms of the number of significantly differentially expressed genes identified, these results were similar to those of the equivalent differential expression analyses, with a large number of transcripts deemed differentially expressed over infection (12,876) and relatively few differentially expressed due to *IFITM3* genotype (8).

Of those identified over infection status, 2,935 of these transcripts had significantly increased expression with infection, and 9,941 had significantly decreased expression. GO and KEGG enrichment analyses were performed on the genes of these transcripts, as shown in Figure 15A-B. The enrichments of genes associated with the upregulation of isoforms in infection strongly reflected and confirmed those seen in the differential gene expression analysis, including those relating to extracellular processes, immune response (with an even stronger enrichment), cytokine activity, and protein binding. Those associated with downregulated isoforms in infection also reflected previous results (Figure 6.15C-D), including many terms relating to intracellular and metabolic processes.

The 8 transcripts that were identified as differentially spliced in relation to *IFITM3* genotype included 6 upregulated in the CC variant, and 2 upregulated in the TT variant. The 6 transcripts for CC upregulation were HLA-DQB1-005, HLA-DPB1-003, HLA-DRB3-204, HLA-DPB1-002, GSKIP-005, and MMP12-001; while the 2 transcripts for TT upregulation were HLA-DRB3-001 and RAI1-004. Again, we see heavy emphasis on MHC class II cell surface receptor proteins of antigen-presenting cells. The *MMP12* gene is also of great interest in virology studies, with the usual role of its protein pertaining to degrading interferon alpha outside of the cell when no infection is present (Marchant *et al.* 2014). Anti-viral drugs have in fact been developed to aid the immune system by blocking MMP12 (Google Patents, 2015).


A
Observed
Expected
Biological Process Ontologies
immune response
multi-organism process
immune system process
response to cytokine
apoptotic process
Cellular Component Ontologies
extracellular region
extracellular region part
cytosol
ribosomal subunit
vesicle
Molecular Function Ontologies
cytokine activity
protein binding
poly(A) RNA binding
RNA binding
structural constituent of ribosome
0
200
400
600
800
1000
1200
Number of Genes
q value
10−28
10−21
10−14
10−7
10−0


B
Observed
Expected
KEGG Pathways
Cytokine-cytokine receptor interaction
Ribosome
RIG-I-like receptor signaling pathway
Toll-like receptor signaling pathway
Hepatitis C
0
20
40
60
80
Number of Genes
q value
10−14
10−7
10−0

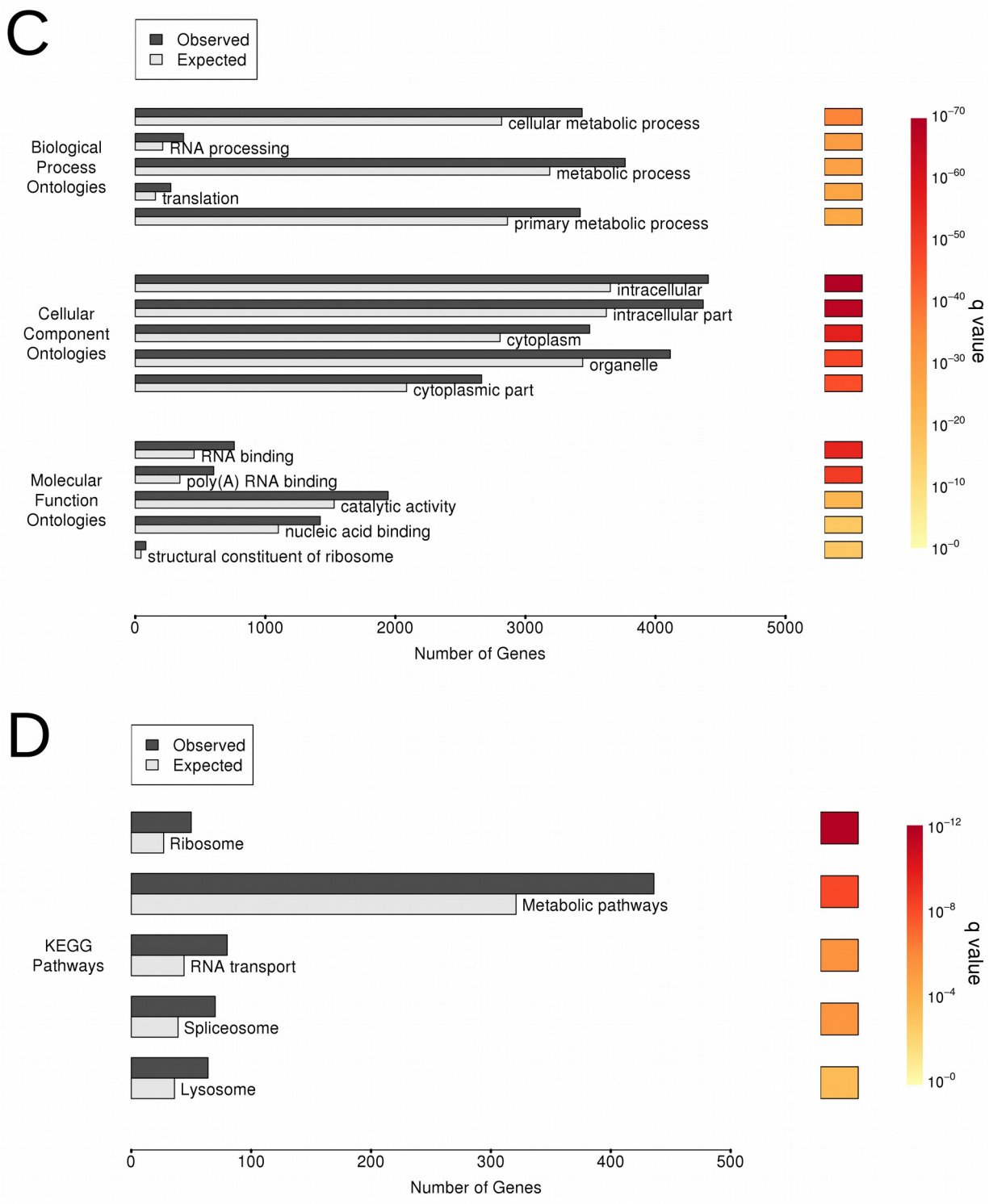


Figure 6.15. Bar plots of the top 5 significant functional annotations (by q-value) for GO terms and KEGG pathway of genes with transcripts upregulated (A-B) and downregulated (C-D) following infection, respectively. Bars represent the number of genes expected and observed for each term. Q-values are displayed in Hinton plots.

Differential Exon Analysis

In addition to the differential splicing analysis, differential exon analysis was executed, with the aim of pinpointing specific exons that may indicate a change in splicing. This was implemented to possibly expand the genotype differential splicing results, and was applied across samples of *IFITM3* CC and TT genotype, and for the interaction of these two.

Significantly differentially expressed (FDR < 0.05) exons indicative of alternative splicing were identified in 14 cases, 11 of which exhibited increased expression in the TT genotype, and 3 in the CC genotype. These 11 exons included: *PCGF1* (exon 6); *SRGAP2C* (exons 5 & 6); *IFITM1* / *IFITM2* (a shared exon, *IFITM1* exon 19); *HLA-DQA1* (exon 14); *RP1-283E3.4* (exons 5, 6 and 8); *HLA-DPB1* (exon 9); *AC093162.5* (exon 5); *RP11-324E6.6* (exon 7). Again, these results show strong associations with MHC class II surface receptor proteins of antigen-presenting cells, and ribosomal protein genes (as seen previously in the likelihood ratio testing of differentially expressed genes across the interaction). *PCGF1* and *SRGAP2C* appear unrelated, with *PCGF1* encoding a transcriptional repressor in development (Gong *et al.* 2005), and *SRGAP2C* being a well-known gene associated with the evolution of human brain development (Dennis *et al.* 2012). The occurrence of an exon shared between *IFITM1* and *IFITM2* which is shown to be upregulated is also of interest, and the relationship of splicing across the IFITM gene family could be worth further investigation. The 3 exons shown to be upregulated in the CC genotype were identified in genes *ATP1B1* (exon 11), *RP1-283E3.4* (exon 3), and *RP11-*

324E6.6 (exon 6). Along with ribosomal protein genes (as observed previously), *ATP1B1* is also observed, which encodes a sodium/potassium-transporting polypeptide and is known to interact with the influenza virus during replication (Mi *et al.* 2010).

Curiously, *IFITM3* is among none of these results. By visualizing the expression across the exons of *IFITM3* (Figure 6.16), we can see that this is likely due to most exons (1-3, 5-7, and 9) being differentially expressed, which would confound the detection of differential splicing.

Finally, differential exon testing was applied to the interaction of *IFITM3* genotype and infection status (FDR < 0.05), which highlighted the exons of 6 genes as sites of potentially influential splicing. These were identified in the following genes: *MAP2K7* (exon 29); *TRMT1* (exon 63); *SLC25A11* (exon 19); *SERHL* (exon 12); *TRAPPC6B* (exon 2); *RP11-324E6.6* (exon 2). As before, this includes the ribosomal protein *RP11-324E6.6*, as well as including genes encoding solute carrier / trafficking proteins, *SLC25A11* and *TRAPPC6B*. *MAP2K7* (otherwise known as *MKK7*) is also a noteworthy gene due to its associations with alternative splicing in mediating the cell response to cytokines and environmental stresses (Tournier *et al.* 2001).

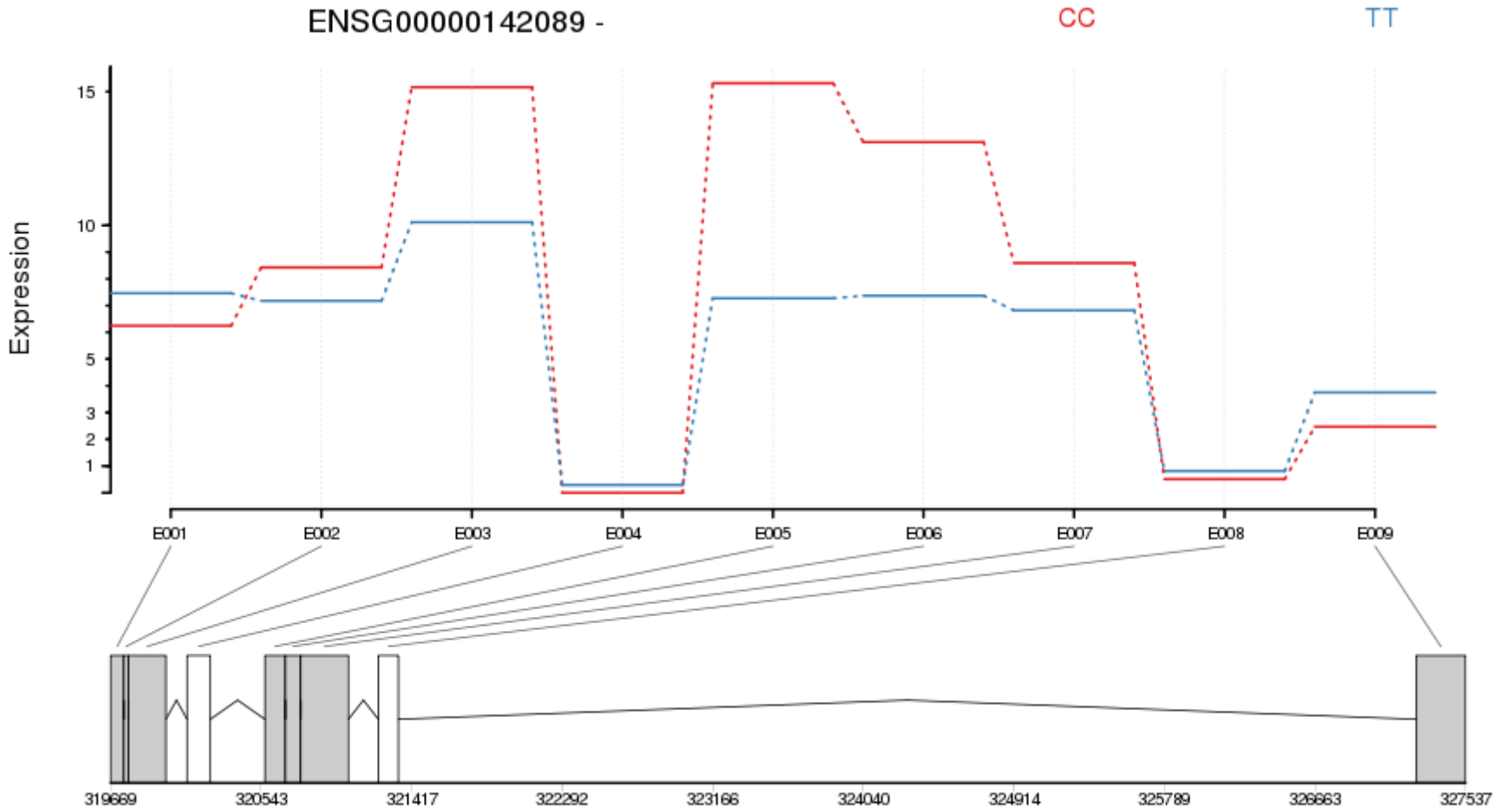


Figure 6.16. Exonic expression across the *IFITM3* gene, as determined by DEXSeq. Differentially expressed exons are represented in gray. The majority of differentially expressed exons display an elevation in the CC genotype, with the exception of the first and final exons, which exhibit an elevation in the TT genotype. The rs12252 SNP (for which the C allele causes truncation) is located at position 320,772.

6.4 Discussion

The IFITM3 protein is well understood to form a key component of the human innate immune system. In addition to this, the SNP rs12252-C variant of the *IFITM3* gene has been associated with increased susceptibility to severe influenza infection. This study takes a first look at the transcriptomes of individuals with differing rs12252 genotypes under infection, and examines gene expression of the human immune system -associated genes. Overall, this exploration of the transcriptomic effects of the variant provides an essential pilot study, presenting multiple routes for further investigation, including the apparent preferential expression of the rs12252 T-variant in infection of heterozygous cells, and the unearthing of multiple associated differential immune system pathways warranting further investigation. Other findings of note include the large transcriptomic-wide effect of infection, with 59.8% of genes reported as being differentially expressed, and the transcriptomic shift upon infection from expression of genes encoding intracellular proteins to expression of genes encoding extracellular proteins.

Investigation into expression of *IFITM3* showed read coverage across the entire gene (Figure 6.16), including in samples of the CC genotype, for which the truncated IFITM3 protein is implicated. Based on the previous literature (Everitt *et al.* 2012), there is an expectation that the CC genotype encodes a transcript that both lacks the first exon (Figure 6.16; E009), and alters the length of the exon containing the SNP (Figure C3). Neither is observed in this study. Instead, we observe even coverage across the locus with either allele.

The rs12252 variant is thought to alter a splice acceptor site, which would result in the truncation of the N-terminal 21 amino acid residues of the IFITM3 protein. These residues represent a region that is essential for the antiviral activity of IFITM3 (Weidner *et al.* 2010). Nevertheless, the original publication (Everitt *et al.* 2012) did not provide transcript sequence evidence for allele-specific alternative transcripts. Instead, it is possible that the effect of this SNP on the resultant protein could be in the efficiency of translation rather than of splicing or transcription. Although this SNP is synonymous, one potential consequence could be ribosomal stalling, which is a phenomenon that has been previously implicated as an outcome of some synonymous mutations, for example in genes such as MDR1 (Tsai *et al.* 2008). This could result in a preferential use of the downstream start codon in IFITM3 that is utilized for the alternative transcript and resultant truncated protein. Whatever the effect of the rs12252 SNP may be, this study does not corroborate the hypothesis of a splicing-transcriptional change. To confirm the lack of observed difference between genotypes observed in this RNA-Seq experiment, qPCR could be performed with primers designed to capture the expression of the alternative transcripts and evaluate the levels of each transcript in each patient. To investigate whether the effect on the resultant protein is occurring at the translational stage, ribosomal profiling (Ingolia *et al.* 2009) could be implemented to indicate which regions of the *IFITM3* transcript in homozygotes are bound by ribosomes and the abundance of different isoforms being translated.

Expression quantification of the two variants in the heterozygous samples

revealed an interesting difference across infected and uninfected samples. The allelic imbalance observed here is not uncommon, with some genome-wide RNA-Seq studies suggesting approximately 4.6% of heterozygous SNPs in humans displaying evidence of allelic imbalance (Heap *et al.* 2010). These imbalances have also been associated with viral infections, for example studies on the viral triggering of Crohn's Disease (also on human monocyte-derived dendritic cells) have shown changes in allelic imbalances in the *NOD2* gene (a gene also implicated in innate immune response) when infected with Newcastle Disease Virus (Hu & Peter, 2013). The allelic imbalance in IFITM3 should be further investigated to confirm this finding and to refine underlying mechanisms. The preliminary statistical testing that indicated this was a significant imbalance is a very promising result, but should be treated with caution. Although this test has been used before in assessing allelic frequencies in SNPs at a population level (Topa *et al.* 2015), it has not been used within a heterozygote individual, and the lack of independence between read counts (as reads are limited in number when sequencing) may have introduced a bias. Future investigation (following this pilot study) should be implemented with a larger sample size (to increase power), and also may require a customized exact test for obtaining a p-value. Specifically, this could be further investigated with targeted RNA-Seq at a greater depth and in more samples, and results could be verified with real time PCR providing higher sensitivity. Studies investigating the mechanisms underlying this allelic imbalance could also unearth greater understanding of the immune response at the gene expression level, which may provide useful information on drug targets.

The recurring theme of differential MHC class II cell surface receptor proteins co-occurring with different rs12252 genotypes is also of great interest. The HLA genes are the human version of the MHC complex observed in many species, which are differentiated into three classes. HLA-A, -B, and –C are observed in class I, and bind foreign peptides in the cell, presenting these on the cell surface and allowing the cell to be recognised as infected and destroyed. HLA-DPA1, -DPB1, -DQA1, -DQB1, -DRA, and -DRB1 are observed in class II, which differ in role by exporting the peptides and presenting antigens on immune system cells. All of the HLA genes we observed in differential analyses were of MHC class II, including the differential expression of the *HLA-DRB1* gene (raised in the TT genotype samples) and the *HLA-DPB1* gene (raised in the CC genotype samples). This was further supported by the discovery of the *RPL32P1* pseudogene, which is located within the *HLA-DPB1* gene (and similarly to the *HLA-DPB1* gene, exhibited increased expression of the CC genotype compared to the TT genotype, upon infection). *HLA-DPB1* also reoccurred in both differential splicing and exon analyses, in addition to the occurrence of *HLA-DQB1* and *HLA-DRB3* in differential splicing.

The HLA genes are known to be the most polymorphic of the human genome and reside in a gene rich region at 6p21.3 (Horton *et al.* 2004). The polymorphism of these genes is principally observed in HLA-A, -B and –C of class I, and HLA- DRB1, -DQB1 and –DPB1 of class II, all playing important role in ensuring pathogen recognition, and with over 12,200 HLA alleles have been identified thus far (Robinson *et al.* 2014). The large diversity of HLA

proteins amongst the human population is thought to have been driven by over-dominant selection, i.e. a heterozygote advantage (Hughes & Nei, 1988). Codominant expression of the HLA alleles provides a resultant increased repertoire of peptide motifs that can be bound and presented to T-cells. The ability of a pathogen to evade presentation by a particular HLA allele may therefore be compensated for by a second. This is further evidenced by the loci of many of the polymorphisms, which show functional relevance. Much of the polymorphism in HLA genes is located in exons encoding the peptide binding groove and the points at which the molecule interacts with the T-cells (Parham & Ohta, 1996).

With HLA genes of HLA-DP and HLA-DR featuring prominently in the results of this study, understanding the differing roles of the HLA genes within classes could provide important insight, and although these roles have not been well distinguished, noticeable differences include these levels of polymorphism. HLAs of class II display the greatest level of polymorphism in the functional region, which is the beta chain. HLA-DPB1 and HLA-DRB1 have vastly differing levels of polymorphism, with HLA-DPB estimated to have 422 alleles, and HLA-DRB estimated to have 1,540 alleles, the highest of all class II HLAs. However, the levels of polymorphism on the alpha chain are in stark contrast to this, with HLA-DPA1 estimated to have 38 alleles, and HLA-DRA estimated to have just 7 (Robinson *et al.* 2008). Although the role of this is unknown, it has been theorized that the higher level of allelic diversity in the functional motifs of HLA-DR indicates higher levels of balancing selection than for other class II HLAs (Parham & Ohta, 1996). Specific alleles of HLA-DR genes are also often

associated to multiple autoimmune disorders, including lupus, multiple sclerosis and rheumatoid arthritis (Jones *et al.* 2006), although the exact reason behind this is also not yet known.

The difference between HLAs observed in this study included an increase in expression of HLA-DRB1 in TT genotype samples, and a decrease in expression of HLA-DPB1 in CC (influenza-susceptible) genotype samples. One potential explanation for an increase in HLA-DPB1 could be that HLA-DP is utilized more in viral over bacterial antigens, and that samples with the CC variant have a greater reliance on the adaptive immune system over the innate system (that includes IFITM3). Indeed, the initial reports of HLA-DP function were all with regard to viral rather than bacterial antigens, including herpes simplex and influenza (Eckels *et al.* 1983), and other reports have associated HLA-DP with viruses such as hepatitis B (Kamatani *et al*. 2009), which has led to the postulation that HLA-DP could be more focussed on viral antigens (Lith *et al.* 2010). However, there is no definitive evidence and considering the case of our study where HLA-DR antigens show the opposing trend, HLA-DR has also been associated with viral immune response, for example in HIV-1 (Eller *et al.* 2011). Overall, more information will be required to draw any absolute conclusions, but further studies into the more precise roles of different HLAs may enable this in future.

This influenza pilot study is novel in both its analysis of the expression of *IFITM3*, including the different rs12252 alleles in heterozygote individuals, and in the identification of changing transcriptional signatures upon infection

dependent on this same *IFITM3* variant. This entirely novel discovery of a potential allelic bias could be indicative of an immune mechanism for increasing expression of the fully functional isoform in heterozygote individuals. Further study into any variation surrounding this occurrence, for example, through identification of pathways whose expression correlates with this allelic imbalance, could potentially reveal useful information on an immune system compensatory mechanism. Similarly, the identification of possible differences in HLA gene expression and splicing with differing *IFITM3* alleles has the potential to uncover entirely novel HLA gene role information. Both of these phenomena should be confirmed in a larger study (see Chapter 7).

Further investigation into these phenomena that have been uncovered in this study could provide essential insight into the genetic variants underlying influenza susceptibility in humans. This would include further observations of the expression of the *IFITM3* rs12252 variant at a cellular level, as well as providing novel insight into associated components of the human immune system. This chapter presents a successful pilot study, which should be followed by a larger study with greater numbers of heterozygous samples. Overall these findings can be used to guide future studies in increasing our understanding in fighting influenza, with the long term aim of developing targeted therapies as a new generation of influenza vaccinations.

Chapter 7 – Conclusions and Future Perspectives

In this thesis I present three novel projects investigating gene expression in various disease relevant tissues and structures, as well as an R package facilitating gene expression clustering and enrichment analyses that can be used in future projects. I implement these by utilizing the advantageous qualities of RNA-Seq to perform a series of bioinformatics analyses and uncover key findings relating these tissues to various diseases and disorders.

Future Directions

RNA-sequencing has been a central platform for analysis in all of the projects throughout this thesis, the results of which would benefit from further interrogation with a variety of NGS and non-NGS techniques.

One example of an informative future direction for the high confidence markers identified in placodes would be to investigate the temporal specificity of these markers using a time-series experiment. This could be performed with similar methods as were used in this thesis, with RNA-Seq and/or in-situ validation, or potentially with a sequencing method that provides greater information on mRNA translation, such as Ribo-Seq. By implementing these methods from stages just before the initial formation of the placodes through to full cranial nerve formation, results would confirm at which stage each marker could be

utilized, and hence would provide a much more in-depth resource for marker selection at a specific developmental stage. This may also shed light on the relationships of many of these markers to one another. A second future direction example would be to utilize a specific technique for spatially resolved transcriptomics, for example tomography-sequencing (Tomo-Seq). Tomo-Seq is designed to combine RNA-Seq with spatial information to give multi-dimensional expression pattern results (Junker *et al.* 2014). This is performed through sectioning of tissues into hundreds of thin slices before performing RNA-sequencing on each of these individuals sections. Subsequently, mathematical image reconstruction is used to determine genome-wide 3D expression patterns. This would allow for spatial results similar to the in-situ validations performed, but transcriptome-wide, thereby providing confirmation of marker localization on a much greater scale.

Future directions in analyses for the autism-enriched gene clusters of the Purkinje cells would likely be largely centred on conditional gene knockouts in mice. Conditional gene knockouts allow for the elimination of specific genes in specific cell types or developmental stages. This could be performed with knockout via the commonly used Cre-lox recombination system (by spanning the gene of interest with what are known as lox DNA sequences, Cre recombinase can be used to delete or invert the gene). Although Cre can be exogenously delivered to the target tissue to achieve this, in the case of this thesis it would be suitable to express Cre under a Purkinje cell specific promoter, such as that for *Calb1* or *Itpr1* (used as Purkinje cell specific markers in Figure 4.7). This would enable the use of conditional gene

knockouts specifically in the Purkinje cells during development, and could be performed in many of the autism-associated genes of the Purkinje cell gene cluster with significant increase in expression over developmental timing that was identified in this thesis. The resultant mice could be phenotyped (including identification of behaviour such as social avoidance, reduced ultrasonic vocalizations, etc; see Silverman *et al.* 2010 for a more comprehensive list of autism-associated mouse phenotypes) to confirm the exact effects of the knockout of these genes from the Purkinje cells.

Finally, future directions for the influenza *IFITM3* rs12252 project will likely include expanding from a pilot study to a much larger study. In designing this experiment, it would ideally involve a much greater number of replicates and a number of different strains of influenza virus. In general this larger study would provide much greater power for analyses in subsequent experimental procedures. To specifically address the differential expression/splicing of HLAs and the allelic bias of rs12252 identified in this thesis, targeted RNA-Seq and/or qPCR could then be used. Targeted RNA-Seq would allow for a much greater depth to be achieved (both on the HLA genes and *IFITM3*) for a more powerful analysis of expression/splicing. Similarly, qPCR can be used to measure abundance of mRNA reliably in these targets. A final potential future direction for this project would be to perform RNA-Seq in single cells, as this could reflect the heterogeneous nature of individual immune cells (see *Future Directions of RNA-Seq*).

Future Directions of Next Generation Sequencing

As discussed in the introduction to this thesis, the novel qualities of next generation sequencing allow the transcriptome to be interrogated at great depth. Nevertheless, these are still met with multiple complications, including biases (for example in GC-rich and GC-poor regions) and the loss of information through short read sequencing. There are multiple technologies, either developed or in development, which are aimed at competing with the more commonly used platforms (namely from Illumina) through offering solutions to many of these disadvantages. Major examples of these include those offered by Ion Torrent Systems (Ion Torrent), Pacific Biosciences of California (PacBio), and Oxford Nanopore Technologies (Oxford Nanopore).

Ion Torrent platforms are based on the same sequencing by synthesis method as that of Illumina systems, but implement synthesis through the detection of hydrogen ions that are released upon DNA polymerization (Rothberg *et al.* 2011). The detection occurs through hydrogen ions triggering an ion-sensitive field-effect transistor. Subsequently, no modified nucleotides or optics are required, which potentially allows the sequencing to be both faster and cheaper than Illumina sequencers. However, these systems have suffered from problems with lower throughput and shorter read length (Rusk, 2011), and do not address the biases and performance issues experienced with sequencing by synthesis. PacBio systems are aimed at reducing some of these issues, offering single molecule real time sequencing. This utilizes a large number of optical waveguides (a structure that guides waves in the optical spectrum, such as an

optical fibre) small enough to observe a single nucleotide of DNA labelled with fluorescent dye being incorporated by DNA polymerase (Korlach *et al.* 2010). These are known as zero-mode waveguide wells, and provide the main advantage of allowing sequencing of much longer reads than that of Illumina, with typical runs averaging 10-15kb reads, and the longest reads exceeding 40kb (Pacific Biosciences, 2014). These longer reads, spanning much larger distances of the transcriptome, can allow for the true quantification of entire isoforms, as well as the ability to cover repeating regions. PacBio systems have also been shown to have much lower GC bias than Illumina systems (Ferrarini *et al.* 2013); however disadvantages to this method of sequencing include a higher error rate and a greater cost per nucleotide (Ferrarini *et al.* 2013). Finally, Oxford Nanopore also offers a very promising method of sequencing, involving the use of nanopores through which an ionic current is passed. As DNA strands pass through the nanopore, the current is disrupted, and the distinctive disruption caused by each of the DNA bases can be used in sequencing (Howorka *et al.* 2001). There are many advantages provided by such a relatively simple method, including the sequencing of longer reads in both a faster and cheaper manner. Read lengths for the Oxford Nanopore MinION have been reported to average 1-2kb, with the longest reads reaching up to 98kb (Laver *et al.* 2015). Much like PacBio systems, Oxford Nanopore systems have also been found to return relatively little GC bias; however these have been shown to be susceptible to a high error rate (Laver *et al.* 2015).

