bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CUT&Tag recovers up to half of ENCODE ChIP-seq peaks

Di Hu[1,2*], Leyla Abbasova[1,2*], Brian M Schilder[1,2], Alexi Nott[1,2], Nathan G Skene[1,2+], Sarah J Marzi[1,2+#]

[1] UK Dementia Research Institute, Imperial College London, London, UK
[2] Department of Brain Sciences, Imperial College London, London, UK
[*] These authors contributed equally: Di Hu, Leyla Abbasova
[+] These authors are joint senior authors: Nathan G Skene, Sarah J Marzi
[#] Correspondence: s.marzi@imperial.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Techniques for genome-wide epigenetic profiling have been undergoing rapid development toward recovery of high quality data from bulk and single cell samples. DNA-protein interactions have traditionally been profiled via chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq), which has become the gold standard for studying histone modifications or transcription factor binding. Cleavage Under Targets & Tagmentation (CUT&Tag) is a promising new technique, which enables profiling of such interactions *in situ* at high sensitivity and is adaptable to single cell applications. However thorough evaluation and benchmarking against established ChIP-seq datasets are still lacking. Here we comprehensively benchmarked CUT&Tag for H3K27ac and H3K27me3 against published ChIP-seq profiles from ENCODE in K562 cells. Across a total of 30 new and 6 published CUT&Tag datasets we found that no experiment recovers more than 50% of known ENCODE peaks, regardless of the histone mark. We tested peak callers MACS2 and SEACR, identifying optimal peak calling parameters. Balancing both precision and recall of known ENCODE peaks, SEACR without retention of duplicates showed the best performance. We found that reducing PCR cycles during library preparation lowered duplication rates at the expense of ENCODE peak recovery. Despite the moderate ENCODE peak recovery, peaks identified by CUT&Tag represent the strongest ENCODE peaks and show the same functional and biological enrichments as ChIP-seq peaks identified by ENCODE. Our workflow systematically evaluates the merits of methodological adjustments and will facilitate future efforts to apply CUT&Tag in human tissues and single cells.

Keywords

Epigenomics, histone modifications, CUT&Tag, ChIP-seq, NGS, ENCODE

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background

In recent years the field of epigenetics has garnered immense interest as a source of new insights into the mechanisms underlying human disease. Human disease risk variants identified through genome-wide association studies (GWAS) overwhelmingly localise to non-coding regions of the genome[1–3]. These risk variants appear to be enriched in gene regulatory regions. Chromatin dynamics at regulatory regions are governed by nucleosomes and their post-translational modifications, as well as interacting chromatin-associated complexes and transcription factors. Chromatin marks can define regions of activation and silencing, and mark transcriptional regulatory elements. These can be cell type-specific and are known to be dynamic during the course of embryonic development, aging, and disease progression[4]. Disease risk variants appear to be specifically enriched in active regulatory elements, which can be mapped using histone marks, such as H3K27ac. H3K27ac is a highly cell type specific histone modification and a marker of active enhancers and promoters[5]. In the brain it has been implicated in neurodegenerative and neuropsychiatric disorders, including Alzheimer's disease[6–8]. However, understanding the precise regulatory mechanisms underlying epigenetic regulation in brain disease and linking non-coding variants to disease phenotypes has been impeded by a lack of epigenomic annotations in disease and control tissue. Furthermore, the resources that do exist tend to use bulk tissues of heterogeneous organs, which is predominantly influenced by cell type composition and obscures cell type specific regulatory landscapes.

For many years, chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) has served as the gold standard method for epigenomic profiling. In ChIP-seq, chromatin is first cross-linked and solubilised, after which a primary antibody specific for the histone mark of interest enables immunoprecipitation of bound DNA[9]. However, it has potential limitations, such as low signal-to-noise ratio, epitope masking from fixation and cross-linking, and heterochromatin bias from chromatin sonication[10,11]. ChIP-seq poses challenges when working with low cell numbers, requiring approximately 1-10 million cells as input, with high demands on sequencing coverage, due to the low signal-to-noise ratio. In addition, ChIP-seq does not adapt well to single-cell applications due to its high cell input requirements and poor signal specificity. Cleavage Under Targets & Tagmentation (CUT&Tag) is a novel enzyme-tethering approach that has been presented as a streamlined, easily scalable, and cost-effective alternative to ChIP-seq. CUT&Tag has been reported to have superior chromatin mapping capabilities as compared to ChIP-seq at approximately 200-fold reduced cellular input and 10-fold reduced sequencing depth requirements[12]. CUT&Tag uses permeabilised nuclei to allow antibodies to bind chromatin associated factors, which enables tethering of protein A-Tn5 transposase fusion protein (pA-Tn5). Upon activation of pA-Tn5, cleavage of intact DNA and insertion of adapters (tagmentation) occurs for paired-end DNA sequencing. Following tagmentation, DNA fragments remain inside the nucleus making the

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

method amenable to single cell chromatin profiling applications, for example enabling individual sorting of nuclei and PCR barcoding. The increased signal-to-noise ratio of CUT&Tag for histone marks is attributed to the direct antibody tethering of pA-Tn5 and its integration of adapters *in situ* while it stays bound to the antibody target of interest during incubation. The process involves minimal sample loss with direct enzymatic end-polishing and ligation compared to regular library preparation protocols that result in sample loss, e.g. in ChIP-seq and CUT&RUN[12].

For ChIP-seq, experimental and analytical guidelines as well as datasets generated by the Encyclopedia of DNA Elements (ENCODE) consortium have served as a de facto gold standard reference panel to the field for years[13]. In contrast, as a relatively new method, CUT&Tag lacks equivalent systematic optimisation or benchmarking against existing datasets and there is little established consensus regarding experimental recommendations and data analysis workflows. Systematic benchmarking of bulk CUT&Tag will serve as a foundation to optimally guide experimental design and analysis of its single cell applications. Here we undertook experimental optimisations and systematic benchmarking of CUT&Tag against ENCODE in human K562 cells for histone modifications H3K27ac and H3K27me3 to serve as a guide for the design and analysis of future CUT&Tag studies. Since the development of CUT&Tag has primarily assessed methyl marks[14], we focused in-depth on underexplored H3K27ac, testing multiple ChIP-grade antibody sources[6,7,15,16], antibody dilutions, histone deacetylase inhibitors (HDACi), as well as PCR parameters, and DNA extraction methods for library preparation (**Fig. 1a**). Experimental outcomes were evaluated by quantitative polymerase chain reaction (qPCR) and paired-end genomic sequencing. Our computational workflow served to iteratively guide experimental optimisations, appraise CUT&Tag data quality, and benchmark CUT&Tag performance against ENCODE ChIP-seq profiles (**Fig. 1b**). We explored the suitability of different peak calling approaches (SEACR and MACS2), and the effects of inclusion versus exclusion of duplicate reads. We characterised the similarities and differences between CUT&Tag and ENCODE based on parameters including read- and peak-level correlation, regulatory element annotation, gene ontology enrichment, and transcription factor binding motif (TFBM) analysis.

Results

Overview of systematic H3K27 CUT&Tag experimental design and analysis

To benchmark the performance of CUT&Tag against established ENCODE ChIP-seq, we profiled histone modifications H3K27ac, a marker of active enhancers and promoters, and H3K27me3, associated with heterochromatin and assessed in the original series of papers

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

introducing CUT&Tag[12,14,17]. Both histone modifications were characterized in K562 cells, generating a total of 30 new CUT&Tag sequencing datasets. We undertook systematic experimental optimisations for H3K27ac CUT&Tag testing multiple ChIP-grade antibody sources[6,7,15,16], antibody dilutions (1:50, 1:100, 1:200), as well as different PCR cycle numbers, DNA extraction methods for library preparation, and histone deacetylase inhibitors (HDACi) (**Fig. 1a**). Primary conditions were first validated by performing qPCR using positive and negative control primers designed based on ENCODE ChIP-seq peaks (**Table 1**). The best conditions were subsequently subjected to paired-end sequencing. Our computational workflow iteratively guided experimental optimisations, assessed data quality, and benchmarked CUT&Tag performance against ENCODE ChIP-seq (**Fig. 1b**).

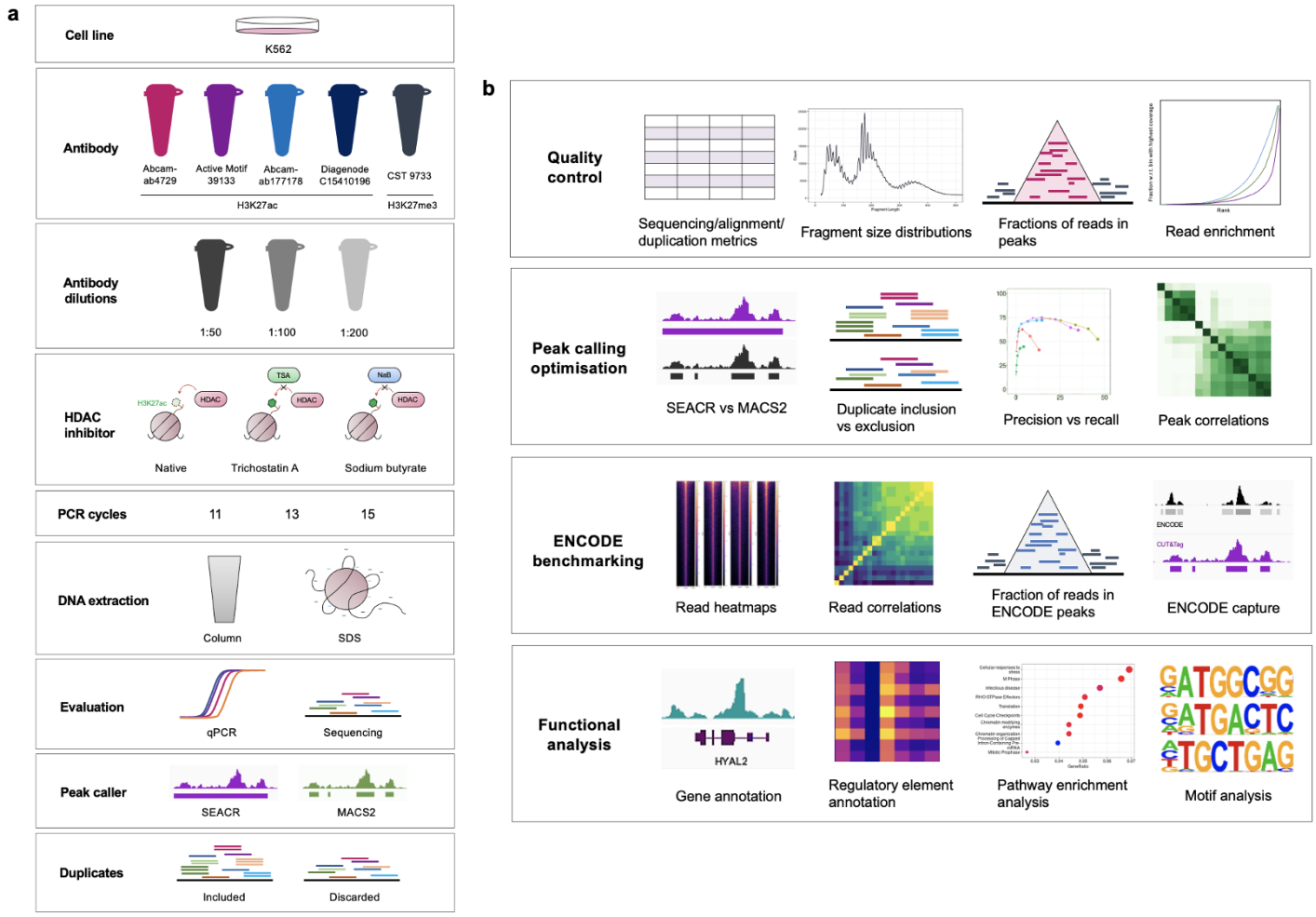


Figure 1. Overview of experimental design and computational benchmarking. a, Summary of experimental design: Five antibodies were tested at dilutions 1:50, 1:100 and 1:200, and 11, 13, or 15 PCR cycles for library preparation. H3K27ac libraries were assessed with and without

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HDAC inhibitor Trichostatin A (TSA; 1 μM), or sodium butyrate (NaB; 5 mM). Column- and SDS-based DNA extraction methods were compared. Antibody performance was assessed by qPCR and sequencing, and sequenced reads were processed with and without duplicates using peak callers SEACR and MACS2. **b**, Summary of analytical approaches: Analysis comprised quality control of sequencing data, optimisation of peak calling approaches with both peak callers, and comparison between CUT&Tag and ENCODE datasets at the level of reads, peaks, and functional annotation.

Quality control of CUT&Tag data

We first assessed four ChIP-seq grade H3K27ac antibodies across three dilutions (1:50, 1:100 and 1:200) by qPCR, using primers designed to amplify regions corresponding to genes falling into the most significant ENCODE peaks (positive controls: *ARGHAP22*, *COX4I2*, *MTHFR*, *ZMYND8*) versus least significant ENCODE peaks (negative controls: *KLHL11*, *SIGIRR*) (**Methods**; **Table 1**; **Fig. 2a**). Based on the outcome, we selected Abcam-ab4729 (1:100), Diagenode C15410196 (1:50 and 1:100), Abcam-ab177178 (1:100), and Active Motif 39133 (1:100) for sequencing. These antibodies will be henceforth referred to as Abcam-ab4729, Diagenode, Abcam-ab177178, and Active Motif. H3K27me3 CUT&Tag was profiled using ChIP-grade antibody Cell Signaling Technology-9733 at dilution 1:100 as previously recommended[14]. In house samples were compared with published CUT&Tag[14] and CUT&RUN[18] data from the research group who originally developed these methods.

Preliminary analysis of sequencing data revealed high duplication rates across samples (min: 55.49%; max: 98.45%; mean: 82.25%; **Supplementary Table 1**). We therefore asked whether such data could still be interpreted by CUT&Tag analysis, as high duplication rates have been observed with different sample types, for example mouse frozen brain tissue (assessed in house). We first quantified fragment length as a quality control metric. Fragment sizes observed in our samples were comparable to CUT&Tag in human nuclei, with an abundance of fragments at around 180bp in size, reflecting the length of DNA from a single nucleosome (**Fig. 2b; Supplementary Fig. 1a**)[19,20]. We also observed short fragments (<100bp) similar to previous CUT&Tag data[14], potentially caused by tagmentation of open chromatin[21]. Shorter fragments were not more abundant in duplicate-containing samples, suggesting that these are not a consequence of PCR amplification bias[22].

