Analytic Validation: ATAC epigenome profiling

Performance Lab:
Bendall Lab,
Stanford University, Department of Pathology.
3373 Hillview Ave. Palo Alto, CA. 94304

Table 1. Summary of analytical validation findings for ATAC profiling

Accuracy	Equivalent to standard bulk ATAC (Figure 1)
Precision: Inter-assay	Peak intensities between technical replicates (n=3) have a correlation R value above 0.95 (Buenrostro et. al., 2013)
Precision: Intra-assay	ATAC readout should significantly correlate with and represent relative levels of chromatin accessibility across the genome in a bulk population. This might be affected by stochastic differences in chromatin accessibility due to fixation. Biological reproducibility - correlation of read counts in naïve T cells between healthy donors is around r = 0.7-0.8 (Figure 3).
Analytical sensitivity	500 – 100K cells depending on collection and preparation method. Target is 50k FACS isolation (Corces *et. al.* 2017)
Analytical specificity including interfering substances	Specificity affected by Nextera kit transposase batch and fixation effects on transposase binding and subsequent activity
Reportable range	Relative chromatin accessibility over the whole genome as read out by level of transposition
Reference interval (normal range)	Normal range defined as accessibility profile of ENCODE cell line GM12878
Standardization, harmonization, reproducibility, and ruggedness	Standardized with 24-plex barcode for a pooled library on a paired end 76 bp HiSeq4000 sequencing lane to minimize technical variation. All analysis for a given sample set performed against the same reference DB.
Quality control and improvement procedures	Experiments carried out with 3 technical and ideally 2 biological replicates. ATAC sequencing data processing was done with the kundaje pipeline. Several QC checks were made such as raw fastq data with high quality stringency (>30) selected for downstream bowtie2 alignment and a 95% genome alignment standard (Figure 2). Final quality filtering will also be assessed through transcriptional start site score (TSS). Ideally, only samples with average TSS scores above 8 will be taken for further analysis.
Any other performance data	Accurate transposition determined by fragment size distribution using Bioanalyzer before sequencing (Figure 1).

Analytic Validation: ATAC epigenome profiling

Materials and methods for generation of the example validation data. Validation was carried out on GM12878 cell line and matched with publicly available chromatin accessibility information for this cell line. The detailed methodology is included in supplement.

Table 1. Antibodies used for intercellular FACS analysis

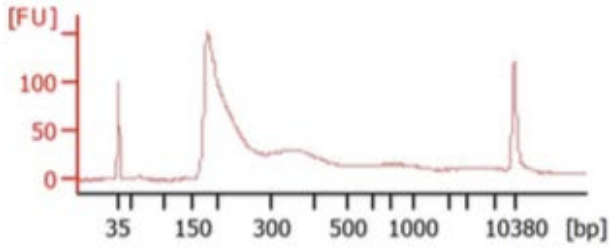
Antibody	Clone, Catalog number
H3K4me3	C42D8, CST#9751

5 million live cells were checked with tryphan blue for viability greater than 98% before 5min of fixation using the ebioscience intercellular staining kit. Cells were stained with antibody for 30min and were run through FACS (Table 1). Low and high populations were collected separately. We included a live cell control as well as a stained but not FACS-gated control.

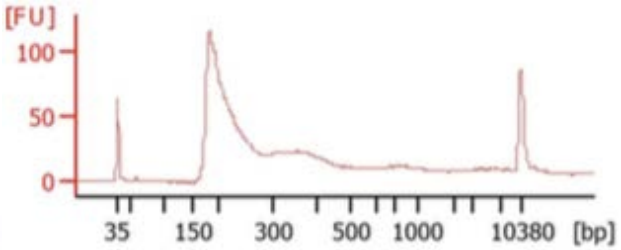
Cells were then lysed under standard ATAC protocol conditions (Supplement). Transposition was carried out with the Nextera kit for 30min at 37C on a shaking heating block. Upon transposition, reverse cross-linking buffer with proteinase K was added to the cells and was incubated at 65C for 6h. This step was optimized to ensure maximal fixation reversal with minimal DNA degradation. DNA was then extracted and amplified for 5 cycles under the PCR protocol for standard ATAC. qPCR was used to check for DNA products and calculate additional PCR cycle number. Total PCR cycles are to be kept under 15 to ensure minimal PCR bias. DNA was then purified to remove primer contamination and checked for fragment length distribution using Bioanalyzer (Figure 1).

