OpenPBTA: An Open Pediatric Brain Tumor Atlas

This manuscript (permalink) was automatically generated from AlexsLemonade/OpenPBTA-manuscript@6ebef2d on September 12, 2022.

Authors

- **Joshua A. Shapiro** 0000-0002-6224-0347 · jashapiro · jashapiro Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Bala Cynwyd, PA, USA · Funded by Alex's Lemonade Stand Foundation Childhood Cancer Data Lab (CCDL)

- **Krutika S. Gaonkar** 0000-0003-0838-2405 · kgaonkar6 · aggokittu Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia; Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia

- **Candace L. Savonen** 0000-0001-6331-7070 · cansavvy · cansavvy Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Bala Cynwyd, PA, USA; Fred Hutchinson Cancer Center, Seattle, WA, USA · Funded by Alex's Lemonade Stand Foundation Childhood Cancer Data Lab (CCDL)

- **Stephanie J. Spielman** 0000-0002-9090-4788 · sjspielman · stephspiel Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Bala Cynwyd, PA, USA+; Rowan University, Glassboro, NJ, USA · Funded by Alex's Lemonade Stand Foundation Childhood Cancer Data Lab (CCDL) +Current affiliation

- **Chante J. Bethell** 0000-0001-9653-8128 · cbethell · cjbethell Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Bala Cynwyd, PA, USA · Funded by Alex's Lemonade Stand Foundation Childhood Cancer Data Lab (CCDL)

- **Run Jin** 0000-0002-8958-9266 · runjin326 · runjin Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Komal S. Rathi** 0000-0001-5534-6904 · komalsrathi · komalsrathi Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia

- **Yuankun Zhu** 0000-0002-2455-9525 · yuankunzhu · zhuyuankun Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Laura E. Egolf** 0000-0002-7103-4801 · LauraEgolf · LauraEgolf Cell and Molecular Biology Graduate Group, Perelman School of Medicine at the University of Pennsylvania; Division of Oncology, Children's Hospital of Philadelphia

- **Bailey K. Farrow** 0000-0001-6727-6333 · baileyckelly Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Daniel P. Miller** 0000-0002-2032-4358 · dmiller15 Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Yang Yang** · yangyangclover Ben May Department for Cancer Research, University of Chicago, Chicago IL, USA

- **Tejaswi Koganti** 0000-0002-7733-6480 · tkoganti Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Nighat Noureen** 0000-0001-7495-8201 · NNoureen Greehey Children's Cancer Research Institute, UT Health San Antonio

- **Mateusz P. Koptyra** 0000-0002-3857-6633 · mkoptyra · koptyram Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Nhat Duong** 0000-0003-2852-4263 · fingerfen · asiannhat Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia

- **Adam A. Kraya** 0000-0002-8526-5694 · aadamk Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Cassie Kline** 0000-0001-7765-7690 · cassiekmd Division of Oncology, Children's Hospital of Philadelphia

- **Yiran Guo** 0000-0002-6549-8589 · Yiran-Guo · YiranGuo3 Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia

- **Phillip B. Storm** 0000-0002-7964-2449 Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia · Funded by Alex's Lemonade Stand Foundation (Catalyst); Children's Hospital of Philadelphia Division of Neurosurgery

- **Javad Nazarian** 0000-0002-1951-9828 Children's National Research Institute, Washington, D.C.; George Washington University School of Medicine and Health Sciences, Washington, D.C.

- **Stephen C. Mack** 0000-0001-9620-4742 Department of Developmental Neurobiology, St. Jude Children's Research Hospital

- **Pichai Raman** 0000-0001-6948-2157 · pichairaman · PichaiRaman Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia

- **Siyuan Zheng** 0000-0002-1031-9424 · syzheng · zhengsiyuan Greehey Children's Cancer Research Institute, UT Health San Antonio

- **Peter J. Madsen** 0000-0001-9266-3685 · petermadsenmd Division of Neurosurgery, Children's Hospital of Philadelphia; Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia

- **Nicholas Van Kuren** 0000-0002-7414-9516 · nicholasvk Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Shannon Robins** 0000-0003-0594-1953 Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Xiaoyan Huang** 0000-0001-7267-4512 · HuangXiaoyan0106 Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Angela J. Waanders** 0000-0002-0571-2889 · awaanders Division of Hematology, Oncology, Neuro-Oncology, and Stem Cell Transplant, Ann & Robert H Lurie Children's Hospital of Chicago; Department of Pediatrics, Northwestern University Feinberg School of Medicine

- **Jung Kim** 0000-0001-6274-2841 Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute

- **Derek Hanson** 0000-0002-0024-5142 Hackensack Meridian School of Medicine; Hackensack University Medical Center

- **Carl J. Koschmann** 0000-0002-0825-7615 Department of Pediatrics, University of Michigan Health, Ann Arbor, MI; Pediatric Hematology Oncology, Mott Children's Hospital, Ann Arbor, MI

- **Sharon J. Diskin** 0000-0002-7200-8939 · sdiskin · sjdiskin Division of Oncology, Children's Hospital of Philadelphia; Department of Pediatrics, University of Pennsylvania

- **Rishi R. Lulla** 0000-0003-4109-2207 Division of Hematology/Oncology, Hasbro Children's Hospital; Department of Pediatrics, The Warren Alpert School of Brown University, Providence, Rhode Island

- **Miguel A. Brown** 0000-0001-6782-1442 · migbro · migbro Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Jessica Wong** 0000-0003-1508-7631 · wongjessica93 · jessicawongbfx Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Jennifer L. Mason** · jenn0307 · jenn0307 Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Laura Scolaro** Division of Oncology, Children's Hospital of Philadelphia

- **Meen Chul Kim** 0000-0002-0308-783X · liberaliscomputing Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Hongbo M. Xie** 0000-0003-2223-0029 · xiehongbo · xiehb Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia

- **Brian R. Rood** Children's National Research Institute, Washington, D.C.; George Washington University School of Medicine and Health Sciences, Washington, D.C.

- **Kristina A. Cole** 0000-0003-0064-2882 Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, PA; Department of Pediatrics, University of Pennsylvania, Philadelphia, PA; Abramson Family Cancer Research Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

- **Ashley S. Margol** 0000-0002-3038-8005 Division of Hematology and Oncology, Children's Hospital Los Angeles; Department of Pediatrics, Keck School of Medicine of University of Southern California

- **Zalman Vaksman** Division of Oncology, Children's Hospital of Philadelphia

- **Rebecca S. Kaufman** 0000-0001-8535-9730 · rebkau Division of Oncology, Children's Hospital of Philadelphia; Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia

- **Allison P. Heath** 0000-0002-2583-9668 · allisonheath · allig8r Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia · Funded by NIH U2C HL138346-03; NCI/NIH Contract No. 75N91019D00024, Task Order No. 75N91020F00003; Australian Government, Department of Education

- **Brian M. Ennis** 0000-0002-2653-5009 · bmennis Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Mariarita Santi** 0000-0002-6728-3450 Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia; Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine

- **Anna R. Poetsch** 0000-0003-3056-4360 · arpoe · APoetsch Biotechnology Center, Technical University Dresden, Germany; National Center for Tumor Diseases, Dresden, Germany · Funded by The St. Anna Kinderkrebsforschung, Austria; The Mildred Scheel Early Career Center Dresden P2, funded by the German Cancer Aid

- **Angela N. Viaene** 0000-0001-6430-8360 Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia; Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine

- **Douglas R. Stewart** 0000-0001-8193-1488 Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute

- **Steven M. Foltz** 0000-0002-9526-8194 · envest Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Bala Cynwyd, PA, USA · Funded by Alex's Lemonade Stand Foundation GR-000002471; National Institutes of Health K12GM081259

- **Payal Jain** 0000-0002-5914-9083 · jainpayal022 · jainpayal022 Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Bo Zhang** 0000-0002-0743-5379 · zhangb1 Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Shrivats Kannan** 0000-0002-1460-920X · shrivatsk · kshrivats Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Michael Prados** 0000-0002-9630-2075 University of California, San Francisco, San Francisco, CA, USA

- **Jena V. Lilly** 0000-0003-1439-6045 · jvlilly · jvlilly Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia

- **Sabine Mueller** 0000-0002-3452-5150 University of California, San Francisco, San Francisco, CA

- **Adam C. Resnick** 0000-0003-0436-4189 · adamcresnick · adamcresnick Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia · Funded by Alex's Lemonade Stand Foundation (Catalyst); Children's Brain Tumor Network; NIH 3P30 CA016520-44S5, U2C HL138346-03, U24 CA220457-03; NCI/NIH Contract No. 75N91019D00024, Task Order No. 75N91020F00003; Children's Hospital of Philadelphia Division of Neurosurgery

- **Casey S. Greene** 0000-0001-8713-9213 · cgreene · greenescientist Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Bala Cynwyd, PA, USA; Center for Health AI, University of Colorado School of Medicine, Aurora, CO, USA; Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA · Funded by Alex's Lemonade Stand Foundation Childhood Cancer Data Lab (CCDL)

- **Jo Lynne Rokita** — 0000-0003-2171-3627 · jharenza · jolynnerokita Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia; Division of Neurosurgery, Children's Hospital of Philadelphia; Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia · Funded by Alex's Lemonade

Stand Foundation (Young Investigator, Catalyst); NCI/NIH Contract No. 75N91019D00024, Task Order No. 75N91020F00003

- **Jaclyn N. Taroni** — 0000-0003-4734-4508 · jaclyn-taroni · jaclyn_taroni
  Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Bala Cynwyd, PA, USA · Funded by Alex's Lemonade Stand Foundation Childhood Cancer Data Lab (CCDL)
- **Children's Brain Tumor Network**
- **Pacific Pediatric Neuro-Oncology Consortium**

Contact information

Correspondence: Jo Lynne Rokita <rokita@chop.edu>, Jaclyn N. Taroni <jaclyn.taroni@ccdatalab.org>.

In Brief

The OpenPBTA is a global, collaborative open-science initiative which brought together researchers and clinicians to genomically characterize 1,074 pediatric brain tumors and 22 patient-derived cell lines. Shapiro, et. al create over 40 open-source, scalable modules to perform cancer genomics analyses and provide a richly-annotated somatic dataset across 58 brain tumor histologies. The OpenPBTA framework can be used as a model for large-scale data integration to inform basic research, therapeutic target identification, and clinical translation.

Highlights

OpenPBTA collaborative analyses establish resource for 1,074 pediatric brain tumors

NGS-based WHO-aligned integrated diagnoses generated for 641 of 1,074 tumors

RNA-Seq analysis infers medulloblastoma subtypes, *TP53* status, and telomerase activity

OpenPBTA will accelerate therapeutic translation of genomic insights

Summary

Pediatric brain and spinal cancer are the leading disease-related cause of death in children, thus we urgently need curative therapeutic strategies for these tumors. To accelerate such discoveries, the Children's Brain Tumor Network and Pacific Pediatric Neuro-Oncology Consortium created a systematic process for tumor biobanking, model generation, and sequencing with immediate access to harmonized data. We leverage these data to create OpenPBTA, an open collaborative project which establishes over 40 scalable analysis modules to genomically characterize 1,074 pediatric brain tumors. Transcriptomic classification reveals

that *TP53* loss is a significant marker for poor overall survival in ependymomas and H3 K28-altered diffuse midline gliomas and further identifies universal *TP53* dysregulation in mismatch repair-deficient hypermutant high-grade gliomas. OpenPBTA is a foundational analysis platform actively being applied to other pediatric cancers and inform molecular tumor board decision-making, making it an invaluable resource to the pediatric oncology community.

Keywords

pediatric cancer, brain tumors, somatic variation, open science, reproducibility, classification, tumor atlas

Introduction

Pediatric brain and spinal cord tumors are collectively the second most common malignancy in children after leukemia, and they represent the leading disease-related cause of death in children[1]. Five-year survival rates vary widely across different histologic and molecular classifications of brain tumors. For example, most high-grade gliomas carry a universally fatal prognosis, while children with pilocytic astrocytoma have an estimated 10-year survival rate of 92%[2]. Moreover, estimates from 2009 suggest that children and adolescents aged 0-19 with brain tumors in the United States have lost an average of 47,631 years of potential life[3].

The low survival rates for some pediatric tumors are clearly multifactorial, explained partly by our lack of comprehensive understanding of the ever-evolving array of brain tumor molecular subtypes, difficulty drugging these tumors, and the shortage of drugs specifically labeled for pediatric malignancies. Historically, some of the most fatal, inoperable brain tumors, such as diffuse intrinsic pontine gliomas (DIPGs), were not routinely biopsied due to perceived risks of biopsy and the paucity of therapeutic options that would require tissue. Limited access to tissue to develop patient-derived cell lines and mouse models has been a barrier to research. Furthermore, the incidence of any single brain tumor molecular subtype is relatively low due to the rarity of pediatric tumors in general.

To address these long-standing barriers, multiple national and international consortia have come together to uniformly collect clinically-annotated surgical biosamples and associated germline materials as part of both observational and interventional clinical trials. Such accessible, centralized resources enable collaborative sharing of specimens and data across rare cancer subtypes to accelerate breakthroughs and clinical translation. The creation of the Pediatric Brain Tumor Atlas (PBTA) in 2018, led by the Children's Brain Tumor Network (CBTN, cbtn.org) and the Pacific Pediatric Neuro-Oncology Consortium (PNOC, PNOC.us) is one such effort that builds on nearly 10 years of multi-institutional enrollment, sample collection, and clinical followup across more than 30 institutions. Just as cooperation is required to share specimens and data, rigorous cancer genomic analysis requires collaboration among researchers with distinct expertise, such as computational scientists, bench scientists, clinicians, and pathologists.

Although there has been significant progress in recent years to elucidate the landscape of somatic variation responsible for pediatric brain tumor formation and progression, translation of

therapeutic agents to phase II or III clinical trials and subsequent FDA approvals have not kept pace. Within the last 20 years, the FDA has approved only five drugs for the treatment of pediatric brain tumors: mTOR inhibitor everolimus, for subependymal giant cell astrocytoma; anti-PD-1 immunotherapy pembrolizumab, for microsatellite instability–high or mismatch repair–deficient tumors; NTRK inhibitors larotrectinib and entrectinib, for tumors with an NTRK 1/2/3 gene fusions; and MEK1/2 inhibitor selumetinib, for neurofibromatosis type 1 (NF1) and symptomatic, inoperable plexiform neurofibromas[4].

This is, in part, due to pharmaceutical company priorities and concerns regarding toxicity, making it challenging for researchers to obtain to new therapeutic agents for pediatric clinical trials. Critically, as of August 18, 2020, an amendment to the Pediatric Research Equity Act called the "Research to Accelerate Cures and Equity (RACE) for Children Act" mandates that all new adult oncology drugs also be tested in children when the molecular targets are relevant to a particular childhood cancer. The regulatory change introduced by the RACE Act, coupled with the identification of putative molecular targets in pediatric cancers through genomic characterization, is poised to accelerate identification of novel and effective therapeutic for pediatric diseases that have otherwise been overlooked.

To leverage diverse scientific and analytical expertise to analyze the PBTA data, we created an open science model and incorporated features such as analytical code review[5,6] and continuous integration to test data and code[6,7] to improve reproducibility throughout the life cycle of our project, termed OpenPBTA.

We anticipated that a model of open collaboration would enhance the value of our effort to the pediatric brain tumor research community and provide a framework for continuous, accelerated translation of pediatric brain tumor datasets. Openly sharing data and code in real time allows others to build upon the work more rapidly, and publications that include data and code sharing are poised for greater impact[8,9]. Here, we present a comprehensive, collaborative, open genomic analysis of 1,074 tumors and 22 cell lines, comprised of 58 distinct brain tumor histologies from 943 patients. The data and containerized infrastructure of OpenPBTA have been instrumental for discovery and translational research studies[10–12], are actively integrated into PNOC molecular tumor board decision-making, and are a foundational layer for the NCI's Childhood Cancer Data Initiative's (CCDI) pediatric Molecular Targets Platform (https://moleculartargets.ccdi.cancer.gov/) recently built in support of the RACE Act[13]. We anticipate OpenPBTA will be an invaluable resource to the pediatric oncology community.

Results

Crowd-sourced Somatic Analyses to Create an Open Pediatric Brain Tumor Atlas

We previously performed whole genome sequencing (WGS), whole exome sequencing (WXS), and RNA sequencing (RNA-Seq) on matched tumor and normal tissues as well as selected cell lines[14] from 943 patients from the Pediatric Brain Tumor Atlas (PBTA), consisting of samples from the Children's Brain Tumor Network (CBTN) and the PNOC003 DIPG clinical trial[12,15] of the Pacific Pediatric Neuro-Oncology Consortium (PNOC) (**Figure 1A**). We then harnessed the

benchmarking efforts of the Gabriella Miller Kids First Data Resource Center to develop robust and reproducible data analysis workflows within the CAVATICA platform to perform primary somatic analyses including calling of single nucleotide variants (SNVs), copy number variants (CNVs), structural variants (SVs), and gene fusions, often implementing multiple complementary methods (**Figure S1**) and **STAR Methods**).

To facilitate analysis and visualization of this large, diverse cohort, we further categorized tumor broad histologies (i.e., broad 2016 WHO classifications) into smaller groupings we denote "cancer groups." A summarized view of the number of biospecimens per phase of therapy across different broad histologies and cancer groups is shown in (**Figure 1B**). We maintained a data release folder on Amazon S3, downloadable directly from S3 or through the open-access CAVATICA project, with merged files for each analysis (See **Data and code availability** section). As new analytical results (e.g., tumor mutation burden calculations) that we expected to be used across multiple analyses were produced, or issues with the data were identified, new data releases were made available in a versioned manner.

A key innovative feature of this project has been its open contribution model used for both analyses (i.e., analytical code) and scientific manuscript writing. We created a public Github analysis repository (https://github.com/AlexsLemonade/OpenPBTA-analysis) to hold all code associated with analyses downstream of the Kids First Data Resource Center workflows and a GitHub manuscript repository (https://github.com/AlexsLemonade/OpenPBTA-manuscript) with Manubot[16] integration to enable real-time manuscript creation using Markdown within GitHub. Importantly, all analyses and manuscript writing were conducted openly throughout the research project, allowing any researcher in the world the opportunity to contribute.

The process for analysis and manuscript contributions is outlined in **Figure 1C**. First, a potential contributor would propose an analysis by filing an issue in the GitHub analysis repository. Next, organizers for the project, or other contributors with expertise, had the opportunity to provide feedback about the proposed analysis (**Figure 1C**). The contributor then made a copy (fork) of the analysis repository and added their proposed analysis code and results to their fork. The contributor would formally request to include their analytical code and results to the main OpenPBTA analysis repository by filing a pull request on GitHub. All pull requests to the analysis repository underwent peer review by organizers and/or other contributors to ensure scientific accuracy, maintainability, and readability of code and documentation (**Figure 1C-D**).

The collaborative nature of the project required additional steps beyond peer review of analytical code to ensure consistent results for all collaborators and over time (**Figure 1D**). We leveraged Docker®[17] and the Rocker project[18] to maintain a consistent software development environment, creating a monolithic image that contained all dependencies necessary for analyses. To ensure that new code would execute in the development environment, we used the continuous integration (CI) service CircleCI® to run analytical code on a small subset of data for testing before formal code review, allowing us to detect code bugs or sensitivity to changes in the underlying data.

We followed a similar process in our Manubot-powered[16] manuscript repository for additions to the manuscript (**Figure 1C**). Contributors forked the manuscript repository, added proposed content to their branch, and filed pull requests to the main manuscript repository with their

changes. Similarly, pull requests underwent a peer review process for clarity and correctness, agreement with interpretation, and spell checking via Manubot.

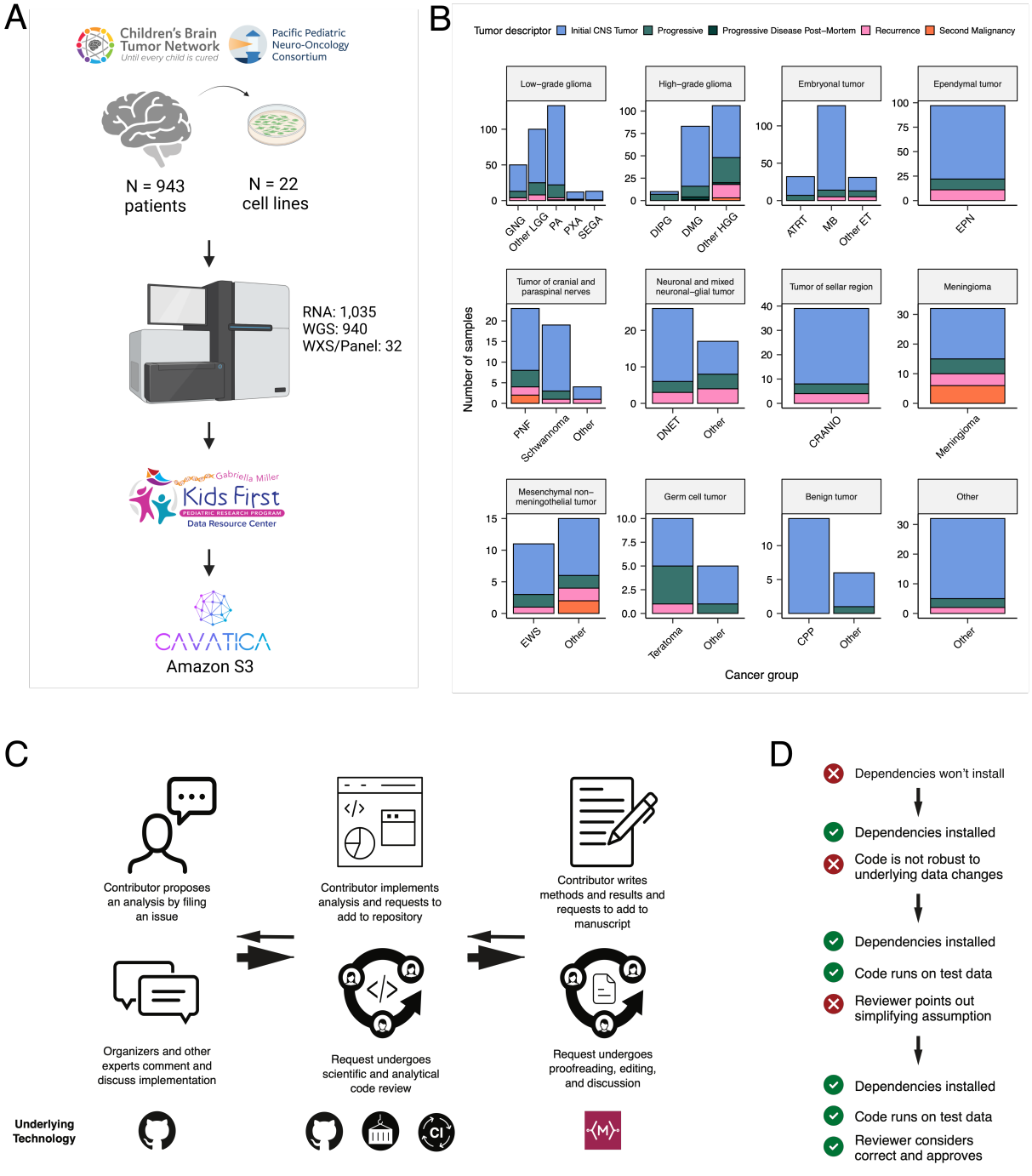


Figure 1: ***Overview of the OpenPBTA Project.*** *A, The Children's Brain Tumor Network and the Pacific Pediatric Neuro-Oncology Consortium collected tumor samples from 943 patients. To date, 22 cell lines were created from tumor tissue, and over 2000 specimens were sequenced (N = 1035 RNA-Seq, N = 940 WGS, and N = 32 WXS or targeted panel). Data was harmonized by the Kids First Data Resource Center using an Amazon S3 framework within CAVATICA. B,*

Stacked bar plot summary of the number of biospecimens per phase of therapy. Each panel denotes a broad histology and each bar denotes a cancer group. (Abbreviations: GNG = ganglioglioma, Other LGG = other low-grade glioma, PA = pilocytic astrocytoma, PXA = pleomorphic xanthoastrocytoma, SEGA = subependymal giant cell astrocytoma, DIPG = diffuse intrinsic pontine glioma, DMG = diffuse midline glioma, Other HGG = other high-grade glioma, ATRT = atypical teratoid rhabdoid tumor, MB = medulloblastoma, Other ET = other embryonal tumor, EPN = ependymoma, PNF = plexiform neurofibroma, DNET = dysembryoplastic neuroepithelial tumor, CRANIO = craniopharyngioma, EWS = Ewing sarcoma, CPP = choroid plexus papilloma). Only samples with available descriptors were included. C, Overview of the open analysis and manuscript contribution model. In the analysis GitHub repository, a contributor would propose an analysis that other participants can comment on. Contributors would then implement the analysis and file a request to add their changes to the analysis repository ("pull request"). Pull requests underwent review for scientific rigor and correctness of implementation. Pull requests were additionally checked to ensure that all software dependencies were included and the code was not sensitive to underlying data changes using container and continuous integration technologies. Finally, a contributor would file a pull request documenting their methods and results to the Manubot-powered manuscript repository. Pull requests in the manuscript repository were also subject to review. D, A potential path for an analytical pull request. Arrows indicate revisions to a pull request. Prior to review, a pull request was tested for dependency installation and whether or not the code would execute. Pull requests also required approval by organizers and/or other contributors, who checked for scientific correctness. Panel A created with BioRender.com.

Molecular Subtyping of OpenPBTA CNS Tumors

Over the past two decades, experts in neuro-oncology have worked with the World Health Organization (WHO) to iteratively redefine the classifications of central nervous system (CNS) tumors[19,20]. More recently, in 2016 and 2021[21,22], molecular subtypes have been integrated into these classifications. In 2011, the Children's Brain Tumor Tissue Consortium, now known as the Children's Brain Tumor Network (CBTN), opened its protocol for brain tumor and matched normal sample collection. Since the CBTN opened its collection protocol in 2011, before molecular data were integrated into classifications, the majority of the samples within the OpenPBTA lacked molecular subtype annotations at the time of tissue collection. Moreover, the OpenPBTA data does not yet feature methylation arrays which are increasingly used to inform molecular subtyping. Therefore, we jointly considered key genomic features of tumor entities described by the WHO in 2016, low-grade glioma (LGG) subtypes described by Ryall and colleagues[23], as well as clinician and pathologist review, to generate research-grade integrated diagnoses for 60% (641/1074) of tumor samples with high confidence (**Table S1**).

Importantly, this collaborative molecular subtyping process allowed us to identify potential data entry errors (e.g., an ETMR incorrectly entered as a medulloblastoma) and histologically mis-identified specimens (e.g., Ewing sarcoma sample labeled as a craniopharyngioma), update diagnoses using current WHO terms (e.g., tumors formerly ascribed primitive neuro-ectodermal tumor [PNET] diagnoses), and discover rarer tumor entities within the OpenPBTA (e.g., H3-mutant ependymoma, meningioma with *YAP1::FAM118B* fusion). **Table 1** lists the subtypes we defined within OpenPBTA, comprising low-grade gliomas (N = 290), high-grade gliomas (N =

141), embryonal tumors (N = 126), ependymomas (N = 30), tumors of sellar region (N = 27), mesenchymal non-meningothelial tumors (N = 11), glialneuronal tumors (N = 10), and chordomas (N = 6). For detailed methods, see **STAR Methods** and **Figure S1**.

