Comparative analysis of multiple single-cell RNASeq platforms

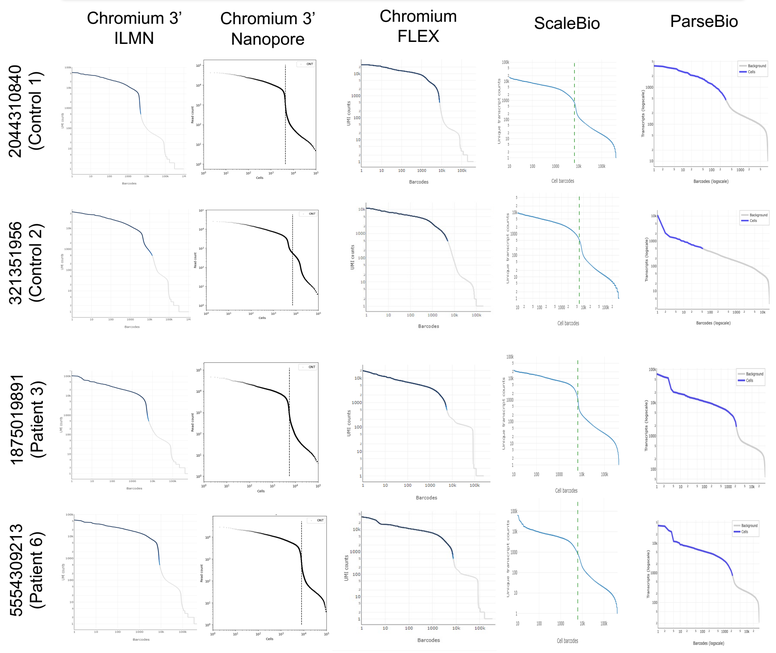
C. Kao[1], H. Qu[1], C. Hou[1], M. Hufnell[1], D. Slater[1], F. Wang[1], K. Wang[1], H. Hakonarson[1,2]

[1]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA; [2]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

