# The Open Pediatric Cancer Project

This manuscript (<u>permalink</u>) was automatically generated from <u>d3b-center/OpenPedCan-methods@713b5e1</u> on June 13, 2023.

# **Authors**

### • Eric Wafula

### **(D)** 0000-0001-8073-3797

Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

#### Krutika S. Gaonkar

### **(D)** 0000-0003-0838-2405

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Run Jin

### © 0000-0002-8958-9266

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

#### Komal S. Rathi

### © 0000-0001-5534-6904

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Yuankun Zhu

### © 0000-0002-2455-9525

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Bailey K. Farrow

### © 0000-0001-6727-6333

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Daniel P. Miller

### © 0000-0002-2032-4358

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Mariarita Santi

### **1** 0000-0002-6728-3450

Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, 19104, USA

### · Adam A. Kraya

### © 0000-0002-8526-5694

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Xiaoyan Huang

**(D)** 0000-0001-7267-4512

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Bo Zhang

© 0000-0002-0743-5379

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

#### Brian M. Ennis

**(D)** 0000-0002-2653-5009

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### • Ryan J. Corbett

© 0000-0002-3478-0784

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Sharon J. Diskin

© 0000-0002-7200-8939

Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Department of Pediatrics, University of Pennsylvania, Philadelphia, PA, 19104, USA

#### Nicholas Van Kuren

**1** 0000-0002-7414-9516

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Noel Coleman

**(D)** 0000-0001-6454-1285

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### • Christopher Blackden

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### • Jennifer L. Mason

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Saksham Phul

**(D)** 0000-0002-2771-2572

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Miguel A. Brown

**1** 0000-0001-6782-1442

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

#### Alex Sickler

**(D)** 0000-0001-7830-7537

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

### Adam C. Resnick

**(D)** 0000-0003-0436-4189

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA · Funded by 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

# • Jo Lynne Rokita^ <sup>™</sup>

© 0000-0003-2171-3627

Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA · Funded by NCI/NIH Contract No. 75N91019D00024, Task Order No. 75N91020F00003

### Kelsey Keith

© 0000-0002-7451-5117

Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

# **Contact information**

^Lead Contact: Jo Lynne Rokita <a href="mailto:rokita@chop.edu">rokita@chop.edu</a>

⊠Correspondence: Jo Lynne Rokita <u>rokita@chop.edu</u>

### In Brief

# **Highlights**

# Summary

# **Keywords**

# Introduction

# **Results**

# Discussion

# **Acknowledgments**

# **Author Contributions**

Author	Contributions
Eric Wafula	Formal analysis, Software
Krutika S. Gaonkar	Data curation, Formal analysis, Investigation, Methodology, Software, Writing – Original draft, Writing – Review and editing
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
Bailey K. Farrow	Data curation, Software
Daniel P. Miller	Formal analysis
Mariarita Santi	Investigation, Validation, Writing - Review and editing
Adam A. Kraya	Methodology
Xiaoyan Huang	Formal analysis
Bo Zhang	Data curation, Formal analysis
Brian M. Ennis	Data curation, Formal analysis
Ryan J. Corbett	Formal analysis
Sharon J. Diskin	Investigation, Supervision, Validation, Funding acquisition, Writing - Review and editing
Nicholas Van Kuren	Data curation, Software
Noel Coleman	Data curation
Christopher Blackden	Resources
Jennifer L. Mason	Supervision
Saksham Phul	Data curation, Methodology, Formal analysis
Miguel A. Brown	Data curation, Methodology, Formal analysis
Alex Sickler	Methodology, Formal analysis
Adam C. Resnick	Conceptualization, Funding acquisition, Resources, Supervision
Jo Lynne Rokita^	Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Writing – Original draft, Writing - Review and editing
Kelsey Keith	Software, Writing - original draft, API, Formal Analysis, Data Curation, Visualization

