



Real-Time Reads, Real-World Impact: Studying Tumor Classification with Nanopore and the Surgeon Classifier

Combining Data Science and Biology

Manasa Praveen

University of Massachusetts, Lowell

Interning at the Hovestadt Lab, Dana Farber Cancer Institute



Table of Contents

| | |
|---|-----------|
| Table of Contents..... | 1 |
| Interning at Dana Farber Cancer Institute..... | 2 |
| Overview..... | 2 |
| Background..... | 3 |
| Methylation..... | 3 |
| What?..... | 3 |
| How?..... | 3 |
| Why?..... | 3 |
| Nanopore Sequencing..... | 3 |
| What?..... | 3 |
| How?..... | 4 |
| Sturgeon..... | 4 |
| What?..... | 4 |
| Why?..... | 4 |
| Real World Application..... | 5 |
| Technical Workflow..... | 5 |
| Dorado..... | 5 |
| Samtools..... | 6 |
| Modkit..... | 6 |
| Sturgeon..... | 7 |
| Visual Snapshot..... | 8 |
| Integrative Genomics Viewer (IGV)..... | 8 |
| Challenges Faced..... | 9 |
| I. Understanding Nanopore Sequencing..... | 9 |
| II. Basecalling Blunder..... | 9 |
| III. T2T Genome Download Glitch..... | 9 |
| Key Takeaways..... | 9 |
| Interdisciplinary Collaboration is Powerful..... | 9 |
| Refining Bash/Terminal Skills..... | 10 |
| Improved Problem-Solving Skills..... | 10 |
| Reflection..... | 10 |
| Bibliography..... | 11 |



Interning at Dana Farber Cancer Institute

Over the summer I had been given the opportunity to take on a summer internship at the Dana Farber Cancer Institute. I worked at the Hovestadt Lab, working directly under the mentorship of Dr. Volker Hovestadt and Dr. Salvatore Benfatto. The Lab is focused on Pediatric Oncology and is renowned for its contributions to extensive research on DNA methylation in brain tumors. My internship work focused on understanding the emerging nanopore sequencing technology and the Sturgeon classifier that has shown exciting promise in improving speed and accuracy in subclassification of central nervous system tumors.

Overview

There exists over 120 different types of brain tumors, which are usually classified based on the type of cell they originate from and where they are located within the brain. Further classifications about the malignancy of tumors and where they fall on the World Health Organization grade can also be noted. Currently, techniques that include MRI scans, tissue analysis, and methylation arrays like the Illumina 450K array dominate brain tumor diagnostics. However, all of these tools have particular drawbacks that prevent them from being efficient diagnostic techniques.

MRI scans are extremely useful for detecting abnormalities within the brain, but fall short in its ability to classify the brain tumor type. Histopathological analysis, which is considered the “gold standard” for brain tumor type classification, is heavily reliant on the pathologists’ subjective interpretation of tumor cell structures. Lastly, currently employed methylation arrays, despite their ability to examine molecular data very closely, are both time and cost inefficient. Analysis from the 450K array can take anywhere from 7-14 days, delaying diagnosis.

With the existence of numerous brain tumor types, understanding the subtype and grade is extremely important to administer proper treatment in a timely manner. Research is being thoroughly conducted to present a new way of classifying brain tumors—nanopore sequencing paired with machine learning classifiers like Sturgeon. This new approach is a potential breakthrough that trumps all methods by providing fast, inexpensive, and accurate diagnosis in less than 2 hours.



Background

Methylation

What?

Methylation is a special kind of chemical modification of DNA. In DNA, methylation tags can alter gene expression. In the process of methylation, chemical tags called methyl groups attach themselves to certain nucleotide pairs to either suppress or encourage the corresponding gene's expression. Methylation patterns are especially useful to study the molecular DNA patterns within a cell. In the case of central nervous system tumor classification, we are particularly interested in examining the methylation patterns in a Cytosine-phosphate-Guanine (CpG) context.

How?

The methyl group (CH₃) is added to a regular cytosine to create a methylated cytosine (5mC). We then look for places in the DNA where these methylated cytosines are paired with guanine bases to form 5mCG.

Why?

Extensive DNA methylation profiling study has shown DNA methylation for more than 80% of the CpG sites in human brains. Realizing the repeated occurrence of methylation patterns in a CpG context, researchers have shifted focus on studying the manner in which these chemical patterns can indicate and classify different brain tumor types.

Nanopore Sequencing

What?

