NAR Genomics and Bioinformatics, 2024, **6**, lqae012
https://doi.org/10.1093/nargab/lqae012
Advance access publication date: 7 February 2024
Methods Article

ChEC-seq2: an improved chromatin endogenous cleavage sequencing method and bioinformatic analysis pipeline for mapping in vivo protein–DNA interactions

Jake VanBelzen, Chengzhe Duan, Donna Garvey Brickner and Jason H. Brickner*

Department of Molecular Biosciences, Northwestern University, Evanston, Illinois, 60208, USA

*To whom correspondence should be addressed. Tel: +1 847 467 0210; Email: j-brickner@northwestern.edu

Abstract

Defining the *in vivo* DNA binding specificity of transcription factors (TFs) has relied nearly exclusively on chromatin immunoprecipitation (ChIP). While ChIP reveals TF binding patterns, its resolution is low. Higher resolution methods employing nucleases such as ChIP-exo, chromatin endogenous cleavage (ChEC-seq) and CUT&RUN resolve both TF occupancy and binding site protection. ChEC-seq, in which an endogenous TF is fused to micrococcal nuclease, requires neither fixation nor antibodies. However, the specificity of DNA cleavage during ChEC has been suggested to be lower than the specificity of the peaks identified by ChIP or ChIP-exo, perhaps reflecting non-specific binding of transcription factors to DNA. We have simplified the ChEC-seq protocol to minimize nuclease digestion while increasing the yield of cleaved DNA. ChEC-seq2 cleavage patterns were highly reproducible between replicates and with published ChEC-seq data. Combined with DoubleChEC, a new bioinformatic pipeline that removes non-specific cleavage sites, ChEC-seq2 identified high-confidence cleavage sites for three different yeast TFs that are strongly enriched for their known binding sites and adjacent to known target genes.

Introduction

Chromatin immunoprecipitation is a powerful and widely used method that captures *in vivo* protein–DNA interactions through formaldehyde fixation and immunoprecipitation (1,2). When combined with next generation sequencing (i.e. ChIP-seq), this method provides an unbiased approach to identifying sites genome-wide and, in the case of sequence-specific DNA binding proteins, to define their specificity (3). However, the resolution of ChIP-seq is limited by the length of the isolated DNA fragments - typically hundreds of base pairs long and much longer than the size of transcription factor binding sites. An alternative approach, ChIP-exo, whereby formaldehyde crosslinked chromatin is digested with exonucleases to reveal protected regions, provides much higher resolution (4). Likewise, the CUT&RUN method, whereby Protein A is fused to micrococcal nuclease (MNase) and chromatin from purified nuclei is treated with a primary antibody against the factor of interest and Protein A-MNase to digest nearby DNA, followed by immunoprecipitation, greatly improves the resolution of ChIP-seq (5). However, these latter methods are technically challenging, produce small amounts of material and are dependent on formaldehyde fixation and antibody quality.

An alternative method to map protein–DNA interactions *in vivo* is Chromatin Endogenous Cleavage (ChEC; (6,7)), which utilizes micrococcal nuclease (MNase) fused to the protein of interest. MNase activity is maximal at 10 mM Ca^{2+}, far greater than the ∼150 nM concentration found in the nucleus of budding yeast (8), so cells can tolerate the expression of MNase fusion proteins. However, upon permeabilization of cells in the presence of millimolar calcium, MNase rapidly cleaves DNA nearby (7). This method was adapted for next generation sequencing (ChEC-seq; (6)) by isolating small DNA fragments through negative selection with AMPure Beads, repairing DNA ends, and ligation of Illumina TruSeq adapters for sequencing (Figure 1A). Small DNA fragments arise from DNA that was cleaved at two adjacent sites, which are enriched for the binding site of the protein of interest (6). However, these small DNA fragments are low in abundance and therefore difficult to reproducibly purify, and their abundance is strongly influenced by the number of binding sites and the concentration of the protein of interest. Proteins that sparsely interact with the genome yield relatively few small DNA fragments, making ChEC-seq less feasible. Extending the cleavage time seems to increase non-specific cleavage rather than increasing the yield of specifically cleaved fragments (6,9).

To address this issue, we have developed ChEC-seq2, which accepts all of the digested genomic DNA as input, marks free ends through the ligation of a custom adapter, followed by Tn5 transposase-mediated library construction (Figure 1A). Library amplification with Nextera index primers amplifies DNA fragments flanked by heterologous adapters, and the final product is a DNA library compatible with Illumina sequencing. This method generates a much higher yield of DNA, minimizing the amplification needed to generate the library. This is particularly true for transcription factors that bind fewer sites. The cleavage patterns are highly reproducible and agree well with data from the original ChEC-seq method.

A significant concern with ChEC-seq is the specificity of the cleavage pattern. ChEC-seq with transcription factors (TFs) identifies cleavage both at specific sites with recognizable

Received: November 30, 2023. **Editorial Decision:** January 12, 2024. **Accepted:** January 24, 2024

