APPLICATION NOTE

Single Cell Sequencing Performance Using the Parse Biosciences Evercode™ WT Kit and the G4™ Sequencing Platform

- Generate accurate, reproducible scRNA-Seq data with the Evercode WT Kit and the G4 Platform.
- G4 sequencing data integrates seamlessly into Parse Biosciences data analysis.
- Rapid, flexible, and cost-effective scRNA-Seq experiments with the G4 Platform.

Introduction

Single cell RNA sequencing (scRNA-Seq) has revolutionized basic and translational research in immunology, developmental biology, and cancer by enabling the resolution of distinct cell populations within heterogeneous samples.[1,2] The resolution and scale of scRNA-seq is made possible by advances in next-generation sequencing (NGS), where greater throughput and faster turnaround times have dramatically reduced the cost of experimentation.

Despite these strengths, there is a growing need for scRNA-Seq library preparation technologies that support greater cell throughput and sensitivity with lower background noise. To address these challenges, Parse Bioscience has developed Evercode™ split-pool combinatorial barcoding to enable scalable scRNA-Seq analysis.

In this application note, we highlight the use of the Parse Biosciences Evercode™ WT kit in combination with the novel G4 Sequencing Platform for rapid, flexible sequencing by synthesis (SBS). We demonstrate excellent accuracy and high concordance of the G4 Platform with the industry-leading NGS platform, all while delivering a faster turnaround time.

G4 Specifications for scRNA-Seq

The G4™ Sequencing Platform is a highly versatile benchtop sequencing platform that is well-suited for demanding scRNA-Seq applications. The G4 Platform leverages a novel, 4-color Rapid SBS chemistry to deliver highly accurate reads in less than a day for typical Evercode scRNA-seq experiments (74x86 bp reads). To maximize flexibility, the G4 Platform enables users to load 1 to 4 flow cells at a time, with each flow cell comprising 4 fluidically independent lanes to facilitate sample multiplexing.

The G4 Platform outputs FASTQ format files that integrate seamlessly with existing bioinformatics tools. Users may elect to automatically demultiplex samples on-instrument via sample indices provided by the sample sheet or off-instrument using the Singular Genomics rapid demultiplexing tool.

More information about G4 specifications, such as run time, accuracy, and quality metrics, can be found on the Singular Genomics website.

	G4	NovaSeq 6000
Estimated Number of Cells	15,128	15,136
Median Transcripts/Cell	3,332	3,000
Median Genes/Cell	1,756	1,642
Mean Reads/Cell	16,598	16,589
Number of Reads	251,088,023	251,088,023
Sequencing Saturation	0.428	0.492
BC1 (RT) > Q30	0.915	0.906
BC2 > Q30	0.953	0.930
BC3 > Q30	0.957	0.940
cDNA > Q30	0.939	0.918

Table 1. Singular Genomics G4™ and Illumina® NovaSeq 6000 read quality and library metrics.
Sample: all-well
Wells: A1-D12

Methods

Library Preparation, Sequencing, and Analysis

Evercode WT scRNA-Seq libraries were prepared using frozen peripheral blood mononuclear cells (PBMCs) collected from four healthy donors. A single, 15k sub-library was split between the different sequencing arms of this effort of this work. A fraction of the resulting sequencing-ready library was converted for sequencing on the G4 using the Singular Genomics library conversion kit with non-indexed PCR primers (Singular Genomics, Part #700119). Library quality was assessed via Qubit and Tapestation HSD5000 kit. The Singular Genomics library was sequenced on the G4 using a single F2 flow cell and 74x86 reads to yield 251M paired reads. The Illumina® library was sequenced on the NovaSeq 6000 using two lanes of an S4 flow cell, with 74x86 reads, and then downsampled to 251M reads prior to downstream processing.

