Analyzing BioRad-Illumina Single Cell RNA-Seq data with open source tools

Lucas Moron Dalla Tor[1], Cristina Patuzzo[1], Michela Deiana[1-3], Samuele Cheri[1-3], Laura Veschetti[1], Monica Castellucci[2], Francesca Griggio[2], Maria Teresa Valenti[3], Elisabetta Trabetti[1], Giovanni Malerba[1]

[1]Lab. Computazionale di Genetica Medica e Molecolare, Sez. Biologia e Genetica, Università degli studi di Verona, Verona
[2]Centro Piattaforme Tecnologiche, Università degli studi di Verona, Verona
[3]Dip. Medicina, Sez. Medicina Interna D, Università degli studi di Verona, Verona

lucas.morondallator@univr.it

Background: Single cell RNA-Seq is a powerful technique that is becoming more popular since it enables to sequence the transcriptome of each cell within a population of different cell types in a single experiment. Currently, there are a few different technologies, like BioRad-Illumina ddSeq and 10X Chromium.

Methods: We studied 6 human PBMC samples (for the purpose of this poster we are going to show the best ones) using ddSeq and sequencing on NextSeq500 (SureCell WTA 3' protocol), generating 30M reads/sample, capturing from 300 to 2000 cells/sample and counting ~5000 transcripts/cell. In this paired end based method the first mate contains a cellular and molecular barcode, while the second contains the 3' portion of the transcript. The amount of valid cell/transcript barcodes, the efficiency of mapping and gene annotation were estimated and then compared using the open source (DST) DropSeqTools together with ddSeeker, and the pay-for-use Illumina BaseSpace (BS) Cloud. Differential expression of cell markers and cluster analyses were performed with the Seurat R package fine tuning every parameter in order to find the better set of values for each sample. Our goal is to investigate the versatility in problem-backtracking as well as the differences in results between DST and BS.

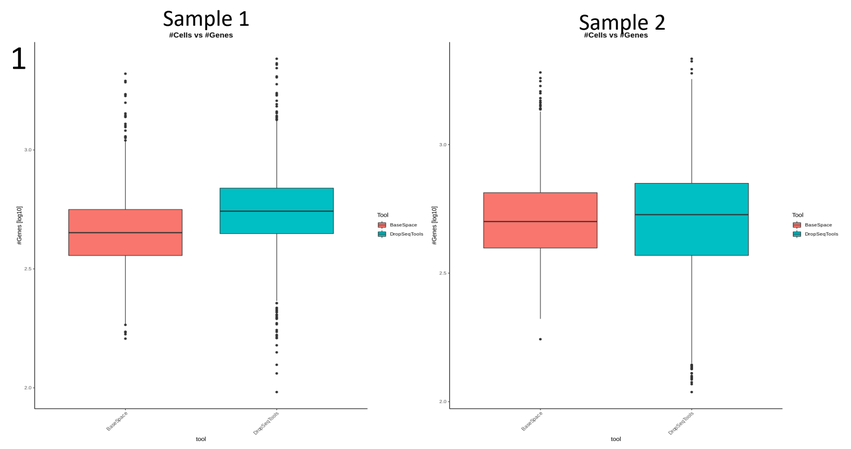

Figure 1: Comparison between the number of cells and transcripts for the 2 best samples, comparing the two analysis strategies.

Results: The two methods showed some differences in the workflows such as the mapping and barcode calling strategies. The BS platform performs a read filtering step that should enhance the outcome of the analysis even though the type of filter used and its parameters are unknown. The preprocessing step performed using ddSeeker leads to a higher number of recovered barcodes (from 77% by BS to 81% by DST), while after the mapping and refinement steps, the number of reads suitable for differential gene expression is slightly higher in BS (25M reads) than in DST (23M reads). The DST tools detected an overall higher number of cells with at least 100 non-duplicate transcripts each, some of them were filtered out during quality controls in R. DST tool results also give better clustering than BS, e.g. Naive, Memory and transitioning CD4+ T cells were identified as 3 different clusters in Sample 2 (Table 3 and Figure 3) only using DST tools thus suggesting some kind of information loss during the BS preprocessing.

