

An Open Pediatric Brain Tumor Atlas

This manuscript ([permalink](#)) was automatically generated from [AlexsLemonade/OpenPBTA-manuscript@5f5c17f](#) on November 26, 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 directly 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 incubated with activated papain 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 available. 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 Oncology 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, WXS, and targeted DNA panel (panel). 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. The targeted DNA panel developed by Ashion (formerly known as the GEM Cancer panel) consisted of exonic probes against 541 cancer genes. Both panel and WXS assays contained 44,000 probes across evenly spaced genomic loci used for genome-wide copy number analysis. For the panel, additional probes tiled across intronic regions of 22 known tumor suppressor genes and 22 genes involved in common cancer translocations for structural analysis. All extractions and library preparations were performed according to manufacturer's instructions.

CBTTC samples

Blood, tissue, and cell line DNA/RNA extractions were 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 treated with RNase A (cat#, 19101, Qiagen); 0.4-1ml was processed using the Qiagen QIAasympy automated platform (Qiagen) using the QIAasympy 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. The panel tumor sample was sequenced to 470X and the normal panel sample was sequenced to 308X. RNA samples were sequenced to an average of 200M reads. All samples were sequenced on the Illumina HiSeq platform (X/400) (Illumina) with 2×150 bp read length.

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) or targeted DNA panel (panel) 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 further 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 [11] was performed 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[12] 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.

Germline Variant Calling

SNV calling for B-allele Frequency (BAF) generation

Germline haplotype calls were performed following the [GATK Joint Genotyping Workflow](#), except the workflow was run on an individual sample basis. Using only SNPs, we applied the [GATK generic hard filter suggestions](#) to the VCF, with an additional requirement of 10 reads minimum depth per SNP. This filtered VCF was used as input to ControlFreeC and CNVKit (below) for generation of BAF files. GATK v4.0.12.0 was used for all steps except `VariantFiltration`, which used 3.8.0 because as of GATK 4.0.12.0, this tool was beta and known to be unreliable for this purpose. This single-sample workflow can be found in the [Kids First GitHub repository](#).

Somatic Mutation 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 input interval BED files for both panel and WXS data provided by the manufacturers were padded by 100 bp on each side during all variant calling algorithm runs. For WGS calling, we utilized the non-padded BROAD Institute interval calling list [wgs_calling_regions.hg38.interval_list](#), comprised of the full genome minus N bases, unless otherwise noted below. Strelka2 [13] v2.9.3 was run using default parameters for canonical chromosomes (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) [14]. The final Mutect2 VCF was filtered for PASS variants. To manage memory issues, VarDictJava [15] v1.58 [16] was run using 20Kb interval chunks of the input BED, padded by 100 bp on each side, such that if an INDEL occurred in between intervals, it would be captured. 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. Lancet [17] v1.0.7 [18] 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, augmented as recommended with PASS variant calls from Strelka2 and Mutect2 [19]. These intervals were then padded by 300 bp on each side during Lancet variant calling. Per recommendations by the New York Genome Center [19], the Lancet input intervals described above were augmented with PASS variant calls from Strelka2 and Mutect2 as validation.

VCF annotation and MAF creation

We filtered outputs from both callers on the “PASS” filter, and annotated using The ENSEMBL Variant Effect Predictor [20], reference release 93, and created MAFs using MSKCC’s vcf2maf [21] v1.6.17.

Somatic Copy Number Variant Calling

We used Control-FREEC [22,23] v11.6 and CNVkit [24] v0.9.3 for copy number variant calls. For both algorithms, the `germline_sex_estimate` (described below) was used as input for sample sex and germline variant calls (above) were used as input for BAF estimation. ControlFreeC was run on human genome reference hg38 using the optional parameters of a 0.05 coefficient of variation, ploidy choice of 2-4, and BAF adjustment for tumor-normal pairs. Theta2 [25] used VarDict germline and somatic calls, filtered on PASS and strongly somatic, to infer tumor purity. Theta2 purity was added as an optional parameter to CNVkit to adjust copy number calls. CNVkit was run on human genome reference hg38 using the optional parameters of Theta2 purity and BAF adjustment for tumor-normal pairs.

Somatic Structural Variant Calling

We used Manta SV [26] v1.4.0 for structural variant (SV) calls. Manta SV calling was also limited to regions used in Strelka2. The hg38 reference for SV calling used was 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](#). Manta SV output was annotated using [AnnotSV v2.1](#) [27] and the workflow can be found in the [D3b GitHub repository](#).

