St. Jude Children's Research Hospital®

CSIGEP: A scalable, GPU-based unsupervised machine learning tool for recovering gene expression programs in atlas-scale single-cell RNA-seq data

Xueying Liu[1], Richard H. Chapple[1], Paul Geeleher[1]

[1]Department of Computational Biology,
St. Jude Children's Research Hospital, Memphis, TN, USA

BACKGROUND

Single-cell RNA-seq can now be cheaply collected in millions of cells and are interrogated using tools such as UMAP, performing hard clustering over two dimensional projections of the data. However, conventional methods have been heavily criticized, with data distorted by arbitrary choices of parameters, and fidelity to ground truth easily lost.

1.

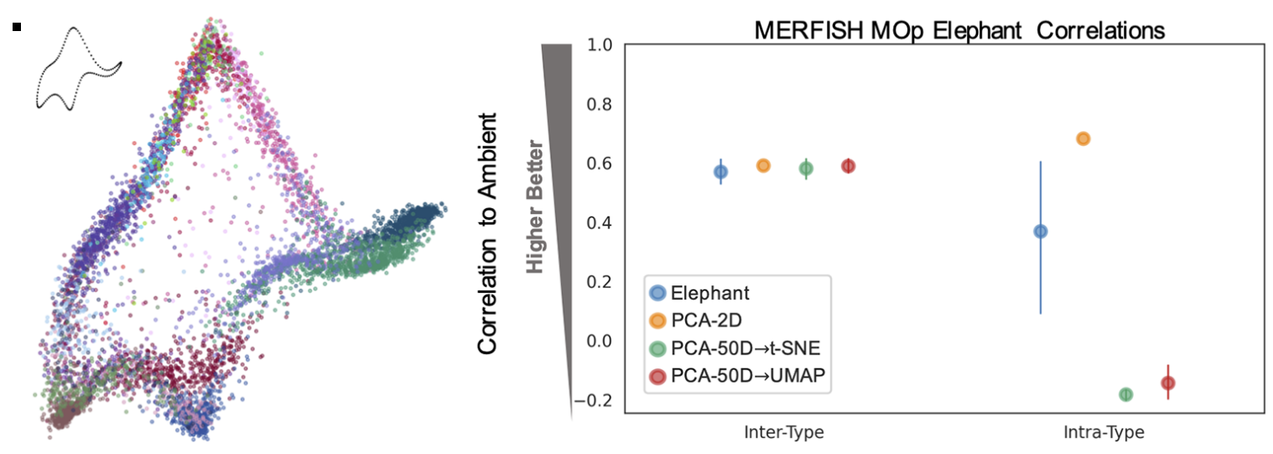


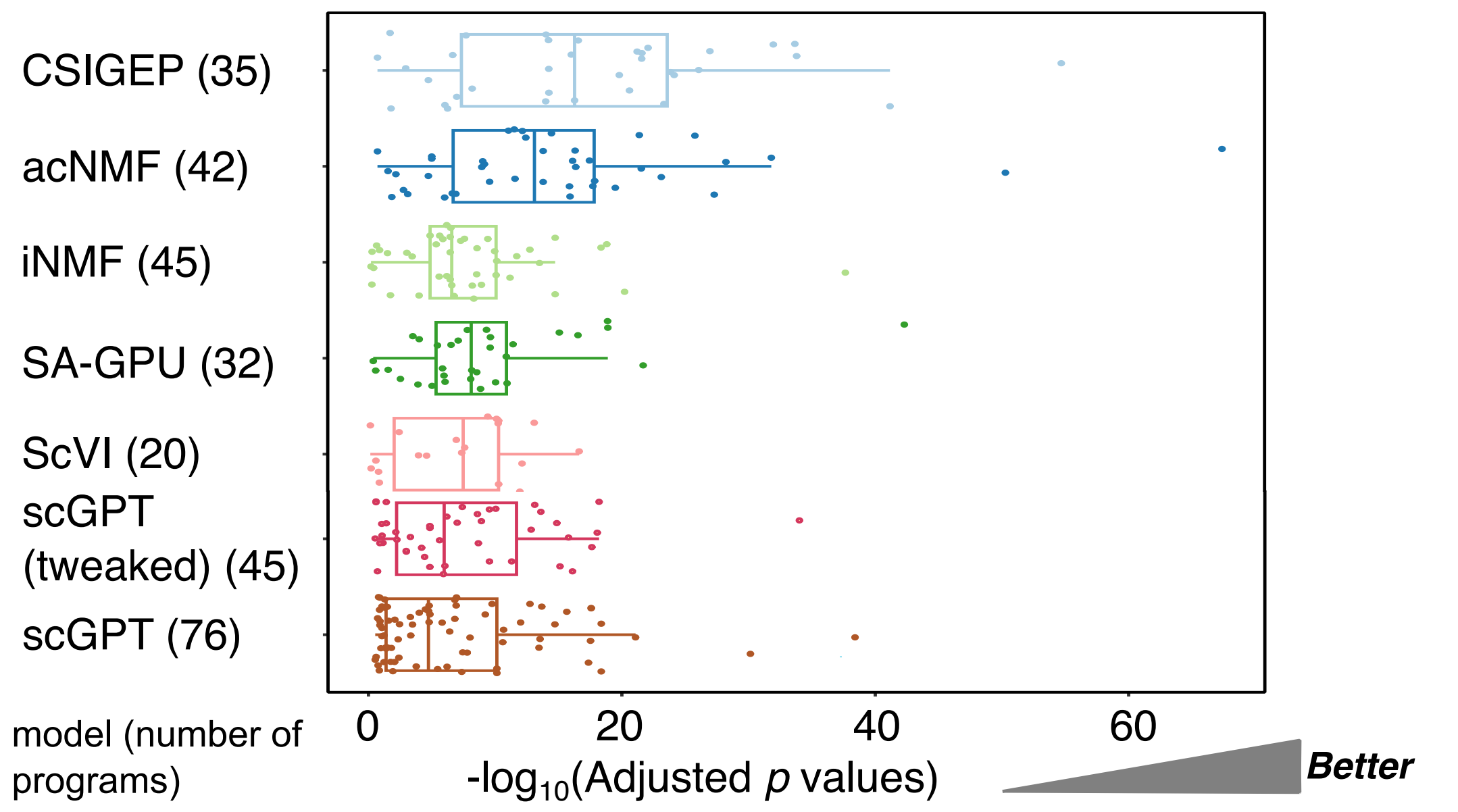
Fig.1: Tara Chari and Lior Pachter[1] have demonstrated the indeterminate nature of 2D UMAP and t-SNE embeddings by fitting cells and embedding data in the shape of an "von Neumann elephant".