Abstract

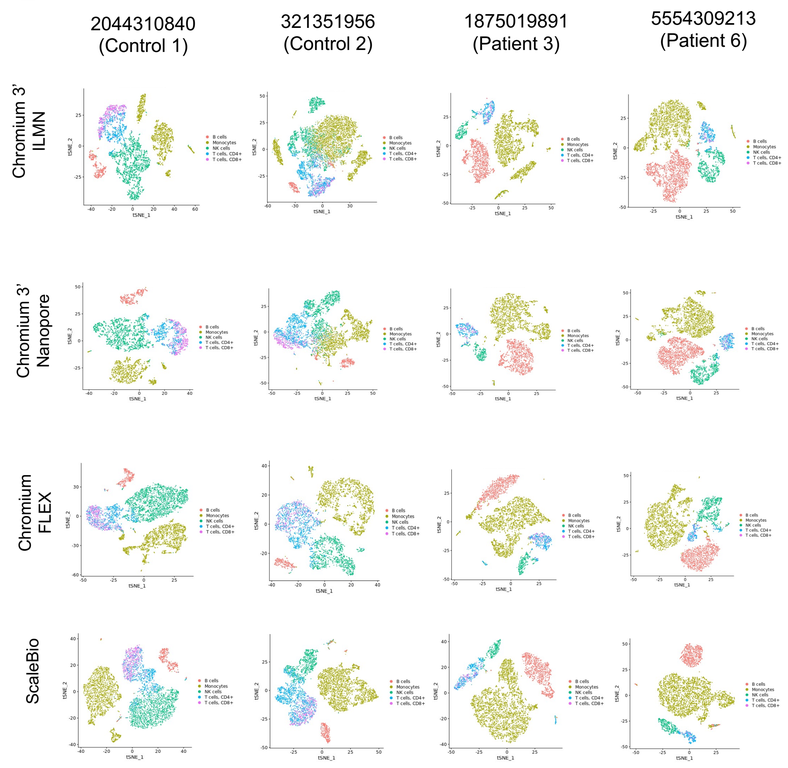
Single cell transcriptomics, or scRNASeq, has experienced transformative progress especially within the last ~5 years, in large part driven by the accessibility of droplet microfluidics technology to partition each cell robustly into uniquely barcoded, individual gel-beads and encapsulating the cell/bead pair in an emulsion containing the requisite cDNA-generating reagents. The impact of scRNASeq on basic and translational studies in genomics and cellular biology has been exponential. However, a microfluidics platform requires the use of separate, specialized instrumentation which can be a bottleneck from both cost and technical perspectives. Recently, alternative scRNASeq platforms have become commercially available based on either successive split-pool barcoding techniques or proprietary emulsion reagents that generate single cell/hydrogel droplets via self-assembly by mixing or vortexing. We performed scRNASeq using the following assays/platforms: 1) Parse Biosciences split-pool barcoding assay, 2) Scale Biosciences split-pool barcoding assay, 3) Particle-templated instant partition sequencing (PIPSeq) from Fluent BioSciences, 4) 10x Genomics (10xG) Chromium single-cell 3' gene-expression assay, v3.1, and 5) 10xG Chromium single-cell flex assay (v1). About 2,500-5,000 cells from 4 peripheral blood mononuclear cell (PBMC) samples derived from pediatric subjects with hypoplastic left heart syndrome (HLHS) were processed on each platform. Fixed cells were used for all assays except the standard 10xG 3' gene expression assay. The cDNA libraries generated from each type of assay were processed further according to manufacturers' instructions to generate Illumina-compatible sequencing libraries, then subsequently sequenced on the NovaSeq. For 4 of the 5 platforms, full-length cDNA libraries were generated, with the exception of the 10xG flex assay (which uses a probe-based method instead of priming and extension from 3' poly-A tails). Thus, an aliquot from each full-length cDNA library was also processed and sequenced on Oxford Nanopore to generate long-read transcriptome datasets. The quality of both the short- and long-read transcriptome datasets were analyzed and compared, using the 10xG 3' gene-expression assay as the benchmark. Additionally, the advantages and disadvantages/limitations of each platform with regards to processing and capture efficiency/input cell numbers were also assessed.

Barcode Rank "Knee" Plots across Platforms



t-SNE Clustering (Seurat)

Filter: 200 < Features < 5000; Mito content < 15%



QC Summary Table

Chromium 3' ILMN

Sample ID	Number of reads	Valid barcodes	Passing Sample Reads	Reads Mapped to Genome	Reads Mapped to Transcriptome	Exonic reads (of reads mapped to transcriptome)	Exonic Reads (of total mapped)	Sequencing saturation	Fraction of reads in cells	Mean reads per cell	UMI threshold	Initial Cell Input	Cells (Pre-filter)	Cells (Post-filter)	Median reads per cell	Median genes per cell	Total genes detected	Median UMI counts per cell	Seurat Features (Transcripts)	Total UMI count
2044310840	296,885,391	96.60%		93.50%	73.90%		54.20%	77.90%	92.80%	63,600	500	20,000	4,599	4,320		2,778	27,831	7,610	23,893	
321351956	372,301,311	96.10%		91.70%	70.20%		50.70%	75.00%	91.20%	30,607	500	20,000	12,012	11,629		975	28,818	1,434	25,192	
1875019891	313,271,568	96.40%		92.70%	74.90%		56.40%	71.70%	92.60%	51,542	500	20,000	5,834	5,518		2,723	27,735	8,376	23,731	
5554309213	308,786,488	95.90%		93.60%	73.00%		52.00%	65.70%	93.10%	37,145	500	20,000	8,304	7,970		2,597	28,788	6,893	25,190	
1550930418	296,930,531	97.10%		94.90%	74.10%		56.10%	85.10%	85.40%	57,212	500	20,000	4,067	3,889		1,158	26,516	2,687	22,016	
5257830638	278,244,030	97.30%		97.40%	79.50%		60.30%	78.00%	88.20%	33,120	500	20,000	8,170	7,775		1,723	27,340	4,246	23,308	

