A clustering algorithm to identify latent cell types from DNA methylation patterns

Alex H. Williams[1,*], Eran A. Mukamel[1,2]

UC San Diego, Neurosciences[1] and Cognitive Science[2], La Jolla, CA
** Current Affiliation, Stanford University, Neuroscience*

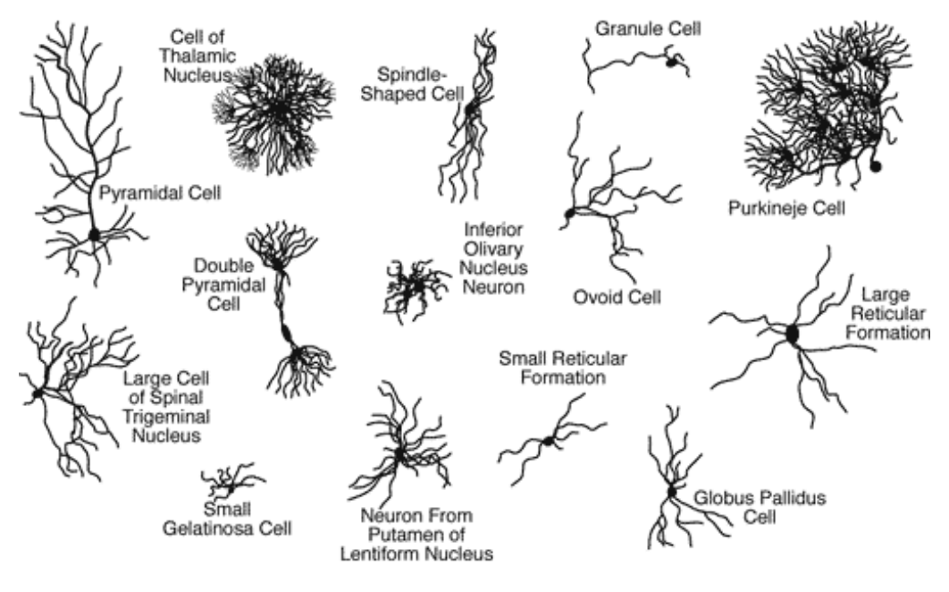
Abstract

The brain is composed of cells with remarkably diverse functions and morphologies. Defining and characterizing these cell types is a long-standing goal in neuroscience. DNA methylation, an epigenetic modification involving the addition of methyl groups to genomic cytosine nucleotides, is a potentially useful marker of cell identity and function. In particular, methylation patterns are thought to be stable in adult animals and can be mapped using whole-genome bisulfite sequencing (WGBS). Furthermore, the addition and removal of DNA methylation may be involved in cell and tissue differentiation; methylation is targeted to different genes in different cells during development and typically silences those genes, resulting in cell type-specific gene expression. Methylation patterns vary among major classes of cells in the brain, and across cell samples from different tissues and organs.

While it is sometimes possible to separate cells using genetic or molecular markers, many cell types cannot be isolated in large enough quantities for WGBS. As a result, these experiments are typically performed on large tissue samples. Methylation patterns are measured across many short DNA sequencing reads, each originating from a cell of unknown type. We developed an unsupervised learning algorithm to cluster these reads, and thus infer cell type-specific methylation profiles from tissue samples with mixed cell types. We represent the methylation pattern across bisulfite sequencing reads as an incomplete binary matrix and obtain a low-rank matrix factorization with alternate convex programming. By appropriately constraining and regularizing the matrix factorization, we can efficiently obtain a soft clustering of the bisulfite sequencing reads into cell types. These results can be used as a heuristic to initialize a hard clustering algorithm, such as k-means clustering, if desired.