Nanopore sequencing is an emerging diagnostic process that is seeing rapid development. Oxford Nanopore Technologies was founded in 2005 and released its first nanopore sequencing device in 2014, making real-time analysis of native DNA accessible. Since then several different research efforts have been made to improve the accuracy of the diagnosis. The employment of nanopore sequencing is so beneficial in oncology diagnosis specifically concerning brain tumors, saving us time, money, and creating ease.



How?

Nanopore sequencing devices use flow cells with an embedded membrane to form stable, consistent pores at the molecular level. Each nanopore has its own electrode connected to a channel and sensor chip, which measures the electric current that flows through the nanopore. When a DNA strand passes through the nanopore, it causes disruptions in the flow of electricity. These disruptions are measured as signals that are then fed to the computer to examine the waves and distinguish the various nucleotides, uncovering the DNA sequence.

Why?

The employment of nanopore sequencing is so beneficial in oncology diagnosis specifically concerning brain tumors, saving us time, money, and creating ease. Nanopore sequencing devices like the MinION are able to provide real time reads with a raw read accuracy rate of 99.2% in 5mC/5hmC modifications in the CpG molecular contexts.

Sturgeon

What?

Published in 2023 by researchers at the Princess Máxima Center for Pediatric Oncology, OncoCode Institute, and UMC Utrecht, Sturgeon is a powerful tool that processes DNA reads with methylation patterns to predict brain tumor type.

How?

Sturgeon classifier was trained on approximately 36.8 million simulated nanopore sequencing runs and further validated on 4.2 million simulated runs. This allows the machine learning based tool to take sparse data and profile the molecular subclassification of brain tumor type without needing patient-specific training.

Why?

Sturgeon is able to produce robust diagnoses with high confidence in as little as 40 minutes. When applied intraoperatively, nanopore sequencing and the Sturgeon classifier work together to produce a classification within 90 minutes. Hence, making Sturgeon classifier a cutting edge diagnostic tool that can address the current time inefficiencies other diagnostic techniques have.

Real World Application

To highlight the potential of this technology, consider the case of a 74-year-old patient with neurological symptoms and no clear diagnosis through conventional methods. Despite multiple tests, including cerebrospinal fluid (CSF) cytology, results remained inconclusive. However, using a mere 2mL of cerebrospinal fluid and 100ng of DNA, nanopore sequencing produced 18 million reads. These reads detected methylation signals pointing to IDH-wildtype glioblastoma. This diagnosis, confirmed later by other molecular tests, helped guide the decision for palliative care—providing clarity to both patient and clinicians when it mattered most.

Technical Workflow

The process of starting from signals obtained by the nanopore sequencing device to the subclassification of the brain tumor type using the Sturgeon classifier features the following softwares: Dorado, Samtools, Modkit, and finally Sturgeon.

Dorado

Dorado is a high-performance, easy-to-use, open source analysis engine for Oxford Nanopore reads. Dorado is a vital component in the pipeline responsible for analyzing the signals produced by the nanopore device to then convert them to nucleotides.

Dorado Basecaller

Command:

```
dorado basecaller dna_r10.4.1_e8.2_400bps_hac@v5.2.0  
--modified-bases 5mCG_5hmCG --kit-name SQK-  
RBK114-24 --reference input.pod5 > calls.bam
```

Figure 1: Dorado Basecaller Command

The demultiplexer then splits these sequences into separate files for each barcode. Essentially generating files for each of the patients.

The first step involves Dorado Basecaller which converts pod5 files into readable DNA sequences. In addition to decoding the nucleotides, it also detects methylation, and effectively recognizes the barcode sequences pertaining to different patients.

Dorado Demultiplexing

Command:

```
dorado demux --output-dir /demultiplexing/output  
--no-classify calls.bam
```

Figure 2: Dorado Demultiplexing Command

Samtools

Samtools is a command-line software that helps manage and process sequencing data, particularly in BAM (Binary Alignment Map) formats. It's a core part of our analysis pipeline because it lets us sort, index, and inspect our read alignments efficiently.