1

	DST			BS		
	Cluster	#Cells	Relative %	Cluster	#Cells	Relative%
Sample 1	0	332	31.43939	0	332	30.74074
	1	256	24.24242	1	263	24.35185
	2	248	23.48485	2	159	14.72222
	3	96	9.090909	3	99	9.166667
	4	88	8.333333	4	94	8.703704
	5	36	3.409091	5	94	8.703704
	Total	**1056**		6	39	3.611111
				Total	**1080**	
Sample 2	Cluster	#Cells	Relative %	Cluster	#Cells	Relative%
	0	501	28.72706	0	360	21.17647
	1	338	19.38073	1	329	19.35294
	2	259	14.85092	2	295	17.35294
	3	177	10.14908	3	233	13.70588
	4	143	8.199541	4	178	10.47059
	5	142	8.142202	5	155	9.117647
	6	117	6.708716	6	75	4.411765
	7	36	2.06422	7	37	2.176471
	8	19	1.08945	8	19	1.117647
	9	12	0.6880734	9	19	1.117647
	Total	**1744**		**Total**	**1700**	

Table 1: Number of cell in each cluster with the relative proportion of the 2 best samples. DST and BS columns indicate data coming from the respective tools.

2

DST Cluster IDs	BS Cluster IDs	p-value adj DST	p-value adj BS	Markers	Cell Type
0	1	0.035882585	1	IL7R	T Cells
1	2,3	1.3871E-155	4.13575E-51	NKG7	CD8+ T Cells
			1.48627E-83		
1	2,3	6.0968E-142	1.06231E-46	GNLY	
			3.53138E-83		
1	2,3	1.8015E-140	2.86522E-55	CCL5	
			5.39191E-58		
1	2,3	9.22378E-24	1.27162E-10	CD8A	
			5.02965E-06		
2	0	1.3388E-14	3.84607E-25	IL7R	Naive CD4+ T Cells
2	NA	2.1582E-08	NA	TRAC	
2	0	3.24301E-05	1.22343E-06	CD3E	
2	0	0.000350304	8.21059E-05	CCR7	
3	5	1.7967E-171	5.2161E-186	LYZ	CD14+ Monocytes
3	5	8.5786E-124	1.8648E-129	CD14	
4	4	1.45208E-92	1.97773E-91	CD79A	B Cells
4	4	7.62722E-50	1.93243E-54	MS4A1	
4	4	1.50585E-44	2.96099E-48	CD79B	
5	6	2.69722E-55	1.19451E-62	GP9	Megakaryocytes
5	6	2.64246E-49	5.14632E-78	PPBP	
5	6	5.37199E-38	9.38847E-66	PF4	

3

DST Cluster IDs	BS Cluster IDs	p-value adj DST	p-value adj BS	Markers	Cell Type
0	NA	1.81573E-15	NA	TRAC	Naive CD4+ T Cells
0	NA	2.79065E-13	NA	IL7R	
0	NA	4.87658E-12	NA	CD3D	
0	NA	0.000172775	NA	CCR7	
0	NA	0.012077819	NA	MAL	
1	1	1.2331E-174	5.5785E-169	GNLY	CD8+ T Cells
1	1	2.4102E-144	3.8839E-129	NKG7	
1	1	2.1237E-126	2.5613E-126	CCL5	
1	1	1.67756E-11	NA	TRDC	
1	1	7.83517E-05	1.45408E-05	CD8A	
2	3	9.4953E-182	7.2394E-179	LYZ	CD14+ Monocytes
2	3	7.9132E-179	1.7759E-170	CD14	
3	4	4.8059E-104	3.45819E-86	CD79A	B Cells
3	4	2.67087E-93	8.67126E-92	MS4A1	
3	4	5.50036E-76	5.73677E-78	CD79B	
4	5	3.74129E-87	7.46104E-85	PPBP	Megakaryocytes
4	5	7.49872E-77	5.92584E-79	GP9	
4	5	5.87628E-72	2.83449E-69	PF4	
5	NA	7.65481E-13	NA	IL7R	Mature CD4+ T cells
5	NA	7.35555E-11	NA	TRAC	
5	NA	1.68796E-05	NA	S100A4	
6	0,6	3.12495E-19	7.58736E-47	IL7R	CD4+ T Cells (in maturation)
			7.72825E-11		
6	0	6.91608E-14	NA	CD3G	
6	0	NA	8.1017E-15	CD3D	
6	0	2.99216E-11	NA	TRAC	
6	0,6	3.341E-11	0.00095616	THEMIS	
			9.41038E-05		
6	0,6	5.96813E-09	1.93761E-06	ITK	
			0.000532456		
6	0	7.85499E-09	3.2266E-11	ICOS	
7	7	2.4972E-99	5.19021E-86	MS4A7	CD16+ Monocytes
7	7	8.1411E-65	1.07247E-64	FCGR3A	
8	8	7.7383E-237	6.7282E-248	FCER1A	Monocyte derived DCs
8	8	1.58686E-29	2.45356E-31	CST3	
9	9	3.14515E-81	4.80531E-32	IL3RA	Plasmacytoid DCs
9	9	2.54444E-22	NA	NRP1	
9	9	9.20627E-11	0.000501666	GZMB	
NA	2	NA	0.235469177	CD3D	Unidentified T Cells

