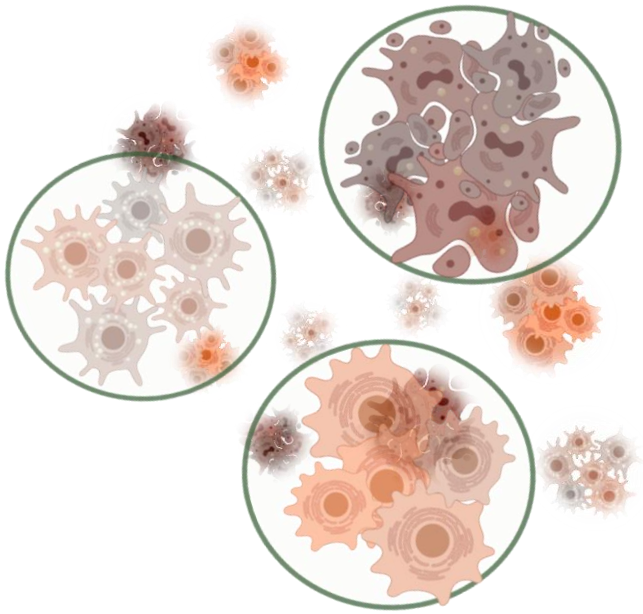
Thyroid Microenvironment Deconvolution – A Comparative Analysis with Single-Cell Reference Input in Cancer

Guilherme Vale Puga
Master in Bioinformatics and Computational Biology
Department of Computer Science | Department of Biology
Faculty of Sciences of the University of Porto - FCUP
2024



Thyroid Microenvironment Deconvolution – A Comparative Analysis with Single-Cell Reference Input in Cancer

Guilherme Vale Puga

Dissertation carried out as part of the Master in Bioinformatics and Computational Biology

Department of Computer Science | Department of Biology

2024

Professor Pedro G. Ferreira

Supervisor, Researcher at FCUP and INESCTEC

Professor José Carlos Machado

Co-supervisor, Researcher at I3S and IPATIMUP

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

INSTITUTO DE INVESTIGAÇÃO
E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

To my grandfather,

Thank you!

Acknowledgements

I would like to express my deepest gratitude to Pedro G. Ferreira for the opportunity to enrol in the Bioinformatics and Computational Biology Master's program at FCUP and whose advice, support, and insights were invaluable throughout this project and academic course. To my co-supervisor J. C. Machado and my course colleagues, I would also like to thank you for the valuable guidance and support.

I am also grateful to University of Porto for providing the resources and academic environment that made this research possible.

Special thanks to Inês Melo for her feedback, and encouragement during the course of this work. I would also like to show my heartfelt gratitude to my parents and brother, whose guidance and encouragement shaped me into who I am today. They have been a constant source of strength and inspiration during the most challenging moments of this project. Finally, I would like to extend my appreciation to the rest of my family and friends for their unwavering support and understanding throughout my academic journey.

Resumo

O cancro da tiroide (TC) é um dos tipos de cancro mais prevalentes e o seu panorama imunológico revela-se especialmente complexo. O microambiente tumoral (TME) desempenha um papel crucial na progressão e na resposta ao tratamento, sendo fundamental compreender a sua composição celular para identificar alvos terapêuticos e melhorar prognósticos. Neste estudo, comparamos vários métodos computacionais de desconvolução, incluindo abordagens de *machine learning* (ML) e estatísticas, para estimar as proporções de células imunes e estromais a partir de dados de sequenciação de RNA (RNA-seq) em massa. Aplicámos estes métodos tanto a tecidos saudáveis da tiroide (coorte GTEX) como a tecidos de TC (coorte GDC) para avaliar o seu desempenho em diferentes contextos. Através de análises de correlação com métodos de enriquecimento, como *xCell,* e do uso de genes marcadores estabelecidos, avaliámos a precisão de cada método na identificação de tipos celulares, como as células T CD4+, células T CD8+, células B, macrófagos e células NK.

O projeto recorreu a dados de RNA-seq de célula única como referência para orientar o processo de desconvolução e superar as limitações inerentes ao RNA-seq em massa.

Os resultados destacam a variabilidade de desempenho entre os métodos de desconvolução, com as ferramentas baseadas em ML, *Scaden* e *CIBERSORTx*, a se destacarem em termos de precisão e robustez. No entanto, métodos tradicionais, como *non-negative least squares* (*NNLS)* e *support vector regression NNLS* (*SVR-NNLS)*, também forneceram estimativas consistentes para determinados tipos celulares. Embora este estudo se tenha focado principalmente no TC, as abordagens e descobertas são aplicáveis a outros cancros, nomeadamente cancro gástrico ou do cólon, que poderão ser objeto de futuras investigações.

Foram encontrados desafios metodológicos, como erros de normalização e dicionários de genes marcadores incompletos, propondo-se estratégias para resolver essas limitações. Em última análise, este trabalho contribui para o avanço das abordagens de desconvolução e estabelece bases promissoras para o desenvolvimento de um sistema automatizado e análise futura de ecótipos tumorais, o que poderá ter implicações para terapias oncológicas personalizadas e prognósticos clínicos.

Palavras-chave: Sequenciação de RNA, Neoplasias da Tiroide, Microambiente Tumoral, Sequenciação de células únicas, Desconvolução de Microambiente Tumoral

Abstract

Thyroid cancer (TC) stands as one of the most incident cancer types and its immune landscape as one of the most complex. The tumour microenvironment (TME) plays a critical role in progression and treatment response, and understanding the cellular composition is vital for identifying therapeutic targets and predicting patient outcomes. In this study, we compared various computational deconvolution methods, including machine learning (ML) and statistical approaches, to estimate the proportions of key immune and stromal cell types from bulk RNA-sequencing data. We applied these methods to both healthy thyroid tissue (GTEX) and TC tissue (GDC) to evaluate their performance in different contexts. Through correlation analyses with *xCell* enrichment scores and the use of established marker genes, we assessed the accuracy of each method in identifying cell types such as CD4+ T-cells, CD8+ T-cells, B-cells, macrophages, and NK cells.

The project leveraged single-cell RNA sequencing data as a reference to guide the deconvolution process and address the limitations of bulk RNA sequencing.

The results highlight the variability in performance between deconvolution methods, with ML-based tools like *Scaden* and *CIBERSORTx* often outperforming others in terms of accuracy. However, traditional methods such as non-negative least squares (*NNLS)* and support vector regression followed by *NNLS* (*SVR-NNLS)* provided consistent estimates in certain cell types. While this study focused primarily on TC, the approaches and findings are applicable to other cancers, particularly gastric or colon cancer, which may be explored in future research.

Methodological challenges, such as normalization errors and incomplete marker gene dictionaries, were encountered, and strategies for addressing these limitations are proposed. Ultimately, this work contributes to advancing deconvolution approaches and sets the foundation for developing an automated pipeline and further analysis of tumour ecotypes, which could have implications for personalized cancer therapies and patient prognosis.

Keywords: Tumour Microenvironment, RNA Sequencing, Thyroid Neoplasms, Single-Cell Analysis, TME Deconvolution, Cell-Type Deconvolution

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A	
API	APPLICATION PROGRAMMING INTERFACE
C	
CAF	CANCER ASSOCIATED FIBROBLAST
CCA	CANONICAL CORRELATION ANALYSIS
CCC	CONCORDANCE CORRELATION COEFFICIENT
CTLA-4	CYTOTOXIC T-LYMPHOCYTE ASSOCIATED PROTEIN 4
CTS	CELL-TYPE-SPECIFIC
D	
DEG	DIFFERENTIALLY EXPRESSED GENE
E	
EM	EXPECTATION-MAXIMIZATION
EMT	EPITHELIAL-MESENCHYMAL TRANSITION
F	
FCUP	FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO
FTC	FOLLICULAR THYROID CARCINOMA
G	
GDC	GENOMIC DATA COMMONS
GEO	GENE EXPRESSION OMNIBUS
GTEX	GNOTYPE-TISSUE EXRESSION
H	
HUGO	HUMAN GENOME ORGANISATION
I	
ICA	INDEPENDENT COMPONENT ANALISYS
INESCTEC	INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES, TECNOLOGIA E CIÊNCIA
IPATIMUP	INSTITUTO DE PATOLOGIA E IMUNOLOGIA MOLECULAR DA UNIVERSIDADE DO PORTO
K	
KNN	K-NEAREST NEIGHBOUR
L	
LASSO	LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR
LLS	LINEAR LEAST SQUARES

LN	LYMPH-NODES
LS	LEAST SQUARES
M	
MAD	MEDIAN ABSOLUTE DEVIATION
MAE	MEAN ABSOLUTE ERROR
MCMC	MARKOV CHAIN MONTE CARLO
ML	MACHINE LEARNING
N	
NN	NEURAL NETWORKS
NK	NATURAL KILLER
NNLS	NON-NEGATIVE LEAST SQUARES
NMF	NON-NEGATIVE MATRIX FACTORIZATION
O	
OLS	ORDINARY LINEAR SQUARES
P	
PCA	PRINCIPAL COMPONENT ANALYSIS
PD-1	PROGRAMMED CELL DEATH PROTEIN 1
PD-L1	PROGRAMMED CELL DEATH PROTEIN 1 LIGAND
PTC	PAPILLARY THYROID CARCINOMA
R	
RMSE	ROOT MEAN SQUARED ERROR
RNA-SEQ	RNA-SEQUENCING
S	
SC	SUBCUTANEOUS METASTATIC LOCI
SCDNAM	SINGLE-CELL DNA METHYLATION
SCMD	SINGLE-CELL METHYLATION DECONVOLUTION
SCRNA-SEQ	SINGLE-CELL RNA-SEQUENCING
SCVI	SINGLE-CELL VARIATIONAL INFERENCE
SRT	SPATIALLY RESOLVED TRANSCRIPTOMICS
V-SVR | SVR	SUPPORT VECTOR REGRESSION
V	
VS	VERSUS
T	
TC	THYROID CANCER

TCGA	THE CANCER GENOME ATLAS
THCA	THYROID CANCER PROJECT
TME	TUMOUR MICROENVIRONMENT
TPM	TRANSCRIPTS PER MILLION
TREGS	REGULATORY T-CELLS
T-SNE	T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING
U	
UMAP	UNIFORM MANIFOLD APPROXIMATION AND PROJECTION
UMI	UNIQUE MOLECULAR IDENTIFIERS
UP	UNIVERSIDADE DO PORTO
W	
WLS	WEIGHTED LEAST SQUARES

1. Introduction

1.1. Thyroid Cancer and TME research

The thyroid gland plays a central role in regulating essential metabolic processes through hormone production, while also maintaining a close relationship with the immune system. Immune cells such as T cells, B cells, and macrophages are constantly present in thyroid tissue to preserve homeostasis and respond to infections or inflammation. Disruptions in this immune balance can lead to autoimmune conditions like Hashimoto's thyroiditis and Graves' disease, underscoring the importance of immune regulation in thyroid health (1).

This is especially relevant in thyroid cancer (TC), regarded as the most prevalent endocrine malignancy and the 5th most incident cancer in Portugal, in 2022 (2, 3). Particularly in the well differentiated TC types - papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma (FTC) - the tumour microenvironment (TME) becomes a key determinant of cancer progression (4). The TME is composed of a heterogeneous mixture of, immune, stromal, and cancer cells, blood vessels, and signalling molecules that interact in intricate ways to shape the biology of the tumour (5, 6). Immune cells within the TME can either promote or suppress tumour growth, depending on their infiltration levels, types, and expression of immune checkpoint molecules like programmed cell death protein 1 (PD-1) and cytotoxic T-lymphocyte associated protein 4 (CTLA-4). For example, regulatory T-cells (Tregs) often exhibit immune-suppressive activity in the TME, allowing tumours to evade immune surveillance, while cytotoxic CD8+ T-cells and natural killer (NK) cells are key players in antitumor immunity (1, 7).

The complexity of the TME extends beyond immune cells to include cancer-associated fibroblasts (CAFs), extracellular matrix components, and angiogenic factors that promote tumour growth and metastasis (8). These non-immune components of the TME can also act as barriers to effective treatment, making it essential to study both the cellular and molecular composition of the TME as a whole (4). Research has shown that an altered immune microenvironment is associated with poor prognosis in TC, and the presence of immune checkpoint molecules, such as PD-1 and PD-1 ligand (PD-L1), suggests that immune evasion plays a significant role in advanced TCs (9).

Therefore, investigating the cellular and molecular composition of the TME in TC is crucial for identifying potential therapeutic targets, including those involved in immune

modulation. The development of therapies such as immune checkpoint inhibitors has underscored the importance of understanding how immune cells within the TME contribute to tumour progression. By dissecting the components of the TME and their interactions, especially immune detection evasion, researchers can identify biomarkers of prognosis and therapeutic response, as well as uncover novel targets for precision medicine approaches in TC (5, 6, 10).

1.2. Computational approaches

Bulk RNA-sequencing (RNA-seq) has become a cornerstone in cancer research, providing a comprehensive outline of gene expression in tissue samples. For TC, it has enabled researchers to study global gene expression patterns and identify molecular signatures associated with tumorigenesis, metastasis, and immune evasion (11). However, one of the key challenges in analysing bulk RNA-seq data is that the results represent the average gene expression across all cells within a tissue, masking the heterogeneity of cell populations within the TME. While powerful, the lacking resolution of bulk RNA-seq makes the detection of the singular contributions of different cell types difficult, which can lead to challenges in interpreting gene expression changes that are critical to understand tumour biology and complex immune responses. This is especially relevant in TC, where the balance between immune-suppressive cells (e.g., regulatory T-cells, tumour-associated macrophages) and immune-activating cells (e.g., cytotoxic CD8+ T-cells, NK cells) can determine the progression and outcome of the disease (9, 11).

To overcome this limitation, computational deconvolution approaches have been developed to infer the relative proportions of different cell types in bulk RNA-seq data (11). These methods use reference gene expression profiles of known cell types (often derived from single-cell RNA-seq data) to estimate the cellular composition of heterogeneous tissue samples. By applying deconvolution techniques, researchers can untangle the TME in TC, identifying the specific contributions of immune, stromal, and cancer cells to the overall gene expression landscape (12-14).

Thus, deconvolution methods allow researchers to understand if the differences in gene expression data may stem from underlying differences in cellular proportions between

samples or from changes in the clinical condition. The problem can essentially be formulated with an equation, as follows:

$$\boxed{T = C * P}$$

- **T** represents a matrix with measured expression values from samples (M genes by N samples)
- **C** denotes the cell-type specific average expression value (M genes by K cell-types)
- **P** symbolizes the mixing proportions (relative composition)

As the equation above suggests, the level of expression of a gene in a heterogenous tissue can be transcribed as the linear combination of the expression values from each cell type in the mixture, with the assumption that each one has similar expression values across samples (15).

Deconvolution frameworks differ in complexity by the amount of system components they consider, ranging from simple two-component systems that separate heterogenous samples as tumour and non-tumour to more complex models that combine multiple cell types (16, 17).

Single-Cell RNA sequencing (scRNA-seq) has further refined the process of cell type deconvolution (18). Unlike bulk RNA-seq, which averages gene expression across all cells in a sample (**Figure 1**, step 2), scRNA-seq provides a high-resolution view of individual cell types (**Figure 1**, step 3). This granular data enhances the accuracy of deconvolution algorithms, allowing for a more precise estimation of cell type proportions along with uncovering previously unrecognized cellular populations (15, 19, 20).

In a more intricate analytical stage of the deconvolution pipeline (**Figure 1**, step 5), *in-silico* purification involves extracting cell-type-specific gene expression information from bulk RNA-seq data. This is achieved by leveraging the knowledge of cell-type-specific gene signatures obtained from scRNA-seq and isolating these signatures from the averaged bulk data. This step recovers the gene expression of individual cell types, even though they were originally mixed in the bulk tissue sample, allowing for a virtual dissection of the heterogeneous tissue sample into its constituent cell types, which enables a more detailed study of individual cellular roles and interactions within the TME (21).

Furthermore, the data derived from *in-silico* purification, may facilitate tumour ecotype classification (based on the cellular composition and gene expression patterns revealed through deconvolution) (**Figure 1**, step 6). Identifying different ecotypes can give insights

into the tumour's heterogenous cellular composition and allow identification of different subtypes, with divergent behaviour patterns and prognostic implications (22).

Different deconvolution problems require suitable formulations depending on the input data, whether it originates from microarrays, RNA-seq data or high-throughput sequencing (17). With the integration of scRNA-seq data an additional complexity layer and potential are introduced. In order to tackle a problem as mathematically complex as deconvolution of cell types in heterogenous samples, like those found in TC for example, a variety of approaches have been developed, each with its own methodology, input requirements and applications (3, 23).

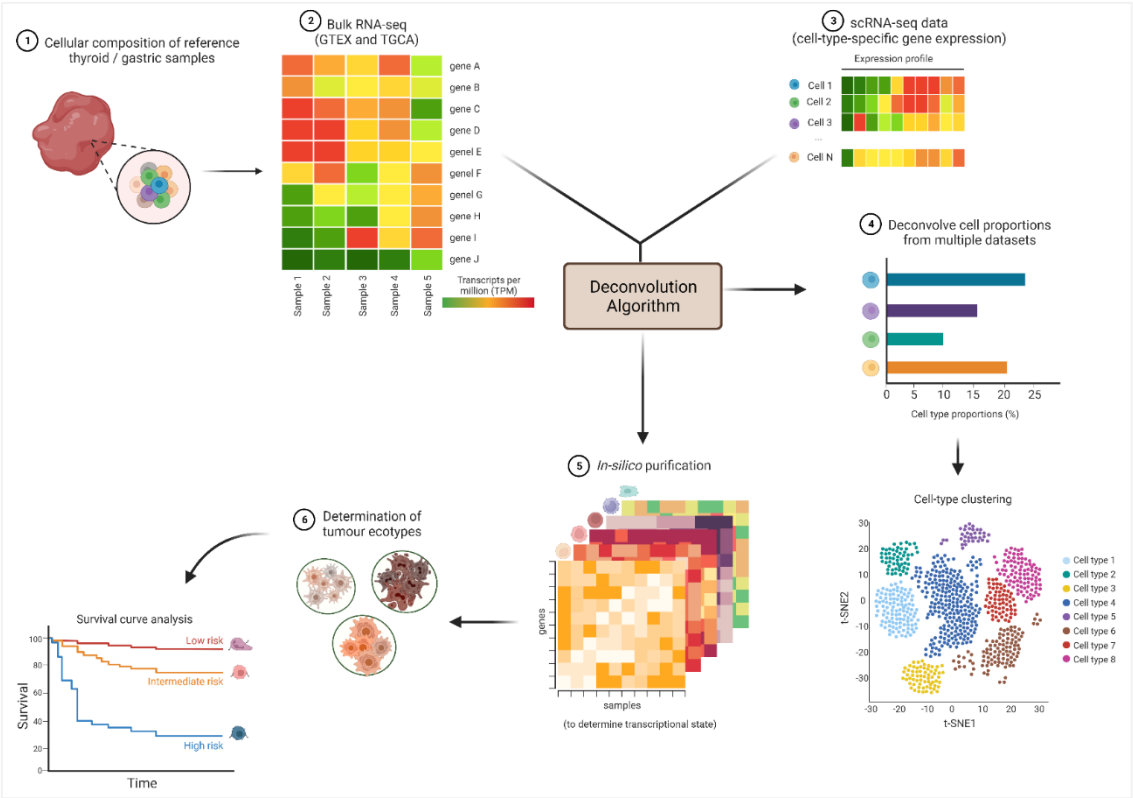


Figure 1: Workflow and deconvolution applications for tumour microenvironment research. The process begins with identification of the cellular composition of reference samples (1), followed by retrieval of bulk RNA-sequencing of datasets such as GTEX and GDC's TCGA (2). Single-cell RNA-sequencing (scRNA-seq) data is used to establish cell-type-specific gene expression profiles (3). Deconvolution algorithms estimate cell-type proportions (4), and downstream analysis include *in silico* purification (5) and leveraging metadata for determining tumour ecotypes and survival analysis to assess risk and therapeutic targets for TC prognosis (6).

1.3. Deconvolution Approaches

1.3.1. Least Squares Methods:

The goal when applying ordinary linear, linear, or simple least squares (*OLS*, *LLS*, and *LS*, respectively) methods is to minimize the Euclidean distance, which involves finding the smallest possible discrepancy between the observed bulk tissue gene expression profiles (C*P) and the estimated profiles based on the predicted cell type proportions (T). This method does not assume a specific distribution of the error terms, making it highly flexible for gene expression data. When the error terms follow a normal distribution, a maximum likelihood estimation approach can be employed for a more accurate prediction. A key advancement in the area is the recognition of the need for non-negative constraints to prevent nonsensical negative proportions. This has led to the rise of non-negative LS methods (*NNLS*), which ensure that all proportions are strictly positive and in total sum to one within each sample (24). Weighted LS methods (*WLS*) have been adapted from gene expression data to cell-type proportion estimation, incorporating adjustments like dampening constants that prevent infinite weights resulting from low cell-type proportions. Dampened WLS (*DWLS*) represents an important method to enhance rare cell types in the TME. This has proved to be highly effective in accurately estimating rare cell types (25, 26).