We next evaluated signal-to-noise quality by calculating the fractions of reads in peaks (FRiPs) using peaks we defined with our dataset, as well as pre-defined ENCODE peaks. Specifically, we compared our data with ENCODE H3K27ac narrow and H3K27me3 broad peak sets (**Fig. 2c**). To identify peaks in our CUT&Tag data we used two analytical approaches: 1) MACS2, a standard peak caller for ChIP-seq data used by ENCODE and also applied to recent CUT&Tag datasets, and 2) SEACR, an algorithm developed specifically to detect peaks

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in high signal-to-noise data, such as CUT&RUN and CUT&Tag[23,24]. With few exceptions, the fractions of reads in CUT&Tag called peaks and ENCODE peaks of the same mark were comparable. Of the antibodies tested, Abcam-ab4729 and Diagenode at 1:50 dilution showed the highest percentage of reads falling into published ENCODE H3K27ac peaks (29.9% and 27.4%, respectively), as well as newly identified CUT&Tag peaks (SEACR: 35.3% and 32.9%, MACS: 34.1% and 30.9%, respectively; **Fig. 2c; Supplementary Fig. 1b**). However, all H3K27ac CUT&Tag FRiP scores were markedly lower than the ENCODE H3K27ac ChIP-seq score of 42%. Only H3K27me3 CUT&Tag FRiPs in ENCODE peaks slightly outperformed ENCODE H3K27me3's FRiP score of 66%, but this was not reflected in the enrichment across empirically identified SEACR and MACS2 peaks (**Fig. 2c**). Of note, H3K37me3 CUT&Tag shows highly specific enrichment for ENCODE H3K27me3 peaks, while H3K27ac shows more unspecific enrichment at ENCODE H3K27ac and H3K27me3 peaks, both for in house and published data. Duplicate inclusion did not make a significant difference to peak calling except in cases where peaks of highly duplicated samples were called using MACS2. Visualisation in the Integrative Genomics Viewer[25] showed variable noise levels for H3K27ac CUT&Tag relative to ENCODE ChIP-seq, while H3K27me3 CUT&Tag exhibited consistently lower background noise (**Fig. 2d**).

Since CUT&Tag and other Tn5 transposase-based methods may be susceptible to open chromatin bias[26], we assessed the proportion of reads falling into open chromatin ATAC-seq peaks (**Fig. 2e; Supplementary Fig. 1c**). Nearly 70% of ENCODE H3K27ac peaks overlapped with ATAC-seq peaks. Therefore ENCODE H3K27ac peaks were further subset to obtain those exclusive to the histone modification, exclusive to ATAC, and shared by both (**Fig. 2e**). This revealed that nearly all H3K27ac regions profiled by CUT&Tag fall into open chromatin regions shared with ATAC, but not into ATAC-only regions. For H3K27me3, around 8% of reads fell into ATAC peaks and this percentage was similar across comparison data. Removing short fragments (<100 bp) reduced the proportion of overlap with ATAC peaks to 4.6% for H3K27me3 (42.5% reduction), while short fragment exclusion had a negligible effect on H3K27ac/ATAC overlap.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

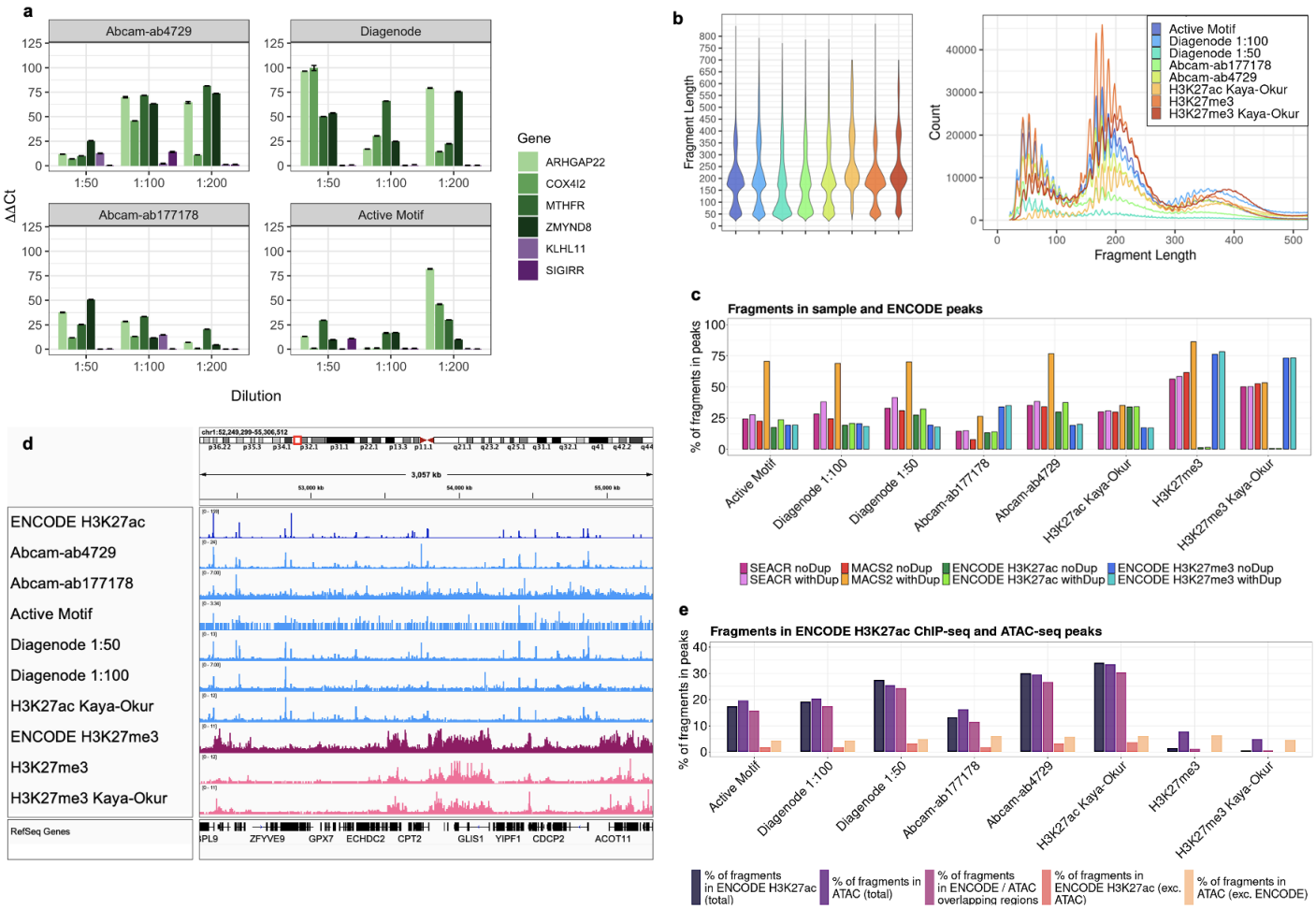


Figure 2. Quality control metrics for CUT&Tag and comparison data. a, Results of qPCR amplification of genes falling into most significant ENCODE H3K27ac peak regions in CUT&Tag samples; **b**, CUT&Tag fragment size distributions; **c**, Fractions of reads in peaks (FRiPs) called on each sample with SEACR and MACS2, as well as in ENCODE reference H3K27ac and H3K27me3 peaks; **d**, Integrative Genomics Viewer (IGV)[25] tracks showing sample read pileups; **e**, Fractions of sample reads in ENCODE H3K27ac ChIP-seq and ATAC-seq peaks.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CUT&Tag peak calling with SEACR and MACS2

We next assessed different peak callers and their settings to identify which would be most suitable for CUT&Tag. We evaluated the performance of both SEACR and MACS2, which were developed for CUT&RUN[24] and ChIP-seq[27], respectively. Parameter optimisation was conducted based on precision (the proportion of CUT&Tag peaks falling into ENCODE peaks of the same histone modification) and recall (the proportion of ENCODE peaks captured by CUT&Tag) (**Fig. 3a; Supplementary Fig. 2**), with the aim of maximising ENCODE capture while maintaining high precision (>80%). SEACR peaks were called using the stringent setting and thresholds of 0.01, 0.03, 0.05 and 0.1, as the relaxed setting was found to be too permissive, with precision scores consistently falling below the 80% threshold. MACS2 peaks were called using the narrow peak setting, with p- and q-values between $1x10^{-5}$ and 0.1. These settings were also tested with local lambda deactivated to replicate the global background approximation employed by SEACR. On the basis of precision and recall analysis, optimum SEACR H3K27ac peaks were called using the stringent setting and a threshold of 0.01. For MACS2, H3K27ac narrow peaks were called with local lambda deactivated and a q-value of $1x10^{-5}$. As a broader histone mark, H3K27me3 peaks were called with the same settings, but using the broad flag in MACS2, or by increasing the SEACR threshold to 0.1. Based on these parameters, SEACR peaks were called with slightly higher precision compared to MACS2. Furthermore, analysis of FRiPs revealed that SEACR peak calling was unaffected by duplicate inclusion, while MACS2 showed very high duplicate sensitivity, identifying an excessive number of spurious peaks despite stringent parameters (**Supplementary Table 2**). CUT&Tag peaks identified with both peak callers appeared to miss some ENCODE peaks of lower signal intensity, particularly in regions of elevated background noise (**Fig. 3b**). While levels of noise in H3K27ac samples were variable and elevated relative to ENCODE, H3K27me3 CUT&Tag showed lower background compared to both H3K27ac CUT&Tag and H3K27me3 ChIP-seq. Another difference between the peak callers was that SEACR defines wider peaks than MACS2, where multiple MACS2 peaks can correspond to a single SEACR peak (**Fig. 3b**). This might make it difficult to detect subtle, local changes in histone modifications, but produces peak ranges more suitable for broader marks such as H3K27me3. Peak correlations, quantified with DiffBind[28], revealed that SEACR calls peaks with higher consistency than MACS2, whereas MACS2 CUT&Tag peaks show greater similarity to ENCODE peaks, which are also called using MACS2 (**Fig. 3c**). Following subsampling to the same read depths, read profiles around peak summits confirmed that at the selected settings, peaks called by SEACR and MACS2 possessed similar read densities (**Fig. 3d**). Furthermore, the best-performing CUT&Tag samples showed greater peak read enrichment compared to ENCODE. However, this does not necessarily translate to higher FRiPs, which depend on abundance of reads, a metric not incorporated into peaks (**Fig. 3d**). From this overall assessment, we chose to call peaks predominantly using SEACR from this point onward.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

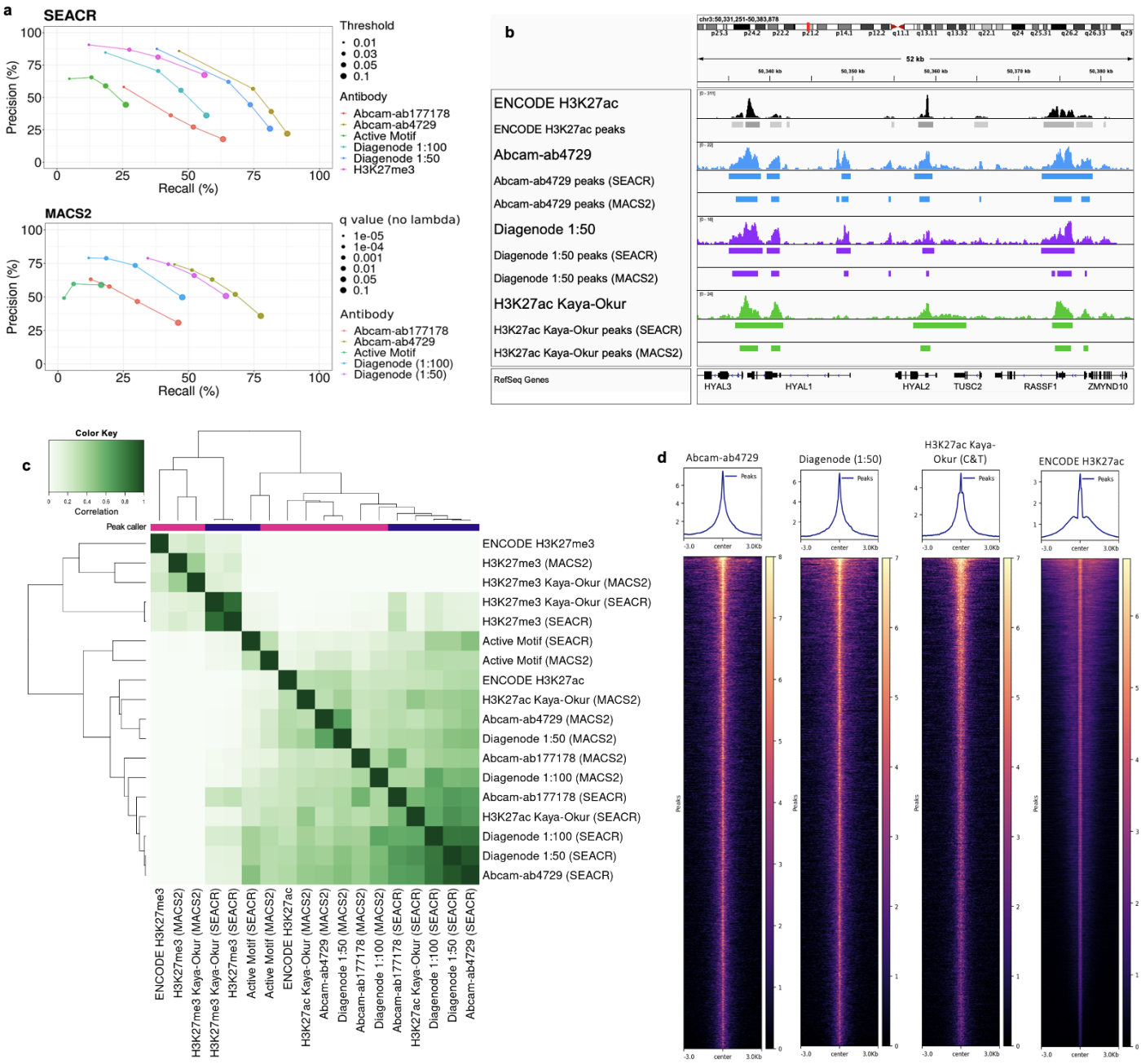


Figure 3. Peak calling with SEACR and MACS2. a, Panel showing precision/recall plots for SEACR and one of several MACS2 tests employed with different parameters; **b**, Select H3K27ac sample tracks visualised alongside called SEACR and MACS2 peak ranges; **c**, Correlations between sample peaks called with both peak callers (pink corresponds to MACS2, and blue corresponds to SEACR peaks); **d**, Average read enrichments around summits of select peaks called with SEACR and ENCODE H3K27ac, all subsampled to 2 million reads.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Benchmarking of CUT&Tag against ENCODE ChIP-seq

We next proceeded to benchmark CUT&Tag against current gold standard ENCODE ChIP-seq profiles. First, in an attempt to minimise any bias potentially incurred by peak calling, samples were correlated on the basis of read counts in different genomic regions: ENCODE H3K27ac peak ranges (**Fig. 4a**), the hg19 reference genome[29] partitioned into 500bp bins (**Supplementary Fig. 3a**), and ENCODE H3K27me3 peak ranges (**Supplementary Fig. 3b**). While genome-wide correlation revealed that similarity between CUT&Tag (and CUT&RUN) samples was markedly higher than that between CUT&Tag and ENCODE ChIP-seq, CUT&Tag-ENCODE correlations were much enhanced when the analysis was restricted to ENCODE H3K27ac and H3K27me3 peak regions for the corresponding mark. This is likely because genome-wide comparison factors in many regions that are devoid of true signal or contain noise, adding unwanted variability to the correlation analysis. On the whole, there was good correspondence between read- (**Fig. 4a**) and peak-level (**Fig. 3c**) correlations. To determine the extent to which CUT&Tag recovers known ChIP-seq peaks, the GenomicRanges[30] package was used to calculate the proportion of ENCODE peaks overlapping with CUT&Tag (recall), and the proportion of sample peaks overlapping with ENCODE (precision; **Fig. 4b; Supplementary Fig. 3c, 3d**). Overall, MACS2 achieved slightly higher recall at the expense of precision, which could be matched by marginally raising the SEACR threshold. The maximum ENCODE capture for H3K27ac was approximately 45% attained using the best-performing antibody Abcam-ab4729. This empirical ceiling was also observed across the comparison of previously published H3K27ac CUT&Tag and CUT&RUN samples (**Supplementary Fig. 3c, 3d**). Despite the much greater number of peaks called, H3K27me3 CUT&Tag only reached an ENCODE coverage ceiling of approximately 50%. Therefore CUT&Tag recovers up to half of ENCODE ChIP-seq peaks.

