The human transcription factor occupancy landscape viewed using high-resolution *in situ* base-conversion strand-specific single-molecule chromatin accessibility mapping

Georgi K. Marinov[1,#], Benjamin R. Doughty[1], Julia M. Schaepe[2], Tong Wang[1], Madison M. Smith[3], Minyong Chen[3], Anshul Kundaje[1,4], Zhiyi Sun[3], and William J. Greenleaf[1,5,6,7,#]

[1] *Department of Genetics, Stanford University, Stanford, CA 94305, USA*
[2] *Department of Bioengineering, Stanford University, Stanford, CA 94305, USA*
[3] *New England Biolabs Inc., 240 County Road, Ipswich, MA 01938, USA*
[4] *Department of Computer Science, Stanford University, Stanford, CA 94305, USA*
[5] *Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA*
[6] *Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[7] *Chan Zuckerberg Biohub, San Francisco, California, USA*
[#] *Corresponding author*

Abstract

Chromatin accessibility profiling is a key tool for mapping the location of *cis*-regulatory elements (cREs) in the genome and tracking chromatin state dynamics during development, in response to various external and internal stimuli, and in disease contexts. Single-molecule footprinting (SMF) methods that rely on the labeling of individual accessible DNA bases have emerged in recent years as a powerful chromatin accessibility mapping approach, as they provide not just an average readout of accessibility over a given genomic position but also the distribution of accessibility states within a population at the level of individual original DNA molecules. However, SMF approaches have been limited either in their resolution or in labeling readout accuracy. To address these limitations, we have developed a high-resolution strand-specific single molecule footprinting approach (C→U/T-ssSMF) based on the application of highly active sequence context-independent endogenous methylation-insensitive double-strand DNA (dsDNA) deaminases (CseDa01 and LbDa02), which convert accessible cytosines into uracils (and in turn into thymine after amplification). We demonstrate the application of the method to mapping fine-grained single-molecule accessibility states in human cell lines in both a short and a long-read format, and quantifying the occupancy states of individual transcription factors (TFs) as well as TF co-accessibility and strand-specific accessibility patterns.

Introduction

In the vast majority of eukaryotes, the majority of genomic DNA is wrapped around nucleosomal particles, making it physically inaccessible to transcription factors and transcription initiation machinery. In turn, active *cis*-regulatory regions, such as promoters, enhancers, and insulators, are characterized by an open-chromatin, nucleosome-depleted configuration, and this property often defines the position and status of active cREs along the genome[1,2]. The physical protection of nucleosome-bound DNA was appreciated very early in the development of the molecular biology toolkit, demonstrated by the use of nucleases such as DNase I to identify open chromatin at promoter and enhancer elements around individual genes[3–5]. Later technological advances enabled the coupling of DNAse I-based chromatin accessibility profiling to microarray[6,7] and high-throughput sequencing[8,9] readouts, eventually allowing truly genome-wide open chromatin profiling. Finally, the Tn5 transposase, in the form of ATAC-seq[10] (**A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing), became the tool of choice for cleavage-based chromatin accessibility mapping due to its high sensitivity and ease of application. Both DNase-seq and ATAC-seq allow not just for the mapping of open chromatin regions, but also for the footprinting of occupancy by individual TFs, at least in some cases[9–11].

While cleavage-based methods efficiently identify candidate cREs, little information about the long-range physical organization of and co-accessibility patterns within individual DNA molecules can be derived from these short-

read data, and because they are enrichment-based, they also do not allow for the measurement of absolute occupancy/protection levels. Single-molecule footprinting is an orthogonal method for measuring chromatin accessibility along single DNA molecules that addresses these shortcomings. Initial SMF-like methods used DNA methyltransferases (MTases) such as the Dam DNA MTase[12,13] (which generates m^6A in GATC contexts), the CpG context-specific M.SssI 5-methylcytosine (5mC) MTase[13,14] (applicable in yeast, where there is no endogenous methylation), and the GpC context-specific M.CvPI MTase[15,16] . Subsequently, M.CvPI GpC-context 5mC labeling was coupled to bisulfite conversion and high-throughput sequencing readout as NOMe-seq[17,18]. While powerful, the resolution of these methods is limited by the sparseness of the sequence contexts targeted for modification, as CpG or GpC dinucleotides are found on average only once more than every 20 bp in the genome, which is considerably longer than the length of the recognition sites of most eukaryotic transcription factors. To improve this resolution, a combination of both 5mC MTases was used for higher-resolution SMF in organisms lacking endogenous methylation, such as *Drosophila melanogaster*[19]. More recently, CpC context-specific 5mC MTases such as M.CviQIX[20] have also been described; however, equalizing the activities of three different enzymes is challenging. Another drawback of traditional NOMe-seq is the use of bisulfite conversion, which is highly destructive to DNA and thus limits read lengths; this limitation is to an extent addressed by non-destructive enzymatic conversion methods (e.g. EM-seq[21,22]) enabling high-quality SMF readouts[23,24].

Higher resolution and longer read lengths have more recently been achieved by combining m^6A methylation and long-read single-molecule sequencing, first using the m^6A MTase EcoGII[25] coupled with long-read nanopore sequencing for long-range single-molecule profiling (SMAC-seq[26]), later followed by Hia5[27] coupled with PacBio sequencing (Fiber-seq[28]). However, these approaches too suffer from shortcomings, including the inability to amplify the m^6A-labeled DNA, the relative inefficiency of the m^6A MTases (which only label 50-70% of accessible adenines), and the high error rates of modification detection using long-read single-molecule sequencing platforms.

The ideal labeling reagent would exhibit near-100% activity for accessible substrates, allow amplification, exhibit no sequence preference, and be unaffected by the presence of endogenous methylation. The use of DNA deaminases that are active on dsDNA substrates has long been proposed as a promising such approach[26], as these enzymes can convert accessible bases *in situ* enabling both subsequent amplification and direct readouts in a variety of formats. Here we describe the use of the recently identified CseDa01 and LbDa02 dsDNA cytosine deaminases as a tool for high-resolution chromatin accessibility profiling, in several different readout formats, as C→T-SMF. We apply strand-specific (C→T-ssSMF) to map and quantify TF footprinting and co-accessibility patterns in human lymphoblastoid cells.

Results

High-resolution chromatin accessibility profiling using the non-specific CseDa01 dsDNA deaminase

The use of cytosine deaminases for chromatin accessibility profiling[26] has been hampered by the lack of available enzymes that act highly efficiently on dsDNA[21,22,29]. A large set of deaminases were recently characterized for their activity on dsDNA[30], revealing two – CseDa01 and LdDa02 – with ideal properties for SMF, i.e. very high activity on dsDNA, no sequence specificity, and complete insensitivity to 5mC methylation. The latter is important because while CpG positions can be filtered out computationally (as they are endogenously methylated in many eukaryotes), that results in decrease of the effective resolution. Furthermore, CpG positions are often concentrated in regulatory elements in mammals, there are transcription factors, such as NRF1[31], that are specifically methylation sensitive, and some eukaryotes also exhibit extensive 5mC methylation in non-CpG contexts[32,33].

Previous work (reproduced in Supplementary Figure 1) demonstrated that the CseDa01 and LbDa02 enzymes exhibit the highest activity out of all deaminases on dsDNA (as well as on ssDNA), have no sequence specificity, and exhibit equal activity on both unmodified substrates and common cytosine modifications. For comparison, enzymes such as MsddA show relatively high activity on unmodified C, but very low activity on 5mC, while SsdA and DddA show much lower absolute activity, and exhibit even lower activity on 5mC compared with unmodified C.

We thus set out to develop a high-resolution version of the SMF assay based on the *in situ* conversion of cytosines to uracils using either of these two enzymes (Figure 1a). Such an assay would provide a resolution of ∼5 bp for mammalian genomes, which is slightly lower than the theoretical resolution of m^6A-based SMF (Figure 1b), but with the advantages of allowing for reliable base calling, more efficient modification of accessible positions, and the ability to carry the accessibility tags through PCR amplification.

We used two different preparations of the two enzymes in our work – either in vitro translation (using the PURExpress kit) products (as will be shown further, these enzymes are sufficiently highly active even for unpurified PURExpress preps to be viable tools for probing accessibility) and purified recombinant CseDa01 (see the Methods section for details).

We first measured the activity of CseDa01 on naked human genomic DNA (gDNA), by deaminating 1 μg of gDNA with 1 μL of PURExpress CseDa01 prep over a time course. After generating Tn5 transposition libraries from the deaminated DNA (see the Methods section for details), we observed ∼70% conversion within the first five minutes, with complete deamination achieved by 90 minutes. We observed

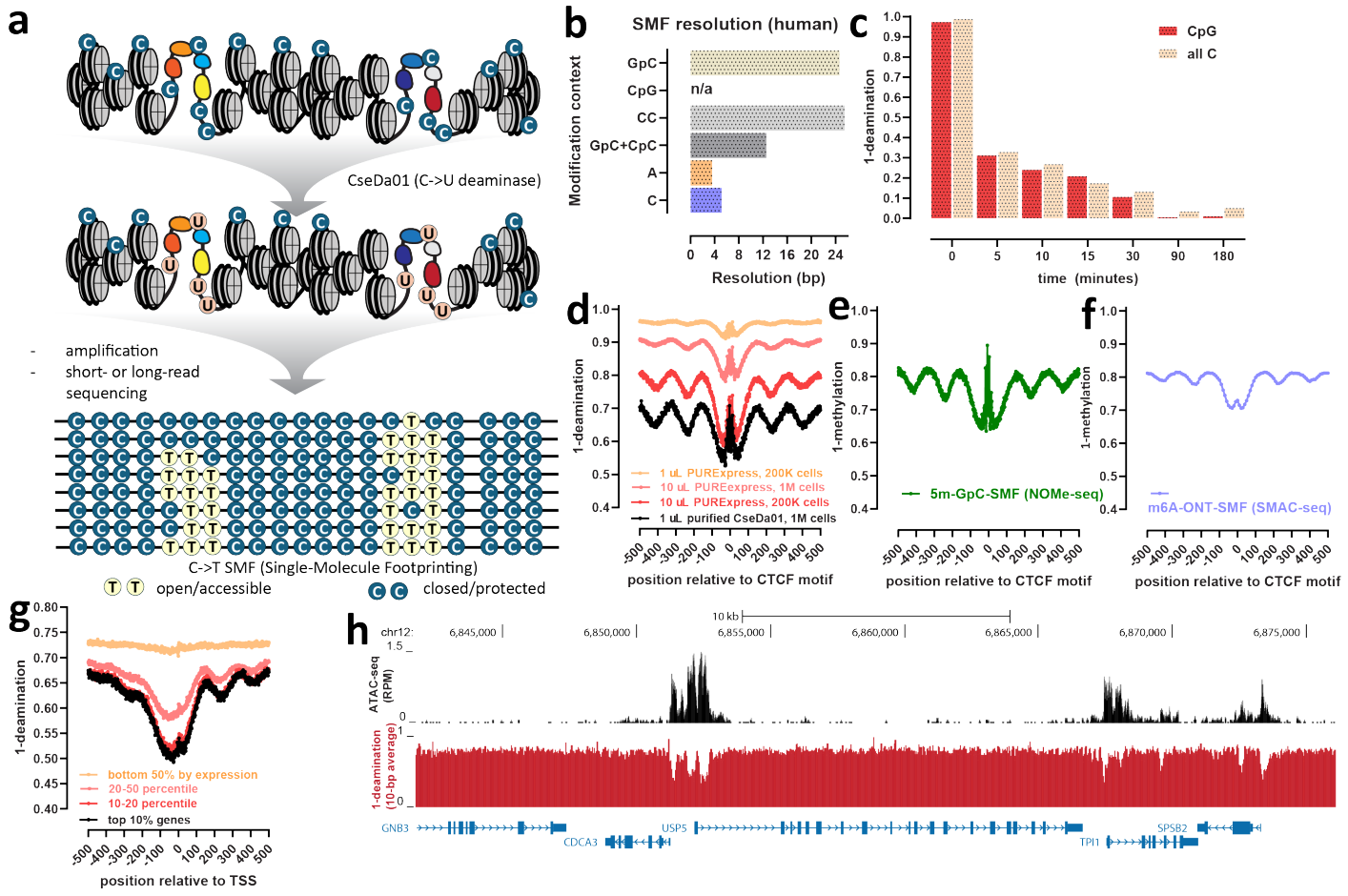


Figure 1: High-resolution chromatin accessibility profiling using the non-specific methylation-insensitive CseDa01 dsDNA deaminase. (a) Outline of the C→T-SMF assay. (b) Average resolution of accessibility profiling using different enzymatic modifications. Shown is the average distance between instances of each sequence context (or combination of) in the human genome. (c) The CseDa01 enzyme efficiently deaminates modified and unmodified cytosines. Human gDNA was isolated and 1 μg of it was incubated with 1 μL of CseDa01 PUREexpress prep for the indicated number of minutes. The average genome-wide deamination levels around all CpG and all C positions in the human genome are shown. (d-g) CseDa01 C→T-SMF accurately captures chromatin accessibility patterns in mammalian cells. (d) Short-read C→T-SMF metaprofiles around CTCF motifs within CTCF ChIP-seq peaks for experiments carried out with PURExpress CseDa01 preps or purified CseDa01 and the indicated amount of enzyme and number of cells (K562 cells, crosslinked). (e) Metaprofiles around CTCF motifs within CTCF ChIP-seq peaks for GpC methylation-based SMF (5mC-GpC-SMF; K562 cells, crosslinked). (f) Metaprofiles around CTCF motifs within CTCF ChIP-seq peaks for m^6A nanopore-based SMF (m^6A-ONT-SMF; GM12878 cells, crosslinked). (g) Short-read C→T-SMF metaprofiles around TSSs divided by expression levels (K562 cells, crosslinked). (h) Correspondence between ATAC-seq and short-read C→T-SMF profiling of chromatin accessibility. Shown is an example UCSC Genome Browser snapshot of ATAC-seq accessibility levels, measured in reads per million (RPM), and C→T-SMF accessibility measurements (shown as a smoothed 10-bp average of the inverse of deamination levels).

no meaningful difference in conversion levels between CpG positions and non-CpG positions (Figure 1c), confirming CseDa01's methylation insensitivity.

Next, we aimed to identify the amount of enzyme required to accurately map chromatin accessibility *in vivo*. Because recent studies by us and others[24] have shown that the dissociation kinetics of TFs can be very fast relative to the MTase treatment in traditional NOMe-seq/SMF, making it impossible to see footprints for such TFs, we carried out most of our experiments on formaldehyde-crosslinked cells (which we refer to as XL-SMF). We used Tn5 transposition to generate sequencing libraries for these tests (see Methods for details). As we observed slightly more *in vivo* activity with CseDa01 than with LbDa02 (Supplementary Figure 2), and due to the reported much higher activity of CseDa01 for more exotic modifications such as gluco-

sylated 5hmC (5gmC), even if those are rare in the human genome[30], we focused on CseDa01 as the enzyme of choice for subsequent experiments. We determined optimal enzyme levels by comparison to GpC-based NOMe-seq datasets using genome-wide metaprofiles around CTCF motifs within ChIP-seq peaks (as these are characterized by very high chromatin accessibility and a stereotypical nucleosome positioning pattern flanking them) and around transcription start sites (TSSs). By comparing C→T-SMF CTCF metaprofiles to NOMe-seq metaprofiles we determined that 1 μL of purified CseDa01 (at ∼0.58 ng/μL) provides similar measures of accessibility when deaminating $\sim 1\times10^6$ crosslinked human cells, while 10 μL of CseDa01 PURExpress prep produce an equivalent accessibility profile from ∼200,000 human cells (Figure 1d). Metaprofiles around CTCF sites for these conditions are comparable to NOMe-seq metaprofiles (Figure 1e) and SMAC-seq profiles (Figure 1f), though we note that we consistently observe 5-10% lower baseline protection levels over nucleosome-occupied DNA than is measured using GpC methylation. These metaprofiles also accurately capture the key aspects of CTCF occupancy – a highly accessible region around the CTCF binding site, a protection footprint over it, and strongly positioned nucleosomes nearby. Other basic aspects of chromatin accessibility are also captured – highly expressed genes are more accessible on average over their promoters than genes that are expressed at lower levels (Figure 1g), and average deamination profiles closely match ATAC-seq profiles (Figure 1h and Supplementary Figure 4). We also successfully generated high-quality deamination-based SMF data from native unfixed samples (Supplementary Figure 3). Overall, these results demonstrated that the CseDa01 deaminase can be used to produce high-quality, high-resolution, single molecule chromatin accessibility data.

As a side note, we propose a unified descriptive nomenclature of single-molecule assays that provides information on the enzymes and modifications used and the readout. This is due to the expansion of the toolkit of molecular probes used to map accessibility and the proliferation of assays that are essentially the same in principle (i.e. m^6A methylation as a tool for accessibility mapping has been published under the names SMAC-seq[26,36], Fiber-seq[28], SAMOSA[37], SAM-seq[38], MATAC-seq[39] and STAM-seq[40], nanopore-based NOMe-seq/SMF under the names NanoNOMe[34] and MeSMLR-seq[35], and while this manuscript was in preparation, several other dsDNA deaminases were used as probes for chromatin accessibility under the names FOODIE[41], ACCESS-ATAC[42], TDAC-seq[43], and DAF-seq[44]). These are all non-descriptive names often referring to the same thing. Instead, we propose the following descriptive naming convention for the full names of SMF assays:

[enzyme*]-[modification/conversion]-[readout*]-[XL*]-SMF

Where $*$ refers to optional information, i.e. short-read sequencing needs not be specified directly (but long-read platform do) and "XL" refers to crosslinked SMF, and the enzyme name need only be mentioned when introducing the assay in full.

Under this convention dsDNA deamination-based accessibility mapping is referred to C→T-SMF if the DNA is amplified after deamination, and C→U-SMF if DNA is read out natively using single-molecule sequencing. C→T-XL-SMF implies crosslinking. CseDa01-C→T-XL-SMF is the full name of the experiments shown in Figure 1.

Traditional NOMe-seq is to be referred to as M.CvPI-5mC-GpC-SMF (or 5mC-GpC-SMF in short) while long-read m^6A-based assays as EcoGII-m^6A-ONT-SMF (or m^6A-ONT-SMF in short) for SMAC-seq and Hia5-m^6A-PB-SMF for Fiber-seq. We believe that having a unified nomenclature will greatly help streamline the literature in the future as the molecular toolkit of accessibility probes expands further.

Single-molecule accessibility profiling using short- and long-read C→T-SMF

We then turned our attention to examining single-molecule C→T-SMF profiles using different library building protocols and sequencing readouts. We obtained ∼300× sequencing coverage for the GM12878 lymphoblastoid cell line by preparing strand-specific libraries that were sequenced on a Ultima Genomics sequencer (see the Methods section for details), which we refer to as C→T-ssSMF (i.e. strand-specific SMF). Figures 2a and 2b show C→T-ssSMF single molecules over representative occupancy sites for the CTCF and NRF1 transcription factors. We observe footprints over the occupancy site both at the level of single molecules and in aggregate profiles for both proteins.

We also generated low-coverage C→T-ONT-SMF for the K562 cell line by carrying out whole-genome amplification of deaminated DNA and sequencing the resulting DNA using nanopore sequencing on a MinIon flowcell. This readout reproduces metaprofiles over CTCF sites and TSSs that are similar to short-read based C→T-SMF (2c-d) demonstrating the viability of using long reads together with CseDa01 SMF inputs.

Finally, we generated high-coverage targeted C→T-ONT-SMF for the promoters of three genes (*IFIT6*, *IFI3* and *USP18*) in K562 cells with and without treatment with interferon (IFN). We used primers to amplify these amplicons that were carefully designed to avoid potentially deaminated bases (see the Methods section for details). These promoters are partially (*IFIT6* and *USP18*) or fully (*IFI3*) accessible in resting cells, with accessibility levels increasing upon IFN treatment (Figure 2e-g). Overall, these data demonstrate the generation of CseDa01-based SMF data in a variety of both short- and long-read based formats.

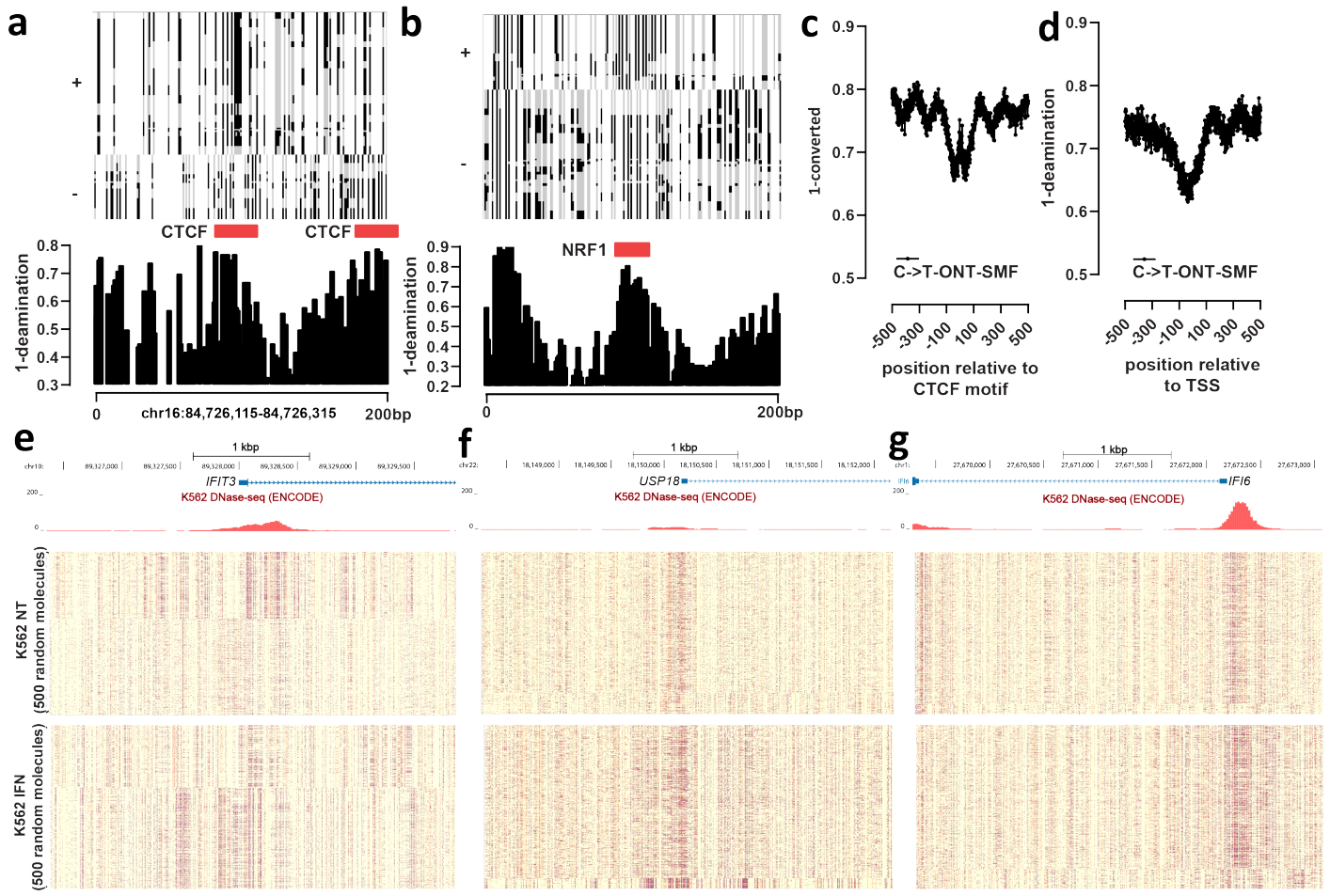


Figure 2: Single-molecule accessibility profiling using short- and long-read C→T-SMF. (a) Representative C→T-SMF footprint patterns around a CTCF binding site. (b) Representative C→T-SMF footprint patterns around an NRF1 binding site. (c) C→T-ONT-SMF metaprofiles around CTCF motifs within CTCF ChIP-seq peaks (K562 cells, crosslinked) (d) C→T-ONT-SMF metaprofiles around TSSs (K562 cells, crosslinked) (e-f) Targeted C→T-ONT-SMF amplicons around the *IFIT3*, *USP18*, and *IFI6* genes. Shown are 500 random single molecules (at 5-bp resolution) for untreated ("NT") and interferon-treated ("IFN") K562 cells.

Transcription factor footprinting patterns in the human genome

We used the high-depth C→T-ssSMF dataset to address several questions – first, we aimed to measure aggregate levels of TF footprinting in the human genome for a collection of TFs; second, we aimed to measure the relationship between this footprinting strength and TF occupancy levels as measured by ChIP-seq; third, we asked if TF footprinting exhibits differential patterns on the two strands of DNA. To these ends, we compiled sets of TF recognition motifs likely to be highly occupied for 64 TFs for which ChIP-seq data exists from the ENCODE Project Consortium[1,2], by identifying binding motifs that are both within ≤100 bp of the ChIP-seq peak, and also the nearest to the peak. These selection criteria aim to exclude potential "bystander" motif matches that might be in the vicinity of the actually occupied site but not bound by the TF. We split these sets of motif matches into quintiles based on ChIP-seq peak strength, and calculated two sets of metaprofiles for each TF – deamination levels over all sites for the forward/plus and reverse/minus strands, and overall deamination levels for each quintile. Detailed results for all TFs are shown in Supplementary Figures 5 to 68. Examining these metaprofiles revealed six general footprinting patterns.

The first pattern is exemplified by the CTCF insulator protein, which is well known for its ability to drive extremely strong nucleosome positioning in its vicinity[45,46]. This positioning is also observed in our data, as expected; however, we also observe strong strand asymmetry in deamination profiles over the positioned nucleosomes around CTCF (Figure 3a). Such strand assymetry of protection was also been observed in yeast using EcoGII-m^{6}A-ONT-SMF[26] based on the highly accurate chemical maps of nucleosomes that are available for *Saccharomyces cerevisiae*[47].

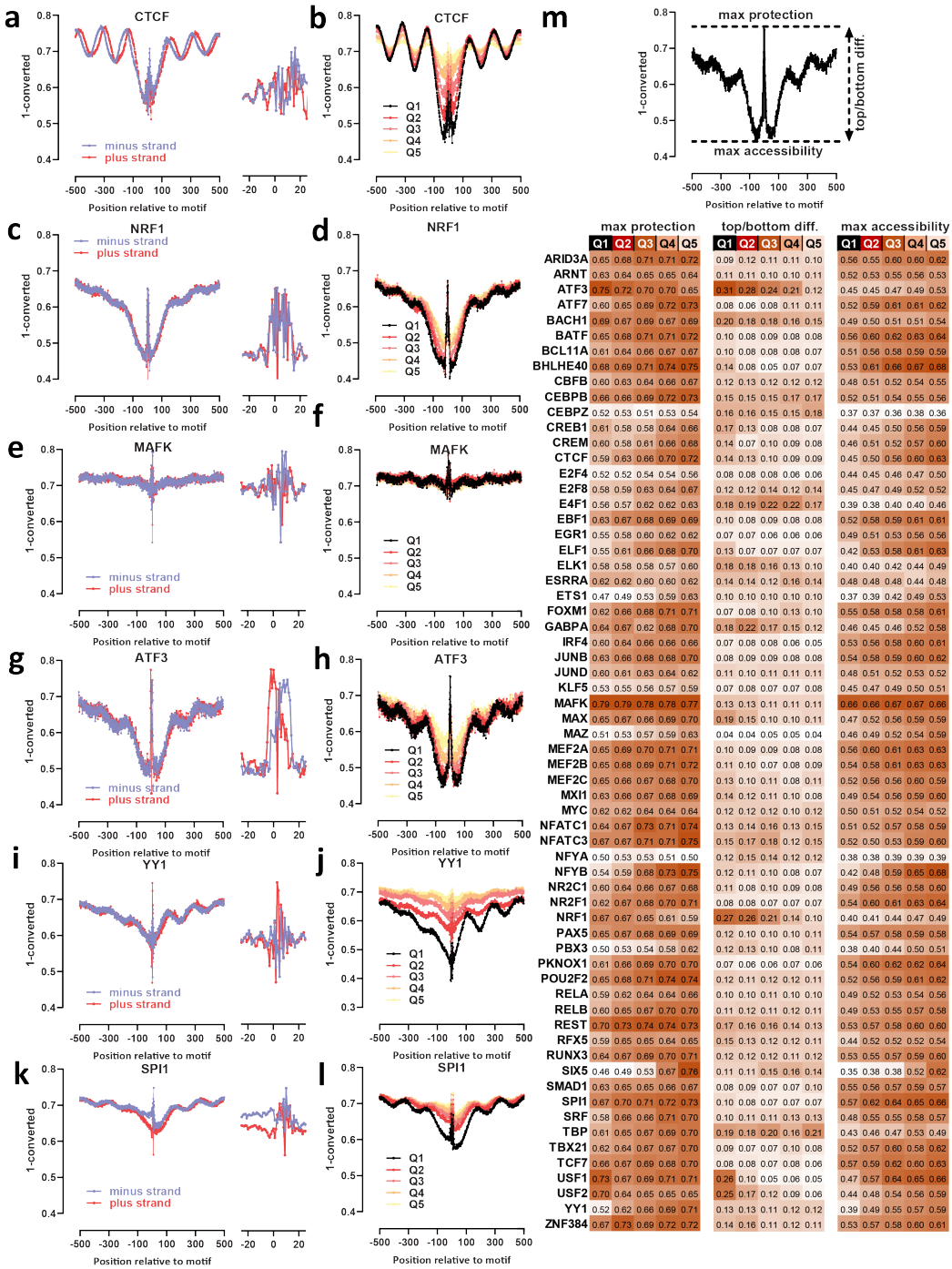


Figure 3: Transcription factor footprinting patterns in the human genome. (a-l) Metaprofiles around the motifs for each TF within ENCODE ChIP-seq peaks and closest to the ChIP-seq peak summit. For each TF, the forward- and reverse-strand C→T-ssSMF metaprofiles are shown on the left, while on the right the profiles are split into quintiles according to the strength of ChIP signal. A zoomed-in forward- and reverse-strand metaprofile is also shown. (m) Quantification of footprinting strength over motifs, split by ChIP peak strength. Three different scores are shown – "maximum protection" shown in the left columns measures the distance from zero to the protection peak over the motif, while "top versus bottom difference" measures the difference between the protection peak over the motif and the lowest protection level around the motif in the metaprofile. The "maximum accessibility" metric shows the depth of the accessibility well and thus average local accessibility associated with each TF.

It arises as a result of the physical structure of the nucleosome which preferentially exposes one strand of DNA on one side of the nucleosomal dyad, and the other strand in a mirrored pattern on the other side[26]. We do not have chemical maps that precisely position CTCF-flanking nucleosomes in GM12878 cells, but despite that potential lack of precision the magnitude of the strand asymmetry observed is larger than that found in yeast (∼5% difference compared to ∼1% difference). We discuss the likely reasons for these observations in a later section. We do not observe substantial strand asymmetry within the CTCF occupancy site itself (Supplementary Figure 18)

As an example of our second observed pattern, the NRF1 TF exhibits extremely strong footprinting over the TF recognition motif itself (Figure 3c-d), but weak positioning of nearby nucleosomes and little strand asymmetry over the occupancy site (Supplementary Figure 48).

The MAFK factor is known to have a repressive role[48,49] and to occupy sites in closed chromatin, and is an example of our third observed pattern of very weak open chromatin

signature with a footprint largely overlaid on top of the baseline protection levels (Figure 3e-f). The NRSF/REST transcription factor is also a well-known repressor, specifically of neuronal genes[50], and can strongly position nucleosomes[51,52]. Its footprinting profiles are a mixture of that of CTCF and MAFK, i.e. inaccessible chromatin and open chromatin regions with strongly positioned nucleosomes nearby (Supplementary Figures 55).

The fourth general type of footprint pattern is exemplified by ATF3, a member of the large family of bZIP TFs. ATF3 exhibits both very strong footprints and extreme strand asymmetry (Figure 3g-h), with as much as 25% difference in absolute accessibility levels between the two strands. This asymmetry is consistent with the structure of bZIP TFs, which typically function as homodimers and/or heterodimers[53,54], or oligomers[55]. This asymmetric protection pattern is consistent with two TFs in a dimer binding DNA on opposite strands in such a manner that the bound strand is well protected while the opposite strand is much more exposed (see also Figure 4b for more details).

Strand-specific accessibility is observed for many other TFs, though not always as clearly as for ATF3. Most bZIP factors, e.g. ATF7 (Supplementary Figure 8), BACH1 (Supplementary Figure 9), CREB1 (Supplementary Figure 16), and CREM (Supplementary Figure 17), exhibit some amount of asymmetry. Some bZIP factors show weaker strand asymmetry, e.g. JUNB (Supplementary Figure 31), JUND (Supplementary Figure 32) and CEBPZ (Supplementary Figure 15).

Other TFs that are known to function as dimers, for example bHLH factors such as USF1/USF2[56,57], also exhibit very strong strand asymmetry of a similar mirror pattern (Supplementary Figures 65 and 66); asymmetry is also observed, though less distinctly, for MYC (Supplementary Figure 41), MAX (Supplementary Figure 35), E4F1 (Supplementary Figure 21), MXI1 (Supplementary Figure 40), and BHLHE40 (Supplementary Figure 12).

We also identify some TFs that exhibit strand-specific footprinting patterns, but only on one strand without a mirrored footprint on the other strand. These include GABPA (Supplementary Figure 29), BATF (Supplementary Figure 10), CBFB (Supplementary Figure 13), CEBPB (Supplementary Figure 14), NFATC3 (Supplementary Figure 43) and the members of the NFY complex[58,59] (Supplementary Figures 44 and 45).

A fifth accessibility/footprinting pattern is unique to the YY1 TF (Figure 3i-j), which exhibits very strong footprints combined with strong nucleosome positioning downstream of the YY1 motif, with the top quintile of YY1 occupancy sites exhibiting this pattern most clearly. YY1 is a structural component of enhancer-promoter loops[60], and thus its occupancy sites are strongly enriched for promoters of active genes. The positioned nucleosomes observed are likely largely associated with proximal active TSSs. However, YY1 sites nevertheless appear to be a strong positioning factor themselves, because nucleosomes are positioned downstream of a YY1 binding site approximately as strongly as they are around annotated TSS for the top 10% most active genes (compare Figure 3j to Figure 1g). Finally, YY1 exhibits orientation dependence, positioning mucleosomes specifically downstream (to the right) relative to the orientation of the YY1 motif.

We also identify a final unique pattern, exhibited most strongly by the SPI1 TF (Figure 3k-l and Supplementary Figure 60), but also by MAZ (Supplementary Figure 36), BCL11A (Supplementary Figure 11), and to a lesser extent ELF1 (Supplementary Figure 24). In these cases the region upstream (to the left) of the motif is more accessible due to preferential accessibility on the minus strand. This region of increased accessibility is broad and not easily attributable to specific DNA sequence elements.

Finally, we quantified the absolute occupancy of the TF based on the levels of protection against deamination. We note that footprint protection levels arise from a combination of TF occupancy and protection from modification/conversion due to occupancy by other proteins, including nucleosomes and other chromatin-associated factors. The complex origins of protection signals are clearly seen in the case of CTCF, where sites in the top ChIP quintile are much more accessible around the motif than sites in the bottom ChIP quintile (Figure 3b), while the absolute height of the protection footprints at the motif is lowest for the top quintile and highest for the bottom, likely due to protection by nucleosomes occupying sites that are not bound by CTCF. In the extreme, factors like MAFK actually bind to closed chromatin, making their occupancy levels impossible to measure, as the nucleosome itself is already protecting DNA over the binding site.

Given these considerations, we measured occupancy levels using two different metrics that provide lower and upper bounds on occupancy levels – first, the absolute height of footprints, and second, the difference between the absolute height of the footprint and the bottom of the "accessibility well". The latter metric aims to provide a minimum estimate of occupancy levels, while the former sets the upper bound. Figure 3m summarizes these metrics for all TFs. For many TFs, such as CTCF mentioned above, absolute protection levels are lowest for the strongest ChIP sites, as those are also most accessible as a baseline. TFs in this class also exhibit weak or no correlation with ChIP strength by the top-versus-bottom difference metric. For other very strongly footprinting factors, such as USF1, USF2, NRF1, REST, and ATF3, the top ChIP sites also exhibit the highest levels of protection by both metrics. Thus, for example, for ATF3 and NRF1, we estimate that a minimum of ∼30% of top-quintile sites are protected from deamination. We also measured the overall levels of local accessibility for each TF (calculated as the difference between the nucleosomal protection baseline and the bottom of the "accessibility well"), and this metric correlates well with occupancy for all TFs except for MAFK.

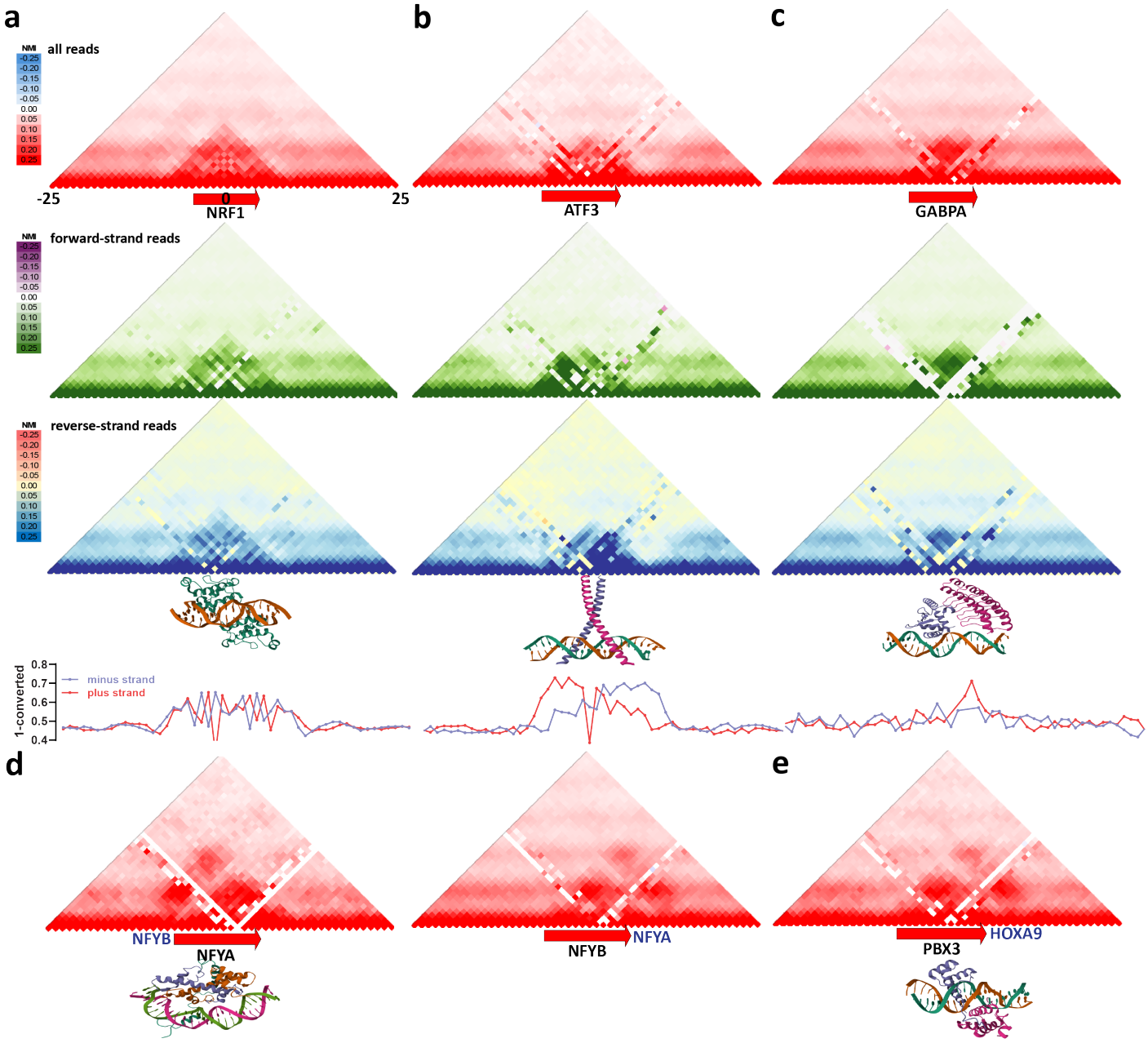


Figure 4: Co-accessibility patterns around transcription factor binding sites. (a) Average NMI values around NRF1 occupancy sites for all reads and forward- and reverse-strand (with respect to the TF motif) reads only. Also shown is the crystal structure of an NRF1 DNA binding domain (DBD) dimer (PDB accession 8K3D), as well as strand-specific footprint profiles around NRF1 occupancy sites. (b) Average NMI values around ATF3 occupancy sites for all reads and forward- and reverse-strand (with respect to the TF motif) reads only. Also shown is the crystal structure of a bZIP dimer (PDB accession 1JNM), as well as strand-specific footprint profiles around ATF3 occupancy sites. (c) Average NMI values around GABPA occupancy sites for all reads and forward- and reverse-strand (with respect to the TF motif) reads only. Also shown is the crystal structure of a GABP dimer (PDB accession 1AWC), as well as strand-specific footprint profiles around GABPA occupancy sites. (d) Average NMI values around NFY occupancy sites. Also shown is the crystal structure of an NFY complex (PDB accession 4AWL) (e) Average NMI values around PBX3 occupancy sites. Also shown is the crystal structure of a PBX-HOXA dimer (PDB accession 1PUF)

Co-accessibility patterns around transcription factor binding sites

Next, we quantified co-accessibility patterns around TF occupancy sites in the human genome using the GM12878 C→T-ssSMF dataset. To this end, we calculated the Normalized Mutual Information (NMI) metric that we previously developed[26], which aims to measure the extent of co-accessibility observed above that expected by chance given average protection/accessibility levels. We computed this metric genome-wide for all pairs of informative positions for all TFs examined (Supplementary Figures 69-72). In all these co-accessibility maps, we observed a globally increased level of co-accessibility at a distance of approximately 10-bp; this likely reflects the length of the DNA double helix turn.

Co-accessibility maps, together with with crystal structures, helped us further rationalize our observations of strand-specific TF occupancy. For example, Figure 4a shows average NMI profiles around NRF1 occupancy sites. NRF1 functions as a homodimer[61], but unlike many bZIP factors exhibit only limited footprint strand asymmetry. We find that positions extending to a few basepairs outside the NRF1 recognition motif are strongly preferentially co-accessible. Consistent with this extended co-accessibilty pattern, the NRF1 crystal structure demonstrates that monomers each contact these positions, while also likely protecting substantially both strands over much of the central binding region. Strand-specific co-accessibility maps that show only a slight offset between the co-accessibility footprints on the forward- and reverse-strands are consistent with the uniform protection of both DNA strands observed in the crystal structure.

In contrast, bZIP TF dimers are structured as thin α-helices that bind to the DNA major groove, likely leaving the other strand prominently exposed (Figure 4b). For ATF3 we observe a co-accessibility pattern that peaks within the monomer binding sites, with a lower-intensity co-accessibility peak between the two. This monomer-centric pattern is supported by strand-specific co-accessibility maps, which exhibit a strong offset between the co-accessibility footprints on the forward- and reverse-strands with respect to the ATF3 motif.

A distinct pattern is exemplified by the GABPA TF motif, which exhibits increased accessibility over only one strand, with no mirror pattern around its binding site. These patterns might be explained by GABP[62] dimers preferentially only contacting one strand, consistent with the GABP-DNA co-crystal structure. This asymmetric, strand-specific binding mode is consistent with the lack of substructure within the co-accessibility signature (Figure 4c), as well as in the stronger and wider co-accessibility footprint on the forward (with respect to the motif) strand.