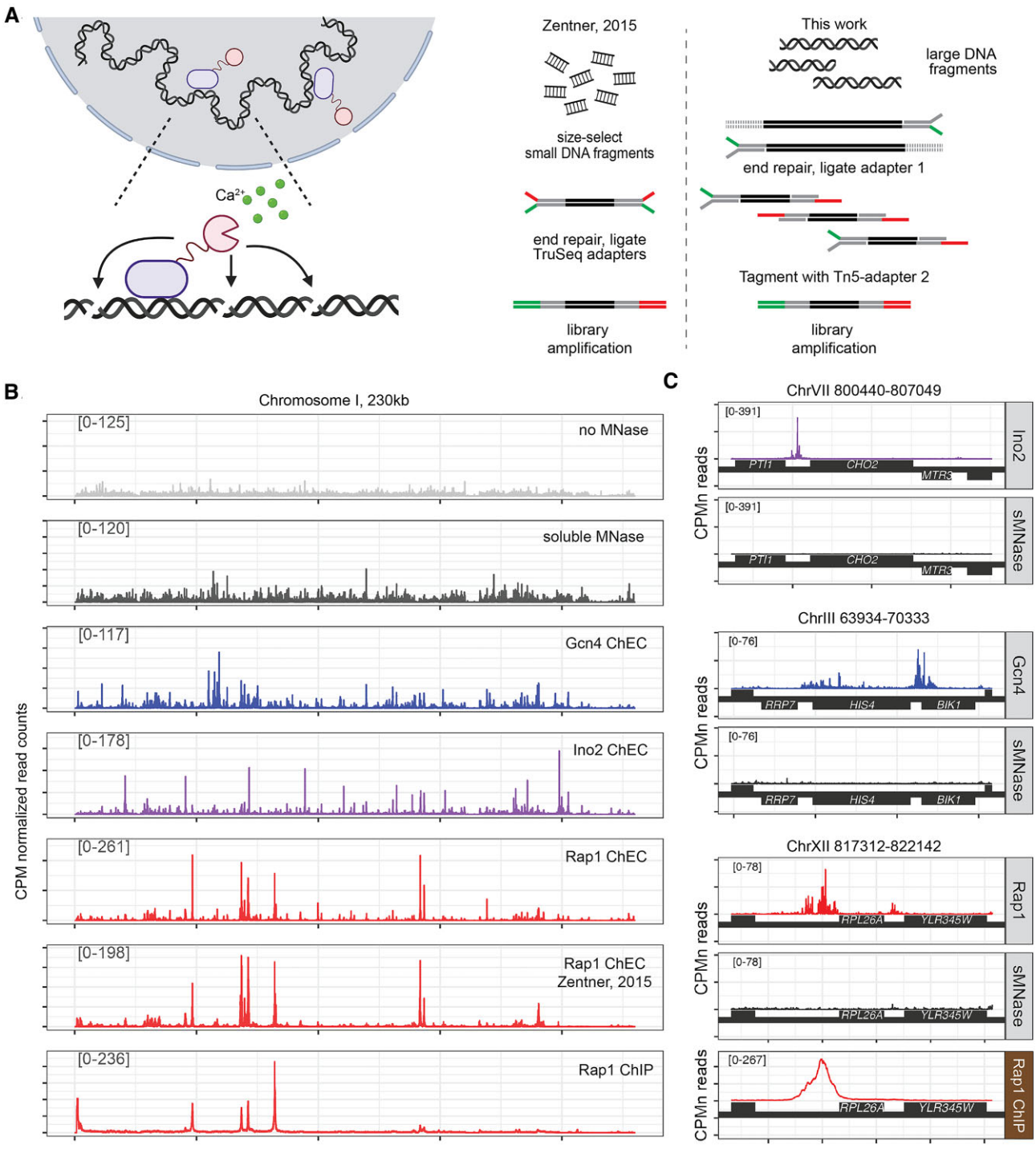

Figure 1. An improved method for chromatin endogenous cleavage (ChEC-seq2). (**A**) Schematic of modifications of ChEC-seq method. Left: a DNA-bound protein fused with micrococcal nuclease (MNase; pink). Upon permeabilization of the cells and addition of calcium, MNase cleaves adjacent DNA. Right: whereas the original method requires size-selection of small quantities of small fragments that have been cleaved twice, the current work uses mild digestion to generate large, cleaved fragments (Supplementary Figure 1). These large fragments are repaired and ligated to Read adapter 1 (green). Treatment with Tn5 transposase loaded with Read adapter 2 (red) results in small fragments that are amplified for single-end sequencing from adapter 1. (**B**) CPM-normalized read coverage from ChEC-seq experiments with Gcn4, Ino2, Rap1, soluble MNase (sMNase), and no MNase over *S. cerevisiae* Chromosome 1. The average of 3-biological replicates is shown. For ChEC-seq data, mapped reads were trimmed to the first base pair adjacent to the cleaved DNA. Published Rap1 ChEC-seq and Rap1 ChIP-seq datasets are included for comparison in the bottom two panels. (**C**) Representative targets of each transcription factor displayed as in (B) with soluble MNase (grey) for comparison, along with Rap1 ChIP-seq data (bottom panel).

motifs as well as many more non-specific sites (6), although cleavage of specific sites was observed preferentially at shorter cleavage times. This was interpreted to represent cleavage by both specifically bound TFs and non-specifically scanning TFs, although that conclusion has been challenged (10). Furthermore, non-specific cleavage of unprotected DNA is a source of background that must be accounted for to interpret ChEC data (11). We have developed a bioinformatic filtering approach called DoubleChEC that identifies high-confidence targets based on (i) the enrichment of cleavage compared with a negative control that evaluates chromatin accessibility (similar to previous work; (9,12–14), (ii) the structure of the cleavage pattern and (iii) the protection of the binding site. This filtering produces very robust identification of TF binding motifs and target genes and is not affected by the abundance or type of transcription factor. ChEC-seq2 with three different TFs identified known binding sites and target genes and agrees well with published ChIP-seq and ChIP-exo datasets. This method and bioinformatic pipeline provide a simple and robust method for mapping protein–DNA interactions *in vivo*.

Materials and methods

Yeast strains

All yeast strains were derived from BY4741 and are listed in Supplementary Table 2. A C-terminal MNase fusion was introduced to the protein of interest through transformation and homologous recombination of PCR-amplified DNA. Primers were designed with 50-bp of homology to the 3′ end of the coding sequence of interest. The 3× FLAG-MNase with a Kan^R marker was amplified from pGZ108 (6) and transformed into BY4741 as previously described. Successful transformation was confirmed by immunoblotting and PCR, followed by sequencing.

Lyophilized DNA oligonucleotides were resuspended in molecular-grade water to a concentration of 100 μM. For ligation, the following pair of oligonucleotides were annealed to produce the Y-adapter: Tn5ME-A (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′) and Y-Adapt-i5 R (5′-CTGTCTCTTATACACATCTTCAT AGTAATCATC-3′). For Tn5 Tagmentation, the following i7 oligonucleotides were annealed: Tn5ME-B (5′-GTCTC GTGGGCTCGGAGATGTGTATAAGAGACAG-3′) and Tn5MErev, (5′-PO_4-CTGTCTCTTATACACATCT-3′). Pairs of oligonucleotides were annealed as follows: 45 μl of each oligo (100 μM) was combined with 10 μl of 1 M potassium acetate, 300 mM HEPES, pH 7.5 in a 0.2 ml PCR tube. In a thermocycler, the mixture was heated to 95°C for 4 min, cooled 1°C/min until 50°C, incubated at 50°C for 5 min, and then cooled 1°C/minute until 4°C. Hybridized oligos were stored in 15 μl aliquots at –20°C.

Tn5 purification and adapter loading

Tn5 E54K L372P was purified as previously described (15). We found that Tn5 was sufficiently pure following purification on Ni^{2+}-chromatography and we therefore omitted the final gel filtration step. Purified Tn5 was aliquoted and stored at –80°C. Optimal Tn5 activity was determined by cleaving genomic DNA and assessing fragmentation using the Femto Pulse (Figure S2d) and resulting DNA libraries were confirmed to be of appropriate length for Illumina Sequencing by TapeStation (Figure S2e).