To facilitate comparison between different samples, precision and recall were compounded into a single metric representing a weighted average of the two measures, the F1-score (**Methods**; **Supplementary Fig. 3e**). This approach excludes true negative peaks, which might distort the score since they occupy the vast majority of the genome. We confirmed Abcam-ab4729 as the best performing antibody, followed by Diagenode at 1:50 dilution, while also highlighting that peak calling with SEACR resulted in better F1-scores than MACS2 on average (**Supplementary Fig. 3e**). For antibody selection, ENCODE coverage was re-calculated with all samples subsampled to the same read depth. This confirmed that Abcam-ab4729 still outperformed other tested antibodies (**Supplementary Fig. 3f**). Duplicates in CUT&Tag data may have biological relevance, potentially arising from tagmentation events that recur in the same place by chance. Thus, peaks were called with versus without duplicates. Duplicate inclusion in samples with relatively low duplication rates had little to no effect on precision or recall; however, in samples more abundant in duplicates, peak calling

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with duplicates using MACS2 was compromised by detection of abundant spurious peaks (**Fig. 4b**). Any improvements seen for ENCODE recall came at the expense of detection of a large number of false positive peaks. We thus recommend exclusion of duplicates.

To further characterise the ENCODE H3K27ac peaks that were captured and missed by each CUT&Tag sample, the -log(q) values (significance) of the ENCODE peaks (from original peak calling performed in ENCODE) falling into these groups were compared (**Fig. 4c**). This revealed that CUT&Tag captures the most significant peaks, i.e. those with lower q-values. We supplemented this by analysing ATAC-seq read counts (**Fig. 4d**), as in principle the H3K27ac mark should coincide with open chromatin regions. This showed that the ENCODE peaks captured by CUT&Tag samples contain more ATAC reads even when corrected for total base count of the captured and missed ENCODE peak sets, supporting the notion that CUT&Tag detects more prominent H3K27ac peaks, or at least those that are more likely to also be detected by an orthogonal epigenomic method. In all cases the differences between the q-values and ATAC-seq read counts in captured and missed ENCODE peaks were statistically significant ($p<0.0001$ across all q-value pairs and $p=0.0002$, respectively; t-test). Due to lower background, CUT&Tag should allow for higher data quality at read depths lower than those required for ChIP-seq, as previously shown for methyl histone marks[14]. To test whether H3K27ac CUT&Tag might have an advantage at lower read depths, FRiPs were calculated at 0.5, 1, 1.5 and 2 million unique reads (**Fig. 4e**). This analysis showed that the best H3K27ac CUT&Tag antibodies still produced fewer reads in peaks than ENCODE ChIP-seq at low read depth, although CUT&Tag peaks showed greater read enrichment around peak summits compared to ChIP-seq peaks (**Fig. 4f**). Cumulative sample read enrichments at equal read depths revealed that the read distributions of H3K27ac CUT&Tag and H3K27ac ENCODE ChIP-seq samples were comparable, while H3K27me3 CUT&Tag samples showed more constrained read distributions than ENCODE H3K27me3 (**Fig. 4g**). In agreement with the FRiPs analyses, this indicates that even at equal read depths, H3K27ac – but not H3K27me3 – CUT&Tag samples contain more off-target reads than ENCODE ChIP-seq.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

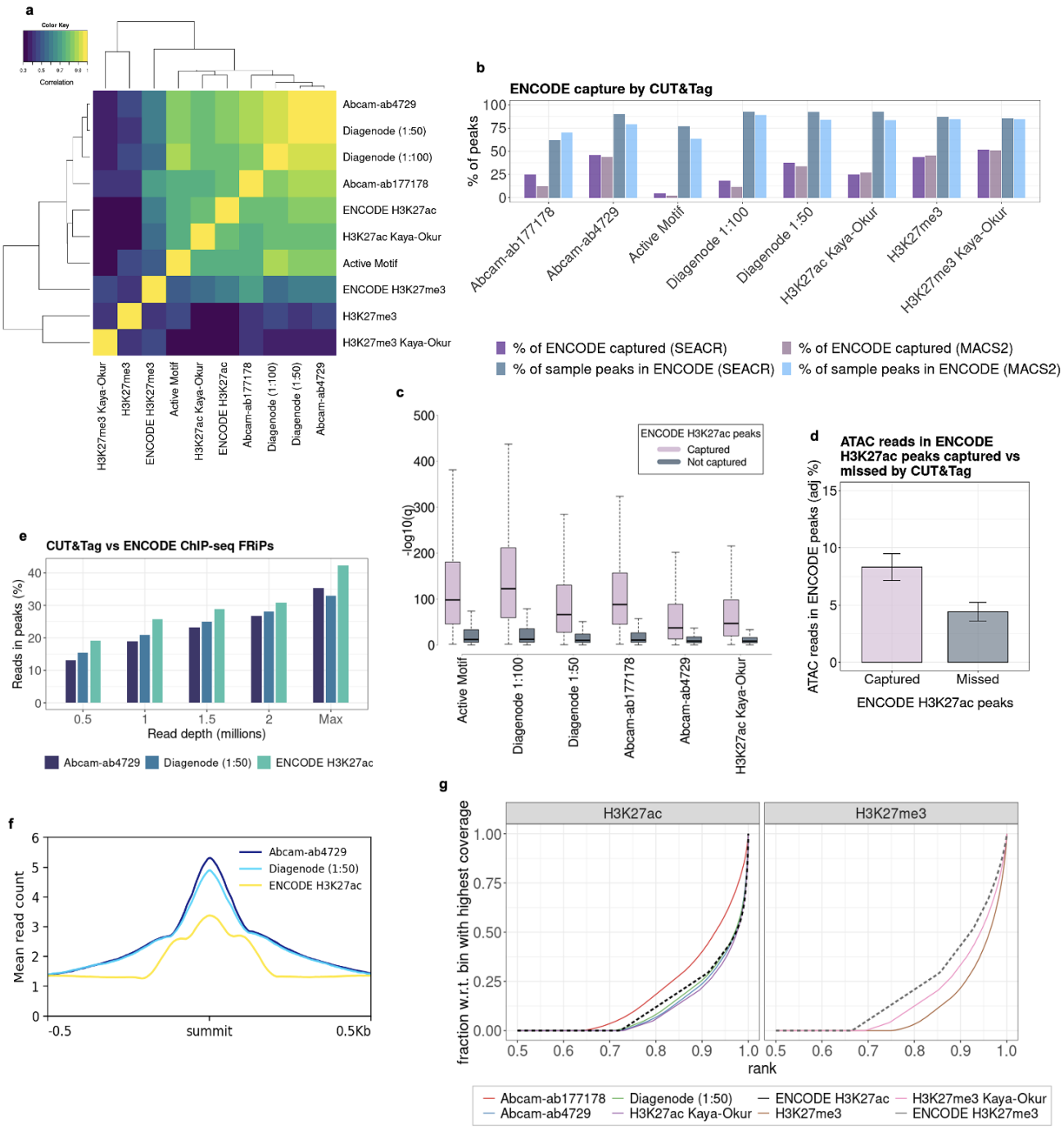


Figure 4. Benchmarking of CUT&Tag against ENCODE. a, Correlation of sample read counts across ENCODE H3K27ac peak ranges; **b**, Percentages of CUT&Tag peaks falling into ENCODE H3K27ac, and of ENCODE H3K27ac peaks captured by CUT&Tag, with and without duplicates; **c**, Comparison of -$\log_{10}$(q) values of ENCODE H3K27ac peaks captured and not captured by CUT&Tag; **d**, Measures of relative abundance of ATAC-seq reads in ENCODE H3K27ac peaks captured and not captured by CUT&Tag; **e**, FRiPs of select top H3K27ac

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CUT&Tag sample peaks and ENCODE H3K27ac peaks at different read depths, and at the original sequencing depth ('Max'); **f**, Average read enrichments around peak summits of top H3K27ac CUT&Tag and ENCODE ChIP-seq peaks; **g**, Fingerprint plots of H3K27ac (left) and H3K27me3 (right) CUT&Tag and ENCODE, all subsampled to 2 million reads. The bin with the highest coverage refers to the 1000bp interval containing the most reads; graphs show the cumulative read counts within ranked 1000bp bins, as a fraction of the read count in the highest-scoring bin.

Experimental optimisation of CUT&Tag

Since H3K27ac is dynamically deposited and removed by histone acetylases and deacetylases (HDACs), its chromatin mapping methods can potentially benefit from adding HDAC inhibitors (HDACi) to eliminate residual deacetylase activity and thereby stabilise acetyl marks. This is particularly relevant for CUT&Tag, which is carried out under native conditions where residual HDAC activity may have a greater impact. To test whether the addition of a potent HDAC inhibitor improves data quality and ENCODE coverage of previously tested antibodies, H3K27ac CUT&Tag was performed with addition of Trichostatin A (TSA; 1 μM). This data was compared to original samples scaled to the same read depths. Addition of HDACi TSA improved neither signal to noise ratio (**Fig. 5a**) nor ENCODE coverage (**Fig. 5b, 5c**). H3K27ac CUT&Tag was also attempted with addition of sodium butyrate (NaB; 5mM) and libraries were evaluated by qPCR amplification to find no improvement in CUT&Tag binding signal (**Supplementary Fig. 4**).

CUT&Tag library preparation was initially carried out with 15 PCR cycles, as per the original protocol[14]. To test whether this contributed to high numbers of duplicate reads we carried out CUT&Tag library preparation at 11 and 13 PCR cycles. In addition to varying cycle numbers, we also appraised SDS-based vs column-based methods of DNA extraction (see methods). All samples were downsampled to the shared minimum read depth (2.6 million paired-end reads) to compare duplication rates and ENCODE coverage, while total unique fragments were compared at maximum read depth (**Fig. 6**). Varying PCR cycles while employing SDS-based DNA extraction produced mixed changes in duplication rate, whereas samples obtained with column-based extraction predominantly showed an increase in duplication rate from 11 to 13 PCR cycles (**Fig. 6a**). Overall, the greatest numbers of unique fragments were generated using 15 PCR cycles and SDS-based DNA extraction (**Fig. 6b**). Almost all samples captured ENCODE peaks with high precision (**Fig. 6c**) while the superior unique fragment yield at 15 PCR cycles did not translate into improved ENCODE coverage after downsampling (**Fig. 6d**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

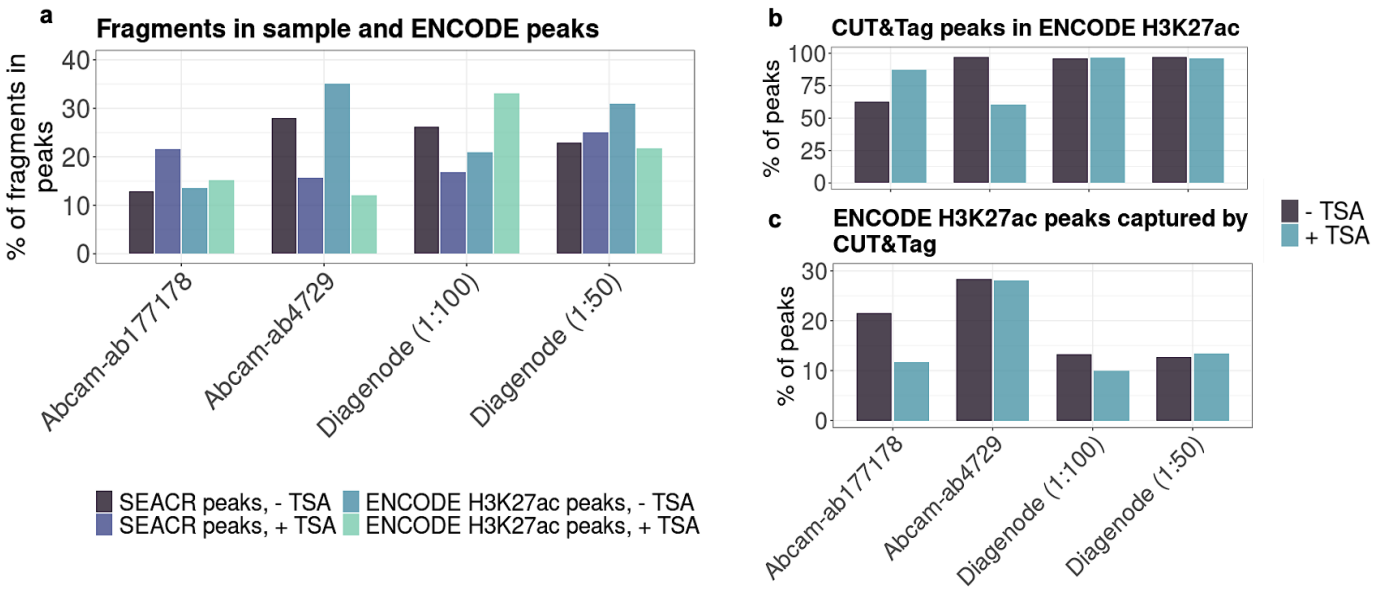


Figure 5. H3K27ac CUT&Tag performance with and without TSA. a) Comparisons between fractions of reads in sample peaks called with SEACR, and in ENCODE H3K27ac ChIP-seq peaks; b) and c) show ENCODE H3K27ac capture metrics with and without TSA. Reads from each antibody tested without TSA were subsampled to the same read depth as the same antibody tested with TSA.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

