Differentially expressed genes from contrasting bulk RNA-seq profiles of ADD3 OE Onda11 versus control, 72 h after transfection. Z-scores of differentially expressed genes (absolute log FC > 0.5 and adjusted P < 0.05) are grouped row-wise according to differential expression sign, with samples hierarchically clustered based on Euclidean similarity.

To investigate the cell-autonomous effects of ADD3 overexpression (OE), we conducted bulk RNA sequencing on GFP+ FACS-sorted Onda-11 cells cotransfected with ADD3 or control plasmids. Count data were regularized and log-transformed using the *rld* built-in DESeq2 function and samples were clustered based on Euclidean distances. Differential expression analysis was performed using DESeq2 using raw counts as input. Differentially expressed genes were identified using a cutoff of absolute $\log_2$ fold change ($\log_2$ FC) ≥ 0.5 and False Discovery Rate (FDR) < 0.05.

2.3.1.7. Basal expression levels of differentially expressed genes upon ADD3 correlates with ADD3 basal expression levels

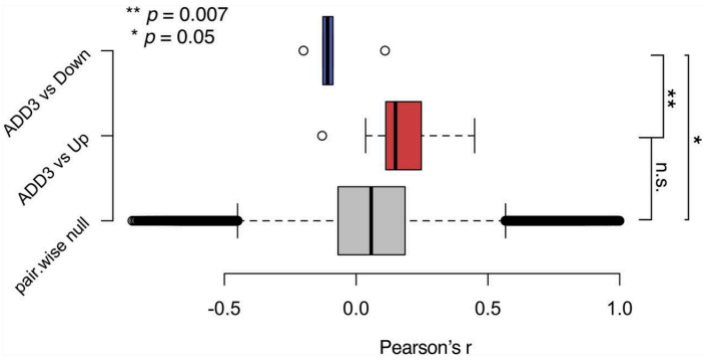


Figure 20 - Quality assessment of differentially expression analysis: Differential gene expression upon ADD3 OE correlates with basal expression patterns in relation to ADD3. Pairwise correlation scores were calculated across basal expression patterns of all possible gene pairs (gray distribution), ADD3 and up-regulated genes in the ADD3 overexpression signature (red distribution), and ADD3 and down-regulated genes in the ADD3 overexpression signature (blue distribution). Error bars, 95% CI; ***P* = 0.007; **P* = 0.05; n.s., not statistically significant; *t* test.

To comprehensively evaluate the outcomes of the Differential Expression Analysis, we employed Cancer Cell Line Encyclopedia (CCLE) profiles - standardized to achieve zero mean and unit variance – across 48 GBM cell lines. We calculated pair-wise correlation scores across all genes, considering the upper triangle of this matrix as a null distribution of scores. Pair-wise Pearson's correlation scores between ADD3 and DEGs were extracted and compared to the null with a t-test. These differentially expressed genes exhibited consistent patterns in other GBM cell lines, showing that upregulated genes are positively correlated and downregulated genes are anti-correlated with ADD3 expression at the basal level, underscoring the robustness of the ADD3 OE signature.

2.3.2. Single-cell analysis of primary GBM tumors show strong association of ADD3 with the OPC-like transcriptional subtype

2.3.2.1. Single-cell RNA expression analysis of primary tumors reveals enrichment of different transcriptional subtypes

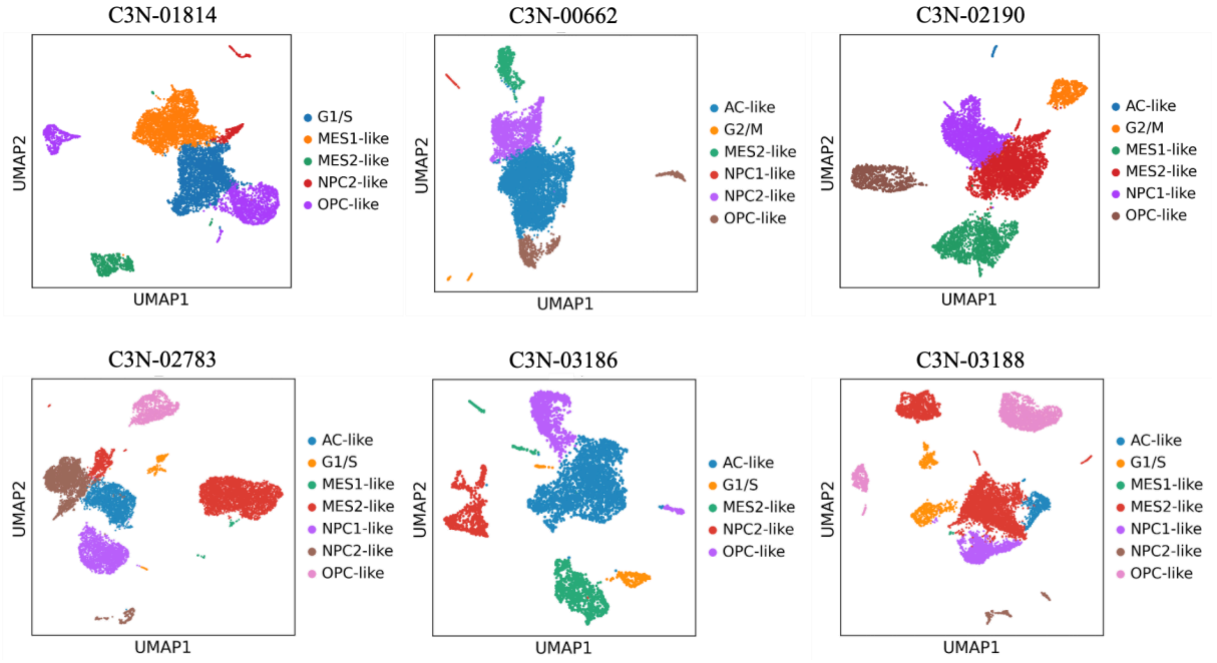


Figure 21 - Leveraging scRNA-seq to investigate GBM heterogeneity: UMAP scatterplots of the six single-cell RNA-seq datasets annotated using transcriptional subtype-specific markers from Neftel et al., 2019.

Given the parallels between neurodevelopment and GBM progression, investigating the cell-type-specific expression of ADD3 and its potential contribution to maintaining stemness and cellular plasticity could uncover novel insights into GBM biology. Here, we leveraged single-cell RNA sequencing (scRNA-seq) datasets and integrative analyses to elucidate the role of ADD3 in GBM, particularly its association with stem-like cellular states and tumor progression. We collected eighteen 10x single cell datasets of primary tumors from the CPTAC-3 project through the GDC data portal. Twelve of the eighteen datasets were excluded from the analysis as they did not meet the baseline quality criteria required for further investigation. For the

six datasets retained (i.e., "C3N-01814", "C3N-00662", "C3N-02783", "C3N-02190", "C3N-03186", "C3N-03188") we applied standard pre-processing procedures for single cell RNA-seq data analysis. We computed the nearest-neighbors distance matrix for each dataset using the first thirty principal components (PCs) and applied the Leiden clustering algorithm. To identify the optimal resolution for clustering, silhouette scores were computed on pre-computed Leiden clusters across a range of resolution parameters. We annotated these clusters based on their transcriptional profiles using prior known gene sets from Neftel et al., 2019 (Neftel et al. 2019) and identified several transcriptionally distinct tumoral cell populations that potentially resemble their normal counterparts.

2.3.2.2. ADD3 expression exhibit localized expression

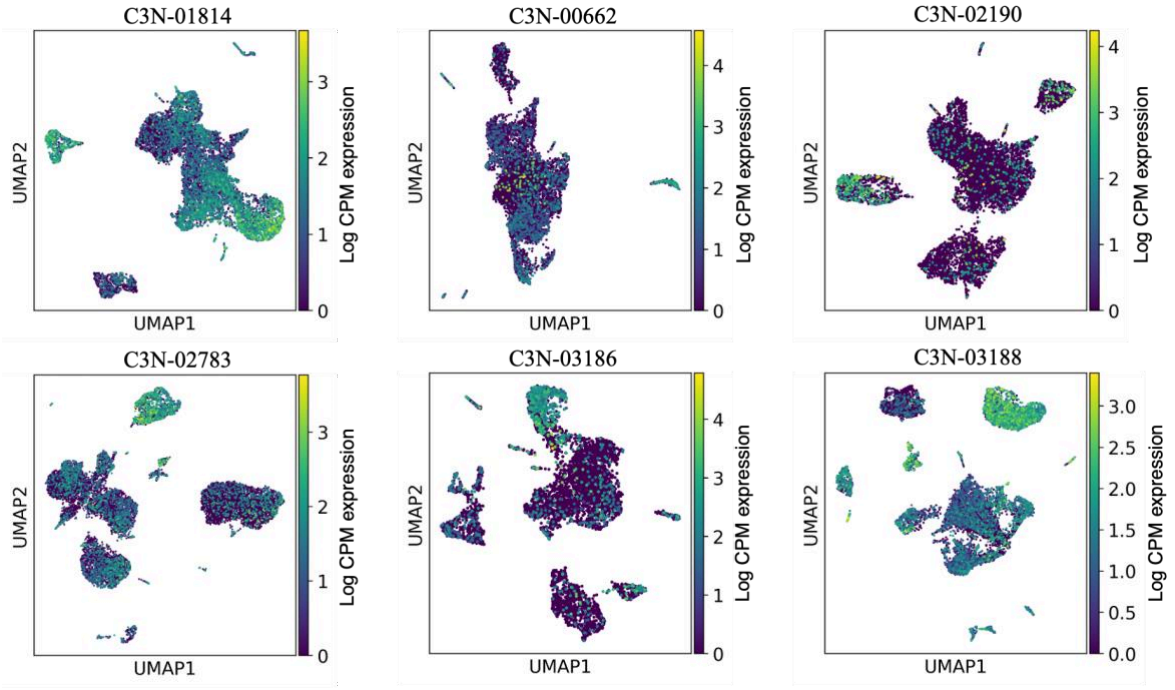
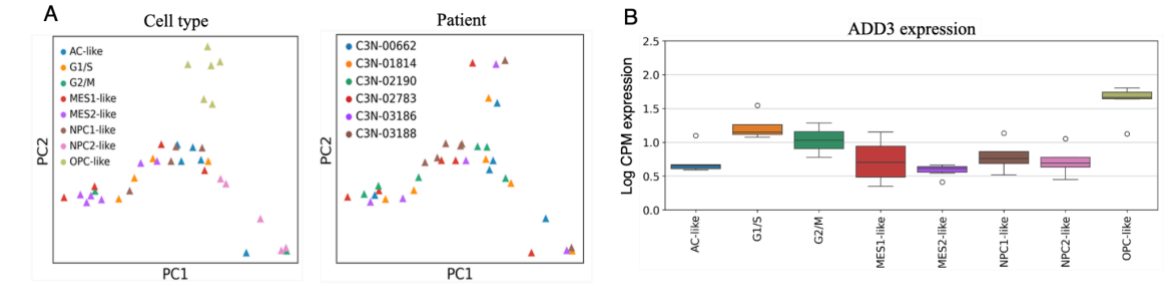


Figure 22 - Visualization of ADD3 expression in single cells across scRNA-seq GBM datasets: UMAP scatterplots of the six single-cell RNA-seq datasets showing ADD3 log CPM expression.

To visualize ADD3 expression across clusters, we plotted its expression levels on UMAP embeddings, highlighting the distribution of ADD3 within the annotated clusters. Visual inspection of the UMAP revealed that ADD3 expression is predominantly localized in the OPC-like signature, indicating its enrichment in this population.

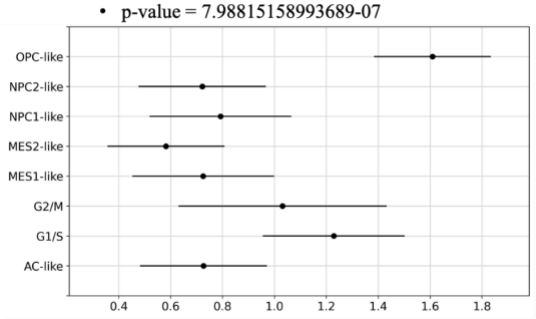
2.3.2.3. Tukey's test results show robust association of ADD3 with the OPC-like signature.



C

ANOVA results:

- F-statistic = 11.5910000188583
- p-value = 7.98815158993689-07



Multiple comparisons between significant pairs

group1	group2	meandiff	P-adj	reject
AC-like	OPC-like	0.8829	0.0	True
G1/S	MES2-like	-0.6466	0.0048	True
MES1-like	OPC-like	0.8839	0.0001	True
MES2-like	OPC-like	1.0271	0.0	True
NPC1-like	OPC-like	0.8169	0.0003	True
NPC2-like	OPC-like	0.8876	0.0	True

Figure 23 - Pseudobulk analysis of single-cell RNA-seq GBM datasets: Principal Component Analysis (PCA) was performed to assess the separation of the main cell states (A, left) and scRNA-seq datasets (A, right) on the two major axes of variation (PCs). Legends in the top left of each PCA scatterplot show the colour code mapping the covariate of interest. (B) Boxplots showing ADD3 log CPM expression (x-axis) at the pseudobulk level across cell states. The ANOVA results are presented (C, left), along with the results of the multiple comparison Tukey's test (C, right), where only the significant pairs are displayed. Results were statistically validated using Family-Wise Error Rate (FWER) correction for p-values, further supporting the reliability of these associations.

We generated pseudobulks for all datasets and annotated cell types. In Figure 23A (left), pseudo-bulks are visualized on the first two Principal Components,

with triangles colored based on their annotated cell types. On the right, the same projection is displayed, but triangles are instead colored by their dataset of origin. The separation of pseudo-bulks appears to be primarily driven by cell type rather than dataset, indicating that the dataset effect is minimized. We generated patient-specific pseudo-bulk transcriptomes for each single-cell cluster by aggregating raw counts by summation. To further assess the relationship between ADD3 expression and cell types, we performed an ANOVA analysis followed by multi-comparison Tukey's HSD test (Figure 23C). Tukey's test revealed six significant associations, with ADD3 expression being significantly higher in OPC-like compared to NPC1-like/NPC2-like, AClike, and MES1-like/MES2-like and in G1/S signature compared to the MES2like. Taken together, these results underscore the robust association of ADD3 expression with the OPC-like signature, suggesting a specific role for ADD3 in this cell state.

2.3.2.4. Associating gene dependency estimates with Differentially Expressed Genes (DEGs) in the OPC-like signature

By leveraging results from large-scale CRISPR-screens and enrichment analysis, we estimated gene dependencies for each transcriptional subtype using its derived upregulated and downregulated genes as query lists against the Genomic Profile of Gene Essentiality (GPGE) ranks (Pellecchia et al. 2023). GPGE scores were calculated by correlating gene essentiality with the expression levels of all messenger RNA (mRNA) across various cancer cell lines (please refer to the "Material and Methods" section).

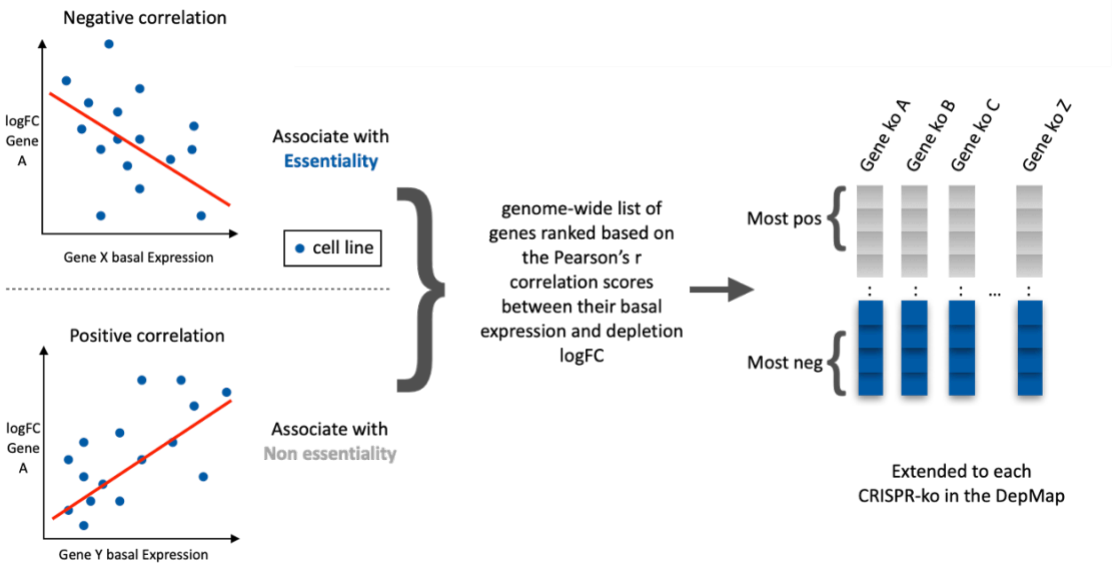


Figure 24 - Building genome-wide signature of gene dependency across ~1,000 cancer cell lines (pan-cancer): In this framework, RNA-seq provides the baseline gene expression data for each cell line, while gene essentiality is assessed based on gene depletion or fitness scores. These scores reflect the importance of a gene for cell survival or proliferation, with negative values indicating greater dependency on the gene and positive values suggesting that the gene is non-essential. Genes whose expression negatively correlates with essentiality scores across cell lines can be considered essential for survival (i.e., higher gene expression corresponds to greater dependency on the gene). In contrast, genes with a positive correlation are considered non-essential (i.e., higher gene expression is linked to reduced dependency on the gene).

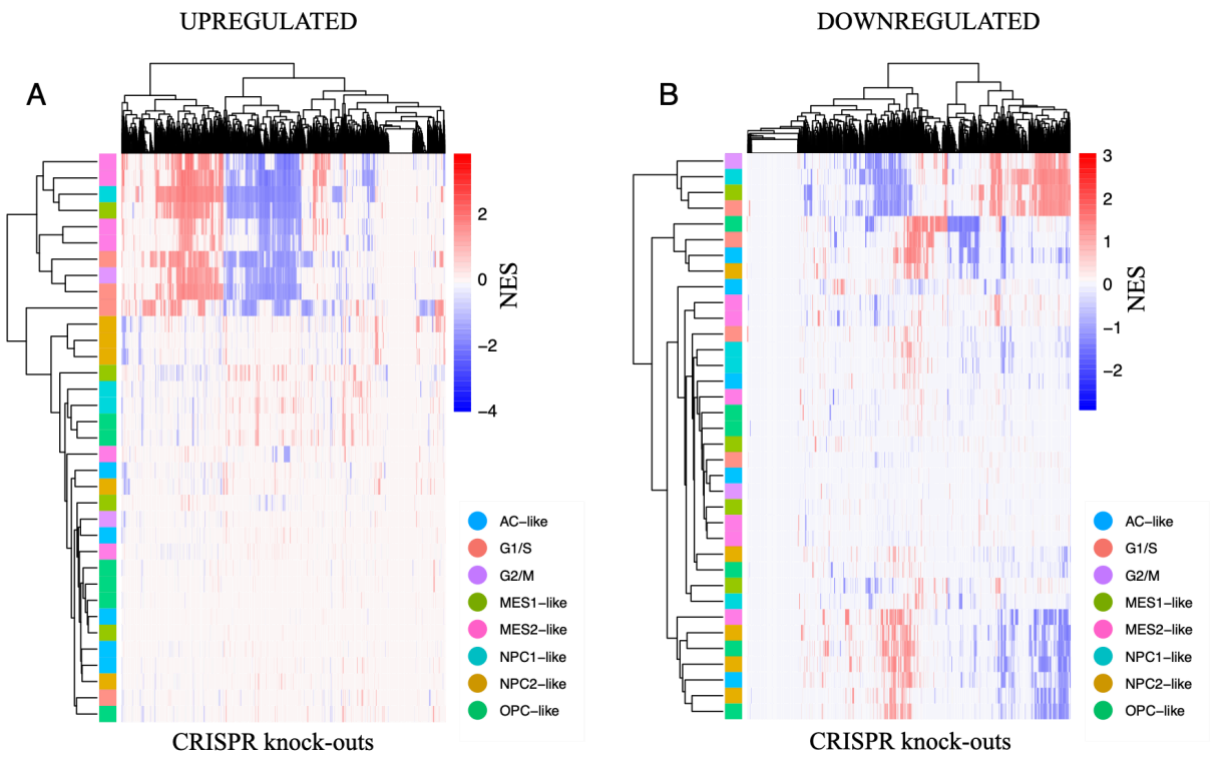


Figure 25 - Heatmaps of estimated essentiality profiles in transcriptional subtypes: Normalized enrichment scores (NES) are shown separately for upregulated (A) and downregulated (B) gene signatures of transcriptional subtypes annotated across the six single-cell datasets used. NES were computed by applying Gene Set Enrichment Analysis (GSEA)(Subramanian et al. 2005) against each Genomic Profile of Gene Essentiality (GPGE) ranked list (Pellecchia et al. 2023). Following (Pellecchia et al. 2023) approach, we assigned a NES equal to 0 if the estimation was predicted not significant (adjusted p value < 0.05). Clustering of both rows (cell types) and columns (gene knockouts from DEPMAP(Tsherniak et al. 2017; Trastulla et al. 2023; Boehm and Golub 2015)) was performed using Euclidean distance, enabling the identification of patterns of genetic dependencies across the different cell types and perturbations.

