illumina®

QC and rebalancing of single-cell gene expression libraries using the iSeq™ 100 System

Assess key metrics of multiplexed single-cell libraries to evaluate quality and enable accurate rebalancing before high-depth runs on the NextSeq™ 550 or NovaSeq™ 6000 Systems.

Introduction

Understanding the biological contribution of individual single cells to complex, multicellular tissues and processes is rapidly becoming a leading application for next-generation sequencing (NGS) technology. Data from single-cell sequencing experiments is driving new discoveries in diverse fields such as oncology, immunology, developmental biology, and many others. NGS provides high discovery power with the ability to profile thousands to tens of thousands of individual cells in parallel in a single high-throughput sequencing run. Depending on the single-cell assay being used, the sample type, and the experimental design, the sequencing requirements of any given single-cell library can vary greatly.[1]

Accurate characterization of multiplexed library composition is a vital step in optimizing library loading, particularly for single-cell gene expression profiling (sc-GEX) libraries. Important metrics of sc-GEX libraries, such as estimated cell counts, intergenic/intronic/exonic content, and fraction of reads in cells cannot be assessed before sequencing using traditional quantitation methods like Qubit or qPCR. To maximize the efficiency of high-throughput sc-GEX experiments, sequencing these libraries first at shallow depths enables characterization of key metrics and subsequent rebalancing before a high-depth NGS run. Library quality control (QC) not only saves time and money, but also leads to more consistent sequencing results, which can simplify data analysis and interpretation.

This application note demonstrates use of the iSeq 100 System (Cat no. 20021532) to perform sc-GEX library* QC and rebalancing before full-scale, high-depth sequencing runs on either the NextSeq 550 System (Cat no. SY-415-1002) or NovaSeq 6000 System (Cat no. 20012850) (Figure 1). As part of this application note, Illumina collaborated with 10x Genomics Inc. for cell isolation and library prep.

Library QC with the iSeq 100 System

Using the iSeq 100 System and a simple, streamlined workflow, multiple sc-GEX libraries can be pooled and sequenced to shallow depths and analyzed for important sc-GEX library metrics. These metrics provide valuable information on the libraries in the pool. First, they act as a screen of library generation success, enabling elimination of libraries with low cell yield or content. Secondly, these metrics can be used to inform library rebalancing to normalize per index or per cell sequencing depths across several sc-GEX libraries. Proper balancing of pooled sc-GEX libraries allows for more uniformity across reads per index or reads per cell. Finally, iSeq 100 sequencing results can be used as a prediction of certain sc-GEX library metrics when sequencing on the NextSeq 550 or NovaSeq 6000 Systems.

Methods

Library preparation and pooling

Sixteen individual sc-GEX libraries were prepared from ~1 K peripheral blood mononuclear cells (PBMCs) and eight sc-GEX libraries were prepared from ~5 K PBMCs using the Chromium Single Cell Gene Expression v3 Solution on the 10x Genomics Chromium Controller (Cat no. 120223) (Figure 2). Library concentrations were determined with the Agilent 2100 Bioanalyzer Instrument (Cat no. G2939BA)

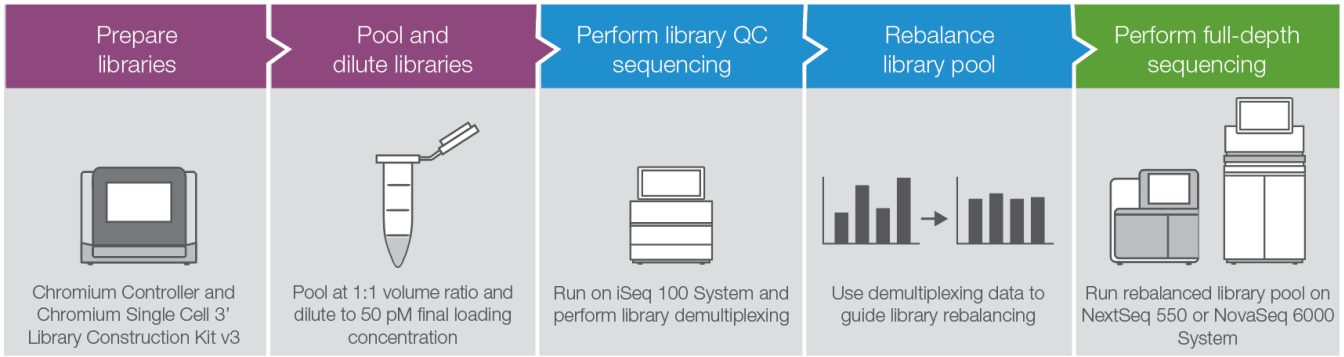

Figure 1: Single-cell library QC with the iSeq 100 System—The iSeq 100 System is part of a fast and simple workflow for performing quality control on sc-GEX libraries prior to high-depth sequencing.

* While there are numerous emerging single cell profiling applications that use NGS (eg, sc-ATAC, sc-CNV, sc-immune profiling) this application note focuses on sc-GEX. For best practices in applying the techniques described here for other library types, please reach out to your Illumina Field Applications Scientist or Illumina Technical Support.

(quantification via qPCR is a viable alternative) and then diluted to 2 nM each. The 16 ~1 K PBMC sc-GEX libraries were pooled in a 1:1 volume ratio into four distinct subpools of four samples each and diluted to 50 pM final loading concentration. Additionally the 16 ~1K PBMC sc-GEX libraries were pooled in a 1:1 volume ratio into one pool and diluted to 50 pM final loading concentration. The eight ~5 K PBMC libraries were pooled in a 1:1 volume ratio and diluted to 50 pM final loading concentration.

