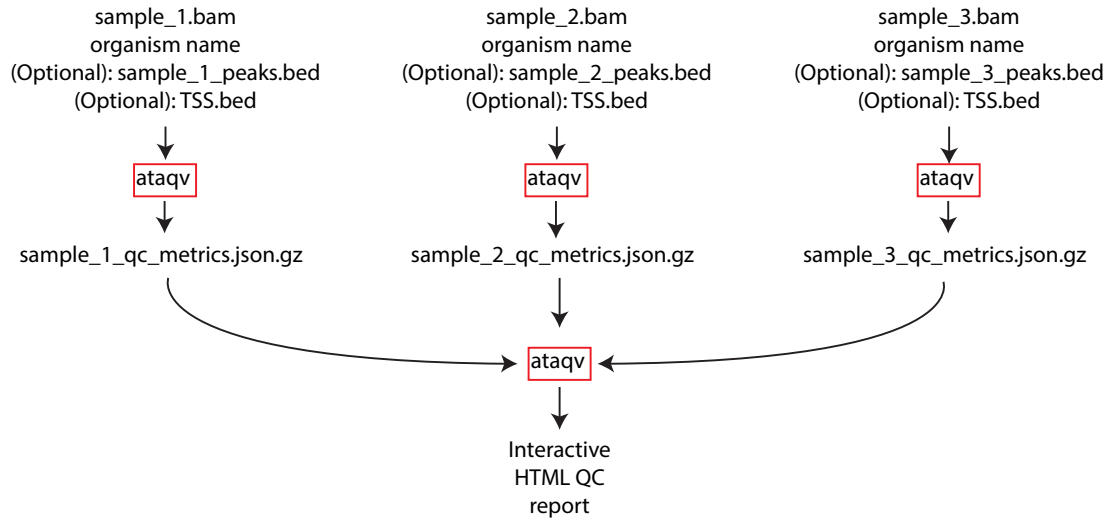
A



B

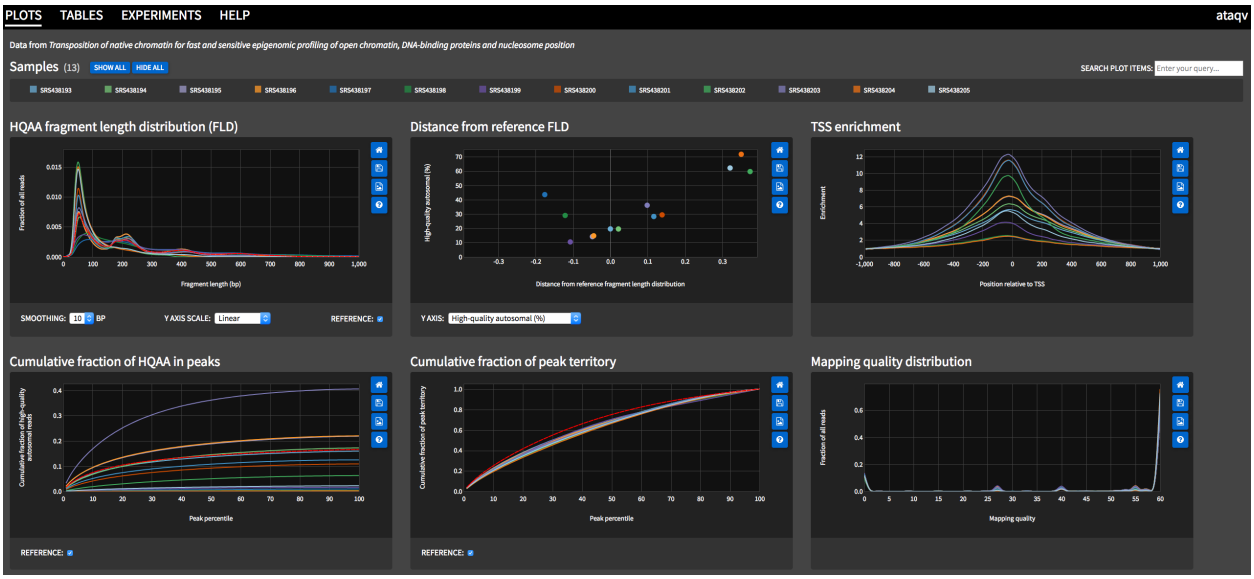


Figure S1. Ataqv workflow, Related to STAR Methods. Aligned reads (BAM format), the organism name, and optionally a file of peak calls and TSS annotations are passed to ataqv, which generates a JSON-formatted file of quality control metrics. JSON files for multiple bam files can be passed back to ataqv, which then creates an interactive HTML report displaying the samples jointly.

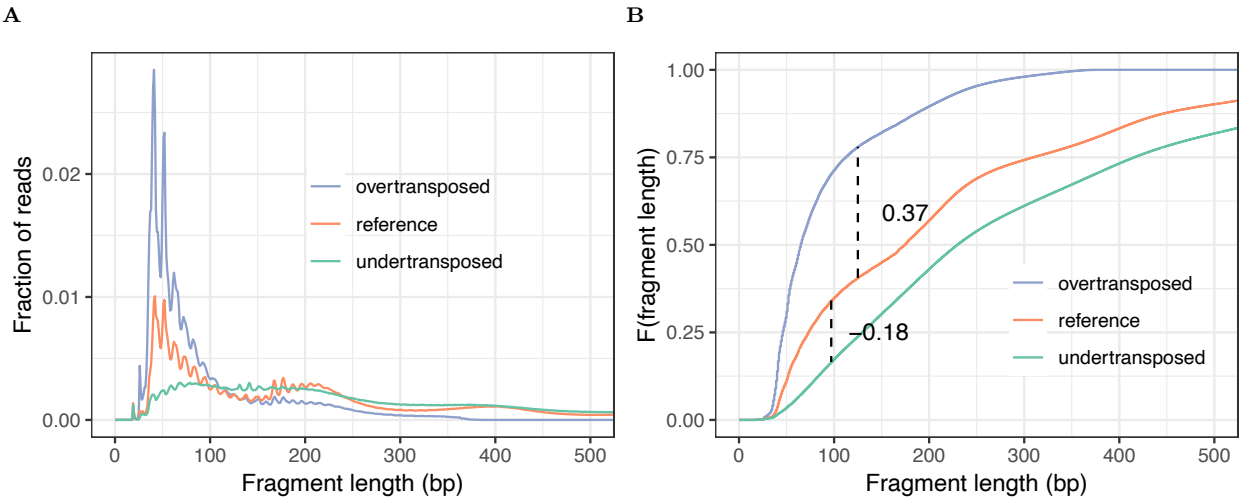


Figure S2. Calculation of the fragment length distribution distance, Related to STAR Methods. Three hypothetical distributions are displayed in panel (A); one is chosen as the reference sample. Relative to the reference sample, one sample has an abundance of shorter fragments (as would be expected for a high Tn5 : nuclei ratio) and one has an abundance of longer fragments (as would be expected for a low Tn5 : nuclei ratio). The fragment length distributions are converted to cumulative distributions (B), and the Kolmogorov Smirnov statistic is calculated (the maximum vertical distance between two cumulative distributions). The sign is then set according to whether or not the sample of interests cumulative distribution takes a value greater than (positive) or less than (negative) the reference cumulative distribution at the point of maximum vertical distance.

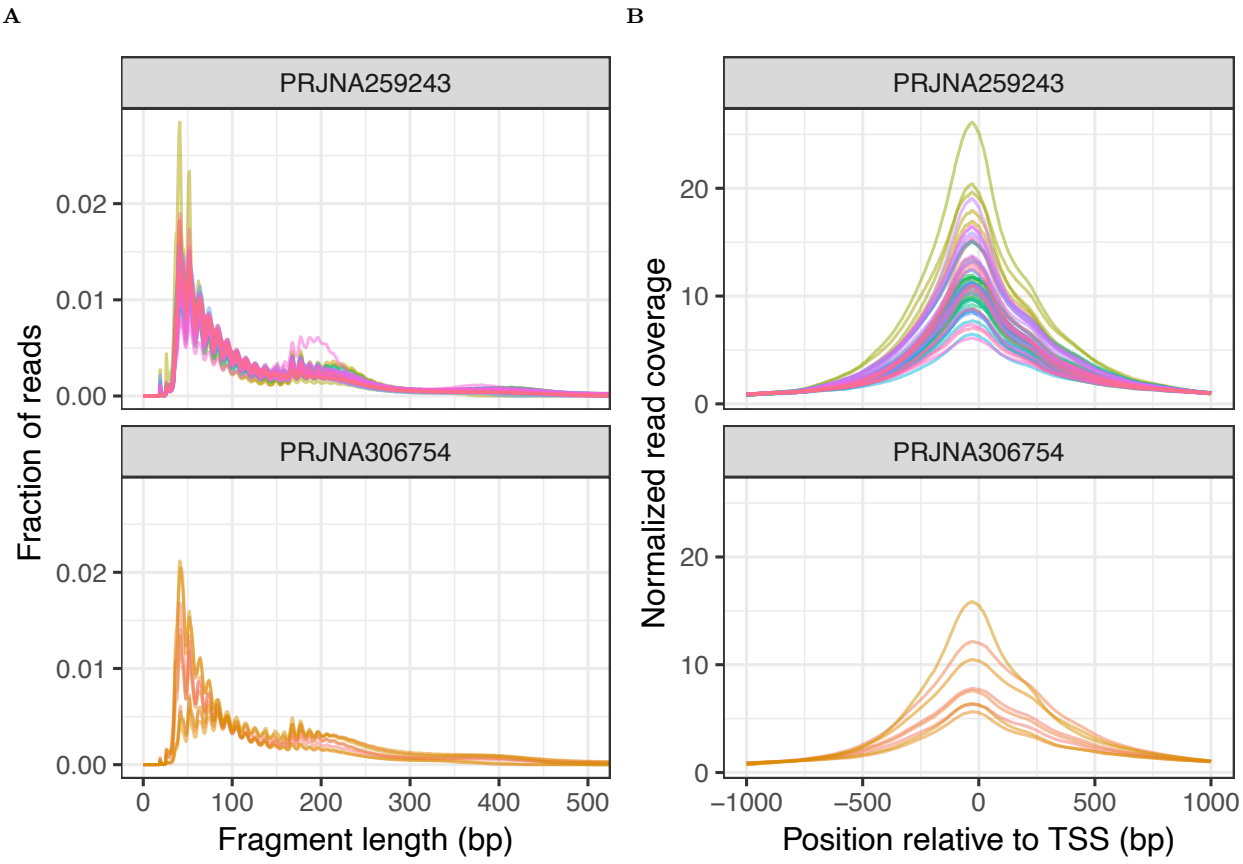


Figure S3. Intrastudy heterogeneity in ATAC-seq data, Related to Figure 1. (A) Fragment length distributions and (B) TSS enrichment for two public datasets (facet header labels).