FASTQ files from each platform were processed using Parse Biosciences data analysis solution with default settings, then the resulting anndata.h5ad files were further processed using Scanpy[3] (v1.8.2). First, doublets were removed using Scanpy's wrapper of Scrublet[4] (v0.2.3) with an expected doublet rate of 7%. Next, cells were removed that either fell below the 5th percentile of the number of unique transcripts per cell and the number of unique genes per cell in the dataset or had greater than 10% of their unique transcripts assigned to mitochondrially encoded genes. Genes were removed that were not detected in at least five cells, then the top 2,000 highly variable genes (HVGs) were determined using Scanpy's implementation of the Seurat v3 HVG detection method.[5] Finally, a latent representation of the datasets was generated using the scvi-tools[6,7] scVI model (v0.14.5; layers = 2, epochs = 400) on the top 2,000 HVGs using raw count data. Nearest-neighbor graphs, UMAPs, and Leiden clusters were generated using Scanpy's modules on the scVI latent representation. Automated cell type identification was performed using CellTypist[8] (v1.2.0) and the "Immune_all_low" model with majority voting applied. Differential gene expression analysis was performed using Scanpy's rank genes module with the Wilcoxon method on the library size-normalized, log-transformed counts.

Results

Sequencing of the Evercode WT library on one Singular Genomics F2 flow cell yielded 251M high quality read pairs (%Q30>90 for Read 1 and Read 2). Read quality and library metrics were highly concordant across platforms, with the estimated number of cells closely matching expectation based on the experimental design **(Table 1)**. Uniform manifold approximation and projection (UMAP) and Leiden clustering revealed a nearly identical grouping of single cell transcriptomes across platforms **(Figure 1A)**, with strongly correlated pseudo-bulk transcription profiles across the platforms (R^2 = 0.9932) **(Figure 1B)** and similar representation of T-Cells, Monocytes, and B-Cells, as inferred by marker gene expression **(Figure 1C)**.

To explore the datasets in greater depth, we used CellTypist to perform unsupervised clustering and automated cell type annotation on the combined dataset. The adjusted rand index (ARI), a measure of similarity between sets of labels, indicated nearly identical unsupervised cluster labels and cluster transcription profiles across the datasets **(Figure 2A, 2B)**.