Gene Expression Abundance Estimation

We used STAR [28] 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 [29], “Comprehensive gene annotation.” We used RSEM [30] v1.3.1 for both FPKM and TPM transcript- and gene-level quantification. We also added a second method of quantification using kallisto [31] v0.43.1. This method differs in that it uses pseudoalignments 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 [32] 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.plugin-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. We used [FusionAnnotator](#) on Arriba fusion calls in order to harmonize annotations with those of STAR-Fusion. The RNA expression and fusion workflows can be found in the [KidsFirst Github repository](#) and the FusionAnnotator workflow found in the [D3b GitHub repository](#).

Fusion prioritization

We performed artifact filtering and additional annotation on fusion calls to prioritize putative oncogenic fusions. Briefly, we considered all inframe and frameshift fusion calls with a minimum of 1 junction reads and at least one gene partner expressed (TPM > 1) 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), we removed these calls as potential false positives. We prioritized a union of fusion calls as true calls if the fused genes were detected by both callers, the same fusion was recurrent within a `broad_histology` (>2 samples) or the fusion was specific to the `broad_histology`. If either 5' or 3' genes fused to more than five different genes within a sample, we removed these calls as potential false positives. We annotated putative driver fusions and prioritized fusions based on partners containing known [kinases](#), [oncogenes](#), [tumor suppressors](#), curated transcription factors [33], [COSMIC genes](#), and/or known [TCGA fusions](#) from curated [references](#). *MYBL1* [34], *SNCAIP* [35], *FOXR2* [36], *TTYH1* [37], and *TERT* [38,39,40,41] were added to the oncogene list and *BCOR* [36] and *QKI* [42] were added to the tumor suppressor gene list based on pediatric cancer literature review. The fusion filtering workflow can be found in the [OpenPBTA Analysis repository](#).

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`. The `glioma_brain_region` was subtyped into hemispheric, midline, mixed, or other based on specimen location (see table below).

Meta data	Definition	Possible values
<code>age_at_diagnosis_days</code>	Patient age at diagnosis in days	numeric
<code>age_last_update_days</code>	Patient age at the last clinical event/update in days	numeric
<code>aliquot_id</code>	External aliquot identifier	variable
<code>broad_composition</code>	Broad classification of sample type	cell-line;cyst;non-tumor;tumor
<code>broad_histology</code>	Broad WHO 2016 classification of cancer type	text
<code>cancer_predispositions</code>	Reported cancer predisposition syndromes	text
<code>cohort</code>	Scientific cohort	CBTTC;PNOC

Meta data	Definition	Possible values
comp ositio n	Sample composition	Derived Cell Line;Not Reported;Peripheral Whole Blood;Saliva;Solid Tissue
diseas e_type _new	Updated and/or integrated molecular diagnosis	text
diseas e_type _old	Reported patient diagnosis from pathology reports	text
ethnic ity	Patient reported ethnicity	text
experi menta l_strat egy	Sequencing strategy	WGS;WXS;RNA-Seq;Panel
germli ne_se x_esti mate	Predicted sex of patient based on germline X and Y ratio calculation (described in methods)	Female;Male;Unknown
gliom a_brai n_regi on	Brain region for all tumors classified as LGAT or HGAT	midline (Thalamus
Kids_F irst_Bi ospeci men_I D	KidsFirst Biopecimen identifier	BS_#####
Kids_F irst_P articip ant_ID	KidsFirst patient identifier	PT_#####
molec ular_s ubtyp e	Molecular subtype defined by WHO 2016 guidelines	text
norm al_frac tion	Theta2 normal DNA fraction estimate	numeric
Notes	Free text field describing changes from diagnosis_old to diagnosis_new or manner in which molecular_subtype was determined	text
OS_da ys	Overall survival in days	numeric
OS_st atus	Overall survival status	DECEASED;LIVING
prima ry_site	Bodily site(s) from which specimen was derived	text
race	Patient reported race	text

Meta data	Definition	Possible values
report ed_gender	Patient reported gender	text
RNA_library	Type of RNA-Sequencing library preparation	stranded;poly-A
sample_id	External biospecimen identifier	variable
sample_type	Broad sample type	Normal;Tumor
seq_center	Sequencing center	BGI@CHOP Genome Center;Genomic Clinical Core at Sidra Medical and Research Center;NantOmics;TGEN
short_histology	Abbreviated disease_type_new	text
tumor_descriptor	Phase of therapy from which tumor was derived	Initial CNS Tumor;Progressive Progressive Disease Post-Mortem;Recurrence;Second Malignancy;Unavailable
tumor_fraction	Theta2 tumor DNA fraction estimate	numeric
tumor_ploidy	ControlFreeC ploidy	numeric

Table S1. Clinical metadata collected for OpenPBTA. {#tbl:S1}

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, denoted `OS_days`, 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 [43] 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`.