# **Declarations of Interest**

# **Figure Titles and Legends**

# **Table Titles and Legends**

## **OPENPEDCAN METHODS**

### **RESOURCE AVAILABILITY**

### **Lead contact**

Requests for access to OpenPedCan 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

Within OpenPedCan (OPC), we harmonized, aggregated, and analyzed data from multiple sources. We harmonized data from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET cohort) Initiative, an NCI-funded collection of disease-specific projects that seeks to identify the genomic changes of pediatric cancers [1]. We included already harmonized neuroblastoma samples from the Gabriella Miller Kids First (GMKF cohort) Pediatric Research Program, a large-scale effort to accelerate research and gene discovery in pediatric cancers and structural birth defects [2]. Additionally, we re-harmonized all samples from the Open Pediatric Brain Tumor Atlas (OpenPBTA, PBTA cohort), an open science initiative led by Alex's Lemonade Stand Foundation Childhood Cancer Data Lab and the Center for Data-Driven Discovery (D3B) at the Children's Hospital of Philadelphia (CHOP), which genomically characterized pediatric brain tumor data from the Children's Brain Tumor Network (CBTN), and the Pacific Pediatric Neuro-oncology Consortium (PNOC) [3,4]. Building on the work of OpenPBTA, OPC added the PBTA X01 data [5], the Chordoma Foundation data [6/], and the MI-ONCOSEQ Study [7], donated to CBTN by the University of Michigan, to the PBTA cohort. Finally, OPC includes the Children's Hospital of Philadelphia (CHOP) P30 Panel data generated by CHOP's Division of Genomic Diagnostics (DGD cohort) which includes fusion panel data [3]. In addition to pediatric cancer data, OpenPedCan contains adult data from large science consortiums as references. For normal gene expression, GTEx [8] was used, and for comparison with adult cancers, The Cancer Genome Atlas (TCGA) [9] was included.

Merged summary files for OpenPedCan v12 are openly accessible in <u>CAVATICA</u> or via download script in the <a href="https://github.com/PediatricOpenTargets/OpenPedCan-analysis">https://github.com/PediatricOpenTargets/OpenPedCan-analysis</a> repository. Cancer group summary data are visible within the NCI's pediatric <a href="Molecular Targets Platform">Molecular Targets Platform</a> and cohort, cancer group, and individual data are visible within <a href="PedcBioPortal">PedcBioPortal</a>

OpenPedCan analysis modules were developed within OpenPBTA [4], modified based on OpenPBTA, or newly created and can be found within the following publicly available repositories. OpenPBTA module analyses can be found at <a href="https://github.com/AlexsLemonade/OpenPBTA-analysis">https://github.com/AlexsLemonade/OpenPBTA-analysis</a>. OpenPedCan module analyses can be found at

https://github.com/PediatricOpenTargets/OpenPedCan-analysis. OpenPedCan api code can be found at https://github.com/PediatricOpenTargets/OpenPedCan-api.

Software versions are documented in **Table XX**.

### **Data releases**

### **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**

### **METHOD DETAILS**

# Nucleic acids extraction and library preparation

## **Data generation**

## **DNA WGS Alignment**

Please refer to the OpenPBTA manuscript for details [4].

# **Quality Control of Sequencing Data**

Please refer to the OpenPBTA manuscript for details [4].

## SNP calling for B-allele Frequency (BAF) generation

Please refer to the OpenPBTA manuscript for details [4].

# **Somatic Mutation Calling**

## SNV and indel calling

Please refer to the OpenPBTA manuscript for details [4].

## **VCF annotation and MAF creation**

**Gather SNV and INDEL Hotspots** 

**Consensus SNV Calling** 

**Somatic Copy Number Variant Calling (WGS samples only)** 

**Consensus CNV Calling** 

Somatic Structural Variant Calling (WGS samples only)

Please refer to the OpenPBTA manuscript for details [4].