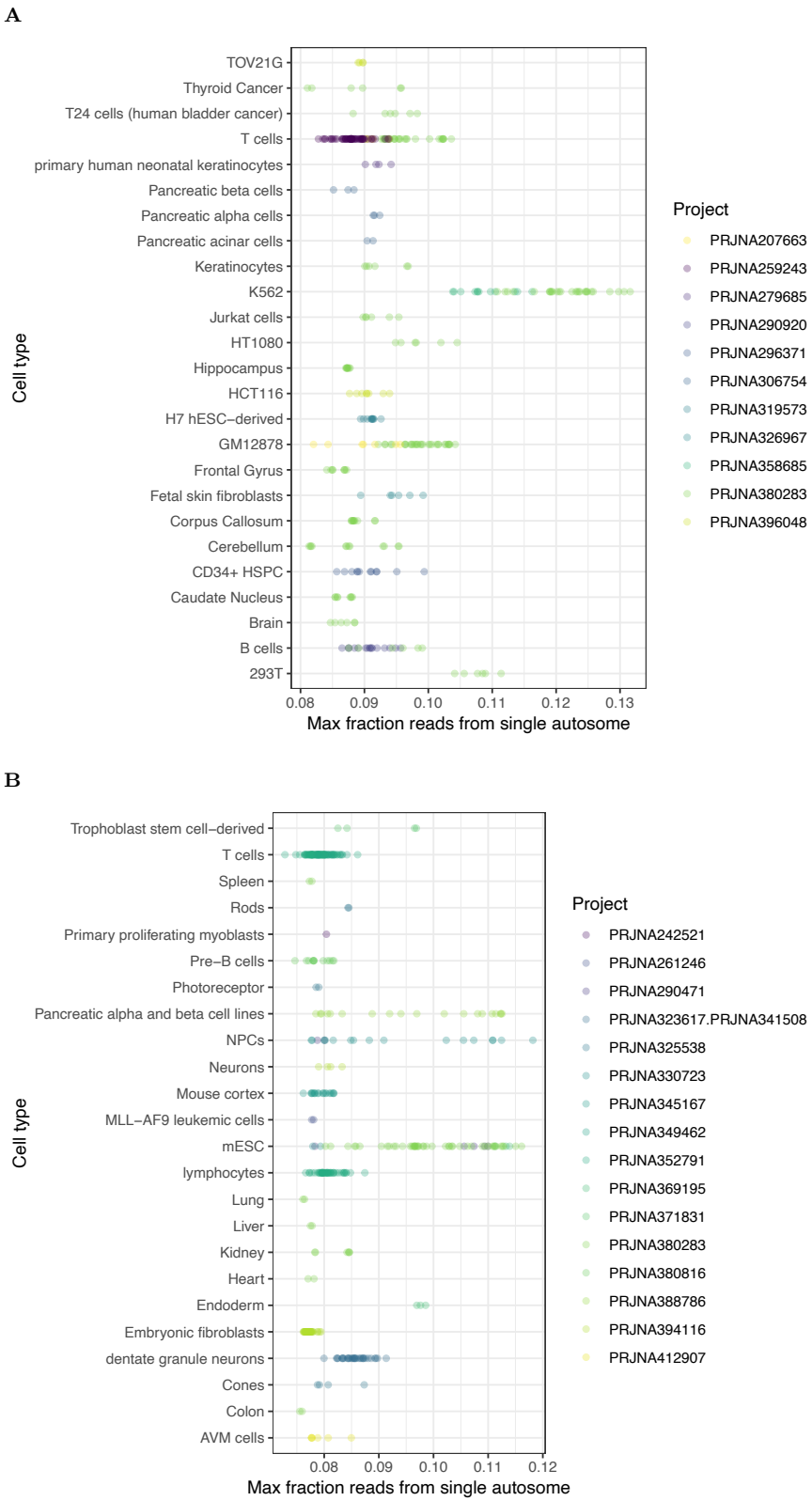


Figure S4. Maximum fraction of autosomal reads from a single chromosome in public bulk ATAC-seq data, Related to Figure 1. Maximum fraction of autosomal reads from a single chromosome in public bulk ATAC-seq data from (A) human and (B) mouse. Each point represents one library. Outliers tend to be cell lines with known abnormal karyotypes.

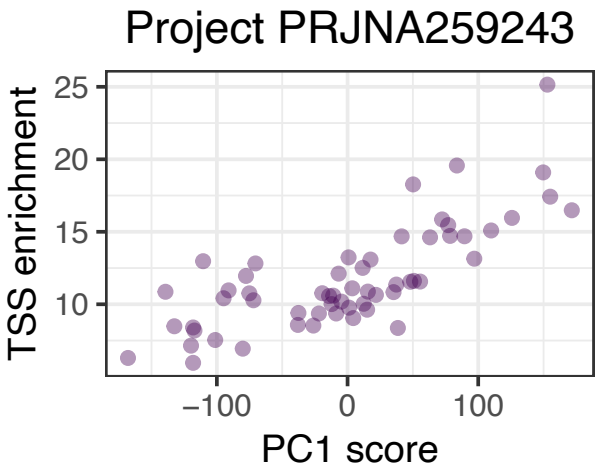


Figure S5. Correlation between PC1 and TSS enrichment in project PRJNA259243, Related to Figure 1. Each point represents one library.

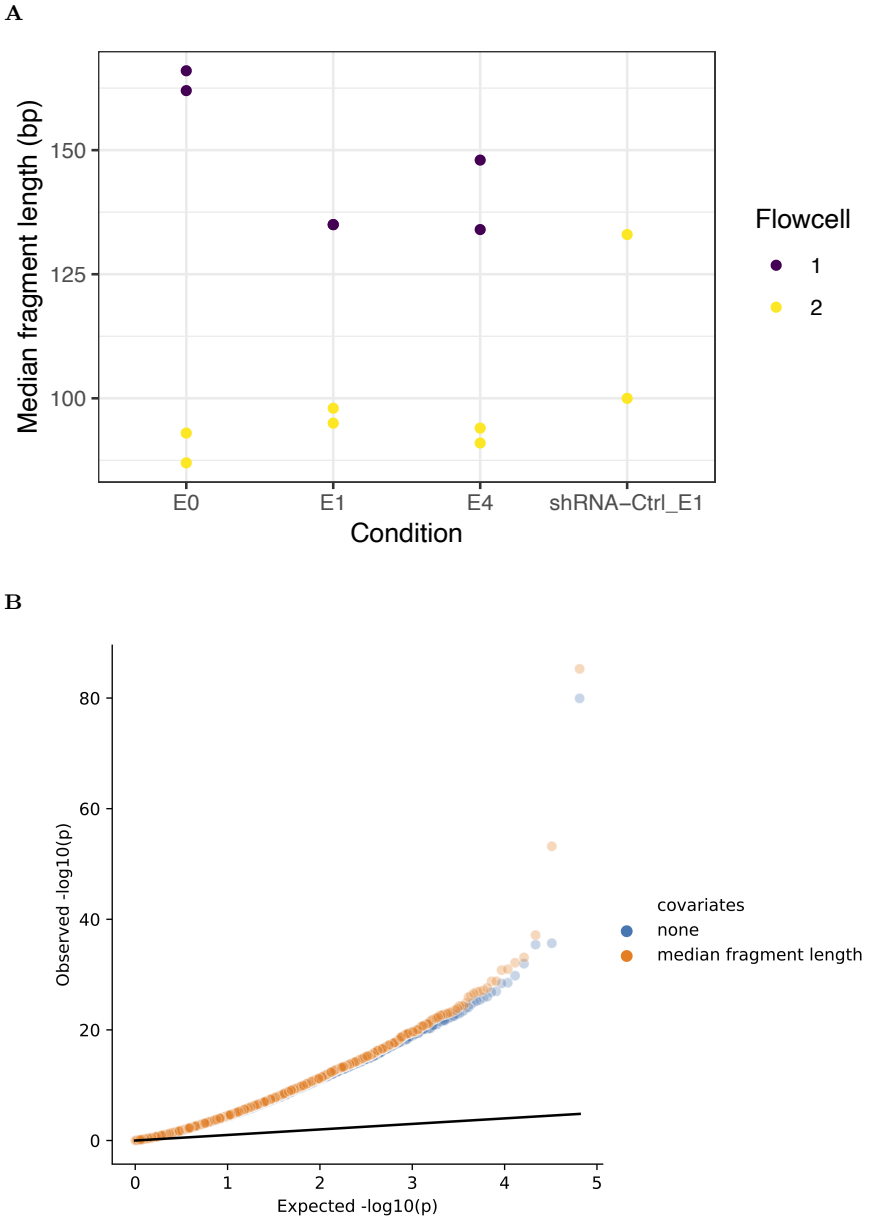


Figure S6. Quality control metrics may track latent variables, Related to Figure 1. (A) Median fragment length vs sequencing run flowcell for a subset of libraries from project PRJNA323617/PRJNA341508. Median fragment length clearly correlates with flowcell, indicating that median fragment length may correlate with processing/sequencing batch or another related experimental variable. (B) Adding median fragment length as a covariate to a differential peak analysis between E0 and E1 results in more extreme p-values.

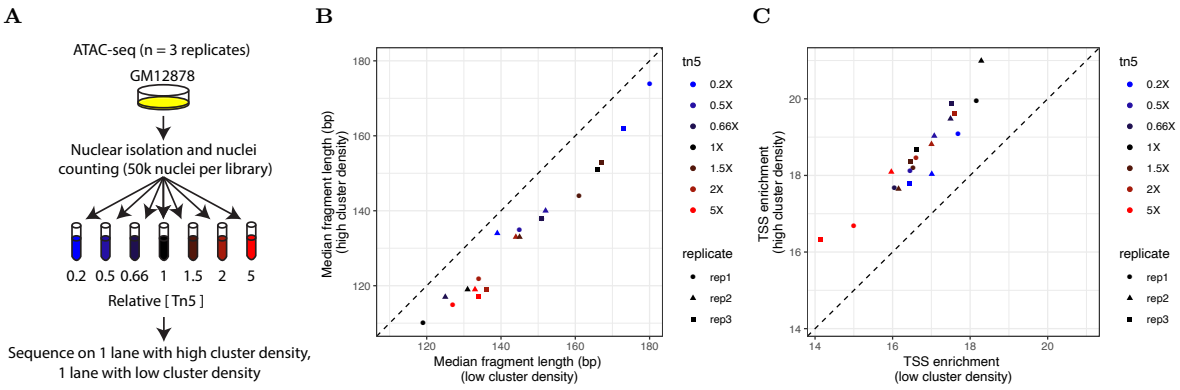


Figure S7. Sequencing lane cluster density systematically alters ATAC-seq results, Related to STAR Methods. (A) Study design for an experiment exploring the influence of sequencing flow cell cluster density. ATAC-seq was performed using 7 concentrations of Tn5; 3 replicates were used, resulting in 21 total libraries. All libraries were sequenced together in 2 sequencing runs, each on the same NextSeq 500 instrument. For one sequencing run, the flow cell loading concentration was low (about 411M clusters passed filtering), for the second the flow cell loading concentration was high (about 508M clusters passed filtering). (B) Median fragment length for each library in either sequencing run. The high cluster density run results in consistently smaller fragments, consistent with shorter fragments increasingly outcompeting longer fragments in the high-cluster-density flowcell. (C) TSS enrichment for each library using results from the high cluster density or low cluster density run. The high cluster density run results in consistently greater TSS enrichment.

