SCIENCE MEETS LIFE

19-20 November 2020

Emerging Technologies in Single Cell Research

Benchmarking of single cell ATAC-seq technologies yields insight into reproducibility

Christopher Flerin[1,2], Albert Rafels Ybern[4], Paula Soler Vila[3], Florian de Rop[1,2], Stein Aerts[1,2,#], Holger Heyn[4,#]

Abstract: We compare scATAC-seq experimental methods by examining differences across technologies, replicates, and sequencing facilities. We use a common biological sample, peripheral blood mononuclear cells (PBMCs), pooled in equal ratio from two donors. On this sample, five centers worldwide have performed scATAC-seq. We find evidence for variation in data quality due to a number of factors, including the scATAC-seq technology used, preparation of the sequencing libraries, and total sequencing depth.

Affiliations:

1 VIB Center for Brain & Disease Research, KU Leuven, Belgium

2 Department of Human Genetics KU Leuven, Belgium

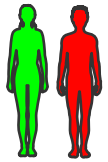
3 Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

4 CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain

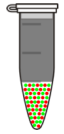
Corresponding authors

Unified bioinformatics pipeline

With each sequencing technology having its own processing tool (10x – cellranger; BioRad – BAP, etc.), we process each sample using the same steps for a fair comparison. This pipeline is implemented in Nextflow with software packaged in Docker containers for maximum reproducibility.



PBMCs: 50% mixture of donor cells. Sequenced in duplicate



Centers and Technologies

Broad Institute	ddSEQ mtDNA Genotyping with ATAC-Seq (10x v1.1)
Stanford	10x Genomics (v1.0)
VIB	10x Genomics (v1.0)
CNAG-CRG	10x Genomics (v1.1) scRNA-seq (10x Genomics)
University of Washington	sciATAC
Wellcome Sanger Institute	10x multiome