# **Methylation Analysis**

# Methylation array preprocessing

We preprocessed raw Illumina 450K and EPIC 850K Infinium Human Methylation Bead Array intensities using the array preprocessing methods implemented in the minfi Bioconductor package [10]. We utilized either preprocessFunnorm when an array dataset had both tumor and normal samples or multiple OpenPedcan-defined cancer\_groups and preprocessQuantile when an array dataset had only tumor samples from a single OpenPedcan-defined cancer\_group to estimate usable methylation measurements (beta-values and m-values) and copy number (cn-values). Some Illumina Infinium array probes targeting CpG loci contain single-nucleotide

polymorphisms (SNPs) near or within the probe [11], which could affect DNA methylation measurements [12]. As the minfi preprocessing workflow recommends, we dropped probes containing common SNPs in dbSNP (minor allele frequency > 1%) at the CpG interrogation or the single nucleotide extensions.

Details of methylation array preprocessing are available in the <u>OpenPedCan methylation-preprocessing module</u>.

## Methylation beta-values summaries

We comprehensively summarized gene-level and isoform-level metrics for the methylation betavalues estimated by array preprocessing to provide insight into the variations in overall genomic DNA methylation levels observed across different pediatric tumors by computing CpG probe-level summary metrics in each cancer group within a cohort, including 1) beta-values quantiles, 2) gene expression (TPM) and methylation (beta-values) correlation, 3) TPM median expression, and 4) transcript representation - a proxy for percent isoform expression in a gene. In addition, each CpG probe was annotated with a gene feature to identify the genomic regions likely involved in regulating gene expression.

Details of the analysis are available in the OpenPedCan methylation-summary module.

## Methylation sample classification

We ran the <u>dkfz's brain classifier version 12.5</u>, a comprehensive DNA methylation-based classification of CNS tumors across all entities and age groups [13]. Unprocessed IDAT-files from the <u>Children's Brain Tumor Network (CBTN)</u> Infinium Human Methylation EPIC (850k) BeadChip arrays were used as input and the following information was compiled into the histologies.tsv file: dkfz\_v12\_methylation\_subclass (predicted methylation subtype), dkfz\_v12\_methylation\_subclass\_score (classification score), dkfz\_v12\_methylation\_mgmt\_status (*MGMT* methylation status), and dkfz\_v12\_methylation\_mgmt\_estimated (estimated *MGMT* methylation fraction).

# **Gene Expression**

### **Abundance Estimation**

### **Gene Expression Matrices with Unique HUGO Symbols**

## **Gene Expression Summary Statistics**

We generated RNA-Seq gene expression (TPM) summary statistics for independent tumor samples from the combined OpenPedCan gene expression matrices, including cancers from pediatric cohorts (PBTA, GMKF, and TARGET) and adult cancers from the TCGA cohort. We grouped selected samples into two groups containing samples from a cancer group in either each cohort or all cohorts, and calculated TPM means, standard deviations, gene-wise z-scores, group-wise z-scores, and ranks for each group as described in the <a href="OpenPedCan rna-seq-expression-summary-stats module">OpenPedCan rna-seq-expression-summary-stats module</a> in detail. The resulting gene-wise and group-wise summary statistics tables were annotated with EFO and MONDO disease codes associated with the cancer groups.

### Gene fusion detection

# **QUANTIFICATION AND STATISTICAL ANALYSIS**

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

Please refer to the OpenPBTA manuscript for details on assignment of copy number status values to CNV segments, cytobands, and genes [4]. We applied criteria to resolve instances of multiple conflicting status calls for the same gene and sample, which are described in detail in the <u>focal-cn-file-preparation</u> module. Briefly, we prioritized 1) non-neutral status calls, 2) calls made from dominant segments with respect to gene overlap, and 3) amplification and deep deletion status calls over gain and loss calls, respectively, when selecting a dominant status call per gene and sample. These methods resolved >99% of duplicated gene-level status calls.

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

Please refer to the OpenPBTA manuscript for details [4].

