Transcription Regulatory Networks in Medulloblastoma

Thesis submitted to:

The Biotechnology Graduate Program

In partial fulfillment of the requirements for the degree of Master of Science

By:

Yasmeen Hesham Amin Howeedy

BSc. Pharmaceutical and clinical sciences, Faculty of Pharmacy, Ain Shams University

Under supervision of:

Prof. /Ahmed Moustafa

Associate professor, biology department, the American University in Cairo

Dr. / Shahenda El Naggar

Head of basic research department, Children’s Cancer Hospital in Egypt 57357

January 2015
Dedication

I dedicate this work to the closest to my heart ever, the ones I can never survive without. To my beautiful and forever supportive family...

To my Father, for always being there for me even through the hardest times.

To my Mother, for making life as joyful as it can be with you always beside me.

To my Brother, for your love respect and BRILLIANT advices.

To my Sister, for being my best friend. I trust you the most and cherish your presence in my life.

To Nanna and Doody no words can describe how I miss you.

And finally, for you showing up in my life...Ihab.
Acknowledgement

I would like to thank my mentor and of course THE advisor Prof. Ahmed Moustafa, graduate program director, the American University in Cairo, for teaching me almost everything I have learned for the past three years both academically and socially. His support and continuous guidance throughout this research trip have been and will always be most appreciated. I would like also to thank my co-advisor Dr Shahenda El Naggar, head of basic research department, the children’s cancer hospital in Egypt, for her unlimited support and encouragement to reach high quality research findings. I am most thankful for her insistence on me presenting my work in conference and for guiding me through the knowhow of this research.

I’m also deeply grateful for the grant support office, specifically Dean Amr Shaarawi, for their financial support throughout my participation in the European Molecular Biology Organization’s conference with my poster. I would like also to acknowledge the office of the dean of graduate studies, specifically Dean Adham Ramadan, for granting me the laboratory instruction award that aided me through my educational trip.

The AUC has become a delightful place because of them. I thereby would like to thank all my professors who shared with me part of their knowledge and their time for which I am most grateful. Many thanks to Dr. Asma Amleh, Dr. Hamza Al Dorry, Dr. Rania Siam, Dr. Walid Fouad and of course Dr. Ahmed Moustafa.
I also thank my colleagues Mariam Rizkallah, Hadeel El Baradisy, Aya Medhat, Lana Murad, Riham Kabbani, Ahmed Samir, Noha Nagdy, Mustafa Adel, Heba Adel, Sarah El Alawi, Hazem Sharaf, Mohamed Nagy, MM Lotaiief, Nahla Hussein, Laila Ziko, Mustafa Abo Hawaya, Sherriff Ammar, Karim Abdelhady, Marwan Felafel, Marwa Nassar, Nihal Abd El Nabi, Ghada Mustafa, Alyaa Mahmoud, Salma Al Shafie and last but not least Amgad Ouf. Throughout those years, they have been my guiding family, a good reason to draw a smile and a good promise for a promising future.

Finally, I’m more than pleased to be a member of this team radiating with hope and intelligence in the CCHE. I thank my team Myrette, Heba, Marwa, Magi and Maha for their support.
Abstract

Pediatric central nervous system tumors are the second most frequent tumor after leukemia and the most common form of childhood solid tumor. They account for 22% and 10% of all malignancies among children aged up to 14 years and 15-19 years respectively. Medulloblastoma is the most common pediatric primary malignant intra-cranial neoplasm, accounting for 20-25% of all childhood brain tumors. Diagnosis and grading of medulloblastoma have mainly been dependent on histological assessment and clinical presentation. In recent years it has been becoming more evident that medulloblastoma tumors display heterogeneity in genetics and clinical response that cannot simply be explained by the histological classification. Recent analysis using microarray, and whole genome sequencing identified five core subgroups: WNT (WNT/wingless), SHH (Sonic hedgehog), Subgroup 3, Subgroup 4 and Subgroup 5 according to their molecular signatures.

In the current study the gene expression signature of 76 medulloblastoma samples previously described by the St. Jude team was unitized to construct a protein-protein interaction network for differentially expressed genes of each molecular subgroup. Also, transcriptional regulatory network for each subgroup was demonstrated, providing novel insights into the complex transcriptional regulation of the genes involved in medulloblastoma. This work introduces transcriptional signatures for the medulloblastoma subgroups, which can be potentially utilized for diagnostic and therapeutic applications.

The present study show that WIF1 (WNT inhibitory factor 1), LEF1 (lymphoid enhancer binding factor1) and FZD10 (frizzled receptors 10) are potential gene biomarkers for WNT subgroup. The WNT subgroup also demonstrated an up regulation in genes involved in MAPK (mitogen activating protein kinase) signaling pathway as well as WNT pathway. Transcription factors (TF)
regulating the differential expression in this subgroup are SUZ12 (suppressor of zest 12) and NANOG. SHH pathway marker gene profiles belong to Gli3 (gli family zinc finger 3), PPARA (peroxisome proliferator-activated receptor alpha), LRRC7 (leucine rich repeat containing 7), BMP2/4 (Bone morphogenetic peptide 2, 4) and FOXG1 (forkhead box 1). With the involvement of this subtype gene behavior with the SHH signaling pathway, these expression profiles are suggested to be transcriptionally regulated by E2F1 and MYC.

Subgroup 3 displayed a down regulation in FOXG1 with associated TTR1 (transthyretin) and DUSP (Dual specificity phosphatase). PRLR (prolactin receptor) and MET (met proto oncogene) up regulations supported the involvement of these patient profiles with retinopathy progression pathways as well as pathways. TFs regulating cancer progression in this subgroup included TBX1 (T box 1), CEBP (CCAAT enhancer binding protein) and CLK1 (CDC-like kinase), while the Apoptotic transcriptional regulation is through AVEN (apoptosis caspase activation inhibitor).

Subgroup 4 exclusively showed TPN1 (Transport of pyridoxine protein 1) down regulation not observed in other profiles. Cytogenetic pathways involvement is witnessed in this subgroup with down regulation of cell cycle genes. TFs regulating cell cycle pathways are the E2F1 and SKP2 (S phase kinas associated protein 2) Subgroup 5 showed a FOXG1 down regulation with no differential expression of the MYC. TEX1S down regulation was unique to this subgroup with no clear mechanism. Involvement in cAMP biosynthesis is a trend in this subgroup. TFs suggested to regulate this behavior are ZNF281 and E2F1/ETS1 regulation.

From this we conclude that, microarray gene expression profiling provides a comprehensive platform for the study of different gene behaviors in medulloblastoma. We also suggest that utilization of RNA-Seq high throughput technology can assist in detecting altered gene expressions as well as SNPs (single nucleotide polymorphisms) that might be an underlying cause for medulloblastoma progression.
# TABLE OF CONTENTS:

1. List of abbreviations: ........................................................................................................... 11

2. Chapter 1: Introduction ........................................................................................................ 13
   - Central nervous system (CNS) malignant tumors: ........................................................... 13
   - Pediatric brain tumors: ........................................................................................................ 14
   - Medulloblastoma: ............................................................................................................... 15
     - Medulloblastoma Diagnosis: ............................................................................................ 15
     - Medulloblastoma subgroups: .......................................................................................... 18
   - Microarray analysis: ............................................................................................................ 28
     - Preprocessing of microarrays: ......................................................................................... 30
     - Data Processing: .............................................................................................................. 30
     - Systems approach: ......................................................................................................... 31

3. Chapter 2: Research hypothesis and objectives .................................................................... 34
   - Objective 1 ............................................................................................................................ 34
     - Identify differentially expressed genes for each molecular subgroup in medulloblastoma
       patients .............................................................................................................................. 34
   - Objective 2 ............................................................................................................................ 34
   - Objective 3 ............................................................................................................................ 34
   - Objective 4 ............................................................................................................................ 34

4. Chapter 3: Materials and methods ...................................................................................... 35
4.1. The dataset: ........................................................................................................... 35

4.2. Expression profiling using R: .............................................................................. 35
   4.2.1. Preprocessing: ............................................................................................ 35
   4.2.2. Normalization: ........................................................................................... 36
   4.2.3. Outlier array detection: .............................................................................. 36
   4.2.4. Gene filtration: .......................................................................................... 37
   4.2.5. Data statistical analysis: .............................................................................. 37

4.3. Functional Annotation: ....................................................................................... 39

4.4. String tool for Protein-Protein interaction: ......................................................... 39

4.5. Cytoscape network visualization: ......................................................................... 40

5. Chapter 4: Results and Discussion ........................................................................... 41
   5.1. Microarray raw data preprocessing: ................................................................. 41
      5.1.1. Testing the sample quality: ...................................................................... 42
      5.1.2. Testing the hybridization and overall signal quality: ............................... 45
      5.1.3. Signal comparability and biases diagnostic: .............................................. 47
      5.1.4. Intensity dependant biases: ....................................................................... 50
      5.1.5. Spatial biases: ......................................................................................... 50
   5.2. Data Normalization: ......................................................................................... 50
   5.3. Evaluation of preprocessing: ............................................................................. 51
      5.3.1. Signal comparability and biases diagnostic: .............................................. 51
5.3.2. Array correlation: ................................................................. 55

5.4. Microarray data processing: ......................................................... 67

5.4.1. Gene filtration: ................................................................. 67

5.4.2. Statistical analysis: ............................................................ 70

5.4.3. Filtration: ................................................................. 72

5.5. Annotation: ........................................................................... 79

5.6. Specific gene markers detection: ........................................... 79

5.7. Protein-protein interactions: ..................................................... 83

WNT subgroup protein-protein interaction network: .................. 84

SHH subgroup protein-protein interaction network: ................. 87

Subgroup 3 protein-protein interaction network: ......................... 90

Subgroup 4 protein-protein interaction network: ................. 93

Subgroup 5 protein-protein interaction network: ......................... 98

5.7. Transcription regulatory networks: ......................................... 102

WNT subgroup transcription regulatory network: .................. 102

SHH subgroup transcription regulatory network: .................. 106

Subgroup 3 transcription regulatory network: ......................... 108

Subgroup 4 transcription regulatory network: ......................... 113

Subgroup 5 transcription regulatory network: ......................... 115

6. Chapter 6: Conclusion ................................................................. 117
| Objective 1: ........................................................................................................................................ | 117 |
| Objective 2: ........................................................................................................................................ | 118 |
| Objective 3: ........................................................................................................................................ | 119 |
| Objective 4: ........................................................................................................................................ | 120 |
| 7. Future directions: ......................................................................................................................... | 122 |
| 8. Appendix: ....................................................................................................................................... | 124 |
| 9. References: .................................................................................................................................... | 152 |
1. **List of abbreviations:**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>µ</td>
<td>Mean of the sample</td>
</tr>
<tr>
<td>A</td>
<td>Absent</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis Of Variance</td>
</tr>
<tr>
<td>AUC</td>
<td>The American University in Cairo</td>
</tr>
<tr>
<td>BMP4</td>
<td>bone morphogenetic protein 4</td>
</tr>
<tr>
<td>BOC</td>
<td>cell adhesion-associated oncogene</td>
</tr>
<tr>
<td>CCHE</td>
<td>Children’s Cancer Hospital in Egypt</td>
</tr>
<tr>
<td>CRX</td>
<td>cone rod homeobox</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>catenin beta 1 (cadherin associated)</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of Variance</td>
</tr>
<tr>
<td>DDX3X</td>
<td>DEAD(Asp-Glu-Ala-Asp)box helicase 3, X-linked,</td>
</tr>
<tr>
<td>DE</td>
<td>Differentially expressed</td>
</tr>
<tr>
<td>DKK1</td>
<td>dikkopf-related protein 1</td>
</tr>
<tr>
<td>DUSP</td>
<td>dual specificity phosphatase</td>
</tr>
<tr>
<td>EOMES</td>
<td>Eomesodermin</td>
</tr>
<tr>
<td>EPHA8</td>
<td>EPH receptor A8</td>
</tr>
<tr>
<td>ETN1</td>
<td>early transposon insertion site 1</td>
</tr>
<tr>
<td>FGF</td>
<td>fibroblast growth factor</td>
</tr>
<tr>
<td>FOXG1B</td>
<td>forkhead box G1</td>
</tr>
<tr>
<td>GABRA5</td>
<td>gamma-aminobutyric acid (GABA) A receptor, alpha 5</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene Expression Omnibus</td>
</tr>
<tr>
<td>Gli</td>
<td>glioma associated oncogene homolog 1/2/3 (Zn Finger)</td>
</tr>
<tr>
<td>GRB</td>
<td>Growth factor receptor bound protein</td>
</tr>
<tr>
<td>GSK3</td>
<td>Glycogen synthase kinase</td>
</tr>
<tr>
<td>HHIP</td>
<td>hedgehog interacting protein</td>
</tr>
<tr>
<td>LEF1</td>
<td>lymphoid enhancer binding factor 1</td>
</tr>
<tr>
<td>MAD</td>
<td>Median Absolute Deviation</td>
</tr>
<tr>
<td>MAP2</td>
<td>microtubule-associated protein 2</td>
</tr>
<tr>
<td>MAPK</td>
<td>Mitogen-activated protein kinase pathway</td>
</tr>
<tr>
<td>MAS5.0</td>
<td>Microarray Affymetrix Suite 5.0 algorithm</td>
</tr>
<tr>
<td>MAST1</td>
<td>microtubule associated serine/threonine kinase 1</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MM</td>
<td>Mismatch</td>
</tr>
<tr>
<td>MPP3</td>
<td>membrane protein palmitoylated 3</td>
</tr>
<tr>
<td>MYC</td>
<td>myelocytomatosis viral oncogene homolog</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NEUROD1</td>
<td>Neurogenic differentiation 1</td>
</tr>
<tr>
<td>NPR3</td>
<td>natriuritic peptide receptor 3</td>
</tr>
<tr>
<td>NR2E3</td>
<td>nuclear receptor subfamily 2 group E3.</td>
</tr>
<tr>
<td>NRL</td>
<td>neural retinal leucine zipper</td>
</tr>
<tr>
<td>OTX3</td>
<td>orthodenticle drosophila homolog 3</td>
</tr>
<tr>
<td>PM</td>
<td>Perfect Match</td>
</tr>
<tr>
<td>PP</td>
<td>Percent Present</td>
</tr>
<tr>
<td>PTCH1</td>
<td>patched homolog 1</td>
</tr>
<tr>
<td>PWMs</td>
<td>Position weight matrices</td>
</tr>
<tr>
<td>RB1</td>
<td>retinoblastoma 1</td>
</tr>
<tr>
<td>RTN4R</td>
<td>Reticulon 4 receptor</td>
</tr>
<tr>
<td>SCNAs</td>
<td>Somatic copy number aberrations</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation of the sample</td>
</tr>
<tr>
<td>SHH</td>
<td>Sonic Hedgehog pathway</td>
</tr>
<tr>
<td>SMARCA4</td>
<td>SWI/SNF Related, Matrix Associated Actin Dependent Regulator of Chromatin, Subfamily A, Member 4</td>
</tr>
<tr>
<td>SMARCD</td>
<td>SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d</td>
</tr>
<tr>
<td>SUFU</td>
<td>supressor of fused homolog</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>UNC5D</td>
<td>unc-5 homolog D</td>
</tr>
<tr>
<td>WIF1</td>
<td>WNT inhibitor factor 1</td>
</tr>
<tr>
<td>WNT</td>
<td>Wingless-Type MMTV Integration Site Family, Member 2B.</td>
</tr>
</tbody>
</table>
2. **Chapter 1: Introduction**

Cancer expresses group of diseases involved in abnormal cell growth that might even propagate and invade other tissues (Teicher, Linehan, & Helman, 2012). Malignant tumors and benign tumors describe over growing tissues that spread or not spread to other parts respectively. Benign tumors receive less attention compared to malignant ones; as a matter of fact they are not considered a type of cancer. Benign tumors resemble an overgrowth in size of certain tissue without spreading thus providing a local defect. Malignant tumors however present a bad prognosis with a spread that might be uncontrolled resulting in continuous defective cellular performance followed by an organ malfunctioning eventually ending with arrested body functions and death. In 2014, the American Cancer Society has estimated the number of new cancer cases to be 1,665,540 with 585,720 cases as deaths from cancer in United States of America (USA). Despite the fact that new cancer cases’ number has declined 20% from the estimated numbers in 1996, yet it is not even close to approaching zero cases (Parker, Tong, Bolden, & Wingo, 1996; Siegel, Ma, Zou, & Jemal). Accordingly, the concern of how to control over the disease progression remains there.

It is the malignant tumors that formulate this growing concern worldwide due to difficulty in accurate diagnosis, treatment or even controlling this spread. This type of disease can occur anywhere in the body. One example of these organs is the brain.

**Central nervous system (CNS) malignant tumors:**
These are tumors that arise in the brain due to an abnormal cell growth in the brain tissue. These tumors are expected to affect 23,380 patients in 2014 with an estimated 14,230 deaths.
expected in USA. Interestingly, 4,200 pediatric cases are expected to arise in 2014 including patients of >15 years old age. The five year rate of survival in these cases is only 75%. In other words, quarter the pediatric patients diagnosed for brain malignant tumor is expected to die within 5 years of diagnosis(.Siegel et al., 2006). Although the rate of inflation in the mortality due to these tumors is stabilized, the percentage remains huge and poses a need for search for more sophisticated techniques for speedy diagnosis and safe recovery (Kohler et al., 2011).

CNS tumors can be divided into two subgroups according to its origin, primary and metastatic. Primary brain tumors originate in the brain and start spreading to other body parts through the cerebrospinal fluid. Metastatic tumors arise in other body parts and then move to the brain(UK, 2014). Primary tumors can be divided into 130 tumors depending on the type of the cell they originated form. One type of the primary tumors is the primitive neuroectodermal tumors (PNETs) arising from the residual cells left over from early embryo development of the body (neuroectoderm). PNETs arise in different brain sites including cerebral hemispheres, cerebellum, pineal gland, spinal cord and brain stem(Lester et al., 2014).

**Pediatric brain tumors:**
Pediatric are patients aged 19 years or younger. CNS tumors in these patients vary between 6 main malignant types: glioblastoma, PNETs, Medulloblastoma, neurofibroma, atypical teratoid rhaboid tumors and congenital tumors. The different types of malignant tumors arising in pediatrics are dependent on the cell originating the tumor (Fleming & Chi, 2012). Glioblastomas originate from glial cells connecting between the neurons in the brain (Verhaak et al., 2010). Neurofibromas have a malignant growth in the nerve fibers originating from to genetic disorders: neurofibromatosis 1 and 2 (NF1/2) (Sawada et al., 1996). Atypical teratoid rhaboid tumors (ATRT) arise from the cerebllum and brain stem rhaboid tissue. These are characterized by rapid growth and invasiveness to other CNS tissues. ATRT are considered most common
cause of CNS malignant tumor cells in pediatrics below age of 3 years old (Tekautz et al., 2005). The second most deathly CNS tumor, and third after acute lymphoblastic leukemia across all cancer types, is the medulloblastoma (Mealey & Hall, 1977). Medulloblastoma are primitive neuroectodermal tumors arising in the posterior fossa in the cerebellum. Compared to PNETs, these tumors are located only in the cerebellum and their growth in the fourth ventricle is characteristic. PNETs however carry the same characteristics as medulloblastoma except for their growth in the cerebral hemispheres (Climate, Dyck, Phylogenies, & Compare, 2004). Finally congenital CNS tumors arise from the germ cells located in the brain tissue (Wakai, Arai, & Nagai, 1984).

