Research use only

10x Genomics Chromium Raw Data

Report

February 2023

Humanizing Genomics
macrogen

ver. MGSC_10X_Raw

Project Information

Client Name	박다소미
Company/Institution	서울대학교
Order Number	HN00189091
Read Length	R1(28) - I1(10) - I2(10) - R2(90)
Number of Samples	12
Library Kit	Chromium Next GEM Single Cell 3p RNA library v3.1

Table of Contents

1. Chromium™ Technology

1. 1. Overall Workflow

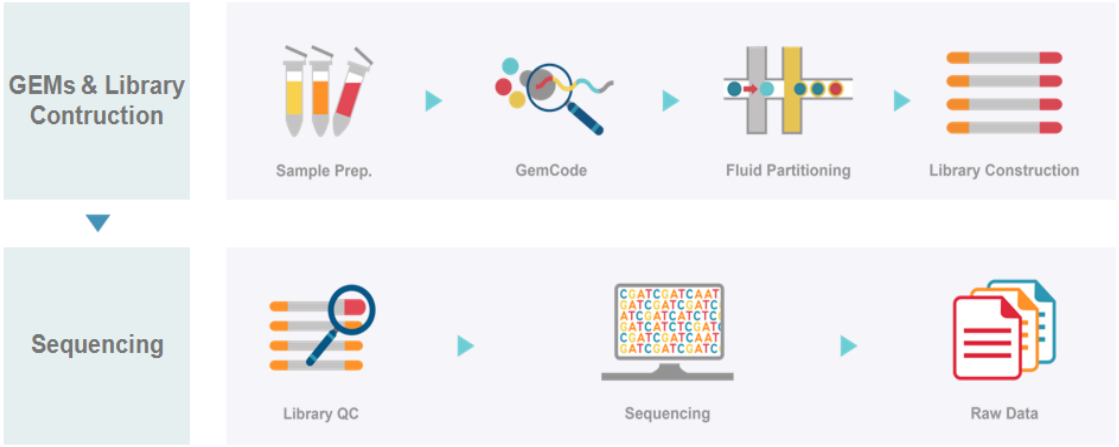


Figure 1. Overall workflow of Analysis

1. 2. GEMS & Library Construction

- The Chromium Single Cell 3' Protocol enables to profile 3' digital gene expression of 500 - 10,000 individual cells per sample.
- Single cells, reagents and a single Gel Bead containing barcoded oligonucleotides are encapsulated into nanoliter-scale GEMs (Gel Bead in emulsion) using the Next GEM Technology.
- Next GEM Technology samples a pool of about 3,500,000 10x Barcodes to separately index each cell's transcriptome and cell surface protein.
- It does so by partitioning thousands of cells into nanoliter-scale Gel Beads-in-emulsion (GEMs), where all generated DNA molecules share a common 10x Barcode.
- The poly(dT) primer captures poly-adenylated mRNA and barcoded, full-length cDNA is produced. The Single Cell 3' v3 Gel Beads include two additional primer sequences, that facilitate capture and priming of targets containing interested feature.

1. 3. Sequencing

- Single Cell 3' GEX and feature barcode libraries are sequenced on the Illumina sequencing system.
- For paired-end sequencing, Read 1 and Read 2 are sequenced from both ends of the fragment.

1. 4. Analysis

1. 4. 1. cellranger mkfastq

- 'mkfastq' module in cellranger demultiplexes raw base call (BCL) files generated by Illumina sequencers into FASTQ files.

- The 3' Gene Expression libraries and the Cell Surface Protein libraries with Feature Barcode will include the paired-end Read 1 (containing the 16 bp 10x™ Barcode and 12 bp UMI), Read 2 (90 bp insert) and the sample index in the i7-i5 dual index (10-10 bp) reads.

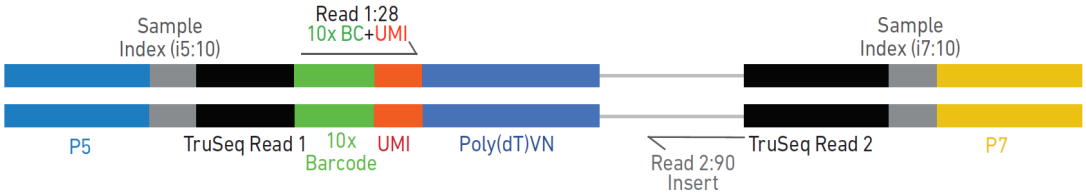


Figure 2. Chromium Next GEM Single Cell 3p RNA library v3.1

If you want to see more detailed information, please refer to the 10X User Guide Documents.