Fusion prioritization (fusion\_filtering analysis module)

Mutational Signatures (mutational-signatures analysis module)

Tumor Mutation Burden (snv-callers analysis module)

**Clinical Data Harmonization** 

**WHO Classification of Disease Types** 

# **Molecular Subtyping**

Here, we build upon the molecular subtyping performed in OpenPBTA [4].

High-grade gliomas..

Atypical teratoid rhabdoid tumors..

Neuroblastoma tumors...

Integration of brain tumor methylation classifications

TP53 Alteration Annotation (tp53\_nf1\_score analysis module)

Please refer to the OpenPBTA manuscript for details [4].

# Prediction of participants' genetic sex

Please refer to the OpenPBTA manuscript for details [4].

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

For analyses that require all input biospecimens to be independent, we use the OpenPedCan-analysis <u>independent-samples</u> module to select only one biospecimen from each input participant. For each input participant of an analysis, the independent biospecimen is selected based on the analysis-

specific filters and preferences for the biospecimen metadata, such as experimental strategy, cancer group, and tumor descriptor.

# **Supplemental Information Titles and Legends**

# Consortia

# References

- 1. dbGaP Study <a href="https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?">https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?</a> study id=phs000218.v23.p8
- 2. dbGaP Study <a href="https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?">https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?</a> study id=phs001436.v1.p1
- dbGaP Study https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi? 3. study id=phs002517.v1.p1
- **OpenPBTA: The Open Pediatric Brain Tumor Atlas** 4.

Joshua A Shapiro, Krutika S Gaonkar, Stephanie J Spielman, Candace L Savonen, Chante J Bethell, Run Jin, Komal S Rathi, Yuankun Zhu, Laura E Egolf, Bailey K Farrow, ... Jaclyn N Taroni Cell Genomics (2023-05) https://doi.org/gr92p6

DOI: 10.1016/j.xgen.2023.100340

- FY21 X01 Projects for the Gabriella Miller Kids First Program (2021-09-02) 5. https://commonfund.nih.gov/kidsfirst/2021x01projects
- 6. Home

Chordoma Foundation https://www.chordomafoundation.org/

- 7. Michigan Center for Translational Pathology <a href="https://mctp.med.umich.edu">https://mctp.med.umich.edu</a>
- dbGaP Study <a href="https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?">https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?</a> 8. study\_id=phs000424.v9.p2
- 9. The Cancer Genome Atlas Program (TCGA) - NCI (2022-05-13) https://www.cancer.gov/ccg/research/genome-sequencing/tcga
- 10. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi

Jean-Philippe Fortin, Timothy J Triche Jr, Kasper D Hansen

Bioinformatics (2016-11-29) https://doi.org/f9x7kd

DOI: <u>10.1093/bioinformatics/btw691</u> · PMID: <u>28035024</u> · PMCID: <u>PMC5408810</u>

Review of processing and analysis methods for DNA methylation array data 11.

CS Wilhelm-Benartzi, DC Koestler, MR Karagas, JM Flanagan, BC Christensen, KT Kelsey, CJ Marsit, EA Houseman, R Brown

British Journal of Cancer (2013-08-27) https://doi.org/gb9qvv

DOI: 10.1038/bjc.2013.496 · PMID: 23982603 · PMCID: PMC3777004

Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 12. BeadChip Array: implications for comparative population studies

Patrycja Daca-Roszak, Aleksandra Pfeifer, Jadwiga Żebracka-Gala, Dagmara Rusinek, Aleksandra Szybińska, Barbara Jarząb, Michał Witt, Ewa Ziętkiewicz

BMC Genomics (2015-11-25) https://doi.org/gb3h5r

DOI: 10.1186/s12864-015-2202-0 · PMID: 26607064 · PMCID: PMC4659175

13. DNA methylation-based classification of central nervous system tumours

David Capper, David TW Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, Felix Sahm, Lukas Chavez, David E Reuss, ... Stefan M Pfister Nature (2018-03-14) https://doi.org/gc5t36

DOI: <u>10.1038/nature26000</u> · PMID: <u>29539639</u> · PMCID: <u>PMC6093218</u>