Bioinformatics, 35(6), 2019, 953–961
doi: 10.1093/bioinformatics/bty702
Advance Access Publication Date: 28 August 2018
Original Paper

Gene expression

Parallel clustering of single cell transcriptomic data with split-merge sampling on Dirichlet process mixtures

Tiehang Duan[1,*], José P. Pinto[2] and Xiaohui Xie[1,*]

[1]Department of Computer Science, University of California, Irvine, CA 92617, USA and [2]SysBioLab, Centre for Biomedical Research (CBMR), University of Algarve, Faro, Algarve 8005-139, Portugal

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on February 27, 2018; revised on July 20, 2018; editorial decision on August 13, 2018; accepted on August 22, 2018

Abstract

Motivation: With the development of droplet based systems, massive single cell transcriptome data has become available, which enables analysis of cellular and molecular processes at single cell resolution and is instrumental to understanding many biological processes. While state-of-the-art clustering methods have been applied to the data, they face challenges in the following aspects: (i) the clustering quality still needs to be improved; (ii) most models need prior knowledge on number of clusters, which is not always available; (iii) there is a demand for faster computational speed.
Results: We propose to tackle these challenges with **Para**llelized Split Merge Sampling on **D**irichlet **P**rocess **M**ixture **M**odel (the Para-DPMM model). Unlike classic DPMM methods that perform sampling on each single data point, the split merge mechanism samples on the cluster level, which significantly improves convergence and optimality of the result. The model is highly parallelized and can utilize the computing power of high performance computing (HPC) clusters, enabling massive inference on huge datasets. Experiment results show the model outperforms current widely used models in both clustering quality and computational speed.
Availability and implementation: Source code is publicly available on https://github.com/tiehangd/Para_DPMM/tree/master/Para_DPMM_package.
Contact: xhx@ics.uci.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Parallelized droplet based single cell transcriptomic profiling has achieved significant progress in recent years (Zheng *et al.*, 2017). Compared to traditional methods, parallelized droplet based systems utilize Gel bead in Emulsion (GEM) to capture single cells in parallel (the co-occurrence of multiple cells in one GEM is eliminated by controlling the dilution in the reagent oil). The 3' messenger RNA digital counting is performed through the reading of unique molecular identifiers (UMI) in each GEM. Massive parallelized droplet based systems have the following properties: (i) Samples are processed in parallel in microfluidic chip with multiple channels, allowing the analysis of a much larger number of cells. (ii) The multiplet rate (rate of multiple cells in one GEM) is controlled to be less than 2% by limiting dilution, and performs direct counting of molecule copies using UMI. (iii) The detection result of UMI is minimally affected by the composition of nucleobases and gene length, resulting in low transcript bias. Because of these properties, parallelized droplet based single cell transcriptomic profiling has resulted in the creation of mass single cell genomic datasets and lead to a number of advancements such as better approaches for transplant monitoring (Athanasiadis *et al.*, 2017) and detection of rare cell populations (Proserpio and Lönnberg, 2016).

Cell clustering based on transcriptomic profiles plays an important role in single cell analysis. It identifies and characterizes cell

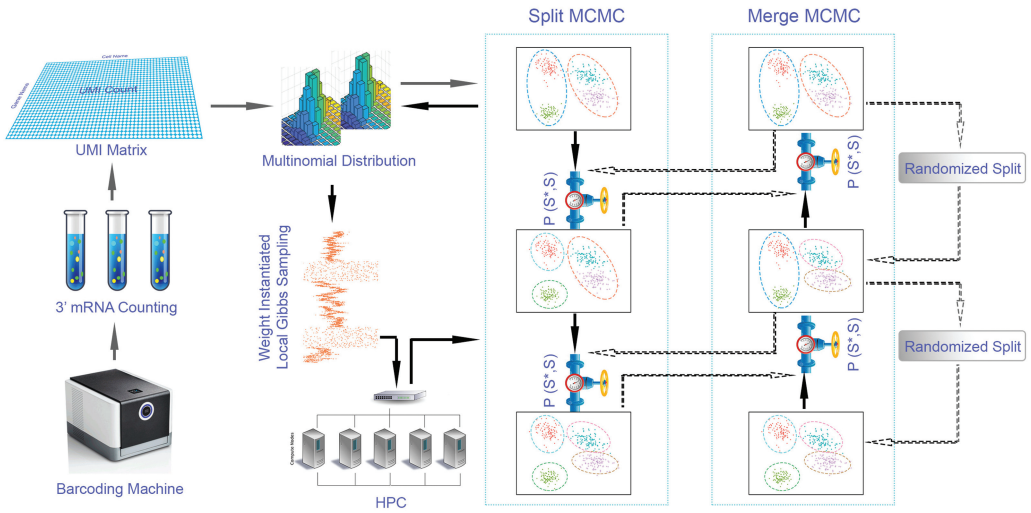

Fig. 1. Workflow of para-DPMM model

subtypes from heterogeneous tissues and enhances understanding of cell identity and functionality. Classic clustering methods such as K-means (Kanungo *et al.*, 2002), hierarchical clustering (Manning *et al.*, 2008), spectral clustering (Ng *et al.*, 2001) can be applied directly to single cell clustering. Given the high dimensionality of single cell data, a widely adopted approach involves combining dimension reduction with classic clustering. Common combinations of methods include t-SNE with K-means (Grün *et al.*, 2015), PCA with hierarchical clustering (Žurauskienė and Yau, 2016) and Rt-SNE with model based clustering (Fraley and Raftery, 2002; Žurauskienė and Yau, 2016). The high dimensionality problem can also be tackled by replacing Euclidean distances with similarity measures that are robust in sparse high dimension space such as ranking on shared nearest neighbors (SNN) (Satija *et al.*, 2015), ward linkage (Guo *et al.*, 2015) and graph based clustering methods which perform graph partition by finding maximal cliques on the similarity matrix (Xu and Su, 2015). Other recent works proposed to solve the problem with consensus clustering (Kiselev *et al.*, 2017), regulon formation (Aibar *et al.*, 2017), multi-kernel learning (Wang *et al.*, 2017). Imputation is shown to be effective for performance improvement (Lin *et al.*, 2017). Dirichlet Mixture Model (DMM) is well suited for single cell clustering as the discrete counting information in the UMI matrix can be directly modeled through Multi-nomial distribution and conjugate prior likelihood pairs result in efficient inference (Blei *et al.*, 2003). Recent applications of DMM to single cell analysis have achieved good results (DuVerle *et al.*, 2016; Sun *et al.*, 2017). However, there are still challenges to be addressed: (i) There is demand for faster computational speed for newly created mass single cell datasets, which can be realized through parallelization and utilization of HPC clusters. However, standard DPMM methods are difficult to parallelize. (ii) For challenging tasks, as shown in the Experiment Section, clustering quality can be significantly improved. (iii) Most methods are designed for continuous data, while the scRNA-Seq data is formed of discrete UMI counts. Conversion of the UMI counts to continuous measure would alter the straight-forward interpretation and it is more appealing to directly model discrete data. (iv) Most methods need prior knowledge on the number of clusters (DuVerle *et al.*, 2016; Wang and Xu, 2015), which is not always available for rawly processed single cell data and limits their ability to identify cellular heterogeneity within the same cluster.

