An Open Pediatric Brain Tumor Atlas

This manuscript (permalink) was automatically generated from AlexsLemonade/OpenPBTA-manuscript@a14cb09 on August 22, 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). Blood and tumor biospecimens from patients enrolled within the CBTTC were sent to the Children's Hospital of Philadelphia for RNA and/or DNA extraction. PNOC collected blood and tumor biospecimens from newly-diagnosed DIPG patients as part of the clinical trial PNOC003/NCT02274987 [1].

Nucleic acids extraction

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. 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 [2] 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 [3]. Alignments were futher processed using following the Broad Institute's Best Practices [4] for processing BAMs in preparation for variant discovery. Duplicates were marked using Samblaster[5] v0.1.24, BAMs merged and sorted using Sambamba [6] v0.6.3. Lastly, resultant BAMs were processing using Broad's Genome Analysis Tool Kit (GATK) [7] v4.0.3.0, BaseRecalibrator submodule.

Germ Line Single Nucleotide Variant Calling

Somatic Single Nucleotide Variant Calling

SNV and INDEL calling

We used Strelka2 [8] v2.9.3 and Mutect2 from GATK v4.1.1.0. Strelka2 was run using default parameters on human genome reference hg38, canonical chromosomes only (chr1-22, X,Y,M), as recommended by the author. Mutect2 was run following Broad best practices outlined from their Workflow Description Language (WDL) [9].

VCF annotation and MAF creation

We filtered outputs from both callers on the “PASS” filter, and annotated using The ENSEMBL Variant Effect Predictor [10], reference release 93, and created MAFs using MSKCC’s vcf2maf [11] v1.6.16.

Somatic Copy Number Variant Calling

We used Control-FREEC [12,13] v8.7 and CNVkit [14] 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 [15] v1.4.0 for structural variant (SV) calls. Manta SV calling was also limited to regions used in Strelka2. We also ran LUMPY SV [16] v0.2.13 in express mode using default parameters. The hg38 reference used was also limited to canonical chromosome regions.

Gene Expression Abundance Estimation

We used STAR [17] 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 [18], “Comprehensive gene annotation.” We used RSEM [19] v1.3.1 for transcript- and gene-level quantification. We also added a second method of quantification using kallisto [20] 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 [21] 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.

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 [22] annotations to both driver and prioritized fusion list.

Results

Results section stub.

Conclusions

Stub in conclusions section

References

1. **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

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

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

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

5. **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

6. **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

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

8. **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

9. **Official code repository for GATK versions 4 and up: broadinstitute/gatk**
Broad Institute
(2019-08-22) https://github.com/broadinstitute/gatk

10. **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

11. **Convert a VCF into a MAF, where each variant is annotated to only one of all possible gene isoforms: mskcc/vcf2maf**
Memorial Sloan Kettering
(2019-08-20) https://github.com/mskcc/vcf2maf

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

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

19. **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

20. **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

21. **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

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