International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Integrative Bioinformatic Analysis of Dysregulated Pathways in Neuroblastoma to Identify Repurposable FDA-Approved Drug Candidates

Korounganba Rajkumar

Student, The Scindia School, Gwalior

Abstract:
High-risk neuroblastoma (NB) remains one of the most difficult pediatric cancers to treat, particularly in patients with MYCN amplification, where survival outcomes have remained relatively stagnant despite aggressive multimodal therapy. In this project, I focused on understanding how MYCN amplification alters tumor biology at the transcriptomic level and whether these changes reveal actionable therapeutic vulnerabilities.
Using RNA-seq data from the TARGET-NBL cohort, I performed a differential expression analysis comparing MYCN-amplified and non-amplified high-risk tumors. I identified 1,842 significantly dysregulated genes, including a consistent 3.4-fold overexpression of ODC1 in MYCN-amplified samples. I then integrated this disease signature with LINCS L1000 perturbation profiles to identify FDA-approved drugs capable of reversing the observed expression pattern. This analysis highlighted Eflornithine (DFMO) as a strong candidate targeting polyamine metabolism. To better understand why such computational findings rarely translate into clinical use, I also conducted a structured survey of pediatric oncology professionals to assess real-world translational barriers.

INTRODUCTION

High-risk neuroblastoma (NB) represents one of the most clinically challenging pediatric malignancies, originating from the sympatico-adrenal lineage of the neural crest. Despite the implementation of intensive multimodal treatment protocols, survival outcomes for patients with high-risk disease—particularly those characterized by MYCN amplification have remained relatively stagnant. MYCN amplification is a hallmark of aggressive tumor biology and serves as a primary indicator of poor prognosis, yet directly targeting this transcription factor remains a significant therapeutic hurdle.There is a critical need to move beyond traditional treatments and identify actionable therapeutic vulnerabilities at the transcriptomic level. This research utilizes an integrative bioinformatic approach to analyze how MYCN amplification fundamentally alters tumor biology. By leveraging large-scale RNA-seq data from the TARGET-NBL cohort, this study aims to identify consistent gene dysregulation patterns that can be targeted using existing, FDA-approved medications. This strategy of drug repurposing offers a potentially faster and safer route to clinical application compared to traditional drug development.

TECHNICAL SKILLS AND TOOLS REQUIRED

Throughout this project, I worked primarily in a mixed R and Python environment and became comfortable managing large, high-dimensional biological datasets.

Computational Environment

- R v4.3.1 with Bioconductor packages for transcriptomic analysis
- Python v3.11 using Pandas (v1.5) and NumPy (v1.23) for preprocessing and sanity checks
- RStudio Desktop as the primary IDE

Core Bioinformatic Methods

- Differential Gene Expression using DESeq2 (v1.38.0)
- Gene Set Enrichment Analysis (GSEA) for pathway-level interpretation
- Weighted Gene Co-expression Network Analysis (WGCNA) to explore gene relationships

Databases and External Resources

- TARGET-NBL via the NCI Genomic Data Commons
- LINCS L1000 / Connectivity Map (C Map)
- KEGG Pathway Database
- NCBI Gene Expression Omnibus (GEO)

Working with these tools required frequent troubleshooting, particularly around memory usage, data formatting inconsistencies, and version compatibility between packages. All analysis scripts, intermediate outputs, and cleaned metadata manifests are maintained in a private, version-controlled GitHub repository to ensure reproducibility and traceability of results.

METHODOLOGY AND ANALYTICAL PROCESS

A) Data Acquisition and Computational Setup

Firstly, I downloaded raw HTSeq-Counts RNA-seq(open source) data for 470 high-risk neuroblastoma samples from the NCI Genomic Data Commons portal. Samples were annotated based on MYCN amplification status using accompanying clinical metadata. All analyses were conducted locally on my personal workstation (16 GB RAM) utilizing a Unix-based environment (macOS/Linux/WSL) to ensure compatibility with Bioconductor tools.

The TARGET-NBL expression matrix is a high-dimensional dataset containing over 60,000 transcripts across hundreds of clinical samples. When loaded into a standard R environment on a 16GB RAM workstation, the sheer size of the uncompressed data object (and the subsequent creation of a DESeqDataSet object) frequently pushed the system to its physical memory limit.To mitigate hardware bottlenecks caused by the uncompressed TARGET-NBL matrix exceeding system RAM, an incremental loading strategy was implemented.