The Para-DPMM model (Fig. 1) proposed in this paper addresses these limitations. Its inference is highly parallelized and can be readily implemented on large HPC clusters, which results in high computational speed. For large scaled datasets with tens of thousands of genes and cells, such as the fresh PBMC 68 K dataset used in our second case study, the clustering is completed in a couple of minutes using 32 cores. The model is able to automatically determine the number of clusters with its non-parametric Bayesian setting. Its sampling is highly efficient. New clusters are created by splitting existing clusters instead of setting aside a single data point, which avoids going through the low probability density regions in the sampling space and achieves fast convergence and improved optimality. The model achieved more than 20% improvement on ARI (adjusted rand index) for large challenging tasks over current widely used models in the experiment.

These improvements are due to a split-merge Markov Chain Monte Carlo (MCMC) inference algorithm that we developed for this problem. Unlike variational approximation (Blei and Jordan, 2006; Ji *et al.*, 2017; Kurihara *et al.*, 2007) or collapsed Gibbs sampling (Escobar and West, 1995; Neal, 1992), the inference algorithm is a weight-instantiated sampling method, in which cluster parameters are explicitly instantiated as variables (Ishwaran and James, 2001; Ishwaran and Zarepour, 2002). Variational approximation algorithms lend themselves to parallelization, but are not guaranteed to converge to ground truth distribution. Collapsed Gibbs sampling enables intra-cluster parallelization (Lovell *et al.*, 2013; Williamson *et al.*, 2013), where the number of processes is parallelized to be of the same order as the number of clusters. Its parallelization level is relatively low. The split-merge sampling in Para-DPMM enables inter-cluster parallelization (Chang and Fisher, 2013; Favaro and Teh, 2013; Papaspiliopoulos and Roberts, 2008), in which threads running in parallel are of the same order as data points, resulting in a high level of parallelization. To improve sampling efficiency, new clusters are formed by either splitting an existing cluster or merging two clusters together. Local Gibbs sampling is performed inside each cluster to propose reasonable split proposals with high acceptance ratio.

2 Materials and methods

2.1 Data and model framework

The output of the droplet-based single cell profiling pipeline is a matrix storing UMI counts with rows indexing genes and columns indexing cells. Each entry in this UMI matrix x_i^u is the UMI count of gene u barcoded in cell i. We use $\vec{x}_i$ to denote the expression of all genes in cell i measured in terms of read counts. Single cell clustering is performed on the UMI matrix with size $V \times N$, where V is the total number of genes and N is the total number of cells.

In the transcriptomic clustering model, the cluster assignment c_i of cell i is the discrete hidden variable to be inferred based on observed gene expression $\vec{x}_i$. The model is built on the Dirichlet process mixture model (DPMM), which is the infinite form of the Dirichlet mixture model (DMM). For detailed description of DPMM model please refer to Görür and Edward Rasmussen (2010). In the generative form of DPMM model, with parameters $\vec{\theta}_k \in \mathrm{R}^V$, gene expression $\vec{x}_i$ is generated based on the Multi-nomial distribution

$$p(\vec{x}_i|c_i = k, \vec{\theta}_k) = \text{Multinomial}(\vec{x}_i|\vec{\theta}_k) \sim \prod_{u=1}^{V} \theta_{k,u}^{x_i^u} \tag{1}$$

where $\sum_{u=1}^{V} \theta_{k,u} = 1$. Notation meaning is listed in Table 1. Priors for $\vec{\theta}_k$ are accordingly set to be Dirichlet distribution with hyper parameter λ

$$\text{Dirichlet}(\vec{\theta}_k|\lambda) = \frac{\Gamma(\lambda V)}{\Gamma(\lambda)^V} \prod_{u=1}^{V} \theta_{k,u}^{\lambda-1} \tag{2}$$

For posterior inference of c_i given gene expression x_i, the iterative inference process can be described as

$$(\pi_1, \ldots, \pi_K, \pi_{K+1}) \sim p(\pi|\vec{c}, \alpha) \tag{3}$$

$$\vec{\theta}_k \propto p(\vec{x}_{\{k\}}|\vec{\theta}_k) p(\vec{\theta}_k|\lambda) \; \forall k \in \{1, \ldots, K, K+1\} \tag{4}$$

$$c_i \propto p(c_i|\pi) p(\vec{x}_i|c_i = k, \vec{\theta}_k) \tag{5}$$

where $\{\pi_1, \ldots, \pi_K\}$ represents the mixing proportions of existing clusters and π_{K+1} represents the proportion of next new cluster to be generated.

2.2 Efficient parallel sampling for the DPMM model

Implementing parallel inference for the DPMM model is not trivial. Careful examination of the dependence relationships among the variables is necessary. While collapse Gibbs sampling (Neal, 2000) simplifies the sampling process (when priors are conjugate to the likelihood), its parallelization is not straight forward (Chang and Fisher, 2013) as data points become directly dependent on each other after the cluster parameters are integrated out. The cluster indicators $\vec{c}$ can be seen as a fully connected Markov Random Field (MRF) and can't be parallelized based on proofs in Gonzalez *et al.* (2011).

For the split merge sampling adopted in this paper, the cluster parameters $\vec{\theta}$ are explicitly instantiated as variables. The cluster assignments $\vec{c}$ and cluster parameters $\vec{\theta}$ can be mapped to a two coloring MRF with one color being $\vec{c}$ and the other being $\vec{\theta}$. Based on theorems in Gonzalez *et al.* (2011), all cluster assignments $\vec{c}$ can then be sampled in parallel, as they are conditionally independent of each other given $\vec{\theta}$. Theoretically, the maximum number of computing cores that can be utilized in parallel equals the number of data points.

