# Deciphering Genetic Drivers in Primary and Metastatic Medulloblastoma

by

Patryk Skowron

A thesis submitted in conformity with the requirements

for the degree of Doctor of Philosophy

Laboratory Medicine and Pathobiology
University of Toronto

# Deciphering Genetic Drivers in Primary and Metastatic Medulloblastoma

Patryk Skowron

Doctor of Philosophy

Laboratory Medicine and Pathobiology

University of Toronto

2020

## Abstract

Sonic hedgehog medulloblastoma (Shh-MB) encompasses a clinically and molecularly diverse group of cancers of the developing central nervous system. It initiates within the cerebellum and, in 20% of cases, disseminates throughout the brain and spinal cord. Current therapy consists of maximal safe resection, radiotherapy in patients over 36 months, and cytotoxic chemotherapy. Unbiased sequencing of the transcriptome across a large cohort of 250 primary tumors reveals differences between molecular subtypes of the disease, with a previously unappreciated importance of non-coding RNA transcripts. Analysis of a large cohort of a single molecular type of cancer allows for identification of novel genes with single nucleotide variants (*MYCN*, *GNAS*, *IKBKAP*, and *KDM6A*) as well as gene fusions, some of which are secondary to rearrangement of the genome (*ZBTB20*, *NCOR1*), while others appear to arise through trans-splicing (*RALGAPA2*, and *GNAS*). Integration of genetic and transcriptomic data allows further differentiation of driver from passenger genes. Molecular convergence on a core of specific genes by nucleotide variants, copy number aberrations, and gene fusion further emphasize the key role of specific pathways in the pathogenesis of primary Shh-MB. Little is known about genes driving

metastatic progression since matching human primary and metastatic samples are rare. The Shh-MB Sleeping Beauty (SB) mouse model uses random integration of transposons to initiate tumorigenesis and drive the metastatic cascade providing valuable insight onto the human disease. Common insertion site analysis using 549 metastatic tumors from 131 mice reveal networks of recurrent metastatic drivers (n = 336) and demonstrate extensive heterogeneity between metastasis. A subset of drivers, such as loss-of-function events in *Crebbp* and *Ctnna3,* arise independently between metastasis and are under the pressure of convergent evolution. Recurrent gain-of-function insertions in *Lgalg3* suggest an oncogenic role in metastatic progression which was validated using *Lglas3* knockout experiments in multiple models. Mice missing copies of *Lgals3* had no change in metastatic burden in the brain but showed significantly less metastasis along the spinal cord suggesting a site-specific role as a metastasis driver gene. These findings enhance our understanding of the genomic complexity and heterogeneity underlying Shh-MB pathogenesis and highlight several targets for therapeutic development.

# Acknowledgments

I would like to thank funding from the Terry Fox Foundation and Genome Canada for supporting these important projects.

I am thankful for my supervisors Dr. Michael Taylor and Dr. Gary Bader for the excellent mentorship and scientific guidance through the years. I would also like to thank my committee members, Dr. Sean Egan and Dr. Mathieu Lupien for their important insights and advice.

I would have never been able to finish my projects without the full support of my colleagues. First, I would like to thank Dr. Livia Garzia and Dr. Sorana Morrissy for taking me under their wing and training me when I first started in the lab. Also, thank you Hamza Farooq, Dr. Florence Cavalli, Dr. Hiromichi Suzuki, Michelle Ly, Raul Suarez, Betty Luu, and Dr. Kevin Wang for all the help bringing these projects to life.

I would lastly like to thank my family and friends for standing by me for all these years. Thank you to my girlfriend, Alice Zhang for all the love and support. Thank you to my parents and brother for always pushing me to do the absolute best that I possibly can.

# Table of Contents

# List of Tables

# List of Figures

# List of Appendices

# CHAPTER 1
Medulloblastoma molecular biology and mouse models

## 1.1 MEDULLOBLASTOMA MOLECULAR GENETICS[1]

Patryk Skowron*, Vijay Ramaswamy, Michael D. Taylor

Medulloblastoma is the most common malignant paediatric brain cancer, having an incidence of approximately 0.74/100,000 person-year[2,3]. This tumor is located in the cerebellum and 30% of cases present with metastatic dissemination over the cranial and spinal leptomeninges (Figure 1.1)[4]. Initial treatment for medulloblastoma is maximal safe surgical resection followed by adjuvant craniospinal irradiation and/or high dose cytotoxic platinum based chemotherapy. Radiation as per current protocols in North America and Western Europe is risk adapted, in that, metastatic patients receive 36 Gy and non-metastatic patients receive 23.4 Gy of craniospinal irradiation with a boost to the tumour bed[5]. With the current standard of care, overall patient survival has reached 70%, however, metastatic patients and infants are both high risk groups with poor survival[4,6–9]. Despite successful completion of treatment, patients frequently present with neurocognitive sequelae — long term neurological deficits in cognition[10,11]. As such, there is an urgent need for more specific targeted therapies which minimize impact on the developing brain.

Recent integrated genomic studies have now shown that medulloblastoma is not one single morphological entity, and is in fact at the molecular level, comprised of several different diseases. Large scale efforts focused on studying the transcriptional landscape have revealed 4 distinct subgroups (WNT, SHH, Group 3, Group 4), each with their own unique survival, age demographics, and genetic aberrations[12,13]. These subgroups are stable at recurrence and across tumour compartments[6,14], and are likely reminiscent of the cell of origin[15]. The next generation of clinical trials is already taking subgroup into account to rationally stratify patients and tailor

therapy. Accurate, robust, and inexpensive subgroup prediction methods are essential; molecular subgroups can be reliably assigned by either expression profiling or through the use of genome wide methylation arrays[16–18]. Further investigation into the molecular genetics of medulloblastoma will hopefully pave the way for new targeted therapeutic strategies to cure this devastating childhood disease.

### 1.1.1 Familial Predisposition Syndromes

Initial insights into the pathways driving medulloblastoma were inferred from familial predispositions associated with medulloblastoma. The most common being Li-Framumeni syndrome with germline mutations in *TP53*[19]. These mutations can drive a variety of other cancers, but in medulloblastoma both somatic and germline *TP53* mutations are frequently present in childhood SHH patients and are known to facilitate catastrophic large scale rearrangements via chromothripsis[20–22]. Somatic *TP53* mutations can also occur in the WNT subgroup. Less frequent is Gorlin syndrome which is an autosomal dominate disease characterised by mutations of the transmembrane receptor Patched1 (*PTCH1*). The majority of these patients will acquire basal cell carcinoma, while about 5-20% will get medulloblastoma[23,24]. Deletion of the *PTCH1* locus results in higher Smoothened (SMO) activity and upregulation of the Sonic Hedgehog (Shh) signalling pathway, a marker of the SHH subgroup. Less common predispositions are: i) Turcot Syndrome adenomatous polyposis coli (*APC*) germline mutations which are associated with a multitude of other central nervous system tumours and colorectal cancer[25,26] , and ii) autosomal dominant mutations in CREB binding protein (*CREBBP*) causing Rubinstein-Taybi syndrome[27]. Familial predispositions are not all encompassing and only account for about 6% of medulloblastoma cases[28]. There are many other genetic factors which can lead to the development of medulloblastoma which will be covered in the sections that follow.

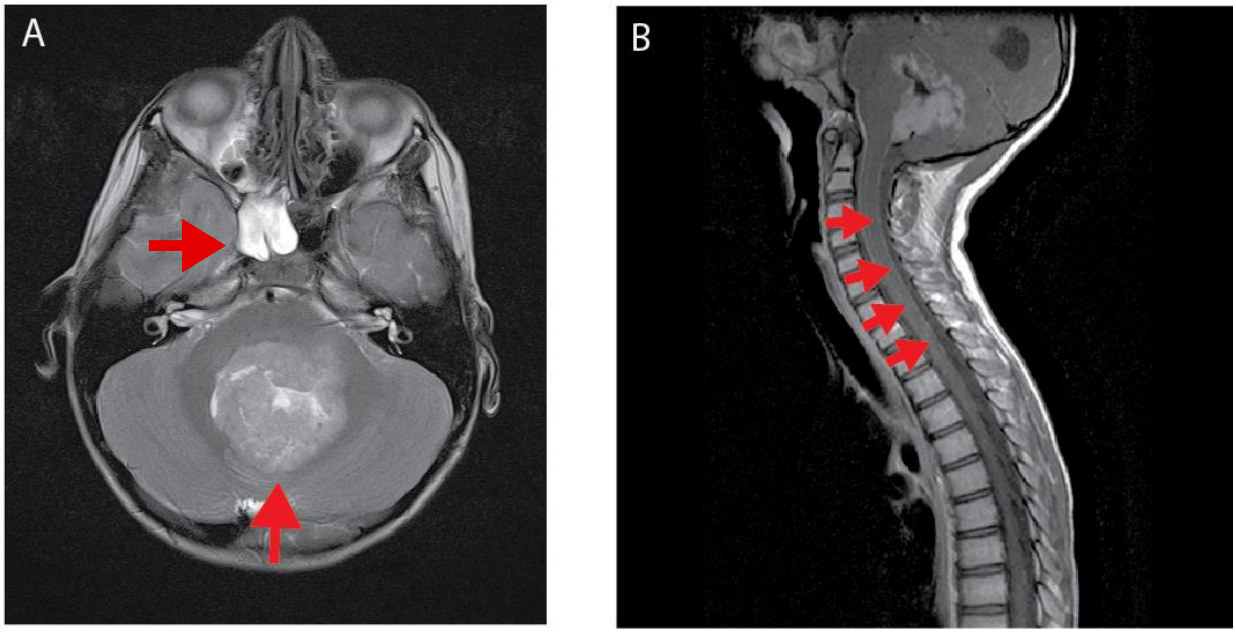



**Figure 1.1 Medulloblastoma primary and metastatic tumors**
**(a)** MRI of pediatric medulloblastoma primary tumor (lower arrow) and the secondary brain metastatic site (upper arrow). **(b)** Spinal metastatic spread along spinal cord in the same patient as **(a)**.

## 1.1.2 Wnt subgroup

### *1.1.2.1 Clinical Attributes*

Of all subgroups, the WNT subgroup has the most favourable prognosis with over 95% of patients surviving their disease (Table 1). WNT tumors exhibit classic histology, are rarely metastatic and have an even gender predisposition. WNT is the least common subtype, with a rate of 10% among medulloblastoma patients. The hallmark alteration in WNT tumors is somatic activating mutations in exon 3 of β-catenin (*CTNNB1*). Monosomy 6 is the main recurrent structural alteration and is usually found in an otherwise balanced genome[29–31].

### *1.1.2.2 Molecular Biology*

The Wnt signalling pathway plays an essential role in embryonic development, controlling cell fate specification, cell proliferation, cell migration and body axis patterning. In the developing

brain, the Wnt pathway has broad regulatory effects on neuronal maturation and synapse formation. This pathway is activated through binding of WNT ligands to Frizzled receptors, which relay signals into the nucleus through induced stabilization of β-catenin (Figure 1.2a). Important negative regulators of this pathway are APC and SUFU which normally limit β-catenin accumulation and translocation into the nucleus[32,33]. Nearly all (90%) of WNT tumors have somatic missense mutations in *CTNNB1,* the gene coding for β-catenin, which promote protein stabilization. The next most common mutation is in *DDX3X*, with mutations clustering in its two helicase domains hypothesized to alter its RNA binding capacity rather than abolish it. *In vivo* and *in vitro* functional studies on *DDX3X* suggest that it enhances and/or maintains proliferation of the WNT progenitor cells. It is also possible that *DDX3X* mutations corporates with β-catenin activation[34–36]. Also commonly found in WNT are missense mutations in *TP53*. Despite being a marker of high risk in the SHH subgroup and other cancers, *TP53* mutations confer no difference in survival for patients diagnosed with WNT subgroup medulloblastomas[22].

#### *1.1.2.3 Models*

Progenitors of the lower rhombic lip are the likely cell of origin for WNT tumours. β-catenin stabilization and nuclear localization is the most characteristic feature of WNT subgroup tumors and in mouse models its action is not sufficient to transform external granule cells, which are the cells of origin for SHH subtype tumors. Furthermore, WNT tumours in humans are found adjacent to the brainstem, unlike SHH tumors which arise from within the cerebellum. During development, postmitotic mossy-fibre neuron precursors in the dorsal brainstem migrate into the central brainstem. Targeted expression of activated beta-catenin in mouse postmitotic mossy-fibre neuron precursors using a brain lipid-binding protein (*Blbp*) promoter, coupled with a knockout of

*TP53* leads to formation of a WNT tumour with long latency and low penetrance[37]. Subsequent work established that through addition of a phosphoinositide 3-kinase (*PI3K*) catalytic-α polypeptide mutant allele (*Pik3ca*$^{E545K}$) identified in WNT medulloblastomas, the penetrance in the mouse model was increased to 100% with highly representative WNT tumours forming within 3 months[35,38].

## 1.1.3 SHH subgroup

### *1.1.3.1 Clinical attributes*

The SHH subgroup accounts for a third of all medulloblastoma tumors and has an intermediate prognosis with a five year survival ranging between 60-80%. The age distribution for this tumor is bimodal, with the majority of infant and adult medulloblastomas being SHH. The histological classification can be any of the 5 described variants from the WHO classification system; however the desmoplastic variant is more common in children and adults compared to infants. Large cell and/or anaplastic histology is common in children harbouring germline or somatic mutations in *TP53*. SHH patients commonly have focal amplifications of *GLI2*, and *MYCN*, as well as loss of 17p (Table 1)[5,13,30,39].

### *1.1.3.2 Molecular Biology*

During early cerebellar development, Purkinje cells release Shh ligand and stimulate the proliferation and subsequent migration of granule cells into the internal granule cell layer. Excessive activation of the Shh pathway overdrives the expression of GLI2 transcription factor targets which induce uncontrolled proliferation of granule cells and the formation of tumour[33,40]. Alterations in this subgroup most often fall within the Shh signalling pathway and, less frequently, in cooperating pathways such as PI3K and mTOR (Figure 1.2b). The most common events are

somatic or germline inactivating alterations of *PTCH1* or *SUFU*, or somatic missense mutations activating SMO[31,34–36]. A subset of high-risk patients present with co-amplification of *MYCN* and *GLI2* accompanied by inactivation of *TP53*. Within the SHH subgroup there is also a difference in molecular biology and risk factors for different age groups. *SUFU* mutations are found predominantly in infants, while the high risk *GLI2* amplifications are found in older children and teenagers[21,41]. In adults, the most common are somatic mutations in *SMO* and in the *TERT* promoter (C228T or C250T)[42], which creates an E-twenty-six binding motif [43,44].

#### *1.1.3.3 Models*

There are a large variety of mouse models that recapitulate SHH subgroup tumors, and these function mainly through dysregulation of the hedgehog signalling pathway. For example, the first medulloblastoma mouse model involved a single allele knockout of the *PTCH1* gene, a negative inhibitor of SMO, which drives tumorigenesis in granule cells[45]. Since then there have been other models whereby *Ptch1*$^{+/-}$ was crossed with other aberrations that confer a more aggressive phenotype, such as deletions of cyclin-dependant kinases *Ink4c* and *Kip1*[46,47], or the master regulator *TP53*[48]. NeuroD2 dependant overexpression of mutant SMO in granule cells is also able to drive highly penetrant tumours with leptomeningeal metastasis[49,50]. In addition, even though SHH medulloblastoma are traditionally thought to arise from granule cells, there have been mouse models that demonstrate aberrant Shh signalling induced tumors in cochlear nuclei and neural stem cells[51,52].

A model that has shown great utility in screening for novel driver genes and cooperating events has been the medulloblastoma Sleeping Beauty (SB) mouse model[53] which utilizes random transposon integration to drive tumorigenesis. SB transposons contain elements which are capable

of overexpressing or truncating genes depending on the insertion location and orientation. Insertion events are mediated by a transposase, which is limited to granule cell precursors through use of the *MATH1* promoter. Nearly all the mice in this model develop tumours with a high rate of leptomeningeal metastasis by 3 months. The SB system has identified a large number of primary tumour drivers such as MyoD[54] and Nfia[55], and has also revealed the large degree of divergence between primary and metastatic tumours (discussed below).

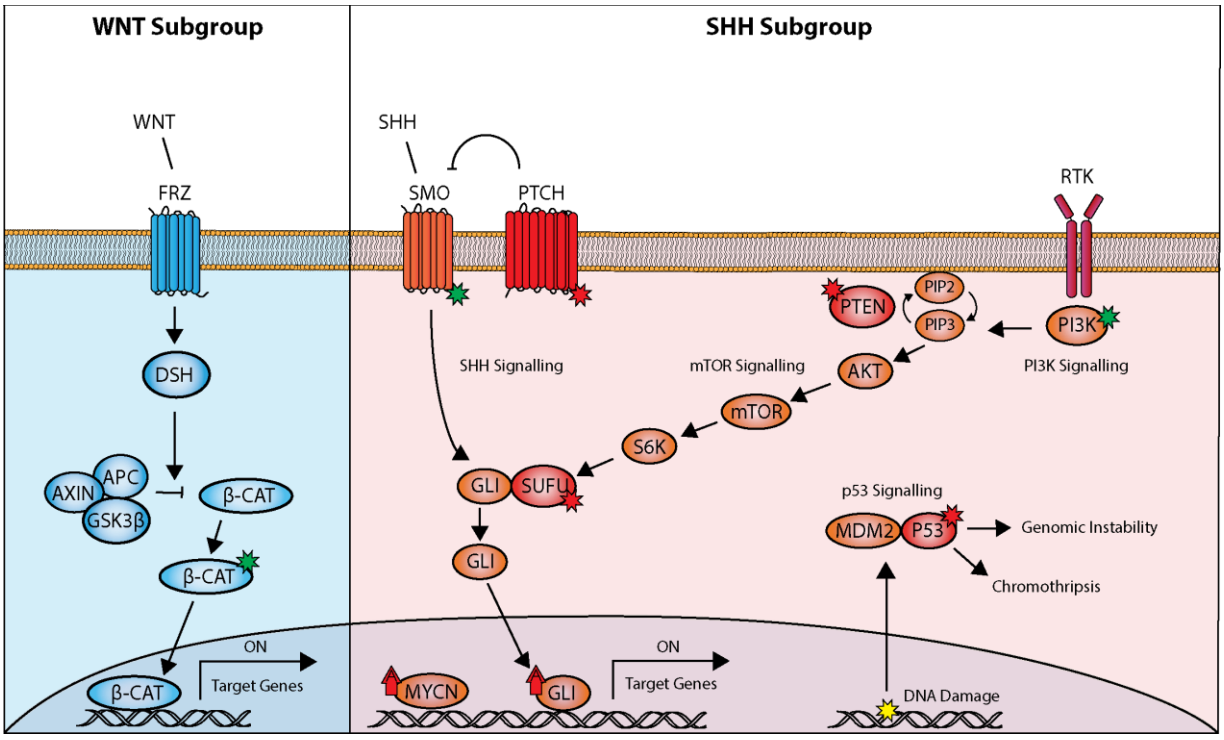



**Figure 1.2 Dysregulated pathways in WNT and SHH medulloblastoma**
**(a)** WNT tumors normally have activating alteration in β-cat which promote its stabilization and allow it to upregulate target genes. **(b)** Alterations in the SHH subgroup usually fall within the Shh signalling as well as cooperating PI3K/mTOR pathways and converge on upregulation of GLI. The most common are inactivating alterations in PTCH or SUFU or activating mutations in SMO. High risk patients typically have co-amplifications of MYCN, GLI2 and mutations in P53 which results in genomic instability and/or chromothripsis[31,34–36,56]. Activating mutations (green star); inactivating mutations (red star); amplifications (red arrow); DNA damage (yellow star); amplification (up arrow).

## 1.1.4 Group 3

### *1.1.4.1 Clinical Attributes*

Group 3 medulloblastoma comprise about 20% of all cases (Table 1). These patients have the worst survival and the highest rate of metastatic dissemination. Group 3 tumors recur almost exclusively with metastatic dissemination with a clean tumour bed[6]. Patients diagnosed with this subgroup are commonly infants and younger children with a male to female discordance of 2:1. The histology of this tumour is commonly classic or large cell anaplastic (LCA) and the genome of these tumours is very unbalanced with a large number of broad alterations such as gain of chromosome 7 and isochromosome 17q. Many of these alterations are also shared with Group 4.

### *1.1.4.2 Molecular Biology*

There are several recurrent somatic copy-number alterations in Group 3 medulloblastoma, but unusually few recurrent single nucleotide variants and indels. The Group 3 transcriptome is dominated by photoreceptor and GABAergic expression signatures[12]. The most common event is amplification of *MYC* in 10-20% of patients, which in many cases occurs with a fusion between the promoter of *PVT1* and the second exon of *MYC*[31]. In many cancers the *MYC* locus is co-amplified with the non-coding RNA *PVT1*, which is able to stabilize MYC protein[57]. In medulloblastoma, these fusions likely create a positive feedback loop since the *PVT1* promoter contains canonical E-boxes which are activated by MYC[58]. Amplification of the gene coding for the transcription factor OTX2 is another common copy number alteration occurring in 10% of patients, and mutually exclusive of *MYC* amplification. OTX2 is known to play an important role in controlling cell fate and differentiation of various progenitors in the developing brain and is able to repress the myogenic differentiation of medulloblastoma cells. It also plays a role in the TGF-

B signalling pathway which contains numerous other genes showing less recurrent copy-number alterations indicating that deregulation of this pathway may be a driver event[59–61].

#### *1.1.4.3 Models*

Two orthotopic transplantation models of Group 3 have been created which couple overexpression of MYC with inactivation of *TP53*[62,63]. MYC expression leads to a higher rate of proliferation as well as a higher rate of TP53 mediated apoptosis which necessitates the need to inactivate the *TP53* locus. In the first model, Pei *et al* isolated mouse CD133-positive and glial lineage marker-negative neural stem cells from the postnatal cerebellum[64]. These cells were unresponsive to Shh stimulation and capable of differentiating into neurons, astrocytes and oligodendrocytes. Infection of these cells with a stabilized MYC (MycT58A) followed by transplantation into a mouse led to formation of transient hyperplastic lesions with a high rate of apoptosis. By introducing a dominate negative *TP53* virus (DNp53), the apoptotic effects were abolished and tumours were formed with LCA histology, prominent necrosis, and nuclear molding. The second model was produced by isolating GFP florescent granule progenitor cells from postnatal TP53 deficient Atoh1-GFP mice. Atoh1-GFP isolated cells transduced with MYC were able to form aggressive tumours with LCA histology, even after multiple passages in mice. In Group 3, focal events in *TP53* are exceedingly rare but there is frequent loss of 17p (where TP53 resides). The resistance to Shh pathway inhibition and the similarity in Group 3 signature genes suggest that these two models are highly representative of the human disease.

## 1.1.5 Group 4

### *1.1.5.1 Clinical Attributes*

Group 4 is the most common of the medulloblastoma subgroups and has an intermediate 5 year overall survival of 75% (Table 1). Group 4 tumor histology is most commonly classic. It has a high rate of metastasis and a 2:1 male to female discordance. The Group 4 genome is commonly tetraploid, and the most common structural alteration is isochromosome 17q, which is found in 80% of tumors.

### *1.1.5.2 Molecular Biology*

Group 4 has a neuronal and glutaminergic expression signature and, like Group 3, few recurrent single nucleotide variants and indels. The most frequently mutated somatic gene in Group 4 medulloblastoma is *KDM6A*, a histone H3 Lys27 (H3K27) demethylase, with nonsense mutations in 13% of patients[65–67]. KDM6A belongs to the Jumonji C family of histone demethylases along with KDM6B, which is also mutated in medulloblastoma. The proto-oncogenes *MYCN* and cyclin-dependant kinase *CDK6* are recurrently amplified in Group 4. More common are focal amplifications/tandem duplications of the alpha-synuclein interacting protein (*SNCAIP*) gene at chromosome 5q23[31], which encodes a protein previously implicated in Parkinson's disease[68]. It is still unknown if *SNCAIP* is the driver for these patients and more research needs to be done to uncover its specific role in Group 4 medulloblastoma.

### *1.1.5.3 Models*

Due to the low number of focal events and many broad rearrangements, the search for a Group 4 model has proven elusive. *MYCN* is commonly upregulated in medulloblastoma and is the site of one of the most recurrent focal amplifications in Group 4. Swartling *et al*. created a

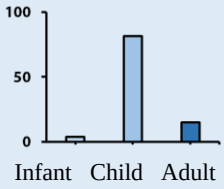
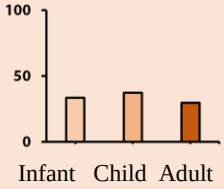
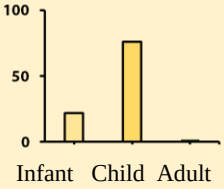
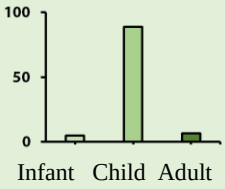
mouse model for *MYCN* driven medulloblastoma by targeting its expression with *Glt1*, a brain specific promoter highly expressed in the cerebellum throughout development until adulthood. Tumors formed with a long latency, had a low metastatic rate, and had either classic or LCA histology. *MYCN* was required for both initiation and maintenance of these tumours and was likely cooperating with other events since the genome was unbalanced and had a number of recurrent copy number alterations. These mice are showing great promise as a Group 4 model, but additional studies need to be performed in order to characterize expression signatures and identify the cell of origin[38].

### 1.1.6 Epigenetics

Across all subtypes there have been a number of recurrent somatic single nucleotide and copy number variants identified within genes coding for chromatin modifiers. Most common are truncating mutations in myeloid/lymphoid or mixed-lineage leukemia protein 2 (*MLL2*) and *MLL3* suggesting a role as an oncogenic driver. In Group 3 and Group 4 there are a large variety of recurrent somatic mutations in *SMARCA4* (exclusive to Group 3), and *KDM6A* (exclusive to Group 4), and less commonly in *CHD7*, *ARID1B*, *KDM4C*, and *ZMYM3*[31,34–36]. There is also over-expression of EZH2, a H3K27 methyltransferase that is part of the polycomb repressive complex essential for regulating development and differentiation. These events are largely mutually exclusive of each other and with amplifications of *MYC* or *MYCN*. The mechanism of their pathogenesis is still a subject of intense investigation, but it's possible that these events preserve Group 3 and Group 4 tumours in a stem cell-like state by maintaining high levels of the H3K27me3 repressive mark (EZH2 upregulation or KDM6A inactivation) and/or disruption of H3K4me3 associated transcription (ZMYM3 and CHD7 inactivation)[60,69,70].

**Table 1 Clinical and Genomic Characteristics of Medulloblastoma Subgroups**

Percentages indicate the recurrence within the respective subgroup. In the gender distribution, pink is female and blue is male. Data acquired from various sources[6,21,30,31,34–36].

| | WNT | SHH | Group 3 | Group 4 |
|---|---|---|---|---|
| **Age Distribution** |  |  |  |  |
| **Gender (f\|m)** |  |  |  |  |
| **Histology** | Classic, Rarely LCA | Desmoplastic, Classic, LCA | Classic, LCA | Classic, LCA |
| **Metastatic Rate** | Low | Low | High | High |
| **Prognosis** | Excellent | Intermediate | Poor | Intermediate |
| **SCNA** | - | MYCN (12%)<br>GLI2 (8%) | MYC (17%)<br>PVT1 (12%)<br>OTX2 (8%) | SNCAIP (10%)<br>MYCN (6%)<br>CDK6 (5%) |
| **SNVs** | CTNNB1 (91%)<br>DDX3X (50%)<br>SMARCA4 (26%)<br>MLL2 (13%)<br>TP53 (13%) | TERT (60%)<br>PTCH1 (46%)<br>SUFU (24%)<br>MLL2 (16%)<br>SMO (14%)<br>TP53 (13%) | SMARCA4 (11%)<br>MLL2 (4%) | KDM6A (13%)<br>MLL3 (5%) |
| **Broad Events** | 6 Loss | 3q Gain<br>9q, 10q, 14q Loss | 1q, 7, 17q, 18q Gain<br>8, 10q, 11, 16p, 17p Loss | 7, 17q, 18q Gain<br>8, 11p, X Loss |
| **Expression** | WNT Signaling | SHH Signaling | MYC/Retinal Signature | Neuronal Signature |
| **Recurrence** | - | Local | Metastatic | Metastatic |

Enhancer-promoter interactions play an essential role in tissue specific regulation of genes and development[71]. The three-dimensional localization of active enhancers ultimately determines which genes can be activated by the enhancer and any disruption can lead to aberrant regulation of genes. In Group 3 and Group 4 medulloblastoma it has recently been demonstrated that structural rearrangements can alter the intended targets of enhancers to drive tumorigenesis[72]. In particular, a series of seemingly unrelated deletion, duplication and translocation events were able to activate expression of transcription factors *GFI1* or *GFI1B* through repositioning of distal enhancers. These somatic events were highly recurrent, constituting a third of Group 3 and 10% of Group 4 tumours. When these genes were co-expressed with *MYC* in murine neural stem cells, they induced the formation of an aggressive tumour in recipient mice with a high rate of metastasis.

### 1.1.7 Metastasis

In medulloblastoma patients, metastasis is a sign of dismal prognosis. It is most commonly seen in patients with Group 3 and Group 4 tumors, both of which almost exclusively recur with metastatic dissemination[6]. Little is known about the genes driving dissemination and the context in which they operate since matching patient primary and metastatic samples are exceedingly rare. Studies with the SB mouse model have shown a large genetic divergence between the primary and metastatic compartments with almost no overlap in common insertion sites[53] suggesting that the primary tumour is a poor indicator of the therapeutic targets present in the metastatic lesions. These *Ptch1*-driven SB models have revealed a number of metastasis drivers such as *Eras*, *Lhx1*, *Ccrk*, and *Gdi2* which likely drive dissemination in SHH patients[73,74]. Tumour-stromal interactions also appear to play an essential role in medulloblastoma tumorigenesis and metastatic dissemination. For example, tumour cell-induced expression of the placental growth factor (PIGF) in the stroma

was shown to activate pro-survival pathways through the Nrp1 receptor[75] and promote tumour growth and metastasis.

### 1.1.8 Therapies

WNT subgroup tumors have the best prognosis out of all the subgroups with nearly all patients surviving after surgery, radiation and chemotherapy. For this subgroup there have been international efforts to de-escalate therapy to help reduce the long term cognitive deficits common in children after receiving radiotherapy[76].

There are a number of proposed therapies for SHH patients all of which aim to lower aberrant activation of Shh signalling. One of the most promising class of drugs are SMO inhibitors, which are already in phase II clinical trials for a number of cancers including medulloblastoma[77–79]. Unfortunately, acquired resistance inevitably occurs in both animals and humans. A recurrent mutation in a conserved aspartic acid residue within the G protein–coupled receptor domain of SMO has been shown to disrupt functionality of inhibitors while leaving Shh signalling intact[77,78]. Furthermore, the drug is only effective for patients with alterations within or upstream of SMO, and therefore high-risk children with amplifications of *MYCN* and *GLI2* would not benefit from such treatment[21]. Another class of drugs called bromodomain inhibitors may circumvent this problem by inhibiting the Shh signalling at the level of GLI2. BRD4 is a bromodomain protein which binds to ε-N-lysine acetylation motifs on open chromatin and is known to facilitate transcription at promoter regions of gene including GLI2 and MYC. Treatment of SHH with BRD4 inhibitors has shown great promise in pre-clinical models even in the presence of *SMO* drug resistance mutations[80,81].

The search for Group 3 and 4 tumor specific therapies has proven elusive. In Group 3, TGF-beta signaling is commonly dysregulated and pathway antagonists are already being explored for a multitude of cancers, including glioblastoma with varying success[82]. MYC inhibition is another potential but challenging therapeutic strategy. While there are no known direct inhibitors of MYC — studies have focused on inhibiting expression of MYC RNA[83] or inhibiting its heterodimer MAX[84,85]. So far the most promising approach is based on BRD4 inhibition using bromodomain inhibitor JQ1 to reduce the *MYC* transcription[86]. There is also some evidence that JQ1 may be effective for treatment of MYCN driven neuroblastoma in pre-clinical models; suggesting that it could also be effective for Group 4 patients[87]. In both Group 3 and Group 4, epigenetic alterations are a characteristic feature and there are a number of approved drugs in clinical trials for several adult and pediatric brain tumours. In particular, there are several inhibitors of the polycomb repressive complex 2 as well as EZH2 which act to decrease the level of H3K27me3[88].

## 1.2 SLEEPING BEAUTY TRANSPOSITION SYSTEM

### 1.2.1 Transposon Design

A transposon is a mobile genetic element which has the ability to "jump" around the genome. These translocation events are mediated through either reverse transcription of transposon mRNA into DNA or through a transposase enzyme mediated "cut and paste" mechanism (

Figure 1.3). Transposon classes differ by the DNA motifs in which insertion events can occur. Members of the Tc1/mariner "cut-and-paste" transposon superfamily were isolated from fish and were initially transpositionally inactive due to gradual acquisition of mutations. By comparing related inactive transposons across fish species, *Ivics et al*. generated a consensus

sequence for the active ancestor transposon. Then a new artificial element was made in a step-wise fashion to match the consensus creating the first generation of Sleeping Beauty (SB) transposons[89]. This system was later modified to increase transposition rates and carry elements which can either overexpress or truncate mRNA if inserted into a gene. The T2onc2 (aka SB100) sleeping beauty system contains a Murine Stem Cell Virus (MSCV) long terminal repeat (LTR) to drive expression of downstream transcripts in one orientation, as well as an SV40 polyadenylation sequence to truncate mRNA from either orientation. The transposase enzyme mediates "cut-and-paste" integration into TA dinucleotides through interaction with right and left inverted repeat regions (IRR and IRL) of the transposon (Figure 1.4a).

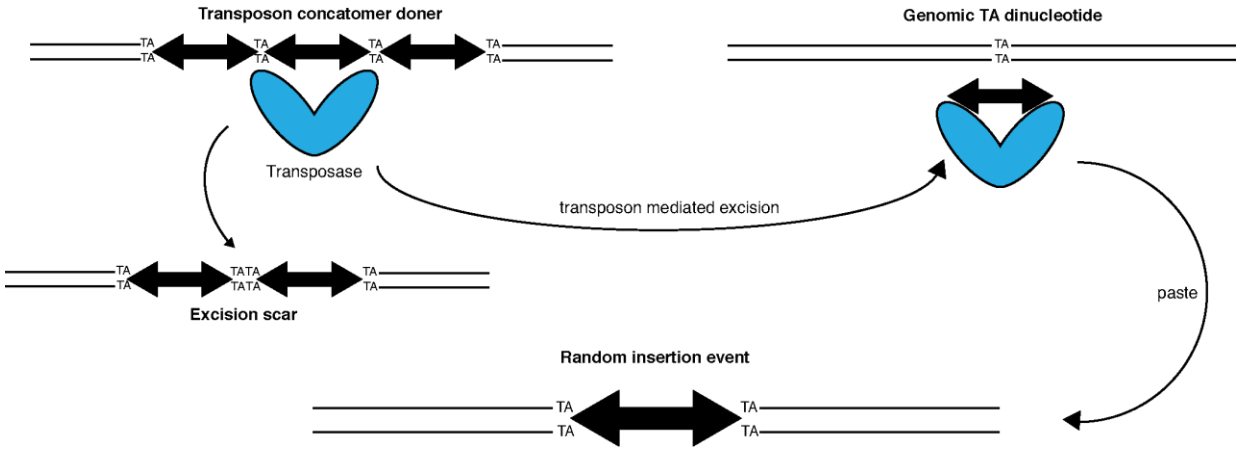



**Figure 1.3 Sleeping Beauty Insertion Event**
The transposase DNA-binding domain recognizes the SB transposon sequence and "cuts" it out, leaving behind a TATA dinucleotide scar sequence. This excised transposon binds to the transposase and is then randomly reinserted at a TA dinucleotide site in the host genome.