Chromium 3' Nanopore

Sample ID	Number of reads	Valid barcodes	Passing Sample Reads (>Q10)	Reads Mapped to Genome	Reads Mapped to Transcriptome	Exonic reads (of reads mapped to transcriptome)	Exonic Reads (of total mapped)	Sequencing saturation	Fraction of reads in cells	Mean reads per cell	UMI threshold	Initial Cell Input	Cells (Pre-filter)	Cells (Post-filter)	Median reads per cell	Median genes per cell	Total genes detected	Median UMI counts per cell	Seurat Features (Transcripts)	Total UMI count
2044310840	135,840,519		134,393,198	45.61%	23.20%			43.00%	56.16%	17,457	1,483	20,000	4,269	4,103	15,304	2,582	24,751	8,489	22,359	68,064
321351956	108,458,093		107,213,867	41.46%	21.89%			32.00%	50.44%	7,624	862	20,000	7,039	6,898	7,027	1,752	23,809	4,704	21,885	65,180
1875019891	124,891,815		123,276,129	50.00%	28.02%			36.00%	60.60%	13,576	1,010	20,000	5,186	5,011	12,356	2,398	23,809	7,726	21,532	65,180
5554309213	134,010,609		132,392,128	48.36%	23.08%			30.00%	59.35%	9,868	884	20,000	7,988	7,792	8,879	2,177	25,135	6,098	22,983	68,813

Chromium FLEX

Sample ID	Number of reads	Valid barcodes	Passing Sample Reads	Reads Mapped to Genome	Reads Mapped to Transcriptome	Exonic reads (of reads mapped to transcriptome)	Exonic Reads (of total mapped)	Sequencing saturation	Confidently mapped reads in cells	Mean reads per cell	UMI threshold	Initial Cell Input	Cells (Pre-filter)	Cells (Post-filter)	Median reads per cell	Median genes per cell	Total genes detected	Median UMI counts per cell	Seurat Features (Transcripts)	Total UMI count
2044310840	192,387,990	98.45%	96,581,171					72.49%	93.47%	24,837	500	20,000	7,746	5,886	13,098	1,813	14,712	3,160	13,625	
321351956	186,299,564	98.01%	92,796,652					85.81%	83.27%	37,904	500	20,000	4,915	4,829	12,265	933	13,881	1,319	12,741	
1875019891	179,638,060	97.90%	90,317,503					66.79%	60.22%	32,590	500	20,000	5,512	5,245	8,495	1,293	13,987	2,056	12,791	
5554309213	159,323,355	98.43%	79,188,721					67.42%	80.19%	21,703	500	20,000	7,341	5,668	9,835	1,572	14,839	2,704	13,520	
6282850618	158,512,001	98.03%	79,681,958					78.96%	77.31%	87,190	500	20,000	1,818	1,024	28,463	2,280	13,830	5,248	12,636	
8458411467	165,564,538	98.48%	83,255,326					76.68%	90.86%	29,607	500	20,000	5,592	4,259	14,316	1,670	14,499	2,880	13,436	

