An Open Pediatric Brain Tumor Atlas

This manuscript (permalink) was automatically generated from AlexsLemonade/OpenPBTA-manuscript@dd26041 on September 29, 2019.

Authors

- **John Doe**
  XXXX-XXXX-XXXX-XXXX · johndoe · johndoe
  Department of Something, University of Whatever · Funded by Grant XXXXXXXX

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

Introduction

Introduction will go here.

Materials and Methods

Biospecimen collection

The Pediatric Brain Tumor Atlas specimens are comprised of samples from Children's Brain Tumor Tissue Consortium (CBTTC) and the Pediatric Pacific Neuro-oncology Consortium (PNOC).

Children's Brain Tumor Tissue Consortium (CBTTC)

The CBTTC [1] is a collaborative, multi-institutional (16 institutions worldwide) research program dedicated to the study of childhood brain tumors. All CBTTC data can be download from the Gabriella Miller Kids First Data Resource Center (KF-DRC), [2]. The deidentified patient's blood and tumor tissue were prospectively collected by the consortium from patients enrolled within the CBTTC.

The cell lines were generated by the CBTTC from either fresh tumor tissue obtained firectly from surgery performed at Children's Hospital of Philadelphia (CHOP) or from prospectively collected tumor specimens stored in Recover Cell Culture Freezing media (cat# 12648010, Gibco). The tissue was dissociated using enzymatic method with papain as described [3]. Briefly, tissue was washed with HBSS (cat# 14175095, Gibco), minced and icubated with activated papin solution (cat# LS003124, SciQuest) for up to 45 minutes. The papain was inactivated using ovomucoid solution (cat# 542000, SciQuest), tissue was briefly treated with DNase (cat# 10104159001, Sigma) and passed through the 100µm cell strainer (cat# 542000, Greiner Bio-One). Two cell culture conditions were initiated based on the number of cells avialble. For cultures utilizing the fetal bovine serum (FBS), a minimum density of 3×10^5 cells/ml were plated 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 the serum-free media conditions cells were plated at minimum density of 1×10^6 cells/ml in DMEM/F12 media 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.

Pacific Pediatric Neuro Onclology Consortium (PNOC)

The Pacific Pediatric Neuro-Oncology Consortium (PNOC) is an international consortium dedicated to bringing new therapies to children and young adults with brain tumors. PNOC collected blood and tumor biospecimens from newly-diagnosed DIPG patients as part of the clinical trial PNOC003/NCT02274987 [4].

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 7 but no QC thresholds were implemented for the DNA. For library

preparation, 500ng of nucleic acids were used as input for RNA-Seq and WXS. The RNA prep was performed using the TruSeq RNA Sample Prep Kit (Illumina, #FC-122-1001) and the exome prep was performed using KAPA Library Preparation Kit (Kapa Biosystems, #KK8201) using Agilent's SureSelect Human All Exon V5 backbone with custom probes. These probes include CGH probes that target 44,000 evenly spaced genomic loci to assess copy number changes, as well as probes that tile across tumor suppressor genes and genes involved in common cancer translocations. All extractions and library preparations were performed according to manufacturer's instructions.

CBTTC samples

Blood, tissue and cell line DNA/RNA extraction was performed at Biorepository Core (BioRC) at CHOP. Briefly, 10-20 mg frozen tissue, 0.4-1ml of blood or 2×10^6 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 threated 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 RINe and DINe (RNA Integrity Number equivalent and DNA Integrity Number equivalent respectively). Library preparation and sequencing was performed by the NantHealth sequencing center. Briefly, DNA sequencing libraries were prepared for tumor and matched-normal DNA using the KAPA Hyper prep kit (cat# KK8541, Roche); tumor RNA-Seq libraries were prepared using KAPA Stranded RNA-Seq with RiboErase kit (cat# KK8484, Roche). Whole genome sequencing (WGS) was performed at an average depth of coverage of 60X for tumor samples and 30X for germline. RNA samples were sequenced to an average of 200M reads. All samples were sequenced on the Illumina HiSeq platform (X/400) (Illumina) with 2 × 150bp read length. For the cell line sequencing, samples labelled with "CL-adh" correspond to the adherent FBS cell lines and those labelled "CL-susp" are the serum-free lines.

