Automated, highly scalable RNA-seq analysis

by

Rory Kirchner

B.S., Rochester Institute of Technology (1999)

Submitted to the Department of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Health Sciences and Technology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

Signature redacted

Author . . . . . .
Department of Health Sciences and Technology
September 1, 2015

Signature redacted

Certified by . . . . . .
Martha Constantine-Paton
Professor of Brain and Cognitive Science
Thesis Supervisor

Signature redacted

Accepted by . . . . . .
Emery N. Brown
Director, Harvard-MIT Program in Health Sciences and Technology
Professor of Computational Neuroscience and Health Sciences and Technology

Automated, highly scalable RNA-seq analysis

by

Rory Kirchner

Submitted to the Department of Health Sciences and Technology
on September 1, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Health Sciences and Technology

Abstract

RNA-sequencing is a sensitive method for inferring gene expression and provides additional information regarding splice variants, polymorphisms and novel genes and isoforms. Using this extra information greatly increases the complexity of an analysis and prevents novice investigators from analyzing their own data. The first chapter of this work introduces a solution to this issue. It describes a community-curated, scalable RNA-seq analysis framework for performing differential transcriptome expression, transcriptome assembly, variant and RNA-editing calling. It handles the entire stack of an analysis, from downloading and installing hundreds of tools, libraries and genomes to running an analysis that is able to be scaled to handle thousands of samples simultaneously. It can be run on a local machine, any high performance cluster or on the cloud and new tools can be plugged in at will. The second chapter of this work uses this software to examine transcriptome changes in the cortex of a mouse model of tuberous sclerosis with a neuron-specific knockout of Tsc1. We show that upregulation of the serotonin receptor Htr2c causes aberrant calcium spiking in the Tsc1 knockout mouse, and implicate it as a novel therapeutic target for tuberous sclerosis. The third chapter of this work investigates transcriptome regulation in the superior colliculus with prolonged eye closure. We show that while the colliculus undergoes long term anatomical changes with light deprivation, the gene expression in the colliculus is unchanged, barring a module of genes involved in energy production. We use the gene expression data to resolve a long-standing debate regarding the expression of dopamine receptors in the superior colliculus and found a striking segregation of the Drd1 and Drd2 dopamine receptors into distinct functional zones.

Thesis Supervisor: Martha Constantine-Paton

Title: Professor of Brain and Cognitive Science

Acknowledgments

Thank you to the members of the Constantine-Paton Lab, Robert and Ellen Kirchner, my parents, Dr. Melcher for her unwavering support and to the Imaginary Bridges Group.

Contents

1 Automated RNA-seq analysis **11**

1.1 Background 12

1.1.1 RNA-seq 12

1.1.2 RNA-seq processing challenges 13

1.1.3 RNA-seq analysis pipelines 15

1.1.4 Scalability 17

1.1.5 Reproducibility 23

1.1.6 Configuration 24

1.1.7 Quantifiable 25

1.1.8 RNA-seq implementation 31

1.1.9 Variant calling and RNA editing 45

1.2 Discussion 47

1.3 Future development 48

2 Transcriptome defects in a mouse model of tuberous sclerosis **50**

2.1 Background 50

2.1.1 Tuberous sclerosis 50

2.1.2 Tuberous sclerosis is a tractible autism model 59

2.1.3 The second order effects of mTor activation are important . . 61
2.2 Methods . . . 63
2.2.1 Cortex collection . . . 63
2.2.2 Library preparation and sequencing . . . 64
2.2.3 Informatics analysis . . . 65
2.2.4 Calcium Imaging . . . 67
2.3 Results . . . 68
2.3.1 Differential expression . . . 68
2.3.2 Pathway analysis . . . 75
2.3.3 Isoform differential expression . . . 75
2.3.4 Disease association . . . 76
2.3.5 RNA editing . . . 80
2.3.6 Syn1-Tsc1$^{-/-}$ mice have aberrant calcium signaling . . . 81
2.3.7 Htr2c blocker halts aberrant calcium signaling . . . 83
2.4 Discussion . . . 83

3 Transcriptome independent retained plasticity of the corticocollicular projection of the mouse 87

3.1 Background . . . 87
3.1.1 Superior colliculus . . . 87
3.1.2 Retinotopographic map formation in the colliculus . . . 91
3.2 Methods . . . 100
3.2.1 Eye closure manipulation and colliculus dissection . . . 100
3.2.2 Total RNA isolation and quality control . . . 101
3.2.3 cDNA Library creation and sequencing . . . 104
3.2.4 Informatics analysis . . . 105

3.2.5 Differential expression 105
3.2.6 Corticollicular projection mapping and quantitation 106
3.3 Results 107
3.3.1 Corticocollicular projection remodelling 107
3.3.2 Sequencing quality control 108
3.3.3 Differential gene expression 108
3.3.4 Possible X-linked cofactors in LHON 113
3.3.5 Differential exon expression 114
3.3.6 Dopamine receptor expression in the superior colliculus 114
3.4 Discussion 116

List of Figures

1-1 collectl benchmarks of hourly memory, disk and CPU usage during a RNA-seq analysis. 19

1-2 Schematic of parallelization abstraction provided by **ipython-cluster-helper**. 20

1-3 The Poisson distribution is overdispersed for RNA-seq count data. . . 30

1-4 RNA-seq differential expression concordance calculation from two simulated experiments. 32

1-5 Schematic of RNA-seq analysis. 34

1-6 Tuning adapter trimming with cutadapt 36

1-7 Gene level quantification can introduce errors. 39

1-8 RNA-seq improves the rat transcriptome annotation. 44

2-1 CNS manifestations of tuberous sclerosis. 51

2-2 The mTor pathway is affected in many disorders with autism as a phenotype. 54

2-3 The laminar structure of cortex in Syn1-Tsc1$^{-/-}$ mice is undisturbed. 58

2-4 MA-plot of gene expression in the cortex of wild type vs. Syn1-Tsc1$^{-/-}$ littermates. 69

2-5 Culurued Syn1-Tsc1$^{-/-}$ neurons have synchronized calcium flux. . . . 82
2-6 Blocking the Htr2c receptor halts aberrant calcium spiking in Syn1-Tsc1$^{-/-}$ neurons. 84

3-1 Cell types of the superior colliculus 90
3-2 Eye opening refines the corticollicular projection. 99
3-3 Schematic of superior colliculus dissection. 102
3-4 Effect of eye closure state on corticocollicular arbor development. . . 109
3-5 Trimmed mean of M-values (TMM) normalization effectively normalizes gene expression. 110
3-6 Clustering of gene expression in the rat colliculus. 111
3-7 MA-plot (mean expression vs $\log_2$ fold change) between eyes open and closed rats. 112
3-8 MA-plot of differential exon analysis. 115
3-9 Dopamine receptor subtypes are segregated into distinct layers of the superior colliculus. 117

List of Tables

1.1 Cleaning the raw Cufflinks reference-guided transcriptome assembly improves the false positive rate. 43

2.1 Differentially expressed genes in Syn1-Tsc1$^{-/-}$ mice. 75
2.2 KEGG pathway analysis of differentially regulated genes 76
2.3 Gene Ontology (GO) term analysis of differentially regulated genes. . 77
2.4 mTor signaling is differentially regulated in the Syn1-Tsc$^{-/-}$ mouse. . 78
2.5 Autism, intellectual disability and epilepsy genes differentially regulated in the Syn1-Tsc1$^{-/-}$ mouse. 79
2.6 RNA editing events found in Syn1-Tsc$^{-/-}$ and WT mice. 80

3.1 Power estimation of the eyes open vs. eyes closed experiment at a range of fold changes. 106
3.2 Differentially expressed genes with FDR < 0.05 113

Chapter 1

Automated RNA-seq analysis

RNA-sequencing (RNA-seq) has largely supplanted array-based methods for inferring differential expression at the gene, isoform and event level as RNA-seq is more sensitive and produces more information than array-based methods at a similar cost. However processing RNA-seq data and making use of the extra information that comes from having sequence data is complex and experiment dependent. Furthermore what is considered best practice is constantly changing which presents a researcher with many choices of how to perform each step of an analysis. Analyzing a large RNA-seq dataset requires access to a high performance computing cluster, either locally or via the cloud. Even a moderately sized RNA-seq experiment can be hundreds of gigabytes in size and if care is not taken with the choices of tools, they may not operate properly on data of this scale. This chapter describes the RNA-seq module implemented for this thesis as part of the **bcbio-nextgen** project, a

community-developed, highly scalable and easily installable set of analyses of whole genome and exome variant calling and RNA-seq data. This software is in use in academic research labs, core facilities and pharmaceutical companies all over the world, is downloaded thousands of times every month and has processed thousands of RNA-seq samples for hundreds of researchers all over the world[37].

1.1 Background

1.1.1 RNA-seq

RNA-sequencing has largely supplanted array-based methods for quantifying gene-expression. Array based methods estimate gene expression by attaching small antisense probe sequences designed to bind a subset of the transcriptome to a slide. Query sequences are fragmented, labelled with a fluoroescent dye and hybridized to the array. The luminance at each spot on the array is measured and used as a proxy for the expression of the gene the probe is designed to target, assuming a one to one relationship between the probe and the target gene[125]. This approach has some limitations. The first limitation is that only known genes can be assayed, because the probe sequences must be designed from known transcripts. The second is that because the quantitation is based on luminance, the dynamic range of the measurements is constrained to just a few orders of magnitude, as at one end the luminance falls below the level of noise and at the other end the camera is saturated. The third limitation is that, for the most part, only gene expression can be assayed using this approach. RNA-seq removes some of the limitations of array analysis. RNA-seq assays gene expression via the sequencing of small fragments of cDNA up to hundreds of bases in size called reads, aligning the reads to the genome of interest

and estimating the overall expression by counting the number of reads aligning to each gene[186]. Since alignment to the genome is not dependent on the state of the transcriptome annotation it is theoretically possible to infer new genes and isoforms strictly from alignments of the reads to the genome. As new genome builds are created and new annotations layered on top of them old data can be further mined for new information by realigning to the new genome build and reanalyzing the gene expression using the updated annotation. Since quantitation of gene expression is based on counting the number of reads aligning to a gene, the dynamic range is limited by only how many reads are sequenced, lending a much higher sensitivity at the low and high end of the expression spectrum over array based methods. Finally, by examining differences in the sequence of the reads aligning to the genome, it is possible to call variants and RNA-editing events from RNA-seq data. The increased power does not come for free, however. Making use of this extra information makes the analysis of RNA-seq data much more complex and places more computational demands on researchers.

1.1.2 RNA-seq processing challenges

There are several major challenges when handling RNA-seq data. RNA-seq data is very heterogeneous, there are many choices at each stage of performing a RNA-seq experiment, starting at choices of how to extract and purify and fragment the RNA, the type of library preparation performed and the type of sequencing to be performed. Sequencing can be from one end of the RNA fragment (single) or both ends (paired), with a range of lengths of reads and insert sizes of the RNA fragment between the pairs. In addition, RNA-seq data can be from one of several strand-specific protocols where information regarding the strand a gene is on is preserved during sequencing,

which affects the downstream analysis. Finally there are several types of sequencers and each type of sequencer produces reads with varying characteristics and error profiles, all of which need to be taken into account during an analysis. As a result of these complexities it is difficult to have a simple one-size fits all analysis pipeline that can handle a wide variety of experiments. Thus, RNA-seq analysis in published papers tends to be experiment-specific and difficult to reproduce. We have solved this problem by implementing a flexible analysis pipeline that can handle many different types of RNA-seq experiments.

In addition to being very heterogeneous, RNA-seq data can be very large and computationally complex which adds to the difficulty in analysis. A single lane of data can be tens of gigabytes in size and a large scale RNA-seq experiment can encompass hundreds of lanes or more. Memory and CPU requirements for most programs mean most analyses will not be able to be run on common lab computers and will have to be run on either large-scale shared cluster compute environments or cloud-based compute environments. These compute environments themselves are very heterogeneous, with differing storage backends, shared and non-shared filesystems, ways to distribute jobs to a cluster via cluster schedulers, heterogeneous CPU and memory availability on nodes, operating systems and other complexities relating to the hardware the analysis is run on. Not only must an analysis pipline be able to run in varied compute environments, it has to be able to process data across a wide range of scales from single samples to thousands of samples. We have solved these problems in **bcbio-nextgen** by abstracting the compute environment away from the analysis with a small library called **ipython-cluster-helper**. The **ipython-cluster-helper** library has been used in several other projects to provide scaling across all available high performance computing schedulers[129][128].

Another area of challenge is the large choice of tools for performing each step

in an analysis. There are a myriad of programs available to perform every step in an analysis, from quality control of the raw reads, aligning[54], assembly[58][163] and quantification[46] of transcripts, differential expression[144] and final quality control[45][62] of the results. Often times it is not clear what the tradeoffs are when choosing one tool over another, and some tools may not function properly with large datasets or with data from a particular sequencer or type of experiment. Each tool for each step in an analysis is configurable and determining how to properly configure each tool for a specific analysis is complex, time consuming and prone to error. For some tools it can be difficult to even install the tool, especially on shared compute environments where necessary libraries may not be installed or administrator privileges might be needed to install some programs. In addition each tool may need special data files in a wide variety of formats to work properly and producing these data files from the known genome and transcriptome can be a challenging task.

All of these challenges make it extremely difficult for a researcher without programming experience and without a deep knowledge of UNIX to even get started analyzing their own data. In addition, all of these challenges make performing analyses in a reproducible way very difficult; at the end of an experiment it is often impossible to trace what was run on each sample much less reproduce the analysis using the same tools with the same data.

1.1.3 RNA-seq analysis pipelines

As of July, 2015 there are 200,000 Google results for 'RNA-seq analysis pipeline' but very few of those are under active development and have kept up with the state of the art and none of them install everything you need to run a RNA-seq analysis. A poll on the RNA-seq blog in July 13th, 2013 asked 'What is the greatest immediate

need facing the RNA-seq community?' with $\sim 70\,\%$ of the responses being either a need for a standardized data analysis pipeline or more skilled bioinformatics experts to process the data. The rate of generation of sequencing data is far outpacing the generation of skilled bioinformaticians so it will be important to have an accurate, flexible, validated RNA-seq pipeline that can be run by a naive researcher in order to keep up with the rapidly increasing amount of generated sequencing data. The need for this is reflected in industry; in December, 2014 Bina, a company that has implemented a variant calling and simple RNA-seq analysis pipeline was bought by Roche and a myriad of startups have sprung up, promising to handle the analysis of second generation sequencing data either on local appliances or the cloud. It is important for a free, community developed, open source alternative to these industry pipelines to exist because many academic labs will not have funding to run on these commercial appliances.

It is important to specify what characteristics a standardized, scalable and reproducible RNA-seq analysis pipeline should have. The first is that it should be configurable at a high level of abstraction. A proper pipeline should translate a small set of high level options to the appropriate low-level settings for each tool in a pipeline. These exposed options should be a minimal set to accurately run an analysis, with as many of the low-level options as possible derived from the data. A researcher should not have to be familiar with the intricacies of each tool in order to run an analysis. The second characteristic of a standardized RNA-seq analysis pipeline is it should be reproducible. A full accounting of the versions of all third party tools that were used along with all of the commands that were run should be provided at the end of an analysis so the analysis can be reproduced. Ideally a Docker container with all of the relevant software installed would be distributed along with the raw data so any researcher could reproduce the analysis. The third

important property of the pipeline is that it should be scalable. The pipeline should be able to be run on small pilot experiments of a couple of samples and be able to be scaled up to a full experiment with thousands of samples, assuming there is sufficient compute to run the analysis. Running on the cloud is a requirement for a generally useful analysis pipeline, because a research can scale up or down their compute depending on the size of their analysis. The fourth important property is that the analysis pipeline should be open and flexible. The pipeline should not be a black box, and it should be able to be changed easily to keep up with new innovations. When choices are made for each step in an analysis, the reasoning should be laid out in a document so a researcher can understand why the choices were made. The fifth important property is that the pipeline should be validated. There should be a set or sets of independent benchmark data against which changes to the pipeline are tested, to catch regressions in new versions of the pipeline and to measure what is an optimal way to process a dataset. Finally the pipeline should produce output data in an easily understandable, parseable format in the form of a report and well structured output data for dissemination with downstream informaticians and bench researchers.

In this work we implemented three tools **bcbio-nextgen**, **bcbio.rnaseq** and **ipython-cluster-helper** which together produce a RNA-seq pipeline with each of these characteristics. It has been used to process thousands of samples in dozens of laboratories all over the world.

1.1.4 Scalability

An analysis is scalable if it can run across a broad range of sizes of experiments, from a small pilot experiment to experiments with thousands of samples. The issue of scaling

up is a complex one involving many areas of possible difficulty. At the most basic level, many third party tools that work on small experiments fail when confronted with a large number of samples or extremely large datasets. We solve this issue by either providing tools known to scale well or providing fixes upstream to existing tools to enhance scalability. An example of the former is replacing the commonly used **samtools** suite of tools used for manipulating BAM files with a high performance alternative sambamba where possible. An example of the latter is providing parallel implementations of previously existing tools such as GEMINI[128]. Other areas where scalability can break down is in overwhelming the I/O of the machine; we choose algorithms which minimize the amount of I/O they perform and stream as much as possible between steps in the analysis to improve performance. Members of the **bcbio-nextgen** community have implemented[36] support for collection of performance data in terms of disk I/O and memory, network and CPU usage at each stage in the analysis using **collectl**[154]. Figure 1-1 on the following page shows an example of the usefulness of collecting these metrics when writing a scalable analysis.

We provide scalability across compute architectures by using the **IPython.parallel**[133] framework for parallel and distributed computing. We developed the **ipython-cluster-helper** library to provide a set of abstractions on top of cluster schedulers much like the Distributed Resource Management Application API (DRMMA). DRMMA provides a unified API for submitting jobs to an array of cluster schedulers but has the limitation that you must have a scheduling system installed on the compute environment. **ipython-cluster-helper** expands the possible compute configurations by allowing machines to work as an ad-hoc cluster with no cluster scheduler, as long as the machines can be accessed via secure shell (SSH). **ipython-cluster-helper** provides a view to a cluster of machines which uses a simple unified interface to distribute jobs to the cluster of machines. This allows our pipeline to distribute jobs

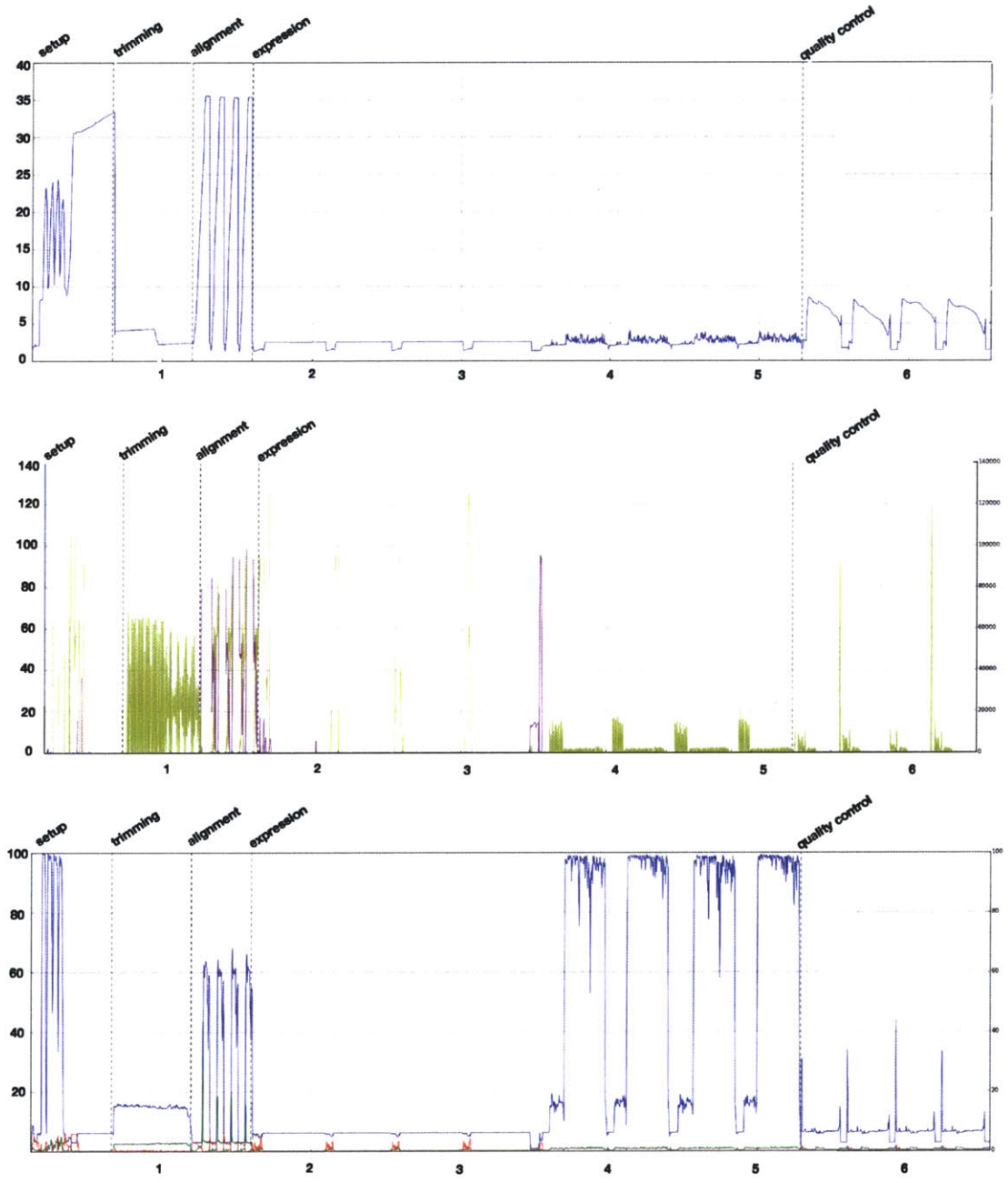


Figure 1-1: collectl benchmarks of memory, disk and CPU usage during a RNA-seq analysis. **top)** memory (in gigabytes), **middle)** disk (in kilosectors/second, yellow is reads, purple is writes) and **bottom)** CPU (in percent CPU usage). Having benchmark statistics for each step in the analysis helps when optimizing performance. For example, the period of low CPU usage in this run during **expression** was due to DEXseq running serially; fixing this and running DEXseq in parallel cut the time to estimate transcript expression in half. Similarly the disk intensive portions during **trimming** were eliminated by moving to streaming between the trimming steps, (see Figure 1-6 on page 36).

across a wide array of compute architectures; running on a laptop in local mode or running across a cluster with thousands of computers both use the same interface to distribute work. Figure 1-2 shows a schematic of the **IPython.parallel** architecture that **ipython-cluster-helper** uses for performing parallel computations. An IPython cluster is spun up consisting of a controller process and a set of IPython engines running on arbitrary nodes of a compute architecture. The controller acts as a scheduler, distributing jobs to be run on worker engines. Communication between the controller and the engines happens over ZeroMQ message queues; the controller places a job to be run in the form of a serialized Python function onto a ZeroMQ queue and the worker unserializes the function, runs it locally, saves the result and places a serialized version of the result back on the message queue.

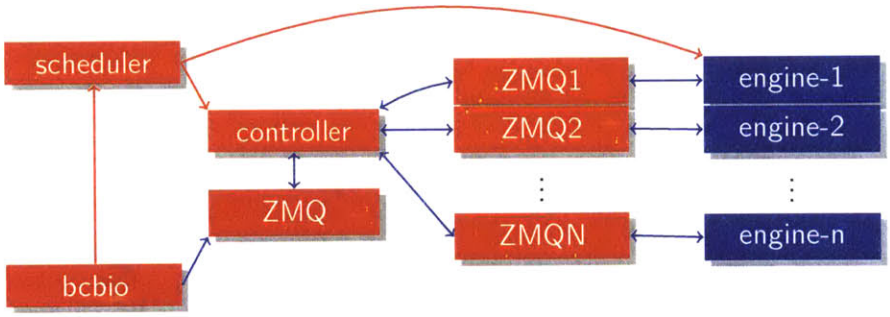


Figure 1-2: Schematic of parallelization abstraction provided by **ipython-cluster-helper**. For each step in the analysis an **IPython.parallel** cluster is set up using **ipython-cluster-helper**, consisting of a controller process which distributes the computation and a set of worker processes called engines which perform the computation. The controller is controlled via a client process, in this schematic bcbio. Communication between the client, the controller and the engines occurs over ZeroMQ message queues and require only that a port be open for two-way commmunication between the components of the **IPython.parallel** cluster. This allows for the compute layer to be abstracted away so that **bcbio-nextgen** can work with a simple interface to the compute.

ipython-cluster-helper is not limited to use in **bcbio-nextgen**, it can be used to run any Python program in parallel across all cluster schedulers. Below is an

example Python code listing showing ease of taking a simple non-parallel program and parallelizing it using **ipython-cluster-helper**. This library has been used in other projects that required scaling across multiple machines[128].

Listing 1.1: Parallelizing a serial function with ipython-cluster-helper.

```
from cluster_helper.cluster import cluster_view
from yourmodule import long_running_function
import sys

if __name__ == "__main__":
    # serial version
    for f in sys.argv[1:]:
        long_running_function(f)
    # parallel version with ipython-cluster-helper
    with cluster_view(scheduler="lsf", queue="hsph", num_jobs=5) as view:
        view.map(long_running_function, sys.argv[1:])
```

All of the scalability in the world is useless if the analysis infrastructure is not able to be relocated to the compute. Thus having a highly scalable pipeline also requires having a pipeline that is easily relocatable and that means having an installation process which installs everything that is necessary for an analysis. This includes all of the necessary data including the genome sequences and metadata about the genome sequences including annotation of gene boundaries and other genomic features of interest. It also involves installing the correct versions of hundreds of tools, libraries and other programs to run the analysis. These tools have to be installed whether or not the user has administrative access to the machine as many researchers using shared compute environments will not have administrative access. We handle

installation by contributing back to two major projects for installing bioinformatics tools, the cloudbiolinux[2] project and the Homebrew project. Both projects provide simple to specify formula for getting and installing tools on UNIX and MacOS machines. **bcbio-nextgen** uses these two projects to compile and install the tools necessary to run an analysis.

The combination of abstraction of the underlying compute with **ipython-cluster-helper** and the ability to install all of the necessary tools in an environment-agnostic way allows **bcbio-nextgen** to be able to easily installed and run in a wide variety of compute environments. Recent work on **bcbio-nextgen** by Brad Chapman and John Morrissey, in collaboration with groups at Biogen, AstraZeneca and Intel have implemented two extremely useful scalability features in **bcbio-nextgen**. The first is enabling **bcbio-nextgen** to be run using the container system Docker[8]. Rather than running the installation script to install all of the tools, installation requires just downloading a Docker container with bcbio-nextgen and all of the dependencies already installed. The second scalability feature is a set of tools to use the Docker images to run an analysis on Amazon Web Services (AWS), pushing and pulling data from Amazon's storage solution, S3. This is an example of one of the major benefits of working on an open source, community driven project: infrastructure like Docker and AWS support only needs to be implemented once for all analyses to take advantage of it. Since the analysis is abstracted away from the compute architecture it also means rerunning an old analysis can be done on more modern hardware or by other researchers to reproduce an old result.

1.1.5 Reproducibility

High throughput genomics has come under scrutiny in the past couple of years for having issues with reproducibility. Here, our definition of reproducibility is broad; we do not mean reproducing the exact numerical result of an analysis but reproducing the major finding of a paper. Even with this relaxed definition, many genomics experiments cannot be reproduced. This problem is so widespread that the NIH has undertaken an initiative to enhance reproducibility in genomics[40]. One pressing issue is there are often no standards for experimental design or analysis[104], and this is especially true for gene expression experiments. The lack of community derived standards results in widespread reproducibility problems with both microarray[50] and RNA-seq experiments[124].

For an informatics analysis to be reproducible, more than just the data has to be made available. Complete recounting of the command lines of all software, the versions of all software, the versions of all intermediate plumbing type code and code generating the downstream summary statistics must be provided. Not only must this software be made available, it must have to work in the compute environment of the person reproducing the analysis. To be reproducible, the code generating the results must be open; a description of the algorithm is not sufficient to reproduce the results of the code[77]. We address this problem of reproducibility in several ways. The first is that we install all third party tools needed to run an analysis on a wide variety of computing platforms. This allows a person attempting to reproduce results to start from the same base environment. The second is that we record all versions of all third party tools and all command lines run so that an analysis can be reproduced by running the commands if necessary. The third is that if an analysis is run via Docker, a Docker container that contains everything necessary to reproduce

an analysis pre-installed which can be run on a local cluster computer or on Amazon. Each version of a container is specified by a unique id which can be accessed later, meaning an analysis can be completely reproduced by any researcher able to use Docket containers at a later time. The Docker and AWS implementation in **bcbio-nextgen** further expands reproducibility, since with that integration a researcher only needs to pay Amazon for the compute and storage and can rerun a published analysis on their own. In these ways, **bcbio-nextgen** provides a solution to the reproducibility problem.

1.1.6 Configuration

RNA-seq analyses are complex in part due to a wide array of choices that one can make regarding how the RNA is extracted, how the libraries are prepared, how the sequencing is performed, what organisms are being studied, if the libraries are stranded or not or aimed at tagging a specific portion of a transcript instead of the entire transcript. Each of these options has an effect on the downstream analysis. For example a stranded experiment must take the strand information into account when aligning, quantifying and assembling the reads, so each tool that is run must have the appropriate options set. We provide a carefully considered set of high level options that describe a RNA-seq experiment that an experimenter must set with many of the low level configuration options for each tool set to appropriate default values or appropriate values learned from the data. In this manner we simplify setting up a RNA-seq analysis down to setting a minimum set of parameters that can drive a wide variety of tools underyling the analysis.

These parameters describe whether the library is strand specific or not, which kit was used to produce the library, and which genome the library was from. Parameters

controlling expensive, optional analyses such as whether or not to call variants or assemble the transcriptome can be enabled if those were part of the design of the experiment. We also handle a difficult type of experiment that is often used in cancer research, where a tumor from one organism is grown in another organism, often a human tumor in mouse tissue. When the tumor is sequenced, some mouse tissue can be included. There is an option to disambiguate reads of two organisms in order to remove this type of contamination. Finally, arbitrary metadata about each sample can be provided, such as which batch it is from, if it is treated or not, anything the user can provide. During downstream analyses, this metadata is automatically made available for model fitting and differential expression calling in a separate RNA-seq reporting tool we created called **bcbio.rnaseq**.

1.1.7 Quantifiable

It is important to have validation datasets for an analysis, both to benchmark the speed of the analysis as well as validate the results of the analysis. Validation datasets allow for an analysis to be fine tuned in terms of new tools or configuration of existing tools to improve results without fear of introducing errors into the data. More importantly it also places the downstream results in a greater experimental context. A validation data set allows a researcher to understand where they are making mistakes or where the analysis is blind and can help guide downstream users of the data towards the most salient results.

Having a standardized, quantifiable analysis allows a researcher to treat their analysis as an optimization problem and improve it. Benchmark data sets serve as type of integration test for the software making up an analysis pipeline. When a new tool is published, it is standard for the work to show a comparison to the standard

tools of the day and show how the new tool outperforms the old in some aspect. Occasionally review papers compare several tools to each other with a validation dataset to determine in a less biased way the strengths and weaknesses of each tool[98][144]. These papers are valuable but often there has been no fine tuning of any of the comparison algorithms. Consequently, they are often run with their default values and are compared against each other, whereas a researcher with experience with a particular tool would be able to tune it to get better results. **bcbio-nextgen** is tool agnostic and tools have gone through a process of fine tuning the configuration parameters already. Tools are chosen based on the consensus of the users regarding which tool works best and the tools are tuned or improved to give the best results. This allows for a fair comparison across tools when evaluating making changes to an analysis pipeline.