DNA concentrations of samples were determined using Bioanalyzer but future efforts with include a more quantitative Qubit measurement. Samples were pooled into 1 library for paired end 76bp sequencing on a HiSeq4000 machine at the Stanford Functional Genomics Core. Data processing was carried out with the Kundaje pipeline on a server. A summary of the QC output shows the quality of the data obtained from the ATAC protocol (Figure 2).

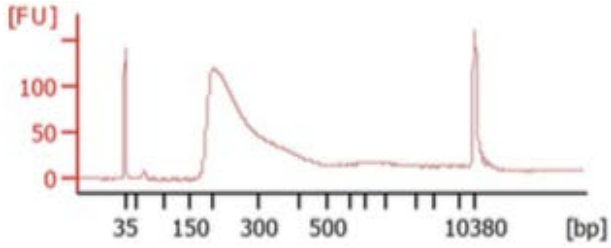
Analytic Validation: ATAC epigenome profiling



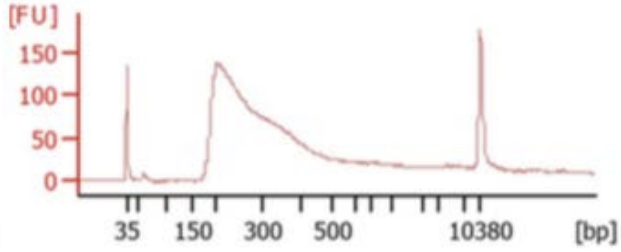
GM Live replicate 1



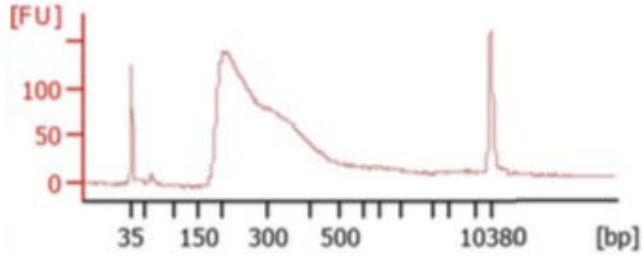
GM Live replicate 2



GM unsorted sample



GM high H3K4me3 sample



GM low H3K4me3 sample

Figure 1. Bioanalyzer peaks of fragment length distribution. GM cells were FACS-sorted for high and low abundance of H3k4me3. GM live cells with standard protocol, GM unsorted cells, GM High H3K4me3 and GM low H3K4me3 with inTAC protocol were analyzed for fragment length distribution using a bioanalyzer.

Accuracy. Accuracy and yield of inTAC protocol was confirmed by processing data from sequencing run. Kundaje pipeline was deployed on a server and run on 15 cores (cite). Quality control statistics of sequencing is shown in Figure 2.