Data generation

NantHealth Sequencing Center (Culver City, CA) performed whole genome sequencing (WGS) on all paired tumor (~60X) and constitutive (~30X) DNA samples. WGS libraries were 2x150 bp and sequenced on an Illumina X/400. NantHealth Sequencing Center performed ribosomal-depleted whole transcriptome stranded RNA-Seq to an average depth of 100M reads for CBTTC tumor samples. The Translational Genomic Research Institute (TGEN; Phoenix, AZ) performed paired tumor (~200X) and constitutive whole exome sequencing (WXS) and poly-A selected RNA-Seq (~200M reads) for PNOC tumor samples. PNOC WXS and RNA-Seq libraries 2x100 bp and sequenced on an Illumina HiSeq 2500.

DNA WGS Alignment

We used BWA-MEM [5] v0.7.17 for alignment of paired-end DNA-seq reads. The alignment reference that we used was Homo Sapiens Human Genome (hg) version 38, patch release 12, fasta file obtained from UCSC [6]. Alignments were futher processed using following the Broad Institute's Best Practices [7] for processing BAMs in preparation for variant discovery. Duplicates were marked using Samblaster[8] v0.1.24, BAMs merged and sorted using Sambamba [9] v0.6.3. Lastly, resultant BAMs were processing using Broad's Genome Analysis Tool Kit (GATK) [10] v4.0.3.0, BaseRecalibrator submodule.

Quality Control of Sequencing Data

NGSCheckmate [doi:10.1093/nar/gkx193] was peformed on matched tumor/normal CRAMs to confirm sample matches and remove mis-matched samples from the dataset. CRAM inputs were preprocessed using bcftools to filter and call 20k common SNPs using default parameters[11] and the resulting VCFs were used to run NGSCheckmate using this workflow in the D3b GitHub repository. Per author guidelines, <= 0.61 was used as a correlation coefficient cutoff at sequencing depths >10 to predict mismatched samples. For RNA-Seq, read strandedness was determined by running the `infer_experiment.py` script on the first 200k mapped reads. If calculated strandedness did not match strandedness information received from the sequencing center, samples were removed from analysis. We required at least 60% of RNA-Seq reads mapped to the human reference or samples were removed from analysis.

Somatic Single Nucleotide Variant Calling

SNV and INDEL calling

We used four variant callers to call SNVs and INDELS from targeted DNA panel, WXS, and WGS data: Strelka2, Mutect2, Lancet, and VarDict. The same input interval BED files were used for both panel and WXS data and intervals were padded by 100 bp on each side for all variant calling algorithm runs. The BED files for WGS were not padded for Mutect2 and Strelka2 runs, were padded by 300 bp on each side for Lancet, and by 100 bp on each side for VarDict. Strelka2 [12] v2.9.3 was run using default parameters on human genome reference hg38, canonical chromosomes only (chr1-22, X,Y,M), as recommended by the authors. The final Strelka2 VCF was filtered for PASS variants. Mutect2 from GATK v4.1.1.0 was run following Broad best practices outlined from their Workflow Description Language (WDL) [13]. The final Mutect2 VCF was filtered for PASS variants. Lancet [14] v1.0.7 [15] was run using default parameters, unless noted below. For input intervals to Lancet, a reference BED was created by using only the UTR, exome, and start/stop codon features of the GENCODE 31 reference. Per recommendations by the New York Genome Center, the Lancet input intervals were augmented with PASS variant calls from Strelka2 and Mutect2 as validation. VarDictJava [16] v1.58 [17] was run using the hg38 fasta reference with the same BED intervals used for Mutect2.
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 VarDict in consensus calling. The final VCF was filtered for PASS variants with TYPE=StronglySomatic.