Sampling is inefficient in this naive parallel approach. It is difficult to open new clusters as parameters sampled directly from the prior are usually a poor fit of the data. Also, extremely large number of sampling steps are needed for common scenarios such as: (i) dividing the current cluster into more fine grained clusters; (ii) transferring a significant portion of data points in the current cluster to another cluster and (iii) merging two clusters. The naive approach has to go through a series of low probabilistic density intermediate steps in the sampling space to reach the more optimized setting. In real world applications where sampling time is limited, this approach leads to sub-optimality.

Table 1. Notations

Notation	Meaning
$\vec{x}$	Collection of cells
$\vec{c}$	Cluster assignments of cells
$\vec{\theta}$	Cluster parameters
π	Mixing proportions in the Dirichlet process
λ	Dirichlet hyper parameter for cluster parameters $\vec{\theta}$
α	Parameter for Chinese restaurant process
c_i	Cluster assignment for cell i
$\vec{\theta}_k$	Collection of parameters for the multi-nomial distribution in cluster k
θ_k^u	Parameter for multi-nomial distribution of gene u in cluster k
$\vec{x}_i$	The gene expression of ith cell
x_i^u	The UMI count of gene u in cell i
$\vec{x}_{\{k\}}$	The gene expression of cells assigned to cluster k
$\bar{c}$	Local split sub-cluster assignment
$\bar{\theta}_r$	Parameters for local sub-clusters, $r \in \{0, 1\}$
n_k	Number of cells in cluster k
$\bar{n}_r$	Number of cells in sub-cluster r, $r \in \{0, 1\}$
N	Total number of cells
K	Current number of clusters in the model
V	Total number of genes

The split merge sampling mechanism was adopted to solve this problem. New clusters are created by splitting existing clusters, instead of setting aside a single data point. This endows newly created clusters with sensible parameters and data membership from the very beginning, and avoids going through low probability intermediate states, thus leading to faster convergence. To guarantee that the process converges to the desired stationary state, a MCMC is built to satisfy the detailed balance by either accepting or rejecting the splitting proposal. Merge moves are introduced to make the Markov chain ergodic, its proposal is accepted based on a separate acceptance ratio.

2.3 Inference through split/merge MCMC sampling

The MCMC sampler is characterized by the states and acceptance ratio of state transitions. For the Para-DPMM model, each state is defined as $S = \{\vec{\pi}, \vec{\theta}, \vec{c}, \vec{x}\}$. For each update, the algorithm proposes a new state $S_* = \{\vec{\pi}_*, \vec{\theta}_*, \vec{c}_*, \vec{x}_*\}$ which is reachable from the old state by either a split or merge move. As the derivation for the two moves are similar, here we take split move as example. The proposed state is either accepted or rejected based on the acceptance ratio:

$$p(S_*, S) = \min\left[1, \frac{p(S_*)}{p(S)} \frac{q(S|S_*)}{q(S_*|S)}\right] \tag{6}$$

where $p(S)$ is the likelihood of the old state, $p(S_*)$ is the likelihood of the new state, $q(S_*|S)$ is the transition probability from old state to new state and $q(S|S_*)$ is the reversed transition probability. Updates with this acceptance ratio satisfy the detailed balance of Markov chain and are guaranteed to converge to the stationary state.

Derivation of the acceptance ratio is based on the specific split merge mechanism we choose. The random split with binomial

Table 2. Performance comparison on different data scales

	S-Set			M-Set			L-Set		
	ARI	RI	HI	ARI	RI	HI	ARI	RI	HI
Para-DPMM	**0.654 ± 0.021**	**0.849 ± 0.011**	**0.699 ± 0.023**	**0.670 ± 0.012**	**0.855 ± 0.004**	**0.711 ± 0.008**	**0.688 ± 0.016**	**0.863 ± 0.008**	**0.726 ± 0.016**
DIMM-SC	0.578 ± 0.029	0.803 ± 0.006	0.606 ± 0.012	0.352 ± 0.009	0.662 ± 0.018	0.324 ± 0.036	0.331 ± 0.013	0.650 ± 0.023	0.301 ± 0.047
CellTree	0.270 ± 0.006	0.637 ± 0.015	0.274 ± 0.031	0.289 ± 0.009	0.643 ± 0.016	0.285 ± 0.032	0.273 ± 0.008	0.634 ± 0.024	0.268 ± 0.048
Seurat	0.503 ± 0.017	0.776 ± 0.010	0.553 ± 0.019	0.576 ± 0.032	0.815 ± 0.008	0.630 ± 0.015	0.463 ± 0.028	0.785 ± 0.018	0.569 ± 0.036
PCA-Reduce	0.294 ± 0.015	0.684 ± 0.018	0.368 ± 0.036	0.284 ± 0.016	0.681 ± 0.021	0.363 ± 0.041	0.302 ± 0.014	0.688 ± 0.016	0.376 ± 0.032
K-means	0.312 ± 0.014	0.680 ± 0.004	0.360 ± 0.008	0.302 ± 0.007	0.678 ± 0.012	0.355 ± 0.023	0.312 ± 0.019	0.683 ± 0.005	0.367 ± 0.010
SC3	0.602 ± 0.018	0.823 ± 0.006	0.646 ± 0.012	0.614 ± 0.026	0.828 ± 0.018	0.657 ± 0.036	0.640 ± 0.017	0.840 ± 0.010	0.680 ± 0.020
SIMLR	0.203 ± 0.014	0.606 ± 0.006	0.212 ± 0.012	0.334 ± 0.011	0.699 ± 0.013	0.398 ± 0.026	0.381 ± 0.008	0.724 ± 0.012	0.449 ± 0.024
CIDR	0.222 ± 0.011	0.605 ± 0.014	0.209 ± 0.028	0.196 ± 0.009	0.617 ± 0.015	0.235 ± 0.030	0.205 ± 0.016	0.628 ± 0.009	0.255 ± 0.018

Note: Para-DPMM outperformed all comparison methods for a large margin on all experiment settings.

distribution is straight forward, yet its performance is not satisfactory, as it doesn't utilize any information in the data points and the proposals are unlikely to be reasonable. The acceptance ratio is usually very low in this scenario.

