|  |
| --- |
| YADA - Reference Free Deconvolution of RNA Sequencing Data  Dan Livne1\*, Tom Snir1, and Sol Efroni1  1Department of Life Sciences, Bar Ilan University Israel.  \*To whom correspondence should be addressed.  **Abstract**  We present YADA, a cellular content deconvolution algorithm for estimating cell type proportions in heterogeneous cell mixtures based on gene expression data. YADA utilizes curated gene signatures of cell type-specific marker genes, either obtained intrinsically from pure cell type expression matrices or provided by the user.  YADA implements an accessible and extensible deconvolution framework uniquely capable of handling marker genes alone as inputs. By relying solely on literature-supported cell type-specific signatures rather than full transcriptomic profiles from purified isolates, adoption barriers lower significantly. However, flexible inputs do not necessitate sacrificing rigor - predictions match metrics of current methodologies through an integrated optimization scheme balancing multiple inference algorithms. Efficiency optimizations via compiled runtimes enable rapid execution. Packaging as an importable Python toolkit promotes community enhancement while retaining codebase extensibility.  Validation studies demonstrate that YADA matches or exceeds the performance of current deconvolution methods on benchmark datasets. To demonstrate utility and enable immediate usage, we provide an online Jupyter notebook implementation coupled with tutorials. YADA provides an accurate, efficient, and extensible Python-based toolkit for cellular deconvolution analysis of heterogeneous gene expression data.  **Availability:** https://github.com/zurkin1/YADA  **Contact:** dani.livne@yahoo.com  **Supplementary information:** |

# Introduction and Historical Overview

Heterogeneity is an inherent attribute of tissues, which are composed of diverse cell types that carry out specialized functions. However, the analysis of tissue gene expression data frequently ignores this heterogeneity by portraying overall expression profiles in an oversimplified, unidimensional manner. While single-cell RNA sequencing (scRNA-seq) can reveal cell type-specific expression patterns, there are often barriers preventing its utilization, such as cost, sample viability concerns, or challenges inherent to isolating certain cell populations. To circumvent these limitations, computational approaches termed deconvolution methods have been developed to leverage transcriptomic data from bulk tissue samples and infer the proportional representation or enrichment score of individual component cell types.

The development of computational approaches to deconvolute heterogeneous gene expression profiles into constituent cell type contributions is rooted in early work on the mathematical separation of mixed tissue expression signals. These deconvolution methods, initially proposed by (Wolslegel et al., 2009) and (Clarke et al., 2010), use various algorithms ranging from quadratic programming (Gong et al., 2011),(Zhong et al., 2013) to nonnegative matrix factorization (Shen-Orr & Gaujoux, 2013). Initial methodologies relied on prior knowledge of signature gene sets exhibiting cell type-specific expression from microarray data or curated databases. For example, early seminal efforts leveraged immune cell-specific signature genes to deconvolve blood or peripheral blood mononuclear cell (PBMC) samples. Notably, (Aran et al., 2017) developed a deconvolution approach termed xCell that leverages a compensation algorithm to transform large-scale signature gene sets identified from isolated immune and stromal cell populations into cell type enrichment scores. By compensating for signature overlap, xCell outperformed earlier methods in deconvolving heterogeneous bulk RNA-sequencing data. (Erkkilä et al., 2010) introduced DSection, a highly accurate deconvolution algorithm leveraging a Bayesian statistical framework to estimate cell type proportions. DSection models the deconvolution process by assigning prior probabilities to cellular fractions and expression profiles, with the mixture data represented as a likelihood function. By iteratively transitioning between prior and posterior distributions, DSection refines estimates of proportions. Originally implemented in MATLAB with an accompanying web application, DSection demonstrated the potential of Bayesian methodologies for heterogeneous expression decomposition using Gibbs sampling.

Over time, the sophistication of deconvolution algorithms has progressively advanced. For instance, some later techniques incorporated reference expression data from pure cell populations to inform deconvolution analyses via linear unmixing models or utilization of reference profile similarity metrics. (Newman et al., 2015) presented CIBERSORT, a deconvolution algorithm using support vector regression. CIBERSORT provided a new gene signature map of 22 immune cell types for microarray datasets. Other methods, such as CIBERSORT, require the use of a full, pure gene expression matrix. LASSO regression calculates a linear regression for each mixture as a sum of its pure proportions. Other methods involve a nonnegative least square factoring of the mixture matrix using optimization under the assumption that the rows in the proportion matrix sum to one. This approach was used in deconRNASeq (Gong & Szustakowski, 2018) and CellMix (Gaujoux & Seoighe, 2013).