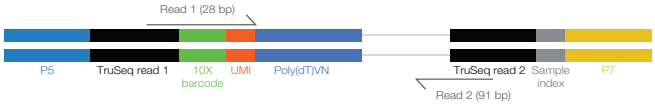

Figure 2: Schematic of Chromium Single Cell 3' Gene Expression Library— Read 1 is 28 bp long and encodes the 10x Genomics cellular barcode as well as a Unique Molecular Identifier (UMI). Index 1 is 8 bp and allows for multiplexing of several 10x libraries on the same sequencing run. Read 2 is 91 bp long and contains the insert derived from the expressed RNA. While data presented in this application note are based on Single Cell Gene Expression v3 libraries, 10x Genomics do expect comparable performance with Single Cell Gene Expression v3.1 (Next GEM).

Sequencing

All sequencing runs described in this application note are publicly available in BaseSpace™ Sequence Hub (see Appendix for details). Each ~1 K PBMC subpool was run separately on the iSeq 100 System (Software v1.4) targeting 1M reads per library as higher depth QC (Runs 1–4). The 16 combined ~1 K PBMC libraries and the eight pooled ~5K PBMC libraries were run separately on the iSeq 100 System (Runs 5, 7, and 9). Rebalanced library pools were run on the NextSeq 550 System (RTA Version 2.4.11) using the NextSeq 500/550 High Output Kit v2.5 (150 Cycles, Cat no. 20024907, Runs 6 and 8) or on the NovaSeq 6000 System (RTA Version 3.4.4) using the NovaSeq S2 Reagent Kit (200 cycles, Cat no. 20012861, Run 10).

Data analysis

The Cellranger software package (v3.0.2, 10x Genomics Inc.), mkfastq command was used to demultiplex and convert BCL to FASTQ files. After FASTQ generation, demultiplexed reads from iSeq 100, NextSeq 550, and NovaSeq 6000 runs were subsampled using the software tool at seqtk[2] at various read depths. Read depths for the iSeq 100 System ranged from ~1.5 K reads per cell down to 5 reads per cell, for the NextSeq 550 System from ~25 K down to 5 reads per cell, and for the NovaSeq 6000 System from ~40 K down to 5 reads per cell. The Cellranger count command was used, specifying a reference transcriptome of refdata-cellranger-GRCh38-3.0.0 and default cell count estimate of 3000 to generate single cell feature counts.

Results

Library rebalancing based on index representation

Sequencing results from the combined pool of all 16 ~1 K PBMC library pools run on the iSeq 100 System (Run 5) showed > 95% of bases ≥ Q30 (Figure 3). Demultiplexed library representation values obtained from the combined run on the iSeq 100 System were used to rebalance the 16 ~1 K PBMC libraries in the pool. First, normalized index representations were calculated for each library by dividing the reads passing filter (PF) for each index by the sum of reads PF across all 16 indexes. Second, loading factors were calculated for each library as the ratio between the highest normalized index representation across all samples and the index representation for the current library (Table 1).

$$Norm.\%\ Index_i = \frac{reads\ PF_i}{\sum_{i=1}^{n} reads\ PF_n}$$

$$Loading\ Factor_i = \frac{max(Norm.\%\ Index[n])}{Norm.\%\ Index_i}$$

For each library, multiplying the original input volume by its calculated loading factor provides the new volume for index rebalanced pooling prior to high-depth sequencing on the NextSeq 550 System. The index rebalanced pool was initially sequenced at 1.8 pM final loading concentration, per Bioanalyzer quantification; however, in an effort to optimize cluster density, the concentration was increased to 3.6 pM. Users should adhere to the 1.8 pM final loading concentration, as recommended by 10x Genomics and adjust accordingly if desired. Sequencing results from the run with 3.6 pM showed > 90% of bases ≥ Q30 (Run 6, Figure 4). Comparing the coefficient of variation (CV) of index representation across the 16 ~1 K PBMC libraries before and after rebalancing showed a decrease from 10.41% to 3.65% (Figure 5).

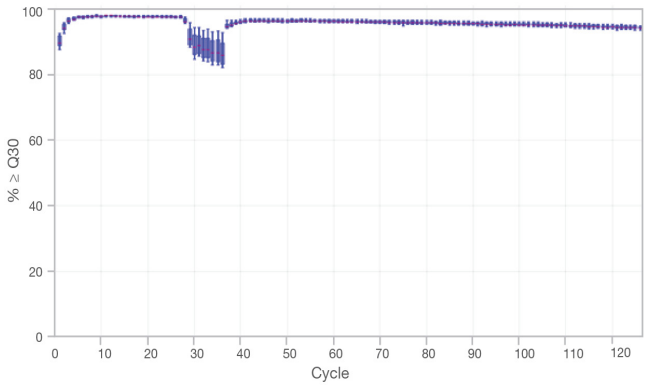

Figure 3: Q30 plot from run 5 on the iSeq 100 System—Sequencing results from the combined pool of all 16 ~1 K PBMC libraries run on the iSeq 100 System showed > 95% of bases ≥ Q30.

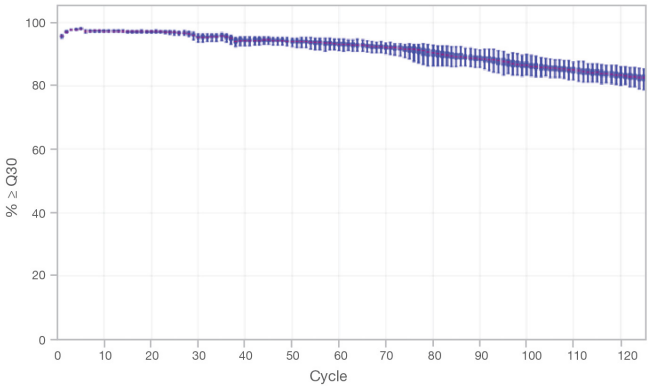

Figure 4: Q30 plot from run 6 on the NextSeq 550 System—Sequencing results from the index rebalanced pool of all 16 ~1 K PBMC libraries run on the NextSeq 550 System showed > 90% of bases ≥ Q30.

