An Open Pediatric Brain Tumor Atlas

This manuscript (permalink) was automatically generated from AlexsLemonade/OpenPBTA-manuscript@c802c65 on May 29, 2020.

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, [2]. The de-identified 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 diffuse intrinsic pontine glioma (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 the Biorepository Core 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 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 HyperPrep 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 × 150bp 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. We used version 38, patch release 12 of the *Homo sapiens* genome as our alignment reference, which we obtained as a FASTA file from UCSC [6]. Alignments were further processed using following the Broad Institute's Best Practices [7] for processing Binary Alignment/Map files (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. Resultant BAMs were processing using Broad's Genome Analysis Tool Kit [GATK] (https://software.broadinstitute.org/gatk/) v4.0.3.0, BaseRecalibrator submodule. Lastly, for normal/germline input, we run the GATK HaplotypeCaller [10] submodule on the recalibrated BAM,

generating a genomic variant call format (GVCF) file. This file is used as the basis for germline calling, described in the "SNV calling for B-allele Frequency (BAF) generation" section. References can be obtained from the Broad Genome References on AWS bucket, with a general description of references here [11].

Quality Control of Sequencing Data

NGSCheckmate [12] was performed on matched tumor/normal CRAM files to confirm sample matches and remove mismatched samples from the dataset. CRAM inputs were preprocessed using BCFtools to filter and call 20k common single nucleotide polymorphisms (SNPs) using default parameters[13] 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. MEND QC [14] was performed on aligned RNA-Seq reads using this workflow to identify mapped exonic non-duplicate reads.

Germline Variant Calling

SNP 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. This workflow was applied to the GVCF output from the alignment workflow on normal/germline samples. 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 Control-FREEC 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. References can be obtained from the Broad Genome References on AWS bucket, with a general description of references here [11].

Somatic Mutation Calling

SNV and indel calling

For PBTA samples, we used four variant callers to call SNVs and indels from targeted DNA panel, WXS, and WGS data: Strelka2 [15], Mutect2 [16], Lancet [17], and VarDict [18]. WXS samples from TCGA were run using Strelka2, Mutect2 and Lancet. TCGA samples were captured using different WXS target capture kits and all the BED files were downloaded from `GDC portal`. The input interval BED files for both panel and WXS data for PBTA samples were provided by the manufacturers. For both PBTA and TCGA, all panel and WXS BED files were padded by 100 bp on each side during Strelka2, Mutect2, and VarDict runs and 400 bp for the Lancet run.
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 [15] 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) [19]. The final Mutect2 VCF was filtered for PASS variants. To manage memory issues, VarDictJava [18] v1.58 [20] 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 v1.0.7 was run using default parameters, except for those noted below. For input intervals to Lancet WGS, 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 [21]. These intervals were then padded by 300 bp on each side during Lancet variant calling. Per recommendations by the New York Genome Center [21], for WGS samples, 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 [22], reference release 93, and created MAFs using MSKCC's vcf2maf [23] v1.6.17.

Consensus SNV Calling

Our SNV calling process led to separate sets of predicted mutations for each caller. We considered mutations to describe the same change if they were identical for the following MAF fields: `Chromosome`, `Start_Position`, `Reference_Allele`, `Allele`, and `Tumor_Sample_Barcode`. Strelka2 does not call multinucleotide variants (MNV), but instead calls each component SNV as a separate mutation, so we separated MNV calls from Mutect2 and Lancet into consecutive SNVs before comparing them with Strelka2. We examined the variant allele frequencies produced by each caller and compared their overlap with each other [24]. VarDict calls included many variants that were not identified by other callers [25], while the other callers produced results that were relatively consistent with one another. Many of these VarDict-specific calls were variants with low allele frequency [26]. We termed mutations shared among the other three callers (Strelka2, Mutect2, and Lancet) to be consensus mutation calls and dropped VarDict due to concerns about it calling a large number of false positives. In practice, because our filtered set was based on the intersection of these three sets and because VarDict called nearly every mutation from the other three callers plus many that were unique to it, the decision to not consider VarDict calls has little impact on the results.

For some downstream analyses, only coding sequence SNVs (based on GENCODE v27 [27]) are used, to enhance comparability to other studies. We considered base pairs to be *effectively surveyed* if they were in the intersection of the genomic ranges considered by the callers used to generate the consensus and where appropriate, regions of interest, such as coding sequences. This definition of *effectively surveyed* base pairs is what is used to calculate effective genome size for calculations for tumor mutation burden and mutational signatures.

Recurrently mutated genes and co-occurrence of gene mutations

Using the consensus SNV calls, we identified genes that were recurrently mutated in the cohort, including nonsynonymous mutations with a variant allele frequency greater than 5% among the set of independent samples. The set of nonsynonymous mutations was determined using ENSEMBL Variant Effect Predictor [22] annotations, including High and Moderate consequence types as defined in `maftools` [28]. For each gene, we then tallied the number of samples that had at least one nonsynonymous mutation.

For genes that contained nonsynonymous mutations in multiple samples, we calculated pairwise mutation co-occurrence scores. This score was defined as the $I \times -\log_{10}(P)$ where I is 1 when the odds ratio is > 1 (indicating co-occurrence), and -1 when the odds ratio is < 1 (indicating mutual exclusivity), with P defined by Fisher's Exact Test.

Somatic Copy Number Variant Calling (WGS samples only)

We used Control-FREEC [29,30] v11.6 and CNVkit [31] 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. Control-FREEC 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 [32] 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. We used GISTIC [33] v.2.0.23 on the CNVkit and the consensus CNV segmentation files to generate gene-level copy number abundance (Log R Ratio) as well as chromosomal arm copy number alterations using the parameters specified in the OpenPBTA Analysis repository.

Consensus CNV Calling

For each caller and sample, CNVs were called based on consensus among Control-FREEC [29,30], CNVkit [31], and Manta [34]. CNVs called significant by Control-FREEC (p-value < 0.01) were included in the consensus calling.

Sample and caller combination files with more than 2500 CNVs called were removed from the set; we expect these to be noisy and poor quality samples based on cutoffs used in GISTIC [33]. For each sample, regions with reciprocal overlap of at least 50% between two of the three callers were included in the final consensus set. CNV regions within 10,000 bp of each other with the same direction of gain or loss were merged into single region. We filtered out any CNVs that overlapped 50% or more with immunoglobulin, telomeric, centromeric, segment duplicated regions or were shorter than 3000 bp.

Focal Copy Number Calling

We added the ploidy inferred via Control-FREEC to the consensus CNV segmentation file and used the ploidy and copy number values to define gain and loss values broadly at the chromosome level. We used bedtools coverage [35,36] to add cytoband status using the UCSC cytoband file [37,38]. The output status call fractions, which are values of the loss, gain and callable fractions of each cytoband region, were used to define dominant status at the cytoband-level. The weighted means of each status call fraction were calculated using band length. We used the weighted means to define the dominant status at the chromosome arm-level.