The proliferation of large-scale single-cell RNA sequencing (scRNA-seq) atlases has also contributed to the expansion of cell type-specific gene signatures for deconvolution. More recent methods have investigated machine learning-based approaches as well as statistical learning (Qiao et al., 2012), (Liebner et al., 2014).

Overall, deconvolution methodology has transformed from reliance on manual gene set curation to integrative strategies capitalizing on expanding single-cell and bulk tissue omics resources. The trajectory of innovation appears poised to further improve resolution and accuracy by interweaving biological knowledge of cell identity with advanced computational approaches. Moving forward, the integration of multiomics data and benchmarking across diverse tissue contexts will help solidify best practices. In parallel, translational applications of deconvolution will likely expand into precision medicine and evaluation of tissue microenvironment cell type dynamics.

Gaujoux and Shen-Orr introduced CellMix (Gaujoux & Seoighe, 2013), which incorporates iterative nonnegative matrix factorization (NNMF) with sum-to-one constraints on mixture fractions. This approach enabled stricter enforcement of nonnegativity than earlier regression-dependent methods. Gene set enrichment analysis (GSEA) constitutes another broadly adopted technique for assigning enrichment scores to marker gene sets based on their signal consistency and rankings within complex mixtures. Alternatively, deconICA (Czerwinska PhD Thesis, n.d.) imports blind source separation methods from signal processing, namely, independent component analysis, to minimize shared information across isolated expression components.

More recently, advanced solutions such as CIBERSORTx, AutoGeneS, Scaden, and DWLS have continued pushing boundaries through integrative machine learning. For example, Scaden trains deep neural networks on public single-cell RNA sequencing datasets to predict mixture fractions. Collectively, these innovations underscore the multifaceted algorithmic developments within the continually evolving field.

We present YADA, an accessible and high-performing toolkit for transcriptional deconvolution that uniquely accepts curated gene signatures alone as input rather than necessitating full profile matrices from purified reference populations. Such flexibility confers compatibility with real-world datasets where availability constraints or financial feasibility concerns often preclude assaying idealized “pure” cell types. By accommodating standard inputs, YADA removes adoption barriers that hinder the translational utility of preceding techniques. However, flexible inputs do not come at the cost of accuracy. Validated across an array of public benchmark suites - which we have carefully curated, documented, and standardized into a community resource - YADA matches or surpasses the capabilities of existing methodologies. This balanced combination of flexibility and performance derives from a rigorous mathematical framework and streamlined software implementation. Specifically, approximation algorithms infer surrogate profiles, enabling a cascade of computations to ultimately estimate mixture proportions matching ground truth labels. Tight integration with underlying C libraries ensures rapid execution even on basic hardware. An interactive Python interface promotes accessibility while retaining extensibility to additional single-cell data resources or custom marker sets. YADA advances the adoption of cellular deconvolution by eliminating unrealistic input stipulations while achieving best-in-class results, thereby bringing compositional dissection of complex tissue samples closer to realization for the broader research community.

# The Deconvolution Problem

The deconvolution problem refers to computationally estimating the cellular composition and proportional contributions of constituent cell types within heterogeneous biological samples. Complex tissues comprise diverse specialized cell populations that carry out distinct functions. However, standard genome-wide measurements of the gene expression profiles of these complex mixtures have not been performed.

Mathematically, the gene expression profile of a heterogeneous biological sample can be modeled as a weighted sum of the expression signals from each comprising cell type:

Mixture Expression Profile = Σ (Proportion Cell Type 1 × Expression Profile Cell Type 1) + Σ (Proportion Cell Type 2 × Expression Profile Cell Type 2) +...

The aim of deconvolution is to leverage this linear model to infer the proportions of each cell type, as well as their distinct expression profiles, within bulk tissue samples. This allows resolution of the composite profile into its component contributions. A variety of computational approaches, ranging from the utilization of cell type-specific expression markers to regression-based techniques, have been applied to solve this inverse problem.

This is a simple toy setup of this problem, with two mixtures, 1 and 2, and three cell types, A, B and C. We profiled the gene expression in 2 biologically heterogeneous mixtures, which we know consist of cells from 3 distinct cell types - Type A, Type B, and Type C.

