Title
Computational methods for the analysis of high throughput genomic data in cancer and development

Author
Pankov, Aleksandr

Publication Date
2016

Peer reviewed|Thesis/dissertation

Computational methods for the analysis of high throughput genomic data in cancer and development

by

Aleksandr Pankov

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Dedications and acknowledgements

The text of chapter three of this dissertation are a reprint of the material as it appears in Mazor T., Pankov A., et al., *DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors.* Cancer Cell, 2015. **28**(3): p. 307-317 and Mazor T., Pankov A, Song J S, Costello J F. *Intratumoral Heterogeneity of the Epigenome.* Accepted at Cancer Cell, 2016.

I would like to start by thanking the fantastic researchers that I had the great priviledge of working with throughout my graduate degree. I would specifically like to thank Tali Mazor, Julia Ye, and Jacob Freimer for sharing their wisdom and technical expertise to create an extraordinary collaborative environment that combined the experimental and computational aspects of genomics. The successful completion of our projects would not have been possible without their extensive knowledge.

I would also like to expressed my gratitude to Dr. Robert Blelloch for his dedicated mentorship and ongoing support in helping to design the novel techniques used for our projects with his lab.

A special thanks to Dr. Howard Rosen and Dr. John Kornak for providing continuing guidance on a non-genomics project to identify shrinking brain regions in dementia. They allowed me to push the boundaries of my learning and expand how I think about the applications of my work.

I would especially like to thank the many researchers whose incredible skills have inspired me to become a better, more well-rounded scientist. I was incredibly priviledged to have Dr. Adam Olshen, Dr. Annette Molinaro, and Dr. Henrik Bengtsson working so closely on these projects, their feedback to my approach helped me focus on computationally-feasible solutions. I was lucky to have Daniel Himmelstein and Dr. Aleksandra Olow as close friends and colleagues. Not only were they always willing to listen and help me with both scientific and daily challenges, but their own scientific dilligence and dedication pushed me to learn new languages and statistical methods to keep up.

A special thanks to Dr. Javier Arsuaga and Dr. Mariel Vazquez for believing in me and giving me a chance to do research as an undergraduate student and helping me find my passion for bioinformatics science.

I would also like to gratefully thank Dr. Mark Segal and Dr. Katie Pollard for their mentorship since the start of my graduate degree. They always found the time to follow up about my projects' progress and help lead me through the challenges.

Finally and most importantly, I would like to thank my mentors Dr. Jun Song and Dr. Joseph Costello. I was incredibly lucky to have worked with such encouraging, supportive, and always innovative mentors. Their passion for science and meticulous attention to detail created an environment that allowed my inquisitive nature to thrive and conduct projects fundamental to biology. They have inspired me to not only pursue scientific research further throughout my career but also taught me to always be excited to challenge myself and get out of my comfort zone.

Computational methods for the analysis of high throughput genomic data in cancer and development

Aleksandr Pankov

ABSTRACT

This dissertation describes the research carried out within the scope of two projects that deal with novel biomedical technologies and their use for advancing medical knowledge through statistical integration of genomics assays. I have explored and characterized epigenetic intratumoral heterogeneity and brain cancer evolution by creating customized statistical analyses and novel methodology for understanding and integrating RNA-seq, methylation arrays, and exome-seq data. To explore functional effects of the Ilf2 RNA-binding protein (RBP) through embryonic stem cell differentiation in mice, I created statistical pipelines to remove data-generation artifacts, applied various testing methods, and integrated the information utilizing small RNA-seq, ribosome profiling, and RNA-seq technologies.

Table of contents

List of tables

List of figures

Chapter 1: Introduction

BACKGROUND

Computational methods are paramount to gaining insights from high-throughput biological data. Increased accessibility of fast processors and high-memory computers created the opportunity to develop powerful high-throughput analytical strategies for scientists from both experimental and computational biology fields. These methods are essential for the fields of cancer and development where research and clinical decisions are directly based on conclusions from the data analysis techniques that are applied. The false positive and true positive rate of predictions will depend on the exact methods and significance definitions used. Therefore, when a method is developed, it is important to consider the accuracy and sensitivity of our results and how our biological conclusion will change when applying diverse analytical approaches.

To tackle a biological question, it is essential to not only understand the underlying cellular and molecular mechanisms, or the techniques that are being used to generate the data, but it is also important to understand what computational methods are most appropriate to interpret the generated data. Throughout my training as a bioinformatician, I have learned to integrate the mechanics and applications of different computational techniques, as well as to create necessary modifications and method innovations in order to address the questions with the most appropriate approach. Along these lines, to understand the evolution and tumor heterogeneity of Glioma, I adapted classical methods to integrate the high-throughput genomic data available from somatic mutations, DNA methylation arrays, and RNA expression measures under the complex experimental design employed in that project. Likewise, to understand how removing a RBP changes

the differentiation potential of embryonic stem cells, I created novel analytical strategies by extending currently used methods and developing new techniques for genomic analysis.

USE OF STATISTICS IN GENOMICS

Statistics is used to characterize and model the random variables in the population of interest. In fact, many methods are based on underlying models that simplify the biological process involved. While these models could be applicable in specific situations and apply well to certain scenarios, their performance is not guaranteed across all similar data types or experimental designs. To understand when a method is appropriate, it is essential to understand the underlying assumptions and understand the outcome of making such assumptions. Importantly, even though only a handful of data points are observed, conclusions are being inferred about the entire population. Thus, the observations produced by an experiment are extrapolated using an assumption of the underlying distribution of the data in order to understand the important parameters of the population.

Statistical methods used in the analysis of genomic data can be broadly classified into 2 main categories: supervised and unsupervised learning. Supervised methods attempt to identify the functional relationship between a predictor variable and a response. By contrast, unsupervised methods are used to understand the similarity of the sampled data points and observed variables and find the hidden structure within the data without any response variable that can help guide the separation.

CLUSTERING

A fundamental problem in biology is to identify groups of data points that are more similar to each other than they are to all other data points in a sample. To understand how high-dimensional data points are grouped together, clustering is a general technique that can be used. Throughout this work, I have specifically used hierarchical clustering. This method uses the distances between data points to identify the two most similar data points, join those points into a group, recalculate the distance between all data points and groups, and continuously find the most similar data points or collections and join them until all points are contained in a single group. This allows us to understand the higher order similarity between data points. I have extensively used this technique in understanding the similarity between tumor samples when studying how low-grade glioma undergoes malignant progression.

PRINCIPAL COMPONENT ANALYSIS

In a related problem, it is often inefficient to perform computation in the original high dimensional space that each sample is represented in. For example, every sample has over 30,000 gene expression measurements, and while comparing pairs of samples, it is often not feasible to compute the high-dimensional distributions that represent the entire sample population. Thus, principal component analysis (PCA) was developed to represent the high-dimensional data as accurately as possible using only a smaller number of dimensions. PCA removes any redundant information across the high-dimensional space and condenses the data to a smaller number of dimensions that is able to faithfully

recreate the value of each sample and maintain the relative distance between samples. The PCA procedure first finds the direction that can best separate all samples of a data set, called the first principal component; this direction is some linear combination of the initial features of the data. It then identifies the next principal component as one that is orthogonal to the first principal component and is able to best separate the samples after the separation from the first principle component has been taken into account. This process is repeated until the desired number of principal components needed to approximate the data is extracted. PCA is a valuable method to reduce the dimensionality of a dataset while maintaining important information necessary to conserve separation between the data points. This technique provided the foundation for identifying which variables are most important for creating a branching event in a phylo-epigenetic process.

CLASSICAL T-TEST

A common supervised learning question that is being asked in biology is whether there exists a difference in the distributions of variables between two different groups. Even though this particular question remains ambiguous as to how a difference between distributions is defined, statisticians have been studying a related question for more than a hundred years since the formation of the t-test. To further define how two distributions are different, the question was narrowed down to ask if the central tendency, specifically the mean, is different between the two distributions. Specifically, the classical t-test uses the assumption that there is only a single underlying distribution that generates both of the distributions for the two groups, and that it computes the difference between the means of the two groups, determines a common standard deviation and a scaling factor to

adjust the influence of the standard deviation based on the observed sample size, and estimates what the probability of observing a difference of such a magnitude while considering the spread of the data. However, when there is a small sample size and the two groups are not generated from a normal, the classical t-test could give misleading results.

CLASSICAL T-TEST IN GENOMICS

The classical t-test has provided an important step in understanding how to interpret biological data, but is insufficient for determining differences between groups in genomics data. Unlike traditional experimental design, where an emphasis was placed on gathering a larger sample size to better estimate the parameters of the distribution, experiments with genomic data often emphasize measuring many variables over collecting a larger sample size. To account for the reduced sample size and the technical artifacts that are found in high-throughput technologies collecting multivariate measurements, a novel scenario was created and lead to the development of a new approach in the testing for a difference in distributions.

LIMMA: AN EXTENSION TO THE CLASSICAL T-TEST

Limma (Linear Models for Microarray Data) [1] testing is the procedure that was created to help alleviate such complications associated with genomic data, and is an expansion of the classical t-test. The limma software also incorporates a general framework for analyzing gene expression experiments from both microarray and sequencing data. This method smooths out the variance of each gene by a scaling factor that pushes each

estimated variance toward a global gene variance. This method offsets any errors in estimating the standard deviation due to the particular signature of data generated from various genomic assays. The smoothing factor is in turn estimated from the data itself by an empirical Bayes estimation method that assumes a shared hyperparameter across all variables being measured [2]. This methodology is now prominently used in microarray and sequencing data analysis, and has provided the framework to understand the changes in methylation and transcriptional profiles upon malignant transformation.

T-TEST EXTENSIONS TO DISCRETE SEQUENCE DATA

While both the limma and classical t-test are meant to calculate the difference between two groups with normal, continuous distributions, sequencing data is discrete. Thus, for count-represented sequencing outputs, a Poisson regression method is more appropriate to model the parameters and tests for a difference between groups. On the downside, the Poisson regression approach makes a strong assumption that the mean is equal to the variance which has been shown not true for most sequencing data. Thus, a negative binomial regression approach has been suggested to be more appropriate, since it allows for an additional parameter that has a larger variance than allowed by Poisson regression. This approach has been further extended similar to how limma extended on the classical t-test, and models the variance with an additional parameter that makes a variable have a more similar variance to the average variance of other variables with similar mean counts. This methodology has been implemented in DESeq2, a tool that was prominently used in this dissertation [3]. These extensions have been specifically incorporated when

looking for changes in translation as well as transcription throughout embryonic stem cell differentiation.

CORRECTING FOR MULTIPLE TESTING

A further complication in genomic data analysis occurs when estimating probability of a difference between groups or distributions across a large number of variables. Previously, when testing was done on an individual variable, the probability of observing a difference under the null hypothesis was relatively straightforward to assess. However, the probability of observing a small difference is likely to occur even when randomly generating data from the exact null hypothesis. This has become especially important when the number of variables that are being tested is more than 10,000. To visualize this, we could imagine data randomly generated from the null distribution, with probabilities uniformly distributed between 0 and 1. If there are variables that are truly different, they will form a uni-modal distribution with a mean trending towards 0. Storey's method [4] and the Benjamini-Hochberg procedure [5] for adjusting the probabilities takes advantage of these observations and is further applied to genomic data to estimate more accurate probabilities when testing for a large number of variables. This is a fundamental step in deciding which genomic variables are important and need to be experimentally validated.

INTEGRATED GENOMICS ANALYSIS

Moreover, correcting for multiple hypothesis testing is also an essential aspect of integrating multiple genomic data sets. In particular, since different genomic assays produce a different numbers of variables that are being tested, it is necessary to adjust

their probability estimates to be comparable before making conclusions that are biologically meaningful. This key concept of integrated genomics data analysis is further developed and incorporated throughout the projects of this thesis.

Chapter 2: Understanding the role of the RNA-binding protein Ilf2 in differentiation

INTRODUCTION

Understanding the regulation underlying how embryonic stem cells (ESCs) differentiate to an early somatic lineage is pinnacle to biological science. The regulation of cell fate transition is important not only for expanding our basic understanding of how mammalian blastocysts develop into different tissue types, but also for various clinical applications. A better understanding of pluripotency regulation and embryonic development allows ESCs to be effectively utilized for disease modeling [6], drug discovery [7], and tissue regeneration [8].

While epigenetic and transcriptional regulation has been a primary focus [9-11], post-transcriptional regulation of ESCs has only started to be studied [12]. It has been previously shown that individual microRNAs (miRNAs), which destabilize and inhibit translation of their messenger RNA targets, and long non-coding RNAs, which have been shown to act both as an activator and a repressor, can drive profound cell fate shifts between pluripotency and differentiation [13] [14]; however, differentiation regulation by miRNAs and lincRNAs do not capture the entire complexity of post-transcriptional control in ESC development.

RNA-binding proteins (RBPs) are important for regulating different machinery within the cell; they have been implicated to play a major role in regulating splicing, nuclear export, stability/storage, localization (most studied in highly polarized cells like neurons and oocytes), translation, decay (deadenylases, decapping enzymes, exonucleases), as well as other process [15]. For this project, we are interested in discovering which RBPs are

important for pluripotency and differentiation; to this extent, it is vital to determine which RBPs play a role in the ESC to Epiblast (EpiSC) transition.

REPORTER SYSTEM

To evaluate the role of RBPs in ESC differentiation, my collaborators in the Blelloch lab have developed a dual reporter system to probe ESC differentiation. In recent findings [16], the Blelloch lab has discovered that during normal embryonic development, all cells initially transition from expressing only the miR-290 cluster to expressing both the miR-290 and miR-302 clusters and then to only expressing the miR-302 cluster. Thus, they implemented a knock-in fluorescent reporter system containing the promoters of these two miRNA clusters, the miR-290, which is labeled with mCherry, and miR-302, which is labeled with GFP. This fluorescence system allows embryos to express Red (R) at the time of the inner cell mass from which ESCs are derived and then turn Yellow (Y), when expressing both mCherry and GFP, at the epiblast stage and finally only Green at gastrulation. Monitoring the well-defined ESC-to-EpiSC transition in a culture dish minimizes issues of cellular heterogeneity (Figure 2.1).

STUDY DESIGN

With this well-established reporter system, it has become possible to systematically track the ESC-to-EpiSC transition. This allows a direct way to monitor phenotypic information while also associating the molecular signature promoting that state. To study which RBPs are important to differentiation, we decided to identify all potential RBPs whose expression changes through the cells transitions. After directly perturbing each protein's

expression and understanding what effect it had on the transition phenotype, we would select a single protein to functionally profile.

RBP SCREEN

To select the specific RBPs for further evaluation of their functional role in ESC differentiation, the Blelloch lab first subset the entire portfolio of previously identified RBPs to 356 proteins that are differentially expressed between any stage of the ESC differentiation system. Each potential RBP was then inhibited utilizing a 96-well siRNA pools screen. The effect of siRNA on differentiation was monitored through a GFP-mCherry system. The RBP siRNAs were arranged alphabetically in the C – 3 to F – 10 (row – column) section of a 96-well plate and each plate is replicated three times (Figure 2.2). The experiment had four different negative siRNA controls ("sictrl1" – "sictrl4"), each of which has three technical replicates on each biological replicate of each plate. Additionally, the siOct4 were used as negative controls due to high toxicity and resulting complete cell death. The additional controls were procedural controls and included the negative GFP controls and viability controls (e.g. delivery agent without an siRNA); for this analysis they were referred to as "other" controls. Overall, there were 12 plates each one in triplicate (Figure 2.3).

Each of the wells was followed everyday using the InCell high throughput, high content microscopy system, providing a time course for the transition from miR-290 expressing ESCs to miR-302 expressing EPLCs. After 4 days, the experiment was stopped and cells

were stained with DAPI. Using InCell software with custom designed macros, wells were evaluated for relative levels of red to green.

Computational challenges

From this RBP screen, we were interested in identifying the proteins that most severely affected the differentiation phenotype. To complete this task, I first needed to establish which plate readouts are the most informative. Every plate has a certain amount of GFP and RFP expression that corresponds to the size of the cells, where a larger cell will be expected to emit more fluorescence. I first had to identify which variables are the best measure of fluorescence and then how to adjust for a change in cell size. After creating a normalized measure of fluorescent intensity, I would need to create a testing strategy that would identify proteins that are changing more than would be expected in the experimental control. However, due to the small number of replicates, it would be beneficial to pool the information across all plates in this experiment. Thus, my next steps were to determine the per-plate and per-replicate batch effects and, consequently, to adjust out that non-biological influence. The final steps were to identify which proteins have the most influence on differentiation and determine which protein of interest would be a candidate for functional profiling.

Exploratory analysis

Fluorescence area relationship with mean density x area

First, by examining scatterplots between *fluorescence area* and *fluorescence mean-density x area* (Figure 2.4), I concluded that because a simple (but non-linear)

relationship could describe how the two variables are related, using just the *fluorescence area* would be sufficient for determining which RBPs influence differentiation.

Fluorescence area relationship with DAPI area

Next, visually examining the *fluorescence area* and *DAPI area* relationship showed a nearly linear relationship between the variables. While the *RFP* vs. *DAPI* plots showed a highly linear relationship, the *GFP* vs. *DAPI* plots showed much more variance (both within GFP control wells, as well as excluding those)(Figure 2.5). Even when excluding "other" controls, the *RFP* and *DAPI* had a mean Pearson correlation of 0.986 (ranging from 0.968 to 0.999) and the *GFP* and *DAPI* had a mean Pearson correlation of 0.934 (ranging from 0.824 to 0.987)(Figure 2.6). This suggested that little additional information is gathered from using the *RFP area* in addition to the *DAPI area*.

Quality control

Before further analysis was done, it was important to remove the wells that "failed", i.e. where the majority of cells died. The simplest approach would be to remove any siRNAs that are below the maximum of the *DAPI* control within each replicate. Another approach was to normalize the *DAPI* measure (through the normalization method described below) across replicates and across plates, then to discard any siRNA's values falling within a threshold defined by the distribution built from the *DAPI* controls (Figure 2.7). For this analysis, the *DAPI* readouts were normalized across all plates and all entries whose *DAPI area* was less than the maximum from the *DAPI* controls were discarded. This

conservative approach aimed at minimizing the chance of identifying RBPs whose knockdown produced dead cells (Figure 2.8).

Normalization

Across replicates normalization

To combine the data for a plate, we first needed to normalize it across all its biological replicates. To do this, for each replicate, I would first shift the data by the location estimator defined in [17] and scale the data by the scale estimator from the same source (using the scaleTau2 function defined in R's robustbase package [18]), thus normalizing the data across the replicates for each plate.

However, this exact procedure could not be used to normalize data across all plates because it contains the implicit assumption that the distributions are the same across plates, which might not be valid for this experimental design.

Across plates normalization

To normalize across all plates, I made the assumption that the distribution of all the controls within themselves (not including the RBP siRNA knockdowns) was identical on each plate. Therefore, the distribution of the control siRNAs for each plate was used to calculate a robust measurement of location and scale by which to shift and divide that plate's data. This approach re-scaled the data to describe each observation in terms of robust control deviations away from the control center.

Identifying significant RBPs

Finally, to identify the significant RBPs that influence ESC differentiation, I kept all data points with at least three entries with *DAPI Area* above controls. The difference of the log *fluorescence area* and log *DAPI area* is calculated to account for the total amount of cells available. I then proposed the following approaches:

1. Normalizing the log *fluorescence data* across all replicates. Then, for each plate separately, applying a simple t-test with equal variance for each RBP siRNA against the combined siRNA control of that plate. Then calculating the q-value of each test using the qvalue package in Bioconductor. (Figure 2.9)

2. Normalizing the log *fluorescence data* across all replicates and all plates. Then applying a simple t-test with equal variance for each RBP siRNA against the control siRNA combined over all plates. Then calculating the q-value of each test using the qvalue package in bioconductor. (Figure 2.9)

3. Creating a null test statistic distribution by taking the four different siRNA controls and comparing them against each other to identify the background distribution of t-statistics. Similarly, we could be subsampling the siRNA controls to determine a background distribution in that way. However, I do not believe that this would create a major difference for the majority of the RBPs. Thus, I left this approach as a future direction of the project.

4. It is important to identify the optimal cut-off for determining significant RBPs. The simplest approach was to use a cutoff of the q-value. However, this approach ignored other controls that could be useful in assessing an RBP's importance. For example, the dfect control, which measures the effect of transfecting the cells with a reagent without any siRNA, often seemed to show up as "significantly" different from the siRNA controls. So, I only considered RBPs significant if they were below a q-value cutoff and had a q-value less than the dfect control (Figure 2.10).

5. We also had additional controls that were not used in the analysis, but might contain information to distinguish the signal from noise. Specifically, we had the siGFP knockdown that always reduced the GFP fluorescence and was highly significant. Similarly, we had the 2li control, which is a media additive that prevents differentiation and also highly significantly reduced the GFP fluorescence. These two controls consistently showed the most significant change compared to the siRNA controls. Finally, we had the fluorescence signal resulting from just the media in the "nothing" control. This control could have also been used like the dfect control, as an additional cutoff to identify significant RBPs. This aspect, of using positive controls in addition to the negative controls, was not been implemented.

Results

The t-statistics, p-values, and significant RBPs are included in (Tables 2.1 and 2.2). The results from testing siRNA controls against RBPs per plate, as well as across all plates are presented in (Figure 2.11). In the across all plate analysis [2 in the above section], using a q-value of less than .01 and requiring that the q-value was less than the q-value of the "dfect" control identifies 25 "significant" siRNAs (including controls); however, requiring that the q-value was less than the q-value of the "nothing" control and less than .01 identifies only 5 "significant" siRNAs (including controls) out of 150 siRNAs total. In the by plate analysis [1 in the above section], using a q-value of less than .001 and requiring that the q-value was less than the q-value of the "dfect" control identified 83 "significant" RBP siRNAs; however, requiring that the q-value was less than the q-value of the "nothing" control and less than .001 identified only 49 "significant" RBP siRNAs out of 146 total RBP siRNAs. From these results, it seemed the across plates analysis using the dfect control cutoff gives the most biologically-relevant, yet slightly conservative, significant RBPs.

Caveat

The biology suggests that it would be easier to prevent the transition from ESC to EpiSC using siRNAs than it would be to promote it. In other words, it should be expected that most siRNAs will disrupt the transition process. This phenomena can also be suggested from an empirical evaluation of the t-statistic of the GFP measure in (Figure 2.12). The shift of the center of the distribution towards negative values suggests that most siRNAs only have a small change from the ESC state. However, the current version of the analysis did not take this assumption into account. One way to account for it would be to

use a non-symmetric distribution when testing for the significance of a *fluorescence area* difference.

IDENTIFYING ILF2 AS A MASTER REGULATOR

After a critical evaluation of the statistical analysis that was performed, our collaborators in the Blelloch lab identified Ilf2 as an essential RBP to evaluate its role through differentiation. Among the screen hits, knockdowns leading to accelerated transition to the EpiSC state were especially important and Ilf2 was one of those hits. Ilf2 was first discovered as a transcriptional activator of IL2 in T cells [19] and has since been identified as an RBP in three independent mammalian RBP profiling studies [20-22]. Ilf2 has been proposed to function in post-transcriptional and translational regulation of RNA, suggesting that it may be acting an essential role in ESC development. Additionally, Ilf2 has been described to regulate miRNA biogenesis, preventing the processing of pri-let-7 to pre-let-7 [23], affect splicing by being a part of the exon junction complex (EJC) [24], and has been connected to an internal ribosomal entry site (IRES) trans-acting factor that binds to AU-rich sequences in IRES in the 5' UTR of several transcripts to modulate the translation of both viral and cellular proteins [25-28]. Furthermore, Ilf2 has been found in a complex with the pluripotency transcription factor Nanog [29] and is down-regulated in Nanog-depleted ESCs [30].

To understand the effect of functional role of Ilf2, our collaborators in the Blelloch lab acquired Ilf2 knockout (KO) cells from Dr. Kyoji Horie who had homozygosed ESCs with single allele gene traps. To begin to dissect the molecular roles of Ilf2, we have

performed PAR-CLIP, RNA-seq, small RNA-seq, and ribosome profiling on wild-type [31] and Ilf2-knockout ESCs.

TRANSCRIPTOME PROFILING OF WT AND KO ILF2 CELLS

Using a strand-specific protocol that selectively sequences poly-A-tailed mRNA transcripts, a library of the transcriptome sequences of both the WT and KO cells was created. After applying the Kallisto [32] method for isoform quantification, I identified 1046 isoform changes in the knockout cells (q-value < 0.05 and an absolute log2 fold change greater than 1) by the Sleuth testing method followed by an adjustment for multiple hypothesis testing. The most-enriched molecular functions and biological processes associated with both the upregulated and downregulated genes can be found in (Figure 2.13).

Due to the Ilf2's role in the EJC and presumed importance in regulating RNA splicing, it is important to detect alternative splicing events that are occurring between the WT and KO cells. This is substantially different from identifying differential isoform expression between the two conditions because it involves a change from one isoform to another of a particular gene rather than just an expression difference. To study this change, I applied the MATS software [33] to determine that identified the following list of events in (Table 2.3).

RIBOSOME PROFILING OF WT AND KO ILF2 CELLS

Due to Ilf2's role in regulating translation, it was essential to measure the translation of the transcripts present in a cell. To measures changes in translation rates, ribosome profiling was performed in both the WT and KO cells. The technique involves a short pulse of cycloheximide followed by cell lysis, DNase treatment, and gentle RNase treatment. Ribosome-bound fragments of the RNAs are protected from the RNase digestion and therefore are enriched in nuclease treated versus untreated fractions. Both fractions are transformed into RNA-seq libraries following rRNA depletion. To analysis the sequencing data produced by ribosome profiling, we adapted a commonly used protocol.

Alignment

1. Clipped adapter sequences off the ends of all reads and trim all bases at base quality of 1 or below
    a. Only kept reads with an adapter in the Ribosome Protected samples
    b. Kept all reads in the mRNA samples
2. Aligned all reads against possible contaminants using bowtie [34]
    a. Contaminants include: polyA, polyC, adapter, ChrM, phix, Ribosomal RNA, tRNA
3. Aligned all the reads that did not align to contaminants to the mm10 genome using transcriptome-guided tophat2 alignment [35].
    a. Transcriptome model used was Gencode annotation [36]
4. Filtered all aligned reads to those that did not:
    a. Fail QC (bit flag set by the aligner)

b. Map to random chromosome

c. Aligned to multiple positions on the genome

Assess Experiment Efficacy

After alignment, we assessed the quality of the data as follows:

1. I investigated the coverage along the CDS of genes
    a. For all regions in the transcriptome annotation, I added a pseudocount of 1 as locations previously identified as being expressed.
    b. To avoid any bias from using multiple isoforms of the same gene, I collapsed each gene to only represent a single isoform. I first ranked the average isoform coverage within each sample; then, I chose the isoform with the highest average rank across all samples. Thus representing a single isoform for each gene and displaying the same isoform in all samples.
    c. The genes were then ranked according their average coverage.
    d. To normalize for the length of the genes, I divided each gene into 50 bins and calculated the average coverage per bin.
    e. To represent the change of coverage along the gene, I calculated the Z-score of each bin (Figure 2.14).

2. I also investigated the codons that occur with the most enriched 3-mers across all genes.

a. For regions in the transcriptome annotation, I added a pseudocount of 1 as locations previously identified as being expressed.
b. I calculated the coverage in a running window of size 3 divided by the mean of that coverage per isoform.
c. I then identified the maximum enrichment score and its location along the isoform.
d. Per gene, I used only the isoform that has the highest maximum enrichment score.
e. These calculations were done within frame for only the CDS data.
f. I plotted the enrichment of codons associated with the maximum enrichment scores for the CDS regions (Figure 2.15 and 2.16).

Identified genes with differential translation rates

After confirming that the experiment has worked correctly, we went on to measure expression in the CDS region of each gene for both the mRNA samples and ribosome protected samples. We then used the following linear model to test for differential translation rates between the two conditions:

$\log(mu) = a + b_{\text{ribo}}\, \text{I}[37] + b_{\text{KO}}\, \text{I(KO)} + b_{\text{cross}}\, (\text{I(ribo) x I(KO)})$,

where I(.) is an indicator that equals 1 when the sample satisfies the condition and 0 otherwise. Here, I modeled the log average of the expression as a linear combination of the overall average expression, a, the difference between the WT and KO cells, b_{KO}, the difference for the ribosome protected fragments, b_{ribo}, and the difference of the ribosome protected fragments within the KO condition, b_{cross}. Then by assigning a p-value to the

b_{cross} coefficient, I was able to determine which genes are consistently changing in their translation rates between the conditions after accounting for their expression levels by DESeq2 [3] (Figure 2.17). This analysis identified 252 significant genes (p-value < .05).

