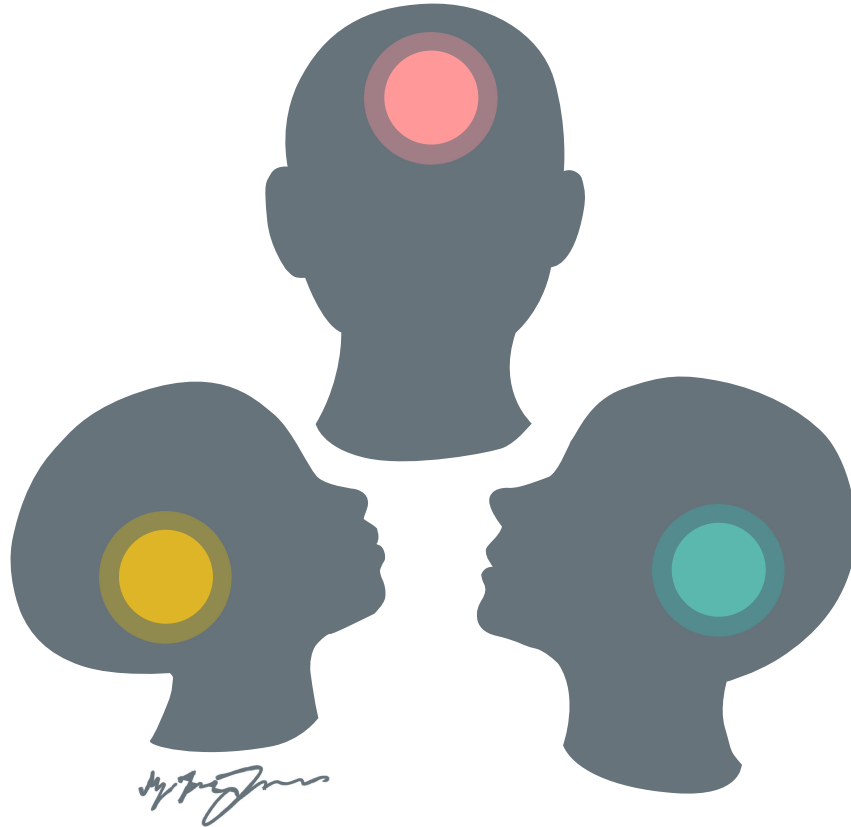




CHALMERS
UNIVERSITY OF TECHNOLOGY



Exploring Tumor Evolution in Pediatric CNS Tumors

Evaluation of EVOFLUX-Derived Parameters
and Posterior Summary Statistics from DNA Methylation Data

MY IRENHEIM JARLHED

MASTER'S THESIS

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
2026

MASTER'S THESIS 2026

Exploring Tumor Evolution in Pediatric CNS Tumors

Evaluation of EVOFLUX-Derived Parameters and Posterior
Summary Statistics from DNA Methylation Data

MY IRENHEIM JARLHED



CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026

Exploring Tumor Evolution in Pediatric CNS Tumors
Evaluation of EVOFLUX-Derived Parameters and Posterior Summary Statistics
from DNA Methylation Data
My Irenheim Jarlhed

© My Irenheim Jarlhed, 2026.

Master's thesis 2026
Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden

Cover: Illustration by My Irenheim Jarlhed.

Typeset in L^AT_EX

Exploring Tumor Evolution in Pediatric CNS Tumors
Evaluation of EVOFLUX-Derived Parameters and Posterior Summary Statistics
from DNA Methylation Data
My Irenheim Jarlhed
Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden

Abstract

In this thesis, the EVOFLUX framework was investigated as a method for inferring evolutionary features from DNA methylation data. The DNA methylation data was taken from pediatric central nervous system CNS tumor samples. EVOFLUX was originally developed for lymphoid malignancies and uses fluctuating CpG sites (fCpGs) to estimate parameters describing tumor evolution.

The framework was applied to pediatric CNS tumor samples from Barntumörbanken. The fCpG selection and filtering procedure was evaluated for the tumor samples, after which EVOFLUX was used to infer parameters including tumor growth rate, tumor age, methylation switching rates, and effective population size. The resulting posterior distributions were summarized using the median, median absolute deviation (MAD), skewness, and kurtosis. The inferred parameters and summary statistics were compared between age groups and tumor types and were evaluated in relation to time to first relapse.

The analyses showed that the age of the patients were associated with differences of the inferred parameters. And that this is needed to take into account when doing analyses. Summary statistics describing the shape of the posterior distributions were evaluated and showed to also have importance meaning that using only the median of the posterior distributions would leave out possibly important information.

The results indicate that EVOFLUX can be used for analysis of evolutionary features in pediatric CNS tumor methylation data and that statistics describing shape of the posterior distributions can provide further insights. However, the survival analyses were limited by small sample sizes and median-based cutoffs. Larger cohorts and additional clinical data are needed to validate these findings and determine whether EVOFLUX-derived features can be useful for relapse prediction.

Keywords: EVOFLUX, pediatric CNS tumors, DNA methylation, fluctuating CpGs, tumor evolution, relapse, Bayesian inference

Acknowledgements

The computations were enabled by resources in project [NAISS YY/A-BBB] provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX, founded by the Swedish Research Council through grant agreement no. 2022-06725.

Thanks to the Carén research group for funny and interesting lunch conversations. To my supervisor Ryan and to my friend Yuna for support during the project and for explaining biological concepts. To my family for always being there and believing in me even when I didn't believe in myself. To all friends that made the years at Chalmers fun and interesting. Och till Farmor, mormor och morfar, för ni under hela min utbildning fått mig att hitta nya och kreativa sätt att förklara vad det är jag håller på med.

My Irenheim Jarlhed, Gothenburg, Oktober 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

BTB	Barntumörbanken
CNV	Copy Number Variation
CNS	Central Nervous System
CpG	Cytosine–phosphate–Guanine site
DMG	Diffuse Midline Glioma
DNA	Deoxyribonucleic Acid
EPIC	Infinium MethylationEPIC BeadChip Array
fCpG	Fluctuating CpG site
HGG	High-Grade Glioma
IDAT	Intensity Data file
MAD	Median Absolute Deviation
MB	Medulloblastoma
MNP	Molecular Neuropathology
MRCA	Most Recent Common Ancestor
PA	Pilocytic Astrocytoma
pedHGG	Pediatric-type High-Grade Glioma
SNP	Single-Nucleotide Polymorphism
t-SNE	t-Distributed Stochastic Neighbour Embedding
TCGA	The Cancer Genome Atlas
WHO	World Health Organization

Nomenclature

Below is the nomenclature of the parameters, variables, and quantities used throughout this thesis.

Indices

i	Index for a sample or tumor profile
j	Index for a CpG site or fCpG site

Sets and data

$fCpGs$	Fluctuating CpG sites used as input to EVOFLUX
$\mathcal{D}$	Input data consisting of fCpGs, patient age, and tumor purity

Evolutionary parameters

Θ	Vector of EVOFLUX parameters, $(\theta, \tau, \mu, \gamma, \nu, \zeta, \beta_n, \delta, \epsilon, \kappa)$
θ	Initial tumor growth rate per year
τ	Patient age at the most recent common ancestor (MRCA)
T	Patient age when the sample was collected
τ/T	Patient age at tumor initiation divided by age at sampling; the fraction of life elapsed before tumor initiation
$T - \tau$	Tumor age at sampling, i.e., time from tumor initiation (MRCA) to sample collection
N_e	Tumor effective population size at sampling
$\log(N_e)$	Logarithm of the tumor effective population size
μ, γ	Initial epigenetic switching rates per allele per year
ν, ζ	Epigenetic switching rates for transitions between methylation states

ν_{rel}	Relative switching rate, ν/μ
ζ_{rel}	Relative switching rate, ζ/γ
β_n	Normal methylation fraction
δ, ϵ	Array saturation offset parameters
κ	Noise precision parameter

Methylation variables

M	Methylated signal
U	Unmethylated signal
β	DNA methylation beta value, calculated from the methylated and unmethylated signals

Summary statistics

$\tilde{x}$	Median of a posterior distribution
MAD	Median absolute deviation of a posterior distribution
skewness	Measure of the asymmetry of a posterior distribution
kurtosis	Measure of the tail behaviour and peakedness of a posterior distribution

Contents

List of Acronyms	ix
Nomenclature	xi
List of Figures	xv
List of Tables	xvii
1 Introduction	1
1.1 Background	1
1.2 Aim	2
1.3 Scope	2
2 Theory	3
2.1 Biological background	3
2.1.1 DNA Methylation	4
2.1.2 Pediatric Central Nervous System, CNS, Tumors	5
2.1.2.1 Tumor taxonomy	5
2.2 Sequencing DNA	7
2.3 EVOFLUX framework	8
2.3.1 Bayesian Modeling	8
2.3.2 DNA Methylation as a Markov Chain	9
2.3.3 Evolutionary Parameters	10
2.3.3.1 Priors	11
2.3.4 Sampling	12
2.3.5 Model assumptions and simplifications	13
2.4 Metrics	13
2.4.1 Summary statistics	13
3 Method	15
3.1 Data	15
3.1.1 Methylation data	16
3.1.2 Tumor purity	16

3.2	Quality filtering	16
3.3	Excluding chromosome with CNV	17
3.4	Finding fCpG sites	18
3.4.1	Patient age at time of sample	19
3.5	Running simulation	19
3.6	Analysis of simulation	20
4	Results	23
4.1	Simulation	24
4.2	Age Dependency	28
4.3	Survival Analysis	29
4.4	Additional findings: Inferred switching rates	33
5	Discussion	35
5.1	Interpreting posterior summary statistics in relation to relapse	36
5.2	Switching rates and age	36
5.3	Further Research	36
5.3.1	Validation and Evaluation	36
5.3.2	Extensions of the Project	37
5.3.3	Gaining Insights into Pediatric CNS Tumors	38
6	Conclusions	39
	Bibliography	40
A	CNV analysis implementation	I
B	Boxplots showing age separation	V

List of Figures

2.1	Image summarizing how DNA is stored and organized in cell. File:Chromosome-en.svg [5] CC-BY-3.0	3
2.2	Illustration of CpG sites with different methylation status and their corresponding beta values. The C nucleotides with a "M" connected to them are methylated. The beta value is the ratio of methylated C nucleotides to the total number of C nucleotides at the CpG site. . .	4
2.3	Illustration of CpG sites with different methylation status and their corresponding beta values. The C nucleotides with a "M" connected to them are methylated. The beta value is the ratio of methylated C nucleotides to the total number of C nucleotides at the CpG site. . .	7
2.4	Overview over inputs and outputs of EVOFLUX	8
2.5	Markov chain for methylation states. Filled circles indicate methylated CpG sites. The arrows indicate the transition rates between states. The switching parameters μ , γ , ν , and ζ are defined in Section 2.3.3.	10
2.6	Prior distributions for all model parameters.	12
3.1	Schematic overview of the fCpG selection procedure. Green boxes show data-processing steps, orange boxes show filtering or selection criteria.	18
4.1	Summary of samples that converged during the simulation and patient age distribution. Diamond markers indicate deceased patients and round markers indicate alive patients. X marks non-converged samples.	24
4.2	t-SNE plots for different data colored on tumor type. ¹	25
4.3	t-SNE plots for different data colored by patient age at sampling. ¹ . .	25
4.4	t-SNE plots for different data colored by tumor purity of the sample. ¹	26
4.5	Median posterior growth rate (θ_{median}) versus patient age at sampling, with 95% credible intervals. Points are colored by tumor type.	27
4.6	Median posterior relative tumor-initiation age (τ/T) versus patient age at sampling, with 95% credible intervals. Points are colored by tumor type.	27

4.7	Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low τ/T median, in relation to the median within that tumor groups.	30
4.8	Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low θ median, in relation to the median within that tumor type.	30
4.9	Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low $\log(N_e)$ median, in relation to the median within that tumor groups.	31
4.10	Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low $\log(N_e)$ skewness, in relation to the median within that tumor groups.	31
4.11	Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low $\log(N_e)$ excess kurtosis, in relation to the median within that tumor groups. .	32
4.12	Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low τ/T MAD, in relation to the median within that tumor groups.	33
4.13	Boxplots of the median relative switching rates, ν_{rel} and ζ_{rel} , for the analysed tumor types.	34
4.14	Boxplots of the median average switching rates for the analysed tumor types.	34
B.1	Boxplots showing separation between two age groups for summary statistics median and MAD for all infered parameters and all tumor types. Images with gray background have a p-value over 0.05. Images with with background is with 95% probability actually separated and the results did not just happen by chance.	VI
B.2	Boxplots showing separation between two age groups for summary statistics skewness and kurtosis for all infered parameters and all tumor types. Images with gray background have a p-value over 0.05. Images with with background is with 95% probability actually separated and the results did not just happen by chance.	VII

List of Tables

2.1	Taxonomy of Pilocytic Astrocytoma according to its MNP classification.	5
2.2	Taxonomy of Medulloblastoma subtypes according to their MNP classification.	6
2.3	Taxonomy of High Grade Glioma subtypes according to their MNP classification.	6
3.1	Chromosomal arms excluded from the fCpG data for each tumor type based on the two-standard-deviation criterion. The letters p and q denote the short (p) and long (q) arms of the chromosome.	18
4.1	Age-separation p-values for posterior summary statistics (patients under nine versus over nine years of age). Gray cells indicate p-values ≥ 0.05 ; uncolored cells indicate p-values < 0.05	28

1

Introduction

Central nervous system, CNS, tumors, brain tumors, are the most common cause of cancer related death among children. Pediatric cancer develop over a much shorter time frame then most cancers in adults. Children have not been alive for enough time for the genetic mutations which are mostly the cause of cancer in adults to accumulate. Instead pediatric brain cancer is believed to be caused by errors in the epigenetic processes. Epigenetic processes control which genes in the DNA are expressed without altering the DNA. The Group Helena Carén at Sahlgrenska Center for Cancer Research investigates pediatric central nervous system tumors from an epigenetics perspective. A clinical challenge in pediatric brain tumors is that there is recurrence after initially successful treatment. Tumor recurrence is thought to be driven by cancer stem cells. As long as these cells persist, the tumor has the potential to reappear [1]. Understanding the evolutionary history of a tumor could reveal which cell populations are driving recurrence, potentially improving diagnostics and treatment. This thesis will explore a new model for inferring the evolutionary history of tumors based on DNA methylation data. The model, called EVOFLUX, was developed for lymphoid malignancies and has not yet been applied to pediatric brain tumors [2].

1.1 Background

The genes in our DNA contains instructions for the cells, but not all instructions are active at the same time. Epigenetic processes act as a control system that determines whether a gene is 'turned on' or 'turned off' without altering the underlying DNA [3]. One of these regulating factors is DNA methylation, which happens at CpG sites in the DNA [3]. Some of the CpG sites in the DNA can change methylation status randomly over time with some probability [2]. These are known as fluctuating CpGs, fCpGs, [2]. The stochastic methylation fluctuations of the fCpGs can be used as a factor when modeling a tumor. The fCpGs can provide a record of the tumor's developmental history by being a measurement of time [2]. In 2025 the EVOFLUX model was published as a way to model lymphoid malignancies. The model uses the independent fluctuations of fCpGs, together with the patient's age and tumor purity, to infer the tumor's evolutionary timeline. This process enables the identification of underlying parameters that may serve as indicators for aggressive or recurrent disease. The model uses Bayesian inference to estimate these latent parameters such as tumor growth rate, fluctuation rates for methylation, and time

since the most recent common ancestor (MRCA) [2]. EVOFLUX has not yet been applied to pediatric brain tumors [2].

