TECHNICAL NOTE

Sequencing Metrics & Base Composition of Sequencing Reads of Chromium Single Cell DNA Libraries

Introduction

The Chromium Single Cell DNA Reagent Kits workflow produces Single Cell DNA libraries for copy number variation (CNV) detection. This Technical Note presents a comparison of sequencing metrics for a control Single Cell DNA library across Illumina platforms, and describes expected base percentage profiles and Phred quality scores. This document is intended to provide general guidance on the expected range of sequencing metrics on multiple platforms based on a control library. Individual results may vary depending on the specific sequencing instrument and/or particular sample and loading characteristics.

Method

A Single Cell DNA library was generated from the diploid human cell line, BJ (ATCC CRL-2522), with a 10% cell spike-in of the gastric cancer cell line MKN-45 following the Single Cell DNA Reagent Kits User Guide (Document CG000153). A total of 500 cells were targeted. The library fragments contain a genomic insert and 10x Barcode, along with standard Illumina paired-end construct sequences, which begin and end with P5 and P7 (Figure 1). The library was quantified using a KAPA DNA Quantification Kit and sequenced with 1% PhiX on various Illumina sequencers to evaluate performance. The following platforms were tested in the noted layout:

- NextSeq 500 - one flowcell
- HiSeq 2500 (RR) - one flowcell
- HiSeq 4000 - one lane
- HiSeq X - one lane
- NovaSeq - one S4 flowcell (pooled with three other Single Cell DNA libraries)

The sequencing configuration for the platforms tested is noted in Table 1.

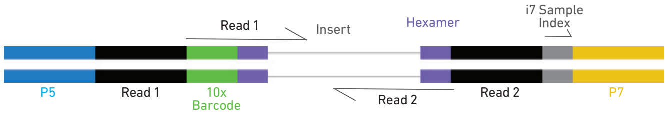

Figure 1. Schematic of final library construct from a Single Cell DNA library. The 10x Barcode (16 bp) and the genomic insert are sequenced in Read 1, the sample index (8 bp) is sequenced on the i7 index read, and the opposite end of the genomic fragment is captured in Read 2. The genomic insert ends include the hexamer sequence used for genome amplification.

Table 1. Sequencing configuration.

Sequencing Read	Number of Cycles
Read 1	100 or 150 cycles[1]
i7 Index	8 cycles
i5 Index	0 cycles
Read 2	100 or 150 cycles[1]

[1]For all sequencing platforms except HiSeq X, 100 cycles were sequenced for Read 1 and Read 2. For HiSeq X, 150 cycles were used, which is the supported Illumina read length for this platform.

Data were analyzed using the Cell Ranger DNA pipeline (1.0), which processes the sequencing data to align reads, detect cells, and identify CNVs.

Results

To assess performance across Illumina platforms, metrics for the same Single Cell DNA library run on different sequencers are shown in Table 2. The loading concentration and cluster density for each sequencer is noted. The sequencing yield for each run correlates with expected yield for each sequencing platform.

Q30 quality scores: Overall, Q30 quality scores were higher for Read 1 and the sample index read (i7) compared to Read 2 on all sequencing platforms (Table 2). Q30 quality scores were marginally lower on NextSeq 500, although the reduced quality had little impact on downstream performance (see section on Median effective reads per Mb).

Fraction of mapped de-duplicated reads: As part of the pre-processing steps in the Cell Ranger DNA pipeline, paired-end reads mapping to the same genomic region and containing the same 10x Barcode are considered duplicates. Duplicates can arise either at the Sample Index PCR step during library construction or as part of the amplification that occurs on the sequencing flowcell. Cell Ranger DNA reports this information in the metric 'Fraction of mapped de-duplicated reads.' This metric was similar across the NextSeq 500, HiSeq 2500 (RR), HiSeq 4000, and NovaSeq (63-68%) platforms (Table 2). The HiSeq X had a lower fraction of mapped de-duplicated reads (47%) due to higher duplication rate on the sequencer.