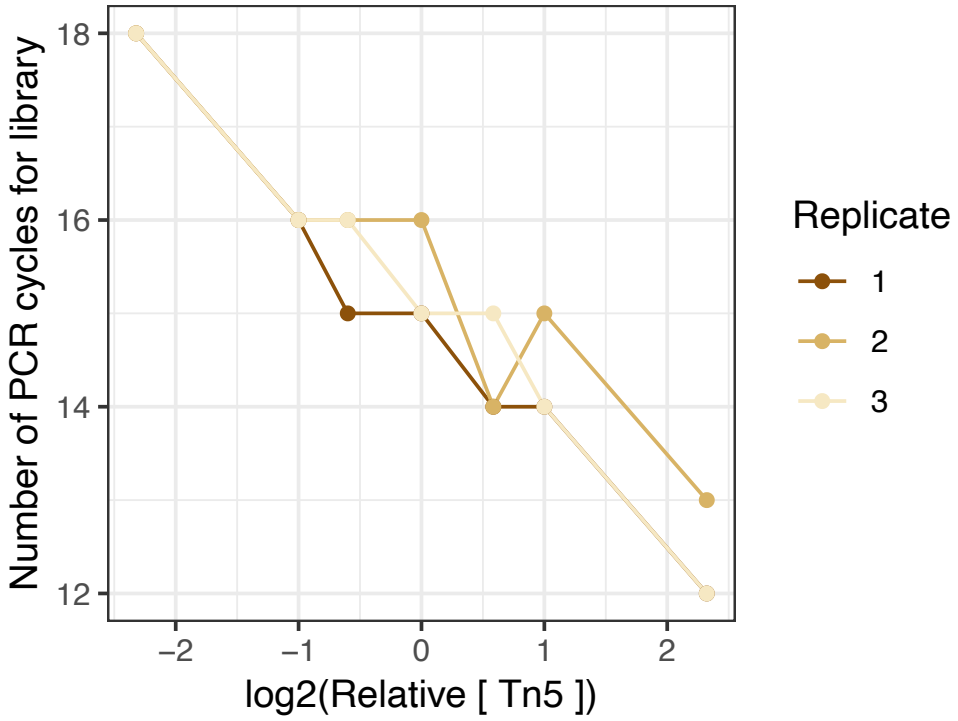


Figure S8. PCR cycles correlates with Tn5 concentration, Related to STAR Methods. Number of PCR cycles for all libraries in the PCR-variable experiment, by nuclear isolation replicate and Tn5 concentration.

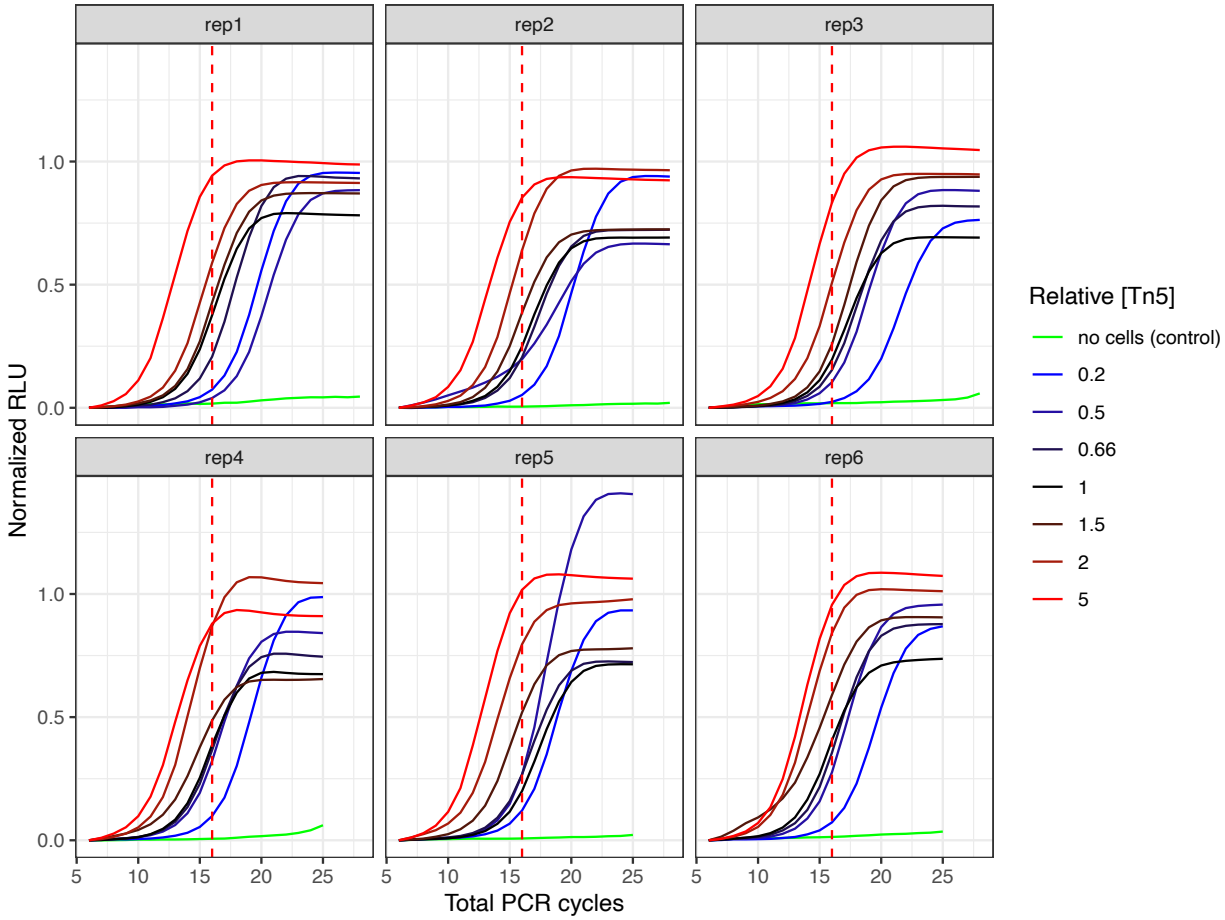


Figure S9. qPCR curves used for the selection of PCR amplification cycles in the PCR-constant experiment, Related to Figure 2. Facets represent nuclear isolation replicates. Dashed red line represents the number of PCR cycles (16) used for all libraries.

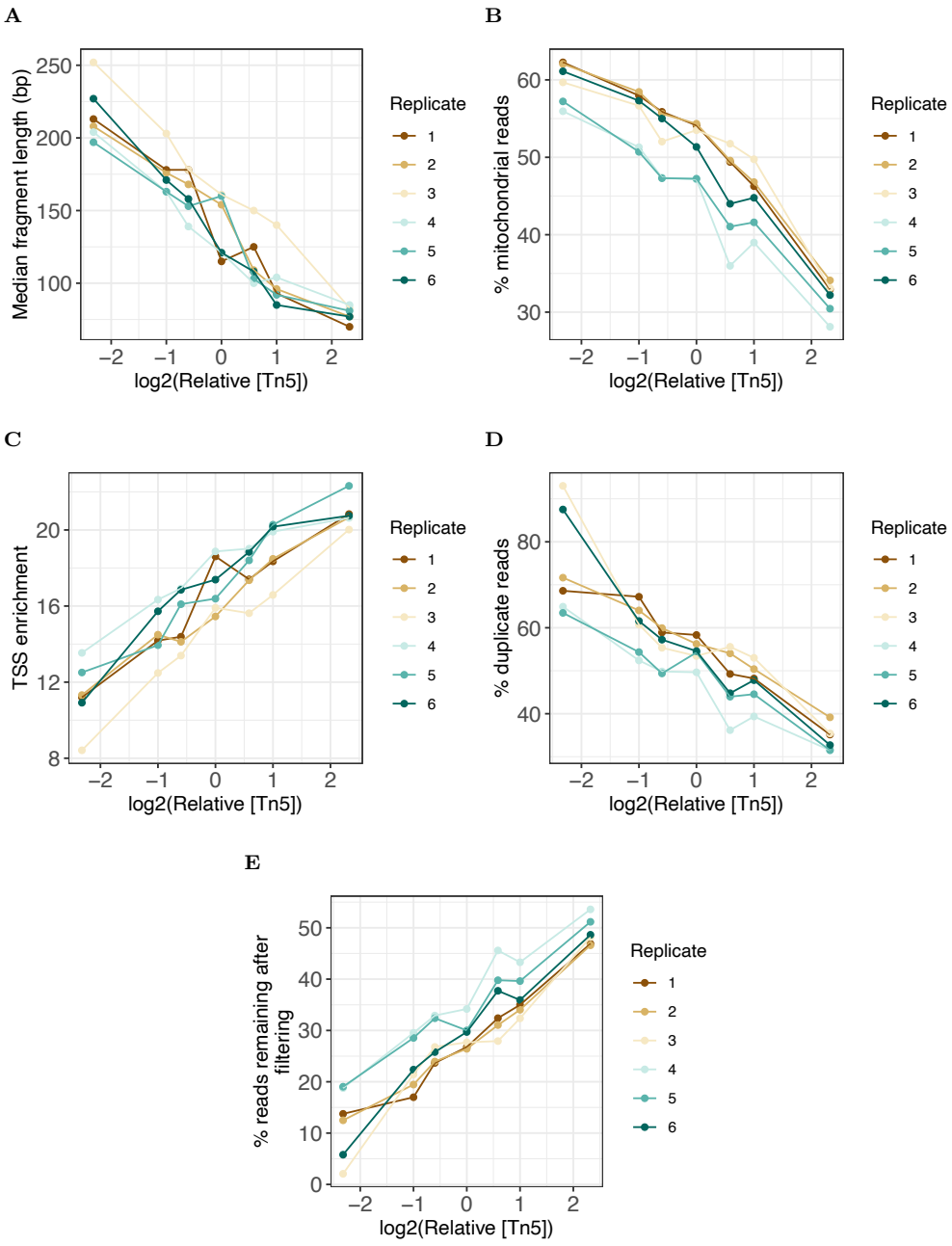


Figure S10. Tn5 concentration correlates with QC metrics in the PCR-constant experiment, Related to Figure 2. Correlation between Tn5 concentration and (A) median fragment length, (B) percent of mitochondrial reads, (C) TSS enrichment, (D) duplication rate, and (E) proportion of reads remaining after deduplication and filtering.