Co-accessibility maps can also reveal co-occupancy patterns between TF co-factor pairs. For example, Figure 4d shows average NMI profiles for the NFY-A and NFY-B TFs. We observe a co-accessibility signature between the NFY-A motif and the region to the left of it, and a similar co-accessibility pattern at the NFY-B motif, but to the right/downstream of the motif. These proteins function as part of the NFY complex[58,59], and bind DNA together. Another example is the pairing of the PBX and HOXA homeodomain factors[63–65]. Figure 4e shows the average NMI profile for the PBX3 TF in GM12878, exhibiting a prominent co-accessibility peak to the right of the PBX3 recognition motif corresponding to the HOXA co-factor. TFs with such clear co-accessibility signatures also include CEBPZ (Supplementary Figure 69k), USF1 and USF2 (Supplementary Figure 72g-h), and others.

Discussion

In this study, we first hypothesized that CseDa01 and LbDa02 deaminases might be the ideal molecular probes for mapping chromatin accessibility at high-resolution using *in situ* base conversion due to them exhibitin the strongest reported activity on dsDNA substrates out of all known cytosine deaminases and essentially complete lack of sensitivity to endogenous cytosine methylation. We developed and optimized a variety of protocols for the use of these enzymes to profile chromatin accessibility at the single-molecule level in different short-read and long-read formats, as C→T-SMF. We then focused on C→T-ssSMF in the GM12878 lymphoblastoid cell line to map the fine-grained strand-specific features of transcription factor occupancy.

Several emerging features of the dsDNA deaminase that we use (which are likely to be shared by all such enzymes) are notable. We find very strong strand-specific protection/accessibility patterns both for nucleosomes and individual transcription factor occupancy sites, much stronger than those previously observed with m^6A methylation. However, we consistently find that CseDa01 treatment of chromatin results in the conversion of a somewhat higher fraction of nucleotides at baseline, i.e. over nucleosomes, than is observed by GpC methylation of chromatin using the M.CviPI enzyme. These differences probably have to do with the different mechanisms of actions of MTases and deaminases. Methylation by M.CviPI methyltransferase involves making contact with both strands, likely requiring more complete physical access to DNA for efficient labeling. While the m^6A MTases are not sequence-specific and will label any A positions on just one strand, they belong to the same general class of enzymes and may have similar requirements for access to dsDNA. In contrast, deaminases such as CseDa01 act on both dsDNA and on ssDNA, and are therefore likely more active on DNA substrates with more limited exposure, or exposure to only one strand. Therefore efficient probing of strand-specific accessibility comes at the cost of higher conversion rates over otherwise strongly protected substrates, such as nucleosomes.

We used C→T-ssSMF to provide lower-bound and upper-bound estimates for the occupancy of 64 humans TFs in the GM12878 cell line. However, the final quan-

titative measurement of TF occupancy across the genome will require a more comprehensive analysis, possibly using different approaches and tools. Because of the rapid dissociation kinetics of many human TFs, we used formaldehyde crosslinking for the experiments presented in this study. However, fixation does not strictly "freeze" a snapshot of the accessibility landscape at a given moment, but integrates it over an extended period of time (in this case, formaldehyde crosslinking was carried out for 15 minutes). We expect a combination of much more highly active/concentrated probes on native chromatin and/or more rapid crosslinking approaches[66,67] to address these remaining open questions.

Materials and Methods

Except for when explicitly stated otherwise, all analyses were carried out using custom-written Python or R scripts.

Cell lines and cell culture

GM12878 and K562 cells were grown following established protocols[1,2], i.e. RPMI 1640 supplemented with 15% and 10% FBS, respectively, and 1% penn-strep.

Interferon induction was peformed by adding 10 ng/mL of IFNβ (Peprotech, 300-02BC) for 6 hours as previously described[23].

dsDNA deaminase production (PURExpress)

CseDa01 and LbDa02 enzymes were produced using PURExpress as follows. First, codon-optimized coding sequences as described by Vaisvila et al. 2024 were ordered as linear gBlocks from IDT with flanking sequences added containing the T7 promoter at 5' end (GCGAATTAATACGACTCACTATAGGGCTTAAGTATAAGGAGGAAAAAAT) and T7 terminator at 3' end (CTAGCATAACCCCTCTCTAAACGGAGGGGTTTATTTGG). The gBlock sequences were the following:

CsaDa01:

GCGAATTAATACGACTCACTATAGGGCTTAAGTATAAGG
AGGAAAAAATATGGATGATGCACTGGCGGAAAAGGCGCG
TCAAGCCCACAACGTGCTGAAACGCAACATCCCGGGTAA
ACCGGAAATCGGCTTTAACCAGTCCACGGTGTCTATTGG
CGAAGGTACTCTGCCGTCTGGCGAGAAACAACTGTTTGC
TAGCGGTAACGGTGCTAAACTGACCCCGGCCCAGCGTGC
AAAACTGATCGAGCTGGGTGTTCCACCAGAAAACATTTA
TTCTGGCGTAGCGTACATGAAGAAACCAGTTAACAAATC
CAAAGTTAAATCCGTGCGTAAAGCCTACAACCTGGAAAA
CCACGCTGAGCGTGTAATTATTCGCAACGCACCGAAAGG
TACGAAATTCAAGAAATGGGGTATCTCCTGGGCGAAATA
CCAGAAGAACGAACCGTGTTCTAACTGTAAACCTCATGT
ACAGTGTGCATCCTAACTAGCATAACCCCTCTCTAAACG
GAGGGGTTTATTTGG

LbDa02:

GCGAATTAATACGACTCACTATAGGGCTTAAGTATAAGG
AGGAAAAAATATGAACGACGGTAACTCCGAATCTGGCTC
CGATGCGCTGGACAAGAAAATTAAACAGGCAGAACAGGC
CCAAAAGAAAGCGTCCGAACTGAGCGACAAATACGGCTA
TGGCACTGCTCCTAAGAAAACTGTGGCTTCTGACGGTAC
CATCCATACTCTGAGCGGTTGGAAACCACTGGATGAGAA
CAGCGAATTCACCCGTGTCTCTACTGATGACGTAATTAA
TAAAAGCAACGATATTGGCCACAAACTGCGTAACGCGGG
TGCGAACGATCAGGGTGTGCCAGGTCGTTACAACGCCAG
CCACGCTGAAAAGCAACTGTCTATTATCTCTGATAAACC
TATTGGTATCAGCCAGCCGATGTGTACCGATTGTCAAGG
TTATTTCAAATCCCTGGCTATCGGCGAGGGTAAAGAATA
CGTAACTGCTGATCCGAACTACATCCGCATCTTTTCTCC
AGATGGTACTGTGAACGAGATTGAACGTCAACGTTAACT
AGCATAACCCCTCTCTAAACGGAGGGGTTTATTTGG

The gBlock sequences were PCR-amplified using flanking sequence primers, then purified using the MinElute PCR Purification Kit (Qiagen), quantified using a Qubit 1X dsDNA HS Assay Kit (ThermoFisher) and confirmed to be the right size using an Agilent TapeStation.

Then 100-200 ng of input DNA per reaction were used to run multiple PUREexpress (NEB #E6800S) reactions for $\geq$2 hours at 37 °C. The reactions were pooled and deamination activity tested for each such batch on naked DNA and/or chromatin.

Purification of recombinant CseDa01 enzyme

To express the CseDa01 enzyme in *E. coli*, a pQE-30-based (Qiagen) expression vector was constructed, incorporating a T7 promoter-driven, N-terminal 6×His-tagged CseDa01 gene along with a T5 promoter-driven CseDa immunity protein gene. This plasmid was transformed into the *E. coli* expression strain T7 Express lysY/Iq (C3013, NEB). Protein purification was carried out using a denaturation-refolding protocol adapted from Mok et al.[68].

Deamination of naked DNA

Genomic DNA isolation

Naked gDNA was isolated from K562 cells using the NEB Monarch HMW DNA Extraction Kit kit (NEB #T3050L).

Enzymatic treatment

For tests of CseDa01 and LbDa02 activity, 1 μg of purified HMW gDNA was incubated with 1 μL of enzyme for 3 hours at 37 °C in a total volume of 100 μL of deamination buffer[30].

The reaction was stopped by adding 5× volume of PB Buffer and then the Qiagen MinElute kit was used for DNA purification.

DNA concentration was measured using a Qubit 1X dsDNA kit (Thermo Fisher Scientific).

Deaminated gDNA transposition libraries generation

Short-read libraries were prepared using Tn5 transposition as follows. Purified SMF DNA was diluted to a total

amount of 20 ng in a volume of 11.5 μL, then 12.5 μL of 2× TD Buffer and 1 μL Tn5 (produced locally following the protocol described by Picelli et al.[69]) was added, and the mixture was incubated at 55 °C for 5 minutes. The reaction was stopped by adding 250 μL of PB Buffer and the transposition product was purified using the Qiagen MinElute kit, eluting in 20 μL EB Buffer.

PCR amplification was carried out by mixing 20 μL transposed DNA, 2.5 μL i5 and 2.5 μL i7 PCR primers (previously described[70]), and 25 μL 2× Q5U PCR Master Mix (i.e. a uracil-tolerant polymerase), with the following settings:

72 °C for 5 minutes
98 °C for 30 seconds
10 cycles of:
  98 °C for 10 seconds
  63 °C for 30 seconds
  72 °C for 45 seconds
Hold at 4 °C

Final libraries were purified and QC-ed using a Qubit Qubit 1X dsDNA kit (Thermo Fisher Scientific) and an Agilent TapeStation, and sequenced on a NextSeq 550 instrument.

Data processing

Adapters were trimmed from reads using Trimmomatic[71] (version 0.36). Trimmed reads were aligned against the `hg38` version of the human genome assembly using `bwa-meth` with default settings. Duplicate reads were removed using `picard-tools` (version 1.99) if sequencing in a paired-end format. Methylation calls were extracted using `MethylDackel` (`https://github.com/dpryan79/MethylDackel`).

C→T-SMF experiments

Native SMF

For native SMF, 1×10^6 cells were first centrifuged at 1,000 *g* for 2 minutes, then resuspended in 200 μL RSB-Lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630, 0.1% Tween-20, 0.01% Digitonin) and incubated on ice for 3 minutes. Subsequently, 800 μL RSB-wash Buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20) were added, and nuclei were pelleted by centrifugation at 500 *g* for 5 minutes at 4 °C.

Nuclei were then resuspended in 250 μL 1×APOBEC Buffer (from NEB EM-seq kit: NEB #E7120). Deamination treatment was carried out by adding 10 μL concentrated CseDa01, then incubating at 37 °C for 15 minutes.

The reaction was stopped by adding 800 μL PB Buffer and then the Qiagen MinElute kit was used to purify deaminated DNA.

XL-SMF

For crosslinked SMF, cells were first centrifuged at 1,000 *g* for 2 minutes, then resuspended in 1 mL 1×PBS with 27 μL 37% formaldehyde added (to a final concentration of 1%). Crosslinking was carried out for 15 minutes at room temperature on a rotator, then quenched with 100 μL 2.5 M glycine solution. Crosslinked cells were centrifuged at 1,000 *g* for 2 minutes, resuspended in 1 mL 1×PBS, and centrifuged again.

Permeabilization was carried out either by resuspending cells in Lysis Buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA630) and incubating on ice for 10 minutes, then centrifuging for 2 minutes at 1,000 g, or by following the omni-ATAC lysis protocol[70,72].

Cells (200,000 cells were used as input) were then resuspended in 1 mL of 1×APOBEC Buffer (from the NEB EM-seq kit). Deamination treatment was carried out by adding 2 μL RNAse A and 10 μL PURExpress CseDa01 or LbDa01, then incubating at 37 °C for 3 hours with shaking at 800 rpm in a Thermomixer.

For purified CseDa01, 1×10^6 cells and 1 μL of concentrated CseDa01 were used.

After deamination, cells were again pelleted by centrifugation at 1,500 *g* for 2 minutes, then resuspended in 200 μL IP Elution Buffer (1% SDS, 0.1 M $NaHCO_3$); 2 μL Proteinase K were added, and reverse crosslinking was carried out overnight at 65 °C.

DNA was then isolated using the MinElute kit (Qiagen) using 600 μL PB buffer and following the manufacturer's instructions, eluting in 52 μL EB buffer. DNA concentration was measured using the Qubit 1X dsDNA kit (Thermo Fisher Scientific).

Transposition short-read library preparation

Short-read libraries were prepared using Tn5 transposition as described above. Purified SMF DNA was diluted to a total amount of 20 ng in a volume of 11.5 μL, then 12.5 μL of 2× TD Buffer and 1 μL Tn5 were added, and the mixture was incubated at 55 °C for 5 minutes. The reaction was stopped by adding 125 μL of PB Buffer and the transposition product was purified using the Qiagen MinElute kit, eluting in 20 μL EB Buffer.

PCR amplification was carried out by mixing 20 μL transposed DNA, 2.5 μL i5 and 2.5 μL i7 PCR primers, and 25 μL 2× Q5U PCR Master Mix (i.e. a uracil-tolerant polymerase).

Final libraries were purified and QC-ed using a Qubit Qubit 1X dsDNA kit (Thermo Fisher Scientific) and an Agilent TapeStation, and sequenced on a NextSeq 550 instrument to assess global deamination profile properties.

Illumina ssDNA library preparation

For single-stranded libraries, SMF DNA was first sheared using a Covaris E220 instrument to an average size of 400-500 bp.

The xGen Methyl-Seq DNA Library Prep Kit (IDT)) was then used to generated Illumina-compatible libraries, which were sequenced on a sequenced on a NextSeq 550 instrument in a 1×300 bp format. SMF samples were split into four equal subsamples, and each of these samples was converted into a separate library, in order to maximize the complexity of the final dataset.

Ultima library conversion

Illumina libraries were converted into Ultima-compatible libraries using a constant primer (CTGTGTGCCTTGGCAGTCTCAGCTCAGACGTGTGCTCTTCCGATC*T) and an index primer (CCATCTCATCCCTGCGTGTCTCCGACTGCACXXXXXXXXXXXGATCTACACGACGCTCTTCCGATC*T, where the portion indicated as "XXXXXXXXXXX" corresponds to the index sequences). For each library, ∼10 ng were used as input, diluted in a 20 μL volume, then mixed with 2.5 μL of each primer and 25 μL of 2× NEB Next High Fidelity PCR Master Mix (NEB Cat # M0541L). Libraries were amplified using the following settings:

98 °C for 45 seconds
7 cycles of:
  98 °C for 15 seconds
  60 °C for 30 seconds
  72 °C for 60 seconds
72 °C for 10 minutes
Hold at 4 °C

Ultima libraries were sequenced by Ultima Genomics.

Whole-genome DNA amplification

Whole-genome amplification (WGA) of SMF DNA was carried out using the Qiagen REPLI-g kit following manufacturer's instructions.

Targeted amplification

Targeted amplification was carried out starting with 100 ng amplified SMF gDNA diluted in 8 μL total volume with ultrapute H_2O, mixed with 1 μL of forward primer, 1 μL of reverse primer, and 10 μL of 2× Q5U PCR Master Mix, with the following settings:

98 °C for 45 seconds
28 cycles of:
  98 °C for 20 seconds
  T_m for 30 seconds
  72 °C for 5 minutes
72 °C for 10 minutes
Hold at 4 °C

The primer pairs used were GGTAGGATGGAGGAATGA and CTCCCTCTCACCAAAAC for the *IFI6* locus (T_m = 62 °C), GAAGGAYGGAAGGAAGA and CACTTCCCCTRCCTTA for the *USP18* locus (T_m = 60 °C), and GAATGGGTGAAGAGAGAGG and TCCTTTACCTACCCAAATCT for the *IFIT3* locus (T_m = 60 °C).

Amplification was verified using a D5000 TapeStation kit.

Nanopore sequencing

Amplicons were mixed and sequenced using the Oxford Nanopore Technologies Flongle sequencing platform (FLG-114 with the SQK-LSK114 kit).

C→T-SMF data processing

Data processing of short-read libraries

Short read SMF sequencing data was processed as previously described[73] with the modification that accessible/protected C positions were flipped (while in the EM-seq and bisulfite readouts used in cytosine methylation-based SMF unmethylated Cs are converted, in C→T-SMF base conversion is direct).

Briefly, adapters were trimmed from reads using Trimmomatic[71] (version 0.36). Trimmed reads were aligned against the `hg38` version of the human genome using `bwa-meth` with default settings. Duplicate reads were removed using `picard-tools` (version 1.99) if sequencing in a paired-end format. Methylation calls were extracted using `MethylDackel` (`https://github.com/dpryan79/MethylDackel`). Additional analyses and visualization were carried out using custom-written Python scripts (`https://github.com/georgimarinov/GeorgiScripts`) as previously described[73].

Nanopore sequencing data processing

Nanopore reads were mapped against the `hg38` version of the human genome using `minimap2`[74].

Additional analyses and visualization were carried out using custom-written Python scripts (`https://github.com/georgimarinov/GeorgiScripts`) as previously described[36].

NOMe-seq/GpC-SMF

NOMe-seq/GpC-SMF experiments were carried out as previously described[73] with some modifications.

Briefly, 1×10^6 cells were pelleted by centrifugation, resuspended in 1 mL 1× PBS and pelleted again, then resuspended in 200 μL RSB-Lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630, 0.1% Tween-20, 0.01% Digitonin) and incubated on ice for 3 minutes. Then 1200 μL RSB-wash Buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20) were added, and nuclei were pelleted by centrifugation at 500 g for 10 minutes at 4 °C. Nuclei were then resuspended in 100 μL M. CviPI Reaction Buffer and 200 U of M.CviPI GpC MTase (NEB) were added, together with SAM (S-adenosylmethionine) at a final concentration of 0.6 mM. The reaction was incubated at 37 °C for 7.5 minutes

in a Thermomixer at 1,000 rpm, then another 100 U of enzyme and more SAM were added and the incubation was continued for another 7.5 minutes.

The reaction was stopped by adding 175 μL of the Cell Lysis Buffer from the NEB Monarch gDNA extraction kit together with 3 μL RNase A and 1 μL Proteinase K, and gDNA purification was carried out following the instructions for the NEB Monarch kit.

Final sequencing libraries were prepared on Covaris-sheared DNA using the NEB EM-seq kit, and sequenced as 2×150mers on an Illumina NovaSeq instrument (through Novogene).

NOMe-seq/GpC-SMF data processing

NOMe-seq/GpC-SMF sequencing data was processed as previously described[73]. Briefly, adapters were trimmed from reads using Trimmomatic[71] (version 0.36). Trimmed reads were aligned against the `hg38` version of the human genome using `bwa-meth` with default settings. Duplicate reads were removed using `picard-tools` (version 1.99) if sequencing in a paired-end format. Methylation calls were extracted using `MethylDackel` (`https://github.com/dpryan79/MethylDackel`). Additional analyses and visualization were carried out using custom-written Python scripts (`https://github.com/georgimarinov/GeorgiScripts`) as previously described[73].

ATAC-seq experiments

ATAC-seq experiments were carried out as previously described[70].

Briefly, 5×10^4 cells were pelleted by centrifugation, resuspended in 200 μL 1× PBS and pelleted again, then resuspended in 50 μL RSB-Lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630, 0.1% Tween-20, 0.01% Digitonin) and incubated on ice for 3 minutes. Then 450 μL RSB-wash Buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20) were added, and nuclei were pelleted by centrifugation at 500 g for 5 minutes at 4 °C.

Tagmentation of chromatin followed by resuspending nuclei with 25 μL 2× TD buffer (20 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 20% Dimethyl Formamide), 2.5 μL transposase (custom produced[69]) and 22.5 μL nuclease-free H_2O, and incubating at 37 °C for 30 min in a Thermomixer at 1000 RPM. Transposed DNA was isolated using the Qiagen MinElute kit and PCR amplified as descrbed above using the following procedure:

- 72 °C for 5 minutes
- 98 °C for 30 seconds
- 10 cycles of:
  - 98 °C for 10 seconds
  - 63 °C for 30 seconds
  - 72 °C for 30 seconds
- Hold at 4 °C

Libraries were then sequenced on an Illumina NextSeq instrument as 2×36mers.

ATAC-seq data processing

ATAC-seq data processing was carried out as previously described[75]. Briefly, demultipexed fastq files were mapped to the `hg38` assembly of the human genome as 2×36mers using Bowtie[76] with the following settings: `-v 2 -k 2 -m 1 --best --strata`. Duplicate reads were removed using `picard-tools` (version 1.99). Peak calling was carried out using MACS2[77] as previously described[75]. Additional analyses and visualization were carried out using custom-written Python scripts (`https://github.com/georgimarinov/GeorgiScripts`).

Transcription factor motif mapping

Motif annotations were obtained form the CIS-BP database[78] (version 1.0.2). Motifs were mapped against the `hg38` version of the human genome assembly using the FIMO program[79] in the MEME-SUITE of motif analysis tools[80].

Co-accessibility assessment

Co-accessibility was measured between each pair of informative positions using the normalized mutual information (NMI) metric previously described in Shipony & Marinov et al.[26].

External sequencing datasets

Additional datasets and annotations were obtained from the ENCODE portal as indicated in the corresponding figures and figure legends.

Author contributions

G.K.M. conceived project, designed and carried out experiments and data analysis and wrote the manuscript with input from all authors. T.W. produced deaminase enzymes. M.M.S., M.C. and Z.S provided key reagents. J.M.S. and B.D. designed and carried out experiments. W.J.G. and A.K. supervised the project.

Acknowledgments

This work was supported by NIH grants U01HG011762, R01 NS128028, DP1HG013599, R01HL171611, and R01HG013317 to W.J.G.

References

1. ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J,

Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kelllis M, Khatun J, Kheradpour P, Kundaje A, Lassman T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SC, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LA, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elinitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Starnatoyannopoulos JA, Tennebaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandu KS, Schaeffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee BK, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Kim SK, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge EC, Trout D, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, King B, Muratet MA, Antoshechkin I, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Gunter C, Newberry JS, Levy SE, Absher DM, Mortazavi A, Wong WH, Wold B, Blow MJ, Visel A, Pennachio LA, Elnitski L, Margulies EH, Parker SC, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Chrast J, Davidson C, Derrien T, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Howald C, Hunt T, Jungreis I, Kay M, Khurana E, Kokocinski F, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tanzer A, Tapanan E, Tress ML, van Baren MJ, Walters N, Washieti S, Wilming L, Zadissa A, Zhengdong Z, Brent M, Haussler D, Kellis M, Valencia A, Gerstein M, Raymond A, Guigó R, Harrow J, Hubbard TJ, Landt SG, Frietze S, Abyzov A, Addleman N, Alexander RP, Auerbach RK, Balasubramanian S, Bettinger K, Bhardwaj N, Boyle AP, Cao AR, Cayting P, Charos A, Cheng Y, Cheng C, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyenger S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Larnarre-Vincent N, Leng J, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Pei B, Raha D, Ramirez L, Reed B, Rozowsky J, Sboner A, Shi M, Sisu C, Slifer T, Witt H, Wu L, Xu X, Yan KK, Yang X, Yip KY, Zhang Z, Struhl K, Weissman SM, Gerstein M, Farnham PJ, Snyder M, Tenebaum SA, Penalva LO, Doyle F, Karmakar S, Landt SG, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Patacsil D, Slifer T, Victorsen A, Yang X, Snyder M, White KP, Auer T, Centarin L, Eichenlaub M, Gruhl F, Heerman S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Weng Z, Whitfield TW, Wang J, Collins PJ, Aldred SF, Trinklein ND, Partridge EC, Myers RM, Dekker J, Jain G, Lajoie BR, Sanyal A, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Ebersol AK, Frum T, Garg K, Gist E, Hansen RS, Boatman L, Haugen E, Humbert R, Jain G, Johnson AK, Johnson EM, Kutyavin TM, Lajoie BR, Lee K, Lotakis D, Maurano MT, Neph SJ, Neri FV, Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sabo P, Sanchez ME, Sandstrom RS, Sanyal

A, Shafer AO, Stergachis AB, Thomas S, Thurman RE, Vernot B, Vierstra J, Vong S, Wang H, Weaver MA, Yan Y, Zhang M, Akey JA, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Kaul R, Dekker J, Stamatoyannopoulos JA, Dunham I, Beal K, Brazma A, Flicek P, Herrero J, Johnson N, Keefe D, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Wilder SP, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Kundaje A, Hardison RC, Miller W, Giardine B, Harris RS, Wu W, Bickel PJ, Banfai B, Boley NP, Brown JB, Huang H, Li Q, Li JJ, Noble WS, Bilmes JA, Buske OJ, Hoffman MM, Sahu AO, Kharchenko PV, Park PJ, Baker D, Taylor J, Weng Z, Iyer S, Dong X, Greven M, Lin X, Wang J, Xi HS, Zhuang J, Gerstein M, Alexander RP, Balasubramanian S, Cheng C, Harmanci A, Lochovsky L, Min R, Mu XJ, Rozowsky J, Yan KK, Yip KY, Birney E. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414):57–74.

2. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, Mortazavi A, Keller CA, Zhang XO, Elhajjajy SI, Huey J, Dickel DE, Snetkova V, Wei X, Wang X, Rivera-Mulia JC, Rozowsky J, Zhang J, Chhetri SB, Zhang J, Victorsen A, White KP, Visel A, Yeo GW, Burge CB, Lécuyer E, Gilbert DM, Dekker J, Rinn J, Mendenhall EM, Ecker JR, Kellis M, Klein RJ, Noble WS, Kundaje A, Guigó R, Farnham PJ, Cherry JM, Myers RM, Ren B, Graveley BR, Gerstein MB, Pennacchio LA, Snyder MP, Bernstein BE, Wold B, Hardison RC, Gingeras TR, Stamatoyannopoulos JA, Weng Z. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**(7818):699–710.
3. Wu C. 1980. The 5′ ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**(5776):854–860.
4. Keene MA, Corces V, Lowenhaupt K, Elgin SC. 1981. DNase I hypersensitive sites in Drosophila chromatin occur at the 5′ ends of regions of transcription. *Proc Natl Acad Sci U S A* **78**(1):143–146.
5. McGhee JD, Wood WI, Dolan M, Engel JD, Felsenfeld G. 1981. A 200 base pair region at the 5′ end of the chicken adult β-globin gene is accessible to nuclease digestion. *Cell* **27**(1 Pt 2):45–55.
6. Dorschner MO, Hawrylycz M, Humbert R, Wallace JC, Shafer A, Kawamoto J, Mack J, Hall R, Goldy J, Sabo PJ, Kohli A, Li Q, McArthur M, Stamatoyannopoulos JA. 2004. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods* **1**(3):219–225.
7. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, Green RD, Navas PA, Noble WS, Stamatoyannopoulos JA. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* **3**(7):511–518.
8. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2):311–322.
9. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6**(4):283–289.
10. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**(12):1213–1218.
11. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutyavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, Stamatoyannopoulos JA. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**(7414):83–90.
12. Gottschling DE. 1992. Telomere-proximal DNA in *Saccharomyces cerevisiae* is refractory to methyltransferase activity in vivo. *Proc Natl Acad Sci U S A.* **89**(9):4062-5(9):4062–4065.
13. Singh J, Klar AJ. 1992. Active genes in budding yeast display enhanced in vivo accessibility to foreign DNA methylases: a novel in vivo probe for chromatin structure of yeast. *Genes Dev* **6**(2):186–196.
14. Kladde MP, Xu M, Simpson RT. 1996. Direct study of DNA-protein interactions in repressed and active chromatin in living cells. *EMBO J* **15**(22):6290–6300
15. Xu M, Kladde MP, Van Etten JL, Simpson RT. 1998. Cloning, characterization and expression of the gene coding for a cytosine-5-DNA methyltransferase recognizing GpC. *Nucleic Acids Res* **26**(17):3961–3966.
16. Jessen WJ, Dhasarathy A, Hoose SA, Carvin CD, Risinger AL, Kladde MP. 2004. Mapping chromatin structure in vivo using DNA methyltransferases. *Methods* **33**(1):68–80.
17. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. 2012. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**(12):2497–2506.
18. Nabilsi NH, Deleyrolle LP, Darst RP, Riva A,

Reynolds BA, Kladde MP. 2014. Multiplex mapping of chromatin accessibility and DNA methylation within targeted single molecules identifies epigenetic heterogeneity in neural stem cells and glioblastoma. *Genome Res* **24**(2):329–339.

19. Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. 2017. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell* **67**(3):411–422.e4.
20. Miura F, Miura M, Shibata Y, Furuta Y, Miyamura K, Ino Y, Bayoumi AMA, Oba U, Ito T. 2022. Identification, expression, and purification of DNA cytosine 5-methyltransferases with short recognition sequences. *BMC Biotechnol* **22**(1):33.
21. Sun Z, Vaisvila R, Hussong LM, Yan B, Baum C, Saleh L, Samaranayake M, Guan S, Dai N, Corrêa IR Jr, Pradhan S, Davis TB, Evans TC Jr, Ettwiller LM. 2021. Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res* **31**(2):291–300.
22. Vaisvila R, Ponnaluri VKC, Sun Z, Langhorst BW, Saleh L, Guan S, Dai N, Campbell MA, Sexton BS, Marks K, Samaranayake M, Samuelson JC, Church HE, Tamanaha E, Corrêa IR Jr, Pradhan S, Dimalanta ET, Evans TC Jr, Williams L, Davis TB. 2021. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res* **31**(7):1280–1289.
23. Doughty BR, Hinks MM, Schaepe JM, Marinov GK, Thurm AR, Rios-Martinez C, Parks BE, Tan Y, Marklund E, Dubocanin D, Bintu L, Greenleaf WJ. 2024. Single-molecule states link transcription factor binding to gene expression. *Nature* **636**(8043):745–754.
24. Schaepe JM, Fries T, Doughty BR, Crocker OJ, Hinks MM, Marklund E, Greenleaf WJ. 2025. Thermodynamic principles link in vitro transcription factor affinities to single-molecule chromatin states in cells. *bioRxiv* 2025.01.27.635162.
25. Murray IA, Morgan RD, Luyten Y, Fomenkov A, Corrêa IR Jr, Dai N, Allaw MB, Zhang X, Cheng X, Roberts RJ. 2018. The non-specific adenine DNA methyltransferase M.EcoGII. *Nucleic Acids Res* **46**(2):840–848.
26. Shipony Z, Marinov GK, Swaffer MP, Sinott–Armstrong NA, Skotheim JM, Kundaje A, Greenleaf WJ. 2020. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods* **17**:319–327.
27. Drozdz M, Piekarowicz A, Bujnicki JM, Radlinska M. 2012. Novel non-specific DNA adenine methyltransferases. *Nucleic Acids Res* **40**(5):2119–2130.
28. Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. 2020. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**(6498):1449–1454.
29. Salter JD, Bennett RP, Smith HC. 2016. The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends Biochem Sci* **41**(7):578–594.
30. Vaisvila R, Johnson SR, Yan B, Dai N, Bourkia BM, Chen M, Corrêa IR Jr, Yigit E, Sun Z. 2024. Discovery of cytosine deaminases enables base-resolution methylome mapping using a single enzyme. *Mol Cell* **84**(5):854–866.e7.
31. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. 2015. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**(7583):575–579.
32. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE. 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* **107**(19):8689–8694.
33. Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**(5980):916–919.
34. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, Timp W. 2020. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat Methods* **17**(12):1191–1199.
35. Wang Y, Wang A, Liu Z, Thurman AL, Powers LS, Zou M, Zhao Y, Hefel A, Li Y, Zabner J, Au KF. 2019. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res* **29**(8):1329–1342.
36. Marinov GK, Shipony Z, Kundaje A, Greenleaf WJ. 2022. Single-Molecule Multikilobase-Scale Profiling of Chromatin Accessibility Using m6A-SMAC-Seq and m6A-CpG-GpC-SMAC-Seq. *Methods Mol Biol* **2458**:269–298.
37. Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, Karimzadeh M, Underwood JG, Goodarzi H, Narlikar GJ, Ramani V. 2020. Massively multiplex single-molecule oligonucleosome footprinting. *Elife* **9**:e5940
38. Leduque B, Edera A, Vitte C, Quadrana L. 2024. Simultaneous profiling of chromatin accessibility and DNA methylation in complete plant genomes using long-read sequencing. *Nucleic Acids Res* **52**(11):6285–6297.
39. Chanou A, WeißM, Holler K, Sajid A, Straub T, Krietsch J, Sanchi A, Ummethum H, Lee CSK, Kruse E, Trauner M, Werner M, Lalonde M, Lopes M, Scialdone A, Hamperl S. 2023. Single molecule MATAC-seq reveals key determinants of DNA replication origin efficiency. *Nucleic Acids Res* **51**(22):12303–12324.

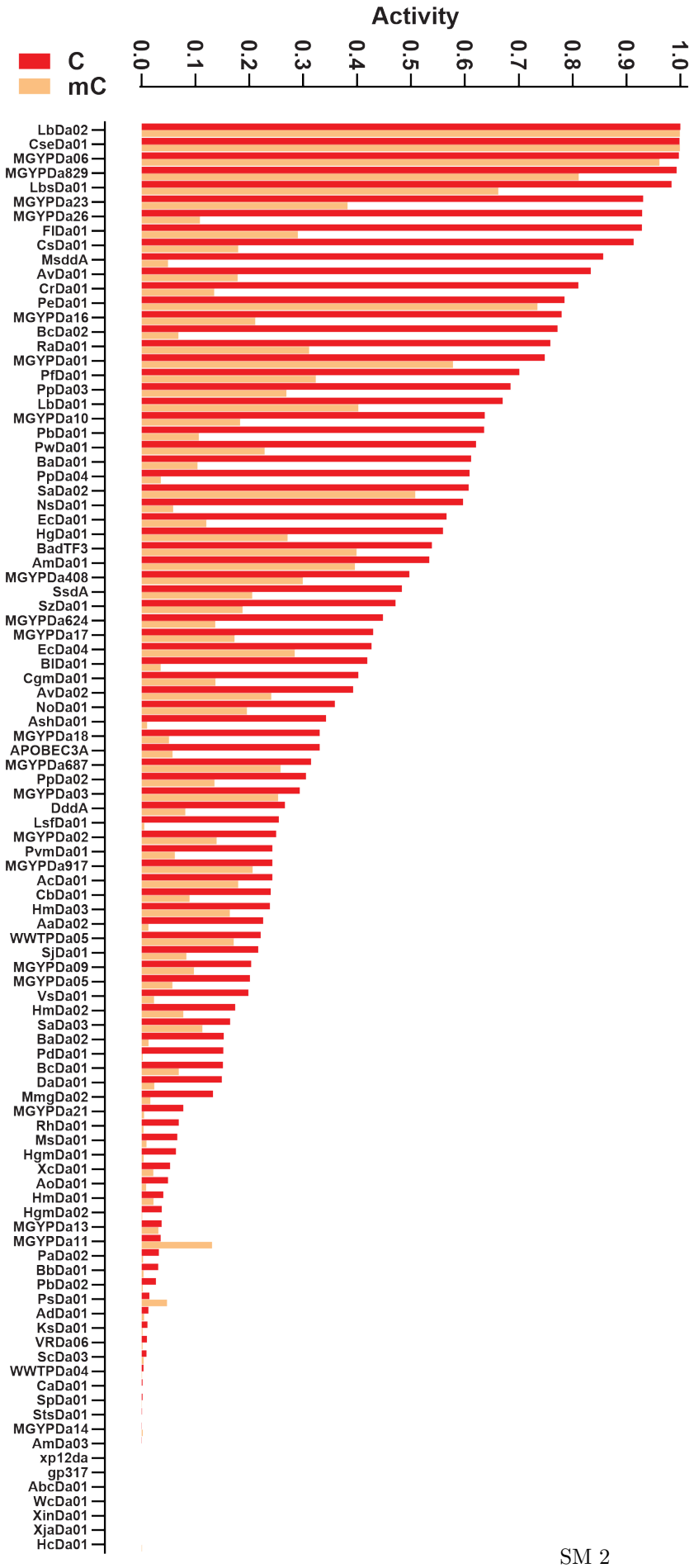
40. Mo W, Shu Y, Liu B, Long Y, Li T, Cao X, Deng X, Zhai J. 2023. Single-molecule targeted accessibility and methylation sequencing of centromeres, telomeres and rDNAs in *Arabidopsis*. *Nat Plants*
41. He R, Dong W, Wang Z, Xie C, Gao L, Ma W, Shen K, Li D, Pang Y, Jian F, Zhang J, Yuan Y, Wang X, Zhang Z, Zheng Y, Liu S, Luo C, Chai X, Ren J, Zhu Z, Xie XS. 2024. Genome-wide single-cell and single-molecule footprinting of transcription factors with deaminase. *Proc Natl Acad Sci U S A* **121**(52):e2423270121.
42. Yu T, Li Z, Gibbs E, Iwase R, Francoeur MJ, Phan QV, Zhao J, Rosin J, Cole PA, Pinello L, Sherwood RI. 2024. Deaminase-mediated chromatin accessibility profiling with single-allele resolution. *bioRxiv* 2024.12.17.628768.
43. Roh H, Shen SP, Hu Y, Kwok HS, Siegenfeld AP, Lee C, Zepeda MA, Guo CJ, Roseman SA, Sankaran VJ, Buenrostro JD, Liau BB. 2024. Coupling CRISPR Scanning with Targeted Chromatin Accessibility Profiling using a Double-Stranded DNA Deaminase. *bioRxiv* 2024.12.17.628791.
44. Swanson EG, Mao Y, Mallory BJ, Vollger MR, Ranchalis J, Bohaczuk SC, Parmalee NL, Bennett JT, Stergachis AB. 2024. Deaminase-assisted single-molecule and single-cell chromatin fiber sequencing. *bioRxiv* 2024.11.06.622310
45. Fu Y, Sinha M, Peterson CL, Weng Z. 2008. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4**(7):e1000138
46. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**(1):24–32.
47. Brogaard K, Xi L, Wang JP, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**(7404):496–501.
48. Fourier N, Zolty M, Azriel A, Tedesco D, Levi BZ. 2020. MafK Mediates Chromatin Remodeling to Silence IRF8 Expression in Non-immune Cells in a Cell Type-Specific Manner. *J Mol Biol* **432**(16):4544–4560.
49. Moyers BA, Partridge EC, Mackiewicz M, Betti MJ, Darji R, Meadows SK, Newberry KM, Brandsmeier LA, Wold BJ, Mendenhall EM, Myers RM. 2023. Characterization of human transcription factor function and patterns of gene regulation in HepG2 cells. *Genome Res* **33**(11):1879–1892.
50. Schoenherr CJ, Anderson DJ. 1995. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267**(5202):1360–1363.
51. Grand RS, Pregnolato M, Baumgartner L, Hoerner L, Burger L, Schübeler D. 2024. Genome access is transcription factor-specific and defined by nucleosome position. *Mol Cell* **84**(18):3455–3468.e6.
52. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. *Nature* **474**(7352):516–520.
53. Franza BR Jr, Rauscher FJ 3rd, Josephs SF, Curran T. 1988. The Fos complex and Fos-related antigens recognize sequence elements that contain AP-1 binding sites. *Science* **239**(4844):1150–1153.
54. Reinke AW, Baek J, Ashenberg O, Keating AE. 2013. Networks of bZIP protein-protein interactions diversified over a billion years of evolution. *Science* **340**(6133):730–734.
55. Jain J, McCaffrey PG, Valge-Archer VE, Rao A. 1992. Nuclear factor of activated T cells contains Fos and Jun. *Nature* **356**(6372):801—804.
56. Corre S, Galibert MD. 2005. Upstream stimulating factors: highly versatile stress-responsive transcription factors. *Pigment Cell Res* **18**(5):337–348
57. Viollet B, Lefrancois-Martinez AM, Henrion A, Kahn A, Raymondjean M, Martinez A. 1996. Immunochemical characterization and transacting properties of upstream stimulatory factor isoforms. *J Biol Chem* **271**(3):1405–1415.
58. Romier C, Cocchiarella F, Mantovani R, Moras D. 2003. The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. *J Biol Chem* **278**(2):1336–1345.
59. Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S, Jothi R. 2014. Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. *Mol Cell* **55**(5):708–722.
60. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, Guo YE, Hnisz D, Jaenisch R, Bradner JE, Gray NS, Young RA. 2017. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**(7):1573–1588.e28.
61. Liu K, Li W, Xiao Y, Lei M, Zhang M, Min J. 2024. Molecular mechanism of specific DNA sequence recognition by NRF1. *Nucleic Acids Res* **52**(2):953–966.
62. Rosmarin AG, Resendes KK, Yang Z, McMillan JN, Fleming SL. 2004. GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions. *Blood Cells Mol Dis* **32**(1):143–154.
63. Neuteboom ST, Murre C. 1997. Pbx raises the DNA binding specificity but not the selectivity of antennapedia Hox proteins. *Mol Cell Biol* **17**(8):4696–4706.
64. LaRonde-LeBlanc NA, Wolberger C. 2003. Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. *Genes Dev* **17**(16):2060–2072.
65. Li Z, Zhang Z, Li Y, Arnovitz S, Chen P, Huang H,

Jiang X, Hong GM, Kunjamma RB, Ren H, He C, Wang CZ, Elkahloun AG, Valk PJ, Döhner K, Neilly MB, Bullinger L, Delwel R, Löwenberg B, Liu PP, Morgan R, Rowley JD, Yuan CS, Chen J. 2013. PBX3 is an important cofactor of HOXA9 in leukemogenesis. *Blood* **121**(8):1422–1431

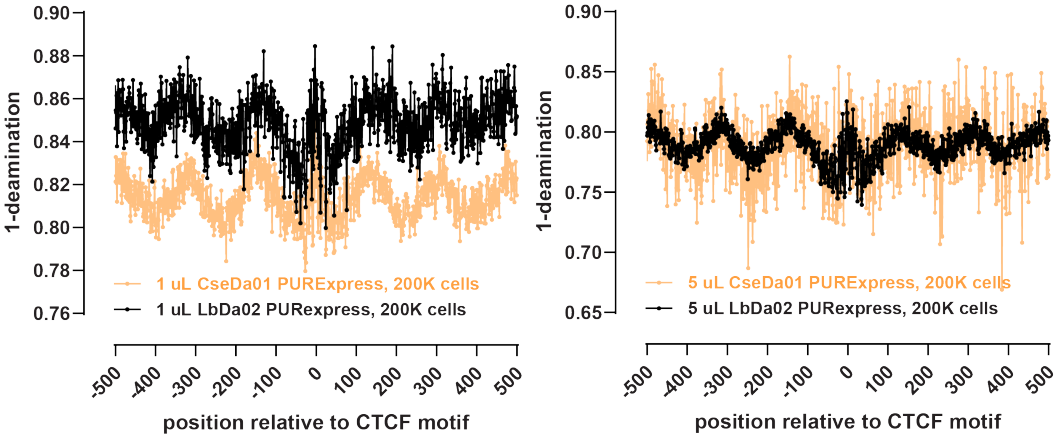
66. Steube A, Schenk T, Tretyakov A, Saluz HP. 2017. High-intensity UV laser ChIP-seq for the study of protein-DNA interactions in living cells. *Nat Commun* **8**(1):1303
67. Nagaich AK, Walker DA, Wolford R, Hager GL. 2004. Rapid periodic binding and displacement of the glucocorticoid receptor during chromatin remodeling. *Mol Cell* **14**(2):163–174.
68. Mok BY, de Moraes MH, Zeng J, Bosch DE, Kotrys AV, Raguram A, Hsu F, Radey MC, Peterson SB, Mootha VK, Mougous JD, Liu DR. 2020. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* **583**(7817):631–637.
69. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* **24**(12):2033–2040.
70. Marinov GK, Shipony Z, Kundaje A, Greenleaf WJ. 2023. Genome-Wide Mapping of Active Regulatory Elements Using ATAC-seq. *Methods Mol Biol* **2611**:3–19.
71. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15):2114–2120.
72. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA, Risca VI, Kundaje A, Khavari PA, Montine TJ, Greenleaf WJ, Chang HY. 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**(10):959–962.
73. Hinks M, Marinov GK, Kundaje A, Bintu L, Greenleaf WJ. 2023. Single-Molecule Mapping of Chromatin Accessibility Using NOMe-seq/dSMF. *Methods Mol Biol* **2611**:101–119.
74. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18):3094–3100.
75. Marinov GK, Shipony Z. 2021. Interrogating the Accessible Chromatin Landscape of Eukaryote Genomes Using ATAC-seq. *Methods Mol Biol* **2243**:183–226.
76. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3):R25.
77. Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**(9):1728–1740.
78. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJ, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**(6):1431–1443.
79. Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**(7):1017–1018.
80. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**(Web Server issue):W202–208.

