1-11-2024

Applications of Bioinformatics for Exploring the Genomic Landscape of Cancer and Characterizing Treatment Response for Precision Oncology

Sharon Freshour
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Human and Statistical Genetics

Dissertation Examination Committee:
Malachi Griffith, Chair
Obi L. Griffith, Co-Chair
Vivek K. Arora
Megan A. Cooper
Allegra A. Petti
Joshua B. Rubin
Ting Wang

Applications of Bioinformatics for Exploring the Genomic Landscape of Cancer and Characterizing Treatment Response for Precision Oncology

by

Sharon L. Freshour

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2024
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

To Malachi and Obi, thank you so much for all your support and mentorship throughout my PhD. I am truly lucky to have had two great PIs whose enthusiasm for research is genuinely inspiring.

To my thesis committee, Allegra, Josh, Megan, Ting, and Vivek, thank you for all the expertise and advice you have shared with me; I have learned so much from all of you. Thank you especially to Josh and Vivek for giving me the opportunity to work with you on your projects.

To the members of Griffith lab, many of whom have become good friends, thank you for all the support, guidance, and good memories. To the past members of Griffith lab, especially Cody, Huiming, Kelsy, Megan, and Zach, thank you for welcoming me into Griffith lab and making me feel like I belonged there from day one. To the current members of Griffith lab, especially Evelyn, Kartik, Mariam, and My, thank you for making it so much fun to come into lab every day.

To the friends I made outside of Griffith lab, especially Emily, Kara, and Sidi, thank you for all the good meals, movie nights, and adventures. Grad school would not have been the same without you.

Lastly, to my family, thank you so much for your endless support, encouragement, and belief in me, not just during grad school, but throughout my life.

Sharon L. Freshour

Washington University in St. Louis

May 2024

ABSTRACT OF THE DISSERTATION

Applications of Bioinformatics for Exploring the Genomic Landscape of Cancer and Characterizing Treatment Response for Precision Oncology

by

Sharon Laura Freshour

Doctor of Philosophy in Biology and Biomedical Sciences

Human and Statistical Genetics

Washington University in St. Louis, 2024

Malachi Griffith, Chair

Obi L. Griffith, Co-Chair

While the integration of next-generation sequencing and bioinformatics into cancer research has been immeasurably useful for advancing the field, there are still many gaps in understanding the genomic landscape of cancer due to the complexity of cancer development and evolution. Likewise, there are still many challenges for identifying and applying effective therapeutics for cancer treatment. Thus, there remains a need to continue characterizing the genomic landscape of cancer, to explore the evolution of cancer, and to understand how tumors respond to treatment. The projects described in this dissertation cover several applications of next-generation sequencing, bioinformatic analysis, and database resources for cancer research and precision oncology. The first project described involves the use of single-cell RNA sequencing, along with whole genome and whole exome sequencing, to characterize the genomics of a mouse model of bladder cancer and explore mechanisms of response to immune checkpoint blockade treatment. Differential expression and gene set enrichment analysis performed on multiple immune and

stromal cell types within the tumor microenvironment revealed that IFN-g response in endothelial cells was upregulated in response to treatment, suggesting that IFN-g response in endothelial cells may play a crucial role in treatment response. Functional analysis confirmed that knocking out *IFNgR1* in endothelial cells negated the treatment response observed in *IFNgR1*-intact mice, further indicating that IFN-g response in endothelial cells is a key mediator of effective treatment response. The second project described focuses on the use of whole genome sequencing to explore the landscape of copy number variation in over 250 pediatric brain tumors across four diagnosis groups (ATRT, Ependymoma, High-Grade Glioma, and Medulloblastoma). This analysis revealed that copy number alterations within pediatric brain tumors were quite common and could be extensive, particularly in Ependymoma, High-Grade Glioma, and Medulloblastoma. Exploration of the relationship between copy number variant (CNV) burden and overall survival suggested that CNV burden may have prognostic value within specific diagnosis groups. Likewise, analysis of the relationship between recurrently altered genomic regions and overall survival indicated that particular recurrent alterations within certain diagnosis groups could be significantly correlated with changes in overall survival. Lastly, the third project described covers updates to the Drug-Gene Interaction database (DGIdb, dgidb.org) implemented in the DGIdb 4.0 release. This resource allows researchers to explore drugs, genes, and known or predicted drug-gene interactions gathered from multiple sources in a single, harmonized database. The updates presented here (DGIdb 4.0) include the addition of several new sources, integration with crowdsource efforts, and improvements to the normalization and grouping of interactions. Collectively, this dissertation describes the use of next-generation sequencing, bioinformatic analysis approaches, and development of public

database resources for cancer research that have improved the understanding and treatment of cancer.

Chapter 1: Introduction

1.1 Next-generation sequencing and cancer research

Cancer is a complex disease that encompasses numerous different cancer types and is one of the leading causes of death worldwide (Bray et al., 2021). As of 2020, there were an estimated 19.3 million new cancer cases and 10 million cancer-related deaths per year worldwide (Sung et al., 2021). While many of the hallmark functional capabilities that cells acquire to drive development and proliferation of cancer cells have been defined (e.g. avoiding immune destruction, resisting cell death, and activating invasion and metastasis), the specific mechanisms by which these functions are acquired can vary greatly across, and within, cancer types (Hanahan, 2022).

The classic model of cancer development involves a multistep process by which somatic mutations (e.g. single nucleotide variants, insertions, deletions, translocations) and/or epigenetic changes (e.g. DNA methylation, histone modifications) accrue over time as a result of environmental stressors and improperly repaired DNA damage (Diori Karidio & Sanlier, 2021; Miles & Tadi, 2023). Understanding the exact combination of somatic events that leads to the initiation and progression of cancer presents a considerable challenge in understanding cancer biology and developing effective therapies.

Traditional cancer treatments typically involve surgical resection (for solid tumors) and the use of chemotherapy and/or radiotherapy, which generally function by inducing DNA damage that halts cell growth and division, eventually leading to cell death, through exposure to radiation or specific classes of drugs (e.g. alkylating agents) (Liu et al., 2021; Tilsed et al., 2022). While these therapies have contributed greatly to improving survival and are still widely used, these therapies (particularly chemotherapy) also cause damage to normal cells due to their non-

selective mechanisms of action and are associated with severe, sometimes life-threatening, side effects (Min & Lee, 2022; Schirrmacher, 2019).

The drawbacks associated with these therapies have led to an interest in discovering and developing targeted therapy methods which can more selectively target cancer cells and reduce harm to normal cells. These therapies can include molecular agents (e.g. small molecule inhibitors), hormonal agents (e.g. estrogen receptor blockers), and immunotherapies (e.g. immune checkpoint blockade inhibitors or personalized neoantigen vaccines). However, for these targeted therapies to be utilized effectively, the cancer must harbor the appropriate targets (e.g. specific mutations or receptors) (Min & Lee, 2022). This has led to the rise of precision oncology, which involves using genomic and transcriptomic profiling of a patient's cancer to detect targetable mutations (or other features) and then developing a personalized treatment plan for the patient based on these profiles (Tsimberidou et al., 2020). One of the driving forces behind many of the recent breakthroughs in the understanding of cancer biology and the discovery of therapeutic targets has been the use of next-generation sequencing for profiling the genomic and transcriptomic landscape of cancer.

1.1.1 Generating next-generation sequencing data

Next-generation sequencing (NGS) is often defined as massively parallel sequencing that allows millions of DNA (or RNA) fragments to be sequenced at the same time (Satam et al., 2023). More generally, NGS can be defined as post-Sanger sequencing technologies (Datto & Lundblad, 2016). These technologies are often split into different generations based on their short- or long-read capabilities. Short-read technologies, i.e. technologies that capture ~50 to ~800 bp reads, are often called second-generation technologies, while long-read technologies,

i.e. technologies that can capture read lengths up to 10,000 bp or more, are often considered to be third-generation technologies (Hu et al., 2021). Many NGS technologies can also be broadly separated into DNA sequencing (e.g. bulk whole genome or whole exome) and RNA sequencing (e.g. bulk mRNA) technologies. The basic steps for generating data from any NGS technology always include nucleic acid library preparation and sequencing.

Library preparation for NGS typically starts with a fragmentation step. During the fragmentation step, DNA or RNA is either physically fragmented, e.g. through sonication, or enzymatically fragmented, e.g. via digestion by DNase I (Head et al., 2014). Generally, the method and target size for fragmentation are determined based on the NGS technology being used for sequencing (Hess et al., 2020). For RNA sequencing, additional steps are included after fragmentation to capture the target RNA, e.g. to remove rRNA and capture mRNA, and reverse transcribe the RNA fragments into cDNA fragments (Hu et al., 2021). Once a sample has been fragmented and size selected, adapter sequences are then ligated to the 3' and 5' ends of the fragments. These adapters contain known sequences that are used to capture the fragments for amplification and sequencing (Datto & Lundblad, 2016).

After fragmentation and adapter ligation, many NGS technologies have an amplification step, typically using a polymerase chain reaction (PCR) method. This PCR amplification step is used to ensure that there are enough input fragments to be reliably detected and sequenced (Hess et al., 2020). When sequencing is performed, it can be either paired-end or single-end sequencing. In paired-end sequencing, the fragments are sequenced from both the 3' and 5' ends, which typically allows improved sequencing accuracy compared to single-end sequencing (Pervez et al., 2022).

1.1.2 Aligning NGS data

Once raw sequencing reads have been generated, the next step is generally to align reads to a reference genome, such as GRCh38 for humans or GRCm39 for mice. The goal of most common alignment algorithms (e.g. BWA-Mem or Bowtie) is to align reads with a balance of speed and accuracy (H. Ye et al., 2015). For example, BWA-Mem starts by utilizing a Burrows-Wheeler Transform (a commonly used method) to index and compress the reference genome, which allows more efficient searching of the genome during sequence alignment and reduces the memory requirements for alignment (H. Li & Durbin, 2009). For alignment, BWA-Mem uses a maximal exact match method to attempt to produce accurate alignments. This method involves looking for the longest possible substring in a read that matches a sequence of the reference genome exactly and then using the Smith-Waterman algorithm to extend the alignment with affine gap penalties (*bwa.1*). Some regions of the genome, such as regions containing repeat nucleotide sequences, are difficult to accurately align reads to, particularly when using short-read sequencing. The accuracy of alignments is also affected by the limitations of the reference genome itself. For example, GRCh38 contains millions of bases that are unresolved, represented by N's, and numerous regions that are unplaced, represented separately from the main chromosomes (Aganezov et al., 2022).

While improvements in sequencing technologies, the use of paired-end sequencing, and targeting a higher sequencing depth can help improve the overall accuracy of alignments, the caveats listed above should be considered when interpreting sequencing and alignment data for downstream analysis. Typically, before beginning downstream analysis, aligned reads also go through additional processing to further improve the quality and accuracy of the final alignment data. These steps can include marking reads that appear to be duplicate reads introduced during

PCR amplification and recalibrating the base quality scores (i.e. the scores which reflect the confidence that each base was called correctly) to account for technical biases that can occur during preparation or sequencing of samples (Zverinova & Guryev, 2022).

1.1.3 Variant calling with NGS data

The next step in the analysis of next-generation sequencing data is usually to perform variant calling. In the context of cancer informatics, this ideally will involve comparison of a tumor sample and a matched normal sample. Normal samples can be used to identify inherited germline variants, such as single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and other variants that are expected to be found in all cells and are not unique to tumor cells. Tumor samples can be used to identify somatic variants, such as single nucleotide variants (SNVs), indels, copy number variants (CNVs), structural variants (SVs), and other genomic rearrangement events that are acquired specifically in tumor cells. While cancer is more commonly thought to be driven by acquired somatic mutations, there are examples of heritable cancer syndromes that are linked to germline mutations, such as BRCA1 and BRCA2 mutations, which can greatly increase a person's risk of developing cancer (Petrucelli et al., 1998).

The appropriate tools to use for calling variants depend on whether you are calling germline or somatic variants and what variant type you are trying to identify. For example, GATK's HaplotypeCaller is specifically designed for detection of germline SNPs and indels, while Mutect2 is designed for calling somatic SNVs and indels (Benjamin et al., 2019; Poplin et al., 2018). As different variant callers can produce different results, the best practice for calling a high-quality set of SNVs and indels generally includes running multiple variant callers (such as Mutect2, Varscan, and Strelka2) and identifying the variants called by multiple callers (Z. Chen

et al., 2020). Separate tools are used for calling larger variants, including copy number variants, e.g. CNVkit, and structural variants, e.g. Manta (X. Chen et al., 2016; Talevich et al., 2016). Once variant calling results have been generated, manual review is often done to verify the quality and accuracy of variant calls (Barnell et al., 2019).

1.1.4 Large-scale NGS databases for cancer research

The rise of next-generation sequencing has allowed great improvement in understanding the mutational landscape of cancer (e.g. in the discovery of driver mutations or in the molecular subtyping of cancer types) and in the adoption of precision oncology approaches for diagnosing and treating cancer (e.g. in the use of targeted treatments based on the specific alterations found in a tumor) (Avila & Meric-Bernstam, 2019; Nones & Patch, 2020). While initial studies often focused on sequencing samples from individual patients, the success of these individual studies gave rise to large-scale projects that aimed to sequence and analyze genomic and transcriptomic data from thousands of samples across numerous cancer types (Mardis, 2019). The Cancer Genome Atlas (TCGA) is perhaps the most notable example of these efforts, generating multi-omic pan-cancer and single-cancer datasets of over 20,000 cancer samples across more than 30 cancer types (Cancer Genome Atlas Research Network et al., 2013). In addition to reporting their findings, these large-scale sequencing and analysis projects often make their omics data publicly available (or available upon request), so the research community can leverage these databases to explore their own projects in cancer research.

1.1.5 Survival analysis in cancer research

Generically, survival analysis refers to methods of analysis used to assess the time until an event of interest occurs (i.e. the survival time). In particular, survival analysis is used for studies in which only some individuals have experienced the event of interest at the conclusion of the study. This means the survival time for the individuals who did not experience the event of interest is unknown. This is referred to as censoring. Censoring can also include patients who dropped out of a study before its conclusion, but were still alive at the time they dropped out (Clark et al., 2003). In the context of cancer research, the events can include time to death (i.e. overall survival) or time to recurrence or relapse (i.e. disease-free or event-free survival). Survival analysis is often used to compare event-free or overall survival between treatment groups (e.g. untreated patients versus patients that received a specific treatment) or groups of patients with or without certain disease characteristics (e.g. specific mutations or disease subtypes). Two of the most commonly used survival analysis methods are Kaplan-Meier survival estimates and the Cox proportional hazards model.

Kaplan-Meier survival probability is estimated based on both censored and uncensored survival times and relies on the assumption that each event occurs independently of all other events. Because of this independence of events, the cumulative probability of being alive at a given time point can be calculated by multiplying the probabilities of being alive at each previous time point. For each individual time point, the survival probability can be calculated by taking the number of subjects still alive (and still enrolled in the study) at the end of the time point and dividing it by the number alive (and enrolled) at the beginning of the time point (Goel et al., 2010; E. L. Kaplan & Meier, 1958). Kaplan-Meier curves are typically used to visualize the survival probability over time and estimate metrics such as median survival.

The Cox proportional hazards model is a regression model that can be used to explore the relationship between survival and one or more variables. The Cox model is often favored for multivariate analysis, since it allows the potential impact of each individual variable in the model to be assessed in the context of all the other variables included in the model (Bewick et al., 2004). Functions used to fit Cox proportional hazards regression models (e.g. the *coxph* function in R) output regression coefficients for each variable, p-values for each variable (which indicate whether the variable is significantly associated with changes in survival), and hazard ratios (the exponential of the corresponding coefficient) for each variable. For categorical variables, the hazard ratio represents the change in the probability of experiencing an event (e.g. death) at a specific time in a given group versus the baseline group at that same time (Deo et al., 2021). For continuous variables, the hazard ratio represents the change in the probability of experiencing an event given a unit of increase (or decrease) in the continuous variable (Abd ElHafeez et al., 2021). Hazard ratios greater than 1 indicate that the probability of an event increases for a given group or given unit change (even accounting for any other variables in the model), while hazard ratios less than 1 indicate the probability of an event decreases for a given group or given unit change (again, even accounting for any other variables in the model). Chapter 3 will discuss the use of whole genome sequencing (WGS) data to characterize the landscape of copy number alteration in pediatric brain tumors and the use of survival analysis methods to evaluate the potential prognostic value of copy number alteration.

1.2 Single-cell RNA sequencing and cancer research

In addition to bulk next-generation sequencing (NGS), such as whole genome sequencing or targeted RNA sequencing, NGS also encompasses single-cell sequencing methods, such as

single-cell RNA sequencing. While bulk RNA sequencing methods can provide a picture of the global average gene expression across a sample containing hundreds of thousands of cells and a mixture of cell types, bulk RNA sequencing can obscure signals from rarer cell types within the sample (X. Li & Wang, 2021). In contrast, single-cell RNA sequencing (scRNA-seq) can distinguish individual cells and provide gene expression profiles for each of those individual cells which allows identification of the various cell types, including rare cell types, present within a sample. With regard to cancer research, scRNA-seq can be particularly useful for characterizing heterogeneity within a tumor or identifying signals from different populations of immune or stromal cells within the tumor microenvironment (Huang et al., 2023). Nevertheless, scRNA-seq has several limitations. In particular, scRNA-seq data can be quite sparse (typically capturing at most a few thousand genes per cell), meaning the complete expression profiles of cells may not be captured (X. Li & Wang, 2021). This sparsity can also inflate the number of genes with zero expression and make it difficult to distinguish between true biological lack of expression and lack of detection due to technical limitations (Kharchenko, 2021).

1.2.1 Generating scRNA-seq data

Currently, one of the most commonly used high-throughput methods for generating scRNA-seq data is droplet-based microfluidics (e.g. Drop-Seq or 10x Genomics). For example, the 10x Genomics sequencing platform relies on a technology called GEM (Gel bead in EMulsion). In addition to the adapters and oligo dT primers used to capture mRNAs and initiate reverse transcription, these beads also contain barcode sequences and unique molecular identifier (UMI) sequences (Zheng et al., 2017). Barcodes serve to identify individual cells, i.e. all transcripts from the same cell will have the same barcode, while UMIs serve to identify individual

transcripts within a cell. Within a channel on a microfluidics plate, individual beads are combined with individual cells (from a single-cell suspension) in oil droplets, where the process of reverse transcription takes place to produce cDNA.

Once this process is completed, the cDNA from individual cells is pooled into one sample, which then undergoes library preparation, PCR amplification, sequencing, and alignment, similar to the preparation and alignment discussed for bulk DNA and RNA samples (Zheng et al., 2017). For the 10x Genomics workflow, reads are then aligned to a reference genome using STAR, an alignment tool designed specifically for aligning RNA sequencing data (Dobin et al., 2013). This workflow also includes steps to correct for sequencing errors in barcode and UMI sequences, which allows the cells and transcripts corresponding to the corrected barcodes and UMIs to be included in downstream analysis. The final output for the pipeline is a gene-barcode matrix. This matrix includes all valid barcodes (i.e. barcodes corresponding to cells) and unique UMI counts (i.e. UMI counts excluding PCR duplicates) for each gene in each cell (Zheng et al., 2017). This gene-barcode matrix is the input used for most downstream analysis of scRNA-seq.

1.2.2 Preparing scRNA-seq data for analysis

Before beginning analysis of scRNA-seq data, several pre-processing steps are usually performed, including filtering low-quality cells, detecting and removing doublets (or multiplets), and correcting for ambient RNA contamination. The two main metrics often used to filter low-quality cells are low feature count (typically gene count) per cell, which can indicate low underlying quality of the cell itself or technical issues sequencing the cell, and high percentage of

UMIs mapping to mitochondrial genes per cell, which can indicate a dying cell (Heumos et al., 2023).

Doublets or multiplets refer to droplets that have captured more than one cell, making it appear as if the transcripts from multiple cells are originating from one cell. When multiple cells of the same cell type are captured, homotypic multiplets are formed. Homotypic multiplets are typically more difficult to detect. When multiple cells of different cell types are captured, heterotypic multiplets are formed. Heterotypic multiplets are generally easier to detect and what most tools for detection focus on identifying (Xi & Li, 2021). Tools (e.g. DoubletFinder) that create artificial doublets from the user's input dataset and then identify doublets by comparing the distance between the expression profiles of artificial doublets and the expression profiles of each cell in the input dataset generally perform best at accurately detecting doublets (McGinnis et al., 2019; Xi & Li, 2021).

Ambient RNA refers to extracellular RNA contamination in a single cell suspension that likely originates from burst cells, such as those that have undergone apoptosis. Ambient RNA can be captured in a droplet along with the RNA from an intact cell and be barcoded as if it originates from that intact cell. When this ambient RNA is captured, it can artificially inflate the UMI counts for the gene expression profile of the actual cell or make it appear as if genes which are not truly expressed in the cell are being expressed (S. Yang et al., 2020). Several tools have been developed to identify and correct for ambient RNA expression in scRNA-seq data. These tools typically either quantify ambient RNA contamination on a global basis for an entire sample (e.g. SoupX) or on a per cell basis for each cell in a sample (e.g. DecontX) (S. Yang et al., 2020; Young & Behjati, 2020).

After cells have been filtered, the final preprocessing step before beginning any analysis is typically to normalize and scale the raw UMI counts from the gene-barcode matrix for the remaining cells. For example, the default method used by Seurat (one of the most commonly used tools for scRNA-seq analysis) normalizes UMI counts for each gene within a cell by the total UMI count for that cell, scales the counts by a factor of 10,000, and then log-transforms that value. To scale the data across cells, Seurat shifts the expression of each gene such that the mean expression across cells is 0 and the variance across cells is 1 (Stuart et al., 2019). Once these preprocessing steps are completed, the scRNA-seq dataset can be used for downstream analysis.

1.2.3 Clustering and visualizing scRNA-seq data

The first steps for analysis of scRNA-seq data usually involve linear dimensionality reduction, clustering, and visualization of clustering using non-linear dimensionality reduction. One common method used for linear dimensionality reduction is principal component analysis (PCA), which is used to identify highly variable sets of genes (i.e. the principal components) across the dataset. Ideally, the top principal components identified will serve as a fairly comprehensive compression of the dataset. The number of principal components to use for further processing and analysis of the dataset can be chosen based on the amount of variance within the dataset explained by each component (Satija et al., 2015; Shalek et al., 2013). For example, if the first ten principal components capture the majority of variance in the dataset, then principal components one through ten can be used for downstream analysis.

The principal components chosen can then be used to construct a k-nearest neighbors (KNN) graph. In the KNN graph, edges are drawn between cells with similar expression profiles based on their Euclidean distance and weights for these edges are calculated based on the Jaccard

similarity (a measure of shared overlap in local neighbors) between the cells (Levine et al., 2015). Clusters are identified from this graph using modularity optimization methods like the Louvain algorithm. Modularity refers to the measure of the quality of the division of a network (or graph) into communities (or clusters) (Newman, 2006). The Louvain algorithm starts with placing each cell in its own cluster and then moves each cell into shared clusters based on the moves that increase the quality measure the most. These clusters are then combined to make an aggregated graph. The clusters in this aggregated graph are then moved to generate the highest quality measure again. These steps are repeated iteratively until the quality can no longer be increased and the final clustering is produced (Traag et al., 2019).

Next, once clustering has been generated, clusters are often visualized in 2D using non-linear dimensionality reduction methods like t-distributed stochastic neighbor embeddings (tSNEs) or uniform manifold approximation and projections (UMAPs) (Clarke et al., 2021). tSNE projections generally preserve the local relationships between similar cells, meaning that tSNEs can be useful for visualizing distinct groups of cells, but are not good for interpreting the relationships between different clusters. In contrast, UMAPs are generally considered to be better for visualizing the global relationships between clusters of cells (Clarke et al., 2021). However, while these projections may reveal some broad characteristics of the relationships within and across clusters of cells, these projections should not be relied on for interpretation of the underlying biology of a dataset as they are only 2D representations of extremely complex datasets.

1.2.4 Analyzing scRNA-seq data

Another common initial step for analysis of scRNA-seq data is to assign cell types to each individual cell. While this can be done manually based on known marker expression (e.g. using *CD79A* expression to distinguish B cells), many tools exist that automate assigning cell types by leveraging existing datasets that contain expression profiles for annotated cell types (Pasquini et al., 2021). One commonly used tool for automated cell typing is SingleR. SingleR assigns cell type labels based on the Spearman correlation between the expression profiles of individual cells in a dataset and reference expression profiles of cell types sequenced using methods such as microarray or RNA-seq (Aran et al., 2019). In addition to allowing the user to specify their own reference dataset, SingleR includes several built-in, well-curated reference datasets for both human and mouse cell typing. For example, built-in reference data from the Immunological Genome Project (ImmGen) can be used to annotate mouse cell types in SingleR. The ImmGen dataset contains gene expression profiles for hundreds of immune cell types, as well as stromal cell types, characterized using a combination of Affymetrix arrays, ultra-low-input RNA-seq, and scRNA-seq (Heng et al., 2008; Immunological Genome Project, 2020). Once cell types have been assigned to each cell, there are many types of downstream analysis that can be performed within each cell type population (in addition to being performed across the entire dataset).

Some of the next steps for analysis of scRNA-seq data often include identification of mutations within cells (particularly in the context of cancer research), differential expression analysis, and gene set enrichment analysis (Conesa et al., 2016; Petti et al., 2019; Wang et al., 2022). Methods for identifying variants within cells can include methods which leverage sets of known variants called using bulk sequencing to look for evidence of variant-supporting reads within scRNA-seq data at the known variant positions and methods which attempt to call

variants from scRNA-seq data itself. VarTrix is an example of a tool which uses known SNV and indel calls to look for evidence of SNVs and indels in single cell sequencing data generated using 10x Genomics platforms. The input required for VarTrix includes a VCF of pre-called variants, a bam file of reads captured from single cell sequencing, and the cell barcodes from the single cell sequencing. VarTrix then evaluates reads mapping to the pre-called variant positions and outputs matrices containing information on each variant position in each cell. These matrices can be in a binary format, indicating whether a variant was detected and if more reads supported the reference or alternate allele, can contain read counts for the reference and alternate alleles, or can report a variant allele fraction (*Vartrix: Single-Cell Genotyping Tool*). The results from VarTrix can then be used to explore variant-positive cells and cell populations within a single cell dataset. In the context of cancer research, detection of variants within cells can be useful for confidently identifying tumor cell populations and for characterizing subclonal tumor populations.

Differential expression analysis is one of the most common analyses performed for scRNA-seq data. Approaches used for differential expression analysis are often grouped as either pseudo bulk methods that use existing bulk RNA-seq methods (e.g. DESeq2 or edgeR) with aggregated expression values from populations of single cells or as single-cell methods which use expression values of individual cells (e.g. the Wilcoxon rank-sum test) (Soneson & Robinson, 2018; Squair et al., 2021). The Wilcoxon rank-sum (Wilcox) test is one of the most commonly used differential expression testing methods for scRNA-seq and is the default method used by Seurat (Squair et al., 2021). The Wilcox test is a non-parametric test which (in the context of single cell differential expression testing) tests whether the mean expression of a gene across two populations of cells is significantly different, comparing the rank of the expression

values from each population of cells (Das et al., 2021). For each gene tested, Seurat's default differential expression method (the FindMarkers function) reports a p-value indicating whether gene expression was significantly different (per the Wilcox test) between two cell populations of interest, a log2 fold change of the average expression between the two populations, the percent of cells expressing the given gene in each population, and an adjusted p-value, which is corrected for multiple testing using Bonferroni correction. These differential expression results can then be used to explore how specific cell populations are changing in response to perturbations such as treatment methods or cellular stress.

After differential expression results have been generated, another common step is to perform gene set enrichment analysis. Gene set enrichment analysis (GSEA) can be used to assess whether differentially expressed genes are significantly enriched for up- or down-regulated genes that are functionally or biologically related (A. Subramanian et al., 2005). To test for enrichment, differentially expressed genes are compared to published gene sets that have been grouped based on common functionality and/or biology. One commonly used source for these test gene sets is the Molecular Signatures Database (MSigDB), which offers access to thousands of well-curated, well-annotated gene sets generated specifically with gene set enrichment analysis in mind (Liberzon et al., 2011, 2015). While gene set enrichment analysis was originally developed for analysis of microarray data, it is now commonly used to analyze RNA-seq data.

For scRNA-seq data, a pre-ranked GSEA method is typically used. This method involves providing a list of genes that were analyzed for differential expression, ranked by a single metric reflecting the degree of differential expression (often the log2 fold change). The ranked gene list is then tested for either positive enrichment, (e.g. enrichment of upregulated genes) or negative

enrichment (e.g. enrichment of downregulated genes) against a specified category of gene sets (e.g. Hallmark gene sets). For each gene set tested, an enrichment score is calculated using a running-sum statistic. The value for this running-sum statistic is generated by going down the input gene list and increasing the score when a gene in the list is found in a given gene set and decreasing the score when a gene in the list is not found in a given gene set. The significance of the enrichment score for each gene set tested is determined by using permutation testing to generate a null distribution of the enrichment scores and calculating the p-value of the observed enrichment score. A normalized enrichment score is then calculated for each gene set to adjust for the size of the gene set. Finally, a false discovery rate (FDR) is calculated for each normalized enrichment score to correct for testing multiple gene sets (A. Subramanian et al., 2005). Significantly enriched gene sets can often provide insight into the biological and functional mechanisms that are being perturbed by changes in conditions (e.g. control versus treated tumors). Chapter 2 will discuss the use of single-cell RNA sequencing and analysis to characterize response to immune checkpoint blockade treatment in a mouse model of bladder cancer.

1.3 The druggable genome

The concept of "the druggable genome" was first coined in 2002 by Hopkins & Groom and was defined as the set of genes in the human genome that express proteins which can be bound by drug-like molecules (Hopkins & Groom, 2002). Over the years, this concept has grown to include both established druggable targets and predicted druggable targets. There are multiple methods that have been employed to attempt to predict druggability. These have included predicting druggability based on the known druggability of other molecules in the same families,

attempting to predict binding between proteins and drug-like molecules based on 3D protein structures, and developing predictive computational methods, such as machine learning models (Hajduk et al., 2005; Radoux et al., 2022; Raies et al., 2022).

Numerous publications and sources cataloging known and predicted drug-gene interactions have also been released over time. While some of these sources make their information readily accessible, some do not. Additionally, there is often a lack of standardization in the way drugs, genes, and drug-gene interactions are represented across sources. To address these issues, the Drug-Gene Interaction database (DGIdb) was released in 2013 to provide a single curated resource that mines existing drug, gene, and drug-gene interaction sources and presents them in one publicly available database with a user-friendly interface and standardized data model (Griffith et al., 2013). Since its first release, DGIdb has released several updates to include additional sources, update existing sources, improve the user interface, and refine the information presented for drugs, genes, and drug-gene interactions (Cannon et al., 2023; Cotto et al., 2018; Freshour et al., 2021; Wagner et al., 2016). Chapter 4 will discuss updates to DGIdb (DGIdb 4.0) published in 2021.

Chapter 2: Endothelial cells are a key target of IFN-g during response to combined PD-1/CTLA-4 ICB treatment in a mouse model of bladder cancer

2.1 Preamble

The following chapter has been published as a peer reviewed manuscript with the following citation:

Freshour SL, Chen THP, Fisk B, Shen H, Mosior M, Skidmore ZL, Fronick C, Bolzenius JK, Griffith OL, Arora VK, Griffith M. Endothelial cells are a key target of IFN-g during response to combined PD-1/CTLA-4 ICB treatment in a mouse model of bladder cancer. 2023. iScience. DOI: 10.1016/j.isci.2023.107937

As an author of the published manuscript, and in compliance with the editorial policies at iScience, the cited publication is included in full in the following chapter. As the first author of this manuscript, I performed bioinformatic analysis of sequencing data, generated figures, assembled the manuscript for submission, and addressed reviewer responses for final publication. A complete list of author contributions is included within the publication (Chapter 2.7).

2.2 Summary

To explore mechanisms of response to combined PD-1/CTLA-4 immune checkpoint blockade (ICB) treatment in individual cell types, we generated scRNA-seq using a mouse model of invasive urothelial carcinoma with three conditions: untreated tumor, treated tumor, and tumor treated after CD4+ T cell depletion. After classifying tumor cells based on detection of somatic variants and assigning non-tumor cell types using SingleR, we performed differential expression analysis, overrepresentation analysis, and gene set enrichment analysis (GSEA) within each cell type. GSEA revealed that endothelial cells were enriched for upregulated IFN-g response genes when comparing treated cells to both untreated cells and cells treated after CD4+ T cell depletion. Functional analysis showed that knocking out *IFNgR1* in endothelial cells inhibited treatment response. Together, these results indicated that IFN-g signaling in endothelial cells is a key mediator of ICB induced anti-tumor activity.

2.3 Introduction

Bladder cancer is a common malignancy worldwide (6th most common among men and 17th most common among women) and accounts for over 500,000 new cancer diagnoses and 200,000 cancer-related deaths per year (Woods, 2022). While over 95% of bladder cancer cases are classified as urothelial carcinomas, they encompass a range of molecular subtypes, which are primarily distinguished by differential expression of differentiation markers and may predict for response to specific treatments (Compérat et al., 2022; Robertson et al., 2017). Initial diagnosis stages can be broadly grouped into non-muscle invasive (NMIBC), muscle invasive (MIBC), and metastatic disease. About 75% of cases are initially diagnosed as NMIBC, 20% as MIBC, and

the remaining 5% as metastatic. Depending on the initial degree of invasiveness and metastasis, 5-year survival rates can range from 96% to 6% (*Survival Rates for Bladder Cancer*).

Standard treatment recommendations likewise depend on the initial degree of invasiveness as well as the risk stratification of recurrence and progression. NMIBC is typically treated with a transurethral resection followed by either chemotherapy, Bacillus Calmette-Guerin immunotherapy, or radical cystectomy in high-risk cases (Compérat et al., 2022). MIBC is typically treated with neoadjuvant chemotherapy followed by radical cystectomy and, in some cases, adjuvant immunotherapy. Previous research has suggested that response to treatment may differ by subtype. For example, basal/squamous bladder cancers may have better response to neoadjuvant chemotherapy than luminal-infiltrated tumors (Robertson et al., 2017). While neoadjuvant chemotherapy has historically been used most commonly, clinical trials looking at the use of neoadjuvant immune checkpoint blockade (ICB) treatments have shown promise as well (Lenis et al., 2020; *Muscle-Invasive and Metastatic Bladder Cancer*).

Currently, there are several ICB treatments approved by the FDA for treatment of bladder cancer, all of which are either PD-1 or PD-L1 inhibitors (Lopez-Beltran et al., 2021; Rhea & Aragon-Ching, 2021). Initially, these treatments were approved specifically for treatment of advanced disease, targeting patients who were ineligible for cisplatin treatment (Suzman et al., 2019). Over time, ICB use has become more widespread and has been applied across the range of bladder cancer stages from NMIBC to metastatic disease (Albisinni et al., 2021; Kartolo et al., 2021; Lopez-Beltran et al., 2021). While ICB therapy shows great promise for treatment of bladder cancer, there are still many patients who do not receive benefit from ICB treatment. Thus, there remains a need to improve treatment methods, determine which patients will respond

well to treatment, understand mechanisms of response to treatment, and identify potential predictors of response (Lenis et al., 2020).

Clinical trials examining the benefit of PD-1/PD-L1 inhibitors in urothelial carcinoma have found that high IFN-g expression is associated with treatment response, suggesting that IFN-g signatures could serve as a predictor of response (Sharma et al., 2017). Additionally, treatment response has been associated with high expression of *CXCL9* and *CXCL10*, two IFN-g induced chemokines that have been associated with increased T cell infiltration in multiple tumor types (Kohli et al., 2022; Rosenberg et al., 2016). However, these trials did not fully explore how or where IFN-g may be acting to help induce or improve treatment response. Clinical trials have also looked at improving treatment response by combining PD-1/PD-L1 inhibitor treatments with CTLA-4 inhibitor treatments. These trials have shown greater response rates compared to PD-1/PD-L1 monotherapy (Gao et al., 2020; Roviello et al., 2021; Sharma et al., 2016; van Dijk et al., 2020). Nevertheless, the challenges of identifying ideal patients for treatment as well as identifying mechanisms and predictors of response remain.

To study mechanisms of response to combined PD-1/CTLA-4 ICB treatment, we used a murine muscle-invasive urothelial carcinoma cell line generated by exposing mice to 4-hydroxybutyl(butyl)nitrosamine (BBN), which caused them to develop areas of invasive disease. These tumor bearing bladders were then resected and used to propagate an organoid cell line, MCB6C (Sato et al., 2018). Previous analysis showed that MCB6C is responsive to ICB treatments and achieves the best treatment response with combined PD-1/CTLA-4 ICB treatment. Additionally, this previous work showed that treatment response was dependent on CD4+ T cells and not dependent CD8+ T cells, consistent with research showing that CD4 T cells may be the primary mediators of anti-tumor activity in human bladder cancer (Oh et al.,

2020; Sato et al., 2018). Analysis of the MCB6C model also showed that ICB treatment led to an increase of IFN-g producing CD4+ T cells with a Th-1 like phenotype. Neutralizing IFN-g in the tumor negated the anti-tumor activity of combined treatment, indicating that IFN-g was a key mediator of response. Surprisingly, this research showed that knocking out *IFNgR1* in the tumor cells themselves did not affect treatment response, suggesting that IFN-g was mediating treatment response through non-tumoral cells in the tumor microenvironment (TME) (Sato et al., 2018).

To better understand mechanisms of treatment response in this model, we performed single cell RNA sequencing (scRNA-seq) on MCB6C tumors isolated from mice under three conditions: untreated tumor, tumor treated with combined PD-1/CTLA-4 ICB treatment, and tumor treated with combined ICB treatment after CD4+ T cell depletion (Figure 2.1A). For each condition in each replicate, tumors from three mice were resected and pooled to generate single cell suspensions for 10x Genomics 5′ gene expression sequencing as well as B cell receptor (BCR) and T cell receptor (TCR) sequencing (Figure 2.1B, STAR Methods). This sequencing was performed for five biological replicates. In addition to scRNA-seq, whole genome and exome sequencing of the tumor cell line were performed, along with matched normal whole genome and exome sequencing of a tail sample (Figure 2.1C, STAR Methods).

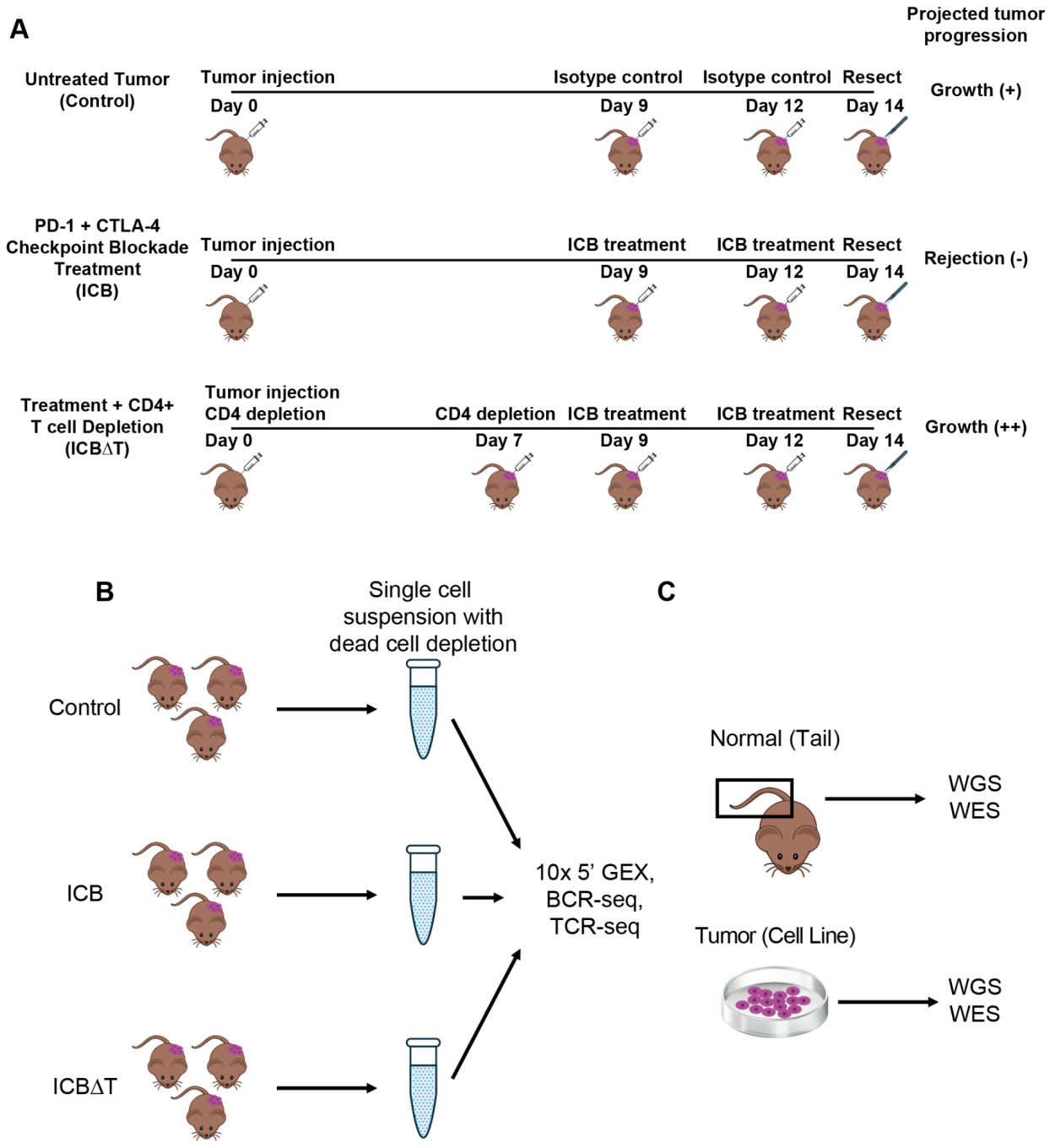


Figure 2.1: Experimental design for single cell RNA and bulk DNA sequencing

(A) *Timelines for generating tumor samples for individual mice for each condition.* **(B)** *Workflow for generating single cell suspensions for single cell RNA sequencing for one of five biological replicates sequenced. For each condition in each replicate, tumors from three*

individual mice were pooled into one suspension and used to create libraries for 10x Genomics 5' single cell gene expression (GEX), B cell receptor (BCR), and T cell receptor (TCR) sequencing. **(C)** *Sources for normal and tumor bulk DNA sequencing. DNA was isolated from a normal mouse tail sample and an MCB6C tumor cell line sample for whole genome sequencing (WGS) and whole exome sequencing (WES). ICB = combined PD-1/CTLA-4 immune checkpoint blockade treatment, ICBΔT = combined PD-1/CTLA-4 immune checkpoint blockade treatment received after CD4+ T cell depletion, Isotype control = rat IgG2a and mouse IgG2b.*

2.4 Results

2.4.1 Bulk DNA sequencing shows that the MCB6C cell line has a high mutation burden, normal ploidy, and a stable genome

Bulk whole genome sequencing (WGS) of the tumor cell line generated over one billion paired reads, 88% of which produced high quality alignments (i.e., had a mapping score of Q20 or greater). Bulk WGS of the normal tail sample produced over 1.1 billion reads, with approximately 91% of reads having high quality alignments. Bulk whole exome sequencing (WES) of the tumor cell line produced over 55 million reads with over 90% of reads having high quality alignments, while WES of the tail sample produced over 77 million reads with over 90% of reads having high quality alignments (Table S1A).

