Supplemental figures

	trimming	read QC	mapping	deduplication	filtering	signal generation	peak calling	QC	downstream analyses
AIAP	cutadapt	FastQC	bwa	picard	samtools methylQA	UCSC tools	MACS2	MultiQC	DESeq2
ATAC2GRN	NA	NA	bowtie2	NA	NA	NA	HOMER	NA	HINT
ATAC-pipe	custom python	custom python	bowtie2	picard	samtools	UCSC tools	MACS2	custom python	CENTIPEDE DESeq2 HOMER
ATACProc	trim_adapters.py*	NA	bowtie2	picard DeepTools	samtools DeepTools	UCSC tools DeepTools	MACS2	ataqv	HINT-ATAC HOMER DeepTools custom python
CIPHER	BBDUK	FastQC	bbmap bowtie2 bwa hisat2 star	NA	samtools	DeepTools	MACS2 epic	MultiQC	NA
ENCODE	trimmomatic cutadapt	NA	bowtie2 bwa	picard	samtools bedtools	UCSC tools	MACS2	custom code	IDR
esATAC	AdapterRemoval	NA	Rbowtie2	custom R	NA	custom R	F-Seq	custom R	ChIPpeakAnno
GUAVA	cutadapt	FastQC	bowtie2	NA	NA	UCSC tools	MACS2	custom code	DESeq2 ChIPpeakAnno
I-ATAC	trimmomatic	FastQC	bwa	picard	NA	NA	MACS2	NA	NA
nfcore/atacseq	Trim Galore!†	FastQC	bwa	picard	samtools bedtools pysam bamtools	bedtools UCSC tools	MACS2	ataqv	DESeq2
PEPATAC	skewer trimmomatic trim_adapters.py‡	FastQC	bowtie2 bwa	samblaster picard samtools	samtools bedtools	custom python	MACS2 F-Seq2 Genrich HMMRATAC HOMER	custom code	HOMER custom code
pyflow-ATAC-seq	atactk§	FastQC	bowtie2	samblaster	samtools	DeepTools	MACS2	ataqv MultiQC	CENTIPEDE
seq2science	Trim Galore!†	FastQC	bowtie2 bwa hisat2 star	picard	samtools	DeepTools	MACS2 Genrich HMMRATAC	MultiQC	custom code
snakePipes ATAC-seq	cutadapt	FastQC	bowtie2	sambamba	samtools	DeepTools	MACS2 Genrich HMMRATAC	MultiQC	CSAW
Tobias Rausch	cutadapt	FastQC	bowtie2	biobambam2	samtools	Alfred	MACS2	Alfred	HOMER custom R tutorial
OVERALL	cutadapt	FastQC	bowtie2	picard	samtools	UCSC tools	MACS2	MultiQC	HOMER DESeq2

Fig. S1: ATAC-seq pipelines universally require several common bioinformatic tools. *While all pipelines require a number of common bioinformatic tools, PEPATAC offers the greatest flexibility and includes a number of the most popular tools.*

Supplemental files

Supplemental_file_1.csv

Supplemental_file_1.csv is the PEP-formatted sample table for the primary dataset. Samples are defined by protocol, whether standard, fast, or omni, and include accession numbers for access through the Gene Expression Omnibus (63).

Supplemental_file_2.xlsx

Supplemental_file_2.xlsx contains two sheets. The "jaccard_similarities" sheet includes tables representing the results of `bedtools intersect` between each independent peak caller software for 1) the PEPATAC derived consensus peak set, and 2) for an individual sample (SRR5210416) between each peak caller. This sheet also includes the average jaccard statistic for each peak caller. The "blacklisted_regions" sheet compares the number of peaks generated by each peak caller that overlap blacklisted regions (35).