Supplementary Materials

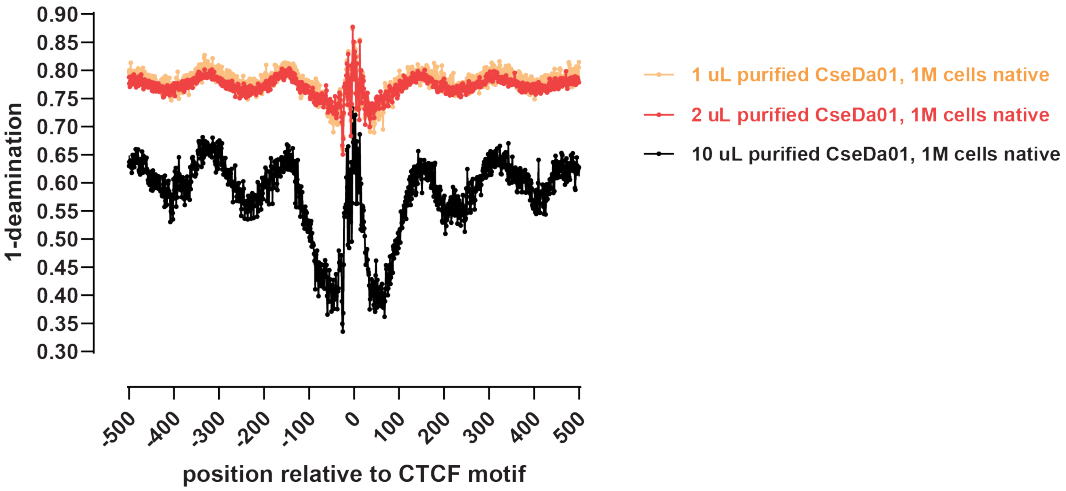
Supplementary Figures



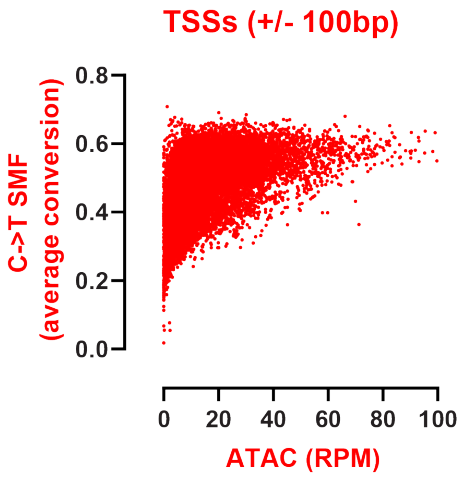
Supplementary Figure 1: Relative activity of various deaminases against unmethylated and methylated ds-DNA substrates. Data was obtained from Vaisvila et al. 2024[30].



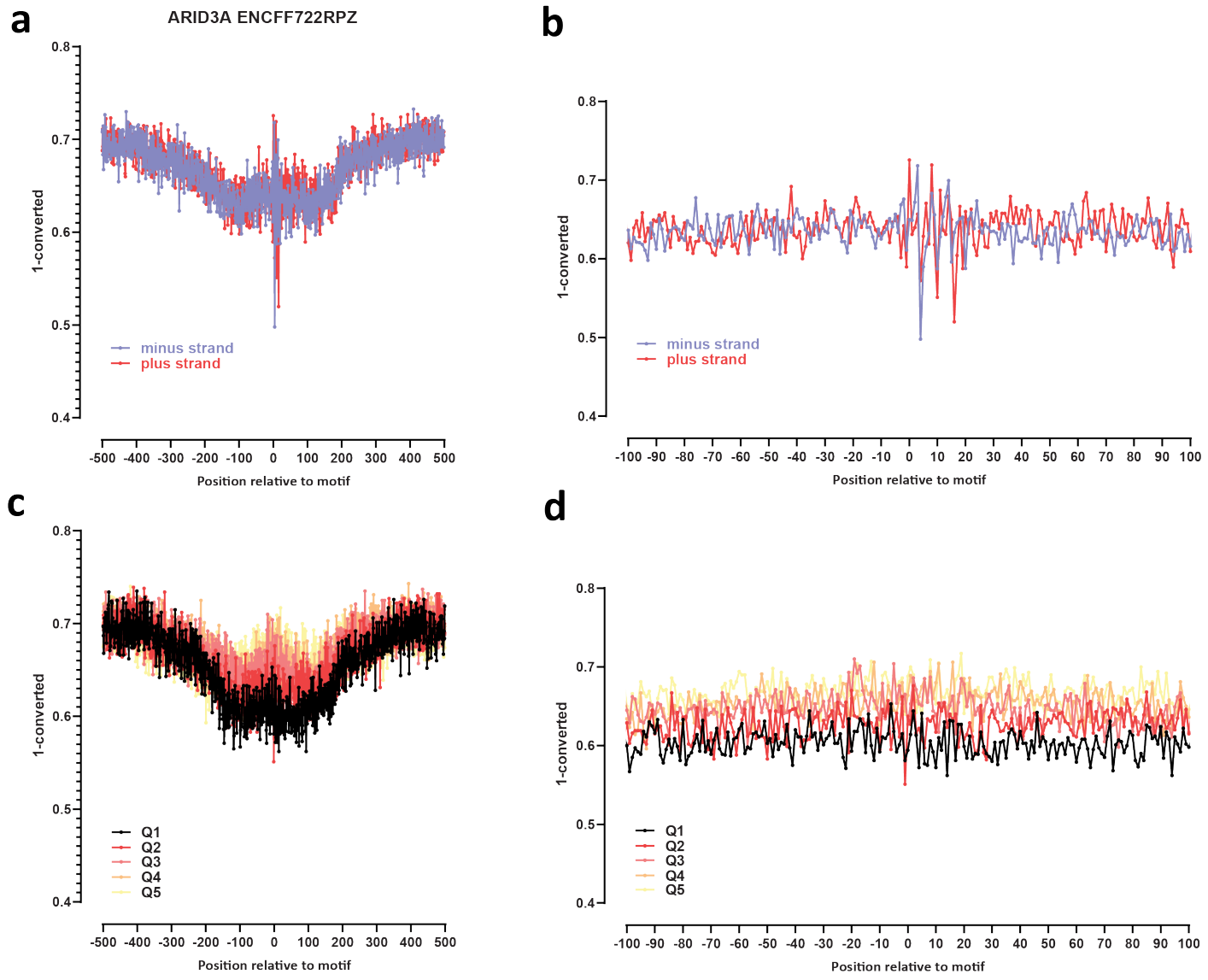
Supplementary Figure 2: High-resolution chromatin accessibility profiling using the non-specific CseDa01 and LbDa02 dsDNA deaminases. Shown are metaprofiles around CTCF motifs within CTCF ChIP-seq peaks for experiments carried out with PURExpress CseDa01 preps or PURExpress LbDa02 and the indicated amount of enzyme and number of cells (K562 cells, crosslinked).



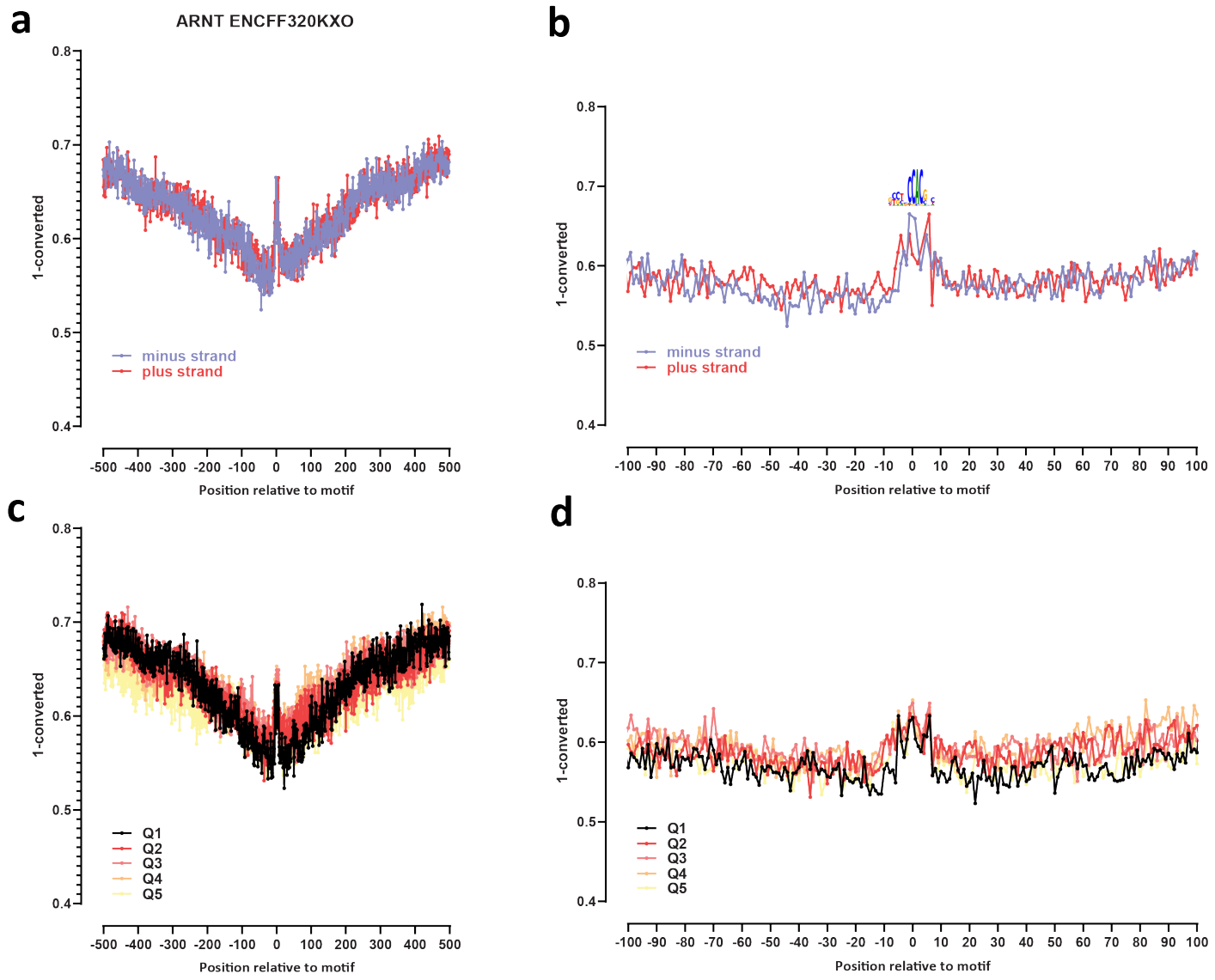
Supplementary Figure 3: Native high-resolution chromatin accessibility profiling using the non-specific CseDa01 dsDNA deaminase. Shown are metaprofiles around CTCF motifs within CTCF ChIP-seq peaks for experiments carried out with purified CseDa01 and the indicated amount of enzyme and number of cells (K562 cells, native).



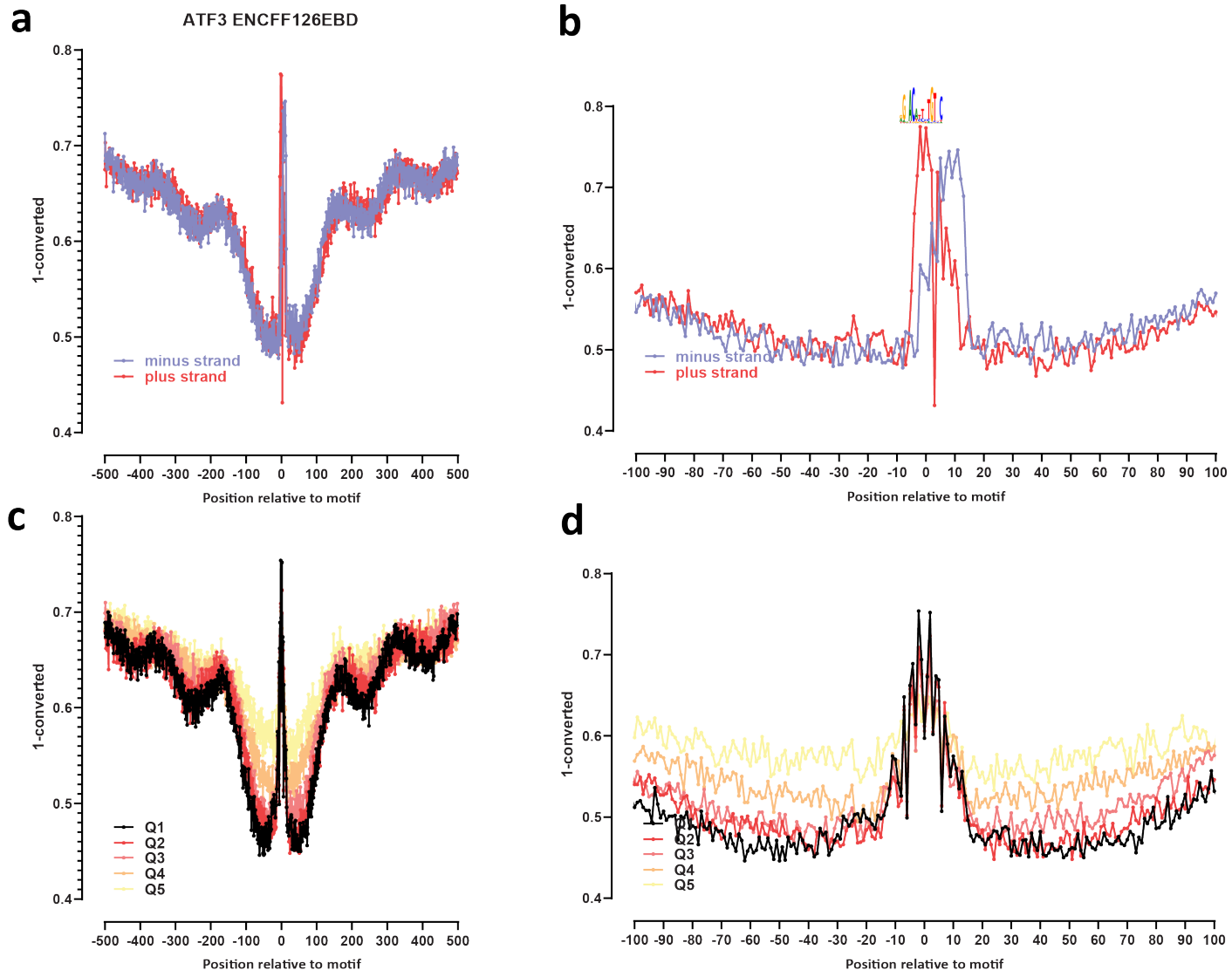
Supplementary Figure 4: Correspondence between short-read CseDa01 C→T-SMF and ATAC-seq genome-wide profiles. Shown are RPKM values over annotated RefSeq TSSs (±100 bp) and average deamination rates for the same windows.



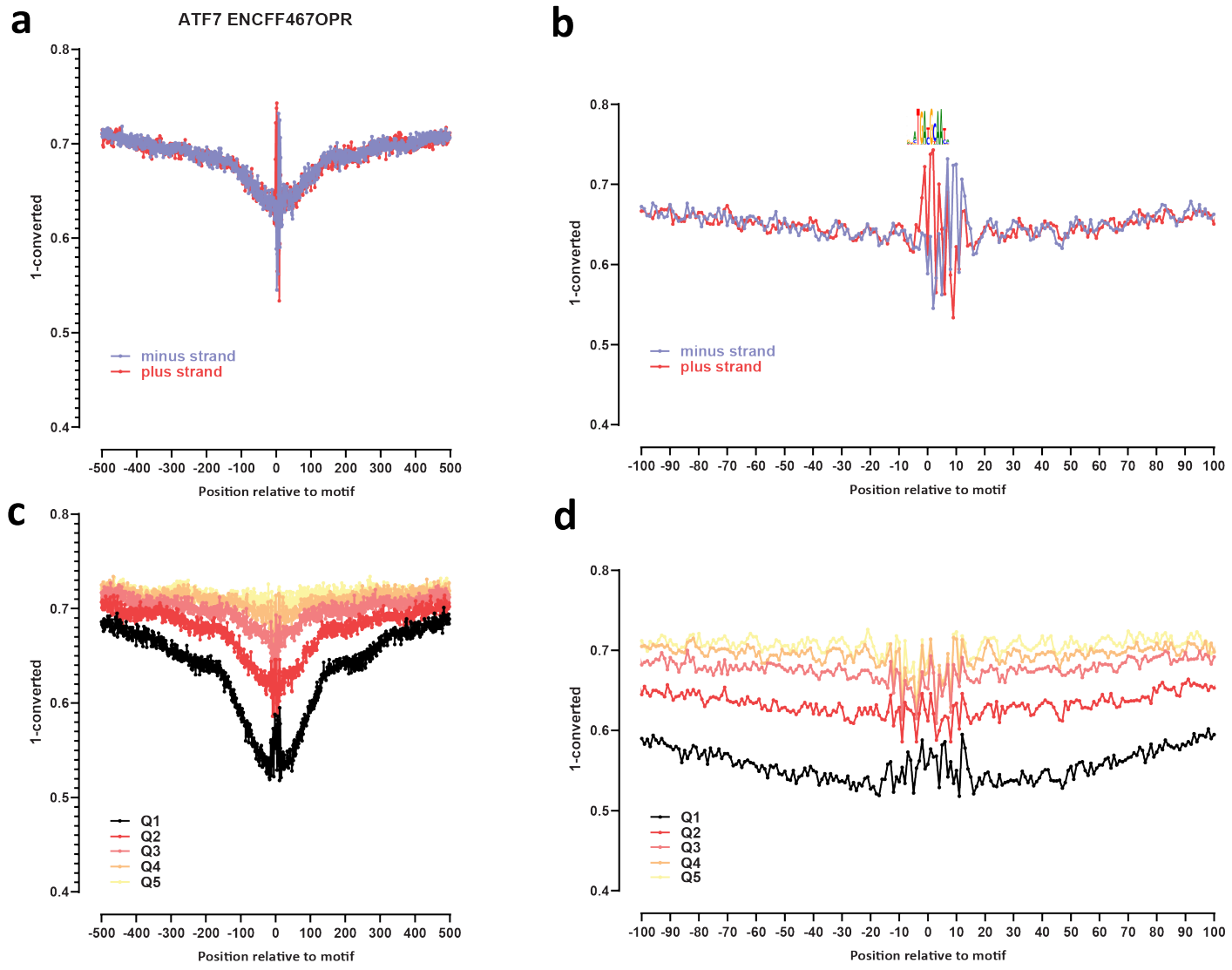
Supplementary Figure 5: Genome-wide footprint profiles for the ARID3A transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



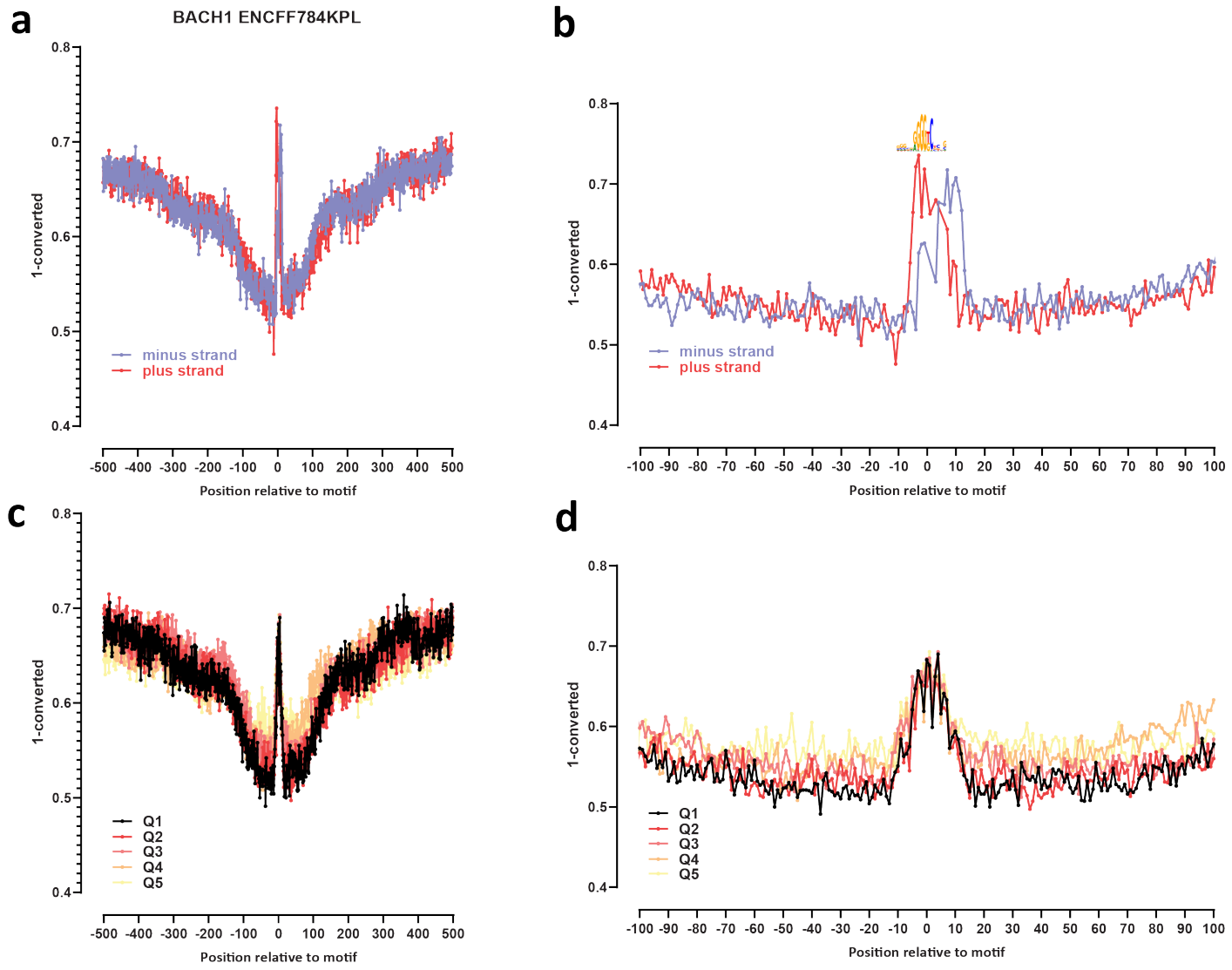
Supplementary Figure 6: Genome-wide footprint profiles for the ARNT transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



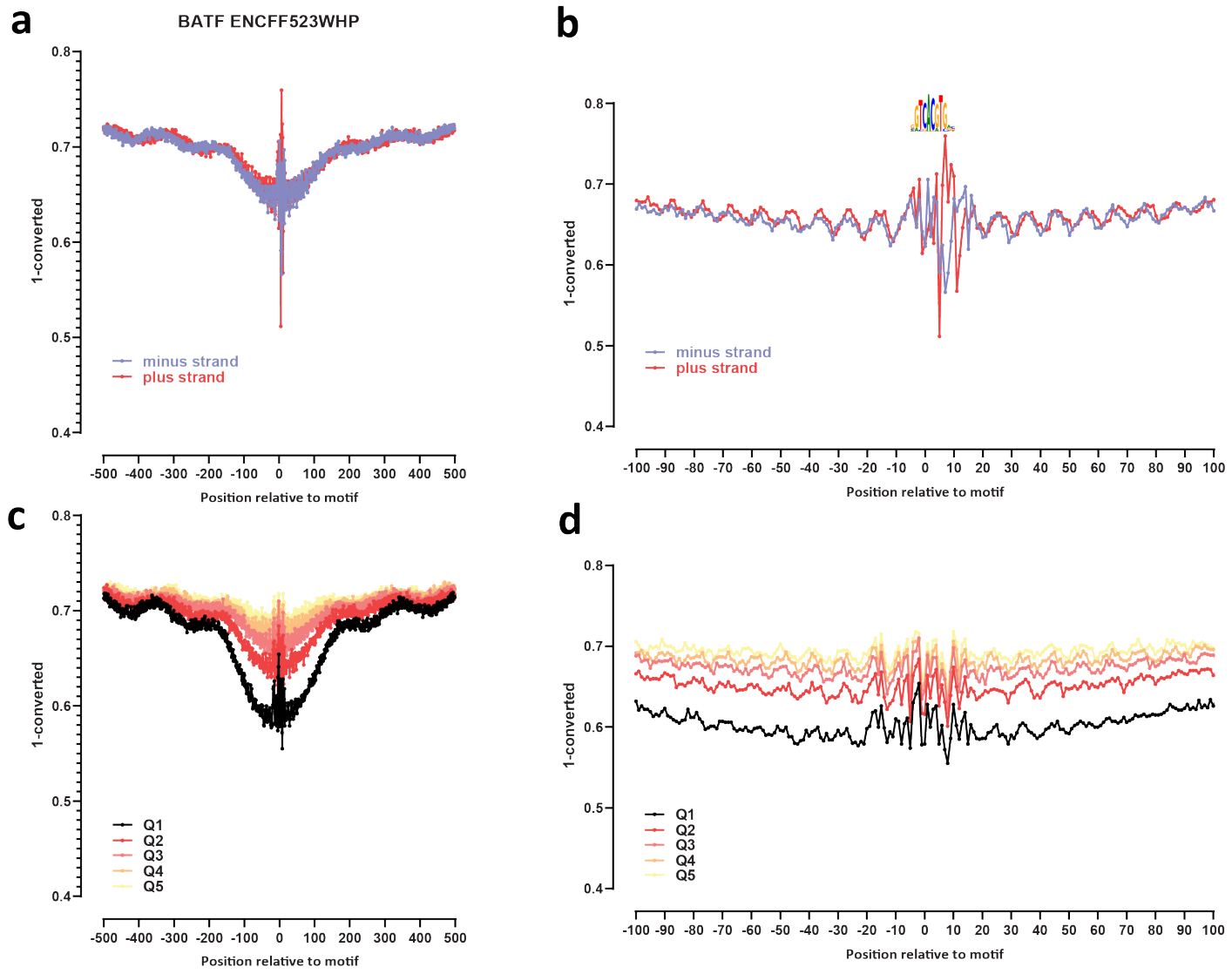
Supplementary Figure 7: Genome-wide footprint profiles for the ATF3 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



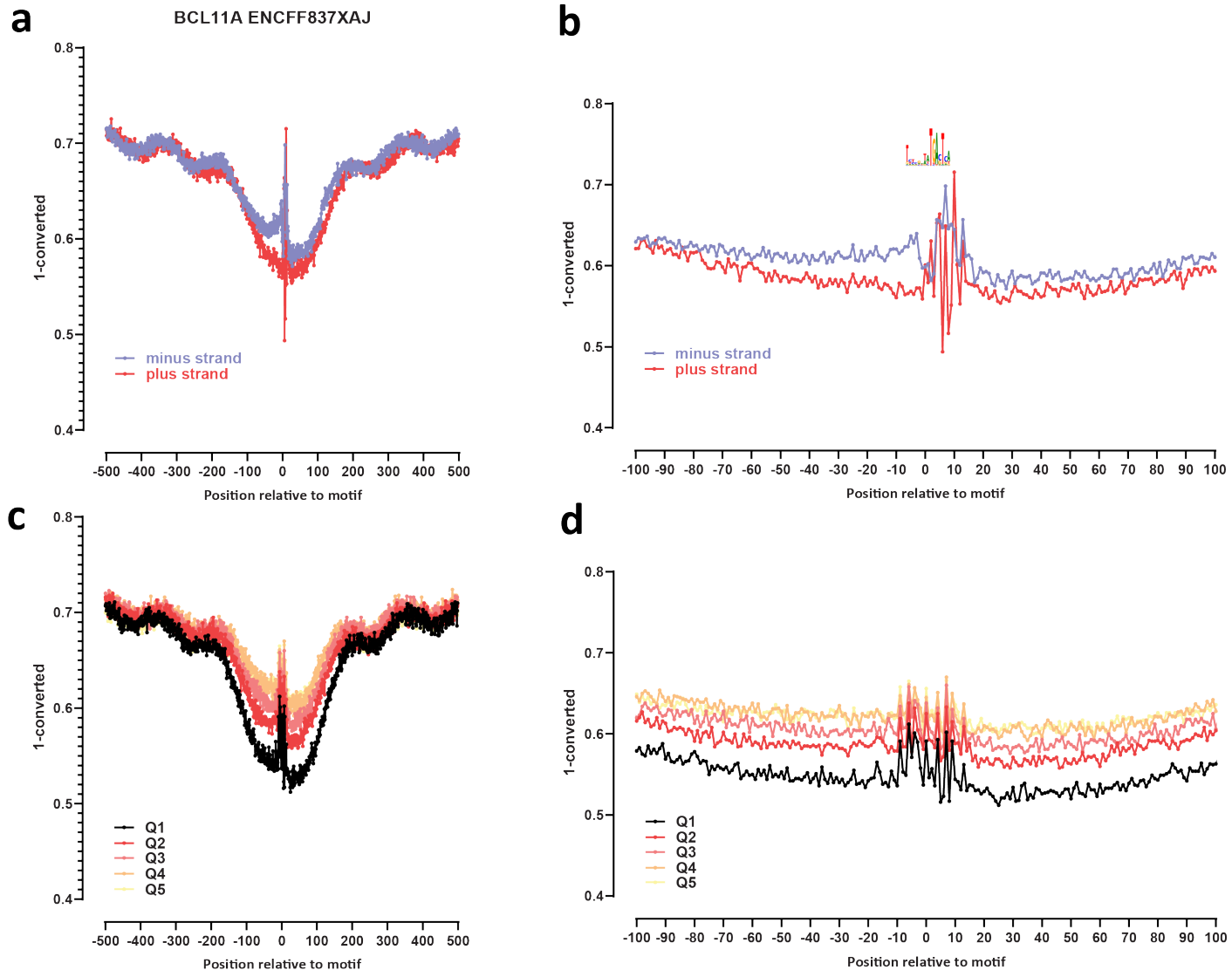
Supplementary Figure 8: Genome-wide footprint profiles for the ATF7 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



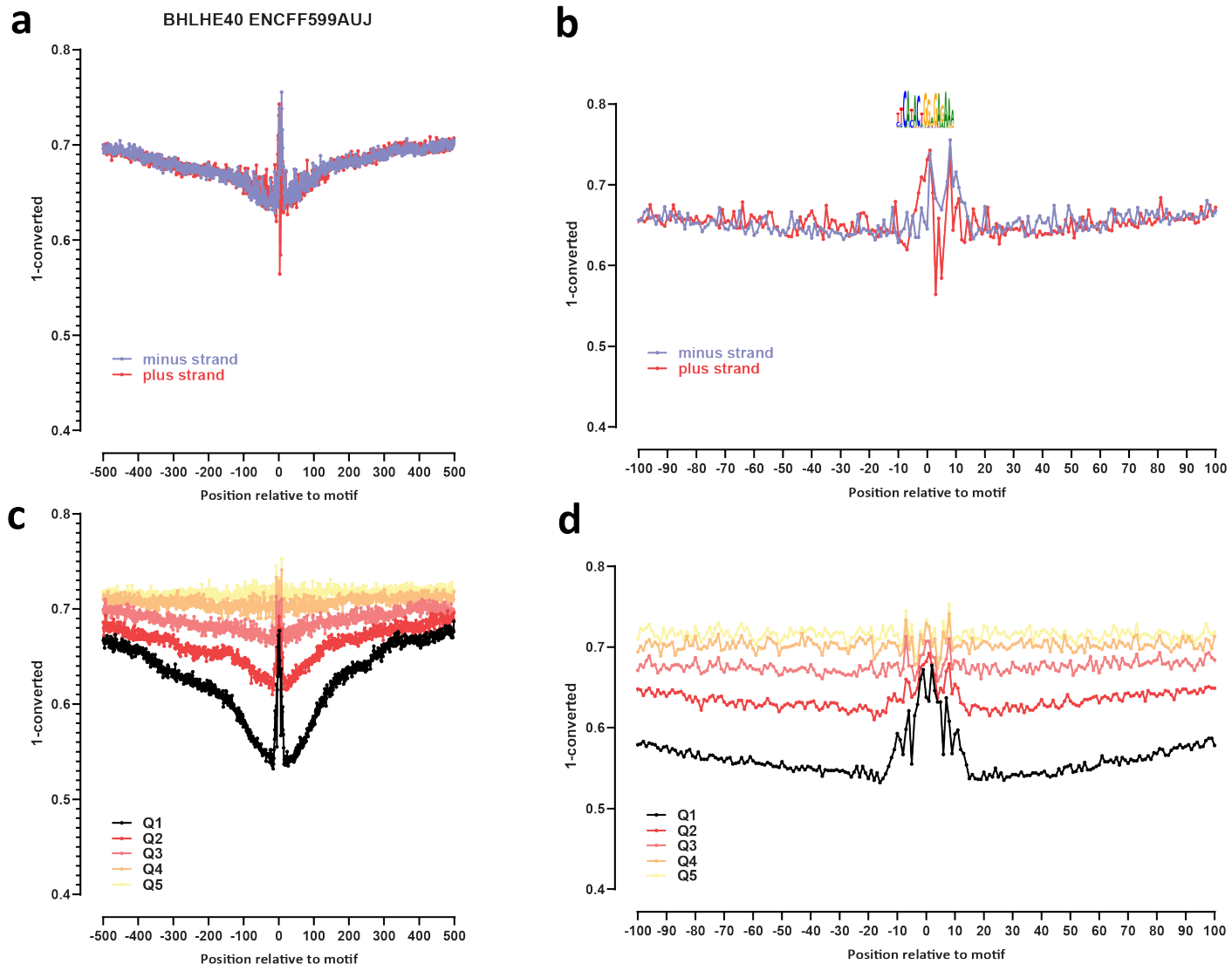
Supplementary Figure 9: Genome-wide footprint profiles for the BACH1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



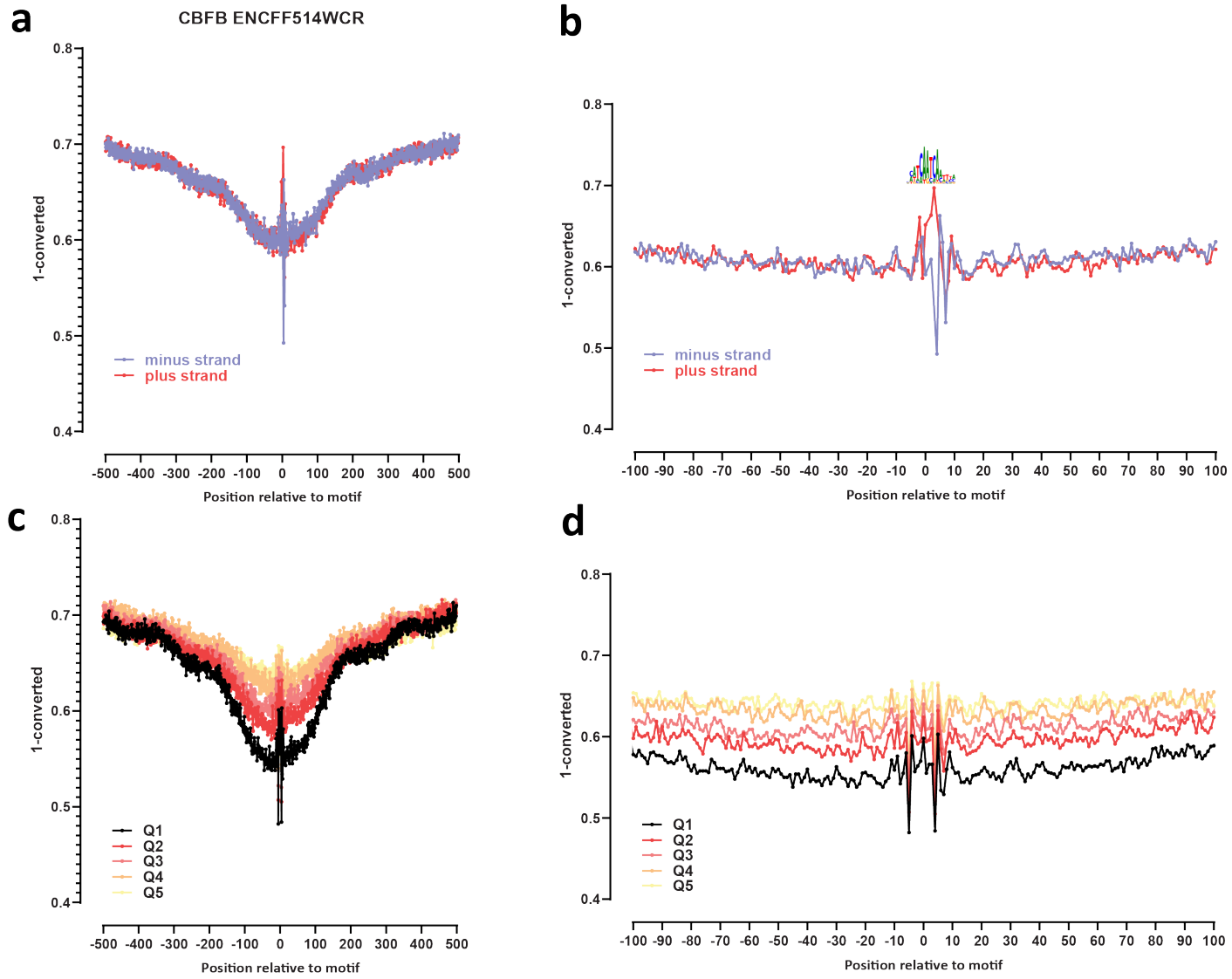
Supplementary Figure 10: Genome-wide footprint profiles for the BATF transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



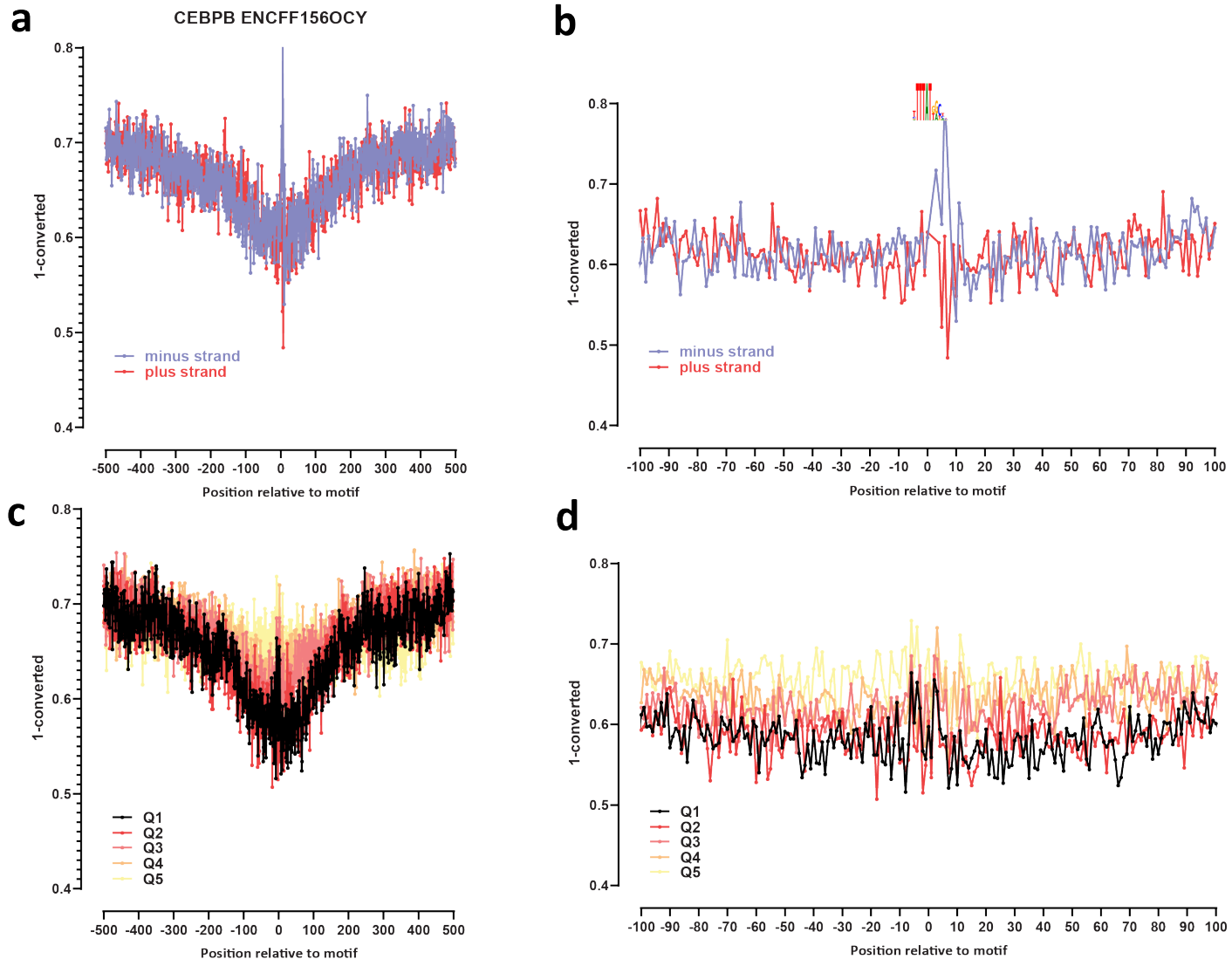
Supplementary Figure 11: Genome-wide footprint profiles for the BCL11A transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



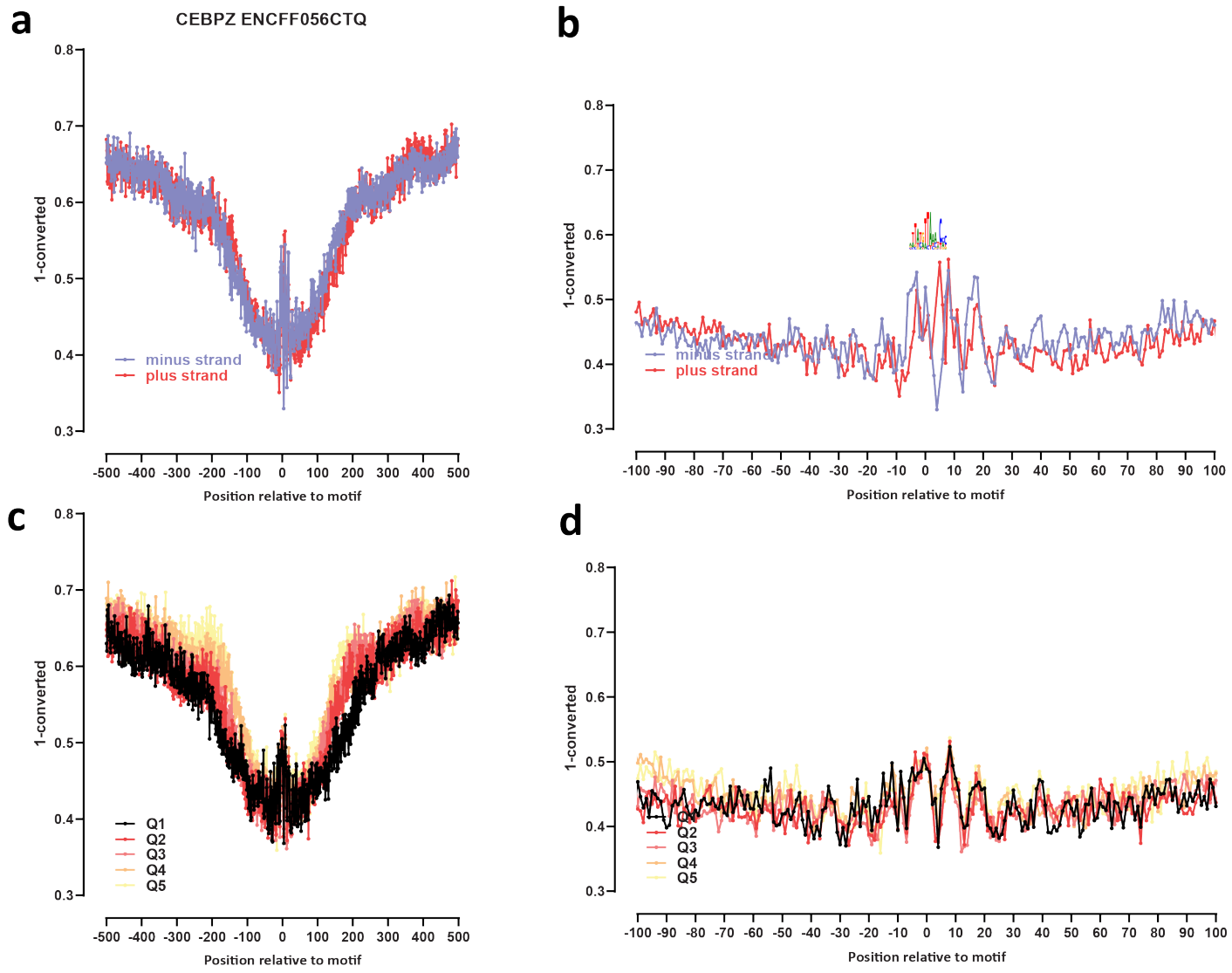
Supplementary Figure 12: Genome-wide footprint profiles for the BHLHE40 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