Because of its use of Bayesian inference EVOFLUX provides a posterior distribution for each of the inferred parameters instead of a single value. A posterior distribution is a probability distribution over all possible values of a parameter, given the observed data. Distributions can have many different shapes which can be described in many different ways. One way to describe the shape of a distribution is by using summary statistics. Summary statistics are single values that describe a specific aspect of the distribution. The median is one of the summary statistic that can describe the center of a distribution. Other summary statistics describing different aspects of shape are median absolute distance, skewness and kurtosis.

1.2 Aim

The aim of this thesis is to investigate whether the EVOFLUX model can be applied to pediatric CNS tumors. The study focuses on evaluating the existing identification of fluctuating CpG sites to pediatric brain-tumor samples and evaluating whether EVOFLUX can infer evolutionary parameters from these data.

A further aim is to investigate whether the inferred parameters and summary statistics of their posterior distributions are associated with patient age, tumor type and time to first relapse which refers to the time between the sample was taken and relapse. Patients without a recorded relapse are treated as censored observations at the end of follow-up. The analysis is exploratory and is intended to assess the potential usefulness of the EVOFLUX model and its derived features for studying tumor evolution and relapse in pediatric CNS tumors.

1.3 Scope

This thesis is limited to the analysis of existing DNA methylation data from pediatric CNS tumors obtained from Barntumörbanken, BTB [4]. No new biological samples are collected. The analysis focuses on the application and evaluation of the EVOFLUX model rather than on reimplementing the underlying Bayesian inference method.

The study investigates inferred evolutionary parameters, including tumor growth rate, tumor age, methylation switching rate, and related quantities, together with summary statistics such as the median, skewness, kurtosis and MAD. These features are compared between age groups and tumor types and are evaluated in relation to time to first relapse.

The survival analyses are exploratory and are not intended to provide clinically validated prognostic models. In particular, the study does not establish causal relationships between inferred parameters and patient outcome, nor does it determine whether the parameters can be used for clinical decision making.

2

Theory

2.1 Biological background

DNA is built up of nucleotides which are usually referred to as A, T, C, and G [1]. They form base pairs, where A and T are connected and C and G are connected. The nucleotides are placed on the two strands which form the double helix of the DNA, see figure 2.1. A gene is a specific sequence of base pairs that carries information about how the cell is supposed to function. The DNA makes up the chromosome which are stored in the nucleus of our cells [1]. The chromosomes are usually drawn as an "X". The part of the chromosome where the two arms intersect is called the centromere. The centromere divides the chromosomes into two arms, one shorter called p and one longer called q.

Even though each human have unique DNA most of the DNA is identical between humans. Sometimes the DNA which is usually shared between humans gets variations, where the DNA looks different in one person compared to the rest of the population. One of these variations is called single nucleotide polymorphisms, SNP - which means that there is a single base pair that differs [1]. Another error is copy number variation, CNV, which means that a chromosome or a big part of it is duplicated or deleted [1].

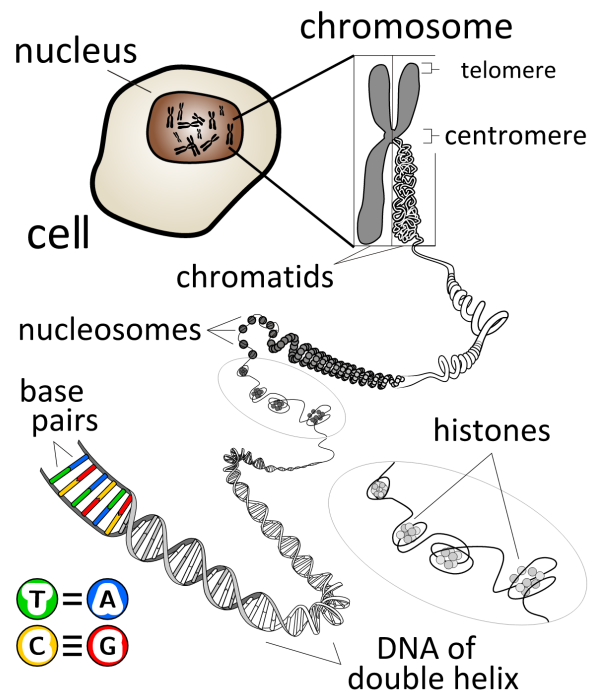


Figure 2.1: Image summarizing how DNA is stored and organized in cell. File:Chromosome-en.svg [5] CC-BY-3.0

2.1.1 DNA Methylation

DNA methylation happens at CpG sites which are locations where a C nucleotide is followed by a G nucleotide. The site becomes methylated when a methyl group connects to the C nucleotide. Typically, when a CpG site is methylated, the gene connected to that site is turned off and becomes inactive [3]. At a CpG site there are two basepairs: C connected to G and G connected to C, figure 2.2. One CpG site can therefore be unmethylated (none of the Cs are methylated), hypomethylated (one of the Cs are methylated), or hypermethylated (both of the Cs are methylated).

In tumors, the methylation status of CpG sites linked to specific genes is frequently altered in a non-random manner, ensuring that genes promoting tumor growth remain active while those restricting it are silenced [3]. By contrast, many CpG sites have no impact on tumor survival. Because these sites confer neither selective advantage nor disadvantage, their methylation state is not actively fixed or altered by the tumor and may therefore change randomly over time [2]. Such sites are termed fluctuating CpGs, or fCpGs [2]. While the tumor microenvironment can influence the overall rate of stochastic methylation switches at these sites, as well as the rate of cell division, it does not determine the direction of individual fCpG changes. In this respect, fCpGs differ fundamentally from the non-random methylation alterations at driver genes, which the tumor actively shapes to promote proliferation. Independent of direct tumor influence, fCpGs undergo stochastic methylation fluctuations that provide a record of the tumor's developmental history [2] [6].

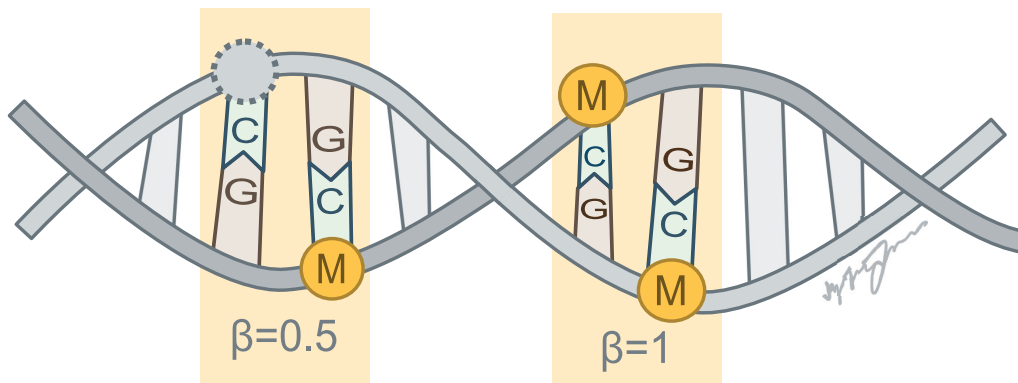


Figure 2.2: Illustration of CpG sites with different methylation status and their corresponding beta values. The C nucleotides with a "M" connected to them are methylated. The beta value is the ratio of methylated C nucleotides to the total number of C nucleotides at the CpG site.

2.1.2 Pediatric Central Nervous System, CNS, Tumors

Central Nervous System Tumors, CNS, are tumors that are located in the brain or spinal cord. CNS tumors are the most common cause of death in pediatric cancer patients [7]. Although rare, tumors can develop before birth [8]. The growth rate of the tumor is an important factor for how aggressive the tumor is. Growth rate of normal cells is also highest during the first years of life and then decreases with age [9]. This developmental context could be important when modeling pediatric tumors, as tumor growth is a clonal process in which all cells descend from a single origin cell, known as the Most Recent Common Ancestor (MRCA) [2]. A tumor sample is a mix of different tissues, some tumor tissue, some healthy tissue and some dead cells. A sample of a tumor can therefore contain a mix of different cells and the tumor purity is the proportion of tumor cells in the sample [7].

The tumor types that are investigated in this thesis are divided into three main groups using the classification system used in the MNP classifier [10]. The groups are Pilocytic Astrocytomas, PA, Medulloblastoma, MB and High Grade Gliomas, HGG. High Grade Glioma are in some figures named only Glioma which is incorrect because Pilocytic Astrocytomas also are a type of Glioma. The WHO has established about the CNS tumor taxonomy since the 70s. From the start tumors were diagnosed based on the look of the tissue and the cells. The way tumors have been classified have changed several times and since 2021 the WHO classification criteria also include methylation data [11].

2.1.2.1 Tumor taxonomy

The tumor taxonomy used in this project is the same as is used in the MNP classifier [10] which also have been used to identify the tumor type of the sample. Each sample used in the project have a Family classification score of at least 0.7 meaning the classifier is at least 70% sure it belongs within that Family in the MNP taxonomy. The subtypes of PA and number of samples are stated in table 2.1, subtypes of MB are stated in table 2.2 and subtypes of HGG are stated in table 2.3

Tumor Type	Pilocytic Astrocytoma	
Superfamily	Low-grade glial/glioneuronal/neuroepithelial tumours	
Family	Pilocytic Astrocytoma	
Class	Infratentorial	Supratentorial midline
Subclass		
N. Samples	67	21

Table 2.1: Taxonomy of Pilocytic Astrocytoma according to its MNP classification.

2. Theory

Tumor Type	Medulloblastoma									
Superfamily	Medulloblastoma									
Family	non-WNT/non-SHH activated						SHH-activated		WNT-activated	
Class	Group 3		Group 4				1	2	3	
Subclass	II	IV	V	VI	VII	VIII				
N. Samples	2	2	4	9	6	9	3	4	5	8

Table 2.2: Taxonomy of Medulloblastoma subtypes according to their MNP classification.

Tumor Type	High Grade Glioma				
Superfamily	Pediatric diffuse high-grade gliomas				
Family	H3-wildtype and IDH-wildtype			DHG, H3 G34-mutant	DMG, H3K27-altered
Class	MYCN	RTK1A	RTK1C	DHG-G34	H3K27-mut/EZHIP+
Subclass					DMG_K27
N. Samples	2	2	1	2	9

Table 2.3: Taxonomy of High Grade Glioma subtypes according to their MNP classification.

2.2 Sequencing DNA

The methylation status is given as a fraction even though DNA methylation only have three states, unmethylated, hypomethylated and hypermethylated. This is an artifact of how DNA is sequenced. The measured methylation status for each CpG site represents the methylation observed at that site in the DNA across all of the cells in the sample [12]. An illustration of this is shown in Figure 2.3. The method used for measuring methylation in this study is the Illumina MethylationEPIC array [12]. The technique uses many probes which can read the methylation status from a specific locations of the DNA. The resulting sequence contains the average methylation statuses across all cells in the sample, beta values, for each probe/CpG site.

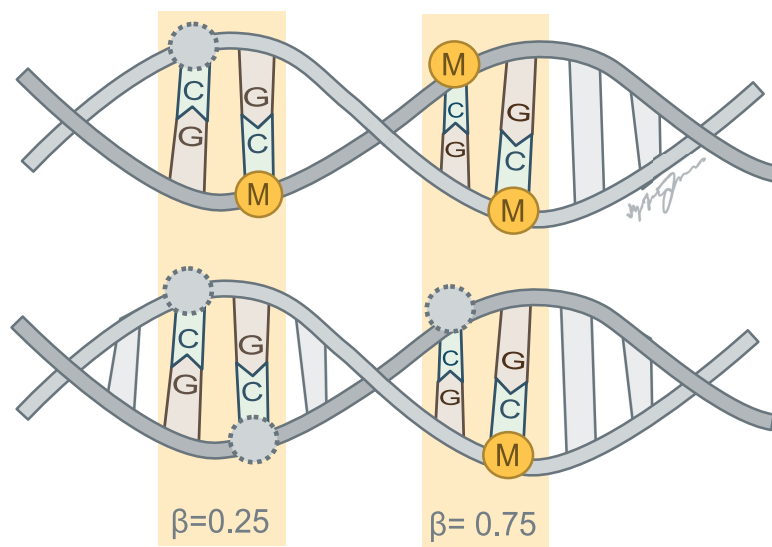


Figure 2.3: Illustration of CpG sites with different methylation status and their corresponding beta values. The C nucleotides with a "M" connected to them are methylated. The beta value is the ratio of methylated C nucleotides to the total number of C nucleotides at the CpG site.

When methylation is detected this is done by two probes. One that detects methylated sites and one that detects unmethylated sites. Beta values represent the proportion of methylated signal at each CpG sites and are calculated as $M/(M+U+100)$, where M and U denote the methylated and unmethylated signals, respectively. The offset is added to stabilize estimates at low intensities. Beta values range from 0 to 1, where 1 is fully methylated [12]. Different dyes are used on the sample to measure the methylated and unmethylated signals. Differences between the dye channels can introduce technical bias, which is why dye-bias correction is included during pre-processing [13]. There are also control probes used to measure background signal and assess the quality of the measurements [12]. The detection-p values are calculating to assess the quality of the measurements. These are calculated by comparing the total signal from the methylation-probe and the unmethylation-probe with the background signal, obtained from the control probes [14].

Some probes can bind to multiple sites in the genome, these are called cross reactive probes and are filtered out during preprocessing [7] [12]. The probes used by Illumina MethylationEPIC v1 array do not cover the whole genome. It measures 850 000 CpG sites [12].

On the EPIC methylation array, the measured intensity at a probe reflects the combined signal from all physical copies of the site. CNS tumors are prone to copy number variations, CNV's, of the chromosomes which could be a problem. A duplicated chromosomeal arm could inflate a CpG site's contribution relative to CpGs on other chromosomes, while a deleted arm could reduce or remove it. This could possibly distort the EVOFLUX model. Furthermore, arms that are frequently deleted or multiplied in a tumor type are more likely to contain regions under tumor selection, so excluding them also removes sites whose methylation patterns may not be truly fluctuating.

2.3 EVOFLUX framework

The EVOFLUX model uses the independent fluctuations of fCpGs to infer a tumor's evolutionary timeline. In addition to the fCpGs, the model also takes patient age and tumor purity as input. An overview of the inputs and outputs of EVOFLUX is seen in Figure 2.4. The process enables the estimation of underlying parameters that may serve as indicators for aggressive or recurrent disease. Due to the use of Bayesian inference in the model, the output is a posterior distribution for each of the underlying parameters [2].

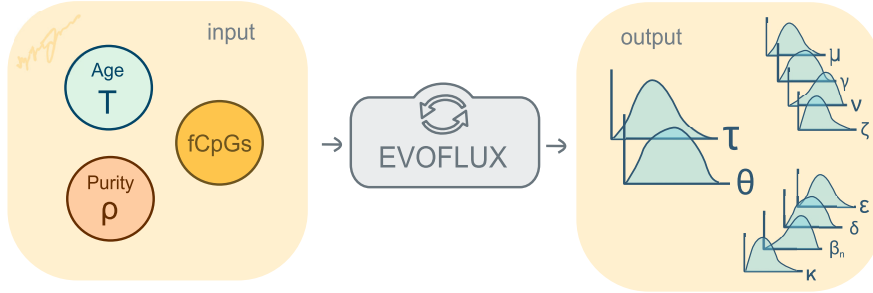


Figure 2.4: Overview over inputs and outputs of EVOFLUX

2.3.1 Bayesian Modeling

Bayesian modelling is a statistical approach for updating the probability of a parameter using observed data. In its basic form, Bayes' theorem can be written as

$$P(\Theta | D) = \frac{P(D | \Theta)P(\Theta)}{P(D)}, \quad (2.1)$$

where Θ denotes the parameters of interest and D denotes the observed data. The term $P(\Theta)$ is the prior distribution and describes the known information about

the parameters before observing the data, such as the plausible range of values, the expected magnitude, and the uncertainty. For example, the prior can encode whether a parameter should lie between 0 and 1, but it also specifies the shape of the distribution and how strongly values near the boundaries are penalised. The likelihood, $P(D \mid \Theta)$, describes how probable the observed data are for a given set of parameter values. The posterior distribution, $P(\Theta \mid D)$, combines the prior information with the information in the data. The denominator, $P(D)$, is called the evidence and normalizes the posterior distribution [15].