**Medulloblastoma:**

It is a highly invasive brain tumor that is more likely to disseminate through CNS early in its course. 80-90% of the children with no dissemination are expected to be cured using current treatments. However, only 75% of disseminated cases survive with combined aggressive treatment including surgery (P. a Northcott, Korshunov, Pfister, & Taylor, 2012). This aggressive treatment causes significant endocrinal and intellectual consequences after resolution (obey MacDonald, 2014). Medulloblastoma presents one of the most aggressive CNS tumors in children.

**Medulloblastoma Diagnosis:**

Since the first diagnosis of the disease in early 1970s (Mealey & Hall, 1977), medulloblastoma patients were either diagnosed as standard risk patients with age over 3 years having no leptomeningeal dissemination or as high risk if other than that (Ellison, 2010; P. a Northcott et al., 2011a). Absence of a proper diagnostic stratification increased the risk of failure of current treatment approaches. Survivors are often left with significant neurological, intellectual and physical disabilities secondary to the effects of non specific cytotoxic therapies on the
developing brain (P. a Northcott, Shih, et al., 2012). These facts lead to the need for better disease-risk stratification and a molecular level understanding of medulloblastoma (Kool et al., 2008).

Currently, a histological examination of a medulloblastoma tissue sample is used as mode of improved diagnostic tool for these tumors. Histologically, medulloblastoma can be subdivided into 4 different classes: tumors showing classic histology, desmoplastic/ nodular histology, desmoplastic with extreme nodularity and large cell anaplastic tumors. Classic medulloblastoma presents diffuse masses of small, undifferentiated oval or round cells. Neuronal differentiation is manifested by neuropil and rosette formation. These structures present cells’ arrangements in a circle around a fibrillary center (see Figure 44). Desmoplastic/ nodular tissues are permeated by fine collagen (reticulin) fibers that give it a firm consistency. The centers of the nodules (pale islands) are reticulin-free. They are less cellular than the surrounding densely packed small cells and are composed of larger cells with neuronal differentiation. The fibroblastic (desmoplastic) reaction occurs when the tumor extends into the subarachnoid space. Extensive nodularity subgroup is difficult to differentiate from this subgroup. Large cell anaplastic subgroup displays large cell nuclei with a high rate of mitosis and apoptosis.

With this histological classification in hand, improved diagnostic protocol is developed, however, treatment outcomes remains non satisfactory. Accordingly, with the arising new technologies and improvements in the science field, new techniques for disease studying have been developed. Scientists’ efforts to improve the understanding of the disease mechanism led to subdividing medulloblastoma into at least 4 separated subgroups differing in their cell of origin, clinicopathological features and disease (Jones et al., 2012), (P. a Northcott, Shih, et al., 2012) and(Ellison et al., 2011a). The subgroups are namely the Wnt (WNT/wingless signaling), SHH (Sonic Hedge Hog), the subgroups 3 and 4 patients that are still under characterization and
present the greatest challenge (Northcott et al., 2012). These findings succeeded to relate specific somatic copy number aberrations (CNA) and single nucleotide variables (SNV) as well as gene expression profiles within each subgroup that could be further studied for possible treatment targets. A fifth subgroup remains under study for significant molecular signature (Kool et al. 2008, Thompson et al. 2006).

These classifications were obtained utilizing novel approaches for molecular studying. The approaches applied can be numbered into:

1. Genome sequencing: for identifying the sequences of each gene involved in the disease pathogenesis to detect possible sequence mutations (Rausch et al., 2012).
2. Genome wide association studies (GWAS): to combine data from different sources in one informative dataset and associate genotype of each sample to an overall phenotype across the whole dataset (Montasr, Aziz, & Ahmed, 2013).
3. Somatic copy number aberrations (SCNAs): applying whole exome sequencing techniques allows for detecting of specific changes in certain genes that might underlie a certain phenotype (Pugh et al., 2012b).
4. Microarray and genomics: chips carrying oligo-nucleotides/ mRNA/ cDNA complementary probes target certain genes to determine their intensity of expression in compare to control samples (Schwalbe et al., 2011).

The analysis of the gene expression patterns can also help show the type of regulation applied on each gene under the disease conditions giving a chance to relate different gene expression patterns to new diagnostic and treatment approaches.
**Medulloblastoma subgroups:**

As mentioned earlier, five medulloblastoma subgroups comprise the different clinical and pathological characteristic of medulloblastoma. By relating specific (SCNA) and (SNV) as well as gene expression profiles, the following subgroups were identified as follows:

**Wnt subgroup:**

Wnt subgroup children have the best prognosis compared to other subgroups. Nearly all Wnt medulloblastomas studied have a classic histology (Massimino et al., 2013; P. A. Northcott et al., 2011). The activation of the Wnt/ wingless pathway is associated with chromosome 6 monosomy with hardly other genetic variations observed in this subgroup (Clifford et al., 2006). The canonical Wnt pathway activation generally plays a role in increasing the proliferation of the neural cells in the CNS (Nishihara, Tsuda, & Ogura, 2003). Also, the activation underlies an over expression of the EAG2 voltage-gated potassium channels that is also believed to play a role in controlling cell proliferation (X. Huang et al., 2012; Nishihara et al., 2003). On the other hand, chromosome 6 carries several genes (WISP3, T-gene, Syne-1, EFHC-1, and others) that play a distinctive role in neural cells proliferation and differentiation as well as transcription regulation. (Clifford et al., 2006; Leach, DeMars, Hasstedt, & White, 1986; SATOH et al., 2000) Combining both statuses, it is suggested that the relatively good prognosis associated with these subgroup patients might be related to this association.

The CNNB1 gene is one of the genes whose expression level is manipulated in Wnt subgroup medulloblastoma. Increased nuclear levels of its protein β catenin that is known to normally increase the proliferation of the cancerous cells is over expressed in the cytoplasm suggesting a bad prognosis (David W. Ellison, Olabisi E. Onilude, Janet C. Lindsey, Meryl E. Lusher & Roger E. Taylor, Andrew D. Pearson, 2005). However, in Wnt activated patients,
cancerous cells seem to acquire a promoted apoptosis character that renders them more liable to radiotherapy. β-catenin is normally broken down through ubiquitination mechanism to avoid its nuclear translocation. When Wnt signaling is active, β catenin is stabilized and translocates into the nucleus to interact with transcription factors of the LEF/TCF1 family and regulate expression of target genes (See Figure 1). In the absence of Wnt signals, β catenin is phosphorylated by casein kinase I (CK1) and GSK3b. Phosphorylated β-catenin is ubiquitinated and degraded by proteosomes. Thus, conditional mutation of β catenin leads to cell loss. In contrast, an over expression of activated β catenin induces expansion of neural tissues. Also, transcriptional control mediated by the β catenin and the LEF/TCF transcription factors is pivotal for the proliferation of neural stem cells. On another hand, medulloblastoma Wnt subgroup model generated by Gibson et al., 2012 where a CTNNB1 allele was conditionally knocked out and crossed with TP53 mutated mice developed into medulloblastoma.

Besides the CTNNB1 expression profile, integrative genomics applied by (Kool et al., 2008) showed that many genes in Wnt pathway are over expressed, including AXIN2, LEF1, WIF1, DKK1, DKK2, WNT11 and WNT16. Several of them are inhibitors of the WNT pathway and are up regulated in a negative feedback loop upon pathway activation.

Also an over expression of members of the TGFB pathway (BMP4, BMP7, BAMBI, AMHR2, SMAD3, TGFBI and INHBA) was observed. An increased expression of genes involved in protein biosynthesis, cell cycle, NOTCH and PDGF pathways was detected (also in SHH subgroup as to be introduced soon); All supporting the involvement of the activated Wnt pathway in the tumor progression.
Figure 1: Diagram showing genes involved in Wnt and SHH signaling pathways. (Nature reviews, 2006)
**SHH subgroup:**

These patients are biologically very similar to the Wnt patients to the extent that they cluster together away from other subgroups. Patients are presented with desmoplastic or large cell anaplastic histology (P. A. Northcott et al., 2011).

On the molecular level, SHH subgroup displayed a mutation in the PTCH1 and SUFU which are suggested to underlie the activation of the pathway (Schwalbe et al., 2011). Expression profiles revealed an up regulation in the PTCH1 and BOC and GLI1/2 but not GLI3. HHIP was also significantly up regulated (P. a Northcott et al., 2011b; Robinson et al., 2012; Romer et al., 2004; Verhaak et al., 2010). On another hand, BMP2/4 were down regulated in this subgroup. (See Figure 1) Proofs have demonstrated that a mutation in PTCH1 is sufficient to activate the pathway by activating the Smo (Romer et al., 2004). Thus these patients’ gene expression profiles might not present an over expression of the HHIP or the PTCH1 (Schwalbe et al., 2011).

NOTCH and PDGFA pathways also displayed an up regulation in their genes. SHH pathway activation has been associated with up regulation of GLI1/2 (E. Y. Lee et al., 2010; Villavicencio, Walterhouse, & Iannaccone, 2000). GLI is involved in the transcriptional control of N-MYC and Cyclin D and is required for robust mitogenic responses in regions of high SHH signaling. Post transcription SHH mechanism inhibits the GSK3, thus restrict N-MYC to malnutrition (glutamate deficient) areas to coordinate cell cycle progression. This is also associated with the B catenin that can be witnessed frequently up regulated expression in association with this mechanism (Di et al., 2014).

**Subgroup 3:**

No clear genetic marker has yet identified this subgroup. However several activation/deactivation gene profiles can be related. The subgroup presents the lowest survival
percentage compared to other subtypes placing it as the poorest prognosis patients. These tumors are a classic medulloblastoma with large cell anaplasia as a histological image. Over expressed genes observed in these tumors are involved in neuronal differentiation (NEUROD2, ETN1, RTN4R, DPYSL5, NEURL and NPAS2) (Daves, Hilsenbeck, Lau, & Mann, 2011; P. Huang et al., 2011; Verhaak et al., 2010). Also members of cytoskeleton structuring (MAST1, MAP2, and MPP3) are over expressed. Up regulation of GRM1, 2, 8 was also observed as indication of glutamate biosynthesis cycle activation (Taylor et al., 2012).

High level of expression of MYC is observed in this subgroup. But the characteristic marker for this subgroup utilized in immunohistochemical positivity testing is the NPR3. These patients are associated with up regulation in retinopathy forming genes (Pugh et al., 2012b; Taylor et al., 2012).

**Subgroup 4:**

This subgroup is often clustered with the later subgroup due to high similarity between their expression profiles. The most significant difference between these subgroups is the MYC expression profile. MYC is over expressed in subgroup 3 yet no differential expression is observed in subgroup 4 (Pugh et al., 2012b; Taylor et al., 2012). Subgroup 4 patients are characterized by 17q isochromosome that is used a diagnostic approach for this subgroup. Also, a notable loss of chromosome X is observed placing a hypothesis for the reason behind its high incidence in females (Kool et al., 2008; Rausch et al., 2012; Taylor et al., 2012). These tumors are classic type with desmoplastic histology (Northcott et al., 2011). Slight retinopathy incidence is frequently witnessed in these patients (Schwalbe et al., 2011).

Differentially expressed genes (DEGs) in this subgroup include: MAPK signaling (also in Subgroup 3) as KRAS, DUSP4, DUSP5, GRB2, FGF1, FGF9, FGF13, and FGF14 (Taylor et al.,
2012) (See Figure 2). Also retinal transcription factors, such as NRL, CRX and NR2E3 were expressed at moderate to high levels in this subgroup as well as in subgroup 5 (Kool et al., 2008).
Figure 2: MAPK signalling pathway

Obtained from: http://www.novusbio.com/mapkerkpathway.html
Subgroup 5:

The subgroup is usually identified to belong to subgroup 4, however, it clusters separate from subgroup 4 as presented in work of (Kool et al., 2008; P. A. Northcott et al., 2009; Taylor et al., 2012) Transcription factors, FOXG1B and EOMES are differentially expressed (DE) in this subgroup. Also, several genes involved in neuronal migration, like UNC5D and EPHA8 were up regulated (Kool et al., 2008).

Like subgroup 3 and 4, retina regulating transcription factors (RB1, S-antigen and opsin) as well as (GABRA5, SMARCD, PLXN and FGF9) are over expressed (Kool et al., 2008).

The interest in the proposed topic in this thesis is based on the molecular findings of the three research subgroups: the University of Heidelberg in Germany (Jones et al., 2012), the St. Jude children’s cancer hospital teams (P. A Northcott, Shih, et al., 2012) and (Ellison et al., 2011a). These teams studied the single nucleotide variations (SNVs) and copy number aberrations (CNAs) associated with several different primary tumor samples. These findings clearly separated the 4 subgroups depending on sequencing data as well as expression profiles as illustrated in Figure 3. Figure 4 also shows the SCNAs results for the 4 medulloblastoma subgroups. Accordingly, microarray data analysis results are needed to support the previously discussed cytogenetic events.
Figure 3: Clinical and genomic features for 4 medulloblastoma subgroups as discussed by Northcott et al., 2012.

The figure demonstrates the different clinical and genomic features that subgroup medulloblastoma into 4 main subgroups. It also demonstrates the age associated with each subgroup and the histological picture.
Figure 4: Subgroup specificity of common genetic aberrations as discussed by Jones et al. 2012.

The combined cohort clinical study microarray clinical results done on n=126. The first panel is the Wnt subgroup; chromosome 6 aberrations together with CTNNB1, DDX3X, TP53 and SMARCA4 genes are observed. The second panel is the SHH subgroup showing 9q chromosome aberrations in expression and PTCH1. Subgroup3 and 4 demonstrate 17q, 17p chromosome gene expression changes with aberrations in X chromosomal gene expression (more common in females than males). MYC expression pattern is changed in subgroup 3 studies. The last panel represents insignificant gene expression patterns found. (UPD=Uni-parental disomy, ND= non distinct data)(Jones et al., 2012)
Microarray analysis:
Microarray has become a highly used technique that is needed by biologists to monitor expression patterns for genes across different conditions. The idea behind microarray analysis is the original central dogma; DNA is transcribed into mRNA that is translated into proteins. DNA can be replicated into DNA as well as mRNA can be reverse transcribed into DNA. (Babu, 2008.)

Hybridization sets the basis for the operation of these chips. Microarrays are basically slides carrying oligonucleotide/mRNA/cDNA probes. These are sequences of bases complementary to specific genes in order to bind them and fluorescence. Oligonucleotide probes are usually 25-mers that are directly synthesized to the glass slide. These chips may carry about 900,000 oligos where each oligo is present in millions of copies. Knowing the location of each probe, it is easy to conclude the amount of a certain gene in certain sample by allowing them to hybridize with their complementary probes (See Figure 5)(McCall, 2011).

Each gene is represented by a probeset. A probeset is a perfect match probe and a mismatch probe in pairs. Perfect matches are identical sequences to the target gene, while mismatches carry single base substitution at the center (See Figure 6). This design aims at determining the background and non specific hybridization that might contribute to the light signal produced upon measuring. Added to that, different experimental conditions, lab variations and chip manufacturing differences share in an erroneous signal(Archer & Reese, 2010). Accordingly, normalization algorithms have been developed to adjust for these variations in order to obtain reliable signals across all chips for a single gene. In other words, preprocessing of the signals is carried out before their analysis.
Figure 5: Design of GeneChip arrays produced by Affymetrix co.
Obtained from: http://www.vsni.co.uk/software/genstat/htmlhelp/marray/AffymetrixChips.htm

Figure 6: Probeset design in an mRNA microarray GeneChip.
Obtained from: http://www.vsni.co.uk/software/genstat/htmlhelp/marray/AffymetrixChips.htm
**Preprocessing of microarrays:**

The Affymetrix GeneChip operating system (GCOS) is software that translates the fluorescent bands emerging from the hybridized chip into a grid image. This produces image files in form of (.CEL) extension or scalar data about probe position and target sequence in form of (.cdf) file (Y. H. Yang, Buckley, & Speed, 2001). Oligonucleotide assays are one colored assays (no red / green fluorescence) thus provide absolute expression values. These expression values are of usually variable from one group to another. The variation is due to biologically significance or additional insignificant factors. Normalization is performed by adding a correction factor for all intensities across the chips in order to decrease variations due to non–hybridization sources thus preserving biological variations and minimizing experimental ones (C. Argyropoulus, 2006). An assumption is placed that most genes/probes do not change between two conditions in order to perform the calculations. Several algorithms have been developed to achieve this step. These include robust microarray analysis (RMA), GC-RMA as well as Microarray Analysis Suite (MAS5.0) …etc. They differ in their way of background correction, scaling and calculation of the normalization factor (Harr & Schlötterer, 2006; Hoffmann, Seidl, & Dugas, 2002). Normalized data are presented with less variations from chip to chip which can be tested for quality before starting the processing. Post normalization QC must provide a proof for absence of biologically insignificant variations through tests as boxplots, histograms, MA-plots and coefficient of variation vs mean plots (Pepper, Saunders, Edwards, Wilson, & Miller, 2007).

**Data Processing:**

In this section of analysis, normalized data are inspected for biologically significant variations that are statistically significant. Biological significance is displayed by a fold change value. Fold change is a number calculated as the ratio between log expression intensity of a certain gene compared to its normal condition/control expression value. The higher this value, the more
biological significance it is. On the other hand, statistical significance is expressed as a p-value. P-value is defined in this context as the probability of accepting your NULL hypothesis when it should be rejected. The higher the p-value, the least significant it is. Accordingly, gene filtration criteria are based on two cutoffs: the fold change cutoff and the p-value cutoff.

A lot of analysis protocols have been designed to reach these two values significantly. Most famous remains the LIMMA (linear models for microarray data) approach (G. Smyth, Thorne, & Wettenhall, 2005). LIMMA comprises the ability to analyze comparisons between many oligo targets simultaneously in complicated designed experiments. To provide stable results even when the sample size is small, it utilizes Empirical Bayesian methods. This fits the log expression data from each probe into a linear model. Coefficients of these fitted models describe then the differences in sources of oligos hybridized to the arrays. These coefficients are then correlated and scaled in contrast to other experiments. These contrasts are used to compute moderated t statistics and F statistics (as several groups are tested and not just pairwise testing) (Gordon et al., 2012; Jeanmougin et al., 2010) and (G. K. Smyth, 2004).

Applying these statistical tests provide a set of genes that are DE. DEGs are significantly interesting as they present high statistical significance across all experimental conditions tested as well as exhibit a strong over expression or down expression compared to their controls (Scholtens & Heydebreck, n.d.; G. K. Smyth, 2004).

**Systems approach:**
Differential gene expression profiles present a target for many molecular researches nowadays. These profiles place a seed for deep understanding of genotypic changes accompanying a specific condition. Applying the systems approach on these genes facilitates the tracking of their interactions with different genes in a certain organism/ condition. Thus, increasing the data
needed to design a diagnostic approach or even a treatment target. Based on that, systems level of analysis can be applied in to identify molecular targets and pathways involved in medulloblastoma. Not only that, specification of these findings to each of the five subgroups currently identified can provide better diagnosis as well as se base for proper targeted treatment (Kapoor, 2006).

One of the systems approaches utilized in cancer research is to construct a protein-protein/ gene-gene interaction networks. A network consists of nodes (the genes/ proteins/ RNA/ ...etc) that are connected with edges (arrows that can be directional/ non directional) to display the interactions that were previously proven to take place between a set of nodes. Utilizing databases carrying information about the currently identified interactions (as STRING, BIOGRID, KEGG,...etc) is thus needed to construct these networks in order to approach the level complete awareness of the mechanisms involved (Bapat, Krishnan, Ghanate, Kusumbe, & Kalra, 2010; Kapoor, 2006).

Another approach that places itself nowadays is the transcriptional regulatory networks. These networks have arisen with the current demand for understanding the type of regulation practiced on specific genes of interest. In these networks, transcription factors’ (TFs) databases (as the TRANSFAC, JASPAR, SWISSREGULON,...etc) are utilized to obtain information about the binding motif sites related to these TFs as well as the potential interacting genes as well as experimentally proven regulons (Carro et al., 2010a; Janky et al., 2014).

