E171782 10x scRNAseq Summary Report

1. Project Information

Investigator: Scott Waddell
Report by: Richard Clark
Report Date: 30/03/2021

Number of Samples: 7

2. Summary of Sequencing Protocol

Library Preparation

The Chromium Single Cell Gene Expression solution allows the 3' digital gene expression profiling of 500-10,000 individual cells from a single sample. GemCode Technology samples a pool of ~3,500,000 10x barcodes to separately index each cell's transcriptome, by partitioning thousands of cells into nanolitre-scale Gel Beads-in-emulsion (GEMs), where all generated cDNA share a common barcode.

Seven barcoded and amplified cDNA samples prepared from the Investigator's samples using the Chromium Single Cell 3' GEM, Library & Gel Bead Kit v3 (10x Genomics, #1000128) were provided by the IGMM Flow Cytometry Service. cDNA samples were quantified using the Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc, #Q32866) and the Qubit dsDNA HS assay Kit (#Q32851). Electropherogram traces generated on the Agilent Bioanalyser 2100 Capillary Electrophoresis instrument were provided by the IGMM Flow Cytometry Service to allow assessment of cDNA fragment size distribution.

The Illumina TruSeq Read 1 (Read 1 primer sequence) is added to the cDNA molecules during GEM incubation. Chromium Single Cell 3' Gene Expression libraries were generated from the barcoded and amplified cDNA by the Edinburgh Clinical Research Facility Genetics Core using the Chromium Next GEM Single Cell 3' Library Construction Kit v3.1 (#1000157). Enzymatic fragmentation and size-selection were used to optimise cDNA amplicon size. P5, P7, a sample index, and TruSeq Read 2 (read 2 primer sequence) were added via End-repair, A-tailing, adapter-ligation and PCR (13 cycles).

A Chromium Single Cell 3' Gene Expression library comprises standard Illumina paired-end constructs which begin and end with P5 and P7. The 16 bp 10x Barcode and 12 bp UMI are encoded in Read 1, while Read 2 is used to sequence the cDNA fragment. Sample index sequences are incorporated as the i7 index read. TruSeq Read 1 and TruSeq Read 2 are standard Illumina sequencing primer sites used in paired-end sequencing.

Seven chromium Single Cell 3' Gene Expression libraries were assessed for size distribution on the Agilent Bioanalyser with the DNA HS Kit, and quantified using the Qubit dsDNA HS assay Kit.

Sequencing on iSeq 100

Sequencing was performed initially on the iSeq 100 System (Illumina Inc, #20021532) using the iSeq 100 i1 Reagent v2 (300 cycle) Kit Kit (#20031371) over a single flow cell. Library

molarity for sequencing was calculated using the Qubit dsDNA quantification results and the fragment size information from the Bioanalyser results. PhiX v3 Control (Illumina Inc, # FC-110-3001) was spiked in at a concentration of 2% to increase library diversity and enable troubleshooting in the event of any run issues.

Data from the run was transferred from the iSeq 100 instrument to a computer running Linux and the Cell Ranger (version 5.0.0) set of analysis pipelines provided by 10x Genomics. Single-cell RNA libraries generated using the Chromium Single Cell 3' Library Kit v3 are assigned index sets containing four sample index oligonucleotides.

Cellranger mkfastq demultiplexes raw base call (BCL) files generated by Illumina sequencers into FASTQ files. It is a wrapper around Illumina's Bcl2fastq2 (Version 2.20), 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. Output from cellranger count provided an estimate of the number of cells represented in each library.

Output from both cellranger mkfastq and cellranger count was archived and uploaded to the University FTP site for sharing with the Investigator.

Sequencing on NextSeq 2000

Sequencing was performed on the NextSeq 2000 platform (Illumina Inc, #SY-415-1002) using the NextSeq 1000/2000 P3 Reagents (100 cycles) Kit (#20040559). PhiX Control v3 (#FC-110-3001) library was spiked in to the run at a concentration of 2% to help with cluster resolution and facilitate troubleshooting in case of any problems with the run. Pools for running on the NextSeq 2000 were balanced using the output from the iSeq 100 run, detailed in the results summary below.

3. Results Summary

Table 1 below shows key iSeq 100 System performance parameters for different run configurations.