1.3.2. Support Vector Regression (v-SVR / SVR):

SVR methods, such as *CIBERSORT* and *CIBERSORTx*, have been refined to be more robust against noise and multicollinearity, involving the minimization of both a linear loss function and a L2-normalization function. These methods use predefined gene signatures specific to each cell type for deconvolution. This makes them particularly effective for complex samples like tumours where cell types are often correlated. The *CIBERSORTx* algorithm is an improvement of *CIBERSORT* and it is widely used for deconvoluting cell type proportions from bulk tissue gene expression data. It leverages a linear support vector regression framework to provide insights into cellular composition, particularly in TME due to its flexibility in adjusting for unknown mixtures (27-29).

1.3.3. Dimensionality Reduction Approaches:

These approaches treat the separation of heterogenous samples as dimensionality reduction problems with principal component analysis (PCA) being a widely used technique. PCA reduces high-dimensional data to its most essential components, simplifying cell type differentiation. However, recent studies have shown that factors beyond cell type identity, such as batch effects, technical noise, or biological variation, may significantly influence the explained variance. Consequently, more advanced methods, like Independent Component Analysis (ICA) or nonlinear approaches such as t-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), have emerged (30, 31). These techniques better preserve local data structure, offering more accurate cell type estimation, particularly for complex blends like the TME. Additionally, factor analysis and latent variable models have proven useful in isolating true biological signals from confounding technical factors, leading to a clearer separation of cell types in complex datasets (32).

1.3.4. Unsupervised Non-Negative Matrix Factorization (NMF) and Bayesian Approaches:

These methods estimate both gene expression profiles of pure cell populations (C) and mixing percentages (P) solely from expression data (T), without prior information. NMF factorizes T as the product of C and P, incorporating non-negativity constraints that ensure biologically meaningful estimates. Bayesian methods maximize a likelihood function with varying parameters and hyper-parameters, making use of techniques like Markov Chain Monte Carlo (MCMC) or expectation-maximization (EM) algorithms for tractability (33).

Both approaches are probabilistic, allowing them to effectively manage mixed cell populations, particularly in complex datasets like the TME, where cell heterogeneity is a significant challenge. For example, *BayesPrism* is a robust method that enables Bayesian integrative analysis across bulk RNA-seq and scRNA-seq. This method infers a joint posterior distribution of cell state proportion and gene expression. It utilizes Gibbs sampling and is particularly effective against technical batch effects and biological variation, making it robust in scenarios where there is substantial gene expression variation between bulk and reference data. *BayesPrism* has demonstrated high accuracy in various settings, including peripheral blood mononuclear cells and different cancer types, outperforming other methods in benchmarking tests (34-36).

InstaPrism, another notable method, is highly efficient in scenarios where reference data are unavailable, making it suitable for deconvolution without relying on pre-established cell type signatures. These methods are particularly useful for discovering novel cell types and subtypes in complex diseases, such as pancreatic ductal adenocarcinoma. Their ability to handle unknown cell types and mixed populations makes them invaluable in TME research, where identifying new cell subtypes can lead to a better understanding of tumour biology and therapeutic strategies (37).

1.3.5. Deep Learning Approaches:

Recently, emerging methods like *scpDeconv* have been introduced, employing deep learning-based techniques like autoencoders and domain adversarial architectures for neural networks (NN). This combination of techniques is tailored for proteomic data deconvolution, where the autoencoder leverages information from bulk RNA-seq data to enhance the quality of single-cell proteomic data, while the NN bridges the single-cell and bulk data distributions, transferring labels. This approach is particularly relevant for interpreting TME and clinical diagnosis/prognosis, highlighting the potential of deep learning in tackling the challenges associated with proteomic data deconvolution (38).

1.3.6. Machine Learning-based Algorithms:

These methods are key in tumour cellular deconvolution for dissecting the intricate composition of cancer tissues. Unsupervised learning algorithms like clustering are commonly used to identify patterns in the data without predefined labels, frequently revealing novel cell types or states within the TME. On the other hand, supervised learning algorithms require labelled training data but may deconvolve unknown tumour samples' cell proportions optimally given enough training information. Additionally, these methods may integrate RNA-seq data with other data types to improve deconvolution accuracy, such as proteomic (protein activity) or genomic (genetic variations) data. Semi-supervised learning approaches like *DSTG* and graph convolutional networks employ techniques like canonical correlation analysis (CCA) and k-nearest neighbour (KNN) algorithms to predict unknown cell proportions for the capture location (39, 40).

1.3.7. Advanced Statistical Methods:

xCell stands out in this category for its ability to integrate diverse cell type signatures, offering a comprehensive tool for cell type deconvolution in cancer research. Its application is crucial in characterizing the TME and immune landscape in various cancers (41).

Multi-task learning algorithms have shown superior performance over single-task algorithms in single-cell network inference. Methods like *scMTNI* and *MRTLE*, which are multi-task learning algorithms, outperform traditional single-task learning algorithms (like LASSO regression and *SCENIC*) in network inference from sparse datasets, such as single-cell transcriptomic data (42).

Reference-free deconvolution approaches like *STdeconvolve* have also been growing. These methods are particularly useful when a suitable single-cell transcriptomics reference is not available since they offer competitive performance to reference-based approaches and are less sensitive to differences in cell-type composition between the dataset to be deconvolved and the single-cell transcriptomics reference (34).

Algorithms like *spatialDWLS*, *SPOTlight*, *RCTD*, and stereoscope leverage cell type profiles learned from scRNA-seq data to decompose cell mixtures for spatial transcriptomics data. These methods use approaches such as *NNLS* regression, maximum likelihood estimation, and probabilistic modeling to infer cell type proportions at different capture locations (39, 43).

Tangram and *Cell2location* represent recent developments in spatial decomposition. *Tangram* is an optimization-based approach compatible with various spatial transcriptomics data, while *Cell2location* is a hierarchical Bayesian model that maps the spatial distribution of cell types. These methods have been systematically evaluated and reported to outperform other methods in detecting the presence of cell types across locations (39).

In the context of single-cell DNA methylation (scDNAm) data an innovative approach to deconvolution, single-cell methylation deconvolution (*scMD*) was proposed. This approach is particularly significant for tissues like brain, which lack cell-type references. *scMD's* ability to estimate cellular fractions has been validated in Alzheimer's disease-related cell types. It can deconfound tissue-level analyses and enable cell-type-specific (CTS) analyses (43, 44).

1.3.8. Comparative Evaluations and Benchmarking:

A recent high impact study conducted an extensive evaluation of deconvolution methods using three adult human brain datasets. The assessment compared algorithms like *CIBERSORT*, *DeconRNASeq*, *dtangle*, and *MuSiC*, along with enrichment scores using *xCell* and *BrainInABlender*. It was found that *CIBERSORT* generally showed the best performance based on both Pearson correlation coefficients and normalized mean absolute error values. However, the study also noted that enrichment methods were less accurate than partial deconvolution methods, with *xCell* showing particularly low accuracy (45, 46).

A bioRxiv preprint introduced a heterogeneous simulation strategy for more realistic benchmarking of cell-type deconvolution methods (15). This approach addresses the limitations of standard bulk simulation pipelines that generate synthetic bulk expression values lacking biological variance. Their findings indicated that marker-based methods and *BayesPrism* are most robust under heterogeneous scenarios, suggesting the need for further methodological development to accommodate unmodeled heterogeneity (47).

1.3.9. Ensemble Learning in Deconvolution:

A significant advancement in ensemble learning for cell type deconvolution is the development of *EnDecon*, a weighted ensemble learning method. *EnDecon* integrates results from multiple base deconvolution methods using a weighted optimization model, aiming to enhance accuracy in predicting cell-type compositions, particularly in spatially resolved transcriptomics (SRT) data. This approach was proven effective in distinguishing distinct spatial colocalization and enrichment patterns. It has been particularly valuable in analysing cancer tissues, like pancreatic ductal adenocarcinoma and breast cancer, offering nuanced views of cell-type compositions and their spatial distribution within these complex tissues (48).

Ensemble learning represents a significant step forward in the field of cell type deconvolution. Methods like *EnDecon* exemplify how combining multiple deconvolution techniques can lead to more accurate, robust, and insightful analyses, especially in complex biological datasets (48).

These methodologies represent significant strides in cell type deconvolution. They utilize advanced statistical models, leveraging scRNA-seq and bulk RNA-seq data to accurately estimate cellular proportions in complex tissues. However, choosing how to approach

the deconvolution problem depends on specific characteristics of the data and the research question. Supervised approaches, where either matrix C or P is available along with T, generally yield more accurate results compared to unsupervised methods, even though the latter are invaluable in scenarios where no prior information is available, since they can uncover novel insights into the cellular composition of tumours (21).

1.4. Comparative analysis

A comprehensive comparative analysis of deconvolution methods helps researchers evaluate the accuracy, reliability, and biases of different approaches in estimating cell type proportions. Some methods may excel at detecting rare cell types, while others might provide more robust estimates for more abundant cell types. By comparing multiple methods side by side (**Supplementary Table 1**), researchers can identify which approaches perform best in specific contexts, such as TC or healthy tissue deconvolution (17, 21). Additionally, this analysis allows for the identification of consistent patterns and discrepancies in cell type estimation, offering insights into the strengths and limitations of each method.

For example, methods like *CIBERSORTx* and *Scaden*, which leverage machine learning (ML), have demonstrated strong performance in estimating immune cell populations in TMEs, whereas traditional methods like *NNLS* or *SVR-NNLS* may struggle to capture the heterogeneity and complexity of tumour-associated cell populations (49). Comparative studies also help uncover potential biases, such as the tendency of certain methods to overestimate or underestimate specific cell types, or to be influenced by technical noise and batch effects present in the data (17, 20, 29). Given the diverse range of deconvolution methods available, the complexity of their implementation, and the need to adapt them to different datasets, there is a growing need for streamlined workflows and automated pipelines (19, 20). Currently, many deconvolution analyses require manual preprocessing steps, such as normalization, gene filtering, and selection of appropriate reference datasets. These processes are not only time-consuming but also introduce variability, as different researchers may apply different preprocessing strategies, leading to inconsistent results (17, 20).

By developing standardized, automated pipelines, researchers can ensure that deconvolution analyses are more reproducible, efficient, and scalable. Automated

pipelines can also help researchers handle large-scale datasets, such as those generated from the Cancer Genome Atlas (TCGA) or Genotype-Tissue Expression (GTEX) projects, which contain thousands of TC and healthy tissue samples (50, 51). These pipelines can then integrate key steps such as data preprocessing, normalization, selection of reference data, deconvolution, and downstream analysis, streamlining the entire workflow and enabling researchers to focus on interpreting the results.

TC presents specific challenges for deconvolution due to its unique immune and stromal composition. As previously suggested, tumour-associated immune cells, fibroblasts, and endothelial cells are among the key players in TC, but their proportions and roles may vary significantly between patients, making it difficult to generalize findings across studies (3, 5, 7, 9). Comparative benchmarking of deconvolution methods in TC is thus essential for understanding which tools are most suitable for analysing this specific tumour type (52-54). By benchmarking methods across datasets from different sources, such as the GTEX, for healthy thyroid tissue, and Genomic Data Commons (GDC), for thyroid cancer tissue, researchers can identify which deconvolution approaches provide the most consistent and biologically meaningful results for TC studies (20, 50, 51).

The ultimate goal of these efforts is to develop more robust, accurate, and reproducible deconvolution methodologies that can be applied to large-scale studies of the TME. This would allow researchers to generate more reliable insights into the cellular composition of tumours, enabling the identification of potential therapeutic targets, biomarkers, and prognostic indicators (17).

2. Objectives

2.1. Main goal

The main objectives of this dissertation are to evaluate and compare the performance of various deconvolution methods in the context of TME analysis for TC and enhance the understanding of TME composition by identifying the key immune and stromal cell types and their roles in cancer progression.

2.1.1. Specific goals

A] Evaluate deconvolution methods by assessing their performance across both healthy and cancerous tissue.

B] Analyse the TME composition across deconvolution methods.

C] Use deconvolution results and metadata to categorize tumours into distinct ecotypes.

D] Investigate the relationship between ecotypes and clinical factors to assess treatment response and overall patient outcomes.

E] Refine deconvolution methodologies and aim to develop an automated and scalable pipeline for TME analysis.

F] Benchmark cell type-specific performance with enrichment methods.

G] Investigate methodological limitations.

3. Materials and methods

3.1. Environment and Data management

The project was conducted using a cloud-based environment, primarily leveraging Google Colab for computational resources, which provided access to GPUs and high memory runtime. This setup facilitated the execution of resource-intensive tasks such as the large-scale data analysis and running deconvolution algorithms. It also enabled the integration of both Python and R, allowing the use of packages and tools from both programming languages to maximize the analysis pipeline's flexibility and capabilities. Google Drive was utilized for efficient data management, allowing seamless access to datasets, results, and backups across sections. Package installation and management were handled dynamically within Colab's environment, ensuring up-to-date and customized installations and essential libraries.

3.2. Population selection

For this study the bulk RNA-sequencing data was sourced from two major repositories: the GTEX project and the GDC database, specifically from TCGA (50, 51). The version of GTEX data selected was the GTEX's 'Analysis V8' release, which was the latest version by February 2024 and provided a wide range of normal tissues expression in a Transcripts per Million (TPM) format. This data was later filtered to contain only thyroid expression and to represent the cohort that would serve as baseline for non-cancerous gene expression profiles. The final cohort was represented by 673 samples (**Table 1**).

The TCGA data, on the other hand, provides comprehensive genomic, transcriptomic, and even epigenomic information from several cancer types. This data was filtered from the Thyroid Cancer (THCA) project in GDC to include only TC samples. From this project all open-access entries in the transcriptome profiling category with the experimental strategy RNA-seq were selected for the cancer cohort, for a total of 572 subjects as of February 2024.

Table 1: **Summary of datasets and type of data utilized in the study.**
Table summarizing the number of samples and number of genes before and after data processing, type of data (bulk or single-cell), origin the datasets and type of cohort.

Dataset	Number of samples / cells (raw vs pre-processed data)		Number of genes (raw vs pre-processed data)		Data type	Source	Cohort
GTEX	838 samples	653 samples	54,592	11,208	Bulk RNA-seq	GTEX Portal's Analysis V8	Healthy
GDC	572 samples		59,427	11,208	Bulk RNA-seq	TCGA's THCA project	Thyroid Cancer
Single-cell reference	91,231 cells	74,042 cells	33,694	11,208	scRNA-seq	GSE184362	Thyroid Cancer

3.3. Reference selection

The single-cell reference dataset was collected from Gene Expression Omnibus (GEO) with the accession number GSE184362 (23). This dataset contains 10X Genomics single cell data from 23 single-cell sequencing samples of eleven cancer patients, totalling 158.577 cells from seven primary tumours, six paratumours, eight metastatic lymph-nodes (LN) and two subcutaneous metastatic loci (SC). Due to computational and memory constraints only ten individual samples were selected (**Table 2**) for a total of 91.231 cells. The selected samples contain each of the categories described above with the intent of capturing an accurate reference of the TC's microenvironment, since it is often reprised of a diverse array of cancer cells, immune cells and stromal cells.

Table 2: **Samples in scRNA-seq reference dataset GSE184362**.
Samples highlighted in green represent the samples selected to construct the signature matrix. From these sample names, the sample ID (brown), patient (green), and specimen (red) information was retrieved for downstream analysis. In the table below # is a placeholder for each cell identifier.

GSE184362 Sample List:	
GSM5585102 Patient 1 tumour	GSM5585102_PTC1_T_#
GSM5585103 Patient 1 paratumour	GSM5585103_PTC1_P_#
GSM5585104 Patient 2 tumour	GSM5585104_PTC2_T_#
GSM5585105 Patient 2 paratumour	GSM5585105_PTC2_P_#
GSM5585106 Patient 2 left lymph node	GSM5585106_PTC2_LeftLN_#
GSM5585107 Patient 3 tumour	GSM5585107_PTC3_T_#
GSM5585108 Patient 3 paratumour	GSM5585108_PTC3_P_#
GSM5585109 Patient 3 left lymph node	GSM5585109_PTC3_LeftLN_#
GSM5585110 Patient 3 right lymph node	GSM5585110_PTC3_RightLN_#
GSM5585111 Patient 4 subcutaneous metastase	GSM5585111_PTC4_SC_#
GSM5585112 Patient 5 tumour	GSM5585112_PTC5_T_#
GSM5585113 Patient 5 paratumour	GSM5585113_PTC5_P_#

GSM5585114 Patient 5 right lymph node	GSM5585114_PTC5_RightLN_#
GSM5585115 Patient 6 right lymph node	GSM5585115_PTC6_RightLN_#
GSM5585116 Patient 7 right lymph node	GSM5585116_PTC7_RightLN_#
GSM5585117 Patient 8 tumour	GSM5585117_PTC8_T_#
GSM5585118 Patient 8 paratumour	GSM5585118_PTC8_P_#
GSM5585119 Patient 9 tumour	GSM5585119_PTC9_T_#
GSM5585120 Patient 9 paratumour	GSM5585120_PTC9_P_#
GSM5585121 Patient 10 tumour	GSM5585121_PTC10_T_#
GSM5585122 Patient 10 right lymph node	GSM5585122_PTC10_RightLN_#
GSM5585123 Patient 11 right lymph node	GSM5585123_PTC11_RightLN_#
GSM5585124 Patient 11 subcutaneous metastase	GSM5585124_PTC11_SC_#

3.4. Data preprocessing

Each selected deconvolution method made some suggestions on what the ideal approach should be to process the samples before generating the count matrices from the raw data, both from the bulk RNA-seq datasets and the scRNA-seq reference. This lack of consensus coupled with high dimensionality complex datasets lead to selecting more secure options as detailed below for the preprocessing steps common to all deconvolution approaches.

3.4.1. Bulk RNA-seq preprocessing

The gene expression data preprocessing started by loading the bulk RNA-seq datasets, reading the gene expression data and associated metadata. The GTEX cohort preprocessing steps included the extraction of thyroid-only sample data from the mix of tissues using its available metadata, while for the cancer cohort the data was downloaded via the GDC's application programming interface (API) and processed in batches to handle the larger file sizes. Next, the datasets were reshaped to ensure consistency with the healthy cohort's structure and guarantee both are compatible for downstream analysis. Both cohorts were henceforth applied the same preprocessing steps, including comparing gene identifiers, swapping the *Ensembl* identifiers with the official gene symbol (Human Genome Organisation nomenclature - HUGO) for each gene, summing the expression of isoforms, and ensuring all genes were consistent and aligned across cohorts. The latter ensured the genes unique to either source were effectively removed and won't be considered for downstream analyses, which was

important for avoiding biases in the comparative analysis with the detriment of not encapsulating the whole tumour microenvironment.

Batch effect removal with *ComBat*, and posterior PCA analysis (**Figure 2**), was tested at this stage but ultimately not applied to the bulk datasets for three reasons: 1) applying batch effect removal after incorporating the scRNA-seq data could allow a more comprehensive effect correction across all datasets; 2) it could potentially remove some important biological variations that could be key for an effective deconvolution of the heterogenous microenvironment of the thyroid tissue; 3) some deconvolution methods address this as part of their algorithms, meaning that adopting this approach could generate a significant overcorrection (17, 31, 55).

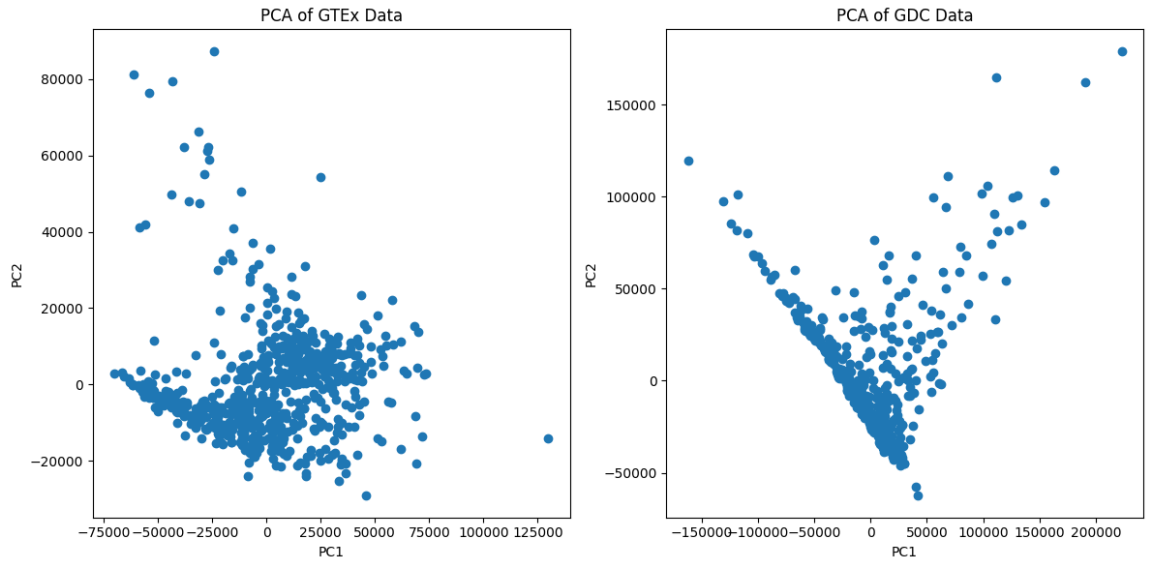


Figure 2: **Principal Component Analysis (PCA) of the GTEX cohort (left) and GDC cohort (right).**
Each point represents a sample, with position determined by the first two principal components (PC1 and PC2), which explain most of the variance. GTEX samples seem more spread-out, possibly due to the inherent biological variability in the normal thyroid tissue samples, while GDC samples show a linear pattern with a sharp spread along the x-axis (PC1), reflecting significant heterogeneity in the cancer tissues, or it can also be indicative of untreated batch effects.