Supplemental_file_3.xlsx

Supplemental_file_3.xlsx includes three sheets for a standard ATAC (SRR5427804), fast ATAC (SRR2920492), and omni ATAC (SRR5427806) sample that has been run through PEPATAC with 1) no prealignments, 2) mitochondrial

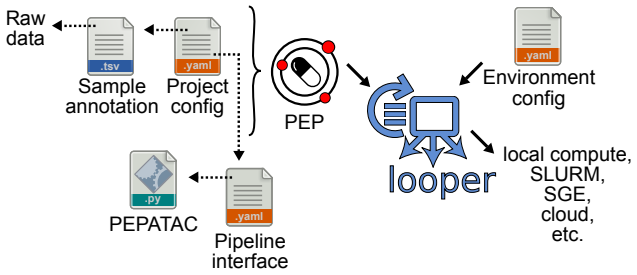

Fig. S2: Deploying PEPATAC across multiple samples using looper. *The PEPATAC pipeline can be easily run across multiple samples in any computing environment using looper.*

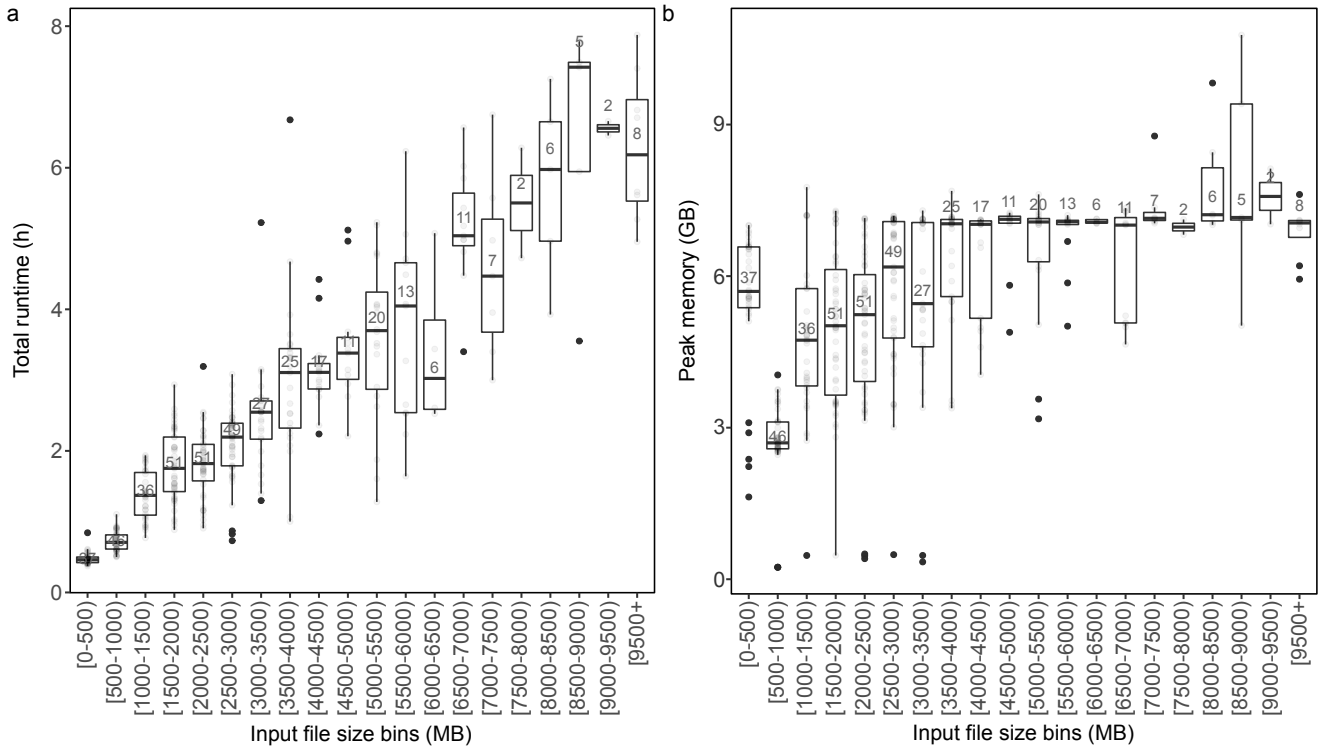

Fig. S3: PEPATAC is computational efficient. *(a) Pipeline runtime scales linearly with input file size. (b) Pipeline memory use peaks between 5-9GB.*

prealignment (rCRSd: the revised Cambridge Reference Sequence doubled genome), and 3) mitochondrial, human repeats, and rDNA prealignments. In each sheet, for the highest scoring peaks, individual peak fasta sequences (included) were aligned with BLAST (60) and top scoring annotations recorded. If the peak overlaps a known blacklisted region (35), this is also marked.