Table 1: iSeq 100 System performance parameters

Run Configuration	Reads Passing Filter (PF) / run	Output	Average Quality Scores
1 x 36bp	4M	144Mb	>85%
1 x 50bp	4M	200Mb	>85%
1 x 75bp	4M	300Mb	>80%
2 x 75bp	4M	600Mb	>80%
2 x 150bp	4M	1.2Mb	>80%

When multiplexing 7 libraries we would therefore expect to generate 570K paired-end (PE) reads per library.

75.1% of clusters passed quality filters (PF) generating 1.1Gb of data with 92.9% ≥Q30. Coverage of each library was variable but output was higher than expected so all generated >700K PE reads (Min: 695K, Max: 884K, Mean: 778K).

Table 2 below shows the number of clusters PF per indexed library and the mean reads per cell calculated by the cellranger count pipeline.

Table 2: iSeq 100 Number of Clusters PF

Sample ID	Number of Clusters PF	Mean reads per cell
K19Wdr3512m_155	743,351	3,260
K19Wdr3512m_158	865,553	445
K19Wdr3512m_168	733,523	1,833
K19Wdr3512m_177	695,374	1,570
K19Wdr3512m_186	763,339	1,346
K19Wdr3512m_176	763,778	1,266
K19Wdr3512m_200	884,414	129

It was reasoned that the mean reads per cell estimated by cellranger count from the iSeq 100 data could be used to rebalance the pool for the higher-output NextSeq 2000 runs in order to achieve even coverage of each library that took into account the different numbers of cells represented in each library. However, discussions with 10x Genomics technical support cast doubt on whether these estimates would be accurate having been generated from such a relatively low number of reads.

Instead a library rebalancing strategy based on index representation was adopted. A loading factor was calculated for each library as the ratio between the highest normalised index representation across all libraries and the index representation for the current library. Multiplying the original input volume by the calculated loading factor provided a new volume for index rebalanced pooling prior to sequencing on the NextSeq 2000.

A 100 cycle sequencing run on the Nextseq 2000 using a P3 flow cell is expected to generate up to 1.1B paired-end (PE) reads (40Gb) with a data quality of >85% higher than Q30.

The first run on the NextSeq 2000 went well. 1.4B PE reads passed quality filters (PF) generating 174.4Gb of data with 95.2% ≥Q30. Coverage of each index was considerably more even than the iSeq 100 run. The number of clusters PF for each indexed library is shown in Table 3 below, along with the mean reads per cell calculated using cellranger count.

Table 3: NextSeq 2000 Run 1 Number of Clusters PF

Sample ID	Loading Factor	Number of Clusters PF	Mean reads per cell
K19Wdr3512m_155	1.2	191,522,293	459,286
K19Wdr3512m_158	1.0	183,647,688	55,282
K19Wdr3512m_168	1.2	183,959,341	400,782
K19Wdr3512m_177	1.3	189,494,948	293,335
K19Wdr3512m_186	1.2	184,462,297	268,895
K19Wdr3512m_176	1.2	182,345,440	157,058
K19Wdr3512m_200	1.0	187,488,056	11,376

The larger data set gave greater confidence in the cellranger count mean reads per cell calculation so a strategy for a second NextSeq 2000 run was proposed that calculated a

loading factor as the ratio between the highest mean reads per cell across all libraries and the mean reads per cell for the current library. This loading factor is expressed in table 4 below as a desired percentage of the reads, along with the revised target percentage for each library requested by the Investigator and the proportion of the reads achieved. At the Investigator's request K19Wdr3512m_155 was omitted from the second run due to the relatively high mean reads per cell for this library.

Table 4:

Sample ID	Desired %	Revised target %	Number of Clusters PF	Proportion of reads (%)
K19Wdr3512m_158	14.8	16.8	301,533,539	23.8
K19Wdr3512m_168	2.0	5.2	75,865,107	6.0
K19Wdr3512m_177	2.8	5.8	80,297,997	6.3
K19Wdr3512m_186	3.0	5.0	70,009,004	5.5
K19Wdr3512m_176	5.2	7.2	116,497,438	9.2
K19Wdr3512m_200	72.1	60.0	621,146,194	49.1

Richard Clark
30th March 2021