Table 1: Rebalancing ~1 K PBMC sc-GEX libraries based on iSeq 100 System index representation

Library Name	Reads PF	Normalized % index	Loading factor
S1_W1_R1	264,639	5.51%	1.43
S1_W1_R2	310,292	6.46%	1.22
S1_W1_R3	289,073	6.02%	1.31
S1_W1_R4	252,327	5.25%	1.50
S1_W2_R1	295,547	6.15%	1.28
S1_W2_R2	284,929	5.93%	1.33
S1_W2_R3	293,135	6.10%	1.29
S1_W2_R4	299,982	6.25%	1.26
S2_W1_R1	379,568	**7.90%**	1.00
S2_W1_R2	359,625	7.49%	1.05
S2_W1_R3	303,996	6.33%	1.25
S2_W1_R4	310,468	6.47%	1.22
S2_W2_R1	281,004	5.85%	1.35
S2_W2_R2	272,773	5.68%	1.39
S2_W2_R3	317,334	6.61%	1.19
S2_W2_R4	287,337	5.98%	1.32

Highest normalized index representation is indicated in red.

Library rebalancing based on reads per cell

An alternative method of rebalancing can be used to achieve a more even distribution of mean reads per cell, rather than reads per index. To demonstrate, a second library pool was rebalanced for mean reads per cell output based on Cellranger analysis taken from a second combined 16 ~1 K PBMC iSeq 100 run (Run 7). For each library, the loading factor was calculated as the ratio between the highest mean reads per cell across all samples and the mean reads per cell for the current library (Table 2).

$$Loading\ Factor_i = \frac{max(Mean\ Reads\ per\ Cell[n])}{Mean\ Reads\ per\ Cell_i}$$

Multiplying the original input volume by the calculated loading factor for each library provides the new volumes for mean reads per cell rebalanced pooling prior to sequencing on the NextSeq 550 System.

The library pool rebalanced for mean reads per cell was also diluted to 3.6 pM final loading concentration and run on the NextSeq 550 System (Run 8). Comparing the CV of mean reads per cell across the 16 ~1 K PBMC libraries before and after rebalancing showed a decrease from 13.87% to 5.43% (Figure 6).

Library rebalancing of more complex sc-GEX libraries

The same method was tested on more complex PBMC libraries. Eight ~5 K PBMC libraries were pooled and run on the iSeq 100 System, as described (Run 9). This run was used to rebalance the libraries for either index representation (Table 3) or mean reads per cell (Table 4). Each rebalanced ~5 K PBMC sc-GEX pool was diluted to 133 pM and loaded into separate lanes of an S2 flow cell using the NovaSeq XP loading protocol, and run on the NovaSeq 6000 System (Run 10). Sequencing results showed > 90% of bases ≥ Q30 (Figure 7). Comparing the CVs across the eight libraries before and after rebalancing showed decreases from 15.01% to 6.79% for index representation (Figure 8A) and 15.56% to 8.86% for mean reads per cell (Figure 8B).

Table 2: Rebalancing ~ 1 K PBMC sc-GEX libraries for mean reads per cell from iSeq 100 System run

Library name	Reads PF	Estimated no. of cells	Mean reads per cell	Loading factor
S1_W1_R1	218,691	841	260	1.64
S1_W1_R2	256,639	733	350	1.22
S1_W1_R3	248,325	707	351	1.21
S1_W1_R4	284,751	729	390	1.09
S1_W2_R1	265,240	857	309	1.38
S1_W2_R2	200,026	757	264	1.61
S1_W2_R3	266,009	800	332	1.28
S1_W2_R4	246,575	874	282	1.51
S2_W1_R1	242,390	846	286	1.49
S2_W1_R2	248,237	739	335	1.27
S2_W1_R3	204,978	715	286	1.49
S2_W1_R4	262,200	731	358	1.19
S2_W2_R1	275,822	854	322	1.32
S2_W2_R2	330,183	774	**426**	1.00
S2_W2_R3	262,099	807	324	1.31
S2_W2_R4	242,838	867	280	1.52

Highest mean reads per cell is indicated in red.

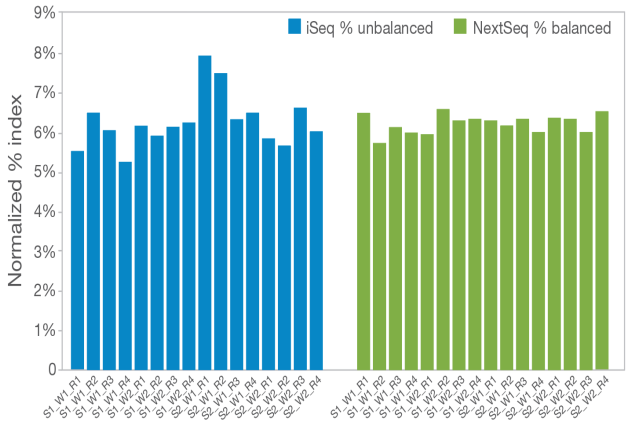

Figure 5: Index rebalancing of ~1 K PBMC sc-GEX libraries—The pool of 16 libraries run on the iSeq 100 System (Run 5) produced a CV of index representation of 10.41% (mean=6.25%, SD=0.65%). After they were index rebalanced and run on the NextSeq 550 System (Run 6), the CV was 3.65% (mean=6.25%, SD=0.23%).