We measured the expression levels of 5 different genes, genes 1-5, in each of the mixtures and cell types:

| Gene | Type A | Type B | Type C | Mixture 1| Mixture 2 |

|-------------|-------------|-------------|-------------|-------------|-------------|

| Gene 1 | 2 | 4 | 1 | 2.25 | 2.75 |

| Gene 2 | 1 | 3 | 5 | 2.5 | 3 |

| Gene 3 | 3 | 1 | 2 | 1.5 | 1.75 |

| Gene 4 | 5 | 2 | 4 | 2.5 | 3.25 |

| Gene 5 | 3 | 5 | 1 | 1.5 | 3.73 |

*Table 1 - Pure and mixed datasets.*

We wanted to deconvolute mixtures 1 and 2 to determine what proportion of cells in each mixture came from types A, B and C. One possible deconvolution solution would be:

* Mixture 1 = 50% Type A cells, 25% Type B cells, 25% Type C cells
* Mixture 2 = 25% Type A cells, 50% Type B cells, and 25% Type C cells.

Indeed, if we take these suggested proportions, the mixture columns are the sum of the pure types of columns multiplied by the proportions.

The goal of the deconvolution analysis is to infer these proportions mathematically using the given data. In practice, there are many mixing columns representing patients at different time points.

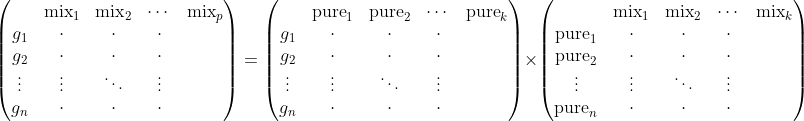
We can formulate the deconvolution problem mathematically as the following matrix decomposition problem. We define the following three matrices:

* Let *P* be an *n × k* pure gene expression matrix that contains *k* cell types and *n* genes.
* Let *W* be the unknown *k × p* matrix where each element of *W* contains the proportion of a particular cell type in a particular mixture.
* Let *M* be an *n × p* mixture expression matrix that contains the observed *n* gene expression levels in different mixtures, where *p* is the number of tissue samples (mixtures).

We assume the following linear relationship:

(1)

Or in matrix notation:



Thus, given *M* and *P* (or just a list of marker genes)*,* our task is to find *W*. In a typical gene expression profiling setting, *M* and *P* are measured using a gene expression microarray or RNA-seq. Given the known *M* and *P* measured by gene expression profiling, the task is to solve for *W*. Since *W* contains biological proportions, its elements must satisfy nonnegativity constraints. However, this constraint is not inherently enforced in naive matrix decomposition. Thus, deconvolution methods must implement algorithms that impose nonnegativity, often through iterative update procedures. Some authors (e.g., (Wolslegel et al., 2009) or (Gaujoux & Seoighe, 2013)) use an iterative process of matrix decomposition to enforce nonnegativity. YADA implements the Karush-Kuhn-Tucker solver under nonnegative constraints.

Deconvolution methodologies confront multiple challenges, the chief of which is a lack of reliable reference profiles *P*. The assumed availability of a comprehensive expression matrix *P* from purified cell constituents fails to reflect practical reality. More often, heterogeneity within cellular source material precludes the isolation of genuine in vivo “pure” populations. Even for artificially selected subtypes, requisite scaling and normalization remain frequently incompatible or unknown relative to target mixture data *M*. Consequently, typical reliance on limited marker gene sets poses a substantially greater computational hurdle. Successful deconvolution under such incomplete information constraints demands algorithms that are sufficiently robust to ignore absent full transcriptional characterizations. However, most of the current techniques presuppose this unavailable knowledge a priori. The development of revised methods capable of deducing cellular admixtures from only sparsely indicative genetic signatures, rather than absolute profiles, necessitates an altogether different conceptual and mathematical approach. Ultimately, universally accessible benchmark subpopulations may not exist. Even viable candidates would reflect artifacts of selection and processing absent in naturally co-occurring cells. Progress lies in embracing these realities rather than ignoring inherent biological complexity.

# Deconvolution Challenges

The foundational linear mixture model does not fully capture biological realities. The inclusion of an uncharacterized noise factor precludes fixed error distribution assumptions in microarray data. RNA sequencing profiles violate linearity through normalization, which obscures interrelationships between component cell types. While the core biochemical parameters underpinning Affymetrix microarrays appear linear, assuming pure superposition fails to account for confounding variables that distort additive directness in practice. The isolation and processing of intact RNA from tissue samples can lead to contamination and degradation, compromising accurate expression measurements. Sample amplification and hybridization protocols introduce further uncertainties separate from true transcriptional activity. The affinity of probe sequences varies beyond nucleotides alone, with spatial artifacts such as uneven hybridization or chip printing influencing the observed intensities. Optical excitation and emission inefficiencies are coupled with probe-specific quenching to skew fluorescence readouts.

Quality control and background noise removal are necessary. Often, a normalization transformation is performed, moving the data to a normal scale. We cannot guarantee that the error is normal and standard, but often, standard random noise is added to the model.