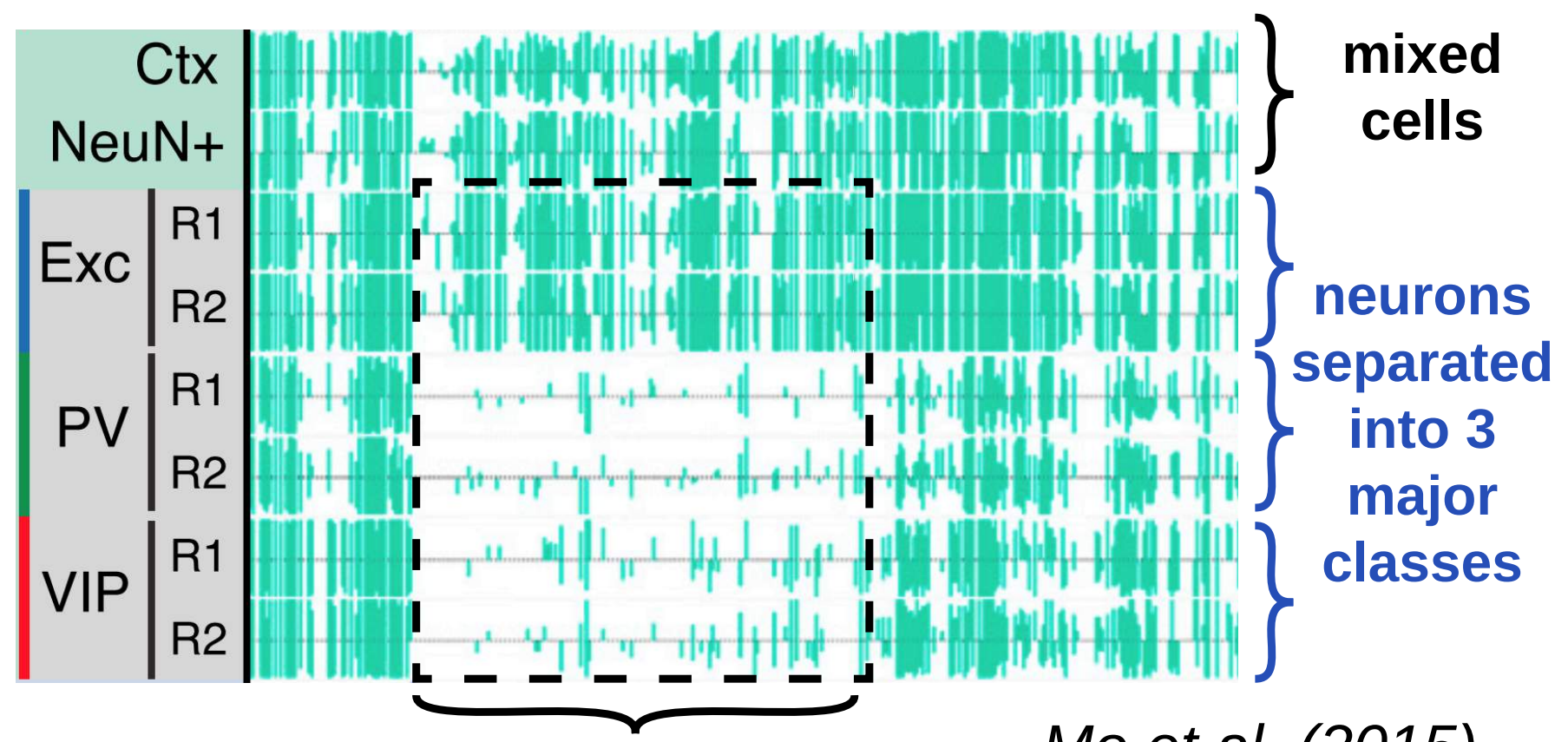
Classifying brain cells with functional, structural, and genetic markers is fundamental to understanding neural circuit function

DNA methylation is a stable, reliably measurable epigenetic marker that is implicated in cell differentiation. Methylation patterns are known to differ among a few high-level neuron types, which can be separated by genetic markers (*lower right figure*, Mo et al., 2015). Because separating brain cell types for epigenomic profiling remains experimentally challenging, we sought to develop techniques for inferring latent methylation profiles in samples containing mixed cell types.