Analytic Validation: ATAC epigenome profiling

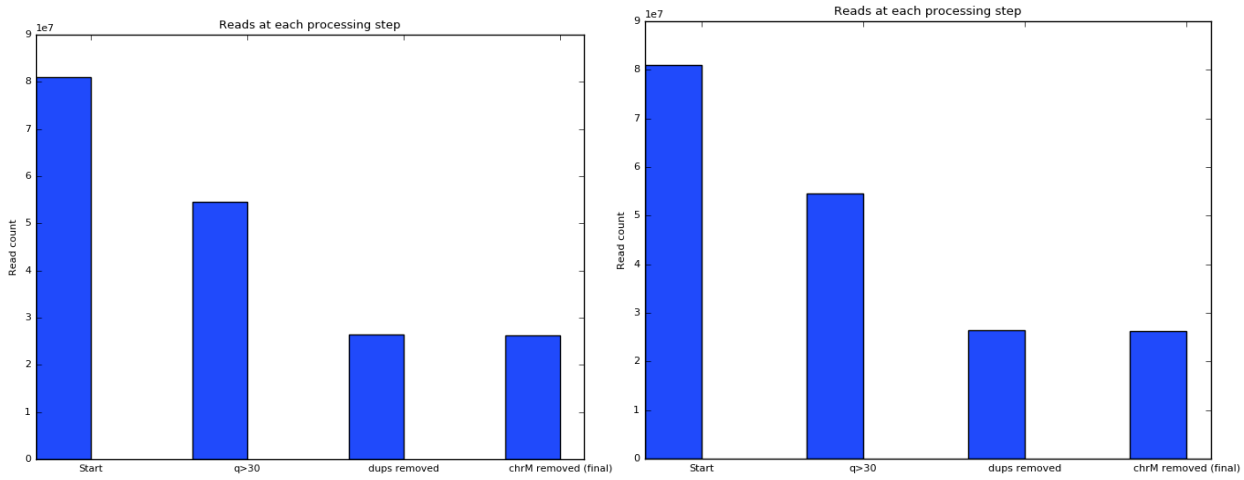


Figure 2A. Reads remaining after each filtering step comprising of quality score cutoff of 30, read duplication removal and blacklist DNA sequence on mitochondrial DNA removal. (QC plots for duplicate GM high H3k4me3 inTAC sequencing data)

Bowtie alignment log

```
40478682 reads; of these:
  40478682 (100.00%) were paired; of these:
    1899422 (4.69%) aligned concordantly 0 times
    12858473 (31.77%) aligned concordantly exactly 1 time
    25720787 (63.54%) aligned concordantly >1 times
    ----
    1899422 pairs aligned concordantly 0 times; of these:
      97331 (5.12%) aligned discordantly 1 time
    ----
    1802091 pairs aligned 0 times concordantly or discordantly; of these:
      3604182 mates make up the pairs; of these:
        3088214 (85.68%) aligned 0 times
        150014 (4.16%) aligned exactly 1 time
        365954 (10.15%) aligned >1 times
96.19% overall alignment rate
```

Figure 2B. Bowtie alignment statistics showing greater than 95% alignment to hg19. (QC plots for GM high H3k4me3 inTAC sequencing data)

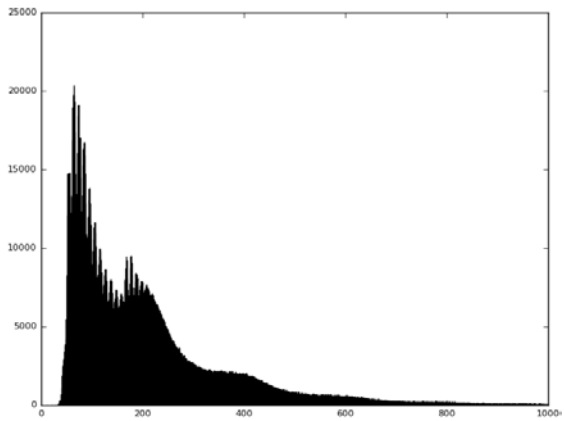


Figure 2C. Fragment length distribution calculated from paired end sequencing output with x axis as read counts and y axis as base pair length. (QC plots for GM high H3k4me3 inTAC sequencing data)