For continuous parameters, the corresponding formulation uses probability densities rather than probabilities. This is the formulation used in EVOFLUX and is shown in equation 2.2. EVOFLUX uses Bayesian modelling to estimate parameters from observed data, D . D consists of methylation values at fCpG sites, patient age and estimated tumor purity, $D = (\text{fCpGs}, \text{age}, \text{tumor purity})$.

$$\underbrace{f(\Theta \mid D)}_{\text{Posterior}} = \frac{\overbrace{f(D \mid \Theta)}^{\text{Likelihood}} \cdot \overbrace{f(\Theta)}^{\text{Prior}}}{\underbrace{\int f(D \mid \Theta) f(\Theta) d\Theta}_{\text{Evidence}}} \quad (2.2)$$

In the EVOFLUX model, Θ is the vector of evolutionary and construction parameters that are estimated by the model, $\Theta = (\theta, \tau, \mu, \gamma, \nu, \zeta, \beta_n, \delta, \epsilon, \kappa)$. The posterior distributions represents both the estimated parameter values and their uncertainty. The specific likelihood function and prior distributions used by EVOFLUX are detailed in Sections 2.3.4 and 2.3.3.

2.3.2 DNA Methylation as a Markov Chain

The methylation fraction of a CpG site can be seen as a state. An unmethylated CpG site can either stay where it is or become hypomethylated; a hypomethylated CpG site can stay as it is, become unmethylated, or become hypermethylated. These transitions are shown in figure 2.5, where the arrows and variables indicate the transition rates to different states. In a Markov model, the future evolution depends only on the present state and not on how that state was reached. The switching parameters μ, γ, ν and ζ are rates per fCpG per year. When the process is simulated with a small discrete time step δt , the probability of making a transition in one step is approximately the corresponding rate multiplied by δt [2].

$$\delta t = \frac{0.1}{\max(\theta, 2\mu, 2\gamma, 2\nu, 2\zeta)}$$

Figure 2.5 also shows that the rate from the unmethylated state to the hypomethylated state is 2μ , because there are two C nucleotides that can become methylated, each with rate μ . Similarly, the rate from the hypermethylated state to the hypomethylated state is 2γ , because either of the two methylated C nucleotides can become demethylated. In principle, direct transitions between the unmethylated and hypermethylated states are possible, but EVOFLUX uses a simplified model in

which these transitions are not included. This simplification is justified by having a small enough time step δt that making a double jump is unlikely [2].

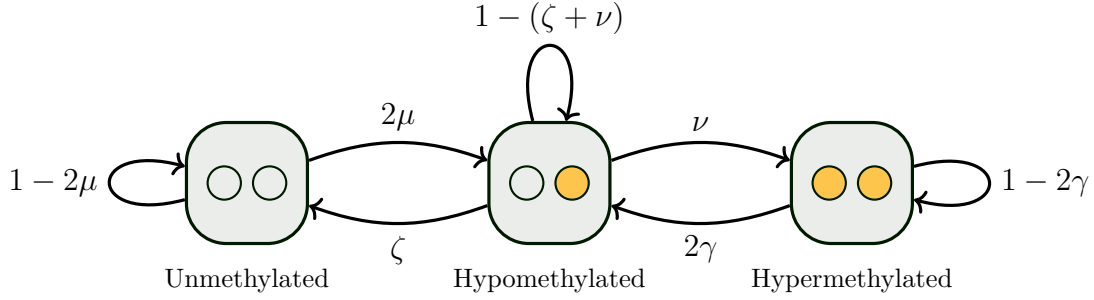


Figure 2.5: Markov chain for methylation states. Filled circles indicate methylated CpG sites. The arrows indicate the transition rates between states. The switching parameters μ , γ , ν , and ζ are defined in Section 2.3.3.

2.3.3 Evolutionary Parameters

EVOFLUX estimates a vector of parameters for each tumor sample, $\Theta = (\theta, \tau, \mu, \gamma, \nu, \zeta, \beta_n, \delta, \epsilon, \kappa)$. The parameters describe tumor growth, tumor age, methylation switching, and measurement-related effects. The main evolutionary quantities used in this thesis are described below, the measurement-related parameters β_n , δ , ϵ , and κ are summarized at the end of this section.

Initial tumor growth rate (θ): The estimated growth rate per year, assuming that the tumor population grows exponentially. The model assumes a constant growth rate over the analysed period. Therefore, θ should be interpreted as an average growth rate over the time between the MRCA and sample collection [2].

Relative age at tumor initiation (τ/T): τ is the patient's estimated age when the tumor originated and T is the patient's age at sample collection. Thus, τ/T is the fraction of the patient's age at sampling that had elapsed before tumor initiation. The tumor age at sampling is $T - \tau$.

Tumor effective population size (N_e): The number of long-lived tumor lineages at sampling, calculated under the model's exponential-growth assumption using the growth rate and tumor age, $T - \tau$:

$$N_e = e^{\theta(T-\tau)}.$$

Epigenetic switching rates (μ , ν , γ , and ζ): The rates per allele per year for transitions between the three possible methylation states (0%, 50%, and 100%) at fluctuating CpG sites. The parameters ν and ζ are represented relative to μ and γ , respectively, to reduce redundancy in the model [2]:

$$\nu_{\text{rel}} = \frac{\nu}{\mu}, \quad \zeta_{\text{rel}} = \frac{\zeta}{\gamma}.$$

Higher switching rates mean faster stochastic turnover of methylation states and lower switching rates mean slower change of methylation states.

The above parameters are the main focus of the analyses in this thesis. They are also the ones used in the Markov model for estimating the likelihood which will be discussed in section 2.3.4. The remaining parameters account for tumor purity, normal-tissue contamination, and technical measurement effects. And are used in the model for correction of biases and noise. These parameters are required by the model but were not the primary focus of the analyses in this thesis.

Normal methylation fraction (β_n): The expected methylation level in contaminating non-tumor cells; the subscript n indicates normal (non-tumor) tissue. It is combined with tumor purity to adjust the observed methylation fractions toward a pure tumor signal.

Array saturation offsets (δ, ϵ): Technical parameters that model methylation-array saturation at the lower (δ) and upper (ϵ) ends of the beta scale which is an effect of the sequencing process.

Noise precision (κ): The precision (inverse variance) of the beta-distributed measurement noise of the sequencing. Larger values of κ indicate that observed methylation fractions cluster more tightly around the latent cancer methylation value.

2.3.3.1 Priors

The priors are preset in the EVOFLUX model. The strictly positive parameters have LogNormal priors. Parameters with values between 0 and 1 have beta distributions as priors [2]. The prior scale for the growth-rate parameter was set using `thetastd = 1.0`, which determines the spread of the prior distribution assigned to θ . In EVOFLUX the prior for θ was LogNormal(3,2) but after test runs it was determined that LogNormal(3,1) was more suitable for the data. The initial switching rates and the noise precision use HalfNormal priors, as listed below. These priors, except for the prior for κ , are intended to be weakly informative. The prior for κ is set to represent the typical precision of the methylation measurements [2]. Figure 2.6 shows the shape of each prior.

- Growth rate per year (θ): $\theta \sim \text{LogNormal}(3, 1)$
- Normalized tumor age (τ/T): $\frac{\tau}{T} \sim \text{Beta}(2, 2)$
- Initial switching rates (μ, γ): $\mu, \gamma \sim \text{HalfNormal}(0, 0.05)$
- Normalized switching rates (ν, ζ): $\frac{\nu}{\mu}, \frac{\zeta}{\gamma} \sim \text{LogNormal}(1, 0.7)$
- Normal methylation fraction (β_n): $\beta_n \sim \text{Beta}(2, 2)$
- Array saturation offsets (δ, ϵ): $\delta \sim \text{Beta}(5, 95), \epsilon \sim \text{Beta}(95, 5)$
- Noise precision (κ): $\kappa \sim \text{HalfNormal}(100, 30)$

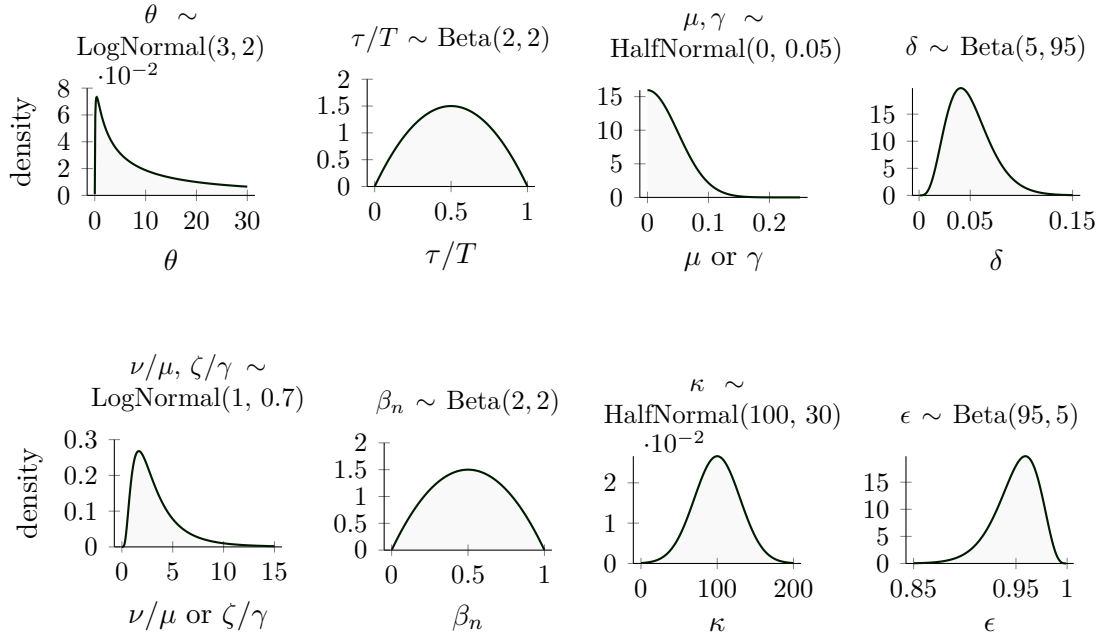


Figure 2.6: Prior distributions for all model parameters.

2.3.4 Sampling

For the sampling EVOFLUX uses the package **dynesty**, which implements nested sampling [2], [16], [17]. Nested sampling begins by drawing a set of **live points**, which in the end will be what makes up the posterior distribution, from the prior distribution. In this case 500 10-dimensional live points, one dimension for each parameter. In each iteration, the live point with the lowest likelihood is excluded. The algorithm then draws an ellipse around the remaining live points and draws a new point from within this ellipse. This ellipse is called a contour. The process is repeated, which builds a sequence of nested likelihood contours and simultaneously estimates the Bayesian evidence, $P(D)$, as the integral of likelihood over prior volume. The process stops when the estimated remaining contribution to the evidence, reported as **dlogz**, falls below a threshold. The surviving live points then approximate the posterior distribution. The number of live points controls how well nested sampling can identify separated regions of high posterior probability. Nested sampling handles multi-modal posterior distributions and provides direct estimates of the Bayesian evidence [16], [17].

For each proposed parameter vector, EVOFLUX runs **NSIM** independent stochastic forward simulations of the three-state methylation process. Each simulation tracks the number of cells in each methylation state rather than simulating individual cells, and produces a possible bulk cancer methylation fraction $\beta_c(t)$ at the sampling time. The simulated fractions are adjusted for tumor purity and the average methylation of normal cells, β_n , and then binned to estimate the latent distribution of cancer methylation fractions, $P(\beta_j \mid \Theta)$. The observed methylation fraction y_i at each fCpG is evaluated under a beta noise model that accounts for array background effects and measurement precision κ . Because the true latent bin is unknown, the

per-fCpG likelihood is the mixture

$$P(y_i | \Theta) = \sum_j P(y_i | \beta_j, \delta, \epsilon, \kappa) P(\beta_j | \Theta).$$

Assuming conditional independence across fCpGs, the total log-likelihood is the sum of the log probabilities over all observed sites, evaluated with a log-sum-exp calculation for numerical stability. A small finite-simulation bias correction is applied because the latent distribution is estimated from a finite number of simulations [2, Supplementary Information].

2.3.5 Model assumptions and simplifications

The model assumes that the tumor growth rate is constant over time. Consequently, the inferred value of θ should be interpreted as an average growth rate over the modelled period. The methylation-state model also only includes transitions between neighbouring methylation states. These assumptions simplify the model and should be considered when interpreting the inferred parameters [2].

2.4 Metrics

2.4.1 Summary statistics

The output from EVOFLUX is not single values, but a posterior probability distribution for each parameter. This distribution describes the relative probability of different values of the parameter and also represents the uncertainty in the inference. A posterior distribution contains more information than a single summary value, but it is more difficult to use directly when comparing samples. Therefore, each inferred parameter can be summarized using four statistics: the median, median absolute deviation (MAD), skewness, and kurtosis.

Median is the 50th percentile, dividing the distribution into two equal halves. In bioinformatics, the median is commonly reported by convention. From a Bayesian perspective, however, the mode, the value with the highest posterior probability, is arguably a more natural point estimate for a unimodal distribution because it represents the most probable value of the parameter. The median was retained here to remain consistent with standard practice in the field, but this choice could be revisited in future work.

The MAD is calculated using the `scipy.stats.median_abs_deviation` function [18]. It is defined in equation 2.3, where $\tilde{X}$ is the median of the data. The MAD measures the spread of the data around the median and is less affected by extreme values than the variance. An example of this is in figure 2.7a. It was selected as a summary statistic because it is directly related to the median which is the main statistic used in EVOFLUX [2].

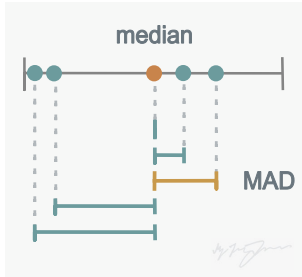
$$\text{MAD} = \text{median}(|X_i - \tilde{X}|) \quad (2.3)$$

Skewness, calculated as equation 2.4 [19], measures the asymmetry of a distribution. Positive skewness indicates a longer right tail, while negative skewness indicates a longer left tail. An example of this is in figure 2.7b

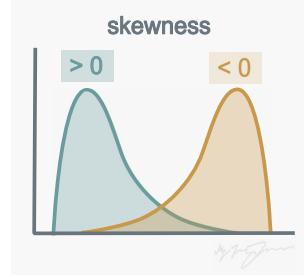
$$\text{skewness} = \frac{E[(X - \mu)^3]}{\sigma^3} = \frac{E[(X - \mu)^3]}{E[(X - \mu)^2]^{3/2}} \quad (2.4)$$

Kurtosis, calculated using the equation 2.5 [19], measures the peakedness of a distribution and the amount of information in its tails. Excess kurtosis subtracts 3, so that the normal distribution has a kurtosis of 0 [18]. High kurtosis means heavier tails and pointier. Low kurtosis means that the distribution is chubbier. An example of this is in figure 2.7c.

