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 Bratwurst available in Github.

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

We applied i2NMF to:

1. recover cell specific signatures between different species

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

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

2. identify rare cell populations

Human embryos
Morula and blastocyst
scCAT-seq data

- 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 considerable fraction of the variance.

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

- The set of chromatin accessible regions associated with the ATAC-seq Sig. 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.

- 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.

i2NMF workflow:

1. Starting from two or more non-negative matrices, i2NMF initially decomposes the shared effect across them. Using i2NMF (INMF).

Solving the following problem:

\[
\min_{H,W,X} \sum_{i=1}^{N} \left| \sum_{j} W_{ij} H_{j} - X_{i} \right|_F^2 + \sum_{j} \left| H_{j} \right|_F^2
\]

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

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

\[
\begin{align*}
\text{Residual input matrix} & = X - WH \\
\text{H and W are decomposed into:} & \end{align*}
\]

i2NMF advantages:

- The feature exposure matrices, W1 and W2, 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:

\[
\text{Exp}^2_{\text{stage 1}} = \frac{1}{n} \sum_{i=1}^{n} (W_{1i} H_{i})^2 - (\bar{X}_i)^2
\]

\[
\text{Exp}^2_{\text{stage 2}} = \frac{1}{n} \sum_{i=1}^{n} (W_{2i} H_{i})^2 - (\bar{X}_i)^2
\]

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

---

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

---

[1] Liu et al., 2019

[2] Saunders et al., 2018

[3] Welch et al., 2019

[4] 4,127 genes in common

[5] Welch et al., 2019

---

**Human embryos**

**Morula and blastocyst**

**scCAT-seq data**

- 72 Cells gene expression and chromatin accessibility for every cell
- 16,501 expressed genes (RNA-seq)
- 42,713 identified peaks (ATAC-seq)

**Human & Mouse substantia nigra (SN)**

**scRNA-seq data**

- 40,453 Cells
- 51,912 Cells
- 5,127 genes in common

---

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

\[
\text{Exp}^2_{\text{stage 1}} = \frac{1}{n} \sum_{i=1}^{n} (W_{1i} H_{i})^2 - (\bar{X}_i)^2
\]

\[
\text{Exp}^2_{\text{stage 2}} = \frac{1}{n} \sum_{i=1}^{n} (W_{2i} H_{i})^2 - (\bar{X}_i)^2
\]

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

---

**Human embryos**

**Morula and blastocyst**

**scCAT-seq data**

- 72 Cells gene expression and chromatin accessibility for every cell
- 16,501 expressed genes (RNA-seq)
- 42,713 identified peaks (ATAC-seq)