Since non-essential biomarkers are ranked at the top and essential biomarkers at the bottom of the GPGE lists, a negative normalized enrichment score (NES) indicates that the putative subtype is predicted to be dependent on a gene. Conversely, a positive NES suggests that the subtype is not dependent on the gene. For the purpose of this analysis, we have chosen to focus specifically on the OPC-like subtype, based on our previous findings. This targeted approach allows us to explore the genetic dependencies and

vulnerabilities specific to this subtype, but the methodology can be applied to other subtypes as needed for future studies. By examining the ranked predictions associated with the upregulated genes in the OPC-like signature, we aim to identify key genes that may play a critical role in the OPC-like's dependency.

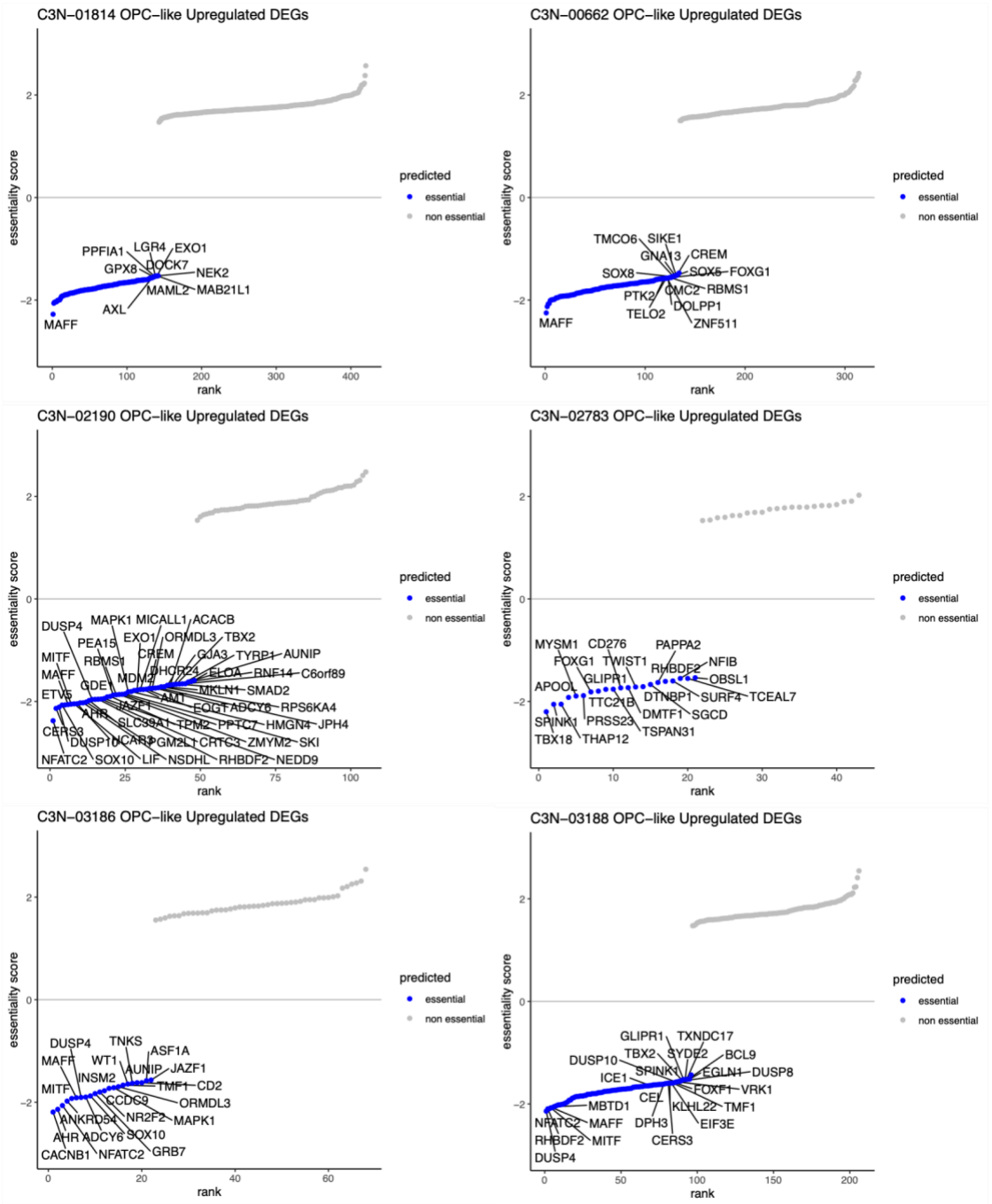


Figure 26 - Visualization of ranked predictions for the upregulated genes in the OPClike signature: By sorting the genes based on their normalized enrichment scores (NES) to which we refer for simplicity as "essentiality score", we can highlight which upregulated genes are most strongly associated with a subtype of interest across all

single-cell datasets. Genes with negative NES values are predicted to be essential for the survival or proliferation of the cell subtype, while those with positive NES values indicate a reduced dependency on the gene.

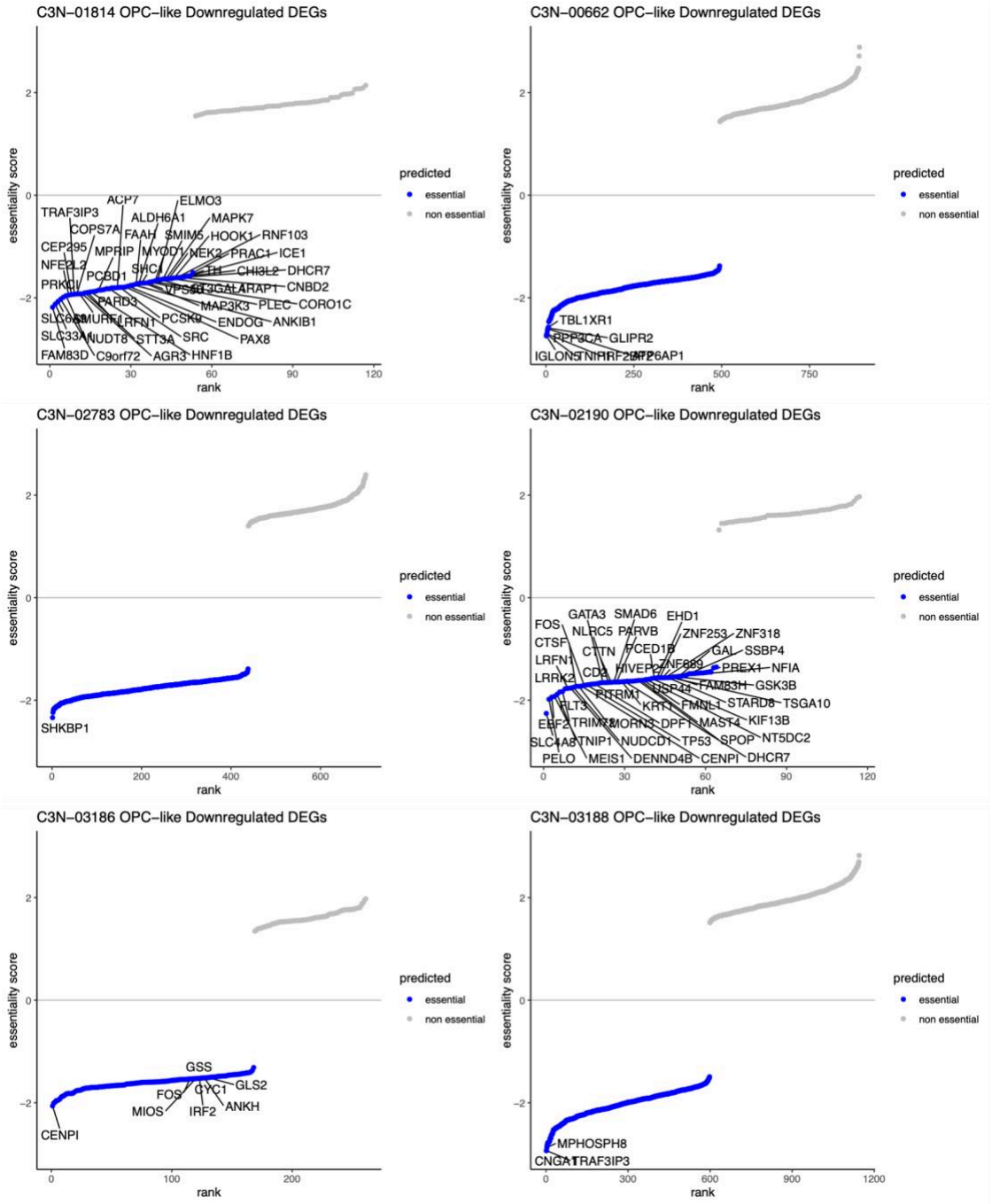


Figure 27 - Visualization of ranked predictions for the downregulated genes in the OPC-like signature. By sorting the genes based on their normalized enrichment scores (NES) to which we refer for simplicity as "essentiality score", we can highlight which downregulated genes are most strongly associated with a subtype of interest across all singlecell datasets. Genes with negative NES values are predicted to be essential for the

survival or proliferation of the cell subtype, while those with positive NES values indicate a reduced dependency on the gene.

2.3.2.5. Characterization of the predicted essentialities associated with an upregulated signature in the OPC-like subtype

Figure 28 - Pathway enrichment analysis for upregulated and predicted essential genes: Bar plots depicting the Normalized Enrichment Scores (NES) for enriched pathways. Positive NES values indicate the enrichment of upregulated genes that are

predicted to be essential for the specified pathways. The color map represents the significance of the association, with an adjusted p-value < 0.05.

2.3.2.6. Survival analysis of upregulated genes and predicted essential genes aids in prioritizing potential gene targets

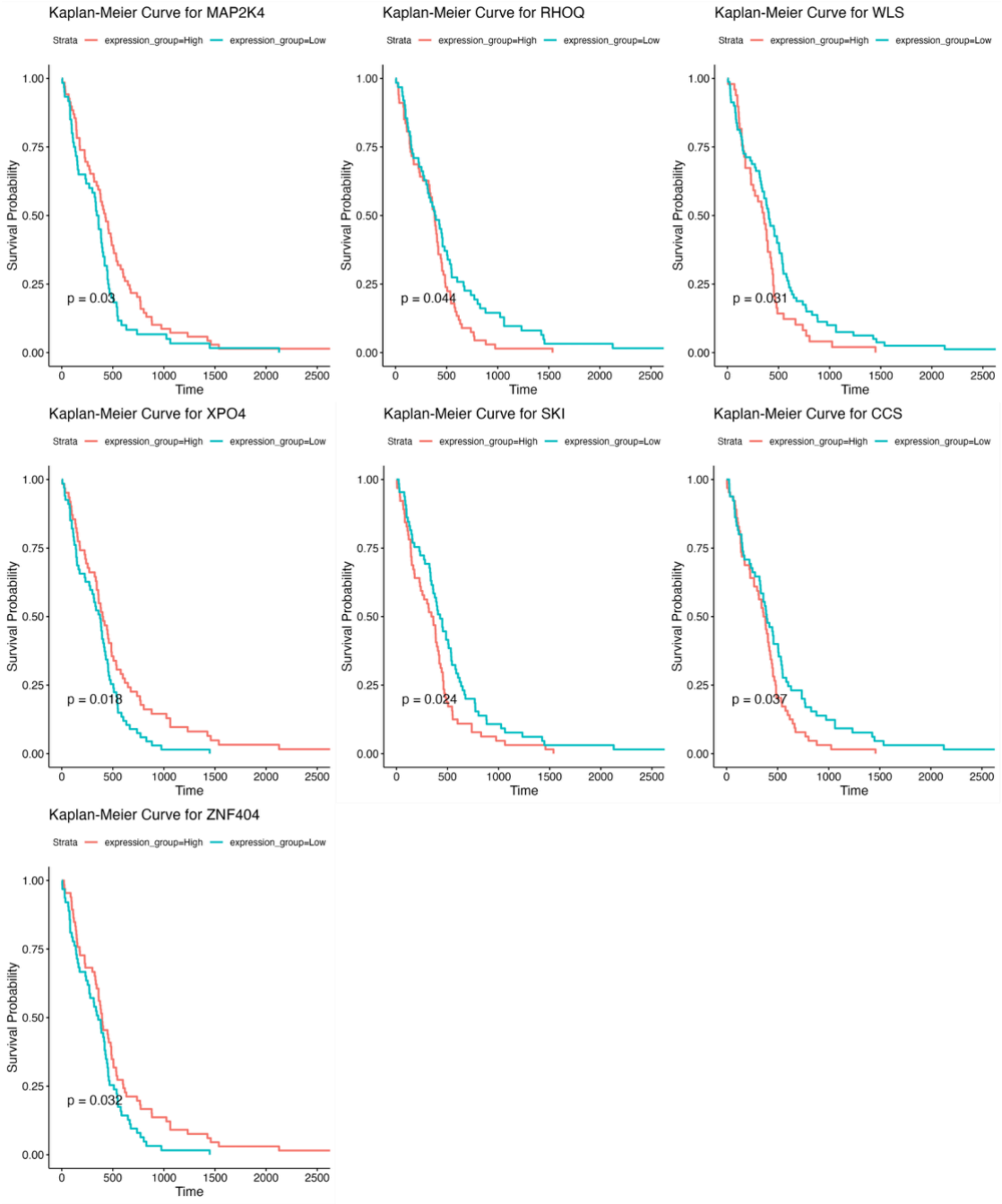


Figure 29 - Kaplan-Meyers curves: To prioritize genes relevant to glioblastoma progression, we performed Kaplan–Meier survival analysis across 166 bulk RNA-seq GBM patient datasets collected from the TCGA-GBM database. We stratified patients into high- and low-expression groups for each gene based on median expression values and assessed differences in overall survival using the log-rank test. This analysis identified a set of genes (i.e., XPO4, SKI, CCS, ZNF404, WLS, RHOQ, and MAP2K) that are not only

upregulated and predicted essential within the OPC-like subtype but also significantly associated with shorter patient survival ($p < 0.05$ for all comparisons).

2.3.2.7. Characterization of the predicted essentialities associated with an downregulated signature in the OPC-like subtype

Figure 30 - Pathway enrichment analysis for downregulated and predicted essential genes: Bar plots depicting the Normalized Enrichment Scores (NES) for enriched pathways. Positive NES values indicate the enrichment of downregulated genes that are predicted to be essential for the specified pathways. The color map represents the significance of the association, with an adjusted p-value < 0.05.

2.3.2.8. Downregulated and predicted essential genes are enriched in processes involved in oligodendrogenesis

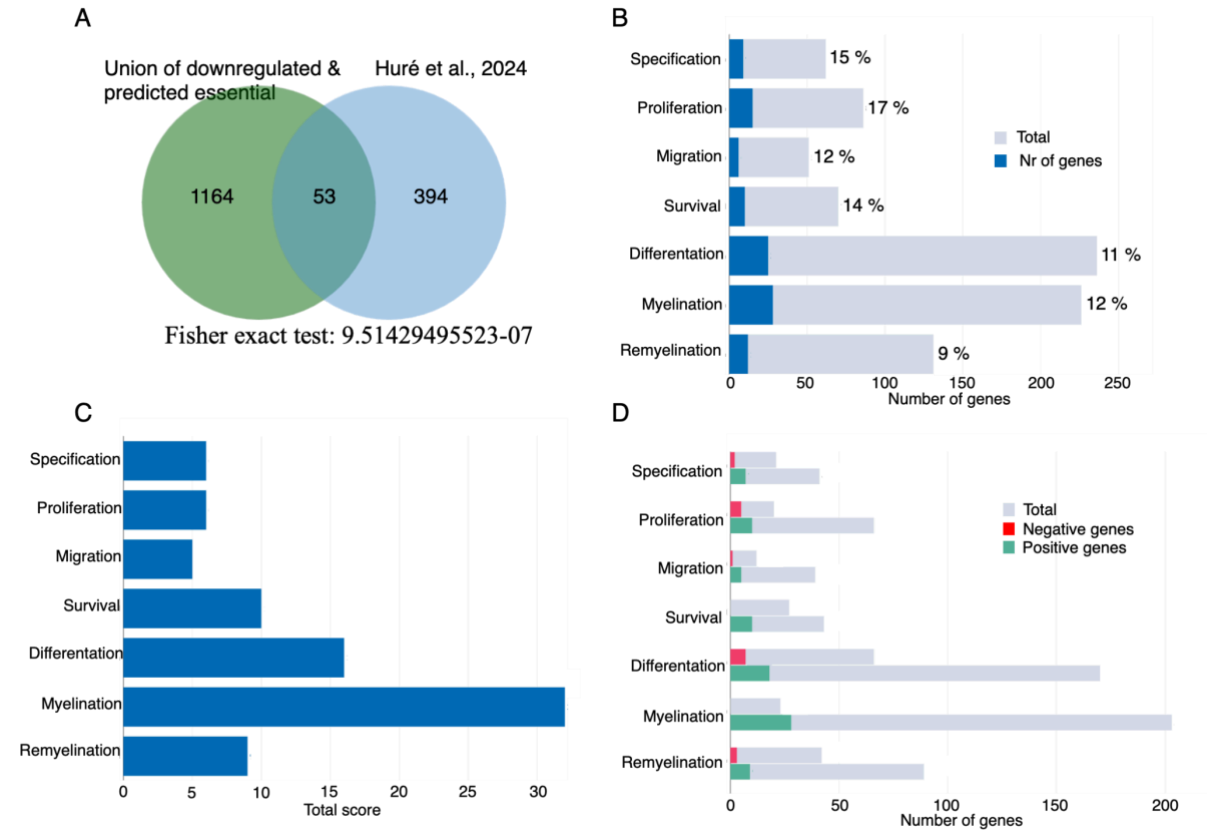


Figure 31 - Results from OligoScore: To evaluate the impact of the downregulated predicted essential in key biological processes involved in oligodendrogenesis and (re)myelination we have used the OligoScore web application (Huré et al. 2024). (A) Intersection between a list of 1217 downregulated genes in the OPC-like signature and 454 genes from Huré et al., resulting in shared genes. Each gene was assigned an activity score from 1 to 3 (low, medium, high) in each process The total score reflects the cumulative number of genes that contribute to promoting a given process overall (C). In (D) scores are defined in "Positive genes" values, indicating promotion, and "Negative genes" values indicating inhibition (Figure D), based on the severity of gain- or loss-of-function phenotypes.

The OPC-like signature, characterized by the expression profile resembling oligodendrocyte precursor cells (OPCs), shares several hallmarks with normal oligodendrocyte function, including processes such as myelination, remyelination, and cell migration. Given that these processes are essential for the proper function and maintenance of the CNS, we hypothesized that genes

predicted to be downregulated and essential in this transcriptional subtype could be specifically involved in these critical functions. Indeed, myelination and re-myelination are fundamental processes in oligodendrocytes, and disturbances in these pathways could play a significant role in GBM progression. Driven by this, we collected the union of the downregulated and predicted essential genes across the six single-cell datasets and leveraged an expert curation of 454 genes previously involved in oligodendroglial biology (Huré et al. 2024). By leveraging this curated gene set, we identified a striking intersection of 53 genes (Figure 30A), which were further annotated using a knowledge-driven scoring procedure implemented by Huré et al., 2024. Among the key findings, myelination was the process that received the highest score (Figure 30B,C,D), suggesting that impairments in this process might be playing a role in the progression of GBM. Interestingly, among the 53 identified genes, several are implicated in the ERBB and EGFR signaling pathways. EGFR is altered in almost 50% of GBM tumors(Zahonero and Sánchez-Gómez 2014), and it currently represents one among the most promising therapeutic targets for GBM treating. In fact, it has been associated with several distinct steps in tumorigenesis, from tumor initiation to tumor growth and survival, and also with the regulation of cell migration and angiogenesis. ERBB receptors (especially ERBB2, ERBB3, and ERBB1) are critical for OPC proliferation, migration and differentation. For example, the ERBB3/4–NRG1 axis plays a key role in regulating maturation and myelin sheath formation, particularly in the peripheral nervous system, while also being active in the CNS.(Mei and Nave 2014; Mansour, Khattab, and El-Khatib 2023; Loos et al. 2016). Overall, the downregulation suggests that these tumor cells lose features of mature oligodendrocytes, becoming more like progenitors or stem-like cells. Moreover, dedifferentiation can be typically associated with higher proliferative potential, increased plasticity and therapy resistance(PérezGonzález, Bévant, and Blanpain 2023). Such a state not only supports tumor progression but also hinders normal CNS

integration and repair. Emerging therapeutic strategies(Suvà et al. 2014) aim to reverse this transcriptional program by promoting glial differentiation through epigenetic modulation, retinoid or thyroid hormone signaling, or manipulation of pathways such as Wnt and BMP. While these "pro-differentiation" approaches show promise in preclinical models, their translation into effective and safe therapies in glioblastoma remains an ongoing challenge.