An improved method is to run local Gibbs sampling in each cluster to learn cluster sub-structures before the split proposal. An additional indicator variable $\bar{c} = \{0, 1\}$ is assigned to each data point in cluster k to denote which data points will be in the sub-clusters after the possible split. Local Gibbs sampling computes the probability of assigning data points to either side of the split:

$$p(\bar{c}_i = r|\bar{c}_{\{r\},\neg i}, \rightarrow x_{\{r\}}, \bar{\theta}) = \frac{\bar{n}_{\{r\},\neg i} p(\rightarrow x_i|\bar{\theta}_r, \bar{c}_i = r)}{\bar{n}_{\{0\},\neg i} p(x_i|\bar{\theta}_0, \bar{c}_i = 0) + \bar{n}_{\{1\},\neg i} p(\rightarrow x_i|\bar{\theta}_1, \bar{c}_i = 1)} \ \forall r \in \{0, 1\} \tag{7}$$

where $\bar{c}_{\{r\},\neg i}$ are the assignments to sub-cluster r excluding cell i and $\bar{n}_{\{r\},\neg i}$ is the number of cells in sub-cluster r excluding cell i. Parameters for local sub-clusters are then updated based on

$$\bar{\theta}_r \propto p(\vec{x}_{\{r\}}|\bar{\theta}_r) p(\bar{\theta}_r|\bar{\lambda}) \ \forall r \in \{0, 1\} \tag{8}$$

where $\bar{\lambda}$ is the Dirichlet hyper parameter for sub-cluster parameters $\bar{\theta}$.

The number of iterations for local Gibbs sampling involves a trade off between accuracy and computational cost. In practice we found one iteration is already enough for the model to achieve decent performance. Transition probability $q(S^*|S)$ based on the local Gibbs sampling is a product of conditional probabilities of assigning each observation $i \in \{k\}$ to a split mixture component as given by Equation (7). The transition probability from the new state back to old state $q(S|S^*)$ is also needed. This reverse transition is the merge operation. In contrast to the split operation which has diversified splitting choices, the merge operation is deterministic as there is only one way to merge two components into one component, so $q(S|S^*) = 1$.

To calculate the acceptance ratio in Equation (6), we also need to evaluate the ratio of likelihood between the new state and the old state $\frac{p(S^*)}{p(S)}$. According to the generative procedure of DPMM, $\frac{p(S^*)}{p(S)}$ can be decomposed as

$$\frac{p(S^*)}{p(S)} = \frac{p(\vec{\pi}_*, \vec{\theta}_*, \vec{c}_*, \vec{x}_*)}{p(\vec{\pi}, \vec{\theta}, \vec{c}, \vec{x})} = \frac{p(\vec{\pi}_*)}{p(\vec{\pi})} \frac{p(\vec{c}_*|\vec{\pi}_*)}{p(\vec{c}|\vec{\pi})} \frac{p(\vec{\theta}_*|\lambda)}{p(\vec{\theta}|\lambda)} \frac{p(\vec{x}|\vec{c}_*, \vec{\theta}_*)}{p(\vec{x}|\vec{c}, \vec{\theta})} \tag{9}$$

$\frac{p(S^*)}{p(S)}$ can be readily derived from Equation (9) to be

$$\frac{p(S^*)}{p(S)} = \alpha \frac{\bar{\pi}_{k_0}^{\bar{n}_{k_0}-1} \bar{\pi}_{k_1}^{\bar{n}_{k_1}-1}}{\pi_k^{n_k-1}} \frac{\Gamma(\lambda V)}{\Gamma(\lambda)^V} \frac{\prod_{u=1}^V \theta_{k_0,u}^{\lambda-1} \prod_{u=1}^V \theta_{k_1,u}^{\lambda-1}}{\prod_{u=1}^V \theta_{k,u}^{\lambda-1}} \times \frac{(\prod_{i\in\{k_0\}} \prod_{u=1}^V \theta_{k_0,u}^{x_i^u})(\prod_{i\in\{k_1\}} \prod_{u=1}^V \theta_{k_1,u}^{x_i^u})}{(\prod_{i\in\{k\}} \prod_{u=1}^V \theta_{k,u}^{x_i^u})} \tag{10}$$

The detailed derivation is included in the Supplementary Material.

2.4 Random splits in merge moves

A key consideration when constructing the MCMC sampler is to avoid the acceptance rate to be too small. For this reason, as mentioned in the previous section, we replaced random split with local Gibbs sampling when designing split moves. When the split is more reasonable, the likelihood of the new state $p(S^*)$ significantly increases, thus increasing the acceptance rate. Merge moves can be seen as split moves going from the new state back to the old state. To increase the acceptance rate of merge moves, we should do exactly the opposite. And we included in the model a separate pair of merge/split moves which is randomized to propose good merges (as here the splitted cluster is the old state whose likelihood we are trying to decrease). For randomized merge moves, as $p(\bar{c}_i = r|\bar{c}_{\{r\},\neg i}, \vec{x}_{\{r\}})$ is simply $\frac{1}{2}$, the ratio of transition probability becomes

$$\frac{q(S|S_*)}{q(S_*|S)} = \left(\frac{1}{2}\right)^{n_{k_0}+n_{k_1}-2} \tag{11}$$

The derivation of $\frac{p(S_*)}{p(S)}$ is similar to the split move.

Please note the split moves and merge moves that take place in the model belong to two independent MCMC chains. The integrated dynamic process thus formed is a rational MCMC with guaranteed convergence as long as the atomic moves are selected randomly from the two chains and each of the chains satisfies detailed balance (Tierney, 1994).

3 Performance in cellular heterogeneity analysis

The Para-DPMM model was applied to the challenging task of distinguishing three T cell types (CD4+/CD25+ regulatory T cells, CD4+/CD45RA+/CD25- naive T cells and CD8+/CD45RA+ naive cytotoxic T cells) similar to Sun *et al.* (2017). The data was provided by 10X Genomics and is publicly available (Zheng *et al.*, 2017). Three datasets of different scales were used: (i) a set of 1200 cells with the 1000 top variable genes (small scale, referred to as S-Set below), (ii) a set of 3000 cells with the 3000 top variable genes (medium scale, referred to as M-Set) and (iii) a set of 6000 cells with the 5000 top variable genes (large scale, referred to as L-Set). In these

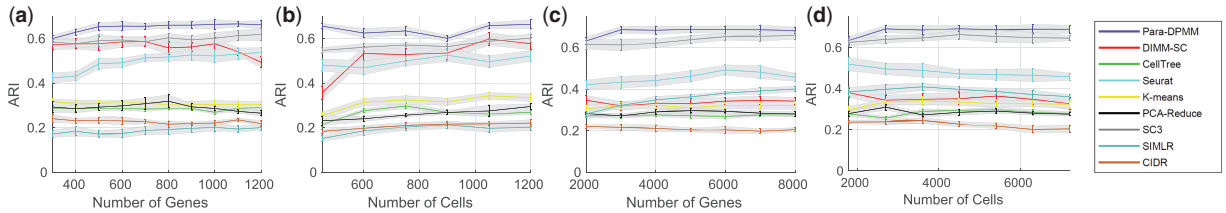

Fig. 2. (**a**) Performance (ARI) with respect to different number of genes on S-Set. (**b**) Performance with respect to different number of cells on S-Set. (**c**) Performance with respect to different number of genes on L-Set. (**d**) Performance with respect to different number of cells on L-Set

