# The Open Pediatric Cancer Project

This manuscript (<u>permalink</u>) was automatically generated from <u>d3b-center/OpenPedCan-methods@bbbb01b</u> on June 27, 2023.

# **Authors**

### Eric Wafula

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

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

### Sangeeta Shukla

© 0000-0002-3727-9602

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

### · Aditya Lahiri

© 0000-0001-9352-1312

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

### Krutika S. Gaonkar

© 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

**D** 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

**(D)** 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

**(D)** 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

© 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

# **D** 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

### Chuwei Zhong

# © 0000-0003-2406-2735

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

# Zhuangzhuang Geng

## **D** 0009-0007-6883-0691

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

### © 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

# **(D)** <u>0000-0002-3478-0784</u>

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

### **(D)** 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

### **(D)** 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

### Matthew R. Lueder

### 🔟 0009-0002-7370-102X

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 Pathology and Laboratory Medicine, 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

### Kelsey Keith

**(iii)** 0000-0002-7451-5117

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

### Dave Hill

© 0000-0002-1337-1789

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

### • Asif T Chinwalla

© 0000-0001-7831-3996

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

### Yuanchao Zhang

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

### • Adam C. Resnick

© 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

### Alvin Farrel

**(D)** 0000-0003-1087-9840

Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA · Funded by NCI/NIH Contract No. 75N91019D00024, Task Order No. 75N91020F00003

### Deanne Taylor

© 0000-0002-3302-4610

Department of Bioinformatics and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA; Department of Pediatrics, University of Pennsylvania Perelman Medical School, Philadelphia, PA, 19104, USA · Funded by NCI/NIH Contract No. 75N91019D00024, Task Order No. 75N91020F00003

# • Jo Lynne Rokita^ <sup>™</sup>

**1** 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

# **Contact information**

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

# In Brief Highlights Summary Keywords Introduction Results

# **Discussion**

# **Acknowledgments**

# **Author Contributions**

Author	Contributions
Eric Wafula	Formal analysis, Software
Sangeeta Shukla	Data curation, Formal analysis, Investigation, Methodology, Software, Writing – Original draft, Writing - Review and editing
Aditya Lahiri	Data curation, Formal analysis, Investigation, Methodology, Software, Writing – Original draft, Writing - Review and editing
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
Chuwei Zhong	Formal analysis
Zhuangzhuang Geng	Data curation, Formal analysis, Investigation, Methodology, Software, Writing – Original draft
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
Matthew R. Lueder	Data curation
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
Kelsey Keith	Software, Writing - original draft, API, Formal Analysis, Data Curation, Visualization
Dave Hill	Software, Writing - original draft, API, Data Curation
Asif T Chinwalla	Project Administration, Supervision, Methodology, Investigation, Validation, API

Author	Contributions
Yuanchao Zhang	API, Software, Formal Analysis, Writing - original draft
Adam C. Resnick	Conceptualization, Funding acquisition, Resources, Supervision
Alvin Farrel	Supervision, Investigation, Visualization, Methodology, Funding acquisition
Deanne Taylor	Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Writing – Original draft, Writing - Review and editing
Jo Lynne Rokita^	Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Writing – Original draft, Writing - Review and editing

# **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-data.sh 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.

All original code was developed within the following modules in the OpenPedCan analyses repository as listed below. Links to the modules are available here, and within each module is a detailed README that describes the purpose and intended usage of the scripts, along with pointers to the results from the data those scripts process.

List of OpenPedCan Analyses Modules
<u>chromosomal-instability</u>
<u>cnv-frequencies</u>
<u>collapse-rnaseq</u>
<u>compare-gistic</u>
copy number consensus call
<u>create-subset-files</u>
data-pre-release-qc
efo-mondo-mapping
<u>filter-mtp-tables</u>
focal-cn-file-preparation
<u>fusion-frequencies</u>
<u>fusion-summary</u>
<u>fusion filtering</u>
gene-set-enrichment-analysis
gene match
<u>immune-deconv</u>
<u>independent-samples</u>
long-format-table-utils
methylation-preprocessing