2.3.2.9. Associating drug response estimates with differentially expressed genes of annotated transcriptional subtypes

We leveraged data from the GDSC database and basal mRNA expression level to derive GPDS ranked lists. These were derived by correlating the basal RNAseq expression levels with the in vitro response to small-molecule compounds across a panel of cancer cell lines for which drug potency data were available.

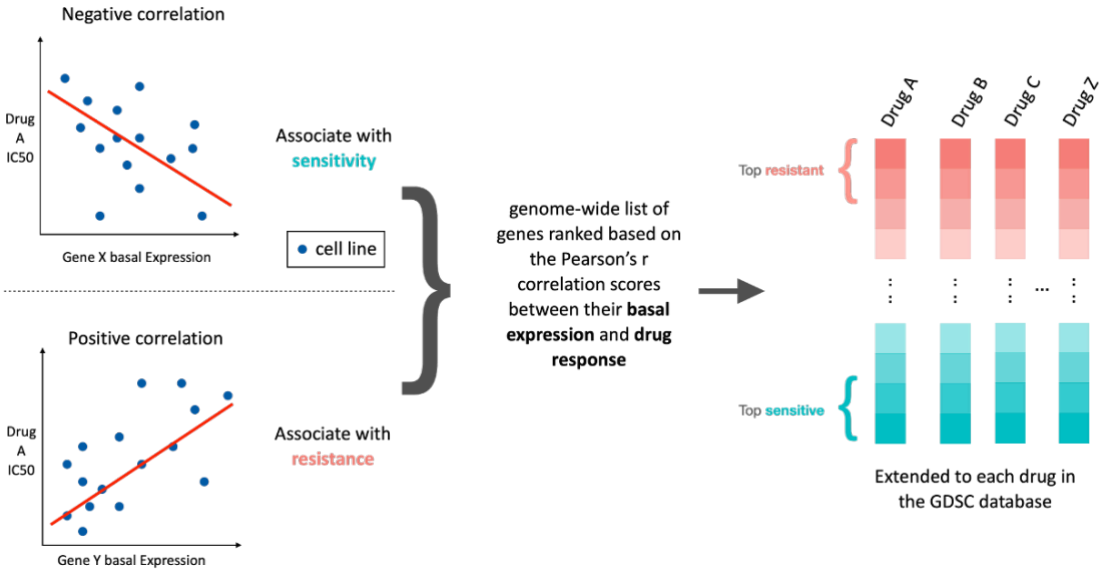


Figure 32 - Building genome-wide signature of drug sensitivity/resistance across ~1,000 cancer cell lines (pan-cancer): In this framework, RNA-seq captures the baseline transcriptional state of each cell line, while drug response is assessed using the half maximal inhibitory concentration (IC_{50}) obtained from dose–response experiments. The IC_{50} value reflects the drug concentration required to inhibit cell growth by 50% after 72 hours of treatment: lower

IC_{50} values indicate greater sensitivity, whereas higher values suggest resistance. Accordingly, genes whose expression positively correlates with IC_{50} are considered predictive of drug resistance, as higher expression is associated with reduced drug efficacy. In contrast, genes that negatively correlate with IC_{50} are viewed as predictive of drug sensitivity, meaning their elevated expression enhances the drug's inhibitory effect.

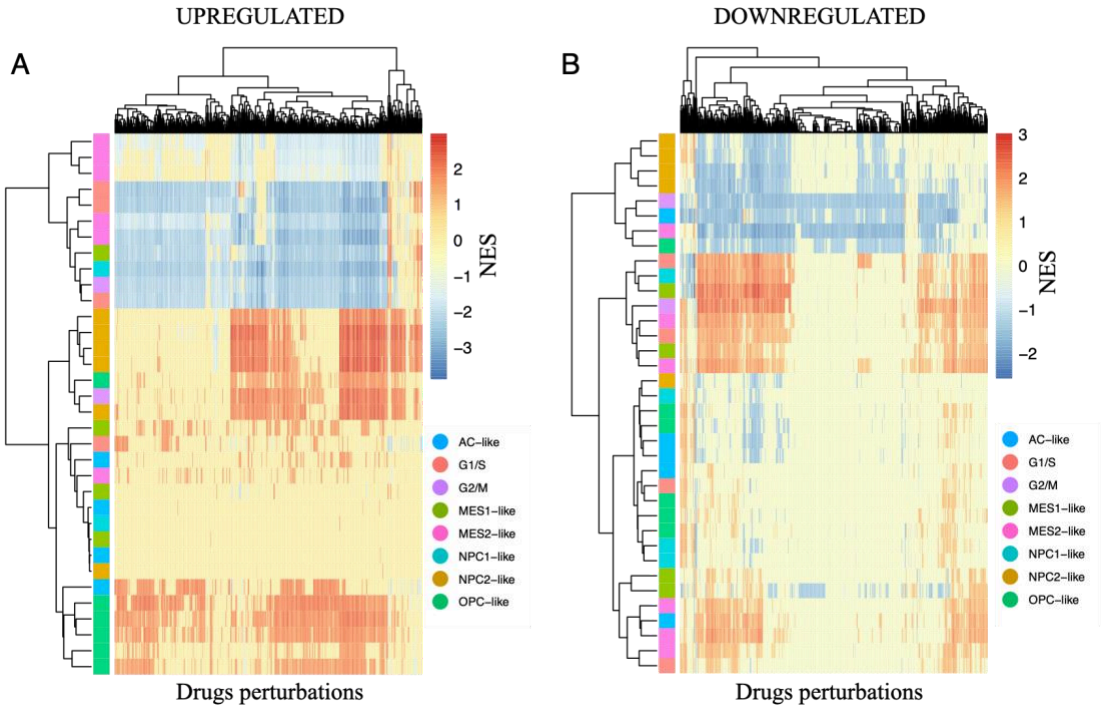


Figure 33 - Heatmaps of estimated drug response in transcriptional subtypes: Normalized enrichment scores (NES) are shown separately for upregulated (A) and downregulated (B) gene signatures of transcriptional subtypes annotated across the six single-cell datasets used. NES were computed by applying Gene Set Enrichment Analysis (GSEA)(Subramanian et al. 2005) against each Genomic Profiles of Drug Sensitivity (GPDS) ranked list (please refer to the "Material and Methods" section)(Pellecchia et al. 2023). Following (Pellecchia et al. 2023) approach, we assigned a NES equal to 0 if the estimation was predicted not significant (adjusted p-value < 0.05). Clustering of both rows (cell types) and columns (drugs perturbation from the Genomic of Drug Sensitivity in Cancer (GDSC) database(Yang et al. 2013)) was performed using Euclidean distance, enabling the identification of patterns of genetic dependencies across the different cell types and perturbations.

Each value in the heatmap indicates the NES for a given cell type and drug perturbation, with negative NES values suggesting that the cell type is

sensitive to the corresponding drug, and positive values indicating that the cell type is resistant to the drug. We computed the NES score by considering upregulated and downregulated genes separately.

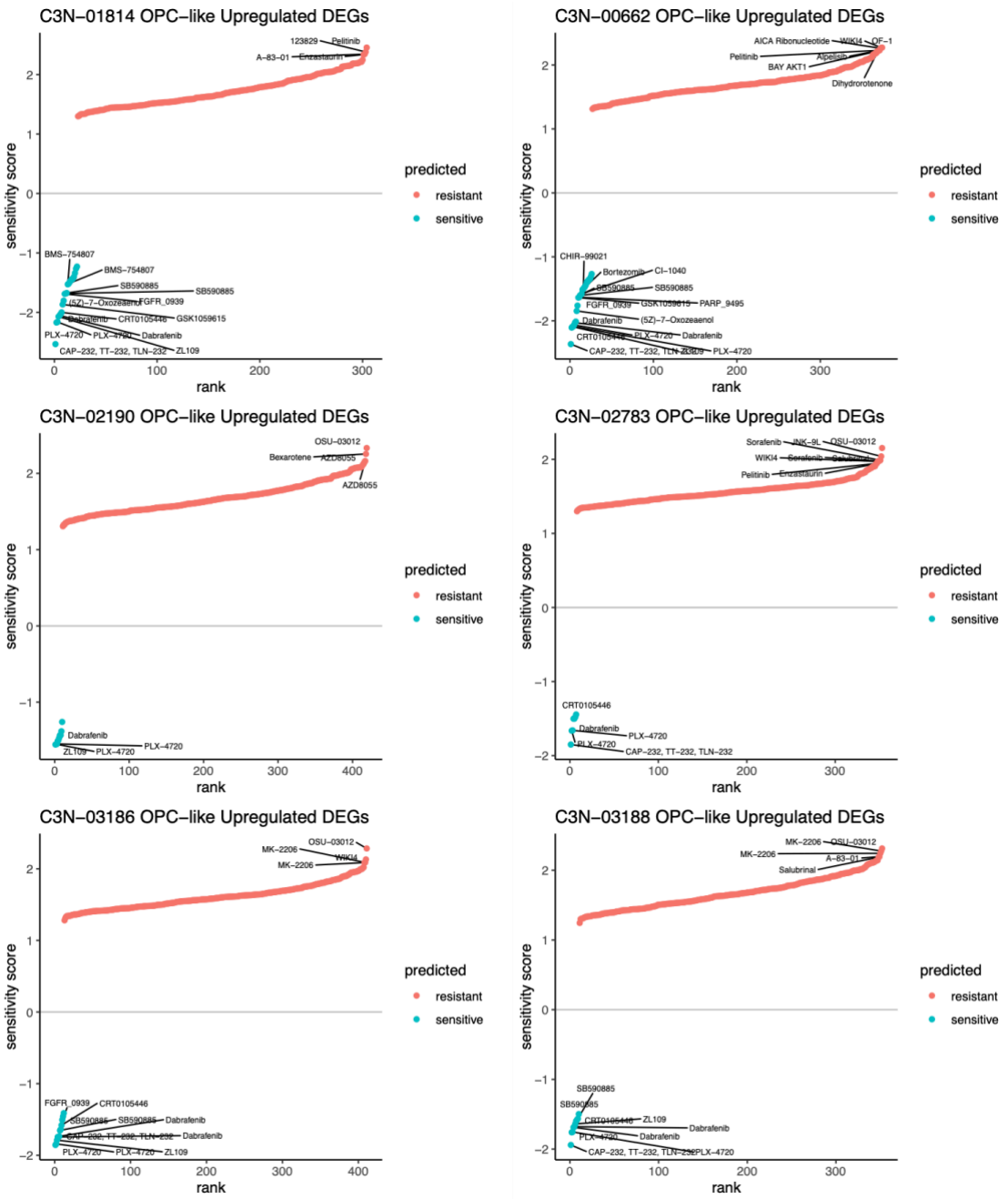


Figure 34 - Visualization of ranked predictions for the upregulated genes in the OPClike signature. By sorting the genes based on their normalized enrichment scores (NES) to which we refer for simplicity as "sensitivity score", we can highlight which upregulated genes are most strongly associated with a subtype of interest across all single-cell datasets. Genes with negative NES values are predicted to be sensitive for the survival or proliferation of the cell subtype, while those with positive NES values indicate resistance dependency on the gene.

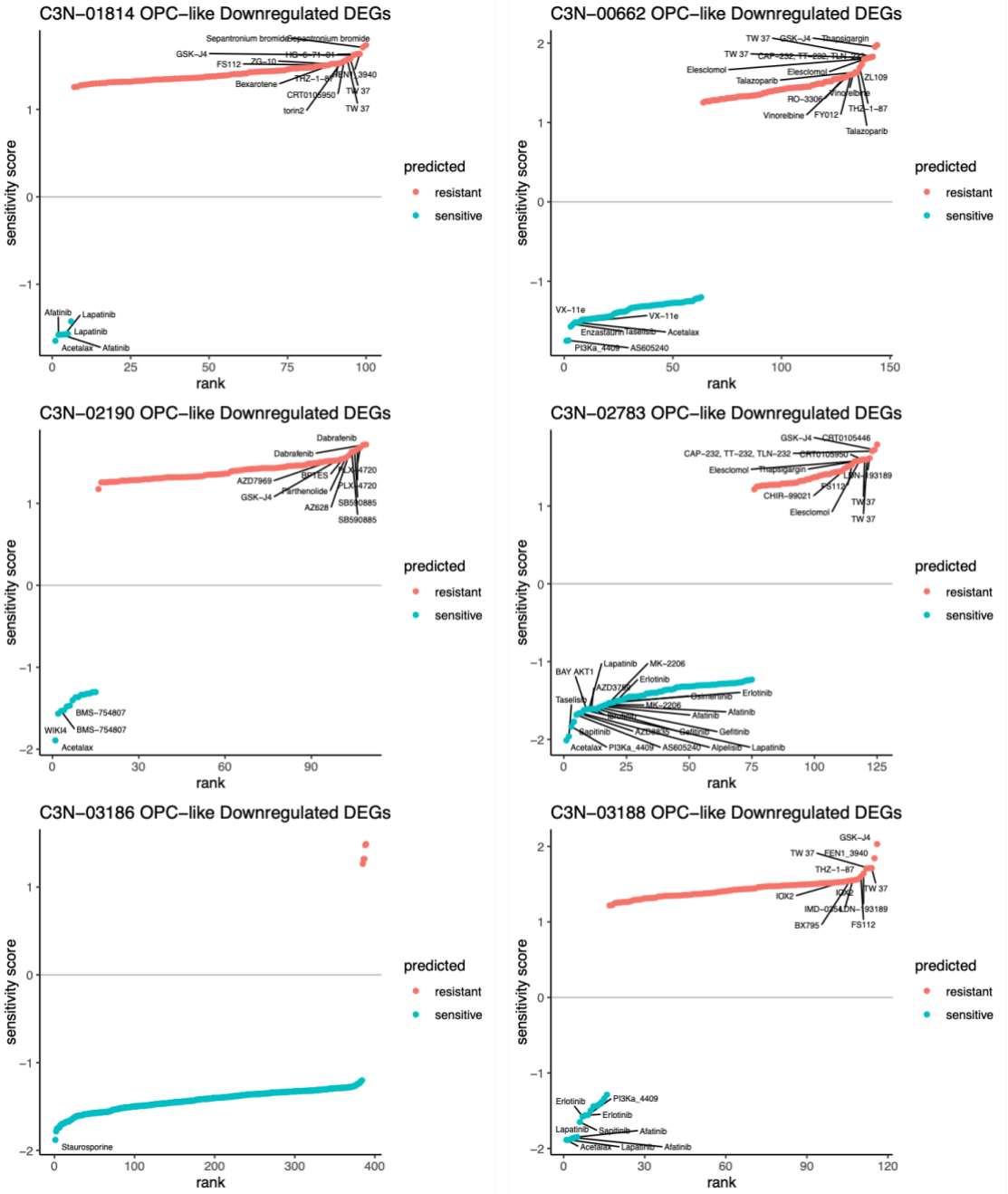


Figure 35 - Visualization of ranked predictions for the downregulated genes in the OPC-like signature. By sorting the genes based on their normalized enrichment scores (NES) to which we refer for simplicity as "sensitivity score", we can highlight which downregulated genes are most strongly associated with a subtype of interest across all singlecell datasets. Genes with negative NES values are predicted to be sensitive for the survival or proliferation of the cell subtype, while those with positive NES values indicate resistance dependency on the gene.

2.3.2.10. The OPC-like downregulated genes are predicted to be associated to sensitivity to ERBB2, EFGR, ERK2 and PI3K inhibitors

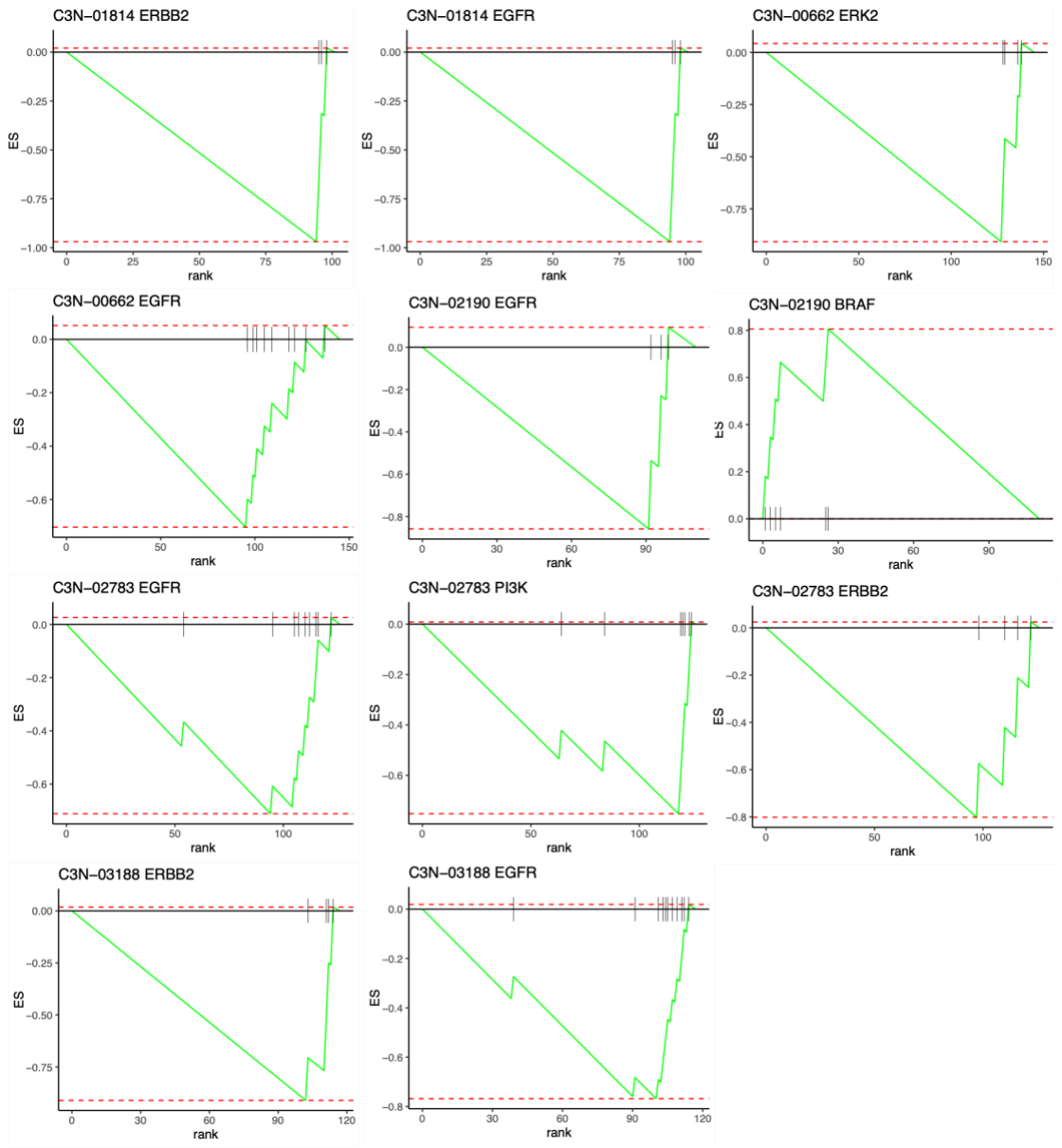


Figure 36 - Targets enrichment analysis: We performed gene set enrichment analysis (GSEA) using curated drug-target gene sets from the Genomics of Drug Sensitivity in Cancer (GDSC) database to identify OPC-like-specific pharmacological vulnerabilities across six glioblastoma patients. The OPC-like transcriptional signature in samples C3N00662, C3N-02190, C3N-02783, and C3N-03188 was significantly associated with sensitivity to EGFR inhibitors (adjusted p-value < 0.05). Notably, samples C3N-02783 and C3N-03188 also showed significant enrichment for sensitivity to ERBB and PI3K inhibitors (adjusted p-value < 0.05). In contrast, the OPC-like signature in sample C3N-02190 was significantly associated with resistance to BRAF inhibitors (adjusted p-value < 0.05).

We assessed whether genes downregulated in OPC-like tumors were enriched in drugs targeting specific genes or pathways. Strikingly, the OPC-like

downregulated genes were significantly enriched in drug sets associated with increased sensitivity to ERBB2, EGFR, and ERK pathway inhibitors, yet with resistance to BRAF inhibitors. This pattern suggests that, although ERBB2, EGFR, and ERK2 are transcriptionally suppressed in this subtype, the OPClike expression profile nonetheless mirrors that of cancer cell lines that respond favorably to inhibitors targeting these pathways. At the same time, the resistance to BRAF inhibition implies that MAPK signaling in OPC-like tumors may be activated upstream or independently of BRAF, potentially via alternative RTKs such as EGFR or PDGFRA. This apparent contradiction, transcriptional downregulation combined with predicted essentiality and drug sensitivity, suggests a phenomenon consistent with non-oncogene addiction, in which tumor cells remain dependent on the basal or residual activity of critical signaling pathways despite their low expression. Another possibility is that transcriptional repression does not equate to loss of function at the protein level, due to post-transcriptional regulation, pathway compensation, or feedback loops that sustain MAPK signaling output. Alternatively, the expression pattern itself may serve as a biomarker of vulnerability, capturing a broader molecular context associated with sensitivity to RTK or MEK inhibition. Taken together, these findings highlight the complex relationship between gene expression, pathway dependence, and drug response in MAPK, and point to potential therapeutic opportunities targeting MAPK signaling in OPC-like tumors, despite their atypical molecular signature.