In addition to being tool agnostic, **bcbio-nextgen** is also dataset agnostic. If a new benchmark dataset comes out for a particular aspect of a pipeline, the correct values for the benchmark dataset can be added and the benchmark dataset run through the pipeline and compared. In this way researchers can improve both the analysis itself and the measurement of the analysis simultaneously through the incorporation of new benchmark datasets.

Researchers can ask many more questions of their RNA-seq data than with microarray analyses. Expression can be summarized and differential expression called at the gene, transcript, exon and splicing event level. Rearrangements and variants, gene fusion events and RNA-editing events can all be assayed with RNA-seq. Novel genes and novel isoforms can be assembled, and differential usage of promoter sites can be analyzed. Each of these aspects of an analysis have several tools which handle them and each needs a benchmark dataset to test against when testing iterations of a pipeline. **bcbio-nextgen** gathers available benchmark datasets from the commu-

nity to test each aspect of RNA-seq analysis. We provide benchmark datasets in the form of data from actual experiments and simulated data to benchmark gene-level and transcript-level expression calling, transcriptome assembly with and without a reference data set and fusion gene calling on cancer datasets. As the community develops alternate benchmarks we fold them into **bcbio-nextgen**.

Gene and transcript expression

bcbio-nextgen uses a combination of validation datasets from real data and simulation to assess the results of the gene-wise differential expression analysis. The first validation dataset is from phase three of the Sequencing Quality Control (SEQC)[41] project from the US Food and Drug Administration. The goal of releasing these datasets was to provide laboratories with known reference standards to use to tease out the types of technical variation introduced across laboratories, sequencers and protocols. This dataset consists of a two sets of samples: the first is RNA sequenced from the Universal Human Reference RNA (UHRR) from Agilent. This sample is composed of total RNA from ten human cell lines to be used as a reference panel. The second sample is from the Human Brain Reference RNA (HBRR) panel from Ambion which consists of brain samples from all regions in several subjects pooled together to form a reference panel. The SEQC project provides qPCR results from one thousand genes in the UHRR and HBRR panels, to serve as a proxy for a truth dataset. The validation dataset we analyze is fifteen million randomly selected reads from five replicates of each SEQC sample.

The SEQC data have a few major limitations as a validation data set. The first limitation is that since the RNA comes from stock of pooled RNA, the replicates are technical replicates of only the library preparation and sequencing and do not take

into account variability induced by RNA-extraction or biological variability such as samples from different mice or humans. Proper handling of biological variability is important component in a RNA-seq analysis and the SEQC dataset is not an appropriate way to assess it. The second major limitation is that the validation data is gene-level qPCR data. RNA-seq data is capable of assessing expression at the level of the transcript but the SEQC data set will not be useful for assessing quantitation at anything other than the level of the gene.

To address these limitations the we also produce simulated datasets of raw RNA-seq reads using Flux Simulator[67], an in-silico RNA-seq experiment simulator. Using Flux Simulator, reads are generated from a given reference transcriptome and many steps and biases introduced during RNA-seq library preparation are simulated including fragmentation at the RNA or cDNA level, reverse transcription, PCR amplification, gene expression and sequencing. Flux Simulator does not produce biological replicates, so we implemented a simulator to add biological noise and fold change spike ins to the data generated by Flux Simulator to mimic biological replicates and differential expression in a real experiment. This package is available online as **flux-replicates**.

RNA-seq count and read simulator

Flux Simulator outputs a simulated relative expression level, p which is the proportion of the total RNA molecules in the simulated sample that are from a given transcript. These values are used as a baseline level of proportions from which a single factor differential expression experiment is simulated by spiking in fold changes between replicates of a sample. We implemented a RNA-seq read count simulator[94] by setting a proportion of the simulated transcripts to be differentially expressed at

range of fold changes and drawing counts for each gene from the negative binomial distribution with the dispersion parameter set to increase as the expression level decreases, to mimic real data(Figure 1-3 on the following page).

Starting from a given level of biological coefficient variation, BCV, a relative expression level for gene y with proportion of the total expression level p_y and a library size l for a sample, we simulate c_y, the counts of transcript y from the negative binomial distribution by composing a gamma distribution and a poisson distribution to add biological and technical noise, respectively:

$$\mu_y = p_y l \tag{1.1}$$

$$BCV_y = (BCV + \frac{1}{\sqrt{\mu_y}})\sqrt{\chi^2(40)} \tag{1.2}$$

$$c_y \sim T(B(\mu_y, BCV_y)) \tag{1.3}$$

B is a function that draws values from a gamma distribution of mean μ_y and variance ${BCV_y}^2$ and T draws values from a poisson distribution with mean $B(\mu_y, BCV_y)$. We add random noise dependent on expression level of the gene to BCV to simulate the higher dispersion of the negative binomial at low counts (figure 1-3 on the next page).

The simulator can be run in a mode where it is supplied with an existing set of counts from which the parameters for the simulation are estimated to match the existing data. This allows the user to take a previously run experiment and estimate what would happen if they sequenced less reads or changed the number of replicates. It also allows for some rough post-hoc power calculations to inform future experimental design choices and to make stronger negative result claims rather than failing to reject the null hypothesis of no differential expression.

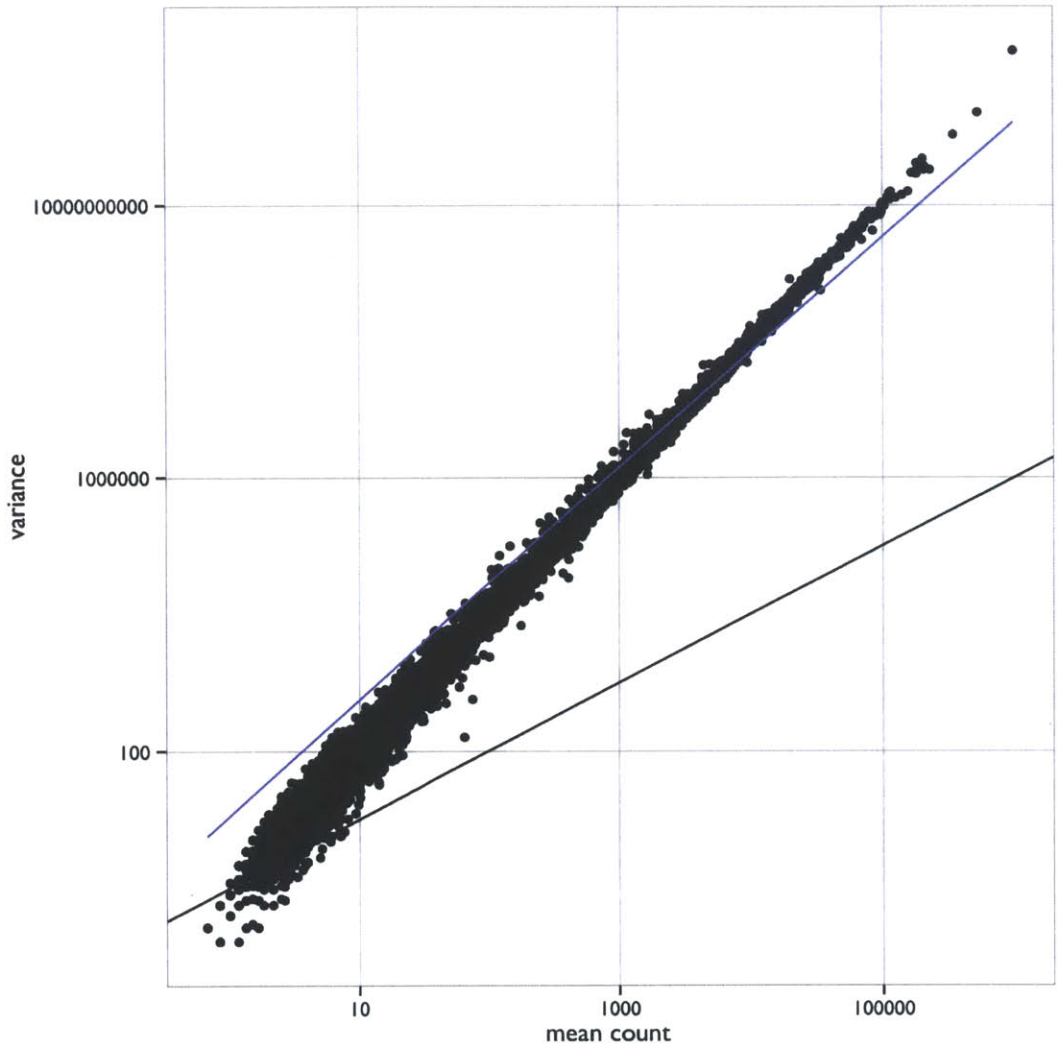


Figure 1-3: The Poisson distribution is overdispersed for RNA-seq count data. Variance vs. mean counts for a RNA-seq experiment (black dots) with variance estimated by the Poisson (black line) and negative binomial (blue line) distributions. Each point on this graph is the mean and variance for a single gene across biological replicates. RNA-seq data is more noisy than expected from the Poisson due to additional technical and biological variability.

As an example, RNA sequencing is very expensive and time consuming and often experimenters are looking to maximize the power of their experiment and minimize the cost. One potential optimization involves determining how many reads to sequence against how many replicates to run. An experiment running at most on two HiSeq lanes can expect at least 300 million reads. This leaves a case-control experiment with three replicates at 50 million reads per sample. Alternatively, rather than sequencing six samples at 50 million reads per sample, the read depth could be cut in half and twice as many replicates could be sequenced. Using the simulator to compare these two experiments against each other, we can make a recommendation to run more replicates at lower depth if the goal is to identify differentially expressed genes at moderate-to-high fold change (Figure 1-4 on the following page). In addition, if the genes of interest are at the low end of the differential expression spectrum, we could recommend to not run the experimental at all.

1.1.8 RNA-seq implementation

Figure 1-5 on page 34 shows a high level overview of the RNA-seq pipeline implemented in **bcbio-nextgen** version 0.8.5a. The implementation covers quality control of the data, adapter sequence removal, aligning to the genomes, quantifying at the gene, isoform and exon level, calling variants and RNA-editing events, transcriptome assembly, classification and filtering and calling differential expression events at the gene, isoform and exon level using several commonly used tools. There are a multiplicity of third party tools that could be used to perform each step and tools are chosen based on considering their accuracy, licensing, scalability in terms of CPU, IO or memory bottlenecks and how well they are actively maintained. When compute constraints may be a roadblock, alternatives are provided; for example the STAR[48]

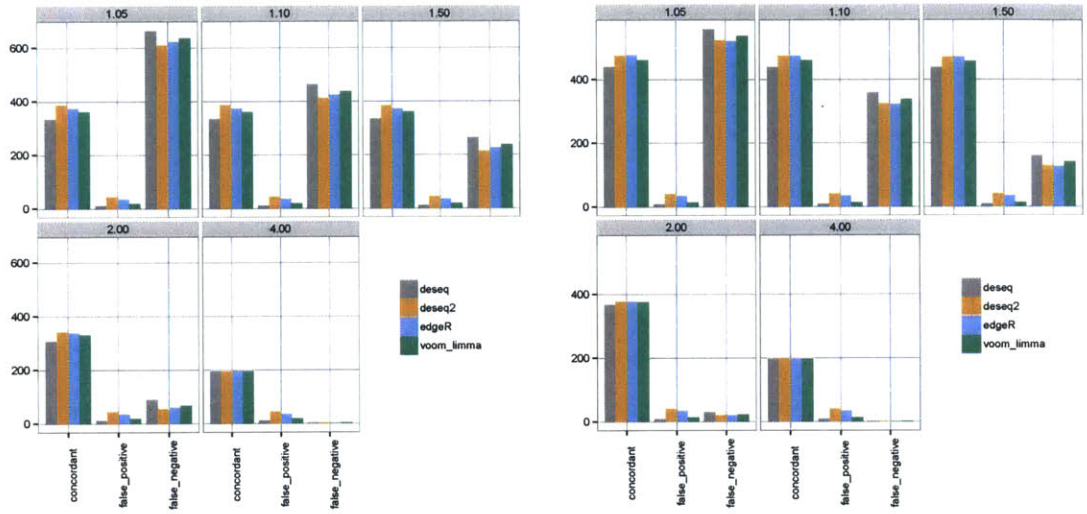


Figure 1-4: RNA-seq differential expression concordance calculation from two simulated experiments. An RNA-seq experiment was simulated with a sample size of three and a library depth of 50 million reads (a) and a sample size of six and a library size of 25 million reads (b). We simulated 200 genes as differentially expressed up or down across each of five fold changes: 1.05, 1.10, 1.50 and 4.0. Each facet of the graph calculates concordance, false positive an false negative rate, considering only genes at the specified fold change or greater. Four commonly-used negative binomial based differential expression callers (DESeq2[101], DESeq[6], edgeR[149] and limma[94]) were run on the data and their concordance with the known fold changes from the spike ins were simulated. At large fold changes most of the spike in genes are correctly called differentially expressed in both experiments but at fold changes below four, the experiment with a higher number of replicates calls more concordant differentially expressed genes.

aligner, while very fast and accurate has a large memory requirement which may make it not feasible to use in some compute environments. In that instance a slower, less memory intensive alternative in Tophat2[170] is offered (see Figure 1-5, alignment box). In the following sections a brief, non-comprehensive rationale behind choosing the tools for each step in the analysis is covered. Where appropriate there is a discussion of optimizations we implemented to improve the accuracy and speed of the chosen tools. This discussion isn't intended to be an exhaustive comparison of every tool for each step or a conclusive statement about which tools are better than others, but more a breakdown of the tradeoffs, possible pitfalls and other important considerations for each step in the analysis. The design philosophy of the bcbio-nextgen RNA-seq pipeline implementation is to use the data to tune as many parameters as possible to optimize for the particular data set being processed while maintaining a balance between accuracy and throughput and the tool choices reflect that philosophy.

Read trimming

RNA-seq reads can have contaminating sequences at the ends of the reads in the form of adapter sequences, poly-A tails or other sequences, which may cause issues during alignment or in downstream analyses, resulting in loss of information. Some spliced-read aligners such as STAR will handle these reads by soft clipping the homopolymer or adapter sequences that don't align to the genome but other aligners such as Tophat do not handle these reads well. For RNA-seq, adapter contamination on the ends of reads can often be a problem; under most library preparation protocols the RNA is fragmented before conversion to cDNA and there may be small pieces of RNA for which the read length is longer than the fragment. For these small RNAs with long

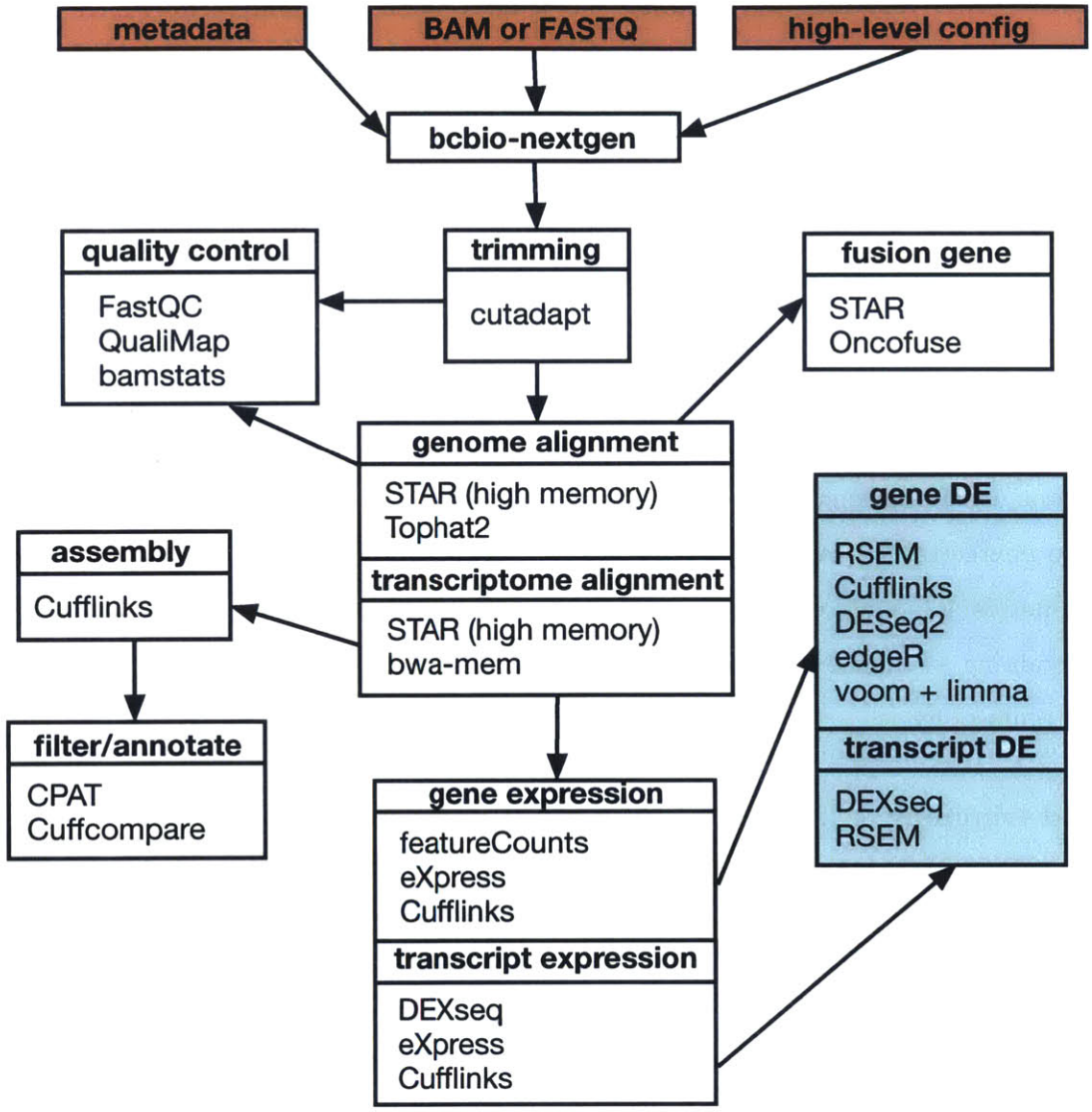
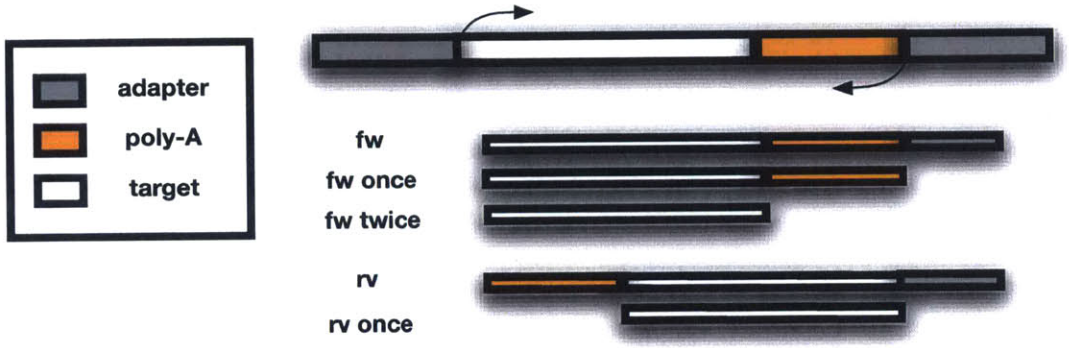


Figure 1-5: Schematic of RNA-seq analysis. A high level overview of the implementation of the RNA-seq pipeline in **bcbio-nextgen** version 0.8.5a, with the external programs used in each module listed. Included are modules to do gene, isoform and exon-level differential expression calling, variant calling on RNA-seq data, fusion gene calling for assaying structural variation in cancers, transcriptome assembly, quality control of the raw data and the alignments, clustering and sample outlier detection and automatic report generation. Red colored boxes indicate user-supplied data and green boxes are functionality supplied by **bcbio.rnaseq**.

read lengths, the read will be long enough to continue on past the RNA sequence into the adapter sequence. Several tools that have been created to fix this adapter read through issue, with tradeoffs regarding specificity, sensitivity and speed[83].

Careful tuning of the parameters of the read trimmer beyond the default values can improve sensitivity and specificity. Figure 1-6a on the following page shows that tuning cutadapt[109] to trim twice on the forward reads, once to trim possible adapter read-through and once to trim polyA sequences that are masked by the adapter sequence can rescue reads that would have been lost from the 3' end of RNAs. Beyond tuning for accuracy, paired-end read trimming using cutadapt can be tuned to run faster on compute environments with spinning disks by a simple architecture change. When run in paired-end mode, cutadapt must write two temporary files to disk and be run twice, this I/O operation is very slow and takes time. If running thousands of samples simulataneously, this can become a large bottleneck in an analysis. If instead of writing temporary files to disk, the command is constructed to use named pipes[164] instead, a 30% gain in performance can be realized, even on a very small dataset of a single lane of a million reads.

There has been some debate regarding if it is useful to trim low quality ends of reads before aligning. Most, but not all, modern aligners can take into account the low quality scores so most analysis pipelines include a quality trimming step. The commonly used threshold for RNA-seq is to to trim bases with PHRED quality scores less than 20, but a more gentle threshold of 5 results in much less information lost[105].



(a)

hard drive	streaming	time (seconds)
SSD	no	61
SSD	yes	64
platter	no	211
platter	yes	154

(b)

Figure 1-6: Tuning of adapter trimming with cutadapt. **a)** Trimming of adapter, polyA tails and other non-informative contaminant sequences from the ends of reads is necessary for compatibility with downstream tools that cannot handle reads with contaminated ends, such as aligners that do no soft clipping or kmer counting algorithms. cutadapt needs the flag set to try trimming twice when handling RNA-seq data to be able to trim polyA tails masked by adapter sequence. **b)** cutadapt requires intermediate files to be written out when handling paired end data and cannot natively stream the files from one step to another. For small amounts of data on fast disks (SSDs), this does not contribute to the processing time at all. On slow disks writing the intermediate files is the bottleneck in the process. Replacing the intermediate files with named pipes, thus re-enabling streaming, speeds up processing by 30% on a million paired-end reads.

Alignment

RNA-seq reads may cross exon-exon boundaries and when aligned to the genome will appear to be split across an exon. RNA-seq aligners have to be split-read aware or be able to take a gene model and create a proper alignment for these reads. The STAR aligner[48] is very fast and accurate aligner for RNA-seq that can map reads up to 50 times faster than Tophat2 but requires a machine with 50 GB of memory to run[54]. In addition to mapping reads quickly and accurately, STAR can simultaneously generate a mapping to the transcriptome for use with downstream quantitation tools such as eXpress[146], saving a step in the downstream analysis.

When high-memory compute is not available, Tophat2 is run instead of STAR. For paired-end reads, a small subset of the reads are mapped with Bowtie2 to determine an estimate of the mean and standard deviation, using the median and median absolute deviation as proxies for the mean and standard deviation of the insert size. The transcriptome-only mapping is made using bwa-mem since there is not a need to handle the intron spanning reads and bwa-mem is extremely fast and sensitive.

Transcriptome expression quantification

The primary goal of most RNA-seq experiments is to examine differences in the transcriptome between two or more experimental conditions. In order to assay the differences in transcription, the transcriptome must first be quantified. Depending on the organism, quantifying the expression of the transcriptome can be more complex than it initially seems. In organisms that have little to no alternative splicing of transcripts, this task is conceptually simple: count up the number of reads mapping to each gene and determine if the number of reads mapping to the gene is systematically different between conditions. For organisms such as the human with

transcriptomes that undergo intensive alternative splicing[181], the task of calling differences between conditions is much more complex. The complexity arises because what is sequenced are small pieces of transcripts which, when alternatively spliced, can be very similar to each other. Figure 1-7 on the next page shows an example where a single gene has multiple transcripts. Determining which isoform to assign reads that could come from multiple transcripts is a complex chicken-and-egg problem since it requires knowledge of how much of each transcript is expressed, which is what you are trying to estimate.

Many approaches ignore the read assignment issue altogether by quantifying at the level of the gene; the reads aligning to all transcripts of a gene are counted and combined to be the total reads mapping to the gene. The trouble with this approach is it doesn't reflect reality, the reads came from individual transcripts, not a gene and quantifying expression in this way can lead to errors. Figure 1-7 on the following page shows an example of one type of error, where a splicing event results in the expression of a much smaller transcript in one condition resulting in a false differential expression call.

Other practical concerns make quantifying the expression of the transcriptome difficult. Quantifying transcripts is dependent on the state of the annotation of the transcriptome because isoforms of genes that do not occur in the transcriptome will not have reads assigned to them, and reads that were sampled from unannoted isoforms will be misattributed to related isoforms, an issue illustrated in cartoon form in Figure 1-7 on the next page. The state of the transcriptome annotation for model organisms varies widely even for closely related organisms that should show similar degrees of alternative splicing (e.g. Figure 1-8 on page 44). Thus, for most organisms, augmenting the existing transcriptome assembly is necessary for accurate differential isoform calls. However this thesis, (Table 1.1 on page 43) and work of

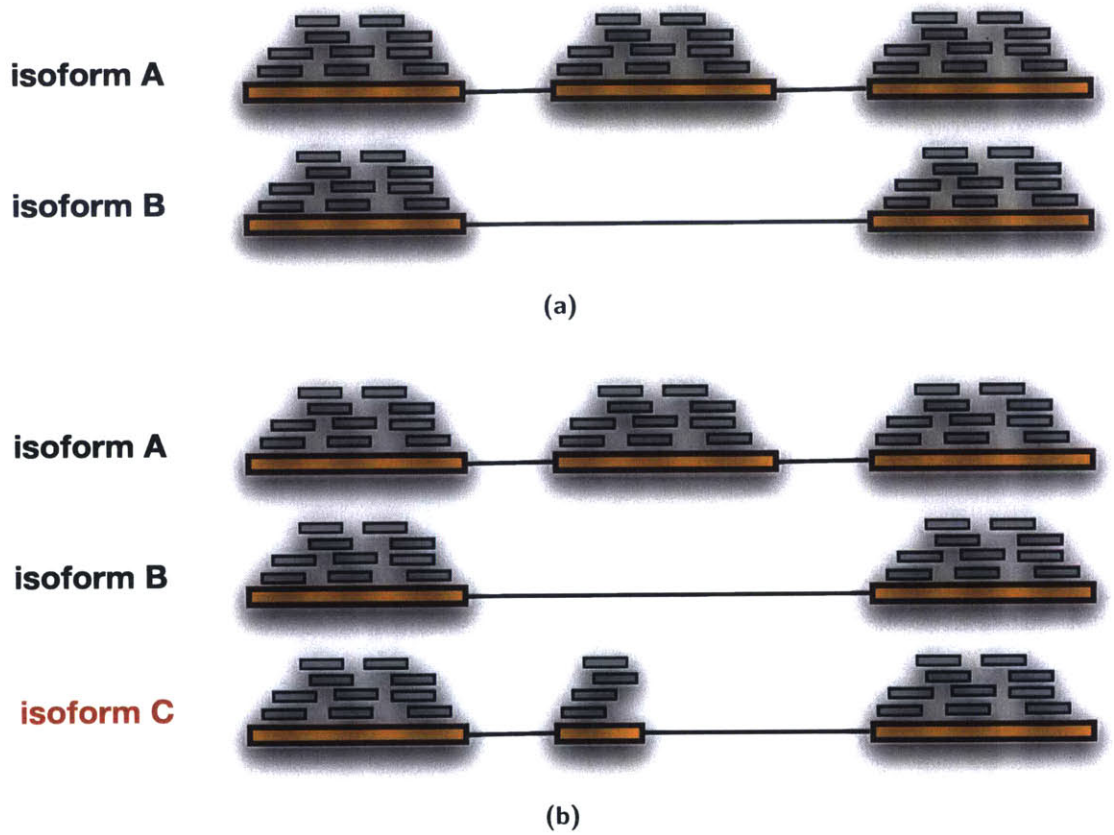


Figure 1-7: Gene level quantification can introduce errors. **a)** Quantification at the gene level can introduce quantification errors. If in one condition isoform A is expressed and in another condition isoform B is expressed, quantification at the gene level by counting the number of reads aligning to each gene will show a 1.5x fold change, even though the expression of the gene is unchanged. **b)** Missing isoforms can introduce errors in isoform-level differential expression calls. In the illustration, isoform C is missing from the annotation which will lead to reads being incorrectly assigned to isoform A. This also illstrates the complexities of assigning reads to specific isoforms; if a read aligns to the first exon, determining to which transcript of the gene it should be assigned is not a trivial problem.

others[3] have shown that transcriptome assembly is often incomplete and rife with false positives. Including these error prone assemblies introduces a major source of noise and into quantifying the transcriptome expression.

The challenges in quantifying at the isoform level have lead to the exploration of algorithms quantitating individual splicing events, exons and parts of exons instead of isoforms[87][7]. Quantitating at this level requires much less accurate transcriptome information and only requires enumeration of the exons and splicing events that can occur in the data, a much more tractable problem than a complete enumeration of all possible isoforms. In addition, for incomplete transcriptome annotations, assembling exons is much more successful than assembling entire isoforms[3], especially for organisms in which splicing is complex, and the enumeration of the exons is generally more complete in exisiting transcriptome annotations. Recently a method called derfinder[60] was developed which takes the resolution of the transcriptome quantitation to the extreme and quantitates transcriptome expression at the level of a single base.

Transcriptome quantification is implemented at three levels of resolution in **bcbio-nextgen**, at the level of the gene with featureCounts[97], eXpress[146] and Cufflinks[172], the isoform level with eXpress and Cufflinks and the sub-exon level with DEXseq[7]. featureCounts and eXpress both generate gene-level estimated counts of reads mapping to genes, suitable for use in count-based differential expression callers (see Section 1.1.8 on the following page); featureCounts only counts reads which can be uniquely assigned to a gene whereas eXpress assigns ambiguous reads probabilistically based on the overall expression of the gene. There are other tools with similar functionality to featureCounts and eXpress, but featureCounts and eXpress are both extremely fast, up to 30 times faster than similar tools[97]. For isoform expression, both eXpress and Cufflinks produce estimates of the gene-level and isoform level

expression, eXpress summarizing with estimated counts suitable for use in count-based callers and Cufflinks with FPKM, a gene-length normalized expression measure, which a companion program Cuffdiff uses to call differential isoform expression. Finally DEXseq quantitates differential expression at the level of the exon fragment; exons are broken into the smallest fragments unique to an isoform in the existing annotation and DEXseq quantitates the expression of those fragments. Each of these levels of quantitation are used to make differential expression calls in a companion tool released with **bcbio-nextgen**, **bcbio.rnaseq**, this gives the researcher flexibility to choose the resolution of quantitation that is most appropriate to their experiment.

Differential expression

Differential expression calling on the gene, isoform and splicing event level is performed with a companion program to **bcbio-nextgen** called **bcbio.rnaseq**. **bcbio.rnaseq** runs baySeq[72], DESeq2[101], edgeR[149], Cufflinks[171], voom+limma[94], edgeRun[47] and EBSeq[95] to call gene-level differential expression, Cufflinks and EBSeq to call isoform-level differential expression and DEXSeq to call splicing event level differential expression using the estimated expression values calculated from the **bcbio-nextgen** RNA-seq pipeline. Running several tools is important, as different tools perform well on specific types of RNA-seq data. DESeq2 and limma are great choices for most RNA-seq experiments, but they can be outperformed in specific conditions. For example for low replicate, low count experiments with under ten million reads per sample or less, we have created an improved algorithm called edgeRun[47], which is much more sensitive for these specific types of experiments.

Transcriptome augmentation

A strength of RNA-seq over microarray analysis is that novel isoforms, novel genes and other novel coding or noncoding RNA can be discovered. With a microarray analysis the gene fragments assayed are generally restricted to the set of what is already known. There has been a lot of work on how to augment the known transcriptome with novel isoforms and genes, through either de-novo assembly of the transcriptome or using the known transcriptome as a reference to guide the discovery of new isoforms. This problem is extremely difficult to solve, and the state of transcriptome assembly software reflects that difficulty. For complex transcriptomes with many splicing events such as human, only 30% of the full length transcripts can be recovered from simulated data, with a similarly low precision[163].