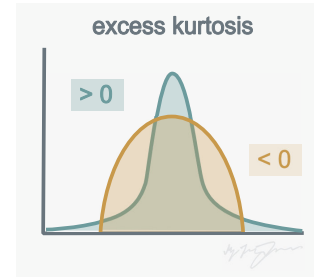
$$\text{excess kurtosis} = \frac{E[(X - \mu)^4]}{\sigma^4} - 3 = \frac{E[(X - \mu)^4]}{E[(X - \mu)^2]^2} - 3 \quad (2.5)$$



(a) Visualization of median and MAD



(b) Visualization of distributions with different skewness.



(c) Visualization of distributions with different excess kurtosis

3

Method

This chapter describes the steps used to process the DNA methylation data, running the simulation and how to investigate the inferred posteriors of evolutionary features of pediatric CNS tumors. First, the clinical and methylation data are described. The raw methylation data are then quality filtered and normalized to reduce technical variation and remove unreliable measurements. Chromosomal regions affected by copy-number variation are excluded because they can influence the methylation distributions used by EVOFLUX.

After preprocessing, fCpG sites are identified separately for the different tumor families. These sites are selected because their methylation patterns contain variation that can be used to study tumor evolution, while sites that may be associated with tumor type, technical effects, or copy-number changes are removed. The selected fCpGs, together with patient age and tumor purity, are then used as input to EVOFLUX. The simulations are run separately for each sample and posterior distributions of the evolutionary parameters are obtained. Finally, the posterior distributions are summarized using the median, median absolute deviation, skewness, and kurtosis.

In the results chapter, these summary statistics are compared between tumor types and age groups and are evaluated in relation to time to first relapse. This approach makes it possible to investigate both the inferred evolutionary parameters and the information contained in the shape of their posterior distributions. The survival analysis is exploratory and is not intended to provide a clinically validated prediction model.

3.1 Data

The data were collected as part of an initiative called Barntumörbanken (BTB) [4]. It consists of clinical data from the patients from whom the samples were collected and methylation data extracted from tumor samples.

The clinical data contain the date of birth, vital status, date of vital status, date of diagnosis, treatment region, predisposing diseases, previous cancer diagnoses, previous radiotherapy, age at diagnosis, clinical diagnosis, the result of the primary tumor surgery, dates of subsequent surgeries, treatment with radiotherapy or chemotherapy, and information about first and subsequent relapses.

3.1.1 Methylation data

The initial methylation data consisted of 156 samples, with approximately 850,000 CpG sites covered by probes. After quality filtering, 770,417 probes remained. The samples were analysed using the Illumina MethylationEPIC v1 array. For each sample, there were two ‘.idat’ files containing the red and green methylation signals. Sample diagnoses had been established using the DNA methylation-based classification of central nervous system tumours Molecular Neuropathology (MNP) framework [10]. Only samples with a family classification score of at least 0.7 were retained. The MNP classifier outputs a calibrated confidence score for each tumor class. A score ≥ 0.7 means that, the classifier assigns at least 70% confidence to the reported tumor family.

3.1.2 Tumor purity

The tumor purity score was estimated using the R package `RF_Purify` with the `ESTIMATE` setting [20]. With the `ESTIMATE` setting the package can estimate the tumor purity of a sample without a reference sample. Tumor purity was used as an input to EVOFLUX because methylation measurements are made on samples containing DNA from both tumor and non-tumor cells.

3.2 Quality filtering

The initial filtering of the methylation data is done to remove probes and samples that have bad quality. The quality filtering that has been done has followed the same steps as in [12]. The library used is `minfi` [14].

Samples with median methylation and unmethylation intensities smaller than 10.5 in log2 scale were discarded. Samples with small intensities were discarded

Samples were read in batches of 20 but evaluated one by one. For each sample raw intensity data were imported as an `RGChannelSet` object and detection p-values were calculated for all probes using the `detectionP` function from `minfi`. A detection p-value of 0.01 was used as the threshold for insufficient signal, consistent with the quality-control procedures described in Kling2019 and the EVOFLUX paper [2], [12]. In this study, samples were excluded when more than 1% of probes had detection p-values greater than or equal to 0.01. In Kling2019 the threshold was set at 10% and the difference to this report is due to an error in the implementation. This should not have too much of an impact on the results because samples with low detection should be hard to classify and probably have a low confidencescore in the MNP classification and therefore be excluded. This sample-level criterion differs from the mean detection p-value criterion used in the original EVOFLUX preprocessing pipeline.

Normalization was subsequently performed using the `preprocessNoob` function with the single-sample approach and dye-bias correction [13]. The single-sample approach was used to avoid batch biases.

Several categories of probes were then removed to reduce technical noise and hidden dependencies. First, probes overlapping known single nucleotide polymorphisms, SNPs, were removed using `dropLociWithSnps` [14]. Probes overlapping SNPs can reflect genetic variation rather than variation in methylation. Probes associated with non-CpG methylation loci were also excluded using `dropMethylationLoci`. Probes mapping to sex chromosomes (X and Y) or lacking a definitive chromosomal mapping were removed to not have a sex-dependent variation. This was done so that genetic differences between male and female patients would not affect the analysis. The `getAnnotation` function contained the chromosomal mapping [14].

Probes known to bind to multiple genomic locations, cross-reactive probes, were filtered out using the `xreactive` function from the `maxprobe` library. This function compares the probes to lists from prior work on the EPICv1 arrays [21].

Following the probe filtering, beta values were extracted using `getBeta` [14].

3.3 Excluding chromosome with CNV

Chromosomal arms with the most variation within each tumor type family were excluded from the fCpG data as described below. The CNV analysis was performed using the `conumee` R package [22]. The algorithm segments each chromosome into bins and estimates the average copy number within each bin based on normalized intensity data.

Since the segmentation results in overlapping bins of varying lengths and copy number values [22], a weighted summary statistic, $\bar{x}_w$, was computed for each chromosomal arm. To account for overlaps, the full range of bins were partitioned into a set of non-overlapping bins defined by the unique sorted boundary points of all segments start and end positions. For each bin, all segments that overlap are identified and their average is computed. This results in a piecewise representation where each bin has a defined length and an assigned mean value. The overall weighted mean for the chromosome is calculated as

$$\bar{x}_w = \frac{\sum_i \ell_i m_i}{L},$$

where ℓ_i represent the length, in base pairs, of bin i , m_i is its corresponding bin copy number average and L is the total length covered by bins of the chromosome [19]. The weighted standard deviation is computed as

$$\sigma_w = \sqrt{\frac{\sum_i \ell_i (m_i - \bar{x}_w)^2}{L}}$$

To distinguish between chromosomal arms, bins overlapping the centromeric region were split such that separate estimates could be obtained for the p and q arms. The weighted mean was then computed independently for each arm, with weights proportional to the bin lengths relative to the total length of the chromosome arm. The R implementation of the weighted mean and the chromosomal-arm processing is provided in Appendix A. Chromosomal arms were excluded separately for each

tumor type when their arm mean was more than two standard deviations from the mean arm value for that tumor type. The excluded arms are shown in Table 3.1.

Tumor type	Excluded chromosome arms
Pilocytic Astrocytoma	5p, 5q
Medulloblastoma, WNT-activated	6p, 6q
Medulloblastoma, SHH-activated	9q, 10q
Medulloblastoma, non-WNT/non-SHH, Group 3	11q
Medulloblastoma, non-WNT/non-SHH, Group 4	8p, 17q
Diffuse midline glioma, H3K27-altered	1q, 14q
Diffuse pediatric-type high-grade glioma, H3-wildtype and IDH-wildtype	4p, 4q, 13q

Table 3.1: Chromosomal arms excluded from the fCpG data for each tumor type based on the two-standard-deviation criterion. The letters p and q denote the short (p) and long (q) arms of the chromosome.

3.4 Finding fCpG sites

The fCpGs were identified using the EVOFLUX function `ef.filter_for_fcpgs` [2]. For each CpG site, the beta values across all samples within the same tumor family were used to compute two quantities: the standard deviation of the beta values and the mean absolute distance of the beta values from 0.5. Sites were retained if their standard deviation was greater than 0.16 and their mean absolute distance from 0.5 was smaller than 0.1. This mean absolute distance should not be confused with the median absolute distance used as a metric to describe the posterior distributions. These criteria select sites that show sufficient variation between samples while remaining centred close to a beta value of 0.5. Without the criteria for standard deviation the algorithm would select sites that only had beta value 0.5 and therefore not the fluctuating ones.

The threshold for the mean absolute distance from 0.5 (0.1) is the default value used by the EVOFLUX filtering function. The standard-deviation threshold (0.16) and the 20% Laplacian-feature-score cutoff were tuned in this project. The standard-deviation threshold was chosen by inspecting the resulting fCpG beta-value histograms and the number of

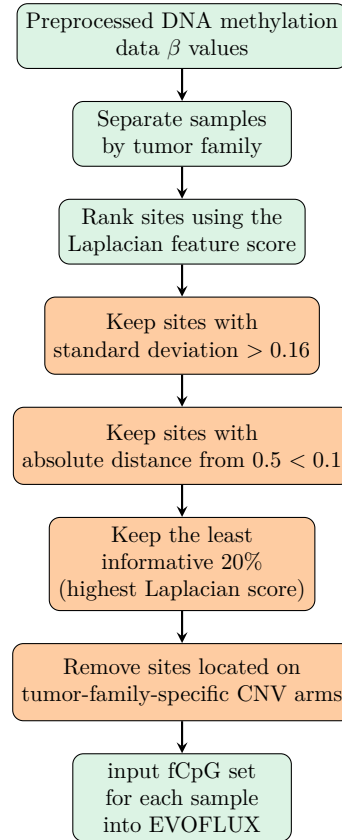


Figure 3.1: Schematic overview of the fCpG selection procedure. Green boxes show data-processing steps, orange boxes show filtering or selection criteria.

retained sites: too low a threshold produced distributions with too few probes at the unmethylated and hypermethylated extremes, while too high a threshold excluded too many probes. The 20% cutoff for the Laplacian feature score was chosen to retain enough fCpGs for reliable EVOFLUX inference, because the original EVOFLUX report indicates that several hundred fCpGs are needed [2]. Using the package defaults without tuning retained fewer than 300 fCpGs in these data, which was considered insufficient. The Laplacian feature score, is a measure used in the EVOFLUX pipeline that ranks CpG sites by how informative they are for separating the tumor families [2]. Sites highly informative for tumor type are removed because they are likely to be influenced by factors other than stochastic methylation fluctuations. Only the 20% which were the least informative were kept. Finally, fCpGs located on chromosomal arms that show frequent copy-number variation were removed. The specific arms excluded for each tumor family are listed in Table 3.1.

3.4.1 Patient age at time of sample

To the patient’s age at sampling 280 days (40 weeks) were added. This adjustment extends the model’s age timeline to allow for tumors that may originate before birth; fetal brain tumors have been reported in the literature [8]. EVOFLUX requires the inferred age at the MRCA, τ , to be positive. Measuring age only from birth would not allow a prenatal origin to be represented. The 280-day offset is an approximation using the conventional gestational-age interval applied in standard due-date calculations, equation 3.1, [23]. The offset was converted to years by dividing by 365.25.

$$\text{AGE_SAMPLING} = \text{AGE} + \frac{280}{365.25} \quad (3.1)$$

3.5 Running simulation

EVOFLUX was run separately for each sample. The fCpGs, tumor purity, and patient age were provided as input for one sample at a time. The simulation therefore produced one posterior distribution for each evolutionary parameter for each sample.

The simulations used the neutral model with 500 live points (`nlive` = 500). For each likelihood evaluation, 3000 stochastic simulations were performed (`NSIM` = 3000). Computations were parallelized across 16 CPU cores (`Ncores` = 16). Sampling was terminated when the estimated remaining uncertainty in the logarithm of the Bayesian evidence fell below 1 (`dlogz` = 1). The resulting posterior samples were subsequently used to calculate summary statistics such as the median, MAD, skewness, and kurtosis.

The choice of 500 live points was conservative for the ten-dimensional parameter space and larger than the 100 live points used in the original EVOFLUX study. Under the common guideline of approximately 50 live points per posterior mode [16]. A mode is a peak in a distribution. 500 live points can represent roughly ten comparably sized modes. Because the original study used the same general

framework with 100 live points, and because additional convergence analyses were not possible within the scope of this project, the choice of 500 live points should be regarded as a conservative methodological setting rather than an empirically validated optimum.

The number of forward simulations was set to `NSIM=3000`, increased from the original 2000 in order to improve the resolution of the estimated likelihood. The stopping criterion `dlogz=1` terminates nested sampling when the estimated remaining contribution to the logarithm of the Bayesian evidence falls below 1. This threshold is coarser than the default of 0.01. It makes the posterior approximation rougher but reduces computation time substantially.

3.6 Analysis of simulation

The output from the simulation is a set of simulated values for each parameter and sample. These values make up the probability distribution of the parameter. In the analysis the median, MAD, skewness and kurtosis for parameter distributions will be used as summary statistics. These statistics will be calculated with the package `scipy.stats` [18]. Using only these statistics still means that a some of the information stored in the shape of the posterior distribution is missed. These statistics doesn't capture multimodality of the posterior distributions.

Throughout the analysis, if not otherwise stated the prenatal state is included in the age. The samples will be split into groups based on age when sample was taken. The younger age groups are patients that were under age of nine and the older age group over the age of nine. There will be a test using SciPy's Kruskal-Wallis to see if there is a significant difference of the inferred parameters in the two age groups [18].

A survival analysis, or time-to-first-relapse analysis, using time between operation and first relapse as the event time will be done. Samples from patients without a documented relapse will be treated as censored observations. θ , τ/T , $\text{Log}(N_e)$, γ , μ , ζ/γ and ν/μ with the different summary statistics will be used as explanatory variables. For the survival analysis the samples will be split into two groups based on the median value of the explanatory variable within each tumor type. The group with explanatory variable higher than the median among the tumor type will be drawn with a solid line and the group with explanatory variable lower than the median will be drawn with a dashed line. The x-axis will show time in years and the y-axis will show the probability of not having a relapse. All lines start at 1 and decrease for each relapse event. If a patient drop out of the study without having a relapse, the drop out is marked with a line. A p-value will be calculated using the log-rank test to see if there is a significant difference between the two lines for each tumor type; a p-value lower than 0.05 indicates that the lines are significantly different at the 5% significance level.

The fCpGs were chosen such that they were uninformative of what tumor type the sample was. If the EVOFLUX simulation catches underlying tumor properties there should be some separation between tumor types within the posterior parameters.

A t-SNE (t-distributed Stochastic Neighbor Embedding) plot will be made to see if the inferred parameters can be used to differentiate between tumor types. t-SNE is a dimensionality-reduction technique that visualizes high-dimensional data in two dimensions by preserving local neighborhoods. This is done as an initial check to see if the simulated parameters hold any information related to tumor type specific properties.