3. Discussion

Cancers exhibit remarkable heterogeneity, manifesting through diverse genetic mutations, molecular pathways, and phenotypic traits. This contributes to their complexity and adaptability. Among brain cancers, Glioblastoma (GBM) is

characterized by pronounced heterogeneity, creating significant challenges in identifying effective therapeutic vulnerabilities. This heterogeneity underscores the urgent need for enhanced predictive biomarkers that account for GBM's diverse cellular states, as well as novel therapeutic targets associated with these biomarkers. This prompted us to think about data-driven integration strategies based on the use of DEPMAP, a comprehensive collection of cancer vulnerabilities systematically assembled employing pharmacological and genome-wide CRISPR-Cas9 fitness screens, across human cancers. These data served as the mainstay, guiding the analysis and forming the core of the methodologies employed throughout this thesis. Therefore, we have introduced CRISPR-Cas9 fitness screens and their pivotal role in identifying genetic dependencies across CCLs. A starting point for discussion is the usability of CCLs as primary source of investigation to map key phenotypes from *in vitro* to patients' tumours. On this note, DEPMAP offers the advantage of coupling molecular profiles of large panels of CCLs with functional and/or viability data, with scale and annotation needed for machine learning research in multimodal data analysis and integration. However, it is important to acknowledge the limitations associated with these models. Indeed, CCLs lack key components of the TME that surrounds patients' cancer cells *in vivo*, and therefore are considered overly simplistic models when compared to patient-derived xenografts (PDxs), genetic engineered mice (GEMM) or tumor organoids, as extensively supported in Peng et al. (Peng et al. 2021). Nevertheless, what makes CCLs an indispensable tool is their unlimited availability, ease of manipulation and scalability for HTS platforms. CRISPR screens are relatively straightforward when performed on CCLs compared to tumor samples. As such, tumor heterogeneity poses a significant challenge, as tumor cells often exhibit variability in their susceptibility to transfection or viral infection, making efficient manipulation difficult. In contrast, CCLs are typically more homogeneous and are better adapted to standardized experimental conditions,

which facilitates more consistent and reproducible readouts. Nonetheless, the CRISPR data generated from CCLs must be carefully assessed and properly calibrated before use. As shown, we addressed CRISPR screens' reproducibility and quality assessment, and CRISPR bias correction in two separate works. In the first work, we designed and implemented the HT29 benchmark R package for assessing the reliability of a newly established experimental CRISPR-Cas9 screening pipeline(Iannuzzi et al. 2024). We have provided accompanying high-quality reference data from screening the HT-29 cell line multiple times, and intuitive quality control metrics to assess the userprovided screen for cross-replicate reproducibility, similarity with our reference data set, and reliability in detecting known essential and nonessential genes. As highlighted in Iannuzzi et al. (Iannuzzi et al. 2024), the calibration screen requires adherence to the experimental settings outlined in Behan et al(Behan et al. 2019). Specifically, we have constrained experimentalists to using the Human Improved Genome-wide Knockout CRISPR sgRNA library (commonly known as the Sanger library), although other libraries are also available for use. While an integrated, high-quality consensus across sgRNA libraries could potentially serve as a universal reference for such evaluation, our approach offers a more feasible and standardized method to test for experimental reproducibility. Behan et al. successfully transfected the Sanger library to 339 cell lines and identified a set of 838 most informative sgRNAs which were ideal for quality control assessment and baseline filtering. These same sgRNAs have been implemented in our package and serve the same purpose, providing a low-level quality assessment tool for users. In addition, the HT-29 colorectal cancer cell line is a widely characterized and established model. Its adaptability to standardized cell culture protocols, coupled with its ease of growth and manipulation, makes it an ideal reference for such largescale experimental studies. In the second work, we benchmarked eight computational methods, rigorously evaluating their ability to reduce both CN

and proximity bias in the two largest publicly available cell-line-based CRISPR-Cas9 screens to date(Vinceti et al. 2024). The CN bias(Munoz et al. 2016; Gonçalves et al. 2019), arises when genes reside in amplified genomic regions. In such cases, Cas9-induced double-strand breaks occur at multiple sites within the amplified region, leading to excessive DNA damage and subsequent cell death, independent of the gene's actual essentiality. This phenomenon skews the results of CRISPR screens, as genes within these regions tend to exhibit disproportionately low fold changes. The proximity bias(Iorio et al. 2018; Lazar et al. 2024) in contrast, depends on the phenomenon where a Cas9-induced double-strand break results in the detachment of an entire chromosome arm. In some cases, the cell may tolerate this chromosomal loss and continue replicating, albeit at a reduced rate. However, this results in a systematic down-representation of sgRNAs targeting genes located on the lost chromosomal fragment, confounding the interpretation of gene essentiality. Numerous algorithms have been proposed to correct biases in CRISPR-Cas9 screening data *in silico*. In this study, we conducted a benchmark analysis of eight of the most recent methods(Gonçalves et al. 2021, 2019; Iorio et al. 2018; Lazar et al. 2024; J. M. Dempster et al. 2021; de Weck et al. 2018; Li et al. 2015, 2014; Vinceti et al. 2023), evaluating their performance in bias correction and assessing their impact on data quality across different use cases. While prior studies have benchmarked common algorithms for analyzing pooled CRISPR screens(Bodapati et al. 2020), our analysis is the first to thoroughly assess recently developed methods specifically designed to address biases in CRISPR-Cas9 data, particularly those associated with established (Aguirre et al. 2016) or recently identified (Iorio et al. 2018; Lazar et al. 2024) structural features of targeted genomic regions. Among the tested methods, we found that Chronos exhibited lower effectiveness in correcting proximity bias. In contrast, CCR, GAM, Crispy, and particularly AC-Chronos demonstrated superior performance in addressing CN bias. Interestingly, we observed no

clear distinction in performance between supervised methods (those requiring CN or other omics data) and unsupervised methods. One possible explanation for this is that CRISPR-Cas9 screening data are influenced by multiple sources of bias, and supervised methods are often tailored to correct only one (i.e., CN bias) or two biases (i.e., CN and proximity biases), potentially overlooking other sources of error, such as off-target effects or variability in cellular responses. However, we observed that AC-Chronos emerged as the most effective method for addressing both CN and proximity biases. Based on our findings, AC-Chronos stands out as the preferred choice for correcting both CN and proximity biases, particularly for its ability to minimize the impact on data quality when processing multiple screens with available CN data. Additionally, AC-Chronos enables the identification of a broader range of clinically relevant biomarker/dependency associations. However, CCR also performs well in correcting CN and proximity biases, and its ability to operate on a single-screen basis without requiring additional omics data makes it an ideal option when data availability is limited. In contrast, the datasets corrected using AC-Chronos and Chronos, which borrow information across screens, showed a lower impact on data quality, allowing us to more accurately recapitulate known essential and non-essential genes. Ultimately, no single method emerged as the clear winner across all analyses. Instead, researchers should carefully select the method that best aligns with their specific needs and use cases. Our analyses can also be applied to future CRISPR-Cas9 datasets, enabling a more comprehensive evaluation of each method's performance in different contexts.

In the second part of the thesis, we firstly developed a data-driven strategy to integrate molecular with functional/ viability data from the DEPMAP with the aim of prioritizing gene targets that regulate GSCs morphology and survival(Barelli et al. 2025). As such, we identified a new potential target, ADD3, which is essential for Onda-11 GBM cell line viability. In this collaborative effort, we revealed that ADD3 is associated with GSCs'

morphology and transcription. In addition, we were able to identify a robust association of ADD3 expression with the OPC-like signature, which was previously described by others (Neftel et al. 2019), suggesting a specific role for ADD3 in this transcriptional subtype. Although not strictly a GSC subtype but rather a more differentiated one, the connection between ADD3 expression and the OPC-like signature offers valuable insights. Recent studies have observed that a loss of myelin sheaths around nerve fibers can influence tumor behavior. For instance, in head and neck cancers, demyelination has been associated with the expression of neurodegenerative markers such as those found in Parkinson's and Alzheimer's diseases(Zirngibl et al. 2022; Lee and Kim 2024; Tan et al. 2024). This suggests that demyelination may not only be a consequence of tumor growth but could also actively contribute to tumor progression by altering the neural microenvironment and promoting tumor cell invasion along nerve pathways. Indeed, demyelination can occur as a result of tumor compression on surrounding tissue, as well as vascular disruptions. Additionally, the administration of whole-brain or focal ionizing radiation further exacerbates demyelination within the tumor and adjacent tissues (Lampert and Davis 1964). These demyelination-induced effects may vary across individual tumor patients due to genetic differences and variable responses to treatments. It is crucial, therefore, to account for these individual differences in order to develop personalized treatment strategies and identify predictive biomarkers linked to specific genotypic and phenotypic characteristics. Given these observations, it can be speculated that the association between ADD3 and oligodendrocyte biology in cancer may offer mechanistic insights into these disruptions, potentially unveiling new therapeutic strategies for both GBM and myelin-related pathologies. The expression of ADD3 in the OPC-like subtype, despite the concurrent downregulation of EGFR, ERBB2, and the oligodendrocyte signature, may reflect a compensatory or alternative mechanism supporting cytoskeletal organization and cellular motility. This would explain the constitutive

expression of ADD3 in a demyelinated microenvironment. Indeed, ADD3 is involved in membrane stability, cell–cell junctions, and motility functions that are often co-opted by tumor cells to promote invasion and plasticity. Overall, a comprehensive understanding of the ADD3 role and its potential contribution to demyelination in GBM could be essential to improving patient outcomes and informing future therapeutic approaches.

This collaboration has been instrumental in my doctoral thesis, as it formed the foundation for designing a transfer learning framework that aims to integrate single-cell gene expression profiles of CCLs with drug response data from the GDSC database. Specifically, motivated by evidence suggesting CCLs can evolve to form distinct strains, therefore exhibiting transcriptional heterogeneity, we implemented a two-step machine-learning framework that integrates scRNA-seq data from CCLs into the model training process, alongside bulk data. If intra-cell-line heterogeneity is linked to variations in therapeutic responses, these cell lines could serve as valuable integrative models to identify novel drug targets or combination therapies, insights that may be disregarded when relying solely on bulk data. While additional evaluations are required, this preliminary finding is promising and lays a solid foundation for a data-integration strategy that could enhance current state-ofthe-art transfer learning models that aim to adapt their predictions to the heterogeneous nature of primary tumors, as captured at the single-cell level. Another future perspective would be to integrate genetic dependency data into the training phase of such frameworks by modelling *gene expression-gene dependency* relationships, after harmonizing scRNA-seq data with bulk data from CCLs. As tumors evolve and develop resistance to treatment, sub clonal dependencies may emerge that are unique to specific subclones within the tumor. These dependencies may differ from those of the primary tumor, with some genetic vulnerabilities being gained and others lost as the tumor adapts to therapeutic pressure. For example, a subclone might acquire a mutation that makes it resistant to a particular drug, but this same

mutation might render the subclone reliant on a previously non-essential gene. This newly acquired genetic dependency could represent a potential therapeutic target, whereas the bulk of the tumor may not exhibit the same dependency. Moreover, the evolving nature of subclonal dependencies might complicate the drug discovery process, as it introduces dynamic and context-dependent vulnerabilities that are not always captured in traditional drug screening models. Incorporating genetic dependency data would help identify contextspecific genetic vulnerabilities that emerge in resistant subclones and provide mechanistically insights of acquired drug resistance mechanisms and drugs mode of action (MoA). Moreover, it would facilitate the identification of new potential synthetic lethal interactions, enabling a more effective, context-aware prioritization, potentially uncovering combination therapies that could improve treatment outcomes by targeting multiple genetic pathways simultaneously.

4. Material and Methods

4.1. Benchmark Software and Data for Evaluating CRISPR-Cas9 Experimental Pipeline Through the Assessment of a Calibration Screen

4.1.1. Reference data set preprocessing

We quantified and preprocessed post library-transduction and control libraryplasmid sgRNA read counts as described in Behan et al., removing sgRNAs with <30 reads in the library-plasmid and keeping only sgRNAs in common between the two versions of the Sanger libraries. Subsequently, we normalized counts across technical replicates, scaling each sample by the total number of reads. Post-normalization, we computed sgRNA log fold-changes (LFCs) between individual replicate read counts and library-plasmid read counts for each experiment, keeping the technical replicates separated. These preprocessing steps were performed with the ccr.NormfoldChanges function of our previously published *CRISPRcleanR* R package, using default parameters(Iorio et al. 2018). The same preprocessing steps can be now performed through our recently published, user-friendly, interactive web frontend to *CRISPRcleanR*: *CRISPRcleanR*^{WebApp} (publicly accessible at https://crisprcleanr-webapp.fht.org), which does not require any bioinformatics/programming knowledge and can be used via the web browser(Vinceti et al. 2023). Resulting data at all the intermediate preprocessing levels are included in our reference data set (available at: https://score.depmap.sanger.ac.uk/downloads, at https://groupsdashboards.fht.org/iorio/, and on FigShare).

4.1.2. **Example of user provided data**

To compute receiver operating characteristic (ROC) and precision/recall (PrRc) curves, required to perform high-level quality control assessment of CRISPRCas9 screens, we used the HT29R.individualROC function of the *HT29benchmark* R package, which implements the ROC_Curve and PrRc_Curve functions of the *CRISPRcleanR* package (version 2.2.1), which itself implements the roc and coords functions of the pROC open-source R package (version 1.18.0)(Iorio et al. 2018).

4.1.3. **Fitness effect threshold**

Following the approach outlined in Pacini et al., we employed a rank-based method to determine a fitness effect significance threshold for each HT-29 reference screen. This allowed us to identify a set of significantly depleted (or essential) genes, using a fixed false discovery rate (FDR) of 5%, based on their depletion log-fold changes (LFCs). Specifically, for each screen, we first ranked all genes in ascending order of their average depletion LFCs, which were calculated based on the differential abundance of their targeting sgRNAs at the end of the assay compared to the plasmid control. Next, we scanned the ranked list from the most depleted gene to the least, considering the depletion LFC (r) of each encountered gene as a potential threshold. Any gene with a depletion LFC < r was classified as significantly depleted. Among the genes deemed significantly depleted at a candidate threshold (r), we focused only on those that belonged to one of two pre-established sets: essential (E) and nonessential (N) genes. Using these two sets as reference controls (positive and negative, respectively), we computed the positive predictive value (PPV) and, from this, derived the FDR (FDR = 1 − PPV). The fitness-effect significance threshold was selected as the largest r that yielded an FDR ≤ 0.05. This procedure was implemented using the roc and coords functions of the pROC open-source R package (version 1.18.0), within the

HT29R.ROCanalysis and HT29R.FDRconsensus functions of the *HT29benchmark* R package.

4.1.4. **Data visualization**

We used R base graphics plus the following R libraries and packages (listed in alphabetical order), all available on Bioconductor or on The Comprehensive R Archive Network (CRAN) repository: *crayon* version 1.5.1; *enrichPlot* version 1.14.2; *GGally* version 2.1.2; *ggplot2* version 3.3.6; *ggrastr* version 1.0.1; *grid* version 4.1.0; *gridExtra* version 2.3; *gtable* version 0.3.0; *RcolorBrewer* version 1.1.3; *VennDiagram* version 1.7.3; and *vioplot* version 0.3.7.

4.1.5. **Enrichment analysis**

We performed Gene Ontology (GO) enrichment analysis to identify biological processes (BP) overrepresented in the list of HT-29-specific fitness genes. For this analysis, we used the *org.Hs.eg.db* R package (version 3.14.0) to retrieve the gene universe and the *clusterProfiler* R package (version 4.2.2) to perform the enrichment analysis of the HT-29-specific genes.

4.1.6. **Data records**

The entire HT-29 reference data set described here is available at different intermediate levels of preprocessing on the Project Score website https://score.depmap.sanger.ac.uk/downloads, at https://groups-dashboards.fht.org/iorio/, and on FigShare. The main data folder contains four subfolders:

- 00_rawCounts assembled—Containing one tsv file for each HT-29 screen. Each file comprises the control library-plasmid sgRNA counts, as well as 14 days postselection sgRNA counts across technical replicates;

- 01_normalised_and_FCs—Containing Rdata files of normalized counts and depletion LFCs for the six screens, plots of counts' distribution pre- and post normalization, and boxplots showing LFCs' distributions (PDF files);
- 02_lowLev_QC—subdivided in the following four subfolders:
    - FC_distr—LFC distribution plots for each of the six screens, in PDF;
    - FC_Rep_corr—Between-technical replicate correlation plots for each of the six screens, in PDF;
    - PrRc_curves_ind_rep—Plots of technical replicate's PrRc curves quantifying essential/nonessential gene classification performances across the six screens, in PDF;
    - ROC_curves_ind_rep—Plots of technical replicates' ROC curves quantifying essential/nonessential gene classification performances across the six screens, in PDF;
    - 03_HL_QC_Stats—Density plots of depletion LFCs for reference gene sets across the six experiments with quality control values, in PDF.

4.2. A Benchmark of Computational Methods for Correcting Biases of Established and Unknown Origin in CRISPR-Cas9 Screening Data

4.2.1. Pre-processing of cancer dependency datasets

We focused on the two largest genome-wide CRISPR-Cas9 screen datasets available to date: Project Score (release 1), which used the KY library(Tzelepis et al. 2016; Behan et al. 2019), and Project Achilles (release 23Q2), which employed the AVANA library(Doench et al. 2016). The raw read count versions of both datasets were downloaded from the DEPMAP portal (https://depmap.org/portal/download/all/). These datasets had already been pre-processed through the DEPMAP pipeline up to the computation of raw read counts (README.txt, https://depmap.org/portal/download/all/), resulting in 1,018 cell lines and 18,535 genes for the Project Achilles dataset, and 339 cell lines and 18,009 genes for the Project Score dataset. We then followed the procedure outlined below:

- Single-guide RNAs (sgRNAs) and technical replicates that did not meet quality control standards were excluded based on the quality report files available on the DEPMAP web portal. Specifically, sgRNA quality was assessed using the [Lib]GuideMap.csv file, where [Lib] refers to the library used for generating the dataset (either “Avana” or “KY”). The quality of technical replicates for both datasets was evaluated using the AchillesSequenceQCReport.csv file. Low-quality sgRNAs were defined as those with multiple alignments to the same gene or to multiple genes, no alignments, only intergenic alignments, or being the sole passing guide targeting a gene.

Low-quality replicates were defined as those with fewer than 185 mean read counts per guide, a Pearson correlation coefficient < 0.41 with at least one other replicate when examining genes with the highest variance in gene effect across cell lines, and a null-normalized median difference (NNMD) from the LFC > 1.25. The NNMD is calculated as follows:

$$NNMD = \frac{\tilde{\text{E}}\text{G} - \text{N}\tilde{\text{E}}\text{G}}{MAD_{NEG}}$$

where ẼG and NẼG are the median LFC values of the essential and non-essential reference gene sets, and MAD_{NEG} is the median absolute deviation of the non-essential reference genes as computed in Pacini et al. (Pacini et al. 2021). This yielded 971 cell lines and 17,787 for the Project Achilles dataset and 317 cell lines and 17,349 genes for the Project Score dataset.

- For raw read counts, NAs were replaced with 0.

Given that each method has distinct input requirements, we generated the corrected datasets as follows: we ran *MAGeCK MLE* directly on the raw read counts, as it requires a design matrix contrasting the technical replicates for each cell line after library transduction against the reference plasmid DNA counts. For *CRISPRcleanR* and *Crispy,* we used the sgRNA-level raw LFCs and applied the following steps:

- LFCs were calculated by normalizing the raw read counts after library transduction against the reference plasmid DNA counts for each replicate.
- LFC scores per guide were averaged across high-quality replicates to obtain sgRNA-level LFCs for each cell line.

Geometric, *LDO* and *GAM* methods were applied to gene-level LFCs and the datasets were further processed as follows: Raw gene-level LFCs were obtained by taking the median value among the same gene-targeting sgRNAs

per screened cell line. *Chronos*, on the other hand, requires training on dependency data to optimize the parameters of its mechanistic model of cell population dynamics. This involves removing clonal outgrowths and inferring uncorrected gene effects(J. M. Dempster et al. 2021). To avoid retraining the model since the results were already available, we downloaded the inferred gene effects from the DEPMAP portal (ScreenGeneEffectUncorrected.csv,https://depmap.org/portal/download/all/). We then applied the *Chronos* correction strategy for copy number bias to each dataset. Finally, we intersected the corrected dependency datasets, focusing only on the common cell lines and genes for bias quantification and data distortion analysis. This resulted in 956 cell lines and 17,596 genes for Project Achilles, and 249 cell lines and 17,161 genes for Project Score.

4.2.2. **Arm-corrected Chronos**

To correct for proximity bias post-*Chronos* processing, the median gene effect of each chromosome arm is aligned to be the same across all screens. This involves calculating the median gene effect m_s for each screen, for all genes g in a given chromosome arm within a cell line. The difference between m_s and the median med(m_s) across all cell lines is then subtracted from the gene effect of all genes on the arm e_g, yielding the corrected gene effect for that arm, as follows:

$$e^c_g = e_g - (m_s - med(m_s))$$

This correction is applied to all chromosome arms with > 5 genes.