A status is considered dominant if more than half of the region was callable and the status call fraction was greater than 0.9 for that region. The 0.9 threshold was chosen to ensure that the dominant status fraction call is greater than the remaining status fraction calls in a region.

We also wanted to define focal copy number units to avoid calling adjacent genes in the same cytoband or arm as copy number losses or gains where it would be more appropriate to call the broader region a loss or gain. For the determination of the most focal units, we first considered the dominant status calls at the chromosome arm-level. If the chromosome arm dominant status was not clearly defined as a gain or loss (and was callable) we looked to include the cytoband-level status call. Similarly, if a cytoband dominant status call was not clearly defined as a gain or loss (and was callable) we looked to include the gene-level status call. To obtain the gene-level data, we used the mergeByOverlaps function [39] from the IRanges package [40] to find overlaps between the segments in the consensus CNV file and the exons in the GENCODE v27 annotation file [27].

Somatic Structural Variant Calling (WGS samples only)

We used Manta SV [34] 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 [41] and the workflow can be found in the D3b GitHub repository.

Gene Expression

Abundance Estimation

We used STAR [42] v2.6.1d to align paired-end RNA-seq reads. This output was used for all subsequent RNA analysis. We used Ensembl GENCODE 27 [27], "Comprehensive gene annotation" as a reference. We used RSEM [43] v1.3.1 for both FPKM and TPM transcript- and gene-level quantification. We also added a second method of quantification using kallisto [44] v0.43.1. This method differs in that it uses pseudoaligments using FASTQ reads directly to the aforementioned GENCODE 27 reference.

Gene Expression Matrices with Unique HUGO Symbols

Algorithms that perform gene set enrichment, molecular subtyping, or immune-profiling, for example, require an RNA-seq gene expression matrix as input, with HUGO gene symbols as row names and sample names as column names. There is a small proportion of gene symbols that map to multiple Ensembl gene identifiers (in GENCODE v27, 212 gene symbols map to 1866 Ensembl gene identifiers), termed multi-mapped gene symbols.

We first removed genes with no expression from the RSEM abundance data using a cut-off of FPKM > 0 in at least 1 sample across the PBTA cohort. We computed the mean FPKM across all samples per gene and for each multi-mapped gene symbol, we chose the Ensembl identifier corresponding to the maximum mean FPKM with the goal of choosing the identifier that best represented the expression of the gene. After collapsing gene identifiers, there were a total of 46,400 unique expressed genes in the poly-A dataset and a total of 53,011 unique expressed genes remaining in the stranded dataset. More detail can be found in the `collapse-rnaseq` analysis module.

Immune Profiling/Deconvolution

We used the R package immunedeconv [45,46] to deconvolute, quantify and compare various immune cell types across 21 histologies from the PBTA cohort with xCell [47] and CIBERSORT [48] in the stranded and poly-A collapsed FPKM RNA-seq datasets (`immune-deconv` analysis module). Both methods allow between samples (inter-sample), between cell types (intra-sample) and between cancer type (inter-histology) comparisons.

Gene Set Variation Analysis

We performed Gene Set Variation Analysis (GSVA) [49] on collapsed, log2-transformed RSEM FPKM data using the GSVA Bioconductor package [50] with setting `mx.diff=TRUE` to obtain Gaussian-distributed scores (`gene-set-enrichment-analysis` analysis module) for each of the MSigDB hallmark gene sets [51]. We compared GSVA scores among histology groups (`short_histology`) using ANOVA and subsequent Tukey tests; p-values were Bonferroni-corrected for multiple hypothesis testing.

Dimension reduction

We applied Uniform Manifold Approximation and Projection (UMAP) [52] to log2-transformed FPKM data using the umap R package [53]. We set the number of neighbors to 15 (`transcriptomic-dimension-reduction` analysis module).

RNA Fusion Calling and Prioritization

Gene fusion detection

We set up Arriba v1.1.0 and STAR-Fusion 1.5.0 [54] 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 poly-A 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 in frame 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 [55], COSMIC genes, and/or known TCGA fusions from curated references. *MYBL1* [56], *SNCAIP* [57], *FOXR2* [58], *TTYH1* [59], and *TERT* [60,61,62,63] were added to the oncogene list and *BCOR* [58] and *QKI* [64] 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.

Mutational Signatures

We obtained weights for signature sets by applying deconstructSigs [65,66] to consensus SNVs with the BSgenome.Hsapiens.UCSC.hg38 annotations [67]. We estimated how many mutations contributed to each signature for each sample using each sample's signature weights. Weights for signatures fall in the range zero to one inclusive. For a given sample and signature combination, we estimated the number of contributing mutations per Mb of the genome by multiplying the signature weight by the total number of trinucleotide mutations identified by deconstructSigs and then dividing by the size of the effectively surveyed genome.

$$\frac{\#\text{ of contributing mutations}}{Mb} = \frac{\text{weight} * \sum \#\text{ Trinucleotide mutations}}{\text{Size in Mb of effectively surveyed genome}}$$

These results do not include signatures with small contributions; deconstructSigs drops signature weights that are less than 6% [65]. We used these methods to calculate signature scores for each

sample with both COSMIC [68] and Alexandrov et al, 2013 [69] signature sets.

PBTA Tumor Mutation Burden

We consider tumor mutation burden (TMB) to be the number of consensus SNVs per *effectively surveyed* base of the genome.

$$\text{TMB} = \frac{\backslash\#\text{ of coding sequence SNVs}}{\text{Size in Mb of }\{\backslash\text{em effectively surveyed}\}\text{ genome}}$$

We used the total number coding sequence consensus SNVs for the numerator and the size of the intersection of the regions considered by Lancet, Strelka2, and Mutect2 with coding regions (CDS from GENCODE v27 annotation [27]) as the denominator.

TCGA Tumor Mutation Burden

We calculated tumor mutation burden in TCGA using MC3 mutation calls [70] for TCGA brain-related tumor projects including: LGG (lower-grade glioma) [71], GBM (glioblastoma multiforme) [72], and PCPG (pheochromocytoma and paraganglioma) [73]. The MC3 project provided an exome BED file. All SNVs fell within these regions. We considered the regions covered by the MC3 BED file (based on GENCODE v19 annotation [74]) to have been effectively surveyed.