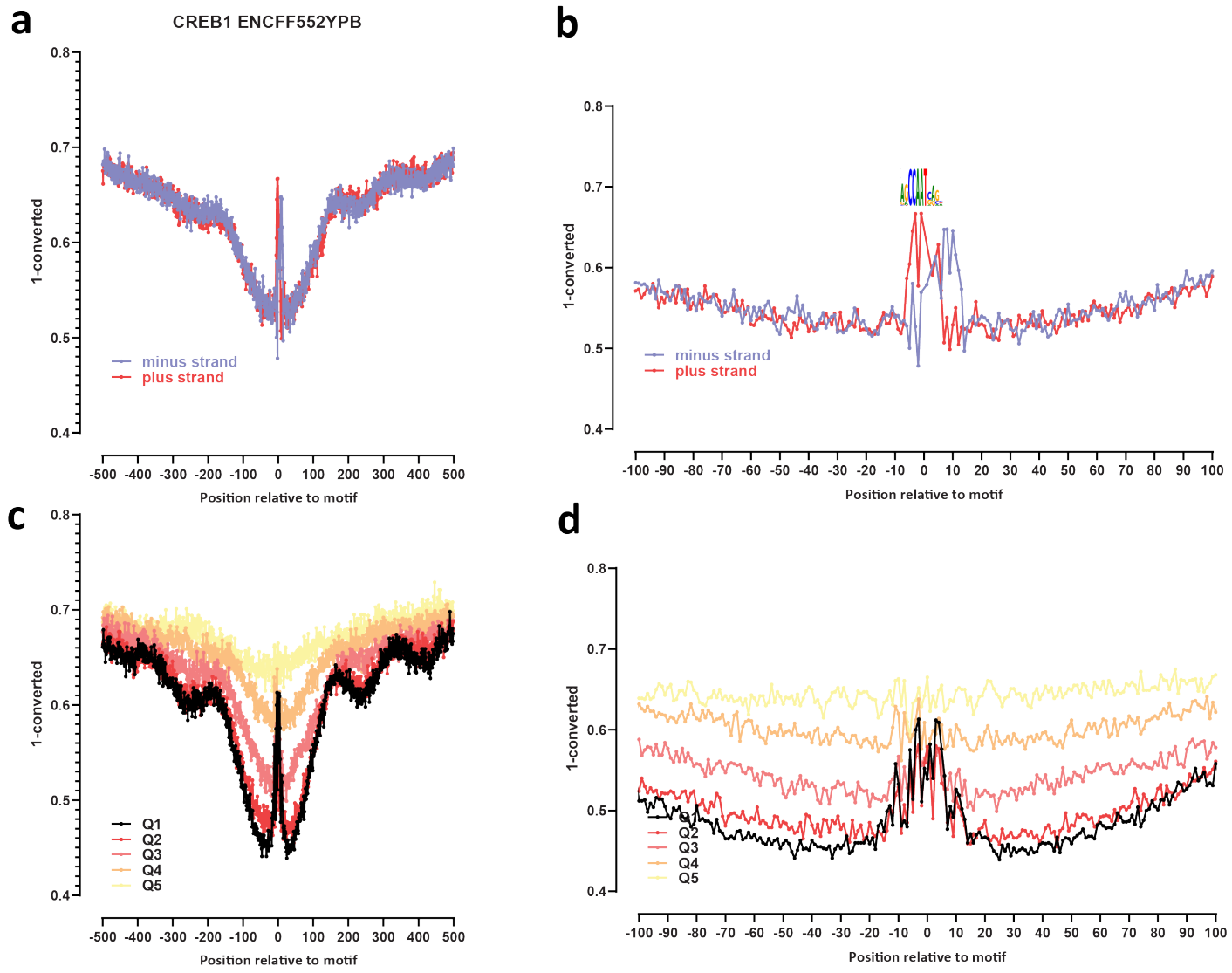
Supplementary Figure 13: Genome-wide footprint profiles for the CBFB transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



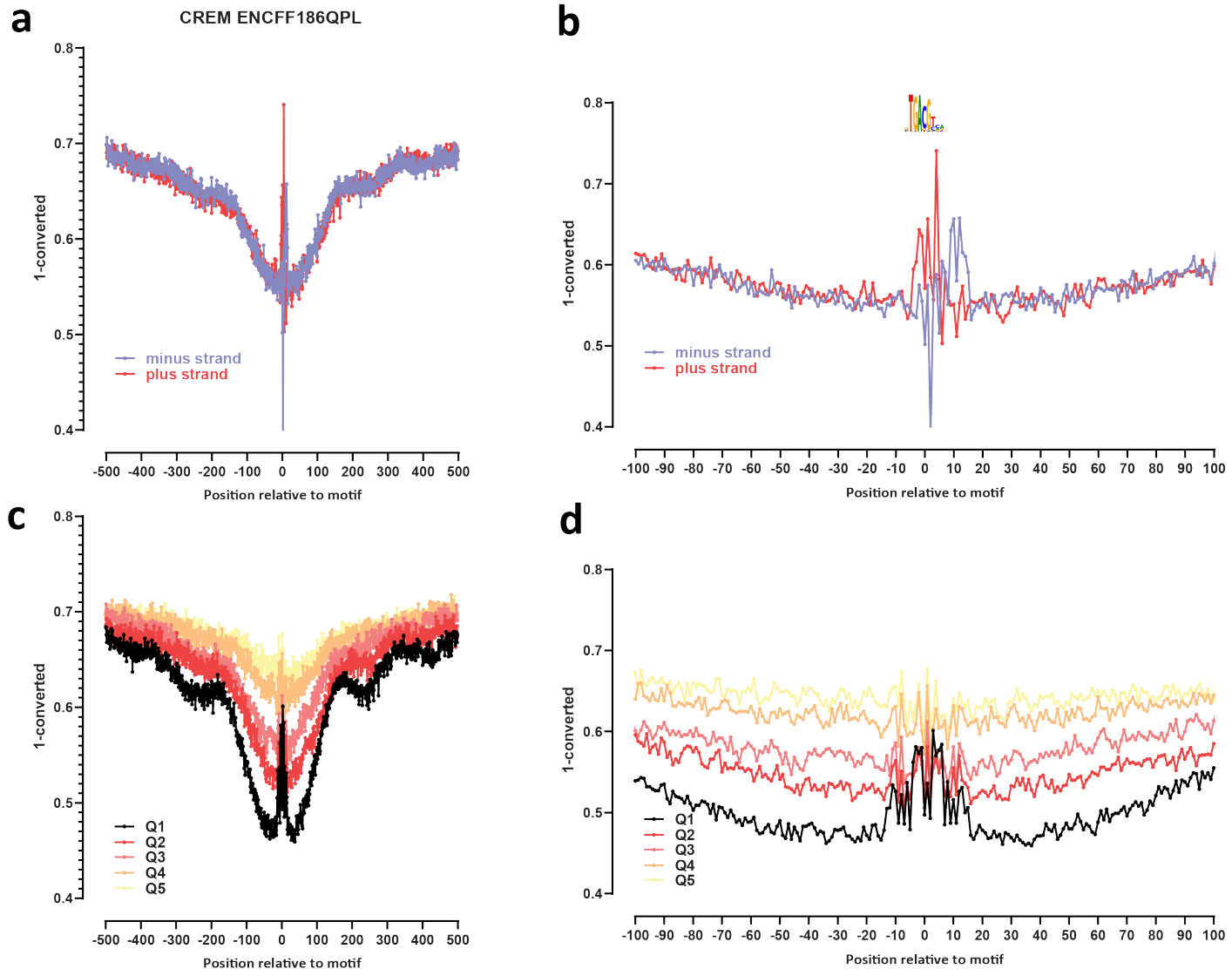
Supplementary Figure 14: Genome-wide footprint profiles for the CEBPB transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



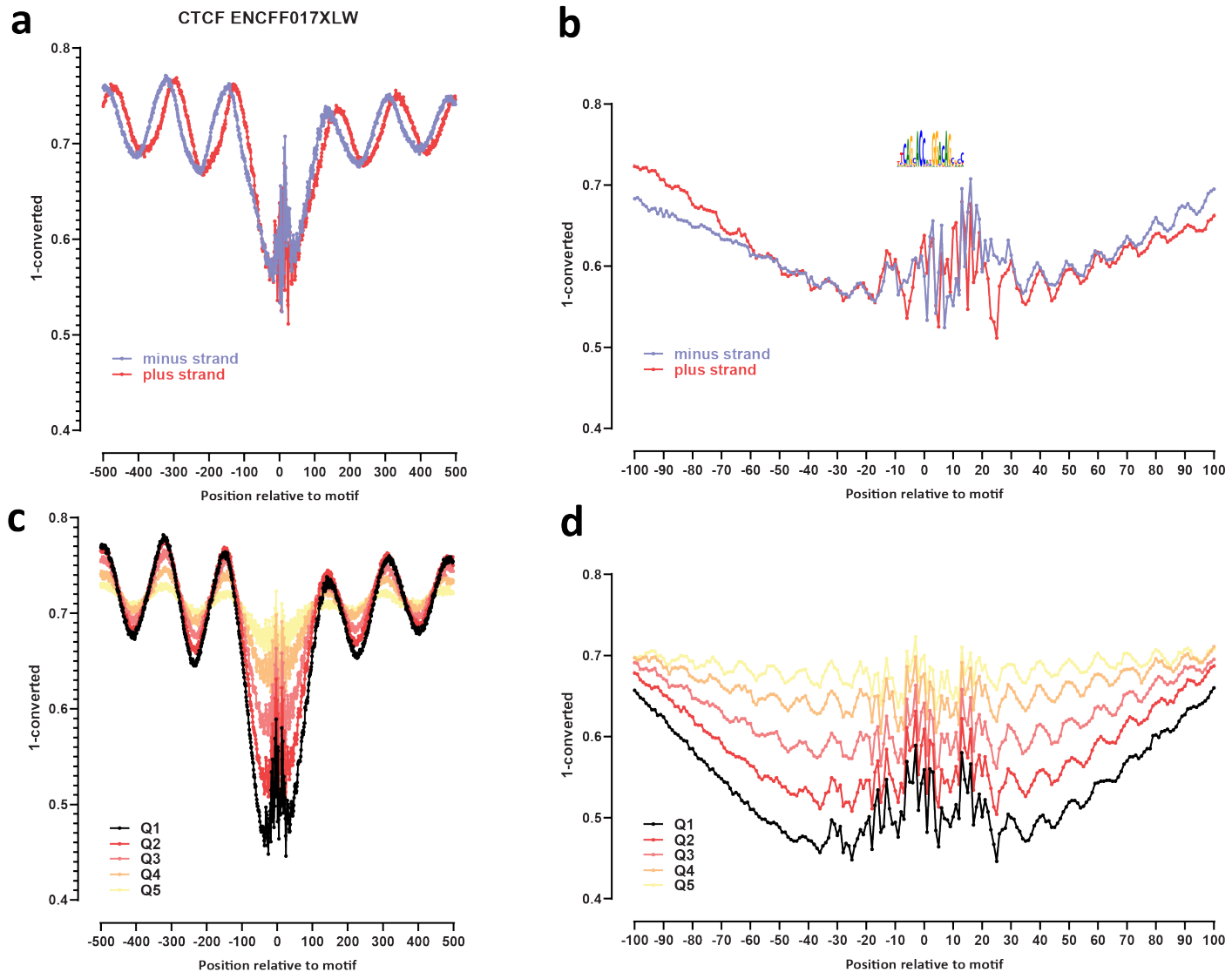
Supplementary Figure 15: Genome-wide footprint profiles for the CEBPZ transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



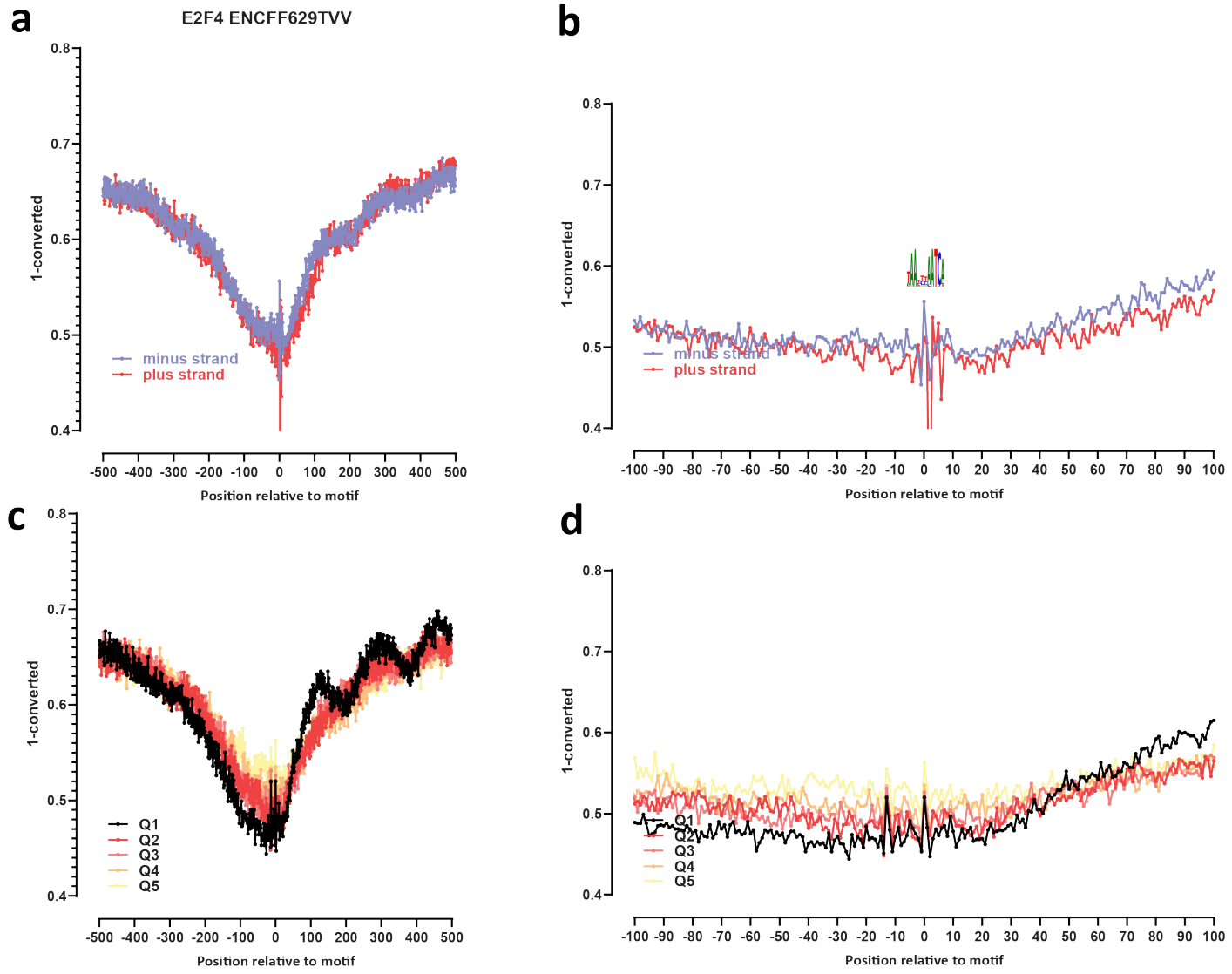
Supplementary Figure 16: Genome-wide footprint profiles for the CREB1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



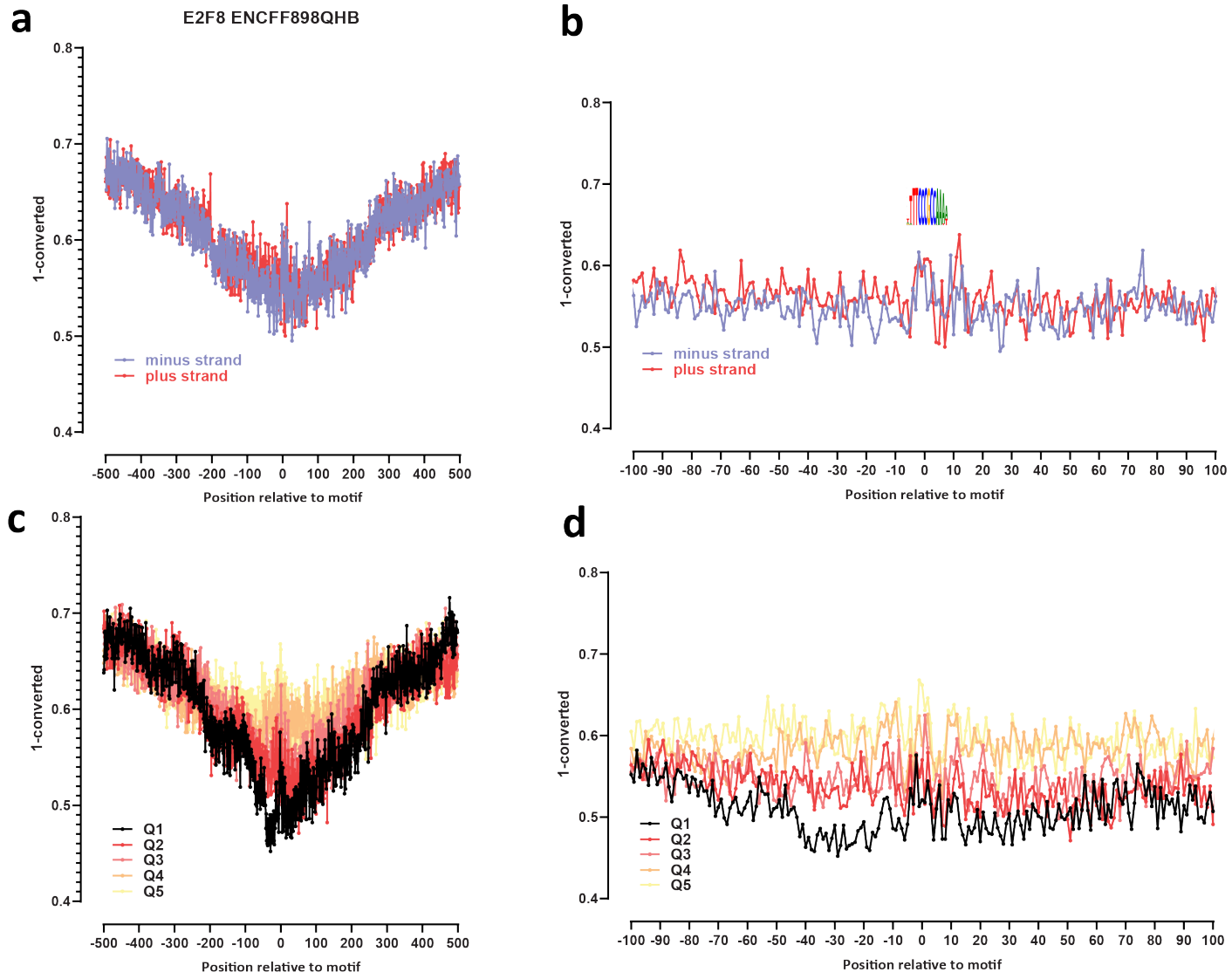
Supplementary Figure 17: Genome-wide footprint profiles for the CREM transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



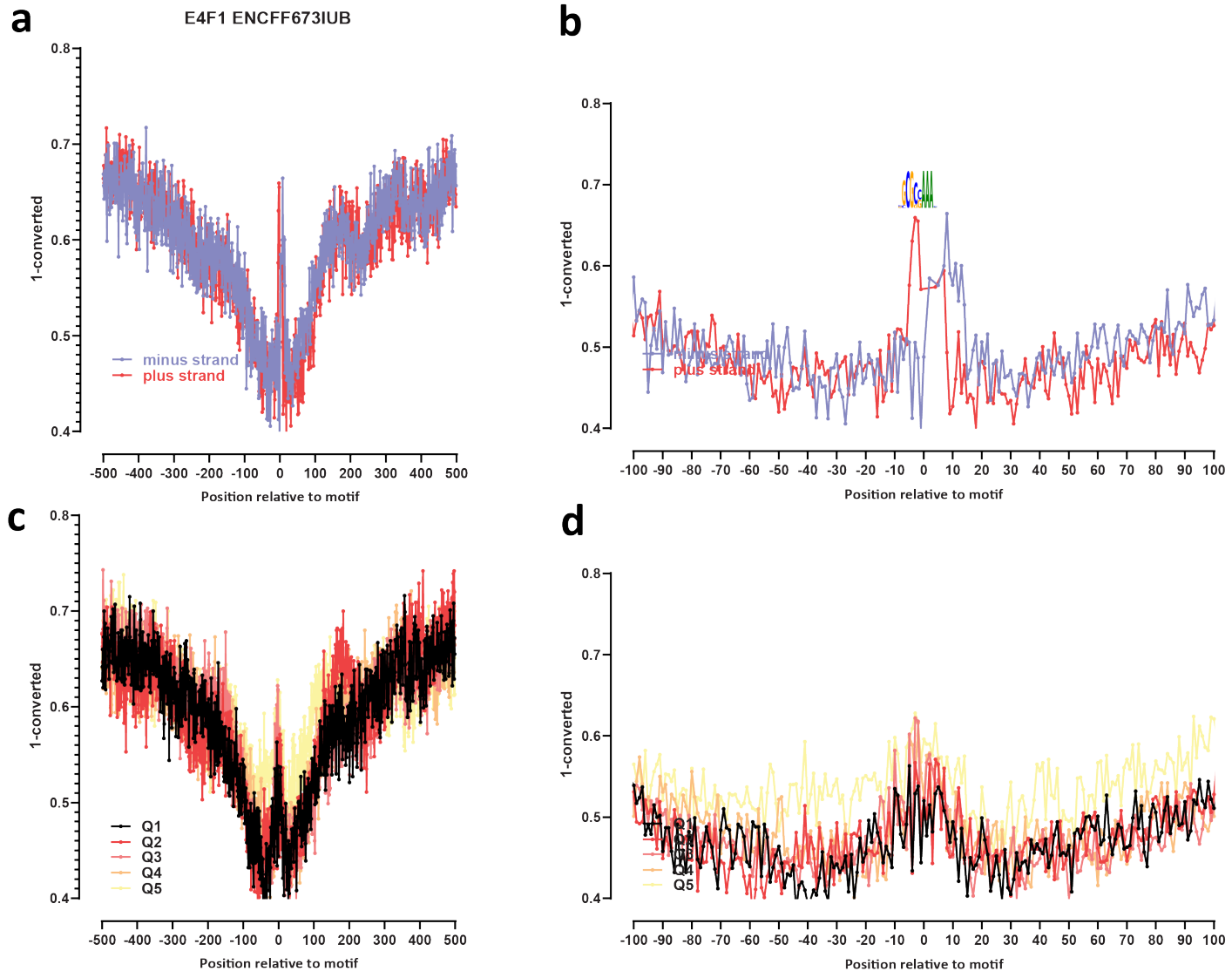
Supplementary Figure 18: Genome-wide footprint profiles for the CTCF transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



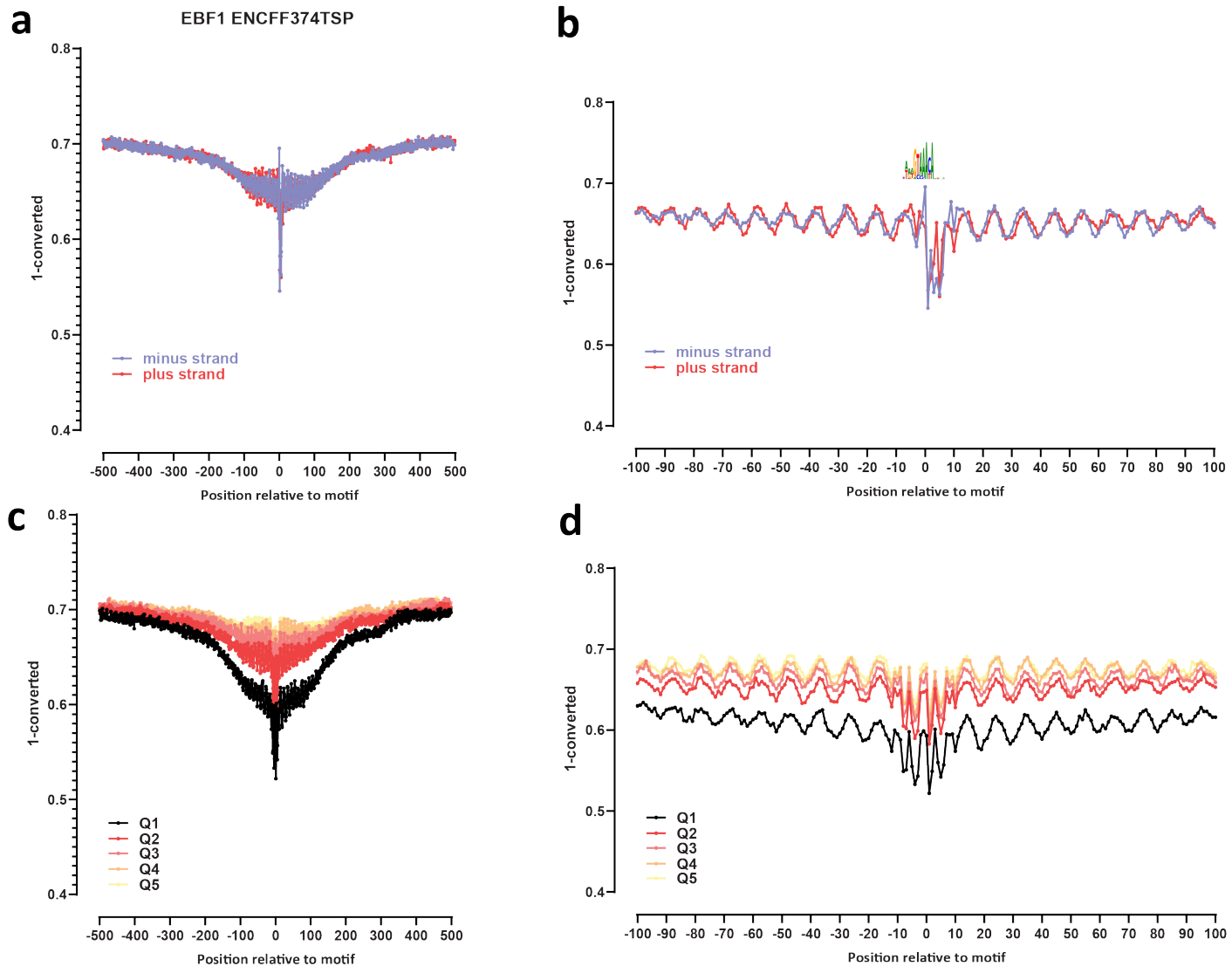
Supplementary Figure 19: Genome-wide footprint profiles for the E2F4 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



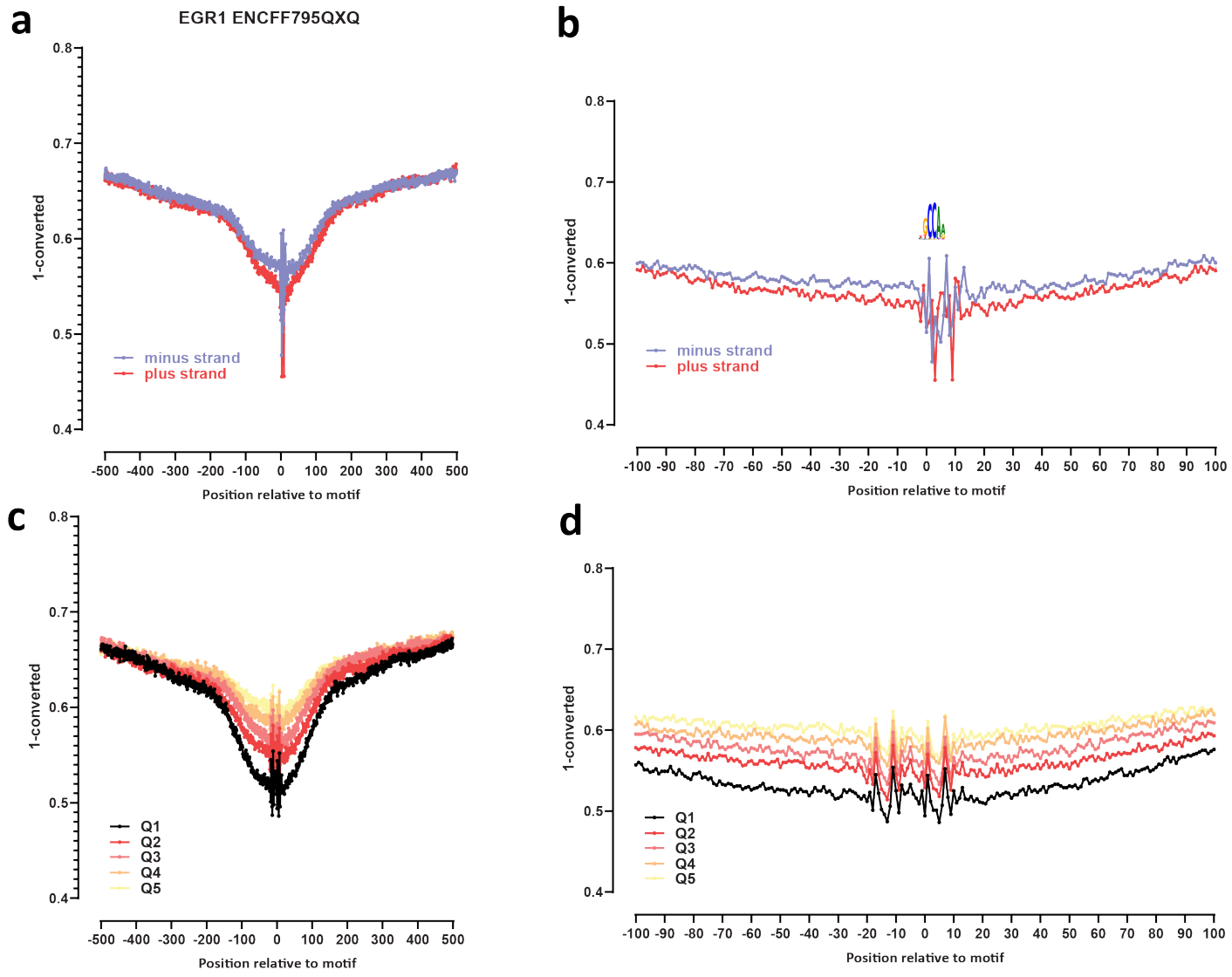
Supplementary Figure 20: Genome-wide footprint profiles for the E2F8 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



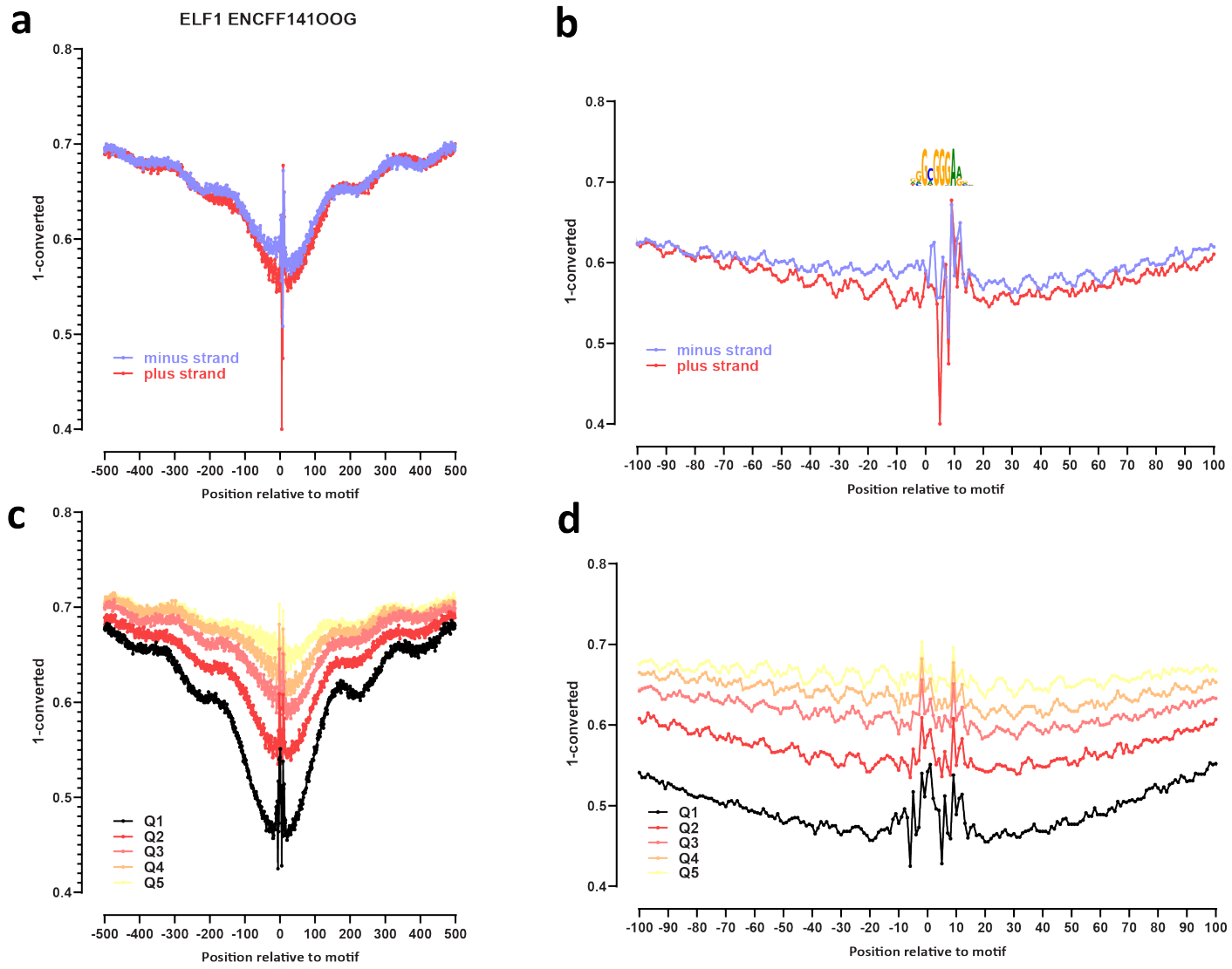
Supplementary Figure 21: Genome-wide footprint profiles for the E4F1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



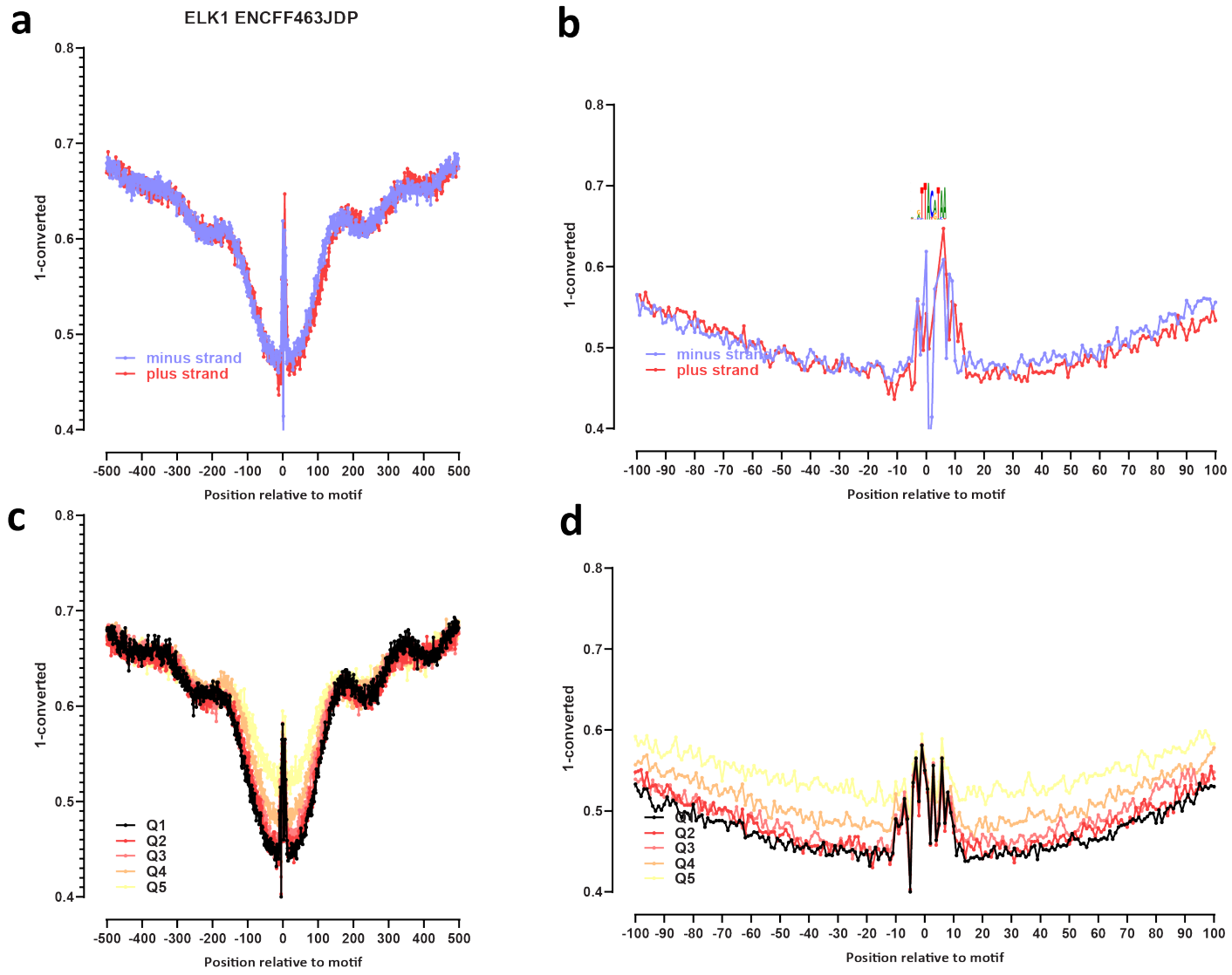
Supplementary Figure 22: Genome-wide footprint profiles for the EBF1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



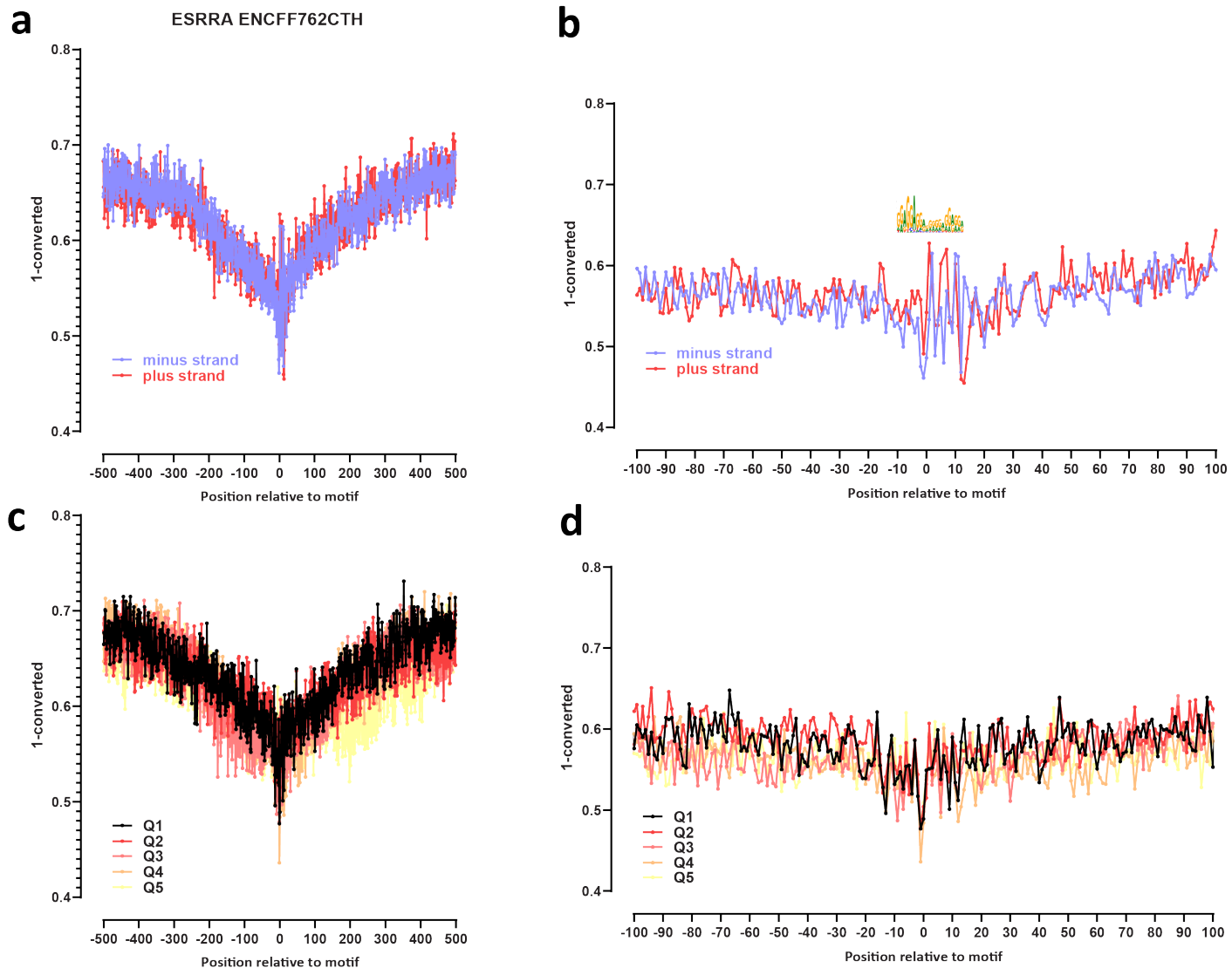
Supplementary Figure 23: Genome-wide footprint profiles for the EGR1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