## 1.2.2 Cancer gene discovery using SB mediated insertional mutagenesis

If inserted into a gene, the transposon can either upregulate or truncate gene expression depending on the orientation and location of the insertion relative to the gene promoter (Figure 1.4b). If any insertions increase proliferative potential, it can help trigger cell transformation. There are many different mouse models for cancer based on SB-mediated mutagenesis. By spatially and temporally

controlling expression of the transposase, a variety of cancers can be faithfully modeled. In the original SB model, transposase was knocked into the ubiquitously expressed *Rosa26* locus generating a variety of cancers including T-cell and B-cell lymphoma, intestinal neoplasia, and medulloblastoma. Most commonly observed in the T-cell tumors were alterations in the Notch signalling pathway with *Notch1* and *Rasgrp1* insertions in more than half of the tumors[90]. More recently, in an osteosarcoma SB model, expression of transposase was restricted to the osteoblast progenitors through the use of a an *Osterix* promoter. Control mice were already predisposed to osteosarcoma through conditional expression of $Trp53^{R270H}$, but with the help of SB, mice developed tumors with a much shorter latency. Interestingly, these mice also had a lower number of structural events since SB mutagenesis acted as substitute for alterations otherwise caused by the *Tp53* dysfunction. The most commonly inserted genes in this screen were *Nf2* (26%), *Pten* (24%) and *Nf1* (19%). *Pten* was subsequently shown to corporate with *Tp53* to promote tumorigeneses in both in-vitro and in-vivo mouse models[91].

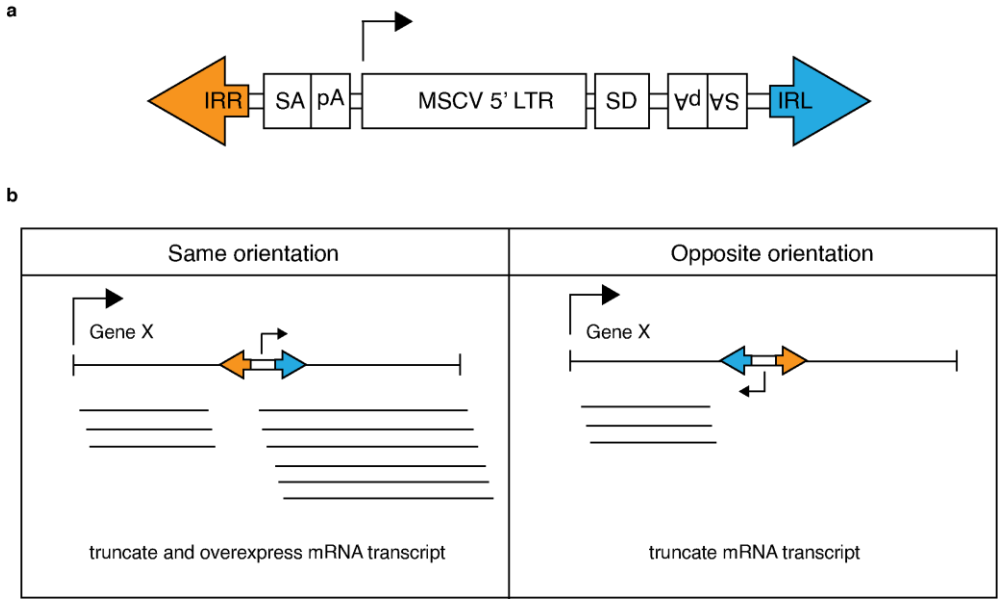



**Figure 1.4 Sleeping Beauty transposon model and mechanism**
**(a)** Schematic of the T2onc2 transposon with MSCV promoter transcriptional start site indicated. **(b)** Transposon effects on gene transcription in each possible orientation. mRNA are indicated as lines under the gene body. IRR - Inverted repeat right, SA - splice acceptor, pA - polyadenylation signal, SD - splice doner, IRL - inverted repeat left.

### 1.2.3 Statistical methods for analyzing insertions

As random insertions accumulate in cells expressing transposase, insertions that increase proliferation and survival of cell are selected for and become part of the dominant clone in the expanding cancerous mass. These insertions can be amplified using a modified splinkerette PCR method (for shear-SPINK method see 3.3.3) and sequenced. To increase efficiency and number of transpositions, T2onc2 exists as a concatemer (usually >40 copies) all of which are capable of integrating into genes. On the scale of individual tumors, this translates into thousands of detected insertion sites. It is therefore necessary to employ statistical models to find genes under positive selection. There are three main methods commonly used to call significant common insertion sites (CIS) and infer drivers in a cohort of tumors: (1) Gaussin Kernal Convolution[92], (2) Monte Carlo Simulation[93], and (3) Gene centric common insertion sites[94].

#### *1.2.3.1 Gaussian Kernel Convolution (GKC)*

A gaussian kernel (i.e smooth normal distribution peak) is placed on each insertion detected along the genome. To get an estimate of the insertion density, overlapping peaks are summed up with each other. The kernel width parameter smooths the area around insertions to infer information from their neighborhood. By altering the kernel width, CISs of different sizes can be detected. A background or null distribution is generated by permuting insertions randomly throughout the genome in TA locations and calculating the summed gaussian kernel density for each iteration. By comparing observed peaks to the expected density distribution at each position, significant peaks can be called. This method is gene naïve, only calling significant regions at different kernel scale factors. It is then up to the investigator to inspect peaks and decide the functional consequence of an insertion cluster.

#### *1.2.3.2 Monte Carlo (MC) Bootstrapping*

Unlike gaussian kernel methodology, the MC method breaks up the genome into equal size windows. Over thousands of iterations, it randomly permutes insertions across all TA locations and takes a count for each window. Statistical tests are done on a per-window basis, comparing the observed insertion count to the permuted count distribution in order to find significant regions. This process is repeated for multiple window lengths and offsets to ensure that all important cancer driver regions are found. Like the GKC method, it is relatively unbiased with respect to identification of CISs since it does not consider functional elements in the genome. Thus MC bootstrapping is suited for detection of poorly characterized genes or regulatory regions. Unfortunately, both GKC and MC can output significant regions that are either too big or too small making many CISs difficult to characterize.

#### *1.2.3.3 Gene centric common insertion (gCIS)*

The gCIS method sets windows for each annotated gene rather than across the entire genome. For each mouse it calculates the expected number of insertions taking into account the number of TA dinucleotides in each gene and the number of insertions detected in each tumor. It uses the Chi Squared test to compare observed and expected counts across the entire tumor cohort (refer to methods 3.3.6 for detailed model). As this method only looks at gene windows, it greatly decreases the number of tests (therefore increasing power) after multiple test correction compared to GKC and MC. It also focuses on genes, making the results easier to interpret and functionally validate. Lastly, unlike the prior methods it controls for the number of insertions detected in each tumor ensuring that each sample is equally represented. The gCIS method has great utility, as it can detect the same genes as GKC and MC, and many more due to its higher power.

#### *1.2.3.4 Sources of Bias*

There are multiple sources of bias that must be considered when running CIS analysis and interpreting the results. Firstly, SB transposons leaving the donor location are more likely to re-insert in a nearby location on the same chromosome. This is contrary to the statistical assumption used in CIS methodology that all TA dinucleotide locations have equal probably of insertion. Therefore, all insertions from the donor chromosomes are filtered before running the CIS analysis. It is important to have multiple mouse lines with different donor chromosomes to ensure that important drivers are not missed. Secondly, the number of reads detected for each insertion are not always correlated with the relative number of insertions in the tumor samples (i.e. clonality) due to sequence-dependence efficiency of amplification[94]. A better estimate of clonality uses the number of unique fragments generated by sonication in the shear-SPLINK protocol (for shear-SPINK method see 3.3.3) since the searing process is not sequence dependant[95]. Lastly, there are a number of false positive genes usually present after running any of the statistical models which need to be manually curated[96]. Most common are the genes *En2* and *Foxf2* which are themselves used in the construction of the SB transposon and therefore are overrepresented in the genome. Likewise, the false positive *Sfi1* has many more copies in the genome than is known in mm9 mouse assembly annotation.

## 1.3 THESIS OVERVIEW

Medulloblastoma is the most common malignant pediatric brain cancer and a significant cause of cancer related mortality in children[2]. It initiates within the cerebellum and in 30% of cases presents with dissemination throughout the brain and spinal cord[4]. Based on various expression studies and a multi-institutional consensus, medulloblastoma has been shown to have at least 4 distinct subgroups (SHH, WNT, Group 3, Group 4)[39]. These subgroups are spatially and temporally stable and have significant prognostic utility in stratifying patients[97,6]. Each of these subgroups has been shown to have unique copy number alterations, methylation and expression profiles which suggest a different cell of origin[12]. Despite numerous genomic studies there is still a lapse of knowledge in the specific primary and metastatic genetic drivers in SHH medulloblastoma (Shh-MB) and its subtypes. New sequencing modalities and mouse models will help decipher the full driver landscape in Shh-MB.

### 1.3.1 Hypothesis

Novel human and mouse datasets will uncover primary and metastatic driver genes acting upon various stages of sonic hedgehog medulloblastoma progression.

### 1.3.2 Study aims

1. Profile and analyze a large set of primary Shh-MB RNAseq libraries, while leveraging complimentary datasets, to identify copy number aberrant, mutated, and fusion driver genes (CHAPTER 2)
2. Use the SB Shh-MB mouse model to identify metastatic drivers under the influence of convergent evolution (CHAPTER 3)
3. Identify genes responsible for Shh-MB primary tumor maintenance using a hybrid SB and PiggyBac transposition system (APPENDIX**)**

# CHAPTER 2
## The Transcriptional Landscape of Sonic Hedgehog Medulloblastoma

Patryk Skowron*, Hamza Farooq*, Florence M.G. Cavalli*, A. Sorana Morrissy*, Michelle Ly, Liam D. Hendrikse, Evan Y. Wang, Haig Djambazian, Helen Zhu, Karen L. Mungall, Quang M. Trinh, Tina Zheng, Shizhong Dai, Ana G. Stucklin, Maria C. Vladoiu, Vernon Fong, Borja L. Holgado, Carolina Nor, Xiaochong Wu, Diala Abd-Rabbo, Yu Chang Wang, Betty Luu, Raul A. Suarez, Avesta Rastan, Aaron H. Gillmor, John J.Y. Lee, Xiao Yun Zhang, Craig Daniels, Peter Dirks, David Malkin, Eric Bouffet, Pierre Bérubé, Uri Tabori, James Loukides, François Doz, Franck Bourdeaut, Olivier Delattre, Julien Masliah-Planchon, Olivier Ayrault, Seung-Ki Kim, David Meyronet, Wieslawa A. Grajkowska, Carlos G. Carlotti, Carmen de Torres, Jaume Mora, Charles G. Eberhart, Erwin G. Van Meir, Toshihiro Kumabe, Pim J. French, Johan M. Kros, Nada Jabado, Boleslaw Lach, Ian F. Pollack, Ronald L. Hamilton, Amulya A. Nageswara Rao, Caterina Giannini, James M. Olson, Bognár László, Almos Klekner, Karel Zitterbart, Joanna J. Phillips, Reid C. Thompson, Michael K. Cooper, Joshua B. Rubin, Linda M. Liau, Miklós Garami, Peter Hauser, Kay Ka Wai Li, Ho-Keung Ng, Wai Sang Poon, G. Yancey Gillespie, Jennifer A. Chan, Shin Jung, Roger E. McLendon, Eric M. Thompson, David Zagzag, Rajeev Vibhakar, Young Shin Ra, Maria Luisa Garre, Ulrich Schüller, Tomoko Shofuda, Claudia C. Faria, Enrique López-Aguilar, Gelareh Zadeh, Chi-Chung Hui, Vijay Ramaswamy, Swneke D. Bailey, Steven J.M. Jones, Andrew J. Mungall, Richard A. Moore, John Calarco, Lincoln Stein, Gary D. Bader, Jüri Reimand, Jiannis Ragoussis, William A. Weiss, Marco A. Marra, Hiromichi Suzuki, Michael D. Taylor

### 2.1.1 Abstract

Sonic hedgehog medulloblastoma encompasses a clinically and molecularly diverse group of cancers of the developing central nervous system. Unbiased sequencing of the transcriptome across a large cohort of 250 tumors reveals differences among molecular subtypes of the disease, demonstrating the previously unappreciated importance of non-coding RNA transcripts. We identify a number of novel genes with mutations (*MYCN, GNAS*, *IKBKAP*, and *KDM6A*), somatic copy number aberrations, and gene fusions. Furthermore, we show that many fusions arise secondary to rearrangement of the genome (*PTCH1*, *NCOR1),* while others through trans-splicing (*RALGAPA2,* and *GNAS*). Molecular convergence on a core of specific genes by nucleotide variants, copy number aberrations, and gene fusions highlights key roles of specific pathways in the pathogenesis of Sonic hedgehog medulloblastoma.

## 2.1.2 Introduction

Medulloblastoma (MB) is the most common malignant pediatric brain tumor, and a major cause of morbidity and mortality in the pediatric population[98]. Current therapy consists of maximal safe resection, radiotherapy in patients over 36 months, and cytotoxic chemotherapy. MB is thought to comprise a group of four molecularly distinct diseases: Wnt, Sonic Hedgehog (Shh), Group 3, and Group 4[12]. Shh-MB is clinically heterogeneous with infants, teenagers and adults affected. Shh-MB likely comprises four molecular subtypes, Shh-α (adolescents), Shh-β (babies with a poor prognosis), Shh-γ (babies with a good prognosis), and Shh-δ (adults)[99]. The vast difference in the host (babies versus adolescents versus adults) dictates different treatment approaches for different molecular subtypes. Prior delineation of Shh-MB subtypes used expression microarrays[100], and/or DNA methylation arrays[99], and the biology underlying the differences among the subtypes is poorly understood.

To further understand the biology of Shh-MB and its molecular subtypes, we studied 250 human Shh-MB using strand-specific RNA sequencing. This non-biased approach to the Shh-MB transcriptome allows us to understand the transcriptional basis and underlying biology of Shh-MB, and reveals a previously unsuspected role for many non-coding RNAs. It also identifies a number of novel fusion transcripts in Shh-MB, some of which are due to structural events in the genome, while others appear to arise secondary to trans-splicing events. This analysis of a large cohort of similar tumors highlights previously unsuspected examples of molecular convergence where the same gene or pathway is activated through diverse molecular mechanisms, emphasizing the importance of those drivers in Shh-MB. Genetic events in Shh-MB do not assort randomly across the cohort, but rather show very restricted patterns of mutual exclusivity, suggesting a specific biology, with implications for Shh-MB modeling, and perhaps for the design of synthetic lethal

approaches to therapy. This very large cohort allows unprecedented insights into the transcriptome of this disease.

## 2.2 RESULTS

### 2.2.1 Importance of the non-coding transcriptome in Shh-MB.

Our Shh-MB strand-specific RNA-seq samples (n = 250) were additionally characterized with whole genome sequencing (WGS) (n = 26), 450K methylation arrays (n = 196), Affymetrix HuGene 1.1 expression arrays (n = 173), and Affymetrix SNP 6.0 arrays (n = 130) (Figure 2.1a). Integrative analysis and unsupervised clustering of both RNA-seq and 450K methylation data allowed us to assign Shh-MB samples to their appropriate molecular subtype[99]. Subtype assignment based on RNA-seq and 450K methylation data overlaps highly with subtyping using Affymetrix expression and 450K methylation arrays (Figure 2.1b, c). While protein coding genes make up only 35% of the transcriptome in GENCODE (v19), 95% of subtype-specific genes identified using expression arrays are protein coding genes (Figure 2.1d). However, Shh-MB subtype-specific transcripts identified with RNA-seq encompass many non-coding RNA species, including long non-coding RNAs, expressed pseudogenes, and microRNAs (Figure 2.1d). Indeed, the majority of the genes differentially expressed between subtypes using RNA-seq data are non-coding transcripts, which are not evaluated by expression arrays (Figure 2.1e). While many of these non-protein coding genes are poorly annotated, pathways analysis reveals divergent biological mechanisms among Shh-MB subtypes (Figure 2.1f). We conclude that non-protein coding genes likely play a hitherto unexpected and important role in the biology of Shh-MB.

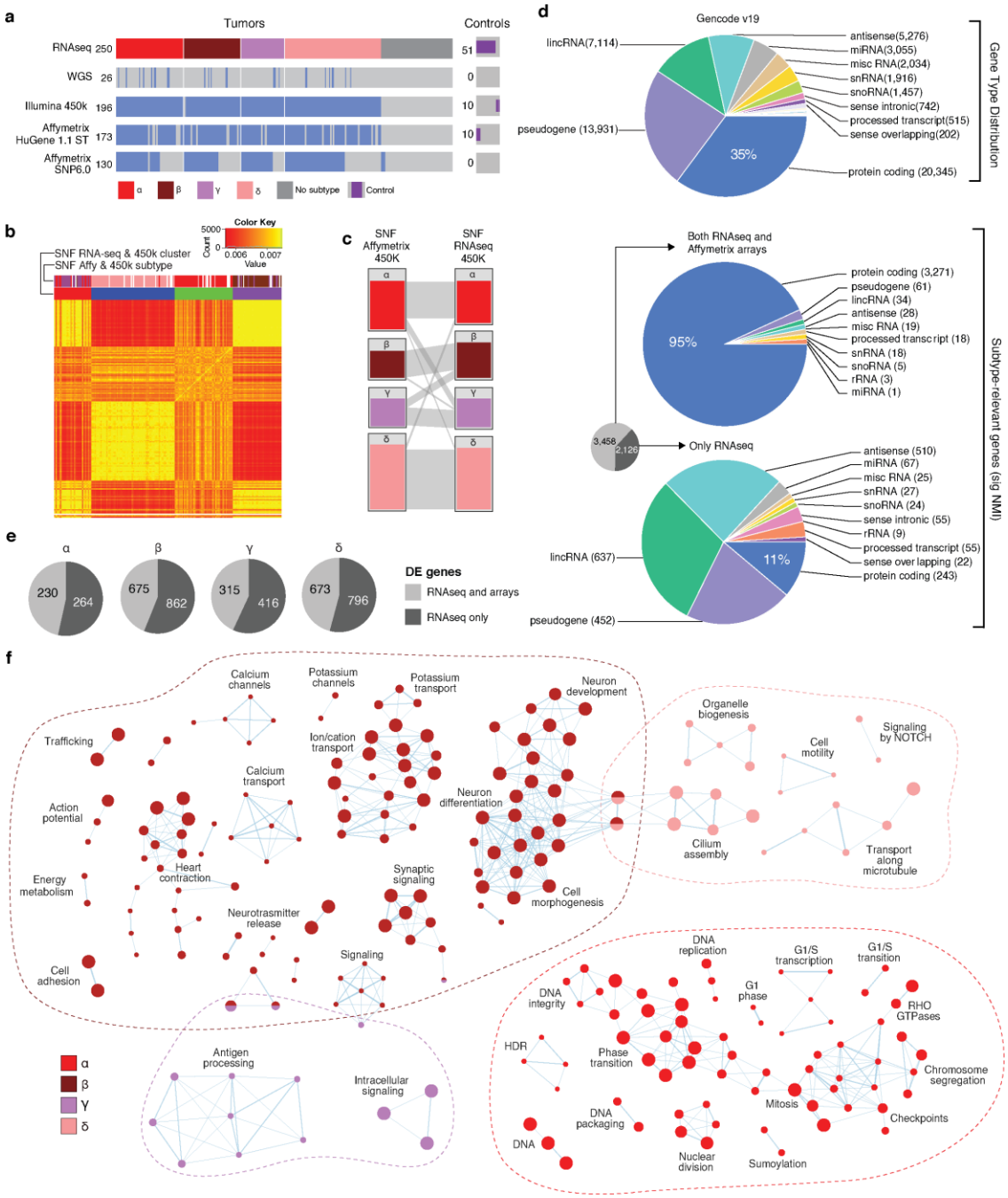



**Figure 2.1 Importance of the non-coding transcriptome in Shh-MB**

**(a)** Overview of Shh-MB RNA-seq samples and overlapping data sources. **(b)** Heatmap of sample-to-sample fused network by cluster (k = 4, n = 250). Sample similarity is represented by red (less similar) to yellow (more similar) coloring inside the heatmap. **(c)** Subtype clusters obtained by SNF (k = 4) using Affymetrix + 450K methylation and RNA-seq + 450K methylation (n = 196). Relationships between clustering methods are indicated by gray bars between columns. **(d)** Biotype distribution amongst all genes (top) as compared to genes that differentiate subtypes (significant NMI from SNF RNA-seq + 450K methylation), in both RNA-seq and microarray datasets (middle) or restricted to only the RNA-seq dataset (bottom). **(e)** Differentially expressed genes per subtype (RNA-seq). Genes found only with RNA-seq data are indicated. **(f)** Enrichment map of biological processes and pathways in Shh-MB subtypes. Each node represents a pathway or process and connecting lines represent common genes between them. Nodes with many shared genes are grouped together and labeled with a biological theme. The color of the nodes refers to the subtype(s) in which the process is enriched. The size of the node is proportional to the number of genes in the process.

### 2.2.2 Identification of known and novel indels and single nucleotide variants

We identified the incidence and patterns of mutations in known Shh-MB driver genes (i.e., *PTCH1*, *SUFU*, *DDX3X*, *TP53*) in a subtype-specific manner. Several Shh-MB drivers previously identified as amplified, (i.e., *GLI2*, *MYCN*, and *PPM1D*) also harbor novel damaging mutations in a subset of patients (Figure 2.2; Figure 2.3a–h). As we did not have germline gDNA for all patients, a subset of these mutations could be germline mutations. Many *GLI2* single nucleotide variants (SNVs) are found within the activation domain (p.P1028L, p.H1073Y, p.Q1323H, p.A1514V)[101] (Figure 2.3a) and can disrupt PKA phosphorylation sites (p.A896D)[102]. Other SNVs can disrupt binding to *SUFU* (p.G274R)[103]. *GLI2* SNVs are largely exclusive of *GLI2* amplification or fusions (Figure 2.3b).

We also detect a cluster of SNVs in *MYCN* within the phospho-degron containing MBI domain (p.T43I, p.P44L, p.T58K, p.T58M, p.P59L) (Figure 2.3c, d). *MYCN* Amplifications and SNVs are mutually exclusive (Figure 2.3e). Phosphorylation of MYCN at S62 primes for a second phosphorylation at T58 by glycogen synthase kinase-3 (GSK3). Subsequent dephosphorylation at S62 leads to recruitment of the FBXW7 E3 ubiquitin ligase complex to a phosphodegron motif that includes amino acids both N-terminal and C-terminal to pT5898, and the subsequent ubiquitination of MYCN[104,105]. Mutations in this region of MYCN disrupt FBXW7 binding and/or ubiquitination, and are predicted to stabilize MYCN[106] (Figure 2.3d). Remarkably, we also identify missense mutations of *FBXW7* within tryptophan-aspartic acid motif (WD40)[107–110] that binds MYCN, in >10% of Shh-MB, which are mutually exclusive of *MYCN* amplification or SNVs. Therefore, 17% of Shh-MB patients have a genetic event that directly target the abundance and/or stability of MYCN (Figure 2.3e, f).

*PPM1D*, a negative regulator of the p53 DNA damage response pathway[111] undergoes nonsense and frameshift mutations at its C-terminus (Figure 2.3g, h), all of which are predicted to leave its phosphatase activity intact while significantly increasing protein stability[112–114]. We also identified several novel recurrent mutations in *GNAS*, *IKBKAP*, and *KDM6A* (Figure 2.2). *GNAS*, encoding a heterotrimeric Gs protein α subunit (Gαs), is mutated (4% Shh-MB) between the GTPase and helical domains which is predicted to reduce GTP binding (Figure 2.3i). GNAS activates adenylyl cyclase which increases intracellular cAMP, there-by activating protein kinase (PKA), a negative regulator of Shh signalling[115]. Correspondingly, we also observe mutations mutually exclusive of *GNAS* in *PRKAR1A*, a critical component of the PKA complex (Figure 2.3j). This is in line with the phenotype of *Gnas* knockout mice which develop Shh-MBs[115].

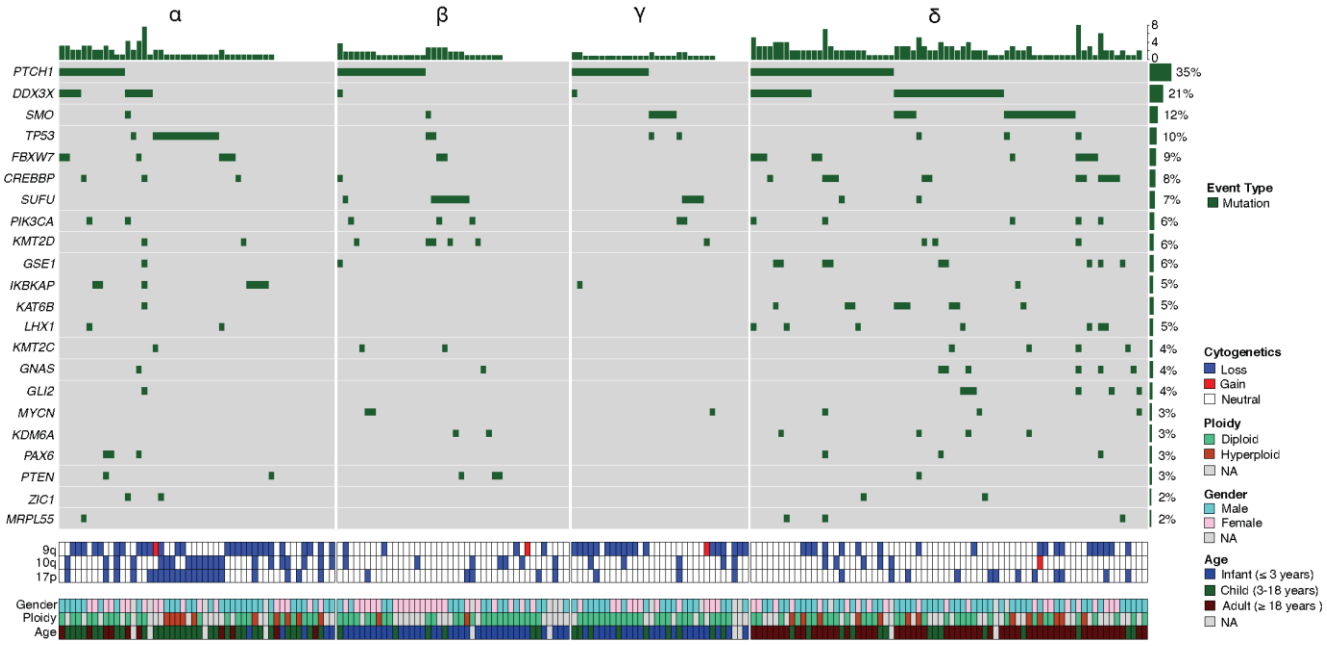



**Figure 2.2 Mutation Landscape**
Oncoprint summary of mutations detected across Shh-MB subtypes (n = 196). Subtypes are denoted above. NA, not available.

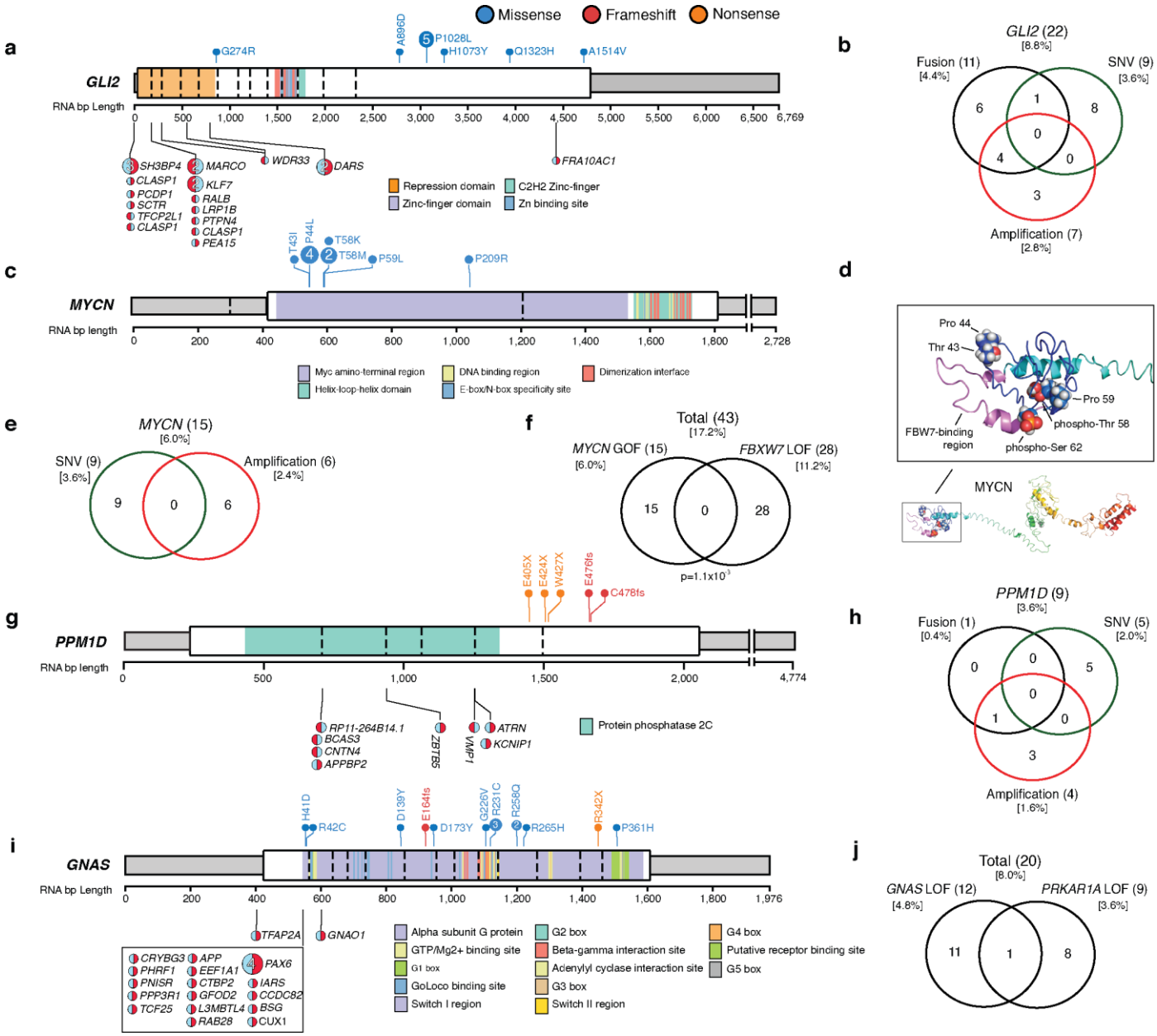



**Figure 2.3 Identification of known and novel indels and single nucleotide variants**

**(a, b)** Gene-level summary of **(b)** *GLI2* events and **(c)** their overlaps. Mutations in **(a)** are shown as lollipop diagrams above the gene schematic and fusion events are shown below. The 5'/3' orientation of the fusion transcript is indicated by the color orientation. In cases where *GLI2* is the 3' partner the fusion lollipop is red on the right. **(c)** Gene-level summary of *MYCN* events. **(d)** Structural model of MYCN highlighting positions affected by hotspot mutations (blue) near the FBWX7 protein binding region (purple), and phospho-degron positions (red).**(e)** Overlap of *MYCN* amplification and SNV events. **(f)** Mutual exclusivity of *MYCN* gain-of-function (GOF) and *FBXW7* loss-of-function (LOF) events. P-value calculated using the DISCOVER package. GOF/LOF events include both high-level CNA and mutation events. **(g, h)** Gene-level summary of **(h)** *PPM1D* events and **(i)** their overlaps. **(i)** Gene-level summary of *GNAS* events. **(j)** Mutual exclusivity of *GNAS* and *PRKAR1A* LOF events. LOF events include mutations and high-level deletions.

### 2.2.3 Somatic copy number aberrations in Shh-MB

Regions of recurrent genomic gain and loss identify both known Shh-MB driver genes (i.e., *MYCN*, *GLI2*, *PPM1D*, *PTEN*)[31], as well as some novel putative drivers (i.e., *PRMT2*, *HECTD1*, *SOX11*, and *LHX1*) (Figure 2.4a). Several recurrent somatic copy number aberrations (CNAs) that do not contain any genes when studied by expression arrays, do contain transcripts when studied by RNA-sequencing (Figure 2.4b). Regions of focal amplification are much more likely to show concomitant changes in gene transcription as compared to larger, broad copy number changes (Figure 2.4c). A number of putative Shh-MB driver genes encompassed by focal gains or deletions demonstrate copy number driven expression, further supporting their role as drivers (Figure 2.4d). Notably, only 15% (378/2,536) of genes identified within GISTIC regions show copy number driven expression (Figure 2.4e, Figure A1a–c). In many cases, the copy number responsive genes are poorly annotated non-coding RNAs that might first be overlooked (Figure 2.4e–h, Figure A1d–f). We also observe significant deletions at 9q34.11 encompassing the copy number responsive gene *GPR107* (Figure 2.4f). This region is usually lost in the context of chromosome 9q loss along with *PTCH1* and *IKBKAP* (Figure A1g, h). A substantial minority (24%) of Shh-MB are aneuploid; their transcriptome differs from diploid tumors by over-expression of genes involved in RNA processing and translation (Figure A2a–d). We conclude that regions of focal CNAs in the Shh-MB genome contain both copy number responsive and non-responsive genes, that many events focus on poorly characterized non-coding transcripts, and that non-copy number responsive genes within CNAs are likely a poor choice for the development of targeted therapy.

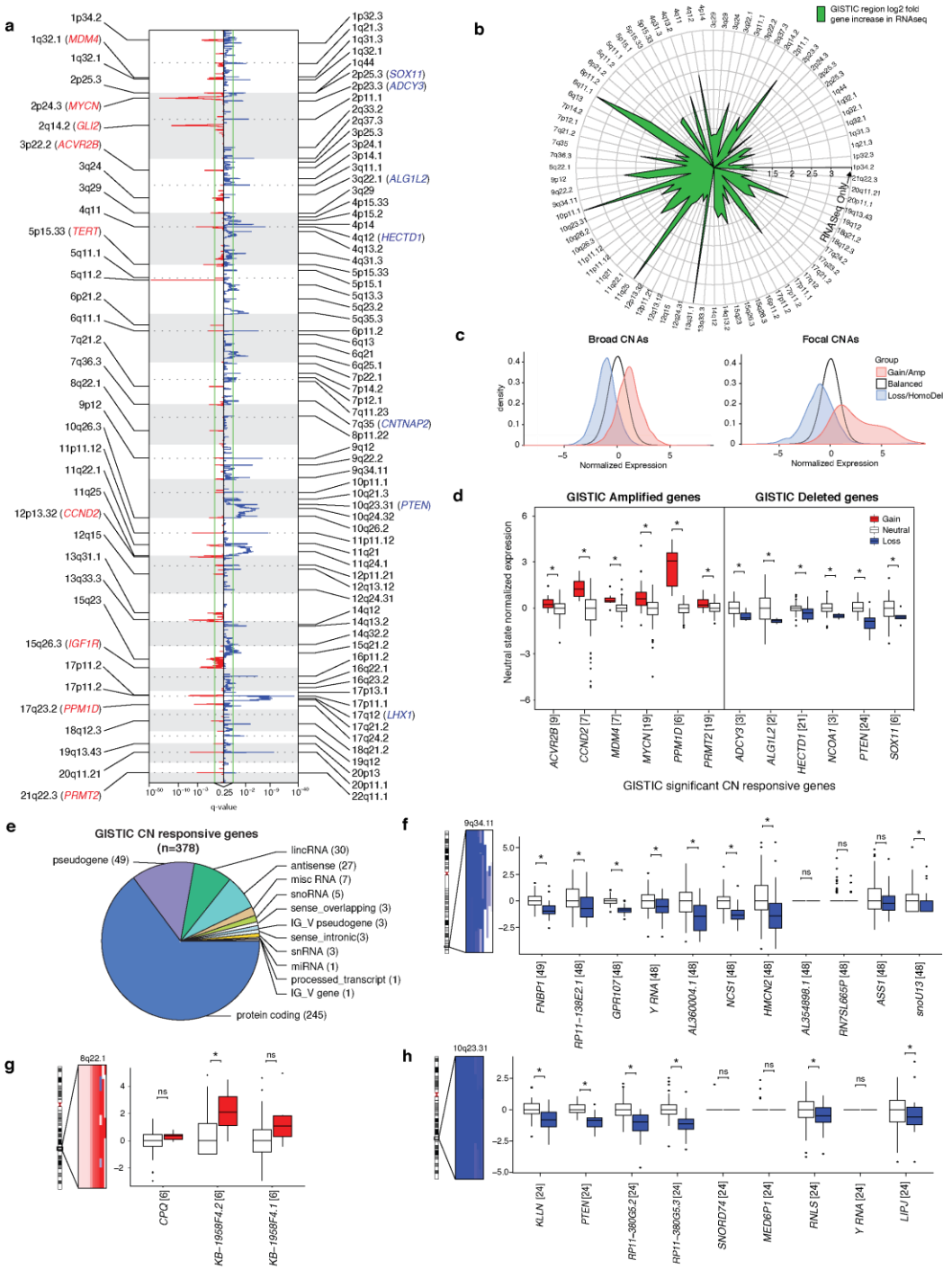



**Figure 2.4 Somatic copy number aberrations in Shh-MB**

**(a)** GISTIC significant amplifications (red) and deletions (blue) observed in Shh-MB (n = 126). **(b)** Log2 fold increase of known annotated gene in GISTIC regions using RNA-seq compared to expression arrays. GISTIC regions with genes only found in the RNAseq dataset have points on the outermost circle. **(c)** Normalized expression density across broad and focal CNAs. **(d)** Expression difference between copy number neutral and aberrant states in GISTIC region copy number responsive genes. Numbers in square brackets denote the number of patients detected with the CNA. **(e)** GISTIC copy number responsive gene types. **(f–h)** Expression difference between copy number neutral and aberrant states in **(f)** 9q34.11, **(g)** 8q22.1, and **(h)** 10q23.31. Asterisks annotates copy number responsive genes (Kruskal-Wallis adjusted p-value < 0.05). The SNP 6.0 copy number segments are shown to the left of each graph. Expression of each gene was normalized by the expression median of the neutral copy number state.