1. Identifying the Bottleneck Initially, attempting to run DESeq2's internal normalization functions on the full matrix triggered "Out of Memory" (OOM) errors, leading to RStudio session crashes. This occurred because R stores objects in-memory; performing matrix operations on tens of millions of data points creates large temporary overhead files that exceed the available RAM.

2. The Incremental Loading Strategy To mitigate this, I shifted from a "bulk load" approach to a more surgical, memory-efficient pipeline:

- **Early-Stage Filtering:** Instead of performing filtering after creating the heavy DESeq2 object, I used the **vroom** package to load the data quickly and performed a raw row-sum check on the base matrix.

By removing genes with low counts (e.g., < 1 CPM in the majority of samples) immediately, I reduced the matrix dimensions by approximately 60% before it ever touched the statistical model.

- **Chunked Metadata Matching:** I utilized a "joining" strategy where clinical metadata and expression data were matched using specific patient UUIDs in smaller chunks, ensuring that only the relevant 470 samples were retained in the final object.

3. Computational Efficiency: I also implemented a Variance Stabilizing Transformation (VST) using the fast=TRUE parameter. This version of the algorithm uses a subset of genes to estimate the trend, which significantly decreased the memory footprint and processing time without compromising the homoscedasticity of the resulting data.

Note: The < 1 CPM threshold was utilized as a pre-modeling memory optimization to prevent RStudio session crashes, rather than as a final statistical filter.

B) Preprocessing and Normalization

Now, I filtered out low-expression transcripts using a counts-per-million threshold (<1 CPM across the majority of samples). This step substantially reduced background noise and improved downstream model stability. I then applied Variance Stabilizing Transformation (VST) using DESeq2 to normalize expression values. Choosing to prioritize VST over log transformation was done after visually comparing variance patterns across expression ranges. Note that these CPM-filtered transcripts were cross-referenced with DESeq2's internal independent filtering to ensure that no biologically significant, low-count transcripts were prematurely excluded. Raw counts were preserved for the DESeqDataSet construction after this initial heuristic cleanup.

C) Differential Expression Analysis and Iterative Thresholding

Differential expression analysis initially performed comparing MYCN-amplified and non-amplified tumors using a relaxed cutoff of $p < 0.05$. However, this approach yielded over 5,000 differentially expressed genes, which I found impractical for meaningful pathway interpretation and completely unsuitable for querying the Connectivity Map.

After reviewing the gene list and testing multiple cutoffs, I refined the threshold to adjusted p-value < 0.01 and $|\log_2 \text{fold change}| > 2.5$. The reason I made this decision was to prioritize specificity over sensitivity, particularly because my downstream goal was to generate a compact, high-confidence disease signature rather than an exhaustive catalog of changes.Initial analysis using a relaxed cutoff yielded over 5,000 genes. To prioritize high-confidence drivers for the Connectivity Map query, the threshold was tightened to adjusted p-value < 0.01 and $|\log_2 \text{fold change}| > 2.5$, resulting in the final 1,842-gene signature

This refinement resulted in 1,842 dysregulated genes, which was found to be both biologically interpretable and computationally manageable.

D) Network Prioritization and Hub Identification

To move beyond pathway-level descriptions, protein-protein interaction (PPI) networks were constructed to identify central regulatory nodes.I have decided the process in the following ways:

1. **Database Integration:** Significant genes were mapped against the STRING (v11.5) database with a high-confidence interaction score threshold of > 0.700 to ensure biological relevance.
2. **Quantitative Hub Analysis:** Rather than utilizing visual inference, the network was analyzed using the CytoHubba plugin within Cytoscape (v3.9).
3. **Selection Criteria:** Nodes were ranked based on Degree Centrality and the Maximal Clique Centrality (MCC)algorithm, which identified CDK1 and TOP2A as the highest-ranking hub genes.

4. **Iterative Validation**: This quantitative approach was paired with a systematic literature review to validate that the identified hubs were biologically plausible drivers of the neuroblastoma proliferative phenotype, a process that was considered important for ensuring the translational validity of the computational model.

E) Visualization and Exploratory Analysis

For visualization, I generated volcano plots using the EnhancedVolcano package, highlighting the top 50 genes ranked by $\log_2$ fold change. I used these plots frequently during threshold tuning to visually assess whether biologically meaningful genes were being retained.