After alignment, we performed somatic variant calling with the WES data and identified 16,449 possible somatic variants, including 16,315 single nucleotide variants (SNVs) and 134 small insertions or deletions (indels), before filtering. These variants were then filtered using several metrics, including total coverage, variant allele frequency (VAF), and consensus across somatic variant callers (STAR Methods). 10,427 variants remained after filtering, of which

10,407 were SNVs and 20 were indels, showing that the MCB6C cell line has a high SNV burden (approximately 4.17 mutations per Mb) consistent with a mutagen induced tumor model. We then characterized the clonality of the cell line by examination of the VAF distribution (STAR Methods) (Zhang et al., 2016). This distribution appeared to be centered close to 50%, with a median VAF of ~ 48.0%, and had a near-normal distribution (Figure 2.2A; Table S1B). However, there was a small group of low VAF (i.e., VAF of 20% or less) variants detected. Out of the 2,128 variants used to assess clonality based on VAF distribution, 31 variants had low VAF values. These variants could represent a subclone within the cell line.

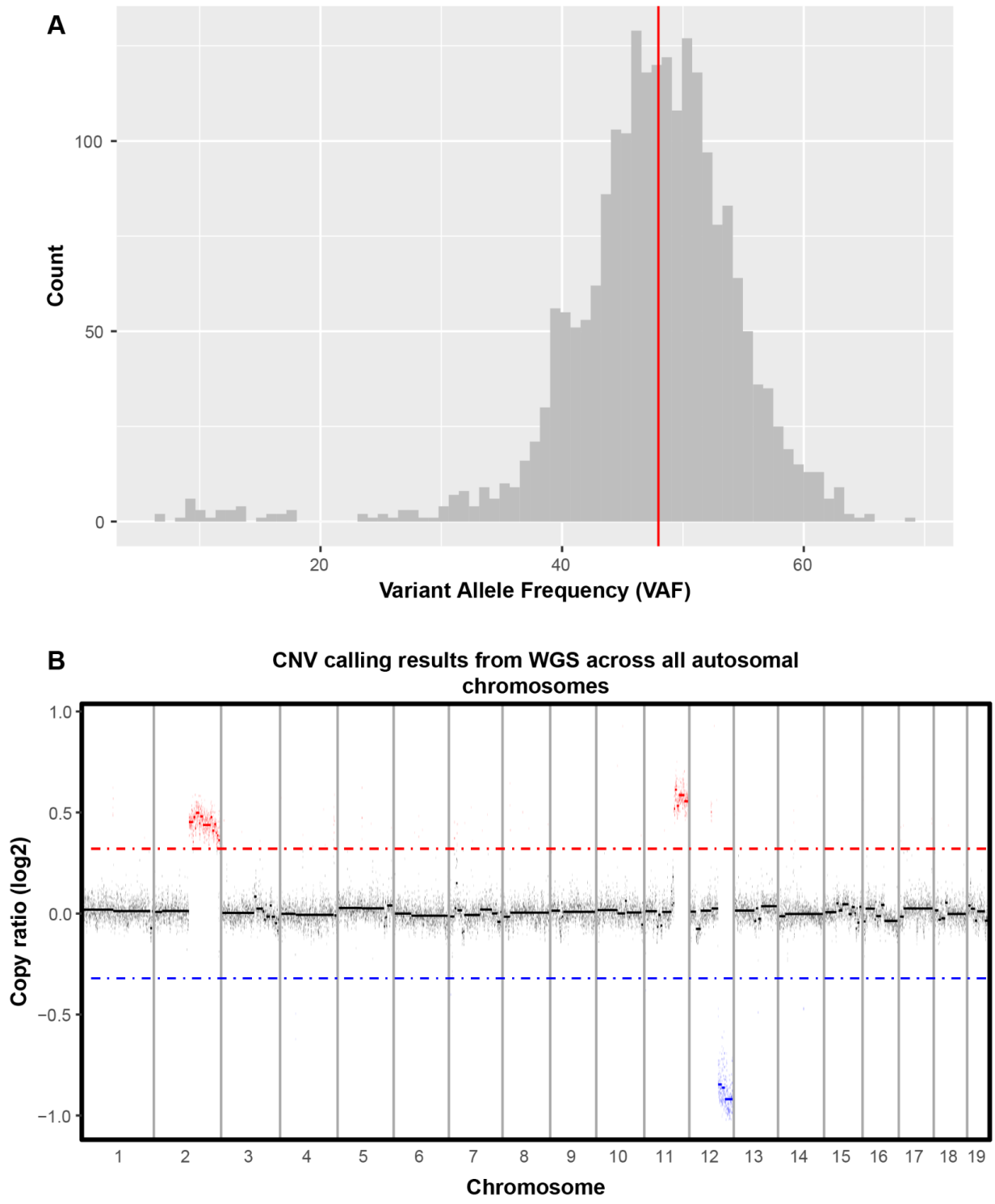


Figure 2.2: Bulk DNA sequencing shows that the MCB6C cell line has a high mutation burden, normal ploidy, and a stable genome

(A) *Distribution of variant allele frequencies (VAFs) for the set of somatic SNVs used to assess clonality. Only variants with 100x normal and tumor coverage, normal VAF = 0%, tumor VAF >5%, and no overlap of CNV regions were used. SNVs were detected using matched tumor and normal whole exome sequencing of the MCB6C cell line. Red line indicates overall median VAF (~48.0%).* **(B)** *Visualization of copy number variants detected using matched tumor and normal whole genome sequencing of the MCB6C cell line. Copy number calling was performed using CNVkit with 100k bin size. Dots represent read count differences in bins. Solid lines represent segments identified by CNVkit using circular binary segmentation. The red dotted line corresponds to a log2 copy ratio of 0.32. Bins and segments falling above this line are classified as copy gains. The blue dotted line corresponds to a log2 copy ratio of −0.32. Bins and segments falling below this line are classified as copy losses. Three chromosomes exhibited large segments of copy number alteration: chr2 (copy gain of ~86.9 Mb), chr11 (copy gain of ~41.2 Mb), and chr12 (copy loss of ~43.0 Mb). See also Tables S1B and S1D.*

Looking at individual somatic mutations, we confirmed three driver mutations (*Kras* G12D, *Trp53* T122K, and *Kdm6a* H1146Y) for the MCB6C cell line, which were previously reported from analysis of bulk whole transcriptome sequencing (RNA-seq) (Sato et al., 2018). Along with these three mutations, we identified 31 additional mutations across 20 previously reported driver genes in human bladder cancer, including a second *Trp53* mutation (a splice donor variant) and a second missense *Kdm6a* mutation (Table S1C, STAR Methods) (Martínez-Jiménez et al., 2020). This set also included mutations in *Atm* (S1884T), *Fat1* (two missense, one stop-gained mutations), *Kmt2a* (H1067Q), and *Kmt2c* (one splice region mutation), which have each been shown to harbor mutations in over 10% of bladder cancers, although none of the

specific mutations identified appear to have been previously reported in bladder cancer (Robertson et al., 2017). In addition to the stop-gained and splice region mutations identified in *Atm* and *Kmt2c*, respectively, two additional stop-gained mutations (one in *Birc6* and one in *Rnf213*) and three additional splice region variants (one in *Birc6*, one in *Brca2*, and one in *Sf3b1*) were identified. Finally, a mutation in *Sf3b1* (E873K), which was identified as a possible driver of a similar mouse urothelial carcinoma cell line, but was not previously detected in MCB6C using RNA-seq, was detected using WES (Sato et al., 2018).

In addition to calling SNVs and indels, we called copy number variants using the WGS data (STAR Methods). These results indicated that the MCB6C cell line has a relatively stable genome with only a few larger regions of copy number alteration consisting of copy gains on chromosomes 2 and 11 and a single copy loss on chromosome 12 (Figure 2.2B; Table S1D). Together, these results indicated that the MCB6C cell line has high SNV burden and low CNV burden. Previous research has shown that metastatic urothelial carcinoma patients with high SNV/low CNV tumor profiles may benefit more from ICB therapy. While high SNV/low CNV status has been associated with greater chance of response, the utility of SNV and CNV status is still being evaluated as a possible predictor of treatment response in bladder cancer (Roviello et al., 2020).

2.4.2 scRNA-seq was generated for over 64,000 cells, with over 59,000 cells passing filtering

10x Genomics 5′ single cell gene expression sequencing (scRNA-seq) was performed for five biological replicates, with each replicate consisting of three conditions: untreated tumor, treated tumor, and tumor treated after CD4+ T cell depletion (Figure 2.1A). In total, fifteen samples

were sequenced, generating ~8.3 billion reads across 64,049 cells (Table S2A). In addition to gene expression sequencing, 10x Genomics V(D)J B cell receptor (BCR) and T cell receptor (TCR) sequencing was also performed for all fifteen samples (Figure 2.1B; Tables S2B and S2C).

Before analyzing the scRNA-seq data, we aggregated all three conditions for each replicate and performed basic filtering on each of the five replicates to remove cells that appeared to be low quality based on mitochondrial gene expression per cell, detected gene count per cell, and/or total UMI count per cell. Briefly, cells expressing high percentages of mitochondrial genes, cells with low gene counts, and cells with high UMI counts were removed (Table S3, STAR Methods). Ultimately, 4,708 cells across all fifteen samples were removed, with 59,341 cells remaining.

2.4.3 scRNA-seq allows identification of lymphocyte, myeloid, and stromal cell populations in the tumor microenvironment

After completing basic filtering of cells, we used SingleR with the ImmGen dataset to assign fine label cell types to all remaining cells from each replicate (Aran et al., 2019; Heng et al., 2008; Shay & Kang, 2013). We then further filtered the set of remaining cells, removing all cells marked as "pruned" by SingleR. "Pruned cells" are those cells that have received poor-quality cell type assignments, potentially because of underlying poor quality of the cell itself. Once we removed all pruned cells, we were left with 57,818 cells total across all conditions and all replicates, which we aggregated into a single gene-barcode matrix for downstream analysis.

Next, we performed manual curation of SingleR's fine label cell type assignments to group fine labels of the same broad cell type and to identify subtypes within certain broad cell

types, e.g., to identify naive CD4 and CD8 T cells within the broader CD4 and CD8 T cell populations (STAR Methods). We also confirmed the general accuracy of the cell type assignments. First, for several cell types, we picked one reported marker for each cell type (e.g., *Cd79a* for B cells, *Epcam* for epithelial cells, and *Col3a1* for fibroblasts) and compared the expression of each marker in the cell type expected to express it (based on the SingleR cell type assignment) versus all other cell types (Figures S1A–S1D; Table S4). These plots confirmed that the expected cell types generally showed more common and higher expression of their markers than non-expected cell types. Additionally, cells identified as expressing BCR sequences (i.e., cells that are likely B cells) or TCR sequences (i.e., cells that are likely T or NK cells) were compared to cells labeled as either B or T/NK cells, respectively, according to their gene expression signatures. These results showed that approximately 92.8% of cells identified as expressing BCR sequences were labeled as B cells by SingleR and 98.9% of cells identified as expressing TCR sequences were labeled as some type of T or NK cell, i.e., CD4, CD8, NK, NKT, Tgd, or Treg cells (Figures S1E and S1F; Table S4).

Before beginning any additional analysis, we also filtered out lowly expressed genes. For a gene to pass filtering, we required the gene to be detected in two or more cells in each replicate, with a UMI count of at least two in each cell. After filtering genes based on these criteria, we were left with 11,398 genes. Finally, we generated a tSNE projection for the aggregated dataset and colored cells by their manually curated cell type labels (Figure 2.3; Table S4, STAR Methods). This tSNE projection suggested that epithelial cells formed two distinct clusters, indicating there may be two transcriptionally distinct populations of epithelial cells within the dataset (discussed extensively in the following section).

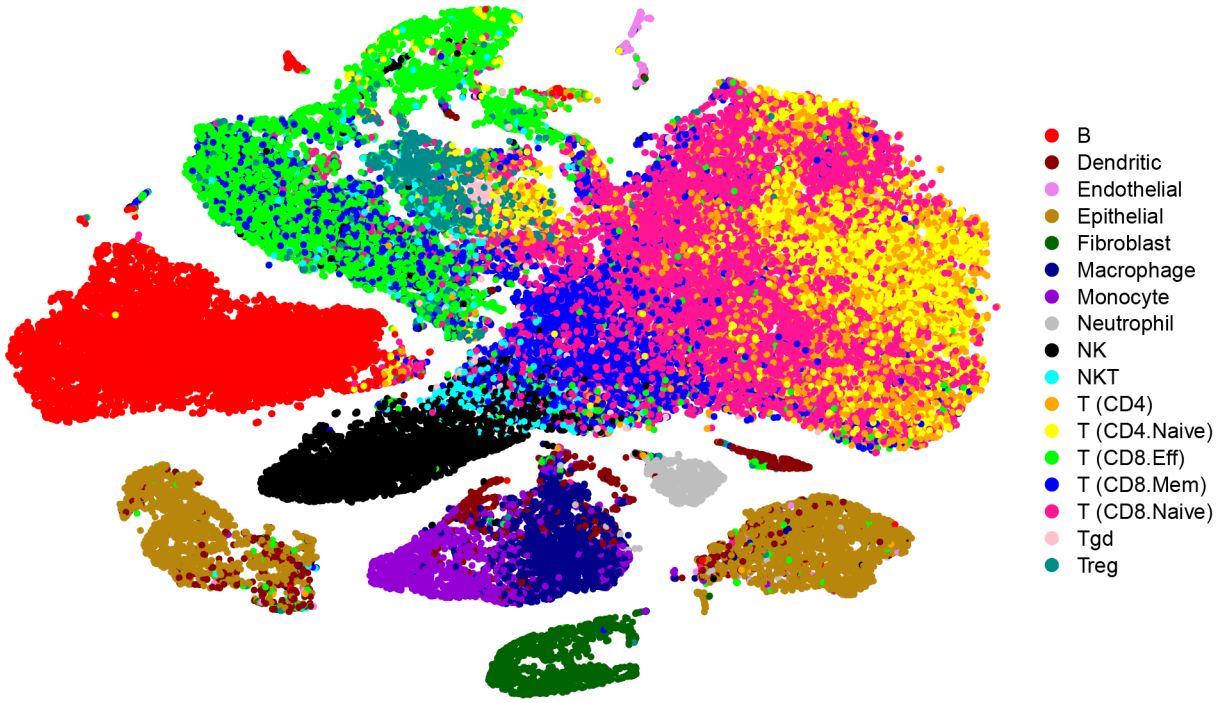


Figure 2.3: scRNA-seq allows identification of lymphocyte, myeloid, and stromal cell populations in the tumor microenvironment

tSNE projection of the aggregated dataset containing 57,818 cells, across all replicates and conditions, that passed filtering and were not "pruned" by SingleR. Cells were clustered using PCs = 20. Cells are colored by manually curated SingleR cell types. See also Figure S1 and Table S4.

After cell typing was completed, we also explored the BCR and TCR sequencing results. This analysis revealed little evidence of clonotype expansion of BCRs or TCRs across any of the conditions or replicates. For the TCR analysis, we excluded samples containing less than 1,500 TCR+ cells after filtering. We found that the majority of samples with less than 1,500 cells did not appear to have had their naive T cell populations captured during sequencing, which could lead to a skewed appearance of clonal expansion among the T cell populations that were

captured (Figures S2A and S2B). For the remaining samples, there did not appear to be any dominant clonotypes detected (Figure S3A). The most commonly observed clonotype in any sample was present in only 41 cells and the vast majority of clonotypes were present in only one cell (Figure S3B). While no samples had evidence of dominant clonotypes, several samples did show evidence of cells with modestly expanded clonotypes (i.e., clonotypes detected in more than one cell) clustering together on the tSNE projection (Figure S4A). These clonotypes were largely found to be expressed in CD8 effector T cells, which showed evidence of co-expression of some exhaustion markers (Figures S4B, S5A, and S5B). For BCR analysis, most samples (11 of 15) had less than 1,000 BCR+ cells that passed filtering (Figure S6A). For samples where more than 1,000 cells passed filtering, there was little evidence of clonal expansion (Figure S6B).

2.4.4 Somatic variation can be used to identify tumor cell populations with high confidence

Since we expected tumor tissue to be epithelial, we expected that the epithelial populations identified by SingleR would correspond to the tumor cell populations. To verify this expectation, we classified cells as tumor or non-tumor based on the presence or absence of somatic mutations as follows.

With the 10,427 somatic variants identified from WES, we used VarTrix to detect supporting reads for the reference and alternate alleles at each variant position in each individual cell in the aggregated dataset (Figure 2.4A). To identify a high confidence set of variant-containing cells, we required a cell to have at least two variant positions with greater than 20x

total coverage, greater than five reads supporting the alternate allele, and a VAF over 10%. Using these criteria, we classified 4,628 cells as somatic variant-containing cells.

These variant-containing cells largely formed two distinct clusters on the tSNE projection, which heavily overlapped the two clusters identified as epithelial clusters using SingleR's cell type labels (Figure 2.4B; Table S5). Since we expected the tumor tissue to be epithelial tissue, this extensive overlap appeared to confirm that variant-positive status could be used to identify tumor cells with high confidence. Additionally, we compared the overlap of variant-positive cells, cells that were assigned as epithelial cells by SingleR, and cells that were expressing *Epcam*, a marker of epithelial tissue. While the two variant-positive, epithelial-typed clusters showed high, widespread expression of *Epcam* as expected, there was also expression of *Epcam* detected across numerous other clusters (Figure 2.4B). These results indicated that variant status could be used to distinguish *Epcam*+ cells that are epithelial tumor cells from cells that appear to be *Epcam*+, but are not likely to be tumor cells (i.e. variant-negative, non-epithelial labeled cells) and may have simply been contaminated by ambient *Epcam* RNA.

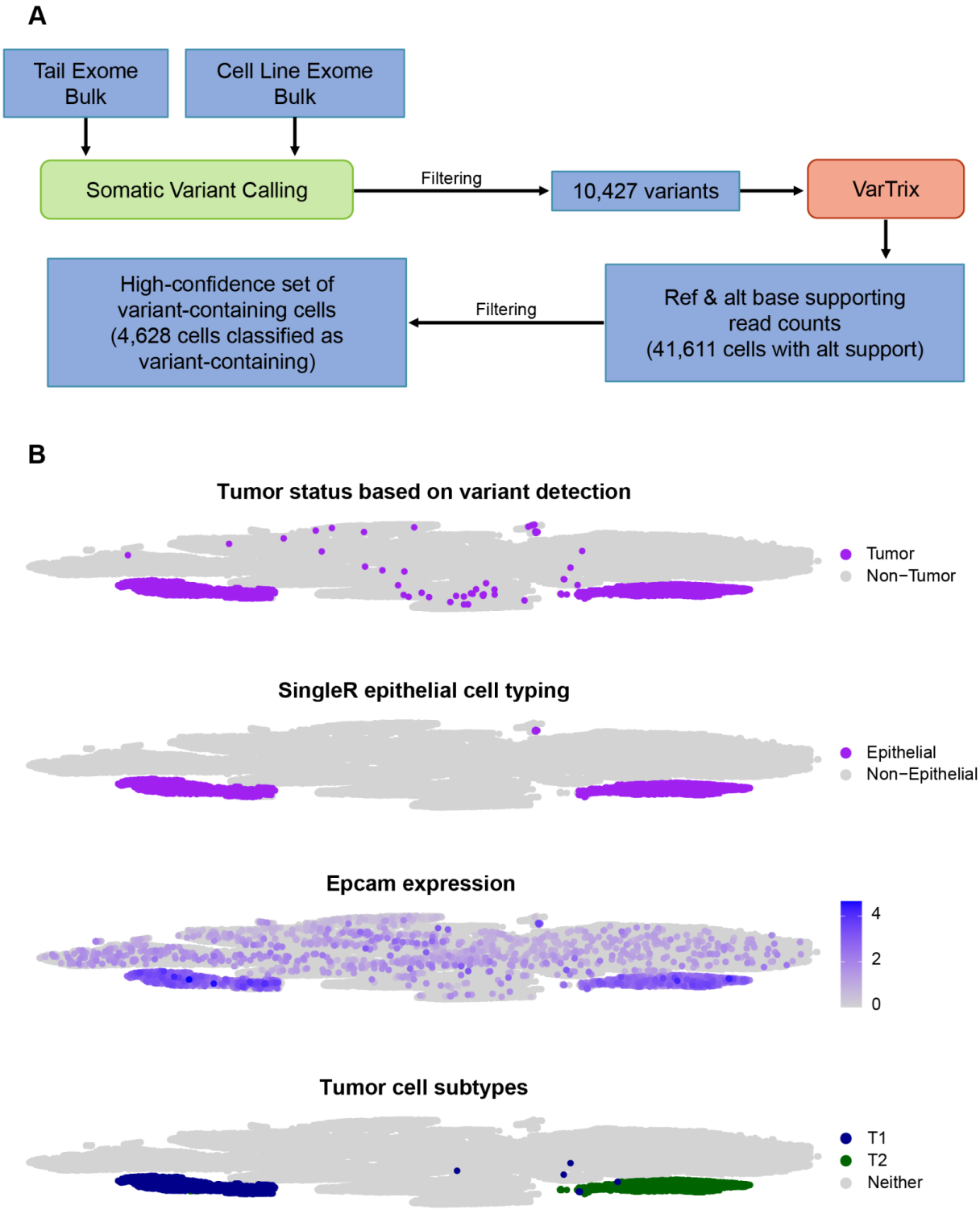


Figure 2.4: Somatic variation can be used to identify tumor cell populations with high confidence

(A) *Workflow for detecting somatic variation in scRNA-seq data to identify variant-containing cells.* **(B)** *tSNE projections showing the classification of tumor cells based on variant detection, epithelial cell typing from SingleR, Epcam expression, and labeling of tumor subpopulations (T1 and T2) based on clustering, respectively. See also Table S5.*

2.4.5 Tumor cell populations show evidence of two distinct subpopulations

After confirming which cells and clusters corresponded to tumor cell populations, we investigated why tumor cells appeared to form two distinct clusters, which we labeled T1 and T2 (Figure 2.4B). To further explore whether these populations truly represented distinct tumor cell populations, we separated the tumor clusters from the rest of the aggregated dataset and reclustered them (STAR Methods). The tSNE projection again revealed distinct clustering of each of the two subpopulations (Figure 2.5A). We then assigned relative differentiation scores to each cell using CytoTRACE (Gulati et al., 2020). We also performed differential expression analysis comparing the T1 subpopulation, containing all three conditions, to the T2 subpopulation, also containing all three conditions (STAR Methods).

The relative differentiation scores revealed that the T2 cells largely corresponded to the most highly differentiated cells, while the T1 cells appeared to form two groups of cells—one which corresponded to the least differentiated cells and one which corresponded to slightly more differentiated, but still relatively lowly differentiated cells (Figure 2.5A; Table S6A). As previous literature has established that luminal bladder cancers display characteristics of greater differentiation than basal bladder cancers, we next wanted to determine if these two subpopulations might have different basal-like and luminal-like expression patterns (Robertson et al., 2017).

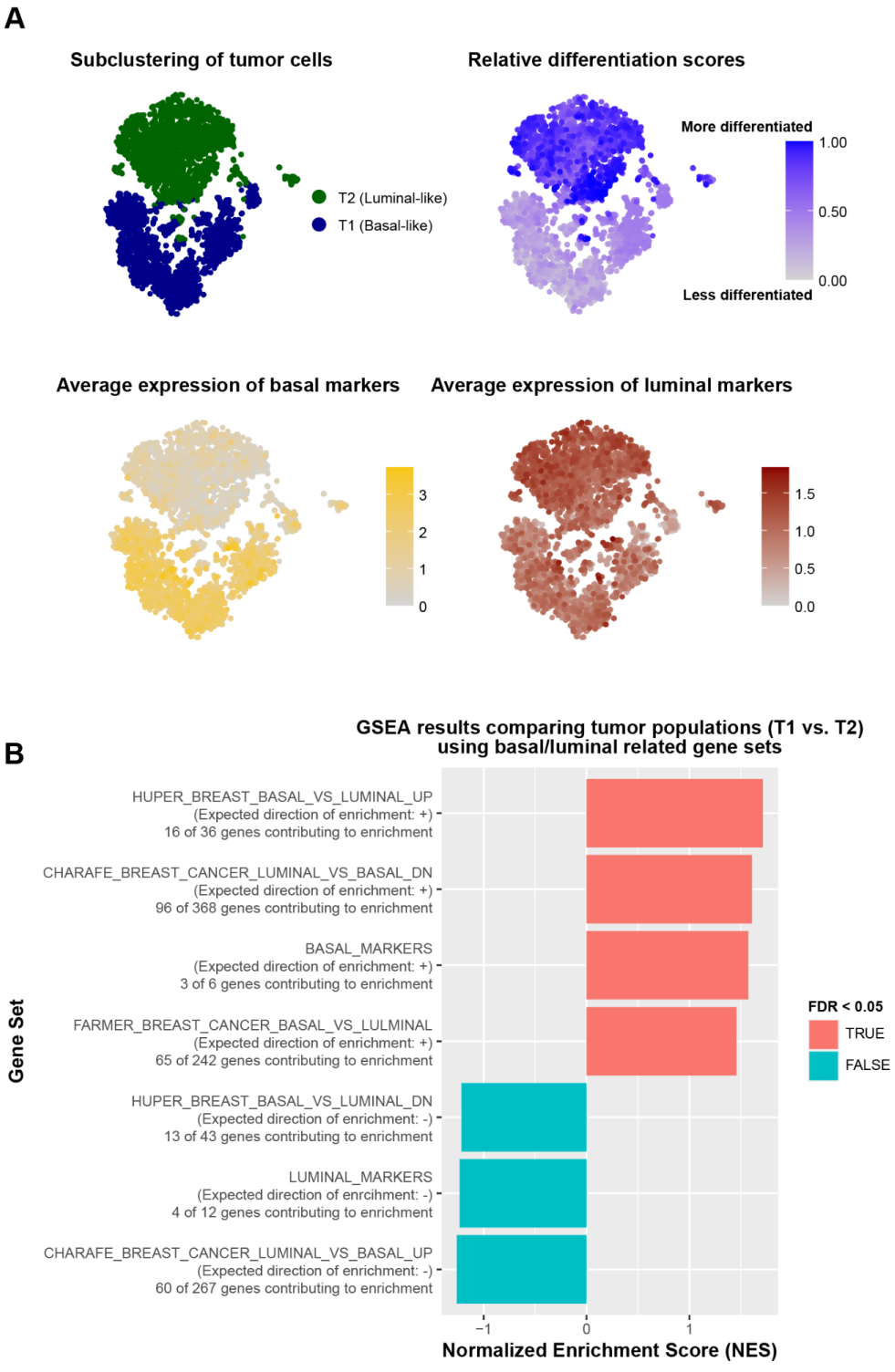


Figure 2.5: Tumor cell populations show evidence of two distinct subpopulations

(A) *tSNE projections of tumor cell populations showing subpopulation labels, differentiation scores (1 - CytoTRACE scores), average expression of basal bladder cancer markers, and average expression of luminal bladder cancer markers. Differentiation scores indicate the relative differentiation states of each cell within the full tumor cell population. Differentiation scores close to 1.00 indicate cells are relatively more differentiated. Differentiation scores close to 0.00 indicate cells are relatively less differentiated. Cells were clustered using PCs = 20.* **(B)** *Bar plot showing the normalized enrichment scores (NES) for GSEA of gene sets related to basal and luminal bladder and breast cancer gene expression. Bars are colored by their FDR q-value status. Salmon pink indicates a significant FDR q-value (<0.05). Blue indicates a non-significant FDR q-value (>0.05). See also Tables S6A and S6B.*

First, we calculated the average expression of reported basal (*Cd44*, *Krt14*, *Krt5*, *Krt16*, *Krt6a*) and luminal (*Cd24a*, *Erbb2*, *Erbb3*, *Foxa1*, *Gata3*, *Gpx2*, *Krt18*, *Krt19*, *Krt7*, *Krt8*, *Upk1a*) bladder cancer markers in each tumor cell and overlaid the values on the tSNE projection of the tumor populations (Figure 2.5A; Table S6A) (W. Choi et al., 2014; Guo et al., 2020). These results indicated that the less differentiated cells (the T1 population) had stronger and more widespread expression of basal markers than the more differentiated cells (the T2 population), consistent with the expectation that basal-like bladder cancer cells would be less differentiated than luminal-like bladder cancer cells (Figure 2.5A). Both populations showed expression of luminal markers. However, the less differentiated T1 population appeared to have lower levels of luminal marker expression overall, suggesting that the T1 population could represent a population with more basal-like expression, while the T2 population could represent a population with more luminal-like expression.

Next, after generating differential expression analysis results for comparing the full basal-like population (T1) to the full luminal-like population (T2), we performed gene set enrichment analysis (GSEA) using gene sets generated by comparing basal and luminal breast cancers, which have been shown to have highly similar expression profiles to basal and luminal bladder cancers (STAR Methods) (Dadhania et al., 2016). We also included two gene sets that we generated from reported lists of basal and luminal bladder cancer markers, respectively (W. Choi et al., 2014; Guo et al., 2020). These results indicated that upregulated genes in the T1 population were significantly enriched (FDR <0.05) for basal bladder cancer markers. These upregulated genes were also significantly enriched for genes that were found to be upregulated in basal breast cancers compared to luminal breast cancers (Figure 2.5B; Table S6B). Both of these observations were consistent with the T1 population being more basal-like than the T2 population.

By contrast, genes that were downregulated in the T1 population showed enrichment of luminal bladder cancer markers as well as genes that were found to be downregulated in basal breast cancers compared to luminal breast cancers. While this enrichment was not significant at an FDR cutoff of 0.05, the direction of enrichment was consistent with the observation that the T2 population appeared to have stronger expression of luminal markers than the T1 population (Figure 2.5B; Table S6B). These results further indicated that there were two distinct subpopulations of tumor cells within the full tumor cell population, i.e., a population with more basal-like characteristics and a population with more luminal-like characteristics.

2.4.6 Overrepresentation and gene set enrichment analysis identify IFN-g response as a commonly perturbed gene set across immune and tumor cell types upon ICB treatment

After assigning cell types and subtypes, where appropriate, to all cells, we explored how each individual cell type was responding to ICB treatment. To do this, we performed differential expression analysis comparing each possible pair of conditions within each cell type (STAR Methods). We then used the results of these differential expression analyses to perform overrepresentation analysis and GSEA.

Overrepresentation analysis showed that the top 5 most commonly overrepresented hallmark gene sets across cell types and comparisons were related to immune response (Table S7A). The IFN-g response gene set was the second most commonly overrepresented gene set (Figure 2.6A). Given that prior research suggested that IFN-g within the TME may be an important mediator of treatment response, we chose to explore the IFN-g response gene set further (Sato et al., 2018).

Since the overrepresentation analysis did not include information about the directionality or magnitude of overrepresentation, we generated a quantitative metric that captured both these aspects. Specifically, we summed the average log2 fold changes reported by Seurat for each detected gene in the IFN-g response gene set. With this method, a positive value indicates that genes from the gene set skew toward upregulation in the first condition of a given comparison in a given cell type, while a negative value indicates that genes skew toward downregulation. These sums indicated that ICB-treated endothelial cells experience upregulation of IFN-g response genes when compared to both untreated endothelial cells and endothelial cells treated after CD4+ T cell depletion (Figure 2.6B). These sums also suggested that untreated endothelial cells

experience upregulation of IFN-g response genes when compared to endothelial cells treated after CD4+ T cell depletion. Endothelial cells appeared to be the only non-tumor cell type that experienced upregulation across all three comparisons (Figure S7A; Table S7B).

To explore enrichment of up- and downregulated genes more formally, we performed ranked GSEA, using average log2 fold changes as the ranking metric and MSigDB's hallmark gene sets as the test set (STAR Methods). Similar to the results seen with the overrepresentation analysis, we found that the IFN-g response gene set was commonly enriched across multiple cell types. Furthermore, the enrichment results followed similar patterns to those seen using the "sum of fold changes" metric. When looking at the ICB treated condition versus both the control and CD4+ T cell depleted conditions, endothelial cells showed significant enrichment of upregulated IFN-g response genes (Figure S7B). Additionally, endothelial cells were the only cell type to show significant positive enrichment across all three comparisons (Figure S7B). Examination of the top genes contributing to enrichment in each comparison within endothelial cells indicated that upregulation of chemokines, such as *Cxcl9* and *Cxcl10*, and adhesion molecules, such as *Vcam1* and *Icam1*, was common across comparisons (Figure 2.6C; Tables S7C–S7E). Since IFN-g signaling in endothelial cells has been suggested to play multiple roles in the tumor immune response, but its role in ICB treatment response has not been documented, we examined the role of IFN-g signaling in endothelial cells further (Kammertoens et al., 2017; Ni & Lu, 2018).

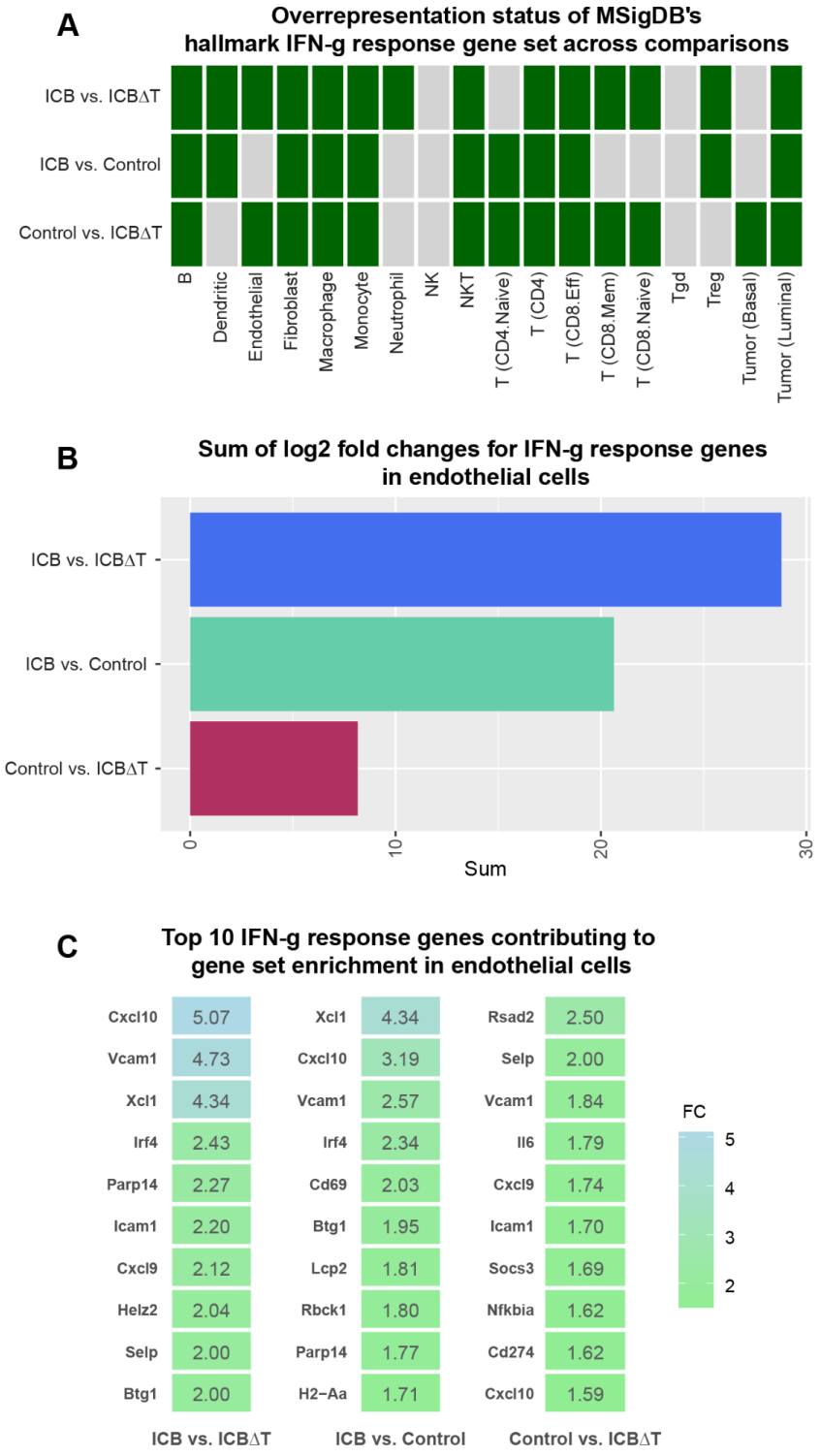


Figure 2.6: Overrepresentation and gene set enrichment analysis identify IFN-g response as a commonly perturbed gene set across immune and tumor cell types upon ICB treatment

(A) *Heatmap showing the overrepresentation status of MSigDB's hallmark IFN-g response gene set across each pairwise comparison of conditions in each cell type. A green square indicates IFN-g response genes were significantly overrepresented for the given comparison and cell type. A gray square indicates IFN-g response genes were not significantly overrepresented for the given comparison and cell type.* **(B)** *Sum of fold changes for hallmark IFN-g response genes for each pairwise comparison of conditions in endothelial cells. Positive values indicate that IFN-g response genes skew toward upregulation in the first condition of a given comparison.* **(C)** *Fold changes for the top 10 genes contributing to gene set enrichment of MSigDB's hallmark IFN-g response gene set for each pairwise comparison of conditions in endothelial cells. ICB = combined PD-1/CTLA-4 immune checkpoint blockade treatment, ICBΔT = combined PD-1/CTLA-4 immune checkpoint blockade treatment received after CD4+ T cell depletion. See also Figures S7A and S7B and Table S7.*

2.4.7 Functional analysis confirms endothelial cells are a principal target of IFN-g and a key mediator of treatment response

To test the role of IFN-g signaling in endothelial cells in response to ICB treatment, we generated a mouse model system where *IFNgR1* could be knocked out specifically in endothelial cells with tamoxifen treatment by crossing CDH5-ERT2-Cre+ mice with *IFNgR1* flox/flox (f/f) mice. Using flow cytometry, we confirmed *IFNgR1* expression was significantly reduced in CD31+ endothelial cells from mice in the knockout conditions compared to intact mice lacking the Cre-expressing allele (Figures S8A, S9A, and S9B).

After establishing this model system, we compared tumor growth in *IFNgR1* intact mice with and without ICB treatment to tumor growth in endothelial *IFNgR1* knockout mice with and

without ICB treatment (STAR Methods). This comparison revealed that ICB treated knockout mice had tumor growth patterns nearly identical to untreated intact mice, demonstrating that significantly reducing *IFNgR1* expression in endothelial cells negated the anti-tumor effects of ICB treatment (Figure 2.7A; Table S8A). Thus, IFN-g response in endothelial cells is necessary for an effective ICB treatment response. Furthermore, untreated tumors in the knockout mice grew more quickly than untreated tumors in the intact mice (Figure 2.7A; Table S8A). These findings are analogous to previously reported findings which showed that CD4+ T cell depletion in the MCB6C model not only prevented ICB induced tumor rejection, but also led to increased tumor growth even in the absence of ICB treatment, indicating that a basal level of T cell activity restrains tumor growth (Sato et al., 2018). Similarly, the findings presented here indicated that basal levels of IFN-g signaling in endothelial cells restrained tumor growth and that upregulation of IFN-g activity in endothelial cells was necessary for tumor rejection upon ICB treatment.