### 2.2.4 Identification of Shh-MB fusion genes

We identified known and novel fusion transcripts in the Shh-MB transcriptome using three distinct assembly and alignment-based callers (STAR-fusion, InFusion, Trans-Abyss)[116–118]. We filtered any readthrough transcripts, or fusion contigs that were also observed in libraries of non-cancerous brain tissue (Figure A3; Figure A4). A subset of Shh-MB patients (12/126, 10%) harbor a high number (top 25$^{th}$ percentile) of both fusions and copy number events, and are significantly associated with both aneuploidy (10/12; $p = 7.4 \times 10^{-7}$, two-sided Fisher's exact test) and *TP53* mutations (6/12; $p = 1.2 \times 10^{-4}$, two-sided Fisher's exact test) (Figure A5a). Only a subset of fusion transcripts demonstrate substantial evidence of an underlying structural variant (SV) in the genome due to the presence of breakpoints in matching WGS or SNP 6.0 data and/or the identification of multiple splice variants of the same fusion transcript. The amount of SV supported fusions per patient was significantly different among subtypes ($p= 4.7x10^{-8}$; Kruskal-Wallis rank sum test), with Shh-α showing the highest number of fusions per tumor.

A large number of SV supported fusions coincide with focal amplifications of *GLI2* (2q14.2), *MYCN* (2p24.3), *CCND2* (12p13.32), and *PPM1D* (17q23.2) (Figure 2.5a,b). Most recurrently, we observe *GLI2* fusion transcripts (11/250 Shh-MB) fused in the 5' end of the mRNA which houses the repressor domain of the encoded protein, suggesting that the fusions could lead to an overactive protein (Figure 2.3a). We additionally observe recurrent fusion transcripts at nearby genomic loci, such as *EPB41L5*, *NBAS*, *BCAS3*, and *GLIS3* which are likely a result of chromothripsis, and/or the formation of extrachromosomal double minutes (Figure 2.5c-f)[20,119]. It is unclear the extent to which amplification versus the formation of a fusion transcript contributes to clonal selection (Figure A5b–g), nor is it obvious whether the fusion transcripts in other nearby

genes are drivers or passengers. Conversely, we now identify novel fusions of *ZBTB20* (14/250 patients), which are not usually found in context of amplification (Figure 2.6a, b).

We also identify novel 'fusion transcripts' of known Shh-MB tumor suppressor genes such as *PTCH1* and *SUFU*, (Figure 2.6c–h), which are accompanied by decreased expression of the gene immediately following the 'breakpoint'. These are likely markers of chromosomal events that result in loss of gene function and are largely mutually exclusive of tumors with mutations or large chromosomal deletions, supporting their functional role (Figure 2.6g, h). *NCOR1*, a transcriptional regulator of neural stem cell differentiation[120,121] harbors similar loss-of-function (LOF) fusion transcripts and damaging mutations (13/250) (Figure 2.6i, j). We conclude that >20% of Shh-MB patients exhibit fusion transcripts with structural support for an event in the genome.

### 2.2.5 Promiscuous recurrent chimeric transcripts in Shh-MB

While some chimeric fusion transcripts exhibit strong evidence for a genomic rearrangement (i.e., *GLI2*, *PTCH1*), others (i.e., *RALGAPA2*, *GNAS*, *NOC4L*, *CHMP1A*) showed no evidence of a genomic DNA rearrangement (Figure A4). However, these latter fusions transcripts are likely *bona fide*, as they can be validated by RT-PCR and Sanger sequencing (Figure 2.7a, b; Figure A6a, b). Both *RALGAPA2* and *GNAS* were almost always the 3' partner of the fusion, and have a single exonic breakpoint position (Figure 2.7a, Figure A6a). After closer examination of the fusion junction, it became evident that most patients harbored chimeric reads at the fusion junction in *RALGAPA2* (200/250 Shh-MB) and *GNAS* (245/250 Shh-MB), most of which were missed by the fusion callers. Curiously, both genes are transcriptionally fused with a large number of different partner genes (Figure 2.7a–d, Figure A6a–d). Within a given Shh-MB tumor, several different 5' fusion partner genes can be identified (range = [0–34], median = 4).

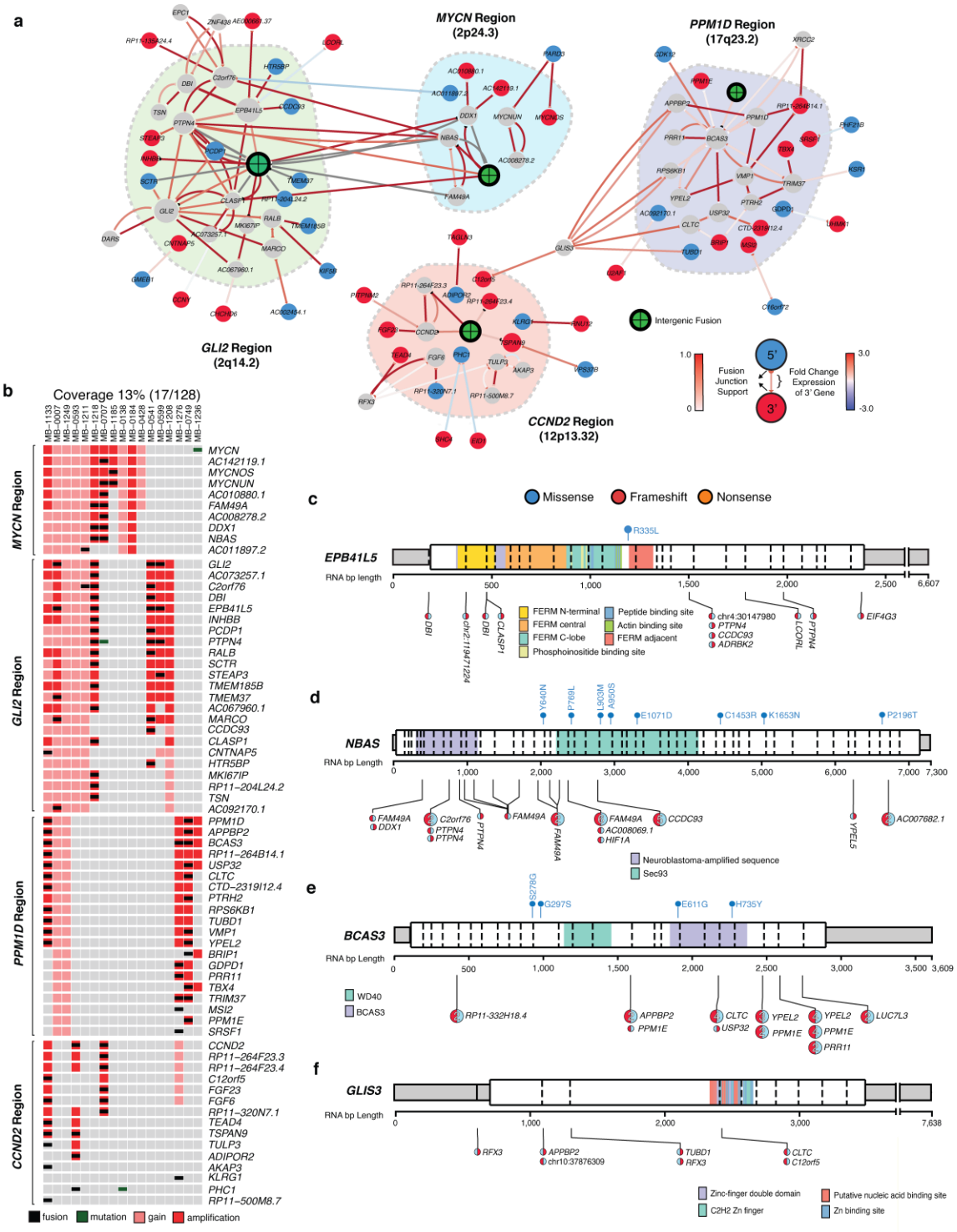



**Figure 2.5 Identification of Shh-MB fusion genes**

**(a)** Network of gene fusions in focally amplified regions. Node color signifies the most common orientation of the fusion gene, 5' (blue), 3' (red), or both (gray). The arrow and base color show the proportion of chimeric reads compared to wildtype supporting the fusion. The arrow line color shows the difference in expression of the 3' fusion partner compared to patients without the detected fusion. **(b)** Oncoprint of fusions depicted in focally amplified regions illustrated in (a). **(c–f)** Gene-level summary of **(c)** *EPB41L5*, **(d)** *NBAS*, **(e)** *BCAS3*, and **(f)** *GLIS3* events. Refer to Fig. 2b for schema description.

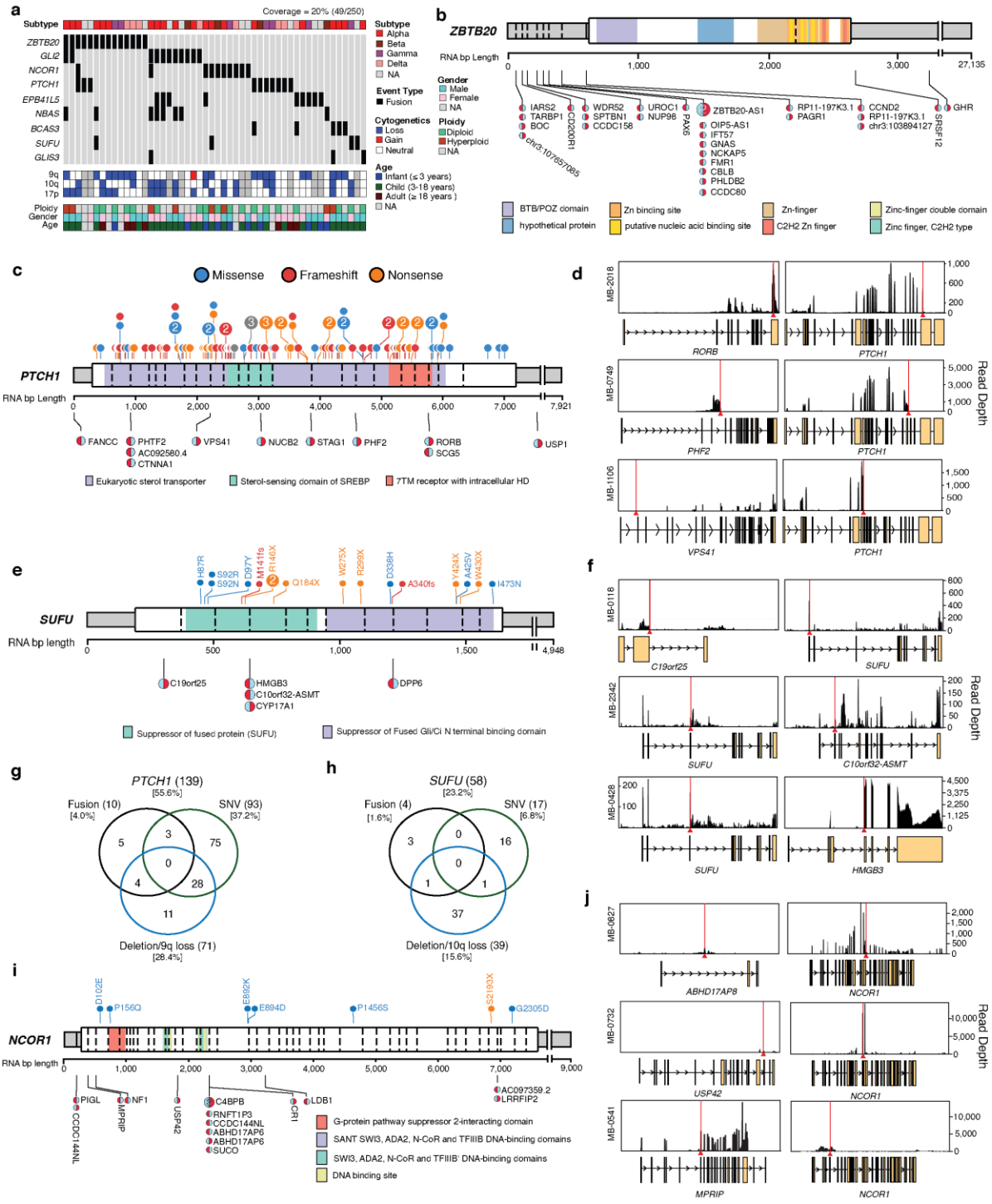



**Figure 2.6 Novel recurrent fusions in Shh-MB**

**(a)** Oncoprint of fusions detected in focally amplified regions and known Shh-MB tumor suppressors. **(b)** Gene-level summary of *ZBTB20* events. Mutations are shown as lollipop diagrams above the gene schematic and fusion events are shown below. The 5'/3' orientation of the fusion transcript is indicated by the color orientation. In cases where *ZBTB20* is the 3' partner the fusion lollipop is red on the right. **(c)** Gene-level summary of *PTCH1* events. **(d)** Read depth diagrams of representative *PTCH1* fusion events. **(e)** Gene-level summary *SUFU* events. **(f)** Read depth diagrams of representative *SUFU* fusion events. **(g)** Overlap of *PTCH1* fusion, amplification and mutation events. **(h)** Overlap of *SUFU* fusion, amplification and mutation events. **(i)** Gene-level summary of *NCOR1* events. **(j)** Read depth diagrams of representative *NCOR1* fusion events.

The vast majority of chimeric transcripts in *RALGAPA2* have their breakpoint in exon 37, and indeed most Shh-MB tumors exhibit some chimeric reads at this locus, fused to a large number of different 5' partner genes (Figure 2.7c). A subset of the fusion partner genes is recurrent (i.e., *CRINKL1, ZDHHC8, DYNC1L12, UBXN4*), while others are limited to a single sample (Figure 2.7c, d). While the majority of the chimeric reads map to exons, nearly all chimeric junctions exhibit a strong U12 splicing signal immediately preceding the 5' partner breakpoint (Figure 2.7e).

As chimeric transcripts can be artifacts of template switching by reverse transcriptase during cDNA preparation[122], recurrent chimeric fusions were validated using the same (Figure 2.7f), or a different (Figure 2.7i) reverse transcriptase, followed by PCR amplification and Sanger sequencing across a panel of Shh-MB tumors and controls. Chimeric *RALGAPA2* fusions were found in a subset of Shh-MBs, but not in control Group 3-MB, Group 4-MB, or normal cerebella (Figure 2.7f). PacBio Iso-Seq long reads further validate full-length chimeric *RALGAPA2* transcripts with high confidence (Figure A7). These fusions are not purely a result of high expression since both *RALGAPA2* and *GNAS* were not the most highly expressed genes in Shh-MB (Figure 2.7j). Similar chimeric transcripts were seen in *GNAS* (Figure A6a–d), with a strong U2 splicing signal (Figure A6e), and long read sequencing validations (Figure A7). *GNAS* fusions were not completely restricted to Shh-MB since they were also detected in Group 3 MB (*PAX6-GNAS*) and normal cerebellar controls (*FGFR1-GNAS*). (Figure A6f). Shh-MBs have additional genes exhibiting chimeric transcripts, but without good evidence of a structural event in the genome, often with a variety of fusion partners (Figure A4). A recent PAN-CAN report suggests that up to 18% of fusion transcripts in cancer are generated through trans-splicing[123]. As chimeric fusion transcripts of *RALGAPA2* and *GNAS* are seen in multiple patients, lack support for a

structural event in the genome, are fused to numerous partner genes (even within a single tumor), and have a splicing signal, we hypothesize that these chimeric transcripts have arisen through trans-splicing. We conclude that these chimeric Shh-MB transcripts are *bona fide*, although their biological importance, let alone their role as Shh-MB drivers will require additional functional validation, likely *in vivo*.

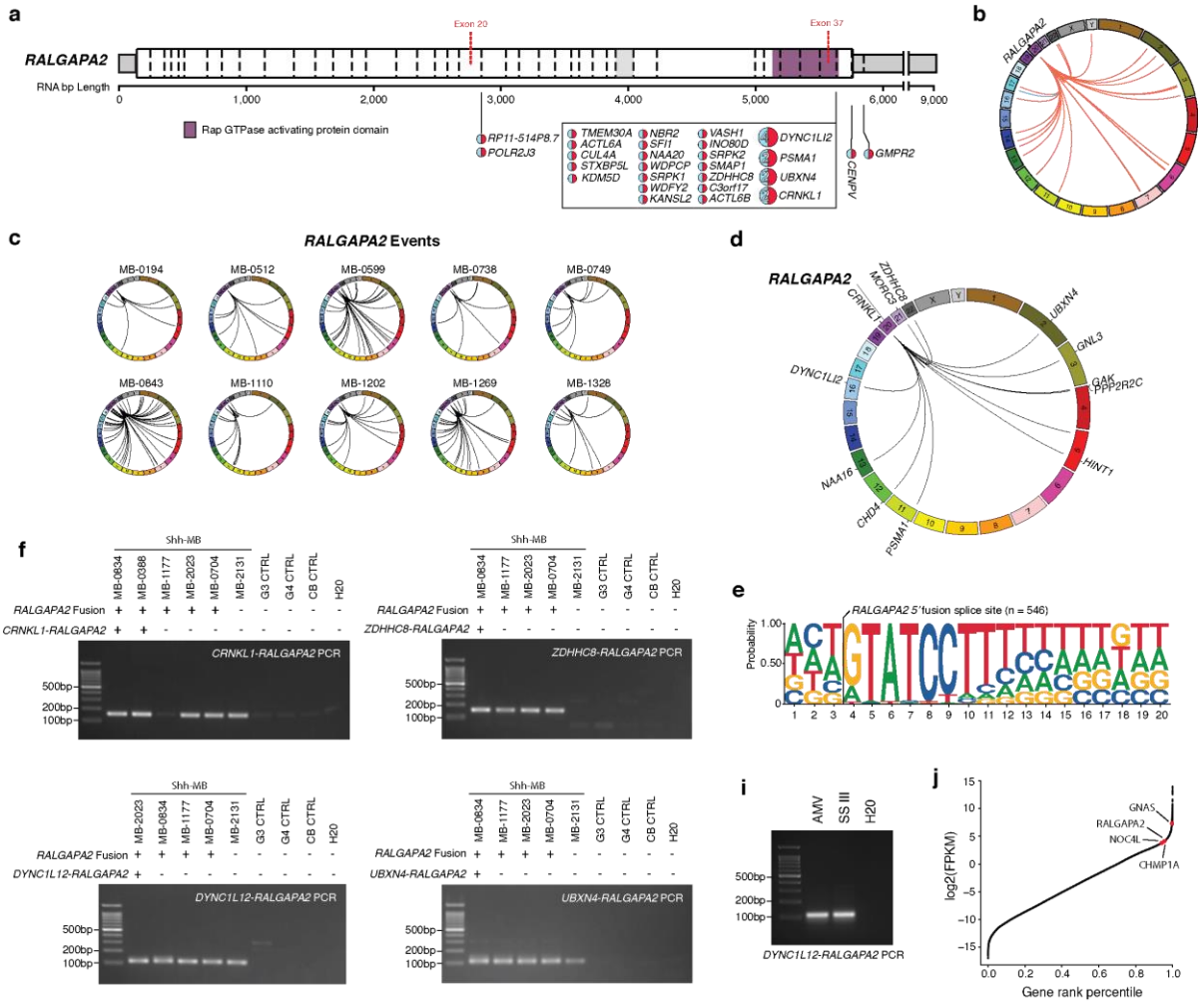



**Figure 2.7 Promiscuous recurrent *RALGAPA2* chimeric transcript breakpoints**
**(a, b)** Gene-level summary of **(a)** *RALGAPA2* fusions detected by fusion-callers, and **(b)** their distribution across the genome. Refer to Fig. 2b for schema description. **(c, d)** Distribution of *RALGAPA2* exon 37 chimeric junction spanning reads across the genome with **(d)** genes found in >10 samples indicated. Chimeric reads were extracted from STAR alignments. **(e)** Splice site consensus sequence of *RALGAPA2* 5' chimeric fusion partner transcripts (n = 546). **(f)** PCR validation of *RALGAPA2* fusion RNA transcripts in human Shh-MB samples with and without detected fusions (by RNA-seq) compared to Group 3-MB, Group 4-MB and normal cerebellar controls. Patients with any detected chimeric transcripts at exon 37 in *RALGAPA2* (by RNA-seq) are indicated as *RALGAPA2* fusion positive (+). **(i)** Validation of *DYNC1L12-RALGAPA2* using SuperScript III transcriptase (SSIII), and avian myeloblastosis virus reverse transcriptase (AMV). **(j)** FPKM Expression ranking of genes with recurrent chimeric transcripts in Shh-MB.

### 2.2.6 Landscape of oncogenic alterations across Shh-MB

Transcriptional profiling of this large cohort of a single molecular tumor type permits identification of both known and novel Shh-MB driver genes, and their patterns of mutual exclusivity. Most Shh-MBs (86%) have an identifiable event activating the Sonic Hedgehog signaling pathway, including mutations of *PTCH1* (42%), *SMO* (12%), *SUFU* (10%), or *GLI2* (9%) (Figure 2.8). About 11% patients have previously unappreciated inactivating (i.e., *SUFU* or *PTCH1*), or activating (i.e., *GLI2*) fusion transcripts affecting Shh pathway genes. Pathways discovered using copy number aberrations, mutations, or fusion transcripts were numerous in Shh-α and Shh-δ, but limited for Shh-β or Shh-γ due to their low number of mutational events (Figure A8). There is strong mutational convergence on genes important for Shh signaling, neuronal development, cell cycle progression, and modification of the epigenome (Figure 2.9).

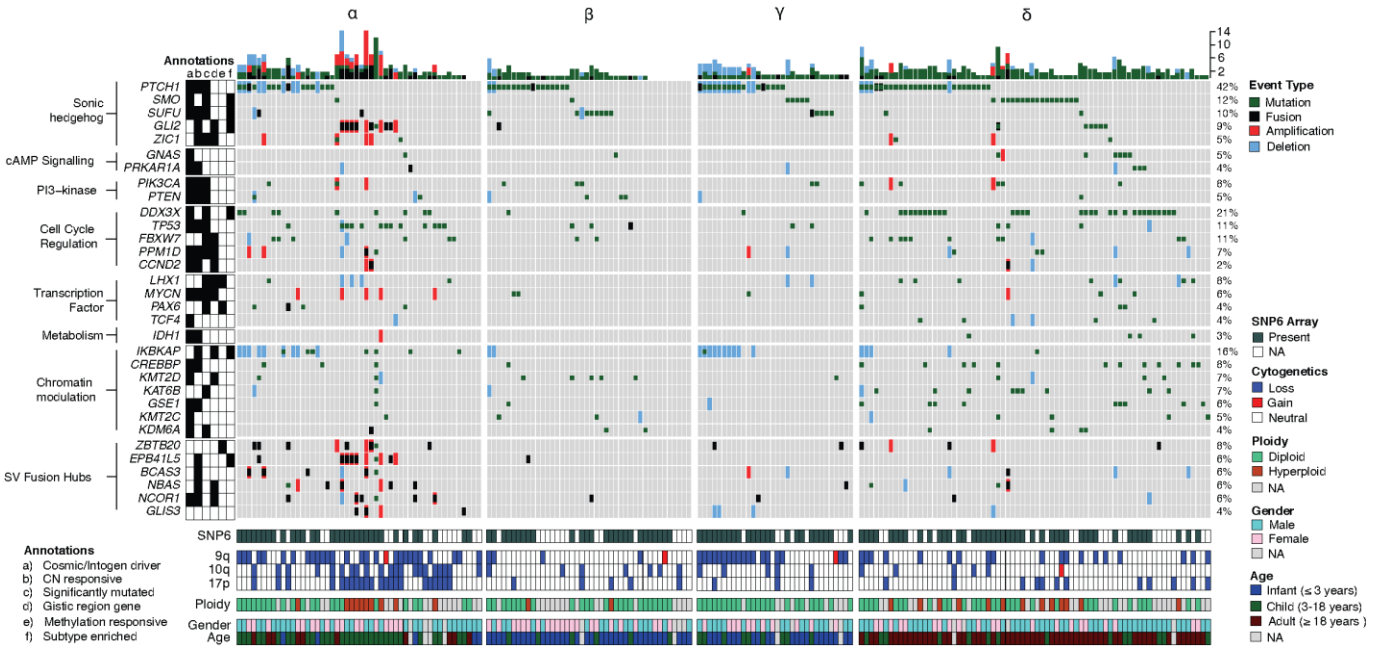



**Figure 2.8 Landscape of oncogenic alterations across Shh-MB**
Oncoprint summaries of all fusion, mutation and copy number data converging on known and novel pathways (n = 196). Subtypes are denoted above. NA, not available.

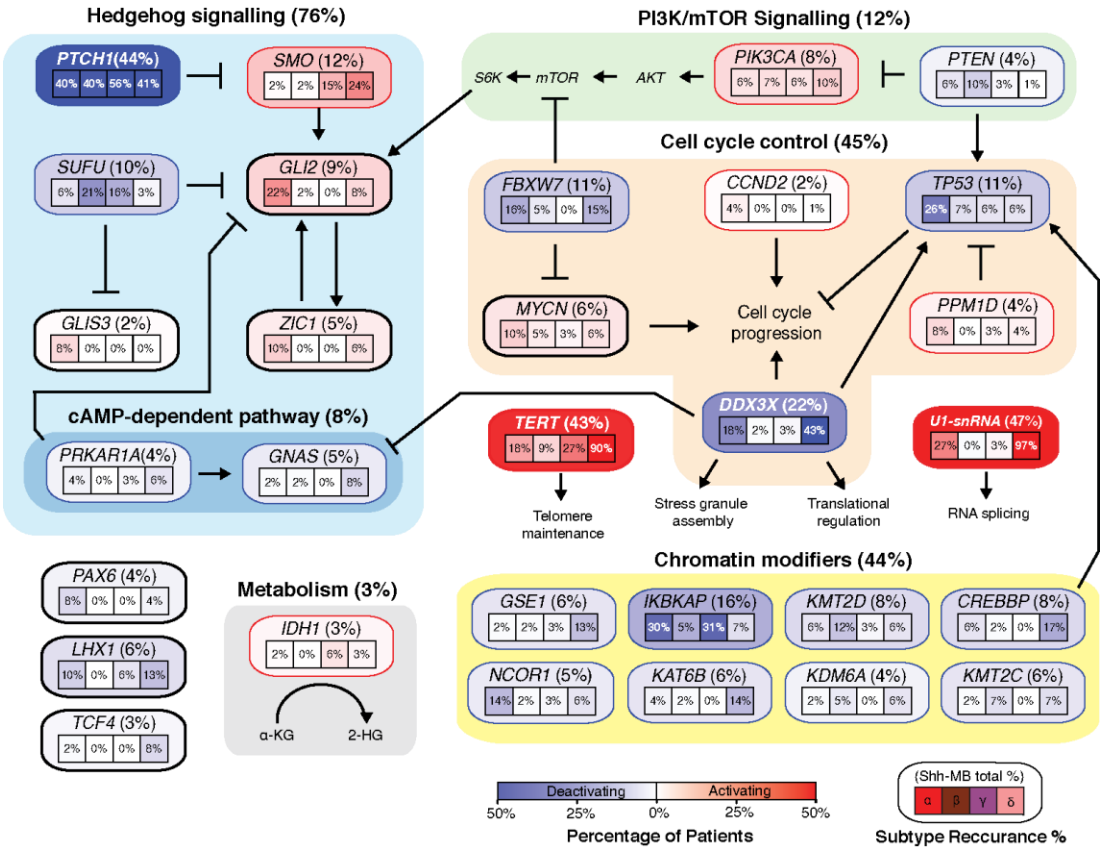



**Figure 2.9 Shh-MB oncogenic pathways**
Percentage of altered genes and pathways integrating mutation, high-level copy number and fusion data. Alteration frequencies are expressed as percentages of all cases per subtype (n = 196) in the boxes and total percentage across Shh-MB (n = 250) in parenthesis beside each gene name. Red indicates activating alterations while blue indicates inactivating alterations. *TERT* and *U1-snRNA* alternation percentages obtained from earlier published studies[99,100].

Of Shh-MBs without detected events that canonically lead to excess Shh signaling (*PTCH1*, *SMO*, *SUFU*, *TP53*, *GLI2*, 9q, 10q, and 17p loss) (45/250), the most recurrent mutational events involved *DDX3X* (n = 12), *KMT2D* (n = 6), *PRKAR1A*, *GNAS*, *GSE1* and *CREBBP* (each n = 5) (Figure 2.8); all of which have been previously shown to interact with or potentiate Shh signaling[115,124,125].

We used MethylMix[126] to identify potential Shh-MB driver genes affected by promoter CpG hypo- or hypermethylation, for which there is a correlative change in gene expression (Figure A9). This approach identifies a number of known cancer genes (i.e., *FOXL2*, *RUNX1T1*), transcription factors (i.e., *MEIS2*), as well as *LHX1* and *PAX6 (*which are also recurrently affected by mutations).

Transcriptional silencing of *PAX6* through promoter CpG methylation, versus somatic mutations of *PAX6* appear to be largely mutually exclusive ($p = 7.3 \times 10^{-4}$, multinomial exact test), suggesting convergence on PAX6 loss of function (Figure A9c–h). We observe significant mutual exclusivity of genetic events affecting genes in the Shh signaling, PI3-Kinase, cell cycle and chromatin modifier pathways (i.e., *MYCN*;*FBXW7*, *PTCH1;SUFU, SUFU;SMO*). Chromosomal deletions of 9q, 10q, and 17p are mutually exclusive with each other, as well as focal events affecting genes in the Shh signaling pathway. We conclude that Shh-MB mutational events exhibit marked patterns of mutual exclusivity which offer insights for modeling of Shh-MB, and suggest avenues for synthetic lethal approaches to therapy.

## 2.3 DISCUSSION

Initial efforts to subdivide cancers through unsupervised clustering primarily used expression microarrays that focused on the protein coding elements of the genome. Through an unbiased approach using whole transcriptome sequencing, we now identify a large number of non-coding genes as differentially expressed between the molecular subtypes of Shh-MB. This is complementary to our prior discoveries of the most common mutations in Shh-MB, mutations of the *TERT* promoter[42] and mutations of the *U1-snRNA*[100], both of which are non-coding. Assigning biological functions to either individual or groups of non-coding RNA transcripts is obviously more difficult than it is for protein coding genes, and thus the importance and specific biological role of most of these differentially expressed non-coding transcripts will need to be addressed in the future through additional functional experiments.

Shh-MBs harbor few mutations, but frequently have more structural and copy number aberrations in their genomes[31]. For many of these CNAs, the specific resident genes driving clonal

selection were not previously apparent. Indeed, many of the minimally amplified/deleted intervals appeared to be devoid of transcripts when studied with microarrays. Our unbiased transcriptional approach identifies novel transcripts within almost all intervals, and further demonstrates that only a subset of genes within a given region of recurrent CNAs have copy number driven expression, and thus are possible drivers. Discerning the driver genes within regions of recurrent CNAs might allow for the design of rationally targeted therapies.

Transcriptional profiling of such a large cohort of a single molecular type of cancer allows unprecedented understanding of the tumor's genomic landscape, including the identification of novel genes affected by mutations (*GNAS*, *MYCN*, *SETD1B*, *IKBKAP*, and *KDM6A*), and fusion transcripts (*ZBTB20* and *NCOR1*). We also report fusion transcripts in known Shh-MB driver genes, that are likely actually 'tombstones' of large genomic events leading to gene inactivation (i.e., *PTCH1*, and *SUFU*). Other drivers previously known to be amplified in Shh-MB are now identified in additional patients as activated through the creation of fusion transcripts (i.e., *GLI2*), and/or point mutations (i.e., *MYCN* and *GLI2*). These latter events in *GLI2* and *MYCN* further support the driver role for these genes in Shh-MB, and are clinically important as their presence in a tumor will likely render them unresponsive to Sonic Hedgehog pathway inhibition using small molecules. The intriguing finding of highly recurrent fusion transcripts for which there is no support for a structural event in the genome (i.e., *RALGAPA2*, *GNAS*, *NOC4L*, *CHMP1A*) might arise through trans-splicing and requires further functional understanding of their role in Shh-MB biology. Diverse molecular events do appear to converge on a limited set of pathways in Shh-MB, with the different genes showing clear patterns of mutual exclusivity, perhaps telling us about the molecular events that initiate and sustain Shh-MB growth.

## 2.4 METHODS

### 2.4.1 Patient consent

Samples were obtained from the Medulloblastoma Advanced Genomics International Consortium (MAGIC), and from the International Cancer Genome Consortium (ICGC). All patient material was collected after receiving informed consent, under approval and oversight by their respective internal review boards. Control brain RNA was acquired from commercial suppliers (Brainchain, USA) and control RNA-seq libraries were obtained from the Genotype-Tissue Expression (GTEx) project (phs000424.v7.p2)[127].

### 2.4.2 Material processing

Samples were obtained fresh from patients at time of diagnosis and stored at -80°C. Tissues were either manually homogenized using a mortar and pestle after freezing in liquid nitrogen or processed in an automated manner using a Precellys 24 tissue homogenizer (Bertin Technologies, France), following manufacturer's instructions. DNA was extracted by SDS/Proteinase K digestion followed by 2–3 phenol extractions and ethanol precipitation. Total RNA was isolated using the Trizol method (Invitrogen, USA) using standard protocols. DNA and RNA were quantified using a NanoDrop 1000 instrument (Thermo Scientific, USA) and integrity assessed either by agarose gel electrophoresis (DNA) or Agilent 2100 Bioanalyzer (RNA; Agilent, USA) at The Centre for Applied Genomics (TCAG, Toronto, Canada).

### 2.4.3 Messenger RNA library construction and sequencing

Strand-specific transcriptome library construction and sequencing was performed as previously described[128]. Briefly, total RNA samples (2 µg) were arrayed into 96-well plates, and polyadenylated mRNA was purified with a MultiMACS mRNA isolation kit as per the

manufacturer's instructions. First-strand cDNA was synthesized using a SuperScript cDNA Synthesis kit with random hexamer primers. Second strand cDNA was synthesized following the SuperScript cDNA Synthesis protocol by replacing the dTTP with dUTP in dNTP mix, allowing the second strand to be digested using UNG (Uracil-N-Glycosylase, Life Technologies, USA) in the post-adaptor ligation reaction, and thus achieving strand specificity. The cDNA was quantified and checked for quality before fragmentation. Plate-based libraries were prepared following the BC Cancer Agency's Michael Smith Genome Sciences Centre (BCGSC) paired-end (PE) protocol[128]. The libraries, 2×100 PE lanes, were sequenced on the Illumina HiSeq 2000/2500 platform using v3 chemistry and HiSeq Control Software version 2.0.10.

### 2.4.4 RNA-seq alignment

The hs37d5 reference genome FASTA (1000 Genomes Project Phase II) was appended to the C1_2 ERCC spike-in sequences used for C1 Fluidigm, as well as Caltech profile 3 spike-ins sequences by ENCODE. A STAR assembly was then built with this reference and GENCODE (v19) gene annotations using parameter `*--sjdbOverhang 124*`. RNA-seq library reads were then mapped with the built assembly using STAR (2.5.1b) and parameters '*--outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignMatesGapMax 200000 --alignIntronMax 200000 --alignSJDBoverhangMin 10 --alignSJstitchMismatchNmax 5 -1 5 5 --outSAMmultNmax 20 --twopassMode Basic*'.