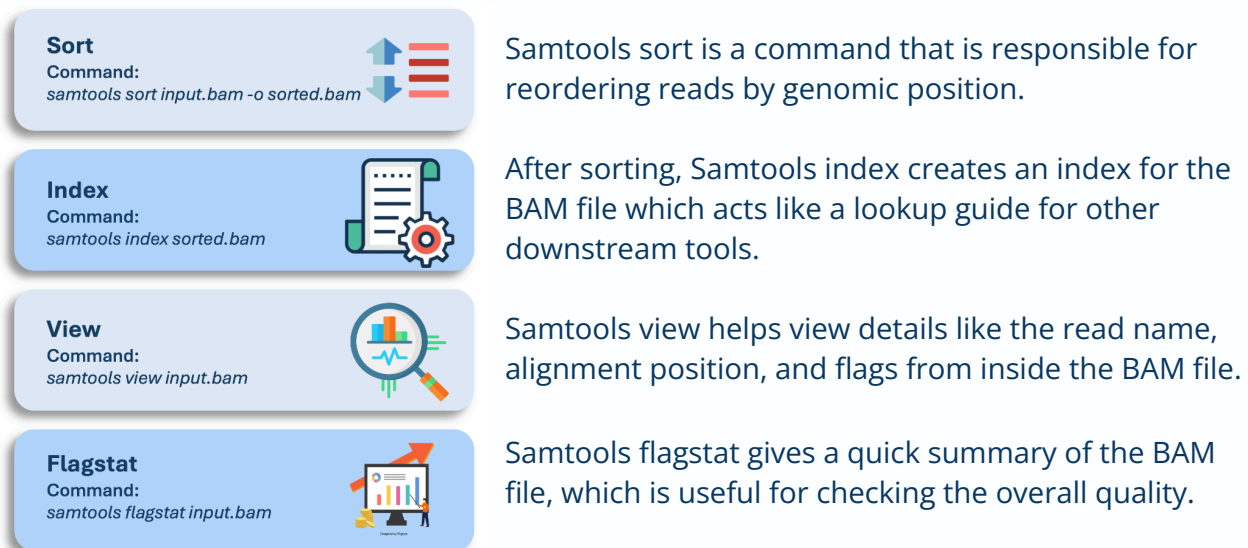


Figure 3: Samtools Commands

Modkit

Modkit is a command-line tool designed to extract methylation information from aligned reads. This tool allows us to focus specifically on isolating modified bases, such as methylated cytosines.

We used two Modkit commands in our workflow:

Modkit Adjust-Mods converts the 5hmC methylation modifications to 5mC. The Sturgeon classifier is trained on 5mCG data so the conversion ensures the analysis is focused and consistent.

Once the modifications are standardized, the **Modkit Extract** extracts the methylation information into a text file.

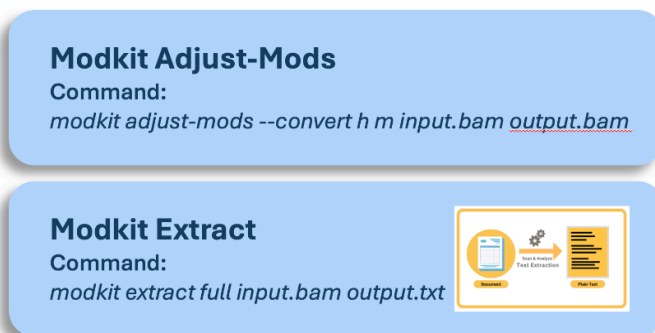


Figure 4: Modkit Commands

This presents a comprehensive output with columns detailing unique read identifier, genomic position of read, chromosome name, methylation likelihood, methylation type for each CpG site detected. The table below displays details from the output.



| Read_ID | Ref_Position | Chrom | Read_Length | Mod_Qual | Mod_Code |
|--------------------------------------|--------------|-------|-------------|-----------|----------|
| a34637d6-adca-423a-b263-6a0b0ed62458 | 38545250 | chr21 | 862 | 0.9980469 | m |
| 6515131f-f6da-46d9-a0b0f23e27fb7bbf | 45674746 | chr5 | 3295 | 0.9980469 | m |

