Single-cell transcriptome and chromatin accessibility data integration reveals cell specific signatures

Andrés Quintero[1,2,°], Anne-Claire Kröger[2] and Carl Herrmann[2,]*

[1]Division of Neuroblastoma Genomics, German Cancer Research Center, Heidelberg, Germany.

[2]Health Data Science Unit, Medical Faculty Heidelberg and BioQuant.

**Correspondence: carl.herrmann@uni-heidelberg.de*

GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

The ability to integrate multiple layers of omics data plays an essential role in understanding the complex interplay of different molecular mechanisms that give rise to cellular diversity.

To address this challenge we implemented Integrative Iterative Non-negative Matrix Factorization (i2NMF), a computational method to dissect genomic signatures from multi-omics data sets.

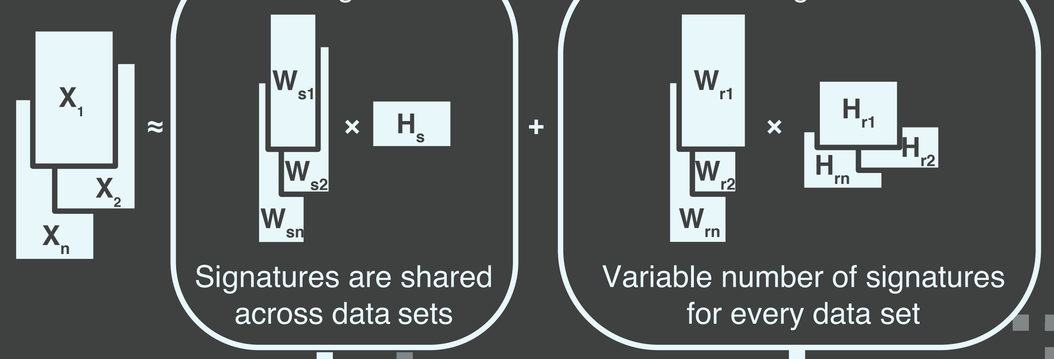
i2NMF was implemented as an extension of the R package Brat-wurst available in Github.

https://github.com/wurst- theke/bratwurst

We applied i2NMF to :

i2NMF workflow:

Common features should be shared across columns, e.g. gene or sample IDs.



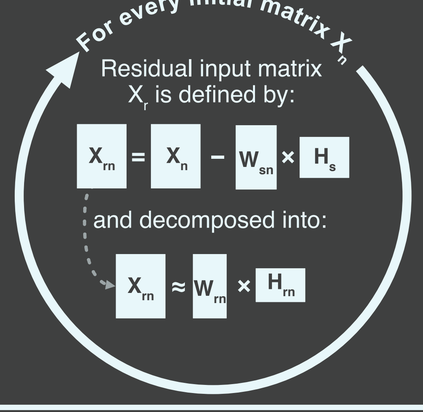
1. Starting from two or more non-negative matrices, i2NMF initially decomposes the shared effect across them, using integrative NMF (iNMF).

Solving the following problem:

$$\min_{\substack{W_s\geq 0, H_s\geq 0,\\ H_v\geq 0}} \sum_{n=1}^{N} ||X_n - W_{s_n}(H_s + H_{v_n})||_F^2 + \lambda \sum ||W_{s_n} H_{v_n}||_F^2$$

The shared effect is recovered in the H_s matrix, and the exposure of the features explaining this effect are contained in the W_s matrices (Yang and Michailidis, 2016).

2. On a second iteration, i2NMF decomposes the residual effect, which was not explained by the shared decomposition



i2NMF advantages

- The feature exposure matrices W_s and W_r are different, recovering unique signatures between stage 1 and 2
- The number of inferred signatures in stage 2 can vary across matrices, allowing a better resolution of specific effects.
- All solvers were implemented on TensorFlow, allowing scalability between platforms.