### 2.4.5 Shh-MB subtype identification

The Similarity Network Fusion (SNF) method[129] was run on 196 primary tumor samples using both RNA-seq gene expression and DNA methylation data as previously described[99] to determine Shh-MB subtypes. The full gene expression and methylation matrix was used since the

SNF method does not require any prior feature selection. The SNFtool R package (v2.2.0) was used with parameters '*K = 40, alpha = 0.6, T = 50'* and then spectral clustering, implemented in the SNFtool package, was run on the SNF fused similarity matrix to obtain the groups corresponding to k = 2–12. The four clusters obtained at k = 4 corresponded to the four Shh-MB medulloblastoma subtypes, α (n = 50), β (n = 42), γ (n = 32) and δ (n = 72).

### 2.4.6 Shh-MB subtype relevant genes (NMI)

The Normalized Mutual Information (NMI) score (as part of the SNFtool package) was identified for each feature (i.e., each gene and methylation probe). For each feature, a patient network based on the feature alone was constructed and subsequently used in spectral clustering. This was then compared to the whole fused similarity matrix through computation of NMI scores as previously described[129]. All features were then ranked according to their NMI scores, representing their importance for the fused network (a score of 1 indicates that the network of patients based on the given feature leads to the same groups as the fused network, whereas 0 means no agreement). The top 10% of features (called subtype-relevant genes) were considered for subsequent analysis.

### 2.4.7 Shh-MB subtype differentially expressed genes

Differential expression analysis was performed using DESeq2 R Bioconductor package[130] comparing samples from one Shh-MB subtype to the samples from the remaining 3 Shh-MB subtypes, considering significant genes with an FDR adjusted p-value <0.05.

### 2.4.8 RNA-seq mutation analysis

RNA-seq mutation calls were performed using GATK (v3.8.0) as previously described[100]. Detected variants were filtered using a panel of normal controls (9 Brainchain and 42 GTEx RNA-

seq libraries), multiallelic mutations, and if candidates had <5 variant reads. Annotation was performed using ANNOVAR software[131].

Mutations with a frequency greater than 0.01 in 1000 Genomes, dnSNP138, Exome Aggregation Consortium database, NHLBI-ESP project, Kaviar Genomic Variant Database, Haplotype Reference Consortium database, Greater Middle East Variome, Brazilian Genomic Variants database, and from an inhouse SNP database (356 sequenced whole genomes) were discarded. Suspected RNA editing events registered in the RADAR database[132] were also discarded. Any deletions which were completely matched with an intron registered in the GENCODE (v19) database were also removed since splice junctions caused by canonical splicing were often miscalled as deletions.

Reads were split into intron-exon segments, however since there remained unsplit-reads overlapping splice junctions, the splice site variant read numbers were re-calculated using a modified 'realignment' function of the GenomonMutationFilter package. The default algorithm remapped reads around detected mutations into reference genomic sequences with and without detected variants. Isoform sequences constructed from the GENCODE (v19) database were added as well as non-annotated isoforms detected using LeafCutter[133] since Shh-MB often contain *U1-snRNA* mutations which cause cryptic splicing. Variants on splice sites were calculated using a modified GenomonMutationFilter and any splice sites with < 5 variants were removed.

Candidates on homopolymer sites were filtered out using the following criteria. (1) homopolymer sequence is ≥ 5 bps, (2) Insertions or deletions, (3) deleted or inserted bases were the same or consecutive base(s) with the homopolymer base. Any mutations only supported by soft-clipped reads were discarded. Additionally, SNPs were filtered if: (1) they were present in

germline SNP clusters which were defined as any regions ≥ 10 bps where SNPs were registered on all the positions in dbSNP150. (2) Any missense or synonymous mutations and non-frameshift indels registered in any of the SNP databases listed above and registered with less than 10 samples or, (3) they were not registered in COSMIC v87. Mutations were also classified as non-pathogenic and removed if: (1) they registered with less than 10 samples in COSMIC v87, (2) the SIFT score was ≥ 0.05, PolyPhen-2 HDIV ≤ 0.908, PolyPhen-2 HVAR ≤ 0.956, "polymorphism" or, (3) "polymorphism_automatic" by MutationTaster, and "predicted non-functional" by MutationAssessor.

Lastly, EBCall[134] was run using the same normal panel. Candidates with $<10^{-3}$ p-value calculated by EBCall were discarded. EBCall uses the samtools mpileup function, so a subset of mutations detected by local-realignment can not be evaluated correctly. Therefore, any mutations which samtools mpileup could call with <5 variant reads, or less than a half of variants reads detected by GATK were not filtered out. Significantly mutated genes ($q < 0.05$) were identified using MutSigCV[135] with its default setting.

### 2.4.9 SNP 6.0 Processing

Affymetrix Power Tools (v1.18.2) was used to process and normalize the probe intensities to generate LRR and BAF using the PennCNV-Affy pipeline[136]. The affygw6.hg19.pfb file was used to map the probes onto the hg19 genome. All other parameters were left on default.

### 2.4.10 Copy number determination and ploidy estimation

The resulting probe level LRR and BAF data were input into ASCAT (v2.4.3)[137]. GC wave correction was then performed, followed by germline genotype prediction. Lastly, the ASCAT algorithm was run to determine copy number values for each genomic region as well as the overall

ploidy and purity of the sample. Samples whose model fit was less than 80% failed their ASCAT processing stage.

### 2.4.11 Copy number post processing

Log ratios for each segment were calculated using the copy number of each segment as well as the average ploidy of the sample, according to the equation: log2((Copy Number)/Ploidy). Adjacent segments whose log ratios differed by <0.25 were then merged using their size weighted mean.

### 2.4.12 Filtering common variants

To derive filtered lists, the gold standard variants listed in DGV release 2016-05-15 for GRCh37 found in at least 1% of samples were used to remove any segments with a 50% reciprocal overlap with segments produced by ASCAT. Once removed, the remaining segments were merged using their size weighted means as before. Further filtering was also done using the list variants in the supporting variants list in the DGV release 2016-05-15 for GRCh37. Studies that had at least 50 subjects as well as variants found in at least 1% of the study were used, and ASCAT segments which had a reciprocal overlap of 80% with these variants were removed. This was performed after removing variants from the Gold Standard list. The resulting segments were then merged using their size-weighted means. Copy number states were assigned to each segment based on their log ratio and their ploidy values. Segments were then grouped into either broad or focal segments depending on whether the segment spanned a length greater than 12Mb, or equal to and less than 12Mb. These broad and focal segments were then used to determine gene level states.

### 2.4.13 GISTIC analysis and increased genes in RNA-seq

The filtered and size-weighted merged segments were then input into the GISTIC 2.0 module on GenePattern[138] and run with slight changes to the default parameters: `*focal length cutoff=0.5, confidence level=0.9, q-value=0.25, remove X=false, run broad analysis=yes*`. The amplified and deleted segments were then extracted from the filtered file and used to determine which genes fell within the region using bedtools (v2.27.1). Microarray annotations and RNA-seq annotations were used to determine the number of detectable genes captured by each method.

### 2.4.14 Gene level determination of copy number state

The copy number segments for each patient were then intersected with the list of GENCODE (v19) genes. The segment that overlapped the greatest amount of the gene was the copy number ratio/state assigned to that gene (e.g. if segment A overlapped with 25% of the gene, while segment B overlapped with 45% of the gene, the gene would be given the ratio/state of segment B. A majority of the gene does not have to be overlapped by a segment to assign it to that ratio/state – similar to "first past the post"). Further to this, for a gene to be gained or amplified, it must overlap at least 50% of the gene, whereas any loss or deletion that overlaps a gene would give that gene this status.

### 2.4.15 Copy number responsive gene

Gene expression was categorized based on either having an amplification, neutral, or with a loss. The Kruskal-Wallis test was performed on each gene to determine if the gene copy number state corresponded with a significant difference in expression. The significance values were adjusted for multiple testing using the Benjamini-Hochberg method, and genes whose adjusted p-values $< 0.05$ were flagged as being copy number responsive.

### 2.4.16 Fusion calling

Multiple fusion callers were used to maximize sensitivity. *Star-Fusion*: STAR RNA-seq read alignment outputs, bam and the 'Chimeric.out.junction' file were input into STAR-Fusion[116] (v0.8.0) using default parameters. STAR fusion results were then further filtered with FusionInspector (v0.8.0) using default settings. *InFusion*: Bowtie2 (v2.2.1)[139] genome assembly was created using hs37d5 (appended to the C1_2 ERCC spike-in, as well as Caltech profile 3 spike-ins sequences) and GENCODE (v19). Infusion[118] (v0.7.3) was ran twice for each sample, firstly with parameters `*--allow-intronic --allow-intergenic --allow-non-coding --allow-all-biotypes*` from which only gene-gene fusions were kept for further filtering. Infusion was run a second time with the addition of more stringent parameters '*--min-split-reads 3 --min-span-pairs 2 --min-fragments 4'* , from which only gene-intergenic or intergenic-intergenic fusions were kept. Afterward both Infusion lists were concatenated. Trans-Abyss: fusions were identified as previously described[31] through *de novo* assembly of each library using Trans-ABySS[117] and then further analyzed to determine fusion orientation. Predicted fusion contigs were split into two sequences by gene and aligned to the reference (hg19) using BLAT (v35). The predicted orientation was determined to be that which allowed fusion partner genes to be in a sense-sense orientation, similar to what is done in STAR-Fusion. Predicted orientations which were not compatible with both fusion partner genes being in a sense-sense orientation were flagged as low confidence orientations.

### 2.4.17 Control sample fusion filtering

A list of blacklisted fusion pairs and breakpoints were created from control GTEx and Brainchain RNA-seq libraries using a (1) fusion contig alignment, and (2) control sample fusion calling strategy. (1) From each detected event, fusion contigs were extracted (110bp from both the

5’ and 3’ partner side where possible) using the scripts supplied by the respective fusion caller. These contigs were then used as a reference for alignment of the normal brain RNA-seq libraries using bbmap (v37.33) with parameters `*mappedonly semiperfectmode qin=33 boundstag=t saa=f g maxsites=1000000 minaveragequality=30 ambiguous=all*`. A fusion was blacklisted if a high-quality control sample read (bp quality average > 30) aligned perfectly with the fusion contig with at least a 20bp overhang past the fusion junction. If the same fusion gene-pair was found in >=2 control samples, it was also subsequently blacklisted. (2) STAR, InFusion, and Trans-Abyss fusion callers were used on all fetal and adult control brain samples using the same parameters as the tumor libraries. Any fusion pairs detected in the fetal MAGIC control and at least 2 adult samples were blacklisted. Furthermore, all fusion breakpoints detected in any control samples callers were blacklisted.

### 2.4.18 Fusion filtering

Any fusions in the control sample breakpoint and gene pair blacklists were filtered out as well as fusions where both fusion breakpoints were called within the same gene (circular RNA artifacts). In an effort to minimize the number of readthrough fusions, fusion pairs within 50 kb and fusions with highly recurrent breakpoints (> 15 samples with the same event) were filtered out unless there were other fusion breakpoints detected in the same genes. Highly expressed genes often contained readthrough fusions so the ratio of ((fusion reads)/200bp)/(gene RPKM) was calculated and any fusions where either partner had a ratio of < 0.01 were removed. Fusions where the read proportion supporting the fusion junction was less than 0.05 for both partners were also removed. From this filtered list, an event was further characterized as a structural variant (SV) based fusion if it was validated by WGS or SNP 6.0 (see Fusion validation method), or if there were multiple fusion isoforms detected with both spanning reads and bridging reads > 0 and

spanning + bridging sum > 20 in at least one partner. For highly recurrent fusion genes, the unfiltered events were manually inspected and salvaged if there was a change in read depth at the fusion junction or WGS/SNP 6.0 support. Gviz (v1.18.2)[140] was used to visualize the change in read depth associated with each fusion event.

### 2.4.19 Fusion validation

*Sanger sequencing*: Primary patient RNA was reversed transcribed into cDNA with SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific). Nested PCRs were performed with primers designed in the 5' and 3' partners of each fusion and Taq polymerase. PCR products were gel extracted using the GenepHlow Gel/PCR kit (Geneaid). Purified PCR products were cloned into DH5α cells by the TA cloning method (TOPO-TA Cloning, Thermo Fisher Scientific) and prepared for Sanger sequencing.

*WGS*: There were different assigned validation states based on the location of the two partner genes relative to the location of WGS detected breakpoints: (1) fused exon is first or last exon and the breakpoint falls into the intergenic region between the gene and adjacent gene, (2) fused exon is the middle exon and the WGS breakpoint falls within an adjacent intron (3) breakpoint falls within a 100kbp window from the edge of the fused exon. Confidence levels were assigned as follows: High - Both partner genes meet conditions (1) or (2), Intermediate - One partner meets condition (1) or (2) and the other partner fulfilled (3), Low - Both partners meet condition (3).

*SNP 6.0*: The position of RNA fusion breakpoints was compared to SNP 6.0 predicted breakpoints corresponding to a change in copy number. The SNP 6.0 breakpoints were padded with a 250 kbp window upstream and downstream, and then each RNA fusion breakpoint in a pair

was checked for support (i.e., support for each breakpoint of a fusion was done respectively) using bedtools (v2.27.1). Support of each fusion was reported as left sided (only the first breakpoint of the fusion was detected), right sided (only the second breakpoint of the fusion was detected), both (both breakpoints of the fusion were detected), or none.

### 2.4.20 PacBio long read cDNA synthesis

cDNA synthesis and library preparation were carried out as described previously[141]. In brief, cDNA was synthesized using SMART-Seq, as follows: Quality and concentration of RNA samples were measured using Agilent RNA ScreenTape (only RIN of >7 was used), after which 9uL of RNA samples (ranging from 590 to 1288 ng) plus a negative control sample (9µL water) were cleaned by adding 18.2µL (2.2X) of room temperature RNAClean XP beads in 0.2mL PCR tubes. RNA samples and beads were pipette mixed and incubated at room temperature for 10 minutes, then the RNA bound beads were pulled down on magnet for 5 minutes followed by two rounds of 250µL 80% ethanol washes. The beads were dried for 5 minutes and then 6µL of Pre-RT mix [Water 2.2µL, RNAse inhibitor 0.1µL, PolyT Primer (12uM) 1.4µL, Triton-X100 (0.4% vol/vol) 1.18µL and dNTP Mix (10 mM each) 1.12µL] was added. RNA beads were re-suspended with a brief vortex and spun down followed by a pre-RT protocol on a thermocycler [72°C 3 min, 4°C 10 min, 25°C 1 min, 4°C Hold]. For each 6µL of Pre-RT reaction, 8µL of RT mix [Water 0.50µL, SS4 first-strand buffer (5×) 2.8µL, DTT (100 mM) 0.35µL, TSO (12 µM) 1.4µL, RNAse inhibitor 0.35µL, SS4 reverse transcriptase 0.70µL, Betaine (5M) 1.4µL, $MgCl_2$ (100 mM) 0.50µL] was added followed by RT on a thermocycler. For each 14 µL of the RT reaction, 5.6 µL of ExoSap-IT was added and mixed well with a brief vortex then spun down, followed by the ExoSap-IT protocol on the thermocycler [37°C 15 min, 85°C 15 min, 4°C Hold]. For each 19.6 µL of the ExoSap-IT reaction, 30.4µL of PCR Mix [PCR-Grade Water 19.4µL, 10X Advantage 2

PCR Buffer (not SA, short amplicon) 5µL, 50X dNTP Mix (Advantage 2 PCR Kit) 2µL, PCR primer (12uM PCR primers) 2µL, 50X Advantage 2 Polymerase Mix (Advantage 2 PCR Kit) 2µL] was added and mixed well with a brief vortex then spun down, followed by 12 cycles of PCR. The fµll-length cDNA product was purified using 0.7x SPRI beads.

### 2.4.21 PacBio long read library preparation and sequencing

Between 373 ng and 912 ng of cDNA (average 660 ng) was used to generate PacBio libraries with a gDNA PacBio Library Preparation kit for cDNA and the manufacturer's 2-kb-template preparation-and-Sequencing protocol and sequenced using a PacBio Sequel2 instrument by performing diffusion sample loading and Sequel Sequencing Kit 2.1 v2 chemistry. Between 2-3 SMRT cells were used per library. On average 1.1M raw subreads with 138K circular consensus reads were sequenced per sample. Long read RNA sequences generated were initially processed using Pacific Biosciences SMRT analysis (V3 smrtlink-release_5.1.0.26412) software. Consensus sequences and a secondary *de novo* transcriptome assembly were produced for each read using the dataset and pbsmrtpipe isoseq tools. SMRT analysis was also used to output all raw subreads from the raw data produced by the sequencer. Raw reads, consensus reads, and transcriptome assembly contigs were then used to validate known fusions or fusion partner genes. Fusions were found by mapping different partial or full-length RNA datasets to the transcriptome reference and filtering to include reads that map to one or more of the known fusion genes. With this strategy both fusion junctions and, in some cases, full length maps of each gene were acquired from long reads. This isoseq pipeline produced on average 6,415 low quality contigs and 399 high quality contigs per sample.

### 2.4.22 Exon chimeric read analysis

Individual STAR aligned chimeric reads with breakpoints matching *RALGAPA2* and *GNAS* recurrent breakpoint locations were combined. Chimeric reads were filtered to only include reads spanning the exon breakpoint junction and to exclude circular RNA transcripts within *RALGAPA2* and *GNAS*. The remaining events were used to generate per-patient circos plots. Genomic sequences flanking the 5' fusion partner were used to calculate the splice consensus sequence using the R package ggseqlogo (v0.1)[142].

### 2.4.23 Whole-genome library construction

Samples were sequenced on the Illumina HiSeq 2000/2500 platform at Canada's Michael Smith Genome Science Centre in the BC Cancer Agency. Sequencing methods are as previously described[128].

### 2.4.24 WGS alignment

Whole genome sequencing reads were aligned to the human reference genome "hs37d5" by 1000 Genomes Project Phase II using Burrows-Wheeler Aligner (BWA) - MEM, (v0.7.8) with '-T 0' parameter. Duplicates were marked using biobambam (v0.0.148).

### 2.4.25 WGS structural variant calling

Somatic structural variant calling was performed using two softwares: Genomon-SV (v0.4.1)[143] and DELLY2 (v0.7.5)[144]. Genomon-SV was run with its default setting. Detected candidates were filtered with *'--min_tumor_allele_freq 0.02 --max_control_variant_read_pair 1 --control_depth_thres 10 --inversion_size_thres 1000 --min_overhang_size 50 --remove_simple_repeat'*. DELLY2 was run using its default setting. The following filter was used for somatic structural calls: *'-m 15 -a 0.1'* for deletion, *'-m 400 -a 0.1'* for tandem duplication and

inversion, *'-m 0 -a 0.1'* for translocation. DELLY2 results were filtered using 341 control whole genome sequence data using 'filter' function of DELLY2 using its default setting. Both results were merged and detected candidate mutations were reanalyzed using velvet *de novo* assembler[145]. Soft-clipped and one-anchor reads were extracted within 1,000 bp of detected breakpoints from tumor and matched control whole genome sequence. Then, contigs were generated using velvet with *'-short'* option and hash length *'11, 72, 10'* (from 11 to 72 with a step of 10). Reference sequences were prepared for remapping which contained reference sequences ±1,200 bp around both paired breakpoints and expected variant sequences with the somatic structural variant. Contigs were mapped to the references using blat version 35 with *'-fine'* function. Only the candidates where contigs from tumor were mapped on the variant sequences and not found mapped in the control were used.

### 2.4.26 MYCN protein structural model

To predict protein structure, the weighted existing structural information of some MYCN and MYC regions from the RSCB PDB (5G1X, 6G6J, 1NKP, 2A93) were used in i-TASSER[146,147]. These models were subsequently visualized and modified in PyMOL (v2.3) and UCSF Chimera (v1.13.1). The prediction is imprecise, as the structure of the N-terminus of MYCN shows intrinsic disorder.

### 2.4.27 Mutual and co-occurrence analysis

The DISCOVER[148] R package (v1.1.0) was used to calculate mutual exclusivity and co-occurrence on high-level copy number, mutation, SV fusion events, as well as arm level gains/losses using default parameters on all patients and on a per-subtype basis. Only known

drivers, significantly mutated, GISTIC copy number responsive genes, and arm level events were included and a corrected p-value < 0.01 was used for downstream analysis.

### 2.4.28 Pathway analysis

*Subtype driving genes*: Enriched pathways were identified using the gProfileR R package[149]. Four gene lists corresponding to the four Shh-MB subgroups were generated by selecting the top 10% of genes having the highest NMI scores and a positive Z-score. Each gene list was ranked by Z-score in decreasing order and analyzed by the gProfileR function with the ordered query setting. Pathways from the Reactome pathway database and biological processes (BP) from Gene Ontology that have between 5 and 1000 associated genes with at least 3 associated genes belonging to gene lists were included in the enrichment analysis. Electronically annotated (IEA) BPs were excluded from the enrichment analysis. P-values of enriched pathways and BPs were corrected using the default multiple-hypotheses testing method (g:SCS) of gProfileR; those with an adjusted p-value < 0.05 were retained.

*Ploidy*: Gene set enrichment analysis was performed using GSEA software. Genes were ranked using the sign of log2FC * -log10(p-value) and analyzed using the pre-ranked option. Gene sets from MSigDB, pathways from Reactome, and biological processes from Gene Ontology were included in the analysis. Gene sets larger than 200 were excluded. Significantly enriched pathways were corrected with FDR and only genes with q-value <0.01 were retained.

*Integrative*: Genes were ranked by the number of patients with a mutation, focal copy number events or SV fusion event in a given gene. Pathway analysis was conducted using gProfileR with the following parameters *'ordered_query = TRUE, exclude_iea = TRUE, min_set_size = 5, max_set_size = 1000, min_isect_size = 2, max_p_value = 0.05 and,*

*correction_method = "analytical"'*. The GMT file was retrieved from gProfileR on March 12, 2019 and included gene sets from Gene Ontology and Reactome.

### 2.4.29 Cytoscape network visualization

*Pathway Enrichment*: Visualization of enriched pathways and biological processes (BPs) was generated with the Enrichment Map plugin of Cytoscape[150,151]. Enriched pathways and BPs are organized into a network, in which similar pathways or BPs cluster together. Nodes represent an enriched pathway or BP; node size is proportional to the number of genes associated to the node; and node colors correspond to the Shh-MB subgroup in which they are enriched. Nodes that are connected by an edge have shared genes in common. Edge thickness is proportional to the number of shared genes among the connected nodes and edges having a Jaccard and Overlap coefficient combined greater than 0.66 were shown.

*Fusion Network*: A curated list of Tier 1 exon-exon and salvaged SV fusions was input into Cytoscape. This network was further filtered to include fusions hubs with a minimum of 5 events as well as their first-degree partners. The network was then manually curated to focus on fusions with SV and/or validation support.

### 2.4.30 Methylation array arm level copy number analysis

The copy number was inferenced using methylation arrays (Illumina Infinium HumanMethylation450 BeadChips). Copy number segmentation was performed from genome-wide methylation arrays using the conumee package (v0.99.4) in the R statistical environment (v3.2.3) as previously described[152,153]. Arm level gains or losses were identified using GISTIC and manually curated by visual inspection of whole genome profiles.

### 2.4.31 Identification of promoter methylation responsive genes

The MethylMix R Bioconductor package[126] was used to identify potential cancer driver genes affected by hypo- or hypermethylation changes (i.e. looking for anti-correlation between methylation level and gene expression levels across samples). Probes were annotated[154] and filtered to only include regions within 1500 bp of transcription start sites. Promoter probes that were correlated were grouped as probe set, then each promoter probe or probe set were considered per gene. Methylation clusters based on mixture model were then identified for each probe or probe set. These were further filtered based on the following criteria: 1) remove promoter probe-gene pairs if one of the methylation clusters has less than 5% of the samples and for pairs with two methylation clusters, 2) pairs were filtered out if the difference of the mean methylation value between the 2 groups was < 0.25 and 3) if the difference of the mean expression value between the two groups was < 0.75. The pairs were further ranked according to a score defined as diff mean * diff exp (difference computed between the 2 extreme clusters). Z-score expression values were used to compute the mean expression differences mentioned above.

### 2.4.32 Illustrations

Oncoprint landscape figures were generated in R (v3.5.1) using the ComplexHeatmap (v2.0.0) library[155]. Gene mutation, fusion summary lollipop type figures were generated using ProteinPaint[156]. Circos plots were generated in CIRCOS[157] (v0.69).

# CHAPTER 3
## Convergent Evolution of Medulloblastoma Metastatic Tumours

Patryk Skowron*, Livia Garzia, Raul A. Suarez, A. Sorana Morrissy, Evan Y. Wang, Betty Luu, Michael, D. Taylor

Medulloblastoma initiates within the cerebellum and spreads throughout the spine and frontal lobe. Lempatminagial dissemination is present in about 30% of patients at presentation and is a marker of poor prognosis. Metastatic tumors are likely seeded early in the course of disease from rare subclones of the primary tumor. In both humans and mice, accumulation of somatic mutations with subsequent selection leads to divergent evolution between the primary and metastatic sites (Figure 3.1)[53]. As a result, targeted therapy directed against the primary tumor is unlikely to be very effective against the metastatic compartment. Little is known about genes that drive dissemination and the context in which they operate since matching patient primary and metastatic samples are virtually nonexistent. Indeed, biopsy of the leptomeninges are rare. High throughput forward genetic screens utilizing transposons have tremendous potential to address this issue in medulloblastoma.

The sleeping beauty medulloblastoma mouse model is a highly penetrant and metastatic model of Shh-MB[53]. Presenting with aggressive dissemination of the spine and brain, it allows for the independent sampling and analysis of multiple metastatic tumors in every mouse. A convergent evolution model was applied on independent metastatic samples to discover high confidence drivers in Shh-MB. The metastatic landscape was highly diverse with a large number of independent drivers across metastasis, even in the same animal. Functional validation emphasized the importance of timing in *Crebbp* genetic alterations and the essential role of *Lgals3* in spinal metastasis.

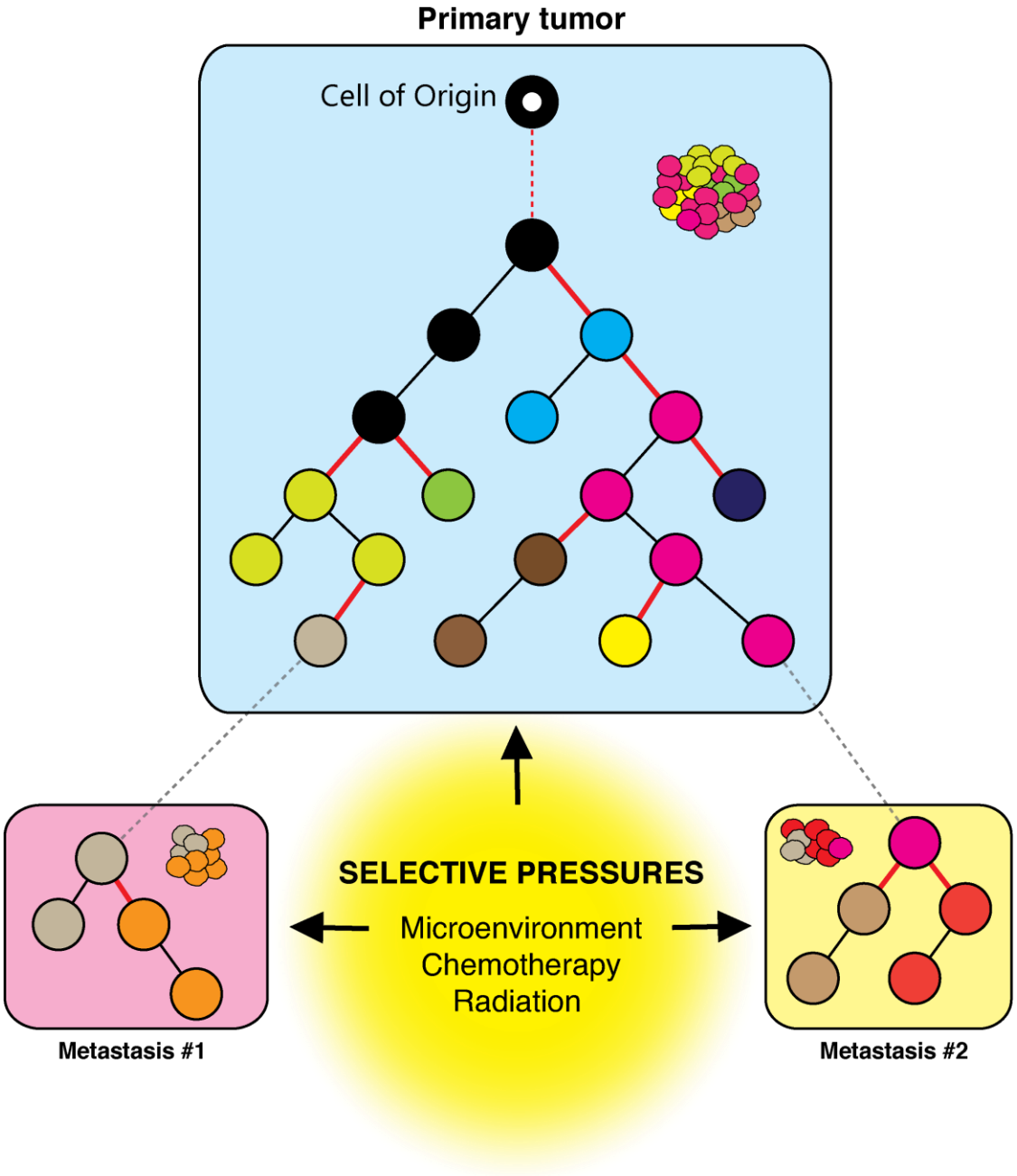



**Figure 3.1 Clonal evolution of metastatic tumors**
Overview of clonal evolution in primary metastatic tumors. Different colors depict different clones.

## 3.1 RESULTS

### 3.1.1 Driver discovery across multiple metastasis

A *Ptch1* heterozygous knockout medulloblastoma model was crossed to mice with Sleeping Beauty (SB) transposition machinery to increase penetrance and metastatic propensity (Figure 3.2a, b). In this model, SB transposase was expressed from a *Math1* gene promoter to localize transposition to cerebellar precursor cells[53]. Multiple brain and spinal locations were biopsied and sequenced using the shear-SPINK protocol (see methods 3.3.3) (Figure 3.2c, d). The majority of biopsied mice had detectable metastatic insertions (108/130 mice), producing a total of 549 metastatic samples; the largest cohort of SB metastasis to date.

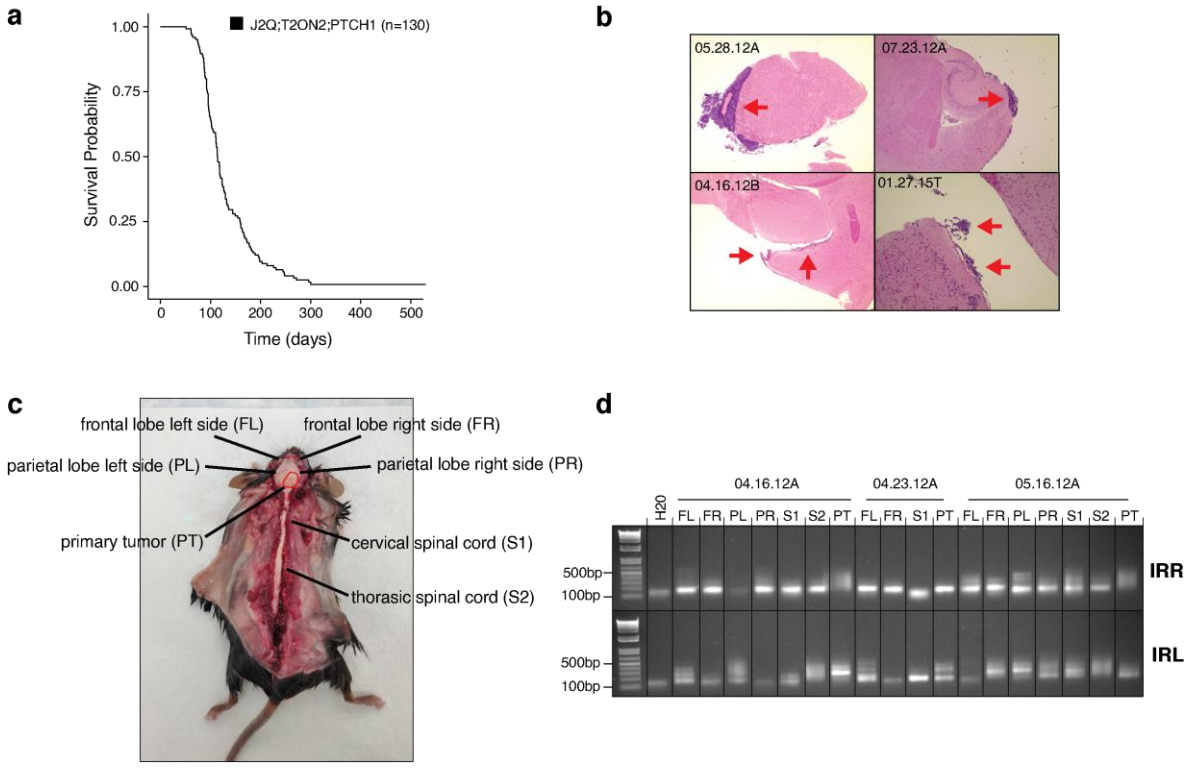



**Figure 3.2 Sleeping Beauty medulloblastoma mouse model**
**(a)** Survival curve of *J2Q;T2onc2;PTCH1* mice. **(b)** Representative H&E histology of frontal and spinal *J2Q;T2onc2;PTCH1* mouse sections at endpoint. Metastatic locations are indicated with red arrows. **(c)** Location of metastatic and primary biopsy sites in the *J2Q;T2onc2;PTCH1* mice at endpoint. **(d)** Shear-SPLINK library preparation quality control of IRR and IRL SB transposon orientations with a water negative control.

During metastatic dissemination, rare primary tumor subclones enter the circulation and/or the cerebrospinal fluid[158]. Through the process of convergent evolution, genes essential to metastasis are selected for independently across multiple metastatic locations[159]. A convergent event is defined as a different structural and/or mutation event in the same gene in different metastatic tumors within the same patient. In our SB model, a convergent event instead describes different SB insertions targeting the same gene in independent metastasis (Figure 3.3a). Not every convergent event is necessarily under the influence of a strong selective pressure, since transposition is a stochastic process these events can happen by chance alone. Therefore, for each gene it is necessary to develop a model to find the expected number of convergent events in a cohort of mice (assuming no evolutionary selection) and compare that to the observed number of events. The methods which are currently used for finding significant genes in SB mice such as Gaussin Kernal Convolution[92], Monte Carlo Simulation[93], and Gene centric common insertion site (gCIS)[94] analysis treat every tumor as an independent sample and are not ideally suited for mice with multiple metastatic sites without modification because they can share early clonal insertions[94].

For each mouse the probability of a random convergent event was modelled using a binomial distribution. Then the expected number of convergent events in a population were compared to the observed number with a Poisson binomial distribution to identify genes undergoing convergent selective pressure (Figure 3.3b). In parallel, the gCIS method was modified by merging all metastatic samples together before running the analysis thus ensuring that early clonal events are not counted more then once per mouse (Figure 3.3c). This was performed at the expense of power due to the smaller cohort size. Lastly, gCIS genes were further examined to find convergent insertion events. (See method 3.3.6 and 3.3.7 for model details).

