Spring 2019

An Integrative and Comparative Analysis of Transcriptome and Targetome Data of Medulloblastoma

Blaine Nelson

AN INTEGRATIVE AND COMPARATIVE ANALYSIS OF TRANSCRIPTOME AND TARGETOME DATA OF MEDULLOBLASTOMA

By

Blaine Nelson

A Thesis Submitted in Partial Fulfillment
Of the Requirements for the
University Honors Program

Department of Biology
The University of South Dakota
May 2019

The members of the Honors Thesis Committee appointed
to examine the thesis of Blaine Nelson
find it satisfactory and recommend that it be accepted.

_____

Dr. Dan Van Peursem
Mathematical Sciences Department Chair
Professor of Mathematical Sciences
Director of the Committee

_____

Dr. Erliang Zeng
Associate Professor, Division of Biostatistics and Computational Biology
University of Iowa

_____

Beate Wone
Instructor of Biology

ABSTRACT

An Integrative and Comparative Analysis of Transcriptome and Targetome Data of Medulloblastoma

Blaine Nelson

Director: Dan Van Peursem, Ph.D.

Medulloblastoma (MB) arises in the cerebellum and is the most common brain tumor seen in the field of pediatrics. Primary and recurrent MBs are often found to contain deregulated *Atonal Homolog 1* (*ATOH1*) expression among SHH/PTCH signals. Therefore, mice models were generated for research by inducing expression of the Atoh1 transgene in the cerebellum of *Ptch1*$^{+/-}$ mice. The overexpression of the Atoh1 transgene in the animals transform the non-metastatic brain tumor to a metastatic tumor that disseminates to the spinal cord and other parts of the brain. In order to understand the molecular and cellular events involved in the cascade of metastatic MB, statistical analysis of the transcriptome and targetome were applied. RNA-Sequencing was run first to generate a common list of shared differentially expressed genes and then followed by the addition of chromatin immunoprecipitation sequencing. From the data obtained, pathway analysis was applied. The data from the mice were then subject to comparison to a cohort of human data on MB to further investigate the similarities and differences in the biological causes for the formation of the disease.

Das Medulloblastom entstammt im Kleinhirn und ist der häufigste pädiatrische Gehirntumor. Es wird häufig festgestellt, dass primäre und rezidivierende Medulloblastome deregulierte *atonale Homolog 1* (*ATOH1*)-Expression unter SHH-PTCH-Signalen enthalten. Darum wurden Mäusemodelle in der Forschung erstellt, indem die Expression des Atoh1-Transgens im Kleinhirn von *Ptch1*$^{+/-}$ Mäusen induziert wurde. Die Überexpression dieses Transgens in den Tieren wandelt den gutartigen Gehirntumor in einen metastatischen Tumor um, der sich auf das Rückenmark und andere Teile des Gehirns verbreitet. Um die molekularen und zellulären Ereignisse nachzuvollziehen, die an der Kaskade metastatisches Medulloblastoms beteiligt sind, wurden statistische Analysen des Transkriptoms und des Targetoms durchgeführt. Die RNA-Sequenzierung wurde zuerst durchgeführt, um eine gemeinsame Liste von differentiell exprimierten Genen zu erstellen, gefolgt von dem Zusatz der Chromatin-Immunopräzipitationssequenzierung. Von den erhaltenen Daten wurde eine Weganalyse durchgeführt. Die Daten der Mäuse wurden dann einem Vergleich mit einer Kohorte menschlicher Daten zum MB unterzogen, um die Ähnlichkeiten und Unterschiede in den biologischen Ursachen für die Entstehung der Krankheit weiter zu untersuchen.

KEYWORDS: Medulloblastoma, RNA-Sequencing, genetics, immunoprecipitation

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This project was funded by the University of South Dakota Undergraduate Research Excellence Award, Sanford Research, NIH Centers of Biomedical Research Excellence (COBRE), and the South Dakota Biomedical Research Infrastructure Network (SD BRIN). Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103443.

An Integrative and Comparative Analysis of Transcriptome and Targetome Data of Medulloblastoma

Blaine A. Nelson [1,2], Hasitha Premathilake [1,2], Alex Heglin [1,2], Katie B. Grausam [6,7], Haotian Zhao [8], and Erliang Zeng [3,4,5]

[1] Department of Biology, University of South Dakota, Vermillion, SD 57069
[2] Bioinformatics and Computational Systems Biology Laboratory (BioComs Lab), University of South Dakota, Vermillion, SD 57069
[3] Division of Biostatistics and Computational Biology, [4] Iowa Institute for Oral Health Research, College of Dentistry, University of Iowa, Iowa City, IA 52242
[5] Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, IA 52242
[6] Sanford Research, Sioux Falls, SD 57104
[7] Division of Basic Biomedical Sciences, Sanford School of Medicine, University of South Dakota, Vermillion, SD 57069
[8] Department of Biomedical Sciences, New York Institute of Technology College of Osteopathic Medicine, Old Westbury, NY 11568

I. Introduction

Medulloblastoma (MB) constitutes nearly 20% of all childhood brain tumors making it the most common pediatric brain malignancy. MB stems from the posterior fossa, a region of the brain located at the base of the skull, or more specifically within the cerebellum, which is the part of the posterior fossa controlling coordination and balance (St. Jude, 2019). The malignant tumor disseminates through the cerebrospinal fluid (CSF) to the area of the meninges and subarachnoid space sheathing the brain and spinal cord termed the leptomeninges. Leptomeningeal metastases, which are often discovered at the time of diagnosis or during recurrence, are associated with a poor clinical prognosis. In order to treat the tumor, an aggressive treatment plan including the use of surgery, craniospinal radiation, chemotherapy, or a combination of the aforementioned is used to resect or destroy the malignancy; however, in spite of the aggressive treatments, improvement in the survival of patients with the metastatic

disease has progressed slowly, which may be due to the expanse and intricacy of the condition. This cancerous tumor is further categorized into one of four molecular subgroups, each of which are based on different types of gene mutations and are distinguished from one another by their aberrations, transcriptional profiles, and clinical outcomes. These subgroups that compose the entirety of medulloblastoma include: wingless (WNT), Sonic Hedgehog (SHH), Group 3, and Group 4; however, little is known about the last two groups. While each subgroup of medulloblastoma is distinct from one another, the extent of the effects on the cellular mechanisms resulting in medulloblastoma formation is not clear. It is important to explore the molecular and cellular events involved in the ATOH1-driven cascade of metastatic MB so that potential therapeutics may be safer and more effective, assuring that there will be no detrimental effects on the developing brain.

II. Methods

II.1 Overview

A previous study performed by Sanford Research has shown that Atonal homolog 1 (ATOH1) not only plays a vital role in normal development of the cerebellum, but also plays a critical role for murine models in the initiation and progression of MB in the SHH subgroup. The research demonstrated an acceleration of MB development in mice with the *protein patched homolog 1* gene (*Ptch1*$^{+/-}$) when Atoh1 expression was induced, transforming the benign tumors into highly metastatic tumors. Further research for the transgenic overexpression of Atoh1 in different mice strains has been conducted by Sanford Research to understand the role of Atoh1 in leptomeningeal dissemination and metastasis.

Dr. Haotian Zhao, a former researcher in the Sanford Health's Research Center at Sanford Research now located at the New York Institute of Technology, provided the gene

expression profiles with a set of 22,557 genes for statistical analysis of the mice models. There were three strains of mice taken into account for this project in order to investigate the molecular cascade of MB. These strains include medulloblastomas from transgenic mice with an overexpression of Atoh1 and *Ptch1*$^{+/-}$ and thus present with a metastatic tumor (these mice are designated with a **T**), mice with overexpressed *Ptch1*$^{+/-}$ only (these mice are designated with a **P**), and the control group, or wild-type mice (designated with a **C**). Each strain of mice had three replicates, ensuring validity and accounting for error that may have arisen during a trial.

In this study, functional genomic statistical analyses of RNA-Sequencing (RNA-Seq) and chromatin immunoprecipitation sequencing (ChIP-Seq) data from the obtained gene expressions values from Sanford were run and applied to uncover differentially expressed genes (DEGs) of the MB SHH subgroup in the mice models. By doing such procedures, we are able to unveil the cellular and molecular mechanisms involved in MB formation. RNA-Seq is biomedically relevant as this sequencing is used to interpret the function of the genome as well as to understand how the disease develops at the level of gene expression, otherwise known as the transcriptome. Meanwhile, ChIP-Seq was applied to reveal the DNA binding sites of the transcription factors and ultimately the gene regulation events. ChIP-Seq allowed us to assess the targetome, or all the microRNA targets of an organism. The results of the targetome and transcriptome were compared and integrated, and enriched pathways were detected and analyzed using Ingenuity Pathway Analysis (QIAGEN®, 2017).

We applied multiple RNA-Seq statistical analysis tools using R to identify the significant DEGs from the data. The tools used include baySeq (Hardcastle et al, 2010), Cuffdiff (Trapnell, 2017), DESeq (Anders et al, 2010), edgeR (McCarthy et al, 2012), limma (Ritchie et al, 2015), and PoissonSeq (Li, 2011).