The transcriptome assembly implementation in bcbio-nextgen accounts for the difficulties in de-novo assembling the transcriptome by using the reference guided assembly mode in Cufflinks, which uses the known transcriptome as a guide to call novel isoforms and novel genes[148]. Even with this crutch, the false positive rate of Cufflinks assembly is very high[3]. We confirmed this by simulating reads from chromosome 22 of the human genome using gene models from Ensembl release 75 and assembling the transcriptome with Cufflinks v2.1.1, using an annotation where half of the known annotated transcripts are dropped at random. This resulted in an assembly where nearly half of the novel transcripts are false positives. We implemented a simple filtering algorithm by training a logistical model[183] to predict coding/non-coding status of the novel transcripts, training against the known transcripts, and keeping isoforms only if the coding status matches the known coding status of the known gene. This filtering greatly reduces the false positive rate at the cost of a moderate hit in sensitivity on simulated data(Table 1.1 on the next page).

status	specificity	sensitivity
unfiltered	0.51	0.70
filtered	0.85	0.63

Table 1.1: Cleaning the raw Cufflinks reference-guided transcriptome assembly improves the false positive rate of the assembly. Unfiltered Cufflinks-assembled transcripts have a sensitivity of 0.7 but a specificity of 0.5, indicating that half of the assembled transcripts are false positives. Filtering the assembled transcripts by removing transcripts with a predicted coding status differing from the known coding status of the parent gene in the annotation improves the specificity to 0.8 at a minor hit to the sensitivity. Reads were simulated from chr21 of the human genome using gene models from Ensembl release 75 and Cufflinks was run in reference-guided mode with an assembly missing half of the known annotated transcripts.

Using data from a real experiment in Chapter 3 of this work, we reasoned that the rat and mouse should have similar numbers of transcripts per gene since the two species are highly related. The mouse annotation has many more genes per transcript identified, so we assembled the rat transcriptome and compared the number of transcripts per gene identified before and after filtering. Prior to filtering, more transcripts were assembled from the rat transcriptome, even though the samples should capture transcripts expressed in the brain, about 50% of the total transcripts in the organism (see Chapter 3 on page 87). Our filter greatly reduced the total number of novel transcripts identified, while still increasing the number of transcripts by 20% (Figure 1-8 on the next page). The coding/noncoding prediction part of the filter is also used to classify novel genes, tagging novel genes with a low coding score as ncRNAs or lncRNAs, depending on their length, and genes with a high coding score as protein coding. The assembly, model training, filtering and classification are run automatically during the RNA-seq pipeline implemented in **bcbio-nextgen**.

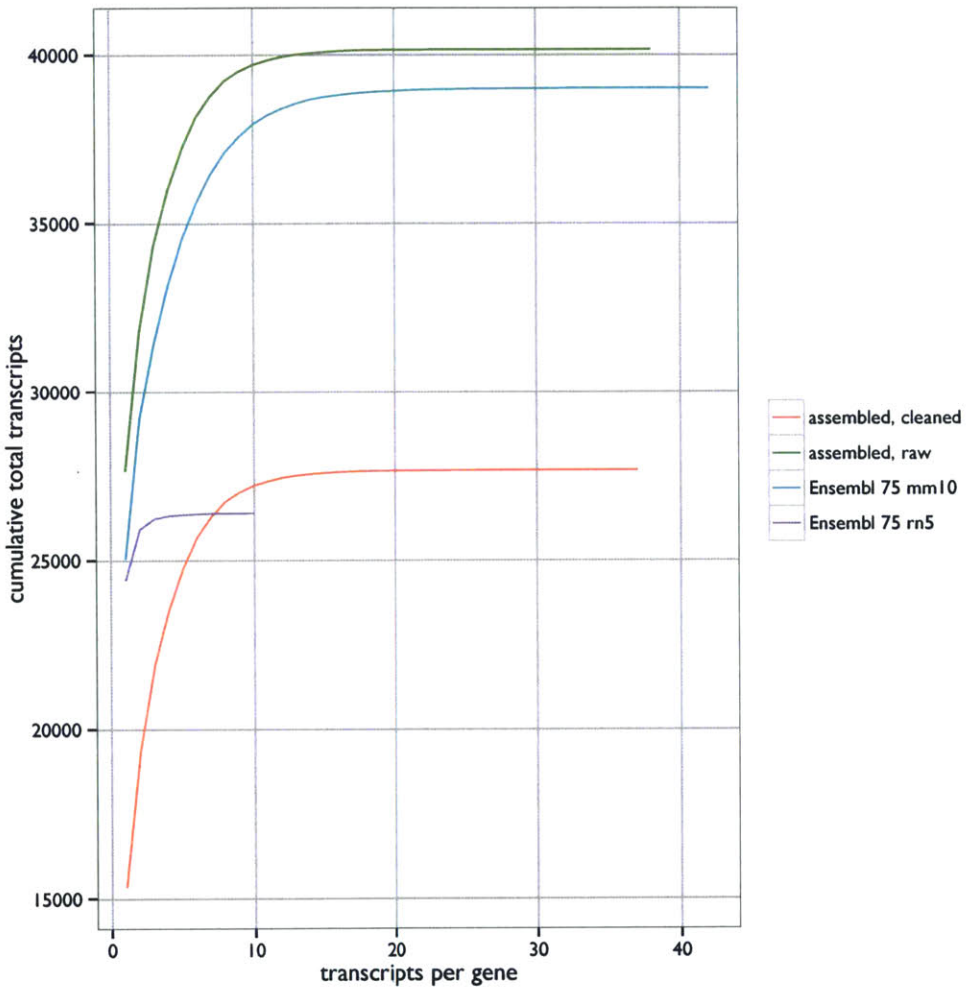


Figure 1-8: RNA-seq improves the rate transcriptome annotation. The rat (rn5) Ensembl gene annotation lags behind its close relative, the mouse (mm10), annotation in numbers of transcripts per gene identified. Using RNA-seq data from the rat superior colliculus, the unfiltered Cufflinks assembly identifies more transcripts than exist in the mouse gene annotation even though only 50% of the transcripts of the rat are expressed in the superior colliculus. Filtering the transcripts by keeping only those transcripts that agree with the coding status of their parent gene greatly reduces the number of novel isoforms identified, and is a much more reasonable result.

1.1.9 Variant calling and RNA editing

Another benefit of sequencing for gene expression analysis over microarrays is the ability to call variants from the RNA-seq data. The reads are aligned to the genome and differences in the RNA compared to the reference sequences are called. Calling variants from RNA-seq data is prone to errors, however.

Variant calling from RNA-seq data is implemented in bcbio-nextgen following recommendations[78] from the Broad Institute. This is supplemented with custom filters to remove variants that are likely to be technical artifacts. When aligned to the genome, RNA-seq reads that cross exon-exon junctions are soft clipped, meaning those bases are masked during the alignment and HaplotypeCaller, the Broad's variant caller, can't handle the soft clipped bases. We split those reads into two reads, one for each end of the exon and then call variants with HaplotypeCaller. We then filter the called variants, throwing out variants that have a large strand bias, as these are likely to be false positives caused by incorrect mapping. We also drop variants that are biased in terms of where they appear in a read as these are also likely due to incorrect mappings. Other sources of error come from misaligned reads to the genome. The transcriptome is much more complex than the genome and a given exon can be attached to many different exons through alternative splicing. This can cause many false positive variants to be called where two exons are spliced together [139]. We filter out variants that are within 5 bp of an exon-exon junction in our final set of variant calls. During this process we also keep all variants that are known to exist for that organism using the most current dbSNP release and pass them through unfiltered.

Variants can also be identified in the RNA sequences not due to variants in the DNA sequence, but by variants in the RNA sequence introduced by A-to-I RNA-

editing by a class of enzymes called ADARs[152]. For supported organisms, we separate out variants that are likely to be RNA-editing events from DNA variants using a combination of preexisting annotations and custom filters. The majority of RNA editing events verified to date are A-to-I edits, where the I is interpreted as G during translation. For mouse and human samples, possible RNA-editing events that appear in the the Rigorously Annotated Database of A-to-I Editing (RADAR) and the Database of RNA Editing in Humans (DARNED) are kept.

Functional effects of the cleaned DNA and RNA-editing events are calculated using the Variant Effect Predictor from Ensembl. Human calls are then loaded into a GEMINI[128] database, which loads the variants into a database and annotates the variants with information from their genomic context, using existing variant annotations from ENCODE, OMIM, dbSNP, KEGG, HPRD and other databases.

Experiment summary

bcbio.rnaseq produces a summary report of the output from the **bcbio-nextgen** pipeline in the Rmarkdown language, a hybrid of the markdown document formatting language and the R statistical analysis language. The report includes a wealth of information for each sample including the number of reads mapped, the mapping rate, the number of genes and isoforms detected, the amount of rRNA contamination, the overall complexity of the library, and how well the transcriptome was covered. Heatmaps of correlations between samples are included in the plot as well as multidimensional scaling (MDS) plot; these plots along with the basic metrics regarding each sample are helpful for eliminating outlier samples from the analysis. If a model formula is included in the **bcbio-nextgen** run, **bcbio.rnaseq** will run the differential expression callers on the output and generate a summary report of

the differentially expressed genes found in the experiment. The RMarkdown report serves as a starting point for digging deeper into an experiment. It has all of the metadata supplied about each sample, the calculated quality metrics, the normalized count data, the differential expression calls and all of the R code to perform the analysis and make all of the figures. Given the output from a **bcbio-nextgen** run and the Rmarkdown report, a researcher can regenerate and tweak the analysis and figures for a project.

1.2 Discussion

This chapter described an open source, community curated analysis for RNA-seq data that is being used by researchers in academia, core facilities, startups and pharmaceutical companies around the world. It builds on top of and extends previously implemented infrastructure work that supports the automated deployment of all necessary software and data necessary to run an analysis. It is highly scalable, capable of running thousands of samples simultaneously across a wide range of computing environments including local machines, high performance clusters and on Amazon and other cloud computing platforms. It provides a set of validation tools to justify the implementation choices and to allow for future new methods to be compared to existing methods in an unbiased manner. It implements features not found in preexisting commercial and open source solutions, including filtering of isoform calls by using the consensus of multiple tools, calling and classifying variants, and calling existing RNA-editing events from RNA-seq data. It can be configured to work in concert with a laboratory information management system (LIMS) to automatically kick off analyses and can be configured to automatically upload results to Amazon S3 or to a Galaxy instance for dissemintation to collaborators. A report designed as a

summary and a starting point for more in depth analysis is automatically generated at the end of the process.

Importantly, this work is open source and under active development by a vibrant, thorough community of next generation sequencing experts. As part of an umbrella project, features that get incorporated into the infrastructure by the community, such as adding the ability to use Docker images, are able to be used by all analysis implementations. This ensures that as progress moves forward the analyses will keep up with the current science.

1.3 Future development

There are many areas of active development, both in terms of the infrastructure driving **bcbio-nextgen** and **bcbio.rnaseq** and improving the accuracy and speed of the implemented analyses. An important feature that is missing from the RNA-seq pipeline is support for finding and quantitating both known and novel very small exons called microexons. Recent work has found microexons to be misregulated in the brains of autistic patients and microexons to be preferentially expressed in brains in general[79]. With the kick off of the BRAIN initiative soon to start, there should be enormous interest in assaying transcription in the brain and microexons are likely to be an important feature of the transcriptome to assay.

In addition, single-cell RNA-seq has become an exciting new technique and analyzing the data from single-cell experiments is challenging[81]. There is a large increase in technical noise caused by incomplete sampling of the transcriptome during library preparation from a single cell[68]. The most sensitive protocols only capture 40% of the RNA molecules in each cell and so there are many artificial zero counts due to the incomplete capture[68]. **bcbio-nextgen** and **bcbio.rnaseq** have been

used to analyze single-cell data but at present the implementation is not robus and validation data is not available. As this type of data becomes more common, there will be a need to develop an option to handle single-cell RNA-seq data.

We are in the process of swapping in much faster tools for transcriptome reconstruction[135] and quantification[23][130]. These tools will allow for these tasks to be done on the order of minutes instead of hours. The current work has highlighted the importance of having validated, reproducible analyses. We have discovered bugs that escaped reviewers in some of these programs that severely affected their output (see [134]).

Finally, the **bcbio-nextgen** community has been watching and participating in the formation of the common workflow language (CWL). The CWL is an in-progress specification for a language of describing data-intensive workflows in a language agnostic way. The common workflow language is a high level language for describing steps in a workflow by specifying the expected inputs and outputs and the tools to run and linking them together to form an analysis pipeline [162]. This allows the informatics workflow to be abstracted away from the compute and storage and allows workflows to be shared across different infrastructure. We plan to replace the **IPython.parallel** method of parallelizing a **bcbio-nextgen** workflow with the CWL when it is complete.

Chapter 2

Transcriptome defects in a mouse model of tuberous sclerosis

2.1 Background

2.1.1 Tuberous sclerosis

Tuberous sclerosis is an autosomal dominant genetic disorder characterized by growth of benign tumors of the skin (angiofibromas), lung (lymphangiomyomatosis), heart (cardiac rhabdomyomas), kidney (renal angiomyolipomas) and brain (cerebral cortical deformations called cortical tubers). Tuberous sclerosis affected individuals have a normal lifespan but have debilitating neurological symptoms including autism, epilepsy and delays in intellectual development [84][42]. Clinical signs of tuberous

sclerosis vary between patients with no one sign being diagnostic, but 80 % of patients present with cortical tubers. Tuberous sclerosis, Latin for 'hard swelling' is named after this characteristic symptom of the disease.

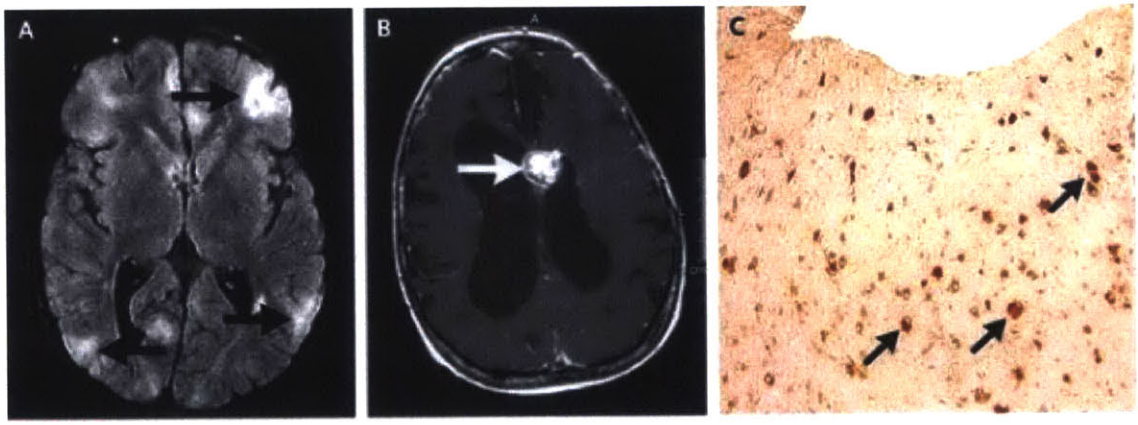


Figure 2-1: CNS manifestations of tuberous sclerosis. **A)** an MRI showing several cortical tubers with the apex of the tubers pointing towards the ventricle. This shape is due to disruption in migration of cells from the ventricle to their proper cortical layer. **B)** shows another neurological tumor that can develop, a subependymal giant cell astrocytoma (SEGA). This is a non malignant tumor in the lateral ventricle. **C)** shows a magnified area of a cortical tuber stained with an antibody against the translation marker ribosomal S6. Arrows point to giant cells with overactive translation. Reproduced with permission [42], Copyright Massachusetts Medical Society.

The cortical tubers are areas of the brain that develop abnormally due to disruption of outward migration of presumptive cortical cells from the ventricle. This disruption of migration causes the characteristic triangular shape of the cortical tuber, with the apex pointing at the ventricle (Figure 2-1). The tuber itself is made up of cells called giant cells which have large, swollen cell bodies and stain strongly for ribosomal S6 protein, indicating overactive protein translation. The giant cells can also stain for neuronal or glial markers, and occasionally for both markers, indicating the giant cells are undifferentiated or partially differentiated neurons or glia [117]. In addition to the presence of giant cells, the area within and around the

tuber shows disruption of normal cortical lamination, a phenomenon called cortical dysplasia. This abnormality occurs in other diseases such as focal cortical dysplasia and is most often associated with seizures. Intracranial EEG studies of children with tuberous sclerosis confirm the link between the tubers and seizures. Subdural electrodes placed over tuberous and non-tuberous cortical tissue indicate that the tuber itself is often, but not always, the locus of seizure generation Resection of tubers and surrounding dysplastic cortex often halts intractable seizures, further supporting the connection between tubers and seizure initiation.[118]. However, despite the strong association between cortical tubers and seizures, some epileptic tuberous sclerosis patients do not have any discernable cortical tubers or cortical dysplasia [185]. Mouse models of tuberous sclerosis also reflect this heterogeneity in the seizure phenotype as the presence of both seizures and tubers can vary depending on the nature of the knockout. Seizures without tubers or gross cortical abnormality suggest that changes at the molecular level may also be contributing to the seizure phenoptype.

Tuberous sclerosis is a genetically simple syndrome. In 85 % of cases of tuberous sclerosis, the root cause can be linked to mutations in one of two genes, TSC1 and TSC2, that encode the proteins hamartin and tuberin, respectively. These two proteins bind together to form the tuberous sclerosis complex (TSC) that governs the rate of translation in the cell via inhibition of the mammalian target of rapamycin (mTor) pathway. Tuberous sclerosis is a monogenetic disorder because loss of function mutations in TSC1 or TSC2 cause the mTor pathway to be overactive and allows protein translation to run out of control. Specifically, the TSC governs the rate of protein translation through inhibition of Rheb-GTP activity and loss of TSC function causes Rheb-GTP to be overactive.[99] Activated Rheb-GTP triggers the mTORC1 kinase branch of the mTor pathway, which acts as a positive regulator of protein synthesis by phosphorylating p70S6 kinase and 4E-binding protein 1 (4E-

BP1), both are involved in governing the rate of translation in the cell (Figure 2-2 on the next page). Phosphorylated p70S6 kinase in turn phosophorylates ribosomal protein S6 kinase (S6K1), an important step in synthesis of the translation machinery [66] while phosphorylation of 4E-BP1 causes it to dissociate from eukaryotic translation initiation factor 4E (eIF4E), allowing protein synthesis to occur[88]. Thus these two phosphorylation events are essential to activation of the mTor pathway. They cause the synthesis of a key component of the scaffolding of translation and also releases a major translation factor from inhibition. Thus the pathway consequently has a profound effect on the fate of the cell.

Increased protein production causes benign tumor-like growths called hamartomas in tuberous sclerosis patients through causing cells to either fail to enter G0 and remain preliforating or leave G0 and re-enter the cell cycle[159]. This results in the formation of the hamartomas such as angiofibromas in the tuberous sclerosis patients. Inhibiting mTor with rapamycin slows the rate of growth of these hamartomas in tuberous sclerosis patients[75]. Deactivation of the tuberous sclerosis complex not only affects the proliferation but also the differentiation of cells. This interference with differentiation is the cause of the poorly differentiated giant cells seen in the cortical tubers and treatment with rapamycin can cause neurons to properly differentiate after Tsc2 inhibition[158].

In addition to disrupting the migration and differentiation of neurons and glia, overactivation of the mTor pathway has more subtle effects on many aspects neural circuit formation ranging from axonal arborization to spine formation and synaptic potentiation. These molecular and circuit effects may be the locus of the more subtle neurological symptoms of tuberous sclerosis such as autism. Overexpression of either Tsc1 or Tsc2 in cultured hippocampal neurons greatly reduces the number of neurons sprouting axons while knocking down either Tsc1 or Tsc2 has the opposite effect,

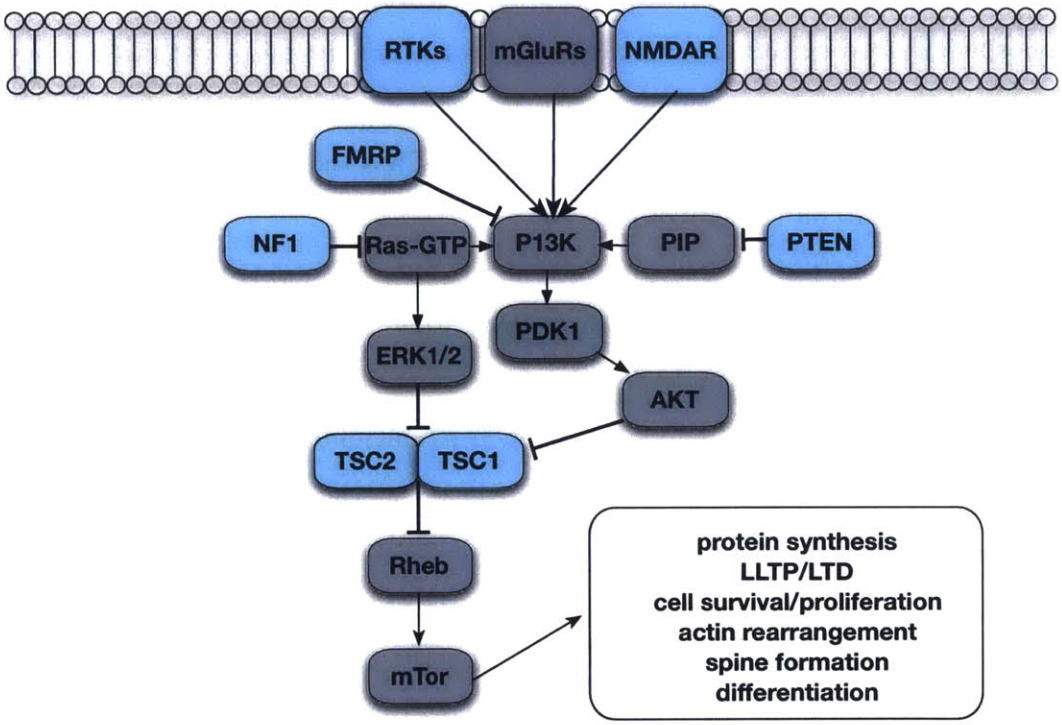


Figure 2-2: The mTor pathway is affected in many disorders with autism as a phenotype. Genes colored blue are genes known to be associated with autism. TSC1 and TSC2 are disrupted in tuberous sclerosis, NF1 in neurofibromatosis, FMRP in fragile X syndrome and mutations in PTEN, receptor tyrosine kinases and NMDA receptors are all known to be disrupted in some autism spectrum disorders.

causing, in culture, an increase in the number of neurons with multiple axons[39]. Knocking out Tsc1 also disrupts myelination of neurons[114]. These results demonstrate that disruption of the activity of the tuberous sclerosis complex affects the structure and functional capabilities of neural circuits. Supporting this notion, other forms of local circuit disruption are found when the tuberous sclerosis complex is not fully functional. $Tsc2^{+/-}$ mice have an aberrant retinotopic map which is formed by a process of axon guidance to the correct termination zone via interaction of the ephrins and the Eph receptors (see Chapter 3 on page 87) of this work for a brief review)[127], indicating that proper regulation of the mTor pathway is necessary for both the outgrowth of the axon and guidance of the axon to its proper termination zone. Overactivation of the mTor pathway can also affect the fine tuning of the synaptic contacts. It has long been known that blocking the NMDA receptor with low levels of ketamine can induce new synapse formation in a protein synthesis dependent manner. Blocking the mTor pathway with rapamycin blocks this effect of ketamine on new synapse formation [96]. More subtly, mTor activation is important in protein synthesis dependent long long-term potentiation and LTD. Mice heterozygous in loss of function mutations of Tsc1 or Tsc2 have a reduced threshold for protein synthesis dependent long long-term potentiation[52] due to overactivation of the mTor pathway. The decreased threshold for long long-term potentiation in tuberous sclerosis heterozygote knockout mice is likely the cause of the deficits in learning which can be reversed with rapamycin treatment. Knocking out Tsc1 can also affect signaling in the neuropil through overactivate translation in glia. A Tsc1 knockout specific to glia has excessive glutamate signaling which can be reversed by partially blocking the NMDA receptor[192]. Disrupting Tsc1 function in neurons causes a weakening of inhibition in the hippocampus, leading to overexcitability[12]. These results show that disruption of the careful balance of translation in the cell

has wide ranging consequences for the architecture of the brain and can affect many phases of wiring of the neuropil, from the initial migration of the cells to their proper positions, to controllng and guiding the axons via chemotactic signals. The defects involve both the initial wiring of the brain, as with the retinotopic map disruption, and the fine tuning of established circuits, as with the learning deficits. These multiple levels of brain connectivity disruption from overactive protein synthesis have devastating neurological consequences for affected individuals.

There is no cure for tuberous sclerosis and treatments are focused on management of the most severe symptoms. A promising avenue is chronic treatment with low levels of rapamycin, the inhibitor of the mTor pathway that the mTor (mammalian target of rapamycin) pathway is named after. Treatment with rapamycin can slow the growth of tuberous sclerosis tumors, [53], help control epilepsy[25] reverse autism-like symptoms and, in a $Tsc2^{+/-}$ mouse model of tuberous sclerosis, reverse short term learning deficits[151][52]. The success of rapamycin treatment on the symptoms of tuberous sclerosis has caused an interest in investigating other treatments targeting both the mTor pathway and pathways affected downstream of mTor[150]. To enable these studies, many mouse models of tuberous sclerosis have been developed which display subsets of the anatomical and neurological symptoms of tuberous sclerosis, depending on the nature of the knockout.

Tsc1 and Tsc2 null mice are embryonic lethal, with renal tumors and failure to close the neural tube[91], so much of the work on rodent models of tuberous sclerosis focuses on less severe conditional knockouts and heterozygotes. Conditional knockouts of Tsc1 and Tsc2 have been made in both glia and post-mitotic neurons. Glia-specific Tsc1/2 knockouts were created by crossing floxed Tsc1/Tsc2 alleles to mice expressing Cre recombinase under control of the human glial fibrillary acidic protein promoter (GFAP-Cre) to generate the glia-specific Tsc knockout

mice (GFAP-Tsc1/2$^{-/-}$)[177][193]. Neuron-specific Tsc1/2 knockouts were created by crossing floxed Tsc1/Tsc2 mice to mice expressing Cre recombinase driven by the rat synapsin I promoter to create Syn1-Cre mice to generate neuron specific knockouts (Syn1-Tsc1/2$^{-/-}$)[114]. Interestingly, both the GFAP-Tsc1$^{-/-}$ and GFAP-Tsc2$^{-/-}$knockout mice both have seizures and disruption of the laminar structure of the hippocampus, despite a lack of tuber formation[193]. The Syn1-Tsc1$^{-/-}$ mouse has seizures, autism-like symptoms but no tubers or disruption of cortical lamination (Figure 2-3 on the next page). The severity of the seizure and anatomical phenotype is temporally related to when Tsc1 function is lost in neurons, as a Emx1-Tsc1$^{-/-}$ conditional knockout, that removes Tsc1 expression in neural progenitor cells, has seizures and severe deficits in cortical lamination[32]. Since proper mTor activity is necessary for neurons to differentiate and leave the cell cycle, overactive mTor prevents neurons from properly differentiating and migrating, resulting in the creation of the poorly differentiated giant cells, cortical tubers and lamination disruption. The Syn1-Tsc1$^{-/-}$ mouse loses Tsc1 after cells have differentiated and migrated to their final location, so the cortical tubers and cortical dyplasia never form.

The above is a summary of the major neurological effects caused by overactivation of the mTor pathway in tuberous sclerosis. These effects operate simultaneously on several different levels of neuronal organization. At the highest level, aberrant mTor activity affects differentiation of neurons and glia, resulting in cortical tubers, major disruption of the cortex and seizures. At this level the neurons that make up the neuropil are not properly formed or not properly localized. Knockouts where the activity of the tuberous sclerosis complex is disrupted either later on in development or in a subset of already differentiated cells have phenotypes that are a subset of the tuberous sclerosis symptoms. In these milder models the more subtle aberrant wiring of the neuropil is revealed, with disruptions both in chemotactic and activity

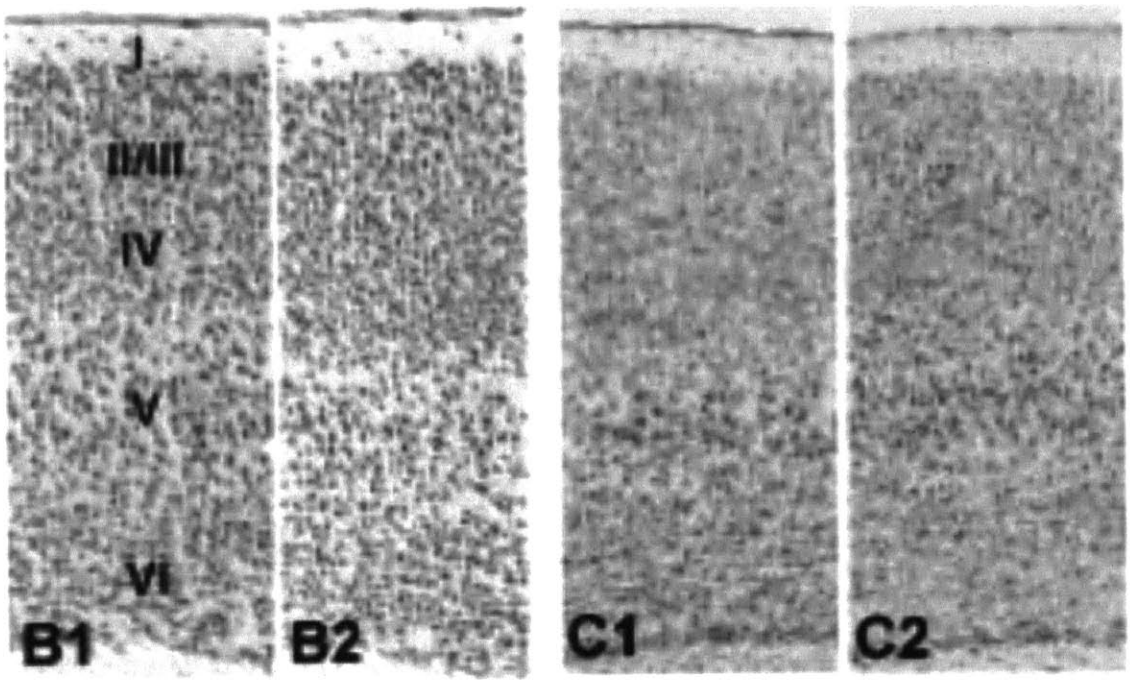


Figure 2-3: The laminar structure of the cortex in Syn1-Tsc1$^{-/-}$ mice is undisturbed. Syn1-Tsc1$^{-/-}$ mice, a neuron-specific knockout of Tsc1, has seizure activity, giant cells and dysmorphic neurons and glia, but no tubers and no disruption of the laminar structure of the cortex. **B1)** NeuN staining shows a normal cortical structure in the controls and **B2)** Syn1-Tsc1$^{-/-}$ mice. cresyl violet staining also shows no difference between **C1)** controls and **C2)** Syn1-Tsc1$^{-/-}$ mice. Used with permission from [185], copyright John Wiley and Sons.

dependent wiring. These two features of tuberous sclerosis, that it is a monogenic disorder causing autism and that mice can be generated which show only a subset of the phenotypes make it a powerful model system for the study of autism.

2.1.2 Tuberous sclerosis is a tractible autism model

Autism spectrum disorder is a complex multigenic disorder with evidence of hundreds of succeptibility loci that, with some exceptions, only confer a small amount of risk. A recent whole genome study of quartet families where both parents and two ASD affected siblings were sequenced was a dramatic demonstration of the genetic complexity: siblings from the same family, both with autism spectrum disorders were found not to share the same putatively causitive mutations in over 70 % of the cases. [191]. Most diagnosed autism disorders are genetically complex with no single causative mutation, however there are monogenic diseases which feature autism as one of the predominant symptoms. These monogenic disorders include neurofibromatosis, mutations in Pten, Fragile-X disorder and tuberous sclerosis. Monogenic disorders causing autism are valuable as systems to study and treat the symptoms of autism spectrum disorders as knockout animal models can be generated that have an autism phenotype[10]. The efficacy of autism spectrum disorder interventions can then be assayed using a social behavioral test for autism spectrum disorder in mice[157].