Consensus-NMF-based models, which allow scRNA-seq data to be represented as “gene expression programs” (GEPs), perform vastly better in preserving the underlying structure of single-cell RNA-seq, but currently, these methods are not scalable. Here, we present a computationally efficient GPU-based consensus-NMF method. We implemented the method using a machine learning technique called “mini-batching”, where the model is trained on iterative subsets of the data, and is thus scalable to a dataset of any size.

RESULTS

Comparison on proposed model to other candidate models on GOSH, a scRNA-seq dataset of 5 neuroblastoma samples. For each model, we ran cell type inference against Descartes cell and tissue database, and assigned the lowest P-value to the program. Fig. 6 displays resulting p values for each model, with number of programs each of them has recovered. cNMF-based models outperformed other models in general.

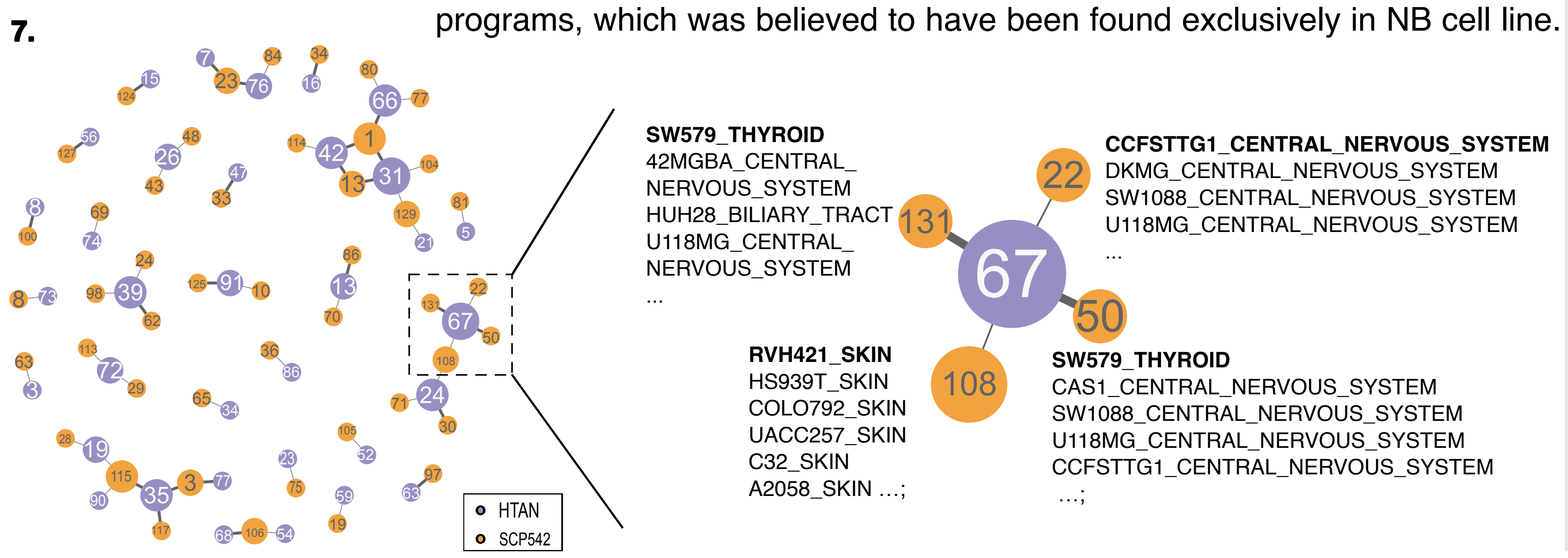
6.