Tn5 was thawed on ice and 100 μl Tn5 was added to 10 μl i7 (45 μM) in a 1.7 ml tube and mixed by gently pipetting. The mixture was incubated at 23°C, mixing at 350 rpm for 45 min. Adapter-loaded Tn5 was stored at –20°C and used within 24 h.

Chromatin endogenous cleavage and library preparation

Chromatin cleavage, DNA purification, end-repair, ligation, Tagmentation and library amplification was performed as described in the detailed protocol in the Supplementary Materials. Libraries were sequenced on an Illumina Hiseq and the Northwestern NUSeq core facility using the 50-bp, single-end option.

Bioinformatic analysis

Quality control, trimming and mapping

Read quality and sequencer performance was evaluated with FASTQC. Reads were adapter and quality trimmed with Trimmomatic (16) using single-end settings. Bases at either end of a read were trimmed if base-call quality was <30, and only reads of length ≥25 bp were retained. Trimmed reads were mapped to the *Saccharomyces cerevisiae* genome (17), version R64-4-1 with Bowtie2 (18) and mapped reads with a MAPQ <10 were removed with Samtools (19).

DoubleChEC identification of high-confidence TF binding sites

For peak calling analysis, BAM files for three or more biological replicates of the TF-MNase and soluble MNase were read and trimmed to the first base pair. Unnormalized counts and normalized counts per million (CPMn) were tallied for each base pair in the yeast genome and the average CPMn values among replicates were calculated for each position. Next, mean CPMn values were smoothed using a sliding window of 3 and a step size of 2. Windows with CPMn values less than three times the genome average were filtered out. After this filtering, local maxima (windows with values greater than their immediate neighbors) were identified. Unnormalized reads were smoothed, retaining positions that were identified as local maxima, and inputted them in DESeq2 (version 1.36.0) to identify windows with values significantly higher than those in the soluble MNase control. Only TF-MNase peaks with a greater $\log_2$-fold change of 1.7 and an adjusted *P*-value <0.0001 over soluble MNase were retained. Finally, the peaks were filtered again to identify doublet peaks that are between 15 and 50 bp apart, which were merged to single peaks.

Gene Ontology term analysis

A list of genes whose 700 bp upstream regions overlap with peaks identified by the peak finder was input to enrichGO (20) to generate GO term plots based on biological functions. The 10 most significant GO terms with adjusted *P*-values <0.05 were plotted.

MEME analyses

The MEME Suite (version 5.5.1) was installed onto the local computer and two custom wrapper functions were written in R for the local bed2fasta and meme programs. These functions were then used to convert bed files, generated from peak calling, into FASTA files. These FASTA files were subsequently

to generate motif logos. Both bed2fasta and meme programs were run using their default parameter values.

Results

A detailed ChEC-seq2 protocol is provided as a Supplementary file. Partial MNase digestion was optimized for each fusion protein to produce large fragments (>5 kb; mean size ∼15 kb; Figures S2a–c). DNA was purified from cells, cleaved ends were repaired and ligated to an P5-compatible Y-adapter (containing the Read 1 adapter sequence; Figure 1A). Repaired DNA fragments were then fragmented with recombinant Tn5 transposase (15) loaded with Read 2 adapter (i.e. 'Tagmented', Figure 1A). As with Tagmentation in Nextera Library Preparation, this yielded a mixture of small DNA fragments flanked by adapter sequences (<1 kb; Figure S2d). DNA fragments bearing both Read 1 and Read 2 adapter regions were enriched by 15 cycles of PCR amplification using Nextera XT primers. The indexed libraries were pooled and sequenced using Illumina HiSeq 4000 single-end 50 bp reads. The reads are from the Read 1 direction, which represents the bases adjacent to the cleavage site.

ChEC-seq2 is highly reproducible and recapitulates the patterns produced by ChEC-seq

To evaluate ChEC-seq2, we fused MNase to three well-characterized budding yeast transcription factors, Rap1, Gcn4, and Ino2 at their endogenous loci. Rap1, Gcn4 and Ino2 represent a good test of the robustness of ChEC-seq2 because they bind to distinct, well-defined sequences (https://jaspar.elixir.no/; Figure S1; (21)) and have well-understood biological functions. Furthermore, they differ in abundance (Rap1 ∼4400 molecules per cell; Gcn4 <100 molecules per cell; Ino2 ∼784 molecules per cell (22)); and regulate different numbers of target genes (Rap1, 906 targets; Gcn4, 1257 targets, Ino2, 86 targets; Saccharomyces Genome Database). ChEC-seq2 was also performed with a strain expressing soluble, nuclear MNase (sMNase) expressed from the *CYC1* promoter to control for chromatin accessibility and the sequence bias of MNase digestion (23). Each TF cleaved at a different rate, likely reflecting its abundance and number of DNA-binding sites. Time points that produce clear, partial digestion were selected for both the TFs and sMNase from cells grown under matching conditions (Figure S2).

The first base of each sequencing read corresponds to the genomic location adjacent to a MNase cleavage event. Therefore, following read-mapping, genome-coverage data was calculated from only the first base of each read, which facilitates resolution of the fine structure of the cleavage peaks (Figure S3). Genome coverage data was then normalized to the library size (counts per million reads, CPM) and averaged from three biological replicates. The pattern of cleavage across chromosome I revealed patterns for each transcription factor (Figure 1b) that were distinct from each other and from libraries made from a strain lacking MNase (representing mechanical shearing) or expressing sMNase (Figure 1B). Comparison to previously published Rap1 ChEC-seq data revealed that ChEC-seq2 produces a highly similar cleavage pattern (Figure 1B). A high-quality Rap1 ChIP-seq dataset (24) showed a similar, but not identical, pattern of peaks to the ChEC-seq or ChEC-seq2 datasets (Figure 1B). Finally, the peaks of cleavage from ChEC are much narrower than those produced by ChIP-seq (Figure 1C).

DoubleChEC: a filtering approach to identify high-confidence transcription factor binding sites

ChEC-seq and ChEC-seq2 reveal cleavage peaks of different intensity and only a fraction of these sites are *bona fide* binding sites that resemble the consensus motif (6). The biological significance of such 'off-target' cleavage events is unclear. They may represent transcription factor scanning or chromatin accessibility to randomly diffusing DNA binding proteins. Regardless, identifying sites with high occupancy is critical to define the sequence specificity and biological role of DNA binding proteins. Previous work distinguished between high occupancy and low occupancy sites based on relative peak height and how rapidly the peaks appeared in time course experiments (6,9). Consistent with previous studies (9,12–14), we reasoned that true binding sites should be associated with peaks that are significantly larger than those produced by soluble MNase, which controls for DNA accessibility and MNase bias (23), and that such peaks should be found in pairs, flanking protected binding sites. To test this hypothesis, we measured the mean Rap1-MN or sMNase cleavage patterns over putative Rap1 binding sites (2366 instances of 5′-GNNNGGGTG-3′; Figure 2A). Indeed, Rap1-MN cleavage frequency produced strong peaks immediately adjacent to the binding site with a protected region of 8–10 bp over the query sequence (Figure 2A). The cleavage frequency of sMNase was generally low adjacent to these sequences but also showed protection of the query sequence (Figure 2A). Thus, *bona fide* TF binding sites should produce pairs of peaks flanking a protected sequence.