3.4.2. scRNA-seq preprocessing

The single-cell data downloaded from GEO was processed entirely in python within the same environment using *Scanpy, Anndata, scVI-tools*, and *doubletdetection*. Starting from the TAR file, the data was missing a gene expression column which was added to ensure compatibility with *cellbender*, a tool for ambient RNA removal, which revealed the scRNA-seq data was previously filtered (56-58).

The individual 10x Genomics data were then loaded into several *AnnData* objects with *Scanpy*, each processed individually to endure proper handling and annotation. A

function was defined to extract relevant information from the sample ID which would be crucial for downstream analysis, especially in the visualization of clusters after integration and dimensionality reduction using single-cell variational inference (*scVI*) and UMAP (31, 57).

Cells expressing fewer than 200 gene were filtered out since they likely presented dying cells or empty droplets (57). Additionally, some quality control metrics were calculated, namely:

i. **Mitochondrial content**: high values may indicate stressed or dying cells. This was expected since it is a cancer dataset but calculated regardless (**Figure 3**).
ii. **Ribosomal content**: this can vary depending on cell state and could be useful later on for downstream analysis where we compare different cell states.
iii. **Haemoglobin content**: may indicate red blood cell contamination.
iv. **Number of genes by counts**: Total number of detected genes in a cell (**Figure 4**).
v. **Total number of counts**: total unique molecular identifiers (UMI) detected in a cell.
vi. **Percentage of counts in the top 20 genes**: a measure of gene expression dominance that would be useful for cell type annotations.

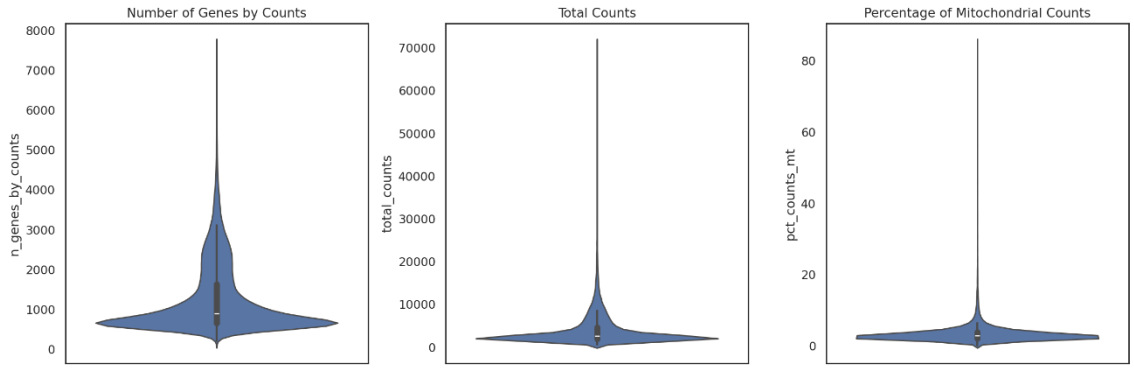


Figure 3: Violin plots displaying the distribution of the number of detected genes per cell across the reference dataset (left), the distribution of the total number of RNA counts per cell (middle plot), and the distribution of mitochondrial gene expression in each cell (right). The plots highlight the variability, overall transcriptional activity and percentage of stressed or dying cells in the reference data, respectively.

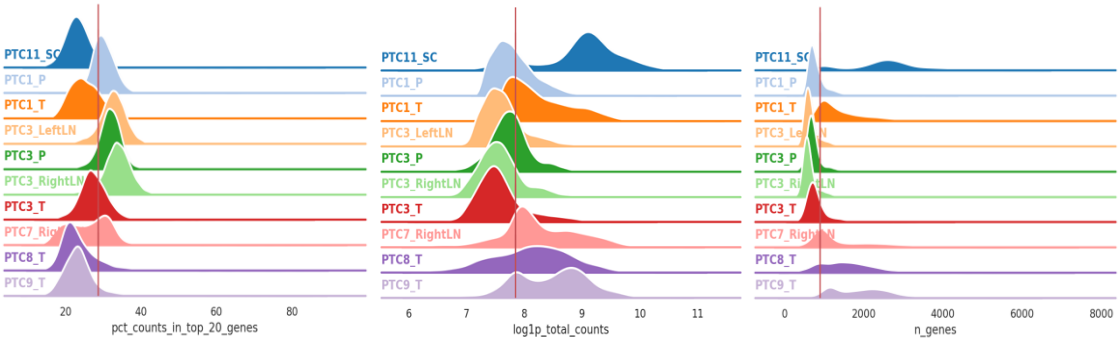


Figure 4: Density plots illustrating the proportion of total counts derived from the top 20 highly expressed genes by each sample (left), the distribution of total counts (log-transformed) per cell (middle), and the number of genes detected per cell (right). These density plots offer insights into the transcriptional activity and diversity across samples.

Outliers were removed by calculating the median absolute deviation (MAD) and setting a threshold for five MADs above the median. Cells five MADs below the median were not removed as they could represent meaningful rare cell types. The distributions of the gene counts were plotted pre- and post-processing for comparison (**Figure 5**).

In regard to normalization algorithms, several were explored before settling for library-size normalization to 10.000 UMI as it is simple to implement and a very effective strategy:

i. **Shifted logarithm (Delta method)**: fast normalization technique that applies a non-linear function to the raw counts; particularly effective when combined with PCA to uncover the latent structure of the data.
ii. **Scran's pooling-based size factor estimation**: uses a deconvolution approach to estimate the size factors based on linear regression over genes for pools of cells; aims to account for differences in count depths across samples.
iii. **Analytic Pearson residuals**: models count depth as a covariate in a generalized linear model, removing sampling effects, while preserving heterogeneity; complex and requires a large amount of memory for a heavy dataset such as the one selected.

This essential step was to account for differences in sequencing depth between cells and samples and was followed by a log1p transformation, while preserving the normalized counts for use with deconvolution methods that required raw count data.

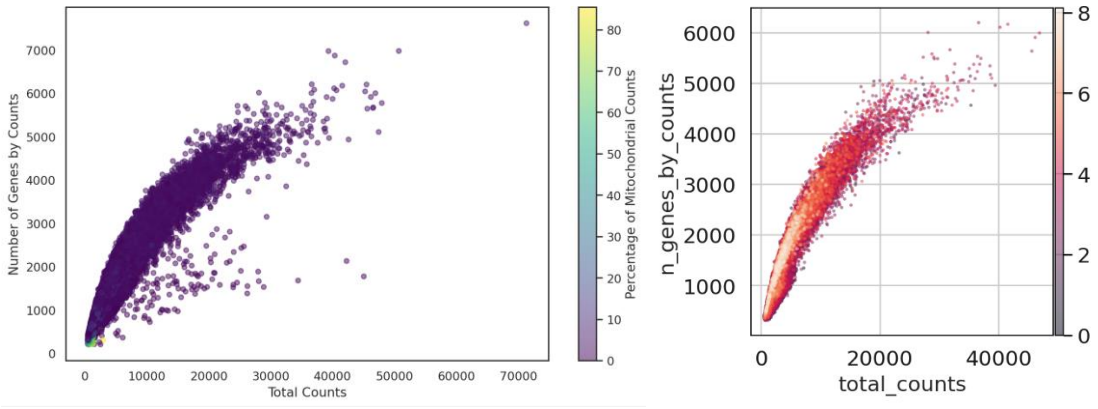


Figure 5: **Number of detected genes per sample, pre- (left scatter plot) and post- (right scatter plot) preprocessing.** The left scatter plot shows the number of detected genes per sample as a function of total counts for each cell before preprocessing. The colour intensity reflects the percentage of mitochondrial counts, with brighter regions indicating a higher proportion of mitochondrial genes; The right scatter plot displays the counts post quality control and filtering, exhibiting a similar relationship between the number of genes and total counts but with reduced mitochondrial percentages and fewer outliers, which indicates a successful preprocessing and cell filtering. Colour intensity corresponds to mitochondrial counts but with a compressed scale due to removal of cells with high mitochondrial content.

In order to detect doublets, which occur when two or more cells are captured in the same droplet during scRNA-seq, conceiving an artificial cell with a mixed transcriptome, the package *doubletdetection* was employed. This ML approach based on a boost classifier was trained on the gene expression data and used to predict doublets based on scores calculated for each cell after clustering with the Louvain algorithm. Cells not meeting the specified voting and probability thresholds of 0.5 and 1^{-16}, respectively, were removed from the datasets. The thresholds were set as recommended by the package developers (58).

Running *scVI*, a deep learning method for data integration the individually processed *anndata* objects were then incorporated into a single h5ad file. This step aimed to remove batch effects while preserving biological variation, allowing for an accurate comparison between the different samples and datasets (57).

The Wilcoxon rank-sum test was then used as a differentially expressed gene (DEG) analysis to identify marker genes for different cell types (56).

3.5. Cell type clustering and annotation

The immune landscape of the thyroid is complex (**Figure 6**), encompassing closely related cell types with similar gene expression profiles and tumour heterogeneity only deepens this complexity.

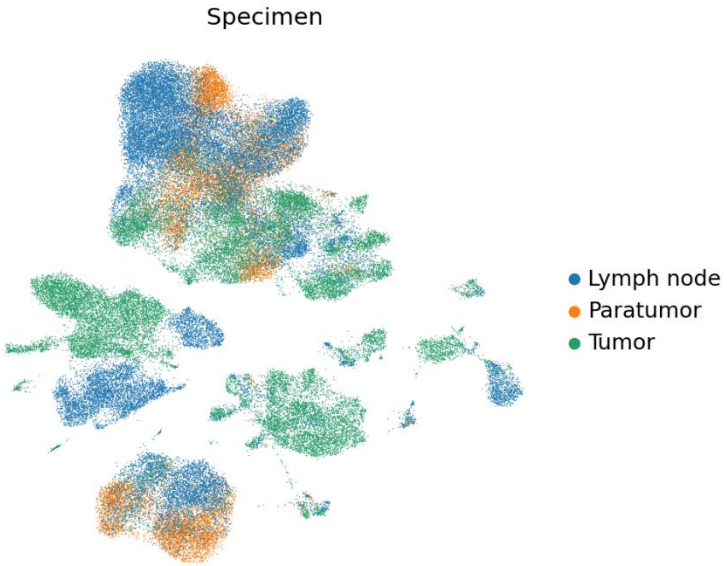


Figure 6: **UMAP plot representing the distribution of cells from each tissue in a lower-dimensional space according to their source.** The colour represents the sample type – blue for lymph nodes, likely containing a variety of cells important in immune surveillance; orange for paratumour, which is expected to contain a mix of fibroblasts, epithelial cells and immune cells due to proximity to tumour tissue; and green for tumour, where we expect to find a variety of immune cells and possibly cancer associated fibroblasts. Note: Tumour cells represented also include the cells from metastatic tissue.

Adding to this, memory and storage constraints in the workspace resulted in this step being conducted several times with different packages. Below is a detailed report of the first and last attempts, using different approaches (57, 59-61).

3.5.1. Manual cell type annotation with online databases

The Leiden algorithm is a graph-based clustering method that build upon the previously employed Louvain algorithm, aiming for higher resolution and more stable clustering (62). For both instances, a moderate resolution of 0.7 was employed, resulting in 25 cell type clusters for this first attempt which were later annotated as 23 putative cell types using the marker dictionary curated from genes, ranked using *Scanpy*, with significant

adjusted p-values and high log fold-change (**Figure 7**). This task was supported by online databases with known cell markers such as *Panglao DB* and *David DB* (56, 60, 61).

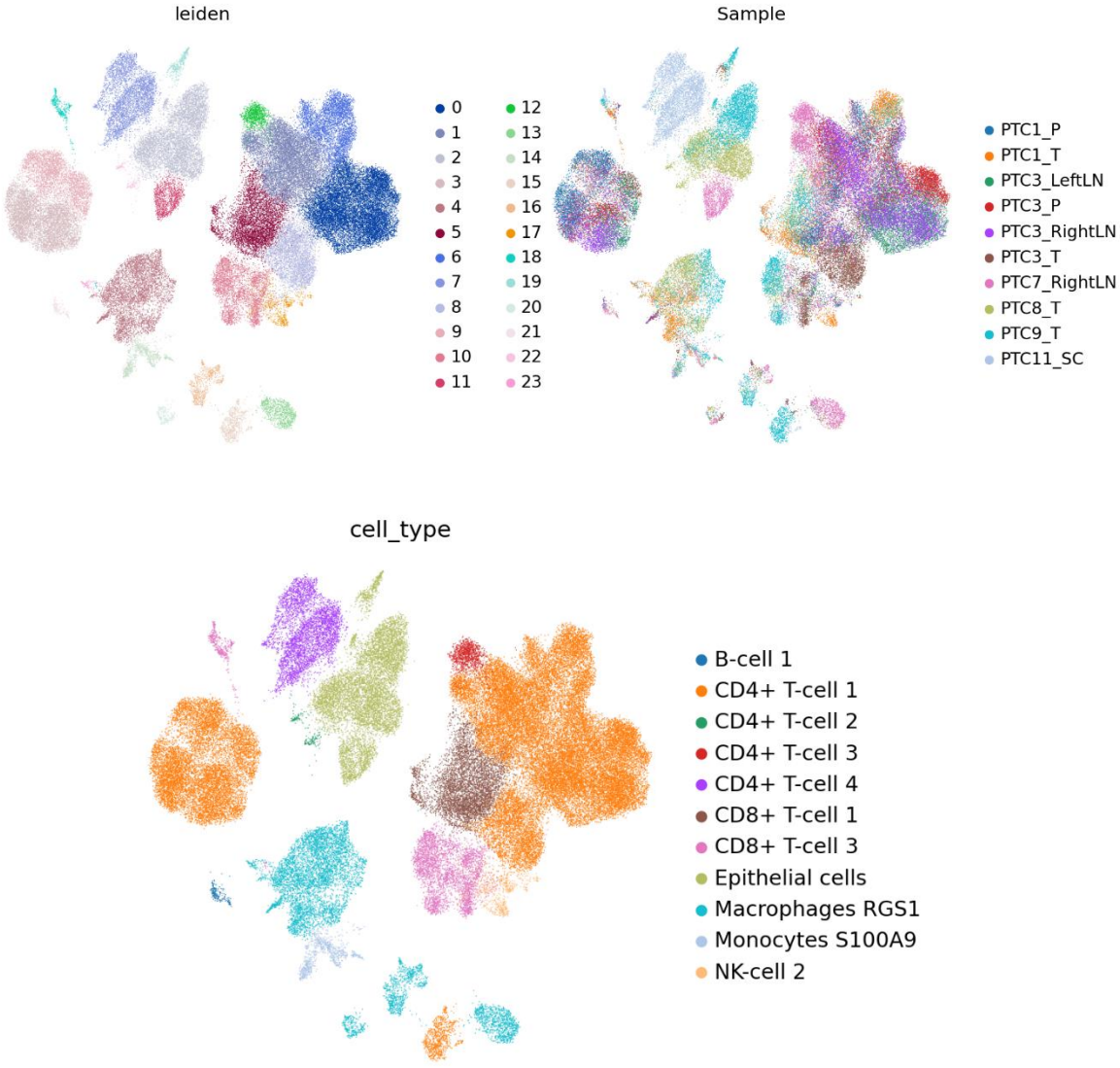


Figure 7: **Single-cell annotation and clustering**.
Top left corner (Leiden clustering): Cells are clustered using the Leiden algorithm based on their transcriptional profiles. Different clusters are represented by distinct colours, labelled from 0 to 23; Top right corner (Sample origin): The same cells labelled based on tissue of origin, with colours representing various patient sample, including tumour (T), paratumour (P), lymph nodes (LN) and metastatic (SC) samples; Bottom panel (Cell type annotation): each cluster is assigned a cell type based on marker gene expression. Colour labels indicate the various cell types annotated.

3.5.2. Automatic cell type annotation with celltypist and GSE184362 metadata

A faster and more systematic cell type annotation approach was explored later on using *celltypist*, however it lacked a thyroid reference dataset which resulted in selecting a set of marker genes from a combination of those utilized by Pu et al (7) for the same dataset and *scVI*'s DEG results (**Table 3**). This approach relies entirely on the accuracy of the marker gene dictionary, elaborated on the assumption that the cell types in the

heterogeneous selected mixture of 91.231 cells were likely the same as the ones present in the whole data, which included an extra 67.346 cells.

To lessen the impact of this assumption a score was calculated for each putative cell type within each cluster based on average expression of the corresponding marker genes, and then the highest scoring cell type was assigned to the group, settling for a total of 11 different cell types (**Figure 8**). The scoring function used provided a quantitative measure of how well a cluster's gene expression profile matches the predefined marker gene sets for different cell types.

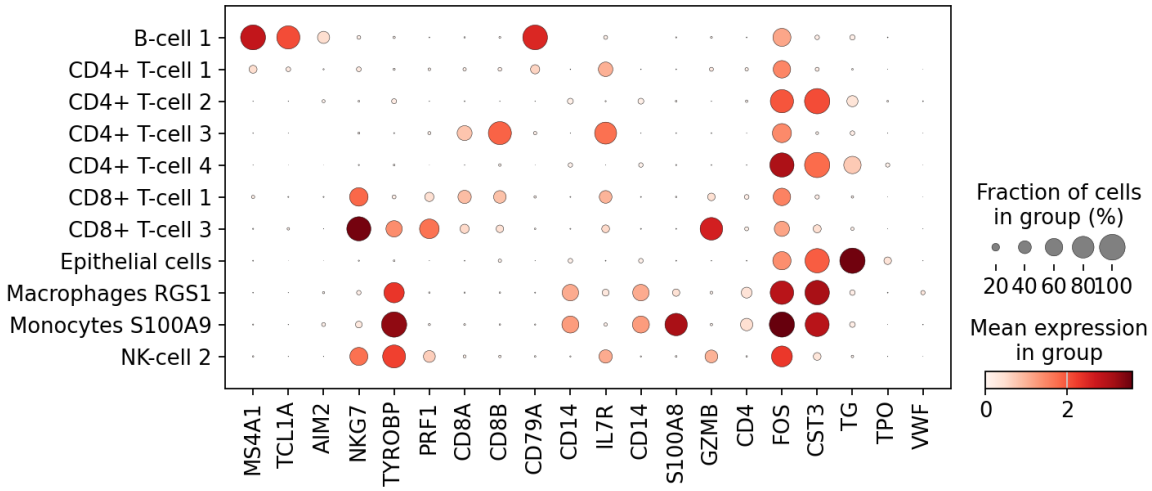


Figure 8: **Dotplot of Marker Genes across Cell Types in scRNA-seq Reference Data**.
Each dot represents the average expression of a marker gene in a specific cell type. The intensity of the colour indicates the level of expression while the size represents the percentage of cells within a cell type expressing that gene. By reading the plot one row at a time and comparing the genes' expression for each cell type we can get a good sense of the gene signatures for each type and subtype of the cells depicted. Gene signatures between subtypes are often similar and preventing inaccurate representation of each cell type is a major challenge the deconvolution approaches face, acknowledging the need for a good scRNA-seq reference matrix when available.
Note: A more comprehensive dotplot of the marker genes for each cell type and their levels of expression can be found in the Attachments section.

Following this final attempt and to further refine and add validation a manual inspection of the specific genes in each cluster ensued using UMAP visualizations.

Table 3: **Marker gene dictionary curated from (Pu et al, 2021) markers in addition to *PanglaoDb* and *DavidDb*.** Each gene selected represents a gene specific for each cell type according to the databases consulted and according to the p-values and log fold-changes consulted after applying DEG analysis to the obtained clusters.