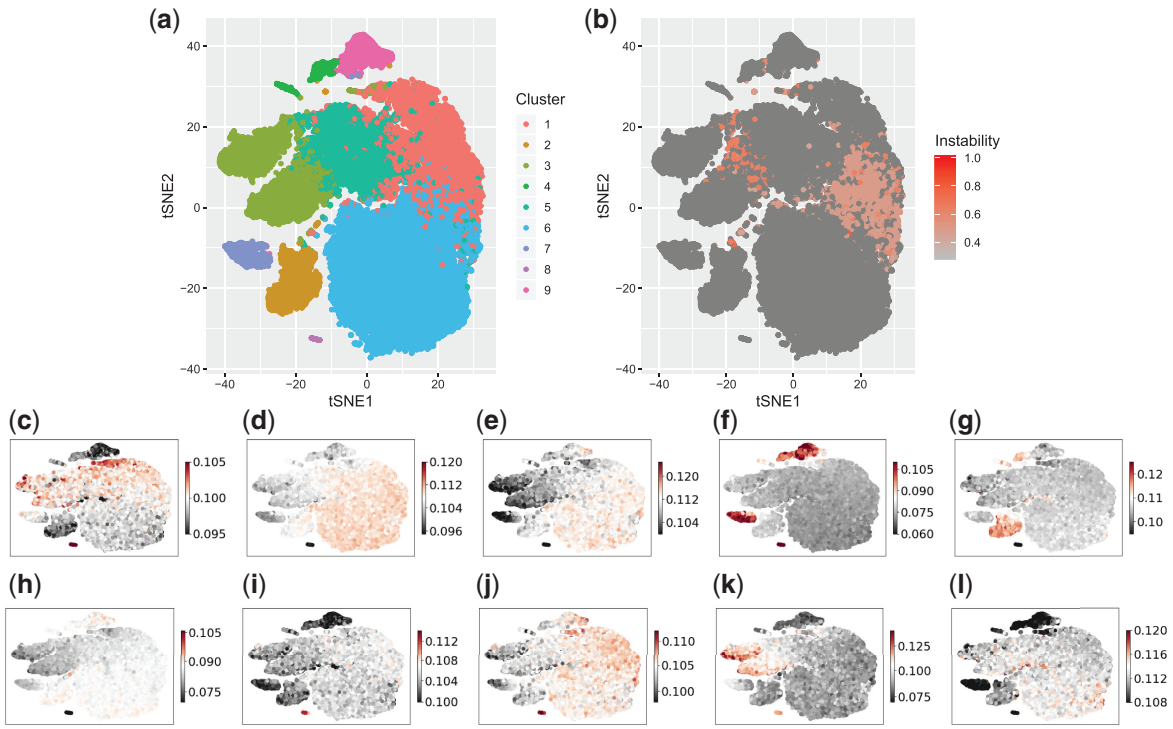

Fig. 3. (**a**) t-SNE visualization of Para-DPMM clustering on Fresh PBMC 68K dataset; (**b**) stability of the clustering result; (**c**) CD4+/CD25+ regulatory T cell distribution; (**d**) CD4+/CD45ra+/CD25- naive T cell distribution; (**e**) CD8+/CD45ra+ naive cytotoxic T cell distribution; (**f**) CD14+ monocytes distribution; (**g**) CD19+ B cell distribution; (**h**) CD34+ cell distribution; (**i**) CD4+ helper T cell distribution; (**j**) CD4+/CD45ro+ memory T cell distribution; (**k**) CD56+ Natural Killer cell distribution and (**l**) CD8+ cytotoxic T cell distribution

datasets, cells were randomly selected from the population, we ensured that each cell type was equally represented in the datasets. The top variable genes were selected based on their standard deviations across the cell transcriptome profiles in the UMI matrix.

We compared Para-DPMM's performance with other currently widely used models, including Seurat (Satija *et al.*, 2015), CellTree (DuVerle *et al.*, 2016), PCA-Reduce (Žurauskienė and Yau, 2016), SC3 (Kiselev *et al.*, 2017), SIMLR (Wang *et al.*, 2017), CIDR (Lin *et al.*, 2017) and DIMM-SC (Sun *et al.*, 2017). For models needing prior knowledge on the number of clusters, we set it to the ground truth value. The results are shown in Table 2. The model's performance was measured with three benchmarks: Adjusted Rand Index (ARI), Rand Index (RI) and Hubert's Index (HI). Rand Index (RI) measures the similarity between two clusterings, it ranges between 0 and 1 with a perfect match being scored 1. Adjusted Rand Index (ARI) is the corrected-for-chance version of Rand Index, it scores 0 for random matches. Hubert's Index (HI) (Hubert and Arabie, 1985) is another popular metric for comparing partitions. It has the advantage of probabilistic interpretation in addition to being corrected for chance. Its value ranges between –1 and 1. The analysis below mainly refers to ARI due to its wide adoption in the field.

As shown in Table 2, Para-DPMM outperformed all comparison methods for a large margin on all experiment settings, and the trend is more significant in the large data setting (L-Set), where it achieved approximately 5% improvement on ARI compared to SC3 and is more than 20% better than the other comparison methods. We further applied Para-DPMM to the full dataset, which includes 32 695 cells and 32738 genes, where the model achieved a 71.47% score on ARI.

As mentioned in the previous section, the performance improvement is due to the split merge mechanism which enables the model to make efficient moves in the sampling space and avoid being trapped in sub-optimal situations. The underlying Dirichlet Process allows the model to automatically decide the most appropriate number of clusters for the data, and the parallelized sampling enhances the convergence speed.