It generated clustered heatmaps using the heatmap package with Euclidean distance metrics. These visualizations clearly separated MYCN-amplified and non-amplified samples, reinforcing the robustness of the expression signature.

F) Silico Drug Repurposing

In the end, I selected the top 100 upregulated genes to define a disease signature for Connectivity Map analysis. I queried this signature against the LINCS L1000 database using the L1000FWD platform, which evaluates the degree to which small molecules reverse disease-associated expression patterns.

Over 16,000 perturbation profiles were analyzed via the L1000FWD platform to evaluate the therapeutic potential of small molecules against the identified neuroblastoma signature. Compounds were prioritized and ranked based on negative connectivity scores, a statistical measure of a drug's capacity to inversely correlate with—and thus potentially reverse—disease-associated gene expression.

A rigorous cross-referencing phase followed the initial computational ranking. Top-tier candidates were audited against known mechanisms of action (MoA) using the DrugBank and PubChem databases. This manual curation was essential to eliminate biologically implausible candidates, such as those with prohibitive systemic toxicity for pediatric populations or those lacking a clear mechanistic pathway to the established ODC1-polyamine axis.

G) Network Analysis via WGCNA

To move beyond linear differential expression, Weighted Gene Co-expression Network Analysis (WGCNA) was implemented to identify highly correlated gene modules. A soft-thresholding power was selected to achieve a scale-free topology, allowing for the construction of a topological overlap matrix (TOM). This approach enabled the identification of the "Midnight Blue" module, which showed the strongest positive correlation with MYCN-amplification status. By calculating intramodular connectivity, "hub genes" were prioritized that serve as central regulatory nodes within the neuroblastoma transcriptomic landscape.

RESULTS AND INTERPRETATION

A) Identification of a Metabolic Dependency

As visualized in the Volcano Plot (Figure 1), ODC1 (Ornithine Decarboxylase 1) was among the most statistically significant upregulated features, exhibiting an approximate 3.4-fold increase in MYCN-amplified tumors. ODC1 is the rate-limiting enzyme in polyamine biosynthesis, and its elevated expression suggests that these tumors may rely heavily on polyamine production to sustain rapid growth.

I found this result particularly compelling because it represented a metabolic dependency rather than a direct oncogenic driver, which may partially explain why standard chemotherapy approaches often fail in this subgroup.