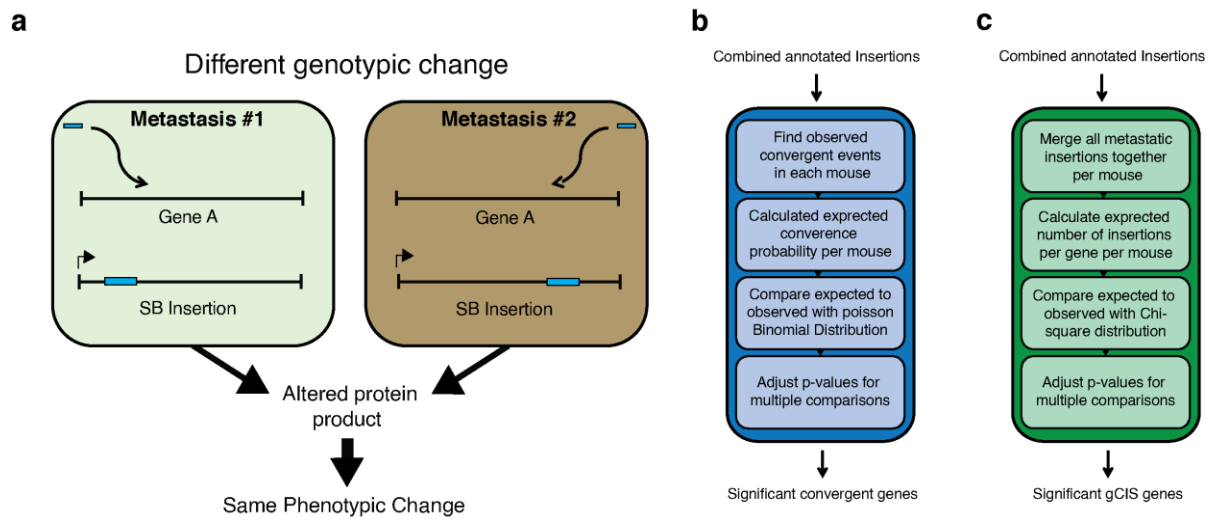



**Figure 3.3 Sleeping beauty metastasis analysis statistical models**
**(a)** Schematic of convergent evolution between two metastatic tumors in the same mouse. A clonal insertion in Gene A is present in both tumors but the location is different suggesting a different parental clone. In both case, alteration of the gene leads to an altered protein product and phenotypic change. **(b)** Overview of the convergent evolution model pipeline **(c)** Overview of the modified gCIS analysis for metastatic driver discovery. See methods section 3.3.7 for model details.

### 3.1.2 Landscape of metastatic alterations in Shh-MB

A large number of metastatic driver genes were discovered by both the convergent and modified gCIS methods (n = 431), some of which were also found in the primary compartment (n = 36) (Figure 3.4a). There was a high degree of overlap in gene lists between the two metastatic driver gene discovery methods (Figure 3.4b). Across samples, the most recurrent metastatic drivers were *Dlg2*, *Cntnap2*, and *Ppp1r12a* (Figure 3.4c). The landscape of metastatic alterations is highly complex with few mutually exclusive driver events (Figure 3.4c). Pathway analysis of significant drivers reveals a convergence of genes in cell adhesion and migration pathways, as well as the unsuspected importance of EGFR and ILK signalling (Figure A10). Each mouse presents with a large number of different metastatic driver genes, most of which are not shared across compartments (Figure 3.5). Most mice have multiple significant convergent genes. In conclusion, Shh-MB has a large number of metastatic drivers and presents with a heterogenous driver profile between samples in the same mouse. Convergent genes would make better targets for therapy since they are by definition found in multiple metastatic samples.


**Figure 3.4 Metastatic and primary driver gene overlaps**

**(a)** Overlap of all metastatic and primary drivers discovered in the Sleeping Beauty mouse cohort (n = 108). **(b)** Overlap of genes detected by gCIS and convergence metastatic driver models. **(c)** Oncoprint summary of drivers found in >= 10 metastatic samples combining the gCIS and convergence methods. Clonal insertions found to be under convergent selection with other metastatic tumors in the same mouse are indicated in red.

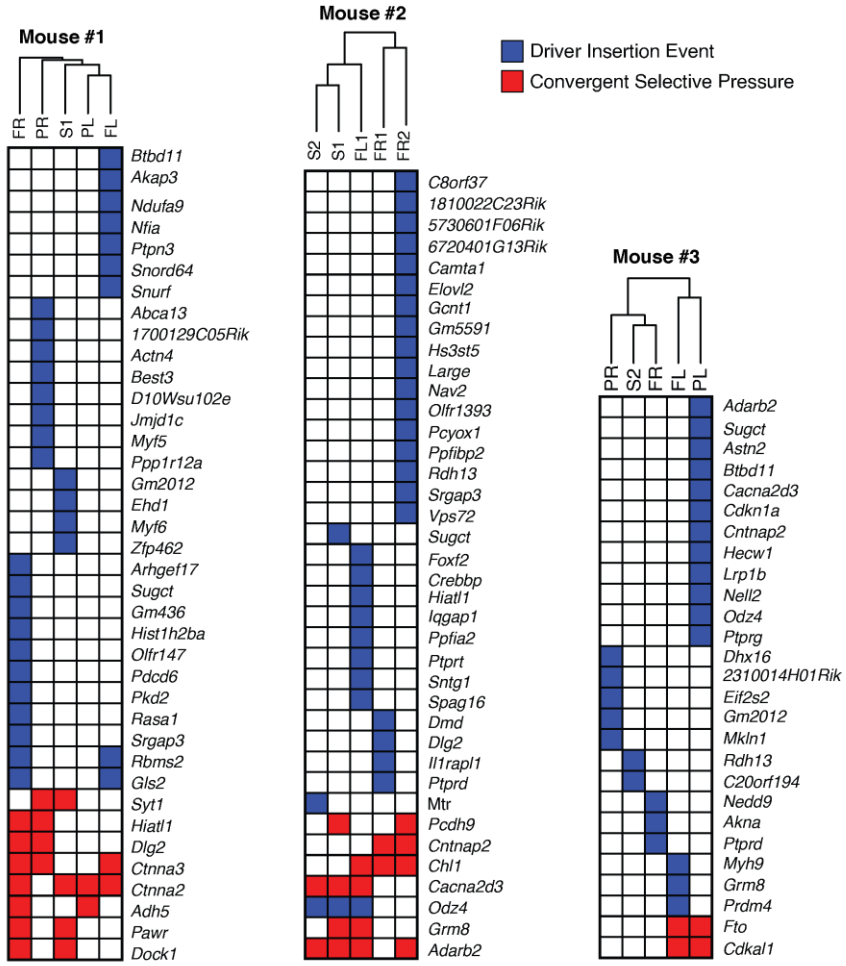



**Figure 3.5 Mouse metastatic driver insertion profiles**
Driver genes found using convergence and gCIS methods in multiple representative mice. Genes under the influence of convergent selective pressure are indicated in red. Each column represents a different metastatic sample in the same mouse. FR - right frontal lobe, FL - left frontal lobe, PL - left parietal lobe, PR - right parietal lobe, S1 - cervical spine, S2 - thoracic spine.

### *3.1.3* Functional validation of *Crebbp* loss-of-function insertions

One of the most recurrent genes that came up in convergence and gCIS screens, and also found to be a driver in primary tumors, was *Crebbp* (17% of mice). CREBBP is ubiquitously expressed and involved in the transcriptional coactivation of a number of different transcription factors through modification of lysine residues on both histone and nonhistone nuclear proteins[160]. It's loss has also been shown to have a prominent role in lymphoma initiation as well as adult Shh-MB tumor growth[125]. The insertion profile in *Crebbp* suggests that this gene is also a Shh-MB metastasis tumor suppressor (Figure 3.6a). Furthermore, IHC for *Crebbp* in primary and metastatic

tissue demonstrate more positive nuclei in primary as compared to metastatic lesions (Figure 3.6b). *NCOA3* is a nuclear receptor which can interact with nuclear hormone receptors to enhance their transcriptional activator functions. It has histone acetyltransferase activity and can also recruit *CREBBP* as part of a multisubunit coactivation complex[161]. *Ncoa3* insertions are mostly clustered in the first intron positioning the transposon promoter in the same direction as gene expression (Figure 3.6c). This is consistent with previous studies where *NCOA3* was found to be amplified in breast as well as ovarian cancers[162]. Interestingly, *Crebbp* and *Ncoa3* insertions were often found in the same mouse, frequently co-occurring within the same metastatic sample (Figure 3.6d). It is therefore hypothesized that the *Crebbp/Ncoa3* complex can regulate aspects of the metastatic phenotype and the knockdown of *Crebbp* can increase metastatic burden in Shh-MB.

Next, a *Crebbp*$^{(flox)}$ mouse was bred with a *Ptch1*$^{(flox)}$, a *Math1*-GFP reporter, and *Math1*-CRE mouse lines (Figure 3.6e). *Crebbp* and *Ptch1* are knocked out early in cerebellar development due to expression of *Math1*-CRE. These mice go on to develop metastatic tumors. Homozygous loss of *Crebbp* at this stage did not change the morphology of the cerebellum (Figure 3.6f). The *Ptch1*$^{(flox)}$ status had the biggest effect on survival (Figure 3.6g). *Ptch1* alteration seemed to synergize with *Crebbp* as suggested by previous literature[163]. GFP fluorescent metastases were clearly observable under the florescence stereoscope at end-point (Figure 3.6h). Unfortunately, there was no significant decrease in metastatic area or count when comparing Crebbp$^{(flox)}$ wildtype to homozygous or heterozygous knockout mice (Figure 3.6h, i). Although, there does seem to be a trend towards a lower metastatic propensity in *Crebbp* knockouts, contrary to our hypothesis. It is known that the timing of *Crebbp* loss is important for Shh-MB primary tumor formation[125], and this can play an important role in metastatic dissemination.

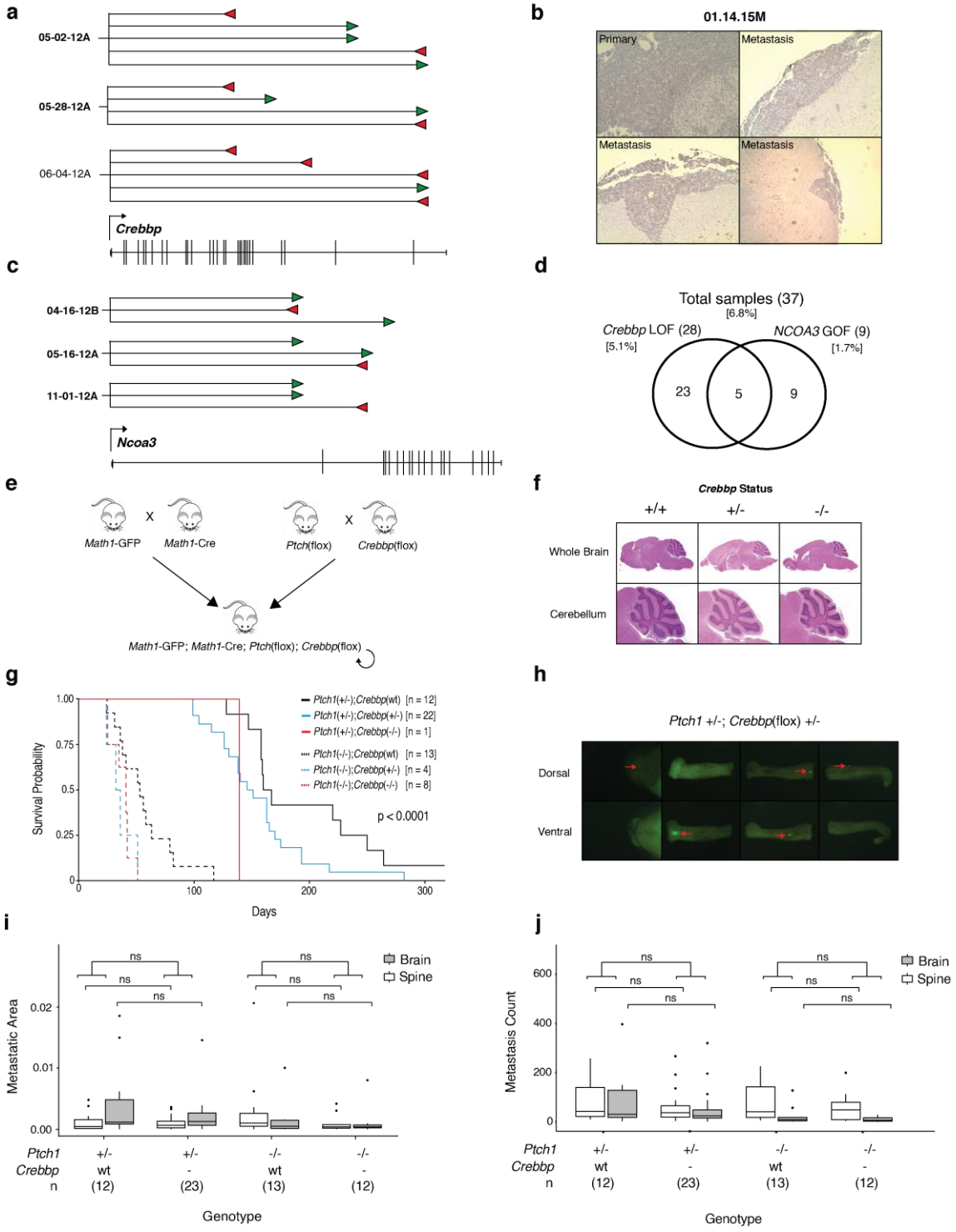



**Figure 3.6 Functional validation of *Crebbp* loss-of-function insertions**

**(a)** *Crebbp* SB insertion profile. Each connected horizontal line represents a different metastatic sample in the indicated mouse. Red arrows are truncating insertions whereas greens arrows indicate overexpression insertions. **(b)** IHC staining of *Crebbp*. **(c)** *Ncoa3* insertion profile. **(d)** Sample overlap of *Crebbp* and *Ncoa3* loss of function (LOF) and gain-of-function (GOF) insertions. **(e)** Schematic of the metastatic Shh-MB *Crebbp* mouse model breeding strategy. **(f)** Brain morphology of *Crebbp* knockout mice. **(g)** Survival of *Ptch1* and *Crebbp* knockout mice. **(h)** Florescence imaging of a *Crebbp* knockout mouse. **(i)** Metastatic area and **(j)** metastatic count comparisons between *Ptch1* and *Crebbp* genotype combinations.

### *3.1.4* Functional validation of *Lgals3* gain-of-function insertions

Another recurrent metastatic driver gene was *Lgals3*, which contains a carbohydrate-recognition domain allowing it to specifically bind β-galactosides. This protein can shuttle between the cytoplasm and nucleus and is secreted onto the cell surface. It plays many roles including inhibition of apoptosis, spliceosome assembly, and cell surface molecule associated signalling. In the context of cancer, overexpression of *LGALS3* has been shown to promote angiogenesis and to enhance tumor cell adhesion to the extracellular matrix[164]. In Shh-MB metastasis it appears that *Lgals3* is likely activated and overexpressed in response to SB insertional mutagenesis (Figure 3.7a). There is little evidence of Lgals3 protein seen in the primary tumor using IHC. Positive signal is usually found in regions of contact between normal and metastatic tissue (Figure 3.7b). To test for the importance of this gene on metastatic burden, *Lgals3* was knocked out in two highly metastatic Shh-MB mouse models (Figure 3.7c, d) with *Math1*-GFP as a reporter. In the *SmoaA1* model, there was a modest change in survival of *Lgals3* knockout mice (Figure 3.7e, f). Florescence imaging of resected spinal and brain tissue revealed that loss of *Lgals3* leads to significantly less metastasis, particularly in the spinal cord regions (Figure 3.7i). Little change in metastatic burden was seen in frontal brain regions suggesting a site-specific effect of *Lgals3* in Shh-MB metastasis. Thus far, no significant difference in metastatic burden has been found in the *Ptch1* SB model. Although the size of these cohorts is still small, there is a trend towards lower spinal metastasis in the spine of SB *Lgals3 -/-* mice.

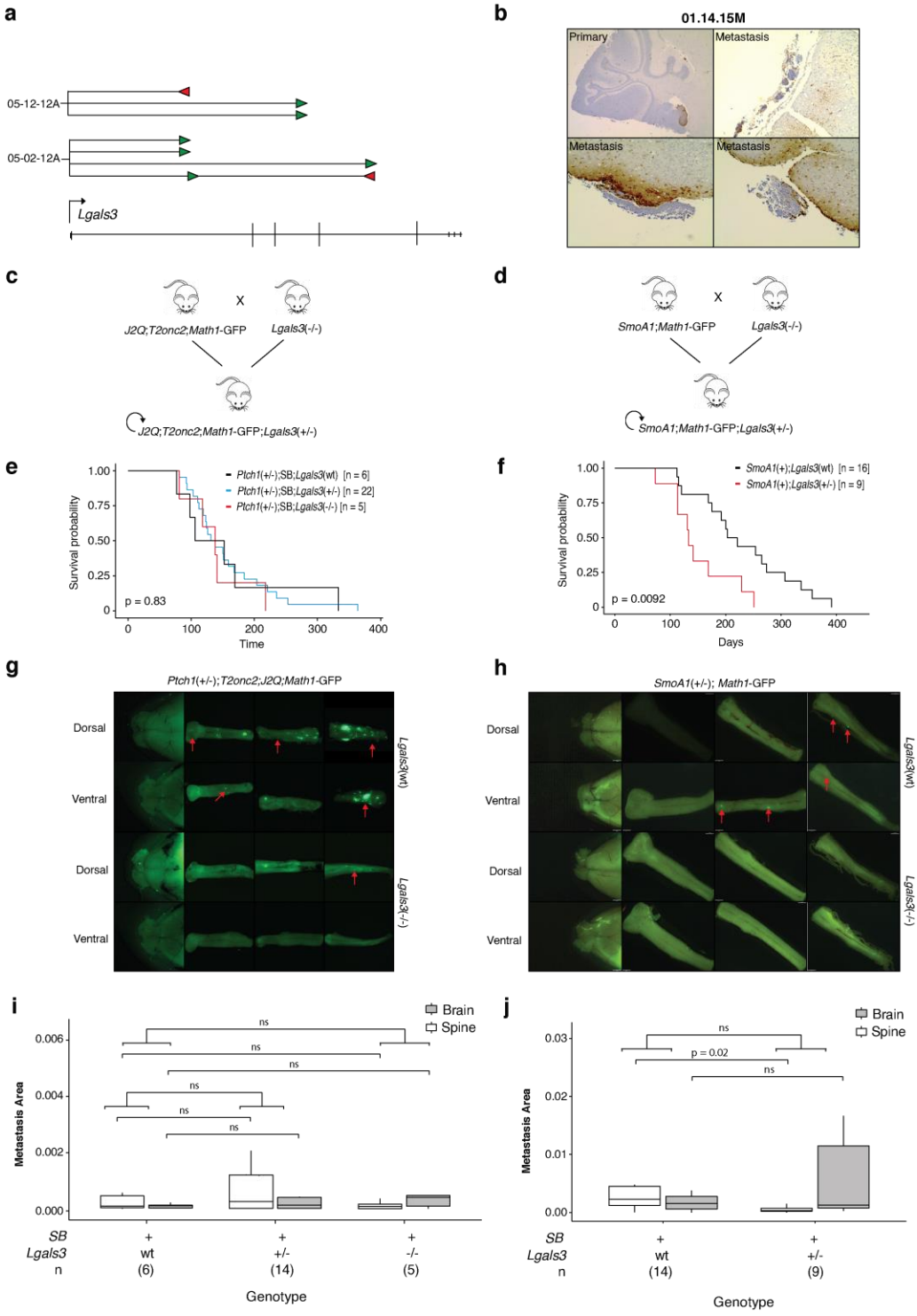



**Figure 3.7 Functional validation of *Lgals3* gain-of-function insertions**

**(a)** *Lgals3* SB insertion profile. Each connected horizontal line represents a different metastatic sample in the indicated mouse. Red arrows are truncating insertions whereas greens arrows indicate overexpression insertions. **(b)** IHC staining of *Lgals3*. **(c–d)** Breeding strategies of the metastatic Shh-MB **(c)** *Ptch1* Sleeping Beauty and **(d)** *SmoaA1* models. **(e)** Florescence imaging of a SB *Lgals3* knockout mouse. **(f)** Florescence imaging of a SB *Lgals3* knockout mouse. **(i–j)** Metastatic area of **(i)** *Ptch1* Sleeping beauty and **(j)** *SmoaA1 Lglals3* mouse models.

## 3.2 DISCUSSION

Regardless of medulloblastoma subgroup, the presence of metastasis is dismal for patient survival. Indeed, metastasis often spreads throughout the leptomeninges and brain rendering it inoperable. The advancement of targeted therapies for metastasis has been lagging due to emphasis on primary and recurrent tumors. This is partly due to the more common surgical resection and biobanking of these tumors. There is a large need for programs and initiatives to biopsy and collect metastatic tumors despite complicated logistics. The Shh-MB SB model has been used in the past to demonstrate vast differences between drivers in primary and metastatic tumors suggesting that current developments in primary tumor targeted therapy will not be able to cure patients of metastasis[53]. Unfortunately, it is also now clear that there is a large degree of heterogeneity among metastasis in the same patient suggesting that multiple approaches will need to be taken to effectively treat this disease. Despite this, there are drivers with a convergent selective pressure for different events occurring within the same gene (i.e. *Crebbp*, *Pawr*, and *Lgals3*). Since these are by definition found in multiple metastasis in the same mouse, treatments focusing on such targets would be more likely to make a lasting impact on patient survival.

*CREBBP* is commonly mutated in Shh-MB, particularly in adult patients. Studies have shown a large phenotypic difference between patients with somatic versus germline alterations of *CREBBP*, suggesting that the timing of mutation is important. This concept may extend to its role as a metastatic driver. *Math1* expression starts early in brain development particularly in neural progenitors of the cerebellar rhombic lip, which eventually differentiate to form granule cells. Knockout of *Crebbp* at this stage does not change the morphology of the cerebellum but it may alter the evolutionary trajectory of the primary tumor such that it is less metastatic. Since *Crebbp*

is a primary tumor driver it may also be possible that the mice die before developing metastasis. Ideally, there is a need for a conditional knockout of *Crebbp* after tumor formation. This would be possible in a tumor expressing luciferase from a *Math1* promoter. Here *Crebbp* could be deleted during tumor progression through application of a tamoxifen inducible CRE system. More functional work needs to be done to demonstrate the role of *Crebbp* as a bona fide metastatic driver.

On the other hand, it is clear that *Lgals3* plays an important role in metastasis. Even the loss of a single allele significantly decreases metastasis on the spine of *SmoA1* mice. It seems likely that this gene is important for attachment of metastasis onto the spine since it was found to be secreted into normal tissue at sites of contact. This is in line with research in breast cancer metastasis which show that overexpression of *LGALS3* promotes angiogenesis and more importantly enhances tumor cell adhesion to the extracellular matrix[164]. It is less likely that *Lgals3* plays a role in the initial extravasation process since there is just as much (if not slightly more) brain metastasis present in these mice. In this scenario, there would be a similar burden of circulating tumor cells, but without *Lgals3* they would be unable to seed the spine. More experiments need to be done to tease out the mechanism, and potentially inhibit the *Lgals3* driven metastasis.

## 3.3 METHODS

Unless otherwise stated all chemical reagents where obtained from Sigma.

### 3.3.1 Genotyping

Mouse eye or tail clippings were lysed in and incubated overnight at 56°C, then incubated for 5min at 95°C. 1µL of DNA lysate was used in PCR reaction along with 2.5µL 10X PCR buffer, 1µL $MgCl_2$, 0.5µL 10mM dNTPs, 0.5µL primer 1, 0.5µL primer 2, 2µL PCR loading Dye, 16.8µL $ddH_20$, and Taq polymerase. Refer to Table 2 for all genotyping primers used.

### 3.3.2 Tissue processing

Tissue samples were frozen in liquid nitrogen and pulverized with a mortar and pedestal. Powdered samples were then incubated overnight in 500ml lysis buffer (10mM Tris-Cl and 0.1M EDTA with pH 8.0, 0.5% SDS) and 2.5µL proteinase-K (final concentration 100µg/ml) at 50°C. The solution was cooled to room temperature and 500µL phenol, equilibrated with 0.1M Tris-Cl (pH 8.0), was added. The two phases were gently mixed for a minimum of 1 min by inverting the sample tube until the two phases have formed an emulsion. samples were then centrifuged at 10,000g for 10min at room temperature. The viscous aqueous phase was transferred to a new tube and organic phase was discarded. The extraction was repeated with phenol once more. Into the aqueous sample, 50 µL 3M sodium acetate, 1000 µL ethanol, and 1 µL glycogen was added. The samples were gently shaken, and then incubated at -20°C for 20min. A DNA precipitate was present after the incubation. The samples were centrifuged for 10min at maximum speed. The supernatant was removed and 500 µL of 70% ethanol was added without disturbing the pellet. The samples were centrifuged at max speed for 10min. The ethanol washing step was repeated once more and, after supernatant removal, the pellet was left to dry at room temperature in an open tube

until all visible traces of ethanol had disappeared. Molecular grade water was added (25 µL) and the pellet was dissolved (fig. 9). Concentration readings were taken using a Nanodrop spectrometer.

## 3.3.3 SB insertion sequencing Shear-SPLINK

### *3.3.3.1 DNA shearing*

A Covaris S220/E220 Focused-ultrasonicator (Covaris Inc. , USA) was used to shear 100ul of each DNA sample with parameters: Peak Incident Power (W) - 140, Duty Factor - 10%, Cycles per Burst – 200, Treatment Time – 80, Temperature 7 °C, Water Level - 12cm. Quality control was run on random samples (1 in 10 samples) on a 2% agarose gel in TEA buffer (pH: 7.4, 0.2M Triethanolamine, 1mM MgSO4, 1mM EDTA, 0.01% Azide) with a 1kb protein ladder (Invitrogen) for 20min. A wide band in the 300bp region was indicative of successful sonication.

### *3.3.3.2 End repair*

Epicenter End repair kit (Lucigen Corporation, USA) was used with 20 µL of Sonicated DNA, 0.5 µL dd$H_20$, 3µL kit buffer, 3µL dNTP, 3µL ATP and 0.5 µL kit enzyme mix. Sample was incubated at RT for 45min and then 10min at 70 °C

### *3.3.3.3 Adaptor ligation*

Linker+ and linker- primers (Table 3) (100 µM) were mixed at 1: 1 ratio in Sodium-Tris-EDTA buffer (50mM NaCl, 10mM Tris-Cl - pH 8.0, 1 mM EDTA - pH 8.0). Primer solution was heated to 95 °C for 5min and slowly cooled to room temperature to facilitate formation of the double stranded adaptor. Fast-link ligase kit (Lucigen Corporation, USA) was used with 30µL end-repaired DNA, 1.75µL ATP, 1.64ul adaptor mix, 0.5µL kit buffer, and 1.11µL Fast-Link ligase.

Solution was incubated at RT for 45min and then the enzyme was inactivated with an incubation at 70 °C for 15min.

#### *3.3.3.4 Concatemer digestion*

The ligation solution was digested with *Bam*HI to break apart the transposon concatomers in the normal tissue and prevent their amplification. Then 35µL of the adaptor ligation solution from previous step, 1 µL High Fidelity (HF) *Bam*HI, 1.5µL NEB buffer 4, 5µL 10X bovine serum albumin (BSA), and 4µL dd$H_20$ were incubated overnight at 37 °C.

#### *3.3.3.5 Primary PCR*

Two primary PCR reactions were set up for each side of the SB transposons (IRR and IRL). 5 µL DNA mix from previous step, 12.25µL dd$H_20$, 5µL 5x Phusion buffer, 0.75µL 10mM $MgCl_2$, 0.5µL 10mM dNTPs, 0.5µL 10Mm IRR or IRL primer, 0.5µL 10Mm Linker-A1 primer, and 0.5µL Phusion Taq (Sigma, USA) were mixed together (Refer to Table 3 for primer sequences). The sample was run using the following PCR cycle protocol: 1) 98°C (30s), 2) 98°C (20s), 3) 55°C (30s), 4) 72°C (60s), Steps 2,3,4 repeated 25 times, 5) 72°C (60s), 6) 4°C (hold). 3µL of the primary PCR samples were diluted 1:50, vortexed and incubated at RT for 30min.

#### *3.3.3.6 Secondary PCR*

PCR mix was made with 4 µL DNA mix from previous step, 32.5µL dd$H_20$, 10µL 5x Phusion buffer, 1µL 10mM dNTPs, 2µL 2.5µM IR-barcoded transposon primer, 0.25µL 10 µM Linker-A2 primer, and 1µL Phusion Taq. A touch down PCR cycling protocol was used: 1) 98°C (180s), 2) 95°C (30s), 3) 49°C (30s), 4) 72°C (60s), Steps 2,3,4 repeated 10 times, 5) 95°C (30s), 6) 53.3°C (60s), 7) 72°C (120s), Steps 5,6,7 repeated 25 times, 8) 72°C (60s), 9) 4°C (hold). Refer to Table 3 for primer sequences.

#### *3.3.3.7 Sequencing preparation and submission*

Secondary PCR sample (10 µL) was analyzed on a 1.5% agarose gel. A bright band in the 300bp region was expected. The PCR products were pooled and purified using Qiagen purification kit and resuspended in 50µL TE buffer. A Nanodrop was used to determine the concentration of purified DNA. A maximum of 96 samples were pooled together from the IRL and IRR libraries per lane with a final concentration of 20-25ng/ µL. This pool was incubated at 40 °C for 30min and submitted for sequencing on the Hiseq (Illumina, USA) paired-end 2 x 126bp.

### 3.3.4 Read preprocessing and alignment

Adaptors were trimmed with cutadapt (v1.8) with parameters *'-m 5 --no-indels --discard-untrimmed -g R1_5prime=^NNNNNNNNTGTATGTAAACTTCCGACTTCAACTG'* from read 1 (R1) for each sample. Since the SB insertions recognize and insert into a TA dinucleotide, only the reads starting with a TA were kept for downstream steps. Read 1 reads were then paired with their respective paired reads (R2) and aligned with novoalign (v3.05.01) using parameters '*-r ALL 1 -R 0 -c 8 -o SAM*' with the mm9 mouse genome assembly. Aligned sam files were then converted to bams for downstream analysis.

### 3.3.5 Insertion read processing and filtering

#### *3.3.5.1 Annotation*

Each detected insertion was annotated using refFlat tables from the UCSC genome database. Using the chromosomal address the following information was extracted: [tumor ID], [gene name], [region of gene hit (e.g. intron, exon, and promoter)], [predicted affect of insertion on the expression of the gene], [number of reads on this insertion site within the sample],

[orientation of the transposon relative to the gene]. Some insertion events were not annotated because they did not occur within a known gene.

#### *3.3.5.2 Insertion clonality estimates*

Clonality was estimated using Shear-SPLINK's unique ligation point (ULP) score which quantifies the number of unique positions in the ligation point between genomic DNA and the adaptor for every given insertion[95]. For different reads mapping to the same insertion a different sequenced fragment length is indicative of a unique ligation point (ULP). Each bam file was converted into a bed file using samtools (v1.9) command '*bedtools bamtobed -bedpe -i stdin*'. The length of each fragment was then calculated by extracting the start coordinates of R1 and the end of R2. In cases where only one read mapped the fragment, the mapping read length was extracted. Fragment Lengths greater than 700 are not possible with paired end sequencing on the Hiseq platform and were therefore likely alignment artifacts, these lengths were all set to 0. The number of unique fragments lengths (i.e ULP count) was then calculated and appended to the annotated insertion list. A ULP score was calculated by dividing the insertion ULP count by the highest ULP in each library (i.e. range is [0,1] and a score of 1 would represent the most clonal insertion in a given library)

#### *3.3.5.3 Clustering*

Insertion locations within 5bp are stitched together as a cluster in the same library and then merged since they more likely represent the same insertion with an alignment artifact. The insertion with the highest number of mapped reads is assumed to be the 'true' insertion location. The read count of the insertion cluster was summed up with the highest ULP is used for the cluster. The IRL and IRR libraries were then merged together. If an insertion was detected in both libraries

(i.e. transposon orientations) the read and ULP counts from the higher ULP score insertion was used after merging.

#### *3.3.5.4 Filtering*

An insertion was filtered out if (1) found in more than 1 control biological or technical library, (2) ULP count of 1, (3) ULP score less than 0.05, (4) found on the donor chromosome (SB76 = chr1 and SB68=chr15). (5) for every pooled set of libraries if there is more than one merged library with the same insertion (in different mice), only the insertions in the mouse containing with the highest read count is kept. This is to ensure that there is no cross contamination between libraries pooled together.

### 3.3.6 Gene centric common insertion (gCIS) analysis

Refer to Benjamin *et al.* for detailed explanation of the gCIS method[94]. A Sleeping Beauty transposon can only insert into a TA dinucleotide, so the probability of an insertion in a particular gene ($p$) is the number of TA sites in the gene ($TA_G$) divided by the number of TA sites in the entire genome ($TA_T$).

$$p = \frac{TA_G}{TA_T} \quad (1)$$

Therefore, if there are N insertions in tumor I, then the expected number of insertions in the gene is:

$$E = {}_p(N_i) \quad (2)$$

Using the Chi-squared distribution we can compare the expected number, $E$, of insertions in tumor $i$ to the observed number of insertions, $O$, in tumor $i$ using the Chi squired distribution.

This is assuming that O is discrete, and can have a value of either 1, if an insertion is present, or 0, if absent.

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

This test is repeated for every gene. P-values are adjusted using the stringent Bonferroni group-wise correction. Corrected p-values (Q) <0.05 are called significant.

### 3.3.7 Metastatic convergent evolution model

Within a tumor (*X*) the probability of x random insertions for a gene follows a binomial distribution where *n* is the number of high abundance (i.e. clonal) insertions found in the tumor and *p* (from (1) above) is the probability of a single gene insertion event:

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (4)$$

In a mouse with a primary tumor and multiple metastatic sites each compartment is independent of one another therefore the probability of a convergent event (*C*) is all the combinations by which $P_X(x = 0)$ in the primary tumor and $P_X(x > 0)$ in greater or equal to 2 metastatic sites. In the case of one primary X and two metastatic sites Y and Z this becomes:

$$P(C) = P_X(x = 0) * P_Y(x > 0) * P_Z(x > 0) \quad (5)$$

Using the binomial distribution this can be expanded to this form (assuming that $n_x$, $n_y$, $n_z$ is the total number of clonal insertions in tumor X,Y,Z respectfully):

$$P(C) = P_X(x = 0) * (1 - P_Y(x = 0)) * (1 - P_Z(x = 0)) \quad (6)$$

$$P(C) = \binom{n_x}{0} p^0 (1-p)^{n_x - 0} * (1 - \binom{n_y}{0} p^0 (1-p)^{n_y - 0}) * (1 - \binom{n_z}{0} p^0 (1-p)^{n_z - 0}) \quad (7)$$

$$P(C) = (1-p)^{n_x} * (1 - (1-p)^{n_y}) * (1 - (1-p)^{n_z}) \quad (8)$$

For each mouse $i$ and in for a particular gene, the convergent probability ($P(C) = p_i$) can have a different value due to differences in the number of metastatic sites and number of insertions. To compare the expected number of convergent events with the observed number of convergent events, the Poisson binomial distribution was used where the probability of having *x* convergent mice out of a total of n can be written as the sum below and $F_k$ is the set of all subsets of *x* integers that can be selected from {1,2,3,...,n}.

$$P_X(x) = \sum_{A \in F_k} \prod_{i \in A} p_i \prod_{A^c} (1 - p_i) \quad (9)$$

Using this, the probability of x or more convergent mice can be calculated to find the p-value for each convergent gene. False discovery rate (FDR) is used to adjust the p-values and genes with Q<0.05 are called significant (i.e. likely under the influence of convergent selection).

## 3.3.8 Metastasis imaging

### *3.3.8.1 Tissue Dissection and Imaging*

The brain and spine are dissected out of the mouse at end-point in one piece and transported in phosphate-buffered saline (PBS). Tissues are then washed with PBS and placed on a petri dish. The contiguous tissue is then separated into 4 regions (i.e. brain, cervical spine, thoracic spine, and lumber spine) to ensure there is no overlap between image acquisitions. The *Crebbp* and SB *Lgals3 Math1*-GFP mice were imaged on the Leica Florence stereomicroscope (Wetzlar, Germany) using the Velocity imaging suite (v6.3). The *SmoA1 Lgals3 Math1*-GFP mice were imaged on the Nikon SMZ25 stereoscope (Minato, Tokyo, Japan) using the NIS elements imaging suite (v5.0.2.01).

#### *3.3.8.2 Metastatic area quantification*

All metastatic images were imported into the Met NIS elements imaging suite as either .tiff or .nd2 image formats and calibrated using the true size of each pixel. Each individual florescent met was manually circled as well as the total spine/brain area. In each mouse the metastatic spine area was summed together and divided by total spine area to calculate the spine met proportion. This was repeated for the brain lobe metastasis to calculate the brain met proportion. The total count of individual metastasis was summed and normalized by the size of each respective spine. Wilcox test was used for all statistical comparisons between groups.