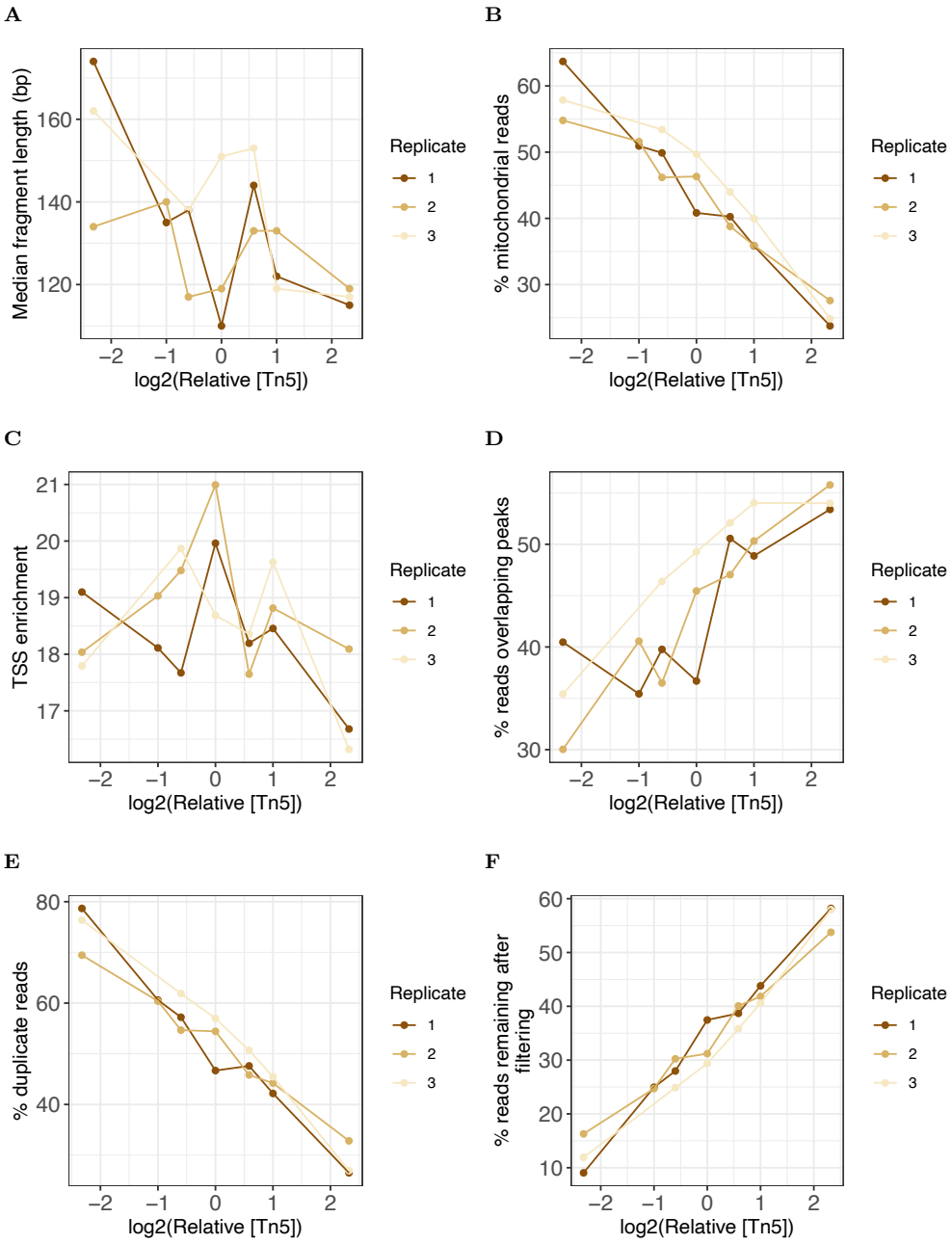


Figure S11. Tn5 concentration correlates with QC metrics in the PCR-variable experiment, Related to Figure 2. Correlation between Tn5 concentration and (A) median fragment length, (B) percent of mitochondrial reads, (C) TSS enrichment, (D) percent of filtered reads overlapping peaks, (E) duplication rate, and (F) proportion of reads remaining after deduplication and filtering. Data shown is from the high cluster density sequencing run.

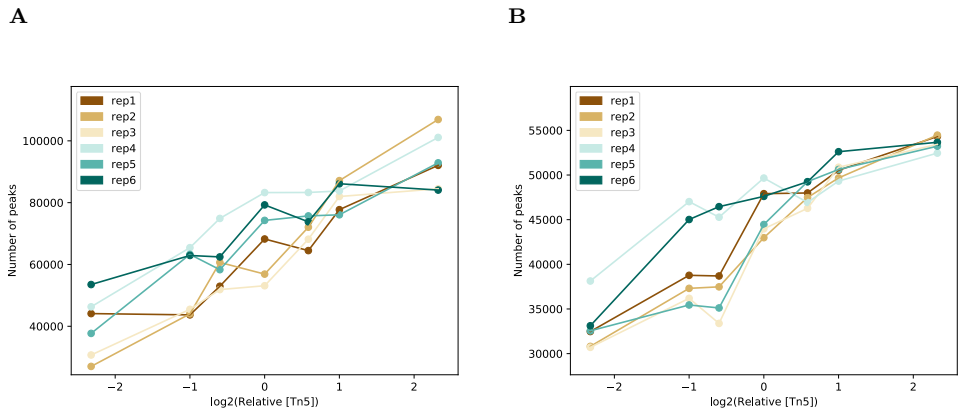


Figure S12. Tn5 concentration correlates with number of peaks called, Related to Figure 2. Correlation between Tn5 concentration and number of peaks called (A) without and (B) with subsampling to an equal number of post-filtering reads (PCR-constant experiment).

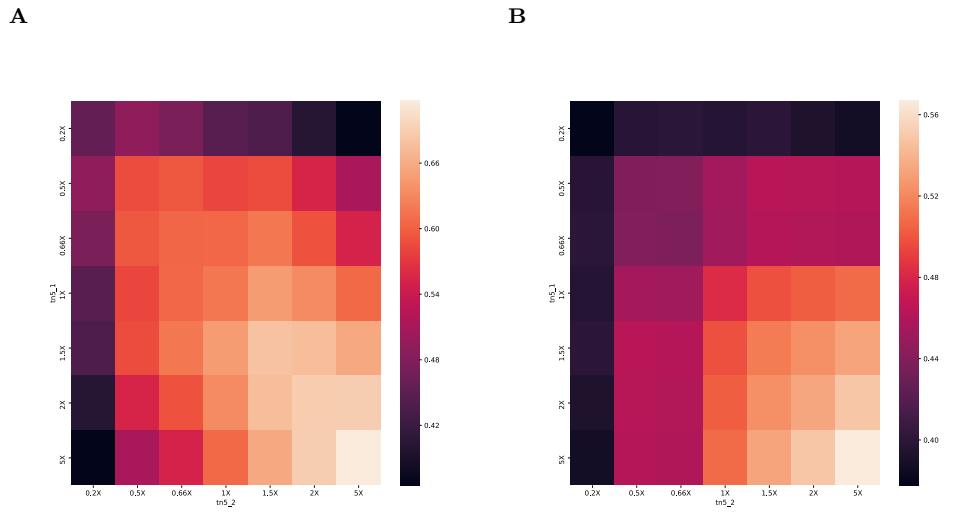


Figure S13. Peak call reproducibility increases with Tn5 concentration, Related to Figure 2. The Jaccard index between peak calls from all pairs of libraries from the PCR-constant experiment was calculated by first taking the union of peak calls in any two libraries to generate a set of merged peaks, and then dividing the number of those merged peaks that appear in both libraries by the total number of merged peaks. A merged peak was considered to appear in a library if there was any overlap between a peak from that library and a merged peak. (A) The average Jaccard index between two replicates with given Tn5 concentrations, using all reads for peak calling. (B) The average Jaccard index between two replicates after subsampling reads to ensure that the same number of reads is used for peak calling for each library.

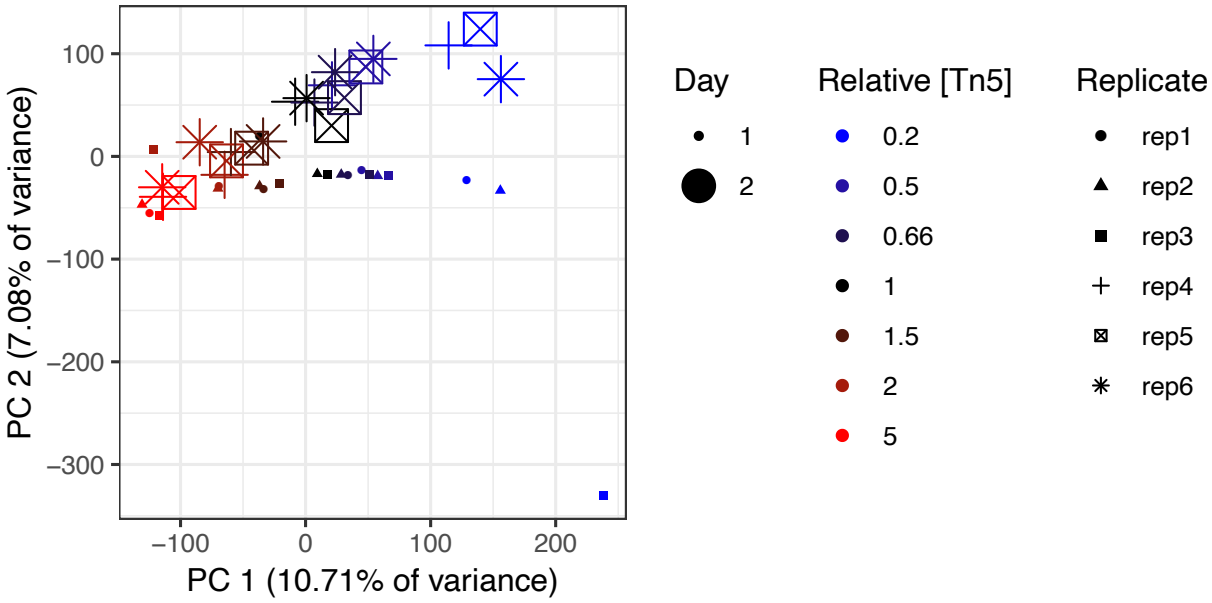


Figure S14. Principal component analysis of PCR-constant ATAC-seq datasets, Related to Figure 2. Six replicates were used (reps 1-3 produced on day 1, reps 4-6 on day 2). The first principal component captures Tn5 concentration. The second principal component captures the day on which the experiment was performed.