Conclusions: Being a proprietary platform, BS doesn't have the same versatility and transparency of the DST solution, which allows the user to better set the parameters according to the biological sample. In addition to this, DST shows a higher capability of retaining information than BS (as shown in Table 2,3 and Figures 2,3) hence giving better clustering results In conclusion, the DST approach could be a valid alternative to the pay-for-use method for bioinformaticians using ddSeq.

Table 2 & 3: Markers tables with p-value (Bonferroni adjusted) which indicates the probability of the marker to be a true marker for that cluster, used to identify cluster cell types In Sample 1 (2) and Sample 2 (3). NA is used when either the marker or cluster can't be found with the other method. In red it is also possible to see the only available marker for BS Cluster #2 (3) is not statistically significant.

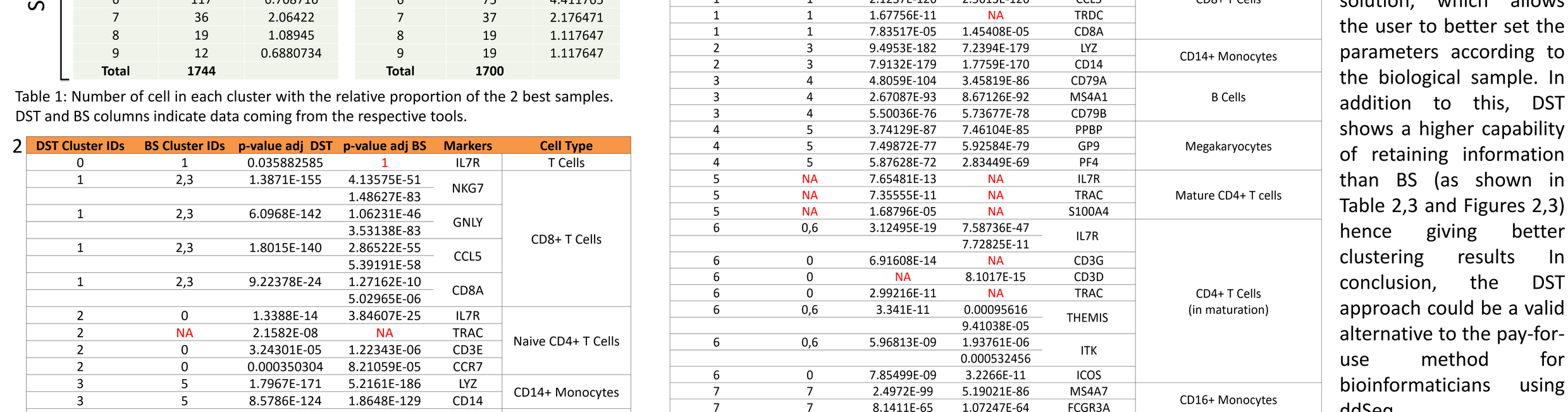

Figure 2 & 3: Uniform Manifold Approximation and Projection clustering plot (UMAP) for Sample 1 (2) and Sample 2 (3). The left side of both plots shows results from DST while the right side shows results form BS. Cell types were assigned using markers from Table 2 and 3.

Romagnoli D. et al. ., *ddSeeker: a tool for processing Bio-Rad ddSEQ single cell RNA-seq data.* BMC Genomics. 2018; 19:960
Macosko *et al*., Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 2015.
Satija R. et al., A toolkit for quality control, analysis, and exploration of single cell RNA sequencing data. *Nature Biotechnology* volume33, pages495–502 (2015)

https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/surecell/surecell-wta3-library-prep-reference-guide-1000000021452-01.pdf