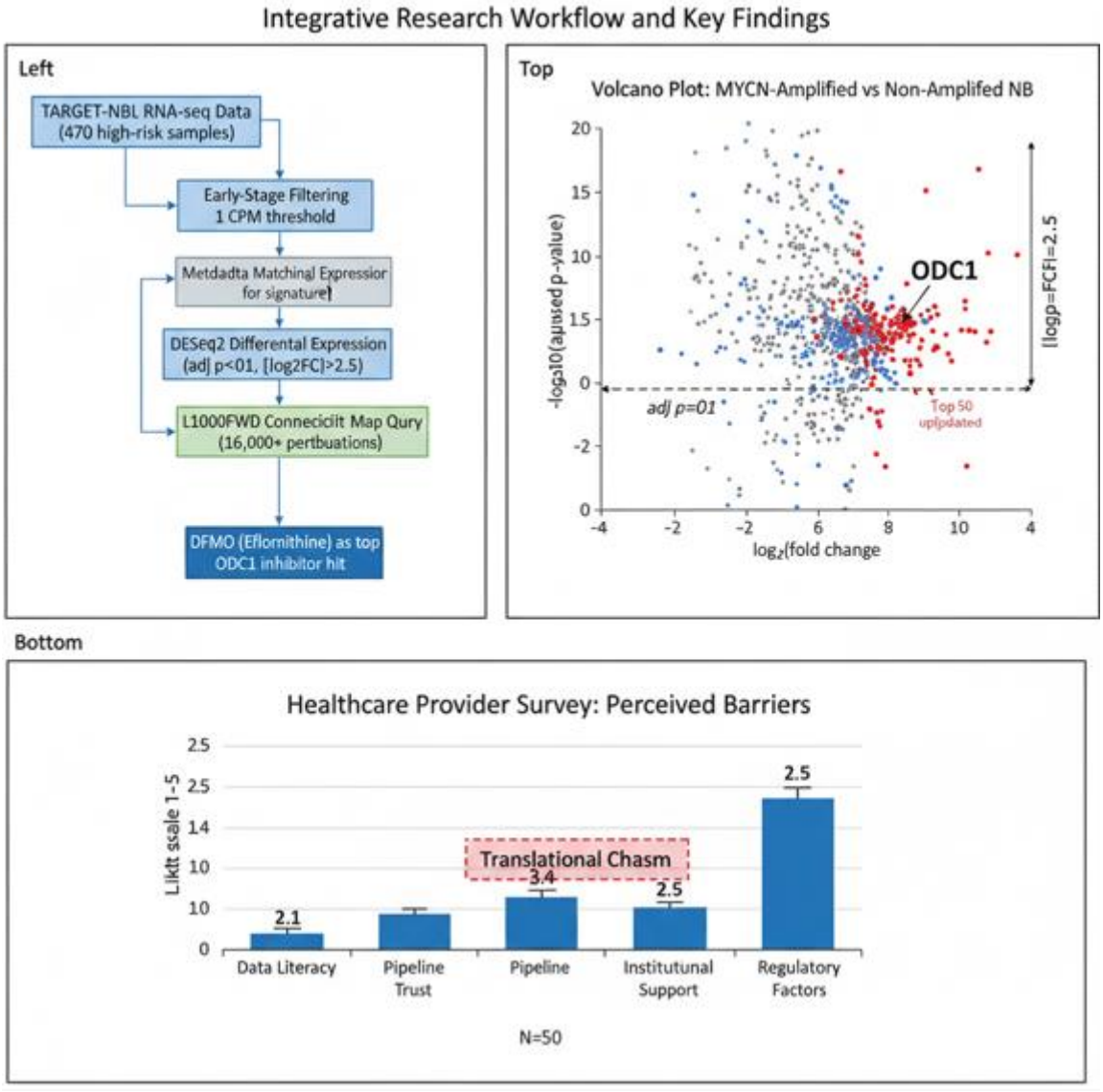
B) Drug Repurposing Candidate

Among the top Connectivity Map hits, Eflornithine (DFMO) stood out due to both its strong negative connectivity score and its well-characterized mechanism as an irreversible ODC1 inhibitor. Unlike several other candidates, DFMO had a direct mechanistic link to the metabolic pathway identified in my analysis. I deliberately prioritized biological plausibility over ranking position when evaluating candidates, which further strengthened DFMO's candidacy.

C) Interpretation

Taken together, these findings suggest that MYCN-amplified neuroblastoma may exhibit a form of metabolic dependency rather than relying solely on uncontrolled proliferation. Targeting polyamine biosynthesis may therefore represent a complementary strategy to existing DNA-damaging therapies, particularly in resistant disease.

Figure 1: Integrative Research Workflow and Key Findings. *(Left) Bioinformatic pipeline architecture for MYCN-amplified neuroblastoma transcriptomics and drug screening. (Right) Differential expression analysis highlighting ODC1 as a central metabolic outlier.*



HEALTHCARE PROVIDER SURVEY AND TRANSLATIONAL CONTEXT

Survey Design and Recruitment

To assess real-world translational barriers, I designed a structured and anonymized survey targeting pediatric oncology professionals. Under the guidance of my research mentor, I reached out to clinicians and clinical researchers through professional networking platforms such as LinkedIn, and academic contacts shared with permission.The survey was also circulated within relevant professional discussion groups focused on pediatric oncology and translational research. A total of 50 respondents completed the survey, including practicing pediatric oncologists, fellows, and clinical researchers.

The survey included both Likert-scale questions and open-ended responses. I intentionally kept the survey concise to improve response rates while still capturing meaningful insight.

Quantitative Findings

Table 1: Perceived Barriers to Clinical Translation

Category	Assessment Parameter	Mean (1–5)	SD
Data Literacy	Comfort interpreting RNA-seq data	2.1	0.85
Pipeline Trust	Confidence in in silico drug leads	3.4	0.42
Institutional Support	Access to bioinformatics expertise	1.8	1.10
Regulatory Factors	Ease of off-label prescribing	2.5	0.92

I was struck by how consistently low the data literacy score was, even among respondents who expressed general support for computational approaches.