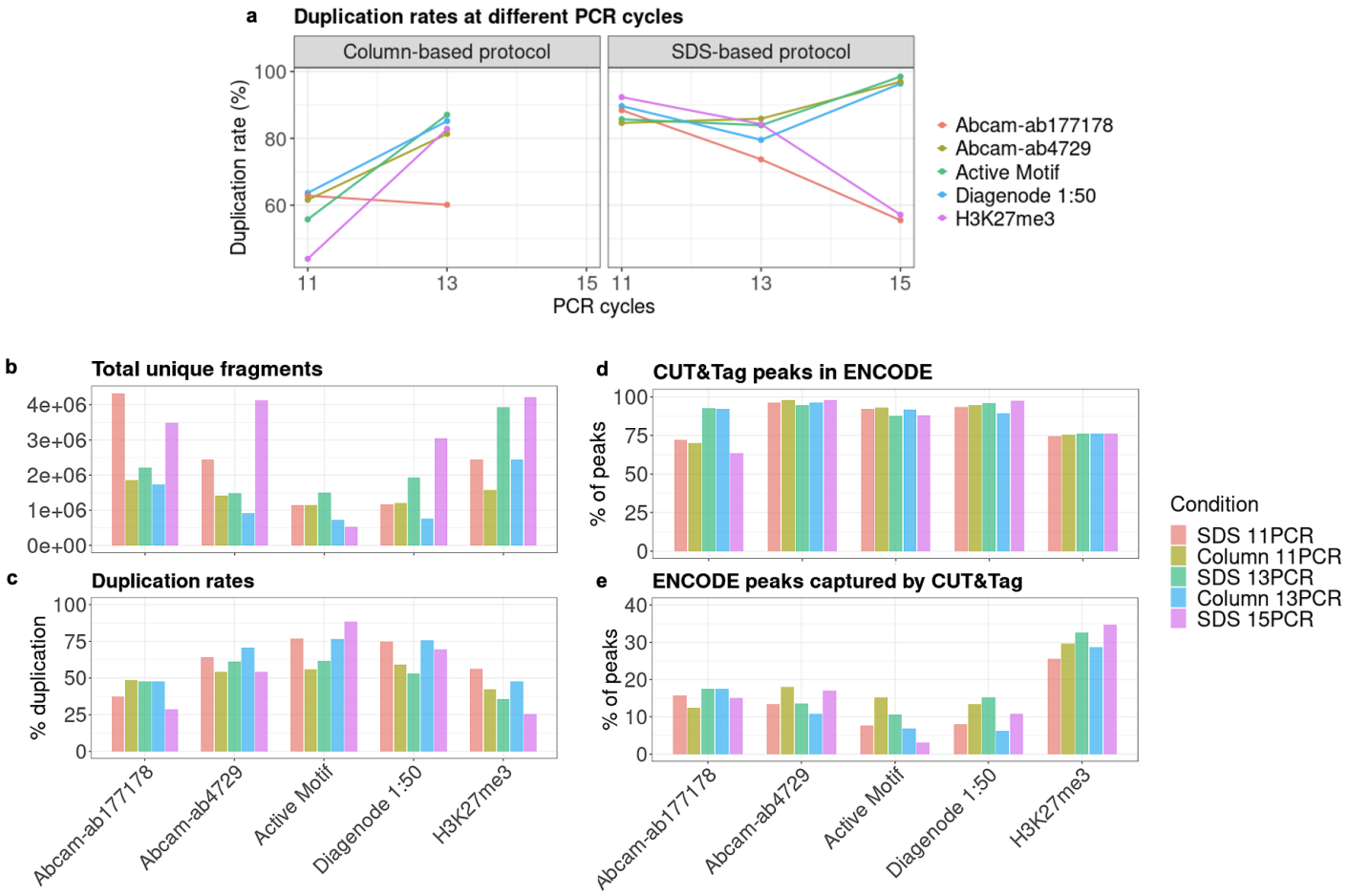


Figure 6. CUT&Tag performance using different numbers of PCR cycles during library amplification. a, Duplication rates obtained with different PCR cycles comparing SDS-based and column-based extraction, without downsampling; **b**, Total unique fragments obtained using the two extraction methods and 11, 13 or 15 PCR cycles, without downsampling; **c**, Duplication rates of reads across antibodies, PCR cycles and DNA extraction methods; **d**, Proportion of CUT&Tag peaks falling into ENCODE H3K27ac; **e**, Proportion of ENCODE H3K27ac peaks captured by CUT&Tag. Samples in **c-e** were subsampled to the same read depth.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Functional analysis of CUT&Tag peaks

To investigate functional similarities of peaks identified by CUT&Tag compared to ENCODE ChIP-seq, we assessed the genomic distribution of CUT&Tag peaks in relation to genes and chromatin states. Using ChIPseeker, peaks were annotated with and mapped in relation to their proximal genes. This revealed a strong skew towards promoter proximal regions for H3K27ac CUT&Tag samples, and a corresponding depletion in promoter regions for H3K27me3 (**Fig. 7a**). H3K27ac CUT&Tag exhibited a stronger promoter preference than H3K27ac ENCODE ChIP-seq, particularly for peaks called by SEACR (**Supplementary Fig. 5a**). H3K27ac CUT&Tag also showed enrichments for distal intergenic regions, which likely harbour a significant fraction of enhancers[31]. Next we explored an alternative gene-independent breakdown of functional genomic elements by assigning peaks to ChromHMM-derived chromatin states[32] using the genomation R package[33]. This confirmed a predominance of promoters and enhancers amongst the regions mapped by H3K27ac CUT&Tag (**Fig. 7b; Supplementary Fig. 5b**). In contrast, H3K27me3 overwhelmingly localised to heterochromatic and repressed chromatin regions, as would be expected. The two peak callers showed slight differences in regulatory element enrichment. While MACS2 corresponded better to ENCODE H3K27ac, SEACR peaks were more enriched at weak enhancers, weak promoters, weakly transcribed regions, and regions of transcriptional transition from initiation to elongation than their MACS2-called counterparts (**Fig. 7b**). This could be in part attributed to the fact that SEACR peaks are broader than MACS2 peaks and thus more likely to extend to neighbouring elements.

Peaks specific to duplicate-containing samples were functionally annotated to reveal that a significant portion of excess MACS2 peaks called upon inclusion of duplicates fall into heterochromatic regions, even among H3K27ac CUT&Tag samples (**Supplementary Fig. 5c**). This suggests that duplicates should not be retained when calling peaks with MACS2, as it can lead to artifacts. On the other hand, the few extra peaks called by including duplicates in SEACR match the regulatory element distribution of the corresponding deduplicated peaks. Finally, CUT&Tag peaks that did not overlap with ENCODE spanned diverse element types, with tested antibodies showing an enrichment in areas of weak transcription, weak enhancers and heterochromatin, while published CUT&Tag and CUT&RUN data showed an enrichment for transcription elongation, weak transcription, weak enhancer and transcription transition categories (**Supplementary Fig. 5d**). For H3K27me3, these peaks were still almost exclusively located in heterochromatic regions. Read distributions around transcription start sites (TSS) obtained from NCBI RefSeq[34] were visualised with heatmaps, which showed enrichment around TSS for H3K27ac (**Supplementary Fig. 5e**). Although CUT&Tag was found to capture slightly fewer promoters overall compared to ChIP-seq when subsampled to the same read depth (not shown), it showed higher average read densities in these regions for H3K27ac (**Supplementary Fig. 5e**). Reads from H3K27me3 samples did not co-localise with promoters.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we performed gene ontology enrichment analysis visualising the top 10 enriched categories with clusterProfiler[35]. Overall, CUT&Tag recovered almost all top ENCODE K562 H3K27ac ChIP-seq ontology terms, including cell adhesion molecule binding, RNA-acting catalytic activity, and cadherin binding, with the exception of the most poorly performing antibodies (**Fig. 7c**). The correspondence of enriched terms indicates that although CUT&Tag may not recover all K562 ChIP-seq peaks, it performs sufficiently well to approximate the K562 regulatory landscape. We further conducted motif analysis using HOMER[36]. Plotting the union of the top 15 enriched transcription factors in each sample revealed that most CUT&Tag samples detected key ENCODE H3K27ac TFBM, including those for the TFs *Bach1*, *Bach2*, *Elk1*, *Elk4*, *Gata1*, *Gata2*, *Gata4*, *Gata6*, *Jun-AP1*, *Nrf2* and *NF-E2* (**Fig. 7d**), which relate to cell growth[37] and hematological cell fate[38,39]. H3K27me3 samples showed more variable and modest TFBM enrichment, which is expected given that the vast majority of transcription factors bind in regions of open chromatin.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

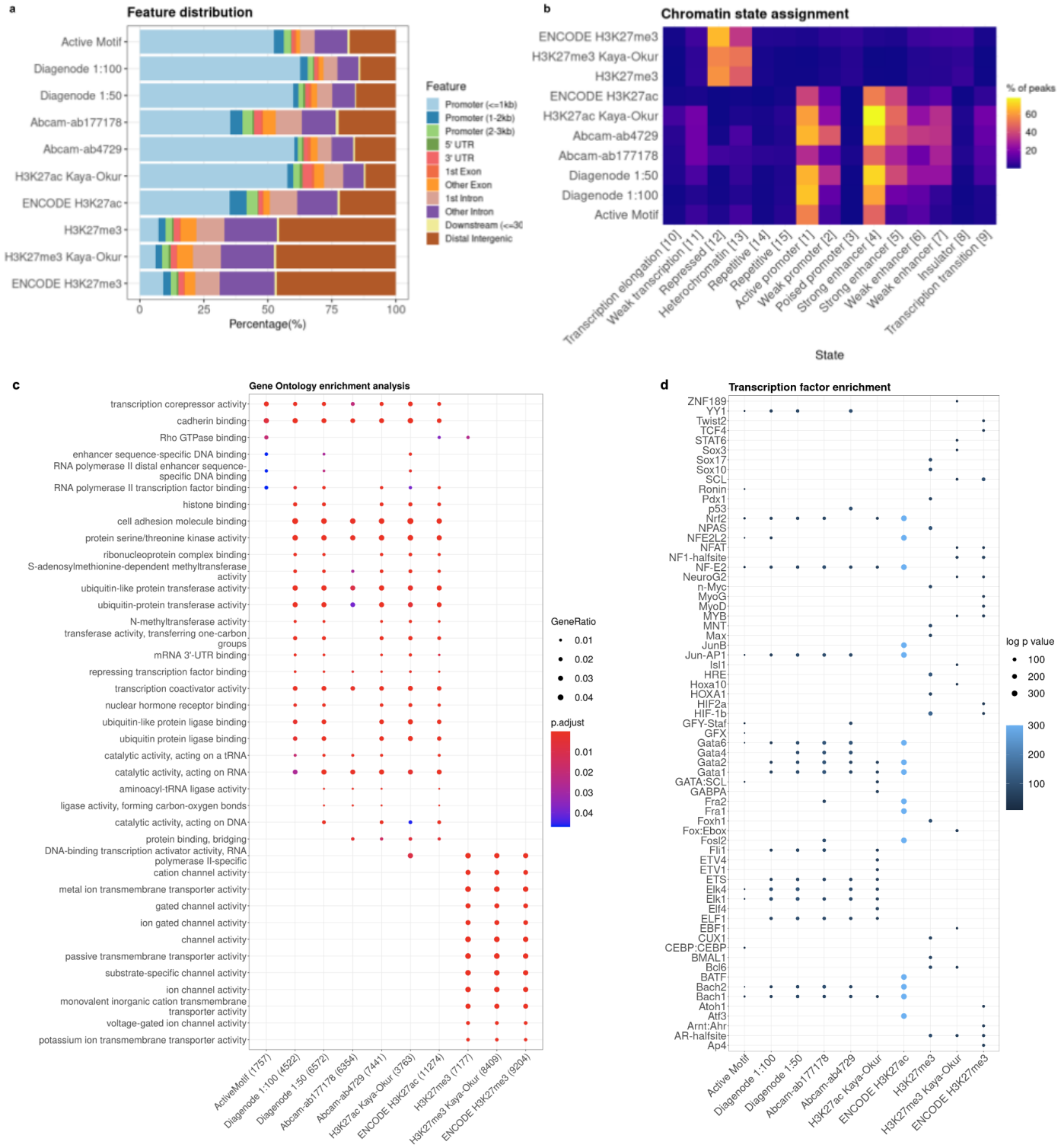


Figure 7. Functional analysis of CUT&Tag peaks. a, ChIPseeker assignment of peaks with regulatory elements; **b**, ChromHMM assignment of peaks with chromatin states, showing the

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

relative percentages of total peaks falling into each category (note that peaks can fall into multiple categories simultaneously); **c**, Gene ontology enrichment analysis results of genes assigned to sample peaks ('gene ratio' reflects the proportion of total differentially expressed genes falling into a particular GO term); **d**, Top significantly enriched motifs across all samples.

Discussion

Here we optimised conditions for the execution and analysis of CUT&Tag for H3K27 histone marks, benchmarking its performance against matched ENCODE ChIP-seq reference datasets. We studied H3K27ac in depth, due to its functional co-localisation with active promoters and enhancers, relevance for mapping risk variants in complex human disease, and the lack of previous literature optimising CUT&Tag for acetylation marks. We systematically assessed experimental optimisations including antibody selection, antibody concentration, DNA extraction method, use of enzymatic inhibitors of deacetylases and PCR cycles. We assessed analysis methods exploring the performance of different peak callers SEACR and MACS2, peak calling parameters, and inclusion versus exclusion of duplicates, as a consensus regarding these has been lacking in the field.

On the whole, H3K27ac CUT&Tag successfully recovers many features of ENCODE ChIP-seq and captures the most significant ENCODE peaks. However, overall CUT&Tag only recovers up to half of ENCODE peaks. Additionally, CUT&Tag appears to generate distinct peak profiles that favor H3K27ac domains coinciding with open chromatin regions and does not capture less significant ENCODE peaks, which are less enriched in open chromatin. It is uncertain whether this is a result of CUT&Tag failing to capture some finer but nevertheless relevant ChIP-seq peaks, or an indicator that ENCODE ChIP-seq may detect less relevant H3K27ac domains that have lower incidence of open chromatin, an important feature of active regulatory elements. Thus, although ENCODE ChIP-seq is often used as a dataset to benchmark against, it is unclear exactly how well ENCODE data reflects the ground truth. Investigation into new methods of chromatin profiling would significantly benefit from inclusion of orthogonal approaches to mapping chromatin modifications and regulatory elements. It would also be helpful to determine whether H3K27ac ChIP-seq peaks that are not captured can be functionally validated. Benchmarking was attempted with massively parallel reporter assay (MPRA) data[40], but the most significant regulatory assay quantitative trait loci (raQTLs) did not appear to be enriched in enhancer and promoter elements, and their coordinates could not be used as proxies for their genomic locations. Thus currently available MPRA data cannot serve this benchmarking purpose.

The capabilities of CUT&Tag likely also vary depending on histone mark, given the better performance of H3K27me3 (**Fig. 2d**). Both current and comparison data suggest that (despite only achieving a maximum of ~50% ENCODE coverage across all samples) this

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

method performs better for methyl marks than for H3K27ac, and in the literature, the superiority of CUT&Tag over ChIP-seq was demonstrated on methyl marks[14]. We noted that peak calling parameters can greatly influence quality measures, such as FRiPs, and should be adjusted depending on histone modification. Nevertheless, it is uncertain why H3K27ac did not yield itself as well to CUT&Tag, and why it may be a more challenging mark to profile even when using an HDACi. It would be interesting to see whether the same issues are encountered when profiling other acetyl marks, which has not yet been systematically addressed.