II.2 baySeq

The Bioconductor package baySeq uses empirical Bayesian methods to identify differential expression in high-throughput data. This package is able to find the identification for differential expression by calculating estimated differential expression posterior likelihoods (Hardcastle, 2009). This method assumes a negative binomial distribution and estimates priors. BaySeq begins by defining the set of data in terms of similarities and differences between the samples and their replicates. For a given set of data, baySeq then seeks to analyze which samples behave similar to one another and which sets of samples behave identifiably different. Because baySeq uses numerical methods with an empirical Bayesian approach, this allowed the real data to be retained and the library size to be used as a scaling factor. The library size can be defined as the total number of mapped reads during a run of data (Hardcastle et al, 2010). As a package, baySeq is not intended for use with normalized data; therefore, raw count data were used as the input. The samples were then paired up with one another [ex: the control group with the overexpressed *Ptch1*$^{+/-}$ group (CP), the control group with the mice who have the metastatic tumor (CT), and the overexpressed *Ptch1*$^{+/-}$ group with the mice who have the metastatic tumor (TP)] in order to investigate the differences and similarities between gene expression. Each replicate of the same sample [ex: C1, C2, C3] share the same set of underlying parameters, but the sets of parameters between two different samples are not identical, allowing for pairwise comparison.

II.3 Cuffdiff

Not only does Cuffdiff find differently expressed genes and transcripts through substantial changes in the expression of the transcript, splicing, and promoter use, but the program also finds genes being differentially regulated at the level of transcription. The program

identifies genes that are differentially regulated at the transcriptional level by grouping transcripts into groups of biological meaning. In order to find which genes or transcripts are differentially expressed, Cuffdiff tests the expression of an observed log-fold change against that of the null hypothesis with no change; however, Cuffdiff must also measure the significance of comparison because an observed change that is nonzero may occur on the behalf of variability within both technical and cross-replicate biological aspects though the gene or transcript is not actually differentially expressed (Trapnell, 2014). This package takes SAM files that contain data from two or more samples as the input. By accepting and analyzing the data from two or more biological conditions, Cuffdiff aids in the exploration of transcriptional regulation under differing conditions (Ghosch et al, 2016).

II.4 DESeq

DESeq is another Bioconductor package. This package is able to estimate variance-mean dependence using raw count data from high-throughput sequencing assays. Like baySeq, DESeq also uses a negative binomial distribution to test for differential expressions (Anders et al, 2010); however, while DESeq does make the assumption of negative binomial distribution, the package also adds an assumption that there is a local linear relationship between the mean expression levels and an over-dispersion of the data (Hardcastle et al, 2010). With DESeq, digital gene expression analysis is performed on raw read counts, not transformed or normalized data for sequencing depth. If anything other than raw read counts is used, nonsensical results may occur; therefore, raw count data were used as the input. Comparisons between the different conditions were run using the respective codes and then analyzed.

II.5 edgeR

Another tool that was utilized in this project to discover DEGs was edgeR, which is described as an empirical analysis of DGE in R. This program was designed for analyzing expression data of replicated counts. EdgeR is able to find changes between two or more groups by implementing a large quantity of statistical methodologies as long as one of the groups has a phenotypical or experimental condition that has been replicated (Robinson et al, 2010). These methodologies are based on a negative binomial distribution and include but are not limited to an empirical Bayes estimation, exact tests, generalized linear models, and quasi-likelihood tests. The quasi-likelihood tests account for the uncertainty in the dispersion estimation and thus, give this package a stronger and firmer control on error rates (Chen et al, 2019). Currently, pairwise comparisons are supported by edgeR to test for differential expression. We therefore had to specify which two groups we were going to compare at a time, though the end result was still a comparison between CP, CT, and TP. When keyed into the statistical tool, rows of the data corresponded to genes, while the columns corresponded to the independent libraries, or the replicates. Just like the other packages, raw read counts were used as the original input.

II.6 limma

As a Bioconductor software tool, limma analyzes data from a variety of platforms. These include experiments that involve microarrays, protein arrays, and high-throughput polymerase chain reaction (PCR). Rather than breaking the treatments down individually and then making piece by piece comparisons between pairs of samples, limma analyzes experiments on an integrated whole level using linear models (Ritchie et al, 2015). This approach is useful as the technique provides us the ability to model correlations that may exist between samples with differences in their transcriptome. Limma also has the unique ability to incorporate quantitative

weights into all its levels of analysis. Power to detect differently expressed genes is increased with the usage of weights. RNA-seq read counts are able to be analyzed through limma with high precision. This works through the function voom, which converts the read counts that have been processed to the log-scale, thus estimating the mean-variance relationship in an empirical fashion (Ritchie et al, 2015). Because this software package accepts RNA-seq data in the form of a matrix and thus operates on a matrix of expression values, data was input with the rows accounting for the genomic features, otherwise known as genes, and the columns corresponding to RNA samples. However, if any problems arise with how data is plugged in, it is possible for limma to accept results, specifically the DGE list, from edgeR so that the analysis may be properly run.

II.7 PoissonSeq

PoissonSeq is an R package for RNA-sequencing data, implementing statistical methods like normalization, estimation of false discovery rate, and testing to recall a list of significant genes as well as a list for the possibility of a false discovery. Like the binomial distribution, the Poisson distribution is discrete; however, the distribution only has a single parameter, which is composed of both the mean and the variance. The Poisson distribution gives a good approximation to binomial distribution when the sample size, *n*, is larger (Larget, 2005). In general, the usage of PoissonSeq can be quite useful in data analysis as it can be used not only for data with two types of outcomes, but also for data with multiple-class outcomes. Doing so provides a more complete sense of comparison. PoissonSeq uses the Poisson goodness-of-fit test to estimate sequence depth (Li, 2011). Sequencing depth, which not only characterizes the importance of inference data founded on sequencing data but also serves as a scaling factor between experiments, may be estimated through the implementation of PoissonSeq.

II.8 Statistical Analysis of the Differential Expressed Genes

Each RNA-Seq statistical tool resulted in their own respectable selective number of DEGs. The lists of DEGs from each software were then integrated into new data sets that were then analyzed with R to compute the q-values. Originally, p-values were obtained for comparison; however, because each software package contain different normalization factors that need to be taken into consideration, the p-values were converted to q-values to justly compare.

Often in statistical analysis, there will be the possibility of obtaining a false positive due of chance. The false discovery rate (FDR) refers to a proportion of false positives that are expected among the hypotheses of significance, or the likelihood of a gene to be deemed falsely positive among the entire pool of significant genes (Nonlinear Dynamics, n.d). Q-values use an FDR method to adjust p-values, resulting in fewer false positives. For example, a q-value of 0.01 suggests that 1% of significant tests will actually result in a false positive. Therefore, q-values were used because fewer false positives result with q-values.

The goal of differential analysis is to figure out what could be involved in the biological process of interest by finding compounds and molecular events that show a great quality of difference between experimental groups. Therefore, a q-value of 0.05 was used as the cut-off. Later, a cut-off q-value of 0.01 was used in order to be more selective.

II.9 Visualization and Ingenuity Pathway Analysis

Once the q-values were set, the lists of DEGs used for the comparison were finalized. The identified DEGS from each analytical tool were then subject to comparison using InteractiVenn (Heberle et al, 2015). Comparisons were made between overexpressed Atoh1

induced mice and the control, the expression of *Ptch1*$^{+/-}$ mice and the control, as well as the two groups of mice with the tumors (Atoh1 vs. *Ptch1*$^{+/-}$). Using this program results in the number and the identity of DEGs shared between each RNA-Seq tool and between different conditions, thus narrowing in on what genes, or as equally important, what pathways may be affected in the grand scheme of metastatic tumor formation of medulloblastoma with the overexpression of Atoh1 or *Ptch1*$^{+/-}$.

In order to analyze, integrate, and interpret the omics data, Ingenuity Pathway Analysis (IPA) was applied (QIAGEN®, 2017). Using IPA allowed us to return gene pathways that were significantly enriched within each group of DEGs. IPA, through powerful analysis, identifies targets or biomarkers of biological systems and reveals the significance of the data. The program is able to do so through algorithms that discover mechanisms, functions, and pathways that are relevant to those changes being observed in an analyzed dataset. IPA also makes use of BioProfiler, a component identifying potential therapeutic targets by surfacing molecules relevant to the disease of interest or the phenotype of interest (QIAGEN®, 2017). The IPA analysis provides a greater interpretation of the impact DEGs have on phenotypic effects.

II.10 Comparative Analysis of Metastatic Medulloblastoma between *Mus musculus* and *Homo sapiens*

Comparative genomics allows researchers to compare the genomic features of different organisms to one another. In this project, comparative analysis was performed between tumor samples obtained from humans and mice as a way to explore the potential affected genes that may be similar or different among the different subgroups. A large cohort of primary medulloblastoma samples that was uploaded to the National Center for Biotechnology Information (NCBI) website by The Hospital for Sick Children in Toronto, Canada was obtained

from NCBI under Accession number GSE85212 for a comparative genomic analysis (Cavalli et al, 2017). Comparative genomics allowed us to study the biological similarities and differences between humans and mice and can be used as a way to understand the underlying causes of disease formation. This cohort dataset consists of a total of 763 samples comprised of the four distinct subgroups of medulloblastoma: WNT, SHH, Group 3, and Group 4. The samples were divided into their respective subtypes and an ANOVA test was used to identify the significant differentially expressed genes. After the DEGs of each subgroup were returned, the genes were integrated for comparison to study the degree of overlap between the tumor subgroups and later the overlap between mice and humans.

III. Results

III.1 Comparisons of DEGs returned by the statistical tools

Statistical analysis is important to understanding how phenotypes are affected by molecular and cellular mechanisms as the analysis is able to detect DEGs between different conditions. As mentioned in section II, six methods of statistical methods for RNA-Seq data were applied, each resulting in their respectable amount of DEGs identified. The DEGs from each package were then placed into the InteractiVenn program to be compared to one another on the basis of their experimental conditions (Heberle, 2015).