Table 1: ***Molecular subtypes generated through the OpenPBTA project.*** *Listed are broad tumor histologies, molecular subtypes generated, and number of specimens subtyped within the OpenPBTA project.*

Broad histology group	OpenPBTA molecular subtype	n
Chordoma	CHDM, conventional	2
Chordoma	CHDM, poorly differentiated	4
Embryonal tumor	CNS Embryonal, NOS	13
Embryonal tumor	CNS HGNET-MN1	1
Embryonal tumor	CNS NB-FOXR2	3
Embryonal tumor	ETMR, C19MC-altered	5
Embryonal tumor	ETMR, NOS	1
Embryonal tumor	MB, Group3	14
Embryonal tumor	MB, Group4	49
Embryonal tumor	MB, SHH	30
Embryonal tumor	MB, WNT	10
Ependymal tumor	EPN, H3 K28	1
Ependymal tumor	EPN, ST RELA	28
Ependymal tumor	EPN, ST YAP1	1
High-grade glioma	DMG, H3 K28	24
High-grade glioma	DMG, H3 K28, TP53 activated	13
High-grade glioma	DMG, H3 K28, TP53 loss	40
High-grade glioma	HGG, H3 G35	3
High-grade glioma	HGG, H3 G35, TP53 loss	1
High-grade glioma	HGG, H3 wildtype	31
High-grade glioma	HGG, H3 wildtype, TP53 activated	5
High-grade glioma	HGG, H3 wildtype, TP53 loss	21
High-grade glioma	HGG, IDH, TP53 activated	2
High-grade glioma	HGG, IDH, TP53 loss	1
Low-grade glioma	GNG, BRAF V600E	13
Low-grade glioma	GNG, BRAF V600E, CDKN2A/B	1
Low-grade glioma	GNG, FGFR	1
Low-grade glioma	GNG, H3	1
Low-grade glioma	GNG, IDH	2
Low-grade glioma	GNG, KIAA1549-BRAF	5
Low-grade glioma	GNG, MYB/MYBL1	1

Broad histology group	OpenPBTA molecular subtype	n
Low-grade glioma	GNG, NF1-germline	1
Low-grade glioma	GNG, NF1-somatic, BRAF V600E	1
Low-grade glioma	GNG, other MAPK	4
Low-grade glioma	GNG, other MAPK, IDH	1
Low-grade glioma	GNG, RTK	3
Low-grade glioma	GNG, wildtype	14
Low-grade glioma	LGG, BRAF V600E	27
Low-grade glioma	LGG, BRAF V600E, CDKN2A/B	5
Low-grade glioma	LGG, FGFR	8
Low-grade glioma	LGG, IDH	3
Low-grade glioma	LGG, KIAA1549-BRAF	113
Low-grade glioma	LGG, KIAA1549-BRAF, NF1-germline	1
Low-grade glioma	LGG, KIAA1549-BRAF, other MAPK	1
Low-grade glioma	LGG, MYB/MYBL1	2
Low-grade glioma	LGG, NF1-germline	6
Low-grade glioma	LGG, NF1-germline, CDKN2A/B	1
Low-grade glioma	LGG, NF1-germline, FGFR	2
Low-grade glioma	LGG, NF1-somatic	2
Low-grade glioma	LGG, NF1-somatic, FGFR	1
Low-grade glioma	LGG, NF1-somatic, NF1-germline, CDKN2A/B	1
Low-grade glioma	LGG, other MAPK	12
Low-grade glioma	LGG, RTK	10
Low-grade glioma	LGG, RTK, CDKN2A/B	1
Low-grade glioma	LGG, wildtype	34
Low-grade glioma	SEGA, RTK	1
Low-grade glioma	SEGA, wildtype	11
Mesenchymal non-meningothelial tumor	EWS	11
Neuronal and mixed neuronal-glial tumor	CNC	2
Neuronal and mixed neuronal-glial tumor	EVN	1
Neuronal and mixed neuronal-glial tumor	GNT, BRAF V600E	1
Neuronal and mixed neuronal-glial tumor	GNT, KIAA1549-BRAF	2
Neuronal and mixed neuronal-glial tumor	GNT, other MAPK	1
Neuronal and mixed neuronal-glial tumor	GNT, other MAPK, FGFR	1
Neuronal and mixed neuronal-glial tumor	GNT, RTK	2
Tumor of sellar region	CRANIO, ADAM	27
	Total	641

Somatic Mutational Landscape of Pediatric Brain Tumors

We performed a comprehensive genomic analysis of somatic SNVs, CNVs, SVs, and fusions across 1,074 tumors (N = 1,019 RNA-Seq, N = 918 WGS, N = 32 WXS/Panel) and 22 cell lines (N = 16 RNA-Seq, N = 22 WGS), from 943 patients, 833 with paired normal specimens (N = 801 WGS, N = 32 WXS/Panel). Following SNV consensus calling (**Figure S1** and **Figure S2A-G**), we observed as expected lower tumor mutation burden (TMB) **Figure S2H** in pediatric tumors compared to adult brain tumors from The Cancer Genome Atlas (TCGA), **Figure S2I**, with hypermutant (> 10 Mut/Mb) and ultra-hypermutant (> 100 Mut/Mb) tumors[24] only found within HGGs. **Figure 2** and **Figure S3A** depict oncoprints of histology-specific driver genes across PBTA histologies.

Low-grade gliomas

As expected, the majority (62%, 140/227) of LGGs harbored a somatic alteration in *BRAF*, with canonical *BRAF::KIAA1549* fusions as the major oncogenic driver[25] (**Figure 2A**). We observed additional mutations in *FGFR1* (2%), *PIK3CA* (2%), *KRAS* (2%), *TP53* (1%), and *ATRX* (1%) and fusions in *NTRK2* (2%), *RAF1* (2%), *MYB* (1%), *QKI* (1%), *ROS1* (1%), and *FGFR2* (1%), concordant with previous studies reporting the near universal upregulation of the RAS/MAPK pathway in these tumors resulting from activating mutations and/or oncogenic fusions[23,25]. Indeed, we observed significant upregulation (ANOVA Bonferroni-corrected $p < 0.01$) of the KRAS signaling pathway in LGGs (**Figure 5B**).

Embryonal tumors

The majority (N = 95) of embryonal tumor samples were medulloblastomas that spanned the spectrum of molecular subtypes (WNT, SHH, Group3, and Group 4; see **Molecular Subtyping of CNS Tumors**), as identified by subtype-specific canonical mutations (**Figure 2B**). We detected canonical *SMARCB1/SMARCA4* deletions or inactivating mutations in atypical teratoid rhabdoid tumors (ATRTs) and C19MC amplification in the embryonal tumors with multilayer rosettes (ETMRs, displayed as other embryonal tumors)[26–29].

High-grade gliomas

Across HGGs, we found that *TP53* (57%, 35/61) and *H3F3A* (52%, 32/61) were both most mutated and co-occurring genes (**Figure 2A and C**), followed by frequent mutations in *ATRX* (30%, 18/61). We observed recurrent amplifications and fusions in *EGFR*, *MET*, *PDGFRA*, and *KIT*, highlighting that these tumors utilize multiple oncogenic mechanisms to activate tyrosine kinases, as has been previously reported[15,30,31]. Gene set enrichment analysis showed upregulation (ANOVA Bonferroni-corrected $p < 0.01$) of DNA repair, G2M checkpoint, and MYC pathways as well as downregulation of the TP53 pathway (**Figure 5B**). The two tumors with ultra-high TMB (> 100 Mutations/Mb) were from patients with known mismatch repair deficiency syndrome[14].

Other CNS tumors

We observed that 25% (15/60) of ependymoma tumors were *C11orf95::RELA* (now, *ZFTA::RELA*) fusion-positive ependymomas and that 68% (21/31) of craniopharyngiomas were driven by mutations in *CTNNB1* (**Figure 2D**). Multiple histologies contained somatic mutations or fusions in *NF2*: 41% (7/17) of meningiomas, 5% (3/60) of ependymomas, and 27% (3/11) schwannomas. Rare fusions in *ERBB4*, *YAP1*, *KRAS*, and *MAML2* were observed in 10% (6/60) of ependymoma tumors. DNETs harbored alterations in MAPK/PI3K pathway genes as previously reported[32], including *FGFR1* (21%, 4/19), *PDGFRA* (10%, 2/19), and *BRAF* (5%, 1/19). Frequent mutations in additional rare brain tumor histologies are depicted in **Figure S3A**.

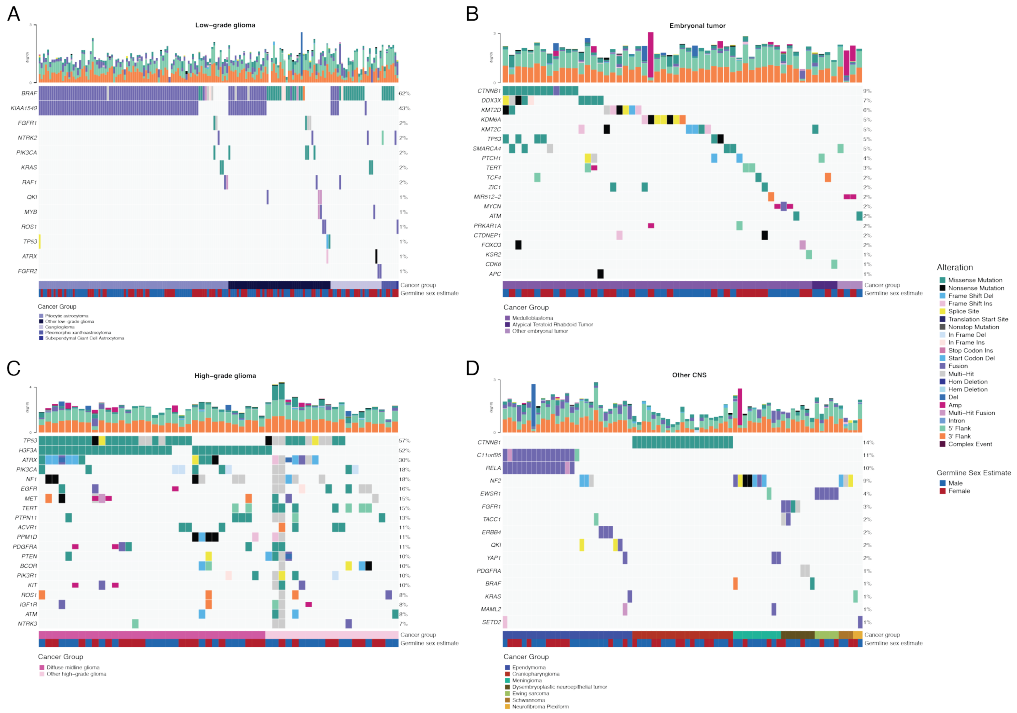


Figure 2: ***Mutational landscape of PBTA tumors.*** *Shown are frequencies of canonical somatic gene mutations, CNVs, fusions, and TMB (top bar plot) for the top 20 genes mutated across primary tumors within the OpenPBTA dataset. A, Low-grade astrocytic tumors (N = 227): pilocytic astrocytoma (N = 104), other low-grade glioma (N = 69), ganglioglioma (N = 35), pleomorphic xanthoastrocytoma (N = 9), subependymal giant cell astrocytoma (N = 10); B, Embryonal tumors (N = 128): medulloblastomas (N = 95), atypical teratoid rhabdoid tumors (N = 24), other embryonal tumors (N = 9); C, Diffuse astrocytic and oligodendroglial tumors (N = 61): diffuse midline gliomas (N = 34) and other high-grade gliomas (N = 27); D, Other CNS tumors (N = 194): ependymomas (N = 60), craniopharyngiomas (N = 31), meningiomas (N = 17), dysembryoplastic neuroepithelial tumors (N = 19), Ewing sarcomas (N = 7), schwannomas (N = 11), and neurofibroma plexiforms (N = 7). Additional, rare CNS tumors are displayed in* ***Figure***

***S3A**. Tumor histology (`Cancer Group`) and patient sex (`Germline sex estimate`) are displayed as annotations at the bottom of each plot. Only samples with mutations in the listed genes are shown. Multiple CNVs are denoted as a complex event.*

Mutational co-occurrence, CNV, and signatures highlight key oncogenic drivers

We analyzed mutational co-occurrence among OpenPBTA tumors, using a single sequencing sample from each individual with available WGS (N = 666). The top 50 mutated genes (see **STAR Methods** for details) in primary tumors are shown in **Figure 3** by tumor type (**A**, bar plots), with co-occurrence scores illustrated in the heatmap (**B**). *TP53* was the most frequently mutated gene across OpenPBTA tumors (8.4%, 56/666), significantly co-occurring with *H3F3A* (OR = 32, 95% CI: 15.3 - 66.7, q = 8.46e-17), *ATRX* (OR = 20, 95% CI: 8.4 - 47.7, q = 4.43e-8), *NF1* (OR = 8.62, 95% CI: 3.7 - 20.2, q = 5.45e-5), and *EGFR* (OR = 18.2, 95% CI: 5 - 66.5, q = 1.6e-4). Other canonical cancer driver genes that were frequently mutated included *BRAF*, *H3F3A*, *CTNNB1*, *NF1*, *ATRX*, *FGFR1*, and *PIK3CA*.

At the broad histology level, mutations in *CTNNB1* significantly co-occurred with mutations in *TP53* (OR = 42.9, 95% CI: 7 - 261.4, q = 1.63e-3) and *DDX3X* (OR = 21.1, 95% CI: 4.6 - 96.3, q = 4.46e-3) in embryonal tumors. Mutations in *FGFR1* and *PIK3CA* significantly co-occurred in LGGs (OR = 76.1, 95% CI: 9.85 - 588.1, q = 3.26e-3), consistent with previous findings[33,34]. Of HGG tumors with mutations in *TP53* or *PPM1D*, 52/54 (96.3%) had mutations in only one of these genes (OR = 0.188, 95% CI: 0.04 - 0.94, p = 0.0413, q = 0.0587). This trend recapitulates previous observations that *TP53* and *PPM1D* mutations tend to be mutually exclusive in HGGs[35].

We summarized broad CNV and SV and observed that HGGs and DMGs, followed by medulloblastomas, had the most unstable genomes (**Figure S3B**). By contrast, craniopharyngiomas and schwannomas generally lacked somatic CNV. Together, these CNV patterns largely aligned with our estimates of tumor mutational burden (**Figure S2H**). The breakpoint density estimated from SV and CNV data was significantly correlated across tumors (p = 1.08e-37) (**Figure 3C**) and as expected, the number of chromothripsis regions called increased as breakpoint density increased (**Figure S3B-C**). We identified chromothripsis events in 28% (N = 11/39) of diffuse midline gliomas and in 40% (N = 19/48) of other HGGs (non-midline HGGs) (**Figure 3D**). We also found evidence of chromothripsis in over 15% of sarcomas, PXAs, metastatic secondary tumors, chordomas, glial-neuronal tumors, germinomas, meningiomas, ependymomas, medulloblastomas, ATRTs, and other embryonal tumors, highlighting the genomic instability and complexity of these pediatric brain tumors.

We next assessed the contributions of eight previously identified adult CNS-specific mutational signatures from the RefSig database[36] across samples (**Figure 3E** and **Figure S4A**). Stage 0 and/or 1 tumors characterized by low TMBs (**Figure S2H**) such as pilocytic astrocytomas, gangliogliomas, other LGGs, and craniopharyngiomas, were dominated by Signature 1 (**Figure S4A**), which results from the normal process of spontaneous deamination of 5-methylcytosine. Signature N6 is a CNS-specific signature which we observed nearly universally across samples. Drivers of Signature 18, *TP53*, *APC*, *NOTCH1* (found at

https://signal.mutationalsignatures.com/explore/referenceCancerSignature/31/drivers), are also canonical drivers of medulloblastoma, and indeed, we observed Signature 18 as the signature with the highest weight in medulloblastoma tumors. Signatures 3, 8, 18, and MMR2 were prevalent in HGGs, including DMGs. Finally, we found that the Signature 1 weight was higher at diagnosis (pre-treatment) and was almost always lower in tumors at later phases of therapy (progression, recurrence, post-mortem, secondary malignancy; **Figure S4B**). This trend may have resulted from therapy-induced mutations that produced additional signatures (e.g., temozolomide treatment has been suggested to drive Signature 11[37]), subclonal expansion, and/or acquisition of additional driver mutations during tumor progression, leading to higher overall TMBs and additional signatures.

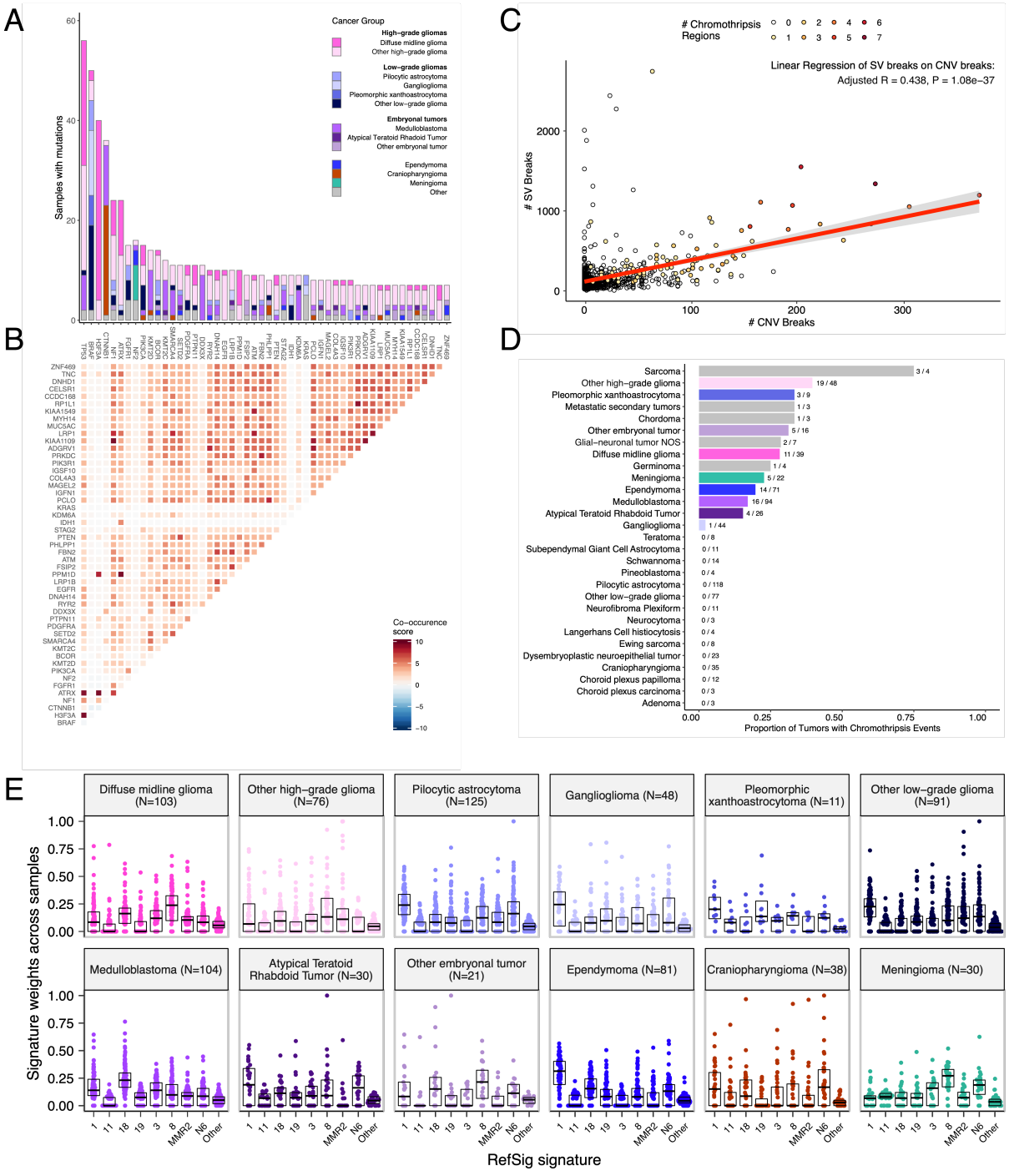


Figure 3: ***Mutational co-occurrence and signatures highlight key oncogenic drivers.*** *A, Bar plot of occurrence and co-occurrence of nonsynonymous mutations for the 50 most commonly mutated genes across all tumor types, which are denoted as "Other" when there are fewer than 10 samples per grouping; B, Co-occurrence and mutual exclusivity of nonsynonymous mutations between genes; The co-occurrence score is defined as* $I(-log_{10}(P))$ *where* P *is defined by Fisher's exact test and* I *is 1 when mutations co-occur more often than expected and -1 when exclusivity is more common; C, The number of SV breaks significantly correlate with*

CNV breaks (Adjusted R = 0.436, p = 1.08e-37). D, Chromothripsis frequency across pediatric brain tumors for all cancer groups with N >= 3 samples. E, Sina plots of RefSig signature weights for signatures 1, 11, 18, 19, 3, 8, N6, MMR2, and Other across cancer groups. Box plot lines represent the first quartile, median, and third quartile.

Transcriptomic Landscape of Pediatric Brain Tumors

Prediction of *TP53* oncogenicity and telomerase activity

To understand the *TP53* phenotype in each tumor, we ran a classifier previously trained on TCGA[38] to calculate a *TP53* score and infer *TP53* inactivation status. We compared results of this classifier to "true positive" alterations derived using high-confidence SNVs, CNVs, SVs, and fusions in *TP53*. Specifically, we annotated *TP53* alterations as "activated" if samples harbored one of p.R273C or p.R248W gain-of-function mutations[39], or "lost" if the given patient either had a Li Fraumeni Syndrome (LFS) predisposition diagnosis, the tumor harbored a known hotspot mutation, or the tumor contained two hits (e.g. both SNV and CNV), which would suggest both alleles had been affected. If the *TP53* mutation did not reside within the DNA-binding domain or we did not detect any alteration in *TP53*, we annotate the tumor as "other," reflecting its unknown *TP53* alteration status. The classifier achieved a high accuracy (AUROC = 0.85) for rRNA-depleted, stranded samples compared to randomly shuffled *TP53* scores (**Figure 4A**). By contrast, while this classifier has previously shown strong performance on poly-A data from both adult[38] tumors and pediatric patient-derived xenografts[40], it did not perform as well on the poly-A samples in this cohort (AUROC = 0.62; **Figure S5A**).

While we expected that samples annotated as "lost" would have higher *TP53* scores than would samples annotated as "other," we observed that samples annotated as "activated" had similar *TP53* scores to those annotated as "lost" (**Figure 4B**, Wilcoxon p = 0.23). This result suggests that the classifier actually detects an oncogenic, or altered, *TP53* phenotype (scores > 0.5) rather than solely *TP53* inactivation, as interpreted previously[38]. Moreover, tumors with "activating" *TP53* mutations showed higher *TP53* expression compared to those with *TP53* "loss" mutations (Wilcoxon p = 3.5e-3, **Figure 4C**). Tumor types with the highest median *TP53* scores were those known to harbor somatic *TP53* alterations and included DMGs, medulloblastomas, HGGs, DNETs, ependymomas, and craniopharyngiomas (**Figure 4D**), while gangliogliomas, LGGs, meningiomas, and schwannomas had the lowest median scores.

To further validate the classifier's accuracy, we assessed *TP53* scores for patients with LFS, hypothesizing that all of these tumors would have high scores. Indeed, we observed higher scores in LFS tumors (N = 8) for which we detected high-confidence *TP53* somatic alterations (**Tables S1 and S3**). Although we did not detect canonical somatic *TP53* mutations in two patients whose tumors had low *TP53* scores (BS_DEHJF4C7 with a score of 0.09 and BS_ZD5HN296 with a score of 0.28), we confirmed from pathology reports these patients were both diagnosed with LFS and had pathogenic germline variants in *TP53*. In addition, the tumor purity of these two LFS samples was low (16% and 37%, respectively), suggesting the classifier may require a certain level of tumor purity to achieve good performance, as we expect *TP53* to be intact in normal cells.

We next used gene expression data to predict telomerase activity using EXpression-based Telomerase ENzymatic activity Detection (EXTEND)[41] as a surrogate measure of malignant potential[41,42], such that higher EXTEND scores suggest increased malignant potential. As expected, EXTEND scores significantly correlated with *TERC* (R = 0.619, p < 0.01) and *TERT* (R = 0.491, p < 0.01) expression (**Figure S5B-C**). We found aggressive tumors such as HGGs (DMGs and other high-grade gliomas) and MB had high EXTEND scores (**Figure 4D**), while low-grade lesions such as schwannomas, GNGs, DNETs, and other low-grade gliomas had among the lowest scores (**Table S3**). These findings support previous reports of a more aggressive phenotype in tumors with higher telomerase activity[43–46].

Hypermutant tumors share mutational signatures and have dysregulated *TP53*

We further investigated the mutational signature profiles of the hypermutant (TMB > 10 Mut/Mb; N = 3) and ultra-hypermutant (TMB > 100 Mut/Mb; N = 4) tumors and/or derived cell lines from six patients in the OpenPBTA cohort (**Figure 4E**). Five of six tumors were diagnosed as HGGs and one was a brain metastasis of a MYCN non-amplified neuroblastoma tumor. Signature 11, which is associated with exposure to temozolomide plus *MGMT* promoter and/or mismatch repair deficiency[47], was indeed present in tumors with previous exposure to the drug (**Table 2**). We detected the MMR2 signature in tumors of four patients (PT_0SPKM4S8, PT_3CHB9PK5, PT_JNEV57VK, and PT_VTM2STE3) diagnosed with either constitutional mismatch repair deficiency (CMMRD) or Lynch syndrome (**Table 2**), genetic predisposition syndromes caused by a variant in a mismatch repair gene such as *PMS2*, *MLH1*, *MSH2*, *MSH6*, or others[48]. Three of these patients harbored pathogenic germline variants in one of the aforementioned genes. While we did not find a *known* pathogenic variant in the germline of PT_VTM2STE3, this patient had a self-reported *PMS2* variant noted in their pathology report and we did find 19 intronic variants of unknown significance (VUS) in *PMS2*. This is not surprising since an estimated 49% of germline *PMS2* variants in patients with CMMRD and/or Lynch syndrome are VUS[48]. Interestingly, while the cell line derived from patient PT_VTM2STE3's tumor at progression was not hypermutated (TMB = 5.7 Mut/Mb), it solely showed the MMR2 signature of the eight CNS signatures examined, suggesting selective pressure to maintain a mismatch repair (MMR) phenotype *in vitro*. From patient PT_JNEV57VK, only one of the two cell lines derived from the progressive tumor was hypermutated (TMB = 35.9 Mut/Mb). This hypermutated cell line was strongly weighted towards signature 11, while this patient's non-hypermutated cell line showed a number of lesser signature weights (1, 11, 18, 19, MMR2; Table S2), highlighting the plasticity of mutational processes and the need to carefully genomically characterize and select models for preclinical studies based on research objectives.

We observed that signature 18, which has been associated with high genomic instability and can lead to a hypermutator phenotype[36], was uniformly represented among hypermutant solid tumors. Additionally, we found that all of the HGG tumors or cell lines had dysfunctional *TP53* (**Table 2**), consistent with a previous report showing *TP53* dysregulation is a dependency in tumors with high genomic instability[36]. With one exception, hypermutant and ultra-hypermutant tumors had high *TP53* scores (> 0.5) and telomerase activity. Interestingly, none of the hypermutant samples showed evidence of signature 3 (present in homologous recombination deficient tumors), signature 8 (arises from double nucleotide substitutions/unknown etiology), or

signature N6 (a universal CNS tumor signature). The mutual exclusivity of signatures 3 and MMR2 corroborates a previous report suggesting tumors do not tend to feature both deficient homologous repair and mismatch repair[38].