Clinical Data Harmonization

WHO Classification of Disease Types

The `pathology_diagnosis` field in the `pbta-histologies.tsv` file contains the diagnosis denoted from the patient's pathology report. The `integrated_diagnosis` field in the `pbta-histologies.tsv` file includes updates to `pathology_diagnosis` and these changes are documented in the `Notes`. For instance, any diagnosis denoted as "Other" in `pathology_diagnosis` was modified to capture the pathology report diagnosis in `integrated_diagnosis`. Additionally, `pathology_diagnosis` was modified to `integrated_diagnosis` if the presence of specific molecular alterations defined a biospecimen as having an alternate diagnosis. The `broad_histology` denotes the broad 2016 WHO classification [75] 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
age_at_diagnosis_days	Patient age at diagnosis in days	numeric
age_last_update_days	Patient age at the last clinical event/update in days	numeric
aliquot_id	External aliquot identifier	variable

Meta data	Definition	Possible values
broad_composition	Broad classification of sample type	cell-line;cyst;non-tumor;tumor
broad_histology	Broad WHO 2016 classification of cancer type	text
cancer_predispositions	Reported cancer predisposition syndromes	text
cohort	Scientific cohort	CBTTC;PNOC
composition	Sample composition	Derived Cell Line;Not Reported;Peripheral Whole Blood;Saliva;Solid Tissue
ethnicity	Patient reported ethnicity	text
experimental_strategy	Sequencing strategy	WGS;WXS;RNA-Seq;Panel
germline_sex_estimate	Predicted sex of patient based on germline X and Y ratio calculation (described in methods)	Female;Male;Unknown
glioma_brain_region	Brain region for all tumors classified as LGAT or HGAT	midline (Thalamus
`integrated_diagnosis`	Updated and/or integrated molecular diagnosis	text
Kids_First_Biospecimen_ID	KidsFirst biospecimen identifier	BS_########
Kids_First_Participant_ID	KidsFirst patient identifier	PT_########
molecular_subtype	Molecular subtype defined by WHO 2016 guidelines	text
normal_fraction	Theta2 normal DNA fraction estimate	numeric

Meta data	Definition	Possible values
Notes	Free text field describing changes from `pathology_diagnosis` to `integrated_diagnosis` or manner in which molecular_subtype was determined	text
OS_days	Overall survival in days	numeric
OS_status	Overall survival status	DECEASED;LIVING
parent_aliquot_id	External identifier combining sample_id, sample_type, aliquot_id, and sequencing_strategy for some samples	text
pathology_diagnosis	Reported patient diagnosis from pathology reports	text
primary_site	Bodily site(s) from which specimen was derived	text
race	Patient reported race	text
reported_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 `integrated_diagnosis`	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	Control-FREEC 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, following World Health Organization 2016 classification criteria [75].

Medulloblastoma (MB) subtypes SHH, MYC, Group 3, and Group 4 were predicted using an RNA expression classifier on the RSEM FPKM data.

High-grade glioma (HGG) subtypes were derived using the criteria below (additional details in the analysis README):

1. If any sample contained an H3F3A K28M, HIST1H3B K28M, HIST1H3C K28M, or HIST2H3C K28M mutation and no BRAF V600E mutation, it was subtyped as `DMG, H3K28`.
2. If any sample contained an HIST1H3B K28M, HIST1H3C K28M, or HIST2H3C K28M mutation and a BRAF V600E mutation, it was subtyped as `DMG, H3 K28, BRAF V600E`.
3. If any sample contained an H3F3A G35V or G35R mutation, it was subtyped as `HGG, H3 G35`.
4. If any high-grade glioma sample contained an IDH1 R132 mutation, it was subtyped as `HGG, IDH`.
5. If a sample was initially classified as HGAT, had no defining histone mutations, and a BRAF V600E mutation, it was subtyped as `BRAF V600E`.
6. All other high-grade glioma samples that did not meet any of these criteria were subtyped as `HGG, H3 wildtype`.

Non-MB and non-ATRT embryonal (`Embryonal tumor` in the `broad_histology` column of the metadata pbta-histologies.tsv) subtypes were derived using the criteria below [76,77,78,79]. Additional details can be found in the analysis notebook.

1. Any RNA-seq biospecimen with *LIN28A*a overexpression, plus a *TTYH1* fusion (5' partner) with a gene adjacent or within the C19MC miRNA cluster and/or copy number amplification of the C19MC region was subtyped as `ETMR, C19MC-altered` (Embryonal tumor with multilayer rosettes, chromosome 19 miRNA cluster altered).
2. Any RNA-seq biospecimen with *LIN28A*a overexpression, a *TTYH1* fusion (5' partner) with a gene adjacent or within the C19MC miRNA cluster but no evidence of copy number amplification of the C19MC region was subtyped as `ETMR, NOS` (Embryonal tumor with multilayer rosettes, not otherwise specified).
3. Any RNA-seq biospecimen with a fusion having a 5' *MN1* and 3' *BEND2* or *CXXC5* partner were subtyped as `CNS HGNET-MN1` (Central nervous system (CNS) high-grade neuroepithelial tumor with *MN1* alteration).
4. Non-MB and non-ATRT embryonal tumors with internal tandem duplication of *BCOR* were subtyped as `CNS HGNET-BCOR` (CNS high-grade neuroepithelial tumor with *BCOR* alteration).
5. Non-MB and non-ATRT embryonal tumors with over-expression and/or gene fusions in *FOXR2* were subtyped as `CNS NB-FOXR2` (CNS neuroblastoma with *FOXR2* activation).
6. Non-MB and non-ATRT embryonal tumors with *CIC-NUTM1* or other *CIC* fusions, were subtyped as `CNS EFT-CIC` (CNS Ewing sarcoma family tumor with *CIC* alteration).
7. Non-MB and non-ATRT embryonal tumors that did not fit any of the above categories were subtyped as CNS Embryonal, NOS (CNS Embryonal tumor, not otherwise specified).

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 [80] 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.

The Open Pediatric Brain Tumor Atlas

This section will introduce the dataset (e.g., the histologies represented and what data types are included; Figure 1A-B) and the process for contributing analytical code and to the manuscript (Figure 1C-D).

A

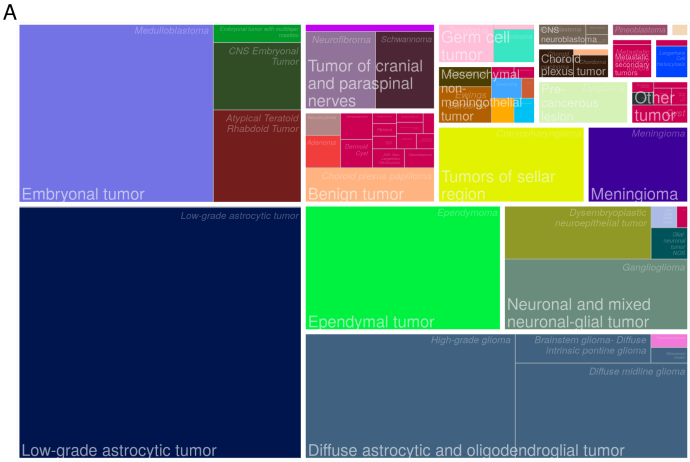

Figure 1: An overview of the OpenPBTA project. A) The distribution of unique participant samples across short histologies and integrated diagnoses.