### 3.3.9 Pathway enrichment analysis

Pathway analysis was performed using gProfiler. Genes were ranked by frequency of recurrence. Mouse gene sets from MSigDB, pathways from Reactome, and biological processes from Gene Ontology were included in the analysis. Significantly enriched pathways (FDR qvalue <0.05) were visualized using Enrichment Map in Cytoscape. Node sizes are proportional to the number of genes and edge weight (Jaccard and overlap coefficient set at 0.66 cutoff) represents the number of shared genes between each gene set.

### 3.3.10 Illustrations

All plots were generated using R (v3.5.1) Oncoprint landscape figures were generated using the ComplexHeatmap library[155].

# CHAPTER 4
## Sonic Hedgehog Medulloblastoma: Where do we go from here?

#### *4.1.1.1 Shh-MB landscape studies*

Medulloblastoma is one of the most common types of cancer in the developing brain and a significant cause of morbidity in children. It has been the subject of intense investigation in multiple groups throughout the world. Many hypothesis generating studies have focused on deciphering the pattern of somatic alterations across medulloblastoma. Each of these studies have revealed a more complete picture of this cancer, opening up new avenues for biological investigation. Up to date, large scale bioinformatic studies have utilized expression microarrays, SNP 6.0 copy number arrays[31], methylation arrays[99], proteomics[165], and whole genome sequencing[34–36,166]. No technology is perfect, and each is best suited for a particular angle of investigation. In Chapters 2 and 3, a large variety of both human and mouse datasets were integrated together leveraging each of their strengths in order to paint the most comprehensive picture of alterations across Shh-MB. A number of therapeutic targets have been revealed which will be the subject of intense functional investigation for years to come.

#### *4.1.1.2 Shh-MB functional validation model systems*

There were a number of novel primary and metastatic driver genes discovered in this study that will need to be validated. Mouse transgenic models are most ideal, but they are time consuming, with a risk for embryonic lethality. Furthermore, the timing of alteration often plays an important role in cancer biology. Xenografts present an attractive alternative, but unfortunately in Shh-MB the choices are limited. Many established Shh-MB lines (i.e. DAOY, ONS76)[167] have very little resemblance to the original tumor, since in-vitro culture conditions select for a different

set of traits over the countless passages. Although there are now a number of in-vivo lines maintained in NSG mice (i.e. Med-1712FH, MED-813FH), these are unfortunately unable to grow in-vitro long enough for efficient genetic manipulation. Luckily, a recent report provides a valuable alternative through the use of human derived pluripotent stem cells reprogrammed to generate Shh-MB tumors[124]. Neuroepithelial stem cell (NES) lines can be generated from human pluripotent stem-cell derived neural rosettes propagated in long time culture. Even if extracted from an adult they show similar characteristics to fetal NES cells. These progenitors differentiate to cerebellar granule neural precursors which are the cells of origin for Shh-MB[15]. Through introduction of *PTCH1* or *MYCN* alterations, commonly found in Shh-MB, these cells form faithful orthopedic transplantation models. These would make ideal models for functional validation of candidate primary and metastatic driver genes because they are easy to propagate in culture, can be genetically modified, and don't have any confounding somatic alterations.

## 4.1.2 Shh-MB primary tumors

### *4.1.2.1 RNAseq*

RNAseq can reveal both the quantity and presence of RNA in a biological sample at a given moment in time. It is unbiased in that it can look at the entire repertoire of expressed genes without any presumptions, which is an advantage compared to microarray based analysis. Since mRNA is transcribed directly from DNA, it can also serve as a proxy for somatic alterations in DNA. Mutations and structural variants (in particular fusions) can be confidently called from RNAseq. The 30-40x whole genome analysis commonly reported does not provide uniform coverage meaning fusion events can be missed even if they are highly expressed. Even if a structural variant is detected through such analysis, it is not clear what it's doing without

corroboration at the mRNA or protein level. Likewise, since RNAseq can only evaluate expressed genes, it cannot call events from unexpressed genes. RNA editing also posses a technical challenge and serves as a confounding factor while calling point mutations. Lastly, like any short-read technology, RNAseq fragments can be erroneously filtered out/removed if a confident match is not present in the genome, potentially omitting important somatic events.

#### *4.1.2.2 RNAseq Study Summary*

In Chapter 2, RNAseq was utilized to acquire a comprehensive picture of the transcriptional landscape in human Shh-MB. Despite the use of poly-A enriched RNA it was clear that noncoding RNA play an important role in this disease. Of course, many non-coding transcripts that don't utilize a poly-A tail would be missed by this analysis. It was also concluded that the fusion landscape of Shh-MB was much more complex than previously theorized. Fusions in Shh-MB are very common, especially in *Tp53* mutated Shh-α patients. There were also a number of loss-of-function fusions discovered in tumor suppressor genes such as *PTCH, SUFU,* and *NCOR1*. Collectively ≥20% of Shh-MB patients were shown to have fusions. Mutations were called despite the lack of germline controls. The most interesting were events in *MYCN*, *GLI2*, *PPM1D, GNAS,* and *IKBKAP*. Mutations in *MYCN* were found in the region binding the ubiquitin ligase *FBXW7* which was also mutated in a fraction of tumors. Therefore, ~20% percent of Shh-MB patients had alterations that stabilize or overexpress MYCN protein. Another surprising discovery was the presence of recurrent fusions hubs in *RALGPA2* and *GNAS* which are not a consequence of structural alterations in the DNA, but rather formed through trans-splicing.

#### *4.1.2.3 Functional validation of GLI2 and MYCN*

One of the most common focal amplifications in Shh-MB encompasses *GLI2*, an important mediator of Shh signalling. This is the first study detecting recurrent mutations in *GLI2*. Despite being within the 'activation' domain of *GLI2*[101] it is not clear if these missense events lead to a more active *GLI2* protein (Figure 2.3a). Mutations can be introduced into the neuroepithelial stem cell (NES) Shh-MB model using CRISPR/cas9 to generate the appropriate amino acid changes (i.e. p.P1028L, p.H1073Y, p.Q1323H, p.A1514V). Then a luciferase reporter assay could be used to get a readout of *GLI2* mutant activity. Specifically, the NES line can be transfected with a vector containing GLI binding sites and a δ-cristallin basal promoter driving expression of a luciferase gene allowing for a readout of GLI2 transcription factor activity[168]. In parallel, *GLI2* mutant and wildtype NES cell lines could be orthopedically transplanted into mouse brains and monitored for tumor growth. RNAseq sequencing and classification could be used to infer SHH pathway activity and ensure that resulting tumors closely resemble Shh-MB.

Another gene to validate is *MYCN*, which was shown to be mutated around the FBWX7 binding motif in 4% of patients (Figure 2.3c). It is hypothesized that these mutations would lead to a more stable MYCN since they would no longer be targeted for ubiquitin mediated protein degradation. In support of this hypothesis, there were also loss-of-function events in *FBXW7* which were mutually exclusive of *MYCN* alterations. Like in *GLI2*, mutations can be introduced into the NES Shh-MB line using CRISPR to generate the appropriate amino acid changes. Western blot for MYCN can be utilized to get a readout of protein stability compared to normal controls. The tumorigenicity can also be assessed after orthotopic transplantation into the mouse cerebellum. It would be important to show that the increased stability is mediated through a decrease in FBXW7

mediated ubiquitination rather than other factors. FBXW7 can be blotted through co-immunoprecipitation of MYCN to check for decreased binding. Furthermore, a ubiquitination assay can be used to show that MYCN is no longer targeted for degradation. Alternatively, it is possible to use transient expression of flag tagged *MYCN* and *FBXW7* vectors in NES cells to streamline co-immunoprecipitation of the protein complex.

#### *4.1.2.4 Trans-splicing*

Alternative splicing is present in ~90–95% of all human genes. This process enhances genetic diversity by generating multiple protein isoforms from the same set of exons[169]. Most common is cis-canonical splicing which generates mRNA diversity through the use of exon skipping, exon retention, and alternative 3' and 5' splice site selection. Less commonly seen is cis-splicing of adjacent genes (i.e. readthroughs), cis-backsplicing (i.e. circular RNA), intron retention, and cryptic splicing. In contrast to cis-splicing which involves a single or adjacent gene, trans-splicing brings together transcripts from spatially distinct genomic loci[170]. Spliced leader (SL) trans-splicing involves splicing of small nuclear RNAs (i.e. splice leaders) onto select pre-mRNA. This process occurs in diverse groups of eukaryotic organisms, including nematodes and flatworms, and serves as an alternative way to cap mRNAs[171]. SL independent trans-splicing chimeras have been detected in higher eukaryote organisms including fruit flies, mice and even humans but its purpose and mechanism has remained elusive. In Drosophila, *mod(mdg4)* is commonly trans-spliced to a variety of 3' partners[172]. Using tiling deletion on the *mod(mdg4)* 3' intron immediately after the trans-spliced exon, it was shown that there is a highly conserved motif able to bind U1 snRNP, and together with an enhancer motif mediate trans-splicing. Canonical spicing machinery was used in this context to generate a Y-structured outron instead of the typical

lariat. There have been numerous accounts of trans-splicing in human cell lines [170]. Trans-splicing was also investigated as a means of gene therapy to replace mutated or deleted exons[170], although this application has been limited by the difficulty involved in introducing genetic vectors into tumors. In human endometrial stromal tumors *JAZF1-SUZ12* is a common fusion mediated through structural arrangements at the DNA level. Strikingly, the chimeric transcript is also commonly found in stromal cells of non-cancerous individuals[173] suggesting that trans-splicing could be a precondition for RNA-mediated DNA recombination. A similar phenomenon is also observed in prostate cancer (*SLC45A3-ELK4*)[174] and neoplastic haematopoietic cells (*IGH-BCL2*)[175]. With progressing technology and large scale sequencing studies there is an increasing number of trans-spliced chimeric fusions being detected and categorized[176]. Most recently a PAN-CAN report attempted to assign structural DNA rearrangements to each recurrent fusion across a large panel of tumors and suggested that up to 18% of fusion transcripts in cancer are generated through trans-splicing[123].

#### *4.1.2.5 Deciphering the mechanism of trans-splicing in Shh-MB*

There is an extensive network of fusion transcripts found in Shh-MB (Figure A4), likely to be tumor specific since fusions were filtered against a large library of gTEX and fetal cerebellum controls. Many of these fusions have no support in overlapping WGS, which might be due to the limited sample size and/or sensitivity of short-read technologies. Notable were fusions in *RALGAPA2* and *GNAS* (Figure 2.7;Figure A6) which were hypothesized to result from trans-splicing. These genes are unique in that they have a large multitude of 5' partners even in the same tumor, which has not been reported in any other model. My study was focused on validating the presence of trans-splicing. Going forward, a number of steps need to be taken to decipher the

mechanism responsible for trans-splicing in Shh-MB and to test for a role for this process in tumorigenesis.

Initially, differences in expression can be checked between patients with a high number of *RALGAPA2* or *GNAS* chimeric reads compared to those with a low count (top $10^{th}$ versus bottom $10^{th}$ percentile) to find genes and/or pathways differentially regulated in patients with trans-splicing. Similarity, the proportion of chimeric reads can be correlated with the expression of all genes across Shh-MB to find significant hits. Despite the large number of detected 5' genes, there does seem to be a subset of highly recurrent trans-spliced genes suggesting some sort of common splicing signal. It is also likely that the process is spliceosome mediated due to the presence of a strong U12 splicing signal. Just like in the Drosophila *mod(mdg4)*, there could be specific motifs that bind and/or enhance spliceosome formation between *RALGAPA2*/*GNAS* and their respective 5' partners. To find possible binding motifs, the introns of recurrent 5' partners can be compared against each other to find conserved sequences[177]. It is also possible that trans-splicing is in part mediated by close spatial proximity of fusion partner pairs in the nucleus. Available Shh-MB Hi-C libraries can be used to test this hypothesis by calculating interaction scores between trans-spliced partners and comparing it to random permuted gene pairs. The most difficult question to answer is whether trans-splicing plays a role in Shh-MB tumorigenesis. Without a mechanism it's impossible to increase the amount of trans-splicing in *RALGAPA2* or *GNAS*. However, it may be possible to completely abolish it through genetic manipulation. This can be achieved with knockout of the intron immediately following the recurrent splice junction forming a large contiguous exon, thereby removing the need for splicing without altering the canonical gene transcript. Specifically for *RALGAPA2*, this would require using CRISPR/Cas9 to remove the intron between exons 37 and 38 in both alleles in the *PTCH1* Shh-MB NES line and then

measuring the difference in primary tumor growth (and *RALGAPA2* trans-splicing) compared to the control. The presence of trans-splicing would first need to be demonstrated in the NES Shh-MB cell line before attempting this approach.

### 4.1.3 Shh-MB metastatic tumors

#### *4.1.3.1 Study summary*

Metastatic dissemination yields a dismal prognosis for any medulloblastoma patient. Although Shh-MB patients are not as commonly metastatic as Group 3 and Group 4, it is still an important subject of study. Mouse models are essential to study metastasis since human biopsies of metastasis are rare. Multiple metastatic locations were sampled from a large cohort of Shh-MB SB mice. Human genomics studies have shown that potent cancer driver genes often have several spatially independent alterations due to convergent evolutionary pressure[178]. This model of convergent evolution was extended to SB mice to identify important, and potentially actionable, drivers in metastatic lesions. Using this method there was a large overlap in genes identified to those found using the gold standard gCIS analysis approach. Furthermore, there was a high degree of spatial heterogeneity between metastasis in the same animal, just like it was observed within medulloblastoma primary tumors[179]. *Crebbp* and *Lgals3* were shown to be prominent metastatic drivers and were knocked out in various different transgenic Shh-MB models. A *Math1* mediated *Crebbp* knockout was not shown to increase metastasis in *Ptch1* driven Shh-MB, likely due to the timing of the *Crebbp* knockout. In contrast, mice missing one or both alleles of *Lgals3* had significantly less metastasis of the spine in a SmoA1 model.

#### *4.1.3.2 Functional validation of Shh-MB metastasis drivers*

Both the *Crebbp* and *Lgals3* genes were chosen based on their recurrence and what is known in literature. There were many other drivers discovered and it would be beneficial to unbiasedly compare and validate their metastatic propensity. With the careful inspection of insertion profiles, it is clear that the majority of driver genes undergo loss-of-function in the metastasis. Therefore, it would be possible to use a CRISPRi library[180,181] of all significant gCIS and convergent metastatic drivers in a Shh-MB NES model. There are *PTCH1* and *MYCN* mutated versions of this model available, but the *PTCH1* model would be best since the SB model was also driven by *PTCH1* loss-of-function alterations. In order to fully represent the library of metastatic drivers in a CRISPRi screen, there needs to be sufficient representation of each set of CRISPR guides in the injected cell line. Millions of cells can be injected into the flank, whereas a maximum of 50,000 cells can be injected into the cerebellum of NSG mice without overflowing. This imposes a smaller theoretical maximum of driver gene guides that can be fully represented in the cerebellar injections. Ideally there is a need for two CRISPRi experiments. One using the entire driver CRISPRi guide library (n = 431) with cells injected into the flank of NSG mice. In the other, cells would be injected into the cerebellum with a small high confidence subset of driver genes targeted by the CRISPRi guide library. Although Shh-MB grows in the cerebellum, the flank experiment would still be relevant because MB circulating tumor cells (CTC) can disseminate through both the blood and cerebrospinal fluid[182]. Furthermore, cerebellar injections are not 100% accurate and often result in cells being injected into the cerebrospinal fluid, thus skipping the primary tumor extravasation process. In both experimental setups, the metastatic tumors can be collected at endpoint and profiled to see which metastatic drivers are most represented. One drawback with

this system is the inability to look at synergistic effects between multiple drivers since each cell in the CRISPRi would have an average of one gene knockout.

#### *4.1.3.3 Functional validation of Crebbp and Lgals3*

*Crebbp* was found to be a highly recurrent driver in both the metastatic and primary tumors of SB mice. Furthermore, there were insertions in *Ncoa3*, which codes for an important component of the *Crebbp* protein complex. Using the current mouse models there was no significant difference in metastasis when both *Ptch1* and *Crebbp* were targeted in *Math1* expressing cells, which mark granule cells and their precursors (Figure 3.6). The timing of the *CREBBP* deletion has been shown to play an important role in Shh-MB primary tumor progression in humans[163] and it is possible that such a early knockout would not lead to a higher propensity of metastasis. In future experiments, it would be better to first allow for primary tumor development and then knock out *Crebbp*. This can be achieved using a *Ptch1*;*Crebbp*(+/flox or flox/flox); *Math1*-GFP mouse along with a tamoxifen inducible CRE system. After formation of tumor (confirming by MRI), *Crebbp* can be knocked out through ingestion of tamoxifen. It is also possible that treatment in an established tumor with *CREBBP* inhibitors could push the mouse toward a more metastatic phenotype. Unfortunately current available CREBBP inhibitors don't readily cross the blood brain barrier[183]. Lastly, rather than relying on costly and time-consuming transgenic experiments, NES cells can be used with *Crebbp*(flox) and a tamoxifen inducible CRE system.

Galectins are a class of secreted lectins which contain a carbohydrate recognition domain. These proteins play numerous roles in development and homeostasis. *Lgals3* is a unique galectin in that it also contains another domain allowing it to interact with non-carbohydrate ligands. *Lgals3* is expressed throughout the developing CNS system, in the meninges, choroid plexus, as well as

cerebellar cortex microglial and astrocyte subpopulations[184]. It is not clear whether Lgals3 is expressed in meninges of children and adults. *Lgals3* is involved in a large number of normal processes including growth, adhesion, differentiation, cell-cycle, immune response, and apoptosis. In the context of metastasis, surface expression of *Lgals3* on tumor cells can mediate homotypic cell adhesion by binding to soluble complementary glycoconjugates[185] which allow it to bind and maneuver through endothelial cell layers into the circulation. *Lgals3* interactions also allow CTCs to dock within distal metastatic sites.

*Lgals3* has shown great potential as a metastasis driver in Shh-MB. In the *SmoA1* Shh-MB mouse model, *Lgals3* (+/-) mice have a lower metastatic burden in the spine (Figure 3.7i). In SB mice, there is a trend towards less metastasis in *Lgals3* homozygous knockout mice, but it is not significant (Figure 3.7j). The SB metastasis model has a large amount of tumor heterogeneity and is more aggressive compared to the *SmoA1* model (Figure 3.7e–f). Both factors may allow SB metastasis to easily escape the restriction imposed by *Lglas3* and utilize other pathways. More SB mice need to be studied to confidently assess the statistical significance of this data. It is possible that *Lgals3* knockout mice are not metastatic in the spine because *Lgals3* plays a role in the metastatic cell docking process along the spine meninges. The rate of extravasation does not seem to be changed in the *SmoA1* model since tumor bearing mice with/without *Lgals3* have a similar metastasis disease burden. It is unclear whether expression of *Lgals3* in the meninges (rather than in the CTCs) mediates metastasis since the mouse model used a whole-body (constitutive) knockout of *Lgals3*. Lastly, the use of a *Lgals3* inhibitor[186] can be explored with *SmoA1* or SB *Math1*-GFP mice as a means to inhibit spinal metastasis.

### 4.1.4 The difficult path to a cure

Despite the enormous investment of time and money into medulloblastoma research the path towards a cure is still fraught with challenges. Medulloblastoma is comprised of four distinct subgroups, each with different clinical, methylation, transcriptional, and mutational characteristics. These subgroups should not be considered as one disease, but rather, therapy should be tailored for each subgroup. This is especially important for Wnt patients who would benefit from less radiation than the current standard of care (which is detrimental in the developing brain)[76]. WHO has recognized the importance of MB subtypes[187], but their adaption is slow to follow in the medical community. More initiative is needed by clinicians to bring next generation technology into the clinic for diagnosis and patient stratification. NanoString, which measures expression of a selected gene panel, offers a cheap quick method to define subtype and is already used for clinical diagnosis in some centers[17]. More recently, methylation arrays have been used to classify brain cancers and subtypes, challenging the need for pathological characterization[18]. It is clear from previous chapters that there is enormous heterogeneity of somatic alterations within the same subgroup (Figure 2.8) and even within the same patient due to spatial heterogeneity[179]. In the approaching era of personalized medicine clinicians need to use subgroup and genomic profiling on multiple biopsies (to ensure gene targets are spatially ubiquitous) in order to make informed decisions on therapy. In Shh-MB progress has been made through the use of SMO inhibitors. Unfortunately, this therapy can only work for patients with alterations within or upstream of SMO[77–79], emphasizing the need for genetic profiling before treatment. It is also important to consider combination therapy to circumvent resistance, as is commonly observed in SMO. Metastatic dissemination presents its own challenges since its typically too dangerous to biopsy and is spatially heterogenous. More research needs to be done to characterize common

pathways and to determine how metastatic cells interact with the local niche in order to uncover rationale therapies.

Although, Medulloblastoma is a common pediatric brain tumor it is still a relatively rare cancer. In Canada there is an incidence of 4.82 per 1,000,000 resulting in about 181 medulloblastoma cases per year[188] which is in stark contrast to the ~26,900 women diagnosed with breast cancer each year. When further stratifying by subgroup and genetic targets it becomes impossible to conduct a clinical trail within a single center in a realistic amount of time. Therefore, it is extremely important that the brain tumor research community collaborate at an international scale – by doing so it becomes possible to recruit enough patients. There would be great benefit in the use of adaptive trail designs to allow opportunities to modify ongoing studies and integrate new hypothesises as data is accumulated and analyzed[189]. Due to the limited number of patients, it is also important to find the best possible therapeutic alternatives before going forward with trials. This is where pre-clinal models have great utility. Shh-MB has a number of transgenic and orthopedic mouse models which closely resemble their human counterparts. Most studies use Shh-MB models with monotherapy which does not allow assessment of survival compared to existing combinations of neurosurgery, radiotherapy and/or chemotherapy used in humans. Although more involved, studies using the 'mouse hospital' approach would allow for more confident selection of therapies in human trials. It is also essential that models focus on study and treatment of tumors in the brain, rather than the flank, due to the blood brain barrier which posses a unique challenge in the brain cancer field. Despite all these issues, there has been enormous progress in the last decade. As the field continues to move forward, more and more effort will be put towards overcoming these hurdles in search for a cure.

# References


1. Skowron, P., Ramaswamy, V. & Taylor, M. D. Genetic and molecular alterations across medulloblastoma subgroups. *J. Mol. Med.* (2015). doi:10.1007/s00109-015-1333-8
2. Pui, C.-H., Gajjar, A. J., Kane, J. R., Qaddoumi, I. a & Pappo, A. S. Challenging issues in pediatric oncology. *Nat. Rev. Clin. Oncol.* **8**, 540–549 (2011).
3. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer Statistics, 2015. *CA Cancer J Clin* **65**, 5–29 (2015).
4. Rutkowski, S. *et al.* Survival and prognostic factors of early childhood medulloblastoma: An international meta-analysis. *J. Clin. Oncol.* **28**, 4961–4968 (2010).
5. Ramaswamy, V. *et al.* Medulloblastoma subgroup-specific outcomes in irradiated children: who are the true high-risk patients? *Neuro. Oncol.* 1–7 (2015). doi:10.1093/neuonc/nou357
6. Ramaswamy, V. *et al.* Recurrence patterns across medulloblastoma subgroups: an integrated clinical and molecular analysis. *Lancet Oncol.* **14**, 1200–7 (2013).
7. Gajjar, A. *et al.* Risk-adapted craniospinal radiotherapy followed by high-dose chemotherapy and stem-cell rescue in children with newly diagnosed medulloblastoma (St Jude Medulloblastoma-96): long-term results from a prospective, multicentre trial. *Lancet Oncol.* **7**, 813–820 (2006).
8. Shih, D. J. H. *et al.* Cytogenetic prognostication within medulloblastoma subgroups. *J. Clin. Oncol.* **32**, 886–896 (2014).
9. Packer, R. J. *et al.* Phase III study of craniospinal radiation therapy followed by adjuvant chemotherapy for newly diagnosed average-risk medulloblastoma. *J. Clin. Oncol.* **24**, 4202–4208 (2006).
10. Mabbott, D. J. *et al.* Serial evaluation of academic and behavioral outcome after treatment with cranial radiation in childhood. *J. Clin. Oncol.* **23**, 2256–2263 (2005).
11. Spiegler, B. J., Bouffet, E., Greenberg, M. L., Rutka, J. T. & Mabbott, D. J. Change in neurocognitive functioning after treatment with cranial radiation in childhood. *J. Clin. Oncol.* **22**, 706–713 (2004).
12. Taylor, M. D. *et al.* Molecular subgroups of medulloblastoma: The current consensus. *Acta Neuropathol.* **123**, 465–472 (2012).
13. Cho, Y. J. *et al.* Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. *J. Clin. Oncol.* **29**, 1424–1430 (2011).
14. Wang, X. *et al.* Medulloblastoma subgroups remain stable across primary and metastatic compartments. *Acta Neuropathol.* (2015). doi:10.1007/s00401-015-1389-0
15. Vladoiu, M. C. *et al.* Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature* (2019). doi:10.1038/s41586-019-1158-7
16. Pietsch, T. *et al.* Prognostic significance of clinical, histopathological, and molecular characteristics of medulloblastomas in the prospective HIT2000 multicenter clinical trial cohort. *Acta Neuropathol.* **128**, 137–149 (2014).
17. Northcott, P. a *et al.* Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. *Acta Neuropathol.* **123**, 615–26 (2012).
18. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
19. Varley, J. M., Evans, D. G. & Birch, J. M. Li-Fraumeni syndrome--a molecular and clinical review. *Br. J. Cancer* **76**, 1–14 (1997).
20. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59–71 (2012).
21. Kool, M. *et al.* Genome Sequencing of SHH Medulloblastoma Predicts Genotype-Related Response to Smoothened Inhibition. *Cancer Cell* **25**, 393–405 (2014).
22. Zhukova, N. *et al.* Subgroup-specific prognostic implications of TP53 mutation in medulloblastoma. *J. Clin. Oncol.* **31**, 2927–2935 (2013).
23. Evans, D. G., Farndon, P. a, Burnell, L. D., Gattamaneni, H. R. & Birch, J. M. The incidence of Gorlin syndrome in 173 consecutive cases of medulloblastoma. *Br. J. Cancer* **64**, 959–961 (1991).
24. Farndon, P. a, Del Mastro, R. G., Evans, D. G. & Kilpatrick, M. W. Location of gene for Gorlin syndrome. *Lancet* **339**, 581–582 (1992).
25. Huang, H. *et al.* APC mutations in sporadic medulloblastomas. *Am. J. Pathol.* **156**, 433–7 (2000).
26. Hamilton, S. R. *et al.* The molecular basis of Turcot's syndrome. *N. Engl. J. Med.* **332**, 839–47 (1995).
27. Kushner, B. H. *et al.* Rubinstein–Taybi Syndrome Predisposing to Non-WNT, Non-SHH, Group 3 Medulloblastoma. *J. Clin. Oncol.* **14**, 1526–1531 (1996).

28. Waszak, S. M. *et al.* Spectrum and prevalence of genetic predisposition in medulloblastoma: a retrospective genetic study and prospective validation in a clinical trial cohort. *Lancet Oncol.* **19**, 785–798 (2018).
29. Ellison, D. W. *et al.* ??-catenin status predicts a favorable outcome in childhood medulloblastoma: The United Kingdom Children's Cancer Study Group Brain Tumour Committee. *J. Clin. Oncol.* **23**, 7951–7957 (2005).
30. Kool, M. *et al.* Molecular subgroups of medulloblastoma: An international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. *Acta Neuropathol.* **123**, 473–484 (2012).
31. Northcott, P. a *et al.* Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* **488**, 49–56 (2012).
32. Patapoutian, A. & Reichardt, L. F. Roles of Wnt proteins in neural development and maintenance. *Curr. Opin. Neurobiol.* **10**, 392–399 (2000).
33. Marino, S. Medulloblastoma: Developmental mechanisms out of control. *Trends Mol. Med.* **11**, 17–22 (2005).
34. Jones, D. T. W. *et al.* Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**, 100–105 (2012).
35. Robinson, G. *et al.* Novel mutations target distinct subgroups of medulloblastoma. *Nature* **488**, 43–48 (2012).
36. Pugh, T. J. *et al.* Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**, 106–110 (2012).
37. Gibson, P. *et al.* Subtypes of medulloblastoma have distinct developmental origins. *Nature* **468**, 1095–9 (2010).
38. Pöschl, J. *et al.* Genomic and transcriptomic analyses match medulloblastoma mouse models to their human counterparts. *Acta Neuropathol.* **128**, 123–136 (2014).
39. Northcott, P. a. *et al.* Medulloblastoma comprises four distinct molecular variants. *J. Clin. Oncol.* **29**, 1408–1414 (2011).
40. Hatten, M. E. & Roussel, M. F. Development and cancer of the cerebellum. *Trends Neurosci.* **34**, 134–142 (2011).
41. Northcott, P. a. *et al.* Pediatric and adult sonic hedgehog medulloblastomas are clinically and molecularly distinct. *Acta Neuropathol.* **122**, 231–240 (2011).
42. Remke, M. *et al.* TERT promoter mutations are highly recurrent in SHH subgroup medulloblastoma. *Acta Neuropathol.* **126**, 917–929 (2013).
43. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–61 (2013).
44. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–9 (2013).
45. Goodrich, L. V, Milenković, L., Higgins, K. M. & Scott, M. P. Altered neural cell fates and medulloblastoma in mouse patched mutants. *Science* **277**, 1109–13 (1997).
46. Uziel, T. *et al.* The tumor suppressors Ink4c and p53 collaborate independently with Patched to suppress medulloblastoma formation. *Genes Dev.* **19**, 2656–67 (2005).
47. Ayrault, O., Zindy, F., Rehg, J., Sherr, C. J. & Roussel, M. F. Two tumor suppressors, p27Kip1 and patched-1, collaborate to prevent medulloblastoma. *Mol. Cancer Res.* **7**, 33–40 (2009).
48. Wetmore, C., Eberhart, D. E. & Curran, T. Loss of p53 but not ARF accelerates medulloblastoma in mice heterozygous for patched. *Cancer Res.* **61**, 513–6 (2001).
49. Hatton, B. a. *et al.* The Smo/Smo model: Hedgehog-induced medulloblastoma with 90% incidence and leptomeningeal spread. *Cancer Res.* **68**, 1768–1776 (2008).
50. Hallahan, A. R. *et al.* The SmoA1 mouse model reveals that notch signaling is critical for the growth and survival of Sonic Hedgehog-induced medulloblastomas. *Cancer Res.* **64**, 7794–7800 (2004).
51. Grammel, D. *et al.* Sonic hedgehog-associated medullobla stoma arising from the cochlear nuclei of the brainstem. *Acta Neuropathol.* **123**, 601–614 (2012).
52. Yang, Z. J. *et al.* Medulloblastoma Can Be Initiated by Deletion of Patched in Lineage-Restricted Progenitors or Stem Cells. *Cancer Cell* **14**, 135–145 (2008).
53. Wu, X. *et al.* Clonal selection drives genetic divergence of metastatic medulloblastoma. *Nature* **482**, 529–33 (2012).
54. Dey, J. *et al.* MyoD is a tumor suppressor gene in medulloblastoma. *Cancer Res.* **73**, 6828–6837 (2013).
55. Genovesi, L. a *et al.* Sleeping Beauty mutagenesis in a mouse medulloblastoma model defines networks that discriminate between human molecular subgroups. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4325-34 (2013).
56. Northcott, P. a *et al.* Medulloblastomics: the end of the beginning. *Nat. Rev. Cancer* **12**, 818–34 (2012).

57. Tseng, Y.-Y. *et al.* PVT1 dependence in cancer with MYC copy-number increase. *Nature* (2014). doi:10.1038/nature13311
58. CARRAMUSA, L. *et al.* The PVT-1 Oncogene Is a Myc Protein Target That Is Overexpressed in Transformed Cells. *J. Cell. Physiol.* **213**, 440–444 (2007).
59. Bai, R. Y., Staedtke, V., Lidov, H. G., Eberhart, C. G. & Riggins, G. J. OTX2 represses myogenic and neuronal differentiation in medulloblastoma cells. *Cancer Res.* **72**, 5988–6001 (2012).
60. Bunt, J. *et al.* OTX2 sustains a bivalent-like state of OTX2-bound promoters in medulloblastoma by maintaining their H3K27me3 levels. *Acta Neuropathol.* **125**, 385–394 (2013).
61. Young, M. D. *et al.* ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.* **39**, 7415–7427 (2011).
62. Pei, Y. *et al.* An Animal Model of MYC-Driven Medulloblastoma. *Cancer Cell* **21**, 155–167 (2012).
63. Kawauchi, D. *et al.* A Mouse Model of the Most Aggressive Subgroup of Human Medulloblastoma. *Cancer Cell* **21**, 168–180 (2012).
64. Lee, A. *et al.* Isolation of neural stem cells from the postnatal cerebellum. *Nat. Neurosci.* **8**, 723–729 (2005).
65. Sengoku, T. & Yokoyama, S. Structural basis for histone H3 lys 27 demethylation by UTX/KDM6A. *Genes Dev.* **25**, 2266–2277 (2011).
66. Kim, E. & Song, J. J. Diverse ways to be specific: A novel zn-binding domain confers substrate specificity to UTX/KDM6a histone H3 lys 27 demethylase. *Genes Dev.* **25**, 223–2226 (2011).
67. Kooistra, S. M. & Helin, K. Molecular mechanisms and potential functions of histone demethylases. *Nat. Rev. Mol. Cell Biol.* **13**, 297–311 (2012).
68. Wan, O. W. & Chung, K. K. K. The role of alpha-synuclein oligomerization and aggregation in cellular and animal models of Parkinson's disease. *PLoS One* **7**, 1–14 (2012).
69. Dubuc, A. M. *et al.* Aberrant patterns of H3K4 and H3K27 histone lysine methylation occur across subgroups in medulloblastoma. *Acta Neuropathol.* **125**, 373–384 (2013).
70. Northcott, P. a *et al.* Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma. *Nat. Genet.* **41**, 465–472 (2009).
71. Ong, C.-T. & Corces, V. G. Enhancers: emerging roles in cell fate specification. *EMBO Rep.* **13**, 423–30 (2012).
72. Northcott, P. a. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* (2014). doi:10.1038/nature13379
73. Mumert, M. *et al.* Functional genomics identifies drivers of medulloblastoma dissemination. *Cancer Res.* **72**, 4944–53 (2012).
74. Jenkins, N. C. *et al.* Genetic drivers of metastatic dissemination in sonic hedgehog medulloblastoma. *Acta Neuropathol. Commun.* **2**, 85 (2014).
75. Snuderl, M. *et al.* Targeting placental growth factor/neuropilin 1 pathway inhibits growth and spread of medulloblastoma. *Cell* **152**, 1065–76 (2013).
76. Aldape, K. *et al.* Challenges to curing primary brain tumours. *Nat. Rev. Clin. Oncol.* **16**, (2019).
77. Rudin, C. M. *et al.* Treatment of medulloblastoma with hedgehog pathway inhibitor GDC-0449. *N. Engl. J. Med.* **361**, 1173–1178 (2009).
78. LoRusso, P. M. *et al.* Phase I Trial of Hedgehog Pathway Inhibitor Vismodegib (GDC-0449) in Patients with Refractory, Locally Advanced or Metastatic Solid Tumors. *Clin. Cancer Res.* **17**, 2502–2511 (2011).
79. Rodon, J. *et al.* A phase I, multicenter, open-label, first-in-human, dose-escalation study of the oral smoothened inhibitor Sonidegib (LDE225) in patients with advanced solid tumors. *Clin. Cancer Res.* **20**, 1900–9 (2014).
80. Long, J. *et al.* The BET Bromodomain Inhibitor I-BET151 Acts Downstream of Smoothened Protein to Abrogate the Growth of Hedgehog Protein-driven Cancers. *J. Biol. Chem.* **289**, 35494–35502 (2014).
81. Tang, Y. *et al.* Epigenetic targeting of Hedgehog pathway transcriptional output through BET bromodomain inhibition. *Nat. Med.* **20**, 732–40 (2014).
82. Akhurst, R. J. & Hata, A. Targeting the TGFβ signalling pathway in disease. *Nat. Rev. Drug Discov.* **11**, 790–811 (2012).
83. Brooks, T. a. & Hurley, L. H. Targeting MYC Expression through G-Quadruplexes. *Genes Cancer* **1**, 641–649 (2010).
84. Yin, X., Giap, C., Lazo, J. S. & Prochownik, E. V. Low molecular weight inhibitors of Myc-Max interaction and function. *Oncogene* **22**, 6151–6159 (2003).