RNA-seq data also suffer from nonlinearity but for different reasons. Unlike microarray, RNA-seq is a competitive assay; that is, different transcripts compete over the same limited set of reads. This limitation has major effects. First, some competing genes might exist in two different marker gene sets. In this case, we need a way to “compensate” one of the genes in the pair for the existence of the second gene.

Second, the pool of genes was generated by sharing all the genes in the mixture. For example, in TPM normalization, the size of the library is one million reads, and the gene expression values are the number of reads identified for that gene. Since these reads are now “missing” from other competing genes, the TPM reads resemble more of a ratio or proportion of a gene out of one million reads. Ratios, as opposed to percentages, are not additive; i.e., they suffer from Simpson’s paradox, which implies high nonlinearity.

Some researchers (Monaco et al., 2019) suggest a preprocessing step to assess the abundance of mRNA produced by each cell type for each gene. (Aran et al., 2017) developed a compensation matrix. YADA uses an ensemble approach to handle noise. The ensemble learning method is a standard machine learning method that improves performance and prevents overfitting (Hansen et al., n.d.), (Kuncheva, 2004). The main idea is to split the learning data between multiple models and aggregate their results to one answer. YADA uses this approach, splitting the large list of marker genes into hundreds of independent workers that estimate the proportions. Aggregation is performed by averaging the results. Our tests show that this approach is the most efficient way to handle the noise in the data. Deconvolution algorithms must address both missing reference data and deviations from simplistic linear mixtures. The former complicates model inference, while the latter reduces descriptive accuracy if unaddressed. Ongoing efforts exist to improve imputation and generalization. However, integrative strategies combining multiomic data or mechanistic knowledge may prove necessary to fully surmount these challenges.

# YADA Design and Implementation

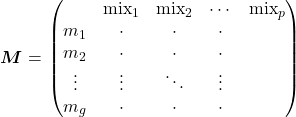
The primary input is an expression matrix *M* with gene IDs as row labels and sample IDs as column labels. This mixture data should follow standard RNA-sequencing normalization procedures (e.g., transcripts per million). The secondary input is a cell type marker gene list with one gene set per estimated constituent population. Ideally, markers should be manually curated based on the literature. However, these predicted markers can suffice for preliminary analyses when accompanied by confidence rankings and checked against known databases. Each list should provide as many specific and selective markers as possible, with a minimum of approximately 50 genes per group. Overlaps between lists are permissible as long as the main drivers of list uniqueness remain.

Given only the mixture expression matrix *M*, the principal computational challenge becomes calculating the unknown proportion matrix *W* in equation (1), without access to reference profiles in *P*. YADA's core algorithm tackles this by generating an approximation for the unknown matrix *W*. Next, we describe the high-level algorithm of YADA.

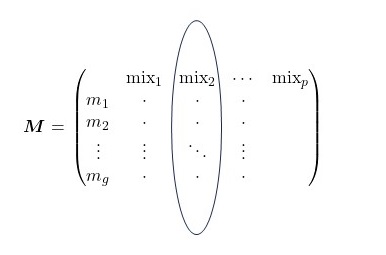
# YADA Algorithm

1. Data preparation and cleanup of the pure (if given) and the mixture matrix. Removal of missing data. Anti-log. Cleanup of duplicate genes.
2. If a pure matrix was given instead of a marker list, gene list selection is performed for each cell type.
3. Approximate the proportion matrix *W* using columns from *M* and the names of the marker genes.

Let *M* be:



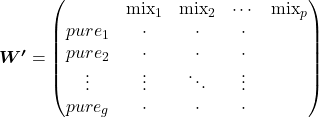
It has dimensions *p \* g,* where *p* is the number of mixtures and *g* is the total number of marker genes. We use the matrix *M* to approximate the matrix *W*. For each cell type *k*, we select the mixture (the column) whose gene expression across these cell type marker genes is maximal. In other words, we looped over each cell type, and for each cell type, we selected a specific mixture column with the maximum expression. That mixture satisfies that on that cell type group of marker genes its expression value is significantly greater than the other mixtures. Assuming that enough mixtures are provided, one such mixture with high expression values should exist. This mixture replaces the cell type of the unknown pure column.



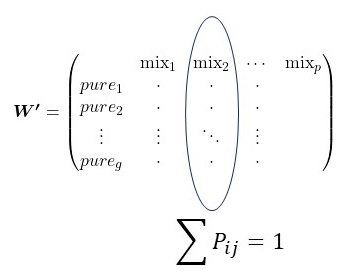
1. We then measured the distances of all the mixtures in *M* from that specific mixture (only on the set of the cell type marker genes) to create the row of estimated proportions for cell type *k*.