4

Results

4.1 Simulation

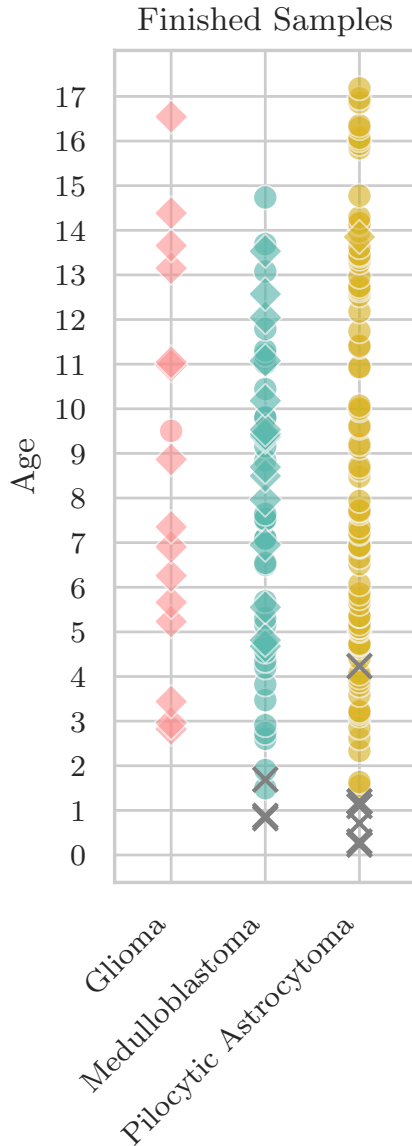


Figure 4.1: Summary of samples that converged during the simulation and patient age distribution. Diamond markers indicate deceased patients and round markers indicate alive patients. X marks non-converged samples.

Age was an input into the simulation, and this indicates that the output from the simulation depends on age. It might also indicate that the model captures general tendencies for growth rather than tumor behaviour. Further tests are necessary to assess the importance of this and validate it.

Figure 4.1 shows an overview over what samples converged in the simulation. The crosses indicate that the samples didn't converge. In figure 4.1 it is visible that only a few samples from patients under the age of two converged.

As mentioned in the Method the t-SNE plots in this section are used as exploratory sanity checks to verify that the inferred parameters contain any type of possibly relevant information and to give a hint of what might be interesting.

In figure 4.2 there are three t-SNE plots. The first is made with the methylation data as it was after preprocessing. There is a clear separation between tumor types. The second figure in figure 4.2 uses the methylation data from the fCpGs, the data methylation data that was used as input into EVOFLUX, and does not have any apparent kind of order. In the third figure each sample uses 50 sampled points from each of its posterior distributions. The samples from Pilocytic Astrocytoma are mostly located in the lower half of the figure with the exception of a smaller group at the very top. The medulloblastoma samples seem to be mainly located in the top, with about half the samples mixed in with the PA samples. On the right side of the figure there is a cluster of MB and Glioma samples, many of whose patients are deceased. The grouping of samples is not as clear as when all the methylation data were used, but some tendencies are still visible.

Figure 4.3 is made with the same data as figure 4.2 but colored by patient age when the sample was taken. The first two images show no apparent pattern. In the third image, the one with inferred data, there is a tendency for samples to be located near samples from patients of similar age. Samples from young patients are mostly in the lower left corner, while samples from older patients are located in the upper right quadrant.

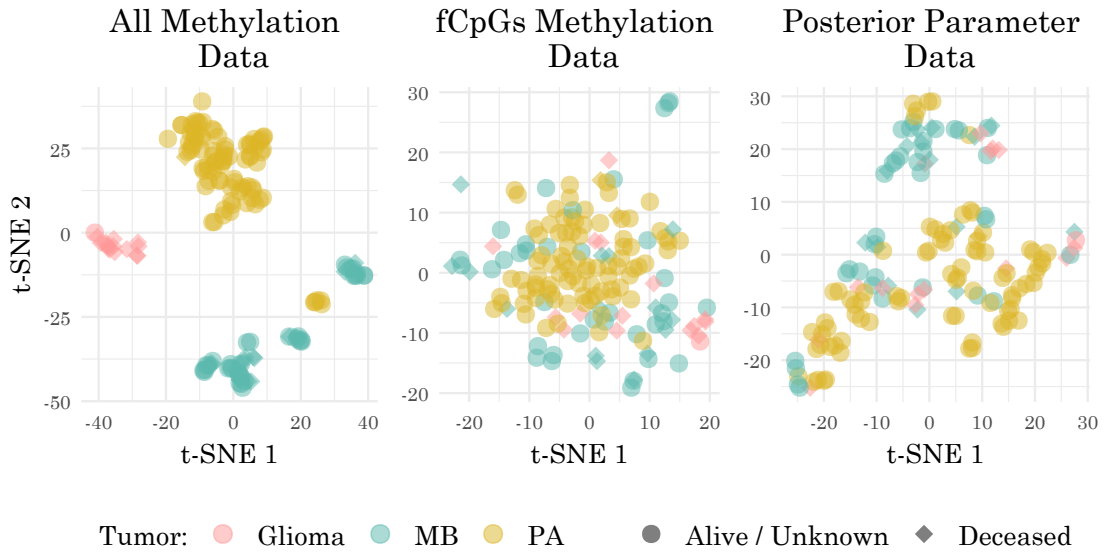


Figure 4.2: t-SNE plots for different data colored on tumor type.¹

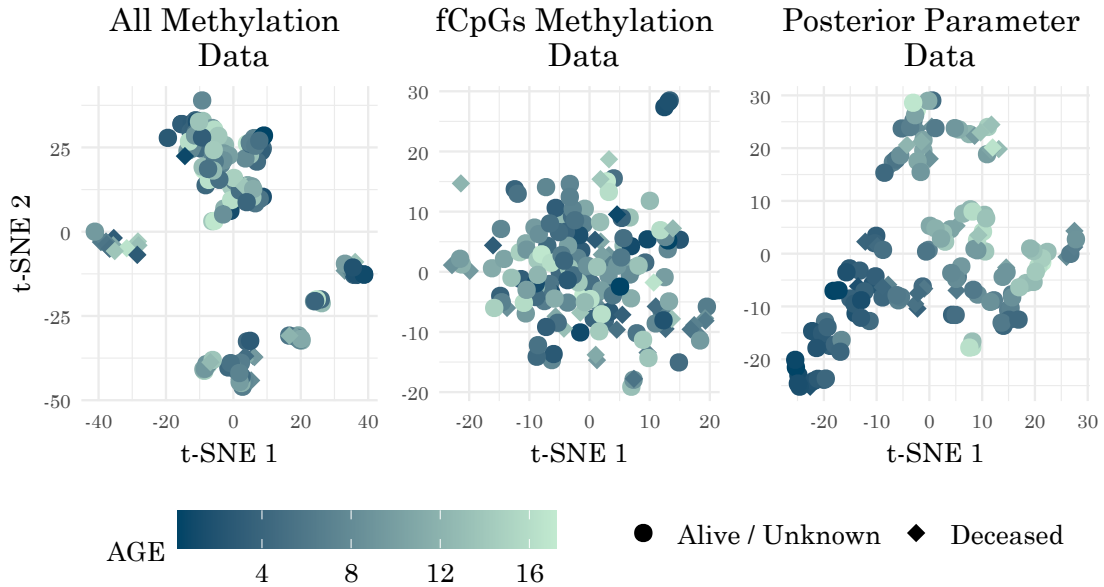


Figure 4.3: t-SNE plots for different data colored by patient age at sampling.¹

In figure 4.4 the samples have been colored based on tumor purity, which also was an input into the model. In the first image, showing the t-SNE based on all methylation data, it is possible to see that tumor purity is strongly related to tumor type, recall figure 4.2. The t-SNE plots show that the EVOFLUX simulation produces outputs that are more informative then what was given as input.

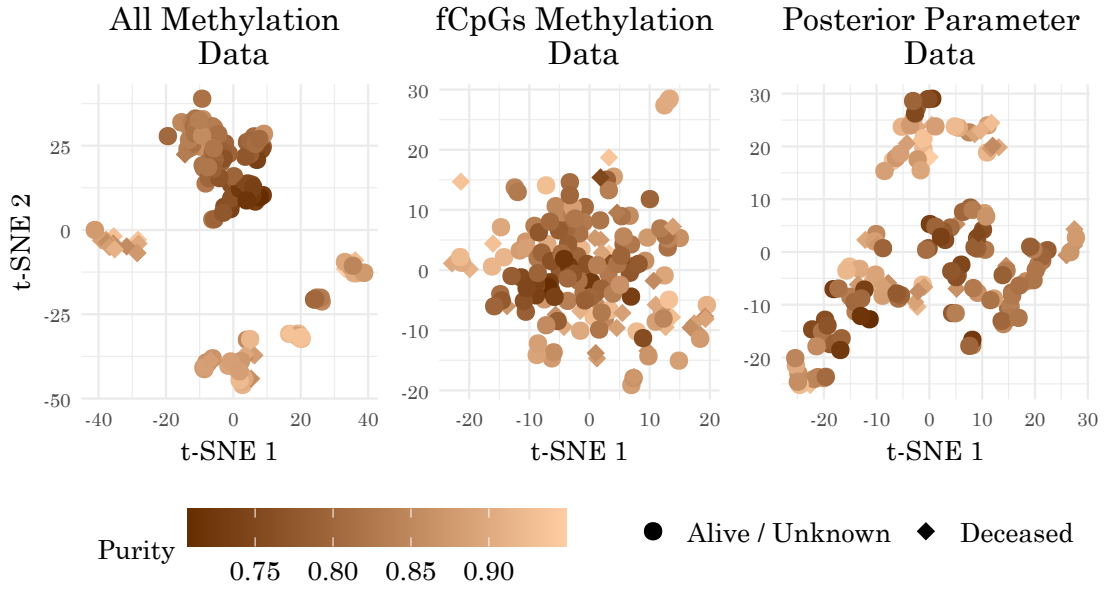


Figure 4.4: t-SNE plots for different data colored by tumor purity of the sample.¹

Figure 4.5 shows θ_{median} for different ages with 95% credible intervals. For samples under the age of nine a steep downward trend is visible. For this group the θ_{median} appears to reflect the age at which the sample was taken rather than any pattern connected to the tumor's properties. The parameter τ/T shows the reversed pattern for the younger samples (figure 4.6): the youngest samples have the lowest τ/T (the smallest percentage of life without a tumor). Although τ/T is age-normalized, a systematic trend with patient age suggests that the model infers tumors to have initiated at a relatively earlier developmental stage in younger patients, or that the model conflates age with tumor biology. The credible intervals are wide for some points, making overlap between tumor types difficult to assess.

¹Patient vital status (alive, unknown, or deceased) was recorded during data preparation but is not relevant to the present analysis. These markers were retained from an earlier version of the figure, before the decision to use time to first relapse rather than overall survival as the clinical endpoint.

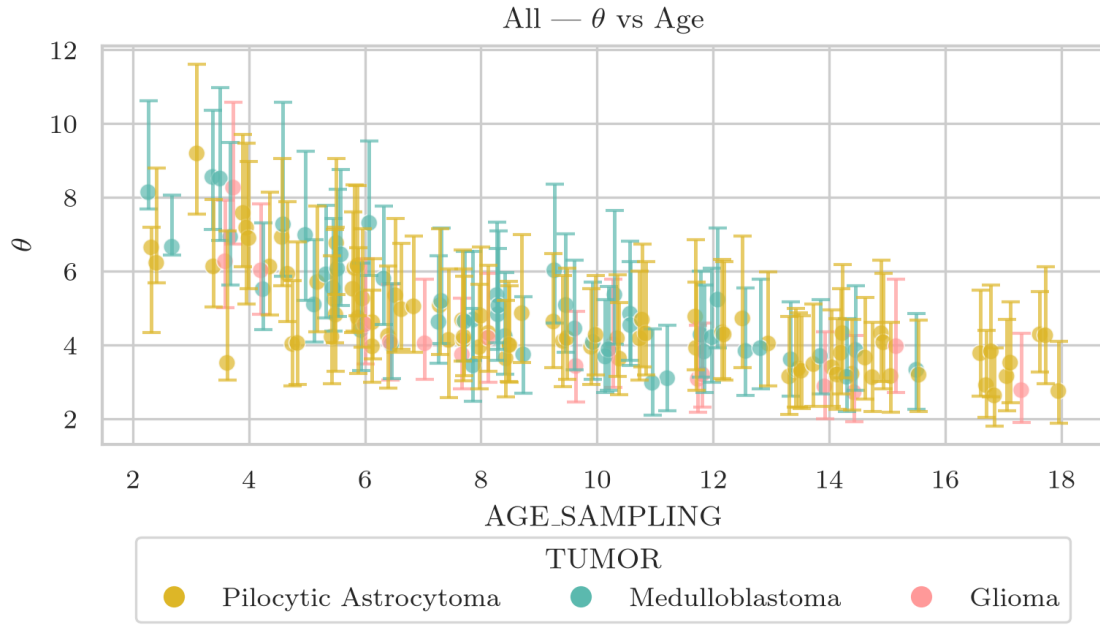


Figure 4.5: Median posterior growth rate (θ_{median}) versus patient age at sampling, with 95% credible intervals. Points are colored by tumor type.

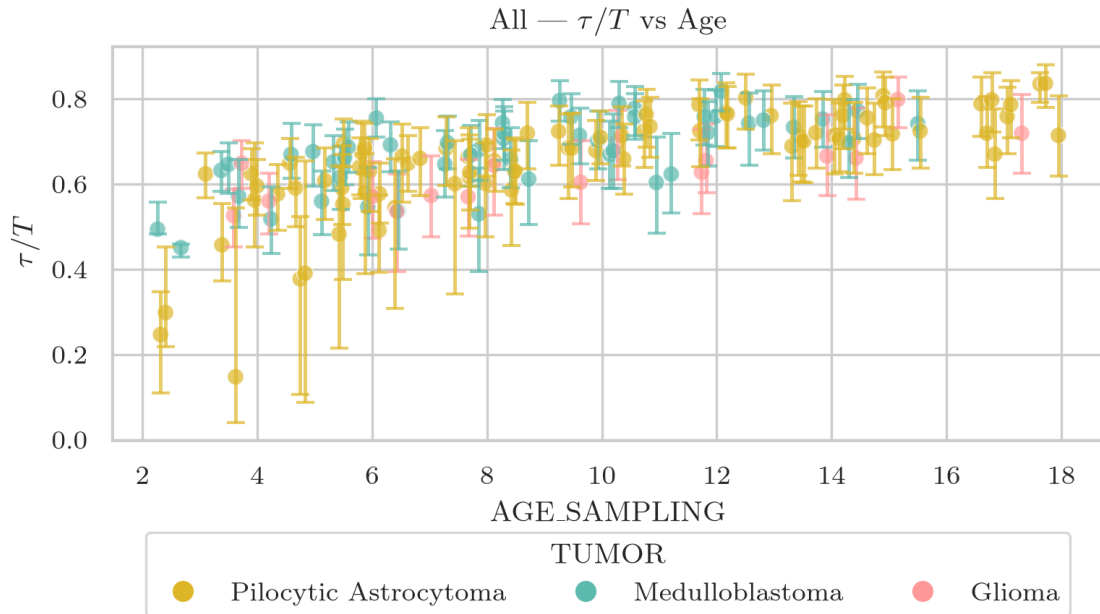


Figure 4.6: Median posterior relative tumor-initiation age (τ/T) versus patient age at sampling, with 95% credible intervals. Points are colored by tumor type.