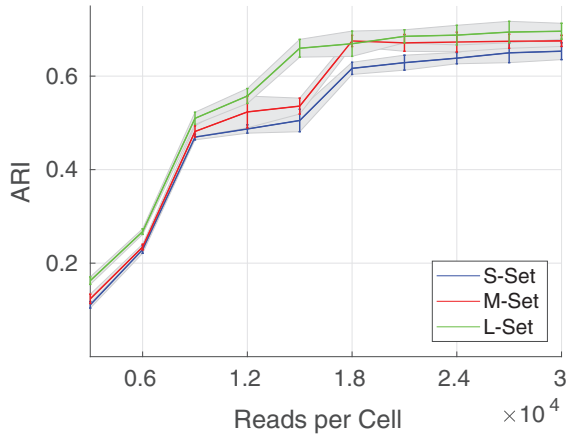

Fig. 4. Influence of sequencing depth on model performance

We further explored the relationship of model performance with different number of genes and cells. Results are presented in Figure 2. For the small scale setting, the performance slightly increased with gene number (Fig. 2a), as the cell clusters are more distinguishable with the added information. This result shows Para-DPMM's ability to handle the increasing dimensionality in data, as posterior inference of multi-nomial model only involves multiplying one dimension at a time and naturally circumvents the high dimensionality challenge. The DIMM-SC model achieved good performance with number of genes less than 1000. The Seurat algorithm performed better with the increase of the number of genes. Its clustering is based on embedding cells to graphs and analyzing the cliques formed. Increasing the number of genes made the edge weight more accurate. The performance of other comparison methods is not significantly influenced by number of genes. For large scale setting, the performance of Para-DPMM remained stable (Fig. 2c and d). The performance slightly improved when more genes were involved, as more UMI counts are accumulated in the process and clusters becomes more distinguishable.

4 Analysis on fresh PBMC 68 K dataset

In order to demonstrate our model's ability to deal with real world large datasets, in this case study we applied Para-DPMM to a publicly available fresh PBMC 68 K dataset (Publicly available on https://support.10xgenomics.com/single-cell-gene-expression/datasets). The dataset is composed of 68 K freshly processed peripheral blood mononuclear cells obtained from one donor. Samples are divided between T cells(> 80%), NK cells(∼6%), B cells(∼6%) and myeloid cells(∼7%). Clustering analysis on the data reveals proportion of each cell types, identifies cell types with similar transcriptome profiles, finds finer grained subtypes in existing categories and discovers rare cell populations.

The results of the Para-DPMM clustering can be seen in Figure 3a. Our model divided the data points into 9 clusters, a result close to the 10 clusters identified with human expert knowledge (Zheng *et al.*, 2017). The clustering is in accordance with the boundaries of clusters visualized in the t-SNE plot. To test the stability of the clustering we repeated the process 50 times and measured the probability of each cell being assigned to different clusters. As illustrated in Figure 3b, the clusters were quite stable, though there was some uncertainty on the intersection regions of cluster 1 with 6 and cluster 3 with 5. We also tested the influence of hyper parameter α on the clustering result and found different values of α had little

Table 3. Performance comparison on pairwise PBMC cell types

	CD4+CD45ro+/CD34+			CD8+/CD4+CD45ra+CD25-			CD56+/CD4+CD25+		
	ARI	RI	HI	ARI	RI	HI	ARI	RI	HI
Para-DPMM	**0.706 ± 0.037**	**0.853 ± 0.019**	**0.706 ± 0.037**	**0.750 ± 0.035**	**0.875 ± 0.018**	**0.750 ± 0.035**	**0.990 ± 0.004**	**0.995 ± 0.002**	**0.990 ± 0.004**
DIMM-SC	0.672 ± 0.042	0.836 ± 0.021	0.672 ± 0.042	0.562 ± 0.048	0.781 ± 0.024	0.562 ± 0.048	0.971 ± 0.007	0.985 ± 0.003	0.971 ± 0.007
CellTree	0.250 ± 0.031	0.625 ± 0.016	0.250 ± 0.031	0.161 ± 0.034	0.580 ± 0.017	0.161 ± 0.034	0.782 ± 0.038	0.891 ± 0.019	0.782 ± 0.038
Seurat	0.432 ± 0.048	0.716 ± 0.024	0.432 ± 0.048	0.286 ± 0.012	0.643 ± 0.006	0.286 ± 0.012	0.581 ± 0.054	0.790 ± 0.027	0.581 ± 0.054
PCA-Reduce	0.621 ± 0.040	0.811 ± 0.020	0.621 ± 0.040	0.459 ± 0.038	0.729 ± 0.019	0.459 ± 0.038	0.528 ± 0.032	0.764 ± 0.016	0.528 ± 0.032
K-Means	0.202 ± 0.010	0.601 ± 0.005	0.202 ± 0.010	0.143 ± 0.008	0.572 ± 0.004	0.143 ± 0.008	0.746 ± 0.034	0.873 ± 0.017	0.746 ± 0.034
SC3	**0.695 ± 0.026**	**0.847 ± 0.013**	**0.695 ± 0.026**	0.709 ± 0.016	0.855 ± 0.008	0.709 ± 0.016	0.980 ± 0.004	0.991 ± 0.002	0.980 ± 0.004
SIMLR	0.465 ± 0.034	0.761 ± 0.017	0.465 ± 0.034	0.376 ± 0.017	0.721 ± 0.008	0.376 ± 0.017	0.726 ± 0.026	0.878 ± 0.013	0.726 ± 0.026
CIDR	0.684 ± 0.014	0.859 ± 0.007	0.684 ± 0.014	0.430 ± 0.012	0.745 ± 0.006	0.430 ± 0.012	0.823 ± 0.011	0.921 ± 0.005	0.823 ± 0.011

Note: The performance of SC3 was comparable to Para-DPMM for the CD4+CD45ro+/CD34+ pair. Para-DPMM achieved better performance than all comparison methods in the other two pairs.

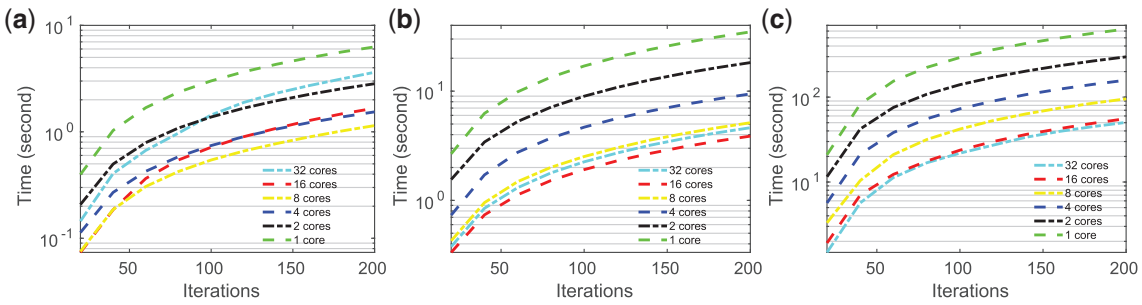

Fig. 5. (**a**) Comparison of computing time on S-set; (**b**) comparison of computing time on L-set and (**c**) comparison of computing time on PBMC 68K dataset