Table 2: ***Patients with hypermutant tumors.*** *Listed are patients with at least one hypermutant or ultra-hypermutant tumor or cell line. Pathogenic (P) or likely pathogenic (LP) germline variants, coding region TMB, phase of therapy, therapeutic interventions, cancer predisposition (CMMRD = Constitutional mismatch repair deficiency), and molecular subtypes are included.*

Kids First Participant ID	Kids First Biospecimen ID	CBTN ID	Phase of therapy	Composition	Therapy post-biopsy	Cancer predisposition	Pathogenic germline variant	TMB	OpenPBTA molecular subtype
PT_0SPKM4S8	BS_VW4XN9Y7	7316-2640	Initial CNS Tumor	Solid Tissue	Radiation, Temozolomide, CCNU	None documented	NM_000535.7(PMS2):c.137G>T (p.Ser46Ile) (LP)	187.4	HGG, H3 wildtype, TP53 activated
PT_3CHB9PK5	BS_20TBZG09	7316-515	Initial CNS Tumor	Solid Tissue	Radiation, Temozolomide, Irinotecan, Bevacizumab	CMMRD	NM_000179.3(MSH6):c.3439-2A>G (LP)	307	HGG, H3 wildtype, TP53 loss
PT_3CHB9PK5	BS_8AY2GM4G	7316-2085	Progressive	Solid Tissue	Radiation, Temozolomide, Irinotecan, Bevacizumab	CMMRD	NM_000179.3(MSH6):c.3439-2A>G (LP)	321.6	HGG, H3 wildtype, TP53 loss
PT_EB0D3BXG	BS_F0GNWEJJ	7316-3311	Progressive	Solid Tissue	Radiation, Nivolumab	None documented	None detected	26.3	Metastatic NBL, MYCN non-amplified
PT_JNEV57VK	BS_85Q5P8GF	7316-2594	Initial CNS Tumor	Solid Tissue	Radiation, Temozolomide	Lynch Syndrome	NM_000251.3(MSH2):c.1906G>C (p.Ala636Pro) (P)	4.7	DMG, H3 K28, TP53 loss
PT_JNEV57VK	BS_HM5GFJN8	7316-3058	Progressive	Derived Cell Line	Radiation, Temozolomide, Nivolumab	Lynch Syndrome	NM_000251.3(MSH2):c.1906G>C (p.Ala636Pro) (P)	35.9	DMG, H3 K28, TP53 loss
PT_JNEV57VK	BS_QWM9BPDY	7316-3058	Progressive	Derived Cell Line	Radiation, Temozolomide, Nivolumab	Lynch Syndrome	NM_000251.3(MSH2):c.1906G>C (p.Ala636Pro) (P)	7.4	DMG, H3 K28, TP53 loss
PT_JNEV57VK	BS_P0QJ1QAH	7316-3058	Progressive	Solid Tissue	Radiation, Temozolomide, Nivolumab	Lynch Syndrome	NM_000251.3(MSH2):c.1906G>C (p.Ala636Pro) (P)	6.3	DMG, H3 K28, TP53 activated
PT_S0Q27J13	BS_P3PF53V8	7316-2307	Initial CNS Tumor	Solid Tissue	Radiation, Temozolomide, Irinotecan	None documented	None detected	15.5	HGG, H3 wildtype, TP53 activated
PT_VTM2STE3	BS_ERFMPQN3	7316-2189	Progressive	Derived Cell Line	Unknown	Lynch Syndrome	None detected	5.7	HGG, H3 wildtype, TP53 loss
PT_VTM2STE3	BS_02YBZSBY	7316-2189	Progressive	Solid Tissue	Unknown	Lynch Syndrome	None detected	274.5	HGG, H3 wildtype, TP53 activated

Next, we asked whether transcriptomic classification of *TP53* dysregulation and/or telomerase activity recapitulate the known prognostic influence of these oncogenic biomarkers. To this end, we conducted a multivariate Cox regression on overall survival (**Figure 4F**; **STAR Methods**), controlling for extent of tumor resection and whether a tumor was low-grade (LGG group) or high-grade (HGG group). We identified several expected trends, including a significant overall survival benefit if the tumor had been fully resected (HR = 0.35, 95% CI = 0.2 - 0.62, $p < 0.001$) or if the tumor belonged to the LGG group (HR = 0.046, 95% CI = 0.0062 - 0.34, $p = 0.003$) as well as a significant risk if the tumor belonged to the HGG group (HR = 6.2, 95% CI = 4.0 - 9.5, $p < 0.001$). High telomerase scores were associated with poor prognosis across brain tumor histologies (HR = 20, 95% CI = 6.4 - 62, $p < 0.001$), demonstrating that `EXTEND` scores calculated from RNA-Seq are an effective rapid surrogate measure for telomerase activity. Although higher *TP53* scores, which predict *TP53* gene or pathway dysregulation, were not a significant predictor of risk across the entire OpenPBTA cohort (**Table S4**), we did find a

significant survival risk associated with higher *TP53* scores within DMGs (HR = 6436, 95% CI = 2.67 - 1.55e7, p = 0.03) and ependymomas (HR = 2003, 95% CI = 9.9 - 4.05e5, p = 0.005). Since we observed the negative prognostic effect of *TP53* scores for HGGs, we assessed the effect of molecular subtypes within HGG samples on survival risk. We found that DMG H3 K28 tumors with *TP53* loss had significantly worse prognosis (HR = 2.8, CI = 1.4-5.6, p = 0.003) than did DMG H3 K28 tumors with wildtype *TP53* (**Figure 4G** and **Figure 4H**). This finding was also recently reported in a restrospective analysis of DIPG tumors from the PNOC003 clinical trial[12].

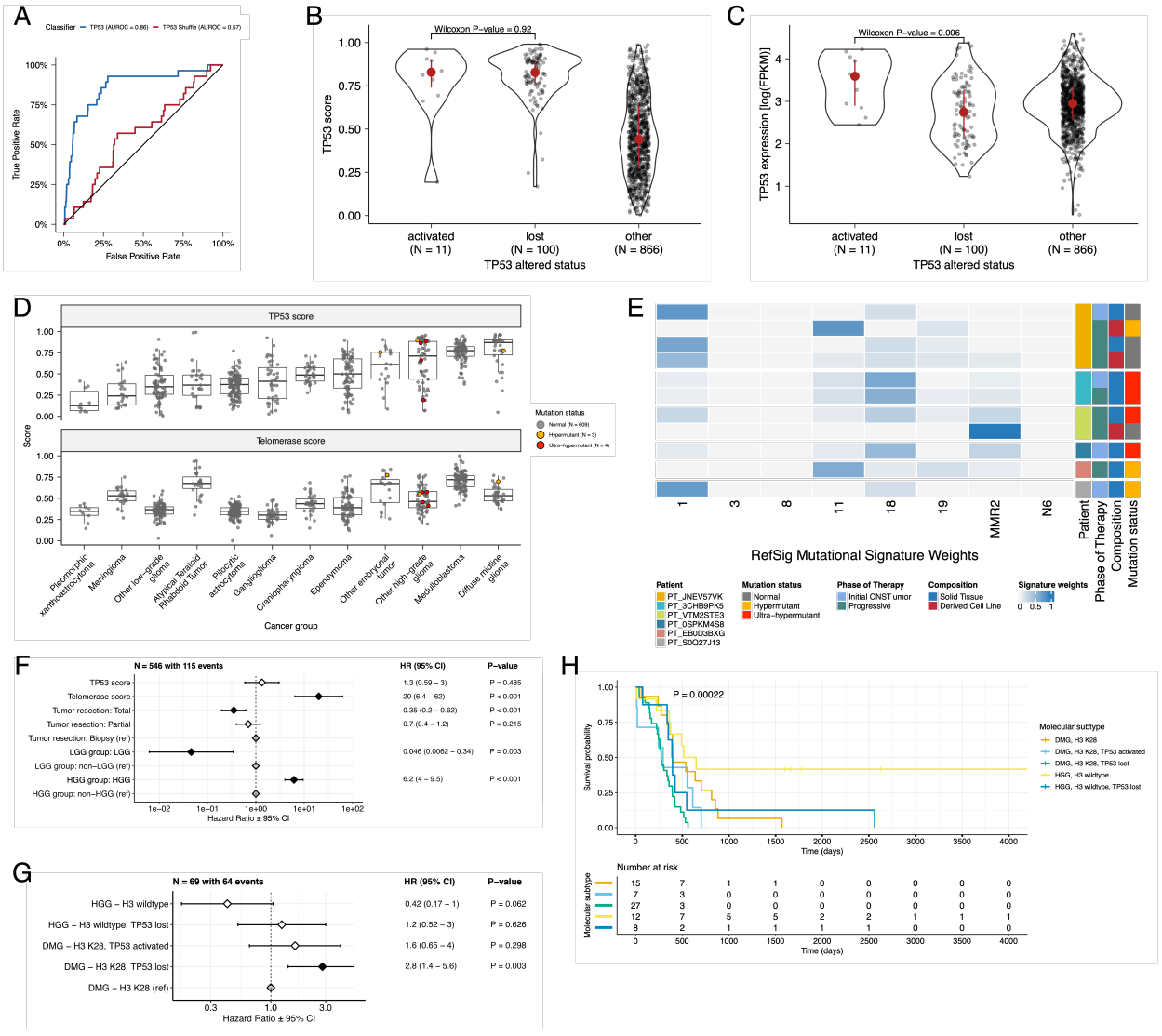


Figure 4: ***TP53 and telomerase activity*** *A, Receiver Operating Characteristic for TP53 classifier run on FPKM of stranded RNA-Seq samples. B, Violin and strip plots of TP53 scores from stranded RNA-Seq samples plotted by TP53 alteration type (Nactivated = 11, Nlost = 100, Nother = 866). C, Violin and strip plots of TP53 RNA expression from stranded RNA-Seq samples plotted by TP53 activation status (Nactivated = 11, Nlost = 100, Nother = 866). D, Box plots of TP53 and telomerase (EXTEND) scores across cancer groups. Mutation status is highlighted in orange (hypermutant) or red (ultra-hypermutant). E, Heatmap of RefSig mutational signatures for patients who have least one tumor or cell line with a hypermutant*

phenotype. F, Forest plot depicting the prognostic effects of TP53 and telomerase scores on overall survival, controlling for extent of tumor resection, LGG group, and HGG group. G, Forest plot depicting the effect of molecular subtype on overall survival of HGGs. For F and G, hazard ratios (HR) with 95% confidence intervals and p-values are listed. Significant p-values are denoted with black diamonds. Reference groups are denoted by grey diamonds. H, Kaplan-Meier curve of HGG tumors by molecular subtype.

Histologic and oncogenic pathway clustering

UMAP visualization of gene expression variation across brain tumors (**Figure 5A**) showed the expected clustering of brain tumors by histology. We additionally explored UMAP projections of gene expression within molecular subtypes for certain cancer groups. We observed that, except for three outliers, *C11orf95::RELA* (*ZFTA::RELA*) fusion-positive ependymomas fell within distinct clusters (**Figure S6A**). Medulloblastoma (MB) samples cluster by molecular subtype, with WNT and SHH in distinct clusters and Groups 3 and 4 showing some overlap (**Figure S6B**), as expected. Of note, two MB samples annotated as the SHH subtype did not cluster with the other MB samples, and one clustered with Group 3 and 4 samples, suggesting potential subtype misclassification or different underlying biology of these two tumors. *BRAF*-driven low-grade gliomas (**Figure S6C**) were present in three separate clusters, suggesting that there might be additional shared biology within each cluster. Histone H3 G35-mutant high-grade gliomas generally clustered together and away from K28-mutant tumors (**Figure S6D**). Interestingly, although H3 K28-mutant tumors have different biological drivers than do H3 wildtype tumors[49], they did not form distinct clusters. This pattern suggests these subtypes may be driven by common transcriptional programs, have other much stronger biological drivers than their known distinct epigenetic drivers, or our sample size is too small to detect transcriptional differences.

We next performed gene set variant analysis (GSVA) for Hallmark cancer gene sets to demonstrate activation of underlying oncogenic pathways (**Figure 5B** and quantified immune cell fractions across OpenPBTA tumors using quanTIseq (**Figure 5C** and **Figure S6E**). Through these analyses, we were able to recapitulate previously-described tumor biology. For example, HGG, DMG, MB, and ATRT tumors are known to upregulate *MYC*[50] which in turn activates *E2F* and S phase[51]. Indeed, we detected significant (Bonferroni-corrected $p < 0.05$) upregulation of *MYC* and *E2F* targets, as well as G2M (cell cycle phase following S phase) in MBs, ATRTs, and HGGs compared to several other cancer groups. In contrast, LGGs showed significant downregulation (Bonferroni-corrected $p < 0.05$) of these pathways. Schwannomas and neurofibromas, which have a documented inflammatory immune microenvironment of T and B lymphocytes as well as tumor-associated macrophages (TAMs), are driven by upregulation of cytokines such as IFNγ, IL-1, and IL-6, and TNFα[52]. Indeed, we observed significant upregulation of these cytokines in GSVA hallmark pathways (Bonferroni-corrected $p < 0.05$) (**Figure 5B**) and found immune cell types dominated by monocytes in these tumors (**Figure 5C**). We also observed significant upregulation of pro-inflammatory cytokines IFNα and IFNγ in LGGs and craniopharyngiomas compared to medulloblastoma and ependymoma tumors (Bonferroni-corrected $p < 0.05$), both of which showed significant down-regulation of these cytokines (**Figure 5B**). Together, these results supported previous proteogenomic findings of

lower immune infiltration in aggressive medulloblastomas and ependymomas versus higher immune infiltration in *BRAF*-driven LGG and craniopharyngiomas[53].

Although CD8+ T-cell infiltration across all cancer groups was quite low (**Figure 5C**), we observed some signal in specific cancer molecular subtypes (Groups 3 and 4 medulloblastoma) as well as outlier tumors (BRAF-driven LGG, BRAF-driven and wildtype ganglioglioma, and CNS embryonal NOS; **Figure S6E**) Surprisingly, the classically immunologically-cold HGG and DMG tumors[54,55] contained higher overall fractions of immune cells, where monocytes, dendritic cells, and NK cells were the most prevalent (**Figure 5C**). Thus, we suspect that quanTIseq might actually have captured microglia within these immune cell fractions.

While we did not detect notable prognostic effects of immune cell infiltration on overall survival in HGG or DMG tumors, we did find that high levels of macrophage M1 and monocytes were associated with poorer overall survival (monocyte HR = 2.1e18, 95% CI = 3.80e5 - 1.2e31, p = 0.005) in medulloblastoma tumors (**Figure 5D**). We further reproduced previous findings (**Figure 5E**) that medulloblastomas typically have low expression of *CD274* (PD-L1)[56]. However, we also found that higher expression of *CD274* was significantly associated with improved overall prognosis for medulloblastoma samples, although with a marginal effect size (HR = 0.0012, 95% CI = 7.5e−06 - 0.18, p = 0.008) (**Figure 5D**). This result may be explained by the higher expression of *CD274* found in WNT subtype tumors by us and others[57], as this diagnosis carries the best prognosis of all medulloblastoma subgroups (**Figure 5E**).

Finally, we asked whether any molecular subtypes might show an immunologically-hot phenotype, as roughly defined by a greater proportion of CD8+ to CD4+ T cells[58,59]. While adamantinomatous craniopharyngiomas and Group 3 and Group 4 medulloblastomas had the highest CD8+ to CD4+ T cell ratios (**Figure S6F**), very few tumors had ratios greater than 1, highlighting an urgent need to identify novel therapeutics for these immunologically-cold pediatric brain tumors with poor prognosis.

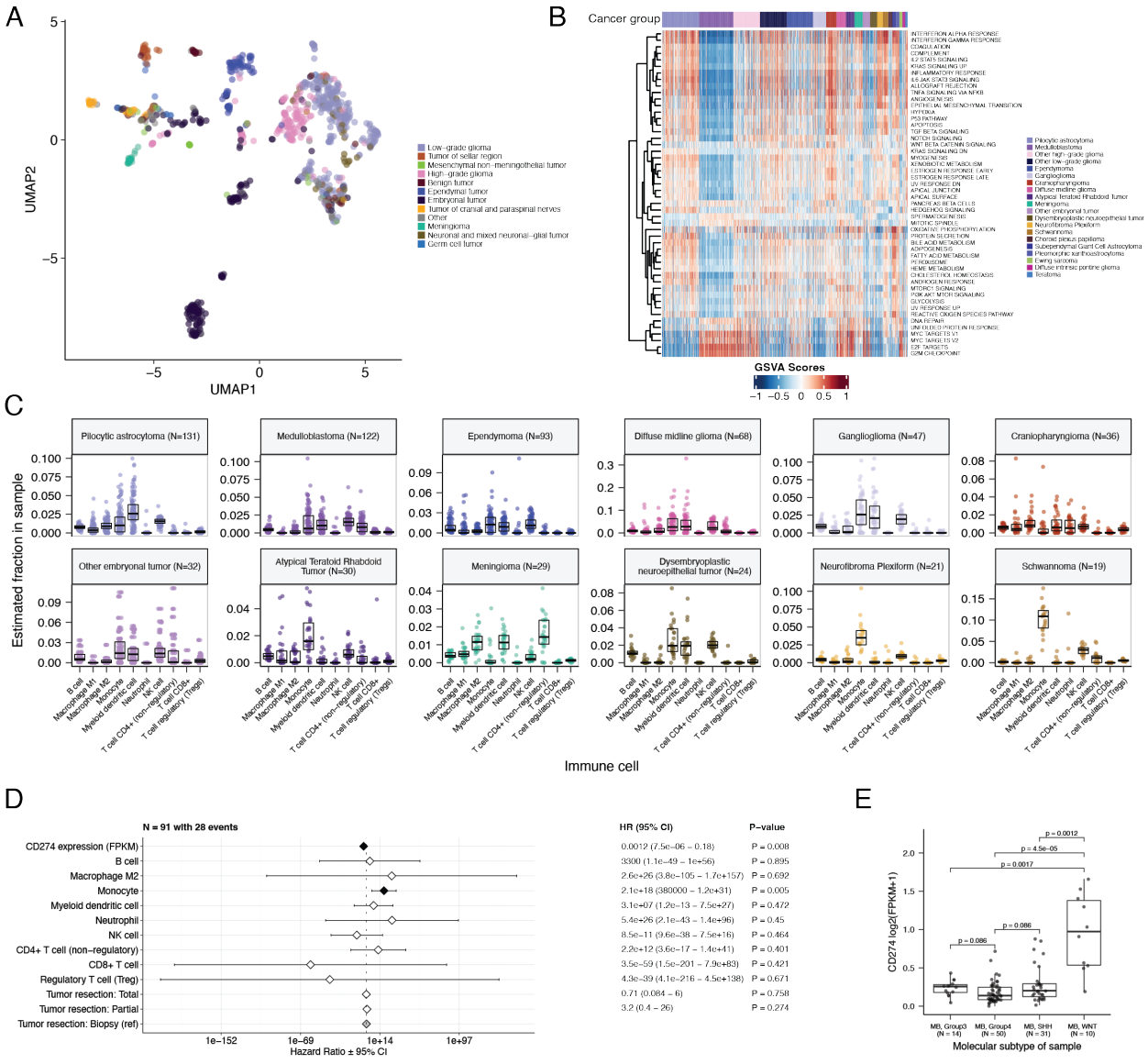


Figure 5: ***Transcriptomic and immune landscape of pediatric brain tumors*** *A, First two dimensions from UMAP of sample transcriptome data. Points are colored by the broad histology of the samples they represent. B, Heatmap of GSVA scores for Hallmark gene sets with significant differences, with samples ordered by cancer group. C, Box plots of quanTIseq estimates of immune cell proportions in select cancer groups with N > 15 samples. Note: Other HGGs and other LGGs have immune cell proportions similar to DMG and pilocytic astrocytoma, respectively, and are not shown. D, Forest plot depicting the additive effects of CD274 expression, immune cell proportion, and extent of tumor resection on overall survival of medulloblastoma patients. Hazard ratios (HR) with 95% confidence intervals and p-values are listed. Significant p-values are denoted with black diamonds. Reference groups are denoted by grey diamonds. Of note, the Macrophage M1 HR was 0 (coefficient = -9.90e+4) with infinite upper and lower CIs, and thus it was not included in the figure. E, Box plot of CD274 expression*

(log2 FPKM) for medulloblastoma samples grouped by molecular subtype. Bonferroni-corrected p-values from Wilcoxon tests are shown.

Discussion

We created OpenPBTA to define an open, real-time, reproducible analysis framework to genomically characterize pediatric brain tumors that brings together basic and translational researchers, clinicians, and data scientists on behalf of accelerated discovery and clinical impact. We provide robust reusable code and data resources, paired with cloud-based availability of source and derived data resources, to the pediatric oncology community, encouraging interdisciplinary scientists to collaborate on new analyses in order to accelerate therapeutic translation for children with cancer, goals we are seeing play out in real-time. To our knowledge, this initiative represents the first large-scale, collaborative, open analysis of genomic data coupled with open manuscript writing, in which we comprehensively analyzed the largest cohort of pediatric brain tumors to date, comprising 1,074 tumors across 58 distinct histologies. We used available WGS, WXS, and RNA-Seq data to generate high-confidence consensus SNV and CNV calls, prioritize putative oncogenic fusions, and establish over 40 scalable modules to perform common downstream cancer genomics analyses, all of which have undergone rigorous scientific and analytical code review. We detected and showed expected patterns of genomic lesions, mutational signatures, and aberrantly regulated signaling pathways across multiple pediatric brain tumor histologies.

Assembling large, pan-histology cohorts of fresh frozen samples and associated clinical phenotypes and outcomes requires a multi-year, multi-institutional framework, like those provided by CBTN and PNOC. As such, uniform clinical molecular subtyping was largely not performed for most of this cohort at the time of diagnosis and/or at surgery, and when available (e.g., sparse medulloblastoma subtypes), it required manual curation from pathology reports and/or free text clinical data fields. Furthermore, rapid classification to derive molecular subtypes could not be immediately performed since research-based DNA methylation data for these samples are not yet available. Thus, to enable biological interrogation of specific tumor subtypes, we created RNA- and DNA-based subtyping modules aligned with WHO molecularly-defined diagnoses. We worked closely with pathologists and clinicians to build modules from which we determined a research-grade integrated diagnosis for 60% of samples while discovering incorrectly diagnosed or mis-identified samples in the OpenPBTA cohort.

We harnessed RNA expression data for a number of analyses, yielding important biological insights across multiple brain tumor histologies. For example, we performed subtyping of medulloblastoma tumors, for which only 35% (43/122) had subtype information from pathology reports. Among the subtyped tumors, we accurately recapitulated subtypes using MM2S (91%; 39/43) or medulloPackage (95%; 41/43)[60,61]. We then applied the consensus of these methods to subtype all medulloblastoma tumors lacking pathology-based subtypes.

We advanced the integrative analyses and cross-cohort comparison via a number of validated modules. We used an expression classifier to determine whether tumors have dysfunctional *TP53*[38] and the EXTEND algorithm to determine their degree of telomerase activity using a 13-gene signature[41]. Interestingly, in contrast to adult colorectal cancer and gastric

adenocarcinoma, in which *TP53* loss of function is less frequent in hypermutated tumors[62,63], we found that hypermutant HGG tumors universally displayed dysregulation of *TP53*. Furthermore, high *TP53* scores were a significant prognostic marker for poor overall survival for patients with certain tumor types, such as H3 K28-altered DMGs and ependymomas. We also show that EXTEND scores are a robust surrogate measure for telomerase activity in pediatric brain tumors. By assessing *TP53* and telomerase activity prospectively from expression data, information usually only attainable with DNA sequencing and/or qPCR, we can quickly incorporate oncogenic biomarker and prognostic knowledge and expand our biological understanding of these tumors.

We identified enrichment of hallmark cancer pathways and characterized the immune cell landscape across pediatric brain tumors, demonstrating tumors in some histologies, such as schwannomas, craniopharyngiomas, and low-grade gliomas, may have a inflammatory tumor microenvironment. Of note, we observed upregulation of IFNγ, IL-1, and IL-6, and TNFα in craniopharyngiomas, tumors difficult to resect due to their anatomical location and critical surrounding structures. Neurotoxic side effects have been reported when interferon alpha immunotherapy is administered to reduce cystic craniopharyngioma tumor size and/or delay progression[64,65]. Thus, additional immune vulnerabilities, such as IL-6 inhibition and immune checkpoint blockade, have recently been proposed as therapies for cystic adamantinomatous craniopharyngiomas[66–70] and our results noted above support this approach. Finally, our study reproduced the overall known poor infiltration of CD8+ T cells and general low expression of *CD274* (PD-L1) in pediatric brain tumors, further highlighting the urgent need to identify novel therapeutic strategies for these immunologically cold tumors.

OpenPBTA has rapidly become a foundational data analysis and processing layer for a number of discovery research and translational projects which will continue to add other genomic modalities and analyses, such as germline, methylation, single cell, epigenomic, mRNA splicing, imaging, and model drug response data. For example, the RNA fusion filtering module created within OpenPBTA set the stage for development of the R package *annoFuse*[71] and an R Shiny application *shinyFuse*. Using medulloblastoma subtyping and immune deconvolution analyses performed herein, Dang and colleagues showed enrichment of monocyte and microglia-derived macrophages within the SHH subgroup which they suggest may accumulate following radiation therapy[10]. Expression and copy number analyses were used to demonstrate that *GPC2* is a highly expressed and copy number gained immunotherapeutic target in ETMRs, medulloblastomas, choroid plexus carcinomas, H3 wildtype high-grade gliomas, as well as DMGs. This led Foster and colleagues to subsequently develop a chimeric antigen receptor (CAR) directed against *GPC2*, for which they show preclinical efficacy in mouse models[11]. Moreover, OpenPBTA has enabled a framework to support real-time integration of clinical trial subjects as each was enrolled on the PNOC008 high-grade glioma clinical trial[72], allowing researchers and clinicians to link tumor biology to translational impact through clinical decision support during tumor board discussions. Finally, as part of the the NCI's Childhood Cancer Data Initiative (CCDI), the OpenPBTA project was recently expanded into a pan-pediatric cancer effort (https://github.com/PediatricOpenTargets/OpenPedCan-analysis) to build the Molecular Targets Platform (https://moleculartargets.ccdi.cancer.gov/) in support of the RACE Act. An additional, large-scale cohort of >2,500 tumor samples and associated germline DNA is in the process of undergoing sequence data generation as part of CBTN CCDI-Kids First NCI and

Common Fund project (https://commonfund.nih.gov/kidsfirst/2021X01projects#FY21_Resnick). Like the original OpenPBTA cohort, data will be processed and released in near real-time via the Kids First Data Resource and integrated with OpenPBTA. The OpenPBTA project has paved the way for new modes of collaborative data-driven discovery, open, reproducible, and scalable analyses that will extend beyond the current research described herein, and we anticipate this foundational work will continue to have a long-term impact within the pediatric brain tumor translational research community and beyond, ultimately leading to accelerated impact and improved outcomes for children with cancer.

All code and processed data are openly available through GitHub, CAVATICA, and PedcBioPortal (see **STAR METHODS**).