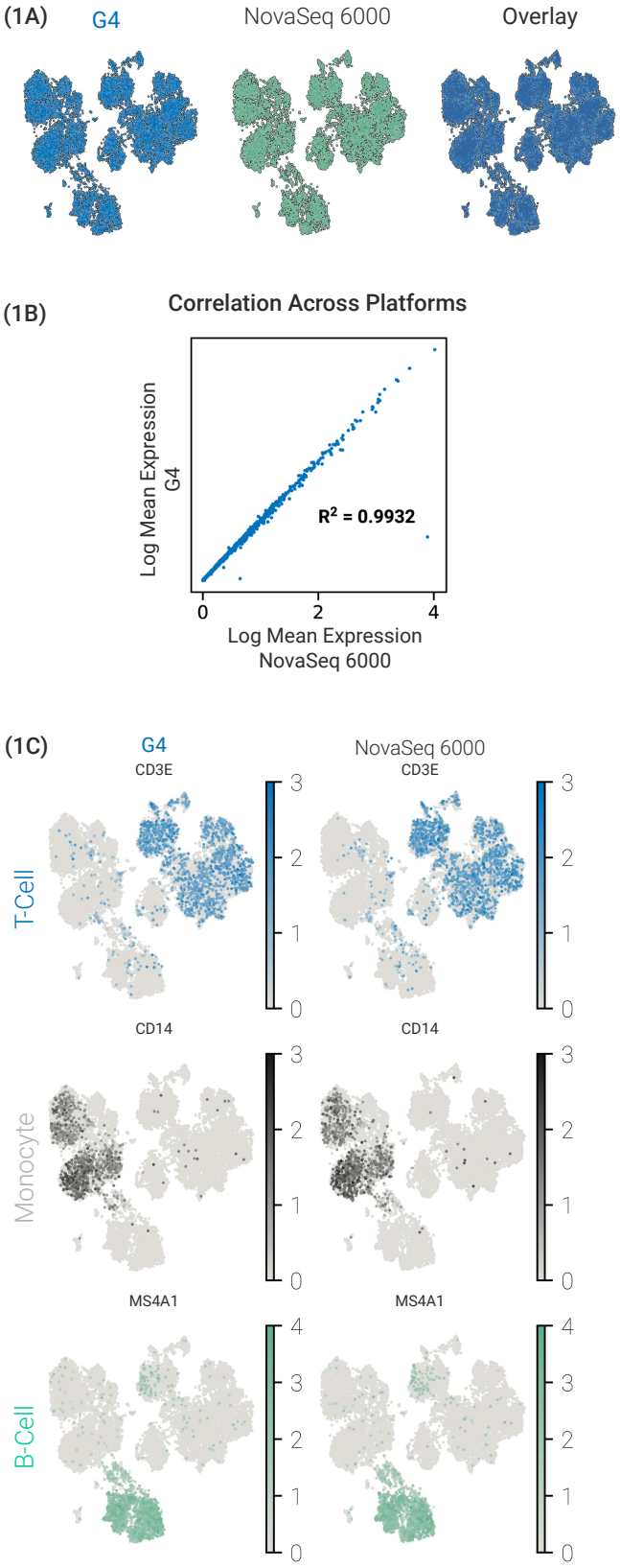

Figure 1 (A) UMAP embedding of single cell gene expression profiles obtained from sequencing of a PBMC library on the G4 and NovaSeq 6000 platforms. **(B)** Spearman's correlation of average gene expression across platforms. **(C)** Expression of lineage-defining markers for T-Cells, Monocytes, and B-Cells in G4 and NovaSeq 6000 datasets overlaid on the UMAP embedding.

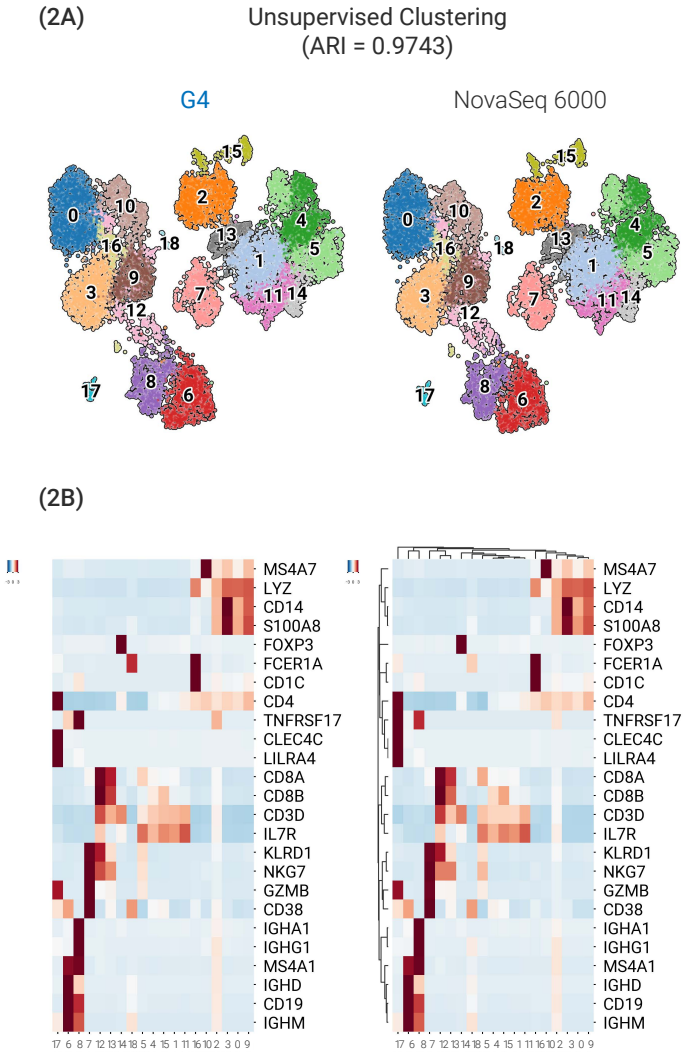

Figure 2 (A) Unsupervised Leiden clustering and for the NovaSeq 6000 and G4 datasets overlaid on the UMAP embedding.
(B) Heatmap of average expression profiles of a panel of well-known PBMC phenotyping markers for each identified cluster/cell type. Cluster maps were generated by performing hierarchical clustering of the NovaSeq 6000 dataset and then applying the same row/column ordering to the G4 dataset. Z-scoring was applied to each column of the cluster map and values were clipped to the range −3 to +3.

