**Supplementary Information**

**St. Jude Cloud**—**a Pediatric Cancer Genomic Data Sharing Ecosystem**

Clay McLeod1\*, Alexander M. Gout1\*, Xin Zhou1, Andrew Thrasher1, Delaram Rahbarinia1, Samuel W. Brady1, Michael Macias1, Kirby Birch1, David Finkelstein1, Jobin Sunny1, Rahul Mudunuri1, Brent A. Orr2, Madison Treadway1, Bob Davidson3, Tracy K. Ard3, Arthur Chiao1, Andrew Swistak1, Stephanie Wiggins1, Scott Foy1, Jian Wang1, Edgar Sioson1, Shuoguo Wang1, J. Robert Michael1, Yu Liu1, Xiaotu Ma1, Aman Patel1, Michael N. Edmonson1, Mark R. Wilkinson1, Andrew M. Frantz1, Ti-Cheng Chang1, Liqing Tian1, Shaohua Lei1, S. M. Ashiqul Islam4, Christopher Meyer5, Naina Thangaraj5, Pamella Tater5, Vijay Kandali5, Singer Ma5, Tuan Nguyen5, Omar Serang5, Irina McGuire6, Nedra Robison6, Darrell Gentry6, Xing Tang7, Lance E. Palmer7, Gang Wu1, Ed Suh6, Leigh Tanner6, James McMurry6, Matthew Lear2, Alberto S. Pappo8, Zhaoming Wang1,9, Carmen L. Wilson9, Yong Cheng7, Soheil Meshinchi10, Ludmil B. Alexandrov4, Mitch Weiss7, Gregory T. Armstrong9, Leslie L. Robison9, Yutaka Yasui9, Kim E. Nichols8, David W. Ellison2, Chaitanya Bangur3, Charles G. Mullighan2, Suzanne J. Baker11, Michael A. Dyer11, Geralyn Miller3, Scott Newman1, Michael Rusch1, Richard Daly5, Keith Perry6$, James R. Downing2$ and Jinghui Zhang1$

**AFFILIATIONS:**

1. Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, 38105, Tennessee, USA

2. Department of Pathology, St. Jude Children's Research Hospital, Memphis, 38105, Tennessee, USA

3. Microsoft Research, Redmond, 98052, Washington, USA

4. Department of Cellular and Molecular Medicine and Department of Bioengineering, Moores Cancer Center, University of California San Diego, La Jolla, 92093, California, USA

5. DNAnexus, Mountain View, 94040, California, USA

6. Department of Information Services, St. Jude Children's Research Hospital, Memphis, 38105, Tennessee, USA

7. Department of Hematology, St. Jude Children's Research Hospital, Memphis, 38105, Tennessee, USA

8. Department of Oncology, St. Jude Children's Research Hospital, Memphis, 38105, Tennessee, USA

9. Department of Epidemiology & Cancer Control, St. Jude Children's Research Hospital, Memphis, 38105, Tennessee, USA

10. Fred Hutchinson Cancer Research Center Professor of Pediatrics, University of Washington School of Medicine, Seattle, 98109, Washington, USA

11. Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, 38105, Tennessee, USA

\* Contributed Equally

$ Correspondence addressed to Keith Perry ([keith.perry@stjude.org](mailto:keith.perry@stjude.org)), James R. Downing ([james.downing@stjude.org](mailto:james.downing@stjude.org)) and Jinghui Zhang ([jinghui.zhang@stjude.org](mailto:jinghui.zhang@stjude.org))

# SUPPLEMENTARY TABLES

Supplementary Table S1. Data Sets in PeCan Knowledgebase.

Supplementary Table S2. **(A)** Gene set enrichment analysis following differential gene expression analysis between two adamantinomatous craniopharyngioma groups (ACPG group1 and ACPG group 2) identified in Figure 4C. Table displays Human Gene Atlas gene groups over-represented in upregulated genes in ACPG group 1 following ENRICHR analysis. **(B)** Gene set enrichment analysis following differential gene expression analysis between ACPG group1 and ACPG group 2. Table displays WikiPathways 2019 Human gene groups over-represented in upregulated genes in ACPG group 1 following ENRICHR analysis. **(C)** Gene set enrichment analysis following differential gene expression analysis between ACPG group1 and ACPG group 2. Table displays GO Molecular Function 2018 gene groups over-represented in upregulated genes in ACPG group 2 following ENRICHR analysis. **(D)** Gene set enrichment analysis following differential gene expression analysis between metastatic osteosarcomas (circled in Figure 4B) and other and osteosarcomas. Table displays Human Gene Atlas gene groups over-represented in upregulated genes in metastatic osteosarcomas following ENRICHR analysis.