ScaleBio

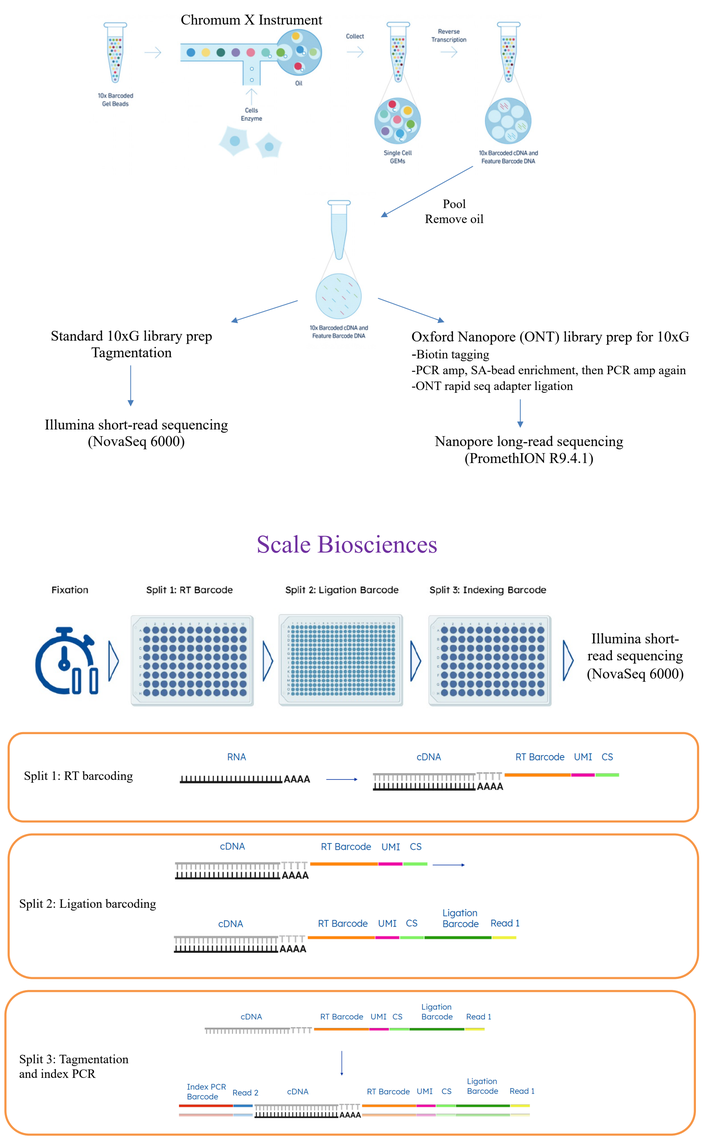
Sample ID	Total Sample Reads	Valid barcode fraction	Passing Sample Reads	Reads Mapped to Genome	Reads Mapped to Transcriptome	Exonic reads (of reads mapped to transcriptome)	Exonic Reads (of total mapped)	Saturation	(Fraction of) reads in cells	Mean passing reads per cell	Unique Transcript Counts Threshold	Initial Cell Input	Cells above Threshold	Cells (Post-filter)	Median reads per cell	Median genes per cell	Total genes detected	Median UTC per cell	Seurat Features (Transcripts)	Total UTC
2044310840	231,984,784		216,155,464	92.10%	78.40%	27.81%	21.80%	84.00%	85.90%	27,470	776	19,250	6,592	5,788		1,418		2,092	29,598	
321351956	146,928,426		135,829,557	90.20%	77.90%	25.80%	20.10%	84.00%	84.20%	17,228	535	19,250	6,396	6,238		1,039		1,381	28,355	
1875019891	396,519,379		368,747,401	93.30%	83.00%	30.12%	25.00%	84.00%	89.10%	48,786	1,171	19,250	6,643	4,062		2,278		4,731	30,608	
8458411467	161,023,185		149,047,390	91.00%	74.80%	28.48%	21.30%	84.00%	83.70%	24,846	663	19,250	4,825	4,458		1,381		1,972	29,241	
6282850618	73,817,865		68,179,368	92.10%	77.30%	31.05%	24.00%	84.00%	83.10%	43,389	1,104	19,250	1,214	869		2,063		4,037	24,096	
5554309213	291,772,900		272,262,607	92.00%	73.20%	27.32%	20.00%	84.00%	84.80%	36,610	844	19,250	6,160	5,279		1,631		2,698	30,954	
2044310840	106,028,395		98,436,749	91.20%	77.40%	24.16%	18.70%	84.00%	83.30%	13,372	416	19,250	5,813	5,772		884		1,109	26,850	
321351956	107,070,159		98,429,739	90.40%	75.60%	26.72%	20.20%	83.00%	80.00%	13,881	441	19,250	5,350	5,286		897		1,121	27,256	