Landscape of Genomic Alterations

The OncoPrint will provide a visualization of the genomic alterations found in the analyses implemented throughout the OpenPBTA project.

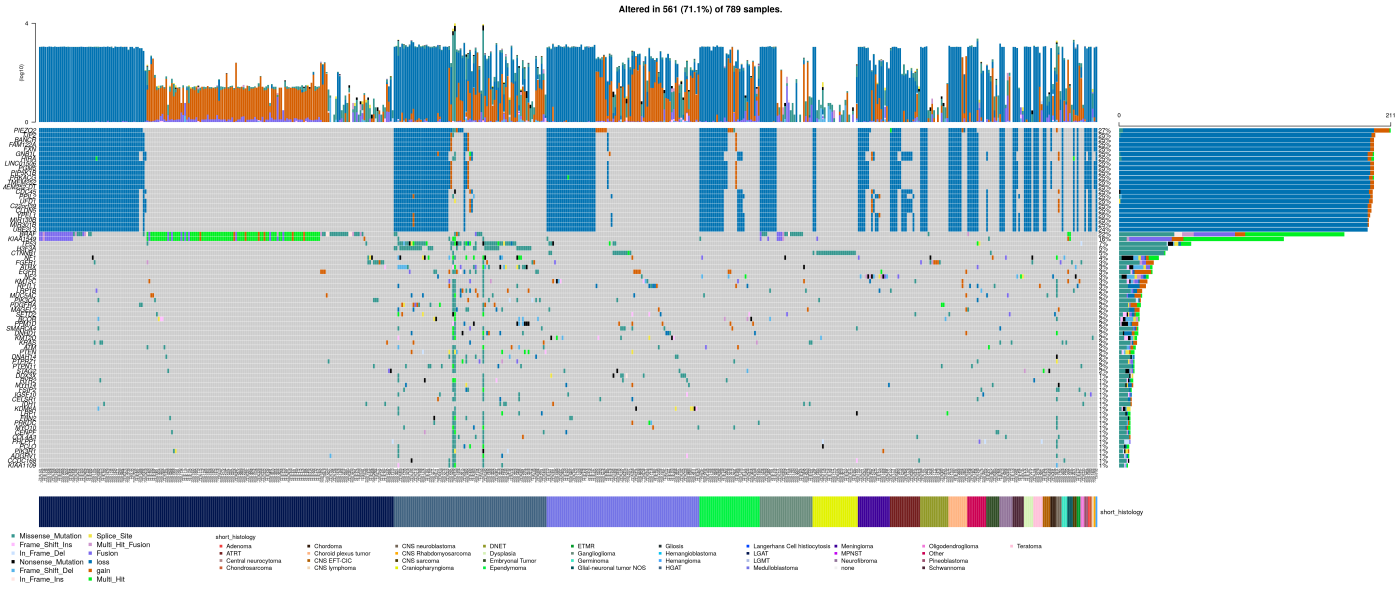

Figure 2: OncoPrint displaying genes most frequently altered across the OpenPBTA dataset. Genes include the top 50 most frequently mutated genes and the top 20 genes with copy number alterations. Samples were filtered to primary samples or, when no primary sample from an individual participant was available, a randomly selected sample with whole genome sequencing (WGS) data. Sample histology (`short_histology`) is displayed in the annotation bar at the bottom of the plot.

Landscape of Mutational Processes

This section summarizes the mutational landscape of the pediatric brain tumor samples of this dataset. Figure 1A shows the tumor mutation burden as compared to adult TCGA brain-related tumors. Figure 3B-C show concordance of these samples with mutational signatures from COSMIC [68] and Alexandrov et al, 2013 [69] signature sets.