85. Wang, H. *et al.* Improved low molecular weight Myc-Max inhibitors. *Mol. Cancer Ther.* **6**, 2399–2408 (2007).
86. Bandopadhayay, P. *et al.* BET bromodomain inhibition of MYC-amplified medulloblastoma. *Clin. Cancer Res.* **20**, 912–925 (2014).
87. Puissant, A. *et al.* Targeting MYCN in neuroblastoma by BET bromodomain inhibition. *Cancer Discov.* **3**, 309–323 (2013).
88. McCabe, M. T. *et al.* EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature* **492**, 108–112 (2012).
89. Plasterk, R. H. & Izsva, Z. Molecular Reconstruction of Sleeping Beauty , a Tc1 -like Transposon from Fish , and Its Transposition in Human Cells. **91**, 501–510 (1997).
90. Dupuy, A. J., Akagi, K., Largaespada, D. a, Copeland, N. G. & Jenkins, N. a. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**, 221–6 (2005).
91. Moriarity, B. S. *et al.* A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and metastasis. *Nat. Genet.* **47**, 615–624 (2015).
92. de Ridder, J., Uren, A., Kool, J., Reinders, M. & Wessels, L. Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS Comput. Biol.* **2**, e166 (2006).
93. Dupuy, A. J. *et al.* A modified sleeping beauty transposon system that can be used to model a wide variety of human cancers in mice. *Cancer Res.* **69**, 8150–6 (2009).
94. Brett, B. T. *et al.* Novel molecular and computational methods improve the accuracy of insertion site analysis in Sleeping Beauty-induced tumors. *PLoS One* **6**, e24668 (2011).
95. Koudijs, M. J. *et al.* High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors. *Genome Res.* **21**, 2181–2189 (2011).
96. Riordan, J. D. *et al.* Sequencing methods and datasets to improve functional interpretation of sleeping beauty mutagenesis screens. *BMC Genomics* **15**, 1–15 (2014).
97. Wang, X. *et al.* Medulloblastoma subgroups remain stable across primary and metastatic compartments. *Acta Neuropathol.* (2015). doi:10.1007/s00401-015-1389-0
98. Stucklin, A. S. G., Ramaswamy, V., Daniels, C. & Taylor, M. D. Review of molecular classification and treatment implications of pediatric brain tumors. *Curr. Opin. Pediatr.* **30**, 3–9 (2018).
99. Cavalli, F. M. G. *et al.* Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell* **31**, 737-754.e6 (2017).
100. Suzuki, H. *et al.* Recurrent non-coding U1-snRNA mutations drive cryptic splicing in Shh medulloblastoma. *Nature* **0**, (2019).
101. Sasaki, H., Nishizaki, Y., Hui, C., Nakafuku, M. & Kondoh, H. Regulation of Gli2 and Gli3 activities by an amino-terminal repression domain: implication of Gli2 and Gli3 as primary mediators of Shh signaling. *Development* **126**, 3915–24 (1999).
102. Niewiadomski, P. *et al.* Gli protein activity is controlled by multisite phosphorylation in vertebrate hedgehog signaling. *Cell Rep.* **6**, 168–181 (2014).
103. Zhang, Y. *et al.* Structural insight into the mutual recognition and regulation between Suppressor of Fused and Gli/Ci. *Nat. Commun.* **4**, 1–12 (2013).
104. Richards, M. W. *et al.* Structural basis of N-Myc binding by Aurora-A and its destabilization by kinase inhibitors. *Proc. Natl. Acad. Sci.* **113**, 13726–13731 (2016).
105. Adhikary, S. & Eilers, M. Transcriptional regulation and transformation by Myc proteins. *Nat. Rev. Mol. Cell Biol.* **6**, 635–645 (2005).
106. Farrell, A. S. & Sears, R. C. MYC degradation. *Cold Spring Harb. Perspect. Med.* **4**, 1–15 (2014).
107. Welcker, M. & Clurman, B. E. FBW7 ubiquitin ligase: a tumour suppressor at the crossroads of cell division, growth and differentiation. *Nat. Rev. Cancer* **8**, 83–93 (2008).
108. Thompson, B. J. *et al.* The SCF FBW7 ubiquitin ligase complex as a tumor suppressor in T cell leukemia. *J. Exp. Med.* **204**, 1825–1835 (2007).
109. O'Neil, J. *et al.* FBW7 mutations in leukemic cells mediate NOTCH pathway activation and resistance to γ-secretase inhibitors. *J. Exp. Med.* **204**, 1813–1824 (2007).
110. Close, V. *et al.* FBXW7 mutations reduce binding of NOTCH1, leading to cleaved NOTCH1 accumulation and target gene activation in CLL. *Blood* **133**, 830–839 (2019).
111. Oghabi Bakhshaiesh, T., Majidzadeh-A, K. & Esmaeili, R. Wip1: A candidate phosphatase for cancer diagnosis and treatment. *DNA Repair (Amst).* **54**, 63–66 (2017).
112. Kleiblova, P. *et al.* Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint. *J.*

*Cell Biol.* **201**, 511–521 (2013).
113. Zajkowicz, A. *et al.* Truncating mutations of PPM1D are found in blood DNA samples of lung cancer patients. *Br. J. Cancer* **112**, 1114–1120 (2015).
114. Zhang, L. *et al.* Exome sequencing identifies somatic gain-of-function PPM1D mutations in brainstem gliomas. *Nat. Genet.* **46**, 726–730 (2014).
115. He, X. *et al.* The G protein α subunit $G\alpha_s$ is a tumor suppressor in Sonic hedgehog−driven medulloblastoma. *Nat. Med.* **20**, 1035–1042 (2014).
116. Haas, B. J. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* (2017). doi:10.1101/120295
117. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
118. Okonechnikov, K. *et al.* InFusion: Advancing Discovery of Fusion Genes and Chimeric Transcripts from Deep RNA-Sequencing Data. *PLoS One* **11**, e0167417 (2016).
119. Ratnaparkhe, M. *et al.* Defective DNA damage repair leads to frequent catastrophic genomic events in murine and human tumors. *Nat. Commun.* **9**, 4760 (2018).
120. Jepsen, K. *et al.* Combinatorial roles of the nuclear receptor corepressor in transcription and development. *Cell* **102**, 753–763 (2000).
121. Hermanson, O., Jepsen, K. & Rosenfeld, M. G. N-CoR controls differentiation of neural stem cells into astrocytes. *Nature* **419**, 934–939 (2002).
122. Houseley, J. & Tollervey, D. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One* **5**, (2010).
123. Fonseca, N. A. *et al.* Pan-cancer study of heterogeneous RNA aberrations. *bioRxiv* 183889 (2017). doi:10.1101/183889
124. Huang, M. *et al.* Engineering Genetic Predisposition in Human Neuroepithelial Stem Cells Recapitulates Medulloblastoma Tumorigenesis. *Cell Stem Cell* 1–14 (2019). doi:10.1016/j.stem.2019.05.013
125. Merk, D. J. *et al.* Opposing Effects of CREBBP Mutations Govern the Phenotype of Rubinstein-Taybi Syndrome and Adult SHH Medulloblastoma. *Dev. Cell* **44**, 709-724.e6 (2018).
126. Cedoz, P. L., Prunello, M., Brennan, K. & Gevaert, O. MethylMix 2.0: An R package for identifying DNA methylation genes. *Bioinformatics* **34**, 3044–3046 (2018).
127. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
128. Morrissy, A. S. *et al.* Divergent clonal selection dominates medulloblastoma at recurrence. *Nature* **529**, 351–357 (2016).
129. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–7 (2014).
130. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
131. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
132. Ramaswami, G. & Li, J. B. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* **42**, D109–D113 (2014).
133. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
134. Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89–e89 (2013).
135. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
136. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
137. Loo, P. Van *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910–16915 (2010).
138. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
139. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
140. Wu, Z. & Wu, H. *Visualizing Genomic Data Using Gviz and Bioconductor*. *Methods in Molecular Biology* **1418**, (2016).

141. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* **6**, 31602 (2016).
142. Wagih, O. Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
143. Kataoka, K. *et al.* Aberrant PD-L1 expression through 3′-UTR disruption in multiple cancers. *Nature* **534**, (2016).
144. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. **28**, 333–339 (2012).
145. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
146. Yang, J. & Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).
147. Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: Improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.* **45**, W291–W299 (2017).
148. Canisius, S., Martens, J. W. M. & Wessels, L. F. A. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol.* **17**, 1–17 (2016).
149. Reimand, J. *et al.* g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).
150. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).
151. Paul Shannon, 1 *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 6 (2003).
152. Sturm, D. *et al.* Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. *Cancer Cell* **22**, 425–437 (2012).
153. Hovestadt, V. *et al.* Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. *Acta Neuropathol.* **125**, 913–916 (2013).
154. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
155. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
156. Zhou, X. *et al.* Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat. Genet.* **48**, 4–6 (2015).
157. Connors, J. *et al.* Circos : An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
158. Phi, J. H. *et al.* Cerebrospinal fluid M staging for medulloblastoma: Reappraisal of Chang's M staging based on the CSF flow. *Neuro. Oncol.* **13**, 334–344 (2011).
159. Cancer Genome Atlas Research Network *et al.* Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N. Engl. J. Med.* **374**, 135–45 (2016).
160. Zhang, J. *et al.* The CREBBP Acetyltransferase Is a Haploinsufficient Tumor Suppressor in B-cell Lymphoma. (2017). doi:10.1158/2159-8290.CD-16-1417
161. Jaber, B. M., Mukopadhyay, R. & Smith, C. L. Estrogen receptor- interaction with the CREB binding protein coactivator is regulated by the cellular environment. 307–323 (1998).
162. Levin, M. L. *et al.* AIB1 , a Steroid Receptor Coactivator Amplified in Breast and Ovarian Cancer binds to the transcriptional integrators. (1988).
163. Medulloblastoma, S. H. H. *et al.* Opposing Effects of CREBBP Mutations Govern the Phenotype of Rubinstein-Taybi Syndrome and Adult Article Opposing Effects of CREBBP Mutations Govern the Phenotype of Rubinstein-Taybi Syndrome and Adult SHH Medulloblastoma. 709–724 (2018). doi:10.1016/j.devcel.2018.02.012
164. Newlaczyl, A. U. & Yu, L. Galectin-3 – A jack-of-all-trades in cancer. *Cancer Lett.* **313**, 123–128 (2011).
165. Rivero-Hinojosa, S. *et al.* Proteomic analysis of Medulloblastoma reveals functional biology with translational potential. *Acta Neuropathol. Commun.* **6**, 48 (2018).
166. Sidiropoulos, N. *et al.* Article The whole-genome landscape of medulloblastoma subtypes. *Nat. Publ. Gr.* **547**,

311–317 (2017).
167. Ivanov, D. P., Coyle, B., Walker, D. A. & Grabowska, A. M. In vitro models of medulloblastoma: Choosing the right tool for the job. *J. Biotechnol.* **236**, 10–25 (2016).
168. Pandolfi, S. & Stecca, B. Luciferase Reporter Assays to Study Transcriptional Activity of Hedgehog Signaling in Normal and Cancer Cells. in (ed. Wang, K.) **1224**, 71–79 (Springer New York, 2015).
169. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
170. Chwalenia, K., Facemire, L. & Li, H. Chimeric RNAs in cancer and normal physiology. *Wiley Interdiscip. Rev. RNA* **8**, (2017).
171. Lei, Q. *et al.* Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biol. Evol.* **8**, 562–577 (2016).
172. Gao, J. L. *et al.* A conserved intronic U1 snRNP-binding sequence promotes trans-splicing in Drosophila. *Genes Dev.* **29**, 760–771 (2015).
173. Li, H., Wang, J., Mor, G. & Sklar, J. A Neoplastic Gene Fusion Mimics Trans-Splicing of RNAs in Normal Human Cells. *Science (80-. ).* **321**, 1357–1361 (2008).
174. Rickman, D. S. *et al.* SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer Res.* **69**, 2737–2738 (2009).
175. Janz, S., Potter, M. & Rabkin, C. S. Lymphoma- and leukemia-associated chromosomal translocations in healthy individuals. *Genes Chromosom. Cancer* **36**, 211–223 (2003).
176. Frenkel-Morgenstern, M. *et al.* ChiTaRS: A database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res.* **41**, 142–151 (2013).
177. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
178. Gerlinger, M. & Rowan, A. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *… Engl. J. …* (2012).
179. Morrissy, A. S. *et al.* Spatial heterogeneity in medulloblastoma. *Nat. Genet.* **49**, 780–788 (2017).
180. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16**, 299–311 (2015).
181. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8**, 2180–2196 (2013).
182. Garzia, L. *et al.* A Hematogenous Route for Medulloblastoma Leptomeningeal Metastases. *Cell* **172**, 1050-1062.e14 (2018).
183. Lasko, L. M. *et al.* Discovery of a selective catalytic p300/CBP inhibitor that targets lineage-specific tumours. *Nature* **550**, 128–132 (2017).
184. Selimi, F. Expression and role of Galectin-3 in the postnatal development of the cerebellum. *bioRxiv* (2018). doi:10.1101/364760
185. Ahmed, H. & Alsadek, D. M. M. Galectin-3 as a Potential Target to Prevent Cancer Metastasis. *Clin. Med. Insights Oncol.* **9**, CMO.S29462 (2015).
186. Blanchard, H., Yu, X., Collins, P. M. & Bum-Erdene, K. Galectin-3 inhibitors: A patent review (2008-present). *Expert Opin. Ther. Pat.* **24**, 1053–1065 (2014).
187. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
188. Johnston, D. L. *et al.* Incidence of medulloblastoma in Canadian children. *J. Neurooncol.* **120**, 575–579 (2014).
189. Pallmann, P. *et al.* Adaptive designs in clinical trials: Why use them, and how to run and report them. *BMC Med.* **16**, 1–15 (2018).
190. Temiz, N. A. *et al.* RNA-sequencing of Sleeping Beauty Transposon Induced Tumors Detects Transposon-RNA Fusions Allowing Precision Analyses of Forward Genetic Cancer Screens. *Genome Res.* 1–11 (2015). doi:10.1101/gr.188649.114.7

# Appendix

## THE TRANSCRIPTIONAL LANDSCAPE OF SONIC HEDGEHOG MEDULLOBLASTOMA

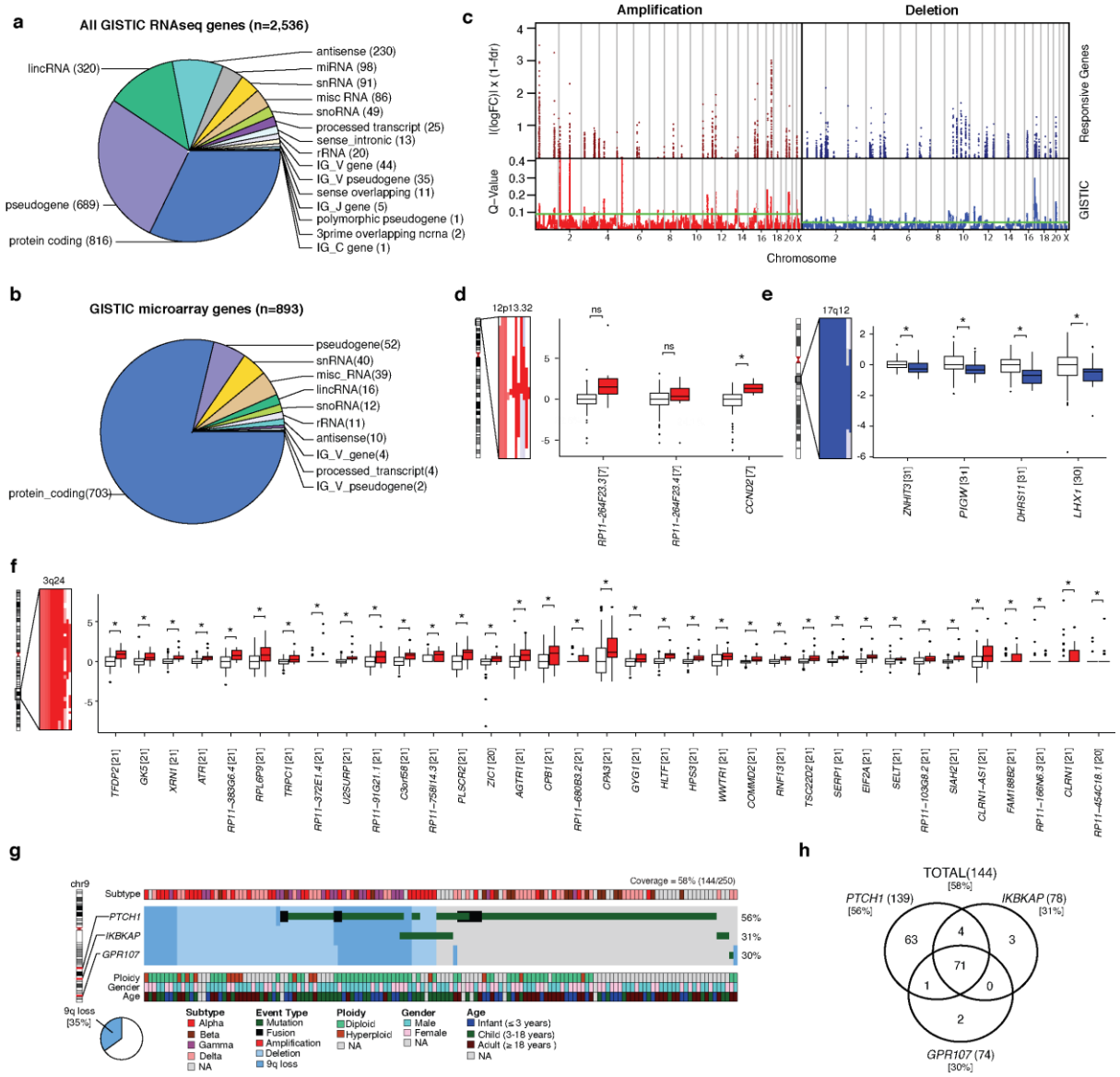



**Figure A1 Copy number responsive Genes in GISTIC regions**

**(a, b)** Gene type distribution of all genes found within GISTIC regions using **(a)** RNA-seq and **(b)** within the subset of genes found only on microarrays. **(c)** Distribution of significant copy number responsive genes (top) GISTIC peaks (bottom). Scores are calculated based on the absolute value of the fold change in GISTIC region expression and copy number responsive analysis FDR adjusted p-value. The threshold for GISTIC significant peaks is shown as a horizontal line. **(d-f)** Expression difference between copy number neutral and aberrant states in GISTIC regions. **(c)** All recurrently amplified genes in 12p13.32. **(e)** Copy number responsive genes in the 17q12 deletion GISTIC region. **(f)** Copy number responsive genes in 3q24 recurrently gained GISTIC region. Genes annotated with an asterisk are copy number responsive with (Kruskal-Wallis adjusted p-value < 0.1). The GISTIC region copy number segments are shown to the left of each graph. Numbers in square brackets denote number of tumors detected with the CNA. Expression of each gene was normalized by the expression median of the neutral copy number state.

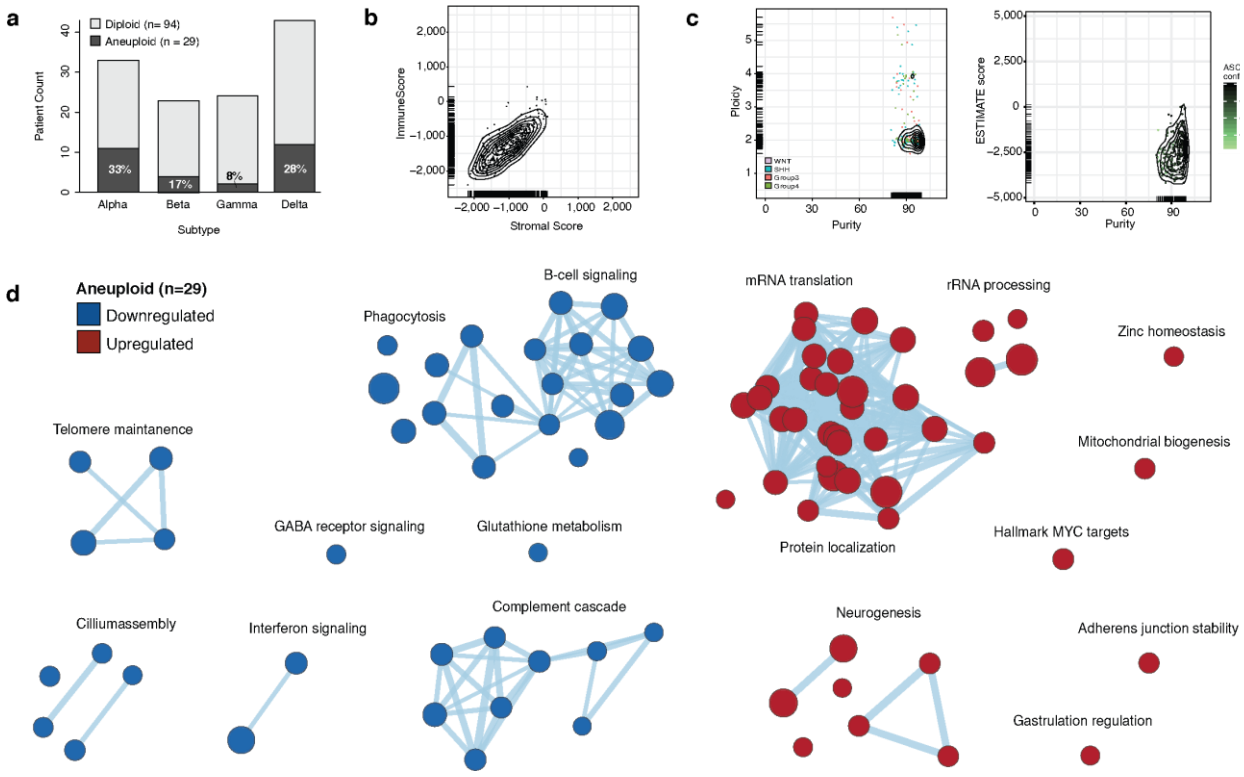



**Figure A2 Transcriptional landscape of aneuploid tumors**

**(a)** Number of diploid and aneuploid tumors across Shh-MB subtypes. **(b)** Tumor Purity assessment using RNA-seq ESTIMATE SNP6 Stromal and Immune scores. Scores below zero denote higher tumor purity. **(c)** SNP 6.0 calculated purity compared to ploidy (left) and RNA-seq derived ESTIMATE scores (right). **(d)** GSEA enrichment map of genes differentially expressed in aneuploid (n = 29) compared to diploid (n = 94) Shh-MB tumors (FDR q-value <0.01). Node size is proportional to the number of genes and edge weight represents the number of shared genes between each gene set. The color represents where the pathway was found to be overexpressed; either diploid (blue) or aneuploid tumors (red).

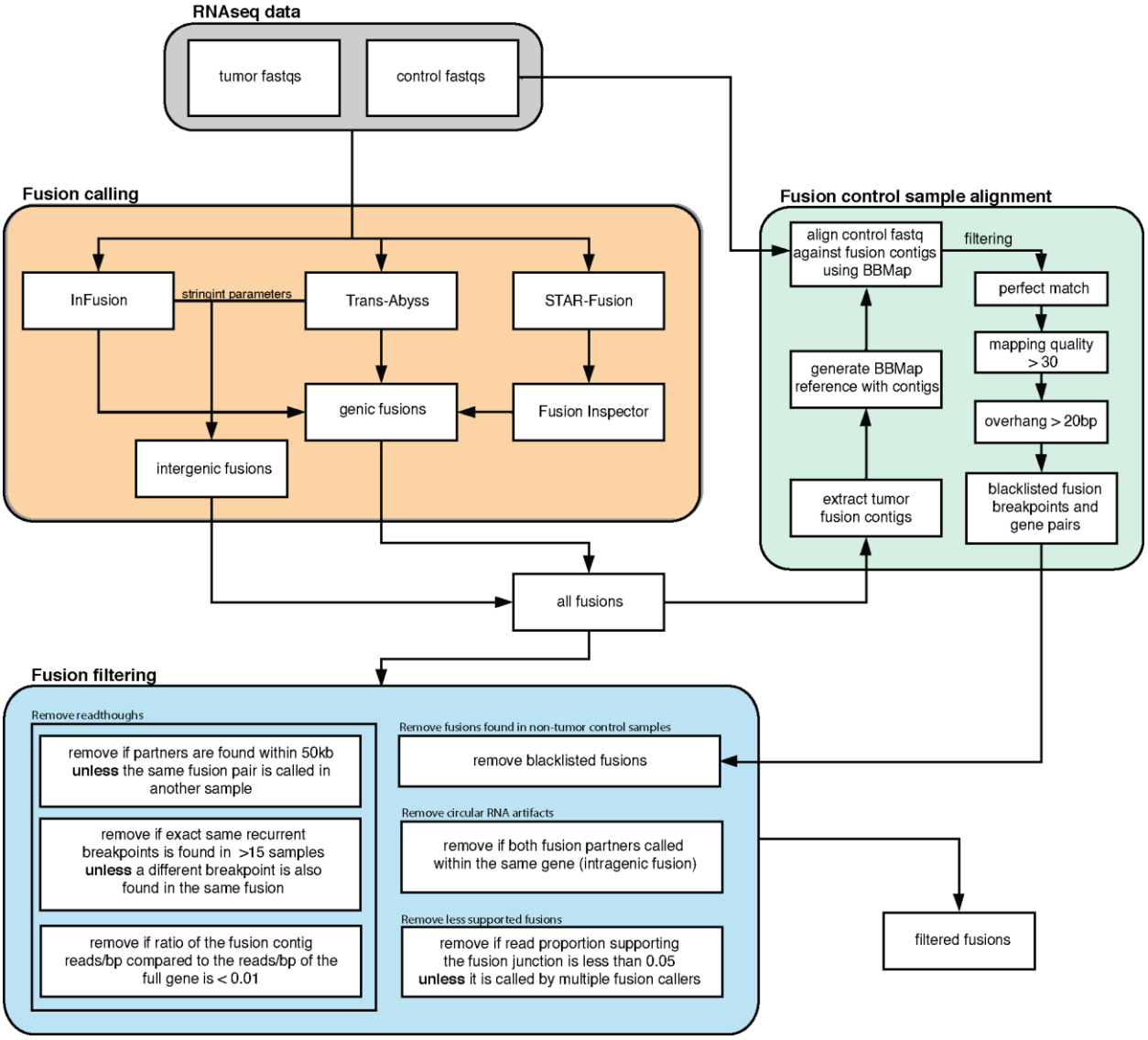



**Figure A3 Fusion calling overview**
Flowchart of the fusion calling method (n = 250) using InFusion, STAR-Fusion and Trans-ABYSS. Fusion contigs matching reads in the GTEx and Biotech control samples (n = 51) were filtered as well as any read through, subclonal, and intragenic fusions.

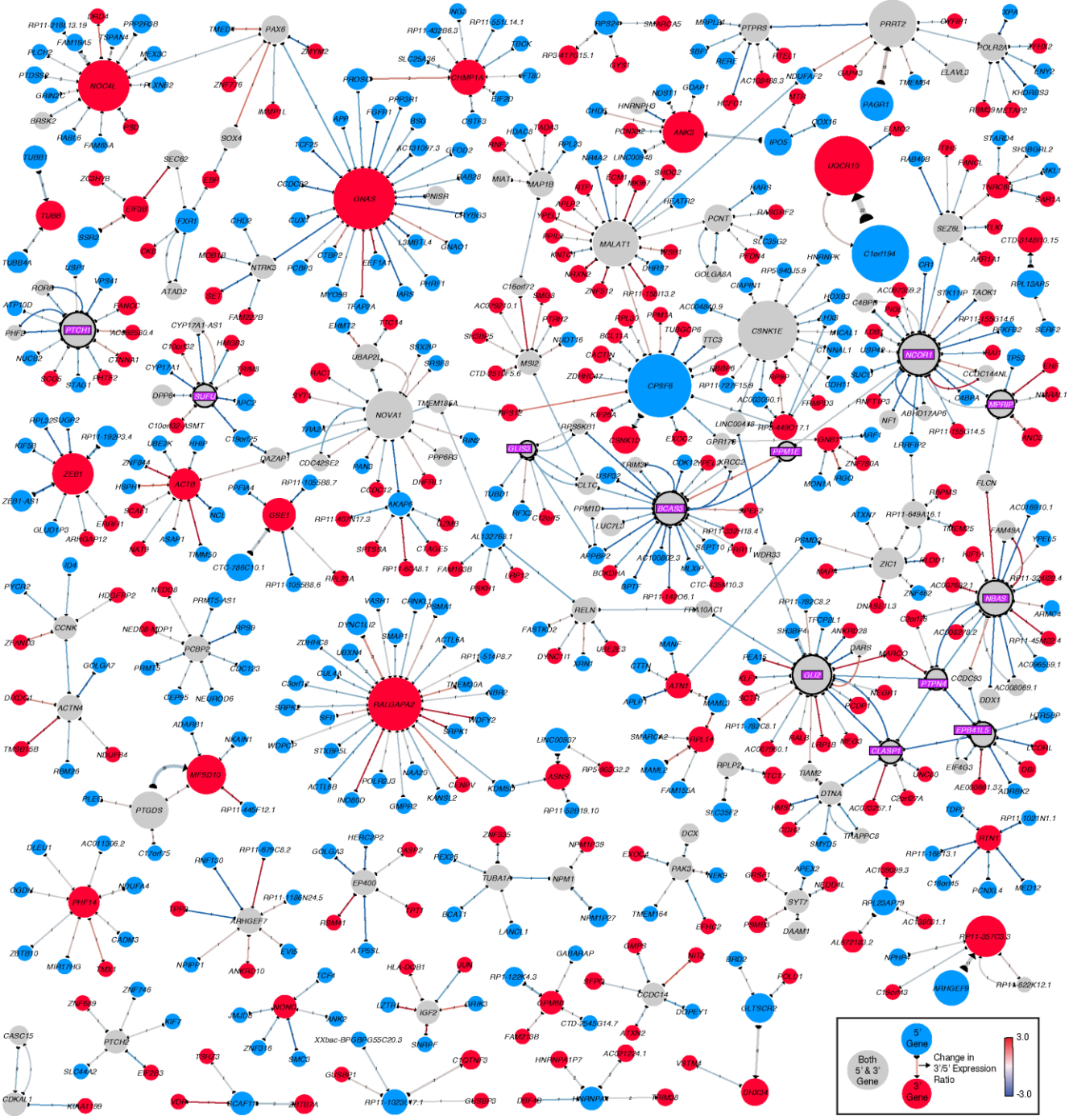



**Figure A4 Fusion landscape**

Exon-exon fusion network in Shh-MB. The color represents orientation of the gene (5’ is blue , 3’ is red, and both is grey), while the size of the node is proportional to the recurrence of the gene. The color of the line shows fold change difference in gene 3’/5’ expression ratio fusion positive compared to negative patients, while line thickness is proportional to the recurrence. Fusion hubs supported by structural variants are encircled with highlighted gene names.

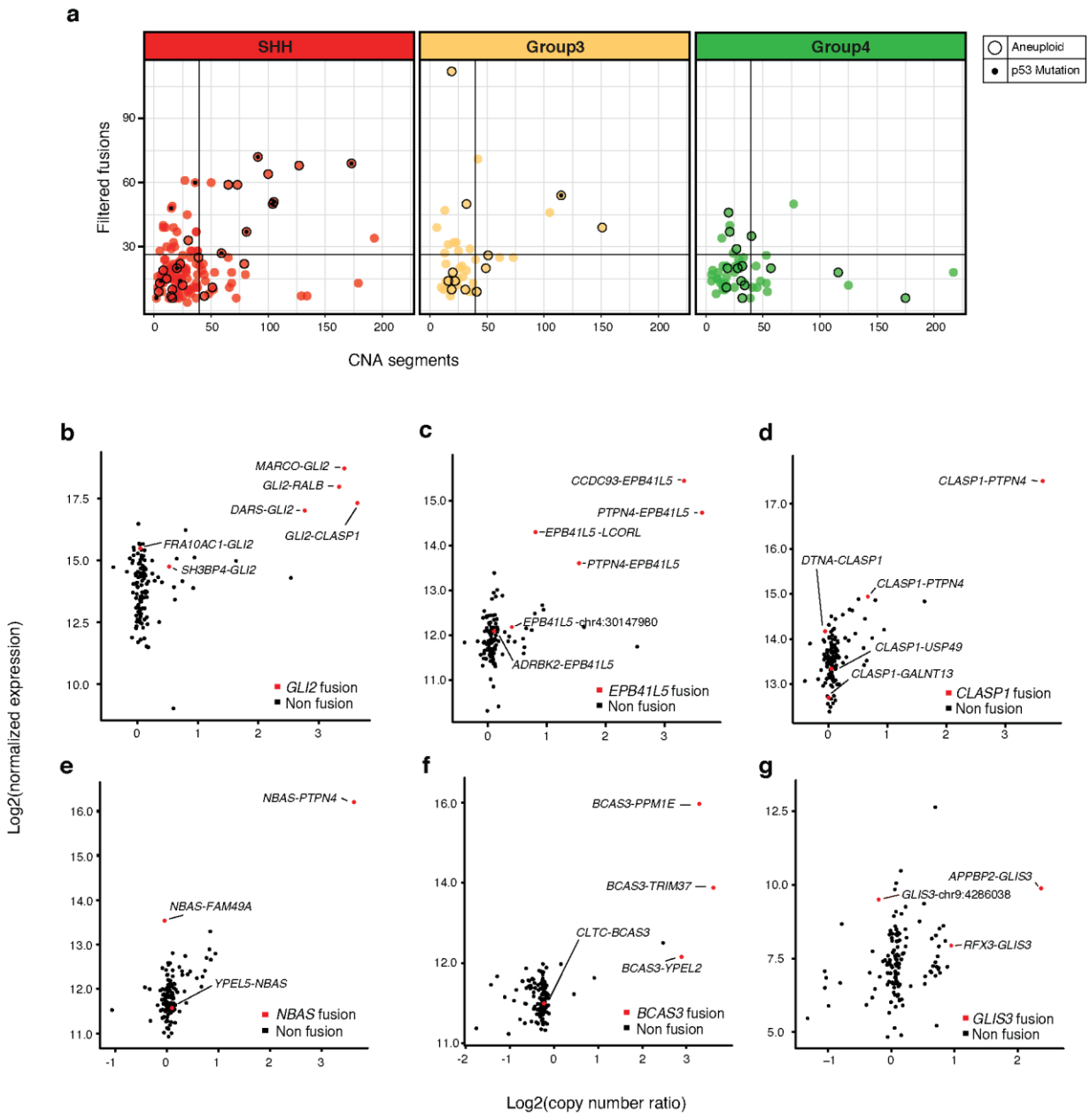



**Figure A5 Copy number alterations in fusion hubs**

**(a)** Correlation between the number of copy number segments and the number of fusions in Shh-MB (n = 250), Group 3-MB (n= 56), and Group 4-MB (n= 61). Aneuploidy status is shown across all subgroups. p53 mutation status is only known for Shh-MB tumors. **(b-e)** Correlation between expression and copy number for **(b)** *GLI2*, **(c)** *NBAS*, **(d)** *BCAS3*, and **(e)** *EPB41L*. Fusion patients are indicated in red. The fusion with the most read support in its respective patient is shown.