MIRNA PROFILING OF WT AND KO ILF2 CELLS

As a crucial first step to exploring the miRNA profile of our KO and WT samples, a quantification of the miRNAs within the cell must be computed. To quantify the miRNAs, I wrote an aligner specifically for the characteristics involved in small RNA sequencing that takes into account redundant reads and a small reference set to quickly align the sequencing data on a desktop machine. First, I determined that the length of each sequencing read is fairly short, usually between 17 and 24 nucleotides in length (Figure 2.18). Next, I recognized that the MirBase microRNA reference database [38] is small with only 1193 precursors and 1915 mature for the mouse genome, where each hairpin is less than 100 nucleotides. Additionally, since the target space is so small, we can expect to have many identical short reads being sequenced. Thus it is best to align each read only once and maintain quick access to the mapping information. To preserve memory usage, I used Hoffman encoding to compress each read's sequence content before storing it.

Alignment

The following is the algorithm for mapping small RNA sequencing reads against a miRNA reference:

1. Read in the mature miRNA reference, break up each reference miRNA into *k*-mers, and hash each *k*-mer with a pointer back to the miRNA sequence that it originated from.
2. Redo #1 for the precursor sequences as well.
3. Map the location of the mature miRNA sequences within the precursor sequences.
4. For each read in a sequencing library do the following
    a. Break the read in *k*-mers
    b. Check the *k*-mers against the known hash tables generated from the mature sequences
    c. If all the *k*-mers overlap with a specific mature miRNA, do an exact pairwise alignment against that mature miRNA to determine if the read originated from there and assign the read as belonging to that mature miRNA.
    d. If the *k*-mers do not overlap with a specific mature miRNA or the read does not originate from the mature miRNAs, repeat the mapping procedure against the hash table of the precursor miRNAs and similarly determine if the read originates from any precursor.
    e. If a read originates from the precursor, determine if it is a read that spans part of the mature and precursor miRNA and assign it to the mature.
    f. Compress the read and add it to a set
    g. Quantify for all reads.

Using this simple mapping procedure, I was able to achieve fast and accurate quantification of gigabyte-scale small RNA sequencing libraries.

Testing for differential expression

After I aligned each read against the miRbase database, I quantified the reads falling into the following categories:

- Reads mapping perfectly to a single mature miRNA reference
- Reads mapping above 90% and below 99.9% similarity to mature miRNA but mapping perfectly to a single hairpin reference

This allowed me to estimate the expression at the most unique level quantifiable. After comparing different normalization techniques, I conclude that even though the log CPM transform has the best variance stabilizing properties when comparing to CPM and anscombe CPM (Figure 2.19), the anscombe CPM should be used because the number of false negatives with large changes in expression decreases using that transform.

To guarantee that the miRNAs that I was testing had a large enough concentration in the cells to have a functional effect, I subset the miRNA to only test the ones that have an expression measurement of CPM > 50 (Figure 2.20). After running a simple limma model to determine which miRNAs are changing between the KO and WT conditions, I was able to narrow down the list to a handful of significant miRNAs in (Table 2.4).

Testing cis-transcribed miRNAs for differential expression

Here, I tested different groupings of miRNAs. To do this, I used anscombe CPM as the normalized expression measure and used the same moderated t-test as before. However, I used the expression values of all hairpins associated with a grouping and created

indicator variables for each individual hairpin (in addition to the WT vs. KO state) to account for individual hairpin expression values. For a grouping *j* with two hairpin members, the equation is estimated:

$CPM_j = b_{ko,j}\, x_j + b_{h1,j} h1 + b_{h2,j} h2 + \epsilon_j$

where CPM_j is a vector of CPM expressions for that grouping, x_j is an indicator variable with 1 if the measurement is for the KO condition and 0 otherwise, h1 is an indicator variable with 1 if the measurement is for the first hairpin in the grouping and 0 otherwise, and h2 is an indicator variable with 1 if the measurement is for the second hairpin in the grouping. After running this test across all groupings, I was able to identify the cis-transcribed miRNA groupings that are different between conditions in (Table 2.5).

Testing functional families of miRNAs for differential expression

After identifying the differentially expressed cis-transcribed miRNA clusters, we then asked whether expression differs from functional families of miRNAs which are defined through target scan. Unlike the cis-transcribed clusters, since these miRNAs are acting on the same target element, it is more appropriate to aggregate the counts across all miRNAs of the same family. After aggregating the miRNA quantities over the functional family, we were able to test for differential expression and identify the most significantly changing targetscan-defined families of miRNAs in (Table 2.6):

INTEGRATING MIRNA AND ISOFORM EXPRESSION CHANGES

While we have shown that a number of changes occur in the expression of a particular isoform as well as the quantity of a miRNA between the KO and WT conditions, an

integrative analysis would confer information about direct targets of Ilf2. It has been previously shown that RBPs provide the regulation to express either an isoform or a miRNA within the intron of that isoform while downregulating the other member of the interaction. Thus, I extracted all the miRNAs within the intron of an isoform. To identify the changing isoforms associated with a changing miRNA, I first checked how comparable the miRNA analysis is to the RNA-seq differential expression. I created the qq-plots comparing the p-values, coefficients, and test statistics between the hairpin testing and the isoform testing procedures. One thing to notice is the departure of the p-value relationship from the y=x line, indicating that the p-values are considerably different between hairpin and isoform testing (Figure 2.21).

To correct for this difference in p-value distributions, I normalized the p-values of both test statistics to be comparable to each other. It is important to rank-normalize all the hairpin p-values and all the isoform p-values, not just the ones containing an association between the two groups. This provided a much more similar distribution of p-values between the two testing sets (Figure 2.22).

After associating all the intronic hairpins with the isoforms where those hairpins are found in introns, I plotted the p-value, coefficient, and test statistic relationships (left to right) of all the hairpins and their isoforms (Figure 2.23). (Since this is a many-to-many relationship, a hairpin may appear multiple times and the same for isoforms). From the plots below, there did not seem to be a clear relationship between intronic hairpin expression and the expression of isoform where it. In the left most plot, there did

not seem to be a trend p-value in hairpins and p-values in isoforms. In the middle plot, there were a handful of points where a positive change estimate in the hairpin is associated with a negative expression estimate in its isoform (bottom right quadrant).

In the right plot, most points fall uniformly in a square around (0,0) and there are only a small number of instances that were outside that square.

In the middle and left plots, the size of each point is proportional to how significant the hairpin is (larger points had more significant hairpins) and the hue of the blue is proportional to the significance of the isoform (the more blue, the more significant the isoform was).

However, when I subset to only the hairpins that were significantly differentially expressed and examine their relationship with the isoforms that overlap that hairpin's genomic location, we can see a modest positive correlation between the isoform and hairpin changes across the cell types.

CONCLUSION

These preliminary studies provide much of the data required to develop and validate a bioinformatically derived network of Ilf2-regulated genes. However, these experiments did not provide conclusive evidence of whether these changes are direct or indirect effect of Ilf2 inactivation. To understand which genes are directly affected by Ilf2, the Blelloch lab sequenced the RNA fragments bound by the Ilf2 protein. That experiment showed us that there were very few bound RNA fragments and not more than would be expected by

chance. This led us to suggest that Ilf2 was in fact not directly binding RNA targets but was working through its binding partner Ilf3 to control differentiation. Ilf3 has already been shown to be a DNA-binding protein [39] and is likely a master-regulator RBP. Further works is still required to determine the function of Ilf3 in these cells, as well as to understand the importance of the interaction of Ilf2 and Ilf3 to differentiation.

FIGURES

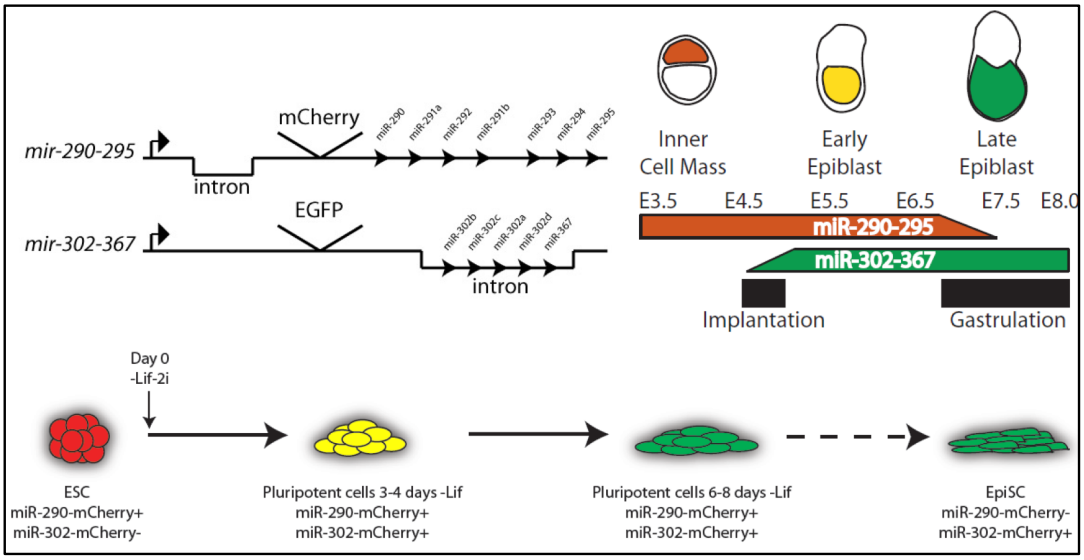


Figure 2.1. ESC to EpiSC Reporter system.
The Blelloch lab has developed knock-in fluorescent reporters for two miRNA clusters associated, each of which has been shown to be associated with the ESC to EpiSC transition. [From Julia's Presentation]

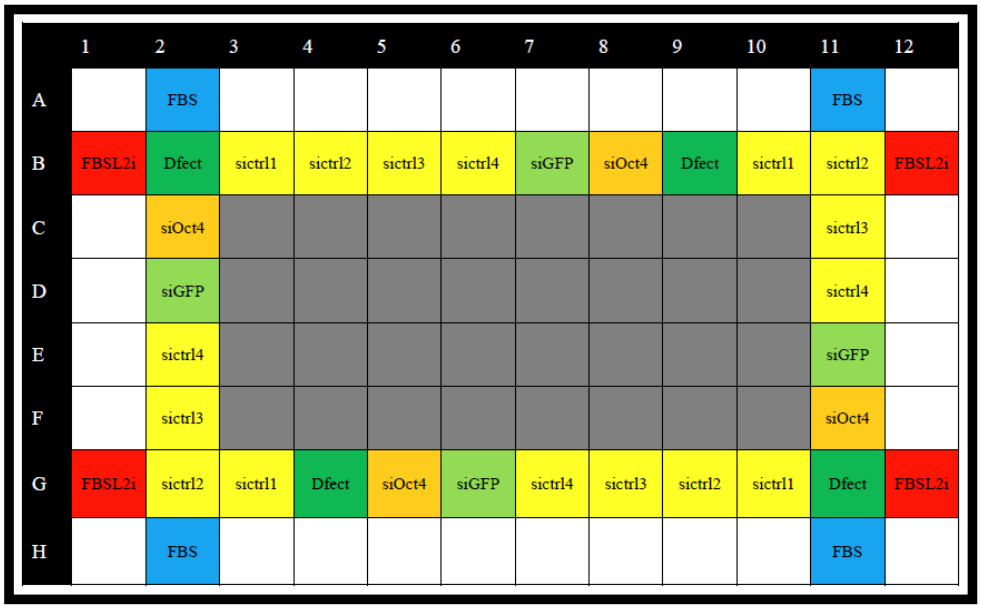


Figure 2.2. Plate design visualization.
The blue wells are media controls, labeled "nothing" throughout the analysis; the red wells contain media with an additive that prevents differentiation; the light green wells contain siGFP; the dark green wells indicate controls to measure the effect of transfecting reagent without an siRNA; the orange marks wells where all the cells died and are used as DAPI controls; the yellow wells contain non-RBP siRNAs; and the grey wells contain the RBP siRNAs.

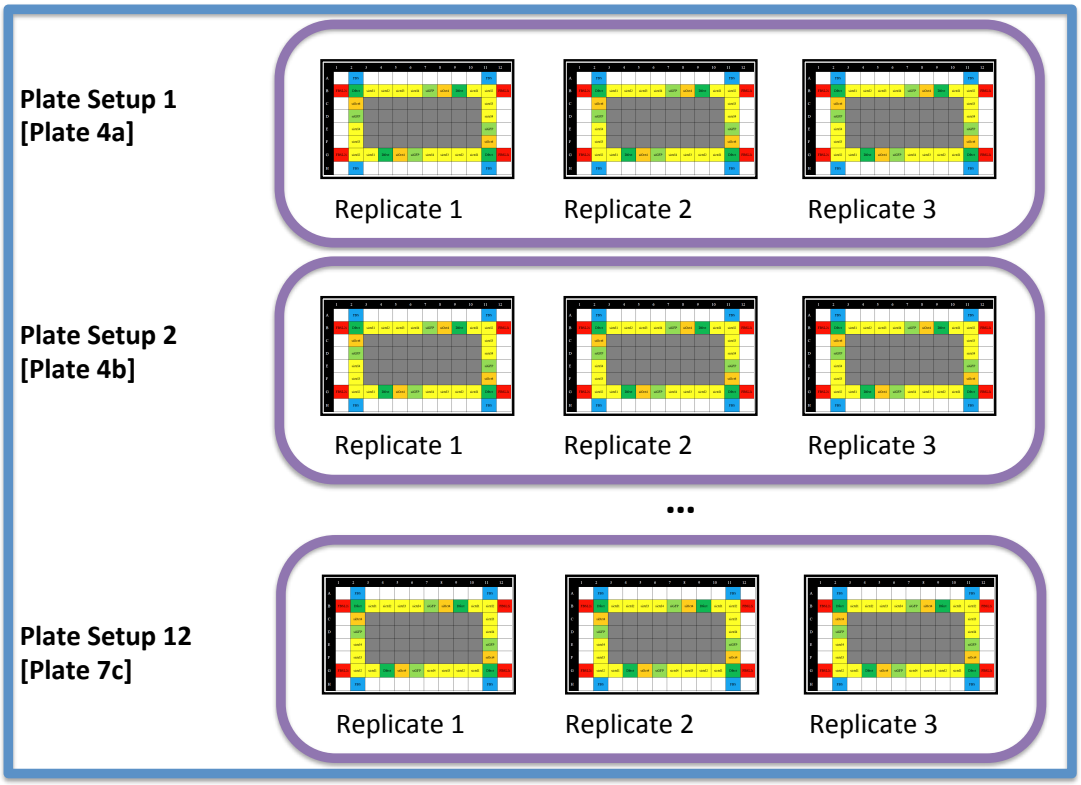


Figure 2.3. Experiment Design.

Each experiment consists of 12 unique plate setups, performed in triplicate.

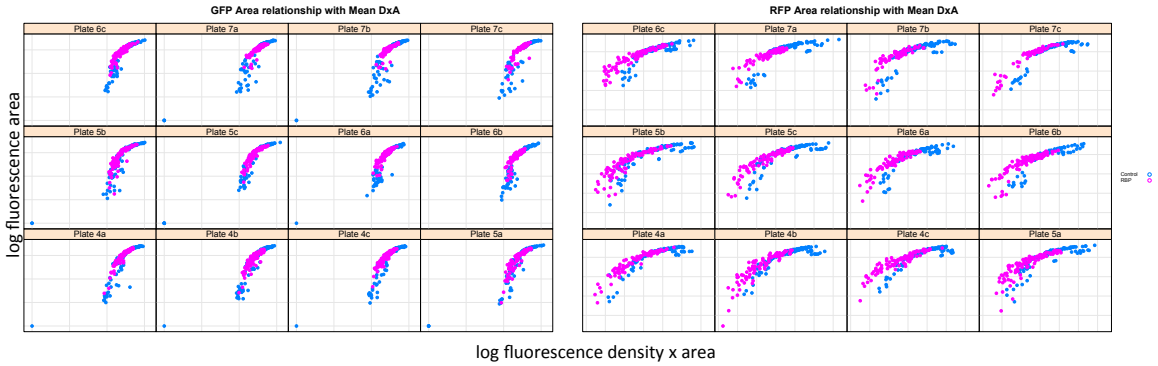


Figure 2.4. Scatterplots of fluorescence area and mean density x area relationship across plates.
The left panel shows the relationship for the GFP reporter, while the right panels shows it for the RFP. Here we can see that a simple, but non-linear, relationship could describe how the two variables are related.

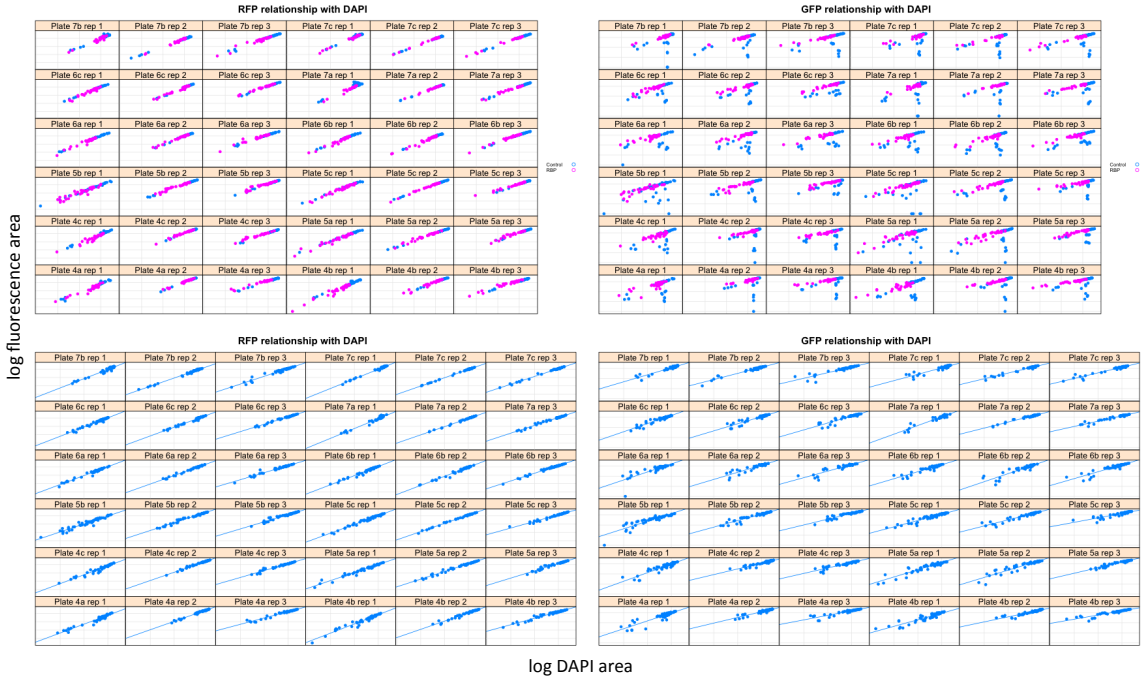


Figure 2.5. Scatterplots of *fluorescence area* and *DAPI area* relationship across replicates.
The top left panel shows the relationship for the RFP reporter, while the top right panel shows it for the GFP; the blue color signifies the controls and the RBPs are marked in magenta. The bottom left panel shows the relationship for just the RBP siRNAs for the RFP reporter and the bottom right panel shows it for the GFP reporter. The *RFP* has a very strong correlation with *DAPI*, suggesting that knowing the reporter Area gives minimal additional information above just the *DAPI area*.

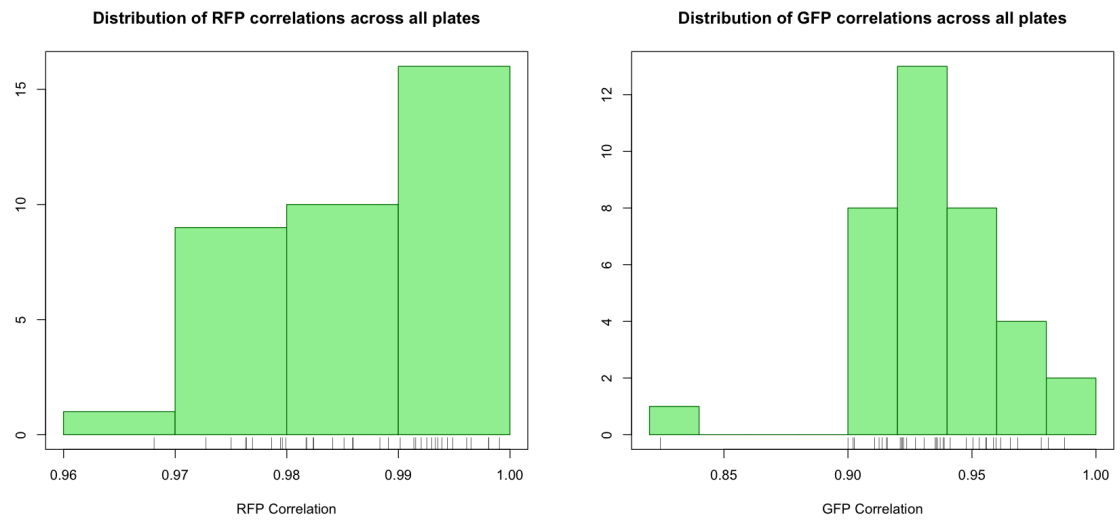


Figure 2.6. Histogram of correlation between *fluorescence area* and *DAPI area* across all replicates.
The left panel shows the distribution for the RFP reporter, while the right panel shows it for the GFP. The high correlation between the *RFP area* and *DAPI area* suggests that not much additional information can be incorporated from the reporter above what is available in the *DAPI*. The *GFP* is less correlated and could contain more information than using the *DAPI* only.

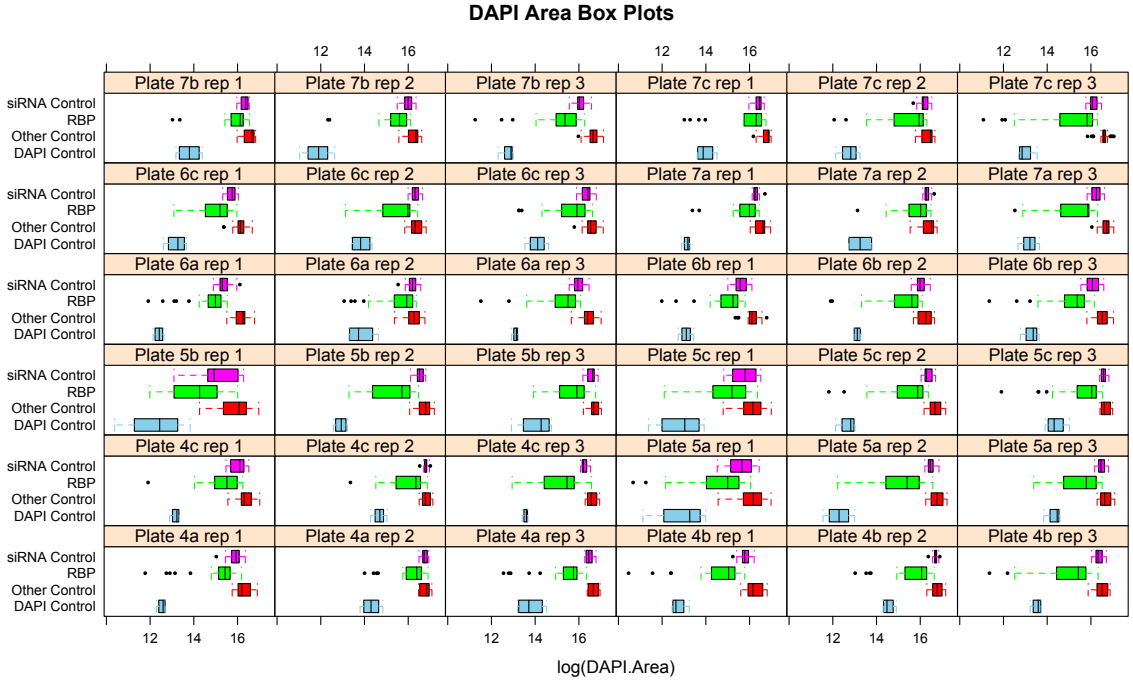


Figure 2.7. Boxplots of the log *DAPI area* per replicate.
The siRNA controls are in purple, the RBPs are in green, the *DAPI* controls are in blue, and the "other" controls are in red.

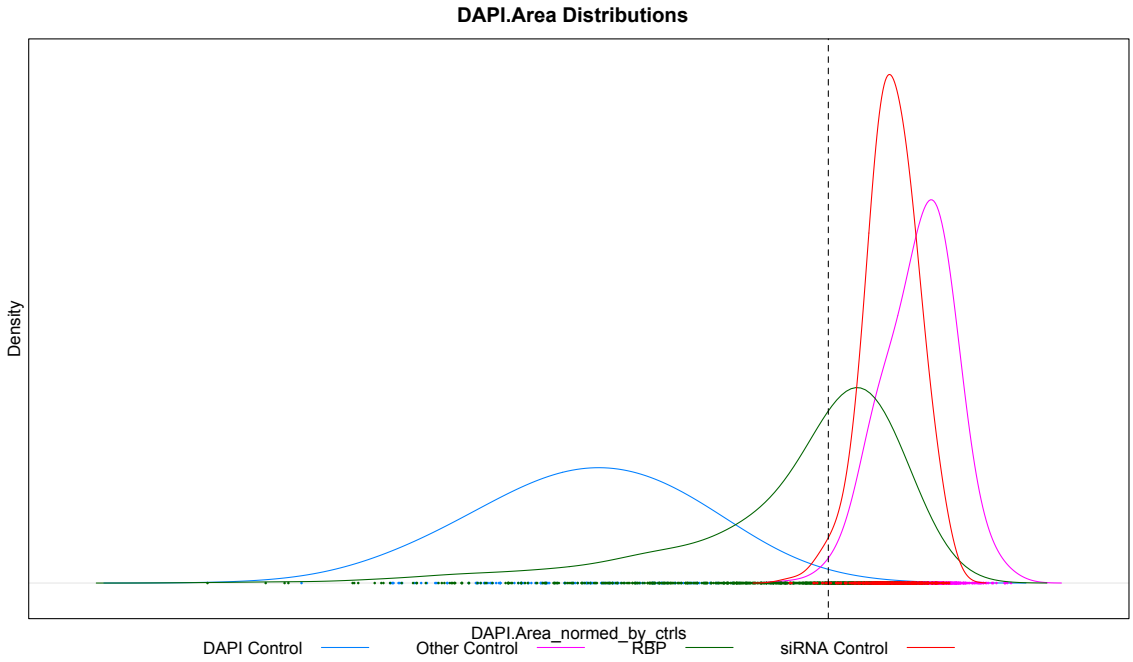


Figure 2.8. Density Plots of the log *DAPI area* normalized across all plates.
The siRNA controls are in red, the RBPs are in green, the *DAPI* controls are in blue, and the "other" controls are in purple. The dashed vertical line identifies the *DAPI* cut-off location, below which all observations were discarded. This approach discards ~26% of the data out of which 6% are DAPI control measurements..

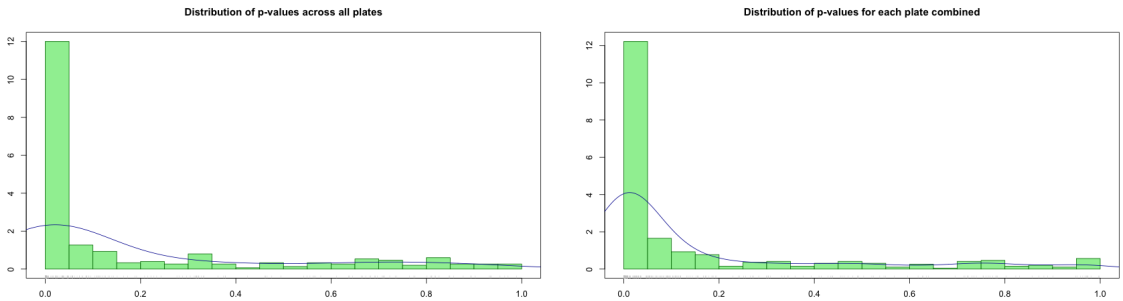


Figure 2.9. Distributions of p-values.
The left panel shows the p-value distribution for the per plate analysis. The right panels shows the distribution for the across all plate analysis.