The numerous advantages of this next generation of sequencers show great promise for the sequencing of RNA. Potential for a reduction in bias and in cost

are welcome improvements. As previously mentioned, an increase in read length can also provide many more options for advanced analysis of RNA, including more in-depth scrutiny of splicing isoforms (not relying on probabilistic models) due to the spanning of multiple splice junctions, the ability to analyse repeat regions due to the potential for reads to span these regions (and hence align uniquely to), and an increased ability to identify the phase of observed variants. Overall, this third generation of sequencers will likely bring more utilizable, accurate and numerous RNA-Seq data sets in the near future.

Future Directions of RNA-Seq

Alongside the further development of next generation sequencing platforms, more sophisticated techniques for manipulation of RNA for sequencing are also emerging, most notably with the advent of single-cell RNA-Seq (scRNA-Seq). Through the development of techniques for the isolation of single cells, and amplification of the subsequent extracted RNA, it has been made possible to perform sequencing on the RNA of a single cell (Hebenstreit, 2012). This provides multiple avenues for potential analyses, for example in a developmental setting, in which there are cases where only a small number of cells convey a distinct role (Ohnishi *et al.* 2014). Additionally, it will allow for the analysis of highly complex tissues, for example in the brain, which contains a vast number of cell types that are difficult to scrutinize experimentally (Belgard *et al.* 2011).

The benefits of scRNA-Seq could prove advantageous in many studies regardless of the objectives. Every RNA-Seq project in this thesis could have been improved or advantageously supplemented by scRNA-Seq. In every case, the resolution of the studies in this thesis was maximized as much as possible within the limitations of bulk RNA-Seq, with attempts to obtain the purest cell-type population possible, for example through fluorescence-activated cell sorting in the placode and influenza projects (Chapters 3 and 6 respectively) and laser-capture microdissection in the autism project (Chapter 4). Utilizing scRNA-Seq would allow for purer samples for analysis and ensure accurate identification of cell types and subsequent results. The use of scRNA-Seq could also provide additional distinct potential analyses for each project. For the placode project this could include the identification of markers in single cells of specific stages of development. Studies in placodes have already identified expression of regulatory genes at a single cell resolution, for example *Pax3* was shown to correlate with neuronal specification at the single cell level in the ophthalmic placode (Baker & Bronner-Fraser, 2000). Investigation of single cells at a transcriptome-wide level would likely uncover many more regulatory genes or genetic markers, which could be utilized to fully understand the transition to specific cell types. For example in the otic placode, determining regulatory networks underlying the specific development of inner ear hair cells could be useful for the development of these *in vitro*, which could subsequently aid in studies on (for example) hearing loss. For the Purkinje project, scRNA-Seq could be used similarly in identifying the gene expression signatures underlying Purkinje cell development. The use of scRNA-Seq in this project could also have ensured a purer Purkinje cell population, through profiling of

each cell sequenced and utilizing markers to identify each. Subsequent analysis on the Purkinje cells only could affirm any findings as directly associated with Purkinje cells. The capture of cells would occur without laser-capture microdissection, meaning the RNA could have incurred less damage (with higher RIN values) and potentially could have provided data of higher quality for analysis. Previous investigation of neurons at a single cell level with RNA-Seq has also shown individual neurons to have distinct gene expression patterns, despite being morphologically identical (Qiu *et al.* 2012). Single-cell RNA-Seq could be used to assess the levels of heterogeneity amongst Purkinje cells, and allow for pinpointing of the same cluster identified in this thesis (which was shown to be enriched for genes associated with autism) amongst specific cells. Finally, for the influenza project, scRNA-Seq would provide advantages in the same vein as both the aforementioned projects. Similarly to the placode project, scRNA-Seq could provide the ability to track changes at a single cell resolution for an in depth analysis of the changes in gene expression patterns. This could identify specific associations (for example correlations) between *IFITM3* expression levels, infection status, and expression levels of other genes associated with immune response, in different cells. Similarly to the Purkinje project, scRNA-Seq could unearth detailed information on the diversity of the immune response across heterogeneous immune cells. The heterogeneity of immune cells is known to be vast, and Chattopadhyay *et al.* (2014) have postulated that to enable truly robust identification of the unique combination of traits defining a subset of cells involved in immunological response, technologies need to emphasize both depth and breadth of measurements at single-cell resolution.

Although scRNA-seq provides many advantages in potential additional analyses, and in offering analysis at a greater resolution, from a bioinformatics perspective this new approach also comes with a number of analytical challenges, which many are still working to address (Stegle *et al.* 2015). For example, in addition to the quality control strategies necessary for bulk RNA-Seq (such as library quality and PCA to confirm correct samples), it is also important to consider the increased likelihood of degradation of the RNA. Biases, such as 3' bias, have also been reported as being increasingly problematic.

Although this is a newly emerging technology, and still presents many challenges to analytical approaches, scRNA-Seq has already produced interesting results in diseases and disorders. One major example of this is in cancer. As cancer is a highly complex and heterogeneous disease, a single tumour can contain many clinically important subpopulations, for which scRNA-Seq has proved effective in understanding (Min *et al.* 2015). The identification and subsequent interrogation of these subpopulations can be important in understanding common and uncommon properties across tumours to aid the development of targeted biological therapies. Another application of scRNA-Seq in cancer is in circulating tumour cells, which are found in the blood of cancer patients, but are notoriously challenging to profile due to commonly being contaminated with other cell types. Single-cell RNA-Seq has been used to isolate and characterize these cells, which have been found to be useful in investigating disease prognosis and drug resistance (Miyamoto *et al.* 2015).

Overall, this is evidence that, while RNA-Seq itself is becoming an increasingly established technique over microarrays, it is also still developing further and becoming progressively more useful in interrogating different aspects of transcriptomes. The future of RNA-Seq shows great promise in unearthing even more information on the mechanisms underlying many diseases and disorders.

References

Abbeduto, L., McDuffie, A. and Thurman, A. J. (2014). The fragile X syndrome-autism comorbidity: what do we really know? Front. Genet. *5: 355.*

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control. *19(6): 716-723.*

Alekseyenko, A. V., Kim, N. and Lee, C. J. (2007). Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. RNA. *13: 661–670.*

Alexa, A., Rahnenführer, J. and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. *22(13): 1600-1607.*

Allison, M. B., Patterson, C. M., Krashes, M. J., Lowell, B. B., Myers, M. G. and Olson, D. P. (2015). TRAP-seq defines markers for novel populations of hypothalamic and brainstem LepRb neurons. Molecular Metabolism. *4(4): 299–309.*

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11(10):R106.*

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W. and Robinson M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat. Protocols. *8:1765-1786.*

Anders, S., Pyl, P. T. and Huber, W. (2015). HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics 32(2):166-169.*

Anders, S., Reyes, A. and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22(10): 2008-2017.*

Anderson, B. J., Alcantara, A. A. and Greenough, W. T. (1996). Motor-skill learning: changes in synaptic organization of the rat cerebellar cortex. Neurobiol. Learn. Mem. *66(2): 221-9.*

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. *Available online at: http://www.bioinformatics.babraham.ac.uk /project/fastqc.*

Arbusow, V., Strupp, M., Wasicky, R., Horn, A. K. E., Schulz, P. and Brandt, T. (2000). Detection of herpes simplex virus type 1 in human vestibular nuclei. Neurology. *55: 880-882.*

Archer, S. K., Inchaustegui, D., Queiroz, R. and Clayton, C. (2011). The Cell Cycle Regulated Transcriptome of Trypanosoma brucei. PLoS ONE

6(3):e18425.

Ariëns-Kappers, C. U., Huber, G. C. and Crosby, E. C. (1960). The comparative anatomy of the nervous system of vertebrates, including man. New York: Hafner Publishing. *11: 433-439.*

Baker, C. V. and Bronner-Fraser, M. (2000). Establishing neuronal identity in vertebrate neurogenic placodes. Development. *127(14): 3045-56.*

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J. and Lipman, D. (2008). The influenza virus resource at the National Center for Biotechnology Information. J. Virol. 82(2): 596-601.

Barlow, L. A. (2002). Cranial Nerve Development: Placodal Neurons Ride the Crest. Current Biology. *12(5): R171-173.*

Basu, S. N., Kollu, R. and Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. Nucleic Acids Res. *37: D832–D836.*

Baumgart, M., Groth, M., Priebe, S., Savino, A., Testa, G., Dix, A., Ripa, R., Spallotta, F., Gaetano, C., Ori, M., Tozzini, E. T., Guthke, R., Platzer, M. and Cellerino, A. (2014). RNA-Seq of the aging brain in the short-lived fish N. furzeri-conserved pathways and novel genes associated with neurogenesis. Aging Cell. *13(6): 965-974.*

Becker, E.B.E. (2014). The Moonwalker mouse: new insights into TRPC3 function, cerebellar development, and ataxia. Cerebellum. *13: 628–636*.

Becker, E. B. E. and Stoodly, C. J. (2013). Autism spectrum disorder and the cerebellum. Int. Rev. Neurobiol. *113: 1-34.*

Becker, E. B. E., Oliver, P. L., Glitsch, M. D., Banks, G. T., Achilli, F., Hardy, A., Nolan, P. M., Fisher, E. M. C. and Davies, K. E. (2009). A point mutation in TRPC3 causes abnormal Purkinje cell development and cerebellar ataxia in moonwalker mice. Proc. Natl. Acad. Sci. USA. *106: 6706–6711.*

Begbie, J., Brunet, J., Rubenstein, J. L. R. and Graham, A. (1999). Induction of the epibranchial placodes. Development. *126: 895-902.*

Belgard, T. G., Marques, A. C., Oliver, P. L., Abaan, H. O., Sirey, T. M., Hoerder-Suabedissen, A., García-Moreno, F., Molnár, Z., Margulies, E. H. and Ponting, C. P. (2011). A Transcriptomic Atlas of Mouse Neocortical Layers. Neuron*. 71: 605–616.*

Bell, D., Streit, A., Gorospe, I., Varela-Nieto, I., Alsina, B. and Giraldez, F. (2008). Spatial and temporal segregation of auditory and vestibular neurons in the otic placode. Developmental Biology. *322(1): 109-120.*

Belonga, E. A., Kieke, B. A., Donahue, J. G., Greenlee, R. T., Balish, A., Foust,

A., Lindstrom, S., Shay, D. K. and The Marshfield Influenza Study Group (2009). Effectiveness of inactivated influenza vaccines varied substantially with antigenic match from the 2004-2005 season to the 2006-2007 season. J. Infect. Dis. *199(2): 159-167.*

Belonga, E. A., Kieke, B. A., Donahue, J. G., Coleman, L. A., Irving, S. A., Meece, J. K., Vandermause, M., Lindstrom, S., Gargiullo, P. and Shay, D. K. (2011). Influenza vaccine effectiveness in Wisconsin during the 2007-08 season: comparison of interim and final results. Vaccine. *29(38): 6558-63.*

Benjamin, A. M., Nichols, M., Burke, T. W., Ginburg, G. S. and Lucas, J. E. (2014). Comparing reference-based RNA-Seq mapping methods for non-human primate data. BMC Genomics *15:570.*

Benjamini, Y. and Hochberg, Y. (1995). Controlling false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B. *57(1): 289-300.*

Ben-Yehudah, G., Guediche, S. and Fiez, J. A. (2007). Cerebellar contributions to verbal working memory: beyond cognitive theory. Cerebellum. *6(3): 193-201.*

Berg, J. M. and Geschwind, D. H. (2012). Autism genetics: searching for specificity and convergence. Genome Biol. *13: 247.*

Blake, J. A., Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E. and The Mouse Genome Database Group (2014). The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res. *42: D810–D817.*

Brass, A. L., Huang, I. C., Benita, Y., John, S. P., Krishnan, M. N., Feeley, E. M., Ryan, B., Weyer, J. L., van der Weyden, L., Fikrig, E., Adams, D. J., Xavier, R. J., Farzan, M. and Elledge, S. J. (2009). IFITM proteins mediate the innate immune response to influenza A H1N1 virus, West Nile virus and dengue virus. Cell. *139(7): 1243-1254.*

Brunet, J. F. and Pattyn, A. (2002). Phox2 genes – from patterning to connectivity. Curr. Opin. Genet. Dev. *12(4):435-40.*

Budimirovic, D. B. and Kaufmann, W. E. (2011). What can we learn about autism from studying fragile X syndrome? Dev. Neurosci. *33(5): 379-394.*

Buschow, S. I., van Balkom, B. W., Aalberts, M., Heck, A. J., Wauben, M. and Stoorvogel, W. (2010). MHC class II-associated proteins in B-cell exosomes and potential functional implications for exosome biogenesis. Immunol. Cell Biol. *88(8): 851-856.*

Buzzi, A., Wu, Y., Frantseva, M. V., Perez Velazquez, J. L., Cortez, M. A., Liu, C. C., Shen, L. Q., Gibson, K. M. and Snead, O. C. (2006). Succinic semialdehyde

dehydrogenase deficiency: GABAB receptor-mediated function. Brain Research. *1090: 15-22.*

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics – theory and methods. *3(1): 1-27.*

Camacho, C. P., Lindsey, S. C., Melo, M. C., Yang, J. H., Germano-Neto, F., Valente, F. O. F., Lima, T. R., Biscolla, R. P., Vieira, J. G., Cerutti, J. M., Dias-da-Silva, M.R. and Maciel, R. M. (2013). Measurement of calcitonin and calcitonin gene-related peptide mRNA refines the management of patients with medullary thyroid cancer and may replace calcitonin-stimulation tests. Thyroid *23(3): 308-316.*

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Research. *32: D262.*

Camon, E. B., Barrell, D. G., Dimmer, E. C., Lee, V., Magrane, M., Maslen, J., Binns, D. and Apweiler, R. (2005). An evaluation of GO annotation retrieval for BioCreAtivE and GOA. BMC Bioinformatics*. 6(Suppl 1).*

Castelli, F. A., Buhot, C., Sanson, A., Zarour, H., Pouvelle-Moratille, S., Nonn, C., Gahery-Ségard, H., Guillet, J. G., Ménez, A., Georges, B. and Maillère, B. (2002). HLA-DP4, the most frequent HLA II molecule, defines a new supertype of peptide-binding specificity. J. Immunol. *169(12): 6928-34.*

Cate, T. R. (1987). Clinical manifestations and consequences of influenza. American Journal of Medicine. *82(6A): 15-19.*

Centers for Disease Control and Prevention (2015). Seasonal Influenza Vaccine Effectiveness, 2005-2015. *Available from http://www.cdc.gov/flu/professionals /vaccination/effectiveness-studies.htm [accessed 3/10/2015].*

Centers for Disease Control and Prevention (2015). CDC presents updated estimates of flu vaccine effectiveness for the 2014-2015 season. *Available from http://www.cdc.gov/flu/news/updated-vaccine-effectiveness-2014-15.htm [accessed 3/10/2015].*

Chattopadhyay, P. K., Gierahn, T. M., Roederer, M. and Love, J. C. (2014). Single-cell technologies for monitoring immune systems. Nature Immunology. *15(2): 128–135.*

Chen, Y., Auer-Grumbach, M., Matsukawa, S., Zitzelsberger, M., Themistocleous, A. C., Strom, T. M., Samara, C., Moore, A. W., Cho, L. T., Young, G. T., Weiss, C., Schabhuttl, M., Stucka, R., Schmid, A. B., Parman, Y., Graul-Neumann, L., Heinritz, W., Passarge, E., Watson, R. M., Hertz, J. M., Moog, U., Baumgartner, M., Valente, E. M., Pereira, D., Restrepo, C. M., Katona, I., Dusl, M., Stendel, C., Wieland, T., Stafford, F., Reimann, F., von Au, K.,

Finke, C., Willems, P. J., Nahorski, M. S., Shaikh, S. S., Carvalho, O. P., Nicholas, A. K., Karbani, G., McAleer, M. A., Cilio, M. R., McHugh, J. C., Murphy, S. M., Irvine, A. D., Jensen, U. B., Windhager, R., Weis, J., Bergmann, C., Rautenstrauss, B., Baets, J., De Jonghe, P., Reilly, M. M., Kropatsch, R., Kurth, I., Chrast, R., Michiue, T., Bennett, D. L., Woods, C. G. and Senderek, J. (2015). Transcriptional regulator PRDM12 is essential for human pain perception. Nature Genet. *47: 803-808.*

Chesson, J. (1976). A Non-Central Multivariate Hypergeometric Distribution Arising from Biased Sampling with Application to Selective Predation. Journal of Applied Probability. *13(4): 795-797.*

Chico, L. K., Van Eldik, L. J. and Watterson, D. M. (2009). Targeting protein kinases in central nervous system disorders. Nature Reviews Drug Discovery. *8(11): 892–909.*

Chiu, I. M., Barrett, L. B., Williams, E. K., Strochlic, D. E., Lee, S., Weyer, A. D., Lou, S., Bryman, G. S., Roberson, D. P., Ghasemlou, N., Piccoli, C., Ahat, E., Wang, V., Cobos, E. J., Stucky, C. L., Ma, Q., Liberles, S. D. and Woolf, C. J. (2014). Transcriptional profiling at whole population and single cell levels reveals somatosensory neuron molecular diversity. Elife. *3.*

Cho, S. G., Lee, J. W., Heo, J. S. and Kim S. Y. (2014). Gene expression change in human dental pulp cells exposed to a low-level toxic concentration of triethylene glycol dimethacrylcate: an RNA-Seq analysis. Basic Clin. Pharmacol. Toxicol. *115(3): 282-90.*

Chopra, R. and Shakkottai, V. G. (2014). Translating cerebellar Purkinje neuron physiology to progress in dominantly inherited ataxia. Future Neurol. *9: 187–196.*

Chow, M. L., Pramparo, T., Winn, M. E., Barnes, C. C., Li, H. R., Weiss, L., Fan, J. B., Murray, S., April, C., Belinson, H., Fu, X. D., Wynshaw-Boris, A., Schork, N. J. and Courchesne, E. (2012). Age-dependent brain gene expression and copy number anomolies in autism suggest distinct pathological processes at young versus mature ages. PLoS Genetics. *8(3): e1002592.*

Clifford, H., Wessely, F., Pendurthi, S. and Emes, R. D. (2011). Comparison of clustering methods for investigation of genome-wide methylation array data. Front. Genet. *2:88.*

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solex/Illumina FASTQ variants. Nucleic Acids Res. *38(6): 1767-1771.*

Cooper, C. W. (1993). Vestibular neuronitis: a review of a common cause of vertigo in general practice. J. Gen. Pract. *43(469): 164-167.*

Costa, V., Aprile, M., Esposito, R. and Ciccodicola, A. (2013). RNA-Seq and

human complex diseases: recent accomplishments and future perspectives. Eur J Hum Genet. *21(2): 134-142.*

Courchesne, E., Mouton, P. R., Calhoun, M. E., Semendeferi, K., Ahrens-Barbeau, C., Hallet, M. J., Barnes, C. C. and Pierce, K. (2011). Neuron number and size in prefrontal cortex of children with autism. JAMA. *306(18): 2001-10.*

Crick, F. (1970). Central dogma of molecular biology. Nature. *227(5258): 561-563.*

D'Amico-Martel, A. and Noden, D. M. (1983). Contributions of placodal and neural crest cells to avian cranial peripheral ganglia. American Journal of Anatomy. *166(4): 445-468.*

D'Autréaux, F., Coppola, E., Hirschm M., Birchmeier, C. and Brunet, J. (2011). Homeoprotein Phox2b commands a somatic-to-visceral switch in cranial sensory pathways. Proc. Nat. Acad. Sci. *108(50): 20018-20023.*

D'Mello, A. M., Crocetti, D., Mostofsky, S. H. and Stoodley, C. J. (2015). Cerebellar gray matter and lobular volumes correlate with core autism symptoms. NeuroImage: Clinical. *7: 631-9.*

Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y. S., Mele, A., Fraser, C. E., Stone, E. F., Chen, C., Fak, J. J., Chi, S. W., Licatalosi, D. D., Richter, J. D. and Darnell, R. B. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell. *146(2): 247-261.*

Deluca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M. D., Williams, C., Reich, M., Winckler, W. and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics. 28(11): 1530-1532.*

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B. *39(1): 1-38.*

Dennis, M. Y., Nuttle, X., Sudmant, P. H., Antonacci, F., Graves, T. A., Nefedov, M., Rosenfeld, J. A., Sajjadian, S., Malig, M., Kotkiewicz, H., Curry, C. J., Shafer, S., Shaffer, L. G., de Jong, P. J., Wilson, R. K. and Eichler, E. E. (2012). Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. Cell. *149(4): 912-922.*

Denoth, F., Magherini, P. C., Pompeiano, O. and Stanojević, M. (1979). Responses of Purkinje cells of the cerebellar vermis to neck and macular vestibular inputs. Pflügers Arch. *381: 87-98.*

Desai, T. M., Marin, M., Chin, C. R., Savidis, G., Brass, A. L. and Melikyan, G. B. (2014). IFITM3 restricts influenza A virus entry by blocking the formation of fusion pores following virus-endosome hemifusion. PloS Pathog. *10(4):*

e1004048.

Devlin, B. and Scherer, S. W. (2012). Genetic architecture in autism spectrum disorder. Curr. Opin. Genet. Dev. *22: 229–237.*

Di Zazzo, E., De Rosa, C., Abbondanza, C. and Moncharmont, B. (2013). PRDM Proteins: Molecular Mechanisms in Signal Transduction and Transcriptional Regulation. Biology. *2(1): 107–141.*

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M. and Jaffrezic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics*. 14: 671–683.*

Dinger, M. E., Pang, K. C., Mercer, T. R. and Mattick, J. S. (2008). Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. PloS Comput Biol *4(11):e1000176.*

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics. *29(1): 15-21.*

Draghici, S., Khatri, P., Eklund, A. C. and Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. *22(2): 101-109.*

Dude, C. M., Kelly Kuan, C. Y., Bradshaw, J. R., Greene, N. D. E., Relaix, F., Stark, M. R. and Baker, C. V. H. (2010). Activation of Pax3 target genes is necessary but not sufficient for neurogenesis in the ophthalmic trigeminal placode. Dev. Biol. *326(2): 314-326*

Eckels, D. D., Lake, P., Lamb, J. R., Johnson, A. H., Shaw, S., Woody, J. N. and Hartzman, R. (1983). *SB*-restricted presentation of influenza and herpes simplex virus antigens to human T-Iymphocyte clones. Nature. *301: 716–718.*

El Amrani, M., Ducros, A., Boulan, P., Aidi, S., Crassard, I., Visy, J. M., Tournier-Lasserve, E. and Bousser, M. G. (2002). Familial cluster headache: a series of 186 index patients. Headache. *42: 974-977.*

Eller, M. A., Blom, K. G., Gonzalez, V. D., Eller, L. A., Naluyima, P., Laeyendecker, O., Quinn, T. C., Kiwanuka, N., Serwadda, D., Sewankambo, N. K., Tasseneetrithep, B., Wawer, M. J., Gray, R. H., Marovich, M. A., Michael, N. L., de Souza, M. S., Wabwire-Mangen, F., Robb, M. L., Currier, J. R. and Sandberg, J. K. (2011). Innate and Adaptive Immune Responses Both Contribute to Pathological CD4 T Cell Activation in HIV-1 Infected Ugandans. PLoS ONE. *6(4): e18779.*

Eswaran, J., Horvath, A., Godbole, S., Reddy, S. D., Mudvari, P., Ohshiro, K., Cyanam, D., Nair, S., Fuqua, S. A., Polyak, K., Florea, L. D. and Kumar, R. (2013) RNA sequencing of cancer reveals novel splicing alterations. Sci. Rep. *3:1689.*

Everitt, A. R., Clare, S., Pertel, T., John, S. P., Wash, R. S., Smith, S. E., Chin, C. R., Feeley, E. M., Sims, J. S., Adams, D. J., Wise, H. M., Kane, L., Goulding, D., Digard, P., Anttila, V., Baillie, J. K., Walsh, T. S., Hume, D. A., Palotie, A., Xue, Y., Colonna, V., Tyler-Smith, C., Dunning, J., Gordon. S. B., The GenISIS Investigators, The MOSAIC Investigators, Smyth, R. L., Openshaw, P. J., Dougan, G., Brass, A. L. and Kellam, P. (2012). IFITM3 restricts the morbidity and mortality associated with influenza. Nature. *484(7395): 519-23.*

Faghihi, M. A. and Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. Nat. Rev. Mol. Cell. Biol. *10: 637–643.*

Fare, T. L., Coffrey, E. M., Dai, H., He, Y. D., Kessler, D. A., Kilian, K. A., Koch, J. E., Le Proust E., Marton, M. J., Meyer, M. R., Stoughton, R. B., Tokiwa, G. Y. and Wang, Y. (2003). Effects of atmospheric ozone on microarray data quality. Anal. Chem. *75(17): 4672-4675.*

Feng, Y. and Xu, Q. (2009). Pivotal role of hmx2 and hmx3 in zebrafish inner ear and lateral line development. Dev. Biol. *339(2): 507-18.*

Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A. and Sargent, D. J. (2013). An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics. *14: 670.*

Fertuzinhos, S., Li, M., Kawasawa, Y. I., Ivic, V., Franjic, D., Singh, D., Crair, M. and Šestan, N. (2014). Laminar and temporal expression dynamics of coding and noncoding RNAs in the mouse neocortex. Cell Reports. *6: 938–950.*

Fisher, R. A. (1922). On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society. *85(1): 87-94.*

Flannery, B., Clippard, J., Zimmerman, R. K., Norwalk, M. P., Jackson, M. L., Jackson, L. A., Monto, A. S., Petrie, J. G., McLean, H. Q., Belongia, E. A., Gaglani, M., Berman, L., Foust, A., Sessions, W., Thaker, S. N., Spencer, S. and Fry, A. M. (2015). Early estimates of seasonal influenza vaccine effectiveness – United States, January 2015. MMWR Morb Mortal Wkly Rep*. 64(1): 10-15.*

Flicek, P., Ridwan Amode, P. M, Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F.,

Harrow, J., Herrero, J., Hubbard, T. J. P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R. and Searle, S. M. J. (2014). Ensembl 2014. *Nuc Acids Res. 42:D749-D755.*

Fonseca, N. A., Rung, J., Brazma, A. and Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. Bioinformatics *28(24): 3169-3177.*

Frazee, A. C., Langmead, B. and Leek, J. T. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics. *12: 449.*

Gao, Q., Brydon, E. W. A. and Palese, P. (2008). A seven-segmented influenza A virus expressing the influenza C virus glycoprotein HEF. J. Virol. *82(13): 6419-6426.*

Geschwind, D. H. (2009). Advances in autism. Annu. Rev. Med. *60: 367-80.*

Glaus, P., Honkela, A. and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics. *28(13): 1721-8.*

Glennon, N. B., Jabado, O., Lo, M. K. and Shaw, M. L. (2015). Transcriptome profiling of the virus-induced innate immune response in Pteropus vampyrus and its attenuation by Nipah virus interferon antagonist functions. J. Virol. *00302-15 [Epub ahead of print].*

Gong, Y., Wang, X., Liu, J., Shi, L., Yin, B., Peng, X., Qiang, B. and Yuan, J. (2005). NSPc1, a mainly nuclear localized protein of novel PcG family members, has a transcription repression activity related to its PKC phosphorylation site at S183. FEBS Lett. *579(1): 115-121.*

Goodstadt, L. (2010). Ruffus: a lightweight Python library for computational pipelines. Bioinformatics *26(21): 2778-2779.*

Google Patents - the designation of the inventor has not yet been filed (2015). Inhibitors of MMP-12 as antiviral agents*. EP 2907512 A1. Available from http://www.google.com/patents/EP2907512A1 [accessed 30/10/15].*

Gottwald, B., Mihajlovic, Z., Wilde, B. and Mehdorn, H. M. (2003). Does the cerebellum contribute to specific aspects of attention? Neuropsychologia*. 41(11): 1452-1460.*

Graham, A. and Shimeld, S. (2013). The origin and evolution of the ectodermal placodes. Journal of Anatomy. *222(1): 32-40.*

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., Brown, J.B., Cherbas, L., Davis, C.A., Dobin, A., Li, R., Lin, W., Malone, J.H., Mattiuzzo, N.R., Miller, D., Sturgill, D., Tuch, B.B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads,

B., Green, R.E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J.E., Wan, K.H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P.J., Brenner, S.E., Brent, M.R., Cherbas, P., Gingeras, T.R., Hoskins, R.A., Kaufman, T.C., Oliver, B. and Celniker, S.E. (2011). The Developmental Transcriptome of Drosophila melanogaster, Nature. *471(7339): 473-479*.