Flow cytometric analysis indicated that ICB treatment induced recruitment of CD4+ T lymphocytes to the TME in intact mice, consistent with previous work (Figures 2.7B and S8B) (Sato et al., 2018). However, in mice where *IFNgR1* had been knocked out in endothelial cells, this recruitment of CD4+ T lymphocytes after treatment was negated (Figure 2.7B; Table S8B). Furthermore, recruitment of Tbet+, IFN-g+, CD4+ T lymphocytes (i.e., Th1-like cells) seen after ICB treatment was no longer seen in the knockout condition (Figure 2.7C; Table S8C). Similar to previously reported analysis of the MCB6C model, this analysis showed no significant change in the proportion of CD8+ T cells seen before or after ICB treatment in either intact or knockout mice (Figure S10) (Sato et al., 2018). These results further indicated that IFN-g signaling in endothelial cells is a key mediator of treatment response and that it underlies recruitment of CD4+ effector T cells in the TME.

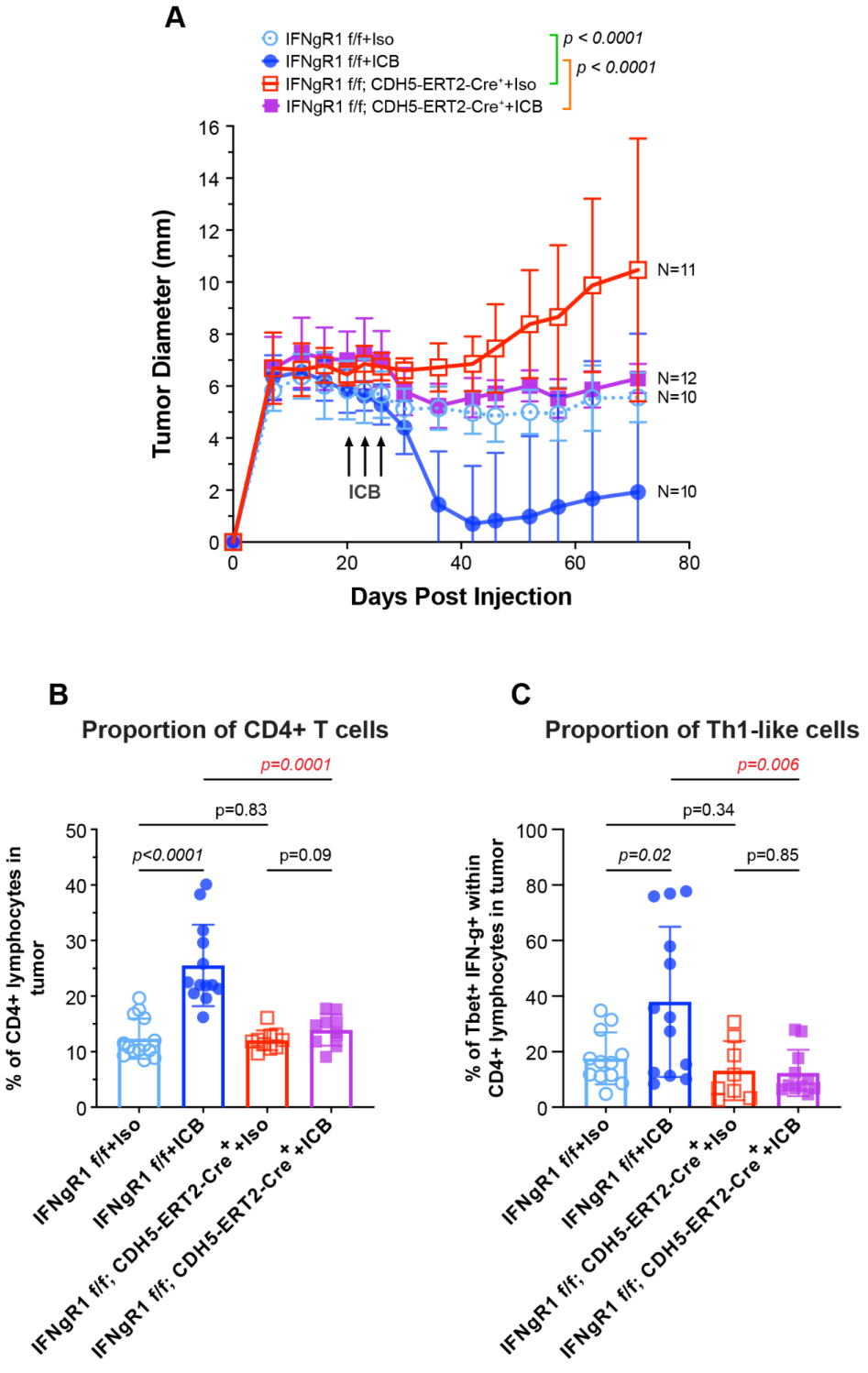


Figure 2.7: Functional analysis confirms endothelial cells are a principal target of IFN-g and a key mediator of treatment response

(A) *Tumor diameter measurements for IFNgR1 intact and endothelial IFNgR1 knockout mice with and without ICB treatment over time (pre- and post-treatment). For all comparisons, a 2-way ANOVA for repeated measures was performed. P-values <0.05 were considered significant. Error bars represent one standard deviation.* **(B)** *Bar graphs displaying the percentage of CD4+ lymphocytes in the tumor microenvironment across the same four conditions as* **(A)**. **(C)** *Bar graphs displaying the percentage of Tbet+, IFNg+ cells detected within the CD4+ T lymphocyte population across the same four conditions as* **(A)**. *For all bar graphs, bar height indicates the average percentage across all mice from the given condition. Each point represents the percentage for an individual mouse. For all comparisons, a two-tailed, unpaired Student's t-test was performed. P-values <0.05 were considered significant. Error bars represent one standard deviation. f/f = flox/flox, ICB = combined PD-1/CTLA-4 immune checkpoint blockade treatment, Iso = rat IgG2a and mouse IgG2b isotype control. See also Figures S8–S10 and Table S8.*

2.5 Discussion

To explore mechanisms of response to combined PD-1/CTLA-4 ICB treatment of bladder cancer in individual cell types, we generated scRNA-seq from a mouse model of urothelial carcinoma. The three sample conditions used in this study were untreated tumor, combined PD-1/CTLA-4 ICB treated tumor, and tumor that received combined ICB treatment after CD4+ T cell depletion. In total, we performed scRNA-seq on fifteen samples (five per each condition) and captured over 57,000 cells that passed filtering and were aggregated into a single dataset for downstream analysis. Within the aggregated dataset, we identified numerous lymphocyte, myeloid, and stromal cell populations. Clustering and visualization of the data revealed two distinct epithelial clusters, which we confirmed corresponded to tumor cell populations based on expression of

somatic variants and, more specifically, appeared to correspond to distinct basal-like and luminal-like subpopulations.

After identifying cell types present within the aggregated dataset, we used differential expression, overrepresentation, and GSEA to explore how individual cell types were responding to treatment. This analysis showed that IFN-g response was commonly perturbed with treatment across multiple cell types, including endothelial cells. Multiple clinical trials exploring human bladder cancer have identified IFN-g pathway activity as being correlated with increased benefit from ICB treatment (Rosenberg et al., 2016; Sakatani et al., 2022; Sharma et al., 2017). Previous work in the MCB6C model established that IFN-g activity is necessary for ICB treatment response (Sato et al., 2018). While previous research of tumor immunosurveillance models has shown that IFN-g signaling can act through both tumor cell intrinsic and extrinsic mechanisms, the role of IFN-g and its key target cells in ICB treatment response has not been completely defined (Alspach et al., 2019; Kammertoens et al., 2017; D. H. Kaplan et al., 1998).

While previous work excluded IFN-g activity in tumor cells as having an essential role in treatment response in the MCB6C model, we had not previously evaluated its role in endothelial cells. Here, we establish endothelial cells as a key target of IFN-g activity and further show that loss of IFN-g signaling in endothelial cells impairs recruitment of IFN-g producing CD4+ T cells to the TME. Notably, *Cxcl9*, *Cxcl10*, *Vcam1*, and *Icam1*, which are mediators of T cell trafficking, were among the most upregulated IFN-g response genes in endothelial cells following ICB treatment, suggesting that a key role of IFN-g activity in endothelial cells may be to enable recruitment of T cells to the TME.

We hypothesize a feedforward model in which ICB treatment induces IFN-g production from CD4+ T cells, which in turn leads to further recruitment of CD4+ T cells to the TME via

upregulation of chemoattractant molecules in endothelial cells. However, other roles of IFN-g signaling in endothelial cells could also contribute to treatment response. For example, IFN-g signaling in endothelial cells could induce tumor ischemia or impact vascular permeability, as shown by previous studies (Chrobak et al., 2013; Kammertoens et al., 2017; Ni & Lu, 2018). Ultimately, these results showed that IFN-g response in endothelial cells is a key mediator of treatment response and suggested that strategies which selectively induce IFN-g signaling in endothelial cells in the TME could favorably impact response to ICB treatment as well as other T cell based therapies.

2.5.1 Limitations of the study

While these findings support the role of IFN-g signaling in endothelial cells as a key node in treatment response, there are limitations to this analysis. In particular, effective treatment response involves a cascade of events which are still not fully defined. For example, the mechanisms by which T cells in the TME actually kill tumor cells are not elucidated in this system. Likewise, the mechanisms by which endothelial cells recruit T cells to the TME have not been fully explored. Additionally, the immune microenvironment arising from subcutaneous injection of a bladder cancer cell line could differ significantly from the immune microenvironment of a bladder cancer grown in bladder tissue. Ultimately, further analysis will be needed to verify and fully characterize the mechanisms underlying effective ICB treatment response. Nevertheless, these results underscore the power of scRNA-seq analysis to inform hypotheses that, when coupled with mouse modeling, can help identify cell-type specific signaling nodes that are key to generating an effective immune response.

2.6 STAR Methods

Detailed STAR Methods are provided in the online version of this paper.

2.6.1 Mice used for MCB6C experiments

All animal experiments were carried out according to the guidelines of the American Association for Laboratory Animal Science under a protocol approved by the Institutional Animal Care and Use Committee at Washington University and performed in Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC)-accredited specific pathogen-free facilities at Washington University School of Medicine in St. Louis. Forty-five 5- to 6-week-old Black 6 (B6NTac) male mice were purchased from Taconic Biosciences and were allowed to acclimate for a week before in vivo experiments were performed. The maximal tumor size/burden permitted by our institutional review board is 15% of body weight (combined burden if more than one mass present) and mean tumor diameter = or >20 mm in adult mice (~25 g). The maximal tumor size/burden permitted by our institutional review board was not exceeded.

2.6.2 CDH5-ERT2-Cre+, IFNgR1 flox/flox (f/f) mice

C57BL/6-*Tg(Cdh5-cre/ERT2)*^{1Rha} mice were originally generated by Dr. Ralf H. Adams and purchased from Taconic Biosciences then bred with C57BL/6N-*Ifngr1*$^{tm1.1Rds}$/J (IFNgR1$^{flox/flox}$) mice that were obtained from Dr. Robert Schreiber at Washington University School of Medicine to generate CDH5-ERT2-Cre+/IFNgR1$^{flox/flox}$ offspring.

2.6.3 Bulk DNA sequencing, alignment, and analysis

Whole genome sequencing (WGS) libraries were constructed from genomic DNA isolated from an MCB6C cell line sample and a black 6 (B6NTac) matched normal tail sample using Automated Kapa HYPER PCR free preparation kits (catalog #7962371001 – KK8505) and sequenced on the Illumina NovaSeq 6000 platform. WGS reads were aligned to the GRCm38 reference genome using BWA-MEM. Copy number variant calling was performed using the CNVkit (v0.9.8) batch pipeline with a target bin size of 100,000 bp (Talevich et al., 2016). Whole exome sequencing (WES) libraries were constructed and sequenced similarly to the WGS experiment following hybrid capture selection with the hybrid reagent SureSelect DNA - Mouse All Exon V1 (Agilent). WES reads were aligned to the GRCm38 reference genome using BWA-MEM. Somatic variant calling was performed using common workflow language pipelines provided by the McDonnell Genome Institute (https://github.com/genome/analysis-workflows). Somatic variants were called with Pindel, VarScan, Mutect, and Strelka and combined (Cibulskis et al., 2013; Griffith, Griffith, et al., 2015; Kim et al., 2018; Koboldt et al., 2012; K. Ye et al., 2009). Variants were then filtered based on the criteria of being called by at least two variant callers, normal coverage > 30X, tumor coverage > 30X, normal VAF < 5%, and tumor VAF > 5%. To explore clonality based on VAF distribution, variants were filtered more stringently based on the criteria of being called by at least two variant callers, having normal and tumor coverage >= 100X, normal VAF = 0%, tumor VAF > 5%, and no overlap with any regions of copy number alteration (i.e. any regions with abs(log2 copy ratio) > 0.32). Additionally, variants with VAF < 30% were manually reviewed to remove possible false variants caused by artifacts such as sequencing errors or misalignments.

2.6.4 Identifying possible driver mutations in WES

After filtering somatic variants, a subset of possible driver mutation positions was determined by further filtering the set of somatic mutations down to mutations found in genes that have been previously reported to harbor driver mutations in human bladder cancer (https://www.intogen.org/search?cancer=BLCA). Human gene names were converted to homologous mouse gene names using the Mouse Genome Informatics human and mouse homology report with mammalian phenotype IDs (https://www.informatics.jax.org/homology.shtml). Each mutation was manually reviewed against mutations reported in ProteinPaint (https://proteinpaint.stjude.org/), IntOGen, and Cancer Hotspots (https://www.cancerhotspots.org/#/home).

2.6.5 Mouse bladder organoid culture for injection

One previously archived frozen vial of singly suspended MCB6C organoid was thawed at least 2 weeks before mouse injection and expanded weekly in culture at least 2 times. For MCB6C organoid culture expansion, growth factor reduced Matrigel was thawed on ice for minimally 1.5 hours. Pelleted MCB6C cells were washed and resuspended in 1 ml of Advanced DMEM/F12+++ medium (Advanced DMEM/F12 medium [Gibco, catalog #12634010] supplemented with 1% penicillin/streptomycin, 1% HEPEs, and Glutamax) and cell concentration was determined by automated cell counter. To establish organoid culture, 50 µl Matrigel tabs with 10,000 cells/tab were generated and plated on 6-well suspension culture plates, 6 tabs wells. Tabs were incubated at 37°C for 15 min until Matrigel was hardened, returned to tissue culture incubator, and cultured with mouse bladder organoid medium (MBO medium - Advanced DMEM/F12+++ medium supplemented with EGF, A-83-01, Noggin, R-

Spondin, N-Acetly-L-cysteine, and Nicotinamide). Organoids were replenished with fresh MBO medium every 3–4 days and also one day before mouse injection.

2.6.6 Mouse injection with MCB6C organoid cells

A single cell suspension of MCB6C organoid was generated by TrypLE Express (Gibco, catalog #12605010) digestion organoid Matrigel tabs at 37°C for 15 min. After digestion, pelleted cells were washed and resuspended in PBS to determine cell concentration. After cell concentration was adjusted to 20 million/ml in PBS, organoid cells were mixed with growth factor reduced Matrigel at 1:1 ratio before being injected subcutaneously into the left flank of the mouse (1 million/100 µl cells each mouse). Tumor development was monitored using digital calipers to assess the length, width, and depth of each tumor. For ICB, each mouse was injected intraperitoneally with 250 µg anti-PD1 (BioXcell, catalog #BE0146, clone #RMP1-14) and 200 µg anti-CTLA-4 (BioXcell, catalog #BE0164, clone 9D9) day 9 and 12 after organoid implantation. For isotype controls, each mouse was injected with 250 µg rat IgG2a (BioXcell, catalog #BE0089, clone 2A3) and 200 µg IgG2b (BioXcell, catalog #BE0086, clone #MPC-11). For CD4+ T cell depletion, each mouse was injected with 250 µg anti-CD4 (BioXcell, catalog #BE0003-1, clone #GK1.5) day 0 and 7 after organoid depletion. Rat IgG2b (BioXcell, catalog #BE0090, clone #LTF-2) was used as an isotype control for anti-CD4.

2.6.7 Harvesting tumors for scRNA-seq, BCR-seq, and TCR-seq

Based on 10x Genomics Demonstrated Protocols, 14 days after organoid implantation, tumors were dissected from euthanized mice, cut into small pieces of ~2–4 mm^3, and further processed into dead-cell depleted single cell suspension following manufacturer's protocol using Tumor

Dissociation Kit and MACS Dead Cell removal Kit (Miltenyi Biotec). Briefly, tumor tissue pieces were transferred to gentleMACS C tube containing enzyme mix before loading onto a gentleMACS Octo Dissociator with Heaters for tissue digestion at 37°C for 80 min. After tissue dissociation was completed, cell suspension was transferred to a new 50 ml conical tube, and supernatant was removed after centrifugation. Cell pellet was resuspended in RPMI 1640 medium, filtered through a prewetted 70 μM cell filter, strained, pelleted, and resuspended in red cell lysis buffer and incubated on ice for 10 min. After adding the wash buffer, the cell suspension was pelted and resuspended in the wash buffer. To remove dead cells, Dead Cell Removal Microbeads were added to resuspend cell pellet (100 μl beads per 107 cells) using a wide-bore pipette tip. After incubation for 15 min at room temperature, the cell-microbead mixture was applied onto a MS column. Dead cells remained in the column and the effluent represented to the live cell fraction. The percentage of viable cells was determined by an automated cell counter. Dead cell removal was repeated if the percentage of viable cells did not reach above 90%. Two rounds of centrifugation/resuspension were carefully performed for two rounds in 1xPBS/0.04% BSA using a wide-bore tip. To submit cell samples for single-cell RNA-seq analysis, cell concentration was determined accurately by sampling cell suspension twice and counting each sampling twice and adjusted to 1167 cells/μl. 40 μl of each cell suspension was submitted to the Genome Technology Access Center/McDonnell Genome Institute (GTAC/MGI) for single-cell RNA-seq analysis using the 5'v2 library kit (10x Genomics catalog #PN-1000263) with BCR and TCR V(D)J enrichment kits (10x Genomics catalog #PN-1000072 and #PN-1000071, respectively). cDNA generation and TCR/BCR enrichment were performed according to the Chromium Single Cell V(D)J Reagent Kits User Guide (CG000086 Rev L). The libraries

were sequenced on the S4 300 cycle kit flow (2x151 paired end reads) using the XP workflow as outlined by Illumina. FASTQ outputs were generated.

2.6.8 Alignment, filtering, and clustering of scRNA-seq

Alignment and gene expression quantification were performed with CellRanger count (v5.0, default parameters). Gene-barcode matrices were then imported into Seurat for filtering cells, QC, clustering, etc (Hao et al., 2021). To filter suspected dying cells, cells were clustered before filtering to identify cells appearing to cluster based on high mitochondrial gene expression (i.e. the percentage of UMIs per cell mapping to mitochondrial genes). The cutoff for mitochondrial gene expression was based on the percentage that captured the majority of these cells. A cutoff of 12.5% was used across all replicates. Doublets were filtered based on high UMI expression and CellRanger's reported doublet rate (0.9% per 1000 cells), with the top 0.9% of cells removed from each condition in each replicate. Cutoffs for filtering cells with low feature detection were determined by assigning cell types to each cell using the CellMatch method, identifying cells that did not have enough features for their cell type to be predicted, and calculating the average number of features detected in those "non-predicted" cells. CellMatch infers cell types by training a nearest-neighbors algorithm on published expression data. Spearman correlation is used as the distance metric. CellMatch's cell typing inference is unsupervised and infers cell types in a marker free manner (Petti et al., 2019). After filtered cells were removed, gene expression values for each gene in the remaining cells were normalized and scaled and variable genes were selected using Seurat with default settings. Principal component (PC) analysis was then performed using the top 2,000 variable genes and npcs = 20. Clustering of cells was performed using 20 PCs and resolution = 0.7. Finally, dimensionality reduction and visualization

were performed using Seurat's tSNE function. B cell and T cell receptors were assembled and identified using the 10x Genomics CellRanger V(D)J pipeline (v5.0, default parameters).

2.6.9 Assigning cell types using SingleR

Cell types for each cell were annotated with SingleR using expression profiles from the ImmGen dataset (https://www.immgen.org/) (Aran et al., 2019; Heng et al., 2008). Cell types were manually simplified to B cell (B), CD4+ T cell (CD4), naive CD4+ T cell (CD4.Naive), naive CD8+ T cell (CD8.Naive), CD8+ effector T cell (CD8.Eff), CD8+ memory T cell (CD8.Mem), dendritic cell, endothelial cell, epithelial cell, fibroblast, macrophage, monocyte, neutrophil, natural killer cell (NK), natural killer T cell (NKT), gamma delta T cell (Tgd), and regulatory T cell (Treg).

2.6.10 Analysis of tumor cell subtypes

To explore tumor cell subpopulations, we labeled tumor cells as either T1 or T2 based on cluster number. Cells that were assigned to the tumor clusters (i.e. clusters 10 and 14) on the left side of the tSNE projection were labeled as T1 tumor cells and cells assigned to the tumor cluster (i.e cluster 7) on the right side were labeled as T2 tumor cells. After assigning these labels, we separated the T1 and T2 cell populations from all other cell populations. We then scaled and normalized gene expression and selected variable genes using Seurat's default methods. Principal component (PC) analysis was then performed using the top 2,000 variable genes and npcs = 20. Clustering of cells was performed using 20 PCs and resolution = 0.7. Finally, dimensionality reduction and visualization were performed using Seurat's tSNE function. We then assigned differentiation scores to each cell using CytoTRACE

(https://cytotrace.stanford.edu/) and calculated the differentiation score as 1 - the CytoTRACE score (Gulati et al., 2020).

2.6.11 DE, overrepresentation, and GSEA in scRNA-seq

All differential expression analyses were performed using Seurat's FindMarkers function with the Wilcoxon Rank Sum method. P-values were adjusted using Benjamini-Hochberg multiple test correction. All reference gene sets used for overrepresentation analysis and gene set enrichment analysis (GSEA) were from MSigDB with the exception of the basal and luminal marker gene sets, which were generated ad-hoc from published lists of basal and luminal bladder cancer markers (Liberzon et al., 2015; A. Subramanian et al., 2005). For all overrepresentation analysis, results were generated using the enricher function from the clusterProfiler package in R (Yu et al., 2012). For comparisons of conditions within each cell type, input gene lists for overrepresentation analysis were generated by taking all genes with adjusted p-value < 0.05 and fold change value > ∼1.2 (i.e. abs(log2FC) > 0.26). GSEA results were generated using UC San Diego and Broad Institute's GSEA software to run GSEAPreranked with genes ranked by the average log2 fold changes reported by Seurat.

2.6.12 Deletion of IFNgR1 in endothelium by tamoxifen

Tamoxifen (Alfa Aesar, catalog #J63509) was dissolved in corn oil (MilliporeSigma, catalog #C8267) at the concentration of 20 mg/ml in a 37°C shaker overnight one day before the treatment began and kept at 4°C during the 5-day treatment.

2.6.13 Subcutaneous engraftment of MCB6C organoids

Tumor experiments were performed following methods established previously with modifications (Sato et al., 2018). To improve the engraftment and growth of the organoid cells on mice, Matrigel with high protein concentration (Corning, catalog #354262) was used instead of growth factor reduced Matrigel (Corning, catalog #356231). After organoids were expanded in culture for > 2 weeks and subsequently harvested by TrypLE Express (Gibco, catalog #12605010) treatment, organoid cells were resuspended in 3:1 PBS/high protein concentration Matrigel (instead of 1:1 PBS/growth factor reduced Matrigel) at 10 million cells/ml. 1 million/100 µl of cell/Matrigel mix was subcutaneously injected into the left flank of the mouse, which was performed one week after the completion of Tamoxifen treatment. Tumor growth was monitored twice a week using digital calipers. The mean of long and short diameters was used for tumor growth curves. For ICB treatment, mice were injected with 250 µg/mouse αPD-1 (BioXcell, catalog #BE0146, clone RMP1-14) and 200 µg/mouse αCTLA-4 (BioXcell, catalog #BE0164, clone 9D9) i.p. every 3 days from day 15 to 18 after organoid implantation for short term studies, and from day 15 through day 21 from long term studies. 250 µg/mouse rat IgG2a (BioXcell, catalog #BE0089, clone 2A3) and 200 µg/mouse IgG2b (BioXcell, catalog #BE0086, clone MPC-11) were used as isotype controls.

2.6.14 Antibodies

The following antibodies were used for flow cytometry: Brilliant Violet 510™ anti-mouse CD45 Antibody (BioLegend, catalog #103137, clone 30-F11), PE/Dazzle™ 594 anti-mouse CD3ε Antibody (BioLegend, catalog #100347, clone 145-2C11), FITC anti-mouse CD4 Antibody (BioLegend, catalog #116003, clone RM4-4), PerCP/Cyanine5.5 anti-mouse CD4 Antibody

(BioLegend, catalog #116011, clone RM4-4), PE/Cyanine7 anti-mouse CD8α Antibody (BioLegend, catalog #100721, clone 53-6.7), Alexa Fluor® 700 anti-mouse CD8α Antibody (BioLegend, catalog #100729, clone 53-6.7), Brilliant Violet 421™ anti-mouse CD19 Antibody (BioLegend, catalog #115537, clone 6D5), APC anti-mouse/human CD11b Antibody (BioLegend, catalog #101211, clone M1/70), PE/Cyanine7 anti-mouse CD11c Antibody (BioLegend, catalog #117317, clone N418), Alexa Fluor® 647 anti-mouse CD326 (Ep-CAM) Antibody (BioLegend, catalog #118211, clone G8.8), PerCP/Cyanine5.5 anti-mouse CD326 (Ep-CAM) Antibody (BioLegend, catalog #118219, clone G8.8), Brilliant Violet 421™ anti-mouse/human CD44 Antibody (BioLegend, catalog #103039, clone IM7), PE anti-mouse CD62L Antibody (BioLegend, catalog #104407, clone MEL-14), PE/Cyanine7 anti-mouse Ly-6C Antibody (BioLegend, catalog #128015, clone HK1.4), PerCP/Cyanine5.5 anti-mouse Ly-6G Antibody (BioLegend, catalog #127615, clone 1A8), PE anti-mouse Siglec-F Antibody (BD Biosciences, catalog #552126, clone E50-2440), APC/Cyanine7 anti-mouse CD335 (NKp46) Antibody (BioLegend, catalog #137645, clone 29A1.4), Alexa Fluor™ 700 anti-Foxp3 Antibody (eBioscience, catalog #56-5773-80, clone FJK-16s), PE/Dazzle™ 594 anti-T-bet Antibody (BioLegend, catalog #644827, clone 4B10), APC/Cyanine7 anti-mouse IFN-γ Antibody (BioLegend, catalog #505849, clone XMG1.2), Alexa Fluor® 647 anti-Ki-67 Antibody (BD Biosciences, catalog #561126 clone B56), Biotin anti-mouse CD119 (IFN-γ R α chain) Antibody (BioLegend, catalog #112803, clone 2E2), PE Streptavidin (BioLegend, catalog #405203).

2.6.15 Flow cytometry

To determine the cellular composition of the tumor, tumors were isolated, minced into small pieces, and digested for 1 hour in DMEM media (MilliporeSigma, catalog #D5796) containing

100 µg/ml Collagenase type IA (Gibco, catalog #17101015), 100 µg/ml Dispase II (MilliporeSigma, catalog #D4693) and 50 U/ml of DNase I (Worthington Biochemical, catalog #LS002006). Cells were washed in ice-cold PBS with 3% FCS and 2 mM EDTA (FACS buffer) and filtered over 70-µm nylon mesh. After red blood cell lysis with ACK solution (Gibco, catalog #A1049201), cells were stained with a Zombie NIR Fixable Viability kit (BioLegend, catalog #423105) for dead cell exclusion followed by Fc-receptor blocking with purified mouse CD16/32 antibody (BioLegend, catalog #101301, clone 93). After cell surface marker staining with fluorescent-conjugated antibodies, cells were fixed and permeabilized using a Foxp3/transcription factor staining kit (eBioscience, catalog #00-5523-00) and intracellularly stained with fluorescent-conjugated antibodies. Flow cytometric data were acquired by Cytek-upgraded 10-color FACScan cytometers at Washington University Siteman Cancer Center Cell Sorting Core facility and analyzed by FlowJo 10 (TreeStar).

2.6.16 Quantification and statistical analysis

Statistical analyses for *IFNgR1* knockout experiments were performed using Prism 8.3.0 (GraphPad). For all tumor growth curve comparisons, a 2-way ANOVA for repeated measures was used. For all other comparisons, an unpaired Student's t test was used. All tests were 2-tailed. P-values of less than 0.05 were considered significant.

2.7 Acknowledgments

Malachi Griffith was supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under Award Number R00HG007940. Malachi Griffith and Obi Griffith were supported by the NIH National Cancer Institute (NCI) under Award

Numbers U01CA209936, U01CA231844, U24CA237719, and U01CA248235. Malachi Griffith was supported by the V Foundation for Cancer Research under Award Number V2018-007. Vivek Arora was supported by the Department of Defense Career Development Award W81XWH-17-1-0562, the Rabushka Bladder Cancer Research Fund, and the Clinical Investigator Award from the Damon Runyon Foundation. We thank Dr. Robert Schreiber at Washington University School of Medicine for gifting us C57BL/6N-*Ifngr1*$^{tm1.1Rds}$/J (IFNgR1$^{flox/flox}$) mice. We thank Dr. Kian H. Lim for providing infrastructure support for mouse experiments. We thank Joshua F. McMichael for designing the graphical abstract.

Author contributions

Conceptualization V.K.A., M.G., and O.L.G.; Methodology T.H.P.C. and J.K.B.; Formal Analysis S.L.F., B.F., H.S., M.M., and Z.L.S.; Investigation T.H.P.C.; Writing - Original Draft S.L.F.; Writing - Review and Editing V.K.A., M.G., O.L.G., T.H.P.C., and S.L.F.; Supervision V.K.A., M.G., and O.L.G.; Funding Acquisition V.K.A., M.G., and O.L.G.

Declaration of interests

V.K.A. currently serves as an employee of Bristol Myers Squibb and has stock options in the company. J.K.B. currently serves as an employee of Pfizer Inc.

Inclusion and diversity

We support inclusive, diverse, and equitable conduct of research.

2.8 Supplemental Information

Supplemental information is provided in the online version of this paper.

Chapter 3: Copy number alterations are commonly seen in pediatric brain tumors and may have prognostic value

3.1 Preamble

The project presented here represents a work in progress and will be edited for journal submission and final publication. My role in this project was to perform copy number calling for a cohort of samples from the Children's Brain Tumor Network (CBTN), explore the landscape of copy alteration for the CBTN cohort and a cohort of samples from Washington University (WashU), perform survival analysis exploring the relationship between CNVs and overall survival, generate figures, and prepare the manuscript for submission with the following proposed author list:

Freshour SL, Fisk B, Miller CA, Dahiya S, Rubin JB, Griffith OL, Griffith M.

3.2 Summary

Brain and central nervous system tumors are the most common form of pediatric solid tumor cancers and the second most common cancer overall among children. Although advances have been made in understanding the genomics of pediatric brain tumors and copy number alteration has been explored in the context of molecular subtyping and in relation to other alterations (e.g. structural variants), the role of copy number alteration in pediatric brain tumors has not been

fully characterized (Shapiro et al., 2023; Y. Yang & Yang, 2023). To examine the landscape of copy number alteration further, we characterized copy number alteration across the genomes of almost 300 pediatric brain tumor samples across four diagnosis groups (ATRT, Ependymoma, High-Grade Glioma, and Medulloblastoma) and found that copy alterations were common and, in some cases, extensive across diagnosis groups. In addition to characterizing copy alteration burden (aka CNV burden), we also characterized the landscape of recurrent copy alterations across samples. Survival analysis of CNV burden and recurrent alterations suggested that both CNV burden and certain recurrent alterations could have prognostic value with respect to specific diagnosis groups. These results indicated that copy number alteration may play an important role in pediatric brain tumors and should be explored further.

3.3 Introduction

Pediatric brain and central nervous system tumors are the second most common cancer type, the most common solid tumor malignancy, and the leading cause of cancer related deaths in children (S. Subramanian & Ahmad, 2023). They encompass a range of tumor types and subtypes with distinct molecular profiles. Likewise, there is a large range in overall survival depending on the tumor type and subtype. For example, High-Grade Glioma (HGG) has one of the lowest 5-year survival rates (~33%) and many HGG patients do not survive more than two years after their diagnosis (Damodharan & Puccetti, 2023; Ostrom et al., 2022). In contrast, Ependymal tumors have been shown to have 5-year survival rates of ~75% (Ostrom et al., 2022). Even within broad tumor types, such as embryonal tumors, there can be great variability in survival rates. For example, the 5-year survival rate for Medulloblastoma, a common type of embryonal tumor, is over 70% (*Medulloblastoma Diagnosis and Treatment*, 2018). In contrast, the 5-year survival

rate for atypical teratoid/rhabdoid tumors (ATRTs), a rare type of embryonal tumor, is only about 33% (Damodharan & Puccetti, 2023).

While survival rates have improved for certain subsets of pediatric brain tumor types over time (e.g. Medulloblastoma and Low-Grade Glioma), survival rates for other types have not seen much improvement (e.g. High-Grade Glioma) (Kulubya et al., 2022). Understanding the molecular landscape of pediatric brain tumors through genomic profiling has been an important approach for improving accuracy of tumor diagnosis, uncovering possible targets for therapy, and formulating more personalized treatment plans. For example, identification of mutations within genes of the MAPK (mitogen-activated protein kinase) signaling pathway as a common driver of Low-Grade Glioma has led to multiple clinical trials looking at the potential utility of inhibitors targeting MAPK signaling related genes, e.g. MEK inhibitors and BRAF inhibitors (Barbato et al., 2023; Manoharan et al., 2023; Mueller et al., 2020). Nevertheless, there remains a need to continue elucidating the molecular biology and genomic landscape of pediatric brain tumors and their potential prognostic value.

One type of genomic alteration that remains under-explored in the context of pediatric brain tumors is copy number alteration. While there have been studies looking at copy number alteration in pediatric brain tumors, which have reported commonly seen alterations and some associations with overall survival, the extent of copy number alteration and its potential prognostic value has not been fully explored. Thus, the aim of this project was to explore the landscape of copy number alteration further and determine what relationship there may be between copy alteration burden and overall survival as well as what relationship there may be between specific copy alterations and overall survival.

3.4 Results

3.4.1 Copy number alterations are common, sometimes extensive in pediatric Ependymoma, High-Grade Glioma, and Medulloblastoma

To characterize the landscape of copy number alteration in pediatric brain tumors, we first analyzed a cohort of diagnostic tumor samples from patients at Washington University School of Medicine in St. Louis (WashU) using low pass (5x coverage) whole genome sequencing. Samples were obtained from 67 patients (median age: 7.09 years) across four different diagnosis groups (Table 3.1). As these patients did not have matched normal samples, a panel of 48 normal samples was used as the unmatched normal for copy number analysis (Methods). We performed copy number calling for each sample using CNVkit (Talevich et al., 2016). We then manually corrected CNVkit's copy number calling results to rescue alterations that may have been missed by the initial copy number calling (Figures 3.1 - 3.3, Methods).

Figure 3.1: Workflow for generating and analyzing copy number calling results

Schematic of the steps involved in data collection, data processing to generate CNV and LOH calls, manual correction of CNV calls, and downstream analysis of CNVs and overall survival.

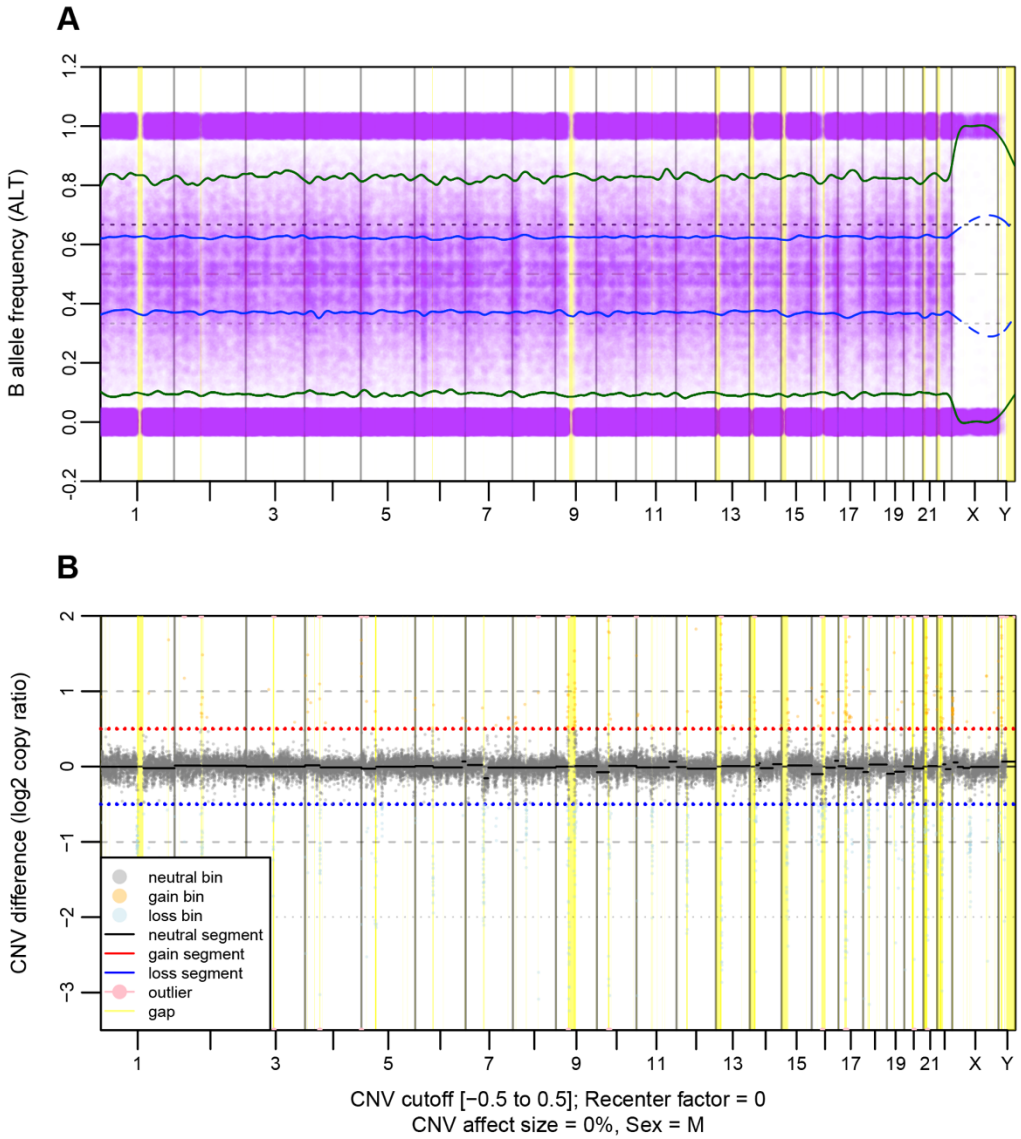


Figure 3.2: Example of LOH and CNV plots generated for a patient with a quiet genome

(A) *Example of an LOH plot generated for an Ependymoma sample which does not exhibit LOH or copy number alteration. These LOH plots are used to inform manual copy number calling corrections based on regions exhibiting both copy alterations and loss-of-heterozygosity. LOH is*

determined by calculating the b allele frequency (BAF) for a subset of heterozygous SNP positions across the genome. Green lines are smoothed lines of BAFs across the genome, including SNP positions with BAFs of 0% and 100%. Blue lines are smoothed lines of BAFs across the genome, excluding SNPs positions with BAFs of 0% and 100%. Lines are smoothed by applying the smooth.spline function in R, which uses cubic regression to estimate values in between the plotted BAFs. Purple dots represent BAFs of individual SNPs. **(B)** *Example of a CNV plot generated from copy number calling results for an Ependymoma sample which does not exhibit LOH or copy number alteration. Copy number calling was performed using CNVkit with 100k bin size. Dots represent read count differences in bins. Solid lines represent segments identified by CNVkit using circular binary segmentation. The dotted red line at 0.5 marks the cutoff above which copy alterations are classified as gains. The dotted blue line at -0.5 marks the cutoff below which copy alterations are classified as losses. The yellow bars correspond to GRCh38 assembly gap positions in the genome (as reported in the UCSC Genome Browser gap track).*

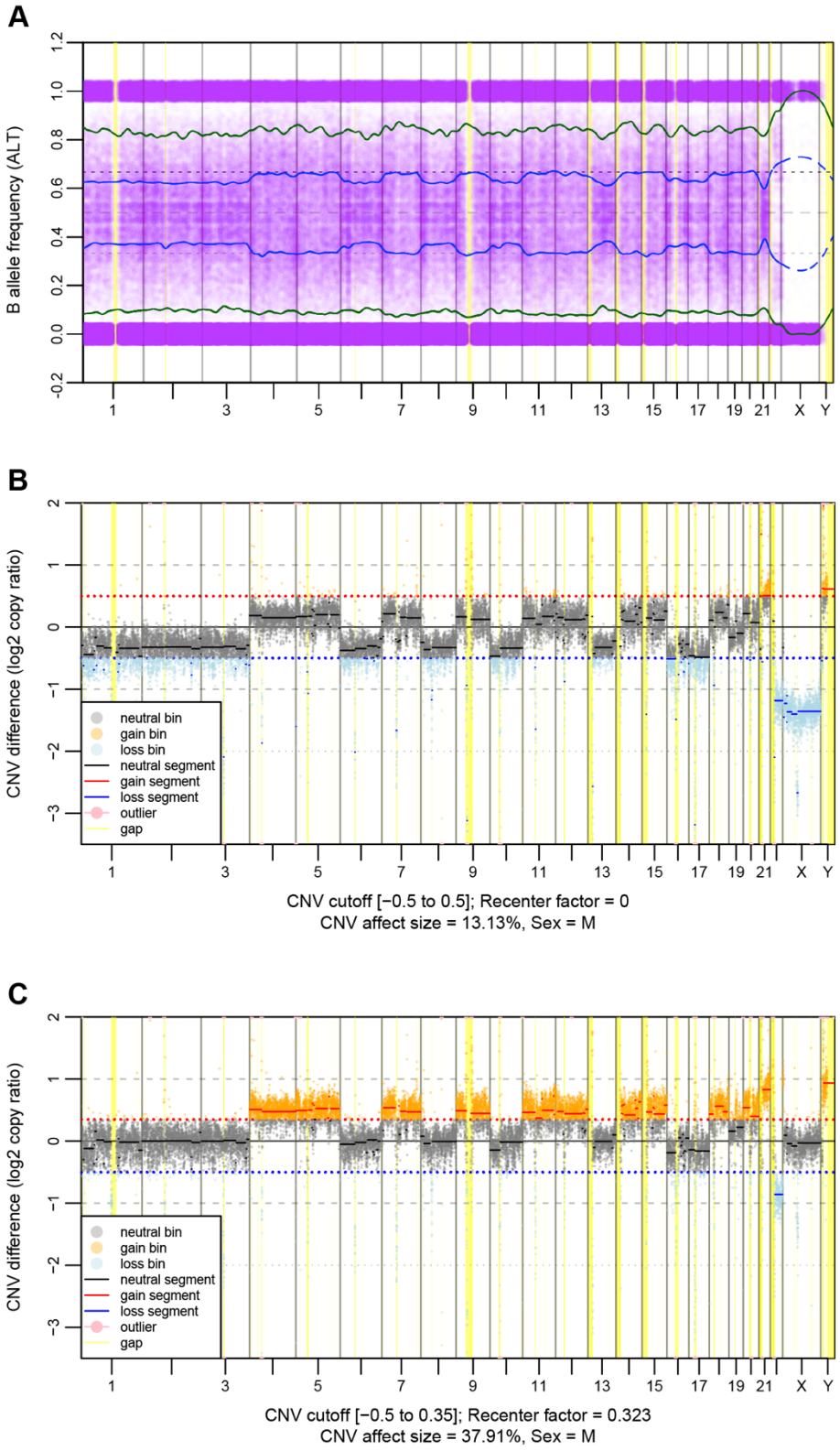

A
B allele frequency (ALT)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
1
3
5
7
9
11
13
15
17
19
21
X
Y
B
CNV difference (log2 copy ratio)
2
1
0
−1
−2
−3
neutral bin
gain bin
loss bin
neutral segment
gain segment
loss segment
outlier
gap
CNV cutoff [−0.5 to 0.5]; Recenter factor = 0
CNV affect size = 13.13%, Sex = M
C
CNV difference (log2 copy ratio)
CNV cutoff [−0.5 to 0.35]; Recenter factor = 0.323
CNV affect size = 37.91%, Sex = M

Figure 3.3: Example of LOH and CNV plots generated for a patient with copy alteration detected

(A) *Example of an LOH plot generated for an Ependymoma sample which does exhibit LOH and copy number alteration and had manual corrections applied to its copy number calling results.* **(B)** *Example of an uncorrected CNV plot generated for an Ependymoma sample which does exhibit LOH and copy number alteration and had manual corrections applied to its copy number calling results.* **(C)** *Example of a corrected CNV plot generated for an Ependymoma sample which does exhibit LOH and copy number alteration and had manual corrections applied to its copy number calling results. This sample had a recentering factor of 0.323 applied and the upper cutoff for calling a copy number gain lowered from 0.5 to 0.35.*

After copy number calling results were generated for all samples, we calculated the extent of copy alteration for each sample using two metrics: the percentage of copy altered base pairs detected across the genome (CNV%) and the total number of copy altered segments called across the genome (CNV segment count) (Methods). Both of these metrics indicated that copy number alteration was common and, in some cases, extensive in the WashU cohort (Figures 3.4A, 3.4C). The majority of samples displayed some level of alteration (~90% samples had at least one copy-altered segment), with multiple samples (17 of 67) having more than 25% of their genomes altered (Figure 3.4A, 3.4C). While the CNV% and CNV segment count metrics were highly correlated, the extent of correlation varied by diagnosis group and was not a perfect correlation (Figure 3.5). Several samples appeared to have disparate extents of copy alteration per the two metrics, i.e. some samples had higher CNV segment count, but lower CNV% and vice versa. Therefore, we chose to include both metrics in our analysis.