To identify high-confidence binding sites, we developed a workflow in R (https://www.r-project.org/) called DoubleChEC that: (i) smoothes the ChEC cleavage frequency using a sliding window of 3 bp, step size of 2 bp, (ii) identifies all local maxima, (iii) compares the cleavage of the protein of interest to sMNase over every peak using DESeq2 (parameters: >1.7 $\log_2$-fold change, adjusted *P*-value <1e-4; (25)) and (iv) identifies pairs of peaks (flowchart in Figure S4a). The optimal distance between pairs of peaks was determined by plotting the frequency of distances between adjacent TF peaks for Rap1, Gcn4 and Ino2. The distances between neighboring local maxima are dominated by peaks that are adjacent to each other (i.e. the peak is in the next neighboring sliding window; Figure S4b). However, for peaks that are significantly enriched over sMNase, a clear secondary maximum between 15 and 50 bp is evident (Figure 2B). Therefore, we selected pairs of peaks with local maxima between 15 and 50 bp apart (DoubleChEC is available at Github (https://github.com/jasonbrickner/DoubleChEC.git) and Dryad (https://doi.org/10.5061/dryad.c866t1gd5)).

For Rap1-MN, 178 843 local maxima were identified from ChEC cleavage frequencies. Filtering for those that were significantly enriched over sMNase reduced this number to 3352 peaks (Figure S4c) and increased average peak height by approximately 10-fold (Figure 2C). Finally, selecting adjacent peaks flanking a protected region produced 896 doublet peaks (Figure 2C), each of which was merged into a single high-confidence site. MEME analysis (26,27) revealed that neither the initial local maxima nor the sMNase filtered peaks were enriched for the Rap1 binding site (Figure 2C, bottom

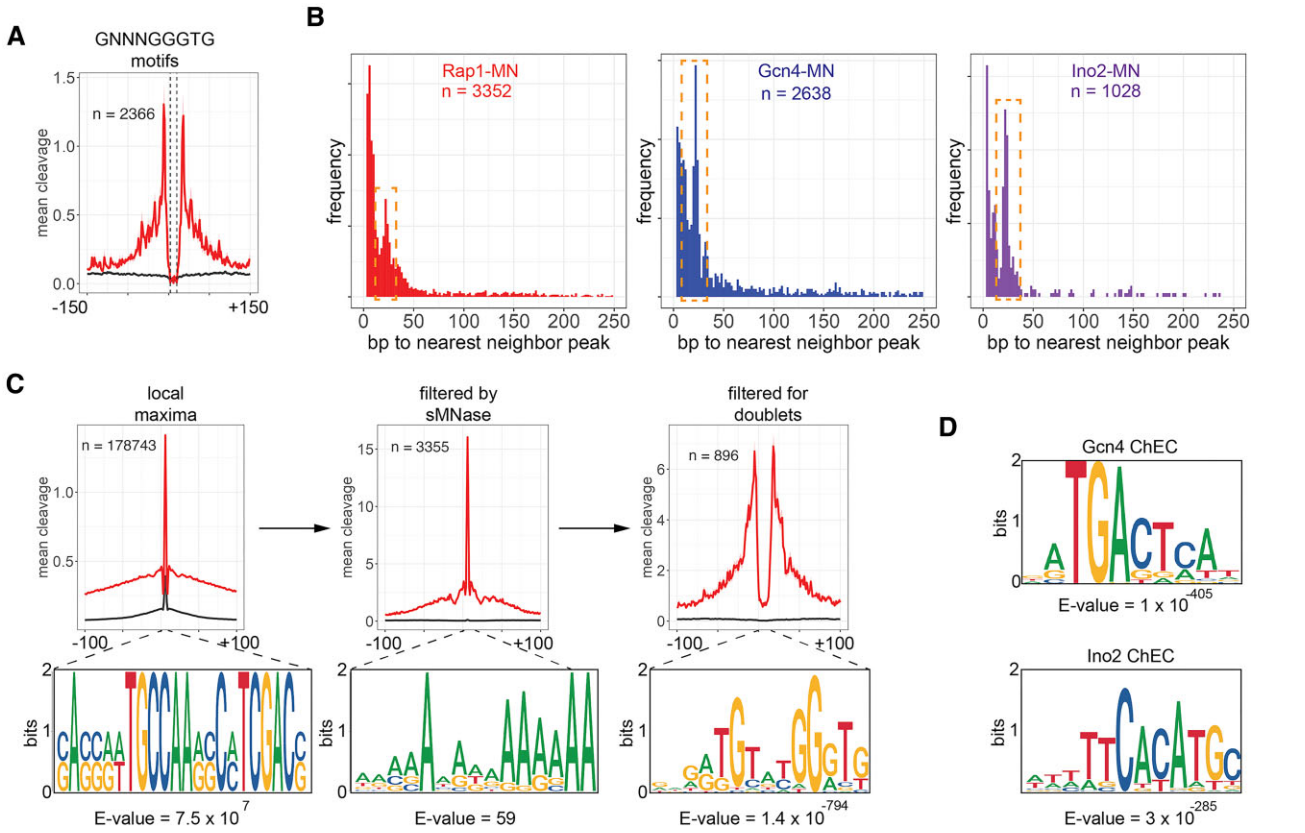

Figure 2. Improved analysis pipeline to identify high-confidence binding sites from ChEC-seq2 data.(**A**) Mean Rap1 ChEC-seq2 cleavage (red) and soluble MNase cleavage (black) over all 5′-GNNNGGGTG-3′ sites in the yeast genome. The average of three biological replicates is displayed. The dashed lines flank an 8bp protected region over the sequence. (**B**) Distribution of distances between adjacent peaks that were significantly enriched over sMNase. A secondary peak between 15bp and 50bp is highlighted with the dashed box. (**C**) Mean cleavage by Rap1 (red) and soluble MNase (black) over local maxima (left), peaks enriched over soluble MNase (middle) or for doublets (right). Each set of peaks (21 bp windows; overlapping peaks were merged) or peak pairs (merged) was analyzed for motif enrichment using MEME and the most enriched motif from each set is shown below. For the motif identified by the doublet peaks, 555 of 896 sites (62%) contained the motif. (**D**) High confidence sites for Gcn4 and Ino2 from ChEC-seq2 data were identified as in (B). Doublet peaks were analyzed by MEME and the top motifs for each Gcn4 (top) and Ino2 (bottom) are shown. For Gcn4, 430 of 627 sites (69%) contained the motif and for Ino2, 167 of 194 sites (86%) contained the motif.