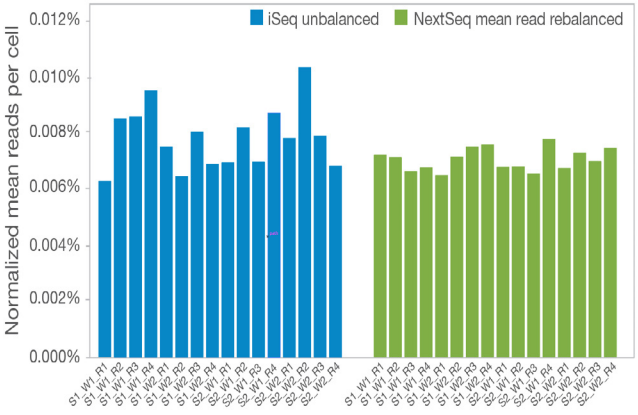

Figure 6: Mean reads per cell rebalancing of ~1 K PBMC sc-GEX libraries—The pool of 16 libraries run on the iSeq 100 System (Run 7) produced a CV of normalized mean reads per cell of 13.87% (mean=7.95E-3%, SD=1.1E-3%). Rebalanced libraries run on the NextSeq 550 System (Run 8) had a CV of 5.43% (mean=7.19E-3%, SD=3.91E-4%).

Table 3: Rebalancing ~ 5 K PBMC sc-GEX libraries based on iSeq 100 System index representation

Library name	Reads PF	Normalized % index	Loading factor
T1_W1_R1	554,168	10.69%	1.54
T1_W1_R2	683,013	13.18%	1.25
T1_W1_R3	853,552	**16.47%**	1.00
T1_W1_R4	625,127	12.06%	1.37
T1_W2_R1	622,075	12.00%	1.37
T1_W2_R2	528,790	10.20%	1.61
T1_W2_R3	596,223	11.50%	1.43
T1_W2_R4	720,780	13.90%	1.18

Highest normalized index representation is indicated in red.

Table 4: Rebalancing ~ 5 K PBMC sc-GEX libraries for mean reads per cell from iSeq 100 System run

Library name	Reads PF	Estimated no. of cells	Mean reads per cell	Loading factor
T1_W1_R1	554,168	4545	121	1.64
T1_W1_R2	683,013	4681	145	1.37
T1_W1_R3	853,552	4286	**199**	1.00
T1_W1_R4	625,127	4194	149	1.34
T1_W2_R1	622,075	4132	150	1.33
T1_W2_R2	528,790	4206	125	1.59
T1_W2_R3	596,223	4148	143	1.39
T1_W2_R4	720,780	4203	171	1.16

Highest mean reads per cell is indicated in red.

Sc-GEX library secondary analysis

Estimated cell counts

Additional Cellranger analysis was conducted without specifying a library cell complexity constraint or estimate to assess the ability of shallow sequencing to predict sc-GEX library metrics. Across the 16 ~1 K PBMC sc-GEX libraries there was a ~7.8-8.9% mean variation between cell counts estimated by the iSeq 100 System (Run 5) with ~20 reads per cell or more and the NextSeq 550 System (Run 6) at ~25 K reads per cell (Figure 9). Across the eight ~5 K PBMC sc-GEX libraries there was a ~21-23% mean variation between cell counts estimated by the iSeq 100 System (Run 9) with ~20 reads per cell or more and the NovaSeq 6000 System (Run 10) at ~40 K reads per cell (Figure 10). Although absolute estimates of cell counts showed this high mean variation between shallow and high-depth sequencing, the relative amounts within a given subsampling held steady from ~20 reads per cell to ~40 K reads per cell.

Comparing cell count estimates across the range of mean reads per cell (Figures 9 and 10) reveals three regions with distinct response to read depth. At extreme low depths below 20 reads per cell, cell count estimates are unstable as every barcode found is identified as a cell (Barcode Rank plot indicates only cells in dark blue, no background). Above 20 reads per cell the cell calling algorithms begin to distinguish cells from background (Barcode Rank plot indicates both cells in dark blue and background in grey). Finally, as read depths approach the full

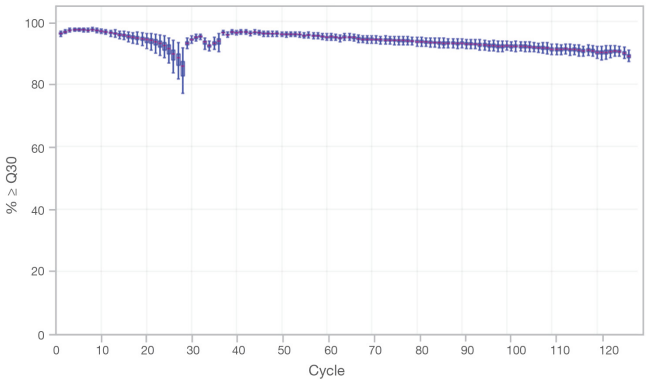

Figure 7: Q30 plot from run 10 on the NovaSeq 6000 System—Sequencing results from the rebalanced pool of all eight ~5 K PBMC libraries run on the NovaSeq 6000 System showed > 90% of bases ≥ Q30.

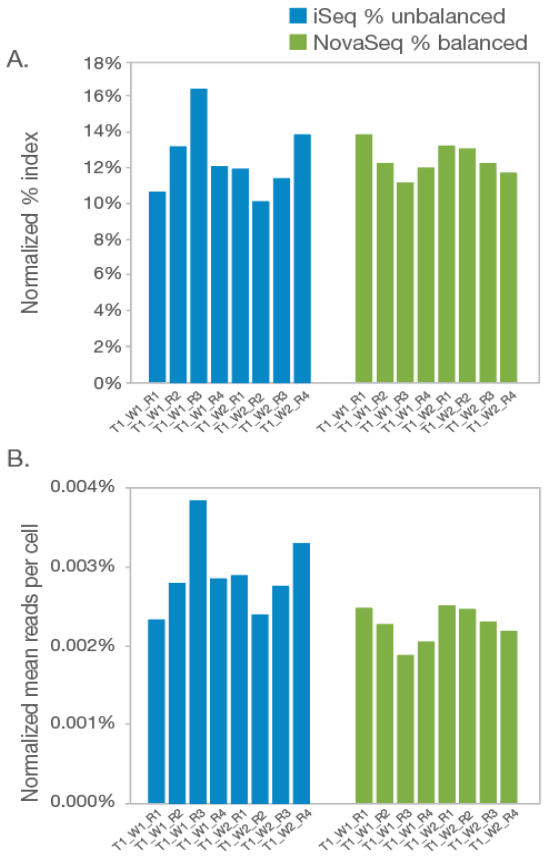

Figure 8: Rebalancing of ~5 K PBMC sc-GEX libraries—The pool of eight libraries run on the iSeq 100 System (Run 9) produced (A) a CV of index representation of 15.01% (mean=12.5%, SD=1.88%) and (B) a normalized mean reads per cell CV of 15.56% (mean=2.9E-3%, SD=4.51E-4%). Rebalanced libraries run on the NovaSeq 6000 System (Run 10) had CVs of 6.79% (mean=12.5%, SD=0.85%) and 8.86% (mean=2.3E-3%, SD=2.04E-4%), respectively. Note: the normalized mean reads per cell on the NovaSeq 6000 System generally decreased as more cells were identified due to higher depth sequencing.