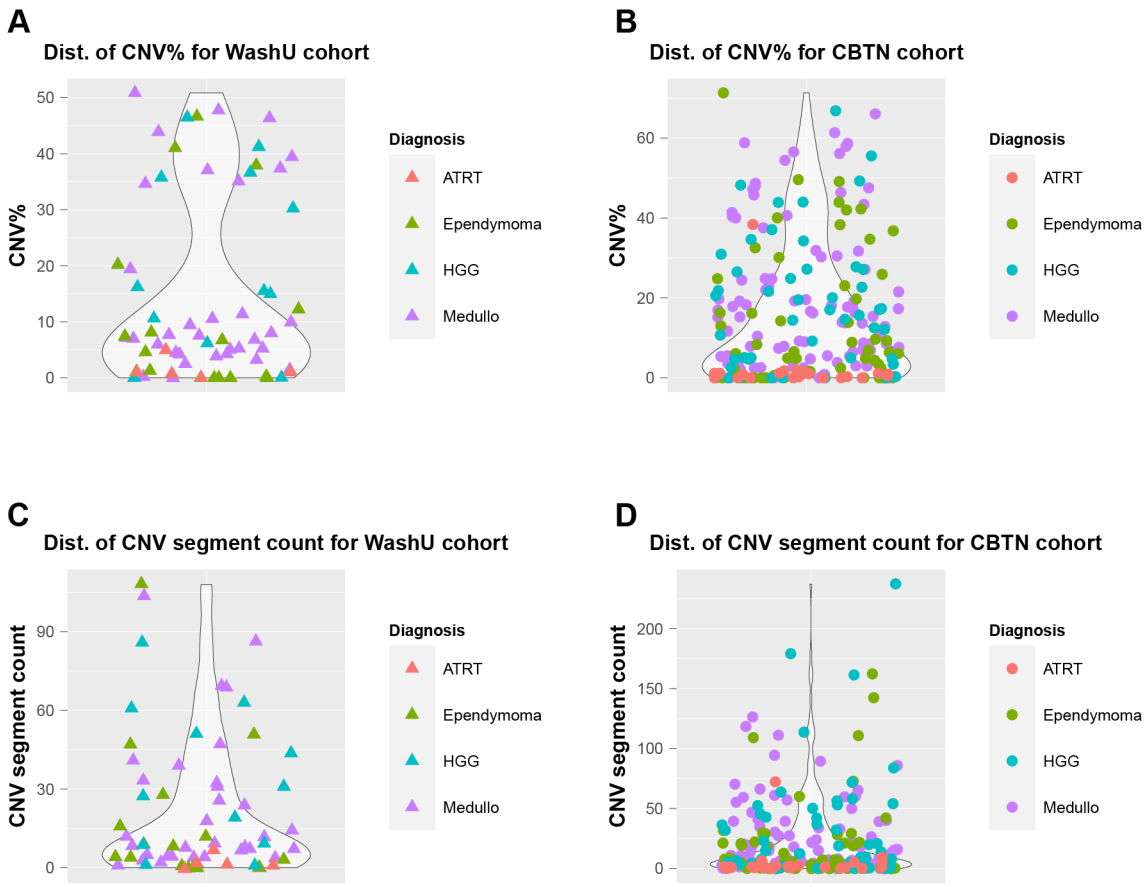


Figure 3.4: Distributions of CNV% and CNV segment count within each cohort

(A) *Distribution of CNV% for all samples in the WashU cohort.* **(B)** *Distribution of CNV% for all samples in the CBTN cohort. CNV% is calculated as the number of copy altered base pairs divided by the approximate size of the human genome (i.e. 3.2 billion base pairs). Each point corresponds to a single sample. The value on the y-axis indicates the CNV%. The point color indicates the diagnosis group. The violin plot shows the overall distribution of CNV% for the given set of samples.* **(C)** *Distribution of CNV segment count for all samples in the WashU cohort.* **(D)** *Distribution of CNV segment count for all samples in the CBTN cohort. Each point corresponds to a single sample. The value on the y-axis indicates the CNV segment count. The*

point color indicates the diagnosis group. The violin plot shows the overall distribution of CNV segment count for the given set of samples.

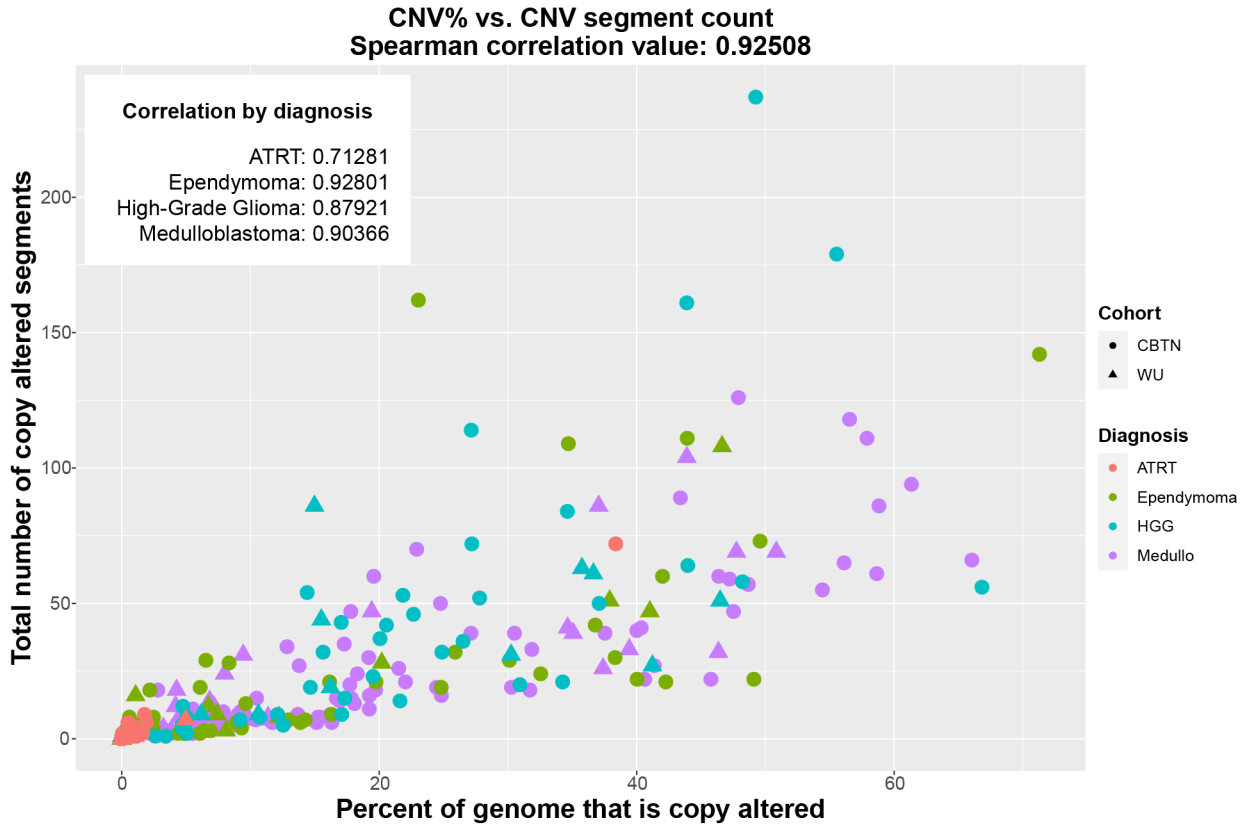


Figure 3.5: CNV% versus CNV segment count with Spearman correlation values

Scatter plot of CNV% versus CNV segment count for all diagnosis groups across both cohorts. Point shape indicates cohort. Point color indicates diagnosis group. The Spearman correlation value between CNV% and CNV segment count for the full dataset was ~0.925. Spearman correlation values within each diagnosis group are listed on the plot.

To increase our sample size and expand on the observations seen in the WashU cohort, we obtained an additional cohort of samples from the Children's Brain Tumor Network (CBTN) with whole genome sequencing available for both tumor and matched normal samples. In total,

we added 226 additional samples (median age: 7.47 years) across the same four diagnosis groups, bringing the grand total of samples to 293 tumors (Table 3.1, Methods). As with the WashU cohort, we generated copy number calling results for these samples using CNVkit and manual correction, then calculated the percentage of copy altered base pairs and the number of copy altered segments for each sample. Similar to the results seen in the WashU cohort, these metrics indicated that copy alteration was common (~84% of samples had at least one copy altered segment) and, in some cases, extensive (53 of 226 samples had over 25% of their genomes altered) in the CBTN cohort (Figures 3.4B, 3.4D). However, the extent of alteration observed varied by diagnosis group. In particular, ATRT samples generally had quiet genomes (with CNV% ranging from 0% to ~1.9% and CNV segment count ranging from 0 to 9), with the exception of one outlier which had ~38% alteration and 72 altered segments (Figures 3.6A - 3.6C).

Diagnosis Group	WashU	CBTN	Total
ATRT	5 (~7.5%)	24 (~10.6%)	29
Ependymoma	17 (~25.4%)	60 (~26.5%)	77
High-Grade Glioma	12 (~17.9%)	46 (~20.4%)	58
Medulloblastoma	33 (~49.3%)	96 (~42.5%)	129
Total	67	226	293

Table 3.1: Number of samples for each diagnosis group within each cohort

While all diagnosis groups had samples with quiet genomes, the Ependymoma, High-Grade Glioma, and Medulloblastoma diagnosis groups had greater variety in the extent of alteration, from samples with no alteration detected to samples with alteration detected in only a

few chromosomes to samples with alteration detected across numerous chromosomes (Figure 3.6C). Clustering of copy number alterations within diagnosis groups also revealed the presence of patterns of alteration within and across diagnosis groups (Figure 3.6C). For example, several distinct patterns within the Medulloblastoma group were chromosome 6 deletions (with little alteration in other chromosomes), 17p deletions paired with 17q gains, and chromosome X losses in female patients, all of which are previously reported patterns of alteration seen in certain types of pediatric Medulloblastoma (Doussouki et al., 2019; Kool et al., 2008). In contrast, other alterations appeared to be fairly common across diagnosis groups, such as 1q gains and chromosome 7 gains in Ependymoma, High-Grade Glioma, and Medulloblastoma. These results suggested that, although copy number alteration appears to be a common occurrence in pediatric brain tumors, extensive copy number alteration is more common in certain diagnosis groups (i.e. Ependymoma, High-Grade Glioma, and Medulloblastoma). Additionally, specific copy number alterations can be distinct to diagnosis groups or detectable across multiple diagnosis groups.

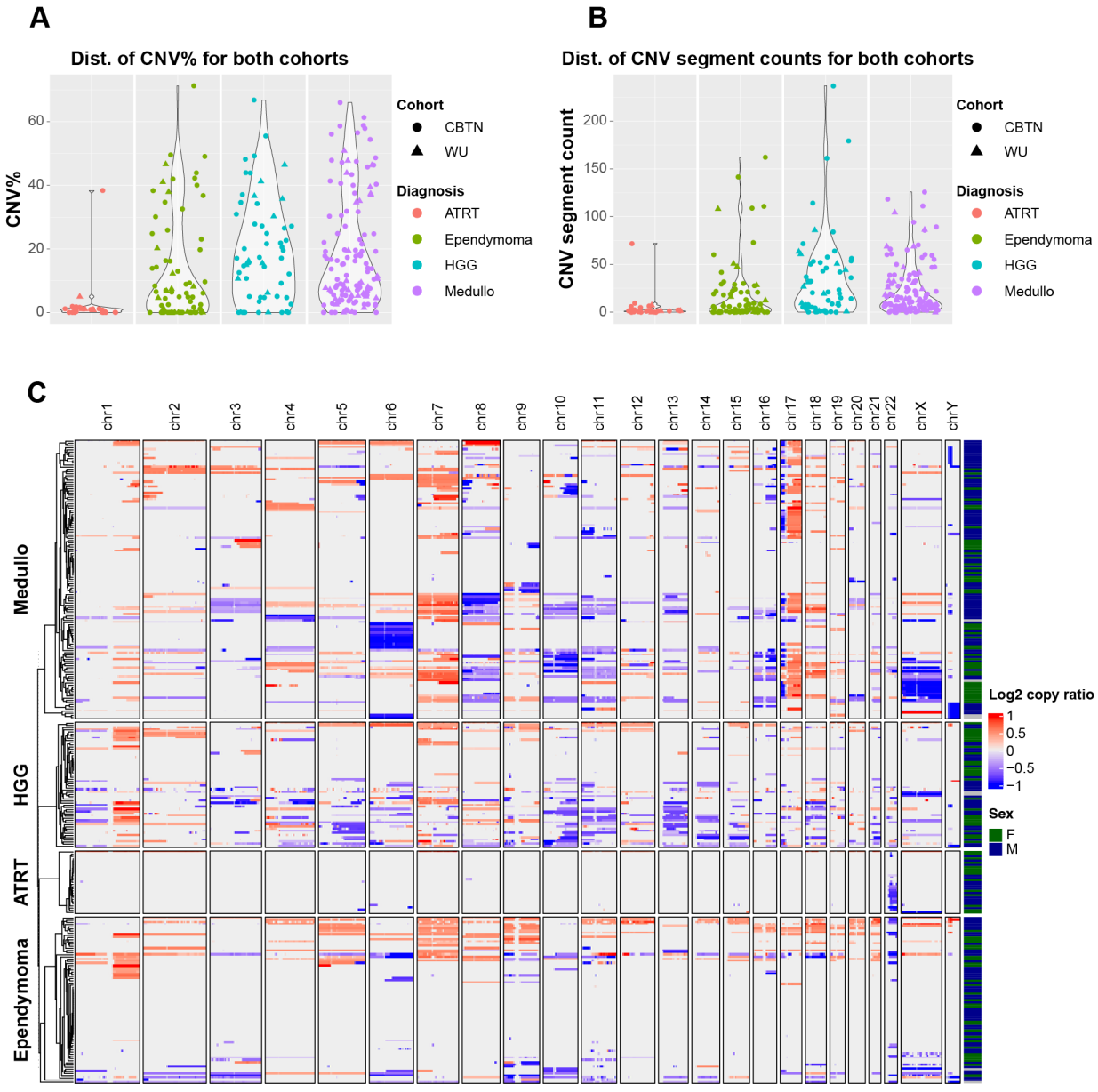


Figure 3.6: Landscape of copy number alteration within diagnosis groups

(A) *Distributions of CNV% for all samples across both cohorts faceted by diagnosis group. The value on the y-axis indicates the CNV%.* **(B)** *Distributions of CNV segment count for all samples across both cohorts faceted by diagnosis group. The value on the y-axis indicates the CNV segment count. Each point corresponds to a single sample. The point color indicates the*

diagnosis group. The point shape indicates the cohort. The violin plots show the overall distribution for the given set of samples. **(C)** *Heatmap of average log2 copy ratios across the genome, split into 1MB sections, for all samples across both cohorts faceted by diagnosis group. Rows correspond to samples. Columns correspond to genomic regions. Within each diagnosis group, samples were clustered by rows. Red indicates a copy number gain. Blue indicates a copy number loss.*

3.4.2 Recurrent alterations are detected across diagnosis groups and show patterns within diagnosis groups

After exploring the full landscape of copy alteration, we next wanted to explore the landscape of recurrent copy number alterations. To identify recurrently altered regions across autosomal chromosomes, we ran GISTIC with the copy number calling results for the full dataset containing all diagnosis groups and with the copy number calling results for each diagnosis group separately (Methods). In the full dataset, 39 recurrently altered regions were identified, 18 of which were classified as recurrent amplifications and 21 of which were classified as recurrent deletions (Figure 3.7).

Clustering samples based on their amplification and deletion patterns in these recurrently altered regions revealed several interesting patterns. For example, across diagnosis groups there was a subset of samples which showed no presence of recurrent alterations (Figure 3.7). Some recurrent alterations appeared to be largely detected in specific diagnosis groups, such as 17q amplifications in Medulloblastoma samples and 22q deletions in ATRT samples, which were also seen in the clustering of all CNVs across the genome (Figures 3.6C, 3.7). Both these observations are consistent with previous reports of 17q amplifications (typically paired with 17p

losses) in Medulloblastoma and 22q deletions in ATRT (J. Y. Choi, 2023; Ho et al., 2020). These results suggested that while some recurrent alterations can be shared across diagnosis groups, certain recurrent alterations can still distinguish specific diagnosis groups.

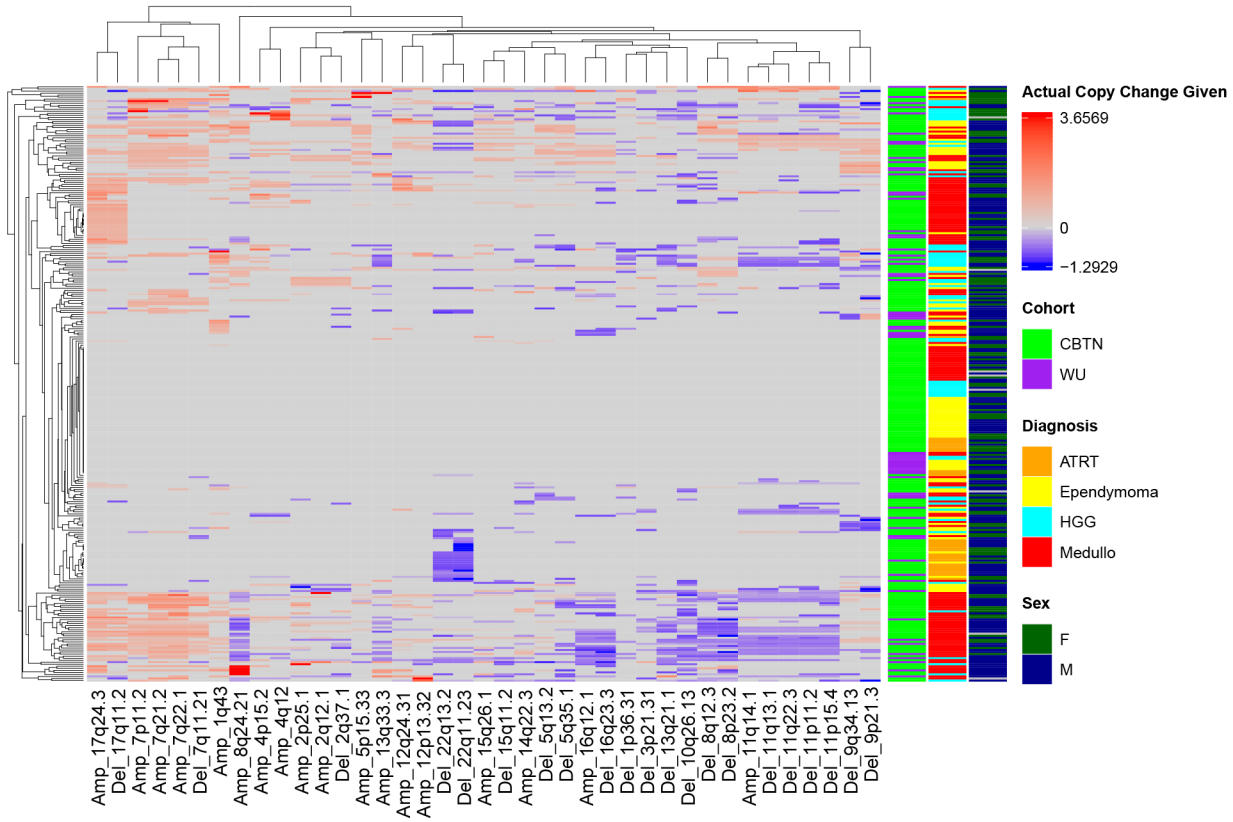


Figure 3.7: Heatmap of recurrent alterations across all diagnosis groups

Heatmap of recurrently altered regions identified by GISTIC in the full dataset containing all diagnosis groups across both cohorts. Rows correspond to samples. Columns correspond to recurrent alterations. The color of the squares indicates "Actual Copy Change Given" as reported by GISTIC. Red indicates a copy gain. Blue indicates a copy loss.

Within each individual diagnosis group, clustering samples based on recurrent amplifications and deletions revealed patterns within diagnosis groups as well. For example, a

subset of Ependymoma samples appeared to harbor 7q amplifications and 9q amplifications together, while another subset of samples had no 7q alterations, but some 9q deletions (Figure 3.8B). In the Medulloblastoma dataset, there appeared to be two subsets of samples with recurrent 17q amplifications: one group which harbored multiple other recurrent alterations and one which appeared to primarily harbor recurrent 17q amplifications and few other recurrent alterations (Figure 3.8D). These samples likely represent group 3 and group 4 Medulloblastoma samples. Group 3 and group 4 Medulloblastomas have both been shown to frequently harbor 17q alterations (although they are more commonly seen in group 4), but have different rates of other alterations (e.g. group 3 is more likely to harbor losses in 10q) (Kijima & Kanemura, 2016; Kool et al., 2012; Thomas & Noël, 2019). The High-Grade Glioma group appeared to have few distinct patterns, with the exception of samples that appeared to harbor no recurrent alterations (Figure 3.8C). Overall, these patterns indicated that, within certain diagnosis groups, there can be distinct patterns of recurrent alteration detected.

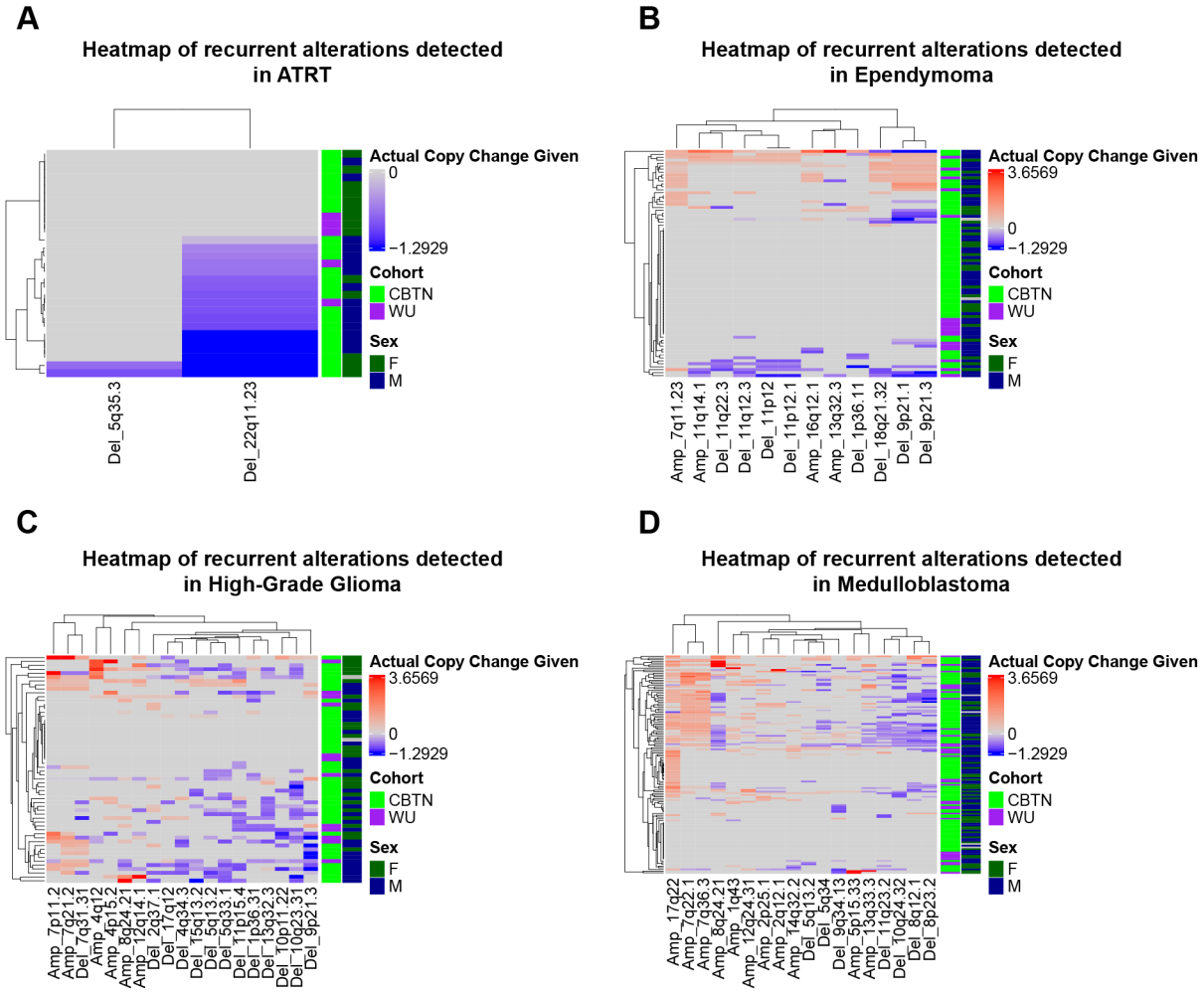


Figure 3.8: Heatmaps of recurrent alterations within diagnosis groups

(A) *Heatmap of recurrently altered regions identified by GISTIC in the ATRT dataset across both cohorts.* **(B)** *Heatmap of recurrently altered regions identified by GISTIC in the Ependymoma dataset across both cohorts.* **(C)** *Heatmap of recurrently altered regions identified by GISTIC in the High-Grade Glioma dataset across both cohorts.* **(D)** *Heatmap of recurrently altered regions identified by GISTIC in the Medulloblastoma dataset across both cohorts. Rows correspond to samples. Columns correspond to recurrent alterations. The color of the squares indicates "Actual Copy Change Given" as reported by GISTIC. Red indicates a copy gain. Blue indicates a copy loss.*

3.4.3 Correlation between CNV burden and overall survival appears to be dependent on diagnosis

Before exploring the relationship between the extent of copy alteration in each sample and overall survival, we first verified differences in overall survival by diagnosis group (Figure 3.9). This analysis showed that the High-Grade Glioma and ATRT groups appeared to have significantly worse overall survival than the Ependymoma and Medulloblastoma groups, as has been previously reported (Ostrom et al., 2015; Pogorzala et al., 2014; Stanić et al., 2021). Since there were significant differences in overall survival by diagnosis group, we performed most of the following analysis within each diagnosis group separately.

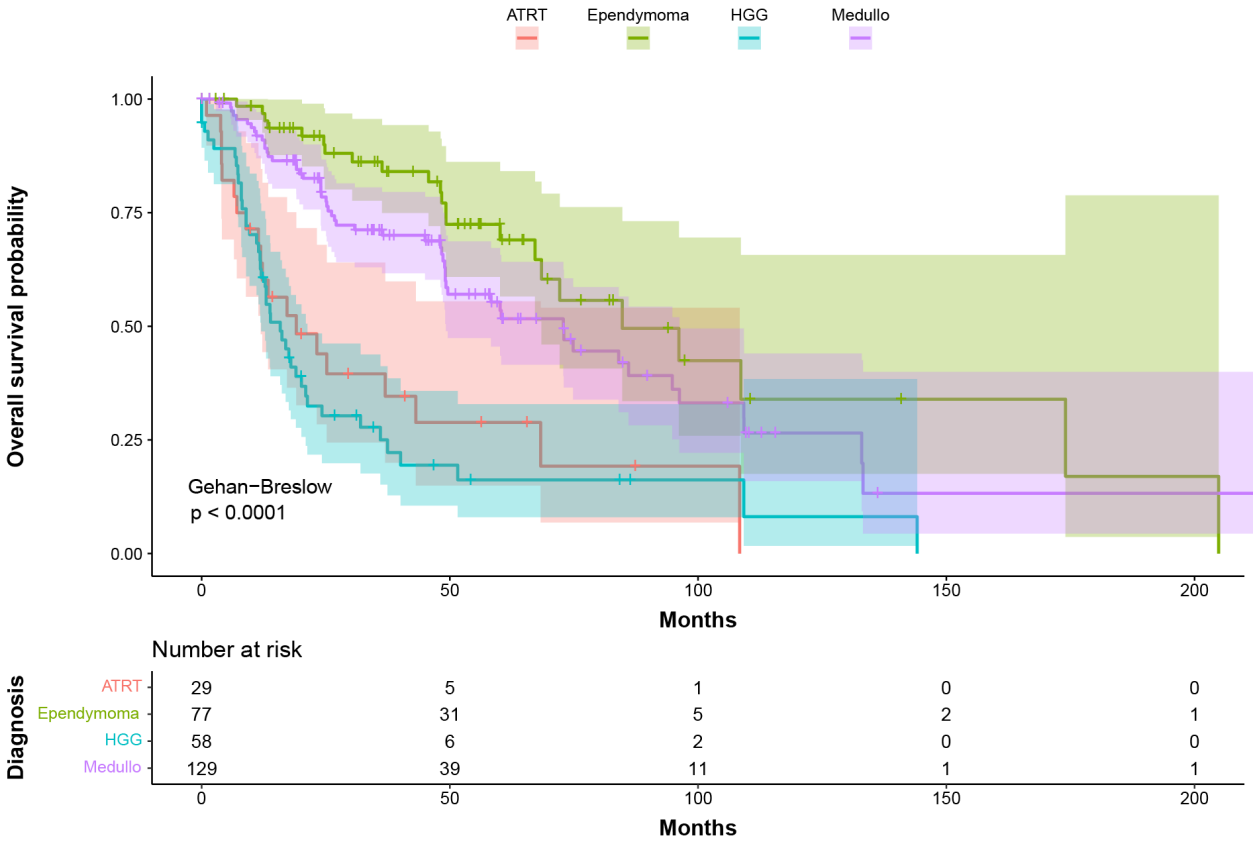


Figure 3.9: Kaplan-Meier survival curves split by diagnosis group

Kaplan-Meier survival curves split by diagnosis group (top) and the corresponding risk table (bottom) for the full dataset containing all diagnosis groups across both cohorts. Curves are colored by the diagnosis group. Boxes around the curves indicate the 95% confidence intervals.

Next, we investigated whether there was any correlation between copy alteration burden (aka CNV burden) and overall survival. To do this, we looked at both CNV% and CNV segment count as two separate metrics of CNV burden and explored the relationship between each of these metrics and overall survival. We performed analysis using these two CNV burden metrics as continuous variables and also as categorical variables based on tertiles (i.e. high, mid, and low burden), based on the average value of each metric (i.e. high and low burden), and based on identifying the optimal cutpoint for each metric (i.e. high and low burden).

For exploration of the relationship between overall survival and CNV burden as a continuous variable, we calculated the Spearman correlation values between CNV% and overall survival as well as CNV segment count and overall survival (Methods). These calculations were done within each diagnosis group separately. Although the correlation values varied by diagnosis group, the correlation values for both CNV% and CNV segment count were fairly low across diagnosis groups. The absolute values for correlation ranged from ~0.036 to ~0.111 for CNV% and from ~0.017 to ~0.157 for CNV segment count, indicating there was no strong correlation between continuous CNV burden and overall survival within any diagnosis group (Figures 3.10A - 3.10D, 3.11A - 3.11D).

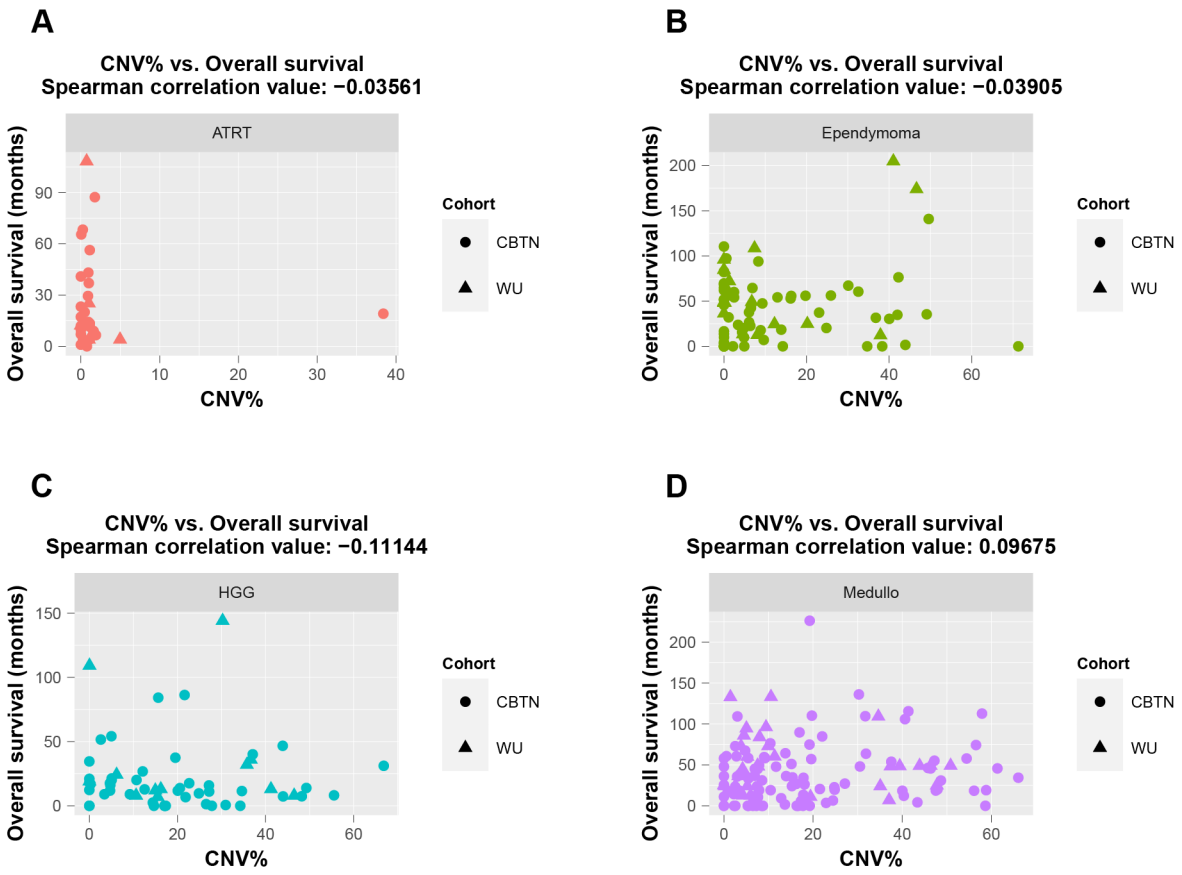


Figure 3.10: CNV% versus overall survival within each diagnosis group

(A) *Scatter plot of CNV% versus overall survival for the ATRT group across both cohorts.* **(B)** *Scatter plot of CNV% versus overall survival for the Ependymoma group across both cohorts.* **(C)** *Scatter plot of CNV% versus overall survival for the High-Grade Glioma group across both cohorts.* **(D)** *Scatter plot of CNV% versus overall survival for the Medulloblastoma group across both cohorts. Point shape indicates cohort. Point color indicates diagnosis group. Spearman correlation values for each group are included in the corresponding plot titles.*

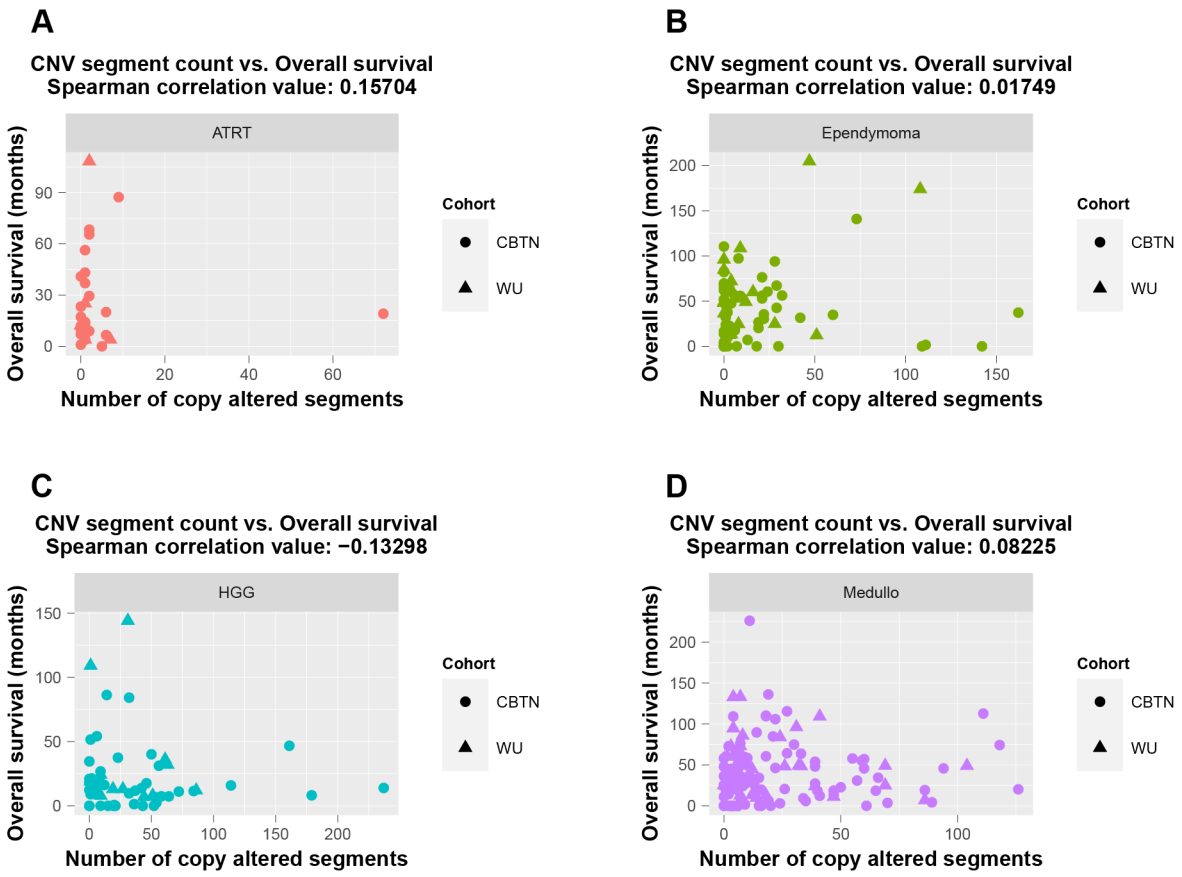


Figure 3.11: CNV segment count versus overall survival within each diagnosis group

(A) *Scatter plot of CNV segment count versus overall survival for the ATRT group across both cohorts.* **(B)** *Scatter plot of CNV segment count versus overall survival for the Ependymoma group across both cohorts.* **(C)** *Scatter plot of CNV segment count versus overall survival for the High-Grade Glioma group across both cohorts.* **(D)** *Scatter plot of CNV segment count versus overall survival for the Medulloblastoma group across both cohorts. Point shape indicates cohort. Point color indicates diagnosis group. Spearman correlation values for each diagnosis group are included in the corresponding plot titles.*

Next, we categorized CNV burden based on CNV% and CNV segment count separately and explored the relationship between overall survival and high/low or high/mid/low CNV

burden within each diagnosis group (with the exception of ATRT as only one sample in that group had high CNV% and CNV segment count). For each metric (i.e. CNV% and CNV segment count), we classified samples as high/low CNV burden based on two methods. The first was by calculating the average for each metric in each diagnosis group (Supplemental Table S3.1, Methods). The second was by identifying a single optimal cutpoint for each metric in each diagnosis group (Table 3.2, Methods). Additionally, for each metric, we classified samples as high/mid/low CNV burden based on identifying tertiles for each metric in each diagnosis group (Supplemental Table S3.2, Methods).