List of OpenPedCan Analyses Modules
methylation-summary
molecular-subtyping-ATRT
molecular-subtyping-CRANIO
molecular-subtyping-EPN
molecular-subtyping-EWS
molecular-subtyping-HGG
molecular-subtyping-LGAT
molecular-subtyping-MB
molecular-subtyping-NBL
molecular-subtyping-chordoma
molecular-subtyping-embryonal
molecular-subtyping-integrate
molecular-subtyping-neurocytoma
molecular-subtyping-pathology
<u>mtp-annotations</u>
mtp-tables-qc-checks
<u>mutational-signatures</u>
pedcbio-cnv-prepare
pedcbio-sample-name
pedot-table-column-display-order-name
<u>rna-seq-expression-summary-stats</u>
<u>rnaseq-batch-correct</u>
<u>run-gistic</u>
<u>snv-frequencies</u>
<u>tmb-calculation</u>
<u>tp53 nf1 score</u>
<u>tumor-gtex-plots</u>
tumor-normal-differential-expression

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

# **Data releases**

We maintained a data release folder on Amazon S3, downloadable directly from S3 or our open-access CAVATICA project, with merged files for each analysis (See data and code availability section). As we produced new results that we expected to be used across multiple analyses, or identified data issues, we created new data releases in a versioned manner.

# **METHOD DETAILS**

# Nucleic acids extraction and library preparation

For the PBTA X01 cohort, libraries were prepped using the Illumina TruSeq Strand-Specific Protocol to pull out poly-adenylated transcripts.

**cDNA Library Construction** Total RNA was quantified using the Quant-iT™ RiboGreen® RNA Assay Kit and normalized to 5ng/ul. Following plating, 2 uL of ERCC controls (using a 1:1000 dilution) were spiked into each sample. An aliquot of 325 ng for each sample was transferred into library preparation. The resultant 400bp cDNA went through dual-indexed library preparation: 'A' base addition, adapter ligation using P7 adapters, and PCR enrichment using P5 adapters. After enrichment, the libraries were quantified using Quant-iT PicoGreen (1:200 dilution). Samples were normalized to 5 ng/uL. The sample set was pooled and quantified using the KAPA Library Quantification Kit for Illumina Sequencing Platforms.

# **Data generation**

**PBTA X01 Illumina Sequencing** Pooled libraries were normalized to 2nM and denatured using 0.1 N NaOH prior to sequencing. Flowcell cluster amplification and sequencing were performed according to the manufacturer's protocols using the NovaSeq 6000. Each run was a 151bp paired-end with an eight-base index barcode read. Data was analyzed using the Broad Picard Pipeline which includes demultiplexing and data aggregation.

# **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 and INDEL Calling**

We used the same mutation calling methods as described in OpenPBTA manuscript for details [4].

### **VCF** annotation and MAF creation

Somatic variants were annotated by the Ensembl Variant Effect Predictor (VEP v.105) [10].

# **Gather SNV and INDEL Hotspots**

# **Consensus SNV Calling**

We adopted the consensus SNV calling method described in OpenPBTA manuscript with adjustment [4]. For SNV calling, we combined four consensus SNV calling algorithms, including Strelka2[11], Mutect2[12], Lancet[13], and VarDict[doi? 10.1093/nar/gkw227]. Strelka2 outputs multi-nucleotide polymorphisms (MNPs) as consecutive single-nucleotide polymorphisms. In order preserve MNPs, we gather MNP calls from the other caller inputs, and search for evidence supporting these consecutive

SNP calls as MNP candidates. Once found, the Strelka2 SNP calls supporting a MNP are converted to a single MNP call. This is done to preserve the predicted gene model as accurately as possible in our consensus calls. Consensus SNV from all four callers were collected and by default, calls that were detected in at least two calling algorithms or marked as "HotSpotAllele" were retained. Potential non-hotspot germline variants were removed if they had a normal depth <= 7 and gnomAD allele frequency > 0.001. Final results were saved in MAF format.

# **Somatic Copy Number Variant Calling**

# **Consensus CNV Calling**

We adopted the consensus CNV calling described in OpenPBTA manuscript [doi:10.1016/j.xgen.2023.100340] with minor adjustments. For each caller and sample with WGS performed, we called CNVs based on consensus among Control-FREEC ([14]; [15]), CNVkit ([doi? 10.1371/journal.pcbi.1004873]), and GATK ([doi? 10.1101/gr.107524.110]). Sample and consensus caller files with more than 2,500 CNVs were removed to de-noise and increase data quality, based on cutoffs used in GISTIC ([16]). For each sample, we included the following 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 were covered by another caller. For GATK, if a panel of normal was not able to be created (required 30 male and 30 female with the same sequencing platform), consensus was not run for tumors with WGS performed on that sequencing platform. 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. Any CNVs that overlapped 50% or more with immunoglobulin, telomeric, centromeric, segment duplicated regions, or that were shorter than 3000 bp were filtered out. The CNVKit calls for WXS samples were appended to the consensus CNV file.