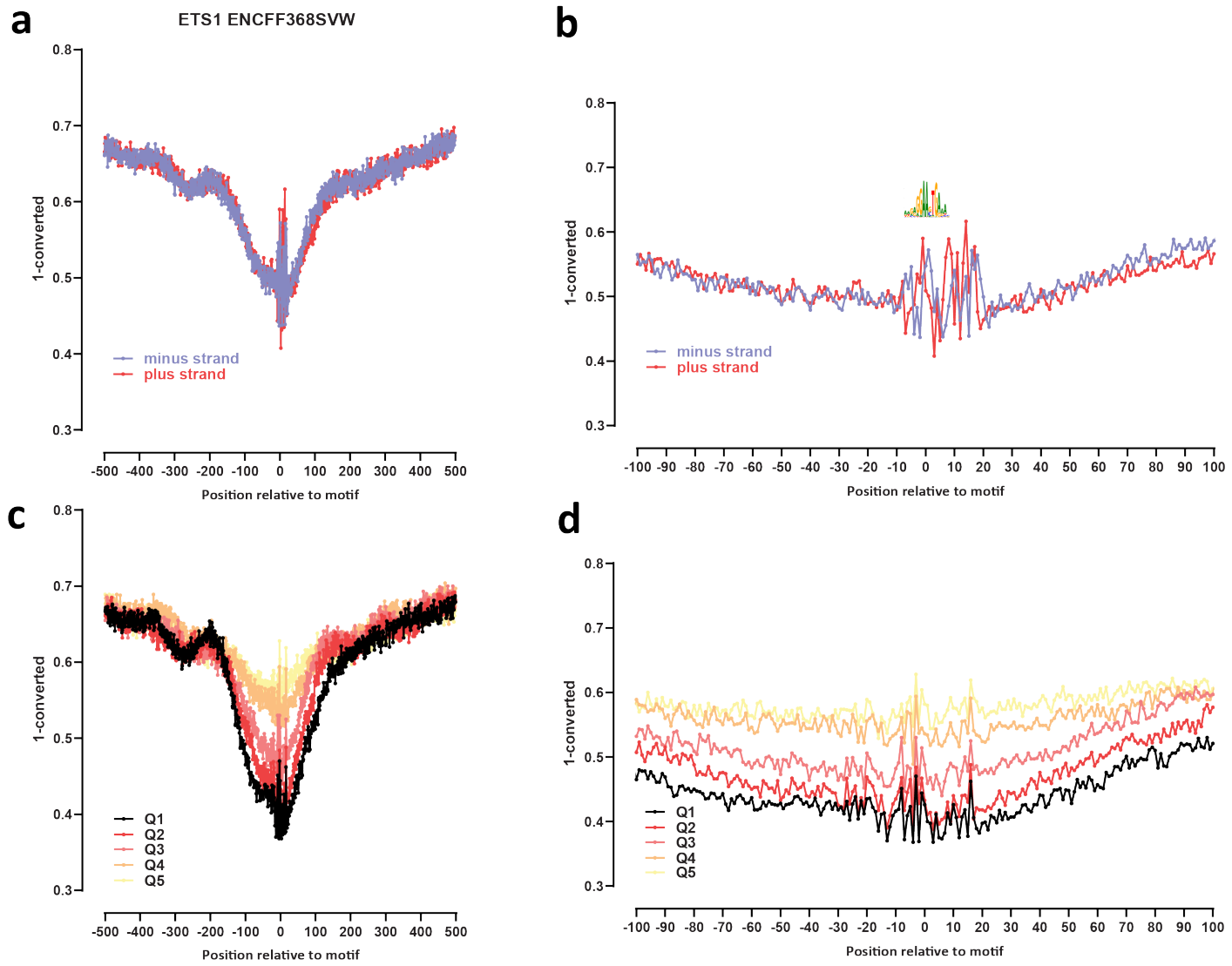
Supplementary Figure 24: Genome-wide footprint profiles for the ELF1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



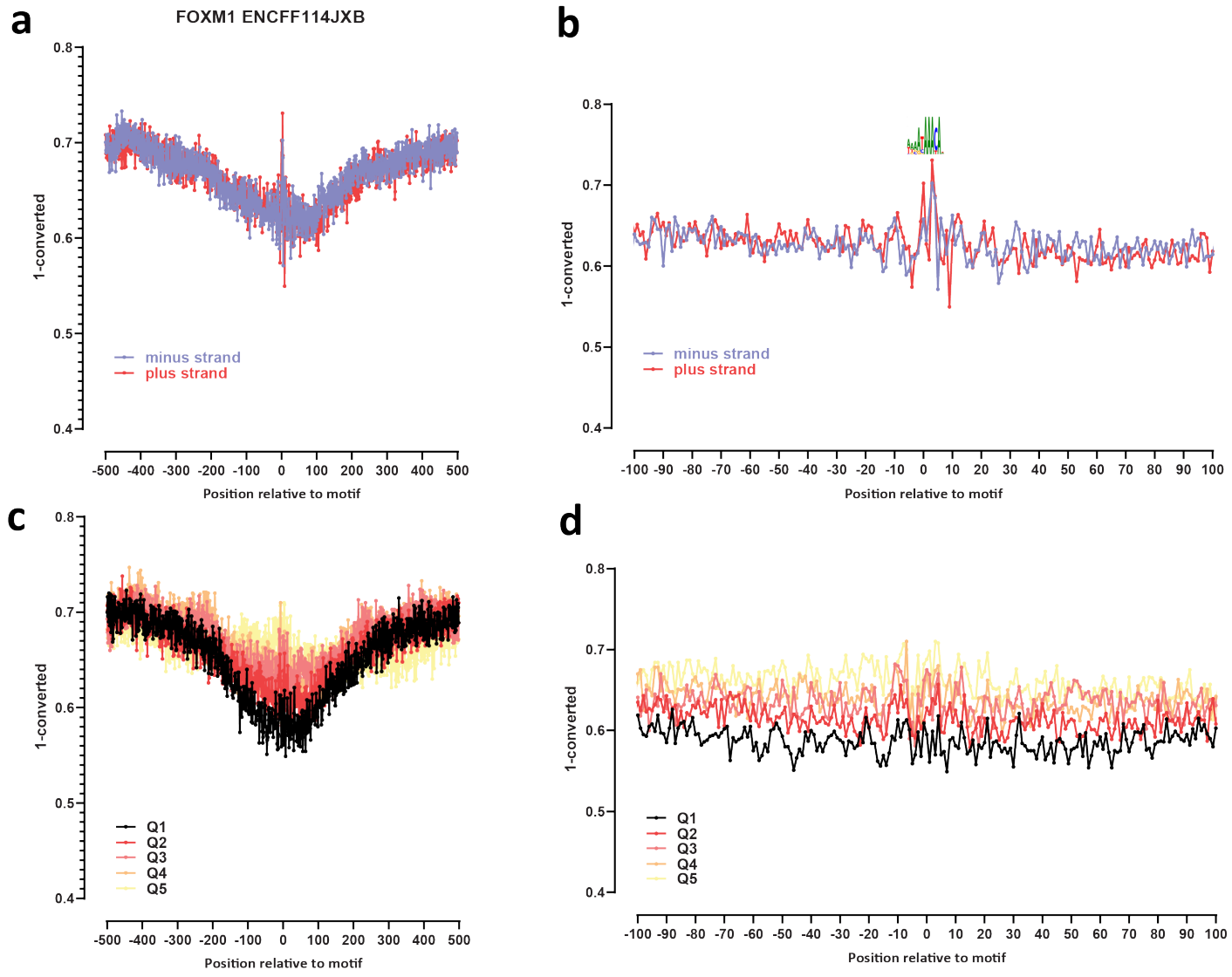
Supplementary Figure 25: Genome-wide footprint profiles for the ELK1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



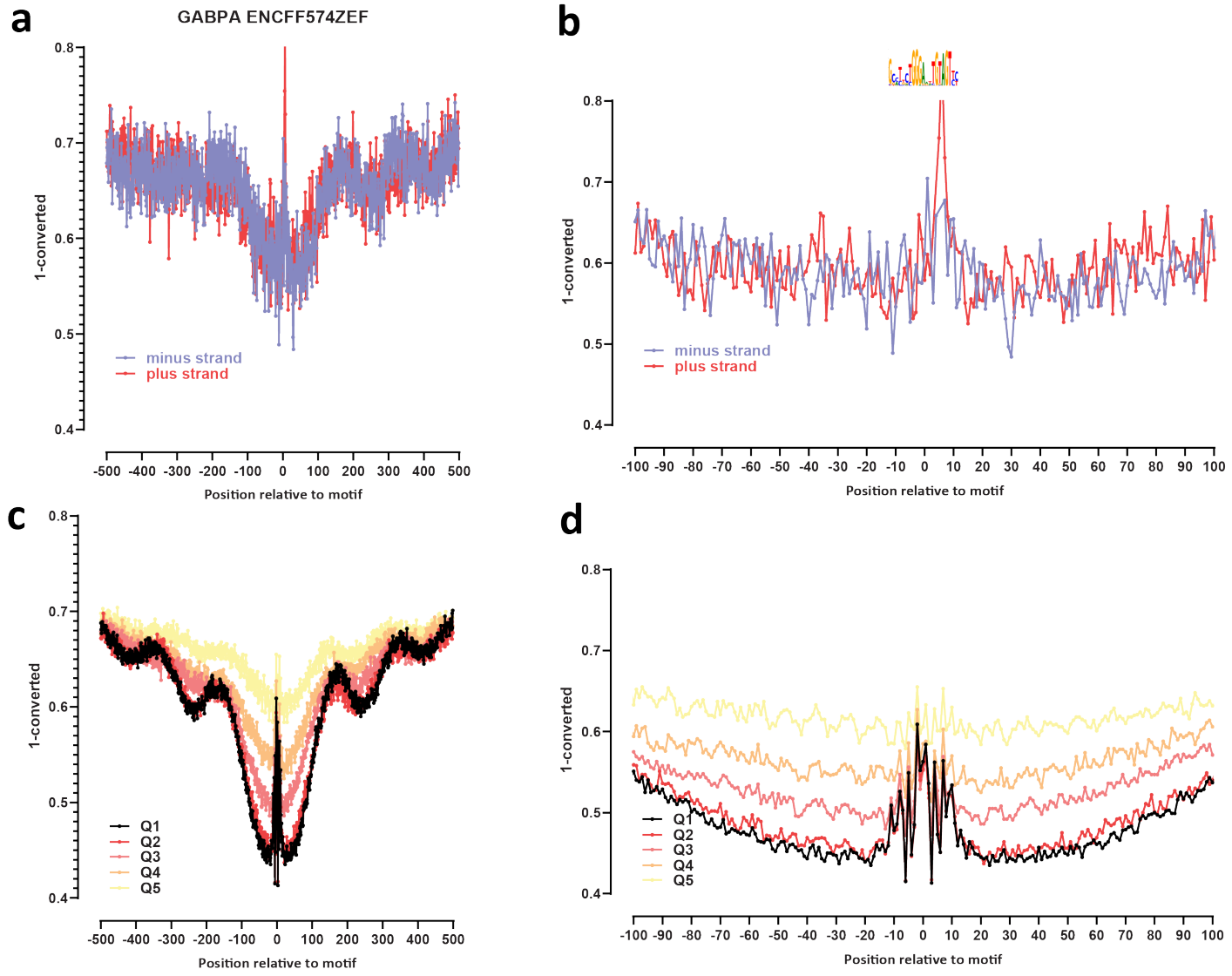
Supplementary Figure 26: Genome-wide footprint profiles for the ESRRA transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



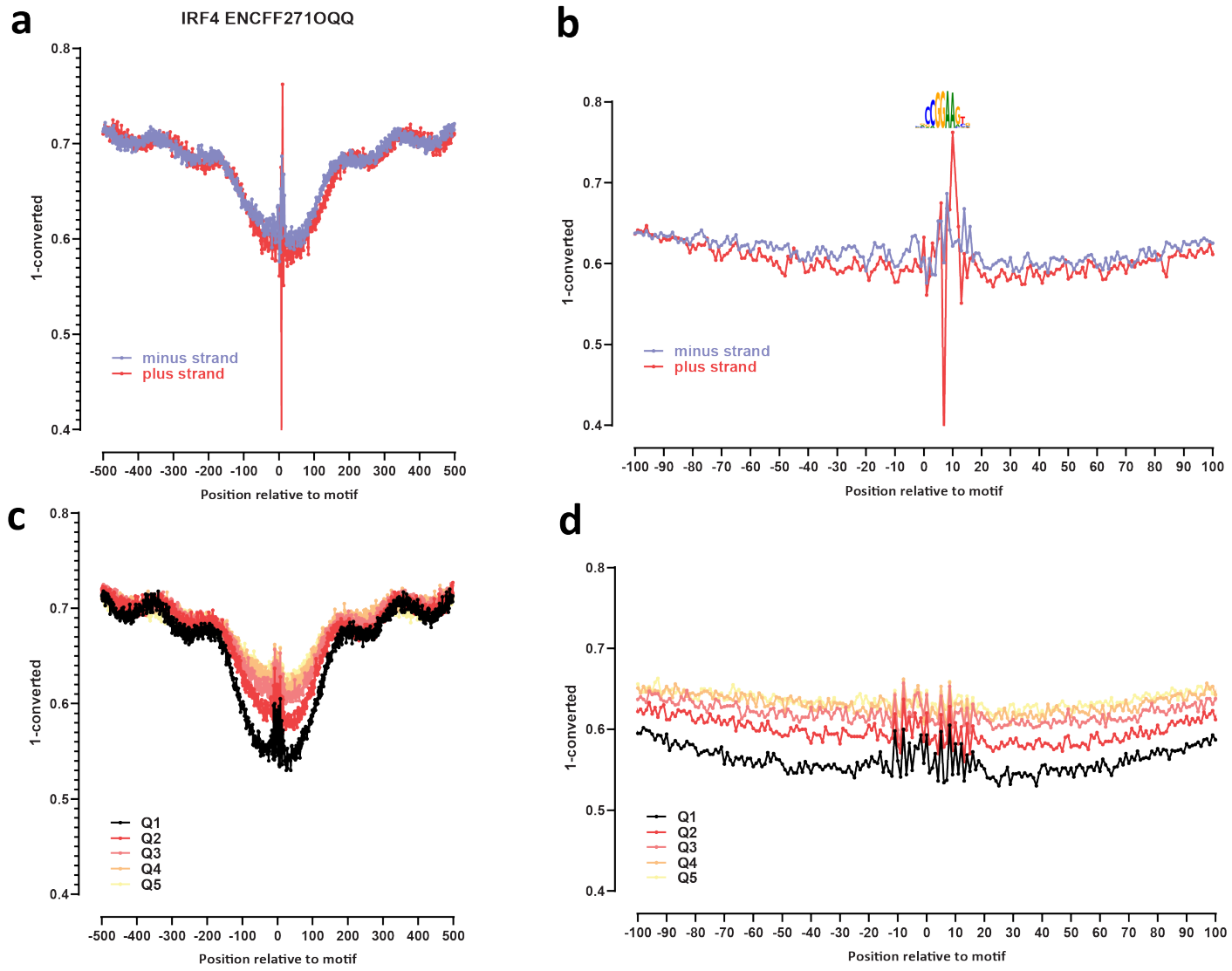
Supplementary Figure 27: Genome-wide footprint profiles for the ETS1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



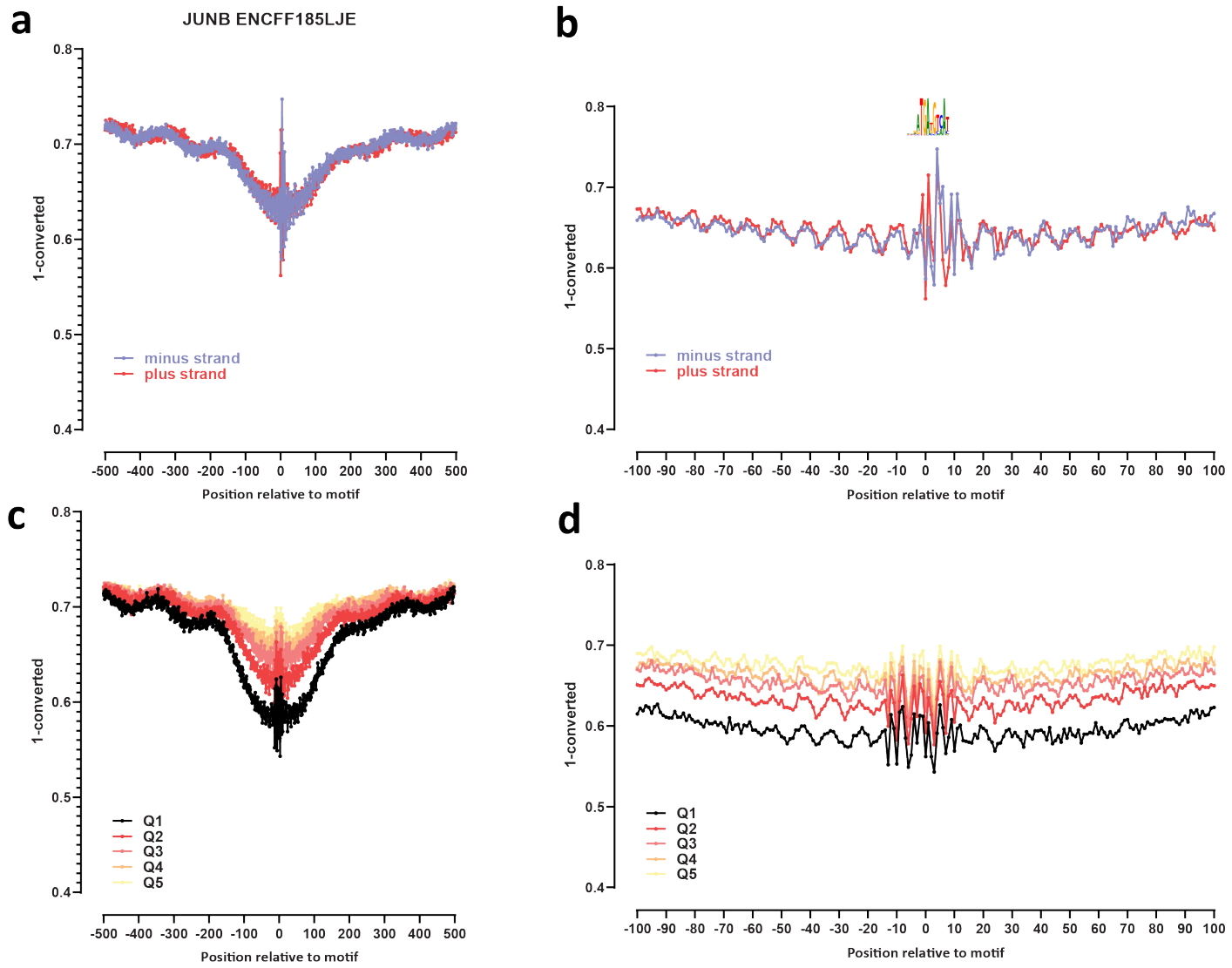
Supplementary Figure 28: Genome-wide footprint profiles for the FOXM1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



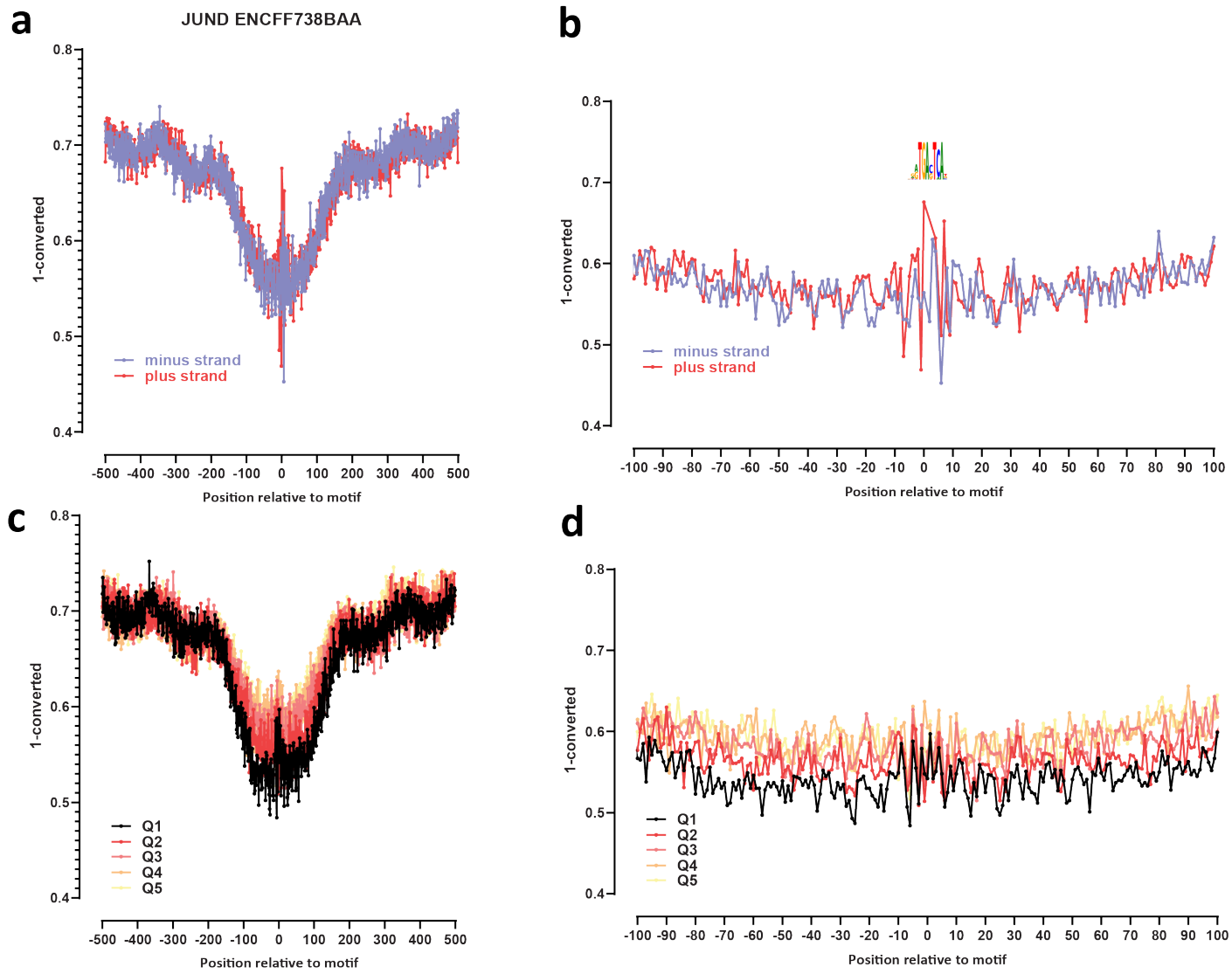
Supplementary Figure 29: Genome-wide footprint profiles for the GABPA transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



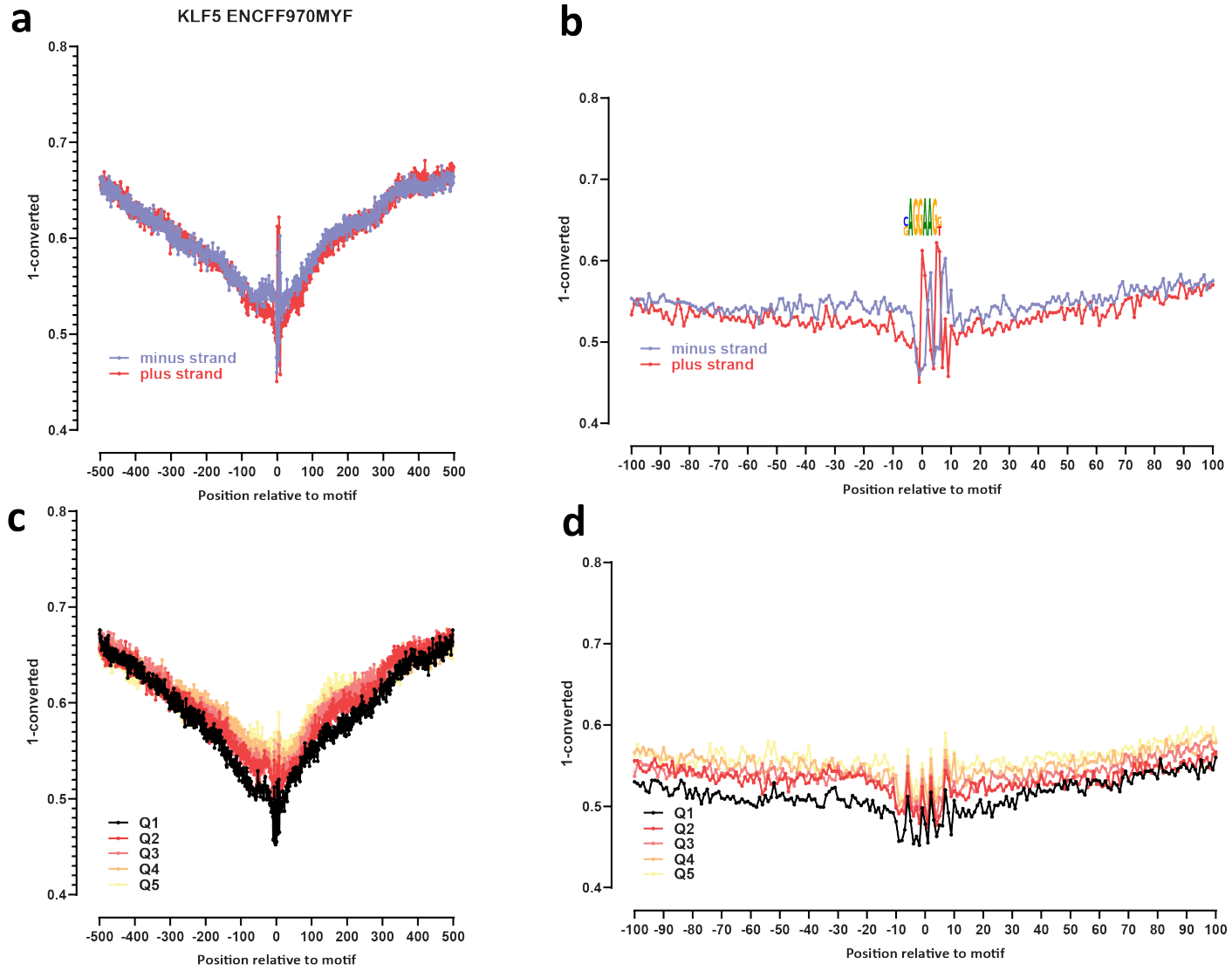
Supplementary Figure 30: Genome-wide footprint profiles for the IRF4 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



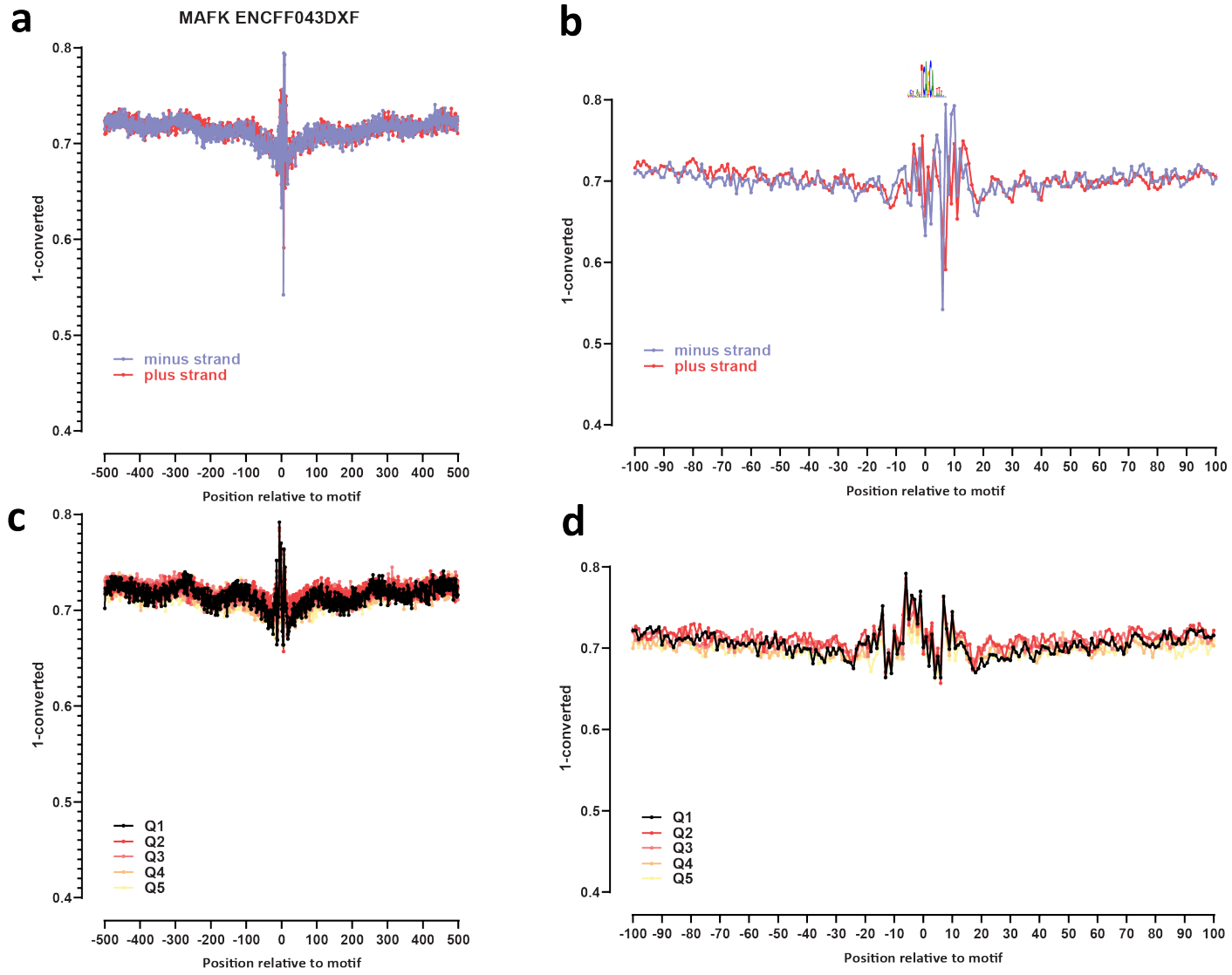
Supplementary Figure 31: Genome-wide footprint profiles for the JUNB transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



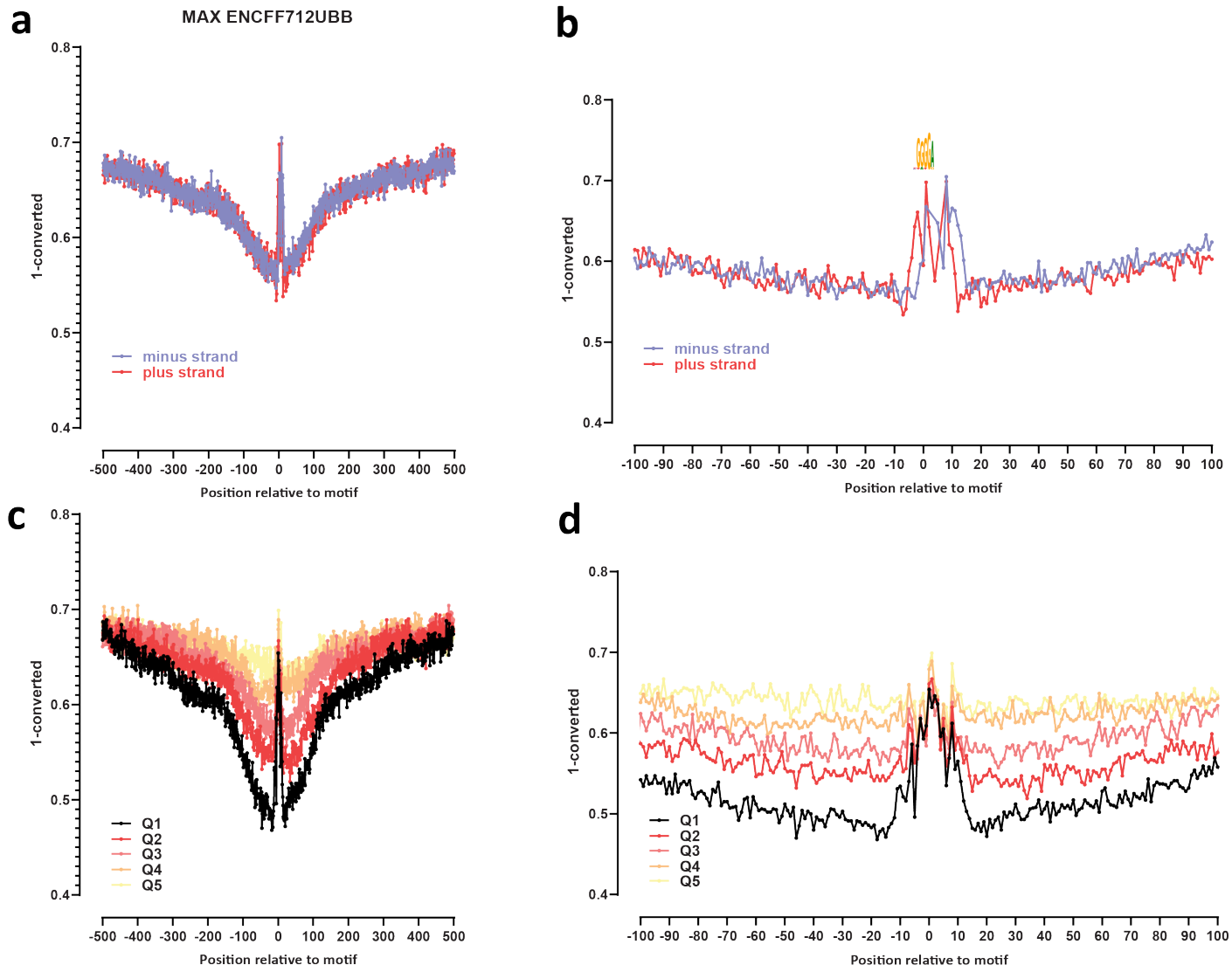
Supplementary Figure 32: Genome-wide footprint profiles for the JUND transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



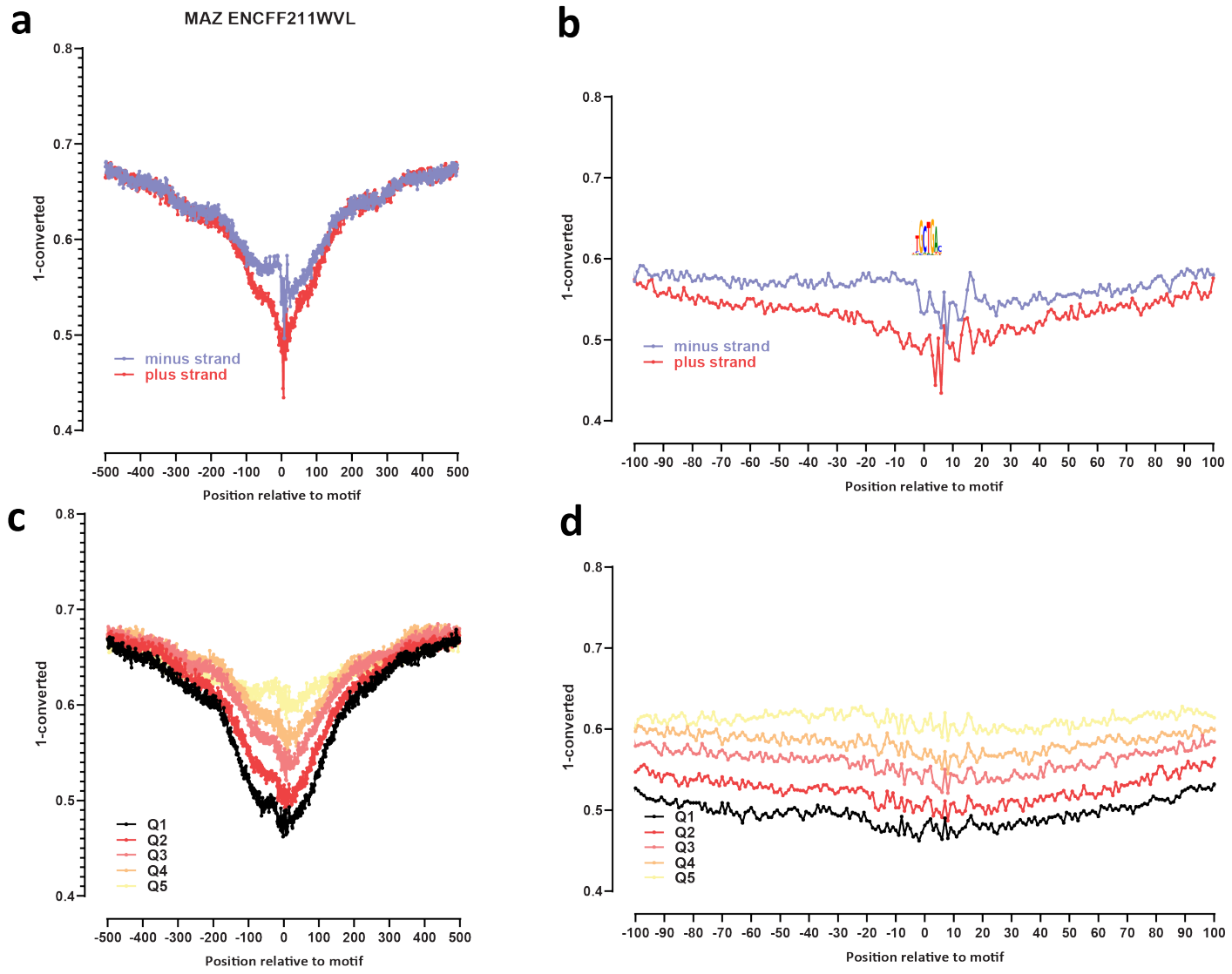
Supplementary Figure 33: Genome-wide footprint profiles for the KLF5 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



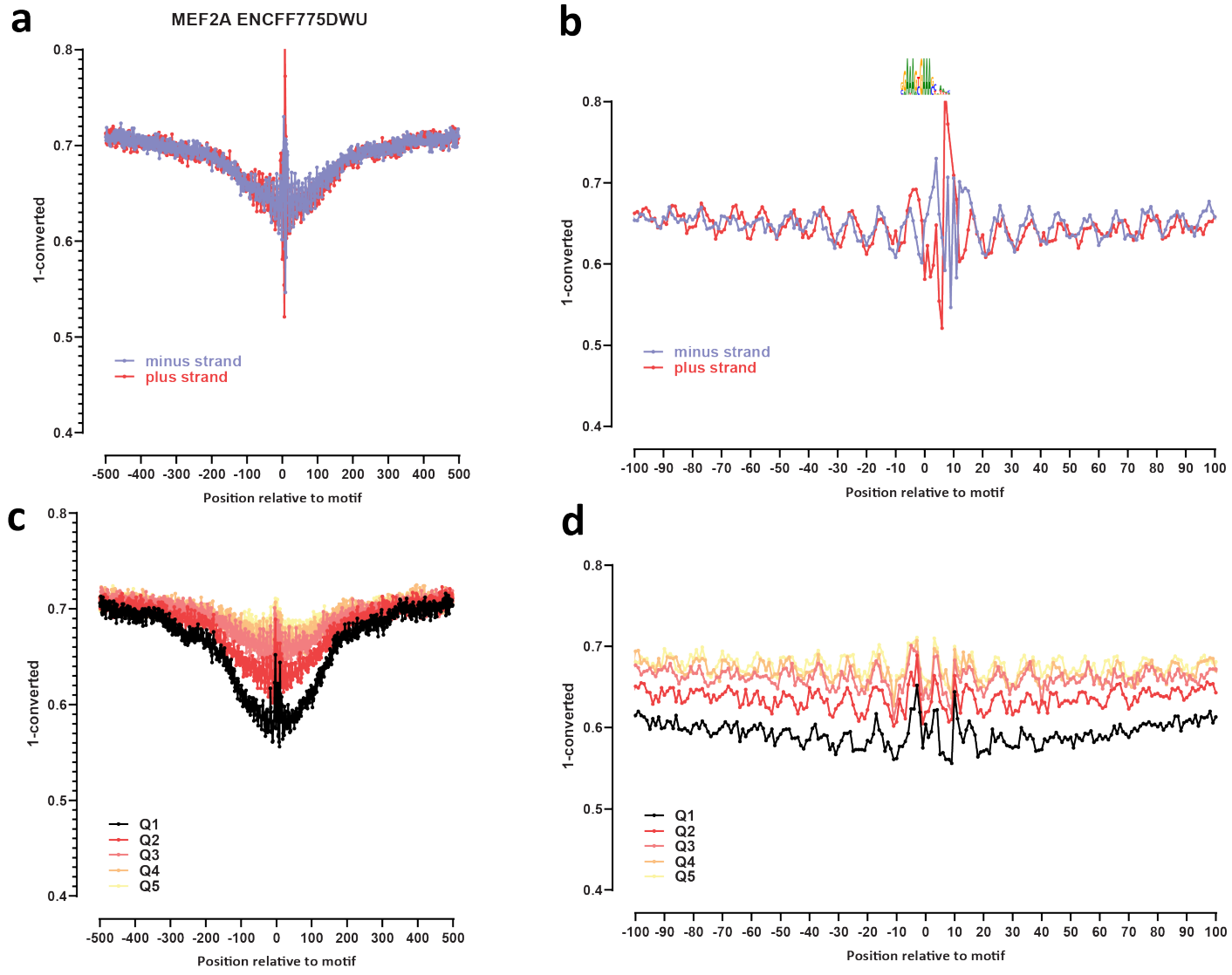
Supplementary Figure 34: Genome-wide footprint profiles for the MAFK transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



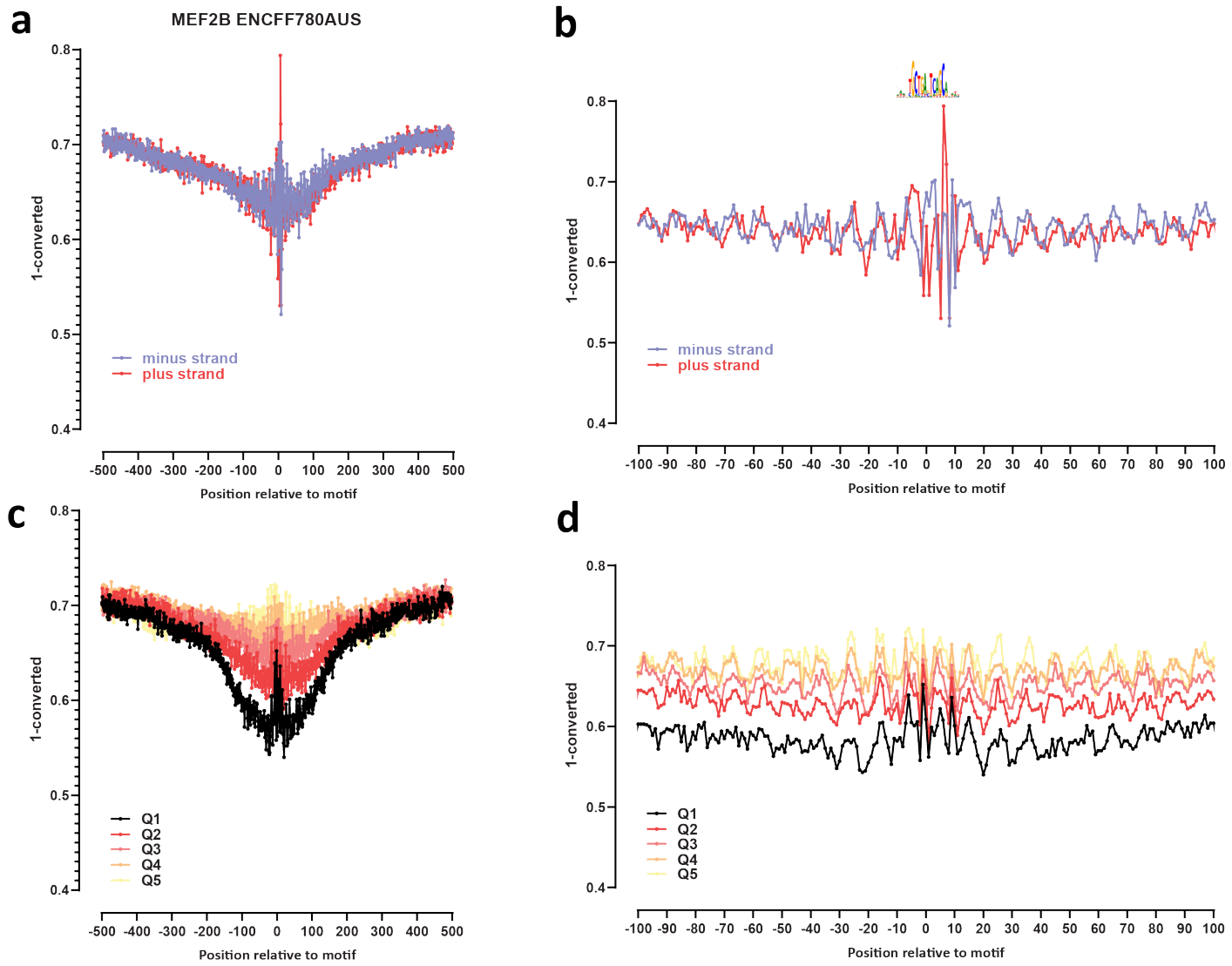
Supplementary Figure 35: Genome-wide footprint profiles for the MAX transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