Acknowledgments

We graciously thank the patients and families who have donated their tumors to the Children's Brain Tumor Network (CBTN) and/or the Pacific Pediatric Neuro-oncology Consortium, without which this research would not be possible.

Philanthropic support has ensured the CBTN's ability to collect, store, manage, and distribute specimen and data since it was founded in 2013. In addition to the support from the CBTN Executive Council members and Brain Tumor Board of Visitors, the following donors have provided leadership level support for CBTN: Children's Brain Tumor Foundation, Easie Family Foundation, Kortney Rose Foundation, Lilabean Foundation, Minnick Family Charitable Fund, Perricelli Family, Psalm 103 Foundation, and Swifty Foundation.

This work was funded through the Alex's Lemonade Stand Foundation (ALSF) Childhood Cancer Data Lab (CSG), ALSF Young Investigator Award (JLR), ALSF Catalyst Award (JLR, ACR, PBS), ALSF Catalyst Award (SJS), ALSF CCDL Postdoctoral Training Grant (SMF), Children's Hospital of Philadelphia Division of Neurosurgery (PBS and ACR), the Australian Government, Department of Education (APH), the St. Anna Kinderkrebsforschung, Austria (ARP), the Mildred Scheel Early Career Center Dresden P2, funded by the German Cancer Aid (ARP), and NIH Grants 3P30 CA016520-44S5 (ACR), U2C HL138346-03 (ACR, APH), U24 CA220457-03 (ACR), K12GM081259 (SMF), R03-CA23036 (SJD), and NIH Contract No. HHSN261200800001E (SJD). This project has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. 75N91019D00024, Task Order No. 75N91020F00003 (JLR, ACR, APH). Additionally, this work was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics of the National Cancer Institute. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government.

The authors would like to thank the following collaborators who contributed or supervised analyses present in the analysis repository that were not included in the manuscript: William Amadio, Holly C. Beale, Ellen T. Kephart, A. Geoffrey Lyle, and Olena M. Vaske. Finally, we would like to thank Yuanchao Zhang and Eric Wafula for adding to the project codebase,

Jessica B. Foster for helpful discussions while drafting the manuscript, and Gina D. Mawla for identifying and reporting issues in OpenPBTA data.

Author Contributions

Author	Contributions
Joshua A. Shapiro	Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original draft, Writing - Review and editing, Visualization, Supervision
Krutika S. Gaonkar	Data curation, Formal analysis, Investigation, Methodology, Software, Writing – Original draft, Writing - Review and editing
Candace L. Savonen	Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original draft, Writing - Review and editing, Visualization
Stephanie J. Spielman	Validation, Formal analysis, Writing - Review and editing, Visualization, Supervision, Funding acquisition
Chante J. Bethell	Methodology, Validation, Formal analysis, Investigation, Writing - Original draft, Visualization
Run Jin	Data curation, Formal analysis, Visualization, Writing – Original draft, Writing - Review and editing
Komal S. Rathi	Formal analysis, Investigation, Methodology, Writing – Original draft
Yuankun Zhu	Data curation, Formal analysis, Investigation, Methodology, Supervision
Laura E. Egolf	Formal analysis, Writing - Original draft
Bailey K. Farrow	Data curation, Software
Daniel P. Miller	Formal analysis
Yang Yang	Formal analysis, Software
Tejaswi Koganti	Formal analysis, Investigation
Nighat Noureen	Formal analysis, Visualization, Writing - Original draft
Mateusz P. Koptyra	Formal analysis, Writing – Original draft
Nhat Duong	Formal analysis, Investigation, Methodology
Adam A. Kraya	Methodology
Cassie Kline	Supervision, Investigation, Writing - Review and editing
Yiran Guo	Formal analysis, Writing - Review and editing
Phillip B. Storm	Conceptualization, Funding acquisition, Resources
Javad Nazarian	Conceptualization
Stephen C. Mack	Writing - Review and editing
Pichai Raman	Conceptualization, Formal analysis, Methodology
Siyuan Zheng	Formal analysis, Visualization, Writing - Original draft, Supervision, Writing - Review and editing
Peter J. Madsen	Writing – Review & editing
Nicolas Van Kuren	Data curation, Software
Shannon Robins	Data curation

Author	Contributions
Xiaoyan Huang	Formal analysis
Angela J. Waanders	Supervision, Conceptualization
Jung Kim	Investigation, Writing - Review and editing
Derek Hanson	Validation
Carl J. Koschmann	Conceptualization
Sharon J. Diskin	Investigation, Supervision, Validation, Funding acquisition, Writing - Review and editing
Rishi R. Lulla	Conceptualization
Miguel A. Brown	Data curation, Methodology, Formal analysis
Jessica Wong	Writing – Original draft
Jennifer L. Mason	Supervision
Laura Scolaro	Data curation
Meen Chul Kim	Data curation
Hongbo M. Xie	Methodology, Supervision
Brian R. Rood	Conceptualization
Kristina A. Cole	Writing - Review and editing
Ashley S. Margol	Writing - Review and editing
Zalman Vaksman	Formal analysis, Investigation
Rebecca S. Kaufman	Formal analysis, Investigation, Validation
Allison P. Heath	Project administration, Funding acquisition
Brian M. Ennis	Data curation, Formal analysis
Mariarita Santi	Investigation, Validation, Writing - Review and editing
Anna R. Poetsch	Formal analysis, Funding acquisition, Writing – Review and editing
Angela N. Viaene	Investigation, Validation
Douglas R. Stewart	Supervision, Writing - Review and editing
Steven M. Foltz	Validation, Funding acquisition
Payal Jain	Data curation, Investigation, Validation
Bo Zhang	Data curation, Formal analysis
Shrivats Kannan	Formal analysis, Methodology, Writing – Original draft
Michael Prados	Conceptualization
Jena V. Lilly	Conceptualization, Funding acquisition, Project administration
Sabine Mueller	Conceptualization
Adam C. Resnick	Conceptualization, Funding acquisition, Resources, Supervision
Casey S. Greene	Conceptualization, Funding acquisition, Methodology, Project administration, Software, Supervision, Writing – Review & editing
Jo Lynne Rokita	Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Writing – Original draft, Writing - Review and editing

Author	Contributions
Jaclyn N. Taroni	Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - Review and editing, Visualization, Supervision, Project administration
Children's Brain Tumor Network	Conceptualization
Pacific Pediatric Neuro-Oncology Consortium	Conceptualization

Except for the first and last four authors, authorship order was determined as follows: Authors who contributed to the OpenPBTA code base are listed based on number of modules included in the manuscript to which that individual contributed and, in the case of ties, a random order is used. All remaining authors are then listed in a random order.

Code for determining authorship order can be found in the `count-contributions` module of the OpenPBTA analysis repository.

Declarations of Interest

CSG's spouse was an employee of Alex's Lemonade Stand Foundation, which was a sponsor of this research. JAS, CLS, CJB, SJS, and JNT are or were employees of Alex's Lemonade Stand Foundation, a sponsor of this research. AJW is a member of the Scientific Advisory boards for Alexion and DayOne Biopharmaceuticals.

Figure Titles and Legends

Figure 1. Overview of the OpenPBTA Project. A, The Children's Brain Tumor Network and the Pacific Pediatric Neuro-Oncology Consortium collected tumor samples from 943 patients. To date, 22 cell lines were created from tumor tissue, and over 2000 specimens were sequenced (N = 1035 RNA-Seq, N = 940 WGS, and N = 32 WXS or targeted panel). Data was harmonized by the Kids First Data Resource Center using an Amazon S3 framework within CAVATICA. B, Stacked bar plot summary of the number of biospecimens per phase of therapy per broad histology (Abbreviations: GNG = ganglioglioma, Other LGG = other low-grade glioma, PA = pilocytic astrocytoma, PXA = pleomorphic xanthoastrocytoma, SEGA = subependymal giant cell astrocytoma, DIPG = diffuse intrinsic pontine glioma, DMG = diffuse midline glioma, Other HGG = other high-grade glioma, ATRT = atypical teratoid rhabdoid tumor, MB = medulloblastoma, Other ET = other embryonal tumor, EPN = ependymoma, PNF = plexiform neurofibroma, DNET = dysembryoplastic neuroepithelial tumor, CRANIO = craniopharyngioma, EWS = Ewing sarcoma, CPP = choroid plexus papilloma). Only samples with available descriptors were included. C, Overview of the open analysis and manuscript contribution model. In the analysis GitHub repository, a contributor would propose an analysis that other participants can comment on. Contributors would then implement the analysis and file a request to add their changes to the analysis repository ("pull request"). Pull requests underwent review for scientific rigor and correctness of implementation. Pull requests were additionally checked to ensure that all software dependencies were included and the code was not sensitive to underlying data

874 changes using container and continuous integration technologies. Finally, a contributor would
875 file a pull request documenting their methods and results to the Manubot-powered manuscript
876 repository. Pull requests in the manuscript repository were also subject to review. D, A potential
877 path for an analytical pull request. Arrows indicate revisions to a pull request. Prior to review, a
878 pull request was tested for dependency installation and whether or not the code would execute.
879 Pull requests also required approval by organizers and/or other contributors, who checked for
880 scientific correctness. Panel A created with BioRender.com.

881 **Figure 2. Mutational landscape of PBTA tumors.** Shown are frequencies of canonical
882 somatic gene mutations, CNVs, fusions, and TMB (top bar plot) for the top 20 genes mutated
883 across primary tumors within the OpenPBTA dataset. A, Low-grade astrocytic tumors (N = 227):
884 pilocytic astrocytoma (N = 104), other low-grade glioma (N = 69), ganglioglioma (N = 35),
885 pleomorphic xanthoastrocytoma (N = 9), subependymal giant cell astrocytoma (N = 10); B,
886 Embryonal tumors (N = 128): medulloblastomas (N = 95), atypical teratoid rhabdoid tumors (N =
887 24), other embryonal tumors (N = 9); C, Diffuse astrocytic and oligodendroglial tumors (N = 61):
888 diffuse midline gliomas (N = 34) and other high-grade gliomas (N = 27); D, Other CNS tumors
889 (N = 194): ependymomas (N = 60), craniopharyngiomas (N = 31), meningiomas (N = 17),
890 dysembryoplastic neuroepithelial tumors (N = 19), Ewing sarcomas (N = 7), schwannomas (N =
891 11), and neurofibroma plexiforms (N = 7). Additional, rare CNS tumors are displayed in **Figure**
892 **S3A**. Tumor histology (`Cancer Group`) and patient sex (`Germline sex estimate`) are
893 displayed as annotations at the bottom of each plot. Only samples with mutations in the listed
894 genes are shown. Multiple CNVs are denoted as a complex event.

895 **Figure 3. Mutational co-occurrence and signatures highlight key oncogenic drivers.** A,
896 Bar plot of occurrence and co-occurrence of nonsynonymous mutations for the 50 most
897 commonly mutated genes across all tumor types (annotated from `cancer_group` if N >= 10 or
898 `Other` if N < 10); B, Co-occurrence and mutual exclusivity of nonsynonymous mutations
899 between genes; The co-occurrence score is defined as $I\left(-\log_{10}(P)\right)$ where P is defined by
900 Fisher's exact test and I is 1 when mutations co-occur more often than expected and -1 when
901 exclusivity is more common; C, The number of SV breaks significantly correlates with the
902 number of CNV breaks (Adjusted R = 0.438, p = 1.08e-37). D, Chromothripsis frequency across
903 pediatric brain tumors shown by `cancer_group` with N >= 3. E, Sina plots of RefSig signature
904 weights for signatures 1, 11, 18, 19, 3, 8, N6, MMR2, and Other across cancer groups. Box plot
905 lines represent the first quartile, median, and third quartile.

906 **Figure 4. *TP53* and telomerase activity** A, Receiver Operating Characteristic for *TP53*
907 classifier run on FPKM of stranded RNA-Seq samples. B, Violin and box plots of *TP53* scores
908 plotted by *TP53* alteration type (Nactivated = 11, Nlost = 100, Nother = 866). C, Violin and box
909 plots of *TP53* RNA expression plotted by *TP53* activation status (Nactivated = 11, Nlost = 100,
910 Nother = 866). D, Box plots of *TP53* and telomerase (EXTEND) scores grouped by
911 `cancer_group`. Mutation status is highlighted in orange (hypermutant) or red (ultra-
912 hypermutant). E, Heatmap of RefSig mutational signatures for patients who have least one
913 tumor or cell line with a TMB >= 10 Mut/Mb. F, Forest plot depicting the prognostic effects of
914 *TP53* and telomerase scores on overall survival, controlling for extent of tumor resection, LGG
915 group, and HGG group. G, Forest plot depicting the effect of molecular subtype on overall
916 survival of HGGs. For F and G, hazard ratios (HR) with 95% confidence intervals and p-values

are listed. Significant p-values are denoted with black diamonds. Reference groups are denoted by grey diamonds. H, Kaplan-Meier curve of HGG tumors by molecular subtype.

Figure 5. Transcriptomic and immune landscape of pediatric brain tumors A, First two dimensions from UMAP of sample transcriptome data. Points are colored by the broad histology of the samples they represent. B, Heatmap of GSVA scores for Hallmark gene sets with significant differences, with samples ordered by cancer group. C, Box plots of quanTIseq estimates of immune cell proportions in select cancer groups with N > 15 samples. Note: Other HGGs and other LGGs have immune cell proportions similar to DMG and pilocytic astrocytoma, respectively, and are not shown. D, Forest plot depicting the additive effects of *CD274* expression, immune cell proportion, and extent of tumor resection on overall survival of medulloblastoma patients. Hazard ratios (HR) with 95% confidence intervals and p-values are listed. Significant p-values are denoted with black diamonds. Reference groups are denoted by grey diamonds. Of note, the Macrophage M1 HR was 0 (coefficient = -9.90e+4) with infinite upper and lower CIs, and thus it was not included in the figure. E, Box plot of *CD274* expression (log2 FPKM) for medulloblastoma samples grouped by molecular subtype. Bonferroni-corrected p-values from Wilcoxon tests are shown.

Table Titles and Legends

Table 1. Molecular subtypes generated through the OpenPBTA project. Listed are broad tumor histologies, molecular subtypes generated, and number of specimens subtyped within the OpenPBTA project.

Table 2. Patients with hypermutant tumors. Listed are patients with at least one hypermutant or ultra-hypermutant tumor or cell line. Pathogenic (P) or likely pathogenic (LP) germline variants, coding region TMB, phase of therapy, therapeutic interventions, cancer predisposition (CMMRD = Constitutional mismatch repair deficiency), and molecular subtypes are included.

STAR METHODS

RESOURCE AVAILABILITY

Lead contact

Requests for access to OpenPBTA raw data and/or specimens may be directed to, and will be fulfilled by Jo Lynne Rokita (rokita@chop.edu).

Materials availability

This study did not create new, unique reagents.

Data and code availability

Raw and harmonized WGS, WXS, and RNA-Seq data derived from human samples are available within the KidsFirst Portal[73] upon access request to the CBTN (https://cbtn.org/) as of

the date of the publication. In addition, merged summary files are openly accessible at https://cavatica.sbgenomics.com/u/cavatica/openpbta or via download script in the https://github.com/AlexsLemonade/OpenPBTA-analysis repository. Summary data are visible within PedcBioPortal at https://pedcbioportal.kidsfirstdrc.org/study/summary?id=openpbta. Associated DOIs are listed in the **Key Resources Table**.

All original code was developed within the following repositories and is publicly available as follows. Primary data analyses can be found at https://github.com/d3b-center/OpenPBTA-workflows. Downstream data analyses can be found at https://github.com/AlexsLemonade/OpenPBTA-analysis. Manuscript code can be found at https://github.com/AlexsLemonade/OpenPBTA-manuscript. Associated DOIs are listed in the **Key Resources Table**. Software versions are documented in **Table S5** as an appendix to the **Key Resources Table**.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Biospecimen Collection

The Pediatric Brain Tumor Atlas specimens are comprised of samples from Children's Brain Tumor Network (CBTN) and the Pediatric Pacific Neuro-Oncology Consortium (PNOC). The CBTN is a collaborative, multi-institutional (26 institutions worldwide) research program dedicated to the study of childhood brain tumors. PNOC is an international consortium dedicated to bringing new therapies to children and young adults with brain tumors. We also include blood and tumor biospecimens from newly-diagnosed diffuse intrinsic pontine glioma (DIPG) patients as part of the PNOC003 clinical trial PNOC003/NCT02274987[15].

The CBTN-generated cell lines were derived from either fresh tumor tissue directly obtained from surgery performed at Children's Hospital of Philadelphia (CHOP) or from prospectively collected tumor specimens stored in Recover Cell Culture Freezing medium (cat# 12648010, Gibco). We dissociated tumor tissue using enzymatic method with papain as described[14]. Briefly, we washed tissue with HBSS (cat# 14175095, Gibco), and we minced and incubated the tissue with activated papain solution (cat# LS003124, SciQuest) for up to 45 minutes. We used ovomucoid solution (cat# 542000, SciQuest) to inactivate the papain, briefly treated tissue with DNase (cat# 10104159001, Roche), passed it through the 100µm cell strainer (cat# 542000, Greiner Bio-One). We initiated two cell culture conditions based on the number of cells available. For cultures utilizing the fetal bovine serum (FBS), we plated a minimum density of 3×105 cells/mL in DMEM/F-12 medium (cat# D8062, Sigma) supplemented with 20% FBS (cat# SH30910.03, Hyclone), 1% GlutaMAX (cat# 35050061, Gibco), Penicillin/Streptomycin-Amphotericin B Mixture (cat# 17-745E, Lonza), and 0.2% Normocin (cat# ant-nr-2, Invivogen). For serum-free media conditions, we plated cells at minimum density of 1×106 cells/mL in DMEM/F12 medium supplemented with 1% GlutaMAX, 1X B-27 supplement minus vitamin A (cat# 12587-010, Gibco), 1x N-2 supplement (cat# 17502001, Gibco), 20 ng/ml epidermal growth factor (cat# PHG0311L, Gibco), 20 ng/mL basic fibroblast growth factor (cat# 100-18B,

PeproTech), 2.5µg/mL heparin (cat# H3149, Sigma), Penicillin/Streptomycin-Amphotericin B Mixture, and 0.2% Normocin.

Nucleic acids extraction and library preparation

PNOC samples

The Translational Genomic Research Institute (TGEN; Phoenix, AZ) performed DNA and RNA extractions on tumor biopsies using a DNA/RNA AllPrep Kit (Qiagen, #80204). All RNA used for library prep had a minimum RIN of seven, but no QC thresholds were implemented for the DNA. For library preparation, 500 ng of nucleic acids were used as input for RNA-Seq, WXS, and targeted DNA panel (panel) sequencing. RNA library preparation was performed using the TruSeq RNA Sample Prep Kit (Illumina, #FC-122-1001) and the exome prep was performed using KAPA Library Preparation Kit (Roche, #KK8201) using Agilent's SureSelect Human All Exon V5 backbone with custom probes. The targeted DNA panel developed by Ashion Analytics (formerly known as the GEM Cancer panel) consisted of exonic probes against 541 cancer genes. Both panel and WXS assays contained 44,000 probes across evenly spaced genomic loci used for genome-wide copy number analysis. For the panel, additional probes tiled across intronic regions of 22 known tumor suppressor genes and 22 genes involved in common cancer translocations for structural analysis. All extractions and library preparations were performed according to manufacturer's instructions.

CBTN samples

Blood, tissue, and cell line DNA/RNA extractions were performed at the Biorepository Core at CHOP. Briefly, 10-20 mg frozen tissue, 0.4-1ml of blood, or 2e6 cells pellet was used for extractions. Tissues were lysed using a Qiagen TissueLyser II (Qiagen) with 2×30 sec at 18Hz settings using 5 mm steel beads (cat# 69989, Qiagen). Both tissue and cell pellets processes included a CHCl3 extraction and were run on the QIACube automated platform (Qiagen) using the AllPrep DNA/RNA/miRNA Universal kit (cat# 80224, Qiagen). Blood was thawed and treated with RNase A (cat#, 19101, Qiagen); 0.4-1ml was processed using the Qiagen QIAsymphony automated platform (Qiagen) using the QIAsymphony DSP DNA Midi Kit (cat# 937255, Qiagen). DNA and RNA quantity and quality was assessed by PerkinElmer DropletQuant UV-VIS spectrophotometer (PerkinElmer) and an Agilent 4200 TapeStation (Agilent, USA) for RIN and DIN (RNA Integrity Number and DNA Integrity Number, respectively). The NantHealth Sequencing Center, BGI at CHOP, or the Genomic Clinical Core at Sidra Medical and Research Center performed library preparation and sequencing. BGI at CHOP and Sidra Medical and Research Center used in house, center-specific workflows for sample preparation. At NantHealth Sequencing Center, DNA sequencing libraries were prepared for tumor and matched-normal DNA using the KAPA HyperPrep kit (cat# 08098107702, Roche), and tumor RNA-Seq libraries were prepared using KAPA Stranded RNA-Seq with RiboErase kit (cat# 07962304001, Roche).

Data generation

NantHealth and Sidra performed 2x150 bp WGS on paired tumor (~60X) and constitutive DNA (~30X) samples on an Illumina X/400. BGI at CHOP performed 2x100 bp WGS sequenced at

60X depth for both tumor and normal samples. NantHealth performed ribosomal-depleted whole transcriptome stranded RNA-Seq to an average depth of 200M. BGI at CHOP performed poly-A or ribosomal-depleted whole transcriptome stranded RNA-Seq to an average depth of 100M. The Translational Genomic Research Institute (TGEN; Phoenix, AZ) performed paired tumor (~200X) and constitutive whole exome sequencing (WXS) or targeted DNA panel (panel) and poly-A selected RNA-Seq (~200M reads) for PNOC tumor samples. The panel tumor sample was sequenced to 470X, and the normal panel sample was sequenced to 308X. PNOC 2x100 bp WXS and RNA-Seq libraries were sequenced on an Illumina HiSeq 2500.

DNA WGS Alignment

We used `BWA-MEM`[74] to align paired-end DNA-seq reads to the version 38 patch release 12 of the *Homo sapiens* genome reference, obtained as a FASTA file from UCSC (see **Key Resources Table**). Next, we used the Broad Institute's Best Practices[75] to process Binary Alignment/Map files (BAMs) in preparation for variant discovery. We marked duplicates using `SAMBLASTER`[76], and we merged and sorted BAMs using `Sambamba`[77] We used the `BaseRecalibrator` submodule of the Broad's Genome Analysis Tool Kit `GATK`[78] to process BAM files. Lastly, for normal/germline input, we used the `GATK HaplotypeCaller`[79] submodule on the recalibrated BAM to generate a genomic variant call format (GVCF) file. This file is used as the basis for germline calling, described in the **SNV calling for B-allele Frequency (BAF) generation** section.

We obtained references from the Broad Genome References on AWS bucket with a general description of references at https://s3.amazonaws.com/broad-references/broad-references-readme.html.

Quality Control of Sequencing Data

To confirm sample matches and remove mis-matched samples from the dataset, we performed `NGSCheckMate`[80] on matched tumor/normal CRAM files. Briefly, we processed CRAMs using `BCFtools` to filter and call 20k common single nucleotide polymorphisms (SNPs) using default parameters. We used the resulting VCFs to run `NGSCheckMate`. Per `NGSCheckMate` author recommendations, we used <= 0.61 as a correlation coefficient cutoff at sequencing depths > 10 to predict mis-matched samples. We determined RNA-Seq read strandedness by running the `infer_experiment.py` script from `RNA-SeQC`[81] on the first 200k mapped reads. We removed any samples whose calculated strandedness did not match strandedness information provided by the sequencing center. We required that at least 60% of RNA-Seq reads mapped to the human reference for samples to be included in analysis.

Germline Variant Calling

SNP calling for B-allele Frequency (BAF) generation

We performed germline haplotype calls using the `GATK` Joint Genotyping Workflow on individual GVCFs from the normal sample alignment workflow. Using only SNPs, we applied the `GATK` generic hard filter suggestions to the VCF, with an additional requirement of 10 reads minimum

depth per SNP. We used the filtered VCF as input to `Control-FREEC` and `CNVkit` (below) to generate B-allele frequency (BAF) files. This single-sample workflow is available in the D3b GitHub repository. References can be obtained from the Broad Genome References on AWS bucket, and a general description of references can be found at https://s3.amazonaws.com/broad-references/broad-references-readme.html.

Assessment of germline variant pathogenicity

For patients with hypermutant samples, we first added population frequency of germline variants using `ANNOVAR`[82] and pathogenicity scoring from ClinVar[83] using `SnpSift`[84]. We then filtered for variants with read depth >= 15, variant allele fraction >= 0.20, and which were observed at < 0.1% allele frequency across each population in the Genome Aggregation Database (see **Key Resources Table**). Finally, we retained variants in genes included in the KEGG MMR gene set (see **Key Resources Table**), *POLE*, and/or *TP53* which were ClinVar-annotated as pathogenic (P) or likely pathogenic (LP) with review status of >= 2 stars. All P/LP variants were manually reviewed by an interdisciplinary team of scientists, clinicians, and genetic counselors. This workflow is available in the D3b GitHub repository.

Somatic Mutation Calling

SNV and indel calling

For PBTA samples, we used four variant callers to call SNVs and indels from panel, WXS, and WGS data: `Strelka2`[85], `Mutect2`[86], `Lancet`[87], and `VarDictJava`[88]. `VarDictJava`-only calls were not retained since ~ 39M calls with low VAF were uniquely called and may be potential false positives. (~1.2M calls were called by `Mutect2`, `Strelka2`, and `Lancet` and included consensus CNV calling as described below.) We used only `Strelka2`, `Mutect2` and `Lancet` to analyze WXS samples from TCGA. TCGA samples were captured using various WXS target capture kits and we downloaded the BED files from the `GDC portal`. The manufacturers provided the input interval BED files for both panel and WXS data for PBTA samples. We padded all panel and WXS BED files were by 100 bp on each side for `Strelka2`, `Mutect2`, and `VarDictJava` runs and by 400 bp for the `Lancet` run. For WGS calling, we utilized the non-padded BROAD Institute interval calling list `wgs_calling_regions.hg38.interval_list`, comprised of the full genome minus N bases, unless otherwise noted below. We ran `Strelka2`[85] using default parameters for canonical chromosomes (chr1-22, X,Y,M), as recommended by the authors, and we filtered the final `Strelka2` VCF for PASS variants. We ran `Mutect2` from `GATK` according to Broad best practices outlined from their Workflow Description Language (WDL), and we filtered the final `Mutect2` VCF for PASS variants. To manage memory issues, we ran `VarDictJava`[88] using 20 Kb interval chunks of the input BED, padded by 100 bp on each side, such that if an indel occurred in between intervals, it would be captured. Parameters and filtering followed BCBIO standards except that variants with a variant allele frequency (VAF) >= 0.05 (instead of >= 0.10) were retained. The 0.05 VAF increased the true positive rate for indels and decreased the false positive rate for SNVs when using `VarDictJava` in consensus calling. We filtered the final `VarDictJava` VCF for PASS variants with `TYPE=StronglySomatic`. We ran `Lancet` using

default parameters, except for those noted below. For input intervals to `Lancet` WGS, we created a reference BED from only the UTR, exome, and start/stop codon features of the GENCODE 31 reference, augmented as recommended with PASS variant calls from `Strelka2` and `Mutect2`. We then padded these intervals by 300 bp on each side during `Lancet` variant calling. Per recommendations for WGS samples, we augmented the Lancet input intervals described above with PASS variant calls from `Strelka2` and `Mutect2` as validation[89].