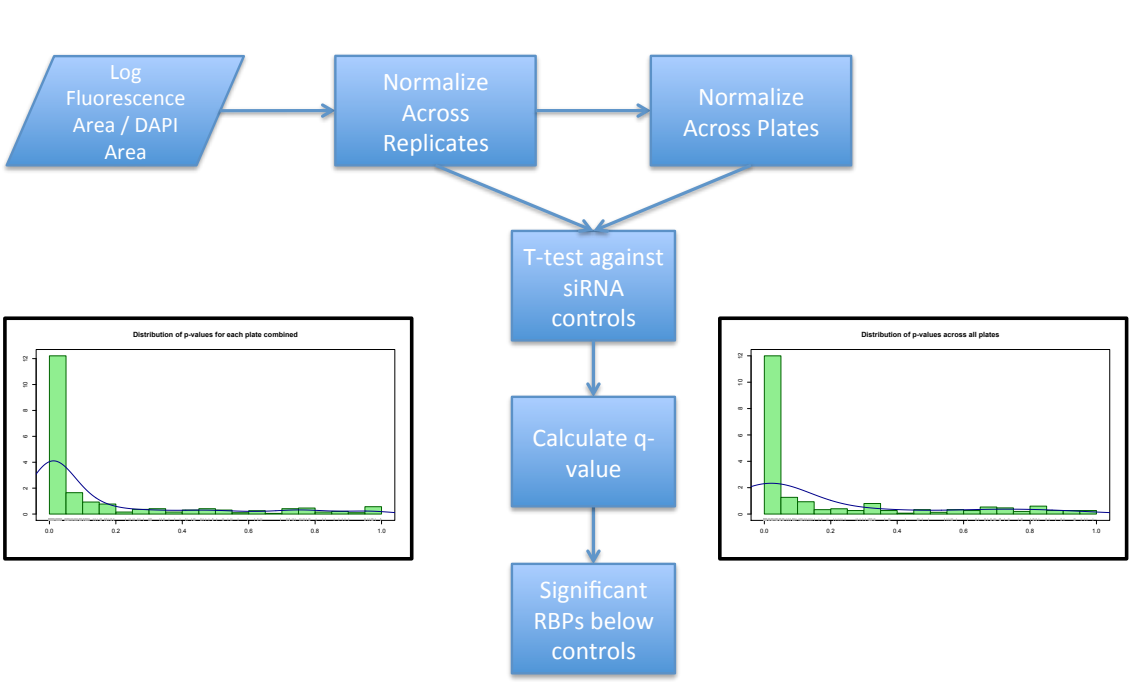


Figure 2.10. Graphical summary of approaches to identify significant RBPs.

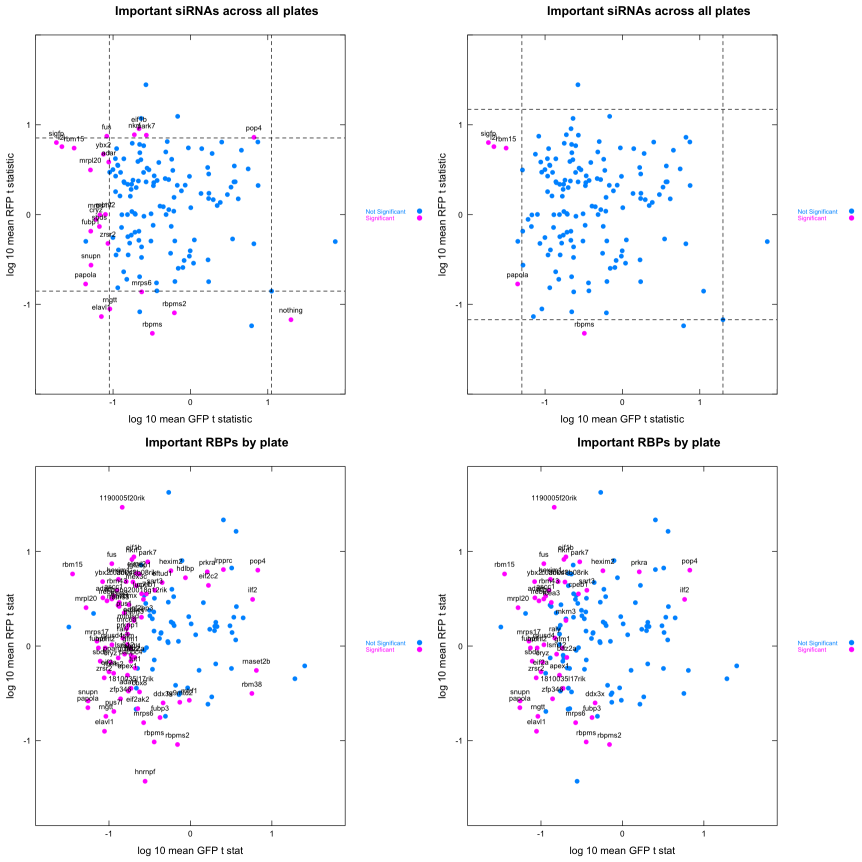


Figure 2.11. Important siRNAs identified.
The scatterplot of siRNAs with the axes on log 10 of the mean RBP divided by the mean control of normalized fluorescence values. The top left panel has the important siRNAs in the across all plates analysis using the "dfect" control as a q-value cutoff in addition to requiring the q-value to be below .01. The top right panel is the same except using the "nothing" control as a q-value cutoff instead of the "dfect". The bottom left panel has the important siRNAs in the by plate analysis using the "dfect" control as a q-value cutoff in addition to requiring the q-value to be below .001. The bottom right panel is the same except using the "nothing" control as a q-value cutoff instead of the "dfect".

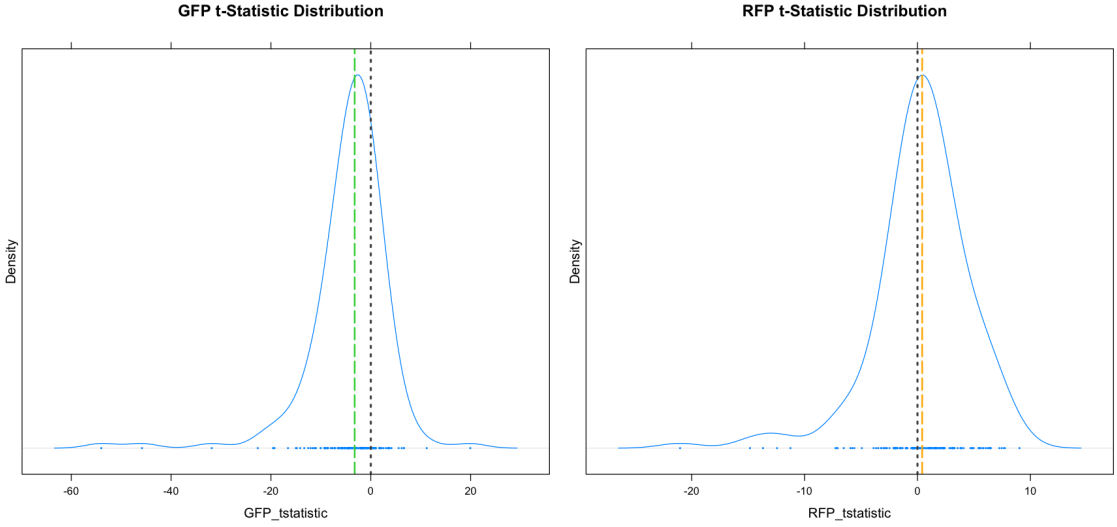


Figure 2.12. Distribution of t-statistics of the RFP and GFP measures.

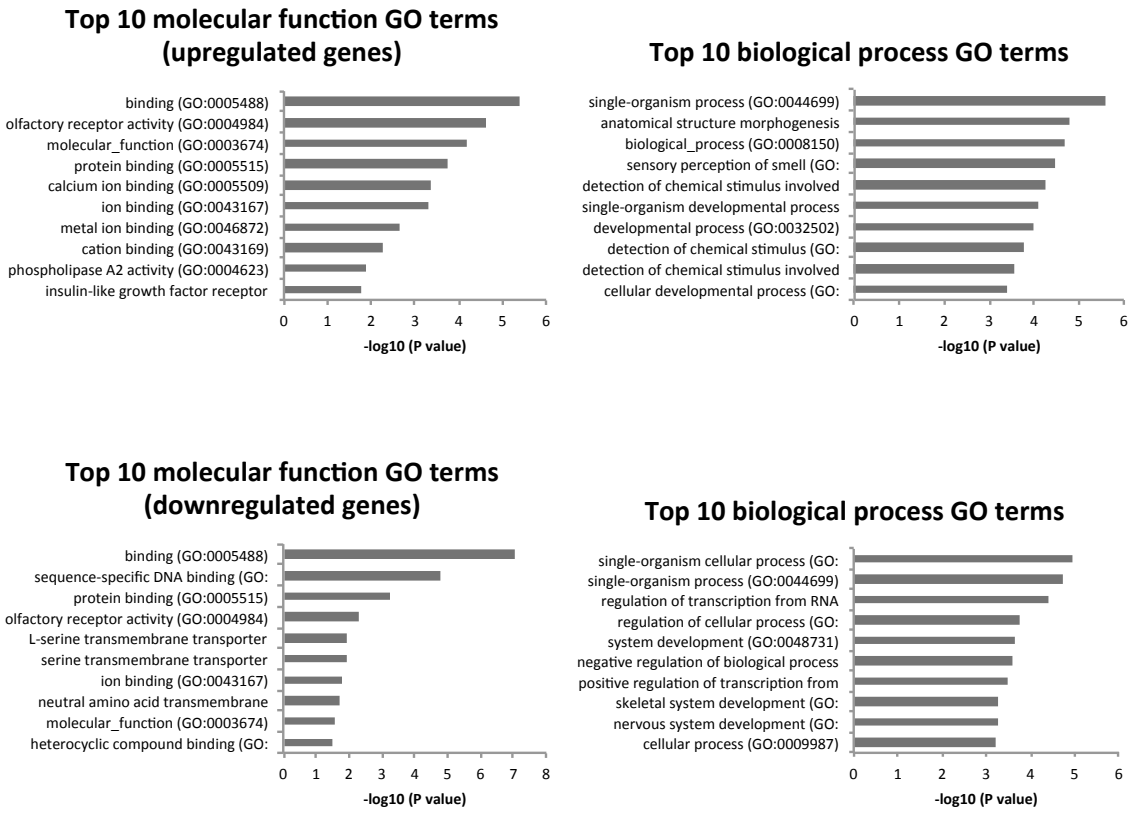


Figure 2.13. Gene set enrichment analysis for the differentially expressed genes.
The top left and top right barplots shows the p-values for the top enriched molecular function and biological processes gene ontology sets for upregulated genes. The bottom two barplots show the same information for downregulated genes.

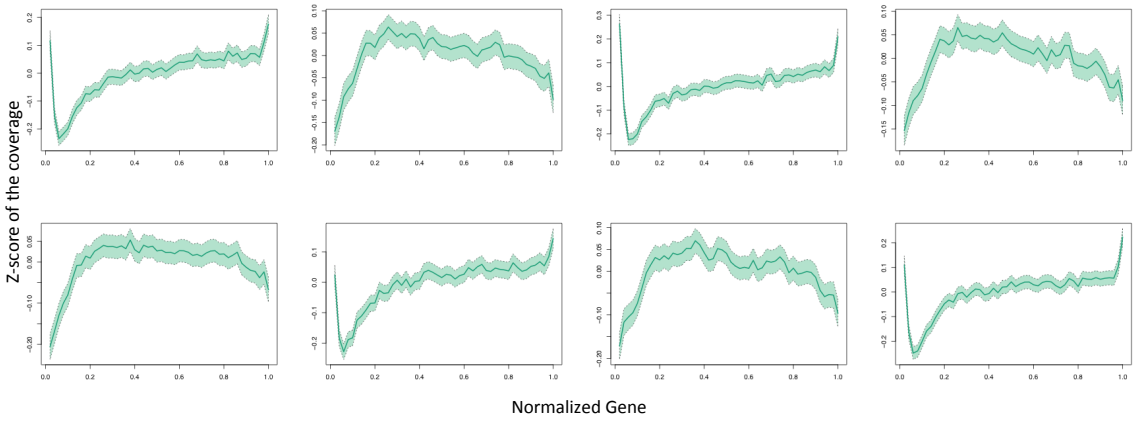


Figure 2.14. Average Z-score for each bin along the normalized gene.
The top 5,000 genes with the highest average coverage are displayed. Here the x-axis is the normalized gene location and the y-axis is the averaged Z-score per location. [Left 4 plots] The replicates for KO cells. [Right 4 plots] The replicates for WT cells. Within each set of 4, the topleft and bottomright are the profiles for the two ribosme protected samples. The topright and bottomleft are the profiles of the mRNA samples.

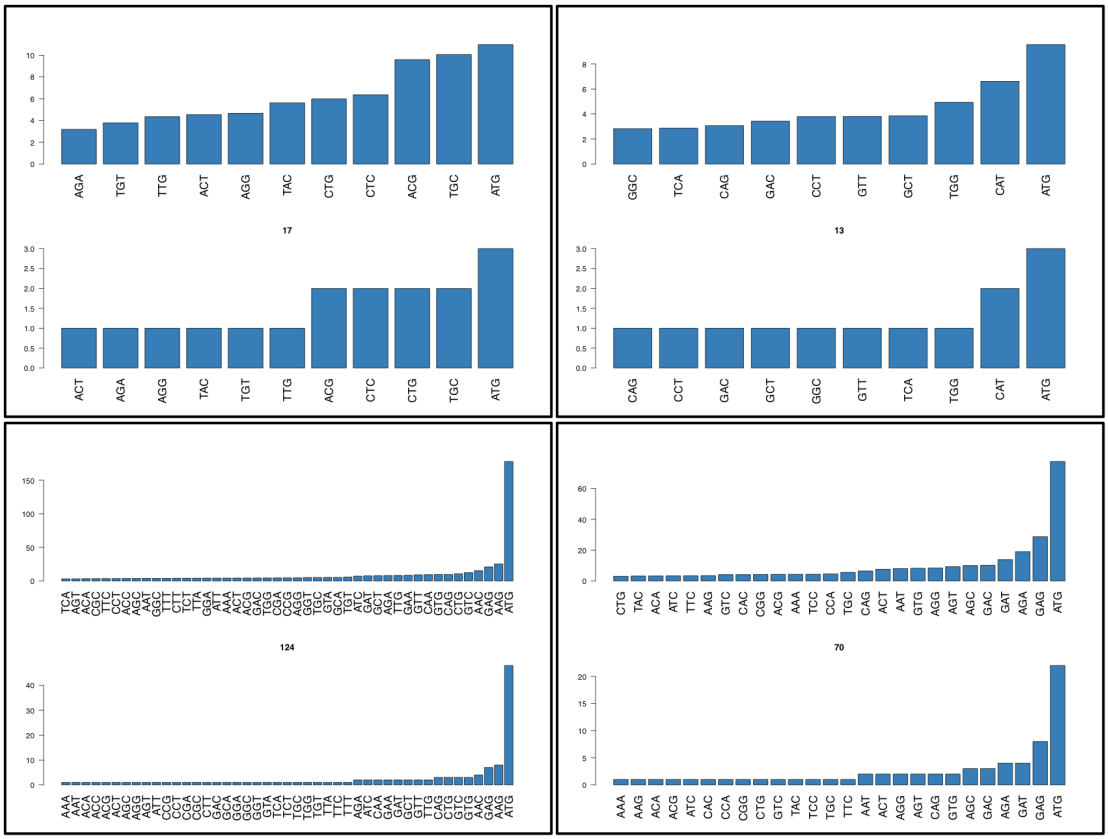


Figure 2.15. Distributions of codons in maximum enrichment locations (KO).
The max enrichment score is above 20. For each plot the top barplot shows the codons counts normalized by the occurrence of each codon in the transcriptome annotation; the bottom plot is the raw counts per codon; the number of genes above the cutoff is in the middle of each plot. [Top Row] The mRNA samples. [Bottom Row] The ribosome protected samples. This data is for the KO cells.

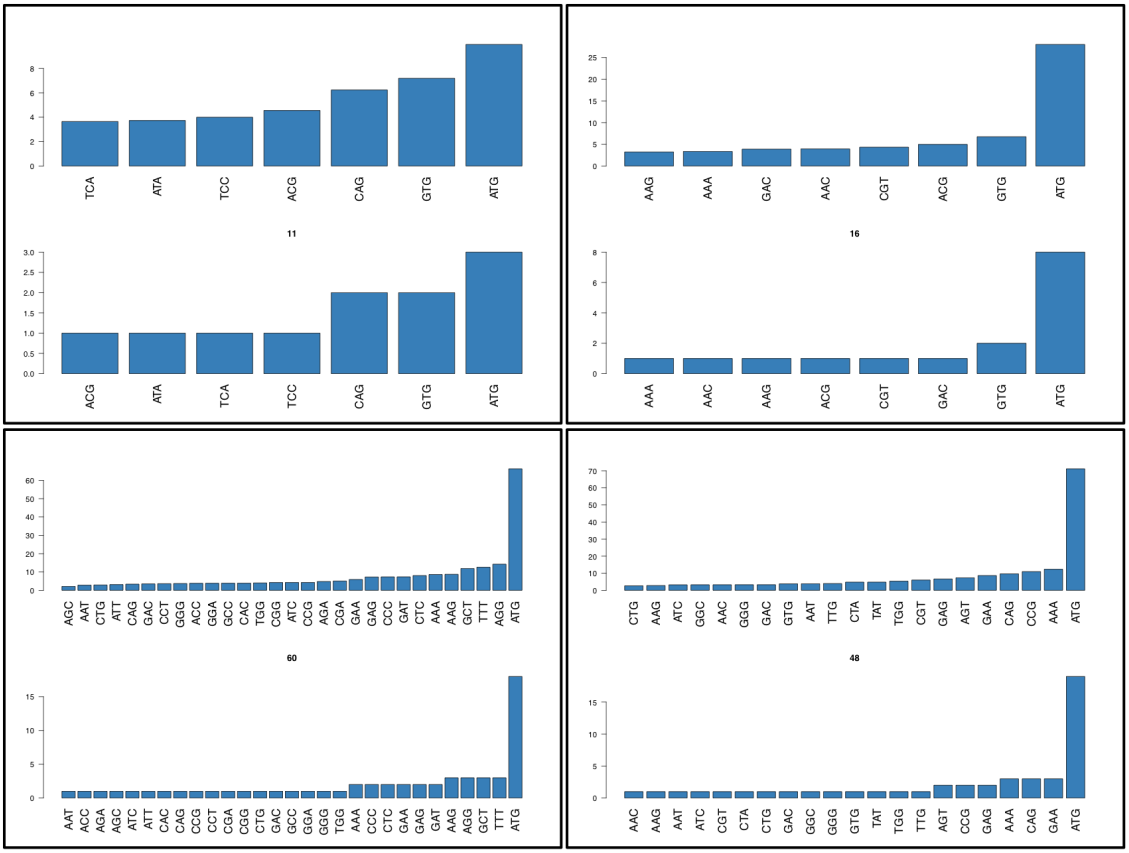


Figure 2.16. Distributions of codons in maximum enrichment locations (WT).
This is the same figure as above but for the WT cells.

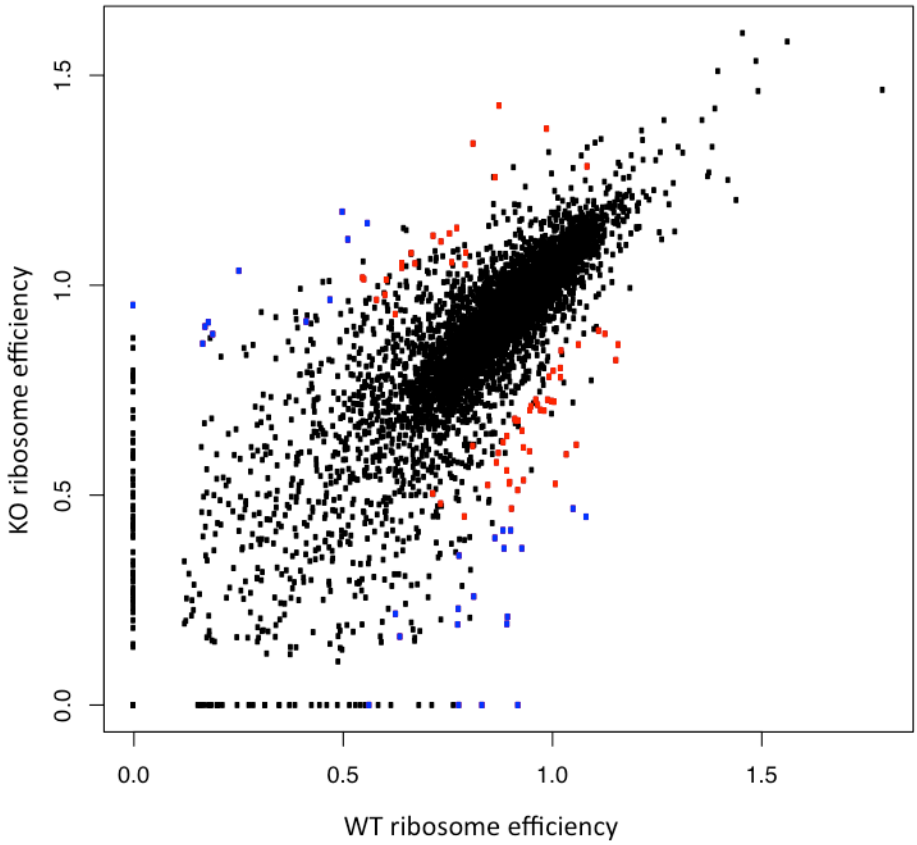


Figure 2.17. Scatterplot of the ribosome efficiency in the KO vs WT cells.
Each point is a gene; the red points are genes significant below a p-value of .05; the blue points are genes significant below a p-value of .001

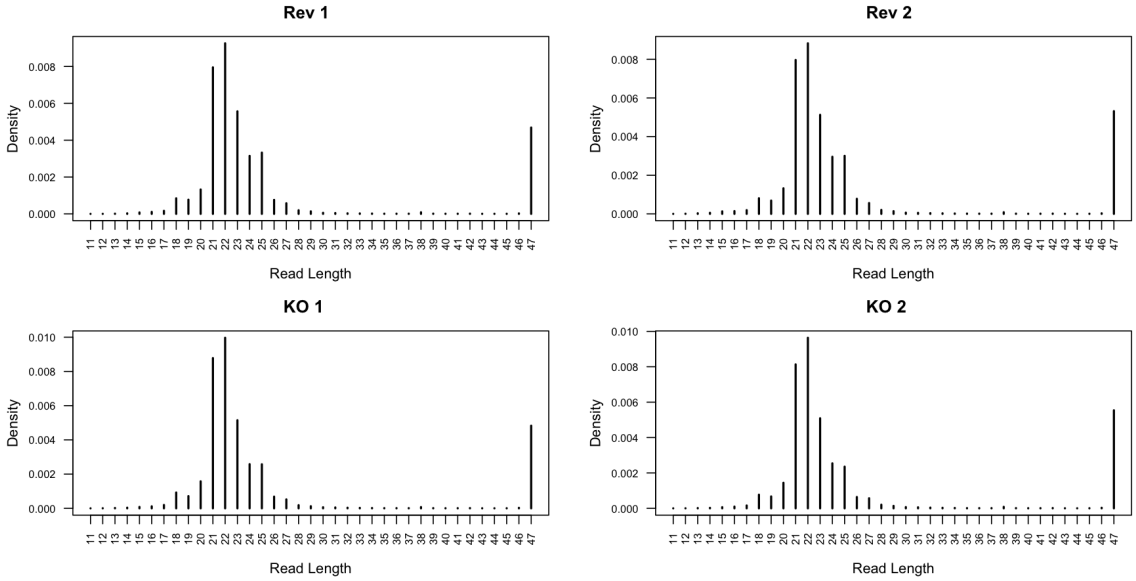


Figure 2.18. Distribution of read length for each replicate in the experiment.

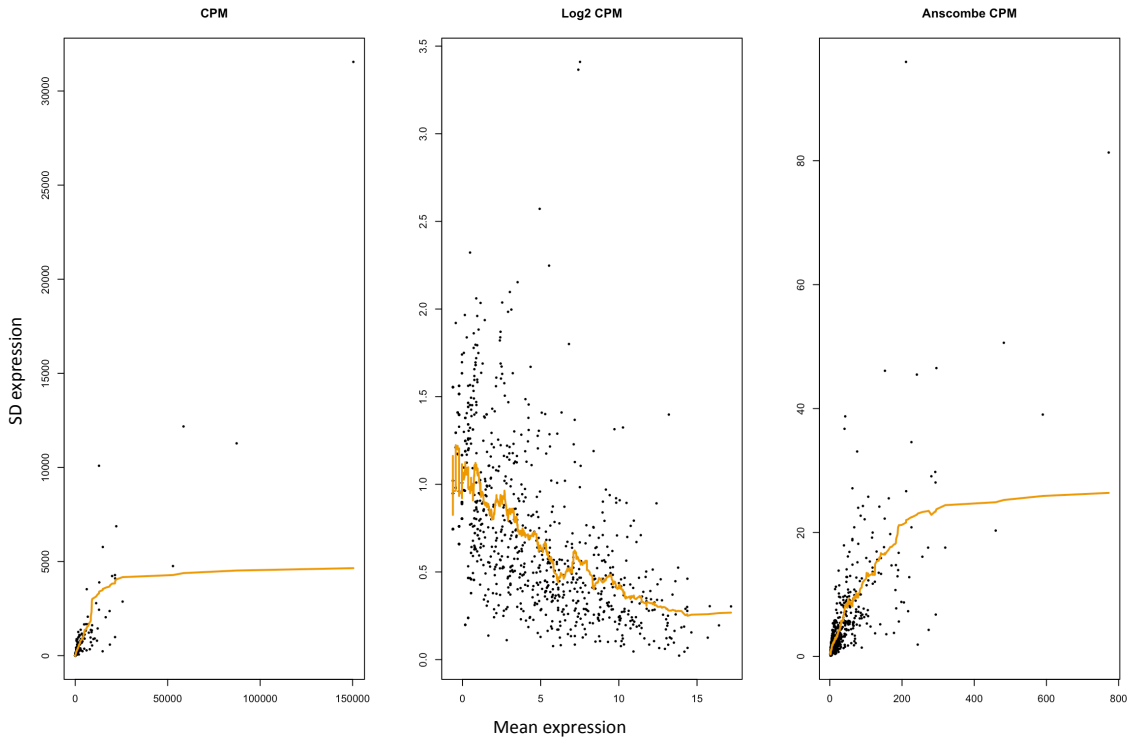


Figure 2.19. The mean-variance relationship when comparing different transforms.
Every point is a hairpin with a loess trend drawn to describe the relationship between the mean and variance.

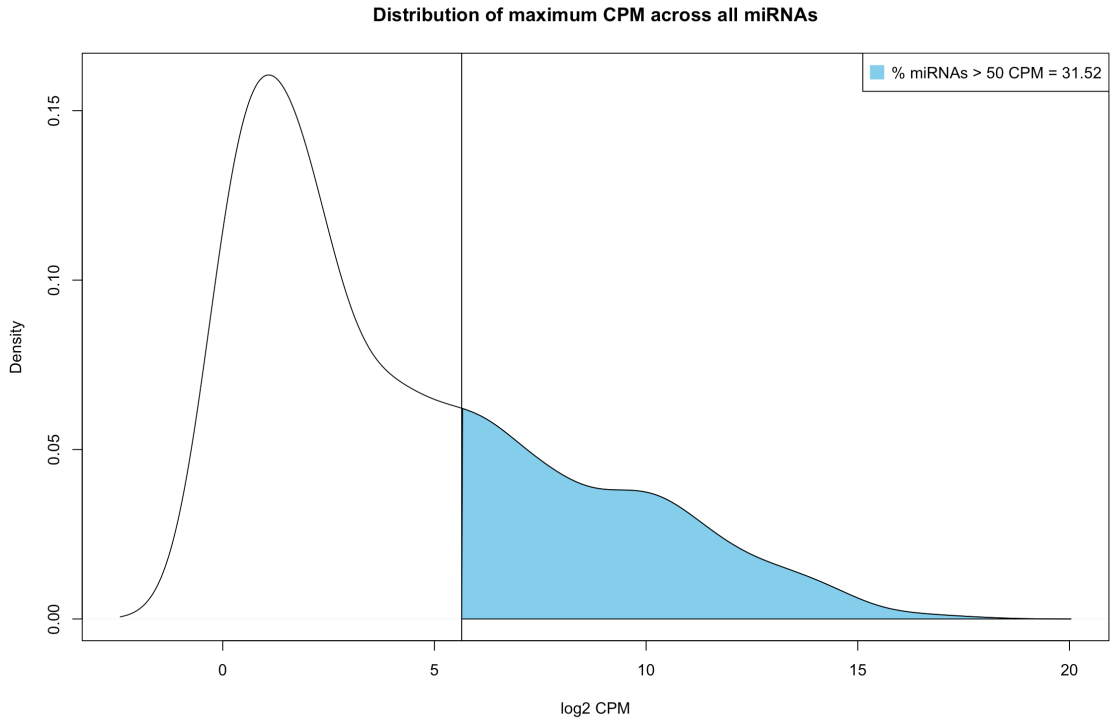


Figure 2.20. Distribution of CPM across all miRNAs and a cutoff at 50.