LINK CG000317.Rev.A.pdf

2. Summary of Data Production

2. 1. Raw Data Statistics

(Refer to Path: 10X_RawData_Outs/RawData.statistics.summary.xlsx)

The total number of bases, reads, GC (%), Q20 (%), Q30 (%) are calculated for 12 samples. For example, in SPF_HC_1, 336,558,590 reads are produced, and total read bases are 19.9Gbp. The GC content (%) is 49.24% and Q30 is 91.18%.

Sample	Read	Total Bases	Read Count	GC(%)	Q20(%)	Q30(%)
SPF_HC_1	Read1	4,711,820,260	168,279,295	49.37	99.07	97.44
	Read2	15,145,136,550	168,279,295	49.20	95.25	89.23
	R1+R2	**19,856,956,810**	**336,558,590**	**49.24**	**96.16**	**91.18**
SPF_HC_2	Read1	5,668,567,296	202,448,832	49.25	99.11	97.50
	Read2	18,220,394,880	202,448,832	48.84	94.87	88.42
	R1+R2	**23,888,962,176**	**404,897,664**	**48.94**	**95.87**	**90.57**
SPF_HC_3	Read1	4,450,884,144	158,960,148	49.23	99.11	97.51
	Read2	14,306,413,320	158,960,148	47.94	95.63	89.90
	R1+R2	**18,757,297,464**	**317,920,296**	**48.24**	**96.46**	**91.71**
GF_HC_1	Read1	5,873,705,656	209,775,202	49.17	99.15	97.60
	Read2	18,879,768,180	209,775,202	47.71	95.79	90.28
	R1+R2	**24,753,473,836**	**419,550,404**	**48.06**	**96.59**	**92.02**
GF_HC_2	Read1	5,834,352,972	208,369,749	49.20	99.12	97.55
	Read2	18,753,277,410	208,369,749	47.00	95.52	89.77
	R1+R2	**24,587,630,382**	**416,739,498**	**47.52**	**96.37**	**91.61**
GF_HC_3	Read1	5,958,605,352	212,807,334	49.16	99.14	97.57
	Read2	19,152,660,060	212,807,334	47.46	95.39	89.46
	R1+R2	**25,111,265,412**	**425,614,668**	**47.86**	**96.28**	**91.39**
SPF_NPC_1	Read1	5,541,405,352	197,907,334	49.42	99.14	97.61
	Read2	17,811,660,060	197,907,334	45.28	95.79	90.48
	R1+R2	**23,353,065,412**	**395,814,668**	**46.26**	**96.58**	**92.17**
SPF_NPC_2	Read1	6,647,638,760	237,415,670	49.30	99.16	97.65
	Read2	21,367,410,300	237,415,670	45.47	95.41	89.72
	R1+R2	**28,015,049,060**	**474,831,340**	**46.38**	**96.30**	**91.60**
SPF_NPC_3	Read1	6,131,148,744	218,969,598	49.38	99.12	97.54
	Read2	19,707,263,820	218,969,598	44.25	94.75	88.48
	R1+R2	**25,838,412,564**	**437,939,196**	**45.47**	**95.78**	**90.63**
GF_NPC_1	Read1	4,819,350,228	172,119,651	49.43	99.20	97.75
	Read2	15,490,768,590	172,119,651	44.67	96.29	91.45

	R1+R2	**20,310,118,818**	**344,239,302**	**45.80**	**96.98**	**92.95**
GF_NPC_2	Read1	4,921,708,624	175,775,308	49.36	99.14	97.58
	Read2	15,819,777,720	175,775,308	45.34	94.69	88.27
	R1+R2	**20,741,486,344**	**351,550,616**	**46.29**	**95.75**	**90.47**
GF_NPC_3	Read1	4,981,379,816	177,906,422	49.18	99.20	97.71
	Read2	16,011,577,980	177,906,422	44.45	96.25	91.33
	R1+R2	**20,992,957,796**	**355,812,844**	**45.57**	**96.95**	**92.84**

- Read: Read category
- Total Bases: Total number of bases sequenced
- Read Count: Total number of reads
- GC (%): GC content
- Q20 (%): Ratio of bases that have phred quality score greater than or equal to 20
- Q30 (%): Ratio of bases that have phred quality score greater than or equal to 30

2. 2. Average Base Quality at Each Cycle

(Refer to Path: 10X_RawData_Outs/[Sample]/[Flowcell]/FASTQC)

The quality of produced data is determined by the phred quality score at each cycle. Box plot containing the average quality at each cycle is created with FastQC.

The x-axis shows number of cycles and y-axis shows phred quality score. Phred quality score 20 means 99% accuracy and reads over score of 20 are accepted as good quality.

