Research use only

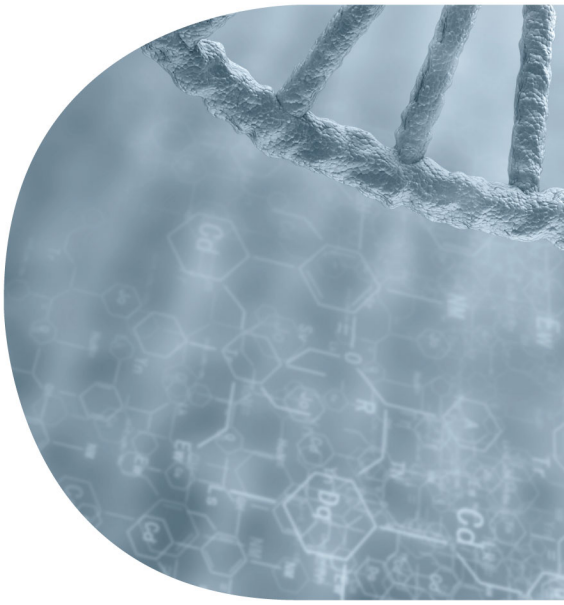
human

10x Genomics Chromium Analysis

Single Cell Gene Expression Flex

Report

November 2023



Humanizing Genomics
macrogen

Project Information

Company/Institution	Testname
Order Number	FLEX_test
Species	*human*
Reference	GRCh38
Read Length	R1(28) - I1(10) - I2(10) - R2(90)
Number of Samples	1
Library Kit	Chromium Next GEM Single Cell Fixed RNA library v1

Table of Contents

1. Chromium™ Technology

1. 1. Overall Workflow

Figure 1. Overall workflow of Analysis

1. 2. GEMS & Library Construction

- Chromium Fixed RNA Profiling offers comprehensive, scalable solutions to measure gene expression in formaldehyde fixed samples.
- Gene expression is measured using probe pairs designed to hybridize to mRNA specifically.
- Using a microfluidic chip, the fixed and probe-hybridized single cell and nuclei suspensions are partitioned into nanoliter-scale Gel Beads-in-emulsion (GEMs).
- A pool of about 737,000 10x GEM Barcodes (also referred to as 10x Barcodes) is sampled separately to index the contents of each partition.
- Inside the GEMs, probes are ligated and the 10x GEM Barcode is added, and all ligated probes within a GEM share a common 10x GEM Barcode.
- Barcoded and ligated probes are then pre-amplified in bulk, after which gene expression libraries are generated and sequenced.

1. 3. Sequencing

- Single Cell 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 Single Cell Gene Expression Flex libraries include the paired-end Read 1 (containing the 16 bp

10x™ Barcode, 12 bp UMI), Read 2 (50 bp Ligated Probe Insert, 16 bp Const.Seq, 12 bp pCS and 8 bp Probe barcode for multiplexing) and the unique dual sample index in the i7, i5 index (each 10 bp) read.

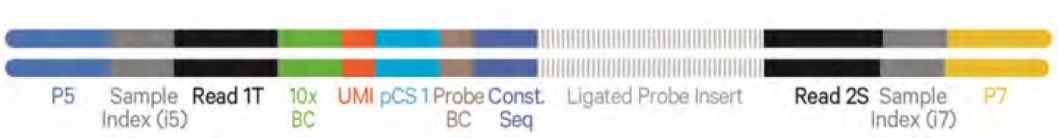


Figure 2. Chromium Next GEM Single Cell Fixed RNA library v1

1. 4. 2. cellranger multi

cellranger multi 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 cellranger multi pipeline also supports the analysis of Feature Barcode data for singleplexed Gene Expression Flex library.

Fixed RNA Profiling Gene Expression relies on the same underlying Cell Ranger analysis software as Single Cell 3' Gene Expression solutions, but uses a short read aligner tailored to probe sequences. The probe aligner assigns probe and gene IDs to reads originating from ligation of correctly paired probe halves, distinguishing from potential artifacts and non-probe reads. For fixed samples, Cell Ranger counts oligo ligation events to build the feature-barcode matrix. As a consequence, if the probe hybridizes to a transcript with a variant, that variant will not be present in the sequencing data.