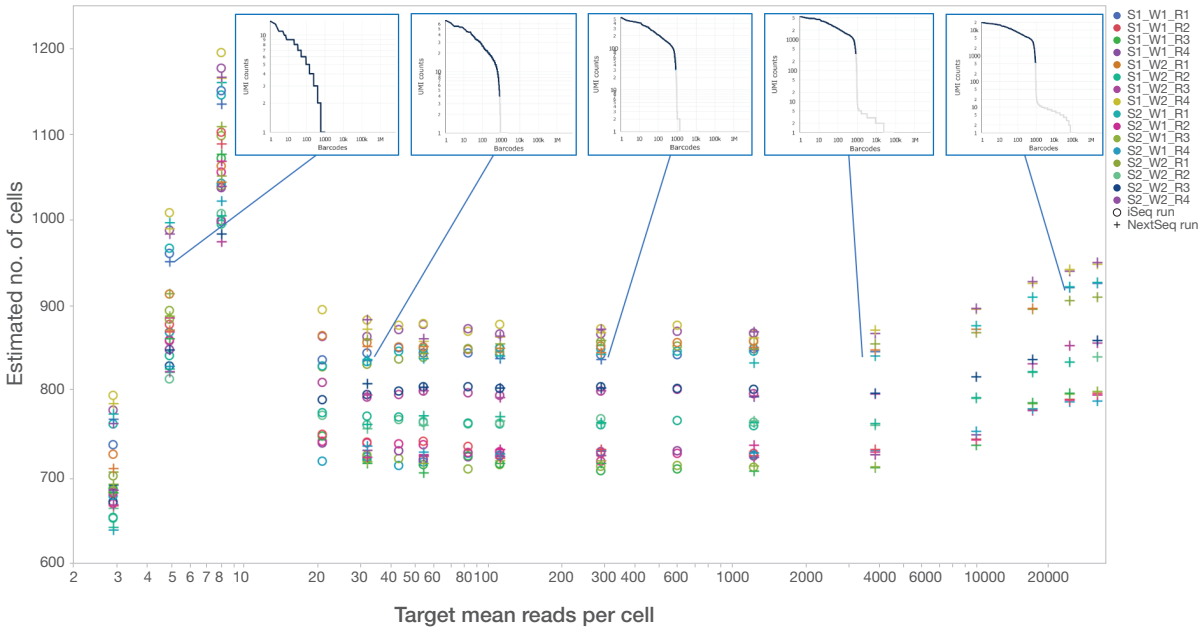

Figure 9: ~1 K PBMC sc-GEX library cell count estimates—Plotting cell count estimates for each library taken from full run and subsampled data from the iSeq 100 (Run 5) and NextSeq 550 System (Run 6). Comparing cell count estimates at ≥ 20 reads per cell from subsampled iSeq 100 and NextSeq 550 runs to a full depth NextSeq 550 runs shows a mean variation of 7.8-8.9%. Barcode Rank Plots for S1_W1_R1 at select read depths shown to illustrate plots response to read depth. Barcode labeled cell are denoted by dark blue line, with background shown in grey.