We have created Venn diagrams representing the shared DEGs between the transgenic Atoh1 model and the control (Figure 1, Figure 4), between the *Ptch1*$^{+/-}$ model and the control (Figure 2, Figure 5), and between the transgenic Atoh1 model and the *Ptch1*$^{+/-}$ model (Figure 3, Figure 6). As we have previously known, there are similarities and differences between the two tumor groups and the control group. Observing the Venn diagrams confirms our thoughts even more as they show a great number of DEGs were identified between the Atoh1 model and the

control, the *Ptch1*$^{+/-}$ model and the control, but relatively few DEGs between the *Ptch1*$^{+/-}$ model and the Atoh1 model. This information suggests that metastatic tumors and primary tumors are similar to each other for they retain similar gene expression profiles

Originally, DEGS were identified using a q-value of less than 0.05; however, the value was decreased to a cut-off of 0.01 in order to be more precise.

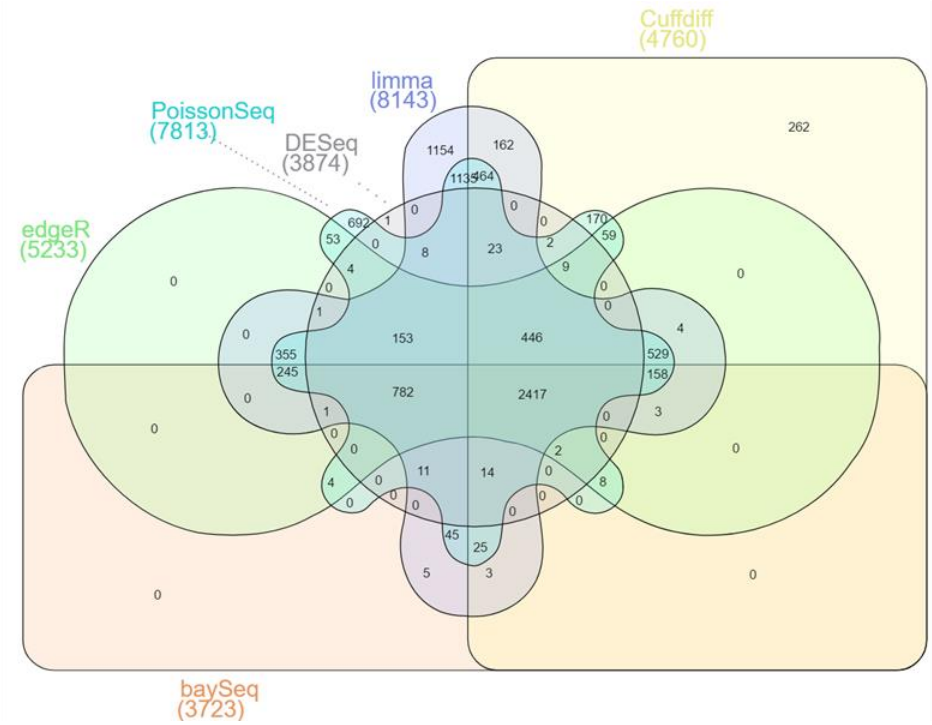

Figure 1: Venn diagram presenting the comparison of DEGs between transgenic mice models Atoh1 and wild-type mice. DEGs were obtained from the previously mentioned six RNA-Seq statistical analysis tools. Results are from those DEGs with a q-value<0.05.

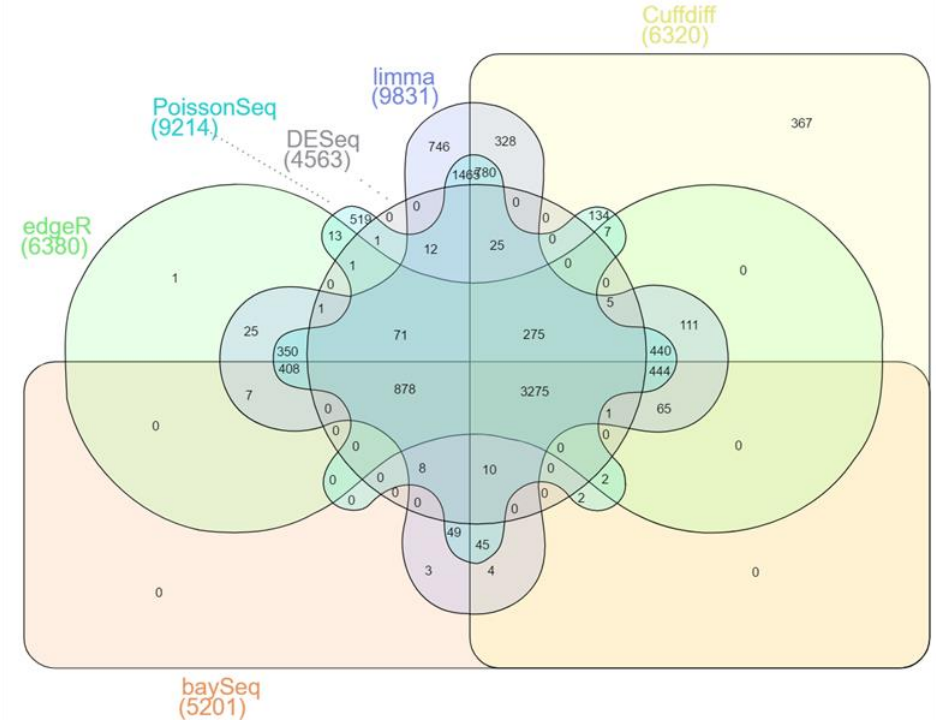

Figure 2: Venn diagram showing the comparison of DEGs between transgenic mice models *Ptch1*$^{+/-}$ and wild-type mice. DEGs were obtained from the previously mentioned six RNA-Seq statistical analysis tools. Results are from those DEGs with a q-value<0.05.

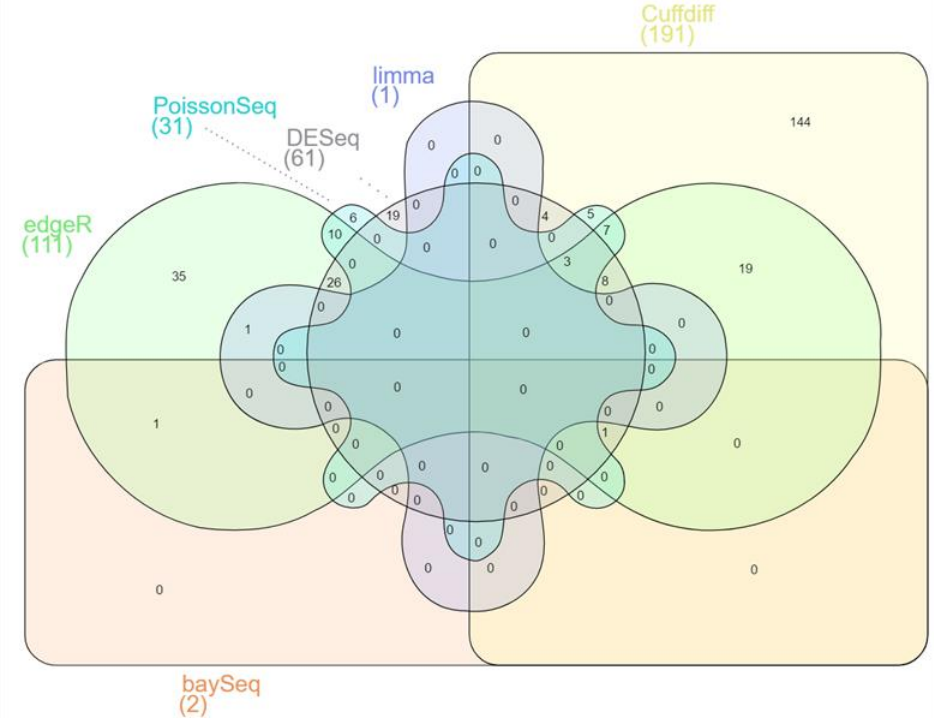

Figure 3: Venn diagram displaying the comparison of DEGs between transgenic mice models Atoh1 and *Ptch1*$^{+/-}$. DEGs were obtained from the previously mentioned six RNA-Seq statistical analysis tools. Results are from those DEGs with a q-value<0.05.