Analytic Validation: ATAC epigenome profiling

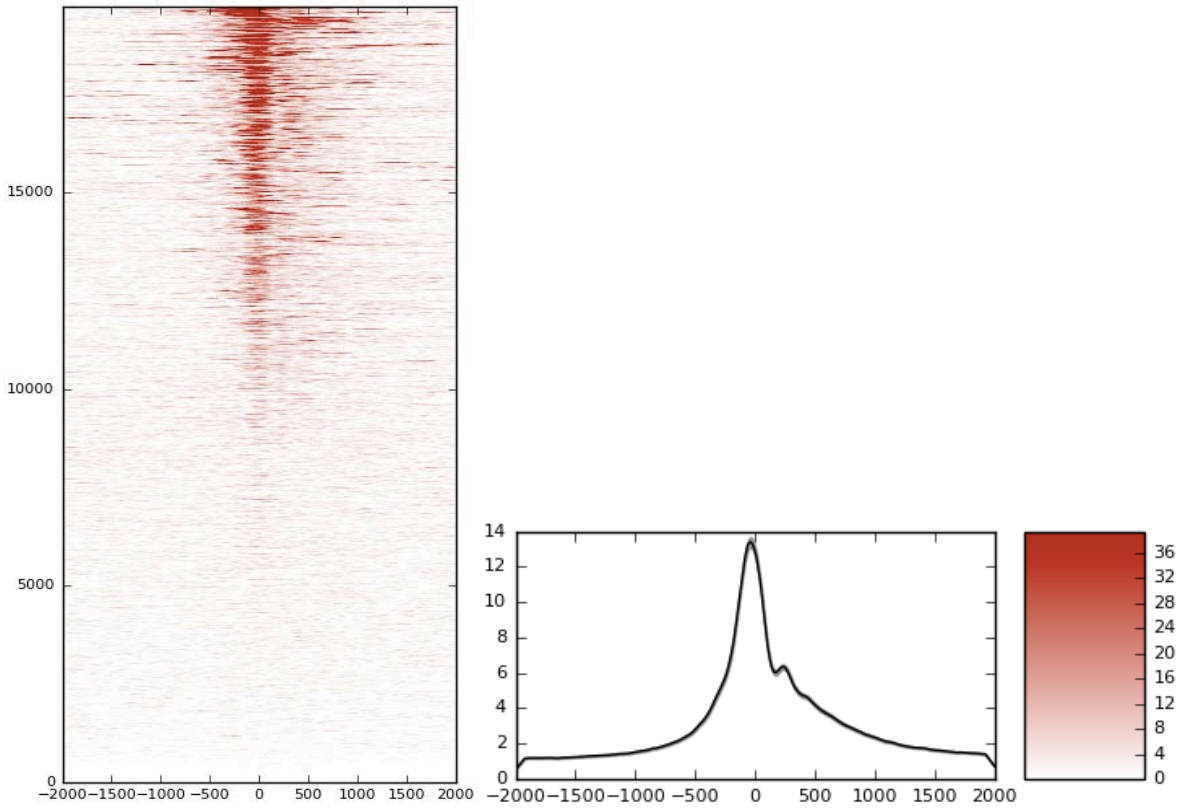


Figure 2D. Transcription start site (TSS) enrichment plot showing the density of reads around TSS visualized as base pair position 0 (Left). Smoothened histogram showing normalized read density at base pair position relative to TSS as 0. (QC plots for GM high H3k4me3 inTAC sequencing data)

Biological Reproducibility Between Healthy Donors:

Even though technical replicates routinely generate >95% concordance between abundance of called peaks we wanted to assess biological variability / reproducibility between similar cell types from different donors. Naïve T cells have a distinctive chromatin accessibility profile in healthy individuals. Here, we show they have 75-80% reproducibility across biological replicates, showing some variation between healthy individuals (Figure 3). Presumably, the variation here is due to fundamental differnces in the cellular composition of the input cell populations despite being based on the same Naïve T cell definition. There profiles should continue to diverge a cells differentiate into T cell subsets and also move into exhaustion. Moreover, this will be further perturbed if a patient's naïve T cells are transduced with CAR vector and proliferated for therapy. With these protocols as demonstrated we will be able to capture the changes in chromatin accessibility and better understanding of T cell responses and immunotherapy products for biomarker development and engineering efforts in the future.