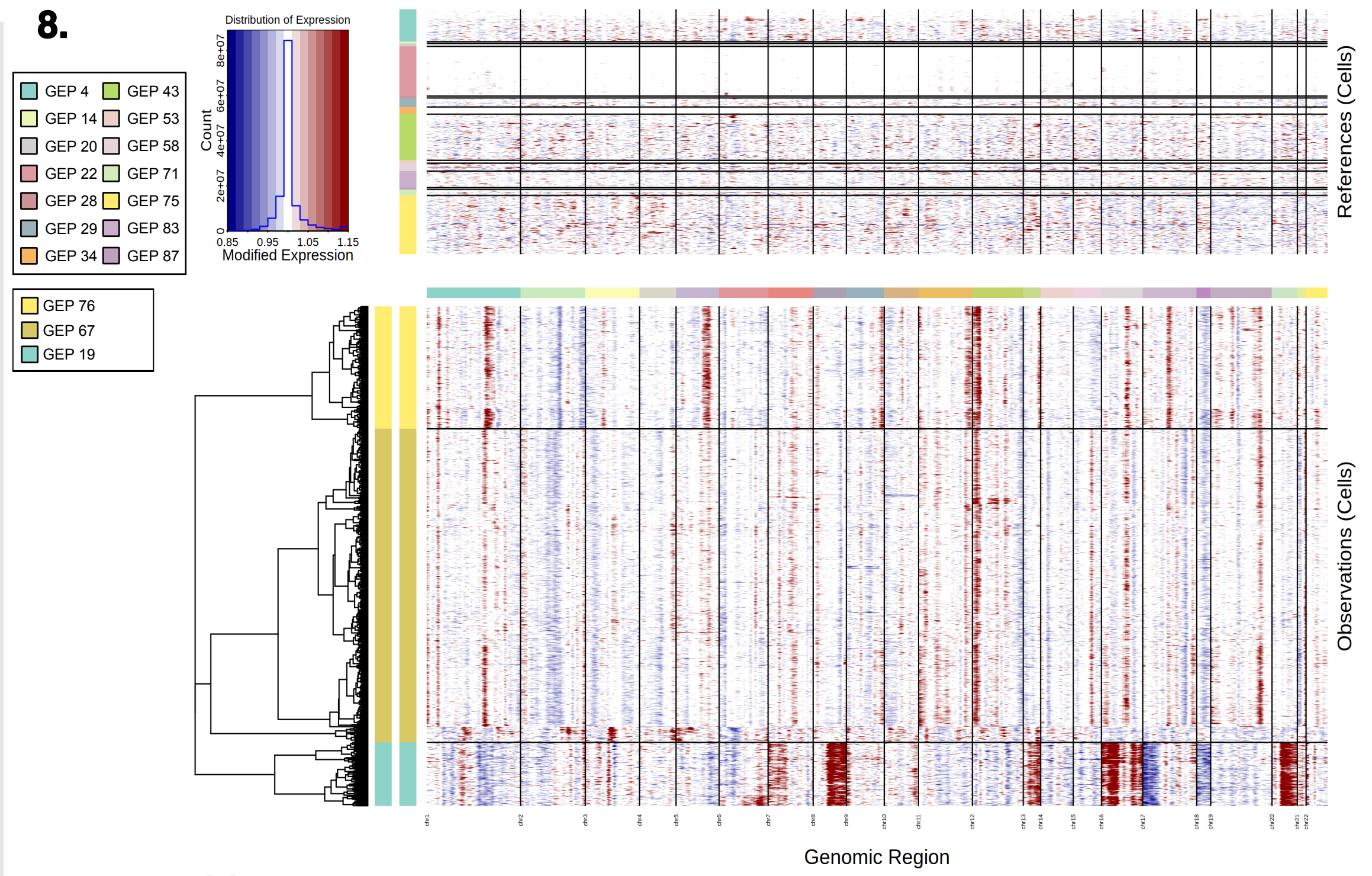
- 115 samples from Human Tumor Atlas Network (HTAN) WUSTL atlas (https://data.humantumoratlas.org), totaling 600,000 cells, from 10 cancer types, recovered 91 programs.
- 198 cancer cell lines of 22 cancer types from DepMap, consistsing of 53.5k cells, identified 135 programs in total.