Label	Key genes
CD4+ T-cell 1	*CD3D, CD3E, CD3G, MT1G, MT1X, MT1E, MT2A, MT1F*
CD4+ T-cell 2	*CD3D, CD3E, CD3G, KLRB1, IL7R, ANXA1, CD40LG, S100A11*
CD4+ T-cell 3	*CD3D, CD3E, CD3G, NOSIP, RGS10, LDHB, TCF7, CCR7*
CD4+ T-cell 4	*CD3D, CD3E, CD3G, HSPA1B, DNAJB1, HSPA1A, JUN, FOS*
CD4+ T-cell 5	*CD3D, CD3E, CD3G, ICA1, TOX2, CD200, CTLA4, TIGIT*
CD4+ T-cell 6	*CD3D, CD3E, CD3G, FOXP3, PDCD1, TNFRSF4, TNFRSF18, IL32, TIGIT, CTLA4*
CD4+ T-cell 7	*CD3D, CD3E, CD3G, PLCG2, IFI44L, MX1, IFI6, STAT1*
CD8+ T-cell 1	*CD3D, CD3E, CD3G, GZMK, CCL4, CCL4L2, GZMA, CCL5*
CD8+ T-cell 2	*CD3D, CD3E, CD3G, XCL1, KLRC1, XCL2, GZMB, CCL5*
CD8+ T-cell 3	*CD3D, CD3E, CD3G, GNLY, FGFBP2, NKG7, GZMH, KLRD1, GZMB*
CD8+ T-cell 4	*CD3D, CD3E, CD3G, CCL3, CXCL13, CCL4L2, IFNG, CCL5, PDCD1*
NK-cell 1	*NKG7, GNLY, TYROBP, FGFBP2, FCGR3A, FCER1G, PRF1*
NK-cell 2	*NKG7, GNLY, TYROBP, XCL1, XCL2, KRT81, GZMK*
B-cell 1	*CD19, CD79A, MS4A1, TCL1A, IGHD, IL4R*
B-cell 2	*CD19, CD79A, MS4A1, CLECL1, AIM2, CD27*
B-cell 3	*CD19, CD79A, MS4A1, LMO2, LRMP*
Dendritic cell	*CST3, LYZ, CD1C, FCER1A, CD1E, CD1A, S100B*
Macrophages *RGS1*	*CST3, LYZ, RGS1, ITM2B, C3, GPR34, CD9, TM6SF1*
Macrophages *FOLR2*	*CST3, LYZ, FOLR2, FCGBP, SEPP1, C1QB, C1QA, C1QC*
Macrophages *CCL18*	*CST3, LYZ, CCL18, CTSL, GPNMB, CTSD, FABP5, LGALS3, MARCO*
Monocytes *THBS1*	*CST3, LYZ, FPR2, VCAN*
Monocytes *S100A9*	*CST3, LYZ, S100A9, S100A8, FCN1, S100A12, VCAN*
Epithelial cells	*TPO, TSHR, TG, NKX2-1*
Fibroblasts	*NOTCH3, PDGFRA, POSTN, MFAP5*
Endothelial cells	*VWF, SELP, SELE*
Follicular cells	*TPO, TSHR, TG, NKX2-1*

3.6. Deconvolution tools

For this study, we selected a diverse set of computational methods to deconvolute the TME and estimate cell type proportions in TC. The chosen methods include *BLADE*, *Scaden*, *CIBERSORTx*, *NNLS*, *SVR-NNLS*, *MuSiC*, and *EPIC*. These methods were selected to represent a broad range of algorithmic approaches, from ML-based

techniques like *Scaden* to simple regression-based models such as *NNLS* and *SVR-NNLS*, and more specialized methods like *CIBERSORTx* and *EPIC*, which are designed to address specific challenges in bulk RNA-seq deconvolution (**Supplementary Table 1**). *BLADE*, known for its ability to leverage multi-sample information, and *MuSiC*, which integrates multiple single-cell reference profiles, were included to enhance the accuracy of cell type identification (35, 63).

For the *NNLS* method, the cell type proportions were estimated by solving the linear system using the *NNLS* algorithm, which constrains the solutions to be non-negative. Specifically, the signature matrix X, representing cell type-specific gene expression profiles, was used as the reference, while Y, the bulk RNA-seq data, was decomposed into a linear combination of X. The cell fractions for each sample were obtained by minimizing the residual sum of squares under the non-negativity constraint. This approach ensures that the estimated fractions are biologically meaningful by disallowing negative contributions from any cell type.

In contrast, the *v-SVR* followed by *NNLS* (*SVR-NNLS*) method employed a two-step process to improve feature selection for deconvolution. Initially, a *SVR* with a linear kernel was applied to the bulk RNA-seq data, fitting the model for each cell type to estimate feature weights. These weights were then used to rank genes, allowing us to select the top features most informative for distinguishing between cell types. A reduced signature matrix was constructed using the top-ranked genes, and the *NNLS* algorithm was subsequently applied to the reduced matrix for deconvolution. This combination of *SVR* and *NNLS* allowed for dimensionality reduction and focused on the most relevant genes, potentially improving accuracy in the presence of noise or irrelevant features (17, 19, 20).

All remaining deconvolution methods discussed were applied as per package instructions, apart from *BLADE* which proved to be a very computationally intensive algorithm and required subsetting of the bulk data (**Figure 9**).

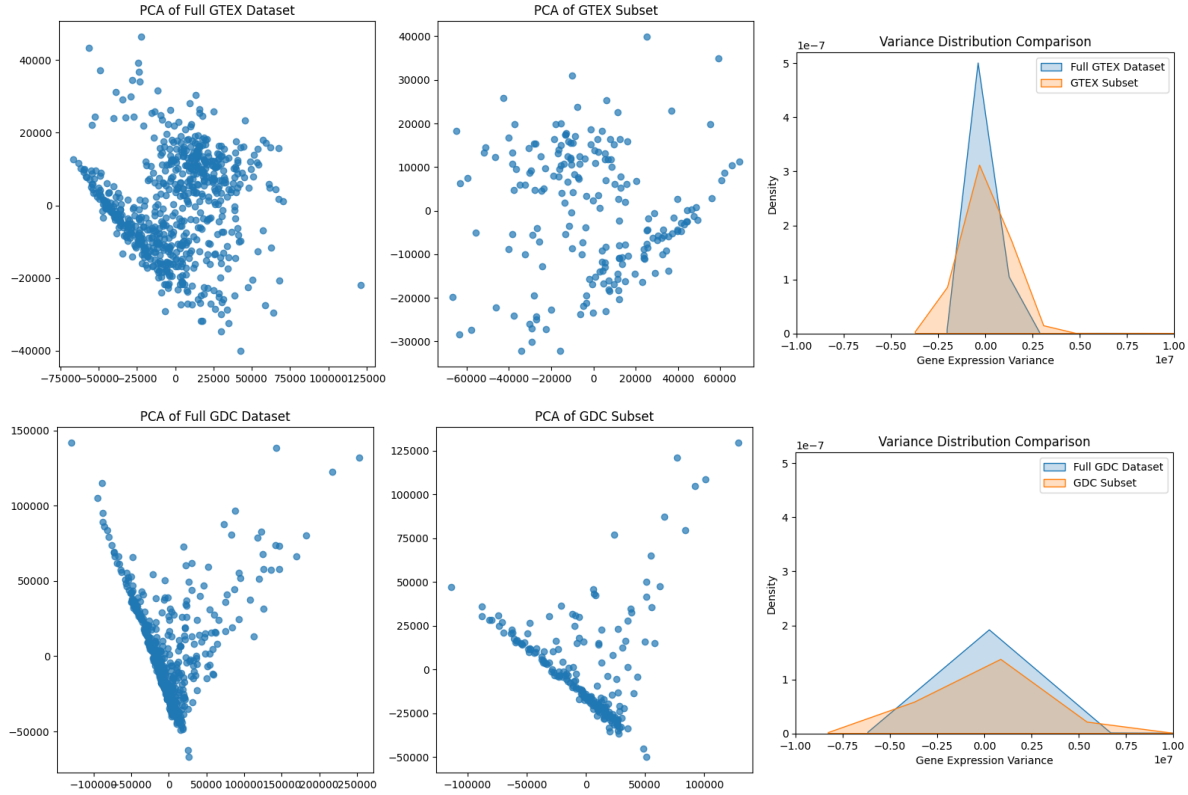


Figure 9: **PCA of GTEX (top row) and GDC (bottom row) subsets.**
The figure illustrates the comparison between the full GTEX and GDC cohorts and their respective subsets of 200 samples and top 5,000 highly variable genes. The right panels display the distribution of gene expression variance for both the full and subset data, allowing for a comparison of the variance lost by subsetting - The subset variance distribution (orange) is slightly narrower compared to the full datasets (blue), indicating a loss of variance in the gene expression profiles after subsetting.

By utilizing these diverse methods, we aimed to achieve a comprehensive understanding of the TC TME, comparing their performance across TCGA-THCA and GTEX datasets.

3.7. Cell enrichment tools

In addition to the deconvolution methods, *xCell*, a cell type enrichment analysis tool, was incorporated into the project. It constitutes a computational method that infers the abundance of various immune and stromal cell types based on gene expression profiles (64). *xCell* utilizes a gene signature-based approach to score the relative enrichment of each cell type across samples instead of directly estimating the cell type fractions. Since it does not rely on a reference, the cell type fractions of the deconvolution methods were mapped to a common key in a dictionary, along with the enriched cells subsequent from the *xCell* analysis to make them comparable (**Table 4**).

Table 4: **Classification of *xCell* populations and reference cell populations mapped to a common cell type.**
The grouping includes various immune and stromal cell types that are of interest in the study of TME. Each category lists populations that are possibly grouped together during deconvolution and are likely to be miscalculated.

Cell Type	Cell Populations
CD8+ T-cell	CD8+ naive T-cells, CD8+ T-cells, CD8+ Tcm, CD8+ Tem, CD8+ T-cell 1, CD8+ T-cell 3
CD4+ T-cells	CD4+ memory T-cells, CD4+ naive T-cells, CD4+ T-cells, CD4+ Tcm, CD4+ Tem, CD4+ T-cell 1, CD4+ T-cell 2, CD4+ T-cell 3, CD4+ T-cell 4, Tgd cells, Th1 cells, Th2 cells, Tregs
B-cells	B-cells, Class-switched memory B-cells, Memory B-cells, naive B-cells, pro B-cells, B-cell 1
Macrophages	Macrophages, Macrophages M1, Macrophages M2, Macrophages *RGS1*
Monocytes	Monocytes, Monocytes *S100A9*
NK cells	NK cells, NK-cell 2
Natural Killer T-cells	NKT
Dendritic cells	aDC, cDC, DC, iDC, pDC
Progenitor cells	CLP, CMP, GMP, MEP, MPP
Eosinophils	Eosinophils
Endothelial cells	Endothelial cells, ly Endothelial cells, mv Endothelial cells
Mast cells	Mast cells
Melanocytes	Melanocytes
Epithelial cells	Epithelial cells
Neutrophils	Neutrophils
Basophils	Basophils
Connective tissue cells	Chondrocytes, Astrocytes, Fibroblasts, Keratinocytes, Myocytes, Skeletal muscle, Smooth muscle, Osteoblast, Pericytes, MSC
Haematopoietic cells	Plasma cells, Platelets, HSC, Erythrocytes, Megakaryocytes
Adipose cells	Adipocytes, Preadipocytes, Sebocytes
Other cells	otherCells, Hepatocytes, Mesangial cells, Neurons

Abbreviations: Tgd, gamma-delta T; Th, helper T; NK, natural killer; NKT, natural killer T; aDC, activated dendritic cells; cDC, conventional dendritic cells; DC, dendritic cells; iDC, immature dendritic cells; pDC, plasmacytoid dendritic cells; CLP, common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte-macrophage progenitor; MEP, megakaryocyte-erythroid progenitor; MPP, multipotent progenitor; ly, lymphatic; mv, microvascular; MSC, mesenchymal stem cells; HSC, haematopoietic stem cells;

This tool is particularly valuable for evaluating the immune landscape of tumours, offering additional insights into the complexity of the TME. In this comparative analysis, *xCell* serves as an important complementary approach, allowing us to assess not only the fraction of cell types but also the relative enrichment across different datasets, including TCGA-THCA and GTEX. Its inclusion provides a broader perspective on the TME, especially when used alongside more traditional deconvolution methods (17, 64).

3.8. Evaluation metrics

To evaluate and compare the deconvolution results across the selected methods, a series of quantitative and visualization-based metrics were employed. Correlation analysis was used to assess the linear relationship between estimated cell type proportions generated by different methods, providing an initial measure of agreement. Bland-Altman plots were then applied to further explore the agreement between methods by plotting the difference in their estimates against the average, enabling us to identify systematic biases and the limits of agreement between each pair of methods. To visually compare the distribution of estimated cell type proportions across methods, we utilized violin plots, which highlight both the spread and central tendency of estimates. PCA was performed to reduce the dimensionality of the data, allowing us to visualize clusters of similar methods and detect potential outliers in a lower-dimensional space. Additionally, we calculated the Concordance Correlation Coefficient (CCC) to measure the agreement between two methods, taking into account both accuracy and precision, where higher values reflect better concordance. Finally, we treated the *xCell* results as a ground truth for comparison, using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared to evaluate the accuracy of each method in relation to the *xCell*-generated cell type enrichment scores.

4. Results

To assess the concordance of these methods, we evaluated the correlation between estimated cell fractions, as visualized in the correlation heatmaps for GTEX (**Figure 10**) and GDC (**Figure 11**). The results highlight key differences in cell type estimation across methods and cohorts, with notable trends emerging for immune and stromal populations. Bland-Altman plots were also used to explore cell type-specific discrepancies between methods. Further analyses incorporate enrichment scores from *xCell* to provide a more comprehensive view of the immune landscape across both cohorts.

4.1. Correlation analysis

Across the heatmaps there isn't much overall concordance between the analysed methods when estimating cell fractions, which suggests they may be assimilating different aspects of the data due to their distinct approaches for deconvolution or dealing with batch effects in different ways. *Scaden* and *CIBERSORTx* stand out, seemingly showing better alignment for most cell types with a slight diagonal line of positive correlations implying they might be estimating similar fractions for specific immune cells, which can be attested by the similar distributions in the corresponding violin plots. This indicates a more robust deconvolution, especially in identifying immune cells such as T-cells, B-cells, or macrophages. Higher CCC of 139.76 and 343.32 indicate strong agreement in the estimation of NK-cells and CD8+ T-cells (type 3), respectively.

High correct and incorrect associations can be found throughout both cohorts for B-cells and macrophages. CCC evaluation for these cells shows relatively high values for methods like *CIBERSORTx* and *Scaden*, while for other methods lower (or even negative values) indicate that they may be misestimated. On the other hand, there seems to be a consistent agreement for CD8+ T-cells and NK cells, supported by high CCC values across deconvolution methods, in both cohorts.

The simple *NNLS* and *SVR-NNLS* approaches have generally poor results, showing weak correlations with other deconvolution methods suggesting they are likely to produce results that deviate from the other techniques, and given their linear approach they may be struggling to capture the complex TME-landscape, especially compared to

the other more complex algorithms. This is further validated by low (close to zero) or negative CCC values across most cell types estimated by these methods, reinforcing the notion that they indeed have weaker performance in terms of concordance with the others.

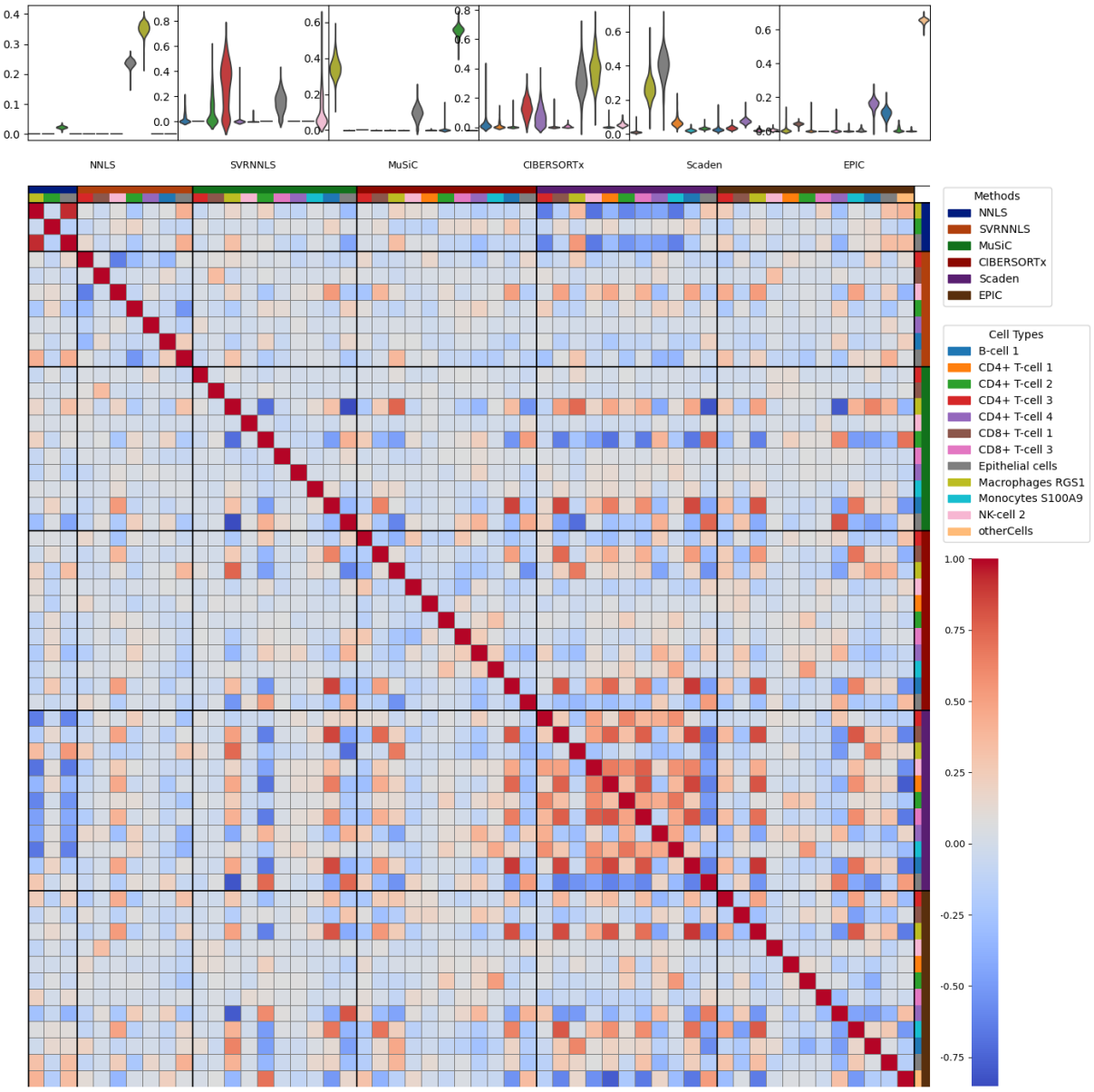


Figure 10: **Correlation Matrix for GTEX Deconvolution Methods with Cellular Fractions.**
The diagonal elements (deep red) show perfect correlations between the same cell type estimates within the same method. Off-diagonal elements reveal correlations between different methods for the same or different cell types. Blue regions indicate negative correlations, meaning methods disagree on specific cell type proportions. Red regions indicate positive correlations, meaning methods show similar estimates for certain cell types. The violin plots visualize the distribution of estimated cell fractions for each method across different cell types and reveal the variability of estimates across methods. The colour bars enable a clear distinction between methods and cell types, for easy association of the colours in the heatmap with the corresponding method and cell type. Methods like *Scaden*, *CIBERSORTx*, and *EPIC* seem to provide more stable estimates for cell fractions across samples, while *NNLS* and *SVR-NNLS* seem to fail at unravelling the complete immune landscape and estimate fewer cell types.

EPIC and *MuSiC* also show relatively low concordance, exhibiting poor performance shown by the lack of strong positive correlations. Their estimates also diverge significantly from those of seemingly more reliable methods like *Scaden* and *CIBERSORTx*. CCC analysis also reveals negative correlations across most cell types supporting the previous statement which hinted at a sizable discordance.

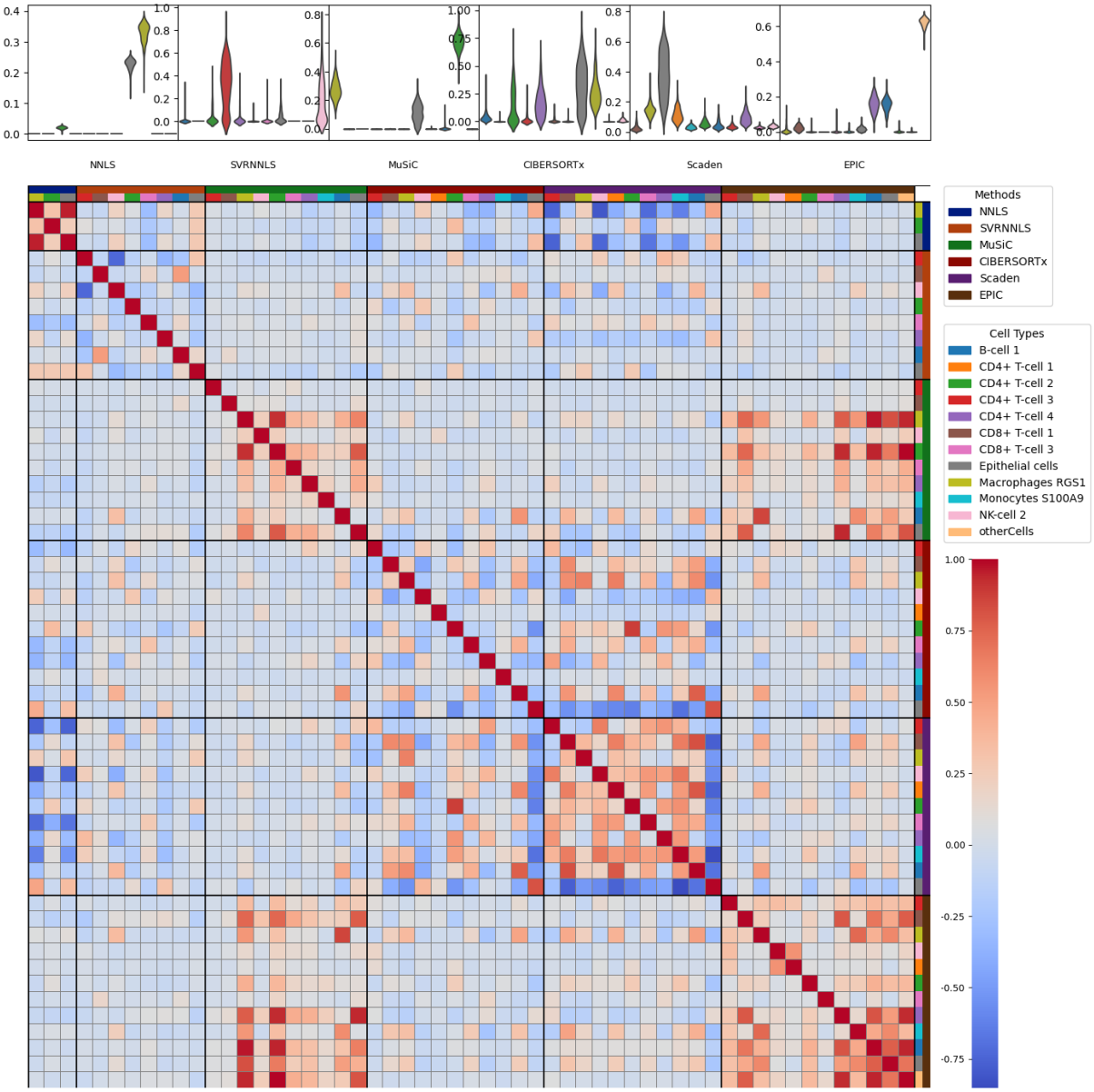


Figure 11: **Correlation Matrix for GDC Deconvolution Methods with Cellular Fractions**.
The diagonal elements (deep red) show perfect correlations between the same cell type estimates within the same method. Off-diagonal elements reveal correlations between different methods for the same or different cell types. Blue regions indicate negative correlations, meaning methods disagree on specific cell type proportions. Red regions indicate positive correlations, meaning methods show similar estimates for certain cell types. The violin plots visualize the distribution of estimated cell fractions for each method across different cell types and reveal the variability of estimates across methods. The colour bars enable a clear distinction between methods and cell types, for easy association of the colours in the heatmap with the corresponding method and cell type. Methos like *MuSiC*, *Scaden*, *CIBERSORTx,* and *EPIC*, seem to do well in a cancer context, showing strong correlations between several cell type estimates. Comparing MuSiC and EPIC estimates, for example, we can see several wrong or spurious strong associations between different cell types, which can be attested by the asymmetrical estimates displayed in the corresponding violin plots.