4.2 Age Dependency

Recall that each tumor type was divided into two age groups—patients under nine and over nine years of age—and that Kruskal–Wallis tests were used to compare the posterior summary statistics between these groups. The results are shown in Table 4.1, where each cell reports the p-value for one parameter-summary-statistic combination and one tumor type. Gray cells indicate p-values ≥ 0.05 , meaning the difference between the age groups is not statistically significant at the 5% significance level. Uncolored cells indicate p-values < 0.05 , meaning the statistic differs significantly between the age groups at the 5% significance level. Only the MAD of τ/T was not age-dependent at the 5% significance level for all tumor groups. This age dependence must be accounted for in downstream analyses. For example, survival comparisons should be performed separately within each age group. The boxplots can be seen in the Appendix in Figures B.1a, B.1b, B.2a, and B.2b.

Table 4.1: Age-separation p-values for posterior summary statistics (patients under nine versus over nine years of age). Gray cells indicate p-values ≥ 0.05 ; uncolored cells indicate p-values < 0.05 .

Median				MAD			
	MB	PA	HGG		MB	PA	HGG
θ	4.5e-05	2.2e-08	1.6e-03	θ	0.213	4.8e-04	0.248
τ/T	8.1e-06	1.2e-13	3.3e-03	τ/T	0.658	0.323	0.345
μ	9.6e-05	5.3e-10	8.7e-03	μ	7.3e-04	2.7e-05	8.7e-03
γ	3.0e-07	1.9e-12	2.3e-03	γ	4.7e-04	1.3e-05	3.2e-03
ν/μ	5.9e-05	0.155	0.0157	ν/μ	0.0423	0.985	0.0357
ζ/γ	0.0244	0.0447	0.0929	ζ/γ	0.0168	0.0713	0.00328
$\text{Log}(N_e)$	3.8e-05	7.2e-10	4.6e-03	$\text{Log}(N_e)$	2.4e-05	1.0e-09	1.1e-03
Skewness				Kurtosis			
	MB	PA	HGG		MB	PA	HGG
θ	0.0331	0.123	0.248	θ	0.296	0.633	0.0209
τ/T	0.763	0.974	0.0209	τ/T	0.244	0.0439	0.0117
μ	1.0e-06	4.2e-11	1.1e-03	μ	9.8e-06	1.5e-06	1.6e-03
γ	7.4e-06	1.3e-08	1.6e-03	γ	9.1e-04	8.7e-09	3.3e-03
ν/μ	0.387	5.5e-04	0.462	ν/μ	0.112	4.7e-03	0.6
ζ/γ	0.325	2.3e-04	0.0742	ζ/γ	0.159	2.3e-03	0.0929
$\text{Log}(N_e)$	5.9e-03	3.2e-06	2.3e-03	$\text{Log}(N_e)$	6.3e-04	4.6e-08	1.6e-03

4.3 Survival Analysis

Survival analyses were made for the two age groups separately using τ/T median, figure 4.7, θ median, figure 4.8, $\log(N_e)$ median, skewness and kurtosis, figures 4.9, 4.10 and 4.11. The MAD of τ/T was also tested for both age groups together, figure 4.12. The median value within each tumor type was used as the cutoff between the high and low groups. The survival probabilities for the different tumor types and age groups are shown in the figures.

For τ_{rel} median only the separation for Medulloblastoma in the older age group was statistically significant, figure 4.7. And higher τ_{rel} was better for survival. In the younger age group, even though the separation was not statistically significant the trend seem to show that lower τ_{rel} was associated with better outcome. For θ median, figure 4.8, none of the tumor types showed a statistically significant separation for any of the age groups. Medulloblastoma however seem to follow a similar trend as for τ_{rel} median, with a lower value better for younger patient and a higher value better for older patients.

Figures 4.9, 4.10, and 4.11 compare time to first relapse using the median, skewness, and kurtosis of the posterior distribution of $\log(N_e)$. In the model, $N_e = e^{\theta(T-\tau)}$, combines the inferred growth rate, θ , with tumor age, $T - \tau$. It is therefore a derived quantity, not an independent parameter. Its posterior distribution reflects the joint posterior uncertainty in these inputs. These comparisons explore whether summaries of this derived posterior are associated with relapse.

Figure 4.9 shows the median of $\log(N_e)$ for the two age groups. In the older group the Medulloblastoma samples show a statistically significant difference between high and low $\log(N_e)$ median for relapses. A low $\log(N_e)$ median make relapse less likely.

Figure 4.10 shows the skewness of $\log(N_e)$ for the two age groups. In the younger group High Grade Glioma show a statistical difference between high and low $\log(N_e)$ skewness. The separation is mostly seen in the p-value and is not as visible as for MB in previous figures. In the older group MB show a statistical difference between high and low $\log(N_e)$ skewness. In both cases a high $\log(N_e)$ skewness is beneficial for survival. A high positive skewness means that the distribution leans more towards the left and that the lower values in the credible interval are more likely than the higher values.

Figure 4.11 shows the excess kurtosis of $\log(N_e)$ for the two age groups. In the younger group PA and HGG show a statistical difference between high and low $\log(N_e)$ excess kurtosis. Both tumor types show that higher excess kurtosis is associated with better survival. A high excess kurtosis means that the distribution is pointier and has heavier tails, which means that the mode value is more likely but also extreme values are more likely than for a normal distribution.

4. Results

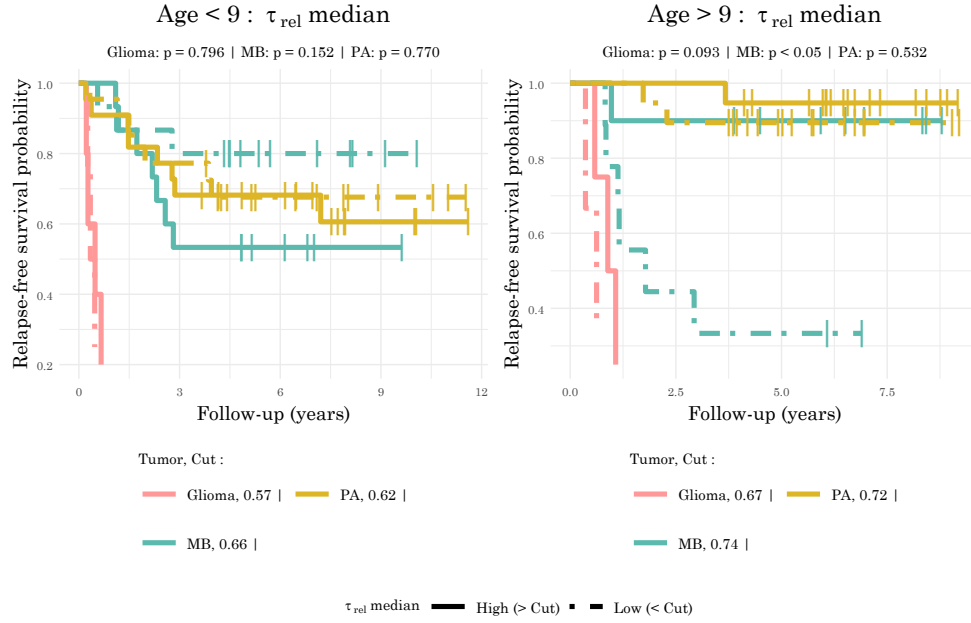


Figure 4.7: Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low τ/T median, in relation to the median within that tumor groups.

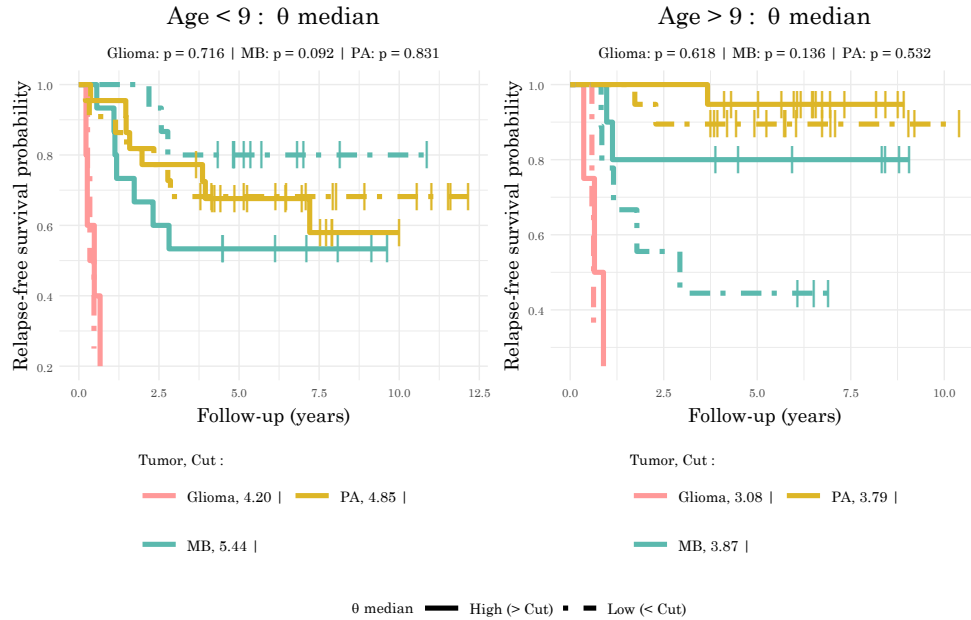


Figure 4.8: Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low θ median, in relation to the median within that tumor type.

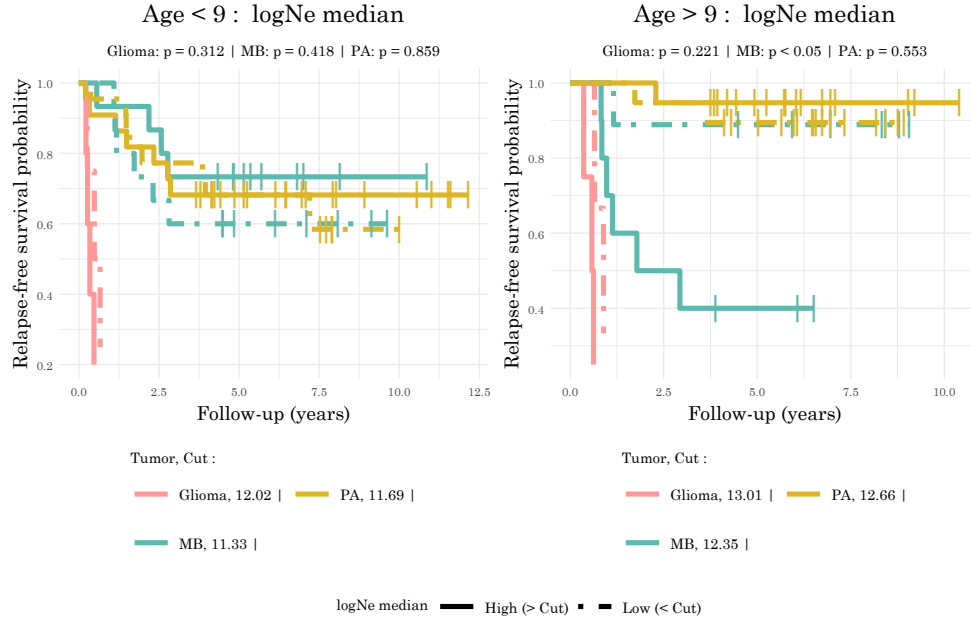


Figure 4.9: Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low $\log(N_e)$ median, in relation to the median within that tumor groups.

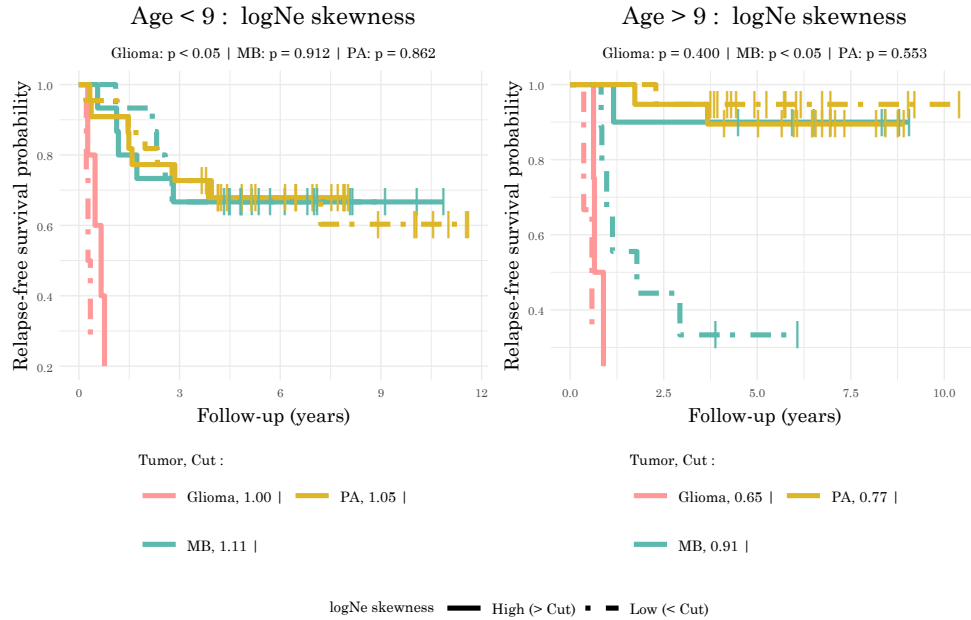


Figure 4.10: Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low $\log(N_e)$ skewness, in relation to the median within that tumor groups.

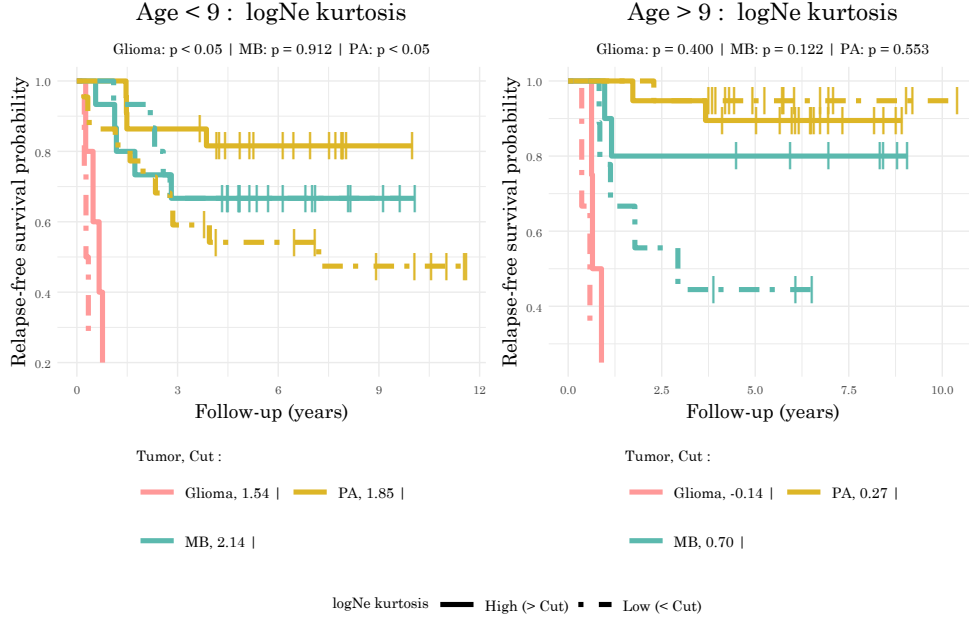


Figure 4.11: Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low $\log(N_e)$ excess kurtosis, in relation to the median within that tumor groups.