Interestingly, medulloblastoma diagnosis and drug targeting place itself as a fast growing field of research. Thus, a need for better analysis of the molecular level interactions sharing in the progression of this tumor is required. Accordingly, my current thesis places a deep study for
these interactions from the transcriptional regulatory point of view in aim to support a hypothesis of higher transcriptional regulation underlying the altered gene expression profiles.
3. **Chapter 2: Research hypothesis and objectives**

In every research, a hypothesis needs to be placed in order to start your analysis asking the correct questions. The hypothesis set in this thesis is:

*Molecular subgroups have variable gene expression profiles that associate their different clinical presentations and accordingly their response to different treatments.*

Consequently, studying these gene expression profiles to predict gene markers for each molecular subgroup is a major concern in this work. This study needs to be translated to the protein – protein interaction level aiming to allocate putative transcription regulatory elements that possibly underlie these different profiles.

Based on this, the present work aims at five objectives to be reached through a computational and statistical approach:

**Objective 1**
Identify differentially expressed genes for each molecular subgroup in medulloblastoma patients.

**Objective 2**
Specify gene biomarkers for each molecular subgroup in medulloblastoma.

**Objective 3**
Analyze the protein –protein interaction network for each molecular subgroup.

**Objective 4**
Compute the type of transcription regulation applied in each network.
4. Chapter 3: Materials and methods

4.1. The dataset:
The microarray dataset was obtained from the NCBI Gene Expression Omnibus (GEO) under the accession of GSE37418.

The data include microarray expression profiles for the human medulloblastomas from the experiment conducted by Robinson et al. in 2012 at St. Jude Children’s Research Hospital.

4.2. Expression profiling using R:
The raw CEL files of the 76 samples were processed using the R statistical computing software version 3.0.3. (R core team, 2013) The experiment utilized Affymetrix U133 Plus 2.0 microarray chip scans. The expression values were processed in three main steps as follow: (see Figure 45 in appendix)

4.2.1. Preprocessing:
Using the Bioconductor package “affy” (Gautier, Irizarry, Cope, & Bolstad, 2013), images for the scans were first inspected using the “image ()” function to check visually for defects. Then a preprocessing quality control was performed where boxplots; histograms; and RNA degradation plots were generated to visually inspect the behavior and quality of the data. Further quality inspection of the scattering between the fold of the change in the signal intensity against its mean was performed with prenormalization MA plots. Assessment of the quality of the Affymetrix batch was confirmed by using the “simpleAffy” package.
4.2.2. Normalization:
The normalization was done using the MA5.0 algorithm and the expression set was then
linearized using the “log ()” function for ease of handling. Post normalization quality control was
tested through boxplots and MA plots for each array.

4.2.3. Outlier array detection:
To confirm whether all arrays are within an acceptable range of variations, the following three
tests were performed:

Dendrogram:
The Euclidean distance was used to calculate the distance matrix between the arrays. The
clustering between the distance trees was in turn calculated with the complete hieratical
clustering. Finally, the dendrogram was plotted to detected outlier arrays.

Correlation plot:
   a. Average correlation plot:
   
To calculate the average correlation, Pearson’s correlation coefficient was calculated for each
pair of variables for all complete observations.

   b. Correlation plot:

Correlation between each pair of arrays was plotted using the “gplots” package.

   c. Coefficient of variance (CV) versus mean plot:
Mean versus coefficient of variation for each array was plotted to investigate the degree of
variation of each array from the mean.
The array GSM918628 was found to be of a substantially different behavior compared to the other 75 arrays under this study. In order to avoid biases, the questionable array GSM918628 was removed from the data.

4.2.4. Gene filtration:
This is vital step in the analysis and necessitates a checkpoint for the genes’ behavior across the sampled arrays. This test was achieved through a CV versus mean plot for each gene across the different arrays. The plot was used to set a cutoff for filtering those genes below a CV = 0.1. The genes lying below a coefficient of variation cut off of 0.2 shows no significant variation along the different arrays. Filtering those genes improves the analysis sensitivity and decreases the noise. Filtration was performed as follows:

a. PM calls were used to filter out the absent probes after normalization.

b. Probes that showed 20% or more absence in all neighbor arrays based on the dendrogram analysis and previous subgroup classification of the St. Jude team were removed manually.

c. “nsfilter” function was used to eliminate probe signals that are duplicated or of probes with no Entrez IDs and of signals that lie below a variance of 0.2 from its mean signal value.

4.2.5. Data statistical analysis:

4.2.5.1. ANOVA:
Analysis of variance of each subgroup was done against the remaining 4 subgroups to visualize the variation in expression behavior of each gene from group to another. Using the package “limma”, a model contrast matrix was designed for the 5 expression subgroups using the cell means. Average variability across all genes was estimated with the “eBayes ()” for the designed
matrix to adjust high variability genes down and low variability genes up. This function is used to rank genes in order of evidence for differential expression. The moderated t-statistic results were plotted as a quality check point. A t test quantile-quantile plot was used to compare between the theoretical sample quantiles and the original data quantiles considering the degree of freedom to test the variation in results.

4.2.5.2. False Discovery Rate (FDR):

To control false positive results whilst multiple comparison hypothesis testing, FDR using the “fdr” algorithm was applied on filtered data. Stringent FWER methods, as the Bonferroni’s, were avoided to decrease the loss of potential DEGs. For each subgroup, FDR was calculated using the “topTable ()” ranking function, using the F-statistic, at significance level of 0.02 for genes above a log F-statistic cutoff (lfc) of 2; where only genes showing more than a doubled log change were included and a list of DEGs selected for each subgroup was prepared.

To visualize these results, a heat map for each subgroup was prepared using the resultant DEGs’ expression values. The procedure was achieved using the “gplots” package. Using the “dist ()” function and applying the Euclidean algorithm, the pairwise distances between the expression values of the genes across the different subgroups were calculated for each subgroup against all four others.

Supporting these results, “decideTests ()” function was used applying the “fdr” algorithm and filtering out genes above lfc = 2 and p-value = 0.02. “VennDiagram” package was used to plot a venn diagram describing the final statistical interaction between DEGs across all medulloblastoma subgroups on trial.
4.2.5.3. Student t test for specific gene selection:

The student paired t test was used for each gene in a certain subgroup in pair with each of other 4 subgroups. Genes demonstrating highest significance (p value of 0.05 or less) in one subgroup in contrast to all the other subgroups are set to be potential markers for this subgroup.

4.3. Functional Annotation:

Lists of DEGs, their expression values and log fold change (logFC) values together with the significance value of the decision were prepared using the data output from the “decideTests()” function. Using the Database for Annotation, Visualization and Integrated Discovery (DAVID), functional annotation of each list was done. Lists of Ensembl IDs coupled with gene symbols were prepared as an input for the String tool. Lists for putative markers resulting from the pairwise t testing at a p value<0.05 were annotated using the same tool.

4.4. String tool for Protein-Protein interaction:

For each medulloblastoma molecular subgroup, the list of DEG IDs was uploaded. In Homo sapiens, protein-protein interaction networks were constructed using the DEGs list as the core network with an additional 500 white nodes and 500 possible interactors shown for every node. The results were set to highest confidence (>0.9). For each of the five subgroups, a list was prepared for the proteins coded by the DEGs and their protein interactions.
4.5. Cytoscape network visualization:

Constructed protein-protein interaction networks were visualized using Cytoscape to detect significant clusters. Five networks were constructed, one for each subgroup. For each, MCODE (Bader & Hogue, 2003) was used to identify protein clusters. MCODE network scoring included loops and was adjusted to a degree cutoff of 20. The cluster finding fluff and haircut algorithm were both applied. The primary clusters (score>15) for each of the subgroups were extracted and transcription regulatory motifs were detected by employing the iRegulon (Janky et al., 2014), which was applied on networks extracted from DEGs and their first neighbors in each of the clusters.

Transcription factors resulting from the analysis were plotted in networks with their target DEGs to track the type of regulation in each cluster. Collectively, union merge of all transcription regulatory networks of each of the clusters was performed to construct a transcription regulatory network for each subgroup. TF hubs were detected and the type of interaction was concluded.
5. Chapter 4: Results and Discussion

The aim of this chapter is to display the results obtained from the 76 arrays’ data analysis starting from the microarray chip scans quality assessment passing by the specific DEGs detected for each subgroup. The chapter also discusses the protein-protein interaction networks obtained and displays the transcription regulatory networks for each of the subgroups pinpointing the key regulatory transcription factors for each network. The objective of these results is to provide a clear genetic signature for each of the five medulloblastoma molecular subgroups.

The results are obtained from the dataset images uploaded as CEL files on the R software. These had no visual deteriorations as seen in appendix Figure 46. Accordingly, the initial prenormalization preprocessing started with the initial quality control inspection of the arrays. The chips were found of accepted quality regarding their box plots, histograms, MA plots and RNA degradation plots. Normalization was performed by multiplying the values of expression of genes on each chip with a constant factor to be able to compare across the different conditions as one normalized sample.

5.1. Microarray raw data preprocessing:
The preprocessing stage is a preliminary processing of the raw images in hand to convert them into expression intensities that can be handled statistically for producing biologically relevant data (Leung & Cavalieri, 2003). In this stage, the sample quality, hybridization and overall binding quality, signal comparability and biases, and array correlation are checked.
5.1.1. Testing the sample quality:

RNA degradation plots test the overall RNA quality control showing the trend of RNA degradation from the 5’ till the 3’ end. Plotting the degradation mean intensity for each probe for each of the 76 arrays is done from the 5’ till the 3’ end. This originally keeps up with the fact that the degradation of the RNA will start from the 5’ till the 3’. The results do support that and the degradation starts minimal at the 5’ and reach its maximum value by the 3’ end (McCall, 2011). The plot showed that all arrays had a curve with a regular slope of 2.5. Although the slope is relatively high (>1.7) yet all arrays showed the same relatively high slope which supports good quality of samples. Figure 7 The RNA degradation calculation statistics summary is displayed in the following table ordered by the arrays’ names (see Table 1 in appendix)

Another assessment applied is the 3’/ 5’ ratio for beta actin and GAPDH. As long genes that are present in most of tissues, they are used as a control for the RNA quality. Three probe-sets are assigned for each of their 3 positions: the 3’, the mid and the 5’. For good quality, the 3’/ 5’ ratio of beta actin should not exceed 3 and of GAPDH 1.25 (Zahurak et al., 2007). This indicates that the transcripts were labeled equally along the sequence and thus was not cut in middle. In Figure 8, red triangles designate the 3’/ 5’ ratio for beta actin that exceeded the accepted limit the blue triangles lie in the accepted boundaries. Also, the red circles designate for the 3’/ 5’ ratio for the GAPDH lying outside the acceptance boundaries, blue circles are conforming with the quality criteria. These results show that most of the arrays have and accepted 3’/5’ ratio for both housekeeping controls however, some arrays showed inconsistency with standards. This necessitates a normalization step.
Figure 7: RNA degradation plot

Figure displaying results of RNA degradation calculations for the microarray experiment held on the 76 samples. The plot compares the average intensity of each of the probes across all probe sets ordered from the 5’ to the 3’ ends (direction of RNA degradation). The results support this piece of information as the degradation starts minimal at the 5’ and reach its maximum value by the 3’ end. All arrays behave similarly in their RNA degradation behavior with no outliers indicating good quality of experimental procedures.
Figure 8: QC statistics summary(simpleaffy)

Summary for the QC statistics calculated by the simpleaffy package function qc(). The left column of the graph carries the array names (in black) with the values of the percent present probesets (up) and the average background (down) for each array (in red). The dotted horizontal lines in the background of the graph act as array separators. The blue region is the spread. This region carries the scale factors that fall within three fold of the mean scale factor for all arrays. Red lines report for arrays having scale factors exceeding the 3 fold change. The red triangles indicate a high beta-actin 3:5' ratio (greater than 3) while red circles indicate a high GAPDH 3:5' ratio (greater than 1.25). Blue colored triangles and circles indicate that the former ratios are within the accepted range.
5.1.2. Testing the hybridization and overall signal quality:

5.1.2.1. Hybridization spike in controls:

The microarray chips have 4 embedded spike-in control probes as a control spike-in probe-set. These hybridize with the complementary oligonucleotide placed during sample preparation. These are mainly the BioB, BioC, BioD and CreX cRNAs extracted from E. coli biotin synthesis pathway and the Cre is the recombinase gene for P1 bacteriophage. The presence of the BioB signal at minimum of 70% of the chips indicates good hybridization (Huber, Wolfgang; Heydebreck, 2002; McCall, 2011).

The QC stats plot supported that the BioB is present in all arrays (no warnings for absence of BioB in all arrays were detected). This designates good quality of hybridization and overall accepted signal quality (as seen in Figure 8).

5.1.2.2. Background intensity:

Background is the fluorescent area surrounding a fluorescent spot. This fluorescence is not due to hybridization between the probe and its complementary oligonucleotide. However, other interfering factors result in this fluorescence; an example of that is a non specific hybridization or fluorescence due to other fluorescent artifacts (Woo et al., 2004). The background intensity is thus calculated as the difference between the original spot intensity recorded and the foreground intensity calculated using the adaptive circle segmentation (Finkelstein, 2005). The values produced are displayed in table 2. These are an average between maximum and minimum reported background intensity for the same spot. Arrays showing high variation from average background intensity are considered of poor quality.
As seen in Figure 8, the arrays have different variations in their background intensities. These variations are however acceptable showing a standard deviation (SD) of 20.39 from a mean (μ) value of 60.47. The median absolute deviation (MAD) as a robust estimation of the deviation of background intensities from the median is 13.88 thus acceptable. The normally distributed sample will originally display a SD/MAD value of 1.4826. This prenormalized sample displays a SD/MAD of 1.469. MAD is used as a robust estimation of deviation not depending on outliers as it is calculated using Median instead of mean and absolute values are used (Huber, Wolfgang; Heydebreck, 2002; McCall, 2011).

5.1.2.3. Percent gene present:

This quality control parameter is calculated utilizing the MAS5.0 algorithm. A “Present” call is given to probeset hybridized to its targeted transcript regarding the perfect match (PM) versus mismatch (MM) ratio. This indicates that this parameter is most sensitive to error causing factors starting from the RNA sampling and the scanning tills the data analysis including the background adjustment itself (Pepper et al., 2007). Thus, the arrays should not show a deviation outside a range of +/- 10% from the overall array percent present average. Some arrays in the dataset did show a variation above the specified range (45.6-55.8) which suggested potential bad quality in one or more of the replicates displayed in these chips (Figure 8). Table 3 presents the percent present probeset values for each array calculated using the MAS5.0 algorithm PM calls.
5.1.3. Signal comparability and biases diagnostic:

5.1.3.1. Scale factor:
A dataset is expected to have variations in their expression behavior that is due to actual expression levels or due to experimental biases. Accordingly, a normalization algorithm adds a scaling factor to these values in aim to normalize the behavior globally across the dataset. This scaling factor, as mentioned before, should not exceed a 3 fold change from the mean (C. Argyropoulus, 2006; Huber, Wolfgang; Heydebreck, 2002).

The data as displayed in Figure 8 showed some variations in their scale factors of an average of 1.278. A 3 fold variation boundary for the scale factors is thus calculated as (+/-3.834). Basically, most of the SF distribution lies within boundaries; however, some arrays display non conformity. This indicates improper signal distribution that might be due to bad quality.

5.1.3.2. Boxplot of prenormalized data:
Between-array boxplots are plotted utilizing the prenormalized expression values. As expected the results are not identical distribution but still not completely different (Figure 9)

5.1.3.3. Density histograms of prenormalized data:
Between-array histograms are plotted in one graph superimposed. This pot shows the data distribution of each array compared to others. Results should be similar before and after normalization. Normalization is expected to increase the similarity but no change in the data distribution is observed. As seen in Figure 10 the 76 arrays show some deviations suggesting need for normalization.
Figure 9: Boxplot for prenormalized medulloblastoma dataset

Figure showing the 76 boxplots of the medulloblastoma microarray samples aligned before normalization of data with deviations from the average of the expression intensity variations of neighboring arrays.
Figure 10: Histogram for 76 medulloblastoma prenormalized samples
5.1.4. Intensity dependant biases:

MA plot:
This is a pairwise comparison between the log-intensity of each array in compare with a reference median array (M) with the average log-intensity of both arrays (A). MA plot is a scatter plot between the M value ($\log_2 R + \log_2 G$) vs the A value ($(\log_2 R + \log_2 G)/2$). M value represents the ratio between the Red and the Green spots on the array while the A value represent the signal intensity (C. Argyropoulus, 2006; Zahurak et al., 2007). This is scattering is expected to center around the Y-axis value of zero since no significant difference between both values should be observed. The prenormalized plots are of higher scattering compared to the normalized plots. This can be seen in the Figure 47 and Figure 48 in appendix section.

5.1.5. Spatial biases:

2D image raw data images:
Two dimensional images showing the different spot intensities distributed on the array are plotted in this step. The images of the arrays did not show any visual spatial trends or biases that can differ from the raw data. Accordingly, the arrays are assumed to be of good homogeneity and thus quality (Figure 46 in appendix)

5.2. Data Normalization:
As discussed earlier in this chapter, the data in hand does have array to array variations that need to be adjusted. This adjustment aims at cancelling variations in the expression intensity that is not due to biological differences. To make meaningful comparisons between expression intensities across the arrays, a normalization step is thus needed.
The Microarray Analysis Suite 5.0 (MAS5.0) algorithm was applied for normalizing the data. The algorithm achieves the 4 preprocessing aims: a. Background correction, b. Normalization, c. PM correction and d. Expression index calculation yet with no logging of the values. Consequently, logging proceeded (C. Argyropoulus, 2006). Compared to other proposed algorithms (as AD, MBEI, RMA, GCRMA and PLIER), the MAS5.0 has an added value of considering PM-MM average ratios in its calculations whilst alternative algorithms completely ignore the MM values (Pepper et al., 2007). Thus MAS5.0 produces more variations in results yet more reliable calculations. This can be seen from the post normalization QC tests practiced on the normalized data.

5.3. Evaluation of preprocessing:

5.3.1. Signal comparability and biases diagnostic:

5.3.1.1. Boxplot of normalized data:

The boxplot showed closer distribution of the arrays around a median intensity for all the 76 arrays. Although some slight variations for some arrays can still be witnessed, yet this is expected when utilizing MAS5.0 as it considers a PM-MM average ratio which present more variations compared to other algorithms. Figure 11 and Figure 12 display the difference between the normalization by MAS5.0 and RMA algorithms.

5.3.1.2. Density histogram of normalized data:

Normalized data of the 76 arrays perfectly superimposed. This demonstrates a successful normalization. Studying the histogram Figure 13, the data are skewed slightly to the right were you can see a tendency of the genes to be over expressed. The shape of the plot is not perfect bell shaped as expected to be after normalization, but still close enough.
Figure 11: Boxplot for logged and normalized data using MAS5.0 algorithm
Figure 12: Boxplot for normalized data using RMA algorithm
Figure 13: Post normalization histogram for medulloblastoma sample
5.3.1.3. MA plots of normalized data:
After normalization, plotting the log intensity of each array to the reference median array against the average log intensity of each displayed a normalized scatter plot. The data scattered around a regression line that is of zero M value. This supports that for each array, the difference between the average log intensities of expression in compare with the ratio between the expression intensity of that array and the median expression value is minimal approaching zero (Figure 48).

5.3.2. Array correlation:
In attempt to discover arrays that show a different expression behavior compared to the rest of the variables (arrays), array to array correlation analysis is conducted. The correlation analysis aims to discover the association between variables in hand, in this case the arrays. An outlier array will fail three out of four correlation tests: a. correlation plot, b. Hierarchical clustering, c. Coefficient of variance and d. PCA analysis.