The explained variance of the decomposed model can be estimated for both stages by:

$$\underset{stage1}{evar} = 1 - \frac{\sum_{i,j}((W_{s_n}H_s)_{ij} - X_{n_{ij}})^2}{\sum(W_{s_n}H_s)_{ij}^2}$$

$$\underset{stage2}{evar} = 1 - \frac{\sum_{i,j}((W_{r_n}H_{r_n})_{ij} - X_{n_{ij}})^2}{\sum(W_{r_n}H_{r_n})_{ij}^2}$$

This is useful to compare the performance between stages and the overall decomposition

1 recover cell specific signatures between different species

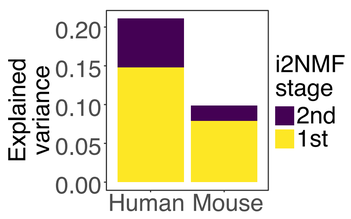
Human & Mouse substantia nigra (SN) scRNA-seq data

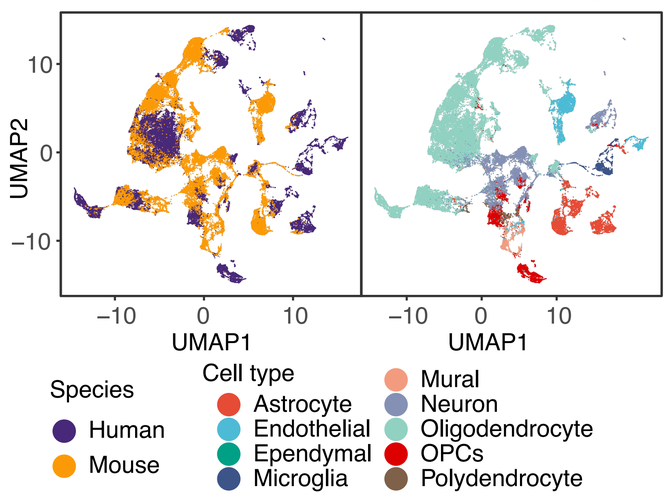
40,453 Cells — Welch et al., 2019

51,912 Cells — Saunders et al., 2018

5,127 genes in common

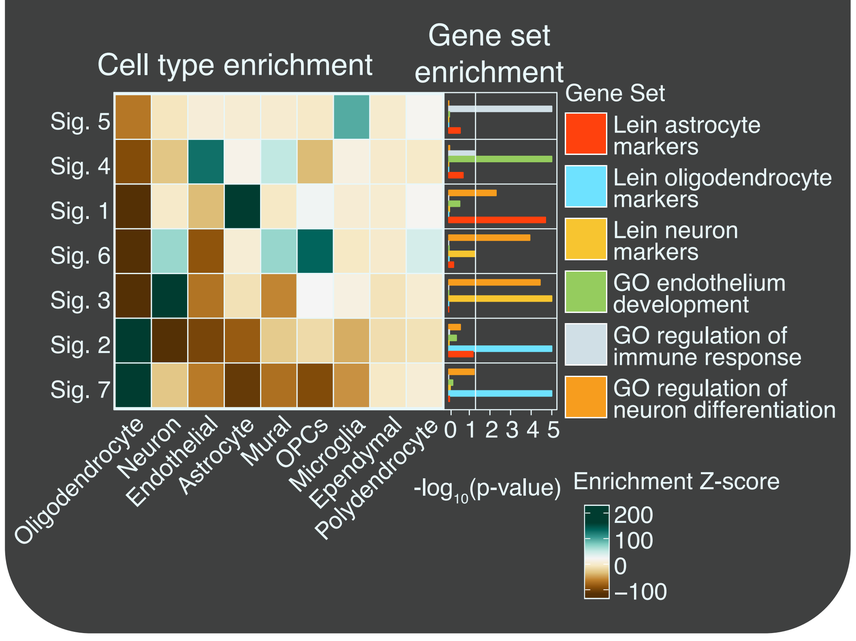
a. The human and mouse SN data sets were integrated over the set of shared genes using i2NMF.