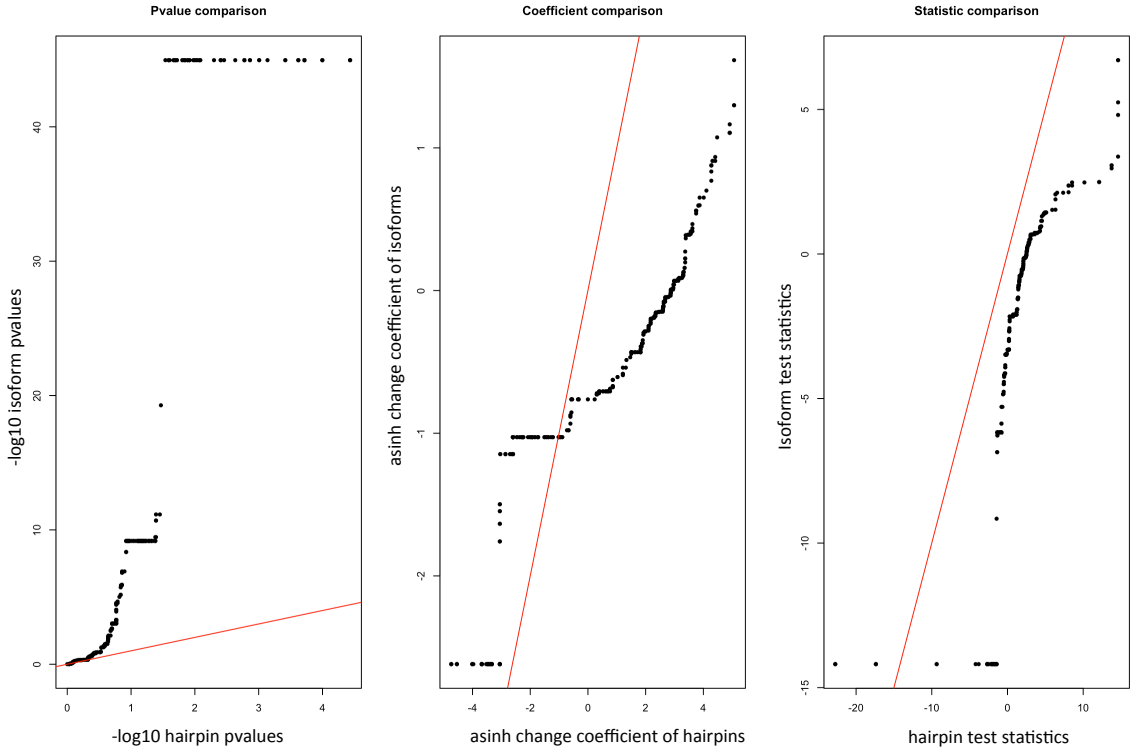


Figure 2.21. Qqplots comparing the distribution of pvalues (left), coefficients (middle), and test statistics (right) across hairpin and isoform groups.

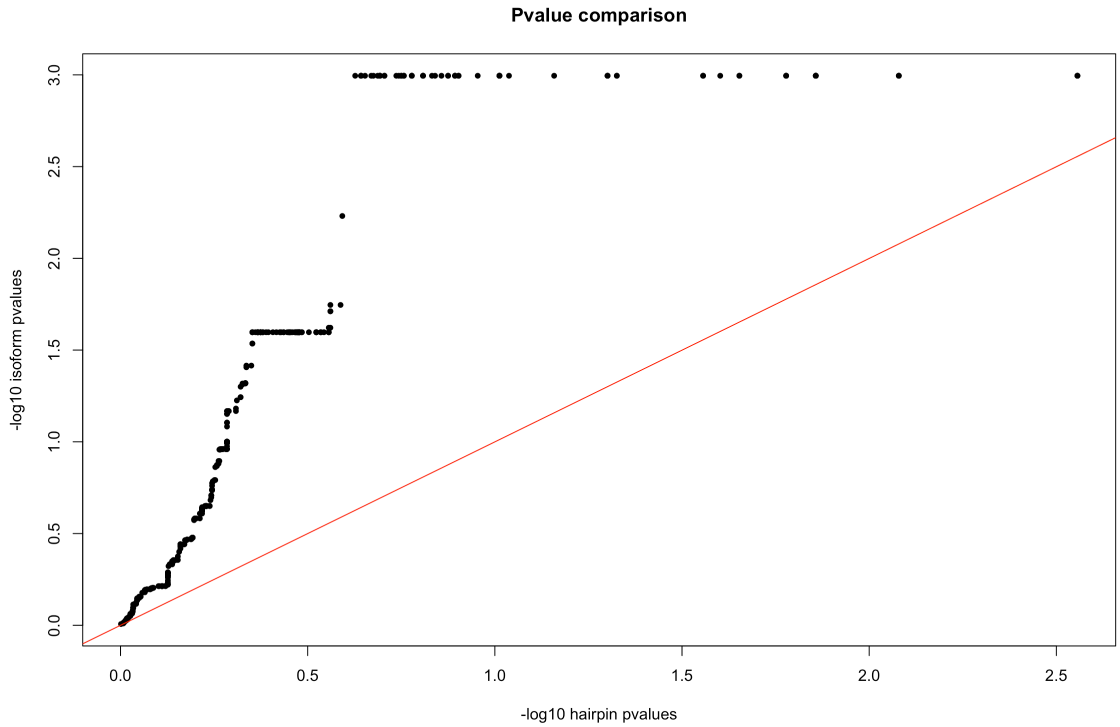


Figure 2.22. Qqplot of the normalized p-values.

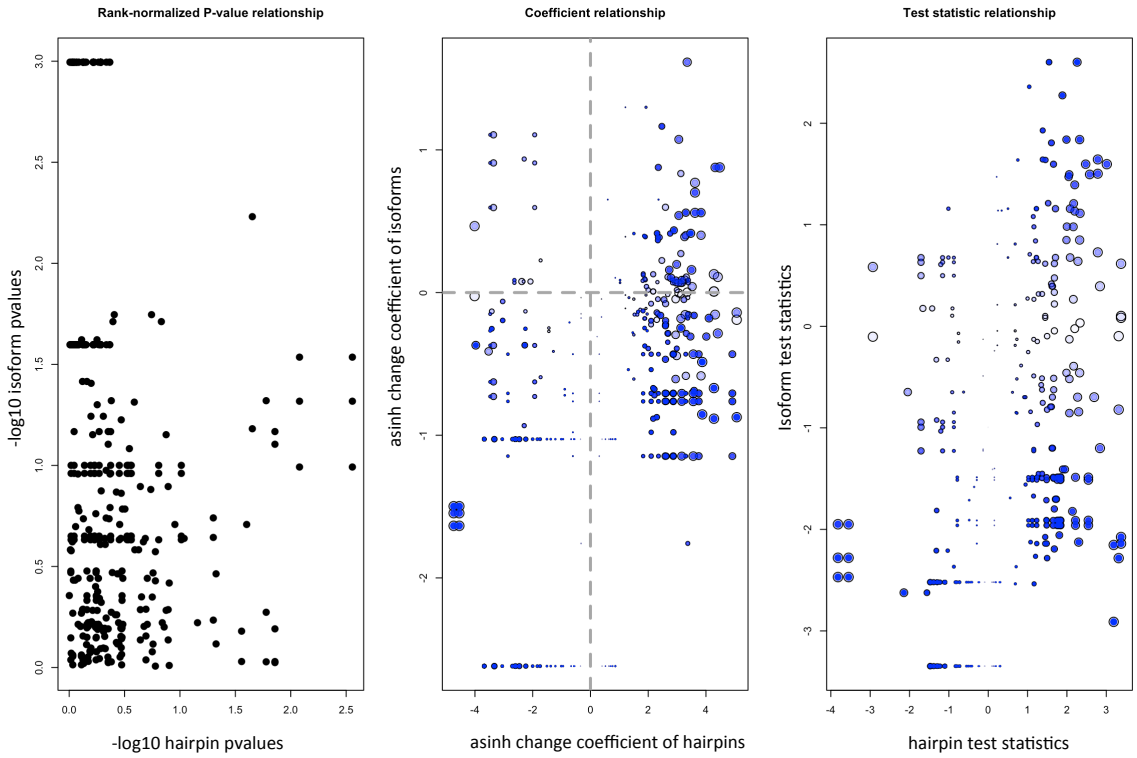


Figure 2.23. Comparing isoform and miRNA changes.
Every point is a hairpin matched with its isoform. The more significant a hairpin is, the larger the size of the point. Similarly the more significant the isoform is, the darker the blue of that point.

TABLES

genes	GFP qvalue	GFP tstatistic	RFP qvalue	RFP tstatistic	Significant RBPs from dfect	Significant RBPs from nothing
nothing	2.95E-69	1.99E+01	1.69E-42	-1.48E+01	3	0
l2i	1.65E-213	-4.58E+01	3.48E-08	5.70E+00	1	1
dfect	2.68E-26	1.12E+01	6.93E-12	-7.13E+00	0	0
sigfp	8.90E-253	-5.40E+01	1.02E-09	6.33E+00	1	1
cstf2	3.99E-01	3.36E-01	2.21E-02	2.27E+00	0	0
b020018g12rik	3.62E-06	-4.78E+00	1.44E-04	3.90E+00	0	0
1190005f20rik	2.03E-16	-8.66E+00	4.16E-01	-2.85E-01	0	0
ddx3x	6.21E-03	-2.75E+00	2.70E-08	-5.77E+00	0	0
1700021f05rik	5.38E-05	4.15E+00	9.57E-02	1.50E+00	0	0
d1pas1	1.31E-01	-1.31E+00	5.31E-02	1.84E+00	0	0
cpeb1	4.30E-01	2.29E-01	7.22E-05	4.08E+00	0	0
baz2a	4.43E-09	-6.10E+00	3.29E-02	-2.08E+00	0	0
1810035l17rik	1.95E-10	-6.65E+00	4.41E-01	1.90E-01	0	0
dis3l2	2.81E-01	-7.30E-01	6.61E-02	1.72E+00	0	0
dhx29	3.26E-01	-5.99E-01	4.13E-01	-2.99E-01	0	0
dclre1b	4.61E-02	-1.91E+00	3.84E-01	-4.08E-01	0	0
anks1	4.58E-01	-1.34E-01	6.80E-10	6.43E+00	0	0
dhx32	4.61E-02	-1.90E+00	2.32E-01	-8.82E-01	0	0
c030048b08rik	4.97E-08	-5.65E+00	2.31E-09	6.22E+00	0	0
apex1	8.88E-10	-6.38E+00	5.04E-03	-2.82E+00	0	0
cryz	1.26E-49	-1.66E+01	2.32E-01	8.83E-01	1	0
csad	2.32E-02	-2.24E+00	1.50E-01	1.22E+00	0	0
csda	3.23E-01	6.10E-01	1.84E-01	1.09E+00	0	0
adar	2.59E-26	-1.14E+01	1.77E-04	3.84E+00	1	0
ascc1	1.87E-18	-9.27E+00	3.89E-01	-3.73E-01	0	0
denr	2.96E-01	6.84E-01	4.73E-01	-8.07E-02	0	0
cbx8	3.06E-05	-4.29E+00	2.91E-03	-3.00E+00	0	0
adat1	3.41E-12	-7.29E+00	4.30E-01	2.30E-01	0	0
igf2bp1	9.04E-05	-4.02E+00	1.87E-08	5.84E+00	0	0
g3bp2	3.89E-01	3.71E-01	1.61E-02	2.39E+00	0	0
mars2	1.54E-01	-1.21E+00	1.42E-03	-3.25E+00	0	0
lrpprc	1.12E-02	2.54E+00	7.46E-10	6.41E+00	0	0
igf2bp2	6.61E-02	1.72E+00	6.74E-08	-5.59E+00	0	0
eefsec	1.16E-06	-5.02E+00	3.93E-02	2.00E+00	0	0
igf2bp3	2.54E-04	-3.74E+00	1.99E-02	2.31E+00	0	0
gfm1	6.09E-07	-5.15E+00	2.03E-01	-9.93E-01	0	0
eftud1	3.29E-02	-2.08E+00	1.43E-06	4.98E+00	0	0
mbnl2	3.93E-31	-1.25E+01	2.03E-01	1.00E+00	1	0
lsm12	6.24E-10	-6.45E+00	6.08E-02	-1.76E+00	0	0
ilf2	7.30E-08	5.57E+00	1.50E-03	3.22E+00	0	0

grb7	1.50E-03	-3.23E+00	2.06E-01	-9.80E-01	0	0
mettl1	4.84E-01	4.44E-02	3.97E-02	-1.99E+00	0	0
elavl1	3.46E-38	-1.41E+01	4.18E-36	-1.37E+01	3	0
hbp1	9.85E-02	-1.48E+00	3.67E-01	-4.73E-01	0	0
elavl2	1.03E-05	-4.54E+00	6.22E-09	6.04E+00	0	0
eif1b	7.36E-06	-4.62E+00	1.18E-17	9.03E+00	2	0
mex3a	3.50E-03	-2.94E+00	3.89E-01	-3.71E-01	0	0
hbs1l	4.58E-01	-1.33E-01	3.18E-02	2.11E+00	0	0
elavl3	6.93E-02	-1.69E+00	2.16E-03	3.10E+00	0	0
eif2a	3.28E-21	-1.00E+01	5.64E-02	-1.80E+00	0	0
mex3c	1.18E-05	-4.51E+00	2.26E-06	4.88E+00	0	0
ireb2	3.40E-21	-1.00E+01	1.81E-03	3.16E+00	0	0
hdlbp	1.90E-01	-1.06E+00	9.36E-08	5.52E+00	0	0
eif2ak2	1.03E-05	-4.54E+00	1.81E-06	-4.93E+00	0	0
jmjd6	7.44E-04	-3.44E+00	1.88E-03	3.15E+00	0	0
hexim1	1.48E-17	-9.00E+00	1.59E-07	5.42E+00	0	0
fubp1	9.19E-64	-1.94E+01	9.01E-02	-1.53E+00	1	0
luc71	5.11E-06	-4.70E+00	3.81E-01	-4.23E-01	0	0
hexim2	3.93E-02	-1.99E+00	4.68E-10	6.50E+00	0	0
fubp3	6.80E-03	-2.71E+00	1.29E-11	-7.07E+00	0	0
eif4b	8.26E-02	-1.59E+00	4.44E-01	1.80E-01	0	0
khdrbs3	3.31E-01	-5.77E-01	6.60E-02	1.72E+00	0	0
hnrnpf	2.35E-05	-4.35E+00	4.73E-01	-8.50E-02	0	0
fus	4.00E-29	-1.21E+01	1.16E-12	7.45E+00	3	0
mkrn3	1.87E-08	-5.84E+00	4.61E-02	1.91E+00	0	0
ict1	3.03E-04	3.69E+00	4.16E-01	-2.88E-01	0	0
eif2c2	4.61E-02	1.91E+00	3.53E-06	4.79E+00	0	0
rg9mtd2	2.59E-01	7.99E-01	1.49E-04	-3.89E+00	0	0
raver2	3.10E-05	-4.28E+00	3.78E-01	4.36E-01	0	0
rnaset2b	2.49E-10	6.61E+00	3.11E-02	-2.12E+00	0	0
rbm5	2.91E-03	3.00E+00	7.19E-02	1.67E+00	0	0
rbm10	2.07E-01	9.69E-01	3.85E-03	-2.91E+00	0	0
mthfsd	2.79E-08	-5.76E+00	3.70E-01	-4.60E-01	0	0
rngtt	6.93E-25	-1.10E+01	7.13E-26	-1.13E+01	2	0
rbm6	4.52E-01	1.55E-01	2.75E-02	2.17E+00	0	0
rmb11	4.08E-02	1.97E+00	1.01E-01	1.46E+00	0	0
pum2	4.60E-01	1.23E-01	7.99E-02	1.61E+00	0	0
mtrf1	2.25E-01	-9.12E-01	6.39E-04	-3.48E+00	0	0
mrpl11	4.18E-12	-7.26E+00	4.29E-01	-2.39E-01	0	0
rnmtl1	3.86E-07	-5.25E+00	1.99E-01	1.03E+00	0	0
rbm12	3.31E-01	-5.80E-01	4.86E-03	2.83E+00	0	0
pus1	8.14E-12	-7.15E+00	2.22E-02	2.26E+00	0	0
pop4	2.25E-10	6.62E+00	4.29E-12	7.25E+00	2	0

obfc2a	2.41E-04	-3.76E+00	4.86E-01	-3.59E-02	0	0
rbms1	3.84E-01	-4.11E-01	1.20E-01	1.36E+00	0	0
rbm15	1.38E-127	-3.19E+01	1.08E-07	5.49E+00	1	1
pus10	3.71E-04	-3.63E+00	1.25E-03	3.29E+00	0	0
ppargc1b	1.38E-13	-7.77E+00	1.14E-01	-1.39E+00	0	0
mrpl16	4.32E-01	-2.19E-01	4.41E-01	-1.92E-01	0	0
rbms2	5.63E-16	-8.52E+00	1.31E-02	-2.48E+00	0	0
pus3	1.10E-05	-4.52E+00	4.73E-01	8.26E-02	0	0
pabpc4	4.55E-07	-5.21E+00	8.43E-02	-1.57E+00	0	0
mrpl2	7.77E-02	-1.63E+00	8.73E-02	1.55E+00	0	0
rbms3	4.29E-12	-7.25E+00	1.96E-02	2.32E+00	0	0
pus7l	1.57E-16	-8.69E+00	3.81E-10	-6.53E+00	0	0
mrpl20	1.45E-64	-1.96E+01	1.96E-03	3.14E+00	1	0
rbmx	9.19E-08	-5.53E+00	1.54E-02	2.42E+00	0	0
rbm3	6.76E-06	-4.64E+00	4.41E-02	-1.94E+00	0	0
prkcsh	3.48E-04	3.65E+00	1.99E-02	2.31E+00	0	0
nkrf	2.84E-07	-5.31E+00	1.68E-13	7.74E+00	2	0
mrps17	3.11E-41	-1.48E+01	2.03E-01	9.90E-01	1	0
rbpms	2.16E-03	-3.11E+00	8.07E-72	-2.10E+01	2	2
rbm38	2.95E-09	6.17E+00	4.80E-01	5.77E-02	0	0
raly	2.23E-13	-7.70E+00	2.03E-01	-1.00E+00	0	0
prkra	2.85E-01	-7.14E-01	1.60E-07	5.42E+00	0	0
papola	4.55E-80	-2.26E+01	1.13E-08	-5.93E+00	1	1
mrps5	1.46E-03	-3.24E+00	4.08E-02	-1.97E+00	0	0
rbpms2	7.99E-02	-1.61E+00	1.04E-30	-1.24E+01	2	0
rbm4	5.87E-04	3.51E+00	4.95E-02	-1.87E+00	0	0
raly1	5.64E-02	-1.80E+00	7.05E-03	-2.70E+00	0	0
prkrip1	4.34E-10	-6.51E+00	2.03E-01	-1.00E+00	0	0
mrps6	3.33E-05	-4.26E+00	4.05E-12	-7.26E+00	2	0
rdbp	4.02E-25	-1.11E+01	3.29E-03	2.96E+00	0	0
rbm43	1.23E-16	-8.73E+00	4.89E-04	3.56E+00	0	0
park7	2.85E-04	-3.71E+00	2.95E-13	7.65E+00	2	0
rdm1	1.26E-18	-9.32E+00	4.29E-01	-2.39E-01	0	0
rbm45	2.31E-01	-8.90E-01	4.57E-02	-1.92E+00	0	0
msi2	2.15E-08	-5.81E+00	3.28E-01	5.90E-01	0	0
yipf1	2.13E-04	3.79E+00	3.79E-01	-4.32E-01	0	0
trim32	6.04E-02	-1.77E+00	2.80E-01	7.38E-01	0	0
trmt1	3.10E-02	-2.12E+00	3.87E-01	-3.97E-01	0	0
stau1	8.21E-02	-1.59E+00	3.54E-01	5.08E-01	0	0
rpusd4	1.77E-18	-9.28E+00	3.36E-01	-5.56E-01	0	0
zcchc12	3.31E-01	5.73E-01	4.29E-01	-2.42E-01	0	0
trmt2a	1.05E-01	-1.44E+00	4.28E-01	2.52E-01	0	0
trove2	2.46E-02	2.22E+00	1.42E-01	1.26E+00	0	0

strbp	2.03E-01	9.93E-01	3.87E-01	3.95E-01	0	0
smarcad1	1.62E-01	-1.17E+00	3.67E-01	-4.71E-01	0	0
tsfm	3.53E-08	-5.71E+00	1.86E-01	1.08E+00	0	0
thoc6	6.61E-02	1.72E+00	9.38E-02	1.51E+00	0	0
sart3	3.87E-01	3.90E-01	2.09E-04	3.80E+00	0	0
zfp346	4.62E-18	-9.15E+00	3.94E-01	3.55E-01	0	0
thumpd1	4.93E-01	-1.34E-02	3.92E-02	-2.00E+00	0	0
sbds	4.01E-42	-1.50E+01	2.80E-01	7.35E-01	1	0
tut1	6.87E-06	-4.63E+00	9.01E-02	-1.53E+00	0	0
thumpd3	2.07E-01	9.72E-01	3.89E-01	-3.76E-01	0	0
scaf1	1.02E-03	3.35E+00	3.78E-02	2.02E+00	0	0
tnrc6a	8.60E-09	-5.98E+00	8.13E-02	1.60E+00	0	0
zrsr2	1.94E-27	-1.16E+01	3.21E-02	-2.10E+00	1	0
upf1	8.57E-05	-4.03E+00	3.48E-03	2.94E+00	0	0
tcea3	2.73E-14	-8.00E+00	3.99E-01	-3.39E-01	0	0
xpo5	2.03E-01	1.01E+00	1.57E-01	1.19E+00	0	0
traf6	1.01E-04	-3.99E+00	1.94E-01	-1.04E+00	0	0
tceal5	1.32E-02	-2.47E+00	1.46E-01	1.24E+00	0	0
snupn	4.37E-63	-1.93E+01	3.39E-04	-3.66E+00	1	0
trdmt1	2.91E-03	-3.01E+00	3.21E-02	2.10E+00	0	0
tdrd3	6.93E-02	1.69E+00	1.64E-03	-3.20E+00	0	0
ybx2	1.52E-34	-1.33E+01	5.16E-06	4.70E+00	1	0
trim3	3.41E-02	-2.06E+00	2.03E-01	9.99E-01	0	0
tdrd7	2.70E-03	-3.03E+00	1.30E-01	-1.32E+00	0	0

Table 2.1. Results for RBP testing combined across all plates.
The RBPs more significant than the dfect are highlighted in Green and the ones more significant than nothing are highlighted in Yellow; the last two columns of each sheet are the significance cutoff columns for the dfect and nothing controls, respectively, with 0 being not significant, 1 being significant for GFP, 2 being significant for RFP, and 3 being significant for both.

Plates	Gene	GFP qvalue	GFP tstatistic	RFP qvalue	RFP tstatistic	Significant RBPs from dfect	Significant RBPs from nothing
Plate 4a	1190005f20rik	2.85E-08	-6.88E+00	4.70E-01	-3.40E-02	1	1
Plate 4a	1700021f05rik	2.29E-03	3.24E+00	1.08E-01	1.39E+00	0	0
Plate 4a	1810035l17rik	4.13E-06	-5.35E+00	3.77E-01	3.57E-01	1	1
Plate 4a	b020018g12rik	5.17E-04	-3.78E+00	3.12E-03	3.12E+00	1	0
Plate 4a	baz2a	2.27E-05	-4.82E+00	1.19E-01	-1.32E+00	1	1
Plate 4a	cpeb1	3.94E-01	2.89E-01	9.40E-04	3.56E+00	2	2
Plate 4a	cstf2	3.82E-01	3.41E-01	4.40E-02	1.94E+00	0	0
Plate 4a	d1pas1	1.95E-01	-9.80E-01	6.48E-02	1.72E+00	0	0
Plate 4a	ddx3x	2.83E-02	-	2.76E-	-4.00E+00	2	2

			2.17E+00	04			
Plate 4a	dfect	6.74E-02	1.68E+00	2.47E-01	-7.81E-01	0	0
Plate 4a	dhx29	3.68E-01	-3.93E-01	4.70E-01	-4.63E-02	0	0
Plate 4a	dis3l2	3.32E-01	-5.02E-01	7.47E-02	1.62E+00	0	0
Plate 4a	l2i	4.88E-15	-1.10E+01	2.17E-02	2.29E+00	1	1
Plate 4a	nothing	4.50E-04	3.78E+00	8.04E-03	-2.71E+00	1	0
Plate 4a	sigfp	2.69E-19	-1.40E+01	7.08E-02	1.65E+00	1	1
Plate 4b	anks1	4.34E-01	1.63E-01	2.31E-05	4.81E+00	2	2
Plate 4b	apex1	6.61E-07	-5.93E+00	3.76E-02	-2.03E+00	1	1
Plate 4b	c030048b08rik	1.01E-05	-5.08E+00	2.67E-05	4.76E+00	3	3
Plate 4b	cryz	3.76E-12	-9.56E+00	2.77E-01	6.74E-01	1	1
Plate 4b	dclre1b	7.61E-02	-1.61E+00	3.92E-01	-3.00E-01	0	0
Plate 4b	dfect	7.43E-03	2.74E+00	1.47E-01	-1.19E+00	0	0
Plate 4b	dhx32	7.84E-02	-1.59E+00	2.97E-01	-6.08E-01	0	0
Plate 4b	l2i	2.86E-13	-9.82E+00	2.57E-02	2.21E+00	1	1
Plate 4b	nothing	4.23E-05	4.54E+00	2.28E-03	-3.22E+00	1	0
Plate 4b	sigfp	1.93E-14	-1.06E+01	4.64E-03	2.94E+00	1	1
Plate 4c	adar	3.02E-15	-1.19E+01	2.32E-03	3.24E+00	1	1
Plate 4c	adat1	1.05E-06	-5.77E+00	3.83E-01	3.34E-01	1	0
Plate 4c	ascc1	5.09E-11	-8.75E+00	3.93E-01	-2.96E-01	1	1
Plate 4c	cbx8	1.38E-04	-4.22E+00	3.65E-03	-3.06E+00	1	0
Plate 4c	csad	2.17E-02	-2.31E+00	1.12E-01	1.36E+00	0	0
Plate 4c	csda	3.21E-01	5.33E-01	1.69E-01	1.09E+00	0	0
Plate 4c	denr	2.87E-01	6.41E-01	4.78E-01	3.90E-03	0	0
Plate 4c	dfect	5.56E-03	2.87E+00	2.95E-02	-2.14E+00	0	0
Plate 4c	l2i	7.48E-19	-1.36E+01	2.38E-02	2.25E+00	1	1
Plate 4c	nothing	5.77E-07	5.83E+00	3.38E-05	-4.61E+00	3	0
Plate 4c	sigfp	2.69E-23	-1.72E+01	1.15E-01	1.34E+00	1	1
Plate 5a	dfect	7.06E-02	1.66E+00	7.01E-05	-4.38E+00	0	0
Plate 5a	eefsec	1.53E-05	-	4.40E-	1.94E+00	1	0