4.2.3. **In-house implementation of the Geometric method**

As the code for the *Geometric* method has not been made publicly available yet, we implemented our in-house version of this method following the description provided in Lazar et al. (Lazar et al. 2024) and considered as

unexpressed those genes with a transcript per million (TPM) < 1 for a given cell line in the DEPMAP release 23Q2, using basal RNAseq bulk profiles of the same cell line available on the DEPMAP portal (https://depmap.org/), preprocessed following the DEPMAP Omics processing pipeline available at: https://github.com/broadinstitute/depmap_omics.

4.2.4. Batch-wise execution of MAGeCK MLE

Unlike the other methods tested, executing *MAGeCK MLE* on the full Project Achilles and Project Score datasets required prohibitive amounts of memory and time, making it impractical to run the correction in a single step. For instance, when tested on the entire Project Achilles dataset using the institute's in-house high-performance computing cluster, allocating 300 GB of memory and the maximum available computing time (i.e., 2 weeks), the job still failed due to an out-of-memory error. To fully leverage *MAGeCK MLE*'s capabilities for processing and correcting CRISPR-Cas9 knock-out screens in a pooled format, we developed a practical solution by running the method in batches of screens. For further details of this process, please refer to the "Methods" section in (Vinceti et al. 2024).

4.2.5. Quantification of copy number bias

For each dataset, we evaluated the ability of each method to correct for copy number (CN) bias by calculating the area under the recall curve (AURC) for the set of amplified non-expressed genes. These genes were defined as those in the top 1% of the highest CN scores that were not expressed (TPM < 1) in a given cell line. The remaining non-expressed genes served as the outgroup. The AURC is independent of any fixed LFC threshold. For each rank position (k), we calculated the corresponding threshold-specific recall. Finally, we applied the *trapz* function from the *pracma* R package (v2.4.2) to the set of recall scores to compute the AURC.

In addition, we grouped gene-level LFCs by their PICNIC absolute CN score(Greenman et al. 2009) from the GDSC resource("Iorio et Al.," 2016). Then, for each method, we computed the average residual difference (ARD) across CN-binned LFCs as follows:

$$ARD = \frac{1}{C} \sum_{i=0}^{C} \left| \frac{\tilde{x}_i}{s_i} \right|$$

Where $\tilde{x}_i$ and s_i are the median score and the standard deviation derived from the LFC distribution binned for i^{th} absolute CN value of a given method, respectively. C is the maximum absolute CN value. This metric indicates the goodness of a method in centering the LFC distributions towards 0.

4.2.6. **Quantification of proximity bias**

For each method and dataset, after correction, we calculated the pairwise cosine similarity between gene targets and ordered them by their chromosomal position along the human genome. The results were then quantile-normalized to achieve a mean of 0 and a standard deviation of 0.2, following the procedure outlined in(Lazar et al. 2024). To assess proximity bias based on TP53 status, we first divided the datasets into TP53 wild-type and mutated cell lines and repeated the pipeline for each subset. Next, we used the Brunner-Munzel test statistic to evaluate the presence of proximity bias on a genome-wide scale. For each chromosome arm, we compared the intra-arm cosine similarity distribution, where both genes are located on the same arm, with the inter-arm distribution, where one gene is located on the arm of interest and the other on a different chromosome arm.

4.2.7. **Recall of common essential genes**

We used two sets of pre-defined common essential and non-essential genes as positive and negative classes to assess the impact of each method's correction on data quality. These sets were downloaded from the DEPMAP portal (https://depmap.org/portal/download/all/): the AchillesCommonEssentialControls.csv file, derived from the intersection of Hart and Blomen's essential gene lists (Hart et al. 2015; Blomen et al. 2015), and the AchillesNonessentialControls.csv file, which contains Hart's reference non-essential genes(Hart et al. 2014). For each dataset, algorithm, and cell line, we calculated the AUROC curve and the AUPRC curve using the ccr.ROC_Curve and ccr.PrRc_Curve functions from the *CRISPRcleanR* package (https://github.com/francescojm/CRISPRcleanR)(Hart et al. 2014; Iorio et al. 2018). The final sets of positive and negative controls were determined by intersecting the common essential and non-essential gene sets with the genes screened in the dataset under consideration (either Project Achilles or Project Score) across all algorithms. Additionally, we calculated the cell-wise NNMD between the two gene sets, defined as the difference in the median LFC of positive and negative controls, normalized by the median absolute deviation of the negative controls.

4.2.8. **Dependency biomarkers' analysis**

We downloaded a comprehensive table of annotated cancer genetic variants in cancer from the OncoKB database (https://www.oncokb.org/cancerGenes, 12/21/2023 as last update). To ensure the relevance of the data, we focused on genes marked as having gain-of-function or switch-of-function alterations (i.e., annotated genes with a clear oncogenic role). First, we scaled the genome-wide essentiality profiles for each method and dataset tested so that the median LFC scores of common essential and non-essential genes were -1 and 0, respectively, across cell lines, allowing cross-screen comparability. For

each specific oncogene, we utilized somatic mutation and fusion calls from the DEPMAP portal release 23Q2: oncogenes found mutated in a given cell line were deemed as positive controls, while those that were found wild-type and not expressed (TPM < 1) were considered as negative controls. In addition, only oncogenes with at least one positive and one negative control were retained. Finally, for each method and dataset we calculated the AUROC curve on the pooled oncogene LFC scores using the ccr.ROC_Curve function in the *CRISPRcleanR* R package (https://github.com/francescojm/CRISPRcleanR)(Iorio et al. 2018) to estimate each method's ability to recall oncogenic addictions.

4.2.9. **Recall of pre-defined gene sets of essential/non-essential genes**

We considered three predefined sets of genes: common essential and nonessential genes derived from the DEPMAP portal and additional essential genes derived from the Molecular Signature Database (MsigDB) (https://www.gseamsigdb.org/gsea/msigdb)(Subramanian et al. 2005). To generate MsigDB sets of prior essential genes, we downloaded gene sets from (Pacini et al. 2021), originally derived from MSigDB (v7.2). The gene sets used from KEGG were KEGG_DNA_REPLICATION, KEGG_RNA_POLYMERASE, KEGG_SPLICEOSOME, KEGG_RIBOSOME, and KEGG_PROTEASOME.

For the histone gene set, we combined two Reactome gene sets, namely REACTOME_HATS_ACETYLATE_HISTONES and REACTOME_HDACS_DEACETYLATE_HISTONES, as well as the curated histones gene set from(Behan et al. 2019). For these gene sets, we computed the recall at 5% false discovery rate (FDR) across methods and datasets. Considering an essentiality profile, we ranked the LFC scores of the reference common essential (E) and non-essential (N) genes in increasing order. For each ranked position *k*, a set of predicted essential genes is found:

$$P_k = \{s \in E \cup N | r_s \leq k\}$$

where r_s indicates the rank position of s, and the corresponding positive predictive value (PPV) is computed as follows:

$$PPV_k = |P_k \cap E| / |P_k|$$

We then determined the lowest threshold of LFC (LFC^*) in rank position k^* with $PPV_k \geq 0.95$. This is equivalent to an FDR ≤ 0.05. The LFC^* threshold corresponds to $r_s = k^*$. All genes with an LFC score $\leq LFC^*$ were deemed essential at 5% FDR. Given one of the three gene sets G (i.e., common essential, MsigDB essential or non-essential), the recall at 5% FDR is defined as:

$$R_{FDR} = |G*| / |G|$$

where G^* is the subset of genes in G below the LFC^* threshold.

4.2.10. Systematic association between cancer functional events and gene dependencies

For each method, dataset, CFE, and lineage, we conducted a systematic twosided unpaired t-test to evaluate the differential essentiality of each strongly selective dependency (SSD) in relation to the status (presence/absence) of predefined cancer functional events (CFEs), such as somatic mutations, recurrently aberrant copy number segments, and hypermethylated sites. SSDs were identified using a likelihood ratio test as described in (J. Dempster et al. 2019; Pacini et al. 2021). The null hypothesis assumed equal means between the compared populations, while the alternative hypothesis suggested an association between the tested CFE/gene-dependency pair. To correct for multiple hypothesis testing, we applied the Benjamini–Hochberg method to adjust the obtained p-values. Associations with a FDR of less than 5% were considered statistically significant.

4.3. Linking Transcriptional and Morphological Heterogeneity To Therapeutic Vulnerabilities in Glioblastoma

4.3.1. Data driven selection of ADD3

To identify genes that potentially regulate GSC morphology, we used a published tumor atlas of differentially expressed genes in primary GBM tumors(Bhaduri et al. 2020). We intersected this dataset with a list of morphoregulatory genes involved in neurodevelopment identified in(Kalebic et al. 2019). This yielded a list of 30 candidate genes. The enrichment of adducins among the 30 genes was calculated through a hypergeometric test with the following parameters: the total number of human protein-coding genes = 19,396 (N), total number of adducins = 3 (n), number of selected genes = 30 (k), and number of hits = 3 (x). We then investigated the expression level of the selected genes (29 of 30 genes as one of the genes, MGEA5, was not analyzed in the datasets mentioned below) in 48 annotated GBM cell lines from DEPMAP dataset (22Q2 version) (Tsherniak et al. 2017; Behan et al. 2019; Pacini et al. 2021) and the Sanger Cell Model Passports(van der Meer et al. 2018) observing a bimodal distribution from which we identified 18 highly expressed genes (whose basal expression was seemingly generated by the distribution with the higher mean). Subsequently, we derived the depletion fold change of these 18 genes upon CRISPR/Cas9 targeting in 48 GBM cell lines using the same resources. We excluded pan-cancer core-fitness genes (as predicted in (Vinceti et al. 2021) and focused our attention on ADD3 as an important morphoregulator during development (Kalebic et al. 2019), differentially expressed in GBM (Bhaduri et al. 2020) and with a strong and context-specific depletion fold change in GBM cell lines. We then identified

Onda-11 as the GBM cell line with the highest dependency on ADD3. U-87 MG was selected as a GBM cell line with low or no ADD3 dependency, whereas H4 was selected as a glioma cell line with mild ADD3 dependency.

4.3.1.1. RNA-sequencing and gene expression analyses

During sorting, GFP+ Onda-11 GSCs were collected in lysis buffer containing RNA inhibitors in nuclease-free water. RNA was extracted through SMARTSeq v4 Ultra Low Input RNA Kit for Sequencing (Takara). The libraries were sequenced with NovaSeq 6000 with SP flow cell and the following read configuration: 150 × 10 × 10 × 150. Reads from the same sample, obtained from different sequencing lanes, were aggregated and subjected to adapter trimming using Trim Galore. Processed reads were aligned to the human reference genome (GRCh38) using STAR, and quantification was performed with Salmon. Count data were regularized and log-transformed using the *rld* built-in DESeq2 function, and samples were clustered based on Euclidean distances. Differential expression analysis was performed using DESeq2 using raw counts as input. Differentially expressed genes were identified using a cutoff of absolute $\log_2$ fold change ($\log_2$ FC) $\geq$ 0.5 and a false discovery rate (FDR) < 0.05. To comprehensively evaluate the outcomes of the differential expression analysis, we employed Cancer Cell Line Encyclopedia (CCLE) profiles, standardized to achieve zero mean and unit variance, across 48 GBM cell lines. We calculated pairwise correlation scores across all genes, considering the upper triangle of this matrix as a null distribution of scores. Pairwise Pearson's correlation scores between ADD3 and DEGs were extracted and compared with the null with a *t* test. The source code of the scripts is available on an online repository (https://github.com/RaffaeleMI91/ADD3project/tree/main).

4.3.1.2. Evaluation of differential expressed genes in Glioblastoma cell lines

To comprehensively evaluate the outcomes of the Differential Expression Analysis, we employed Cancer Cell Line Encyclopedia (CCLE) profiles - standardized to achieve zero mean and unit variance – across 48 GBM cell lines. We calculated pair-wise correlation scores across all genes, considering the upper triangle of this matrix as a null distribution of scores. Pair-wise Pearson's correlation scores between ADD3 and DEGs were extracted and compared to the null with a t-test. These DEGs exhibited consistent patterns in other GBM cell lines, showing that upregulated genes are positively correlated and downregulated genes are anti-correlated with ADD3 expression at the basal level, underscoring the robustness of the ADD3 OE signature.

4.3.2. **Single-cell RNA sequencing data collection and analysis**

To elucidate the role of ADD3 in GBM, particularly its association with stemlike cellular states and tumor progression, we collected eighteen 10x single cell datasets of primary tumors from the CPTAC-3 project through the Genomic Data Portal (GDC) data portal. Twelve of the eighteen datasets were excluded from the analysis as they did not meet the baseline quality criteria required for further investigation. For the six datasets retained (i.e., "C3N-01814", "C3N00662", "C3N-02783", "C3N-02190", "C3N-03186", "C3N-03188") we applied standard pre-processing procedures for single cell RNA-seq data analysis by utilizing the *scanpy* framework(Wolf, Angerer, and Theis 2018).

4.3.2.1. Data pre-processing and quality control assessments

We performed pre-filtering steps on each retained dataset using scanpy.pp.filter_cells to retain only cells with at least 200 genes expressed, ensuring the removal of low-quality cells. Similarly, we applied

scanpy.pp.filter_genes to retain genes that were expressed in at least 5 cells, filtering out genes with minimal expression across the dataset. This ensures that only cells and genes with sufficient expression were retained for downstream analysis. For assessing overall data quality and library complexity, we applied log-transformation (log1p=True) to the data and quantified how the top 20 most highly expressed genes (percent_top=[20]) contribute to the total gene expression in each cell. For each dataset, we have used scanpy.pp.calculate_qc_metrics to calculate QC metrics for both mitochondrial and ribosomal genes, marking them as QC variables. Other QC metrics, such as the total number of counts ("total_counts") and number of genes with positive counts ("total_genes_by_counts") in each cell, were calculated to detect potential quality issues, such as cell stress or damage, and provides a cleaner dataset for downstream analysis. In addition to these pre-filtering and QC assessments, we implemented a custom filtering approach based on the Median Absolute Deviation (MAD). For a given data, the function mad_threshold computes the sample median and the MAD, defined as the median of the absolute deviations from the sample median. The function uses astropy.stats.median_absolute_deviastion. Thresholds are then defined as $median \pm n\_mads \times MAD$ where n_mads is a user-defined scalar determining the stringency of outlier detection. We applied the upper threshold through the function mad_filter to pre-computed QC features "log1p_total_counts", "log1p_n_genes_by_counts" and "pct_counts_mt". For each dataset, we then used scanpy.pp.scrublet for doublet detection and removed cells that were predicted as doublet.

4.3.2.2. Normalization and selection of high variable genes

For each dataset total-count normalization and log transformation were performed to standardize gene expression values across cells. Specifically, raw counts were normalized using scanpy.pp.normalize_total with a target sum of 1e4 per cell. A natural logarithm transformation (log1p) was then

applied to the normalized counts using scanpy.pp.log1p. Subsequently, highly variable genes (HVGs) were identified per dataset using scanpy.pp.high_variable_genes with n_top_genes=2000. To refine HVG selection, genes annotated as mitochondrial or ribosomal were excluded unless they also passed the variability threshold independently.

4.3.2.3. Cell-cycle scoring

To quantify cell cycle activity in each dataset, we applied the scanpy.tl.score_genes_cell_cycle. For each dataset, cells were scored based on canonical marker genes associated with the S and G2/M phases of the cell cycle. The function computes two separate scores per cell (S_score and G2M_score) by averaging the expression of predefined S-phase and G2/Mphase gene sets, respectively. It then assigns each cell to a predicted cell cycle phase (G1, S, or G2M) based on the dominant score.

4.3.2.4. PCA and visualization of technical and cell-cycle effects

Prior to batch correction or regression of confounding variables, each dataset was scaled using scanpy.pp.scale with a clipping value of 10 to limit the influence of extreme expression values. Principal Component Analysis (PCA) was then performed using scanpy.pp.pca with the “arpack” solver, restricting the computation to genes marked as highly variable

(mask_var=”highly_variable”). To assess the contribution of technical and biological covariates to global expression variation, PCA scatter plots were generated with scanpy.pl.pca_scatter, coloring cells by mitochondrial content, ribosomal content, number of expressed genes, total counts per cell, and cell cycle scores. These visualizations were used to evaluate the extent of confounding prior to correction.

4.3.2.5. Quantification of covariates' effect using principal components

To quantify the contribution of technical and biological covariates to global transcriptomic variation, we evaluated the association between covariates and principal components (PCs) derived from log-normalized expression data. For each dataset, Pearson correlation coefficients were computed between the first 10 principal components and six covariates: mitochondrial percentage, ribosomal percentage, number of detected genes, total UMI counts, and cell cycle scores. To more formally assess variance attribution, we performed ANOVA on the first five principal components using ordinary least squares (OLS) models. For each PC, a linear model of the form PC ~ pct_counts_mt + pct_counts_ribo + n_genes + total_counts + S_score + G2M_score was fitted using the ols function from the statsmodels.formula.api module. ANOVA was conducted using anova_lm with Type II sum of squares to quantify the proportion of variance in each PC explained by each covariate. The total variance explained by each covariate was computed as a percentage of the total sum of squares, and residual (unexplained) variance was reported accordingly. Results were summarized per dataset and visualized to identify covariates contributing more than 5% to the variance in any PC.

4.3.2.6. Regression of technical covariates

To mitigate the influence of technical variation on downstream analyses, selected covariates were regressed out from the log-normalized gene expression data in each dataset using scanpy.pp.regress_out. For all datasets except *C3N-01814*, three covariates—total UMI counts per cell (total_counts), number of detected genes (n_genes), and ribosomal content (pct_counts_ribo)—were regressed out. For *C3N-01814*, only total_counts and n_genes were regressed out due to the negligible or inconsistent impact of ribosomal content in this specific sample. This regression step was applied to each gene independently to remove linear effects of the specified covariates,

ensuring that subsequent dimensionality reduction and clustering analyses reflected primarily biological rather than technical sources of variation.

4.3.2.7. Graph construction, clustering and evaluation of cluster quality

Following normalization, regression and PCA, we computed k-nearest neighbors (KNN) (n_neighbors=15) for each dataset using scanpy.pp.neighbors and the top 30 principal components (n_pcs=30). Based on this graph, a two-dimensional UMAP embedding was generated with scanpy.tl.umap with parameters min_dist=0.3 and spread=1.0 to balance preservation of local and global structure. We then performed Leiden clustering at multiple resolutions (0.4, 0.6, 0.7, 0.8, 0.9 and 1.0) using scanpy.tl.leiden. To assess cluster compactness and separation, we computed the silhouette score for each Leiden resolution using sklearn.metrics.silhouette_score applied to the PCA space. The silhouette score quantifies how well each cell fits within its assigned cluster compared to other clusters, with higher values indicating more distinct and well-separated clusters.

4.3.2.8. Cluster annotation

To assign biologically meaningful cell type labels based on transcriptional programs, we leveraged gene sets derived from Neftel et al (Neftel et al. 2019). These included signatures for MES1-like, MES2-like, AC-like, OPC-like, NPC1-like, NPC2-like, and cell cycle phases (G1/S, G2/M). Overrepresentation analysis (ORA) was performed on each dataset using decoupler.run_ora, with the derived gene sets as input. The analysis was conducted on log-normalized data, and the resulting ORA activity scores were stored. To ensure numerical stability, non-finite values in the ORA score matrices were replaced with the maximum observed finite value in each dataset. For each dataset, the top three enriched transcriptional programs per

cluster were identified using decoupler.rank_sources_groups, with cluster assignments provided by previously selected Leiden resolutions. Based on the highestscoring gene set, each cluster was annotated with a dominant cell type identity. This mapping was dataset-specific and derived from the, which linked clusters to their top-ranked transcriptional program. Finally, cell type–specific marker genes were identified for each dataset by applying scanpy.tl.rank_genes_groups, ranking all genes based on their differential expression across the annotated cell types.

4.3.2.9. Pseudobulk aggregation and ANOVA testing for ADD3 expression across cell types

To enable pseudo-bulk analysis, individual datasets were concatenated into a single object using anndata.concat. Pseudobulk profiles were generated using the decoupler.get_pseudobulk function, aggregating raw counts by cell_type within each batch (patient id). Only cell type–batch combinations with at least 10 cells and a minimum of 1000 total counts were retained (min_cells=10, min_counts=1000). The resulting pseudo-bulk matrix was normalized by totalcount scaling to 10,000 reads per sample (scanpy.pp.normalize_total) and logtransformed (scanpy.pp.log1p). Highly variable genes (HVGs) were selected (n_top_genes=2000), and PCA was performed using only HVGs (mask_var="highly_variable"). For downstream differential expression testing, pseudo-bulk expression values for the gene ADD3 were also extracted. A oneway ANOVA was performed using scipy.stats.f_oneway to test for expression differences of ADD3 across annotated cell types. This was followed by Tukey's Honest Significant Difference (HSD) post hoc test (statsmodels.stats.multicomp.pairwise_tukeyhsd) to identify significantly different cell type pairs, controlling for multiple comparisons ($\alpha = 0.05$).