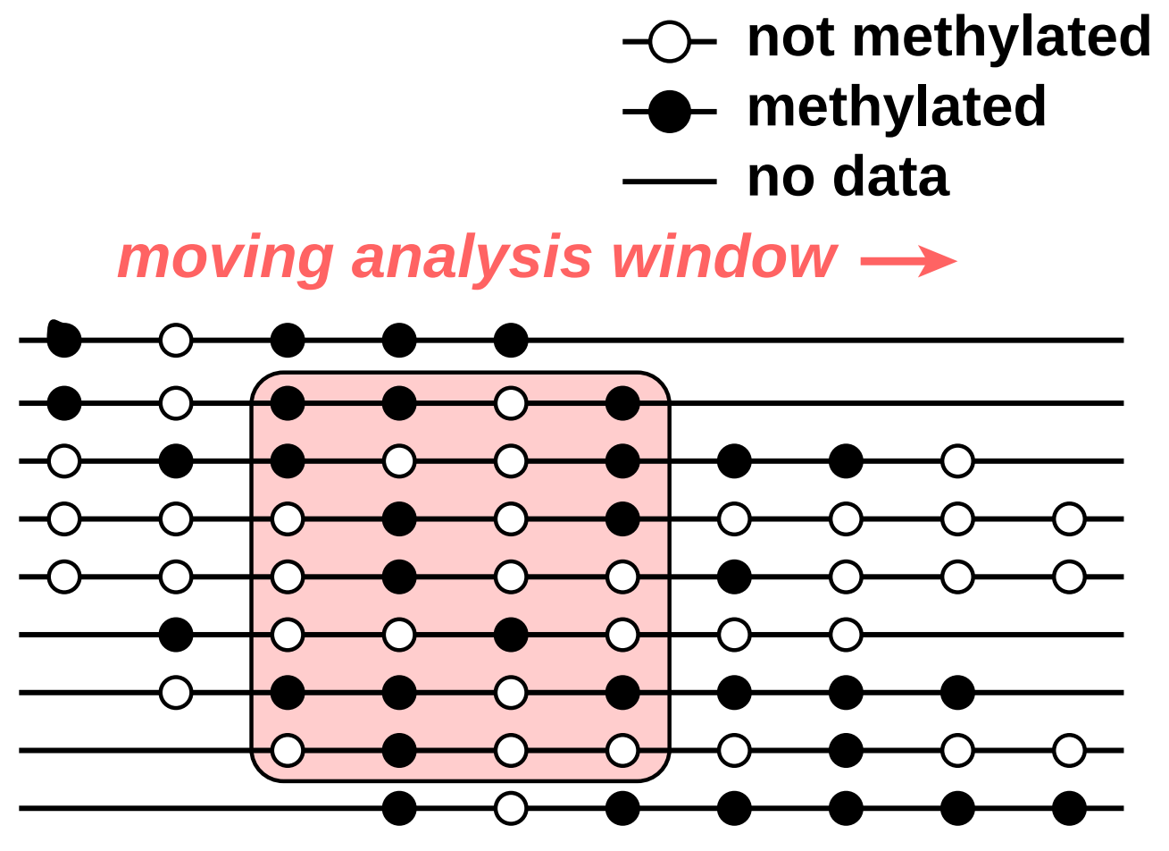
Cell of Thalamic Nucleus
Spindle-Shaped Cell
Granule Cell
Pyramidal Cell
Purkinje Cell
Double Pyramidal Cell
Inferior Olivary Nucleus Neuron
Ovoid Cell
Large Reticular Formation
Small Reticular Formation
Large Cell of Spinal Trigeminal Nucleus
Small Gelatinosa Cell
Neuron From Putamen of Lentiform Nucleus
Globus Pallidus Cell

http://www.mind.ilstu.edu/



Mo et al. (2015) Neuron. 86:1369.

Several techniques exist for detecting differentially methylated regions between samples (e.g. Dolzhenko & Smith, 2014). Fewer studies have developed methods to cluster reads within individual samples.



The diagram to the right illustrates the data provided by bisulfite sequencing. Each read provides a short data sequence on the methylation status of cytosine nucleotides. However, these sequences are not aligned, making it difficult to directly compare and cluster reads. Previous studies (Wu et al., 2015; Lin et al., 2015) have simplified the problem by only considering reads that completely cover a small (~4 sites) window (see red box). Any clustering technique (e.g. K-means) can be applied within this window, and the analysis can be repeated across the genome by shifting the window. However, this approach excludes potentially relevant data points, such as the top and bottom read, which cover all but one cytosine. Additionally, it may be difficult to define a principled procedure to consolidate multiple local cluster assignments into a cohesive global picture.