The median absolute distans, MAD, for τ/T does not have a significant separation with more then 95% confidence between the age group and because of this there was a survival analysis made for all the samples, figure 4.12. There is no significant separation for any of the tumor types, but all seem to indicate a trend of low τ/T MAD to be good for survival. A low τ/T MAD means that there's fewer possible τ_{rel} values and that it's less likely that τ_{rel} will be extreme, which probably means that the simulation is easier and therefore more reliable given the smaller parameter space.

In summary, different infered parameter and summary statistics are significant for different tumor types and age groups. No significant difference of survival was found based on the switching rates.

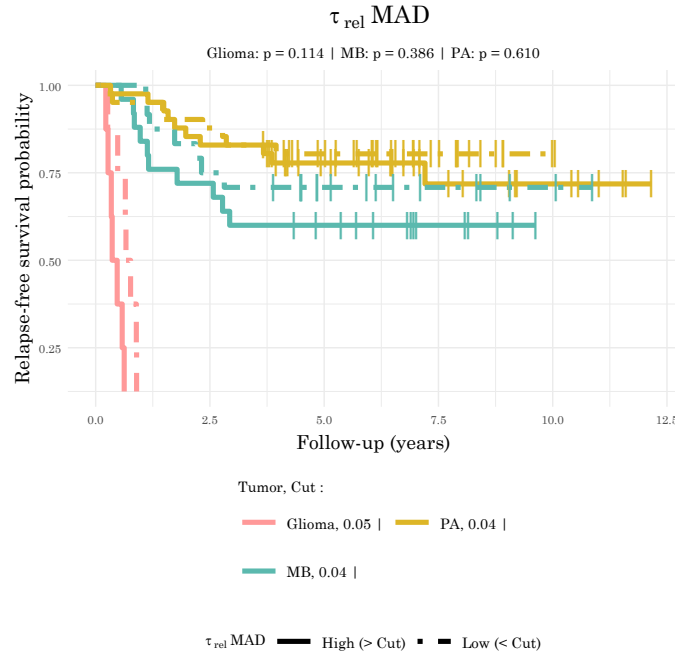


Figure 4.12: Survival plot showing time to first relapse for two age groups. The different tumor types are divide into groups with high or low τ/T MAD, in relation to the median within that tumor groups.

4.4 Additional findings: Inferred switching rates

The switching-rate parameters describe changes between methylation states at the two alleles of a CpG site. The relative switching rates, ν_{rel} and ζ_{rel} , are calculated in relation to the corresponding initial switching rates:

$$\nu_{rel} = \frac{\nu}{\mu}, \quad \zeta_{rel} = \frac{\zeta}{\gamma}.$$

Values below one indicate that the corresponding relative switching rate is lower than its reference rate. The distributions of the relative switching rates for the different tumor types are shown in Figure 4.13. The relative switching rates seem to be consistently below one for all tumor types in the younger age group and closer to one for the older age group.

The median average switching rate is shown in Figure 4.14. The average switching rate appears to decrease with patient age, while the relative switching rates show an increasing tendency with age. These tendencies were not statistically significant for every tumor type. Since age is an input to the EVOFLUX model, the observed age dependence may reflect both biological differences between the samples and the effect of age on the model output.

The results show that switching rates are not independent of patient age. This should be considered when comparing switching rates between tumor types. No significant association between switching rates and time to first relapse was found in the survival analysis.

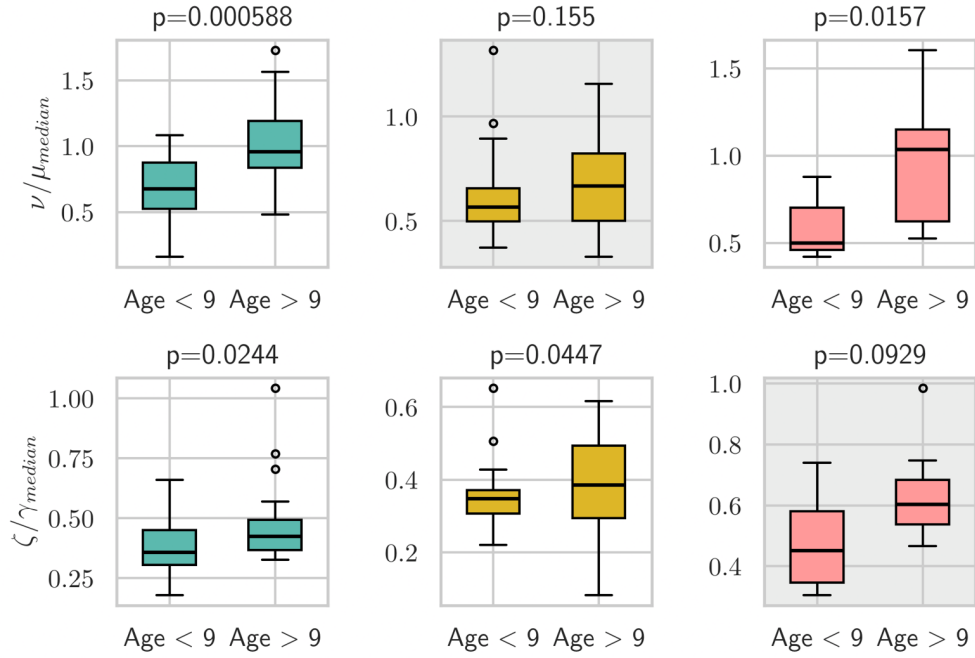


Figure 4.13: Boxplots of the median relative switching rates, ν_{rel} and ζ_{rel} , for the analysed tumor types.

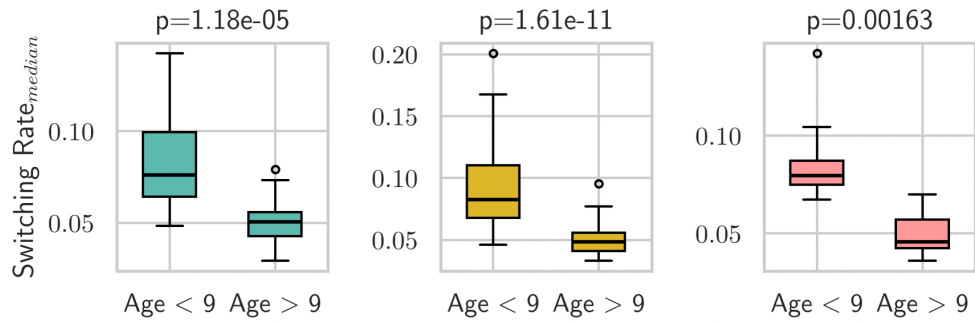


Figure 4.14: Boxplots of the median average switching rates for the analysed tumor types.

5

Discussion

The results showed no consistency in what parameter or summary statistic that was significant for all the tumor types or age groups. The median for τ_{rel} , θ and $\log(N_e)$ showed significant separation for MB but not for the other two tumor types. $\log(N_e)$ skewness and kurtosis for HGG and $\log(N_e)$ kurtosis for PA however showed significant separation for HGG in the younger age group. This should show that only using the median as a summary statistic would leave out possibly important information about the posterior distribution.

The survival groups were split at the median, and this may not be the best cutoff. The number of events was also small in some subgroups: there were only four relapses among older PA patients, and few HGG samples in either age group. As a result, the observed associations are uncertain. Hazard ratios were not calculated or another effect-size estimate, so the Kaplan–Meier comparisons do not tell us how large the difference in relapse risk is.

The survival analysis showed that for older MB patients higher τ_{rel} median was associated with better survival. A high τ_{rel} median would mean that the tumor appeared later in life hence the tumor would have had less time to grow. The results make sense for the older age group. The younger age group however showed a trend of lower τ_{rel} , tumor appearing earlier in life, being better for survival. This seem counter intuitive. For the θ median which also points towards the same trend of flipping high and low for age groups. Higher θ median seems better for survival in the older age group which also seem counter intuitive because a high growth rate should be bad for survival. One explanation for this could be how τ_{rel} and θ interact in the model. A high τ_{rel} could be compensated by a high θ and vice versa. Another explanation could be that when a patient reaches a certain age other treatments are possible and there might be ways that puberty have an effect on the tumor behaviour. I.e. if a tumor uses the same growth mechanisms as normal cells, the tumor should stop growing when the normal cells stop growing. This could explain why it's not beneficial with low growth rate θ for older MB patients, those tumors might not use the same mechanisms as the normal cells and therefore might not be affected by the stopping of growth associated with puberty.

5.1 Interpreting posterior summary statistics in relation to relapse

The survival analyses of skewness, kurtosis and MAD should not be interpreted as showing that these statistical properties themselves cause relapse or correspond directly to a biological process in the tumor. They describe the posterior distributions of quantities inferred by EVOFLUX, rather than directly measured tumor characteristics. These statistics can convey aspects of the model’s uncertainty that are not captured by the median alone: for example, a broad posterior indicates less precise estimates, while an asymmetric or heavy-tailed posterior indicates that some values or alternative inferred histories remain more plausible than others.

For kurtosis of $\log(N_e)$, the survival plots show nominally significant differences between the high- and low-kurtosis groups in the younger HGG and PA groups. They motivate asking whether features of the posterior tails carry information not captured by the median. The younger HGG result is particularly uncertain because the age-specific groups contain very few samples. Its survival estimates and apparent separation may therefore be unstable and should not be treated as reliable evidence of a prognostic association. The apparent tendency toward better outcome with lower MAD of τ/T was also not statistically significant. Overall, these exploratory comparisons across small subgroups and multiple statistics do not establish a biological interpretation or a reliable prognostic marker. Their main value is to motivate validation in larger datasets.

5.2 Switching rates and age

The average switching rate appeared to decrease with patient age, while the relative switching rates showed an increasing tendency with age. These tendencies were not statistically significant for every tumor type. Because age is an input to the EVOFLUX model, the observed age dependence may reflect both biological differences between the samples and the effect of age on the model output. Possible biological explanations include developmental changes in epigenetic maintenance machinery or differences in tumor biology between younger and older children. This study cannot distinguish among these explanations.

5.3 Further Research

5.3.1 Validation and Evaluation

There is currently no known way to establish a ground truth for the parameters inferred by EVOFLUX, since the actual values are not directly observable. It is, however, possible to estimate an upper bound for τ , the age at which the tumor appeared. For patients with multiple relapses, the tumor associated with the second relapse cannot have appeared before the last surgery. Therefore, the patient’s age directly after the previous surgery can be used as a possible upper bound for the

age of the tumor at that time.

The clinical records may also contain information about the size of the tumors that were removed during surgery. These measurements could be compared with the EVOFLUX estimate of $\log(N_e)$, which represents the effective number of tumor cells. There will not be an exact cell count but larger $\log(N_e)$ should correspond to larger tumor size compared to samples with small $\log(N_e)$. Such a comparison could provide an additional way to evaluate the inferred parameters.

The thresholds used to select fCpGs were tuned in an exploratory manner to ensure that enough fCpGs were retained for the EVOFLUX inference to run. A more relaxed approach to choosing the fCpGs was adopted because the primary goal was to test whether the simulation could produce sensible results for pediatric CNS tumors, which was deemed more important than obtaining a fully optimised, 100% reliable result at this stage. A more rigorous approach would systematically compare the effects of different thresholds on the inferred parameters and survival analyses.

The nested-sampling stopping criterion `dlogz=1` is coarser than the default of 0.01 used in the original EVOFLUX study, but it is only twice the `dynesty` package default of 0.5. This setting was chosen to keep computation time manageable while still producing usable posteriors.

The effect of removing chromosomes with CNVs was not tested in this project. It is therefore not known whether this filtering step improves the results or whether it is unnecessary. This could be evaluated by running the analysis with and without the CNV filtering and comparing the inferred parameters and the resulting survival analyses.

5.3.2 Extensions of the Project

The study could be extended by including more samples and additional tumor types. A larger number of samples would also make it possible to divide the patients into smaller age groups and investigate differences between tumor types in more detail.

The analysis could also be extended to other methylation-array types. Many publicly available samples were analysed using the earlier 450K array rather than the EPICv1 array used in this project. It would therefore be useful to investigate whether samples from different array types can be combined after suitable preprocessing. Nanopore sequencing is another possible extension. It has the potential to measure methylation across the whole genome and should cover more fCpG sites compared to other sequencing techniques which focus on sites explicitly connected to tumor regulation.

The importance of the construction parameters used by EVOFLUX could also be investigated. These parameters may affect the inferred posterior distributions and should therefore be evaluated when applying the framework to pediatric CNS tumors.

5.3.3 Gaining Insights into Pediatric CNS Tumors

EVOFLUX also has the option to simulate subclones. Subclones are common in CNS tumors, and including them could make the model more representative of the tumor structure. It would therefore be useful to investigate whether subclonal simulations improve the inference of tumor evolution in these samples.

The relationships between the different summary statistics should also be investigated. If several summary statistics show the same tendencies, they may contain redundant information. On the other hand, summary statistics describing different parts of the posterior distributions may provide complementary information.

Dividing rare tumor types into several age groups may not be possible because of the limited number of samples. One possible alternative would be to estimate the normal growth rate of cells at different ages. The inferred tumor growth rate θ could then be normalized using the expected normal growth rate. If the normal growth rate is not known, the difference between θ and a regression line fitted to all samples could instead be investigated.

6

Conclusions

The aim of this thesis was to investigate if the EVOFLUX framework can be used on DNA methylation data from pediatric CNS tumors. The fCpG selection and filtering was adapted for the tumor samples, and evolutionary parameters were inferred for most of the samples. This shows that EVOFLUX can be used to study evolutionary features in pediatric brain tumors, even though the framework was originally developed for lymphoid malignancies.

Most of the inferred parameters were dependent on the patient's age. This means that age needs to be taken into account when comparing the parameters between samples and when studying their connection to survival. The inferred parameters and their summary statistics were not associated with survival in the same way for all tumor types or age groups.

The clearest survival differences were found for Medulloblastoma which also was the tumor type with the most absolute number of relapses. Younger patients had better survival when the median of θ and τ/T was low. In the older age group, the opposite tendency was observed, where higher medians were associated with better survival. This could be related to differences in tumor growth between younger and older patients, but this explanation needs further investigation.

In younger patients with High Grade Glioma and Pilocytic Astrocytoma, higher kurtosis of $\log(N_e)$ was associated with better survival and for the younger HGG patients higher skewness. Median, skewness and kurtosis showed statistically significant associations in some of the survival analyses. MAD did not show statistically significant survival differences with 95% confidence. However, low MAD for τ/T showed a similar tendency towards better survival for the different tumor types.

The results indicate that no single summary statistic was useful for all tumor types and age groups. The median was easy to interpret, but it was strongly dependent on age. Skewness, Kurtosis and MAD could provide additional information about the inferred distributions, but their usefulness needs to be tested with more samples.

The number of samples was small for some tumor types, especially HGG, and there were only a few relapses among older patients with PA. The median was also used to divide the samples into high and low summary statistics groups and this might not either be the optimal cutoff.

In conclusion, EVOFLUX can be used to infer evolutionary features from pediatric CNS tumor methylation data. Some of these features showed a connection to relapse

6. Conclusions

free survival, but the results depended on both patient age and tumor type. More samples and independent clinical data are needed to validate the inferred parameters and determine if they can be useful for predicting relapse.