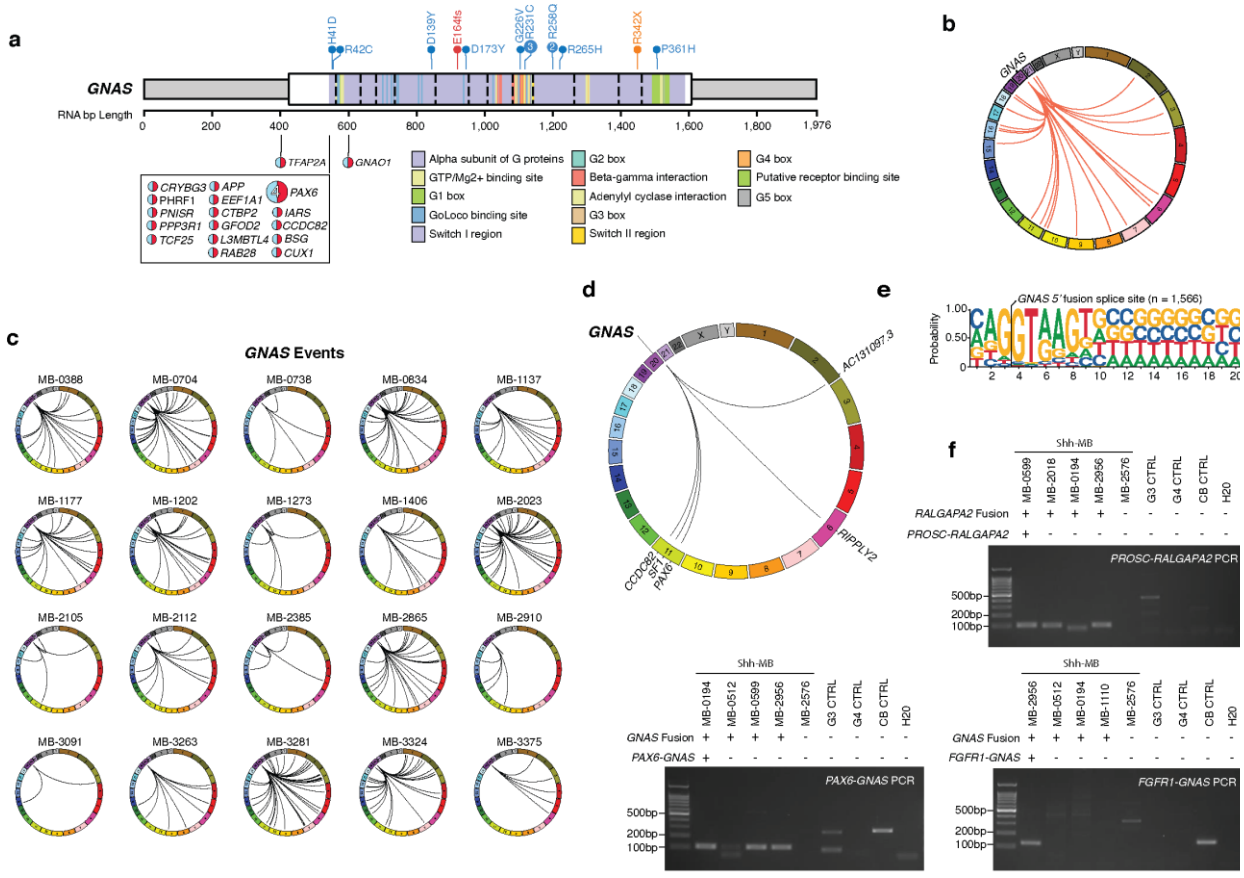



**Figure A6 Promiscuous recurrent GNAS chimeric transcript breakpoints**

**(a, b)** Gene-level summary of **(a)** *GNAS* fusions detected by fusion-callers, and **(b)** their distribution across the genome. Refer to Extended Data Fig. 6b for schema description. **(c, d)** Distribution of *GNAS* exon 1 chimeric junction spanning reads across the genome with **(d)** genes found in >10 samples indicated. Chimeric reads were extracted from STAR alignments. **(e)** Splice site consensus sequence of *GNAS* 5' chimeric fusion partner transcripts (n = 1,566). **(f)** PCR validation of *GNAS* fusion RNA transcripts in human Shh-MB samples with and without detected fusions (by RNA-seq) compared to Group 3-MB, Group 4-MB and normal cerebellar controls. Patients with any detected chimeric transcripts at exon 1 in *GNAS* (by RNA-seq) are indicated as *GNAS* fusion positive (+).

**a**

| Sample Name | Fusion Hub | Partner |
|---|---|---|
| MB-2018 | *GNAS* | *CRELD1* |
| | *GNAS* | *EVAVL2* |
| | *GNAS* | *PMPCB* |
| | *RALGAPA2* | *SPTAN1* |
| MB-2023 | *GNAS* | *RPS25* |
| MB-2054 | *GNAS* | *APCDD1* |
| | *GNAS* | *STAU2* |
| MB-2056 | *GNAS* | *SPECC1L* |
| | *GNAS* | *MLDCD* |
| MB-2971 | *GNAS* | *CYFIP1* |
| | *GNAS* | *APP* |
| MB-3023 | *GNAS* | *CEP350* |
| | *GNAS* | *KIAA1841* |
| | *GNAS* | *RBM10* |
| | *GNAS* | *SCARB2* |
| | *RALGAPA2* | *HERC2P(4,5,8)* |
| MB-3090 | *GNAS* | *MAFIP* |
| MB-3202 | *GNAS* | *KLRD1* |
| | *RALGAPA2* | *SNRNP200* |

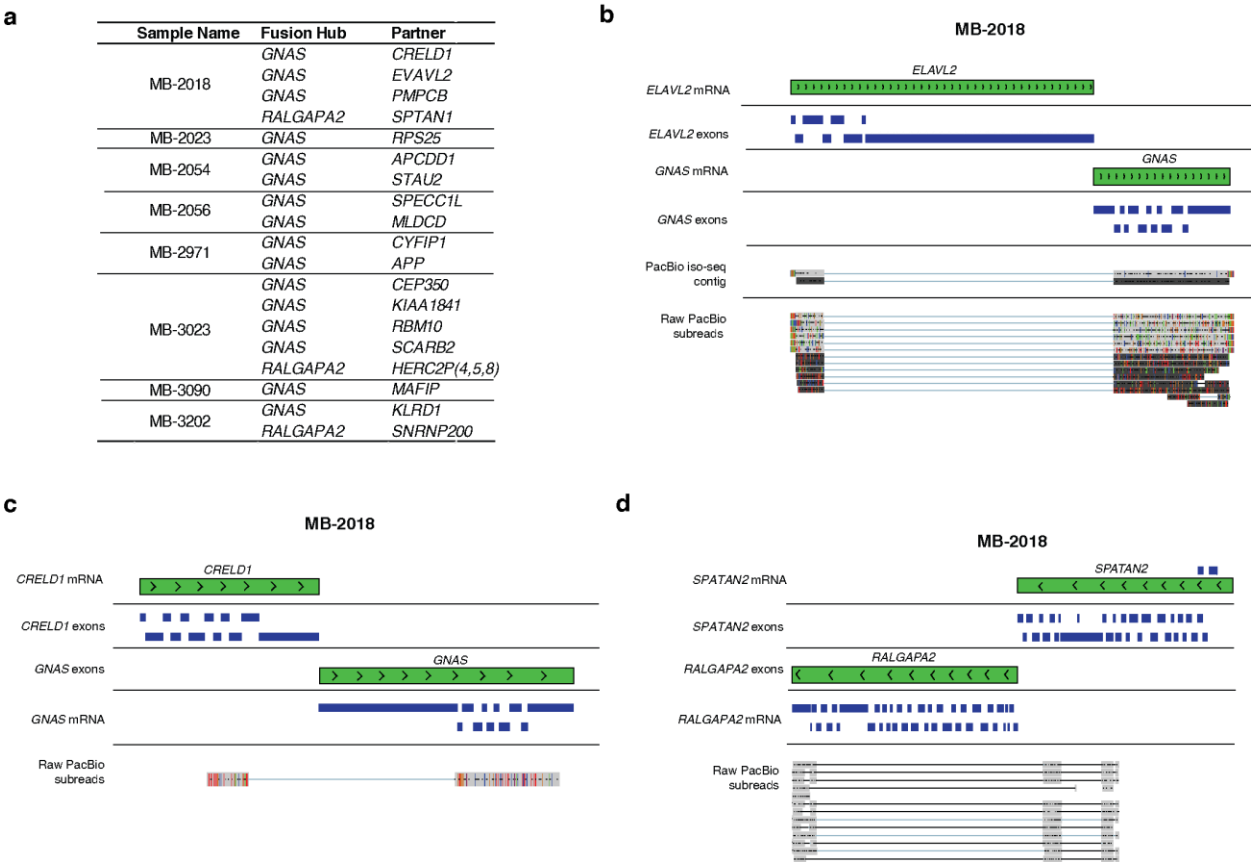



**Figure A7 Pacbio IsoSeq Validations**
**(a)** Table of *GNAS* and *RALGAPA2* fusions validated using Pacbio Iso-seq RNA long read sequencing. **(b, c)** Pacbio Iso-seq RNA long read sequencing validated *GNAS* fusion. **(d)** PacBio Iso-Seq RNA long read sequencing validated *RALGAPA2* fusions.

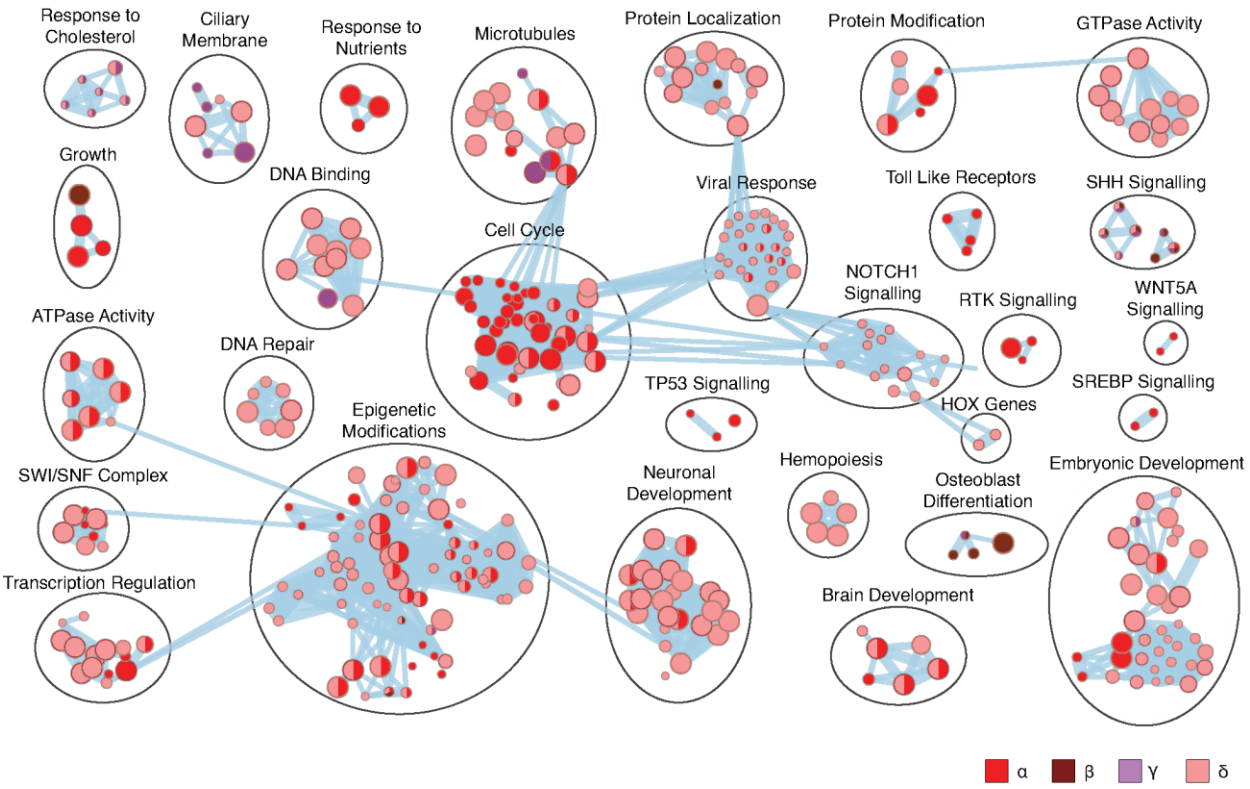



**Figure A8 Shh-MB oncogenic pathways**

Enrichment map of biological processes and pathways affected by mutation or focal amplifications/deletions in Shh-MB subtypes. Each node represents a pathway or process and connecting lines represent common genes between them. Nodes with many shared genes are grouped together and labeled with a biological theme. The color of the nodes refers to the subtype(s) in which the process is enriched. The size of the node is proportional to the number of genes in process. Enriched processes were determined with g:Profiler (FDR-corrected q-value < 0.05) and visualized with the Enrichment Map app in Cytoscape.

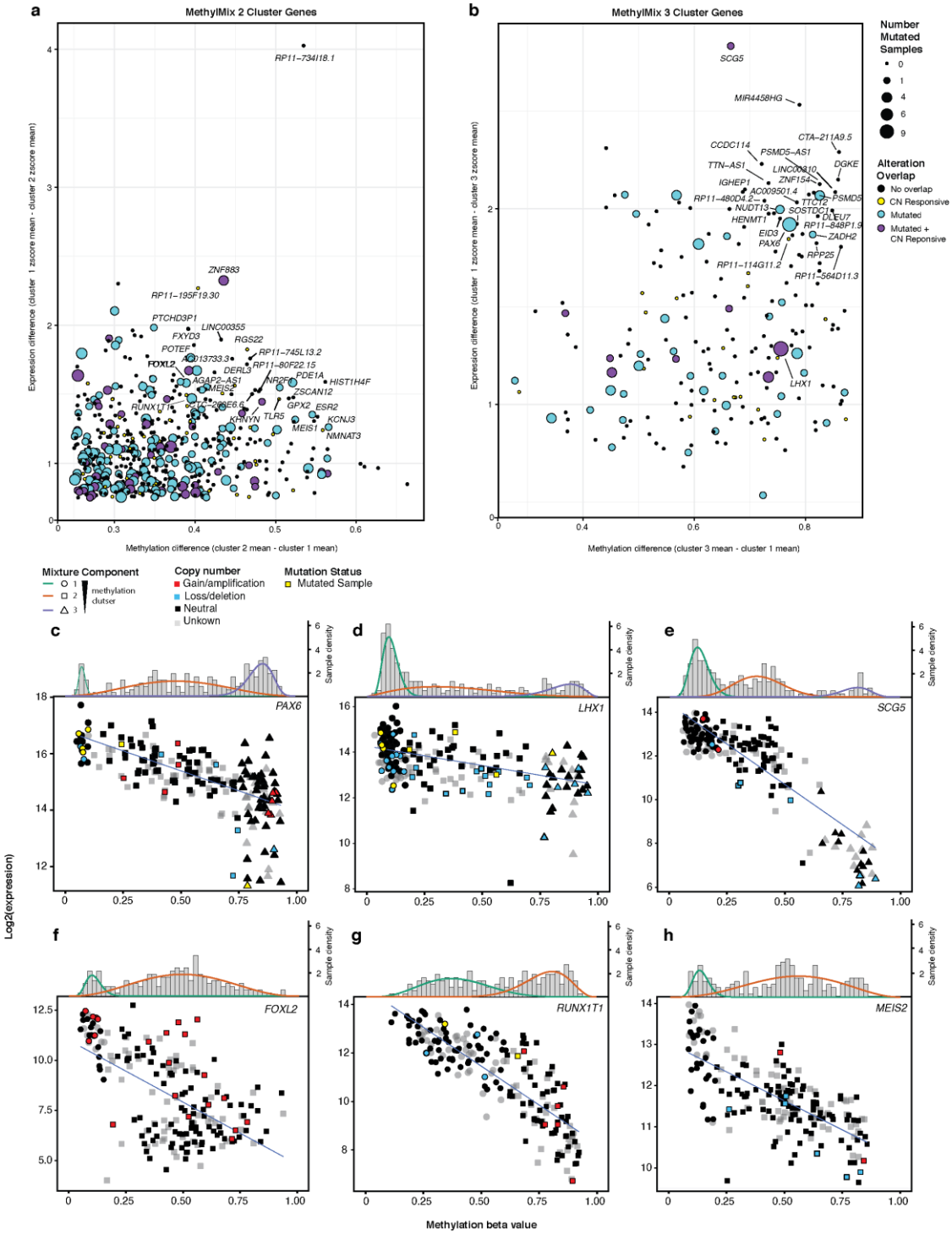



**Figure A9 DNA methylation anticorrelated with change in gene expression across Shh-MB**

**(a, b)** Scatterplot showing the mean difference in expression and methylation between samples with **(a)** two methylation clusters discovered by MethylMix**,** and **(b)** between samples from 3 methylation clusters **(b)**. The size of the points is proportional to the number of patients with mutations found in the corresponding gene. Points are colored by their overlap in mutation events and if they were found to be copy number responsive. **(c-i)** Correlation of gene expression and DNA methylation in genes identified by MethylMix. The methylation clusters are highlighted in a histogram above each scatterplot and are represented by different shapes in the bottom plot. The point border and fill colors correspond to the copy number and mutation state of the given gene, respectively for each Shh-MB sample.

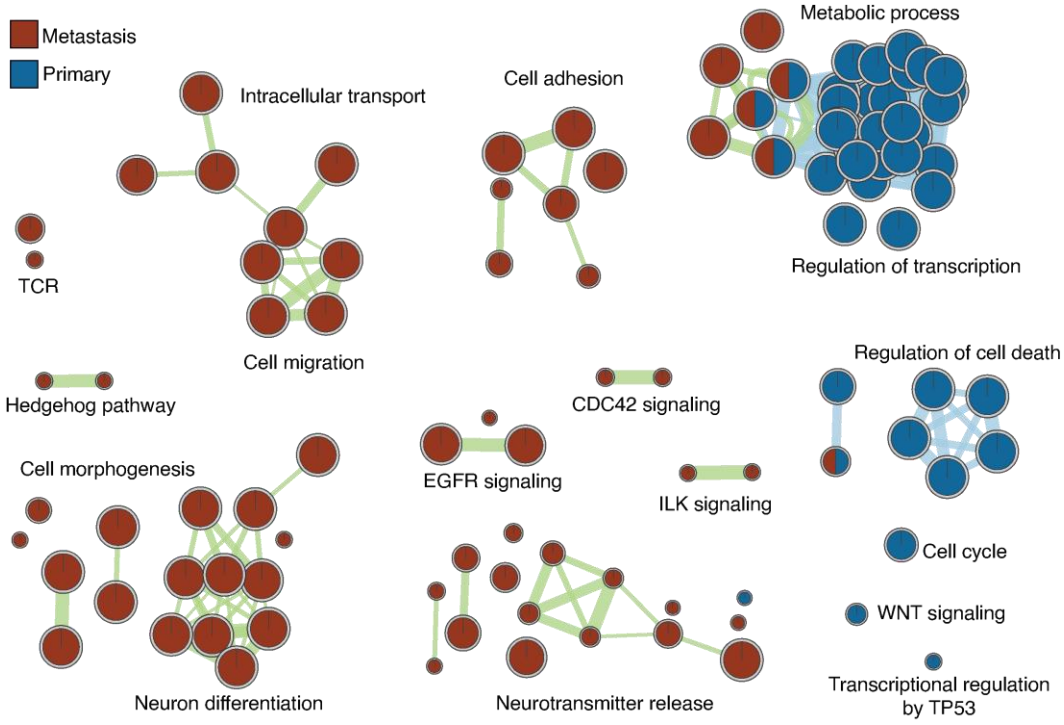



**Figure A10 Primary and metastatic oncogenic pathways**
GSEA enrichment map of primary (blue) and metastatic (red) driver genes in Shh-MB SB model (FDR q-value <0.01). Node size is proportional to the number of genes and edge weight represents the number of shared genes between each gene set.

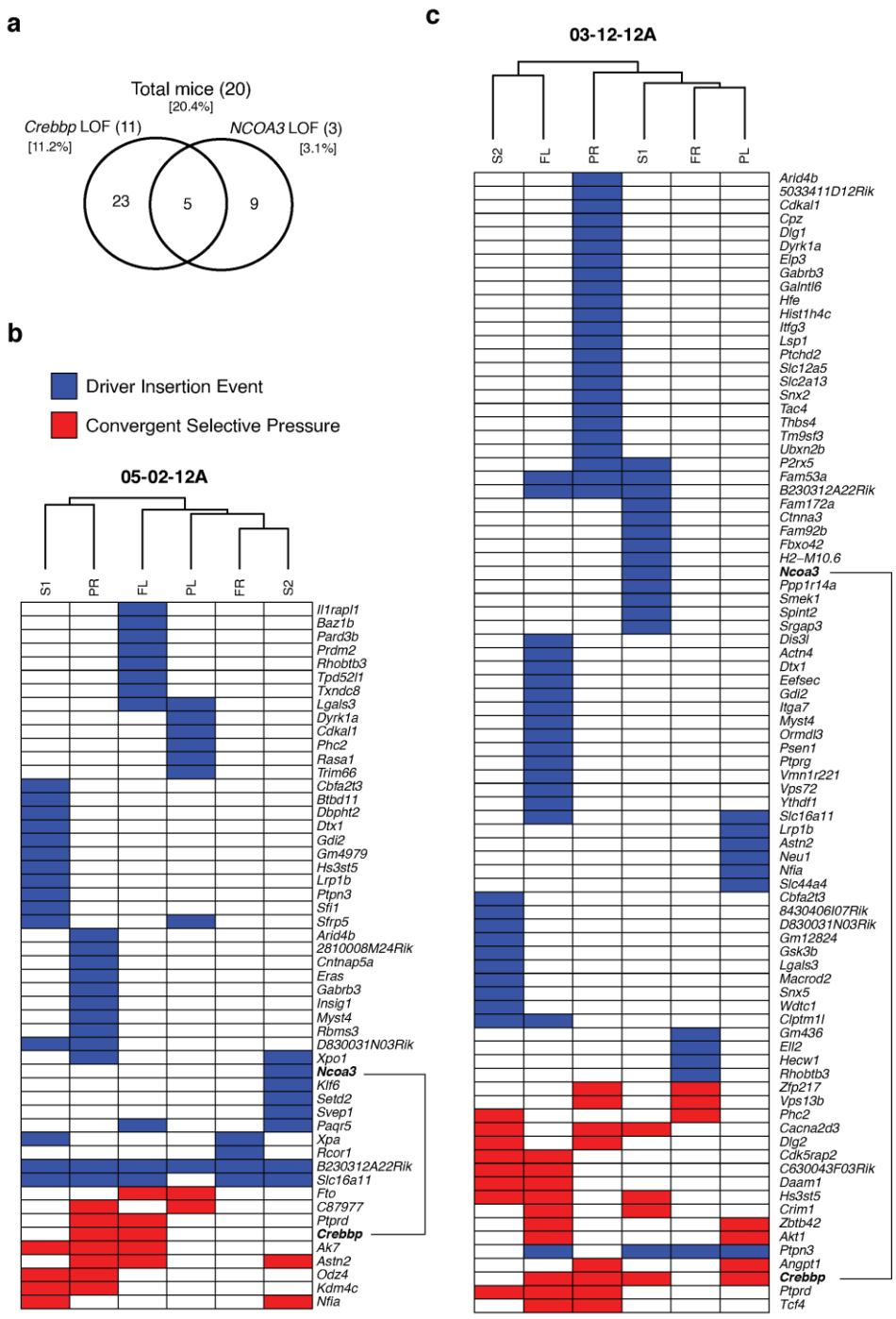



**Figure A11 Convergence of *Crebbp* and *Ncoa3* across SB mice**
**(a)** Recurrence and overlap of *Crebbp* and *Ncoa3* across mice (n = 98). **(b, c)** Driver profiles of mice containing both *Crebbp* and *Ncoa3* insertions (highlighted with a line). Convergent insertions are indicated in red.

**Table 2 Transgenic mice and genotyping PCR primer sequences**
All transgenic mice and genotyping primers used in Chapter 2 and 3.

| Common Name | Full | Description | Direction | Sequence |
|---|---|---|---|---|
| GFP | Tg(Atoh1-GFP)1Jejo | Portion of Math1 enhancer used to drive expression of a nuclear GFP reporter in the Math1 lineage | Forward | 5'-CTGACCCTGAAGTTCATCTGCACC-3' |
| | | | Reverse | 5'-TGGCTGTTGTAGTTGTACTCCAGC-3' |
| SB68/SB76 | TgTn(sb-T2/Onc)68Dla | Sleeping Beauty transposon concatemer. SB68 is in chr15 while SB76 is in chr1 | Forward | 5'-AGTGGGTCAGAAGTTTACATACAC-3' |
| | | | Reverse | 5'-GCTTCAGATCGAATTCCTGCA-3' |
| J2Q | Tg(Atoh1-sb11)Mtay | Transgene SB11 transposase was expressed under regulation of mouse Atoh1 enhancer | Forward | 5'-GCTTGGGGTCATTGTCTTGT-3' |
| | | | Reverse | 5'-CTACGGTTTGCAAGAGCACA-3' |
| PTCH | $Ptch1^{tm1Mps}/J$ | promoterless lacZ-neo fusion gene was inserted into start codon deleting a portion of exon 1 and all of exon 2 | Forward | 5'-TGTCTGTGTGTGCTCCTGAATCAC-3' |
| | | | Reverse | 5'-TGGGGTGGGATTAGATAAATGCC-3' |
| PTCH(flox) | $Ptch1^{tm1Bjw}$ | Exon 3 flanked by a single upstream loxP site in intron 2 and an FRT-neo-FRT-loxP cassette in intron 3 | Forward | 5'-CCACCAGTGATTTCTGCTCA-3' |
| | | | Reverse | 5'-AGTACGAGGCATGCAAGACC-3' |
| SMOA1 | Tg(Neurod2-Smo*A1)199Jols/J | Transgene containing SmoA1 with constitutively active point mutation under control of Neurod2 promter which is specific to granule cells | Forward | 5'-AATCTCTGCTTTTCCTGCGTTGGG-3' |
| | | | Reverse | 5'-CTCGGTCATTCTCACACTTG-3' |
| CREBBP(flox) | $Crebbp^{tm1Jvd}/J$ | loxP sites flanking exon 9 of the Crebbp gene | Forward | 5'-TGGGTGTGTAGATGCAAGGT-3' |
| | | | Reverse | 5'-GGCTTGAACGCTGAAAGAAC-3' |
| LGALS3(flox) | $Lag3^{tm1Doi}$ | 3.7kb of sequence, encompassing exons 2 through 4, was replaced via the insertion of a neomycin selection cassette | Forward | 5'-GACTGGAATTGCCCATGAAC-3' |
| | | | Reverse 1 | 5'-TCGCCTTCTTGACGAGTTCT-3' |
| | | | Reverse 2 | 5'-GAGGAGGGTCAAAGGGAAAG-3' |

**Table 3 Sleeping Beauty sequencing primers**

| Name | Sequence |
|---|---|
| linker+ | 5’-GTAATACGACTCACTATAGGGCTCCGCTTAAGGGAC-3’ |
| linker- | 5’-Phos-GTCCCTTAAGCGGAG-C3spacer-3’ |
| IRR | 5’-GGATTAAATGTCAGGAATTGTGAAAA-3’ |
| IRL | 5’-AAATTTGTGGAGTAGTTGAAAAACGA-3’ |
| PBR | 5’-CTCCAAGCGGCGACTGAG-3’ |
| PBL | 5’-CGATAAAACACATGCGTC-3’ |
| JXR | 5’-GTTGAGTACTAAGCTTGTGCTTAACAAT-3’ |
| JXL | 5’-CTAAGCTTTTAAATTGTTAAGCACAAGC-3’ |
| linker-A1 | 5’-GTAATACGACTCACTATAGGGC-3’ |
| IR-BARCODED | 5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT (BARCODE)TGTATGTAAACTTCCGACTTCAACTG-3’ |
| PB-BARCODED | 5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT (BARCODE)TATCTTTCTAGGGTTAA-3’ |
| Linker-A2 | 5‘-CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTAGGGCTCCGCTTAAGGGAC-3’ |
| PB-Transposon | 5‘-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTTATCTTTCTAGGGTTAA-3’ |

Patryk Skowron*, Kevin Wang*, Raul A. Suarez, A., Xiaochong Wu, Michael, D. Taylor

## Lazy Piggy transposon system

The Lazy Piggy (LP) system is a hybrid transposon system containing both PiggyBac and Sleeping Beauty (SB) excision sequences as well as cargo capable of dysregulating gene expression (Figure A12a). This system is spatially regulated using Nestin driven SB transposase (*Nestin*:Luc-SB100) and temporally regulated with a tamoxifen inducible PibbyBac transposon (Nestin:Cre/ R26:LSL-mPB-Ert2). Mobilization of the hybrid transposon in *Nestin* expression cells generates a highly penetrate model of medulloblastoma. Subsequent activation of R26:LSL-mPB-Ert2 with tamoxifen then gradually depletes insertion events by removing cargo and restoring normal gene function. Initiation and passenger events in cells will be removed without any consequence, but remobilization of a transposon in a maintenance gene will result in cell death or failure to proliferate, thereby stopping its clonal contribution to the tumour. This process gradually depletes initiators and passenger events and enriches for maintenance insertions.

## Design and optimization of Lazy Piggy SPLINK- based library preparation

Quintuple tamoxifen positive (TAM (+)) and litter matched tamoxifen negative (TAM (-)) tumours were used in the design and optimization of the library preparation protocol. A restriction-based SPLINK protocol was designed to identify SB insertion (IR), excision junction events (JX), and PiggyBac maintenance events (PB) in both transposon orientations. There were two donor mice generated, LP-137 and LP-128, with the donor concatemer located on chr10 and chr7 respectively. Only LP-129 had a significantly different survival between TAM (+) and TAM (-) treatments (Figure A12b). These differences were not due to the potential therapeutic effects of tamoxifen treatment since LP mice without active Piggybac transposition show no survival difference between TAM (+) and TAM (-) groups (not shown).

## Analysis of Lazy Piggy TAM+ and TAM- mice

The proportion of clonal insertions that overlap between IR and JX libraries was accessed in a pair-wise manor. The IR insertion set contains all clonal insertions including initiator, passenger, and maintenance events. The JX set contains insertions for which cargo had been excised and normal gene function was restored. It was hypothesized that TAM (-) LP tumours would not contain evidence for JX clonal insertions since the R26:LSL-mPB-Ert2 depends on tamoxifen dependant activation. Unfortunately, JX transposon scars were detected in TAM (+) and TAM (-) tumours, suggesting the system is intrinsically leaky (Figure A12c). Fortunately, there is significantly more overlap between IR and JX in TAM (+) compared to the TAM (-) tumours ($p = 0.034$; T-test) (Figure A12d), suggesting a higher rate of remobilization in TAM (+) tumours. To maximize power, TAM (+) and TAM (-) samples were pooled together for the gCIS analysis (Figure A12e) and genes predominately found in TAM (+) mice were shortlisted for validation. Most notable were genes coding the two potassium channel protein *Kcnh2* and *Kcnb1*, and a recurrent medulloblastoma primary/recurrence tumor gene, *Dyn1h1*.

## RNAseq of Lazy Piggy Tumors

With restriction-SPLINK insertion data it is not immediately clear what the result of integration is on gene expression, especially when multiple insertions are detected in the same gene or when insertions are detected in intergenic space. RNAseq has shown great potential in overcoming these problems[190]. A large cohort of tumours ($n = 60$) was sequenced by RNAseq (10 million reads/sample) deep enough to detect clonal fusions. Clustering of data from RNAseq libraries using PCA analysis demonstrated a tendency for TAM+ samples to cluster closer together (Figure A12f). This was further confirmed using PCA on differentially expressed genes (Figure A12g).

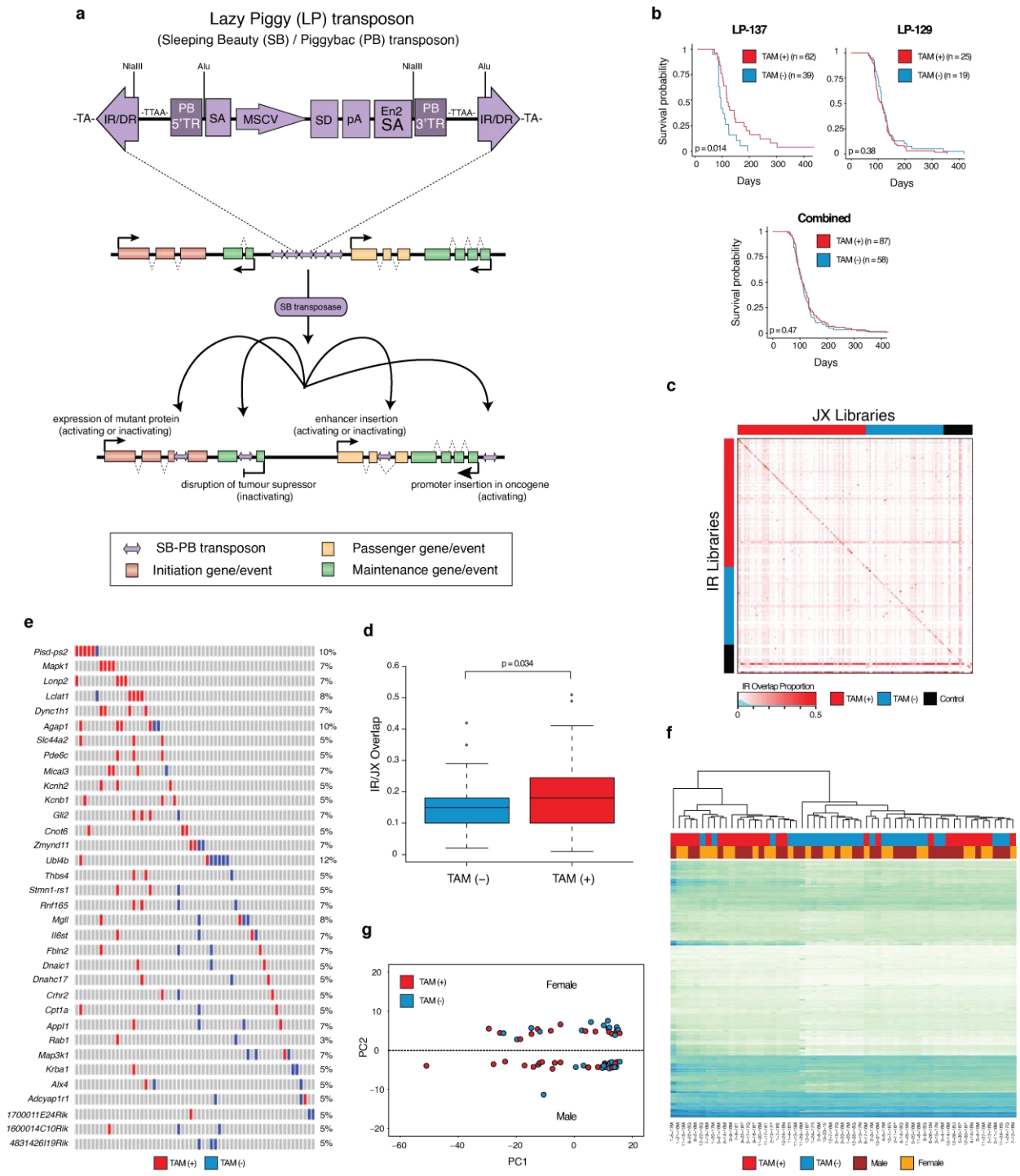



**Figure A12 Lazy Piggy mouse model analysis**

**(a)** Overview of the Lazy Piggy (LP) mouse model transposition system. **(b)** Survival difference between different LP donor mice. **(c)** Clonal insertion overlap matrix between all transposon (IR) and excision scar (JX) events. Red is tamoxifen positive (TAM (+)), blue is tamoxifen negative (TAM (-)) and black is control samples with no LP. **(d)** Differences in clonal excision events between TAM (+) and TAM (-) samples. **(e)** gCIS analysis on combined TAM (+) and TAM (-) LP libraries **(e)** PCA analysis on 60 RNAseq LP samples. **(f)** Hierarchical clustering of LP samples using TAM (+) vs. TAM (-) differentially expressed genes.

# Copyright Acknowledgements

Section 1.1 is adapted from a published review paper. Permissions for use below:

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS
Nov 13, 2019

---

This Agreement between University of Toronto ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

| | |
|---|---|
| License Number | 4654840580077 |
| License date | Aug 23, 2019 |
| Licensed Content Publisher | Springer Nature |
| Licensed Content Publication | Journal of Molecular Medicine |
| Licensed Content Title | Genetic and molecular alterations across medulloblastoma subgroups |
| Licensed Content Author | Patryk Skowron, Vijay Ramaswamy, Michael D. Taylor |
| Licensed Content Date | Jan 1, 2015 |
| Licensed Content Volume | 93 |
| Licensed Content Issue | 10 |
| Type of Use | Thesis/Dissertation |
| Requestor type | academic/university or research institute |
| Format | print and electronic |
| Portion | full article/chapter |
| Will you be translating? | no |
| Circulation/distribution | <501 |
| Author of this Springer Nature content | yes |
| Title | Deciphering Genetic Drivers in Primary and Metastatic Medulloblastoma |
| Institution name | University of Toronto |
| Expected presentation date | Dec 2019 |
| Order reference number | 1 |
| Requestor Location | University of Toronto<br>27 King's College Circle<br><br>Toronto, ON M5S 1A1<br>Canada<br>Attn: University of Toronto |
| Total | 0.00 CAD |