Many of the monogenic disorders with a penetrant autism phenotype involve disruption of genes regulating the mTor pathway (see Figure 2-2 on page 54). Over 50 % of patients with tuberous sclerosis present with symptoms of autism spectrum disorder (ASD). Mouse models disrupting Tsc1 or Tsc2 function recapitulate the social deficits and can be rescued with low levels of rapamycin[166]. Mu-

tations in NF1, the gene coding for neurofibromin, cause neurofibromatosis type 1, with 15–30 % of affected individuals presenting with autism spectrum disorder symptoms[63][140]. Neurofibromin inhibits Ras-GTP and Ras-GTP activates Erk1/2 which inhibits Tsc1/2, leading to overactive mTor. A mouse model knocking out Nf1 has an autism phenotype that is rescued by blocking Pak1[119], a kinase activated by the mTor pathway[69]. Pten mutations cause autism in 10 % of affected patients[61] and again act through the mTor pathway. Pten loss of function leads to overactivation of AKT which blocks the activity of the tuberous sclerosis complex. A mouse model deleting Pten has autism-like behavioral deficits[102], and a neuron specific knockout of Pten has cortical dysplasia and seizures which, like tuberous sclerosis, can be rescued by rapamycin treatment [126]. Finally, Fragile X syndrome is another disorder presenting with autism like symptoms through activation of the mTor pathway, via disruption of the fragile X mental retardation protein (FMRP). FMRP normally inhibits P13 kinase, and release of PI3 kinase from inhibition causes the TSC to be inhibited[156]. All of these autism disorders increase activity of the mTor pathway through indirect inhibition of the activity of the tuberous sclerosis complex (Figure 2-2 on page 54).

The tuberous sclerosis complex is the common gateway through which these monogenic autism disorders exert their effects. The difficulty with using tuberous sclerosis as a model for studying autism is that tuberous sclerosis, while genetically simple, is phenotypically complex. Patients present with a wide range of severe symptoms and teasing apart what is involved in each phenotype is difficult. This complexity is ameliorated, however, by the existence of mouse models that have only a subset of the tuberous sclerosis symptoms. In particular, Syn1-Tsc1$^{-/-}$ mice lack the gross anatomical defects of other tuberous sclerosis models, but retain the seizure and autism phenotypes. This mouse represents the opposite of the autism genome-wide

association studies analyses are restricted to patients with non-syndromic autism with as few comorbid phenotypes as possible but one where the underlying genetic cause is unknown and complex. Studying Syn1-Tsc1$^{-/-}$ mice trades genetic complexity for phenotypic complexity, but the phenotupic complexity is reduced by studying the neuronal knockout of Tsc1 which has only the seizure and the autism spectrum disorder phenotypes.

2.1.3 The second order effects of mTor activation are important

Despite these autism associated disorders all resulting in overactivation of the mTor pathway, the downstream effects of mTor activation are not the same. This can be seen at multiple levels. At the phenotypic level, tuberous sclerosis disorder, Cowden's syndrome, where PTEN is lost, and neurofibromatosis all present with seizures, hamartomas or tumors and autism-like symptoms. Fragile X syndrome, on the other hand, presents with seizures and autism but the hamartomas are absent. The penetration of each endophenotype differs amongst these disorders with varying rates of autism, tumor and seizures in the population. This indicates that either downstream effects or non-mTor related effects of each disorder are driving phenotypic differences. At a molecular level, these syndromes can have vastly different effects despite the common overactivation of the mTor pathway. Multiple studies have shown the importance of dissecting the second order effects specific to each syndrome.

The effect of fragile X syndrome and tuberous sclerosis on the mGluR5 receptor is a salient example. Fmr1$^{-/y}$ mice, a mouse model of the fragile X syndrome, have enhanced mGluR5 dependent LTD which is triggered by enhanced protein synthesis downstream of mGluR5 activation. Tsc2$^{+/-}$ mice also have excessive mTor activation

but have decreased protein synthesis dependent mGluR5 LTD[10]. Crossing $Tsc2^{+/-}$ and $Fmr1^{-/y}$ mice to each other results in normal mGluR5 activity indicating that these two mutations have opposite effects on mGluR5 activation and can be balanced out, despite both resulting in overactivation of the mTor pathway. Both of these knockouts have phenotypes which can be rescued with rapamycin treatment, but this result shows that there are secondary effects of the mTor pathway activation that are specific to a particular genotype. In the case of $Fmr1^{-/y}$ and $Tsc2^{+/-}$, therapies targeting mGluR5 in one disorder would be contraindicated for the other disorder, despite both affecting the same pathway. A more nuanced dissection of the specific molecular effects of overactivating the mTor pathway is important if more focused, disease specific therapies are to be developed.

Most work assaying gene expression differences in autism has focused on large cohorts of non-syndromic autism disorders. Those studies have found differing amounts of evidence for large scale gene expression changes in the brain across non-syndromic autism disorders. A study of moderate size comparing gene expression in postmortem control and autistic human cortices found only MAL and C11ORF30 differentially expressed, both at very moderate fold changes. Pathway analysis in the same study found a gene co-expression module upregulated for immune-related genes[70]. Another study that examined both human cerebellum and cortex found there was scant evidence for gene expression differences in cerebellum but hundreds of differentially expressed genes in the frontal and temporal cortex of autism affected individuals[180].

Limited work has been done exploring the effect of mTor-linked syndromic autism disorders on transcription in the brain. A study examining the transcriptome of the cerebellum of $Tsc2^{+/-}$ and $Fmr1^{-/y}$ mice and found no gene-level significant differences in any comparisons, including $Tsc2^{+/-}$ and $Fmr1^{-/y}$ compared to control mice[92]. This study focused on the cerebellum which has mixed evidence for gene

expression differences in autism spectrum disorder brain[180]. Despite the diverse results, characterizing the transcriptome of specific syndromic autism spectrum disorders is important. Just as cancer is now known to not be a single disorder but to have specific subtypes that have very different molecular pathways disrupted, it is likely that a complex, subtle neurological disorder such as autism also can be caused by multiple different pathways. The differing regulation of mGluR5 in fragile X syndrome and tuberous sclerosis demonstrates the importance of characterizing these disorders at the molecular level.

For this study we used RNA-sequencing to assay gene expression changes in the frontal cortex of Syn1-Tsc1$^{-/-}$ mice that have Tsc1 knocked out only in neurons[114] via use of the rat synapsin I promoter, a membrane-associated protein expressed only in neurons. We examined differential gene expression, differential splicing and differential RNA-editing between wild type and Syn1-Tsc1$^{-/-}$ mice and found hundreds of gene and splicing differences. We found a set of genes known to be involved in autism and epilepsy which are differentially transcribed in these mice. Following up on one of these hits, we show evidence that overexpression of the serotonin receptor Htr2c causes hyperexcitability and synchronized calcium influx into the neuron.

2.2 Methods

2.2.1 Cortex collection

Syn1-Tsc1$^{-/-}$ mice were prepared by crossing Syn1-Tsc1$^{loxp+/-}$-Syn1$^{Cre+/-}$ mice and Syn1-Tsc1$^{loxp+/-}$-Syn1$^{Cre-/-}$ mice of the opposite sex. Four week old Syn1-Tsc1$^{-/-}$ and unaffected control littermates were anesthetized with isoflurane and decapitated. The frontal cortex was dissected, rinsed with ice-cold PBS and stored in RNALater

at -20 °C after being cut into 3-4 mm pieces to ensure penetration of the RNALater. Syn1-Tsc1$^{-/-}$ mice had phenotypes as previously described[114], with tremors, a hind-limb grasping reflex when suspended by their tail, seizures and a much lower weight than control littermates. In addition, Syn1-Tsc1$^{-/-}$ mice, when spun gently by their tails often had a a very large seizure resulting in death. Genotypes were confirmed with PCR using a standard genotyping protocol and a set of previously described primers[114].

All experiments were carried out with the approval of the Committee on Animal Care at the Massachusetts Institute of Technology.

2.2.2 Library preparation and sequencing

Cortices were placed in 1 mL Qiazol and homogenized and total RNA was extracted following the manufacturers protocol. Contaminating genomic DNA was removed from the total RNA by treatment with ten units of DNAse 1 (Roche) following the manufacturers protocol. Total RNA was further cleaned up using the RNeasy® MinElute® cleanup kit (Qiagen #74204) to remove contaminants left over from the extraction process following the manufacturer's protocol with the exception of rinsing the pellet with ethanol three times. Total RNA was eluted in 30 µl of RNAse free water. The purity of the samples was analyzed using a NanoDrop spectrophotometer (Thermo Scientific) and samples with 260/280 ratios less than 1.8 or 260/230 ratios less than 2.0 were subjected to a second round of cleanup. After cleanup the samples were run on an Agilent 2100 Bioanalyzer and samples with a RNA integrity number of less than 9, which indicates degradated RNA, were discarded. Purified total RNA samples were stored at -80 °C until library creation.

Libraries were created for paired-end sequencing on the Illumina HISeq using

the Illumina TruSeq v2 kit, following the manufacturer's protocol. Libraries were created in two batches, with equal control and Syn1-Tsc1$^{-/-}$ samples in each batch. Equal samples from control and Syn1-Tsc1$^{-/-}$ litters were included in each batch. For sequencing, in one batch, two control and two Syn1-Tsc1$^{-/-}$ samples were multiplexed on a single lane. In the other batch, a single control and Syn1-Tsc1$^{-/-}$ sample were run in separate lanes. Samples were sequenced at the BioMicroCenter at MIT, sequencing 40 basepairs from each end of the reads.

2.2.3 Informatics analysis

Alignment and differential expression

The RNA-sequencing reads were processed using the RNA-seq pipeline implemented in version 0.8.3 of the **bcbio-nextgen** analysis project. Briefly, poor quality bases with PHRED scores less than five[105], contaminant adapter sequences and polyA tails were trimmed from the ends of reads with cutadapt[109] version 1.2.1, discarding reads shorter than twenty bases. A STAR[48] index was created from a combination of the *Mus musculus* version 10 (mm10) build of the mouse genome and the Ensembl release 75 gene annotation. Trimmed reads were aligned to the STAR index, discarding reads with ten or more multiple matches to the genome. Quality metrics including mapping percentage, rRNA contamination, average coverage across the length of the genes, read quality, adapter contamination and others were calculated using a combination of FastQC, RNA-SeQC[45] and custom functions from bcbio-nextgen and bcbio.rnaseq. Chapter 1 on page 11 of this work has more specific implementation details.

The features of the transcriptome were quantitated at the gene, isoform and exon level by three separate methods. Reads mapping to genes were counted us-

ing featureCounts[97] version 1.4.4, excluding reads mapping multiple times to the genome and reads that could not be uniquely assigned to a gene. Reads mapping to isoforms were counted using eXpress[147]. Reads aligning to individual exons, broken up into unique features, were counted using DEXSeq[7] version 1.12.1.

Differential expression at the level of the gene was called using DESeq2 version 1.6.3, filtering for genes using a BH corrected[15] false discovery rate (FDR) cutoff of 0.1. Splicing differential expression was called at the isoform level with EBSeq version 1.6.0, using a matrix of normalized effective counts for each isoform from eXpress and at the exon level using DEXseq. To control the false positive rate when calling differential isoforms, a transcript was called differentially spliced only if both EBseq and DEXseq called a differential splice in the same transcript with a FDR cutoff of 0.1.

A background set of 15,807 expressed genes in the cortex was constructed by filtering the normalized count data for genes with a mean count of at least ten. A set of transcriptionally regulated genes was made by compiling all genes differentially expressed with all genes with a splicing event, in total 407. The set of transcriptionally regulated genes were tested for overrepresentation in pathways using the expressed set of genes as a background using WebGestalt[89] with a FDR cutoff of 0.1.

RNA-editing

Using **bcbio-nextgen** version 0.8.3, STAR aligned reads were preprocessed by splitting alignments with deletions into two separate alignments to prevent the variant caller from calling spurious deletions at exon-exon junctions[78]. The genotypes of the preprocessed reads of individual samples were called with the GATK HaplotypeCaller[44] which was configured to output genomic variant calling format

(gVCF) intermediate files for joint calling at all positions with a variant in any sample[178]. These files contain both variants and probability information of reference calls. These intermediate files were then jointly called with GenotypeGVFs to produce a final set of squared off variants for all of the samples. Known mm9 coordinates of RNA editing sites from the DARNED and RADAR databases were lifted over to mm10 coordinates and combined into a set of 17,831 curated RNA editing sites. A→G and T→C variant calls from HaplotypeCaller were intersected with the curated editing sites to produce the final set of edited sites. The functional effect of the RNA edits was annotated with Variant Effects Predictor[190].

There is a lower level of editing that we can reliably detect. Assuming editing follows a binomial distribution, we need to have at least $n = \frac{4p(1-p)}{p^2}$ observatons to detect at least one edited read 95 % of the time for a fraction of edited transcripts p. We set the edit fraction to 0.1 and filtered out any editing events with less than 36 total observations in either the control or Syn1-Tsc1$^{-/-}$ samples to keep a set where we can reliably detect editing, if it is occuring, in both conditions. Editing events were flagged as differentially edited via the binomial test with a cutoff of FDR < 0.1 and a editing percentage difference $> 25\,\%$.

2.2.4 Calcium Imaging

At embryonic day 15.5, fetal brain cortices of Syn1-Tsc1$^{-/-}$ mice were dissected and collected. Concurrently, the rest of the brain tissue were used for genotyping after DNA was extracted using a commercial kit (Sigma Aldrich #XNAT2). Fetal cortices were digested with a solution containing papain and Dase for 25 min. Cells were dissociated using pipets. Cells were plated to a coverslip coated with laminin and poly-D-lysine. A DNA construct encoding GCaMP3 (a gift from Loren Looger,

Addgene plasmid #22692) was transfected at day 6-8 in vitro using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. Neurons were imaged at DIV24-30 using a 40x objective on a Nikon Eclipse E600FN confocal microscope. Tyrode solution without $MgCl_2$, was used, which consisted of (in mM) NaCl, 145; KCl, 3.0; HEPES, 10; Glucose, 10; Glycine, 0.005; $CaCl_2$, 2.6; 300 mOsmol; adjusted to pH 7.4. Serial images were taken at 512 x 512 pixels using a water immersion lens with a large pinhole. In a pilot experiment, various intervals (0.125-1 sec per frame) were tested and patterns of GCaMP3 signal intensity changes were comparable. Consequently, the interval at 1 sec per frame was chosen. For pharmacological treatments, neurons were treated with 0.1 µmol SB242084 (Tocris #2901) for 1-2 hours after baseline images were captured. In the second experiment, neurons were treated with 5 µmol Ro25-6981 (Tocris #1594) for 1 hour, then 0.1 µmol SB242084, and 50 µmol D-APV (Sigma-Aldrich #A8054) were sequentially added to the imaging solution at the interval of one hour between imaging. Ro25-6981 blocks the NR2B subunit of the NMDA receptor, SB242084 blocks the HTR2C receptor and D-APV blocks the NMDA receptor.

2.3 Results

2.3.1 Differential expression

We identified 251 genes at FDR < 0.1 as differentially expressed in the cortex of Syn1-Tsc1$^{-/-}$ mice compared to wild type littermates. (Table 2.1 on page 75). We also found 254 differential splicing events in between the two conditions. EBSeq called 2895 isoforms differentially expressed (FDR < 0.1, representing 2436 genes, whereas DEXseq called 549 exons differentially expressed at FDR < 0.1, representing

489 genes. Taking consensus calls between EBSeq and DEXseq left us with 254 differential splicing calls in 156 genes.

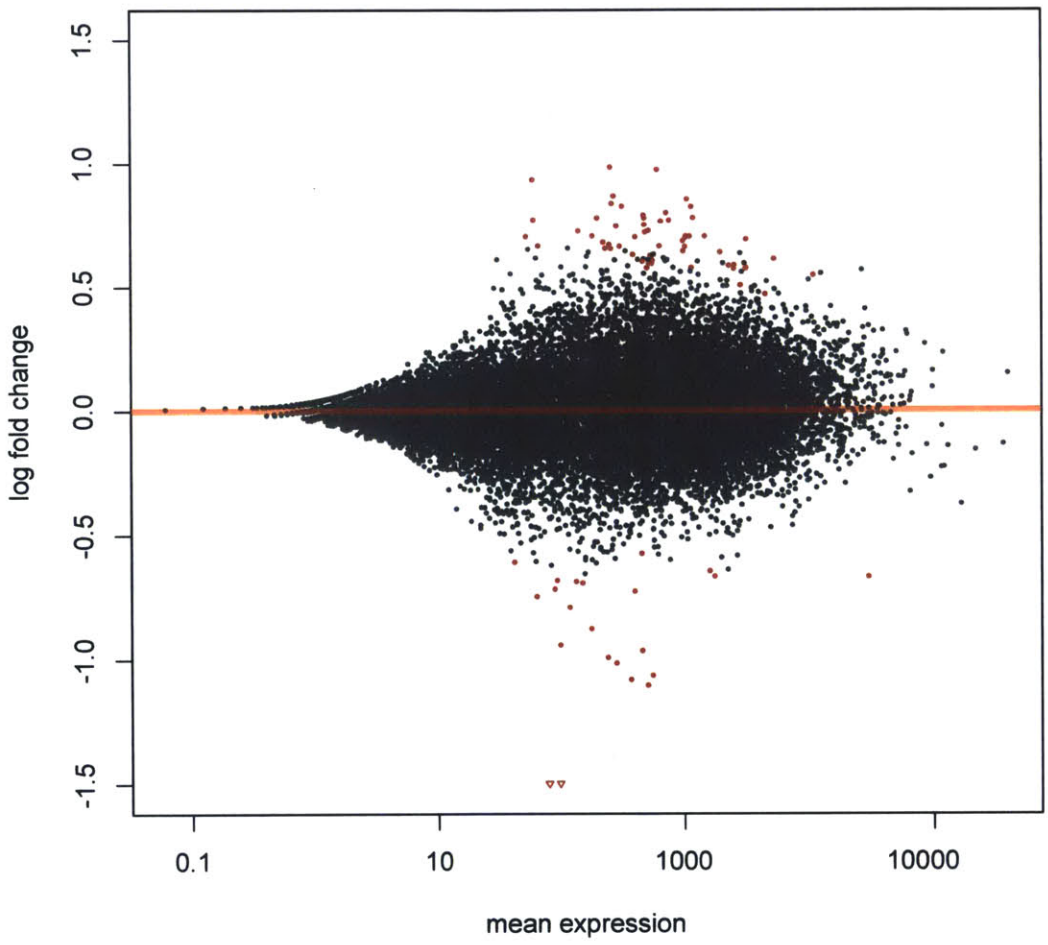


Figure 2-4: MA-plot of gene expression in the cortex of wild type vs. Syn1-Tsc1$^{-/-}$ littermates. 251 genes are differentially expressed at a false discovery rate of 0.10 (colored red). Genes expressed more highly in Tsc$^{-/-}$ mice have a positive $\log_2$ fold change.

symbol	$\log_2$ fold change	FDR	symbol	$\log_2$ fold change	FDR
Col1a1	0.89	0.00	Vwf	-0.56	0.09
Rec8	-1.09	0.00	Il17ra	0.58	0.10
Rab3b	-0.69	0.00	Ddr1	0.56	0.05
Fosb	0.70	0.04	Calr	0.46	0.09
Grm3	0.67	0.01	Hif3a	-1.38	0.00
Sst	-0.78	0.00	Nes	0.85	0.00
Wisp1	0.71	0.07	Ephb3	0.59	0.04
Cbfa2t3	1.08	0.00	Adamts4	0.77	0.01
Itih3	-0.62	0.08	Cnp	0.94	0.00
Homer1	0.79	0.00	Dio2	0.55	0.05
Hnrnph1	0.47	0.10	Bcas1	0.61	0.02
Fcrls	1.27	0.00	Phf21b	0.61	0.07
Scube1	0.63	0.02	Pltp	0.70	0.01
Erbb3	0.91	0.00	Ksr1	-0.84	0.00
Slc13a3	0.64	0.09	Kdm6b	0.52	0.10
Dusp14	0.72	0.00	P4ha1	0.75	0.00
Dcn	0.67	0.09	Dusp6	0.61	0.01
Epb4.1l2	0.50	0.08	Cfap54	-1.07	0.00
Mybpc1	-1.78	0.00	Unc5b	0.83	0.00
Ddit4	-0.70	0.01	Gamt	0.84	0.01
Cobl	0.49	0.08	Pgam2	-0.92	0.00
Trib2	0.69	0.01	Cacna1g	0.69	0.00
Serpina3n	-1.15	0.00	Galnt16	-0.77	0.00
Asb2	-0.84	0.01	Akr1c18	-1.16	0.00

symbol	$\log_2$ fold change	FDR	symbol	$\log_2$ fold change	FDR
Net1	-0.61	0.04	Tppp	0.48	0.07
Pcsk1	0.66	0.02	Crhbp	-0.63	0.09
Galnt15	-1.00	0.00	Dmtn	0.47	0.09
Dct	-0.67	0.04	Kcnv1	0.56	0.03
Myh9	0.60	0.02	Emp2	0.76	0.01
Arc	0.80	0.02	Tfrc	-1.15	0.00
Nr4a1	0.89	0.00	Synj2	0.66	0.00
Xdh	-0.86	0.01	Col11a2	0.97	0.00
Ttc39c	-0.66	0.08	Nr3c1	0.50	0.07
Slc22a6	0.85	0.01	Tnfrsf25	1.23	0.00
Drap1	-0.52	0.08	Gfra1	0.64	0.05
Alas2	0.82	0.01	Gli1	-0.76	0.04
Cpeb1	-0.65	0.01	Pdia4	0.73	0.01
Col3a1	0.99	0.00	Cdk18	0.63	0.08
Tmem63a	0.67	0.05	Mgst3	-0.51	0.08
Myoc	1.95	0.00	Gad2	-0.54	0.04
Lcn2	-1.34	0.00	Ermn	0.98	0.00
Gsn	0.72	0.01	Sord	-0.67	0.07
Lamp5	0.67	0.00	Chgb	-0.51	0.04
Mal	0.58	0.03	Pdyn	-0.63	0.04
Cd93	0.76	0.02	Car2	0.56	0.05
Slc7a11	0.89	0.00	Serpini1	-0.51	0.05
Olfml3	1.23	0.00	Tspan2	0.55	0.06
Trp53inp1	-0.64	0.10	Calb1	0.54	0.04

symbol	$\log_2$ fold change	FDR	symbol	$\log_2$ fold change	FDR
Nr4a3	1.04	0.00	Podn	0.68	0.09
Hivep3	0.54	0.03	Atpaf1	-0.61	0.06
Hspg2	0.76	0.01	Csmd2	0.61	0.01
Map3k6	-0.93	0.00	Sema3a	0.61	0.06
Cort	-0.99	0.00	Fosl2	0.72	0.00
Ociad2	-0.83	0.00	Cenpa	-0.82	0.02
Tpst2	-0.57	0.09	Snx8	-0.64	0.07
Rph3a	0.50	0.09	Col1a2	0.67	0.02
Gpnmb	-1.14	0.00	Klf15	-0.61	0.07
Bcat1	-0.51	0.07	Dmpk	0.68	0.03
Sult1a1	-0.76	0.03	Dgat2	-0.55	0.05
Cox6a2	-0.75	0.05	Crym	-0.72	0.00
Plxnb3	1.01	0.00	Plp1	0.65	0.00
Cdh11	0.56	0.04	Pllp	0.54	0.08
Cx3cl1	0.53	0.02	Mast3	0.49	0.04
Robo3	1.08	0.00	Mcam	0.95	0.00
Dock6	0.58	0.09	Myo1e	-0.61	0.06
Thsd4	-0.63	0.08	Col12a1	0.65	0.08
Trib1	0.81	0.00	Ephb1	0.62	0.03
Ugt8a	0.85	0.00	Fa2h	0.95	0.00
Egr3	0.96	0.00	9430020K01Rik	0.49	0.09
Scn10a	-0.91	0.00	Pla2g3	-0.99	0.00
Tdg	0.82	0.00	Tet3	0.50	0.08
Fam214a	-0.58	0.06	Eef2k	-0.52	0.09

symbol	$\log_2$ fold change	FDR	symbol	$\log_2$ fold change	FDR
Tanc1	-0.67	0.02	Lars2	-0.48	0.08
Pdzrn3	0.56	0.05	Midn	0.66	0.00
Cerk	-0.49	0.09	Dnajb5	0.65	0.00
Tm6sf2	0.69	0.05	Lrrk2	0.57	0.06
P2ry12	0.86	0.00	Rnf39	0.71	0.06
Mag	0.94	0.00	Arhgap33	0.56	0.10
Mycn	0.83	0.01	Clic4	0.58	0.03
Klf10	0.95	0.00	Cldn11	0.80	0.00
Inf2	0.62	0.03	Vstm2l	0.68	0.00
Egr2	1.43	0.00	Ostm1	-0.52	0.09
Sesn1	-0.57	0.05	Pcp4l1	-0.60	0.03
Egr1	1.14	0.00	Lpar1	1.03	0.00
Gpr126	-0.81	0.02	Rnf122	0.71	0.07
Gpr37	0.78	0.00	Thbs1	0.91	0.00
Ankrd6	0.65	0.02	Dnm3	-0.85	0.00
Stac3	0.69	0.09	Bcor	0.58	0.07
Spns2	-0.58	0.04	Ltbp4	0.54	0.04
Chrm4	0.64	0.03	Sh3pxd2b	0.60	0.06
Eml2	-0.50	0.06	Dlk1	-1.02	0.00
Htr2c	-0.75	0.02	Fmod	0.63	0.08
Mbp	0.51	0.03	Pkp2	-0.71	0.07
Map3k9	-0.47	0.09	Gpr153	0.87	0.00
Nrep	0.60	0.02	Tuba1c	0.89	0.00
Gjc2	0.93	0.00	Mdga1	0.67	0.01

symbol	$\log_2$ fold change	FDR	symbol	$\log_2$ fold change	FDR
S1pr1	0.62	0.01	Cirbp	-0.60	0.02
Ifit2	-0.88	0.00	Lrrc75a	0.65	0.08
Mc4r	-0.83	0.01	Sstr2	0.71	0.02
Phldb1	0.85	0.00	Gpr137c	-0.57	0.08
Pdp1	0.55	0.02	Opalin	1.06	0.00
D8Ertd82e	1.07	0.00	Scg2	-0.81	0.00
Islr2	-0.67	0.00	Gpr17	0.49	0.10
Hbb-bs	1.22	0.00	Cx3cr1	0.64	0.04
Hrh1	0.77	0.02	Cntn2	0.56	0.04
Mapk11	0.77	0.00	Spock3	-0.63	0.01
Nell1	0.65	0.01	Fmnl1	0.66	0.00
Prr18	0.71	0.02	Trnp1	-0.74	0.00
Ier5	0.84	0.00	Myh4	-1.16	0.00
Cdh24	0.74	0.02	Plekhh1	0.73	0.01
Cxcl12	0.68	0.00	Ppp1r1b	-0.64	0.01
mt-Nd6	-0.58	0.02	Cacng3	0.58	0.02
Zbtb16	-0.59	0.03	D430041D05Rik	0.55	0.02
Slc7a14	-0.69	0.00	Hba-a1	1.05	0.00
Serpinh1	0.88	0.00	Gad1	-0.56	0.01
Egr4	0.81	0.00	Zhx2	0.65	0.09
Tmem88b	0.76	0.01	Klhl40	-1.01	0.00
Kcng1	0.98	0.00	Thbd	1.04	0.00
Fnbp1	0.50	0.09	Mog	0.77	0.00
Rab26	-0.63	0.07	Wipf3	-0.76	0.00

symbol	$\log_2$ fold change	FDR	symbol	$\log_2$ fold change	FDR
Plcxd2	0.50	0.06	Tnnc1	0.79	0.03
Cecr6	0.56	0.04			

Table 2.1: Differentially expressed genes in Syn1-Tsc1$^{-/-}$ mice. 251 genes are differentially expressed with FDR < 0.1 in cortical neuron Syn1-Tsc1$^{-/-}$ knockout mice compared to sibling controls. Negative $\log_2$ fold change are genes more highly expressed in the Syn1-Tsc$^{-/-}$ mice compared to control mice.

2.3.2 Pathway analysis

The set of 407 differentially regulated genes is highly enriched for genes in the mTor pathway compared to the background set of 15,807 expressed genes (hypergeometric test, FDR < 1e-5), with an even distribution of genes regulated at the expression and splicing level (Table 2.4 on page 78). KEGG[86], Gene Ontology (GO) molecular function and and GO cellular part analyses show an enrichment in genes involving axon guidance, signalling and extracellular matrix interaction (Tables 2.2 on the next page and Table 2.3 on page 77).

2.3.3 Isoform differential expression

Since isoform differential expression is prone to generating false positive results, we combined the calls from two independent methods into one consensus call set. Differential exon usage was called using DEXseq version 1.12.1 using counts generated from featureCounts. 549 exons were differentially expressed at FDR < 0.1, representing

KEGG pathway	pvalue	FDR	genes
ECM-receptor interaction	0.0004	0.0048	Col1a2, Col1a1, Thbs1, Hspg2, Col3a1, Col11a2, Vwf
Protein digestion and absorption	0.0020	0.0120	Col1a2, Col12a1, Col1a1, Col3a1, Col11a2
Neuroactive ligand-receptor interaction	0.0033	0.0132	Grm3, Nr3c1, Htr2c, Sstr2, Chrm4, Hrh1, S1pr1, Lpar1, Mc4r
Axon guidance	0.0068	0.0204	Unc5b, Sema3a, Ephb3, Ephb1, Plxnb3, Cxcl12, Robo3
Amoebiasis	0.0477	0.0847	Col1a2, Col1a1, Col3a1, Col11a2
Tight junction	0.0476	0.0847	Epb4.1l2, Cldn11, Myh4, Myh9, Rab3b
MAPK signaling pathway	0.0494	0.0847	Dusp6, Cacng3, Pla2g3, Mapk11, Map3k6, Nr4a1, Dusp14, Cacna1g
Cytokine-cytokine receptor interaction	0.0620	0.0930	Tnfrsf25, Cxcl12, Cx3cl1, Cx3cr1, Il17ra
Calcium signaling pathway	0.04	0.08	Tnnc1, Prkcg, Cacna1g, Htr2c, Atp2b2, Grin1, Erbb3, Hrh1

Table 2.2: KEGG pathway analysis of differentially regulated genes. Several KEGG pathways are enriched (FDR < 0.1) comparing the set of differentially expressed genes to the background set of expressed genes. Pathways involved in cell motility, signal transduction in neurons, axon guidance and calcium signaling were differentially activated in Syn1-Tsc1$^{-/-}$ mice compared to wildtype controls.

492 different genes. Differential isoform expression was called using EBSeq version 1.6.0 with effective isoform counts generated from eXpress version 1.5.1. EBSeq called 2895 isoforms differentially expressed representing 2436 genes. Consensus calls between EBSeq and DEXseq were constructed by only flagging transcripts where both DEXseq and EBSeq predicted differential splicing as differentially expressed. This process left us with 254 differential splicing calls in 167 different genes.

2.3.4 Disease association

The set of differentially expressed genes and the set of differentially spliced transcripts were combined into a master set of genes undergoing transcriptional regulation. A set of autism, epilepsy and intellectual disability associated genes were compiled from DisGeNET[13], a database of gene-disease associations integrated from several sources. Table 2.5 on page 79 breaks down the genes that are transcriptionally regulated in those diseases.