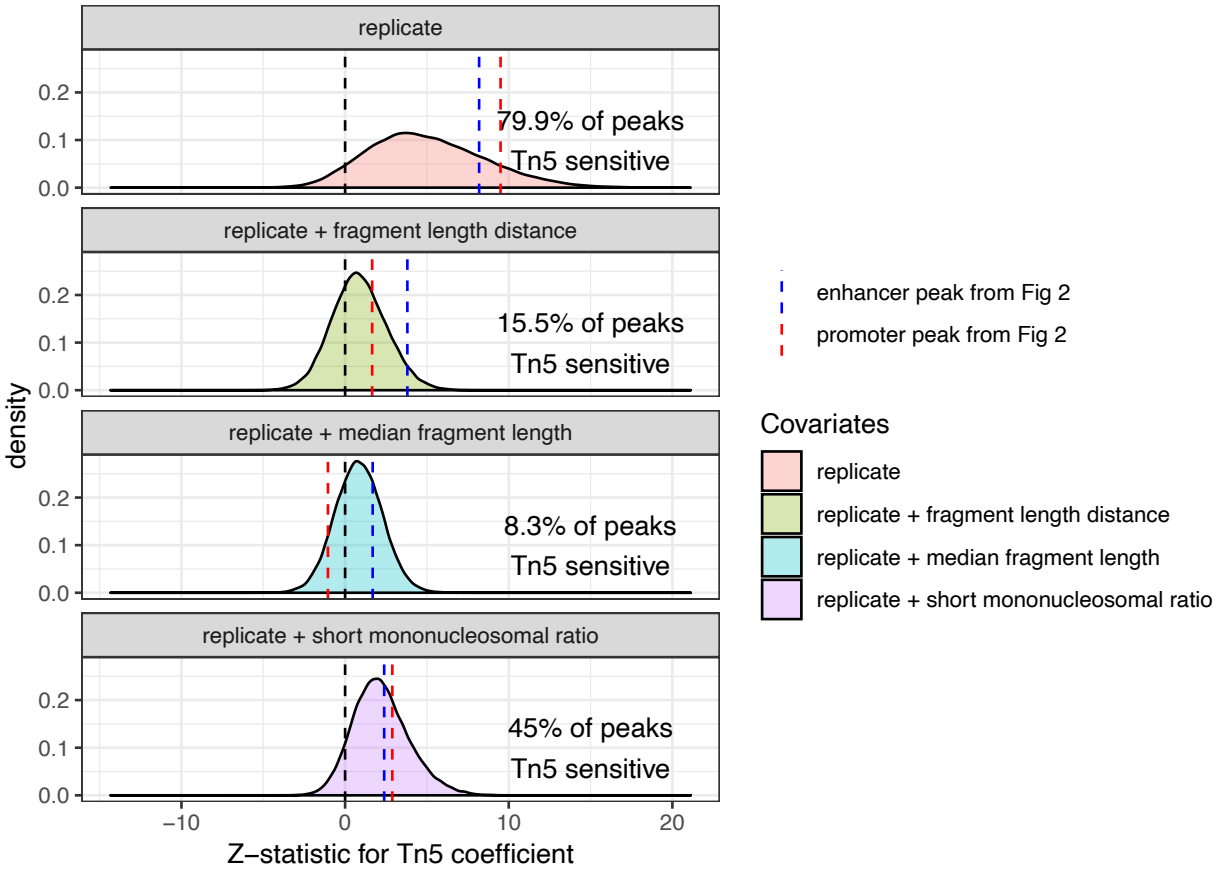


Figure S15. Most ATAC-seq peaks are Tn5 sensitive when PCR-cycles are held constant, Related to Figure 2. Distribution of the Z-statistic for the coefficient of log2(relative Tn5 concentration) in the negative binomial GLM. The distribution is shifted in the positive direction when only the replicate is used as a covariate (top facet), indicating that many ATAC-seq peaks show increased signal (normalized to library size) as the concentration of Tn5 is increased. Adding either one of three covariates summarizing the fragment length distribution weakens the relationship between peak signal and Tn5 concentration. Percentages shown reflect only those peaks that converged in all 4 models (62,576 peaks). The example promoter and enhancer peaks from the genome browser screenshots in Fig. 2f,g are denoted by the red and blue dashed lines, respectively. They are Tn5 sensitive at 5% FDR when no FLD-summarizing variable is used as a covariate (or when short:mononucleosomal is used as a covariate), but no longer Tn5 sensitive when median fragment length or fragment length distance are used as covariates.

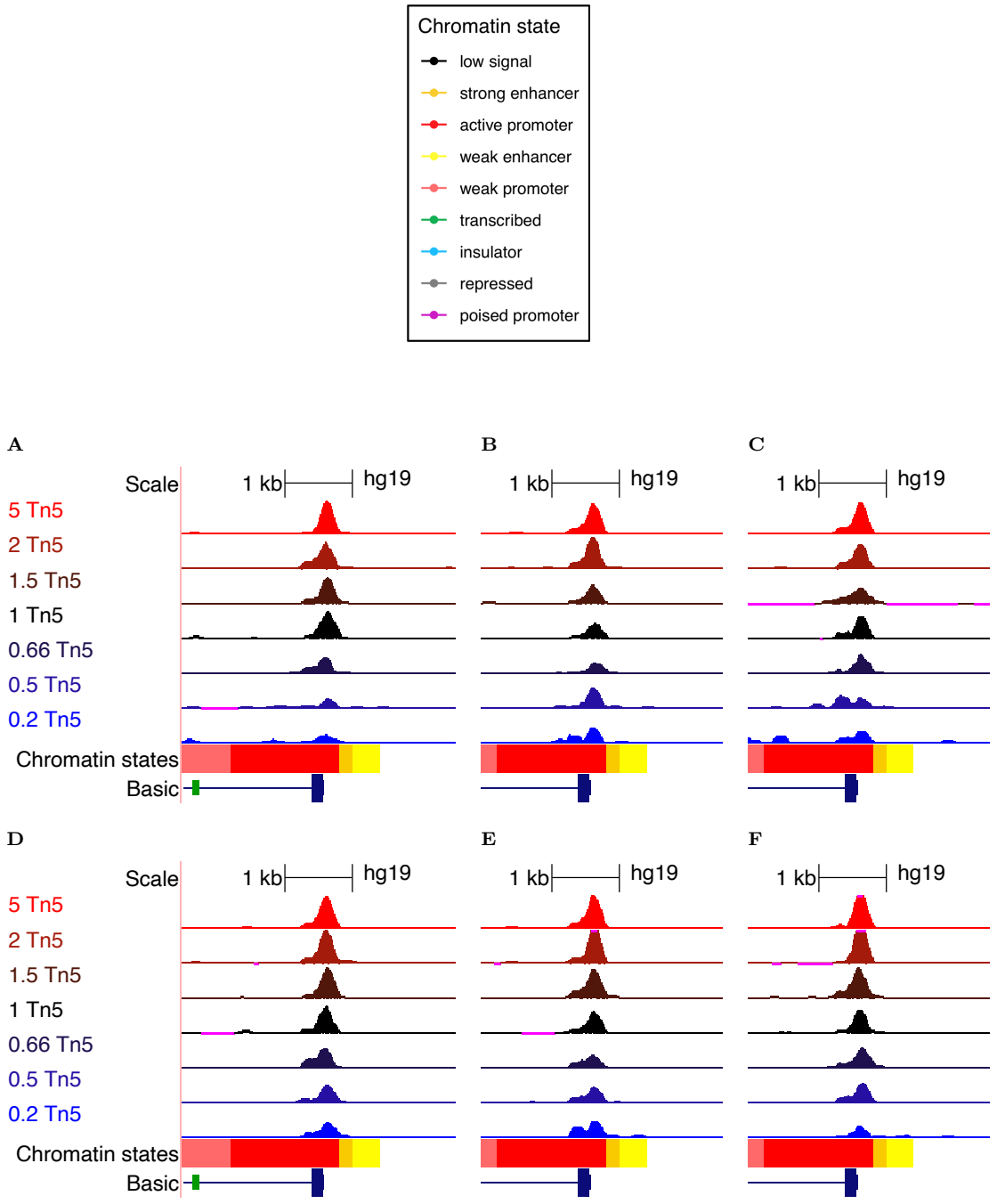


Figure S16. Example Tn5-sensitive promoter peak, Related to Figure 2. UCSC genome browser screenshot displaying a Tn5-sensitive promoter peak in the PCR-constant experiment. The six panels represent the six replicates, displaying high reproducibility (http://genome.ucsc.edu/)(Casper et al., 2018; Kent et al., 2002).

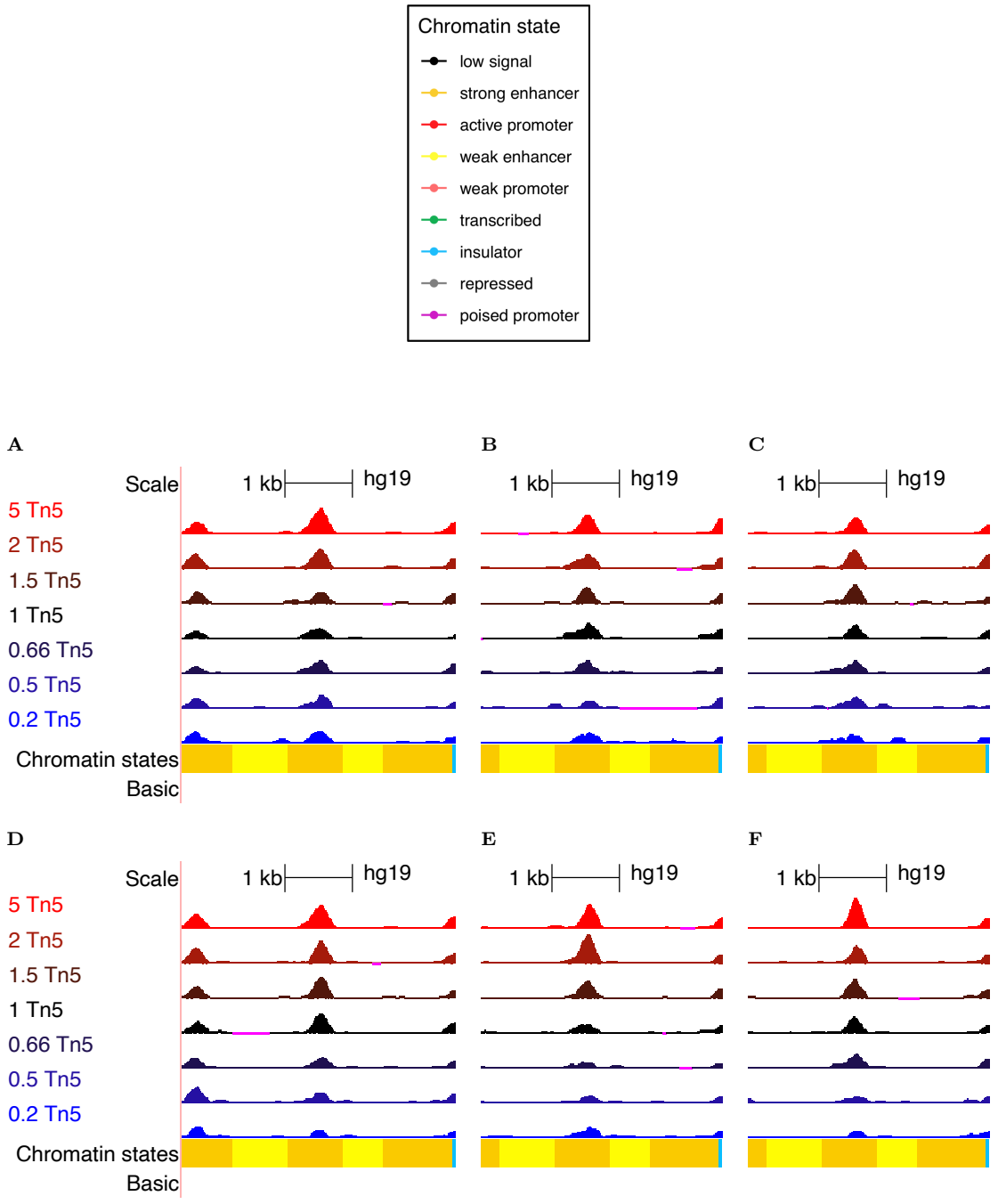


Figure S17. Example Tn5-sensitive enhancer peak, Related to Figure 2. UCSC genome browser screenshot displaying a Tn5-sensitive enhancer peak in the PCR-constant experiment. The six panels represent the six replicates, displaying high reproducibility (http://genome.ucsc.edu/)(Casper et al., 2018; Kent et al., 2002).