Green, D., Charman, T., Pickles, A., Chandler, S., Loucas, T., Simonoff, E. and Baird, G. (2009). Impairment in movement skills of children with autistic spectrum disorder. Developmental medicine & child neurology. *51(4): 311-316.*

Greener, M. (2015). Don't underestimate glial cells. Progress in Neurology and Psychiatry. *19(1): 5-8.*

Griffin, M. R., Monto, A. S., Belongia, E. A., Treanor, J. J., Chen, Q., Chen, J., Talbot, H. K., Ohmit, S. E., Coleman, L. A., Lofthus, G., Petrie, J. G., Meece, J. K., Hall, C. B., Williams, J. V., Gargiullo, P., Berman, L., Shay, D. K. and The U. S. Flu-VE Network (2011). Effectiveness of non-adjuvanted pandemic influenza A vaccines for preventing pandemic influenza acute respiratory illness visits in 4 U. S. communities. PLoS One. *6(8): e23085.*

Gupte S. (2011). Introduction. Influenza: Complete Spectrum-1. Gurgaon (NCR): Elsevier.

Haldane, J. B. (1956). The estimation and significance of the logarithm of a ratio of frequencies. Ann. Hum. Genet. *20: 309–311.*

Hamburger, V. and Hamilton, H. L. (1992). A series of normal stages in the development of the chick embryo. Developmental Dynamics. *195(4): 231-272.*

Hannon Lab (2009). Fastx-Toolkit: FASTQ/A short-reads pre-processing tools. *Available online at: http://www.hannonlab.cshl.edu/fastx_toolkit.*

Harford, C.G. and Hamlin, A. (1952). Effect of influenza virus on cilia and epithelial cells in the bronchi of mice. J. Exp. Med. *95(2): 173-189.*

Hayasaki, H., Sohma, Y., Kanbara, K., Maemura, K., Kubota, T. and Watanabe, M. A. (2006). A local GABAergic system within rat trigeminal ganglion cells. Eur. J. Neurosci*. 23(3): 745-57.*

Hayden, E. C. (2014). Is the $1,000 genome for real? Nature News (online resource). *Available at http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530. Accessed 29 June 2015.*

Heap, G. A., Yang, J. H., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., Franke, L., Dubois, P. C., Mein, C. A., Dobson, R. J., Albert, T. J., Rodesch, M. J., Clayton, D. G., Todd, J. A., van Heel, D. A., and Plagnol, V. (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum. Mol. Genet. *19(1):122-34.*

Hebenstreit, D. (2012). Methods, Challenges and Potentials of Single Cell RNA-seq. Biology. *1(3), 658–667.*

Hermans-Borgmeyer, I., Susens, U. and Borgmeyer, U. (2000). Developmental expression of the estrogen receptor-related receptor gamma in the nervous system during mouse embryogenesis. Mech. Dev. *97: 197-9.*

Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., Lush, M. J., Povey, S., Talbot, C. C., Wright, M. W., Wain, H. M., Trowsdale, J., Ziegler, A. and Beck, S. (2004). Gene map of the extended human MHC. Nat Rev Genet. *5(12):889–99.*

Howorka, S., Cheley, S. and Bayley, H. (2001). Sequence-specific detection of individual DNA strands using engineered nanopores. Nature Biotechnology. *19: 636-39.*

Hu, J. and Peter, I. (2013). Evidence of expression variation and allelic imbalance in Crohn's disease susceptibility genes NOD2 and ATG16L1 in human dendritic cells. Gene. *527(2):496-502.*

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

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. *4(1): 44-57.*

Huang, Q., Sivaramakrishna, R. P., Ludwig, K., Korte, T., Bottcher, C. and Herrmann, A. (2003). Early steps of the conformational change of influenza virus hemagglutinin to a fusion active state: stability and energetics of the hemagglutinin. Biochem. Biophys. Acta. *1614(1):3–13.*

Hughes, A. L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. *335(6186): 167–70.*

Illumina Inc. (2011). Quality scores for next generation sequencing. Illumina Technical Note. *Available at www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf. Accessed 30 December 2015.*

Illumina Inc. (2015). For All You Seq. *Available at http://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf. Webpage last updated 3 June 2015. Accessed 23 May 2016.*

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 324*(5924): 218-223.*

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431: 931-945.*

Isaacs, A. and Lindenmann, J. (1957). Virus interference. I. The interferon. Proc. R. Soc. Lond. B. Biol. Sci. *147(927): 258-67.*

Janetta, P. J. (1967). Arterial compression of the trigeminal nerve at the pons in patients with trigeminal neuralgia. J. Neurosurg. *26: 159-162.*

Janetta, P. J. (1980). Neurovascular compression in cranial nerve and systemic disease. Ann. Surg. *192(4): 518-525.*

Janeway, C. A. Jr., Travers, P., Walport, M. and Shlomchik, M. J. (2001). Immunobiology: The Immune System in Health and Disease. 5th edition. New York, Garland Science. *B-cell activation by armed helper T cells.*

Jiao, Y. and Meyerowitz, E. M. (2010). Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. Molecular Systems Biology. *6: 419*.

Johnson, L. R. and Byrne, J. H. (1998). Essential medical physiology*. Philadelphia: Lippincott-Raven. P890.*

Johnson, W. E., Rabinovic, A. and Li, C. (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics. *8(1): 118-127.*

Jones, E. Y., Fugger, L., Strominger, J. L. and Siebold, C. (2006). MHC class II proteins and disease: a structural perspective. Nature Reviews Immunology. *6: 271-282.*

Jorde, L. B., Hasstedt, S. J., Ritvo, E. R., Mason-Brothers, A., Freeman, B. J., Pingree, C., McMahon, W. M., Peterson, B., Jenson, W. R. and Mo, A. (1991). Complex segregation analysis of autism. Am. J. Hum. Genet. *49(5): 932-938.*

Junker, J. P., Noël, E. S., Guryev, V., Peterson, K. A., Shah, G., Huisken, J., McMahon, A. P., Berezikov, E., Bakkers, J. and van Oudenaarden, A. (2014). Genome-wide RNA Tomography in the zebrafish embryo. Cell. *159(3): 662-75.*

Jurkowska, R. Z., Jurkowski, T. P. And Jeltsch, A. (2011). Structure and function of mammalian DNA methyltransferases. Chembiochem *12(2): 206-222.*

Kamatani, Y., Wattanapokayakit, S., Ochi, H., Kawaguchi, T., Takahashi, A., Hosono, N., Kubo, M., Tsunoda, T., Kamatani, N., Kumada, H., Puseenam, A., Sura, T., Daigo, Y., Chayama, K., Chantratita, W., Nakamura, Y. and Matsuda, K. (2009). A genome-wide association study identifies variants in the L.A-DP locus associated with chronic hepatitis B in Asians. Nat. Genet*. 41: 591–595.*

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28: 27–30.*

Kasai, N., Fukushima, K., Ueki, Y., Prasad, S., Nosakowski, J., Sugata, K., Sugata, A., Nishizaki, K., Mever, N. C. and Smith, R. J. (2001). Genomic structures of SCN2A and SCN3A – candidate genes for deafness at the DFNA16 locus. Gene. *264: 113-22.*

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K. C., Hallinan, J., Mattick, J., Hume, D. A., Lipovich, L., Batalov, S., Engstrom, P. G., Mizuno, Y., Faghihi, M. A., Sandelin, A., Chalk, A. M., Mottagui-Tabar, S., Liang, Z., Lenhard, B. and Wahlestedt C. (2005). Antisense transcription in the mammalian transcriptome. Science. *309: 1564–1566.*

Katoh, M. (2004). Identification and characterization of PDZRN3 and PDZRN4 genes in silico. Int. J. Mol. Med. *13(4): 607-13.*

Katz, Y., Wang, E. T., Airoldi, E. M. and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. *7(12): 1009-1015.*

Kaur, K., Sullivan, M. and Wilson, P. C. (2011). Targeting B cell responses in universal influenza vaccine design. Trends in Immunology. *32(11): 524-531.*

Keren, H., Lev-Maor, G. and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews Genetics. *11: 345-355.*

Kim, E., Goren, A. and Ast, G. (2008). Alternative splicing: current perspectives. Bioessays. *30: 38–47.*

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14: R36.*

Klepikova, A. V., Logacheva, M. D., Dmitriev, S. E. and Penin, A. A. (2015). RNA-seq analysis of an apical meristem time series reveals a critical point in Arabidopsis thaliana flower initiation. BMC Genomics. *16: 466.*

Klockgether, T. (2011). Update on degenerative ataxias. Curr. Opin. Neurol. *24: 339-45*.

Kolle, G. (2007). Microarrays and alternative splicing. ASBMB *38(3).*

Korlach, J., Bjornson, K. P., Chaudhuri, B. P., Cicero, R. L., Flusberg, B. A., Gray, J. J., Holden, D., Saxena, R., Wegener, J. and Turner, S. W. (2010). Real-

time DNA sequencing from single polymerase molecules. Methods Enzymol. *472: 431-55.*

Kornberg, R. D. (2007). The molecular basis of eukaryotic transcription. PNAS *104(32): 12955-12961.*

Koziol, L. F., Budding, D., Andreasen, N., D'Arrigo, S., Bulgheroni, S., Imamizu, H., Ito, M., Manto, M., Marvel, C., Parker, K., Pezzulo, G., Ramnami, N., Riva, D., Schmahmann, J., Vandervert, L. and Yamazaki, T. (2014). Consensus paper: The cerebellum's role in movement and cognition. Cerebellum. *13: 151-77.*

Kuhn, H. and Frank-Kamenetskii, M.D. (2005). *Template-independent ligation of single-stranded DNA by T4 DNA ligase.* FEBS J. *272(23): 5991-6000.*

Kumar, L. and Futschik, M. E. (2007). Mfuzz: a software package for soft clustering of microarray data. Bioinformation. *2(1): 507.*

Kumar, R. A., Marshall, C. R., Badner, J. A., Babatz, T. D., Mukamel, Z., Aldinger, K. A., Sudi, J., Brune, C. W., Goh, G., Karamohamed, S., Sutcliffe, J. S., Cook, E. H., Geschwind, D. H., Dobyns, W. B., Scherer, S. W. and Christian, S. L. (2009). Association and Mutation Analyses of 16p11.2 Autism Candidate Genes. PLoS ONE. *4(2): e4582.*

Lahiry, P., Torkamani, A., Schork, N. J. and Hegele, R. A. (2010). Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat. Rev. Gen. *11: 60-74.*

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

Langfelder, P. and Horvath, S. (2014). WGCNA package FAQ. *Available from https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/faq.html [accessed 2/6/2016].*

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods. *9(4):357-359.*

Lapidot, M. and Pilpel, Y. (2006). Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO Reports. *7(12): 1216–1222.*

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol. Detect. Quantif. *3: 1-8.*

Leone, M., Russell, M. B., Rigamonti, A., Attanasio, A., Grazzi, L., D'Amico, D., Usai, S. and Bussone, G. (2001). Increased familial risk of cluster headache. Neurology. *56: 1233-1236*.

Leroux, E. and Ducros, A. (2008). Cluster headache. Orphanet. J. Rare Dis. *3: 20.*

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and The 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics*. 25(16): 2078-2079.*

Liu, L., Chang, G., Jiao, Y. and Simon, S. (1998). Neuronal nicotinic acetylcholine receptors in rat trigeminal ganglia. Brain Research. *809(2): 238-245.*

Liyou, N. E., Buller, K. M., Tresillian, M. J., Elvin, C. M., Scott, H. L., Dodd, P. R., Tannenberg, A. E. and McManus, M. E. (2003). Localization of a brain sulfotransferase, SULT4A1, in the human and rat brain: an immunohistochemical study. J. Histochem. Cytochem. *51(12): 1655-64.*

Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory*. 28(2): 129-137.*

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. and Darnell, J. (2000). The three roles of RNA in protein synthesis. Molecular Cell Biology. 4th edition. *New York: Freeman, W. H. Section 4.4*

Lorenzo, L., Magnussen, C., Bailey, A., St Louis, M., De Koninck, Y., Ribeiro-da-Silva, A. (2014). Spatial and temporal pattern of changes in the number of GAD65-immunoreactive inhibitory terminals in the rat superficial dorsal horn following peripheral nerve injury. Molecular Pain*. 10: 57–68.*

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15: 550.*

Lu, V. B., Ikeda, S. R. and Puhl, H. L. (2015). A 3.7 kb fragment of the mouse Scn10a gene promoter directs neural crest but not placodal lineage EGFP expression in a transgenic animal. J. Neurosci. *35: 8021-8034.*

Majewski, J. and Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends in Genetics. 27(2):72-79.

Mannion, A., Leader, G. and Healy, O. (2013). An investigation of comorbid psychological disorders,sleep probles, gastrointestinal symptoms and epilepsy in children and adolescents with Autism Spectrum Disorder. Research in Autism Spectrum Disorders*. 7(1): 35-42.*

Marchant, D. J., Bellac, C. L., Moraes, T. J., Wadsworth, S. J., Dufour, A., Butler, G. S., Bilawchuk, L. M., Hendry, R. G., Robertson, A. G., Cheung, C. T., Ng, J., Ang, L., Luo, Z., Heilbron, K., Norris, M. J., Duan, W., Bucyk, T., Karpov, A., Devel, L., Georgiadis, D., Hegele, R. G., Luo, H., Granville, D. J., Dive, V.,

McManus, B. M., Overall, C. M. (2014). A new transcriptional role for matrix metalloproteinase-12 in antiviral immunity. Nat. Med. *20(5): 493-502.*

Marill, M. C. (2015). After flu vaccine mismatch, calls for delayed selection intensify. Nature Medicine. *21: 297-298.*

Marthandan, S., Priebe, S., Baumgart, M., Groth, M., Cellerino, A., Guthke, R., Hemmerich, P. and Diekmann, S. (2015). Similarities in Gene Expression Profiles during In Vitro Aging of Primary Human Embryonic Lung and Foreskin Fibroblasts. Biomed Res Int*. 73198.*

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal*. 17:10-12.*

Maski, K. P., Jeste, S. S. and Spence, S. J. (2011). Common neurological co-morbidities in autism spectrum disorders. Curr. Opin. Pediatr. *23: 609–61.*

Matlin, A. J., Clark, F. and Smith C. W. J. (2005). Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Bio*. 6:386-398.*

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Steven Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R. and Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. *337(6099): 1190–1195.*

McLean, H. Q., Thompson, M. G., Sundaram, M. E., Kieke, B. A., Gaglani, M., Murthy, K., Piedra, P. A., Zimmerman, R. K., Nowalk, M. P., Raviotta, J. M., Jackson, M. L., Jackson, L., Ohmit, S. E., Petrie, J. G., Monto, A. S., Meece, J. K., Thaker, S. N., Clippard, J. R., Spencer, S. M., Fry, A. M. and Belongia, E. A. (2015). Influenza vaccine effectiveness in the United States during 2012-2013: variable protection by age and virus type. Journal of Infectious Diseases. *211(10): 1529-1540.*

Menashe, I., Grange, P., Larsen, E. C., Banerjee-Basu, S. and Mitra, P. P. (2013). Co-expression profiling of autism genes in the mouse brain. PLoS Comput. Biol. *9: e100312.*

Mi, S., Li, Y., Yan, J. and Gao, G. F. (2010). Na(+)/K(+)-ATPase β1 subunit interacts with M2 proteins of influenza A and B viruses and affects the virus replication. Sci. China Life Sci. *53(9): 1098-1105.*

Michel, A. M., Fox, G., Kiran, A., De Bo, C., O'Connor, P. B., Heaphy, S. M., Mullan, J. P., Donohue, C. A., Higgins, D. G. and Baranov, P. V. (2014). GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Research. *42: D859-64.*

Mills, J. D., Kawahara, Y. and Janitz, M. (2013). Strand-Specific RNA-Seq Provides Greater Resolution of Transcriptome Profiling. Current Genomics. *14(3): 173–181*.

Min, J. W., Kim, W. J., Han, J. A., Jung, Y. J., Kim, K. T., Park, W. Y., Lee, H. O. and Choi, S. S. (2015). Identification of Distinct Tumor Subpopulations in Lung Adenocarcinoma via Single-Cell RNA-seq. PLoS ONE. *10(8): e0135817.*

Miyamoto, D. T., Zheng, Y., Wittner, B. S., Lee, R. J., Zhu, H., Broderick, K. T., Desai, R., Fox, D. B., Brannigan, B. W., Trautwein, J., Arora, K. S., Desai, N., Dahl, D. M., Sequist, L. V., Smith, M. R., Kapur, R., Wu, C. L., Shioda, T., Ramaswamy, S., Ting, D. T., Toner, M., Maheswaran, S. and Haber D. A. (2015). RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science. *349(6254): 1351-1356.*

Mortimer, S. A., Trapnell, C., Aviran, S., Pachter, L. and Lucks, J. B. (2012). SHAPE-Seq: High-Throughput RNA Structure Analysis. Curr. Protoc. Chem. Biol*. 4(4): 275-97.*

Napoli, C., Lerman, L. O., Sica, V., Lerman, A., Tajana, G. and de Nigris, F. (2003). Microarray analysis: a novel research tool for cardiovascular scientists and physicians. Heart *89(6): 597-604.*

Nassenstein, C., Taylor-Clark, T. E., Myers, C. A., Ru, F., Nandigama, R., Bettner, W. and Undem, B. J. (2010). Phenotypic distinctions between neural crest and placodal derived vagal C-fibres in mouse lungs. The Journal of Physiology. *588: 4769-4783.*

Nemeroff, M. E., Barabino, S. M., Li, Y., Keller, W. and Krug, R. M. (1998). Influenza virus NS1 protein interacts with the cellular 30 kDa subunit of CPSF and inhibits 3' end formation of cellular pre-mRNAs. Mol. Cell. *1(7): 991-1000.*

Nieuwenhuys, R. Donkelaar, H, J. and Nicholson, C. (1998). The central nervous system of vertebrates. Springer-Verlag Berlin Heidelberg. *Chapter 11: p56.*

Noden, D. M. (1993). Spatial integration among cells forming the cranial peripheral nervous system. J. Neurobiol. *24(2): 248-261.*

Oberstaller, J., Joseph, S. J. and Kissinger, J. C. (2013). Genome-wide upstream motif analysis of Cryptosporidium parvum genes clustered by expression profile. BMC Genomics. *14: 516.*

Ohmit, S. E., Thompson, M. G., Petrie, J. G., Thaker, S. N., Jackson, M. L., Belongia, E. A., Zimmerman, R. K., Gaglani, M., Lamerato, L., Spencer, S. M., Jackson, L., Meece, J. K., Nowalk, M. P., Song, J., Cheng, P. Y., Rinaldo, C. R., Clipper, L., Shay, D. K., Piedra, P. and Monto, A. S. (2014). Influenza vaccine effectiveness in the 2011-2012 season: protection against each circulating

virus and the effect of prior vaccination on estimates. Clin. Infect. Dis. *58(3):319-327.*

Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., Oleś, A. K., Araúzo-Bravo, M. J., Saitou, M., Hadjantonakis, A. K., Hiiragi, T. (2014). Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. Nature Cell Biology. *16(1): 27–37.*

Okoniewski, M. J. and Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics *7: 276.*

Oschlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol Direct*. 4: 14.*

Pacific Biosciences (2014). Pacific Biosciences releases new DNA sequencing chemistry to enhance read length and accuracy for the study of human and other complex genomes. Pacific Biosciences Investor Release *(online resource). Available at http://investor.pacificbiosciences.com/ releasedetail.cfm?ReleaseID=876252. Accessed 29 December 2015.*

Palmen, S. J. M. C. (2004). Neuropathological findings in autism. Brain. *127: 2572–2583.*

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet *40: 1413–1415.*

Pan, C. Y., Tsai, C. L. and Chu, C. H. (2009). Fundamental movement skills in children diagnosed with autism spectrum disorders and attention deficit hyperactivity disorder. J. Autism. Dev. Disord*. 39(12): 1694-1705.*

Parham, P. and Ohta, T. (1996). Population biology of antigen presentation by MHC class I molecules. Science. *272(5258): 67–74.*

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., Horvath, S. and Geschwind, D. H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism*. Cell. 155(5): 1008-21.*

Patefield, W. M. (1981). Algorithm AS159: an efficient method of generating r x c tables with given row and column totals. Applied Statistics*. 30: 91–97.*

Patthey, C., Clifford, H., Haerty, W., Ponting, C. P., Shimeld, S. M. and Begbie, J. (2016). Identification of molecular signatures specific for distinct cranial sensory ganglia in the developing chick. Neural Dev. *11(1): 3.*

Pearl, P. L., Novotny, E. J., Acosta, M. T., Jakobs, C. and Gibson, K. M. (2003).

Succinic semialdehyde dehydrogenase deficiency in children and adults. Ann. Neurol. *54: S73-80.*

Peedicayil, J. (2005). DNA methylation and the central dogma of molecular biology. Med Hypotheses *64(6): 1243-1244.*

Pesquita, C., Faria, D., Falcão, A. O., Lord, P. and Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. PloS Comput Biol*. 5(7): e1000443.*

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. *464(7289): 768-772.*

Pinessi, L., Rainero, I., Rivoiro, C., Rubino, E. and Gallone, S. (2005). Genetics of cluster headache: an update. J. Headache Pain. *6: 234-236.*

Pinto, L. H. and Lamb, R. A. (2006). The M2 proton channels of influenza A and B viruses. J. Biol. Chem. *281(14): 8997–9000.*

Piras, V., Masaru, T. And Selvarajoo, K. Transcriptome-wide variability in single embryonic development cells. Scientific Reports *4: 7137.*

Pirvola, U., Zhang, X., Mantela, J., Ornitz, D. M. and Ylikoski, J. (2004). Fgf9 signaling regulates inner ear morphogenesis through epithelial-mesenchymal interactions. Dev. Biol. *273(2): 350-60.*

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A., O'McNamara, J. and Williams, S. M. (2001). Circuits within the cerebellum. Neuroscience 2nd edition. *Sunderland (MA): Sinauer Associates.*

Qiu, S., Luo, S., Evgrafov, O., Li, R., Schroth, G. P., Levitt, P., Knowles, J. A. and Wang, K. (2012). Single-neuron RNA-Seq: technical feasibility and reproducibility. Frontiers in Genetics. *3: 124.*

Ramsköld, D., Wang, E. T., Burge, C. B. and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PloS Comput. Biol. *5(12): e1000598.*

Rapin, I. and Tuchman, R. F. (2008). Autism: definition, neurobiology, screening, diagnosis. Pediatr. Clin. North Am*. 55(5): 1129-46.*

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. Science. *297(5586): 1551-5.*

Razin, A. and Cedar, H. (1991). DNA methylation and gene expression. Microbiol. Rev. *55(3): 451-458.*

Rebelo, S., Reguenga, C., Osorio, L., Pereira, C., Lopes, C. and Lima, D. (2007). DRG11 immunohistochemical expression during embryonic development in the mouse. Developmental Dynamics. *236: 2653-2660.*

Redcay, E. and Courchesne, E. (2008). Deviant functional megnetic resonance imaging patterns of brain activity to speech in 2-3-year-old children with autism spectrum disorder. Biological Psychiatry*. 64(7): 589-598.*

Reeber, S. L., Otis, T. S. and Sillitoe, R. V. (2013). New roles for the cerebellum in health and disease. Front. Syst. Neurosci. *7: 83.*

Reith, R. M., McKenna, J., Wu, H., Hashmi, S. S., Cho, S. H., Dash, P. K. and Gambello, M. J. (2013). Loss of Tsc2 in Purkinje cells is associated with autistic-like behavior in a mouse model of tuberous sclerosis complex. Neurobiol. Dis. *51: 93–103.*

Revolution Analytics & Weston, S. (2014). doParallel: Foreach parallel adaptor for the parallel package. R package version 1.0.8. *http://CRAN.R-project.org/package=doParallel.*

Richards, A. L., Holmans, P., O'Donovan, M. C., Owen, M. J. and Jones, L. (2008). A comparison of four clustering methods for brain expression microarray data. BMC Bioinformatics*. 9: 490.*

Risso, D., Schwartz, K., Sherlock, G. and Dudoit, S. (2011). GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics. *12: 480.*

Rizo, J. and Rosenmund, C. (2008). Synaptic vesicle fusion. Nat. Struct. Mol. Biol. *15(7): 665-74.*

Robert, C. and Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biology. *16:177.*

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics*. 26(1):139-140.*

Robinson, J., Halliwell, J. A., Hayhurst, J. H., Flicek, P., Parham, P. and Marsh, S. G. E (2015). The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. ***43****: D423-431*.

Robinson, J., Mistry, K., McWilliam, H., Lopez, R., Parham, P. and Marsh, S. G. E. (2011). The IMGT/HLA database. Nucleic Acids Res. *39(database issue): D1171-D1176.*

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E.,

Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., McKernan, K. J., Williams, A., Roth, G. T. and Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature. *475: 348-352.*

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. *20: 53-65.*

Rusk, N. (2011). Torrents of sequence. Nature Methods. *8: 44.*

Sakatani, M., Bonilla, L., Dobbs, K. B., Block, J., Ozawa, M., Shanker, S., Yao, J. and Hansen, P. J. (2013). Changes in the trancriptome of morula-stage bovine embryos caused by heat shock: relationship to developmental acquisition of thermotolerance. Reprod. Biol. Endocrinol. *11: 3.*

Samji, T. (2009). Influenza A: understanding the viral life cycle. Yale J. Biol. Med. *82(4): 153-159.*

Sanz, E., Yang, L., Su, T., Morris, D. R., McKnight, G. S. and Amieux, P. S. (2009). Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. PNAS. *106(33): 13939-13944.*

Sauter, N. K., Bednarski, M. D., Wurzburg, B. A., Hanson, J. E., Whitesides, G. M., Skehel, J. J. and Wiley, D. C. (1989). Hemagglutinins from two influenza virus variants bind to sialic acid derivatives with millimolar dissociation constants: a 500-MHz proton nuclear magnetic resonance study. Biochemistry. *28(21): 8388-96.*

Schena, M., Shalon, D., Davis, R. W. and Brown P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270(5235): 467-470.*

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature. *511: 421–427.*

Schlosser, G. (2006). Induction and specification of cranial placodes. Developmental Biology. *294(2): 303-351.*

Schroeder, R. D., Angelo, L. S. and Kurzrock, R. (2014). NF2/Merlin in hereditary neurofibromatosis 2 versus cancer: biologic mechanisms and clinical associations. Oncotarget. *5(1): 67-77.*

Schulte, E. C., Ellwanger, D. C., Dihanich, S., Manzoni, C., Stangl, K., Schormair, B., Graf, E., Eck, S., Mollenhauer, B., Haubenberger, D., Pirker, W., Zimprich, A., Brücke, T., Lichtner, P., Peters, A., Gieger, C., Trenkwalder, C., Mewes, H., Meitinger, T., Lewis, P. A., Klünemann, H. H. and Winkelmann, J.

(2014). Rare variants in *LRRK1* and Parkinson's disease. Neurogenetics. *15: 49–57.*

Schweet, R. and Heintz, R. (1966). Protein synthesis. Annu. Rev. Biochem*. 35: 723-758.*

SEQC/MAQC-III Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology. *32: 903-914.*

Shah, S. P., Kobel, M., Senz, J., Morin, R. D., Clarke, B. A., Wiegand, K. C., Leung, G., Zayed, A., Mehl, E., Kalloger, S. E., Sun, M., Giuliany, R., Yorida, E., Jones, S., Varhol, R., Swenerton, K. D., Miller, D., Clement, P. B., Crane, C., Madore, J., Provencher, D., Leung, P., DeFazio, A., Khattra, J., Turashvili, G., Zhao, Y., Zeng, T., Glover, J. N., Vanderhyden, B., Zhao, C., Parkinson, C. A., Jimenez-Linan, M., Bowtell, D. D. L., Mes-Masson, A., Brenton, J. D., Aparicio, S. A., Boyd, N., Hirst, M., Gilks, C. B., Marra, M. and Huntsman, D. G. (2009). Mutation of FOXL2 in granulosa-cell tumors of the ovary. N Engl J Med*. 360: 2719–2729.*

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transciptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. USA *100(26): 15776-15781*

Si, Y., Liu, P., Li, P. and Brutnell, T. P. (2014). Model-based clustering for RNA-seq data. Bioinformatics*. 30(2): 197-205.*

Silverman, J. L., Yang, M., Lord, C. and Crawley, J. N. (2010). Behavioural phenotyping assays for mouse models of autism. Nature Reviews Neuroscience. *11: 49—502.*

Simm, S., Fragkostefanakis, S., Paul, P., Keller, M., Einloft, J., Scharf, K. D. and Schleiff, E. (2015). Identification and Expression Analysis of Ribosome Biogenesis Factor Co-orthologs in Solanum lycopersicum. Bioinfom Biol Insights. *9: 1-17.*

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S. and Young, R. A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. Cell *106: 697–708.*

Skehel, J. J. and Wiley, D. C. (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu. Rev. Biochem. *69: 531–569.*

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98(3): 503-517.*

Stegle, O., Teichmann, S. A. and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. *16(3): 133-45.*

Steinberg, J. and Webber, C. (2013). The roles of FMRP-regulated genes in autism spectrum disorder: single- and multiple-hit etiologies. Am. J. Hum. Genet. *93(5): 825-39.*

Stern, C. D. (2004). Gastrulation: Frem Cells to Embryo. CSHL Press.