Median effective reads per Mb: This metric is reported by Cell Ranger DNA and can be used as a proxy for CNV detection performance. It captures the number of usable reads per megabase for CNV calling. The metric's value is dependent on sequencing depth, mapping rate, and genome size. More information on how this metric is related to CNV detection sensitivity can be found on the 10x Genomics Support website. The NextSeq 500, HiSeq 2500, and HiSeq 4000 all have median effective reads per Mb between 331 and 377 (Table 2). For the library run on NovaSeq, which had a higher total number of reads, a higher median effective reads per Mb was observed (471). For HiSeq X, however, the higher sequencing depth as compared to the NextSeq 500, HiSeq 2500, and HiSeq 4000 did not translate into higher median effective reads per Mb due to the lower fraction of mapped de-duplicated reads. This may result in lower CNV detection performance on the HiSeq X compared to other platforms at comparable sequencing depths.

Data by Cycle: We also assessed the 'Data by Cycle' plots from the Illumina Sequencing Analysis Viewer (SAV) software. All sequencers tested had similar profiles. Figure 2 shows examples of 'Data by Cycle' plots for percent base composition (% Base) from NovaSeq. The fluctuations in % Base are the result of the Single Cell DNA library construct (10x Barcode, hexamers) and the genome composition. Table 3 has interpretation of the expected profile for Single Cell DNA libraries. Figure 3 shows the 'Data by Cycle' plot for percentage of bases with a quality score of 30 or higher (Q30) across a NovaSeq sequencing run. The Phred quality score reflects the base calling accuracy and is used to determine how much of the data from a given sequencing run can be used. Sequencing runs with lower quality scores can result in a significant portion of reads being unmappable. The percentages of Q30 were relatively stable across cycles for Read 1, i7, and Read 2 for Single Cell DNA libraries.

Conclusion

The sequencing metrics and base composition of sequencing reads reported in this Technical Note serve as guidelines to assess relative sequencing run quality of Chromium Single Cell DNA libraries. Additional factors that may contribute to overall success of a sequencing run and impact application performance metrics include:

- High quality of sample preparation to obtain an adequate single cell suspension.
- Final libraries with fragment length of 300–1000 bp and a significant number of inserts between 400–600 bp for optimal cluster formation on Illumina flowcells.
- Reliable and accurate library quantification using the KAPA DNA Quantification Kit using a fixed insert size of 550 bp.
- Sequencing platform and loading concentration.

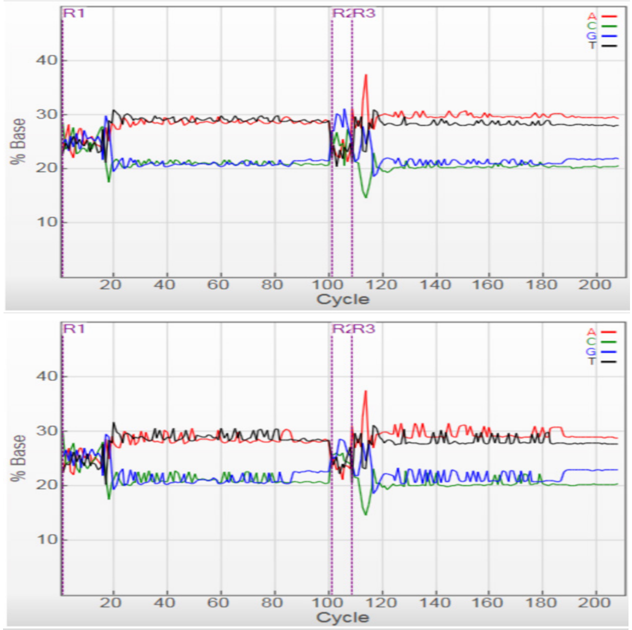

Figure 2. 'Data by Cycle' plot from the Illumina SAV software displaying the percentage of base calls from two representative Single Cell DNA library sequencing runs performed on NovaSeq.