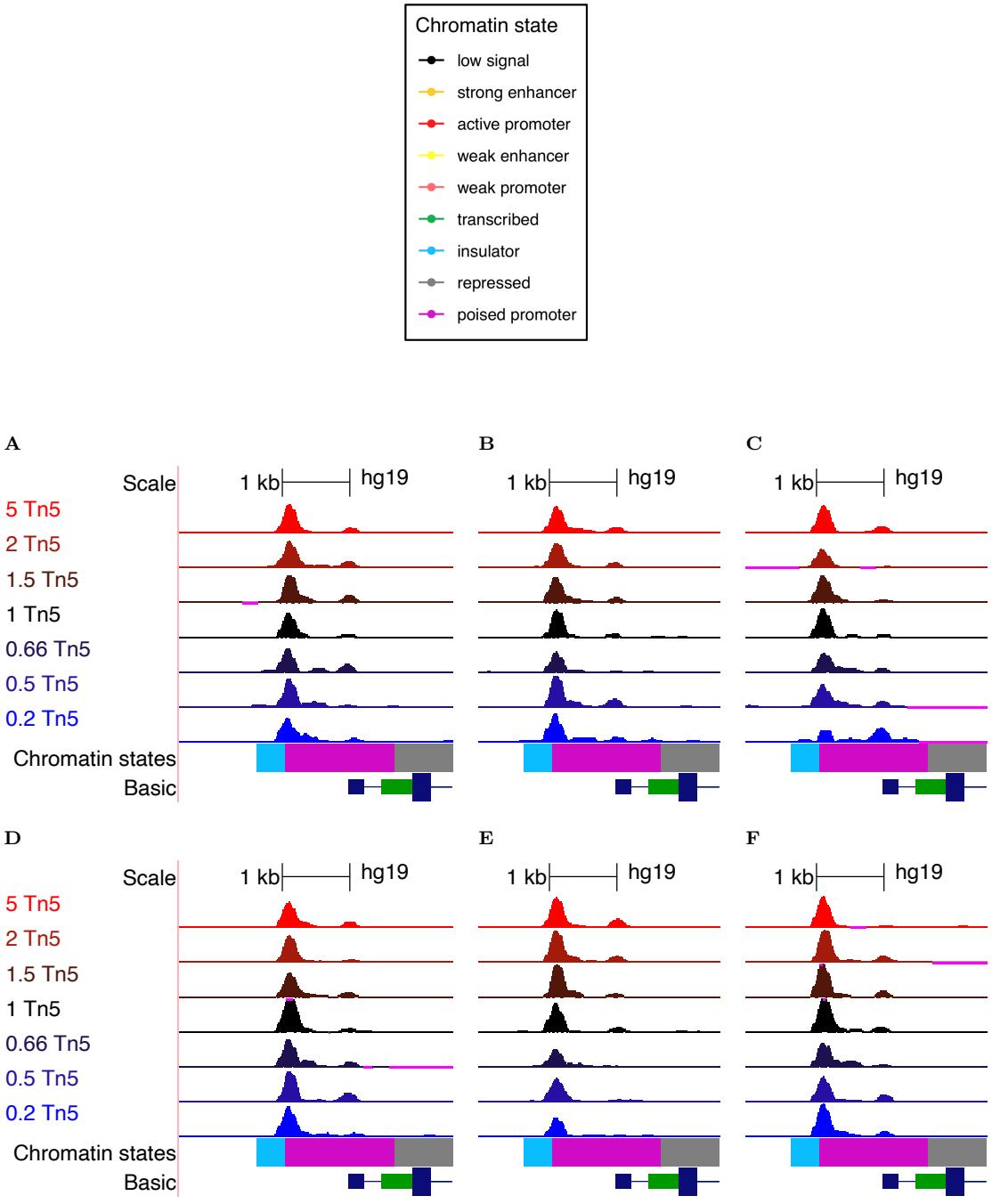


Figure S18. Example Tn5-insensitive peak, Related to Figure 2. UCSC genome browser screenshot displaying a Tn5-insensitive peak in the PCR-constant experiment. The six panels represent the six replicates (http://genome.ucsc.edu/)(Casper et al., 2018; Kent et al., 2002).

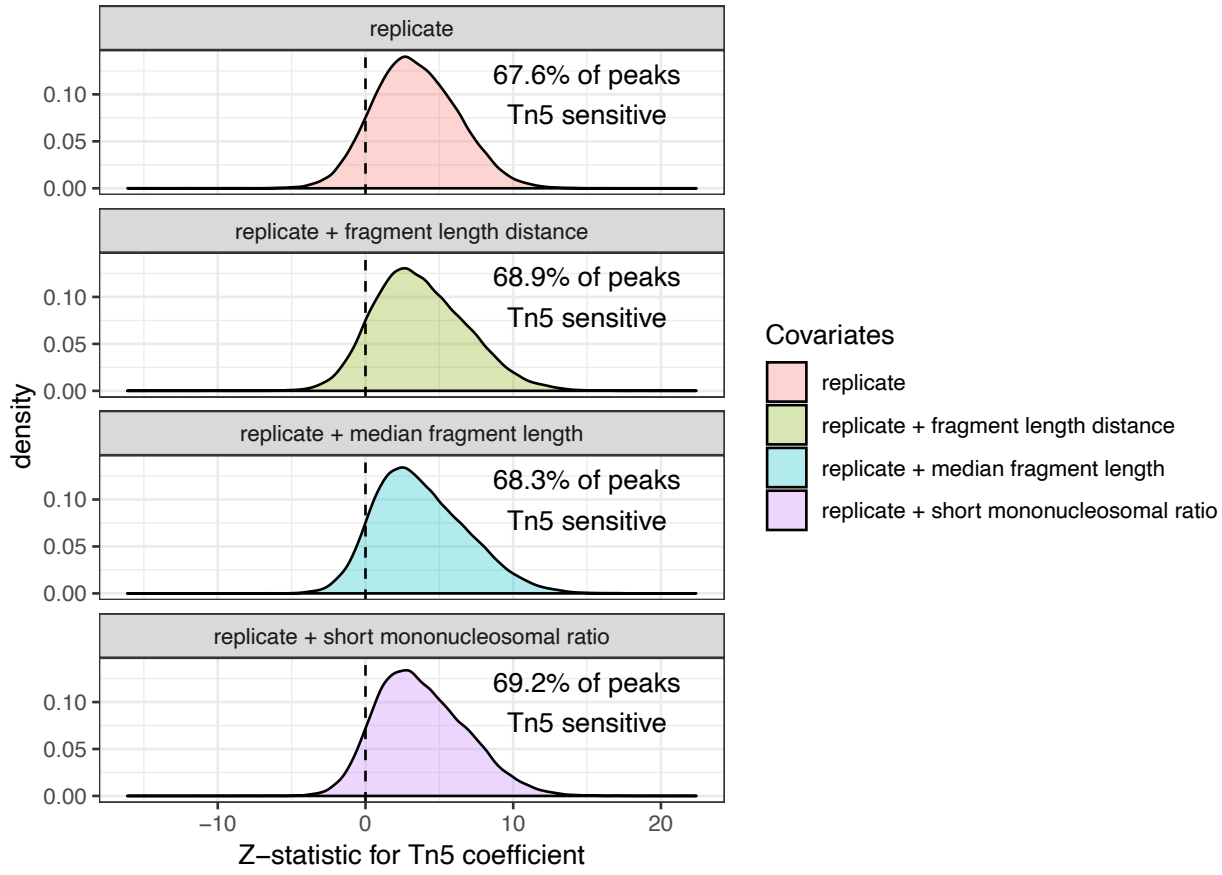


Figure S19. Most ATAC-seq peaks are Tn5 sensitive when PCR-cycles are allowed to vary, Related to Figure 2. Distribution of the Z-statistic for the coefficient of log2(relative Tn5 concentration) in the negative binomial GLM, in the PCR-variable experiment and when including the respective covariates. The distribution is shifted in the positive direction in all cases, indicating that many ATAC-seq peaks show increased signal (normalized to library size) as the concentration of Tn5 is increased.

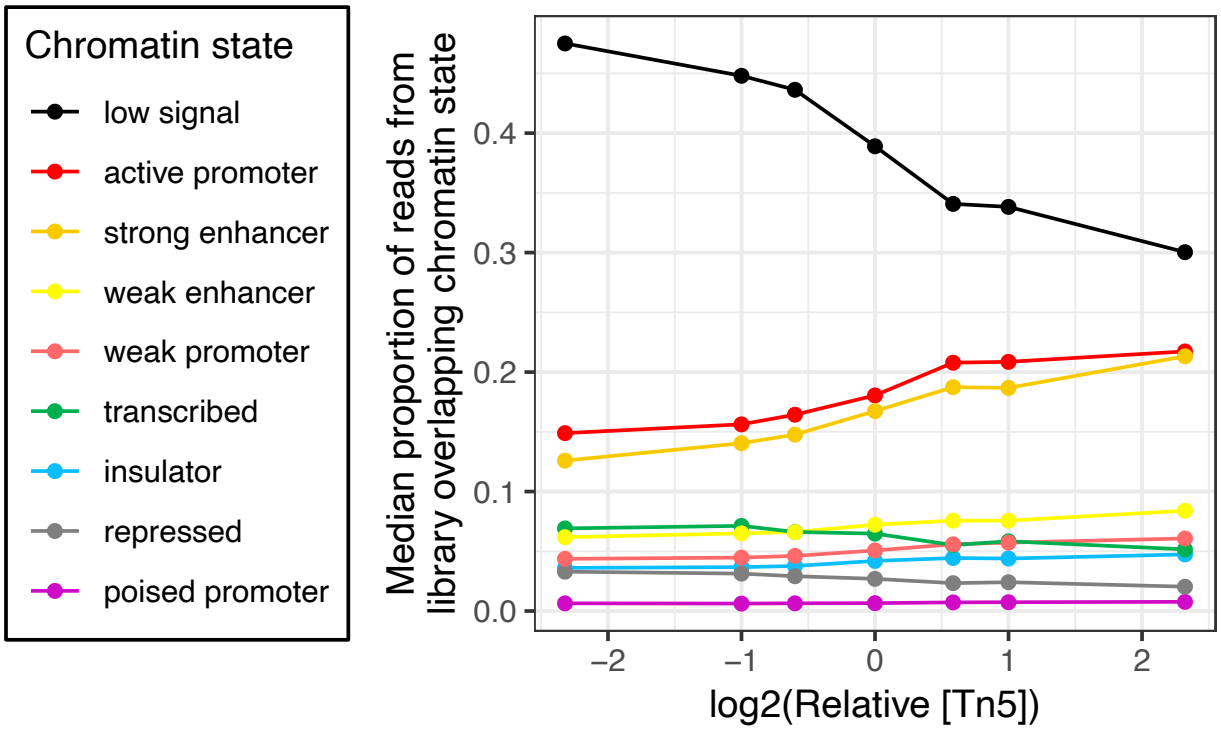


Figure S20. Proportion of reads overlapping with chromatin states as a function of Tn5 concentration (PCR-variable experiment), Related to Figure 2. The percentage of ATAC-seq reads falling into enhancer and active TSS chromatin states increases with increasing Tn5, while the percentage of reads falling into low signal regions decreases.