Fig. 7 illustrates the connectivities of programs between the 2 datasets. ***Interestingly***, program 67 (normal fibroblasts in HTAN) is connected to four cancer cell line programs.
Remarkably, this mirrors the previuosly described expression of chemo-resistent "mesenchymal-like" programs, which was believed to have been found exclusively in NB cell line.

7.


67
131
22
50
108
SW579_THYROID
42MGBA_CENTRAL_NERVOUS_SYSTEM
HUH28_BILIARY_TRACT
U118MG_CENTRAL_NERVOUS_SYSTEM
...
CCFSTTG1_CENTRAL_NERVOUS_SYSTEM
DKMG_CENTRAL_NERVOUS_SYSTEM
SW1088_CENTRAL_NERVOUS_SYSTEM
U118MG_CENTRAL_NERVOUS_SYSTEM
...
RVH421_SKIN
HS939T_SKIN
COLO792_SKIN
UACC257_SKIN
C32_SKIN
A2058_SKIN ...;
SW579_THYROID
CAS1_CENTRAL_NERVOUS_SYSTEM
SW1088_CENTRAL_NERVOUS_SYSTEM
U118MG_CENTRAL_NERVOUS_SYSTEM
CCFSTTG1_CENTRAL_NERVOUS_SYSTEM
...;
HTAN
SCP542

8.

