10x GENOMICS®

TECHNICAL NOTE

Resolving Cell Types as a Function of Read Depth and Cell Number

INTRODUCTION

Chromium™ Single Cell Solutions enable gene expression profiling of 100–10,000 cells per reaction. When sequencing costs or capacity are limiting, there is often a trade-off between sequencing a larger cell number (breadth) and sequencing a smaller cell number with more reads (depth).

Heimberg et al recently reported that the coordinated expression of subsets of genes in transcriptional modules decreases the effective dimensionality of gene expression data from number of genes to number of modules detected. This suggests that major cell types distinguished by gene expression modules can be resolved from single cell gene expression data sequenced at low depth.

Here, we examine how sequencing depth and cell number influence the detection of major cell types in peripheral blood mononuclear cells (PBMCs).

METHODS

A sequencing library was prepared from ~4,000 PBMCs from a healthy donor following the *Chromium Single Cell 3' Reagent Kits v2 User Guide (CG00052).* The library was sequenced on an Illumina® HiSeq 2500 instrument to near saturation (90.5%) using paired-end sequencing (26 bp Read 1 and 98 bp Read 2) with a single sample index (8 bp i7 index) to target over 50K reads per cell.

To study the impact of read depth on PBMC cell type detection, raw reads were randomly sampled to target read depths: 500, 1K, 2.5K, 5K, 7.5K, 10K, 15K, 25K, and 50K *Mean Reads per Cell*. Subsampling at each depth was replicated 3 times. For each subsampled data set, *cellranger count* was run using parameters --expect-cells=5000 and --transcriptome=GRCh38. The Cell Ranger™ pipeline was used to align reads, generate gene-barcode unique molecular identifier (UMI) count matrices, and perform clustering and gene expression analysis.

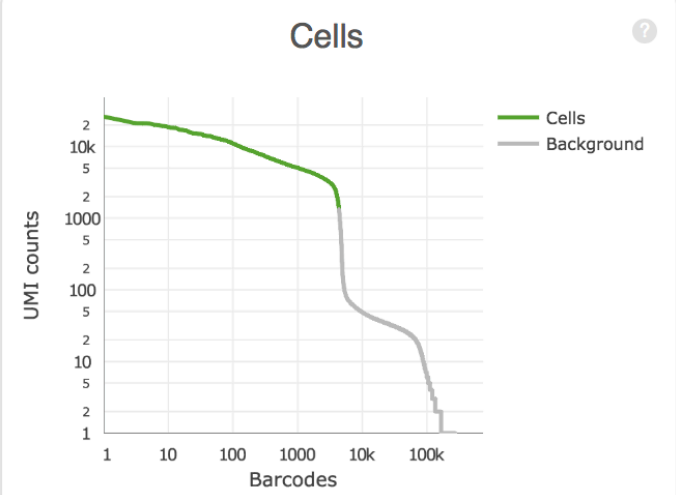
To study the impact of cell number on PBMC cell type detection, subsets of cell barcodes were randomly sampled from each *cellranger count* filtered gene-barcode UMI count matrix at each subsampled read depth. Samples of barcodes were obtained at 100, 200, 400, 600, 800, 1K, 2K, 3K, and 4K cells. Each subsampled list of cell barcodes was used as input to *cellranger reanalyze* to regenerate Cell Ranger secondary analysis, including clustering and gene expression analysis.

RESULTS

Read depth subsampling

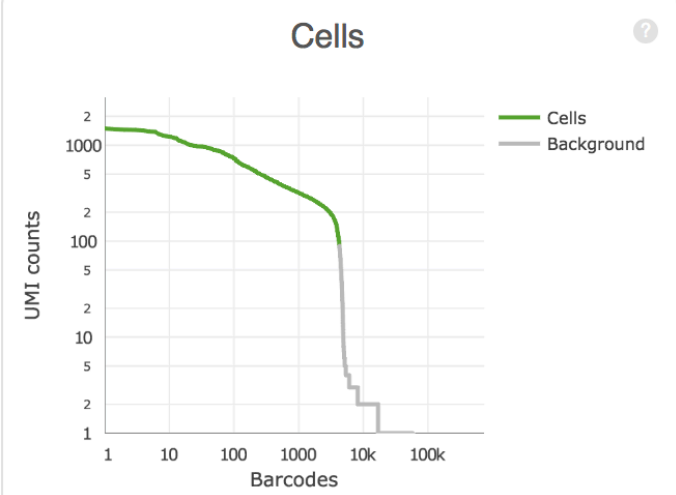
The *Estimated Number of Cells* remained stable (within 2%, range 4,273–4,353) across the range of read depths under investigation (Figure 1). Even at ~500 *Mean Reads per Cell*, the sharp knee observed at the UMI count cutoff for cell calling was similar to the high sequencing depth dataset (86,503 *Mean Reads per Cell*).