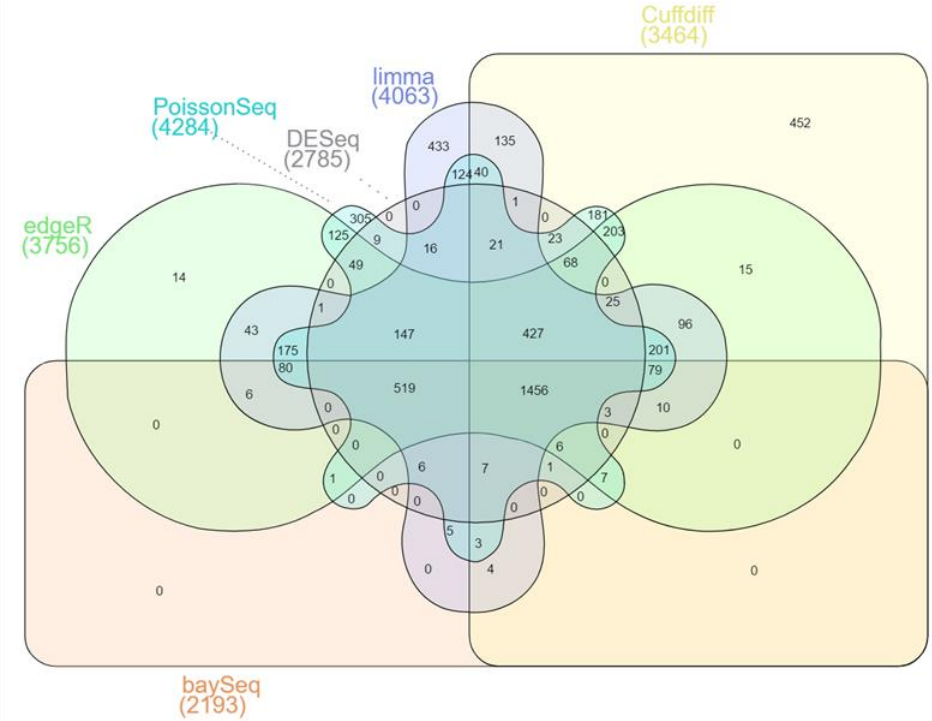

Figure 4: Venn diagram presenting the comparison of DEGs between transgenic mice models Atoh1 and wild-type mice. DEGs were obtained from the previously mentioned six RNA-Seq statistical analysis tools. Results are from those DEGs with a q-value<0.01.

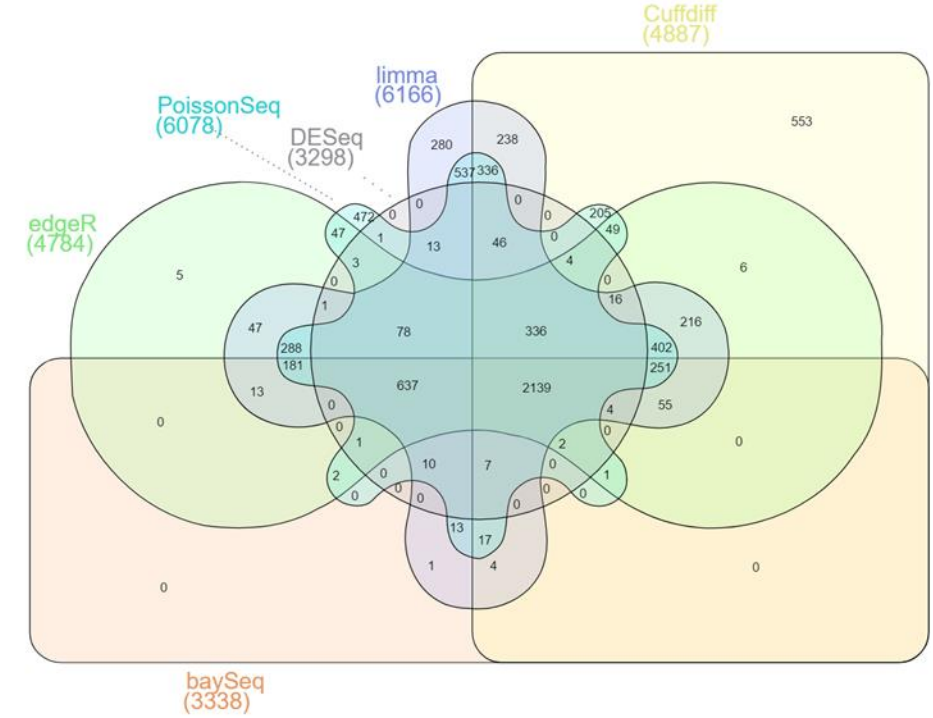

Figure 5: Venn diagram showing the comparison of DEGs between transgenic mice models *Ptch1*$^{+/-}$ and wild-type mice. DEGs were obtained from the previously mentioned six RNA-Seq statistical analysis tools. Results are from those DEGs with a q-value<0.01.

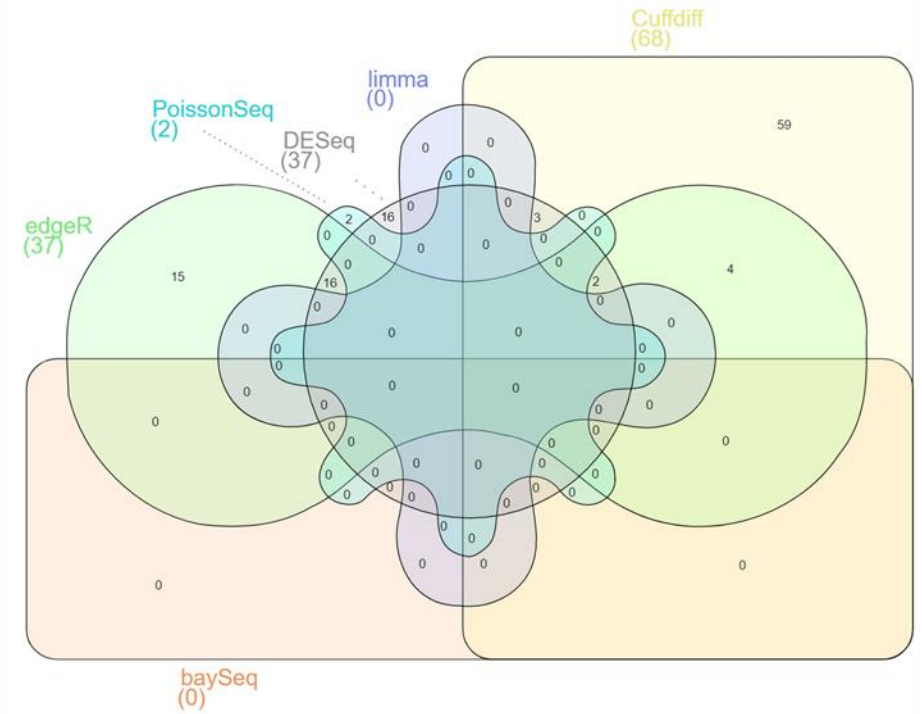

Figure 6: Venn diagram displaying the comparison of DEGs between transgenic mice models Atoh1 and *Ptch1*$^{+/-}$. DEGs were obtained from the previously mentioned six RNA-Seq statistical analysis tools. Results are from those DEGs with a q-value<0.01.

III.2 Implementation of ChIP-Seq and the comparison of DEGs between the transcriptome and targetome

For this study, we desired not only to understand how the transcriptome is affected in MB, but also how the tumor affects the targetome. Therefore, ChIP-Seq was applied. ChIP-Seq captures DNA targets for transcription factors across the entire genome of any organism and reveals gene regulatory networks with RNA sequencing (Illumina, 2015). After ChIP-Seq was run and significant genes were found, the observed data were placed into a comparison with the data from the transcriptome, originating from the RNA-Seq analyses previously conducted, in order to better understand the cascade of the cancer.

Primarily, the data obtained for ChIP-Seq were organized by ATOH1 targets in the metastatic tumor and the ATOH1 targets in the primary tumor and thus, were compared to one another to get a general sense of what ChIP-Seq originally discovered (Figure 7). Through this

comparison, we found 1446 DEGs were shared between the two tumors. ChIP-Seq data, as mentioned above, was then incorporated into the comparison from the DEGs of RNA-Seq. In order to carry out this procedure, two of the RNA-Seq packages were dropped using the false discovery rate calculated by the number of unique genes of the package divided by the total number of genes, or $\frac{U}{T}$. By calculating the FDR of each package using that expression, limma and Cuffdiff were exempted from the analysis with the ChIP-Seq as they provided the highest number of false positives. When ChIP-Seq data was added, it resulted similarly to the outcomes from the transcriptome analysis. Many genes were significant between the metastatic tumor and the control (Figure 9), the primary tumor and the control (Figure 8), while few genes were found between the two tumors (Figures 10-12).

Figure 7: Venn diagram displaying the comparison between two sets of ATOH1 associated genes in the metastatic tumor group and the primary tumor group of medulloblastoma. The genes used were the results from ChIP-Seq.

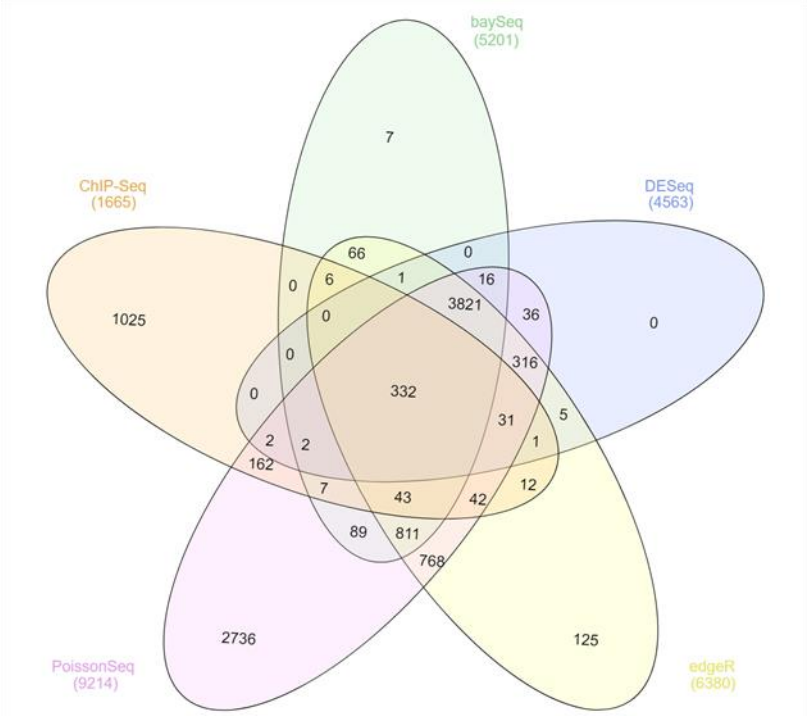

Figure 8: Venn diagram showing the comparison of DEGs from RNA-Seq (involved with the transcriptome) and of ChIP-Seq (involved with the targetome) between mice models *Ptch1*$^{+/-}$ and the control.