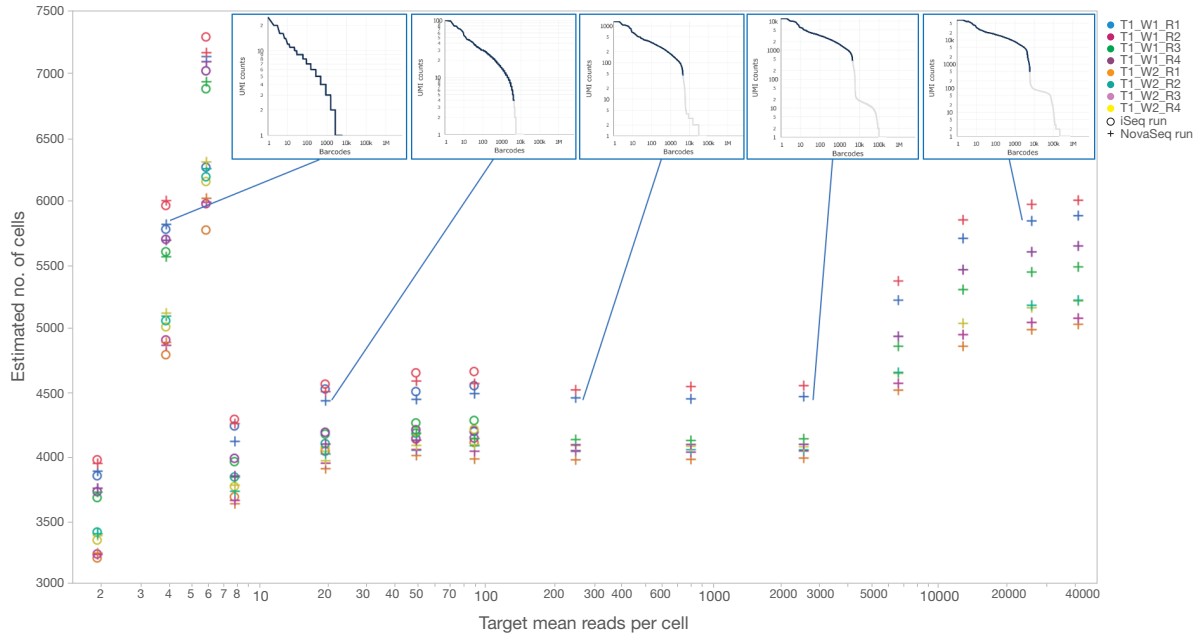

Figure 10: ~5 K PBMC sc-GEX library cell count estimates—Plotting cell count estimates for each library taken from full run and subsampled data from the iSeq 100 (Run 9) and NovaSeq 6000 System (Run 10). Comparing cell count estimates at ≥ 20 reads per cell from subsampled iSeq 100 and NovaSeq 6000 runs to a full depth NovaSeq 6000 run shows a mean variation of 21-23%. Barcode Rank Plots for T1_W1_R1 at select read depths shown to illustrate plots response to read depth. Barcode labeled cell are denoted by dark blue line, with background shown in grey.

run depth, cell count estimates increase asymptotically as the empty drops method identifies barcodes below the UMI cutoff as cells based on their RNA profile (Barcode Rank plot shows cells in dark blue extending to lower UMI count). Given the accuracy of cell count estimates across these three regions, we recommend checking the Barcode Rank plot generated to confirm that a background population is identified. The 20 reads per cell threshold would indicate a theoretical max of 200 K cells across several sc-GEX libraries could be assayed on the iSeq 100 System in a single run. Given that the mean reads per cell across the sc-GEX libraries are not balanced before running on the iSeq 100 System, it is recommended that the upper limit be set at 50 K cells total amongst all sc-GEX libraries run together on the iSeq 100 System.

Please note that the distribution of reads per cell might vary based on the sample type, sample quality and library complexity. While we outline recommendations based on human PBMCs, additional sequencing and/or other quality controls may be required for other sample types. Specifically, examining the expected "knee" of the Barcode Rank plot can support findings from low sequencing QC experiments.

Cell barcode representation across sequencing depths and platforms

To confirm the same cell barcodes are being sequenced in the same relative proportions at different depths and across platforms, correlation plots were created to compare the UMI counts per cell barcode at full depth between the iSeq 100 (Run 5) and NextSeq 550 Systems (Run 6) (Figure 11) and the iSeq 100 (Run 9) and NovaSeq 6000 Systems (Run 10) (Figure 12).

Predictive library metrics

Cellranger software outputs several metrics beyond estimated cell counts when analyzing sc-GEX libraries. These metrics respond at different rates when downsampling to shallow sequencing depths. Some sc-GEX library metrics remain flat across the range of depths sampled, such as percent reads mapped to genomic, intergenic, intronic, and exonic regions (Figure 13). Such stability indicates that even at shallow sequencing depths these specific metrics are still predictive of higher depth run metrics for the same libraries†.

Other sc-GEX library metrics respond in a more linear fashion to sequencing depth before saturating at higher depths, such as sequencing saturation, median genes per cell, and median UMI counts per cell (Figure 14). Such linear dependence at shallow depths indicates these metrics may not be predictive of higher depth run metrics for the same libraries and should be evaluated with care.

Metrics such as fraction of reads in cell respond to sequencing depth similar to estimated cell counts, with relative stability and good performance above 20 reads per cell (Figure 15A). Finally, Total Genes detected exhibits high sensitivity to reads per cell at shallow depths, and flattens out at higher depth (Figure 15B). Total genes detected at shallow depths may not be predictive of higher depth run metrics for the same libraries and should be evaluated with care.

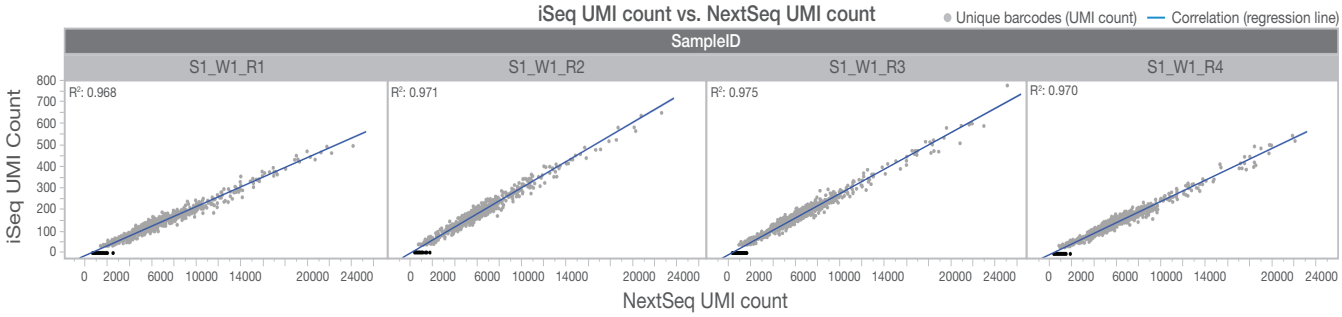

Figure 11: UMI counts per cell from iSeq 100 Run 5 and NextSeq 550 Run 6—Plotting samples S1_W1_R1 through S1_W1_R4 show high R^2 values (> 0.96) between sequencing systems. Cells observed on the NextSeq 550 but not on the iSeq 100 System (black) are all among lowest abundance for a full NextSeq 550 run.