Presence of high duplication rates could be a result of overamplification during library preparation or over-sequencing. In either case duplicates can be removed without compromising data quality. An advantage of high sequencing depths is sample saturation, meaning that the vast majority of unique fragments present in each sample was recovered; however, one intended advantage of CUT&Tag relative to methods such as ChIP-seq is the ability to recover comparable or superior levels of information at lower sequencing depths. Fractions of reads in ENCODE H3K27ac peaks were approximately equal with and without duplicates, suggesting that they are evenly distributed. Consequently, duplicates made little to no difference when calling peaks with SEACR, since genuine H3K27ac reads contributed to peaks that would in any case be called without duplicates, and reads outside genuine peaks did not meet the peak calling threshold. However, duplicates resulted in the detection of a significant number of spurious peaks with MACS2, many of which fell into heterochromatic regions, which should not be marked by H3K27ac[5]. Using fewer PCR cycles during library preparation appears to modestly reduce duplication rates without significantly influencing ENCODE capture.

Peak calling settings can have a significant effect on the perceived outcomes of chromatin profiling experiments. In this study, multiple peak calling parameters were tested and selected on the basis of precision and recall against matched ENCODE ChIP-seq profiles. MACS2 and SEACR performed similarly in spite of the marked differences in peak definition between the two peak callers. However, achieving optimal performance with MACS2 required significantly more optimisation. Which approach is correct is debatable because there is no strict definition as to what qualifies as a 'peak', but one concern is that peak calling with SEACR might make it difficult to detect subtle changes in histone marks, due to its tendency to call wider peaks and combine multiple potentially distinct H3K27ac domains into single peaks. With regards to differences in precision, MACS2 was optimised specifically for ChIP-seq. ChIP-seq samples are normally sequenced to much higher read depths and tend to possess higher levels of background, which is why MACS2 is designed to identify signal in data with high levels of noise[27,41]. In principle, CUT&Tag and CUT&RUN have reduced background as the only DNA fragments that are released are those bound by the protein of interest[14,42]. For a peak caller like MACS2, any off-target reads in samples with low background might be perceived as legitimate peaks, and this may explain why the inclusion of duplicates gave rise to spurious peaks. Notably, our analyses did not confirm higher signal-to-noise ratio for H3K27ac

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CUT&Tag compared to ENCODE ChIP-seq profiles. Rather we found H3K27ac CUT&Tag to display higher noise, in contrast to what is seen for H3K27me3 and other methyl marks in our analyses and previously reported by others. Whether this represents a general challenge for CUT&Tag of histone acetylation marks remains to be explored further. In the future, more specific peak calling methods designed for CUT&Tag data, such as the recently reported method GoPeaks[43], are likely to improve performance of CUT&Tag profiling and should be included in comparisons for benchmarking analysis.

The lack of established metrics to standardise performance makes it challenging to compare peak callers. Precision cutoffs are arbitrary and there is opportunity to significantly increase recall at the expense of precision even within a predetermined boundary. Going forward, it might be worth characterising the CUT&Tag peaks that could be obtained without strict limits on precision to determine whether they could be legitimate peaks that are not captured by H3K27ac ChIP-seq. For example, it has been suggested that the relatively low correlation between CUT&RUN and ChIP-seq may be due to CUT&RUN's superior ability to map repetitive, difficult regions that are typically not covered by ChIP-seq[44]. There is some indication of this as CUT&Tag and CUT&RUN samples processed with SEACR were far more enriched in weak enhancers and weakly transcribed regions than ENCODE ChIP-seq, but MACS2 ChromHMM profiles differed minimally from ENCODE and this analysis indicated that this effect is more likely a result of peak caller selection rather than being intrinsic to CUT&Tag. One limitation associated with the use of ChromHMM annotations is that chromatin states are inferred on the basis of broad ENCODE ChIP-seq peaks, which introduces some circularity into overlap analysis with ENCODE H3K27ac peaks. Thus, states which occur in CUT&Tag but not in ENCODE H3K27ac are those that should in principle not contain the H3K27ac mark. However, these annotations draw upon combinations of histone marks[45], and still give some indication as to where a particular modification might or might not be expected to occur.

The improved sensitivity of CUT&Tag compared to ChIP-seq is due to the use of pA-Tn5 to streamline library preparation through direct insertion of PCR sequencing adapters via *in situ* tagmentation. However, its sensitivity is inherently limited by PCR, since pA-Tn5 inserts its adapters in random orientations such that approximately half of the targets do not have adapters in the correct orientation to amplify. In addition, PCR library preparation is highly sensitive to size variations of amplicons. When two adjacent transposition events occur too far apart, they will not amplify efficiently during PCR or sequencing cluster generation. However, when they are too close, they will exponentially bias library coverage due to increased PCR amplification and clustering efficiency of shorter fragments. One recent approach that may help overcome some of these issues is linear amplification by Targeted Insertion of Promoters (TIP-seq[15,46]). In TIP-seq, a pA-Tn5 fusion protein is used to insert a T7 RNA polymerase promoter near sites occupied by transcription factor or histone mark of interest. The promoter facilitates linear amplification of DNA in its vicinity using a T7 polymerase to create 1,000-fold RNA copies of insertion sites. The distance between two transposition sites does not bias

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

library preparation since only one T7 promoter is needed to amplify the site of interest. Linear amplification generates greater fidelity and uniformity, as mistakes made during amplification do not themselves become templates to exponentially propagate errors – this results in higher mappability of single cell sequencing reads[47]. TIP-seq was shown to generate single cell libraries with higher read coverage, greater library complexity, and contain lower background with a higher proportion of unique, non-duplicated reads per cell compared to CUT&Tag[46]. Comprehensive optimisation and benchmarking of this novel technique will be essential moving forward.

Conclusions

CUT&Tag has been promoted as a more streamlined, cost-effective approach to chromatin profiling, but despite a definite correspondence with ENCODE ChIP-seq, CUT&Tag consistently achieves an ENCODE coverage ceiling of approximately 50%. Furthermore, the performance of this method appears to vary by histone mark. Additional analysis will be required to better characterise the inconsistencies between CUT&Tag and ENCODE ChIP-seq. Optimising experimental parameters, our analysis established Abcam-ab4729 as the top-performing antibody and demonstrated that the use of an HDACi does not improve H3K27ac CUT&Tag performance. Duplicates can and should be discarded, particularly beyond a threshold at which they start to contribute more off-target than on-target information. The optimal choice of peak caller is debatable, but SEACR attains the best possible performance with very little optimisation and is more impervious to identification of spurious peaks. There is also some evidence that it calls peaks with greater consistency than MACS2, even on the basis of poorer-quality reads and lower sequencing depths. We observe that fewer PCR cycles reduced duplication rates, but at the expense of ENCODE recovery and capture. We hope that our systematic optimisations of CUT&Tag will help to facilitate its more widespread adaptation in the field and expedite its application in understanding the epigenetic causes and consequences of complex diseases.

Methods

Biological materials

Human K562 cells were obtained from ATCC (Manassas, VA, Catalog #CCL-243) and cultured according to the supplier's protocol. Mycoplasma was tested to be negative for all cellular input reported using Mycoplasma Detection Kit (Jena Bioscience PP-401) following manufacturer's instructions. The following antibodies were used: Guinea Pig anti-Rabbit IgG (Heavy & Light Chain) Preabsorbed antibody (Antibodies-Online ABIN101961), H3K27me3 (Cell Signaling Technology, 9733, Lot 14), H3K27ac (Abcam ab177178, Lot GR3202987-5), H3K27ac (Active

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Motif 39133, Lot 16119013), H3K27ac (Abcam ab4729, Lot G3374555-1), H3K72ac (Diagenode C15410196, Lot A1723-0041D). The following histone deacetylase inhibitors were used: Sodium butyrate (Merck B5887-250MG; used at 5 mM in CUT&Tag solutions with HDACi treatment), Trichostatin A (Enzo Life Sciences BML-GR309-0001; used at 1 μM in CUT&Tag solutions with HDACi treatment). The following commercial loaded protein A-Tn5 transposase fusion protein (pA-Tn5) were used at recommended dilutions by the manufacturer: CUTANA™ pAG-Tn5 (Epicypher 15-1017; Lot 20142001-C1), or pA-Tn5 Transposase - loaded (Diagenode C01070001; Lot 1/b/b).

CUT&Tag nuclei processing

Bench top CUT&Tag was performed as previously described (https://www.protocols.io/view/bench-top-cut-amp-tag-bcuhiwt6)[14]. Exponentially growing K562 cells were harvested, counted and centrifuged for 3 min at 600*g* at room temperature (RT). 500,000 cells per condition were washed twice in 1 mL Wash Buffer (20 mM HEPES-KOH pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1x Protease inhibitor cocktail; Roche 11836170001). Nuclei were extracted by incubating cells for 10 minutes on ice in 200 μL/sample of cold Nuclei Extraction buffer (NE buffer: 20 mM HEPES-KOH pH 7.9, 10 mM KCl, 0.1% Triton X-100, 20% Glycerol, 0.5 mM Spermidine, 1x Protease Inhibitor cocktail). Following incubation in NE buffer, nuclei were centrifuged for 3 min at 600g at RT, then resuspended in 100 μL cold NE buffer. Concanavalin A coated magnetic beads (Bangs Laboratories BP531) were prepared as previously described [42]and 11 μL of activated beads were added per sample into PCR strip tubes and incubated at RT with gentle rocking for 10 min. Beads were placed on a magnetic rack and unbound supernatant was discarded. Bead bound nuclei were resuspended in 50 μL Dig-wash Buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.05% Digitonin) with 2 mM EDTA and 0.1% BSA. 1:50/1:100/1:200 dilution of primary antibody was added, followed by a gentle vortex and brief spin. Primary antibody incubation was conducted on a rotating platform overnight at 4°C. Primary antibody solution was removed by placing the PCR tube on a magnetic rack, allowing the solution to fully clear, then removing the supernatant. Next, the appropriate secondary antibody, Guinea Pig anti-Rabbit IgG antibody for a rabbit primary antibody, was added at 1:100 in Dig-Wash buffer and incubated at RT with rotation for 30-60 min. Nuclei were washed twice in 200 μL Dig-Wash buffer using a magnetic rack to remove unbound antibodies in supernatant. Nuclei were resuspended in 50 μL Dig-med Buffer (20 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1× Protease inhibitor cocktail, 0.05% Digitonin), then 1:20 CUTANA™ pAG-Tn5 (Epicypher 15-1017), or 1:250 pA-Tn5 Transposase - loaded (Diagenode C01070001) was added, gently mixed and spun down. pA-Tn5 binding occurred at RT for 1 hour on a rotating platform. To remove unbound pA-Tn5, nuclei were washed twice in 200 μL Dig-med Buffer. Nuclei were then resuspended in 50 μL Tagmentation buffer (10 mM $MgCl_2$ in Dig-med Buffer) and incubated at 37°C for 1 hour to activate transposase enzymatic activity.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, either column or sodium dodecyl sulfate (SDS) based DNA extraction was conducted.

Column DNA extraction

To stop tagmentation and solubilise DNA fragments, the following were added to each 50 µL sample: 1.68 µL 0.5M ethylenediaminetetraacetic acid (EDTA), 0.5 µL 10% SDS, 0.44 µL 10 mg/mL Proteinase K. The samples were briefly mixed and vortexed at full speed for ~2 seconds, then incubated at 55°C for 1 hour to digest the DNA. After a quick spin, tubes were placed on a magnetic rack and solution was allowed to clear. Supernatant was carefully transferred to a new 1.5 mL microcentrifuge tube, then sample processing protocol of ChIP DNA Clean & Concentrator (Zymo Research D5205) was executed, eluting with 21 µL Elution Buffer.

SDS-based DNA extraction

Following tagmentation at 37°C for 1 hour, PCR tubes were placed on a magnetic rack and solution was allowed to clear. Supernatant was removed carefully, then beads were resuspended thoroughly in 50 µL [tris(hydroxymethyl)methylamino]propanesulfonic acid (TAPS) Buffer (10 mM TAPS pH 8.5, 0.2 mM EDTA) at RT. Tubes were returned to a magnetic rack and supernatant was removed. 5 µL SDS Release Buffer (10 mM TAPS pH 8.5, 0.1% SDS) was added at RT to each sample and tubes were vortexed at full speed for ~10 seconds. After a quick spin, ensuring no beads are stuck to the side of the tubes, samples were incubated at 58°C for 1 hour. Next, 15 µL SDS Quench Buffer (0.67% Triton-X 100 in Molecular grade H_2O) was added at RT and vortexed at maximum speed to neutralise the SDS prior to PCR library amplification.

CUT&Tag PCR-based library amplification

For library amplification in PCR tube format, 21 µL DNA was combined with 2 µL of universal i5 and uniquely barcoded i7 primer[48] where a different barcode was used for each sample that was intended to be pooled together. 25 µL NEBNext HiFi 2× PCR Master mix was added, then the sample was gently mixed through and spun down. The sample was placed in a Thermocycler with heated lid following these conditions: 72°C for 5 min (gap filling); 98°C for 30 s; 11-15 cycles of 98°C for 10 s and 63°C for 30 s; final extension at 72°C for 1 min; and hold at 4°C. Following PCR, bead cleanup was conducted by addition of 1.1x Ampure XP beads (Beckman Coulter). Library and beads were mixed thoroughly, then spun down and incubated at RT for 10-15 min. Beads were gently washed twice with freshly prepared 80% ethanol using a magnetic rack, then the library was eluted with 20-30 µL 10 mM Tris-HCl pH 8.0 at RT.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequencing

Final library size distributions were assessed by Agilent 2100 Bioanalyser and Agilent 4200 TapeStation for quality control before sequencing. Libraries were pooled to achieve equal representation of the desired final library size range (equimolar pooling based on Bioanalyser/TapeSation signal in the 150bp to 800bp range). Paired-end Illumina sequencing using the HiSeq 4000 PE75 strategy was conducted on barcoded libraries at Imperial Biomedical Research Centre (BRC) Genomics Facility following manufacturer's protocols.

qPCR

Quantitative real-time PCR (qPCR) was performed following manufacturer's instructions (https://www.thermofisher.com/order/catalog/product/4309155#/4309155). Positive and negative control primers were designed based on ENCODE peaks ranked highest to lowest, respectively.

Gene	Control type	Primer sequence (5'→3')
ARHGAP22	Positive	Fw: GCTGAGAAGGAAGGGCTTAAT Rv: GCTAGTCGGGATGATTTACAGG
COX4I2	Positive	Fw: GGATACCTCCAAGGCTTCATAC Rv: GTAGTCACAGAACTAGGGTTGG
MTHFR	Positive	Fw: GGGTGGAACATCTCGAACTATC Rv: GAACGAAGCCAGAGGAAACA
ZMYND8	Positive	Fw: GGATCTACAAACTTCCCTTCCC Rv: GAAGGCATCGCAGGCTAATA
KLHL11	Negative	Fw: GACAAGCAGTGGCTCTACAA Rv: CAGTATCGGAAAGAAGCCTACC
SIGIRR	Negative	Fw: CCAAGCTCAGACCTCAAAGT Rv: TTCTTGCTGTGCTCGTATCC

Table 1. qPCR primer sequences. Control primer sequences based on ENCODE peaks.