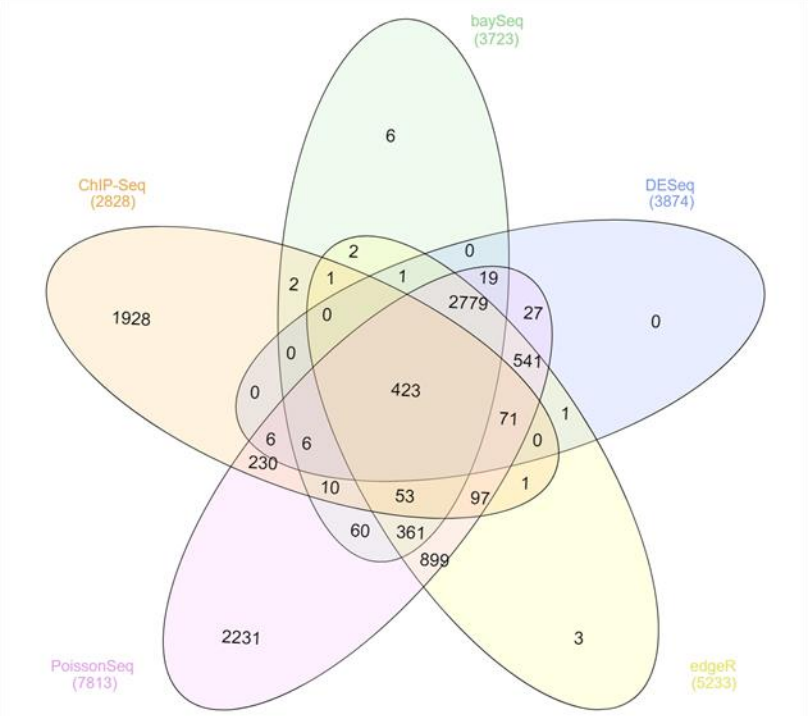

Figure 9: Venn diagram presenting the comparison of DEGs from RNA-Seq (involved with the transcriptome) and of ChIP-Seq (involved with the targetome) between transgenic mice models Atoh1 and the control.

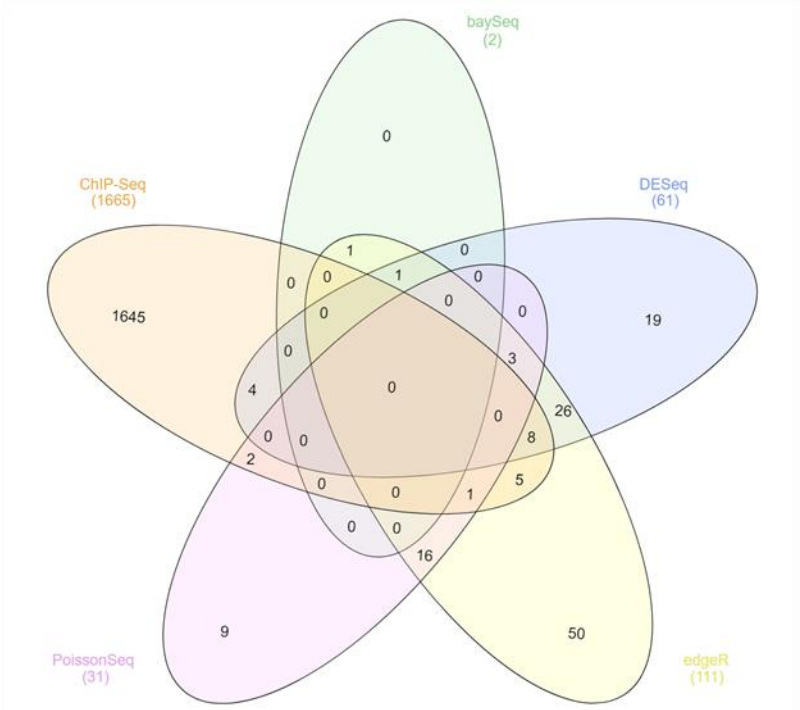

Figure 10: Venn diagram showing the comparison of DEGs from RNA-Seq (involved with the transcriptome) and of ChIP-Seq (involved with the targetome) between transgenic mice models Atoh1 and *Ptch1*$^{+/-}$. The ChIP-Seq data used was that of the primary tumor.

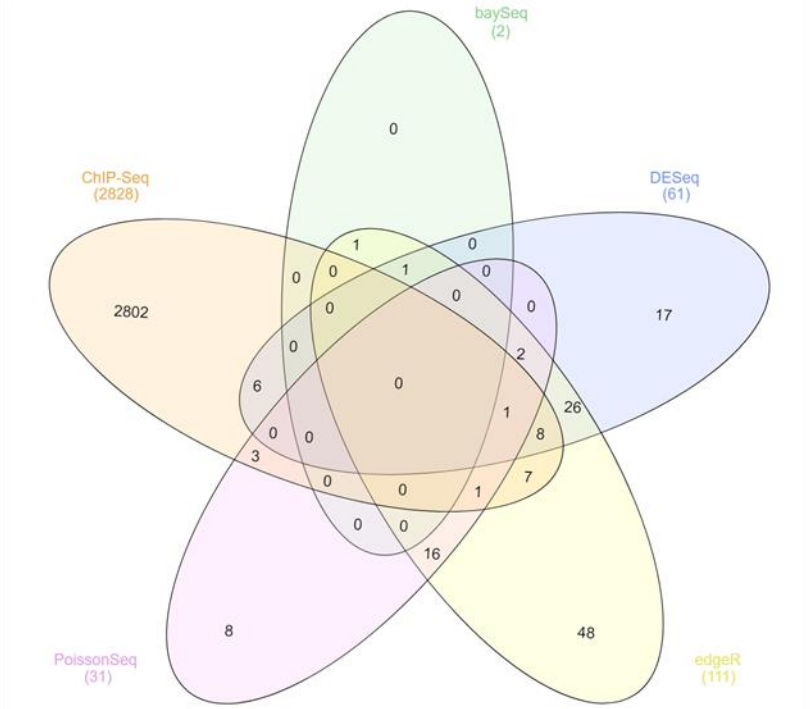

Figure 11: Venn diagram showing the comparison of DEGs from RNA-Seq (involved with the transcriptome) and of ChIP-Seq (involved with the targetome) between transgenic mice models Atoh1 and *Ptch1*$^{+/-}$. The ChIP-Seq data used was that of the metastatic tumor.

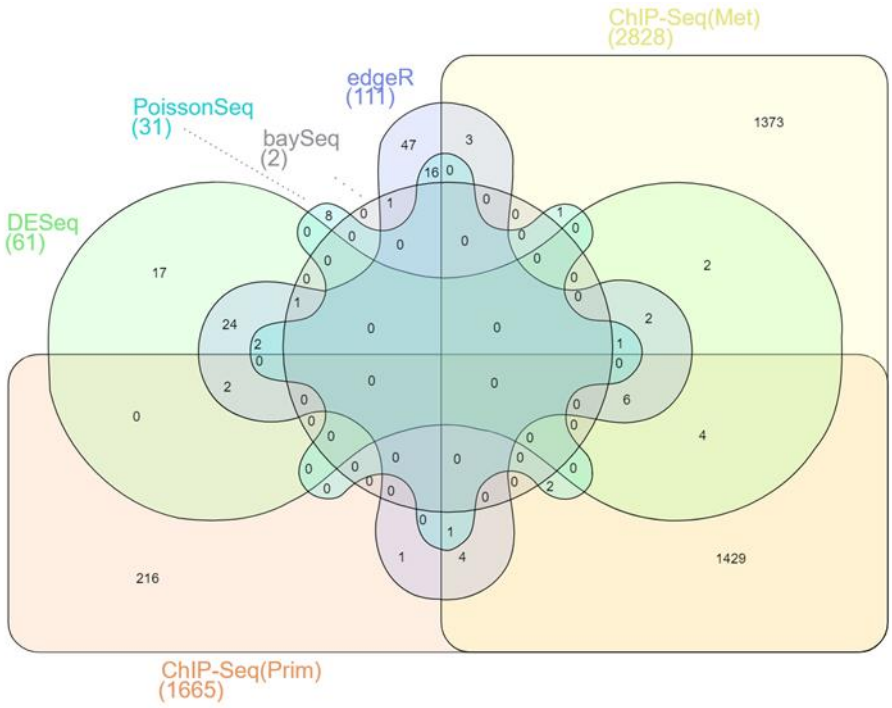

Figure 12: Venn diagram displaying shared DEGs from the Transcriptome and Targetome from both the primary tumor and the metastatic tumor among different tools.