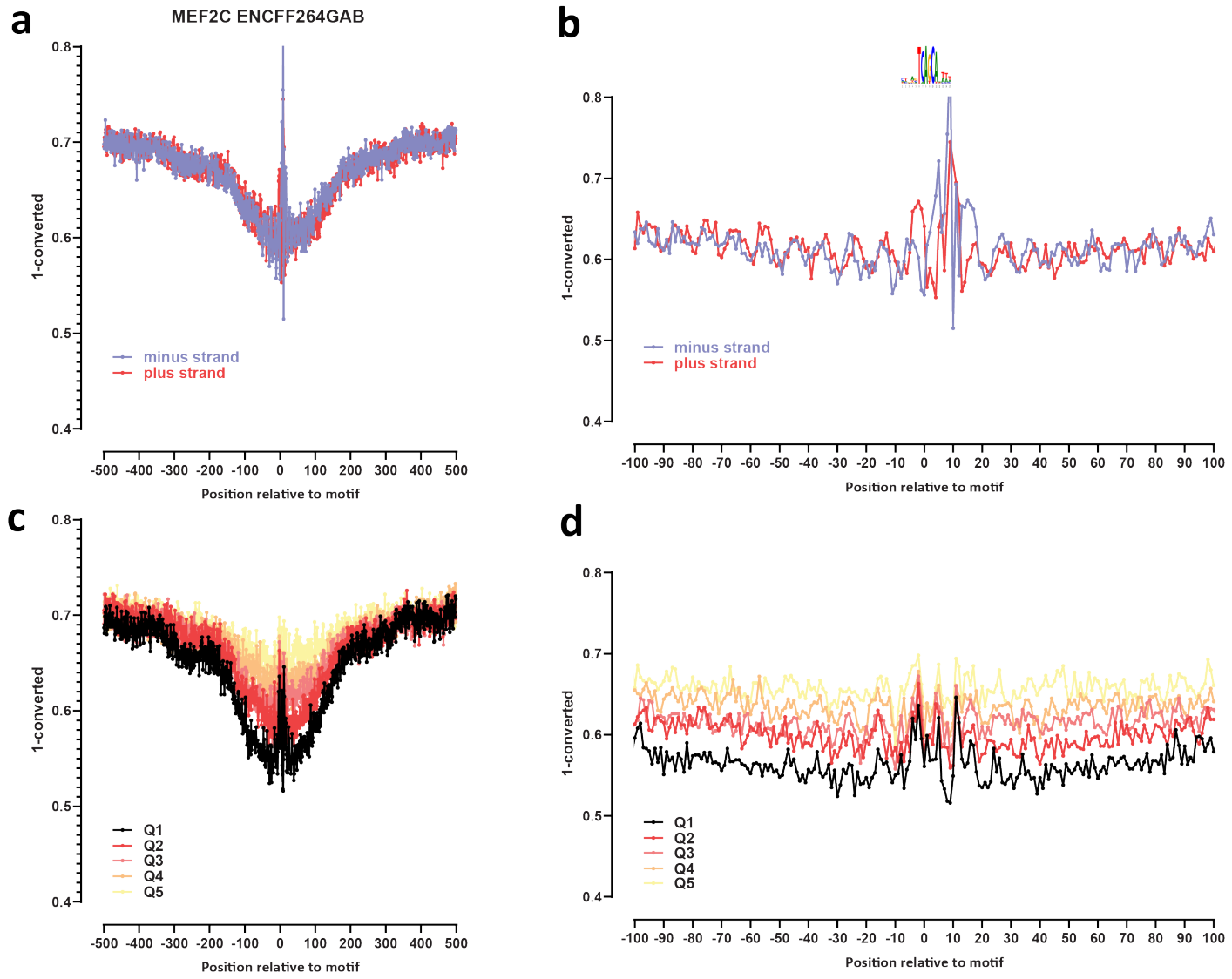
Supplementary Figure 36: Genome-wide footprint profiles for the MAZ transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



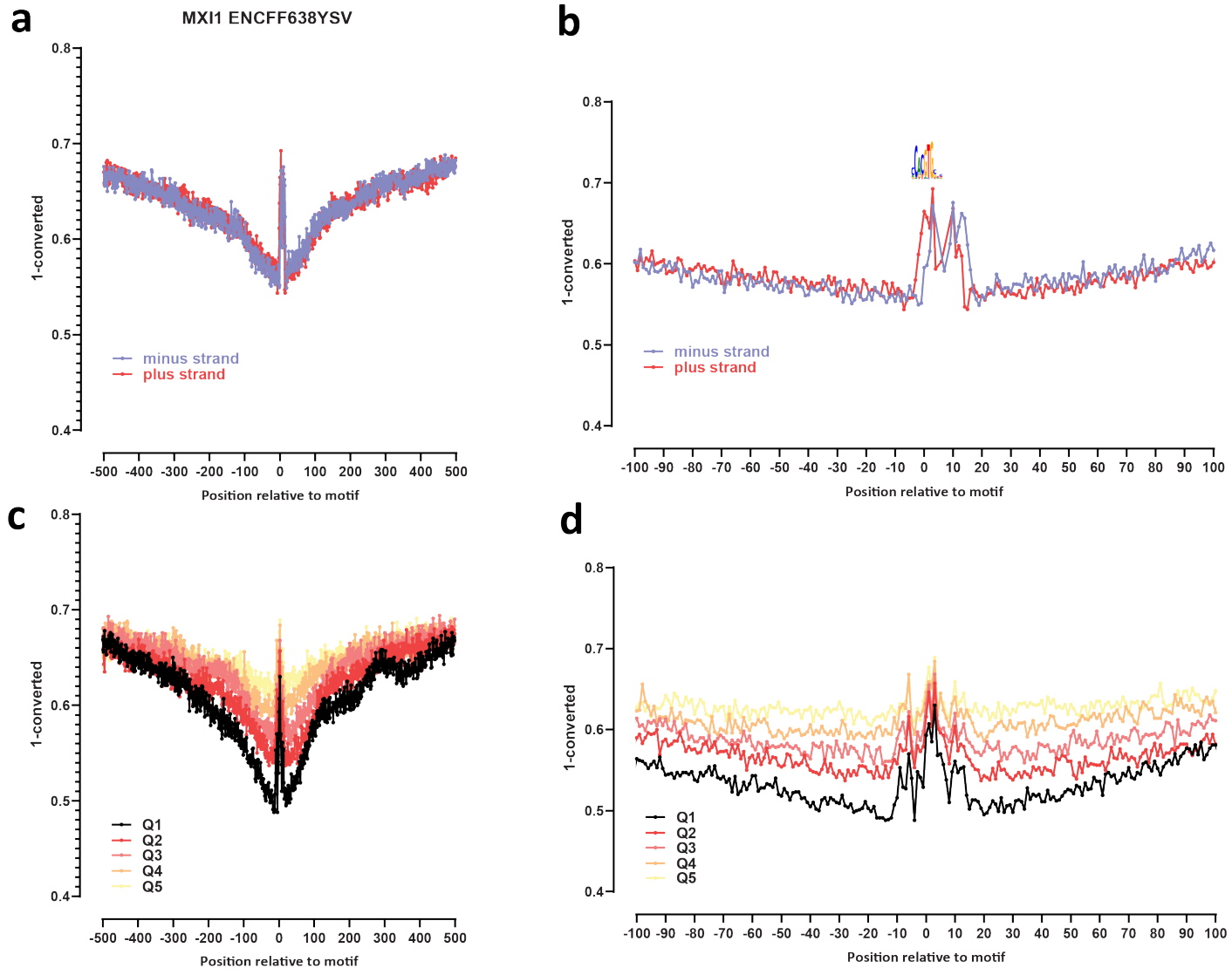
Supplementary Figure 37: Genome-wide footprint profiles for the MEF2A transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



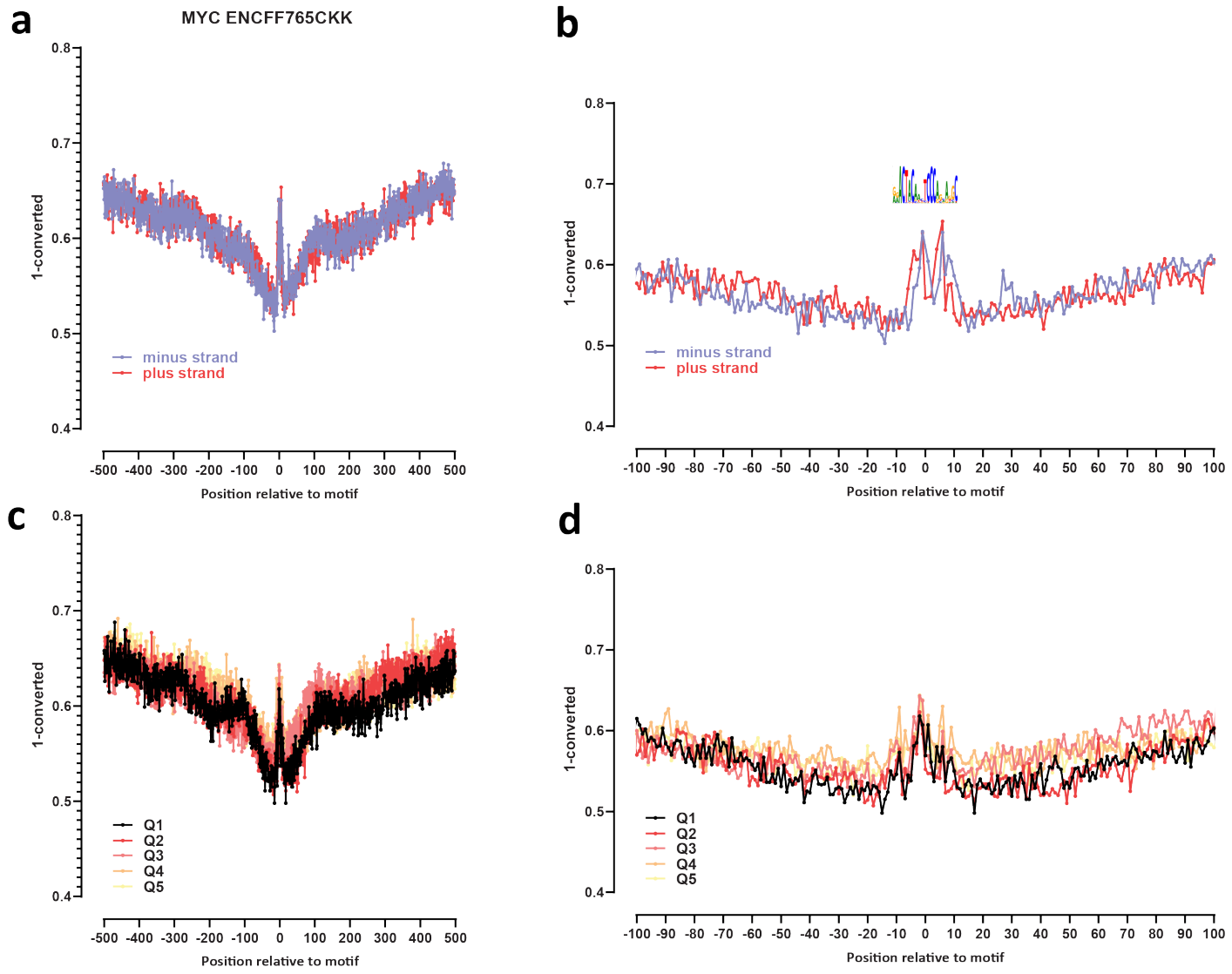
Supplementary Figure 38: Genome-wide footprint profiles for the MEF2B transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



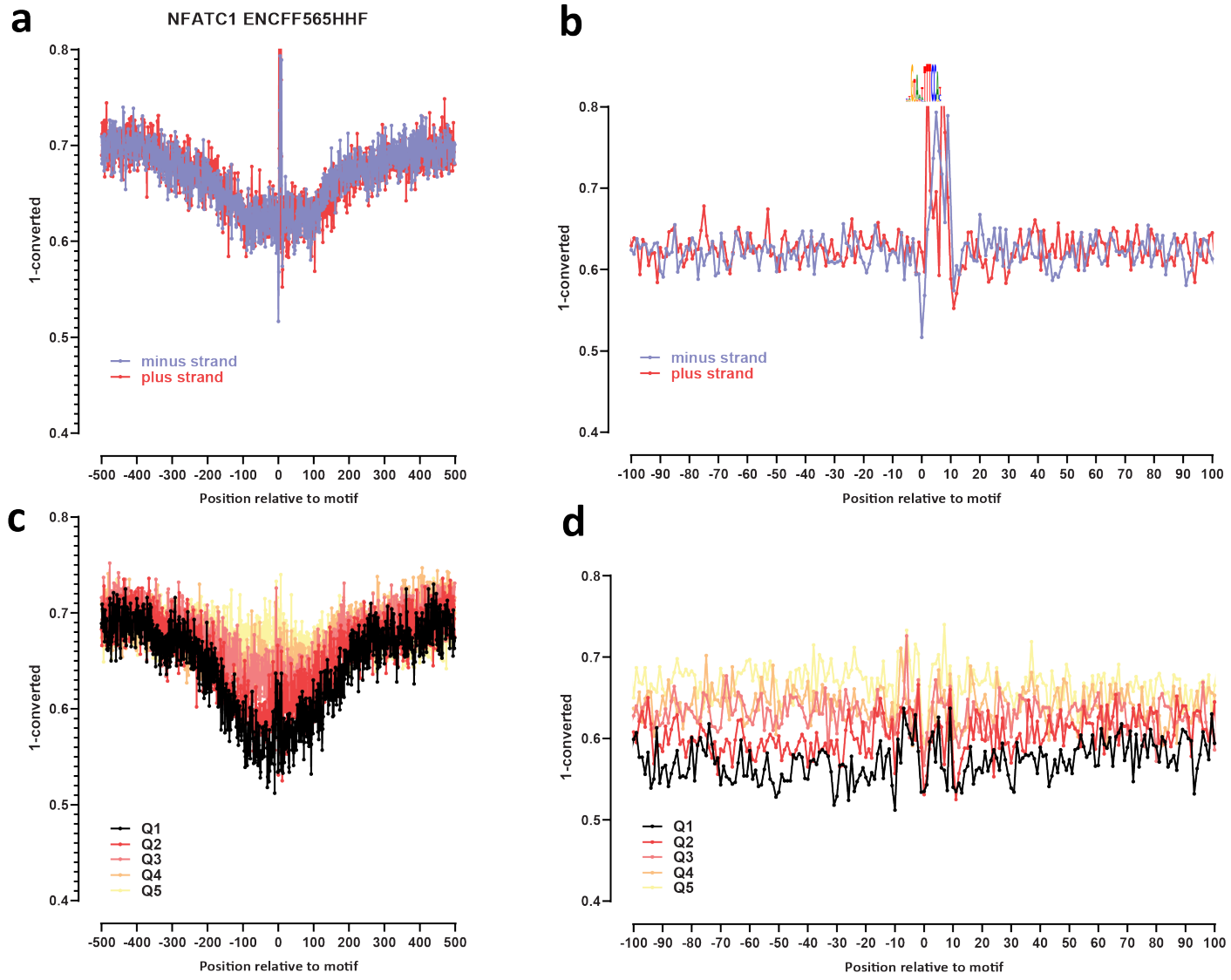
Supplementary Figure 39: Genome-wide footprint profiles for the MEF2C transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



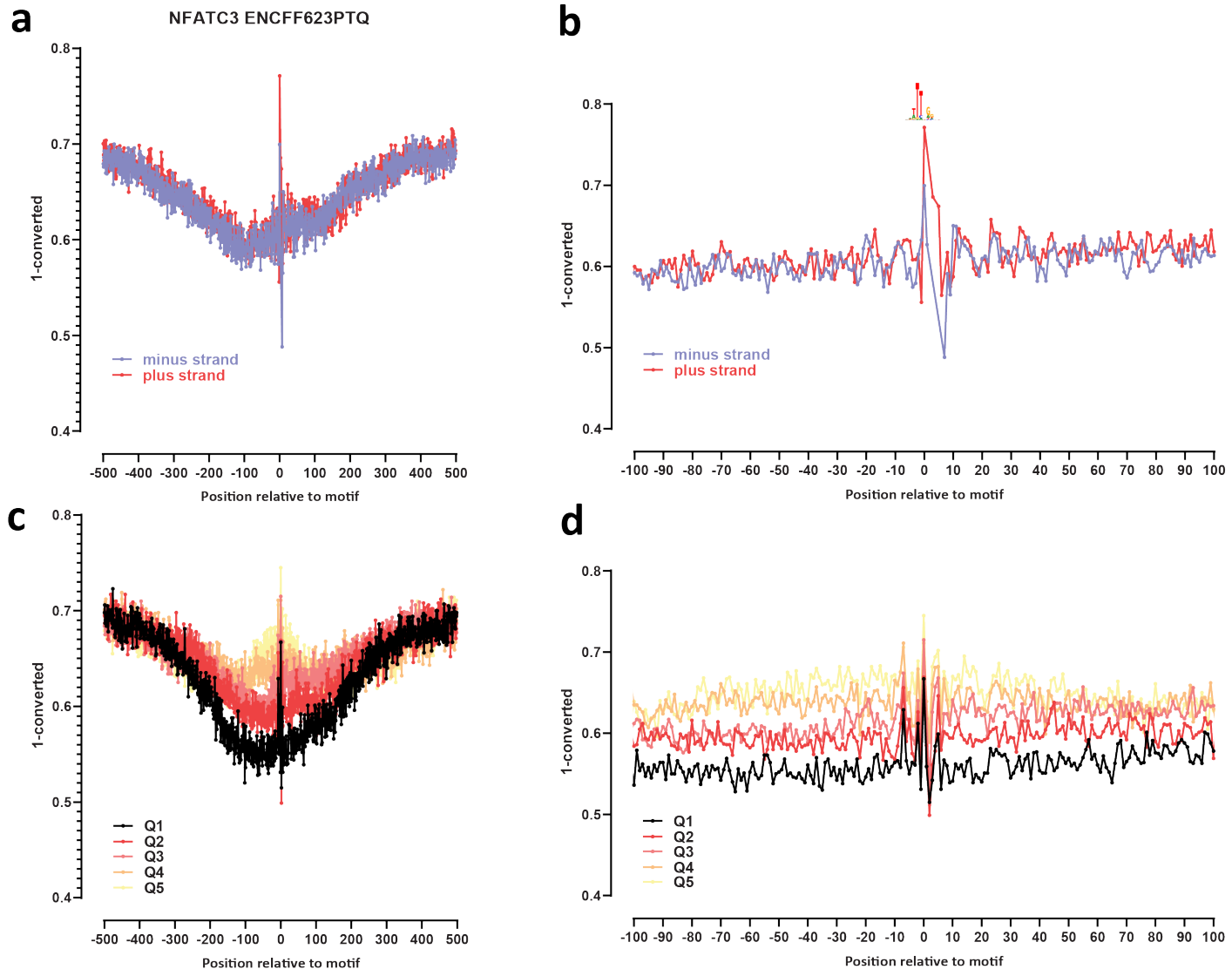
Supplementary Figure 40: Genome-wide footprint profiles for the MXI1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



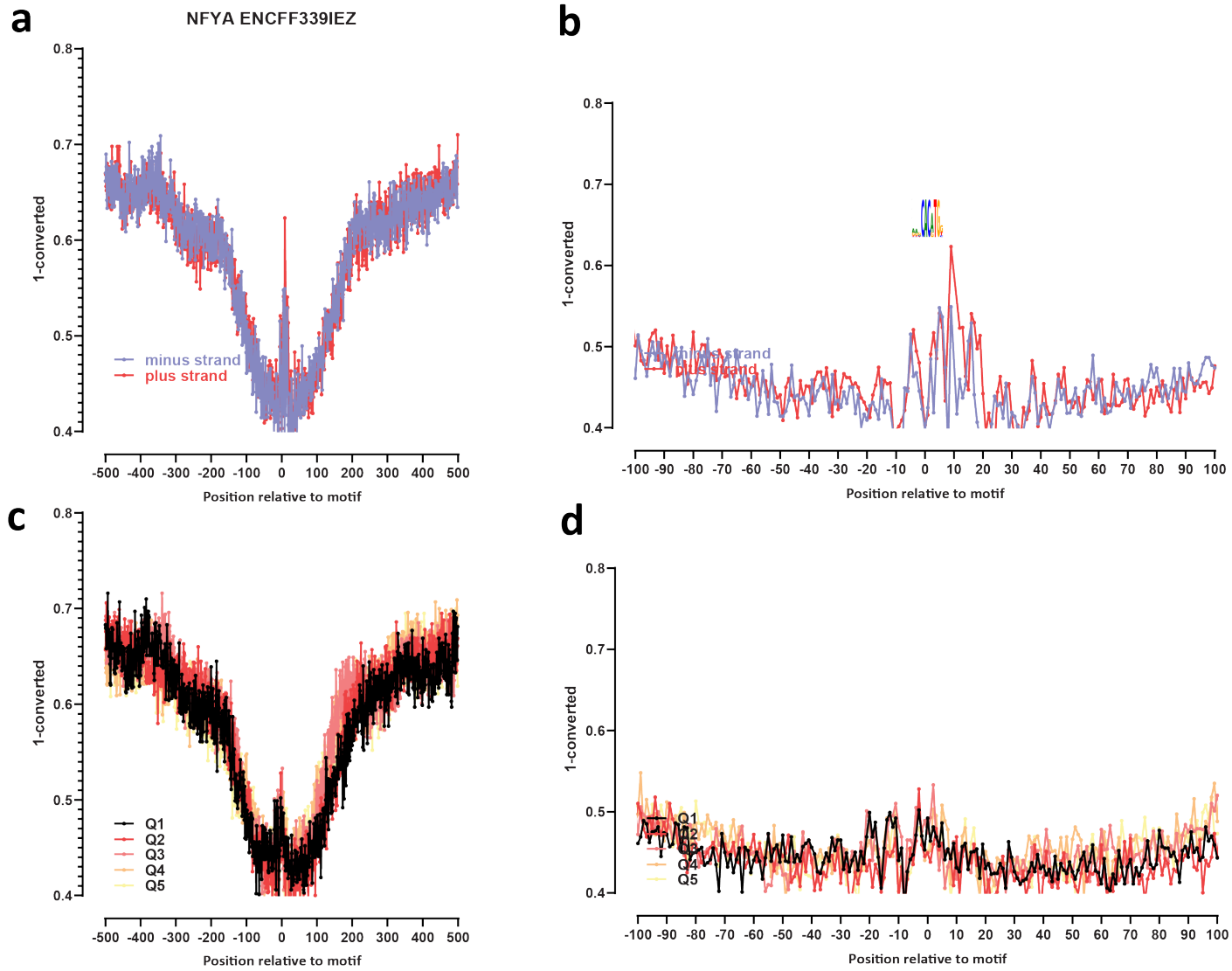
Supplementary Figure 41: Genome-wide footprint profiles for the MYC transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



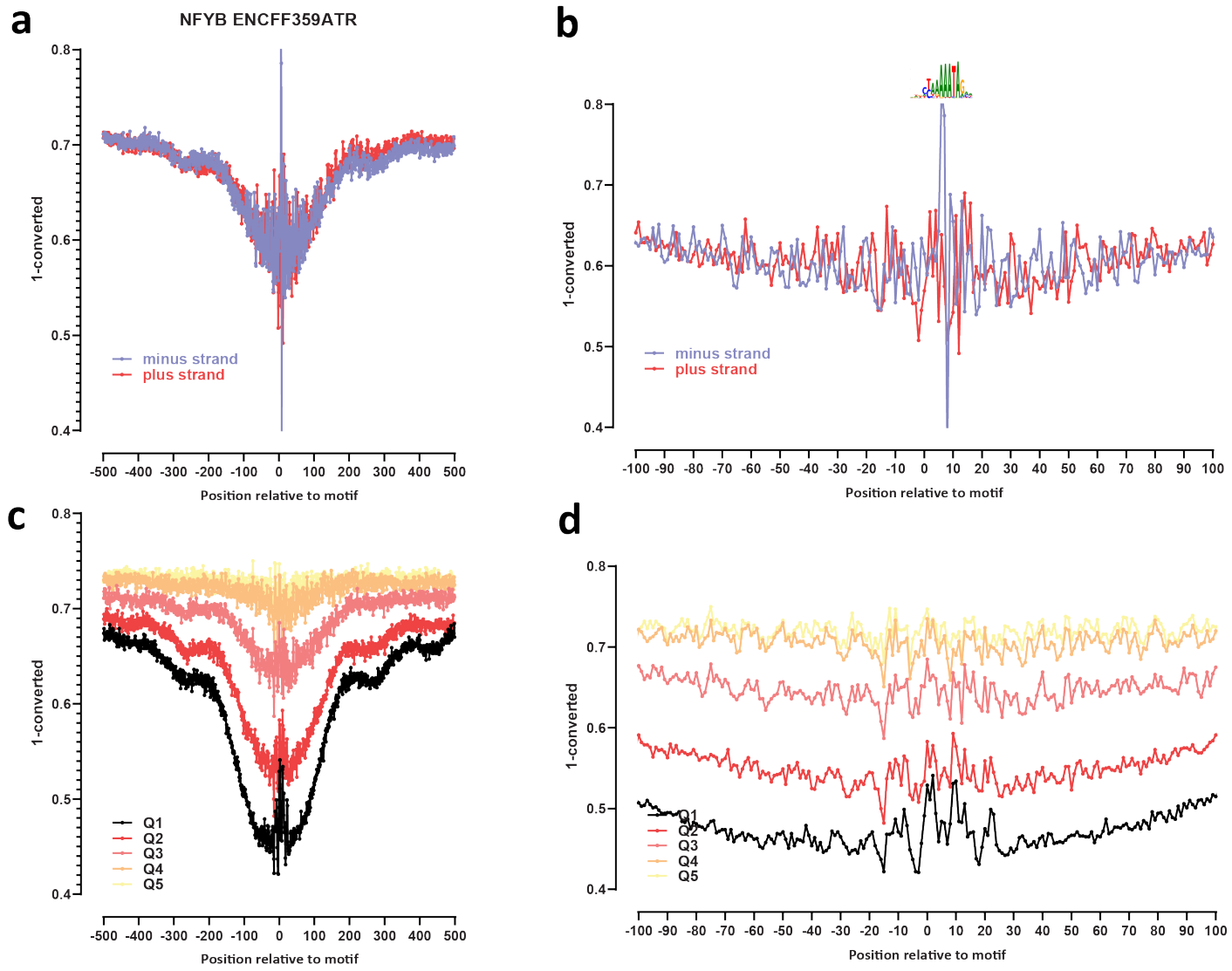
Supplementary Figure 42: Genome-wide footprint profiles for the NFATC1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



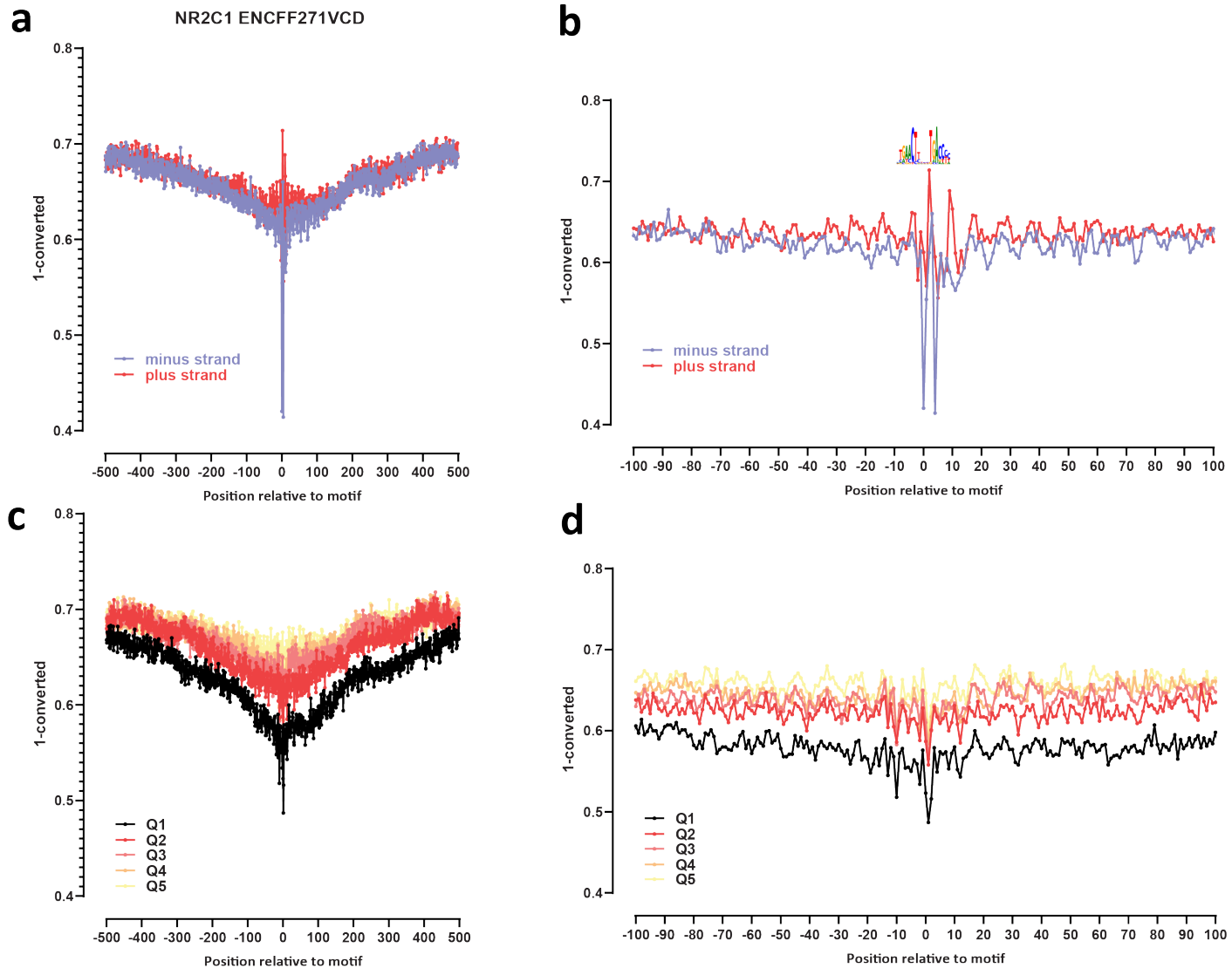
Supplementary Figure 43: Genome-wide footprint profiles for the NFATC3 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



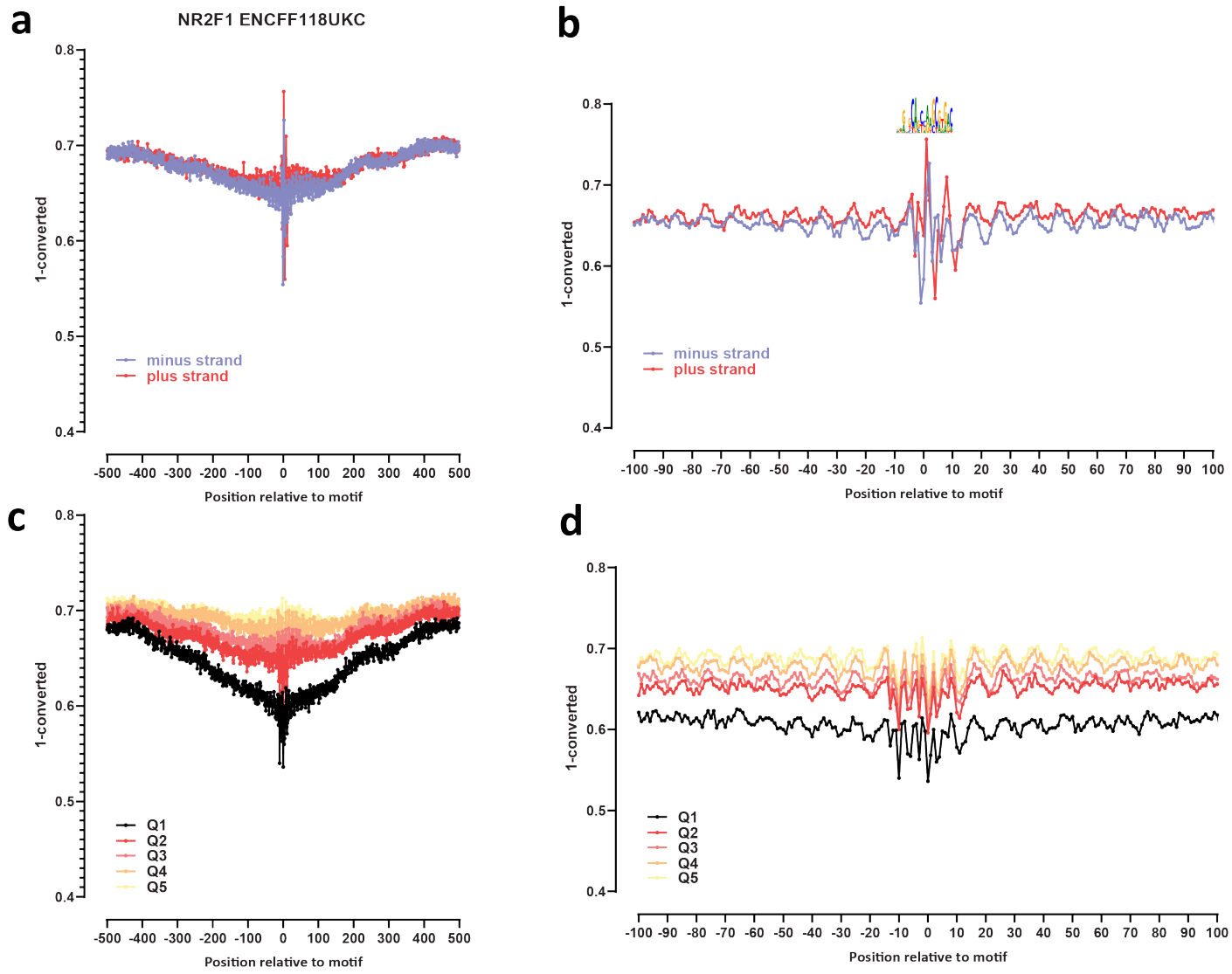
Supplementary Figure 44: Genome-wide footprint profiles for the NFYA transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



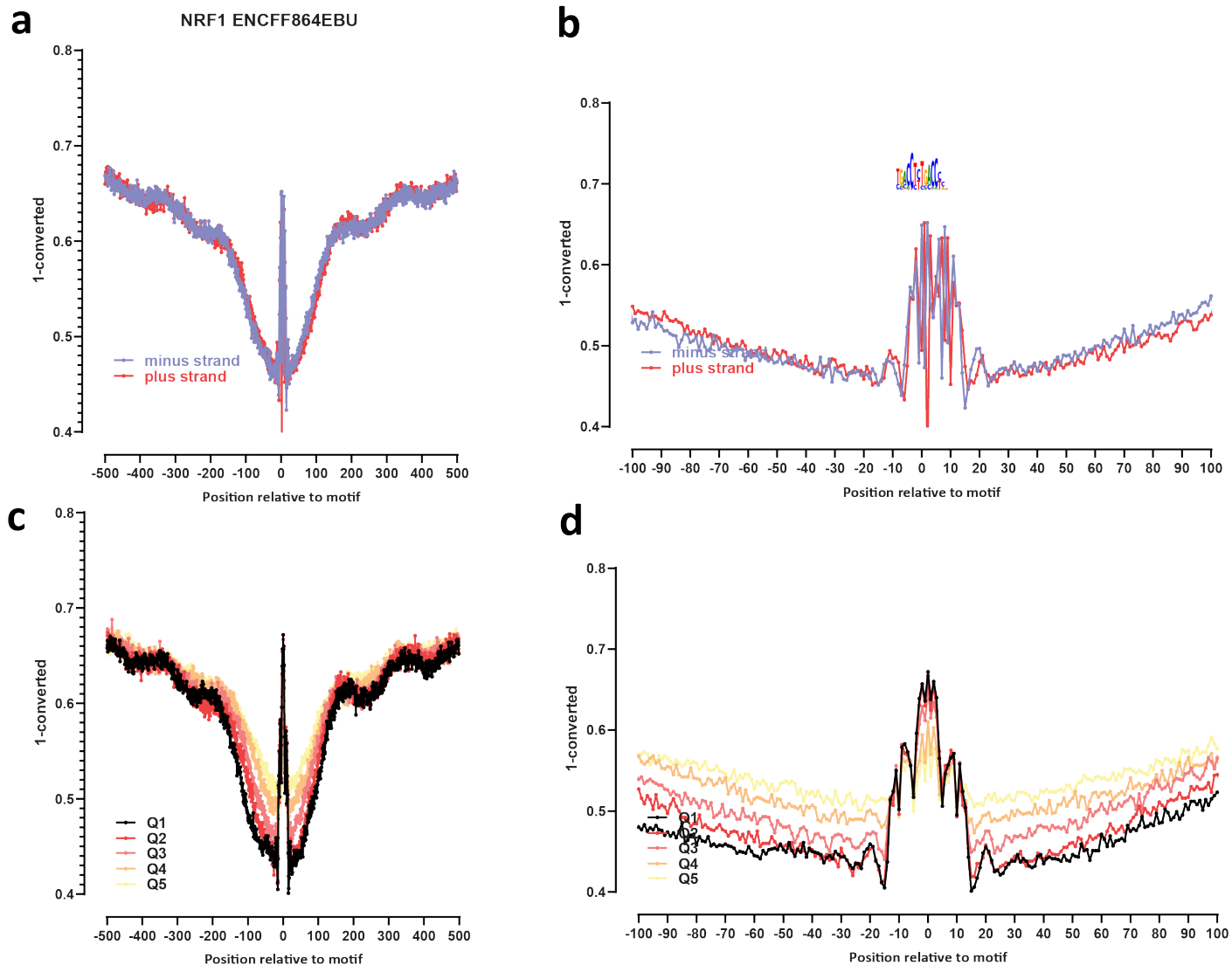
Supplementary Figure 45: Genome-wide footprint profiles for the NFYB transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



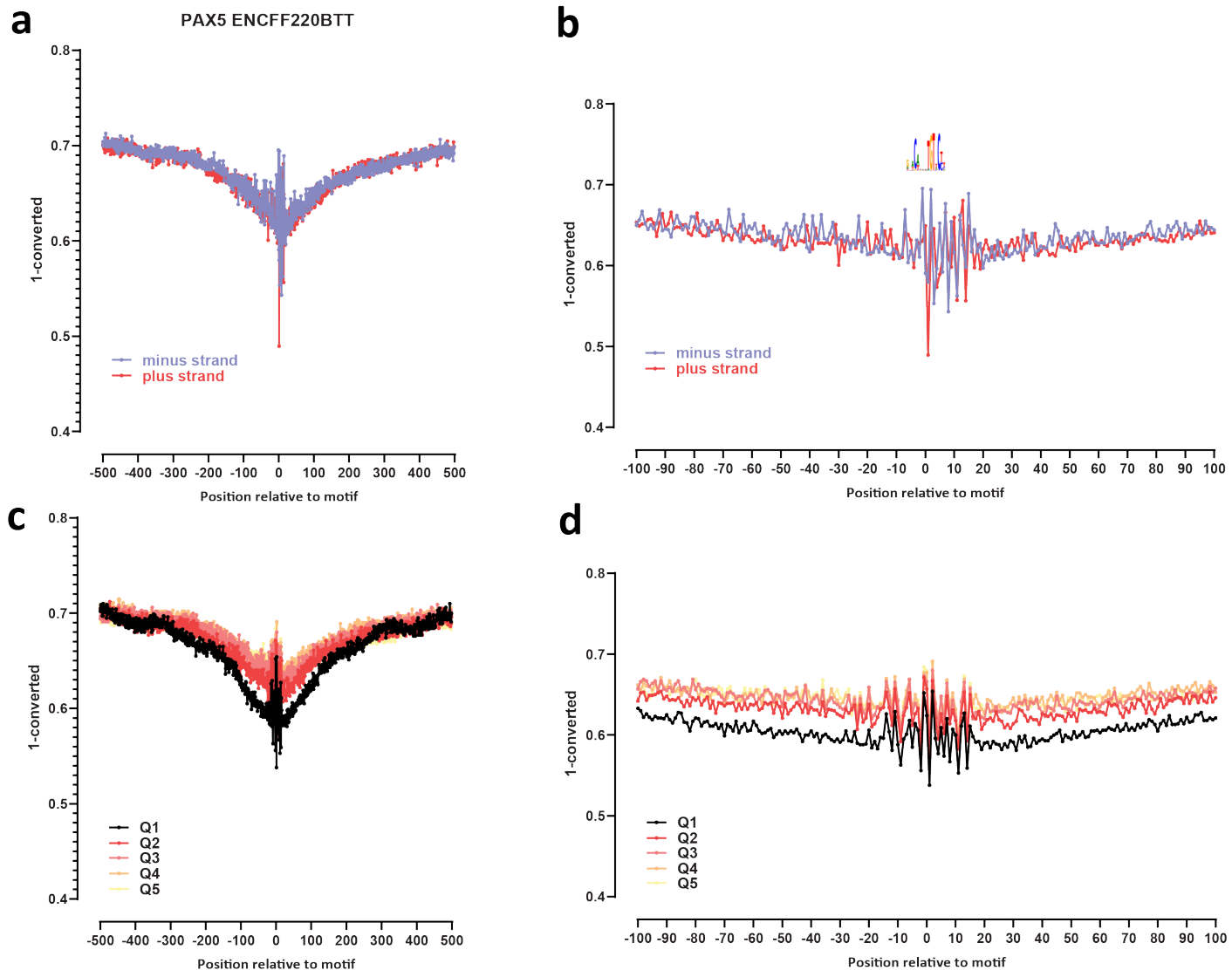
Supplementary Figure 46: Genome-wide footprint profiles for the NR2C1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