5.3.2.1. Correlation plot:
The plot is designed for each array in contrast with the 76 arrays available. The correlation coefficient is calculated utilizing the Pearson’s correlation equation using the “pairwise.complete.obs” algorithm in its calculation.

\[
r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}
\]

(http://www.statmethods.net/stats/correlations.html/)
Accordingly, the correlation matrix was designed for the 76 arrays where each array was compared with the other arrays from all complete pairs of observations (genes) on these variables (arrays). Figure 14 shows the correlation matrix plotted for the 76 arrays.

Observed is a low correlation between array GSM-918580-mbt-006 and some of the 75 arrays (excluding 29 arrays that show some correlation represented by the dark blue color \((r= 0.84)\) and 3 arrays that show higher correlation represented by the brown color \((r=0.87)\)). Array GSM-918628-mbt-136 shows even lower correlation with almost all 75 remaining arrays, making it a potential outlier array.

On the other hand, an average correlation plot displays the correlation coefficients dispersion around a correlation mean calculated as a simple average of the correlation between each array and the other. The plot in Figure 15 displayed a scattering around the average correlation \((\text{Avg} \; r)\) yet array GSM-918628 showed highest scattering as well as array GSM-918580, supporting my observations. Thus GSM-918628 failed the correlation test. Also, GSM-918580 did fail this test.
Figure 14: Correlation plot for 76 normalized arrays

The figure shows a correlation plot for the 76 arrays. The degree of correlation between each two arrays along the whole sampled arrays is displayed. The yellow dots represents highest correlation (r=1) where the correlation is between the array and itself. The light blue dots represents lowest correlation (r=0.7) detecting arrays of lowest correlation to neighboring arrays. Array GSM918628 is of lowest correlation to all other arrays supporting its being an outlier.
Figure 15: Average correlation plot for 76 medulloblastoma arrays

The average correlation plot: each red circle represents the correlation coefficient between each sample and the mean correlation of all samples. All average correlations lie within the same range except for the 1 array, GSM918628.
5.3.2.2. Hierarchical clustering:
Distance trees utilizing the Euclidean algorithm for distance calculation were clustered by different clustering algorithms. Figure 16 shows that in all cases, the array GSM-918628-mbt-136 was clustered alone in a cluster far from the rest of the arrays. Also Figure 16 shows that distance trees, constructed by different algorithms (Manhattan Minkowski, maximum likelihood and Canberra) and clustered by complete clustering, give same results.

It is worth noting that GSM-918653-tbm-143 is also clustered in a relatively far cluster even after changing the distance tree construction algorithms. But still it is clustered in big clusters. Accordingly, GSM-918628 failed the clustering test, whilst GSM-918653 showed unclear evidence. To confirm these results, the remaining two tests needed to be approached.
Figure 16: Different distance trees clustered by complete clustering.

Figures showing different distance trees for the 76 arrays under study clustered by the complete clustering algorithm. The distance trees were structured utilizing the following algorithms consecutively: a. Manhattan, b. Minkowski, c. Maximum likelihood and d. Canberra
Figure 37 (continued next page)
Figure 17: Different clustering for Euclidean distance tree for 76 arrays.

Figures showing different clustering methods applied for the 76 arrays under inspection. The clustering methods displayed are for the distance tree calculated by Euclidean algorithm. These methods are by applying the following algorithms respectively: single and ward.
5.3.2.3. Coefficient of variance:

Coefficient of variation is the ratio between the standard deviation of a variant and the overall mean value. \((CV = \frac{SD}{\mu})\) Thus, it is a standardized measure of dispersion of a frequency distribution. Accordingly, the coefficient of variations of each array is plotted against the mean expression for each. This shows that most arrays show closer deviations to the rest of arrays with GSM-918628 plotted at the extremity away from this array. GSM-918653 is observed closer to the neighboring arrays fading the postulation proposed of it being a potential outlier (Figure 18).

![Coefficient of variation vs. mean plot for the 76 arrays.](image)

**Figure 18: CV versus mean plot for the 76 arrays.**

The figure shows plot of coefficient of variation (CV) against mean calculated for each of the arrays. The array GSM918628 is of highest variation from the mean in relation to other arrays supporting its being an outlier.
5.3.2.4. PCA analysis:
Principal component analysis is an algorithm applied on the logged and normalized dataset to calculate the main components varying in the data by reducing its dimensionality. The variables in this test are the different arrays while the observations are the gene expression values. PCA aims at identifying the “pcaₙ” where n= number of new principal variables in your experiment. The pca is always less than N of the sample, where N is the number of all variables available in the experiment (Affymetrix, 2002; Raychaudhuri, Stuart, & Altman, 2000). PCA produces a list of the rotations applied to the data points in order to redistribute them into a less dimensional space (n<N) with their standard deviations and proportion of variance.

Applying PCA on the 76 arrays displayed 7 principal components with the first 4 being of higher frequency (more influential). One can subdivide the data into 4 main dimensions and the fifth being more or less unspecified with reference to these results. Compared to the clustering discussed earlier in this chapter, the results are consistent with the 5 major clusters obtained after applying different clustering algorithms (Figure 19). A deeper look in the distribution of the rotations applied on the data point flashed the fact that high rotation factor is applied on arrays: GSM-918628 and GSM-918580 in relation to remaining arrays. This supports the hypothesis proposed that GSM-918580 is an outlier array that should be removed from the dataset before the data processing. Also, GSM-918653 cannot be considered an outlier as it succeeded two out of the four tests. Figure 20 further support for the decision. Thus, GSM-918628 is removed from the dataset. The new logged and normalized dataset is quality accepted and ready for processing.
Figure 19: PCA for 76 medulloblastoma samples

Figure showing 7 principle components in the data normalized. The last 3 components plotted are of almost zero contribution. Also, as seen in figure, the first 4 components form recognizable behavior. This supports the clustering protocol applied showing 4 main clusters and the fifth of minimum significance. Supporting these findings, medulloblastoma can be subdivided on the molecular level into 4 significant subgroups with the fifth subgroup being unspecified.
Figure 20: Correlation between PCA1 and PCA2.

Figure displaying a correlation analysis between the degrees of rotation applied on the variance values of each of the first two components. As observed, the two arrays GSM-918580 and GSM-918628 have the most degree of rotation applied on them. This is translated as a high variation for both arrays compared to the remaining.
5.4. Microarray data processing:
In this section, normalized data obtained from previous steps are processed to obtain information about the different expression profiles of the genes. To achieve this, downsizing the number of the genes being studied by filtering out genes having stable expression profiles and keeping the DEGs is done. Several filtering approaches have been developed during the past few years; however selection of the method is completely dependent on the scope of the study and the type of the dataset. Had the data been filtered, annotation is to proceed in order to provide interaction with actual biological information. Several statistical tests utilizing the annotated data are performed to provide answers to the proposed hypothesis presented at the beginning of the research. Results from statistical analysis either lead to supporting or rejecting the hypothesis proposed.

5.4.1. Gene filtration:
As previously discussed, the dataset consists of 75 arrays that carry expression profiles for 54,675 probes representing almost 40,000 genes (Woo et al., 2004). Trying to obtain a biological relevance between genes of interesting expression profiles and each molecular subgroup seems really hard to achieve at this level. This is due to the fact that not all genes are to be expressed at biologically meaningful levels as most tissues will express only 30-40% of their genes. Also, not all these genes do play a role in each experimental condition (Calza et al., 2007). Accordingly, different approaches have arisen to filter out genes that do not play a significant role in each of the experimental conditions (arrays from different molecular subgroups in this case). These approaches are basically dependant on the variation in the gene expression intensities between the different conditions, or average signals, or the detection calls identified
by the MA5.0 algorithm in Affymetrix chips or even a combination between the previously mentioned approaches. The dataset is filtered using a combination between the detection calls and the variance filtrations (Archer & Reese, 2010). Using the Wilcoxon sign test, probe sets are assigned a P (present), M (marginal) and A (absent) (Hackstadt & Hess, 2009). This is the detection calls approached used to filter out the absent probes. This produced a matrix carrying 47,238 probes instead of the 54,675. In other words, 7437 probes were completely absent in this dataset. For each probeset, a limit of 20% presence across all arrays was assigned. A probeset displaying a P call less than 20% was filtered out. Using this data matrix, a CV versus mean plot for the gene expression values across all arrays is plotted to check for the different variances of each gene. This produced a left skewed bell shaped distribution of CVs (Figure 21).

The distribution peak represents genes showing highest variations (genes of interest) while those at the bottom represent genes of little or no variation. A variance cutoff for the gene expression intensities of 0.2 was applied as a filter for genes of low variance across arrays. Filtration was absolute; this indicates that data variances were ranked by utilizing absolute variance values and not considering the variance values as a quantile value.

The final step of filtration included the filtering out of all duplicated probesets. This produced a ready to analyze dataset matrix carrying 75 arrays with 19,844 probes showing significant variance.
Figure 21: CV versus mean plot for each gene across the 75 arrays

Top: To set a variance cut off for filtration CV vs. mean plot was plotted for each gene across all 75 arrays. Accordingly, a variance corresponding to a CV cut off of 0.2 was set as a filtration criterion to filter out the genes that showed no significant variation (below a CV of 0.2) compared to statistically significant genes lying above the threshold. Bottom: a CV vs. mean plot for each gene across 75 arrays after PM calls is displayed showing a CV cutoff of 0.2 as a filtration line.
5.4.2. Statistical analysis:
The data in hand is a matrix filled with numbers corresponding to expression intensity of each gene across the different arrays. Statistical analysis is a critical milestone where these expression values are tested statistically to obtain information about their performance throughout different experimental conditions and to pinpoint genes that are differentially expressed within each subgroup/condition.

5.4.2.1. ANOVA:
This dataset is a multivariate sample of medulloblastoma patients. Accordingly, statistical approaches must consider the different conditions in hand; thus, ANOVA is a first choice in the analysis. Linear models for microarray data (LIMMA) can be applied in this case to linearize the multivariation (Jeffery, Higgins, & Culhane, 2006). A designed model matrix for the data is constructed with the gene expression profiles are structure for each subgroup against the average of the remaining 4 subgroups. Empirical Bayesian statistics (ebayes) are applied on this contrast matrix. The ebayes shrink the probe wise variances towards a common value (G. K. Smyth, 2004). This is done by applying a moderated t statistic test (F statistic) on each gene in the contrast matrix. This matrix is structured as arbitrary numbers that are tested to equal to zero. The F statistic applied is an ANOVA but differs only in that the residual mean squares and residual degrees of freedom have been moderated between the probes to linearize the model. A moderated t statistic plot can be seen in Figure 22.
Figure 22: Student t quantile-quantile plot

The plot shows the proposed moderated versions of the t-statistic by Smyth (2004) where the original student’s t test statistic distribution (sample quantiles) is compared against expression intensity-specific variance after being augmented by a constant that is derived from the data of all variables (theoretical quantiles). As seen both variations show similar quantiles and are aligned around the regression line (red line).
5.4.2.2. FDR:

Each coefficient calculated in the contrast matrix corresponds to a subgroup. For each subgroup, a risk for false positives and negatives is expected as the general variance ranking is assessed across a big sample number (19,844 genes). This ranking depends on adding a shrinkage factor that controls the general variance across all genes with respect to the individual variance level of each. This leads to high risk of false positives and negatives (Type I and II errors). So to make a proper inference (conclusion about the hypothesis using unobserved data), these inferential errors need to be adjusted. The FDR correction method is thus applied. This method is more practical than family wise error rate (FWER) correction method; as the later produces results of very low error rate yet the risk of losing some valuable candidates is more likely, accordingly, FDR is more preferable however error prone it might be (Hackstadt & Hess, 2009). Thus the very basic algorithm FDR (recently updated to the Benjamini-Hochberg “BH”) is applied on the bayesian matrix (Benjamini, D, Hochberg, 1995).

5.4.3. Filtration:

A final filtration step for candidate DEGs is done utilizing statistical approaches. Genes expressing a fold change of more than or equal to 2 and of a significance value less than 0.02 are rendered in the list of the DEGs (Dalman, Deeter, Nimishakavi, & Duan, 2012). In other words, genes of a p-value of <=0.02 and a log expression value that is double or more than double the mean expression value for this gene (whether by increase or decrease) are set as DEGs. A heatmap for each subgroup can be seen in figure. Heatmaps display the list of DEGs of this subgroup plotted against the other subgroups. Clustering using complete algorithm for a distance tree calculated by the Euclidean algorithm is performed. Clear DEGs profiles can be observed for each subgroup in compare to the rest (see Figure 23, Figure 24, Figure 25, Figure 26, Figure 27 and Figure 28).
Figure 23: Heatmap for the 326 DEGs in the WNT subgroup.
Figure 24: Heatmap for the 463 DEGs in SHH subgroup.
Figure 25: Heatmap for 651 DEGs in Group 3 subgroup.
Figure 26: Heatmap for 106 DEGs in Group 4 subgroup.
Figure 27: Heatmap for 670 DEGs of Unspecified subgroup.
Figure 28: Heatmap of 670 DEGs in all medulloblastoma subgroups collectively.
5.5. Annotation:
The expression profiles resulting from previously performed tests belong to probe IDs not yet annotated. DAVID was used for functional annotation of these probes. A list of annotated DEGs for each subgroup is provided in appendix table 4.

5.6. Specific gene markers detection:
A multivariate analysis utilizing ANOVA test was performed on the dataset to obtain the significance of each expression value across the different subgroups. A long list of DEGs was obtained at this level. These DEGs are still questioned for specificity for each subgroup. So, in this section, a different approach is applied. Each gene expression profile is analyzed for specificity across the different subgroups utilizing a simple comparative two sided Student t-testing. The results obtained are vectors carrying the corresponding p-values.

For WNT subgroup, no specific gene can be detected with the resulting p-values; however, a putative specific gene that shows specificity for the WNT in compare to other subgroups except for the SHH is the BTBD17 (The gene codes for the protein BTB (Galectin-3-Binding Protein-Like) /POZ domain containing protein 17 that is secreted in the extracellular region to play a role in transcriptional regulation as well as down regulation of viral genome. The gene is located on chromosome 17q. Utilizing the chromosome 6 monosomy previously recorded in WNT patients still sounds as a better mode of diagnosis for these patients(Ellison et al., 2011b; P. A. Northcott et al., 2011; Remke et al., 2011; Thompson et al., 2006a).

The SHH subgroup on the other hand showed 19 genes that are specifically expressed in this subgroup (ano2, TTC8, CABP2, NDP, OTX2, RAGE, , , TT8K1, CCDC3, TRAC/ V2, RhPn2, SUN2, GAB1, PP2R2B, DAB1,). Out of this list, SUN2 (Sad1 and unc-84 -containing domain 2) is
specifically deregulated in SHH subgroup. The SUN2 has not been previously reported in relation to cancer, however, a study by (Hsieh et al., 2014) was conducted to relate between mir221/222 and neuroectodermal embryonic tumors suggests that suppressing the SUN2 by the miRNAs will propagate atypical teratoid/rhabdoid tumor (AT/RT) tumors but not medulloblastoma specifically. They also suggest the role of SUN2 in regulating cell proliferation and tumor malignancy in embryonic cells. Their results support our finding that over expression of the SUN2 plays a key role in developing the SHH subgroup. Also, CCDC3, the coiled coil domain 3, is down regulated associated with poor prognostic osteoblastoma patients. This matches the situation in this analysis with medulloblastoma SHH subgroup, suggesting further experimentation of the CCDC3 potential as marker for SHH subgroup (O'Donoghue et al., 2010). SUN2 remains under query with no significant over expression in this subgroup relative to other subgroups.

On another hand, subgroup 3 arrays showed no significant genes in compare to other subgroups at a p-value cutoff of 0.05; yet raising the cutoff till 0.052 provided one single specific gene expressed in this subgroup only. The NPW (neuropeptide W) is a signaling peptide involved in neuronal signal transduction when secreted. It also plays a role in G-coupled receptor protein signaling pathway in hypothalamus(Singh & Davenport, 2006). No clear relation with the role of NPW in cancer progression has been recorded. A suggested mechanism arises from the fact that this neuropeptide is a ligand for GPR7 and GPR8 (Takenoya et al., 2012). GPR7 is a receptor expressed mainly in the cerebellum and frontal cortex (Brezillon et al., 2003). This can relate to the significance of the NPW to a medulloblastoma subtype.

For subgroup 4, 3 ribosomal proteins were specifically expressed namely: rpl24 and its pseudogene 6, rpl29 and its pseudogenes 4, 6 and 11 and rps7 and its pseudogene 19, 10. As pseudogenes, these are non functional version of functionally present genes that arise from
either duplication events or frame-shift mutations (Balakirev & Ayala, 2003). These non-functional genes are believed to play a regulatory role in gene expression by either repressing the transcription by competitive interaction on transcription factors or DNA-protein or protein-protein interaction (Kalmoykova et al, 1998). Taking into consideration that subgroup 4 patients are known of high incidence of tetraploidy and duplication events, it is expected to have associated copy number aberrations (CNAs) in their genes. Of these events is a high frequency SCNAs on chromosome 6q24.3, 17q22, 2p25.3 and 3p21.1 (P. a Northcott, Shih, et al., 2012). These regions map for locations of the pseudogenes RPL29P4 and RPS7P11 and genes RPS7 and RPL29 respectively (De Bortoli et al., 2006; P. a Northcott, Shih, et al., 2012) Specific expression of these genes can be translated as an underlying driving factor for altered expression behavior witnessed in this subgroup.

Finally, applying the same test on subgroup 5, with changing the p-value cutoff to 0.01, provided a list of 16 genes that are specifically expressed in this subgroup compared to others displayed in the following table:
<table>
<thead>
<tr>
<th>Probe_ID</th>
<th>Gene_symbol</th>
<th>Species</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>218910_AT</td>
<td>Ano10</td>
<td>Homo sapiens</td>
<td>anoctamin 10</td>
</tr>
<tr>
<td>1559412_AT</td>
<td>C21orf34</td>
<td>Homo sapiens</td>
<td>chromosome 21 open reading frame 34</td>
</tr>
<tr>
<td>209832_S_AT</td>
<td>Cdt1</td>
<td>Homo sapiens</td>
<td>chromatin licensing and DNA replication factor 1</td>
</tr>
<tr>
<td>214079_AT</td>
<td>Dhrs2</td>
<td>Homo sapiens</td>
<td>dehydrogenase/reductase (SDR family) member 2</td>
</tr>
<tr>
<td>236523_AT</td>
<td>LOC285556</td>
<td>Homo sapiens</td>
<td>hypothetical protein LOC285556</td>
</tr>
<tr>
<td>226066_AT</td>
<td>MITF</td>
<td>Homo sapiens</td>
<td>microphthalmia-associated transcription factor</td>
</tr>
<tr>
<td>244444_AT</td>
<td>PKD1L2</td>
<td>Homo sapiens</td>
<td>polycystic kidney disease 1-like 2</td>
</tr>
<tr>
<td>233342_AT</td>
<td>LOC100134040</td>
<td>Homo sapiens</td>
<td>hypothetical LOC100134040</td>
</tr>
<tr>
<td>241382_AT</td>
<td>Pcp4l1</td>
<td>Homo sapiens</td>
<td>Purkinje cell protein 4 like 1</td>
</tr>
<tr>
<td>238788_AT</td>
<td>LOC494150</td>
<td>Homo sapiens</td>
<td>prohibitin pseudogene</td>
</tr>
<tr>
<td>207282_S_AT</td>
<td>MYOG</td>
<td>Homo sapiens</td>
<td>myogenin (myogenic factor 4)</td>
</tr>
<tr>
<td>63009_AT</td>
<td>Shq1</td>
<td>Homo sapiens</td>
<td>SHQ1 homolog (S. cerevisiae)</td>
</tr>
<tr>
<td>244680_AT</td>
<td>GLRB</td>
<td>Homo sapiens</td>
<td>glycine receptor, beta</td>
</tr>
<tr>
<td>236024_AT</td>
<td>gpm6a</td>
<td>Homo sapiens</td>
<td>glycoprotein M6A</td>
</tr>
<tr>
<td>208574_AT</td>
<td>SOX14</td>
<td>Homo sapiens</td>
<td>SRY (sex determining region Y)-box 14</td>
</tr>
<tr>
<td>238878_AT</td>
<td>ARX</td>
<td>Homo sapiens</td>
<td>aristaless related homeobox</td>
</tr>
</tbody>
</table>
Functional annotation of these putative markers displayed their role in several regulatory cycles varying from: regulation of RNA metabolic process (SOX14, MITF, and MYOG), ion transport (Ano10, GLRB, GP M6A, and PKD1L2) and cell cycle regulation (CDT1 and DHR52) as well as neuron development (ARX). These several functions play a key role in cancer progression yet no clear underlying mechanism of how these markers can display a characteristic expression profile for this subgroup. It is worth mentioning that SOX14, ARX, MITF and MYOG all play a role in both regulation of RNA metabolic process as well as DNA-dependent regulation of transcription. MITF and ARX also play a role in cell proliferation regulation supporting their role in tumor development.