VCF annotation and MAF creation

We normalized INDELs with `bcftools norm` on all PASS VCFs using the `kfdrc_annot_vcf_sub_wf.cwl` subworkflow, release v3 (See **Table S5**). The Ensembl Variant Effect Predictor (`VEP`)[90], reference release 93, was used to annotate variants and bcftools was used to add population allele frequency (AF) from gnomAD[91]. We annotated SNV and INDEL hotspots from v2 of Memorial Sloan Kettering Cancer Center's (MSKCC) database (See **Key Resources Table**) as well as the *TERT* promoter mutations C228T and C250T[92]. We annotated SNVs by matching amino acid position (`Protein_position` column in MAF file) with SNVs in the MSKCC database, we matched splice sites to `HGVSp_Short` values in the MSKCC database, and we matched INDELs based on amino acid present within the range of INDEL hotspots values in the MSKCC database. We removed non-hotspot annotated variants with a normal depth less than or equal to 7 and/or gnomAD allele frequency (AF) greater than 0.001 as potential germline variants. We matched *TERT* promoter mutations using hg38 coordinates as indicated in ref.[92]: C228T occurs at 5:1295113 is annotated as existing variant `s1242535815`, `COSM1716563`, or `COSM1716558`, and is 66 bp away from the TSS; C250T occurs at Chr5:1295135, is annotated as existing variant `COSM1716559`, and is 88 bp away from the TSS. We retained variants annotated as `PASS` or `HotSpotAllele=1` in the final set, and we created MAFs using MSKCC's `vcf2maf` tool.

Gather SNV and INDEL Hotspots

We retained all variant calls from `Strelka2`, `Mutect2`, or `Lancet` that overlapped with an SNV or INDEL hotspot in a hotspot-specific MAF file, which we then used for select analyses as described below.

Consensus SNV Calling

Our SNV calling process led to separate sets of predicted mutations for each caller. We considered mutations to describe the same change if they were identical for the following MAF fields: `Chromosome`, `Start_Position`, `Reference_Allele`, `Allele`, and `Tumor_Sample_Barcode`. `Strelka2` does not call multinucleotide variants (MNV), but instead calls each component SNV as a separate mutation, so we separated MNV calls from `Mutect2` and `Lancet` into consecutive SNVs before comparing them to `Strelka2` calls. We examined VAFs produced by each caller and compared their overlap with each other (**Figure S2**). `VarDictJava` calls included many variants that were not identified by other callers (**Figure S2C**), while the other callers produced results that were relatively consistent with one another. Many of these `VarDictJava`-specific calls were variants with low allele frequency

(**Figure S2B**). We therefore derived consensus mutation calls as those shared among the other three callers (`Strelka2`, `Mutect2`, and `Lancet`), and we did not further consider `VarDictJava` calls due to concerns it called a large number of false positives. This decision had minimal impact on results because `VarDictJava` also identified nearly every mutation that the other three callers identified, in addition to many unique mutations.

Somatic Copy Number Variant Calling (WGS samples only)

We used `Control-FREEC`[93,94] and `CNVkit`[95] for copy number variant calls. For both algorithms, the `germline_sex_estimate` (described below) was used as input for sample sex and germline variant calls (above) were used as input for BAF estimation. `Control-FREEC` was run on human genome reference hg38 using the optional parameters of a 0.05 coefficient of variation, ploidy choice of 2-4, and BAF adjustment for tumor-normal pairs. `Theta2`[96] used `VarDictJava` germline and somatic calls, filtered on PASS and strongly somatic, to infer tumor purity. `Theta2` purity was added as an optional parameter to `CNVkit` to adjust copy number calls. `CNVkit` was run on human genome reference hg38 using the optional parameters of Theta2 purity and BAF adjustment for tumor-normal pairs. We used `GISTIC`[97] on the `CNVkit` and the consensus CNV segmentation files to generate gene-level copy number abundance (Log R Ratio) as well as chromosomal arm copy number alterations using the parameters specified in the (`run-gistic` analysis module in the OpenPBTA Analysis repository).

Consensus CNV Calling

For each caller and sample, we called CNVs based on consensus among `Control-FREEC`[93,94], `CNVkit`[95], and `Manta`[98]. We specifically included CNVs called significant by `Control-FREEC` (p-value < 0.01) and `Manta` calls that passed all filters in consensus calling. We removed sample and consensus caller files with more than 2,500 CNVs because we expected these to be noisy and derive poor quality samples based on cutoffs used in `GISTIC`[97]. For each sample, we included the regions in the final consensus set: 1) regions with reciprocal overlap of 50% or more between at least two of the callers; 2) smaller CNV regions in which more than 90% of regions are covered by another caller. We did not include any copy number alteration called by a single algorithm in the consensus file. We defined copy number as `NA` for any regions that had a neutral call for the samples included in the consensus file. We merged CNV regions within 10,000 bp of each other with the same direction of gain or loss into single region. We filtered out any CNVs that overlapped 50% or more with immunoglobulin, telomeric, centromeric, segment duplicated regions, or that were shorter than 3000 bp.

Somatic Structural Variant Calling (WGS samples only)

We used `Manta`[98] for structural variant (SV) calls, and we limited to regions used in `Strelka2`. The hg38 reference for SV calling used was limited to canonical chromosome regions. We used `AnnotSV`[99] to annotate `Manta` output. All associated workflows are available in the workflows GitHub repository.

Gene Expression

Abundance Estimation

We used `STAR`[100] to align paired-end RNA-seq reads, and we used the associated alignment for all subsequent RNA analysis. We used Ensembl GENCODE 27 "Comprehensive gene annotation" (see **Key Resources Table**) as a reference. We used `RSEM`[101] for both FPKM and TPM transcript- and gene-level quantification.

Gene Expression Matrices with Unique HUGO Symbols

To enable downstream analyses, we next identified gene symbols that map to multiple Ensembl gene identifiers (in GENCODE v27, 212 gene symbols map to 1866 Ensembl gene identifiers), known as multi-mapped gene symbols, and ensured unique mappings (`collapse-rnaseq` analysis module in the OpenPBTA Analysis repository). To this end, we first removed genes with no expression from the `RSEM` abundance data by requiring an FPKM > 0 in at least 1 sample across the PBTA cohort. We computed the mean FPKM across all samples per gene. For each multi-mapped gene symbol, we chose the Ensembl identifier corresponding to the maximum mean FPKM, using the assumption that the gene identifier with the highest expression best represented the expression of the gene. After collapsing gene identifiers, 46,400 uniquely-expressed genes remained in the poly-A dataset, and 53,011 uniquely-expressed genes remained in the stranded dataset.

Gene fusion detection

We set up `Arriba`[102] and `STAR-Fusion`[103] fusion detection tools using CWL on CAVATICA. For both of these tools, we used aligned BAM and chimeric SAM files from `STAR` as inputs and `GRCh38_gencode_v27` GTF for gene annotation. We ran `STAR-Fusion` with default parameters and annotated all fusion calls with the `GRCh38_v27_CTAT_lib_Feb092018.plug-n-play.tar.gz` file from the `STAR-Fusion` release. For `Arriba`, we used a blacklist file `blacklist_hg38_GRCh38_2018-11-04.tsv.gz` from the `Arriba` release to remove recurrent fusion artifacts and transcripts present in healthy tissue. We provided `Arriba` with strandedness information for stranded samples, or we set it to auto-detection for poly-A samples. We used `FusionAnnotator` on `Arriba` fusion calls to harmonize annotations with those of `STAR-Fusion`. The RNA expression and fusion workflows can be found in the D3b GitHub repository. The `FusionAnnotator` workflow we used for this analysis can be found in the D3b GitHub repository.

QUANTIFICATION AND STATISTICAL ANALYSIS

Recurrently mutated genes and co-occurrence of gene mutations (`interaction-plots` analysis module)

Using the consensus SNV calls, we identified genes that were recurrently mutated in the OpenPBTA cohort, including nonsynonymous mutations with a VAF > 5% among the set of independent samples. We used `VEP`[90] annotations, including “High” and “Moderate” consequence types as defined in the R package `Maftools`[104], to determine the set of nonsynonymous mutations. For each gene, we then tallied the number of samples that had at least one nonsynonymous mutation.

For genes that contained nonsynonymous mutations in multiple samples, we calculated pairwise mutation co-occurrence scores. This score was defined as $I\left(-\log_{10}(P)\right)$ where I is 1 when the odds ratio is > 1 (indicating co-occurrence), and -1 when the odds ratio is < 1 (indicating mutual exclusivity), with P defined by Fisher’s Exact Test.

Focal Copy Number Calling (`focal-cn-file-preparation` analysis module)

We added the ploidy inferred via `Control-FREEC` to the consensus CNV segmentation file and used the ploidy and copy number values to define gain and loss values broadly at the chromosome level. We used `bedtools coverage`[105] to add cytoband status using the UCSC cytoband file[106] (See **Key Resources Table**). The output status call fractions, which are values of the loss, gain, and callable fractions of each cytoband region, were used to define dominant status at the cytoband-level. We calculated the weighted means of each status call fraction using band length. We used the weighted means to define the dominant status at the chromosome arm-level.

A status was considered dominant if more than half of the region was callable and the status call fraction was greater than 0.9 for that region. We adopted this 0.9 threshold to ensure that the dominant status fraction call was greater than the remaining status fraction calls in a region.

We aimed to define focal copy number units to avoid calling adjacent genes in the same cytoband or arm as copy number losses or gains where it would be more appropriate to call the broader region a loss or gain. To determine the most focal units, we first considered the dominant status calls at the chromosome arm-level. If the chromosome arm dominant status was callable but not clearly defined as a gain or loss, we instead included the cytoband-level status call. Similarly, if a cytoband dominant status call was callable but not clearly defined as a gain or loss, we instead included gene-level status call. To obtain the gene-level data, we used the `IRanges` package in R[107] to find overlaps between the segments in the consensus CNV file and the exons in the GENCODE v27 annotation file (See **Key Resources Table**) . If the copy number value was 0, we set the status to “deep deletion”. For autosomes only, we set the status to “amplification” when the copy number value was greater than two times the ploidy value. We plotted genome-wide gains and losses in (**Figure S3B**) using the R package `ComplexHeatmap`[108].

Breakpoint Density (WGS samples only; `chromosomal-instability` analysis module)

We defined breakpoint density as the number of breaks per genome or exome per sample. For Manta SV calls, we filtered to retain "PASS" variants and used breakpoints from the algorithm. For consensus CNV calls, if |log2 ratio| > log2(1), we annotated the segment as a break. We then calculated breakpoint density as:

$$\text{breakpoint density} = \frac{\text{N breaks}}{\text{Size in Mb of effectively surveyed genome}}$$

Chromothripsis Analysis (WGS samples only; `chromothripsis` analysis module)

Considering only chromosomes 1-22 and X, we identified candidate chromothripsis regions in the set of independent tumor WGS samples with ShatterSeek[109], using Manta SV calls that passed all filters and consensus CNV calls. We modified the consensus CNV data to fit `ShatterSeek` input requirements as follows: we set CNV-neutral or excluded regions as the respective sample's ploidy value from `Control-FREEC`, and we then merged consecutive segments with the same copy number value. We classified candidate chromothripsis regions as high- or low-confidence using the statistical criteria described by the `ShatterSeek` authors.

Immune Profiling and Deconvolution (`immune-deconv` analysis module)

We used the R package `immunedeconv`[110] with the method `quanTIseq`[111] to deconvolute various immune cell types across tumors from the PBTA cohort in the stranded and poly-A collapsed FPKM RNA-seq datasets (`immune-deconv` analysis module). The `quanTIseq` deconvolution method directly estimates absolute fractions of 10 immune cell types that represent inferred proportions of the cell types in the mixture. Therefore, we utilized `quanTIseq` for inter-sample, intra-sample, and inter-histology score comparisons.

Gene Set Variation Analysis (`gene-set-enrichment-analysis` analysis module)

We performed Gene Set Variation Analysis (GSVA) on collapsed, log2-transformed RSEM FPKM data using the `GSVA` Bioconductor package[112]. We specified the parameter `mx.diff=TRUE` to obtain Gaussian-distributed scores for each of the MSigDB hallmark gene sets[113]. We compared GSVA scores among histology groups using ANOVA and subsequent Tukey tests; p-values were Bonferroni-corrected for multiple hypothesis testing. We plotted scores by cancer group using the `ComplexHeatmap` R package (**Figure 5B**)[108].

Transcriptomic Dimension Reduction (`transcriptomic-dimension-reduction` analysis module)

We applied Uniform Manifold Approximation and Projection (UMAP)[114] to log2-transformed FPKM data using the `umap` R package (See **Key Resources Table**). We set the number of neighbors to 15.

Fusion prioritization (`fusion_filtering` analysis module)

We performed artifact filtering and additional annotation on fusion calls to prioritize putative oncogenic fusions. Briefly, we considered all in-frame and frameshift fusion calls with at least one junction read and at least one gene partner expressed (TPM > 1) to be true calls. If a fusion call had a large number of spanning fragment reads compared to junction reads (spanning fragment minus junction read greater than ten), we removed these calls as potential false positives. We prioritized a union of fusion calls as true calls if the fused genes were detected by both callers, the same fusion was recurrent within a broad histology grouping (> 2 samples), or the fusion was specific to the given broad histology. If either 5' or 3' genes fused to more than five different genes within a sample, we removed these calls as potential false positives. We annotated putative driver fusions and prioritized fusions based on partners containing known kinases, oncogenes, tumor suppressors, curated transcription factors[115], COSMIC genes, and/or known TCGA fusions from curated references. Based on pediatric cancer literature review, we added *MYBL1*[116], *SNCAIP*[117], *FOXR2*[118], *TTYH1*[119], and *TERT*[120–123] to the oncogene list, and we added *BCOR*[118] and *QKI*[124] to the tumor suppressor gene list.

Oncoprint figure generation (`oncoprint-landscape` analysis module)

We used `Maftools`[104] to generate oncoprints depicting the frequencies of canonical somatic gene mutations, CNVs, and fusions for the top 20 genes mutated across primary tumors within broad histologies of the OpenPBTA dataset. We collated canonical genes from the literature for low-grade astrocytic tumors[25], embryonal tumors[26,28,29,125,126], diffuse astrocytic and oligodendroglial tumors[15,22,30,31], and other tumors: ependymal tumors, craniopharyngiomas, neuronal-glial mixed tumors, histiocytic tumors, chordoma, meningioma, and choroid plexus tumors[127–136].

Mutational Signatures (`mutational-signatures` analysis module)

We obtained weights (i.e., exposures) for signature sets using the `deconstructSigs` R package function `whichSignatures()`[137] from consensus SNVs with the BSgenome.Hsapiens.UCSC.hg38 annotations (see **Key Resources Table**). Specifically, we estimated signature weights across samples for eight signatures previously identified in the Signal reference set of signatures ("RefSig") as associated with adult central nervous system (CNS) tumors[36]. These eight RefSig signatures are 1, 3, 8, 11, 18, 19, N6, and MMR2. Weights for signatures fall in the range zero to one inclusive. `deconstructSigs` estimates the weights for each signature across samples and allows for a proportion of unassigned weights referred to as "Other" in the text. These results do not include signatures with small contributions; `deconstructSigs` drops signature weights that are less than 6%[137]. We plotted mutational

signatures for patients with hypermutant tumors (**Figure 4E**) using the R package `ComplexHeatmap`[108].

Tumor Mutation Burden (`snv-callers` analysis module)

We consider tumor mutation burden (TMB) to be the number of consensus SNVs per effectively surveyed base of the genome. We considered base pairs to be effectively surveyed if they were in the intersection of the genomic ranges considered by the callers used to generate the consensus and where appropriate, regions of interest, such as coding sequences. We calculated TMB as:

$$\text{TMB} = \frac{\text{\# of coding sequence SNVs}}{\text{Size in Mb of effectively surveyed genome}}$$

We used the total number coding sequence consensus SNVs for the numerator and the size of the intersection of the regions considered by `Strelka2` and `Mutect2` with coding regions (CDS from GENCODE v27 annotation, see **Key Resources Table**) as the denominator.

Clinical Data Harmonization

WHO Classification of Disease Types

Table S1 contains a README, along with sample technical, clinical, and additional metadata used for this study.

Molecular Subtyping

We performed molecular subtyping on tumors in the OpenPBTA to the extent possible. The `molecular_subtype` field in `pbta-histologies.tsv` contains molecular subtypes for tumor types selected from `pathology_diagnosis` and `pathology_free_text_diagnosis` fields as described below, following World Health Organization 2016 classification criteria[21].

Medulloblastoma (MB) subtypes SHH, WNT, Group 3, and Group 4 were predicted using the consensus of two RNA expression classifiers: `MedulloClassifier`[61] and `MM2S`[60] on the RSEM FPKM data (`molecular-subtyping-MB` analysis module).

High-grade glioma (HGG) subtypes were derived (`molecular-subtyping-HGG` analysis module) using the following criteria:

1. If any sample contained an *H3F3A* p.K28M, *HIST1H3B* p.K28M, *HIST1H3C* p.K28M, or *HIST2H3C* p.K28M mutation and no *BRAF* p.V600E mutation, it was subtyped as `DMG, H3 K28`.
2. If any sample contained an *HIST1H3B* p.K28M, *HIST1H3C* p.K28M, or *HIST2H3C* p.K28M mutation and a *BRAF* p.V600E mutation, it was subtyped as `DMG, H3 K28, BRAF V600E`.

3. If any sample contained an *H3F3A* p.G35V or p.G35R mutation, it was subtyped as `HGG, H3 G35`.
4. If any high-grade glioma sample contained an *IDH1* p.R132 mutation, it was subtyped as `HGG, IDH`.
5. If a sample was initially classified as HGG, had no defining histone mutations, and a BRAF p.V600E mutation, it was subtyped as `BRAF V600E`.
6. All other high-grade glioma samples that did not meet any of these criteria were subtyped as `HGG, H3 wildtype`.

Embryonal tumors were included in non-MB and non-ATRT embryonal tumor subtyping (`molecular-subtyping-embryonal` analysis module) if they met any of the following criteria:

1. A *TTYH1* (5' partner) fusion was detected.
2. A *MN1* (5' partner) fusion was detected, with the exception of `MN1::PATZ1` since it is an entity separate of CNS HGNET-MN1 tumors[138].
3. Pathology diagnoses included "Supratentorial or Spinal Cord PNET" or "Embryonal Tumor with Multilayered Rosettes".
4. A pathology diagnosis of "Neuroblastoma", where the tumor was not indicated to be peripheral or metastatic and was located in the CNS.
5. Any sample with "embryonal tumor with multilayer rosettes, ros (who grade iv)", "embryonal tumor, nos, congenital type", "ependymoblastoma" or "medulloepithelioma" in pathology free text.

Non-MB and non-ATRT embryonal tumors identified with the above criteria were further subtyped (`molecular-subtyping-embryonal` analysis module) using the criteria below[139–142].

1. Any RNA-seq biospecimen with *LIN28A* overexpression, plus a *TYH1* fusion (5' partner) with a gene adjacent or within the C19MC miRNA cluster and/or copy number amplification of the C19MC region was subtyped as `ETMR, C19MC-altered` (Embryonal tumor with multilayer rosettes, chromosome 19 miRNA cluster altered)[119,143].
2. Any RNA-seq biospecimen with *LIN28A* overexpression, a *TTYH1* fusion (5' partner) with a gene adjacent or within the C19MC miRNA cluster but no evidence of copy number amplification of the C19MC region was subtyped as `ETMR, NOS` (Embryonal tumor with multilayer rosettes, not otherwise specified)[119,143].
3. Any RNA-seq biospecimen with a fusion having a 5' *MN1* and 3' *BEND2* or *CXXC5* partner were subtyped as `CNS HGNET-MN1` [Central nervous system (CNS) high-grade neuroepithelial tumor with *MN1* alteration].
4. Non-MB and non-ATRT embryonal tumors with internal tandem duplication (as defined in[144]) of *BCOR* were subtyped as `CNS HGNET-BCOR` (CNS high-grade neuroepithelial tumor with *BCOR* alteration).

5. Non-MB and non-ATRT embryonal tumors with over-expression and/or gene fusions in *FOXR2* were subtyped as `CNS NB-FOXR2` (CNS neuroblastoma with *FOXR2* activation).
6. Non-MB and non-ATRT embryonal tumors with *CIC::NUTM1* or other *CIC* fusions, were subtyped as `CNS EFT-CIC` (CNS Ewing sarcoma family tumor with *CIC* alteration)[118]
7. Non-MB and non-ATRT embryonal tumors that did not fit any of the above categories were subtyped as `CNS Embryonal, NOS` (CNS Embryonal tumor, not otherwise specified).

Neurocytoma subtypes central neurocytoma (CNC) and extraventricular neurocytoma (EVN) were assigned (`molecular-subtyping-neurocytoma` analysis module) based on the primary site of the tumor[145]. If the tumor's primary site was "ventricles," we assigned the subtype as CNC; otherwise, we assigned the subtype as EVN.

Craniopharyngiomas (CRANIO) were subtyped (`molecular-subtyping-CRANIO` analysis module) into adamantinomatous (`CRANIO, ADAM`), papillary (`CRANIO, PAP`) or undetermined (`CRANIO, To be classified`) based on the following criteria[146,147]:

1. Craniopharyngiomas from patients over 40 years old with a *BRAF* p.V600E mutation were subtyped as `CRANIO, PAP`.
2. Craniopharyngiomas from patients younger than 40 years old with mutations in exon 3 of *CTNNB1* were subtyped as `CRANIO, ADAM`.
3. Craniopharyngiomas that did not fall into the above two categories were subtyped as `CRANIO, To be classified`.

A molecular subtype of `EWS` was assigned to any tumor with a *EWSR1* fusion or with a `pathology_diagnosis` of `Ewings Sarcoma` (`molecular-subtyping-EWS` analysis module).

Low-grade gliomas (LGG) or glialneuronal tumors (GNT) were subtyped (`molecular-subtyping-LGAT` analysis module). based on SNV, fusion and CNV status based on[23], and as described below.

1. If a sample contained a *NF1* somatic mutation, either nonsense or missense, it was subtyped as `LGG, NF1-somatic`.
2. If a sample contained *NF1* germline mutation, as indicated by a patient having the neurofibromatosis cancer predisposition, it was subtyped as `LGG, NF1-germline`.
3. If a sample contained the *IDH* p.R132 mutation, it was subtyped as `LGG, IDH`.
4. If a sample contained a histone p.K28M mutation in either *H3F3A*, *H3F3B*, *HIST1H3B*, *HIST1H3C*, or *HIST2H3C*, or if it contained a p.G35R or p.G35V mutation in *H3F3A*, it was subtyped as `LGG, H3`.
5. If a sample contained *BRAF* p.V600E or any other non-canonical *BRAF* mutations in the kinase (PK_Tyr_Ser-Thr) domain PF07714 (see **Key Resources Table**), it was subtyped as `LGG, BRAF V600E`.

6. If a sample contained `KIAA1549::BRAF` fusion, it was subtyped as `LGG, KIAA1549::BRAF`.
7. If a sample contained SNV or indel in either *KRAS*, *NRAS*, *HRAS*, *MAP2K1*, *MAP2K2*, *MAP2K1*, *ARAF*, *RAF1*, or non-kinase domain of *BRAF*, or if it contained *RAF1* fusion, or *BRAF* fusion that was not `KIAA1549::BRAF`, it was subtyped as `LGG, other MAPK`.
8. If a sample contained SNV in either *MET*, *KIT* or *PDGFRA*, or if it contained fusion in *ALK*, *ROS1*, *NTRK1*, *NTRK2*, *NTRK3* or *PDGFRA*, it was subtyped as `LGG, RTK`.
9. If a sample contained *FGFR1* p.N546K, p.K656E, p.N577, or p. K687 hotspot mutations, or tyrosine kinase domain tandem duplication (See **Key Resources Table**), or *FGFR1* or *FGFR2* fusions, it was subtyped as `LGG, FGFR`.
10. If a sample contained *MYB* or *MYBL1* fusion, it was subtyped as `LGG, MYB/MYBL1`.
11. If a sample contained focal CDKN2A and/or CDKN2B deletion, it was subtyped as `LGG, CDKN2A/B`.

For LGG tumors that did not have any of the above molecular alterations, if both RNA and DNA samples were available, it was subtyped as `LGG, wildtype`. Otherwise, if either RNA or DNA sample was unavailable, it was subtyped as `LGG, To be classified`.

If pathology diagnosis was `Subependymal Giant Cell Astrocytoma (SEGA)`, the `LGG` portion of molecular subtype was recoded to `SEGA`.

Lastly, for all LGG- and GNT- subtyped samples, if the tumors were glialneuronal in origin, based on `pathology_free_text_diagnosis` entries of `desmoplastic infantile`,`desmoplastic infantile ganglioglioma`, `desmoplastic infantile astrocytoma` or `glioneuronal`, each was recoded as follows: If pathology diagnosis is `Low-grade glioma/astrocytoma (WHO grade I/II)` or `Ganglioglioma`, the `LGG` portion of the molecular subtype was recoded to `GNT`.

Ependymomas (EPN) were subtyped (`molecular-subtyping-EPN` analysis module) into `EPN, ST RELA`, `EPN, ST YAP1`, `EPN, PF A` and `EPN, PF B` based on evidence for these molecular subgroups as described in Pajtler et al.[128]. Briefly, fusion, CNV and gene expression data were used to subtype EPN as follows:

1. Any tumor with fusions containing `RELA` as fusion partner, e.g., `C11orf95::RELA`, `LTBP3::RELA`, was subtyped as `EPN, ST RELA`.
2. Any tumor with fusions containing `YAP1` as fusion partner, such as `C11orf95::YAP1`, `YAP1::MAMLD1` and `YAP1::FAM118B`, was subtyped as `EPN, ST YAP1`.
3. Any tumor with the following molecular characterization would be subtyped as `EPN, PF A`:

- *CXorf67* expression z-score of over 3
- *TKTL1* expression z-score of over 3 and 1q gain

4. Any tumor with the following molecular characterization would be subtyped as `EPN, PF B`:

- *GPBP17* expression z-score of over 3 and loss of 6q or 6p
- *IFT46* expression z-score of over 3 and loss of 6q or 6p

Any tumor with the above molecular characteristics would be exclusively subtyped to the designated group.

For all other remaining EPN tumors without above molecular characteristics, they would be subtyped to `EPN, ST RELA` and `EPN, ST YAP1` in a non-exclusive way (e.g., a tumor could have both `EPN, ST RELA` and `EPN, ST YAP1` subtypes) if any of the following alterations were present.

1. Any tumor with the following alterations was assigned `EPN, ST RELA`:
- `PTEN::TAS2R1` fusion
- chromosome 9 arm (9p or 9q) loss
- *RELA* expression z-score of over 3
- *L1CAM* expression z-score of over 3
2. Any tumor with the following alterations was assigned `EPN, ST YAP1`:
- `C11orf95::MAML2` fusion
- chromosome 11 short arm (11p) loss
- chromosome 11 long arm (11q) gain
- *ARL4D* expression z-score of over 3
- *CLDN1* expression z-score of over 3

After all relevant tumor samples were subtyped by the above molecular subtyping modules, the results from these modules, along with other clinical information (such as pathology diagnosis free text), were compiled in the `molecular-subtyping-pathology` module and integrated into the OpenPBTA data in the `molecular-subtyping-integrate` module.