Goal: Assign each read to a cell type

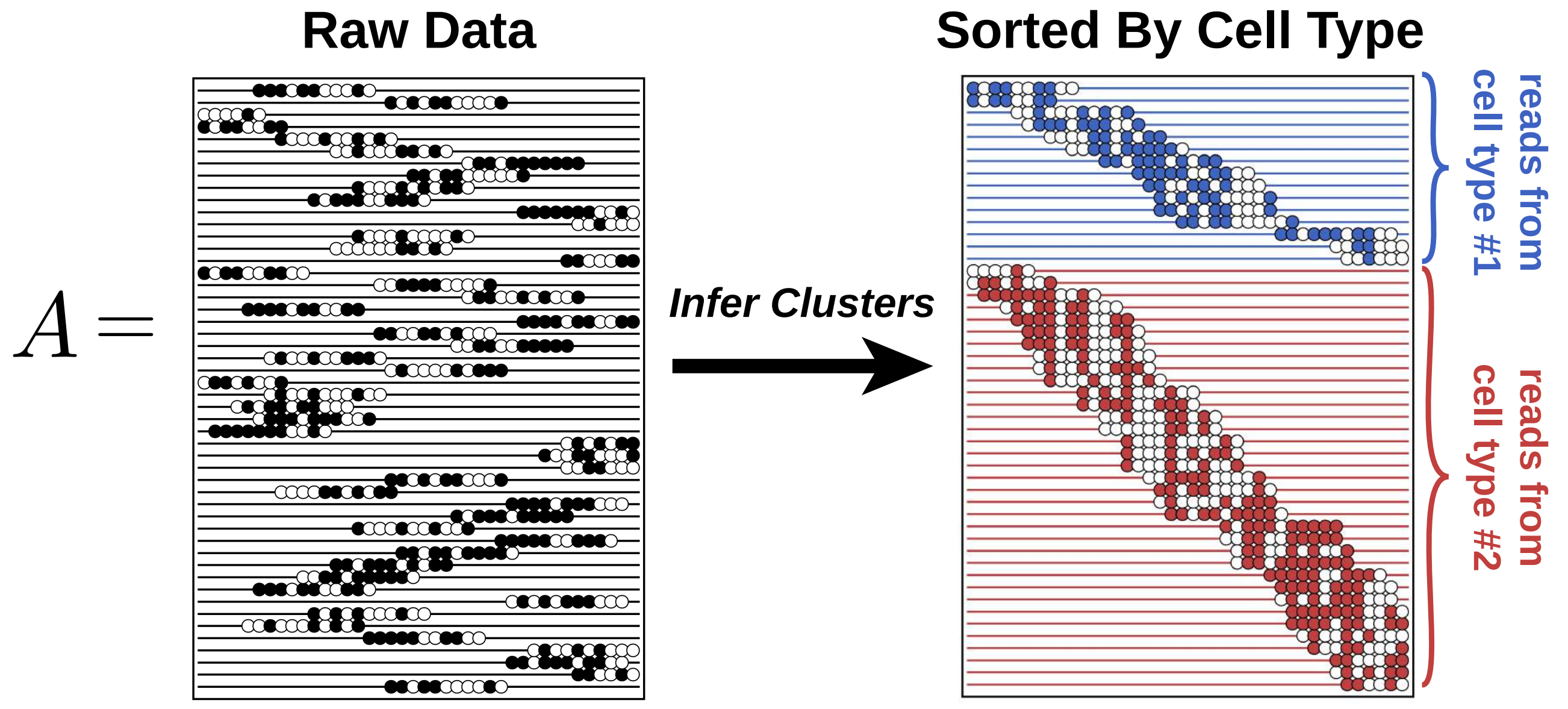
Mathematical Formulation: Determine a low-rank approximation of a binary matrix with many missing entries.

Data Description:

$$A_{ij} = \begin{cases} 1 \ , & \text{methylated} \\ 0 \ , & \text{not methylated} \\ \text{NA} \ , & \text{not observed} \end{cases}$$

Each row of A corresponds to a bisulfite sequencing read.

Each column of A corresponds to a cytosine position.



Approach: Regularized matrix factorization problem

$$A \approx XY \qquad \begin{matrix} X \in \mathbf{R}^{m\times k} \\ Y \in \mathbf{R}^{k\times n} \\ A \in \mathbf{R}^{m\times n} \end{matrix}$$

Optimization problem for a specified loss function (L), and regularization function (r_x) applied to each row of X (x_i).

$$\underset{X,Y}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} L(A_{ij}, x_i y_j) - \gamma_x \sum_i r_x(x_i)$$

Hard Clustering
K-means, NP-hard

$$r_x(x) = \begin{cases} \infty \ , & \mathbf{Card}(x) \neq 1 \\ \infty \ , & ||x||_1 \neq 1 \\ 0 \ , & \text{otherwise} \end{cases}$$