Selection of independent samples

Certain analyses required that we select only a single representative specimen for each individual. In these cases, we prioritized primary tumors and those with whole-genome sequencing available. If this filtering still resulted in multiple specimens, we selected from the remaining set randomly.

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](https://doi.org/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](https://doi.org/10.1002/ijc.32258) · PMID: [30861105](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1093/bioinformatics/btu314) · PMID: [24812344](https://pubmed.ncbi.nlm.nih.gov/24812344/) · PMCID: [PMC4147885](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1093/bioinformatics/btv098) · PMID: [25697820](https://pubmed.ncbi.nlm.nih.gov/25697820/) · PMCID: [PMC4765878](https://pubmed.ncbi.nlm.nih.gov/PMC4765878/)

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

11. NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types

Sejoon Lee, Soohyun Lee, Scott Ouellette, Woong-Yang Park, Eunjung A. Lee, Peter J. Park

Nucleic Acids Research (2017-03-23) <https://doi.org/f9xrq4>

DOI: [10.1093/nar/gkx193](https://doi.org/10.1093/nar/gkx193) · PMID: [28369524](https://pubmed.ncbi.nlm.nih.gov/28369524/) · PMCID: [PMC5499645](https://pubmed.ncbi.nlm.nih.gov/PMC5499645/)

12. Software program for checking sample matching for NGS data: parklab/NGSCheckMate

Park Lab at Harvard Medical School

(2019-11-14) <https://github.com/parklab/NGSCheckMate>

13. **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](https://doi.org/10.1038/s41592-018-0051-x) · PMID: [30013048](https://pubmed.ncbi.nlm.nih.gov/30013048/)

14. **Official code repository for GATK versions 4 and up: broadinstitute/gatk**

Broad Institute

(2019-11-25) <https://github.com/broadinstitute/gatk>

15. **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](https://doi.org/10.1093/nar/gkw227) · PMID: [27060149](https://pubmed.ncbi.nlm.nih.gov/27060149/) · PMCID: [PMC4914105](https://pubmed.ncbi.nlm.nih.gov/PMC4914105/)

16. **VarDict Java port. Contribute to AstraZeneca-NGS/VarDictJava development by creating an account on GitHub**

AstraZeneca - NGS Team

(2019-11-25) <https://github.com/AstraZeneca-NGS/VarDictJava>

17. **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](https://doi.org/10.1038/s42003-018-0023-9) · PMID: [30271907](https://pubmed.ncbi.nlm.nih.gov/30271907/) · PMCID: [PMC6123722](https://pubmed.ncbi.nlm.nih.gov/PMC6123722/)

18. **Microassembly based somatic variant caller for NGS data: nygenome/lancet**

New York Genome Center

(2019-11-16) <https://github.com/nygenome/lancet>

19. **Deep sequencing of 3 cancer cell lines on 2 sequencing platforms**

Kanika Arora, Minita Shah, Molly Johnson, Rashesh Sanghvi, Jennifer Shelton, Kshithija Nagulapalli, Dayna M. Oschwald, Michael C. Zody, Soren Germer, Vaidehi Jobanputra, ... Nicolas Robine

Cold Spring Harbor Laboratory (2019-04-30) <https://doi.org/ggc9vx>

DOI: [10.1101/623702](https://doi.org/10.1101/623702)

20. **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](https://doi.org/10.1186/s13059-016-0974-4) · PMID: [27268795](https://pubmed.ncbi.nlm.nih.gov/27268795/) · PMCID: [PMC4893825](https://pubmed.ncbi.nlm.nih.gov/PMC4893825/)

21. **Convert a VCF into a MAF, where each variant is annotated to only one of all possible gene isoforms: mskcc/vcf2maf**

Memorial Sloan Kettering

(2019-11-20) <https://github.com/mskcc/vcf2maf>

22. **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 Cappel, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, Emmanuel Barillot

Bioinformatics (2011-12-06) <https://doi.org/ckt4vz>

DOI: [10.1093/bioinformatics/btr670](https://doi.org/10.1093/bioinformatics/btr670) · PMID: [22155870](https://pubmed.ncbi.nlm.nih.gov/22155870/) · PMCID: [PMC3268243](https://pubmed.ncbi.nlm.nih.gov/PMC3268243/)