# 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 [17]. 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 [18], which could affect DNA methylation measurements [19]. 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 <a href="dkfz's brain classifier version 12.5">dkfz's brain classifier version 12.5</a>, a comprehensive DNA methylation-based classification of CNS tumors across all entities and age groups [20]. Unprocessed IDAT-files from the <a href="Children's Brain Tumor Network">Children's Brain Tumor Network (CBTN)</a>. Infinium Human Methylation EPIC (850k) BeadChip arrays were used as input and the following information was compiled into the histologies.tsv file: <a href="dkfz\_v12\_methylation\_subclass">dkfz\_v12\_methylation\_subclass</a> (predicted methylation subtype), <a href="dkfz\_v12\_methylation\_mgmt\_status">dkfz\_v12\_methylation\_mgmt\_status</a> (MGMT methylation status), and <a href="dkfz\_v12\_methylation\_mgmt\_estimated">dkfz\_v12\_methylation\_mgmt\_estimated</a> (estimated MGMT methylation fraction).

# **Gene Expression**

The <a href="tumor-normal-differential-expression">tumor-normal-differential-expression</a> module perfoms differential expression analyses for all sets of Disease (cancer\_group) and Dataset (cohort) across all genes found in the gene-expression-rsem-tpm-collapsed.rds table. The purpose of this analysis is to highlight the correlation and understand the variability in gene expression in different cancer conditions across different histological tissues. For OpenPedCan v12 data release, this module performs expression analysis over 102 cancer groups across 52 histological tissues for all 54,346 genes found in the dataset. This analysis was performed on the Children's Hospital of Philadelphia HPC and was configured to use 96G of RAM per CPU, with one task (one iteration of expression analysis for each set of tissue and cancer group) per CPU (total 102\*52=5304 CPUs) using the <a href="RVDESeq2">RVDESeq2</a> package. Please refer to script run-tumor-normal-differential-expression.sh in the module for additional details on Slurm processing configuration. The same analysis can also be performed on CAVATICA, but requires further optimization. The module describes the steps for CAVATICA set up, and scripts to publish an application on the portal. The required data files are also available publicly on CAVATICA under the <a href="Open Pediatric Cancer">Open Pediatric Cancer</a> (OpenPedCan) Open Access. Refer to the module for detailed description and scripts.

# **Abundance Estimation**

Among the data sources used for OpenPedCan, GTEx and TCGA used GENCODE versions v26 and v36, respectively. Moreover, the gene symbols used in these different GENCODE versions also varied. Therefore, the gene symbols had to be harmonized for compatibility to map unique gene identifiers to their gene symbols. ENSG IDs from each data source were pulled and mapped to the GTF/GFF3 file from GENCODE v39 to extract unique gene symbols and remove duplicates. Additionally, the gene expression matrices had some instances where multiple Ensembl gene identifiers mapped to the same gene symbol. This was dealt with by filtering the expression matrix to only genes with [FPKM/TPM] > 0 and then selecting the instance of the gene symbol with the maximum mean [FPKM/TPM/Expected\_count] value across samples. This enabled many downstream modules that

require RNA-seq data have gene symbols as unique gene identifiers. Refer to <u>collapse-rnaseq</u> module for scripts and details.

# **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)

The <u>fusion\_filtering</u> module filters artifacts and annotates fusion calls, with prioritization for oncogenic fusions, for the fusion calls from STAR-Fusion and Arriba. After artifact filtering, fusions were prioritized and annotated as "putative oncogenic fusions" when at least one gene was a known kinase, oncogene, tumor suppressor, curated transcription factor, on the COSMIC Cancer Gene Census List, or observed in TCGA. Fusions were retained in this module if they were called by both callers, recurrent or specific to a cancer group, or annotated as a putative oncogenic fusion. Please refer to the module linked above for more detailed documentation and scripts.