The distance between the vectors of genes can be calculated as the simple average distance per gene expression value across the gene sets.

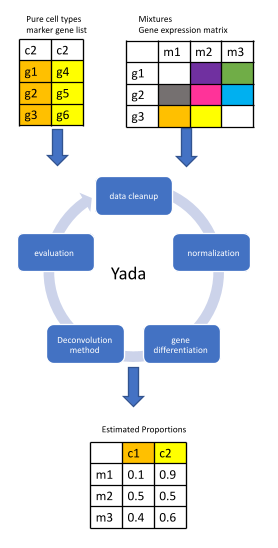
1. We now obtain an approximated proportion matrix *W’*:



1. To further refine the cell type proportion estimates in *W'*, an additional constraint is imposed requiring the fractions for each mixture sample to sum to unity. By solving for this added mixture-level sum constraint through an optimization scheme, calibrated approximations of the genuine underlying proportions matrix emerge. Essentially, after the initial approximation, imposing expected invariant sums refines posterior estimates to better reflect biological reality. The subsequent optimized *W'* matrix constitutes the final output - proportion fractions summing to unity across all samples, while maximizing concordance with input expression and marker evidence.



The deconvolution process is summarized in the following diagram.



*Figure 1 - YADA Algorithm Pipeline*

*Figure 1-Deconvolution pipeline. Cleanup is needed to remove outliers from the data. Normalization guarantees the same range of values in mixtures. Gene differentiation is needed in the case of a complete pure matrix to select marker genes. Deconvolution is the core algorithm. Evaluation was performed using Pearson and Spearman correlation.*

# Results

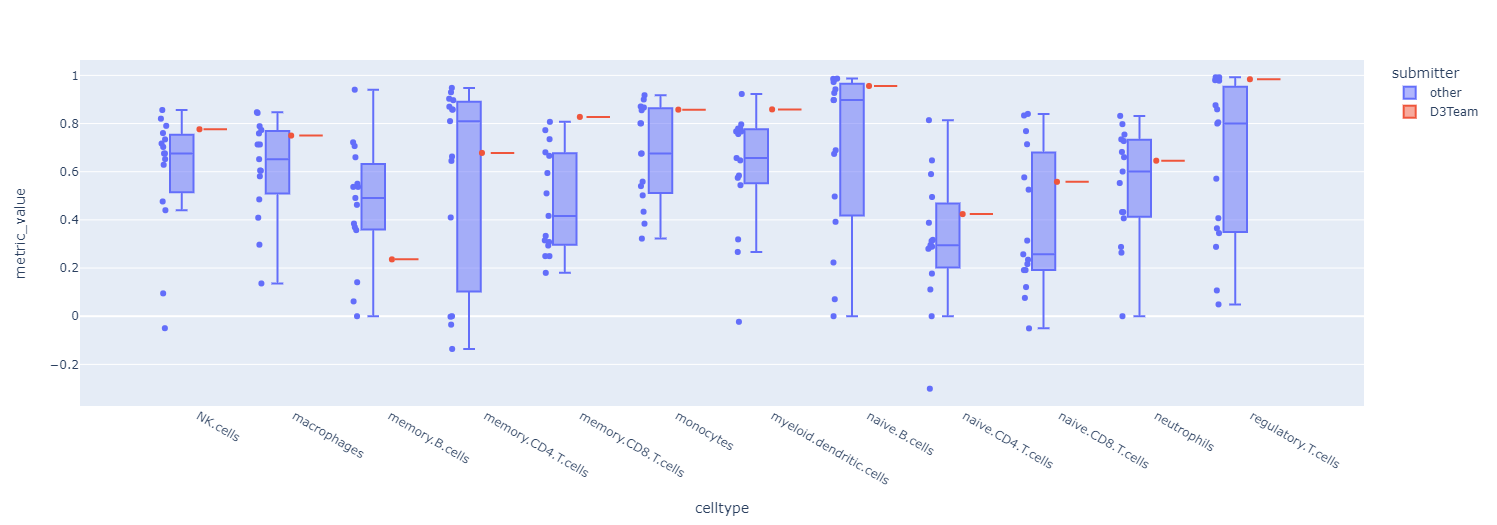
With our GitHub data folder, we compiled renowned datasets comprising mixtures of microarray and RNA sequencing data samples from various cell types, each with their respective ground truth proportions. These datasets encompass diverse instances, such as DSA (Zhong et al., 2013), Abbas *et al*. (Wolslegel et al., 2009), EPIC (Racle et al., 2017), TIMER (Jiang et al., 2016), Pert (MacBeath et al., 2017), Rat Brain (MacBeath et al., 2017), and CIBERSORT (Newman et al., 2015). Our analysis involved the application of YADA’s deconvolution algorithm alongside alternative methodologies, with YADA demonstrating comparable or superior performance across most patients. Notably, in the context of CIBERSORT, YADA exhibited an average mean square error of 0.879, outperforming CIBERSORT’s average mean square error of 1.265.