23. 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](https://doi.org/10.1093/bioinformatics/btq635) · PMID: [21081509](https://pubmed.ncbi.nlm.nih.gov/21081509/) · PMCID: [PMC3018818](https://pubmed.ncbi.nlm.nih.gov/PMC3018818/)

24. 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](https://doi.org/10.1371/journal.pcbi.1004873) · PMID: [27100738](https://pubmed.ncbi.nlm.nih.gov/27100738/) · PMCID: [PMC4839673](https://pubmed.ncbi.nlm.nih.gov/PMC4839673/)

25. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data

Layla Oesper, Gryte Satas, Benjamin J. Raphael

Bioinformatics (2014-10-08) <https://doi.org/f6rmgt>

DOI: [10.1093/bioinformatics/btu651](https://doi.org/10.1093/bioinformatics/btu651) · PMID: [25297070](https://pubmed.ncbi.nlm.nih.gov/25297070/) · PMCID: [PMC4253833](https://pubmed.ncbi.nlm.nih.gov/PMC4253833/)

26. 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](https://doi.org/10.1093/bioinformatics/btv710) · PMID: [26647377](https://pubmed.ncbi.nlm.nih.gov/26647377/)

27. AnnotSV: an integrated tool for structural variations annotation

Véronique Geoffroy, Yvan Herenger, Arnaud Kress, Corinne Stoetzel, Amélie Piton, Hélène Dollfus, Jean Muller

Bioinformatics (2018-04-14) <https://doi.org/gdcsh3>

DOI: [10.1093/bioinformatics/bty304](https://doi.org/10.1093/bioinformatics/bty304) · PMID: [29669011](https://pubmed.ncbi.nlm.nih.gov/29669011/)

28. 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](https://doi.org/10.1093/bioinformatics/bts635) · PMID: [23104886](https://pubmed.ncbi.nlm.nih.gov/23104886/) · PMCID: [PMC3530905](https://pubmed.ncbi.nlm.nih.gov/PMC3530905/)

29. GENCODE - Human Release 27 https://www.encodegenes.org/human/release_27.html

30. 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](https://doi.org/10.1186/1471-2105-12-323) · PMID: [21816040](https://pubmed.ncbi.nlm.nih.gov/21816040/) · PMCID: [PMC3163565](https://pubmed.ncbi.nlm.nih.gov/PMC3163565/)

31. 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](https://doi.org/10.1038/nbt.3519) · PMID: [27043002](https://pubmed.ncbi.nlm.nih.gov/27043002/)

32. 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](https://doi.org/10.1101/120295)

33. The Human Transcription Factors

Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, Matthew T. Weirauch

Cell (2018-02) <https://doi.org/gcw8rb>

DOI: [10.1016/j.cell.2018.01.029](https://doi.org/10.1016/j.cell.2018.01.029) · PMID: [29425488](https://pubmed.ncbi.nlm.nih.gov/29425488/)

34. Genomic analysis of diffuse pediatric low-grade gliomas identifies recurrent oncogenic truncating rearrangements in the transcription factor MYBL1

L. A. Ramkissoon, P. M. Horowitz, J. M. Craig, S. H. Ramkissoon, B. E. Rich, S. E. Schumacher, A. McKenna, M. S. Lawrence, G. Bergthold, P. K. Brastianos, ... K. L. Ligon

Proceedings of the National Academy of Sciences (2013-04-30) <https://doi.org/f42gg4>

DOI: [10.1073/pnas.1300252110](https://doi.org/10.1073/pnas.1300252110) · PMID: [23633565](https://pubmed.ncbi.nlm.nih.gov/23633565/) · PMCID: [PMC3657784](https://pubmed.ncbi.nlm.nih.gov/PMC3657784/)

35. Subgroup-specific structural variation across 1,000 medulloblastoma genomes

Paul A. Northcott, David J. H. Shih, John Peacock, Livia Garzia, A. Sorana Morrissy, Thomas Zichner, Adrian M. Stütz, Andrey Korshunov, Jüri Reimand, Steven E. Schumacher, ... Michael D. Taylor

Nature (2012-07-25) <https://doi.org/ggdhk3>

DOI: [10.1038/nature11327](https://doi.org/10.1038/nature11327) · PMID: [22832581](https://pubmed.ncbi.nlm.nih.gov/22832581/) · PMCID: [PMC3683624](https://pubmed.ncbi.nlm.nih.gov/PMC3683624/)