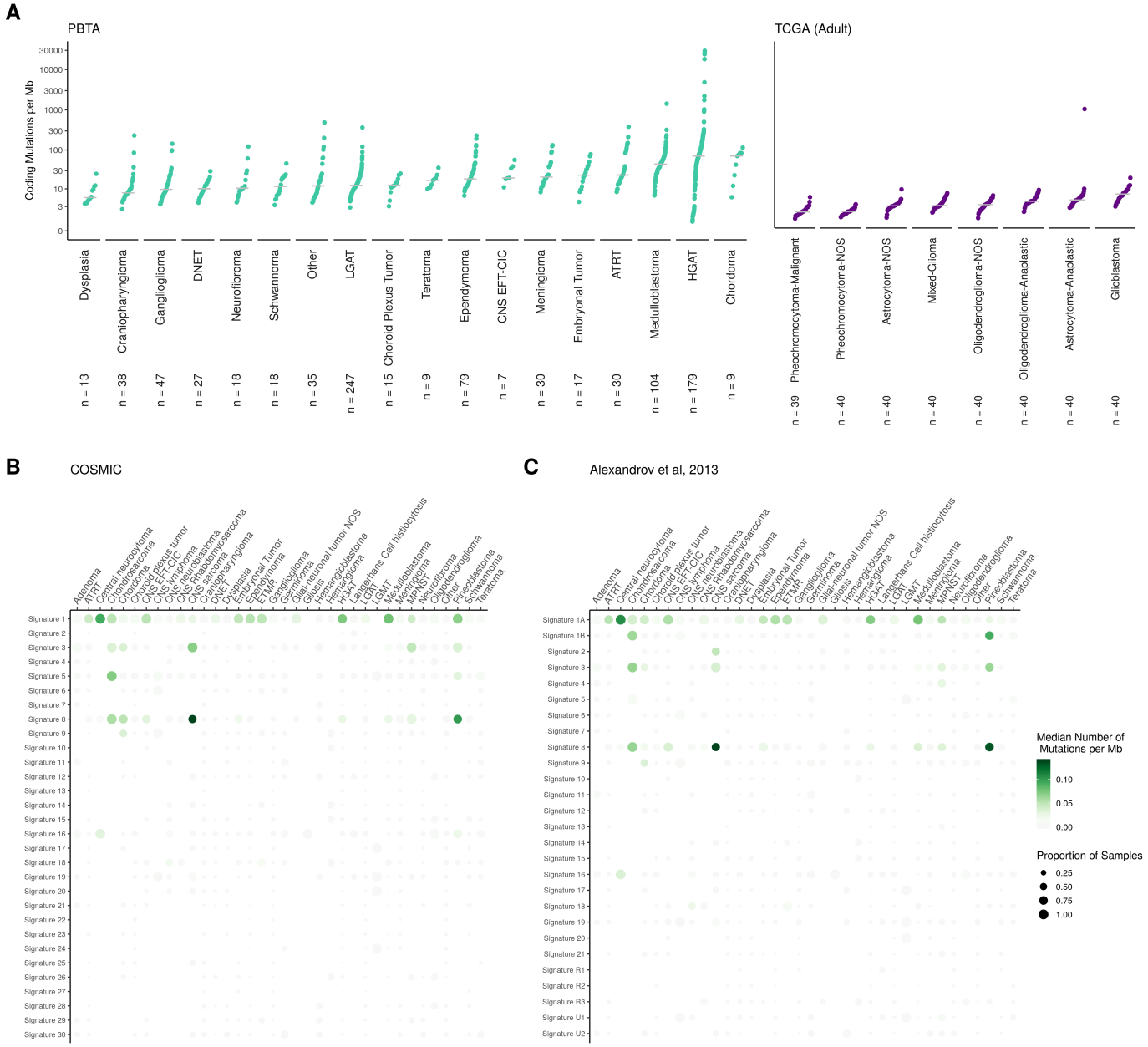

Figure 3: Mutational Landscape

Copy Number Variant Overview

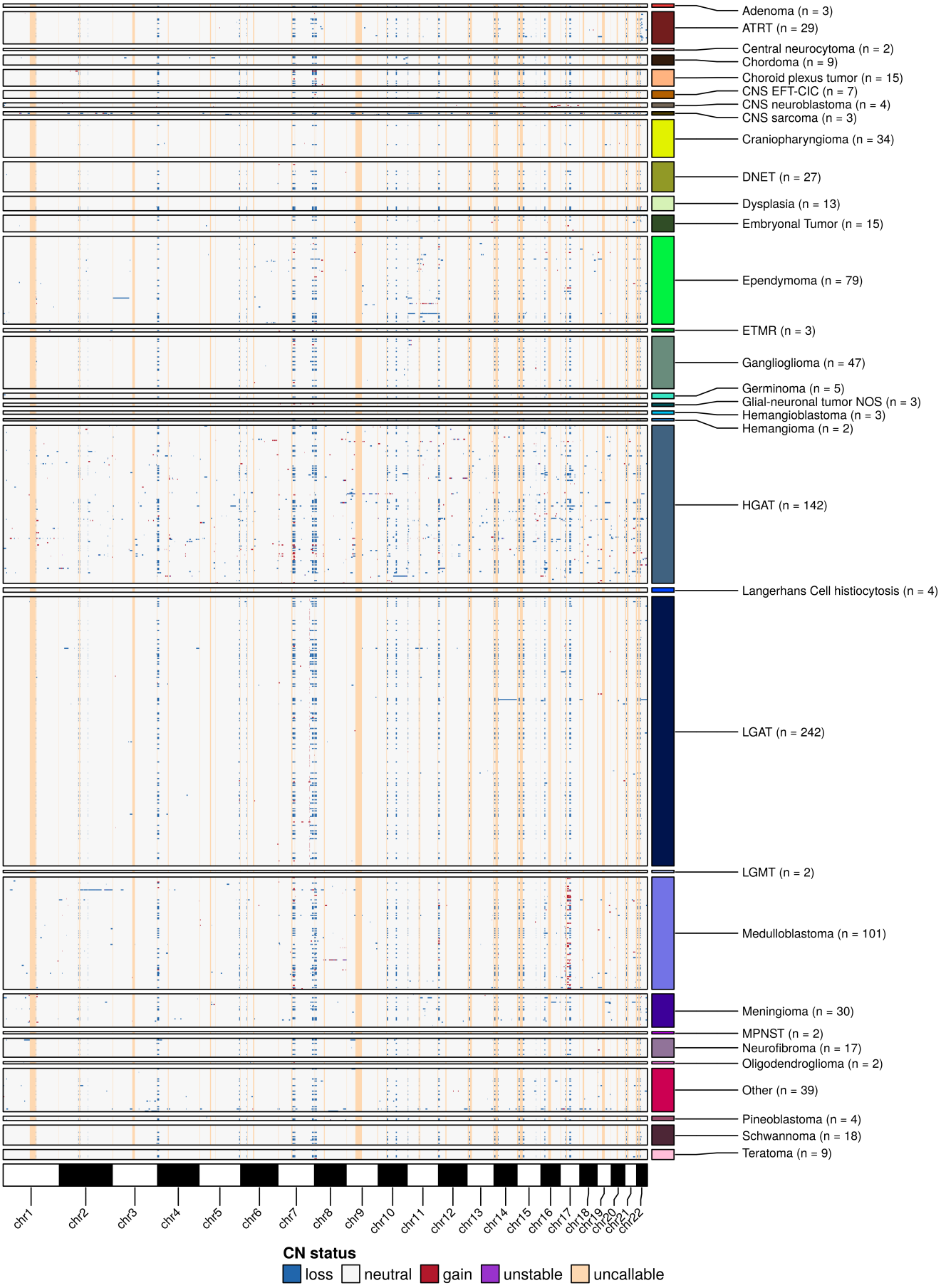

Figure 4: This figure shows dominant statuses for small copy number consensus segments (< 10 Mb) across the genome, where each square represents a ~1 Mb binned section of the genome. A dominant status is declared if one status is >75% coverage. Unstable indicates multiple non-neutral statuses totaling coverage >75%. Copy number segments longer than 10 Mb have been removed from the figure for easier interpretability.

Recurrence and co-occurrence of mutations

This section will discuss the genes and regions that are repeatedly mutated within and between cancer types. The occurrence of mutations in affecting particular genes, separated by tumor type is shown in Figure 5A, with significant co-occurrence across all types illustrated in Figure 5B.