TP53 Alteration Annotation (`tp53_nf1_score` analysis module)

We annotated *TP53* altered HGG samples as either `TP53 lost` or `TP53 activated` and integrated this within the molecular subtype. To this end, we applied a *TP53* inactivation classifier originally trained on TCGA pan-cancer data[38] to the matched RNA expression data for each sample. Along with the *TP53* classifier scores, we collectively used consensus SNV and CNV, SV, and reference databases that list *TP53* hotspot mutations[148,149] and functional domains[150] to determine *TP53* alteration status for each sample. We adopted the following rules for calling either `TP53 lost` or `TP53 activated`:

1. If a sample had either of the two well-characterized *TP53* gain-of-function mutations, p.R273C or p.R248W[39], we assigned `TP53 activated` status.

2. Samples were annotated as `TP53 lost` if they contained i) a *TP53* hotspot mutation as defined by IARC *TP53* database or the MSKCC cancer hotspots database[148,149] (see also, **Key Resources Table**), ii) two *TP53* alterations, including SNV, CNV or SV, indicative of probable bi-allelic alterations; iii) one *TP53* somatic alteration, including

SNV, CNV, or SV or a germline *TP53* mutation indicated by the diagnosis of Li-Fraumeni syndrome (LFS)[151], or iv) one germline *TP53* mutation indicated by LFS and the *TP53* classifier score for matched RNA-Seq was greater than 0.5.

Prediction of participants' genetic sex

Participant metadata included a reported gender. We used WGS germline data, in concert with the reported gender, to predict participant genetic sex so that we could identify sexually dimorphic outcomes. This analysis may also indicate samples that may have been contaminated. We used the `idxstats` utility from `SAMtools`[152] to calculate read lengths, the number of mapped reads, and the corresponding chromosomal location for reads to the X and Y chromosomes. We used the fraction of total normalized X and Y chromosome reads that were attributed to the Y chromosome as a summary statistic. We manually reviewed this statistic in the context of reported gender and determined that a threshold of less than 0.2 clearly delineated female samples. We marked fractions greater than 0.4 as predicted males, and we marked samples with values in the inclusive range 0.2-0.4 as unknown. We performed this analysis through CWL on CAVATICA. We added resulting calls to the histologies file under the column header `germline_sex_estimate`.

Selection of independent samples (`independent-samples` analysis module)

Certain analyses required that we select only a single representative specimen for each individual. In these cases, we identified a single specimen by prioritizing primary tumors and those with whole-genome sequencing available. If this filtering still resulted in multiple specimens, we randomly selected a single specimen from the remaining set.

Quantification of Telomerase Activity using Gene Expression Data (`telomerase-activity-prediction` analysis module)

We predicted telomerase activity of tumor samples using the recently developed `EXTEND` method[41]. Briefly, `EXTEND` estimates telomerase activity based on the expression of a 13-gene signature. We derived this signature by comparing telomerase-positive tumors and tumors with activated alternative lengthening of telomeres pathway, a group presumably negative of telomerase activity.

Survival models (`survival-analysis` analysis module)

We calculated overall survival (OS) as days since initial diagnosis and performed several survival analyses on the OpenPBTA cohort using the `survival` R package. We performed survival analysis for patients by HGG subtype using the Kaplan-Meier estimator[153] and a log-rank test (Mantel-Cox test)[154] on the different HGG subtypes. Next, we used multivariate Cox (proportional hazards) regression analysis[155] to model the following: a) `tp53 scores + telomerase scores + extent of tumor resection + LGG group + HGG group`, in which `tp53 scores` and `telomerase scores` are numeric, `extent of tumor resection` is categorical, and `LGG group` and `HGG group` are binary variables indicating

whether the sample is in either broad histology grouping, b) `tp53 scores + telomerase scores + extent of tumor resection` for each `cancer_group` with an N>=3 deceased patients (DIPG, DMG, HGG, MB, and EPN), and c) `quantiseq cell type fractions + CD274 expression + extent of tumor resection` for each `cancer_group` with an N>=3 deceased patients (DIPG, DMG, HGG, MB, and EPN), in which `quantiseq cell type fractions` and `CD274 expression` are numeric.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
Recover Cell Culture Freezing media	Gibco	12648010
Hank's Balanced Salt Solution (HBSS)	Gibco	14175095
Papain	SciQuest	LS003124
Ovomucoid	SciQuest	542000
DNase	Roche	10104159001
100µm cell strainer	Greiner Bio-One	542000
DMEM/F-12 medium	Sigma	D8062
Fetal Bovine Serum (FBS)	Hyclone	SH30910.03
GlutaMAX	Gibco	35050061
Penicillin/Streptomycin -Amphotericin B	Lonza	17-745E
Normocin	Invivogen	ant-nr-2
B-27 supplement minus vitamin A	Gibco	12587-010
N-2 supplement	Gibco	17502001
Epidermal growth factor	Gibco	PHG0311L
Basic fibroblast growth factor	PeproTech	100-18B
Heparin	Sigma	H3149
DNA/RNA AllPrep Kit	Qiagen	80204
TruSeq RNA Sample Prep Kit	Illumina	FC-122-1001
KAPA Library Preparation Kit	Roche	KK8201

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
AllPrep DNA/RNA/miRNA Universal kit	Qiagen	80224
RNase A	Qiagen	19101
QIAsymphony DSP DNA Midi Kit	Qiagen	937255
KAPA HyperPrep kit	Roche	08098107702
RiboErase kit	Roche	07962304001
Raw and harmonized WGS, WXS, Panel, RNA-Seq	KidsFirst Data Resource Center, this project	[73]
Merged summary files	this project	https://cavatica.sbgenomics.com/u/cavatica/openpbta
Merged summary files and downstream analyses	this project	https://github.com/AlexsLemonade/OpenPBTA-analysis/
Processed data	this project	https://pedcbioportal.kidsfirstdrc.org/study/summary?id=openpbta
Experimental models: Cell lines		
CBTN pediatric brain tumor-derived cell lines	[14]	See **Table S1** for identifiers
Software and algorithms		
Data processing and analysis software	Multiple	See **Table S5** for identifiers
OpenPBTA workflows repository	this project	[156]
OpenPBTA analysis repository	this project	[157]
OpenPBTA manuscript repository	this project	
Other		
TCGA WXS dataset	National Institutes of Health The	dbGAP phs000178.v11.p8

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
	Cancer Genome Atlas (TCGA)	
Cancer hotspots	MSKCC	https://www.cancerhotspots.org/#/download (v2)
Reference genomes	Broad	https://s3.console.aws.amazon.com/s3/buckets/broad-references/hg38/v0/
Reference genome hg38, patch release 12	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/
Human Cytoband file	UCSC	http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/cytoBand.txt.gz
CDS from GENCODE v27 annotation	GENCODE	https://www.gencodegenes.org/human/release_27.html
PFAM domains and locations	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/pfamDesc.txt.gz; https://pfam.xfam.org/family/PF07714
BSgenome.Hsapiens.UCSC.hg38 annotations	Bioconductor	https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.UCSC.hg38.html
gnomAD v2.1.1 (exome and genome)	Genome Aggregation Database	https://gnomad.broadinstitute.org/downloads#v2-liftover-variants
KEGG MMR gene set v7.5.1	BROAD Institute	https://www.gsea-msigdb.org/gsea/msigdb/download_geneset.jsp?geneSetName=KEGG_MISMATCH_REPAIR
ClinVar Database (2022-05-07)	NCBI	https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2022/clinvar_20220507.vcf.gz

Supplemental Information Titles and Legends

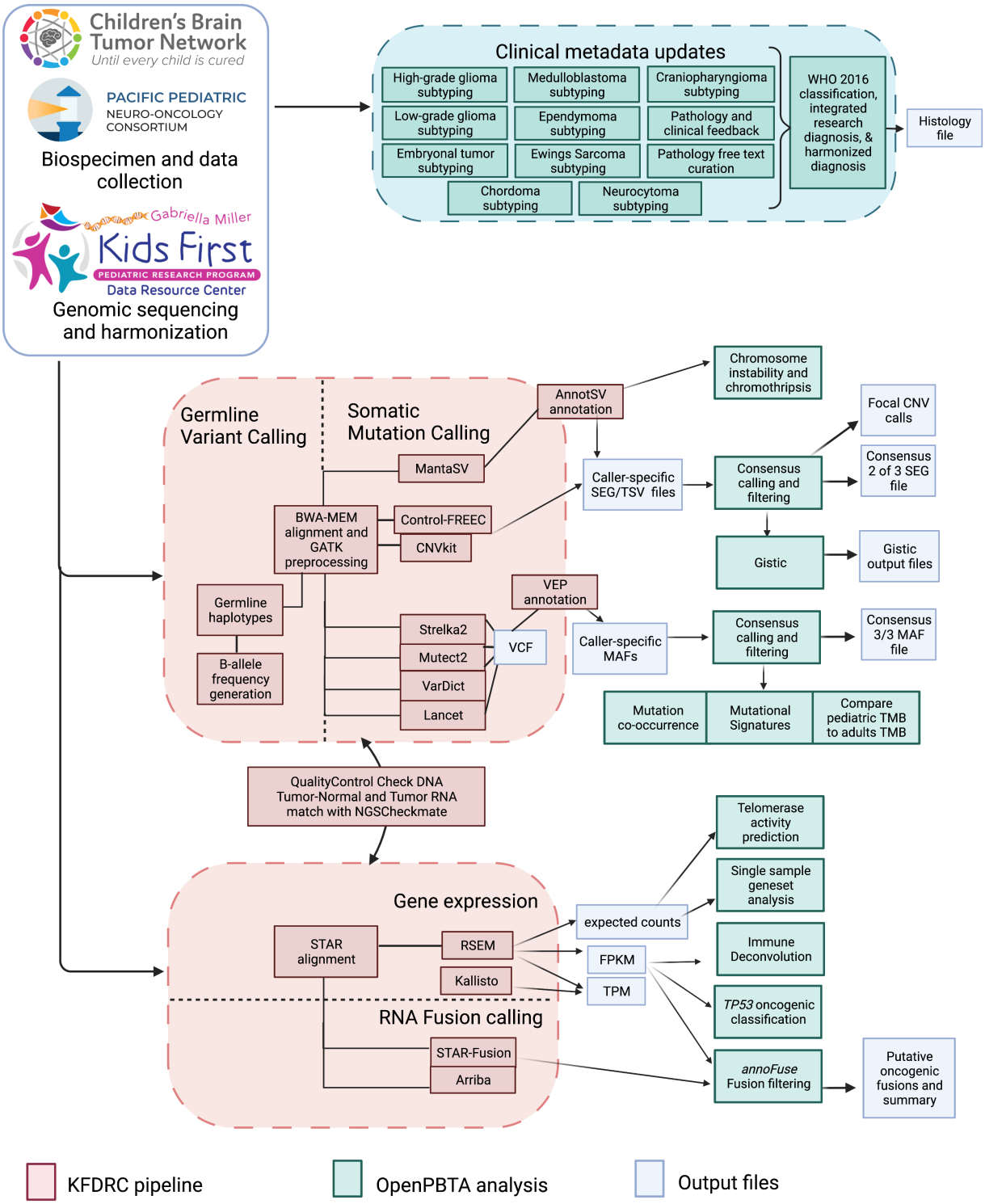


Figure S1: ***OpenPBTA Project Workflow, Related to Figure 1.*** *Biospecimens and data were collected by CBTN and PNOC. Genomic sequencing and harmonization (orange boxes) were*

performed by the Kids First Data Resource Center (KFDRC). Analyses in the green boxes were performed by contributors of the OpenPBTA project. Output files are denoted in blue. Figure created with BioRender.com.

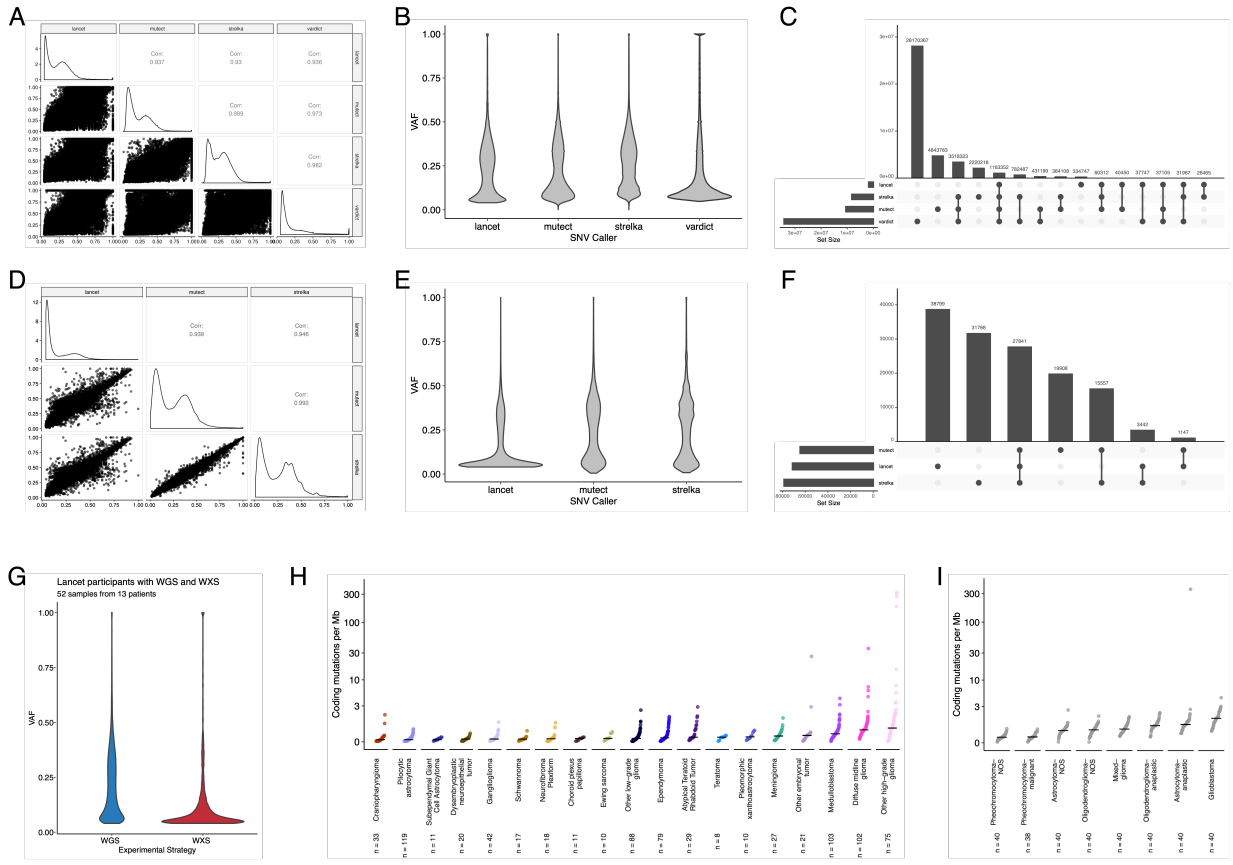


*Figure S2: **Validation of Consensus SNV calls and Tumor Mutation Burden, Related to Figures 2 and 3.** Correlation (A) and violin (B) plots of mutation variant allele frequencies (VAFs) comparing the variant callers (Lancet, Strelka2, Mutect2, and VarDict) used for PBTA samples. Upset plot (C) showing overlap of variant calls. Correlation (D) and violin (E) plots of mutation variant allele frequencies (VAFs) comparing the variant callers (Lancet, Strelka2, and Mutect2) used for TCGA samples. Upset plot (F) showing overlap of variant calls. Violin plots (G) showing VAFs for Lancet calls performed on WGS and WXS from the same tumor (N = 52 samples from 13 patients). Cumulative distribution TMB plots for PBTA (H) and TCGA (I) tumors using consensus SNV calls.*

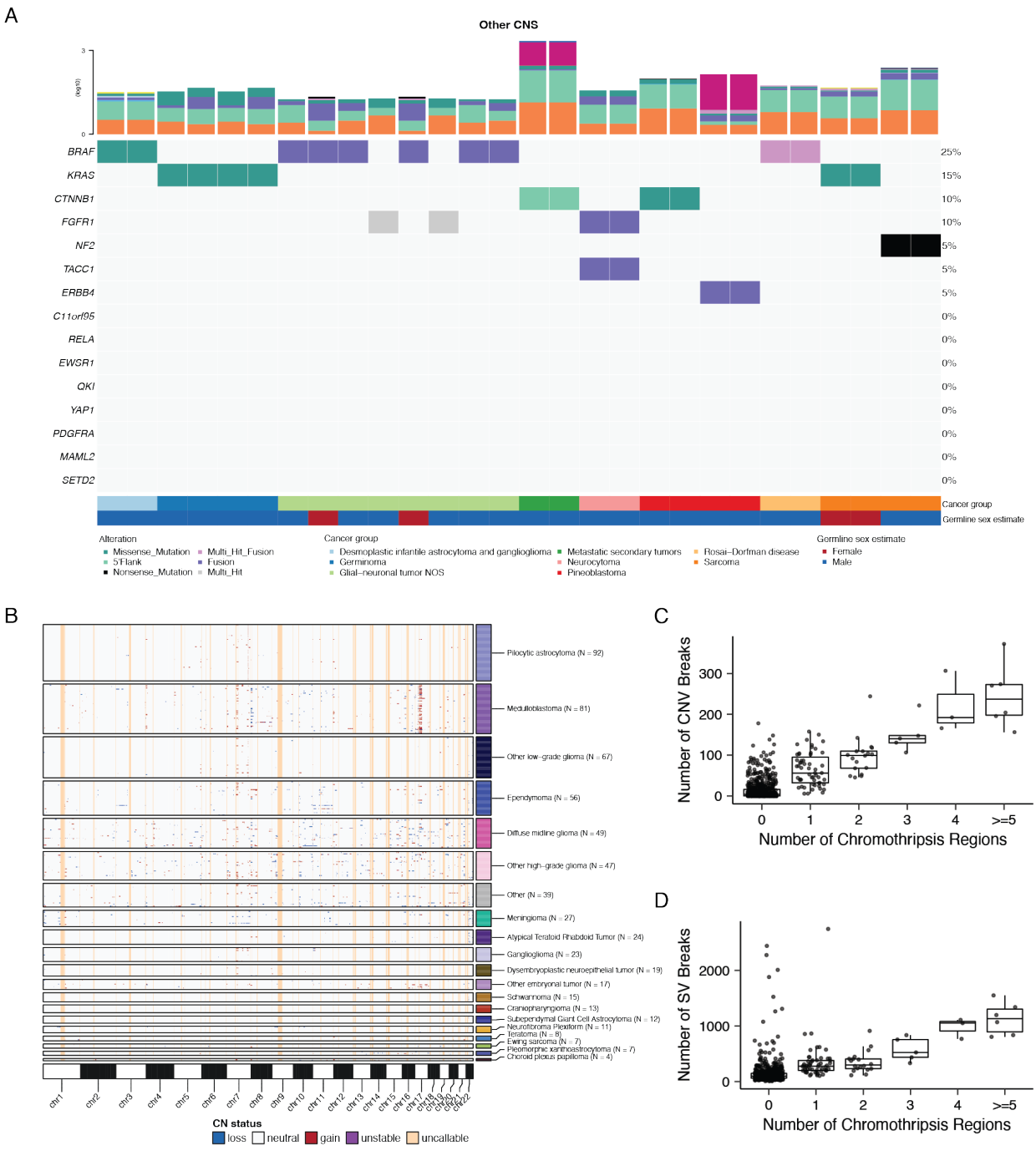
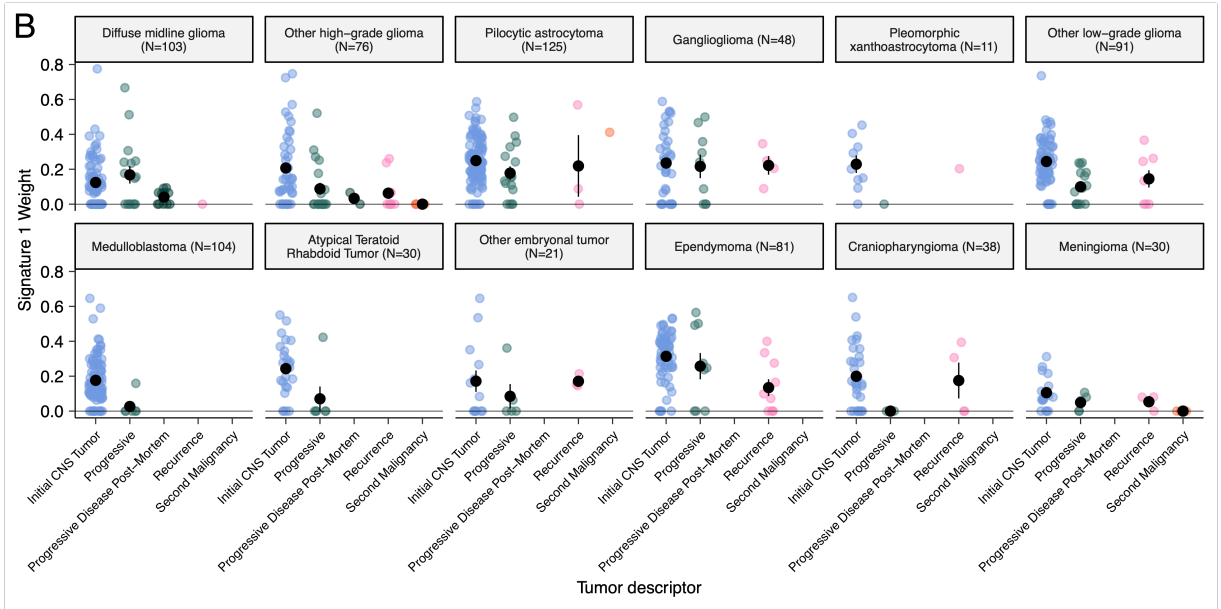


Figure S3: ***Genomic instability of pediatric brain tumors, Related to Figures 2 and 3.*** *(A) Oncoprint of canonical somatic gene mutations, CNVs, fusions, and TMB (top bar plot) for the top 20 genes mutated across rare CNS tumors: desmoplastic infantile astrocytoma and ganglioglioma (N = 2), germinoma (N = 4), glial-neuronal NOS (N = 8), metastatic secondary tumors (N = 2), neurocytoma (N = 2), pineoblastoma (N = 4), Rosai-Dorfman disease (N = 2), and sarcomas (N = 4). Patient sex (`Germline sex estimate`) and tumor histology (`Cancer Group`) are displayed as annotations at the bottom of each plot. Only primary tumors with mutations in the listed genes are shown. Multiple CNVs are denoted as a complex event. (B)*

Genome-wide plot of CNV alterations by broad histology. Each row represents one sample. Box and whisker plots of number of CNV breaks (C) or SV breaks (D) by number of chromothripsis regions.



1582

Figure S4: ***Mutational signatures in pediatric brain tumors, Related to Figure 3.*** *(A) Sample-specific RefSig signature weights across cancer groups ordered by decreasing Signature 1 exposure. (B) Proportion of Signature 1 plotted by phase of therapy for each cancer group.*

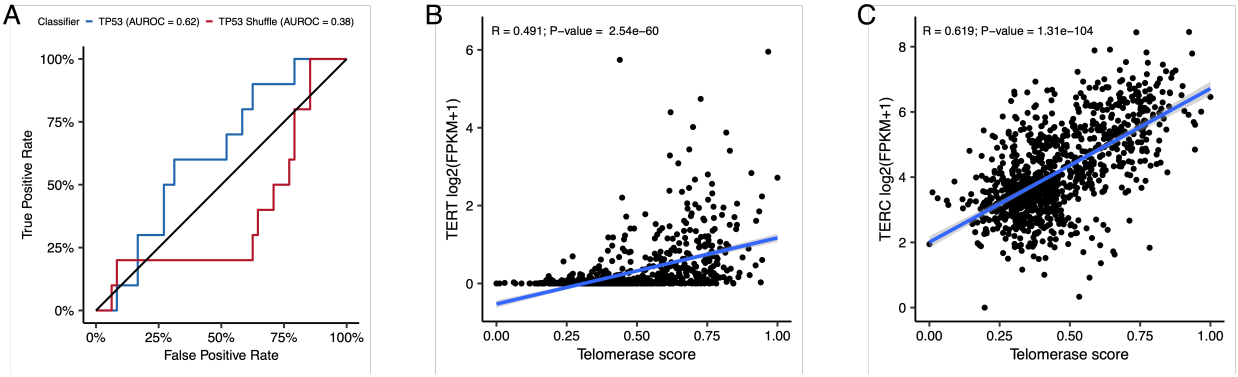


Figure S5: ***Quality control metrics for TP53 and EXTEND scores, Related to Figure 4****. (A) Receiver Operating Characteristic for TP53 classifier run on FPKM of poly-A RNA-Seq samples. Correlation plots for telomerase scores (EXTEND) with RNA expression of TERT (B) and TERC (C).*

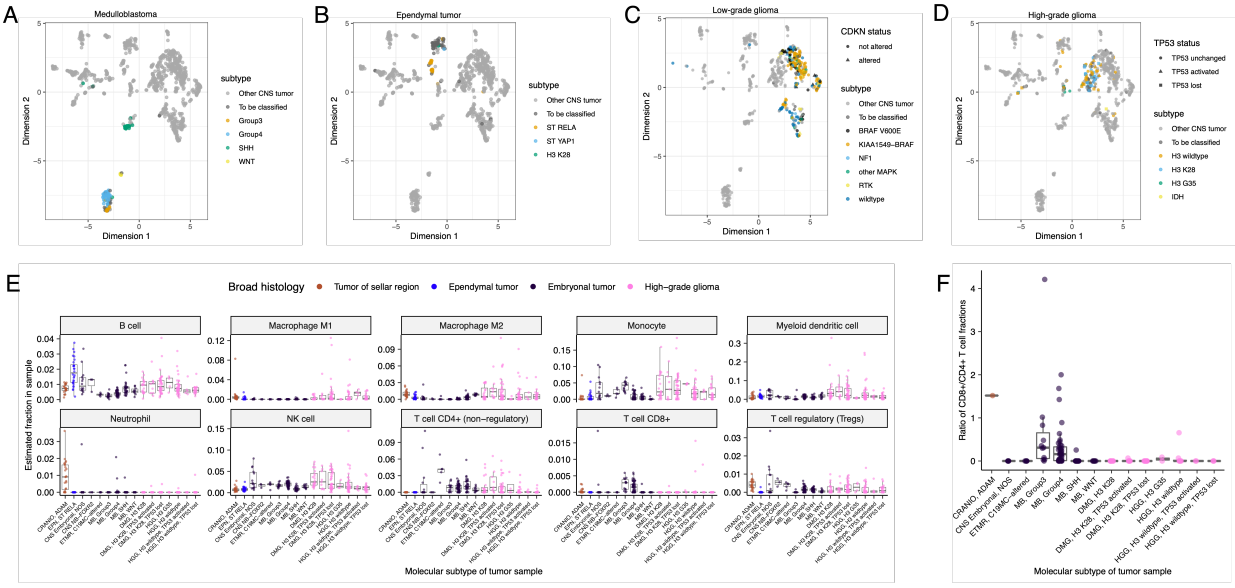


Figure S6: ***Subtype-specific clustering and immune cell fractions, Related to Figure 5****. First two dimensions from UMAP of sample transcriptome data with points colored by* `molecular_subtype` *for medulloblastoma (A), ependymoma (B), low-grade glioma (C), and high-grade glioma (D). (E) Box plots of quanTIseq estimates of immune cell fractions in histologies with more than one molecular subtype with N >=3. (F) Box plots of the ratio of immune cell fractions of CD8+ to CD4+ T cells in histologies with more than one molecular subtype with N >=3.*

Table S1. Related to Figure 1. Table of specimens and associated metadata, clinical data, and histological data utilized in the OpenPBTA project.