Stoner, R., Chow, M. L., Boyle, M. P., Sunkin, S. M., Mouton, P. R., Roy, S., Wynshaw-Boris, A., Colamarino, S. A., Lein, E. S. and Courchesne, E. (2014). Patches of disorganization in the neocortex of children with autism, N. Engl. J. Med. *370:1209-1219.*

Stoodley, C. J. (2014). Distinct regions of the cerebellum show gray matter decreases in autism, ADHD, and developmental dyslexia. Front. Syst. Neurosci. *8.*

Stros, M., Launholt, D. and Grasser, K. D. (2007). The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. Cell. Mol. Life Sci. *64: 2590–606.*

Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS ONE. *6(7): e21800.*

Temin, H. M. and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature *226(5252): 1211-1213.*

Thach, W. T. (1998). Combination, complementarity and automatic control: a role for the cerebellum in learning movement coordination. Novartis Found Symp. *218: 219-28.*

Thorndike, R. L. (1953). Who belongs in the family?. Psychometrika. *18(4): 267-276.*

Topa, H., Jonas, A., Kofler, R., Kosiol, C. and Honkela, A. (2015). Gaussian process test for high-throughput sequencing time series: application to experimental evolution. Bioinformatics. *31(11):1762-70.*

Tournier, C., Dong, C., Turner, T. K., Jones, S. N., Flavell, R. A. and Davis, R. J. (2001). MKK7 is an essential component of the JNK signal transduction pathway activated by proinflammatory cytokines. Genes Dev. *15(11): 1419-26.*

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and

switching among isoforms. Nat Biotechnol. *28(5): 511-515*.

Treanor, J. J., Talbot, H. K., Ohmit, S. E., Coleman, L. A., Thompson, M. G., Cheng, P. Y., Petrie, J. G., Lofthus, G., Meece, J. K., Williams, J. V., Berman, L., Breese Hall, C., Monto, A. S., Griffin, M. R., Belongia, E., Shay, D. K. and The U. S. Flu-VE Network (2012). Effectiveness of seasonal influenza vaccines in the United States during a season with circulation of all three vaccine strains. Clin. Infect. Dis. *55(7): 951-9.*

Tsai, C. J., Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M. M. and Nussinov, R. (2008). Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. Journal of Molecular Biology. *383(2): 281-291.*

Tsai, P. T., Hull, C., Chu, Y., Greene-Colozzi, E., Sadowski, A. R., Leech, J. M., Steinberg, J., Crawley, J. N., Regehr, W. G., and Sahin, M. (2012). Autistic-like behaviour and cerebellar dysfunction in Purkinje cell Tsc1 mutant mice. Nature. *488: 647–651.*

Twine, N. A., Janitz, K., Wilkins, M. R. and Janitz, M. (2011). Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. PloS One*. 6(1):e16266.*

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. and Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147, 1537–1550.*

UniProt Consortium. (2015). UniProt: a hub for protein information. Nucleic Acids Research. *43(Database issue): D204–D212.*

Van Lith, M., McEwen-Smith, R. M. and Benham, A. M. (2010). HLA-DP, HLA-DQ, and HLA-DR Have Different Requirements for Invariant Chain and HLA-DM. The Journal of Biological Chemistry. *285(52): 40800–40808.*

Vensko, N. W., Ferguson, S. M. (2010). Research Discoveries After Kubin. Tomorrow’s technology transfer *2(1): 46-58.*

Venter J. C., M. D. Adams, E. W. Myers, et al. (274 co-authors) (2001). The sequence of the human genome. Science *291:1304-1351.*

Velculescu, V. E., Zhang, L., Vogelstein, B. And Kinzler, K. W. (1995). Serial analysis of gene expression. Science *270(5235): 484-487.*

Verkerk, A. J., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhi, D. P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F. P., Eussen, B. E., van Ommen, G. J. B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A. and Warren, S. T. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell.

65(5): 905-14.

Wadman, M. (2008). James Watson's genome sequenced at high speed. Nature *452: 788.*

Wahl, M. C., Will, C. L., Luhrmann, R. (2009). The spliceosome: Design principles of a dynamic RNP machine. Cell *136: 701–718.*

Walker, H. K. (1990). Cranial Nerve V: The Trigeminal Nerve. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition, Boston: Butterworths. *Chapter 61.*

Walker, H. K. (1990). Cranial Nerve VII: The Facial Nerve and Taste. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition, Boston: Butterworths. *Chapter 62.*

Walker, H. K. (1990). Cranial Nerves IX and X: The Glossopharyngeal and Vagus Nerves. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition, Boston: Butterworths. *Chapter 63.*

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. and Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics. *23(10): 1274-1281.*

Wang, L., Wang, S. and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics. *28(16): 2184-5.*

Wang, S. S. H., Kloth, A. D. and Badura, A. (2014). The cerebellum, sensitive periods, and autism. Neuron. *83: 518–532.*

Wang, X., Li, M., Zheng, H., Muster, T., Palese, P., Beg, A. A. and Garcia-Sastre, A. (2000). Influenza A virus NS1 protein prevents activation of NF-kB and induction of alpha/beta interferon. J. Virol. *74(24): 11566-11573.*

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Gen. *10: 57-63.*

Waterhouse, L. (2013). Rethinking autism: variation and complexity. London: Academic Press. *p329.*

Watson, J. B., Coulter, P. M., II, Xia, Y. and Lusis, A. J. (1994). Mouse chromosomal localization of the cortexin (Ctxn) gene. Genomics. *22: 251-252.*

Weidner, J. M., Jiang, D., Pan, X. B., Chang, J., Block, T. M. and Guo, J. T. (2010). Interferon-induced cell membrane proteins, IFITM3 and tetherin, inhibit vesicular stomatitis virus infection via distinct mechanisms. Journal of Virology. *84(24): 12646-57.*

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm,

A., Flicek, P., Manolio, T., Hindorff, L. and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*. 42: D1001-D1006.

Whitaker, L. (1914). On the Poisson law of small numbers. Biometrika. *10:36–71.*

Wijnen, H., Naef, F., Boothroyd, C., Claridge-Chang, A., & Young, M. W. (2006). Control of Daily Transcript Oscillations in Drosophila by Light and the Circadian Clock. PLoS Genetics *2(3):e39.*

Williams, R. W. and Herrup, K. (1988). The control of neuron number. Annu. Rev. Neurosci. *11:423-53.*

Willsey, A. J. and State, M. W. (2015). Autism spectrum disorders: from genes to neurobiology. Curr. Opin. Neurobiol. *30: 92–99.*

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., Reilly, S. K., Lin, L., Fertuzinhos, S., Miller, J. A., Murtha, M. T., Bichsel, C., Niu, W., Cotney, J., Ercan-Sencicek, A. G., Gockley, J., Gupta, A. R., Han, W., He, X., Hoffman, E. J., Klei, L., Lei, J., Liu, W., Liu, L., Lu, C., Xu, X., Zhu, Y., Mane, S. M., Lein, E. S., Wei, L., Noonan, J. P., Roeder, K., Devlin, B., Sestan, N. and State, M. W. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell. *155: 997–1007.*

Wolpert, L., Tickle, C. and Arias, A. F. (2015). Vertebrate development III: Chick and mouse-completing the body plan. Principles of Development*. Oxford University Press: Chapter 5, 5.4.*

World Health Organization (2002). Outbreak of Influenza, Madagascar, July-August 2002. Weekly epidemiological record. *77: 381-388.*

World Health Organization (2009). Influenza (seasonal). *Available from http://www.who.int/mediacentre/factsheets/fs211/en/ [accessed 2/10/2015].*

Wu, J. Q., Wang, X., Beveridge, N. J., Tooney, P. A., Scott, R. J., Carr, V. J. and Cairns, M. J. (2012). Transcriptome Sequencing Revealed Significant Alteration of Cortical Promoter Usage and Splicing in Schizophrenia. PloS one *7(4):e36351.*

Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping alignment program for mRNA and EST sequences. Bioinformatics *21(5):1859-1875.*

Wysoker, A., Tibbetts, K. and Fennell, T. (2013). Picard Tools – By Broad Institute. *Available online at: http://www.broadinstitute.github.io/picard.*

Yakushin, S. B., Martinelli, G. P., Raphan, T., Xiang, Y., Holstein, G. R. and Cohen, B. (2014). Vasovagal oscillations and vasovagal responses produced by

the vestibulo-sympathetic reflex in the rat. Front. Neurol. *5: 37.*

Yamada, K. and Watanabe, M. (2002). Cytodifferentiation of Bergmann glia and its relationship with Purkinje cells. Anat. Sci. Int. *77(2): 94-108.*

Young, J. A. and Trowsdale, J. (1985). A processed pseudogene in an intron of the HLA-DP beta 1 chain gene is a member of the ribosomal protein L32 gene family. Nucleic Acids Res. *13(24): 8883-91.*

Young, M. D., Wakefield, M. J., Smyth, G. K. and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Gen. Biol. *11:R14*

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y. and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *26(7): 976-8.*

Zandian, A., Osiro, S., Hudson, R., Ali, I. M., Matusz, P., Tubbs, S. R. and Loukas, M. (2014). The neurologist's dilemma: A comprehensive clinical review of Bell's palsy, with emphasis on current management trends. Med. Sci. Monit. *20: 83-90.*

Zhang, Y. H., Zhao, Y., Li, N., Peng, Y. C., Giannoulatou, E., Jin, R. H., Yan, H. P., Wu, H., Liu, J. H., Liu, N., Wang, D. Y., Shu, Y. L., Ho, L. P., Kellam, P., McMichael, A. and Dong, T. (2013). Interferon-induced transmembrane protein-3 genetic variant rs12252-C is associated with severe influenza in Chinese individuals. Nature Communications. *4: 1418.*

Zhou, X. and Su, Z. (2007). EasyGO: Gene Ontology annotation and functional enrichment analysis tool for agronomical species. BMC Genomics*. 8: 246.*

Appendix A – Supplemental Materials for Chapter 4

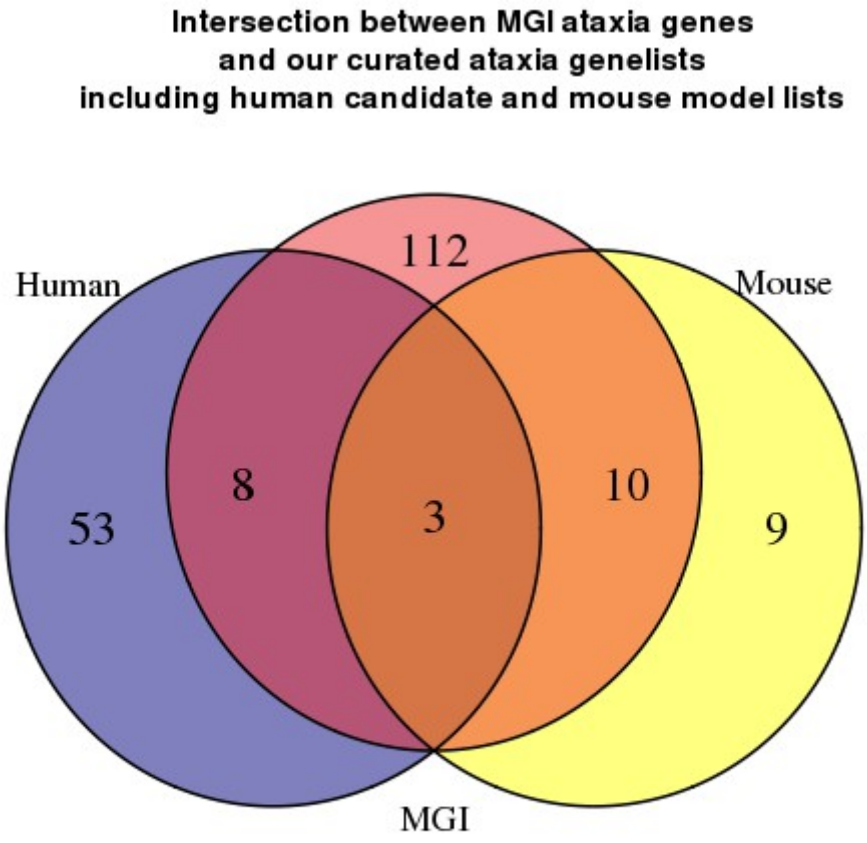
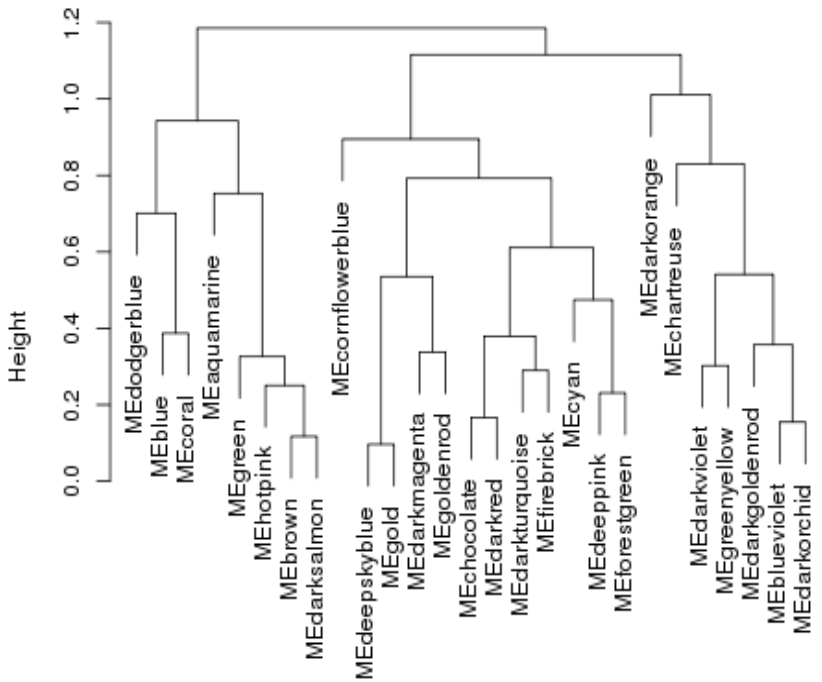
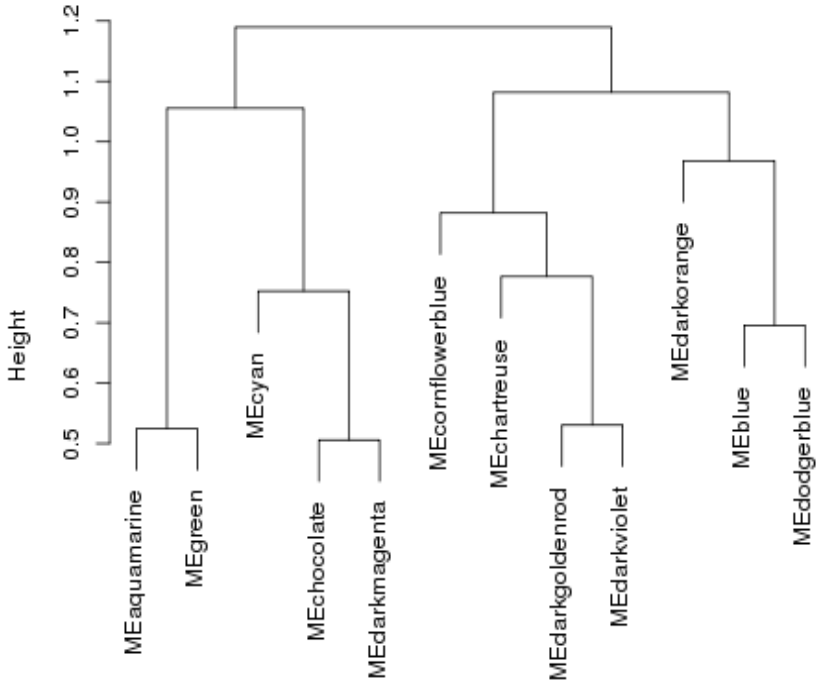


Figure A1. A Venn diagram displaying the overlap between the curated ataxia gene lists used for this study (including human candidate and mouse model gene lists), and the ataxia MGI phenotype genes.

A


Clustering of module eigengenes
Height
0.0
0.2
0.4
0.6
0.8
1.0
1.2
MEdodgerblue
MEblue
MEcoral
MEaquamarine
MEgreen
MEhotpink
MEbrown
MEdarksalmon
MEcornflowerblue
MEdeepskyblue
MEgold
MEdarkmagenta
MEgoldenrod
MEchocolate
MEdarkred
MEdarkturquoise
MEfirebrick
MEcyan
MEdeeppink
MEforestgreen
MEdarkorange
MEchartreuse
MEdarkviolet
MEgreenyellow
MEdarkgoldenrod
MEblueviolet
MEdarkorchid

B


Merged clustering of module eigengenes
Height
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
MEaquamarine
MEgreen
MEcyan
MEchocolate
MEdarkmagenta
MEcornflowerblue
MEchartreuse
MEdarkgoldenrod
MEdarkviolet
MEdarkorange
MEblue
MEdodgerblue

C

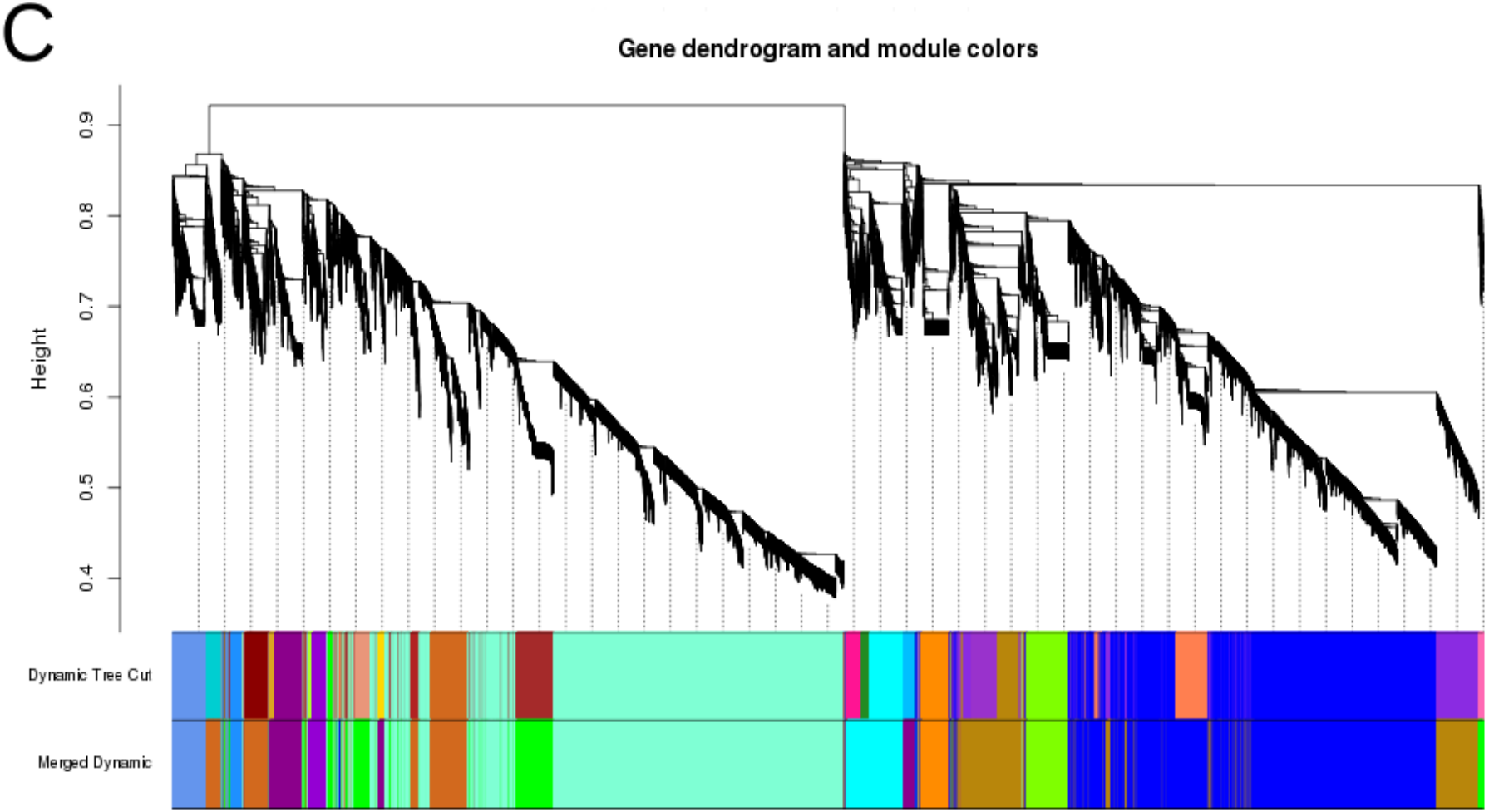
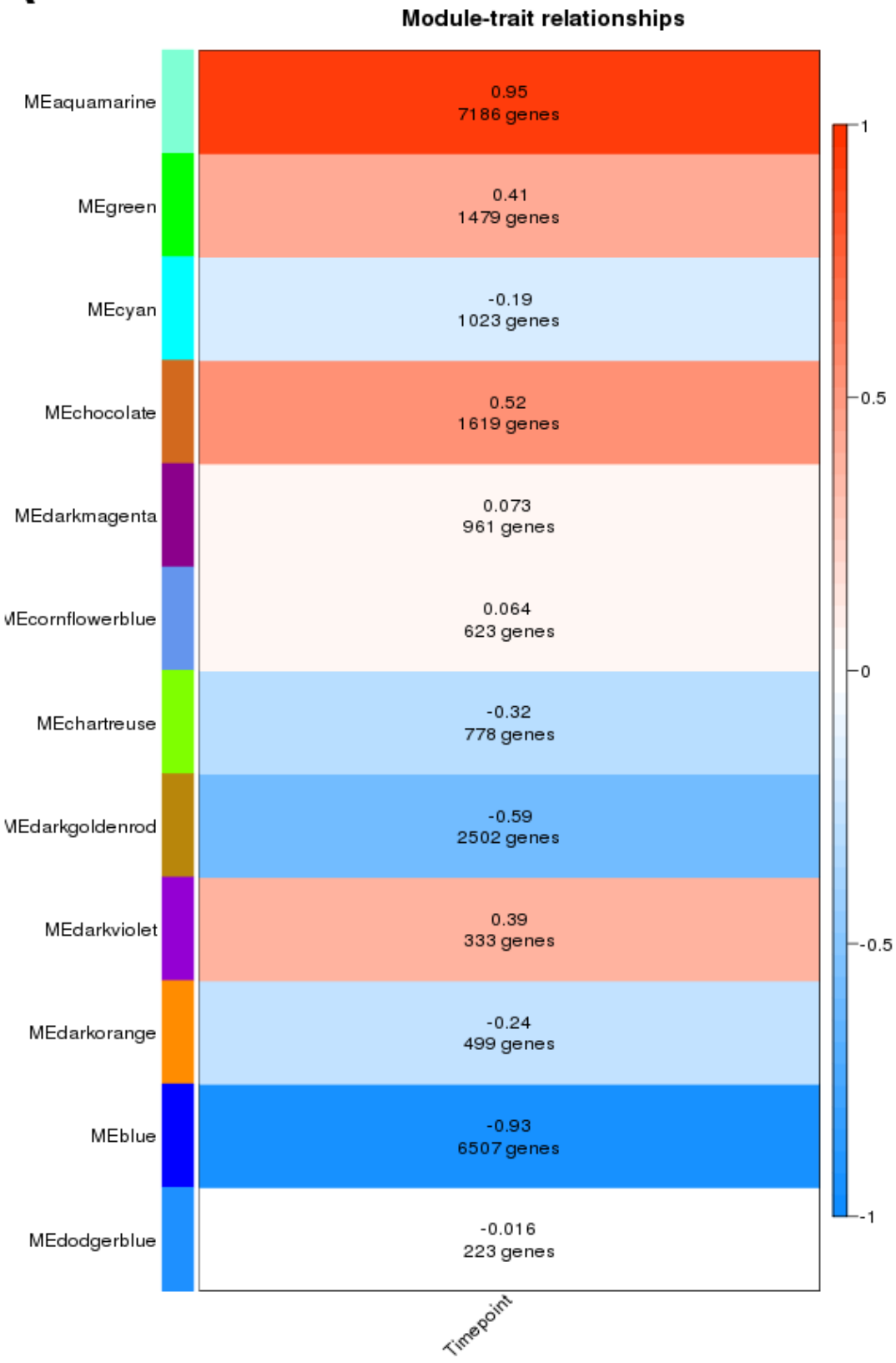

Gene dendrogram and module colors
Height
0.9
0.8
0.7
0.6
0.5
0.4
Dynamic Tree Cut
Merged Dynamic

Figure A2. Dendrograms displaying the clusters formed from WGCNA of the neocortex expression data. Hierarchical clustering of the eigengenes of the original 27 clusters can be seen in (A), followed by the same clustering of eigengenes after merging of similar clusters, leaving 12 clusters, in (B). (C) displays hierarchical clustering of all of the genes in the data-set (post-QC), with assignment to the clusters (as seen in (A) and (B)) displayed in the coloured bars below. The top coloured bar is representative of the 27 clusters (as shown in (A)) following dynamic tree cutting, whereas the lower coloured bar shows this same information following merging of similar clusters (as shown in (B)). All clustering was performed with euclidean distance measure and average linkage method.

A


Module-trait relationships
MEaquamarine
0.95
7186 genes
MEgreen
0.41
1479 genes
MEcyan
-0.19
1023 genes
MEchocolate
0.52
1619 genes
MEdarkmagenta
0.073
961 genes
MEcornflowerblue
0.064
623 genes
MEchartreuse
-0.32
778 genes
MEdarkgoldenrod
-0.59
2502 genes
MEdarkviolet
0.39
333 genes
MEdarkorange
-0.24
499 genes
MEblue
-0.93
6507 genes
MEdodgerblue
-0.016
223 genes
Timepoint
1
0.5
0
-0.5
-1

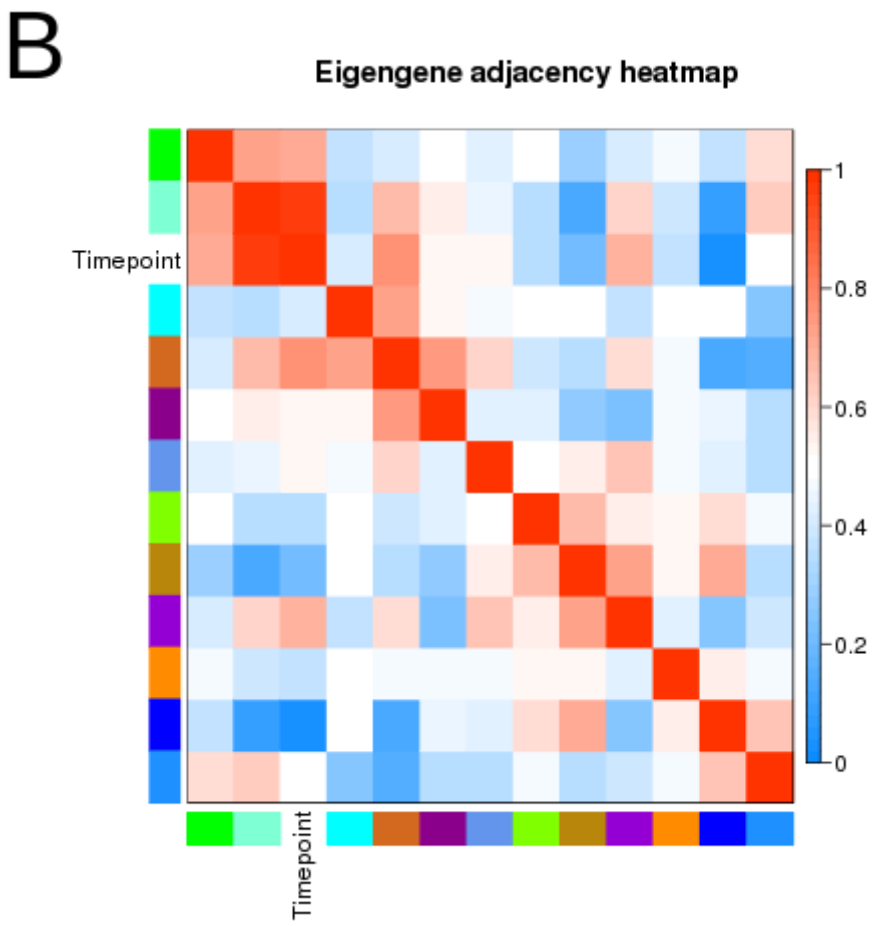
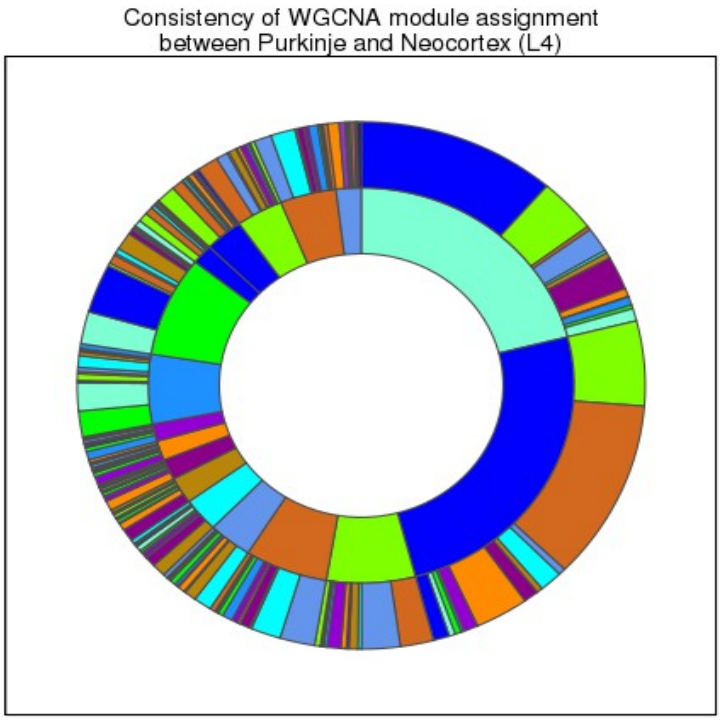


Figure A3. Heat plots displaying correlations between the eigengenes of the neocortex clusters (as defined by WGCNA) and developmental timepoint. (A) displays the correlation (r) between each of the clusters and timepoint, coloured by correlation and including information on the number of genes in each cluster. (B) displays similar correlation information, but in a full heatmap comparing each cluster with one another. Timepoint has also been included in this heatmap.