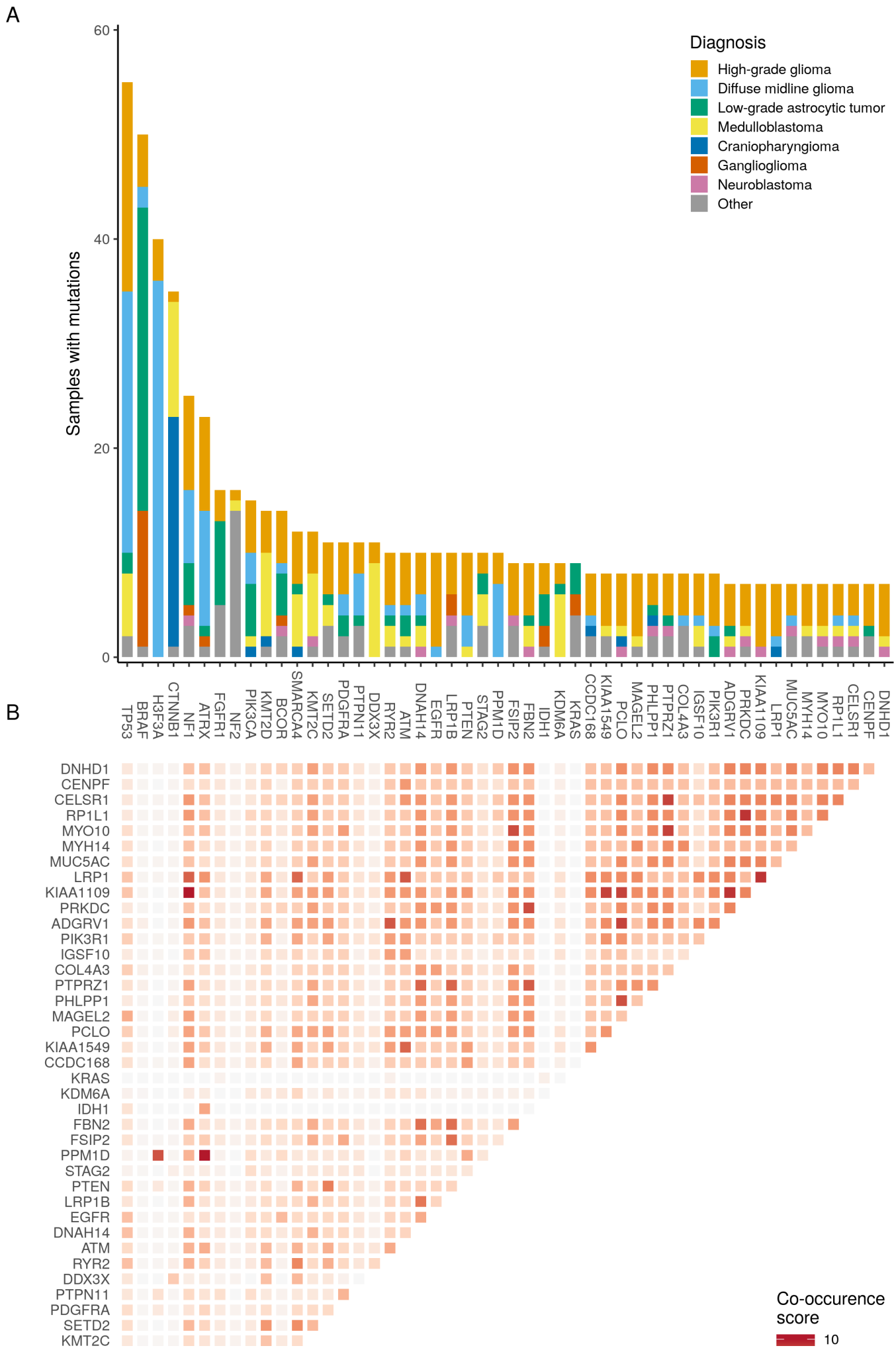
A
Samples with mutations
60
40
20
0
Diagnosis
High-grade glioma
Diffuse midline glioma
Low-grade astrocytic tumor
Medulloblastoma
Craniopharyngioma
Ganglioglioma
Neuroblastoma
Other
TP53
BRAF
H3F3A
CTNNB1
NF1
ATRX
FGFR1
NF2
PIK3CA
KMT2D
BCOR
SMARCA4
KMT2C
SETD2
PDGFRA
PTPN11
DDX3X
RYR2
ATM
DNAH14
EGFR
LRP1B
PTEN
STAG2
PPM1D
FSIP2
FBN2
IDH1
KDM6A
KRAS
CCDC168
KIAA1549
PCLO
MAGEL2
PHLPP1
PTPRZ1
COL4A3
IGSF10
PIK3R1
ADGRV1
PRKDC
KIAA1109
LRP1
MUC5AC
MYH14
MYO10
RP1L1
CELSR1
CENPF
DNHD1
B
DNHD1
CENPF
CELSR1
RP1L1
MYO10
MYH14
MUC5AC
LRP1
KIAA1109
PRKDC
ADGRV1
PIK3R1
IGSF10
COL4A3
PTPRZ1
PHLPP1
MAGEL2
PCLO
KIAA1549
CCDC168
KRAS
KDM6A
IDH1
FBN2
FSIP2
PPM1D
STAG2
PTEN
LRP1B
EGFR
DNAH14
ATM
RYR2
DDX3X
PTPN11
PDGFRA
SETD2
KMT2C
Co-occurence score
10

Figure 5: Occurrence and co-occurrence of nonsynonymous mutations for the 50 most commonly mutated genes across all tumor types. A) Counts of nonsynonymous mutations, colored by tumor type (as defined in `integrated_diagnosis`). B) Co-occurrence and mutual exclusivity of nonsynonymous mutations between genes. The co-occurrence score is defined as $I(-\log_{10}(P))$ where P is defined by Fisher's exact test and I is 1 when mutations co-occur more often than expected and -1 when exclusivity is more common.

Transcriptomic landscape

This section will discuss the overall structure of the transcriptome data, pathway analysis, and immune deconvolution. A dimension reduction plot, a heatmap of GSVA scores [49], and a heatmap of xCell fraction values [47] are shown in Figure 6.

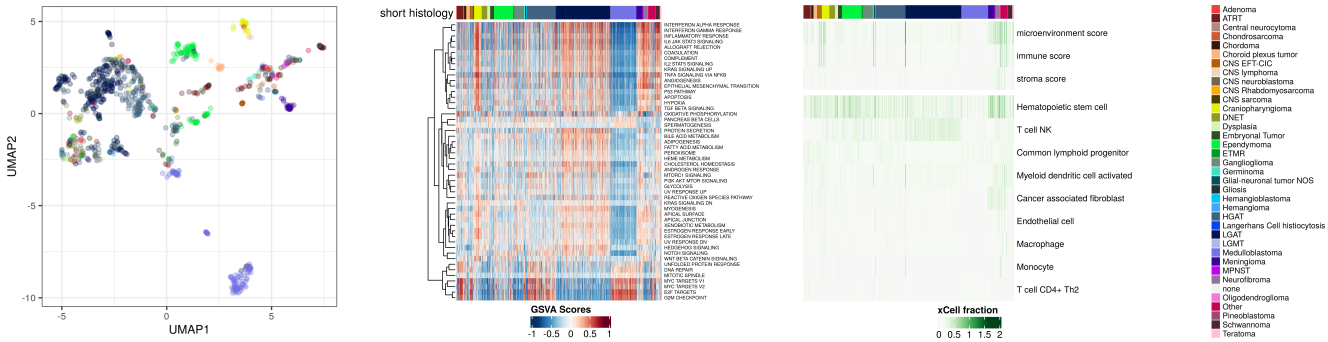

Figure 6: Transcriptomic overview (stranded data only) A) First two dimensions from Uniform Manifold Approximation and Projection (UMAP) of sample transcriptome data. Points are colored by `short_histology` of the samples they represent. B) Heatmap of GSVA scores for Hallmark gene sets with significant differences, with samples ordered by `short_histology`. C) Heatmap of xCell fraction values for scores and cell types with high variance, with samples ordered by `short_histology`.