In words: Restrict each row of X to have one nonzero element equal to 1

Soft Clustering
Sparse NMF, biconvex

$$r_x(x) = \begin{cases} \infty \ , & x \not\geq 0 \\ ||x||_1 \ , & \text{otherwise} \end{cases}$$

In words: Restrict each row of X to be nonnegative and penalize nonzeros

Interpretation:

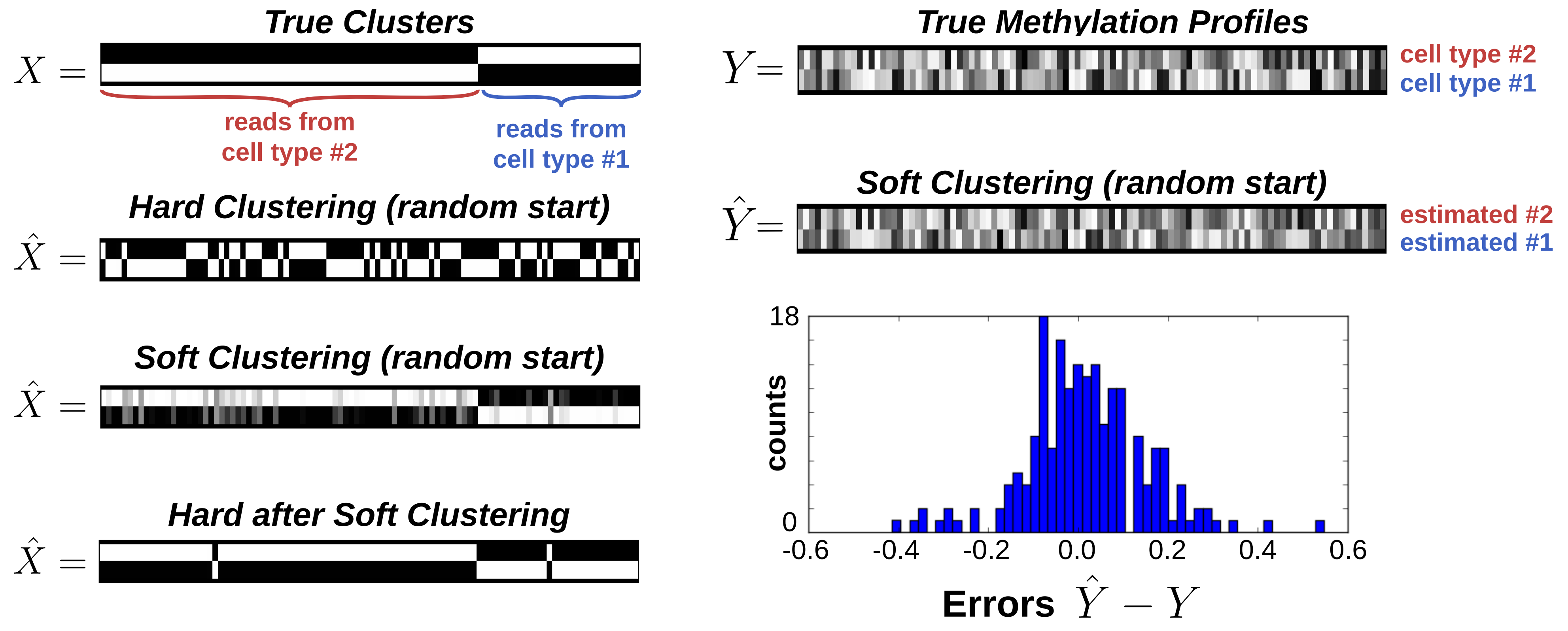
We restrict X and Y to be nonegative and we encourage each row of X to be sparse (few nonzero elements). Thus:

- Each row of Y specifies the methylation profile of a cell-type.
- Each row of X specifies a (possibly approximate) clustering assignment

For example, if $k = 3$ then $x_i = [0 \ \ 1 \ \ 0]$ would assign the i^{th} read to the 2nd cluster, since the product $[0 \ \ 1 \ \ 0]Y$ is the 2nd row of Y.