1) The cell barcode will be composed only of the 10x GEM Barcode and all reads with valid barcodes are considered to be part of the sample.

1) The cell barcode will be composed of both the 10x GEM Barcode and the Probe Barcode. Reads must have valid 10x GEM and Probe Barcodes in order to be assigned to a sample.

2) Reads are aligned with the probe set reference CSV file information specified in the probe-set parameter, and ligation events are counted using the probe aligner algorithm.

3) Removes barcodes not associated with cellular GEM partitions.

4) Generates feature-barcode matrices by counting unique molecules.

5) Probes that are predicted to have off-target activity to homologous genes or sequences are excluded from analysis by default.

6) Cell calling is performed and we utilize the 10x GEM and Probe Barcodes to estimate the mean of the Poisson distribution that should describe the number of Probe Barcodes observed per GEM under optimal chip loading.

7) Runs principal component analysis to reduce gene expression to its most highly variable

components.

8) Runs graph-based clustering and modularity optimization to partition the data into subpopulations based on PCA results.

9) Runs k-means clustering to partition the data into subpopulations based on PCA results.

10) Runs t-Distributed Stochastic Neighbor Embedding to visualize PCA results in two dimensions.

11) Runs Uniform Manifold Approximation and Projection to visualize PCA results in two dimensions.

12) Merge graph-based clusters which have insufficient differential expression between them.

13) Combines the different clustering results into a single file.

14) Identifies the most differentially expressed genes in each cluster relative to other clusters.

15) Estimate gDNA contamination using a subset of transcript probes that have a mix of exon-junction-spanning (spliced) and non-exon-junction-spanning (unspliced) probes for the same gene.

16) Creates CSV and HTML files summarizing sequencing run performance.

17) Produces the input file for Loupe Cell Browser.

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

LINK CG000477.Rev.D.pdf
LINK Fixed_RNA_Profiling_Algorithms_Overview

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 1 sample. For example, in SAMPLE_1, 1,637,532,240 reads are produced, and total read bases are 96.6Gbp. The GC content (%) is 50.77% and Q30 is 94.91%.

Sample	Read	Total Bases	Read Count	GC(%)	Q20(%)	Q30(%)
SAMPLE_1	Read1	22,925,451,360	818,766,120	45.49	98.78	96.07
	Read2	73,688,950,800	818,766,120	52.41	98.14	94.55
	R1+R2	**96,614,402,160**	**1,637,532,240**	**50.77**	**98.30**	**94.91**

- 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

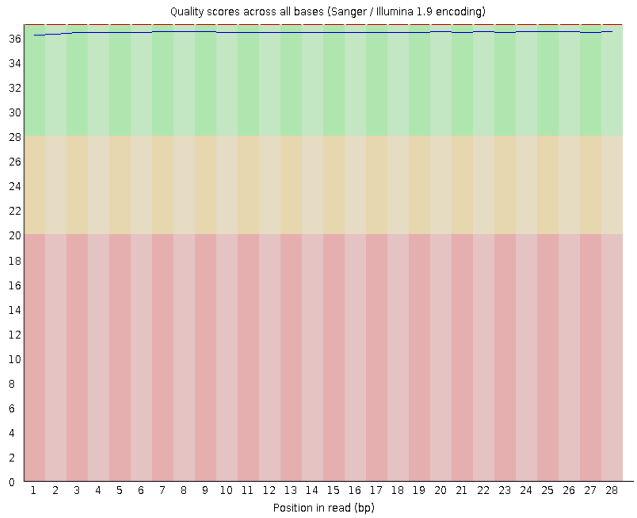


Figure 3. Read quality at each cycle of SAMPLE_1 (read1)