# Mutational Signatures (mutational-signatures analysis module)

# Tumor Mutation Burden [TMB] (tmb-calculation analysis module)

Recent clinical studies have associated high TMB with improved patient response rates and survival benefit from immune checkpoint inhibitors [doi.org/10.1002/gcc.22733?].

The <u>Tumor Mutation Burden (TMB)</u> <u>tmb-calculation</u> module was adapted from the <u>snv-callers module</u> of the OpenPBTA project [4]. Here, we use mutations in the <u>snv-consensus-</u>

Workflow and is included in the OpenPedCan data download. The consensus MAF contains SNVs or MNVs called in at least 2 of the 4 callers (Mutect2, Strelka2, Lancet, and Vardict) plus hotspot mutations if called in 1 of the 4 callers. We calculated TMB for tumor samples sequenced with either WGS or WXS. Briefly, we split the SNV consensus MAF into SNVs and multinucleotide variants (MNVs). We split the MNV subset into SNV calls, merged those back with the SNVs subset, and then removed sample-specific redundant calls. The resulting merged and non-redundant SNV consensus calls were used as input for the TMB calculation. We tallied only nonsynonymous variants with classifications of high/moderate consequence ("Missense\_Mutation", "Frame\_Shift\_Del", "In\_Frame\_Ins", "Frame\_Shift\_Ins", "Splice\_Site", "Nonsense\_Mutation", "In\_Frame\_Del", "Nonstop\_Mutation", and "Translation\_Start\_Site") for the numerator.

All mutation TMB For WGS samples, we calculated the size of the genome covered as the intersection of Strelka2 and Mutect2's effectively surveyed areas, regions common to all variant callers, and used this as the denominator. WGS\_all\_mutations\_TMB = (total # mutations in consensus MAF) / intersection\_strelka\_mutect\_vardict\_genome\_size For WXS samples, we used the size of the WXS bed region file as the denominator. WXS\_all\_mutations\_TMB = (total # mutations in consensus MAF)) / wxs\_genome\_size

Coding only TMB We generated coding only TMB from the consensus MAF as well. We calculated the intersection for Strelka2 and Mutect2 surveyed regions using the coding sequence ranges in the GENCODE v39 gtf supplied in the OpenPedCan data download. We removed SNVs outside of these coding sequences prior to implementing the TMB calculation below: WGS\_coding\_only\_TMB = (total # coding mutations in consensus MAF) / intersection\_wgs\_strelka\_mutect\_vardict\_CDS\_genome\_size For WXS samples, we intersected each WXS bed region file with the GENCODE v39 coding sequence, sum only variants within this region for the numerator, and calculate the size of this region as the denominator. WXS\_coding\_only\_TMB = (total # coding mutations in consensus MAF) / intersection\_wxs\_CDS\_genome\_size

Finally, we include an option (nonsynfilter\_focr) to use specific nonsynonymous mutation variant classifications recommended from the TMB Harmonization Project.

### **Clinical Data Harmonization**

# **WHO Classification of Disease Types**

# **Molecular Subtyping**

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

# High-grade gliomas.

A new high-grade glioma entity called infant-type hemispheric gliomas (IHGs), characterized by distinct gene fusions enriched in receptor tyrosine kinase (RTK) genes including *ALK*, *NTRK1/2/3*, *ROS1* or *MET*, was identified in 2021 [doi? 10.1038/s41467-019-12187-5]. To identify IHG tumors, first, tumors which were classified as "IHG" by the DKFZ methylation classifier or diagnosed as "infant type hemispheric glioma" from pathology\_free\_text\_diagnosis were selected [20]. Then, the corresponding tumor RNA-seq data were utilized to seek the evidence for RTK gene fusion. Based on the specific RTK gene fusion present in the samples, IHGs were further classified as "IHG, ALK-altered", "IHG, NTRK-altered", "IHG, ROS1-altered", or "IHG, MET-altered". If no fusion was observed, the samples were identified as "IHG, To be classified".

# Atypical teratoid rhabdoid tumors.

Atypical teratoid rhabdoid tumors (ATRT) tumors were categorized into three subtypes: "ATRT, MYC", "ATRT, SHH", and "ATRT, TYR" [21]. In OpenPedCan, the molecular subtyping of ATRT was based solely on the DNA methylation data. Briefly, ATRT samples with a high confidence DKFZ methylation subclass score (>= 0.8) were selected and subtypes were assigned based on the DKFZ methylation subclass [doi10.1038/nature26000?]. Samples with low confidence DKFZ methylation subclass scores (< 0.8) were identified as "ATRT, To be classified".