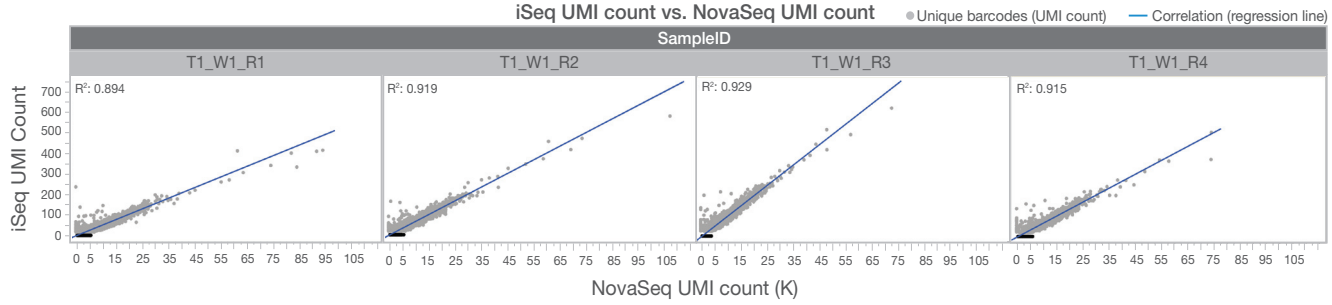

Figure 12: UMI counts per cell from iSeq 100 Run 9 and NovaSeq 6000 Run 10—Plotting samples T1_W1_R1 through T1_W1_R4 show high R^2 values (> 0.89) between sequencing systems. Cells observed on the NovaSeq 6000 but not on the iSeq 100 System (black) are all among lowest abundance for a full NovaSeq 6000 run.

† All remaining graphs (Figures 13–15) show sample S1_W1_R1 run on iSeq 100 and NextSeq 550 Systems; however, trends observed remain the same across all samples and platform pairings (data not shown).

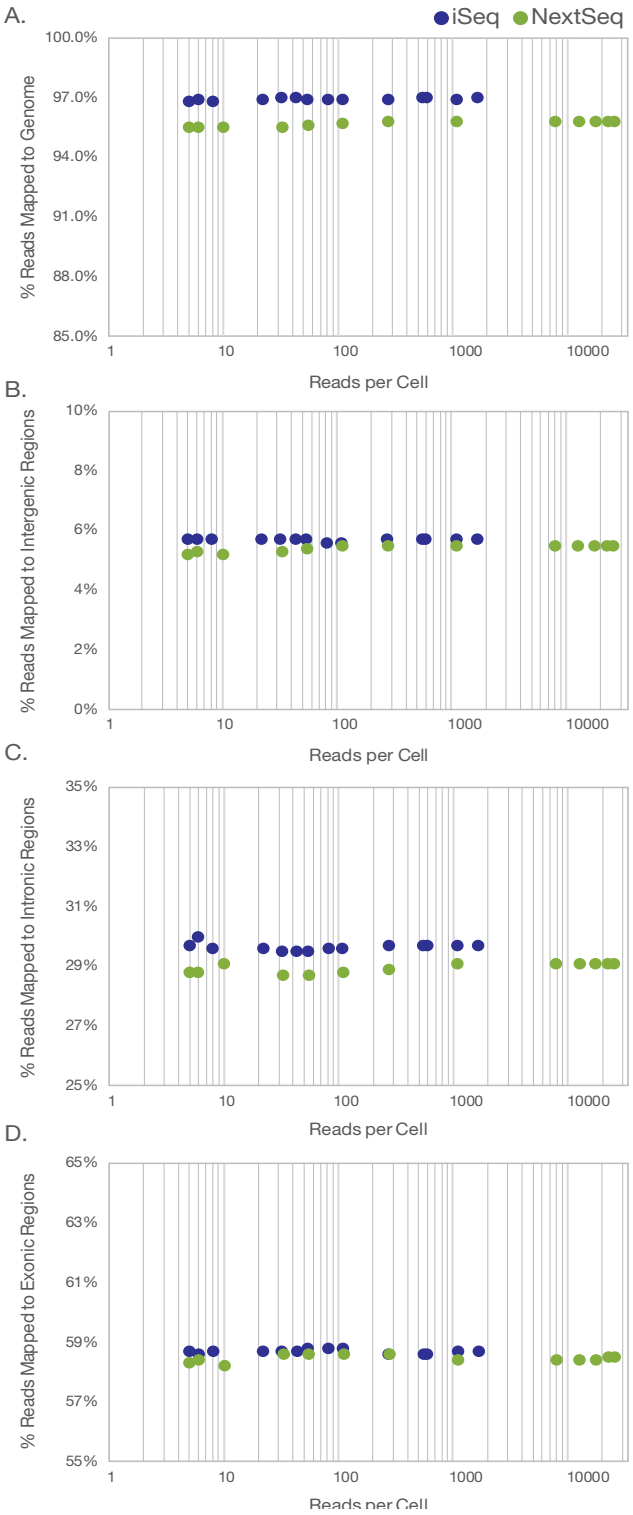

Figure 13: sc-GEX library metrics stable across sequencing depths—Percent reads mapped to (A) genome, (B) intergenic regions, (C) intronic regions, and (D) exonic regions.