Supplemental_file_4.csv

Supplemental_file_4.csv is the PEP-formatted sample table for the performance testing dataset. Accession numbers for file access through the Gene Expression Omnibus (63) are included for each sample.

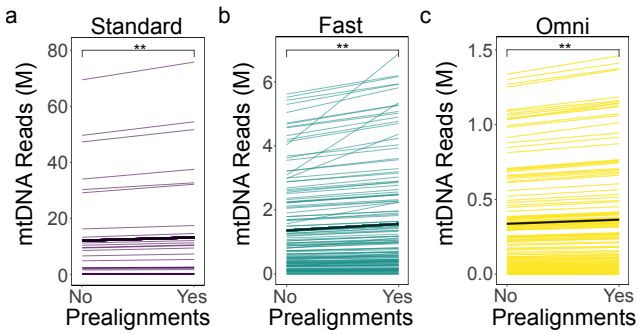

Fig. S4: Prealignment increases mtDNA alignment. *Within Standard (a), Fast (b), and Omni (c) ATAC-seq library preparation protocols, every sample shows increased mtDNA alignment when utilizing prealignments (The gray lines represent the mean increase within each protocol. ** = $p < 0.001$; t-test (mu = 0) with Benjamini-Hochberg correction.)*

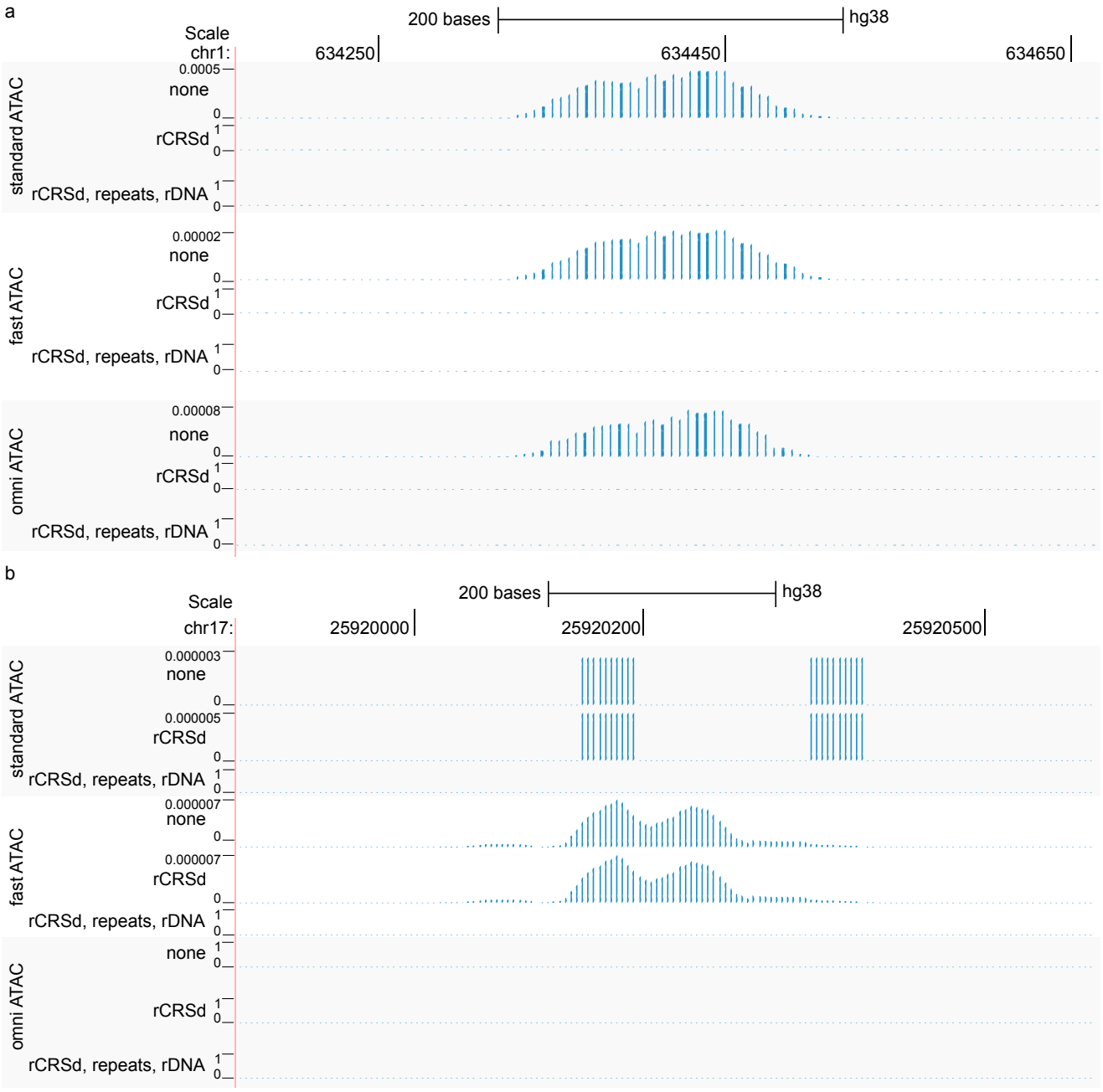

Fig. S5: Prealignment (and improved ATAC-seq library preparation protocols) successfully deplete signal from NuMTs, repeat