ParseBio

Sample ID	Total Reads	Valid barcode fraction	Passing Sample Reads	Reads Mapped to Genome	transcriptome map fraction	Exonic reads (of reads mapped to transcriptome)	Exonic Reads (of total)	Saturation	Fraction reads in cells	mean reads per cell	Cell transcript cutoff	Initial Cell Input	Number of cells	Cells (Post-filter)	Median reads per cell	median genes per cell		median transcripts per cell	Seurat Features (Transcripts)	Number of transcripts
2044310840	13360519.68	71.49%			64.70%	30.24%	19.57%	61.03%	41.82%	21376.83	418	1,668	625	623		689		923	12047	2355163
321351956	16329900.98	71.46%			64.70%	36.19%	23.41%	60.18%	33.99%	3145.81	292	1,668	681	516		128		141	9384	2878599
1875019891	343738289.2	71.46%			64.70%	35.33%	22.86%	62.24%	65.50%	70713.49	1973	7,500	4,861	4,834		2382		5304	28255	60593429
5554309213	163957215.2	71.45%			64.71%	37.50%	24.27%	61.98%	65.35%	33002.66	903	7,500	4,968	4,958		1433.5		2407	25105	28902017
Parse control	238821934.9	71.47%			64.70%	37.34%	24.16%	61.42%	68.55%	143007.15	4519	1,668	1,670	1,406		3918.5		10643.5	27876	42099005
Pool (1+2)	*776207860*	*71.46%*			*64.70%*	*36.34%*	*23.51%*	*61.88%*	*65.36%*	*44828.64*	*2469*	*20,004*	*12,805*	*12,337*		*1475*		*2497*		*136828213*

Platforms

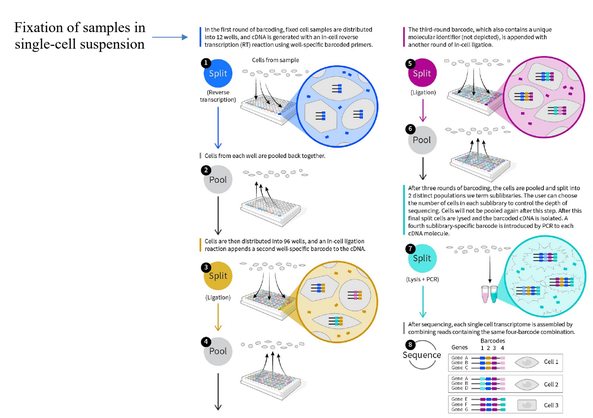
10x Genomics Chromium



Scale Biosciences

Platforms

Parse Biosciences (Mini kit)

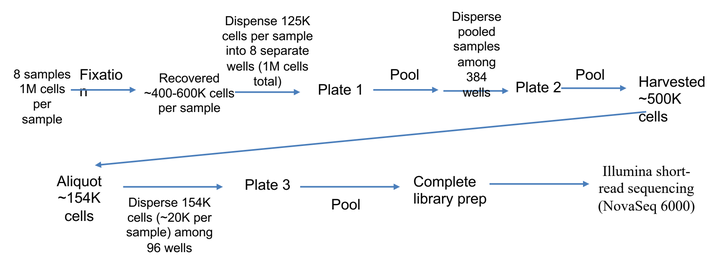


Fluent Biosciences (PIPSeq)

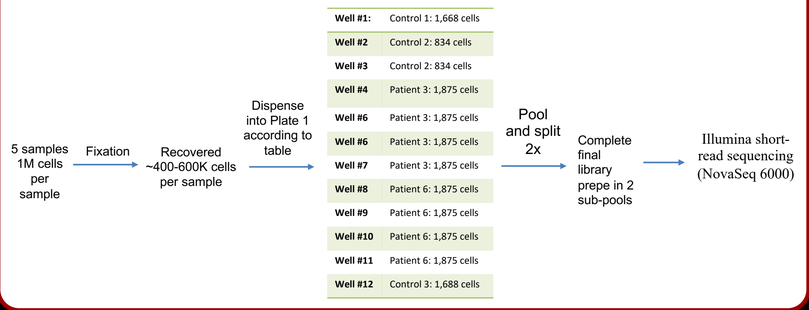
Finding: Unable to generate usable library on any of the 10 different PBMC samples assayed, thus this platform was not pursued further at this time.