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, … Children's Brain Tumor Tissue Consortium
   bioRxiv (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 (2013-05-28) https://arxiv.org/abs/1303.3997

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

7. **GATK**https://gatk.broadinstitute.org/hc/en-us

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 (2015-06-15) https://doi.org/gfzsfw
   DOI: 10.1093/bioinformatics/btv098 · PMID: 25697820 · PMCID: PMC4765878

10. **Scaling accurate genetic variant discovery to tens of thousands of samples**
    Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, … Eric Banks
    bioRxiv (2018-07-24) https://doi.org/ggmrvr
    DOI: 10.1101/201178

11. **Broad Genome References**https://s3.amazonaws.com/broad-references/broad-references-readme.html

12. **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 (2017-06-20) https://doi.org/f9xrq4
DOI: 10.1093/nar/gkx193 · PMID: 28369524 · PMCID: PMC5499645

13. **parklab/NGSCheckMate**
Park Lab at Harvard Medical School
(2020-05-27) https://github.com/parklab/NGSCheckMate

14. **Framework for determining accuracy of RNA sequencing data for gene expression profiling of single samples**
Holly C. Beale, Jacquelyn M. Roger, Matthew A. Cattle, Liam T. McKay, Katrina Learned, A. Geoffrey Lyle, Ellen T. Kephart, Rob Currie, Du Linh Lam, Lauren Sanders, … Olena M. Vaske
bioRxiv (2019-07-30) https://doi.org/gghzj8
DOI: 10.1101/716829

15. **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

16. **Calling Somatic SNVs and Indels with Mutect2**
David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, Lee Lichtenstein
bioRxiv (2019-12-02) https://doi.org/ggntwv
DOI: 10.1101/861054

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 · PMID: 30271907 · PMCID: PMC6123722

18. **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 (2016-06-20) https://doi.org/f8v6qz
DOI: 10.1093/nar/gkw227 · PMID: 27060149 · PMCID: PMC4914105

19. **broadinstitute/gatk**
GitHub
https://github.com/broadinstitute/gatk

20. **AstraZeneca-NGS/VarDictJava**
AstraZeneca - NGS Team
(2020-05-27) https://github.com/AstraZeneca-NGS/VarDictJava

21. **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
bioRxiv (2019-04-30) https://doi.org/ggc9vx
DOI: 10.1101/623702

22. **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

23. **mskcc/vcf2maf**
Memorial Sloan Kettering
(2020-05-25) https://github.com/mskcc/vcf2maf

24. **AlexsLemonade/OpenPBTA-analysis**
GitHub
https://github.com/AlexsLemonade/OpenPBTA-analysis

25. https://github.com/AlexsLemonade/OpenPBTA-analysis/blob/master/analyses/snv-callers/plots/comparison/upset_plot.png

26. https://github.com/AlexsLemonade/OpenPBTA-analysis/blob/master/analyses/snv-callers/plots/comparison/vaf_violin_plot.png

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

28. **Maftools: efficient and comprehensive analysis of somatic variants in cancer**
Anand Mayakonda, De-Chen Lin, Yassen Assenov, Christoph Plass, H. Phillip Koeffler
Genome Research (2018-11) https://doi.org/gfmnwf
DOI: 10.1101/gr.239244.118 · PMID: 30341162 · PMCID: PMC6211645

29. **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 (2012-02-01) https://doi.org/ckt4vz
DOI: 10.1093/bioinformatics/btr670 · PMID: 22155870 · PMCID: PMC3268243

30. **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 (2011-01-15) https://doi.org/c6bcps
DOI: 10.1093/bioinformatics/btq635 · PMID: 21081509 · PMCID: PMC3018818

31. **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

32. **Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data**
Layla Oesper, Gryte Satas, Benjamin J. Raphael
Bioinformatics (2014-12-15) https://doi.org/f6rmgt
DOI: 10.1093/bioinformatics/btu651 · PMID: 25297070 · PMCID: PMC4253833

33. **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/dzhjqh
DOI: 10.1186/gb-2011-12-4-r41 · PMID: 21527027 · PMCID: PMC3218867

34. **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 (2016-04-15) https://doi.org/gf3ggb
DOI: 10.1093/bioinformatics/btv710 · PMID: 26647377

35. **coverage — bedtools 2.29.2 documentation** https://bedtools.readthedocs.io/en/latest/content/tools/coverage.html

36. **BEDTools: a flexible suite of utilities for comparing genomic features**
Aaron R. Quinlan, Ira M. Hall
Bioinformatics (2010-03-15) https://doi.org/cmrms3
DOI: 10.1093/bioinformatics/btq033 · PMID: 20110278 · PMCID: PMC2832824

37. http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/cytoBand.txt.gz

38. **The UCSC Genome Browser database: extensions and updates 2013**
Laurence R. Meyer, Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Robert M. Kuhn, Matthew Wong, Cricket A. Sloan, Kate R. Rosenbloom, Greg Roe, Brooke Rhead, ... W. James Kent
Nucleic Acids Research (2012-11-15) https://doi.org/f4jr4v
DOI: 10.1093/nar/gks1048 · PMID: 23155063 · PMCID: PMC3531082

39. **findOverlaps-methods function | R Documentation** https://www.rdocumentation.org/packages/IRanges/versions/2.6.1/topics/findOverlaps-methods

40. **Software for Computing and Annotating Genomic Ranges**
Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, Vincent J. Carey
PLoS Computational Biology (2013-08-08) https://doi.org/f5cmfg
DOI: 10.1371/journal.pcbi.1003118 · PMID: 23950696 · PMCID: PMC3738458

41. **AnnotSV: an integrated tool for structural variations annotation**
Véronique Geoffroy, Yvan Herenger, Arnaud Kress, Corinne Stoetzel, Amélie Piton, Hélène Dollfus, Jean Muller
Bioinformatics (2018-10-15) https://doi.org/gdcsh3
DOI: 10.1093/bioinformatics/bty304 · PMID: 29669011

42. **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 (2013-01) https://doi.org/f4h523
DOI: 10.1093/bioinformatics/bts635 · PMID: 23104886 · PMCID: PMC3530905

43. **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

44. **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

45. **Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology.**
Gregor Sturm, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H Fridman, Markus List, Tatsiana Aneichyk
Bioinformatics (Oxford, England) (2019-07-15) https://www.ncbi.nlm.nih.gov/pubmed/31510660
DOI: 10.1093/bioinformatics/btz363 · PMID: 31510660 · PMCID: PMC6612828

46. **icbi-lab/immunedeconv**
ICBI
(2020-05-26) https://github.com/icbi-lab/immunedeconv

47. **xCell: digitally portraying the tissue cellular heterogeneity landscape**
Dvir Aran, Zicheng Hu, Atul J. Butte
Genome Biology (2017-11-15) https://doi.org/gckmjs
DOI: 10.1186/s13059-017-1349-1 · PMID: 29141660 · PMCID: PMC5688663

48. **Robust enumeration of cell subsets from tissue expression profiles**
Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, Ash A Alizadeh
Nature Methods (2015-03-30) https://doi.org/gcp3f5
DOI: 10.1038/nmeth.3337 · PMID: 25822800 · PMCID: PMC4739640