A


Consistency of WGCNA module assignment
between Purkinje and Neocortex (L4)

B

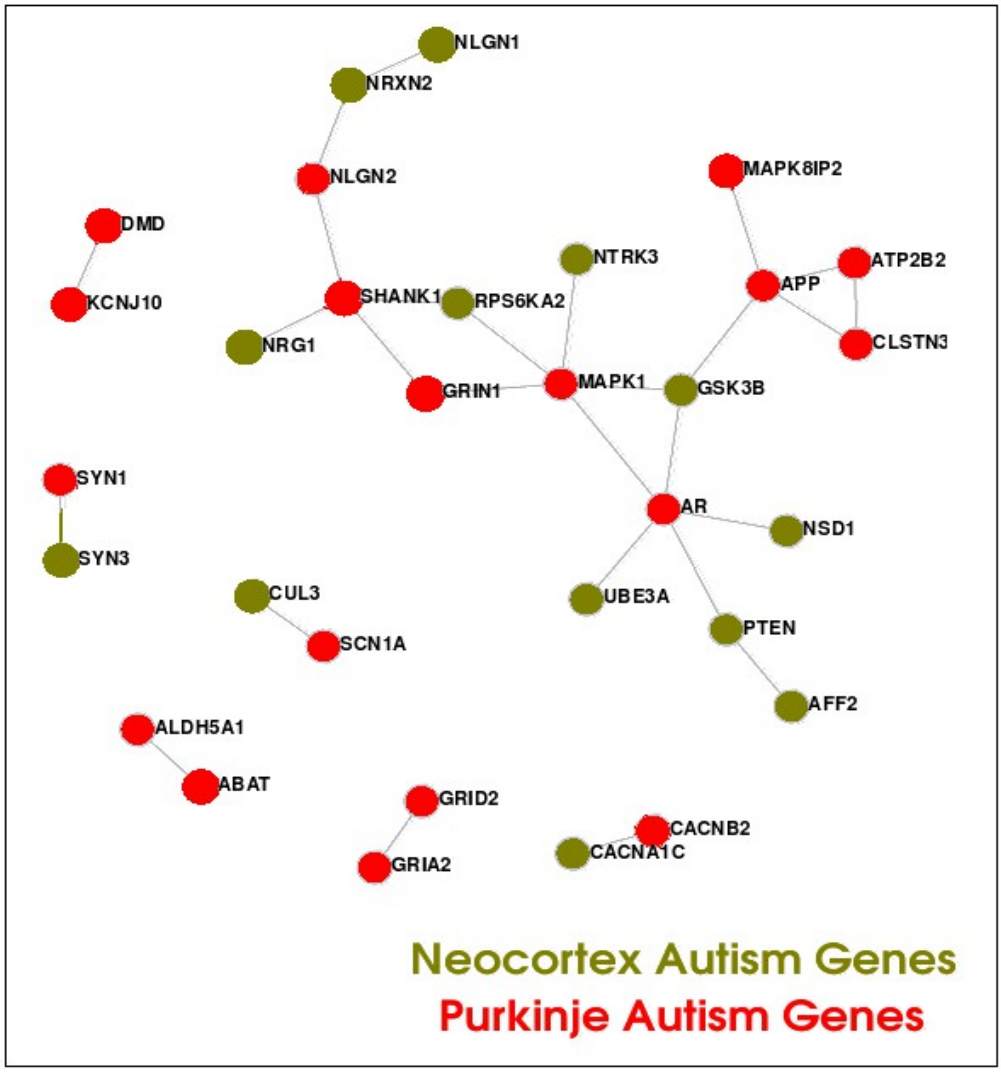

NLGN1
NRXN2
NLGN2
MAPK8IP2
DMD
NTRK3
ATP2B2
KCNJ10
SHANK1
RPS6KA2
APP
NRG1
CLSTN3
GRIN1
MAPK1
GSK3B
SYN1
AR
NSD1
SYN3
CUL3
UBE3A
SCN1A
PTEN
AFF2
ALDH5A1
ABAT
GRID2
CACNB2
CACNA1C
GRIA2
Neocortex Autism Genes
Purkinje Autism Genes

Figure A4. Plots investigating consistencies between the Purkinje cell and neocortex clusters. (A) displays a ring plot of all genes and their cluster assignment. The inner ring displays the cluster assignment for the Purkinje cell analysis genes; the outer ring then displays the cluster assignment for the same genes (in the same order) and their assignment in the neocortex analysis. (B) displays a dapple output for the autism genes identified in the respective ASD-enriched clusters from the Purkinje cell (differentially expressed clustering) and neocortex (WGCNA) analyses, showing no particular known separation between genes from each cluster. Genes have been coloured by tissue.

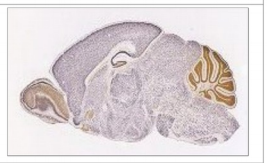
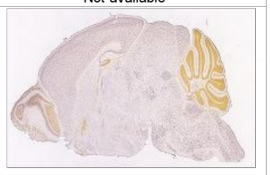
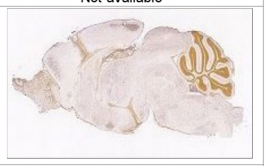
Table A1. A list of the genes used for the curated ataxia human candidate and mouse model lists, including Ensembl gene ID and common gene name, with information on which of the two lists each gene is included in (a 1 indicated inclusion in the list).

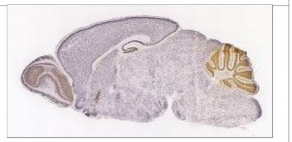
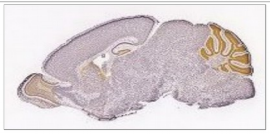
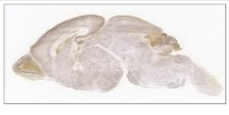
Ensembl Gene ID	Gene Name	Human Candidate	Mouse Model
ENSMUSG00000017412	Cacnb4	1	1
ENSMUSG00000024357	Sil1	1	1
ENSMUSG00000034656	Cacna1a	1	1
ENSMUSG00000000838	Fmr1	1	0
ENSMUSG00000001627	Ifrd1	1	0
ENSMUSG00000003617	Cp	1	0
ENSMUSG00000004263	Atn1	1	0
ENSMUSG00000005360	Slc1a3	1	0
ENSMUSG00000005621	Zfp592	1	0
ENSMUSG00000014767	Tbp	1	0
ENSMUSG00000014782	Plekhg4	1	0
ENSMUSG00000016200	Syt14	1	0
ENSMUSG00000016541	Atxn10	1	0
ENSMUSG00000019971	Cep290	1	0
ENSMUSG00000019986	Ahi1	1	0
ENSMUSG00000020003	Pex7	1	0
ENSMUSG00000021177	Tdp1	1	0
ENSMUSG00000021738	Atxn7	1	0
ENSMUSG00000022620	Arsa	1	0
ENSMUSG00000023033	Scn8a	1	0
ENSMUSG00000024413	Npc1	1	0
ENSMUSG00000024500	Ppp2r2b	1	0
ENSMUSG00000024527	Afg3l2	1	0
ENSMUSG00000024924	Vldlr	1	0
ENSMUSG00000025232	Hexa	1	0
ENSMUSG00000025551	Fgf14	1	0
ENSMUSG00000026170	Cyp27a1	1	0
ENSMUSG00000026489	Adck3	1	0
ENSMUSG00000026664	Phyh	1	0
ENSMUSG00000027405	Nop56	1	0
ENSMUSG00000028411	Aptx	1	0
ENSMUSG00000030102	Itpr1	1	0
ENSMUSG00000031333	Abcb7	1	0
ENSMUSG00000031872	Bean1	1	0
ENSMUSG00000031928	Mre11a	1	0
ENSMUSG00000032046	Abhd12	1	0
ENSMUSG00000034218	Atm	1	0
ENSMUSG00000034958	Atcay	1	0
ENSMUSG00000034994	Eef2	1	0
ENSMUSG00000035735	Dagla	1	0
ENSMUSG00000037949	Ano10	1	0
ENSMUSG00000039176	Polg	1	0
ENSMUSG00000040896	Kcnd3	1	0
ENSMUSG00000041261	Car8	1	0
ENSMUSG00000042605	Atxn2	1	0
ENSMUSG00000043535	Setx	1	0
ENSMUSG00000045594	Glb1	1	0
ENSMUSG00000046230	Vps13a	1	0
ENSMUSG00000046876	Atxn1	1	0
ENSMUSG00000047976	Kcna1	1	0
ENSMUSG00000048279	Sacs	1	0
ENSMUSG00000059363	Fxn	1	0
ENSMUSG00000062785	Kcnc3	1	0
ENSMUSG00000067889	Sptbn2	1	0
ENSMUSG00000073988	Ttpa	1	0
ENSMUSG00000078816	Prkcg	1	0
ENSMUSG00000090100	Ttbk2	1	0
ENSMUSG00000004207	Psap	0	1
ENSMUSG00000010066	Cacna2d2	0	1
ENSMUSG00000019146	Cacng2	0	1
ENSMUSG00000021557	Agtpbp1	0	1
ENSMUSG00000024397	Aif1	0	1
ENSMUSG00000026048	Ercc5	0	1
ENSMUSG00000026131	Dst	0	1
ENSMUSG00000027716	Trpc3	0	1
ENSMUSG00000028222	Calb1	0	1
ENSMUSG00000029313	Aff1	0	1
ENSMUSG00000031960	Aars	0	1
ENSMUSG00000032238	Rora	0	1
ENSMUSG00000033615	Cplx1	0	1
ENSMUSG00000042453	Reln	0	1
ENSMUSG00000043301	Kcnj6	0	1
ENSMUSG00000071424	Grid2	0	1
ENSMUSG00000079037	Prnp	0	1
ENSMUSG00000087408	Cers1	0	1

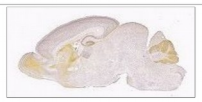
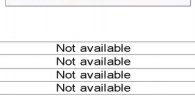
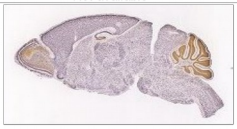
Table A2. Reference genome alignment statistics, including the percentage of reads aligned, and of those aligned, the percentage aligned to a single unique location.

	Aligned (%)	Uniquely Aligned (%)
P0 Rep1	98.96	72.95
P0 Rep2	98.96	73.88
P0 Rep3	99.07	72.17
P4 Rep1	99.00	75.92
P4 Rep2	99.04	76.91
P4 Rep3	98.96	72.95
P8 Rep1	98.96	73.88
P8 Rep2	99.07	72.17
P8 Rep3	99.00	75.92
P14 Rep1	99.04	76.91
P14 Rep2	98.96	72.95
P14 Rep3	98.96	73.88
P21 Rep1	99.07	72.17
P21 Rep2	99.00	75.92
P21 Rep3	99.04	76.91

Table A3. The autism-associated genes identified in the two tissue clusters, and (where available) in-situ hybridization images from stages P4 and P14 of mouse development from the Allen Brain Atlas.

Ensembl Gene ID	Gene Name	Occurrence in Gene List		Occurrence in Cluster		Allen Brain Atlas ISH	
		Human Autism	Mouse Autism	Cerebellum Purkinje	Neocortex L4	P4	P14
ENSMUSG00000025037	Maoa	1	1	1	1		
ENSMUSG00000037217	Syn1	1	1	1	0		
ENSMUSG00000027030	Stk39	1	1	1	0	Not available	Not available
ENSMUSG00000026904	Slc4a10	1	1	1	0	Not available	Not available
ENSMUSG00000037541	Shank2	1	1	1	0	Not available	Not available
ENSMUSG00000038738	Shank1	1	1	1	0	Not available	Not available
ENSMUSG00000064329	Scn1a	1	1	1	0		
ENSMUSG00000008658	Rbfox1	1	1	1	0	Not available	Not available
ENSMUSG00000020893	Per1	1	1	1	0		
ENSMUSG00000063142	Kcnma1	1	1	1	0	Not available	Not available
ENSMUSG00000049511	Htr1b	1	1	1	0		
ENSMUSG00000071424	Grid2	1	1	1	0		
ENSMUSG00000045103	Dmd	1	1	1	0	Not available	Not available
ENSMUSG00000030302	Atp2b2	1	1	1	0	Not available	Not available
ENSMUSG00000032826	Ank2	1	1	1	0	Not available	Not available
ENSMUSG00000035936	Aldh5a1	1	1	1	0	Not available	Not available
ENSMUSG00000027630	Tbl1xr1	1	0	1	0	Not available	Not available
ENSMUSG00000041540	Sox5	1	0	1	0		
ENSMUSG00000019558	Slc6a8	1	0	1	0	Not available	Not available
ENSMUSG00000030683	Sez6l2	1	0	1	0	Not available	Not available
ENSMUSG00000061718	Ppp1r1b	1	0	1	0		
ENSMUSG00000023089	Ndufa5	1	0	1	0	Not available	Not available
ENSMUSG00000029009	Mthfr	1	0	1	0	Not available	Not available
ENSMUSG00000068205	Macrod2	1	0	1	0	Not available	Not available
ENSMUSG00000044708	Kcnj10	1	0	1	0	Not available	Not available

1	0	1	0		
1	0	1	0		
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0	Not available	Not available
1	0	1	0		
1	0	1	0		
1	0	1	0	Not available	Not available
1	0	1	0		
1	0	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0		
0	1	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0		
0	1	1	0		
0	1	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0	Not available	Not available
0	1	1	0		
0	1	1	0	Not available	Not available
0	1	1	0		
1	1	0	1		
1	1	0	1	Not available	Not available

1	1	0	1		
1	1	0	1	Not available	Not available
1	1	0	1	Not available	Not available
1	1	0	1	Not available	Not available
1	1	0	1		
1	1	0	1	Not available	Not available
1	1	0	1		
1	1	0	1	Not available	Not available
1	1	0	1	Not available	Not available
1	1	0	1		
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1		
1	0	0	1		
1	0	0	1		
1	0	0	1		
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1		
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1	Not available	Not available
1	0	0	1		
1	0	0	1	Not available	Not available
0	1	0	1	Not available	Not available
0	1	0	1	Not available	Not available
0	1	0	1		
0	1	0	1		
0	1	0	1	Not available	Not available
0	1	0	1		
0	1	0	1		
0	1	0	1		

Appendix B – Supplemental Materials for Chapter 5

B1 - Main Function Help Pages

R Documentation

cluster.wrapper {ROCKS}

Parallel Negative-Binomial Model-Based Clustering

Description

A wrapper function for all main MBCluster.Seq functions (Si, 2012) to perform model-based clustering on an RNA-Seq data count table, with additional functionality to extract log-likelihood values and calculate Akaike's Information Criterion (AIC). This wrapper also allows this to be executed for multiple numbers of clusters in parallel, both through the forming of socket clusters and through forking.

Usage

cluster.wrapper(counts, treatment, nK.vector, additional.forks=1)

Arguments

counts

a GxP matrix storing the numbers of reads mapped to G genes in P samples. Row names should contain corresponding gene IDs for

downstream ontology analysis.

treatment

a vector of length P specifying the corresponding treatments for each column of the counts matrix. See 'Details' for an example.

nK.vector

a numeric vector specifying the varying numbers of clusters (K values) for which the cluster.wrapper function is to be applied. See 'Details' for an example.

additional.forks

the number of additional forks to be formed from any applications (including those on socket clusters) of the model-based clustering algorithm. Note: forking is not supported by Windows. See 'Details' for an example.

Details

treatment is a variable that allows the function to know which samples are replicates of one another. For example, treatment=c(1,1,1,2,2,2) indicates the first 3 columns of the counts matrix are 3 replicates of the same treatment, with the remaining 3 columns forming the 3 replicates of the second treatment.

nK.vector indicates how many times the model-based clustering algorithm should be run, and with what value of K (the final number of clusters to be formed) each time. For example, nK.vector=c(2,3,4) will mean the cluster.wrapper function applies the algorithm 3 times, forming 2 clusters, 3 clusters, and 4 clusters respectively. If this function is used in conjunction with a socket cluster (created using doParallel's makeCluster and registerDoParallel functions; Revolution Analytics & Weston, 2014), these algorithms can be run in parallel.

additional.forks allows any runs of the algorithm (parallel or not) to themselves be run

in parallel forks. For example, if a cluster of 3 parallel sockets has been created and additional.forks=2 specified, this wrapper will run up to 3 model-based clustering algorithms at once (if there are 3 or more values in nK.vector) and then allow each of these 3 algorithm runs to be split over 2 forks. This would therefore create a grand total of 6 parallel processes. This is useful if the number of cores available is much greater than the number of algorithms to be performed. Note: Forking is not supported by Windows, in which case the value for additional.forks should be left as 1.

Value

cluster.wrapper returns a list (with equal length to nK.vector) of lists. Each list corresponds to a running of the model-based clustering algorithm, with a corresponding K value from the specified nK.vector, and contains the following components:

probability

a matrix containing the probability of each gene belonging to each cluster. Each row corresponds to each gene (as provided in the original counts matrix) and each column corresponds to each cluster (as assigned by the original nK.vector).

centers

a matrix containing estimates of the cluster centres. Each row corresponds to each cluster (as assigned by the original nK.vector) and each column corresponds to each treatment (as provided by the original treatment vector).

cluster

a vector indicating which cluster each gene has been assigned to. Each value corresponds to each gene (as provided in the original counts matrix).

lglk

the log-likelihood of the final model selected by the model-based clustering algorithm.

nparams

the number of parameters in the final model selected by the model-based clustering algorithm.

aic

the Akaike Information Criterion (AIC) of the final model selected by the model-based clustering algorithm.

References

Revolution Analytics & Weston, S. (2014). doParallel: Foreach parallel adaptor for the parallel package. R package version 1.0.8. http://CRAN.R-project.org/package=doParallel.

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Examples

```
# open a 3 socket cluster for running of the K values in parallel
cl <- makeCluster(3)
registerDoParallel(cl)
# load example data
load(url("http://bowtie-
bio.sourceforge.net/recount/ExpressionSets/trapnell_eset.RData"))
expr <- exprs(trapnell.eset)
counts <- expr[,order(trapnell.eset@phenoData@data$time.point.hrs)]
treatment <- 1:4
```

```
# perform model-based clustering 3 times with K values of 2, 3 and 4 respectively
# the 3 socket cluster opened above allows these to be run completely in parallel
# 2 forks per socket allows each model-based clustering run to be run over 2 cores
# the total cores used in this example is 6
output <- cluster.wrapper(counts, treatment, nK.vector=2:4, additional.forks=2)
# close the 3 socket cluster
stopCluster(cl)
# view the structure of the output
str(output)
```

AIC.plot.prep {ROCKS}

Prepare AIC Values for Plotting

Description

A function for extracting Akaike's Information Criterion (AIC) values from an output of the cluster.wrapper function and formatting these to be ready for plotting.

Usage

```
AIC.plot.prep(clustering.output)
```

Arguments

clustering.output

a list of lists as output by the cluster.wrapper function. Each list will contain the variables 'cluster', a vector indicating which cluster each gene is assigned to, and 'aic', an AIC value.

Details

AIC is a statistical measure that can be used to determine the quality of a statistical model compared to other models, based on their log-likelihood values (the goodness of fit of the models) penalized by the number of parameters each model has used (the complexity of the models) (Akaike, 1974). As the complexity of the models increases it has an increasing goodness of fit, however at high levels of complexity the increase in goodness of fit becomes more negligible. This can be seen in the curved plot output by this function. What is usually considered the best model is one with the lowest cluster number (K value) and lowest AIC, and so the user should look for an 'elbow point' in the plot.

Value

AIC.plot.prep returns a matrix where each row represents data from each running of the model-based clustering algorithm by the cluster.wrapper function. The first column states the number of clusters produced in that run, and the second column contains the resultant AIC value. This resultant matrix is suitable for passing directly through a simple plotting function (see 'Examples').

References

Akaike, H. (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* **19**(6): 716–723.

Examples

```
# load example data
library(“Biobase”)
load(url("http://bowtie-
bio.sourceforge.net/recount/ExpressionSets/trapnell_eset.RData"))
expr <- exprs(trapnell.eset)
counts <- expr[,order(trapnell.eset@phenoData@data$time.point.hrs)]
treatment <- 1:4
# perform model-based clustering
output <- cluster.wrapper(counts, treatment, nK.vector=2:10)
# extract AIC information
AIC.matrix <- AIC.plot.prep(output)
# plot AIC information and look for the 'elbow point' for optimal cluster number
plot(AIC.matrix)
```

gene.lengths {ROCKS}

Obtain Gene Length Information for All Clustered Genes

Description

A function for obtaining gene length information for all genes in an output of the cluster.wrapper function.

Usage

```
gene.lengths(input, genome, gene.id)
```

Arguments

input

a list of lists as output by the cluster.wrapper function. Each list will contain a variable 'cluster', a vector indicating which cluster each gene is assigned to.

genome

a string identifying the genome that the data was originally aligned to. For a list of supported genomes, run the supportedGenomes function from the goseq geneLenDataBase package (see 'Examples').

gene.id

a string identifying the gene identifier used in the row names of the original count table run through cluster.wrapper. For a list of supported gene IDs, run the supportedGeneIDs function from the goseq geneLenDataBase package (see 'Examples').

Details

This function extracts gene information from the cluster.wrapper output and applies it to the getlength function of the goseq package (Young *et al.* 2010). This information can then be used with the cluster.enrichment or single.enrichment functions to apply gene ontology enrichment analyses with p-values corrected for gene length bias.

genome and gene.id are also both applied to goseq's getlength function, which will search the built-in database of gene lengths for suitable data to download, or if there is no suitable data, will download this from an online resource.

Value

gene.lengths returns a numeric vector of gene lengths, with names matching the gene ids in the input data.

References

Young, M.D., Wakefield, M.J., Smyth, G.K. and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2): R14.

Examples

```
# perform model-based clustering
output <- cluster.wrapper(counts, treatment, nK.vector=2:4, additional.forks=2)
# view supported genomes and gene ids
library('goseq')
supportedGenomes()
supportedGeneIDs()
# obtain gene length information with supported genomes/geneids
bias.data <- gene.lengths(output, 'mm9', 'ensGene')
```

GO.cats {ROCKS}

Obtain Gene Ontology Information for All Clustered Genes

Description

A function for obtaining gene ontology information for all genes in an output of the cluster.wrapper function.

Usage

GO.cats(input, genome, gene.id)

Arguments

input

a list of lists as output by the cluster.wrapper function. Each list will contain a variable 'cluster', a vector indicating which cluster each gene is assigned to.

genome

a string identifying the genome that the data was originally aligned to. For a list of supported genomes, run the supportedGenomes function from the goseq geneLenDataBase package (see 'Examples').

gene.id

a string identifying the gene identifier used in the row names of the original count table run through cluster.wrapper. For a list of supported gene IDs, run the supportedGeneIDs function from the goseq geneLenDataBase package (see 'Examples').

Details

This function extracts gene information from the cluster.wrapper output and applies it to the getgo function of the goseq package (Young *et al.* 2010). This information can then be used with the cluster.enrichment or single.enrichment functions to apply gene ontology enrichment analyses.

genome and gene.id are also both applied to goseq's getgo function, which will search the built-in database of gene ontology information for suitable data to download, or if there is no suitable data, will download this from an online resource.

Value

GO.cats returns a list of gene ontology codes, with each variable corresponding to a gene in the input data.

References

Young, M.D., Wakefield, M.J., Smyth, G.K. and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2): R14.

Examples

```
# perform model-based clustering
output <- cluster.wrapper(counts, treatment, nK.vector=2:4, additional.forks=2)
# view supported genomes and gene ids
library('goseq')
supportedGenomes()
supportedGeneIDs()
# obtain gene ontology information with supported genomes/geneids
gene2cat <- GO.cats(output, 'mm9', 'ensGene')
```

KEGG.cats {ROCKS}

Obtain KEGG Pathway Information for All Clustered Genes

Description

A function for obtaining KEGG pathway information for all genes in an output of the cluster.wrapper function.

Usage

KEGG.cats(input, genome, gene.id)

Arguments

input

a list of lists as output by the cluster.wrapper function. Each list will contain a variable 'cluster', a vector indicating which cluster each gene is assigned to.

genome

a string identifying the genome that the data was originally aligned to. For a list of supported genomes, run the supportedGenomes function from the goseq geneLenDataBase package (see 'Examples').

gene.id

a string identifying the gene identifier used in the row names of the original count table run through cluster.wrapper. For a list of supported gene IDs, run the supportedGeneIDs function from the goseq geneLenDataBase package (see 'Examples').

Details

This function extracts gene information from the cluster.wrapper output and applies it to the getgo function of the goseq package (Young *et al.* 2010). This information can then be used with the cluster.enrichment or single.enrichment functions to apply KEGG pathway enrichment analyses.

genome and gene.id are also both applied to goseq's getgo function, which will search the built-in database of KEGG pathway information for suitable data to download, or if there is no suitable data, will download this from an online resource.

Value

KEGG.cats returns a list of KEGG pathway number, with each variable corresponding to a gene in the input data.

References

Young, M.D., Wakefield, M.J., Smyth, G.K. and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2): R14.

Examples

```
# perform model-based clustering
output <- cluster.wrapper(counts, treatment, nK.vector=2:4, additional.forks=2)
# view supported genomes and gene ids
library('goseq')
supportedGenomes()
supportedGeneIDs()
# obtain gene ontology information with supported genomes/geneids
gene2cat <- KEGG.cats(output, 'mm9', 'ensGene')
```

cluster.enrichment {ROCKS}

Parallel Enrichment Analysis for Clustering Data

Description

A function for multiple applications of goseq enrichment analyses (Young *et al.* 2010) across an output of the cluster.wrapper function. This function allows this to be executed in parallel through socket clusters.

Usage

cluster.enrichment(input, bias.data, gene2cat, use_genes_without_cat=FALSE, verbose=FALSE)

Arguments

input

a list of lists as output by the cluster.wrapper function. Each list will contain a variable 'cluster', a vector indicating which cluster each gene is assigned to.

bias.data

a numeric vector of gene lengths, with names matching the gene ids in the input data. This can be produced with the gene.lengths function.

gene2cat

a list of gene ontology codes or KEGG pathway numbers, with each variable corresponding to a gene in the input data. This can be produced with either the GO.cats function or the KEGG.cats function.

use_genes_without_cat

a logical flag to indicate whether genes in any given cluster that do not

have GO/KEGG information (in the gene2cat list) should still be used in the enrichment p-value calculation. If this option is set to false, these genes will be ignored (default).

verbose

a logical flag to enable verbose messages.

Details

a parallel socket cluster must be created for the enrichment tests to be run in parallel (see 'Examples'). For more information on the calculations involved, please see the goseq function help page.