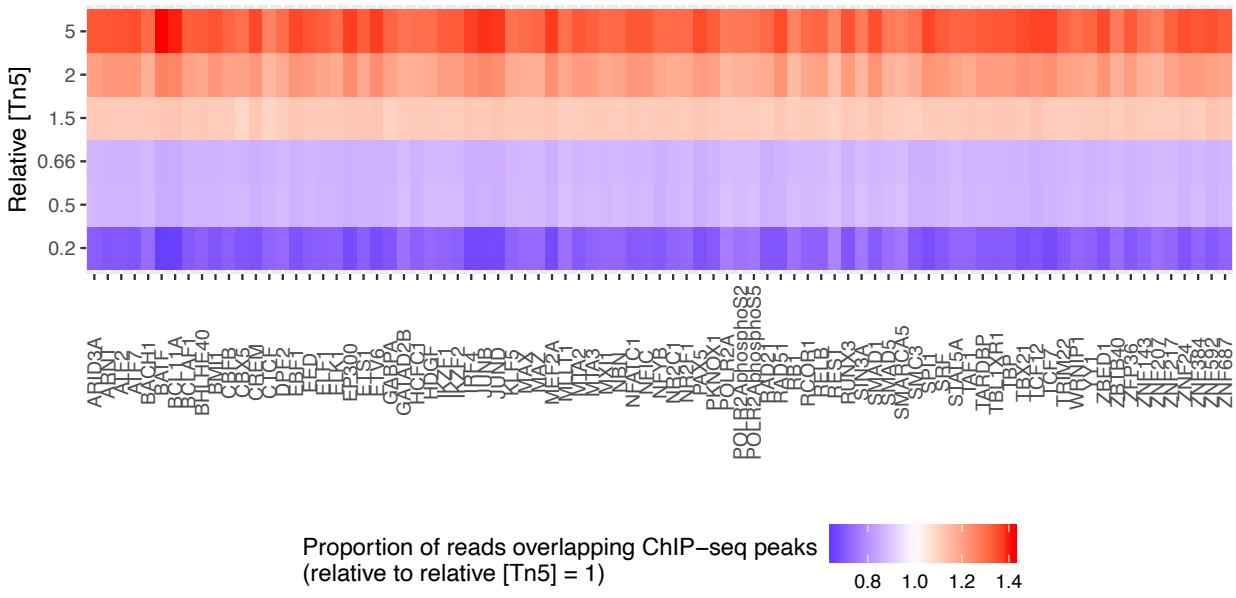


Figure S21. The percentage of ATAC-seq reads overlapping TF ChIP-seq peaks increases with increasing Tn5 (PCR-constant experiment), Related to Figure 2. For each TF and Tn5 concentration, the mean proportion of reads overlapping ChIP-seq peaks was calculated across the six replicates before normalizing to the 1X Tn5 values.

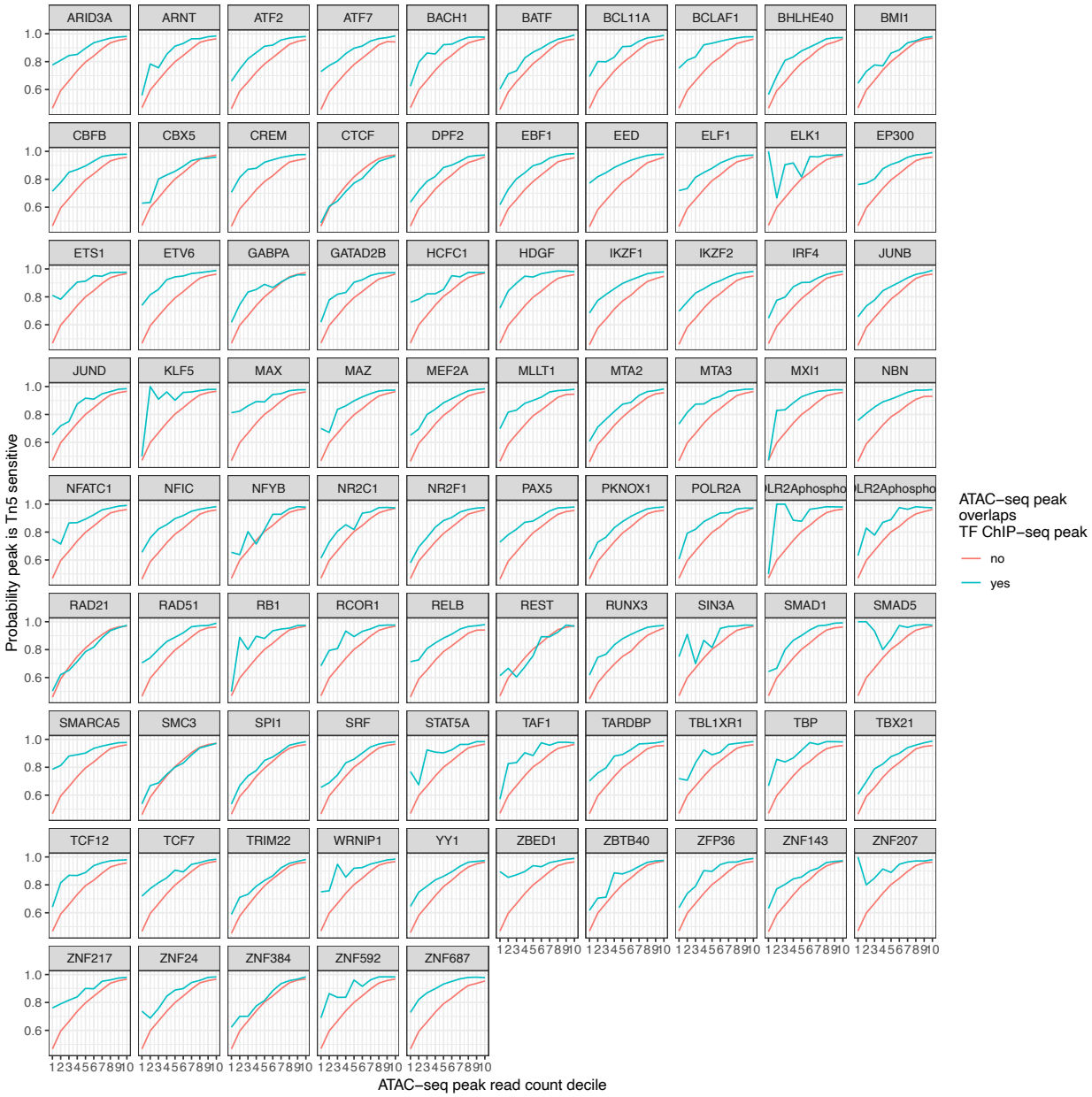


Figure S22. Binding of most TFs is associated with increased peak Tn5 sensitivity, Related to Figure 2. Peaks are binned into deciles based on the median read count at 1X Tn5 in the PCR-constant experiment (i.e., across 6 replicates). Out of all 85 TF ChIP-seq experiments, binding of 82 TFs was significantly (Bonferroni adjusted $p \leq 0.05$) associated with increased Tn5 sensitivity using the logistic regression approach (see Methods). The three exceptions are CTCF, RAD21, and REST.

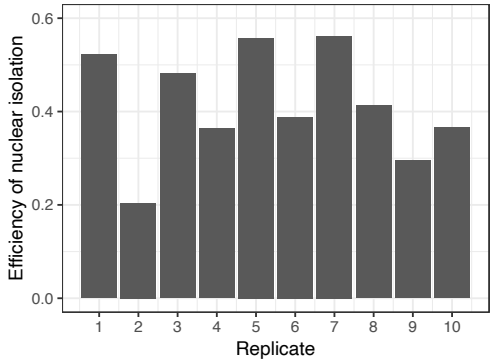


Figure S23. Efficiency of nuclear isolation, Related to STAR Methods. Nuclear isolation was performed on C2C12 cells. 250,000 cells were used as input. Efficiency is computed as (number of nuclei isolated) / (number of input cells).