As described in this section, several markers can be related to each of the different medulloblastoma subgroups described in this dataset. However, added efforts need to be spent to discover the type of interaction between these specific genes and the expression profile related to each subgroup. In trial to answer this research question, protein-protein interaction networks are constructed to observe the suggested mechanisms. In the following two sections, a detailed explanation is provided for the results from these networks.

5.7. Protein-protein interactions:
Genes function primarily by being expressed into proteins. To have a better understanding of how the gene regulation is maintained in each of the five subgroups, a better look on the protein behavior was taken. The STRING tool databases (Franceschini et al., 2013) provide a source for the different types of interaction (direct or indirect) between proteins whether known from previous experiments and co-expression knowledge or predicted through a literature review or genomic context utilization. DEGs for each subgroup are uploaded on the
tool and their interactions in Homo sapiens are mapped in a protein–protein interaction network. These interacting nodes are further mapped for additional 500 nodes at highest confidence (> 0.9) to help visualize alternative interactions that might play a role in regulating the DEGs. Downloaded interactions’ files are uploaded as input files for Cytoscape (Cline et al., 2007) software to visualize the networks and perform advanced modeling.

**WNT subgroup protein-protein interaction network:**

The WNT network comprised 1129 nodes with 9511 edges connecting between them. Applying MCODE clustering algorithm, the network included 38 main clusters, 4 clusters of which scored above 15. The first cluster comprised 1039 nodes (91 are proteins coded by DEGs) connected by 9839 edges with a score 18.056. The second has 759 nodes connected by 7916 edges with a score 20.832. The third included 97 nodes with 821 edges scoring 16.755 and the fourth included 98 nodes with 776 edges scoring 15.677.

Each of the DEGs in a cluster is visualized in a new network utilizing their coded proteins and their first neighbors. These networks demonstrate the first interactors with the DEGs of the WNT subgroup in each cluster.

**Cluster 1 (score: 18.056):**

91 proteins coded by DEGs are directly connected to 529 proteins. The top over expressed genes are (FOXG1, GABRA5, NPR3, TTR, PRL) while the most under expressed proteins are (PAX6, BOC, UBASH3B, BMP4, WIF1, ADCYAP, RUNX2).

It is observed that in general, proteins coded by DEGs in this cluster belong to the following pathways: WNT signaling (6), SHH signaling (4), MAPK (11), TGF beta signaling (4), calcium signaling (4), neuroactive ligand receptor (6) and cancer pathways (10).
Cluster 2 (score: 20.832):

39 proteins coded by DEGs are directly connected to 339 proteins. The top over expressed proteins are (TTR, PRL, PDC, NPW, DOK6) while the least expressed proteins are (UBASH3B, ADCYAP, ADRB1, RUNX2)

The proteins belonged to the cancer pathways as colorectal cancer and prostate cancer as well as general cancer triggering proteins. Again, the Hub of this cluster is the UBC interacting with same proteins coded by DEGs.

Cluster 3 (score: 16.755):

42 proteins are included in the analysis and are directly connected to the only protein coded by a DEG: MYC (up regulated). The MYC plays role in WNT signaling, TGFβ signaling, MAPK signaling as well as cell cycle control.

The main hub (highly connected node in a network) in clusters 1, 2 and 3 in WNT subgroup protein-protein interaction networks is the UBC (ubiquitine conjugating protein E2). (see figure31)
Figure 29: UBC network in cluster 1 of WNT protein-protein interaction network.

APP: Amyloid beta (A4) precursor protein, AR: Androgen receptor, NGFR: nerve growth factor receptor, NEUROD1: neurogenic differentiation D1, NOS1: neuronal nitric oxide synthase 1, PPARGC1A: peroxisome proliferator-activated receptor gamma, coactivator 1 alpha, RUNX2: runt-related transcription factor 2, UBASH3B: ubiquitin associated and SH3 domain containing, B.
SHH subgroup protein-protein interaction network:
The SHH network has 1238 nodes connected by 16602 edges. Out of the 1238 nodes, 233 were DEGs. Clustering of the SHH subgroup network provided 25 clusters, 5 of them are above score threshold of 15.0. The cluster sizes are very big requiring a downsizing for better understanding of the behavior of the DEGs. Accordingly, the sub networks were constructed utilizing the DEGs of each cluster and their first neighbors.

Cluster 1 (score: 31.266):
743 nodes connected through 11631 edges construct this network. The network includes 84 proteins coded by DEGs with the most over expressed FOXG1, PRL, KCNA5 and HTR2C. The most under expressed proteins are the RUNX2, BMP2, BMP4, COL4A5, COL5A2, COL6A3 and GLI3R.

This behavior ascertains the up regulation of the SHH pathway with the down regulation of the GLI3 repressor (Villavicencio et al., 2000), down regulation of BMP2/4 (a case observed only in the neural tube development) (Yandig, 2000) and FOXG1 up regulation (Danesin et al., 2009).

Up regulated genes coding for proteins clustered in this cluster only are: HTR2C, GRM1 NID2, CHRM1 and the COL family. The genes are involved in tumor growth through increasing vascularity. This vascularity is supported by collagen (COL) over expression in connective tissue, and nidogen (NID2) as a cell adhesion protein that increase collagen binding and ion transfer (Anderton, Lindsey, Lusher, Ellison, & Clifford, 2005; Y. Liang, Diehn, Bollen, Israel, & Gupta, 2008). These genes are highly up regulated compared to DEGs in this cluster. Supporting this finding, down regulated genes unique to this cluster are the ANGPTL1, CCD3, NPNT, SYK and HORMAD1 (Roussel & Hatten, 2011; Teicher et al., 2012).
Adding to these findings, MYC is found under expressed in this cluster (and the other cluster of SHH) and of highest connectivity (k=90). Down regulation of MYC is proposed to coincide with lack of nutrients and/or oxygen. The nutrients for a cancer cell include either glucose or glutamate. This suggests that the over expression of GRM1 (Metabotrophic glutamate receptors) can be due to a scarcity in cell glutamate suggesting a malfunctioning GRM system due to higher TF regulation (Nord et al., 2014).

**Cluster 2 (score: 29.163):**

In a cluster of 496 proteins connected by 7274 edges, only 57 were coded by DEGs. This cluster is considered a subset of the first cluster.

Cluster 3 (score: 29.157), cluster 4 (21.074) and cluster 5 (17.029) are all subsets of the first cluster, thus no further discussion is needed.

All five clustered displayed the MYC as the hub (see Figure 30). The main hub for the whole network however is the UBC as seen in Figure 31.
Figure 30: Myc as main hub in SHH subgroup protein protein interaction network

Figure 31: UBC as main hub for entire protein interaction network in SHH subgroup.
**Subgroup 3 protein-protein interaction network:**

Subgroup3 network comprises 1253 nodes, 249 are proteins coded by DEGs. The network carries 10128 edges connecting between the nodes. The clustering resulted in 24 clusters only 2 lie above a clustering score of 15. It is worth noting that genes involved in activation of WNT signaling are not witnessed up regulated (FZD10 down regulated, WNT5A down regulation, LEF1 down regulation) except for the DKK1 and WIF1 (still show down regulation). The later’s under expression still can’t provide WNT pathway activation with the crucial down regulation of the frizzled receptors on the cell membrane responsible for receiving their signal. In addition to this, SHH pathway proteins are under expressed (BMP4/7, PTCH1/2) and extreme over expression of SHH inhibitor HHIP (Sengupta et al., 2012).

**Cluster 1 (score: 22.054)**

The cluster is structured of 371 nodes connected by 4023 edges. It is clear in this subgroup that all genes are highly connected (with an MCODE_K value= 13.4 till 0.00 for end nodes). 54 DEGs coded for 54 proteins constructing this network. The maximum expressed gene is the PPP2R2C, SFRP1, BCHE and CALCB showing increased expression above 5 fold increases. While minimum expression values below 4 folds decrease are witnessed in FOXG1, UNC119, TTR, DAB1 and CAMK2B.

PP2R2C and CXCR 4 are up regulated in this subgroup cluster, although SHH pathway is recorded deactivated. This activation has been experimented as a novel subtype of SHH medulloblastomas where it augments the SHH activation to increase cell proliferation to maximum. This can’t be supported here as the SHH pathway is deactivated. But it provides a new basis for defining this subgroup. CXCR4 suppresses expression of the regulatory unit of
phosphatase PP2A (PP2R2C) that enhances proliferation. (Sengupta et al., 2012) On the other hand, SFRP1 decreases intracellular beta-catenin levels and thus inhibiting WNT induced transcription signaling and overall effect of decreasing proliferation (Kongkham et al., 2010b). Increased level of SFRP1 is also known to inhibit SHH pathway in correlation to BCHE increased levels (Schwalbe et al., 2011). FOXG1 is down regulated unlike other subgroups providing marker identification for this subtype. Also, transthyretin (TTR) is a protein responsible for lipophyllic molecules transport. It is expressed in benign gliomas and meningiomas as well as in retinal pigment epithelium. Down regulation of these genes has been observed in medulloblastoma and causes an associated retinopathy (Albrecht, 1995). DUSP2 down regulation is associated with cancer progression by decreasing epithelial mesenchymal transition (Lin, Hsiao, & Tsai, 2013).

Cluster 2 (score: 17.894)

The cluster comprises 280 proteins with 44 expressed by DEGs. Interconnection is through 2398 edges. Highly expressed proteins are the PRLR, BCHE and KIF26A. prolaction hormone receptor (PRLR) up regulation induces tumor growth by activating MAPK/JAK signaling and thus activating the PRL hormone signals (Plotnikov, 2009). But this also caused lower levels of the serum Prolactin hormone (PRL) that is processed into vasoinhibins that are accumulated in retina inducing the retinopathy (Arnold et al., 2010).

On another level, MET is up regulated. This is a proto oncogene that functions to bind GAB1 (upregulated) inducing a signaling RAS/ERC cascade leading to a variety of cancers (Sachs et al., 2000) The poor prognosis of these patients might be accredited to the down regulation of the CRABP2 gene. CRABP2 gene (cellular retinoic acid binding protein 2) binds retinoic acid and
delivers it to retinoic acid receptors and fatty acid biosynthesis protein 5 for processing. Absence of this mechanism triggers cell proliferation and tumor growth (Győngyösi et al., 2013).

The subgroup 3 protein interaction network displays the UBC as hub. Network in Figure 32 provides support for the important role of the UBC in regulating the gene expression patterns in this disease as a whole and subgroup in specific.

Figure 32: UBC hub in subgroup 3 protein-protein interaction network.
**Subgroup 4 protein-protein interaction network:**

The network includes 1049 protein nodes interacting through 12494 edges. Higher interactions are witnessed suggesting expected strong progression of symptoms, which is the situation in these subgroup patients. 49 DEGs code for proteins in this network. Despite the high connectivity between the nodes, proteins coded by DEGs are separated, except for 3 clusters: troponin regulation (upregulation), centromeric complex CENPK (down regulation) and S phase kinase associated protein 2 (up regulated with corresponding down regulation of FOXM1 and CDT1). The differential expression drives the observations towards involvement in muscle development, cell structure from actin binding to actin cytoskeleton to sarcomere functioning. Studying the network displays a significant over expression in muscle contraction and movement proteins. The genes coding for troponin I type 1 and 2 together with the actin alpha-1 are located on the 1q32:42 chromosome and are up regulated with tropinin I type 2(chr: 11p15.5) and tropnin T type 1 (chr: 19q13.4). There is no clear mechanism why this up regulation is associated with this phenotype, however records of over expression TNNT1 in ovarian serous papillary tumors involved in angiogenesis have been documented (Bapat et al., 2010; Santin et al., 2004).

Regarding the negative regulation of apoptosis, it is lead by 6 genes. The BarH-like homeobox 1 (BAHL1) and the NEUROD1 are under expressed while the S-phase kinase associated protein, the dehydrogenase reductase member 2 and the keratin 18 as well as the Tumor necrosis factor alpha-induced protein 8 are over expressed. Absence of BAHL1 expression in mice activated the NT-3 (neutrophin 3) in cerebellar granule tissue leading to increased death of these cells during embryogenesis. Added to that, the loss of NEUROD1/BETA2 in cerebellar tissue accompanies the loss in the granule numbers with observed granule cell death in the posterior compartment.
These suggest a developmental defect in the cerebellum of a subgroup 4 medulloblastoma patient.

Positive regulation of transcription is led by over expressed GLI3, BMP4, MYOD1, MYOG and TNNI2 and the under expressed CENPK and NEUROD1. No record have been established discussing the role of CENPK in tumor progression, yet its absence does interfere with generating interactions with microtubules (Cheeseman, Hori, Fukagawa, & Desai, 2008).

21 clusters can be detected in the subgroup 4 network utilizing the MCODE algorithm. Four of them scored above 15 (MCODE clustering score cutoff).

**Cluster 1, score=47.078:**

The cluster comprises 176 proteins with 10 coded by DEGs. The nodes are interconnected utilizing 5478 edges. This network is a highly dense structure due to high connectivity between its proteins. The cluster proteins are involved mainly in cell cycle regulation as well as DNA replication. But mostly belong to cancer pathways. In this context, DEGs involved in these expression patterns are: only up regulated genes [DLK1 (6 folds), XPO4 and SKP2 (2 folds)] and extreme down regulated genes [CENPM (4 folds), FOXM1 (3.4 folds) and CDT1 (3 folds)]. In other words, down regulation of Cell cycle controlling genes was observed in subgroup 4 (Cheeseman et al., 2008; Chen et al., 2013; Xouri et al., 2004) supporting the observation that this subgroup is mainly associated with cytogenetic events (DeSouza, Jones, Lowis, & Kurian, 2014; Ellison et al., 2011b; Kool et al., 2008; P. A. Northcott et al., 2011; Thompson et al., 2006a, 2006b). Over expression of SKP2 (S phase kinase 2) has been associated with breast cancer progression with bad prognosis in absence of ERR receptor. DLK1 (delta like homologue 1) controls over neurogenesis and is recorded to be only expressed in tumor cells compared to its zero
concentration normal state. (Falix, Aronson, Lamers, Hiralall, & Seppen, 2012; Signoretti et al., 2002) On the contrary, XPO4 decreased expression in hepatocellular carcinoma accompanies a poor prognosis (X.-T. Liang et al., 2011). XPO4 up regulation in this subgroup is thus beyond the expectation of this study.

**Cluster 2, score = 18.438:**

Over expressed proteins only participated in this cluster. These proteins with their first neighbors comprised a cluster of 37 nodes connected by 131 edges. The five up regulated genes expressed in this cluster are: AKR1B10, NMB, CCK, ACTA1 and TNFAIP8.

These and their neighbors are involved in responses to hormonal stimulus and positive regulation of phosphorylation whether in metabolic process or stimulus propagation. Involvement of in regulation of apoptosis was observed in interactions of TNFAIP8. (Laliberté et al., 2010; Woodward et al., 2010)
Interesting note about this cluster is that it carries one protein coded by a DEG (SKP2). The 25 nodes are interacting with the SKP2 and themselves through 117 edges. S-phase kinase – associated protein 2 is an enzyme member of the F box family functioning in phosphorylation dependant ubiquitination. It recognizes specifically cyclin dependant kinase inhibitor 1B in the S-phase. It is also recorded as a protooncogene in lymphoma. SKP2 gene was proved up regulated in medulloblastoma. (Luotong, 2005). Also, an accompanied androgen increased expression with over expression of SKP2 has been witnessed in many tumors (Frescas & Pagano, 2008).

Subgroup 4 patients are characterized by a loss in chromosome X. Differential expression of genes coding for androgen receptors can be observed in this subgroup (TRIM68, 2.5 fold decrease in its expression) (Kuniyoshi, Terui, Mishima, Matsusaka, & Hatake, 2009).

Cluster 4, score=16.061:

71 proteins interconnected with 325 edges. 6 DEGs coded for over expressed proteins in this cluster and only ne coded for the under expressed NEUROD1. Up regulated genes are involved in cellular biosynthetic processes regulation and in regulation of transcription (CCK, CAV3, ACTA1, TNFAIP8, GLI3 and XPO4) No former record of involvement of CCK in medulloblastomas progression however experimental procedure on different neuroectodermal tumors showed involvement of this regulatory peptide in tumor progression, however this was not observed in medulloblastoma tissues (Schaer & Reubi, 2013). On the other hand, CAV3 over expression was correlated to increased apoptosis of neural cells. But this observation was only valid in presence of over expression of its receptors CAV3.1 and CAV3.2 (Ohkubo & Yamazaki, 2012). Thus the reported over expression of CAV 3 in this subgroup poses a question of significance.
NEUROD1 down regulation causes a decreased risk of tumor formation in neuroblastoma patients. This was due to the induction of SLIT2 expression which increased cell motility and tumor spheres growth. FMRP is a protein coded by the X chromosome located FMR1 gene associated with fragile X syndrome. The gene is a repressor of mRNAs of several transcribed genes and one of these is the NEUROD1. This repression was associated with increased neuronal progenitor cell differentiation into neurons rather than glial cells (P. Huang et al., 2011; Jeon et al., 2014). Cytogenetic events in subgroup 4 include a loss of X chromosome suggesting a relation between this case and the FMR1 distorted performance underlying the repression of NEUROD1.

Studying the connection in the whole network provided the UBC as the main hub (see Figure 35).

Figure 33: UBC main hub in subgroup 4 protein interaction network.
Subgroup 5 protein-protein interaction network:
There are 1232 proteins interconnected by 9794 edges constructing this network. DEGs coding for proteins in this network are 231 genes. Most under expressed proteins are the DCX (-6.9 fold), FOXG1 (-6.3 fold), TEX15 (-5 fold) while the most over expressed are WIF1 (8.9 fold), DKK2 (8.6 fold), FGF20 (6.9 fold) and DKK4 (6.6 fold). It can be seen in this subgroup a clear inhibition of the WNT pathway. Also, an over expression of SHH pathway genes (BMP4, HOXD11/13, ISL1 and FGF8) can be observed. MYC is not differentially expressed in this subgroup as well as in subgroup 4. Although amplification of n-MYC is considered a signature for these subgroups, yet many samples have been recorded of normal MYC behavior, especially in subgroup 5 (Whittier, 2013).