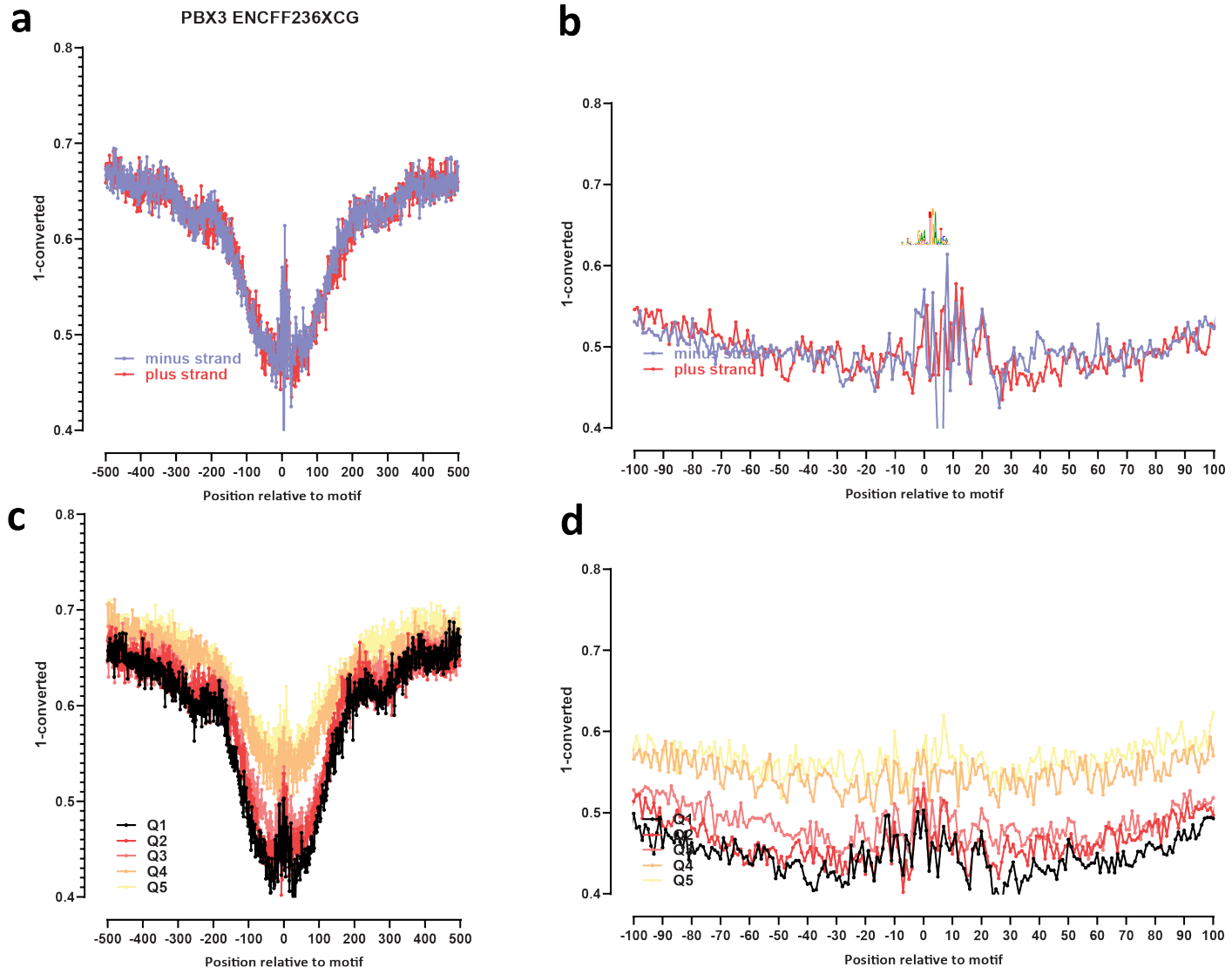
Supplementary Figure 47: Genome-wide footprint profiles for the NR2F1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



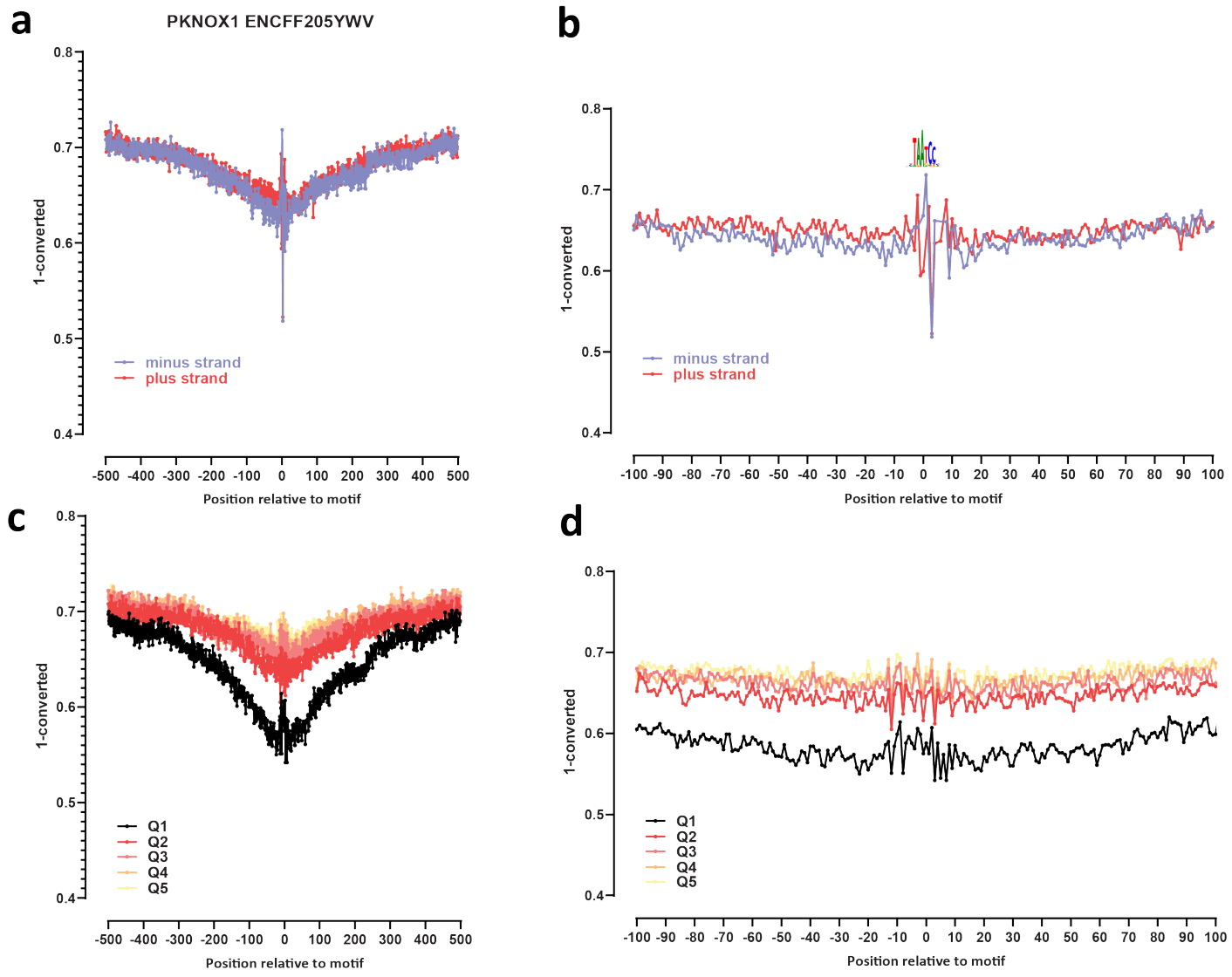
Supplementary Figure 48: Genome-wide footprint profiles for the NRF1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



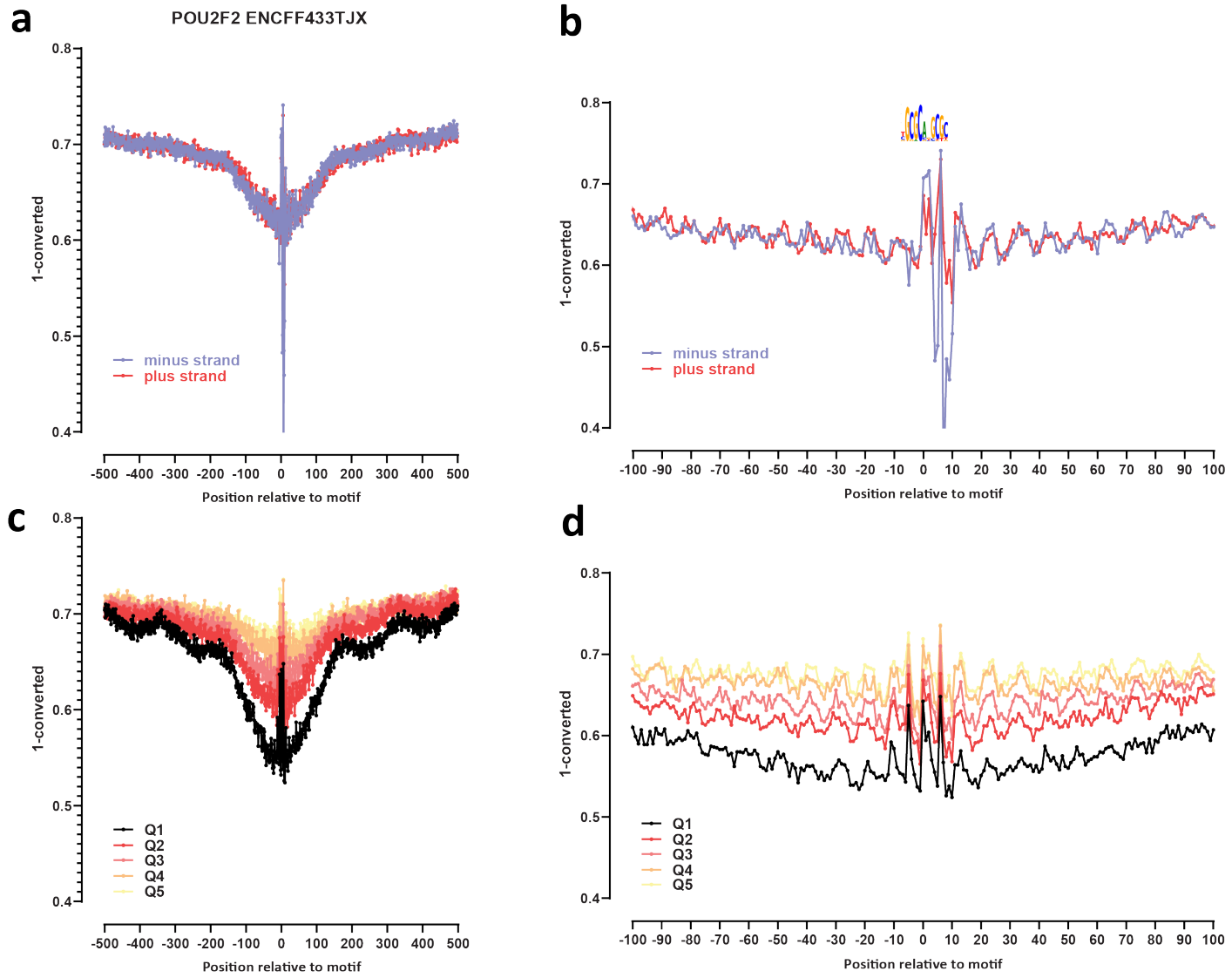
Supplementary Figure 49: Genome-wide footprint profiles for the PAX5 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



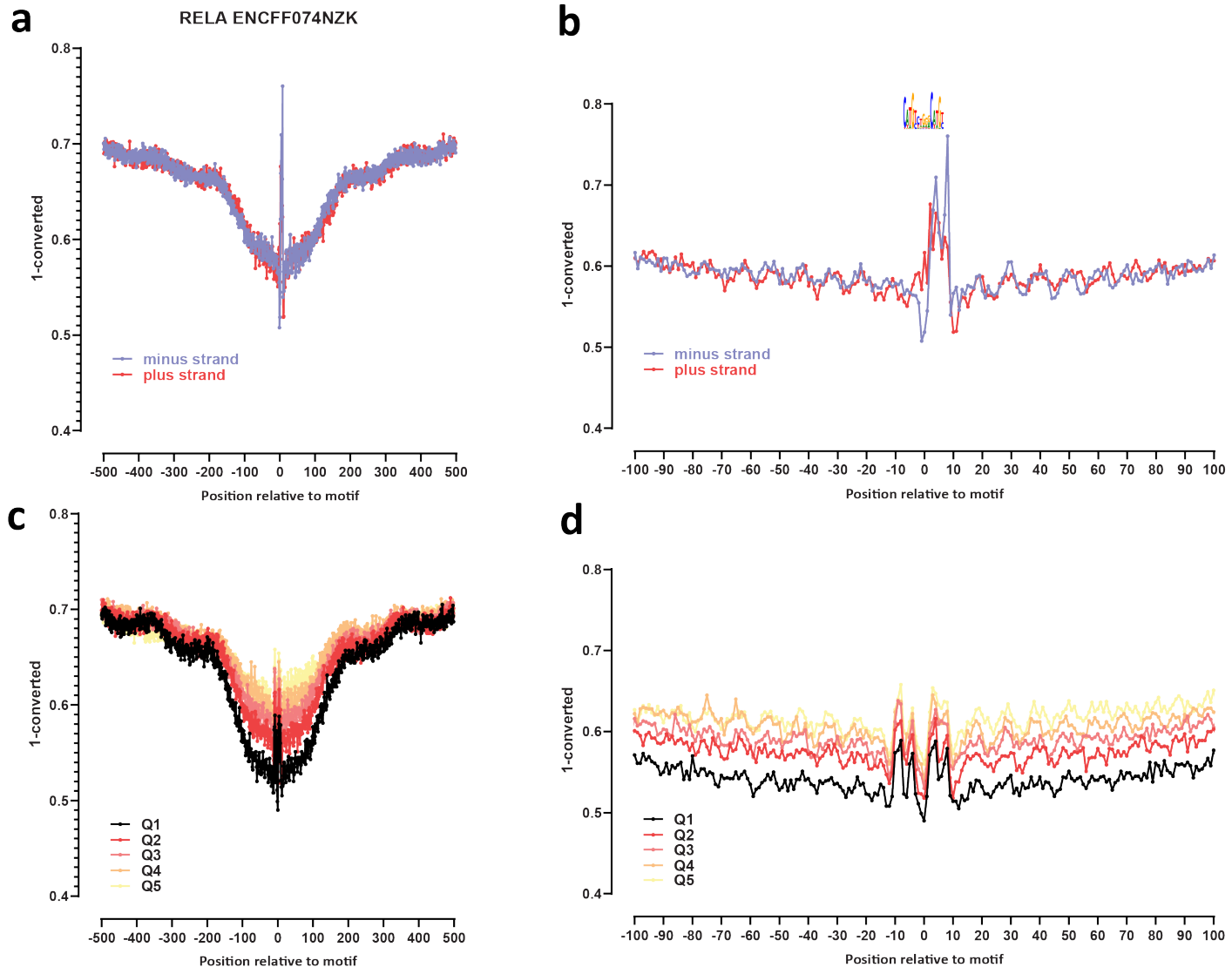
Supplementary Figure 50: Genome-wide footprint profiles for the PBX3 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



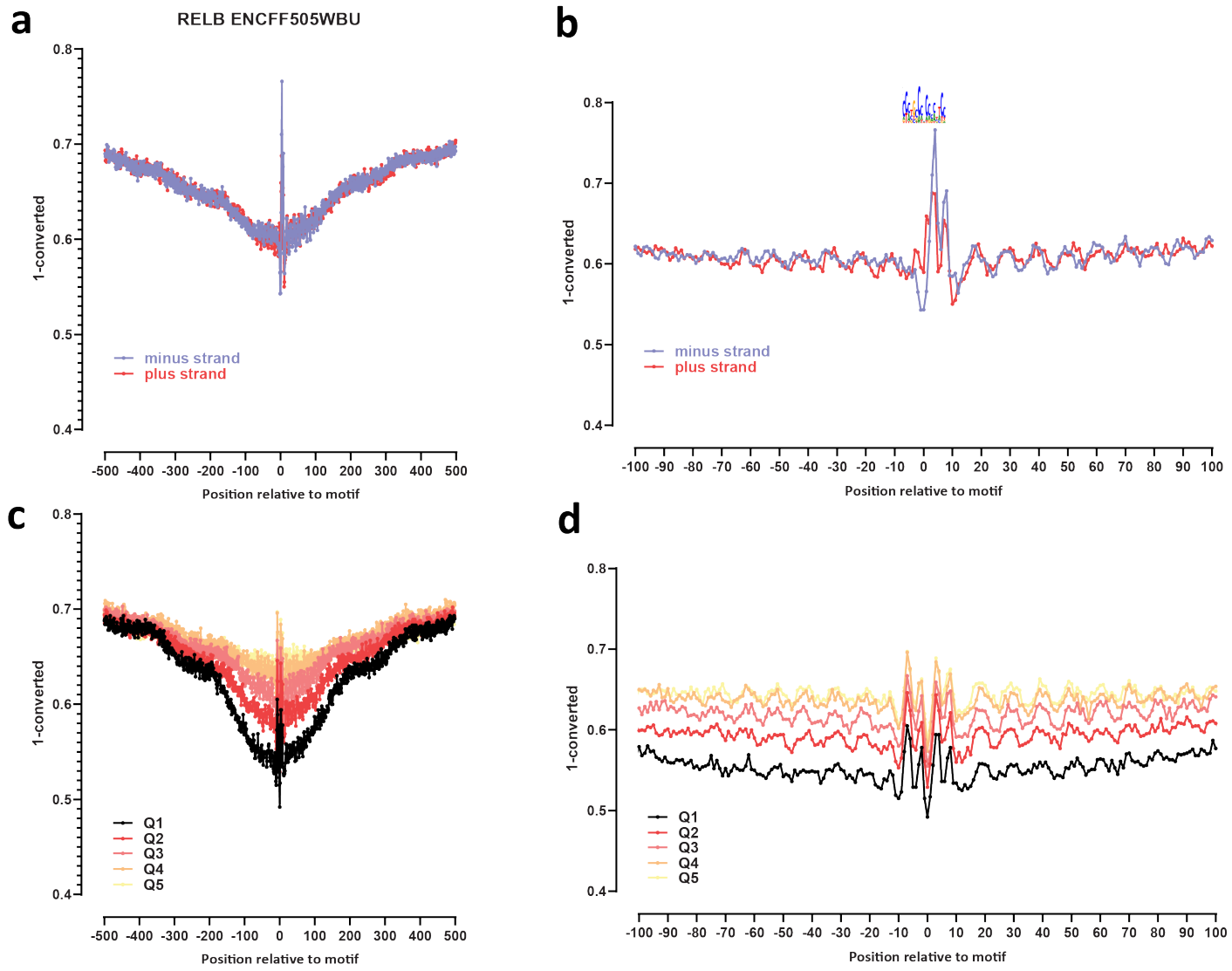
Supplementary Figure 51: Genome-wide footprint profiles for the PKNOX1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



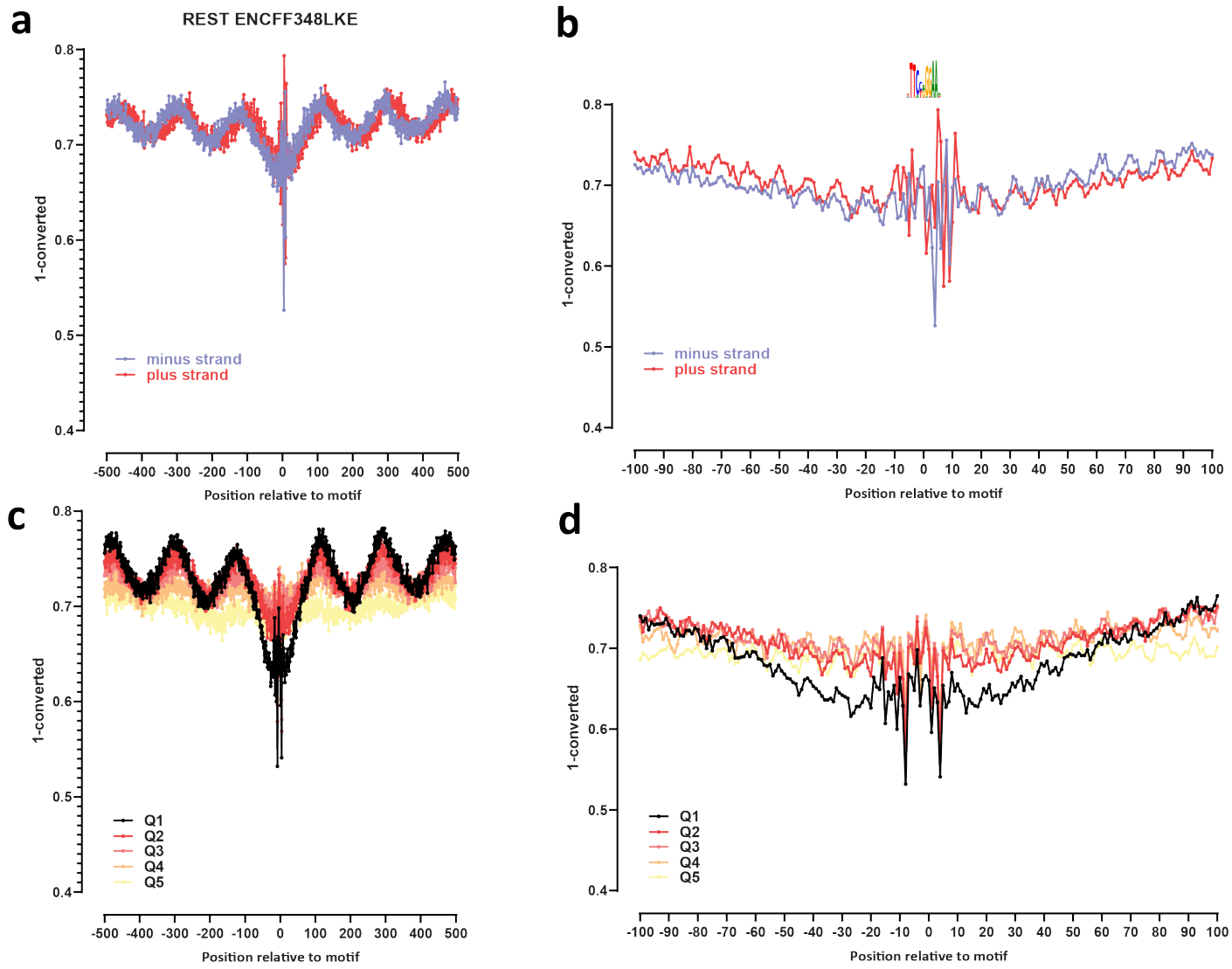
Supplementary Figure 52: Genome-wide footprint profiles for the POU2F2 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



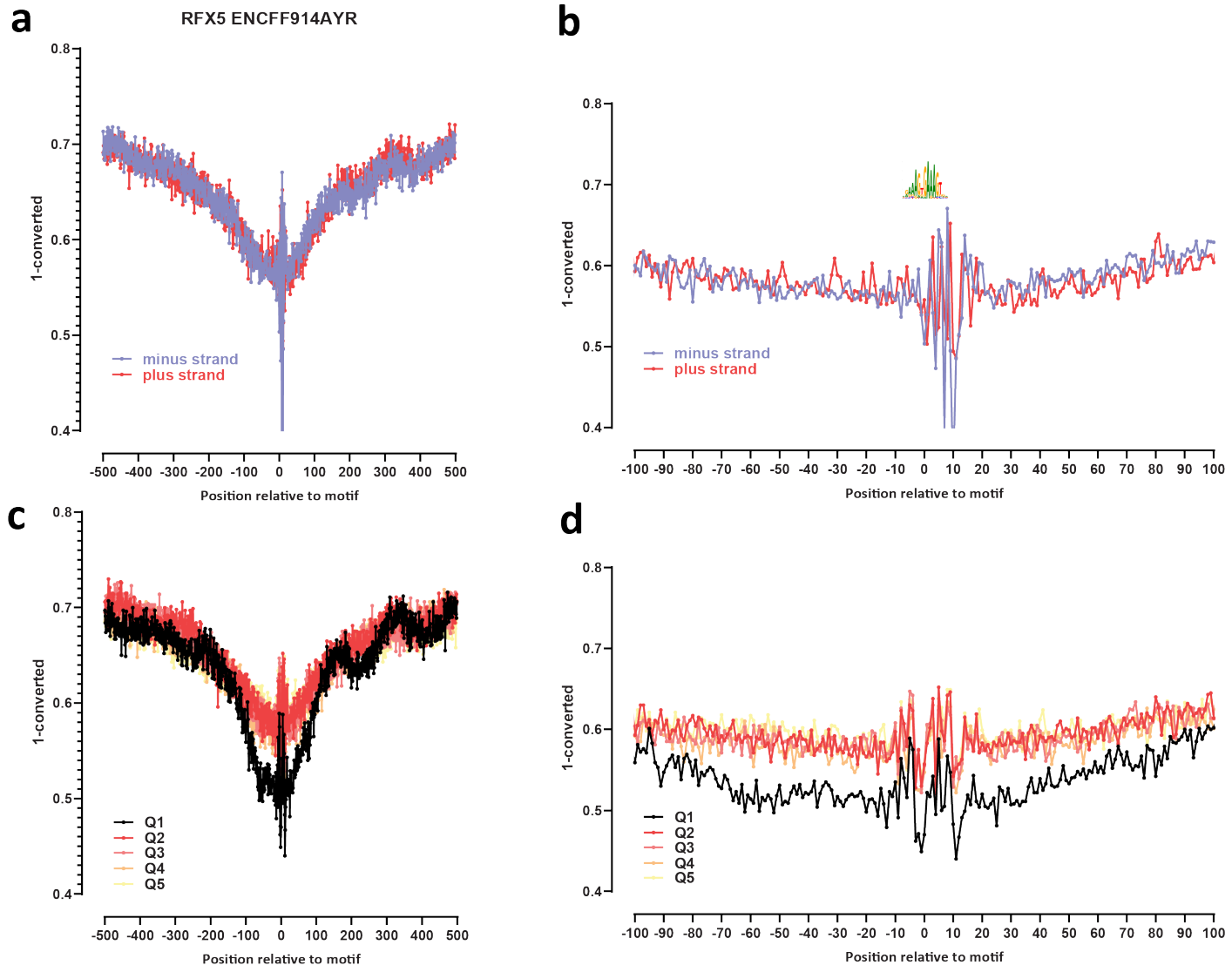
Supplementary Figure 53: Genome-wide footprint profiles for the RELA transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



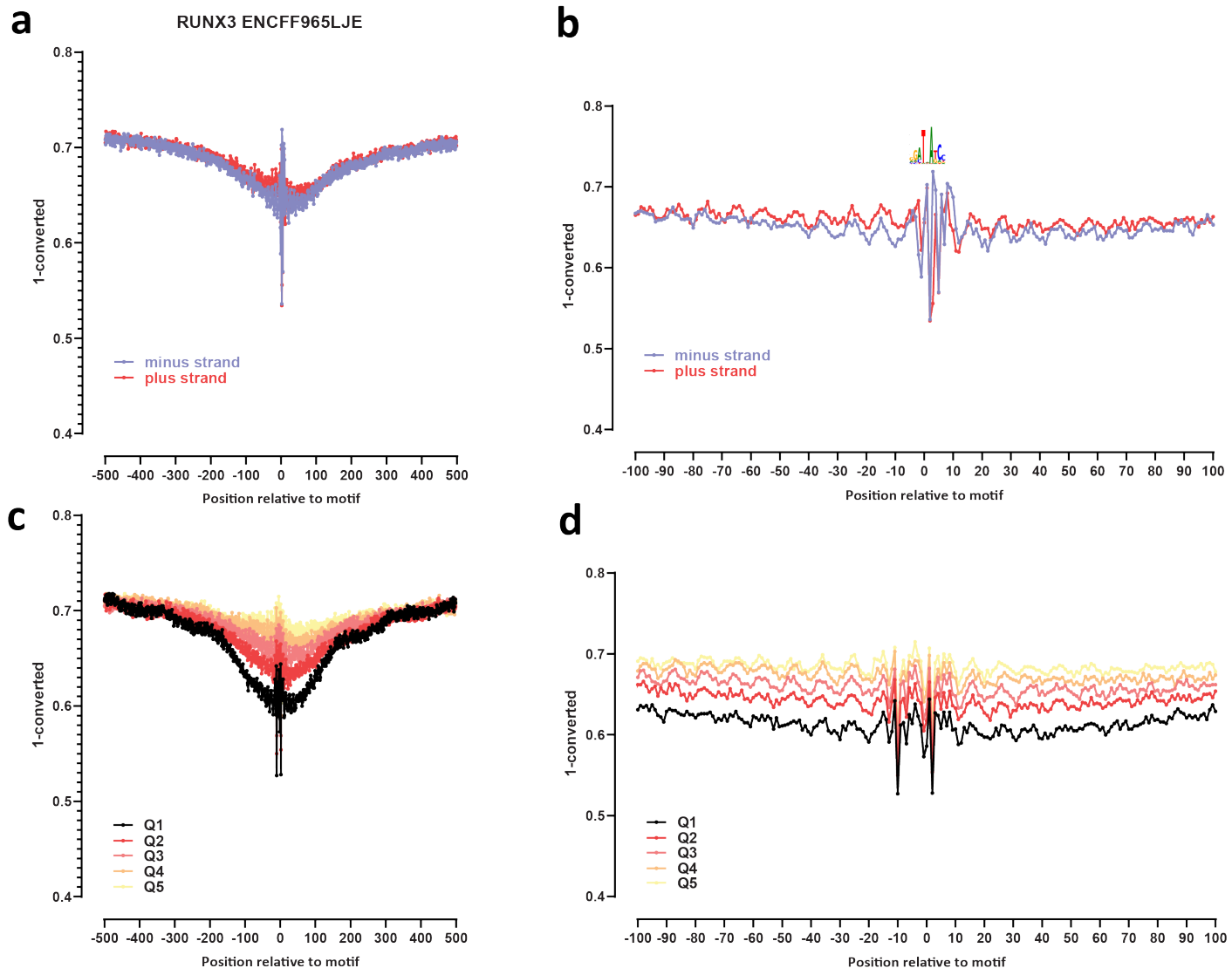
Supplementary Figure 54: Genome-wide footprint profiles for the RELB transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



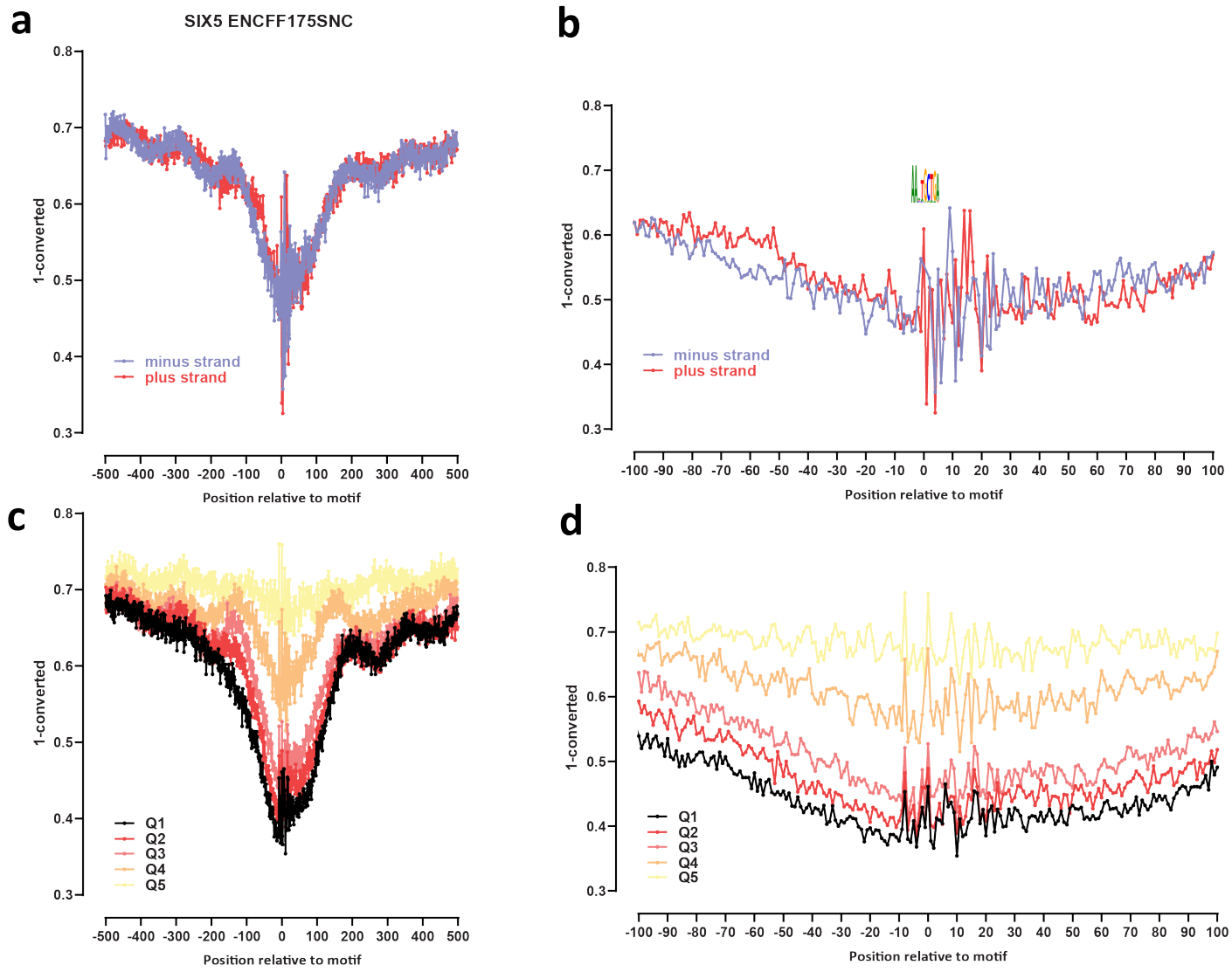
Supplementary Figure 55: Genome-wide footprint profiles for the REST/NRSF transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



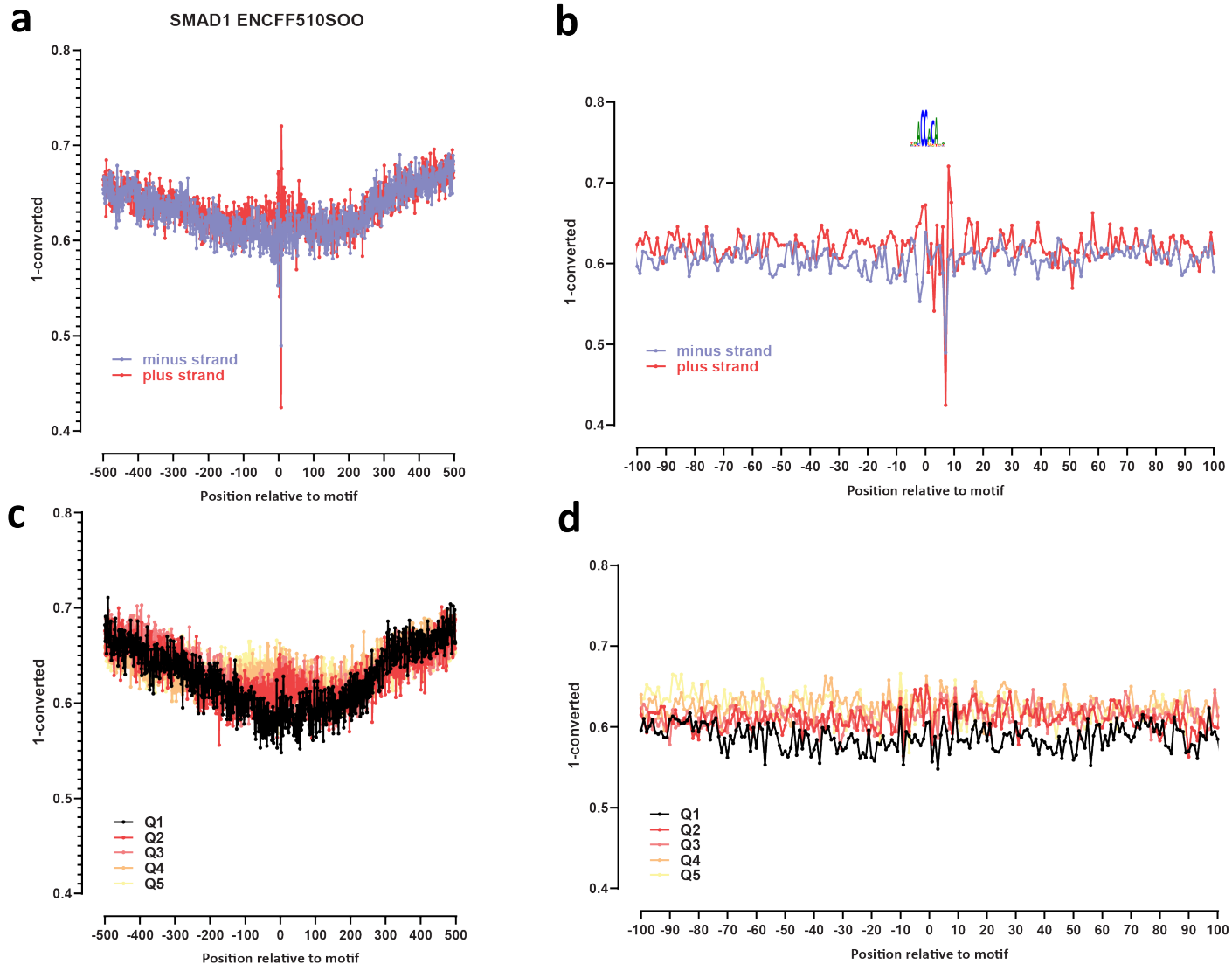
Supplementary Figure 56: Genome-wide footprint profiles for the RFX5 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



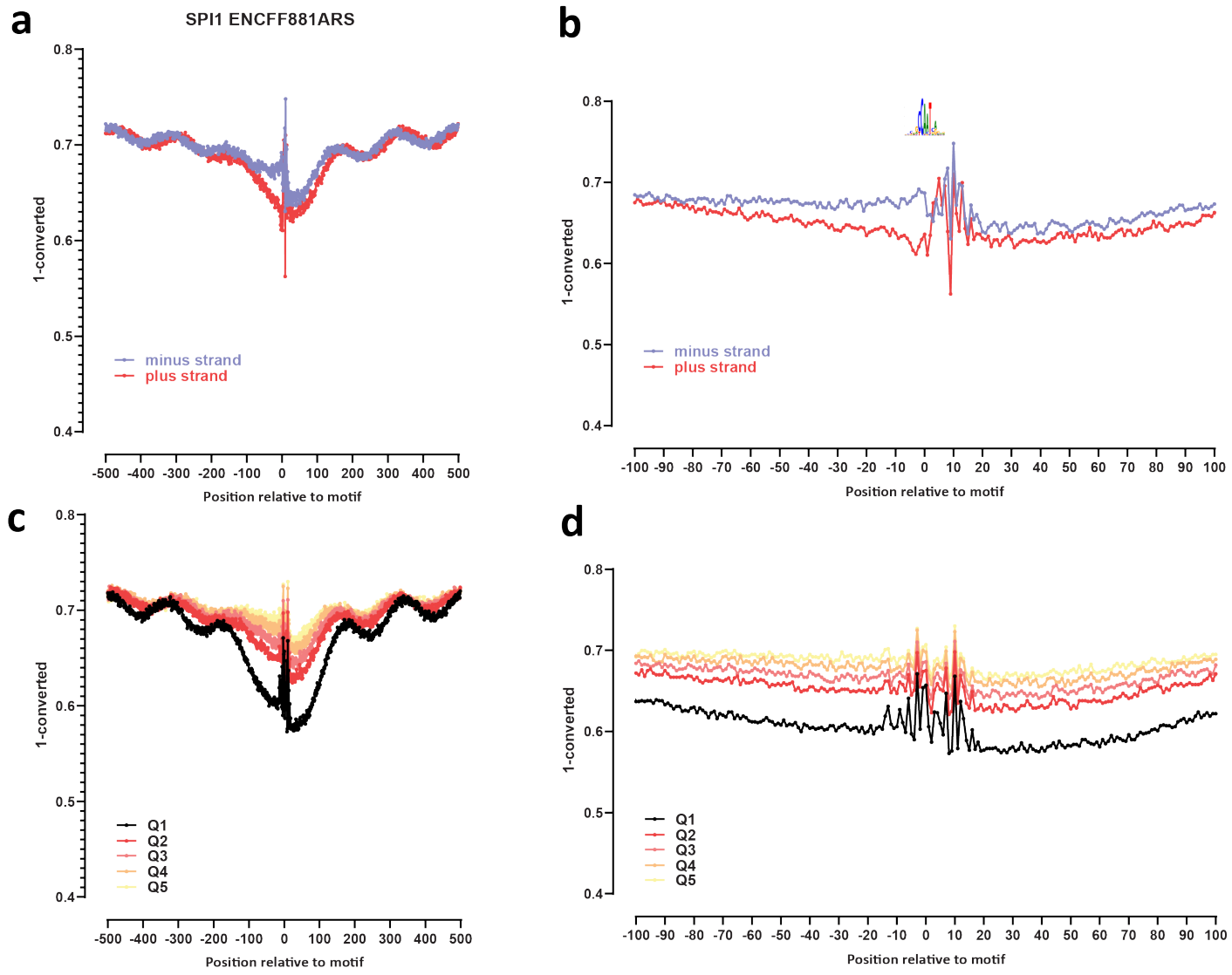
Supplementary Figure 57: Genome-wide footprint profiles for the RUNX3 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



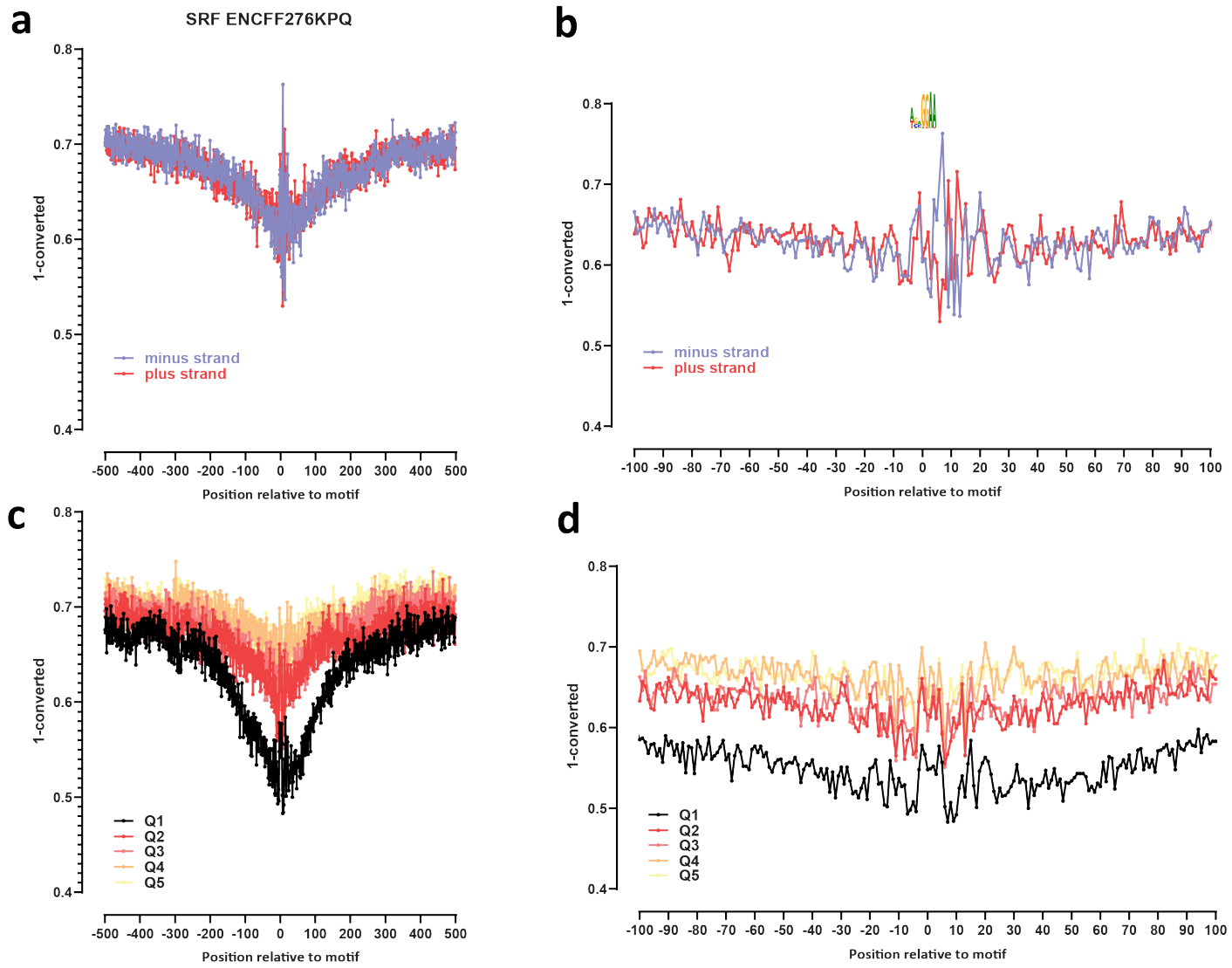
Supplementary Figure 58: Genome-wide footprint profiles for the SIX5 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