Mean Reads per Cell: 86,503



Estimated Number of Cells	4,353
Fraction Reads in Cells	92.7%
Mean Reads per Cell	86,503
Median Genes per Cell	1,234
Total Genes Detected	19,774
Median UMI Counts per Cell	3,858

Mean Reads per Cell: 506



Estimated Number of Cells	4,273
Fraction Reads in Cells	92.3%
Mean Reads per Cell	506
Median Genes per Cell	160
Total Genes Detected	13,309
Median UMI Counts per Cell	243

Figure 1. Comparison of the barcode-UMI count rank plots and cell level metrics at high and low read depths. The green line indicates UMI counts for cell-associated barcodes.

In contrast, Figure 2 illustrates that the percent variations for assay sensitivity metrics across the read depths under investigation were more consistent with the percent variation in read depth. For example, *Median Genes per Cell* varied by 87% (range 161–1,234).

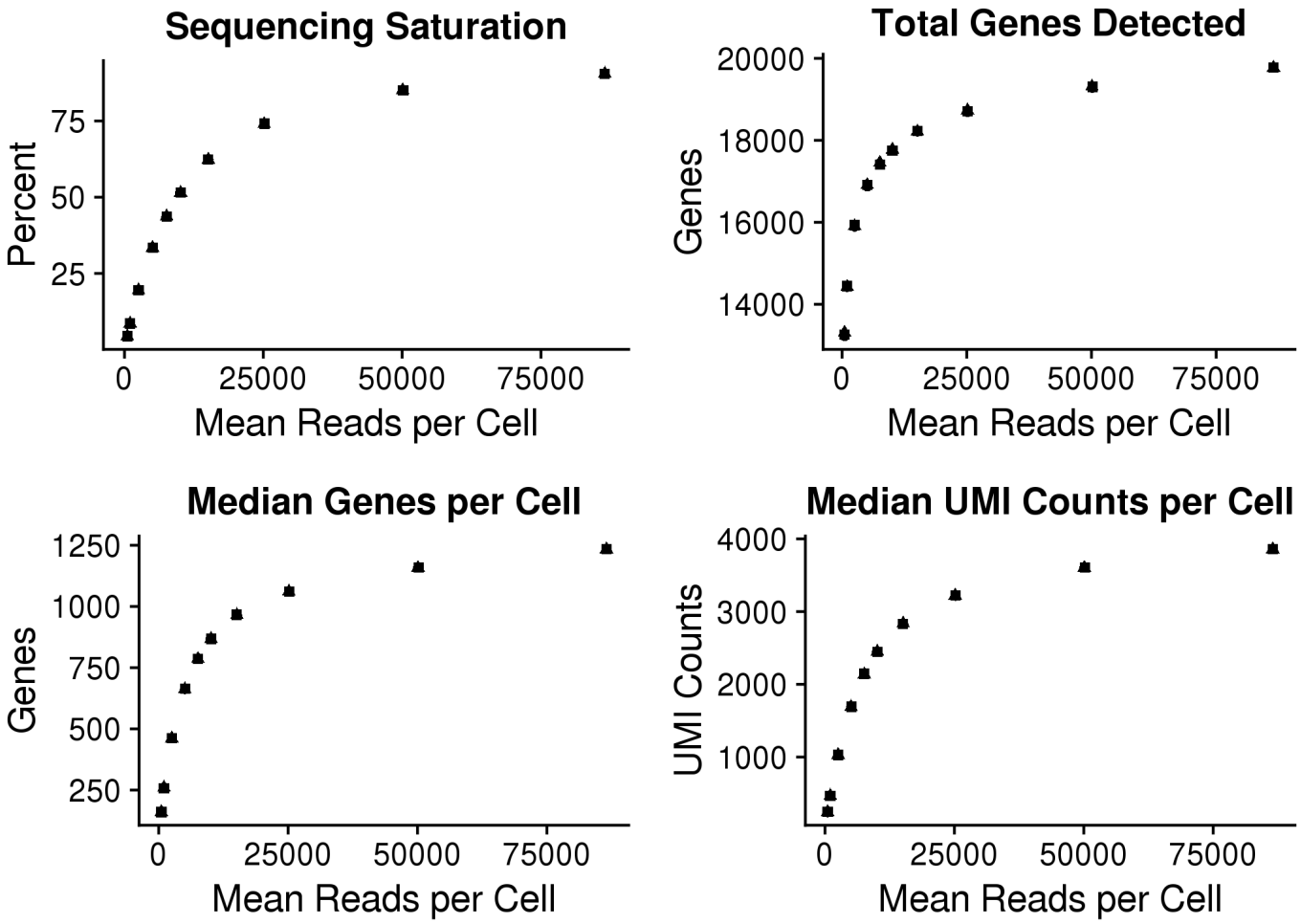


Figure 2. Examples of sensitivity metrics as a function of read depth for a dataset of ~4,000 PBMCs (3 replicates).