Plate 5a	eftud1	2.56E-02	-4.94E+00 -2.22E+00	02 3.23E-05	4.70E+00	2	0
Plate 5a	g3bp2	3.89E-01	3.14E-01	1.81E-02	2.39E+00	0	0
Plate 5a	gfm1	2.89E-06	-5.47E+00	1.78E-01	-1.05E+00	1	1
Plate 5a	grb7	1.26E-03	-3.46E+00	1.79E-01	-1.04E+00	0	0
Plate 5a	igf2bp1	1.32E-04	-4.23E+00	1.05E-06	5.78E+00	3	0
Plate 5a	igf2bp2	7.29E-02	1.64E+00	1.86E-04	-4.12E+00	0	0
Plate 5a	igf2bp3	2.80E-04	-3.99E+00	3.86E-02	2.02E+00	1	0
Plate 5a	ilf2	1.01E-06	5.80E+00	3.13E-03	3.12E+00	1	1
Plate 5a	l2i	1.34E-21	-1.57E+01	1.20E-04	-4.21E+00	1	1
Plate 5a	lrpprc	1.26E-02	2.54E+00	1.21E-07	6.44E+00	2	0
Plate 5a	lsm12	1.24E-07	-6.43E+00	1.40E-01	-1.22E+00	1	1
Plate 5a	mars2	1.15E-01	-1.34E+00	3.04E-03	-3.13E+00	0	0
Plate 5a	mbnl2	2.70E-14	-1.11E+01	2.00E-01	9.54E-01	1	1
Plate 5a	nothing	8.47E-06	5.04E+00	2.73E-14	-1.05E+01	3	0
Plate 5a	sigfp	1.31E-20	-1.49E+01	1.61E-02	2.43E+00	1	1
Plate 5b	dfect	1.91E-02	2.35E+00	6.49E-03	-2.81E+00	0	0
Plate 5b	eif1b	1.53E-05	-4.94E+00	4.81E-11	8.78E+00	3	2
Plate 5b	eif2a	8.73E-13	-1.00E+01	5.02E-02	-1.87E+00	1	1
Plate 5b	eif2ak2	7.21E-05	-4.43E+00	4.38E-05	-4.59E+00	3	0
Plate 5b	elavl1	1.42E-14	-1.14E+01	7.11E-10	-7.97E+00	3	3
Plate 5b	elavl2	4.49E-05	-4.58E+00	8.02E-07	5.87E+00	3	0
Plate 5b	elavl3	5.02E-02	-1.87E+00	5.30E-03	2.91E+00	0	0
Plate 5b	hbp1	7.29E-02	-1.64E+00	3.07E-01	-5.75E-01	0	0
Plate 5b	hbs1l	4.13E-01	-2.25E-01	4.09E-02	1.99E+00	0	0
Plate 5b	hdlbp	1.55E-01	-1.15E+00	5.22E-06	5.28E+00	2	0
Plate 5b	ireb2	1.39E-13	-1.06E+01	4.14E-03	3.01E+00	1	1
Plate 5b	l2i	2.68E-20	-1.47E+01	4.18E-02	-1.96E+00	1	1
Plate 5b	mettl1	4.70E-01	-3.95E-02	7.47E-02	-1.62E+00	0	0
Plate 5b	mex3a	2.94E-03	-	3.38E-	-4.78E-01	0	0

			3.14E+00	01			
Plate 5b	mex3c	4.23E-05	-4.61E+00	9.10E-05	4.35E+00	3	0
Plate 5b	nothing	8.16E-08	6.39E+00	2.63E-08	-6.70E+00	3	0
Plate 5b	sigfp	1.74E-22	-1.65E+01	5.94E-02	1.77E+00	1	1
Plate 5c	dfect	1.83E-02	2.37E+00	2.05E-02	-2.32E+00	0	0
Plate 5c	eif2c2	7.06E-02	1.66E+00	8.32E-05	4.38E+00	2	0
Plate 5c	eif4b	1.09E-01	-1.38E+00	3.90E-01	3.07E-01	0	0
Plate 5c	fubp1	2.79E-17	-1.35E+01	1.78E-01	-1.05E+00	1	1
Plate 5c	fubp3	1.91E-02	-2.36E+00	1.32E-06	-5.71E+00	2	2
Plate 5c	fus	1.13E-11	-9.23E+00	4.19E-09	7.43E+00	3	3
Plate 5c	hexim1	2.06E-09	-7.65E+00	1.01E-05	5.08E+00	3	1
Plate 5c	hexim2	6.36E-02	-1.74E+00	2.21E-07	6.26E+00	2	2
Plate 5c	hnrnpf	8.41E-04	-3.60E+00	4.70E-01	3.71E-02	1	0
Plate 5c	ict1	2.48E-03	3.21E+00	4.38E-01	-1.49E-01	0	0
Plate 5c	jmjd6	7.37E-03	-2.77E+00	2.54E-03	3.20E+00	0	0
Plate 5c	khdrbs3	3.35E-01	-4.88E-01	5.84E-02	1.78E+00	0	0
Plate 5c	l2i	1.35E-27	-2.12E+01	7.29E-02	-1.63E+00	1	1
Plate 5c	luc71	2.88E-04	-3.98E+00	3.96E-01	-2.80E-01	1	0
Plate 5c	mkrn3	1.53E-05	-4.94E+00	5.06E-02	1.86E+00	1	1
Plate 5c	nothing	7.21E-05	4.36E+00	1.32E-06	-5.58E+00	3	0
Plate 5c	sigfp	9.49E-27	-2.03E+01	2.82E-02	2.16E+00	1	1
Plate 6a	dfect	4.74E-06	5.21E+00	4.18E-02	-1.96E+00	0	0
Plate 6a	l2i	2.46E-27	-2.09E+01	4.07E-02	1.98E+00	1	1
Plate 6a	mrpl11	1.53E-09	-7.74E+00	3.77E-01	-3.69E-01	1	0
Plate 6a	mthfsd	1.02E-06	-5.79E+00	2.98E-01	-6.03E-01	1	0
Plate 6a	mtrf1	1.81E-01	-1.03E+00	5.61E-04	-3.75E+00	2	0
Plate 6a	nothing	1.35E-11	8.75E+00	1.02E-05	-4.98E+00	3	0
Plate 6a	obfc2a	2.55E-04	-4.02E+00	4.41E-01	-1.39E-01	0	0
Plate 6a	pop4	4.80E-08	6.72E+00	1.64E-07	6.35E+00	3	2
Plate 6a	pum2	4.70E-01	3.22E-02	7.83E-	1.59E+00	0	0

				02			
Plate 6a	pus1	9.78E-08	-6.51E+00	2.45E-02	2.24E+00	1	0
Plate 6a	raver2	4.24E-05	-4.61E+00	3.77E-01	3.58E-01	0	0
Plate 6a	rbm10	2.08E-01	9.24E-01	7.34E-03	-2.77E+00	0	0
Plate 6a	rbm12	2.72E-01	-6.93E-01	5.73E-03	2.88E+00	0	0
Plate 6a	rbm5	3.65E-03	3.06E+00	7.29E-02	1.64E+00	0	0
Plate 6a	rbm6	4.68E-01	6.46E-02	2.58E-02	2.21E+00	0	0
Plate 6a	rg9mtd2	2.59E-01	7.40E-01	3.28E-04	-3.93E+00	2	0
Plate 6a	rmb11	4.17E-02	1.98E+00	9.82E-02	1.45E+00	0	0
Plate 6a	rnaset2b	1.18E-07	6.45E+00	5.58E-02	-1.81E+00	1	0
Plate 6a	rngtt	4.91E-14	-1.09E+01	2.25E-06	-5.54E+00	3	3
Plate 6a	rnmtl1	4.19E-06	-5.35E+00	2.44E-01	7.99E-01	1	0
Plate 6a	sigfp	2.46E-27	-2.09E+01	3.92E-03	3.01E+00	1	1
Plate 6b	dfect	1.78E-05	4.81E+00	7.84E-02	-1.58E+00	0	0
Plate 6b	l2i	4.94E-51	-5.86E+01	9.33E-02	1.48E+00	1	1
Plate 6b	mrpl16	3.96E-01	-2.75E-01	4.68E-01	-6.12E-02	0	0
Plate 6b	mrpl2	6.40E-02	-1.72E+00	5.71E-02	1.80E+00	0	0
Plate 6b	mrpl20	2.79E-23	-1.91E+01	1.25E-02	2.55E+00	1	1
Plate 6b	mrps17	5.83E-18	-1.41E+01	1.60E-01	1.13E+00	1	1
Plate 6b	nkrf	6.10E-06	-5.23E+00	2.77E-10	8.25E+00	3	2
Plate 6b	nothing	4.35E-11	8.44E+00	5.54E-07	-5.85E+00	3	0
Plate 6b	pabpc4	3.63E-06	-5.39E+00	9.91E-02	-1.45E+00	1	0
Plate 6b	ppargc1b	1.72E-09	-7.70E+00	1.16E-01	-1.34E+00	1	0
Plate 6b	prkcsh	6.95E-04	3.67E+00	1.07E-02	2.61E+00	0	0
Plate 6b	pus10	6.57E-04	-3.69E+00	1.35E-03	3.43E+00	0	0
Plate 6b	pus3	3.78E-05	-4.65E+00	4.16E-01	2.15E-01	0	0
Plate 6b	pus7l	5.66E-11	-8.71E+00	1.60E-05	-4.92E+00	3	0
Plate 6b	rbm15	3.82E-30	-2.80E+01	1.02E-06	5.79E+00	3	1
Plate 6b	rbm3	6.73E-05	-4.46E+00	6.40E-02	-1.73E+00	0	0
Plate 6b	rbms1	3.39E-01	-4.72E-	7.61E-	1.61E+00	0	0

			01	02			
Plate 6b	rbms2	4.87E-11	-8.77E+00	4.40E-02	-1.94E+00	1	0
Plate 6b	rbms3	3.47E-09	-7.50E+00	1.04E-02	2.63E+00	1	0
Plate 6b	rbmx	1.38E-06	-5.68E+00	7.92E-03	2.74E+00	1	0
Plate 6b	sigfp	5.43E-42	-4.01E+01	4.33E-03	2.97E+00	1	1
Plate 6c	dfect	6.54E-03	2.80E+00	2.23E-02	-2.27E+00	0	0
Plate 6c	l2i	2.30E-37	-3.28E+01	6.62E-02	1.70E+00	1	1
Plate 6c	mrps5	5.12E-03	-2.92E+00	6.06E-02	-1.76E+00	0	0
Plate 6c	mrps6	5.48E-04	-3.76E+00	1.14E-07	-6.47E+00	3	2
Plate 6c	msi2	6.00E-06	-5.24E+00	2.56E-01	7.53E-01	1	0
Plate 6c	nothing	4.47E-08	6.55E+00	1.32E-06	-5.58E+00	3	0
Plate 6c	papola	2.28E-22	-1.82E+01	6.10E-05	-4.49E+00	3	1
Plate 6c	park7	1.74E-03	-3.34E+00	1.35E-09	7.78E+00	2	2
Plate 6c	prkra	2.93E-01	-6.21E-01	3.90E-07	6.10E+00	2	2
Plate 6c	prkrip1	6.83E-07	-5.92E+00	2.58E-01	-7.47E-01	1	0
Plate 6c	raly	4.26E-08	-6.76E+00	2.37E-01	-8.23E-01	1	1
Plate 6c	raly1	8.77E-02	-1.52E+00	1.10E-02	-2.60E+00	0	0
Plate 6c	rbm38	1.38E-06	5.68E+00	3.89E-01	3.15E-01	1	0
Plate 6c	rbm4	2.60E-03	3.19E+00	6.40E-02	-1.73E+00	0	0
Plate 6c	rbm43	7.32E-10	-7.96E+00	3.68E-04	3.89E+00	3	1
Plate 6c	rbm45	2.47E-01	-7.83E-01	5.61E-02	-1.81E+00	0	0
Plate 6c	rbpms	7.11E-03	-2.79E+00	2.96E-13	-1.03E+01	2	2
Plate 6c	rbpms2	9.96E-02	-1.44E+00	3.78E-14	-1.10E+01	2	2
Plate 6c	rdbp	1.13E-11	-9.23E+00	2.92E-03	3.15E+00	1	1
Plate 6c	rdm1	5.13E-10	-8.07E+00	4.78E-01	3.12E-03	1	1
Plate 6c	sigfp	6.87E-24	-1.77E+01	2.63E-01	7.26E-01	1	1
Plate 7a	dfect	6.33E-04	3.60E+00	4.54E-02	-1.90E+00	0	0
Plate 7a	l2i	9.29E-21	-1.51E+01	1.11E-05	4.95E+00	3	3
Plate 7a	nothing	4.49E-05	4.51E+00	9.96E-02	-1.44E+00	1	0
Plate 7a	rpusd4	1.21E-11	-	1.97E-	-9.69E-01	1	1

Plate 7a	sigfp	1.03E-23	9.20E+00 - 1.75E+01	01 2.67E- 01	7.11E-01	1	1
Plate 7a	smarcad1	1.06E-01	- 1.40E+00	2.18E- 01	-8.92E-01	0	0
Plate 7a	stau1	5.21E-02	- 1.85E+00	4.72E- 01	-2.37E-02	0	0
Plate 7a	strbp	2.47E-01	7.89E-01	4.45E- 01	-1.25E-01	0	0
Plate 7a	trim32	3.81E-02	- 2.03E+00	4.28E- 01	1.81E-01	0	0
Plate 7a	trmt1	1.78E-02	- 2.40E+00	2.35E- 01	-8.32E-01	0	0
Plate 7a	trmt2a	6.74E-02	- 1.69E+00	4.04E- 01	-2.51E-01	0	0
Plate 7a	trove2	3.69E-02	2.04E+00	2.85E- 01	6.47E-01	0	0
Plate 7a	yipf1	8.54E-04	3.60E+00	2.26E- 01	-8.62E-01	0	0
Plate 7a	zcchc12	3.77E-01	3.57E-01	2.72E- 01	-6.94E-01	0	0
Plate 7b	dfect	3.15E-04	3.83E+00	1.67E- 03	-3.28E+00	0	0
Plate 7b	l2i	1.26E-27	- 2.13E+01	8.35E- 05	4.32E+00	3	3
Plate 7b	nothing	2.85E-08	6.68E+00	1.48E- 01	-1.18E+00	1	0
Plate 7b	sart3	3.77E-01	3.66E-01	3.69E- 04	3.89E+00	2	2
Plate 7b	sbds	1.91E-16	- 1.28E+01	2.72E- 01	6.92E-01	1	1
Plate 7b	scaf1	1.44E-03	3.41E+00	3.86E- 02	2.02E+00	0	0
Plate 7b	sigfp	1.20E-20	- 1.50E+01	4.70E- 01	3.43E-02	1	1
Plate 7b	thoc6	6.40E-02	1.73E+00	9.03E- 02	1.50E+00	0	0
Plate 7b	thumpd1	4.70E-01	-5.21E- 02	2.57E- 02	-2.22E+00	0	0
Plate 7b	thumpd3	2.00E-01	9.54E-01	3.35E- 01	-4.89E-01	0	0
Plate 7b	tnrc6a	4.53E-07	- 6.05E+00	8.62E- 02	1.53E+00	1	0
Plate 7b	tsfm	6.61E-07	- 5.93E+00	1.79E- 01	1.04E+00	1	0
Plate 7b	tut1	2.65E-05	- 4.76E+00	6.56E- 02	-1.71E+00	1	0
Plate 7b	zfp346	9.13E-09	- 7.21E+00	3.96E- 01	2.77E-01	1	1
Plate 7c	dfect	1.34E-05	4.73E+00	6.80E- 03	-2.75E+00	0	0
Plate 7c	l2i	4.59E-30	- 2.38E+01	3.35E- 06	5.31E+00	3	3
Plate 7c	nothing	4.07E-07	5.93E+00	1.15E- 02	-2.57E+00	1	0
Plate 7c	sigfp	1.80E-25	- 1.91E+01	4.17E- 02	1.97E+00	1	1
Plate 7c	snupn	2.89E-22	-	4.73E-	-3.81E+00	3	3

Plate 7c	tcea3	3.82E-09	1.80E+01 - 7.46E+00	04 3.81E- 01	-3.46E-01	1	1
Plate 7c	tceal5	2.22E-02	- 2.29E+00	1.08E- 01	1.39E+00	0	0
Plate 7c	tdrd3	5.42E-02	1.83E+00	1.29E- 03	-3.45E+00	0	0
Plate 7c	tdrd7	7.19E-03	- 2.78E+00	1.06E- 01	-1.40E+00	0	0
Plate 7c	traf6	6.44E-04	- 3.70E+00	1.63E- 01	-1.11E+00	0	0
Plate 7c	trdmt1	6.60E-03	- 2.82E+00	2.19E- 02	2.30E+00	0	0
Plate 7c	trim3	4.76E-02	- 1.90E+00	1.62E- 01	1.12E+00	0	0
Plate 7c	upf1	1.01E-03	- 3.54E+00	6.54E- 03	2.82E+00	0	0
Plate 7c	xpo5	1.53E-01	1.16E+00	1.20E- 01	1.32E+00	0	0
Plate 7c	ybx2	2.32E-15	- 1.20E+01	2.36E- 05	4.80E+00	3	3
Plate 7c	zrsr2	1.08E-14	- 1.15E+01	2.83E- 02	-2.17E+00	1	1

Table 2.2. Results for RBP testing per plate.
The RBPs more significant than the dfect are highlighted in Green and the ones more significant than nothing are highlighted in Yellow; the last two columns of each sheet are the significance cutoff columns for the dfect and nothing controls, respectively, with 0 being not significant, 1 being significant for GFP, 2 being significant for RFP, and 3 being significant for both.

Alternative splicing event type	Number of events	Significant events (different in Rev vs. KO)
Skipped exon	20,249	353
Mutually exclusive exon	2,594	40
Alternative 5' splice site	3,623	24
Alternative 3' splice site	4,306	26
Retained intron	4,071	19

Table 2.3. MATS results for occurrence of isoform events.

mirna	change	statistic	pvalue
mmu-miR-335-5p	-46.98	-22.75	0
mmu-miR-381-3p	165.6	22.186	0
mmu-miR-335-3p	-57.06	-17.386	0.0001
mmu-miR-27a-3p	-32.78	-15.369	0.0002
mmu-miR-7a-5p	35.98	14.61	0.0002
mmu-miR-186-5p	79.15	13.757	0.0002
mmu-miR-181d-5p	-38.14	-13.639	0.0002
mmu-miR-3473d	23.98	12.11	0.0004
mmu-miR-200b-3p	44	10.162	0.0007
mmu-miR-9-5p	-27.61	-9.354	0.001
mmu-miR-544-3p	19.42	9.15	0.0011
mmu-miR-142a-5p	-65.74	-9.112	0.0011
mmu-miR-410-3p	27.14	9.07	0.0011
mmu-miR-293-5p	-85.79	-9.01	0.0011
mmu-miR-142a-3p	-62.29	-8.927	0.0012
mmu-miR-182-5p	-137.55	-8.723	0.0013
mmu-miR-341-3p	41.17	8.533	0.0014
mmu-miR-1954	18.67	8.068	0.0017
mmu-miR-493-3p	19.65	7.981	0.0017
mmu-miR-24-2-5p	-16.61	-7.921	0.0018

Table 2.4. Significant miRNAs

Family Members	change	statistic	pvalue	padj
miR-134-5p; miR-134-3p; miR-154-5p; miR-154-3p; miR-299a-3p; miR-299a-5p; miR-300-3p; miR-323-5p; miR-323-3p; miR-329-3p; miR-329-5p; miR-376a-3p; miR-376a-5p; miR-377-3p; miR-377-5p; miR-379-5p; miR-379-3p; miR-380-5p; miR-380-3p; miR-381-3p; miR-382-5p; miR-382-3p; miR-409-5p; miR-409-3p; miR-410-3p; miR-376b-3p; miR-376b-5p; miR-411-3p; miR-411-5p; miR-412-5p; miR-485-5p; miR-485-3p; miR-543-3p; miR-543-5p; miR-539-5p; miR-541-3p; miR-541-5p; miR-494-3p; miR-376c-3p; miR-487b-3p; miR-369-5p; miR-369-3p; miR-758-5p; miR-758-3p; miR-668-3p; miR-667-3p; miR-667-5p; miR-666-5p; miR-666-3p; miR-496a-3p; miR-679-5p; miR-495-3p; miR-544-3p; miR-1193-5p; miR-1193-3p; miR-1197-3p	15.763	8.506	0	0
miR-470-5p; miR-871-3p; miR-881-3p; miR-465c-5p	-9.859	-12.226	0	0
miR-335-3p; miR-335-5p	-52.018	-20.951	0	0
miR-23a-3p; miR-24-2-5p; miR-27a-3p; miR-27a-5p	-18.204	-7.939	0	0
miR-127-3p; miR-127-5p; miR-136-5p; miR-136-3p; miR-337-5p; miR-337-3p; miR-431-3p; miR-431-5p; miR-433-3p; miR-433-5p; miR-434-5p; miR-434-3p; miR-540-3p; miR-665-3p; miR-673-3p; miR-673-5p; miR-493-3p	13.545	5.43	0	0
miR-142a-3p; miR-142a-5p	-64.013	-13.162	0	0
miR-181c-5p; miR-181c-3p; miR-181d-5p	-31.233	-7.525	0	0.0001
miR-106a-5p; miR-363-3p; miR-363-5p; miR-20b-3p; miR-20b-5p; miR-18b-5p	13.144	5.191	0	0.0001
miR-200b-3p; miR-200a-5p; miR-200a-3p; miR-429-3p	27.089	6.413	0	0.0001
miR-582-3p; miR-582-5p	9.62	6.631	0.0002	0.0024
miR-3473d	23.976	11.725	0.0002	0.0025
miR-196a-1-3p; miR-196a-5p	11.567	5.926	0.0004	0.0043
miR-423-3p; miR-423-5p	12.082	5.803	0.0004	0.0046
miR-9-5p	-27.615	-9.558	0.0005	0.0046
miR-28a-5p; miR-28a-3p	18.295	5.644	0.0005	0.0047
miR-23b-3p; miR-27b-3p; miR-27b-5p; miR-24-3p; miR-24-1-5p	14.033	4.046	0.0007	0.0059
miR-1954	18.666	8.013	0.0009	0.0077
miR-341-3p; miR-341-5p; miR-370-3p	20.744	4.186	0.001	0.0077
miR-1191a	17.459	7.63	0.0012	0.0086
miR-182-5p; miR-183-5p; miR-183-3p; miR-96-3p; miR-96-5p	-42.119	-3.447	0.0027	0.0185

Table 2.5. Significant cis-transcribed miRNA clusters.

miRNA family	change	statistic	pvalue
miR-300/381/539-3p	165.6	21.827	0
miR-335/335-5p	-46.98	-19.662	0.0001
miR-335-3p	-57.06	-16.292	0.0001
miR-186	79.15	13.422	0.0003
miR-7/7ab	35.98	13.117	0.0003
miR-674/674-5p/3473d	22.92	9.755	0.0009
miR-142-5p	-65.74	-8.957	0.0012
miR-142-3p	-62.29	-8.767	0.0013
miR-200bc/429/548a	44.23	8.728	0.0013
miR-182	-137.55	-8.665	0.0013
miR-9/9ab	-27.61	-8.649	0.0013
miR-410/344de/344b-1-3p	27.14	8.402	0.0015
miR-341	41.17	8.255	0.0016
miR-544/544ab/544-3p	19.42	7.961	0.0018
miR-181abcd/4262	-53.52	-7.656	0.0021
miR-28-5p/708/1407/1653/3139	22.35	7.474	0.0023
miR-380-3p	58.06	7.27	0.0025
miR-493/493b	19.65	7.166	0.0026
miR-1954/3158-5p	18.67	7.155	0.0026
miR-673-5p	42.87	7.12	0.0027

Table 2.6. Significant targetscan-defined miRNA clusters.

Chapter 3: Inferring tumor evolution through heterogeneity of the epigenome

IMPORTANCE OF STUDYING CANCER

Cancer is a vastly complex disease exhibiting a plethora of genomic alterations and resulting regulatory failures at the root of its progression. Following the milestone of the Human Genome Project era, researchers are now focusing on integrating the rich genome-wide association studies (GWAS), single-nucleotide polymorphism (SNP) data, and identified gene signatures within the *in vitro*, *in vivo* and clinical frameworks to further our understanding of the molecular mechanisms of carcinogenesis. More recently, the study of epigenetic changes that occur during carcinogenesis is rapidly developing into an important research field. For example, epigenetic silencing that occurs through the CpG island methylator phenotype (CIMP) embodies a novel viewpoint in cancer diagnosis and therapy [40]. The dynamic co-dependencies between genomics, epigenetic signatures, and post-translational modifications contribute to the complicated regulation of tumor progression. Additionally, the regulatory effect of non-coding RNAs, such as microRNAs, has been also implicated in the malignant signaling networks. While the information derived from novel technologies becomes abundant, the current challenge lies in the ability to effectively and rigorously integrate and analyze such diverse data types to create biological insights and actionable translational solutions that improve cancer therapy strategies [41, 42].

TUMOR HETEROGENEITY

At the time a patient is first diagnosed with cancer, the tumor may be composed of tens of millions of cells. These cell populations have already diversified, producing a tumor that can be highly heterogeneous. Such intratumoral heterogeneity (ITH) has been observed in

spatially distinct regions of solid tumors [43-46] and among individual cells in solid tumors or leukemias [47-52]. Profiling ITH provides a powerful opportunity to trace back through the formation of the malignancy and reconstruct the tumor's evolution, allowing for the discovery of tumor initiating events and subsequent stepwise development of malignant subclones [53, 54].

In addition to genomic alterations, tumor formation involves the co-evolution of cancer cells together with its stroma – the extracellular matrix, vasculature and immune cells. Successful outgrowth of tumors and eventual metastasis is thus determined not only by oncogenic mutations, but also by the fitness advantage such alterations confer in a given environment. As fitness is context dependent, evaluating tumors as complete organs, and not simply as masses of transformed cells, becomes essential. Through such studies, it became apparent that the dynamic tumor topography varies drastically even throughout the same lesion, and moreover, that the heterologous cell types within tumors and its environment can actively influence therapeutic response and treatment resistance [55].

It has been shown that ITH present at diagnosis may be altered by cytotoxic or targeted cancer therapies that exert additional selective pressure, promoting outgrowth of one or more therapy-resistant tumor cell clones [56]. Therapeutic interventions could therefore lead to contraction of ITH in some cases or expansion in others, influencing response to subsequent therapies and patient outcome. In most ITH studies, however, only a small fraction of the whole tumor is available for analysis. Furthermore, for most ITH studies

the samples lack information on where within the heterogeneous tumor they were obtained.

GLIOMAS

Diffuse low-grade and intermediate-grade gliomas (World Health Organization [WHO] grades II and III, hereafter called lower-grade gliomas) are infiltrative neoplasms that arise predominantly in the cerebral hemispheres of adults and include astrocytomas, oligodendrogliomas, and oligoastrocytomas [57, 58]. Due to their highly invasive nature, total neurosurgical resection is often not possible. The presence of residual tumor is thus the leading cause of recurrence and malignant progression. The treatment options are decided based on factors such as the extent of the tumor resection, tumor grade and the presence of metastatic disease, and include clinical monitoring, chemo- and radio-therapy [59-63]. Given these strategies, the rate of recurrence is highly variable [64-66]; a subset of low-grade gliomas will progress to glioblastoma (WHO grade IV gliomas) within months, while others could remain stable for years. The survival of low grade-glioma patients ranges from 1 to longer than 15 years, with a subset of patients exhibiting an impressive therapeutic sensitivity [67].

Recent genomic findings of driver mutations in lower-grade gliomas have become paramount in assisting histopathological classification in adequately predicting clinical outcomes [68]. Mutations in *IDH1* and *IDH2* (referred to collectively as *IDH*) characterize the majority of low-grade gliomas in adults and are associated with a favorable prognosis. It has also been reported that low-grade oligodendrogliomas with both an *IDH* mutation and a co-deletion of chromosome arms 1p and 19q have better

responses to radiochemotherapy and are associated with longer survival than diffuse gliomas without these alterations [69, 70]. Similarly, *TP53* and *ATRX* mutations are more frequent in astrocytomas and are likewise important markers of clinical behavior [71]. Mutation of the TERT promoter, which has been frequently reported in grade II oligodendrogliomas and is the most frequent mutation across grade IV GBMs, may be an additional classification aiding defining feature [31].

GENETIC ANALYSIS OF GLIOMAS AND THEIR RECURRENCES

With Glioma recurrence being a relatively frequent event and a major cause of mortality, it is essential to profile the genetic and molecular drivers of the recurrence. Furthermore, to facilitate evaluating which therapies will be most effective in treating the recurrent disease, it becomes critical to determine how genomic drivers differ between the initial and the recurrent tumor. [45] sequenced the initial and recurrent Gliomas from 23 patients. Their work showed that for more than 40% of the patients, the majority of mutations found in the initial tumors were not present in their patient-matched recurrences. This observation suggested that the recurrent tumor was already seeded at during the growth and evolution of the initial Glioma.