49. **GSVA: gene set variation analysis for microarray and RNA-Seq data**
Sonja Hänzelmann, Robert Castelo, Justin Guinney
BMC Bioinformatics (2013) https://doi.org/gb8vx5
DOI: 10.1186/1471-2105-14-7 · PMID: 23323831 · PMCID: PMC3618321

50. **GSVA**
Justin Guinney [Aut, Cre], Robert Castelo [Aut], Joan Fernandez[Ctb]
Bioconductor (2017) https://doi.org/ggxrqs
DOI: 10.18129/b9.bioc.gsva

51. **The Molecular Signatures Database Hallmark Gene Set Collection**
Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, Pablo Tamayo
Cell Systems (2015-12) https://doi.org/gf78hq
DOI: 10.1016/j.cels.2015.12.004 · PMID: 26771021 · PMCID: PMC4707969

52. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**
Leland McInnes, John Healy, James Melville
arXiv (2018-02-09) https://arxiv.org/abs/1802.03426v2

53. https://cran.r-project.org/package

54. **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
bioRxiv (2017-03-24) https://doi.org/gf5pc5
DOI: 10.1101/120295

55. **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 · PMID: 29425488

56. **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 · PMID: 23633565 · PMCID: PMC3657784

57. **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 · PMID: 22832581 · PMCID: PMC3683624

58. **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 · PMID: 26919435 · PMCID: PMC5139621

59. **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, Albena 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 · PMID: 24316981

60. **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 · PMID: 26523776

61. **Recurrent pre-existing and acquired DNA copy number alterations, including focal *TERT* gains, 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 (2013-12) https://doi.org/f5gd94
DOI: 10.1002/gcc.22110 · PMID: 24123354

62. **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 · PMID: 25481751

63. **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 (2017-03-07) https://doi.org/ggdhk2
DOI: 10.1007/s11910-017-0722-5 · PMID: 28271343

64. **MYB-QKI rearrangements in angiocentric glioma drive tumorigenicity through a tripartite mechanism**
Pratiti Bandopadhayay, 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 · PMID: 26829751 · PMCID: PMC4767685

65. **deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution**
Rachel Rosenthal, Nicholas McGranahan, Javier Herrero, Barry S. Taylor, Charles Swanton
Genome Biology (2016-02-22) https://doi.org/f8bdsq
DOI: 10.1186/s13059-016-0893-4 · PMID: 26899170 · PMCID: PMC4762164

66. **raerose01/deconstructSigs**
GitHub
https://github.com/raerose01/deconstructSigs

67. **BSgenome.Hsapiens.UCSC.hg38**
Bioconductor
http://bioconductor.org/packages/BSgenome.Hsapiens.UCSC.hg38/

68. **COSMIC - Catalogue of Somatic Mutations in Cancer**
Cosmic
https://cancer.sanger.ac.uk/cosmic

69. **Signatures of mutational processes in human cancer**
Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, … ICGC PedBrain
Nature (2013-08-14) https://doi.org/f22m2q
DOI: 10.1038/nature12477 · PMID: 23945592 · PMCID: PMC3776390

70. **Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines**
Kyle Ellrott, Matthew H. Bailey, Gordon Saksena, Kyle R. Covington, Cyriac Kandoth, Chip Stewart, Julian Hess, Singer Ma, Kami E. Chiotti, Michael McLellan, … Armaz Mariamidze
Cell Systems (2018-03) https://doi.org/gf9twn
DOI: 10.1016/j.cels.2018.03.002 · PMID: 29596782 · PMCID: PMC6075717

71. **Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas**
The Cancer Genome Atlas Research Network
New England Journal of Medicine (2015-06-25) https://doi.org/f7f82c
DOI: 10.1056/nejmoa1402121 · PMID: 26061751 · PMCID: PMC4530011

72. **The Somatic Genomic Landscape of Glioblastoma**
Cameron W. Brennan, Roel G.W. Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R. Salama, Siyuan Zheng, Debyani Chakravarty, J. Zachary Sanborn, Samuel H. Berman, … Roger McLendon
Cell (2013-10) https://doi.org/f5dbzj
DOI: 10.1016/j.cell.2013.09.034 · PMID: 24120142 · PMCID: PMC3910500

73. **Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma**
Lauren Fishbein, Ignaty Leshchiner, Vonn Walter, Ludmila Danilova, A. Gordon Robertson, Amy R. Johnson, Tara M. Lichtenberg, Bradley A. Murray, Hans K. Ghayee, Tobias Else, … Erik Zmuda
Cancer Cell (2017-02) https://doi.org/f9vcmf
DOI: 10.1016/j.ccell.2017.01.001 · PMID: 28162975 · PMCID: PMC5643159

74. **GENCODE - Human Release 19**https://www.gencodegenes.org/human/release_19.html

75. **The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary**
David N. Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, David W. Ellison
Acta Neuropathologica (2016-05-09) https://doi.org/f8mspx
DOI: 10.1007/s00401-016-1545-1 · PMID: 27157931

76. **Embryonal Tumors of the Central Nervous System in Children: The Era of Targeted Therapeutics.**
David E Kram, Jacob J Henderson, Muhammad Baig, Diya Chakraborty, Morgan A Gardner, Subhasree Biswas, Soumen Khatua
Bioengineering (Basel, Switzerland) (2018-09-23) https://www.ncbi.nlm.nih.gov/pubmed/30249036
DOI: 10.3390/bioengineering5040078 · PMID: 30249036 · PMCID: PMC6315657

77. **LIN28A, a sensitive immunohistochemical marker for Embryonal Tumor with Multilayered Rosettes (ETMR), is also positive in a subset of Atypical Teratoid/Rhabdoid Tumor (AT/RT)**
Shilpa Rao, R. T. Rajeswarie, T. Chickabasaviah Yasha, Bevinahalli N. Nandeesh, Arimappamagan Arivazhagan, Vani Santosh
Child's Nervous System (2017-07-25) https://doi.org/ggnpkn
DOI: 10.1007/s00381-017-3551-6 · PMID: 28744687

78. **Childhood Medulloblastoma and Other Central Nervous System Embryonal Tumors Treatment (PDQ®)–Health Professional Version - National Cancer Institute**(2008-02-13)
https://www.cancer.gov/types/brain/hp/child-cns-embryonal-treatment-pdq

79. **DNA Methylation Profiling for Diagnosing Undifferentiated Sarcoma with Capicua Transcriptional Receptor (CIC) Alterations**
Evelina Miele, Rita De Vito, Andrea Ciolfi, Lucia Pedace, Ida Russo, Maria Debora De Pasquale, Angela Di Giannatale, Alessandro Crocoli, Biagio De Angelis, Marco Tartaglia, … Giuseppe Maria Milano
International Journal of Molecular Sciences (2020-03-06) https://doi.org/ggnn7x
DOI: 10.3390/ijms21051818 · PMID: 32155762 · PMCID: PMC7084764

80. **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