Levels of H3K27ac CUT&Tag binding signal was determined by qPCR amplification carried out with the QuantStudio™ 5 Real-Time PCR System (ThermoFisher A34322) using the Standard Curve experiment type and SYBR Green Master Mix (ThermoFisher 4309155). Each qPCR condition was conducted with triplicate repeats and the data was analysed using the 2^-Δ*ΔCT* method where each CUT&Tag sample was normalised to qPCR levels of K562 genomic DNA (gDNA) run in parallel. qPCR results were calculated using the equation:

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$2^{-(CT\ sample\ -\ CT\ gDNA)}$$

Data processing

Sequencing data was processed according to the CUT&Tag Data Processing and Analysis Tutorial (https://yezhengstat.github.io/CUTTag_tutorial), with some alterations. Raw sequencing reads were trimmed using TrimGalore (version 0.6.6; https://github.com/FelixKrueger/TrimGalore) to remove adapters and low-quality reads. The trimmed fastq files were aligned to hg19 using bowtie (version 2.2.9; [49]) with the following parameters: *--local --very-sensitive --no-mixed --no-discordant --phred33 -I 10 -X 700*. PCR duplicates were removed using Picard (version 2.6.0; http://broadinstitute.github.io/picard/), and bam and fragment bed files from original and deduplicated alignments were generated using samtools (version 1.3.1;[50]) and bedtools (version 2.25.0;[51]), selecting for fragment lengths under 1000 bp. Peaks were called using MACS2 (Model-based Analysis of ChIP-seq; version 2.1.4) [41] and SEACR (Sparse Enrichment Analysis for CUT&RUN; version 1.3)[24]. MACS2 peaks were called as follows: *macs2 callpeak -t input_bam -n sample_name -f BAMPE -g hs -q 1e-5 --keep-dup all –nolambda --nomodel --outdir out_dir*. SEACR peaks were called on the basis of fragment bedgraph files generated with bedtools genomecov. SEACR peaks were called as follows: SEACR_1.3.sh input_bedgraph 0.01 non stringent out_name. In both bases other combinations of peak calling settings were also tested (see **Results**). All peaks overlapping with hg19 blacklisted regions (ENCODE file ID: ENCFF000KJP) were removed prior to downstream analysis. Motifs were identified using HOMER (Hypergeometric Optimization of Motif EnRichment; version 4.10;[36]) as follows: *findMotifsGenome.pl input_bed hg19 out_dir -size 1000*. Downsampled bam files were generated by random sampling of original bam files as follows, where {x} represents the seed value and {y} the fraction of total read pairs to be sampled: *samtools view -bs {x}.{y} input_bam > downsampled_bam*.

Sample comparisons

Published CUT&Tag[14] and CUT&RUN[18] samples were obtained as fastq files from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home; study accessions PRJNA512492 and PRJNA522731, respectively) and processed as described above. Peak-level correlations were obtained with the DiffBind package (version 3.0.15;[28]). Genome-wide sample correlations were carried out using bedtools multicov against hg19 split into 500bp bins. Read counts were then quantile-normalised and rounded to the nearest integer, and heatmaps plotted in R[52] based on sample-by-sample Pearson correlations of the processed counts. Fingerprint plots were generated from sample and ENCODE bam files using deepTools (version 3.5.1;[53]) plotFingerprint, setting genome-wide bin sizes of 1000bp. Heatmaps were plotted using deeptools computeMatrix and plotHeatmap to visualise read enrichment around hg19 transcription start sites (obtained from NCBI RefSeq) and peak

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

summits. For these ends, ENCODE H3K27ac and H3K27me3 samples (ENCSR000AKP and ENCSR000EWB, respectively) were run through the ENCODE histone ChIP-seq pipeline (https://github.com/ENCODE-DCC/chip-seq-pipeline2), with replicates downsampled to 1 million reads per sample and pooled together. Likewise, paired-end CUT&Tag sample bam files were downsampled to 2 million fragments (4 million reads) and only the first of the read mates mapped to yield a total of 2 million mapped reads. As a weighted average of precision and recall, F1-scores were calculated as follows, where *tp*, *fp*, and *fn* represent the numbers of true positive, false positive, and false negative CUT&Tag peaks, respectively:

$$F1 = \frac{tp}{tp + 1/2(fp + fn)}$$

Downstream data analysis

Downstream analysis, including quality control, ENCODE benchmarking, and regulatory element annotation, were performed in R[52]. Peaks falling into mitochondrial chromosomes were removed using BRGenomics (version 1.1.3; https://mdeber.github.io) prior to downstream analysis. Peak overlaps were determined with the GenomicRanges package (version 1.38.0;[30]). Reads in peaks were calculated using the chromVAR package (version 1.8.0;[54]). Regulatory element annotation was performed using ChIPseeker (version 1.22.1;[55]), after annotating peaks with genes using the TxDb.Hsapiens.UCSC.hg19.knownGene database (version 3.2.2;[56]). ChromHMM annotations assigned with genomation (version 1.18.0; [33]). Functional enrichment analysis was carried out with clusterProfiler (version 3.13.4;[57]), using the "enrichGO" function.

Data and code availability

Fastq and peak files are available through GEO under accession GSE199611. Analysis and code used in this study is available in the dedicated GitHub repository: https://github.com/neurogenomics/CUT_n_TAG

Generalised code for performing comparisons between genome-wide histone modification profiles has been made available as an R package via github at https://github.com/neurogenomics/EpiCompare.

Authors' contributions

All authors conceived the study. DH performed all laboratory experiments. LA performed computational analyses with guidance and support of BMS. SJM and NGS acquired funding and supervised the study. AN guided study directions and interpretations. All authors wrote, read and approved the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Funding

This work was funded by an Alzheimer's Association grant [grant number ADSF-21-829660-C] to SJM and a UKRI Future Leaders Fellowship [grant number MR/T04327X/1] awarded to NGS. SJM is funded by the Edmond and Lily Safra Early Career Fellowship Program. SJM, NGS and AN are funded by the UK Dementia Research Institute, which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

Acknowledgements

We thank the UK DRI Neurogenomics Laboratory (https://www.neurogenomics.co.uk), in particular Alan E Murphy, for feedback, troubleshooting, and guidance in experimental design and analysis. We thank Katiuska Pulgar Prieto for help with K562 mycoplasma testing.

Ethics declarations

Competing interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–10 (2015).
2. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. *Nat. Commun.* **12**, 1610 (2021).
3. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
4. Fox, S. *et al.* Hyperacetylated chromatin domains mark cell type-specific genes and suggest distinct modes of enhancer function. *Nat. Commun.* **11**, 4544 (2020).
5. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
6. Marzi, S. J. *et al.* A histone acetylome-wide association study of Alzheimer's disease identifies disease-associated H3K27ac differences in the entorhinal cortex. *Nat. Neurosci.* **21**, 1618–1627 (2018).
7. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
8. Nativio, R. *et al.* An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nat. Genet.* **52**, 1024–1035 (2020).
9. Rodríguez-Ubreva, J. & Ballestar, E. Chromatin Immunoprecipitation. in *Functional Analysis of DNA and Chromatin* (eds. Stockert, J. C., Espada, J. & Blázquez-Castro, A.) 309–318 (Humana Press, 2014).
10. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18602–18607 (2013).

11. Baranello, L., Kouzine, F., Sanford, S. & Levens, D. ChIP bias as a function of cross-linking time. *Chromosome Res.* **24**, 175–181 (2016).
12. Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K. & Henikoff, S. Efficient low-cost chromatin profiling with CUT&Tag. *Nat. Protoc.* **15**, 3264–3283 (2020).
13. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
14. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
15. Zhu, C. *et al.* Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods* **18**, 283–292 (2021).
16. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).
17. Henikoff, S., Henikoff, J. G., Kaya-Okur, H. S. & Ahmad, K. Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation. *Elife* **9**, (2020).
18. Meers, M. P., Bryson, T. D., Henikoff, J. G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. *Elife* **8**, (2019).
19. Woodcock, C. L., Skoultchi, A. I. & Fan, Y. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Res.* **14**, 17–25 (2006).
20. Luger, K., Dechassa, M. L. & Tremethick, D. J. New insights into nucleosome and

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.* **13**, 436–447 (2012).

21. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).
22. Krehenwinkel, H. *et al.* Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Sci. Rep.* **7**, 17668 (2017).
23. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
24. Meers, M. P., Tenenbaum, D. & Henikoff, S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics Chromatin* **12**, 42 (2019).
25. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
26. Wang, M. & Zhang, Y. Tn5 transposase-based epigenomic profiling methods are prone to open chromatin bias. *bioRxiv* 2021.07.09.451758 (2021) doi:10.1101/2021.07.09.451758.
27. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
28. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
29. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).
30. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
31. Stadhouders, R. *et al.* Transcription regulation by distal enhancers. *Transcription* **3**, 181–186 (2012).
32. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nat. Protoc. **12**, 2478–2492 (2017).

33. Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E. & Schübeler, D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**, 1127–1129 (2015).
34. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).
35. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)* **2**, 100141 (2021).
36. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
37. Tanigawa, S. *et al.* Jun dimerization protein 2 is a critical component of the Nrf2/MafK complex regulating the response to ROS homeostasis. *Cell Death Dis.* **4**, e921 (2013).
38. Itoh-Nakadai, A. *et al.* The transcription repressors Bach2 and Bach1 promote B cell development by repressing the myeloid program. *Nat. Immunol.* **15**, 1171–1180 (2014).
39. Gao, J., Chen, Y.-H. & Peterson, L. C. GATA family transcriptional factors: emerging suspects in hematologic disorders. *Exp. Hematol. Oncol.* **4**, 28 (2015).
40. van Arensbergen, J. *et al.* Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153 (2017).
41. Gaspar, J. M. Improved peak-calling with MACS2. *bioRxiv* 496521 (2018) doi:10.1101/496521.
42. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, (2017).
43. Yashar, W. M. *et al.* GoPeaks: Histone Modification Peak Calling for CUT&Tag. *bioRxiv*

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2022.01.10.475735 (2022) doi:10.1101/2022.01.10.475735.

44. Akdogan-Ozdilek, B., Duval, K. L., Meng, F. W., Murphy, P. J. & Goll, M. G. Identification of chromatin states during zebrafish gastrulation using CUT&RUN and CUT&Tag. *bioRxiv* 2021.06.22.447589 (2021) doi:10.1101/2021.06.22.447589.
45. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
46. Bartlett, D. A. *et al.* High-throughput single-cell epigenomic profiling by targeted insertion of promoters (TIP-seq). *J. Cell Biol.* **220**, (2021).
47. Chen, C. *et al.* Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **356**, 189–194 (2017).
48. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
49. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
50. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
52. Team, R. C. R: A language and environment for statistical computing (R Version 4.0. 3, R Foundation for Statistical Computing, Vienna, Austria, 2020). (2021).
53. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).
54. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods **14**, 975–978 (2017).

55. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

56. Carlson, M. *TxDb.Hsapiens.UCSC.hg19.knownGene*. (Bioconductor, 2015). doi:10.18129/B9.BIOC.TXDB.HSAPIENS.UCSC.HG19.KNOWNGENE.

57. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Materials

Supplementary Tables

Sample	Total Fragments	Mapped Fragments	Alignment Rate	Duplication Rate	Unique Fragments
Active Motif	34,467,445	34,246,137	99.36%	98.45%	531,603
Diagenode 1:100	13,970,902	13,884,243	99.38%	89.15%	1,506,260
Diagenode 1:50	90,812,784	83,771,834	92.25%	96.36%	3,051,389
Abcam-ab177178	7,911,303	7,839,891	99.10%	55.49%	3,489,316
Abcam-ab4729	146,621,482	134,760,046	91.91%	96.94%	4,123,523
H3K27me3	9,912,113	9,823,687	99.11%	57.12%	4,211,994
H3K27ac Kaya-Okur SRR8383507 (C&T)	2,471,858	2,293,087	92.77%	20.18%	1,830,390
H3K27ac Kaya-Okur SRR8383508 (C&T)	3,320,561	3,088,577	93.01%	14.87%	2,629,278
H3K27ac Meers SRR8581604 (C&R)	6,777,196	5,523,031	81.49%	7.30%	5,120,026
H3K27me3 Kaya-Okur SRR11074238 (C&T)	3,945,633	3,806,740	96.48%	1.43%	3,752,159
H3K27me3 Kaya-Okur SRR11074239 (C&T)	4,159,984	4,051,439	97.39%	1.36%	3,996,319
H3K27me3 Meers SRR9073702 (C&R)	9,047,596	8,668,833	95.81%	1.58%	8,531,865

Supplementary Table 1. Sequencing and alignment results of CUT&Tag and CUT&RUN data.

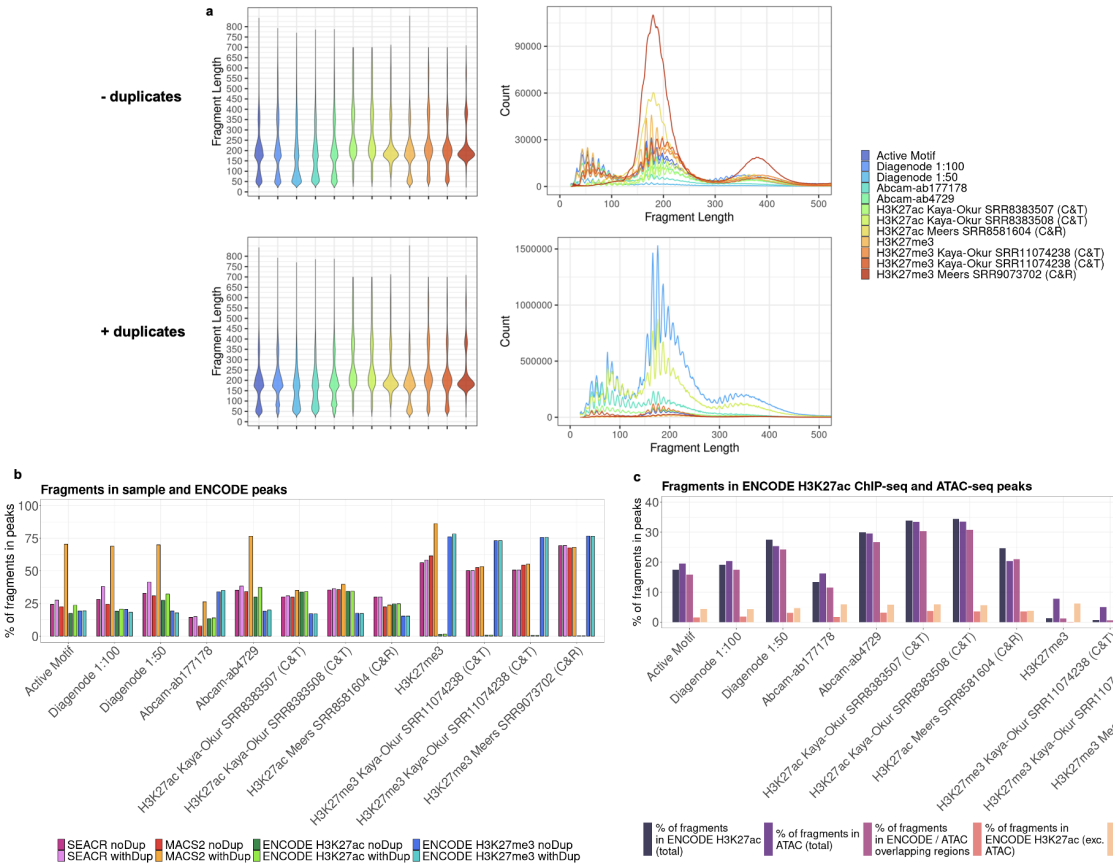
bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sample	Peaks called			
	SEACR		MACS2	
	No duplicates	With duplicates	No duplicates	With duplicates
Active Motif	2,497	3,005	1,762	95,223
Diagenode (1:100)	7,065	7,570	6,717	180,149
Diagenode (1:50)	11,094	11,364	19,672	167,500
Abcam-ab177178	15,458	15,838	9,688	93,453
Abcam-ab4729	12,657	12,924	24,076	275,284
H3K27me3	103,953	104,914	86,965	219,783
H3K27ac Kaya-Okur SRR8383507 (C&T)	6,415	6,451	12,922	18,105
H3K27ac Kaya-Okur SRR8383508 (C&T)	8,044	8,076	18,190	22,789
H3K27ac Meers SRR8581604 (C&R)	16,974	16,983	20,589	23,325
H3K27me3 Kaya-Okur SRR11074238 (C&T)	115,643	115,627	81,134	83,054
H3K27me3 Kaya-Okur SRR11074239 (C&T)	117,329	117,282	87,095	89,009
H3K27me3 Meers SRR9073702 (C&R)	100,777	100,842	81,465	82,758

Supplementary Table 2. Numbers of peaks called with SEACR and MACS2, with and without duplicates included.