GO term	pvalue	FDR
receptor activity	7.67e-06	0.0006
signaling receptor activity	8.35e-06	0.0006
transmembrane signaling receptor activity	2.65e-05	0.0007
signal transducer activity	3.20e-05	0.0007
molecular transducer activity	3.20e-05	0.0007
protein binding	2.15e-05	0.0007
G-protein coupled receptor activity	1.93e-05	0.0007
hormone binding	0.0001	0.0018
structural molecule activity	0.0002	0.0033
extracellular matrix structural constituent	0.0004	0.0053
extracellular space	2.49e-10	2.86e-08
extracellular region	2.83e-09	1.63e-07
extracellular region part	6.64e-09	2.55e-07
myelin sheath	2.72e-08	6.97e-07
axon	3.03e-08	6.97e-07
cell periphery	3.78e-08	7.25e-07
plasma membrane	1.32e-07	2.17e-06
cell surface	1.30e-06	1.87e-05
neuron projection	5.38e-06	6.19e-05
cell projection	5.36e-06	6.19e-05

Table 2.3: Gene Ontology (GO) term analysis of differentially regulated genes. Top ten GO terms for molecular function and cellular component enriched in the set of differentially expressed genes compared to the expressed background; in all there were 40 significant cellular component terms and 26 significant molecular function terms. Terms for signal transduction, interactions with the extracellular space, axon development and myelination are overrepresented.

gene	isoform	both
Thbd, Cxcl12, Gsn, Dcn, Trp53inp1, Gfra1, S1pr1, Ddit4, Fosb, Col1a1, Mapk11, Ksr1, Dusp6, Eef2k, Sesn1, Egr2, Map3k6, Nr3c1, Serpini1, Col1a2, Nr4a1	Trf, Gm20425, Gm12117, Dnm1, Gm11214, Gm3839, Gm7293, Gapdh, Iqsec1, Plekha2, Csnk1g2, 3110039M20Rik, Exoc1, Ppargc1a, Prkcg, Eif4g1, Rb1cc1, Foxg1, Gdi1, Zfyve28, Gnao1, Eng, Mapk10, Hspa8, Arpc3, Dnm2	Cacna1g, Egr1, Tfrc

Table 2.4: mTor signalling is differentially regulated in the Syn1-Tsc$^{-/-}$ mouse. The mTor pathway is overrepresented by genes with expression and splicing differences in the Syn1-Tsc$^{-/-}$ mouse, compared to the background set of expressed genes (hypergeometric test, $FDR < 0.0000135$).

autism	intellectual disability	epilepsy
Gpr37, Htr2c, Zhx2, Cux1, Nrxn2	Fosb, Calr, Dio2, Gamt, Tppp, Nr3c1, Alas2, Car2 Sema3a, Gpnmb, Dmpk, Sult1a1, Inf2, Zbtb16, Hba-a1, Gria3, Gdi1, Lamp2, Foxg1 Stxbp1, Eng, Grin1, Arhgef2, Wdr13, Dnm2, Dbn1, Metap2, Tbce, Iqsec2, Mapk10	Fosb, Calr, Grm3, Sst, Nes, Dusp6, Gamt, Cacna1g, Tppp, Pcsk1, Crhbp, Myh9, Nr3c1 Alas2, Gad2, Lcn2, Pdyn, Car2, Serpini1, Fosl2, Sult1a1, Plp1, Ugt8a, Egr1, Thbs1 Dlk1, Htr2c, Nrep, Sstr2, Cntn2, Cxcl12, Cacng3, Gad1, Gria3, Csnk1g2, Lamp2 Foxg1, Dapk1, Stxbp1, Eng, Dnm1, Grin1, Tpm3, Arhgef2, Cux1, Gnao1, Lphn3 Mapk10

Table 2.5: Autism, intellectual disability and epilepsy genes that are differentially regulated in the Syn1-Tsc1$^{-/-}$ mouse.

non-coding (2851)		coding (54)	
class	percent	class	percent
intron	38	missense	84
downstream	25	synonymous	15
upstream	12	in-frame deletion	1
non-coding	11		
3' UTR	4		

Table 2.6: RNA editing events found in Syn1-Tsc$^{-/-}$ and WT mice. 3169 editing events were identified in the WT and TSC samples, using known edit events from the DARNED and RADAR databases for mm10. VEP annotation revealed very few edit events with functional effects, with most edit events in either introns or genomic regions flanking a gene. Most edits in coding regions are missense mutations.

2.3.5 RNA editing

Distinguishing RNA-editing events from germline variants is a difficult problem, prone to generating many false positives[11] so we instead focused on quantitating known edit sites rather than predicting new sites. Variants were called from the aligned reads with the GATK Haplotype caller and A→G and T→C variants were intersected with a set of 17,831 curated RNA-editing events from the DARNED and RADAR RNA-editing databases. This left a set of 3,169 editing sites in 1,756 genes where at least one editing event was called in a gene in a single sample. The functional effects of these events were annotated with Ensembl's Variant Effect Predictor with the vast majority of edits occuring in the introns or regions flanking genes (Table 2.6).

Differential editing events between Syn1-Tsc1$^{-/-}$ and control mice were called using a binomial test with a FDR cutoff of 0.1 and a fold change cutoff > 0.25,

with the rationale that these events are more likely to be true events and to also be biologically relevant. This process identified 32 differential editing events between the Syn1-Tsc1$^{-/-}$ and control mice. These differential editing events are primarily in introns and the 3' UTR of genes but do not overlap splice sites or known miRNA target regions.

2.3.6 Syn1-Tsc1$^{-/-}$ mice have aberrant calcium signaling

Given that we observed differential expression of genes associated with calcium signaling pathways in the cortex of Syn1-Tsc1$^{-/-}$ mice (see Table 2.2 on page 76, we hypothesized that knocking down Tsc1 may lead to an alteration of calcium signalling in neurons. We cultured cortical neurons from Syn1-Tsc1$^{-/-}$ and control mice and transfected the cultures with the genetically encoded calcium sensor GCaMP3. Figure 2-5 on the next page shows example images of increased calcium flux throughout the dendritic tree in Syn1-Tsc1$^{-/-}$ neurons. Syn1-Tsc1$^{-/-}$ neurons had more frequent bursts of calcium throughout the dendritic tree, spiking close to 300 % more often (interspike interval: 13.5 seconds $\pm$ 3.4 seconds in Syn1-Tsc1$^{-/-}$ cells (n=5), 35.0 seconds $\pm$ 10.0 seconds in wild type cells (n=4), (p=0.047).

Previous work has shown deletion of Tsc2 causes overactive release of calcium from the ER in non-excitable cells in tuberous sclerosis tumors[132], indicating that increased calcium signalling may be a common feature of disrupting the mTor pathway.

Overactive serotonin signaling can cause autism-like symptoms in mice. [179]. In addition, work on the same conditional Syn1-Tsc1$^{-/-}$ mouse model where the deletion was restricted to only serotonin releasing neurons showed that Tsc1 deletion in those neurons alone is sufficient to cause the autism phenotype[113] but not sufficient to

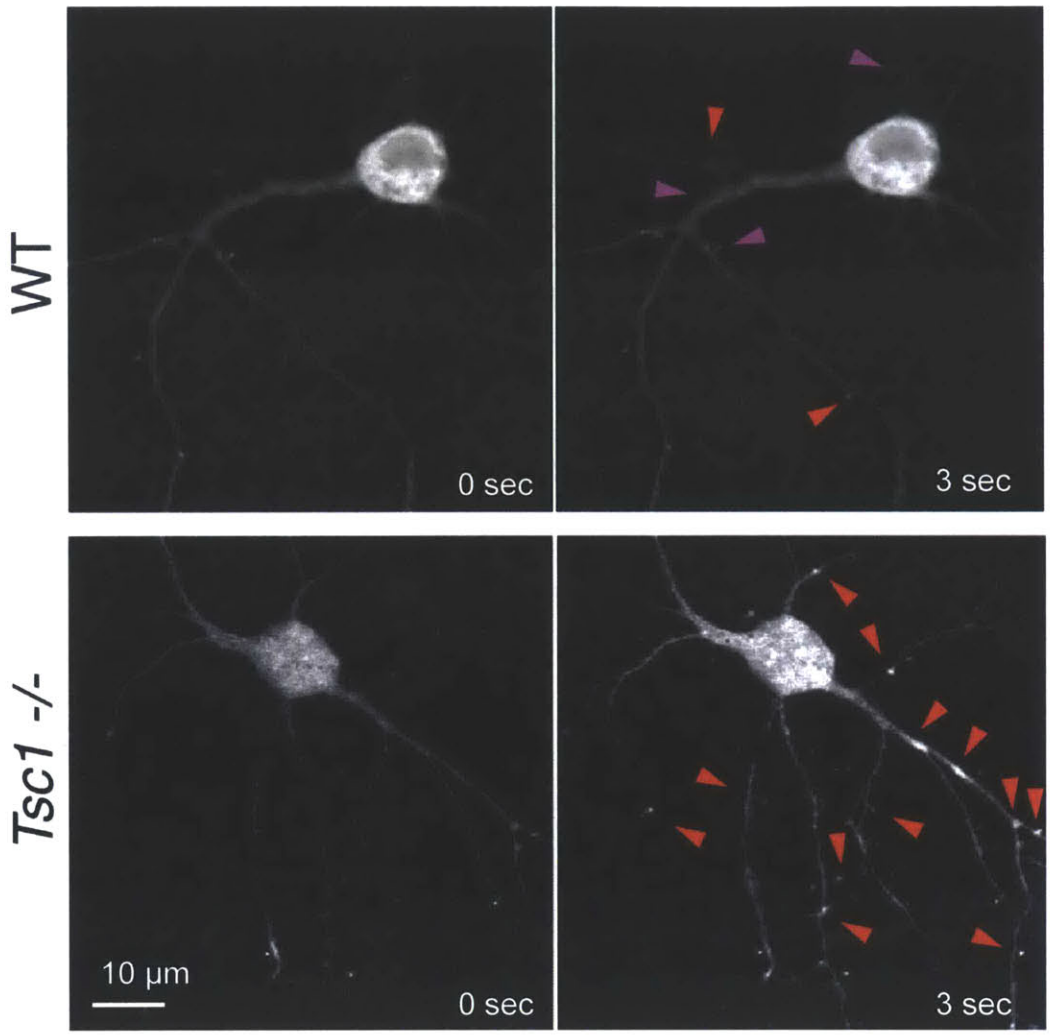


Figure 2-5: Cultured cortical Syn1-Tsc1$^{-/-}$ neurons have synchronized calcium flux throughout the dendritic tree (red arrows) whereas wild type neurons have more localized calcium flux (purple arrows show decreased calcium). Neurons were transfected with GCaMP3 and imaged at three second intervals. Synchronized calcium flux in the dendrites was much more frequent in Syn1-Tsc1$^{-/-}$ neurons with an interspike interval of 13.5 seconds $\pm$ 3.4 seconds (n=5) in Syn1-Tsc1$^{-/-}$ neurons and 35.0 seconds $\pm$ 10.0 seconds (n=4) in wild type neurons (p=0.047).

cause seizures. We hypothesized that overactive Htr2c receptor expression may be the cause of the overactive calcium signaling in the Syn1-Tsc1$^{-/-}$ neurons.

2.3.7 Htr2c blocker halts aberrant calcium signaling

Cultured wildtype and Syn1-Tsc$^{-/-}$ neurons transfected with GCaMP3 were imaged. Wildtype neurons had punctate calcium flux in the spines and not the shaft or soma whereas the Syn1-Tsc$^{-/-}$ neurons had calcium bursts in the soma and long the entire dendritic shaft. Blocking Htr2c in the Syn1-Tsc$^{-/-}$ cultures reversed the aberrant calcium bursts in the shaft and soma and restored the punctate spiking at the spines (Figure 2-6 on the following page).

2.4 Discussion

In this study we surveyed the transcriptional landscape of the cortex of mice lacking Tsc1 in neurons. To our knowledge, this is the first RNA-seq dataset of tuberous sclerosis in any organism. We assayed not only differential expression at the gene level but also splicing and editing differences caused by Tsc1 deletion. From this data we identified a set of 253 differentially expressed genes, 167 genes with differential splicing calls and a small set of 32 differential editing events when compared to mice with the neuronal loss of Tsc1.

Genes with differential transcription events were highly encriched for genes known to interact or be affected by the mTor pathway. We also identified dozens of genes differentially transcribed that are associated with autism, intellectual disability and epilepsy, three of the strongest neurological phenotypes associated with tuberous sclerosis. We identified several genes known to be involved in calcium flux in neu-

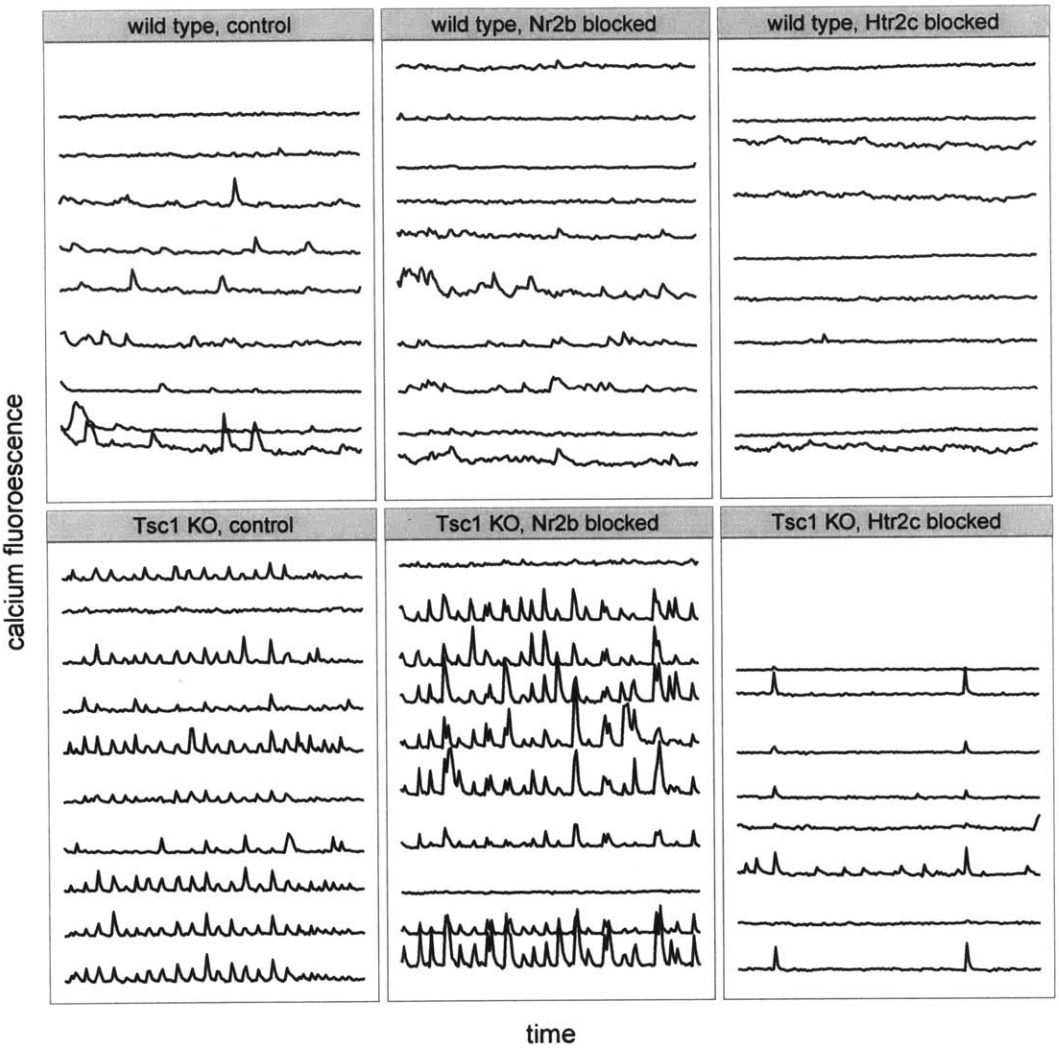


Figure 2-6: Data from a pilot experiment showing blocking the Htr2c receptor halts aberrant calcium spiking in Syn1-Tsc1$^{-/-}$ neurons. Treatment of cultured cells with SB24208, an inhibitor of Htr2c, halts the rapid calcium spiking in cultured Syn1-Tsc1$^{-/-}$ cortical cells. GCaMP3 Each treatment is on the same cell and each trace represents the calcium signal in a small random segment along the dendritic shaft. Treatment with Ro 25-6981, a selective antagonist of NR2b does not halt the calcium spikes.

rons both via the MAP signaling pathway and through surface receptors known to pass calcium across the membrane. We confirmed via calcium imaging that there is increased an increased rate of calcium spiking in neurons lacking Tsc1. The calcium spikes were not only more frequent and more diffuse, spreading throughout the dendritic tree rather than at punctate spots at the synapses.

There is no cure for tuberous sclerosis and current therapies are focused on managing the myriad of symptoms caused by tuberous sclerosis. One promising avenue is the treatment of low levels of rapamycin, an inhibitor of the mTor pathway to alleviate the effects of overactivation of the mTor pathway. This treatment has been shown to reduce autism symptoms and slow the growth of the cortical tubers in both mouse models and patients. The long term effects of rapamycin are unknown, however, and rapamycin treatment does not treat all of the symptoms or completely ameliorate the symptoms it does treat. It may be necessary to treat the symptoms of tuberous sclerosis with a cocktail of drugs. With this in mind, we used the set of differential transcription events we identified as an atlas of potential sites of therapeutic intervention. We focused on Htr2c, as it was strongly upregulated in mice lacking Tsc1 in neurons, has been shown to be be dysregulated in mouse autism models and passes calcium. We showed evidence that blocking Htr2c stops the diffuse calcium activation in tuberous sclerosis neurons, suggesting that Htr2c is responsible, at least in part, for the aberrant calcium activation in neurons lacking Tsc1. Interestingly, Htr2c is a dense source of miRNA, seven different miRNA are encoded in its introns[51]. Examining the affect of Htr2c upregulation on the expression of those miRNA would be a useful follow up experiment.

One aspect of tuberous sclerosis that is not addressed by any current therapies is the cortical disorganization that occurs in and around the cortical tubers. The cortical disorganization is due in part to improper migration during cortical development

and also due to a disorganized axonal arbor within and around the tubers. In the mice lacking Tsc1 only in neurons, the tubers are not seen, yet the mice have severe seizures. We identified a set of genes involved in axonal guidance and cell-matrix interaction that may cause seizures indepent of cortical tubers. It is possible that early intevention treating the expression of these genes could prevent some of the seizure of intellectual disability phenotypes seen in humans.

Future work involves following up on the effect of upregulation of Htr2c in the Syn1-Tsc-1$^{-/-}$ knockout mouse to determine if an Htr2c antagonist could be a therapy for the neurological symptoms of tuberous sclerosis. Several commercially available Htr2c antagonists exist. Risperidone, the first FDA approved compound for the treatment the behavioral symptoms of autism, is an antagonist of Htr2c among other receptors[111]. Further investigation into the effects of these antagonists on the aberrant spiking in cultures and on the epilepsy and autism phenotypes of the Syn1-Tsc1$^{-/-}$ mouse is warranted. In addition, it would be interesting to sequence RNA from postmortem tuberous sclerosis patients to create a more comprehensive catalog of differentially expressed, spliced and edited genes that may be involved in the disorder.

Chapter 3

Transcriptome independent retained plasticity of the corticocollicular projection of the mouse

3.1 Background

3.1.1 Superior colliculus

The superior colliculus in the mammal and the optic tectum in non-mammals integrates information from all visual areas in the brain and is important for the generation of saccades and orienting to stimuli[71]. The superficial layers of the superior colliculus receive projections from the retinal ganglion cells and from the visual cor-

tex. The deeper layers of the superior colliculus receive projections from from many other sensory areas of the brain including the ipsilateral auditory and association cortex, the somatosensory cortex, the inferior colliculus and others. Thus, the superior colliculus is anatomically situated to be the locus of the sensory integration involved identifying and orienting to salient stimuli from multiple senses. In addition to receiving projects with information useful for orientation, the deep layers of the superior colliculus project to motor nuclei in the brainstem controlling head and eye orientation, further supporting the role of the colliculus as a site of sensory integration and subsequent motor output [73]. Information flows to the colliculus from the various senses, computations are performed in the colliculus and the proper motor behavior is output in the form of head movement and eye orientation.

Supporting this notion, ablating the superior colliculus of the primate causes a permanent deficit in saccading to visual stimuli[4]. A study comparing lesions of the frontal eye fields to lesions of the superior colliculus in the primate showed only lesions of the superior colliculus affected saccades to visual stimuli[153], indicating the superior colliculus is important for saccade generation. The superficial layers of the superior colliculus are important for saccading to visual stimuli while the deeper layers are important for saccading to non-visual stimuli. For example, ablating the superior colliculus of the tree shrew, including the deep layers, results in a failure to follow, track or even orient to any stimuli, including putative threats, whereas ablation of only the superficial layers of the colliculus results in normal visually guided behavior [33]. Similarly, deactivating the cat superficial superior colliculus by cooling produces a lack of orientation to visual stimuli while leaving orientation to auditory stimuli unimpaired[100]. The separation of the functions of the superficial and deep layers of the superior colliculus are supported not only by lesion studies but also the anatomy of the colliculus.

Anatomically, the superior colliculus is a highly organized structure with inputs from the senses organized in topographic maps with respect to visual space and with projections from different senses terminating in topographically similar positions. On the surface of the colliculus, the topography of the contralateral retina is represented as a high fidelity map of nasal retina in the anterior region and progresses through the temporal retina in posterior regions[65]. The internal structure of the colliculus is broadly divided into two major laminae: a superficial layer (sSC) comprising of the stratum zonale (SZ), the stratum griseium superficiale (SGS) and the stratum opticum (SO) and deep layer (dSC) comprising of the stratum griseum intermediale (SGI), the stratum album intermediale (SAI), the stratum griseium profundum (SGP) and the stratum album profundum (SAP) [160]. The superficial layers of the superior colliculus recieve input from the retinae and the visual cortex [103] whereas deeper layers receive input from other sensory areas such as the auditory cortex and the somatosensory cortex [160]. This retinotopic organization becomes more diffuse in the deeper layers of the colliculus compared to the high fidelity map of space in the sSC. The connections within and between each layer of the superior colliculus are also highly organized, with each layer containing a small, anatomically disparate set of neurons. In the sSC, there are five major cell types: marginal cells in the SZ, horizontal and stellate cells in the SGS that process intralayer visual signals and wide field vertical and vertical cells that integrate information from the superficial and deeper layers of the colliculus. Visual information enters the sSC mostly from the contralateral retina via the SO. This retinocentric information is combined with information from the ipsilateral visual cortex via stellate cells and wide field vertical cells. This visual information is sent to the deepest layers of the colliculus via the vertical cells that combine it with information from afferents from other sensory systems[80]. Neurons deep in the superior colliculus output to motor nuclei in the

brainstem, eliciting eye movements and other orienting behavior. (Figure 3-1).

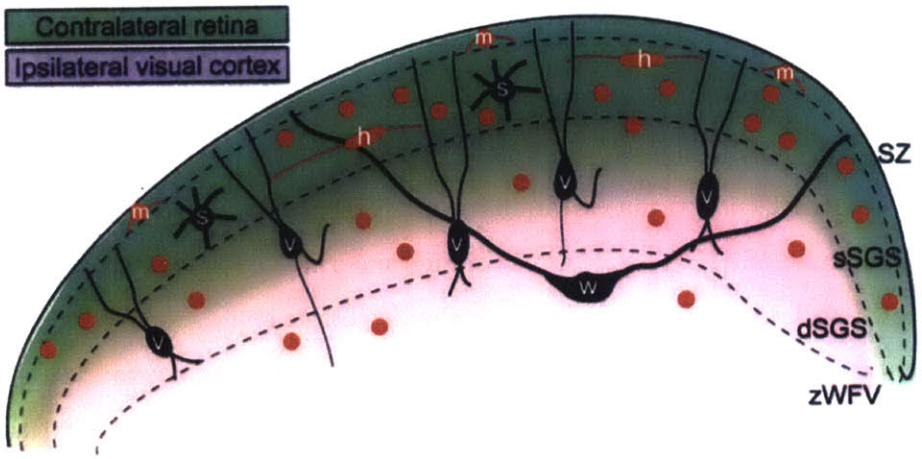


Figure 3-1: Cell types of the superior colliculus. A schematic of the superficial visual layers of the superior colliculus, with cell types labeled in their appropriate layers. Cells marked in red are inhibitory and cells marked in black are excitatory. m = marginal, v = vertical, s = stellate, h = horizontial, w = wide field vertical, SZ = *stratum zonale*, sSGS = *superficial stratum griseium superficiale*, dSGS = deep SGS, zWFW = zone of widefield vertical neurons. Used with permission from [136], Copyright Massachusetts Institute of Technology.

The establishment and registration of the sensory specific maps in the layers of the superior colliculus is important for its function and disruption of the topograph map results in an impaired ability to locate and saccade to many sensory stimuli. Quaia et al. showed that deactivating intermediate layers of the right superior colliculus with the GABA receptor blocker muscimol induced a reversible increase in response latency, a decrease in speed and a decrease in accuracy to stimuli in the left visual field. In addition this effect was strongest with stimuli in the zone of the left visual field that was deactivated in the right colliculus[143]. More recently a similar experiment was performed using optogenetic deactivation of neurons in the

monkey superior colliculus. This allowed for a precise deactivation of subpopulations of neurons in the superior colliculus and showed that saccade endpoints were shifted away from the deactivated visual area[34].

3.1.2 Retinotopographic map formation in the colliculus

Topographic maps are an efficient way to represent information through maximization of wiring economy[38] and many sensory systems use a topographic map as a first step of processing the information from the environment. The somatosensory system employs a topographic map of the body in primary somatosensory cortex[116], the auditory system has a frequency map in the cochlear nucleus [85] representing the locations of resonant frequency on the basilar membrane and the visual system has a topographic map of the retina projected on both the visual cortex and the superficial layers of the superior colliculus. The topographic map from the retina to the superior colliculus is formed and refined to its mature state through several gradual stages involving both activity dependent and independent processes.

The mature retinotopic map in the superior colliculus forms from projections of the optic tract to the stratum opticum with nasal-temporal retinal ganglion cells mapping to the anterior-posterior superior colliculus and dorsal-ventral retina mapping to lateral-medial superior colliculus. The question of how visual axons find the correct zone to terminate in is an interesting one. A solution was suggested by Sperry in the form of the chemotaxic hypothesis where multiple gradients of receptors on the growth cones of axons afferent cells and gradients of ligands for those receptors in the target neuropil could serve as signals for the axons to stop growing and to elaborate terminal arbors[161]. This is an interesting hypothesis because it requires only two orthogonal gradients of molecules to specify unique coordinates across the

neuropil, much like a single integer and a single letter can uniquely specify every square on a chessboard. The chemotaxic hypothesis was first borne out in the mammalian colliculus, with the initial, rough patterning of the map occuring prenatally as the result of two main orthogonal gradients of molecular cues, the Eph receptors and their ligands, the ephrins. EphA expressed on retinal ganglion cell (RGC) axons interacts with ephrin-A expressed in the colliculus, acting to repel the axons. The retinal ganglion cell axons enter the sSC from the anterior end, where ephrin-A is less expressed and RGC axons with high EphA expression stop expanding and terminate. Continuing along the anterior-poterior axis of the colliculus, the gradient of ephrin-A increases causing RGC axons with less and less EphA terminate[112]. A similar process guides the lateral-medial mapping of the dorsal-ventral retina, using gradients of the EphB receptor on RGC axons and ephrin-B expressed in the superior colliculus. The utility of chemotaxic gradients is very robust and is involved in not only the initial retinotopic map formation but in the registration of multimodal sensory maps on top of the retinotopic map in the colliculus. Similar gradients of these same or closely related molecules have been identified in both warm and cold blooded vertebrates in most sensory modalities. For example, a study from the Feldheim lab looking at the somatosensory whisker map to colliculus sensory input showed using knockout mice for ephrinA4 and ephrinA7 that gradients of ephrinA4 and ephrinA7 are necessary to register the somatosensory map on top of the retinotopic map in the superior colliculus[173].

While the chemotaxic hypothesis explains how the intial rough retinotopic map in the colliculus is formed it is does not completely explain how the mature, well-refined topographic map arises. In addition to molecular cues the retinotopic map is further refined through intrinsic spontaneous activity generated in the retina[188][35] and later via visual activation of the retina [29]. These two processes serve to fine-

tune the diffuse, exuberant elaborations generated through the initial chemotactic mapping.

In rodents the immature retinae generate spontaneous bursts of infrequent activity which spread across the retinae and serve to synchronize the firing of neighboring cells[188]. In mice this spontaneous activity begins at P6 and peaks at P10, petering out at P14, after the eyes have opened[120]. These spontaneous waves have been shown to be important for organizing multiple levels of the mouse visual system[1]. In the visual cortex it is well known that if this spontaneous synchronized activity is disrupted through blocking of activity via retinal TTX injection[165] the formation of ocular dominance columns is impaired. Similarly, in the colliculus, either blocking spontaneous activity in the retina or inducing dissynchrony among inputs[112] results in an abnormal retinotopic map formation in the colliculus. The work by McLaughlin is especially interesting– mice lacking the $\beta 2$ receptor of the neuronal nicotinic acetylcholine receptor lack retinal waves but have a normal amount of activity during the first postnatal week, when majority of the 2nd order refinemnt of the retinotopic map in the colliculus is occuring. $\beta 2^{-/-}$ mice do not display a tight termination zone by P8, with diffuse the retinocollicular connection being in approximately the correct area but with a diffuse, unpruned arborization. This indicates that it is the pattern of spontaneous activity in the retina, not the overall amount of activity that is important in the refinement of the retinocollicular connection, likely due to mechanisms of strengthing synaptic connections by co-active inputs as suggested by Hebb[74].

Disruption of spontaneous activity affects not only the anatomy of the retinotopic map map in the colliculus but also has a functional effect. $\beta 2^{-/-}$ mice have abnormal receptive fields in the colliculus but surprisingly cortical cells exhibit normal receptive field size[184]. The same study showed that $\beta 2^{-/-}$ mice perform poorly on an optokinetic task tracking the mouse's head movement following horizontal and

vertical sinusoidal gratings. Such tracking has been shown to test the subcortical visual centers[49]. Surprisingly, the same mice performed normally in a task measuring cortex-dependent spatial vision, suggesting that activity has a larger role in determining the quality of the functioning ouput in the colliculus than in the visual cortex.

The event of eye opening in rodents provides another form of activity that serves to instruct the final refinement of the topographic maps and visual function in the visual cortex and superior colliculus. Mice are born with their eyes closed until P13 but the eyelid is not completely opaque and is capable of transmitting up to 10% of available visible light [17]. This diffuse, attenuated light is capable of driving visually responsive cells in the cortex and the superior colliculus one or two days before eye opening. The light response through the closed eye is strong enough for visual neurons to be able to discern secondary features of the light such as place, orientation and even movement[93]. However, numerous studies in diverse vertebrates have shown that this small amount of visual stimulation is not enough to properly refine the cortical and subcortical visual connections and highly patterened activation after eye opening is very important for the functional development of the visual system in both humans and rodents. The visual neuropil requires light evoked activity within a critical window called the critical period in order to develop properly; if deprived of light evoked activity during this critical period the visual system development is retarded in a profound fashion. The critical period for the visual system ranges in species, for months in rodents to several years in humans [16]. Studies of human infants with uncorrected cataracts, where a cloudiness of the lens prevents most extrinsic light from accessing the retina, is an excellent model for studying the functional effects of early visual visual activity in humans. Lack of extrinsically driven visual activity has a profound effect on visual system func-

tion and children born with cataracts left uncorrected for 7-9 years are functionally blind even after cataract correction[110]. Even brief periods of early deprivation of patterened activity cause defects in lower-order processes such as spatial acuity, orientation selectivity, directional selectivity[22] and higher order cognitive processes such as discernment of implied forms and depth based on visual cues[59]. Subsequent visual exposure after the critical period has passed is unable to correct these deficits. In cats, lack of visual activity through dark rearing or lid sutures for several months induces gross impairments in visuomotor and visual acuitiy[122]; in monkeys dark rearing causes a suppression in the naso-temporal direction as well as a deprivation induced nystagmus[176].