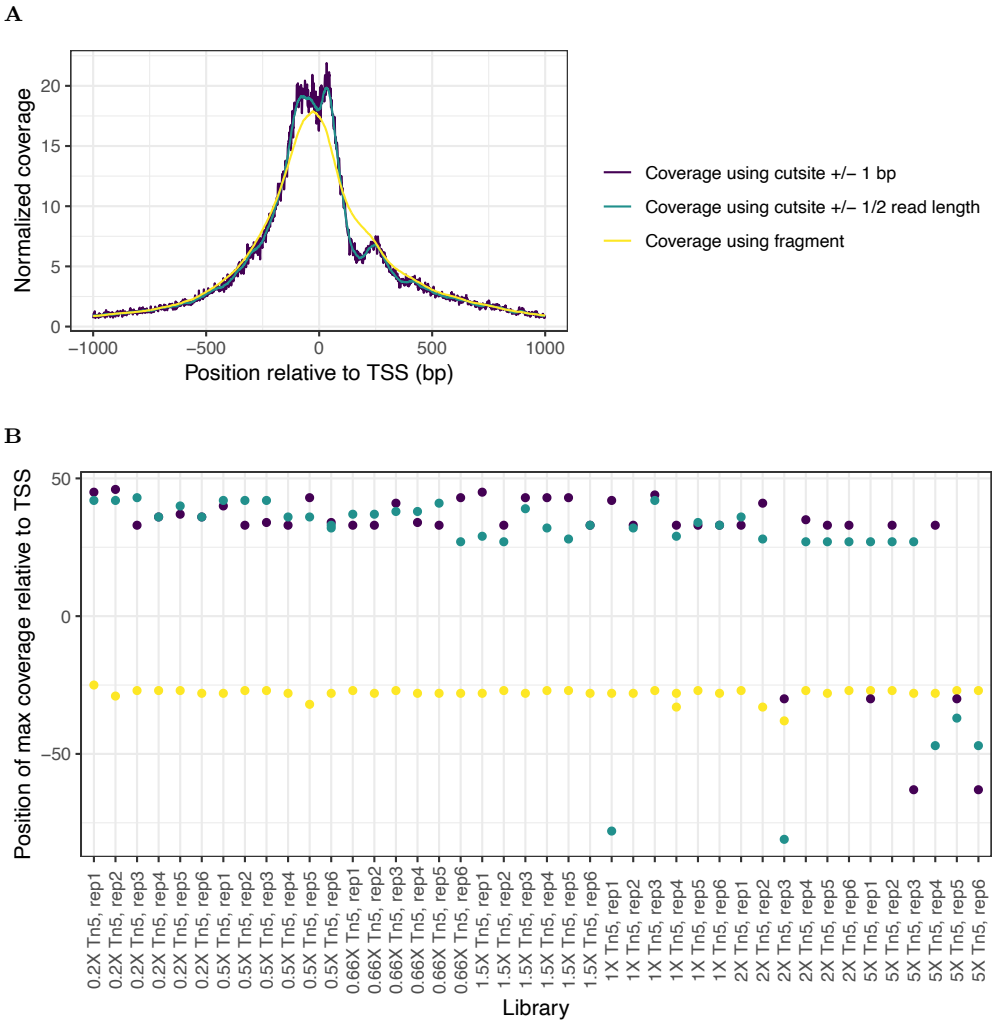


Figure S24. TSS enrichment curves and values are highly sensitive to calculation method, Related to STAR Methods. (A) Aggregate coverage around TSS for one library from the PCR-constant experiment, using different methods for calculating coverage. In one method ('coverage using fragment'), the coverage over each bp around the TSS is calculated using the actual entire ATAC-seq fragment. In a second method ('coverage using cutsite +/- 1bp'), coverage is calculated by determining ATAC-seq cutsites (the exact ends of each ATAC-seq fragment) and placing a fake 3-bp read over each cutsite and calculating TSS enrichment using these fake reads. In another method ('coverage using cutsite +/- 1/2 read length'), cutsite-centered fake reads are again utilized, but the length of these fake cutsite-centered reads are set to the read length. For all methods, 'normalized coverage' is calculated by dividing the coverage of each bp by the average coverage over the flanking regions (-1000 - -900, and 900 - 1000). TSS coverage shows clear differences in shape and smoothness of the curve. (B) Position of max coverage (around TSS) for libraries from the PCR-constant experiment, with coverages calculated as in (A). The maximum point on the curve shows greater stability across the libraries when calculating coverage using the ATAC-seq fragment rather than cutsite-centered coverages.

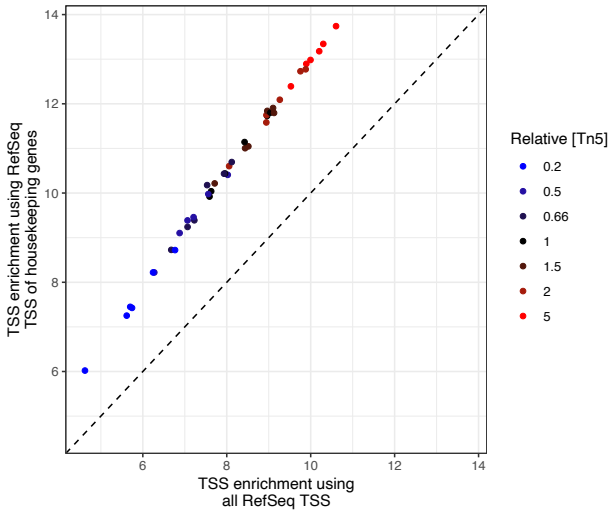


Figure S25. TSS enrichment is highly sensitive to TSS annotations used, Related to STAR Methods. TSS enrichment calculated using all RefSeq TSS or TSS of housekeeping genes only (PCR-constant experiment). Using housekeeping genes only results in systematically larger TSS enrichment values.

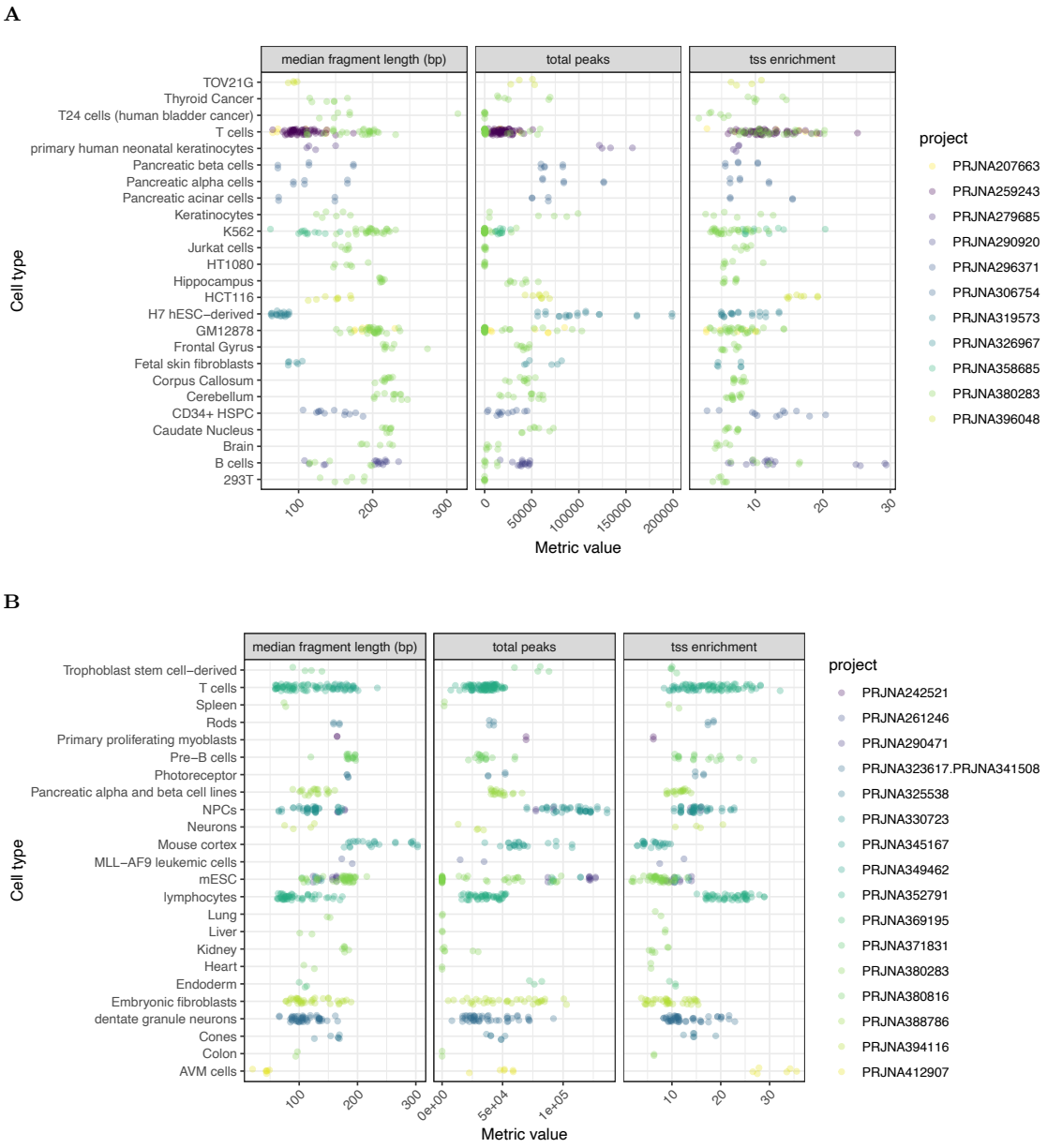


Figure S26. Distribution of several QC metrics across mouse and human cell types. Related to Figure 1. Median fragment length, TSS enrichment, and the number of peaks in the analyzed public bulk ATAC-seq data from (A) human and (B) mouse. Each point represents one library.

Metric	Notes	Abbreviation in Figure 1
Fragment length distribution		N/A (summarized with median_fragment_length)
Percentage of reads that are high-quality and autosomal		percent_hqaa
Percentage of reads properly paired and mapped		percent_properly_paired_and_mapped
Percentage of reads that aligned to autosomes that were duplicates		percent_autosomal_duplicate
Short-to-mononucleosomal-ratio	Ratio of number of fragments 50 - 100 bps long to the number of fragments 150-200 bps long	short_mononucleosomal_ratio
TSS enrichment	Coverage near TSSs (relative to coverage 900-1000 bp from TSSs)	tss_enrichment
Duplicate fraction in peaks	The fraction of properly paired reads that map within peaks and are duplicates (Requires peak file)	duplicate_fraction_in_peaks
Duplicate fraction outside of peaks	The fraction of properly paired reads that map outside of peaks and are duplicates (Requires peak file)	duplicate_fraction_not_in_peaks
Peak duplicate ratio	The ratio of 'duplicate fraction outside of peaks' to 'duplicate fraction in peaks' (Requires peak file)	peak_duplicate_ratio
Cumulative fraction of high-quality autosomal reads in peaks	Requires peak file	hqaa_overlapping_peaks_percent
Cumulative fraction of the genome that falls within peaks	Requires peak file	total_peak_territory
Distribution of mapping qualities		N/A
Number of total reads		total_reads

Percentage of alignments marked secondary		percent_secondary
Percentage of alignments marked supplementary		percent_supplementary
Percentage of alignments marked as duplicates		percent_duplicate
Mean mapping quality		mean_mapq
Median mapping quality		median_mapq
Percentage of reads unmapped		percent_unmapped
Percentage of reads with an unmapped mate		percent_unmapped_mate
Percentage of QC-fail reads		percent_qcfailed
Percentage of unpaired reads		percent_unpaired
Percentage of reads with mapping quality 0		percent_mapq_0
Percentage of reads that paired and mapped but in RF orientation		percent_rf
Percentage of reads that paired and mapped but in FF orientation		percent_ff
Percentage of reads that paired and mapped but in RR orientation		percent_rr
Percentage of reads that paired and mapped but on separate chromosomes		percent_mate_separate_chromosome
Percentage of reads that paired and mapped but too far from mate		percent_mate_too_distant
Percentage of reads that paired and mapped but not properly		percent_improperly_paired
Percentage of reads that aligned to autosomes		percent_autosomal
Percentage of reads that aligned to mitochondria		percent_mitochondrial
Percentage of reads that aligned to mitochondria that were duplicate		percent_mitochondrial_duplicate
Number of peaks called	Requires peak file	total_peaks
Fragment length distribution distance	See methods	fragment_length_distance
Max fraction of reads from a single autosome		max_fraction_reads_from_single_autosome

Table S1. List of ATAC-seq metrics displayed in interactive ataqv HTML report, related to STAR Methods and Figure 1.