Typical results on synthetic data:

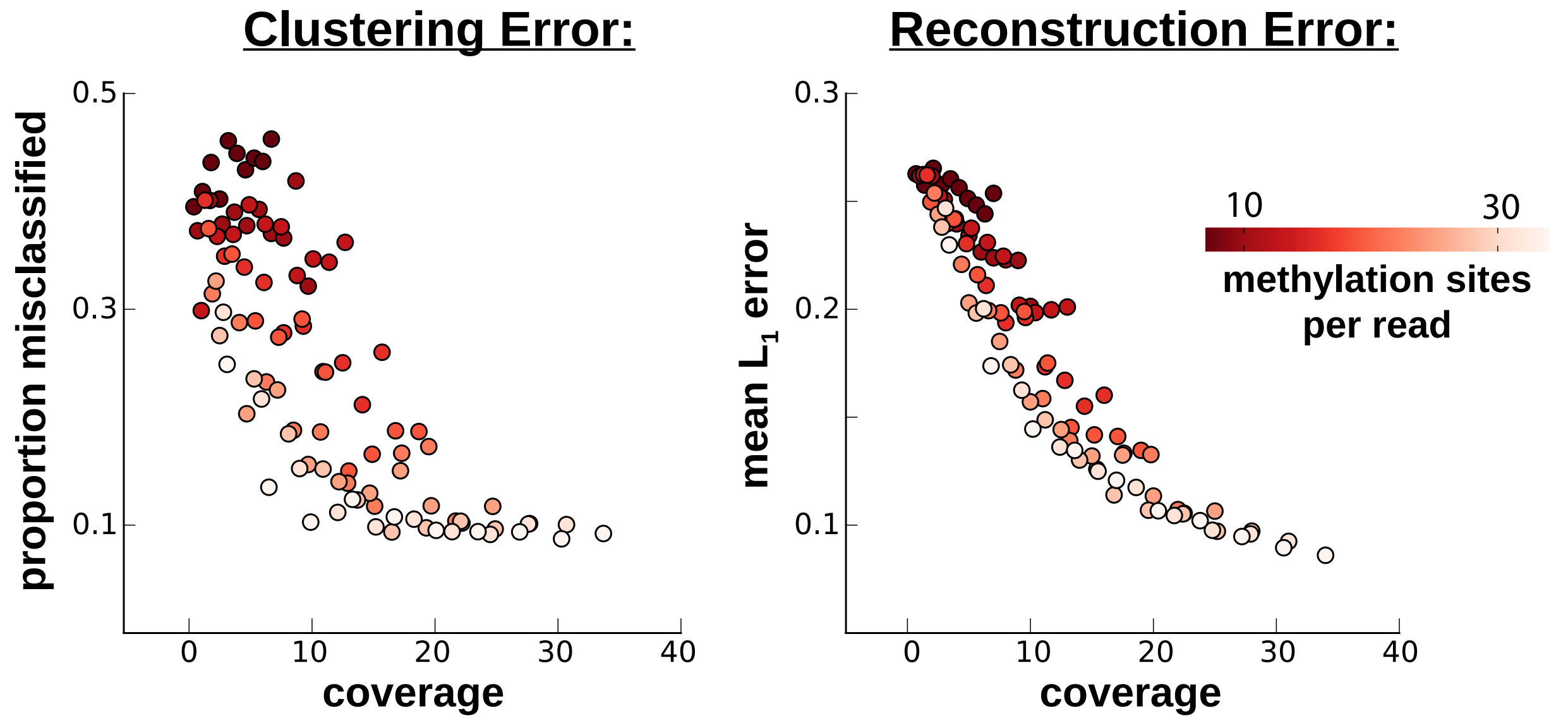
Hard clustering typically fails to recover true cell types. The soft clustering variant is easier to optimize (less likely to be trapped in local minima), and provides an effective initialization for hard clustering if desired.



Implementation details: The estimates $\hat{X}$ and $\hat{Y}$ were obtained by alternating proximal gradient descent as described in Udell et al. (2015). My code is available at: https://github.com/ahwillia/MethylClust