Table 4. Computing speed comparison of different models

	Para-DPMM	DIMM-SC	CellTree	Seurat	PCA-reduce	K-means	SC3	SIMLR	CIDR
S-Set	1.14 s	33.10 s	1.82 s	28.46 s	5.56 s	5.30 s	3.11 min	9.26 min	7.09 s
M-Set	2.16 s	4.77 min	3.06 s	1.23 min	2.07 min	20.12 s	5.21 min	1.28 h	54.39 s
L-Set	3.88 s	16.98 min	6.41 s	2.48 min	11.10 min	48.95 s	8.43 min	6.65 h	6.77 min

effect on the clustering when ranging from 0.1 to 1. The reason for this robustness lies in the relative strength of prior (compared to likelihood) in determining posterior cluster distribution. Given the high dimensionality (number of genes) of the dataset, the likelihood dominates the posterior distribution in the sampling process and the small difference caused by different α in the prior distribution is negligible.

Since there is no available ground truth cell labeling for this dataset to obtain detailed knowledge about the specific cell types which compose the clusters, we resorted to 10 purified cell populations Publicly available on https://support.10xgenomics.com/single-cell-gene-expression/datasets of the cell types that were previously identified in this dataset using human expert knowledge. The cell type's gene expression profile was obtained by averaging the profiles of each purified population. The cell type assignment was based on the covariance between profiles of the cell types and samples. The distribution of each cell type is visualized in Figure 3c–3l. CD14+ monocytes, CD19+ B cells and CD56+ NK cells were easily separated from other cell types. On the other hand, we observed a significant overlap of CD4+/CD45+/CD25- naive T cell, CD8+/CD45ra+ naive cytotoxic T cells and CD4+/CD45+ memory cells on the t-SNE plot.

These cell type distributions easily explain certain clusters, more specifically clusters 2, 3 and 7, which are composed mostly of CD19+ B cells, CD56+ NK cells and CD14+ monocytes, respectively. Other clusters are composed of multiple cell types. Cluster 6 is a combination of CD4+/CD45+/CD25- naive T cells and CD8+/CD45ra+ naive cytotoxic T cells, clusters 1 and 5 also contain a significant amount of these cell types while being mainly composed of CD4+/CD25+ regulatory T cells.

We found that three pairs of cells were largely overlapping in the clusters, namely CD4+/CD45ro+ memory T with CD34+ cells, CD8+ cytotoxic T with CD4+/CD45ra+/CD25- naive T cells and CD56+ Natural Killer with CD4+/CD25+ regulatory T cells. We further tested our model's ability to distinguish these three pairs of cells. 2000 cells from each category were randomly selected and clustered based on the 16 000 genes with top expression variation. Results are presented in Table 3. The performance of SC3 was comparable to Para-DPMM for the CD4+CD45ro+/CD34+ pair. Para-DPMM achieved better performance than all comparison methods in the other two pairs. We found it was significantly easier to distinguish between CD56+ Natural Killer and CD4+/CD25+ regulatory T cells than the other two pairs.

5 Applicable scenario analysis

The Para-DPMM model should be applied to datasets created with UMI based techniques. In UMI labeling based systems, the UMI counts are independent of transcript length and is suitable to model with Multi-nomial distribution. As illustrated in Islam *et al.* (2014) and Phipson *et al.* (2017), earlier non-UMI based techniques introduced bias during the cDNA amplification phase, the resulting expression matrix is correlated with transcript length and normalizations used in RPKM and FPKM are necessary. For these datasets, clustering methods based on continuous similarity measures such as Seurat, SC3 and PCA-Reduce are more appropriate choices.

Current droplet-based single cell sequencing techniques has the drop out phenomenon, where not all transcriptome information is captured during the cell reads. This results in a sparser expression matrix when the sequencing depth is not deep enough. To test the robustness of Para-DPMM regarding to varying sequencing depth, we measured the model performance on different data scales (S-Set, M-Set and L-Set) with sequencing depth ranging from 3000 to 30 000 reads per cell. As shown in Figure 4, the model performance is highly correlated with sequencing depth when reads per cell is less than 10 000 and performance is stable after sequencing depth reaches 18 000 reads per cell. The recommended minimum sequencing depth for 10X platform is 50 000 reads per cell (Baran-Gale *et al.*, 2017), which lies well inside the model's robust region.

6 Scalability analysis on parallel computing clusters

In this section, we analyze the scalability of the model. Para-DPMM was implemented on a HPC cluster built with the BeeGFS system, the model uses the OpenMP framework and is able to run in parallel on multiple cores in one node. We tested the model's scalability with up to 32 cores. Further improvement on parallelization is possible if

the model is extended with the MPI framework, which is not in the scope of this paper. We requested 64 GB RAM for all experiment settings.

We recorded the model's computing time on variating number of cores for different dataset sizes, results are shown in Figure 5. The trade off between the gain and cost of parallelization is clearly exemplified on the small dataset (S-Set, shown in Fig. 5a), where fastest computing speed was achieved with eight computing cores, after which computing became slower as the number of cores further increased. The cost of parallelization came from coordination between different threads, including parallel tasks creation, I/O of the shared memory and communications between threads, which eventually offsets the gains. Figure 5a demonstrates it is not necessary to use more than eight cores for training on the small dataset. The strength of parallelized implementation becomes evident when dealing with large scaled datasets, such as the PBMC 68K data. As shown in Figure 5c, the computing speed is approximately 12 times faster when using 32 cores compared to a single core. The computing time is initially inversely proportional to the number of cores, and then gradually converge to constant time.

6.1 Based on Amdahl's law

$$\text{Speed Up} = \frac{1}{\frac{P}{N} + S} \tag{12}$$

where P denotes the parallelized portion in the code, N denotes number of cores and $S = 1 - P$ denotes the serial portion in the code, the parallelization ratio of the model implementation is as high as 91%.

We also compared other models' computing speed (Please note the computing time is significantly affected by factors at software engineering level. This comparison should only serve as guidance for real world applications, and not to be used for inferring algorithm complexity.) with Para-DPMM (Table 4). For fairness, the measurements include only running time and exclude time for data I/O and dimension reduction (in Seurat). All models were run on eight cores and towards convergence. Para-DPMM and CellTree are significantly faster than other comparison methods. Para-DPMM is about 30% faster than CellTree on small data setting and 40% on large settings.