Supplementary Table S3. Description of pediatric cancer subtypes in the categories of blood (hematologic malignancy), solid, brain, and germ cell tumors available on St. Jude Cloud. Inclusion of the subtype in the t-SNE analysis presented in Figure 4A-C is indicated. Where applicable, categorization of the subtypes for the mutation burden and COSMIC mutation signature analysis presented in Figure 5A is also indicated.

# SUPPLEMENTARY FIGURES

**Diagram

Description automatically generated**

Supplementary Figure S1**.** Workflow for monthly deposition of Real-time Clinical Genomics (RTCG) data onto St. Jude Cloud. For WGS, WES and RNA-seq data generated from our CLIA/CAP-accredited laboratory, a check for informed consent for research use of patient data is performed following an embargo period (30 days) prior to transfer to the research computing environment. Following a further embargo period (90 days) the genomic data will be uploaded to the St. Jude Cloud with data harmonization performed and data quality assessed prior to release on the Genomics Platform of St. Jude Cloud.

Map

Description automatically generatedSupplementary Figure S2**.** Data sharing with the global research community. The number of approved data requests per institution throughout the world are indicated by color. The institutions granted access to clinical genomics (ClinGen) data are indicated with a red dot.

**A picture containing diagram

Description automatically generated**

Supplementary Figure S3. Analysis of two groups of adamantinomatous craniopharyngioma (ACPG) highlighted in Figure 4C. **(A)** Volcano plot displaying differential gene expression (DGE) analysis between ACPG Group 1 and 2. Here DGE analysis shows an over-representation of oligodendrocytic (blue), prefrontal cortex (yellow), and fetal brain (red) genes in ACPG Group 1, while overrepresentation of matrix metallopeptidase (MMP) genes (green) is seen in ACPG Group 2. **(B)** Somatic mutation status and demographic information of the ACPG samples in ACPG Group 1 and 2. CTNNB1 mutational variant allele frequencies (VAF) were determined from whole-genome (WGS) or whole-exome (WES) sequencing data. Where CTNNB1 VAFs were calculated from both WGS and WES, a range is presented. **(C)** Boxplots of Keratin (KRT7, KRT8, and KRT14) gene expression observed in ACPG Group 1 and 2 samples. Note p-values calculated using a t-test. **(D)** Boxplots of select absolute gene expression observed in ACPG group 1 (top) and 2 (middle), in addition to reported DGE for ACPG samples when compared to normal fetal brain tissue (1) (bottom). **(E)** Hematoxylin and Eosin stained sections from ACPG Group 1 and 2 samples at 10X and 20X magnification. All five Group 1 tumors were characterized by brain invasion. The surrounding reactive brain parenchyma is designated by an asterisk (\*). In contrast, Group 2 tumors showed involvement with either tumor cyst walls (“CW”) or meninges (“M”), with minimal or no involvement with neural tissue. Scale bar represents 160 uM for the 10X images and 80 uM for the 20X images.

Chart, scatter chart

Description automatically generated Supplementary Figure S4. Differential gene expression (DGE) analysis of the four metastatic osteosarcoma samples highlighted in Figure 4B versus other osteosarcoma samples. Volcano plot depicting DGE analysis between these two osteosarcoma groups showing an over-representation of Human Gene Atlas ‘Lung’ genes expression (blue) in the four metastatic osteosarcoma samples.

Diagram

Description automatically generated Supplementary Figure S5. St. Jude Cloud RNA-Seq Expression Analysis workflow enabling clustering user-supplied RNA-Seq data with a St. Jude Cloud pediatric cancer cohort via an interactive two-dimensional t-SNE plot. **(A)** User interface for selecting user sample(s), parameters for analysis and pediatric cancer type (which can be one of ‘blood’ (selected in this example), ‘solid’, ‘brain’, or ‘all’ reference cohort). If the input is a BAM file, realignment to hg38 using our inhouse RNA-Seq alignment protocol is used to generate gene counts. The user gene counts are then collated with the reference gene count data and expression analysis is performed. The genes used for the t-SNE analysis are available as a ‘gene\_list.txt’ and t-SNE plot as ‘PAWNXH.html’ (PAWNXH is the COG sample used as an example for use case 1). **(B)**t-SNE plot of blood cancer samples where the user sample (i.e. PAWNXH) is clustered with the St. Jude Cloud AML samples. A user may Pan/Zoom the t-SNE plot, query a sample of interest, and highlight subtypes of interest by clicking on the cancer subtype label on the right. The current display highlighted 1) samples of *KMT2A* rearrangements in all subtypes of blood cancers, 2) AML Core Binding Factor and 3) B-ALLs with *NUTM1* fusions. Mouse over a tumor sample displays the associated metadata for the sample, in this instance SJINF049\_D, a B-ALL with *BRD9-NUTM1* fusion.