4.3.3. **Integration of cell viability data with gene expression in cancer cell lines**

4.3.3.1. Gene effects data processing

CRISPR-Cas9 gene effect scores were obtained from the DEPMAP 24Q4 dataset. To correct for variability in essentiality across genes and cell lines, the dependency matrix was normalized using the CoRe.scale_to_essentials function from the CoRe R package(Vinceti et al. 2021). This scaling approach uses curated BAGEL reference lists of essential and non-essential genes to standardize gene effect scores. A total of 10,130 never essential and 1542 corefitness essential genes (i.e,, genes involved in DNA replication, histone proteins, proteasome and spliceosome components, ribosomal proteins (Vinceti et al. 2021)) were removed from this analysis, finally leading to a gene essentiality matrix of 6404 CRISPR gene knockouts and 1178 cell lines.

4.3.3.2. Bulk RNA-seq dataset processing

Batch-corrected gene expression data from the DEPMAP 24Q4 release was filtered to retain highly variable genes. Genes' standard deviations across cell lines were computed, and the 35th percentile of this distribution were calculated and used as threshold. Only genes with a standard deviation greater than 35th percentile were retained. This resulted in a gene expression matrix of 12,414 genes and 1673 cell lines.

4.3.3.3. Construction of Genomic Profile of Drug Sensitivity (GPDS) ranked lists

Gene expression and the killing potency of a drug (i.e., IC_{50}) were aligned to include only cell lines for which the IC_{50} was measured. Genomic Profiles of

Drug Sensitivity (GPDS) (Pellecchia et al. 2023) were generated by computing gene-wise Pearson correlation coefficients between gene expression and IC_{50} values across cell lines. This process produced ranked lists of expression-based biomarkers across 664 drugs or small molecules, from the GDSC databases(Yang et al. 2013), with the strength of the correlation indicating their power in predicting the molecule's effect.

4.3.3.4. Construction of Genomic Profile of Gene Essentiality (GPGE) ranked lists

Gene expression and dependency matrices were aligned to include only genes and cell lines screened in both assays, finally retaining 4116 genes and 1103 cell lines. As for GPDS (Pellecchia et al. 2023), Genomic Profiles of Gene Essentiality (GPGE) were generated by computing Pearson correlation coefficients between each matched pair of expression and dependency profiles on a gene-wise basis. This process produced 4116 ranked lists of expressionbased biomarkers for each tested gene knockout from DEPMAP (Tsherniak et al. 2017; Boehm et al. 2021; Trastulla et al. 2023) with the strength of the correlation indicating their power in predicting the knockout's effect.

4.3.3.5. Gene set enrichment analysis to map differentially expressed genes against GPDS and GPGE

Gene set enrichment analysis (GSEA) was performed using the *fgsea* R package version 1.30.0. Ranked gene lists were tested against a collection of cell type–specific gene sets (e.g., glioblastoma subtype markers). Multiple testing correction was applied using the Benjamini–Hochberg procedure to control the false discovery rate (FDR). Following the DREEP framework(Pellecchia et al. 2023), non-significant enrichments (adjusted $p >$

0.05) were assigned a normalized enrichment score (NES) of 0.

4.3.3.6. Pathway enrichment analysis

Pathway enrichment analysis of deregulated genes (both upregulated and downregulated) was performed using the *fgsea* package in R version 1.30.0, in combination with hallmark gene sets retrieved via the *msigdbr* package. Specifically, hallmark pathways *Homo sapiens* were obtained using the function msigdbr() and organized into a list format to match gene symbols with their corresponding gene sets. The *fgsea* function was then applied to assess enrichment scores and significance for each pathway based on ranked gene statistics. Pathways with an adjusted p-value < 0.05 were considered significantly enriched.

4.3.3.7. Survival analysis

Survival analysis was performed using the R packages *survival* version 3.8-3 and *survminer* version 0.5.0 on a cohort of 166 bulk GBM tumors obtained from the TCGA-GBM database. For each gene of interest, patients were stratified into high- and low-expression groups based on the median expression value. Kaplan–Meier (KM) survival curves were generated to assess overall survival differences between the groups, and statistical significance was evaluated using the log-rank test. Genes showing significant associations with survival were considered prognostically relevant.

4.3.3.8. Scoring of genes regulating oligodendrogenesis

We assessed the implication of the predicted vulnerabilities in transcriptional programs involved in oligodendrogenesis and (re)myelination by using a knowledge-driven scoring procedure(Huré et al. 2024) implemented into a user-friendly Web Server Application (OligoScore platform). This

userfriendly platform enables the comparison of input gene sets with a curated and continuously updated reference list of 454 genes implicated in oligodendrocyte development, curated by experts in oligodendrocyte biology. For further details, we suggest referring to the original publication.

4.3.3.9. Target enrichment analysis

Target enrichment analysis of drugs predicted to be sensitive/resistant was performed GSEA using the *fgsea* package in R version 1.30.0, in combination with a curated and annotated datasets of compounds screened in the GDSC(Yang et al. 2013). We extracted significantly enriched gene sets (adjusted p-value < 0.05) from prior GSEA results and constructed samplespecific ranked vectors of normalized enrichment scores (NES). Drug targets with an adjusted p-value < 0.05 were considered significantly enriched.

5. Appendix

5.1. Introduction

Cell-type–aware drug discovery models have so far been developed primarily on bulk CCLE pharmacogenomic data, implicitly assuming that cancer cell lines (CCLs) are homogeneous. However, recent single-cell transcriptomic studies have revealed that CCLs often harbor substantial heterogeneity, including distinct subclones shaped by selective pressures and capable of exhibiting differential drug sensitivities(Kinker et al. 2020; U. Ben-David 2018). This complexity poses a challenge for bulk-based models, as it can obscure cell-state–specific vulnerabilities. At the same time, it presents a methodological opportunity: incorporating single-cell transcriptomic data from CCLs could refine bulk approaches and enable more accurate frameworks for target prioritization. As part of a broader effort, my contributions have extended to investigating whether and how single-cell transcriptomic data of CCLs can serve as additional resources for improving drug response models. Although this line of work is still in its early stages, I have shown that integrating singlecell–derived data can enhance the predictive power of pharmacogenomic models, which are typically built without accounting for this additional layer of heterogeneity. Taken together, these efforts illustrate how my thesis seeks to integrate large-scale functional genomics and pharmacogenomics data to improve strategies for cancer target discovery. By bridging bulk and single-cell perspectives, and by exploring both methodological and biological dimensions of tumor heterogeneity, this work aims to push the boundaries of how we model cancer and predict therapeutic outcomes, ultimately laying the groundwork for more effective, context-aware treatments in the future.

5.2. Results

5.2.1. A bioinformatic pipeline to identify polyclonal cancer cell lines

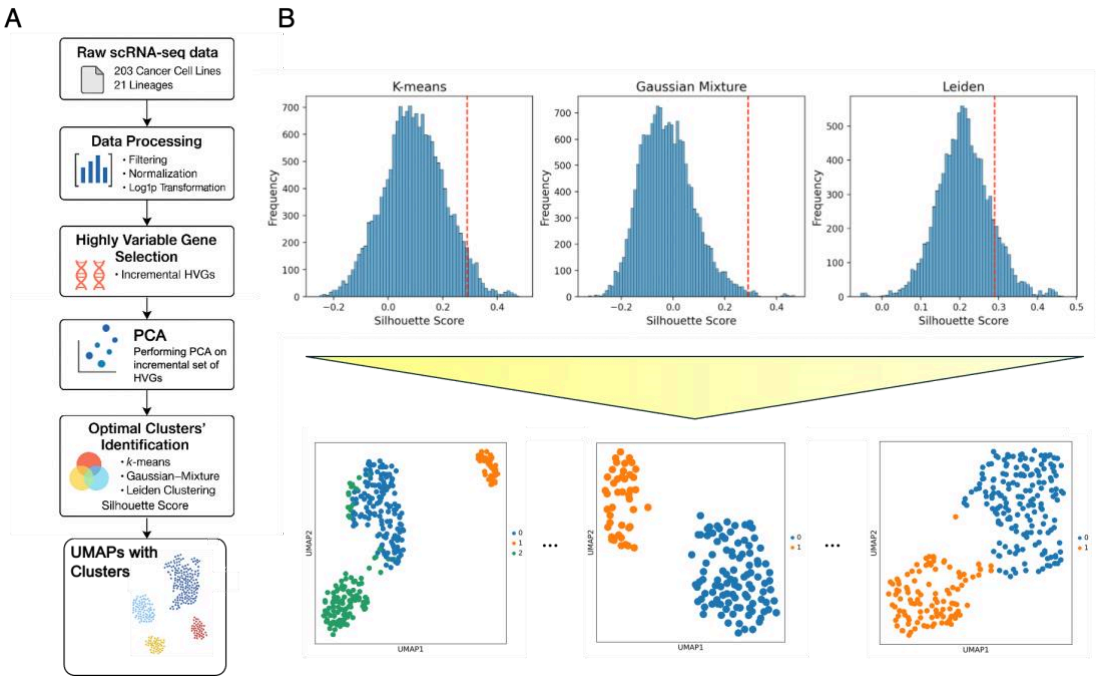


Figure 37 - Schematic illustrating the bioinformatic pipeline adopted: (A) For each of the 203 CCLs, we implemented a data-driven strategy using three clustering algorithms (i.e., *k*-means, Gaussian Mixture Models (GMM), and Leiden Clustering) each capturing distinct and complementary patterns in the data. We iterated this approach across various combinations of hyperparameters and selected the polyclonal CCLs based on the optimization of clustering performance, ensuring that the chosen models reflected the inherent transcriptional heterogeneity within the cell lines. (B) Top: Distribution of the Silhouette scores across the clustering algorithms and hyperparameters tested. The red dashed vertical line represents the threshold of 0.3 used for selecting optimal clusters (see Material and Methods), ensuring that only those clusters with sufficient intra-cluster consistency and separation from other clusters were retained for further analysis. Bottom: Umaps visualization of polyclonal cancer cell lines.

We applied our bioinformatics pipeline to analyze scRNA-seq data from 203 cancer cell lines spanning 21 different lineages, obtained from the Broad Institute's Single-Cell Portal (Kinker et al. 2020). We specifically focused on

a subset of cell lines treated with drugs from the GDSC database(Yang et al. 2013). Using the COSMIC IDs from the scRNA-seq data, we selected cell lines for which drug treatment information was available in the GDSC and identified a total of 139 cell lines that were included in both datasets. This step was needed to build machine learning models aimed at predicting drug response for polyclonal cancer cell lines that were tested in the GDSC. Our pipeline enabled the identification of 78 CCLs, whose transcriptional heterogeneity mirrors diverse cellular subpopulations, offering a valuable resource to expand existing genome-wide expression datasets of CCLs.

5.2.2. A two-step transfer learning framework to enhance drug response prediction using bulk and single-cell RNA expression data

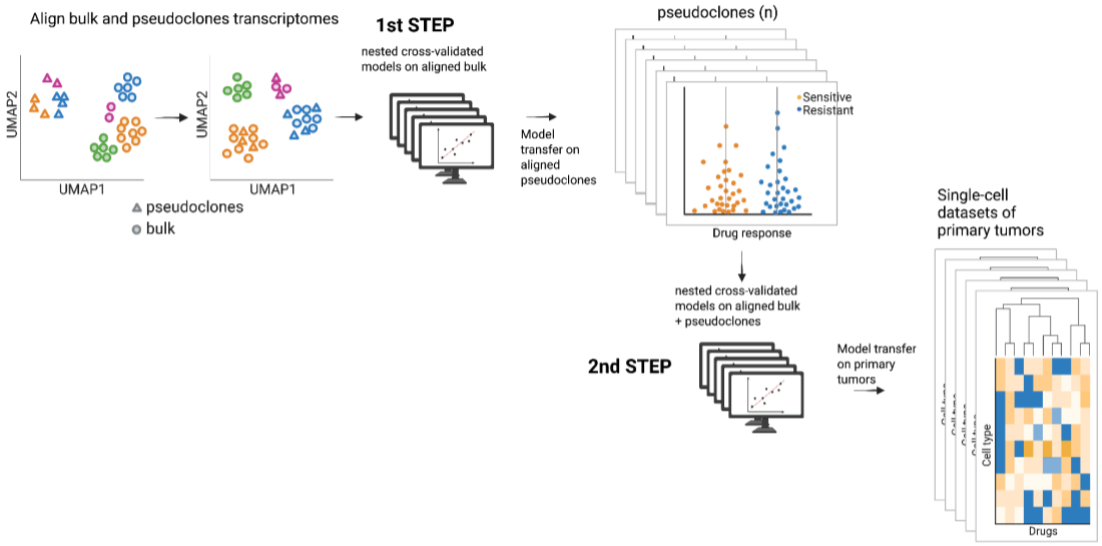


Figure 38 - Schematic illustrating the two-step ML model adaptation from cell lines to heterogeneous tumor biopsies, incorporating genome-wide expression data of polyclonal cancer cell lines. In the first step, we train the elastic-net regression model on the alinged bulk data from the CCLE dataset and predict the IC50 response of 192 pseudoclones' RNA expression profiles to 271 drugs. In the second step, another elastic-net model is trained and cross-validated using the expanded dataset, which accounts for the heterogeneity introduced by the pseudoclones (i.e., aligned bulk and pseudoclones)

whose IC50 responses were predicted in the first step. Finally, the model is adapted to capture the complexity and diversity of cellular subpopulations within tumor biopsies, bridging the gap between preclinical models and real-world clinical data.

5.2.3. Single-cell derived pseudo clones can be integrated into bulk transcriptomic profiles of cancer cell lines

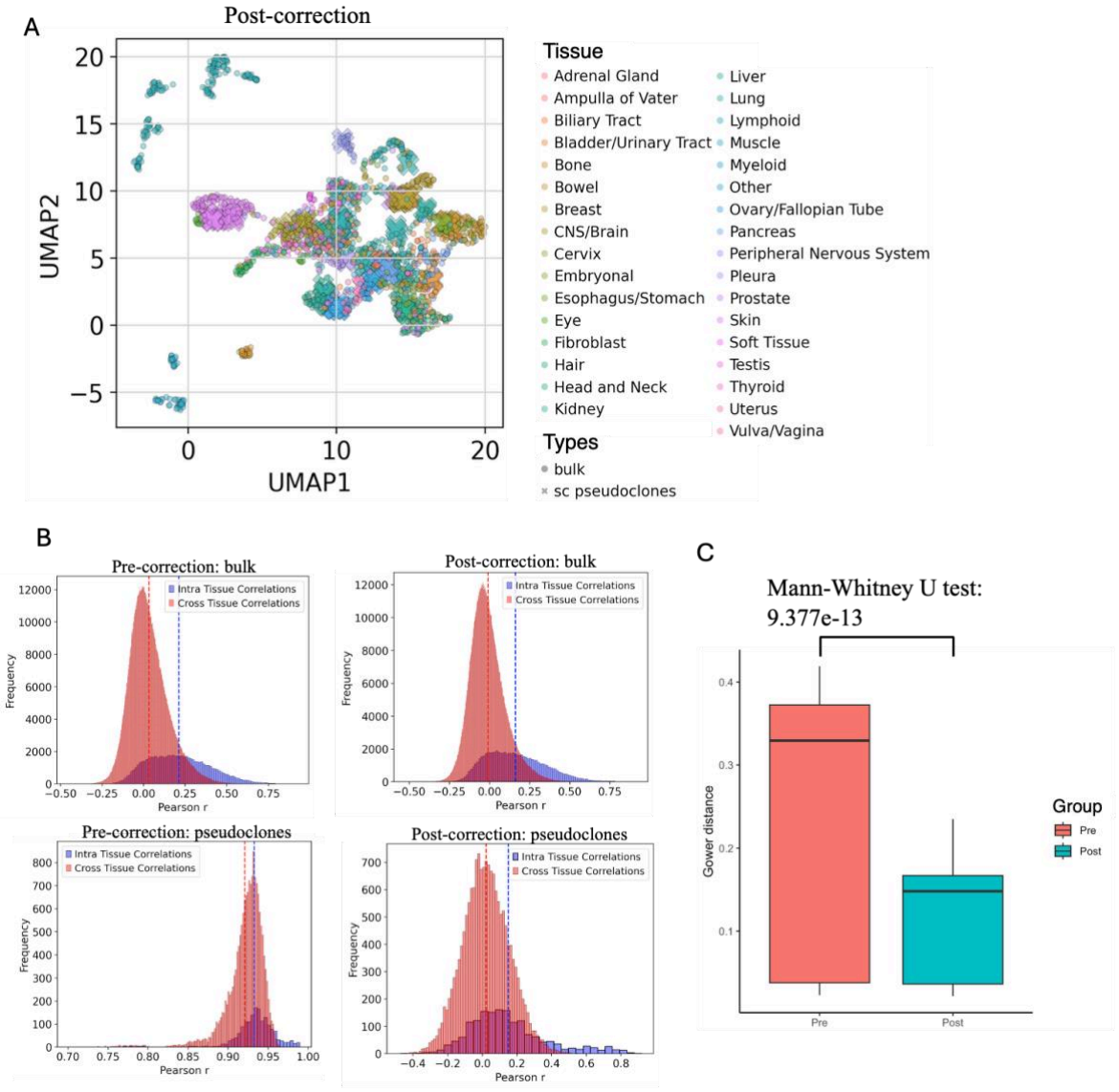


Figure 39 - Alignment of bulk and pseudoclones transcriptomes: (A) UMAP visualization showing the integration of pseudoclones with bulk RNA expression profiles across multiple lineages in the CCLE database. (B) Distribution of intra-tissue (blue) versus cross-tissue (red) Pearson's correlations before and after alignment with ComBat for both bulk RNA (top) and pseudoclone RNA (bottom) expression profiles. (C) Distribution of Gower distances calculated before and after ComBat alignment between pseudoclones

of polyclonal cancer cells and their corresponding bulk counterparts, with a Mann-Whitney U test p-value of 9.77e-13.

We split the genome-wide expression profile of a polyclonal CCL into two or more aggregated profiles, referred to as "pseudoclones," based on the number of clusters identified, ultimately obtaining a set of 192 pseudoclone profiles. These pseudoclone RNA expression profiles were then integrated with the bulk RNA expression data from the Cancer Cell Line Encyclopedia (CCLE) using ComBat, resulting in an expanded dataset of 1709 profiles (1517 cell lines from CCLE + 192 derived pseudoclones). We first assessed whether the tissue/lineage relationships were preserved after the ComBat correction. While the bulk profiles remained stable before and after correction, a notable improvement was observed for the pseudoclone RNA expression profiles postalignment (Figure 39B bottom). The intra- and cross-lineage distribution patterns for pseudoclones closely resembled those of the bulk profiles. Additionally, the distribution of Gower distances (Figure 39C) calculated before and after ComBat alignment between pseudoclones and their corresponding bulk counterparts showed a significant reduction in distance after alignment, with a Mann-Whitney U test p-value of 9.77e-13. This result confirms that the alignment successfully reduced batch effects related to the platform and enhanced data consistency.

5.2.4. Single-cell derived pseudoclones enhances predictive performance of drug response modeling

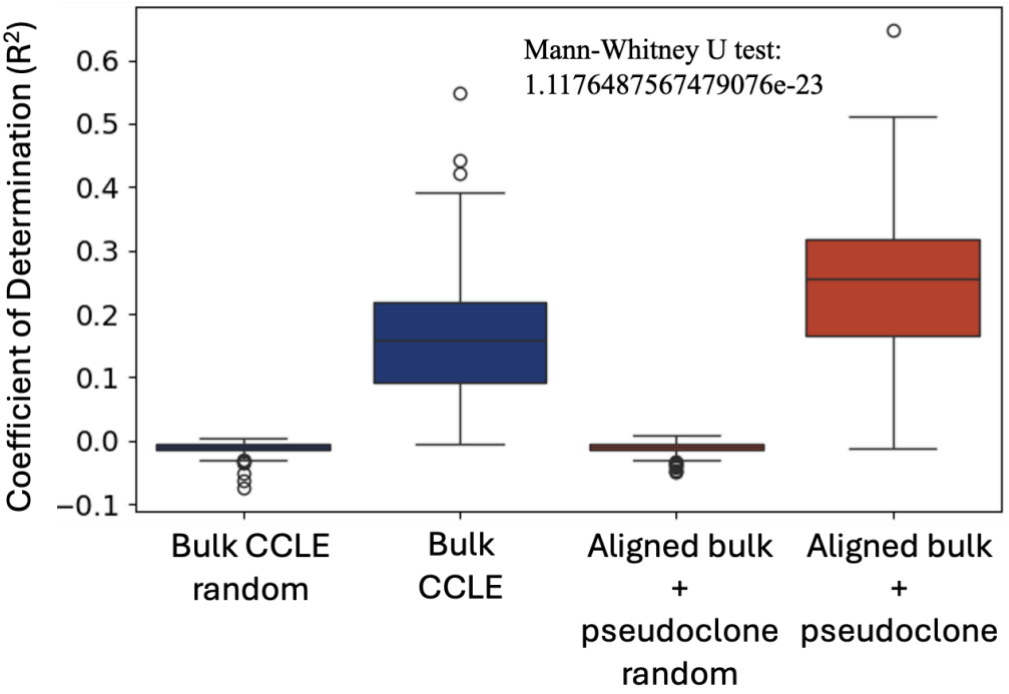


Figure 40: Model performances: Boxplots showing the distribution of the coefficient of determination (R^2) for each model tested with aligned bulk and pseudoclone profiles, demonstrating statistically significant improvements in performance (Mann-Whitney U pvalue = 1.118e-23)

Using aligned bulk RNA and pseudoclone RNA expression data has proven to be more effective than relying solely on CCLE bulk expression data to predict drug response. When comparing the elastic-net regression model trained only on the bulk CCLE dataset with the model built using the aligned (or expanded) dataset, we observe promising results. The integration of pseudoclones derived from single-cell data, did not degrade model performance; rather, it led to a statistically significant improvement. While further evaluations are necessary, this preliminary result is groundbreaking and lays the foundation for an informed data-integration strategy that can enhance current state-of-the-art machine learning models for drug response prediction. By incorporating pseudoclones of polyclonal CCLs, we encourage

the model to better account for variability, improving its ability to generalize to new or unseen data and reducing the risk of overfitting. Additionally, we strongly suggest that for successful model adaptation to single-cell RNA expression profiles of primary tumors, models must be trained on these expanded resources, where single-cell data variability is considered.