III.3 Comparison of Transcriptome and Targetome between the metastatic tumor and the primary tumor with Pathway Analysis

To understand the genetic effects medulloblastoma has, it is critical to run an analysis between the tumors using both transcriptome and targetome data. Doing so provides a general overview of the genes commonly affected as well as the function of such genes as pathways become compromised. Here, we propose comparing the primary tumor vs. the control and the metastatic tumor vs. the control with one another using the results from the analysis of the transcriptome (RNA-Seq) and the results from the analysis of the targetome (ChIP-Seq) that were obtained (Figure 13). A comparison using different sets of DEGs obtained from different tools provides true differentially expressed genes. From each tool used in this analysis, we also compared the pathways found to be enriched by IPA (QIAGEN®, 2017). Comparing enriched pathways provides information on the overlap and differences between the two tumors and is

biologically meaningful as enriched pathways signify biological functions or processes (Figure 14). Because biological information is provided by pathways, interpretation of the overlap of enriched pathways provides insight to the molecular cascade of the cancer. Genes and pathways can be incorporated into a disease and developmental function table signifying their roles in the development of the disease (Figure 15). From Figure 14, 202 enriched pathways were identified among the transcriptome and targetome of the two tumors.

Using Figures 13 and 14, we suggest that gene set analysis through IPA has more strengths to discover hypotheses deemed to be biologically significant. This is indicated in Table 1 as the number of EPs is significantly reduced compared to the number of DEGs.

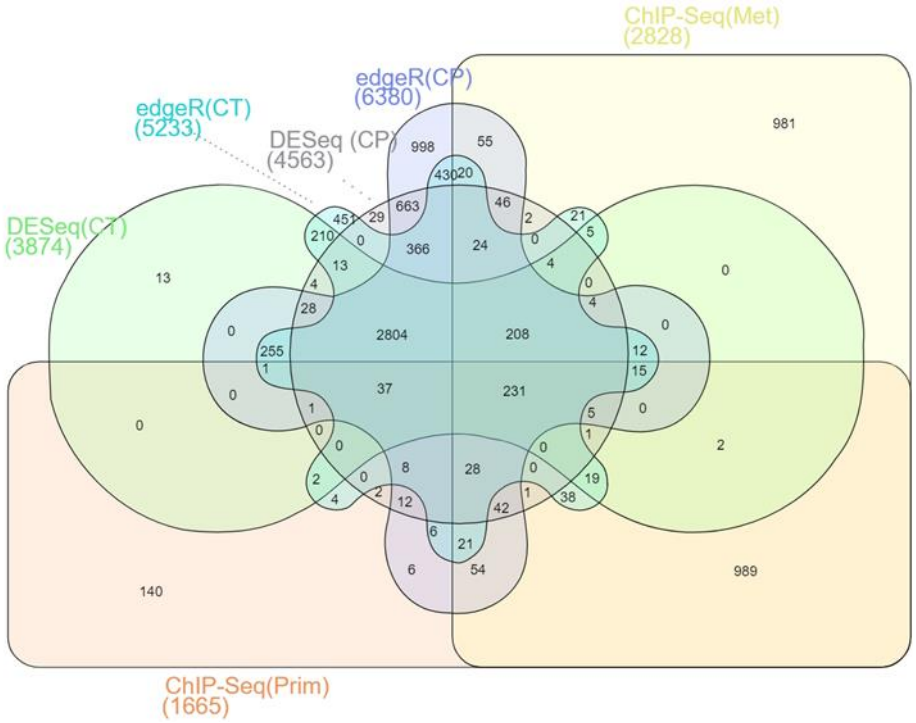

Figure 13: Venn diagram representing the difference between the primary and metastatic tumors. DEGs identified in the Transcriptome and Targetome of transgenic mice model *Ptch1*$^{+/-}$ vs. the control and transgenic mouse model Atoh1 vs. the control. Only edgeR, DESeq, and ChIP-Seq were used.

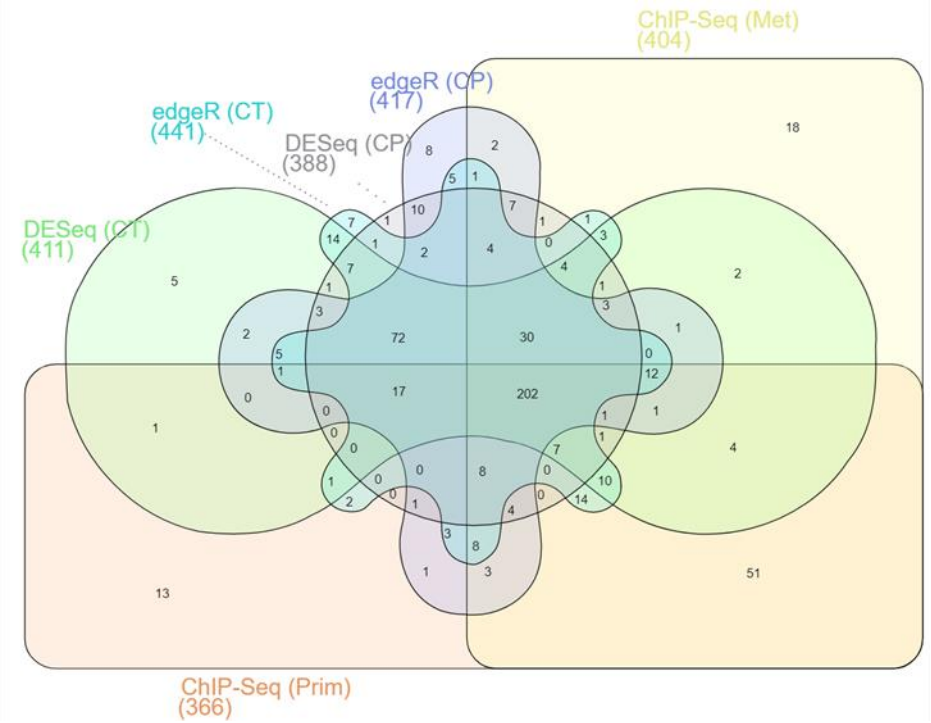

Figure 14: Venn diagram presenting the shared enriched pathways between three tools in both transgenic mouse model *Ptch1*$^{+/-}$ vs. the control and transgenic mouse model Atoh1 vs. the control.

	# of DEGs	# of Unique DEGs	Ratio of Unique DEGs (%)	# of EPs	# of Unique EPs	Ratio of Unique EPs (%)
ChIP-Seq (Primary)	1665	140	8.41	366	13	3.55
ChIP-Seq (Metastatic)	2828	981	34.69	404	18	4.66
DESeq (Primary)	4563	29	0.64	388	1	0.26
DESeq (Metastatic)	3874	13	0.34	411	5	1.22
edgeR (Primary)	6380	998	15.64	417	8	1.92
edgeR (Metastatic)	5233	451	8.62	441	7	1.59

Table 1: Unique DEGs and enriched pathways (EPs) in Primary and Metastatic Tumors models compared to the control. The DEGs and the EPs were observed among three statistical analysis tools.

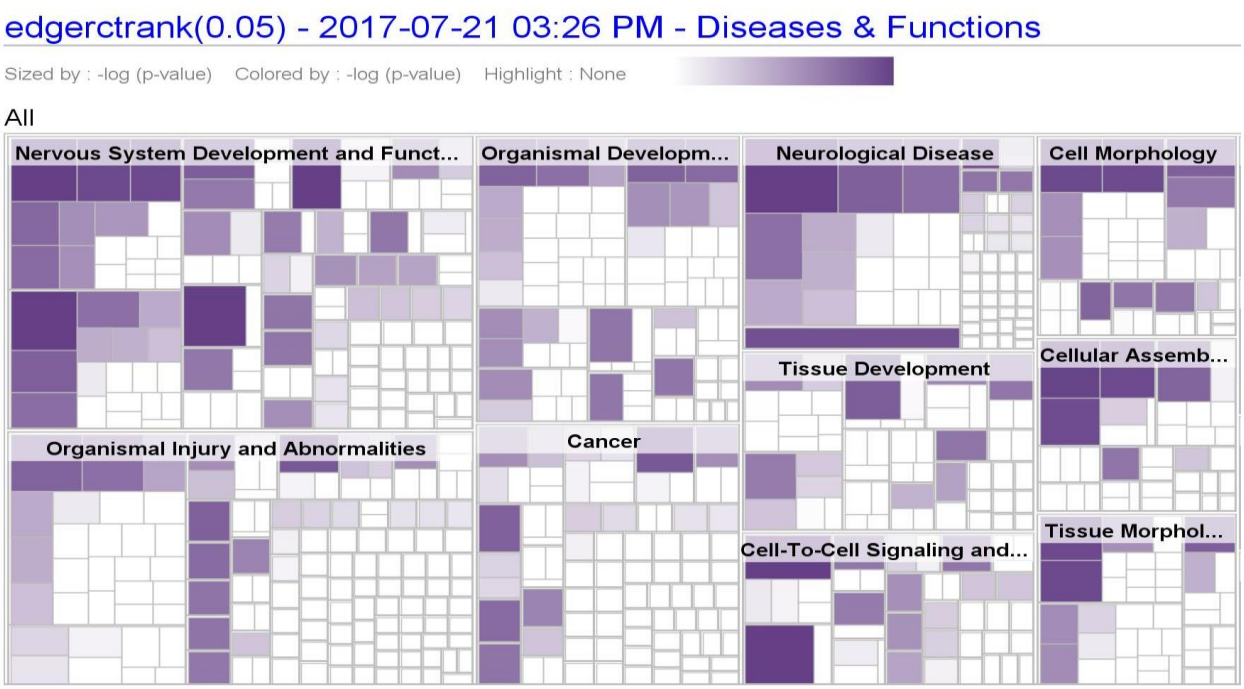

Figure 15: A Disease and Developmental Function table affected by DEGs identified by edgeR in MB Metastatic Tumor vs. the control. The darker violet represents more gene involvement.