The clustering of this subgroup utilizing MCODE algorithm results in 25 clusters. Although huge number of nodes and interactions, only 3 clusters were significant to the analysis specifications applied in this work (at a score of clustering <15.00). The clusters under study are as follows:

Cluster 1, score= 21.85:
599 nodes interconnected with 6555 edges construct this cluster. 46 DEGs coded for proteins in this cluster and extracting those genes with their first neighbors in this cluster give rise to a relatively smaller network (351 nodes, 4003 edges) more feasible to study. The 45 proteins coded by DEGs participated in regulation of cAMP biosynthesis, regulation of cell migration and regulation of apoptosis.

Regulation of cAMP is managed by up regulation of ADORA1, CHRM4, GALR1 and NTRK1 and down regulation of EDN1, GCGR and NPY1R.

Up regulation of ADORA1 results in activation of protein kinase A (PKA) which represses the pathway by GLI3 phosphorylation (activation). It is also an inducer of the extracellular signal
regulated kinase (ERK1/2) which induces the GLI3 activation. These results are confirmed by fold change numbers in the work of Whittier et al. trying to plot the changes occurring in G coupled protein receptors (GPCR) during medulloblastoma (Whittier, 2013). ADORA1 was observed of increased level of expression causing inhibition of SHH pathway as witnessed in this subgroup (Gruber, 2014; Whittier, 2013; Wolff et al., 2013). CHRM4 was one of those GPCRs found up regulated in this study and theirs as well. This is an M cholinergic muscarinic receptor expressed in the brain tissue in response to a decreased level of REST silencing TF (Conforti et al., 2013). This inhibition is not dependent on the cell type, in other words, REST under expression in cells lacking the RE1 (repressor element 1) will not affect levels of expression of target genes (Ooi & Wood, 2007). Ooi et al. also showed that CHRM4 in neuroblastoma and medulloblastoma cell lines was associated with the REST differential expression, suggesting the role of this TF in this subgroup.

On the other hand, EDN1 (endothelin 1) is down regulated. This protein is expressed during neural crest development in a temporal manner (Birks et al., 2011; M. Bonano et al., 2006). A proof of involvement of this gene in angiogenesis is documented and angiogenesis is a character in intracranial tumors (Marcela Bonano et al., 2008; D’Asti, Kool, Pfister, & Rak, 2014; Palm et al., 2009). GALR1 is a galanin receptor expressed in the cerebellar cortex (Jungnickel & Gundlach, 2005) producing peripheral actions in the small intestine as well as central actions as neurotransmitter (O’Donnell, Ahmad, Wahlestedt, & Walker, 1999). This gene is associated with several cancer types including neuroblastomas and oral squamous cell cancer (Jacky, 2011).

Proteins involved in cell migration include: over expressed CXCL12 and PTP4A1 and as well as under expressed EDN3, EDN1, ERBB4 and PDGFRA. Up regulation of PTP4A1 gene has been observed in many types of cancer cell lines. ERBB2/4 inhibition activates the PTP4A1 in breast cancer increasing cell migration (Antoinetta, 2010). In lung cancer, PTP4A1 inactivation showed
less invasiveness of the tumor cells through Src activation (Achiwa & Lazo, 2007) supporting their role in cell migration in this subtype. Down regulation of PDGFRA through Src family kinases inhibition was observed in alveolar rhabdomyosarcoma (Abraham et al., 2012). The gene itself has been proved to carry several aberrations during gastrointestinal stromal tumors, suggesting a dysfunctional regulation of transcription (Abraham et al., 2012; Verhaak et al., 2010).

Finally, the Apoptotic regulation in this cluster is performed by over expression of EPHA7, Adora1, APP, CLU, NGF, NTRK1 and SOCS and under expression of ACTN2, CDH1, GHR and NTF3. This function is though not as significant as those previously discussed.

**Cluster 2, score =16.609:**

518 nodes are interconnected with 4310 edges. Selection of the 54 proteins coded by DEGs together with their first neighbors designated a new network with 311 nodes and 2653 edges. RUNX2, EPHA3, ALK and ISL1 show highest over expression (>4.5) while TEX15, PRKCB and BTC show extremes of under expression in this network (<-4.00).

Proteins are involved in protein kinase ATP binding as well as in responses to hormonal stimuli. Up regulation of EPHA3, EPHA4, EPHA7, ALK, NTRK1, PDGFRA and SYK with under expression of PRKCB, SGK1 and ERBB4 proteins displayed the protein kinase binding function. The EPHAs are tyrosine kinases present in the nervous system. These were proven to be expressed significantly in medulloblastoma causing an increased cell migration and invasiveness (Sikkema et al., 2012). The anaplastic lymphoma kinase (ALK) has also been up regulated in tumor cells in association with NPM (nucleophosphin receptors) (Chiarle, Voena, Ambroio, Piva, & Inghirami, 2008). PRKCB, the protein kinase C beta, is actively expressed in tumor cell phosphorylating the histone H3T6 stopping apoptosis and increasing tumor growth. In medulloblastoma cell lines,
experimentation has proved this gene to be under expressed yet no mechanism is still known (Surdez et al., 2012).

**Cluster 3, score=15.346:**

With 51 nodes and 399 edges, the network comprised 12 proteins coded by DEGs. Over expression peak value is 4.2 fold change experienced by COL6A3, COL23A1 and MMP2 while most under expressed proteins are COL4A1, COL4A2 and EDN1 (-2.5 fold).

The majority of the later proteins are an extracellular matrix structure/ forming. Namely, DEGs coding for these proteins are: COL9A3, COL5A1, COL6A3, COL8A2, COL13A1, MMP2 and TGFBI (up regulated) as well as COL4A1, COL4A2, NID1 and EDN1 (down regulated).

Like other molecular subgroups discussed above, subgroup 5 has the main hun in the protein interaction network being the UBC.

![Diagram of UBC as main hub in subgroup 5 protein-protein interaction networks.](image)

**Figure 34:** UBC as main hub in subgroup 5 protein-protein interaction networks.
5.7. Transcription regulatory networks:

Transcription regulatory networks consist of transcription factors and their direct target genes. These networks aim at detecting the upstream transcription factors regulating the transcription of a certain gene set through certain binding motifs and enriched ChIP-seq data utilizing ENCODE database. iRegulon tool (Janky et al., 2014) is a recently developed computation method to reverse engineer transcriptional regulated network underlying a coexpressed gene set using cis-regulatory sequence analysis. Utilizing the tool on each protein-protein interaction network cluster, a set of TFs regulating the expression in this cluster was displayed with prediction of potential motif binding sites for the TF on the coding genes’ of the proteins. A regulating TF discovery seeks a better understanding of the upper level of regulation applied on the DEGs in each subgroup. The hypothesis proposed suggests an altered TF behavior is behind each expression profile for each of the subgroups.

WNT subgroup transcription regulatory network:

In WNT subgroup, genes in WNT pathway are up regulated/ down regulated leading to an overall activation of the pathway. This behavior can be confirmed by the earlier results confirming the down regulation of the (Wif1, FZD10, LEF1, WNT5A, and LRP5/6) and up regulation of (Myc, TCF/LEF, Prickl). The transcription regulatory network was constructed utilizing the larger motif database (10k PWMs) and ROC threshold for AUC calculation as default (3%). The motifs utilized are from Homo sapiens with maximum FDR for similarity (0.001).
Cluster 1, 620 query nodes:

26 motifs can be targeted by 12 TFs regulating the 620 query nodes in this cluster. The TF with highest NES (normalized enrichment score) is the SIN3A targeting 4 DEGs; MYC (log_FC=2.23), ETV4 (log_FC= -2.07), MAF (log_FC=-2.32) and TFAP2A (log_FC=-2.92). (see Figure 35)

Figure 35: SIN3A regulatory network in WNT subgroup.
SIN3A is a transcription regulatory protein that binds the REST receptors and induces a repression of transcription of the target gene (Lawinger et al., 2000). Its activity has been recorded with HDAC1 (histone deacetylase) in chromatin modification. The transcription factor silencing/knock out causes an increased activation of the MYC (Zelski Ellison, 2010). In this network, the increase in MYC expression supports the possibility of silencing or disruption in the SIN3A gene. Also, MAT 2 alpha protein (methionine adenosyl transferase family) is found in complex with the SIN3A as chromatin regulators together with the HDAC1/2 as well as other repressors (Katoh et al., 2011). MAT2 alpha is also coexpressed with MAFK in heme oxygenase 2 repressions (Sun et al., 2004). This suggests that MAF expression is affected by the SIN3A regulation, with an expected up regulation in presence of SIN3A (Brand, 2004).

A second TF, SUZ12 (3.506), targets 49 genes, 19 are DEGs. The Figure 36 below displays the type of behavior for each gene. This TF score is based on motif similarity. The most extreme DEGs are regulated by this TF, FOXG1 (log_FC=5.96) and PAX9 (log_FC= -4.64). BMI1 is a polycomp genes group member that is complexes with SUN12 to repress the BMP (bone morphogenetic protein) pathway controlling over metastases and invasion in medulloblastomas (Merve et al., 2014). FOXG1 coexpressed with the BMI1 in medulloblastoma stem cells to regulate stem cell renewal and tumorigenesis (Manoranjan et al., 2013). PAX9 over expression in medulloblastoma is postulated to relate with poor prognosis as the PAX9 targets the cell differentiation in the hindbrain cerebellar tissue. Down regulation of this gene might be an answer for the good prognosis associated with medulloblastoma WNT subgroup ascertained by the up regulation of the FOXG1 responsible for the cell renewal (Fotaki, Smith, Pratt, & Price, 2013).
Figure 36: SUZ12 regulatory network in WNT subgroup.
SHH subgroup transcription regulatory network:

Cluster 1, 84 query nodes:

E2F1 is a TF formulating a tumor repressor complex with the Rb. This TF is under expressed in medulloblastomas in the SHH subtype increasing lipogenesis driven by the FASN (fatty acid synthase N) expression (B Bhatia, 2011). This expression profile is associated with the DE profiles described Error! Reference source not found.. This finding suggests the E2F1 plays an important regulatory role in developing SHH subgroup clinical profile. Added to that, a study conducted on E2F1 knockout mice at different embryonic stages proves that deregulation of E2F1 developed medulloblastomas at early age however glioblastoma will be the consequence at older ages if escaped the early stage PNETs(Olson, 2007) as seen in Error! Reference source not found..

Figure 37: E2F1 regulatory network in SHH subgroup.
Other TFs regulating this cluster are: DDX1, FOXk1, CNOT3, MAP2K1 and MYC; all showing same interaction cluster

Cluster 2, cluster 3 and cluster 4 DEGs are found a subset of cluster 1 with same transcriptional regulation acting upon them. Accordingly, this study places its findings considering cluster 1 analyzed protein-protein interactions.
Subgroup 3 transcription regulatory network:

Cluster 1, query nodes= 54:

Three main TFs are postulated to regulate over the DEGs in this cluster. The first is TBX1 which regulates over 11 DEGs (up regulated: SFRP1, ADCY8, CXCR4, PRKCB, DRD2, NFATC1, PDGFRA and down regulated: PTPRR, SYNJ2, FOXG1, DABB1) (Figure 38)

T-box 1 is a TF involved in many developmental and signaling pathways. A special fact lies in the gene position coding for this TF (22q11.2) which is known for mutations causing several birth defects. These defects rely mainly in the distorted performance of the TGF / BMP signaling pathway, Smad signaling, SHH signaling and WNT signaling (Jiang et al., 2014; Norden, Greulich, Rudat, Taketo, & Kispert, 2011; Yamagishi et al., 2003). FOX protein family regulates enhancement of transcription of TBX1 and embryonic stem cells with TBX1 knocked out showed decreased levels of FOXG1 (Esmailpour, 2012). It is worth mentioning that TBX1 has been found down regulated in cancer cells (Esmailpour, 2012; Trempus et al., 2011).
Figure 38: TBX1 regulatory network in subgroup 3.
Cluster 2, query nodes= 44:
Cluster 2 included genes involved in MAPK signaling (TGFB, FGFR, PDGFR, MET, PKC, c-MYC), ERBB signaling (PKC, ERBB2/3, MYC, CAMK, GAB1), Ca signaling pathway (PTK, SPHK, PKC) and axon guidance (Robi2/3, CXCR4, EphA/B, UNC5, MET).

13 TF target this cluster genes yet only 4 are of proved significance regarding their scores (above NES=5). Most TF targeted the same DEGs with extra genes or less. To facilitate the display, all TF clusters were collected in one network. (Figure 39)

Highest scored TF and interacting with 25 DEGs is AVEN. AVEN codes for a protein that plays a role in male and female germ cell development. It is also an antiApoptotic TF that binds and inhibits the pro-apoptotic APAF1 and enhances BCL-xL (B cell lymphoma extra large) activity (O’Shea, Hensey, & Fair, 2013). AVEN up regulation was observed in poor prognostic cases of acute lymphoblastic leukemia (ALL) suggesting its association with poor prognosis in malignancy (J. Choi et al., 2006). On the other hand, loss of heterozygosity due to epigenetic silencing of AVEN was proved by (Eißmann et al., 2013; Sugarbaker et al., 2008) to be associated with malignant pleural mesotheliomas suggesting a potential mutation underlying the altered control of this TF. With these mutations, amplification of c-MYC was observed. In our network a MYC down regulation is seen supporting the altered regulation of AVEN on its regulon. Relating this information to cancer progression can be explained through the need for BCL-2, activated by AVEN, with the c-MYC to induce cell proliferation in CHO-k1 (Chinese hamster ovary k1 cell line) compared to the c-MYC alone which as a matter of fact didn’t prevent apoptosis compared to control cells (ifandi Vasiliki, 2005).
Figure 39: network of Subgroup 5 DEGs in cluster 2 interactions with their TFs.
Figure 40: AVEN regulatory network in subgroup 3.
Subgroup 4 transcription regulatory network:

Cluster 1, query nodes= 176:

Four TFs regulated expression in this cluster, E2F1, NFYC, ZBTB33 and ATF2. Remarkably, E2F1 (NES=15.766) targets all 142 genes, whether DEGs or not, through 45 possible motifs. DEGs targeted by this TF are: DLK1, SKP2 and XPO4 (up regulated) and DEPDC1B, CENPK, CENPM, FOXM1, TK1 and CDT1 (down regulated).

DLK1/Pref1 preadipocyte complex was activated with the activation of transcription regulator E2F1 in adipocyte tissue inducing proliferation and differentiation events (Moon et al., 2002; Shen, 2009). SKP2 expression is upregulated by E2F1 induction with the help of PI3K (phosphoinositide 3 kinase) (Marti, Wirbelauer, Scheffner, & Krek, 1999; Reichert, Saur, Hamacher, Schmid, & Schneider, 2007). XPO4 has been recorded down regulated in hepatocellular carcinoma with a recorded alteration in genomic context of the gene coding for the E2RF1 (Xue, 2009). Down regulation of CENPK/CENPM has been associated with cytogenetic events already proved to exist in Subgroup 4 patients. Literature review has clarified the involvement of E2F1 regulation to these genes as well as other genes participating in this anomaly (Hrubá et al., 2011; Reinhold et al., 2011) (see figure 43).

E2F1 is also regulating transcription of genes in cluster 2, 3 and 4.
Figure 41: E2F1 regulatory network in subgroup 4.
Subgroup 5 transcription regulatory network:

Cluster1, query nodes=46:

13 possible TFs can target this cluster, yet all scored below NES=5.00. However, mentioned below are the highest three putative TFs.

ZNF281 TF (NES=4.695) targets 30 DEGs through 10 possible binding motifs. This TF was recorded at the minimum loss region on chromosome 1q 25:32.2 in retinoblastoma samples. Although the case here is a medulloblastoma, yet further research is needed to find out the role of this aberration in regulatory function of this TF (Corson, Huang, Tsao, & Gallie, 2005; Mol et al., 2014).

Figure 42: ZNF281 regulatory network in cluster 1 subgroup 5 network.
ETS1 and E2F1 are two other TFs controlling transcription in this cluster. Both compete in their binding to MYC promoter. Mutation of the E2F1 site allows the binding of ETS1 (Bunt et al., 2011). E2F1 TF (NES=4.047) targets 24 DEGs in this cluster. As discussed before, the TF is deregulated in most brain tumors including medulloblastoma. Association between this TF and Myc miRNA regulation was witnessed in many types of cancer including neuroblastoma (P. A. Northcott et al., 2009; Olson et al., 2007). (See figure45)

TFs E2F1 and ZNF281 also target cluster 2. Added to these are the MAFB and PARP1 TFs (NES=4.7) which play an important role in E2F1 signal propagation.
6. Chapter 6: Conclusion

While empirical treatments for medulloblastoma are still practiced, it has become clear that medulloblastoma can no longer be considered as a single disease. Adding to that is the need for more precise approaches to diagnose the disease other than depending on histology. In this context, supporting the research hypothesis posed at the beginning of this work, I thereby conclude the following:

**Objective 1:**

“Identify differentially expressed genes for each molecular subgroup in medulloblastoma patients.”

After gene filtering, each subgroup presented a set of differentially expressed genes that differentiate it from other subgroups. Accordingly, the first subgroup comprised up regulated WNT canonical signaling pathway resembled in down regulation of WIF, BMP4, ADRB, FZD10 and LEF1 that inhibit the pathway with up regulation of FOXG1, CTNNB1, GABRA5, GABRA1 and WNT5A genes signaling in the pathway. On the other hand, SHH subgroup displayed a very similar profile with genes involved in SHH signaling pathway directed towards activation of the pathway. Up regulation of FOXG1, LRRC7 and PPARA with down regulation of GLI3, BMP4 and BMP2 indicating an up regulation of the PTCH1 and down regulation of the SMO are regulating this expression (SHH activation).

These results constitute a relatively clear profile for the first two subgroups, however, for the other 3 subgroups the profiles show some blurred characteristics. Considering the subgroup 3 profiles, as a non WNT/SHH subgroup, deactivated later pathways are observed. Characteristic
gene expression behaviors included down regulation of FOXG1 only expressed in this subgroup associated with TTR and DUSP down regulations provoking cancer progression and a retinopathy (TTR). The latter aggravated with the PRLR and MET up regulation. Subgroup 4 provides more stable gene expression profiles compared to its recorded cytogenetic effect. Hence, it is expected to observe down regulated gene expression profiles involved in cell cycle regulation (CENPM/K, FOXM1 and CDT1). Added to that, the expression profile of TPN1 was exclusive to this subtype. Down regulated TPN1 genes are observed only in subgroup 4 medulloblastoma providing a characteristic signature for detection. Lastly, subgroup 5 shared subgroup 3 in the down regulation of FOXG1 and inhibition of WNT pathway but an up regulation of SHH pathway was observed (BMP4, HOXD11/13, ISL1 and FGF8). In general, genes regulating cAMP biosynthesis are differentially expressed in this subgroup. No MYC differential expression and a down regulation of TEX 15 are unique to this subgroup; however absence of proper scientific explanation for these behaviors presents a research concern.

**Objective 2:**

“Specify gene markers for each medulloblastoma molecular subgroup.”

Pairwise t-testing for each gene across all subgroups was used to identify a unique behavior for each gene across these subgroups. This differed from previously applied multiple t testing / ANOVA in that the results are limited to comparing a single gene’s expression value in one group versus every single other group separately. Keeping this in mind, each subgroup is expected to display limited specific genes. The WNT subgroup, for instance, failed to show a specific marker applying this analysis, thus shifting back to chromosome 6 monosomy identification as a diagnostic marker is recommended. On the contrary, SHH subgroup displayed two putative specific markers namely the CCDC3 and the SUN2. CCDC3 has been reported to repress the tumor necrosis factor alpha and kappa in breast tumors, gastrointestinal tumors and glioma.
propagating the growth of the tumor cells (El Yazidi-Belkoura, Adriaenssens, Dollé, Descamps, & Hondermarck, 2003; Kwei, Finch, Ranger-Moore, & Bowden, 2006; Laliberté et al., 2010). This matches the situation in this analysis with medulloblastoma SHH subgroup, suggesting further experimentation of the CCDC3 potential as marker for SHH subgroup. SUN2 remains under query with no significant over expression in this subgroup relative to other subgroups.