Bland-Altman plots were also created to visually inspect cell type-specific differences of estimation between methods and revealed linear patters and cloud-like distributions for most deconvolution methods in the GTEX and GDC cohorts, respectively. These linear patterns can occur if there's a consistent relationship between the methods being compared, suggesting that they are systematically estimating the proportions in a similar manner, leading to minimal variation in differences, while cloud-like distributions may occur if there is more variability in estimates among methods, showcasing less agreement and more inconsistency. These, however, are accompanied by part of the distribution being trailing beyond the limits of concordance (above or below the red lines), which means that for some samples there is poor agreement. Despite this, most methods seem to have the bulk of the distribution between the limits.

In several Bland-Altman plots (**Figure 12**), particularly for the GDC cohort, a linear cluster can be visualized alongside a cloud-like distribution, which could be indicative of a subset of samples where both methods produce more inconsistent estimates.

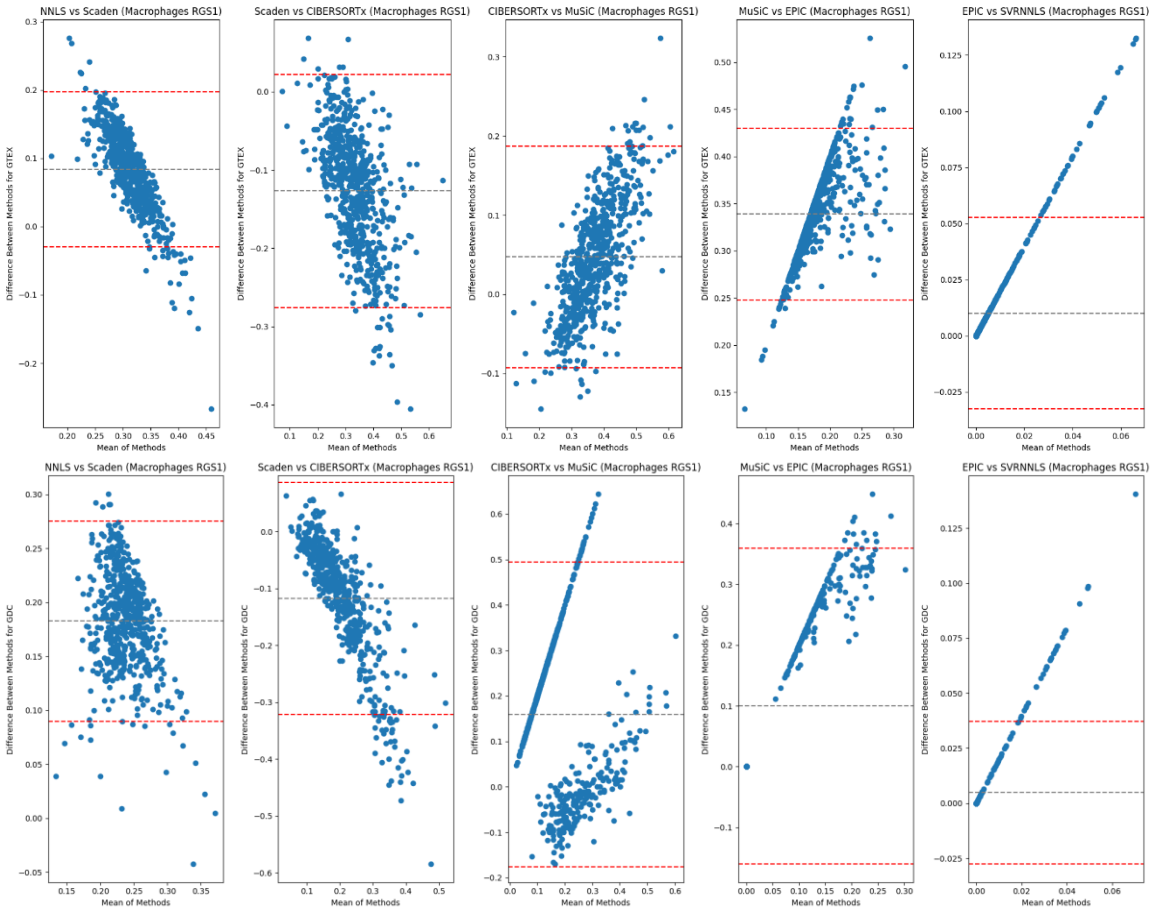


Figure 12: **Bland-Altman Plots for Method Comparison of Macrophages RGS1 Fractions in GTEX and GDC cohorts.** The plots display the difference between cell fraction estimates (y-axis) against the mean of the estimates (x-axis) for each pair of deconvolution methods. The top row represents comparisons for GTEX, while the bottom row represents comparisons for GDC. The dashed red lines indicate the limits of agreement (mean difference ± 1.96 standard

deviations), and the grey dashed line shows the mean difference. Clear patterns, such as linear trends or outlier points exceeding the limits of concordance, suggest the presence of biases or inconsistencies in cell type estimations between the methods.

Even though there is still some room for improvement, *Scaden* and *CIBERSORTx* seem to emerge as more reliable with the highest mean CCC both for the GTEX (0.2469) and GDC (0.1780) cohorts. The pair *CIBERSORTx* and *MuSiC* show a higher mean CCC in the GDC cohort (0.1311) which is also an improvement compared to the GTEX cohort (0.0886). While *MuSiC* generally underperforms in normal tissue samples it seems to do slightly better for TME deconvolution when compared to *CIBERSORTx*. The pair *Scaden* – *MuSiC* also reveals a higher CCC in GDC (0.0592) compared to GTEX (0.0666) but these values are still relatively low, indicating only a moderate agreement between these two approaches.

Negative or very low mean CCC values continue to be observed across most method pairs for the *SVR-NNLS* approach for both cohorts (*MuSiC*: -0.0256; *CIBERSORTx*: -0.0025). Likewise, *EPIC* continues to show poor concordance in both cohorts, with a mean CCC of accordance with *CIBERSORTx* of only 0.039 for GTEX and 0.0157 for GDC. The method pair *NNLS* and *MuSiC* also shows weak alignment with each other in either cohort, with mean CCC values of 0.0071 and -0.0029 for the GTEX and GDC cohorts, respectively.

4.2. Enrichment analysis with *xCell*

In both GTEX and GDC, we observe significant variability in the correlations between *xCell* scores and deconvolution methods, as illustrated by the heatmaps (**Figure 13**). We see consistently strong correlations for B-cells across multiple methods (*Scaden*, *CIBERSORTx*, *EPIC*, and *MuSiC*).

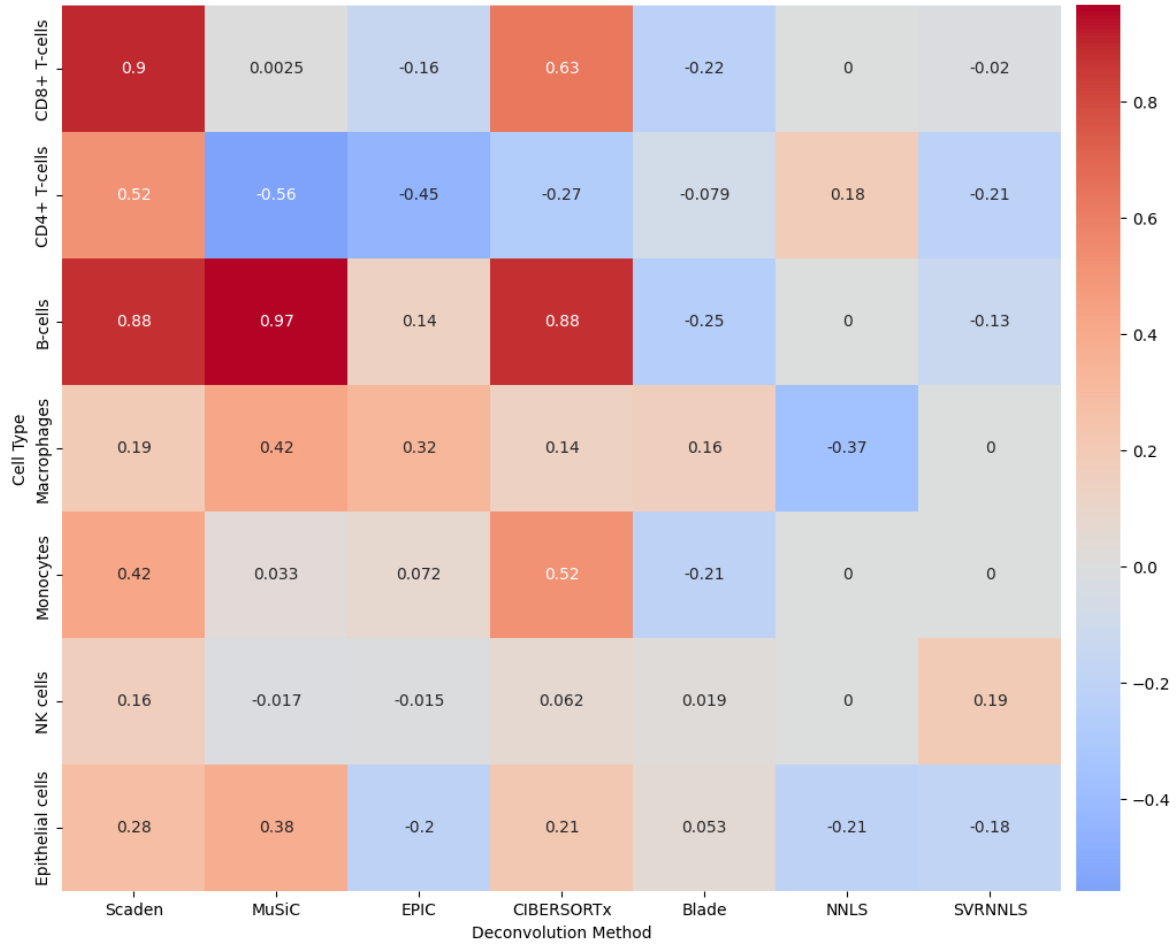


Figure 13: **Correlation Heatmap of *xCell* Enrichment Scores with Estimated GTEX Cell Fractions**.
Blue regions indicate negative correlations, meaning methods disagree on specific cell type proportions. Red regions indicate positive correlations, meaning methods show similar estimates for those cell types. B-cells show strong correlations across methods, standing out when comparing the estimated proportions with the enrichment scores, especially for methods like *CIBERSORTx* and *Scaden*. *NNLS* and *SVR-NNLS* show weak to moderate correlations across all cell types.

This is also observed on monocytes for methods like *Scaden* and *CIBERSORTx*, and in macrophages for most deconvolution methods especially for the GDC cohort. On the other hand, on the GTEX cohort, CD8+ T-cells show decent correlations, slightly higher than with GDC, for methods like *Scaden* and *CIBERSORTx*.

For some cell types like NK cells and epithelial cells, the correlation between *xCell* and most deconvolution methods is weak or even negative, particularly in the GDC cohort. Contrarily, the enrichment scores for the GTEX cohort (**Supplementary figure 1**) and GDC cohort (**Supplementary figure 2**) paint a different scenery, with NK cells and epithelial cells having a higher presence in the TME than in healthy tissue, where the expression is weak. CD4+ T cells appear to have moderate correlations with

deconvolution methods, especially GTEX, but performance drops in the cancer cohort, particularly in *MuSiC* (**Figure 14**).

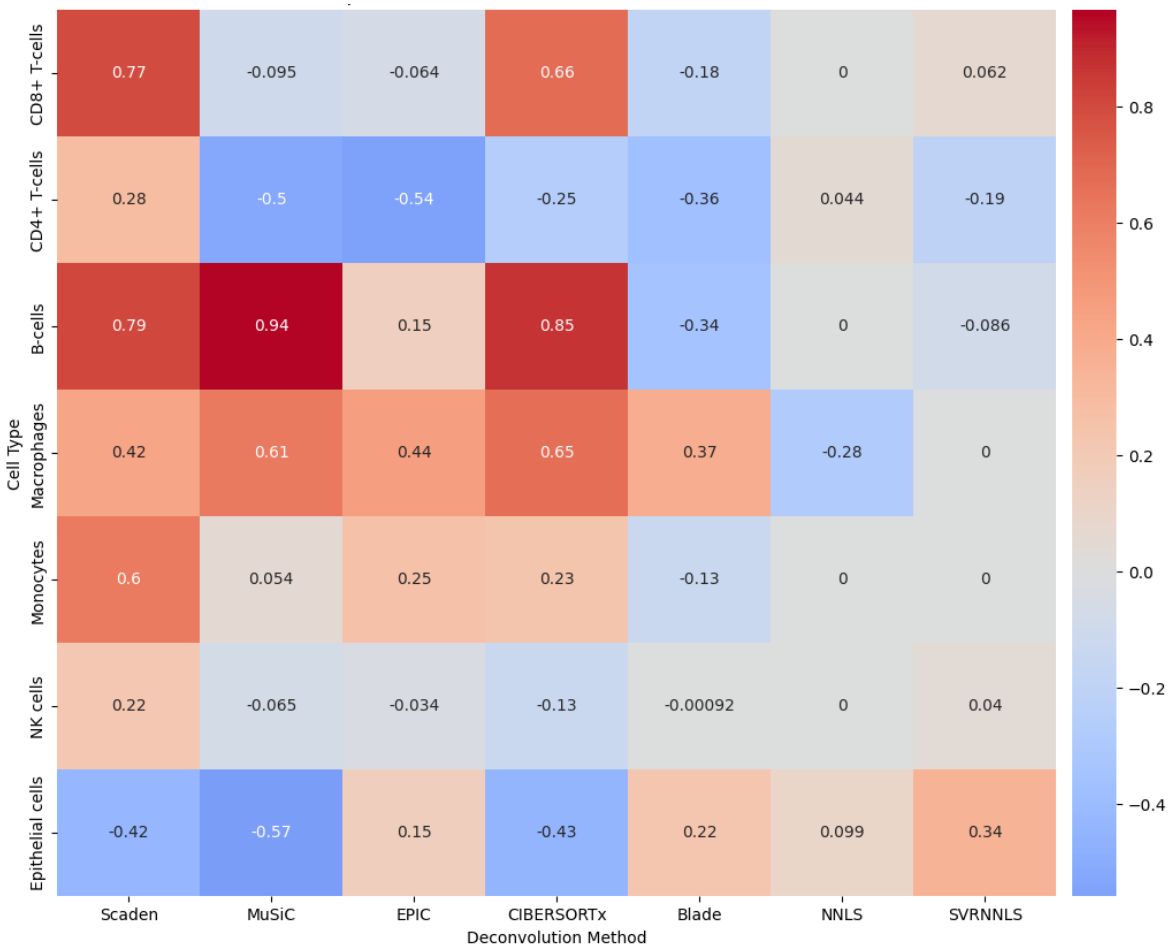


Figure 14: **Correlation Heatmap of *xCell* Enrichment Scores with Estimated GDC Cell Fractions**.
Blue regions indicate negative correlations, meaning methods disagree on specific cell type proportions. Red regions indicate positive correlations, meaning methods show similar estimates for those cell types. B-cells and macrophages stand out as the cell types with strongest positive correlations with the enrichment scores, while epithelial cells and CD4+ T-cells show moderate to strong negative correlations across deconvolution methods.

Scaden and *CIBERSORTx* show consistently high correlations with the enrichment scores across both cohorts, while *MuSiC* shows a surprisingly good performance at estimating B-cells and macrophages, particularly in GTEX unlike previously verified, suggesting it may perform better at deconvolving immune cell fractions in healthy tissue. Moreover, *SVR-NNLS* and *EPIC* seem to show generally weaker correlations with *xCell* scores (**Figure 15**), mainly for specialized immune cells like NK-cells and epithelial cells.

A further in-depth cell type specific analysis of the immune landscape scored by *xCell* reveals some key trends and patterns, especially for the cancer cohort. There are few cell types displaying strong positive correlations in the healthy cohort, in fact, besides

closely related cell types such as macrophages M1 and M2, representing pro- and anti-inflammatory phenotypes, subpopulations of B-cells, such as class-switched memory B-cells and memory B-cells, or microvascular and lymphatic endothelial cells, there are no strong positive correlations or patterns of note in the GTEX cohort.

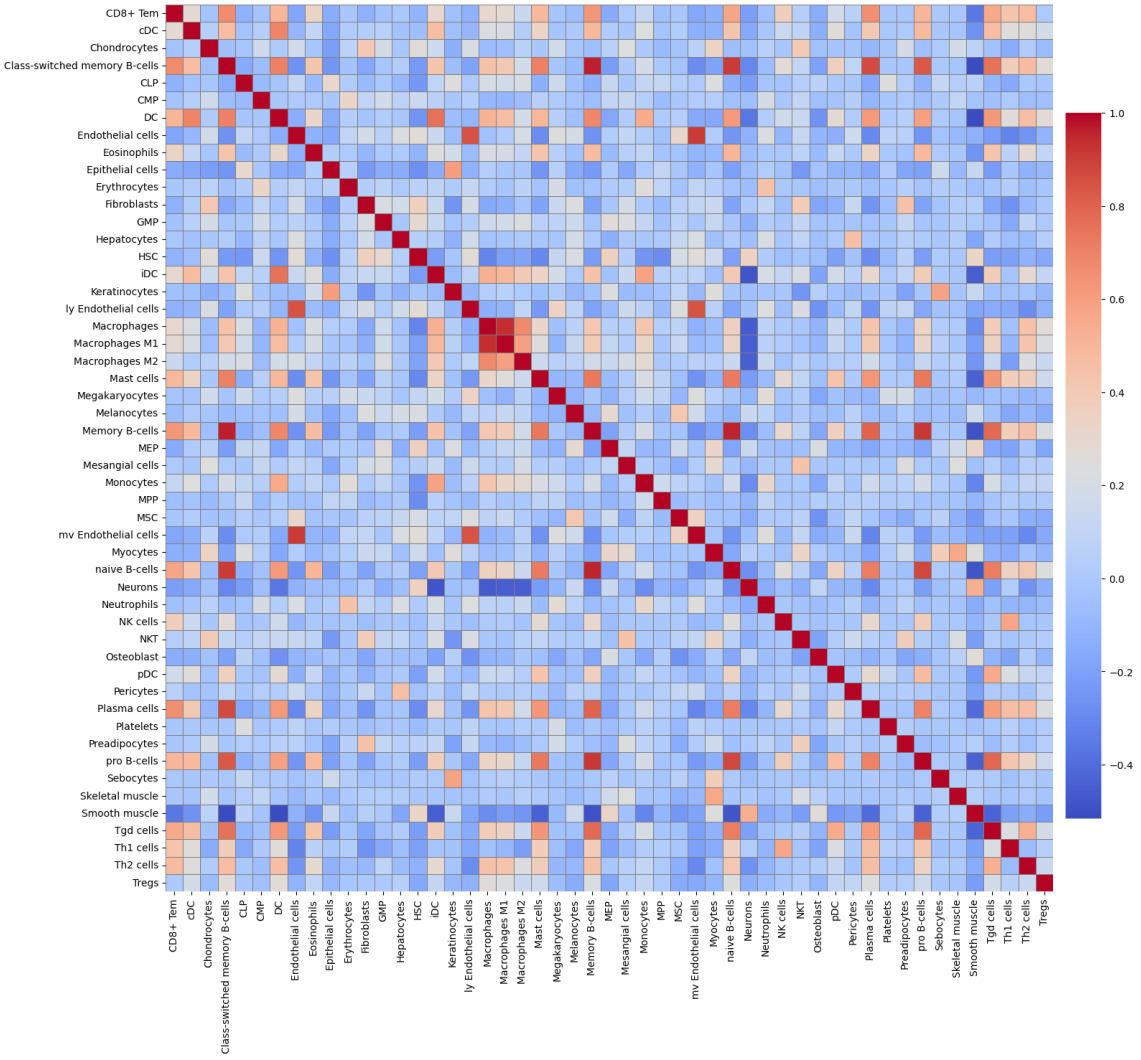


Figure 15: **Correlation Matrix of *xCell* Cell Type Enrichment Scores for the GTEX cohort**.
The colour red indicates positive correlation (higher similarity between cell types), blue indicates negative correlation (inverse relationship between cell types), and white represents near-zero correlation (weak or no relationship between cell types).
Abbreviations: Tgd, gamma-delta T; Th, helper T; NK, natural killer; NKT, natural killer T; aDC, activated dendritic cells; cDC, conventional dendritic cells; DC, dendritic cells; iDC, immature dendritic cells; pDC, plasmacytoid dendritic cells; CLP, common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte-macrophage progenitor; MEP, megakaryocyte-erythroid progenitor; MPP, multipotent progenitor; ly, lymphatic; mv, microvascular; MSC, mesenchymal stem cells; HSC, haematopoietic stem cells.