LINK http://www.bioinformatics.babraham.ac.uk/projects/fastqc

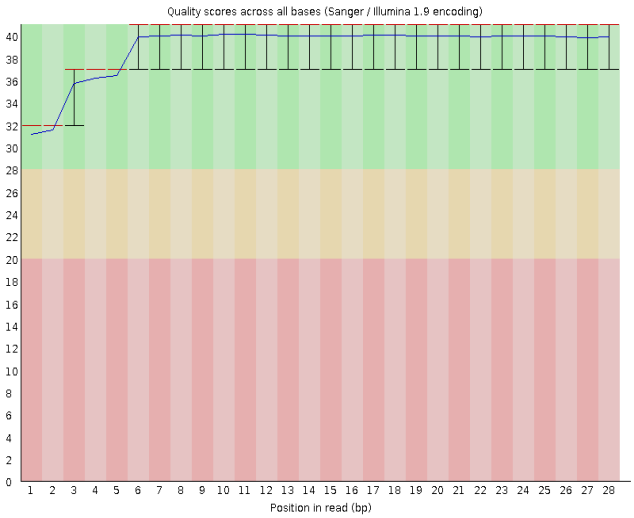


Figure 3. Read quality at each cycle of SPF_HC_1 (read1)

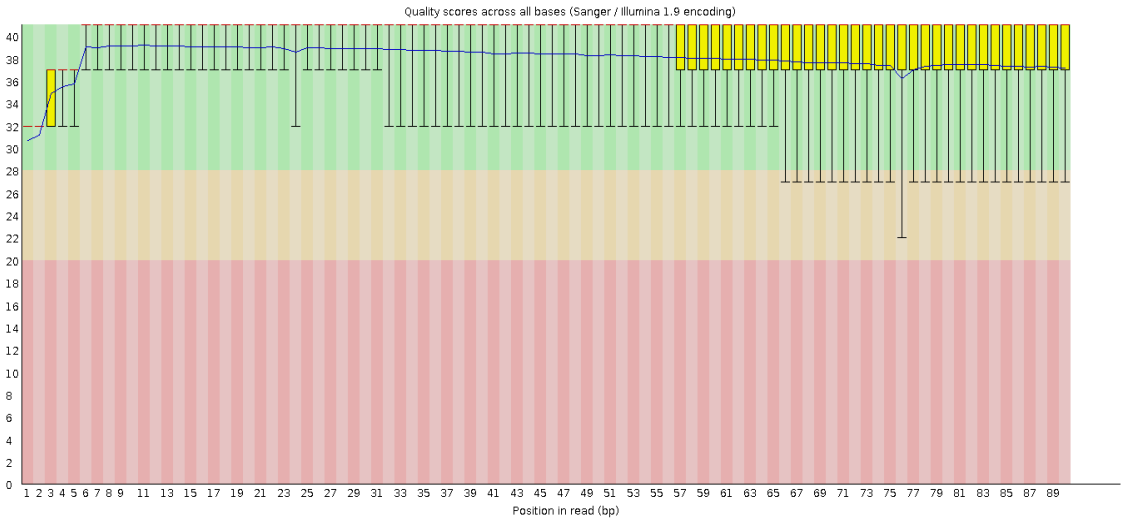
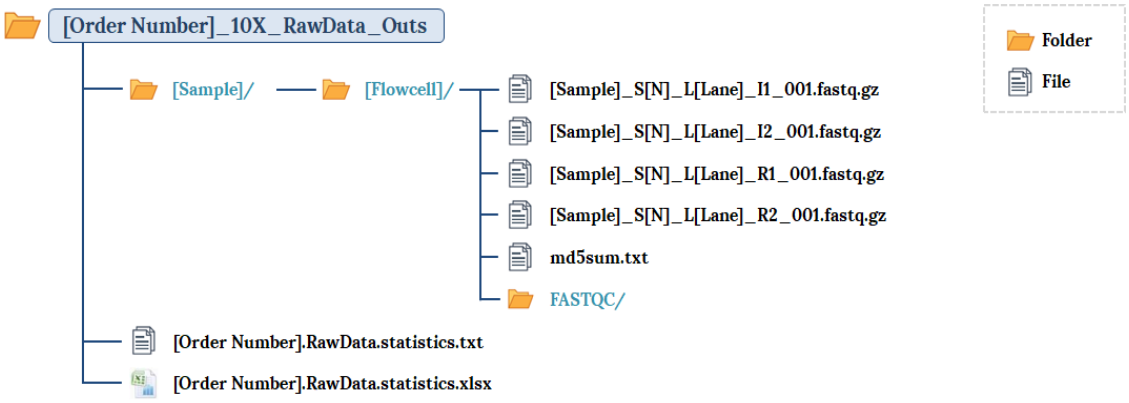


Figure 4. Read quality at each cycle of SPF_HC_1 (read2)

- Yellow box: Interquartile range (25-75%) of phred score at each cycle
- Red line: Median of phred score at each cycle
- Blue line: Average of phred score at each cycle
- Green background: Good quality
- Orange background: Acceptable quality
- Red background: Bad quality

3. Data Information

3. 1. Folder Structure



4. Appendix

4. 1. Phred Quality Score Chart

Phred quality score numerically express the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000.

Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

Phred Quality Score Q is calculated with $-10\log_{10}P$, where P is probability of erroneous base call.

4. 2. Programs used in Analysis

4. 2. 1. FastQC v0.11.7

LINK http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC is a program that performs quality check on the raw sequences before analysis to make sure data integrity. The main function is importing BAM, SAM, FastQ files and providing quick overview on which section has problems. It provides such results as graphs and tables in html files.

4. 2. 2. Cell Ranger 7.0.1, bcl2fastq v2.20.0

LINK what-is-cell-ranger

Cell Ranger is a set of analysis pipelines that process Chromium single-cell RNA-seq output to align reads, generate feature-barcode matrices and perform clustering and gene expression analysis. Cell Ranger includes four pipelines relevant to single-cell gene expression experiments

- cellranger mkfastq: Demultiplexes raw base call (BCL) files generated by Illumina sequencers into FASTQ files. It is a wrapper around Illumina's bcl2fastq, with additional useful features that are specific to 10x libraries and a simplified sample sheet format.
- cellranger count: Takes FASTQ files from cellranger mkfastq and performs alignment, filtering, barcode counting, and UMI counting. It uses the Chromium cellular barcodes to generate feature-barcode matrices, determine clusters, and perform gene expression analysis. The count pipeline can take input from multiple sequencing runs on the same GEM well. cellranger count also processes Feature Barcoding data alongside Gene Expression reads.
- cellranger vdj: Takes FASTQ files from cellranger mkfastq or bcl2fastq for V(D)J libraries and performs sequence assembly and paired clonotype calling. It uses the Chromium cellular barcodes and UMIs to assemble V(D)J transcripts per cell. Clonotypes and CDR3 sequences are output as a .vloupe file which can be loaded into Loupe V(D)J Browser.
- cellranger aggr: Aggregates outputs from multiple runs of cellranger count, normalizing those runs to the same sequencing depth and then recomputing the feature-barcode matrices and analysis on the combined data. The aggr pipeline can be used to combine data from multiple samples into an experiment-wide feature-barcode matrix and analysis.
- cellranger reanalyze: Takes feature-barcode matrices produced by cellranger count or cellranger aggr and reruns the dimensionality reduction, clustering, and gene expression algorithms using tunable parameter settings.
- cellranger multi: Takes FASTQ files from cellranger mkfastq or bcl2fastq for any combination of 5' Gene Expression, Feature Barcode (cell surface protein or antigen) and V(D)J libraries from a single gem-well. It performs alignment, filtering, barcode counting, and UMI counting on the Gene Expression and/or Feature Barcode libraries. It also performs sequence assembly and paired clonotype calling on the V(D)J libraries. Additionally, the cell calls provided by the gene expression data are used to improve the cell calls inferred by the V(D)J library.

Research use only

HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support Center

[08511] 1001, 10F, 254, Beotkkot-ro, Geumcheon-gu, Seoul, Republic of Korea (Gasan-dong, World Meridian 1)
Tel: +82-2-2180-7000
Email1: ngs@macrogen.com(Overseas)
Email2: ngskr@macrogen.com (Republic of Korea)
Web: www.macrogen.com
LIMS: dna.macrogen.com

SUBSIDIARY

Macrogen Europe

Laboratory, Business & Support Center

Meibergdreef 57, 1105 BA, Amsterdam, the Netherlands
Tel: +31-20-333-7563
Email: ngs@macrogen.eu

Macrogen Singapore

Laboratory, Business & Support Center

3 Biopolis Drive #05-18, Synapse, Singapore 138623
Tel: +65-6339-0927
Email: info-sg@macrogen.com

Psomagen (Macrogen USA)

Laboratory, Business & Support Center

1330 Piccard Drive, Suite 103, Rockville, MD 20850, United States
Tel: +1-301-251-1007
Email: inquiry@psomagen.com

Macrogen Japan

Laboratory, Business & Support Center

16F Time24 Building, 2-4-32 Aomi, Koto-ku, Tokyo 135-0064 JAPAN
Tel: +81-3-5962-1124
Email: ngs@macrogen-japan.co.jp

BRANCH

Macrogen Spain

Laboratory, Business & Support Center

Av. Sur del Aeropuerto de Barajas, 28. Office B-2, 28042 Madrid, Spain
Tel: +34-911-138-378
Email: info-spain@macrogen.com

Humanizing Genomics
macrogen