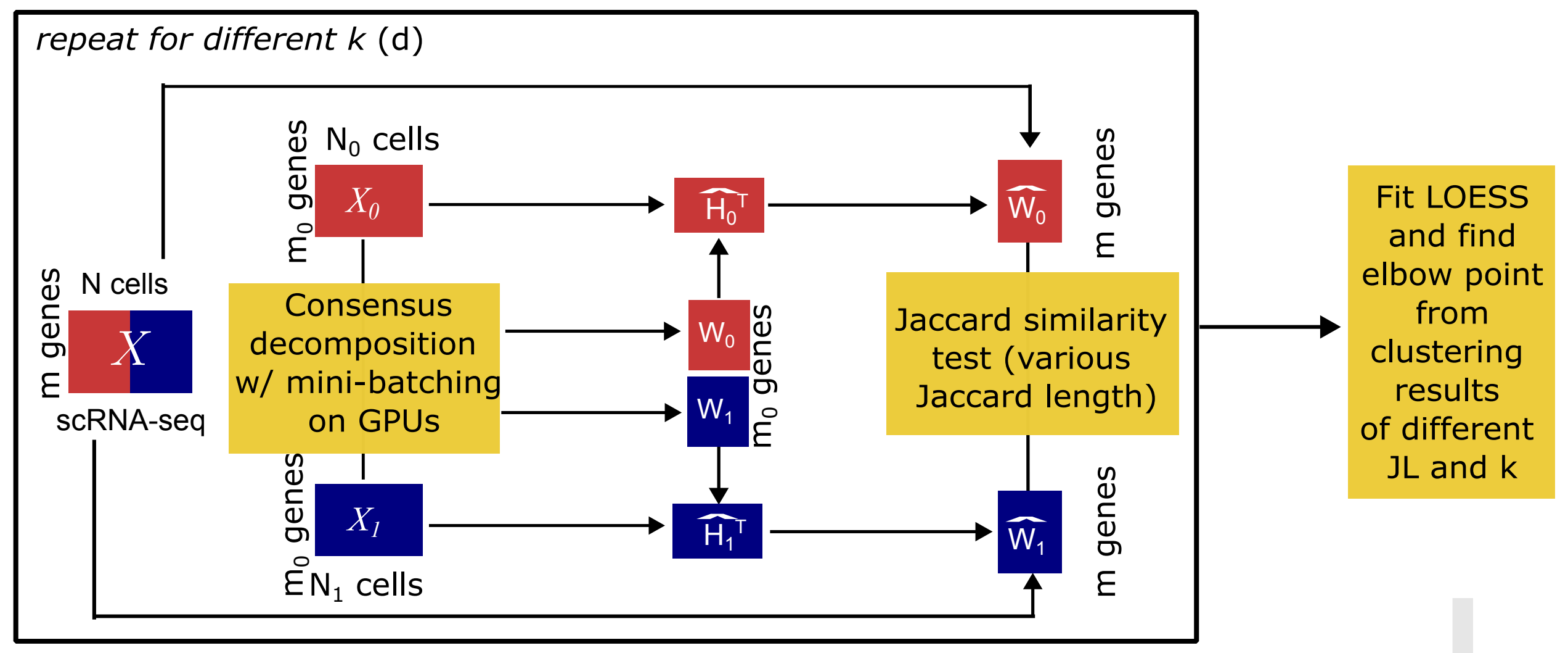


InferCNV[6] ran on cells with high h-values (>100) in GEP67 against cells expressing immune cell signatures in the same samples.
Fig. 8 confirms that these cells of interest retain similar pattern in chromosomal copy number variations as "immune cells", i.e., non-cancer cells.

METHODS

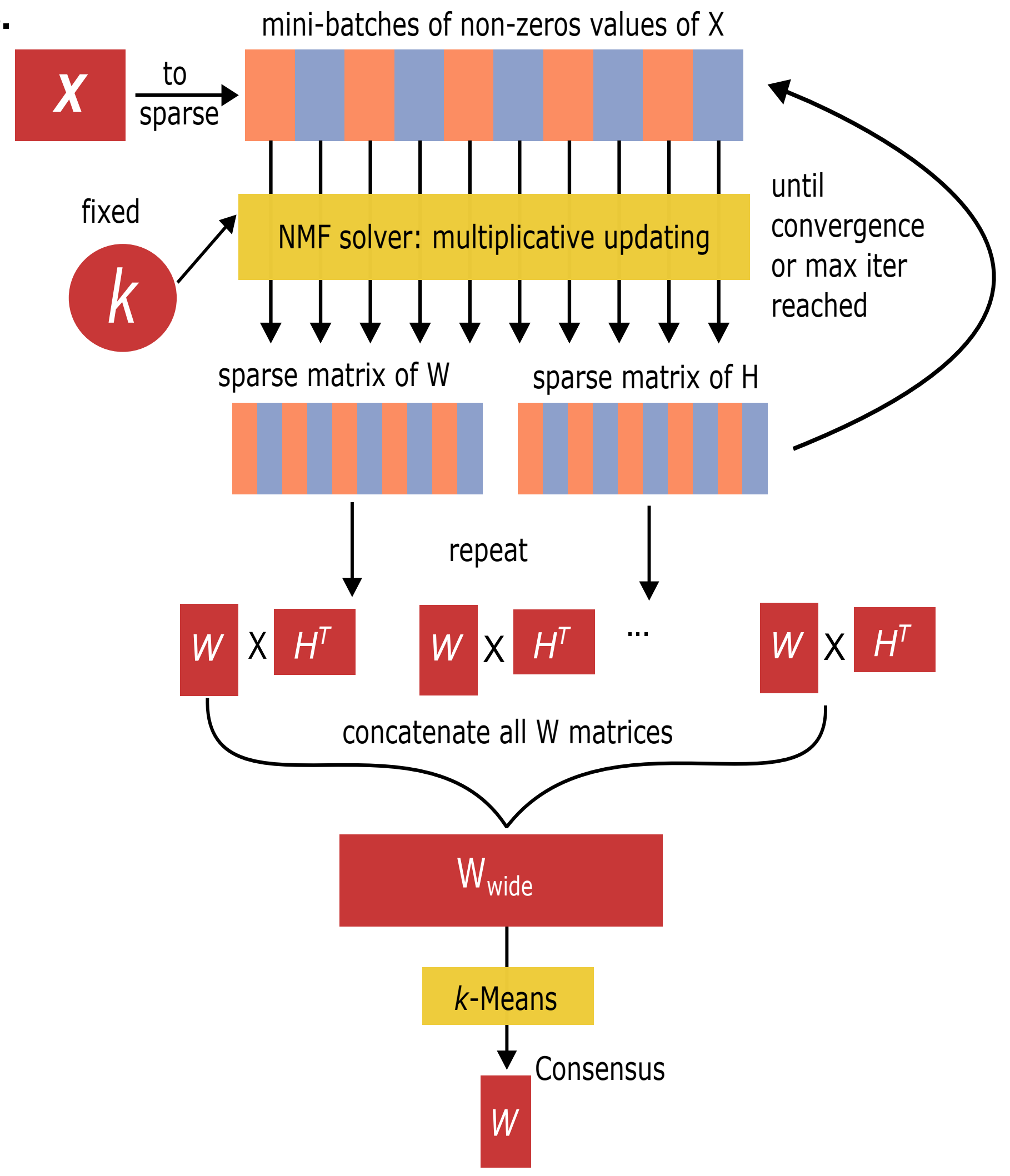
- We have developed a new consensus-NMF-based model **CSIGEP** (***C****onsensus and* ***S****table* ***I****nterpretation of* ***G****ene* ***E****xpression* ***P****rograms*), which can run on specialized hardware (Graphical Processing Units (GPUs)), based on our previous work where we developed a novel NMF-based implementation that outputs precise latent representation of scRNA-seq data.
- Two other models were benchmarked against, both of which have shown success in analyzing large datasets and providing fast and precise latent representations of biological data.
(a). **SignatureAnalyzer-GPU (SA-GPU)[4]** utilizes automatic relevance determination, a Bayesian NMF to determine number of latent representation;
(b). **Single cell Variational Inference (ScVI)[5]** is a deep neural network-based model that learns expression data and minimizes the difference between input and output.
- **scGPT[7],** a recently published foundation model, was also included for comparison on a real dataset (see **Results**, **Fig. 6**).