In contrast, the GDC cohort (**Figure 16**) is characterized by strong correlations and patterns, especially among immune cells, such as CD4+ T-cells, CD8+ T-cells, B-cells, and T-helper cells, which only adds to the previous statements supporting a high immune response and a vastly different microenvironment.

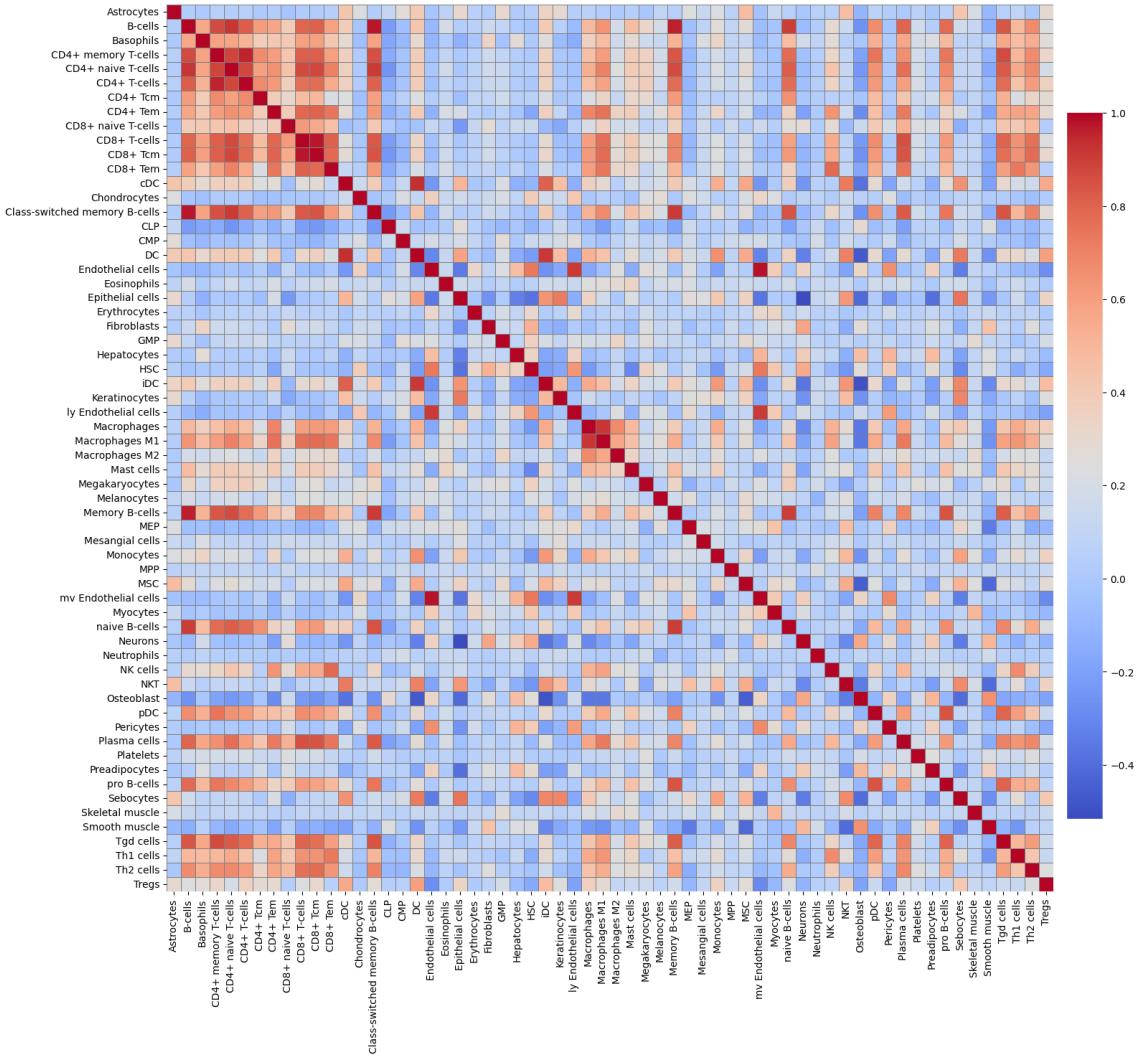


Figure 16: **Correlation Matrix of *xCell* Cell Type Enrichment Scores for the GDC cohort**.
The colour red indicates positive correlation (higher similarity between cell types), blue indicates negative correlation (inverse relationship between cell types), and white represents near-zero correlation (weak or no relationship between cell types).
Abbreviations: Tgd, gamma-delta T; Th, helper T; NK, natural killer; NKT, natural killer T; aDC, activated dendritic cells; cDC, conventional dendritic cells; DC, dendritic cells; iDC, immature dendritic cells; pDC, plasmacytoid dendritic cells; CLP, common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte-macrophage progenitor; MEP, megakaryocyte-erythroid progenitor; MPP, multipotent progenitor; ly, lymphatic; mv, microvascular; MSC, mesenchymal stem cells; HSC, haematopoietic stem cells.

5. Discussion

The complexity of the TME arises from the intricate interplay between diverse cell types, including immune cells, stromal cells, and cancerous cells. Understanding the composition and behaviour of these cell types is critical for gaining insights into cancer progression, immune evasion, and potential therapeutic targets. Deconvolution tools offer a powerful means to untangle the complex cellular architecture of bulk RNA-seq data, enabling insights into the distribution and behaviour of immune and stromal cells within the TME (29, 35). In this study, we employed several deconvolution methods to investigate key immune and stromal cells in the TME, such as CD4+ T-cells, CD8+ T-cells, B-cells, macrophages, NK cells, and epithelial cells. The goal was to assess the performance of these methods by comparing the estimated cell fractions with *xCell* enrichment scores, known marker genes, and DEG data across both healthy thyroid tissue (GTEX) and TC tissue (GDC) cohorts (64, 65).

While most deconvolution methods were applied consistently across the full datasets, *BLADE* was excluded from the comparative analysis due to its computational complexity. Because of its resource-intensive nature, *BLADE* was tested on a subset of 200 samples using the top 5000 highly variable genes (63). Post-processing, PCA revealed that the dataset was not well represented, leading to incomplete and suboptimal results (**Figure 9**). To prevent skewing the interpretation of overall trends, *BLADE* was excluded from the comparative analysis, though its results are available in the supplementary material for reference (**Supplementary Figures 3**-**6**).

Additionally, not all cell types from the marker gene dictionary were captured in the deconvolution results. This can be attributed to the moderate resolution parameter (0.7) applied during clustering with the Leiden algorithm, which resulted in fewer distinct clusters. As a consequence, several cell types were missing from the reference matrix, potentially impacting the detection and estimation of specific cell populations. Moving forward, with such a complex dataset, a resolution of 1.0 may be a good experimentation for attempting to capture a more comprehensive immune landscape. Another good approach could be to create two reference matrices, one from healthy thyroid single-cell tissue and another from TC (62, 66).

Another point of consideration is the interpretation of Bland-Altman plots. In several instances, a trailing tail of data points extended beyond the limits of concordance. This

pattern may be attributed to batch effects or variability in sample quality across the datasets. If this is the case, the linear patterns observed in some comparisons could result from a systematic bias present in one or more methods, rather than true biological variation. Recognizing these limitations is crucial for accurately evaluating the performance of each deconvolution tool (67, 68).

At last, comparing the deconvolution estimates for key immune and stromal populations, supported by marker gene data and DEG analysis (**Supplementary Figure 7**), may provide a clearer view of the strengths and limitations of each method in capturing the distinct features of the TME in both healthy and cancerous tissues (69).

5.1. Cell type-specific analysis

5.1.1. Cytotoxic cells

CD8+ T-cells are one of the most critical immune cell types in both normal and tumour environments, involved in cytotoxic activities against infected and malignant cells. Cytotoxic T-cells correlate relatively well with the deconvolution methods, particularly *Scaden* and *CIBERSORTx* (65). Furthermore, like the enrichment analysis suggests, the correlation seems slightly higher in GDC, indicating potential difficulties in estimating these cells in the healthy tissue. Based on the signature matrix (**Supplementary Figure 8**), *CD8A*, *GZMB*, and *PRF1* are key marker genes associated with CD8+ T-cell subsets in the reference data.

These genes indicate high cytotoxic activity, which is particularly relevant in the context of TC (1, 5). Moreover, the cytotoxic cells' subpopulation 1 (CD8+ T-cell 1), marked by *GZMK*, *GZMA*, *CCL5*, and *CD3D*, and subpopulation 3 (CD8+ T-cell 3), marked by the upregulated *GNLY*, *NKG7*, and *GZMB*, are critical for antitumour activity, possessing NK-like capabilities and often being found in immune-responsive tumours, which should be the case for TC (5, 70). These populations were estimated mainly in the TME, where these infiltrating immune cells are critical for controlling tumour growth and contributing to cancer cell clearance. This suggests an activated immune state, which can be corroborated with the deconvolution results, particularly for methods that show higher correlations in the GDC cohort (like *CIBERSORTx*), and with the aforementioned enrichment scores (23, 70).

In fact, the high correlations observed between the same cell types (correct associations) (**Figures 15-16**), suggest these cell types may be more consistently estimated across

methods, potentially due to strong marker gene expression or simply abundance in the datasets. High correlations between different cell types (incorrect associations) however, may be indicative of poor performance, meaning that the methods are mistaking cell types for others, or they have overlapping signatures that make them difficult to distinguish. For instance, macrophages may be often mistaken for monocytes in some methods, and the same goes for the different types of CD8+ T-cells due to shared gene markers.

5.1.2. CD4+ T-cells

The CD4+ T-cell subpopulations showed notable heterogeneity across both cohorts, with distinct marker expression patterns defining each group. CD4+ T-cell 1 was characterized by metallothionein genes (*MT1G*, *MT1X*, *MT1E*, *MT2A*), indicating immune activation, and was consistently detected by *Scaden*, *CIBERSORTx*, and *EPIC*. CD4+ T-cell 2 exhibited key regulatory markers such as *KLRB1* and *IL7R*, emphasizing its role in immune modulation, with fractions estimated by all deconvolution methods. CD4+ T-cell 3, marked by *CCR7* and *TCF7*, was associated with memory T-cell function, with a consistent presence across methods (except for *NNLS*), particularly in the GDC cohort. Finally, CD4+ T-cell subpopulation 4 was also estimated by all methods apart from *NNLS* and expressed stress-response genes like *JUN* and *FOS*, highlighting its role in immune responses under stress conditions (5, 9).

These subpopulations reflect the diverse functional roles of CD4+ T-cells in both immune surveillance (GTEX) and TME (GDC), with fractions varying based on immune context and method agreement. Nevertheless, the consistent CD4+ T-cell estimation across deconvolution methods suggests a robust detection of these cell types, particularly in TME, while on the other hand, the absence of detection of the subpopulations 5 through 7 emphasize the impact of reference data on identifying specific subpopulations.

5.1.3. B-cells

Expressed by canonical B-cell markers like *CD19*, *CD79A*, and *MS4A1*, B-cell 1 is estimated by *Scaden*, *MuSiC*, *CIBERSORTx*, *EPIC*, and *SVR-NNLS*. These cells are involved in antibody production and antigen presentation. Notably, the putative B-cell 1 subpopulation showed high positive correlation across methods, demonstrating good agreement between the deconvolution methods and *xCell* enrichment scores, especially

in the GDC cohort, where it might reflect tumour-infiltrating B-cells. High marker overlap in the reference data (notably *CD19* and *MS4A1*) reinforces their importance in immunosurveillance and further highlights the ease with which these immune cells can be detected in both cohorts (3, 5).

5.1.4. Macrophages and Monocytes

The subpopulation Macrophages RGS1 is characterized by the upregulated expression of *RGS1*, *C3*, and *ITM2B*, and was consistently estimated by *Scaden*, *MuSiC*, *CIBERSORTx*, *EPIC*, and *NNLS* across both cohorts yet more prominently in the GDC cohort, where immune infiltration in the TME may have made these cells easier to detect. Macrophages play key roles in tumour progression, and their presence, especially in the M2-like polarization (*RGS1* and *C3* markers), indicates immunosuppressive activity (71). Similarly, monocytes expressing *S100A9* and *S100A8* (Monocytes *S100A9* cell type) were estimated by *Scaden*, *MuSiC*, *EPIC*, *CIBERSORTx*, and enriched by *xCell* in the GDC cohort, corresponding to their role in inflammation and recruitment of other cells important in the immune landscape. Moreover, S100 proteins are known to participate in both immune regulation and tumour progression, which may explain the upregulation of these genes in the reference dataset (72).

5.1.5. Epithelial and NK cells

Marked by *TPO*, *TSHR*, and *TG*, this cell type was estimated by all methods, indicating consistent presence, although weak, in both healthy thyroid tissues (GTEX) and TC (GDC). In the TME however, higher estimations of this subpopulation may also represent a shift towards epithelial-mesenchymal transition (EMT) which could be relevant to tumour progression (73). In contrast, NK cells and epithelial cells showed weak or even negative correlations between *xCell* and most deconvolution methods, particularly in the GDC cohort, suggesting they present greater challenges for accurate estimation, likely due to their complex roles within the TME and potential gene overlap with other cell types. The consistent expression of the epithelial markers in both contexts also underlines their role in thyroid function and may suggest follicular and parafollicular cells could also be encompassed by the labelling of this subpopulation, which could also explain the poor correlations across both cohorts, especially for methods like *Scaden* and *CIBERSORTx*.

These results align with the observation that the GDC cohort, characterized by a more complex and heterogeneous TME, poses a greater challenge for accurate cell type estimation, and the lower correlations for most cell types in the GDC cohort reflect this complexity, underscoring the difficulty of deconvoluting tumour-infiltrating cells in cancerous tissues (4, 6, 23).

5.2. Limitations and Future Directions

While this study provides valuable insights into the cellular composition of the TME in TC and healthy tissues, several limitations should be considered when interpreting the results. First and foremost, the reference data used for deconvolution was derived from TC samples, which likely contributed to its suboptimal performance in healthy thyroid tissue. The discrepancies between cancerous and normal tissue in cell type composition and gene expression profiles may have skewed the deconvolution results, especially in the GTEX cohort.

Additionally, there were issues with the normalization strategies applied during preprocessing. The counts assay from the scRNA-seq data was initially normalized to library size (10.000 UMI), but the counts were mistakenly stored as “raw”. Consequently, further normalization steps were applied when performing deconvolution, where the reference data was normalized as TPM to match the bulk data and, in some cases, log-transformed (log2 for *Scaden*, TPM for *CIBERSORTx*, and log1p for the remaining methods). This layered normalization likely impacted the accuracy of cell type estimation and should be rectified in future analyses (20, 74-76).

Another limitation involves the completeness and representativeness of the marker gene dictionary. While it provided a solid foundation for identifying cell types, it did not fully capture the heterogeneity of the cell populations within the reference or bulk data. Expanding and refining the marker gene dictionary would help improve the specificity and accuracy of cell type assignments. This can likely be easily achieved by leveraging separate single-cell reference matrices for cancerous and healthy tissues (66, 77).

Linear patterns in Bland-Altman plots suggest that some methods are systematically estimating cell proportions in a similar manner, leading to minimal variation in differences. In contrast, cloud-like distributions reflect greater variability in estimates, which may indicate less agreement and more inconsistency among methods. This

pattern, particularly evident in the GDC cohort, could be attributed to the inherent heterogeneity of the TME in tumoral tissues, with the bulk of the distributions falling within the limits of concordance (1, 5).

Finally, the different deconvolution methods employed in this study rely on varying estimation strategies, making direct comparisons between them challenging. These methods differ in their underlying assumptions, optimization strategies, and reliance on external reference data, all of which could contribute to varying results across the methods. Furthermore, the methods weren't leveraged to their full extent in the case of *MuSiC* and *CIBERSORTx* for example. *MuSiC* enables the integration of multiple references and *CIBERSORTx* enables eco-type analysis and further hyper parameterization which could be explored for betterment of the results (35).

Overall, the comparative analysis reinforces *CIBERSORTx* and *Scaden* as the most reliable deconvolution methods across both cohorts, providing robust cell fraction estimates. Conversely, *SVR-NNLS* and *EPIC* were found to consistently underperform, making them less reliable for TME analysis. The negative or low mean CCC values observed for the *SVR-NNLS* approach, especially in comparisons with methods like *MuSiC* and *CIBERSORTx*, indicate its inaccuracy across both cohorts. These results highlight that *SVR-NNLS*, while perhaps a step-up from traditional *NNLS*, consistently underperforms compared to other deconvolution approaches, further confirming its limitations in accurately estimating cell fractions (29, 35, 46).

Despite these limitations, this study offers a comprehensive evaluation of deconvolution methods in the context of TC and healthy tissue, contributing valuable insights into the TME's cellular composition. By reviewing and leveraging multiple approaches, integrating marker gene data, *xCell* enrichment scores, and DEG, the strengths and limitations of each method are highlighted, especially in the context of cancer (56, 64).

Conclusion

In this study, we explored the TME of both healthy and cancerous thyroid tissues through a comprehensive comparative analysis of deconvolution methods. Key findings include the strong correlation observed for CD8+ T-cells in the GDC cohort, emphasizing the importance of accurate cell type estimation within cancerous tissues, where immune diversity and infiltration play critical roles. The performance of deconvolution methods across both the GDC and GTEX cohorts revealed the challenges inherent to estimating immune and stromal populations in both healthy and malignant environments. Despite these challenges, the study provides valuable insights that can contribute to refining computational tools and methodologies used for TME deconvolution.

The diverse deconvolution strategies applied, including traditional methods and ML-based approaches like *Scaden* and *CIBERSORTx*, reflect the evolving landscape of bioinformatics in cancer research. ML methods stood out for their precision and ability to handle complex datasets, particularly in the context of TC. However, limitations such as the use of cancer-derived reference data for healthy tissue deconvolution and inconsistencies in normalization impacted the accuracy of cell type estimates, particularly in the GTEX cohort. Nevertheless, the results of this study lay a solid foundation for future work, demonstrating that with proper refinement, deconvolution tools can offer robust solutions for studying the TME.

The findings of this study have significant implications for TC research and TME studies at large and understanding the cellular composition of the thyroid TME is crucial for elucidating the mechanisms of immune evasion, cancer progression, and potential therapeutic targets. As demonstrated, deconvolution of bulk RNA-seq data offers a powerful approach for dissecting the cellular landscape of tumours, and improving the accuracy of these estimations could lead to more precise stratification of patient subgroups and better-tailored treatment options.

Looking ahead, several avenues exist to enhance the robustness and accuracy of deconvolution methods in the context of TME research. First, addressing the limitations identified in this study - such as using a more representative reference dataset and applying consistent normalization strategies - will be critical for improving cell type estimation. Testing a broader array of deconvolution methods with varying computational

approaches will also provide a more comprehensive understanding of the strengths and limitations of these tools in different tissue types and disease contexts.

Additionally, the development of an automated pipeline for preprocessing, normalization, and deconvolution would streamline the analysis process, improving efficiency and reproducibility. Such a pipeline would allow for more consistent application of deconvolution methods across datasets, reducing the impact of methodological variability. Moreover, once optimal deconvolution strategies are identified, future studies could expand into the analysis of cell eco-types and leverage clinical and experimental metadata to explore the functional relevance of specific cell populations in disease progression, patient outcomes, and therapeutic responses.

In conclusion, this study contributes to the field by providing an in-depth comparison of deconvolution methods applied to TC and healthy tissue. The findings underscore the potential of ML approaches to enhance the precision of TME deconvolution, while also highlighting areas for methodological improvement. With further refinement, these tools will play an increasingly pivotal role in cancer research, offering insights into the cellular dynamics of cancer and advancing personalized medicine approaches.