III.4 Comparative Genomic Analysis of Medulloblastoma for *Homo sapiens* and *Mus musculus*

Because MB is complex with its separation into subgroups, each subgroup has to be taken into consideration in order to properly understand the molecular and cellular events that are occurring and that are affected by overexpression. An ANOVA analysis was performed on the data obtained for each subgroup to acquire the respective differentially expressed genes. For each tumor group, the top 1000 genes (based on the relative q-values) were selected for comparison to the rest of the groups (Figure 16).

In the first evaluation, the results of each subgroup from the human primary tumors were compared to one another and results indicated zero intercorrelated genes, suggesting heterogeneity as characterized by each of the subgroups having discrete somatic aberrations, activated pathways, and clinical outcomes. The SHH *Mus musculus* data was then integrated

into the comparison for a second evaluation to figure out any possible overlap between *Mus musculus* and *Homo sapiens* (Figure 17). Although there was overlap between three or four of the five types included in this study, there were zero DEGs common among all five.

We then decided to observe the similarities and differences of the DEGs in *Homo sapiens* and *Mus musculus* categorized under the SHH group. Looking at Figure 18, 142 similar DEGs were uncovered. These genes were loaded into IPA to reveal the enriched pathways (Figure 19), which can provide reference for future research.

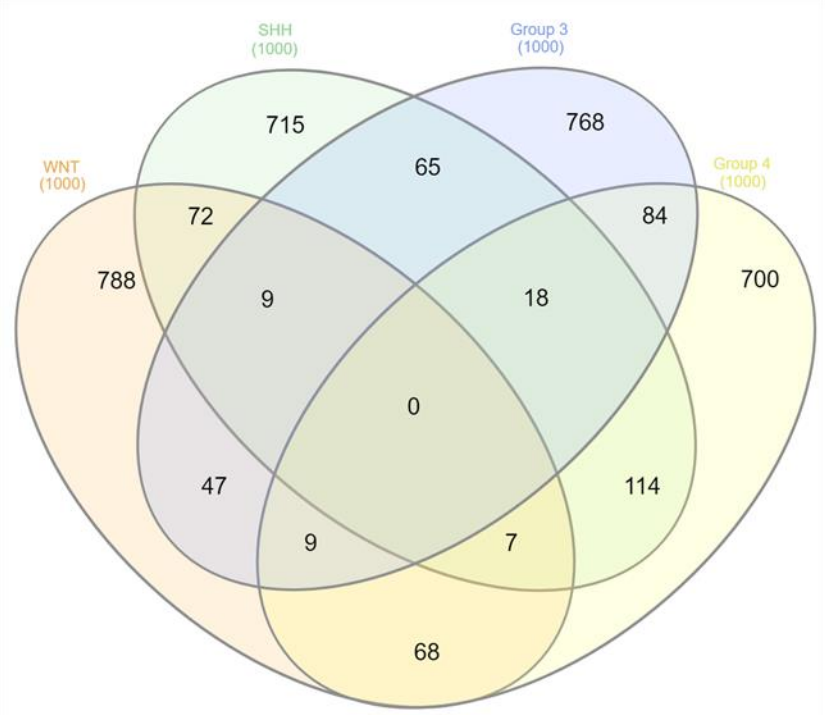

Figure 16: Venn diagram presenting the top 1000 returned medulloblastoma DEGS within the different subgroups.

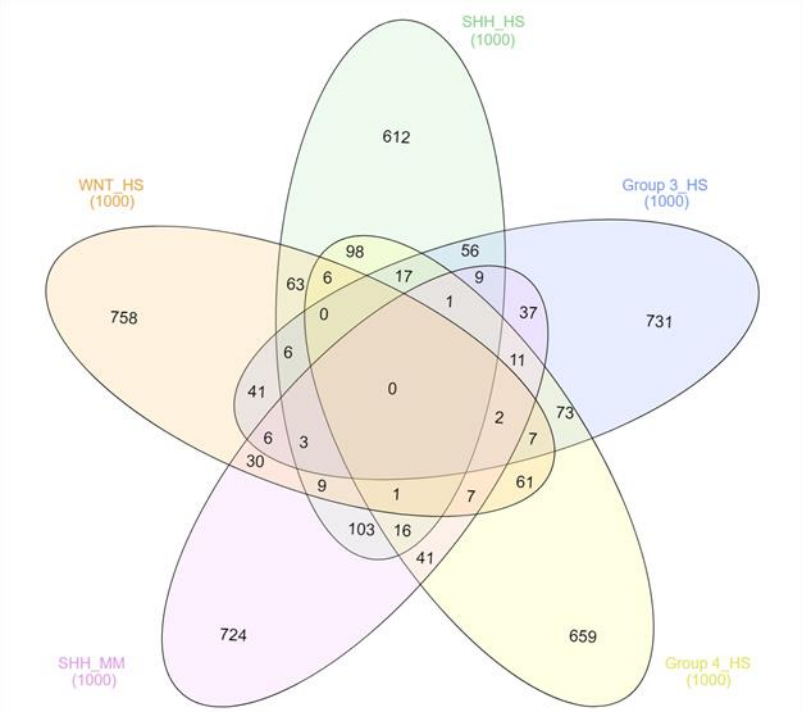

Figure 17: Venn diagram presenting the top 1000 returned medulloblastoma DEGS within the different subgroups in human data with the addition of mouse data.

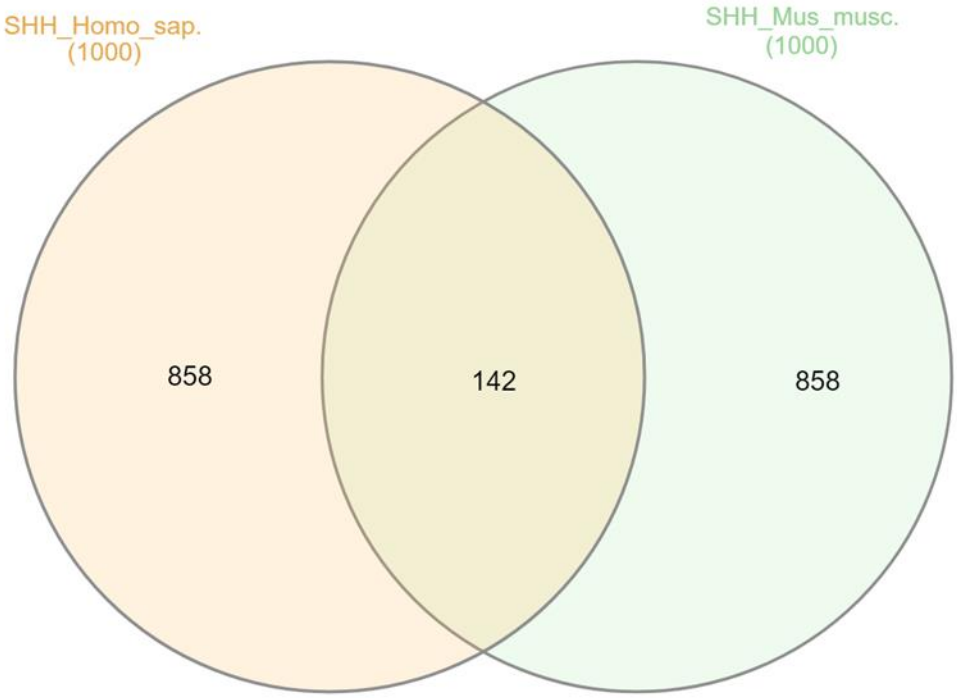

Figure 18: Venn diagram showcasing the Top 1000 returned DEGs in the SHH subgroup of medulloblastoma found in *Homo sapiens* and in the SHH subgroup of medulloblastoma found in *Mus musculus.*

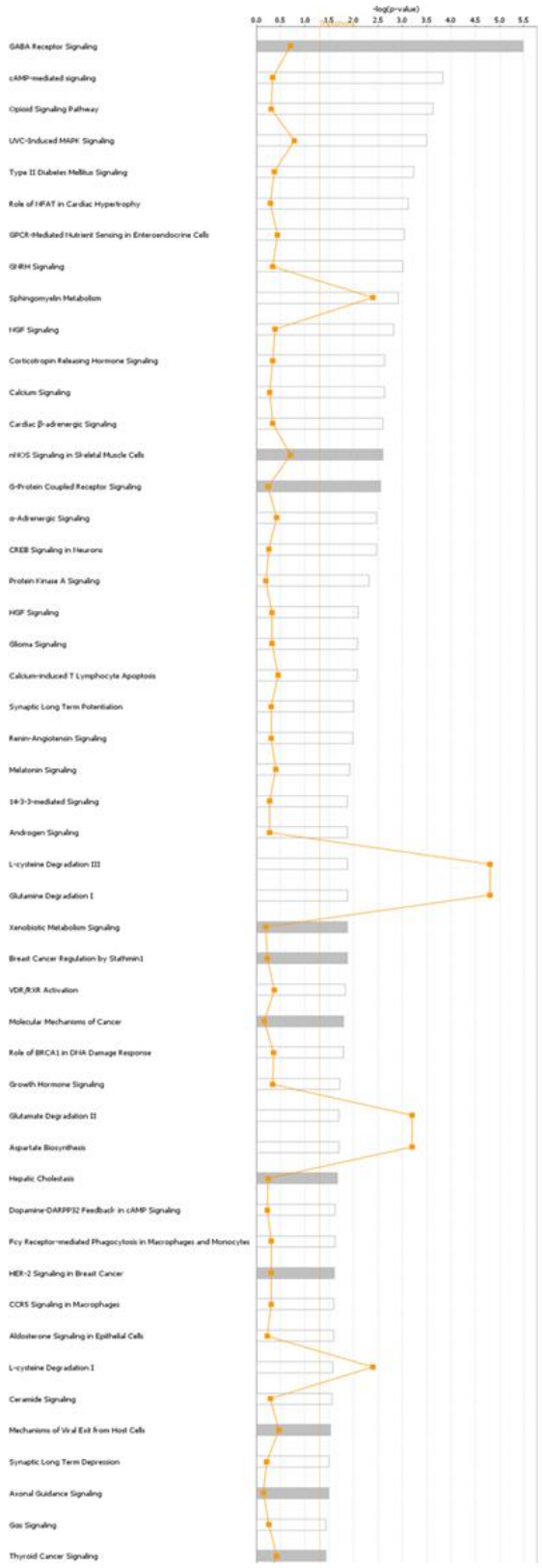

Figure 19: Enriched Canonical Pathways in DEGs identified from the overlap between the SHH subgroup in *Homo sapiens* and in *Mus musculus*.