Cell Distribution

Scale Biosciences

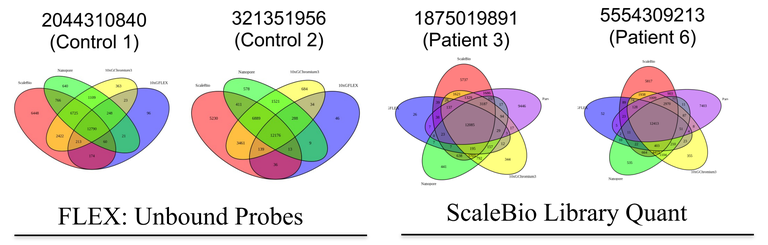


Parse Biosciences

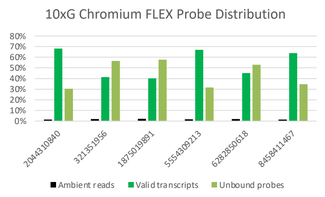


Well #1:	Control 1: 1,668 cells
Well #2	Control 2: 834 cells
Well #3	Control 2: 834 cells
Well #4	Patient 3: 1,875 cells
Well #6	Patient 3: 1,875 cells
Well #6	Patient 3: 1,875 cells
Well #7	Patient 3: 1,875 cells
Well #8	Patient 6: 1,875 cells
Well #9	Patient 6: 1,875 cells
Well #10	Patient 6: 1,875 cells
Well #11	Patient 6: 1,875 cells
Well #12	Control 3: 1,688 cells

Transcripts/Features Overlap

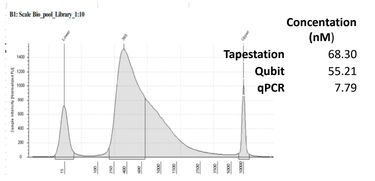


FLEX: Unbound Probes



Finding: Significant amount of unbound probes are sequenced, which severely dampens amount of features (genes/transcripts) detected in cells

ScaleBio Library Quant



Finding: The amount of Illumina-compatible material in ScaleBio library is a minor fraction of the total material generated in the library pool.

Mitochondrial Content

	10xG Chromium 3' ILMN Cell count			ScaleBio Cell count		
Sample ID	No mito filter	<1% mito content	Retention rate	No mito filter	<1% mito content	Retention rate
2044310840	4599	184	**4.00%**	6,592	6578	**99.79%**
321351956	12012	260	**2.16%**	6,396	6395	**99.98%**
1875019891	5834	368	**6.31%**	6,643	6625	**99.73%**
5554309213	8304	114	**1.37%**	6,160	6156	**99.94%**

Finding: Proportion of mitochondrial content in cells prepared via ScaleBio is dramatically lower compared to that found in that generated using 10xG Chromium 3' assay.

Summary and Future Directions

- PBMC stocks used in this comparative study were likely more challenging and lower quality than commercially available reference samples.
- PIPSeq assay from Fluent Biosciences were unable to generate usable library from our PBMC samples (10 different samples failed).
- ParseBio assay also had difficulty with two PBMC samples loaded at lower initial number, but successfully generated single-cell data for other 2 samples.
- ScaleBio assay generated final library material of which only a fraction (~14%) were Ilumina-compatible content; however single-cell data were generated successfully in all samples from this library.
- 10xG Chromium FLEX assay detected significantly lower number of genes/transcripts compared to other platforms, most likely due to the large fraction of unbound probes retained within the library.
- Mitochondrial content in cells detected in ScaleBio library is noticeably lower than that from standard 10xG Chromium 3' assay.
- Usable single-cell datasets from samples from ScaleBio and ParseBio seem to capture a significant number of features (i.e. genes/transcripts) that were not found in 10xG Chromium 3' (~6000-9000 more in ScaleBio/ParseBio compared to 10xG. The identity of these features will be examined further.
- Analyses of isoform distribution in long-read Nanopore data (from 10xG Chromium 3' cDNA library) is ongoing.