Table 1: Modkit Extract Output

Sturgeon

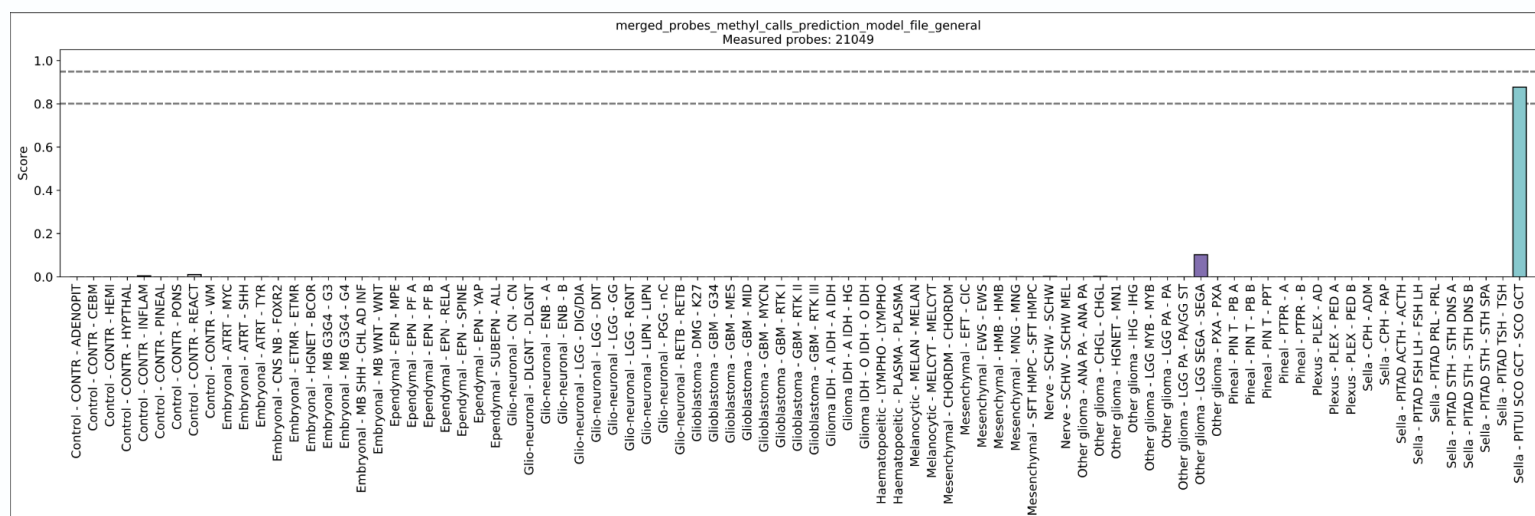
After processing the raw sequencing reads through the different softwares, we finally get to Sturgeon. Sturgeon plays a critical role in our pipeline, as it is responsible for examining the extracted methylation features to provide a tumor classification.

A key step in using Sturgeon is the **inputtobed** command, which takes in methylation data tables produced by Modkit and reformats them for reliable prediction. These bed tables capture essential details such as the chromosome name, start and end positions, methylation call, and probe IDs. The methylation call data is represented in a binary format with 1 corresponding to methylated sites and 0 to unmethylated sites. The probe IDs correspond to those used in the 450k array, which is the platform Sturgeon was trained on—ensuring the input features are aligned with the model’s expectations.

| Chrom | chromStart | chromEnd | Methylation_Call | Probe_ID |
|-------|------------|-----------|------------------|------------|
| 5 | 144762103 | 144762104 | 1 | cg10478863 |
| 5 | 148918435 | 148918436 | 0 | cg17574857 |
| 6 | 9293970 | 9293971 | 1 | cg00874765 |
| 22 | 51281714 | 51281715 | 1 | cg09456760 |