2.


repeat for different k (d)
N cells
m genes
X
scRNA-seq
N_0 cells
m_0 genes
X_0
Consensus decomposition w/ mini-batching on GPUs
X_1
m_0 genes
N_1 cells
$\hat{H}_0^T$
W_0
W_1
m_0 genes
$\hat{H}_1^T$
$\hat{W}_0$
m genes
Jaccard similarity test (various Jaccard length)
$\hat{W}_1$
m genes
Fit LOESS and find elbow point from clustering results of different JL and k

- **Description of the model (Fig. 2)**
(a). Randomly split a scRNA-seq dataset into two halves: X_0 and X_1.
(b). Perform cNMF (see **Core components, Fig. 3**) on the two splits, given desired number of component k.
(c). Align the k columns of W_0 and W_1 using Jaccard similarity test (with different Jaccard length JL). Pair of columns associated with a significant test statistic is selected and forms a community.
(d). Repeat (b)-(c) with different values of k, and keep track of the connectivity given different k and JL.
(e). Fit a local regression (LOWESS) on the number of clusters against k for each JL and find the corresponding k at the elbow point.

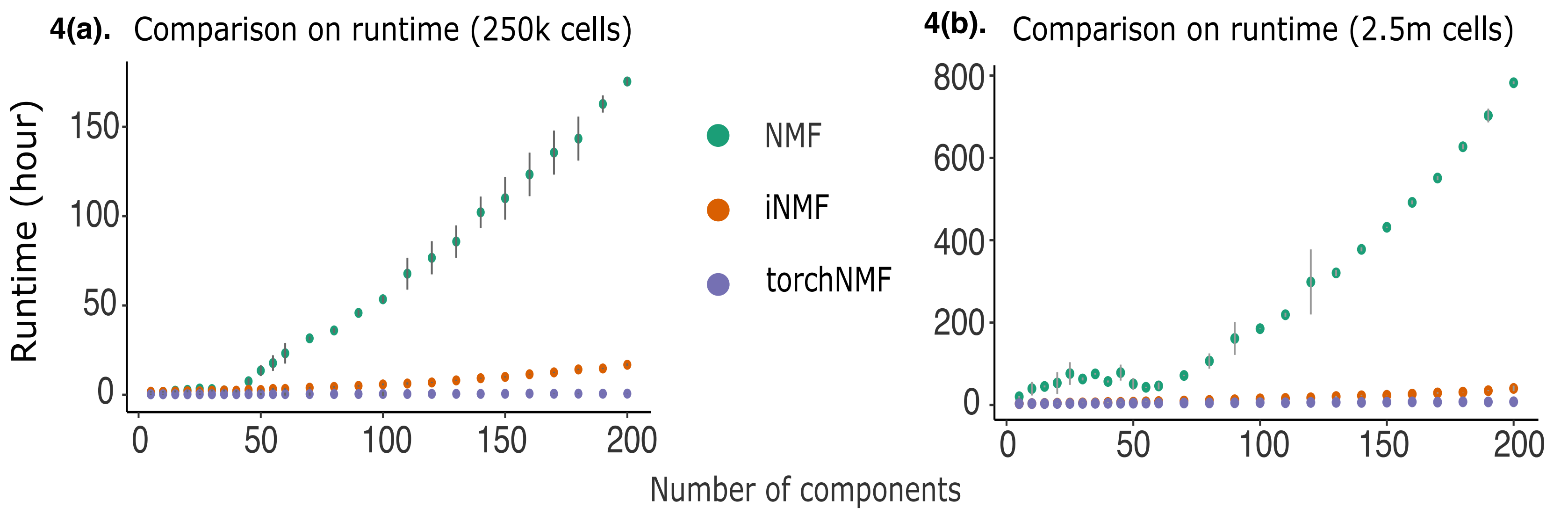
3.