Classifying CNV burden based on tertiles suggested there was no significant association between overall survival and high/mid/low CNV burden within any diagnosis group (Supplemental Figures S3.1 - S3.3, Supplemental Table S3.2). Similarly, classifying CNV burden as high/low based on averages indicated there was no significant association between high/low CNV burden and overall survival in any diagnosis group (Supplemental Figures S3.4 - S3.6, Supplemental Table S3.1). However, the association between high CNV burden based on average CNV segment count and worse overall survival in the High-Grade Glioma group approached significance (Supplemental Figure S3.5, Supplemental Table S3.1).

Classification of high/low CNV burden based on the optimal cut points of CNV% and CNV segment count indicated that the relationship between CNV burden and overall survival may be dependent on both the metric chosen for classification and the disease context. In the Ependymoma group, CNV burden based on CNV% suggested a significant correlation between high CNV burden and better overall survival, while CNV segment count showed no significant relationship between CNV burden and overall survival (Figure 3.12, Table 3.2). However, these results must be considered with the caveat that the Ependymoma group had a limited number of

high CNV burden samples (8 by CNV% and 11 by CNV segment count). In the High-Grade Glioma group, both metrics (CNV% and CNV segment count) indicated that high CNV burden was associated with worse overall survival (Figure 3.13, Table 3.2). Meanwhile, in the Medulloblastoma group, neither metric suggested there was any relationship between CNV burden and overall survival differences (Figure 3.14, Table 3.2). Together, these results indicated that potential associations between CNV burden and overall survival are dependent on disease context and classification metric.

Diag.	Metric	Cutpoint	KM Method	KM Pval
Ependymoma	CNV%	40.038	Log-Rank	0.019
Ependymoma	CNV segment count	30	Gehan-Breslow	0.48
HGG	CNV%	6.187	Gehan-Breslow	0.032
HGG	CNV segment count	32	Gehan-Breslow	0.022
Medullo	CNV%	19.19	Gehan-Breslow	0.66
Medullo	CNV segment count	22	Log-Rank	0.17

Table 3.2: Survival analysis results for categorizing CNV burden based on optimal cutpoints

Table of survival analysis results for categorizing CNV burden based on optimal cutpoints for CNV% and CNV segment count within each diagnosis group across both cohorts. Samples with CNV% or CNV segment count below the corresponding cutpoint value were classified as "CNV Low". Samples with CNV% or CNV segment count above the corresponding cutpoint value were classified as "CNV High". **KM Method** *indicates the statistical method used to test for significant differences in the Kaplan-Meier survival estimates.*

However, plotting the test statistics used in the selection of the optimal cutpoints for each metric within each diagnosis group suggested that the values selected for these cutpoints may be somewhat arbitrary, leading to arbitrary significance as well. For example, in the Ependymoma

group, the test statistics for CNV% appeared to generally increase as the values for CNV% increased until reaching the last value tested (Supplemental Figure S3.7). In the Medulloblastoma group, several different CNV% values seemed to produce similarly high test statistics (Supplemental Figure S3.9). CNV segment count in the High-Grade Glioma group appeared to have the clearest peak in its test statistics, suggesting a less arbitrary selection of the optimal cutpoint (Supplemental Figure S3.8). Regardless, follow-up analysis will be needed in larger, independent cohorts to validate the potential significance of the CNV burden cut points presented here and CNV burden in general.

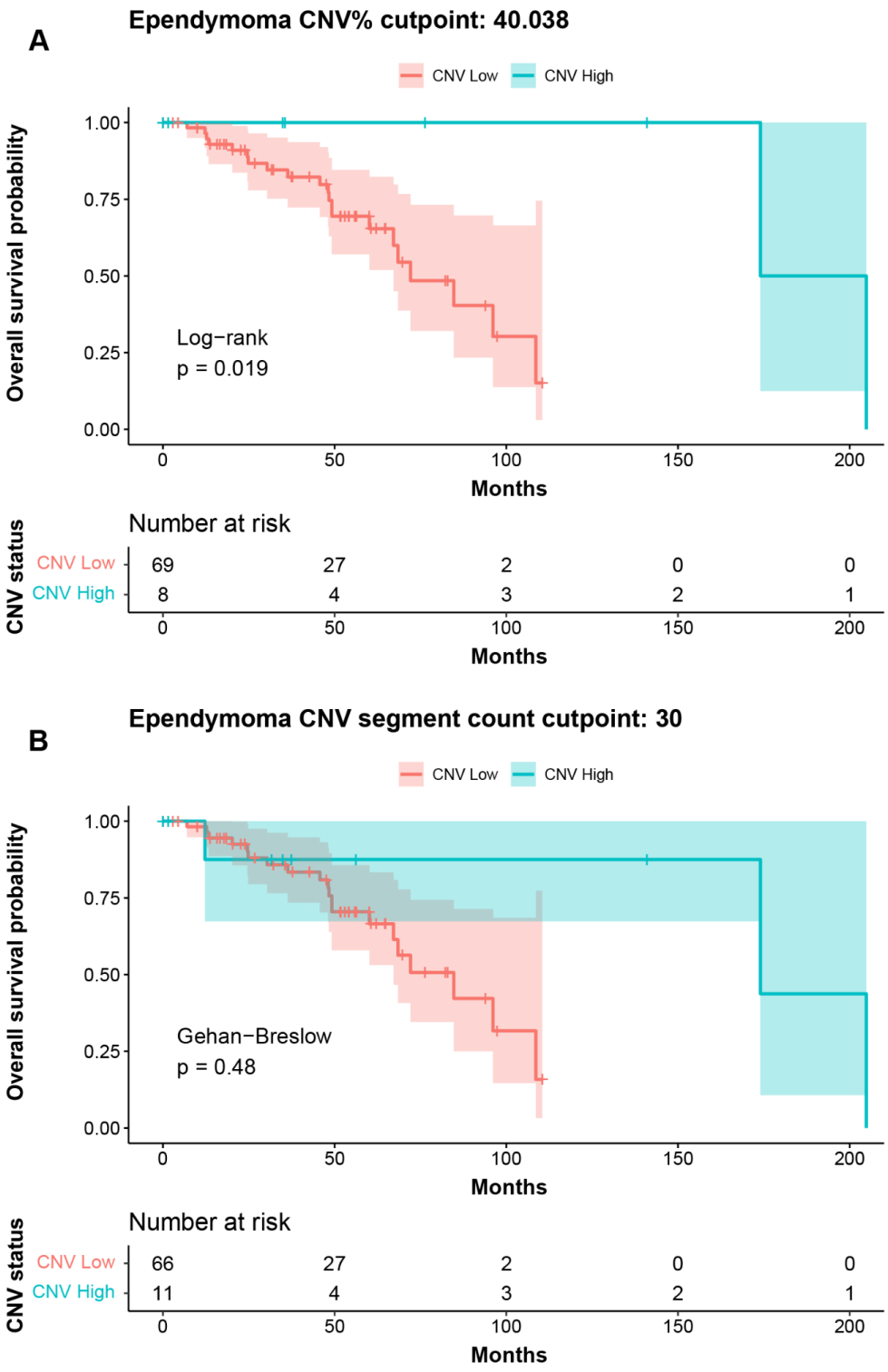


Figure 3.12: Kaplan-Meier survival curves split by CNV burden based on optimal cutpoints for the Ependymoma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on the CNV% optimal cutpoint (top) and the corresponding risk table (bottom) in the Ependymoma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on the CNV segment count optimal cutpoint (top) and the corresponding risk table (bottom) in the Ependymoma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

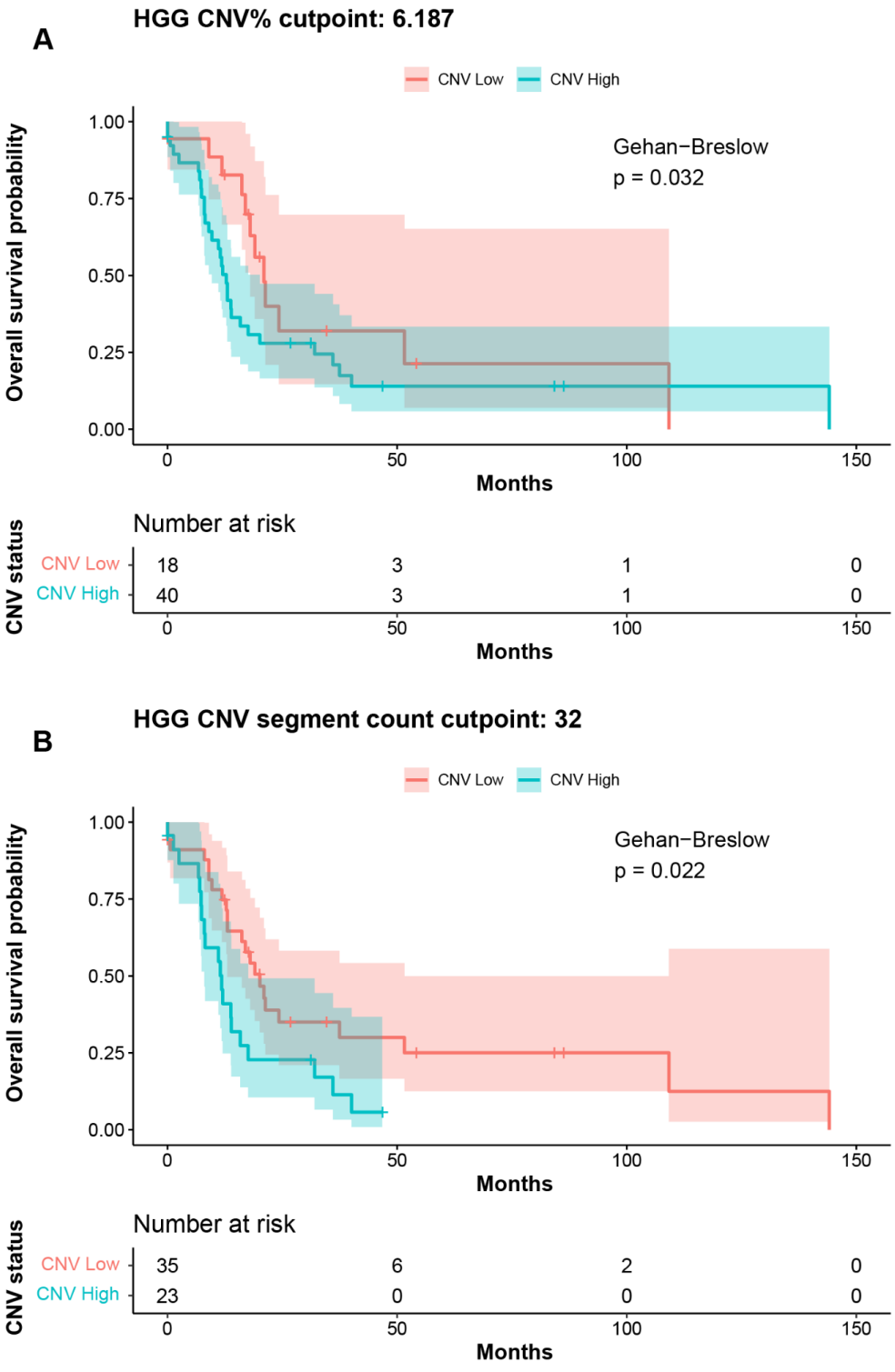


Figure 3.13: Kaplan-Meier survival curves split by CNV burden based on optimal cutpoints for the High-Grade Glioma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on the CNV% optimal cutpoint (top) and the corresponding risk table (bottom) in the High-Grade Glioma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on the CNV segment count optimal cutpoint (top) and the corresponding risk table (bottom) in the High-Grade Glioma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

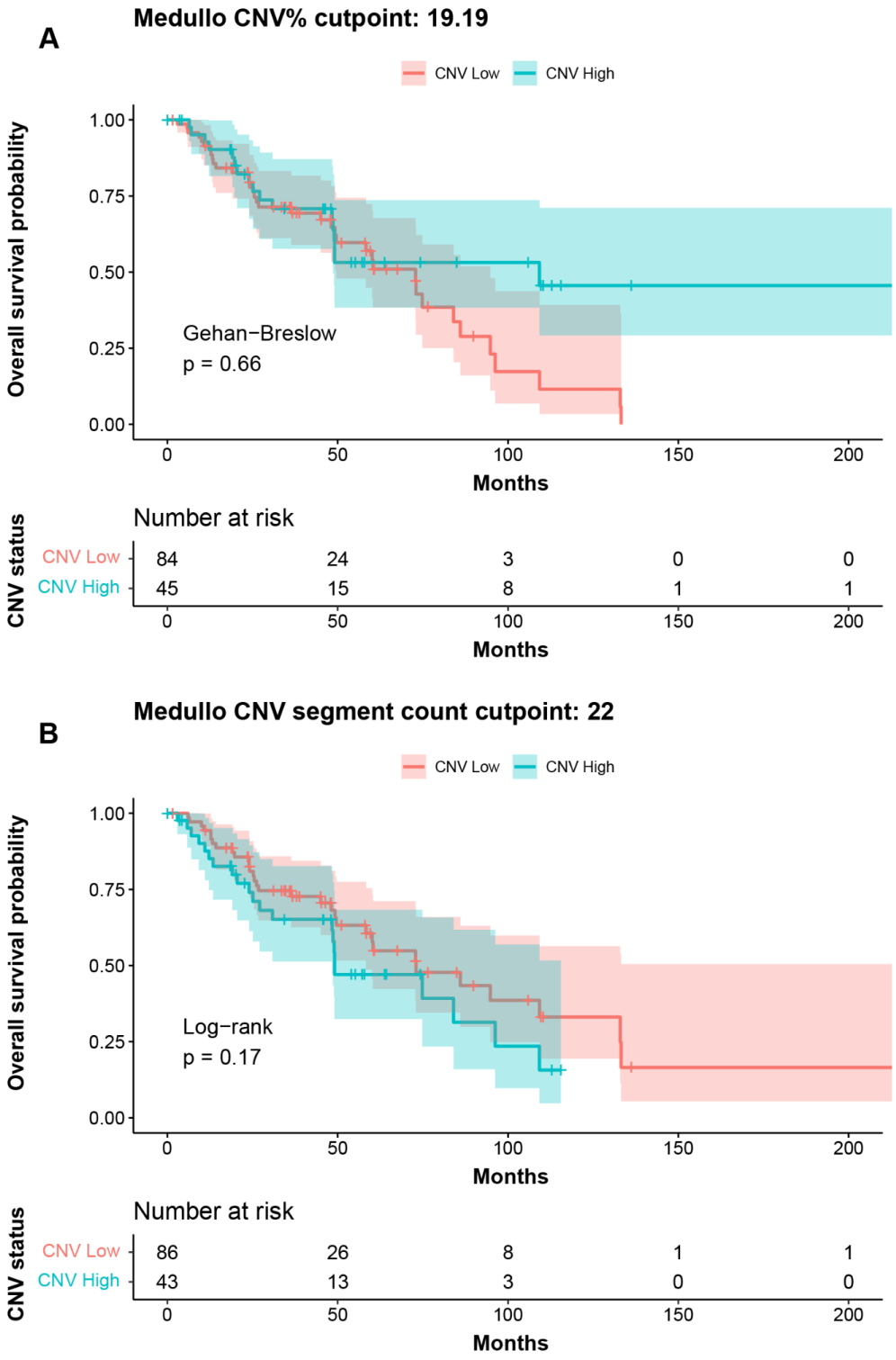


Figure 3.14: Kaplan-Meier survival curves split by CNV burden based on optimal cutpoints for the Medulloblastoma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on the CNV% optimal cutpoint (top) and the corresponding risk table (bottom) in the Medulloblastoma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on the CNV segment count optimal cutpoint (top) and the corresponding risk table (bottom) in the Medulloblastoma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

3.4.4 Recurrent alterations may be associated with changes in survival

Next, we explored the relationship between recurrent copy number alterations and overall survival. First, we used the results from GISTIC to more broadly classify the gain and loss statuses of recurrently altered chromosome arms (Methods). We chose this approach because we found that the results reported by GISTIC could be too focal in some cases, either splitting single copy alteration events into multiple events (e.g. 11p or 11q deletions) or only returning a subsection of larger events for some samples (e.g. 22q deletions). Therefore, instead of using the direct results from GISTIC for our analysis, we collapsed recurrent alterations reported on the same chromosome arm into single recurrent alteration events and classified the gain/loss status of each sample based on the presence of any gain or loss on each chromosome arm of interest (Methods).

Furthermore, for our analysis of recurrent alterations and overall survival, we used samples from the CBTN cohort only. We chose to focus on the CBTN cohort because analysis of the relationship between overall survival and cohort alone indicated that the WashU cohort had significantly worse overall survival (Supplemental Figure S3.10). Additionally, multivariate analysis performed using a Cox model with diagnosis group and cohort as categorical variables

indicated that cohort was significantly associated with differences in overall survival, even when accounting for survival differences by diagnosis (p-value < 0.0001, HR ~2.17). To avoid confounding effects based on cohort, we performed the analysis below using only the larger CBTN cohort. Likewise, as mentioned previously, we separated our analysis by diagnosis group.

After we classified the gain (or no gain) and loss (or no loss) statuses of each recurrently altered chromosome arm in each sample, we tested whether gains or losses on each chromosome arm were correlated with any significant changes in survival. To do this, we used both Kaplan-Meier survival estimates and Cox proportional hazards models (Methods). For each recurrently altered arm, we tested gain status (versus everything else) and loss status (versus everything else) separately. In total, four recurrently amplified regions (one in High-Grade Glioma, three in Medulloblastoma) and one recurrently deleted region (in High-Grade Glioma) were reported to be significantly associated with changes in survival (Table 3.3). Four of these regions (7q, 11p in HGG and 8p, 8q in Medulloblastoma) were reported as significant by both the Kaplan-Meier and the Cox model testing (Table 3.3). The remaining region (1q in Medulloblastoma) was only reported as significant by the Cox model, but was near significant by the Kaplan-Meier testing as well (Table 3.3).

Diag.	Arm	Status	Sig per KM	Sig per Cox	KM Pval Method	KM Pval	KM FDR	Cox Pval	Cox FDR	Cox HR
HGG	7q	Gain	Yes	Yes	Log-Rank	0.017	0.274	0.021	0.34	2.41
HGG	11p	Loss	Yes	Yes	Log-Rank	0.01	0.151	0.012	0.163	2.71
Medullo	1q	Gain	No	Yes	Log-Rank	0.061	0.034	0.046	0.23	2.52
Medullo	8p	Gain	Yes	Yes	Log-Rank	<0.0001	0	<0.0001	<0.0001	12.88
Medullo	8q	Gain	Yes	Yes	Log-Rank	<0.0001	0.00032	0.0002	0.001	5.46

Table 3.3: Survival analysis results for testing gain and loss statuses of recurrently altered chromosome arms

Table of results from Kaplan-Meier survival estimates and Cox proportional hazards models for recurrently altered regions reported to have significant correlation with changes in overall survival within specific diagnosis groups in the CBTN cohort.

While the role of copy number alterations has not been fully characterized in the context of pediatric brain tumors for all recurrent alterations identified in our analysis, most of the significant associations we detected between recurrent alterations and overall survival have been previously reported. In High-Grade Glioma, 7q gains have been previously reported in pediatric tumors (Barrow et al., 2011; Warren et al., 2012). Furthermore, gains on chromosome 7 have been shown to be associated with High-Grade Glioma subtypes that have worse overall survival (Phillips et al., 2006). Similarly, our analysis suggested that 7q gains are associated with worse overall survival (Figure 3.15A, Table 3.3). While 11p losses have been previously reported in adult High-Grade Glioma, they do not appear to have been previously reported in pediatric High-Grade Glioma (Fults et al., 1992; Wemmert et al., 2005). Furthermore, no association between

11p losses and overall survival seems to have been reported in either adult or pediatric High-Grade Glioma. Our results suggested that 11p losses can occur in pediatric cases and may be associated with worse overall survival (Figure 3.15B, Table 3.3). However, one caveat to these results is that the associations seen between recurrent alterations and overall survival in the High-Grade Glioma group no longer reached significance (at a cutoff of 0.05) after multiple testing correction was applied (Table 3.3).

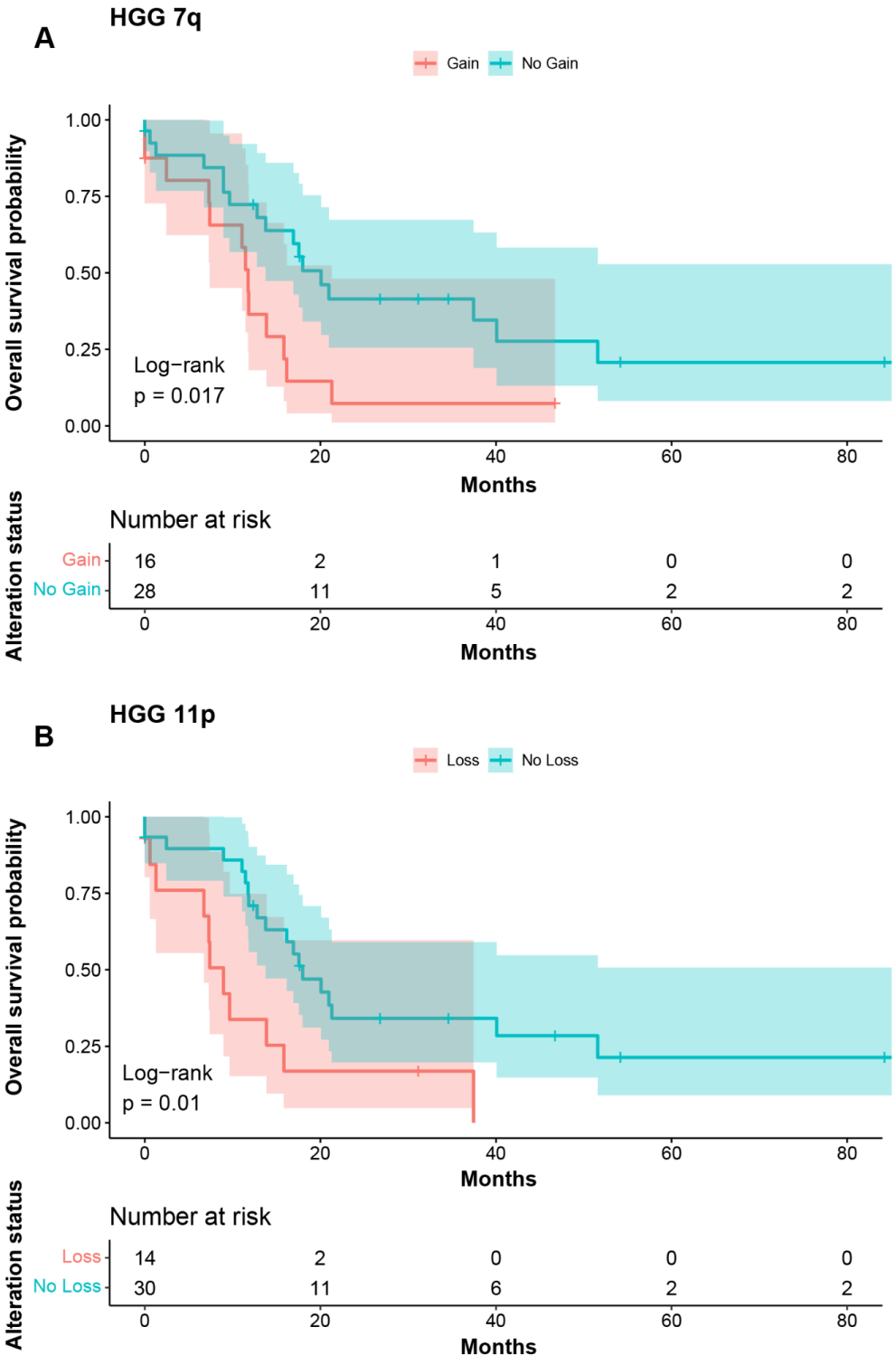


Figure 3.15: Kaplan-Meier survival curves split by alteration status for recurrently altered arms that were significantly associated with changes in survival in the High-Grade Glioma

group

(A) *Kaplan-Meier survival curves split by the gain status of 7q (top) and the corresponding risk table (bottom) in the High-Grade Glioma dataset within the CBTN cohort.* **(B)** *Kaplan-Meier survival curves split by the loss status of 11p (top) and the corresponding risk table (bottom) in the High-Grade Glioma dataset within the CBTN cohort. Curves are colored by their gain or loss status, respectively. Boxes around the curves indicate the 95% confidence intervals.*

In Medulloblastoma, both 1q and 8q gains have been previously reported to be detected in pediatric brain tumors (J. Y. Choi, 2023; Doussouki et al., 2019; Williamson et al., 2022). Moreover, gains of 1q and 8q have been found to be associated with worse overall survival (de Bont et al., 2008; De Bortoli et al., 2006; Lo et al., 2007). Likewise, our results suggested that both 1q and 8q gains are significantly correlated with worse overall survival (Figures 3.16A, 3.16C). 8p gains and their association with overall survival do not appear to have been previously reported in pediatric brain tumors. Our analysis seemed to suggest that 8p gains are also associated with worse overall survival (Figure 3.16B). However, further exploration of this result suggested it may have been confounded by the relationship between 8q gains and overall survival, as these 8p gains always co-occurred with 8q gains in our dataset. Ultimately, our analysis of recurrent alterations confirmed the presence of several previously reported alterations and their associations with changes in overall survival (i.e. 7q gains in High-Grade Glioma, 1q and 8q gains in Medulloblastoma). Furthermore, it suggested that 11p losses in High-Grade Glioma, which seem to be previously unreported in the context of pediatric brain tumors, may be relevant to pediatric cases.

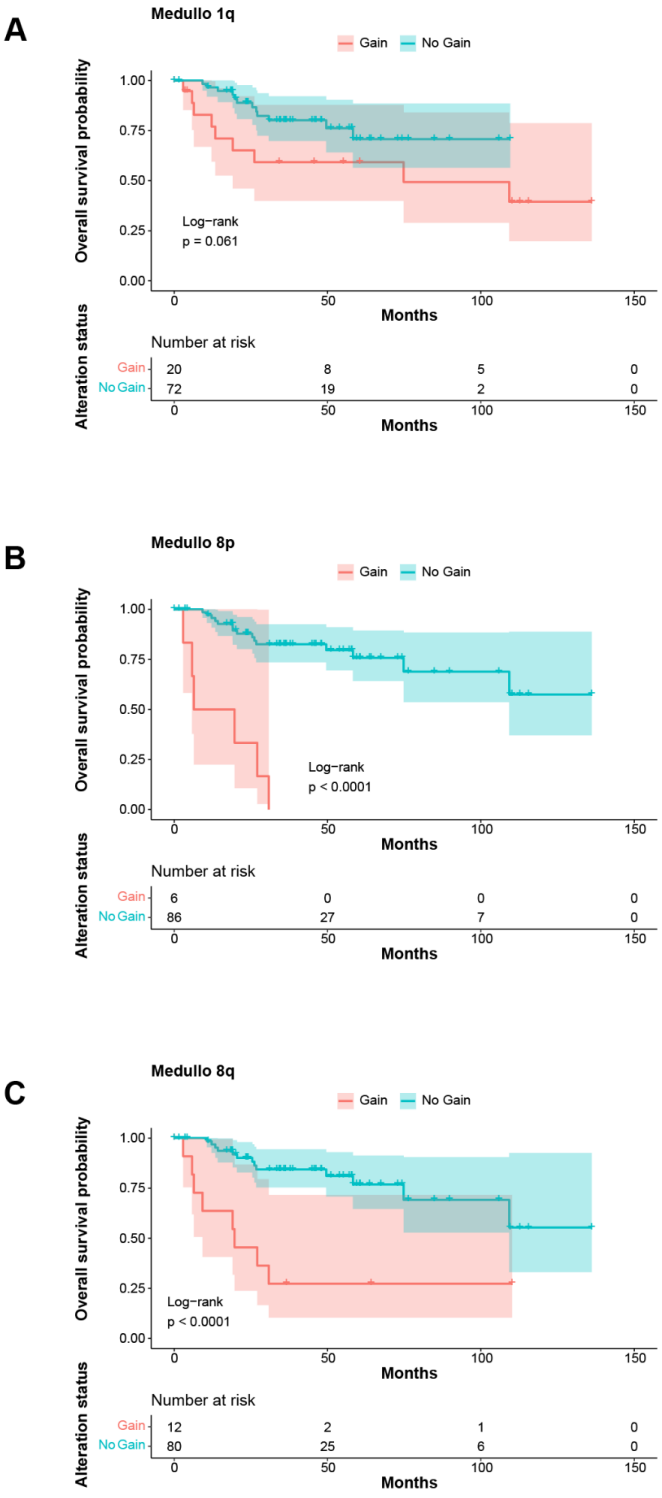


Figure 3.16: Kaplan-Meier survival curves split by alteration status for recurrently altered arms that were significantly associated with changes in survival in the Medulloblastoma

group

(A) *Kaplan-Meier survival curves split by the gain status of 1q (top) and the corresponding risk table (bottom) in the Medulloblastoma dataset within the CBTN cohort.* **(B)** *Kaplan-Meier survival curves split by the gain status of 8p (top) and the corresponding risk table (bottom) in the Medulloblastoma dataset within the CBTN cohort.* **(C)** *Kaplan-Meier survival curves split by the gain status of 8q (top) and the corresponding risk table (bottom) in the Medulloblastoma dataset within the CBTN cohort. Curves are colored by their gain status. Boxes around the curves indicate the 95% confidence intervals.*

In addition to testing recurrently altered arms identified within each diagnosis group, we also used multivariate Cox models (with variables for arm alteration status, cohort, diagnosis group, and sex) to test each recurrently altered region identified in the full dataset and assess whether any recurrent alterations may be significantly associated with changes in survival across diagnosis groups. This testing indicated that gains in three regions (2q, 4p, 11q) may be significantly associated with worse overall survival and losses in two regions (1p, 11q) may be significantly associated with better overall survival across the full dataset (Supplemental Table S3.3). However, the results also indicated that diagnosis group and cohort still had a strong effect on survival differences for all these regions. Furthermore, the variables for diagnosis group and cohort appeared to fail the proportional hazards assumption for the Cox model (Supplemental Figure S3.11, Supplemental Table S3.4). Thus, the associations reported between arm gains or losses and overall survival in the full dataset containing all diagnosis groups may not be reliable and would require additional validation.

3.5 Discussion

To investigate the landscape of copy number alteration in pediatric brain tumors and examine the relationship between CNVs and overall survival, we characterized copy number variations for more than 250 pediatric brain tumor samples across four diagnosis groups (ATRT, Ependymoma, High-Grade Glioma, and Medulloblastoma), calculated CNV burden using two metrics (CNV% and CNV segment count), identified recurrent copy number alterations, and analyzed their associations with overall survival. The results of this analysis suggested that both CNV burden and specific copy number alterations may have prognostic value within particular diagnosis groups.

For example, high CNV burden appeared to be associated with worse overall survival within the High-Grade Glioma group. Furthermore, gains of 7q and losses of 11p in High-Grade Glioma and gains of 1q and 8q in Medulloblastoma seemed to be associated with worse overall survival. While most of these observations have been previously reported in other pediatric brain tumor studies, the association between 11p losses and worse overall survival appeared to be previously unreported in pediatric High-Grade Glioma.

One caveat to the results presented here, particularly those reported within individual diagnosis groups, is that the sample size for our analysis was fairly low, with the largest diagnosis group (Medulloblastoma) having only 129 samples analyzed at most. Nevertheless, the ability to detect known events and associations (as well as seemingly unreported events and associations) within our dataset indicates the approaches utilized for this analysis can be expanded to larger cohorts and may lead to additional detection of previously unknown events and associations, which in turn could be used to inform the diagnosis and treatment of pediatric brain tumors.

Moreover, the common occurrence of CNVs also raises several questions about the role of CNVs in pediatric brain tumors. Namely, with regard to the role copy alteration events play in the initiation and progression of pediatric cancers and how alteration events might be related to treatment responses. For example, can copy number alterations be a driving mechanism of pediatric brain tumor development? Are these copy alteration events typically clonal or subclonal? And how might clonality of copy alterations impact susceptibility or response to treatment? The observations presented here, and the subsequent questions raised, provide an avenue for further exploration of the full impact of copy number alterations in pediatric brain tumors.

In addition, several other topics for exploration include analysis of CNVs in the context of molecular subtypes within diagnosis groups, exploration of potential driver genes and mutations located in regions of recurrent alteration, looking at the presence of co-occurring or mutually exclusive copy alteration events and their correlation with overall survival, examining the relationship between specific gene mutations, tumor mutation burden, and copy number alteration, and exploring patterns of LOH (particularly copy neutral LOH) and their relationship with survival.

3.6 Methods

3.6.1 Selection of pediatric brain tumor samples

For both the Washington University (WashU) and Children's Brain Tumor Network (CBTN) cohorts, tumor samples obtained from patients with ATRT, Ependymoma, High-Grade Glioma, or Medulloblastoma diagnoses were included in the analysis. The High-Grade Glioma group included patients with a diagnosis of Glioblastoma, Anaplastic Astrocytoma, or DIPG. For the

WashU cohort, tumor samples were obtained from a bank of diagnostic pediatric brain tumor samples from patients initially diagnosed and treated at St. Louis Children's Hospital. Permission to access CBTN data as a satellite consortia member was obtained from the Children's Brain Tumor Network (formerly the Children's Brain Tumor Tissue Consortium). Samples for inclusion in the CBTN cohort used for this analysis were identified using the Kids First Data Resource Portal. Only patients with whole genome sequencing (WGS) of initial tumor samples available were included. Aligned WGS reads for tumor and matched normal samples were downloaded in cram format from the Seven Bridges Genomics Cavatica platform. To access and download alignment files from Cavatica, membership access was obtained for the PBTA-CBTN dataset. Once downloaded, cram files were converted to bam files using samtools view -b.

3.6.2 Whole genome sequencing and alignment

Whole genome sequencing (WGS) libraries for the tumor samples from the WashU cohort were constructed using Automated Kapa Hyper PCR Free preparation kits for frozen samples (N=59) and Swift Accel-NGS 2S DNA Library preparation kits for FFPE samples (N=8). These libraries were sequenced on the Illumina NovaSeq 6000 to a target depth of 5x (i.e. low pass WGS). This sequencing was performed by the McDonnell Genome Institute (MGI) at Washington University School of Medicine in St. Louis. Details of library construction and sequencing methods utilized by MGI have been previously reported (Griffith, Miller, et al., 2015). Samples used to generate the panel of normals for the WashU cohort were isolated from matched normal adjacent breast tissue from 48 breast cancer patients. WGS libraries were constructed using Automated TruSeq Nano preparation kits and sequenced on the HiSeqX to a target depth of 30x. For the CBTN samples, tumor samples were sequenced to a target depth of 60x, while matched normal samples

were sequenced to a target depth of either 30x or 60x. Normal samples for each patient in the CBTN cohort consisted of peripheral whole blood samples for all samples except two, which had saliva samples that served as the normal. Full details of CBTN sequencing methods have been previously reported (Shapiro et al., 2023). For all samples from both cohorts, WGS reads were aligned to the GRCh38 reference genome using BWA-MEM (H. Li, 2013).

3.6.3 CNV calling, LOH calling, and manual correction

Copy number variant calling was performed using the CNVkit (v0.9.8) batch pipeline with the WGS method and a target bin size of 100,000 bp (Talevich et al., 2016). CNVkit's segment command was then run with the CNR files generated from this initial copy number variant calling to produce updated CNS files with outliers and low coverage segments dropped (parameters: -t 1e-31 --drop-low-coverage --drop-outliers 10). The filtered CNS files and the original CNR files were used to generate CNV plots across the genome for each sample. LOH calling was performed using a subset of SNP positions reported in GNOMAD (v3.0) (S. Chen et al., 2022). Bam-readcount was run to get reference and alternate allele read counts to calculate b-allele frequencies (BAFs) at each of the selected SNP positions in each tumor sample (Khanna et al., 2021). LOH plots were generated for each sample using these BAFs. The *smooth.spline* function in R was used to generate smoothed lines connecting the calculated BAFs. The CNV and LOH plots were then used to perform manual correction and rescue any missed copy number variant calls. Correction metrics could include a recentering factor, an adjusted upper cutoff for calling gains, an adjusted lower cutoff for calling losses, and adjustment of X and Y copy number values based on the reported sex of a sample. Segments flanking reported gap regions from the UCSC Genome Browser gap track for GRCh38 were filtered as well. These gap types

included short arm, heterochromatin, telomere, contig, and scaffold gaps, as defined by UCSC (*Gap Track Settings*).

3.6.4 Calculation of CNV burden

Total genome-wide CNV burden was summarized by two metrics: CNV% and CNV segment count. CNV% was calculated by taking the total number of altered base pairs and dividing it by the approximate size of the human genome (3.2 billion base pairs). For each sample, the total number of altered base pairs was calculated by subtracting the start position from the end position for each segment remaining after filtering and correction to determine the size of each segment, then summing the sizes of all segments across the genome. For each sample, the CNV segment count was calculated by summing the total number of altered segments remaining after filtering and correction across the genome. CNV burden was categorized separately for each metric (i.e. CNV% and CNV segment count) using three different methods: averages, optimal cutpoints, and tertiles. Averages for CNV% and CNV segment count were calculated within each diagnosis group separately using the *mean* function in R. Optimal cutpoints for CNV% and CNV segment count were identified within each diagnosis group separately using the *surv_cutpoint* function (which is a wrapper for the *maxstat* function) within the survival package in R. Tertile cutoffs for CNV% and CNV segment count were calculated within each diagnosis group using the *quantiles* function in R with probabilities 1/3, 2/3, and 1.

3.6.5 Identification of recurrent alterations

Recurrently altered genomic regions were identified using GISTIC 2.0 with default parameters (Mermel et al., 2011). Recurrently altered regions with FDR q-values of 0.25 or less were

reported. Input segmentation files were generated for the full dataset and within each diagnosis group using CNVkit's export seg command. For testing recurrently altered genomic regions for correlation with changes in survival, focal regions reported by GISTIC were collapsed into single regions based on chromosome arms. For each arm, the gain/no gain and loss/no loss statuses were determined in each sample. The p arm for a given recurrently altered chromosome was classified as harboring a gain in a sample if any portion of a gain segment was detected more than 500kb upstream of the centromere region on the corresponding chromosome in that sample. The p arm for a given recurrently altered chromosome was classified as harboring a loss in a sample if any portion of a loss segment was detected more than 500kb upstream of the centromere region on the corresponding chromosome in that sample. The q arm for a given recurrently altered chromosome was classified as harboring a gain in a sample if any portion of a gain segment was detected more than 500kb downstream of the centromere region on the corresponding chromosome in that sample. The q arm for a given recurrently altered chromosome was classified as harboring a loss in a sample if any portion of a loss segment was detected more than 500kb downstream of the centromere region on the corresponding chromosome in that sample.

3.6.6 Statistical analyses

Spearman correlation values for comparing continuous variables were calculated using the *cor* function in R with complete observations. Kaplan-Meier curves were generated using the *survfit* function in R. The choice of statistical tests used to check for significant differences in Kaplan-Meier survival curves was determined based on crossing of the curves. For crossing curves, the Gehan-Breslow (aka the generalized Wilcoxon) method was used. For non-crossing curves, the

log-rank method was used. Cox models were run using the *coxph* function in R. Tests of the proportional hazards assumption were run using the *cox.zph* function in R. For testing recurrently altered arms within each diagnosis group, Cox models for gains and losses were run separately for each arm. The models included categorical variables for arm (gain/no gain or loss/no loss) and sex. Sex was not significantly associated with changes in survival in any of the models. Samples with no sex reported (N = 8 for the full dataset) were excluded. False discovery rate (FDR) multiple testing correction was done for both the Kaplan-Meier and the Cox model p-values reported within each diagnosis group for testing the relationship between survival and chromosome arm gains or losses.