Value

cluster.enrichment returns a list of lists. The main list corresponds to one variable for each running of the model-based clustering algorithm by the cluster.wrapper function. Within this is a list of enrichment results for each cluster. These enrichment results are in the form of a data frame for which each row represents an enriched category (e.g. a single gene ontology term). The first column gives the name of the category; the second gives the p-value for this category being over represented amongst the cluster. The third column gives the adjusted p-value, corrected for false discovery rate. The fourth column gives the number of genes in both the cluster and the category, with the fifth giving the overall total in the category. For gene ontology terms this table also includes ontology information in the final columns.

References

Young, M.D., Wakefield, M.J., Smyth, G.K. and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2): R14.

Examples

```
# perform model-based clustering
clustering.data <- cluster.wrapper(counts, treatment, nK.vector=2:4,
additional.forks=2)
# determine gene length data and gene ontology category data
bias.data <- gene.lengths(clustering.data , 'mm9', 'ensGene')
gene2cat <- GO.cats(clustering.data , 'mm9', 'ensGene')
# open a 6 socket cluster
library("doParallel")
cl <- makeCluster(6)
registerDoParallel(cl)
# apply cluster enrichment analysis
GO.output <- cluster.enrichment(clustering.data, bias.data, gene2cat)
# close the 6 socket cluster
stopCluster(cl)
# view the structure of the output
str(GO.output)
```

extract.clusters {ROCKS}

Extract Gene Information from All Clusters

Description

a function applied to a cluster.wrapper output, which extracts the names of the genes forming each cluster, performing this for all clustering runs.

Usage

```
extract.clusters(input)
```

Arguments

input

a list of lists as output by the cluster.wrapper function. Each list will contain a variable 'cluster', a vector indicating which cluster each gene is assigned to.

Details

the cluster.wrapper function returns a list of lists, for which each of the outer list is a running of the model-based clustering algorithm. The inner list of these includes a vector of cluster assignment. The extract.clusters function is useful for extracting this information and forming a flatter list in which each entry includes a vector of genes assigned to a particular cluster. This is useful for the user to view the genes of a specific single cluster, or to pass on to the 'single.enrichment' function for applying an enrichment test to any one cluster. See 'Examples' for information on how to do this.

Value

extract.clusters returns a list containing gene name vectors for each of the clusters in the cluster.wrapper output. These list entries are named to reflect their original position in the cluster.wrapper output, for example the entry named 'group1-cluster1' contains the genes of the first cluster in the first run of the model-based clustering algorithm.

Examples

```
# perform model-based clustering
clustering.data <- cluster.wrapper(counts, treatment, nK.vector=2:4,
additional.forks=2)
# extract gene clusters
clusters.genes <- extract.clusters(clustering.data)
# view the structure of the output
str(clusters.genes)
# view the genes of the 1st cluster from the 1st run
as.data.frame( extracted.clusters[[“group1-cluster1”]] )
```

single.enrichment {ROCKS}

Enrichment Analysis for Clustering Data

Description

A function for applying goseq enrichment analyses (Young *et al.* 2010) to an extracted cluster (using the extract.clusters function) from an output of the cluster.wrapper function.

Usage

```
single.enrichment(extracted.cluster, bias.data, gene2cat,
use_genes_without_cat=FALSE)
```

Arguments

input

an extracted cluster (see 'Examples') output by the cluster.wrapper function. This will contain a variable 'cluster', a vector indicating which cluster each gene is assigned to.

bias.data

a numeric vector of gene lengths, with names matching the gene ids in the input data. This can be produced with the gene.lengths function.

gene2cat

a list of gene ontology codes or KEGG pathway numbers, with each variable corresponding to a gene in the input data. This can be produced with either the GO.cats function or the KEGG.cats function.

use_genes_without_cat

a logical flag to indicate whether genes in any given cluster that do not

have GO/KEGG information (in the gene2cat list) should still be used in the enrichment p-value calculation. If this option is set to false, these genes will be ignored (default).

Details

For more information on the calculations involved, please see the goseq function help page.

Value

single.enrichment returns a list with the following components:

ID

the group and cluster number of the extracted cluster

GOresults

a data frame for which each row represents an enriched category (e.g. a single gene ontology term). The first column gives the name of the category, the second gives the p-value for this category being over represented amongst the cluster. The third column gives the adjusted p-value, corrected for false discovery rate. The fourth column gives the number of genes in both the cluster and the category, with the fifth giving the overall total in the category. For gene ontology terms this table also includes ontology information in the final columns.

unknown

the number of genes in the extracted cluster that had no category (absent from the bias.data or gene2cat input variables).

References

Young, M.D., Wakefield, M.J., Smyth, G.K. and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2): R14.

Examples

```
# perform model-based clustering
clustering.data <- cluster.wrapper(counts, treatment, nK.vector=2:4,
additional.forks=2)
# determine gene length data and gene ontology category data
bias.data <- gene.lengths(clustering.data , 'mm9', 'ensGene')
gene2cat <- GO.cats(clustering.data , 'mm9', 'ensGene')
# extract a single cluster
single.clustering.data <- extract.clusters(clustering.data)[[“group1-cluster1”]]
# apply cluster enrichment analysis
GO.output <- single.enrichment(single.clustering.data, bias.data, gene2cat)
# view the structure of the output
str(GO.output)
# view the goseq output
GO.output$GOresults
```

cluster.SS {ROCKS}

Parallel Semantic Similarity Calculator for Clustering Data

Description

A function for multiple applications of GOSemSim's semantic similarity calculation (Yu *et al.* 2010) across an output of the cluster.enrichment function, returning a mean semantic similarity for each cluster. This function allows this to be executed in parallel through socket clusters.

Usage

```
cluster.SS(GOclusters, organism, FDR.lim=0.05, verbose=FALSE)
```

Arguments

GOclusters

a list of lists as output by the cluster.enrichment function. The main list corresponds to one variable for each running of the model-based clustering algorithm by the cluster.wrapper function. Within this is a list of enrichment results for each cluster.

organism

one of the organism names as permitted by the GOSemSim function goSim, including: "anopheles", "arabidopsis", "bovine", "canine", "chicken", "chimp", "coelicolor", "ecolik12", "ecsakai", "fly", "human", "malaria", "mouse", "pig", "rat","rhesus", "worm", "xenopus", "yeast" and "zebrafish".

FDR.lim

The significance cut-off value for selection of gene ontology terms to be

applied. All enriched gene ontology terms with an FDR corrected p-value above this will be excluded. This has a default value of 0.05.

verbose

a logical flag to enable verbose messages.

Details

semantic similarity calculations are applied to all clusters of each model-based clustering algorithm run. The resultant value for each cluster is the mean of the semantic similarities of all pairwise combinations. For more information on the calculations involved, please see the Wang method of the GOSemSim vignette.

Value

cluster.SS returns a list of lists. The main list corresponds to one variable for each group of clusters. Within this is a list of numeric vectors containing the total number of significant gene ontologies, and the mean semantic similarity, for each cluster.

References

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y. and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7): 976-978

Examples

```
# apply semantic similarity calculations
SS.output <- cluster.SS(GO.output, organism='mouse', FDR.lim=0.05)
# view the structure of the output
str(SS.output)
```

combined.plot.prep {ROCKS}

Prepare Semantic Similarity and Gene Ontology Data for Plotting

Description

A function for extracting semantic similarity and gene ontology data from an output of the cluster.SS function and formatting these to be ready for plotting with the scatter.smooth function.

Usage

```
combined.plot.prep(SSclusters)
```

Arguments

SSclusters

a list of lists as output by the cluster.SS function. The main list corresponding to one variable for each group of clusters, and within this a list of numeric vectors containing the total number of significant gene ontologies, and the mean semantic similarity, for each cluster.

Details

When elucidating information from the GO.plot.prep and SS.plot.prep functions, it may be useful to combine these to a single curve. The product of these values can be used to find the cluster number (K) that maximizes significant gene ontology terms, and semantic similarity within the clusters, using the curved plot following the output by this function. As with AIC, the user should look for an 'elbow point' closest to the upper left corner of the plot. Although this data can be plotted with the regular plot function, this data is noisy, and we advise using the scatter.smooth function to fit a

smooth curve to the plot for easier identification of the elbow, followed by selection of K as the individual point closest to the upper left corner of the plot.

Value

combined.plot.prep returns a matrix where each row represents data from each running of the model-based clustering algorithm. The first column states the number of clusters produced in that run, and the second column contains the resultant plotting value, which is the normalized semantic similarity (normalized to a range of zero to one) multiplied by the number of significant gene ontology terms. This resultant matrix is suitable for passing directly through the scatter.smooth function (see 'Examples').

Examples

```
# apply semantic similarity calculations
SS.output <- cluster.SS(GO.output, organism='mouse', FDR.lim=0.05)
# prepare the data for plotting
prepdata <- main.plot.prep(SS.output)
# plot the data
scatter.smooth(prepdata)
```

GO.plot.prep {ROCKS}

Prepare Gene Ontology Data for Plotting

Description

A function for extracting gene ontology data from an output of the cluster.SS function and formatting these to be ready for plotting with the scatter.smooth function.

Usage

```
GO.plot.prep(SSclusters)
```

Arguments

SSclusters

a list of lists as output by the cluster.SS function. The main list corresponding to one variable for each group of clusters, and within this a list of numeric vectors containing the total number of significant gene ontologies, and the mean semantic similarity, for each cluster.

Details

As the number of clusters used increases, the total number of significant gene ontologies will increase before reaching a ceiling, giving a characteristic curve. Although this data can be plotted with the regular plot function, this data is noisy, and we advise using the scatter.smooth function to fit a smooth curve to the plot.

Value

GO.plot.prep returns a matrix where each row represents data from each running of the model-based clustering algorithm. The first column states the number of clusters

produced in that run, and the second column contains the total gene ontologies reported for all clusters. This resultant matrix is suitable for passing directly through the scatter.smooth function (see 'Examples').

Examples

```
SS.output <- cluster.SS(GO.output, organism='mouse', FDR.lim=0.05)
# prepare the data for plotting
prepdata <- GO.plot.prep(SS.output)
# plot the data
scatter.smooth(prepdata)
```

SS.plot.prep {ROCKS}

Prepare Semantic Similarity Data for Plotting

Description

A function for extracting semantic similarity data from an output of the cluster.SS function and formatting these to be ready for plotting with the scatter.smooth function.

Usage

```
SS.plot.prep(SSclusters)
```

Arguments

SSclusters

a list of lists as output by the cluster.SS function. The main list corresponding to one variable for each group of clusters, and within this a list of numeric vectors containing the total number of significant gene ontologies, and the mean semantic similarity, for each cluster.

Details

As the number of clusters used increases, the mean semantic similarity of significant gene ontologies will increase (as the clusters becomes more specified) before reaching a ceiling, giving a characteristic curve. Although this data can be plotted with the regular plot function, this data is noisy, and we advise using the scatter.smooth function to fit a smooth curve to the plot.

Value

SS.plot.prep returns a matrix where each row represents data from each running of the model-based clustering algorithm. The first column states the number of clusters produced in that run, and the second column contains the mean semantic similarity reported for all clusters. This resultant matrix is suitable for passing directly through the scatter.smooth function (see 'Examples').

Examples

```
SS.output <- cluster.SS(GO.output, organism='mouse', FDR.lim=0.05)
# prepare the data for plotting
prepdata <- SS.plot.prep(SS.output)
# plot the data
scatter.smooth(prepdata)
```

cluster.plot.prep {ROCKS}

Prepare Clustering Data for Plotting

Description

A function for standardizing a count table and extracting information from an output of the cluster.wrapper function, for use with the cluster.plot function.

Usage

```
cluster.plot.prep(counts, treatment, clustering.output, ncluster)
```

Arguments

counts

a GxP matrix storing the numbers of reads mapped to G genes in P samples. Row names should contain corresponding gene IDs for downstream ontology analysis.

treatment

a vector of length P specifying the corresponding treatments for each column of the counts matrix. See 'Details' for an example.

clustering.output

a list of lists as output by the cluster.wrapper function. Each list will contain the variables 'cluster', a vector indicating which cluster each gene is assigned to, 'probability', a matrix containing the probability of each gene belonging to each cluster, and 'centers', a matrix containing estimates of the cluster centres.

ncluster

the number of clusters to plot from, i.e., which run of the

clustering.output to plot (this will be one of the values of the nK.vector used in the original cluster.wrapper function).

Details

This function standardizes expression data for each gene in the count table to a mean of zero and standard deviation of one, allowing plotting of all genes within a cluster on the same graph. In addition to this, it subsets the clustering.output to one running of the model-based clustering algorithm (which is selected with the ncluster variable), and forwards the probabilities of each gene belonging to each cluster, as well as the cluster centre, for use with the cluster.plot function.

Value

cluster.plot.prep returns a list with the following components:

counts

a GxP matrix corresponding to the original counts matrix, but with values standardized for each gene to a mean of zero and standard deviation of one.

cluster

a vector indicating which cluster each gene has been assigned to. Each value corresponds to each gene (matching the row names of the counts matrix).

probability

a matrix containing the probability of each gene belonging to each cluster. Each row corresponds to each gene (matching the row names of the counts matrix) and each column corresponds to each cluster (as defined by ncluster).

centers

a matrix containing estimates of the cluster centres. Each row corresponds to each cluster (as defined by ncluster) and each column corresponds to each treatment (as defined by the treatment vector).

Examples

```
# perform model-based clustering
clustering.data <- cluster.wrapper(counts, treatment, nK.vector=2:4,
additional.forks=2)
# prepare the data for plotting clusters from the K=4 run of the model-based
clustering
prepdata <- cluster.plot.prep(counts, treatment, clustering.data, ncluster=4)
# example 1 - plot the gene expressions for a single cluster (the first)
cluster.plot(prepdata, cluster=1)
# example 2 – plot the gene expressions for all three clusters
# (it is much faster to do this directly to a file, such as a png)
png(filename="myplot.png")
par(mfrow=c(2,2))
for(i in 1:4){
   cluster.plot(prepdata, cluster=i)
}
dev.off()
```

cluster.plot {ROCKS}

Plot a Cluster

Description

A function for plotting the gene expression data for a cluster from an output of the cluster.plot.prep function.

Usage

cluster.plot(plot.prep, cluster, RColorBrewer.pal="YlGnBu", centroid.col="red", type="l", lwd=1, main=NULL, xlab="Treatment", ylab="Standardized Expression", ...)

Arguments

plot.prep

a list output by the cluster.plot.prep function, including a standardized counts matrix, a vector of cluster assignments, a matrix of gene probabilities for assignment to each cluster, and a matrix of estimated cluster centres.

cluster

single numeric value of which of the clusters to plot.

RcolorBrewer.pal

a colour palette for use with the brewer.pal function of the RcolorBrewer package (Neuwirth, 2011). This will be used to colour the gene expression lines by p-value. See the brewer.pal help page for palette options.

centroid.col

the colour to plot the cluster centre line. This should be chosen to stand

out from the palette chosen for RcolorBrewer.pal.

...

all other arguments will be passed to the embedded plot function, and may include graphical parameters (see 'par').

Details

This function plots a graph of standardized expression values with one line per gene. The line colours are defined by the palette selected in RcolorBrewer.pal, changing with the p-value for association with the selected cluster. Finally, the cluster centre is plotted over this.

References

Neuwirth, E. (2011). RColorBrewer: ColorBrewer palettes. R package version 1.0-5. http://CRAN.R-project.org/package=RColorBrewer

Examples

```
prepdata <- cluster.plot.prep(counts, treatment, clustering.data, ncluster=4)
# example 1 - plot the gene expressions for a single cluster (the first)
cluster.plot(prepdata, cluster=1)
# example 2 – plot the gene expressions for all three clusters
# (it is much faster to do this directly to a file, such as a png)
png(filename="myplot.png")
par(mfrow=c(2,2))
for(i in 1:4){
   cluster.plot(prepdata, cluster=i)
}
dev.off()
```

B2 - Raw Function Codes

cluster.wrapper

```
cluster.wrapper <- function( counts, treatment, nK.vector, additional.forks=1 ){
        if(length(treatment)!=ncol(counts)){ stop("counts does not match treatment") }
        require("doParallel")
        require("MBCluster.Seq")
        require("parallel")
        counts <- counts[which(apply(counts,1,var)!=0),]
        counts <- counts[which(apply(counts,1,sum)!=0),]
        indata <- para.RNASeq.Data(counts, Treatment=treatment,
                                                GeneID=rownames(counts))
        packs <- c(“ROCKS”,"parallel","MBCluster.Seq")
        message("Applying Expectation-Maximization Algorithms")
        output <- foreach(nK=nK.vector, .packages=packs) %dopar% {
                centre_calc <- para.KmeansPlus.RNASeq(indata, nK=nK,
                                        cores=additional.forks)
                cluster_calc <- Cluster.RNASeq.plus(data=indata,model="nbinom",
                                                centers=centre_calc$centers,method="EM")
                return(cluster_calc)
        } return(output)
}
```

Cluster.RNASeq.plus

```
Cluster.RNASeq.plus <- function (data, model, centers = NULL, method = c("EM", "DA", "SA"), iter.max = 30,
TMP = NULL){
        n = data$Count
        nG = nrow(n)
        nK = nrow(centers)
        s = data$Normalizer
        t = data$Treatment
        logFC = data$logFC
        n[n <= 0] = 1e-10
        nG = nrow(n)
        if (model == "poisson")
                nb.disp = rep(1e-10, nG)
        if (model == "nbinom")
                nb.disp = data$NB.Dispersion
        if (is.vector(nb.disp))
                nb.disp = matrix(rep(nb.disp, ncol(n)), nrow = nG)
        if (is.null(TMP)) {
                method = method[1]
                if (method == "EM")
                        TMP = rep(1, iter.max)
                if (method %in% c("DA", "SA"))
                        TMP = 4 * 0.9^(0:iter.max)
        }
        if (length(centers) == 1)
                centers = logFC[sample(nG, centers), ]
        C = centers
        P = matrix(1/nrow(C), nrow(n), nrow(C))
        M = matrix(0, nrow(n), nrow(C))
        for (k in 1:nrow(C)) M[, k] = log(rowSums(n)/rowSums(exp(sweep(s,
                2, C[k, t], "+"))))
        k = rep(1, nG)
```

```
    for (i in 1:length(TMP)) {
        P.new = cl.mb.est.P(n, s, t, model, P, C, M, nb.disp, a = 1/TMP[i],
            method = method)
        MC = cl.mb.est.MC(n, s, t, model, P.new, C, M, nb.disp,
            method = method)
        M.new = MC$M
        C.new = MC$C
        k.new = apply(P.new, 1, which.max)
        dif = mean(k != k.new)
        if (i > 20 & dif < 0.001)
            break
        M = M.new
        C = C.new
        k = k.new
        P = P.new
    }
    names(k) <- rownames(n)
    print(paste("--->>>> iteration stops after", i, "steps"))
    lglk = matrix(0, nrow = nrow(n), ncol = nK)
    for (l in 1:nK) {
        r = exp(sweep(s + M[, l], 2, C[l, t], "+"))
        lglk[, l] = lglk.nb(n, r, v = nb.disp)
    }
    lglk = sum(P * lglk)/nG
    nT <- length(unique(t))
    nGi <- mean(table(k))
    nparams <- (nGi * (nK + 1)) + (nK * nT) - 1
    aic <- -2*(lglk-nparams)
    return(list(probability = P, centers = C, cluster = k, lglk = lglk, nparams = nparams, aic = aic))
}
```

para.KmeansPlus.RNASeq

```
para.KmeansPlus.RNASeq <- function (data, nK, print.steps = TRUE, cores=1) {
        n = data$Count
        s = data$Normalizer
        t = data$Treatment
        logFC = data$logFC
        v = data$NB.Dispersion
        nG = nrow(n)
        nT = length(unique(t))
        m = sumRow(n, by = t)/sumRow(exp(s), by = t)
        m = m[, t] * exp(s)
        lglk0 = rowSums(lgamma(n + 1/v) - lgamma(n + 1) - lgamma(1/v) -
                        log(1 + m * v)/v - log(1 + 1/m/v) * n)
        C = matrix(0, nK, nT) ; ID = rep(0, nK) ; c0 = rep(0, nT)
        colnames(C) = colnames(logFC)
        lglk = c()
        for (i in 1:nK) {
                if (print.steps){ print(paste("--> centroid", I)) }
                lglk = cbind(lglk, para.lglk.nb.c(n = n, s = s, t, v = v,
                                C = c0, cores=cores))
                d = -2 * (lglk - lglk0)
                if (i == 1){ d = d[, 1] }
                if (i > 1){ d = apply(abs(d), 1, min) }
                pr = d^2/sum(d^2)
                pr = pr - min(pr) + (max(pr) - min(pr)) * 1e-10
                id = sample(nG, 1, prob = pr)
                C[i, ] = c0 = logFC[id, ]
                ID[i] = id
        }
        class(C) = "logFC"
        return(list(centers = C, id = ID))
}
```

para.lglk.nb.c

```
para.lglk.nb.c <- function (n, s, t, v, C, cores=1){
        if (is.vector(C)){ C = matrix(C, nrow = 1) }
        nK = nrow(C)
        nG = nrow(n)
        lglk = matrix(0, nG, nK)
        for (k in 1:nK) {
                lglk[, k] = para.lglk.nb.1c(n = n, s = s, t, v, c = C[k, ], cores=cores)
        }
        return(lglk)
}
```

para1.est.nb.mu.mle.one

```
para1.est.nb.mu.mle.one <- function (n, c, v, mu0 = NULL) {
        if (is.vector(n)){ return(n/c) }
        if (is.null(mu0)){ mu0 = rowSums(n)/rowSums(c) }
        nJ = ncol(n)
        nG = nrow(n)
        nU = 400
        u = seq(-2, 2, length.out = 400)
        tmp.lglk <- foreach(i=as.list(1:nU)) %dopar% {
                m = c * mu0 * exp(u[i])
                lglk = rowSums(lgamma(n + 1/v) - lgamma(n + 1) - lgamma(1/v) - log(1 + m * v)/v - log(1 +
1/m/v) * n)
                return(lglk)
        }
        lglks <- do.call(cbind,tmp.lglk)
        id <- apply(lglks,1,which.max)
        mu = mu0 * exp(u[id])
        return(mu)
}
```

para.lglk.nb.1c

```
para.lglk.nb.1c <- function (n, s, t, v, c, cores=1){
        sc = sweep(s, 2, c[t], "+")
        m = para2.est.nb.mu.mle.one(n, exp(sc), v, cores=cores)
        m = exp(sc) * m
        lglk = rowSums(lgamma(n + 1/v) - lgamma(n + 1) - lgamma(1/v) -
                log(1 + m * v)/v - log(1 + 1/m/v) * n)
        return(lglk)
}
```

para2.est.nb.mu.mle.one

```
para2.est.nb.mu.mle.one <- function (n, c, v, mu0 = NULL, cores=1){
        if (is.vector(n)){ return(n/c) }
        if (is.null(mu0)){ mu0 = rowSums(n)/rowSums(c) }
        nJ = ncol(n)
        nG = nrow(n)
        nU = 400
        u = seq(-2, 2, length.out = 400)
        tmp.lglk <- mclapply (as.list(1:nU), function(i) {
                m = c * mu0 * exp(u[i])
                lglk = rowSums(lgamma(n + 1/v) - lgamma(n + 1) - lgamma(1/v) -
                        log(1 + m * v)/v - log(1 + 1/m/v) * n)
                return(lglk)
        }, mc.cores=cores)
        lglks <- do.call(cbind,tmp.lglk)
        id <- apply(lglks,1,which.max)
        mu = mu0 * exp(u[id])
        return(mu)
}
```

para.RNASeq.data

```
para.RNASeq.Data <- function(Count, Normalizer = NULL, Treatment, GeneID =NULL) {
        Count = data.matrix(Count)
        Count = matrix(Count, ncol = length(Treatment))
        if (is.null(Normalizer)) {
                n = Count
                n[n <= 0] = NA
                q3 = apply(n, 2, quantile, 0.75, na.rm = TRUE)
                Normalizer = log(q3/mean(q3))
        }
        if (is.vector(Normalizer))
                Normalizer = matrix(rep(Normalizer, each = nrow(Count)),
                        ncol = length(Treatment))
        Count = Count[, order(Treatment)]
        Normalizer = Normalizer[, order(Treatment)]
        Treatment = Treatment[order(Treatment)]
        Treatment = cumsum(!duplicated(Treatment))
        if (is.null(GeneID)){ GeneID = paste(1:nrow(Count)) }
        colnames(Count) = colnames(Normalizer) = Treatment
        rownames(Count) = GeneID
        if (nrow(Normalizer) == 1){ Normalizer = c(Normalizer) }
        fc = sumRow(Count, by = Treatment)/sumRow(exp(Normalizer), by = Treatment)
        fc[fc <= 0] = min(fc[fc > 0])/10
        logFC = log(fc) - rowMeans(log(fc))
        colnames(logFC) = unique(Treatment)
        expr = rowMeans(fc)
        disp = para.est.nb.v.QL(Count, Normalizer, Treatment)
        data = list(GeneID = GeneID, Count = Count, Normalizer = Normalizer,
                Treatment = Treatment, Aver.Expr = expr, logFC = logFC,
                NB.Dispersion = disp)
        return(data)
}
```

para.est.nb.v.QL

```
para.est.nb.v.QL <- function(Count, Normalizer, Treatment, v=NULL, max.iter=5) {
        Count = data.matrix(Count) + 1e-10
        Size = exp(Normalizer)
        v = rep(0.1, nrow(Count))
        m = sumRow(Count, by = Treatment)/sumRow(Size, by = Treatment)
        for (i in 1:max.iter) {
                for (j in 1:ncol(m)) m[, j] = para1.est.nb.mu.mle.one(Count[,
                        Treatment == j], Size[, Treatment == j], v, mu0 = m[, j])
                mu = m[, Treatment] * Size
                v = para.est.nb.v.QL.one(Count, mu = mu, v0 = v)
        }
        return(v)
}
```

para.est.nb.v.QL.one

```
para.est.nb.v.QL.one <- function (n, mu, v0 = NULL){
        n = n + 1e-10 ; mu = mu + 1e-10
        nJ = ncol(n) ; nG = nrow(n)
        nU = 400
        if (is.null(v0)){ v0 = rep(0.1, nG) }
        u = seq(-2, 2, length.out = 400)
        tmp.d <- foreach(i=as.list(1:nU)) %dopar% {
                v = v0 * exp(u[i])
                Qlglk = n * log(n/mu) - (n + 1/v) * log((n + 1/v)/(mu + 1/v))
                d = abs(2 * rowSums(Qlglk) - (nJ - 1))
                return(d)
        }
        ds <- do.call(cbind,tmp.d)
        id <- apply(ds,1,which.min)
        return( v0*exp(u[id]) )
}
```

cluster.plot.prep

```
cluster.plot.prep <- function(counts, treatment, clustering.output, ncluster){
        if(length(treatment)!=ncol(counts)){
                stop("counts does not match treatment")
        }
        counts <- counts[which(apply(counts,1,var)!=0),]
        counts <- counts[which(apply(counts,1,sum)!=0),]
        mean.counts <- c()
        for(i in unique(treatment)){
                tmp.counts <- counts[,which(treatment==i),drop=F]
                tmp.means <- apply(tmp.counts,1,mean)
                mean.counts <- cbind(mean.counts,tmp.means)
        }
        standardized.counts <- t(apply( mean.counts,1,function(x)                                    (x-
mean(x,na.rm=T))/sd(x,na.rm=T) ))
        colnames(standardized.counts) <- as.character(unique(treatment))
        clusters <- unlist(lapply( clustering.output, function(y)
                length(unique(y$cluster)) ))
        which.clustering <- which(clusters==ncluster)
        if(length(which.clustering)==0){
                stop("clustering.output does not contain ncluster provided")
        }
        tmp.clustering <- clustering.output[[which.clustering]]
        if(length(tmp.clustering$cluster)!=nrow(counts)){
                stop("counts does not match the clustering.output")
        }
        return(list(counts=standardized.counts,
                                cluster=tmp.clustering$cluster,
                                pvals=tmp.clustering$probability,
                                centres=tmp.clustering$centers))
}
```

cluster.plot

```
cluster.plot <- function(plot.prep, cluster, RColorBrewer.pal="YlGnBu",
        centroid.col="red", type="l", lwd=1, main=NULL, ylim=NULL,
                xlab="Treatment", ylab="Standardized Expression", ...){
        require("RColorBrewer")
        which.c <- which(plot.prep$cluster==cluster)
        if(length(which.c)==0){
                stop("ncluster does not contain cluster requested")
        }
        pval.order <- order(plot.prep$pvals[which.c])
        cluster.counts <- plot.prep$counts[which.c,][pval.order,]
        cluster.counts <- cluster.counts[complete.cases(cluster.counts),]
        centre <- plot.prep$centres[cluster,]
        centre <- (centre-mean(centre))/sd(centre)
        cols <- colorRampPalette(brewer.pal(8,RcolorBrewer.pal))
                (nrow(cluster.counts))
        if(is.null(ylim)){ ylim <- c( min(c(cluster.counts,centre)),
                        max(c(cluster.counts,centre)) ) }
        if(is.null(main)){ main <- paste("Cluster",cluster) }
        plot(cluster.counts[1,],type=type,main=main,xlab=xlab,ylab=ylab,
                ylim=ylim,col=cols[1],lwd=lwd,...)
        for(j in 2:nrow(cluster.counts)){
                lines(cluster.counts[j,],type=type,col=cols[j],lwd=lwd)
        }
        lines(centre,type=type,col=centroid.col,lwd=3*lwd)
}
```