7 Discussion

As shown in the experiments, the Para-DPMM model scales well with different dataset size (Table 2) and with variating data dimensionality (Fig. 2). This scalability and versatility enables its possible wide application on real world genomic systems. Clustering analysis on the fresh PBMC dataset (Fig. 3a) identified cells with similar transcriptome profiles and helped uncover finer grained heterogeneous structures for each cell type. As illustrated in the applicable scenario analysis (Section 5), the model should only be applied to UMI-based datasets.

To cope with the large scaled single cell transcriptomic datasets, the model's inference process is highly parallelized and ready for applications in large computing clusters. This parallelization is achieved by explicitly instantiating the cluster parameters of the model and makes data points conditionally independent of each other. While the model can potentially utilize as many computing cores as the number of data points, 32 cores are generally enough for current large datasets (Fig. 5c).

The split-merge mechanism is adopted in the model to significantly improve convergence and optimality of the result. The integrated split-merge process is formed with two independent MCMC chains which generates high acceptance ratio for both split and merge moves. We performed detailed comparison with current widely used methods, and Para-DPMM model simultaneously achieved significant improvements on both clustering accuracy and computing speed. The model's performance increases with higher dimensionality of the data, and it automatically infers number of clusters from the dataset without using prior knowledge.

Several extensions of the Para-DPMM model are possible. For single cell datasets created from heterogeneous sources (e.g. PBMC cells from multiple individuals), the model could be extended to include hierarchical processes to discover fine grained sub-structures in the clusters. Given the availability of purified cell populations, the clustering accuracy could be further improved with semi-supervised guidance. We will explore these possible extensions in the near future.

Funding

The work is partially supported by NSF IIS-1715017, NSF DMS1763272, and a grant from the Simons Foundation 594598.

Conflict of Interest: none declared.

References

Aibar,S. *et al*. (2017) Scenic: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083 EP.

Athanasiadis,E.I. *et al*. (2017) Single-cell rna-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat. Commun.*, **8**, 2045.

Baran-Gale,J. *et al*. (2017) Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics*, **17**, elx035.

Blei,D.M. and Jordan,M.I. (2006) Variational inference for Dirichlet process mixtures. *Bayesian Anal.*, **1**, 121–143.

Blei,D.M. *et al*. (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.*, **3**, 2003.

Chang,J.F. and Fisher,J.W. (2013) Parallel sampling of dp mixture models using sub-clusters splits. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*, *NIPS'13*. Vol. 1. Curran Associates Inc., USA, pp. 620–628.

DuVerle,D.A. *et al*. (2016) Celltree: an r/bioconductor package to infer the hierarchical structure of cell populations from single-cell rna-seq data. *BMC Bioinformatics*, **17**, 363.

Escobar,M.D. and West,M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, **90**, 577–588.

Favaro,S. and Teh,Y.W. (2013) Mcmc for normalized random measure mixture models. *Statist. Sci.*, **28**, 335–359.

Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Gonzalez,J.E. *et al*. (2011) *Parallel Gibbs Sampling: From Colored Fields to Thin Junction Trees*. PMLR, Sydney, Australia.

Görür,D. and Edward Rasmussen,C. (2010) Dirichlet process gaussian mixture models: choice of the base distribution. *J. Computer Sci. Technol.*, **25**, 653–664.

Grün,D. *et al*. (2015) Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, **525**, 251 EP.

Guo,M. *et al*. (2015) Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLOS Comput. Biol.*, **11**, e1004575–e1004528.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.

Ishwaran,H. and James,L.F. (2001) Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.*, **96**, 161–173.

Ishwaran,H. and Zarepour,M. (2002) Exact and approximate sum representations for the dirichlet process. *Can. J. Stat.*, **30**, 269–283.

Islam,S. *et al.* (2014) Quantitative single-cell rna-seq with unique molecular identifiers. *Nat. Methods*, **11**, 163.

Ji,G. *et al.* (2017) From patches to images: a nonparametric generative model. In: *Icml*. PMLR. Sydney, Australia.

Kanungo,T. *et al.* (2002) An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Machine Intel.*, **24**, 881–892.

Kiselev,V.Y. *et al.* (2017) Sc3: consensus clustering of single-cell rna-seq data. *Nat. Methods*, **14**, 483 EP –.

Kurihara,K. *et al.* (2007). Collapsed variational dirichlet process mixture models. In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 2796–2801.

Lin,P. *et al.* (2017) CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.*, **18**, 59.

Lovell,D. *et al.* (2013) *ClusterCluster: Parallel Markov Chain Monte Carlo for Dirichlet Process Mixtures*. ArXiv e-prints.

Manning,C.D. *et al.* (2008) *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Neal,R.M. (1992) *Bayesian Mixture Modeling*. Dordrecht, Springer Netherlands, pp. 197–211.

Neal,R.M. (2000) Markov chain sampling methods for dirichlet process mixture models. *J. Comput. Graph. Stat.*, **9**, 249–265.

Ng,A.Y. *et al.* (2001) On spectral clustering: analysis and an algorithm. In: *Advances in Neural Information Processing Systems*, MIT Press, pp. 849–856.

Papaspiliopoulos,O. and Roberts,G.O. (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95**, 169–186.

Phipson,B. *et al.* (2017) Gene length and detection bias in single cell RNA sequencing protocols [version 1; referees 4 approved]. *F1000Research*, **6**, 595.

Proserpio,V. and Lönnberg,T. (2016) Single-cell technologies are revolutionizing the approach to rare cells. *Immunol. Cell Biol.*, **94**, 225.

Satija,R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotech.*, **33**, 495–502.

Sun,Z. *et al.* (2017) *Dimm-Sc: A Dirichlet Mixture Model for Clustering Droplet-Based Single Cell Transcriptomic Data* Bioinformatics.

Tierney,L. (1994) Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701–1728.

Wang,B. *et al.* (2017) Visualization and analysis of single-cell RNA-seq data by Kernel-based similarity learning. *Nat. Methods*, **14**, 414.

Wang,X.-F. and Xu,Y. (2015) Fast clustering using adaptive density peak detection. *Stat. Methods Med. Res.*, **1**, 0962280215609948.

Williamson,S.A. *et al.* (2013) Parallel Markov chain Monte Carlo for nonparametric mixture models. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning, ICML'13*. Vol. 28, pp. I–98–I–106. JMLR.org., Atlanta, GA, USA.

Xu,C. and Su,Z. (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, **31**, 1974–1980.

Zheng,G.X.Y. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.

Žurauskienė,J. and Yau,C. (2016) Pcareduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, **17**, 140.