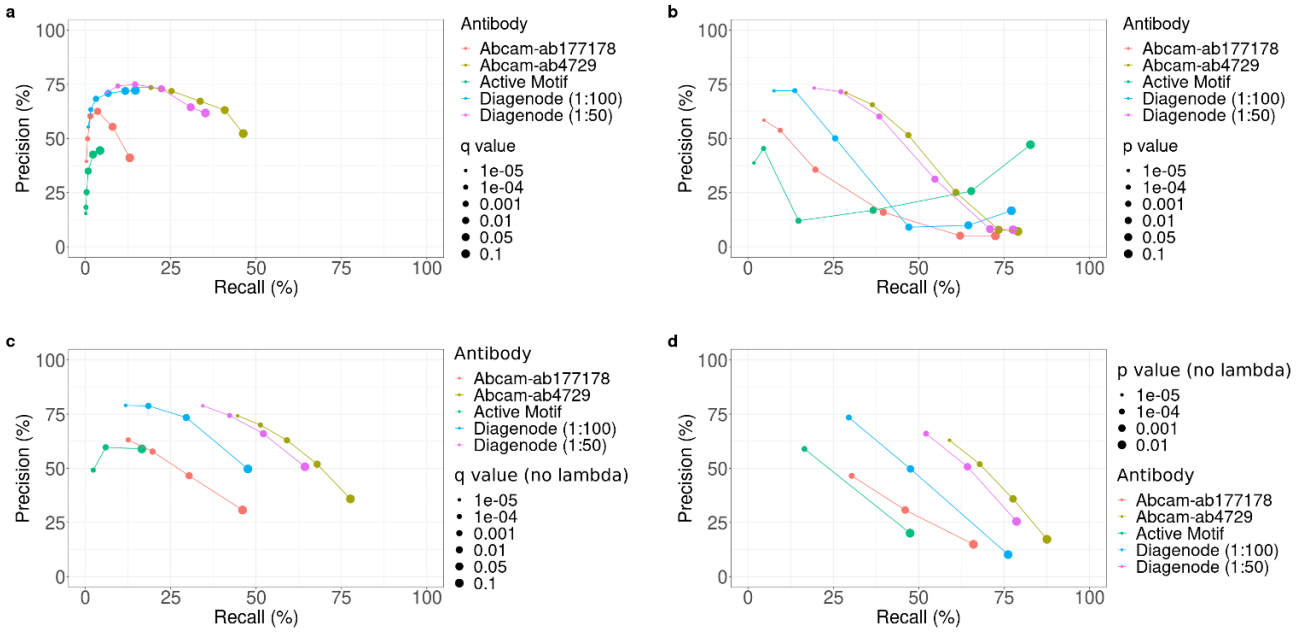
bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figures



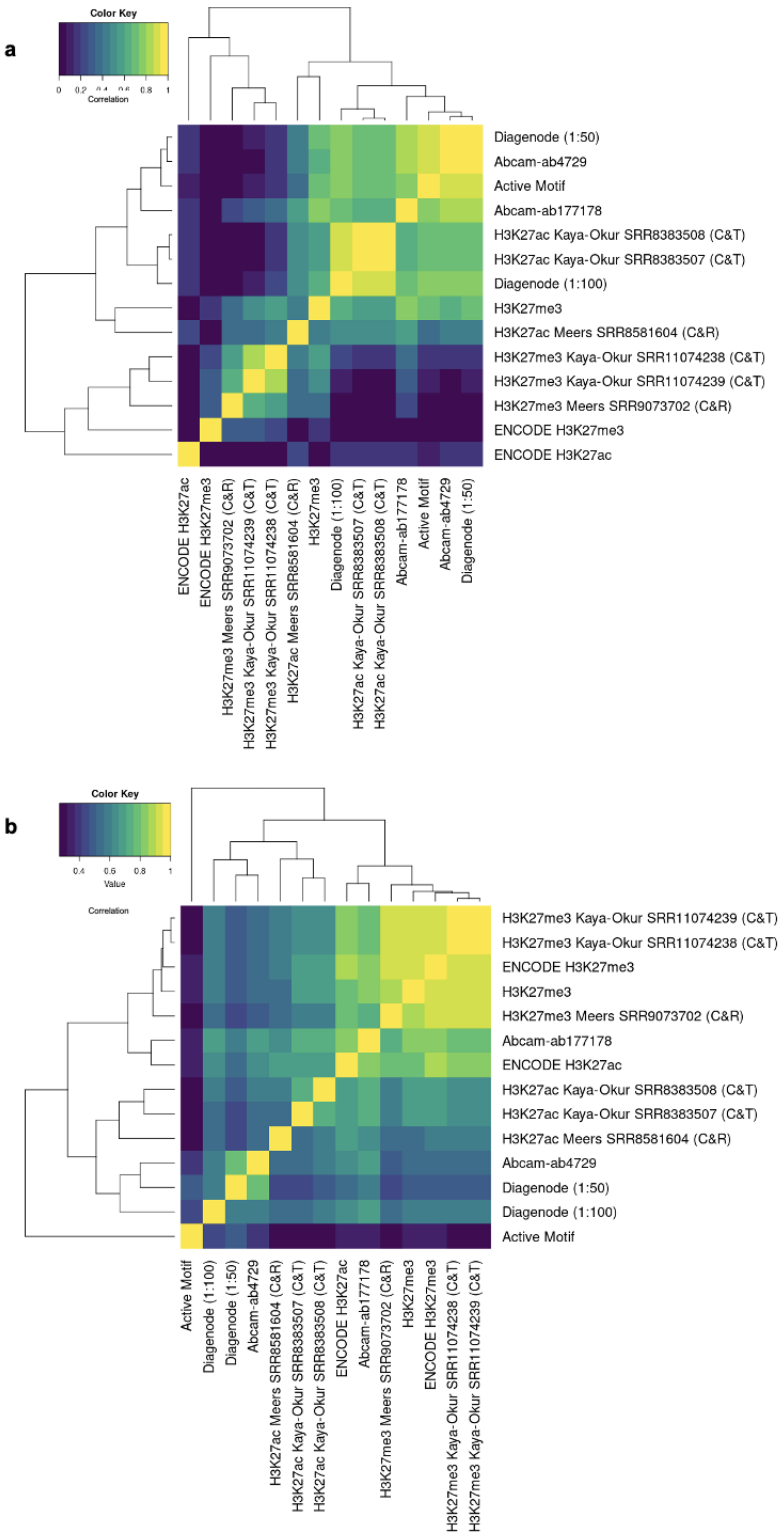
Supplementary Figure 1. Quality control metrics of all tested samples and published data. **a**, Fragment length distributions; **b**, Percentages of reads in peaks called with SEACR or MACS2, with or without duplicates, as well as ENCODE H3K27ac ChIP-seq narrow peaks and ENCODE H3K27me3 ChIP-seq broad peaks; **c**, Percentages of sample reads in ENCODE H3K27ac ChIP-seq and ATAC-seq peaks.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

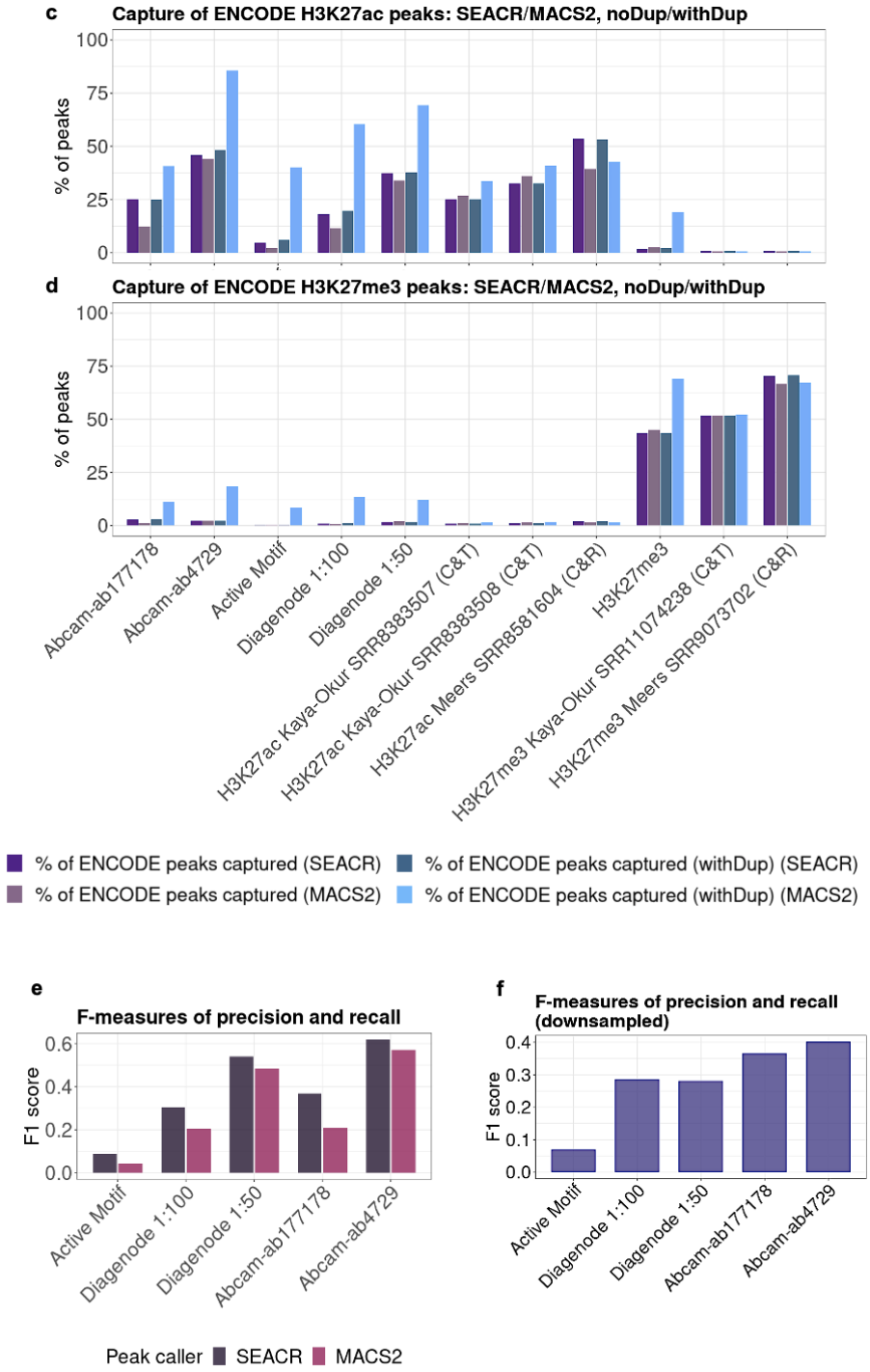


Supplementary Figure 2. Optimisation of peak calling with MACS2, showing precision vs recall when varying **a**, q value; **b**, p value; **c**, q value with local lambda deactivated; **d**, p value with local lambda deactivated.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