AIC.plot.prep

```
AIC.plot.prep <- function(clustering.output){
        K <- unlist( lapply(clustering.output,function(x)
                length(unique(x[["cluster"]])))  )
        AIC <- unlist( lapply(clustering.output,function(x) x[["aic"]]) )
        prep.AIC <- cbind(K,AIC)
        rownames(prep.AIC) <- K
        colnames(prep.AIC)[1] <- c("Number of Clusters")
        return(prep.AIC)
}
```

gene.lengths

```
gene.lengths <- function(input,genome,gene.id){
        require("goseq")
        require("GO.db")
        allgenes <- unique(unlist(lapply( input, function(x) names(x$cluster) )))
        bias.data <- getlength(allgenes,genome,gene.id)
        names(bias.data) <- allgenes
        return(bias.data)
}
```

GO.cats

```
GO.cats <- function(input,genome,gene.id){
        require("goseq")
        require("GO.db")
        allgenes <- unique(unlist(lapply( input, function(x) names(x$cluster) )))
        gene2cat = getgo(allgenes,genome,gene.id,
                        fetch.cats=c("GO:CC","GO:BP","GO:MF"))
        names(gene2cat) = allgenes
        cat2gene = goseq:::reversemapping(gene2cat)
        gene2cat = goseq:::reversemapping(cat2gene)
        return(gene2cat) }
```

cluster.enrichment

```
cluster.enrichment <- function(input, bias.data, gene2cat,
                                    use_genes_without_cat=FALSE, verbose=FALSE){
        require("goseq")
        require("GO.db")
        extracted.clusters <- extract.clusters(input)
        message("Applying GOSeq to clusters")
        GOs <- foreach(i=1:length(extracted.clusters),.verbose=verbose,
                        .packages=c("goseq","GO.db"),.export=c("single.enrichment"))%dopar%{
                single.enrichment(extracted.clusters[[i]],bias.data=bias.data,
                        gene2cat=gene2cat,use_genes_without_cat=use_genes_without_cat)
        }
        message(paste(GOs[[1]]$unknown,"genes had no categories and were excluded"))
        restored.clusters <- restore.clusters(GOs)
        return(restored.clusters)
}
```

restore.clusters

```
restore.clusters <- function(GOs){
        restored.clusters <- list()
        for(b in 1:length(GOs)){
                tmp <- unlist(strsplit(GOs[[b]]$ID,"-"))
                restored.clusters[[tmp[1]]][[tmp[2]]] <- GOs[[b]]$GOresults
        }
        return(restored.clusters)
}
```

single.enrichment

```
single.enrichment <- function(extracted.cluster, bias.data, gene2cat,
                                    use_genes_without_cat=FALSE){
        ingenes <- as.integer(extracted.cluster$clustering)
        NAs <- is.na(bias.data)
        ingenes <- ingenes[!NAs]
        bias.data <- bias.data[!NAs]
        pwf <- suppressMessages(nullp(ingenes, bias.data=bias.data, plot.fit=F))
        rownames(pwf) <- names(bias.data)
        unknown <- nrow(pwf) - length(gene2cat)
        results <- suppressMessages(goseq(pwf, gene2cat=gene2cat,
                        use_genes_without_cat=use_genes_without_cat))
        colnames(results)[grepl("under_represented_pvalue",colnames(results))] <-
                        "over_represented_FDR"
        results$over_represented_FDR <-
                        p.adjust(results$over_represented_pvalue,method="fdr")
        return(list(ID=extracted.cluster$ID,GOresults=results,unknown=unknown))
}
```

extract.clusters

```
extract.clusters <- function(input){
        extracted.clusters <- list()
        for(a in 1:length(input)){
                tmp <- input[[a]]$cluster
                for(b in sort(unique(tmp))){
                        tmp.out1 <- names(tmp) %in% names(tmp[which(tmp==b)])
                        names(tmp.out1) <- names(tmp)
                        tmp.title <- paste("group",a,"-cluster",b,sep="")
                        tmp.out2 <- list(ID=tmp.title,clustering=tmp.out1)
                        extracted.clusters[[tmp.title]] <- tmp.out2
        } }
return(extracted.clusters) }
```

cluster.SS

```
cluster.SS <- function(GOclusters,organism="human",FDR.lim=0.05,verbose=FALSE){
        require("GOSemSim")
        extracted.GOs <- extract.GOs(GOclusters)
        out.clusterSS <- foreach(i=names(extracted.Gos),.packages="GOSemSim",
                                .verbose=verbose,.export=c("single.SS.apply",
                                "single.SS","goSim.apply","goSim.prep")) %dopar% {
          single.SS.apply(extracted.GOs[[i]],organism=organism,FDR.lim=FDR.lim)
        }
        restored.GOs <- restore.GOs(out.clusterSS)
        return(restored.GOs) }
```

restore.GOs

```
restore.GOs <- function(out.clusterSS){
        restored.GOs <- list()
        for(b in 1:length(out.clusterSS)){
                tmp <- unlist(strsplit(out.clusterSS[[b]]$ID,"-"))
                restored.GOs[[tmp[1]]][[tmp[2]]] <- out.clusterSS[[b]]$SSanalysis
        }
return(restored.GOs) }
```

extract.GOs

```
extract.GOs <- function(GOclusters){
        extracted.GOs <- list()
        for(a in 1:length(GOclusters)){
                tmp <- GOclusters[[a]]
                for(b in 1:length(tmp)){
                        tmp.out1 <- tmp[[b]]
                        tmp.title <- paste("group",a,"-cluster",b,sep="")
                        tmp.out2 <- list(ID=tmp.title,GOanalysis=tmp.out1)
                        extracted.GOs[[tmp.title]] <- tmp.out2
        } } return(extracted.GOs) }
```

single.SS.apply

```
single.SS.apply <- function(extracted.GO,organism="human",FDR.lim=0.05){
        which.FDR <- which(extracted.GO$GOanalysis$over_represented_FDR < FDR.lim)
        nGOs <- length(which.FDR)
        if(nGOs>1){
                GOlist <- cbind(extracted.GO$GOanalysis$category,
                        extracted.GO$GOanalysis$ontology)[which.FDR,]
                SS <- single.SS(GOlist,organism=organism)
        }else{ SS <- NA }
        tmp.out.SS <- c(nGOs,SS)
        names(tmp.out.SS) <- c("nGOs","SS")
        out.SS <- list(ID=extracted.GO$ID,SSanalysis=tmp.out.SS)
        return(out.SS)
}
```

single.SS

```
single.SS <- function(GOlist,organism="human"){
        SS.pairs <- goSim.prep(GOlist,organism=organism)
        if(!( length(unique(SS.pairs))==1 && is.na(unique(SS.pairs)) )){
                SS.vals <- apply(SS.pairs,1,goSim.apply)
        }else{ SS.vals <- NA }
        if(!( length(unique(SS.vals))==1 && is.na(unique(SS.vals)) )){
                SS.vals <- SS.vals[!is.na(SS.vals)]
                return(mean(SS.vals))
        }else{ return(NA) }
}
```

goSim.apply

```
goSim.apply <- function(SS.pair){
        x <- goSim(SS.pair[1],SS.pair[2],ont=SS.pair[3],organism=SS.pair[4])
        return(x)
}
```

goSim.prep

```
goSim.prep <- function(GOlist,organism="human"){
        SS.pairs <- c()
        for(ont in unique(GOlist[,2])){
                tmp.which <- which(GOlist[,2]==ont)
                if(length(tmp.which)>1){
                        tmp.codes <- GOlist[tmp.which,1]
                        tmp.SS.pairs1 <- t(combn(tmp.codes,2))
                        tmp.SS.pairs2 <- cbind(tmp.SS.pairs1,rep(ont,nrow(tmp.SS.pairs1)),
                                        rep(organism,nrow(tmp.SS.pairs1)))
                        SS.pairs <- rbind(SS.pairs,tmp.SS.pairs2)
                }
        }
        if(length(SS.pairs)==0){ SS.pairs <- NA }
        return(SS.pairs)
}
```

combined.plot.prep

```
combined.plot.prep <- function(SSclusters){
        SS <- plot.prep(SSclusters,"SS",mean)
        norm.SS <- apply( SS, 1, function(x)
                                (x[2]-min(SS[,2]))/(max(SS[,2])-min(SS[,2])) )
        nGOs <- plot.prep(SSclusters,"nGOs",sum)
        SSGO <- cbind(SS[,1],norm.SS*nGOs[,2])
        colnames(SSGO) <- c("Number of Clusters",
                        "Normalized Semantic Similarity x Total Enriched GO Terms")
        return(SSGO)
}
```

GO.plot.prep

```
GO.plot.prep <- function(SSclusters){
        nGOs <- plot.prep(SSclusters,"nGOs",sum)
        colnames(nGOs) <- c("Number of Clusters",
                            "Total Enriched Gene Ontology Terms")
        return(nGOs)
}
```

SS.plot.prep

```
SS.plot.prep <- function(SSclusters){
        SS <- plot.prep(SSclusters,"SS",mean)
        colnames(SS) <- c("Number of Clusters","Semantic Similarity")
        return(SS)
}
```

plot.prep

```
plot.prep <- function(SSclusters,plot.variable,combine.fun){
        SSlist <- list()
        for(i in names(SSclusters)){
                SScluster <- unlist(lapply(SSclusters[[i]],"[[",plot.variable))
                SSlist[[as.character(length(SScluster))]] <- SScluster
        }
        SSlist <- lapply(SSlist, function(x) x[!is.na(x)])
        result <- unlist(lapply(SSlist,combine.fun))
        k <- as.numeric(names(result))
        output <- cbind(k,result)
}
```

B3 - Internal Function Help Pages

R Documentation

para.RNASeq.Data {ROCKS}

Internal Function for cluster.wrapper

Description

para.RNASeq.Data is an internal function for the cluster.wrapper function. It is modeled on the RNASeq.Data function of the MBCluster.Seq package. The main edit of which is an inner function substituted for one with parallel processing capabilities.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Internal Function for para.RNASeq.Data

Description

para.est.nb.v.QL is an internal function for the para.RNASeq.Data function. It is modeled on the est.nb.v.QL function of the MBCluster.Seq package. The main edit of which is two inner functions substituted for others with parallel processing capabilities.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Internal Function for para.est.nb.v.QL

Description

para1.est.nb.mu.mle.one is an internal function for the para.est.nb.v.QL function. It is modeled on the est.nb.mu.mle.one function of the MBCluster.Seq package. The main edit of which is an inner loop substituted for one with parallel processing capabilities.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Internal Function for para.est.nb.v.QL

Description

para.est.nb.v.QL.one is an internal function for the para.est.nb.v.QL function. It is modeled on the est.nb.v.QL.one function of the MBCluster.Seq package. The main edit of which is an inner loop substituted for one with parallel processing capabilities.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Internal Function for cluster.wrapper

Description

para.KmeansPlus.RNASeq is an internal function for the cluster.wrapper function. It is modeled on the KmeansPlus.RNASeq function of the MBCluster.Seq package. The main edit of which is an inner function substituted for one with forking capabilities.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Internal Function for para.KmeansPlus.RNASeq

Description

para.lglk.nb.c is an internal function for the para.KmeansPlus.RNASeq function. It is modeled on the lglk.nb.c function of the MBCluster.Seq package, with an inner function substituted for one with forking capabilities.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Internal Function for para.lglk.nb.c

Description

para.lglk.nb.1c is an internal function for the para.lglk.nb.c function. It is modeled on the lglk.nb.1c function of the MBCluster.Seq package, with an inner function substituted for one with forking capabilities.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Internal Function for para.lglk.nb.1c

Description

para2.est.nb.mu.mle.one is an internal function for the para.lglk.nb.1c function. It is modeled on the est.nb.mu.mle.one function of the MBCluster.Seq package, with an inner function substituted for one with forking capabilities.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

Internal Function for cluster.wrapper

Description

Cluster.RNASeq.plus is an internal function for the cluster.wrapper function. It is modeled on the Cluster.RNASeq function of the MBCluster.Seq package, with additional functionality for extracting the log-likelihood and calculating the number of paramaters of the final model, as well as for calculating AIC. This function acts in an iterative manner and so was not capable of modifying for forking.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

R Documentation

restore.clusters {ROCKS}

Internal Function for cluster.enrichment

Description

restore.clusters is an internal function for the cluster.enrichment function. It is used to reformat multiple outputs of the single.enrichment function into a list to be returned to the user.

R Documentation

extract.GOs {ROCKS}

Internal Function for cluster.SS

Description

extract.GOs is an internal function for the cluster.SS function. It is used to reformat and flatten the GOclusters list input into the cluster.SS function, to allow easy parallel processing data.

R Documentation

single.SS.apply {ROCKS}

Internal Function for cluster.SS

Description

single.SS.apply is an internal function for the cluster.SS function. It is used to extract significantly enriched gene ontology terms from a single variable of the output from the extract.GOs function, and pass these through the single.SS function.

R Documentation

single.SS {ROCKS}

Internal Function for single.SS.apply

Description

single.SS is an internal function for the single.SS.apply function. It is used to pass gene ontology codes through the goSim.prep and goSim.apply functions to calculate all semantic similarity values, which it then takes the mean of.

Internal Function for single.SS

Description

goSim.prep is an internal function for the single.SS function. It is used to build all pairwise combinations for a vector of gene ontology codes, which can be passed on to the goSim.apply function to calculate semantic similarity for each of these, before being passed back to the single.SS function, which takes the mean value.

Internal Function for cluster.SS

Description

restore.GOs is an internal function for the cluster.SS function. It is used to reformat multiple outputs of the single.SS.apply function into a list to be returned to the user.

Internal Function for single.SS

Description

goSim.apply is an internal function for the single.SS function. It is used to apply the goSim feature from the GOSemSim package (Si, 2012) to calculate semantic similarity values for all pairwise combinations of gene ontology codes provided to it by the goSim.prep function. These semantic similarity values are then passed back to the single.SS function, which takes the mean value.

References

Si, Y. (2012). MBCluster.Seq: Model-Based Clustering for RNA-Seq Data. R package version 1.0. http://CRAN.R-project.org/package=MBCluster.Seq.

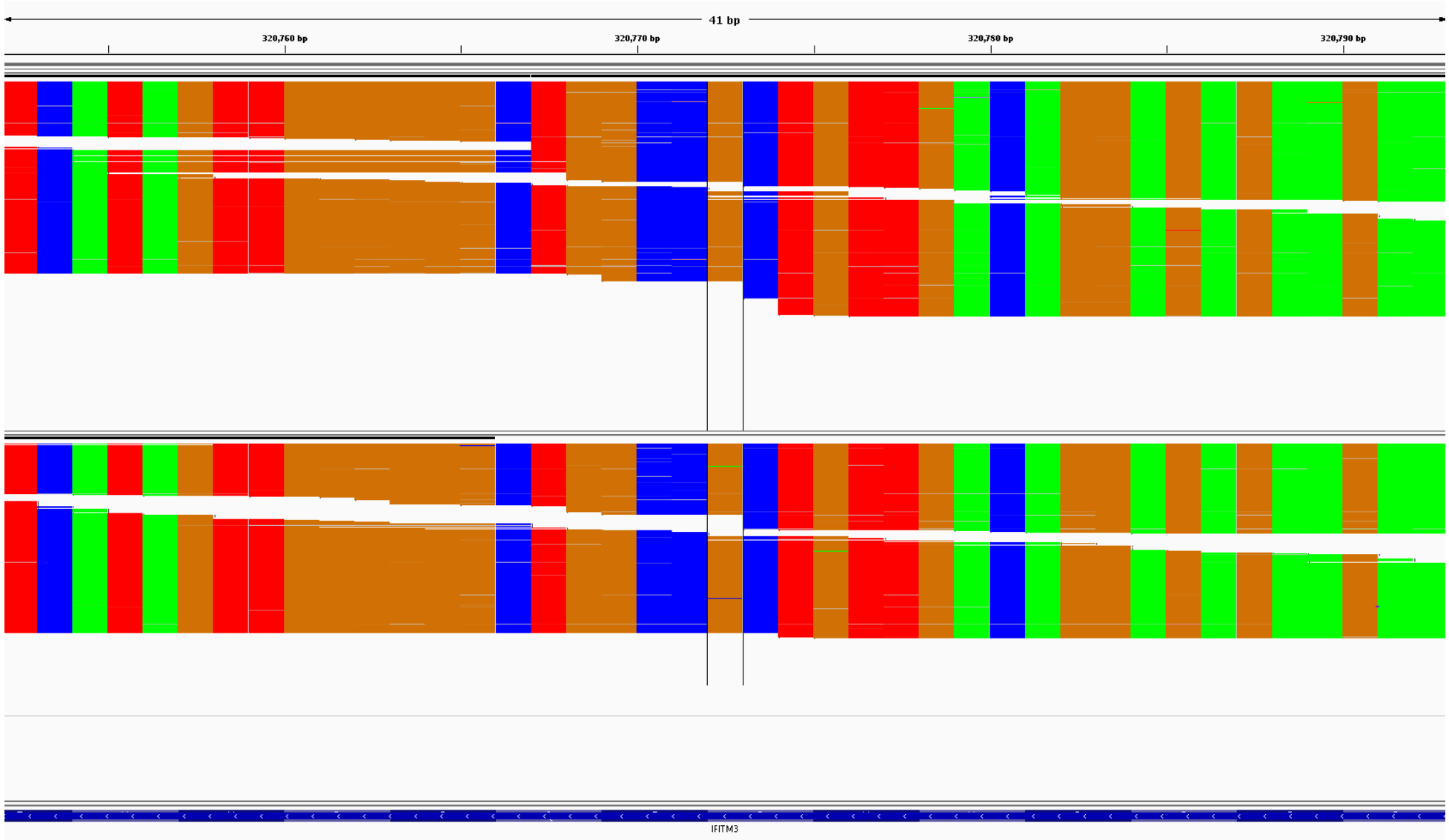
Internal Function for combined.plot.prep, GO.plot.prep, and SS.plot.prep

Description

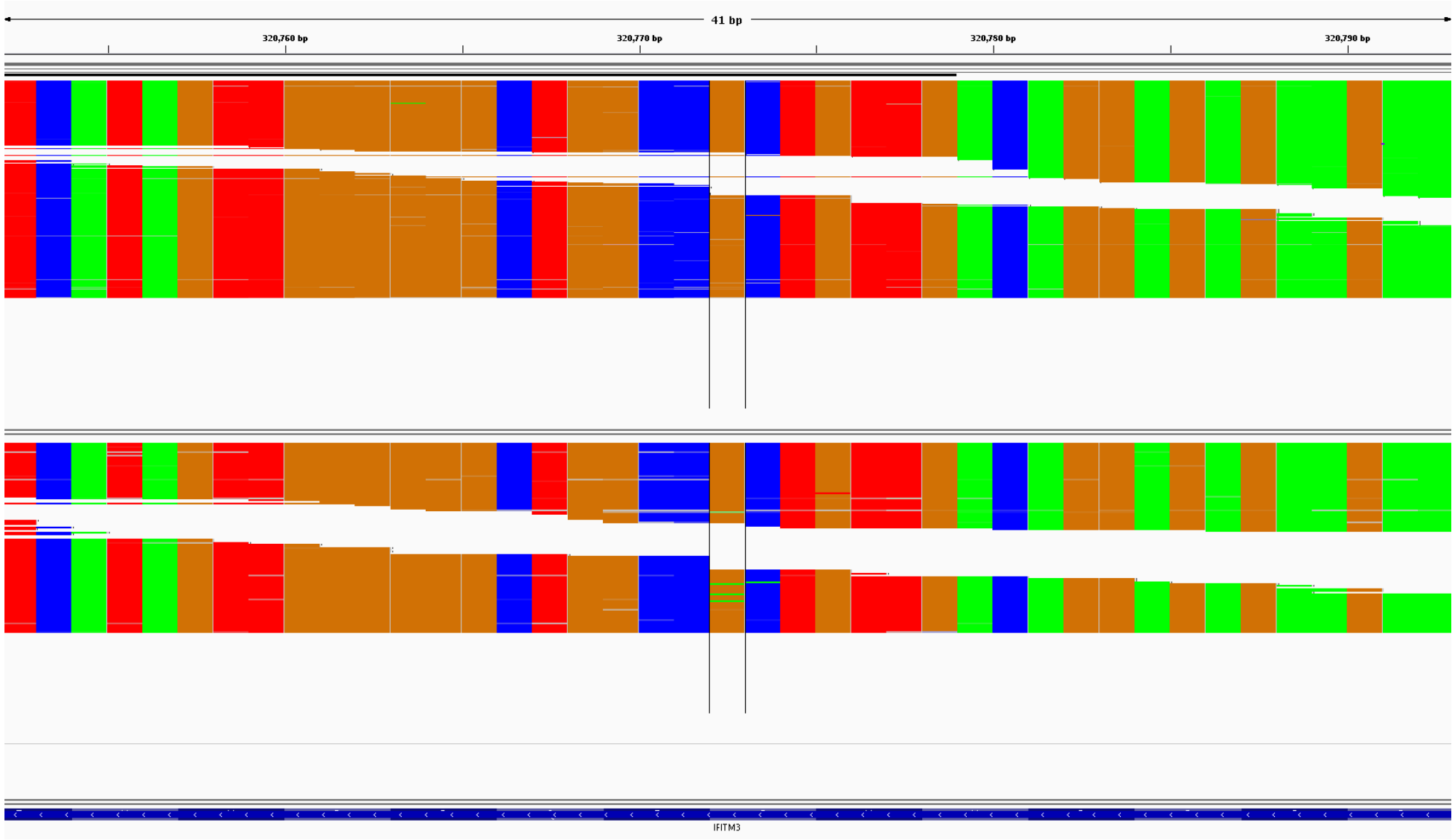
plot.prep is an internal function for the combined.plot.prep, GO.plot.prep, and SS.plot.prep functions. It is used to extract gene ontology or semantic similarity information from a list returned by the cluster.SS function, and reformat this to a matrix suitable for plotting.