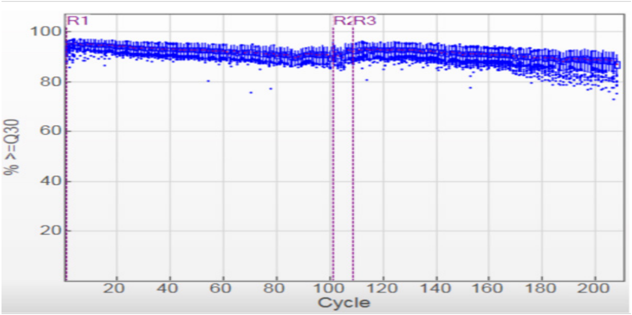

Figure 3. 'Data by Cycle' plot from the Illumina SAV software displaying the Q30 percent bases from a representative Single Cell DNA library sequencing run performed on NovaSeq.

References

- Chromium Single Cell DNA Reagent Kits User Guide (Document CG000153)

Table 2. Sequencing metrics obtained when sequencing the same Single Cell DNA library across different sequencers.

Instrument	Loading Concentration (pM)	Cluster Density (K/mm^2 or %PF)	Yield per Lane (Gb)		%≥ Q30			Total Number of Reads	Fraction of Mapped De-duplicated Reads	Median Effective Reads per Mb**
			Read 1	Read 2	Read 1	i7	Read 2			
NextSeq 500	1.7	240-250	9.5	9.5	85	90	81	895,558,884	63%	377
HiSeq 2500 (RR)	10	1100-1250	20	20	95	92.5	90	790,477,486	68%	356
HiSeq 4000	240	75-80*	39	39	96	96	88.8	771,102,450	64%	331
HiSeq X	240	75-80*	75	75	96.5	96.5	93.5	999,216,748	47%	308
NovaSeq	300	70-80*	300	300	91.5	91	89	1,124,449,548	63%	471

*Percent PF is reported for HiSeq 4000, HiSeq X, and NovaSeq instead of cluster density due to the patterned flowcell.

**Median number of reads per cell with a mapping quality of at least 30 and not duplicated, divided by the genome size in megabases.

Table 3. Interpretation of base percentages from Illumina SAV 'Data by Cycle' plot.

Read Number	Cycle Number	Expected Profile
Read 1	1-16	Base percentages fluctuate due to sequences from the 16 bp 10x Barcode. Each base is represented in roughly equal proportions.
	17-22	Base percentages fluctuate due to the hexamer.
	23-100	Base percentages reflect the expected base composition of the human genome, with a higher representation of A and T bases.
i7 Index	101-108	Base percentages fluctuate due to the 8 bp sample index.
Read 2	109-116	Base percentages fluctuate due to the hexamer.
	117-208	Base percentages reflect the expected base composition of the human genome, with a higher representation of A and T bases.

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. A non-exhaustive list of 10x Genomics' marks, many of which are registered in the United States and other countries can be viewed at: www.10xgenomics.com/trademarks. 10x Genomics may refer to the products or services offered by other companies by their brand name or company name solely for clarity, and does not claim any rights in those third-party marks or names. 10x Genomics products may be covered by one or more of the patents as indicated at: www.10xgenomics.com/patents. All products and services described herein are intended FOR RESEARCH USE ONLY and NOT FOR USE IN DIAGNOSTIC PROCEDURES.

The use of 10x Genomics products in practicing the methods set forth herein has not been validated by 10x Genomics, and such non-validated use is NOT COVERED BY 10X GENOMICS STANDARD WARRANTY, AND 10X GENOMICS 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 terms and conditions of sale for the Chromium Controller or the Chromium Single Cell 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 that it currently or will at any time in the future offer or in any way support any application set forth herein.

Contact:
support@10xgenomics.com
10x Genomics
7068 Koll Center Parkway
Suite 401
Pleasanton, CA 94566 USA