Table S2. Related to Figures 2 and 3. Excel file with three sheets representing tables of TMB, eight CNS mutational signatures, and chromothripsis events per sample, respectively.

Table S3. Related to Figures 4 and 5. Excel file with three sheets representing tables of *TP53* scores, telomerase EXTEND scores, and quanTIseq immune scores, respectively.

Table S4. Related to Figures 4 and 5. Excel file with six sheets representing the survival analyses performed for this manuscript. See **Star Methods** for details.

Table S5. Related to Figure 1. Excel file with four sheets representing of all software and their respective versions used for the OpenPBTA project, including the R packages in the OpenPBTA Docker image, Python packages i the OpenPBTA Docker image, other command line tools in the OpenPBTA Docker image, and all software used in the OpenPBTA workflows, respectively. Note that all software in the OpenPBTA Docker image was utilized within the analysis repository, but not all software was used for the final manuscript.

Consortia

The past and present members of the Children's Brain Tumor Network who contributed to the generation of specimens and data are Adam C. Resnick, Alexa Plisiewicz, Allison M. Morgan, Allison P. Heath, Alyssa Paul, Amanda Saratsis, Amy Smith, Ana Aguilar, Ana Guerreiro Stücklin, Anastasia Arynchyna, Andrea Franson, Angela J. Waanders, Angela N. Viaene, Anita Nirenberg, Anna Maria Buccoliero, Anna Yaffe, Anny Shai, Anthony Bet, Antoinette Price, Arlene Luther, Ashley Plant, Augustine Eze, Bailey K. Farrow, Baoli Hu, Beth Frenkel, Bo Zhang, Bobby Moulder, Bonnie Cole, Brian M. Ennis, Brian R. Rood, Brittany Lebert, Carina A. Leonard, Carl Koschmann, Caroline Caudill, Caroline Drinkwater, Cassie N. Kline, Catherine Sullivan, Chanel Keoni, Chiara Caporalini, Christine Bobick-Butcher, Christopher Mason, Chunde Li, Claire Carter, Claudia MaduroCoronado, Clayton Wiley, Cynthia Wong, David E. Kram, David Haussler, David Kram, David Pisapia, David Ziegler, Denise Morinigo, Derek Hanson, Donald W. Parsons, Elizabeth Appert, Emily Drake, Emily Golbeck, Ena Agbodza, Eric H. Raabe, Eric M. Jackson, Erin Alexander, Esteban Uceda, Eugene Hwang, Fausto Rodriquez, Gabrielle S. Stone, Gary Kohanbash, Gavriella Silverman, George Rafidi, Gerald Grant, Gerri Trooskin, Gilad Evrony, Graham Keyes, Hagop Boyajian, Holly B. Lindsay, Holly C. Beale, Ian F. Pollack, James Johnston, James Palmer, Jane Minturn, Jared Pisapia, Jason E. Cain, Jason R. Fangusaro, Javad Nazarian, Jeanette Haugh, Jeff Stevens, Jeffrey P. Greenfield, Jeffrey Rubens, Jena V. Lilly, Jennifer L. Mason, Jessica B. Foster, Jim Olson, Jo Lynne Rokita, Joanna J. Phillips, Jonathan Waller, Josh Rubin, Judy E. Palma, Justin McCroskey, Justine Rizzo, Kaitlin Lehmann, Kamnaa Arya, Karlene Hall, Katherine Pehlivan, Kenneth Seidl, Kimberly Diamond, Kristen Harnett, Kristina A. Cole, Krutika S. Gaonkar, Lamiya Tauhid, Laura Prolo, Leah Holloway, Leslie Brosig, Lina Lopez, Lionel Chow, Madhuri Kambhampati, Mahdi Sarmady, Margaret Nevins, Mari Groves, Mariarita Santi-Vicini, Marilyn M. Li, Marion Mateos, Mateusz Koptyra, Matija Snuderl, Matthew Miller, Matthew Sklar, Matthew Wood, Meghan Connors, Melissa Williams, Meredith Egan, Michael Fisher, Michael Koldobskiy, Michelle Monje,

Migdalia Martinez, Miguel A. Brown, Mike Prados, Miriam Bornhorst, Mirko Scagnet, Mohamed AbdelBaki, Monique Carrero-Tagle, Nadia Dahmane, Nalin Gupta, Nathan Young, Nicholas A. Vitanza, Nicholas Tassone, Nicholas Van Kuren, Nicolas Gerber, Nithin D. Adappa, Nitin Wadhwani, Noel Coleman, Obi Obayashi, Olena M. Vaske, Olivier Elemento, Oren Becher, Philbert Oliveros, Phillip B. Storm, Pichai Raman, Prajwal Rajappa, Rintaro Hashizume, Rishi R. Lulla, Robert Keating, Robert M. Lober, Ron Firestein, Sabine Mueller, Sameer Agnihotri, Samuel G. Winebrake, Samuel Rivero-Hinojosa, Sarah Diane Black, Sarah Leary, Schuyler Stoller, Shannon Robins, Sharon Gardner, Shelly Wang, Sherri Mayans, Sherry Tutson, Shida Zhu, Sofie R. Salama, Sonia Partap, Sonika Dahiya, Sriram Venneti, Stacie Stapleton, Stephani Campion, Stephanie Stefankiewicz, Stewart Goldman, Swetha Thambireddy, Tatiana S. Patton, Teresa Hidalgo, Theo Nicolaides, Thinh Q. Nguyen, Thomas W. McLean, Tiffany Walker, Toba Niazi, Tobey MacDonald, Valeria Lopez-Gil, Valerie Baubet, Whitney Rife, Xiao-Nan Li, Ximena Cuellar, Yiran Guo, Yuankun Zhu, and Zeinab Helil.

The past and present members of the Pacific Pediatric Neuro-Oncology Consortium who contributed to the generation of specimens and data are Adam C. Resnick, Alicia Lenzen, Alyssa Reddy, Amar Gajjar, Ana Guerreiro Stucklin, Anat Epstein, Andrea Franson, Angela Waanders, Anne Bendel, Anu Banerjee, Ashley Margol, Ashley Plant, Brian Rood, Carl Koschmann, Carol Bruggers, Caroline Hastings, Cassie N. Kline, Christina Coleman Abadi, Christopher Tinkle, Corey Raffel, Dan Runco, Daniel Landi, Daphne Adele Haas-Kogan, David Ashley, David Ziegler, Derek Hanson, Dong Anh Khuong Quang, Duane Mitchell, Elias Sayour, Eric Jackson, Eric Raabe, Eugene Hwang, Fatema Malbari, Geoffrey McCowage, Girish Dhall, Gregory Friedman, Hideho Okada, Ibrahim Qaddoumi, Iris Fried, Jae Cho, Jane Minturn, Jason Blatt, Javad Nazarian, Jeffrey Rubens, Jena V. Lilly, Jennifer Elster, Jennifer L. Mason, Jessica Schulte, Jonathan Schoenfeld, Josh Rubin, Karen Gauvain, Karen Wright, Katharine Offer, Katie Metrock, Kellie Haworth, Ken Cohen, Kristina A. Cole, Lance Governale, Linda Stork, Lindsay Kilburn, Lissa Baird, Maggie Skrypek, Marcia Leonard, Margaret Shatara, Margot Lazow, Mariella Filbin, Maryam Fouladi, Matthew Miller, Megan Paul, Michael Fisher, Michael Koldobskiy, Michael Prados, Michal Yalon Oren, Mimi Bandopadhayay, Miriam Bornhorst, Mohamed AbdelBaki, Nalin Gupta, Nathan Robison, Nicholas Whipple, Nick Gottardo, Nicholas A. Vitanza, Nicolas Gerber, Patricia Robertson, Payal Jain, Peter Sun, Priya Chan, Richard S Lemons, Robert Wechsler-Reya, Roger Packer, Russ Geyer, Ryan Velasco, Sabine Mueller, Sahaja Acharya, Sam Cheshier, Sarah Leary, Scott Coven, Sebastian M. Waszak, Sharon Gardner, Sri Gururangan, Stewart Goldman, Susan Chi, Tab Cooney, Tatiana S. Patton, Theodore Nicolaides, and Tom Belle Davidson.

References

1. Ostrom, Q.T., Cioffi, G., Gittleman, H., Patil, N., Waite, K., Kruchko, C., and Barnholtz-Sloan, J.S. (2019). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012–2016. Neuro-Oncology *21*, v1–v100. 10.1093/neuonc/noz150.

2. Ostrom, Q.T., Gittleman, H., Xu, J., Kromer, C., Wolinsky, Y., Kruchko, C., and Barnholtz-Sloan, J.S. (2016). CBTRUS Statistical Report: Primary Brain and Other Central

1681 Nervous System Tumors Diagnosed in the United States in 2009–2013. Neuro-Oncology *18*,
1682 v1–v75. 10.1093/neuonc/now207.

1683 3. Blank, P.M., Ostrom, Q.T., Rouse, C., Wolinsky, Y., Kruchko, C., Salcido, J., and
1684 Barnholtz-Sloan, J.S. (2015). Years of life lived with disease and years of potential life lost in
1685 children who die of cancer in the United States, 2009. Cancer Med *4*, 608–619.
1686 10.1002/cam4.410.

1687 4. Oncology Center of Excellence, U.S. Food and Drug Administration (2022). Pediatric
1688 Oncology Drug Approvals.

1689 5. Vable, A.M., Diehl, S.F., and Glymour, M.M. (2021). Code Review as a Simple Trick to
1690 Enhance Reproducibility, Accelerate Learning, and Improve the Quality of Your Team's
1691 Research. American Journal of Epidemiology *190*, 2172–2177. 10.1093/aje/kwab092.

1692 6. Parker, H. (2017). Opinionated analysis development. PeerJ Preprints, e3210v1.
1693 10.7287/peerj.preprints.3210v1.

1694 7. Beaulieu-Jones, B.K., and Greene, C.S. (2017). Reproducibility of computational
1695 workflows is automated using continuous analysis. Nat Biotechnol *35*, 342–346.
1696 10.1038/nbt.3780.

1697 8. Piwowar, H.A., Day, R.S., and Fridsma, D.B. (2007). Sharing Detailed Research Data Is
1698 Associated with Increased Citation Rate. PLoS ONE *2*, e308. 10.1371/journal.pone.0000308.

1699 9. Cadwallader, L., Papin, J.A., Mac Gabhann, F., and Kirk, R. (2021). Collaborating with
1700 our community to increase code sharing. PLoS Comput Biol *17*, e1008867.
1701 10.1371/journal.pcbi.1008867.

1702 10. Dang, M.T., Gonzalez, M.V., Gaonkar, K.S., Rathi, K.S., Young, P., Arif, S., Zhai, L.,
1703 Alam, Z., Devalaraja, S., To, T.K.J., et al. (2021). Macrophages in SHH subgroup
1704 medulloblastoma display dynamic heterogeneity that varies with treatment modality. Cell
1705 Reports *34*, 108917. 10.1016/j.celrep.2021.108917.

1706 11. Foster, J.B., Griffin, C., Rokita, J.L., Stern, A., Brimley, C., Rathi, K., Lane, M.V.,
1707 Buongervino, S.N., Smith, T., Madsen, P.J., et al. (2021). Development of GPC2-directed
1708 chimeric antigen receptors using mRNA for pediatric brain tumors. bioRxiv, 2021.07.06.451385.
1709 10.1101/2021.07.06.451385.

1710 12. Kline, C., Jain, P., Kilburn, L., Bonner, E.R., Gupta, N., Crawford, J.R., Banerjee, A.,
1711 Packer, R.J., Villanueva-Meyer, J., Luks, T., et al. (2022). Upfront Biology-Guided Therapy in
1712 Diffuse Intrinsic Pontine Glioma: Therapeutic, Molecular, and Biomarker Outcomes from
1713 PNOC003. Clinical Cancer Research, OF1–OF14. 10.1158/1078-0432.ccr-22-0803.

1714 13. RACE Act Poised to Advance Pediatric Cancer Research (2020). Cancer Discovery *10*,
1715 1434–1434. 10.1158/2159-8290.cd-nb2020-081.

14. Ijaz, H., Koptyra, M., Gaonkar, K.S., Rokita, J.L., Baubet, V.P., Tauhid, L., Zhu, Y., Brown, M., Lopez, G., Zhang, B., et al. (2019). Pediatric high-grade glioma resources from the Children's Brain Tumor Tissue Consortium. Neuro-Oncology *22*, 163–165. 10.1093/neuonc/noz192.

15. Mueller, S., Jain, P., Liang, W.S., Kilburn, L., Kline, C., Gupta, N., Panditharatna, E., Magge, S.N., Zhang, B., Zhu, Y., et al. (2019). A pilot precision medicine trial for children with diffuse intrinsic pontine glioma—PNOC003: A report from the Pacific Pediatric Neuro-Oncology Consortium. Int. J. Cancer. 10.1002/ijc.32258.

16. Himmelstein, D.S., Rubinetti, V., Slochower, D.R., Hu, D., Malladi, V.S., Greene, C.S., and Gitter, A. (2019). Open collaborative writing with Manubot. PLoS Comput Biol *15*, e1007128. 10.1371/journal.pcbi.1007128.

17. Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. Linux Journal *2014*, 2:2.

18. Boettiger, C., and Eddelbuettel, D. (2017). An Introduction to Rocker: Docker Containers for R. arXiv:1710.03675 [cs].

19. Kleihues, P., Louis, D.N., Scheithauer, B.W., Rorke, L.B., Reifenberger, G., Burger, P.C., and Cavenee, W.K. (2002). The WHO classification of tumors of the nervous system. J Neuropathol Exp Neurol *61*, 215-25; discussion 226-9. 10.1093/jnen/61.3.215.

20. Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., Burger, P.C., Jouvet, A., Scheithauer, B.W., and Kleihues, P. (2007). The 2007 WHO Classification of Tumours of the Central Nervous System. Acta Neuropathol *114*, 97–109. 10.1007/s00401-007-0243-4.

21. Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., and Ellison, D.W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol *131*, 803–820. 10.1007/s00401-016-1545-1.

22. Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., et al. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncology *23*, 1231–1251. 10.1093/neuonc/noab106.

23. Ryall, S., Zapotocky, M., Fukuoka, K., Nobre, L., Guerreiro Stucklin, A., Bennett, J., Siddaway, R., Li, C., Pajovic, S., Arnoldo, A., et al. (2020). Integrated Molecular and Clinical Analysis of 1,000 Pediatric Low-Grade Gliomas. Cancer Cell *37*, 569–583.e5. 10.1016/j.ccell.2020.03.011.

24. Campbell, B.B., Light, N., Fabrizio, D., Zatzman, M., Fuligni, F., de Borja, R., Davidson, S., Edwards, M., Elvin, J.A., Hodel, K.P., et al. (2017). Comprehensive Analysis of Hypermutation in Human Cancer. Cell *171*, 1042–1056.e10. 10.1016/j.cell.2017.09.048.

25. Ryall, S., Tabori, U., and Hawkins, C. (2020). Pediatric low-grade glioma in the era of molecular diagnostics. acta neuropathol commun *8*. 10.1186/s40478-020-00902-z.

26. Lambo, S., von Hoff, K., Korshunov, A., Pfister, S.M., and Kool, M. (2020). ETMR: a tumor entity in its infancy. Acta Neuropathol *140*, 249–266. 10.1007/s00401-020-02182-2.

27. Richardson, S., Hill, R.M., Kui, C., Lindsey, J.C., Grabovksa, Y., Keeling, C., Pease, L., Bashton, M., Crosier, S., Vinci, M., et al. (2021). Emergence and maintenance of actionable genetic drivers at medulloblastoma relapse. Neuro-Oncology *24*, 153–165. 10.1093/neuonc/noab178.

28. Łastowska, M., Trubicka, J., Sobocińska, A., Wojtas, B., Niemira, M., Szałkowska, A., Krętowski, A., Karkucińska-Więckowska, A., Kaleta, M., Ejmont, M., et al. (2020). Molecular identification of CNS NB-FOXR2, CNS EFT-CIC, CNS HGNET-MN1 and CNS HGNET-BCOR pediatric brain tumors using tumor-specific signature genes. acta neuropathol commun *8*. 10.1186/s40478-020-00984-9.

29. Northcott, P.A., Buchhalter, I., Morrissy, A.S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V.A., et al. (2017). The whole-genome landscape of medulloblastoma subtypes. Nature *547*, 311–317. 10.1038/nature22973.

30. Mackay, A., Burford, A., Carvalho, D., Izquierdo, E., Fazal-Salom, J., Taylor, K.R., Bjerke, L., Clarke, M., Vinci, M., Nandhabalan, M., et al. (2017). Integrated Molecular Meta-Analysis of 1,000 Pediatric High-Grade and Diffuse Intrinsic Pontine Glioma. Cancer Cell *32*, 520–537.e5. 10.1016/j.ccell.2017.08.017.

31. Pratt, D., Quezado, M., Abdullaev, Z., Hawes, D., Yang, F., Garton, H.J.L., Judkins, A.R., Mody, R., Chinnaiyan, A., Aldape, K., et al. (2020). Diffuse intrinsic pontine glioma-like tumor with EZHIP expression and molecular features of PFA ependymoma. acta neuropathol commun *8*. 10.1186/s40478-020-00905-w.

32. Surrey, L.F., Jain, P., Zhang, B., Straka, J., Zhao, X., Harding, B.N., Resnick, A.C., Storm, P.B., Buccoliero, A.M., Genitori, L., et al. (2019). Genomic Analysis of Dysembryoplastic Neuroepithelial Tumor Spectrum Reveals a Diversity of Molecular Alterations Dysregulating the MAPK and PI3K/mTOR Pathways. Journal of Neuropathology & Experimental Neurology *78*, 1100–1111. 10.1093/jnen/nlz101.

33. Pfaff, E., Remke, M., Sturm, D., Benner, A., Witt, H., Milde, T., von Bueren, A.O., Wittmann, A., Schöttler, A., Jorch, N., et al. (2010). *TP53* Mutation Is Frequently Associated With *CTNNB1* Mutation or *MYCN* Amplification and Is Compatible With Long-Term Survival in Medulloblastoma. JCO *28*, 5188–5196. 10.1200/jco.2010.31.1670.

34. Lucas, C.-H.G., Gupta, R., Doo, P., Lee, J.C., Cadwell, C.R., Ramani, B., Hofmann, J.W., Sloan, E.A., Kleinschmidt-DeMasters, B.K., Lee, H.S., et al. (2020). Comprehensive analysis of diverse low-grade neuroepithelial tumors with FGFR1 alterations reveals a distinct

1789 molecular signature of rosette-forming glioneuronal tumor. acta neuropathol commun *8*.
1790 10.1186/s40478-020-01027-z.

1791 35. Wu, G., Diaz, A.K., Paugh, B.S., Rankin, S.L., Ju, B., Li, Y., Zhu, X., Qu, C., Chen, X.,
1792 Zhang, J., et al. (2014). The genomic landscape of diffuse intrinsic pontine glioma and pediatric
1793 non-brainstem high-grade glioma. Nature Genetics *46*, 444–450. 10.1038/ng.2938.

1794 36. Degasperi, A., Amarante, T.D., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D.,
1795 Morganella, S., Nanda, A.S., Badja, C., Koh, G., et al. (2020). A practical framework and online
1796 tool for mutational signature analyses show intertissue variation and driver dependencies. Nat
1797 Cancer *1*, 249–263. 10.1038/s43018-020-0027-5.

1798 37. Wojciechowicz, K., Cantelli, E., Van Gerwen, B., Plug, M., Van Der Wal, A., Delzenne-
1799 Goette, E., Song, J.-Y., De Vries, S., Dekker, M., and Riele, H.T. (2014). Temozolomide
1800 Increases the Number of Mismatch Repair–Deficient Intestinal Crypts and Accelerates
1801 Tumorigenesis in a Mouse Model of Lynch Syndrome. Gastroenterology *147*, 1064–1072.e5.
1802 10.1053/j.gastro.2014.07.052.

1803 38. Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack,
1804 A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al. (2018). Genomic and Molecular
1805 Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Reports
1806 *23*, 239–254.e6. 10.1016/j.celrep.2018.03.076.

1807 39. Dittmer, D., Pati, S., Zambetti, G., Chu, S., Teresky, A.K., Moore, M., Finlay, C., and
1808 Levine, A.J. (1993). Gain of function mutations in p53. Nat Genet *4*, 42–46. 10.1038/ng0593-42.

1809 40. Rokita, J.L., Rathi, K.S., Cardenas, M.F., Upton, K.A., Jayaseelan, J., Cross, K.L., Pfeil,
1810 J., Egolf, L.E., Way, G.P., Farrel, A., et al. (2019). Genomic Profiling of Childhood Tumor
1811 Patient-Derived Xenograft Models to Enable Rational Clinical Trial Design. Cell Reports *29*,
1812 1675–1689.e9. 10.1016/j.celrep.2019.09.071.

1813 41. Noureen, N., Wu, S., Lv, Y., Yang, J., Alfred Yung, W.K., Gelfond, J., Wang, X., Koul, D.,
1814 Ludlow, A., and Zheng, S. (2021). Integrated analysis of telomerase enzymatic activity unravels
1815 an association with cancer stemness and proliferation. Nat Commun *12*. 10.1038/s41467-020-
1816 20474-9.

1817 42. Artandi, S.E., and DePinho, R.A. (2009). Telomeres and telomerase in cancer.
1818 Carcinogenesis *31*, 9–18. 10.1093/carcin/bgp268.

1819 43. Ceja-Rangel, H.A., Sánchez-Suárez, P., Castellanos-Juárez, E., Peñaroja-Flores, R.,
1820 Arenas-Aranda, D.J., Gariglio, P., and Benítez-Bribiesca, L. (2016). Shorter telomeres and high
1821 telomerase activity correlate with a highly aggressive phenotype in breast cancer cell lines.
1822 Tumor Biol. *37*, 11917–11926. 10.1007/s13277-016-5045-7.

1823 44. Oh, B.-K., Kim, H., Park, Y.N., Yoo, J.E., Choi, J., Kim, K.-S., Lee, J.J., and Park, C.
1824 (2007). High telomerase activity and long telomeres in advanced hepatocellular carcinomas with
1825 poor prognosis. Lab Invest *88*, 144–152. 10.1038/labinvest.3700710.

1826 45. Kulić, A., Plavetić, N.D., Gamulin, S., Jakić-Razumović, J., Vrbanec, D., and Sirotković-
1827 Skerlev, M. (2016). Telomerase activity in breast cancer patients: association with poor
1828 prognosis and more aggressive phenotype. Med Oncol *33*. 10.1007/s12032-016-0736-x.

1829 46. Wong, V.C.H., Morrison, A., Tabori, U., and Hawkins, C.E. (2010). Telomerase Inhibition
1830 as a Novel Therapy for Pediatric Ependymoma. Brain Pathology *20*, 780–786. 10.1111/j.1750-
1831 3639.2010.00372.x.

1832 47. Pich, O., Muiños, F., Lolkema, M.P., Steeghs, N., Gonzalez-Perez, A., and Lopez-Bigas,
1833 N. (2019). The mutational footprints of cancer therapies. Nat Genet *51*, 1732–1740.
1834 10.1038/s41588-019-0525-5.

1835 48. Aronson, M., Colas, C., Shuen, A., Hampel, H., Foulkes, W.D., Baris Feldman, H.,
1836 Goldberg, Y., Muleris, M., Wolfe Schneider, K., McGee, R.B., et al. (2021). Diagnostic criteria
1837 for constitutional mismatch repair deficiency (CMMRD): recommendations from the international
1838 consensus working group. J Med Genet *59*, 318–327. 10.1136/jmedgenet-2020-107627.

1839 49. Lewis, P.W., Müller, M.M., Koletsky, M.S., Cordero, F., Lin, S., Banaszynski, L.A.,
1840 Garcia, B.A., Muir, T.W., Becher, O.J., and Allis, C.D. (2013). Inhibition of PRC2 Activity by a
1841 Gain-of-Function H3 Mutation Found in Pediatric Glioblastoma. Science *340*, 857–861.
1842 10.1126/science.1232245.

1843 50. Hutter, S., Bolin, S., Weishaupt, H., and Swartling, F. (2017). Modeling and Targeting
1844 MYC Genes in Childhood Brain Tumors. Genes *8*, 107. 10.3390/genes8040107.

1845 51. Leone, G., Sears, R., Huang, E., Rempel, R., Nuckolls, F., Park, C.H., Giangrande, P.,
1846 Wu, L., Saavedra, H.I., Field, S.J., et al. (2001). Myc requires distinct E2F activities to induce S
1847 phase and apoptosis. Mol Cell *8*, 105–13. 10.1016/s1097-2765(01)00275-1.

1848 52. Hannan, C.J., Lewis, D., O'Leary, C., Donofrio, C.A., Evans, D.G., Roncaroli, F., Brough,
1849 D., King, A.T., Coope, D., and Pathmanaban, O.N. (2020). The inflammatory microenvironment
1850 in vestibular schwannoma. Neuro-Oncology Advances *2*. 10.1093/noajnl/vdaa023.

1851 53. Petralia, F., Tignor, N., Reva, B., Koptyra, M., Chowdhury, S., Rykunov, D., Krek, A.,
1852 Ma, W., Zhu, Y., Ji, J., et al. (2020). Integrated Proteogenomic Characterization across Major
1853 Histological Types of Pediatric Brain Cancer. Cell *183*, 1962–1985.e31.
1854 10.1016/j.cell.2020.10.044.

1855 54. Lin, G.L., Nagaraja, S., Filbin, M.G., Suvà, M.L., Vogel, H., and Monje, M. (2018). Non-
1856 inflammatory tumor microenvironment of diffuse intrinsic pontine glioma. acta neuropathol
1857 commun *6*. 10.1186/s40478-018-0553-x.

1858 55. Ross, J.L., Velazquez Vega, J., Plant, A., MacDonald, T.J., Becher, O.J., and
1859 Hambardzumyan, D. (2021). Tumour immune landscape of paediatric high-grade gliomas. Brain
1860 *144*, 2594–2609. 10.1093/brain/awab155.

1861 56. Martin, A.M., Nirschl, C.J., Polanczyk, M.J., Bell, W.R., Nirschl, T.R., Harris-Bookman,
1862 S., Phallen, J., Hicks, J., Martinez, D., Ogurtsova, A., et al. (2018). PD-L1 expression in

1863 medulloblastoma: an evaluation by subgroup. Oncotarget *9*, 19177–19191.
1864 10.18632/oncotarget.24951.

1865 57. Bockmayr, M., Mohme, M., Klauschen, F., Winkler, B., Budczies, J., Rutkowski, S., and
1866 Schüller, U. (2018). Subgroup-specific immune and stromal microenvironment in
1867 medulloblastoma. OncoImmunology *7*, e1462430. 10.1080/2162402x.2018.1462430.