3.7 Supplemental Tables

Diag.	Metric	Avg.	KM Method	KM Pval
Ependymoma	CNV%	12.55	Gehan-Breslow	0.41
Ependymoma	CNV segment count	18.48	Gehan-Breslow	0.39
HGG	CNV%	19.56	Gehan-Breslow	0.19
HGG	CNV segment count	36.14	Gehan-Breslow	0.059
Medullo	CNV%	18.52	Gehan-Breslow	0.59
Medullo	CNV segment count	23.31	Log-Rank	0.17

Table S3.1: Survival analysis results for categorizing CNV burden based on averages

Table of survival analysis results for categorizing CNV burden based on averages for CNV% and CNV segment count within each diagnosis group across both cohorts. Samples with CNV% or CNV segment count below the corresponding average were classified as "CNV Low". Samples with CNV% or CNV segment count above the corresponding average were classified as "CNV High". **KM Method** *indicates the statistical method used to test for significant differences in the Kaplan-Meier survival estimates.*

Diag.	Metric	Tertile Low	Tertile Mid	Tertile High	KM Method	KM Pval
Ependymoma	CNV%	0.758	11.373	71.295	Gehan-Breslow	0.56
Ependymoma	CNV segment count	1	15	162	Gehan-Breslow	0.19
HGG	CNV%	10.616	24.87	66.826	Gehan-Breslow	0.16
HGG	CNV segment count	9	43	237	Gehan-Breslow	0.2
Medullo	CNV%	6.967	19.297	66.04	Gehan-Breslow	0.86
Medullo	CNV segment count	6	22.667	126	Gehan-Breslow	0.38

Table S3.2: Survival analysis results for categorizing CNV burden based on tertiles

Table of survival analysis results for categorizing CNV burden based on tertiles for CNV% and CNV segment count within each diagnosis group across both cohorts. Samples with CNV% or CNV segment count below the corresponding low tertile cutoff value were classified as "CNV Low". Samples with CNV% or CNV segment count between the corresponding low and mid tertile cutoff values were classified as "CNV Mid". Samples with CNV% or CNV segment count above the corresponding middle tertile cutoff value were classified as "CNV High". **KM Method** *indicates the statistical method used to test for significant differences in the Kaplan-Meier survival estimates.*

Arm	Status	Arm Pval	Arm HR	ATRT Pval	ATRT HR	HGG Pval	HGG HR	Med Pval	Med HR	Cohort Pval	Cohort HR	Sex Pval	Sex HR
2q	Gain	0.006	2.01	3E-7	5.23	1E-8	4.76	0.065	1.62	2E-5	2.15	0.051	1.42
4p	Gain	0.037	1.77	2E-7	5.3	4E-9	4.95	0.072	1.60	2E-5	2.15	0.025	1.51
11q	Gain	0.026	2.05	2E-7	5.39	9E-10	5.26	0.027	1.78	3E-5	2.09	0.032	1.47
1p	Loss	0.007	0.43	8E-7	4.85	0	6.15	0.037	1.72	4E-6	2.33	0.096	1.35
11q	Loss	0.016	0.56	2E-6	4.57	1E-9	5.15	0.045	1.68	6E-5	2.04	0.114	1.34

Table S3.3: Survival analysis results for testing gain and loss statuses of recurrently altered chromosome arms

Table of Cox model results for recurrently altered arms that were significantly associated with changes in overall survival using Cox models with arm status (gain/no gain or loss/no loss), diagnosis group, cohort, and sex as variables in the full dataset containing all diagnosis groups across both cohorts.

Arm	Status	Diag Pval	Cohort Pval	Sex Pval	Arm Pval	Global Pval
2q	Gain	0.00386	0.00947	0.21076	0.12406	0.00022
4p	Gain	0.00384	0.00737	0.34976	0.12168	0.0003
11q	Gain	0.00243	0.00937	0.62483	0.13525	0.00035
1p	Loss	0.0021	0.02848	0.60712	0.17884	0.00111
11q	Loss	0.00228	0.01183	0.13428	0.14814	0.00024

Table S3.4: Results for checking the proportional hazards assumption for testing gain and loss statuses of recurrently altered chromosome arms

Table of results for checking the proportional hazards assumption of the Cox model (using the cox.zph function in R) for diagnosis group, cohort, sex, arm alteration status, and the model as a whole (global) when testing gain and loss statuses of recurrently altered chromosome arms in the full dataset containing all diagnosis groups across both cohorts. P-values less than 0.05 indicate that the corresponding variable fails the proportional hazards assumption of the Cox model, meaning that the hazard functions of the given variable are not proportional over time.

3.8 Supplemental Figures

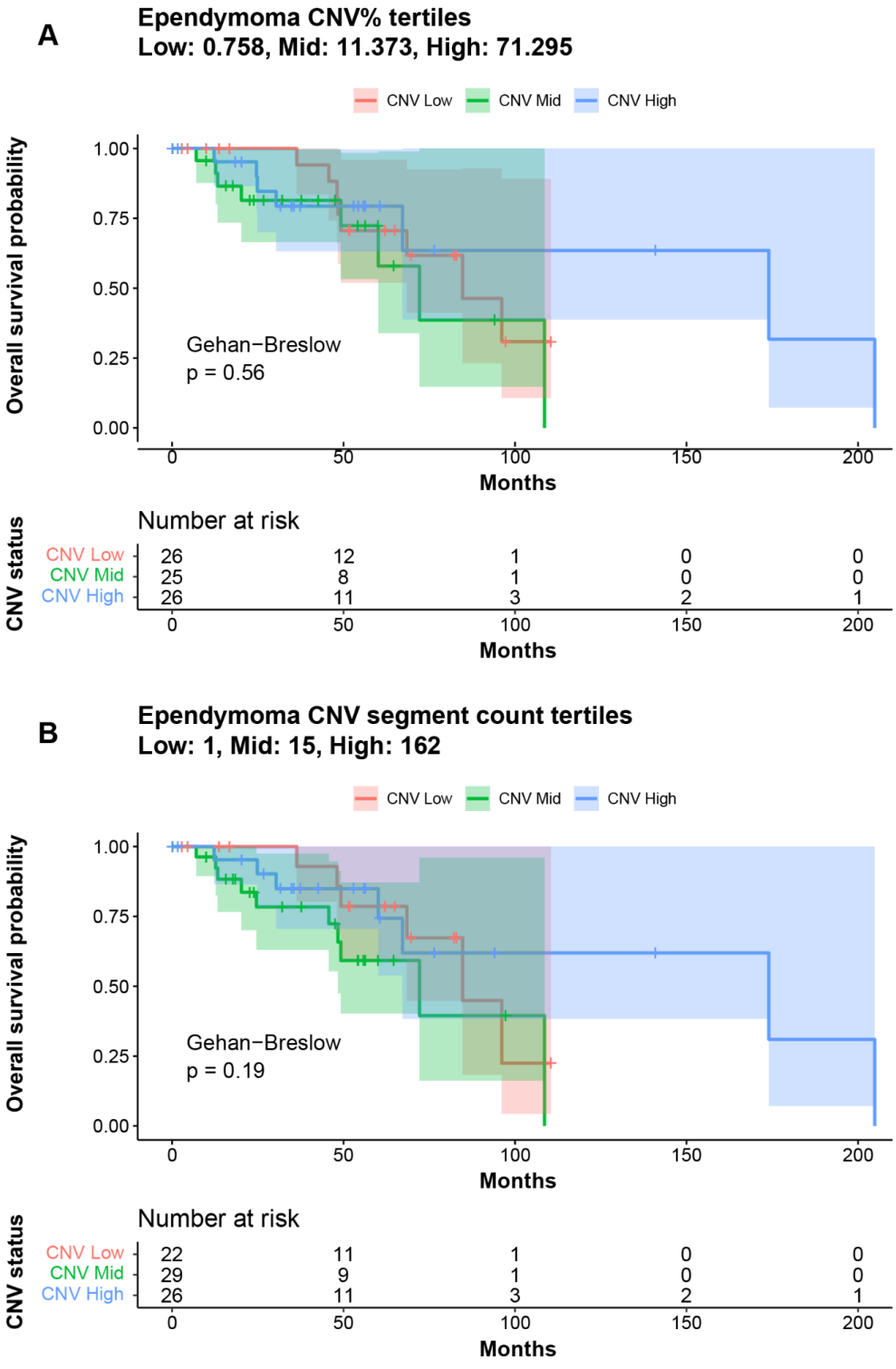


Figure S3.1: Kaplan-Meier survival curves split by CNV burden based on tertiles for the Ependymoma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on CNV% tertiles (top) and the corresponding risk table (bottom) in the Ependymoma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on CNV segment count tertiles (top) and the corresponding risk table (bottom) in the Ependymoma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

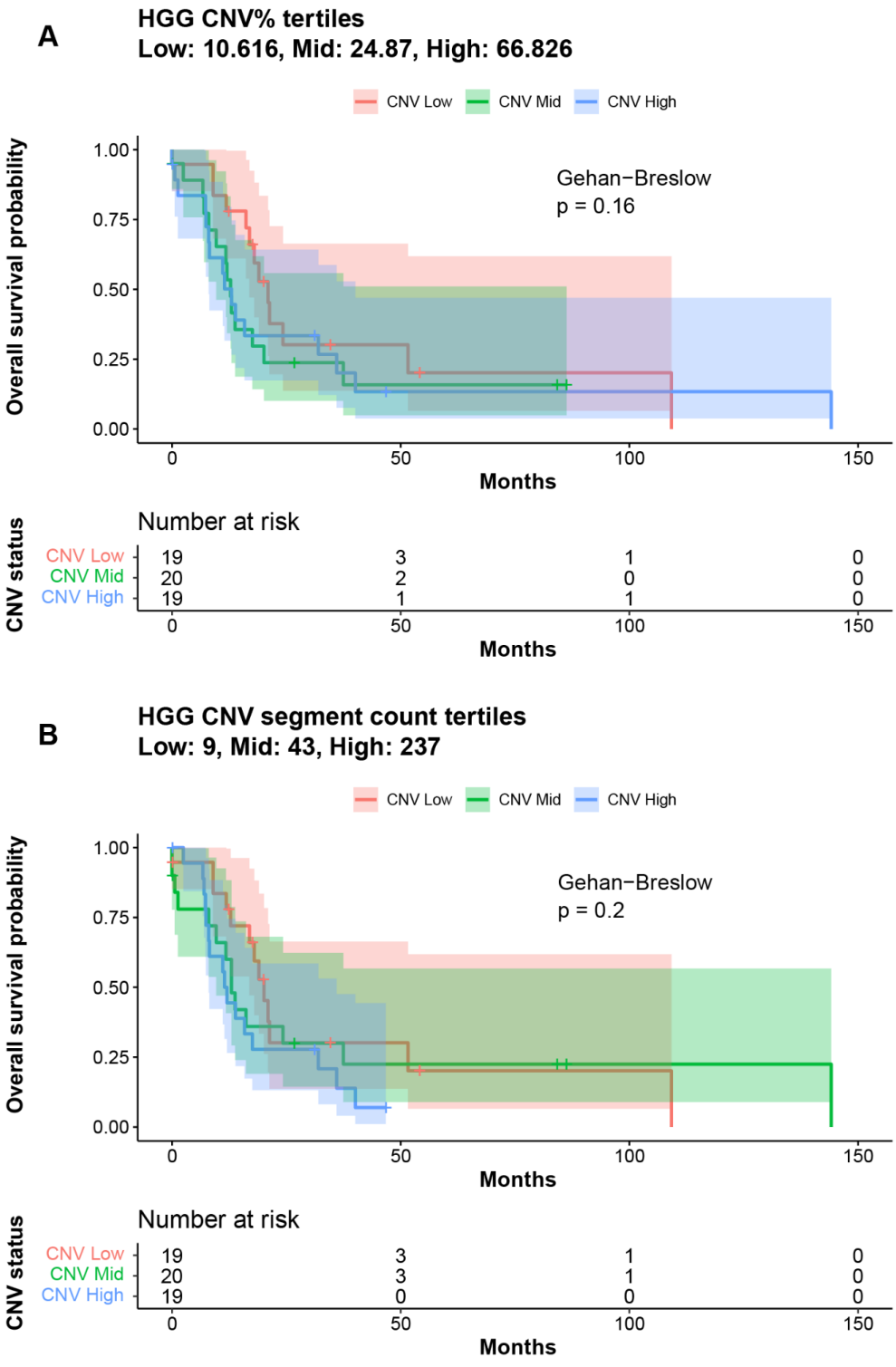


Figure S3.2: Kaplan-Meier survival curves split by CNV burden based on tertiles for the High-Grade Glioma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on CNV% tertiles (top) and the corresponding risk table (bottom) in the High-Grade Glioma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on CNV segment count tertiles (top) and the corresponding risk table (bottom) in the High-Grade Glioma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

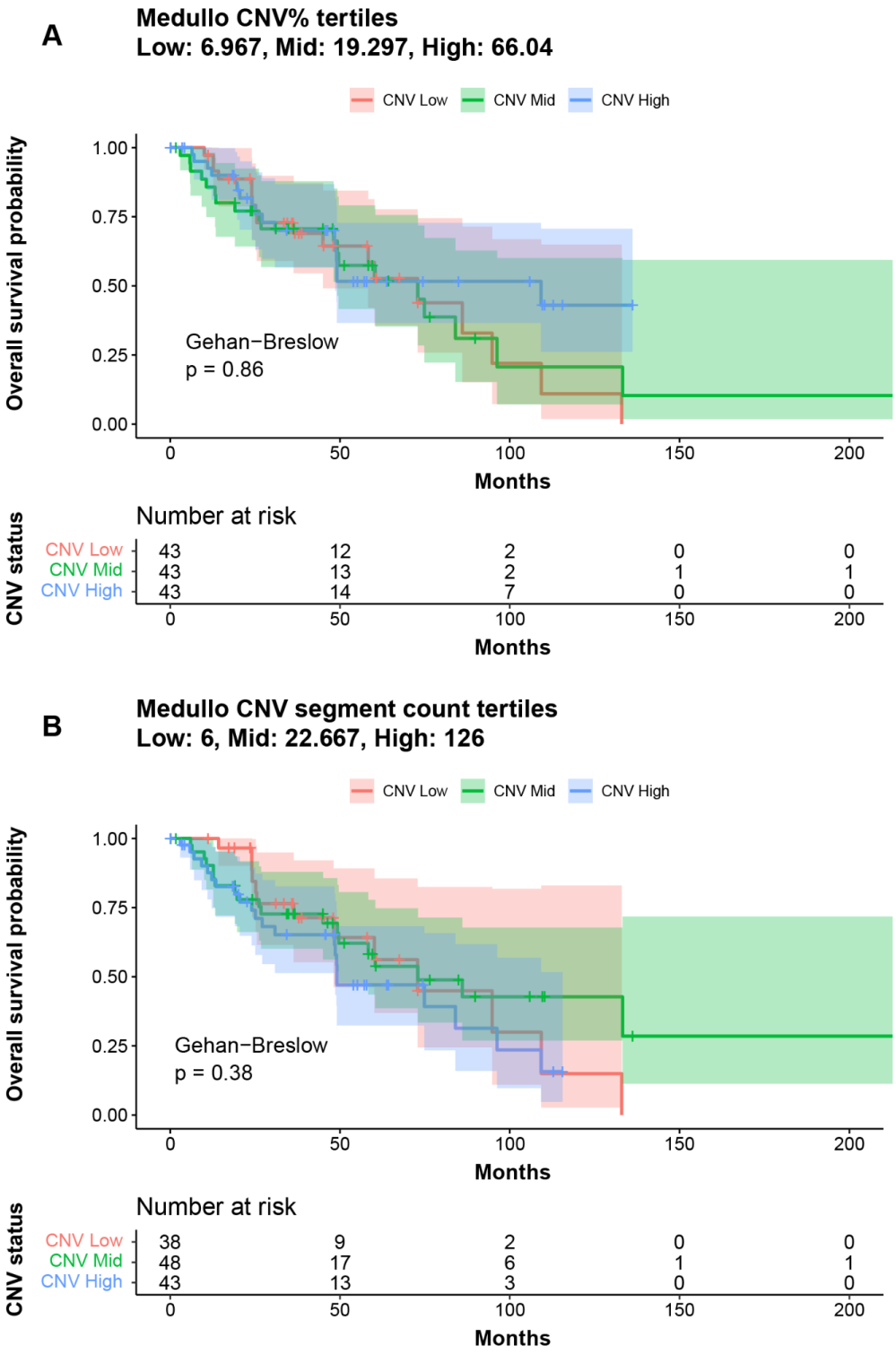


Figure S3.3: Kaplan-Meier survival curves split by CNV burden based on tertiles for the Medulloblastoma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on CNV% tertiles (top) and the corresponding risk table (bottom) in the Medulloblastoma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on CNV segment count tertiles (top) and the corresponding risk table (bottom) in the Medulloblastoma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

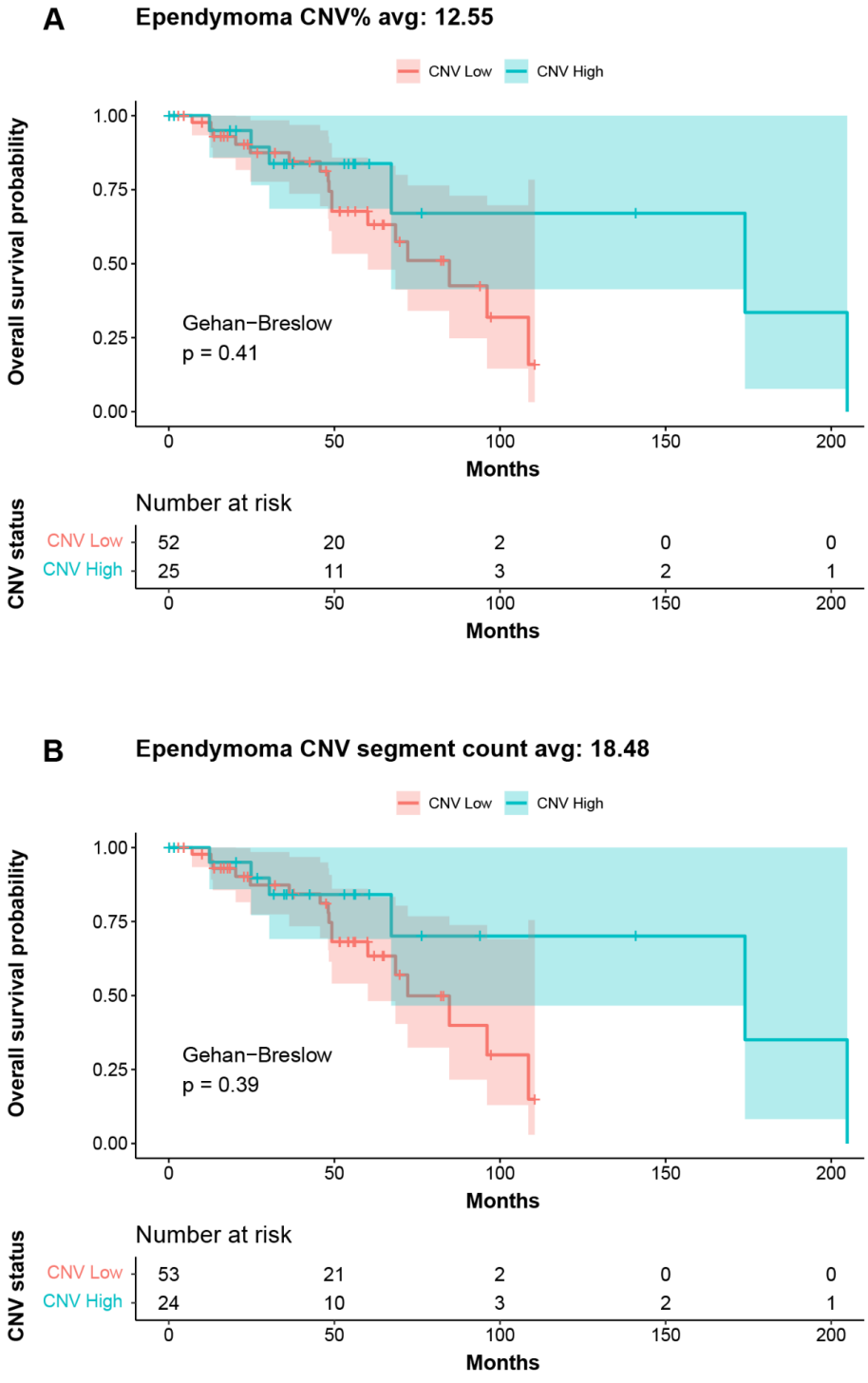


Figure S3.4: Kaplan-Meier survival curves split by CNV burden based on averages for the Ependymoma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on CNV% average (top) and the corresponding risk table (bottom) in the Ependymoma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on CNV segment count average (top) and the corresponding risk table (bottom) in the Ependymoma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

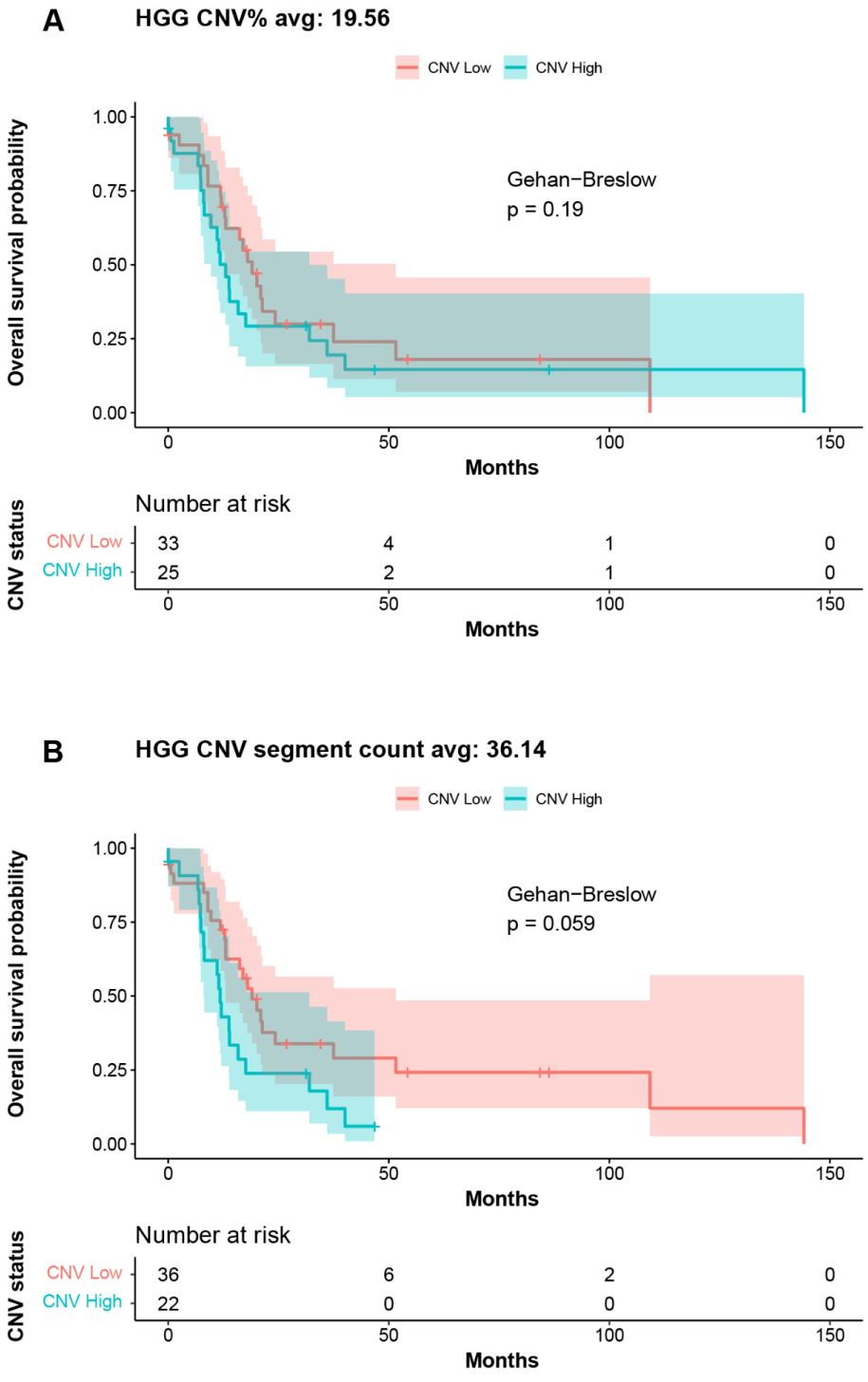


Figure S3.5: Kaplan-Meier survival curves split by CNV burden based on averages for the High-Grade Glioma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on CNV% average (top) and the corresponding risk table (bottom) in the High-Grade Glioma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on CNV segment count average (top) and the corresponding risk table (bottom) in the High-Grade Glioma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

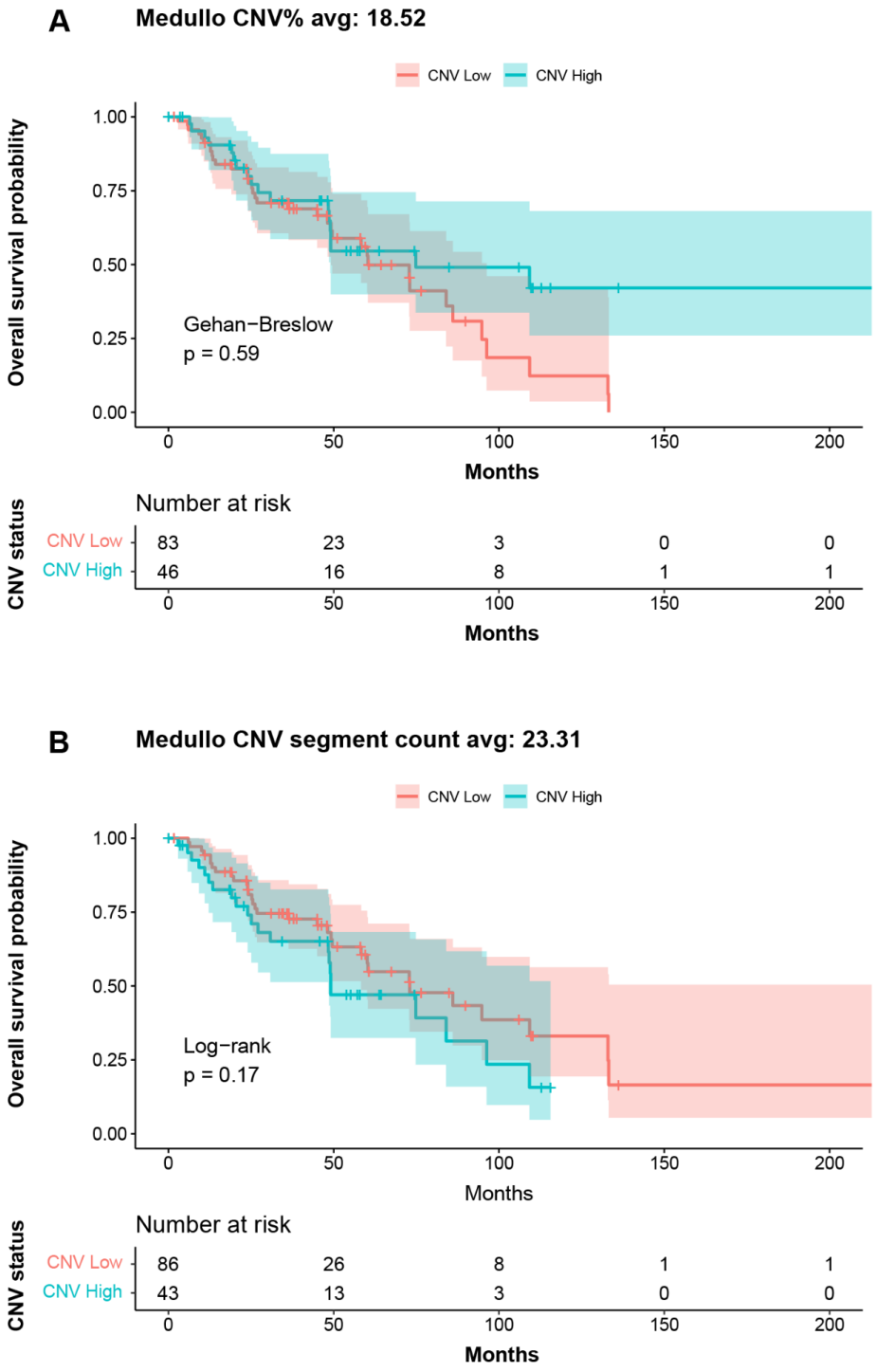


Figure S3.6: Kaplan-Meier survival curves split by CNV burden based on averages for the Medulloblastoma group

(A) *Kaplan-Meier survival curves split by CNV burden classification based on CNV% average (top) and the corresponding risk table (bottom) in the Medulloblastoma group across both cohorts.* **(B)** *Kaplan-Meier survival curves split by CNV burden classification based on CNV segment count average (top) and the corresponding risk table (bottom) in the Medulloblastoma group across both cohorts. Curves are colored by their CNV burden group. Boxes around the curves indicate the 95% confidence intervals.*

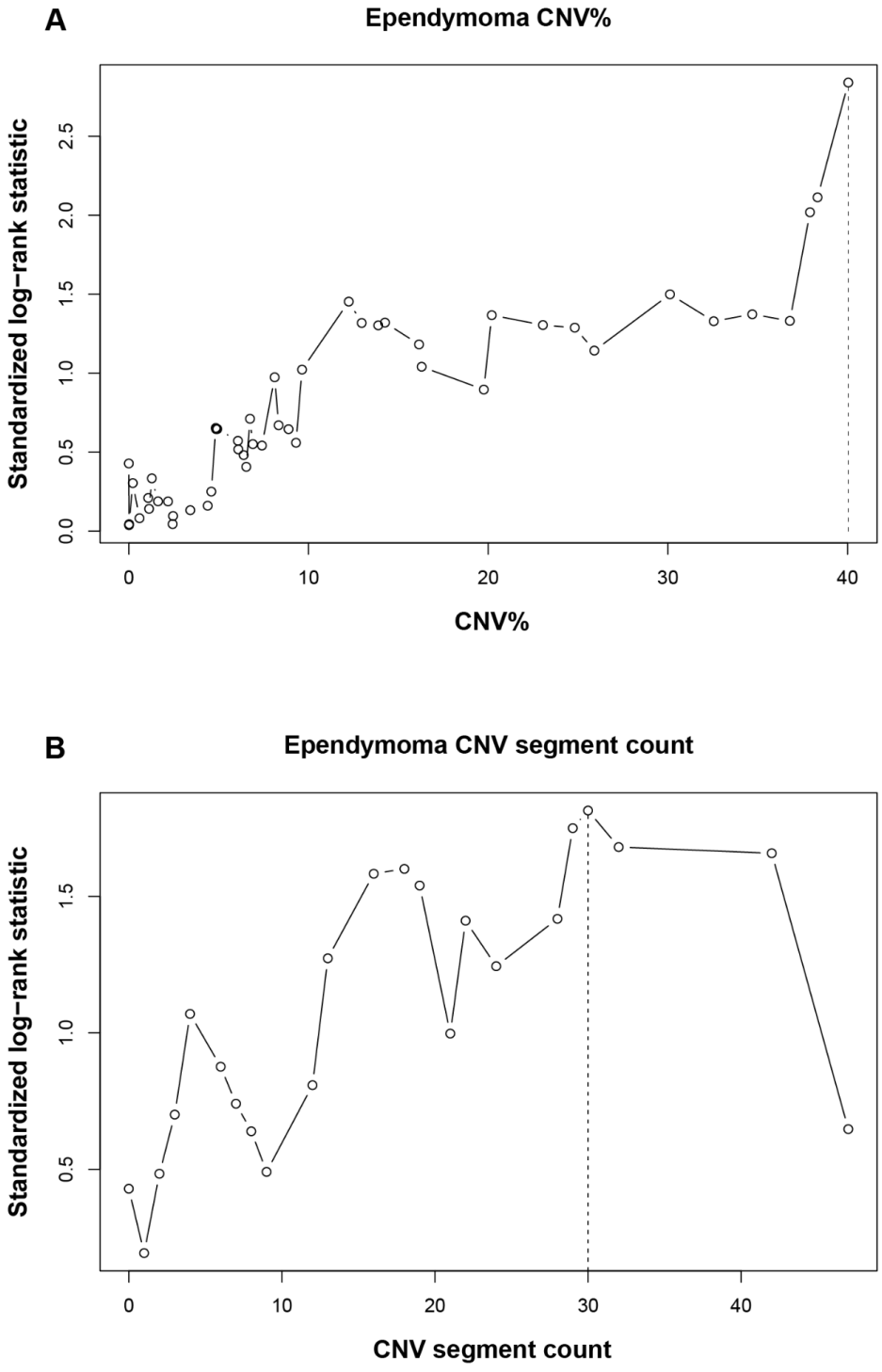


Figure S3.7: Log-rank test statistics used to determine optimal cutpoints for CNV% and CNV segment count in the Ependymoma group

(A) *Scatter plot of CNV% versus standardized log-rank test statistic values generated for testing potential optimal cutpoints, using the maxstat function in R, for the Ependymoma group across both cohorts.* **(B)** *Scatter plot of CNV segment count versus standardized log-rank test statistic values generated for testing potential optimal cutpoints, using the maxstat function in R, for the Ependymoma group across both cohorts.*

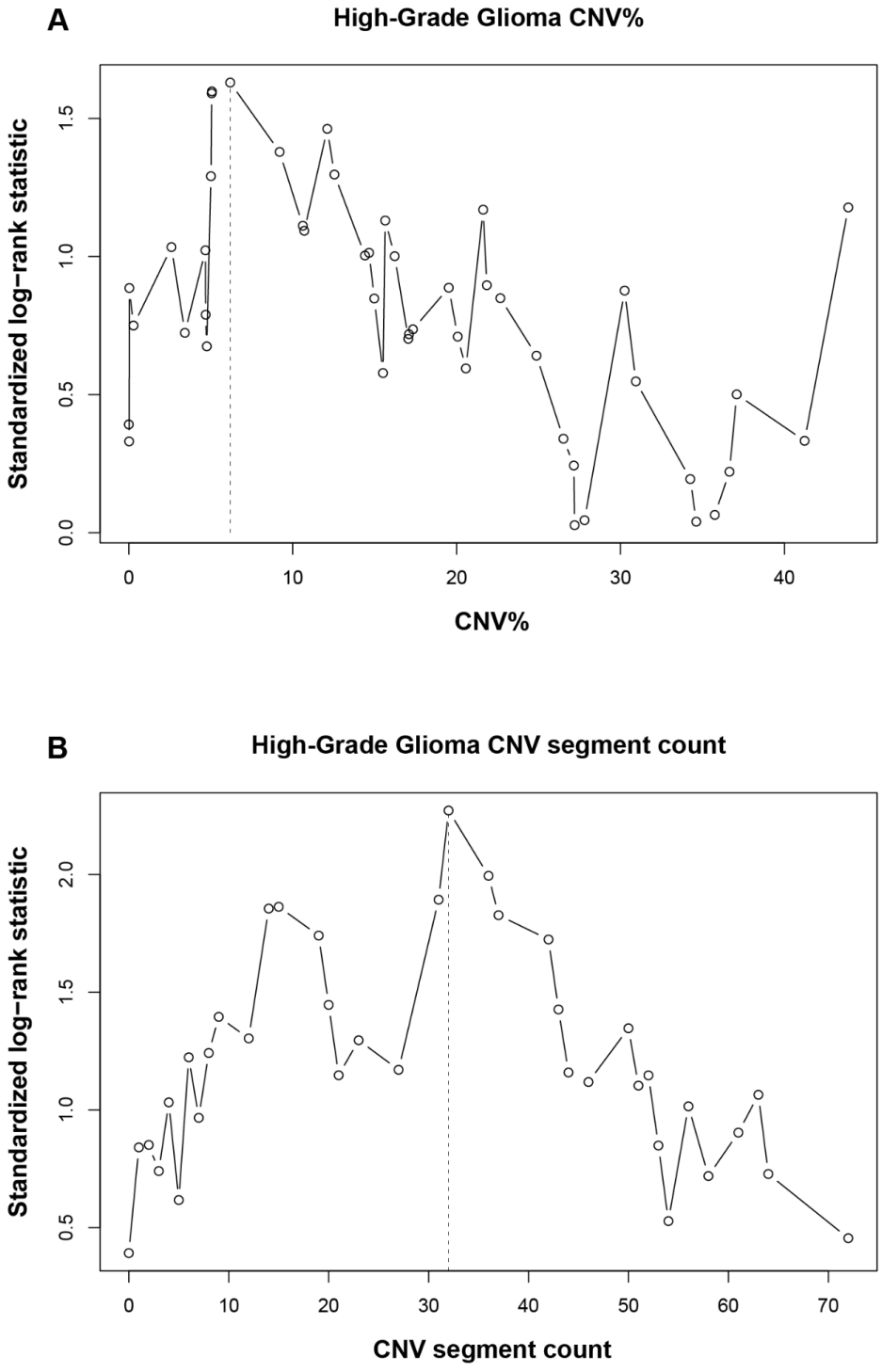


Figure S3.8: Log-rank test statistics used to determine optimal cutpoints for CNV% and CNV segment count in the High-Grade Glioma group

(A) *Scatter plot of CNV% versus standardized log-rank test statistic values generated for testing potential optimal cutpoints, using the maxstat function in R, for the High-Grade Glioma group across both cohorts.* **(B)** *Scatter plot of CNV segment count versus standardized log-rank test statistic values generated for testing potential optimal cutpoints, using the maxstat function in R, for the High-Grade Glioma group across both cohorts.*

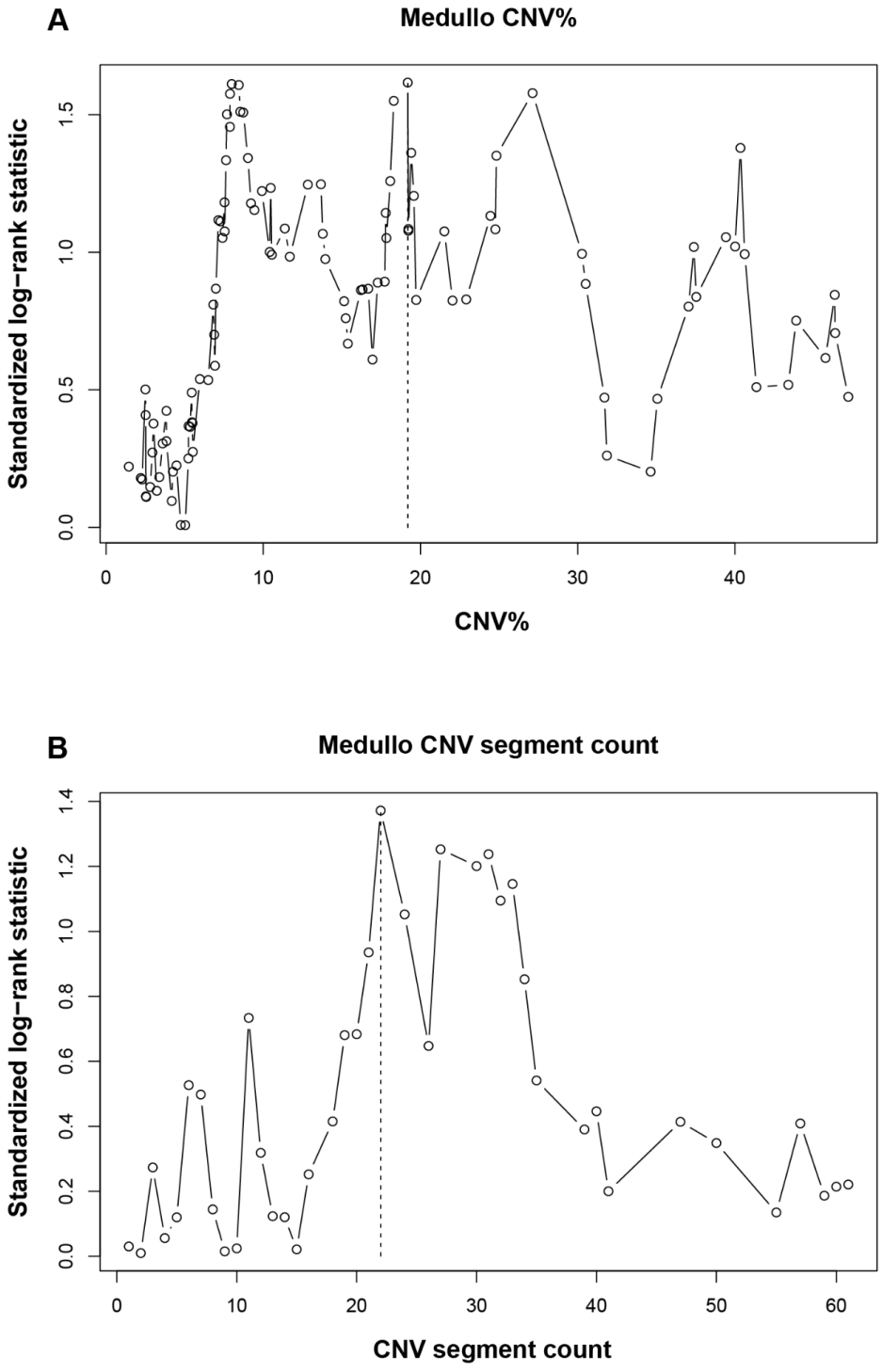


Figure S3.9: Log-rank test statistics used to determine optimal cutpoints for CNV% and CNV segment count in the Medulloblastoma group

(A) *Scatter plot of CNV% versus standardized log-rank test statistic values generated for testing potential optimal cutpoints, using the maxstat function in R, for the Medulloblastoma group across both cohorts.* **(B)** *Scatter plot of CNV segment count versus standardized log-rank test statistic values generated for testing potential optimal cutpoints, using the maxstat function in R, for the Medulloblastoma group across both cohorts.*

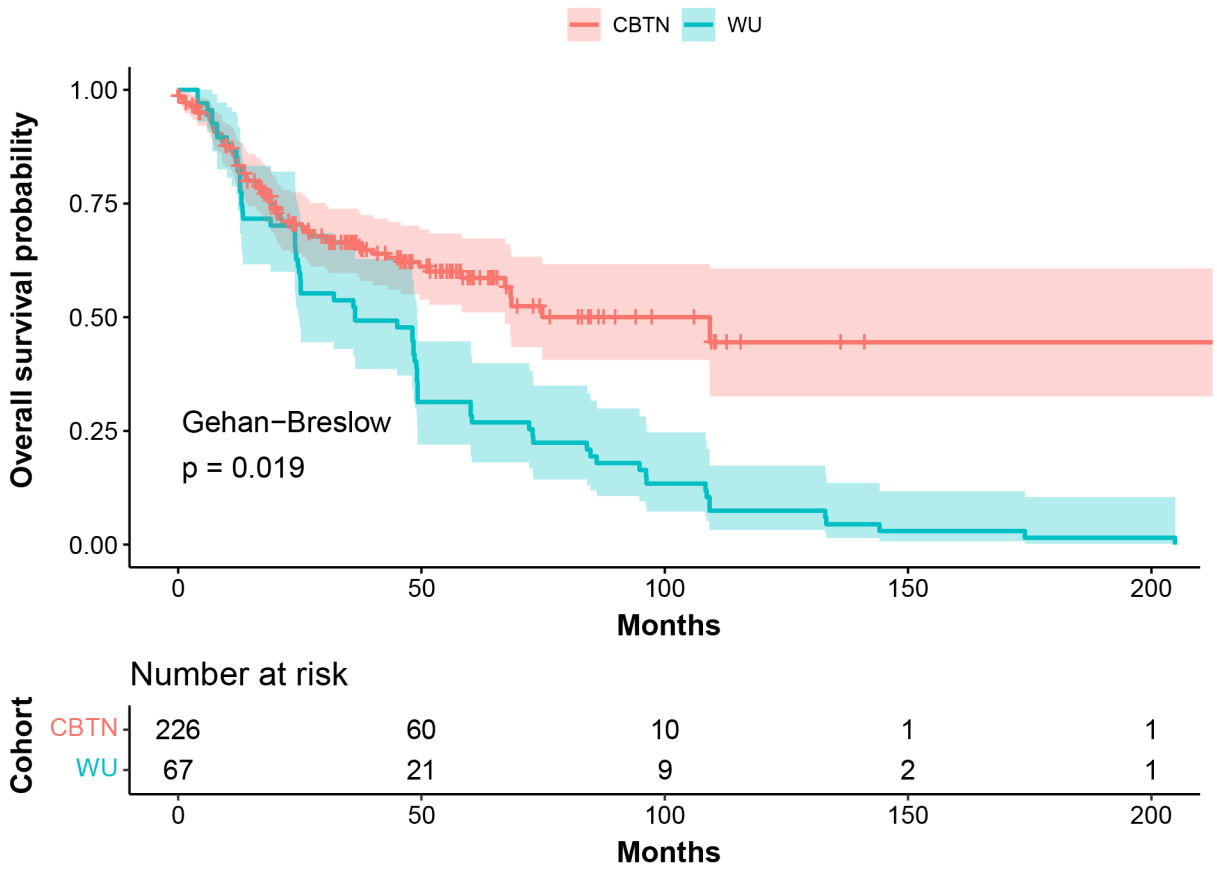


Figure S3.10: Kaplan-Meier survival curves split by cohort

Kaplan-Meier survival curves split by cohort (top) and the corresponding risk table (bottom). Curves are colored by the cohort. Boxes around the curves indicate the 95% confidence intervals.