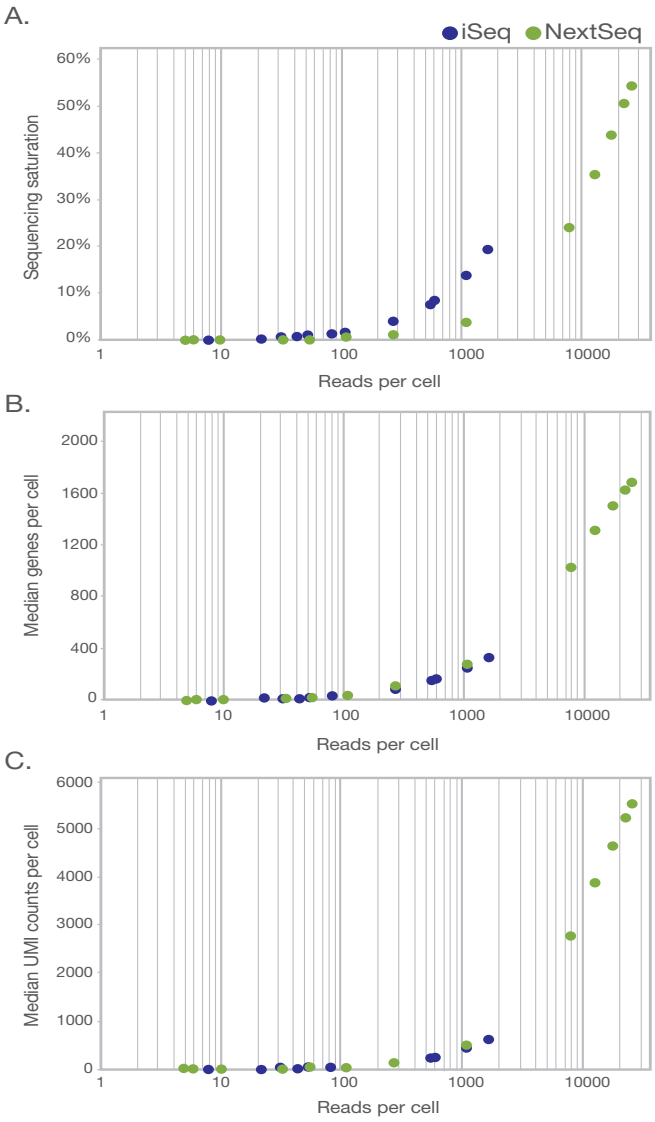

Figure 14: sc-GEX library metrics with linear dependence to sequencing depth—(A) sequencing saturation, (B) median genes per cell, and (C) median UMI counts per cell.

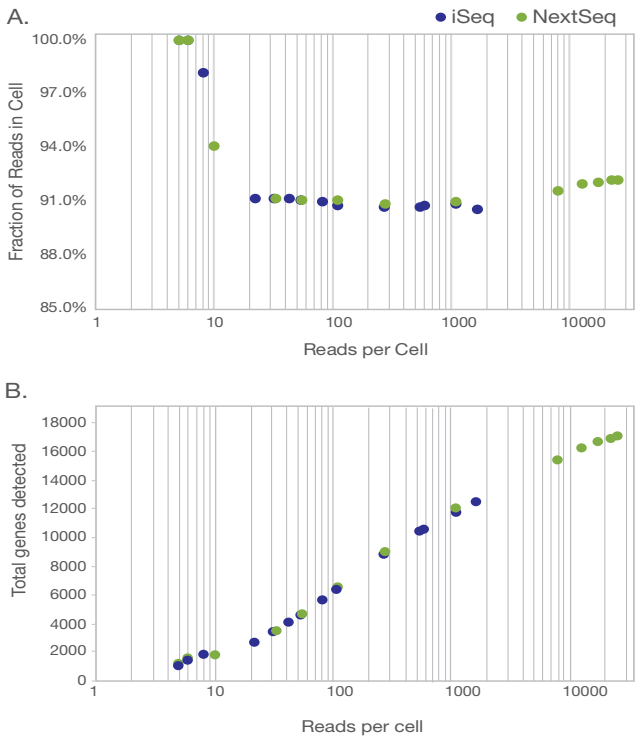

Figure 15: Fraction reads in cell and total genes detected—(A) Fraction of reads in cell responds to read depth similarly to cell estimates, while (B) total genes detected exhibits high sensitivity at shallow depths but flattens out at higher depths.

Summary

The iSeq 100 System enables library pool rebalancing before high-depth multiplexed NGS runs. Rebalancing strategies can either target uniformity for total reads across indexed samples, or uniformity for mean reads per cell across cells within those samples. For sc-GEX libraries, low-depth sequencing on the iSeq 100 System can predict certain library metrics before moving to higher depth sequencing. To explore whether low-depth sequencing on the iSeq 100 System can be applied to other single-cell sequencing assays, and remain predictive for a given sc-library metric of interest, subsampling data from a previous high-depth sequencing run and analyzing with Cellranger software to create response curves as described in this application note (Figures 13–15) is recommended. Given the amount of sample multiplexing and expected cells per sample, an expected read depth on the iSeq 100 System can be determined and evaluated on the previously generated response curves to determine the ability of the iSeq 100 System to predict sc-library metrics for a given pooled set of sc-libraries.

Learn more

To learn more about Illumina sequencing systems, visit www.illumina.com/systems/sequencing-platforms.html

References

1. Satijalab.org. Accessed December 5, 2019.
2. github.com/lh3/seqtk. Accessed December 5, 2019.

Appendix

Run no.	Run name	Link
Run 1	4-plex_iSeq_1k-PBMC-Unbal_subpool1	basespace.illumina.com/s/nyCmXI8LacEK
Run 2	4-plex_iSeq_1k-PBMC-Unbal_subpool2	basespace.illumina.com/s/Hn6TazAxiHKB
Run 3	4-plex_iSeq_1k-PBMC-Unbal_subpool3	basespace.illumina.com/s/7sW98OwhVcHk
Run 4	4-plex_iSeq_1k-PBMC-Unbal_subpool4	basespace.illumina.com/s/D5ulA5nGlPrl
Run 5	16-plex-iSeq_1k-PBMC-Unbal_pool1	basespace.illumina.com/s/HPuoSEsrTv0k
Run 6	16-plex_NextSeq_1k-PBMC-Index-Bal	basespace.illumina.com/s/MUWFoQpDJip7
Run 7	16-plex_iSeq_1k-PBMC-Unbal_pool2	basespace.illumina.com/s/r1zw0mUqABtL
Run 8	16-plex_NextSeq_1k-PBMC_MRPC-Bal	basespace.illumina.com/s/qkChF0pogWPf
Run 9	8-plex_iSeq_5k-PBMC-Unbal	basespace.illumina.com/s/tyxPesQs00g6
Run 10	8-plex_NovaSeq_5k-PBMC_Lane1-Index-Bal_Lane2-MRPC-Bal	basespace.illumina.com/s/fbRPbl3Y59Yf

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®