1868 58. Duchemann, B., Naigeon, M., Auclin, E., Ferrara, R., Cassard, L., Jouniaux, J.-M.,
1869 Boselli, L., Grivel, J., Desnoyer, A., Danlos, F.-X., et al. (2022). CD8+ PD-1+ to CD4+PD-1+
1870 ratio (PERLS) is associated with prognosis of patients with advanced NSCLC treated with PD-
1871 (L)1 blockers. J Immunother Cancer *10*, e004012. 10.1136/jitc-2021-004012.

1872 59. Shindo, G., Endo, T., Onda, M., Goto, S., Miyamoto, Y., and Kaneko, T. (2013). Is the
1873 CD4/CD8 Ratio an Effective Indicator for Clinical Estimation of Adoptive Immunotherapy for
1874 Cancer Treatment? JCT *04*, 1382–1390. 10.4236/jct.2013.48164.

1875 60. Gendoo, D.M.A., and Haibe-Kains, B. (2016). MM2S: personalized diagnosis of
1876 medulloblastoma patients and model systems. Source Code Biol Med *11*. 10.1186/s13029-016-
1877 0053-y.

1878 61. Rathi, K.S., Arif, S., Koptyra, M., Naqvi, A.S., Taylor, D.M., Storm, P.B., Resnick, A.C.,
1879 Rokita, J.L., and Raman, P. (2020). A transcriptome-based classifier to determine molecular
1880 subtypes in medulloblastoma. PLoS Comput Biol *16*, e1008263. 10.1371/journal.pcbi.1008263.

1881 62. Yuza, K., Nagahashi, M., Watanabe, S., Takabe, K., and Wakai, T. (2017).
1882 Hypermutation and microsatellite instability in gastrointestinal cancers. Oncotarget *8*, 112103–
1883 112115. 10.18632/oncotarget.22783.

1884 63. Bass, A.J., Thorsson, V., Shmulevich, I., Reynolds, S.M., Miller, M., Bernard, B., Hinoue,
1885 T., Laird, P.W., Curtis, C., Shen, H., et al. (2014). Comprehensive molecular characterization of
1886 gastric adenocarcinoma. Nature *513*, 202–209. 10.1038/nature13480.

1887 64. Sharma, J., Bonfield, C.M., Singhal, A., Hukin, J., and Steinbok, P. (2015). Intracystic
1888 interferon-α treatment leads to neurotoxicity in craniopharyngioma: case report. PED *16*, 301–
1889 304. 10.3171/2015.2.peds14656.

1890 65. Mohammed, K.E.A., Mike, K.R.A., and Parkes, J. (2013). Unexpected brain atrophy
1891 following administration of intratumoral interferon alpha-2b for cystic craniopharyngioma: A case
1892 report. IJCRI *4*, 719. 10.5348/ijcri-2013-12-419-cr-13.

1893 66. Coy, S., Rashid, R., Lin, J.-R., Du, Z., Donson, A.M., Hankinson, T.C., Foreman, N.K.,
1894 Manley, P.E., Kieran, M.W., Reardon, D.A., et al. (2018). Multiplexed immunofluorescence
1895 reveals potential PD-1/PD-L1 pathway vulnerabilities in craniopharyngioma. Neuro-Oncology
1896 *20*, 1101–1112. 10.1093/neuonc/noy035.

1897 67. Yuan, F., Cai, X., Zhu, J., Yuan, L., Wang, Y., Tang, C., Cong, Z., and Ma, C. (2021). A
1898 Novel Immune Classification for Predicting Immunotherapy Responsiveness in Patients With
1899 Adamantinomatous Craniopharyngioma. Front Neurol *12*, 704130. 10.3389/fneur.2021.704130.

1900 68. Whelan, R., Prince, E., Gilani, A., and Hankinson, T. (2020). The Inflammatory Milieu of
1901 Adamantinomatous Craniopharyngioma and Its Implications for Treatment. J Clin Med *9*.
1902 10.3390/jcm9020519.

1903 69. Apps, J.R., Carreno, G., Gonzalez-Meljem, J.M., Haston, S., Guiho, R., Cooper, J.E.,
1904 Manshaei, S., Jani, N., Hölsken, A., Pettorini, B., et al. (2018). Tumour compartment
1905 transcriptomics demonstrates the activation of inflammatory and odontogenic programmes in
1906 human adamantinomatous craniopharyngioma and identifies the MAPK/ERK pathway as a
1907 novel therapeutic target. Acta Neuropathol *135*, 757–777. 10.1007/s00401-018-1830-2.

1908 70. Grob, S., Mirsky, D.M., Donson, A.M., Dahl, N., Foreman, N.K., Hoffman, L.M.,
1909 Hankinson, T.C., and Mulcahy Levy, J.M. (2019). Targeting IL-6 Is a Potential Treatment for
1910 Primary Cystic Craniopharyngioma. Front. Oncol. *9*. 10.3389/fonc.2019.00791.

1911 71. Gaonkar, K.S., Marini, F., Rathi, K.S., Jain, P., Zhu, Y., Chimicles, N.A., Brown, M.A.,
1912 Naqvi, A.S., Zhang, B., Storm, P.B., et al. (2020). annoFuse: an R Package to annotate,
1913 prioritize, and interactively explore putative oncogenic RNA fusions. BMC Bioinformatics *21*.
1914 10.1186/s12859-020-03922-7.

1915 72. University of California, San Francisco (2022). A Pilot Trial Testing the Clinical Benefit of
1916 Using Molecular Profiling to Determine an Individualized Treatment Plan in Children and Young
1917 Adults With High Grade Glioma (Excluding Diffuse Intrinsic Pontine Glioma) (clinicaltrials.gov).

1918 73. Open Pediatric Brain Tumor Atlas, C.B.T.N., Pediatric Neuro Oncology Consortium
1919 (2022). Open Pediatric Brain Tumor Atlas. 10.24370/openpbta.

1920 74. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with
1921 BWA-MEM. arXiv:1303.3997 [q-bio].

1922 75. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C.,
1923 Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation
1924 discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491–498.
1925 10.1038/ng.806.

1926 76. Faust, G.G., and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural
1927 variant read extraction. Bioinformatics *30*, 2503–2505. 10.1093/bioinformatics/btu314.

1928 77. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast
1929 processing of NGS alignment formats. Bioinformatics *31*, 2032–2034.
1930 10.1093/bioinformatics/btv098.

1931 78. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A.,
1932 Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A
1933 MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*,
1934 1297–1303. 10.1101/gr.107524.110.

1935 79. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Auwera,
1936 G.A.V. der, Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling

1937 accurate genetic variant discovery to tens of thousands of samples. bioRxiv, 201178.
1938 10.1101/201178.

1939 80. Lee, S., Lee, S., Ouellette, S., Park, W.-Y., Lee, E.A., and Park, P.J. (2017).
1940 NGSCheckMate: software for validating sample identity in next-generation sequencing studies
1941 within and across data types. Nucleic Acids Research *45*, e103–e103. 10.1093/nar/gkx193.

1942 81. DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C.,
1943 Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control
1944 and process optimization. Bioinformatics *28*, 1530–1532. 10.1093/bioinformatics/bts196.

1945 82. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of
1946 genetic variants from high-throughput sequencing data. Nucleic Acids Research *38*, e164–e164.
1947 10.1093/nar/gkq603.

1948 83. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and
1949 Maglott, D.R. (2013). ClinVar: public archive of relationships among sequence variation and
1950 human phenotype. Nucl. Acids Res. *42*, D980–D985. 10.1093/nar/gkt1113.

1951 84. Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M., and Lu, X.
1952 (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies
1953 with a New Program, SnpSift. Front. Gene. *3*. 10.3389/fgene.2012.00035.

1954 85. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X.,
1955 Kim, Y., Beyter, D., Krusche, P., et al. (2018). Strelka2: fast and accurate calling of germline
1956 and somatic variants. Nat Methods *15*, 591–594. 10.1038/s41592-018-0051-x.

1957 86. Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019).
1958 Calling Somatic SNVs and Indels with Mutect2. bioRxiv, 861054. 10.1101/861054.

1959 87. Narzisi, G., Corvelo, A., Arora, K., Bergmann, E.A., Shah, M., Musunuri, R., Emde, A.-
1960 K., Robine, N., Vacic, V., and Zody, M.C. (2018). Genome-wide somatic variant calling using
1961 localized colored de Bruijn graphs. Commun Biol *1*. 10.1038/s42003-018-0023-9.

1962 88. Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R.,
1963 Johnson, J., Dougherty, B., Barrett, J.C., and Dry, J.R. (2016). VarDict: a novel and versatile
1964 variant caller for next-generation sequencing in cancer research. Nucleic Acids Res *44*, e108–
1965 e108. 10.1093/nar/gkw227.

1966 89. Arora, K., Shah, M., Johnson, M., Sanghvi, R., Shelton, J., Nagulapalli, K., Oschwald,
1967 D.M., Zody, M.C., Germer, S., Jobanputra, V., et al. (2019). Deep whole-genome sequencing of
1968 3 cancer cell lines on 2 sequencing platforms. Sci Rep *9*. 10.1038/s41598-019-55636-3.

1969 90. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P.,
1970 and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*.
1971 10.1186/s13059-016-0974-4.

1972 91. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins,
1973 R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint
1974 spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. 10.1038/s41586-
1975 020-2308-7.

1976 92. Zvereva, M., Pisarev, E., Hosen, I., Kisil, O., Matskeplishvili, S., Kubareva, E., Kamalov,
1977 D., Tivtikyan, A., Manel, A., Vian, E., et al. (2020). Activating Telomerase TERT Promoter
1978 Mutations and Their Application for the Detection of Bladder Cancer. IJMS *21*, 6034.
1979 10.3390/ijms21176034.

1980 93. Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G.,
1981 Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011). Control-FREEC: a tool for assessing
1982 copy number and allelic content using next-generation sequencing data. Bioinformatics *28*,
1983 423–425. 10.1093/bioinformatics/btr670.

1984 94. Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.-P., Janoueix-Lerosey, I., Delattre, O., and
1985 Barillot, E. (2010). Control-free calling of copy number alterations in deep-sequencing data
1986 using GC-content normalization. Bioinformatics *27*, 268–269. 10.1093/bioinformatics/btq635.

1987 95. Talevich, E., Shain, A.H., Botton, T., and Bastian, B.C. (2016). CNVkit: Genome-Wide
1988 Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol
1989 *12*, e1004873. 10.1371/journal.pcbi.1004873.

1990 96. Oesper, L., Satas, G., and Raphael, B.J. (2014). Quantifying tumor heterogeneity in
1991 whole-genome and whole-exome sequencing data. Bioinformatics *30*, 3532–3540.
1992 10.1093/bioinformatics/btu651.

1993 97. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G.
1994 (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic
1995 copy-number alteration in human cancers. Genome Biol *12*. 10.1186/gb-2011-12-4-r41.

1996 98. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox,
1997 A.J., Kruglyak, S., and Saunders, C.T. (2015). Manta: rapid detection of structural variants and
1998 indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.
1999 10.1093/bioinformatics/btv710.

2000 99. Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., and Muller, J.
2001 (2018). AnnotSV: an integrated tool for structural variations annotation. Bioinformatics *34*,
2002 3572–3574. 10.1093/bioinformatics/bty304.

2003 100. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P.,
2004 Chaisson, M., and Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner.
2005 Bioinformatics *29*, 15–21. 10.1093/bioinformatics/bts635.

2006 101. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq
2007 data with or without a reference genome. BMC Bioinformatics *12*. 10.1186/1471-2105-12-323.

2008 102. Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U.H.,
2009 Neumann, O., Stenzinger, A., Scholl, C., et al. (2021). Accurate and efficient detection of gene
2010 fusions from RNA sequencing data. Genome Res. *31*, 448–460. 10.1101/gr.257246.119.

2011 103. Haas, B.J., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., Bankapur, A., Ganote, C.,
2012 Doak, T.G., Pochet, N., et al. (2017). STAR-Fusion: Fast and Accurate Fusion Transcript
2013 Detection from RNA-Seq. bioRxiv, 120295. 10.1101/120295.

2014 104. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H.P. (2018). Maftools:
2015 efficient and comprehensive analysis of somatic variants in cancer. Genome Res. *28*, 1747–
2016 1756. 10.1101/gr.239244.118.

2017 105. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing
2018 genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.

2019 106. Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan,
2020 C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2012). The UCSC Genome Browser
2021 database: extensions and updates 2013. Nucleic Acids Research *41*, D64–D69.
2022 10.1093/nar/gks1048.

2023 107. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan,
2024 M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLoS
2025 Comput Biol *9*, e1003118. 10.1371/journal.pcbi.1003118.

2026 108. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and
2027 correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849.
2028 10.1093/bioinformatics/btw313.

2029 109. Cortés-Ciriano, I., Lee, J.J.-K., Xi, R., Jain, D., Jung, Y.L., Yang, L., Gordenin, D.,
2030 Klimczak, L.J., Zhang, C.-Z., Pellman, D.S., et al. (2020). Comprehensive analysis of
2031 chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet *52*, 331–
2032 341. 10.1038/s41588-019-0576-7.

2033 110. Sturm, G., Finotello, F., Petitprez, F., Zhang, J.D., Baumbach, J., Fridman, W.H., List,
2034 M., and Aneichyk, T. (2019). Comprehensive evaluation of transcriptome-based cell-type
2035 quantification methods for immuno-oncology. Bioinformatics *35*, i436–i445.
2036 10.1093/bioinformatics/btz363.

2037 111. Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., Krogsdam, A.,
2038 Loncova, Z., Posch, W., Wilflingseder, D., et al. (2019). Molecular and pharmacological
2039 modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data.
2040 Genome Med *11*. 10.1186/s13073-019-0638-6.

2041 112. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis
2042 for microarray and RNA-Seq data. BMC Bioinformatics *14*. 10.1186/1471-2105-14-7.

2043 113. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, Jill P., and Tamayo, P.
2044 (2015). The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems *1*,
2045 417–425. 10.1016/j.cels.2015.12.004.

2046 114. McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation
2047 and Projection for Dimension Reduction. arXiv:1802.03426 [cs, stat].

2048 115. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale,
2049 J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–
2050 665. 10.1016/j.cell.2018.01.029.

2051 116. Ramkissoon, L.A., Horowitz, P.M., Craig, J.M., Ramkissoon, S.H., Rich, B.E.,
2052 Schumacher, S.E., McKenna, A., Lawrence, M.S., Bergthold, G., Brastianos, P.K., et al. (2013).
2053 Genomic analysis of diffuse pediatric low-grade gliomas identifies recurrent oncogenic
2054 truncating rearrangements in the transcription factor *MYBL1*. Proc. Natl. Acad. Sci. U.S.A. *110*,
2055 8188–8193. 10.1073/pnas.1300252110.

2056 117. Northcott, P.A., Shih, D.J.H., Peacock, J., Garzia, L., Sorana Morrissy, A., Zichner, T.,
2057 Stütz, A.M., Korshunov, A., Reimand, J., Schumacher, S.E., et al. (2012). Subgroup-specific
2058 structural variation across 1,000 medulloblastoma genomes. Nature *488*, 49–56.
2059 10.1038/nature11327.

2060 118. Sturm, D., Orr, Brent A., Toprak, Umut H., Hovestadt, V., Jones, David T.W., Capper, D.,
2061 Sill, M., Buchhalter, I., Northcott, Paul A., Leis, I., et al. (2016). New Brain Tumor Entities
2062 Emerge from Molecular Classification of CNS-PNETs. Cell *164*, 1060–1072.
2063 10.1016/j.cell.2016.01.015.

2064 119. Kleinman, C.L., Gerges, N., Papillon-Cavanagh, S., Sin-Chan, P., Pramatarova, A.,
2065 Quang, D.-A.K., Adoue, V., Busche, S., Caron, M., Djambazian, H., et al. (2013). Fusion of
2066 TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B
2067 isoform in the embryonal brain tumor ETMR. Nat Genet *46*, 39–44. 10.1038/ng.2849.

2068 120. Valentijn, L.J., Koster, J., Zwijnenburg, D.A., Hasselt, N.E., van Sluis, P., Volckmann, R.,
2069 van Noesel, M.M., George, R.E., Tytgat, G.A.M., Molenaar, J.J., et al. (2015). TERT
2070 rearrangements are frequent in neuroblastoma and identify aggressive tumors. Nat Genet *47*,
2071 1411–1414. 10.1038/ng.3438.

2072 121. Cobrinik, D., Ostrovnaya, I., Hassimi, M., Tickoo, S.K., Cheung, I.Y., and Cheung, N.-
2073 K.V. (2013). Recurrent pre-existing and acquired DNA copy number alterations, including focal
2074 *TERT* gains, in neuroblastoma central nervous system metastases. Genes Chromosomes
2075 Cancer *52*, 1150–1166. 10.1002/gcc.22110.

2076 122. Karlsson, J., Lilljebjörn, H., Holmquist Mengelbier, L., Valind, A., Rissler, M., Øra, I.,
2077 Fioretos, T., and Gisselsson, D. (2015). Activation of human telomerase reverse transcriptase
2078 through gene fusion in clear cell sarcoma of the kidney. Cancer Letters *357*, 498–501.
2079 10.1016/j.canlet.2014.11.057.

123. Karsy, M., Guan, J., Cohen, A.L., Jensen, R.L., and Colman, H. (2017). New Molecular Considerations for Glioma: IDH, ATRX, BRAF, TERT, H3 K27M. Curr Neurol Neurosci Rep *17*. 10.1007/s11910-017-0722-5.

124. Bandopadhayay, P., Ramkissoon, L.A., Jain, P., Bergthold, G., Wala, J., Zeid, R., Schumacher, S.E., Urbanski, L., O'Rourke, R., Gibson, W.J., et al. (2016). MYB-QKI rearrangements in angiocentric glioma drive tumorigenicity through a tripartite mechanism. Nat Genet *48*, 273–282. 10.1038/ng.3500.

125. Johann, P.D., Erkek, S., Zapatka, M., Kerl, K., Buchhalter, I., Hovestadt, V., Jones, D.T.W., Sturm, D., Hermann, C., Segura Wang, M., et al. (2016). Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes. Cancer Cell *29*, 379–393. 10.1016/j.ccell.2016.02.001.

126. Mong, E.F., Yang, Y., Akat, K.M., Canfield, J., VanWye, J., Lockhart, J., Tsibris, J.C.M., Schatz, F., Lockwood, C.J., Tuschl, T., et al. (2020). Chromosome 19 microRNA cluster enhances cell reprogramming by inhibiting epithelial-to-mesenchymal transition. Sci Rep *10*. 10.1038/s41598-020-59812-8.

127. Ryall, S., Guzman, M., Elbabaa, S.K., Luu, B., Mack, S.C., Zapotocky, M., Taylor, M.D., Hawkins, C., and Ramaswamy, V. (2017). H3 K27M mutations are extremely rare in posterior fossa group A ependymoma. Childs Nerv Syst *33*, 1047–1051. 10.1007/s00381-017-3481-3.

128. Pajtler, Kristian W., Witt, H., Sill, M., Jones, David T.W., Hovestadt, V., Kratochwil, F., Wani, K., Tatevossian, R., Punchihewa, C., Johann, P., et al. (2015). Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. Cancer Cell *27*, 728–743. 10.1016/j.ccell.2015.04.002.

129. Parker, M., Mohankumar, K.M., Punchihewa, C., Weinlich, R., Dalton, J.D., Li, Y., Lee, R., Tatevossian, R.G., Phoenix, T.N., Thiruvenkatam, R., et al. (2014). C11orf95–RELA fusions drive oncogenic NF-κB signalling in ependymoma. Nature *506*, 451–455. 10.1038/nature13109.

130. Bi, W.L., Greenwald, N.F., Abedalthagafi, M., Wala, J., Gibson, W.J., Agarwalla, P.K., Horowitz, P., Schumacher, S.E., Esaulova, E., Mei, Y., et al. (2017). Genomic landscape of high-grade meningiomas. npj Genomic Med *2*. 10.1038/s41525-017-0014-7.

131. Youngblood, M.W., Duran, D., Montejo, J.D., Li, C., Omay, S.B., Özduman, K., Sheth, A.H., Zhao, A.Y., Tyrtova, E., Miyagishima, D.F., et al. (2020). Correlations between genomic subgroup and clinical features in a cohort of more than 3000 meningiomas. Journal of Neurosurgery *133*, 1345–1354. 10.3171/2019.8.jns191266.

132. Qaddoumi, I., Orisme, W., Wen, J., Santiago, T., Gupta, K., Dalton, J.D., Tang, B., Haupfear, K., Punchihewa, C., Easton, J., et al. (2016). Genetic alterations in uncommon low-grade neuroepithelial tumors: BRAF, FGFR1, and MYB mutations occur at high frequency and align with morphology. Acta Neuropathol *131*, 833–845. 10.1007/s00401-016-1539-z.

133. Thomas, C., Soschinski, P., Zwaig, M., Oikonomopoulos, S., Okonechnikov, K., Pajtler, K.W., Sill, M., Schweizer, L., Koch, A., Neumann, J., et al. (2020). The genetic landscape of

choroid plexus tumors in children and adults. Neuro-Oncology *23*, 650–660. 10.1093/neuonc/noaa267.

134. Sekine, S., Shibata, T., Kokubu, A., Morishita, Y., Noguchi, M., Nakanishi, Y., Sakamoto, M., and Hirohashi, S. (2002). Craniopharyngiomas of adamantinomatous type harbor beta-catenin gene mutations. Am J Pathol *161*, 1997–2001. 10.1016/s0002-9440(10)64477-x.

135. Krooks, J., Minkov, M., and Weatherall, A.G. (2018). Langerhans cell histiocytosis in children. Journal of the American Academy of Dermatology *78*, 1035–1044. 10.1016/j.jaad.2017.05.059.

136. Antin, C., Tauziède-Espariat, A., Debily, M.-A., Castel, D., Grill, J., Pagès, M., Ayrault, O., Chrétien, F., Gareton, A., Andreiuolo, F., et al. (2020). EZHIP is a specific diagnostic biomarker for posterior fossa ependymomas, group PFA and diffuse midline gliomas H3-WT with EZHIP overexpression. acta neuropathol commun *8*. 10.1186/s40478-020-01056-8.

137. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol *17*. 10.1186/s13059-016-0893-4.

138. Burel-Vandenbos, F., Pierron, G., Thomas, C., Reynaud, S., Gregoire, V., Duhil de Benaze, G., Croze, S., Chivoret, N., Honavar, M., Figarella-Branger, D., et al. (2020). A polyphenotypic malignant paediatric brain tumour presenting a *MN1-PATZ1* fusion, no epigenetic similarities with CNS High-Grade Neuroepithelial Tumour with *MN1* Alteration (CNS HGNET-MN1) and related to *PATZ1*-fused sarcomas. Neuropathol Appl Neurobiol *46*, 506–509. 10.1111/nan.12626.

139. Kram, D.E., Henderson, J.J., Baig, M., Chakraborty, D., Gardner, M.A., Biswas, S., and Khatua, S. (2018). Embryonal Tumors of the Central Nervous System in Children: The Era of Targeted Therapeutics. Bioengineering (Basel) *5*. 10.3390/bioengineering5040078.

140. Rao, S., Rajeswarie, R.T., Chickabasaviah Yasha, T., Nandeesh, B.N., Arivazhagan, A., and Santosh, V. (2017). LIN28A, a sensitive immunohistochemical marker for Embryonal Tumor with Multilayered Rosettes (ETMR), is also positive in a subset of Atypical Teratoid/Rhabdoid Tumor (AT/RT). Childs Nerv Syst *33*, 1953–1959. 10.1007/s00381-017-3551-6.

141. PDQ® Pediatric Treatment Editorial Board Childhood Medulloblastoma and Other Central Nervous System Embryonal Tumors Treatment (PDQ®): Health Professional Version. In PDQ Cancer Information Summaries (National Cancer Institute).

142. Miele, E., De Vito, R., Ciolfi, A., Pedace, L., Russo, I., De Pasquale, M.D., Di Giannatale, A., Crocoli, A., Angelis, B.D., Tartaglia, M., et al. (2020). DNA Methylation Profiling for Diagnosing Undifferentiated Sarcoma with Capicua Transcriptional Receptor (CIC) Alterations. IJMS *21*, 1818. 10.3390/ijms21051818.

143. Korshunov, A., Ryzhova, M., Jones, D.T.W., Northcott, P.A., van Sluis, P., Volckmann, R., Koster, J., Versteeg, R., Cowdrey, C., Perry, A., et al. (2012). LIN28A immunoreactivity is a

potent diagnostic marker of embryonal tumor with multilayered rosettes (ETMR). Acta Neuropathol *124*, 875–881. 10.1007/s00401-012-1068-3.

144. Rustagi, N., Hampton, O.A., Li, J., Xi, L., Gibbs, R.A., Plon, S.E., Kimmel, M., and Wheeler, D.A. (2016). ITD assembler: an algorithm for internal tandem duplication discovery from short-read sequencing data. BMC Bioinformatics *17*. 10.1186/s12859-016-1031-8.

145. Mohila, C.A., Rauch, R.A., and Adesina, A.M. (2016). Central Neurocytoma and Extraventricular Neurocytoma. In Atlas of Pediatric Brain Tumors (Springer International Publishing), pp. 195–199. 10.1007/978-3-319-33432-5_20.

146. Crotty, T.B., Scheithauer, B.W., Young, W.F., Davis, D.H., Shaw, E.G., Miller, G.M., and Burger, P.C. (1995). Papillary craniopharyngioma: a clinicopathological study of 48 cases. Journal of Neurosurgery *83*, 206–214. 10.3171/jns.1995.83.2.0206.

147. Bunin, G.R., Surawicz, T.S., Witman, P.A., Preston-Martin, S., Davis, F., and Bruner, J.M. (1998). The descriptive epidemiology of craniopharyngioma. Journal of Neurosurgery *89*, 547–551. 10.3171/jns.1998.89.4.0547.

148. Chang, M.T., Bhattarai, T.S., Schram, A.M., Bielski, C.M., Donoghue, M.T.A., Jonsson, P., Chakravarty, D., Phillips, S., Kandoth, C., Penson, A., et al. (2018). Accelerating Discovery of Functional Mutant Alleles in Cancer. Cancer Discovery *8*, 174–183. 10.1158/2159-8290.cd-17-0321.

149. Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., et al. (2015). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol *34*, 155–163. 10.1038/nbt.3391.

150. Harms, K.L., and Chen, X. (2006). The functional domains in p53 family proteins exhibit both common and distinct properties. Cell Death Differ *13*, 890–897. 10.1038/sj.cdd.4401904.

151. Guha, T., and Malkin, D. (2017). Inherited *TP53* Mutations and the Li–Fraumeni Syndrome. Cold Spring Harb Perspect Med *7*, a026187. 10.1101/cshperspect.a026187.

152. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–9. 10.1093/bioinformatics/btp352.

153. Kaplan, E.L., and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association *53*, 457–481. 10.2307/2281868.

154. Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep *50*, 163–70.

155. Cox, D.R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological) *34*, 187–202. 10.1111/j.2517-6161.1972.tb00899.x.

156. Rokita, J.L., and Brown, M. (2022). d3b-center/OpenPBTA-workflows: Release v1.0.4 (Zenodo) 10.5281/zenodo.6968175.

157. Taroni, J., Savonen, C., Krutika Gaonkar, Stephanie, Chante Bethell, Rokita, J.L., Shapiro, J., Greene, C., Yuankun Zhu, Komal Rathi, et al. (2022). AlexsLemonade/OpenPBTA-analysis: Initial Submission (Zenodo) 10.5281/zenodo.7044567.