panels). However, doublet peaks gave a strong enrichment for the consensus Rap1 binding site (Figure 2C). Furthermore, this approach identified 627 doublet peaks for Gcn4 (during histidine starvation) and 194 doublet peaks for Ino2 (during inositol starvation) that were strongly enriched for their respective consensus binding sites (Figure 2D). In contrast, MEME analysis of peaks (± 50 bp) identified from enrichment of cleavage by Gcn4-MN or Ino2-MN over sMNase alone did not identify these sequences (Figure S4d).

To confirm that sMNase cleavage reflects non-specific cleavage, we also mapped the cleavage of chromatin by two chromatin-associated proteins, H2A.Z and Prp20. H2A.Z is enriched in nucleosomes near promoters (28,29) and Prp20 is a general nucleosome binding protein (30). The pattern of cleavage by H2A.Z and Prp20 was qualitatively very similar to that produced by sMNase (Figure S5). Using either Prp20 or H2A.Z as a control instead of sMNase also identified strongly enriched motifs (Figure S5). Therefore, sMNase or chromatin-associated proteins adequately control for non-specific cleavage in ChEC-seq2.

To assess the ability of DoubleChEC to identify target genes, genes with high-confidence sites identified by ChEC-seq2 within 700 bp upstream were identified for Rap1-MN (691 genes), Gcn4-MN (487 genes) and Ino2-MN (203 genes). Gene ontology analysis (31) of these genes revealed that the top 10 most significantly enriched terms for each set matched the known biological functions of Rap1 (regulator of ribosome biosynthesis), Gcn4 (regulator of amino acid biosynthesis) and Ino2 (regulator of phospholipid biosynthesis). These genes were also compared with TF targets reported for 133 transcription factors by the Saccharomyces Genome Database (SGD; https://yeastmine.yeastgenome.org/yeastmine/begin.do; downloaded 09-25-2023; Supplementary Table S1; (32). Enrichment was assessed by Fisher's Exact test (Bonferroni corrected *P*-value). For Rap1-MN and Gcn4-MN, the overlap was most significant with their respective target sets from SGD (Figure 3b). Furthermore, Rap1-MN ChEC-seq2 target genes were strongly enriched for Fhl1 and Ifh1, transcription factors that co-regulate Rap1 targets (33). In contrast, Ino2 targets were only modestly enriched among the genes near Ino2-MN ChEC-seq2 sites (adjusted *P*-value = 0.03). However, the genes near Ino2-MN sites were enriched for targets of Ino4 (with which Ino2 heterodimerizes; adjusted p-value = 3×10^{-3}) and Opi1 (to which Ino2 binds directly; adjusted *P*-value = 7×10^{-4}; (34,35)). This suggests that the list of 86 targets on SGD is incomplete. Indeed, comparison of genes near Ino2-MN sites with genes near Ino2 ChIP-exo

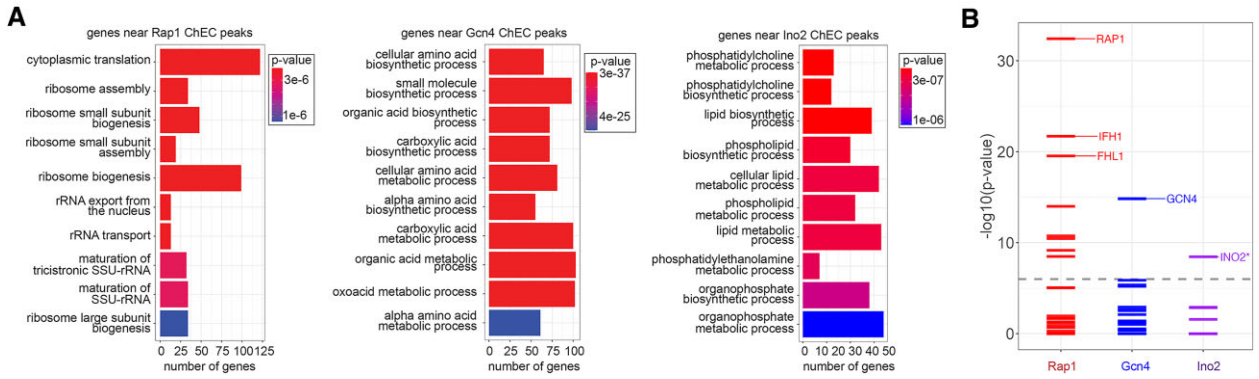

Figure 3. Target gene enrichment.(**A**) Genes with high-confidence sites from Rap1-MN (left; 691 genes), Gcn4-MN (middle; 487 genes), and Ino2-MN (right; 527) ChEC-seq2 within 700bp upstream of their start codon were analyzed for gene ontology (GO) term enrichment. The top 10 GO terms and their adjusted p-values are shown. (**B**) Genes adjacent to high-confidence sites from Rap1-MN, Gcn4-MN or Ino2-MN ChECseq2 were compared with reported target genes from 133 yeast transcription factors (www.yeastgenome.org; Supplementary Table S1). The significance of the overlap was assessed using a Fisher Exact test and the $-\log_{10}$(Bonferroni-adjusted *P*-value) for each comparison was plotted. The dashed line represents an adjusted *P*-value of 1×10^{-6}. The identities of the top three TFs with an adjusted *P*-value $<1 \times 10^{-6}$ is shown.

peaks (36), revealed highly significant overlap (adjusted *P*-value $= 4 \times 10^{-9}$; labeled INO2* in Figure 3B). Thus, ChEC-seq2 is an efficient and robust method for identifying TF binding sites, sequence motifs and target genes.

Unlike Rap1, the Gcn4 and Ino2 TFs function conditionally. Gcn4 protein levels are upregulated by starvation for amino acids (37). Ino2 is regulated by a repressor protein (Opi1), which dissociates upon starvation for inositol (34,35). However, Ino2 expression is also induced by inositol starvation, and the occupancy of Ino2 increases under activating conditions (38). The ChEC cleavage over high-confidence Gcn4 and Ino2 sites increased under inducing conditions, particularly for Gcn4 (Supplementary Figure S6). This suggests that TF occupancy impacts the amount of cleavage detected by ChEC-seq2.