5.3. Discussion

The idea of cell-type aware therapeutic discovery formed the foundation for designing a transfer learning framework that aims to integrate single-cell gene expression profiles of CCLs with drug response data from the GDSC database. Motivated by evidence suggesting CCLs can evolve to form distinct strains, therefore exhibiting transcriptional heterogeneity(Kinker et al. 2020; Uri BenDavid et al. 2018), we implemented a two-step machine-learning framework that integrates scRNA-seq data from CCLs into the model training process, alongside bulk data. If intra-cell-line heterogeneity is linked to variations in therapeutic responses, these cell lines could serve as valuable integrative models to identify novel drug targets or combination therapies, insights that may be disregarded when relying solely on bulk data. While further assessments are required, this preliminary finding is promising and lays a solid foundation for a data-integration strategy that could enhance current state-ofthe-art transfer learning models that aim to adapt their predictions to the heterogeneous nature of primary tumors, as captured at the single-cell level. Another future perspective would be to integrate genetic dependency data into the training phase of such frameworks by modelling *gene expression-gene dependency* relationships, after harmonizing scRNA-seq data with bulk data from CCLs. As tumors evolve and develop resistance to treatment, sub clonal dependencies may emerge that are unique to specific subclones within the tumor. These dependencies may differ from those of the

primary tumor, with some genetic vulnerabilities being gained and others lost as the tumor adapts to therapeutic pressure. For example, a subclone might acquire a mutation that makes it resistant to a particular drug, but this same mutation might render the subclone reliant on a previously non-essential gene. This newly acquired genetic dependency could represent a potential therapeutic target, whereas the bulk of the tumor may not exhibit the same dependency. Moreover, the evolving nature of subclonal dependencies might complicate the drug discovery process, as it introduces dynamic and context-dependent vulnerabilities that are not always captured in traditional drug screening models. Incorporating genetic dependency data would help identify contextspecific genetic vulnerabilities that emerge in resistant subclones and provide mechanistically insights of acquired drug resistance mechanisms and drugs mode of action (MoA). Moreover, it would facilitate the identification of new potential synthetic lethal interactions, enabling a more effective, context-aware prioritization, potentially uncovering combination therapies that could improve treatment outcomes by targeting multiple genetic pathways simultaneously.

5.4. Methods

5.4.1. Integrative analysis of single-cell RNA sequencing data in cancer cell lines

5.4.1.1. Collection and processing of scRNA-seq data

We obtained raw scRNA-seq data for 203 cancer cell lines spanning 21 lineages from the Broad Institute's single-cell portal(Kinker et al. 2020). We then selected cell lines with available drug response profiles from the GDSC database, ultimately focusing on a set of 139 cell lines To ensure data quality, we performed a global filtering procedure using the Scanpy framework(Wolf, Angerer, and Theis 2018). Specifically, we used scanpy.pp.filter_cells to exclude cells expressing fewer than 200 genes and scanpy.pp.filter_genes to remove genes that were expressed in fewer than 5 cells. Raw scRNA-seq data for each cancer cell line were then processed individually. Data were normalized using the scanpy.pp.normalize_total function, scaling the counts to a target sum of 10,000 per cell. To stabilize variance, we applied log1p transformation scanpy.pp.log1p to the normalized data.

5.4.1.2. Hyperparameter tuning for detecting optimal intra-cell-line variation

To identify the most informative genes across each cancer cell line, we performed a hyperparameter tuning approach to compute incremental sets of HVGs. For each cancer cell line, we iterated over a range of potential values for the number of highly variable genes (N), varying from 2,000 to 8,500 in increments of 500. For each value of N, the following steps were performed:

- Highly variable gene selection: Using the scanpy.pp.highly_variable_genes function, we identified the top N genes with the highest variance across cells.

- Remove mitochondrial and ribosomal genes: To avoid potential confounding effects from mitochondrial and ribosomal genes, which are often non-informative or biased by technical variation, we ensured that these genes were excluded from the set of highly variable genes.
- To assess the impact of different sets of HVGs on downstream analyses, we performed PCA on subsets of the data defined by varying numbers of HVGs. For each iteration, we selected a subset of the data containing only the HVGs identified previously for the current value of N. To standardize the gene expression values across all cells, we applied scaling to the selected subset of HVGs using scanpy.pp.scale.
- PCA was then performed on the scaled data using scanpy.pp.pca with the arpack solver.

5.4.1.3. Optimal clusters' identification

Top 20 PCs computed on an incremental range of HVGs were then used as input for three different clustering algorithms (i.e, *k*-means, Gaussian-Mixture, Leiden clustering).

Let:

- *K* be the set of possible cluster numbers (or resolutions, in the case of the Leiden clustering);
- *M* be the set of clustering methods;
- *HVG* be the set of highly variable genes used to compute PCs;
- *SS*(*k, m, h*) be the silhouette score for a specific combination of cluster's parameters *k*, clustering method *m*, and set of highly variable genes *h*.

The optimal combination of the hyperparameter *k*, method *m*, and set *h* of

HVGs, can be defined as the set k_{opt}, m_{opt}, h_{opt} subject to $SS(k_{opt}, m_{opt}, h_{opt}) \geq 0.3$, that maximise the Silhouette Score. We retained cancer cell lines that showed sufficient intra-cell-line variability based on this optimization pipeline and defined this set as polyclonal cancer cell lines.

5.4.1.4. Cross-platform normalization and alignment of pseudoclones RNA expression with bulk RNA-seq data

Pseudoclones profiles were generated averaging raw counts by cluster id across polyclonal cancer cell lines. Data from pseudoclones were then normalized by gene length and scaled to TPM. We applied log transformation to TPM profiles and added a pseudocount. We then concatenated pseudoclones and CCLE RNA expression profiles and scaled prior alignment with pyComBat version 0.3.2.

5.4.1.5. Assessing ComBat alignment

We considered genome-wide expression datasets of concatenated pseudoclones and CCLE RNA expression pre- and post-alignment with ComBat. The two datasets were filtered for bulk RNA and pseudoclones for specific CCLs from the polyclonal CCL list. The same procedure was applied to the post-corrected dataset, ensuring that both datasets were aligned and compatible with the polyclonal CCLs selected for the analysis (n=78). To quantify the similarity between the bulk and pseudoclone RNA expression profiles, we computed the Gower distance for both the pre-corrected and post-corrected datasets. This distance metric was computed using the gower.dist function from the *StatMatch* R package version 1.4.3. To compare the Gower distances before and after correction, we performed Mann-Whitney U between the two distributions of distances (pre and post) extracted from the upper triangular part of the distance matrices.

5.4.1.6. Evaluating overall data distorsion after alignment

To check if the overall tissue/lineage relationships between cell lines were affected by the alignment with ComBat, we considered the genome-wide expression of concatenated pseudoclones (n=192) and CCLE RNA expression profiles (n=1517) pre- and post-alignment, and computed Pearson's correlation scores between cell lines belonging to the same lineage/tissue (i.e., "Intratissue" distribution of Pearson's correlation coefficients) across the two batches. We compared this distribution with a null distribution (i.e., "Crosstissue" distribution of Pearson's correlation coefficients) where all possible tissue/lineage relationships were considered. We visually assessed the separation of the two distributions by plotting histograms of the two distributions.

5.4.1.7. Two-step predictive modelling of drug response using GDSC data

The viability profiles of 78 polyclonal CCLs were available for a set of 271 drugs from the GDSC. Therefore, we developed machine learning models of drug response to firstly predict pseudoclones' IC_{50} profiles for this set of drugs (STEP1). The list of drugs was iterated over, and for each drug, the aligned RNA expression dataset (n=1709) was divided based on the match of cell lines whose IC_{50} profiles were available for the drug considered. Cell lines were retained only if they had corresponding data in both datasets. If fewer than 250 cell lines met these criteria, the drug was excluded from the analysis. Each dataset was further filtered by tissue type. We selected tissue types that had at least 10 cell lines, ensuring that each tissue type had sufficient representation in our models. Next, we reduced the feature space by focusing on the most informative genes. We calculated the coefficient of variation (CV) for each gene by dividing the standard deviation by the mean expression

across cell lines. A threshold was set at the 90th percentile of the CV distribution, and genes with a CV greater than this threshold were classified as highly variable features (HVFs). Furthermore, to address multicollinearity and improve model stability, we extracted and removed features that had a correlation $\geq$ 0.99. Finally, the resulting common set of cell lines and only the set of high variable features were then used as input for a total of 271 nested cross-validated elasticnet regression models in which the bulk aligned data from CCLE, which represented the average gene expression for each cell line, was used as *source* domain to transfer the model on aligned pseudoclones' RNA expression profiles which, in STEP1, represented our *target* domain. In STEP2, the model was expanded by incorporating pseudoclones' IC50 responses of the predicted in the STEP1. To compare the performance of the model trained in STEP2, a third model was trained considering only bulk RNA expression data from the CCLE across the set of 271 drugs. For each of these models, we compared their performances to random baselines, where random performances were computed by shuffling the IC50 values. This process allowed us to assess whether the model's predictions were significantly better than what would be expected by chance, providing a more robust evaluation of model accuracy and performance. The scikit-learn library version 1.7.0 was used to build elastic-net regularized regression models with balanced weights for L1 and L2 norm regularization. We defined a logarithmic range for the alpha hyperparameter to capture a broad range of regularization strengths, allowing the model to explore both very weak and very strong regularization. Similarly, we created a linear range between 0.1 and 1.0 for l1_ratio to explore different degrees of sparsity (from Ridge to Lasso). Models were 5-fold nested cross-validated for both hyperparameter tuning and performance evaluation. Performances were quantified by using the Coefficient of Determination (R^2) score and were averaged across the 5 outer-folds.

6. References

Adikusuma, Fatwa, Sandra Piltz, Mark A. Corbett, Michelle Turvey, Shaun R. McColl,

Karla J. Helbig, Michael R. Beard, James Hughes, Richard T. Pomerantz, and Paul Q. Thomas. 2018. “Large Deletions Induced by Cas9 Cleavage.” *Nature*.

Aguirre, Andrew J., Robin M. Meyers, Barbara A. Weir, Francisca Vazquez, Cheng-Zhong

Zhang, Uri Ben-David, April Cook, et al. 2016. “Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting.” *Cancer Discovery* 6 (8): 914–29.

Ballouz, Sara, and Jesse Gillis. 2016. “AuPairWise: A Method to Estimate RNA-Seq Replicability through Co-Expression.” *PLoS Computational Biology* 12 (4):

e1004868.

Bao, Shideng, Qiulian Wu, Roger E. McLendon, Yueling Hao, Qing Shi, Anita B.

Hjelmeland, Mark W. Dewhirst, Darell D. Bigner, and Jeremy N. Rich. 2006. “Glioma Stem Cells Promote Radioresistance by Preferential Activation of the DNA Damage Response.” *Nature* 444 (7120): 756–60.

Barelli, Carlotta, Flaminia Kaluthantrige Don, Raffaele M. Iannuzzi, Stefania Faletti, Ilaria

Bertani, Isabella Osei, Simona Sorrentino, et al. 2025. “Morphoregulatory ADD3 Underlies Glioblastoma Growth and Formation of Tumor-Tumor Connections.” *Life Science Alliance* 8 (2). https://doi.org/10.26508/lsa.202402823.

Behan, Fiona M., Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M.

Beaver, Giorgia Migliardi, Rita Santos, et al. 2019. “Prioritization of Cancer

Therapeutic Targets Using CRISPR–Cas9 Screens.” *Nature* 568 (7753): 511–16.

Bello, Silvia M., Lucile Crété, Julia Galway-Witham, and Simon A. Parfitt. 2021. “Knapping Tools in Magdalenian Contexts: New Evidence from Gough’s Cave (Somerset, UK).” *PloS One* 16 (12): e0261031.

Ben-David, U. 2018. “Genetic and Transcriptional Evolution Alters Cancer Cell Line Drug Response.” *Nature* 560 (7718): 325–30.

Ben-David, Uri, Benjamin Siranosian, Gavin Ha, Helen Tang, Yaara Oren, Kunihiko Hinohara, Craig A. Strathdee, et al. 2018. “Genetic and Transcriptional Evolution Alters Cancer Cell Line Drug Response.” *Nature* 560 (7718): 325–30.

Bhaduri, Aparna, Elizabeth Di Lullo, Diane Jung, Sören Müller, Elizabeth Erin Crouch,

Carmen Sandoval Espinosa, Tomoko Ozawa, et al. 2020. “Outer Radial Glia-like
Cancer Stem Cells Contribute to Heterogeneity of Glioblastoma.” *Cell Stem Cell* 26 (1): 48-63.e6.

Blomen, Vincent A., Peter Májek, Lucas T. Jae, Johannes W. Bigenzahn, Joppe Nieuwenhuis, Jacqueline Staring, Roberto Sacco, et al. 2015. “Gene Essentiality
and Synthetic Lethality in Haploid Human Cells.” *Science (New York, N.Y.)* 350 (6264): 1092–96.

Bock, Christoph, Paul Datlinger, Florence Chardon, Matthew A. Coelho, Matthew B. Dong, Keith A. Lawson, Tian Lu, et al. 2022. “High-Content CRISPR Screening.” *Nature Reviews Methods Primers* 2 (1): 1–23.

Bodapati, Sunil, Timothy P. Daley, Xueqiu Lin, James Zou, and Lei S. Qi. 2020. “A Benchmark of Algorithms for the Analysis of Pooled CRISPR Screens.” *Genome Biology* 21 (1): 62.

Boehm, Jesse S., Mathew J. Garnett, David J. Adams, Hayley E. Francies, Todd R. Golub, William C. Hahn, Francesco Iorio, James M. McFarland, Leopold Parts, and Francisca Vazquez. 2021. “Cancer Research Needs a Better Map.” *Nature* 589 (7843): 514–16.

Boehm, Jesse S., and Todd R. Golub. 2015. “An Ecosystem of Cancer Cell Line Factories to Support a Cancer Dependency Map.” *Nature Reviews. Genetics* 16 (7): 373–74.

Chan, Edmond M., Tsukasa Shibue, James M. McFarland, Benjamin Gaeta, Mahmoud Ghandi, Nancy Dumont, Alfredo Gonzalez, et al. 2019. “WRN Helicase Is a Synthetic Lethal Target in Microsatellite Unstable Cancers.” *Nature* 568 (7753): 551–56.

Chandrasegaran, Srinivasan, and Dana Carroll. 2016. “Origins of Programmable Nucleases for Genome Engineering.” *Journal of Molecular Biology* 428 (5 Pt B): 963–89.

Chauhan, Raashi, Rama Rao Damerla, and Vijay Shree Dhyani. 2025. “Synthetic Lethality in Cancer: A Protocol for Scoping Review of Gene Interactions from Synthetic Lethal Screens and Functional Studies.” *Systematic Reviews* 14 (1): 81.

Chen, Jian, Yanjiao Li, Tzong-Shiue Yu, Renée M. McKay, Dennis K. Burns, Steven G. Kernie, and Luis F. Parada. 2012. “A Restricted Cell Population Propagates Glioblastoma Growth after Chemotherapy.” *Nature* 488 (7412): 522–26.

Chen, Junyi, Xiaoying Wang, Anjun Ma, Qi-En Wang, Bingqiang Liu, Lang Li, Dong Xu, and Qin Ma. 2022. “Deep Transfer Learning of Cancer Drug Responses by Integrating Bulk and Single-Cell RNA-Seq Data.” *Nature Communications* 13 (1): 6494.

Couturier, Charles P., Shamini Ayyadhury, Phuong U. Le, Javad Nadaf, Jean Monlong,

Gabriele Riva, Redouane Allache, et al. 2020. “Single-Cell RNA-Seq Reveals That

Glioblastoma Recapitulates a Normal Neurodevelopmental Hierarchy.” *Nature Communications* 11 (1): 1–19.

Dempster, J., F. M. Behan, T. Green, and H. Najgebauer. 2019. “Agreement between Two

Large Pan-Cancer Genome-Scale CRISPR Knock-out Datasets.” *Nature Communications*, no. 10:5817. https://doi.org/10.1038/s41467-019-13805-y.

Dempster, Joshua M., Isabella Boyle, Francisca Vazquez, David E. Root, Jesse S.

Boehm, William C. Hahn, Aviad Tsherniak, and James M. McFarland. 2021.

“Chronos: A Cell Population Dynamics Model of CRISPR Experiments That

Improves Inference of Gene Fitness Effects.” *Genome Biology* 22 (1): 1–23.

Doench, John G., Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg,

Katherine F. Donovan, Ian Smith, et al. 2016. “Optimized SgRNA Design to Maximize Activity and Minimize Off-Target Effects of CRISPR-Cas9.” *Nature Biotechnology* 34 (2): 184–91.

Dwane, Lisa, Fiona M. Behan, Emanuel Gonçalves, Howard Lightfoot, Wanjuan Yang, Dieudonne van der Meer, Rebecca Shepherd, Miguel Pignatelli, Francesco Iorio, and Mathew J. Garnett. 2020. “Project Score Database: A Resource for Investigating Cancer Cell Dependencies and Prioritizing Therapeutic Targets.” *Nucleic Acids Research* 49 (D1): D1365–72.

Favreau, Julien. 2023. “Sourcing Oldowan and Acheulean Stone Tools in Eastern Africa: Aims, Methods, Challenges, and State of Knowledge.” *Quaternary Science Advances* 9 (January): 100068.

Fustero-Torre, Coral, María José Jiménez-Santos, Santiago García-Martín, Carlos

Carretero-Puche, Luis García-Jimeno, Vadym Ivanchuk, Tomás Di Domenico,

Gonzalo Gómez-López, and Fátima Al-Shahrour. 2021. “Beyondcell: Targeting

Cancer Therapeutic Heterogeneity in Single-Cell RNA-Seq Data.” *Genome Medicine* 13 (1): 187.

Garofano, Luciano, Simona Migliozzi, Young Taek Oh, Fulvio D’Angelo, Ryan D. Najac,

Aram Ko, Brulinda Frangaj, et al. 2021. “Pathway-Based Classification of Glioblastoma Uncovers a Mitochondrial Subtype with Therapeutic Vulnerabilities.” *Nature Cancer* 2 (2): 141–56.

“Gene Editing at CRISPR Speed.” 2014. *Nature Biotechnology* 32 (4): 309–12.

Geng, Keyi, Lara G. Merino, Linda Wedemann, Aniek Martens, Małgorzata Sobota, Yerma P. Sanchez, Jonas Nørskov Søndergaard, Robert J. White, and Claudia Kutter. 2022. “Target-Enriched Nanopore Sequencing and de Novo Assembly Reveals Co-Occurrences of Complex on-Target Genomic Rearrangements Induced by CRISPR-Cas9 in Human Cells.” *Genome Research* 32 (10): 1876–91.

Gonçalves, Emanuel, Fiona M. Behan, Sandra Louzada, Damien Arnol, Euan A. Stronach, Fengtang Yang, Kosuke Yusa, Oliver Stegle, Francesco Iorio, and Mathew J. Garnett. 2019. “Structural Rearrangements Generate Cell-Specific, Gene-Independent CRISPR-Cas9 Loss of Fitness Effects.” *Genome Biology* 20 (1): 27.

Gonçalves, Emanuel, Mark Thomas, Fiona M. Behan, Gabriele Picco, Clare Pacini, Felicity Allen, Alessandro Vinceti, et al. 2021. “Minimal Genome-Wide Human CRISPR-Cas9 Library.” *Genome Biology* 22 (1): 1–14.

Gostimskaya, Irina. 2022. “CRISPR-Cas9: A History of Its Discovery and Ethical Considerations of Its Use in Genome Editing.” *Biochemistry. Biokhimiia* 87 (8): 777–88.