FASTQ

Preprocessing
- Barcode correction
- Mapping
- Duplicate marking

BAM
- Downsampling

QC & Filtering
- Cell calling
- Doublet removal

Peak matrix generation

Features-by-cells matrix

Dimensionality reduction
- Batch effect correction
- Batch effect evaluation

Batch-corrected dimensionality reduction

Call DARs

TFs

GRN

Peak-to-gene associations

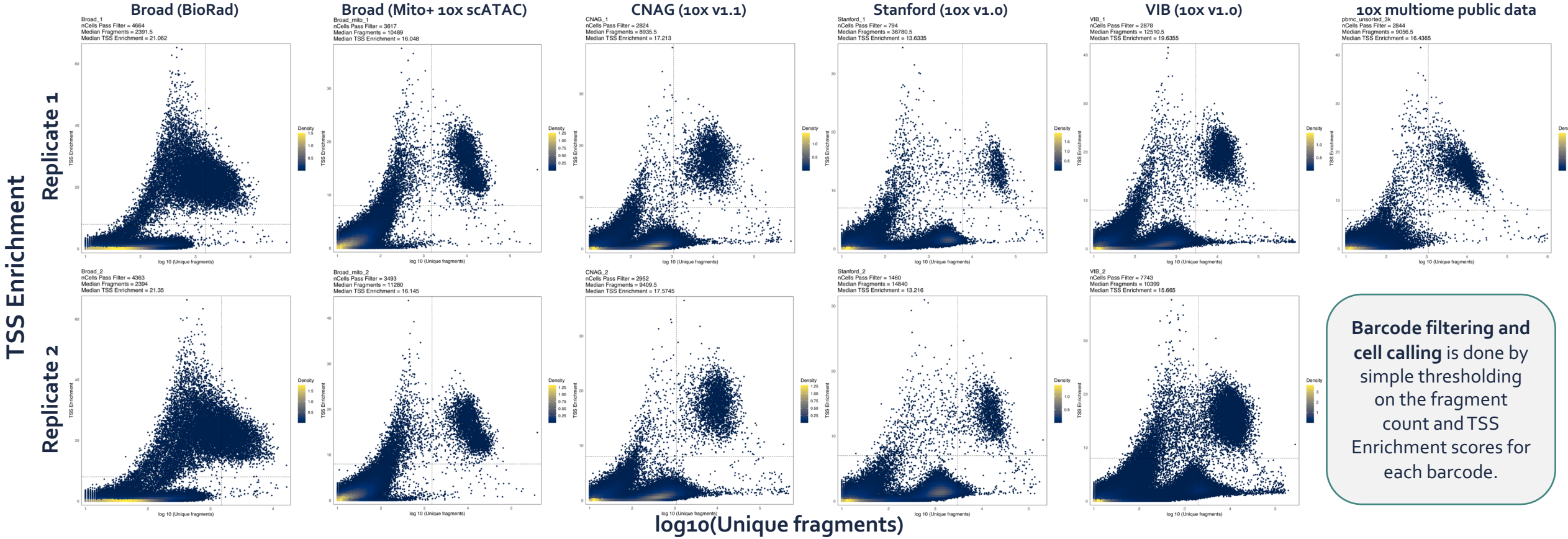
Clustering
- Cell type identification

Clustered & identified cells

Peak calling

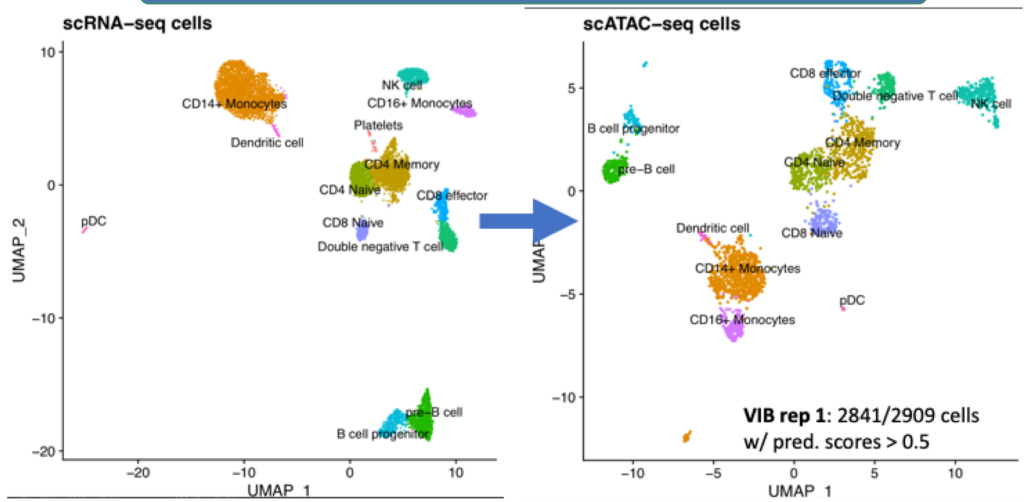
Consensus peaks

SCIENCE MEETS LIFE

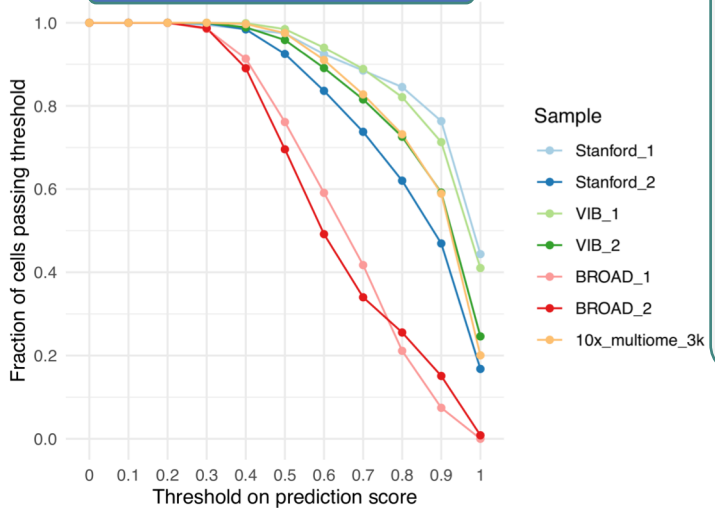


Barcode filtering and cell calling is done by simple thresholding on the fragment count and TSS Enrichment scores for each barcode.