Table 1: Sturgeon Inputtobed Output

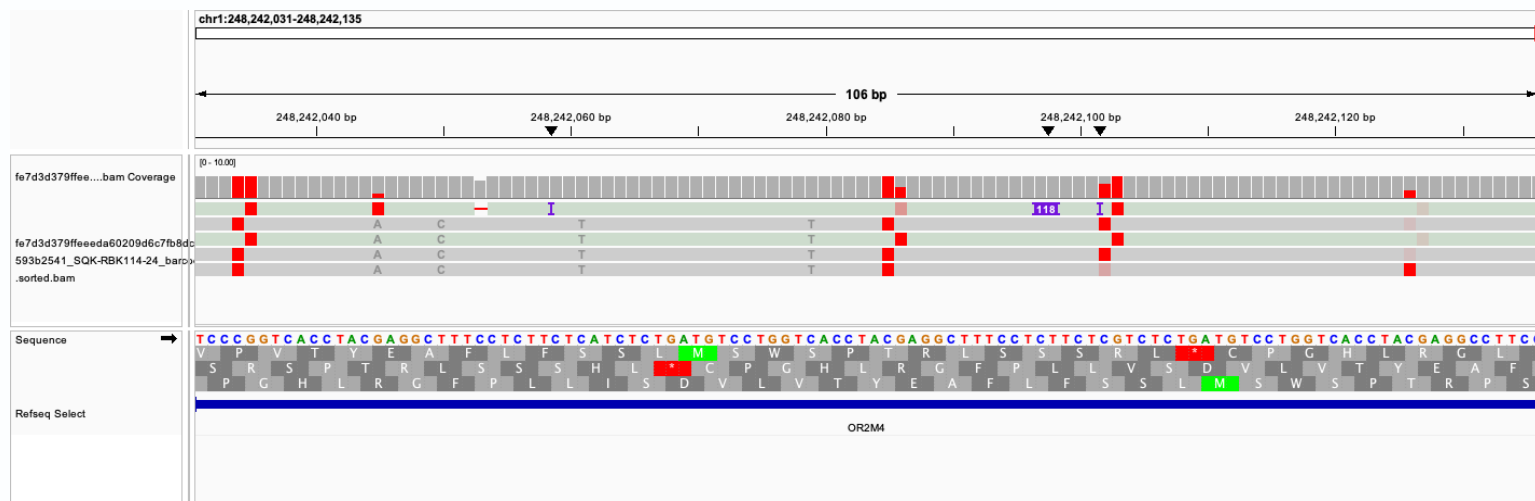
By aggregating methylation patterns at these positions, Sturgeon Predict assigns a tumor subtype and outputs a probability expressing diagnostic confidence. The prediction is presented to the user in a PDF format paired with an excel sheet tabulating the confidence scores. Inside of the PDF, the different subclassifications of central nervous system tumors are labeled on the x-axis mostly clustered by the type of cell they originate from. The y-axis of the bar plot corresponds to the confidence scores, and robust predictions can be derived from the bars that surpass the 0.8 threshold.



Visual Snapshot

Integrative Genomics Viewer (IGV)

The Integrative Genomics Viewer (IGV) is an interactive tool useful for the visual exploration of genomic data. It is useful for exploring the sequenced reads closely. It depicts the sequenced reads in their aligned position while providing information about insertions (purple), deletions (red dashes), and modifications (red blocks) found at various genomic coordinates.





Challenges Faced

I. Understanding Nanopore Sequencing

Coming from a Computer Science background, understanding the biology behind these technologies was initially an uphill battle. Concepts particularly surrounding the library preparation for a nanopore device were very hard to picture. Nevertheless, with a lot of guidance from our mentor, Salvo I was able to finally understand the science behind the data.

II. Basecalling Blunder

Basecalling was the very first and most important step in our pipeline. Understanding the purpose of basecalling and the context of the arguments that make up the command came with time and practice. My use of the 5mC_5hmC model slowed down my basecalling process and disrupted predictions. This was because this modified-based model not only detects methylation in the CpG contexts but in all other cytosine contexts as well. Spotting this mistake, allowed me to gain a deeper understanding into how dorado works and also more broadly about the different types of methylation as well.


III. T2T Genome Download Glitch

Downloading the Telomere-to-Telomere (T2T) Genome was the most straightforward and yet most humbling challenge faced. Getting a file from a download link seemed as easy as it could get but I would soon learn that was not the case. Closer inspection into the size of the file would reveal that my downloaded genome was truncated from a 3GB file to just a couple of bytes. Re-downloading the genome through Chrome instead of Safari eventually solved the problem.

Key Takeaways

Interdisciplinary Collaboration is Powerful

Realizing the value of interdisciplinary thinking was something crucial I learned from this internship. I had always been intimidated about joining the medical/biological fields because in my perspective thriving in those fields required more focused science knowledge. However, learning about the several branches within biological research and



how people with different expertise are able to work collaboratively to make real-world impact was incredibly exciting to learn about.

Refining Bash/Terminal Skills

I have had some prior exposure to the terminal, and using command-line software. However, working in this internship allowed me to practice and refine these skills. In addition to gaining more experience with bash scripting, I was also able to learn new commands and shortcuts whose application is not just limited to the softwares used in the workflow pipeline.

Improved Problem-Solving Skills

Throughout the course of this internship, I have been dealt with several challenges both technical and conceptual. With each difficulty I faced, I learned to not panic or stop trying, but instead think beyond the problem towards the next steps. My problem-solving skills grew stronger with every hiccup along the road and this is something I will take with me in the future.


Reflection

Being my first ever internship, this opportunity will always be extremely valuable. Prior to working at the Hovestadt Lab my knowledge about the application of Computational Biology was extremely limited. As a double major in Computer Science and Mathematics myself, I had never been exposed to softwares like Samtools or Modkit, and did not fully comprehend the extensive and complex research being done in the field of cancer. Prior to joining the lab, I certainly could not have imagined working with real cancer sequencing data to classify something as complex as a brain tumor. Research, to me, felt like something abstract — important, but distant from the world I was used to.