36. New Brain Tumor Entities Emerge from Molecular Classification of CNS-PNETs

Dominik Sturm, Brent A. Orr, Umut H. Toprak, Volker Hovestadt, David T.W. Jones, David Capper, Martin Sill, Ivo Buchhalter, Paul A. Northcott, Irina Leis, ... Marcel Kool

Cell (2016-02) <https://doi.org/f3t869>

DOI: [10.1016/j.cell.2016.01.015](https://doi.org/10.1016/j.cell.2016.01.015) · PMID: [26919435](https://pubmed.ncbi.nlm.nih.gov/26919435/) · PMCID: [PMC5139621](https://pubmed.ncbi.nlm.nih.gov/PMC5139621/)

37. Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR

Claudia L Kleinman, Noha Gerges, Simon Papillon-Cavanagh, Patrick Sin-Chan, Alben Pramatarova, Dong-Anh Khuong Quang, Véronique Adoue, Stephan Busche, Maxime Caron, Haig Djambazian, ... Nada Jabado

Nature Genetics (2013-12-08) <https://doi.org/ggdhk4>

DOI: [10.1038/ng.2849](https://doi.org/10.1038/ng.2849) · PMID: [24316981](https://pubmed.ncbi.nlm.nih.gov/24316981/)

38. TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors

Linda J Valentijn, Jan Koster, Danny A Zwijnenburg, Nancy E Hasselt, Peter van Sluis, Richard Volckmann, Max M van Noesel, Rani E George, Godelieve AM Tytgat, Jan J Molenaar, Rogier Versteeg

Nature Genetics (2015-11-02) <https://doi.org/ggdhk5>

DOI: [10.1038/ng.3438](https://doi.org/10.1038/ng.3438) · PMID: [26523776](https://pubmed.ncbi.nlm.nih.gov/26523776/)

39. Recurrent pre-existing and acquired DNA copy number alterations, including focalTERTgains, in neuroblastoma central nervous system metastases

David Cobrinik, Irina Ostrovnaya, Maryam Hassimi, Satish K. Tickoo, Irene Y. Cheung, Nai-Kong V. Cheung

Genes, Chromosomes and Cancer (2013-10-10) <https://doi.org/f5gd94>

DOI: [10.1002/gcc.22110](https://doi.org/10.1002/gcc.22110) · PMID: [24123354](https://pubmed.ncbi.nlm.nih.gov/24123354/)

40. Activation of human telomerase reverse transcriptase through gene fusion in clear cell sarcoma of the kidney

Jenny Karlsson, Henrik Lilljebjörn, Linda Holmquist Mengelbier, Anders Valind, Marianne Rissler, Ingrid Øra, Thoas Fioretos, David Gisselsson

Cancer Letters (2015-02) <https://doi.org/f25ck5>

DOI: [10.1016/j.canlet.2014.11.057](https://doi.org/10.1016/j.canlet.2014.11.057) · PMID: [25481751](https://pubmed.ncbi.nlm.nih.gov/25481751/)

41. New Molecular Considerations for Glioma: IDH, ATRX, BRAF, TERT, H3 K27M

Michael Karsy, Jian Guan, Adam L. Cohen, Randy L. Jensen, Howard Colman

Current Neurology and Neuroscience Reports (2017-02) <https://doi.org/ggdhk2>

DOI: [10.1007/s11910-017-0722-5](https://doi.org/10.1007/s11910-017-0722-5) · PMID: [28271343](https://pubmed.ncbi.nlm.nih.gov/28271343/)

42. MYB-QKI rearrangements in angiocentric glioma drive tumorigenicity through a tripartite mechanism

Pratiti Bandopadhyay, Lori A Ramkissoon, Payal Jain, Guillaume Bergthold, Jeremiah Wala, Rhamy Zeid, Steven E Schumacher, Laura Urbanski, Ryan O'Rourke, William J Gibson, ... Adam C Resnick

Nature Genetics (2016-02-01) <https://doi.org/f8bwqn>

DOI: [10.1038/ng.3500](https://doi.org/10.1038/ng.3500) · PMID: [26829751](https://pubmed.ncbi.nlm.nih.gov/26829751/) · PMCID: [PMC4767685](https://pubmed.ncbi.nlm.nih.gov/PMC4767685/)

43. 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](https://doi.org/10.1093/bioinformatics/btp352) · PMID: [19505943](https://pubmed.ncbi.nlm.nih.gov/19505943/) · PMCID: [PMC2723002](https://pubmed.ncbi.nlm.nih.gov/PMC2723002/)