Furthermore, during the SAGE bioinformatic DREAM challenge (Tumor Deconvolution DREAM Challenge - Syn15589870, n.d.), YADA achieved the highest scores in microarray datasets. Specifically, our results from the latest round of the microarray fine challenge highlight YADA's attainment of the highest score. These datasets were integral components of this round, and YADA's rank among the 15 different methods used is specified in the last column of the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **dataset.name** | **platform** | **scale** | **normalization** | **YADA’s rank** |
| FIVA1 | Affymetrix HG-U133 Plus 2.0 | Linear | MAS5 | 4 |
| FIAS5 | Affymetrix HG-U133 Plus 2.0 | Log2 | RMA | 7 |
| FIVA2 | Affymetrix HG-U133A | Linear | MAS5 | 1 |
| FIAS6 | Affymetrix HG-U133 Plus 2.0 | Log2 | fRMA | 2 |

*Table 2 - Challenge datasets and YADA’s ranking*

The following box plot shows the per-cell type Spearman correlation results in that round, and the YADA top scores are marked by red lines.



*Figure 2 - Challenge results per cell type and team.*

# YADA’s Theorem

Extensive validation studies across numerous public benchmark datasets reveal a remarkable consistency in YADA’s performance - marker gene inputs alone confer equivalent or exceeding accuracy versus utilization of complete cell type-specific expression reference profiles. This phenomenon reoccurs again and again in our tests and repeated itself also during the training for the DREAM challenge.

This persistent empirical observation during training for DREAM challenges and on independent mixtures prompted deeper interrogation into the drivers of this unexpected parity. The marker set performed just as well as the thousands of genes per cell type. The key insight is that once a signature selectively captures the essence of what uniquely identifies a cell, having more extraneous genes doesn't improve separation ability for deconvolution tasks. We mathematically formalize the thresholds and properties ensuring marker genes suffice for decomposition tasks. The principal implication of this theorem is that access to complete transcriptional profiles from pure cell isolates is unnecessary for accurate deconvolution utilizing YADA. Rather, only a list of literature-supported, selectively expressed marker genes is strictly required as input to recapitulate ground truth mixture proportions across a diversity of samples.

## Theorem 1:

Consider a mixture model with n cell types, where pure population gene expression profiles follow independent negative binomial distributions with parameters r and p. Given:

- x mixture samples, with x > 30.

- Marker gene sets for each cell type covering 0.8% of the total genes.

Then the predicted proportions from the YADA deconvolution algorithm subsequently matched the empirical proportions with ≤ 5% deviation in the Pearson correlation over a series of computational experiments using either mix or pure\* matrix approximation steps.

\*i.e., using either the mix or pure columns in step 3 of the algorithm.

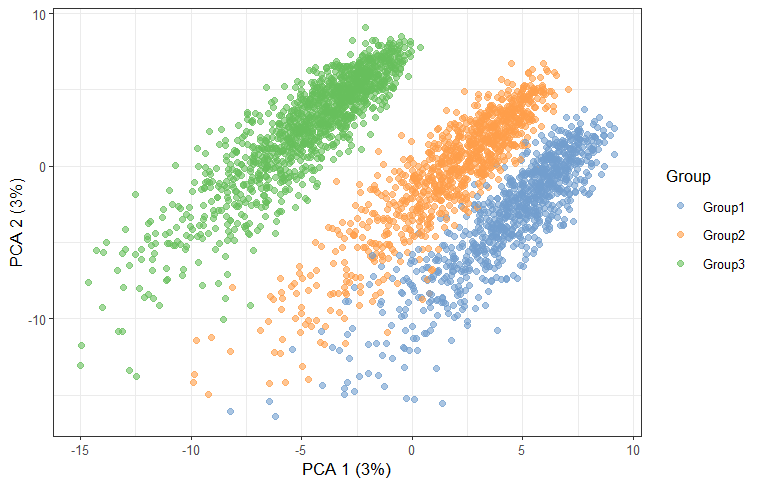
Proof:

- The proof goes by induction on n.