regions, and high signal regions. *(a) Even where improved library preparation protocol leads to a NuMT annotated peak, prealignment successfully removes the spurious signal. (b) Both omni ATAC and prealignment to mitochondria* and *repeats and ribosomal sequence successfully depletes a spurious signal.*

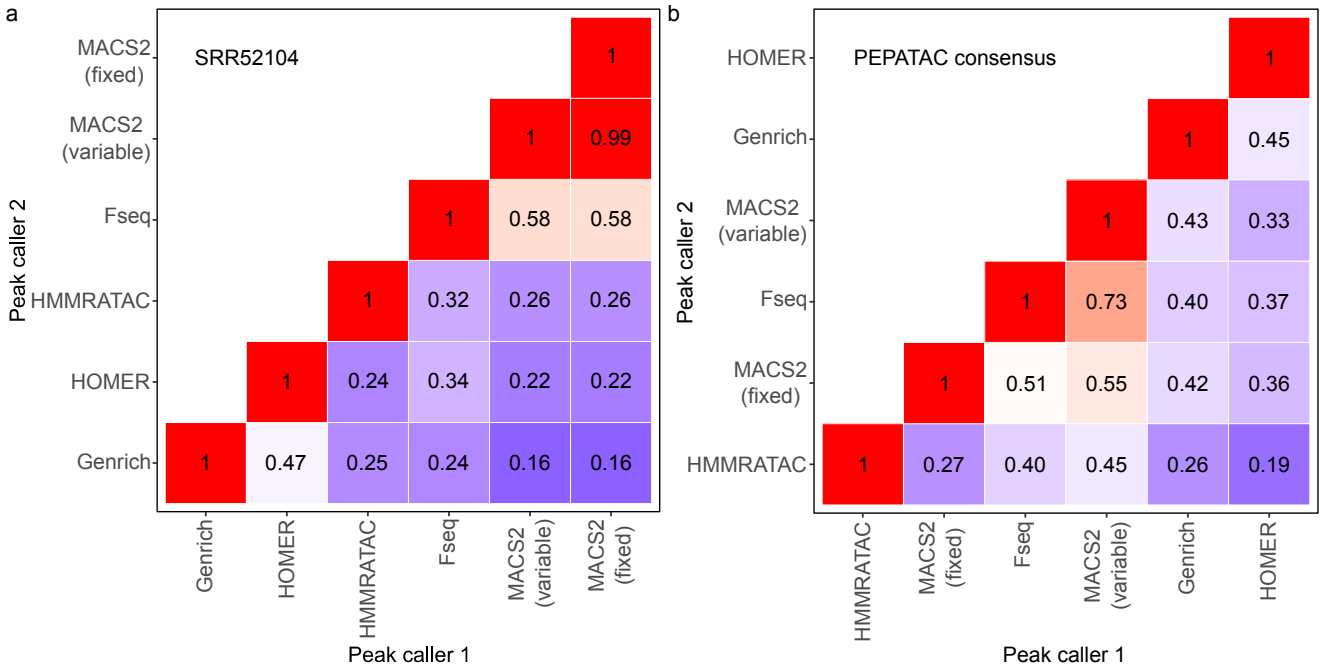

Fig. S6: Peaks are comparatively dissimilar between the five optional peak callers. *(a) For a single sample, MACS2 derived peaks, both with fixed and variable width peaks, are the most similar to Fseq called peaks. Genrich and HMMRATAC are the most unique among peak callers. (b) After PEPATAC consensus peak generation, HMMRATAC becomes even more dissimilar from the results derived from alternative peak callers.*