VCF annotation and MAF creation

We filtered outputs from both callers on the “PASS” filter, and annotated using The ENSEMBL Variant Effect Predictor [18], reference release 93, and created MAFs using MSKCC’s vcf2maf [19] v1.6.16.

Somatic Copy Number Variant Calling

We used Control-FREEC [20,21] v8.7 and CNVkit [22] v0.9.3 for copy number variant calls. CNVkit was run using default parameters on human genome reference hg38 and using the batch command for tumor-normal pairs rather than a panel of normals.

Somatic Structural Variant Calling

We used Manta SV [23] v1.4.0 for structural variant (SV) calls. Manta SV calling was also limited to regions used in Strelka2. We also ran LUMPY SV [24] v0.2.13 in express mode using default parameters. The hg38 reference used was also limited to canonical chromosome regions. The somatic DNA workflow for SNV, INDEL, copy number, and SV calling can be found in the KidsFirst Github repository.

Gene Expression Abundance Estimation

We used STAR [25] v2.6.1d to align paired-end RNA-seq reads. This output was used for all subsequent RNA analysis. The reference we used was that of ENSEMBL's GENCODE 27 [26], "Comprehensive gene annotation." We used RSEM [27] v1.3.1 for transcript- and gene-level quantification. We also added a second method of quantification using kallisto [28] v0.43.1. This method differs in that it uses pseudoaligments using fastq reads directly to the aforementioned GENCODE 27 reference.

RNA Fusion Calling and Prioritization

Gene fusion detection

We set up Arriba v1.1.0 and STAR-Fusion 1.5.0 [29] fusion detection tools using CWL on CAVATICA. For both 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 GRCh38_v27_CTAT_lib_Feb092018.plug-n-play.tar.gz provided in the STAR-fusion release. For Arriba, we used a blacklist file (blacklist_hg38_GRCh38_2018-11-04.tsv.gz) from the Arriba release tarballs to remove recurrent fusion artifacts and transcripts present in healthy tissue. We also provided Arriba with strandedness information or set it to auto-detection for polyA samples. The RNA expression and fusion workflows can be found in the KidsFirst GitHub repository.

Fusion prioritization

We built a fusion prioritization pipeline to filter and annotate fusions. We considered all inframe and frameshift fusion calls with 1 or more junction reads and fused genes expressed with TPM greater than one to be true calls. If a fusion call had large number of spanning fragment reads compared to junction reads (spanning fragment minus junction read greater than ten) or if either 5 or 3 genes fused to more than five different genes we removed these calls as a potential false positive. We also removed fusions if the 5 or 3 ends were the same gene, and these were tagged as non-canonical splicing or duplication. We used a list of curated fusion calls for each histology to capture each occurrence of the fusion as a putative driver fusion. We prioritized a union of fusion calls as true calls if the fused genes were detected by both callers, the same fusion was recurrent in histology (>2 samples) or the fusion was specific to the broad histology. We annotated putative driver fusions and prioritized fusions lists with kinases, oncogenic, tumor suppressor, transcription factor, fused genes and known TCGA fusions from curated datasheets. We also added chimerDB [30] annotations to both driver and prioritized fusion list.

Clinical Data Harmonization

WHO Classification of Disease Types

The `disease_type_old` field in the `pbta-histologies.tsv` file contains the diagnosis denoted from the patient's pathology report. The `disease_type_new` field in the `pbta-histologies.tsv` file includes updates to `disease_type_old` and these changes are documented in the `Notes`. For instance, any diagnosis denoted as "Other" in `disease_type_old` was modified to capture the pathology report diagnosis in `disease_type_new`. Additionally, `disease_type_old` was modified to `disease_type_new` if the presence of specific molecular alterations defined a biospecimen as having an alternate diagnosis. The `broad_histology` denotes the broad 2016 WHO classification [doi:10.1007/s00401-016-1545-1] for each tumor. The `short_histology` is an abbreviated version of the `broad_histology`.

Molecular Subtyping

The `molecular_subtype` column in the `pbta-histologies.tsv` file contains molecular subtype information derived as described below. Medulloblastoma subtypes SHH, MYC, Group 3, and Group 4 were predicted using an RNA expression classifier on the RSEM FPKM data.