- The case from n to n+1. We select two cell types A and B, and treat their merged cells content as a new cell type C. Assuming the induction hypothesis, we perform deconvolution using YADA on n cell types. Using the results for C, we run another deconvolution using YADA for two cell types, A and B, on subsets of size two to obtain the required result. Note that this method requires that n be larger than two.

- Proof induction baseline. n such that for n cells the theorem holds for any number of mixtures. Consider the base case of n=3 cell types. As the theorem constitutes a probability statement, validation relies on empirical demonstration through simulation methods. Splatter (Zappia et al., 2017) represents an established RNA sequencing simulator leveraging validated distributions such as the negative binomial. Adoption as benchmarking infrastructure by hundreds of studies supports suitability. Executing Splatter with following parameters generated in silico cell type expressions and markers exhibiting control over mean-variance dependencies and gene differentiation:

*sim <- splatSimulate(params, nGenes = 10000, group.prob = c(0.3, 0.3, 0.4), method = "groups", batchCells = 3000)*



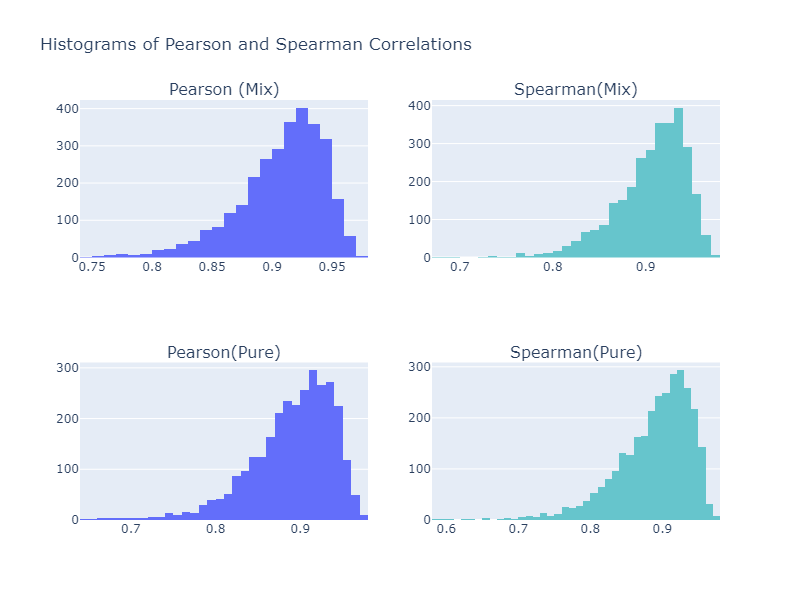
*Figure 3 - PCA results of the Splatter simulation.*

Using the cells from the simulation, we ran the following pipeline.

* Three cells were randomly sampled from each simulated population.
* The transcriptional mixtures were computed by weighing the sampled cell expression profiles using mixing proportions drawn from a Dirichlet distribution.
* Technical noise was added by adding independent normal variates to each gene with μ=1 and σ=0.4.
* YADA deconvolution is executed using both mix and pure matrix approximations at step 3 of the algorithm.
* The pipeline was repeated for 1000 independent trials.

Contrary to initial expectations, utilizing the mixture approximation matrix conferred marginally improved deconvolution accuracy over reliance on simulated pure profiles - a persistent trend across 1000 trials. Specifically, the average Pearson correlation coefficient between the predicted and true proportions increased by 0.2 when leveraging mixtures. Furthermore, Kolmogorov‒Smirnov tests demonstrated statistically significant differences between correlation distributions, with mixture-based shifts toward higher fidelity outcomes. Taken together, the descriptive and inferential statistics strongly support the theorem's assertion; given adequate samples and signature genes, markers alone contain sufficient information to match or exceed full profile inputs.

All simulation scripts, intermediate outputs, and analytical notebooks have been made publicly accessible via our GitHub repository to enable reproducibility as well as extensions exploring wider parameter spaces. The resources packaged aim to facilitate future benchmarking and comparative evaluations by the computational biology community.



*Figure 4 - Comparison of Pearson and Spearman correlation histograms when using either the mixture or the pure matrices.*

*The test results reveal a KS distance of 0.1663583 (p=7.2387309e-37).*

# Discussion

Reliable estimation of cell type fractions within heterogeneous tissue samples has become imperative as transcriptomics permeates precision medicine initiatives. However, conventional RNA sequencing provides only mixture-convolved expression averages, obscuring signals from minor constituents. Deconvolution computationally addresses this via selective regression, disambiguating composite profiles into constituent cell type contributions. This approach empowers the quantification of true underlying biology.