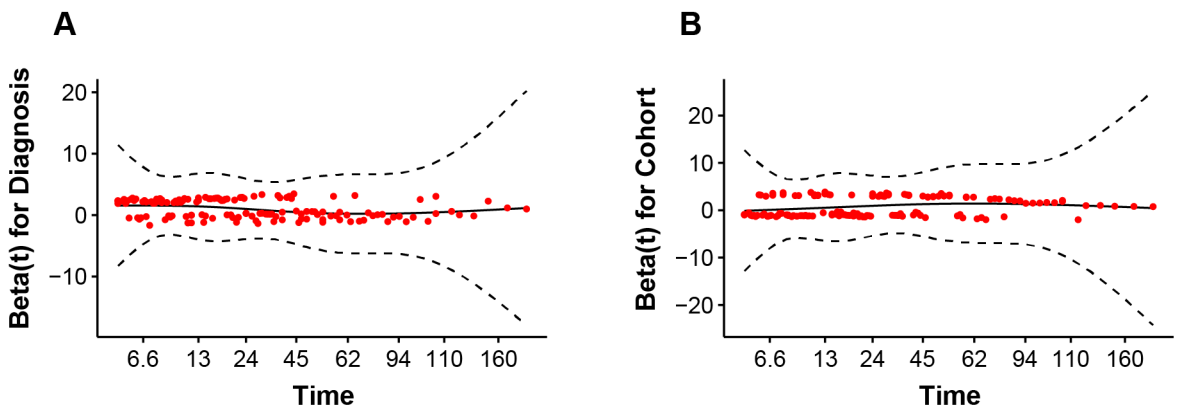


Figure S3.11: Schoenfeld residual plots for cohort and diagnosis variables

(A) *Plot of scaled Schoenfeld residuals for the diagnosis group variable for the Cox model testing arm alteration status in the full dataset containing all diagnosis groups across both cohorts.* **(B)** *Plot of scaled Schoenfeld residuals for the cohort variable for the Cox model testing arm alteration status in the full dataset containing all diagnosis groups across both cohorts. The red points show the residual values. The solid line is a smoothing spline fitted to the residual values. The dotted lines represent the confidence interval of 2 standard deviations. A plot of the Schoenfeld residuals for a variable that fails the proportional hazards assumption of the Cox model will show patterns that vary over time (as opposed to a flat horizontal line that does not vary over time).*

Chapter 4: Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts

4.1 Preamble

The following chapter has been published as a peer reviewed manuscript with the following citation:

Freshour SL*, Kiwala S*, Cotto KC*, Coffman AC, McMichael JM, Song JJ, Griffith M, Griffith OL, Wagner AH. Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. 2021. Nucleic Acids Research. DOI: 10.1093/nar/gkaa1084

* denotes co-first authors

As an author of the published manuscript, and in compliance with the editorial policies at Nucleic Acids Research, the cited publication is included in full in the following chapter. As a co-first author of this manuscript, I assisted with identifying new drug and gene data sources to add and existing sources to update in DGIdb 4.0, contributed to the development of new features to support the concept of interaction directionality, contributed to figure generation, drafting the manuscript, and addressing reviewer comments for resubmission and publication.

4.2 Summary

The Drug-Gene Interaction Database (DGIdb, www.dgidb.org) is a web resource that provides information on drug-gene interactions and druggable genes from publications, databases, and other web-based sources. Drug, gene, and interaction data are normalized and merged into conceptual groups. The information contained in this resource is available to users through a straightforward search interface, an application programming interface (API), and TSV data downloads. DGIdb 4.0 is the latest major version release of this database. A primary focus of this update was integration with crowdsourced efforts, leveraging the Drug Target Commons for community-contributed interaction data, Wikidata to facilitate term normalization, and export to NDEx for drug-gene interaction network representations. Seven new sources have been added since the last major version release, bringing the total number of sources included to 41. Of the previously aggregated sources, 15 have been updated. DGIdb 4.0 also includes improvements to the process of drug normalization and grouping of imported sources. Other notable updates include the introduction of a more sophisticated Query Score for interaction search results, an updated Interaction Score, the inclusion of interaction directionality, and several additional improvements to search features, data releases, licensing documentation and the application framework.

4.3 Introduction

Originally released in 2013, the Drug–Gene Interaction database (DGIdb) serves as a central aggregator of information on drug-gene interactions and druggability from multiple diverse sources (Griffith et al., 2013). The subsequent major updates to DGIdb 2.0 (in 2016) and 3.0 (in

2018) included improvements to the user interface and search response times, the addition of an API, the introduction and improvement of gene and drug grouping methods, and the expansion of source content through the inclusion of new sources and updates of existing sources (Cotto et al., 2018; Wagner et al., 2016). Since the release of DGIdb 3.0, many of the existing sources have been substantially updated and new sources have become available. Here we describe changes made for our most recent major version release, DGIdb 4.0. In this release, we have made an effort to integrate crowdsourced data and sources in several areas, including the addition of the crowdsourced Drug Target Commons as a drug-gene interaction source, and the use of the open, community-curated Wikidata resource for drug normalization (Tanoli et al., 2018; Vrandečić & Krötzsch, 2014). We also illustrate the value of our integration efforts in downstream community tools, through the incorporation of our data into NDEx (Pratt et al., 2017). To keep content offered by DGIdb current, we have developed additional automatic update routines for multiple sources and implemented a new background job management system (Sidekiq, sidekiq.org) for routine job scheduling. Finally, DGIdb 4.0 focuses on numerous improvements of search results, including new and updated scores for interaction search results and improved drug normalization routines.

4.4 Results

4.4.1 Integration with crowdsourced efforts

A primary focus of the DGIdb 4.0 release is the inclusion and utilization of crowdsourced efforts in several aspects of our database. The utility of our database begins with importing relevant drug, gene, and drug-gene interaction records (called claims) from outside resources. We normalize and sort these claims into conceptual groups, and make these concepts searchable via

a web application and API. We also export data for bulk download and use with external resources (Figure 4.1). In this update, we extend these features by integration with crowdsourced drug-gene interaction claims, normalizing drug terms, and integrating with external resources.

For drug–gene interaction claims, we have added Drug Target Commons as a new source in DGIdb 4.0 (Figure 4.1). Drug Target Commons provides an extensive curated database of crowdsourced drug-gene interactions, from which we added a total of 23,879 interaction claims. This represents ~24% (23,879/100,273) of the total interaction claims in DGIdb.

For drug normalization, we now use a Wikidata normalizer in addition to a ChEMBL normalizer from the *thera-py* python package (Figure 4.1; additional detail in 'Drug grouping improvements' section) (Mendez et al., 2019). Wikidata serves as a source of collaborative, crowdsourced drug concepts, and has allowed us to improve normalization in cases where ChEMBL normalization failed. For example, concepts representing the terms *annamycin*, *N-methyl scopolamine* and *Debio 1347* are all found in Wikidata but not ChEMBL.

Finally, we have integrated DGIdb with the Network Data Exchange (NDEx), a community resource that allows sharing and publishing of biological data in a network-based format (Pratt et al., 2017). For DGIdb, export of DGIdb data to the NDEx platform provides a resource for the visual representation of relationships and interactions between drugs and genes present in our database, allowing users to visually explore a global network of drugs and gene interactions of interest. NDEx TSVs are generated monthly and automatically uploaded to the NDEx server to keep the DGIdb network in NDEx up-to-date (Figure 4.1). NDex is the latest in a number of community resources that have integrated DGIdb. Existing data clients include GeneCards, BioGPS, CancerTracer, Gene4Denovo, SL-BioDP, TargetDB and OncoGemini,

among others (De Cesco et al., 2020; Deng et al., 2019; Nicholas et al., 2020; Stelzer et al., 2016; C. Wang et al., 2020; Wu et al., 2016; Zhao et al., 2020).

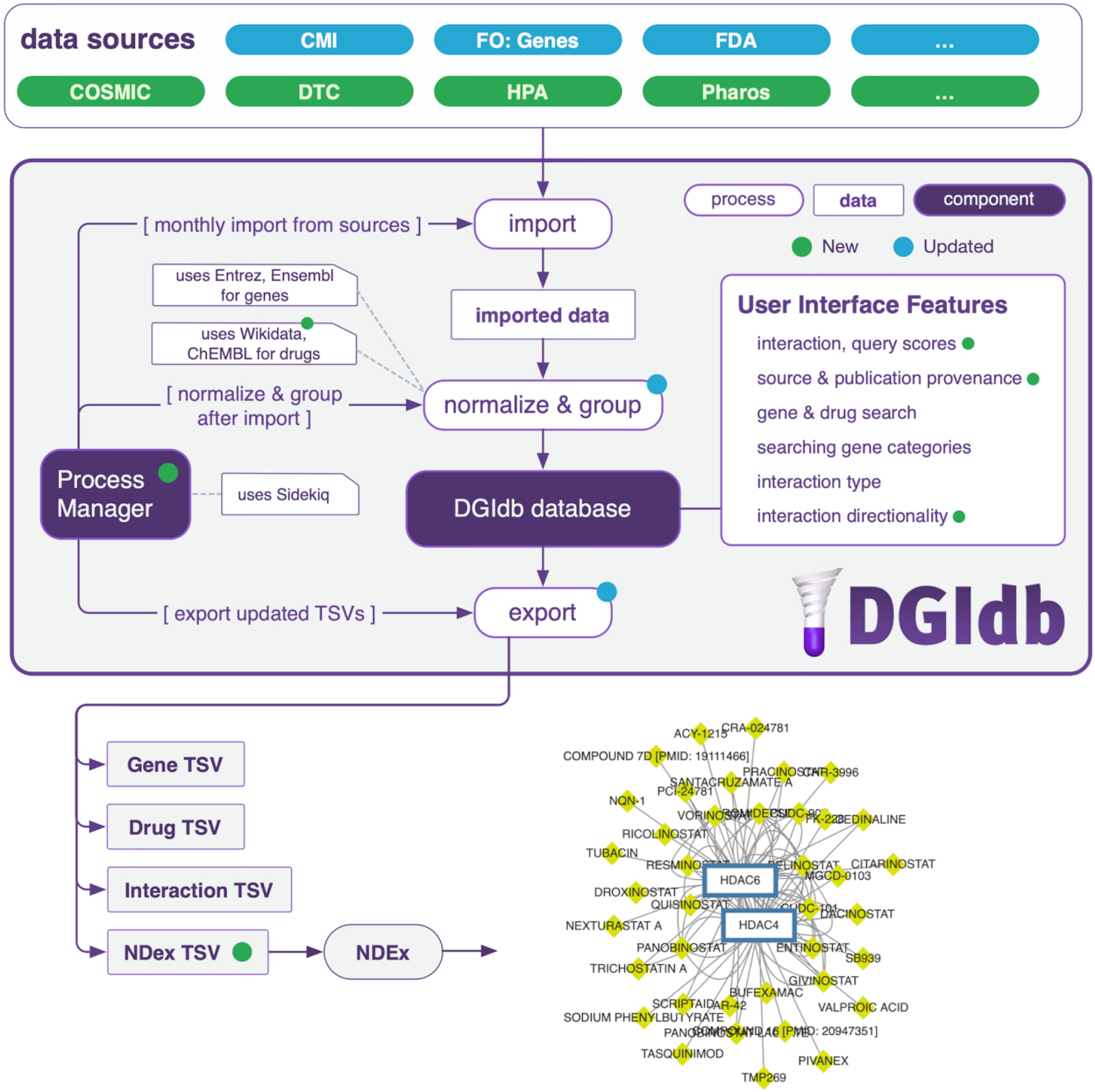


Figure 4.1: Overview of main components of DGIdb

Data sources are imported from outside resources (over 40 as of DGIdb 4.0), normalized and grouped with internal processes to prepare records to be displayed in DGIdb, and exported to TSV for download and integration with other resources. Process management is handled by

Sidekiq for automation of importing, normalization and grouping, and exporting. A subset of new data sources are highlighted in green, a subset of updated pre-existing data sources are highlighted in blue. The updated sources highlighted in this figure are some of the sources that have been updated through manual curation. Information on additional sources and their status in DGIdb 4.0 can be found in Figure 4.2 and Supplementary Table S1. New features and technologies from DGIdb 4.0 are indicated with green dots, pre-existing features and technologies that have been updated are indicated with blue dots. The drug-gene network graph shown in the bottom right is an example of the data visualizations available on NDEx. Abbreviations: CMI = Caris Molecular Intelligence, FO = Foundation One, DTC = Drug Target Commons and HPA = Human Protein Atlas.

4.4.2 New and updated sources

In an effort to ensure that DGIdb offers diverse and contemporary information, we have updated and added several sources to DGIdb 4.0. In addition to the previously mentioned Drug Target Commons, we now also include COSMIC as a new source of drug-gene interaction data (Supplementary Table S1) (Tanoli et al., 2018; Tate et al., 2019). COSMIC also serves as an additional source of curated *Drug Resistance* gene category claims. Other new gene category sources include the Tempus xT panel of actionable cancer therapy target genes, a list of the top priority genes from the Illuminating the Druggable Genome (IDG) Initiative, the Human Protein Atlas, the Oncomine clinical cancer biomarker assay, and understudied targets of the IDG program from Pharos (Supplementary Table S1) (Beaubier et al., 2019; Nguyen et al., 2017; Rodgers et al., 2018; Uhlen et al., 2017; Williams et al., 2018). From these new sources, we have added 23,916 new drug-gene interaction claims and 8,478 new druggable gene category claims

(Figure 4.2). In total, we have added two new sources of drug-gene interactions and five new sources of druggable gene category claims. DGIdb 4.0 now has 100,273 interaction claims and 33,577 druggable gene category claims. In total, there are now 10,606 druggable genes and 54,591 drug–gene interactions, which cover 41,102 genes and 14,449 drugs, within the DGIdb.

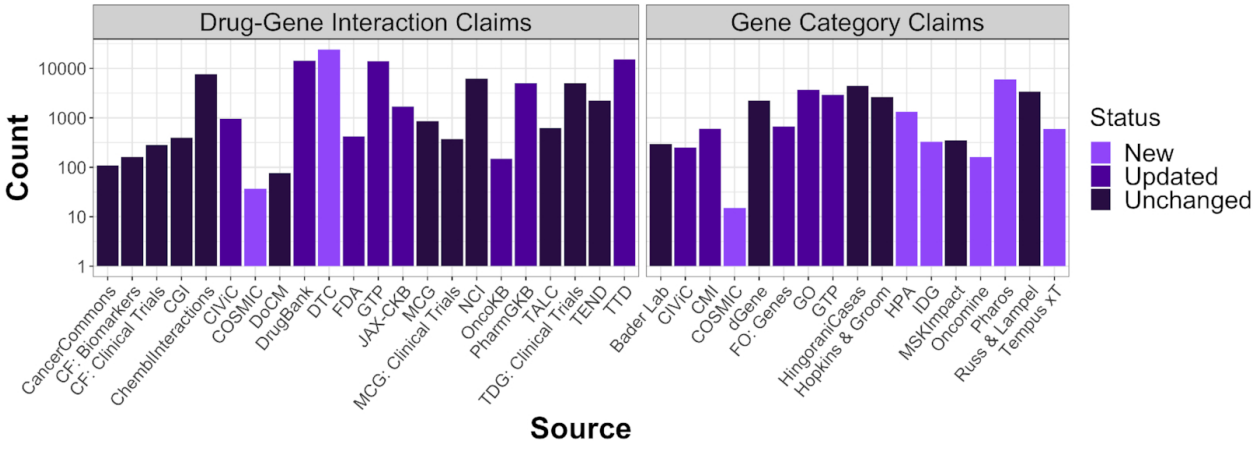


Figure 4.2: DGIdb 4.0 content by source

The number of drug-gene interaction claims (first panel) and druggable gene categories (second panel) are separated into three categories: sources that are new, sources that existed in the DGIdb previously but have been updated for 4.0, or sources that existed previously but have not been updated. Abbreviations: CF = Clearity Foundation, CGI = Cancer Genome Interpreter, CMI = Caris Molecular Intelligence, DTC = Drug Target Commons, FO = Foundation One, GO = Gene Ontology, GTP = Guide to Pharmacology, HPA = Human Protein Atlas, IDG = Illuminating the Druggable Genome, JAX-CKB = JAX-Clinical Knowledgebase, MCG = My Cancer Genome, MSK = Memorial Sloan Kettering, OncoKB = Precision Oncology Knowledge Base, TALC = Targeted Agents in Lung Cancer, TDG = The Druggable Genome, TEND = Trends in the Exploration of Novel Drug targets and TTD = Therapeutic Target Database.

We have also updated multiple sources including large, well-curated sources such as DrugBank, Guide to Pharmacology, Gene Ontology, OncoKB, PharmGKB, and the Therapeutic Target Database (Figure 4.2, Supplementary Table S1) (Armstrong et al., 2020; Ashburner et al., 2000; Chakravarty et al., 2017; The Gene Ontology Consortium, 2019; Y. Wang et al., 2020; Whirl-Carrillo et al., 2012; Wishart et al., 2018). To facilitate routine updates, the importer for PharmGKB has been updated to an online importer that can be run periodically using DGIdb's new automated job scheduling system. Similarly, Pharos, one of our new sources, has been implemented as an online updater (Nguyen et al., 2017). Of the 41 sources in DGIdb, 12 sources are now imported using the online updater format, including Entrez, the core source of gene concepts for gene grouping in DGIdb, and Ensembl, a key source of gene aliases (Brown et al., 2015; Yates et al., 2020). In DGIdb 4.0, the number of genes imported from Entrez has increased from 41,102 to 42,851 and Ensembl has been updated from version 90_38 to version 101_38. We have also migrated several older sources from our original domain-specific language (DSL) importers to the improved TSV importer style implemented in DGIdb 3.0.

In DGIdb 4.0, the database structure and presentation model was updated to allow sources to be imported with multiple source types. This change enables merging of sources that were previously duplicated for each independent claim type (drug, gene, interaction, druggable gene category). For example, we previously imported both interaction claims and druggable gene category claims from Guide to Pharmacology with two separate importers which created two separate sources (GuideToPharmacologyInteractions and GuideToPharmacologyGenes, respectively) (Armstrong et al., 2020). With this update, the import of interaction claims and druggable gene category claims is now handled by one importer and only a single source (GuideToPharmacology) is created. This is intended to simplify and unify claim sources to aid in

downstream interpretation. Additionally, supporting sources that have multiple source types enables easy extension to collect more informative claim type information. For example, some claims from CIViC can be imported with the additional categories of drug resistance and clinically actionable (Griffith et al., 2017). This change results in an additional 150 druggable gene category claims being imported from CIViC. Overall, these changes will increase the efficiency and accuracy of the process of importing and updating sources in DGIdb 4.0 compared to previous versions and will make it easier for users to evaluate individual sources.

4.4.3 Drug grouping improvements

Another notable change in the DGIdb 4.0 update is the improvement to drug grouping and normalization. Previously, we grouped drug claims using a rule-based pairwise association approach. This process was cumbersome, requiring a lengthy and complete re-grouping of all claims whenever we updated sources in order to generate consistent groupings. We revised this approach by creating a normalization component independent of the claims aggregated by DGIdb, that could be run on a per-source basis. When redesigning this part of DGIdb, we took steps to enable reuse of this normalizer as a modular component for other resources. To this end, we leveraged and contributed to a drug normalization service from the Variant Interpretation for Cancer Consortium (the '*thera-py*' Python Package; source code online at https://github.com/cancervariants/therapy-normalization). Among our contributions to *thera-py* was a normalizer for the Wikidata resource, further enabling community contributions to assist in concept normalization both for DGIdb and other resources reliant upon the VICC normalization services (Vrandečić & Krötzsch, 2014).

Drug claims from DGIdb were normalized using the ChEMBL and Wikidata normalizers from *thera-py*. Rules were written to formalize grouper behavior based upon match characteristics of a query. Briefly, these rules prioritize matches to primary labels over aliases, exact case over case-insensitive, and ChEMBL over other normalizers. An algorithm for constructing a merged drug concept from normalizer results was specified, enabling a standardized set of aliases for a given concept identifier. Pseudocode for this algorithm is provided (see Supplementary Data), and all implemented code is available on our public repository (see Data Availability).

4.4.4 New Query Score and updated Interaction Score

One of the main features added in DGIdb 4.0 is the concept of a relative *Query Score* for interaction search results. Previously, interaction search results displayed only a static *Interaction Score* based on evidence of an interaction (i.e. the number of publications and sources supporting an interaction claim). This Interaction Score did not take into account whether the gene and drug involved in a given interaction were also part of a large number of other interactions and, thus, had a low specificity that should be penalized. In addition, when searching for a set of genes or drugs, the Interaction Score does not prioritize results with overlapping interacting drugs or genes, which might be of more interest to the user, particularly in drug discovery and pathway applications.

DGIdb 4.0 now provides a Query Score that is relative to the search set and considers the overlap of interactions in the result set. For interaction searches using a gene list, the Query Score is calculated from the Evidence Scores (publications and sources), the number of genes from the search set that interact with the given drug, and the degree to which the drug has known

interactions with other genes (Figure 4.3). Similarly, for interaction searches using a drug list, the Query Score depends on the Evidence Scores (publications and sources), the number of drugs from the search set that interact with the given gene, and the degree to which the gene has known interactions with other drugs (Figure 4.3). In effect, this means that genes and drugs with many overlapping interactions in the search set will rank more highly, with the caveat that drugs or genes involved in many interactions, in general, will have lowered scores (Figure 4.3).

Our static Interaction Score previously introduced in DGIdb 3.0 has been adjusted in DGIdb 4.0 (Cotto et al., 2018). The Interaction Score now mirrors the Query Score, except it is unaffected by the queried gene or drug sets, instead relying only on Evidence Scores and the degree to which both the gene and drug are involved in other interactions (Figure 4.3). Interaction Scores follow a long-tail distribution, indicative of many highly promiscuous drugs and genes, and relatively few well-supported, highly specific drug-gene interactions (Supplementary Figure S1) (Haupt et al., 2013).

The introduction of the relative Query Score provides users with a score that gives a more intuitive ranking of drugs or genes based on the search set of interest, allowing the prioritization of drugs or genes that have overlapping, specific interactions with the search set. Similarly, the improvements to the static Interaction Score provide a more nuanced scoring system that takes into consideration the number of interactions for a drug-gene pair, in addition to the previous Evidence Scores, giving a more informative static Interaction Score.

As sources are updated and additional interaction claims are added to DGIdb, Interaction Scores and Query Scores are subject to change as a result of the changing measure to which Drug and Gene concepts interact with one another. Query Scores are always variable, dependent upon the set genes or drugs searched.

A

$$\text{relative drug specificity} = \frac{\text{average known gene partners for all drugs}}{\text{known gene partners for drug } d}$$

$$\text{relative gene specificity} = \frac{\text{average known drug partners for all genes}}{\text{known drug partners for gene } g}$$

$$\text{evidence score} = \text{publication count} + \text{source count}$$

Query Score (gene search):

evidence score • queried genes interacting with drug *d* • relative drug specificity

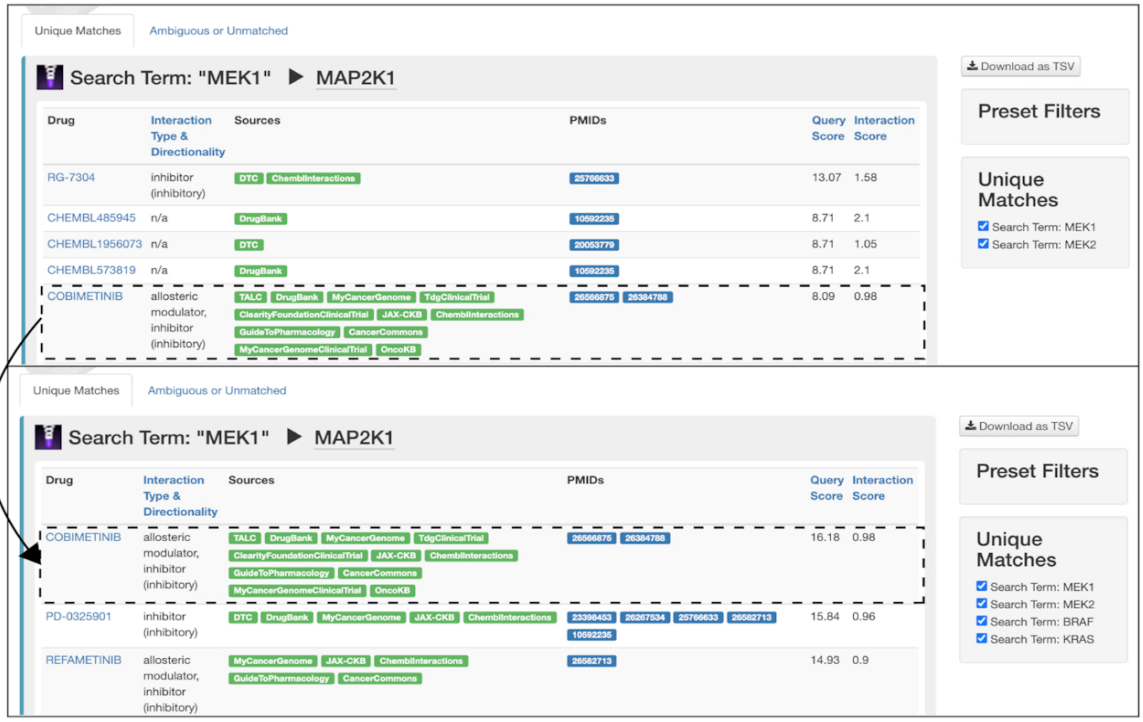
Query Score (drug search):

evidence score • queried drugs interacting with gene *g* • relative gene specificity

Interaction Score (no query):

evidence score • relative drug specificity • relative gene specificity

B



C

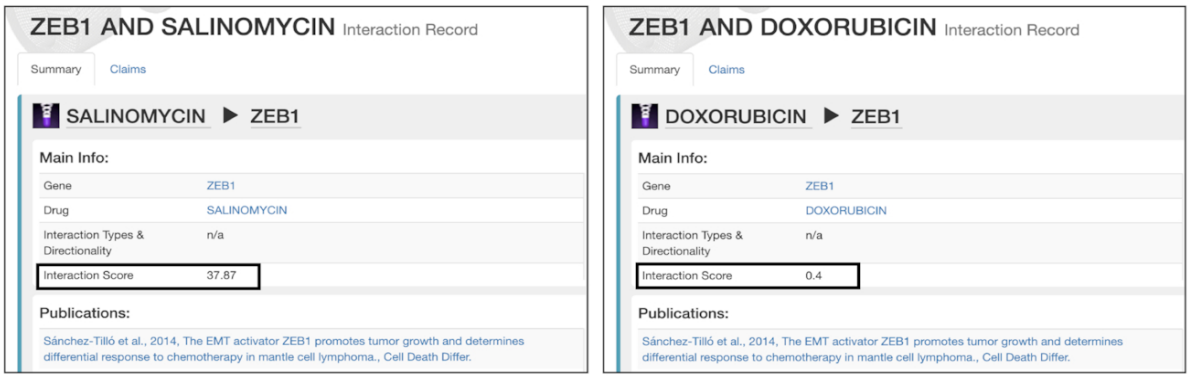


Figure 4.3: Overview of DGIdb's new Query scores and Interaction scores

(A) *Schematic of how each of the new scores is calculated within DGIdb. Gene and drug queries both return a Query Score that is dependent on the search terms. Each interaction has an Interaction Score that is calculated independently of other search terms.* **(B)** *Example of a Query Score changing based on the terms searched. In the first panel, only MEK1 and MEK2 were searched and the Query Score for the interaction between MEK1 and Cobimetinib was 8.09. In the second panel, BRAF and KRAS were added to the search query. These both interact with Cobimetinib and thus raise the Query Score to 16.18.* **(C)** *Example of Interaction Score. The panel on the left shows the interaction between ZEB1 and Salinomycin. This is the only interaction for Salinomycin and thus it has a high Interaction Score. The panel on the right shows the interaction between ZEB1 and Doxorubicin. Doxorubicin is involved in 103 interactions within DGIdb and thus has a much lower Interaction Score. Note that over time, as sources are updated and new claims are added, both Query Scores and Interaction Scores may change.*

4.4.5 Inclusion of interaction directionality

We have also added information on the directionality of interaction types to the interaction search results. Each interaction type in DGIdb now has an indicated directionality of *activating*, meaning the interaction type mechanism has an overall activating effect; *inhibitory*, meaning the interaction type mechanism has an overall inhibitory effect; or *n/a*, meaning the directionality is unclear for the interaction type mechanism (Supplementary Table S2). Determination of the directionality for an interaction type was made from mechanistic definitions provided by drug–gene interaction sources in which the interaction type was observed. Where these definitions

were not available, we instead relied upon community definitions of these interaction types. These interaction directionalities are included on the UI for interaction search results in parentheses next to the interaction type(s) listed for each interaction result (Supplementary Figure S2). While the directionality may be obvious for some interaction types (e.g. activators are activating), some interaction types are not immediately apparent (e.g. chaperones are activating) to those less familiar with mechanisms of drug–gene interactions. Inclusion of directionality can make it easier for users to distinguish interactions that are more relevant for their purposes. For example, a user interested in exploring drugs that inhibit a particular gene will look for drug-gene interactions with inhibitory directionality. Users are also able to limit their interaction search to only interaction types of a desired directionality. Detailed information on each interaction type, the definition of each interaction type, and the directionality of each interaction are also available on the DGIdb Interaction Types page (https://dgidb.org/interaction_types).

4.4.6 Search feature improvements

In DGIdb 4.0, we have introduced several updates related to searching, search results, and information available on the user interface. Among these are the addition of the option to search only cancer-specific sources (meaning sources that report claims relating to cancer only), or disease-agnostic sources (meaning sources that report claims relating to any disease, including cancer) for both interaction searches and druggable gene category searches. Cancer-related searches are a major use-case for DGIdb and cancer-specific sources are well-represented among all sources, with 13 cancer-specific drug-gene interaction sources and five cancer-specific druggable gene category sources. However, DGIdb is not a cancer-specific resource and is

intended to be utilized for non-cancer related research as well. For drug-gene interactions, there are 4,955 interactions supported by cancer-specific sources only, 48,341 interactions supported by disease-agnostic sources only, and 1,295 interactions are supported by both cancer-specific and disease-agnostic sources. Similarly, for druggable gene categories, 233 genes have categories supported by cancer-specific sources only, 17,168 genes have categories supported by disease-agnostic sources only and 2,804 genes have categories supported by both cancer-specific and disease-agnostic sources. These numbers show that although a sizable portion of the sources included in DGIdb are cancer-specific, those types of sources only represent a small proportion of the overall data.

Other improvements to search result features include the addition of linkouts to specific interaction evidence, where available. These will allow users to browse to the primary source for an interaction claim which might provide additional information and context not captured by DGIdb. Also, while we introduced drug and gene filters to the interaction search view in the last major update, we have had several requests to define how these filters are implemented. To address this lack of transparency on the UI, we have now added a link to the FAQ page where these filters are now defined.

4.4.7 Monthly data releases

DGIdb 3.0 implemented online updaters that imported data from dynamic sources (such as CIViC, Guide to Pharmacology, OncoKB, etc.) periodically, usually monthly (Armstrong et al., 2020; Chakravarty et al., 2017; Griffith et al., 2017). As a result, the static TSV data releases available on our Downloads page would quickly become outdated. For DGIdb 4.0, we have implemented monthly data releases of these TSVs to coincide with monthly runs of the online

updaters, to ensure that TSVs available for download reflect the most up to date information in our database. The Downloads page now makes available the current Gene, Drug, Interaction and Category TSVs as well as previous monthly TSVs since the release of DGIdb 4.0. These serve as *de facto* snapshots of the data in DGIdb over time.

4.4.8 Improved transparency and details on licensing of sources

In DGIdb 4.0, we have made a significant effort to update and improve the information we provide on licensing of sources imported into our database through manual curation of data license descriptions and references for every source. This information is now readily available on the sources page. Since DGIdb 3.0, several existing sources have made changes to their licensing, making data from some sources more broadly available and data from other sources more restricted. Notably, PharmGKB has moved to a more permissive Creative Commons Attribution-ShareAlike 4.0 International License and DrugBank has adopted a custom non-commercial license (Whirl-Carrillo et al., 2012; Wishart et al., 2018). In contrast, OncoKB has restricted API access to registered/approved non-commercial research use only, and JAX-CKB has restricted API access to negotiated licenses only (Chakravarty et al., 2017; Patterson et al., 2016). Both resources continue to provide access to a portion of their data for free through their respective web clients.

4.4.9 Application framework updates

To handle increased web traffic and integration with other tools, we have upgraded DGIdb to Rails 6 (from Rails 5), upgraded to Ruby 2.6.5 (from Ruby 2.3), upgraded to PostgreSQL 12 (from PostgreSQL 9.6), and upgraded the server to the latest Ubuntu LTS release (20.04). In

addition to the new features and performance benefits these upgrades bring, they will ensure that we continue to remain on supported software versions that receive regular security updates.

In order to keep DGIdb's underlying source data current, we had previously implemented an automated job scheduling framework using DelayedJob to schedule monthly runs of online updaters. In this release, we switched to using Sidekiq. In contrast to DelayedJob, Sidekiq offers a convenient user interface which makes identifying job failure reasons and rescheduling of failed jobs easier. Furthermore, the addition of Airbrake (https://airbrake.io/), an online tool for exception tracking, gives error reviews and notifies the development team of these errors in real-time (for instance, via email).

To ease future implementation of fixes and new features, we moved testing to a GitHub continuous integration (CI) workflow which allows us to continuously test newly committed code for errors against multiple versions of Ruby and PostgreSQL.

4.5 Discussion

With our most recent release, DGIdb has received significant improvements to source content, functionality such as searching and grouping, and underlying application technology. We have significantly expanded the number of records in our database through the addition of new sources and updates of existing sources. Furthermore, we have improved our ability to maintain regular content updates through the implementation of additional online importers for several sources and the use of Sidekiq for automatic job processing. We have revised our process for drug grouping and normalization to be batched by resource and to leverage continual improvement through community contributions to the VICC *thera-py* normalizers and the Wikidata public-domain crowdsourcing platform. Finally, several updates have been made to

inform users of the relevance of search results through information presented on the UI. We have implemented more sophisticated notations of relative and static interaction scores, improved the relevance of interaction source linkouts wherever possible, and included the concept of directionality for interaction results.

Although the updates in DGIdb 4.0 have improved the usability and content of our resource, we expect there will still be a need for future improvements. One technology improvement on our roadmap is converging the public-facing API with the internal code that powers the web views. Ultimately, we want the APIs available for general use to be the same ones powering our HTML pages. This would provide an even more fully featured API to end users while reducing our overall maintenance burden by eliminating redundant code. We are also evaluating the addition of information on gene-gene relationships. As always, we plan to continue updating sources to online updaters where possible, and migrating TSV-based sources from the legacy DSL importers to the TSV importers introduced in DGIdb 3.0.

4.6 Data Availability

DGIdb is an open access database and web interface (www.dgidb.org) with open source code available on GitHub (https://github.com/griffithlab/dgi-db) under the MIT license. We also provide data downloads for drug claims, gene claims, and interaction claims on the website in addition to a SQL data dump (http://dgidb.org/downloads). Information about the API and its endpoints can also be found on the website (http://dgidb.org/api).

4.7 Supplementary Data

Supplementary Data are available at NAR Online.

4.8 Acknowledgements

We want to thank the creators and maintainers of the seven new resources added to DGIdb and the many previously incorporated resources, as well as our growing community of users for notifying us of minor and major issues and for their suggestions on new features and improvements to DGIdb. We would also like to thank the members of the NDEX team, and in particular Dexter Pratt and Rudolf Pillich for their efforts in integrating DGIdb data into the NDEX resource.

4.9 Funding

National Human Genome Research Institute [K99HG010157 to A.H.W., A.C.C., R00HG007940 to M.G.]; National Cancer Institute [U01CA209936, U01CA248235, U24CA237719 to M.G., O.L.G., S.K., A.C.C., A.H.W., J.F.M.]. Funding for open access charge: departmental funding.

Conflict of interest statement. None declared.

Chapter 5: Conclusion

Understanding the genomic landscape of cancer and how it relates to the development and progression of disease is vitally important for understanding how to effectively target cancer cells and personalize treatments for individual patients. While the rise of next-generation sequencing and bioinformatic analysis of large omics datasets has been immeasurably useful for uncovering mechanisms that drive cancer as well as potential therapeutic targets, there is still much that remains unknown about cancer biology and how it relates to treatment response. Thus, the projects presented in this dissertation focused on the use of next-generation sequencing, bioinformatics, and database resources for furthering our understanding of cancer biology and treatment responses.

The first project presented described the use of single-cell RNA-sequencing, whole genome sequencing, and whole exome sequencing to explore the genomic landscape of a mouse model of bladder cancer and characterize mechanisms of response to a combined PD-1/CTLA-4 immune checkpoint blockade (ICB) treatment. This project revealed that IFN-g signaling in endothelial cells appears to be a key mechanism of effective response to ICB treatment. While the exact role that IFN-g signaling in endothelial cells plays in treatment response will need to be characterized and validated further, this finding suggests that therapeutic strategies that can induce IFN-g signaling in endothelial cells could present an avenue for improving response to ICB treatment in bladder cancer patients.