It was further shown that recurrent Gliomas that progress to a GBM acquire genetic alterations in the RB and AKT-mTOR pathways [45, 72-74]. In fact, [45] linked treatment-associated driver mutations in these two pathways to malignant progression of grade II glioma to GBM, induced by the alkylating chemotherapeutic temozolomide (TMZ).

The treatment associated malignant progression follows selection of tumor cells with epigenetic silencing of the DNA repair protein MGMT [75]. It remains unknown, however, how genome-wide epigenetic alterations contribute to the different courses of evolution of low-grade gliomas and how or if they relate to concurrent mutational evolution.

WHAT IS EPIGENETICS?

All cells throughout the body maintain the same sequence within their DNA, leading to a fundamental question of how does the same starting material lead to different tissue types. It becomes apparent that the genetic code is not the only determining factor in differentiation and development. This can be partially explained by epigenetics, the study of how the tertiary structure of the genome controls gene expression. In short, cells, tissues, and organs differ because they have certain sets of genes that are expressed, as well as others that are inhibited. Epigenetic modification is one such way to regulate the transcription state of a gene and it can contribute to differential expression. Moreover, epigenetics has been shown to play an important role in X-chromosome inactivation in female mammals, which regulates the number of X-chromosome gene products [76].

IMPORTANCE OF EPIGENETICS TO CANCER

How cancer cells harness such epigenetic processes for therapy resistance is an important topic in current research. Powerful *in vitro* and *in vivo* models have shown that epigenetic heterogeneity can drive variable responses to therapy and differences in tumor-propagating potential. Gupta et al. [77] show that upon separation of a breast cancer cell

line into its basal, luminal, and stem-like cell populations, each of the purified cell types expands into a heterogeneous culture that fully recapitulates the initial cell type heterogeneity through cell state interconversions. Kreso et al. [78] further show that after isolating individual cells from the same genetic background and transplanting them into mice, the separate transplants display differences in growth dynamics and treatment-response. Similarly, Sharma et al. [56] find that while the majority of cells in a single cell derived non-small cell lung cancer subline are drug-sensitive, a small subpopulation of cells are drug-tolerant. Following removal of drug, these drug-tolerant persister cells expand and reacquire drug-sensitivity. Persister cells display an altered chromatin landscape, suggesting that epigenetic therapies could block persister cells. Indeed, treatment of cell lines with HDAC inhibitors or knockdown of the histone demethylase KDM5A reduces the emergence of persister cells. This persister cell model mimics observations that some patients respond initially to therapy, develop resistance, and then will respond again to the same chemotherapy after a drug holiday [79]. Thus, one hypothesis is that epigenetic ITH at the single cell level may play a role in therapy responses in patients, and concurrent treatment with epigenetic therapies may improve drug responses [80, 81].

DNA METHYLATION

One of the forms of epigenetic control is DNA methylation, which is a heritable chemical change to DNA that involves attachment of a methyl group to the DNA. This modification is common to a nucleotide sequence site where a cytosine is followed by a guanine and linked by a phosphate, called a CpG site [76, 82, 83]. CpG sites are

methylated by one of three enzymes called DNA methyltransferases (DNMTs) [76, 83]. Methylation changes the appearance and structure of DNA, resulting in modification of gene's interactions with the nuclear transcription machinery. Thus when a gene's promoter CpG dinucleotides become methylated, this results in transcriptional silencing that can be inherited by daughter cells following cell division.

DNA METHYLATION IN CANCER

The finding of increased DNA methylation in colorectal cancer tissue when compared to healthy tissues of the same patients was one of the first links made between epigenetics and cancer [84]. Loss of DNA methylation can cause abnormally high gene activation via lack of methylation-mediated suppression. On the other hand, too much methylation could suppress the protective tumor suppressor genes.

While a majority of CpG cytosines are methylated in mammals, there are stretches of DNA near promoter regions that have higher concentrations of CpG sites (CpG islands) that are free of methylation in normal cells. In early cancer development, it has been reported that these CpG islands become excessively methylated, silencing the genes that should be expressed or maintaining silencing in genes that were already transcriptionally inactive [76, 82, 83]. Moreover, hypermethylation of CpG islands can initiate tumors by turning off tumor-suppressor genes. Such malignant signaling may in fact be more widespread in the cell that DNA sequence mutations.

While epigenetic changes themselves do not alter the nucleotide sequence of DNA, they have been found to cause mutations downstream of the epigenetic gene silencing. For example, hypermethylation of the promoter of MGMT was associated with an increase in the number of G-to-A mutations [85]. Additionally, about half of the genes that cause familial or inherited forms of cancer, and thus driven by their genetics, can also be turned off by aberrant methylation in sporadic cancers.

In another mode of action, hypomethylation can cause instability of microsatellites which as been linked to many cancers, including colorectal, endometrial, ovarian, and gastric cancers [82]. Increased methylation of the promoter MLH1 DNA repair gene can make microsatellites unstable and either lengthen or shorten them.

DNA METHYLATION IN GLIOMA

The critical role that epigenetic alterations play in the development and therapeutic response of gliomas is increasingly being appreciated [86]. Epigenetic mechanisms can alter gene expression and affect tumor suppressors and oncogenes in gliomas [87-92]. Somatic mutation in IDH1 or IDH2 may be the first genetic driver in the development of many low-grade gliomas [45, 93, 94]. Genetic mutations in IDH genes induce a pattern of early epigenetic alterations known as the glioma CpG island methylator phenotype (G-CIMP) characterized by extensive remodeling of the DNA methylome [95-98]. The inactivation of other genes mutated in low-grade gliomas, such as ATRX [99] and SMARCA4 [45], is known to induce specific DNA methylation changes as well [100, 101]. Of clinical importance is DNA hypermethylation of the MGMT promoter, which is

associated with loss of SP1 binding, closed chromatin, and transcriptional silencing in GBM cells [102], and increased survival in GBM patients treated with TMZ [37]. Whether the DNA methylation status at this locus predicts the same survival benefit in patients with low-grade glioma is unclear [75, 103-106]. Although there has been extensive characterization of tumor methylomes using a single sampling per tumor, little is known about intratumoral heterogeneity at the epigenetic level or of temporal evolution of the low-grade glioma methylome and its relationship to the genome. An integrated model of the genomic and epigenomic evolutionary trajectory of initially low-grade gliomas may suggest strategies for delaying or treating recurrent disease, identify biomarkers for predicting the clinical course of a low-grade glioma, and shed light on dynamic relationships between the genome and epigenome in other cancer types.

DNA METHYLATION ANALYSIS

To understand the importance of epigenetics through brain cancer development and progression, 19 initial grade II glioma patients had their initial and recurrent tumors profiled using the Illumina HumanMethylation450 bead array (Illumina 450K). This technology measures approximately 485,000 CpG sites along the genome and estimates the percentage of methylated cytosines at each of those sites. Probe-level signals for individual CpG sites were subject to both background and global dye-bias correction [107]. Probes that map to regions with known germline polymorphisms (Illumina supplementary SNP list v1.2, downloaded Sept. 3, 2013), to multiple genomic loci [108], or to either sex chromosome were filtered out. 297,342 probes remained following filtering. From looking at the DNA methylation profiles of each individual patient, it is

apparent that the sampled sites produce a bimodal distribution for each individual patient sample. Specifically, the two peaks, centered around 0.15 and 0.85, suggest that cytosines are primarily methylated across the entire population of cells or not with fewer sites occurring at a rate of 50% throughout the cells (Figure 3.1).

All gliomas profiled here (Table 3.1) are IDH1 mutant [45, 75] and are therefore expected to possess the characteristic methylation patterns associated with G-CIMP [30, 95, 96]. From these methylation array data (Figure 3.2), we confirmed that the glioma CpG island methylator phenotype (G-CIMP) is present in all tumors profiled here by examining methylation levels at CpGs adjacent to eight previously defined markers (ANKRD43, HFE, MAL, DOCK5, LGALS3, FAS-1, FAS-2, RHOF) [96]. The observation that G-CIMP was present in all initial tumors and always maintained at recurrence highlights that these epigenetic changes arise very early and are potentially tumor-initiating.

Global difference between Grade IV and low-grade methylation profiles

To determine the extent to which these tumors had altered methylomes beyond the ubiquitous G-CIMP methylation patterns, we identified the most variable CpG sites across all initial and recurrent gliomas and performed unsupervised hierarchical clustering. To determine common and specific methylation profiles in the paired initial and recurrent tumors, we performed two-way unsupervised hierarchical clustering using Euclidean distance and Ward linkage on the most variable CpG sites across the cohort,

with variability ranked by standard deviation (0.5% cutoff = 1,486 CpGs; 50% cutoff = 148,572 CpGs).

Initial and recurrent tumors from the same individual clustered together. This result reflects patient-specific methylation patterns, consistent with a previous report on glioma [109], and may be indicative of normal inter-individual epigenetic variation, patient-specific aberrant methylation from early stages of gliomagenesis, or both. Within the clustering, six of the seven patients who recurred with GBM formed a distinct subgroup, suggesting there may be a shared methylation pattern associated with malignant progression to GBM relative to a lower grade of recurrence. To further evaluate this pattern, we performed unsupervised clustering with progressively more lenient selections of variable CpG sites to discover additional global DNA methylation patterns. At intermediate cut-offs, a gradual switch in clustering patterns was evident (Figure 3.3). At the most lenient cutoff, the methylation patterns separated GBM recurrences, as well as two initial tumors that recurred as GBM, from the grade II and III gliomas (Figure 3.4). This further supports a GBM recurrence-specific methylation pattern and suggests extensive evolution of the methylome during malignant progression to GBM (Figure 3.5). This unique pattern of epigenome evolution was prominent across GBM recurrences that arose in the absence of adjuvant therapy as well as in GBMs that arose in a treatment-associated manner, adding to our previous genetic findings that spontaneous and treatment-associated progression to GBM have convergent genetic alterations (Johnson et al., 2014).

TRANSCRIPTOME ANALYSIS

Additionally, strand-specific transcriptome sequencing libraries were prepared for the initial and recurrent tumors of 13 patients as previously described [45]. All transcriptome sequencing data from initial and recurrent tumor pairs were aligned with TopHat (v2.0.12) [110] to the hg19 reference genome using a GENCODE transcriptome-guided alignment. The aligned data were then processed through custom quality-control scripts to remove unmapped, improperly matched, multi-mapping, and chimeric reads, as well as accumulation in non-assembled chromosomes. To estimate transcript abundance, aligned data were processed with the cuffnorm and cuffquant commands from the Cufflinks package (v2.2.1) [111]. For all subsequent statistical analyses, FPKM estimates generated from cuffnorm output for individual genes were made more Gaussian using a log2-transformation.

Interestingly, clustering of the transcriptome at both the top 1% and the top 50% most variably expressed genes segregated some of the grade III recurrences with GBM samples (Figures 3.6), indicating transcriptional changes are complementary to, but not exclusively overlapping with, changes in the DNA methylome during malignant progression. Thus, integrating the methylome and transcriptome may provide important insight into the functional epigenetic events that underlie malignant progression to GBM.

INTEGRATED ANALYSIS

Identification of CpGs that Lose Methylation Specifically during Malignant Progression to GBM

We next examined changes in the methylome and transcriptome to determine whether there is a signature of methylation or expression changes associated with recurrence. We calculated the change in methylation (β value, methylated fraction at a CpG site) from initial to recurrent tumor at each CpG site in each patient, and then identified CpG sites with consistent methylation changes upon recurrence across all patients.

The beta values for individual CpG sites were made more Gaussian using the logit-transformation. We subtracted the transformed beta values between patient-matched recurrent and initial tumors and used Limma [2], an empirical Bayes approach utilizing a moderated t-statistic, to test for significant differences in individual CpG sites between the group of patients that recurred as GBM and the group that did not. Differentially methylated CpGs were defined as those with both a nominal p-value < 0.05 and an average methylation change upon recurrence ≤ -0.2 or ≥ 0.2. The same empirical Bayes approach was also used to compare methylation differences between the GBM and non-GBM groups. Hypomethylated CpGs were defined as those with both a nominal adjusted p-value < 0.05 and an average methylation change upon recurrence as GBM ≤ -0.2 and a difference of the average change between the GBM and non-GBM group of -0.15.

This powerful intra-patient approach controls for differences in DNA methylation that are age-related or reflect germline genetic effects, which confound inter-patient comparisons. DNA methylation differences between normal brain and glioma may be aberrant events in the tumor or may reflect differences between the normal brain tissue sample and the methylation patterns of the tumor's cell of origin [112, 113]. We used Limma to test for

the differential methylation between 33 fetal and 8 adult brain tissues. We selected probes having both a nominal adjusted p-value (derived from the previous analysis) < 0.05 and an average methylation change upon aging ≥ 0.2.

In contrast, the differences we report between initial and recurrent tumors are more likely to be aberrant changes attributable to tumor progression rather than cell of origin. We also applied an equivalent model to the transcriptome sequencing data and identified genes that commonly increase or decrease in expression from initial to recurrent glioma (Figure 3.7). We subtracted the transformed FPKM estimates between patient-matched recurrent and initial tumors and used Limma to test for significant differences among individual genes within the group of patients that recurred as GBM and the group that did not. Differentially expressed genes were defined as those with both a nominal p-value < 0.05 and an average log 2-fold change upon recurrence ≤ -1 or ≥ 1. Limma was again used to compare methylation differences between the GBM and non-GBM groups. Upregulated genes were defined as those with both a nominal p-value < 0.05 and an average log 2-fold change upon recurrence as GBM ≥ 1 and a difference of the average change between the GBM and non-GBM group of at least 1.

The separation by grade in the methylation clustering suggested that a specific pattern of DNA methylation changes may underlie malignant progression to GBM. To discover this pattern in detail, we stratified patients by grade of recurrence. There were few common methylation changes evident in tumors that recurred at grade II or III, whereas a strong pattern of hypomethylation was associated with malignant progression to GBM (Figures

3.8). Patients with tumors that recurred at grade II or III were combined into a single group for further analysis.

To determine which methylation changes were specific to recurrence as GBM, we compared the change in methylation from initial to recurrence in patients who recurred as GBM versus those that recurred at grades II or III. We identified 1,953 CpG sites that were hypomethylated specifically upon recurrence as GBM (Figure 3.9 and 3.10).

Given the G-CIMP-associated hypermethylation in these tumors, we first set out to determine if the hypomethylation in GBM recurrences affected G-CIMP genes. Noushmehr et al. [96] identified 50 genes that were hypermethylated and downregulated in a G-CIMP specific manner. Only two of those genes (ACSS3 and RAB36) showed GBM-specific hypomethylation, but in neither case did the genes show concurrent increased expression. Further examination of these sites of decreasing methylation revealed a surprising enrichment for CpG sites that undergo age-related increased methylation in a comparison of normal fetal and adult brain (odds ratio 4.64, $p < 0.0001$, permutation test). This is contrary to the typical pattern in cancer in which CpG sites that are hypermethylated during aging are also hypermethylated in cancer [97, 114].

To further investigate whether the methylation changes alter gene regulation, we integrated active regulatory regions defined from histone H3K4me3, H3K4me1, and H3K27ac chromatin immunoprecipitation sequencing in adult normal brain and primary GBM tissue and found that sites of GBM-specific DNA hypomethylation were enriched

for candidate active enhancers (odds ratio 1.68, $p < 0.0001$, permutation test). These hypomethylated loci thus may have gene regulatory effects. To enrich for functional methylation changes and exclude passenger events, we next integrated our transcriptome sequencing data with the DNA methylation analysis.

Statistical tests for assessing significant differences in gene expression and methylation status were performed independently. The varying number of tests performed (~300k for methylation and ~25k for expression), makes it difficult to directly compare the resulting p-values. While Storey's false discovery rate controlling for multiple-testing corrections are standard [4], our data show bimodal distribution of the RNA-seq analysis p-values and do not satisfy the assumptions required to apply the method, resulting in incorrect estimation of the number of genes in the null distribution. Thus, we chose our p-value cutoffs by identifying the value at which we would identify an equal number of false positives if all the test cases satisfied a null hypothesis and the p-values had a uniform distribution. Specifically, by using a .05 cut-off in the expression data, under our simplistic assumptions, to identify the same number of false positives in the methylation data, we would need to use a cutoff padjusted–methylation = pmethylation * (N450k probes/ Ngenes). Our use of p-values here is primarily to rank all probes and genes in our study and follow-up by selecting only those with the most consistent difference.

Cell Cycle Genes Are Specifically Hypomethylated upon Malignant Progression

We applied an analysis similar to that of the methylation data and identified 528 genes with GBM-specific overexpression. Of these, 39 genes showed GBM-specific

hypomethylation of at least one CpG site within their promoter regions. Among genes with GBM-specific promoter hypermethylation, only NTSR2 showed consistent transcriptional downregulation. We additionally identified four genes with consistent downregulation and gene body hypomethylation. Strikingly, the set of 39 promoter-hypomethylated and overexpressed genes was significantly enriched for cell cycle genes (Figure 3.11). Ki-67 is a marker of cells in the active stages of the cell cycle, and staining in initial and recurrent tumors confirmed that a statistically significantly higher fraction of positive cells ($p = 0.026$, two-sided Wilcoxon rank sum test) were found among the GBM recurrences (Figure 3.11). Increased proliferation is a hallmark of GBM. These results thus highlight an epigenetic mechanism that may contribute to increased proliferation, concurrent with genetic alterations in key members of the RB pathway [45] that abrogate the G1/S cell cycle checkpoint.

The functional effect of DNA hypomethylation of cell cycle genes specifically upon recurrence as GBM parallels the known GBM-specific genetic events that inactivate the G1/S cell cycle checkpoint [45, 73, 74, 96]. These convergent genetic and epigenetic signals, in addition to the well-characterized functional relationships between genetic and epigenetic aberrations [95, 100, 101, 115, 116], prompted us to explore evolutionary relationships among different tumor cell populations within a tumor, as has been previously done with genetic data, and then compare the relationships inferred from DNA methylation to those inferred from somatic mutation in the same samples.

RECONSTRUCTION OF TUMOR EVOLUTION FROM INTRATUMORAL AND LONGITUDINAL DNA METHYLATION PATTERNS

For the phylogeny analysis of both the genetic and epigenetic data, we used an independent, but parallel, analysis of the methylation data and somatic mutations derived from exome sequencing. For the exome-seq data, we used binary mutation calls to build a distance matrix for all samples from a patient using the Manhattan distance metric, including a normal tissue sample for which all mutations were absent. Similarly, for the methylation data, we used only the probes that had a beta value difference of at least 0.4 between any of the samples from a patient to build a Euclidean distance matrix. Using several other probe selection cut-offs produced similar results. A normal brain sample (adult insula tissue from a different individual) was not included in the probe selection, but was added to the distance matrix calculation to serve as the tree root. To compare the distance matrices from the mutation data and the methylation data, we calculated the Spearman's rho correlation. We then built the phylogeny trees using an ordinary least-squares minimum evolution [117] approach from the ape R package [118] using the distance matrices from the genetic and epigenetic data independently.

We first examined the evolutionary relationships of tumor samples that were previously genetically characterized [45]. We performed methylation profiling of seven spatially distinct pieces of tumor tissue from Patient17, three from the initial tumor and four from the recurrent tumor, and built a phyloepigenetic tree (Figure 3.12). The phyloepigenetic tree presented an intriguing model with early divergence between the initial and recurrent tumors, and more subtle divergences among the samples within each time point (initial A

versus initial B/C; recurrence A/C versus recurrence B/D). We then used exome sequencing data of these same spatially distinct tumor samples to independently construct a phylogenetic tree (Figure 3.12) [45]. The genetically defined relationships (branch length and bifurcation?) among tumor cell clones were consistent with those determined from DNA methylation data. We quantified this surprising degree of similarity as the correlation between the distance matrices that were used to build the phyloepigenetic and phylogenetic trees (Spearman's rho = 0.90).

To identify the CpG sites underlying each branch point in the phyloepigenetic tree, we applied singular value decomposition to the methylation data from each patient to weigh the influence of individual CpG sites on separating particular subsets of samples. The singular value decomposition (SVD) starts with a mean-centered p x n data matrix $\boldsymbol{X}$, where the rows are probes and the columns are samples from a patient. A rank-k approximation of $\boldsymbol{X}$ is obtained from the SVD of $\boldsymbol{X}$ as $\boldsymbol{X_k} = \boldsymbol{UDV^T}$, where $\boldsymbol{U}$ contains the first k left singular vectors as columns, $\boldsymbol{V}$ contains the first k right singular vectors as columns, and $\boldsymbol{D}$ is a diagonal matrix of the first k singular values. We can rewrite $\boldsymbol{X_k}$ as $\boldsymbol{X_k} = [31]\ (\boldsymbol{D^{1-a}}\ \boldsymbol{V^T}) = \boldsymbol{GH}$, where $\boldsymbol{a}$ determines the scaling of the probes and samples. A biplot uses k=2 and plots the rows of $\boldsymbol{G}$ as points and the columns of $\boldsymbol{H}$ as arrows. For the purpose of performing PCA on samples in the probe space, we used the parameter $a = 0$. The axes at the bottom and left of the biplot are the coordinate axes for the probes while the axes at the top and right of the biplot are the coordinate axes for the samples, allowing us to simultaneously represent both the separation of the samples and the magnitude of each probe contributing towards that separation.

For Patient17, the first singular vector (SV1), which accounts for the most methylation variability, mimicked the first major branch point of the phyloepigenetic tree (Figure 3.12). We then selected the most influential CpGs for each singular vector and inferred that these underlie a particular branch point. The most highly weighted CpG sites within SV1 from Patient17 clearly showed differential methylation between the initial and recurrent tumor samples (Figure 3.12). We examined the potential implications of these methylation changes by focusing on those affecting active promoters and enhancers in normal brain and primary GBM tissue and performed a gene ontology enrichment analysis. For Patient17, the CpG sites that underlie the first major branch point were enriched for a variety of developmental, biosynthetic and metabolic processes, indicating that methylation changes during tumor progression may influence cellular metabolic states, in parallel with the genetic events disrupting cell cycle that separate these two main branches on the phylogenetic tree.

We then looked specifically at the evolutionary relationships of tumor samples from patients who underwent chemotherapy-associated malignant progression [45, 75]. We performed methylation profiling of four spatially distinct pieces of the initial tumor and three pieces of recurrent tumor from Patient01 and inferred a phyloepigenetic tree (Figure 3.12). Whereas the four pieces of the initial tumor clustered together, the recurrent tumor consisted of two distinct populations. Recurrence B was relatively closely related to the initial tumor, while a long branch separated it from recurrences A and C, indicating significant evolutionary distance. A phylogenetic tree from these same tumor pieces

(Figure 3D, right; Tables S2 and S5) similarly demonstrates the large evolutionary distance between recurrence B and recurrences A and C (Spearman's rho = 0.83). In the phylogenetic tree, this longest branch corresponds to the development of a hypermutated population in the recurrent tumor. Intriguingly, this same branch is the longest in the phyloepigenetic tree, indicating that the hypermutated cells also have the greatest methylation change. Similarly, in Patient18, the phyloepigenetic tree identified three epigenetically similar pieces of the initial tumor, a piece of the initial tumor that branched off at an earlier evolutionary time point, and a recurrence that diverged even earlier—relationships that are accurately recapitulated in the phylogenetic tree (Spearman's rho = 0.90) (Figure 3.12). Thus, even in extreme evolutionary events such as chemotherapy-associated hypermutation, both DNA methylation changes and mutational landscapes encode similar tumor evolutionary relationships. In these two cases with TMZ-associated hypermutation (Figures 3D and 3E), the longest branch length in both the phyloepigenetic and phylogenetic trees is the hypermutated recurrence. These results suggest a potentially quantitative relationship between the number of mutations and epimutations in each tumor cell clone.

To determine if the strong correlations between phylogenetic and phyloepigenetic trees depend on the large-scale hypomethylation during malignant progression to GBM, we next compared the evolutionary relationships only in lower grade initial and recurrent tumors. Six pieces of tissue from the initial tumor and two pieces of tissue from the grade II recurrence from Patient90 were subjected to DNA methylation profiling. Construction of a phyloepigenetic tree revealed three distinct clusters of samples, with the initial tumor

separating into two populations, and the recurrence forming a third (Figure 3.12). We then performed exome sequencing of these same pieces of tissue to identify somatic mutations and constructed a phylogenetic tree (Figure 3.12). This phylogenetic tree mirrored the evolutionary relationships defined from DNA methylation (Spearman's rho = 0.56). We further pursued this question with Patient49 who underwent a single resection for an initial tumor from which we profiled six spatially distinct pieces. Construction of a phyloepigenetic tree revealed that the six pieces separate into two groups, in agreement with the phylogenetic tree derived from exome sequencing of the same pieces of tissue (Spearman's rho = 0.64) (Figure 3.12). Thus, even in the absence of malignant progression to GBM, DNA methylation changes among tumor cell clones yielded a very similar evolutionary trajectory as was inferred from somatic mutations.

Enhanced Model of Tumor Evolution Derived from Variation between Phyloepigenetic and Phylogenetic Trees

To further address phyloepigenetic relationships over time, we examined tumor samples from Patient04, who underwent four sequential surgical resections over 5 years. We profiled six spatially distinct pieces of tumor from the initial surgery, and one from each of the three subsequent surgeries for tumor recurrence. The phyloepigenetic tree reveals two distinct populations within the initial tumor and an evolutionary trajectory shared among the three recurrences, with a relatively closer relationship between recurrences 2 and 3 (Figure 3.13). The phylogenetic tree again reveals many similar clonal relationships, but also reveals differences that may be informative (Figure 3.13) (Spearman's rho = 0.78). Based on somatic mutations, the first recurrence shares

evolutionary history with the initial tumor, while the second recurrence diverges earlier in the evolution of the tumor and therefore independently progressed to grade III [45]. Despite divergent genetic paths, methylation patterns are shared among the first recurrence and the second and third recurrences. This raises the possibility that the last common ancestor of the first and second recurrences was primed for progression with a set of DNA methylation changes required for progression to a higher grade. This case illustrates how differences in genetic and epigenetic phylogenies may bring to light an enhanced understanding of the evolution of a tumor.

INTEGRATED MODEL OF GLIOMA GENETIC AND EPIGENETIC EVOLUTION

DNA methylation patterns record a remarkable breadth of information about cells, including their chronological age, developmental history, and differentiation potential. Here, we show that despite epigenome plasticity, chemotherapy, and the ubiquitous IDH1 mutation-driven G-CIMP pattern, patient-specific tumor phyloepigenetic analyses replicated and extended tumor phylogenetic analyses. From this striking result, we conclude that the precise chronological order of epigenetic changes, from initiating to late events, can be determined from intratumoral methylation patterns, thus surpassing prior binary categorization of epigenetic events as early or late. While our study is focused on methylation and somatic mutations in IDH1 mutant gliomas, a study of prostate cancer and prostate cancer metastasis showed a complementary unified model of evolution for DNA methylation and copy number alterations [44]. Thus, genomic-epigenomic co-

dependency may be a feature of multiple types of cancer, and may span somatic mutations, copy number, and DNA methylation.

The importance of epigenetic variation within individual human tumors is just beginning to be uncovered. Recent work in chronic lymphocytic leukemia suggests that stochastic changes in the methylome lead to increased heterogeneity, allowing for selection of more malignant epi-phenotypes coupled with an adverse clinical outcome [47]. Somatic genetic events, such as IDH1 mutations, have been directly linked to alterations in the methylome [95, 96], whereas germline variants have been indirectly associated with specific DNA methylation patterns [115, 119, 120]. Consistent with these theories, the widespread correlation between somatic mutations and DNA methylation patterns suggests that in addition to IDH1 mutation and G-CIMP, other epigenetic patterns might be directly or indirectly induced by mutations, or vice versa. It will be of interest to determine the extent to which these findings hold for IDH1-wild-type low-grade gliomas and their recurrences.