In addition to functional effects on behavior, light deprivation causes many anatomical and electrophysiological changes in the visual system neuropil both in the visual cortex and the superior colliculus. Binocular activation of the eyes is necessary for proper development of binocular cells in the visual cortex; depriving an eye of visual experience removes the ability of the deprived eye to drive the binocular cells[9]. Furthermore, long term deprivation through dark rearing decreases the responses of single units in the visual cortex to orientation and direction[57], and causes the receptive field size of neurons in the visual cortex to remain immature[56]. Visual experience in the mouse is necessary for proper retinocollicular projection formation and reduction in visual experience through dark rearing reduces the density of fibres from the retina to the stratum opticum in mice[142]. However, unlike the visual cortex, early visual experience is not necessary for initial receptive field refinement in the superior colliculus but rather visual experience in adulthood past P60 is required for the maintenance of the receptive field size[28]. This finding was supported by work showing that dark rearing affects spatial tuning of single neurons in the superior colliculus, but not their orientation or direction selectivity or receptive field size[182],

even with long term deprivation up to P60. Early exposure to light prevents the destabilization of receptive fields in the colliculus due to long term deprivation in the adult, even though the receptive fields of both the dark reared and light reared animals refine to the same size[31]. This indicates that even relatively brief periods of light exposure can have long-term effects on the colliculus that are not necessarily reflected in the first-order functional or anatomical characteristics. These more subtle second-order effects could be molecular or transcriptional modifications in the neuropil of the superior colliculus.

Evidence for molecular modifications due to eye opening are strong in both the visual cortex and the superior colliculus, especially in the period following eye opening, though most work has been done in the visual cortex. Early studies found that blockade of the NMDA receptor (NMDAR) during the period of eye opening results in disruption of normal ocular dominance column formation and orientation selectivity in the visual cortex[14]. NMDAR activation fluxes Ca^{2+} into the cell, activating cAMP which in turn activates CREB leading to activation of the transcription factor CRE and transcription, raising the possiblity that visual activity resulting in activation the of NMDAR changes the transcriptional or molecular state in the visual system neuropil. Visual manipulations have confirmed this in the cortex, showing that brief visual experience induces the gene expression of a host of the intermediate early gene transcription factors[123]. In addition to activating signal transduction pathways downstream of the NMDAR, light exposure changes the composition of the NMDAR itself, inducing a subunit switch from predominantly NR2B to predominantly NR2A in both the visual cortex[27] and the superior coliculus[18]. In the superior colliculus, early light exposure also decreases phosphorylation of the NR2A receptor, which reduces the decay time of the NMDA current mediated by NR2A[168] and also affects the distribution of the NR2A receptor at the synapse, moving it to

the center of the synapse[169].

The observation that the NMDA receptor is a crucial mediator of light-induced changes in the visual areas and that exposure to light causes activation of several intermediate early gene transcription factors indicate that light exposure may cause the neuropil to enter a new state through modification of the the transcriptional program in the visual neuropil. The advent of relatively inexpensive transcriptome profiling procedures has made performing analyses regarding which transcripts may be differentially regulated feasible. One study showed that dark rearing from birth to P27 or monocularly depriving the animal causes the differential regulation of an thousands of transcripts in the primary visual cortex of mice[174]. Short term monocular deprivation of only four days differentially regulated more transcripts than long term deprivation, indicating that there may be compensatory mechanisms that occur with long term deprivation[174]. The differentially regulated transcripts included GABA receptors, NMDA receptors and AMPAR receptors but not metabotropic glutamate receptors indicating the excitatory and inhibitory circuits in V1 are regulated with light deprivation. Interestingly, the insulin growth factor (IGF) pathway was shown to be upregulated after monocular deprivation and IGF1 injection prevented the ocular dominance shift caused by monocular deprivation, giving a role for the IGF pathway in ocular dominance column formation. A different study using monoenculeation instead of monocular deprivation showed that 4 days of ME regulates a small set of tens of genes in the visual cortex of mice, mostly IEG transcription factors noted in previous, PCR-based studies [106][123]. It is hard to reconcile the results of these two studies, as there is very little overlap between the two gene sets identified as differentually regulated between the two studies, and one study finding two orders of magniude more differentually regulated genes than the other, despite very similar methods[175]. It is possible that CRE activation through ERK has not

had time to ramp up transcription regulation downstream of the IEGs[24]. Nevertheless, transcriptome profiling of the visual cortex has been fruitful, demonstrating the importance of the IGF pathway in ocular dominance column formation[174]. Multiple studies have confirmed, looking at single genes, differential regulation of several individual transcripts in the visual cortex through deprivation[90, 131] including a microRNA[115, 167]. In addition to regulation at the gene level a study showed that visual experience induces differential isoform expression of TrkB in the visual cortex[20]. These studies show that light activation of the retina can induce both transient and long term changes in the molecular composition of the synapse in the superior colliculus and the visual cortex.

Long term light deprivation also induces profound anatomical changes in the projection from the visual cortex to the superior colliculus. Eye closure until P16, three days after eye opening results in corticocollicular axons being stripped of their terminal arbors in the superfificial layers of the superior colliculus. In addition to changes in the axons themselves from the visual cortex, the axons of the visual cortex cause filopodia induction and new synapse formation in the superficial layers of the superior colliculus.[137] Thus eye opening and the onset of patterened activiation induces changes in the function, molecular composition and anatomy of the superior colliculus. Similar work showed these effects were reversible by opening the eyes[64] (Figure 3-2 on the following page).

Despite receiving direct input from the retina and the visual cortex, and evoked activity causing well documented alternations in the molecular composition and anatomy of the colliculus there have been no transcriptome-wide studies of the effect of light deprivation on the superior colliculus. Furthermore, almost no work has been done examining transcriptome-wide splice variant changes in any visual areas of the brain due to eye opening. Splicing differences are important to look at in the brain as

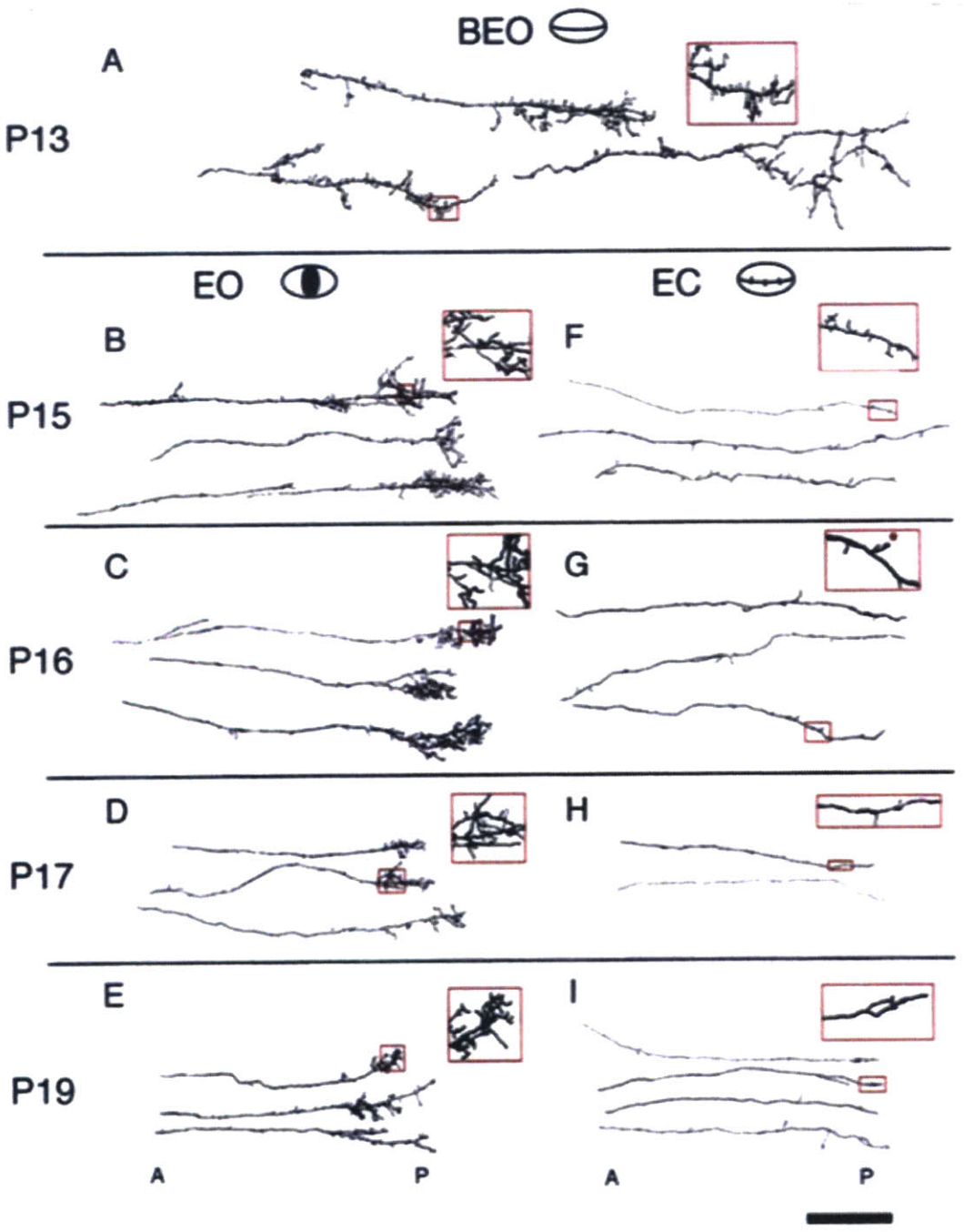


Figure 3-2: Eye opening refines the corticocollicular projection. Tracings of the corticollicular projection show laid out along the anterior-posterior axis show eye opening induces elaboration in the terminal zone and pruning of exuberant branches along the trunk of the axon. Eye closure induces a persistent pruning of the connections along the entire length of the axon. Used with permission from[64], copyright Massachusetts Institute of Technology.

alternative splicing levels area highest in the brain compared to the rest of the body and have been shown to have functional effects in some areas of the brain[21][121].

We chose to further investigate the effects of eye opening on the superior colliculus by examining transcriptional changes in the superficial superior colliculus for much longer periods of deprivation but before the end of the critical period. We hypothesized that long term deprivation may result in the neuropil of the superficial superior colliculus being transitioned to a new transcriptional state, and this new state is in part what causes light induced retina activation to have such profound effects on the molecular composition and anatomical structure of the superior colliculus.

3.2 Methods

3.2.1 Eye closure manipulation and colliculus dissection

Sprague Dawley rats were ordered from Charles River Laboratories and were kept in a clean facility on a 12 hour day/night cycle. The male Sprague Dawley rats in each litter were divided into eyes closed and eyes opened groups. At P12 to P13, the days prior to natural eye opening in the rat, the rats were brought down to our laboratory to perform eye-suturing of the eyes closed group. Rats of the eyes closed group were anesthetized with inhaled isoflurane first applied in a bell-jar and once breathing had slowed, delivered in a controlled fashion through a vaporizer attached to a a nose-cone. We sututed the eyelids closed with black 6-0 monofilament nylon suture using an interrupted stitch with at least five stitches per eyelid. On top of the suture we applied Mastisol® (Ferndale #0523), a translucent liquid adhesive adhesive to make it more difficult for the rats to damage and remove the suture. On top of the suture and the glue we applied Emla® cream, a topical lidocaine-based analgesic to

mitigate pain. We checked the sutured animals health status every day for four days after the surgery and observed no sign of distress in any of the animals. Every two days after P14 we reapplied the glue and after P18 we applied the glue every day as deemed necessary by visual inspection of the sutures. Any rat with the sutures removed was discarded from the study. Rats from the eyes open group and the eyes closed groups from the same litter were anesthetized the day of suturing but were not sutured or glued and Emla® cream was not applied. These procedures followed MIT IACUC-approved protocols.

At P20 we sacrificed both the eyes closed and the eyes open groups of animals and performed a dissection of the superior colliculus in a manner enriched for the superficial layers of the colliculus (Figure 3-3 on the next page). Rats were anesthetized with isoflurane in a bell jar and immediately decapitated and the head placed on dry ice. The brain was dissected on ice with tools treated with RNase*Zap*® (Ambion #AM9780) to protect against RNA degradation by RNases. The skull was removed and the brain was washed with ice-cold RNase free PBS prior to dissection of the colliculus. The cortex was removed and the superior colliculus was dissected and immediately transferred to 1 mL of RNALater®, a solution designed to inactivate RNAses and stabilize RNA to prevent degradation. Dissections were reproducible and uniform, representing a superficial-enriched dissection of the superior colliculus. RNALater treated colliculi were kept at 4°C for 24 hours and were then transferred to -20°C for indefinite storage.

3.2.2 Total RNA isolation and quality control

The superior colliculus was removed from the RNALater® and the and placed in 1 mL Qiazol® (Qiagen #79306) reagent. The colliculus was immediately homogenized

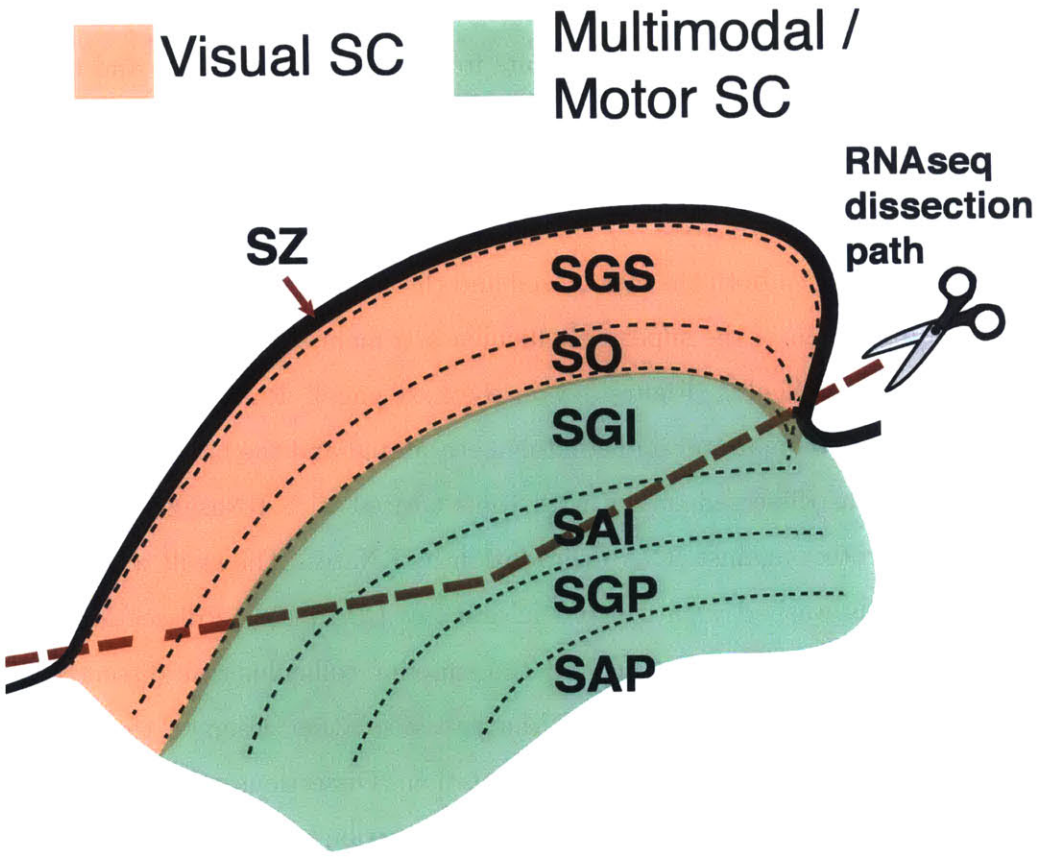


Figure 3-3: A schematic of the superior colliculus dissection, viewed as an idealized, sagittal slice. The superior colliculus was dissected to enrich for the superficial layers of the superior colliculus, capturing the *stratum zonale* (SZ), *stratum griseum superficiale* (SGS) and the *stratum opticum* (SO). Also captured with this dissection was a small amount of the intermediate layers, the *stratum griseum intermediale* (SGI) and the *stratum album intermediale* (SAI) which recieve multisensory inputs.

using a Tissue Tearor™ (Research Products International Corp #985371) for two minutes or until the sample appeared to be completely homogenized. 0.2 mL of chloroform was added to the sample and the sample was vortexed for 15 seconds until the solution was uniform. The sample was centrifuged for 25 minutes at 4°C and the aqueous layer was saved and added to 0.5 mL ice-cold isopropanol with 1 µl of glycogen to help visualize the RNA pellet. The sample was centrifuged for 15 minutes at 4 °C and the supernatant was discarded. The pellet was washed with 75 % ethanol and recentrifuged for 15 minutes at 4 °C. The supernatant was poured off and the pellet was allowed to air-dry inverted for 10 minutes at room temperature. After air-drying the pellet was redissolved in 48 µl DEPC-treated H_2O of 1x DNase buffer (Roche #04-716-728-001).

To clean up the isolated total RNA, we treated the RNA with 1 µl DNase 1 (ten units) (Roche) and 1 µl RNase inhibitor (ten units) for 20 minutes at room temperature to remove the genomic DNA. We then used the RNeasy® MinElute® cleanup kit (Qiagen #74204) to remove the enzymes, leftover DNA, salt and other contaminants from the isolated total RNA following the manufacturers instructions except for washing the RNA with 80 % ethanol three times instead of once. total RNA was eluted in 30 µl of RNAse free H_2O and stored at -80 °C until library creation. This cleanup step removes all RNA fragments that are less than 200 nucleotides in length, an important point to note for downstream analyses as there are many annotated small non-coding RNAs which fall into the group of RNAs under 200 nucleotides.

The purity of total RNA was assayed using a NanoDrop spectrophotometer (Thermo Scientific). The absorbance at 230, 260 and 280 ratios were measured from 1.5 µl of sample and samples with 260/280 ratios less than 1.8 and 260/230 ratios less than 2.0, indicating the presence of protein, phenol or other contaminants, were

re-phenol extracted and purified as above. Samples not conforming to the 260/280 and 260/230 ratios after reextraction were discarded and their sister sample from the litter pair was also discarded. 2 µl of the samples were then run on a Agilent 2100 Bioanalyzer to estimate the concentration and degradation of the isolated total RNA. Samples with RNA integrity numbers less than 9, indicating degradation of the RNA were discarded along with their sister sample. Samples were stored at -80 °C until library creation.

3.2.3 cDNA Library creation and sequencing

cDNA libraries for paired-end sequencing on the Illumina Genome Analyzer II were prepared following the Illumina protocol (#1004898 Rev. D) with a few alterations. Briefly, mRNA was bead-purified from total RNA using Sera-Mag poly-T oligo attached beads and 1 µl of purified mRNA was saved and run on an Agilent Bioanalyzer to check for yield and RNA integrity. RNA was fragmented for 2 minutes at 94 °C, converted to cDNA, end repaired and the 3' ends were adenylated. Adaptors were ligated and the cDNA was run on a 2 % agarose gel. The band at 200 base pairs was cut out and purified. The resulting product was PCR amplified for 10 cycles using primers against the Illumina adaptors. The final library was run on an Agilent Bioanalyzer to confirm proper size selection. A total of six samples, three eyes open control samples and three eyes closed samples were submitted to the BioMicro Center at MIT for sequencing on the Illumina Genome Analyzer II with 36 base pair paired-end reads.

3.2.4 Informatics analysis

We processed the RNA-sequencing data using the RNA-seq pipeline implemented in version 0.7.9a of the bcbio-nextgen analysis package. Briefly, we trimmed off poor quality ends with AlienTrimmer[43], using a cutoff of phred[55] score of 5 or less[105] and trimmed portions of reads and anything after it matching the first 13 bases of the Illumina universal adapter sequence to remove read-through contamination caused by the read length being longer than the insert size for a fragment. We also trimmed polyA and polyT homopolymer sequences from reads the 5' ends of reads.

A STAR[48] index was created from a combination of the *rattus norvegicus* genome, build rn5 and the Ensembl release 74 gene annotation. Reads were aligned with STAR version 2.3.14z using the default settings, with the exception of filtering out reads mapping more than 10 times to the genome. Counts of reads mapping to genes in the Ensembl annotation were calculated using FeatureCounts[97]. Quality metrics including mapping perecentage, rRNA contamination, coverage, read quality, adapter contamination and others were calculated using a combination of FastQC, RNA-SeQC[45] and custom functions using the **bcbio-nextgen** and **bcbio.rnaseq** packages. Chapter 3 of this work has more details.

3.2.5 Differential expression

Differential gene expression was performed with DESeq2 version 1.4.1 on using R version 3.1.0 using a blocked design taking into account litter and deprivation status of the mouse. DESeq2 was chosen based on simulations which showed it has the best performance out of several popular algorithms tested for simple two-factor comparisons and two-factor paired comparisons. Using data simulated to have the same distribution as our count data, we show that these procedures can expect to be

fold change	sensitivity	specificity
2+	0.96	0.95
1.5-2.0	0.37	0.75
1.1-1.5	0	0
1.05-1.1	0	0

Table 3.1: Power estimation of the eyes open vs. eyes closed experiment at a range of fold changes. Simulating a RNA-seq experiment using a similar transcript abundance distribution, sample size and biological variation shows that an experiment similar in size to the colliculus experiment should be expected to pick out fold changes twofold or greater reliably but fail to detect more moderate fold changes. For each range of fold changes the true positive rate (sensitivity) and false positive rate (specificity) are calculated. Fold changes greater than two-fold have high sensitivity and specificity but anything below that is likely to be missed for an experiment of this size.

able to pick out a large percentage of the highest simulated fold changes. However we will miss most moderate changes with our experimental setup (Table 3.1). See Chapter 3 for a description of the simulation with justification for choosing DESeq2 over the other differential expression callers.

Exon level counts were produced and analyzed using DEXSeq version 1.10.8 using a similar model design as the gene-level differential expression. Differentially expressed exons were called with an FDR cutoff of 0.1 and filtered for exons with an absolute fold change of 1.5 or greater, since simulations show that this is the limit of detectability for an experiment of this size.

3.2.6 Corticollicular projection mapping and quantitation

Projections from the visual cortex to the superior colliculus in rats were labelled and reconstructed as described in [64]. Briefly, cortical afferents to the sSC of rats were

labelled with DiI soaked gelfoam inserted below the pial membrane over the visual cortex from P17-P20. After two days of gel foam placement, rats were anesthetized and perfused with 4% PFA and post-fixed for 24 hours. 150–200 µm thick slices were cut from the sagittal plane of the neocortex and midbrain and sections were collected and mounted with Fluoromount. Stacks of images were collected on a Nikon PCM2000 (MVI) confocal scanning microscope using a 40X objective, taking 1.5 µm z-steps. Stacks of images were flattened and compressed into a single 2D projection of the 3D stack. The colliculus was divided into five quadrants from the anterior to posterior end and the number of branch points in each quadrant was counted as a measurement of arborization. Example images from this process appear in previous work[64], here we quantitate the entire dataset from that work, including the EC→EO reopening an EO→EC closure. EC→EO rats had their eyes closed from P13-P19 and were opened for two days from P19-P21. EO→EC rats had their eyes opened from P13-P19 and were closed for two days from P19-P21.

3.3 Results

3.3.1 Corticocollicular projection remodelling

As reported previously[64], eyelid suturing until P20 has a profound effect on the axonal branching of the corticocollicular projection in the colliculus. Normal eye opening results in exuberant branching of the primary axon from the visual cortex to the superficial layers of the superior colliclus; suturing the eyes closed until P20 results in a pruning of the exuberant branches. We quantitated the effect of inducing pruning by closing the eyes of a normal animal or reversing pruning by opening the eyes of an eyes-closed animal on the degree of branching of axons in the corticotectal

projection to the superficial superior colliculus (Figure 3-4 on the next page). An analysis of variance (ANOVA) shows significant variation ($p < 0.01$) among the eye state groups. A post hoc Tukey test showed manipulating the eye state from the normal eye opening (EO) causes significant differences when the eyes are closed (EC) ($p < 0.01$) but not when the eyes are initially closed and then opened (ECtoEO). This supports the conclusion of [64] that reopening the closed eyes restores the structure of the corticocollicular projection to its normal state.

3.3.2 Sequencing quality control

Sequencing libraries contained 50-70 million reads per lane and about 80 % of reads mapping to known genes using the Ensembl release 74 of the rat annotation. After trimmed mean of M-values (TMM) normalization[149], samples had a similar read-count distribution across genes (Figure 3-5 on page 110. Samples had a high Pearson correlation to each other and clustered together based more on the litter they were drawn from, not the experimental condition (Figure 3-6 on page 111). Observing clustering by litter validates the choice to litter-match the samples; without litter matching the major component of the variation between the samples would be unable to be corrected for.

3.3.3 Differential gene expression

The MA-plot (Figure 3-7 on page 112) of closed vs. open samples shows there is are only moderate fold changes in genes between the eyes closed and eyes open state, where the eyes open animals have had five days of eye opening.

A small set of genes that were flagged as differentially expressed is shown in Table 3.2 on page 113. Interestingly two of these genes, Mt-nd4l and Mt-nd6, code

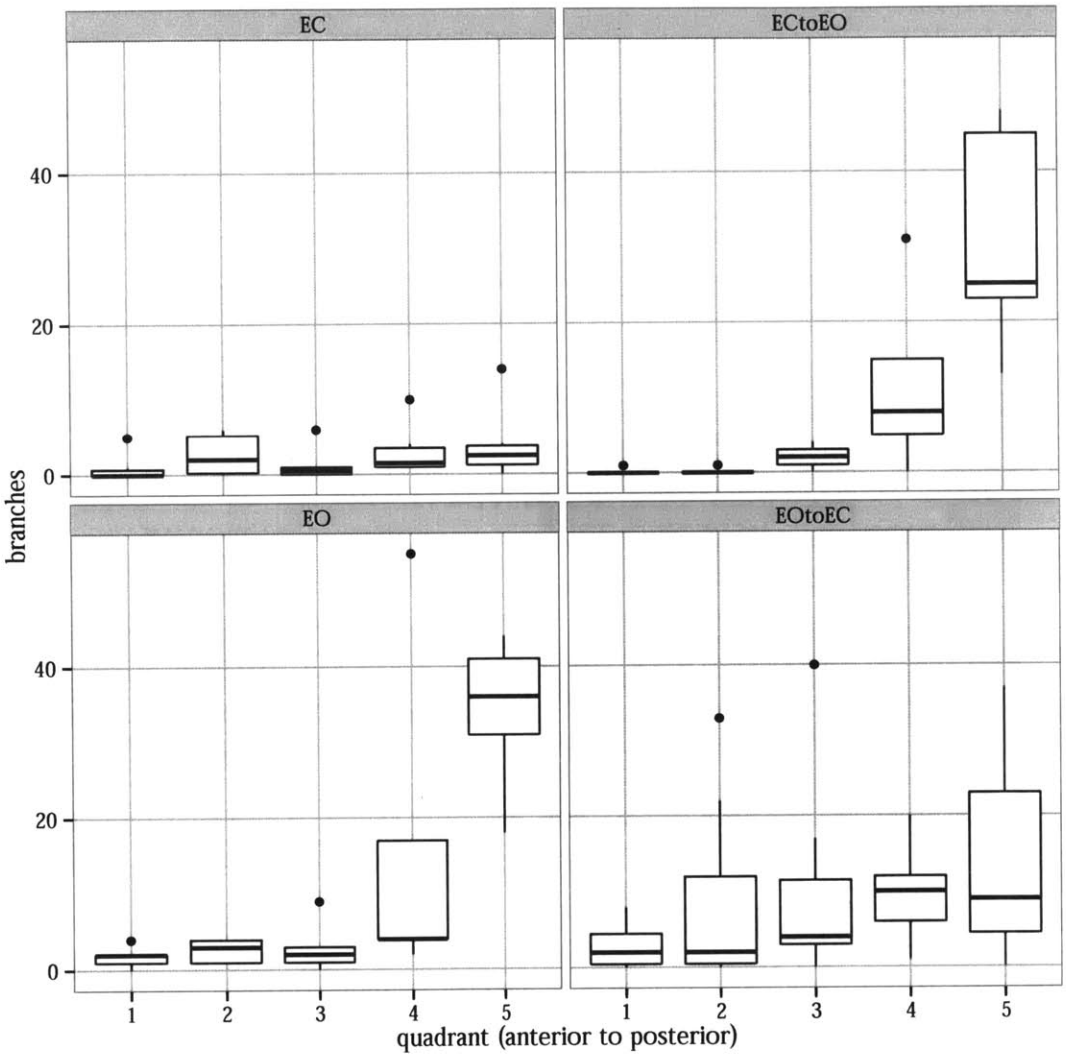
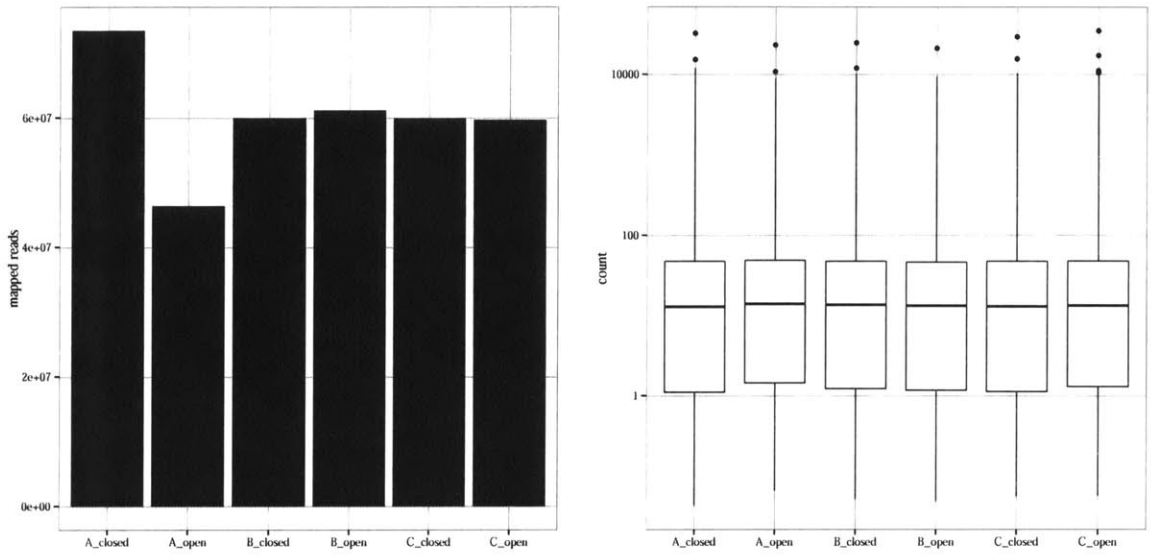


Figure 3-4: Effect of eye closure state on corticocollicular arbor development. Raising rats with eyes sutured closed from P13 until P19 (EC) prevents the exuberant axonal branching of the corticollicular projection from the visual cortex to the posterior superficial superior colliculus compared to the mouse with normal eye opening (EO). Reopening the eyes for two days at P19 (ECtoEO) restores the EO corticocollicular axonal topography. ANOVA with a post hoc Tukey tests shows manipulating the eye state from the normal eye opening (EO) causes significant differences when the eyes are closed (EC, p value < 0.01) but not when eyes are closed and then reopened (ECtoEO).



(a) **(b)**

Figure 3-5: **a)** The libraries have a varying number of total mapped read counts. **b)** Trimmed mean of M-values (TMM) normalization[149] effectively normalizes gene expression. Samples have a (litter)_(closed/open) naming scheme.