X
to sparse
mini-batches of non-zeros values of X
fixed
k
NMF solver: multiplicative updating
until convergence or max iter reached
sparse matrix of W
sparse matrix of H
repeat
W X H^T
W X H^T
...
W X H^T
concatenate all W matrices
W_{wide}
k-Means
Consensus
W

- **Core component (Fig. 3)**

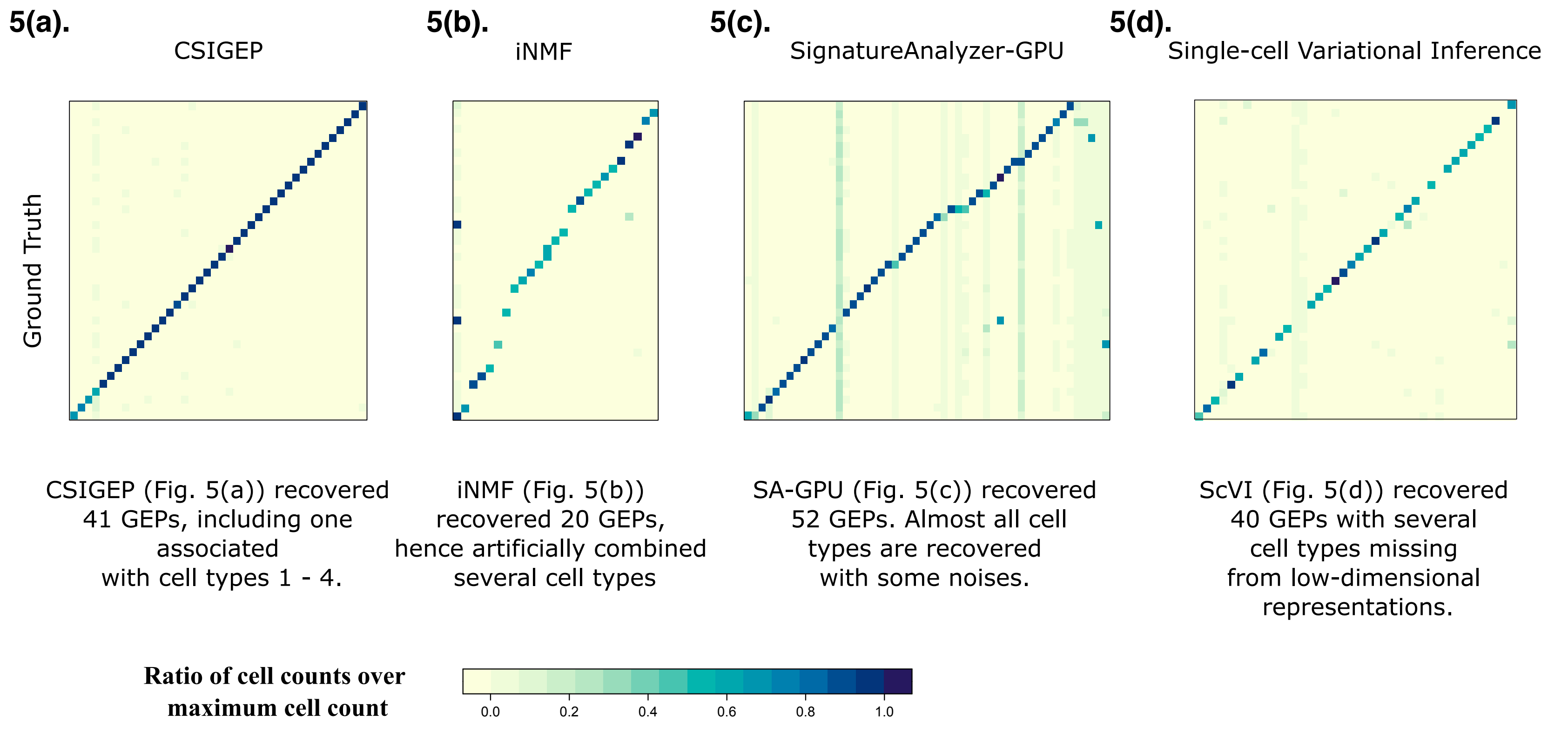
(a). Prepare data matrices by
(1). compute TPM values and (2). pick highly variable genes
(b). Reorder scaled data matrix as a sparse matrix with only non-zero values
(c). Divide into batches
(d). Use multiplicative update rule to compute W and H accumulative values on each batch
(e). Reorder sparse matrices of W and H back to dense matrices with desired shape
(f). Repeat (c)-(e) n_{iter} times
(g). Concatenate n_{iter} W matrices into a wide matrix of k X n_{iter} columns and perform k-Means clustering
(h). Assign median values of each cluster to consensus W
(i). Perform NMF on scaled data to update $\hat{H}^T$
(j). Use inversion of matrix manipulation to get $\hat{W}$