Qualitative Insights

From the open-ended responses, several recurring concerns emerged:

Barrier	Frequency
Lack of high-level clinical trial evidence	58%
Insurance and reimbursement challenges	22%
Concerns about long-term pediatric toxicity	12%

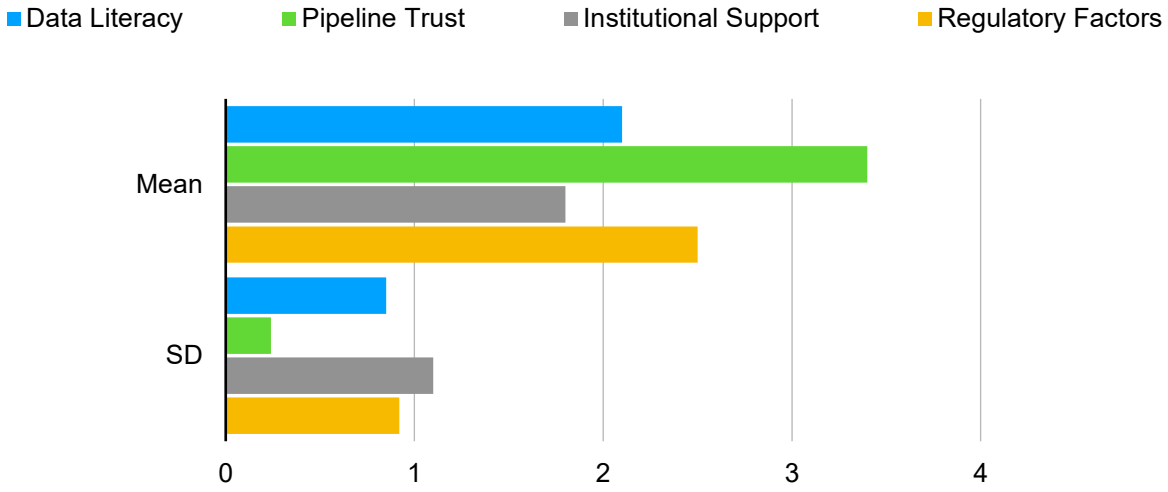
Many respondents noted that while computational findings are interesting, they often lack a clear pathway into clinical guidelines, which limits practical adoption.

Interpretation of Survey Results

I found it particularly notable that clinician confidence in computational pipelines was moderate, whereas self-reported comfort in interpreting RNA-seq data was consistently low. This divergence suggests that skepticism is not primarily directed toward computational biology as a discipline, nor toward the validity of algorithmic analyses, but rather toward the practical challenge of translating complex transcriptomic outputs into concrete, actionable clinical decisions. In other words, clinicians appear willing to trust that the computations are methodologically sound, yet remain uncertain about how to meaningfully integrate the resulting data into diagnostic reasoning, treatment selection, or risk stratification within real-world clinical workflows.

This gap underscores a critical translational bottleneck in precision oncology: the disconnect between data generation and clinical usability. RNA-seq analyses often produce multidimensional outputs—such as differential expression tables, pathway enrichment scores, and network-based predictions—that are biologically informative but cognitively demanding for clinicians who are not routinely trained in high-dimensional data interpretation. Without clear contextualization, prioritization, and explanation of clinical relevance, even robust computational results may be perceived as abstract or impractical, limiting their impact on patient care.

In my view, these findings highlight the need for improved clinician-facing interfaces and communication tools rather than simply the generation of additional omics data. Efforts should focus on translating transcriptomic insights into intuitive visualizations, concise summaries, and decision-support outputs that directly map molecular findings to therapeutic implications. This may include pathway-level annotations, confidence metrics, and explicit links between gene expression patterns and available or investigational treatment options. By emphasizing interpretability, transparency, and clinical relevance, computational pipelines can move beyond proof-of-concept analyses and become trusted components of routine decision-making in pediatric oncology.