Despite the lower sensitivity metrics at lower read depths, putative subpopulations of cells were identified via graph-based clustering and visualized by t-SNE plots (Figure 3A). To determine PBMC subpopulation identities, graph-based clusters were assigned to one of the four major cell types based on the enrichment of known markers (Figure 3B). T cells were identified based on CD3D or CD3E; Natural Killer NK cells on GNLY or NKG7; B cells on CD79A or CD79B and Monocytes on CD14 or FCGR3A.

To validate the clustering-based cell type assignment, results for each subsampled data set were benchmarked against the full data set. For read depths as low as 500 *Mean Reads per Cell*, the four major cell types were classified with median accuracy of 93% (Figure 4A). As shown in the confusion matrix (Figure 4B), B cells and monocytes were classified with over 98% specificity and sensitivity and NK cells and T cells were classified with ~93% sensitivity and ~95% specificity. NK cells and T cells have more closely related gene expression profiles than B cells and monocytes. The accurate classification of these closely related cell types appears to be impacted by a decrease in the number of reads per cell.

Across the range of subsampled read depths, cell type classification accuracy was at least 92% with median values at each subsampled depth ranging from 93–99% (Figure 4A). At 2.5K *Mean Reads per Cell*, accuracy was relatively consistent at 98%, but ranged from 92–97% across replicates at 500 and 1K *Mean Reads per Cell*. The median percent decrease in cell type classification accuracy at ~500 *Mean Reads per Cell* (~7%) was much less than the calculated percent decrease for detection sensitivity metrics such as *Median UMI Counts per Cell* (94%), *Median Genes per Cell* (87%), and *Total Genes Detected* (33%) at the same read depth.

Target Mean Reads per Cell	500	5,000	25,000	86,503
Cell Count	4,273	4,306	4,323	4,353

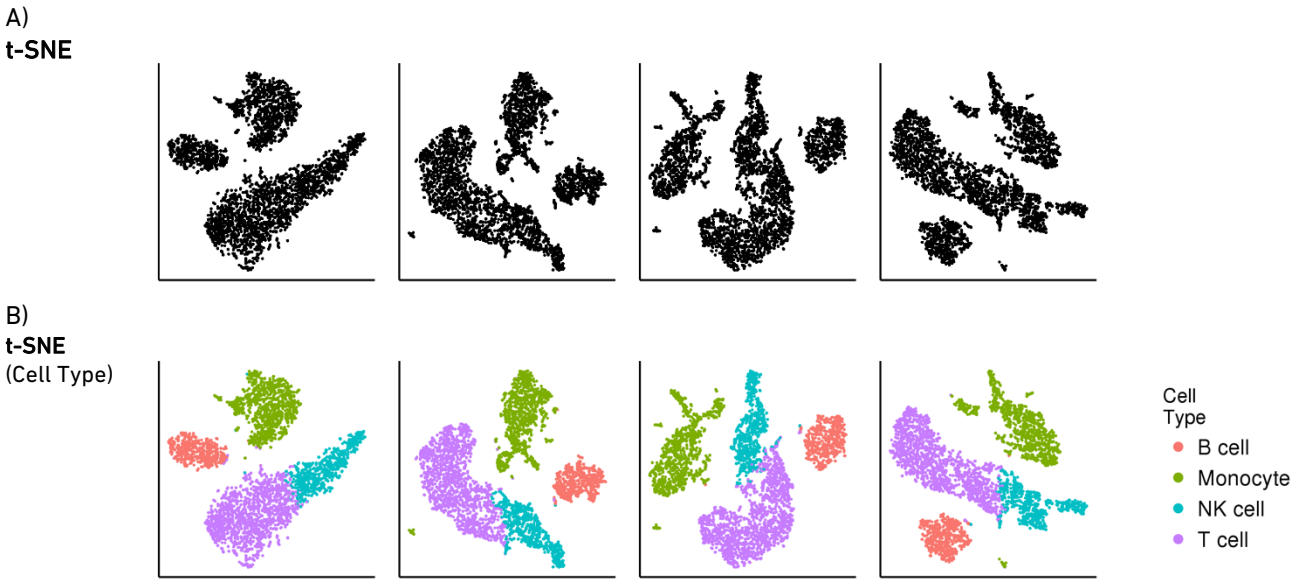


Figure 3. Effect of read depth subsampling on t-SNE visualization. A) t-SNE visualization of each subsampled read depth compared to the full dataset (86,503 Mean Reads per Cell). B) t-SNE visualization colored by cell-type classification.