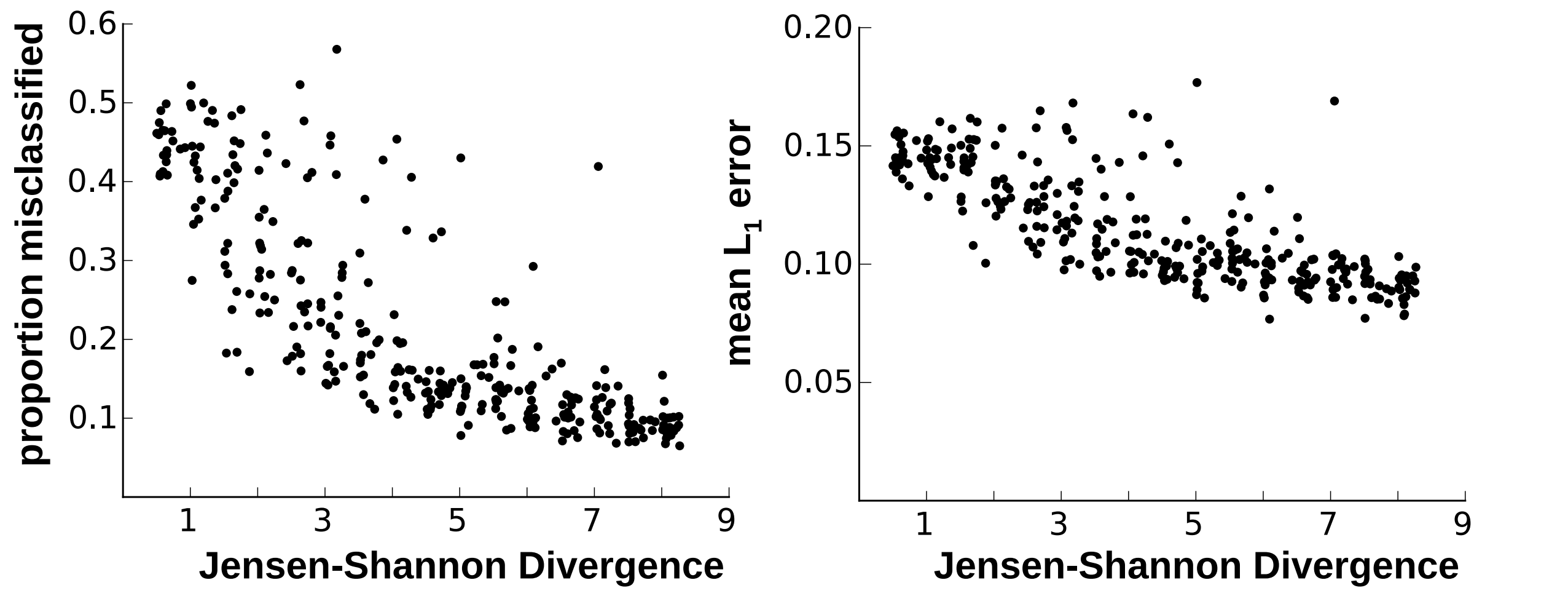
Accurate read sorting/clustering requires moderately deep sequencing

Two important parameters for performance are the average coverage (observed elements in each column of A), and the number of methylation sites covered by each read (observed elements in each row of A).



How distinct do the cell types have to be?

The plots below characterize performance as a function of the divergence between two methylation profiles (rows of Y) varies. For these simulations, A had dimensions 100x100, with ~30 observed elements in each row.



Conclusions

- A new clustering algorithm based on matrix factorization can accomodate larger datasets with partially overlapping reads.
- Local clustering methods utilize fewer observations and exclude potentially relevant data. They also require complex secondary analyses to merge local cluster assignments into a global picture.
- The algorithm performs well on synthetic data with similar characteristics of current sequencing technology.
- It is easy to implement other multivariate analyses (PCA, NMF, etc.) by appropriately modifying the regularization function r_x.

References and Related Work:

Dolzhenko E, Smith AD (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactorial whole-genome bisulfite sequencing experiment. *BMC Bioinform.* 15:215

Lin P, Forêt S, Wilson SR, Burden CJ (2015). Estimation of the methylation pattern distribution from deep sequencing data. *BMC Bioinform.* 16:145

Udell M, Horn C, Zadeh R, Boyd S (2014). Generalized Low Rank Models. *arXiv preprint.* arXiv:1410.0342

Wu X, Sun M, Zhu H, Xie H (2015). Nonparametric Bayesian clustering to detect bipolar methylated genomic loci. *BMC Bioinform.* 16:11

Acknowledgments/Support:

We thank Madeleine Udell for helpful discussions and for developing open-source optimization software packages, which we utilized heavily (see above). This work was supported by Dept. of Energy Computational Science Graduate Fellowship Program (A.H.W.) and NIH/NINDS R00NS080911 (E.A.M.)