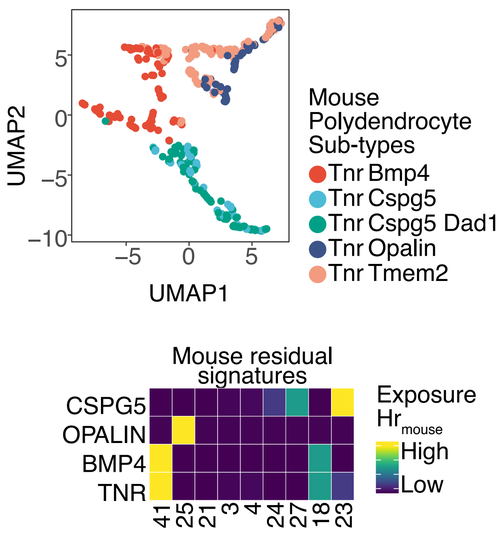


b. The shared Signatures identified in the first integrative step, were able to combine human and mouse cells (left) and resolve groups of the most relevant cell types in the SN.

c. Cell type and gene set enrichment analysis revealed that each shared signature corresponds to cell types.

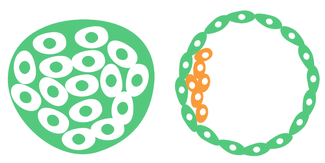


d. The second stage of i2NMF recovered species-specific Signatures, that helped to resolve cellular sub-types (top) and were defined by marker genes (bottom).



2 identify rare cell populations

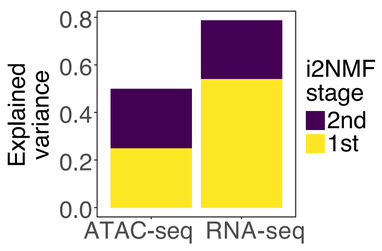
Human embryos Morula and blastocyst scCAT-seq data

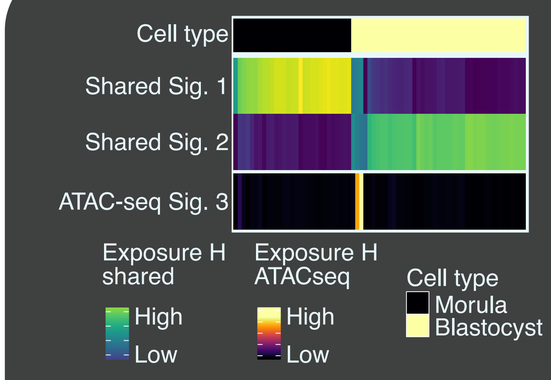


72 Cells
gene expression and chromatin accessibility for every cell
Liu et al., 2019

16,501 expressed genes (RNA-seq) & 42,713 identified peaks (ATAC-seq)

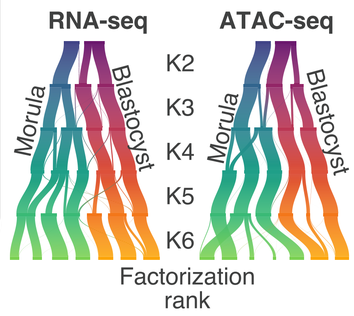
a. The human embryo scCAT-seq data set was integrated over all 72 cells. For the gene expression data, the majority of the explained variance was captured in the first stage of i2NMF, interestingly for the chromatin accessibility the second stage also recovered a a considerable fraction of the variance.





b. The shared H matrix was able to recover two cell specific signatures. On the second iteration for the ATAC-seq data, a defined signature was decomposed for two cells.

c. The decomposed shared signatures where stable across a range of factorization ranks, showing a clear separation between morula and blastocyst cells.



d. The set of chromatin accessible regions associated with the ATAC-seq Sign. 3 and its targets genes showed a specific pattern for two blastocyst cells. These also showed higher expression in marker genes for cells of the inner cell mass (ICM). Thus, allowing the identification of this rare cell type.