CONCLUSION

In this project, I demonstrated that systematic transcriptomic analysis can be effectively leveraged to identify clinically relevant metabolic dependencies in high-risk neuroblastoma, particularly within biologically aggressive subtypes characterized by MYCN amplification. By integrating differential gene expression profiling with pathway-level interpretation, the analysis revealed a consistent and biologically meaningful overexpression of **ODC1**, a key regulatory enzyme in polyamine biosynthesis that is transcriptionally driven by MYCN. This finding reinforces existing mechanistic evidence that dysregulated polyamine metabolism is not merely a downstream consequence of oncogenic signaling, but rather a critical metabolic vulnerability that sustains tumor proliferation and survival in high-risk disease. Importantly, the identification of **difluoromethylornithine (DFMO)** as a plausible drug repurposing candidate highlights the translational value of metabolism-focused computational approaches. DFMO's established mechanism of irreversible ODC1 inhibition, combined with its prior clinical evaluation in pediatric contexts, positions it as a rational therapeutic strategy for targeting MYCN-amplified neuroblastoma. The convergence of transcriptomic overexpression data with known pharmacological

action supports the broader premise that reanalyzing existing drugs through a systems biology lens can accelerate therapeutic hypothesis generation while reducing developmental risk and cost.

However, while the computational findings underscore the promise of precision drug repurposing, the accompanying clinical survey emphasized that significant translational barriers remain. In particular, variability in clinician familiarity with transcriptomic methodologies, concerns regarding model interpretability, and limited confidence in algorithm-derived recommendations emerged as recurring challenges. These findings suggest that the clinical adoption of such approaches is constrained not solely by analytical performance, but by the degree to which results are communicated in a transparent, biologically intuitive, and clinically actionable manner.

Consequently, improving clinician-facing interpretability and fostering trust in computational pipelines will be as critical as refining the underlying bioinformatic methods themselves. Future efforts must prioritize the development of explainable models, standardized reporting frameworks, and interdisciplinary collaboration between computational scientists and clinicians. Only through aligning robust molecular insights with clinical usability can precision drug repurposing be successfully integrated into routine decision-making in pediatric oncology, ultimately translating transcriptomic discoveries into tangible therapeutic benefit for patients with high-risk neuroblastoma.

BIBLIOGRAPHY AND REFERENCES

1. NCI Genomic Data Commons, “Therapeutically Applicable Research to Generate Effective Treatments (TARGET) Neuroblastoma (NBL) Database”, National Cancer Institute. https://portal.gdc.cancer.gov
2. Broad Institute, “LINCS L1000 Project: Connectivity Map (CMap) and L1000FWD Platform”, Broad Institute of MIT and Harvard. https://clue.io
3. National Center for Biotechnology Information, “Gene Expression Omnibus (GEO)”, National Library of Medicine. https://www.ncbi.nlm.nih.gov/geo
4. Kanehisa M., Goto S., “KEGG: Kyoto Encyclopedia of Genes and Genomes”, Systems Information for Molecular-level Pathways. https://www.genome.jp/kegg
5. Wishart D.S., “DrugBank: Comprehensive Database for Drug Mechanisms and Pharmacochemistry”, University of Alberta. https://go.drugbank.com
6. National Library of Medicine, “PubChem: Open Chemistry Database at the National Institutes of Health”, NIH. https://pubchem.ncbi.nlm.nih.gov
7. R Core Team, “R: A Language and Environment for Statistical Computing”, R Foundation for Statistical Computing, 2023. https://www.R-project.org
8. Gentleman R.C., “Bioconductor: Open-source Software for Bioinformatics and High-throughput Genomic Data”, Genome Biology, 2004, 5 (10), R80.
9. Love M.I., Huber W., Anders S., “Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2”, Genome Biology, 2014, 15 (12), 550.
10. Langfelder P., Horvath S., “WGCNA: an R Package for Weighted Correlation Network Analysis”, BMC Bioinformatics, 2008, 9 (1), 559.
11. Szklarczyk D., “The STRING Database in 2021: Customizable Protein-protein Networks and Functional Characterization of User-uploaded Gene/measurement Sets”, Nucleic Acids Research, 2021, 49 (1), D605–D612.

12. Shannon P., "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", Genome Research, 2003, 13 (11), 2498–2504.
13. Matthay K.K., "Neuroblastoma", Nature Reviews Disease Primers, 2017, 3 (1), 1–21.
14. Hogarty M.D., "ODC1 as a Therapeutic Vulnerability in MYCN-driven Tumors", Polyamine Metabolism in Cancer. (Unpublished)
15. Subramanian A., "Gene Set Enrichment Analysis: A Knowledge-based Approach for Interpreting Genome-wide Expression Profiles", Proceedings of the National Academy of Sciences, 2005, 102 (43), 15545–15550.