RESULTS



- Simulated data of single-cell RNA-seq was generated by an engineered Splatter[2]-based model where cells of selected cell types are sampled to share signatures of an activity program.
- Runtime was compared on one simulated dataset with 250k cells, and one with 2.5m cells, respectively.
- **Fig. 4** shows average runtime of 5 repititions with increasing number of components *(k)*. For each experiment, standard deviation was shown with gray bar.

- We assessed ability of ground truth recovery using the simulated data of 1M cells on 4 candidate models: CSIGEP, iNMF[3], SA-GPU, and ScVI.
- Each cell was assigned to one program based on its largest value in H matrix.
- We converted cell counts to a ratio over the maximum, where for each cell type, cells with an above-average value in H matrix was counted.
- **Fig. 5** compared the performance across models, where ground truth cell types are ordered from 1 (bottom) to 40, and recovered GEP IDs are rearranged for visualization purpose.



CSIGEP (Fig. 5(a)) recovered 41 GEPs, including one associated with cell types 1 - 4.

iNMF (Fig. 5(b)) recovered 20 GEPs, hence artificially combined several cell types

SA-GPU (Fig. 5(c)) recovered 52 GEPs. Almost all cell types are recovered with some noises.

ScVI (Fig. 5(d)) recovered 40 GEPs with several cell types missing from low-dimensional representations.

CONCLUSIONS

- Lack of robust and efficient computational tools to interpret data at atlas scale has hindered exploration and interpretation of pan-tumor single-cell RNA-sequencing data.
- We have solved by developing a computationally-efficient implementation of NMF and a set of front-end tools to interpret these results. The new model outperforms state-of-the-art models in terms of both fidelity to ground truth and compute time.
- Because of the scalability of the tools, we have been able to show that by factorizing multiple datasets jointly, we are better powered to discover meaningful patterns of co-expression in the data.
- We have developed a web-based tool (pscb.stjude.org) that allows for interpretation of resulting models by non-technical researchers. We have published analytical results of 4 public human neuroblastoma scRNA-seq datasets, data we generated from cell lines, a generic mouse model and patient derived xenografts. We will continue development and expression of the webbased tool by adding more features and improving user experiences.

REFERENCES

[1] Chari T, Pachter L.
The specious art of single-cell genomics.
PLoS Comput Biol. 2023 Aug 17;19(8):e1011288.
doi: 10.1371/journal.pcbi.1011288. PMID: 37590228; PMCID: PMC10434946.
[2] Zappia L, Phipson B, Oshlack A (2017).
Splatter: simulation of single-cell RNA sequencing data.
Genome Biology. doi:10.1186/s13059-017-1305-0,
https://doi.org/10.1186/s13059-017-1305-0.
[3] Gao, C., Liu, J., Kriebel, A.R. et al.
Iterative single-cell multi-omic integration using online learning.
Nat Biotechnol 39, 1000–1007 (2021).
https://doi.org/10.1038/s41587-021-00867-x
[4] Taylor-Weiner, A., Aguet, F., Haradhvala, N.J. et al.
Scaling computational genomics to millions of individuals with GPUs.
Genome Biol 20, 228 (2019). https://doi.org/10.1186/s13059-019-1836-7
[5] Gayoso, A., Lopez, R., Xing, G. et al.
A Python library for probabilistic analysis of single-cell omics data.
Nat Biotechnol 40, 163–166 (2022).
https://doi.org/10.1038/s41587-021-01206-w
[6] inferCNV of the Trinity CTAT Project.
https://github.com/broadinstitute/inferCNV
[7] Cui, H., Wang, C., Maan, H. et al.
scGPT: toward building a foundation model for single-cell multi-omics using generative AI.
Nat Methods (2024). https://doi.org/10.1038/s41592-024-02201-0