References

1. Mansourian AR. The immune system which adversely alter thyroid functions: a review on the concept of autoimmunity. Pak J Biol Sci. 2010;13(16):765-74.
2. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63.
3. Singh A, Ham J, Po JW, Niles N, Roberts T, Lee CS. The Genomic Landscape of Thyroid Cancer Tumourigenesis and Implications for Immunotherapy. Cells. 2021;10(5).
4. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunology, Immunotherapy. 2018;67(7):1031-40.
5. Menicali E, Guzzetti M, Morelli S, Moretti S, Puxeddu E. Immune Landscape of Thyroid Cancers: New Insights. Front Endocrinol (Lausanne). 2020;11:637826.
6. Lu DN, Zhang WC, Lin YZ, Jiang HY, He R, Li SL, et al. Single-cell and bulk RNA sequencing reveal heterogeneity and diagnostic markers in papillary thyroid carcinoma lymph-node metastasis. Journal of Endocrinological Investigation. 2024;47(6):1513-30.
7. Pu W, Shi X, Yu P, Zhang M, Liu Z, Tan L, et al. Single-cell transcriptomic analysis of the tumor ecosystems underlying initiation and progression of papillary thyroid carcinoma. Nat Commun. 2021;12(1):6058.
8. Ham J, Wang B, Po JW, Singh A, Niles N, Lee CS. Cancer-associated fibroblasts (CAFs) in thyroid papillary carcinoma: molecular networks and interactions. Journal of Clinical Pathology. 2021;74(12):759-65.
9. Avagliano A, Fiume G, Bellevicine C, Troncone G, Venuta A, Acampora V, et al. Thyroid Cancer and Fibroblasts. Cancers (Basel). 2022;14(17).
10. Tan JK, Awuah WA, Roy S, Ferreira T, Ahluwalia A, Guggilapu S, et al. Exploring the advances of single-cell RNA sequencing in thyroid cancer: a narrative review. Med Oncol. 2023;41(1):27.
11. Wusiman D, Li W, Guo L, Huang Z, Zhang Y, Zhang X, et al. Comprehensive analysis of single-cell and bulk RNA-sequencing data identifies B cell marker genes signature that predicts prognosis and analysis of immune checkpoints expression in head and neck squamous cell carcinoma. Heliyon. 2023;9(12):e22656.
12. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Current Opinion in Immunology. 2013;25(5):571-8.
13. Reynolds SR, Zhang Z, Salas LA, Christensen BC. Tumor microenvironment deconvolution identifies cell-type-independent aberrant DNA methylation and gene expression in prostate cancer. Clinical Epigenetics. 2024;16(1):5.
14. Danaher P, Kim Y, Nelson B, Griswold M, Yang Z, Piazza E, Beechem JM. Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. Nature Communications. 2022;13(1):385.
15. Hu M, Chikina M. Heterogeneous pseudobulk simulation enables realistic benchmarking of cell-type deconvolution methods. bioRxiv. 2023:2023.01.05.522919.
16. Dietrich A, Merotto L, Pelz K, Eder B, Zackl C, Reinisch K, et al. Benchmarking second-generation methods for cell-type deconvolution of transcriptomic data. bioRxiv. 2024:2024.06.10.598226.

17. Nguyen H, Nguyen H, Tran D, Draghici S, Nguyen T. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. Nucleic Acids Research. 2024;52(9):4761-83.
18. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nature Reviews Genetics. 2023;24(8):550-72.
19. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34(11):1969-79.
20. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nature Communications. 2020;11(1):5650.
21. Im Y, Kim Y. A Comprehensive Overview of RNA Deconvolution Methods and Their Application. Mol Cells. 2023;46(2):99-105.
22. Zhao W, Dovas A, Spinazzi EF, Levitin HM, Banu MA, Upadhyayula P, et al. Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. Genome Medicine. 2021;13(1):82.
23. Pu W, Shi X, Yu P, Zhang M, Liu Z, Tan L, et al. Single-cell transcriptomic analysis of the tumor ecosystems underlying initiation and progression of papillary thyroid carcinoma. Nature Communications. 2021;12(1):6058.
24. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell reports. 2019;26(6):1627-40. e7.
25. Su C, Xu Z, Shan X, Cai B, Zhao H, Zhang J. Cell-type-specific co-expression inference from single cell RNA-sequencing data. Nature Communications. 2023;14(1):4846.
26. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan G-C. Accurate estimation of cell-type composition from gene expression data. Nature Communications. 2019;10(1):2975.
27. Swapna LS, Huang M, Li Y. Guided-topic modelling of single-cell transcriptomes enables sub-cell-type and disease-subtype deconvolution of bulk transcriptomes. bioRxiv. 2023:2022.12.22.521640.
28. White BS, Reyniès Ad, Newman AM, Waterfall JJ, Lamb A, Petitprez F, et al. Community assessment of methods to deconvolve cellular composition from bulk gene expression. bioRxiv. 2022:2022.06.03.494221.
29. Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. Methods Mol Biol. 2020;2117:135-57.
30. Do VH, Canzar S. A generalization of t-SNE and UMAP to single-cell multimodal omics. Genome Biology. 2021;22(1):130.
31. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology. 2019;37(1):38-44.
32. Kassambara A, Herviou L, Ovejero S, Jourdan M, Thibaut C, Vikova V, et al. RNA-sequencing data-driven dissection of human plasma cell differentiation reveals new potential transcription regulators. Leukemia. 2021;35(5):1451-62.
33. Torregrosa-Cortés G, Oriola D, Trivedi V, Garcia-Ojalvo J. Single-cell Bayesian deconvolution. iScience. 2023;26(10):107941.

34. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nature Cancer. 2022;3(4):505-17.
35. Tran KA, Addala V, Johnston RL, Lovell D, Bradley A, Koufariotis LT, et al. Performance of tumour microenvironment deconvolution methods in breast cancer using single-cell simulated bulk mixtures. Nature Communications. 2023;14(1):5758.
36. Deng W, Li B, Wang J, Jiang W, Yan X, Li N, et al. A novel Bayesian framework for harmonizing information across tissues and studies to increase cell type deconvolution accuracy. Briefings in Bioinformatics. 2023;24(1).
37. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nature Genetics. 2015;47(10):1168-78.
38. Wang F, Yang F, Huang L, Song J, Gasser RB, Aebersold R, et al. Deep Domain Adversarial Neural Network for the Deconvolution of Cell Type Mixtures in Tissue Proteome Profiling. bioRxiv. 2023:2022.11.25.517895.
39. Cable DM, Murray E, Shanmugam V, Zhang S, Zou LS, Diao M, et al. Cell type-specific inference of differential expression in spatial transcriptomics. Nature Methods. 2022;19(9):1076-87.
40. Gynter A, Meistermann D, Lähdesmäki H, Kilpinen H. DeconV: Probabilistic Cell Type Deconvolution from Bulk RNA-sequencing Data. bioRxiv. 2023:2023.12.07.570524.
41. Im Y, Kim Y. A Comprehensive Overview of RNA Deconvolution Methods and Their Application. Molecules and Cells. 2023;46(2):99-105.
42. Zhang S, Pyne S, Pietrzak S, Halberg S, McCalla SG, Siahpirani AF, et al. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. Nature Communications. 2023;14(1):3064.
43. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: a review of recent applications. Human Molecular Genetics. 2017;26(R2):R216-R24.
44. Cai M, Zhou J, McKennan C, Wang J. scMD facilitates cell type deconvolution using single-cell DNA methylation references. Communications Biology. 2024;7(1):1.
45. Sutton GJ, Poppe D, Simmons RK, Walsh K, Nawaz U, Lister R, et al. Comprehensive evaluation of deconvolution methods for human brain gene expression. Nature Communications. 2022;13(1):1358.
46. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications. 2019;10(1):380.
47. Li H, Zhou J, Li Z, Chen S, Liao X, Zhang B, et al. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. Nature Communications. 2023;14(1):1548.
48. Tu J-J, Li H-S, Yan H, Zhang X-F. EnDecon: cell type deconvolution of spatially resolved transcriptomics data via ensemble learning. Bioinformatics. 2022;39(1).
49. Menden K, Marouf M, Oller S, Dalmia A, Magruder DS, Kloiber K, et al. Deep learning–based cell composition analysis from tissue expression profiles. Science Advances. 2020;6(30):eaba2619.
50. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell. 2018;173(2):283-5.

51. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013;45(6):580-5.
52. Feng S, Calinawan A, Pugliese P, Wang P, Ceccarelli M, Petralia F, Gosline SJC. Decomprolute is a benchmarking platform designed for multiomics-based tumor deconvolution. Cell Rep Methods. 2024;4(2):100708.
53. Cobos FA, Panah MJN, Epps J, Long X, Man T-K, Chiu H-S, et al. Effective methods for bulk RNA-seq deconvolution using scnRNA-seq transcriptomes. Genome Biology. 2023;24(1):177.
54. Hu M, Chikina M. Heterogeneous pseudobulk simulation enables realistic benchmarking of cell-type deconvolution methods. Genome Biology. 2024;25(1):169.
55. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics. 2020;2(3).
56. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology. 2018;19(1):15.
57. Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, et al. A Python library for probabilistic analysis of single-cell omics data. Nature Biotechnology. 2022;40(2):163-6.
58. Xi NM, Li JJ. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Systems. 2021;12(2):176-94.e6.
59. Hie BL, Kim S, Rando TA, Bryson B, Berger B. Scanorama: integrating large and diverse single-cell transcriptomic datasets. Nature Protocols. 2024;19(8):2283-97.
60. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford). 2019;2019.
61. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Research. 2022;50(W1):W216-W21.
62. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports. 2019;9(1):5233.
63. Andrade Barbosa B, van Asten SD, Oh JW, Farina-Sarasqueta A, Verheij J, Dijk F, et al. Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data. Nature Communications. 2021;12(1):6106.
64. Aran D. Cell-Type Enrichment Analysis of Bulk Transcriptomes Using xCell. Methods Mol Biol. 2020;2120:263-76.
65. Raskov H, Orhan A, Christensen JP, Gögenur I. Cytotoxic CD8+ T cells in cancer and cancer immunotherapy. British Journal of Cancer. 2021;124(2):359-67.
66. Fan J, Lyu Y, Zhang Q, Wang X, Li M, Xiao R. MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. Briefings in Bioinformatics. 2022;23(6).
67. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 2020;21(1):130.
68. Jeong Y, de Andrade ESLB, Thalmeier D, Toth R, Ganslmeier M, Breuer K, et al. Systematic evaluation of cell-type deconvolution pipelines for sequencing-based bulk DNA methylomes. Brief Bioinform. 2022;23(4).
69. Yi J, Zhong W, Wu H, Feng J, Zouxu X, Huang X, et al. Identification of Key Genes Affecting the Tumor Microenvironment and Prognosis of Triple-Negative Breast Cancer. Front Oncol. 2021;11:746058.

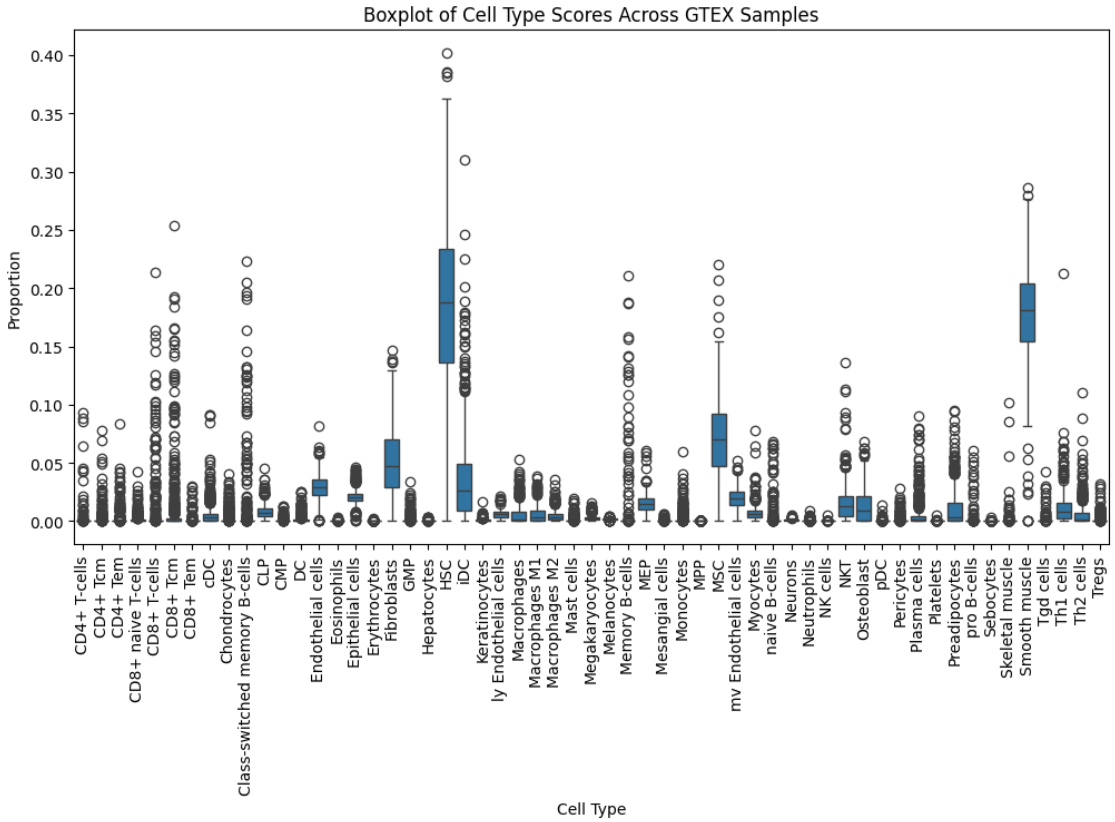
70. Wang K, Zhang Y, Xing Y, Wang H, He M, Guo R. Current and future of immunotherapy for thyroid cancer based on bibliometrics and clinical trials. Discov Oncol. 2024;15(1):50.
71. Yunna C, Mengru H, Lei W, Weidong C. Macrophage M1/M2 polarization. Eur J Pharmacol. 2020;877:173090.
72. Bresnick AR, Weber DJ, Zimmer DB. S100 proteins in cancer. Nature Reviews Cancer. 2015;15(2):96-109.
73. Shakib H, Rajabi S, Dehghan MH, Mashayekhi FJ, Safari-Alighiarloo N, Hedayati M. Epithelial-to-mesenchymal transition in thyroid cancer: a comprehensive review. Endocrine. 2019;66(3):435-55.
74. L. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology. 2016;17(1):75.
75. Lytal N, Ran D, An L. Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey. Frontiers in Genetics. 2020;11.
76. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Reports. 2019;26(6):1627-40.e7.
77. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, Jiang Y. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Briefings in Bioinformatics. 2020;22(1):416-27.

Attachments

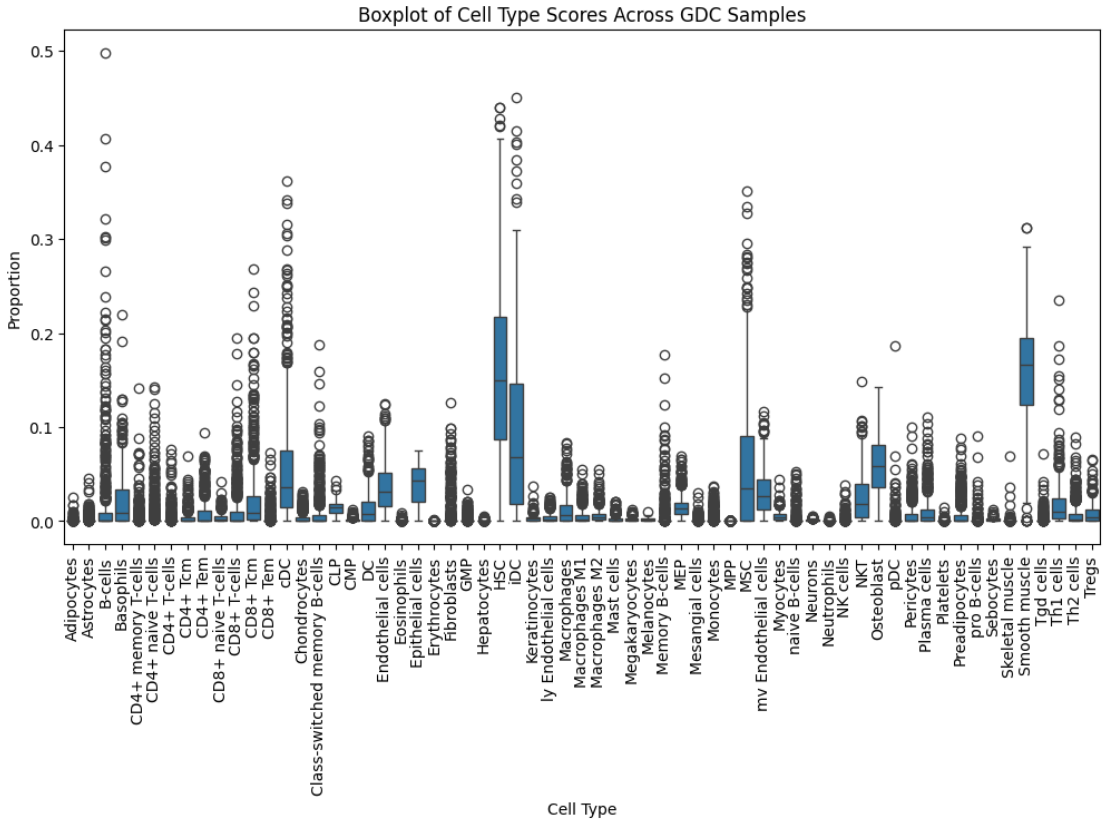
Supplementary Table 1: **Review of Cell-Type Deconvolution Approaches.**
This table summarizes the key characteristics of various deconvolution algorithms used to estimate cell type fractions in bulk RNA-seq data. Each algorithm is categorized based on its underlying computational approach (e.g., least squares methods, support vector regression, deep learning, etc.). Columns provide additional details such as whether the method is supervised or unsupervised, the type of regression model employed (linear or log-linear), whether the algorithm uses a signature or marker gene expression profile for cell type identification, and if scRNA-seq data is required for building the reference signature. Furthermore, the table indicates whether the methods account for gene expression variability, whether they estimate cell type fractions or perform in-silico purification, and the approximate number of cell types each method can handle (flexible or fixed). Algorithms highlighted in yellow are the ones selected for this comparative review.