To further assess the accuracy of CellTypist labels, we quantified the abundance of major immune cell populations including key T-Cell and B-Cell subsets. We observe nearly identical cell type abundance across the platforms, with cell population frequencies and gene expression profiles matching expectations from healthy donor PBMCs[9] (ARI = 0.9932; **Figure 3A, 3B**).

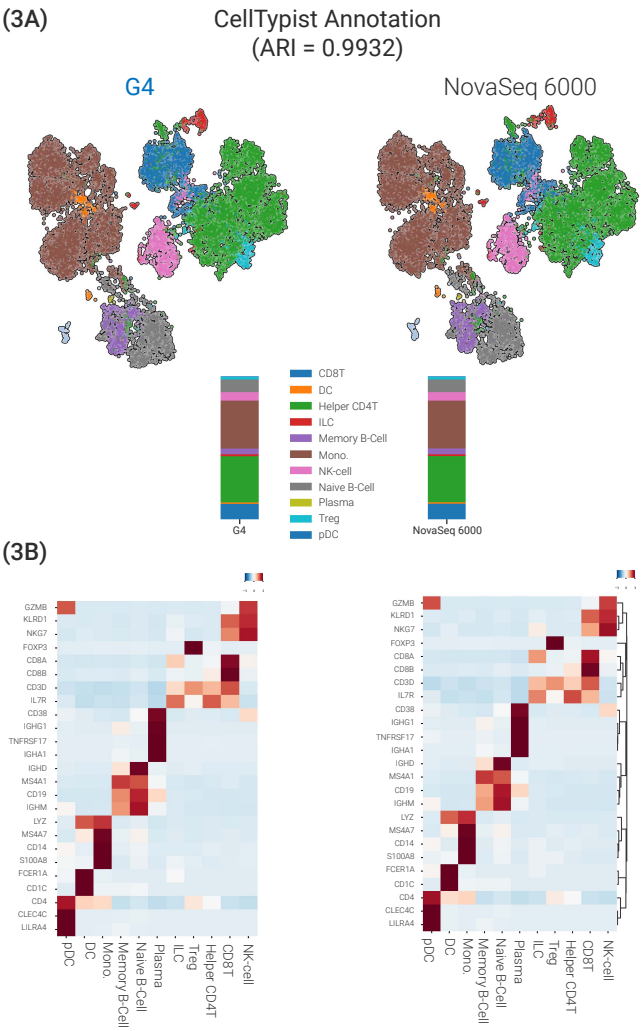

Figure 3 (A) CellTypist annotations for NovaSeq 6000 and G4 datasets overlaid on the UMAP embedding. Annotations were produced using the "Immune_all_low" model with majority voting applied. Stacked bar chart indicates abundance of immune cell types. **(B)** Heatmap indicating gene expression of key cell type-specific markers. Z-scoring was applied to each row of the cluster map and values were clipped to the range –3 to +3.

Finally, we examined pseudo-bulk RNA-seq profiles for T-Cells, Monocytes, and B-Cells by comparing the average gene expression profiles across platforms and log fold changes (logFC; obtained from differential gene expression analysis of each cell types versus all others), again observing highly correlated metrics (Spearman's R^2 > 0.976 and > 0.995, respectively, **Figure 4A, 4B, 4C**).