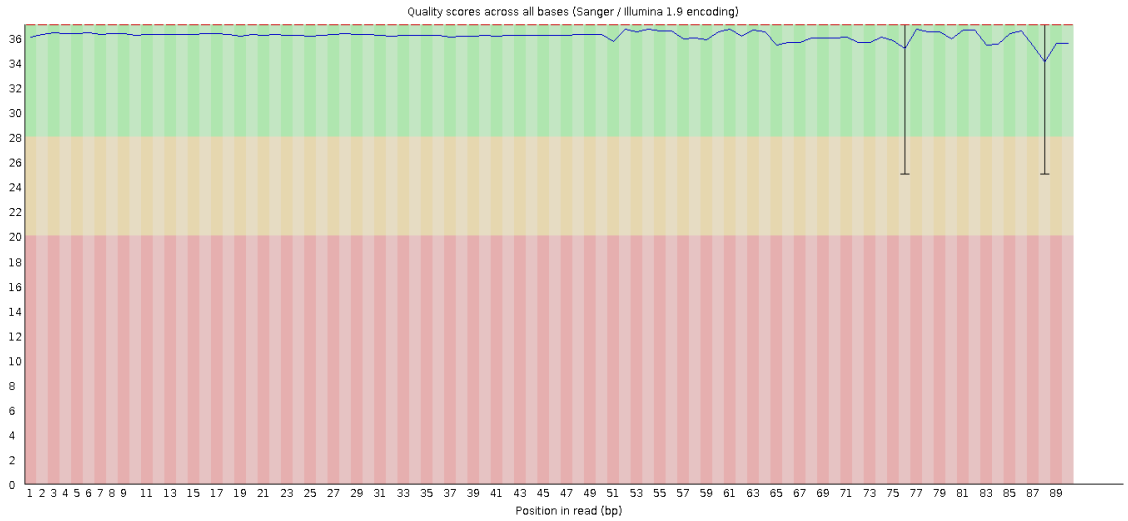


Figure 4. Read quality at each cycle of SAMPLE_1 (read2)

- 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. Cell Ranger Analysis Result

3. 1. cellranger "multi" Analysis Result

(Refer to Path: result_10X/Analysis_statistics/Basic_Analysis_Statistics.xlsx)
(Refer to Path: result_10X/[Group]//per_sample_outs/[Sample]/web_summary.html)

Fixed RNA Profiling (FRP) data was analyzed with cellranger multi pipeline. There are a few differences for singleplex and multiplex Fixed RNA Profiling expreiments. Please refer to the detailed algorithm at the following link.

LINK Fixed RNA Profiling Algorithm

The standard analysis workflow is followed.

1) Extract and correct the cell barcode and probe barcode from R1 and R2 reads.

2) Align reads to the probe set with R2.

3) Correct and filter UMIs

4) Count UMIs and generate raw feature-barcode matrix

5-1) In case of multiplex Fixed RNA library, after 4) step, assign reads to samples and call cells per sample

5-2) In case of singleplex Fixed RNA library, after 4) step, call cells.

6) Filter probes that are predicted to have off-targt activity to homologous genes or sequences

7) In case of multiplex Fixed RNA library, utilizing the number of probe barcodes observed per GEM that follows the Poisson distribution, additionally filter high occupancy GEMs due to high occupancy GEMs may reflect partial clogs/wetting failures or significant deviations from the recommended chip loading protocol

8) As the result, sample filtered feature-barcode matrix is generated.

3. 1. 1. Physical Library Metrics (Library View Summary)

Below tables displays the most important results of the experiment. Summary statistics about cell-associated barcodes. For details, please refer to the web_summary file for each sample.

Table 1. Gene Expression Library Key Metrics

Group Name	Physical library ID	Estimated number of cells	Mean reads per cell	Fraction of initial cell barcodes passing high occupancy GEM filtering
SAMPLE_1	GEX_1	41,071	19,935	99.50%

- Group Name: Pooling group name
- Physical library ID: Library ID of the produced FASTQ
- Estimated number of cells: The number of barcodes identified by the cell-calling algorithm as containing a cell.
- Mean reads per cell: The total number of sequenced read pairs divided by the number of cell-associated barcodes.
- Fraction of initial cell barcodes passing high occupancy GEM filtering: Fraction of cell-associated barcodes from initial cell calls that remain after high occupancy GEM filtering.