Cell type identification (scRNA-seq label transfer)



Cell quality



Cell types are identified by performing label transfer (via Seurat) using well-annotated scRNA-seq data from the same samples. Cell quality can be inferred from the label transfer prediction scores.

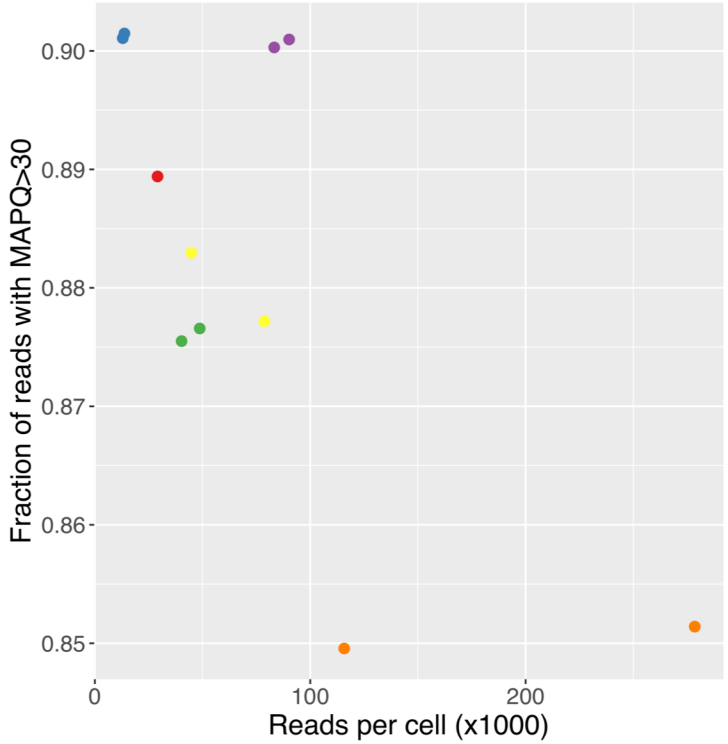
VIB

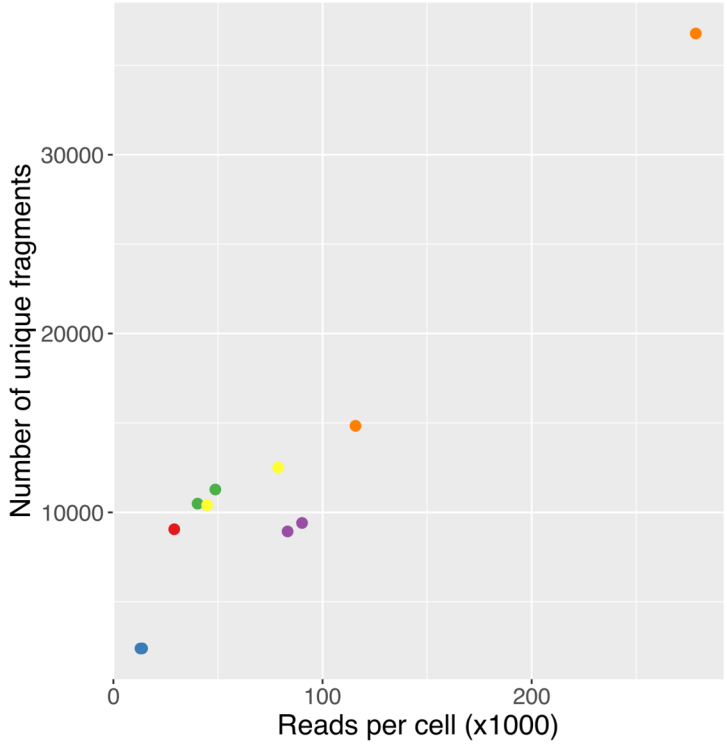
SCIENCE MEETS LIFE

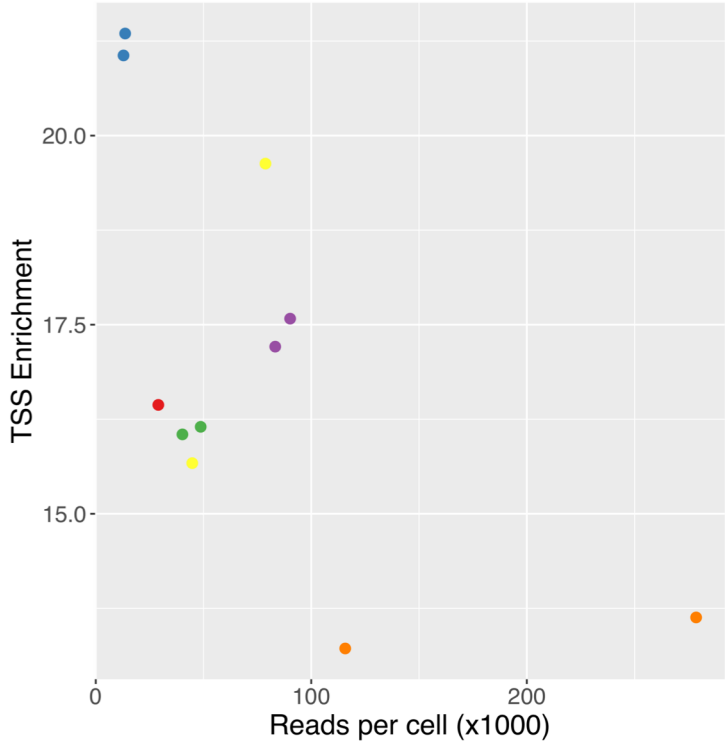
Mapping statistics

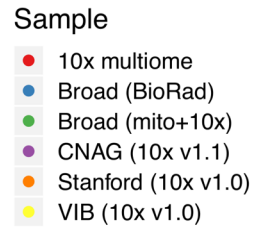
	Technology	Total paired reads	Reads mapped and paired		Uniquely mapped reads		Fraction of reads mapped with MAPQ>30		# of cells	reads/cell	n_fragments	TSS Enrichment
Broad_1	BioRad	60,356,415	59,427,509	98.46%	55,110,049	91.31%	54,385,722	90.11%	4664	12,941	2391	21.06
Broad_2	BioRad	59,794,636	58,940,653	98.57%	54,606,743	91.32%	53,903,092	90.15%	4363	13,705	2394	21.35
Broad_mito_1	mito-scATAC-seq	145,604,100	142,755,255	98.04%	115,271,081	79.17%	127,476,589	87.55%	3617	40,255	10489	16.05
Broad_mito_2	mito-scATAC-seq	170,196,511	167,071,195	98.16%	134,662,744	79.12%	149,189,003	87.66%	3493	48,725	11280	16.15
CNAG_1	10x v1.1	235,210,366	232,977,829	99.05%	214,799,863	91.32%	211,758,118	90.03%	2824	83,290	8935	17.21