Chart, bar chart

Description automatically generated

Supplementary Figure S6. Representative mutational signatures in pediatric cancers. **(A)** APOBEC (COSMIC mutation signature SBS2 and SBS13) signatures identified in pediatric acute megakaryoblastic leukemia samples. **(B)** UV-light (COSMIC mutation signature SBS7a, SBS7b, SBS7c, and SBS7d) associated mutational signatures detected in a B-cell Lymphoblastic Leukemia (B-ALL) subtype with intrachromosomal amplification of chromosome 21 (iAMP21). UV-signature positive samples show an increased mutational burden (right). The key below each plot indicates the COSMIC mutation signatures identified, and mutation burden (total number of SNVs) for each sample represented on the right side barplot.

Chart, waterfall chart

Description automatically generated

Supplementary Figure S7. Identification of COSMIC mutation signature SBS22 in pediatric hepatoblastoma samples. **(A)** Whole-genome sequencing somatic mutation signature landscape of hepatoblastoma patient diagnosis and relapse samples (SJST030137\_D1 and SJST030137\_R1) harboring Aristolochic Acid COSMIC mutation signature SBS22. **(B)** Stacked barplot of somatic mutations (No./Mb) contributing to COSMIC mutation signature SBS1, SBS5, SBS8, SBS18, SBS22, and SBS35 within six hepatoblastoma patient samples. Patient SJST030137 involves both diagnostic and relapse samples (SJST030137\_D1 and SJST030137\_R1), while remaining patients only involve diagnostic samples. Patient SJST030137 is an Asian male while the remaining samples are from non-Asian patients. Data from published (SJST030137\_D1, SJST030246\_D1, https://clinicaltrials.gov/ct2/show/NCT02530658) and unpublished real-time clinical genomics (SJST030137\_R1, SJST030609\_D1, SJST030953\_D1, SJST031109\_D1) datasets.

Graphical user interface, text

Description automatically generated

Supplementary Figure S8. Identification of therapy-related COSMIC mutation signature SBS86 and SBS87 in pediatric tumor samples. **(A)** Whole-genome sequencing somatic mutation signature landscape of relapsed B-cell acute lymphoblastic leukemia sample harboring therapy-related COSMIC mutational signature SBS86. **(B)** Whole-genome sequencing somatic mutation signature landscape of relapsed B-cell acute lymphoblastic leukemia sample harboring therapy-related COSMIC mutational signature SBS87.

Graphical user interface, chart, bar chart

Description automatically generated Supplementary Figure S9. Visualization of output from St. Jude Cloud Mutational Signatures workflow. Summary views of mutation signatures present within a “reference” cohort and identified in a “query cohort” are shown at the top while a per sample view of mutation signature and mutation burden in each cancer genome is shown at the bottom. A user can toggle between the “reference” or the “query” summary to select a cohort of interest for the per sample view. In this display, the query cohort consists of the nine ICGC adult AML samples from “use case 2” while the reference cohort selected is the pediatric AML cohort (top right) on St. Jude Cloud. The sample-level mutation signatures are shown at the bottom is presented for the reference cohort. A user can mouse over each of the stacked bars in order to ascertain the represented mutation signature. Clicking on/selecting any one of the detected mutation signatures in the legend automatically links to Sanger Center’s COSMIC website displaying further information about the particular mutation signature.

# REFERENCES

1. Apps JR, Carreno G, Gonzalez-Meljem JM, Haston S, Guiho R, Cooper JE*, et al.* Tumour compartment transcriptomics demonstrates the activation of inflammatory and odontogenic programmes in human adamantinomatous craniopharyngioma and identifies the MAPK/ERK pathway as a novel therapeutic target. Acta Neuropathol **2018**;135(5):757-77 doi 10.1007/s00401-018-1830-2.