As demonstrated, YADA enables robust and accessible decompositions by harmonizing mathematical rigor with usability. Its easy adoption across platforms, rapid execution harnessing optimized numerical libraries, and coder-friendly Python integration lower barriers for translating deconvolution into mainstream assays. Uniquely, documented code paths and default marker sets facilitate further customization or method swapping to match novel biological contexts.

Critically, reliance solely on literature-curated signatures, rather than purely isolated references, circumvents unrealistic data expectations. By proving marker gene sufficiency given mixture diversity, YADA accommodates real-world constraints. High accuracy is retained under this flexibility, and the benchmark and DREAM challenges result in affirming performance that reaches or exceeds that of the state of the art. This abolishes the purity requirements that previously hindered adoption.

The ensemble algorithm framework also promotes extensibility and evolution as innovation continues within this nascent domain. With strong mathematical foundations in place, bespoke modeling innovations around emerging data types or niche tissue types can readily integrate into established workflows. Therefore, keeping pace with frontier technologies through modular enhancement enabled by YADA's accessible design.

**References**

Aran, D., Hu, Z., & Butte, A. J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*, *18*(1), 220. https://doi.org/10.1186/s13059-017-1349-1

Clarke, J., Seo, P., & Clarke, B. (2010). Statistical expression deconvolution from mixed tissue samples. *Bioinformatics (Oxford, England)*, *26*(8), 1043–1049. https://doi.org/10.1093/bioinformatics/btq097

*Czerwinska PhD thesis*. (n.d.). Retrieved February 25, 2020, from https://urszulaczerwinska.github.io/UCzPhDThesis/

Erkkilä, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I., & Lähdesmäki, H. (2010). Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics (Oxford, England)*, *26*(20), 2571–2577. https://doi.org/10.1093/bioinformatics/btq406

Gaujoux, R., & Seoighe, C. (2013). CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*, *29*(17), 2211–2212. https://doi.org/10.1093/bioinformatics/btt351

Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., & Szustakowski, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027156

Gong, T., & Szustakowski, J. D. (2018). *BIOINFORMATICS APPLICATIONS NOTE DeconRNASeq : a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data*. *29*(8), 1083–1085. https://doi.org/10.1093/bioinformatics/btt090

Hansen, L., & P. S.-I. T. on P. A., & 1990, undefined. (n.d.). Neural network ensembles. *Computer.Org*.

Jiang, P., Pignon, J.-C., Aster, J. C., Shen, H., Novak, J., Liu, J. S., Liu, X. S., Zhao, H., Severson, E., Rodig, S., Li, B., Li, T., & Signoretti, S. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biology*, *17*(1), 1–16. https://doi.org/10.1186/s13059-016-1028-7

Kuncheva, L. (2004). *Combining pattern classifiers: methods and algorithms*.

Liebner, D. A., Huang, K., & Parvin, J. D. (2014). MMAD: Microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics*, *30*(5), 682–689. https://doi.org/10.1093/bioinformatics/btt566

MacBeath, G., Schoeberl, B., Schelker, M., Klipp, E., Feau, S., Du, J., Raue, A., & Ranu, N. (2017). Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-02289-3

Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y. Y., Carré, C., Burdin, N., Visan, L., Ceccarelli, M., Poidinger, M., Zippelius, A., Pedro de Magalhães, J., & Larbi, A. (2019). RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Reports*, *26*(6), 1627-1640.e7. https://doi.org/https://doi.org/10.1016/j.celrep.2019.01.041

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., & Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, *12*(5), 453–457. https://doi.org/10.1038/nmeth.3337

Qiao, W., Quon, G., Csaszar, E., Yu, M., Morris, Q., & Zandstra, P. W. (2012). PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. *PLoS Computational Biology*, *8*(12). https://doi.org/10.1371/journal.pcbi.1002838

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E., & Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *ELife*, *6*, 1–25. https://doi.org/10.7554/eLife.26476

Shen-Orr, S. S., & Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current Opinion in Immunology*, *25*(5), 571–578. https://doi.org/10.1016/j.coi.2013.09.015

*Tumor Deconvolution DREAM Challenge - syn15589870*. (n.d.). Retrieved April 16, 2019, from https://www.synapse.org/#!Synapse:syn15589870/wiki/583086

Wolslegel, K., Seshasayee, D., Modrusan, Z., Clark, H. F., & Abbas, A. R. (2009). Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PLoS ONE*, *4*(7), e6098. https://doi.org/10.1371/journal.pone.0006098

Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, *18*(1), 174. https://doi.org/10.1186/s13059-017-1305-0

Zhong, Y., Wan, Y. W., Pang, K., Chow, L. M. L., & Liu, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, *14*. https://doi.org/10.1186/1471-2105-14-89