Comparison of ChEC-seq2 with other methods

The high-confidence peaks identified for Rap1-MN by ChEC-seq2 were first compared with the Rap1-MN peaks identified using the original ChEC-seq protocol (6). The 7260 peaks identified from those data included nearly all of the 896 peaks identified by ChEC-seq2 (Figure 4A). However, the Rap1 motif (Figure 3A) was not enriched in this larger set of peaks, presumably because they include many non-specific peaks. When the data from 30 s of cleavage with Rap1-MN from Zentner *et al.* were filtered using the sMNase data presented here, followed by selection of doublet peaks, the non-overlapping peaks were largely removed (Figure 4B). The filtered set of 1667 high-confidence sites included 713 peaks from ChEC-seq2 reported here ($P = 3.5 \times 10^{-132}$, Fisher's Exact test) and were strongly enriched for the Rap1 motif (Figure 4b). Thus, ChEC-seq and ChEC-seq2 produce highly concordant data and applying our analysis pipeline to published ChEC-seq data identifies high-confidence sites enriched for the Rap1 sequence motif.

Finally, we compared the 896 high confidence Rap1 peak doublets identified by ChEC-seq2 with 1253 peaks identified by ChIP-seq (24) or 576 peaks identified by ChIP-exo (4). The Rap1 motif is strongly enriched in all three sets of Rap1 sites (Figure 4C), although the degree of enrichment (i.e. *E*-value) was best for the ChEC-seq2 peaks (Figure 4A). Mean cleavage by Rap1-MN from our ChEC-seq2 data peaked over both sites identified by ChIP-seq and sites identified by ChIP-exo (Figure 4D). The protection over the center of these peaks was greater for the peaks identified by ChIP-exo, suggesting that the precision of the sites identified by ChEC and ChIP-exo is higher than that of the peaks identified by ChIP-seq (Figure 4D). Genes near each of these sites ($n = 691$ for ChEC-seq2; $n = 1070$ for ChIP-seq and n = 525 for ChIP-exo) were compared for enrichment of the targets of 133 yeast TFs. All three sets showed strong enrichment for Rap1 as well as Fhl1/Ifh1 (Figure 4E). The target genes were strongly overlapping (Figure 4F). Of the 691 Rap1 targets identified by ChEC-seq2, 83% were also identified by ChIP-seq ($P = 3 \times 10^{-61}$, Fisher's Exact test, Bonferroni corrected) and 54% were identified by ChIP-exo ($P = 7 \times 10^{-80}$).

Discussion

Here we present a modified ChEC-seq protocol and analysis pipeline that provides excellent mapping of *in vivo* transcription factor binding sites. Similar to ChIP-seq and ChIP-exo, ChEC-seq2 identified consensus sequence motifs, target genes and biological roles for three different transcription factors in budding yeast that have distinct protein levels, regulation and DNA binding domains.

As an alternative to ChIP-seq or ChIP-exo, ChEC-seq2 has several advantages. ChEC-seq2 avoids fixation, which can affect chromatin solubility and may not be uniform for all binding sites. Also, ChEC-seq2 does not require antibody-based affinity purification, greatly increasing the yield of starting material and avoiding problems associated with variable accessibility of the epitope. Like ChIP-exo, ChEC-seq2 produces high resolution footprints of DNA binding, but unlike ChIP-exo, cleavage occurs in permeabilized cells, rather than after affinity purification, requires neither fixation nor antibodies and the downstream processing is simpler.

ChEC-seq2 has a disadvantage over ChIP-seq and ChIP-exo: non-specific cleavage of DNA at 'off-target' sites throughout the genome. Such sites tend to be quite reproducible, enriched in nucleosome-depleted regions upstream and downstream of genes. Including these cleavage sites identifies a large number of non-biological sites (*i.e.*, sites without an obvious DNA motif and that are not adjacent to