### Neuroblastoma tumors.

Neuroblastoma (NBL) tumors with a pathology diagnosis of neuroblastoma, ganglioneuroblastoma, or ganglioneuroma were subtyped based on their MYCN copy number status as either "NBL, MYCN amplified" or "NBL, MYCN non-amplified". If pathology\_free\_text\_diagnosis was "NBL, MYCN non-amplified" and the genetic data suggested MYCN amplification, the samples were subtyped as "NBL, MYCN amplified". On the other hand, if pathology\_free\_text\_diagnosis was "NBL, MYCN amplified" and the genetic data suggested MYCN non-amplification, the RNA-Seq gene expression level of MYCN was used as a prediction indicator. In those cases, samples with MYCN gene expression above or below the cutoff (TPM >= 140.83 based on visual inspection of MYCN CNV status) were subtyped as "NBL, MYCN amplified" and "NBL, MYCN non-amplified", respectively. MYCN gene expression was also used to subtype samples without DNA sequencing data. If a sample did not fit none of these situations, it was denoted as "NBL, To be classified".

# 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 <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> 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>
- 8. 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=phs000424.v9.p2
- 9. The Cancer Genome Atlas Program (TCGA) - NCI (2022-05-13) https://www.cancer.gov/ccg/research/genome-sequencing/tcga
- 10. The Ensembl Variant Effect Predictor

William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, Fiona Cunningham

Genome Biology (2016-06-06) https://doi.org/gdz75c

DOI: 10.1186/s13059-016-0974-4 · PMID: 27268795 · PMCID: PMC4893825

Strelka2: fast and accurate calling of germline and somatic variants 11.

Sangtae Kim, Konrad Scheffler, Aaron L Halpern, Mitchell A Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, Christopher T Saunders Nature Methods (2018-07-16) https://doi.org/gdwrp4

DOI: 10.1038/s41592-018-0051-x · PMID: 30013048

12. **Calling Somatic SNVs and Indels with Mutect2** 

> David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, Lee Lichtenstein Cold Spring Harbor Laboratory (2019-12-02) https://doi.org/ggntwv

DOI: 10.1101/861054

13. Genome-wide somatic variant calling using localized colored de Bruijn graphs Giuseppe Narzisi, André Corvelo, Kanika Arora, Ewa A Bergmann, Minita Shah, Rajeeva Musunuri, Anne-Katrin Emde, Nicolas Robine, Vladimir Vacic, Michael C Zody Communications Biology (2018-03-22) https://doi.org/gfcfr8

DOI: 10.1038/s42003-018-0023-9 · PMID: 30271907 · PMCID: PMC6123722

- 14. <a href="https://academic.oup.com/bioinformatics/article/27/2/268/285534?login=false">https://academic.oup.com/bioinformatics/article/27/2/268/285534?login=false</a>
- 15. <a href="https://academic.oup.com/bioinformatics/article/28/3/423/189142">https://academic.oup.com/bioinformatics/article/28/3/423/189142</a>

# 16. **GISTIC2.0** facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers

Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, Gad Getz

Genome Biology (2011-04-28) https://doi.org/10.1186/gb-2011-12-4-r41

DOI: 10.1186/gb-2011-12-4-r41

# 17. 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: 10.1093/bioinformatics/btw691 · PMID: 28035024 · PMCID: PMC5408810

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

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: <u>10.1038/bjc.2013.496</u> · PMID: <u>23982603</u> · PMCID: <u>PMC3777004</u>

# 19. Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 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

# 20. 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) <a href="https://doi.org/gc5t36">https://doi.org/gc5t36</a>

DOI: 10.1038/nature26000 · PMID: 29539639 · PMCID: PMC6093218

# 21. Molecular subgrouping of atypical teratoid/rhabdoid tumors—a reinvestigation and current consensus

Ben Ho, Pascal D Johann, Yura Grabovska, Mamy Jean De Dieu Andrianteranagna, Fupan Yao, Michael Frühwald, Martin Hasselblatt, Franck Bourdeaut, Daniel Williamson, Annie Huang, Marcel Kool

Neuro-Oncology (2019-12-31) https://doi.org/gn3kcm

DOI: 10.1093/neuonc/noz235 · PMID: 31889194 · PMCID: PMC7229260