Table 2. Gene Expression Library Additional Metrics

Group Name	Physical library ID	Valid barcodes	Valid GEM barcodes	Valid probe barcodes	Valid UMIs	Reads mapped to probe set	Confidently mapped reads in cells
SAMPLE_1	GEX_1	96.56%	98.62%	97.85%	99.98%	98.54%	88.58%

- Group Name: Pooling group name
- Physical library ID: Library ID of the produced FASTQ
- Valid barcodes: Fraction of reads with barcodes that are present in the whitelist after barcode correction.
- Valid GEM barcodes: Fraction of reads with GEM barcodes that are present in the whitelist after barcode correction.
- Valid probe barcodes: Fraction of reads with probe barcodes that are present in the whitelist after barcode correction.
- Valid UMIs: Fraction of reads with valid UMIs; i.e. UMI sequences that do not contain Ns and that are not homopolymers.
- Reads mapped to probe set: Fraction of reads that mapped to the probe set.
- Confidently mapped reads in cells: The fraction of valid barcode, valid UMI, confidently mapped to transcriptome reads with cell associated barcodes.

Table 3. Metrics per probe barcode

Group Name	Probe barcode ID	Sample ID	UMIs per probe barcode	Cells per probe barcode
SAMPLE_1	BC001	BC1	84,335,661 (36.27%)	9,439 (22.98%)
	BC002	BC2	80,155,501 (34.47%)	13,286 (32.35%)
	BC003	BC3	28,889,191 (12.42%)	8,001 (19.48%)
	BC004	BC4	39,025,586 (16.78%)	10,345 (25.19%)

- Group Name: Pooling group name
- Probe barcode ID: The identifier of this probe barcode.
- Sample ID: The identifier of the sample associated with this probe barcode.
- UMIs per probe barcode: Number and fraction of UMIs for this probe barcode amongst all UMIs in the raw feature barcode matrix.
- Cells per probe barcode: Number and fraction of cells per probe barcode amongst all cells detected in this GEM well.

3. 1. 2. Samples Metrics (Cells View Summary)

Below tables displays the most important results of each sample for each pooling group. Summary statistics about cell-associated barcodes.

Table 4. Cells Key Metrics

Group Name	Sample	Cells	Median reads per cell	Median genes per cell	Total genes detected	Median UMI counts per cell	Confidently mapped reads in cells
SAMPLE_1	BC1	9,439	6,940	1,261	17,235	1,977	93.33%
	BC2	13,286	5,361	988	17,218	1,515	91.22%
	BC3	8,001	6,781	1,352	16,767	1,934	80.24%
	BC4	10,345	7,485	1,449	16,822	2,152	79.43%

- Group Name: Pooling group name
- Sample: The identifier of the sample associated with this probe barcode.
- Cells: Number of cells called from each samples.
- Median reads per cell: Median number of read pairs sequenced from the cells called from this sample.
- Median genes per cell: The median number of genes detected per cell called from this sample.
- Total genes detected: The number of genes with at least one UMI count in the cells in this sample.
- Median UMI counts per cell: Median number of UMIs obtained from the cells called from this sample.
- Confidently mapped reads in cells: The fraction of valid barcode, valid UMI, confidently mapped to transcriptome reads with cell associated barcodes.

Table 5. Cells Mapping Metrics - Amongst Reads From Cells Assigned To Sample

Group Name	Sample	Number of reads from cells called from this sample	Reads mapped to probe set	Reads confidently mapped to probe set	Reads confidently mapped to filtered probe set
SAMPLE_1	BC1	265,843,064	99.11%	98.42%	94.58%
	BC2	248,243,691	99.13%	98.52%	94.39%
	BC3	79,310,964	98.96%	98.20%	96.03%
	BC4	105,046,019	98.96%	98.36%	95.91%

- Group Name: Pooling group name
- Sample: The identifier of the sample associated with this probe barcode.
- Number of reads from cells called from this sample: The total number of reads from cells called from this sample.
- Reads mapped to probe set: Fraction of reads that mapped to the probe set.
- Reads confidently mapped to probe set: Fraction of reads that mapped uniquely to a probe in the probe set.
- Reads confidently mapped to filtered probe set: Fraction of reads from probes that map to a unique gene. These reads are considered for UMI counting.