Appendix C – Supplemental Materials for Chapter 6

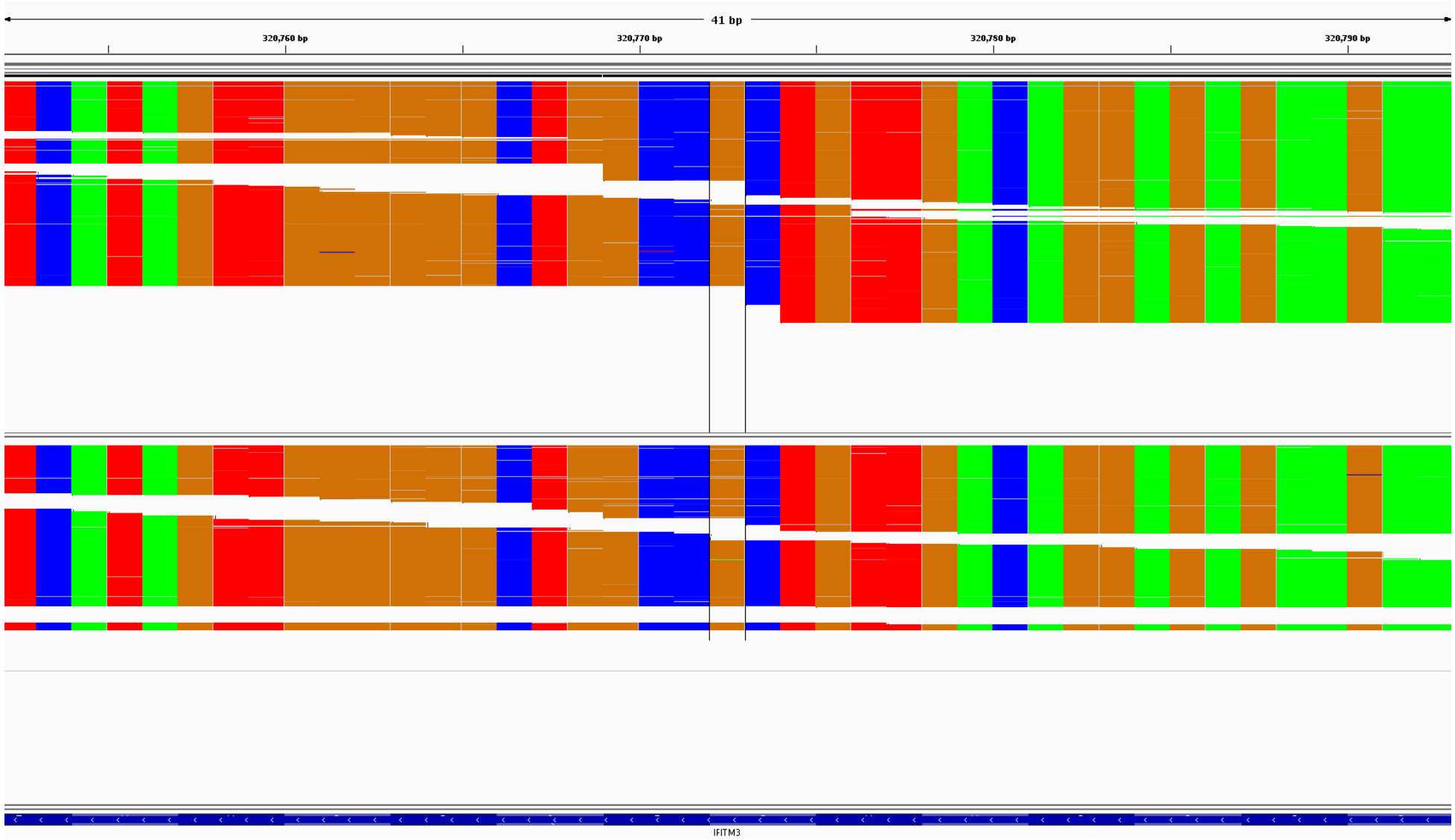
CC Replicate 1



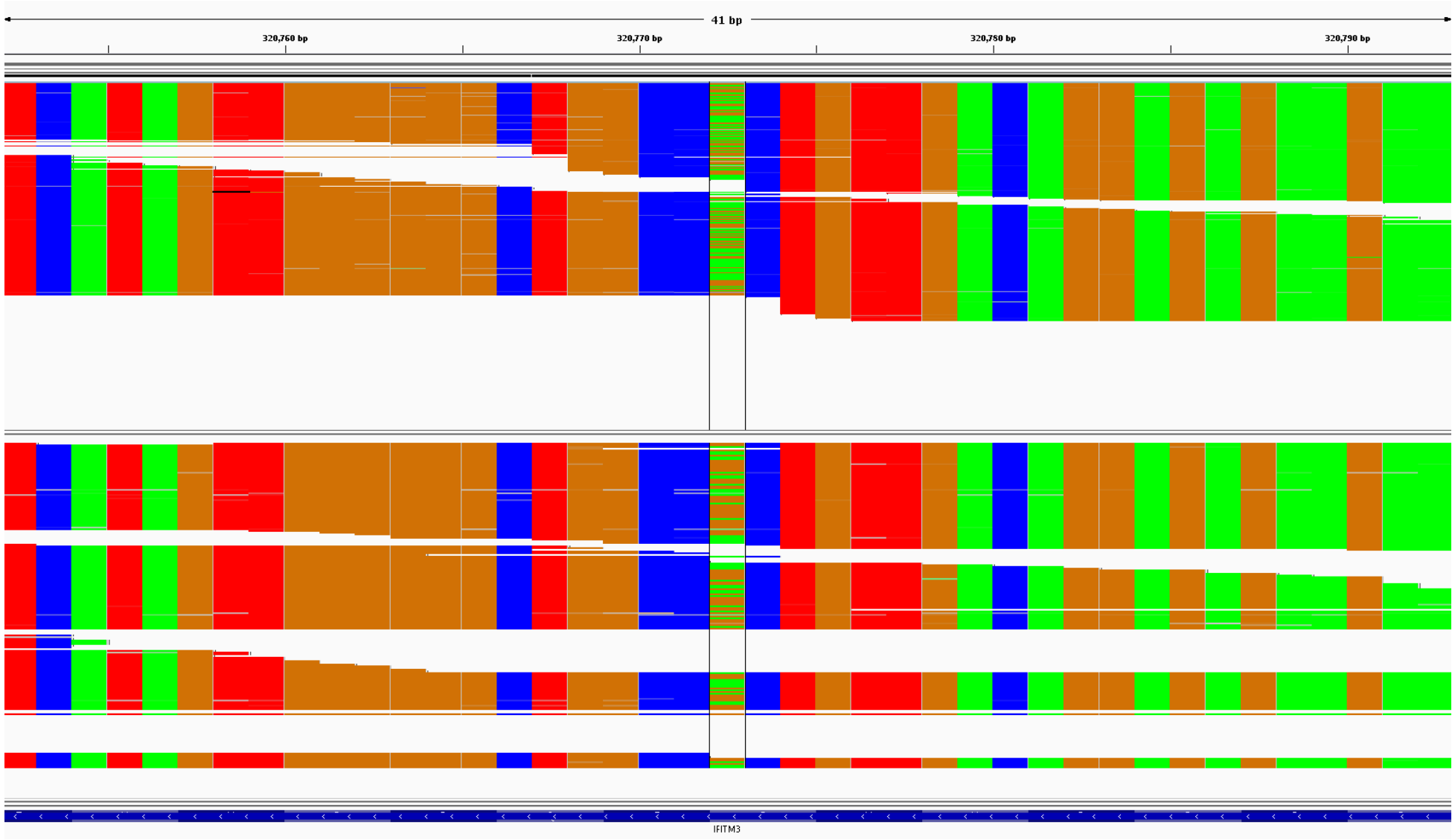
CC Replicate 2



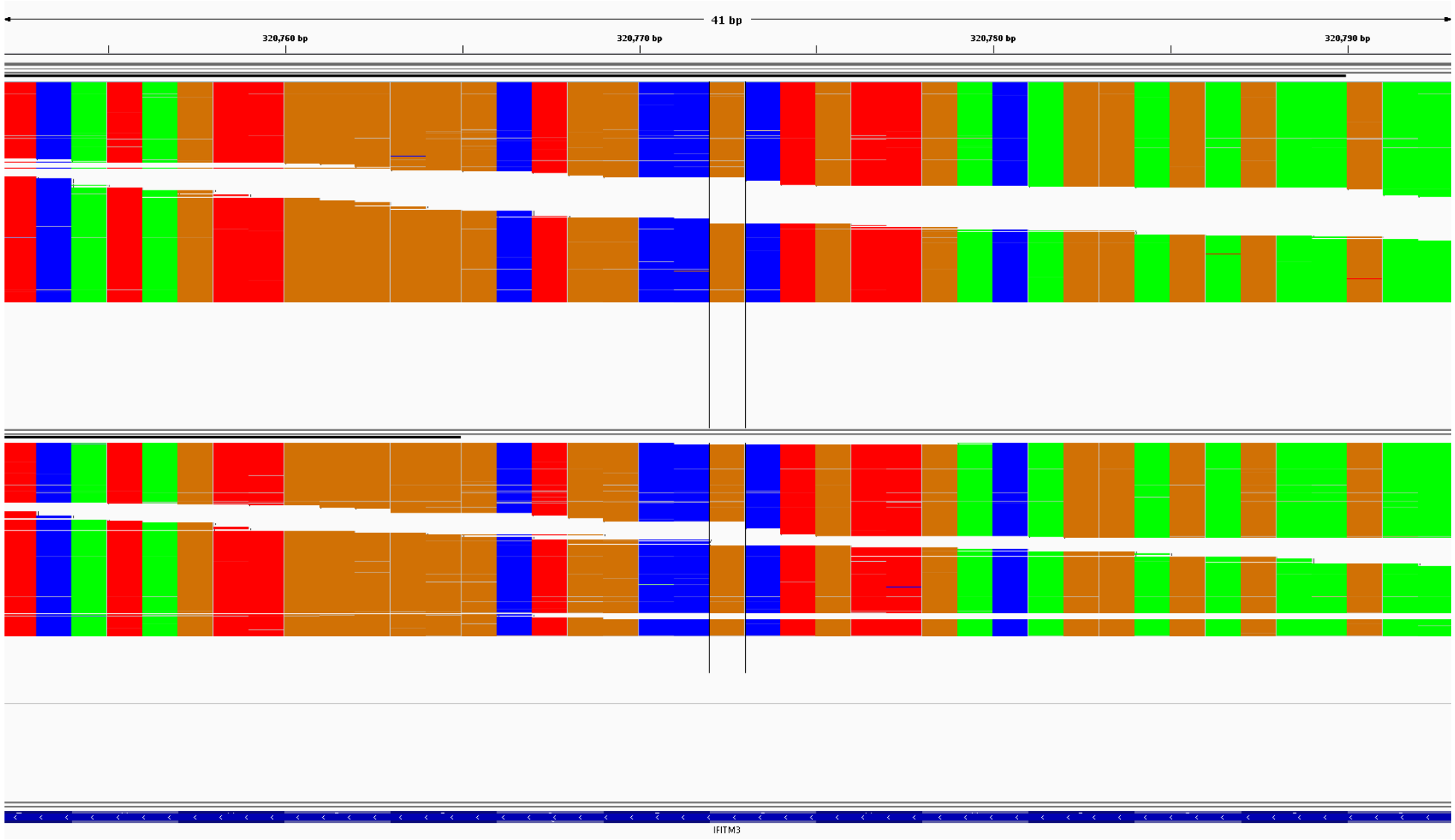
CC Replicate 3



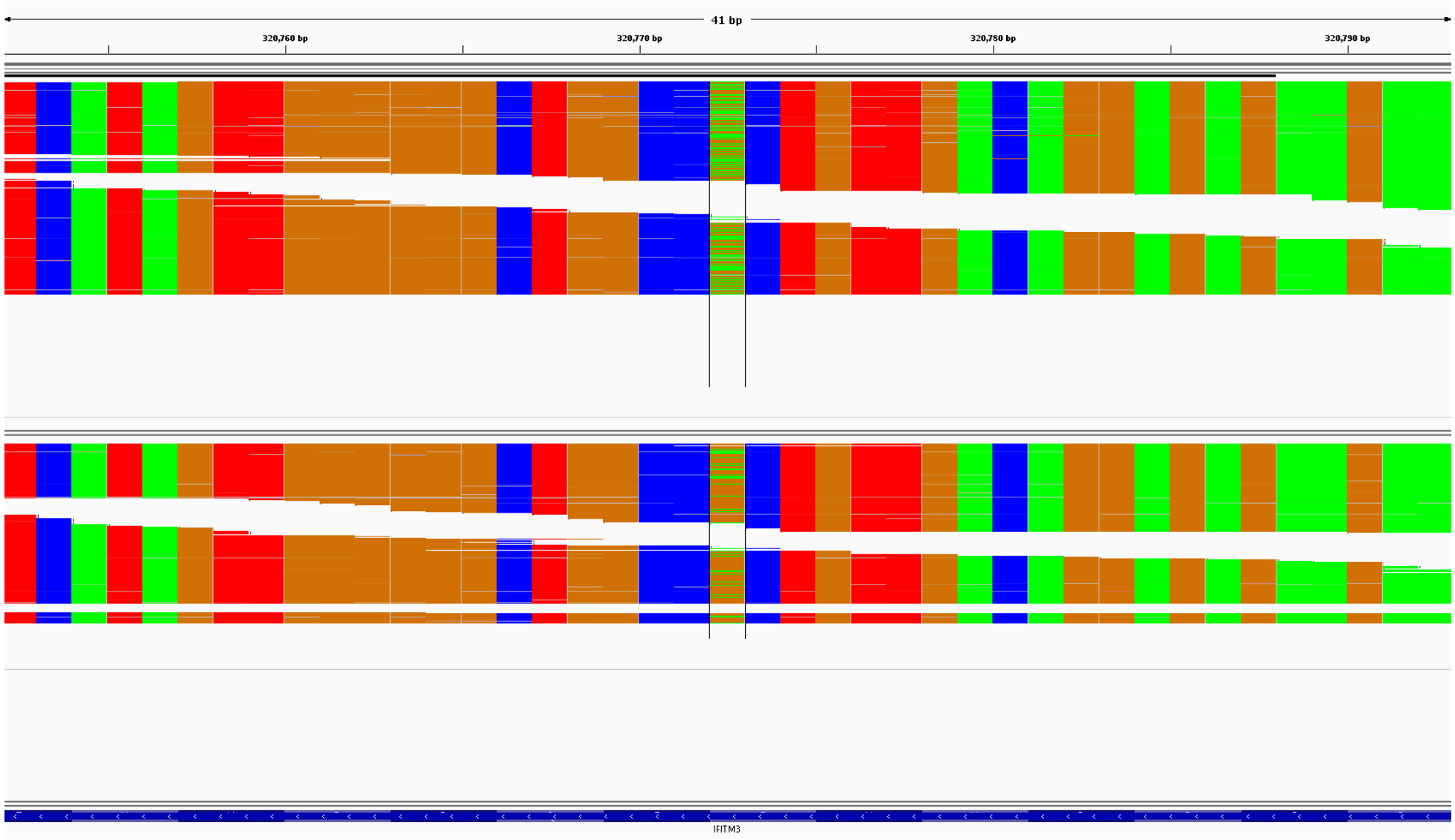
CT Replicate 1



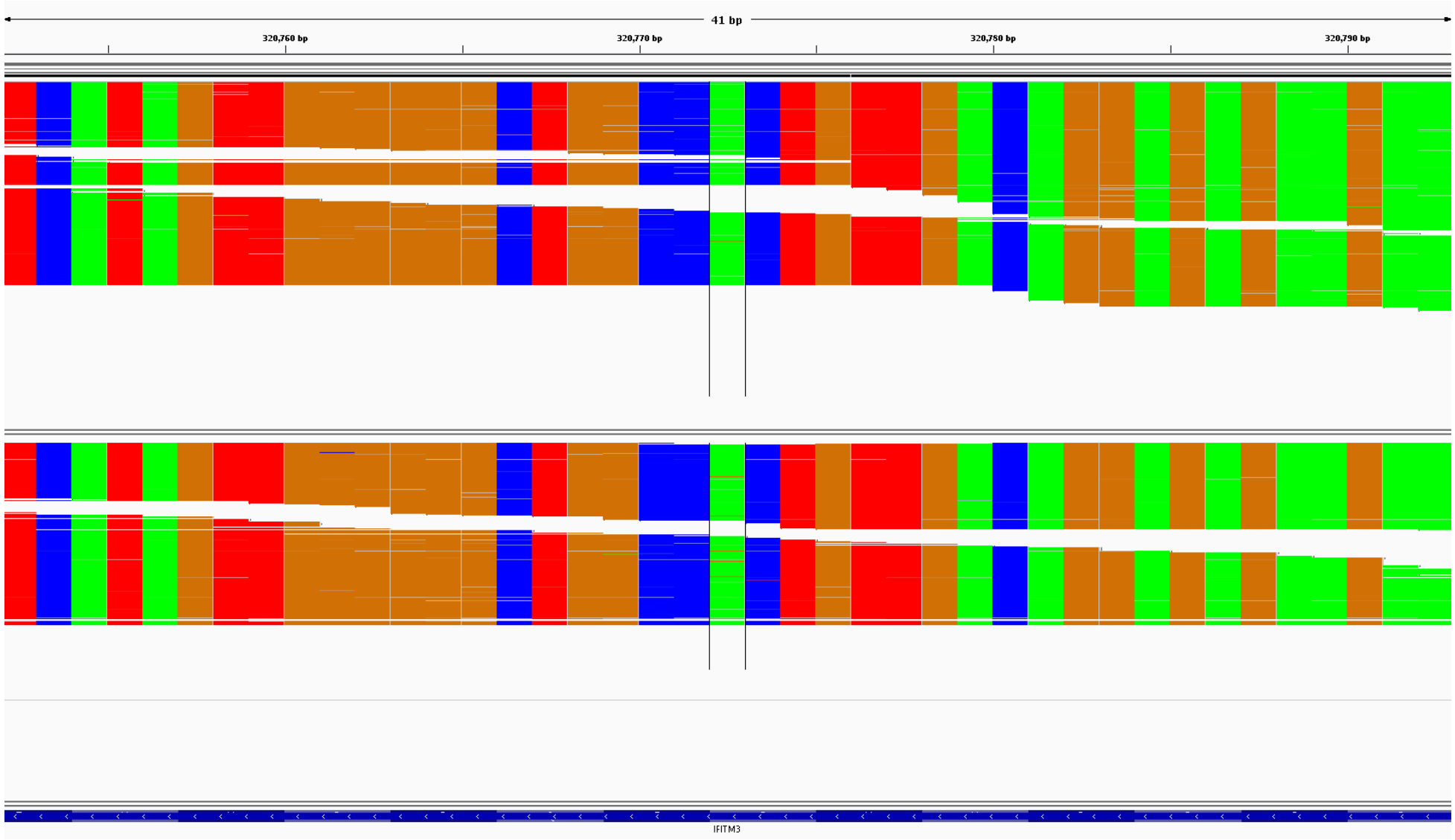
CT Replicate 2 (Later renamed to CC Replicate 4)



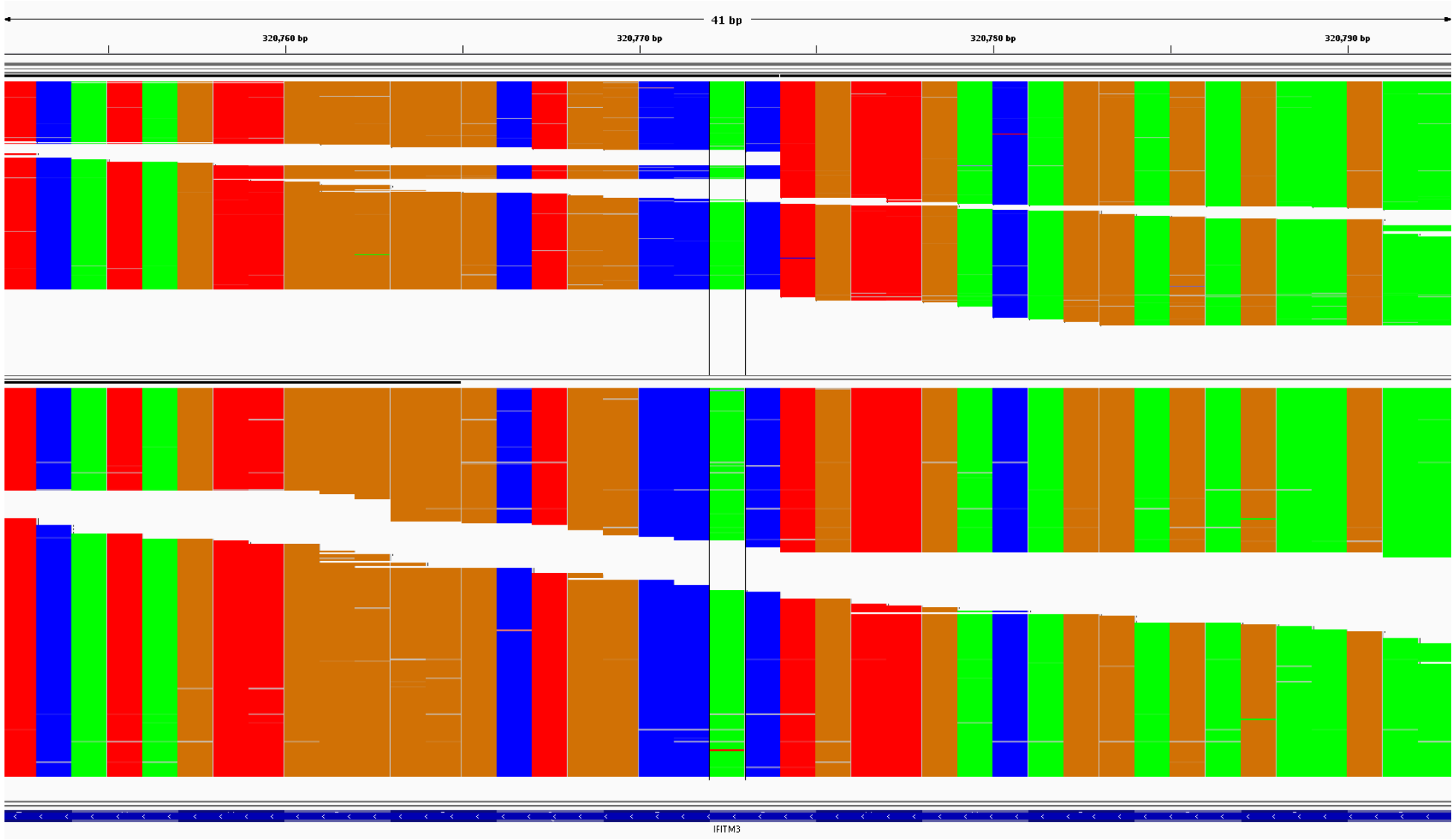
CT Replicate 3



TT Replicate 1



TT Replicate 2



TT Replicate 3

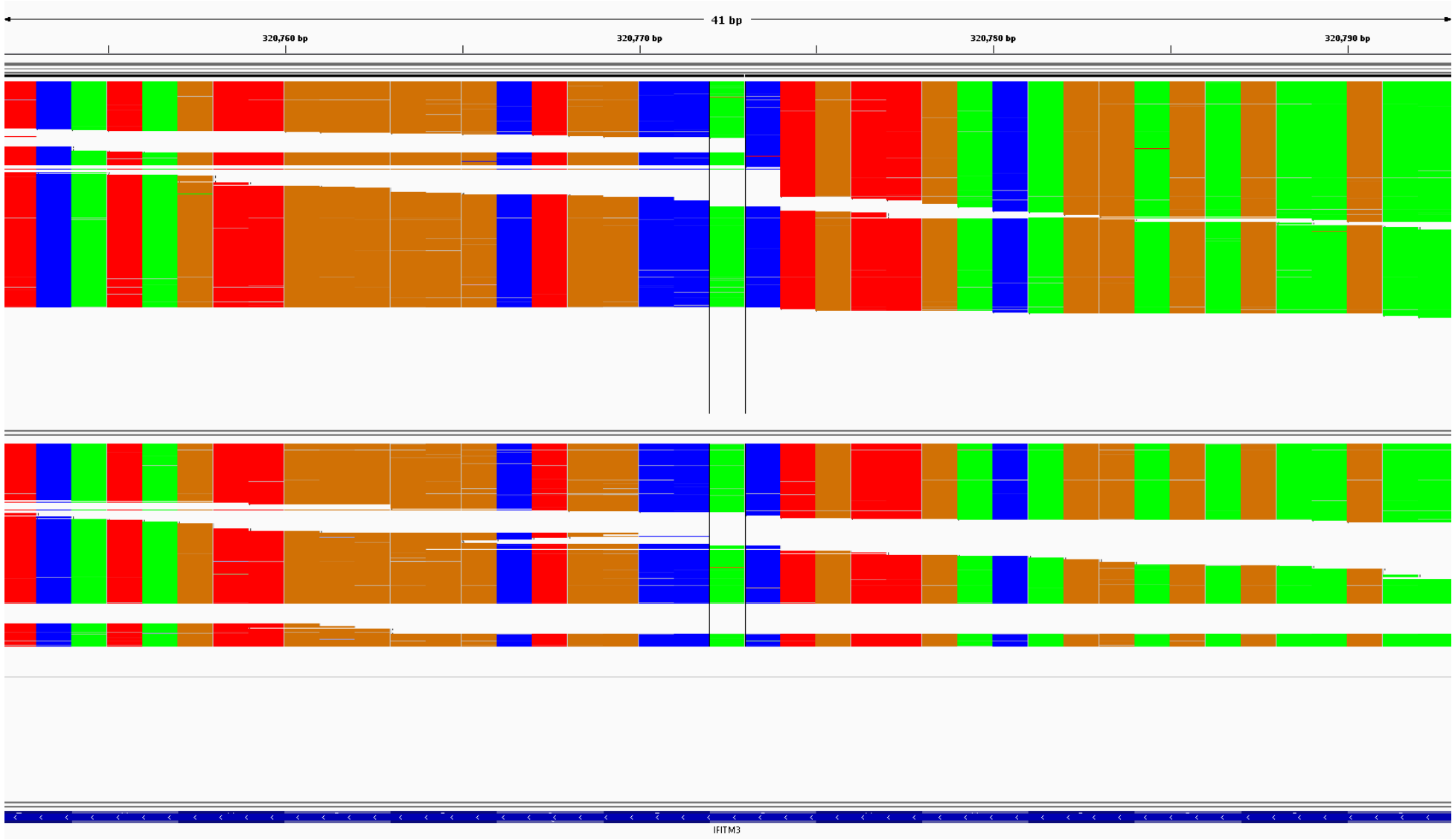


Figure C1. Output from Integrative Genomics Viewer (IGV) for each sample, displaying the reads aligned across the rs12252 SNP. Each output is centred on the rs12252 SNP, as indicated by two vertical black lines. Reads are displayed in a compacted view, with each horizontal line corresponding to a single aligned read. Colours of these reflect the base pairs of the reads (green=A, brown=G, red=T, blue=C). As *IFITM3* is found on the reverse strand, these colours are inverted for these visualizations (green=T allele, brown=C allele). CT replicate 2 was later found to be a homozygous CC replicate, as can be seen in this figure.

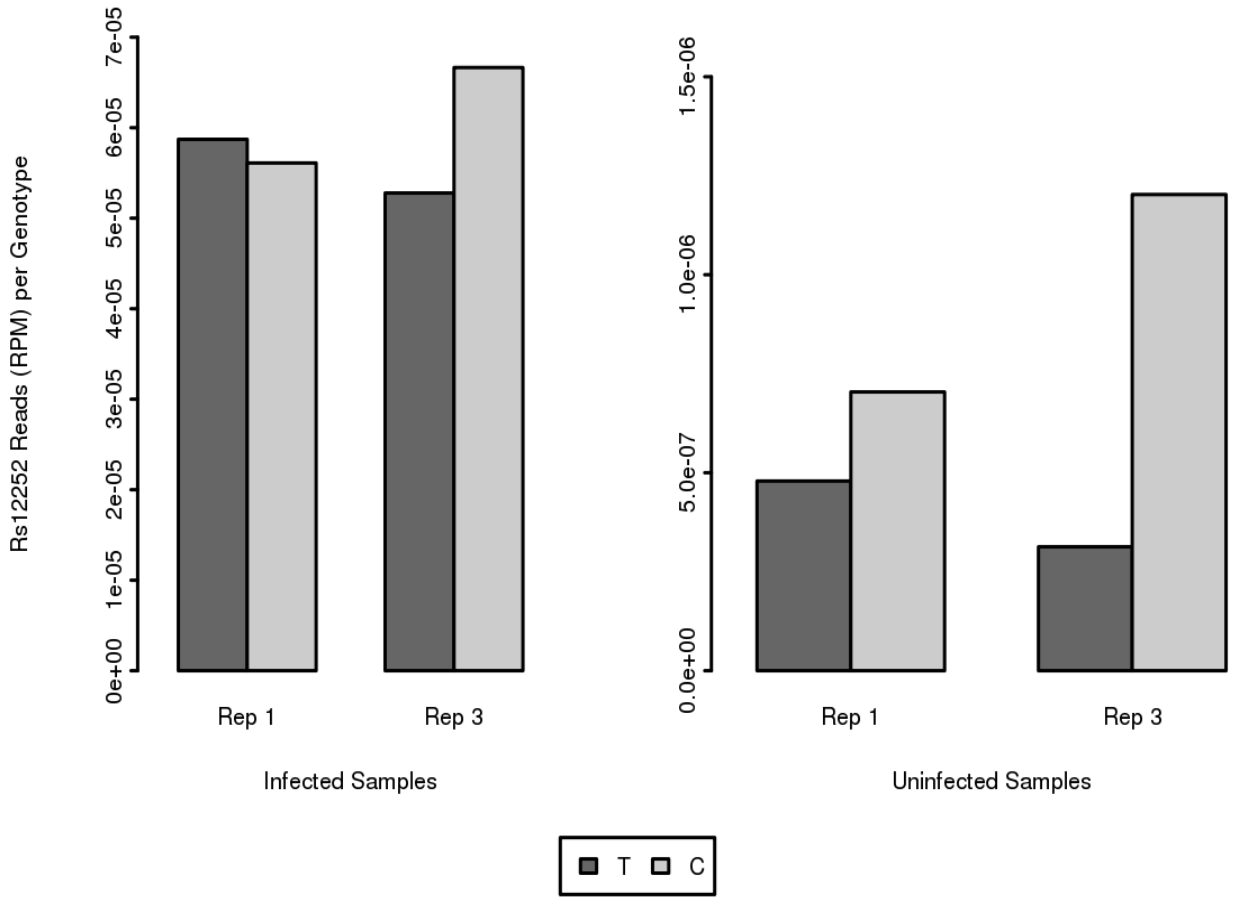


Figure C2. Quantification of reads aligned to the rs12252 SNP in the heterozygote replicates, displaying counts of reads of each allele (corresponding to Figure 9B, displaying the same data as percentages). This shows a reduction in rs12252-C reads in the infected samples. Counting of reads of each variant was implemented following duplicate removal, so as to avoid the possibility of PCR duplicates affecting the proportions.

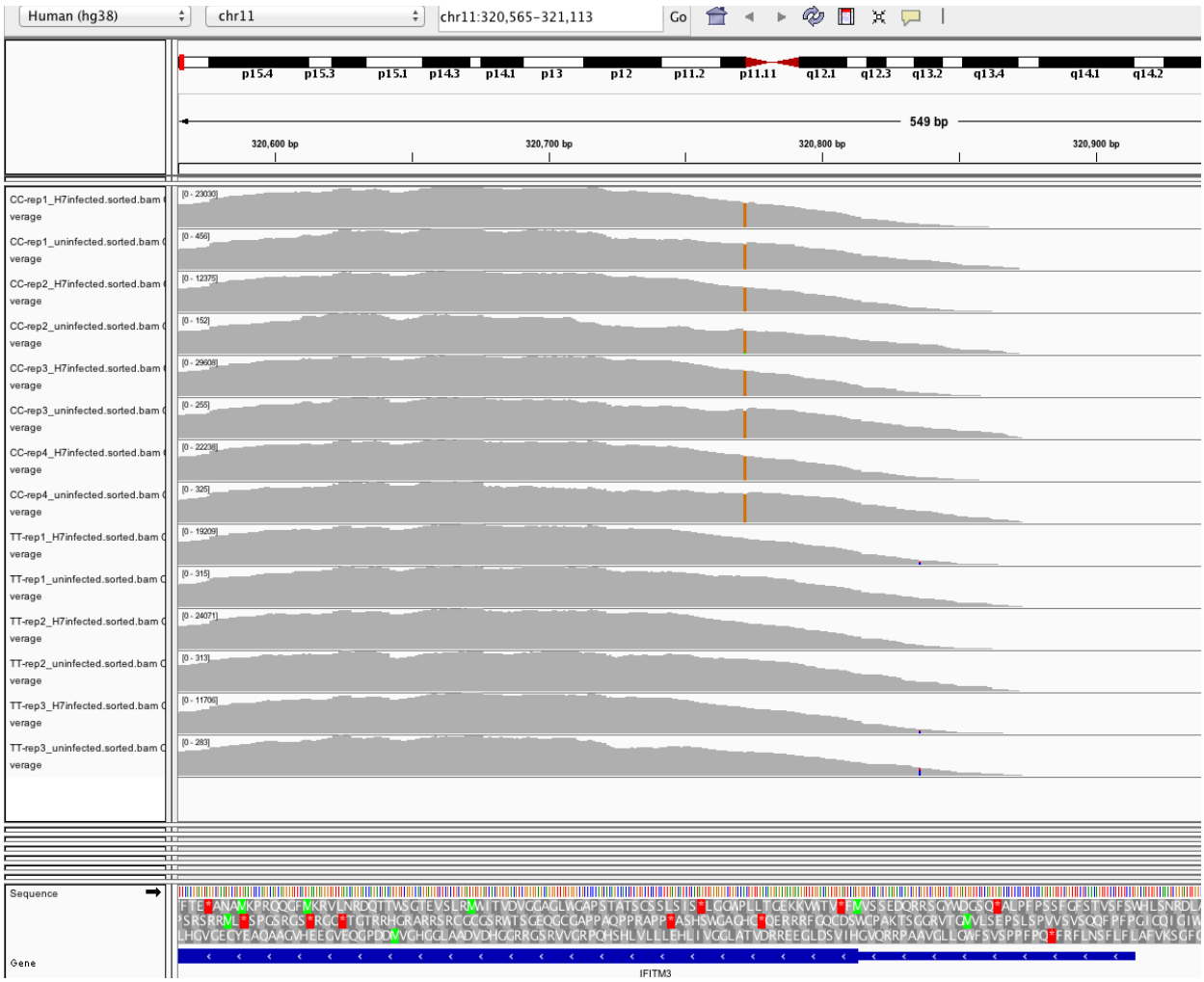


Figure C3. Output from the Integrative Genomics Viewer (IGV) displaying the read coverage of all CC and TT genotype samples across the *IFITM3* exon containing the SNP rs12252. The rs12252-C variant can be observed in the CC genotype in orange at position 11:320772. Read coverage across this exon does not appear to be substantially changed with the CC-genotype, which is associated with a truncated IFITM3 protein and influenza susceptibility.