Survival

Overall survival was calculated as days since initial diagnosis.

Prediction of participants' genetic sex

The clinical metadata provided included a reported gender. We used DNA data, in concert with the reported gender, to predict participant genetic sex so that we could identify sexually dimorphic outcomes. This analysis could also reveal samples that may have been contaminated in certain circumstances. We used the idxstats utility from SAMTOOLS [31] 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 reviewed this statistic in the context of reported gender and determined that a threshold of less than 0.2 clearly delineated female samples. Fractions greater than 0.4 were predicted to be males. Samples with values in the range [0.2, 0.4] were marked as unknown. We ran this analysis through CWL on Cavatica. Resulting calls were added to the clinical metadata as `germline_sex_estimate`.

Results

Results section stub.

Conclusions

Stub in conclusions section

References

1. **Home**
Children's Brain Tumor Tissue Consortium
https://cbttc.org/

2. **Working Together to Put Kids First**https://kidsfirstdrc.org/

3. **Pediatric High Grade Glioma Resources From the Children's Brain Tumor Tissue Consortium (CBTTC) and Pediatric Brain Tumor Atlas (PBTA)**
Heba Ijaz, Mateusz Koptyra, Krutika S. Gaonkar, Jo Lynne Rokita, Valerie P. Baubet, Lamiya Tauhid, Yankun Zhu, Miguel Brown, Gonzalo Lopez, Bo Zhang, ...
Cold Spring Harbor Laboratory (2019-05-31) https://doi.org/gf66qt
DOI: 10.1101/656587

4. **A pilot precision medicine trial for children with diffuse intrinsic pontine glioma—PNOC003: A report from the Pacific Pediatric Neuro-Oncology Consortium**
Sabine Mueller, Payal Jain, Winnie S. Liang, Lindsay Kilburn, Cassie Kline, Nalin Gupta, Eshini Panditharatna, Suresh N. Magge, Bo Zhang, Yuankun Zhu, ... Adam C. Resnick
International Journal of Cancer (2019-04-03) https://doi.org/gf6pfb
DOI: 10.1002/ijc.32258 · PMID: 30861105

5. **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**
Heng Li
arXiv (2013-03-16) https://arxiv.org/abs/1303.3997v2

6. **Index of /goldenPath/hg38/bigZips**http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/

7. **GATK | BP Doc # |**https://software.broadinstitute.org/gatk/best-practices/workflow?id

8. **SAMBLASTER: fast duplicate marking and structural variant read extraction**
G. G. Faust, I. M. Hall
Bioinformatics (2014-05-07) https://doi.org/f6kft3
DOI: 10.1093/bioinformatics/btu314 · PMID: 24812344 · PMCID: PMC4147885

9. **Sambamba: fast processing of NGS alignment formats**
Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, Pjotr Prins
Bioinformatics (2015-02-19) https://doi.org/gfzsfw
DOI: 10.1093/bioinformatics/btv098 · PMID: 25697820 · PMCID: PMC4765878

10. **GATK | Home**https://software.broadinstitute.org/gatk/

11. **Software program for checking sample matching for NGS data: parklab/NGSCheckMate**
Park Lab at Harvard Medical School
(2019-09-12) https://github.com/parklab/NGSCheckMate

12. **Strelka2: fast and accurate calling of germline and somatic variants**
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

13. **Official code repository for GATK versions 4 and up: broadinstitute/gatk**
Broad Institute
(2019-09-28) https://github.com/broadinstitute/gatk

14. **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

15. **Microassembly based somatic variant caller for NGS data: nygenome/lancet**
New York Genome Center
(2019-09-24) https://github.com/nygenome/lancet

16. **VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research**
Zhongwu Lai, Aleksandra Markovets, Miika Ahdesmaki, Brad Chapman, Oliver Hofmann, Robert McEwen, Justin Johnson, Brian Dougherty, J. Carl Barrett, Jonathan R. Dry
Nucleic Acids Research (2016-04-07) https://doi.org/f8v6qz
DOI: 10.1093/nar/gkw227 · PMID: 27060149 · PMCID: PMC4914105