CNAG_2	10x v1.1	266,303,672	263,307,982	98.88%	243,390,399	91.40%	239,931,050	90.10%	2952	90,211	9409	17.58
pbmc_unsorted_3k	10x multiome	82,709,927	81,120,171	98.08%	74,050,662	89.53%	73,562,521	88.94%	2844	29,082	9056	16.44
Stanford_1	10x v1.0	221,145,825	214,403,418	96.95%	196,393,985	88.81%	188,284,784	85.14%	794	278,521	36780	13.63
Stanford_2	10x v1.0	169,040,262	164,589,008	97.37%	150,630,809	89.11%	143,610,491	84.96%	1460	115,781	14840	13.22
VIB_1	10x v1.0	226,657,677	220,278,518	97.19%	203,864,739	89.94%	198,820,185	87.72%	2878	78,755	12510	19.63
VIB_2	10x v1.0	346,959,872	340,895,633	98.25%	314,574,497	90.67%	306,352,370	88.30%	7743	44,809	10399	15.67

Conclusions: Mapping quality and post-processed data quality varies by the scATAC-seq technology used. This is further affected by library preparation, total sequencing depth and additional sample handling steps taken. The most variability is observed between sequencing centers. Preprocessing steps are critical in resolving issues specific to the technology (duplicate reads, doublets, filtering), and producing high-quality usable data for downstream analysis.









SCIENCE MEETS LIFE