IV. Conclusion and Discussion

Observing all the Venn diagrams shows that each statistical tool has its own advantages and limitations as each resulted in not only a different amount of DEGs, but also a variation in the overlap of DEGs. Therefore, to obtain as accurate results as possible, multiple tools have to be used when performing an integrative analysis as no single statistical tool can be ruled a favorite and outed as the best.

Overall, baySeq discovers the fewest number of DEGs for the Atoh1 and wild type comparison, DESeq detects the smallest amount of DEGs for the *Ptch1*$^{+/-}$ and wild type comparison, and limma reveals the greatest amount of DEGs in both. While limma does discover the largest amount of DEGs, as seen in the Venn diagrams, there are quite a few DEGs not discovered within the other packages, thus resulting in no overlap for those genes. PoissonSeq and Cuffdiff also tend to retain unique DEGs. Meanwhile, baySeq, DESeq, and edgeR share the most DEGs with all the other methods, resulting in very close to 100% of their DEGs being shared.

Using the expression $\frac{U}{T}$, where U stands for the number of unique genes of the package and T refers to the total number of genes, the FDR value for each statistical package was calculated. This method was applied to the data from q-values less than 0.05. In the comparison between the *Ptch1*$^{+/-}$ model and the wild-type model, baySeq resulted in an FDR of 0%, Cuffdiff with 5.8%, DESeq with 0%, edgeR with approximately 0%, limma with 7.6% and PoissonSeq with 5.6% FDR. The comparison between the Atoh1 and wild-type models had similar results, with only limma and PoissonSeq increasing their values (baySeq had 0%, Cuffdiff had 0.055%, DESeq had approximately 0%, edgeR had 0%, limma had 14% and PoissonSeq had 8.9%). From this information, we determined to exempt limma and PoissonSeq from analysis with the targetome because of their high FDR values, thus making room for the ChIP-Seq data.

ChIP-Seq data was added to the respective comparison for each experimental condition. We found 332 common DEGs identified by all the packages among the comparison between the primary tumor vs. the control groups (Figure 8) and 423 common DEGs identified for the metastatic tumor vs. the control (Figure 9). We observed that the ratio of uniquely identified DEGS is consistently low when the RNA analysis tools DESeq and edgeR were used, making them seem somewhat stronger than the other methods. Therefore, we used the data from those two packages as incentive to compare the transcriptome and targetome even further between the metastatic tumor and the primary tumor. When we analyzed the two tumors together, with their respective differences from the control, we detected 231 shared DEGs (Figure 13). The pathways from each statistical tool were obtained and compared. As seen in Figure 14, 202 similar pathways were revealed to be enriched. Looking at these pathways may give us better information about the targets and the effects of MB.

Applying comparative genomics to this study allowed us to study the biological similarities and differences between humans and mice, which could then help us apprehend the underlying causes of disease formation for MB. Part of the project was to organize human data into their respective subgroups (WNT, SHH, Group 3, and Group 4) to then be applied to the data we obtained from our mice models. While we were hoping to find DEGs shared between all the subgroups of MB, the analysis came back with zero similarities among all subgroups. Analyzing the subgroups to one another first confirmed our thoughts to the complexity of the development of the disease and suggests heterogeneity. This is thought to be the case as each type of MB has its own discrete somatic aberrations, activated pathways, and clinical outcomes; however, when we compared the SHH subgroup from both the human data and the mouse data, we detected 142 shared DEGs (Figure 18). Detailed investigation to these genes and their pathways will provide the next step in our research.

We have focused on the integrative and comparative analysis of transcriptome and targetome data for medulloblastoma, while also adding the component of comparative analysis between two species. By doing so, we have examined the molecular mechanisms of the cancer and have obtained hypotheses relating to the biological problem itself, thus pointing us in the next step forward. For example, we identified a set of genes that may contribute to the metastasis of medulloblastoma and we determined potential pathways and targets that may be involved in the development of the cancer. The purpose of the latter part of the project was to detect the similarities and differences in the gene expressions and cellular pathways affected in the different tumor subgroups and the similarities and differences in the gene expressions and cellular pathways targeted by the cancer in *Mus musculus* and *Homo sapiens*. Comparative analysis elucidates the dissimilarities between the respective subgroups and organisms, though the extent of similarities are still in need of further investigation

The future plan of this project will be to create a network visualization of the obtained results. Afterwards, we will investigate the data of other subgroups of MB cancer and run RNA-Seq analysis for the other subgroups in a similar fashion. Finally, we will perform a comparative study among all MB subgroups to reveal the molecular and cellular mechanisms for the overall formation of MB. Since the tumor data of humans was obtained from primary tumors, a further step in this research would be to attain gene expression data for metastatic tumors in human samples and perform a comparative analysis between the primary and metastatic tumors of *Homo sapiens* and *Mus musculus*. Doing so may enlighten development of a safer and more effective therapy.

References

Anders S., Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106. doi: 10.1186/gb-2010-11-10-r106. Retrieved from http://genomebiology.com/2010/11/10/R106/.

Cavalli F.M.G, Remke M., Rampasek L., Peacock J. et al (2017, Jun 12). Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell, 31*(6):737-754.e6. PMID: 28609654. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85212.

Chen, Y., McCarthy, D., Ritchie, M., Robinson, M., Smyth, G. (2019, April 24). EdgeR: differential expression analysis of digital gene expression data: User's Guide. Retrieved from https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Ghosh, S., & Chan, C.-K. K. (2016). Analysis of RNA-Seq Data Using Tophat and Cufflinks. *Methods in Molecular Biology (Clifton, N.J.)*, *1374*, 339–361. https://doi.org/10.1007/978-1-4939-3167-5_18. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26519415

Hardcastle, T.J. (2009, May 16). BaySeq: Empirical Bayesian analysis of patterns of differential expression in count data version. Retrieved from https://rdrr.io/bioc/baySeq/

Hardcastle, T. J., & Kelly, K. A. (2010, August 10). BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, *11*(1), 422. https://doi.org/10.1186/1471-2105-11-422. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-422

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinformatics, 16*, 169. Retrieved from http://www.interactivenn.net/

Illumina (2015). Precise analysis of DNA–protein binding sequences: Combining chromatin immunoprecipitation with NGS for genome-wide surveys of gene regulation. Retrieved from https://www.illumina.com/techniques/sequencing/dna-sequencing/chip-seq.html

Larget, Bret (2005, Sept 23). The Poisson Distribution. Retrieved from https://www.stat.wisc.edu/courses/st371-larget/poisson-handout.pdf

Li, J. (2011, Sept 6). Using "PoissonSeq" (Version 1.1) to discover differential expression based on sequencing data. Retrieved from https://www3.nd.edu/~jli9/PoissonSeq/PoissonSeq_instructions.pdf

McCarthy, D. J., Chen, Y., Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288-4297. Retrieved from https://bioconductor.org/packages/release/bioc/html/edgeR.html

Nonlinear Dynamics (n.d.). P-values, False Discovery Rate (FDR), and q-Values. Retrieved from http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/pq-values.aspx

Ritchie M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W., Smyth G.K. (2015, April 20). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic*

Acids Research, *43*(7), e47. https://doi.org/10.1093/nar/gkv007. Retrieved from https://bioconductor.org/packages/release/bioc/html/limma.html

Robinson M.D., McCarthy D.J., Smyth G.K. (2010). EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616. Retrieved from https://academic.oup.com/bioinformatics/article/26/1/139/182458

QIAGEN® (2017). Ingenuity® Pathway Analysis (IPA®): For the analysis and interpretation of 'omics data. Retrieved from http://pages.ingenuity.com/rs/ingenuity/images/IPA_data_sheet.pdf

St. Jude Children's Research Hospital (2019). Medulloblastoma. Retrieved from https://www.stjude.org/disease/medulloblastoma.html

Trapnell, Cole, et al (2014). Cuffdiff (v7). Retrieved from https://software.broadinstitute.org/cancer/software/genepattern/modules/docs/Cuffdiff/7

Trapnell, Cole (2017). Cufflinks: Transcriptome assembly and differential expression analysis for RNA-Seq. Retrieved from http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/