Cell subsampling

To evaluate the effect of the number of cells in a sample on cell type classification accuracy, the PBMCs detected at each read depth were subsampled from 100–4,000 cells. From the Cell Ranger™ analysis of each cell barcode subsampled data set, putative subpopulations of cells were identified via graph-based clustering and annotated by the same marker genes as the read depth subsampling analysis.

At 50K *Mean Reads per Cell*, as read depth approached sequencing saturation (85%), a larger range of median cell type classification accuracy was observed (82–99%) across profiled cell counts (Figure 4C). At 1,000 cells, the accuracy estimates across replicates varied by 10% (87–98%) while at 100 cells, the accuracy estimate was 55–92%, indicating that variability increased with a decrease in cell count.

When profiling cell type classification accuracy across subsampled read depths and cell counts (Figure 4D), the median cell type classification accuracy was 93% at 500 *Mean Reads per Cell* with at least 2,000 cells profiled. In contrast, the accuracy was below 90% when up to 600 cells were profiled at 86,503 *Mean Reads per Cell* (Figures 4C–4D).

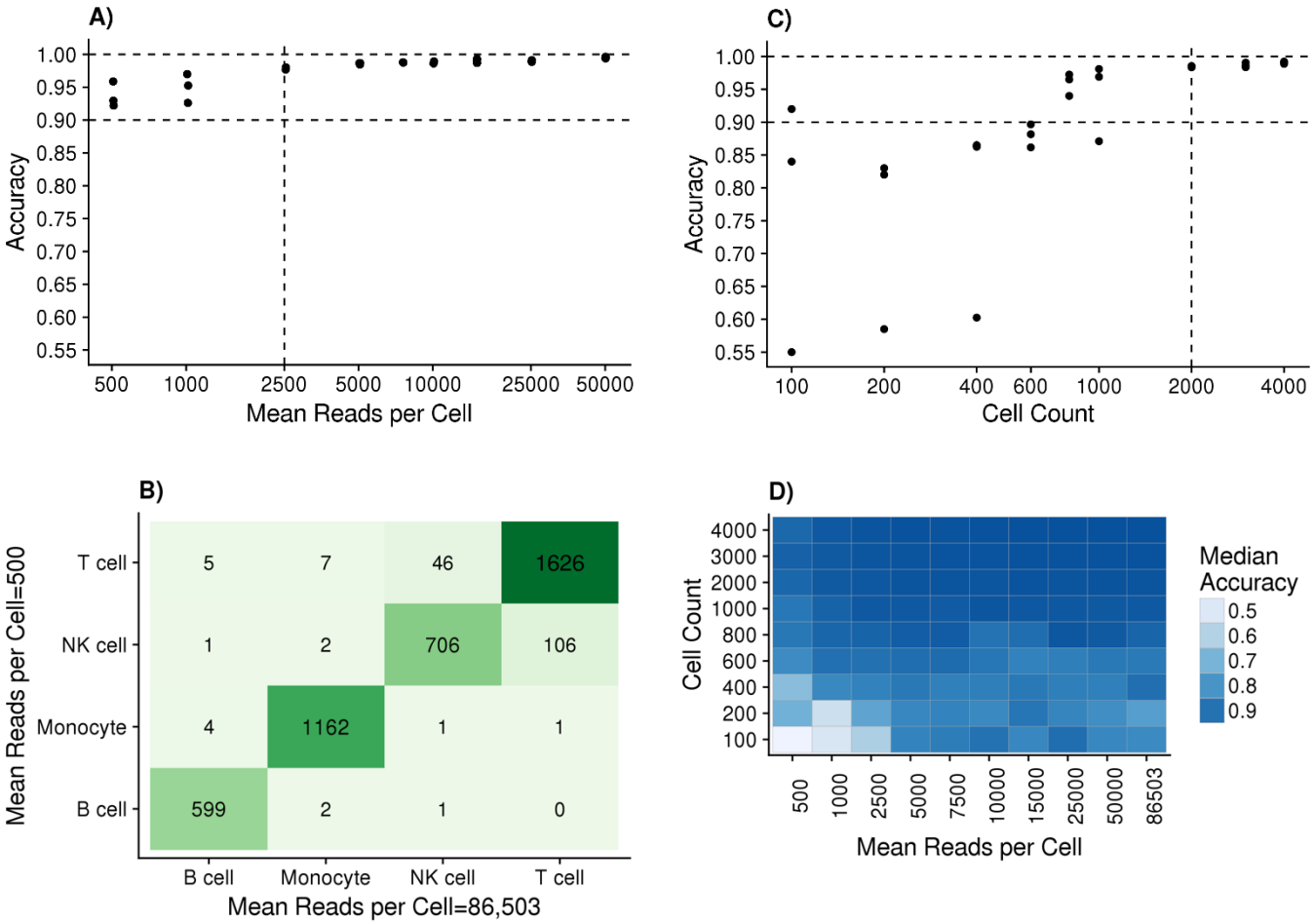


Figure 4. A) Cell type classification accuracy shown as a function of read depth (3 replicates). B) Confusion matrix comparing the classification of ~4,000 filtered barcodes (no subsampling) into the four major PBMC cell types at 500 versus 86,503 Mean Reads per Cell. The diagonal cells (dark green) indicate concordant cell type classifications between the subsampled data (rows) and the high sequencing depth data set (columns). C) Cell type classification accuracy shown as a function of cell count (3 replicates per count at 50K Mean Reads per Cell). D) Heatmap showing median cell type classification accuracy as a function of read depth.

CONCLUSION

The accurate classification of the four major cell types in PBMCs at low read depths indicates that major transcriptional programs can be inferred from shallow sequencing due to low dimensionality in gene expression data sets. Subsampling reads from the ~4,000 PBMCs dataset resulted in cell type classification accuracy of at least 92% at ~500 *Mean Reads per Cell* and was up to 98% at 2.5K *Mean Reads per Cell.* Lower and variable estimates of cell type classification accuracy were observed with a decrease in total cells, particularly below 600 cells even at high sequencing depth (50K *Mean Reads per Cell*). A reduction in the total cells results in a smaller sample size for each cell type, contributing to variability in cell type classification accuracy and underscoring the importance of high cell throughput for single cell gene expression.