17. **VarDict Java port. Contribute to AstraZeneca-NGS/VarDictJava development by creating an account on GitHub**
AstraZeneca - NGS Team
(2019-09-27) https://github.com/AstraZeneca-NGS/VarDictJava

18. **The Ensembl Variant Effect Predictor**
William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. 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

19. **Convert a VCF into a MAF, where each variant is annotated to only one of all possible gene isoforms: mskcc/vcf2maf**
Memorial Sloan Kettering
(2019-09-20) https://github.com/mskcc/vcf2maf

20. **Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data**
Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, Emmanuel Barillot
Bioinformatics (2011-12-06) https://doi.org/ckt4vz
DOI: 10.1093/bioinformatics/btr670 · PMID: 22155870 · PMCID: PMC3268243

21. **Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization**
Valentina Boeva, Andrei Zinovyev, Kevin Bleakley, Jean-Philippe Vert, Isabelle Janoueix-Lerosey, Olivier Delattre, Emmanuel Barillot
Bioinformatics (2010-11-15) https://doi.org/c6bcps
DOI: 10.1093/bioinformatics/btq635 · PMID: 21081509 · PMCID: PMC3018818

22. **CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing**

Eric Talevich, A. Hunter Shain, Thomas Botton, Boris C. Bastian
PLOS Computational Biology (2016-04-21) https://doi.org/c9pd
DOI: 10.1371/journal.pcbi.1004873 · PMID: 27100738 · PMCID: PMC4839673

23. **Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications**
Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, Christopher T. Saunders
Bioinformatics (2015-12-08) https://doi.org/gf3ggb
DOI: 10.1093/bioinformatics/btv710 · PMID: 26647377

24. **LUMPY: a probabilistic framework for structural variant discovery**
Ryan M Layer, Colby Chiang, Aaron R Quinlan, Ira M Hall
Genome Biology (2014) https://doi.org/gf3ggc
DOI: 10.1186/gb-2014-15-6-r84 · PMID: 24970577 · PMCID: PMC4197822

25. **STAR: ultrafast universal RNA-seq aligner**
Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras
Bioinformatics (2012-10-25) https://doi.org/f4h523
DOI: 10.1093/bioinformatics/bts635 · PMID: 23104886 · PMCID: PMC3530905

26. **GENCODE - Human Release 27**https://www.gencodegenes.org/human/release_27.html

27. **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**
Bo Li, Colin N Dewey
BMC Bioinformatics (2011-08-04) https://doi.org/cwg8n5
DOI: 10.1186/1471-2105-12-323 · PMID: 21816040 · PMCID: PMC3163565

28. **Near-optimal probabilistic RNA-seq quantification**
Nicolas L Bray, Harold Pimentel, Páll Melsted, Lior Pachter
Nature Biotechnology (2016-04-04) https://doi.org/f8nvsp
DOI: 10.1038/nbt.3519 · PMID: 27043002

29. **STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq**
Brian J. Haas, Alex Dobin, Nicolas Stransky, Bo Li, Xiao Yang, Timothy Tickle, Asma Bankapur, Carrie Ganote, Thomas G. Doak, Nathalie Pochet, ... Aviv Regev
Cold Spring Harbor Laboratory (2017-03-24) https://doi.org/gf5pc5
DOI: 10.1101/120295

30. **OUP accepted manuscript**
Nucleic Acids Research
(2016) https://doi.org/gf6bx9
DOI: 10.1093/nar/gkw1083 · PMID: 27899563 · PMCID: PMC5210563

31. **The Sequence Alignment/Map format and SAMtools.**
Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin,
Bioinformatics (Oxford, England) (2009-06-08) https://www.ncbi.nlm.nih.gov/pubmed/19505943
DOI: 10.1093/bioinformatics/btp352 · PMID: 19505943 · PMCID: PMC2723002