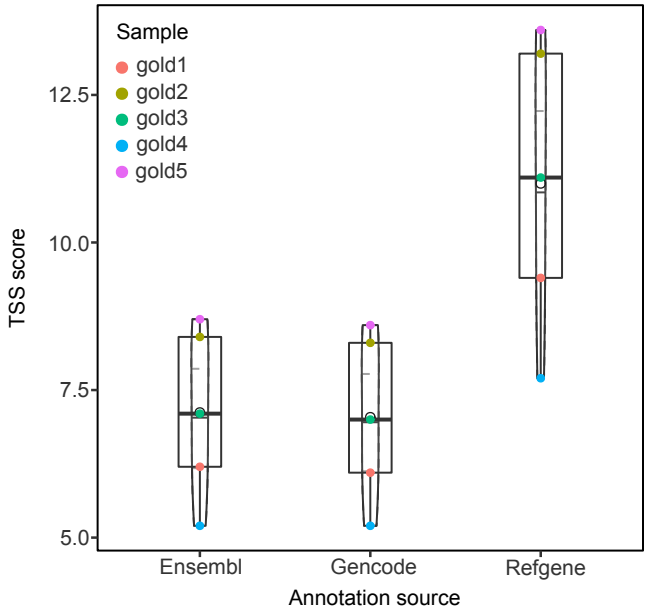

Fig. S7: The TSS enrichment score is dependent on the annotation source. *Refgene TSS annotations, which include the predominant TSS annotation only, produces the highest TSS enrichment score.*

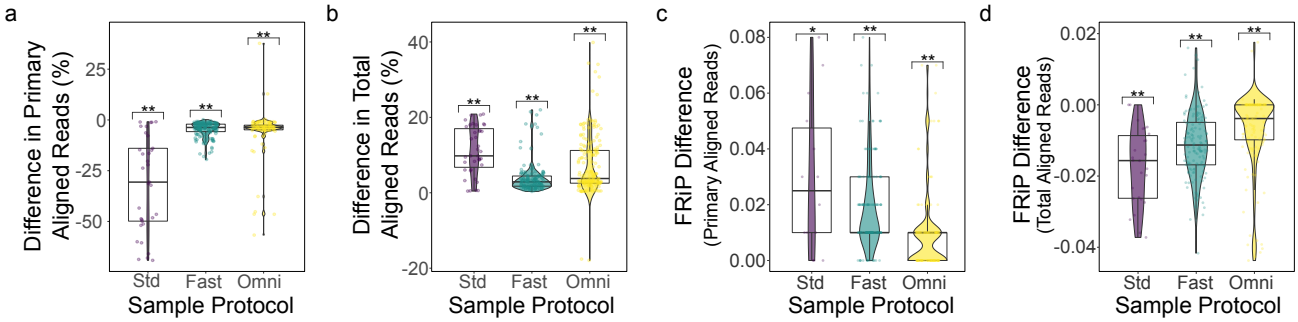

Fig. S8: Prealignment changes the relationship between primary genome and total aligned reads and the fraction of reads in peaks (FRiP) is dependent on mapping strategy. *(a) The number of primary, nuclear genome mapped reads is reduced when using prealignments. (b) However, the total number of mapped reads is increased with prealignments due to more specific read mapping. (c) The FRiP is increased with prealignments when using primary, nuclear genome mapped reads as the denominator. (d) In contrast, when using the total mapped reads the FRiP is reduced when using prealignments due to a larger mapped read pool in the denominator (* = $p < 0.01$; ** = $p < 0.001$; t-test (mu = 0) with Benjamini-Hochberg correction).*