Reaching for subgroup 3, there are no significantly specific markers for this subgroup in compare to others. Thus, relying on association between the non WNT/SHH medulloblastoma and a retinopathy remains as the best choice for diagnosis. On the other hand, RPL 29 and RPS7 and their pseudogenes display a significant expression behavior in subgroup 4 compared to all other groups proposing a hypothesis for their transcription repression for some vital developmental genes. This hypothesis is supported by their observed duplicated chromosomal loci (6q24.3, 17q22, 2p25.3 and 3p21.1). It is worth mentioning that an alternative marker selection can be dependable on a unique expression profile for the TPN1 genes in this subgroup which is not observed in others.

Finally, the unspecified subgroup 5 provides a long list of specific potential markers (16 genes). From this list, MITF is the only gene that is interacting with the DEGs through MAPK pathway regulation of transcription and controlling cell proliferation in cancer. LEF1 over expression activates MITF promoter to increase their expression and propagate melanogenesis in skin. This interaction together with PAX3 up regulation is witnessed to contribute in cancer cell survival as well (Kingo et al., 2008; Scholl et al., 2001). Consequently, MIFT up regulation is proposed as a putative marker for subgroup 5 medulloblastoma patients.

**Objective 3:**
“Network the protein –protein interaction in each molecular subgroup.”
For each subgroup, list of DEGs can be utilized to visualize the behavior of interacting neighboring genes and thus relate it with the progression of a certain phenotype. Protein–protein interactions in WNT subgroups patients are no doubt interacting with mainly WNT signaling pathway as well as the SHH. They show a clear participation in cancer progression through MAPK/TGFβ signaling pathway.

Turning to the SHH pathway, SHH pathway activation is observed with absence of inhibition of the PTCH1. It is also associated with an increased tumor growth through increased vascularity. Pathways for collagen activation and cell adhesion activation associate the increased vascularity for these tumors.

Subgroup 3 DEGs display an activation of cell proliferation resembled in CXCR2/PP2A activation as well as SFRP1. Added to that, an observed disruption in lipophilic molecule homeostasis can be detected which underlies the retinopathy progression. Although subgroup 4 DEGs display an altered retinal tissue stabilization, yet they are more involved in myopathy progression pathways through increased muscle contraction and development. These interactions are coupled with interactions involved in negative regulation of the Apoptotic mechanisms in neural cells. Subgroup 5 DEGs exhibited an involvement in cAMP biosynthesis with unique profiles of GPCRs involved in apoptosis and cell migration (EPHA-family). Also, UBC is observed highly connected along the five subgroups’ networks, suggesting a key role performed by this protein coding enzyme in developing medulloblastoma. By this, studying the protein-protein interactions has been proved to aid in the understanding of these subgroups and helps get a better look on the regulation that can be applied on these gene profiles to improve the prognosis.

**Objective 4:**

“Compute the type of transcriptional regulation applied on each subgroup marker gene.”
Utilization of protein-protein interaction networks not only helps in understanding the behavior of the genes throughout the different conditions but also set a basis for the study of type of regulation applied on different protein clusters in a network. Thereby, studying the WNT network placed the SUZ12 TF as the top interacting TF responsible for the DEG pattern observed in this subgroup. Also, the SHH network presented several interaction clusters with the E2F1 TF playing a significant role in the DEGs profiles along majority of clusters. Added to that, the MYC has represented an interesting expression profile as well as a deep interaction with DEGs in SHH subgroup. For subgroup 3 transcription regulatory networks, TBX1 targets transcription of several DEGs and is observed down regulated in previous research in association with cancer progression. Most significant TF in subgroup3 with over 25 DEGs regulons is the AVEN and targets the apoptotic mechanisms. Subgroup 4 clusters also are regulated by a significant TF E2F1 and SKP2 involved in apoptotic regulation. Subgroup 5 DEGs networks are targeted for regulation by the ZNF281 and the ETS1/ETF1 MYC activated transcription regulatory routes. These findings help support the postulation that medulloblastoma is no longer a single disease, but 5 molecular subgroups subdivide it into 5 phenotypes. A series of gene interactions give rise to these different phenotypes. The type of regulation performed on these interactions identifies the signature of each subgroup. This interaction is suggested to be transcriptional aside to several alternative mechanisms that need to be approached in future research. Also, this work proves the impact of applying a holistic view to study different gene behaviors that occur in medulloblastoma specifically and in cancer generally. Analyzing microarray gene expression platforms helps in visualizing these gene profiles and thus improving the awareness about the roles they play in disease progression.
7. **Future directions:**

Medulloblastoma places itself as an aggressive pediatric tumor in the cancer world that is difficult to diagnose and treat. Accordingly, my work presents an approach to reach a better understanding of the mechanisms of interaction between the genes in this disease. This lead to the confirmation of the subdivision of this major cancer into 5 subgroups of different molecular expression profiles. It is encouraged to:

a. **Classification of medulloblastoma subgroups:**

   The hierarchical clustering utilizing different algorithms presented five main clusters. These clusters are subdivided into smaller ones that might present different subgroups. On the other hand, PCA showed 7 principal components instead of only five, or even 4 as still supported by many researchers. Accordingly, classifying this disease utilizing an increased sample number is required to try to visualize all possible clusters. This requires a GWAS (Genome wide association study) on the transcription expression profiles level to be applied in order to normalize the analysis.

b. **Differentially expressed genes in relation to age/sex:**

   Several TFs regulating network clusters in medulloblastoma subgroups under this study were previously recorded as of differential expression in medulloblastoma cancer or in other cancer types. Accordingly, it is recommended that the gene profiling will be repeated utilizing data about age and sex linked to the subgroups. This analysis will help understand the relation between some gene expression profiles, prognosis and age/sex factors. Subgroup 4 samples analysis are expected to provide promising results.

c. **Transcription factor binding motifs:**

   The current analysis provided general data about potential TFs involved in the progression of the molecular profile. Yet utilizing motif data is intended to provide more
information about possible mutation sites that might affect the transcription regulatory function on the selected genes. Also tracks for TF-gene interaction will provide more data about the type of interaction happening. A high throughput technology as RNA-Seq is a good candidate for this objective with expected improved results through applying sequencing basis instead of the less specific hybridization concept in the microarray platform.

d. Directional transcriptional regulatory networks:

In my work, preliminary data obtained from protein-protein interaction databases (as STRING) are utilized. These data remain non directional. Accordingly, a need for constructing a directional network for these networks and its transcriptional regulatory ones will provide more information and open the gates for new insights for possible involved gene interactions. These networks support the information in hand with the type of interaction (inhibition /activation) as well as its direction (regulator of/ regulated by) between two nodes. Additional direction of analysis will be a weighted edge values depending on type of inhibition/ activation exerted on the affected node by the effector one. In this way, fairly clear data can be obtained from inspecting these networks.
8. **Appendix:**

**Table 1: RNA degradation statistical analysis.**

<table>
<thead>
<tr>
<th>Array</th>
<th>Slope</th>
<th>pvalue</th>
<th>Array</th>
<th>Slope</th>
<th>pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM918578_mbt003</td>
<td>4.78</td>
<td>9.68E-12</td>
<td>GSM918616_mbt098</td>
<td>4.39</td>
<td>1.15E-11</td>
</tr>
<tr>
<td>GSM918579_mbt004</td>
<td>8.04</td>
<td>1.26E-14</td>
<td>GSM918617_mbt099</td>
<td>4.62</td>
<td>3.87E-10</td>
</tr>
<tr>
<td>GSM918580_mbt006</td>
<td>9.67</td>
<td>4.67E-18</td>
<td>GSM918618_mbt102</td>
<td>5.4</td>
<td>3.66E-13</td>
</tr>
<tr>
<td>GSM918581_mbt008</td>
<td>5.89</td>
<td>1.10E-12</td>
<td>GSM918619_mbt103</td>
<td>4.79</td>
<td>2.43E-12</td>
</tr>
<tr>
<td>GSM918582_mbt009</td>
<td>5.62</td>
<td>2.81E-14</td>
<td>GSM918620_mbt105</td>
<td>6.4</td>
<td>9.53E-15</td>
</tr>
<tr>
<td>GSM918583_mbt010</td>
<td>5.94</td>
<td>1.29E-15</td>
<td>GSM918621_mbt106</td>
<td>5.05</td>
<td>3.65E-13</td>
</tr>
<tr>
<td>GSM918584_mbt011</td>
<td>5.08</td>
<td>4.95E-14</td>
<td>GSM918622_mbt109</td>
<td>4.49</td>
<td>2.80E-12</td>
</tr>
<tr>
<td>GSM918585_mbt013</td>
<td>8.18</td>
<td>1.75E-15</td>
<td>GSM918623_mbt110</td>
<td>7.3</td>
<td>2.13E-13</td>
</tr>
<tr>
<td>GSM918586_mbt016</td>
<td>4.37</td>
<td>5.93E-15</td>
<td>GSM918624_mbt124</td>
<td>6.2</td>
<td>8.44E-14</td>
</tr>
<tr>
<td>GSM918587_mbt020</td>
<td>4.84</td>
<td>5.96E-14</td>
<td>GSM918625_mbt126</td>
<td>7.26</td>
<td>1.09E-13</td>
</tr>
<tr>
<td>GSM918588_mbt024</td>
<td>5.69</td>
<td>2.81E-14</td>
<td>GSM918626_mbt127</td>
<td>6.53</td>
<td>4.40E-12</td>
</tr>
<tr>
<td>GSM918589_mbt031</td>
<td>6.3</td>
<td>2.45E-15</td>
<td>GSM918627_mbt135</td>
<td>3.99</td>
<td>6.20E-11</td>
</tr>
<tr>
<td>GSM918590_mbt032</td>
<td>5.39</td>
<td>2.36E-13</td>
<td>GSM918628_mbt136</td>
<td>7.3</td>
<td>1.16E-12</td>
</tr>
<tr>
<td>GSM918591_mbt033</td>
<td>3.68</td>
<td>3.50E-12</td>
<td>GSM918629_mbt140</td>
<td>5.02</td>
<td>1.63E-12</td>
</tr>
<tr>
<td>GSM918592_mbt034</td>
<td>4.26</td>
<td>4.76E-12</td>
<td>GSM918630_mbt141</td>
<td>6.01</td>
<td>1.23E-11</td>
</tr>
<tr>
<td>GSM918593_mbt035</td>
<td>3.61</td>
<td>2.53E-11</td>
<td>GSM918631_mbt144</td>
<td>9.59</td>
<td>3.99E-13</td>
</tr>
<tr>
<td>GSM918594_mbt037</td>
<td>4.19</td>
<td>6.23E-12</td>
<td>GSM918632_mbt145</td>
<td>7.2</td>
<td>3.02E-15</td>
</tr>
<tr>
<td>GSM918595_mbt045</td>
<td>4.66</td>
<td>3.51E-11</td>
<td>GSM918633_mbt146</td>
<td>5.1</td>
<td>3.27E-13</td>
</tr>
<tr>
<td>GSM918596_mbt046</td>
<td>4.73</td>
<td>1.15E-13</td>
<td>GSM918634_mbt147</td>
<td>4.92</td>
<td>1.14E-11</td>
</tr>
<tr>
<td>GSM918597_mbt048</td>
<td>5.15</td>
<td>3.70E-13</td>
<td>GSM918635_mbt148</td>
<td>6.98</td>
<td>3.57E-15</td>
</tr>
<tr>
<td>GSM918598_mbt050</td>
<td>3.47</td>
<td>1.58E-11</td>
<td>GSM918636_mbt149</td>
<td>5.84</td>
<td>5.34E-14</td>
</tr>
<tr>
<td>GSM918599_mbt051</td>
<td>7.4</td>
<td>4.24E-15</td>
<td>GSM918637_mbt150</td>
<td>5.89</td>
<td>1.51E-15</td>
</tr>
<tr>
<td>GSM918600_mbt053</td>
<td>4.43</td>
<td>5.57E-13</td>
<td>GSM918638_mbt151</td>
<td>6.22</td>
<td>1.11E-14</td>
</tr>
<tr>
<td>GSM918601_mbt058</td>
<td>5.37</td>
<td>3.29E-13</td>
<td>GSM918639_mbt156</td>
<td>4.62</td>
<td>4.10E-12</td>
</tr>
<tr>
<td>GSM918602_mbt062</td>
<td>3.88</td>
<td>1.57E-12</td>
<td>GSM918640_mbt158</td>
<td>4.92</td>
<td>2.49E-13</td>
</tr>
<tr>
<td>GSM918603_mbt063</td>
<td>4.47</td>
<td>4.40E-13</td>
<td>GSM918641_mbt161</td>
<td>6.54</td>
<td>1.88E-14</td>
</tr>
<tr>
<td>GSM918604_mbt068</td>
<td>5.21</td>
<td>6.85E-12</td>
<td>GSM918642_mbt166</td>
<td>5.68</td>
<td>8.63E-14</td>
</tr>
<tr>
<td>GSM918605_mbt069</td>
<td>4.5</td>
<td>3.40E-11</td>
<td>GSM918643_mbt167</td>
<td>5.61</td>
<td>5.34E-14</td>
</tr>
<tr>
<td>GSM918606_mbt075</td>
<td>4.25</td>
<td>2.75E-12</td>
<td>GSM918644_mbt168</td>
<td>6.14</td>
<td>2.84E-14</td>
</tr>
<tr>
<td>GSM918607_mbt078</td>
<td>5.43</td>
<td>4.91E-13</td>
<td>GSM918645_tbm055</td>
<td>3.64</td>
<td>5.68E-10</td>
</tr>
<tr>
<td>GSM918608_mbt079</td>
<td>5.4</td>
<td>1.37E-12</td>
<td>GSM918646_tbm061</td>
<td>5.41</td>
<td>4.13E-12</td>
</tr>
<tr>
<td>GSM918609_mbt081</td>
<td>4.15</td>
<td>2.20E-12</td>
<td>GSM918647_tbm082</td>
<td>6.43</td>
<td>1.78E-12</td>
</tr>
<tr>
<td>GSM918610_mbt083</td>
<td>4.9</td>
<td>1.19E-12</td>
<td>GSM918648_tbm084</td>
<td>8.59</td>
<td>9.55E-15</td>
</tr>
<tr>
<td>GSM918611_mbt085</td>
<td>4.05</td>
<td>1.10E-10</td>
<td>GSM918649_tbm091</td>
<td>8.71</td>
<td>1.92E-17</td>
</tr>
<tr>
<td>GSM918612_mbt087</td>
<td>3.71</td>
<td>1.94E-11</td>
<td>GSM918650_tbm092</td>
<td>5.64</td>
<td>1.14E-13</td>
</tr>
<tr>
<td>GSM918613_mbt089</td>
<td>5.27</td>
<td>2.10E-12</td>
<td>GSM918651_tbm107</td>
<td>4.54</td>
<td>1.00E-10</td>
</tr>
<tr>
<td>GSM918614_mbt093</td>
<td>5.58</td>
<td>8.67E-14</td>
<td>GSM918652_tbm111</td>
<td>4.85</td>
<td>5.53E-12</td>
</tr>
<tr>
<td>GSM918615_mbt095</td>
<td>5.24</td>
<td>3.46E-14</td>
<td>GSM918653_tbm143</td>
<td>10.5</td>
<td>2.56E-14</td>
</tr>
<tr>
<td>Array</td>
<td>Background Minimum</td>
<td>Background Maximum</td>
<td>Background Average</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918578_mbt003</td>
<td>41.47927807</td>
<td>44.19097449</td>
<td>42.86221074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918579_mbt004</td>
<td>38.94451872</td>
<td>42.13914174</td>
<td>40.98157403</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918580_mbt006</td>
<td>34.92981283</td>
<td>36.44487678</td>
<td>35.87287786</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918581_mbt008</td>
<td>35.13168449</td>
<td>37.08101102</td>
<td>36.17629111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918582_mbt009</td>
<td>49.52139037</td>
<td>53.88139987</td>
<td>52.09529561</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918583_mbt100</td>
<td>41.27437262</td>
<td>43.87963732</td>
<td>42.96311864</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918584_mbt101</td>
<td>43.20454545</td>
<td>46.12378483</td>
<td>45.00761665</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918585_mbt103</td>
<td>37.09959893</td>
<td>39.40622899</td>
<td>38.53681137</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918586_mbt106</td>
<td>55.17702703</td>
<td>58.41584158</td>
<td>56.60414623</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918587_mbt1020</td>
<td>37.29278075</td>
<td>39.5390117</td>
<td>38.56064739</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918588_mbt1024</td>
<td>42.80815508</td>
<td>46.86964981</td>
<td>45.29261321</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918589_mbt301</td>
<td>77.372973</td>
<td>82.68910891</td>
<td>80.26239002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918590_mbt302</td>
<td>51.58565737</td>
<td>54.81556573</td>
<td>53.19609834</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918591_mbt303</td>
<td>35.32364665</td>
<td>37.01945525</td>
<td>36.21348</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918592_mbt304</td>
<td>39.52277277</td>
<td>42.25506867</td>
<td>41.13948355</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918593_mbt305</td>
<td>31.37864721</td>
<td>32.54153041</td>
<td>32.08616381</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918594_mbt307</td>
<td>61.59157754</td>
<td>65.86984957</td>
<td>63.96570711</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918595_mbt3045</td>
<td>52.02673797</td>
<td>55.77109222</td>
<td>53.98257973</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918596_mbt406</td>
<td>31.19797297</td>
<td>32.86930693</td>
<td>32.17686689</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918597_mbt408</td>
<td>51.2105615</td>
<td>55.01700458</td>
<td>53.43448975</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918598_mbt500</td>
<td>60.08783788</td>
<td>64.3499019</td>
<td>62.08856123</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918599_mbt501</td>
<td>53.43716578</td>
<td>56.79005886</td>
<td>55.68969819</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918600_mbt503</td>
<td>54.07286096</td>
<td>56.63570691</td>
<td>55.60526066</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918601_mbt505</td>
<td>74.40116655</td>
<td>96.08976988</td>
<td>86.07872109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918602_mbt602</td>
<td>54.89043825</td>
<td>57.72827497</td>
<td>56.21755933</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918603_mbt603</td>
<td>51.51854396</td>
<td>54.52010376</td>
<td>53.17219557</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918604_mbt608</td>
<td>51.29376658</td>
<td>54.88620013</td>
<td>53.13757033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918605_mbt609</td>
<td>57.49135638</td>
<td>61.50129702</td>
<td>59.56395025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918606_mbt705</td>
<td>53.70387701</td>
<td>58.03370058</td>
<td>56.55475607</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918607_mbt708</td>
<td>66.72045153</td>
<td>76.23479398</td>
<td>71.12888809</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918608_mbt709</td>
<td>65.48339973</td>
<td>75.21573322</td>
<td>69.88160227</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918609_mbt801</td>
<td>56.2078877</td>
<td>60.81452659</td>
<td>59.31979108</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918610_mbt803</td>
<td>74.2473262</td>
<td>79.75942783</td>
<td>77.20304447</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918611_mbt805</td>
<td>57.93248663</td>
<td>62.56290532</td>
<td>60.73348478</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918612_mbt807</td>
<td>40.96390909</td>
<td>43.73558004</td>
<td>42.50426057</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918613_mbt809</td>
<td>69.70855615</td>
<td>75.40614781</td>
<td>72.99549769</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918614_mbt903</td>
<td>43.99611147</td>
<td>58.4150454</td>
<td>54.8754626</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918615_mbt905</td>
<td>80.97259358</td>
<td>86.79571984</td>
<td>84.05162726</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918616_mbt908</td>
<td>99.21122995</td>
<td>108.4571054</td>
<td>104.783488</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918617_mbt909</td>
<td>122.4661355</td>
<td>134.0039293</td>
<td>128.8102444</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918618_mbt102</td>
<td>73.1054051</td>
<td>80.16666667</td>
<td>76.75283929</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918619_mbt103</td>
<td>82.47927973</td>
<td>90.00792079</td>
<td>86.39261289</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918620_mbt105</td>
<td>75.3937583</td>
<td>83.17004578</td>
<td>78.93135212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918621_mbt106</td>
<td>68.17981283</td>
<td>73.28552718</td>
<td>71.22174828</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918622_mbt109</td>
<td>69.44852941</td>
<td>79.52075227</td>
<td>75.86619588</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918623_mbt110</td>
<td>74.49070385</td>
<td>81.5391432</td>
<td>77.84123244</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918624_mbt124</td>
<td>52.30212483</td>
<td>55.84041857</td>
<td>54.33515045</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918625_mbt126</td>
<td>53.5102674</td>
<td>57.90124264</td>
<td>56.03599346</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918626_mbt127</td>
<td>43.49513994</td>
<td>63.00462046</td>
<td>56.79794339</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918627_mbt135</td>
<td>59.93389501</td>
<td>65.33289732</td>
<td>62.43139961</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918628_mbt136</td>
<td>60.96305898</td>
<td>73.2969261</td>
<td>66.71194302</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM</td>
<td>mbt</td>
<td>Value 1</td>
<td>Value 2</td>
<td>Value 3</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>GSM918629_mbt140</td>
<td>64.05767984</td>
<td>83.70111184</td>
<td>74.00439263</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918630_mbt141</td>
<td>57.15424498</td>
<td>63.86657946</td>
<td>61.56660873</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918631_mbt144</td>
<td>51</td>
<td>54.11510791</td>
<td>52.6574906</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918632_mbt145</td>
<td>45.57219251</td>
<td>48.15434925</td>
<td>47.02875874</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918633_mbt146</td>
<td>38.15708556</td>
<td>41.4587156</td>
<td>39.95035063</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918634_mbt147</td>
<td>35.46524064</td>
<td>37.7464332</td>
<td>36.85128395</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918635_mbt148</td>
<td>42.35227273</td>
<td>45.77719528</td>
<td>44.58651206</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918636_mbt149</td>
<td>42.54459459</td>
<td>45.48959688</td>
<td>44.40162089</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918637_mbt150</td>
<td>38.02567568</td>
<td>40.85149157</td>
<td>39.62113661</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918638_mbt151</td>
<td>39.74197861</td>
<td>43.30978613</td>
<td>41.49433314</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918639_mbt156</td>
<td>72.76203209</td>
<td>82.49675746</td>
<td>78.2611554</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918640_mbt158</td>
<td>59.13404813</td>
<td>64.23071938</td>
<td>62.0345141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918641_mbt161</td>
<td>52.60143043</td>
<td>59.17527796</td>
<td>56.7876606</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918642_mbt166</td>
<td>52.43492479</td>
<td>61.6573472</td>
<td>59.5315897</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918643_mbt167</td>
<td>78.61631016</td>
<td>84.19646365</td>
<td>81.87647822</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918644_mbt168</td>
<td>85.23195187</td>
<td>90.55396619</td>
<td>88.06184329</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918645_tbm055</td>
<td>68.55815508</td>
<td>78.74513619</td>
<td>75.36561527</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918646_tbm061</td>
<td>66.59229748</td>
<td>71.75965946</td>
<td>69.0430544</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918647_tbm082</td>
<td>113.8763514</td>
<td>130.9980198</td>
<td>124.3323962</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918648_tbm084</td>
<td>66.51938503</td>
<td>70.66970091</td>
<td>68.78926554</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918649_tbm091</td>
<td>64.2073882</td>
<td>67.616241</td>
<td>65.44087782</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918650_tbm092</td>
<td>37.12299465</td>
<td>40.92088197</td>
<td>39.22460915</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918651_tbm107</td>
<td>103.0135135</td>
<td>125.2039604</td>
<td>114.5080011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918652_tbm111</td>
<td>46.32987686</td>
<td>111.4636835</td>
<td>96.00047611</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM918653_tbm143</td>
<td>57.5564409</td>
<td>64.3911053</td>
<td>60.76593228</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Percent present probesets for each array