We also discovered a convergence of genetic and epigenetic changes driving aberrant cell cycle function (Figure 3.14). We previously found that recurrent tumors that underwent malignant progression to GBM acquired somatic mutations in the RB pathway that inactivate the G1/S cell cycle checkpoint [45]. Here we identified a pattern of functional DNA hypomethylation specific to recurrence as GBM that alters cell cycle genes. This phenotypic convergence of genetic and epigenetic mechanisms on the same pathway underscores the importance of cell cycle deregulation on the process of malignant

progression, while also raising questions about how these two processes might be connected. Of note, we identify hypomethylation at TP73 as a recurrent event. Transcription of TP73 is upregulated by E2F1 [121], a transcription factor that itself activates cell cycle progression-related genes following inactivation of the RB pathway [122], which is deregulated by genetic mechanisms in these tumors. Further work will be required to deconvolute these relationships. By combining the information from somatic mutations, copy number alteration and DNA methylation patterns, we derived a comprehensive model of glioma evolution (Figure 5). Chronological ordering of IDH1, TP53, and ATRX mutations and copy number alterations was derived from our previous tumor phylogenetic analyses [45], other studies [93, 94], and additional data presented here. This model is derived from 32 patients with paired initial and recurrent samples and includes 70 DNA methylation profiles, 26 mRNA expression profiles and 130 exome sequencing profiles. The model extends from the initiating genetic and epigenetic lesions and captures clinically divergent paths at recurrence, including an evolutionary path driven by treatment.

These findings underscore the power of integrated genetic and epigenetic analyses of tumors. Deregulated cell cycle control is among the essential phenotypes of cancer cells, and we demonstrate that this deregulation is encoded in both the genome and epigenome, raising the question of the extent to which this reflects a functional interaction between genetics and epigenetics. This finding also raises the possibility that other critical molecular phenotypes, such as genomic instability, angiogenesis, or invasion, may leave their imprint on DNA methylation patterns during tumor evolution.

FUTURE DIRECTIONS

While important steps have been taken to understand heterogeneity from multiple spatial samples of a tumor, additional work is still necessary to compute an overall measure of heterogeneity for a spatially-distributed tumor. Thus an essential next step is to develop methods that infer tumor-wide heterogeneity across multiple spatially distinct samples from the same individual. Lui et al. [123] demonstrated the effectiveness of one approach for determining heterogeneity across multiple samples. The authors built a gene co-expression network from 96 serial samplings of normal brain tissue. They then identified modules of genes with similar expression profiles across the 96 samples. Finally, using the number of separate modules that the genes can be separated into, the authors estimated the number of subtypes present within this tissue. So far this method has been applied to gene expression, but the underlying technique can be applied to data types including DNA methylation and other epigenetic marks. While the utility of this approach has already been demonstrated for normal tissue, further work is required to extend this method to apply into cancer cells.

Furthermore, to more fully understand tumor heterogeneity between tumor subclones and to build a comprehensive evolutionary history of cancer progression, a novel analytical approach combining genetic and epigenetic data is required. Although several studies have found a substantial correlation between tumor evolution traced from DNA methylation compared to genetic alterations such as somatic mutations or CNVs, it is not yet possible to create a theoretical mathematical model to understand how much co-

dependency exists between the genetics and epigenetics, as the rate, timing, and location of exact DNA methylation changes is not well known.

FIGURES

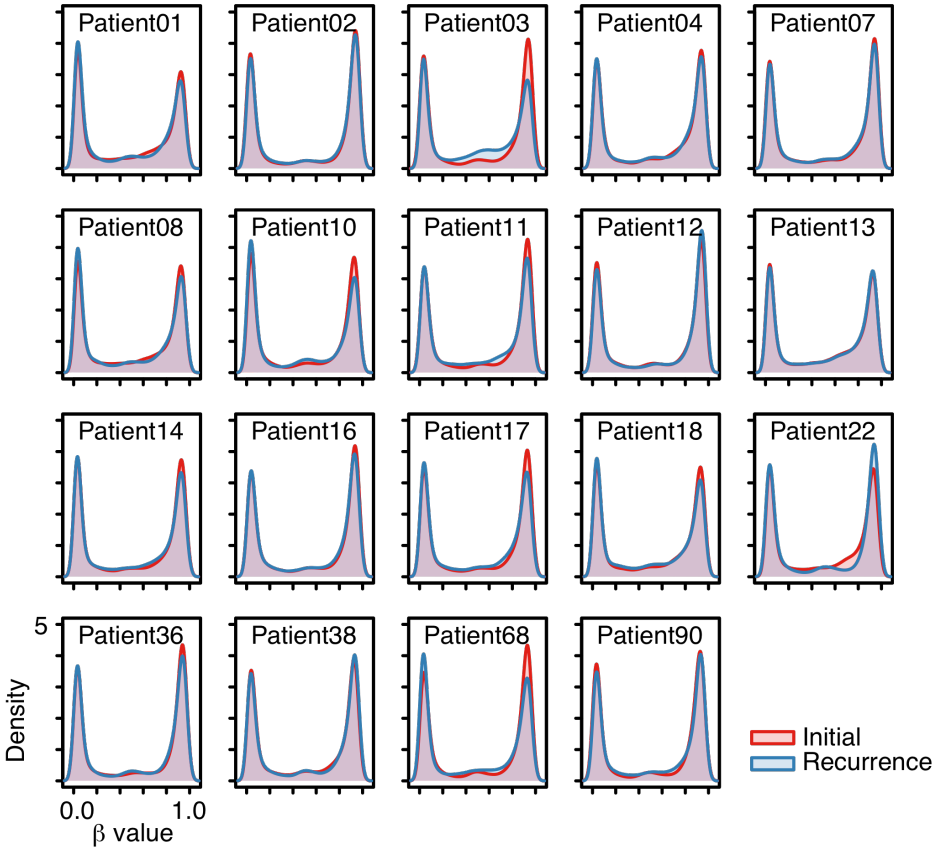


Figure 3.1. Beta value disttributions.
Density plots of background corrected and normalized beta values in each initial and recurrent tumor.

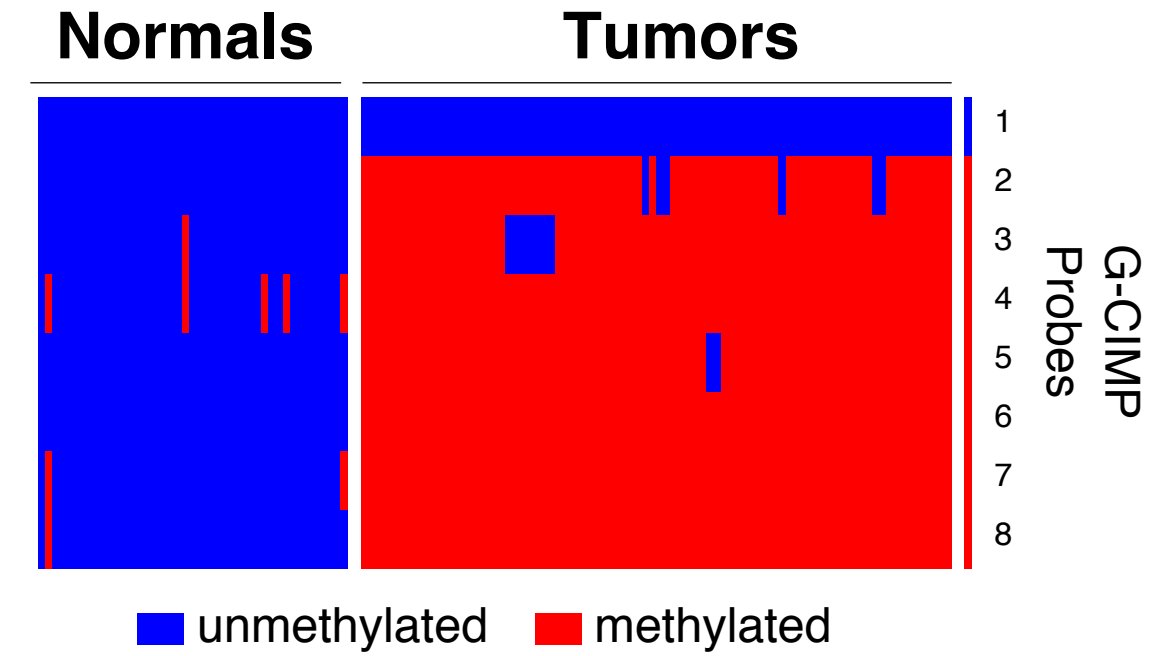


Figure 3.2. G-CIMP signature across tumors and normal samples.
Confirmation of the presence of the glioma CpG island methylatory phenotype (G- CIMP) in all initial tumors and maintenance of G-CIMP at recurrence (tumor N=70). G- CIMP is absent from all normal brain tissues examined (normal brain N=38).

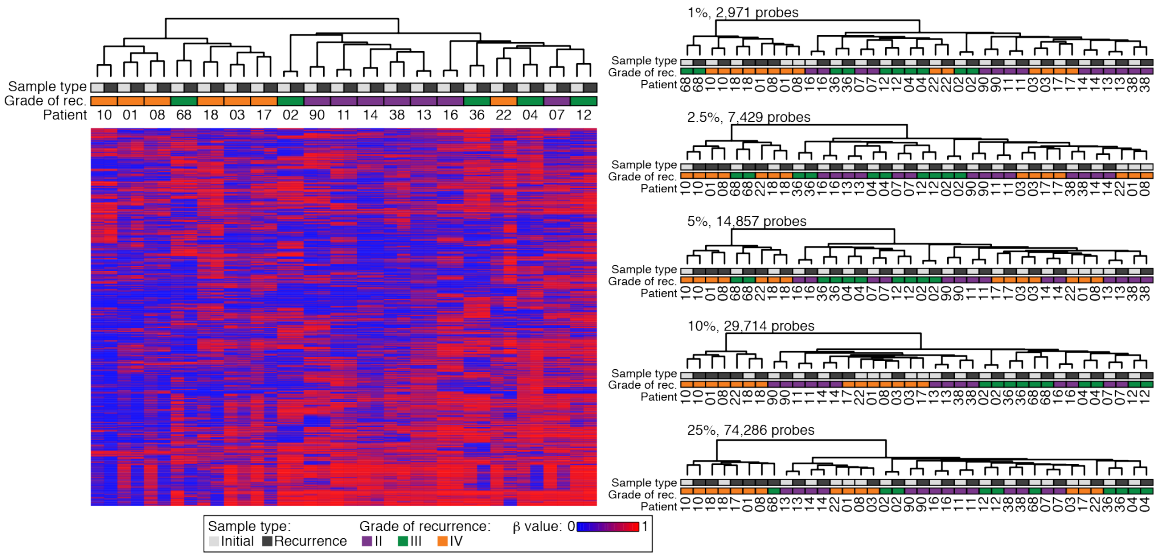


Figure 3.3. Unsupervised clustering of beta values at gradual cutoffs.
(Left) Unsupervised hierarchical clustering of the top 0.5% most variable CpG sites and heatmap of beta values. (Right) Unsupervised hierarchical clustering of the most variable CpG sites at intermediate (top 1%, 2.5%, 5%, 10%, 25%) cutoffs.

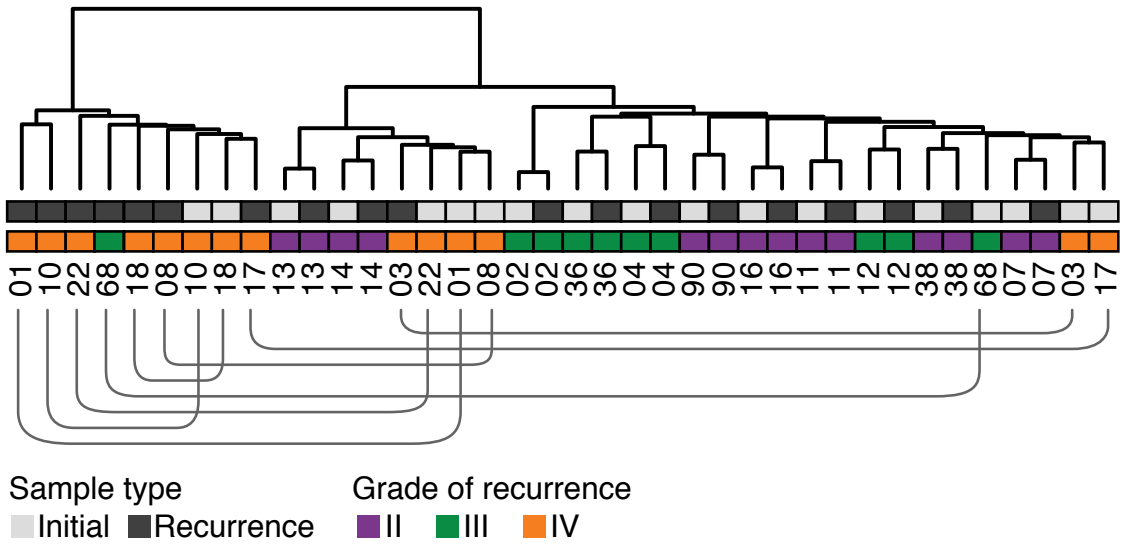


Figure 3.4. Unsupervised hierarchical clustering of the top 50% most variable CpG sites.
Annotations of sample type, grade of recurrence, and patient identification numbers are provided. The lines beneath the patient identification numbers connect initial and recurrent tumors from the same patient that are not adjacent to each other.

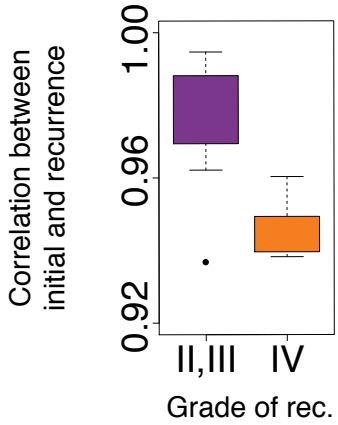


Figure 3.5. Boxplot showing the difference between the correlations of the low-grade and GBM groups.
Boxplot summarizing Pearson correlations of beta values between initial and recurrent tumors for each patient, grouped by the grade of the recurrent tumor, show decreased correlations in patients that recur as GBM.

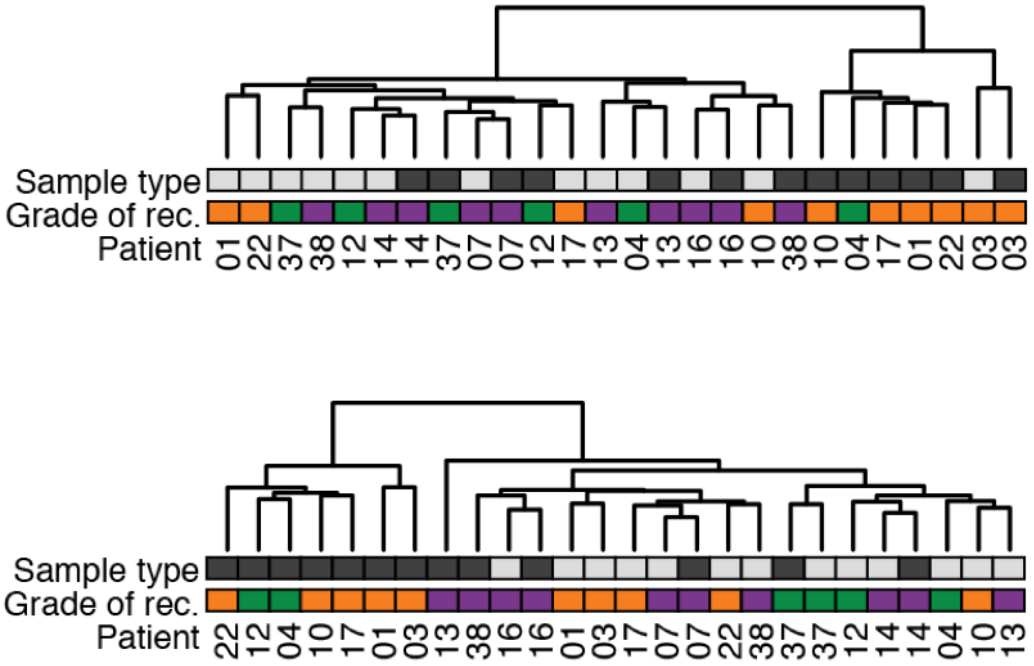


Figure 3.6. Clustering of the samples based on RNA-seq.
Unsupervised hierarchical clustering of the (F) top 1% or (G) top 50% most variably expressed genes across the cohort.

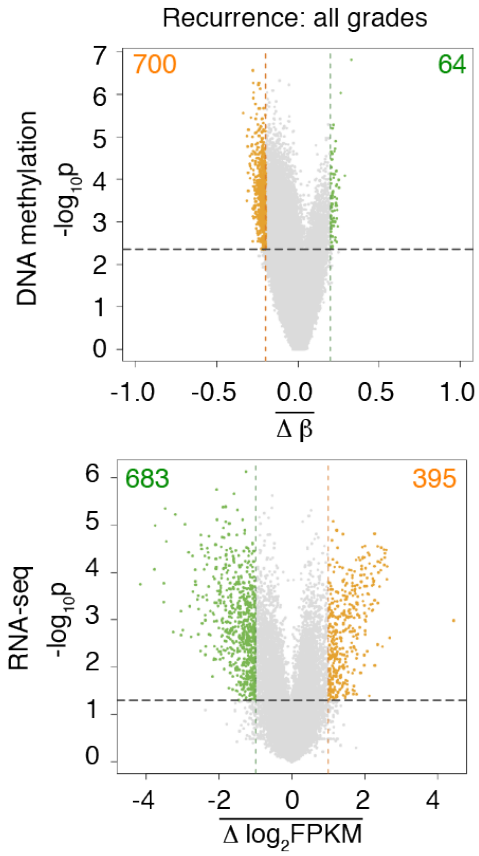


Figure 3.7. Volcano plots for changes across all tumors.
The methylation change (left) and expression change (right) from initial low-grade tumor to recurrence at each CpG site (left) and gene (right), averaged across all patients in the cohort. Colored dots represent CpG sites (left) and genes (right) that show significant changes at recurrence. The number of significant CpG sites (left) and genes (right) are provided at the top of each quadrant.

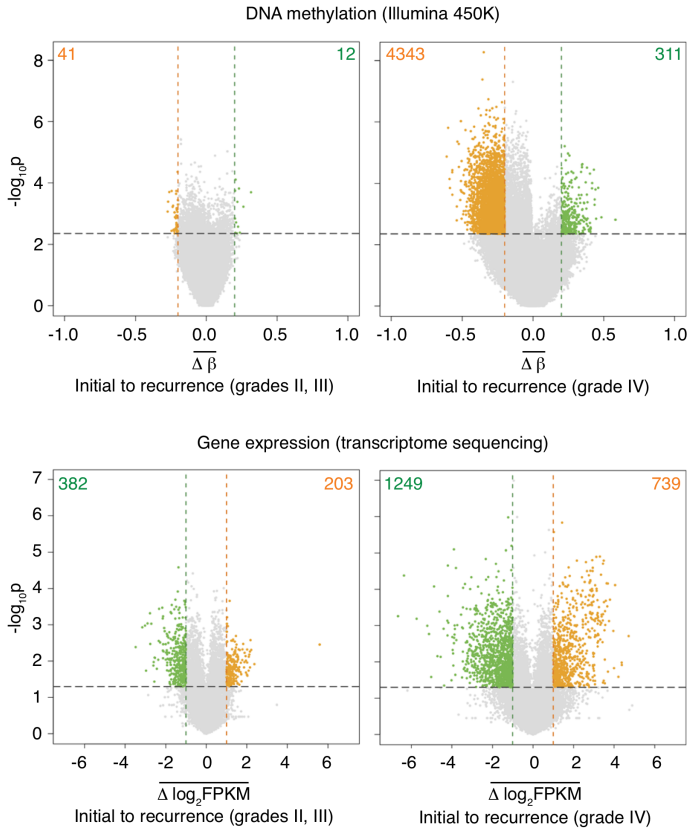


Figure 3.8. Volcano plots for changes separated by grade of recurrence.
Methylation (top) and expression (bottom) changes from initial to recurrent tumor, subdivided by the grade of the recurrent tumor. Tumors that recurred as GBMs show the strongest pattern of common methylation and expression changes.

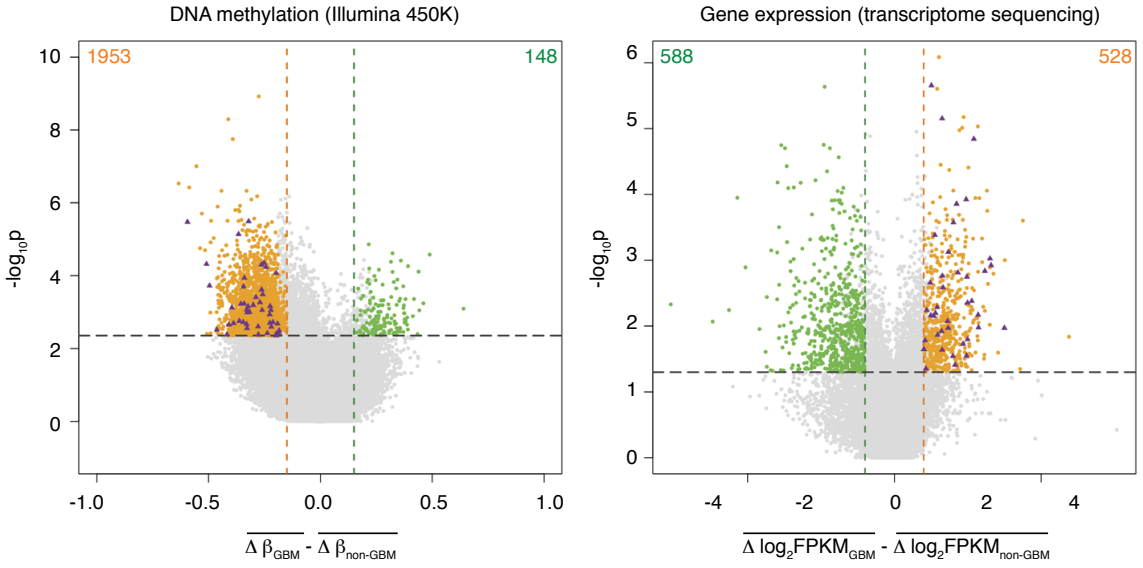


Figure 3.9. Volcano plots for methylation and transcriptome changes specific to malignant progression. Left panel shows a scatter plot of differences between GBM and non-GBM recurrent tumors in methylation changes from initial grade II to recurrent gliomas. Right panel shows an equivalent representation of differences in expression changes between GBM and non-GBM recurrent tumors. Colored points indicate significant differences. Purple triangles highlight genes that become hypomethylated at promoter CpGs (left) and over-expressed (right) during malignant progression to GBM.

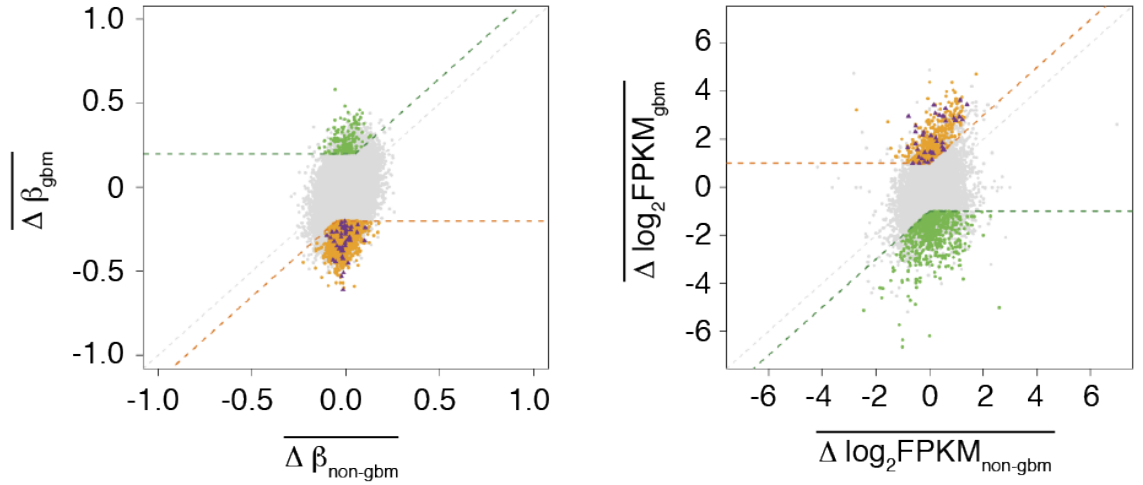


Figure 3.10. Scatterplots show how the average change from initial to recurrent tumor in methylation. (Left) and expression (Right) for each CpG site or gene differs between patients that recur as GBM (y-axis) and those that recur at grades II or III (x-axis). Purple triangles highlight genes that become hypomethylated at promoter CpGs (Left) and over- expressed (Right) during malignant progression to GBM.

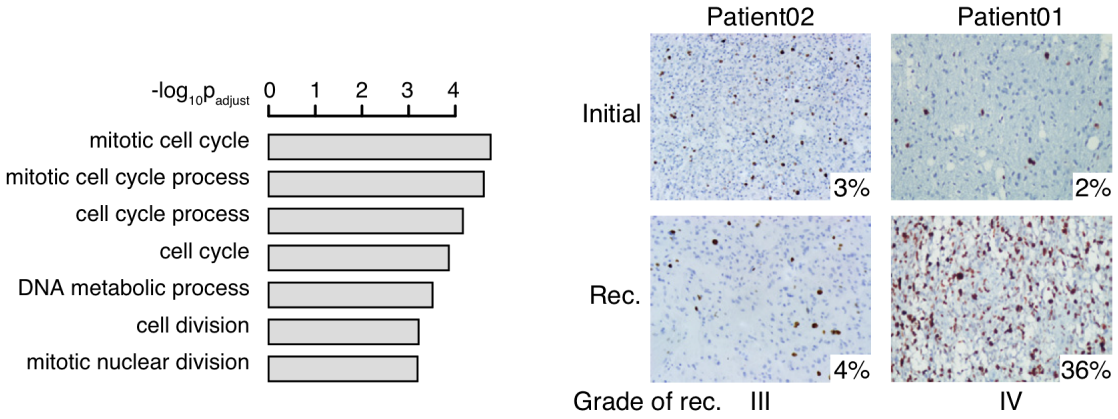


Figure 3.11. Enrichment of hypomethylated and upregulated cell cycle genes.
(Left) Barplot of the top results of a gene ontology analysis of genes that are both significantly hypomethylated and over-expressed specifically upon recurrence as GBM. (Right) Representative staining for Ki-67, a marker of actively cycling cells, in a patient that recurred at grade III (left) and a patient that recurred at grade IV (right).

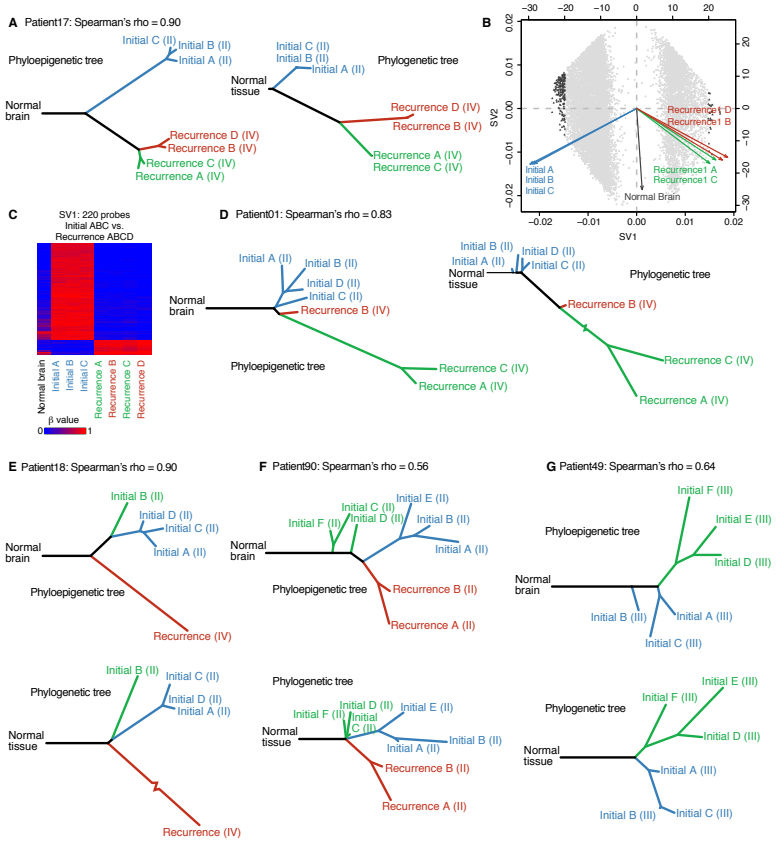


Figure 3.12. Phylogenetic and phyloepigenetic reconstruction of spatial and longitudinal samples of a tumor.
(A) A phyloepigenetic tree constructed from seven samples from Patient17 (left) replicates the structure derived from somatic mutations from exome sequencing of the same DNA samples (right). Tumor grade is provided in parentheses after each sample name.
(B) Singular value decomposition biplot shows the probes involved in separating tumor samples. Each probe used to build the phyloepigenetic tree in (A) is plotted (grey dots). The most highly weighted probes are highlighted (triangles). (C) A heatmap of the beta values at the 220 probes most highly weighted by SV1. (D) A phyloepigenetic tree (left) and a phylogenetic tree (right) were constructed to infer the evolutionary relationships within and between the initial and recurrent tumors of Patient01. Tumor grade is provided in

parentheses after each sample name. (E-G) Phyloepigenetic (top) and phylogenetic trees (bottom) for Patient 18 (E), Patient90 (F) and Patient49 (G) show similar evolutionary relationships. Tumor grade is provided in parentheses after each sample name.