Algorithm	*Classification*	*Supervised/ Unsupervised*	*Linear/Log --linear*	*Marker Gene Expression Profile (Signature)*	*scRNA-seq Data for Signature*	*Account for Gene Expression Variability*	*Cell Type Fractions*	*In-silico Purification*	*No. of Cell Types*
ABIS	1. Least Squares Methods	Supervised	Linear	Yes	No	No	Yes	No	Flexible
csSAM	1. Least Squares Methods	Supervised	Log-linear	No	No	No	No	Yes (group-mode)	Flexible
DSA	1. Least Squares Methods	Supervised	Linear	Yes	No	No	Yes	Yes (group-mode)	Flexible
DWLS	1. Least Squares Methods	Supervised	Linear	Yes	Yes	Yes	Yes	No	Flexible
EPIC	1. Least Squares Methods	Supervised	Linear	Yes	No	Yes	Yes	No	Flexible
MuSiC	1. Least Squares Methods	Supervised	Linear	Yes	Yes	Cross-subject Variance	Yes	No	Flexible
quantTIseq	1. Least Squares Methods	Supervised	Linear	Yes	No	No	Yes	Yes (group-mode)	Flexible
TIMER	1. Least Squares Methods	Supervised	Linear	Yes	No	No	Yes	Yes	Flexible
CIBERSORTx	2. Support Vector Regression (v-SVR)	Supervised	Linear	Yes	Yes	No	Yes	Yes (high resolution)	Flexible
CIBERSORT	2. Support Vector Regression (v-SVR)	Supervised	Linear	Yes	No	No	Yes	No	Flexible
CDSeq	4. Unsupervised NMF and Bayesian Approaches	Unsupervised	Linear	No	No	No	Yes	Yes (group-mode)	Flexible
DECODER	4. Unsupervised NMF and Bayesian Approaches	Unsupervised	Log-linear	No	No	No	Yes	Yes (group-mode)	Flexible
ISOpure	4. Unsupervised NMF and Bayesian Approaches	Semi-supervised	Linear	Yes	No	No	Yes	Yes (high resolution)	2
BayesPrism	4. Unsupervised NMF and Bayesian Approaches	Supervised	Linear	Yes	Yes	No	Yes	Yes (high resolution)	Flexible

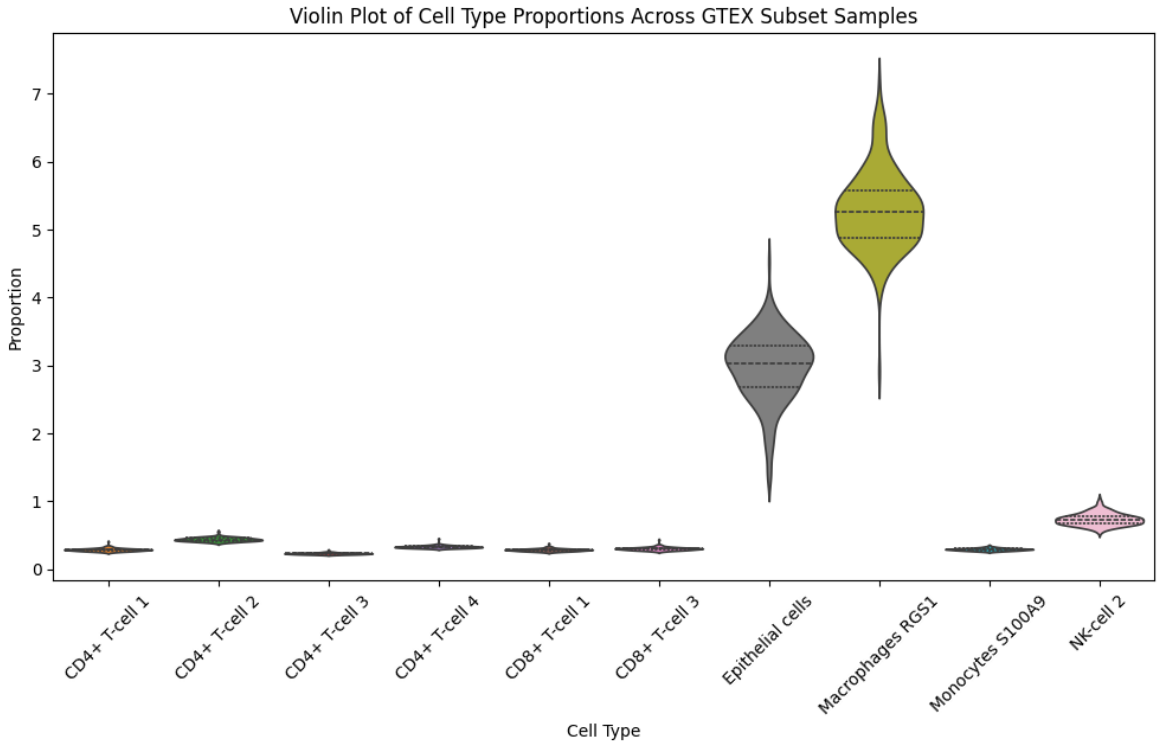
scpDeconv	5. Deep Learning Approaches	Supervised	Linear	Yes	Yes	Yes	Yes	Yes	Flexible
Bseq-SC	1. Least Squares & 4. Bayesian	Supervised	Linear	Yes	Yes	No	Yes	No	Flexible
xCell	3. Dimensionality Reduction Approaches	Supervised	Log-linear	Yes	No	No	No; score	No	Flexible
MCP-counter	7. Advanced Statistical Methods	Supervised	Log-linear	Yes	No	No	No; score	No	Flexible
DemixT	7. Advanced Statistical Methods	Semi-supervised	Linear	Yes	No	Yes	Yes	Yes	3
Demix	7. Advanced Statistical Methods	Semi-supervised	Log-linear	Yes	No	Yes	Yes	Yes	2
BLADE	7. Advanced Statistical Methods	Supervised	Linear	Yes	Yes	Yes	Yes	Yes (high resolution)	Flexible
ESTIMATE	7. Advanced Statistical Methods	Supervised	Log-linear	Yes	No	No	No; score	No	2
InstaPrism	7. Advanced Statistical Methods	Unsupervised	Linear	No	No	No	Yes	Yes	Flexible
DeconRNASeq	1. Least Squares Methods	Supervised	Linear	Yes	No	No	Yes	No	Flexible
scMD	4. Unsupervised NMF and Baysesian Approaches	Unsupervised	Linear	No	Yes	Yes	Yes	No	Flexible
DeconV	5. Deep Learning Approaches	Supervised	Linear	Yes	Yes	Yes	Yes	Yes	Flexible
Scaden	5. Deep Learning Approaches	Supervised	Linear	Yes	Yes	Yes	Yes	No	Flexible



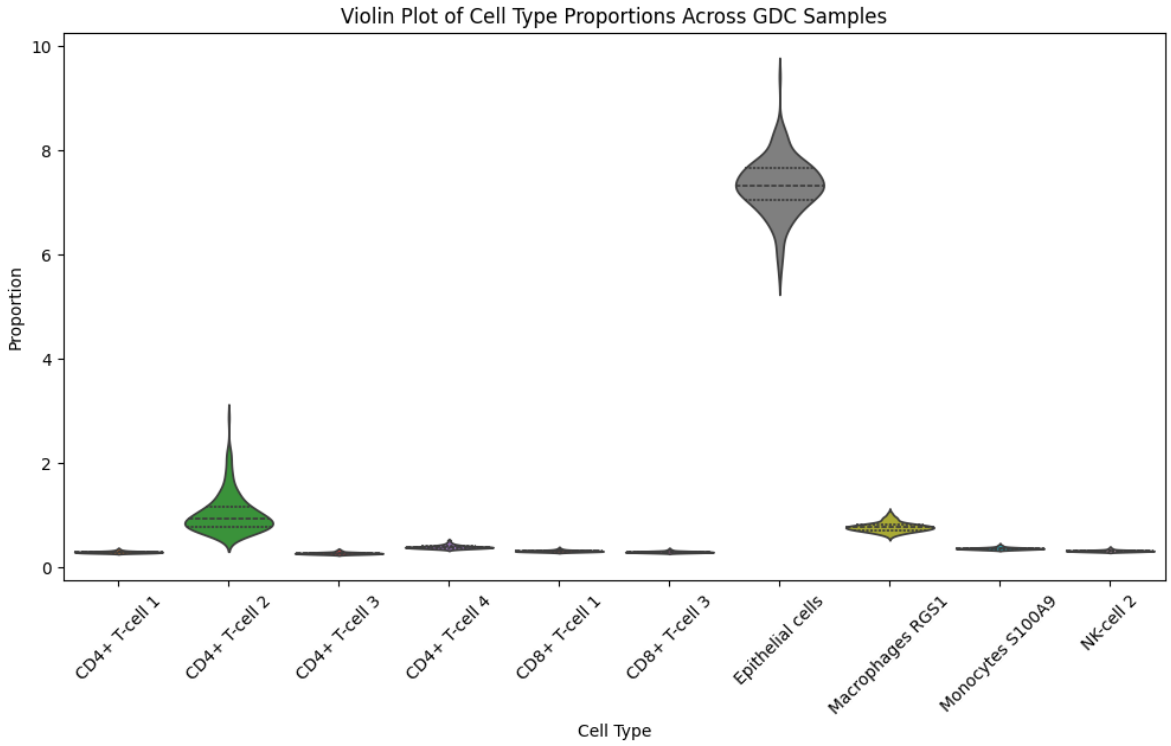
Supplementary Figure 1: *xCell* Enrichment Boxplots (GTEX Samples).
This boxplot depicts the distribution of *xCell* enrichment scores for various cell types across GTEX samples. Each box represents the interquartile range (IQR) of the enrichment scores, with the horizontal line inside the box marking the median score. The whiskers extend to the minimum and maximum values within 1.5 times the IQR. Outliers are displayed as individual points beyond the whiskers. The plot reveals variability in the proportion of cell types, with a great variety of immune cells, such as CD4+ and CD8+ T-cells, showing notable diversity in their distribution across samples. Notably, we can also see a high proportion of haematopoietic stem cells (HSE), mesenchymal stem cells (MSC) – stromal cells, fibroblast, and smooth muscle cells.



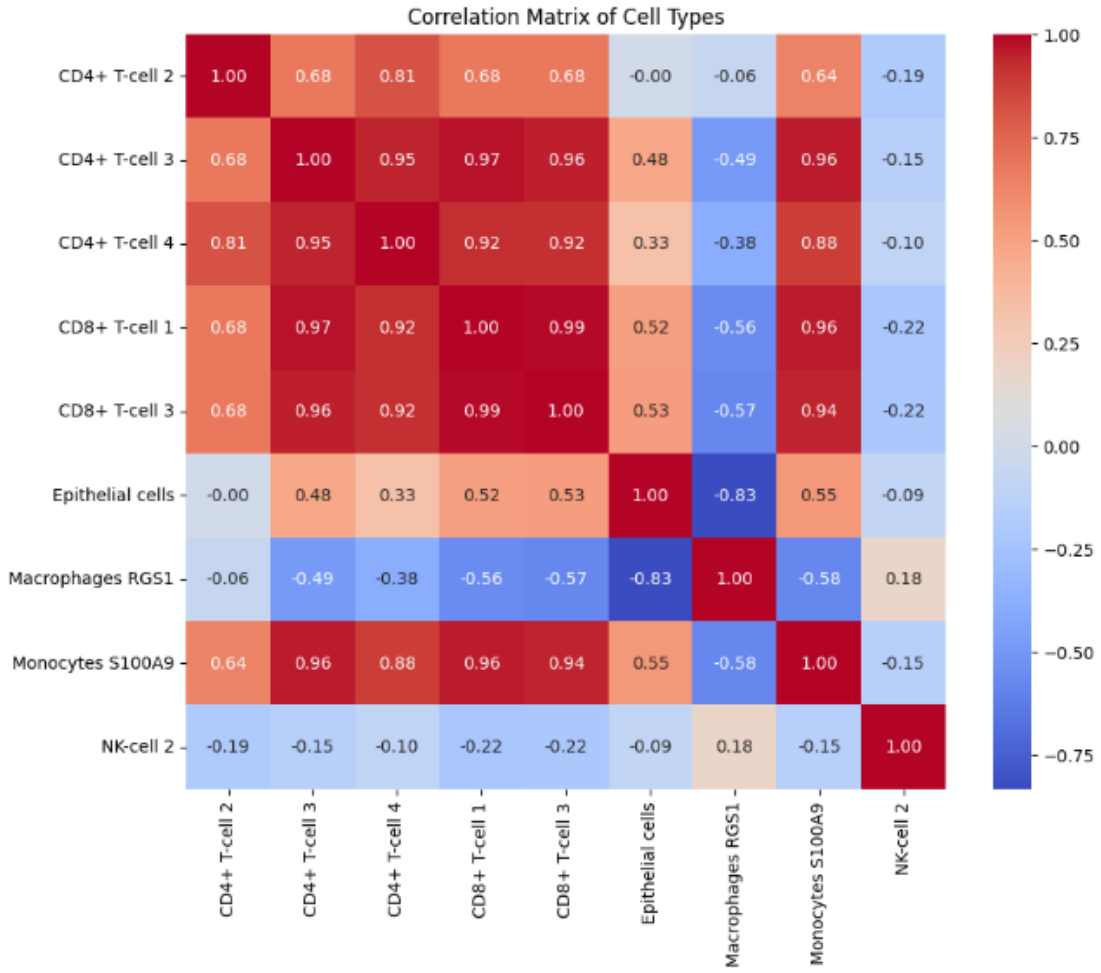
Supplementary Figure 2: *xCell* Enrichment Boxplots (GDC Samples).
This boxplot shows the distribution of cell type scores for each immune and stromal cell type across the GDC cohort. The y-axis represents the proportion of cells in each sample attributed to the specific cell types listed on the x-axis. Each box represents the interquartile range (IQR) of the cell type scores, with whiskers extending to the 1.5*IQR. The circles represent outliers beyond the whiskers. This plot provides insights into the estimated proportions of different cell types within the thyroid cancer microenvironment, highlighting the variation across samples and the prevalence of key cell types like MSC, macrophages, and CD8+ T-cells.



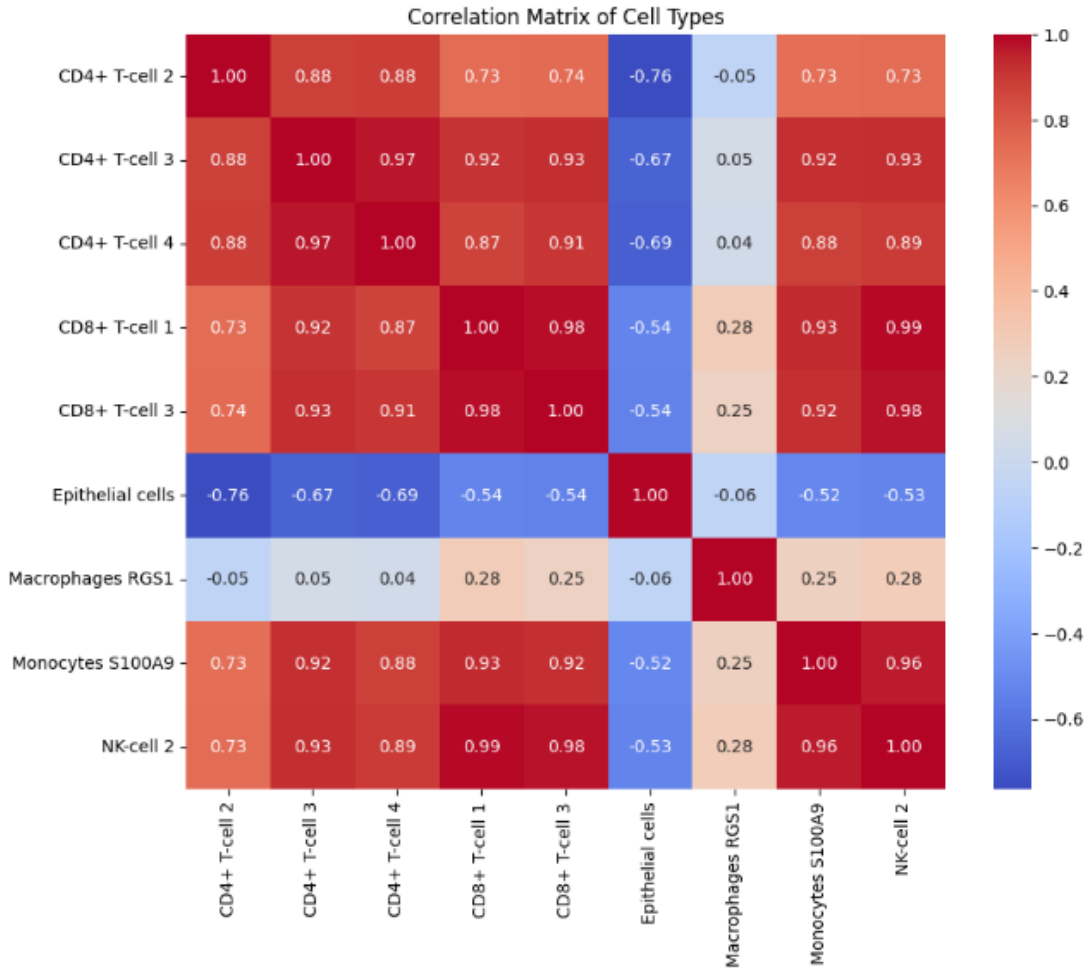
Supplementary Figure 3: Violin Plot with *BLADE* Estimated Cellular Proportions (GTEX Subset).
This violin plot presents the estimated fractions of selected cell types across a subset of GTEX samples, as determined by *Blade* deconvolution. The width of each violin reflects the density of samples for each cell type's proportion, and the dashed lines indicate the median and quartiles. CD4+ T-cells, CD8+ T-cells, and *Monocytes S100A9* are widely underrepresented in these distributions, while B-cell proportions weren't even estimated. Epithelial cells and Macrophages *RGS1* dominate the proportions in this dataset, followed by NK-cells 2 proportions in moderation.



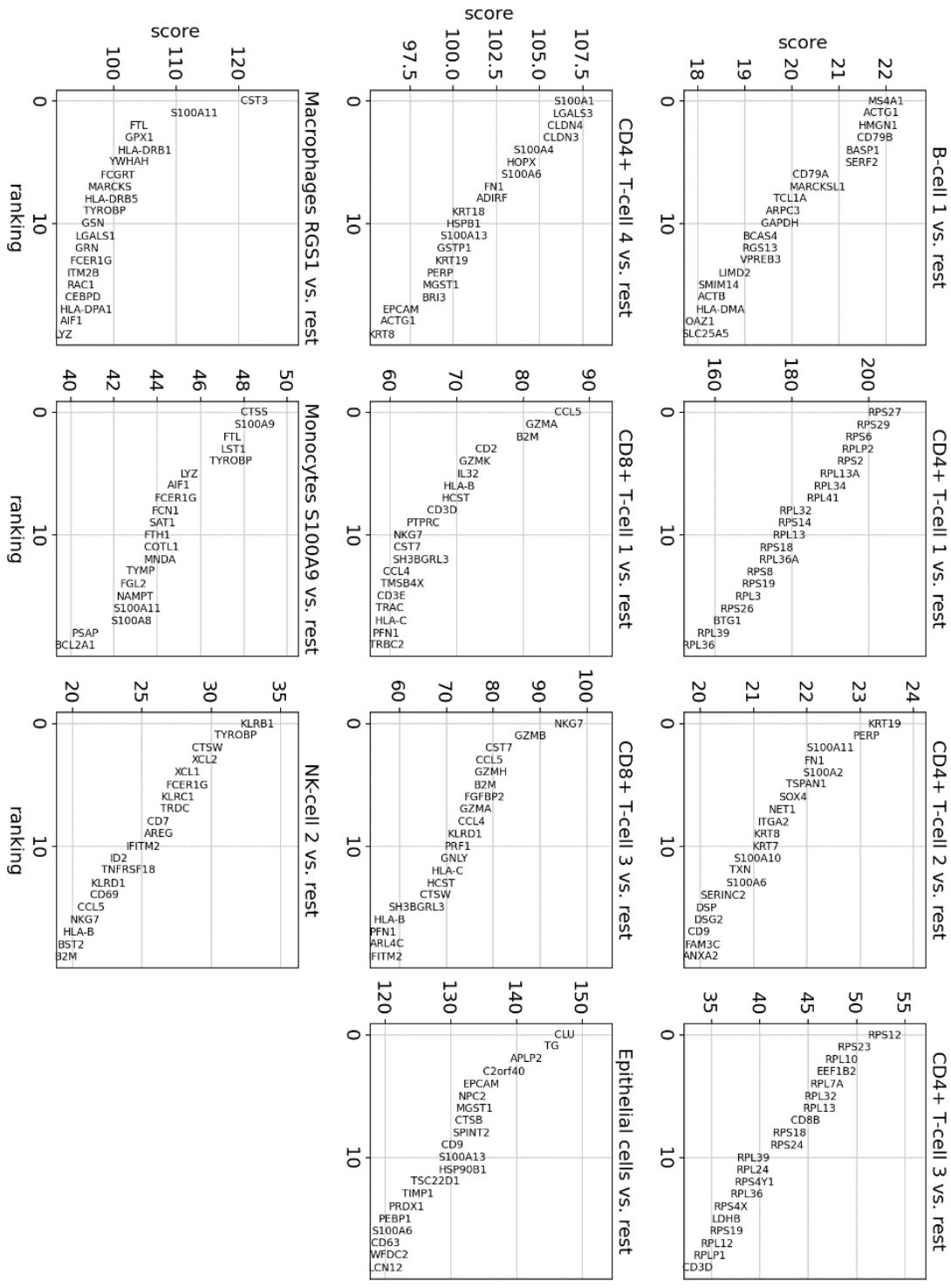
Supplementary Figure 4: Violin Plot with *BLADE* Estimated Cellular Proportions (GDC Subset).
The violin plot displays the distribution of cell type proportions across GDC subset samples, with each plot reflecting a different immune or stromal cell type. The width of each violin represents the density of samples at different cell type proportions, while the dashed line shows the median value. This visualization demonstrates how some cell types, like CD4+ T-cells and epithelial cells, are more prevalent in the GDC cohort than others, suggesting potential immune infiltration and EMT transition, respectively, hinting at their potential role in the TME.



Supplementary Figure 5: *BLADE* Correlation Heatmap of Cell Types (GTEX subset).
This heatmap illustrates the pairwise correlation matrix between cell types estimated by *BLADE* across GTEX samples. The colour scale represents Pearson correlation coefficients, with red indicating positive correlations and blue indicating negative correlations. Cells with high correlations, such as CD4+ and CD8+ T-cell subpopulations, suggest shared expression profiles or overlapping marker genes, while low or negative correlations between cell types such as Epithelial cells, Macrophages, and NK-cells indicate distinct biological roles.



Supplementary Figure 6: *BLADE* Correlation Heatmap of Cell Types (GDC subset).
This heatmap depicts the correlation matrix of cell types estimated across the GDC cohort. The matrix shows pairwise Pearson correlation coefficients between different cell types. Red squares represent strong positive correlations, while blue squares represent negative correlations. Notably, strong correlations are observed among different CD4+ T-cell subpopulations and between CD8+ T-cells, while weaker or negative correlations are seen for macrophages and epithelial cells. These trends highlight the relationships and potential interactions between various immune cells in the TME of the subset.



Supplementary Figure 7: Plot of Highly expressed Genes in each Cell Type.
This plot ranks the most highly expressed genes in each cell type from the single-cell reference data. The x-axis represents the ranking of the genes, while the y-axis displays the expression scores of the corresponding genes. Each panel represents a different cell type (e.g., B-cell 1, CD4+ T-cell 1, Macrophages *RGS1*, etc.), with the top-ranked genes that are most prominent in the reference data. This information serves as a reference point for understanding the gene expression profile of each cell type and is crucial for validating deconvolution methods and interpreting their results.

Supplementary Figure 8: Signature Matrix Dotplot.
This dotplot displays the expression levels of all marker genes across various cell types in the reference dataset. Each dot represents the mean expression of a marker gene within a specific cell type, where the colour intensity indicates the level of expression, and the size of the dot corresponds to the percentage of cells in the cell type that express that gene. The cell types include all curated from the marker gene dictionary in **Table 3**. This visualization aids in assessing which marker genes are most strongly associated with each cell type. Understanding each cell type signature can shed light into why certain cell types are more abundantly estimated than others, providing valuable context for deconvolution analyses.