Analytic Validation: ATAC epigenome profiling

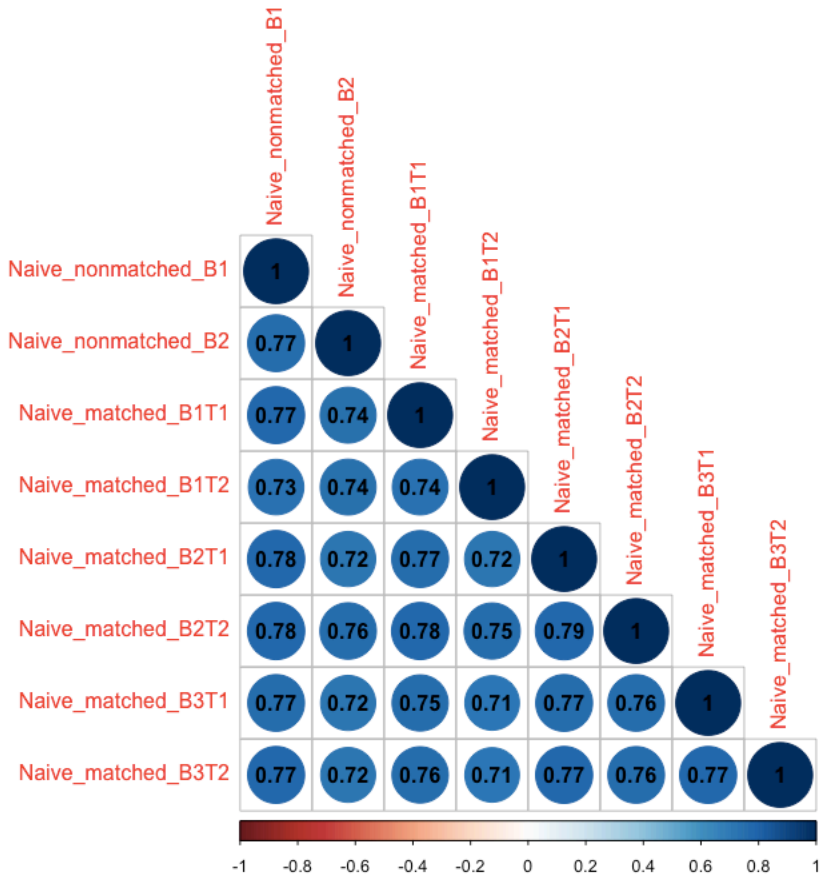


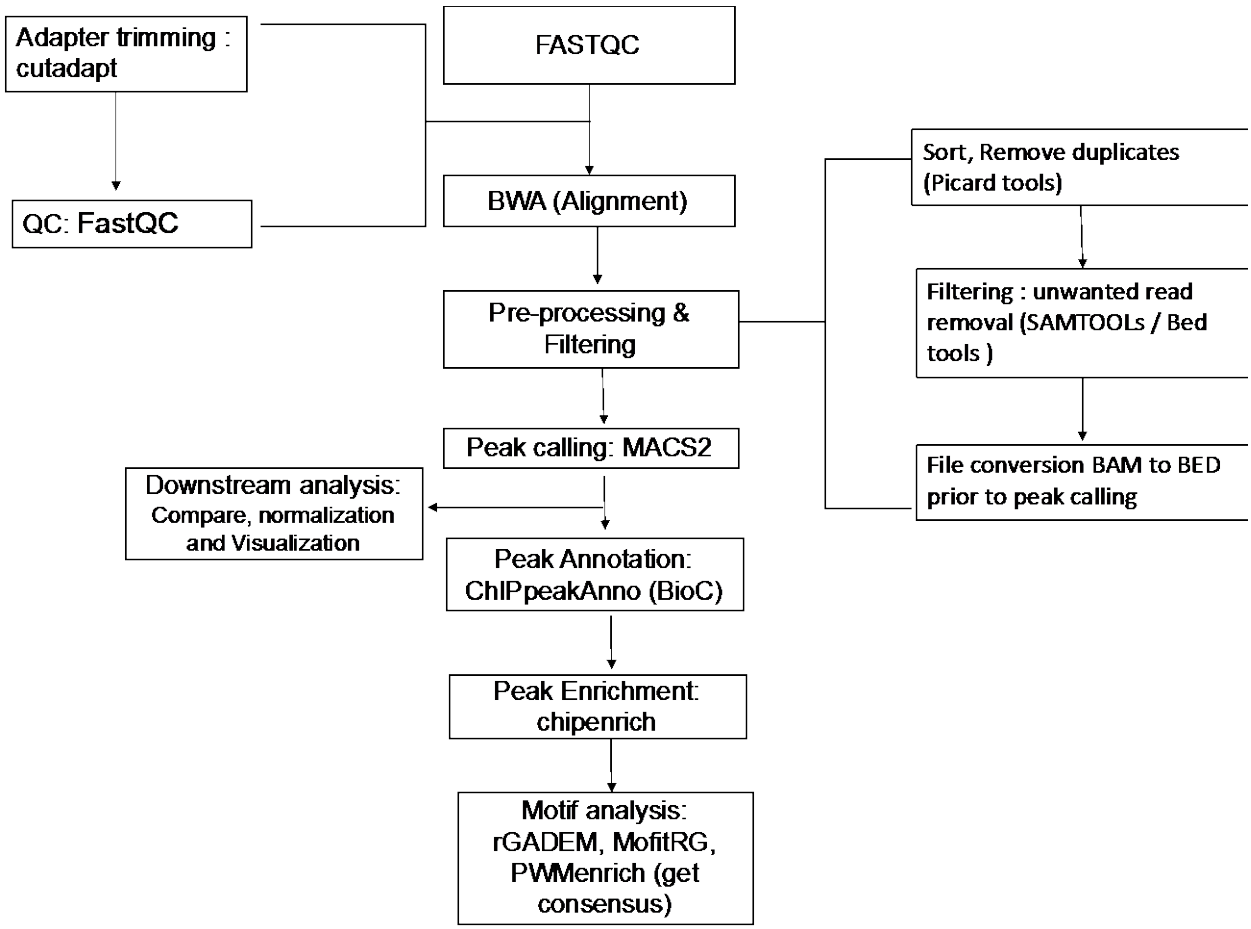
Figure 3. Spearman correlation plot of read counts at called peaks across bulk ATAC sequence replicates of sorted naïve T cells from healthy human donors. (Data from GEO accession GSE101498, Satpathy et. al., 2018)

Overall Description of ATAC-seq analysis pipeline

Raw FASTQ files will be subjected to adapter trimming and quality control metrics will be collected prior to alignment using FastQC and ATAqC (Tool from Kundeja Lab). The adapter trimmed files

will be aligned to the reference genome (hg38) using BWA alignment algorithm. This is followed by processing and filtering steps, including removal of unmapped reads, mitochondrial reads, multi-mapped reads, sorting and removal of PCR duplicates. These steps will be performed using samtools and bedtools. The processed BAM files will be converted to tagAlign files (BED 3+3 format files) and peak calling will be performed with MACS2 software. Peaks will be annotated with ChiPpeak Anno (BioC) which maps peak to nearest feature (TSS, gene, exon, miRNA features) and functional enrichment analysis will be performed with chipenrich. Motif enrichment will be performed with a different tool to find consensus, these include rGADEM, MotifRG and PWMenrich

Flowchart of ATAC-seq analysis workflow



References

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Corces MR, Trevino AE, Hamilton EG,… Greenleaf WJ, Chang HY. (2017) An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods. 2017 Oct; 14(10): 959–962.

Satpathy, A. T., Saligrama, N., Buenrostro, J. D., Wei, Y., Wu, B., Rubin, A. J., … Chang, H. Y. (2018). Transcript-indexed ATAC-seq for precision immune profiling. *Nature Medicine*, *24*(May), 1. https://doi.org/10.1038/s41591-018-0008-8