The second project presented described the use of whole genome sequencing and survival analysis for elucidating the landscape of copy number alterations in pediatric brain tumors and exploring the relationship between copy number alteration and survival across four diagnosis

groups (ATRT, Ependymoma, High-Grade Glioma, and Medulloblastoma). This project revealed that copy number alteration was quite common across diagnosis groups and could be extensive in certain groups. In addition, survival analysis examining the relationship between CNV burden and overall survival suggested that high CNV burden was associated with worse overall survival in certain diagnosis groups and may have prognostic value within those groups. Likewise, certain recurrent alterations showed correlation with worse overall survival in specific diagnosis groups. These results suggested that copy number alteration could play an important role in the development and progression of pediatric brain cancers. Follow-up analysis will be necessary to fully understand the role copy number alteration may play as a driving mechanism of cancer initiation and progression and how those driving mechanisms may relate to treatment susceptibility.

Finally, the third project presented described updates made to the Drug-Gene Interaction database (DGIdb) for the DGIdb 4.0 release to improve the content and usability of DGIdb. These improvements included the addition of several new drug, gene, and drug-gene interaction sources, integration with crowdsourced efforts to provide additional interaction information and refine drug term normalization, the inclusion of more sophisticated Query and Interaction scores, and updates to the user interface to improve functionality and ease of use. As of 2023, DGIdb 5.0 was released, encompassing a major overhaul of the architecture of the database to further expand the content, access, and usability of the DGIdb resource, which continues to be updated and improved. Proposed future enhancements for DGIdb include support for the submission of user curated interaction claims and evidence items, improvements to the data model used to store interaction information to allow better integration with clinical and drug discovery pipelines, and

incorporation of tools to flag retracted interaction evidence to guard against inclusion of false claims.

Together, the projects presented in this dissertation serve to further our understanding of cancer biology and treatment response and present several directions for future research to continue adding to our knowledge of the genomic landscape of cancer and how to develop effective treatment strategies.

References

Abd ElHafeez, S., D'Arrigo, G., Leonardis, D., Fusaro, M., Tripepi, G., & Roumeliotis, S. (2021). Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. *Oxidative Medicine and Cellular Longevity*, *2021*, 1302811.

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D. J., Shafin, K., Shumate, A., Xiao, C., Wagner, J., McDaniel, J., Olson, N. D., Sauria, M. E. G., Vollger, M. R., Rhie, A., Meredith, M., Martin, S., Lee, J., … Schatz, M. C. (2022). A complete reference genome improves analysis of human genetic variation. *Science*, *376*(6588), eabl3533.

Albisinni, S., Martinez Chanza, N., Aoun, F., Diamand, R., Mjaess, G., Azzo, J.-M., Esperto, F., Bellmunt, J., Roumeguère, T., & DE Nunzio, C. (2021). Immune checkpoint inhibitors for BCG-resistant NMIBC: the dawn of a new era. *Minerva Urology and Nephrology*, *73*(3), 292–298.

Alspach, E., Lussier, D. M., & Schreiber, R. D. (2019). Interferon γ and Its Important Roles in Promoting and Inhibiting Spontaneous and Therapeutic Cancer Immunity. *Cold Spring Harbor Perspectives in Biology*, *11*(3). https://doi.org/10.1101/cshperspect.a028480

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, *20*(2), 163–172.

Armstrong, J. F., Faccenda, E., Harding, S. D., Pawson, A. J., Southan, C., Sharman, J. L., Campo, B., Cavanagh, D. R., Alexander, S. P. H., Davenport, A. P., Spedding, M., Davies, J. A., & NC-IUPHAR. (2020). The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Research*, *48*(D1), D1006–D1021.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29.

Avila, M., & Meric-Bernstam, F. (2019). Next-generation sequencing for the general cancer patient. *Clinical Advances in Hematology & Oncology: H&O*, *17*(8), 447–454.

Barbato, M. I., Nashed, J., Bradford, D., Ren, Y., Khasar, S., Miller, C. P., Zolnik, B. S., Zhao, H., Li, Y., Bi, Y., Shord, S. S., Amatya, A. K., Mishra-Kalyani, P. S., Scepura, B., Al-Matari, R. A., Pazdur, R., Kluetz, P. G., Donoghue, M., Singh, H., & Drezner, N. (2023). FDA Approval Summary: Dabrafenib in combination with trametinib for BRAF V600E mutation-positive low-grade glioma. *Clinical Cancer Research: An Official Journal of the*

American Association for Cancer Research. https://doi.org/10.1158/1078-0432.CCR-23-1503

Barnell, E. K., Ronning, P., Campbell, K. M., Krysiak, K., Ainscough, B. J., Sheta, L. M., Pema, S. P., Schmidt, A. D., Richters, M., Cotto, K. C., Danos, A. M., Ramirez, C., Skidmore, Z. L., Spies, N. C., Hundal, J., Sediqzad, M. S., Kunisaki, J., Gomez, F., Trani, L., … Griffith, O. L. (2019). Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *21*(4), 972–981.

Barrow, J., Adamowicz-Brice, M., Cartmill, M., MacArthur, D., Lowe, J., Robson, K., Brundler, M.-A., Walker, D. A., Coyle, B., & Grundy, R. (2011). Homozygous loss of ADAM3A revealed by genome-wide analysis of pediatric high-grade glioma and diffuse intrinsic pontine gliomas. *Neuro-Oncology*, *13*(2), 212–222.

Beaubier, N., Tell, R., Lau, D., Parsons, J. R., Bush, S., Perera, J., Sorrells, S., Baker, T., Chang, A., Michuda, J., Iguartua, C., MacNeil, S., Shah, K., Ellis, P., Yeatts, K., Mahon, B., Taxter, T., Bontrager, M., Khan, A., … White, K. P. (2019). Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. *Oncotarget*, *10*(24), 2384–2396.

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. In *bioRxiv* (p. 861054). https://doi.org/10.1101/861054

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: survival analysis. *Critical Care / the Society of Critical Care Medicine*, *8*(5), 389–394.

Bray, F., Laversanne, M., Weiderpass, E., & Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, *127*(16), 3029–3030.

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., & Murphy, T. D. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, *43*(Database issue), D36–D42.

bwa.1. (n.d.). Retrieved December 5, 2023, from https://bio-bwa.sourceforge.net/bwa.shtml

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120.

Cannon, M., Stevenson, J., Stahl, K., Basu, R., Coffman, A., Kiwala, S., McMichael, J. F., Kuzma, K., Morrissey, D., Cotto, K., Mardis, E. R., Griffith, O. L., Griffith, M., & Wagner, A. H. (2023). DGIdb 5.0: rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkad1040

Chakravarty, D., Gao, J., Phillips, S. M., Kundra, R., Zhang, H., Wang, J., Rudolph, J. E.,

Yaeger, R., Soumerai, T., Nissan, M. H., Chang, M. T., Chandarlapaty, S., Traina, T. A., Paik, P. K., Ho, A. L., Hantash, F. M., Grupe, A., Baxi, S. S., Callahan, M. K., … Schultz, N. (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, *2017*. https://doi.org/10.1200/PO.17.00011

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., … Karczewski, K. J. (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. In *bioRxiv* (p. 2022.03.20.485034). https://doi.org/10.1101/2022.03.20.485034

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* , *32*(8), 1220–1222.

Chen, Z., Yuan, Y., Chen, X., Chen, J., Lin, S., Li, X., & Du, H. (2020). Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Scientific Reports*, *10*(1), 3501.

Choi, J. Y. (2023). Medulloblastoma: Current Perspectives and Recent Advances. *Brain Tumor Research and Treatment*, *11*(1), 28–38.

Choi, W., Czerniak, B., Ochoa, A., Su, X., Siefker-Radtke, A., Dinney, C., & McConkey, D. J. (2014). Intrinsic basal and luminal subtypes of muscle-invasive bladder cancer. *Nature Reviews. Urology*, *11*(7), 400–410.

Chrobak, I., Lenna, S., Stawski, L., & Trojanowska, M. (2013). Interferon-γ promotes vascular remodeling in human microvascular endothelial cells by upregulating endothelin (ET)-1 and transforming growth factor (TGF) β2. *Journal of Cellular Physiology*, *228*(8), 1774–1783.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., & Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–219.

Clarke, Z. A., Andrews, T. S., Atif, J., Pouyabahar, D., Innes, B. T., MacParland, S. A., & Bader, G. D. (2021). Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature Protocols*, *16*(6), 2749–2764.

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. *British Journal of Cancer*, *89*(2), 232–238.

Compérat, E., Amin, M. B., Cathomas, R., Choudhury, A., De Santis, M., Kamat, A., Stenzl, A., Thoeny, H. C., & Witjes, J. A. (2022). Current best practice for bladder cancer: a narrative review of diagnostics and treatments. *The Lancet*, *400*(10364), 1712–1721.

Cotto, K. C., Wagner, A. H., Feng, Y.-Y., Kiwala, S., Coffman, A. C., Spies, G., Wollam, A., Spies, N. C., Griffith, O. L., & Griffith, M. (2018). DGIdb 3.0: a redesign and expansion of

the drug-gene interaction database. *Nucleic Acids Research*, *46*(D1), D1068–D1073.

Dadhania, V., Zhang, M., Zhang, L., Bondaruk, J., Majewski, T., Siefker-Radtke, A., Guo, C. C., Dinney, C., Cogdell, D. E., Zhang, S., Lee, S., Lee, J. G., Weinstein, J. N., Baggerly, K., McConkey, D., & Czerniak, B. (2016). Meta-Analysis of the Luminal and Basal Subtypes of Bladder Cancer and the Identification of Signature Immunohistochemical Markers for Clinical Use. *EBioMedicine*, *12*, 105–117.

Damodharan, S., & Puccetti, D. (2023). Pediatric Central Nervous System Tumor Overview and Emerging Treatment Considerations. *Brain Sciences*, *13*(7). https://doi.org/10.3390/brainsci13071106

Das, S., Rai, A., Merchant, M. L., Cave, M. C., & Rai, S. N. (2021). A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies. *Genes*, *12*(12). https://doi.org/10.3390/genes12121947

Datto, M., & Lundblad, R. L. (2016). DNA, RNA Chemical Properties (Including Sequencing and Next-Generation Sequencing). In R. A. Bradshaw & P. D. Stahl (Eds.), *Encyclopedia of Cell Biology* (pp. 24–35). Academic Press.

de Bont, J. M., Packer, R. J., Michiels, E. M., den Boer, M. L., & Pieters, R. (2008). Biological background of pediatric medulloblastoma and ependymoma: a review from a translational research perspective. *Neuro-Oncology*, *10*(6), 1040–1060.

De Bortoli, M., Castellino, R. C., Lu, X.-Y., Deyo, J., Sturla, L. M., Adesina, A. M., Perlaky, L., Pomeroy, S. L., Lau, C. C., Man, T.-K., Rao, P. H., & Kim, J. Y. H. (2006). Medulloblastoma outcome is adversely associated with overexpression of EEF1D, RPL30, and RPS20 on the long arm of chromosome 8. *BMC Cancer*, *6*, 223.

De Cesco, S., Davis, J. B., & Brennan, P. E. (2020). TargetDB: A target information aggregation tool and tractability predictor. *PloS One*, *15*(9), e0232644.

Deng, X., Das, S., Valdez, K., Camphausen, K., & Shankavaram, U. (2019). SL-BioDP: Multi-Cancer Interactive Tool for Prediction of Synthetic Lethality and Response to Cancer Treatment. *Cancers*, *11*(11). https://doi.org/10.3390/cancers11111682

Deo, S. V., Deo, V., & Sundaram, V. (2021). Survival analysis-part 2: Cox proportional hazards model. *Indian Journal of Thoracic and Cardiovascular Surgery*, *37*(2), 229–233.

Diori Karidio, I., & Sanlier, S. H. (2021). Reviewing cancer's biology: an eclectic approach. *Journal of the Egyptian National Cancer Institute*, *33*(1), 32.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Doussouki, M. E., Gajjar, A., & Chamdine, O. (2019). Molecular genetics of medulloblastoma in children: diagnostic, therapeutic and prognostic implications. *Future Neurology*, *14*(1), FNL8.

Freshour, S. L., Kiwala, S., Cotto, K. C., Coffman, A. C., McMichael, J. F., Song, J. J., Griffith, M., Griffith, O. L., & Wagner, A. H. (2021). Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Research*, *49*(D1), D1144–D1151.

Fults, D., Petronio, J., Noblett, B. D., & Pedone, C. A. (1992). Chromosome 11p15 deletions in human malignant astrocytomas and primitive neuroectodermal tumors. *Genomics*, *14*(3), 799–801.

Gao, J., Navai, N., Alhalabi, O., Siefker-Radtke, A., Campbell, M. T., Tidwell, R. S., Guo, C. C., Kamat, A. M., Matin, S. F., Araujo, J. C., Shah, A. Y., Msaouel, P., Corn, P., Wang, J., Papadopoulos, J. N., Yadav, S. S., Blando, J. M., Duan, F., Basu, S., … Sharma, P. (2020). Neoadjuvant PD-L1 plus CTLA-4 blockade in patients with cisplatin-ineligible operable high-risk urothelial carcinoma. *Nature Medicine*, *26*(12), 1845–1851.

Gap Track Settings. (n.d.). Retrieved December 6, 2023, from https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=gap

Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research*, *1*(4), 274–278.

Griffith, M., Griffith, O. L., Coffman, A. C., Weible, J. V., McMichael, J. F., Spies, N. C., Koval, J., Das, I., Callaway, M. B., Eldred, J. M., Miller, C. A., Subramanian, J., Govindan, R., Kumar, R. D., Bose, R., Ding, L., Walker, J. R., Larson, D. E., Dooling, D. J., … Wilson, R. K. (2013). DGIdb: mining the druggable genome. *Nature Methods*, *10*(12), 1209–1210.

Griffith, M., Griffith, O. L., Smith, S. M., Ramu, A., Callaway, M. B., Brummett, A. M., Kiwala, M. J., Coffman, A. C., Regier, A. A., Oberkfell, B. J., Sanderson, G. E., Mooney, T. P., Nutter, N. G., Belter, E. A., Du, F., Long, R. L., Abbott, T. E., Ferguson, I. T., Morton, D. L., … Wilson, R. K. (2015). Genome Modeling System: A Knowledge Management Platform for Genomics. *PLoS Computational Biology*, *11*(7), e1004274.

Griffith, M., Miller, C. A., Griffith, O. L., Krysiak, K., Skidmore, Z. L., Ramu, A., Walker, J. R., Dang, H. X., Trani, L., Larson, D. E., Demeter, R. T., Wendl, M. C., McMichael, J. F., Austin, R. E., Magrini, V., McGrath, S. D., Ly, A., Kulkarni, S., Cordes, M. G., … Wilson, R. K. (2015). Optimizing cancer genome sequencing and analysis. *Cell Systems*, *1*(3), 210–223.

Griffith, M., Spies, N. C., Krysiak, K., McMichael, J. F., Coffman, A. C., Danos, A. M., Ainscough, B. J., Ramirez, C. A., Rieke, D. T., Kujan, L., Barnell, E. K., Wagner, A. H., Skidmore, Z. L., Wollam, A., Liu, C. J., Jones, M. R., Bilski, R. L., Lesurf, R., Feng, Y.-Y., … Griffith, O. L. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, *49*(2), 170–174.

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., & Newman, A. M. (2020). Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, *367*(6476), 405–

411.

Guo, C. C., Bondaruk, J., Yao, H., Wang, Z., Zhang, L., Lee, S., Lee, J.-G., Cogdell, D., Zhang, M., Yang, G., Dadhania, V., Choi, W., Wei, P., Gao, J., Theodorescu, D., Logothetis, C., Dinney, C., Kimmel, M., Weinstein, J. N., … Czerniak, B. (2020). Assessment of Luminal and Basal Phenotypes in Bladder Cancer. *Scientific Reports*, *10*(1), 9743.

Hajduk, P. J., Huth, J. R., & Tse, C. (2005). Predicting protein druggability. *Drug Discovery Today*, *10*(23-24), 1675–1682.

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573–3587.e29.

Haupt, V. J., Daminelli, S., & Schroeder, M. (2013). Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PloS One*, *8*(6), e65894.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *BioTechniques*, *56*(2), 61–64, 66, 68, passim.

Heng, T. S. P., Painter, M. W., & Immunological Genome Project Consortium. (2008). The Immunological Genome Project: networks of gene expression in immune cells. *Nature Immunology*, *9*(10), 1091–1094.

Hess, J. F., Kohl, T. A., Kotrová, M., Rönsch, K., Paprotka, T., Mohr, V., Hutzenlaub, T., Brüggemann, M., Zengerle, R., Niemann, S., & Paust, N. (2020). Library preparation for next generation sequencing: A review of automation strategies. *Biotechnology Advances*, *41*, 107537.

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Single-cell Best Practices Consortium, Schiller, H. B., & Theis, F. J. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews. Genetics*, *24*(8), 550–572.

Ho, B., Johann, P. D., Grabovska, Y., De Dieu Andrianteranagna, M. J., Yao, F., Frühwald, M., Hasselblatt, M., Bourdeaut, F., Williamson, D., Huang, A., & Kool, M. (2020). Molecular subgrouping of atypical teratoid/rhabdoid tumors-a reinvestigation and current consensus. *Neuro-Oncology*, *22*(5), 613–624.

Hopkins, A. L., & Groom, C. R. (2002). The druggable genome. *Nature Reviews. Drug Discovery*, *1*(9), 727–730.

Huang, D., Ma, N., Li, X., Gou, Y., Duan, Y., Liu, B., Xia, J., Zhao, X., Wang, X., Li, Q., Rao, J., & Zhang, X. (2023). Advances in single-cell RNA sequencing and its applications in cancer research. *Journal of Hematology & Oncology*, *16*(1), 98.

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801–811.

Immunological Genome Project. (2020). ImmGen at 15. *Nature Immunology*, *21*(7), 700–703.

Kammertoens, T., Friese, C., Arina, A., Idel, C., Briesemeister, D., Rothe, M., Ivanov, A., Szymborska, A., Patone, G., Kunz, S., Sommermeyer, D., Engels, B., Leisegang, M., Textor, A., Fehling, H. J., Fruttiger, M., Lohoff, M., Herrmann, A., Yu, H., … Blankenstein, T. (2017). Tumour ischaemia by interferon-γ resembles physiological blood vessel regression. *Nature*, *545*(7652), 98–102.

Kaplan, D. H., Shankaran, V., Dighe, A. S., Stockert, E., Aguet, M., Old, L. J., & Schreiber, R. D. (1998). Demonstration of an interferon γ-dependent tumor surveillance system in immunocompetent mice. *Proceedings of the National Academy of Sciences*, *95*(13), 7556–7561.

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Kartolo, A., Kassouf, W., & Vera-Badillo, F. E. (2021). Adjuvant Immune Checkpoint Inhibition in Muscle-invasive Bladder Cancer: Is It Ready for Prime Time? *European Urology*, *80*(6), 679–681.

Khanna, A., Larson, D. E., Srivatsan, S. N., Mosior, M., Abbott, T. E., Kiwala, S., Ley, T. J., Duncavage, E. J., Walter, M. J., Walker, J. R., Griffith, O. L., Griffith, M., & Miller, C. A. (2021). Bam-readcount - rapid generation of basepair-resolution sequence metrics. *ArXiv*. https://www.ncbi.nlm.nih.gov/pubmed/34341766

Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, *18*(7), 723–732.

Kijima, N., & Kanemura, Y. (2016). Molecular Classification of Medulloblastoma. *Neurologia Medico-Chirurgica*, *56*(11), 687–697.

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., & Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, *15*(8), 591–594.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576.

Kohli, K., Pillarisetty, V. G., & Kim, T. S. (2022). Key chemokines direct migration of immune cells in solid tumors. *Cancer Gene Therapy*, *29*(1), 10–21.

Kool, M., Korshunov, A., Remke, M., Jones, D. T. W., Schlanstein, M., Northcott, P. A., Cho, Y.-J., Koster, J., Schouten-van Meeteren, A., van Vuurden, D., Clifford, S. C., Pietsch, T., von Bueren, A. O., Rutkowski, S., McCabe, M., Collins, V. P., Bäcklund, M. L., Haberler, C., Bourdeaut, F., … Pfister, S. M. (2012). Molecular subgroups of medulloblastoma: an

international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. *Acta Neuropathologica*, *123*(4), 473–484.

Kool, M., Koster, J., Bunt, J., Hasselt, N. E., Lakeman, A., van Sluis, P., Troost, D., Meeteren, N. S., Caron, H. N., Cloos, J., Mrsić, A., Ylstra, B., Grajkowska, W., Hartmann, W., Pietsch, T., Ellison, D., Clifford, S. C., & Versteeg, R. (2008). Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PloS One*, *3*(8), e3088.

Kulubya, E. S., Kercher, M. J., Phillips, H. W., Antony, R., & Edwards, M. S. B. (2022). Advances in the Treatment of Pediatric Brain Tumors. *Children*, *10*(1). https://doi.org/10.3390/children10010062

Lenis, A. T., Lec, P. M., Chamie, K., & Mshs, M. D. (2020). Bladder Cancer: A Review. *JAMA: The Journal of the American Medical Association*, *324*(19), 1980–1991.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E.-A. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, D., & Nolan, G. P. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, *162*(1), 184–197.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems*, *1*(6), 417–425.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* , *27*(12), 1739–1740.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , *25*(14), 1754–1760.

Liu, Y.-P., Zheng, C.-C., Huang, Y.-N., He, M.-L., Xu, W. W., & Li, B. (2021). Molecular mechanisms of chemo- and radiotherapy resistance and the potential implications for cancer treatment. *MedComm*, *2*(3), 315–340.

Li, X., & Wang, C.-Y. (2021). From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science*, *13*(1), 36.

Lo, K. C., Ma, C., Bundy, B. N., Pomeroy, S. L., Eberhart, C. G., & Cowell, J. K. (2007). Gain of 1q is a potential univariate negative prognostic marker for survival in medulloblastoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *13*(23), 7022–7028.

Lopez-Beltran, A., Cimadamore, A., Blanca, A., Massari, F., Vau, N., Scarpelli, M., Cheng, L., & Montironi, R. (2021). Immune Checkpoint Inhibitors for the Treatment of Bladder Cancer. *Cancers*, *13*(1). https://doi.org/10.3390/cancers13010131

Manoharan, N., Liu, K. X., Mueller, S., Haas-Kogan, D. A., & Bandopadhayay, P. (2023). Pediatric low-grade glioma: Targeted therapeutics and clinical trials in the molecular era. *Neoplasia* , *36*, 100857.

Mardis, E. R. (2019). The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic. *Cold Spring Harbor Perspectives in Medicine*, *9*(9). https://doi.org/10.1101/cshperspect.a036269

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., Gonzalez-Perez, A., & Lopez-Bigas, N. (2020). A compendium of mutational cancer driver genes. *Nature Reviews. Cancer*, *20*(10), 555–572.

McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, *8*(4), 329–337.e4.

Medulloblastoma diagnosis and treatment. (2018, September 17). National Cancer Institute. https://www.cancer.gov/rare-brain-spine-tumor/tumors/medulloblastoma

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., … Leach, A. R. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, *47*(D1), D930–D940.

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41.

Miles, B., & Tadi, P. (2023). *Genetics, Somatic Mutation*. StatPearls Publishing.

Min, H.-Y., & Lee, H.-Y. (2022). Molecular targeted therapy for anticancer treatment. *Experimental & Molecular Medicine*, *54*(10), 1670–1694.

Mueller, T., Stucklin, A. S. G., Postlmayr, A., Metzger, S., Gerber, N., Kline, C., Grotzer, M., Nazarian, J., & Mueller, S. (2020). Advances in Targeted Therapies for Pediatric Brain Tumors. *Current Treatment Options in Neurology*, *22*(12), 43.

Muscle-invasive and Metastatic Bladder Cancer. (n.d.). Uroweb - European Association of Urology. Retrieved December 9, 2023, from https://uroweb.org/guidelines/muscle-invasive-and-metastatic-bladder-cancer/chapter/disease-management

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582.

Nguyen, D.-T., Mathias, S., Bologa, C., Brunak, S., Fernandez, N., Gaulton, A., Hersey, A., Holmes, J., Jensen, L. J., Karlsson, A., Liu, G., Ma'ayan, A., Mandava, G., Mani, S., Mehta, S., Overington, J., Patel, J., Rouillard, A. D., Schürer, S., … Guha, R. (2017). Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids*

Research, *45*(D1), D995–D1002.

Nicholas, T. J., Cormier, M. J., Huang, X., Qiao, Y., Marth, G. T., & Quinlan, A. R. (2020). *OncoGEMINI: Software for Investigating Tumor Variants From Multiple Biopsies With Integrated Cancer Annotations* (p. 2020.03.10.979591). https://doi.org/10.1101/2020.03.10.979591

Ni, L., & Lu, J. (2018). Interferon gamma in cancer immunotherapy. *Cancer Medicine*, *7*(9), 4509–4516.

Nones, K., & Patch, A.-M. (2020). The Impact of Next Generation Sequencing in Cancer Research. *Cancers*, *12*(10). https://doi.org/10.3390/cancers12102928

Oh, D. Y., Kwek, S. S., Raju, S. S., Li, T., McCarthy, E., Chow, E., Aran, D., Ilano, A., Pai, C.-C. S., Rancan, C., Allaire, K., Burra, A., Sun, Y., Spitzer, M. H., Mangul, S., Porten, S., Meng, M. V., Friedlander, T. W., Ye, C. J., & Fong, L. (2020). Intratumoral CD4+ T Cells Mediate Anti-tumor Cytotoxicity in Human Bladder Cancer. *Cell*, *181*(7), 1612–1625.e13.

Ostrom, Q. T., de Blank, P. M., Kruchko, C., Petersen, C. M., Liao, P., Finlay, J. L., Stearns, D. S., Wolff, J. E., Wolinsky, Y., Letterio, J. J., & Barnholtz-Sloan, J. S. (2015). Alex's Lemonade Stand Foundation Infant and Childhood Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2007-2011. *Neuro-Oncology*, *16 Suppl 10*(Suppl 10), x1–x36.

Ostrom, Q. T., Price, M., Ryan, K., Edelson, J., Neff, C., Cioffi, G., Waite, K. A., Kruchko, C., & Barnholtz-Sloan, J. S. (2022). CBTRUS Statistical Report: Pediatric Brain Tumor Foundation Childhood and Adolescent Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2014-2018. *Neuro-Oncology*, *24*(Suppl 3), iii1–iii38.

Pasquini, G., Rojo Arias, J. E., Schäfer, P., & Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*, *19*, 961–969.

Patterson, S. E., Liu, R., Statz, C. M., Durkin, D., Lakshminarayana, A., & Mockus, S. M. (2016). The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Human Genomics*, *10*, 4.

Pervez, M. T., Hasnain, M. J. U., Abbas, S. H., Moustafa, M. F., Aslam, N., & Shah, S. S. M. (2022). A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *BioMed Research International*, *2022*, 3457806.

Petrucelli, N., Daly, M. B., & Pal, T. (1998). BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer. In M. P. Adam, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, L. J. H. Bean, K. W. Gripp, & A. Amemiya (Eds.), *GeneReviews*®. University of Washington, Seattle.

Petti, A. A., Williams, S. R., Miller, C. A., Fiddes, I. T., Srivatsan, S. N., Chen, D. Y., Fronick, C. C., Fulton, R. S., Church, D. M., & Ley, T. J. (2019). A general approach for detecting

expressed mutations in AML cells using single cell RNA-sequencing. *Nature Communications*, *10*(1), 3660.

Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., Misra, A., Nigro, J. M., Colman, H., Soroceanu, L., Williams, P. M., Modrusan, Z., Feuerstein, B. G., & Aldape, K. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, *9*(3), 157–173.

Pogorzala, M., Styczynski, J., & Wysocki, M. (2014). Survival and prognostic factors in children with brain tumors: long-term follow-up single center study in Poland. *Anticancer Research*, *34*(1), 323–326.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. In *Cold Spring Harbor Laboratory* (p. 201178). https://doi.org/10.1101/201178

Pratt, D., Chen, J., Pillich, R., Rynkov, V., Gary, A., Demchak, B., & Ideker, T. (2017). NDEx 2.0: A Clearinghouse for Research on Cancer Pathways. *Cancer Research*, *77*(21), e58–e61.

Radoux, C. J., Vianello, F., McGreig, J., Desai, N., & Bradley, A. R. (2022). The druggable genome: Twenty years later. *Frontiers in Bioinformatics*, *2*, 958378.

Raies, A., Tulodziecka, E., Stainer, J., Middleton, L., Dhindsa, R. S., Hill, P., Engkvist, O., Harper, A. R., Petrovski, S., & Vitsios, D. (2022). DrugnomeAI is an ensemble machine-learning framework for predicting druggability of candidate drug targets. *Communications Biology*, *5*(1), 1291.

Rhea, L. P., & Aragon-Ching, J. B. (2021). Advances and Controversies With Checkpoint Inhibitors in Bladder Cancer. *Clinical Medicine Insights. Oncology*, *15*, 11795549211044963.

Robertson, A. G., Kim, J., Al-Ahmadie, H., Bellmunt, J., Guo, G., Cherniack, A. D., Hinoue, T., Laird, P. W., Hoadley, K. A., Akbani, R., Castro, M. A. A., Gibb, E. A., Kanchi, R. S., Gordenin, D. A., Shukla, S. A., Sanchez-Vega, F., Hansel, D. E., Czerniak, B. A., Reuter, V. E., … Lerner, S. P. (2017). Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell*, *171*(3), 540–556.e25.

Rodgers, G., Austin, C., Anderson, J., Pawlyk, A., Colvis, C., Margolis, R., & Baker, J. (2018). Glimmers in illuminating the druggable genome. *Nature Reviews. Drug Discovery*, *17*(5), 301–302.

Rosenberg, J. E., Hoffman-Censits, J., Powles, T., van der Heijden, M. S., Balar, A. V., Necchi, A., Dawson, N., O'Donnell, P. H., Balmanoukian, A., Loriot, Y., Srinivas, S., Retz, M. M., Grivas, P., Joseph, R. W., Galsky, M. D., Fleming, M. T., Petrylak, D. P., Perez-Gracia, J. L., Burris, H. A., … Dreicer, R. (2016). Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-

based chemotherapy: a single-arm, multicentre, phase 2 trial. *The Lancet*, *387*(10031), 1909–1920.

Roviello, G., Catalano, M., Nobili, S., Santi, R., Mini, E., & Nesi, G. (2020). Focus on Biochemical and Clinical Predictors of Response to Immune Checkpoint Inhibitors in Metastatic Urothelial Carcinoma: Where Do We Stand? *International Journal of Molecular Sciences*, *21*(21). https://doi.org/10.3390/ijms21217935

Roviello, G., Catalano, M., Santi, R., Palmieri, V. E., Vannini, G., Galli, I. C., Buttitta, E., Villari, D., Rossi, V., & Nesi, G. (2021). Immune Checkpoint Inhibitors in Urothelial Bladder Cancer: State of the Art and Future Perspectives. *Cancers*, *13*(17). https://doi.org/10.3390/cancers13174411

Sakatani, T., Kita, Y., Fujimoto, M., Sano, T., Hamada, A., Nakamura, K., Takada, H., Goto, T., Sawada, A., Akamatsu, S., & Kobayashi, T. (2022). IFN-Gamma Expression in the Tumor Microenvironment and CD8-Positive Tumor-Infiltrating Lymphocytes as Prognostic Markers in Urothelial Cancer Patients Receiving Pembrolizumab. *Cancers*, *14*(2). https://doi.org/10.3390/cancers14020263

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology*, *12*(7). https://doi.org/10.3390/biology12070997

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502.

Sato, Y., Bolzenius, J. K., Eteleeb, A. M., Su, X., Maher, C. A., Sehn, J. K., & Arora, V. K. (2018). CD4+ T cells induce rejection of urothelial tumors after immune checkpoint blockade. *JCI Insight*, *3*(23). https://doi.org/10.1172/jci.insight.121062

Schirrmacher, V. (2019). From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment (Review). *International Journal of Oncology*, *54*(2), 407–419.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J. J., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J. Z., Park, H., & Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, *498*(7453), 236–240.

Shapiro, J. A., Gaonkar, K. S., Spielman, S. J., Savonen, C. L., Bethell, C. J., Jin, R., Rathi, K. S., Zhu, Y., Egolf, L. E., Farrow, B. K., Miller, D. P., Yang, Y., Koganti, T., Noureen, N., Koptyra, M. P., Duong, N., Santi, M., Kim, J., Robins, S., … Pacific Pediatric Neuro-Oncology Consortium. (2023). OpenPBTA: The Open Pediatric Brain Tumor Atlas. *Cell Genomics*, *3*(7), 100340.

Sharma, P., Bono, P., Kim, J. W., Spiliopoulou, P., Calvo, E., Pillai, R. N., Ott, P. A., De Braud, F. G., Morse, M. A., Le, D. T., Jaeger, D., Chan, E., Harbison, C. T., Lin, C.-S., Tschaika,

M., Azrilevich, A., & Rosenberg, J. (2016). Efficacy and safety of nivolumab monotherapy in metastatic urothelial cancer (mUC): Results from the phase I/II CheckMate 032 study. *Journal of Clinical Orthodontics: JCO*, *34*(15_suppl), 4501–4501.

Sharma, P., Retz, M., Siefker-Radtke, A., Baron, A., Necchi, A., Bedke, J., Plimack, E. R., Vaena, D., Grimm, M.-O., Bracarda, S., Arranz, J. Á., Pal, S., Ohyama, C., Saci, A., Qu, X., Lambert, A., Krishnan, S., Azrilevich, A., & Galsky, M. D. (2017). Nivolumab in metastatic urothelial carcinoma after platinum therapy (CheckMate 275): a multicentre, single-arm, phase 2 trial. *The Lancet Oncology*, *18*(3), 312–322.

Shay, T., & Kang, J. (2013). Immunological Genome Project and systems immunology. *Trends in Immunology*, *34*(12), 602–609.

Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, *15*(4), 255–261.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, *12*(1), 5692.

Stanić, D., Grujičić, D., Pekmezović, T., Bokun, J., Popović-Vuković, M., Janić, D., Paripović, L., Ilić, V., Pudrlja Slović, M., Ilić, R., Raičević, S., Sarić, M., Mišković, I., Nidžović, B., & Nikitović, M. (2021). Clinical profile, treatment and outcome of pediatric brain tumors in Serbia in a 10-year period: A national referral institution experience. *PloS One*, *16*(10), e0259095.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *54*, 1.30.1–1.30.33.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888–1902.e21.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Subramanian, S., & Ahmad, T. (2023). Childhood Brain Tumors. In *StatPearls*. StatPearls Publishing.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3),

209–249.

Survival rates for bladder cancer. (n.d.). Retrieved December 9, 2023, from https://www.cancer.org/cancer/bladder-cancer/detection-diagnosis-staging/survival-rates.html

Suzman, D. L., Agrawal, S., Ning, Y.-M., Maher, V. E., Fernandes, L. L., Karuri, S., Tang, S., Sridhara, R., Schroeder, J., Goldberg, K. B., Ibrahim, A., McKee, A. E., Pazdur, R., & Beaver, J. A. (2019). FDA Approval Summary: Atezolizumab or Pembrolizumab for the Treatment of Patients with Advanced Urothelial Carcinoma Ineligible for Cisplatin-Containing Chemotherapy. *The Oncologist*, *24*(4), 563–569.

Talevich, E., Shain, A. H., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Computational Biology*, *12*(4), e1004873.

Tanoli, Z., Alam, Z., Vähä-Koskela, M., Ravikumar, B., Malyutina, A., Jaiswal, A., Tang, J., Wennerberg, K., & Aittokallio, T. (2018). Drug Target Commons 2.0: a community platform for systematic analysis of drug-target interaction profiles. *Database: The Journal of Biological Databases and Curation*, *2018*, 1–13.

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947.

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338.

Thomas, A., & Noël, G. (2019). Medulloblastoma: optimizing care with a multidisciplinary approach. *Journal of Multidisciplinary Healthcare*, *12*, 335–347.

Tilsed, C. M., Fisher, S. A., Nowak, A. K., Lake, R. A., & Lesterhuis, W. J. (2022). Cancer chemotherapy: insights into cellular and tumor microenvironmental mechanisms of action. *Frontiers in Oncology*, *12*, 960317.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233.

Tsimberidou, A. M., Fountzilas, E., Nikanjam, M., & Kurzrock, R. (2020). Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer Treatment Reviews*, *86*, 102019.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J. M., Brunnström, H., … Ponten, F. (2017). A pathology atlas of the human cancer transcriptome. *Science*, *357*(6352). https://doi.org/10.1126/science.aan2507

van Dijk, N., Gil-Jimenez, A., Silina, K., Hendricksen, K., Smit, L. A., de Feijter, J. M., van

Montfoort, M. L., van Rooijen, C., Peters, D., Broeks, A., van der Poel, H. G., Bruining, A., Lubeck, Y., Sikorska, K., Boellaard, T. N., Kvistborg, P., Vis, D. J., Hooijberg, E., Schumacher, T. N., … van der Heijden, M. S. (2020). Preoperative ipilimumab plus nivolumab in locoregionally advanced urothelial cancer: the NABUCCO trial. *Nature Medicine*, *26*(12), 1839–1844.

vartrix: Single-Cell Genotyping Tool. (n.d.). Github. Retrieved October 15, 2023, from https://github.com/10XGenomics/vartrix

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85.

Wagner, A. H., Coffman, A. C., Ainscough, B. J., Spies, N. C., Skidmore, Z. L., Campbell, K. M., Krysiak, K., Pan, D., McMichael, J. F., Eldred, J. M., Walker, J. R., Wilson, R. K., Mardis, E. R., Griffith, M., & Griffith, O. L. (2016). DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Research*, *44*(D1), D1036–D1044.

Wang, C., Yang, J., Luo, H., Wang, K., Wang, Y., Xiao, Z.-X., Tao, X., Jiang, H., & Cai, H. (2020). CancerTracer: a curated database for intrapatient tumor heterogeneity. *Nucleic Acids Research*, *48*(D1), D797–D806.

Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., Zhang, R., Zhu, J., Ren, Y., Tan, Y., Qin, C., Li, Y., Li, X., Chen, Y., & Zhu, F. (2020). Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Research*, *48*(D1), D1031–D1041.

Warren, K. E., Killian, K., Suuriniemi, M., Wang, Y., Quezado, M., & Meltzer, P. S. (2012). Genomic aberrations in pediatric diffuse intrinsic pontine gliomas. *Neuro-Oncology*, *14*(3), 326–332.

Wemmert, S., Ketter, R., Rahnenführer, J., Beerenwinkel, N., Strowitzki, M., Feiden, W., Hartmann, C., Lengauer, T., Stockhammer, F., Zang, K. D., Meese, E., Steudel, W.-I., von Deimling, A., & Urbschat, S. (2005). Patients with high-grade gliomas harboring deletions of chromosomes 9p and 10q benefit from temozolomide treatment. *Neoplasia* , *7*(10), 883–893.

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., Altman, R. B., & Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*, *92*(4), 414–417.

Williams, H. L., Walsh, K., Diamond, A., Oniscu, A., & Deans, Z. C. (2018). Validation of the Oncomine™ focus panel for next-generation sequencing of clinical tumour samples. *Virchows Archiv: An International Journal of Pathology*, *473*(4), 489–503.

Williamson, D., Schwalbe, E. C., Hicks, D., Aldinger, K. A., Lindsey, J. C., Crosier, S., Richardson, S., Goddard, J., Hill, R. M., Castle, J., Grabovska, Y., Hacking, J., Pizer, B., Wharton, S. B., Jacques, T. S., Joshi, A., Bailey, S., & Clifford, S. C. (2022). Medulloblastoma group 3 and 4 tumors comprise a clinically and biologically significant expression continuum reflecting human cerebellar development. *Cell Reports*, *40*(5),

111162.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., … Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, *46*(D1), D1074–D1082.

Woods, P. (2022, February 22). *Bladder cancer statistics*. WCRF International. https://www.wcrf.org/cancer-trends/bladder-cancer-statistics/

Wu, C., Jin, X., Tsueng, G., Afrasiabi, C., & Su, A. I. (2016). BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic Acids Research*, *44*(D1), D313–D316.

Xi, N. M., & Li, J. J. (2021). Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. *Cell Systems*, *12*(2), 176–194.e6.

Yang, S., Corbett, S. E., Koga, Y., Wang, Z., Johnson, W. E., Yajima, M., & Campbell, J. D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology*, *21*(1), 57.

Yang, Y., & Yang, L. (2023). Somatic structural variation signatures in pediatric brain tumors. *Cell Reports*, *42*(10), 113276.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., … Flicek, P. (2020). Ensembl 2020. *Nucleic Acids Research*, *48*(D1), D682–D688.

Ye, H., Meehan, J., Tong, W., & Hong, H. (2015). Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine. *Pharmaceutics*, *7*(4), 523–541.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* , *25*(21), 2865–2871.

Young, M. D., & Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, *9*(12). https://doi.org/10.1093/gigascience/giaa151

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology*, *16*(5), 284–287.

Zhang, X., Lv, D., Zhang, Y., Liu, Q., & Li, Z. (2016). Clonal evolution of acute myeloid leukemia highlighted by latest genome sequencing studies. *Oncotarget*, *7*(36), 58586–58594.

Zhao, G., Li, K., Li, B., Wang, Z., Fang, Z., Wang, X., Zhang, Y., Luo, T., Zhou, Q., Wang, L.,

Xie, Y., Wang, Y., Chen, Q., Xia, L., Tang, Y., Tang, B., Xia, K., & Li, J. (2020). Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Research*, *48*(D1), D913–D926.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049.

Zverinova, S., & Guryev, V. (2022). Variant calling: Considerations, practices, and developments. *Human Mutation*, *43*(8), 976–985.