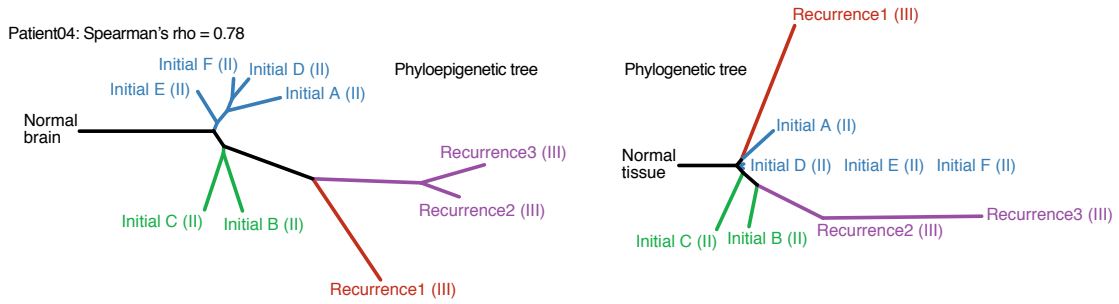


Figure 3.13. Phyloepigenetic (left) and phylogenetic (right) trees of Patient04 present evolutionary relations across four surgical time points.
Tumor grade is provided in parentheses after each sample name.

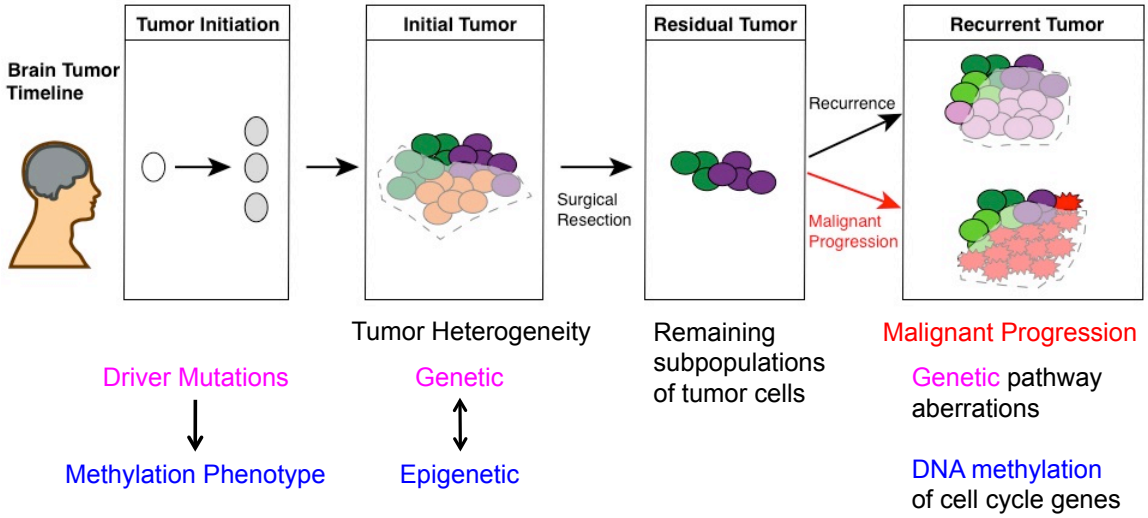


Figure 3.14. Integrated model of glioma evolution.
Low-grade gliomas exhibit intratumoral heterogeneity at initial presentation, with subclones that share the initiating genetic (*IDH1* followed by *TP53* and *ATRX* and copy number alterations, CNA) and epigenetic (*IDH1*-associated glioma CpG island methylator phenotype, G-CIMP) alterations, but further develop distinct genetic and epigenetic characteristics. Following surgical resection, the outgrowth from residual disease may be grade II or III, while still continuing to evolve subclones with genetic and co-dependent epigenetic features that are distinct from the initial tumor. In other patients, residual disease may undergo malignant progression to GBM, either spontaneously or as a consequence of treatment-associated mutations, in either case acquiring genetic defects in the RB and Akt-mTOR pathways and promoter hypomethylation and activation of cell cycle genes. Treatment associated progression to GBM is uniquely associated with an increased epigenetic silencing of MGMT (van Thuijl, 2015) and acquisition of genetic defects in mismatch repair genes.

TABLES

Patient	Gender	Age at diagnosis	Tumor sample	Diagnosis (WHO grade)	Surgical interval (months)	Non-surgical treatment (months)	Overall survival (months)	*IDH1* status	1p19q status	Illumina 405K array	RNA-seq
01	Male	28	Initial tumor	Astrocytoma (II)	31	TMZ (14)[b]	58	R132H	intact	yes	yes
			Recurrence	Glioblastoma (IV)				R132H	intact	yes	yes
02	Female	26	Initial tumor	Oligoastrocytoma (II)	74	None	79[c]	R132H	intact	yes	no
			Recurrence	Anaplastic astrocytoma (III)				R132H	intact	yes	no
03	Female	28	Initial tumor	Astrocytoma (II)	76	TMZ (7), TMZ (11)	85	R132H	intact	yes	yes
			Recurrence	Glioblastoma (IV)				R132H	intact	yes	yes
04	Male	22	Initial tumor	Astrocytoma (II)	15	None	61	R132C	intact	yes	yes
			Recurrence 1	Anaplastic astrocytoma (III)	20	TMZ (7)		R132C	intact	yes	yes
			Recurrence 2	Anaplastic astrocytoma (III)	9	TMZ (6)		R132C	intact	yes	no
			Recurrence 3	Anaplastic astrocytoma (III)				R132C	intact	yes	no
07	Male	30	Initial tumor	Astrocytoma (II)	105	XRT (1)	148	R132H	intact	yes	yes
			Recurrence	Astrocytoma (II)				R132H	intact	yes	yes
08	Male	44	Initial tumor	Oligoastrocytoma (II)	40	None	103	R132H	intact	yes	no
			Recurrence	Glioblastoma (IV)				R132H	intact	yes	no
10	Female	41	Initial tumor	Astrocytoma (II)	25	TMZ (9)	44	R132H	intact	yes	yes
			Recurrence	Glioblastoma (IV)				R132H	intact	yes	yes
11	Female	30	Initial tumor	Oligoastrocytoma (II)	132	XRT (1), TMZ (26)	186	R132H	intact	yes	no
			Recurrence	Oligoastrocytoma (II)				R132H	intact	yes	no
12	Male	35	Initial tumor	Astrocytoma (II)	17	None	82[d]	R132H	intact	yes	yes
			Recurrence	Anaplastic astrocytoma (III)				R132H	intact	yes	yes
13	Male	24	Initial tumor	Oligoastrocytoma (II)	21	None	106	R132G	intact	yes	yes
			Recurrence	Oligoastrocytoma (II)				R132G	intact	yes	yes
14	Male	25	Initial tumor	Astrocytoma (II)	30	None	149[d]	R132H	intact	yes	yes
			Recurrence	Astrocytoma (II)				R132H	intact	yes	yes
16	Female	35	Initial tumor	Astrocytoma (II)	5[a]	None	38	R132H	intact	yes	yes
			Recurrence	Astrocytoma (II)				R132H	intact	yes	yes
17	Male	27	Initial tumor	Oligodendroglioma (II)	30	TMZ (12)	59[d]	R132H	intact	yes	yes
			Recurrence	Glioblastoma (IV)				R132H	intact	yes	yes
18	Male	49	Initial tumor	Oligoastrocytoma (II)	94	TMZ (11)	106[d]	R132H	intact	yes	no
			Recurrence	Glioblastoma (IV)				R132H	intact	yes	no
22	Male	22	Initial tumor	Astrocytoma (II)	56	XRT (1)	70	R132H	intact	yes	yes
			Recurrence	Glioblastoma (IV)				R132H	intact	yes	yes
36	Female	31	Initial tumor	Astrocytoma (II)	71	None	73[d]	R132H	intact	yes	no
			Recurrence	Anaplastic astrocytoma (III)				R132H	intact	yes	no
37	Male	31	Initial tumor	Oligoastrocytoma (II)	57	None	105[d]	R132H	intact	no	yes
			Recurrence	Anaplastic oligoastrocytoma (III)				R132H	intact	no	yes
38	Female	21	Initial tumor	Astrocytoma (II)	20	None	25[d]	R132H	intact	yes	yes
			Recurrence	Astrocytoma (II)				R132H	intact	yes	yes
49	Male	23	Initial tumor	Anaplastic oligodendroglioma (III)			14[d]	R132H	codel	yes	no
68	Female	31	Initial tumor	Oligoastrocytoma (II)	17	TMZ (12)	23[d]	R132H	intact	yes	no
			Recurrence	Anaplastic oligoastrocytoma (III)				R132H	intact	yes	no
90	Female	39	Initial tumor	Oligodendroglioma (II)	22	None	64[d]	R132H	intact	yes	no
			Recurrence 1	Oligoastrocytoma (II)	34	None		R132H	intact	yes	no

a Recurrent surgery for residual disease, no evidence of radiographic progression
b Including a month each of TMZ plus either Accutane or Thalidomide
c Patient lost to follow-up
d Patient alive

Table 3.1. Summary of the data types acquired, clinical features, treatment history and molecular features of each tumor in the cohort.

BIBLIOGRAPHY

1. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
2. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
3. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
4. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.
5. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
6. Sterneckert, J.L., P. Reinhardt, and H.R. Scholer, *Investigating human disease using stem cell models.* Nat Rev Genet, 2014. **15**(9): p. 625-639.
7. Pouton, C.W. and J.M. Haynes, *Embryonic stem cells as a source of models for drug discovery.* Nat Rev Drug Discov, 2007. **6**(8): p. 605-616.
8. Liu, Y., et al., *Generation of functional organs from stem cells.* Cell Regeneration, 2013. **2**(1): p. 1-6.
9. Ng, H.H. and M.A. Surani, *The transcriptional and signalling networks of pluripotency.* Nat Cell Biol, 2011. **13**(5): p. 490-6.
10. Young, R.A., *Control of Embryonic Stem Cell State.* Cell, 2011. **144**(6): p. 940-954.
11. Watanabe, A., Y. Yamada, and S. Yamanaka, *Epigenetic regulation in pluripotent stem cells: a key to breaking the epigenetic barrier.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2013. **368**(1609): p. 20120292.
12. Ye, J. and R. Blelloch, *Regulation of Pluripotency by RNA Binding Proteins.* Cell stem cell, 2014. **15**(3): p. 271-280.
13. Greve, T.S., R.L. Judson, and R. Blelloch, *microRNA Control of Mouse and Human Pluripotent Stem Cell Behavior.* Annual review of cell and developmental biology, 2013. **29**: p. 213-239.
14. Flynn, R.A. and H.Y. Chang, *Long noncoding RNAs in cell fate programming and reprogramming.* Cell stem cell, 2014. **14**(6): p. 752-761.
15. Keene, J.D., *RNA regulons: coordination of post-transcriptional events.* Nat Rev Genet, 2007. **8**(7): p. 533-543.
16. Parchem, R.J., et al., *Two miRNA clusters reveal alternative paths in late-stage reprogramming.* Cell stem cell, 2014. **14**(5): p. 617-631.
17. Maronna, R.A. and R.H. Zamar, *Robust Estimates of Location and Dispersion for High-Dimensional Datasets.* Technometrics, 2002. **44**(4): p. 307-317.
18. Todorov, V. and P. Filzmoser, *An Object-Oriented Framework for Robust Multivariate Analysis.* 2009, 2009. **32**(3): p. 47.
19. Corthesy, B. and P.N. Kao, *Purification by DNA affinity chromatography of two polypeptides that contact the NF-AT DNA binding site in the interleukin 2 promoter.* J Biol Chem, 1994. **269**(32): p. 20682-90.

20. Baltz, A.G., et al., *The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts.* Mol Cell, 2012. **46**(5): p. 674-90.
21. Castello, A., et al., *Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.* Cell, 2012. **149**(6): p. 1393-406.
22. Kwon, S.C., et al., *The RNA-binding protein repertoire of embryonic stem cells.* Nat Struct Mol Biol, 2013. **20**(9): p. 1122-30.
23. Sakamoto, S., et al., *The NF90-NF45 complex functions as a negative regulator in the microRNA processing pathway.* Mol Cell Biol, 2009. **29**(13): p. 3754-69.
24. Singh, G., et al., *The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus.* Cell, 2012. **151**(4): p. 750-64.
25. Faye, M.D., et al., *Nucleotide composition of cellular internal ribosome entry sites defines dependence on NF45 and predicts a posttranscriptional mitotic regulon.* Mol Cell Biol, 2013. **33**(2): p. 307-18.
26. Graber, T.E., et al., *NF45 functions as an IRES trans-acting factor that is required for translation of cIAP1 during the unfolded protein response.* Cell Death Differ, 2010. **17**(4): p. 719-29.
27. Lee, J.W., et al., *Identification of hnRNPH1, NF45, and C14orf166 as novel host interacting partners of the mature hepatitis C virus core protein.* J Proteome Res, 2011. **10**(10): p. 4522-34.
28. Merrill, M.K. and M. Gromeier, *The double-stranded RNA binding protein 76:NF45 heterodimer inhibits translation initiation at the rhinovirus type 2 internal ribosome entry site.* J Virol, 2006. **80**(14): p. 6936-42.
29. Wang, J., et al., *A protein interaction network for pluripotency of embryonic stem cells.* Nature, 2006. **444**(7117): p. 364-8.
30. Lu, R., et al., *Systems-level dynamic analyses of fate change in murine embryonic stem cells.* Nature, 2009. **462**(7271): p. 358-62.
31. Ceccarelli, M., et al., *Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma.* Cell. **164**(3): p. 550-563.
32. Bray, N., et al. *Near-optimal RNA-Seq quantification*. ArXiv e-prints, 2015. **1505**.
33. Shen, S., et al., *MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data.* Nucleic Acids Res, 2012. **40**(8): p. e61.
34. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology, 2009. **10**(3): p. 1-10.
35. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biology, 2013. **14**(4): p. 1-13.
36. Harrow, J., et al., *GENCODE: The reference human genome annotation for The ENCODE Project.* Genome Research, 2012. **22**(9): p. 1760-1774.
37. Hegi , M.E., et al., *MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma.* New England Journal of Medicine, 2005. **352**(10): p. 997-1003.
38. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic Acids Research, 2014. **42**(D1): p. D68-D73.
39. Buaas, F.W., et al., *Cloning and characterization of the mouse interleukin enhancer binding factor 3 (Ilf3) homolog in a screen for RNA binding proteins.* Mamm Genome, 1999. **10**(5): p. 451-6.

40. Teodoridis, J.M., C. Hardie, and R. Brown, *CpG island methylator phenotype (CIMP) in cancer: causes and implications.* Cancer Lett, 2008. **268**(2): p. 177-86.
41. Hanahan, D. and Robert A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.
42. Cao, Y., et al., *Cancer research: past, present and future.* Nat Rev Cancer, 2011. **11**(10): p. 749-754.
43. Mazor, T., et al., *DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors.* Cancer Cell, 2015. **28**(3): p. 307-317.
44. Brocks, D., et al., *Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer.* Cell Reports, 2014. **8**(3): p. 798-806.
45. Johnson, B.E., et al., *Mutational Analysis Reveals the Origin and Therapy-driven Evolution of Recurrent Glioma.* Science (New York, N.Y.), 2014. **343**(6167): p. 189-193.
46. Gerlinger , M., et al., *Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing.* New England Journal of Medicine, 2012. **366**(10): p. 883-892.
47. Landau, D.A., et al., *Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia.* Cancer Cell, 2014. **26**(6): p. 813-25.
48. Landau, D.A., et al., *Evolution and impact of subclonal mutations in chronic lymphocytic leukemia.* Cell, 2013. **152**(4): p. 714-26.
49. Oakes, C.C., et al., *Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia.* Cancer Discovery, 2013.
50. Landan, G., et al., *Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues.* Nat Genet, 2012. **44**(11): p. 1207-14.
51. Aryee, M.J., et al., *DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases.* Sci Transl Med, 2013. **5**(169): p. 169ra10.
52. Pan, H., et al., *Epigenomic evolution in diffuse large B-cell lymphomas.* Nat Commun, 2015. **6**: p. 6921.
53. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306-13.
54. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science (New York, N.Y.), 1976. **194**(4260): p. 23-28.
55. Junttila, M.R. and F.J. de Sauvage, *Influence of tumour micro-environment heterogeneity on therapeutic response.* Nature, 2013. **501**(7467): p. 346-354.
56. Sharma, S.V., et al., *A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations.* Cell, 2010. **141**(1): p. 69-80.
57. Ostrom, Q.T., et al., *CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2006-2010.* Neuro-Oncology, 2013. **15**(suppl 2): p. ii1-ii56.
58. Louis, D.N., et al., *The 2007 WHO Classification of Tumours of the Central Nervous System.* Acta Neuropathologica, 2007. **114**(2): p. 97-109.

59. Bourne, T.D. and D. Schiff, *Update on molecular findings, management and outcome in low-grade gliomas.* Nat Rev Neurol, 2010. **6**(12): p. 695-701.
60. Sanai, N., S. Chang, and M.S. Berger, *Low-grade gliomas in adults.* J Neurosurg, 2011. **115**(5): p. 948-65.
61. Jakola, A.S., et al., *COmparison of a strategy favoring early surgical resection vs a strategy favoring watchful waiting in low-grade gliomas.* JAMA, 2012. **308**(18): p. 1881-1888.
62. Macdonald, D.R., L.E. Gaspar, and J.G. Cairncross, *Successful chemotherapy for newly diagnosed aggressive oligodendroglioma.* Annals of Neurology, 1990. **27**(5): p. 573-574.
63. Erdem-Eraslan, L., et al., *Intrinsic Molecular Subtypes of Glioma Are Prognostic and Predict Benefit From Adjuvant Procarbazine, Lomustine, and Vincristine Chemotherapy in Combination With Other Prognostic Factors in Anaplastic Oligodendroglial Brain Tumors: A Report From EORTC Study 26951.* Journal of Clinical Oncology, 2013. **31**(3): p. 328-336.
64. Shaw, E.G., et al., *Recurrence following neurosurgeon-determined gross-total resection of adult supratentorial low-grade glioma: results of a prospective clinical trial.* J Neurosurg, 2008. **109**(5): p. 835-41.
65. Chaichana, K.L., et al., *Recurrence and malignant degeneration after resection of adult hemispheric low-grade gliomas.* J Neurosurg, 2010. **112**(1): p. 10-7.
66. Ramakrishna, R., et al., *Outcomes in Reoperated Low-Grade Gliomas.* Neurosurgery, 2015. **77**(2): p. 175-84; discussion 184.
67. Network, T.C.G.A.R., *Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas.* New England Journal of Medicine, 2015. **372**(26): p. 2481-2498.
68. Vigneswaran, K., S. Neill, and C.G. Hadjipanayis, *Beyond the World Health Organization grading of infiltrating gliomas: advances in the molecular genetics of glioma classification.* Annals of Translational Medicine, 2015. **3**(7): p. 95.
69. van den Bent, M.J., et al., *Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951.* J Clin Oncol, 2013. **31**(3): p. 344-50.
70. Cairncross, G., et al., *Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402.* J Clin Oncol, 2013. **31**(3): p. 337-43.
71. Liu, X.Y., et al., *Frequent ATRX mutations and loss of expression in adult diffuse astrocytic tumors carrying IDH1/IDH2 and TP53 mutations.* Acta Neuropathol, 2012. **124**(5): p. 615-25.
72. Gladson, C.L., R.A. Prayson, and W. Liu, *The Pathobiology of Glioma Tumors.* Annual review of pathology, 2010. **5**: p. 33-50.
73. Louis, D.N., *Molecular pathology of malignant gliomas.* Annu Rev Pathol, 2006. **1**: p. 97-117.
74. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.

75. van Thuijl, H.F., et al., *Evolution of DNA repair defects during malignant progression of low-grade gliomas after temozolomide treatment.* Acta Neuropathol, 2015. **129**(4): p. 597-607.
76. Egger, G., et al., *Epigenetics in human disease and prospects for epigenetic therapy.* Nature, 2004. **429**(6990): p. 457-463.
77. Gupta, P.B., et al., *Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells.* Cell, 2011. **146**(4): p. 633-644.
78. Kreso, A., et al., *Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer.* Science, 2013. **339**(6119): p. 543-548.
79. Cara, S. and I.F. Tannock, *Retreatment of patients with the same chemotherapy: implications for clinical mechanisms of drug resistance.* Ann Oncol, 2001. **12**(1): p. 23-7.
80. Fang, F., et al., *The novel, small-molecule DNA methylation inhibitor SGI-110 as an ovarian cancer chemosensitizer.* Clin Cancer Res, 2014. **20**(24): p. 6504-16.
81. Matei, D.E. and K.P. Nephew, *Epigenetic therapies for chemoresensitization of epithelial ovarian cancer.* Gynecol Oncol, 2010. **116**(2): p. 195-201.
82. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer.* Nat Rev Genet, 2002. **3**(6): p. 415-28.
83. Robertson, K.D., *DNA methylation and chromatin - unraveling the tangled web.* Oncogene, 2002. **21**(35): p. 5361-79.
84. Feinberg, A.P. and B. Vogelstein, *Hypomethylation distinguishes genes of some human cancers from their normal counterparts.* Nature, 1983. **301**(5895): p. 89-92.
85. Yin, D., et al., *DNA repair gene O6-methylguanine-DNA methyltransferase: promoter hypermethylation associated with decreased expression and G:C to A:T mutations of p53 in brain tumors.* Mol Carcinog, 2003. **36**(1): p. 23-31.
86. Fouse, S.D. and J.F. Costello, *Epigenetics of neurological cancers.* Future Oncol, 2009. **5**(10): p. 1615-29.
87. Baeza, N., et al., *PTEN methylation and expression in glioblastomas.* Acta Neuropathol, 2003. **106**(5): p. 479-85.
88. Costello, J.F., et al., *Silencing of p16/CDKN2 expression in human gliomas by methylation and chromatin condensation.* Cancer Res, 1996. **56**(10): p. 2405-10.
89. Kim, T.Y., et al., *Epigenomic profiling reveals novel and frequent targets of aberrant DNA methylation-mediated silencing in malignant glioma.* Cancer Res, 2006. **66**(15): p. 7490-501.
90. Nagarajan, R.P., et al., *Recurrent epimutations activate gene body promoters in primary glioblastoma.* Genome Research, 2014.
91. Nakamura, M., et al., *Promoter hypermethylation of the RB1 gene in glioblastomas.* Lab. Invest., 2001. **81**(1): p. 77-82.
92. Wiencke, J.K., et al., *Methylation of the PTEN promoter defines low-grade gliomas and secondary glioblastoma.* Neuro-Oncology, 2007. **9**(3): p. 271-279.
93. Lai, A., et al., *Evidence for sequenced molecular evolution of IDH1 mutant glioblastoma from a distinct cell of origin.* J Clin Oncol, 2011. **29**(34): p. 4482-90.

94. Watanabe, T., et al., *IDH1 Mutations Are Early Events in the Development of Astrocytomas and Oligodendrogliomas.* The American Journal of Pathology, 2009. **174**(4): p. 1149-1153.
95. Turcan, S., et al., *IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype.* Nature, 2012. **483**(7390): p. 479-83.
96. Noushmehr, H., et al., *Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma.* Cancer Cell, 2010. **17**(5): p. 510-22.
97. Toyota, M., et al., *CpG island methylator phenotype in colorectal cancer.* Proceedings of the National Academy of Sciences, 1999. **96**(15): p. 8681-8686.
98. Hill, V.K., et al., *Stability of the CpG island methylator phenotype during glioma progression and identification of methylated loci in secondary glioblastomas.* BMC Cancer, 2014. **14**: p. 506.
99. Jiao, Y., et al., *Frequent ATRX, CIC, FUBP1 and IDH1 mutations refine the classification of malignant gliomas.* Oncotarget, 2012. **3**(7): p. 709-22.
100. Banine, F., et al., *SWI/SNF chromatin-remodeling factors induce changes in DNA methylation to promote transcriptional activation.* Cancer Res, 2005. **65**(9): p. 3542-7.
101. Gibbons, R.J., et al., *Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation.* Nat Genet, 2000. **24**(4): p. 368-71.
102. Costello, J.F., et al., *Methylation-related chromatin structure is associated with exclusion of transcription factors from and suppressed expression of the O-6-methylguanine DNA methyltransferase gene in human glioma cell lines.* Mol Cell Biol, 1994. **14**(10): p. 6515-21.
103. Wick, W., et al., *Prognostic or predictive value of MGMT promoter methylation in gliomas depends on IDH1 mutation.* Neurology, 2013. **81**(17): p. 1515-22.
104. Everhard, S., et al., *MGMT methylation: a marker of response to temozolomide in low-grade gliomas.* Ann Neurol, 2006. **60**(6): p. 740-3.
105. Kesari, S., et al., *Phase II study of protracted daily temozolomide for low-grade gliomas in adults.* Clin Cancer Res, 2009. **15**(1): p. 330-7.
106. Taal, W., et al., *First-line temozolomide chemotherapy in progressive low-grade astrocytomas after radiotherapy: molecular characteristics in relation to response.* Neuro Oncol, 2011. **13**(2): p. 235-41.
107. Triche, T.J., Jr., et al., *Low-level processing of Illumina Infinium DNA Methylation BeadArrays.* Nucleic Acids Res, 2013. **41**(7): p. e90.
108. Price, M.E., et al., *Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array.* Epigenetics Chromatin, 2013. **6**(1): p. 4.
109. Laffaire, J., et al., *Methylation profiling identifies 2 groups of gliomas according to their tumorigenesis.* Neuro Oncol, 2011. **13**(1): p. 84-98.
110. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
111. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.

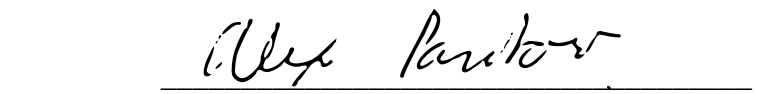
112. Sproul, D., et al., *Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns.* Genome Biol, 2012. **13**(10): p. R84.
113. Witte, T., C. Plass, and C. Gerhauser, *Pan-cancer patterns of DNA methylation.* Genome Medicine, 2014. **6**(8): p. 66.
114. Issa, J.P., et al., *Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon.* Nat Genet, 1994. **7**(4): p. 536-40.
115. Kerkel, K., et al., *Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation.* Nat Genet, 2008. **40**(7): p. 904-8.
116. Lu, C., et al., *IDH mutation impairs histone demethylation and results in a block to cell differentiation.* Nature, 2012. **483**(7390): p. 474-8.
117. Desper, R. and O. Gascuel, *Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle.* J Comput Biol, 2002. **9**(5): p. 687-705.
118. Paradis, E., J. Claude, and K. Strimmer, *APE: Analyses of Phylogenetics and Evolution in R language.* Bioinformatics, 2004. **20**(2): p. 289-90.
119. Shi, J., et al., *Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue.* Nat Commun, 2014. **5**: p. 3365.
120. Heyn, H., et al., *DNA methylation contributes to natural human variation.* Genome Res, 2013. **23**(9): p. 1363-72.
121. Rufini, A., et al., *p73 in Cancer.* Genes Cancer, 2011. **2**(4): p. 491-502.
122. Chen, H.Z., S.Y. Tsai, and G. Leone, *Emerging roles of E2Fs in cancer: an exit from cell cycle control.* Nat Rev Cancer, 2009. **9**(11): p. 785-97.
123. Lui, J.H., et al., *Radial glia require PDGFD-PDGFR[bgr] signalling in human but not mouse neocortex.* Nature, 2014. **515**(7526): p. 264-268.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

03/28/2016

Date