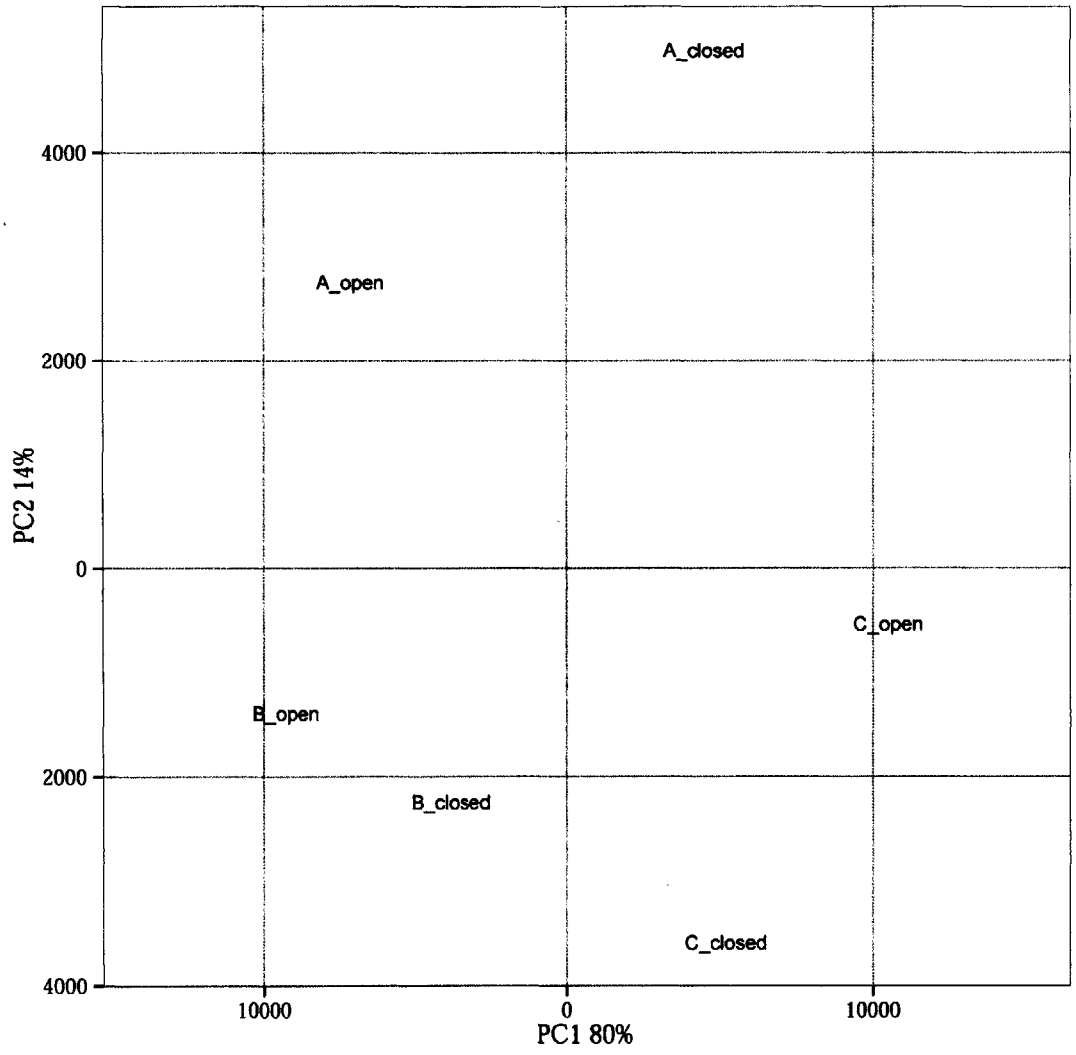


Figure 3-6: Clustering of gene expression in the rat colliculus. Multidimensional scaling (MDS) estimates the similarity between high-dimensional samples. MDS loosely clusters the samples into pairs based on litter (A, B and C), justifying including litter in the differential expression model. Samples have a (litter)_(closed/open) naming scheme.

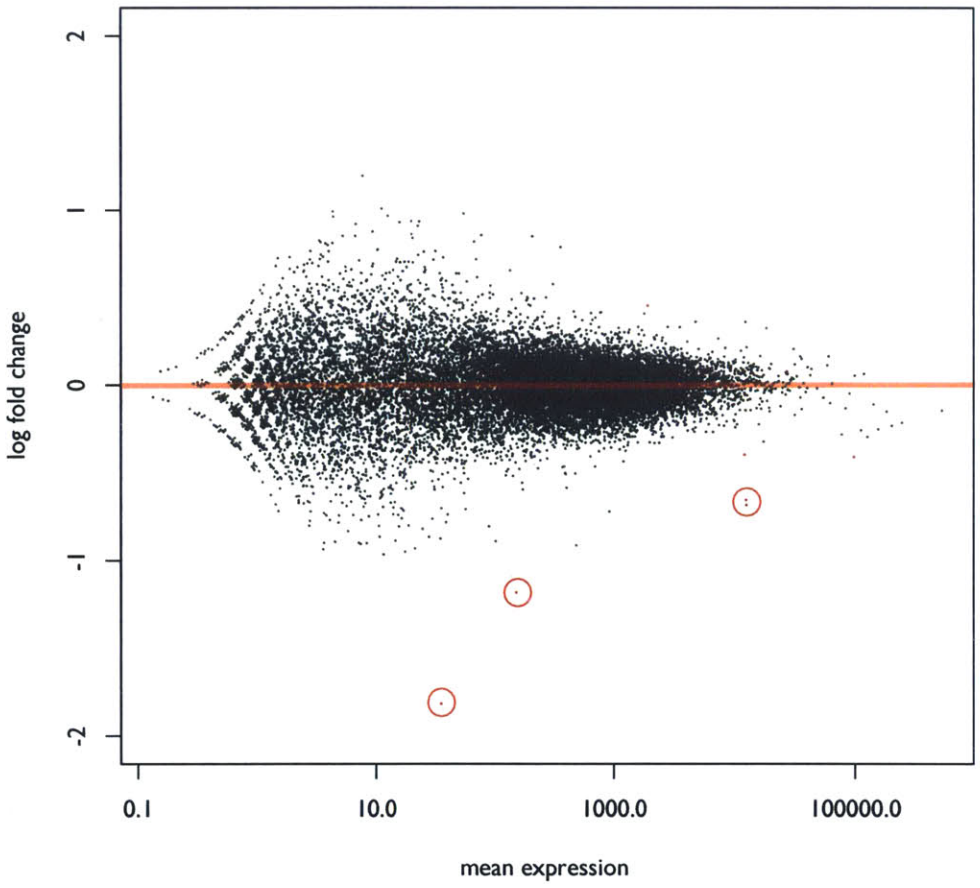


Figure 3-7: MA-plot (mean expression vs $\log_2$ fold change) between eyes open and eyes closed rats. The fold change between the two conditions is very modest. Points colored red are significantly different with a FDR < 0.1. Positive $\log_2$ fold changes are genes that are more highly expressed in the eyes closed compared to the eyes opened animals.

gene id	symbol	$\log_2(\frac{closed}{open})$	FDR
ENSRNOG00000011292	Col1a2	0.4540360	4.087336e-03
ENSRNOG00000033299	Mt-atp8	-0.6846459	1.200520e-08
ENSRNOG00000031053	Mt-nd4l	-0.3978127	3.562899e-03
ENSRNOG00000029042	Mt-nd6	-0.6556634	6.030092e-13

Table 3.2: Differentially expressed genes with with FDR < 0.05. A model was fit using DESeq2, taking into account the within-litter variation and calling difference between eyes closed and eyes open animals. Positive $\log_2$ fold changes are genes that are more highly expressed in the eyes closed compared to the eyes opened rats. Mt-atp8, Mt-nd4l and Mt-nd6 are all involved in the electron transport chain.

for NADH dehydrogenase, an enzyme which starts the electron transport chain by catalyzing the oxidization of NADH by ubiquinone[5]. A third protein, Mt-atp8 codes for mitochondrial ATP-synthase, another protein involved in the electron transport chain.

3.3.4 Possible X-linked cofactors in LHON

Previous work has shown eye opening to increase the energy load on the visual system neuropil[189]. Loss of function mutations in either Mt-nd4l or Mt-nd6 causes Leber's hereditary optic neuropathy (LHON)[26]. LHON causes the selective death of RGCs in the fovea.

An interesting feature of LHON is that although it is primarily a mitochondrial disease there is a much higher prevalence in males and an incomplete penetrance in affected families. One possible explanation for this phenomenon is that there may be a cofactor on the X chromosome that is also mutated in affected male individuals[76], [82],[155]. Using this additional information we looked at genes on the X chromosome

which did not reach significance but had a confidence interval highly skewed in one direction. We identified several genes meeting these criteria, including genes involved in the electron transport chain (Cox7b) actin dynamics (Tmsbx, Tsmbl1, Smarca1) and oncogenes (Rab9b, Mcts1).

3.3.5 Differential exon expression

We found 212 differentially expressed exons at FDR < 0.1, but none differentially expressed at levels 1.5 fold or greater. (Figure 3-8 on the following page) shows very clearly that these samples are highly uniform even when considering individual exons.

3.3.6 Dopamine receptor expression in the superior colliculus

Evidence for the expression of the dopamine receptor Drd1 in the superior colliculus is mixed. Previous work looking at Drd1, Drd2, Drd3 and Drd5 dopamine receptor mRNA expression in the rat CNS found evidence only of Drd2 in the superior colliculus[108]. Another study looking at Drd1 and Drd2 expression found only Drd2 in the superior colliculus[187]. An earlier study looking specifically for Drd1 found no Drd1 mRNA but did find the presense of the Drd1 protein, as it bound a radiolabeled Drd1 antagonist[107]. In this work we show that the composition of dopamine receptor mRNA is much more diverse than previously reported. We detect Drd1, Drd2, Drd3 and Drd5 at low levels in the superficial superior colliculus (Figure 3-9 on page 117). tdTomato expressing Drd1 neurons and EGFP expression in Drd2 neurons show a segregation of Drd1 and Drd2 into distinct layers of the superior colliculus. Drd1 is expressed in the all three sublaminae of the superficial

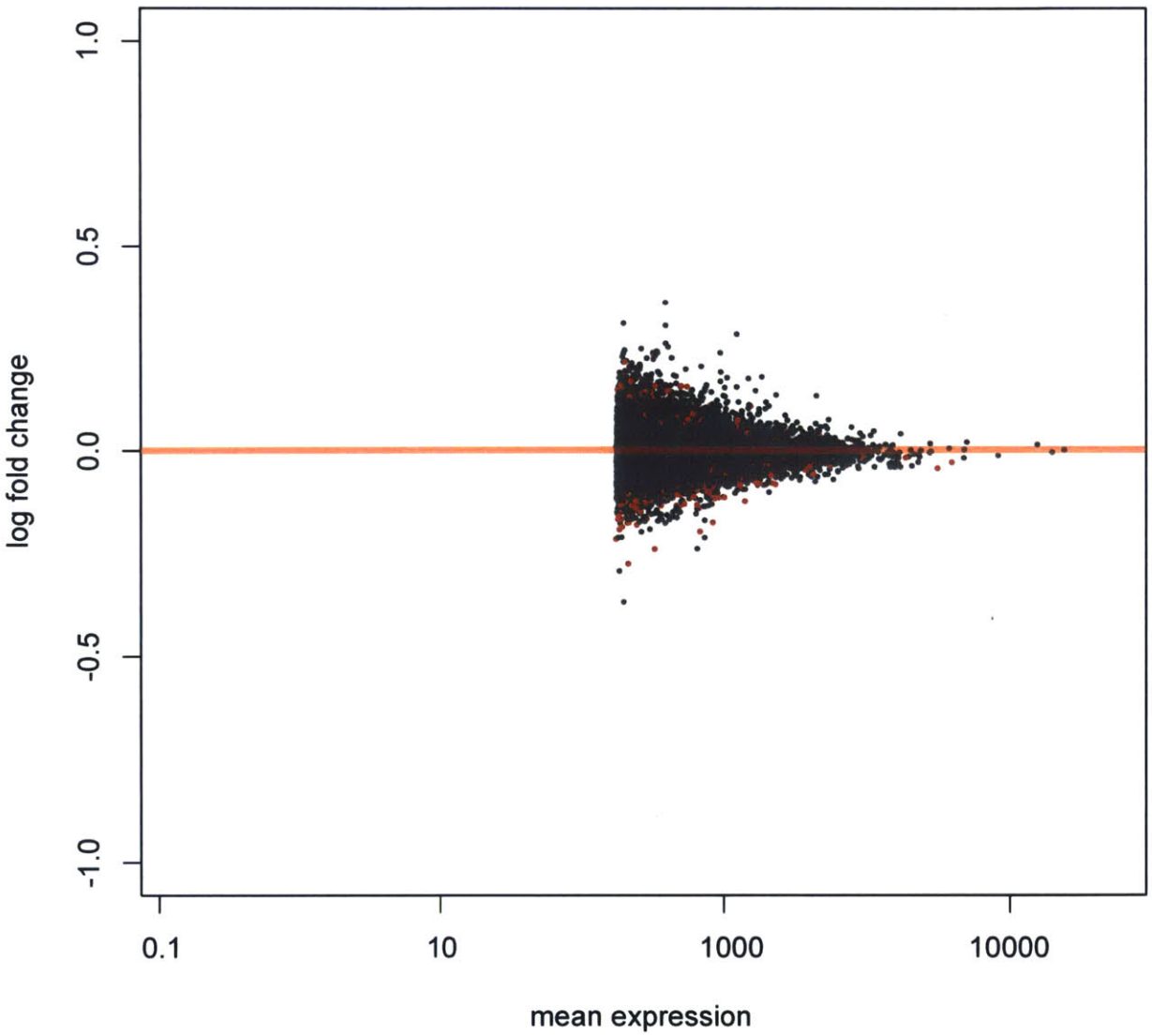


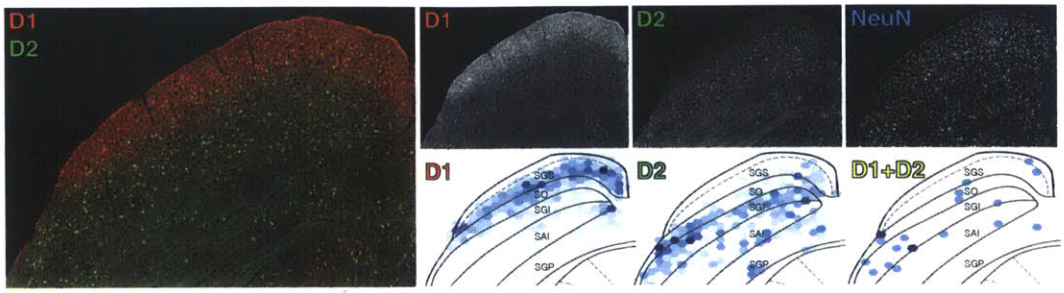
Figure 3-8: MA-plot of differential exon analysis. Fold changes in exon usage are even more moderate than the gene-levels changes. Although exons are called differentially expressed, they are at a very low fold change (red points). These are likely to be false positives so we discarded them. Positive $\log_2$ fold changes indicate exons expressed more highly in the eyes closed samples. Noisy exons with low expression were filtered out from the analysis.

superior colliculus and Drd2 is expressed in intermediate layers of the superior colliculus. There is sporadic overlap between Drd1 and Drd2 positive cells, mostly in the intermediate layers of the superior colliculus (Figure 3-9a on the following page).

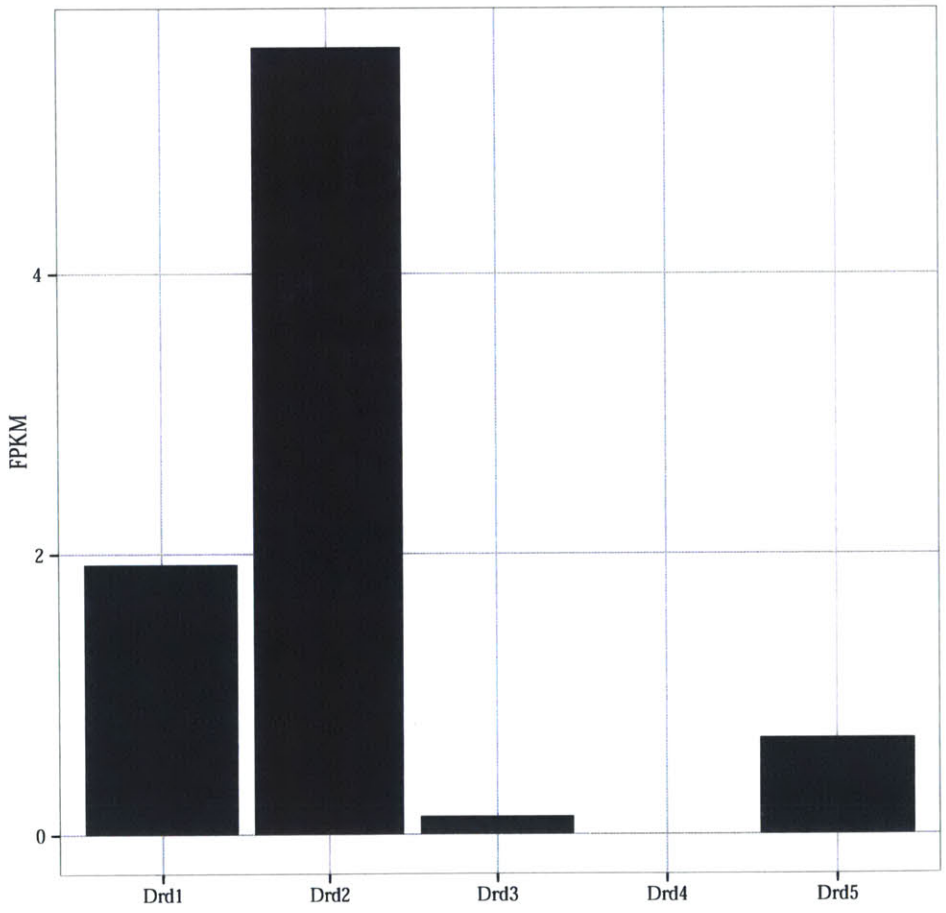
3.4 Discussion

This work shows that despite changes in protein translation immediately after eye opening[169] and persistent structural changes to the corticollicular projection with long term deprivation[64][137] we do not detect many persistent changes in gene expression in the superior colliculus with prolonged visual deprivation. Previous work done with similar sample sizes in the visual cortex[174] and [106] found hundreds of differentially expressed genes in the visual cortex under a similar prolonged light deprivation paradigm. These results and post-hoc power calculations of this experiment support the conclusion that maintenance of the structural changes of projections to the superior colliculus with eye opening do not depend on changes in gene expression to be maintained.

Deprivation on the order of several months reduces the number of projections from the stratum opticum of the rat[141] and reduces the size selectivity [145] and receptive fields of superior colliculus neurons[30]. Superior colliculus receptive field formation is activity-independent[28] whereas the receptive field formation in the visual cortex is activity-dependent[138]. The period we assayed in this experiment falls in between the early chemotaxis-dependent formation of the retinotopic map in the superior colliculus and the late activity dependent maintenace of the retinotopic projection. This work combined with other work is evidence for a process where the initial map is formed via chemotactic cues, after which eye opening causes brief, short term changes in gene expression leading to structural changes in the corticocollicular



(a)



(b)

Figure 3-9: Dopamine receptor subtypes are segregated into distinct layers of the superior colliculus. **a)** Drd1 tdTomato neurons are enriched in the superficial superior colliculus and Drd2 EGFP positive neurons are expressed in the SO and SGI of the superior colliculus[19]. **b)** RNA-seq shows expression of Drd1, Drd2 and Drd5 with possible light expression of Drd3 (Figure 3-9b). FPKM is Fragments Per Kilobase of transcript per Million mapped reads.

projection to the superior colliculus. After eye opening the activity genes involved in structural plasticity return to steady-state, barring a set of mitochondrial genes involved in handling the increased energy load on the superior colliculus caused by the influx of signal to the colliculus from the retina. Long term deprivation on the order of months induces a second round of activity-dependent structural changes in the colliculus leading to functional deficits.

We confirmed the finding that light driven retinal activity causes an increase in energy load on the colliculus and identified a small set of genes which may be coregulated with the mitochondrial genes on the X chromosome. Mcts1 has been shown to have anti anti-apoptotic effects; it is possible that there are additional mutations in Mcts1 or Cox7b that exacerbate the mitochondrial mutations in affected individuals. Most LHON studies have focused on sequencing the mitochondrial genome and it would be interesting to perform exome sequencing on families with affected and unaffected individuals and see if any variants on the X chromosome segregate with affected status in this set of candidate genes.

Finally we showed strong expression of the dopamine receptors Drd1 and Drd2 in the superior colliculus, resolving a conflict in the literature about whether or not Drd1 is expressed in the colliculus. We observed not only expression of Drd1 but a segregation of Drd1 and Drd2 into two distinct zones laminae, with Drd1 expressed mostly in the superficial superior colliculus and Drd2 in the intermediate superior colliculus. This observation lead to a work assessing the functional significance of this uncharacterized dopamine projection to the superior colliculus [19].

Bibliography

[1] James B Ackman, Timothy J Burbridge, and Michael C Crair. Retinal waves coordinate patterned activity throughout the developing visual system. *Nature*, 490(7419):219–225, October 2012. 93

[2] Enis Afgan, Brad Chapman, Margita Jadan, Vedran Franke, and James Taylor. Using cloud computing infrastructure with CloudBioLinux, CloudMan, and Galaxy. *Curr Protoc Bioinformatics*, Chapter 11:Unit11.9–11.9.20, June 2012. 22

[3] Gael P Alamancos, Eneritz Agirre, and Eduardo Eyras. Methods to study splicing from high-throughput RNA sequencing data. *Methods Mol Biol*, 1126(Chapter 26):357–397, 2014. 40, 42

[4] J E Albano and R H Wurtz. Deficits in eye position following ablation of monkey superior colliculus, pretectum, and posterior-medial thalamus. *J Neurophysiol*, 48(2):318–337, August 1982. 88

[5] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential Cell Biology, Fourth Edition*. Garland Science, October 2013. 113

[6] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010. 32

[7] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res*, 22(10):2008–2017, October 2012. 40, 66

[8] Charles Anderson. Docker [Software engineering]. *IEEE Software*, 32(3):102–c3, 2015. 22

[9] A Antonini and M Stryker. Rapid remodeling of axonal arbors in the visual cortex. *Science*, 260(5115):1819–1821, June 1993. 95

[10] Benjamin D Auerbach, Emily K Osterweil, and Mark F Bear. Mutations causing syndromic autism define an axis of synaptic pathophysiology. *Nature*, 480(7375):63–68, December 2011. 59, 62

[11] Brenda Bass, Heather Hundley, Jin Billy Li, Zhiyu Peng, Joe Pickrell, Xinshu Grace Xiao, and Li Yang. The difficult calls in RNA editing. *Nat Biotechnol*, 30(12):1207–1209, December 2012. 80

[12] Helen S Bateup, Caroline A Johnson, Cassandra L Denefrio, Jessica L Saulnier, Karl Kornacker, and Bernardo L Sabatini. Excitatory/Inhibitory Synaptic Imbalance Leads to Hippocampal Hyperexcitability in Mouse Models of Tuberous Sclerosis. *Neuron*, 78(3):510–522, May 2013. 55

[13] Anna Bauer-Mehren, Michael Rautschka, Ferran Sanz, and Laura I Furlong. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics*, 26(22):2924–2926, November 2010. 76

[14] M F Bear, A Kleinschmidt, Q A Gu, and W Singer. Disruption of experience-dependent synaptic modifications in striate cortex by infusion of an NMDA receptor antagonist. *J Neurosci*, 10(3):909–925, March 1990. 96

[15] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met*, 57(1):289–300, 1995. 66

[16] N Berardi and T Pizzorusso. ScienceDirect.com - Current Opinion in Neurobiology - Critical periods during sensory development. *Curr Opin Neurobiol*, 2000. 94

[17] Andrew Bierman, Mariana G Figueiro, and Mark S Rea. Measuring and predicting eyelid spectral transmittance. *J Biomed Opt*, 16(6):067011, June 2011. 94

[18] K E Binns and T E Salt. Developmental changes in NMDA receptor-mediated visual activity in the rat superior colliculus, and the effect of dark rearing. *Exp Brain Res*, 120(3):335–344, May 1998. 96

[19] Andrew Bolton, Yasunobu Murata, Rory Kirchner, and Martha Constantine-Paton. A diencephalic dopamine source projects to the superior colliculus, where D1 and D2 receptors segregate to behaviorally relevant zones. *Cell Rep.* 117, 118

[20] Bethany K Bracken and Gina G Turrigiano. Experience-dependent regulation of TrkB isoforms in rodent visual cortex. *Dev Neurobiol*, 69(5):267–278, April 2009. 98

[21] Bethany K Bracken and Gina G Turrigiano. Experience-dependent regulation of TrkB isoforms in rodent visual cortex. *Dev Neurobiol*, 69(5):267–278, April 2009. 100

[22] Oliver Braddick and Janette Atkinson. Development of human visual function. *Vision Res*, 51(13):1588–1609, July 2011. 95

[23] Nicolas Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal RNA-Seq quantification. *arXiv*, May 2015. 49

[24] L Cancedda, E Putignano, and S Impey. Patterned Vision Causes CRE-Mediated Gene Expression in the Visual Cortex through PKA and ERK. *The Journal of ...*, 2003. 98

[25] Mehmet Canpolat, Huseyin Per, Hakan Gumus, Ali Yikilmaz, Ekrem Unal, Turkan Patiroglu, Levent Cinar, Ali Kurtsoy, and Sefer Kumandas. Rapamycin has a beneficial effect on controlling epilepsy in children with tuberous sclerosis complex: results of 7 children from a cohort of 86. *Childs Nerv Syst*, 30(2):227–240, June 2013. 56

[26] Valerio Carelli, Fred N Ross-Cisneros, and Alfredo A Sadun. Mitochondrial dysfunction as a cause of optic neuropathies. *Prog Retin Eye Res*, 23(1):53–89, January 2004. 113

[27] G Carmignoto. Activity-dependent decrease in NMDA receptor responses during development of the visual cortex. *Science*, 1992. 96

[28] M. M. Carrasco, K. A. Razak, and S. L. Pallas. Visual Experience Is Necessary for Maintenance But Not Development of Receptive Fields in Superior Colliculus. 2005. 95, 116

[29] María M Carrasco, Yu-Ting Mao, Timothy S Balmer, and Sarah L Pallas. Inhibitory plasticity underlies visual deprivation-induced loss of receptive field refinement in the adult superior colliculus. *Eur J Neurosci*, 33(1):58–68, November 2010. 92

[30] María M Carrasco, Yu-Ting Mao, Timothy S Balmer, and Sarah L Pallas. Inhibitory plasticity underlies visual deprivation-induced loss of receptive field refinement in the adult superior colliculus. *Eur J Neurosci*, 33(1):58–68, January 2011. 116

[31] María Magdalena Carrasco and Sarah L Pallas. Early visual experience prevents but cannot reverse deprivation-induced loss of refinement in adult superior colliculus. *Vis Neurosci*, 23(6):845–852, November 2006. 96

[32] Robert P Carson, Dominic L Van Nielen, Peggy A Winzenburger, and Kevin C Ess. Neuronal and glia abnormalities in Tsc1-deficient forebrain and partial rescue by rapamycin. *Neurobiology of Disease*, 45(1):369–380, January 2012. 57

[33] V A Casagrande and I T Diamond. Ablation study of the superior colliculus in the tree shrew (Tupaia glis). *J Comp Neurol*, 156(2):207–237, July 1974. 88

[34] James Cavanaugh, Ilya E Monosov, Kerry McAlonan, Rebecca Berman, Mitchell K Smith, Vania Cao, Kuan H Wang, Edward S Boyden, and Robert H Wurtz. Optogenetic inactivation modifies monkey visuomotor behavior. *Neuron*, 76(5):901–907, December 2012. 91

[35] Anand R Chandrasekaran, Daniel T Plas, Ernesto Gonzalez, and Michael C Crair. Evidence for an instructive role of retinal activity in retinotopic map refinement in the superior colliculus of the mouse. *Journal of Neuroscience*, 25(29):6929–6938, July 2005. 92

[36] Brad A Chapman. Benchmarking variation and RNA-seq analyses on Amazon Web Services with Docker. `http://bcb.io/2014/12/19/awsbench/`. 18

[37] Brad A Chapman and Rory Kirchner. bcbio-nextgen: automated, validated analysis of high throughput sequencing data. `https://pypi.python.org/pypi/bcbio-nextgen/0.9.0`, 2015. 12

[38] Dmitri B Chklovskii and Alexei A Koulakov. Maps in the brain: what can we learn from them? *Annu Rev Neurosci*, 27:369–392, 2004. 91

[39] Yong-Jin Choi, Alessia Di Nardo, Ioannis Kramvis, Lynsey Meikle, David J Kwiatkowski, Mustafa Sahin, and Xi He. Tuberous sclerosis complex proteins control axon formation. *Genes & Development*, 22(18):2485–2495, September 2008. 55

[40] Francis S Collins and Lawrence A Tabak. Policy: NIH plans to enhance reproducibility. *Nature*, 505(7485):612–613, January 2014. 23

[41] SEQC MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol*, 32(9):903–914, September 2014. 27

[42] Peter B Crino, Katherine L Nathanson, and Elizabeth Petri Henske. The Tuberous Sclerosis Complex. *N Engl J Med*, 355(13):1345–1356, September 2006. 50, 51

[43] Alexis Criscuolo and Sylvain Brisse. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*, 102(5-6):500–506, November 2013. 105

[44] G Del Angel, M A Rivas, M Hanna, and A McKenna. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature ...*, 2011. 66

[45] David S DeLuca, Joshua Z Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, 28(11):1530–1532, June 2012. 15, 65, 105

[46] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, Florence Jaffrézic, and French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, November 2013. 15

[47] Emmanuel Dimont, Jiantao Shi, Rory Kirchner, and Winston Hide. edgeRun: an R package for sensitive, functionally relevant differential expression discovery using an unconditional exact test. *Bioinformatics*, page btv209, April 2015. 41

[48] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013. 31, 37, 65, 105

[49] R M Douglas, N M Alam, B D Silver, T J McGill, W W Tschetter, and G T Prusky. Independent visual threshold measurements in the two eyes of freely moving rats and mice using a virtual-reality optokinetic system. *Vis Neurosci*, 22(5):677–684, September 2005. 94

[50] Sorin Draghici, Purvesh Khatri, Aron C Eklund, and Zoltan Szallasi. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.*, 22(2):101–109, February 2006. 23

[51] Stephen M Eacker, Matthew J Keuss, Eugene Berezikov, Valina L Dawson, and Ted M Dawson. Neuronal Activity Regulates Hippocampal miRNA Expression. *PLoS ONE*, 6(10):e25068, October 2011. 85

[52] Dan Ehninger, Sangyeul Han, Carrie Shilyansky, Yu Zhou, Weidong Li, David J Kwiatkowski, Vijaya Ramesh, and Alcino J Silva. Reversal of learning deficits in a Tsc2+/− mouse model of tuberous sclerosis. *Nature medicine*, 14(8):843–848, August 2008. 55, 56

[53] Dan Ehninger and Alcino J Silva. Rapamycin for treating Tuberous sclerosis and Autism spectrum disorders. *Trends in Molecular Medicine*, 17(2):78–87, February 2011. 56

[54] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Tyler Alioto, Jonas Behr, Paul Bertone, Regina Bohnert, Davide Campagna, Carrie A Davis, Alexander Dobin, Thomas R Gingeras, Nick Goldman, Roderic Guigó, Jennifer Harrow, Tim J Hubbard, Géraldine Jean, Peter Kosarev, Sheng Li, Jinze Liu, Christopher E Mason, Vladimir Molodtsov, Zemin Ning, Hannes Ponstingl, Jan F Prins, Gunnar Rätsch, Paolo Ribeca, Igor Seledtsov, Victor Solovyev, Giorgio Valle, Nicola Vitulo, Kai Wang, Thomas D Wu, and Georg Zeller. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Meth*, 10(12):1185–1191, November 2013. 15, 37

[55] B Ewing, L Hillier, M C Wendl, and P Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8(3):175–185, March 1998. 105

[56] M Fagiolini, T Pizzorusso, N Berardi, and L Domenici. ScienceDirect.com - Vision Research - Functional postnatal development of the rat primary visual cortex and the role of visual experience: Dark rearing and monocular deprivation. *Vision Res*, 1994. 95