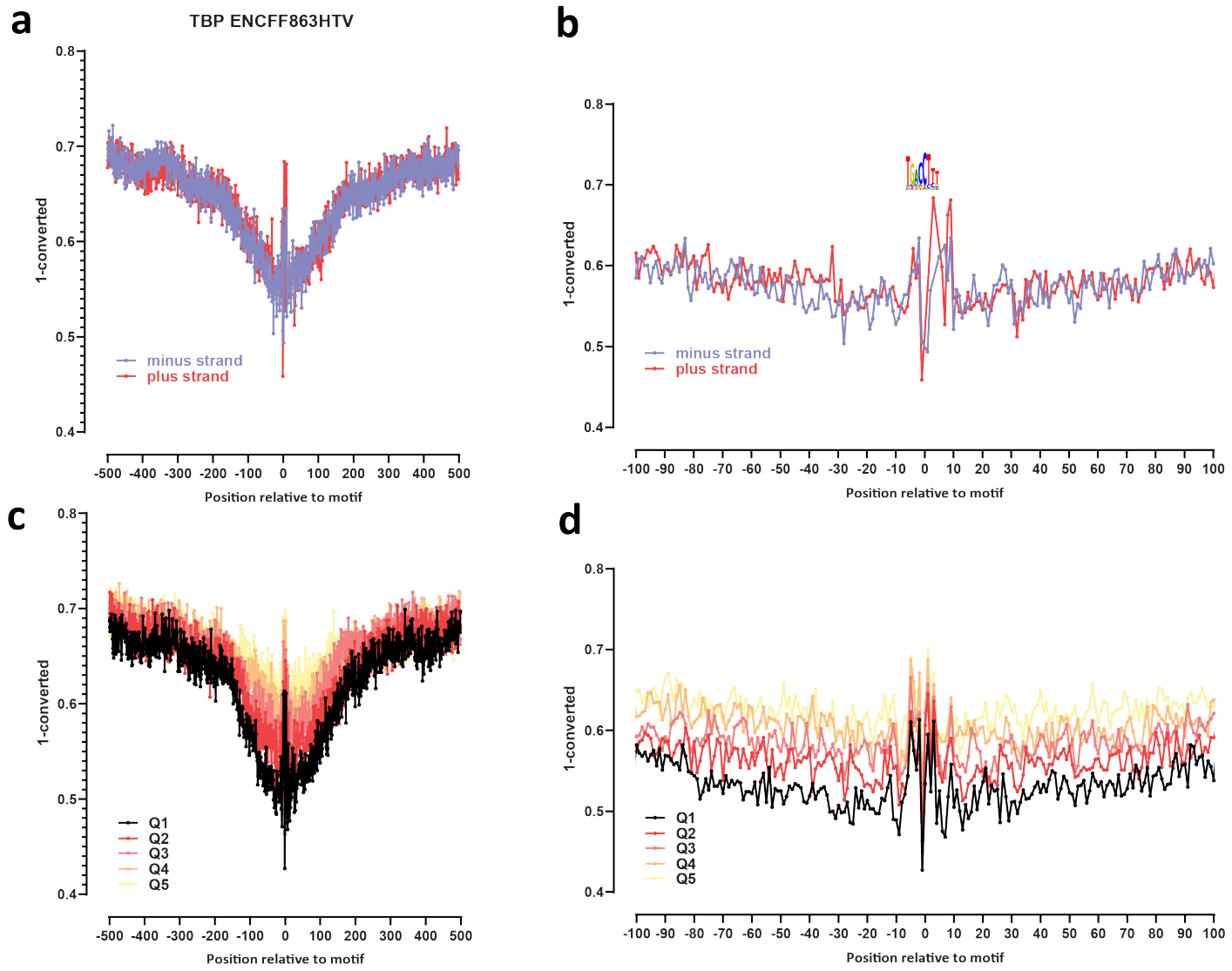
Supplementary Figure 59: Genome-wide footprint profiles for the SMAD1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



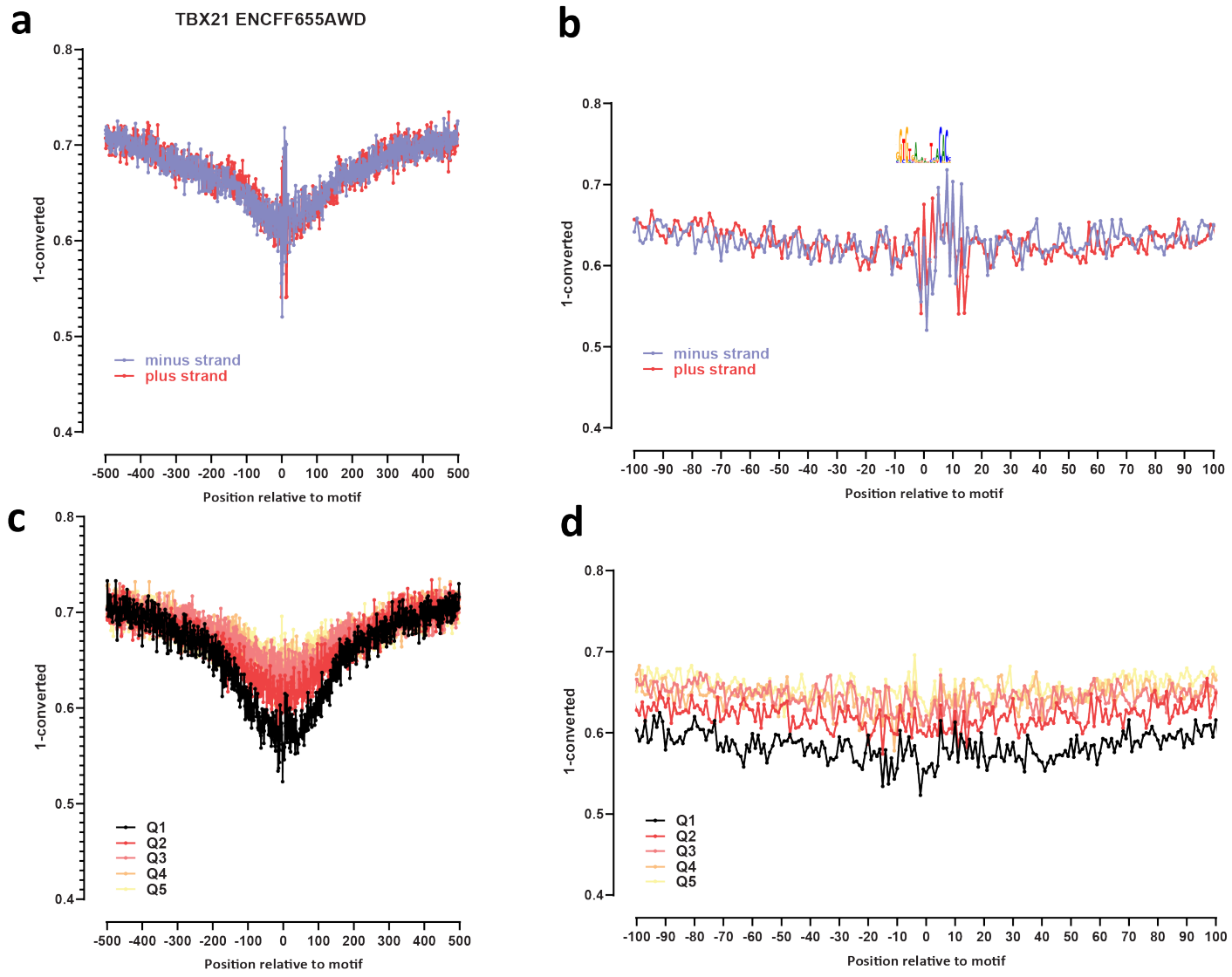
Supplementary Figure 60: Genome-wide footprint profiles for the SPI1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



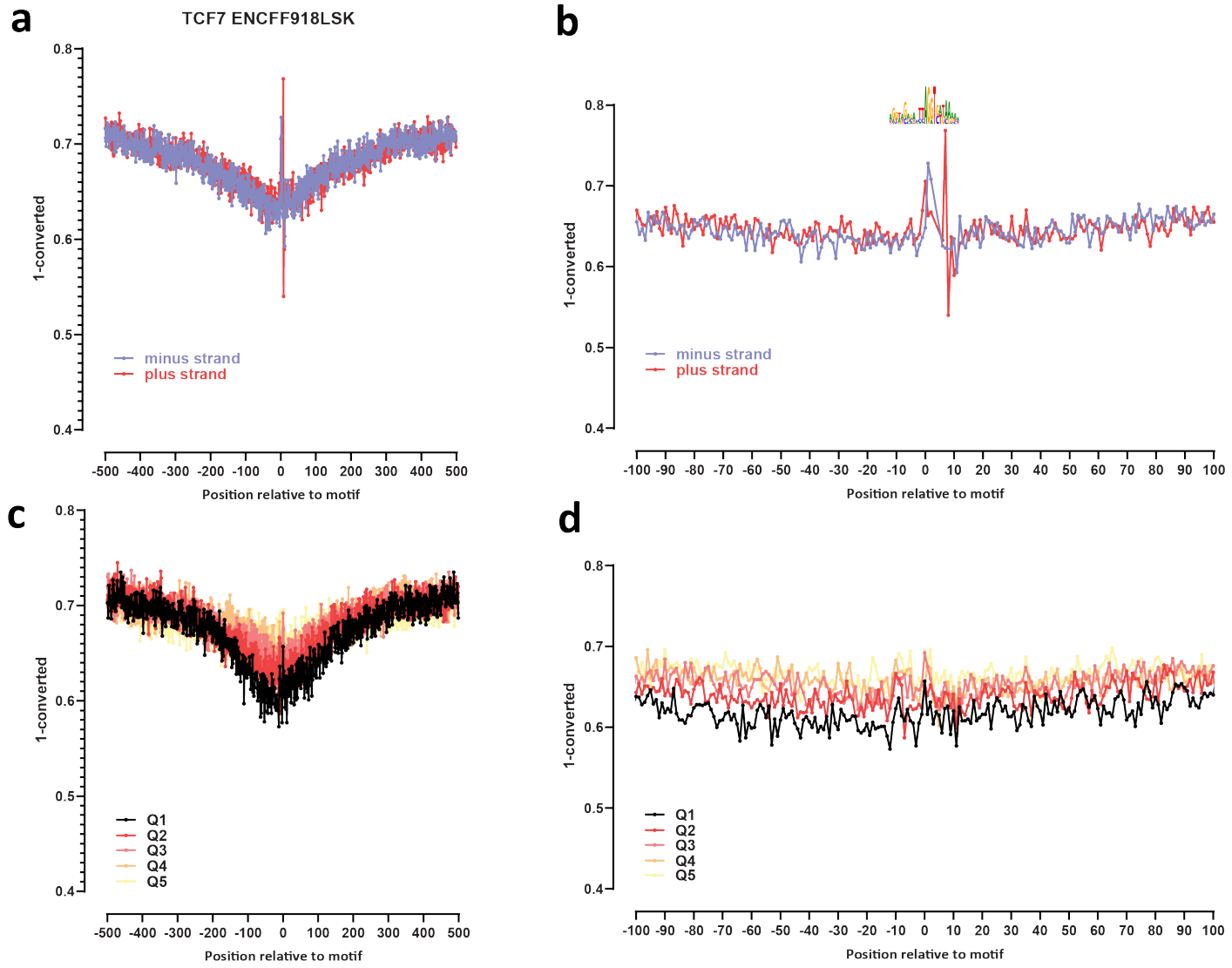
Supplementary Figure 61: Genome-wide footprint profiles for the SRF transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



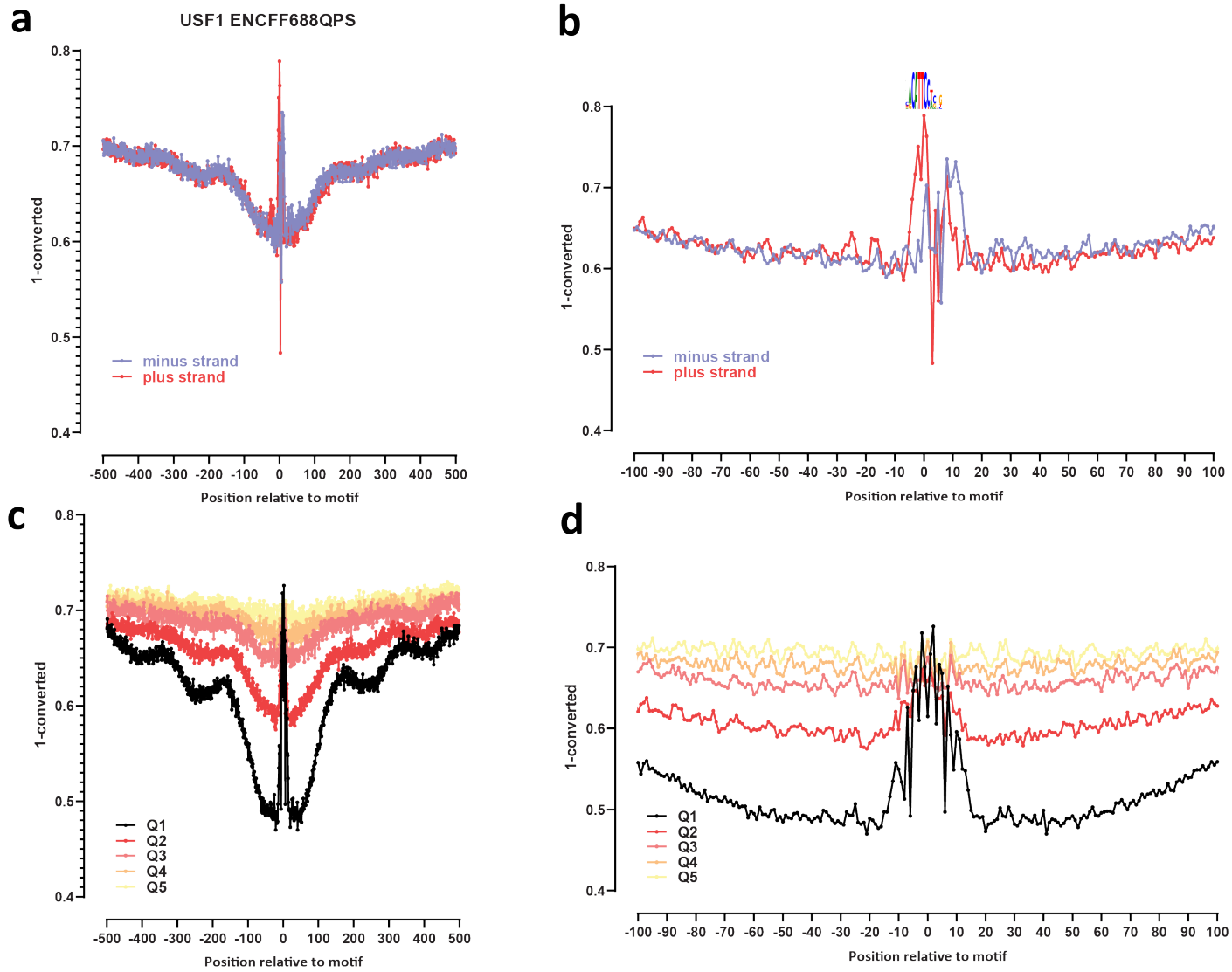
Supplementary Figure 62: Genome-wide footprint profiles for the TBP transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



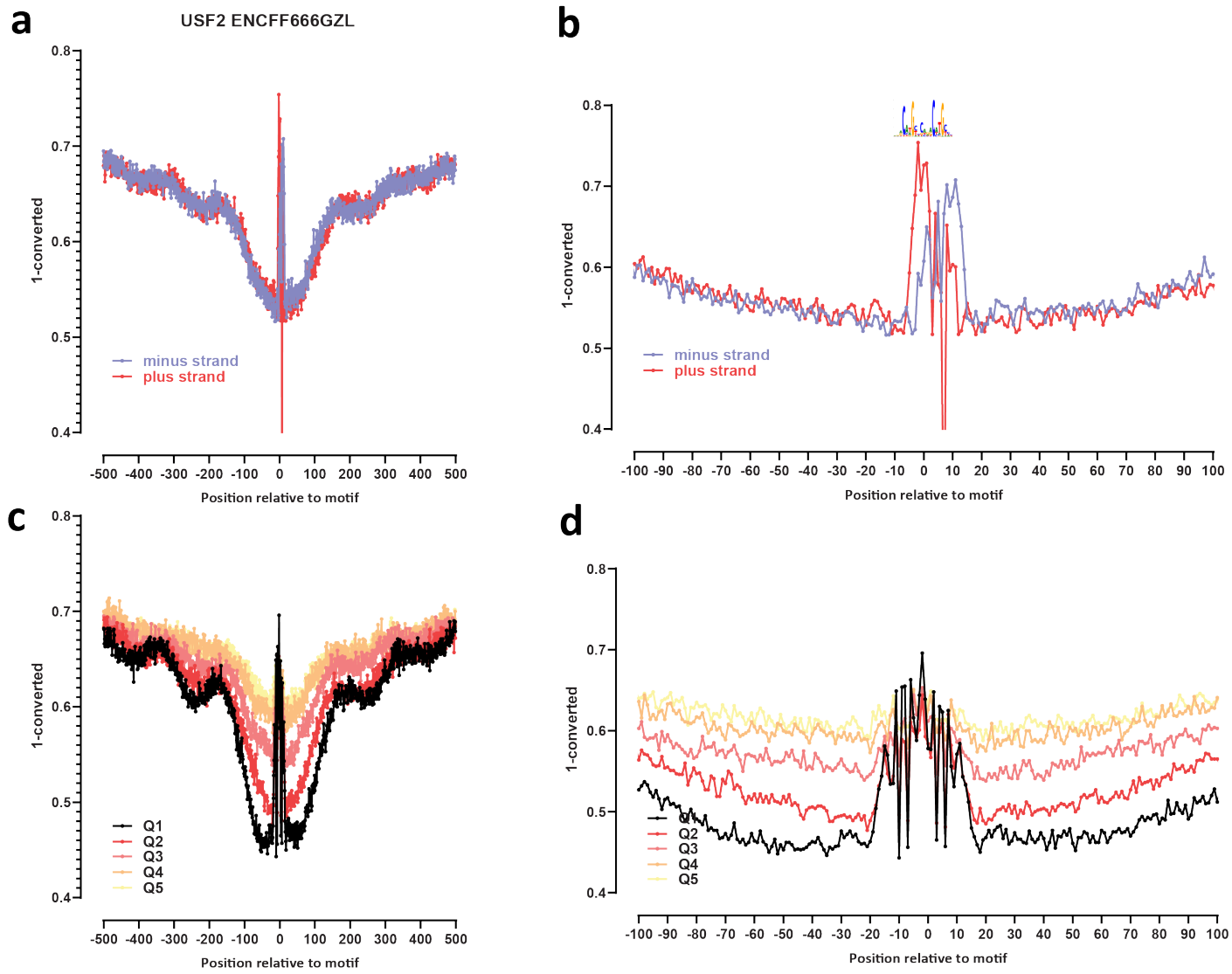
Supplementary Figure 63: Genome-wide footprint profiles for the TBX21 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



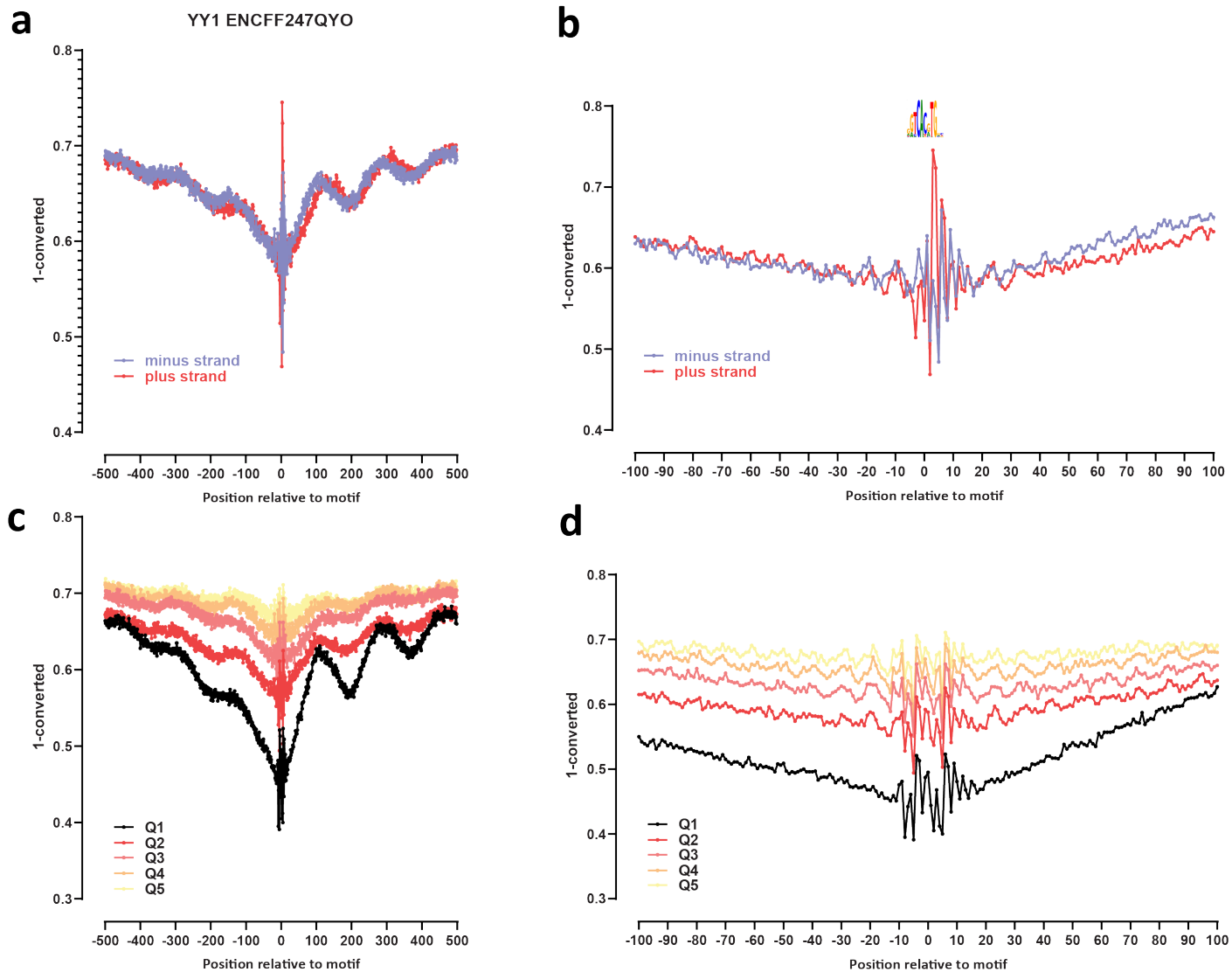
Supplementary Figure 64: Genome-wide footprint profiles for the TCF7 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



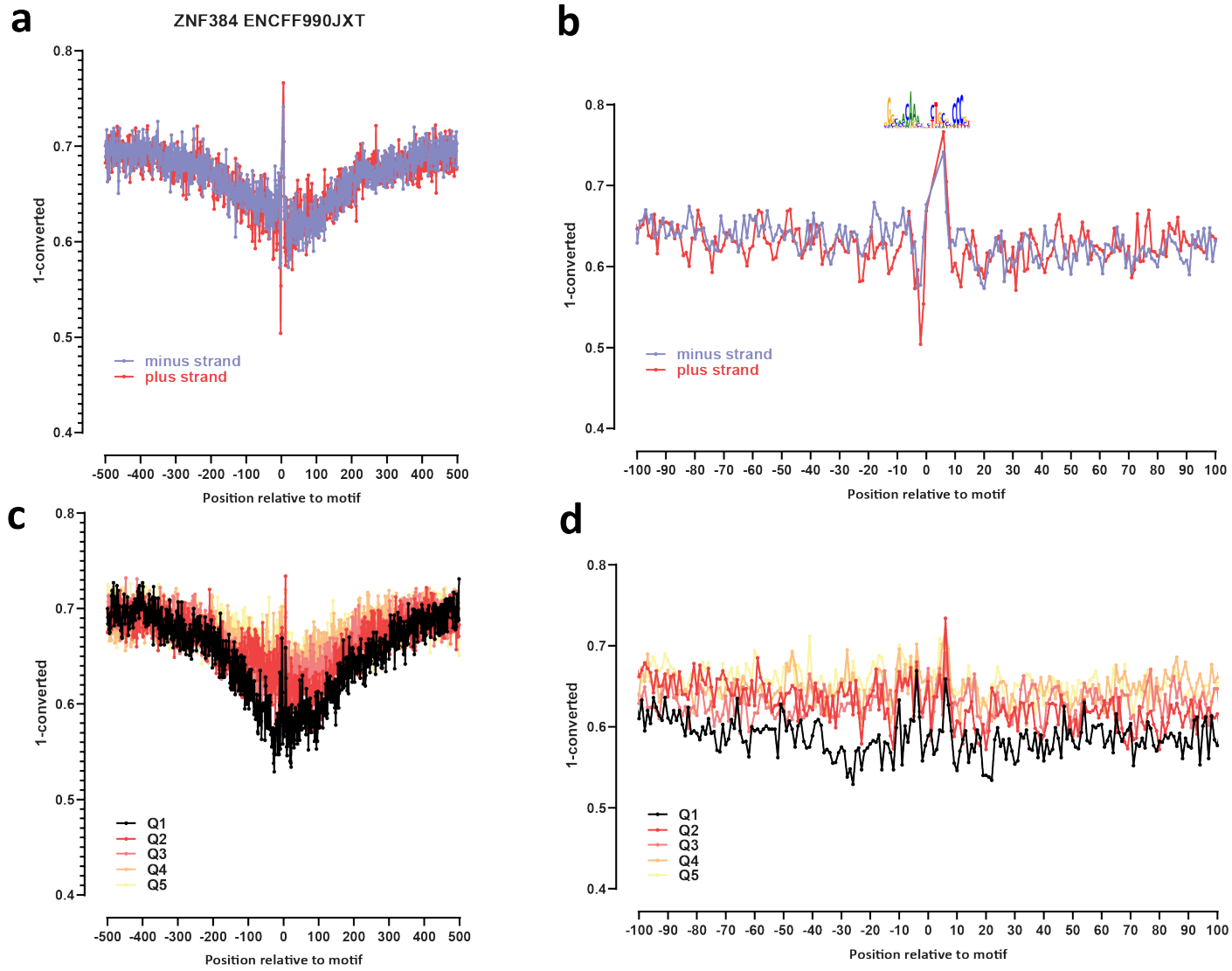
Supplementary Figure 65: Genome-wide footprint profiles for the USF1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



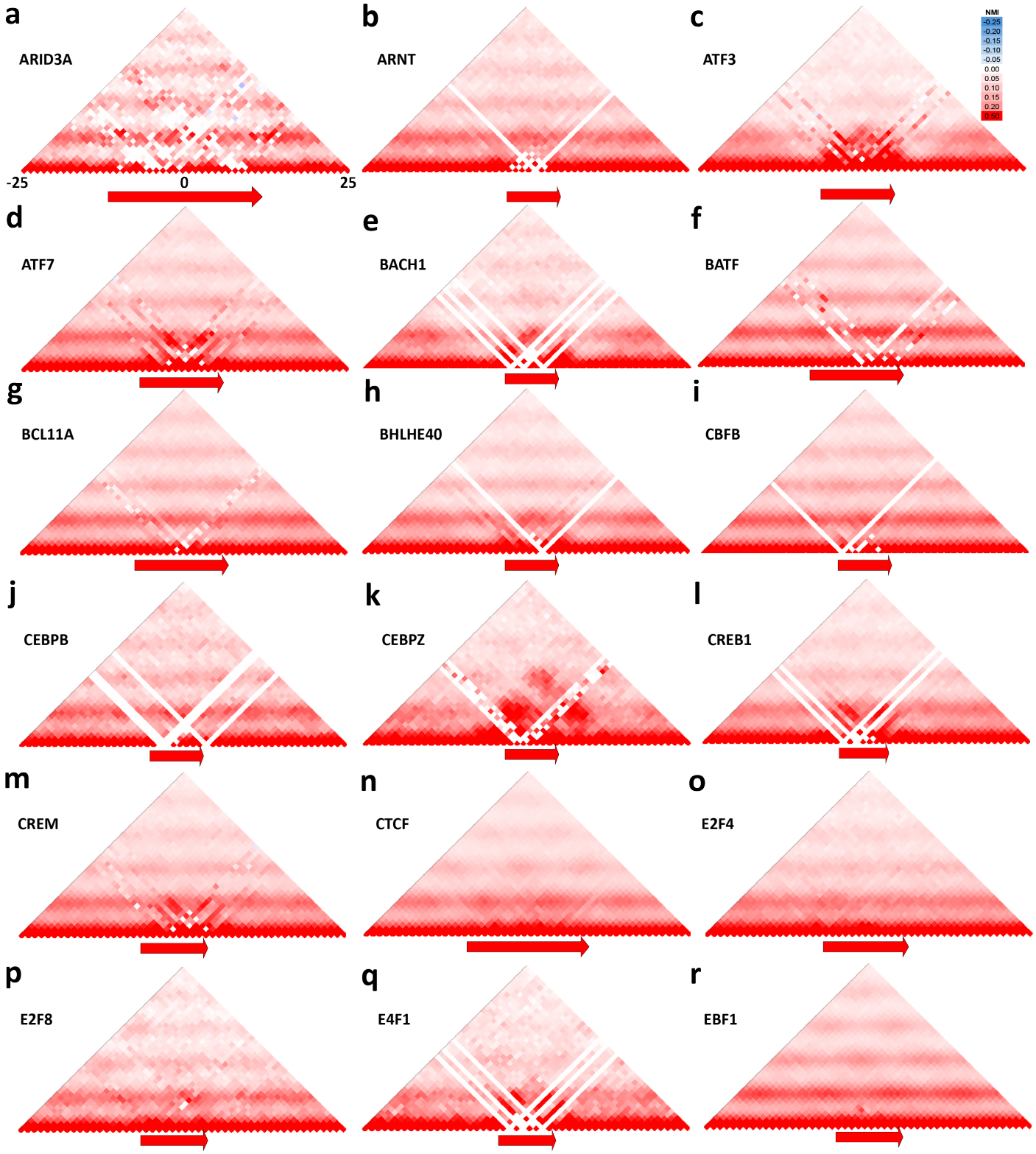
Supplementary Figure 66: Genome-wide footprint profiles for the USF2 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



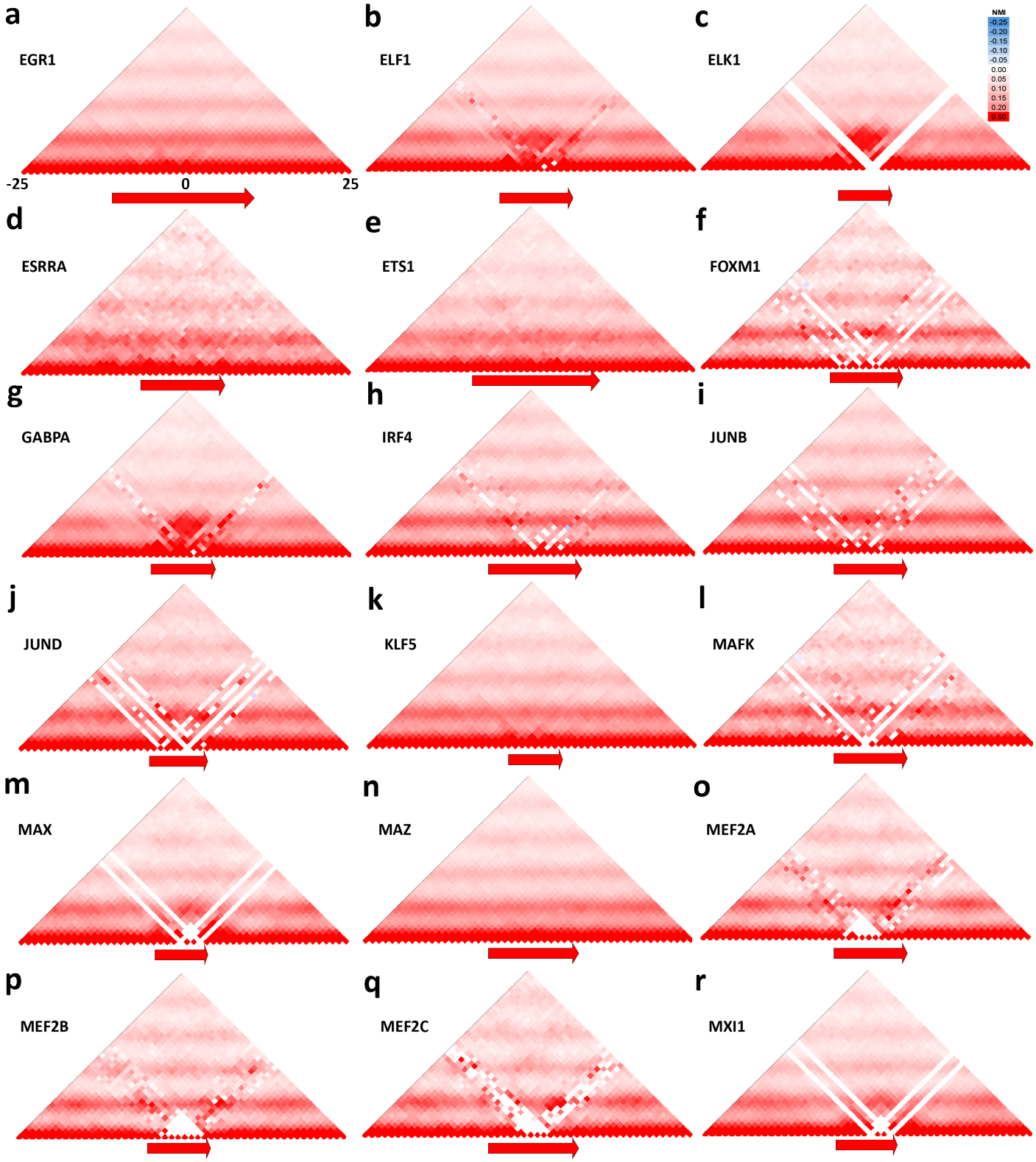
Supplementary Figure 67: Genome-wide footprint profiles for the YY1 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



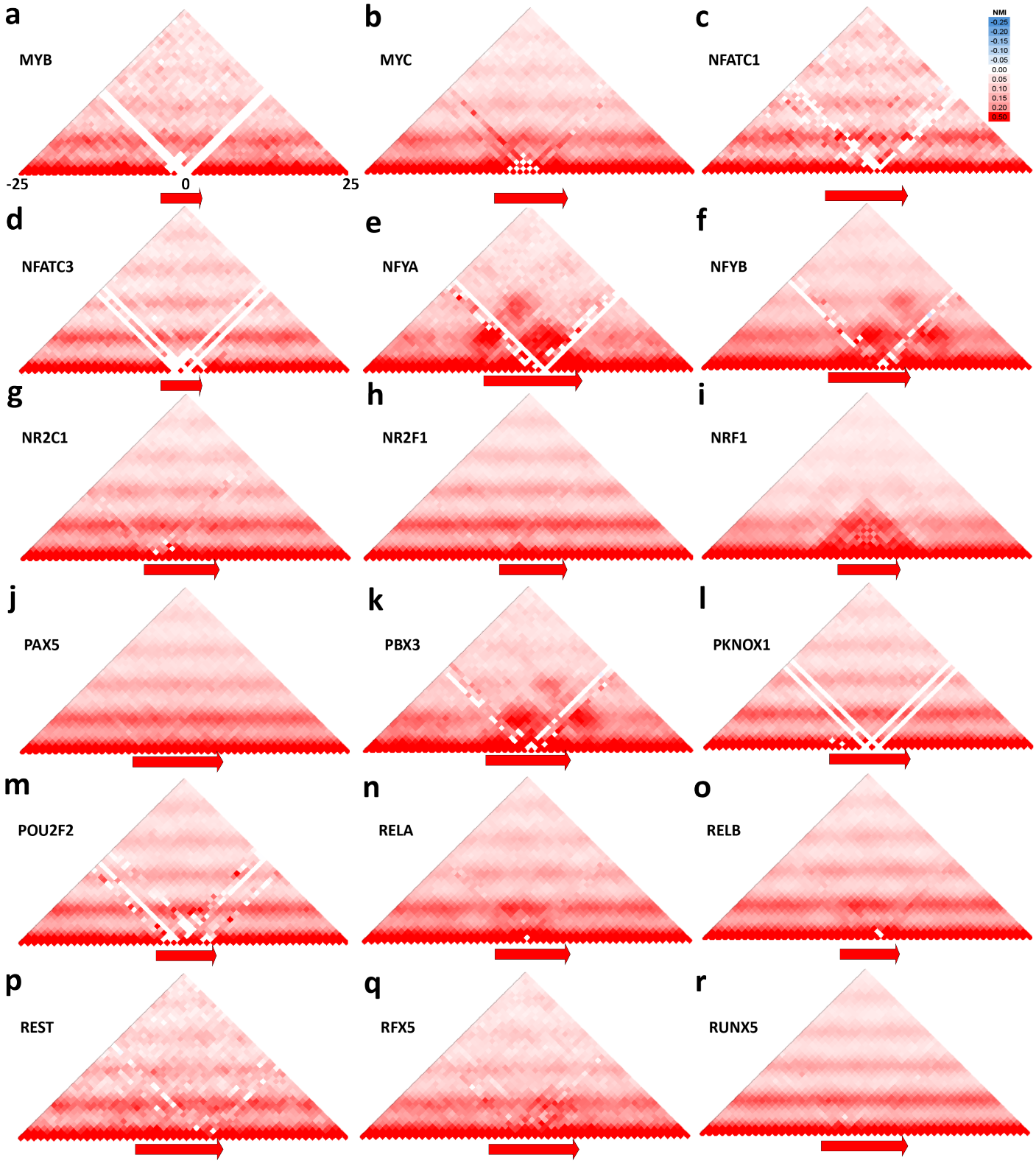
Supplementary Figure 68: Genome-wide footprint profiles for the ZNF384 transcription factor. GM12878 ssDNA-CseDa01-XL-SMF datasets are shown. (a) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks (ENCODE dataset ID indicated on top). Profiles are generated in a strand-aware way with respect to the motif orientation in the genome, and are shown separately for forward and reverse-strand reads. (b) Same as in (a) but zoomed in. The transcription factor sequence recognition motif is shown on top. (c) Genome-wide footprint profiles over the closest motif to ENCODE ChIP-seq peaks for the TF divided into quintiles based on ChIP-seq signal strength. (d) Same as in (c) but zoomed in.



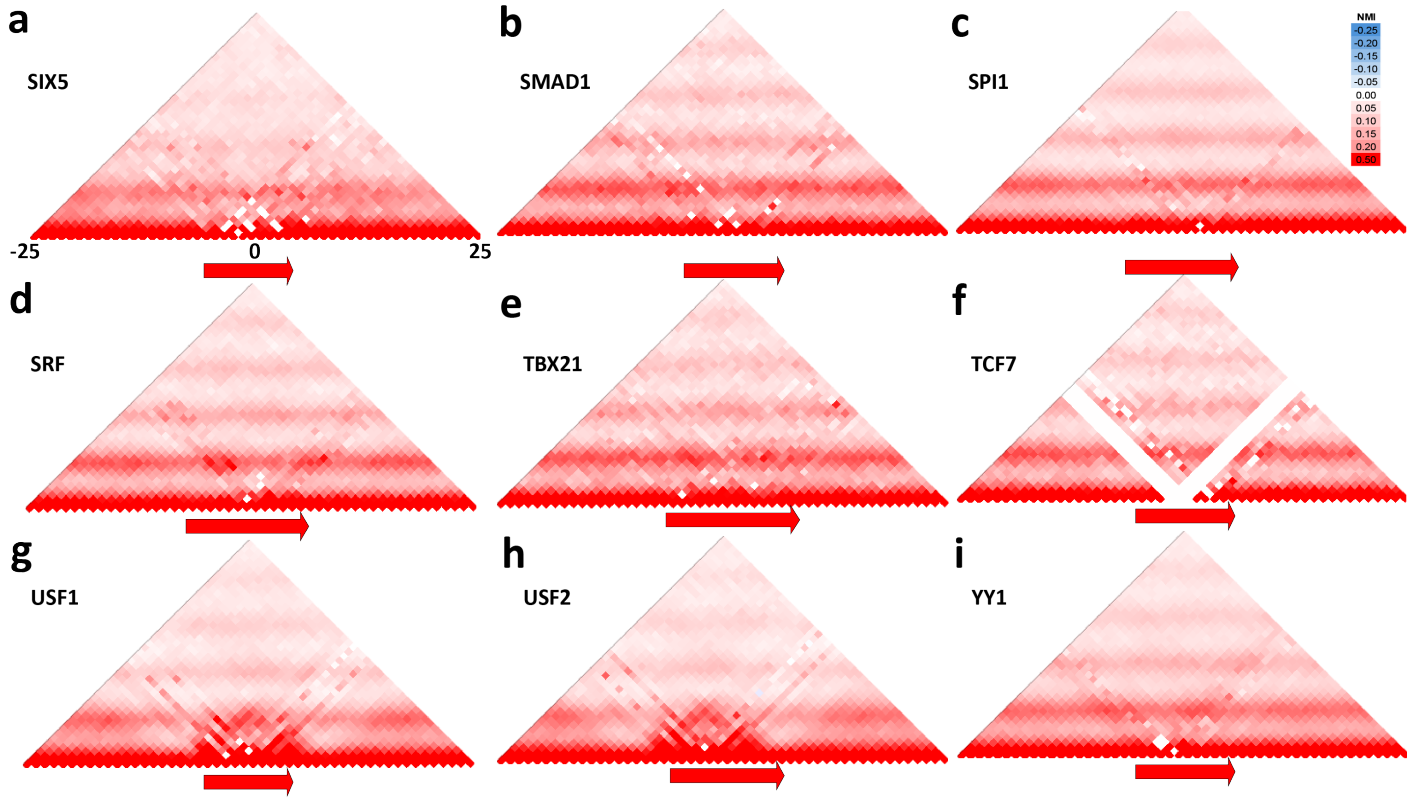
Supplementary Figure 69: Average co-accessibility profiles around occupied TF recognition motifs. Shown are average NMI values for each position in a 50-bp window around the TF recognition motif for the indicated TFs. Only the nearest motifs within ±100bp of ENCODE ChIP-seq peaks were used as input.



Supplementary Figure 70: Average co-accessibility profiles around occupied TF recognition motifs. Shown are average NMI values for each position in a 50-bp window around the TF recognition motif for the indicated TFs. Only the nearest motifs within ±100bp of ENCODE ChIP-seq peaks were used as input.



Supplementary Figure 71: Average co-accessibility profiles around occupied TF recognition motifs. Shown are average NMI values for each position in a 50-bp window around the TF recognition motif for the indicated TFs. Only the nearest motifs within ±100bp of ENCODE ChIP-seq peaks were used as input.



Supplementary Figure 72: Average co-accessibility profiles around occupied TF recognition motifs. Shown are average NMI values for each position in a 50-bp window around the TF recognition motif for the indicated TFs. Only the nearest motifs within ±100bp of ENCODE ChIP-seq peaks were used as input.