REFERENCES

- Heimberg G, Bhatnagar R, El-Samad H, Thomson M. "Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing." *Cell Systems* **2016** 2(4):239-50.
- *Fresh Frozen Human Peripheral Blood Mononuclear Cells for Single Cell RNA Sequencing* (CG00036)
- *Chromium™ Single Cell 3' Reagent Kits v2 User Guide* (CG00052)

Notices

Document Number

CG000148 Rev A *Technical Note*

Legal Notices

© 2018 10x Genomics, Inc. All rights reserved. Duplication and/or reproduction of all or any portion of this document without the express written consent of 10x Genomics, Inc., is strictly forbidden. Nothing contained herein shall constitute any warranty, express or implied, as to the performance of any products described herein. Any and all warranties applicable to any products are set forth in the applicable terms and conditions of sale accompanying the purchase of such product. 10x Genomics provides no warranty and hereby disclaims any and all warranties as to the use of any third party products or protocols described herein. The use of products described herein is subject to certain restrictions as set forth in the applicable terms and conditions of sale accompanying the purchase of such product. "10x", "10x Genomics", "Changing the Definition of Sequencing", "Chromium", "GemCode", "Loupe", "Long Ranger", "Cell Ranger" and "Supernova" are trademarks of 10x Genomics, Inc. All other trademarks are the property of their respective owners. All products and services described herein are intended FOR RESEARCH USE ONLY and NOT FOR USE IN DIAGNOSTIC PROCEDURES.

The use of 10x Product(s) in practicing the methods set forth herein has not been validated by 10x, and such non-validated use is NOT COVERED BY 10X STANDARD WARRANTY, AND 10X HEREBY DISCLAIMS ANY AND ALL WARRANTIES FOR SUCH USE.

Nothing in this document should be construed as altering, waiving or amending in any manner 10x Genomics, Inc., terms and conditions of sale for the Chromium™ Controller, consumables or software, including without limitation such terms and conditions relating to certain use restrictions, limited license, warranty and limitation of liability, and nothing in this document shall be deemed to be Documentation, as that term is set forth in such terms and conditions of sale. Nothing in this document shall be construed as any representation by 10x Genomics, Inc that it currently or will at any time in the future offer or in any way support any application set forth herein.

Customer Information and Feedback

For technical information or advice, please contact our Customer Technical Support Division online at any time.

Email: support@10xgenomics.com

10x Genomics

7068 Koll Center Parkway

Suite 401

Pleasanton, CA 94566 USA