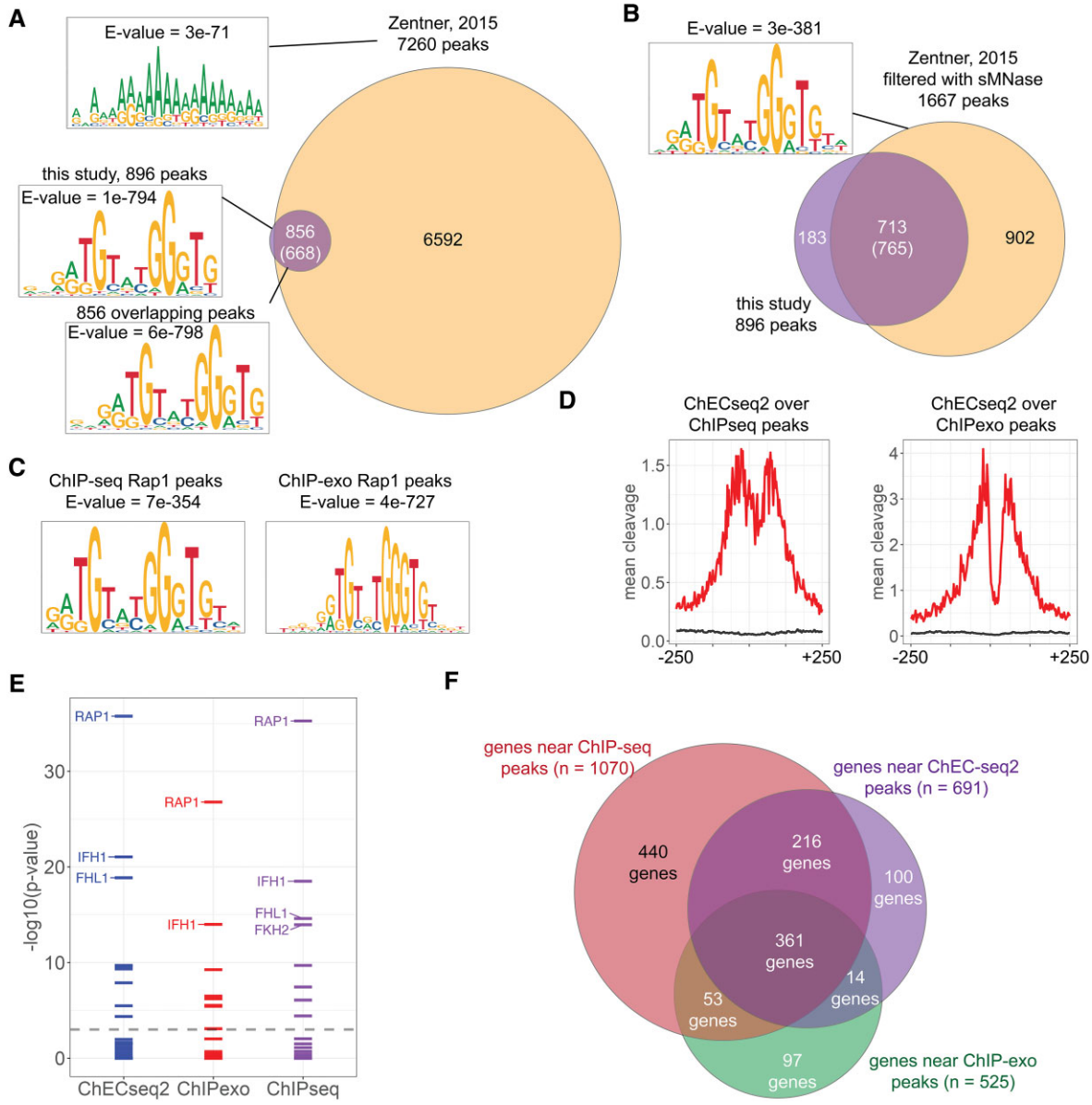

Figure 4. Comparison of ChEC-seq2 to other methods.(**A**) Overlap between 7260 Rap1 sites identified in Zentner *et al.* and 896 sites identified in the current work. Results of MEME analysis against all 7260, all 896 or the overlapping sites are shown. (**B**) High confidence Rap1 sites identified from 30s cleavage from Zentner *et al.* data, filtered using our peak finder (sMNase control from the current study). The top motif from the 1667 sites is shown. (**C**) Top motifs identified from Rap1 ChIP-seq and ChIP-exo sites. (**D**) Mean Rap1-MN cleavage over peaks identified by ChIP-seq and ChIP-exo. **(e)** Genes near ChEC-seq2, ChIP-seq and ChIP-exo peaks were compared for overlap with the targets of 133 TFs from SGD. The $-\log_{10}$(adjusted *P*-value) was plotted and the top three TFs with an adjusted *P*-value <1e-6 were labeled. (**F**) Overlap between genes adjacent to ChEC-seq2 sites, ChIP-seq peaks and ChIP-exo peaks.

known target genes). Selecting larger cleavage peaks or favoring sites that are produced more rapidly only partially addresses this issue (6,9). However, consistent with previous studies (9,12–14,39), the cleavage pattern by soluble nuclear MNase or other chromatin-associated proteins controls for this artifact, allowing these sites to be removed. And together with selecting pairs of peaks flanking protected DNA, ChEC-seq2 identifies high-confidence sites enriched for consensus motifs.

While the sites and target genes identified by ChEC and ChIP are quite similar, some differences are apparent. The source of these differences may be biological (i.e. differences between strains or growth conditions) or technical (i.e. the ability to recover a particular site by ChIP or ChEC). It seems unlikely that variability of ChEC accounts for these differences; the agreement between ChEC replicates is very high. But it is possible that ChEC fails to recover some sites because of local DNA accessibility. For example, if one of the two doublet peaks is very small because of adjacent nucleosomes or DNA binding proteins, or if the mode of DNA binding by a TF favors cleavage on one side of the binding site, it is possible such sites will be removed during filtering by DoubleChEC. Indeed, it is possible that the expectation of doublet peaks may not be appropriate for some transcription factors. Likewise, it is possible that differences in fixation efficiency or epitope availability could reduce recovery of particular sites by ChIP.

The two techniques together seem to capture the most complete set of target sites for each transcription factor.

TFs represent an excellent test case for techniques such as ChIP and ChEC because they bind directly to well-defined DNA sequences in enhancer regions. However, many proteins occupy the genome less specifically. For example, co-activators or co-repressors frequently associate indirectly with enhancers through diverse TFs. This creates a challenge for cross-linking based methods because the antigen is unlikely to be directly cross-linked to DNA. ChEC-seq has been used to map such factors (14,39,40). It seems likely that such factors will enhance the cleavage of nearby DNA and that the significance of this cleavage can be assessed by comparison to sMNase. Although pairs of peaks are expected, the precise nature of the cleavage pattern, in terms of the distance between peaks and the degree of foot-printing, may differ from TFs. Therefore, it will be important to validate the analytical pipeline using genetics to ensure that binding sites are biologically meaningful.

Data availability

Sequencing data has been deposited in the Gene Expression Omnibus at the National Center for Biotechnology Information and can be retrieved with accession number GSE246951. The DoubleChEC peak finder R scripts are available at Github (https://github.com/jasonbrickner/DoubleChEC.git) and Dryad (https://doi.org/10.5061/dryad.c866t1gd5).

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors thank Professor Erik Andersen, Professor Shelby Blythe and members of the Brickner laboratory for helpful comments on the manuscript. We thank Gabe Zentner for helpful discussions and for sharing MNase-tagging plasmids. The Tn5 expression plasmid was the kind gift of Lars Steinmetz. J.V. and D.G.B. performed the experiments; J.V., J.H.B. and C.D. analyzed the data; J.V. and J.H.B. wrote the manuscript.

Funding

J.V. was supported by a National Science Foundation Graduate Fellowship; this work was supported by National Institute of General Medical Sciences [R35GM136419 to J.H.B.].

Conflict of interest statement

None declared.