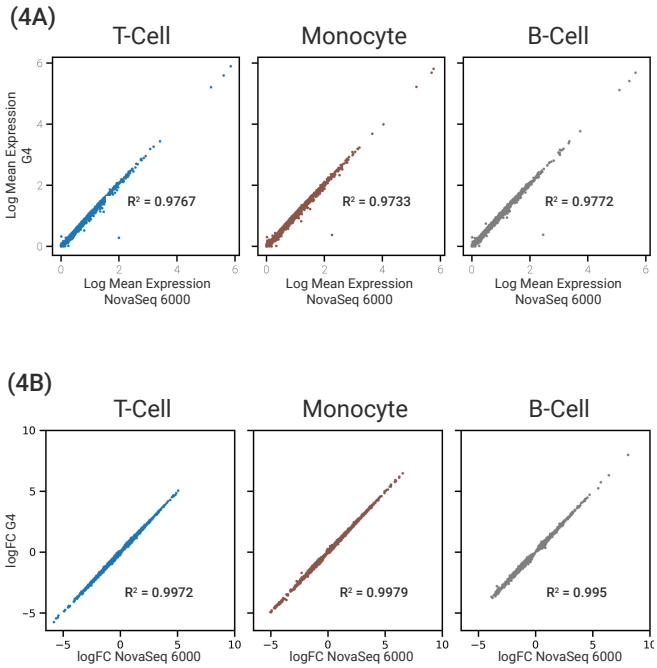

Figure 4 (A) Spearman's correlation of average gene expression for T-Cells, Monocytes, and B-Cells across platforms. Expression is calculated as the library-size normalized, log-transformed UMI counts per gene. **(B)** Pearson's correlation across platforms for log fold changes (logFC) obtained from differential gene expression analysis of T-Cells, Monocytes and B-Cells versus all other cells in the dataset.

Conclusion

scRNA-Seq data generated using the Parse Biosciences Evercode™ WT kit and the G4™ Sequencing Platform demonstrates high accuracy and performance comparable to data derived from the Illumina® NovaSeq 6000 platform. Notably, Singular Genomics G4 Sequencing Platform and Illumina NovaSeq 6000 datasets uncovered nearly identical cell types and unsupervised clusters with comparable gene expression profiles consistent with major PBMC cell types, including prominent T-Cell and B-Cell subpopulations, underscoring the high quality of the sequencing data and the accuracy and sensitivity of the Evercode WT single cell library preparation workflow.

The G4 Sequencing Platform integrates seamlessly with scRNA-Seq workflows while allowing labs to benefit from unique throughput flexibility to minimize batching related delays and optimize the sequencing cost per sample. The combination of the G4 Sequencing Platform with the Parse Evercode WT Kit enables labs to generate fast, flexible, and highly accurate single cell sequencing results.

SINGULAR GENOMICS

Parse BIOSCIENCES

Begin Your Journey with Singular Genomics

Contact us to learn more about the capabilites of the G4 Sequencing Platform

Website: www.singulargenomics.com
Email: care@singulargenomics.com
Call: +1 442-SG-CARES (442-742-2737)
Address: 3010 Science Park Rd, San Diego, CA 92121

Contact Parse Biosciences

Learn more about the Evercode™ WT Kit

Website: www.parsebiosciences.com
Email: info@parsebiosciences.com
Address: 700 Dexter Ave, Suite 600, Seattle, WA 98109

REFERENCES

1. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 9: 1–12 (2017).
2. Sun, G. et al. Single-cell RNA sequencing in cancer: Applications, advances, and emerging challenges. Mol. Ther. - Oncolytics 21: 183–206 (2021).
3. Wolf, F.A., Angerer, P., and Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology 19: 15 (2018).
4. Wolock, S.L., Lopez, R., and Klein, A.M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Systems 8: 281-291 (2019).
5. Stuart, T., Butler, A., et al. Comprehensive Integration of Single-Cell Data. Cell 177(7): 1888-1902 (2019).
6. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. Nature Methods 15(12): 10533-1058 (2018).
7. Gayoso, A., Lopez, R., et al. A Python library for probabilistic analysis of single-cell omics data. Nature Biotechnology 40: 163-166 (2022).