Greenman, Chris D., Graham Bignell, Adam Butler, Sarah Edkins, Jon Hinton, Dave Beare, Sajani Swamy, et al. 2009. “PICNIC: An Algorithm to Predict Absolute Allelic Copy Number Variation with Microarray Cancer Data.” *Biostatistics* 11 (1): 164–75.

Hanna, Ruth E., and John G. Doench. 2020. “Design and Analysis of CRISPR–Cas Experiments.” *Nature Biotechnology* 38 (7): 813–23.

Hara, Toshiro, Rony Chanoch-Myers, Nathan D. Mathewson, Chad Myskiw, Lyla Atta, Lillian Bussema, Stephen W. Eichhorn, et al. 2021. “Interactions between Cancer Cells and Immune Cells Drive Transitions to Mesenchymal-like States in Glioblastoma.” *Cancer Cell* 39 (6): 779-792.e11.

Hart, Traver, Kevin R. Brown, Fabrice Sircoulomb, Robert Rottapel, and Jason Moffat. 2014. “Measuring Error Rates in Genomic Perturbation Screens: Gold Standards for Human Functional Genomics.” *Molecular Systems Biology* 10 (7): 733.

Hart, Traver, Megha Chandrashekhar, Michael Aregger, Zachary Steinhart, Kevin R.

Brown, Graham MacLeod, Monika Mis, et al. 2015. “High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities.” *Cell* 163

(6): 1515–26.

Hart, Traver, and Jason Moffat. 2016. “BAGEL: A Computational Framework for Identifying Essential Genes from Pooled Library Screens.” *BMC Bioinformatics* 17 (April): 164.

He, Bing, Yao Xiao, Haodong Liang, Qianhui Huang, Yuheng Du, Yijun Li, David Garmire,

Duxin Sun, and Lana X. Garmire. 2023. “ASGARD Is A Single-Cell Guided

Pipeline to Aid Repurposing of Drugs.” *Nature Communications* 14 (1): 993.

Hsieh, Chiao-Yu, Jian-Hung Wen, Shih-Ming Lin, Tzu-Yang Tseng, Jia-Hsin Huang,

Hsuan-Cheng Huang, and Hsueh-Fen Juan. 2023. “ScDrug: From Single-Cell

RNA-Seq to Drug Response Prediction.” *Computational and Structural Biotechnology Journal* 21: 150–57.

Huré, Jean-Baptiste, Louis Foucault, Litsa Maria Ghayad, Corentine Marie, Nicolas

Vachoud, Lucas Baudouin, Rihab Azmani, et al. 2024. “Pharmacogenomic Screening Identifies and Repurposes Leucovorin and Dyclonine as ProOligodendrogenic Compounds in Brain Repair.” *Nature Communications* 15 (1):

9837.

Iannuzzi, Raffaele M., Ichcha Manipur, Clare Pacini, Fiona M. Behan, Mario R.

Guarracino, Mathew J. Garnett, Aurora Savino, and Francesco Iorio. 2024.

“Benchmark Software and Data for Evaluating CRISPR-Cas9 Experimental Pipelines through the Assessment of a Calibration Screen.” *The CRISPR Journal*, January. https://doi.org/10.1089/crispr.2023.0040.

“Iorio et Al.,.” 2016. *Cell* 166 (3): 740–54.

Iorio, Francesco, Fiona M. Behan, Emanuel Gonçalves, Shriram G. Bhosle, Elisabeth

Chen, Rebecca Shepherd, Charlotte Beaver, et al. 2018. “Unsupervised Correction of Gene-Independent Cell Responses to CRISPR-Cas9 Targeting.” *BMC Genomics* 19 (1): 1–16.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and

Emmanuelle Charpentier. 2012. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity.” *Science (New York, N.Y.)* 337 (6096): 816–21.

Johnson, Kevin C., Kevin J. Anderson, Elise T. Courtois, Amit D. Gujar, Floris P. Barthel,

Frederick S. Varn, Diane Luo, et al. 2021. “Single-Cell Multimodal Glioma Analyses Identify Epigenetic Regulators of Cellular Plasticity and Environmental Stress Response.” *Nature Genetics* 53 (10): 1456–68.

Joung, J. Keith, and Jeffry D. Sander. 2013. “TALENs: A Widely Applicable Technology for Targeted Genome Editing.” *Nature Reviews. Molecular Cell Biology* 14 (1): 49–

55.

Kalebic, Nereo, Carlotta Gilardi, Barbara Stepien, Michaela Wilsch-Bräuninger, Katherine

R. Long, Takashi Namba, Marta Florio, et al. 2019. “Neocortical Expansion Due to

Increased Proliferation of Basal Progenitors Is Linked to Changes in Their Morphology.” *Cell Stem Cell* 24 (4): 535-550.e9.

Kinker, Gabriela S., Alissa C. Greenwald, Rotem Tal, Zhanna Orlova, Michael S. Cuoco,

James M. McFarland, Allison Warren, et al. 2020. “Pan-Cancer Single-Cell RNASeq Identifies Recurring Programs of Cellular Heterogeneity.” *Nature Genetics* 52

(11): 1208–18.

Lampert, P. W., and R. L. Davis. 1964. “DELAYED EFFECTS OF RADIATION ON THE HUMAN CENTRAL NERVOUS SYSTEM; ‘EARLY’ AND ‘LATE’ DELAYED REACTIONS.” *Neurology* 14 (October): 912–17.

Lander, Eric S. 2016. “The Heroes of CRISPR.” *Cell* 164 (1–2): 18–28.

Lathia, Justin D., Stephen C. Mack, Erin E. Mulkearns-Hubert, Claudia L. L. Valentim, and Jeremy N. Rich. 2015. “Cancer Stem Cells in Glioblastoma.” *Genes & Development* 29 (12): 1203–17.

Lazar, Nathan H., Safiye Celik, Lu Chen, Marta M. Fay, Jonathan C. Irish, James Jensen,

Conor A. Tillinghast, et al. 2024. “High-Resolution Genome-Wide Mapping of Chromosome-Arm-Scale Truncations Induced by CRISPR-Cas9 Editing.” *Nature Genetics* 56 (7): 1482–93.

Lee, Il Hwan, and Dong-Kyu Kim. 2024. “Head and Neck Cancer: A Potential Risk Factor for Parkinson’s Disease?” *Cancers* 16 (13).

https://doi.org/10.3390/cancers16132486.

Li, Wei, Johannes Köster, Han Xu, Chen-Hao Chen, Tengfei Xiao, Jun S. Liu, Myles Brown, and X. Shirley Liu. 2015. “Quality Control, Modeling, and Visualization of CRISPR Screens with MAGeCK-VISPR.” *Genome Biology* 16 (1): 281.

Li, Wei, Han Xu, Tengfei Xiao, Le Cong, Michael I. Love, Feng Zhang, Rafael A. Irizarry, Jun S. Liu, Myles Brown, and X. Shirley Liu. 2014. “MAGeCK Enables Robust Identification of Essential Genes from Genome-Scale CRISPR/Cas9 Knockout Screens.” *Genome Biology* 15 (12): 554.

Loos, Maarten, Dustin Schetters, Myrthe Hoogeland, Sabine Spijker, Taco J. de Vries, and Tommy Pattij. 2016. “Prefrontal Cortical Neuregulin-ErbB Modulation of Inhibitory Control in Rats.” *European Journal of Pharmacology* 781 (June): 157–63.

“Loss of Cytoskeleton Protein ADD3 Promotes Tumor Growth and Angiogenesis in Glioblastoma Multiforme.” 2020. *Cancer Letters* 474 (April): 118–26.

Mansour, Heba Mohamed, Mahmoud Mohamed Khattab, and Aiman Saad El-Khatib. 2023. *Receptor Tyrosine Kinases in Neurodegenerative and Psychiatric Disorders*. Elsevier.

Mariani, L., C. Beaudry, W. S. McDonough, D. B. Hoelzinger, T. Demuth, K. R. Ross, T. Berens, et al. 2001. “Glioma Cell Motility Is Associated with Reduced Transcription of Proapoptotic and Proliferation Genes: A CDNA Microarray Analysis.” *Journal of Neuro-Oncology* 53 (2): 161–76.

Masopust, David, Christine P. Sivula, and Stephen C. Jameson. 2017. “Of Mice, Dirty Mice, and Men: Using Mice To Understand Human Immunology.” *Journal of Immunology (Baltimore, Md. : 1950)* 199 (2): 383–88.

Meer, Dieudonne van der, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E. Francies, and Mathew J. Garnett. 2018. “Cell Model Passports—a Hub for Clinical, Genetic and Functional Datasets of Preclinical Cancer Models.” *Nucleic Acids Research* 47 (D1): D923–29.

Mei, Lin, and Klaus-Armin Nave. 2014. “Neuregulin-ERBB Signaling in the Nervous System and Neuropsychiatric Diseases.” *Neuron* 83 (1): 27–49.

Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. “Computational Correction of Copy Number Effect Improves Specificity of CRISPR–Cas9 Essentiality Screens in Cancer Cells.” *Nature Genetics* 49 (12): 1779–84.

Munoz, Diana M., Pamela J. Cassiani, Li Li, Eric Billy, Joshua M. Korn, Michael D. Jones, Javad Golji, et al. 2016. “CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions.” *Cancer Discovery* 6 (8): 900–913.

Neftel, Cyril, Julie Laffy, Mariella G. Filbin, Toshiro Hara, Marni E. Shore, Gilbert J. Rahme, Alyssa R. Richman, et al. 2019. “An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma.” *Cell* 178 (4): 835-849.e21.

Pacini, Clare, Joshua M. Dempster, Isabella Boyle, Emanuel Gonçalves, Hanna Najgebauer, Emre Karakoc, Dieudonne van der Meer, et al. 2021. “Integrated Cross-Study Datasets of Genetic Dependencies in Cancer.” *Nature Communications* 12 (1): 1–14.

Papathanasiou, Stamatis, Styliani Markoulaki, Logan J. Blaine, Mitchell L. Leibowitz, Cheng-Zhong Zhang, Rudolf Jaenisch, and David Pellman. 2021. “Whole Chromosome Loss and Genomic Instability in Mouse Embryos after CRISPR-Cas9 Genome Editing.” *Nature Communications* 12 (1): 5855.

Patel, Anoop P., Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, et al. 2014. “Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma.” *Science (New York, N.Y.)* 344 (6190): 1396–1401.

Pellecchia, Simona, Gaetano Viscido, Melania Franchini, and Gennaro Gambardella. 2023. “Predicting Drug Response from Single-Cell Expression Profiles of Tumours.” *BMC Medicine* 21 (1): 476.

Peng, Da, Rachel Gleyzer, Wen-Hsin Tai, Pavithra Kumar, Qin Bian, Bradley Isaacs, Edroaldo Lummertz da Rocha, et al. 2021. “Evaluating the Transcriptional Fidelity of Cancer Models.” *Genome Medicine* 13 (1): 1–27.

Pérez-González, Andrea, Kevin Bévant, and Cédric Blanpain. 2023. “Cancer Cell Plasticity during Tumor Progression, Metastasis and Response to Therapy.” *Nature Cancer* 4 (8): 1063–82.

Poon, Ming-Wai, James Tin Fong Zhuang, Stanley Thian Sze Wong, Stella Sun, Xiao-Qin Zhang, and Gilberto Ka Kit Leung. 2015. “Co-Expression of Cytoskeletal Protein Adducin 3 and CD133 in Neurospheres and a Temozolomide-Resistant Subclone of Glioblastoma.” *Anticancer Research* 35 (12): 6487–95.

Rani, Sandhya B., Sachin Shivaji Rathod, Shanmuganandam Karthik, Navjot Kaur, Dattatraya Muzumdar, and Anjali S. Shiras. 2013. “MiR-145 Functions as a TumorSuppressive RNA by Targeting Sox9 and Adducin 3 in Human Glioma Cells.” *Neuro-Oncology* 15 (10): 1302–16.

Ravi, Vidhya M., Kevin Joseph, Julian Wurm, Simon Behringer, Nicklas Garrelfs, Paolo d'Errico, Yashar Naseri, et al. 2019. "Human Organotypic Brain Slice Culture: A Novel Framework for Environmental Research in Neuro-Oncology." *Life Science Alliance* 2 (4). https://doi.org/10.26508/lsa.201900305.

Shalem, Ophir, Neville E. Sanjana, Ella Hartenian, Xi Shi, David A. Scott, Tarjei S. Mikkelsen, Dirk Heckl, et al. 2014. "Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells." *Science* 343 (6166): 84–87.

Shergalis, Andrea, Armand Bankhead 3rd, Urarika Luesakul, Nongnuj Muangsin, and Nouri Neamati. 2018. "Current Challenges and Opportunities in Treating Glioblastoma." *Pharmacological Reviews* 70 (3): 412–45.

Sinha, Sanju, Rahulsimham Vegesna, Sumit Mukherjee, Ashwin V. Kammula, Saugato Rahman Dhruba, Wei Wu, D. Lucas Kerr, et al. 2024. "PERCEPTION Predicts Patient Response and Resistance to Treatment Using Single-Cell Transcriptomics of Their Tumors." *Nature Cancer* 5 (6): 938–52.

Sottoriva, Andrea, Inmaculada Spiteri, Sara G. M. Piccirillo, Anestis Touloumis, V. Peter Collins, John C. Marioni, Christina Curtis, Colin Watts, and Simon Tavaré. 2013. "Intratumor Heterogeneity in Human Glioblastoma Reflects Cancer Evolutionary Dynamics." *Proceedings of the National Academy of Sciences of the United States of America* 110 (10): 4009–14.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 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): 15545–50.

Suphavilai, Chayaporn, Shumei Chia, Ankur Sharma, Lorna Tu, Rafael Peres Da Silva, Aanchal Mongia, Ramanuj DasGupta, and Niranjan Nagarajan. 2021. "Predicting Heterogeneity in Clone-Specific Therapeutic Vulnerabilities Using Single-Cell Transcriptomic Signatures." *Genome Medicine* 13 (1): 1–14.

Suvà, Mario L., Esther Rheinbay, Shawn M. Gillespie, Anoop P. Patel, Hiroaki Wakimoto, Samuel D. Rabkin, Nicolo Riggi, et al. 2014. "Reconstructing and Reprogramming the Tumor-Propagating Potential of Glioblastoma Stem-like Cells." *Cell* 157 (3): 580–94.

Tan, Li Yang, Grace Cunliffe, Michael Patrick Hogan, Xin Yi Yeo, Chansik Oh, Bohwan

Jin, Junmo Kang, et al. 2024. “Emergence of the Brain-Border Immune Niches and Their Contribution to the Development of Neurodegenerative Diseases.” *Frontiers in Immunology* 15 (May): 1380063.

Tirosh, Itay, Andrew S. Venteicher, Christine Hebert, Leah E. Escalante, Anoop P. Patel, Keren Yizhak, Jonathan M. Fisher, et al. 2016. “Single-Cell RNA-Seq Supports a Developmental Hierarchy in Human Oligodendroglioma.” *Nature* 539 (7628): 309–13.

Toth, Nicholas, and Kathy Schick. 2018. “An Overview of the Cognitive Implications of the Oldowan Industrial Complex.” *Azania: Archaeological Research in Africa*, March, 3–39.

Trastulla, Lucia, Aurora Savino, Pedro Beltrao, Isidro Cortés Ciriano, Peter Fenici, Mathew J. Garnett, Ilaria Guerini, et al. 2023. “Highlights from the 1st European Cancer Dependency Map Symposium and Workshop.” *FEBS Letters* 597 (15): 1921–27.

Tsherniak, Aviad, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov, Glenn S. Cowley, Stanley Gill, et al. 2017. “Defining a Cancer Dependency Map.” *Cell* 170 (3): 564-576.e16.

Tsuchida, Connor A., Nadav Brandes, Raymund Bueno, Marena Trinidad, Thomas Mazumder, Bingfei Yu, Byungjin Hwang, et al. 2023. “Mitigation of Chromosome Loss in Clinical CRISPR-Cas9-Engineered T Cells.” *Cell* 186 (21): 4567-4582.e20.

Tzelepis, Konstantinos, Hiroko Koike-Yusa, Etienne De Braekeleer, Yilong Li, Emmanouil Metzakopian, Oliver M. Dovey, Annalisa Mupo, et al. 2016. “A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia.” *Cell Reports* 17 (4): 1193–1205.

Urnov, Fyodor D., Edward J. Rebar, Michael C. Holmes, H. Steve Zhang, and Philip D. Gregory. 2010. “Genome Editing with Engineered Zinc Finger Nucleases.” *Nature Reviews Genetics* 11 (9): 636–46.

Venteicher, Andrew S., Itay Tirosh, Christine Hebert, Keren Yizhak, Cyril Neftel, Mariella G. Filbin, Volker Hovestadt, et al. 2017. “Decoupling Genetics, Lineages, and Microenvironment in IDH-Mutant Gliomas by Single-Cell RNA-Seq.” *Science (New York, N.Y.)* 355 (6332). https://doi.org/10.1126/science.aai8478.

Vinceti, Alessandro, Riccardo Roberto De Lucia, Paolo Cremaschi, Umberto Perron, Emre Karakoc, Luca Mauri, Carlos Fernandez, Krzysztof Henryk Kluczynski, Daniel Stephen Anderson, and Francesco Iorio. 2023. “An Interactive Web Application for

Processing, Correcting, and Visualizing Genome-Wide Pooled CRISPR-Cas9 Screens.” *Cell Reports Methods* 3 (1): 100373.

Vinceti, Alessandro, Raffaele M. Iannuzzi, Isabella Boyle, Lucia Trastulla, Catarina D. Campbell, Francisca Vazquez, Joshua M. Dempster, and Francesco Iorio. 2024. “A Benchmark of Computational Methods for Correcting Biases of Established and Unknown Origin in CRISPR-Cas9 Screening Data.” *Genome Biology* 25 (1): 1–25.

Vinceti, Alessandro, Emre Karakoc, Clare Pacini, Umberto Perron, Riccardo Roberto De Lucia, Mathew J. Garnett, and Francesco Iorio. 2021. “CoRe: A Robustly Benchmarked R Package for Identifying Core-Fitness Genes in Genome-Wide Pooled CRISPR-Cas9 Screens.” *BMC Genomics* 22 (1): 828.

Vinceti, Alessandro, Umberto Perron, Lucia Trastulla, and Francesco Iorio. 2022. “Reduced Gene Templates for Supervised Analysis of Scale-Limited CRISPRCas9 Fitness Screens.” *Cell Reports* 40 (4): 111145.

Wagner, Clifford H. 1982. “Simpson’s Paradox in Real Life.” *The American Statistician* 36 (1): 46–48.

Wang, Qianghu, Baoli Hu, Xin Hu, Hoon Kim, Massimo Squatrito, Lisa Scarpace, Ana C. deCarvalho, et al. 2018. “Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment.” *Cancer Cell* 33 (1): 152.

Watson, J. D., and F. H. C. Crick. 1953. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38.

Weck, Antoine de, Javad Golji, Michael D. Jones, Joshua M. Korn, Eric Billy, E. Robert McDonald 3rd, Tobias Schmelzle, Hans Bitter, and Audrey Kauffmann. 2018. “Correction of Copy Number Induced False Positives in CRISPR Screens.” *PLoS Computational Biology* 14 (7): e1006279.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. “SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis.” *Genome Biology* 19 (1): 1–5.

Wolyniak, Michael J., Donna L. Pattison, Jay N. Pieczynski, and Maria S. Santisteban. 2025. *Introduction to CRISPR-Cas9 Techniques: Strategies for the Laboratory and the Classroom*. Springer Nature.

Yang, Wanjuan, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, et al. 2013. “Genomics of Drug Sensitivity in Cancer

(GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells.” *Nucleic Acids Research* 41 (Database issue): D955-61.

Yingsan, Fang. 1994. “The Pebble Tool or Chopper/Chopping Tool Industry in China.” *Human Evolution* 9 (4): 263–72.

Yoshihama, Maki, Tamayo Uechi, Shuichi Asakawa, Kazuhiko Kawasaki, Seishi Kato,

Sayomi Higa, Noriko Maeda, et al. 2002. “The Human Ribosomal Protein Genes:

Sequencing and Comparative Analysis of 73 Genes.” *Genome Research* 12 (3):

379–90.

Zahonero, Cristina, and Pilar Sánchez-Gómez. 2014. “EGFR-Dependent Mechanisms in Glioblastoma: Towards a Better Therapeutic Strategy.” *Cellular and Molecular Life Sciences : CMLS* 71 (18): 3465–88.

Zirngibl, Martin, Peggy Assinck, Anastasia Sizov, Andrew V. Caprariello, and Jason R.

Plemel. 2022. “Oligodendrocyte Death and Myelin Loss in the Cuprizone Model:

An Updated Overview of the Intrinsic and Extrinsic Causes of Cuprizone Demyelination.” *Molecular Neurodegeneration* 17 (1): 1–28.

Zuccaro, Michael V., Jia Xu, Carl Mitchell, Diego Marin, Raymond Zimmerman, Bhavini Rana, Everett Weinstein, et al. 2020. “Allele-Specific Chromosome Removal after Cas9 Cleavage in Human Embryos.” *Cell* 183 (6): 1650-1664.e15.