References

1. Gilmour,D.S. and Lis,J.T. (1984) Detecting protein–DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc. Natl. Acad. Sci. U.S.A.*, **81**, 4275–4279.
2. Solomon,M.J., Larsen,P.L. and Varshavsky,A. (1988) Mapping protein–DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, **53**, 937–947.
3. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
4. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
5. Skene,P.J. and Henikoff,S. (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, **6**, e21856.
6. Zentner,G.E., Kasinathan,S., Xin,B., Rohs,R. and Henikoff,S. (2015) ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nat. Commun.*, **6**, 8733.
7. Schmid,M., Durussel,T. and Laemmli,U.K. (2004) ChIC and ChEC; genomic mapping of chromatin proteins. *Mol. Cell*, **16**, 147–157.
8. Ohya,Y., Umemoto,N., Tanida,I., Ohta,A., Iida,H. and Anraku,Y. (1991) Calcium-sensitive cls mutants of Saccharomyces cerevisiae showing a Pet- phenotype are ascribable to defects of vacuolar membrane H(+)-ATPase activity., *J. Biol. Chem.*, **266**, 13971–13977.
9. Zentner,G.E., Policastro,R.A. and Henikoff,S. (2021) ChEC-seq produces robust and specific maps of transcriptional regulators. bioRxiv doi: https://doi.org/10.1101/2021.02.11.430831, 12 February 2021, preprint: not peer reviewed.
10. Rossi,M.J., Lai,W.K.M. and Pugh,B.F. (2017) Correspondence: DNA shape is insufficient to explain binding. *Nat. Commun.*, **8**, 15643.
11. Mittal,C., Rossi,M.J. and Pugh,B.F. (2021) High similarity among ChEC-seq datasets. bioRxiv doi: https://doi.org/10.1101/2021.02.04.429774, 05 February 2021, preprint: not peer reviewed.
12. Bruzzone,M.J., Albert,B., Hafner,L., Kubik,S., Lezaja,A., Mattarocci,S. and Shore,D. (2021) ChEC-seq: a robust method to identify protein–DNA interactions genome-wide. bioRxiv doi: https://doi.org/10.1101/2021.02.18.431798, 18 February 2021, preprint: not peer reviewed.
13. Donczew,R., Lalou,A., Devys,D., Tora,L. and Hahn,S. (2021) An improved ChEC-seq method accurately maps the genome-wide binding of transcription coactivators and sequence-specific transcription factors. bioRxiv doi: https://doi.org/10.1101/2021.02.12.430999, 19 February 2021, preprint: not peer reviewed.
14. Kubik,S., O'Duibhir,E., de Jonge,W.J., Mattarocci,S., Albert,B., Falcone,J.-L., Bruzzone,M.J., Holstege,F.C.P. and Shore,D. (2018) Sequence-directed action of RSC remodeler and general regulatory factors modulates +1 nucleosome position to facilitate transcription. *Mol. Cell*, **71**, 89–102.
15. Hennig,B.P., Velten,L., Racke,I., Tu,C.S., Thoms,M., Rybin,V., Besir,H., Remans,K. and Steinmetz,L.M. (2017) Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3: Genes Genomes Genet*, **8**, 79–89.
16. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
17. Engel,S.R., Dietrich,F.S., Fisk,D.G., Binkley,G., Balakrishnan,R., Costanzo,M.C., Dwight,S.S., Hitz,B.C., Karra,K., Nash,R.S., *et al.* (2013) The Reference Genome Sequence of Saccharomyces cerevisiae: then and Now. *G3: Genes Genomes Genet*, **4**, 389–398.
18. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
19. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
20. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L., *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov*, **2**, 100141.

21. Castro-Mondragon,J.A., Riudavets-Puig,R., Rauluseviciute,I., Berhanu Lemma,R., Turchi,L., Blanc-Mathieu,R., Lucas,J., Boddie,P., Khan,A., Manosalva Pérez,N., *et al.* (2021) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **50**, D165–D173.
22. Ghaemmaghami,S., Huh,W.K., Bower,K., Howson,R.W., Belle,A., Dephoure,N., O'Shea,E.K. and Weissman,J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.
23. Horz,W. and Werner,A. (1981) Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic Acids Res.*, **12**, 2643–2658.
24. Bondra,E.R. and Rine,J. (2023) Context dependent function of the transcriptional regulator Rap1 in gene silencing and activation in Saccharomyces cerevisiae. bioRxiv doi: https://doi.org/10.1101/2023.05.08.539937, 11 May 2023, preprint: not peer reviewed.
25. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
26. Bailey,T.L. (2003) Discovering Novel Sequence Motifs with MEME. *Curr. Protoc. Bioinform.*, https://doi.org/10.1002/0471250953.bi0204s00.
27. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol*, **16**, 284–287.
28. Raisner,R.M., Hartley,P.D., Meneghini,M.D., Bao,M.Z., Liu,C.L., Schreiber,S.L., R,O.J.o. and Madhani,H.D. (2005) Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. *Cell*, **123**, 233–248.
29. Guillemette,B., Bataille,A.R., Gevry,N., Adam,M., Blanchette,M., Robert,F. and Gaudreau,L. (2005) Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol.*, **3**, e384.
30. Frasch,M. (1991) The maternally expressed Drosophila gene encoding the chromatin-binding protein BJ1 is a homolog of the vertebrate gene Regulator of Chromatin Condensation, RCC1., *EMBO J.*, **10**, 1225–1236.
31. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
32. Balakrishnan,R., Park,J., Karra,K., Hitz,B.C., Binkley,G., Hong,E.L., Sullivan,J., Micklem,G. and Cherry,J.M. (2012) YeastMine—an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. *Database*, **2012**, bar062.
33. Shore,D., Zencir,S. and Albert,B. (2021) Transcriptional control of ribosome biogenesis in yeast: links to growth and stress signals. *Biochem. Soc. Trans.*, **49**, 1589–1599.
34. Wagner,C., Dietz,M., Wittmann,J., Albrecht,A. and Schuller,H.J. (2001) The negative regulator Opi1 of phospholipid biosynthesis in yeast contacts the pleiotropic repressor Sin3 and the transcriptional activator Ino2. *Mol. Microbiol.*, **41**, 155–166.
35. Heyken,W.T., Repenning,A., Kumme,J. and Schuller,H.J. (2005) Constitutive expression of yeast phospholipid biosynthetic genes by variants of Ino2 activator defective for interaction with Opi1 repressor. *Mol. Microbiol.*, **56**, 696–707.
36. Bergenholm,D., Liu,G., Holland,P. and Nielsen,J. (2018) Reconstruction of a global transcriptional regulatory network for control of lipid metabolism in yeast by using chromatin immunoprecipitation with lambda exonuclease digestion. *MSystems*, **3**, e00215–e17.
37. Hinnebusch,A.G. (1985) A hierarchy of trans-acting factors modulates translation of an activator of amino acid biosynthetic genes in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **5**, 2349–2360.
38. Brickner,J.H. and Walter,P. (2004) Gene recruitment of the activated INO1 locus to the nuclear membrane. *PLoS Biol.*, **2**, e342.
39. Donczew,R., Warfield,L., Pacheco,D., Erijman,A. and Hahn,S. (2020) Two roles for the yeast transcription coactivator SAGA and a set of genes redundantly regulated by TFIID and SAGA. *eLife*, **9**, e50109.
40. Bruzzone,M.J., Grünberg,S., Kubik,S., Zentner,G.E. and Shore,D. (2018) Distinct patterns of histone acetyltransferase and Mediator deployment at yeast protein-coding genes. *Gene. Dev.*, **32**, 1252–1265.

Received: November 30, 2023. **Editorial Decision:** January 12, 2024. **Accepted:** January 24, 2024