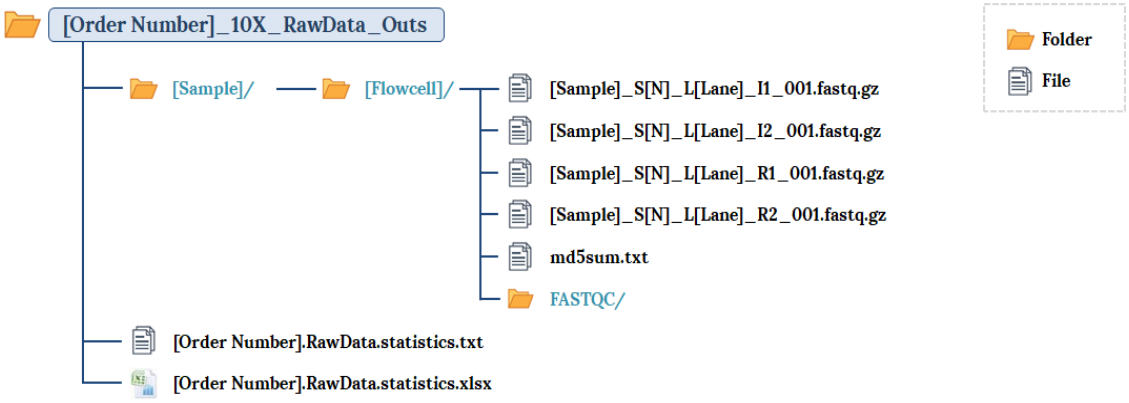
4. Data Information

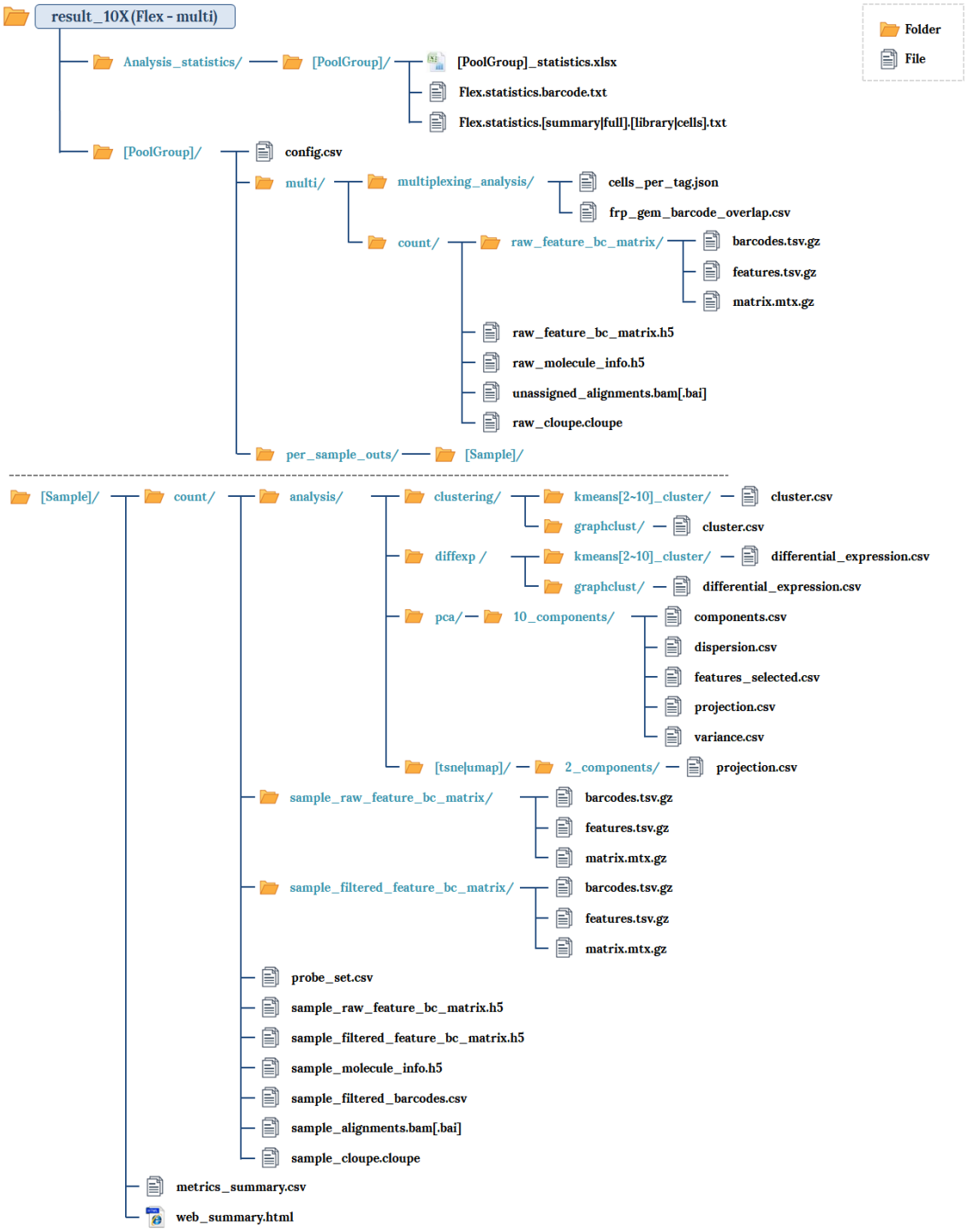
The data for this order will be delivered to you via HDD. Please check the “Folder tree” page for details of the files contained in the HDD.

Once you receive/download the data, please make sure to check the integrity of the files.

Please note that the sequencing files will be deleted from our server 3 months after the analysis report is released. Please contact us within 3 months if you encounter a problem with the data.

4. 1. Folder Structure




result_10X (Flex - multi)
Folder
File
Analysis_statistics/
[PoolGroup]/
[PoolGroup]_statistics.xlsx
Flex.statistics.barcode.txt
Flex.statistics.[summary|full].[library|cells].txt
[PoolGroup]/
config.csv
multi/
multiplexing_analysis/
cells_per_tag.json
frp_gem_barcode_overlap.csv
count/
raw_feature_bc_matrix/
barcodes.tsv.gz
features.tsv.gz
matrix.mtx.gz
raw_feature_bc_matrix.h5
raw_molecule_info.h5
unassigned_alignments.bam[.bai]
raw_cloupe.cloupe
per_sample_outs/
[Sample]/
[Sample]/
count/
analysis/
clustering/
kmeans[2~10]_cluster/
cluster.csv
graphclust/
cluster.csv
diffexp /
kmeans[2~10]_cluster/
differential_expression.csv
graphclust/
differential_expression.csv
pca/
10_components/
components.csv
dispersion.csv
features_selected.csv
projection.csv
variance.csv
[tsne|umap]/
2_components/
projection.csv
sample_raw_feature_bc_matrix/
barcodes.tsv.gz
features.tsv.gz
matrix.mtx.gz
sample_filtered_feature_bc_matrix/
barcodes.tsv.gz
features.tsv.gz
matrix.mtx.gz
probe_set.csv
sample_raw_feature_bc_matrix.h5
sample_filtered_feature_bc_matrix.h5
sample_molecule_info.h5
sample_filtered_barcodes.csv
sample_alignments.bam[.bai]
sample_cloupe.cloupe
metrics_summary.csv
web_summary.html

5. Appendix

5. 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.

5. 2. Programs used in Analysis

5. 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.

5. 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