Bibliography

- [1] L. Harrington, I. Tannock, R. P. Hill, and D. Cescon, *The Basic Science of Oncology*. McGraw Hill, 2021, p. 612, ISBN: 9781259862076.
- [2] C. Gabbutt et al., “Fluctuating dna methylation tracks cancer evolution at clinical scale,” *Nature*, vol. 645, pp. 764–773, 8081 Sep. 2025, ISSN: 14764687. DOI: 10.1038/s41586-025-09374-4
- [3] C. Lengauer, *Dna methylation*, <https://www.accessscience.com/content/article/a801140>, Accessed: February 2026, 2023. DOI: 10.1036/1097-8542.801140
- [4] Institutionen för Onkologi-Patologi Karolinska Institutet, *Barntumörbanken*. [Online]. Available: <https://www.barntumorbanken.se/>
- [5] KES47, *Chromosome en.svg*, Accessed on: Mar 14, 2026. [Illustration]. https://commons.wikimedia.org/wiki/File:Chromosome_en.svg.
- [6] J. L. Doob, “Stochastic process,” *AccessScience*, 2026. DOI: 10.1036/1097-8542.656600 [Online]. Available: <https://www.accessscience.com/content/article/a656600>
- [7] S. F. Vega, “Dna methylation profiling of cns tumors; implications for clinical diagnostics,” Ph.D. dissertation, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, 2022, ISBN: 978-91-8009-900-4.
- [8] H. J. Milani et al., “Fetal brain tumors: Prenatal diagnosis by ultrasound and magnetic resonance imaging,” *World Journal of Radiology*, vol. 7, p. 17, 1 2015, ISSN: 1949-8470. DOI: 10.4329/WJR.V7.I1.17 [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4295174/>
- [9] T. T. Brown and T. L. Jernigan, “Brain development during the preschool years,” *Neuropsychology Review 2012 22:4*, vol. 22, pp. 313–333, 4 Sep. 2012, ISSN: 1573-6660. DOI: 10.1007/S11065-012-9214-1 [Online]. Available: <https://link.springer.com/article/10.1007/s11065-012-9214-1>
- [10] D. Capper et al., “DNA methylation-based classification of central nervous system tumours,” *Nature*, vol. 555, no. 7697, pp. 469–474, 2018. DOI: 10.1038/nature26000
- [11] D. N. Louis et al., “The 2021 who classification of tumors of the central nervous system: A summary,” *Neuro-Oncology*, vol. 23, pp. 1231–1251, 8 Aug. 2021, ISSN: 1522-8517. DOI: 10.1093/neuonc/noab106
- [12] T. Kling and H. Carén, “Methylation analysis using microarrays: Analysis and interpretation,” *Methods in Molecular Biology*, vol. 1908, pp. 205–217, 2019, ISSN: 10643745. DOI: 10.1007/978-1-4939-9004-7_14

- [13] T. J. Triche, D. J. Weisenberger, D. Van Den Berg, P. W. Laird, and K. D. Siegmund, “Low-level processing of Illumina Infinium DNA methylation BeadArrays,” *Nucleic Acids Research*, vol. 41, 7 Apr. 2013, ISSN: 03051048. DOI: 10.1093/nar/gkt090
- [14] M. J. Aryee et al., “Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays,” *Bioinformatics*, vol. 30, p. 1363, 10 May 2014, ISSN: 14602059. DOI: 10.1093/bioinformatics/btu049
- [15] S. M. Lynch, “Basics of Bayesian statistics,” in *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, ser. Statistics for Social and Behavioral Sciences, S. M. Lynch, Ed., New York, NY: Springer, 2007, pp. 47–73. DOI: 10.1007/978-0-387-71265-9_3
- [16] J. S. Speagle, “Dynesty: A dynamic nested sampling package for estimating Bayesian posteriors and evidences,” *Monthly Notices of the Royal Astronomical Society*, vol. 493, pp. 3132–3158, 3 Apr. 2020. DOI: 10.1093/mnras/staa278
- [17] S. Kopolov et al., *JoshSpeagle/dynesty: V3.0.0*, version v3.0.0, Oct. 2025. DOI: 10.5281/zenodo.17268284
- [18] P. Virtanen et al., “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2
- [19] L. Råde and B. Westergren, *Mathematics Handbook: for Science and Engineering* (Mathematics and Statistics). Springer Berlin Heidelberg, 2013, ISBN: 9783662035566.
- [20] P. D. Johann, N. Jäger, S. M. Pfister, and M. Sill, “Rf_purify: A novel tool for comprehensive analysis of tumor-purity in methylation array data based on random forest regression,” *BMC bioinformatics*, vol. 20, p. 428, 1 Aug. 2019, ISSN: 1471-2105. DOI: 10.1186/s12859-019-3014-z
- [21] R. Pidsley et al., “Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling,” *Genome Biology*, vol. 17, p. 208, 1 Oct. 2016, ISSN: 1474760X. DOI: 10.1186/s13059-016-1066-1
- [22] B. Daenekas et al., “Conumee 2.0: Enhanced copy-number variation analysis from dna methylation arrays for humans and mice,” *Bioinformatics*, vol. 40, 2 Feb. 2024, ISSN: 13674811. DOI: 10.1093/BIOINFORMATICS/BTAE029
- [23] American College of Obstetricians and Gynecologists, “Methods for estimating the due date,” *Obstetrics & Gynecology*, vol. 129, no. 5, e150–e154, 2017, Committee Opinion No. 700. DOI: 10.1097/AOG.0000000000002046

A

CNV analysis implementation

The following R code (extracted from `CNV_analysis.R`) implements the length-weighted mean and standard deviation calculation, as well as the per-chromosomal-arm processing used to identify arms with copy-number variation. The first block shows the `w_overlap_mean` helper and the start of the main `CNV` function (lines 71–194). The second block shows the summary-statistic extraction and file export (lines 212–242).

Listing A.1: R implementation of the weighted-overlap mean and chromosomal-arm processing.

```
w_overlap_mean <- function(loc.start, loc.end, seg.mean){
  segments <- data.frame(
    start = loc.start,
    end = loc.end,
    mean = seg.mean
  )
  tot_len <- max(loc.end) - min(loc.start)

  br_points <- sort(unique(c(loc.start, loc.end)))

  bins <- data.frame(
    bin_start = br_points[- length(br_points)],
    bin_end = br_points[-1]
  )

  bins <- bins %>%
    dplyr::mutate(bin_len = bin_end - bin_start) %>%
    rowwise() %>%
    dplyr::mutate(
      bin_mean = {
        overlapping <- segments %>%
          filter(start < bin_end & end > bin_start)
        mean(overlapping$mean)
      }
    ) %>%
    ungroup()

  mean <- sum(bins$bin_len * bins$bin_mean) / tot_len
```

A. CNV analysis implementation

```
sd = sqrt(sum(bins$bin_len * (bins$bin_mean - mean)^2) /
  tot_len)

return(c(mean,sd))
}

CNV <- function(target_tumor_type, tumor_name, samplesheet,
  segments_df, centromere_pos, getFullData=FALSE){

  tumor_dir <- file.path(output_base_dir, tumor_name)
  if (!dir.exists(tumor_dir)) dir.create(tumor_dir, recursive
    = TRUE)

  message(paste("Processing:", target_tumor_type))

  ids_for_tumor <- samplesheet %>%
    filter(CELL_TYPE == target_tumor_type) %>%
    pull(Sentrix_ID)

  tumor_data <- segments_df %>% filter(Sentrix_ID %in% ids_
    for_tumor)

  if(nrow(tumor_data) == 0) return(NULL)

  #-----
  # Prepping
  #-----

  chrom_levels <- c(paste0("chr", 1:22))

  plot_data <- tumor_data %>%
    mutate(chrom = factor(chrom,
      levels = c(paste0("chr", 1:22))),
      chr_num = match(chrom, chrom_levels)) %>%
    filter(!is.na(chrom)) %>%
    arrange(chrom, loc.start) %>%
    dplyr::mutate(
      seg.len = loc.end - loc.start
    ) %>%
    dplyr::group_by(chrom) %>%
    dplyr::mutate(
      chr_length = max(loc.end), # Get chromosome length
      chr_mean = w_overlap_mean(loc.start, loc.end, seg.mean)
        [1],
      chr_sd_mean = w_overlap_mean(loc.start, loc.end, seg.
        mean)[2]
    )
}
```

```

chr_offsets <- plot_data %>%
  dplyr::group_by(chr_num) %>%
  dplyr::summarise(chr_length = unique(chr_length)) %>%
  dplyr::arrange(chr_num) %>%
  dplyr::mutate(
    chr_offset = cumsum(as.numeric(lag(chr_length, default
    = 0)))
  )

plot_data <- plot_data %>%
  arrange(chrom, loc.start) %>%
  left_join(centromere_positions, by="chrom") %>%
  left_join(chr_offsets, by = "chr_num") %>%
  dplyr::group_by(chrom) %>%
  dplyr::mutate(
    arm = case_when(loc.end <= chromStart ~ "p",
                    loc.start >= chromEnd ~ "q",
                    TRUE ~ "c"
  ),
  )

plot_data <- plot_data %>%
  dplyr::group_by(chr_num) %>%
  filter(arm != "c") %>%
  bind_rows(
    plot_data %>%
      filter(arm == "c") %>%
      dplyr::mutate(
        loc.end = chromStart,
        seg.len = loc.end - min(loc.start),
        arm = "p"
      ),
    plot_data %>%
      filter(arm == "c") %>%
      dplyr::mutate(
        loc.start = chromEnd,
        seg.len = max(loc.end) - loc.start,
        arm = "q"
      )
  ) %>%
  dplyr::mutate(
    chromLoc = paste0(chr_num, arm),
    midpoint = (loc.start + loc.end) / 2,
  ) %>%
  ungroup() %>%
  dplyr::mutate(
    cumulative_pos = midpoint + chr_offset
  ) %>%
  dplyr::group_by(chr_num, arm) %>%

```

```

    dplyr::mutate(
      arm_mean = w_overlap_mean(loc.start, loc.end, seg.mean)
        [1],
      arm_sd_mean = w_overlap_mean(loc.start, loc.end, seg.
        mean)[2],
      arm_abs_mean = abs(arm_mean)
    ) %>%
chr_centers <- plot_data %>%
  dplyr::group_by(chrom) %>%
  dplyr::summarize(
    center = (mean(chromEnd) - mean(chromStart))/2 + mean(
      chr_offset) + mean(chromEnd)
  )

# Defining thresholds for middle of arms
lower_abs_mean <- summary(plot_data$arm_abs_mean)[2][1]
upper_abs_mean <- summary(plot_data$arm_abs_mean)[5]
max_sd <- summary(plot_data$arm_sd_mean)[5]

lower_mean <- summary(plot_data$arm_mean)[2][1]
upper_mean <- summary(plot_data$arm_mean)[5]

processed_data <- plot_data %>%
  dplyr::group_by(chromLoc) %>%
  dplyr::summarize(
    n_probes = length(chromLoc),
    arm_mean = unique(arm_mean),
    arm_abs_mean = abs(unique(arm_mean)),
    arm_sd_mean = unique(arm_sd_mean),
    chr = unique(chromLoc), # assuming you have a chr
      column
    .groups = "drop"
  ) %>%
  arrange(arm_abs_mean)

write.csv(processed_data, file.path(tumor_dir, "processed_
  arms.csv"), row.names = FALSE)

if(getFullData==TRUE) write.csv(plot_data, file.path(tumor_
  dir, "full_CNV_analysis.csv"), row.names = FALSE)
}

```

B

Boxplots showing age separation

B. Boxplots showing age separation



Figure B.1: Boxplots showing separation between two age groups for summary statistics median and MAD for all inferred parameters and all tumor types. Images with gray background have a p-value over 0.05. Images with white background have a p-value under 0.05, indicating that the results did not just happen by chance.

B. Boxplots showing age separation

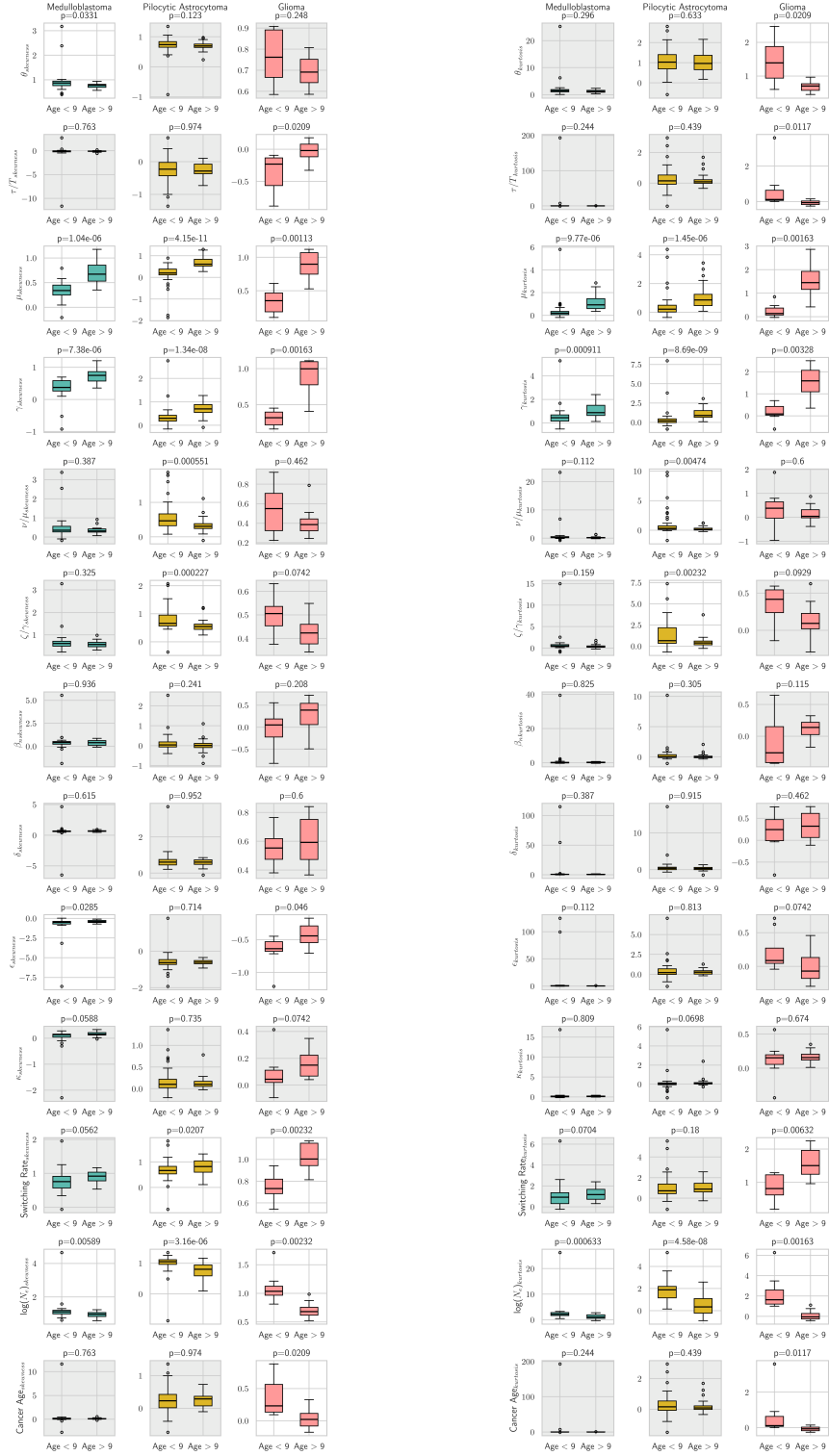


Figure B.2: Boxplots showing separation between two age groups for summary statistics skewness and kurtosis for all inferred parameters and all tumor types. Images with gray background have a p-value over 0.05. Images with white background are with 95% probability actually separated and the results did not just happen by chance.