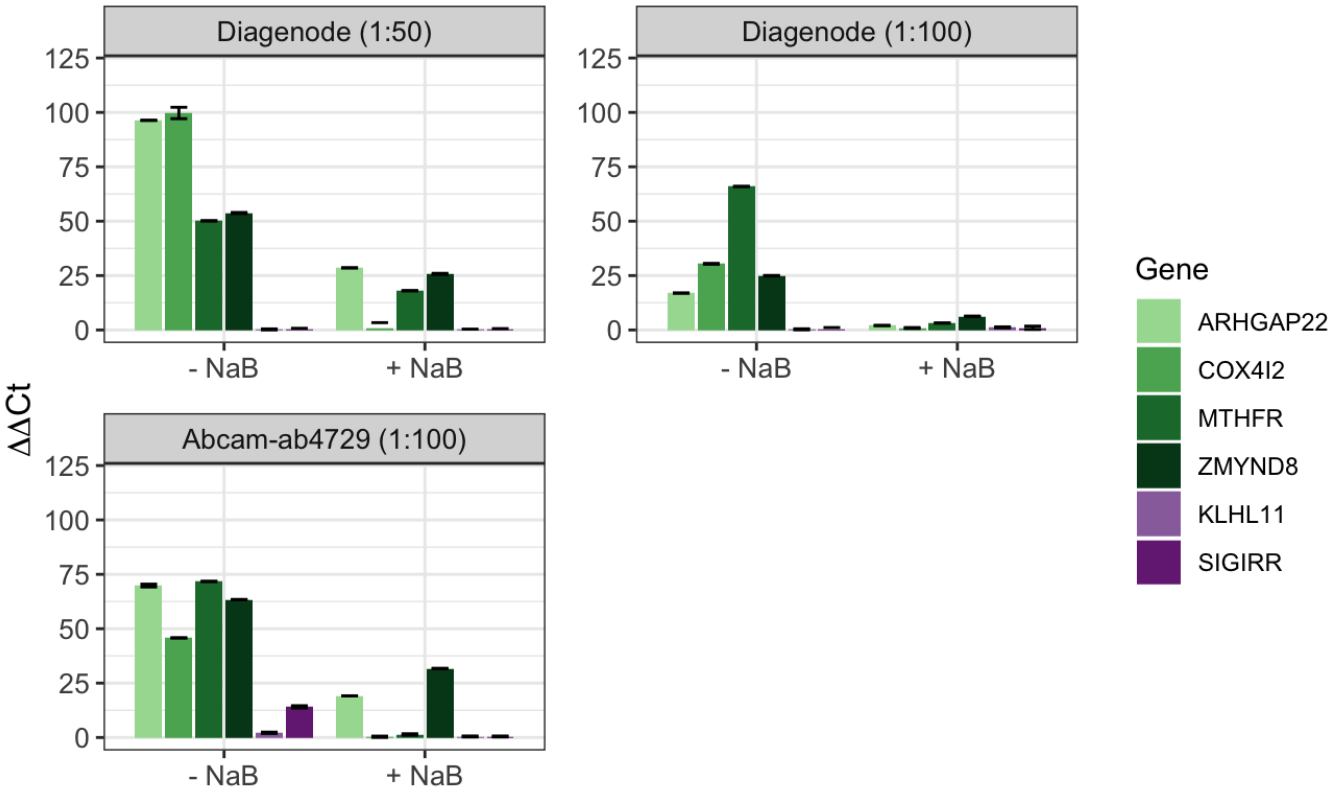


bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



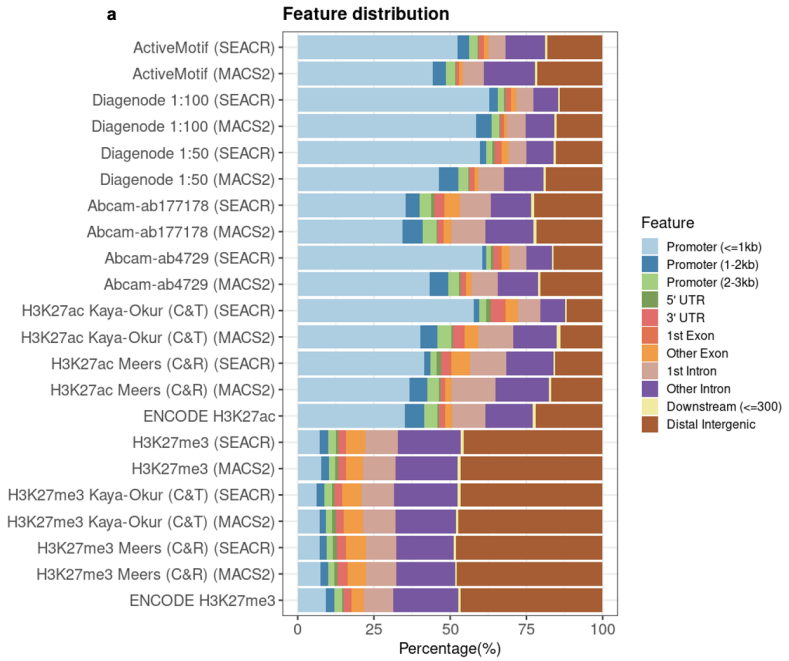
Supplementary Figure 3. Read correlations across **a**, 500bp genome-wide bins and **b**, ENCODE H3K27me3 ChIP-seq peak ranges; capture of ENCODE **c**, H3K27ac and **d**, H3K27me3 ChIP-seq peaks by CUT&Tag and CUT&RUN peaks called with SEACR or MACS2, with or without duplicates; F-measures of precision and recall of tested H3K27ac antibodies at **e**, maximum read depth and **f**, 8 million paired-end reads each.

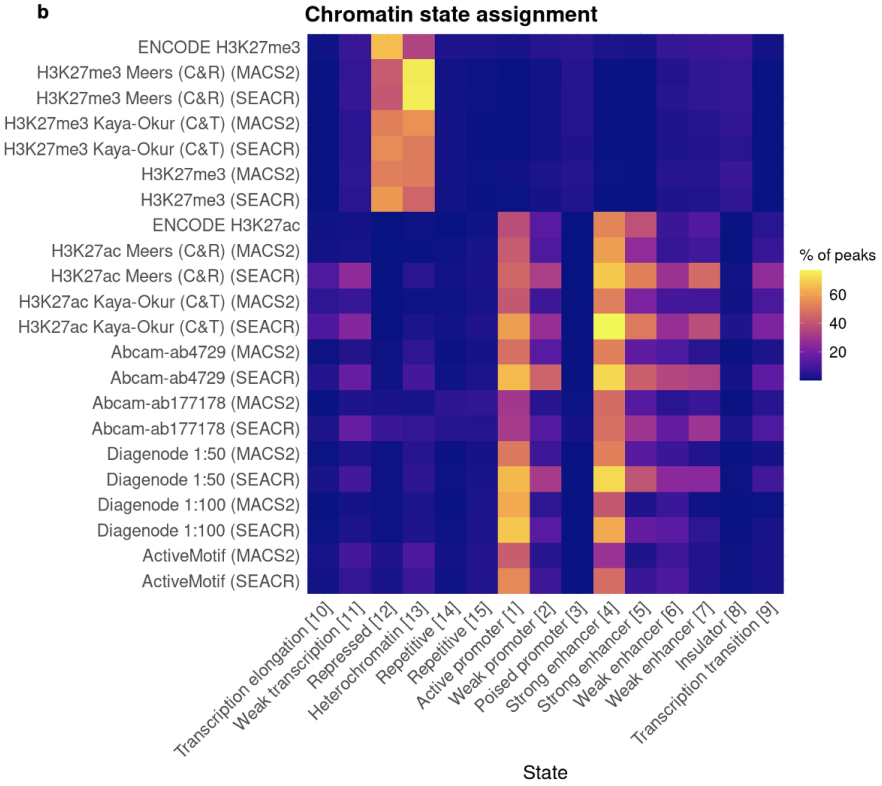
bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



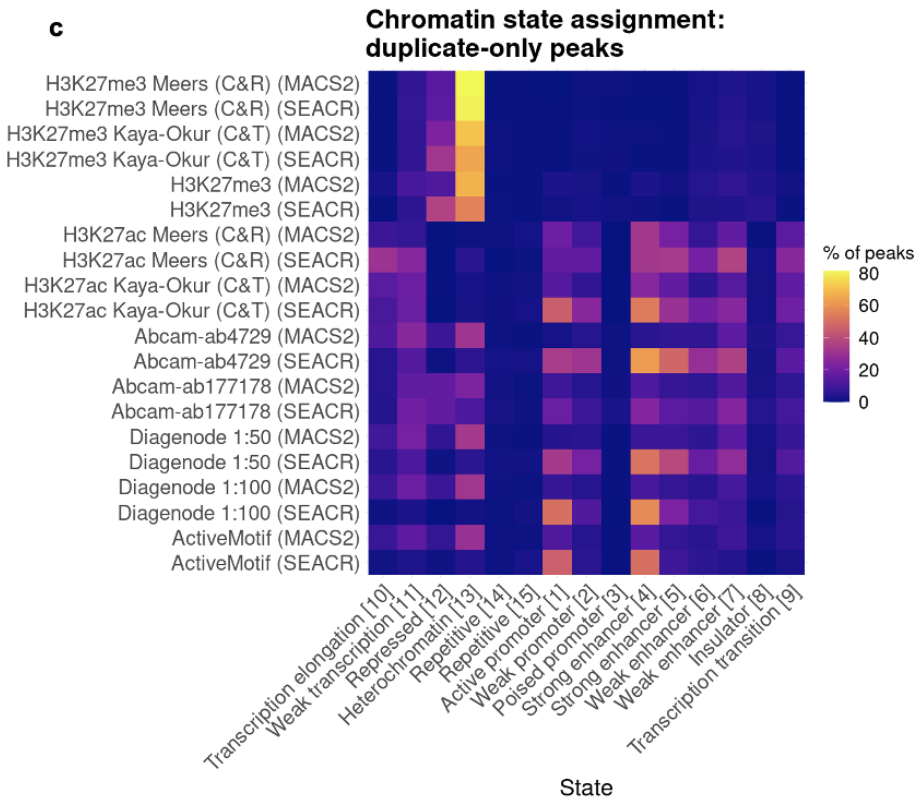
Supplementary Figure 4. Results of qPCR amplification of genes falling into most significant ENCODE H3K27ac peak regions (positive controls; green) versus least significant (negative controls; purple) in CUT&Tag experiments performed with top-performing antibodies, with and without HDAC inhibitor sodium butyrate (NaB; 5 mM).

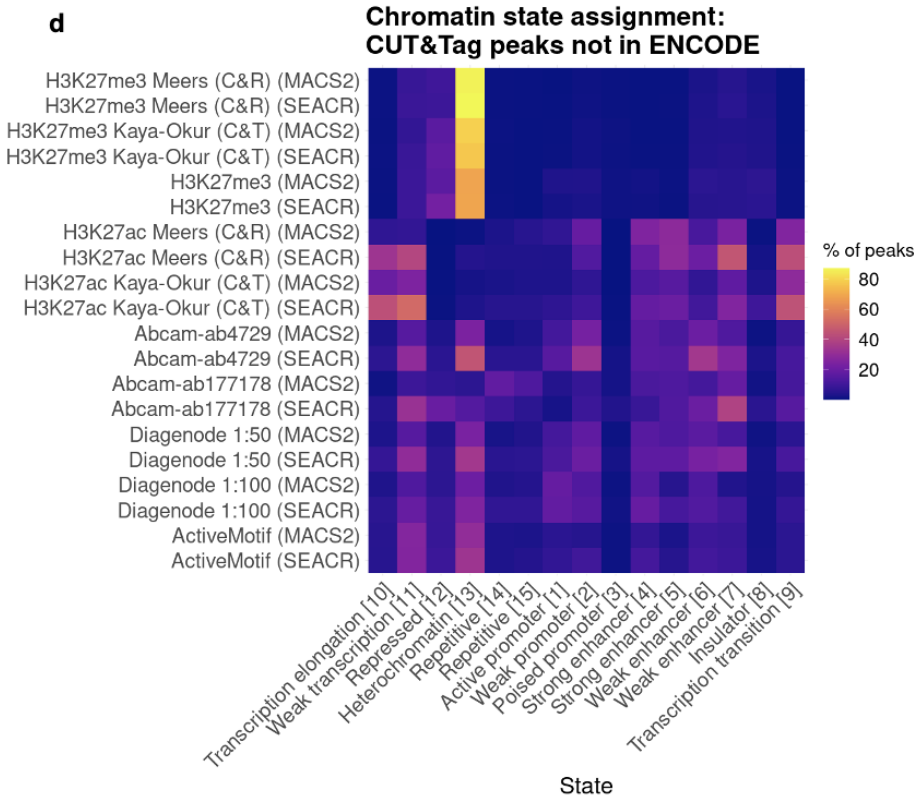
bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



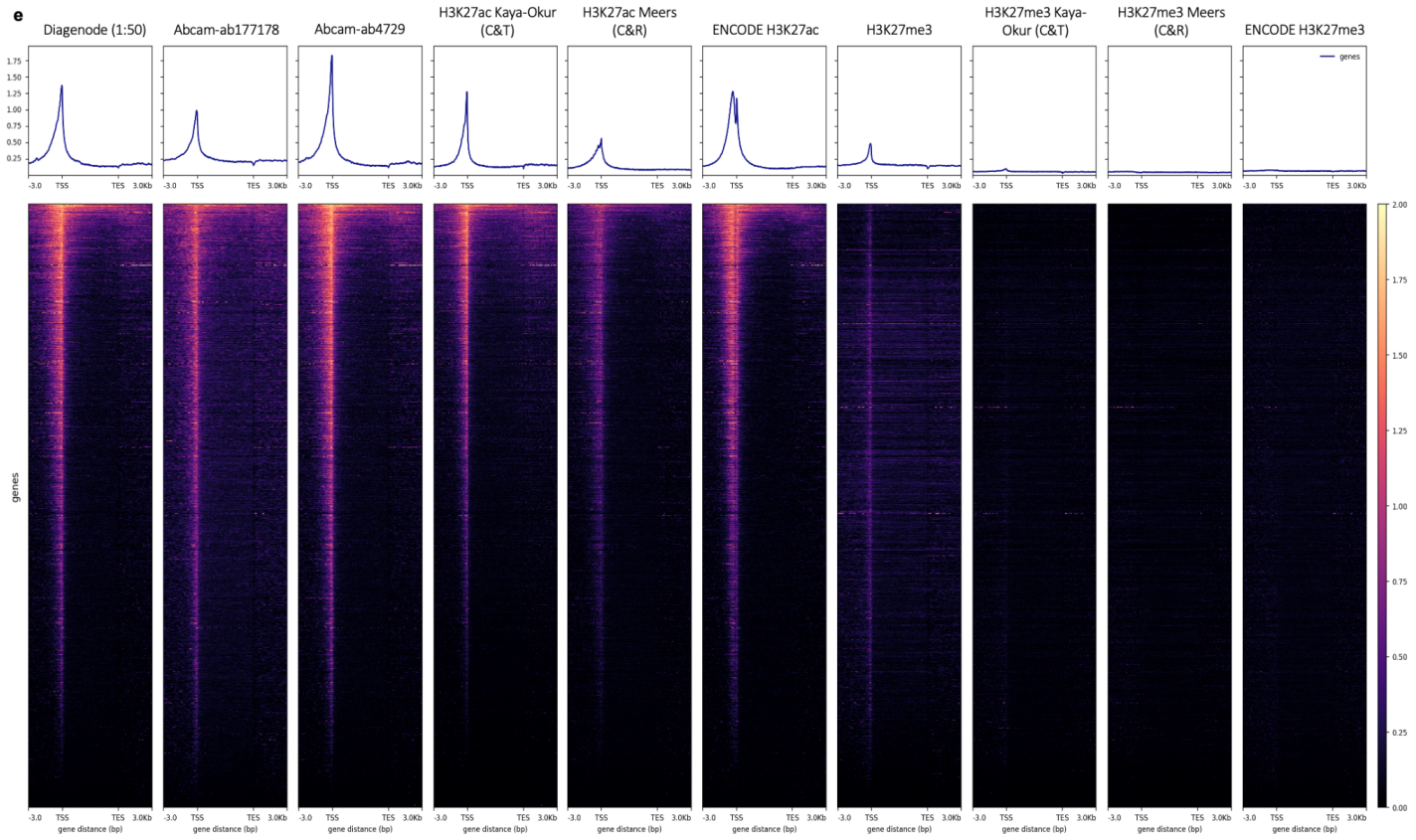


bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.





bioRxiv preprint doi: https://doi.org/10.1101/2022.03.30.486382; this version posted March 31, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 5. a, Regulatory element assignment to CUT&Tag, CUT&RUN, and ENCODE peaks. Chromatin state assignments of: **b**, all peaks, called with SEACR or MACS2, **c**, CUT&Tag peaks specific to duplicate-containing samples, **d**, CUT&Tag peaks not in ENCODE H3K27ac. **e**, Heatmaps showing average read coverage around hg19 transcription start sites, with all samples subsampled to 2 million reads.