Nevertheless, over the course of the internship, my perception changed completely.

Through hands-on experience looking into nanopore sequencing data and then employing the Sturgeon classifier, I was being introduced to a new side of Data Science I had never before considered. I learned how to work with raw DNA signals, and then use bioinformatic tools to convert these signals into DNA sequences and finally classify brain tumor subtypes. Seeing how these tools worked together to diagnose real patients gave me a new sense of purpose behind the code I was writing and the data I was analyzing.

I learned that the work being done in the lab will revolutionize the way cancer is being diagnosed and was overjoyed to be able to learn from the people spearheading this



research. Understanding that pieces of code and few well-trained models could change the course of someone's life was truly eye-opening.

Reading about the 74 year old patient from the Netherlands, for whom nanopore sequencing and Sturgeon helped diagnose his IDH-Wildtype Glioblastoma helped shift my focus from the data to the person behind it.

This internship did not just teach me a set of new tools or help refine my bash scripting skills — it changed my outlook on research and its impacts. It showed me that despite not having a biology background, you could still make meaningful contributions to medical research. I genuinely feel that internship helped open so many doors of opportunities, by showing me the power of interdisciplinary work. It has most certainly evoked my curiosity and pushed me to keep exploring similar opportunities.

Bibliography

"Brain Tumors in Children." *Symptoms, Diagnosis and Treatment at Nationwide Children's Hospital*,
www.nationwidechildrens.org/conditions/brain-tumors#:~:text=Long%2DTerm%20Outlook%20for%20a,New%20developments%20in%20treatment. Accessed 18 July 2025.

"Getting Started." *Dorado Documentation*, dorado-docs.readthedocs.io/en/latest/. Accessed 18 July 2025.


"Methylation." *Genome.Gov*,
www.genome.gov/genetics-glossary/Methylation#:~:text=Methylation%20is%20a%20chemical%20modification,proteins%20that%20the%20gene%20encodes. Accessed 18 July 2025.

"Nanopore Sequencing Accuracy." *Oxford Nanopore Technologies*,
nanoporetech.com/platform/accuracy. Accessed 18 July 2025.

"Overview#." *IGV Desktop Application*, igv.org/doc/desktop/. Accessed 18 July 2025.

"Part Three: Research as Daily Practice, Evidence-Based Practice, and Research Mindedness." *Faculty of Social Work*, 10 July 2023,
socialwork.ucalgary.ca/research/field-education-research/tfel-training-module-course/part-three.

Sol, Nik, et al. "Glioblastoma, IDH-Wildtype with Primarily Leptomeningeal Localization Diagnosed by Nanopore Sequencing of Cell-Free DNA from Cerebrospinal Fluid." *Acta Neuropathologica*, U.S. National Library of Medicine, 3 Sept. 2024,
[pmc.ncbi.nlm.nih.gov/articles/PMC11371860/](https://pubmed.ncbi.nlm.nih.gov/articles/PMC11371860/).



Vermeulen, C., et al. "Ultra-Fast Deep-Learned CNS Tumour Classification during Surgery." *Nature News*, Nature Publishing Group, 11 Oct. 2023, www.nature.com/articles/s41586-023-06615-2.

Walker, David, et al. "Strategies to Accelerate Diagnosis of Primary Brain Tumors at the Primary-Secondary Care Interface in Children and Adults." *CNS Oncology*, U.S. National Library of Medicine, Sept. 2013, [pmc.ncbi.nlm.nih.gov/articles/PMC6136128/#:~:text=Accelerating%20diagnosis&text=The%20symptomatology%20is%20dictated%20by,confidence%20in%20the%20health%20services](https://pubmed.ncbi.nlm.nih.gov/articles/PMC6136128/#:~:text=Accelerating%20diagnosis&text=The%20symptomatology%20is%20dictated%20by,confidence%20in%20the%20health%20services).

Xin, Yurong, et al. "Role of CPG Context and Content in Evolutionary Signatures of Brain DNA Methylation." *Epigenetics*, U.S. National Library of Medicine, Nov. 2011, [pmc.ncbi.nlm.nih.gov/articles/PMC3775885/](https://pubmed.ncbi.nlm.nih.gov/articles/PMC3775885/).