[57] Michela Fagiolini, Hiroyuki Katagiri, Hiroyuki Miyamoto, Hisashi Mori, Seth G N Grant, Masayoshi Mishina, and Takao K Hensch. Separable features of visual cortical plasticity revealed by N-methyl-D-aspartate receptor 2A signaling. *Proc Natl Acad Sci USA*, 100(5):2854–2859, March 2003. 95

[58] N Fillmore, Y Bai, M Collins, J A Thomson, and R Stewart. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *bioRxiv*, 2014. 15

[59] I Fine, AR Wade, AA Brewer, and MG May. Long-term deprivation affects visual perception and cortex - Nature Neuroscience. *Neuroscience*, 2003. 95

[60] Alyssa C Frazee, Sarven Sabunciyan, Kasper D Hansen, Rafael A Irizarry, and Jeffrey T Leek. Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics (Oxford, England)*, 15(3):413–426, July 2014. 40

[61] T W Frazier, R Embacher, A K Tilot, K Koenig, J Mester, and C Eng. Molecular and phenotypic abnormalities in individuals with germline heterozygous PTEN mutations and autism. *Mol Psychiatry*, October 2014. 60

[62] Fernando García-Alcalde, Konstantin Okonechnikov, José Carbonell, Luis M Cruz, Stefan Götz, Sonia Tarazona, Joaquín Dopazo, Thomas F Meyer, and Ana Conesa. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, 28(20):2678–2679, October 2012. 15

[63] Shruti Garg, Jonathan Green, Kathy Leadbitter, Richard Emsley, Annukka Lehtonen, D Gareth Evans, and Susan M Huson. Neurofibromatosis type 1 and autism spectrum disorder. *Pediatrics*, 132(6):e1642–8, December 2013. 60

[64] Julie Goldberg. Eye-opening dependent elaboration and refinement of the cortical projection to the superior colliculus in rats. *Massachusetts Institute of Technology*, pages 1–89, August 2009. 98, 99, 106, 107, 108, 116

[65] Michael E Goldberg and Robert H Wurtz. Activity of superior colliculus in behaving monkey. I. Visual receptive fields of single neurons. *J Neurophysiol*, 35(4):542–559, 1972. 89

[66] Elena A Goncharova, Dmitry A Goncharov, Andrew Eszterhas, Deborah S Hunter, Marilyn K Glassberg, Raymond S Yeung, Cheryl L Walker, Daniel Noonan, David J Kwiatkowski, Margaret M Chou, Reynold A Panettieri, and Vera P Krymskaya. Tuberin regulates p70 S6 kinase activation and ribosomal protein S6 phosphorylation. A role for the TSC2 tumor suppressor gene in pulmonary lymphangioleiomyomatosis (LAM). *J Biol Chem*, 277(34):30958–30967, August 2002. 53

[67] T Griebel, B Zacher, P Ribeca, E Raineri, V Lacroix, R Guigo, and M Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res*, 40(20):10073–10083, November 2012. 28

[68] Dominic Grün, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat Meth*, 11(6):637–640, June 2014. 48

[69] Shuchen Gu, Michalis Kounenidakis, Eva-Maria Schmidt, Divija Deshpande, Saad Alkahtani, Saud Alarifi, Michael Föller, Konstantinos Alevizopoulos, Florian Lang, and Christos Stournaras. Rapid activation of FAK/mTOR/p70S6K/PAK1-signaling controls the early testosterone-induced actin reorganization in colon cancer cells. *Cellular Signalling*, 25(1):66–73, January 2013. 60

[70] Simone Gupta, Shannon E Ellis, Foram N Ashar, Anna Moes, Joel S Bader, Jianan Zhan, Andrew B West, and Dan E Arking. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Communications*, 5:5748, 2014. 62

[71] William C Hall and Adonis K Moschovakis, editors. *The Superior Colliculus: New Approaches for Studying Sensorimotor Integration (Methods and New Frontiers in Neuroscience)*. CRC Press, 1 edition, September 2003. 87

[72] Thomas J Hardcastle and Krystyna A Kelly. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, August 2010. 41

[73] J K Harting, I T Diamond, and W C Hall. Anterograde degeneration study of the cortical projections of the lateral geniculate and pulvinar nuclei in the tree shrew (Tupaia glis). *J Comp Neurol*, 150(4):393–439, 2004. 88

[74] D O Hebb. *The Organization of Behavior.* A Neuropsychological Theory. Psychology Press, January 2002. 93

[75] G F L Hofbauer, A Marcollo-Pini, A Corsenca, A D Kistler, L E French, R P Wüthrich, and A L Serra. The mTOR inhibitor rapamycin significantly improves facial angiofibroma lesions in a patient with tuberous sclerosis. *Br. J. Dermatol.*, 159(2):473–475, August 2008. 53

[76] G HUDSON, S KEERS, P MAN, P GRIFFITHS, K HUOPONEN, M SAVONTAUS, E NIKOSKELAINEN, M ZEVIANI, F CARRARA, and R HORVATH. Identification of an X-Chromosomal Locus and Haplotype Modulating the Phenotype of a Mitochondrial DNA Disorder. *The American Journal of Human Genetics*, 77(6):1086–1091, December 2005. 113

[77] Darrel C Ince, Leslie Hatton, and John Graham-Cumming. The case for open computer programs. *Nature*, 482(7386):485–488, February 2012. 23

[78] Broad Institute. Calling variants in RNAseq. `https://www.broadinstitute.org/gatk/guide/article?id=3891`, 2014. 45, 66

[79] Manuel Irimia, Robert J Weatheritt, Jonathan D Ellis, Neelroop N Parikshak, Thomas Gonatopoulos-Pournatzis, Mariana Babor, Mathieu Quesnel-Vallières, Javier Tapial, Bushra Raj, Dave O'Hanlon, Miriam Barrios-Rodiles, Michael J E Sternberg, Sabine P Cordes, Frederick P Roth, Jeffrey L Wrana, Daniel H Geschwind, and Benjamin J Blencowe. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, 159(7):1511–1523, December 2014. 48

[80] T Isa, T Endo, and Y Saito. The visuo-motor pathway in the local circuit of the rat superior colliculus. *J Neurosci*, 18(20):8496–8504, October 1998. 89

[81] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Meth*, 11(2):163–166, February 2014. 48

[82] Yanli Ji, Xiaoyun Jia, Shiqiang Li, Xueshan Xiao, Xiangming Guo, and Qingjiong Zhang. Evaluation of the X-linked modifier loci for Leber hereditary optic neuropathy with the G11778A mutation in Chinese. *Mol. Vis.*, 16:416–424, 2010. 113

[83] Hongshan Jiang, Rong Lei, Shou-Wei Ding, and Shuifang Zhu. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15(1):182, 2014. 35

[84] C JOINSON, F J O'CALLAGHAN, J P OSBORNE, C MARTYN, T HARRIS, and P F BOLTON. Learning disability and epilepsy in an epidemiological sample of individuals with tuberous sclerosis complex. *Psychological Medicine*, 33(02):335–344, February 2003. 50

[85] James A Kaltenbach and John Lazor. Tonotopic maps obtained from the surface of the dorsal cochlear nucleus of the hamster and rat. *Hearing research*, 51(1):149–160, January 1991. 91

[86] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 2000. 75

[87] Y Katz, E Wang, and E Airoldi. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Meth*, 2010. 40

[88] Do-Hyung Kim, Dos D Sarbassov, Siraj M Ali, Jessie E King, Robert R Latek, Hediye Erdjument-Bromage, Paul Tempst, and David M Sabatini. mTOR Interacts with Raptor to Form a Nutrient-Sensitive Complex that Signals to the Cell Growth Machinery. *Cell*, 110(2):163–175, July 2002. 53

[89] Stefan Kirov, Ruiru Ji, Jing Wang, and Bing Zhang. Functional annotation of differentially regulated gene set using WebGestalt: a gene set predictive of response to ipilimumab in tumor biopsies. *Methods Mol Biol*, 1101(Chapter 3):31–42, 2014. 66

[90] Paul J Kiser, Zijing Liu, Steven D Wilt, and George D Mower. Cellular and laminar expression of Dab-1 during the postnatal critical period in cat visual cortex and the effects of dark rearing. *Brain Res*, 1383:81–89, April 2011. 98

[91] T Kobayashi, O Minowa, Y Sugitani, S Takai, H Mitani, E Kobayashi, T Noda, and O Hino. A germ-line Tsc1 mutation causes tumor development and embryonic lethality that are similar, but not identical to, those caused by Tsc2 mutation in mice. *Proc Natl Acad Sci USA*, 98(15):8762–8767, July 2001. 56

[92] Sek Won Kong, Mustafa Sahin, Christin D Collins, Mary H Wertz, Malcolm G Campbell, Jarrett D Leech, Dilja Krueger, Mark F Bear, Louis M Kunkel, and Isaac S Kohane. Divergent dysregulation of gene expression in murine models of fragile X syndrome and tuberous sclerosis. *Mol Autism*, 5(1):16, 2014. 62

[93] K Krug, C J Akerman, and I D Thompson. Responses of neurons in neonatal cortex and thalamus to patterned visual stimulation through the naturally closed lids. *J Neurophysiol*, 85(4):1436–1443, April 2001. 94

[94] C W Law, Y Chen, W Shi, and G K Smyth. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Preprint 2013*, 2013. 28, 32, 41

[95] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart M G Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–1043, April 2013. 41

[96] Nanxin Li, Boyoung Lee, Rong-Jian Liu, Mounira Banasr, Jason M Dwyer, Masaaki Iwata, Xiao-Yuan Li, George Aghajanian, and Ronald S Duman. mTOR-dependent synapse formation underlies the rapid antidepressant effects of NMDA antagonists. *Science*, 329(5994):959–964, August 2010. 55

[97] Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014. 40, 66, 105

[98] Robert Lindner and Caroline C Friedel. A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq. *PLoS ONE*, 7(12):e52403, 2012. 26

[99] Jonathan O Lipton and Mustafa Sahin. The Neurology of mTOR. *Neuron*, 84(2):275–291, October 2014. 52

[100] S G Lomber, B R Payne, and P Cornwell. Role of the superior colliculus in analyses of space: superficial and intermediate layer contributions to visual orienting, auditory orienting, and visuospatial discriminations during unilateral and bilateral deactivations. *J Comp Neurol*, 441(1):44–57, December 2001. 88

[101] M I Love, W Huber, and S Anders. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol*, 2014. 32, 41

[102] Joaquin N Lugo, Gregory D Smith, Erin P Arbuckle, Jessika White, Andrew J Holley, Crina M Floruta, Nowrin Ahmed, Maribel C Gomez, and Obi Okonkwo. Deletion of PTEN produces autism-like behavioral deficits and alterations in synaptic proteins. *Front Mol Neurosci*, 7:27, 2014. 60

[103] R D Lund. Terminal Distribution in the Superior Colliculus of Fibres Originating in the Visual Cortex. *Nature*, 204(4965):1283–1285, December 1964. 89

[104] Daniel Macarthur. Methods: Face up to false positives. *Nature*, 487(7408):427–428, July 2012. 23

[105] Matthew D Macmanes. On the optimal trimming of high-throughput mRNA sequence data. *Front Genet*, 5:13, 2014. 35, 65, 105

[106] Marta Majdan and Carla J Shatz. Effects of visual experience on activity-dependent gene regulation in cortex. *Nat Neurosci*, 9(5):650–659, May 2006. 97, 116

[107] A Mansour, J H Meador-Woodruff, Q Zhou, O Civelli, H Akil, and S J Watson. A comparison of D1 receptor binding and mRNA in rat brain using receptor autoradiographic and in situ hybridization techniques. *Neuroscience*, 46(4):959–971, February 1992. 114

[108] A Mansour and S J Watson. Dopamine receptor expression in the central nervous system. *Psychopharmacology: the fourth generation of ...*, 1995. 114

[109] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12, February 2011. 35, 65

[110] Daphne Maurer, Terri L Lewis, and Catherine J Mondloch. Missing sights: consequences for visual cognitive development. *Trends Cogn. Sci. (Regul. Ed.)*, 9(3):144–151, March 2005. 95

[111] Christopher J McDougle, Lawrence Scahill, Michael G Aman, James T McCracken, Elaine Tierney, Mark Davies, L Eugene Arnold, David J Posey, Andrès Martin, Jaswinder K Ghuman, Bhavik Shah, Shirley Z Chuang, Naomi B Swiezy, Nilda M Gonzalez, Jill Hollway, Kathleen Koenig, James J McGough, Louise Ritz, and Benedetto Vitiello. Risperidone for the Core Symptom Domains of Autism: Results From the Study by the Autism Network of the Research Units on Pediatric Psychopharmacology. *American Journal of Psychiatry*, 162(6):1142–1148, November 2014. 86

[112] Todd McLaughlin, Christine L Torborg, Marla B Feller, and Dennis D M O'Leary. Retinotopic map refinement requires spontaneous retinal waves during a brief critical period of development. *Neuron*, 40(6):1147–1160, December 2003. 92, 93

[113] John J McMahon, Wilson Yu, Jun Yang, Haihua Feng, Meghan Helm, Elizabeth McMahon, Xinjun Zhu, Damian Shin, and Yunfei Huang. Seizure-dependent mTOR activation in 5-HT neurons promotes autism-like behaviors in mice. *Neurobiology of Disease*, 73:296–306, January 2015. 81

[114] Lynsey Meikle, Delia M Talos, Hiroaki Onda, Kristen Pollizzi, Alexander Rotenberg, Mustafa Sahin, Frances E Jensen, and David J Kwiatkowski. A mouse model of tuberous sclerosis: neuronal loss of Tsc1 causes dysplastic and ectopic neurons, reduced myelination, seizure activity, and limited survival. *Journal of Neuroscience*, 27(21):5546–5558, May 2007. 55, 57, 63, 64

[115] Nikolaos Mellios, Hiroki Sugihara, Jorge Castro, Abhishek Banerjee, Chuong Le, Arooshi Kumar, Benjamin Crawford, Julia Strathmann, Daniela Tropea, Stuart S Levine, Dieter Edbauer, and Mriganka Sur. miR-132, an experience-dependent microRNA, is essential for visual cortex plasticity. *Nat Neurosci*, 14(10):1240–1242, October 2011. 98

[116] M M Merzenich, J H Kaas, J Wall, R J Nelson, M Sur, and D Felleman. Topographic reorganization of somatosensory cortical areas 3b and 1 in adult monkeys following restricted deafferentation. *Neuroscience*, 8(1):33–55, January 1983. 91

[117] Masashi Mizuguchi and Sachio Takashima. Neuropathology of tuberous sclerosis. *Brain Dev*, 23(7):508–515, November 2001. 51

[118] Ahmad R Mohamed, Catherine A Bailey, Jeremy L Freeman, Wirginia Maixner, Graeme D Jackson, and A Simon Harvey. Intrinsic epileptogenicity of cortical tubers revealed by intracranial EEG monitoring. *Neurology*, 79(23):2249–2257, December 2012. 52

[119] Andrei I Molosh, Philip L Johnson, John P Spence, David Arendt, Lauren M Federici, Cristian Bernabe, Steven P Janasik, Zaneer M Segu, Rajesh Khanna, Chirayu Goswami, Weiguo Zhu, Su-Jung Park, Lang Li, Yehia S Mechref, D Wade Clapp, and Anantha Shekhar. Social learning and amygdala disruptions in Nf1 mice are rescued by blocking p21-activated kinase. *Nat Neurosci*, 17(11):1583–1590, November 2014. 60

[120] S Molotchnikoff and S K Itaya. Functional development of the neonatal rat retinotectal pathway. *Brain Res. Dev. Brain Res.*, 72(2):300–304, April 1993. 93

[121] Ali Mortazavi, Brian A Williams, Kenneth Mccue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7):621, May 2008. 100

[122] G D Mower, C J Caplan, and G Letsou. Behavioral recovery from binocular deprivation in the cat. *Behav Brain Res*, 4(2):209–215, February 1982. 95

[123] Elly Nedivi. Molecular analysis of developmental plasticity in neocortex. *J. Neurobiol.*, 41(1):135–147, October 1999. 96, 97

[124] Anton Nekrutenko and James Taylor. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet*, 13(9):667–672, September 2012. 23

[125] Stanley F Nelson. [Book Review: Microarray Gene Expression Data Analysis: A Beginner's Guide]. *The Quarterly Review of Biology*, 79(3):308–309, September 2004. 12

[126] Lena H Nguyen, Amy L Brewster, Madeline E Clark, Angelique Regnier-Golanov, C Nicole Sunnen, Vinit V Patil, Gabriella D'Arcangelo, and Anne E Anderson. mTOR inhibition suppresses established epilepsy in a mouse model of cortical dysplasia. *Epilepsia*, March 2015. 60

[127] Duyu Nie, Alessia Di Nardo, Juliette M Han, Hasani Baharanyi, Ioannis Kramvis, ThanhThao Huynh, Sandra Dabora, Simone Codeluppi, Pier Paolo Pandolfi, Elena B Pasquale, and Mustafa Sahin. Tsc2-Rheb signaling regulates EphA-mediated axon guidance. *Nat Neurosci*, 13(2):163–172, January 2010. 55

[128] Umadevi Paila, Brad A Chapman, Rory Kirchner, and Aaron R Quinlan. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.*, 9(7):e1003153, 2013. 14, 18, 21, 46

[129] L Pantano, X Estivill, and E Martí. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic acids research*, 2010. 14

[130] Rob Patro. Salmon. `http://sailfish.readthedocs.org/en/latest/salmon.html`, 2015. 49

[131] M Paulussen, L Van Brussel, and L Arckens. Monocular enucleation profoundly reduces secretogranin II expression in adult mouse visual cortex. *Neurochem. Int.*, 59(7):1082–1094, December 2011. 98

[132] H Peng, J Liu, Q Sun, R Chen, Y Wang, J Duan, C Li, B Li, Y Jing, X Chen, Q Mao, K-F Xu, C L Walker, J Li, J. Wang, and H Zhang. mTORC1 enhancement of STIM1-mediated store-operated Ca2+ entry constrains tuberous sclerosis complex-related tumor development. *Oncogene*, 32(39):4702–4711, September 2013. 81

[133] Fernando Pérez and Brian E Granger. IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, 9(3):21–29, 2007. 18

[134] Geo Pertea. high FPKM called with few reads mapping to transcript · Issue #12 · gpertea/stringtie. `https://github.com/gpertea/stringtie/issues/12`. 49

[135] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*, February 2015. 49

[136] Marnie A Phillips. Eye-opening and the control of visual synapse development in the mouse superior colliculus. *Massachusetts Institute of Technology*, pages 1–128, September 2007. 90

[137] Marnie A Phillips, Matthew T Colonnese, Julie Goldberg, Laura D Lewis, Emery N Brown, and Martha Constantine-Paton. A synaptic strategy for consolidation of convergent visuotopic maps. *Neuron*, 71(4):710–724, August 2011. 98, 116

[138] Nathalie Picard, Jennifer H Leslie, Sara K Trowbridge, Jaichandar Subramanian, Elly Nedivi, and Michela Fagiolini. Aberrant development and plasticity of excitatory visual cortical networks in the absence of cpg15. *Journal of Neuroscience*, 34(10):3517–3522, March 2014. 116

[139] Joseph K Pickrell, Yoav Gilad, and Jonathan K Pritchard. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, 335(6074):1302–author reply 1302, March 2012. 45

[140] Ellen Plasschaert, Mie-Jef Descheemaeker, Lien Van Eylen, Ilse Noens, Jean Steyaert, and Eric Legius. Prevalence of Autism Spectrum Disorder symptoms in children with neurofibromatosis type 1. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 168B(1):72–80, January 2015. 60

[141] J Roxanne Prichard, Hilda S Armacanqui, Ruth M Benca, and Mary Behan. Light-dependent retinal innervation of the rat superior colliculus. *Anat Rec (Hoboken)*, 290(3):341–348, March 2007. 116

[142] JR Prichard and HS Armacanqui. Light-dependent retinal innervation of the rat superior colliculus - Prichard - 2007 - The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology - Wiley Online Library. ... *Record: Advances in ...*, 2007. 95

[143] C Quaia, H Aizawa, L M Optican, and R H Wurtz. Reversible inactivation of monkey superior colliculus. II. Maps of saccadic deficits. *J Neurophysiol*, 79(4):2097–2110, April 1998. 90

[144] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*, 14(9):R95, 2013. 15, 26

[145] Khaleel A Razak and Sarah L Pallas. Dark rearing reveals the mechanism underlying stimulus size tuning of superior colliculus neurons. *Vis Neurosci*, 23(5):741–748, September 2006. 116

[146] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Meth*, 10(1):71–73, January 2013. 37, 40

[147] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Meth*, 10(1):71–73, January 2013. 66

[148] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329, September 2011. 42

[149] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. 32, 41, 108, 110

[150] Natalia S Rozas, John B Redell, James McKenna, Anthony N Moore, Michael J Gambello, and Pramod K Dash. Prolonging the survival of Tsc2 conditional knockout mice by glutamine supplementation. *Biochem. Biophys. Res. Commun.*, January 2015. 56

[151] Atsushi Sato, Shinya Kasai, Toshiyuki Kobayashi, Yukio Takamatsu, Okio Hino, Kazutaka Ikeda, and Masashi Mizuguchi. Rapamycin reverses impaired social interaction in mouse models of tuberous sclerosis complex. *Nature Communications*, 3:1292, December 2012. 56

[152] Yiannis A Savva, Leila E Rieder, and Robert A Reenan. The ADAR protein family. *Genome Biol*, 13(12):252, 2012. 46

[153] P H Schiller, J H Sandell, and J H Maunsell. The effect of frontal eye field and superior colliculus lesions on saccadic latencies in the rhesus monkey. *J Neurophysiol*, 57(4):1033–1049, April 1987. 88

[154] Mark Seger. collectl. `http://collectl.sourceforge.net`, 2015. 18

[155] Suma P Shankar, John H Fingert, Valerio Carelli, Maria L Valentino, Terri M King, Stephen P Daiger, Solange R Salomao, Adriana Berezovsky, Jr Rubens Belfort, Terri A Braun, Val C Sheffield, Alfredo A Sadun, and Edwin M Stone. Evidence for a Novel X-Linked Modifier Locus for Leber Hereditary Optic Neuropathy. *http://dx.doi.org/10.1080/13816810701867607*, 29(1):17–24, July 2009. 113

[156] Ali Sharma, Charles A Hoeffer, Yukihiro Takayasu, Takahiro Miyawaki, Sean M McBride, Eric Klann, and R Suzanne Zukin. Dysregulation of mTOR Signaling in Fragile X Syndrome. *J Neurosci*, 30(2):694–702, January 2010. 60

[157] Jill L Silverman, Mu Yang, Catherine Lord, and Jacqueline N Crawley. Behavioural phenotyping assays for mouse models of autism. *Nat Rev Neurosci*, 11(7):490–502, July 2010. 59

[158] T Soucek, G Hölzl, G Bernaschek, and M Hengstschläger. A role of the tuberous sclerosis gene-2 product during neuronal differentiation. *Oncogene*, 16(17):2197–2204, April 1998. 53

[159] Thomas Soucek, Oliver Pusch, Ralf Wienecke, Jeffrey E DeClue, and Markus Hengstschläger. Role of the Tuberous Sclerosis Gene-2 Product in Cell Cycle Control: loss of the tuberous sclerosis gene-2 induces quiescent cells to enter s phase. *J Biol Chem*, 272(46):29301–29308, November 1997. 53

[160] D.L. Sparks and R Hartwich-Young. The deep layers of the superior colliculus. *Reviews of oculomotor research*, 3:213, 1989. 89

[161] R W Sperry. Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proc Natl Acad Sci USA*, 50(4):703, October 1963. 91

[162] CWL standards team. Common Workflow Language, Draft 2. `http://common-workflow-language.github.io/#/`, 2015. 49

[163] Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, Tim J Hubbard, Roderic Guigó, Jennifer Harrow, Paul Bertone, and RGASP Consortium. Assessment of transcript reconstruction methods for RNA-seq. *Nat Meth*, 10(12):1177–1184, December 2013. 15, 42

[164] W Richard Stevens and Stephen A Rago. *Advanced Programming in the UNIX Environment*. Addison-Wesley, June 2013. 35

[165] M P Stryker and W A Harris. Binocular impulse blockade prevents the formation of ocular dominance columns in cat visual cortex. *J Neurosci*, 6(8):2117–2133, August 1986. 93

[166] Guomei Tang, Kathryn Gudsnuk, Sheng-Han Kuo, Marisa L Cotrina, Gorazd Rosoklija, Alexander Sosunov, Mark S Sonders, Ellen Kanter, Candace Castagna, Ai Yamamoto, Zhenyu Yue, Ottavio Arancio, Bradley S Peterson, Frances Champagne, Andrew J Dwork, James Goldman, and David Sulzer. Loss of mTOR-dependent macroautophagy causes autistic-like synaptic pruning deficits. *Neuron*, 83(5):1131–1143, September 2014. 59

[167] Paola Tognini, Elena Putignano, Alessandro Coatti, and Tommaso Pizzorusso. Experience-dependent expression of miR-132 regulates ocular dominance plasticity. *Nat Neurosci*, 14(10):1237–1239, October 2011. 98

[168] M Townsend. Retina-driven dephosphorylation of the NR2A subunit correlates with faster NMDA receptor kinetics at developing retinocollicular synapses. *Journal of Neuroscience*, 24(49):11098–11107, December 2004. 96

[169] M Townsend, A Yoshii, M Mishina, and M Constantine-Paton. Developmental loss of miniature N-methyl-D-aspartate receptor currents in NR2A knockout mice. *Proc Natl Acad Sci USA*, 100(3):1340–1345, February 2003. 97, 116

[170] Col Trapnell, Lio Pachter, and Steven Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):0, May 2009. 33

[171] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3):562–578, March 2012. 41

[172] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511, May 2010. 40

[173] Jason W Triplett, An Phan, Jena Yamada, and David A Feldheim. Alignment of Multimodal Sensory Input in the Superior Colliculus through a Gradient-Matching Mechanism. *Journal of Neuroscience*, 32(15):5264–5271, April 2012. 92

[174] Daniela Tropea, Gabriel Kreiman, Alvin Lyckman, Sayan Mukherjee, Hongbo Yu, Sam Horng, and Mriganka Sur. Gene expression changes and molecular pathways mediating activity-dependent plasticity in visual cortex. *Nat Neurosci*, 9(5):660–668, May 2006. 97, 98, 116

[175] Daniela Tropea, Audra Van Wart, and Mriganka Sur. Molecular mechanisms of experience-dependent plasticity in visual cortex. *Philos Trans R Soc Lond, B, Biol Sci*, 364(1515):341–355, February 2009. 97

[176] R J Tusa, M J Mustari, A F Burrows, and A F Fuchs. Gaze-stabilizing deficits and latent nystagmus in monkeys with brief, early-onset visual deprivation: eye movement recordings. *J Neurophysiol*, 86(2):651–661, August 2001. 95

[177] Erik J Uhlmann, Michael Wong, Rebecca L Baldwin, M Livia Bajenaru, Hiroaki Onda, David J Kwiatkowski, Kelvin Yamada, and David H Gutmann. Astrocyte-specific TSC1 conditional knockout mice exhibit abnormal neuronal organization and seizures. *Ann Neurol.*, 52(3):285–296, September 2002. 57

[178] Geraldine Van der Auwera. Calling variants on cohorts of samples using the HaplotypeCaller in GVCF mode. `http://gatkforums.broadinstitute.org/discussion/3893/calling-variants-on-cohorts-of-samples-using-the-haplotypecaller-in-gvcf-mode`, 2014. 67

[179] Jeremy Veenstra-VanderWeele, Christopher L Muller, Hideki Iwamoto, Jennifer E Sauer, W Anthony Owens, Charisma R Shah, Jordan Cohen, Padmanabhan Mannangatti, Tammy Jessen, Brent J Thompson, Ran Ye, Travis M Kerr, Ana M Carneiro, Jacqueline N Crawley, Elaine Sanders-Bush, Douglas G McMahon, Sammanda Ramamoorthy, Lynette C Daws, James S Sutcliffe, and Randy D Blakely. Autism gene variant causes hyperserotonemia, serotonin receptor hypersensitivity, social impairment and repetitive behavior. *Proc Natl Acad Sci USA*, 109(14):5469–5474, April 2012. 81

[180] Irina Voineagu, Xinchen Wang, Patrick Johnston, Jennifer K Lowe, Yuan Tian, Steve Horvath, Jonathan Mill, Rita M Cantor, Benjamin J Blencowe, and Daniel H Geschwind. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351):380–384, June 2011. 62, 63

[181] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008. 38

[182] L Wang, R Sarnaik, and K Rangarajan. Visual Receptive Field Properties of Neurons in the Superficial Superior Colliculus of the Mouse. *The Journal of ...*, 2010. 95

[183] Liguo Wang, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*, 41(6):e74–e74, April 2013. 42

[184] Lupeng Wang, Krsna V Rangarajan, Courtney A Lawhn-Heath, Rashmi Sarnaik, Bor-Shuen Wang, Xiaorong Liu, and Jianhua Cang. Direction-specific disruption of subcortical visual behavior and receptive fields in mice lacking the beta2 subunit of nicotinic acetylcholine receptor. *Journal of Neuroscience*, 29(41):12909–12918, October 2009. 93

[185] Yanling Wang, Joel S F Greenwood, Maria Elisa Calcagnotto, Heidi E Kirsch, Nicholas M Barbaro, and Scott C Baraban. Neocortical hyperexcitability in a human case of tuberous sclerosis complex and mice lacking neuronal expression of TSC1. *Ann Neurol.*, 61(2):139–152, February 2007. 52, 58

[186] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009. 13

[187] D M Weiner, A I Levey, R K Sunahara, H B Niznik, B F O'Dowd, P Seeman, and M R Brann. D1 and D2 dopamine receptor mRNA in rat brain. *Proc Natl Acad Sci USA*, 88(5):1859–1863, March 1991. 114

[188] R.O.L. Wong. Retinal waves and visual system development. *Annu Rev Neurosci*, 22(1):29–47, 1999. 92, 93

[189] Margaret T T Wong-Riley. Energy metabolism of the visual system. *Eye Brain*, 2:99–116, 2010. 113

[190] Michael Yourshaw, S Paige Taylor, Aliz R Rao, Martín G Martín, and Stanley F Nelson. Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Briefings in Bioinformatics*, page bbu008, March 2014. 67

[191] Ryan K C Yuen, Bhooma Thiruvahindrapuram, Daniele Merico, Susan Walker, Kristiina Tammimies, Ny Hoang, Christina Chrysler, Thomas Nalpathamkalam, Giovanna Pellecchia, Yi Liu, Matthew J Gazzellone, Lia D'Abate, Eric Deneault, Jennifer L Howe, Richard S C Liu, Ann Thompson, Mehdi Zarrei, Mohammed Uddin, Christian R Marshall, Robert H Ring, Lonnie Zwaigenbaum, Peter N Ray, Rosanna Weksberg, Melissa T Carter, Bridget A Fernandez, Wendy Roberts, Peter Szatmari, and Stephen W Scherer. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature medicine*, 21(2):185–191, January 2015. 59

[192] Ling-Hui Zeng, Yannan Ouyang, Vered Gazit, John R Cirrito, Laura A Jansen, Kevin C Ess, Kelvin A Yamada, David F Wozniak, David M Holtzman, David H Gutmann, and Michael Wong. Abnormal glutamate homeostasis and impaired synaptic plasticity and learning in a mouse model of tuberous sclerosis complex. *Neurobiology of Disease*, 28(2):184–196, November 2007. 55

[193] Ling-Hui Zeng, Nicholas R Rensing, Bo Zhang, David H Gutmann, Michael J Gambello, and Michael Wong. Tsc2 gene inactivation causes a more severe epilepsy phenotype than Tsc1 inactivation in a mouse model of tuberous sclerosis complex. *Hum Mol Genet*, 20(3):445–454, February 2011. 57