<table>
<thead>
<tr>
<th>Array</th>
<th>Percent present</th>
<th>Array</th>
<th>Percent present</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM918578_mbt003.CEL.present</td>
<td>44.4040668</td>
<td>GSM918616_mbt098.CEL.present</td>
<td>51.4348425</td>
</tr>
<tr>
<td>GSM918579_mbt004.CEL.present</td>
<td>41.7576889</td>
<td>GSM918617_mbt099.CEL.present</td>
<td>49.4046392</td>
</tr>
<tr>
<td>GSM918580_mbt006.CEL.present</td>
<td>42.1474668</td>
<td>GSM918618_mbt102.CEL.present</td>
<td>55.74394147</td>
</tr>
<tr>
<td>GSM918581_mbt008.CEL.present</td>
<td>48.96021948</td>
<td>GSM918619_mbt103.CEL.present</td>
<td>52.10242341</td>
</tr>
<tr>
<td>GSM918582_mbt009.CEL.present</td>
<td>49.90580704</td>
<td>GSM918620_mbt105.CEL.present</td>
<td>49.70278921</td>
</tr>
<tr>
<td>GSM918583_mbt100.CEL.present</td>
<td>44.54869684</td>
<td>GSM918621_mbt106.CEL.present</td>
<td>53.31504344</td>
</tr>
<tr>
<td>GSM918584_mbt011.CEL.present</td>
<td>48.65477824</td>
<td>GSM918622_mbt109.CEL.present</td>
<td>55.52084076</td>
</tr>
<tr>
<td>GSM918585_mbt013.CEL.present</td>
<td>43.54092364</td>
<td>GSM918623_mbt110.CEL.present</td>
<td>49.01326017</td>
</tr>
<tr>
<td>GSM918586_mbt016.CEL.present</td>
<td>48.02194787</td>
<td>GSM918624_mbt124.CEL.present</td>
<td>50.44535894</td>
</tr>
<tr>
<td>GSM918587_mbt020.CEL.present</td>
<td>49.76863283</td>
<td>GSM918625_mbt126.CEL.present</td>
<td>50.12711477</td>
</tr>
<tr>
<td>GSM918588_mbt024.CEL.present</td>
<td>45.10105167</td>
<td>GSM918626_mbt127.CEL.present</td>
<td>48.94192958</td>
</tr>
<tr>
<td>GSM918589_mbt031.CEL.present</td>
<td>54.47828075</td>
<td>GSM918627_mbt135.CEL.present</td>
<td>53.90965020</td>
</tr>
<tr>
<td>GSM918590_mbt032.CEL.present</td>
<td>54.26794606</td>
<td>GSM918628_mbt136.CEL.present</td>
<td>46.37583734</td>
</tr>
<tr>
<td>GSM918591_mbt033.CEL.present</td>
<td>53.28395062</td>
<td>GSM918629_mbt140.CEL.present</td>
<td>46.65569273</td>
</tr>
<tr>
<td>GSM918592_mbt034.CEL.present</td>
<td>52.57064472</td>
<td>GSM918630_mbt141.CEL.present</td>
<td>49.43209877</td>
</tr>
<tr>
<td>GSM918593_mbt035.CEL.present</td>
<td>52.0877915</td>
<td>GSM918631_mbt144.CEL.present</td>
<td>44.32373114</td>
</tr>
<tr>
<td>GSM918594_mbt037.CEL.present</td>
<td>51.41838134</td>
<td>GSM918632_mbt145.CEL.present</td>
<td>47.3997257</td>
</tr>
<tr>
<td>GSM918595_mbt045.CEL.present</td>
<td>50.83127572</td>
<td>GSM918633_mbt146.CEL.present</td>
<td>53.1486054</td>
</tr>
<tr>
<td>GSM918596_mbt046.CEL.present</td>
<td>52.0877915</td>
<td>GSM918634_mbt147.CEL.present</td>
<td>51.1824417</td>
</tr>
<tr>
<td>GSM918597_mbt048.CEL.present</td>
<td>51.16780979</td>
<td>GSM918635_mbt148.CEL.present</td>
<td>48.8996283</td>
</tr>
<tr>
<td>GSM918598_mbt050.CEL.present</td>
<td>50.75628715</td>
<td>GSM918636_mbt149.CEL.present</td>
<td>52.87242798</td>
</tr>
<tr>
<td>GSM918599_mbt051.CEL.present</td>
<td>49.09190672</td>
<td>GSM918637_mbt150.CEL.present</td>
<td>51.3397348</td>
</tr>
<tr>
<td>GSM918600_mbt053.CEL.present</td>
<td>50.5294246</td>
<td>GSM918638_mbt151.CEL.present</td>
<td>51.57567444</td>
</tr>
<tr>
<td>GSM918601_mbt058.CEL.present</td>
<td>49.79972565</td>
<td>GSM918639_mbt156.CEL.present</td>
<td>47.08734245</td>
</tr>
<tr>
<td>GSM918602_mbt062.CEL.present</td>
<td>51.74577046</td>
<td>GSM918640_mbt158.CEL.present</td>
<td>51.92880402</td>
</tr>
<tr>
<td>GSM918603_mbt063.CEL.present</td>
<td>51.84087791</td>
<td>GSM918641_mbt161.CEL.present</td>
<td>58.77091907</td>
</tr>
<tr>
<td>GSM918604_mbt068.CEL.present</td>
<td>53.9478738</td>
<td>GSM918642_mbt166.CEL.present</td>
<td>52.64929127</td>
</tr>
<tr>
<td>GSM918605_mbt069.CEL.present</td>
<td>50.9236397</td>
<td>GSM918643_mbt167.CEL.present</td>
<td>54.15820759</td>
</tr>
<tr>
<td>GSM918606_mbt075.CEL.present</td>
<td>55.7968217</td>
<td>GSM918644_mbt168.CEL.present</td>
<td>52.10791038</td>
</tr>
<tr>
<td>GSM918607_mbt078.CEL.present</td>
<td>57.15043439</td>
<td>GSM918645_tbm055.CEL.present</td>
<td>54.95930498</td>
</tr>
<tr>
<td>GSM918608_mbt079.CEL.present</td>
<td>51.5635848</td>
<td>GSM918646_tbm061.CEL.present</td>
<td>50.47462277</td>
</tr>
<tr>
<td>GSM918609_mbt081.CEL.present</td>
<td>54.48925469</td>
<td>GSM918647_tbm082.CEL.present</td>
<td>48.75537266</td>
</tr>
<tr>
<td>GSM918610_mbt083.CEL.present</td>
<td>50.40329218</td>
<td>GSM918648_tbm084.CEL.present</td>
<td>44.06035665</td>
</tr>
<tr>
<td>GSM918611_mbt085.CEL.present</td>
<td>56.54869684</td>
<td>GSM918649_tbm091.CEL.present</td>
<td>43.58847337</td>
</tr>
<tr>
<td>GSM918612_mbt087.CEL.present</td>
<td>55.20987654</td>
<td>GSM918650_tbm092.CEL.present</td>
<td>51.22450846</td>
</tr>
<tr>
<td>GSM918613_mbt089.CEL.present</td>
<td>53.55647005</td>
<td>GSM918651_tbm107.CEL.present</td>
<td>50.11979881</td>
</tr>
<tr>
<td>GSM918614_mbt093.CEL.present</td>
<td>48.1156836</td>
<td>GSM918652_tbm111.CEL.present</td>
<td>51.16963877</td>
</tr>
<tr>
<td>GSM918615_mbt095.CEL.present</td>
<td>53.42844079</td>
<td>GSM918653_tbm143.CEL.present</td>
<td>38.95564701</td>
</tr>
</tbody>
</table>

The table displays the percent present probesets across all arrays. As displayed, an average percent present value is 50.7%. Good quality arrays have a PP value lying in a range of +/- 10% of the average witnessed percent. Arrays shaded in red have their PP values outside this range (44.6-55.8). This suggests potential bad quality of one of the replicates in those arrays. Normalization is a must to correct for erroneous factors and provide better quality.
Table 4: Top 10 DEGs in each subgroup:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>FC</th>
<th>Gene Name</th>
<th>FC</th>
<th>Gene Name</th>
<th>FC</th>
<th>Gene Name</th>
<th>FC</th>
<th>Gene Name</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAX6</td>
<td>-4.6422863</td>
<td>FOXG1</td>
<td>5.961362</td>
<td>KCNA5</td>
<td>-7.16143</td>
<td>SOSTDC1</td>
<td>4.989189</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCERG1L</td>
<td>-3.8065412</td>
<td>FSTL5</td>
<td>5.9670528</td>
<td>RD3</td>
<td>-5.96129</td>
<td>SFRP1</td>
<td>5.087735</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMNAT2</td>
<td>-3.7607834</td>
<td>GABRA5</td>
<td>4.8109953</td>
<td>OTX2</td>
<td>-5.84138</td>
<td>PRLR</td>
<td>4.920684</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOC</td>
<td>-3.7358627</td>
<td>HLX</td>
<td>3.9979123</td>
<td>USH2A</td>
<td>-5.43604</td>
<td>PPP2R2C</td>
<td>5.613351</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UBASH3B</td>
<td>-3.6893604</td>
<td>IMPG2</td>
<td>5.358301</td>
<td>DAB1</td>
<td>-5.41913</td>
<td>NRIP2</td>
<td>4.710584</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAIM2</td>
<td>-3.6189982</td>
<td>NPR3</td>
<td>4.4193034</td>
<td>ZNF385B</td>
<td>-5.13656</td>
<td>NDST3</td>
<td>4.862195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMNAT2</td>
<td>-3.7607834</td>
<td>GABRA5</td>
<td>4.8109953</td>
<td>OTX2</td>
<td>-5.84138</td>
<td>PRLR</td>
<td>4.920684</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOP1MT</td>
<td>3.85143</td>
<td>RALYL</td>
<td>4.544383</td>
<td>TLE2</td>
<td>-4.41868</td>
<td>RBM20</td>
<td>5.242945</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNFRSF19</td>
<td>-3.80551</td>
<td>PRL</td>
<td>4.671902</td>
<td>AMER3</td>
<td>-4.03934</td>
<td>PDE6C</td>
<td>5.491733</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOL4</td>
<td>-3.50522</td>
<td>KCNA1</td>
<td>4.103352</td>
<td>PTPRO</td>
<td>-3.83057</td>
<td>DHRS2</td>
<td>5.679356</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCX</td>
<td>-6.85893</td>
<td>WIFI1</td>
<td>8.792614</td>
<td>ENC1</td>
<td>-4.73562</td>
<td>RUNX2</td>
<td>6.290047</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOXG1</td>
<td>-6.262</td>
<td>TRDC</td>
<td>8.192017</td>
<td>INSM2</td>
<td>-4.71569</td>
<td>LAMP5</td>
<td>5.688442</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOL4</td>
<td>-5.08418</td>
<td>TNFRSF11B</td>
<td>5.569707</td>
<td>CAMSAP3</td>
<td>-4.48595</td>
<td>GAD1</td>
<td>6.551833</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNRN1</td>
<td>-4.93759</td>
<td>SHOX2</td>
<td>6.58027</td>
<td>ZNF804A</td>
<td>-4.31666</td>
<td>EMX2</td>
<td>6.570035</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Medulloblastoma molecular subgroups

<table>
<thead>
<tr>
<th>Points of difference</th>
<th>WNT subgroup</th>
<th>SHH subgroup</th>
<th>Subgroup 3</th>
<th>Subgroup 4</th>
<th>Subgroup 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prognosis</strong></td>
<td>Best</td>
<td>Second best</td>
<td>Poorest</td>
<td>Second poor</td>
<td>as subgroup 4</td>
</tr>
<tr>
<td><strong>Molecular signature</strong></td>
<td>WNT pathway activation</td>
<td>SHH pathway activation</td>
<td>upregulation of genes involved in retinopathy</td>
<td>slight upregulation of genes involved in retinopathy</td>
<td></td>
</tr>
<tr>
<td><strong>Histology</strong></td>
<td>Classic</td>
<td>Desmoplastic/LCA</td>
<td>LCA/classic</td>
<td>Desmoplastic/classic</td>
<td></td>
</tr>
<tr>
<td><strong>Cytogenetic events</strong></td>
<td>chromosome 6 monosomy</td>
<td>3q+, 9p+, 9q-, 17p-</td>
<td>1q+, 7+, 17q+, 8-, 11-, 16q-, 17p-</td>
<td>7+, 17q+, 17p- (17q isochromosome)</td>
<td>7+, 17q+, 17p- (17q isochromosome)</td>
</tr>
<tr>
<td><strong>Gene mutations</strong></td>
<td>DKK1, DDX3X, CTNNB1, SMARCA4</td>
<td>PTCH1, SUFU</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Upregulated genes</strong></td>
<td>CTNNB1, LEF1, WNT13, WNT16, BMP4</td>
<td>PTCH1, HHIP, BOC, GLI1/2</td>
<td>NEUROD1, ETV1, RTN4R, And MAST1, MAP2, MPP3, High MYC, NPR3</td>
<td>UNC5D, EPHA8, RB1, S-antigen, Opsin (retinal TF), SMARCD, GABRAS</td>
<td></td>
</tr>
<tr>
<td><strong>Down regulated genes</strong></td>
<td>DKK1, DKK2, WIF1</td>
<td>BMP2/4, GSK3</td>
<td>–</td>
<td>No MYC differential expression</td>
<td>FOXG1B, EOMES</td>
</tr>
</tbody>
</table>

Table 5: Summary for current markers associated with molecular subgroups of medulloblastoma
Figure 44: Differences between histological medulloblastoma subgroups.

Figure showing different histological patterns for each medulloblastoma histological subgroup. First photo from the left resembles large cell anaplastic medulloblastoma. In the middle is desmoplastic medulloblastoma. At the right is classical type medulloblastoma tissue.

Figure 45: Steps of gene expression profiling applied in this work.
Figure 46: 2D colored images of 76 microarray CEL files.
Figure 47: Prenormalization MA plots (pages 159:167)
GSE37418: MA plots for prenormalized data.
GSE37418: MA plots for prenormalized data.
GSE37418: MA plots for prenormalized data.
GSE37418: MA plots for prenormalized data.
Figure 48: Post normalization MA plots for 76 medulloblastoma arrays. (Pages 167-174)
GSE37418: MA plots for postnormalized data.
GSE37418: MA plots for postnormalized data.
GSE37418: MA plots for postnormalized data.
GSE37418: MA plots for postnormalized data.
9. References:


161

failing heart. *Circulation Research, 102*(6), 633–6. doi:10.1161/CIRCRESAHA.107.165183


196. Montasr, M. M., Aziz, F. E., & Ahmed, D. (2013). Β-Catenin Expression in Medulloblastoma: Prognosis and Favorable Outcome. *Research in Neuroscience*. Scientific & Academic Publishing. Retrieved from In our era of technology and progress in medical molecular biological science, it’s now obligatory to think extraordinary to deal with the brain tumors, reaching the best prognosis for a better outcome. We focused on proving the strong relation between the subtypes of medulloblastoma and the magnificent rule of B-catan in nucleo-positivity in a Wnt/ wg pathway, associated with mutations like APC, CTNNB1 as a good diagnosis and prognosis of Medulloblastoma, It’s mandatory applying these good prog


genes. *The Journal of Clinical Investigation, 116*(12), 3171–82. doi:10.1172/JCI29401


